The Interstate-24 3D Dataset: a new benchmark for 3D multi-camera vehicle tracking

Derek Gloudemans[1,2]
derek.gloudemans@vanderbilt.edu
Gracie Gumm[1,2]
gracie.gumm@vanderbilt.edu
Yanbing Wang[1,2]
yanbing.wang@vanderbilt.edu
Will Barbour[1,2]
william.w.barbour@vanderbilt.edu
Daniel B. Work[1,2]
dan.work@vanderbilt.edu

[1] Vanderbilt University
2201 West End Ave
Nashville, TN 37235

[2] Vanderbilt University
Institute for Software Integrated Systems
1025 16th Ave S
Nashville, TN 37212

Abstract

This work presents a novel video dataset recorded from overlapping highway traffic cameras along an urban interstate, enabling multi-camera 3D object tracking in a traffic monitoring context. Data is released from 3 scenes containing video from at least 16 cameras each, totaling 57 minutes in length. 877,000 3D bounding boxes and corresponding object tracklets are fully and accurately annotated for each camera field of view and are combined into a spatially and temporally continuous set of vehicle trajectories for each scene. Lastly, existing algorithms are combined to benchmark a number of 3D multi-camera tracking pipelines on the dataset, with results indicating that the dataset is challenging due to the difficulty of matching objects travelling at high speeds across cameras and heavy object occlusion, potentially for hundreds of frames, during congested traffic. This work aims to enable the development of accurate and automatic vehicle trajectory extraction algorithms, which will play a vital role in understanding impacts of autonomous vehicle technologies on the safety and efficiency of traffic. Data is available at i24motion.org.

1 Introduction

In recent years, 3D detection and tracking datasets in the autonomous vehicle domain have led to marked advancements in perception and planning algorithms and AV technology more generally [6, 16, 45]. But designing autonomous technologies from an ego-vehicle perspective alone is not enough. Studies have shown that control algorithms designed for an individual vehicle's objectives can cause rippling instabilities in traffic [20], while controllers designed with global traffic objectives in mind can significantly reduce congestion [41, 52].

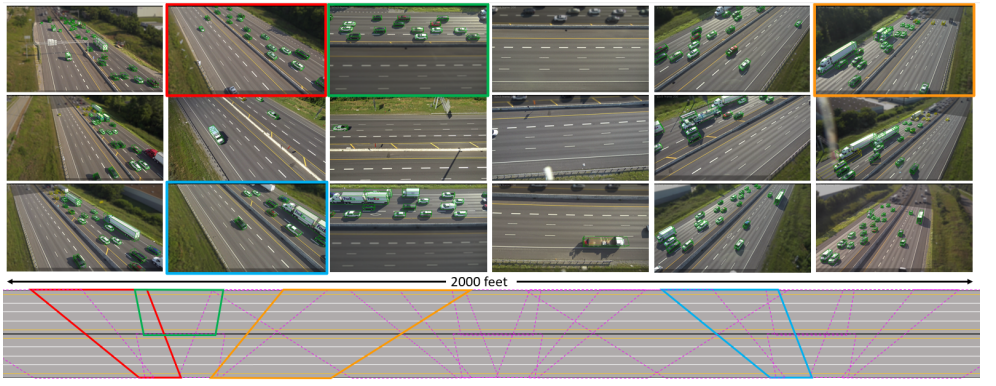

Figure 1: Example annotated (green boxes) frames from each camera field of view for one scene of the I24-3D Dataset. The approximate field of view for each camera is shown on the overhead roadway diagram below (some cameras shown in unique colors as examples). Regions outside of the considered field of view for each camera are blurred for this visualization. Cameras provide coverage of 2000 feet of Interstate-24 near Nashville, TN.

Automatic traffic monitoring offers a tremendous but under-exploited opportunity to address this issue. Computer vision research has progressed sufficiently in other fields such that efficient algorithms for traffic monitoring at scale likely exist, and state and federal transportation agencies maintain camera networks with tens of thousands of cameras nationally; increasingly ubiquitous edge sensing devices only add to the number of potentially useful traffic cameras. Moreover, in several cases, multi-camera systems have been deployed at considerable scale specifically to study the effects of *intelligent transportation systems* (ITS) and AVs on traffic [7, 18, 25, 38, 48]. Similarly, work on autonomous management of city-scale traffic will benefit immensely from the ability to track vehicle movements precisely (often requiring 3D detection) across many cameras [47]. It yet remains to be explored whether existing algorithms can achieve tracking performance suitable for fine-grained traffic analysis (i.e. HOTA above 75% and over 95% mostly tracked objects), where small localization errors or a single ID switch can be damaging in understanding a scenario [12].

We seek to enable research on precise vehicle tracking in the traffic monitoring context, with emphasis on the challenges of multi-camera tracking faced in systems such as [7, 18, 25, 38, 48]. Work in this field has been slowed by a lack of 3D multi-camera tracking data; this work addresses this shortage to enable development and evaluation of tracking methods to meet the needs of the next generation of intelligent traffic systems and AV research.

The primary contribution of this work is the introduction of a novel dataset suitable for multi-camera tracking, consisting of 877,000 3D vehicle bounding boxes annotated across 16-17 cameras with dense viewpoints covering 2000 feet of interstate roadway near Nashville, TN. The *Interstate-24-3D Dataset* (I24-3D) introduced in this work is comprised of 3 *scenes* (sets of videos recorded at the same time from different cameras), recorded at 4K resolution and 30 frames per second. Vehicle 3D bounding boxes are annotated by hand for 720 unique vehicles. I24-3D is the first 3D multiple-camera dataset in a traffic monitoring context with real videos and tracks objects across a larger set of cameras than any other multi-camera tracking dataset. **The secondary contribution of this work is the benchmarking of a number of existing algorithm combinations to assess the difficulty of 3D multi-camera tracking on this dataset,** with results (best performance of 44.8% HOTA and 62% mostly tracked objects) showing that the implemented methods achieve good per-

formance but do not produce data suitable for fine-grained traffic analysis.

The rest of this paper is organized as follows: Section 2 reviews the most analogous existing datasets. Section 3 describes the data and annotations included in I24-3D. Section 4 provides details of benchmarking experiments using the dataset, and Section 5 describes the results. Additional details on the dataset including example video links, annotation details, file format and accuracy metrics, timestamp synchronization efforts, additional experimental settings and implementation details, unabridged results, and privacy considerations are included in Appendices I through VII.

2 Related Work

Vehicle Trajectories: Traffic trajectory data consists of vehicle positional data for each vehicle within a traffic stream. Such data is required for myriad traffic analysis and modeling applications, yet sources are limited: the NGSIM dataset [1], known to contain large vehicle positional errors [12], and the HighD dataset [26] are the two main datasets and are limited in time and space. Some additional works utilize sensor-equipped vehicles [21] or GPS data [30, 39] to collect individual vehicle trajectories, but do not provide data for the majority of vehicles. Such a shortage of traffic trajectory data requires that researchers rely on models to approximate human driving behavior [27, 32, 44]. Recent efforts have sought to provide additional trajectory data using video data and computer vision [25, 38, 48].

Trajectory generation methods: To address the trajectory data shortage, several methods have been proposed to automatically extract vehicle trajectory data from existing traffic cameras. [14] proposes a method to detect vehicle 3D rectangular prism bounding boxes using background subtraction and blob segmentation, relying on automatic parameter extraction of scene homography proposed in [13]. The results are validated on derivative data products (vehicle speeds and lane positions). [37] uses 2D object detectors to roughly estimate vehicle positions on the road plane, and [43] uses ground plane projection of vehicle pixels from multiple cameras to estimate the vehicle's position, validating with turning movement counts. Other solutions use re-identification of 2D tracked objects, without addressing 2D annotation position ambiguity [11, 46]. Lastly, 3D vehicle multiple object detection and tracking methods such as [8, 28, 40, 53, 56] can also be applied to produce vehicle trajectories. A few works have addressed the multi-camera 3D tracking problem, either relying on fusing detections in a shared space [6, 34, 42] or else fusing tracklets from individual cameras after tracking [23, 49].

Multiple object tracking datasets: The task of single camera 2D *multiple object tracking* (MOT) is well-studied in varied contexts, including pedestrian and vehicle tracking from stationary and moving cameras (MOT16) [36], tracking from drone footage (VISDRONE) [58], and traffic monitoring (UA-DETRAC) [50]. 3D single camera (or stereo camera for depth) MOT is also well-addressed within the domain of *autonomous vehicle* (AV) or ego-vehicle data (KITTI, Waymo OpenDrive, and NuScenes) [6, 16, 45]. Data annotation in this context is aided by rich LIDAR data from on-vehicle sensors. Rich 3D data in the traffic monitoring (overhead traffic camera) domain is sparse, in part because LIDAR sensors are not collocated with cameras to aid in annotation. Only the BoxCars116k dataset [40] provides 3D monocular bounding boxes. Thus, research on 3D vehicle tracking from overhead cameras must use simulated or partially synthetic data [5, 55, 57].

Multi-camera tracking datasets: *Multiple camera multiple object tracking* are few in number and are mostly in the context of pedestrian tracking. The Duke-MTMC dataset asso-

ciated 2D object tracklets for pedestrians across 8 cameras [19], and the PETS dataset [15], EPFL Terrace [31], EPFL-RLC [9], and WILDTRACK [10] synchronize up to 7 cameras for pedestrian multi-camera tracking [31], These datasets provide annotations in a unified ground plane, with pedestrians represented as points [10] or grid cell occupants [9] on the ground plane. In a vehicle context, the CityFlow dataset [47] associates 2D MOT data in a traffic monitoring context across multiple cameras throughout a city, with up to 25 cameras covering a scene, but object dimensions and space are not modeled. NuScenes contains multiple frontal, side and rear-facing, frame-capture synchronized cameras enabling 3D multiple-camera tracking in an AV context. To the best of our knowledge, no multiple-camera real-world traffic monitoring dataset with 3D object tracking annotations exists, with the closest exception of Synthehicle [24] published around the submission time of this work using simulated scenes.

3 The I24-3D Dataset

This section introduces the I24-3D Dataset, detailing the location of the cameras, describing the annotations, vehicle classes, and suitable uses, and providing annotation quality metrics. Data is available at `i24motion.org`, and code demonstrating usage is available at `github.com/DerekGloudemans/I24-3D-dataset`.

3.1 Overview

The I24-3D Dataset consists of 3 *scenes*, or collections of video data recorded simultaneously from 16-17 cameras, densely covering a section of roughly 2000 feet of roadway. Each scene is 60-90 seconds long, recorded at 4K resolution and 30 frames per second, and features manually annotated 3D bounding boxes on every vehicle visible within the field of view of each camera suitable for vehicle re-identification, 3D object detection, tracking, and multi-camera tracking tasks (see Table 1.) Over 275 person hours were spent annotating the data. A full description of the dataset file structure and format is included in Appendix I, and example videos are included in supplementary material for review.

3.2 Location

I24-3D was recorded using I-24 MOTION [18], an open-road testbed along Interstate 24 near Nashville, Tennessee. The utilized portion of this testbed contains 18 cameras mounted on three 110-foot tall roadside poles, spaced at roughly 500 feet and covering an approximately

Scene	Time (s)	Cameras	Frames	Boxes	IDs	VMT	Description
1	90	17	45900	291k	324	118	Free-flow traffic
2	60	16	30600	146k	114	24.4	Slow traffic, snow conditions
3	60	16	28800	440k	282	67.0	Congested traffic
Total	210	-	105300	877k	720	209	-

Table 1: Summary of scene data for I24-3D dataset. *Time* indicates the total global duration of a scene (each video segment for the scene has that duration). *Frame* count is aggregated across all cameras in the scene, *cameras* indicates the number of active cameras for the scene. *Boxes* indicates number of 3D bounding boxes, *IDs* indicates unique vehicle trajectories.

Figure 2: Example single annotation. The annotation is stored in roadway coordinates (left) but can be projected into cameras 5 and 6 on pole 1 (p1c5 and p1c6).

2000 foot field of view on the interstate [2]. (Due to periodic camera outages, each scene contains footage from only 16-17 cameras).

3.3 Annotation Description

Annotations are provided in a roadway-aligned coordinate plane, where x-coordinate indicates distance along the roadway and y-coordinate indicates lateral (lane) position, of the bottom center rear of the vehicle. For each direction of travel in each camera field of view, a *homography* relates the roadway coordinate system to the pixel coordinates of the field of view. We rely on standard perspective transforms [22] for this conversion (see Appendix II), assuming the roadway visible in each field of view can be reasonably represented by a flat plane with a relevant *field of view* (FOV) comprising most of each image (masks are provided for regions falling outside of the FOV for each camera). All distances are given feet, as the geometry of the roadway is laid out in feet (e.g. lanes are 12 feet wide).

A single vehicle 3D bounding box annotation includes *vehicle class*, unique *vehicle ID*, bounding box *length*, *width*, and *height* (fixed for all annotations for a single vehicle), *vehicle roadway position*, *originating camera*, *timestamp*, and *frame index*. This information is sufficient to losslessly project the annotation into the originating camera, or into any other camera in which it is visible. Figure 2 shows an example. **Object localization is precise, with 1.24 ft average positional error between annotations of the same vehicle labeled in multiple cameras, and 0.5 ft average dimensional error**. (See Appendix III and IV.)

3.4 Vehicle Classes

Vehicles are classified into six classes: *sedan*, *midsize* (minivan, SUV or compact SUV), *van*, *pickup*, *semi* (tractor-trailer), or *truck*). Figure 3 depicts example annotations for each class as well as the total number of annotations for each class. We make one additional distinction: vehicles other than semis that tow trailers are classified with the towing vehicle's class, but bounding boxes are drawn to include the trailer. This choice reflects that a vehicle and trailer behave as a single semi-rigid body. Vehicle IDs with trailers include: Scene 1: [288, 133, 7, 138, 43, 270, 245, 216], Scene 3: [225, 105, 15, 148, 247, 219].

3.5 Dataset Uses and Comparison

The I24-3D dataset provides annotations of sufficient richness for a variety of canonical computer vision problems, including object reidentification, 2D and 3D detection and tracking. Most notably, multiple videos from a single scene can be used for multiple-camera tracking

Figure 3: Example vehicles and vehicle class annotation counts for the I24-3D dataset.

Dataset	Resolution	Detection		MOT		MCT		Boxes	Frames	Cameras
		2D	3D	2D	3D	2D	3D			
WILDTRACK [10]	1920×1080	✓		✓		✓	✓	38k	61k	7
KITTI [16]	1382×512	✓	✓	✓	✓			200k	15k	1
NuScenes [6]	1600×900	✓	✓	✓	✓	✓	✓	12M	40k	6
BoxCars116k [40]	*varies*	✓	✓					116k	116k	1
UA-DETRAC [50]	1920×1080	✓		✓				1.2M	140k	1
CityFlow [47]	960×540	✓		✓		✓		229k	117k	25*
Synthehicle [24]	1920×1080	✓	✓	✓	✓	✓	✓	4.62M	6.7k	7
I24-3D (Ours)	3840×2160	✓	✓	✓	✓	✓	✓	877k	105k	16-17

Table 2: Suitable uses and metrics for comparable MOT and 3D vehicle detection datasets, grouped by traffic monitoring (bottom) and other contexts (top). *MOT* indicates multiple object tracking (in 2D or 3D), and *MCT* indicates multiple camera tracking (with 3D tracking requiring a unified tracking space). *Boxes* indicates the total number of monocular bounding box view annotations, *Frames* indicates the total number of annotated frames in the dataset, *Cameras* indicates the number of camera views in a single scene. A * indicates some camera fields of view have large gaps between them (non-overlapping).

tasks, and the presence of 3D labels in this dataset enables explicit modeling of a shared 3D space for object tracking. Table 2 provides a comparison of the suitable uses of the I24-3D dataset and the most similar existing datasets. Notably, I24-3D is the only dataset in a traffic monitoring context that allows for 3D multi-camera tracking.

4 Benchmarking Experiments

To provide an initial gauge of tracking difficulty and existing algorithm performance on I24-3D, we benchmark a set of tracking methods on this dataset. Experimental protocol, metrics for evaluation, and implemented algorithms are briefly described in this section.

4.1 Experimental Protocol

Each scene is split into temporally contiguous training and validation partitions (the first 80% and the last 20% of each scene, respectively). Training is performed exclusively using the training partition. All training is performed locally on RTX6000 GPUs, and detection models are trained until convergence. During tracking we maintain tight 1/60th second synchronization between each video using corrected frame timestamps (see Appendix III), skipping frames as necessary to maintain a 15 Hz nominal frame rate.

For tracking evaluation, we find a best-fit 3rd order polynomial spline for each ground truth vehicle to obtain a continuous object representation in roadway coordinates. Predicted vehicle trajectories are compared against boxes sampled from the best-fit spline for each object. We linearly interpolate between the spline-sampled boxes and the tracker-output pre-

dictions at 30Hz to produce object sets at the same discrete times. Additional experimental details are given in Appendix V.

4.2 Metrics

We compare tracker performance using the clearMOT metrics [3], the MT/ML metrics used in [51], and HOTA [33]. We also consider the percentage of ground truth (*GT%*) and predicted objects (*Pred%*) matched to at least one predicted or ground truth object, respectively). To account for time-synchronization errors, we use a requisite 30% 2D-footprint IOU threshold between predicted and ground truth objects.

4.3 Algorithms Implemented

A variety of multi-camera 3D MOT pipelines are assembled, each requiring 3 algorithmic components: i.) a 3D object detector, ii.) an object tracker / association method, and iii.) a method for combining objects across cameras. We briefly describe algorithms implemented for each stage (implementation and parameter details can can be found in Appendix V).

3D Detectors:

- **Monocular 3D Detector (Single3D)** - a Retinanet model with Resnet34-FPN backbone [29]. The formulation is camera-agnostic (as training a separate model for each camera FOV is infeasible both from data scarcity and scalability standpoints.) *Average Precision* (AP) scores for this detector: $AP_{30} = 0.718$, $AP_{50} = 0.598$, $AP_{70} = 0.254$. (See Appendix V for experimental details.)
- **Monocular 3D Multi-frame Detector (Dual3D)** - Inspired by recent works utilizing multiple frames for detection and tracking [55], we add the previous frame as detection input. AP scores for this detector: $AP_{30} = 0.810$, $AP_{50} = 0.714$, $AP_{70} = 0.572$.
- **Monocular 3D Crop Detector (CBT)** - as described in [17], we train a Retinanet Model with Resnet34-FPN backbone for detecting objects in cropped portions of full frames. AP scores for this detector: $AP_{30} = 0.767$, $AP_{50} = 0.700$, $AP_{70} = 0.464$.
- **Ground Truth Detections (GT)** - perfect ground-truth detections.

Object Trackers:

- **Kalman-Filter IOU Tracker (KIOU)** - as described in [4]. We utilize a contant velocity roadway-coordinate Kalman filter for object position prediction.
- **ByteTracker (Byte)** - noting this tracker's state of the art performance on the MOTChallenge benchmarks [36], we utilize the two-stage association method described in [54], using IOU as both primary and secondary matching criterion and utilizing a Kalman filter as suggested by authors.
- **Crop-based Tracking (CBT)** - as proposed in [17], detection on some frames is performed by re-detecting priors in cropped subsets of the overall frame, and object associations are implicit for these frames.
- **Ground Truth Single Camera Tracklets** - perfect single-camera tracklets.

Cross-Camera Rectification Methods:

- **Detection Fusion (DF)** - as preferred in the AV context [6], detections from all cameras are combined online in roadway coordinates and non-maximal supression with a stringent 1% IOU threshold utilized to eliminate overlapping detections.

Detector	Tracker	DF	TF	HOTA	MOTA	Rec	Prec	GT%	Pred%	MT	ML	Sw/GT
Crop	Byte	✓	✓	23.6	21.3	53.4	64.0	90.5	72.9	25.6	25.0	1.1
Crop	KIOU	✓	✓	24.6	21.4	54.4	64.2	90.5	71.2	27.6	22.3	1.1
Dual3D	Byte	✓	✓	30.9	50.0	65.6	81.9	90.6	93.4	35.9	15.0	0.9
Dual3D	KIOU	✓	✓	39.7	71.6	76.5	93.7	91.5	95.3	52.5	10.4	0.7
Single3D	Byte	✓	✓	27.5	49.3	62.8	83.9	92.1	91.8	29.7	15.4	0.9
Single3D	KIOU	✓	✓	39.9	71.6	76.3	94.1	93.4	**95.6**	51.6	8.8	0.7
Crop	Byte		✓	15.2	-16.5	43.5	47.0	90.4	59.8	14.5	32.2	1.5
Crop	KIOU		✓	20.7	-2.1	51.6	52.9	90.2	53.7	25.5	26.4	1.5
Dual3D	Byte		✓	38.7	75.0	80.2	93.8	93.0	93.6	59.0	8.0	0.8
Dual3D	KIOU		✓	**44.8**	77.0	**83.0**	93.2	91.7	92.3	**63.8**	8.8	**0.5**
Single3D	Byte		✓	38.2	72.6	80.6	90.9	**94.9**	92.5	58.7	**4.7**	1.1
Single3D	KIOU		✓	**44.8**	**77.1**	**83.0**	93.4	93.3	91.3	62.2	7.8	**0.5**
Crop	Byte	✓		19.2	34.7	58.3	72.8	91.9	81.4	27.1	16.6	2.4
Crop	KIOU	✓		19.3	32.1	57.9	71.4	91.9	79.7	25.9	17.3	2.4
Dual3D	Byte	✓		20.9	60.2	64.2	94.8	92.7	94.7	29.5	8.7	2.8
Dual3D	KIOU	✓		21.1	60.4	64.0	95.2	92.6	94.7	29.7	8.9	2.7
Single3D	Byte	✓		21.3	60.3	63.9	95.3	93.8	94.2	27.1	7.5	2.6
Single3D	KIOU	✓		21.4	60.3	63.7	**95.5**	94.2	93.9	26.5	7.5	2.6
Crop	Byte			17.6	18.7	59.8	64.0	91.9	73.0	30.4	15.1	3.0
Crop	KIOU			16.9	10.8	57.5	60.0	91.9	66.0	28.2	18.5	3.2
Dual3D	Byte			15.0	55.1	72.8	81.7	93.2	87.5	42.5	6.8	7.3
Dual3D	KIOU			15.1	55.6	72.7	82.2	93.1	87.8	42.3	7.0	7.3
Single3D	Byte			15.1	54.0	72.3	80.8	94.3	85.9	40.5	5.8	7.2
Single3D	KIOU			15.2	54.4	72.2	81.3	94.5	86.1	39.4	5.6	7.1

Table 3: Tracking results for each multi-camera tracking pipeline. **Sw/GT** indicates object ID switches per ground truth object. Best result for each metric shown in bold.

- **Trajectory Fusion (TF)**- as proposed in [49], single camera tracklets are compared for spatio-temporal overlap offline, stitched together when a matching criteria is met, and refined to optimally describe the observed set of tracked object positions.
- **None** - as a baseline, object tracklets from each camera are output with no fusion.
- **Both (DF+TF)** - Tracking uses detection fusion, and a subsequent trajectory stitching step is performed to deal with remaining object fragmentations.

5 Results

Table 3 reports results for each of the above implemented pipelines. The best performing pipeline combines Dual3D detection with KIOU tracking and trajectory fusion (HOTA 44.8%). In general, trajectory fusion alone performs best (across otherwise equal run settings) and no cross-camera rectification strategy (baseline) performs worst. While relatively high MOTA scores are achievable at a low 0.3 IOU threshold (77.1% maximum), HOTA scores are still relatively low when compared to top performing algorithms on MOTchallenge and KITTI [16, 36]. This is primarily driven by relatively low localization accuracy, especially for fast moving vehicles (where a 1-frame timing error results in dropping below a 70% threshold for localization accuracy for an otherwise perfect detection.) See Appendix VI for an example HOTA plot at varying localization thresholds.

Even the best pipelines miss 5% of ground truth objects entirely (GT%), and track only 64% of objects for 80% of overall duration (MT). This result demonstrates the difficulty of tracking most or all of the vehicles in a traffic scene at the level of granularity and

completeness necessary for in-depth traffic analysis. Even utilizing ground truth detections or single camera tracklets cannot fully mitigate these failures. For brevity, pipelines utilizing ground truth inputs are included in Appendix VI; the best-performing pipeline utilizing ground truth detections achieves HOTA 59.6%, and the best-performing pipeline utilizing ground-truth single-camera tracklets achieves HOTA 61.6%. This indicates that the cross-camera tracklet rectification problem is difficult even with great single-camera tracklets.

Table 4 reports results for the best pipeline per scene. Scene 1 is easiest across a variety of metrics, with Scene 2 being easier on MOTP (slow-moving objects due to snowy conditions minimizes localization inaccuracies). Per-scene results for all methods are included in Appendix VI. Figure 4 shows the best performing pipeline's outputs evaluated against ground truth object annotations for Scene 3. Lanes farther from cameras and with high object densities have a much higher rate of false negatives (e.g. westbound (WB) lane 4). Slow-moving, un-occluded objects (e.g. WB Lane 1) are tracked relatively accurately. Faster moving objects (e.g. EB Lane 2) are often tracked, but not accurately enough to surpass the IOU threshold requirement. Results on Scene 3 demonstrate the difficulty of tracking all objects in dense stop-and-go traffic, when many objects are occluded for long periods of time.

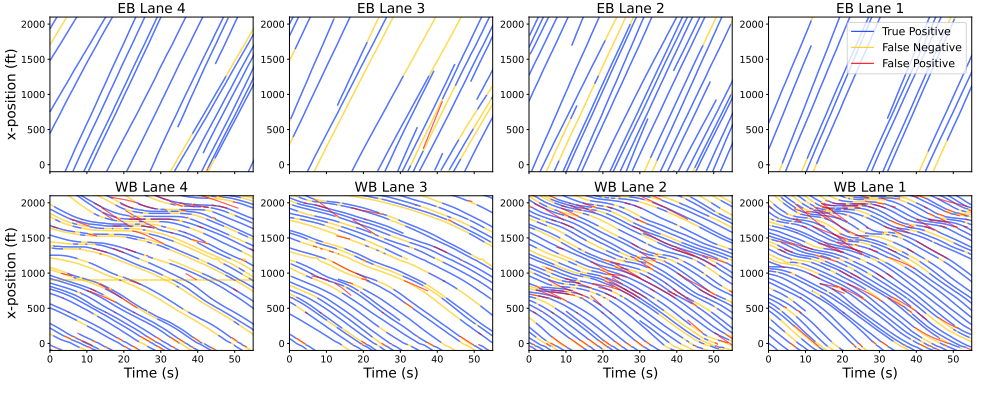

Figure 4: Time-space diagrams (object x-position vs time) for each lane for Dual3D +KIOU+TF pipeline on Scene 3 (Lane 4 is rightmost lane in direction of travel). False negatives (yellow), false positives (red) and true positives (blue) shown. In this case, most false positives are closely paired with a false negative, indicating that an object was tracked below the IOU threshold. Lanes farthest from cameras (EB lane 1 and WB lane 4) have the more false negatives in general, likely due to smaller object size and greater object occlusion. In some cases, a predicted object that falls below the IOU threshold with a ground truth object results in a parallel false positive and false negative track.

Scene	HOTA	MOTA	MOTP	Rec	Prec	GT%	Pred%	MT	ML	Sw/GT
1	**58.5**	**89.7**	69.2	**92.9**	**96.7**	**95.3**	**98.4**	**86.3**	**2.2**	**0.02**
2	46.9	77.7	**74.5**	86.2	91.1	90.4	82.4	64.0	9.6	0.49
3	29.1	63.5	64.8	69.9	91.7	89.3	96.1	40.9	14.6	1.05
avg	44.8	77.0	69.5	83.0	93.2	91.7	92.3	63.8	8.8	0.52

Table 4: Tracking results for Dual3D + KIOU + TF for each scene. Best score for each metric shown in bold, generally suggesting an easier scene.

6 Conclusion

This work introduced the I24-3D dataset, a multi-camera 3D vehicle tracking dataset with a total of 57 minutes of video and 877,000 vehicle annotations across 16-17 cameras. It also provided an initial benchmarking of some multi-camera 3D tracking pipelines from existing algorithms, demonstrating the difficulty of tracking on this dataset.

The benchmarking performed in this work represents a first step towards developing and evaluating efficient and accurate 3D multi-camera tracking pipelines. Moreover, though none of the benchmarked pipelines achieved performance suitable for fine-grained traffic analysis (i.e. HOTA > 0.75, mostly tracked objects > 95%), we suspect that there do exist methods or combinations of methods that will perform better than the implemented methods from this work, especially those that better utilize the 3D scene information stemming from multiple cameras in an intensely occlusion-aware manner. We encourage interested researchers to report their results on this benchmark utilizing the protocol described in Section 4 and Appendix V. In the future, we look forward to developing such scene-aware MOT methods, armed with a new enabling dataset. We also intend to release a 3D multi-camera tracking challenge with new scenes and cameras from the I-24 MOTION system [18].

Acknowledgements

This work is supported by the National Science Foundation (NSF) under Grant No. 2135579, the NSF Graduate Research Fellowship Grant No. DGE-1937963 and the USDOT Dwight D. Eisenhower Fellowship program under Grant No. 693JJ32245006 (Gloudemans) and No. 693JJ322NF5201 (Wang). This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) award number CID DE-EE0008872. This material is based upon work supported by the CMAQ award number TN20210003. The views expressed herein do not necessarily represent the views of the Tennessee Department of Transportation, U.S. Department of Energy, or the United States Government.

References

[1] Vassili Alexiadis, James Colyar, John Halkias, Rob Hranac, and Gene McHale. The next generation simulation program. *Institute of Transportation Engineers. ITE Journal*, 74(8):22, 2004.

[2] W. Barbour, D. Gloudemans, M. Cebelak, P. Freeze, and D. Work. Interstate 24 motion open road testbed. In *Proceedings of the ITS America Annual Meeting, to appear*, location, 12 2021.

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008: 1–10, 2008.

[4] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[5] Thibault Buhet, Emilie Wirbel, and Xavier Perrotton. Conditional vehicle trajectories prediction in carla urban environment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[6] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[7] Collin Castle. Michigan department of transportation cav corridor, 2023. URL https://www.michigan.gov/mdot/travel/mobility/initiatives/cav-corridor.

[8] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teuliere, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2040–2049, 2017.

[9] Tatjana Chavdarova and François Fleuret. Deep multi-camera people detection. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 848–853. IEEE, 2017.

[10] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5030–5039, 2018.

[11] Yucheng Chen, Longlong Jing, Elahe Vahdani, Ling Zhang, Mingyi He, and Yingli Tian. Multi-camera vehicle tracking and re-identification on ai city challenge 2019. In *CVPR Workshops*, volume 2, pages 324–332, 2019.

[12] Benjamin Coifman and Lizhe Li. A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset. *Transportation Research Part B: Methodological*, 105:362–377, 2017.

[13] Markéta Dubská, Adam Herout, Roman Juránek, and Jakub Sochor. Fully automatic roadside camera calibration for traffic surveillance. *IEEE Transactions on Intelligent Transportation Systems*, 16:1162–1171, 2014.

[14] Markéta Dubská, Adam Herout, and Jakub Sochor. Automatic camera calibration for traffic understanding. In *BMVC*, volume 4, page 8, 2014.

[15] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–6. IEEE, 2009.

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237, 2013.

[17] Derek Gloudemans and Daniel B Work. Vehicle tracking with crop-based detection. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 312–319. IEEE, 2021.

[18] Derek Gloudemans, Yanbing Wang, Junyi Ji, Gergely Zachar, Will Barbour, and Daniel B Work. I-24 motion: An instrument for freeway traffic science. *arXiv preprint arXiv:2301.11198*, 2023.

[19] Mengran Gou, Srikrishna Karanam, Wenqian Liu, Octavia Camps, and Richard J Radke. Dukemtmc4reid: A large-scale multi-camera person re-identification dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–19, 2017.

[20] George Gunter, Derek Gloudemans, Raphael E Stern, Sean McQuade, Rahul Bhadani, Matt Bunting, Maria Laura Delle Monache, Roman Lysecky, Benjamin Seibold, Jonathan Sprinkle, et al. Are commercially implemented adaptive cruise control systems string stable? *IEEE Transactions on Intelligent Transportation Systems*, 22(11): 6992–7003, 2020.

[21] Jonathan M Hankey, Miguel A Perez, and Julie A McClafferty. Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets. Technical report, Virginia Tech Transportation Institute, 2016.

[22] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[23] Yuhang He, Xing Wei, Xiaopeng Hong, Weiwei Shi, and Yihong Gong. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE Transactions on Image Processing*, 29:5191–5205, 2020.

[24] Fabian Herzog, Junpeng Chen, Torben Teepe, Johannes Gilg, Stefan Hörmann, and Gerhard Rigoll. Synthehicle: Multi-vehicle multi-camera tracking in virtual cities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1–11, 2023.

[25] Rachel James. Third generation simulation: A closer look at the impact of automated driving systems on traffic, 2023. URL https://highways.dot.gov/research/projects/third-generation-simulation-closer-look-impact-automated-driving-systems-traffic.

[26] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125. IEEE, 2018.

[27] Gunwoo Lee, Soyoung You, Stephen G Ritchie, Jean-Daniel Saphores, Mana Sangkapichai, and R Jayakrishnan. Environmental impacts of a major freight corridor: a study of i-710 in california. *Transportation Research Record*, 2123(1):119–128, 2009.

[28] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In *European Conference on Computer Vision*, pages 644–660. Springer, 2020.

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[30] Siyuan Liu, Ce Liu, Qiong Luo, Lionel M Ni, and Ramayya Krishnan. Calibrating large scale vehicle trajectory data. In *2012 IEEE 13th International Conference on Mobile Data Management*, pages 222–231. IEEE, 2012.

[31] Wenqian Liu, Octavia Camps, and Mario Sznaier. Multi-camera multi-object tracking. *arXiv preprint arXiv:1709.07065*, 2017.

[32] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2575–2582. IEEE, 2018.

[33] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578, 2021.

[34] Elena Luna, Juan C SanMiguel, José M Martínez, and Marcos Escudero-Viñolo. Online clustering-based multi-camera vehicle tracking in scenarios with overlapping fovs. *Multimedia Tools and Applications*, pages 1–21, 2022.

[35] Hui Miao, Feixiang Lu, Zongdai Liu, Liangjun Zhang, Dinesh Manocha, and Bin Zhou. Robust 2d/3d vehicle parsing in arbitrary camera views for cvis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15631–15640, 2021.

[36] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[37] Xinhe Ren, David Wang, Michael Laskey, and Ken Goldberg. Learning traffic behaviors by extracting vehicle trajectories from online video streams. In *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, pages 1276–1283. IEEE, 2018.

[38] Toru Seo, Yusuke Tago, Norihito Shinkai, Masakazu Nakanishi, Jun Tanabe, Daisuke Ushirogochi, Shota Kanamori, Atsushi Abe, Takashi Kodama, Satoshi Yoshimura, et al. Evaluation of large-scale complete vehicle trajectories dataset on two kilometers highway segment for one hour duration: Zen traffic data. In *2020 International Symposium on Transportation Data and Modelling*, 2020.

[39] Wenhuan Shi, Shuhan Shen, and Yuncai Liu. Automatic generation of road network map from massive gps, vehicle trajectories. In *2009 12th international IEEE conference on intelligent transportation systems*, pages 1–6. IEEE, 2009.

[40] Jakub Sochor, Jakub Špaňhel, and Adam Herout. Boxcars: Improving fine-grained recognition of vehicles using 3-d bounding boxes in traffic surveillance. *IEEE transactions on intelligent transportation systems*, 20(1):97–108, 2018.

[41] Raphael E Stern, Shumo Cui, Maria Laura Delle Monache, Rahul Bhadani, Matt Bunting, Miles Churchill, Nathaniel Hamilton, Hannah Pohlmann, Fangyu Wu, Benedetto Piccoli, et al. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments. *Transportation Research Part C: Emerging Technologies*, 89:205–221, 2018.

[42] Elias Strigel, Daniel Meissner, and Klaus Dietmayer. Vehicle detection and tracking at intersections by fusing multiple camera views. In *2013 IEEE Intelligent Vehicles Symposium (IV)*, pages 882–887. IEEE, 2013.

[43] Sukriti Subedi and Hua Tang. Development of a multiple-camera 3d vehicle tracking system for traffic data collection at intersections. *IET Intelligent Transport Systems*, 13 (4):614–621, 2019.

[44] Jian Sun, Han Liu, and Zian Ma. Modelling and simulation of highly mixed traffic flow on two-lane two-way urban streets. *Simulation Modelling Practice and Theory*, 95:16–35, 2019.

[45] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[46] Zheng Tang, Gaoang Wang, Hao Xiao, Aotian Zheng, and Jenq-Neng Hwang. Single-camera and inter-camera vehicle tracking and 3d speed estimation based on fusion of visual and semantic features. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 108–115, 2018.

[47] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8797–8806, 2019.

[48] Antje von Schmidt, María López Díaz, and Alain Schengen. Creating a baseline scenario for simulating travel demand: A case study for preparing the region test bed lower saxony, germany. In *International Conference on Advances in System Simulation (SIMUL)*, pages 51–57. ThinkMind, 2021.

[49] Yanbing Wang, Derek Gloudemans, Zi Nean Teoh, Lisa Liu, Gergely Zachár, William Barbour, and Daniel Work. Automatic vehicle trajectory data reconstruction at scale. *arXiv preprint arXiv:2212.07907*, 2022.

[50] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193:102907, 2020.

[51] Bo Wu and Ram Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 951–958. IEEE, 2006.

[52] Cathy Wu, Aboudy Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M Bayen. Flow: Architecture and benchmarking for reinforcement learning in traffic control. *arXiv preprint arXiv:1710.05465*, 10, 2017.

[53] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Data-driven 3d voxel patterns for object category recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1903–1911, 2015.

[54] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 1–21. Springer, 2022.

[55] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pages 474–490. Springer, 2020.

[56] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7556–7566, 2021.

[57] Minghan Zhu, Songan Zhang, Yuanxin Zhong, Pingping Lu, Huei Peng, and John Lenneman. Monocular 3d vehicle detection using uncalibrated traffic cameras through homography. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3814–3821. IEEE, 2021.

[58] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Haibin Ling, Qinghua Hu, Haotian Wu, Qinqin Nie, Hao Cheng, Chenfeng Liu, et al. Visdrone-vdt2018: The vision meets drone video detection and tracking challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.