

May 2, 2022

DRAFT

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.*

May 2, 2022
DRAFT

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.

May 2, 2022
DRAFT

Contents

1	Introduction	1
2	Background	3
2.1	Urban Big Data	3
2.2	Spatial-Temporal Data	4
2.3	Work Zones	6
2.3.1	Building Roadside Work Zone Understanding	8
2.3.2	Challenges in Work Zones Understanding	9
2.3.3	Hierarchical Scene Understanding of Work Zones	10
2.4	Object Detection	11
3	Structure of Bus Data	13
3.1	Data Source and Collection Process	13
3.2	Raw Data Definition	14
3.3	Data Cleaning	15
3.4	Align Images using Spatial Coordinates	16
3.4.1	Generating a Baseline Trajectory	17
3.4.2	Downsampling the Baseline Trajectory	18
3.4.3	Assigning Data Points to the Baseline Trajectory	18
4	Work Zone Detection	21
4.1	Work Zone Related Datasets	21
4.1.1	Manually Labeled Dataset	21
4.1.2	NuScenes	22
4.2	Model Architecture	23
5	Results	27
5.1	Sample Outputs	27
5.1.1	Single Image Results	27
5.1.2	Spatial Changes	27
5.1.3	Temporal Changes	27
5.1.4	Spatial-Temporal Visualization	27
5.2	Analysis	27
5.3	Spatio-Temporal Attributes	27

5.3.1 Variations across Time	28
6 Conclusion	29
Bibliography	31

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 [17]. 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

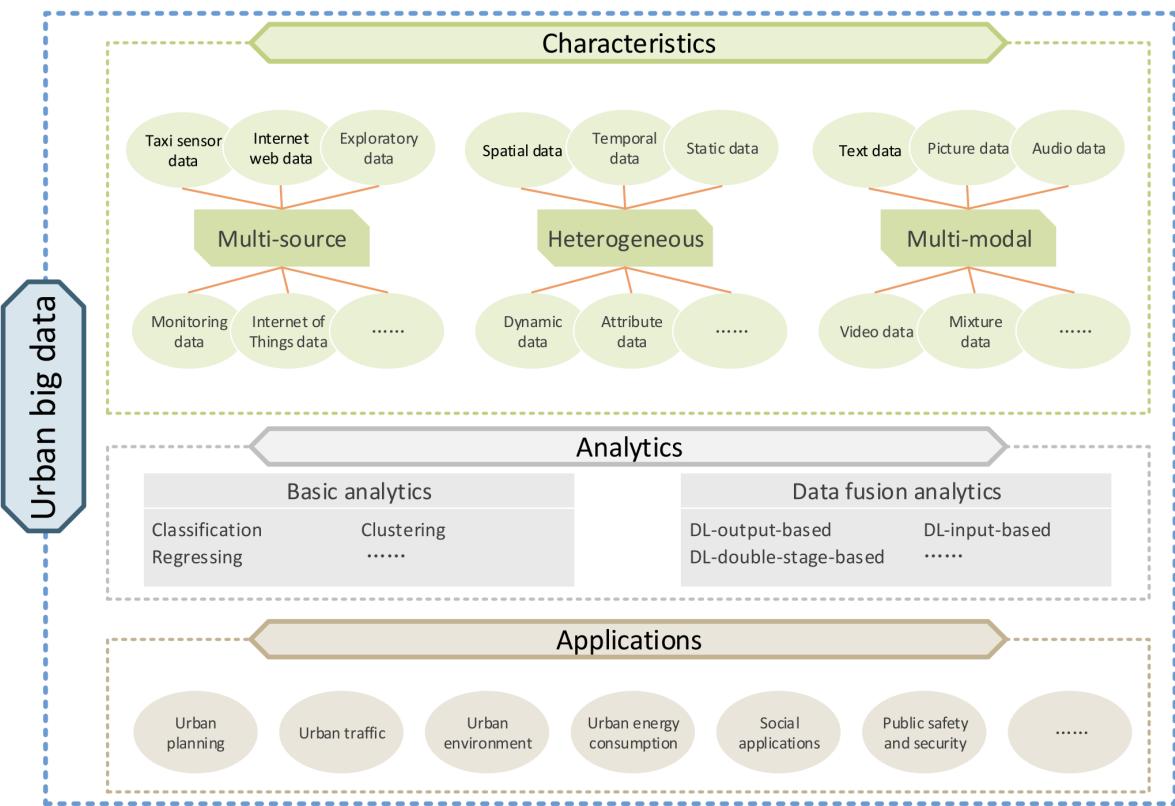


Figure 2.1: This diagram [17] breaks down the definition of urban big data by its characteristics, methods of analysis, and its downstream applications.

Urban big data refers to a combination of structured or unstructured data collected from various urban environments. Zhang et al. [32] 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 [20] notes, urban big data is very complex, and we only extract a small part of its knowledge.

From Figure 2.1, we see that [17] 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. [19] 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 [17]. Lastly, we see that urban big data is multimodal and may include visual, textual, or numeric data. Yi et al. [30] 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 [5], 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 [5]. For example, [5] gave the example of air pollution measures. Air pollution levels and name of location are one example of non-spatial-temporal attributes. [14] 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. [23] 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.

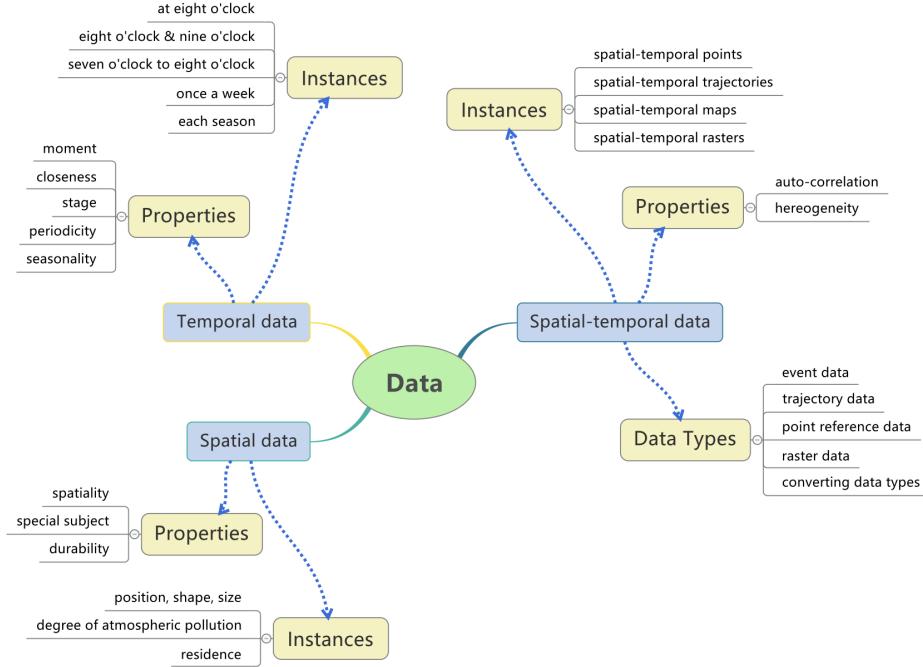


Figure 2.2: Based on the spatial and temporal dimensions, a given data can be divided into data with temporal attribute, data with spatial attribute, and data with both temporal and spatial attributes (spatial-temporal data). [17]

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 [5].

Spatial-temporal data mining (STDM) [11] 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 [11] 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 [3], 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 [2]. 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. [15] 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.

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

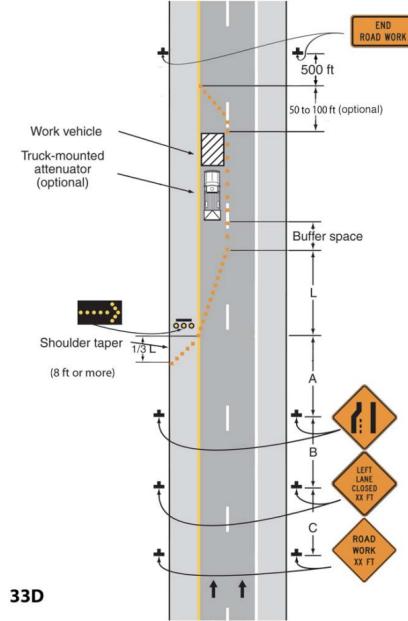


Figure 2.3: An illustration of a short-term stationary setup. The distances A, B, C should be proportional to the work zone speed limit. TTC Signs are placed some distance prior to the actual work zone. Delinators are placed surrounding the boundaries of the work zone. [2]

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 [26] 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 [31] and Cityscapes [9] contain minimal to no labeled instances of road construction object data. NuScenes [6] contains labels for some work zone delineators, 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 [6] contains a lower number of construction-related instances. This presents the challenge of class imbalance in the NuScenes dataset [6].

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, NuScenes [6] 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 [28] 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.

2.3.3 Hierarchical Scene Understanding of Work Zones

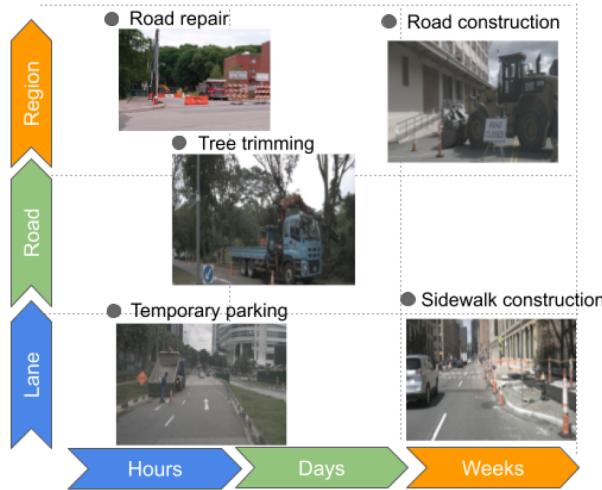


Figure 2.4: Based on the spatial and temporal dimensions, a given data can be divided into data with temporal attribute, data with spatial attribute, and data with both temporal and spatial attributes (spatial-temporal data). [26]

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 [21]. 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 [26], 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. [8] 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 gain an understanding of the scene. Spatial and geometric relationships between the detected objects can provide information about the image category.

2.4 Object Detection

In this work, we focus on exploring bottom-up approaches for the scene understanding task. As previously mentioned, bottom-up approaches uses detectors to identify key points or key objects, then extract semantics about the scene from the the set of key points. In the following sections, we provide an overview of previous object detection approaches, challenges in object detection, and how it relates to the task of work zone 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 [10], 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 [24] 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 such as ResNet-101 [13] 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 [29].

As mentioned above, one of the largest challenges 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 to expand the annotations toward some of the work zone labels. For example, in the most recent release of Argoverse 2.0 [28], the authors added annotations for the construction cone and the construction barrel. Similarly, [6] also contains annotations for construction cones and barricades. However, annotations across datasets cannot be easily combined in training. [33] 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.

We noticed that most objects appear at work zones at different frequencies and counts. For example, channelizing devices such as construction cones and vertical panels appear in 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 the quantities of each object, which introduces a class imbalance in 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.

Within the domain of work zone objects there lie many challenges that may be observed. In order for these objects to be salient, many of these objects 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 the labels between work zones.

Another common example is when our objects of interest are subclasses of common objects. For instance, the class "worker" is a specific instance of "person". Some of the instances of a "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 in the situation where there are less instances of our interest.

Chapter 3

Structure of Bus Data

In this work, we are interested in exploiting spatial and temporal attributes within large-scale urban visual data to detect of interest. Specifically, we structure unprocessed bus data based on their spatial and temporal attributes to mine patterns about work zones. As mentioned above, visual data from public transit systems is that a good source for studying work zones for many reasons. Bus data traverse fixed and repeated routes. As such, this allows us to track changes across specific spatial regions of the route over time. This property is especially beneficial for observers who observe the patterns of work zones, since long-term work zones span a long period of time at a fixed location.

Previous works have used bus data for statistical analysis [22], [25], [27]. However, most of these works focus on attributes directly related to the transportation task, such as the number of stops visited, the time elapsed at each stop, the number of pedestrians on board, etc. A limited amount of work places the focus on using visual data collected from outfacing cameras from the bus to perform an analysis of its surroundings. In this work, we will use data collected through the BusEdge [29] 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

The data used in this work were collected from a bus traveling between downtown Pittsburgh, PA and Washington County, PA. Two round trips are collected every working day. The route traversed is drawn in Figure 3.2. The inbound trip starts at the intersection of Wylie and Allison Avenues and ends at East Busway in downtown Pittsburgh. The outbound trip starts on the East Busway in downtown Pittsburgh and ends at the Pittsburgh Transit Center.

In total, five cameras are 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 interior camera placed behind the windshield of the bus. The technical specifications of the cameras are listed in Table 3.1 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 built-in antennas, one of them being a GPS antenna. The

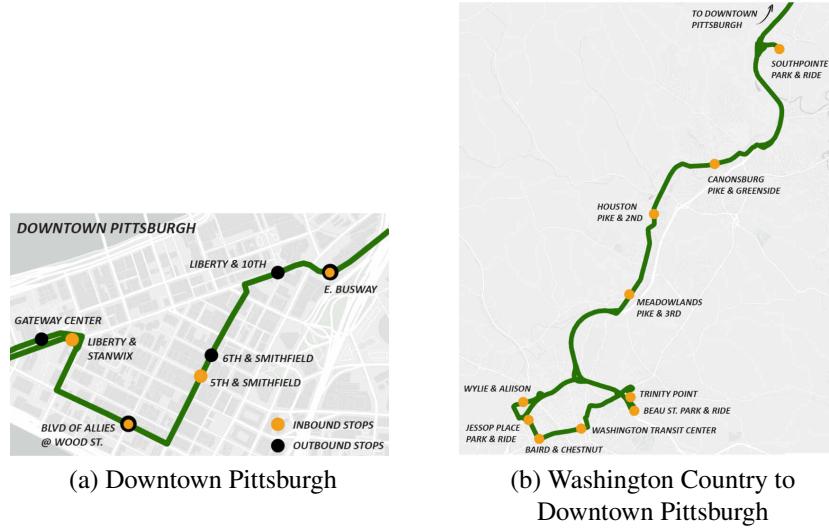


Figure 3.1: Visualization of the trajectory taken by the bus [1]



Figure 3.2: Pictures of the Transit Bus and its Exterior Cameras [29]

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

collected raw data is uploaded to a cloud storage and can be accessed from there. Data are collected continuously through the bus system.

3.2 Raw Data Definition

Our data source traverses between two endpoints: Easy busway in downtown Pittsburgh and East Chestnut Street Transit Center. We define a bus run R as a sequence of data points that

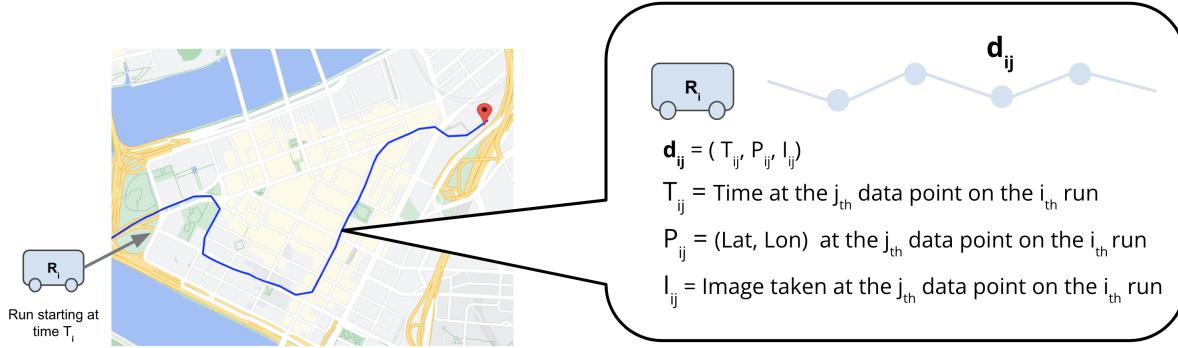


Figure 3.3: Illustration of the data points collected along the i_{th} run of the bus

traverse between the end points. Each run travels 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 $R_i = [d_{i,0}, \dots, d_{i,j}, \dots, d_{i,n}]$. Let d_{ij} be the instance of j -th raw data 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 coordinates of the data instance, $t_{i,j} \in \mathbb{N}$ represents the time in the unix timestamp, and $I_{i,j}$ being the RGB image.

Our goal is to detect changes in work zones from the collection of runs. Specifically, we are looking for changes along the spatial or temporal axis. To observe changes in these axes, we should first define an ordering for the collection of runs along the temporal and spatial axes. Based on the above definition, we see that a natural derivation of the temporal axis is the indices of the runs, since the collections 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 subsample these points.

3.3 Data Cleaning

Based on prior knowledge about the bus trajectories, we expect the collected data points to be continuous and the trajectories to be coarsely aligned. However, data collected in the wild do not always abide by this assumption. There are rare instances when there are missing images 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 pre-processing to remove low-quality images. For example, at low illumination scenarios (before sunrise), the images collected tend to be overly blurry. When the bus is stationary (waiting for the red light), the collected images tend to be highly similar to its neighbors. We remove these instances by thresholding to some blur and duplicate scores. For each data instance $d_{i,j}$ from run R_i , we filter out images whose $blur_score$ and dup_score are above some predefined threshold.

$$\begin{aligned} blur_score &= blur(I_{i,j}) \\ dup_score &= dup(I_{i,j-1}, I_{i,j}) \text{ with } j > 0 \end{aligned}$$

Here, $blur(I)$ is a measure of the variance of the Laplacian in the image I . For blurry images, the variance is expected to be low. $dup(I_{j-1}, I_j)$ measures the pixel-wise L_1 distance of the two images. Repeating images are expected to have a low L_1 distance measure. To ensure the quality of the captured images, in the following experiments, we only used images taken during the day (between 9 am and 5 pm).



Figure 3.4: (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.

3.4 Align Images using Spatial Coordinates

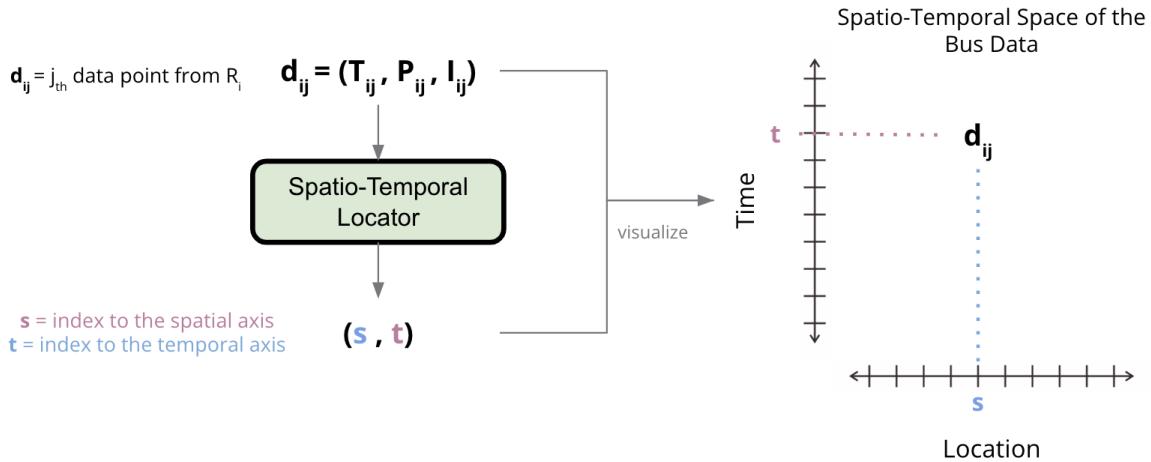


Figure 3.5: Overview of the alignment process: for each data point d_{ij} , we map it to a location on the discretized 2D space formed by a spatial and temporal axis.

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 example, the weekday and weekend routes traverse between different points. Another example

is that when part of the road is blocked, there may be a small detour in that section. Direct alignment of the bus data may not be adapted 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 $K \subseteq \mathbb{R}^2$ coordinates as a representation of the average path that the bus data traverse.

3.4.1 Generating a Baseline Trajectory

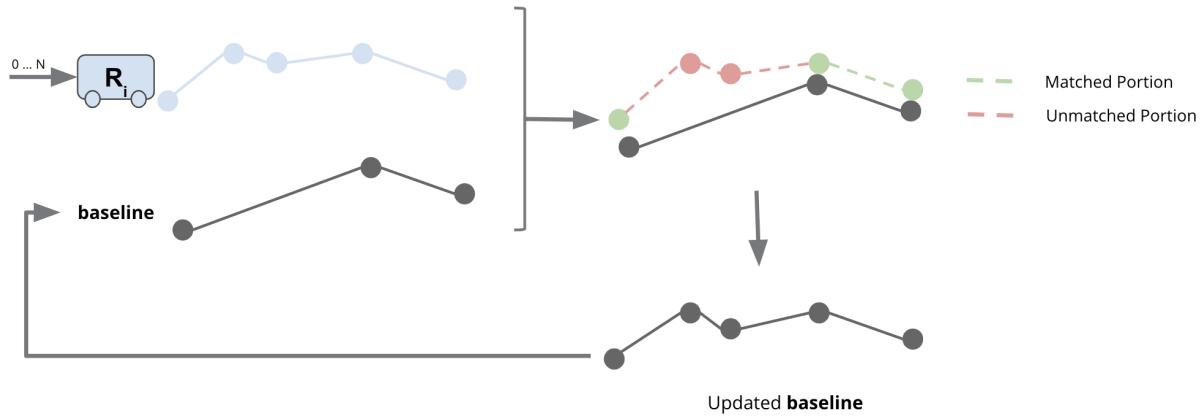


Figure 3.6: The collection of bus runs are iteratively clustered into a dense baseline trajectory.

The work [4] organized a set of algorithms for map construction. They defined map construction as a task that automatically produces or updates street map datasets using vehicle tracking data. This matches our purpose of building 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 best fits the idea of incremental track insertion. Incremental track insertion uses ideas from map matching, where the tracks are clustered and refined based on a rough baseline. This fits our use case since bus trajectories adhere to the assumption that most of the trajectories are roughly aligned on most of the segments.

We used the implementation of the Cao and Krumm's [7] incremental track insertion algorithm. This incremental track insertion approach proceeds in two stages. In the first stage, a 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, these much cleaner data are 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 contain the cluster centers of the original runs. In addition, we can impose an order on the coordinates in K in a sequence whose coordinate orders follow those of a bus trajectory ordering.

3.4.2 Downsampling the Baseline Trajectory

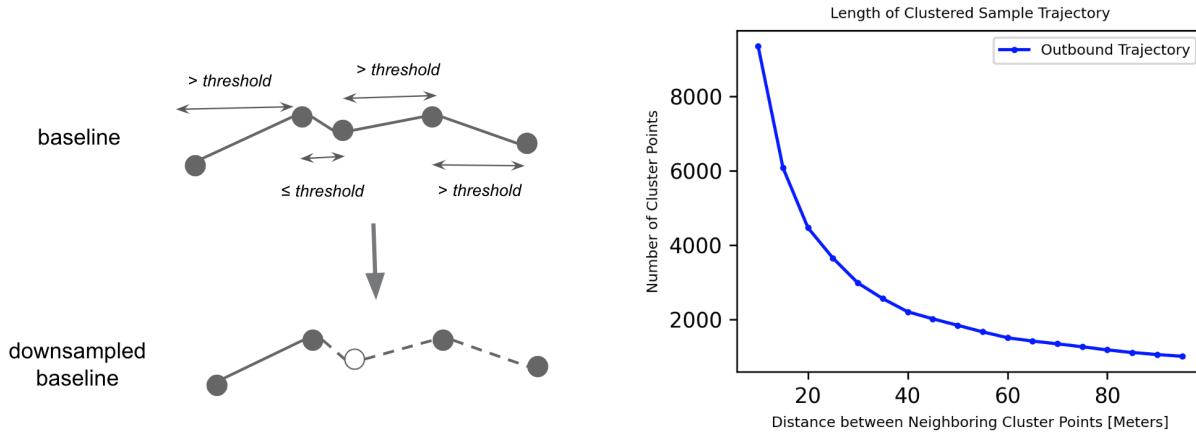


Figure 3.7: (Left) The dense baseline trajectory is downsampled based on a distance threshold to neighboring points. (Right) The number of points on the baseline trajectory decreases exponentially as we increase the distance threshold.

One challenge in working with spatial-temporal data comes from the large amount of data. Unlike other independently and identically distributed datasets. The time-series data set lies on a continuous scale. One practical approach to working with these continuous data is to downsample the data to a coarse representation. Ego vehicle data sets, such as NuScenes [6], do so by extracting keyframes among the continuous stream of data. However, in our case, we are interested in the relationship between neighboring data on both the spatial and temporal axes. Thus, it is important to find a sampling rate that is able to extract a workable set of key points that retains interesting spatial and temporal attributes.

Given the densely clustered baseline trajectory, we want to downsample the points so that the sample clusters can retain spatial information. In the context of bus data, this means that neighboring clusters should share similar visual 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 Assigning Data Points 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. Recall that a run is defined as $R_i = [d_{i,0}, \dots, d_{i,j}, \dots, d_{i,n}]$ 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 $k = \operatorname{argmin}_{k \in K} \operatorname{dist}(c_{i,j}, k)$. As noted above, we can order the coordinates in K so that they follow the order of the bus trajectory. The ordered sequence of coordinates in K is the spatial axis of the bus data.

From the above sections, we define a structure to organize the vast collection of bus data. First, we split the data into a series of runs R_i . Each run consists of an ordered sequence of

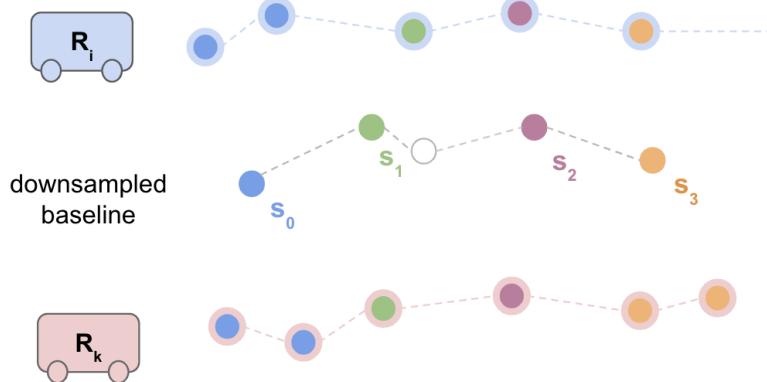


Figure 3.8: Data points on each run is assigned to its nearest point on the downsampled baseline trajectory.

data instances. Each data instance stores an image $I_{i,j}$, the time $t_{i,j}$ at which the image is taken, and the location $c_{i,j}$ at which the image is taken. Each location $c_{i,j}$ is further mapped to one of the coordinates $k \in K$, an averaged and downsampled representation of the bus trajectory. The order by start time of each run formed the temporal axis, while the ordered sequence K formed the spatial axis.

May 2, 2022
DRAFT

Chapter 4

Work Zone Detection

Work zones are a common source of traffic disruption, causing great inconvenience for commuters. Gaining a better understanding of work zones can benefit various downstream applications. Our goal is to detect work zone events among an unlabeled set of large-scale bus data. To do so, we exploit the spatial and temporal properties of the bus data to find patterns for the work zones along these axes. Anomalies can occur across the spatial axis (changes across time) or on a fixed starting time (across regions in space). As such, we are interested in answering the following questions:

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

Previously, we discussed several works that cover approaches to define a work zone. [15] uses a probabilistic model to model a binary classifier for work zones. Each factor related to the work zone, such as the speed of the vehicle and the existence of temporary traffic control objects, is treated as a weighted input to the model. Although this approach considers many related factors, it does not capture the spatial and temporal attributes of these related objects. In [26], 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. This work captures the spatial information of a work zone, but does not capture the temporal changes of a work zone. In our work, we aim to bridge this gap by detecting work zones from data points mapped onto a 2D spatial-temporal space. To do so, we use a bottom-up approach to compute an anomaly score for each data point, which measures the likelihood of a work zone at that data point. For each data point, we also consider its spatial and temporal neighbors to capture more spatial and temporal attributes. In this section, we focus on our approach of assigning each data point a corresponding anomaly score.

4.1 Work Zone Related Datasets

4.1.1 Manually Labeled Dataset

In an effort to make progress in research on work zone detection, we collected and annotated a set of images from work zones. We scrapped 4,400 work zone images from the Internet and

selected 200 images that contain work zones from the bus. In total, we currently have 4,600 work zone images. These images are manually annotated with objects listed in Table 2.1.

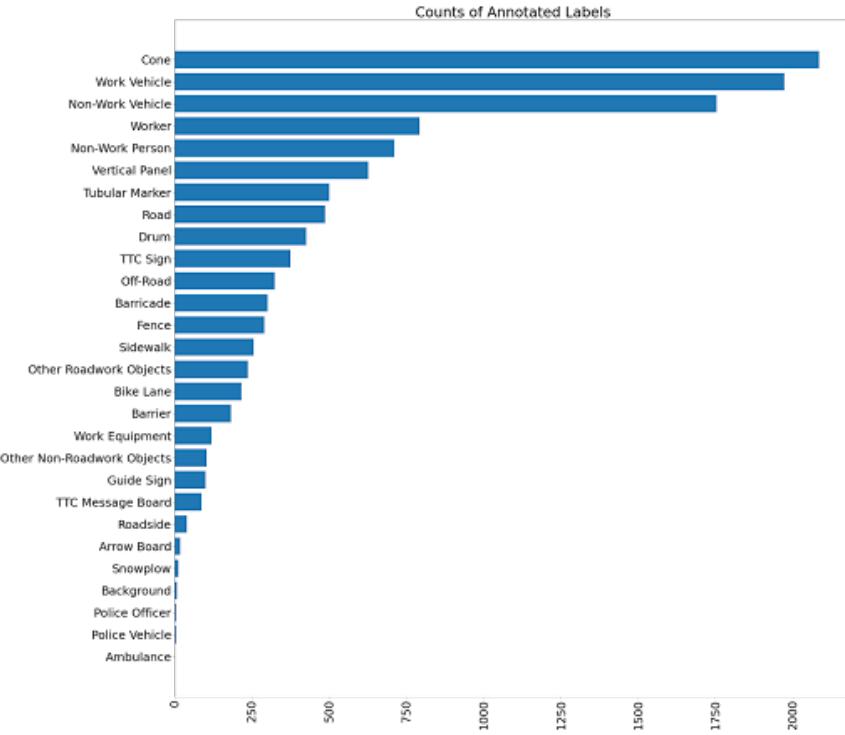


Figure 4.1: Visualization of current annotation progress of work zone related objects.

The annotation of this dataset is in ongoing progress. The instance counts in Figure 4.2 give us information about the frequency at which each object is present in the work zones. The top counts, such as cones, work vehicles, worker, and vertical panels, are observed more frequently in work zone images.

4.1.2 NuScenes

NuScenes [6] is a public large-scale dataset for autonomous driving. It enables researchers to study challenging urban driving situations through densely captured visual inputs. NuScenes contains four class labels that are related to work zones: barriers, traffic cone, construction vehicle, and construction worker.

Barrier	Traffic Cone	Construction Vehicle	Construction Worker
88,545 (12.76%)	87,603 (12.63%)	6,071 (0.88%)	13,582 (1.96%)

Table 4.1: Number of instances and ratio of all annotations in each class label the NuScenes 2D dataset.

For preliminary experiments, we trained a multiclass object detection model using the four class labels related to the work zone: barrier, traffic cone, construction vehicle, and construction

worker. In this experiment, we used a Faster R-CNN model [24] with a ResNet-50 backbone. From the training results, we noticed some challenges in the current experiment setup. First, the model performs worse on the construction vehicle and construction worker classes compared to the barrier and traffic cone classes. The imbalance in the number of instances among the four classes makes it difficult to train a multiclass object detector, especially when there are 15 times more instances of barrier than those of construction vehicles. Next, the model struggles to differentiate between normal vehicles (trucks) and construction vehicles (trucks in the construction zone). A similar problem persists between pedestrians and construction workers. This suggests that a hierarchical detection structure for these subclasses may be needed. From the current set of available annotations, the detection model performs consistently when performing single-class traffic cone detection.

For the purpose of this work, we use the number of traffic cones detected as a measure of the likelihood that the given region contains a work zone. We hope to improve this detector to a multiclass detector as we continue to annotate the work zone dataset. To train our object detection model, we use the traffic cone class [6], and tested our result on the traffic cone class of bus data.

	Number of Images	Traffic Cone
NuScenes	11,867	87,603
Bus Data	158	1,042

Table 4.2: Number of images and number of instances for NuImages and Bus Data datasets.

4.2 Model Architecture

For the purpose of this work, we are interested in detecting work zones and analyzing their spatial-temporal relationship. As such, we are more interested in obtaining accurate bottom-up detection than in efficient real-time performance. The two-stage proposal generation network generally outperforms the 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.

To evaluate the performance of the object detector, we measured the following metrics: AP^{bb} , AP_{50}^{bb} , AP_{75}^{bb} , AP_S^{bb} , AP_M^{bb} , and AP_L^{bb} . Average precision (AP) measures the mean precision values set at some recall threshold level (0 to 1 with a step size of 0.1). The numeric subscript of the AP symbols refers to the threshold of the IoU score. AP_{50}^{bb} means that a prediction with $\text{IoU} > 0.5$ is considered a true positive prediction. Usually, the mean average precision (mAP) is averaged across all class labels. In our case, we are working with single-class object detection, so we do not take the mean. The subscript S , M , L refers to the sizes of the detected objects. AP_S^{bb} measures the average precision score of objects whose area of the ground truth bounding box is less than 32^2 . AP_M^{bb} consists of boxes whose area is in the range of 32^2 and 96^2 . AP_L^{bb} contains boxes whose area $> 96^2$.

	Backbone	AP^{bb}	AP_{50}^{bb}	AP_{75}^{bb}	AP_S^{bb}	AP_M^{bb}	AP_L^{bb}
Faster R-CNN [24]	ResNet-50 FPN [12, 16]	46.8	83.8	45.5	35.6	57.1	65.2
Faster R-CNN [24]	Swin-T FPN [16, 18]	55.7	90.8	59.0	44.7	65.1	75.0

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

Based on Table 4.3, Faster R-CNN with a Swin-T outperforms the model using a ResNet-50 backbone. Next, we will test the performance of this traffic cone detector on a set of manually annotated bus data using images that contain traffic cone instances. We expect the performance of the traffic cone detector to drop on the bus data, since the training and testing distribution has changed.

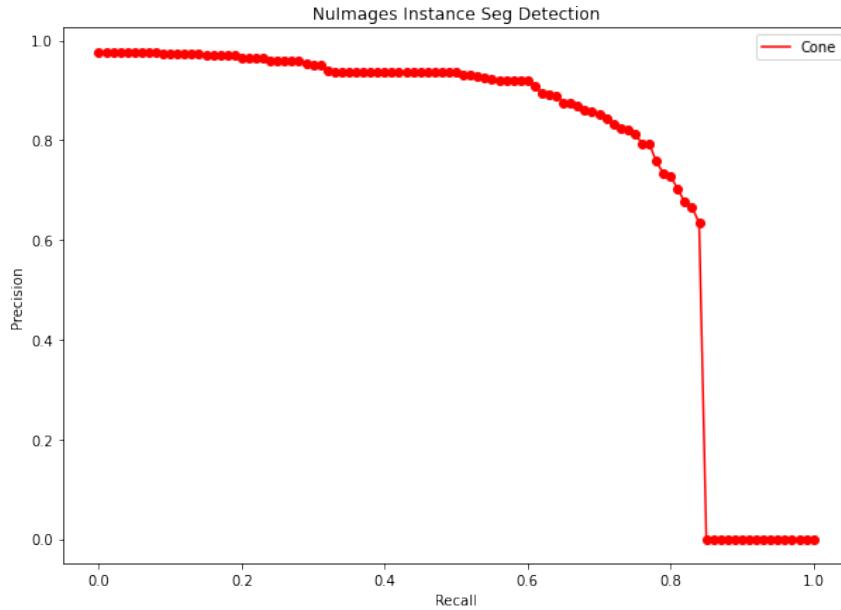


Figure 4.2: Precision-Recall curve of Faster R-CNN Swin-T FPN model on Bus Data.

Data	AP^{bb}	AP_{50}^{bb}	AP_{75}^{bb}	AP_S^{bb}	AP_M^{bb}	AP_L^{bb}
NuScenes [Train]	46.8	83.8	45.5	35.6	57.1	65.2
Bus Data [Evaluation]	35.8	81.4	25.6	32.5	64.7	-

Table 4.4: Traffic cone detection results using Faster R-CNN with Swin-T backbone.

Our goal is to detect work zone events among an unlabeled set of large-scale bus data. To do so, we design a model that assigns an anomaly score to each input data point. The anomaly score indicates the likelihood of a work zone at that data point. Currently, we use the predicted traffic cone counts as the anomaly score. Since the goal is to mine and detect patterns of work zones in bus data, we value high recall over high precision. We want to capture as many instances of the work zone as possible to find patterns in the spatial and temporal domain. From the

precision recall curve, we see that the model is capable of achieving a recall greater than 0.8 while maintaining an accuracy greater than 0.6. In the next section, we show that with this performance, we can extract patterns about work zones from the large amount of bus data.

May 2, 2022
DRAFT

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 interest 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

May 2, 2022
DRAFT

Bibliography

- [1] 3.1
- [2] 2.3, 2.3
- [3] Work/construction zones. URL <https://www.nhtsa.gov/>. 2.3
- [4] 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
- [5] 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
- [6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liang, 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, 2.4, 3.4.2, 4.1.2, 4.1.2
- [7] 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
- [8] 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
- [9] 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
- [10] 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
- [11] 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
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. ??
- [13] 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

- [14] 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
- [15] 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
- [16] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. ??, ??
- [17] 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
- [18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL <https://arxiv.org/abs/2103.14030>. ??
- [19] 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
- [20] 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
- [21] 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
- [22] 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
- [23] Amudapuram Mohan Rao and Kalaga Ramachandra Rao. Measuring urban traffic congestion-a review. *International Journal for Traffic & Transport Engineering*, 2(4), 2012. 2.2
- [24] 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, 4.1.2, ??, ??
- [25] 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

- [26] 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
- [27] 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
- [28] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemeyer 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, 2.4
- [29] Canbo Ye. Busedge: Efficient live video analytics for transit buses via edge computing. Master’s thesis, Pittsburgh, PA, July 2021. 2.4, 3, 3.2
- [30] 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
- [31] 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
- [32] 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
- [33] 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. 2.4