So you think you can track?

Derek Gloudemans† Gergely Zachár† Yanbing Wang† Junyi Ji† Matt Nice†
Matt Bunting† William W. Barbour† Jonathan Sprinkle† Benedetto Piccoli‡
Maria Laura Delle Monache§ Alexandre Bayen§ Benjamin Seibold*
Daniel B. Work†

Vanderbilt University †
1625 16th Ave S, Nashville, TN 37212
derek.gloudemans@vanderbilt.edu

Rutgers University-Camden ‡
303 Cooper St, Camden, NJ 08102

UC Berkeley §
University Ave and Oxford St, Berkeley, CA 94720

Temple University *
1801 N Broad St, Philadelphia, PA 19122

Abstract

This work introduces a multi-camera tracking dataset consisting of 234 hours of video data recorded concurrently from 234 overlapping HD cameras covering a 4.2 mile stretch of 8-10 lane interstate highway near Nashville, TN. Video is recorded in cooperation with Tennessee State Department of Transportation and its policies. The video is recorded during a period of high traffic density with 500+ objects typically visible within the scene and typical object longevities of 3-15 minutes. GPS trajectories from 270 vehicle passes through the scene are manually corrected in the video data to provide a set of ground-truth trajectories for recall-oriented tracking metrics, and object detections are provided for each camera in the scene (159 million total before cross-camera fusion). Initial benchmarking of tracking-by-detection algorithms is performed against the GPS trajectories, and a best HOTA of only 9.5% is obtained (best recall 75.9% at IOU 0.1, 47.9 average IDs per ground truth object), indicating the benchmarked trackers do not perform sufficiently well at the long temporal and spatial durations required for traffic scene understanding. Video data, scene information, and vehicle trajectories are made publicly available at i24motion.org.

1. Introduction

Much concerted work has been spent on multiple object tracking benchmarks in recent years, primarily from the perspective of pedestrian tracking in crowds [11, 35] or vehicle tracking from an AV perspective [8, 16]. These datasets generally have high object density, short scenes (1-2 minutes), and short object longevity (∼10 seconds), focusing on high localization accuracy, precision and recall. As a result they do not emphasize challenging aspects of long-term tracking: appearance changes, long-term occlusions, and increasing chance of fragmentation or ID swaps with increasing track length.

Crucially, there is no existing multiple object tracking dataset with a high object density (over 250), long moving object durations (over 5 minutes), and more than 25 overlapping cameras covering a single scene or scenario at the same time. As a result, researchers cannot answer whether existing tracking algorithms are suitable for tracking objects through dense scenes over tens of thousands of frames, because there is no dataset to perform this evaluation on. Such tracking is crucial in the context of traffic science, where origin-destination information for individual vehicles and long-term vehicle behavior are paramount for designing well-fitting models of human driver behavior [24, 27]. It is our goal to provide a video dataset of a different spatial and temporal scale than previous works to enable object tracking research in this vein.

To this end, we present the *Interstate 24 Video* (I24V) dataset. The dataset consists of a single scene, 1 hour in duration, of 4.2 miles of interstate roadway, covered by 234 cameras with overlapping fields of view. Given the scale of this dataset (over 2000 times the video duration of MOTChallenge [11], 500x the duration of KITTI [16] and 80x the scale of CityFlow [44]) traditional manual annotation of objects is infeasible. To combat this difficulty, we provide a set of 270 manually-corrected GPS trajectories from over 100 instrumented vehicles on the roadway during the recording duration. Objects persist for an average of 6.6 minutes (11880 frames average at 30 *frames per second* (FPS)) and a high object density ($>$ 500 across the scene) is typically observable. This annotation set is suit-

Figure 1. Example fields of view from each of the 234 cameras included in the I24V dataset. Each camera is recorded in 1920 $\times$ 1080 resolution and at 30 frames per second. Scene information is provided for each roadway direction of travel in each camera.

able for assessing object tracking algorithms along recall-oriented metrics. Initial experiments show that existing high-performing trackers fall well short of acceptable tracking performance on data of this scale, and further work is needed to develop suitable algorithms for long-term tracking tasks. We take considerable care to make the data useful for computer vision applications, developing new techniques for keeping camera homographies more accurately aligned than existing stabilization methods allow. Succinctly, the contributions of this work are:

1. The largest multi-camera video dataset (234 cameras and 234 hours of video covering a scene with high object density and long object durations).
2. A sparse set of 270 GPS-produced annotations corresponding to 1782 minutes of labeled vehicle trajectory.
3. Preliminary benchmarking of existing object tracking algorithms on this dataset.
4. Precise scene information and a unified curvilinear coordinate system for the entire scene, useful for filter-based tracking and downstream traffic science.
5. Methods for precisely re-aligning camera homographies to account for drift outperforming existing image stabilization techniques in over 99% of cases.

The rest of this paper is organized as follows: Section 2 situates this work within existing literature. Section 3 introduces the dataset, its attributes, and methods used to ensure its fidelity. Section 4 describes the numerical experiments and Section 5 the results for homography re-estimation methods and for object tracking algorithms benchmarked on the dataset. Much additional explanation and analysis omitted for brevity is available in the supplement.

2. Related Work

This section provides a brief overview of existing multiple object tracking and multiple camera datasets, and briefly explores existing multi-camera tracking approaches.

Multiple Object Tracking Datasets: The task of *multiple object tracking* (MOT) has been well studied, thanks to a number of MOT datasets in different contexts including traffic monitoring [48], drone footage [57], crowded pedestrian scenes [11, 35], and autonomous vehicle scenarios [8, 16, 42]. Objects are annotated either as 2D rectangular bounding boxes [11, 35, 48, 57] or 3D rectangular prisms projected from a from a ground-plane into the image [8, 16, 42]. 3D annotations are primarily seen in the AV context as the collocation of cameras with rich sensor suites such as LIDAR and depth sensors makes the semi-automated production of annotations possible [8]. As a result of this abundance of datasets, a huge variety of accurate MOT methods have been developed in these contexts [33]. A variety of simple yet successful approaches to this task work in an online *tracking-by-detection* paradigm [4,5,10,49]. Many modern approaches to the MOT problem solve tracking and object detection *jointly*, sharing information such as object priors, multiple frames, or scene hidden states to aid in detection [2, 34, 37, 56].

Multi-camera Multiple Object Tracking Datasets: Relatively less work has been devoted to the task of *multiple-camera, multiple object/target tracking* (MCMT), perhaps due to the fact that such datasets are harder to produce and until recently relatively few have been available. The PETS dataset [14], CamNeT [53] and EFPL [15] datasets each track a few pedestrians across up to 8 cameras for relatively short durations in each of a few scenes [9, 15]. The Wildtrack Dataset [9] and DukeMCMT dataset [19]

Dataset	Cameras	Video (hr)	Scene (min)
DukeMTMC [19]	8	11.3	85
Wildtrack [9]	7	1.0	8.6
CityFlow [44]	25	3.3	6.5
Synthehicle [22]	7	17	3
EPFL-Terrace [15]	4	14	3.5
PETS [14]	8	0.2	0.3
pNEUMA Vision [26]	10	3.9	13
I24-3D [17]	17	1.0	1.5
I24-Video (proposed)	234	234	60

Table 1. This table summarizes the most comparable existing multi-camera datasets according to *Cameras*, the total number of camera fields of view covering a single scene, *Video*, the total length of all included video, and typical *Scene* duration as estimated from available information for each work.

track pedestrians across 7 and 8 cameras for much longer scene durations (up to 85 minutes). Modern AV datasets include a variety of cameras onboard so can be utilized as multi-camera datasets [8], but have short scene durations and object longevity. More recently, the CityFlow dataset [44] contains 40 total cameras (up to 25 for a single scene, with some non-overlapping) totaling over 4 hours of video data in a traffic monitoring context. The pNEUMA Vision dataset [26] provides up to 10 drone-mounted camera views and scenes of up to 13 minutes in duration, though has known annotation shortcomings. Synthehicle [22] contains synthetic 3-minute scenes with up to 7 cameras in a traffic monitoring context, totalling over 17 hours of video footage, and the I24-MC3D dataset has scenes with up to 17 cameras and 1.5 minutes in length [17] Table 1 summarizes existing works. Crucially, there is no multi-camera dataset with a high object density (over 100), long object durations (5+ minutes), and more than 25 overlapping cameras.

Multi-camera Tracking Approaches generally handle multiple inputs by one of 3 methods: i.) *Detector input fusion* performs object detection utilizing frames from all cameras simultaneously [25, 39, 50, 54]. ii.) *Detection fusion* combines all object detections in a shared space via non-maximal suppression [8], hierarchical clustering of detections [32,43], Gaussian mixture models [28,41], or other methods [12] before performing object tracking via traditional approaches. iii.) *Tracklet fusion* methods combine single-camera MOT tracklets with graph-based formulations [40, 51, 53], trajectory to tracklet matching [21], or greedy methods based on trajectory similarity measures [23, 45], often also incorporating camera-link based models [21,23,52] or performing smoothing to output more feasible trajectories [47].

3. Dataset

This section describes the data released in this work. This dataset includes: i.) 234 hours of video concurrently recorded from 234 cameras. ii.) Scene information for each roadway direction of travel in each camera. iii.) A unified curvilinear coordinate system aligned with the primary roadway direction of travel. iv.) Ground truth GPS trajectories for 270 vehicle runs through the camera fields of view v.) Object detections produced at 30Hz on the video scene. Each is described in more detail in the following sections.

3.1. Video Data

3.1.1 Location

Video of a single complex traffic scene was recorded using the I-24 MOTION traffic testbed [18]. Briefly, this system is comprised of 294 IP pan-tilt-zoom cameras densely covering a 4.2 mile stretch of 8-10 lane interstate roadway near Nashville, Tennessee. The main system features 40 ∼110-foot tall traffic poles, each with six cameras mounted to provide seamless coverage of roughly 500 feet of the interstate. The primary goal of this camera system is to provide accurate, anonymized vehicle trajectory and dimension information to enable traffic science. See [18] for more details. Figure 2 provides an overview of system features and a typical camera coverage layout for a single camera pole. Due to the layout of the cameras, any object passing through the whole system is visible in a minimum of 185 cameras, and roughly 1-3 cameras at any point in time with a few exceptions for overpasses and camera pole outages.

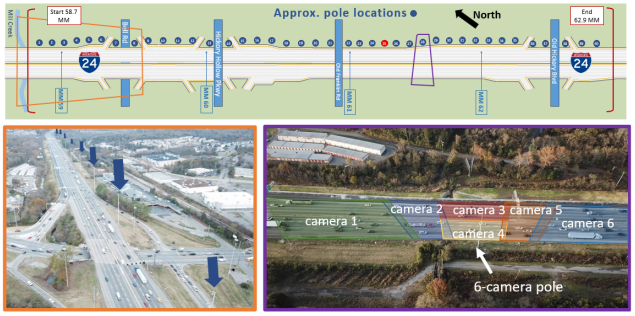
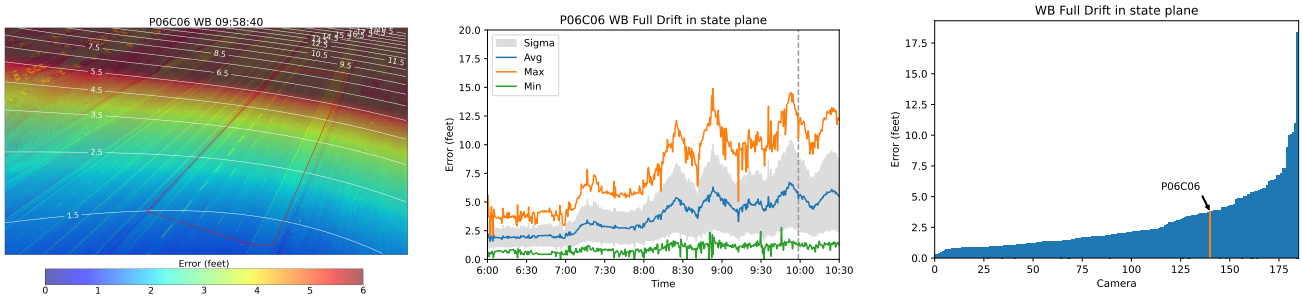


Figure 2. **(top)** Graphical overview of the I-24 MOTION system. Each blue dot represents a camera pole with 6 cameras. Red dot indicates a camera pole outage (Pole 25). **(orange)** drone image showing 8 of the 40 system camera poles. **(purple)** Typical 6 camera per pole coverage layout. Best viewed zoomed-in.

3.1.2 Recording Details

On a morning in November 2022, video data was recorded from 234 of the 296 cameras simultaneously from 6:00AM to 10:30AM, roughly covering the morning rush hours. The 7:00-8:00AM hour is published here. HD video (1920 × 1080 pixels) was recorded at 30 frames per second from each of the cameras and stored in H.264 compressed format, totaling ∼1 TB. Example videos can be found in the supplement. As in [44], any visible license plates are redacted using [38]. Each video is then manually inspected, to remove



(a) (b) (c)

Figure 3. Typical homography error dynamic and the representation of the Sunflower Effect: **(a.)** uncorrected displacement of image points showing the magnitude of error (in feet) that using the original homography without accounting for drift would cause. The red polygon area represents the camera FOV, **(b.)** The displacement error of a typical camera over the day. Gray vertical line indicates the time instant shown in (a). **(c.)** Mean average displacement of all the cameras for the westbound roadway side, sorted by magnitude of error.

any pedestrians, private property, or other personally identifiable information (see supplement). The one-hour scene has notable features, including i.) several anomalous events, including at least 10 stopped vehicles, ii.) high object density (>500 objects present at most times during the recording), and iii.) significant occlusion of vehicles by taller vehicles with moderate frequency.

3.2. Scene Homography

Particular care with scene information is taken in this work as an accurate transformation from image pixel space to a unified coordinate system is a vital pre-requisite for precise multi-camera tracking. The standard approach [20] utilizes a *homography*, which relates two planar surfaces via a linear transformation, in this case the road surface visible within camera frames and a suitable world coordinate system (We use Tennessee State Plane coordinates (EPSG:2274), which are preferred to other systems such as WSG84 (standard GPS convention) in that they utilize a globally orthonormal basis.) The road surface is treated as a planar surface (for each direction of travel) within a limited *field of view* (FOV) for each camera. Intrinsic-extrinsic camera calibration as used in the AV context [8, 16, 42] is infeasible here as cameras were not accessible prior to installation, can be replaced or moved, the focus is not fixed, and in-situ intrinsic camera calibration is not possible.

To compute each homography, lane marking corners are utilized as well-defined, semantically meaningful features. World coordinate system points are obtained by manually labeling aerial survey footage (∼1 inch/pixel), while the corresponding image coordinates are produced with semi-automatic labeling on the recorded images. Manual aid was required because the lane markings are identical and repetitive, so additional visual clues were required to uniquely label each lane marking. The homography matrix is fit to these correspondence points via a least-squares formulation as implemented in OpenCV [6]. See supplement for details.

3.2.1 Homography Re-estimation

Ideally, homographies ensure that multiple views of the same point map to a single unique point on the state plane. In reality, camera fields-of-view are constantly changing due to inaccuracies in the pan-tilt mechanism during homing, settlement of the foundation, and most significantly the *sunflower effect* (the tilting of metal infrastructure poles away from the sun due to differential heating of the sun and shade-facing sides of the pole) [13]. Uncorrected, these factors produce significant homography errors sometimes greater than 10 feet. Figure 3a shows the magnitude of these shifts at one time for a typical camera homography. Figure 3b illustrates how the average shift for a camera changes over time, due to both the initial error (due to long term phenomena since the initial camera calibration) and the fluctuations in error over a single morning due to the rising morning temperature and changing cloud cover (peaks and valleys).

Repeated manual correction is not feasible, and proper correction of the camera movement is challenging because traditional video stabilization methods (utilizing feature-matching techniques, based on e.g. SIFT [30] or SURF [1]) are ill suited for our scenes; a.) a large portion of the image corresponds to "noise" (e.g. trees, grass), producing hard-to-match feature points, b.) feature points are usually not semantically meaningful and potentially do not lie on the plane of the road surface, thus are unsuitable for homography estimation, c.) the relevant features on the ground plane in the region of interest are frequently occluded by vehicles, d.) a large number of co-moving vehicles can skew the calculation of optical flow along the direction of vehicle travel. To circumvent these issues, we propose the following homography re-estimation procedure:

1. Average frames for a suitable time (∼ 1 min) to remove vehicles from the scene.
2. Find an initial, rough alignment based on a SIFT and a FLANN-based matcher [36] (as in OpenCV [6]).

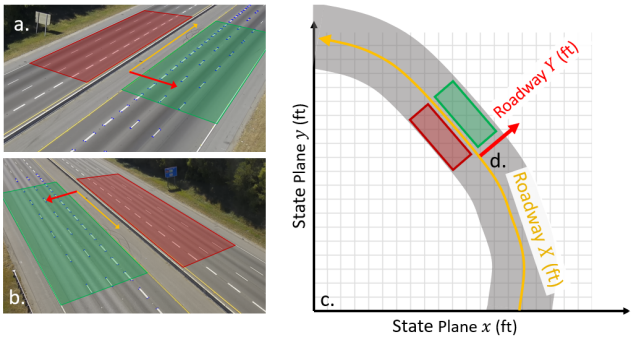


Figure 4. Camera fields of view (a and b) are related to (c) state plane coordinates, a rectilinear coordinate system, via perspective transforms. State plane coordinates are related to curvilinear roadway coordinates (d) via straightforward mathematical equations.

3. Shift original correspondence points using rough alignment. Use to seed re-detection of lane markers.
4. Filter and refine the detected lane marker corner points.
5. Calculate the homography matrix using successfully re-identified corner points.

For a specific time instance this automatic re-detection and homography re-estimation often fails, either due to i.) lane marking occlusion in heavy traffic or ii.) failure of FLANN matcher. To provide a robust homography in spite of these failures, two methods are proposed and implemented: i.) calculation of a single, *static homography* for an extended period (e.g. all-day) by filtering and averaging homographies over the period, and ii.) a *dynamic, time-varying homography*. The later a time-varying kernel-based filter of the homography parameters, with a variable window size. Each method is computed offline (utilizing all information for the whole day). Additional details are given in the supplement. The effectiveness of each solution is compared to the existing approach (FLANN-based matcher) in Section 4.1.

3.3. Roadway Coordinate System

We define an additional roadway coordinate system with the primary (X) axis aligned with the roadway direction of travel, and the secondary(Y) axis always perpendicular to the roadway direction of travel. Since the roadway is not perfectly straight, a *curvilinear coordinate system* is required to achieve the desired attributes, resulting in a locally orthonormal coordinate basis (see Figure 4 for a comparison). Such a coordinate system enables strongly domain-informed filter-based trackers [4] to be implemented trivially (e.g assume that the primary direction of motion for objects is along the primary axis and enforce reasonable vehicle physics). This coordinate formulation is also preferred for traffic science because quantities such as lane position and inter-vehicle spacing within a lane can be easily computed. A full description is given in the supplement.

3.4. GPS Tracks and Correction

Concurrent with video recording, a fleet of 103 GPS instrumented vehicles was driven through the portion of roadway observed by the I-24 MOTION testbed. Details on vehicle instrumentation can be found in [7]. On these vehicles, positional data was recorded at 0.1s intervals. A total of 270 vehicle passes through the roadway were made during the recording period, providing the same number of vehicle trajectories for comparison.

3.4.1 GPS Track Refinement

Initial attempts to compare GPS track data against known, ground truth object positions (manually annotated) revealed that GPS data contained positional errors (mainly bias along primary direction of travel, and mainly high variance perpendicular to direction of travel), consistent with the GPS sensor's reported error metric of 2.5m *circular error probable* (CEP) (see Figure 5). Additionally, a small time discrepancy between some GPS track data and the camera network is observable. The following protocol was utilized to make GPS trajectories suitable for direct comparison against object tracking outputs from camera data:

1. Manually annotate a 'perfect' position for each GPS track, once per camera pole (e.g. 37+ annotations for a GPS track that travels the full length of the camera system). See Figure 5.
2. Correct GPS bias in the roadway coordinate system primary (longitudinal) axis direction by finding the mean offset between GPS positions and manually annotated object positions.
3. Determine the time offset in the range [-2s,2s] that minimizes the variance in GPS positional offsets relative to manually annotated object positions.
4. Correct residual error in the longitudinal direction by linearly interpolating the required offset between consecutively labeled offsets between GPS and manually annotated object positions.
5. Linearly interpolate lateral coordinate between manually annotated object positions for each GPS track.

Figure 5 shows the alignment between manually annotated object positions (circles) and GPS positions (lines) for a single typical GPS track. In total 7885 manual annotations are made. Figure 6 shows a histogram of GPS intersection-over-union alignment with object detections (see Section 3.5) before and after correction. Corrected GPS tracks align more closely with CNN-produced object detections than original GPS tracks (IOU of 45% vs 8%). After correction, 270 vehicle trajectories were produced with an average length of 6.6 minutes and 17560 feet. Each object is virtually always visible in at least one camera, corresponding to a minimum of 3207600 roughly annotated bounding boxes.

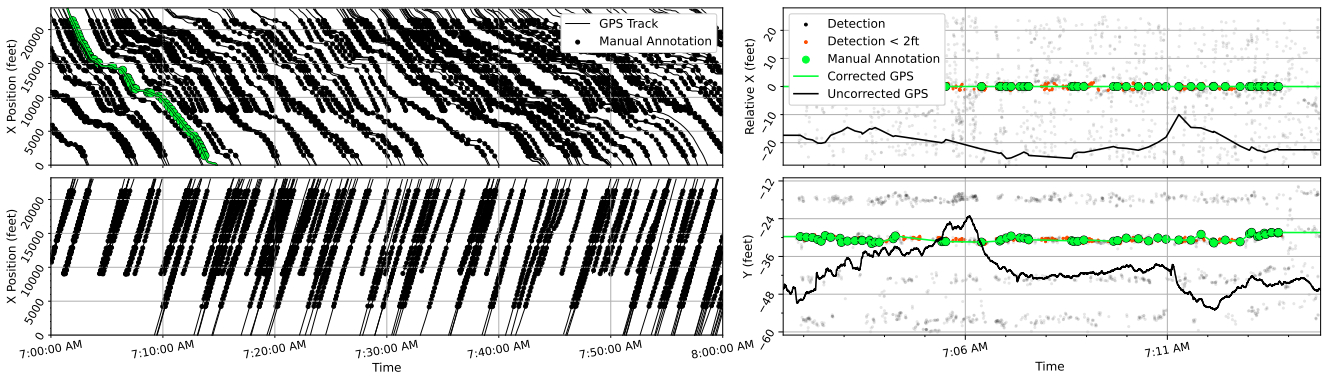


Figure 5. **(left)** GPS tracks (lines) and corresponding manual annotations (circles) for westbound (top) and eastbound (bottom) roadway directions of travel. One GPS trajectory is highlighted in green. **(right)** Detail for highlighted trajectory, showing relative x-position (top) and y-position (bottom) of nearby object detections (black dots), manual annotations (green circles), and the uncorrected corresponding GPS track. Deviations of over 20ft x / 12ft y position can be seen. Detections closely matching corrected GPS track shown in red. (Detections for every 30th frame are plotted for clarity.)

3.5. Detections

To allow preliminary analysis of existing object tracking methods, a baseline set of object detections was produced. Because the cameras in this dataset have widely varying fields of view, a viewpoint agnostic monocular object detector was utilized (i.e. an object detector that does not explicitly or implicitly code scene information into its structure or parameter weights). This allows a single set of network parameters to be utilized for all camera fields of view (rather than training a separate model for each camera field of view, which was infeasible based on storage, implementation, and training time constraints). This work utilized a Retinanet ResNet50-FPN backbone object detector [29] to provide detections. The network outputs were parameterized to produce rectangular prism representations for 3D bounding boxes in addition to 2D bounding box outputs for predicted objects (see supplement). Detections are nominally produced at 30 Hz with some frames skipped to provide $\pm$ 1/60s synchronization across all cameras. The resulting dataset contains 158,976,915 detections, each including a 3D bounding box defined in the roadway coordinate system, a 2D bounding box in image coordinates, vehicle class (*sedan*, *midsize*, *van*, *pickup*, *semi truck* or *other truck*), timestamp, camera, and detection confidence.

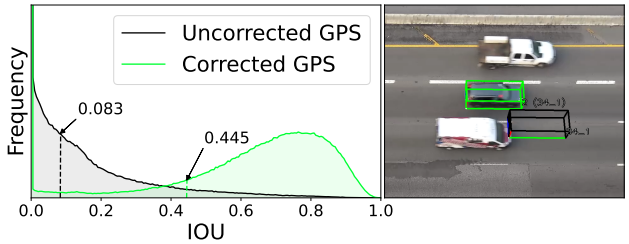


Figure 6. **(left)** Intersection-over-union histogram between GPS and closest automatically detected object position, before (black, mean 0.083) and after (green, mean 0.445) manual correction. **(right)** Examples of corrected (green) and uncorrected (black) GPS positions in a camera field of view.

4. Experiments

This section first describes experiments used to assess the accuracy of the homography re-estimation method proposed in this work, then describes initial MOT algorithm benchmarking performed using baseline object detections.

4.1. Homography Re-estimation

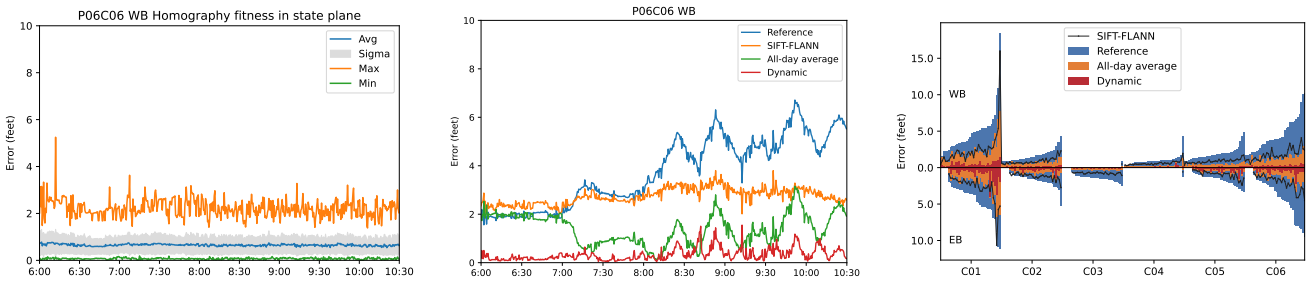
To assess the effectiveness of homography re-estimation methods proposed in Section 3.2.1, we utilize the homography goodness-of-fit (equation 1) which indicates how well the homography maps between the image plane and state plane, and the error metric defined in equation 2 which indicates average the positional error in points translated between the image plane and state plane via the computed homography. For each method, homographies are computed at 1 minute intervals overlapping by 50%. The computed homography's fitness is assessed according to:

$$fitness(t) = ||\mathcal{A}_t, \mathcal{I}'_t \xrightarrow{H_t} ||_2 \quad (1)$$

where $\mathcal{I}'_t$ is the subset of correspondence points successfully rediscovered in the image at time t, A_t is the corresponding subset of points in state plane coordinates, $\xrightarrow{H_t}$ indicates a linear transform between coordinate spaces using H_t, the homography matrix fit directly to the rediscovered points at time t. Error is computed as:

$$error(t) = ||\mathcal{I} \xrightarrow{H_t}, \mathcal{I} \xrightarrow{H^*_t} ||_2 \quad (2)$$

where $\mathcal{I}$ is the full set of correspondence points labeled in the original reference image, H_t is the homography fit directly to time t between the rediscovered points I'_t and the



(a) (b) (c)

Figure 7. **(a.)** Typical homography *fitness* for a single camera, **(b.)** error dynamics for a single camera over time with each homography re-estimation methods, **(c.)** Remaining error for each camera after (black) SIFT-FLANN feature-matching, (orange) one-day best fit homography re-estimation, and (red) dynamic homography re-estimation methods relative to orignal reference homography baseline (blue). Cameras are grouped by position on pole (see Figure 2) and by side of roadway (westbound homographies on top, eastbound on bottom).

corresponding state plane points A_t, and H_t^* is the homography for time t produced by the selected method. Because H_t is prone to error, any reported error may come either from the instantaneous homography H_t or the method-fit homography H_t^* (i.e. H_t is a good baseline when sufficiently many correspondence points are rediscovered.) We report other metrics independent from H_t in supplement.

4.2. MOT Algorithm Benchmarking

A limited set of detection-fusion tracking algorithms (SORT [4], IOUT [5], KIOU [10], and ByteTrack with both Euclidean distance and IOU as similarity metric [55]) is implemented based on the criteria that i.) algorithms must not require retraining on the tracking data as no training data for object detection is made available, ii.) must not require additional inputs (e.g. appearance embeddings), and iii.) must be tracking by detection-based (not joint detection and tracking-based) methods. These criteria are necessary because, on a dataset of this size, generating auxiliary information or conducting one-off algorithm runs on all videos is prohibitively time-intensive. For comparison, an *oracle* tracker is implemented which selects all detections close to a GPS trajectory and linearly interpolates tracklet positions between these selected positions. The oracle represents performance theoretically obtainable using the existing set of object detections with a perfect motion model. This evaluation is merely a first step at gauging the difficulty of this dataset; we make annotations and evaluation protocols public so that researchers may evaluate their own algorithms and report state of the art performance.

Tracking methods are evaluated using recall, assigned IDs per ground truth trajectory, and *Multiple Object Tracking Precision* (MOTP) in terms of both IOU and Euclidean distance from [3], *Longest Consecutive Subsequence* (LCSS) by distance and time from [46], and DetA, AssA, and HOTA from [31]. Because the dataset does not densely label objects, a false positive count cannot be obtained. Thus, the DetA metric from [31] is modified:

$$DetA_{\alpha}^{*} = \frac{TP}{TP + FN} \tag{3}$$

where TP represents the number of object positions that are matched to a ground truth position with at least α IOU overlap, and FN represents the number of ground truth object positions with no such match. We follow the rest of the protocol from [31] for calculating AssA and HOTA.

4.2.1 Evaluation Protocol

Each object tracker is run using the detection set from Section 3.5. GPS trajectories and detections from each camera are obtained at slightly different times. To account for this, tracking evaluation is performed at fixed 0.1 second intervals, and each GPS trajectory and object tracklet position is linearly interpolated at each evaluation time. Evaluation is performed as in [3]. For all metrics other than HOTA metrics, a lax IOU of 0.1 is required for an object tracklet and GPS trajectory to be matched.

5. Results

5.1. Homography Re-estimation Performance

Figure 7a reports the homography (H_t) goodness-of-fit metric (equation 1 over time for a typical camera using the homography re-estimation process defined in Section 3.2.1. This $\sim$2ft tightness is guaranteed by outlier removal processes during homography fitting; remaining error is due primarily to camera lens distortions and errors in the flat-plane assumption. The fitness of H_t represents an "error floor" for a homography based on the same assumptions.

Figure 7b show the additional error above the error floor for different homography re-estimation methods utilizing the error metric from equation 2. The reference (blue) indicates the resulting error without any mitigation, showing both long term (high mean) and short term (high variance) error (3.78 feet whole-day average). The SIFT-FLANN

Tracker	HOTA	DetA	AssA	Recall	IDs/GT ↓	LCSS$_t$ (s)	LCSS$_d$ (ft)	MOTP$_i$	MOTP$_e$ (ft) ↓	TD (s)
SORT [4]	**9.5**	51.3	**1.8**	73.6	53.1	**51.9**	2609	68.0	**2.70**	12.3
IOU [5]	1.1	7.4	0.2	20.4	60.0	16.8	53.2	36.7	7.31	8.4
KIOU [5, 10]	8.5	51.2	1.4	73.9	**47.9**	40.6	2181	66.9	2.72	**15.1**
ByteTrack (L2) [55]	**9.5**	51.5	**1.8**	73.6	53.3	51.5	**2575**	**70.0**	2.71	12.4
ByteTrack (IOU) [55]	8.5	**53.1**	1.4	**75.9**	50.3	44.1	2390	67.1	2.72	14.9
Oracle	53.1	55.1	51.0	86.4	1.2	636	14699	75.3	2.53	690

Table 2. Tracking results for limited benchmark algorithm set. For each metric, a higher score is better unless indicated with a ↓. DetA and AssA indicate the detection and association components of HOTA, respectively. LCSS denotes the average longest consecutive subsequence (in seconds or feet) averaged across all trajectories. MOTP indicates the average precision (by IOU of object footprint or Euclidean distance) for all matched object bounding boxes, averaged over all trajectories. TD indicates mean tracklet duration.

method (the existing optical flow-based "camera stabilization" [6]), is inferior (2.74 feet whole-day average) in almost all cases to the proposed methods utilizing semantically meaningful lane markers. The static, all-day average homography removes the long term error, although it is mostly unable to remove the error caused by the *sunflower effect* especially in highly fluctuating cases (1.39 feet whole-day average). Lastly, the dynamic homography utilizes nearby (temporally) homography estimations for a given time instance, and can cope with short-term fluctuations caused by camera pole movement, substantially reducing (0.33 feet whole-day average) the residual error caused by the static homography.

Figure 7c compares the whole-day average error, per camera, for each homography re-estimation method. The SIFT-FLANN based method (black line) improves on the reference homography (baseline) for 98.6% of cameras. The static all-day reestimated homography (green) improves on the baseline for 100% of cameras and outperforms the SIFT-FLANN method for 88.1% of cameras. The dynamic homography method (red) improves upon the baseline in 100% of cases and on the SIFT-FLANN method in 99.7% of cases. The mean average error over all cameras is 2.78 feet for the reference homography , 1.42ft for the SIFT-FLANN method (49% reduction), 1.03ft for the all-day average method (63% reduction), and 0.33 for the dynamic method (88% reduction).

5.2. Multiple Object Tracking Performance

Table 2 shows multiple object tracking performance for the implemented trackers. First, note that HOTA is quite low for all trackers; driven primarily by low AssA scores. This indicates that object tracklets are not strongly persistent (this is also supported by the relatively low LCSS and mean tracklet durations compared to the 6.6 minute mean trajectory length, and high average IDs per ground truth). Such high fragmentation means the tracking outputs are not useful for traffic science applications requiring long and accurate object tracklets. All trackers with a motion model (all but IOU) achieve higher mean recall than raw object detections of 44.5% (see Figure 6), which indicates that the motion model is crucial for filling in object positional information when detections are missing.

The purpose of this initial benchmarking is not to claim that no existing tracker can perform well on the I24V dataset, but rather to show that popular off-the-shelf methods are not suitable without substantial enhancement such as more strongly physics and scene-informed models. For instance ByteTrack [55] achieves high performance on datasets such as MOTChallenge, where ID switches and fragmentations play a relatively smaller role in overall scores, but performs poorly on this dataset where object persistence plays a more outsized role in overall tracking performance, especially in the AssA component of HOTA.

6. Conclusion

This work introduced the I24-Video Dataset, with concurrent video from 234 cameras recorded for one continuous hour capturing rush-hour traffic along 4.2 miles of interstate roadway, scene information for each camera, and 270 manually corrected GPS trajectories within the video data. These GPS trajectories were used to perform a preliminary benchmarking of object tracking algorithms, indicating that trackers utilizing stronger motion and appearance models are crucial for high performance on this dataset. The work also introduced new methods for keeping traffic camera homographies more precisely synchronized over time than existing methods allow. In the future, we plan to use this dataset to explore and design additional tracking algorithms that prioritize long term (10 minute, 18000 frame) object persistence, necessary for many traffic science applications. Several additional hours of GPS data are also recorded for future public benchmark competitions.

Acknowledgements

This work is supported by NSF grant Nos. 2135579 and DGE-1937963, USDOT Grant Nos. 693JJ32245006 and 693JJ322NF5201, DOE award No. CID DE-EE0008872, and CMAQ award No. 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] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 4

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 2

[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. 7

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016. 2, 5, 7, 8

[5] 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. 2, 7, 8

[6] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 4, 8

[7] Matthew Bunting, Rahul Bhadani, and Jonathan Sprinkle. Libpanda: A high performance library for vehicle data collection. In *Proceedings of the Workshop on Data-Driven and Intelligent Cyber-Physical Systems*, DI-CPS'21, page 32–40, New York, NY, USA, 2021. Association for Computing Machinery. 5

[8] 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. 1, 2, 3, 4

[9] 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. 2, 3

[10] S. Chen and C. Shao. Python implementation of the kalman-iou tracker., 2017. https://github.com/siyuanc2/kiout. 2, 7, 8

[11] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1, 2

[12] Shiloh L Dockstader and A Murat Tekalp. Multiple camera tracking of interacting and occluded human motion. *Proceedings of the IEEE*, 89(10):1441–1455, 2001. 3

[13] Wireless Estimator. A safe out of plumb monopole is most likely caused by the thermal 'sunflower effect', 2020. https://wirelessestimator.com/articles/2020/a-safe-out-of-plumb-monopole-is-most-likely-caused-by-the-thermal-sunflower-effect/. 4

[14] 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. 2, 3

[15] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):267–282, 2007. 2, 3

[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. 1, 2, 4

[17] Derek Gloudemans, Gracie Gumm, Yanbing Wang, William Barbour, and Daniel B. Work. The interstate-24 3d dataset: a new benchmark for 3d multi-camera vehicle tracking. *arXiv preprint arXiv:2308.14833*, 2023. 3

[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. 3

[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. 2, 3

[20] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[21] 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. 3

[22] 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. 3

[23] Hung-Min Hsu, Jiarui Cai, Yizhou Wang, Jenq-Neng Hwang, and Kwang-Ju Kim. Multi-target multi-camera tracking of vehicles using metadata-aided re-id and trajectory-based camera link model. *IEEE Transactions on Image Processing*, 30:5198–5210, 2021. 3

[24] Willie D Jones. Keeping cars from crashing. *IEEE spectrum*, 38(9):40–45, 2001. 1

[25] Nimet Kaygusuz, Oscar Mendez, and Richard Bowden. Multi-camera sensor fusion for visual odometry using deep uncertainty estimation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2944–2949. IEEE, 2021. 3

[26] Sohyeong Kim, Georg Anagnostopoulos, Emmanouil Barmpounakis, and Nikolas Geroliminis. Visual extensions and anomaly detection in the pneuma experiment with a swarm of drones. *Transportation Research Part C: Emerging Technologies*, 147:103966, 2023. 3

[27] Li Li, Rui Jiang, Zhengbing He, Xiqun Michael Chen, and Xuesong Zhou. Trajectory data-based traffic flow studies: A revisit. *Transportation Research Part C: Emerging Technologies*, 114:225–240, 2020. 1

[28] Martijn C Liem and Dariu M Gavrila. Joint multi-person detection and tracking from overlapping cameras. *Computer Vision and Image Understanding*, 128:36–50, 2014. 3

[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. 6

[30] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 4

[31] 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. 7

[32] 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. 3

[33] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple object tracking: A literature review. *Artificial intelligence*, 293:103448, 2021. 2

[34] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8844–8854, 2022. 2

[35] 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. 1, 2

[36] Marius Muja and David G Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331-340):2, 2009. 4

[37] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 145–161. Springer, 2020. 2

[38] Sergio M Silva and Cláudio Rosito Jung. A flexible approach for automatic license plate recognition in unconstrained scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5693–5703, 2021. 3

[39] Apoorv Singh. Transformer-based sensor fusion for autonomous driving: A survey. *arXiv preprint arXiv:2302.11481*, 2023. 3

[40] Andreas Specker, Daniel Stadler, Lucas Florin, and Jurgen Beyerer. An occlusion-aware multi-target multi-camera tracking system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4173–4182, 2021. 3

[41] 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. 3

[42] 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. 2, 4

[43] Hua Tang. Development of a multiple-camera tracking system for accurate traffic performance measurements at intersections, 2013. 3

[44] 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. 1, 3

[45] 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. 3

[46] Kevin Toohey and Matt Duckham. Trajectory similarity measures. *Sigspatial Special*, 7(1):43–50, 2015. 7

[47] 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. 3

[48] 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. 2

[49] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2

[50] Hao Wu, Xinxiang Zhang, Brett Story, and Dinesh Rajan. Accurate vehicle detection using multi-camera data fusion and machine learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3767–3771. IEEE, 2019. 3

[51] Minye Wu, Guli Zhang, Ning Bi, Ling Xie, Yuanquan Hu, and Zhiru Shi. Multiview vehicle tracking by graph matching model. In *CVPR Workshops*, pages 29–36, 2019. 3

[52] Kai-Siang Yang, Yu-Kai Chen, Tsai-Shien Chen, Chih-Ting Liu, and Shao-Yi Chien. Tracklet-refined multi-camera tracking based on balanced cross-domain re-identification for vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3983–3992, 2021. 3

[53] Shu Zhang, Elliot Staudt, Tim Faltemier, and Amit K Roy-Chowdhury. A camera network tracking (camnet) dataset

and performance baseline. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 365–372. IEEE, 2015. 2, 3

[54] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4537–4546, 2022. 3

[55] 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. 7, 8

[56] 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. 2

[57] 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. 2