Supplemental Material for ROADWork: A Dataset and Benchmark for Learning to Recognize, Observe, Analyze and Drive Through Work Zones

Anurag Ghosh Shen Zheng Robert Tamburo Khiem Vuong Juan Alvarez-Padilla
Hailiang Zhu[*] Michael Cardei[*] Nicholas Dunn[*] Christoph Mertz Srinivasa G. Narasimhan
Carnegie Mellon University
https://www.cs.cmu.edu/~roadwork/

A. Related Works

Long-Tail Scenarios in Autonomous Driving. Driving datasets have evolved from KITTI [18] to more diverse collections like BDD100K [71], nuScenes [5], Mapillary [50] and Cityscapes [13], incorporating advanced sensor suites [7, 17, 59]. However, these datasets provide limited representation of long-tailed scenarios such as work zones - for instance, nuScenes contains only 19 driven sequences with work zones [58] out of 1000 scenes. Commercial self-driving vehicle deployments, while impressive in common situations, also find it difficult to navigate work zones, see Figure 1 for some failure examples collected from social media.

Prior research on long-tailed driving scenarios has largely focused on scene understanding. Datasets like CODA [38] (with 1500 scenes containing long-tailed objects), WildDash [72, 73] (with global weather and lighting variations), SegmentMeIfYouCan [6], and BDD-Anomaly [23] focus almost exclusively on recognition, rather than holistically addressing perception and navigation in scenarios like work zones. Another well-studied long-tailed scenario is driving in adverse weather. Despite data collection challenges, specialized datasets exist for fog [3, 56], night [55, 71], and snow [3, 4], although these also primarily target recognition. Figure 2 illustrates why recognition alone is insufficient for self-driving in work zones.

Work zones are complex, dynamic environments requiring multi-level understanding, yet they've received little attention due to the challenges in data mining [40] and task formulation [58]. To our best knowledge, no large-scale public dataset has specifically addressed work zones before our contribution. While the MMI Open Dataset [27] provides raw videos collected for road inspection, we develop scenario taxonomies and annotated work zones to create the ROADWork Dataset.

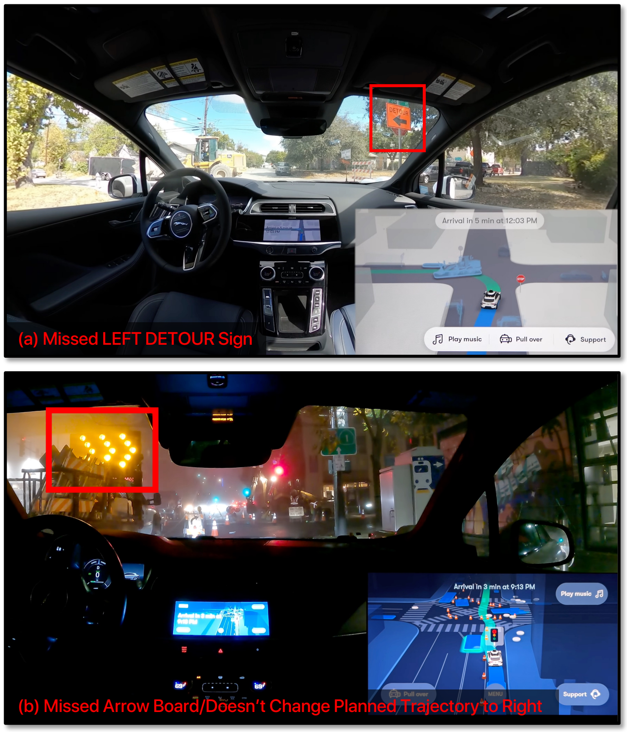

Arrival in 5 min at 12:03 PM
Play music
Pull over
Support
(a) Missed LEFT DETOUR Sign
Arrival in 3 min at 9:13 PM
Play music
Pull over
Support
(b) Missed Arrow Board/Doesn't Change Planned Trajectory to Right

Figure 1. **Examples of work zone failures in a commercial self-driving vehicle.** While obtaining detailed failure reports of self-driving cars is infeasible, customers of these companies regularly post failure cases on social media. (a) The car failed to recognize and observe a sign that mentions "DETOUR" and has a left arrow graphic (Link). (b) The car fails to recognize and observe the Arrow Board, then fails to analyze the situation and finally does not change the predicted pathway in response (Link).

Work Zones in Autonomous Driving. Prior research has addressed isolated work zone edge cases. For example, [20] recognize safety barriers using a laser scanner while [19] attempt to determine which lane lines define a valid lane

[*] Equal contribution.Work done at CMU.

in work zones. Later works [49, 58] attempted to classify and localize work zones, while others updated HD maps [48, 52] with additional work zone information. Concurrent work [31] has proposed segmenting construction areas in videos to detect continuous zones from a distance. However, no prior work systematically categorizes work zones, formulates tasks, or curates data for autonomous driving in these environments.

Language and Navigation in Work Zones. Unseen scenarios, such as newly appearing work zones along a route, pose a major challenge for autonomous driving. Work zones are a classic example of navigation in open-ended driving scenes, requiring a higher level of semantic generalization. Linguistic representations can help generalization, enabling introspective explanations [30] that improve action predictions [70].

Recently, Vision-Language Models (VLMs) [43] (and Large Language Models (LLMs)) have been increasingly applied to scene understanding, demonstrating state-of-the-art generalization and reasoning capabilities. Recent efforts [35, 47, 60, 66] have leveraged these VLMs and LLMs to redefine scene understanding and subsequently, motion planning. Navigating work zones require both visuospatial and linguistic abilities. To address this, we propose a work zone description benchmark to aid global scene understanding in workzones.

For navigation in work zones, we argue that long horizon trajectory forecasting is essential, as traditional structural cues like lanes may be unreliable. Prior works [21, 39, 46] explored a related setting: long horizon human trajectory forecasting. Inspired by this line of work, we propose a new pathway prediction problem and baselines to address tackle this challenge.

B. ROADWork Dataset Description

We describe specific information regarding annotations protocol, data cleaning and processing procedures and other relevant details.

B.1. Image Acquisition

Visual data were acquired from cameras mounted inside a vehicle while driving through 18 US cities, resulting in 9650 images from three sources: *(a)* images that we captured in Pittsburgh *(b)* images that were semi-automatically extracted from the Michelin Mobility Intelligence (MMI) Open Dataset (formerly RoadBotics) [27] *(c)* Images that were discovered in Mapillary [50], BDD100K [71] and other other data sources by our models trained on data from the first two sources.

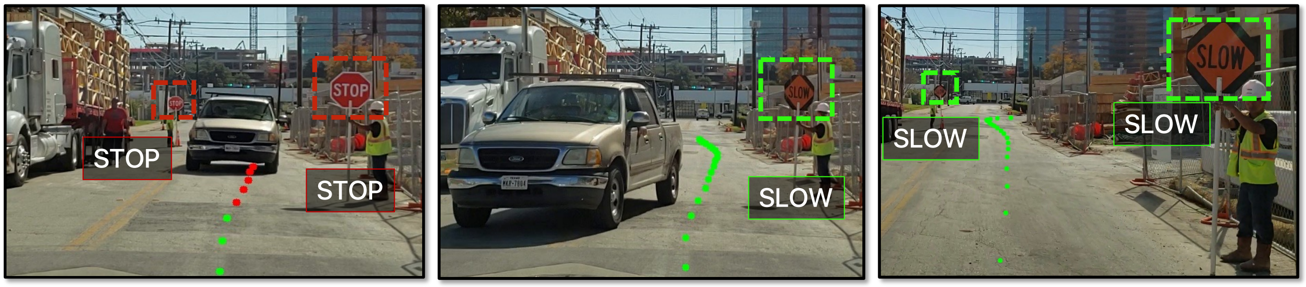
Main Data Sources. To collect the first data subset of the main dataset, we drove on urban, suburban, and rural roads in Pittsburgh and captured 2,338 (32%) images with an iPhone 14 Pro Max paired with a Bluetooth remote trigger. Next, images from other U.S. cities were sourced from videos in the MMI Open Dataset. A combination of Detic [77] and a cone detector trained on NuScenes [5] were used to mine frames presumed to contain roadwork zones with detector confidence at 25% – ensuring high recall with the expense of low precision. This process yielded approximately 100000 candidate images. We then manually selected 5078 (68%) images containing unique road objects or roadwork zones, prioritizing individual scene diversity. The distribution of images across U.S. cities is shown in Figure 3.

Discovered Data Sources. In Section 3 of the main manuscript, we described our model and the work zone classification rule that we use to discover images. We discovered work zone images from common driving datasets (a) 558 images from Mappilary [50] and (b) 411 images from BDD100K [71]. Additionally, we exploit other data sources to curate 1265 images into various subsets containing work zones (See Figure 4 for examples). These subsets were further manually filtered to remove redundant images. We describe the subsets below,

- **Vehicle-Pittsburgh Discovered Subset.** We drove a vehicle in Pittsburgh during various weather and lighting conditions, collecting approximately 157 images with work zones. This subset specifically includes examples captured in rain, fog, snow, and at night.
- **Vehicle-Rural Discovered Subset.** We collected 308 images by driving on rural roads and highways across multiple U.S. states. This subset includes work zones on two-lane roads, interstate highways, and in small towns, captured during both day and night conditions. This subset was captured using a dashcam, and shows significant radial distortion.
- **Bus-Pittsburgh Discovered Subset.** We obtained 800 work zone images from a commuter bus that followed a fixed route in Pittsburgh over the course of two years. This includes 272 images from the front-facing camera and 528 images from side-mounted cameras with unqiue viewpoints, capturing work zones in all weather conditions and times of day.

B.2. Annotations

Scene Tags. We labeled images with scene tags to capture weather, time of day, travel alterations, road environment, and whether the work zone is active (See Table 1). The presence of roadwork objects in a scene does not necessarily indicate an active work zone, e.g., *a cone in a parking lot*. Work zones are labeled as active work zone, not active work zone, or unsure. An active work zone includes roadwork as well as any activity that could potentially impact vehicles or pedestrians mobility. To qualify, objects must be located on a road or sidewalk where a vehicle or pedestrian could



Work vehicle on left side of road. Worker on left side of road. Worker on right side of road. Fence around work zone on right side of road and fully blocking right sidewalk. Work vehicle on right side of road.

Fence around work zone on right side of road. Worker and TTC sign on right sidewalk.

Worker holding TTC sign on right side of road. Work vehicle on left side of road. Barriers and fence around work zone on right side of road.

Figure 2. **Recognition is Not Enough for Navigating Work Zones.** Work zones are dynamic and rare occurrences, thus it is challenging to navigate through them. Depicted is a work zone navigation sequence with sign text detected by Glass [54], work zone descriptions generated by fine-tuned LLaVA-1.5A [44] (incorrect description indicated in red) and car trajectory estimated via COLMAP [57]. Observe that initially the worker is holding a "STOP" sign, but later switches to a "SLOW" sign as the truck passes, indicating that the road is open for traversal by the ego-vehicle. *This example shows mere object recognition is not enough for navigation; continuous fine grained scene observation and global scene analysis are both necessary.*

travel. Approximately 80% images were labeled as active work zones.

Scene Descriptions. The associated descriptions detail key work zone elements, their locations, and relationships within the scene. They specify the approximate locations of work zones and objects on the road or sidewalk while also conveying the relative positioning of objects in relation to the work zone and other scene elements. To ensure consistency, all descriptions were written by a single annotator using a standardized vocabulary.

Object Annotations. We identified 15 categories of objects commonly found in work zones. These include objects that define temporary traffic control pathways, such as cones and tubular markers, fences, barriers, and drums. Additionally, we annotated objects that help navigation, including temporary traffic control (TTC) signs, TTC message boards and arrow boards. We also annotated Workers, Work Vehicles, Police officers and Police Vehicles, since they influence and direct traffic in work zones. See Table 1 for the full list of annotated work zone objects.

The object annotation workflow combined automatic and manual labeling, followed by manual verification. To reduce annotation effort, we used Detic [77] with a custom vocabulary of "cone, drum, vehicle, traffic sign" to bootstrap annotations on our captured images. However, category predictions from Detic [77] were discarded as due to frequent classification errors. Polygons were simplified using the Vishwalingam-Wyatt algorithm [62] to facilitate editing. All object categories were manually assigned, and any additional objects in these images, as well as objects in all other images, were manually segmented and categorized. Finally, all annotations were manually verified by one person.

Object Categories		Weather	Alteration	Time	Env.
Cone	Tubular Marker	Partly Cloudy	Fully Blocked	Dark	Urban
Fence	Vertical Panel	Sunny	Lane Shift	Light	Suburban
Worker	Work Equipment	Unknown	Partially Blckd.	Twilight	Highway
Work Vehicle	Arrow Board	Wet	Other	Unknown	Rural
TTC Sign	TTC Msg. Board	Cloudy	None	Other	Unknown
Drum	Police Vehicle	Fog or Mist			Other
Barricade	Police Officer	Ice			
Barrier	Othr Rdwork Objs	Other			

Table 1. **Work Zone Object Categories and Scene Level Tags.** The left side lists manually annotated object categories, while the right side present scene-level tags that describe various work zone properties.

Fine-Grained Object Annotations. Objects that are partially blocked by other objects or truncated were labeled as "occluded". A few object categories, including arrow boards, TTC signs, and TTC message boards, have additional annotations. For example, arrow board states ("OFF", "LEFT", "RIGHT", "NONE") is annotated (See Figure 3 for the distribution of arrow board states).

TTC sign and TTC message boards generally contain both "text" and "graphics". We also annotated graphic descriptions (e.g. "LEFT ARROW") and associated text for each sign (e.g. "DETOUR AHEAD"). Additionally, text or graphics were marked as "occluded" if the object is partially occluded or truncated by the image boundary. Sign text and graphic descriptions were parsed to identify common types of TTC signs (See Figure 3). The distribution of TTC sign graphics and text follows a long-tailed pattern (See Figure 3), with 62 and 360 different types annotated, respectively.

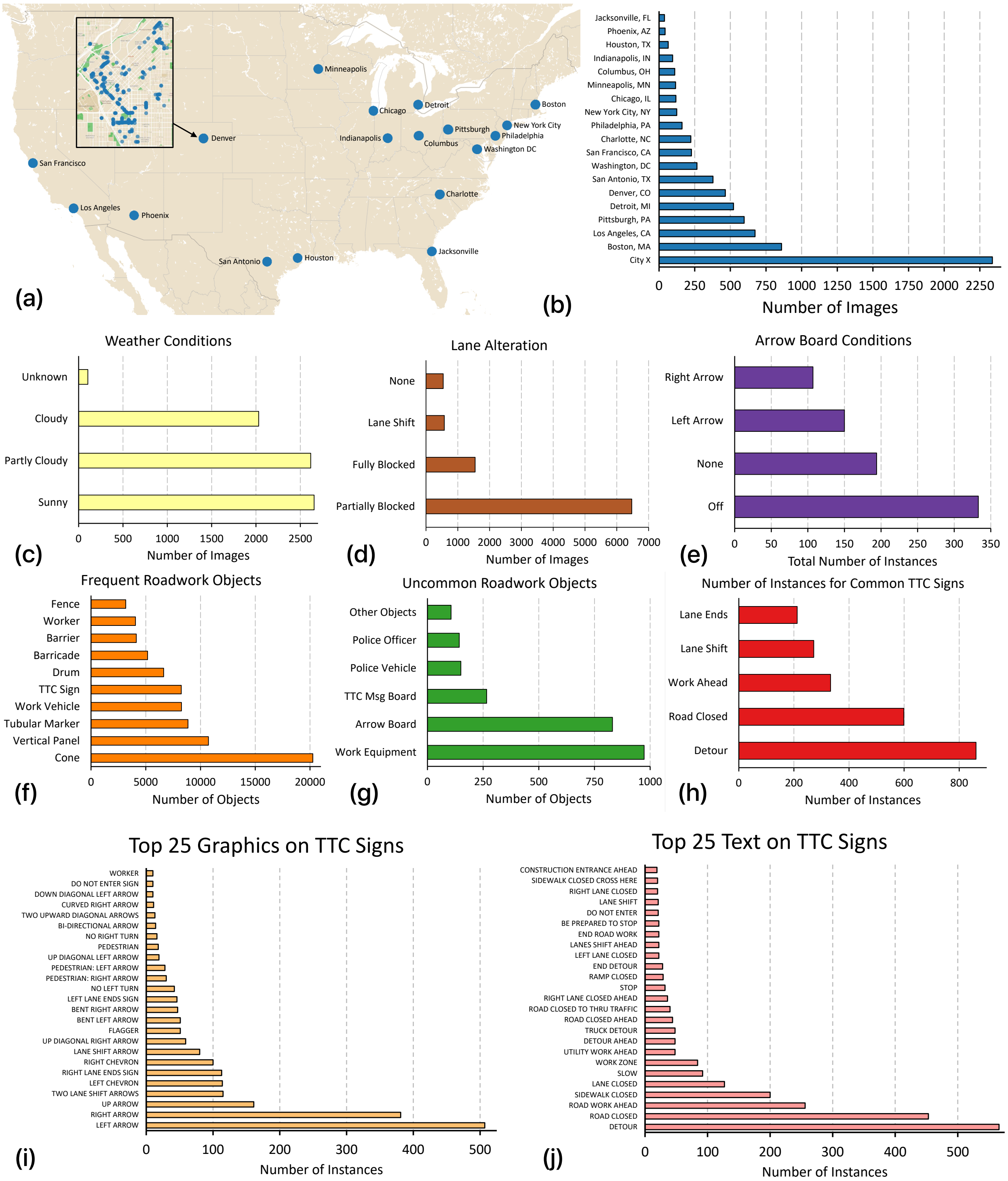


Figure 3. **ROADWork Dataset Statistics.** (a) U.S. cities represented in the dataset, with geotagged images shown for Denver, Colorado. (b) Number of dataset images for each city. (c) Distribution of weather conditions. (d) Distribution of road-network alterations for work zones. (e) Arrow board conditions, where "None" indicates that the arrow board's LEDs are not visible. (f) Distribution of frequent roadwork objects, which are of the order of thousands of total instances. (g) Distribution of uncommon roadwork objects which have a few hundred instances. (h) Distribution of the most common TTC signs (both text and graphics), which have a few hundred instances each. (i-j) Distribution of the top 25 observed TTC signs by graphics and text.

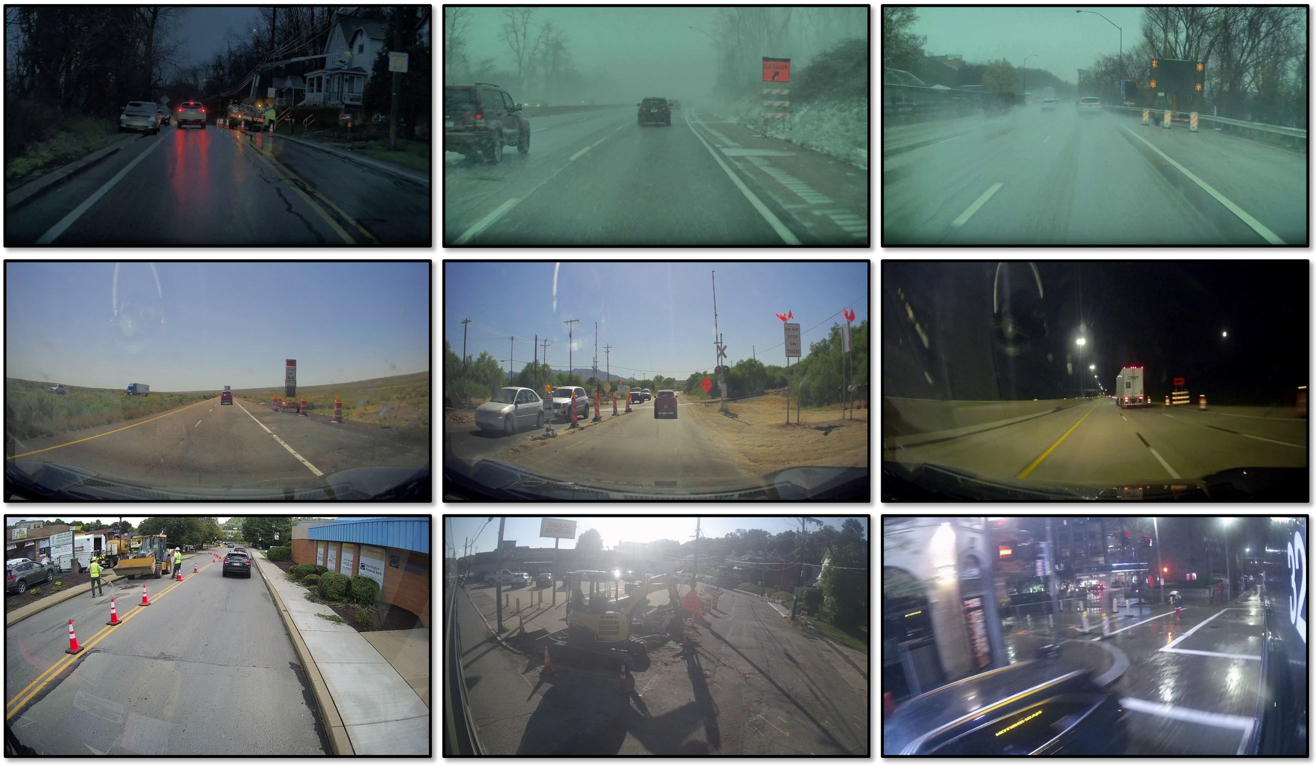


Figure 4. **ROADWork: Additional In-The-Wild Discovered and Annotated Work Zone Images.** Following the discovery process described in Section 3 of the main manuscript, we discovered and annotated an additional 1265 work zones images from a variety of sources apart from 969 images discovered in BDD100K [71] and Mapillary [50]. **(a)** The top row depicts Vehicle-Pittsburgh subset images we discovered from driving in Pittsburgh (around 157 images). The subset consists of work zone images taken in bad weather and night. **(b)** The middle row depicts Vehicle-Rural subset images we driving on rural areas and highways in the US (308 images). The subset consists of work zone images taken in both day and night. **(c)** The bottom row depicts images discovered from a Bus that was driven on a fixed route in Pittsburgh. We discovered 272 images captured from the front camera, while 528 images were captured from other cameras mounted on the bus. Images were captured in all conditions, including bad weather and night.

Semantic Segmentation. We manually segmented roads, sidewalks, and a sparse sampling of bicycle lanes to provide contextual localization for work zone objects.

B.3. Metric 3D Reconstruction and Pathway Generation from Smartphone Videos

Leveraging Smartphone-As-Dashcam Videos. Our work utilizes the MMI Open Data Set [27], which contains extensive video footage captured from a Samsung Galaxy S9 smartphone, for which camera intrinsics are known. From this dataset, we extract 30-second video snippets corresponding to our annotated work zones. These snippets are then downsampled to 5 FPS to yield the final set of smartphone images for our 3D reconstruction pipeline.

Leverging 3D Reconstruction As Anchor. Our primary goal is to produce an accurate, metric-scale, and spatially-aligned 3D reconstruction from the collection of smartphone videos, which have weakly-aligned GPS metadata. With recent advances in Visual Place Recognition [1], it's likely that the weakly-aligned GPS metadata might also be superfluous in the future.

The core of our approach is to anchor our reconstruction to a set of images with high-quality pose information [64, 65]. To achieve this, we use the initial, coarse GPS from each video to query and retrieve nearby Google Street View panoramas. We then generate multiple perspective views from each panorama, following the systematic sampling strategy described in [64]. These views, along with the panoramas' accurate GPS and pose data from large-scale SfM pipelines [33] that Google Street View is based on, serve as the high-quality georeferencing anchor for our 3D reconstructions.

Feature Matching and SfM. To reconstruct from this heterogeneous set of smartphone and Street View images, we must establish robust feature matches. As a brute-force all-pairs matching approach is computationally infeasible, we adopt a retrieval-based strategy. We first compute a global descriptor for every image using EigenPlaces [2]. We then use these features to find the top 20 nearest neighbors for each image using the Faiss library [15], efficiently identifying pairs with likely visual overlap. For these candidate pairs, we perform local feature matching by extracting keypoints and descriptors with SuperPoint [14] and matching them with LightGlue [42]. With this graph of matched images, we perform Structure-from-Motion (SfM) using the global solver GLOMAP [51] to recover camera poses and a sparse 3D point cloud. COLMAP [57] is also applicable but GLOMAP [51] is an order-of-magnitude faster.

Georeferencing and Trajectory Generation. A key step is georeferencing the resulting 3D reconstruction such that we align the reconstruction to a real-world coordinate system via a 7-DoF similarity transformation. Following the procedure described in [63, 64], this is computed by minimizing the discrepancy between the recovered poses of the Street View images and their ground-truth GPS data, which we project into an Earth-Centered, Earth-Fixed (ECEF) global coordinate frame. We explicitly discard the noisy GPS from the smartphone videos during this alignment, relying solely on the high-quality Street View data for metric accuracy [33]. The result is a single, georeferenced sparse reconstruction where the poses for all smartphone images are accurately localized, forming precise 3D trajectories. We then fit a ground plane to the reconstruction by using a Mask2Former [12] semantic segmentation model to identify 3D points corresponding to the road surface. By projecting the 3D camera poses onto this fitted plane, we define the vehicle's 3D path, which is then projected back into the source images to create 2D drivable trajectories. Visualization of our trajectories can be viewed in Figure 5.

Trajectories to Waypoints. To standardize the trajectories for our prediction task, we convert them into a fixed number of waypoints. For each sequence, we identify the longest continuous segment of the 2D trajectory that remains on the road. We then fit a spline to this segment and sample 20 equidistant waypoints. For the pathway prediction problem discussed in the main manuscript, the first five waypoints serve as the observed path (input), the final waypoint represents the goal, and the intermediate 14 points constitute the future pathway to be predicted.

B.4. Other Details

Number of Workzones. Counting work zones is challenging, as they could extend for miles. Should such long stretches be considered a single work zone? Additionally, workzones resemble the Ship of Theseus – they evolve over time while remaining at the same location for months or even years. These spatio-temporal factors make it difficult to define and count work zones accurately.

In our analysis, we counted workzones based on locations alone, clustering images within a 20m radius as a single work zone, regardless of when they were captured. As a result, ROADWork dataset contains instances of work zones at the same location observed across months or years. Besides, we used the DBSCAN algorithm to cluster workzones images based on the the noisy GPS locations. Consequently, we obtained 5024 clusters at a 20m threshold and 4759 such clusters at a 30m threshold. Based on these results, we estimate that our dataset contains approximately 5,000 work zones.

More Visualizations and Details. A full description of all annotated categories is provided in Table 12, while the distribution of each class across all cities is shown in Table 13.

C. Additional Analysis and Results

C.1. Recognizing Work Zones

Method	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l
Open Vocabulary Detectors						
Grounding DINO (O365) [45]	6.6	9.5	7.0	4.0	7.1	10.5
Supervised with ROADWork Dataset						
Faster R-CNN [53]	25.0	42.4	25.8	12.7	30.6	36.0
DiffusionDet [9]	31.1	50.1	32.2	18.3	30.8	42.0
Grounding DINO [45]	37.9	54.2	39.8	21.7	39.0	51.9
DINO [74]	**39.9**	**57.2**	**42.2**	**24.0**	**38.6**	**52.1**

Table 2. **Detecting Work Zone Objects.** We train detection models [9, 74] on the ROADWork dataset using a coarse vocabulary. The open-vocabulary detector [45] struggle to recognize work zone objects, but incorporating our data significantly improves its performance. Overall, our supervised models achieve substantially better results (**+33.3** AP).

Detecting Work Zone Objects. As mentioned in Section 3 of the main manuscript, open-vocabulary detectors such as Grounding DINO [45] follow similar trends to Detic [77] and OpenSeeD [69]. As shown in Table 2, supervised models like DiffusionDet [9] and DINO [74] significantly outperform Grounding DINO (**+33.3** AP). Fortunately, finetuning Grounding DINO [45] on our data almost matches the performance to DINO [74]. However, DINO [74] is still better than Grounding DINO [45] by **+2.0** AP, likely reflecting the trade-off between a specialized detector and an open-vocabulary detector that generalizes well across a larger number of categories.

Zero-Shot Detection On Discovered Workzones. In Section 3 of the main manuscript, we discovered 969 images in BDD100K [71] and Mapillary [50] datasets. While detectors trained on the ROADWork dataset facilitated the discovery of work zones around the world, their performance on these in-the-wild images has not been evaluated. To assess generalization, we manually annotated the 969 in-the-wild work zone images discovered in BDD and Mapillary (See Table 4 of the main manuscript for workzone discovery experiments). As shown in Table 3, the open-vocabulary detector Grounding DINO achieves significantly better zero-shot performance after being fine-tuned

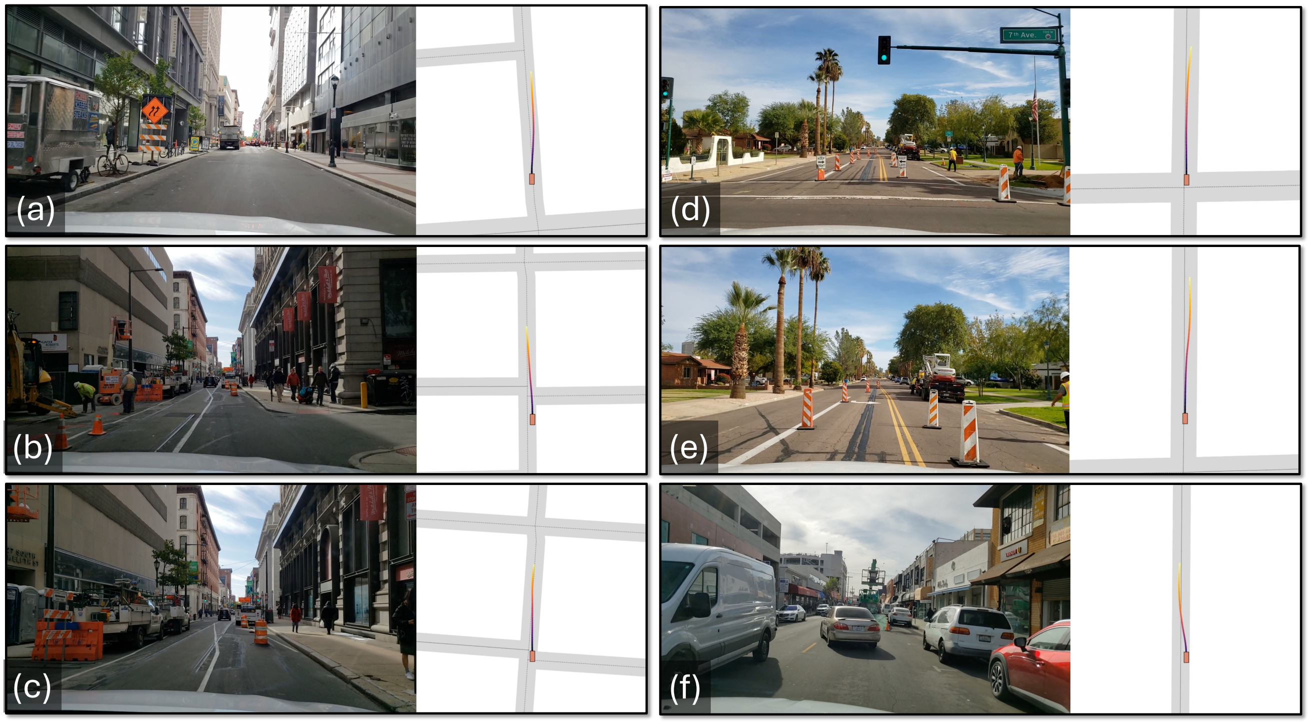


Figure 5. **Metric Geo-referenced Trajectories from our 3D Reconstruction Pipeline.** We show examples of trajectories obtained from our reconstruction pipeline overlayed on birds-eye-view maps retrieved from OpenStreetMaps. We show some interesting situations for planning which require all aspects of scene perception, the trajectories are shown for a 4 second future horizon. **(a-c)** Depict a construction zone with two lane changes, first lane change to the right is marked by a TTC sign, while the second lane change is marked with drums. **(d-e)** Depict a construction zone marked with TTC signs and Vertical Panels. While there exists "free" space to navigate to the right most lane (where the workers are), the objects helpfully mark the actual drivable regions. Identifying drivable regions is still challenging for self-driving cars [16, 34]. **(f)** Cones behind a work zone vehicle mark it as a static object blocking the lane. This cue guides the cars to change the lane towards oncoming traffic to pass this work zone vehicle.

Method	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l
	Mapillary (Discovered In-The-Wild)					
Grounding DINO [45] **(pre-trained)**	5.1	7.9	5.3	2.2	5.3	9.1
DiffusionDet [9]	13.1	24.3	12.5	5.2	12.6	23.4
DINO [74]	19.7	32.1	20.2	**10.0**	18.0	31.8
Grounding DINO [45]	**22.8**	**35.2**	**23.2**	7.9	**19.6**	**37.6**
	BDD100K (Discovered In-The-Wild)					
Grounding DINO [45] **(pre-trained)**	8.8	13.0	9.5	6.6	10.9	11.7
DiffusionDet [9]	18.5	33.2	17.9	12.1	20.7	27.9
DINO [74]	27.3	43.0	28.2	17.0	28.8	36.2
Grounding DINO [45]	**28.5**	**43.2**	**29.4**	**20.1**	**31.8**	**38.6**

Table 3. **Zero-Shot Detection On Discovered Workzones From BDD100K And Mapillary.** For discovered-in-the-wild work zone images, fine-tuning the open-vocabulary detector Grounding DINO on our ROADWork dataset improves performance by **+17.7** AP on Mapillary and **+19.7** AP on BDD100K. Additionally, the supervised detectors DiffusionDet [9] and DINO [74] achieve promising performance.

on our dataset, while supervised detectors also delivers promising zero-shot performance. Interestingly, compared to in-distribution performance (Table 2) where DINO [74] is better than Grounding DINO [45] by **+2.0** AP, in this case Grounding DINO [45] shows improved generalization (**+3.1** AP) over DINO [74].

Method	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l
	Mapillary (Discovered In-The-Wild)					
Detic [77] **(pre-trained)**	2.9	4.5	2.9	0.6	3.3	5.4
Mask R-CNN [22]	14.4	25.4	14.2	2.8	14.1	26.9
Mask DINO [37]	21.6	35.5	22.5	6.9	19	37.2
	BDD100K (Discovered In-The-Wild)					
Detic [77] **(pre-trained)**	3.7	5.8	4	3	5.1	4
Mask R-CNN [22]	19.8	33.8	21	12.8	23.3	28.1
Mask DINO [37]	29.1	46.6	31.3	18.1	31.4	45.5

Table 4. **Zero-Shot Instance Segmentation on Discovered Workzones from BDD100K and Mapillary.** As we noted in Section3, Detic [77] performed miserably for discovering work zones. We also observe that Detic's zero-shot performance on work zone images from Mapillary [50] and BDD100K [71] follows the trends from the main manuscript. Similarly, Mask DINO [37] performs significantly better on both out-of-distribution datasets.

Zero-Shot Segmentation on Discovered Workzones. We evaluate open-vocabulary detectors and ROADWork supervised models on discovered images (See Section3) – which are out-of-distribution for all the models. Pre-trained Detic performs poorly on both Mapillary (**2.9 AP**) and BDD100K (**3.7 AP**), reinforcing our observation that work zone objects are severely underrepresented in foundation model training data. In contrast, models trained on the ROADWork dataset show substantial improvements. Mask DINO [37] achieves **21.6 AP** on Mapillary and **29.1 AP** on BDD100K, representing gains of **+18.7 AP** and **+25.4 AP** respectively over pre-trained Detic. Even the simpler Mask R-CNN architecture demonstrates significant improvements when trained on our dataset.

Method	AP	$AP75$	AP	$AP75$	AP	$AP75$	AP	$AP75$
	Vehicle - Pittsburgh		**Vehicle - Rural**		**Pittsburgh Bus - Front Cam.**		**Pittsburgh Bus - Side Cam.**	
Detic (**pre-trained**)	5.1	5.9	2.4	2.5	3.6	3.4	3.7	3.8
Mask R-CNN	28.1	30.9	20.5	20.9	20	20.9	20.1	21.9
Mask DINO	**38**	**39.6**	**30.1**	**30.0**	**29.4**	**30.4**	**32.6**	**35.6**

Table 5. **Zero-Shot Instance Segmentation Results On Other Discovered Work Zone Images.** We evaluate instance segmentation models on additional discovered work zone subsets (See Figure 13) from various sources. Consistent with our prior findings, pre-trained Detic struggles on these specialized subsets, while Mask DINO trained on ROADWork significantly outperforms both pre-trained models and simpler architectures like Mask R-CNN. The performance gap is particularly noticeable in more challenging conditions like rural areas and bus-mounted camera views.

Zero-Shot Instance Segmentation Results On Other Discovered Work Zone Images. We further evaluate our models on additional discovered work zone subsets (Vehicle-Pittsburgh, Vehicle-Rural, and Bus-Pittsburgh) to assess generalization under varying conditions. As shown in Table 5 pre-trained Detic [77] performs miserably across all subsets, with AP values ranging from **2.4** to **5.1**. This performance is particularly poor in the Vehicle-Rural subset, where the AP is merely **2.4**, highlighting the difficulty of segmenting work zones in rural environments. In contrast, models trained on ROADWork show substantially better performance, with Mask DINO achieving the best results across all subsets (**+32.9** AP on Vehicle-Pittsburgh, **+27.7** AP on Vehicle-Rural, **+25.8** AP on Bus-Pittsburgh Front Camera, and **+28.9** AP on Bus-Pittsburgh Side Camera compared to pre-trained Detic). These results further validate our observations from the main manuscript, confirming that foundation models struggle with work zone recognition in diverse conditions, while our ROADWork -trained models generalize effectively across various scenarios and viewpoints.

Does fine-tuning a open-vocabulary foundation model on ROADWork cause overfitting? Large-scale foundation open-vocabulary model trained on millions of images may forget previously learned distributions when fine-tuned on additional data [26]. This effect is more pronounced when the dataset used for fine-tuning is small, leading to overfitting on the target data. To assess whether if ROADWork dataset is large enough to mitigate overfitting, we evaluate Grounding DINO [45] on common driving datasets Cityscapes [13] and BDD100K [71] with their label set as the vocabulary. We then finetune the model on ROADWork , and re-evaluate the detector on the same datasets with their vocabulary. As shown in Table 6, fine-tuning significantly improves performance on ROADWork (**+31.3** AP), while performance marginally improves on Cityscapes [13] (**+0.2** AP) or BDD100K [71] (AP). We hypothesize that this is due to the small domain gap between ROADWork dataset and common driving datasets. Hence, image features might remain consistent even when fine-tuned on our data. We leave further exploration of this phenomenon for future work.

Training Data	**Test Data**	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l
Obj365	ROADWork	6.6	9.5	7.0	4.0	7.1	10.5
Obj365 + ROADWork	ROADWork	37.9	54.2	39.8	21.7	39.0	51.9
Obj365	Cityscapes	34.2	50.2	35.9	13.6	36.0	56.2
Obj365 + ROADWork	Cityscapes	34.4	52.2	34.2	11.7	33.4	54.0
Obj365	BDD100K	23.6	40.6	23.2	9.2	27.8	49.5
Obj365 + ROADWork	BDD100K	23.7	40.9	22.8	8.4	27.2	50.0

Table 6. **Does fine-tuning a open-vocabulary foundation model on ROADWork cause overfitting?** Large foundation Models are prone to overfitting when trained on small datasets. To assess whether our dataset is large enough to mitigate overfitting, we finetune Grounding DINO [45] and evaluate it on common driving datasets including Cityscapes [13] and BDD100K [71], using their respective categories. We observe that fine-tuning significantly improves performance on ROADWork dataset (**+31.3** AP), while performance on Cityscapes (**+0.2** AP) and BDD100K (**+0.1** AP) does not degrade.

Supervision	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l
	Psuedo-Segmentations from SAM [32]					
Bbox	22.6	44.7	20.9	14.3	29.6	30.9
Bbox + 5 pts	23.3	44.9	22.1	17.7	29.6	30.4
Bbox + 10 pts	23.5	45.6	22.4	15.3	30.0	30.2
Ground Truth	**27.6**	**47.2**	**29.1**	**18.7**	**33.5**	**35.9**

Table 7. **Are Manual Segmentations Still Needed? Results with Boundary IOU.** We train instance segmentation models [22] with varying levels of supervision from the ROADWork dataset using SAM [32]. Unlike the results in the main paper, we evaluate performance using Boundary IOU [11], a metric more sensitive to boundary errors than standard IOU. We observe a larger improvement using Boundary IOU at higher thresholds (**+6.7** $AP75_{IOU(B)}$), which indicates boundary quality improvements with manual annotations compared to psuedo ground truth annotations.

Are Manual Segmentations Still Needed? We posited in Section 3 of the main manuscript that some of the object categories in ROADWork exhibit irregular shapes. We present additional results in Table 7, computing AP using the Boundary IOU [11] metric ($AP_{IOU(B)}$). This metric penalizes boundary errors more strictly, making it more suitable for evaluating segmentation quality, particularly for irregularly shaped work zone objects such as arrow boards and work vehicles (e.g., “cranes”). Compared to the results in Section 3 of the main manuscript, we find that the performance gap at tighter thresholds is even more pronounced (**+6.7** $AP75_{IOU(B)}$) compared to **+5.3** $AP75$ from Table 3 of the main manuscript. This further underscores that manual ground-truth masks yield higher quality boundaries than those predicted by SAM [32], reinforcing the need for manual segmentations of rare work zone objects.

C.1.1. Other Interesting Recognition Scenarios

ROADWork dataset enables the study of various scene understanding challenges beyond those considered of the main manuscript.

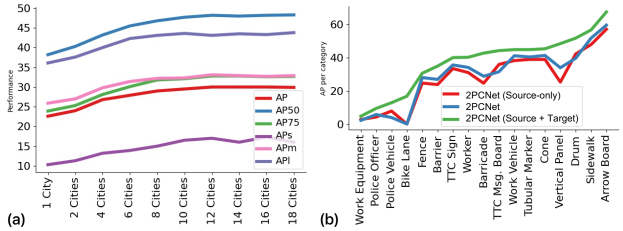


Figure 6. **Adapting to New Geographies.** (a) Starting from Pittsburgh, we progressively add data from new cities to train a detector, leading to significant accuracy gains (**+7.4** AP with all cities). (b) Geographic adaptation remains challenging. Training on source data only serves as our **baseline**, while training on both source and target data represents the **upper bound**. **Adaptation methods** such as 2PCNet [29] provide limited improvement over the baseline. For example, the upper-bound gap of adaptation method for “barricade” (**-13.8** AP_{50}) and for “vertical panel” (**-20.7** AP_{50}) is very large.

Adapting to New Geographies. While we discovered work zones in new geographies (Figure 5 of the main manuscript), does our recognition model maintain the same performance? Domain adaptation methods have explored geographic adaptation, but mainly across countries [24, 28, 61, 67, 76] and mostly for common objects like cars [67]. Obtaining supervised data for new geographies , such as new cities in our case, is expensive to scale. We make two observations: (a) A geographic domain gap exists in our data, and (b) state-of-the-art adaptation methods do not address this gap.

To demonstrate these observations, we conduct a simple experiment. We train work zone detector using data from Pittsburgh and test it on all cities. After that, we add data from the city with most samples and retrain the model. Accuracy improves by **+1.4** AP. We continue adding data from other cities, leading to a final improvement of **+7.4** AP (Figure 6 (a)).

Next, we consider the unsupervised domain adaptation problem. We treat data from Pittsburgh as source domain, where both images and labels are available during training. Data from other cities (excluding Pittsburgh) forms the target domain, where only images are available during training while performing adaptation. We evaluate the model on the target domain.

Following state-of-the-art adaptation methods [29, 76], we use a Faster R-CNN Resnet50 backbone for training, assuming different levels of available target domain data. Training on source data only is our baseline (**red** in Figure 6 (b)), whereas upper bound is trained on both labeled source and labeled target data (depicted in **green**). State-of-the-art adaptation methods [29, 76] (depicted in **blue**) do not significantly improve adaptation over baseline. Compared to the upper bound in Figure 6 (b), performance gaps remain for heavily represented objects like cones (**-5.4** AP_{50}) and drums (**-12.6** AP_{50}), and also for rare objects like barricades (**-13.8** AP_{50}) and vertical panels (**-20.7** AP_{50}). Our ROADWork dataset highlights the geographic domain gap problem, underscoring the need for new algorithms to bridge this gap.

Method	AP	$AP50$	$AP75$	AP_s	AP_m	AP_l
CS Psuedo Labels	35.3	62.2	35.4	**17.2**	36.9	50.6
UniDetector [68]	**37.3**	**63.6**	N/A	14.6	**37.9**	**56.1**
Cityscapes (Pretrained)	40.3	65.3	42.1	17.2	40.9	61.4

Table 8. **Label Unification.** We train a bounding box detector on unified Cityscapes [13] and ROADWork label space by (a) pseudo-labeling our dataset using a pretrained Cityscapes [13] (CS) model (b) via UniDetector [68]. Testing on Cityscapes [13] and compared to a pretrained model solely trained on Cityscapes [13] labels, we observe significant degradation when trained on unified label space while UniDetector [68] improves the performance over naive pseudo-labeling.

Label Unification. Suppose we aim to train a unified detector that detects both common objects like cars from a common driving dataset and rare objects from the ROADWork dataset simultaneously. This requires label unification. While practical, unifying the label space is a challenging task [68, 75] – training a model on a unified set of categories reduces performance compared to specialized models individually trained on each dataset. One reason is due to the presence of *unlabeled* instances of a particular category in the unified dataset, another could be due concept overlaps between two labels in the unified dataset. To assess this observation, we consider the Cityscapes [13] bounding box dataset in addition to our ROADWork dataset, using UniDetector [68] with a Faster R-CNN model. We train our model on the unified label space of Cityscapes (common objects) and ROADWork dataset (long-tailed objects) – (a) by employing a pre-trained detector (see Section 3) to pseudo-label the ROADWork dataset. (b) employing the method proposed by UniDetector [68]. When testing our unified models on the Cityscapes validation set, we observe a considerable drop in performance (-5.0 AP) when naive pseudo-labeling is used. However, UniDetector [68] closes the gap with the Cityscapes pre-trained model by improving the performance by **+2.0** AP over naive pseudo-labeling.

C.2. Analyzing Work Zones

Pretrained	Size	BLEU@4	METEOR	ROUGE	CIDEr
LLaVA-1.5 [44]	7B	0.4	11.0	9.4	0
LLaVA-1.5 [44]	13B	0.3	9.8	8.0	0
LLaVA-NEXT [36, 44]	13B	0.2	9.4	6.9	0
LLaVA-NEXT [36, 44]	34B	0.3	9.3	6.9	0
Fine-tuned					
LLaVA-1.5 [44]	7B	27.0	24.7	48.0	112.1
LLaVA-1.5 [44]	13B	27.7	25.1	**48.6**	113.1
LLaVA-NEXT [36]	13B	28.2	25.3	48.3	**116.4**
LLaVA-NEXT [36]	34B	**28.4**	**26.1**	47.2	113.2

Table 9. **Newer and Larger Vision-Language Models (VLMs).** Consistent with the poor performance trends in Section 5 of the main manuscript, larger pretrained VLMs (13B–34B parameters) also fail to describe work zones. Switching to a newer generation of VLMs [36] does not improve performance, reinforcing the under-representation of work zones in existing large-scale training datasets. Fine-tuning helps, but even a 34B model provides only a marginal improvement (**+1.4 METEOR**).

Newer and Larger Vision-Language Models (VLMs). We performed our experiments in Section 5 of the main manuscript using LLaVA-1.5-7B. However, two key questions arise: **(a)** Do larger VLMs also struggle with work zones descriptions? **(b)** Are newer VLMs such as LLaVA-NEXT [36] better at describing work zones than older models?

As shown in Table 9, pre-trained VLMs of all sizes perform poorly on work zones unless they are fine-tuned on ROADWork dataset. Unfortunately, even the larger LLaVA-NEXT-34B model provides only a marginally performance gain (e.g. **+1.4 METEOR**) over LLaVA-1.5-7B [44]. Moreover, newer VLMs like LLaVA-NEXT [36] do not significantly outperform the previous generation of VLMs, likely because they still lack exposure to work zone images. Our ROADWork dataset fills that gap, advancing high-level scene understanding in work zones.

Does fine-tuning a vision-language foundation model on ROADWork cause overfitting? In Section C.1 we asked if fine-tuned open vocabulary models forget previously learned distributions when trained on our ROADWork dataset. A similar question applies to vision-language foundation models, which we investigate here. We evaluate two models on COCO-Captions [10], a dataset that provides captions for many real-world images. To test our hypothesis, we evaluate two models (a) a pretrained LLaVA-1.5-7B [44] model (b) a LLaVA-1.5-7B additionally fine-tuned on the ROADWork dataset. The input prompt to both the model is *"Describe the given image in detail."* Our findings are reported in Table 10. Surprisingly, contrary to our expectations, the fine-tuned model performs better on almost all metrics. Further analysis suggests that the captioning style of COCO-Captions [10] validation set is more similar to the *terse* and *direct* scene descriptions of the ROADWork dataset, whereas the original LLaVA-1.5-7B [44] training data is more *detailed* and *flowery*. We hypothesize that fine-tuning on our ROADWork dataset improved model alignment for captioning tasks. We leave further investigation to future work.

Methods	Dataset	BLEU@4	METEOR	ROUGE	CIDER
LLaVA-1.5-7B	LLaVA	4.7	**18.1**	18.9	0
LLaVA-1.5-7B	LLaVA + ROADWork	**12.6**	16.9	**37.5**	**59.0**

Table 10. **Does fine-tuning a vision-language foundation model on ROADWork cause overfitting?** Large-scale vision-language models trained on millions of images risk overfit and catastrophic forgetting of prior learned distributions when trained on small target datasets. We test this by evaluating LLaVA-7B [44] models on COCO-Captions [10]. We evaluate using the pretrained model and model fine-tuned on ROADWork data. Surprisingly, the fine-tuned model does not degrade in performance and instead shows significant improvements (**+18.6 ROUGE**).

C.3. Driving through Work Zones

Metric $AE\% < \theta$ vs Pixel level metrics [46]. [46] presents results on pixel level metrics like Average Displacement Error (ADE) and Final Displacement Error (FDE), we also report those metrics. However, we believe $AE\% < \theta$ measures model performance more fairly in autonomous driving situations. This is because pixel level metrics like average displacement error [46] but do not account for camera's field of view in the ego-car's viewpoint. We have access to the camera instrinsics K and the angular error is computed by finding the angle between ground truth point p and predicted point $\hat{p}$ in pixel coordinates,

$$AE(p, \hat{p}) = \cos^{-1}\left(\frac{(K^{-1}p) \cdot (K^{-1}\hat{p})}{\|K^{-1}p\| \, \|K^{-1}\hat{p}\|}\right)$$

Now, we define $AE\% < \theta$ as the percentage of predictions whose angular error is within a threshold θ. Do note horizontal field of view of our images are around $50°$.

Pathway Prediction Results with Pixel Level metrics. Nevertheless, like [46], we also report pixel level displacement metrics. Table 11 shows goal and pathway results, Final Displacement Error (FDE) is the pixel error between the predicted goal and the actual goal while Average Displacement Error (ADE) is the error between predicted pathway and actual pathway. We observe that ROADWork improves both FDE (**-21.3**$\%$) and ADE (**-27.5**$\%$). We also bin the pathways in terms of curvature, and observe that paths with higher curvature are difficult to predict, however, model trained on ROADWork dataset improves FDE

Method	All Paths		Low Curvature		Medium Curvature		High Curvature	
	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE
YNet [46] w/ Pretrained Segm. [13]	31.28	102.7	28.28	95.92	29.84	102.39	40.76	113.38
YNet [46] w/ ROADWork Segm.	**22.68**	**80.78**	**22.41**	**75.33**	**21.82**	**83.28**	**30.21**	**84.58**

Table 11. **Pathway Prediction in Images.** We employ YNet [46] with a segmentation model trained on Cityscapes [13] as our baseline, and train a segmentation model with ROADWork dataset, and we observe that work zone object segmentations improve pathway and goal predictions. Displacement Error (ADE) and Final Displacement Error (FDE) captures the error of predicted pathway and goal from the ground truth pathway and goal respectively. We also report results for different thresholds of average curvatures, hypothesizing that it is more difficult to navigate workzones where pathways are more irregular. We do observe that displacement errors of both predicted pathway and goal is higher at the higher curvature threshold.

(**-25.4%**) and ADE (**-25.9%**) in those cases.

Visual Results Figure 7 shows predictions in a sequence – we observe that the trajectory heatmap is dynamic and stochastic employing different scene level cues while forecasting trajectories (such as locations of other vehicles navigating the same work zone or the available free space in the work zone). Figure 9 shows some of the failure cases. For instance, Figure 9 (c) shows predicting multiple goals is difficult and the models fails at an intersection.

D. Implementation Details

Detecting Work Zone Objects. We employ the pre-trained open vocabulary models [45, 69, 77] as is and follow their custom vocabulary protocol. For training Mask R-CNN [22], we use the mmdetection [8] library initialized with COCO [41] weights. We use the default model zoo parameters with the 1x schedule. For DINO [74], Mask DINO [37] and DiffusionDet [9], we use official codebases and weights. We employ the simple copy-paste implementation from mmdetection [8] and use the default parameters.

Adapting to New Geographies We follow the same pretrain and adaptation protocol described in 2PCNet [29].

Generating Work Zone Descriptions. To circumvent memory constraints, we train models via low rank adaptation (LORA) [25]. We also hypothesized utilzing object predictions as context would improve description quality. We compose the coarse vocabulary work zone object detector (from Section 3 in the main paper) to align our descriptions. We employ rank $R = 128$ and alpha $\alpha = 256$ while performing LORA fine-tuning on LLaVA-7B [43] for 4 epochs. We keep the rest of the parameters as is, following LLaVA's training schedule. The model prompt to generate descriptions is *"You are the planner of an autonomous vehicle, ONLY describe the workzone in the scene identifying and describing the spatial relationship of relevant objects to plan and navigate a route"*. While training with additional object context, we use ground truth to append a programmatic prompt for each object – *"(`object_category`: `confidence`) at `[(x1, y1), (x2, y2)]`"*. While testing with additional object context, we use detector predictions.

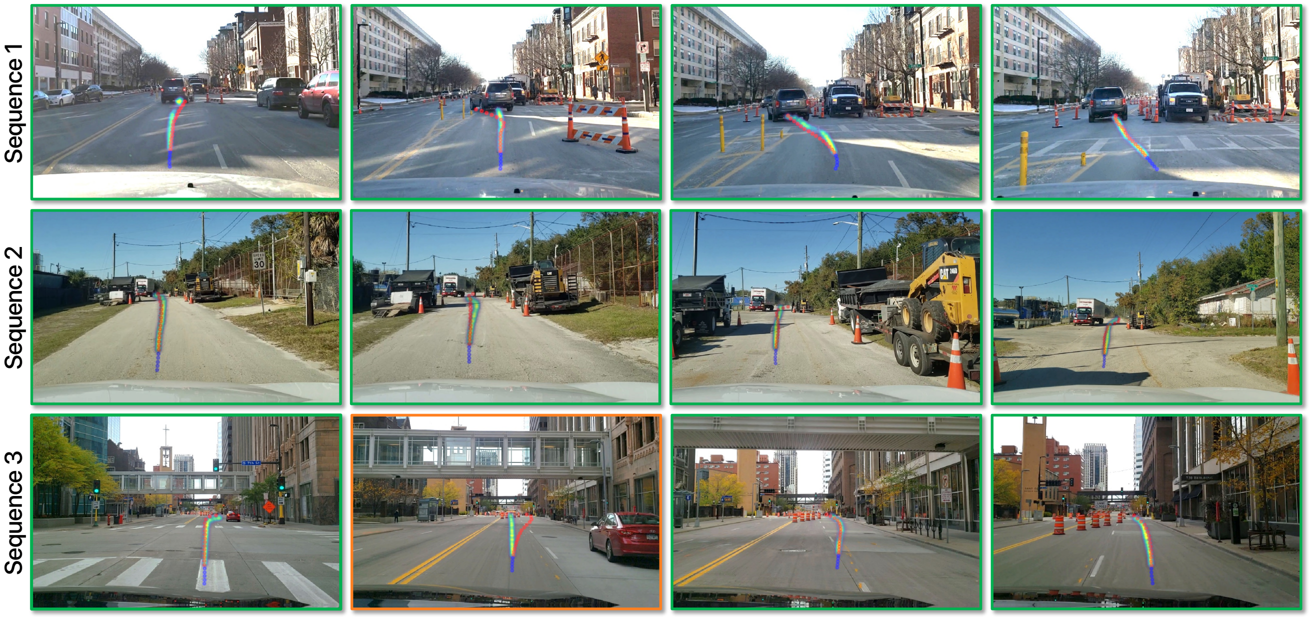


Figure 7. **Pathway Prediction for Work Zone Image Sequences.** We show examples of trajectory heatmaps predicted by YNet [46] for video sequences. Input to the model is the image and **observed pathway**, also shown is the **future pathway** (computed from actual driving). Frames are outlined indicating **plausible pathway heatmap** and **colliding pathway heatmap**. *(Sequence 1)* Following vehicles is a learned cue. *(Sequence 2)* Exploiting available free space is also learned. *(Sequence 3)* Even if the initial goal is plausible, model predicts an unsafe trajectory that would collide with work zone objects. Later, model course-corrects the trajectory when closer to work zone objects.

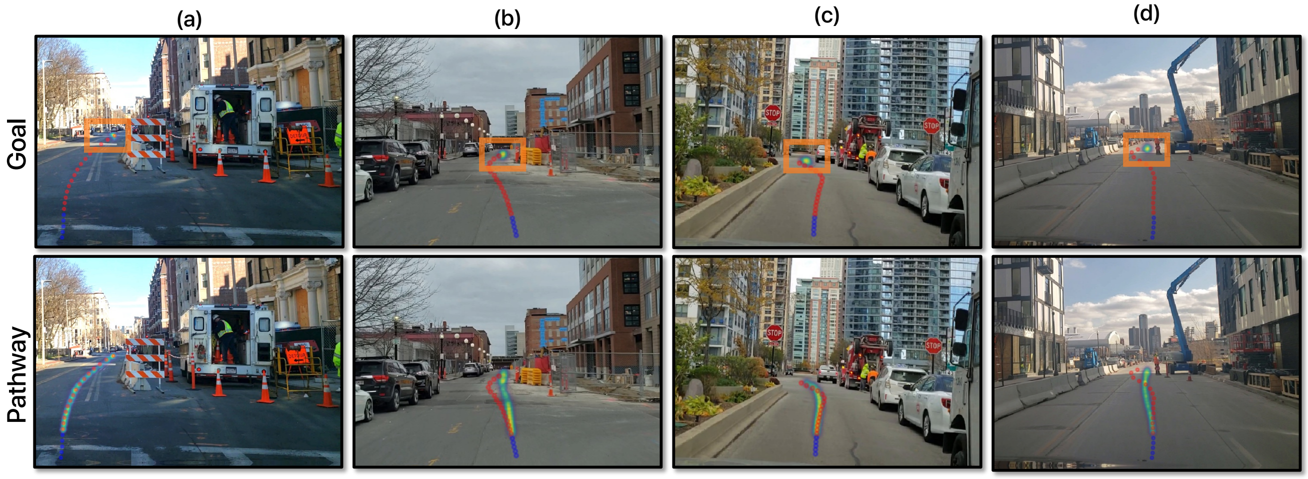


Figure 8. **Pathway Prediction in Work Zone Images.** We show examples of goal and trajectory heatmaps predicted by YNet [46]. Input to the model is the image and **observed pathway**, also shown is the **future pathway** (both computed from actual driving videos). Top row shows the predicted goal heatmaps while the bottom row shows the predicted pathway heatmaps, conditioned on a sampled goal. We observe that the predicted goal heatmap (marked with an **orange box** for clarity) is close to the ground truth goal, and the predicted pathway is plausible.

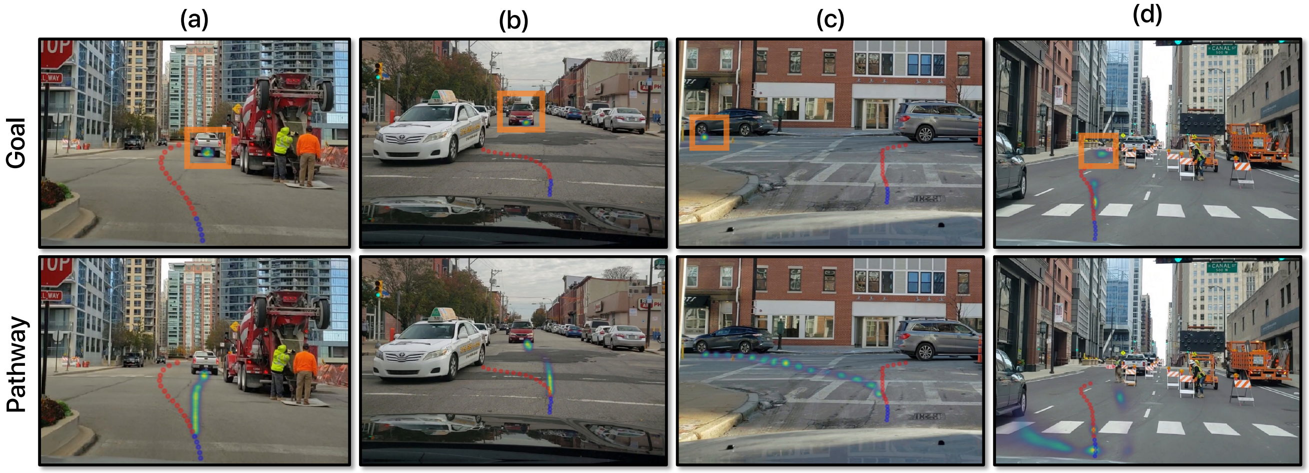


Figure 9. **Pathway Prediction in Work Zone Images: Failure cases.** We show examples of goal and trajectory heatmaps predicted by YNet [46] where the model fails. Input to the model is the image and **observed pathway**, also shown is the **future pathway** (both computed from actual driving). Top row shows the predicted goal heatmaps (marked with an **orange box** for clarity) while the bottom row shows the predicted pathway heatmaps, conditioned on a sampled goal. We observe, (a-b) the model selects vehicles in front as goal without considering global semantics. (c) Modelling multimodality of goals is a challenge, model is unable to predict all goals at an intersection. (d) Even if the goal is valid, the pathway prediction fails for heavily blocked work zones.

Table 12. Description and Examples of Roadwork Objects in ROADWork Dataset.

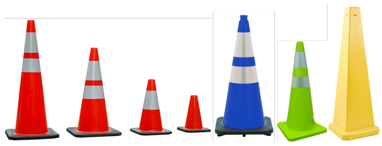
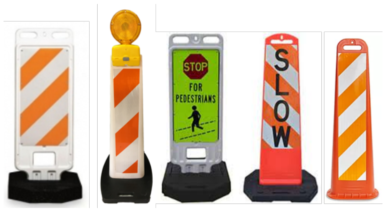
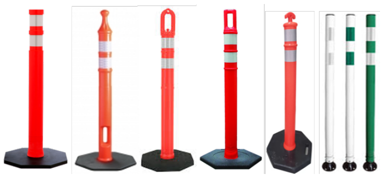
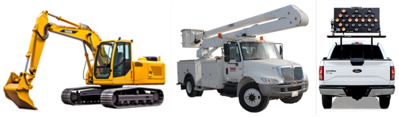
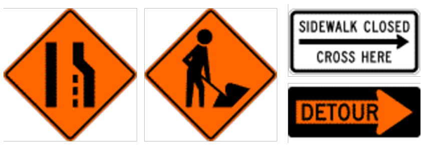
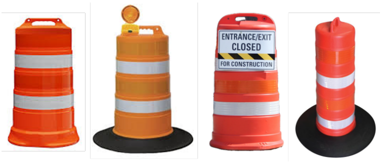
Object Name	Description	Examples
Cone	A cone shaped marker. Usually orange in color, but may be yellow, lime green, blue, red, pink or white. One or more white or retro-reflective collars around the top. May have four flat sides instead of a cone shape.	
Vertical Panel	Rectangular shaped marker. Orange or white with alternating orange and white retro-reflective stripes sloping at an angle. May have text over downward sloping stripes or text and graphics instead of downward sloping stripes. May have light on top.	
Tubular Marker	Long and round tube shaped markers. Predominately orange in color. Typically white or green when used for protected bike lanes. Top may have white or retro-reflective bands on top. Top may become flattened or a loop.	
Work Vehicle	Heavy duty and light duty vehicles that are driven and operated in order to perform roadwork related functions. Also includes traffic control vehicles and passengers vehicles that may be modified for use on the road and in work zones.	
TTC Sign	Placed temporarily in and around work zones to increase motorist and pedestrian awareness and provide information about work zones. Usually orange, but can also be white or yellow.	
Drum	Bright orange cylindrical object with horizontal retro-reflective orange and white stripes around the circumference. May have a warning light or a temporary traffic control sign mounted on top.	

Table continued on following page.

Table 12. Description and Examples of Roadwork Objects in ROADWork Dataset, Continued.

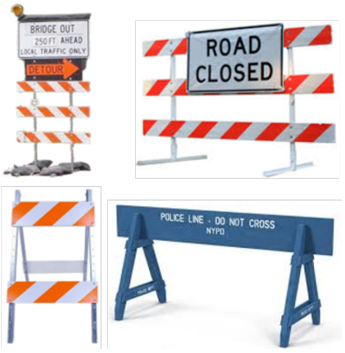
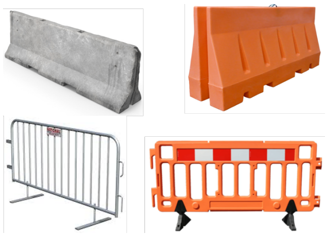
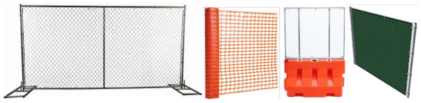
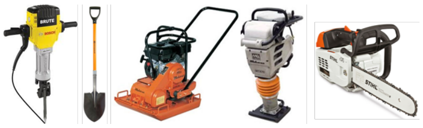
Object Name	**Description**	**Examples**
Barricade	Marker often used to indicate road or sidewalk closer or used as a channeling device. Consists of one to three horizontal boards with alternating orange and white retro-reflective stripes sloping at an angle. Single board barricades, commonly referred to as saw horse or roadblock horse, are often painted in a single color when used by local municipalities and police departments. May have a mounted warning light and/or temporary traffic control sign.	 BRIDGE OUT 250 FT AHEAD LOCAL TRAFFIC ONLY DETOUR ROAD CLOSED POLICE LINE - DO NOT CROSS NYPD
Barrier	Longitudinal channeling device used as a temporary traffic control device for merging traffic, closing roads, and to provide guidance and warning. Also used to protect workers in a work zone. Made of concrete, plastic, or metal. May be solid (e.g., concrete barriers on highway median) or have open vertical space.	
Worker	People that performing duties related to their job in the road environment. Workers may be within a confined roadwork zone or in the area outside of a work zone. Workers may be operating or inside of a vehicle. Usually identifiable by a high visibility vest and hard hat.	
Fence	Temporary structure used around a work zone. Usually a temporary chain link fence or safety fence (usually orange). Chain link fence may have privacy screen and mounted on top of a barrier.	
Work Equipment	Broadly encapsulates equipment (not including work vehicles) commonly found in roadwork zones. Includes manual and power equipment whether. May be actively in use by worker.	

Table continued on following page.

Table 12. Description and Examples of Roadwork Objects in ROADWork Dataset, Continued.

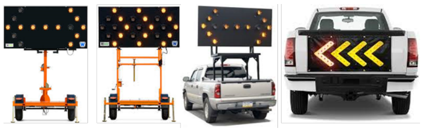
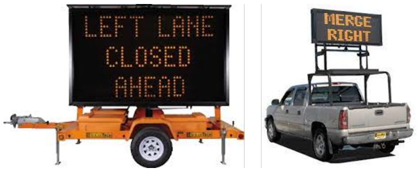
Object Name	**Description**	**Examples**
Arrow Board	Digital sign with a matrix of elements capable of displaying static, sequential, or flashing arrows used for providing warning and directional information to assist with merging and directing road users through or around roadwork zone. Usually on a dedicated trailer or may be mounted on a vehicle.	
TTC Message Board	Digital sign with the flexibility of displaying static, sequential, or flashing messages and symbols. Primarily used to advise road users of unexpected situations, displaying real-time information, and providing information to assist in decision making. Usually on a dedicated trailer or may be mounted on a vehicle.	
Police Vehicle	A vehicle used by police and law enforcement to respond to service calls. Usually a sedan, sports utility vehicle, or pick-up truck fitted with a light bar. Paint color and markings vary between states and municipalities.	
Police Officer	Uniformed officers are often in the area of work zones to help manage traffic around work sites. May be wearing high visibility vest or safety sash belt.	

Table 13. Number of annotated object instances for each city in the dataset. The categories are ordered by their totals in all the cities.

	Bike Lane	**Other Roadwork Objects**	**Police Officer**	**Police Vehicle**	**TTC Message Board**	**Arrow Board**	**Work Equipment**	**Fence**	**Worker**	**Total**
Boston, MA	16	27	27	21	18	15	108	268	263	763
Charlotte, NC	0	0	9	12	10	11	29	113	210	394
Chicago, IL	0	0	3	3	12	26	16	22	106	188
Columbus, OH	0	0	3	6	5	15	19	77	79	204
Denver, CO	40	13	5	3	9	105	29	223	157	584
Detroit, MI	0	0	1	2	9	95	38	381	113	639
Houston, TX	0	0	0	3	11	23	12	38	63	150
Indianapolis, IN	0	0	0	6	7	16	37	64	57	187
Jacksonville, FL	0	0	0	1	0	3	6	10	36	56
Los Angeles, CA	0	0	6	8	19	96	26	203	433	791
Minneapolis, MN	0	0	2	2	2	0	20	69	38	133
New York City, NY	0	0	7	3	1	6	116	59	126	318
Philadelphia, PA	0	0	3	12	4	12	38	117	155	341
Phoenix, AZ	0	0	1	0	0	6	10	39	57	113
Pittsburgh, PA	40	49	22	22	83	220	308	707	930	2381
San Antonio, TX	2	3	42	24	7	10	7	174	350	619
San Francisco, CA	0	0	3	2	6	39	23	79	251	403
Washington, DC	0	0	6	20	34	58	72	136	178	504
Total	98	106	143	150	266	831	973	3171	4060	9798

	Barrier	**Barricade**	**Drum**	**TTC Sign**	**Work Vehicle**	**Tubular Marker**	**Vertical Panel**	**Cone**	**Total**
Boston, MA	568	169	594	225	845	3591	13	1565	7570
Charlotte, NC	98	117	448	155	334	714	0	757	2623
Chicago, IL	35	178	194	44	172	40	0	386	1049
Columbus, OH	65	106	430	193	131	256	68	314	1563
Denver, CO	149	186	183	429	355	487	1004	1663	4456
Detroit, MI	317	227	1042	206	564	304	1	231	2892
Houston, TX	78	95	598	169	126	57	19	111	1253
Indianapolis, IN	58	19	194	52	120	43	1	344	831
Jacksonville, FL	10	29	15	31	57	3	0	175	320
Los Angeles, CA	237	768	23	703	864	128	11	584	3318
Minneapolis, MN	118	214	364	248	108	409	1	195	1657
New York City, NY	119	60	115	71	190	49	10	448	1062
Philadelphia, PA	184	136	396	187	299	21	0	695	1918
Phoenix, AZ	48	75	0	95	58	3	245	81	605
Pittsburgh, PA	1214	1857	471	4071	2265	1328	6906	6332	24444
San Antonio, TX	77	224	915	485	400	491	218	1202	4012
San Francisco, CA	171	133	2	124	419	211	0	1326	2386
Washington, DC	250	24	606	214	248	49	4	1432	2827
Total	4137	5167	6632	8242	8258	8859	10725	20261	72281

Acknowledgements: This work was supported by a research contract from General Motors Research-Israel, NSF Grant CNS-2038612, a US DOT grant 69A3551747111 through the Mobility21 UTC and grants 69A3552344811 and 69A3552348316 through the Safety21 UTC. We thank N. Dinesh Reddy, Shefali Srivastava, Neha Boloor, Tiffany Ma for insightful discussions.

References

[1] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *CVPR*, 2025. 5

[2] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *ICCV*, 2023. 5

[3] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *CVPR*, 2020. 1

[4] Keenan Burnett, David J Yoon, Yuchen Wu, Andrew Z Li, Haowei Zhang, Shichen Lu, Jingxing Qian, Wei-Kang Tseng, Andrew Lambert, Keith YK Leung, et al. Boreas: A multi-season autonomous driving dataset. *IJRR*, 2023. 1

[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. 1, 2

[6] Robin Kien-Wei Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. *NeurIPS Track on Datasets and Benchmarks*, 2021. 1

[7] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. 1

[8] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 11

[9] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, 2023. 6, 7, 11

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 10

[11] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. Boundary iou: Improving object-centric image segmentation evaluation. In *CVPR*, 2021. 8

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 6

[13] 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 *CVPR*, 2016. 1, 8, 9, 11

[14] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 6

[15] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024. 6

[16] Joe Eskenazi. Waymo rolls toward san francisco airport. a showdown is brewing. https://missionlocal.org/2024/12/waymo-rolls-toward-san-francisco-airport-showdown-brewing/, 2024. 7

[17] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aur'elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021. 1

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[19] Regine Graf, Andreas Wimmer, and Klaus CJ Dietmayer. Probabilistic estimation of temporary lanes at road work zones. In *ITSC*, 2012. 1

[20] Thomas Gumpp, Dennis Nienhuser, Rebecca Liebig, and J Marius Zollner. Recognition and tracking of temporary lanes in motorway construction sites. In *Intelligent Vehicles Symposium*, 2009. 1

[21] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2018. 2

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 7, 8, 11

[23] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 1

[24] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 9

[25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 11

[26] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali

Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. *NeurIPS*, 2022. 8

[27] Michelin Mobility Intelligence. Roadbotics open data set. https://www.roadbotics.com/2021/03/15/roadbotics-open-data-set/, 2021. 1, 2, 5

[28] Tarun Kalluri, Wangdong Xu, and Manmohan Chandraker. Geonet: Benchmarking unsupervised adaptation across geographies. In *CVPR*, 2023. 9

[29] Mikhail Kennerley, Jian-Gang Wang, Bharadwaj Veeravalli, and Robby T Tan. 2pcnet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. In *CVPR*, 2023. 9, 11

[30] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *ECCV*, 2018. 2

[31] Jinwoo Kim, Kyounghwan An, and Donghwan Lee. Rosa dataset: Road construct zone segmentation for autonomous driving. In *ECCV 2024 Workshop on Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving*, 2024. 2

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 8, 9

[33] Bryan Klingner, David Martin, and James Roseborough. Street view motion-from-structure-from-motion. In *ICCV*, 2013. 5, 6

[34] Michael Levenson. Driverless car gets stuck in wet concrete in san francisco. https://www.nytimes.com/2023/08/17/us/driverless-car-accident-sf.html, 2023. 7

[35] Boyi Li, Yue Wang, Jiageng Mao, Boris Ivanovic, Sushant Veer, Karen Leung, and Marco Pavone. Driving everywhere with large language model policy adaptation. In *CVPR*, 2024. 2

[36] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 10

[37] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, 2023. 7, 8, 11

[38] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *ECCV*, 2022. 1

[39] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *CVPR*, 2020. 2

[40] Mingfu Liang, Jong-Chyi Su, Samuel Schulter, Sparsh Garg, Shiyu Zhao, Ying Wu, and Manmohan Chandraker. Aide: An automatic data engine for object detection in autonomous driving. In *CVPR*, 2024. 1

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 11

[42] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, 2023. 6

[43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2, 11

[44] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 3, 10

[45] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 6, 7, 8, 11

[46] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *ICCV*, 2021. 2, 10, 11, 12, 13

[47] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 2

[48] Bonolo Mathibela, Michael A Osborne, Ingmar Posner, and Paul Newman. Can priors be trusted? learning to anticipate roadworks. In *ITSC*, 2012. 2

[49] Bonolo Mathibela, Ingmar Posner, and Paul Newman. A roadwork scene signature based on the opponent colour model. In *IROS*, 2013. 2

[50] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 1, 2, 5, 6, 7

[51] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *ECCV*, 2024. 6

[52] David Pannen, Martin Liebner, Wolfgang Hempel, and Wolfram Burgard. How to keep hd maps for automated driving up to date. In *ICRA*, 2020. 2

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 2016. 6

[54] Roi Ronen, Shahar Tsiper, Oron Anschel, Inbal Lavi, Amir Markovitz, and R Manmatha. Glass: Global to local attention for scene-text spotting. In *ECCV*, 2022. 3

[55] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019. 1

[56] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 1

[57] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3, 6

[58] Weijing Shi and Ragunathan Raj Rajkumar. Work zone detection for autonomous vehicles. In *ITSC*, 2021. 1, 2

[59] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1

[60] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2

[61] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*, 2019. 9

[62] Maheswari Visvalingam and James D Whyatt. Line generalization by repeated elimination of points. In *Landmarks in Mapping*. 2017. 3

[63] Khiem Vuong, N Dinesh Reddy, Robert Tamburo, and Srinivasa G. Narasimhan. Walt3d: Generating realistic training data from time-lapse imagery for reconstructing dynamic objects under occlusion. In *CVPR*, 2024. 6

[64] Khiem Vuong, Robert Tamburo, and Srinivasa G. Narasimhan. Toward planet-wide traffic camera calibration. In *WACV*, 2024. 5, 6

[65] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *CVPR*, 2025. 5

[66] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. *arXiv preprint arXiv:2310.17642*, 2023. 2

[67] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *CVPR*, 2020. 9

[68] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *CVPR*, 2023. 9, 10

[69] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 6, 11

[70] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *CVPR*, 2020. 2

[71] 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 *CVPR*, 2020. 1, 2, 5, 6, 7, 8

[72] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash-creating hazard-aware benchmarks. In *ECCV*, 2018. 1

[73] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying panoptic segmentation for autonomous driving. In *CVPR*, 2022. 1

[74] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023. 6, 7, 11

[75] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. Object detection with a unified label space from multiple datasets. In *ECCV*, 2020. 9

[76] Shen Zheng, Anurag Ghosh, and Srinivasa G Narasimhan. Instance-warp: Saliency guided image warping for unsupervised domain adaptation. In *WACV*, 2025. 9

[77] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 3, 6, 7, 8, 11