2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

CAROM Air - Vehicle Localization and Traffic Scene Reconstruction from Aerial Videos

Duo Lu[1], Eric Eaton[1], Matt Weg[1], Wei Wang[2], Steven Como[2], Jeffrey Wishart[2], Hongbin Yu[2], Yezhou Yang[2]

***Abstract*— Road traffic scene reconstruction from videos has been desirable by road safety regulators, city planners, researchers, and autonomous driving technology developers. However, it is expensive and unnecessary to cover every mile of the road with cameras mounted on the road infrastructure. This paper presents a method that can process aerial videos to vehicle trajectory data so that a traffic scene can be automatically reconstructed and accurately re-simulated using computers. On average, the vehicle localization error is about 0.1 m to 0.3 m using a consumer-grade drone flying at 120 meters. This project also compiles a dataset of 50 reconstructed road traffic scenes from about 100 hours of aerial videos to enable various downstream traffic analysis applications and facilitate further road traffic related research. The dataset is available at https://github.com/duolu/CAROM.**

I. Introduction

Road traffic has created many problems that need to be studied with real-world road traffic data. For example, local Departments of Transportation (DOTs) need to count the vehicles on every major road segment for traffic management purposes. It is desirable that each counted vehicle in the data can have fine-grained attributes, such as vehicle type, speed, lane, etc. City planners and transportation system engineers also want to use detailed road traffic data for better decision-making and resource provisioning. Additionally, for researchers and regulators interested in road safety analysis and driver behavior modeling, it is more valuable to capture the comprehensive motion states of vehicles passing through a specific traffic scene instead of obtaining just a count (which does not carry much information) or a crash report (which happens infrequently). For example, aggressive lane switches and frequent close call incidents on a highway segment may indicate that the traffic is reaching the designed capacity. Similarly, reckless driving behaviors can reveal more insights on road safety than reported accidents. Besides the policy makers, vehicle manufacturers and insurance companies can also benefit from datasets of vehicle trajectories, especially if such data can be used to accurately reconstruct and re-simulate the captured traffic scenes.

Traditionally, such road traffic data is collected and managed by the DOTs using devices installed on the road infrastructure. There are a few problems. First, it is expensive and unnecessary to cover every mile of the road with sensors and cameras. As a result, many interesting traffic scenes are not captured. Second, although there are many cameras deployed in strategic locations in major cities, it is challenging to process the videos or deliver them over the network due to the sheer volume of the videos. Hence, the operational cost of these cameras further hinders large-scale deployment. Meanwhile, since the cameras cannot move, the captured video data contain redundant information from repeated patterns. Third, the vehicle localization accuracy of infrastructure-based sensors can significantly degrade when a tracked vehicle is far away or occluded by other vehicles. Fourth, due to regulations and privacy issues, it is challenging for researchers outside the DOTs and industrial partners to access these video data for open research purposes. On the other hand, for independent researchers or companies, it is expensive to collect and manage road traffic data. Last, for the researchers and companies who can afford to collect data on the road using cameras and LIDAR sensors, it is time-consuming to label the data to train neural network models for vehicle detection and tracking. This is particularly an issue if the labeling must be done in the 3D space.

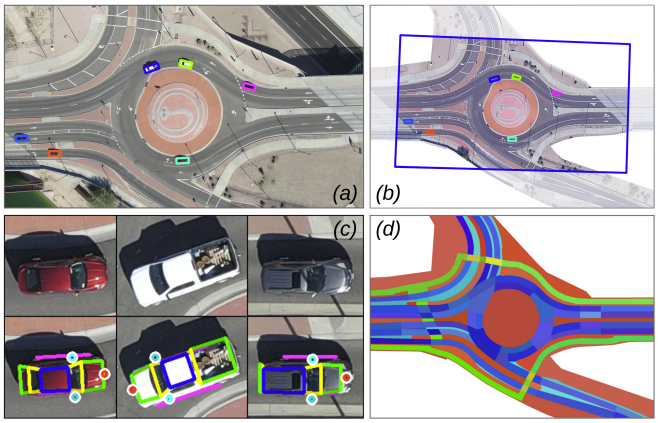


Fig. 1: An overview of CAROM Air: (a) tracked vehicles on the aerial video, (b) the reconstructed traffic scene, (c) vehicle keypoints, and (d) the map with semantic annotation.

To address these issues, we propose a framework named CAROM Air ("CARs On the Map tracked from the Air"), which digitizes and reconstructs road traffic scenes from aerial videos taken by drones, as shown in Fig. 1. It is inexpensive and flexible since it does not require any support from road infrastructure. The core of this framework is a pipeline that can track vehicles on the aerial videos and localize them on the map accurately through the detection

[1]D. Lu, E. Eaton, and M. Weg are with Rider University. {dlu, eatone, wegma}@rider.edu
[2]W. Wang, S. Como, J. Wishart, H. Yu, and Y. Yang are with Arizona State University. {wwang266, scomo, jeffrey.wishart, Hongbin.Yu, yz.yang}@asu.edu
This research is sponsored by Rider University Faculty Research Fellowship, MacMillan Scientific Research Fellowship, and the Institute of Automated Mobility (Arizona). We thank Arizona DOT, Greg Leeming, Marisa Paula Walker, Larry Head, Varun Chandra Jammula, Lu Zhao, Baihan Chen, Ke Fan, Nannan Zhang, Zhichao Li for their help.

of vehicle keypoints. This allows us to convert the aerial videos to vehicle trajectory data which can be delivered over communication networks for reconstruction or further analysis using programs. Such vehicle trajectory data does not have any personal identifiable information, and hence, they can be shared without causing privacy issues. Moreover, we demonstrate that our data can be used as reference measurements or 3D labels for videos and LIDAR point clouds captured by devices on the road infrastructure. This work is a continuation of the ongoing research conducted by the Institute of Automated Mobility (IAM) [1] to support the development and validation of an operational safety assessment methodology [2][3][4][5] and intelligent road traffic infrastructure [6][7][8]. In summary, our contributions are as follows.

1) We developed a keypoint-based vehicle tracking and localization pipeline for aerial videos. The average vehicle localization error is from 0.1 m to 0.3 m using a drone flying at 120 meters in various conditions.
2) We built a dataset of vehicle trajectories obtained from about 100 hours of drone video in 50 different road traffic scenes.
3) For each scene, we also provide the map with semantic segmentation at the lane level, which enables further automated traffic analysis and statistics.
4) We demonstrated several downstream applications to show the practicality of our framework.

II. Related Work

Unmanned Aerial Vehicles (UAVs), commonly called drones, have been used in 3D mapping of road infrastructures [9], traffic monitoring [10][11][12][13][14][15][16][17][18], road safety analysis [19], and transportation of humans or goods [20]. They are gaining popularity as an inexpensive and flexible method of obtaining aerial videos of road traffic scenes. To further process the videos, a pipeline of vehicle detection and tracking can be applied, typically with deep neural networks [21]. With such methods, researchers have constructed datasets of vehicle trajectories obtained from drone videos [22][23][24][25][26][27][28], which supplement existing large-scale autonomous driving datasets [29][30][31][32][33] and road infrastructure based traffic monitoring datasets [34][35][36]. These vehicle trajectory datasets further enable a series of analysis tasks, often with the help of a map containing lane-level traffic semantics [37][38]. Compared to existing works, our method provides better vehicle localization accuracy with more rigorous and more extensive evaluations. Besides, thanks to our keypoint-based vehicle localization algorithm, our framework has better flexibility in drone camera angles rather than requiring the camera to always look downward. Similar keypoint-based methods have been explored in 3D reconstruction [39][40][41][42][43][44][45] and autonomous driving from a ego vehicle's view point [46][47][48][49], but not from a drone camera. Additionally, our dataset is larger and more diverse than similar datasets from existing works.

III. The CAROM Air Framework

The CAROM Air framework contains three layers, as shown in Fig. 2. The foundational layer is a pipeline to track and localize vehicles captured on the aerial video (detailed in this section). The middle layer is the dataset of tracked vehicle trajectories and traffic scene maps with lane-level semantic annotation (section V). After that, the downstream applications form the third layer (section VI).

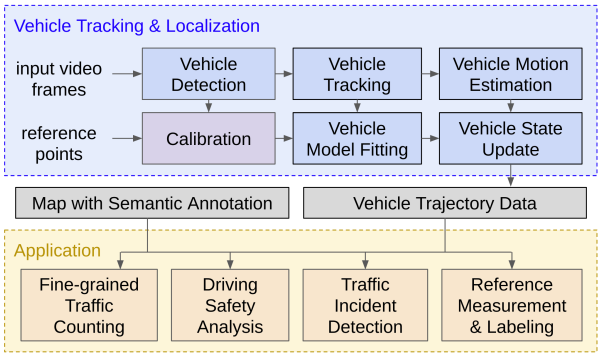


Fig. 2: The CAROM Framework Architecture.

A. Camera Calibration

The camera calibration step is illustrated in Fig. 3. We use a pinhole camera model with no distortion and a flat ground model, which can usually achieve enough accuracy. For each video track, we usually annotate 8 to 16 point correspondences on a satellite map (e.g., a screenshot on Google Maps) and a reference aerial image (typically the first image in a video). With these point correspondences, the 3D pose of the camera is solved through Perspective-n-Points (PnP) [50] given the camera intrinsics (calibrated in the lab) and the 3D coordinates of the points (computed using the scale of the map by assuming the annotated points are on the flat ground). Different from a stationary camera installed on the road infrastructure, the pose of the drone camera can drift. Hence, recalibration is needed for each video image. To achieve this, we detect the corners features [51] on the ground (denoted as the reference points in Fig. 2) for the reference aerial image, track them across the whole video, and recompute the camera pose using PnP. The semantic annotation of the map helps to determine those points on the ground, e.g., the map shown in Fig. 1(d). With the camera parameters and the map, we can back-project any image pixel to a 3D location if that pixel is on the ground.

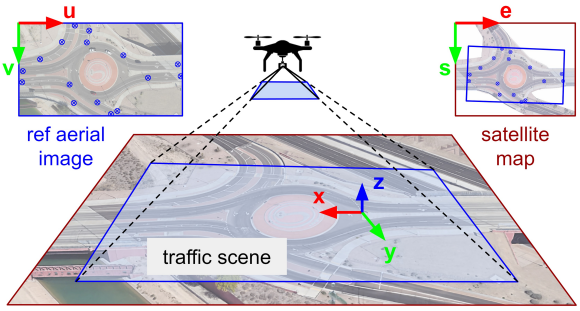


Fig. 3: An illustration of camera calibration.

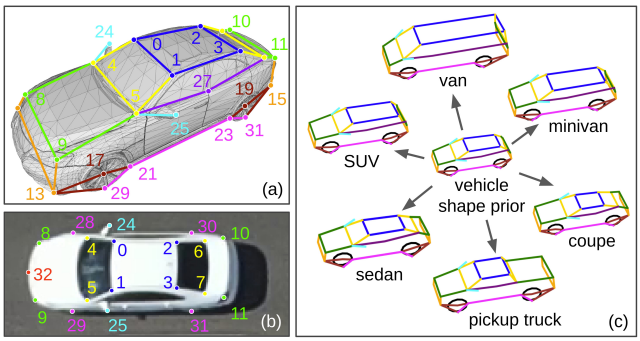


Fig. 4: Vehicle keypoints: (a) defined in 3D, (b) detected on the image, and (c) generated from a vehicle shape prior.

B. Vehicle Detection and Tracking

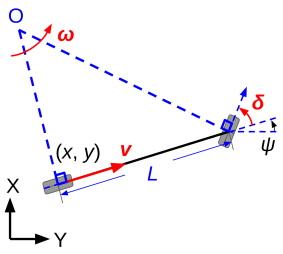
We use a Keypoint RCNN [52] to detect vehicle keypoints and bounding boxes on each image. We define 33 keypoints in total, as shown in TABLE I and illustrated in Fig. 4 (a). Keypoints are usually defined in groups of two (i.e., right-left) or four (i.e., front-right, front-left, rear-right, and rear-left). Among them, 19 keypoints are detected on the image, as shown in Fig. 4 (b). These keypoints are usually related to observable features such as corners. Hence, they can be reliably detected in most cases using a Keypoint RCNN trained from a small dataset constructed by us (4,386 images, about 12,000 vehicles in total). With the detected vehicle object instances on two adjacent video frames, we associate them if the intersection-over-union (IoU) of their bounding boxes exceeds a certain threshold (i.e., tracking by detection).

TABLE I: Definition of vehicle keypoints.

ID	detected?	keypoint definition
0 - 3	Yes	corners of roof top
4 - 7	Yes	corners of font and rear windshields
8 - 11	Yes	centers of front and rear lights
12 - 15	No	corners of front and rear bumpers
16 - 19	No	centers of wheels
20 - 23	No	corners of chassis bottom surface
24 - 25	Yes	outermost corners of side mirrors
26 - 27	No	corners of the front door windows
28 - 31	Yes	wheel-ground contact points
32	Yes	center of the brand logo in the front

C. Vehicle Model Fitting

We collected 200 vehicle 3D models from the Internet and annotated all 33 keypoints in 3D for each model. These 3D models include vehicles of various types, and an example is shown in Fig. 4 (a). We also preprocessed the 3D models to the actual scale of real-world vehicles. For each vehicle model, we concatenate the (x,y,z) coordinates of all 33 annotated 3D keypoints as a long vector (denoted as the shape vector $\mathbf{s}_i$). After that, we run Principal Component Analysis (PCA) [53] on the set of shape vectors $\{\mathbf{s}_i\}$ of all vehicles to find the mean shape (denoted as $\mathbf{s}_m$), the k basis vectors (denoted as the columns of a matrix W) corresponding to the k largest eigen values, and the k-dimensional parameter vectors $\{\mathbf{b}_i\}$, such that the reconstructed shapes



$$\delta_{k+1} = \delta_k \quad \text{- (1)}$$
$$v_{k+1} = v_k \quad \text{- (2)}$$
$$\omega_{k+1} = v_k \tan\delta_k / L \quad \text{- (3)}$$
$$\psi_{k+1} = \psi_k + \omega_k \Delta t \quad \text{- (4)}$$
$$x_{k+1} = x_k + v_k \cos(\psi_k) \Delta t \quad \text{- (5)}$$
$$y_{k+1} = y_k + v_k \sin(\psi_k) \Delta t \quad \text{- (6)}$$

Fig. 5: The simplified vehicle kinematic bicycle model (left) and state prediction rules (right).

$\{\hat{\mathbf{s}}_i = W\mathbf{b}_i + \mathbf{s}_m\}$ can approximate the original shapes $\{\mathbf{s}_i\}$. Similarly, we can generate a vehicle shape $\hat{\mathbf{s}}^* = W\mathbf{b}^* + \mathbf{s}_m$ from an arbitrary parameter vector $\mathbf{b}$. Shapes of vehicles of various types can be generated in this way, as demonstrated in Fig. 4 (c). The mean shape vector $\mathbf{s}_m$ and the matrix W are collectively called the **vehicle shape prior**.

Given a vehicle on an image, we try to find a parameter vector $\mathbf{b}$ and the vehicle pose $(R,\mathbf{t})$, such that the generated vehicle shape best fits the detected keypoints $\mathbf{p}$ under the camera projection $\Pi()$ obtained from recalibration, *i.e.*,

$$\arg\min_{\mathbf{b},R,\mathbf{t}} \sum_j^N \alpha^{(j)} ||\mathbf{p}^{(j)} - \Pi(R(W^{(j)}\mathbf{b} + \mathbf{s}_m^{(j)}) + \mathbf{t})|| + \lambda ||\mathbf{b} - \mathbf{b}_t||.$$

Here, N is the total number of detected keypoints (19 in our case); $\alpha^{(j)}$ is the visibility of the *jth* keypoint reported by the detector, i.e., either 1 (visible) or 0 (invisible); $\mathbf{p}^{(j)}$ is the pixel coordinates of the *jth* keypoint; $W^{(j)}$ and $\mathbf{s}_m^{(j)}$ are the vehicle shape prior components for the *jth* keypoint.

Assuming the vehicle is always on the flat ground (*i.e.*, the XOY plane), there are essentially three unknown variables in the vehicle pose, *i.e.*, the vehicle position (x,y) in $\mathbf{t}$ and the heading angle ψ in R (R is the rotation matrix along the z-axis by the angle ψ). With this parameterization, the model fitting problem is simplified to an unconstrained nonlinear least square problem, which can be solved efficiently using the Levenberg-Marquardt method [54]. The initial position of the vehicle is approximated by the center of the bounding box, and the initial heading of the vehicle is obtained using a set of vectors through random sample consensus (RANSAC) [55]. These vectors are derived from a set of keypoint pairs pointing in the vehicle's forward direction, such as {(0, 2), (1, 3), (4, 6), (8, 10), ...}. In fact, since $\Pi()$ is close to a weak perspective projection for aerial videos, if the initial estimation of the vehicle heading is reasonably accurate (which is usually the case), this problem is very close to a linear least square problem. Hence, it generally converges very fast (sub-millisecond in our implementation).

The last term $\lambda||\mathbf{b} - \mathbf{b}_t||$ is a regularizer, where $\mathbf{b}_t$ is the categorical "template" parameter vector. For examples, if the vehicle is detected as a sedan, $\mathbf{b}_t$ is the average of $\{\mathbf{b}_i\}$ from all sedans among the 200 vehicle 3D models that are used to construct the vehicle shape prior. Meanwhile, $\mathbf{b}_t$ is also used as the initial value of $\mathbf{b}$ in the optimization procedure. After the model fitting, we find the k-nearest-neighbor of $\mathbf{b}$ in $\{\mathbf{b}_i\}$ and use them to determine the type of the vehicle.

TABLE II: Tracking evaluation results.

Videos	MOTA (mask)	MME (mask)	FP (mask)	FN (mask)	MOTA (kp)	MME (kp)	FP (kp)	FN (kp)	#Objects	#Images	#Veh	IDE	MT	ML	VFP
track 1	98.1%	512	16	639	88.5%	430	1	11808	107140	29300	195	1	193	1	0
track 2	99.2%	0	73	1399	90.1%	0	3	17887	180438	42390	650	0	648	2	0
track 3	97.4%	35	943	1503	89.6%	10	405	9681	96975	42796	498	2	495	1	6

D. Vehicle State Estimation

The model fitting step provides us the position and orientation of each detected vehicle on every image of the aerial video. We further run an Extended Kalman Filter (EKF) with a simplified kinematic bicycle model, as illustrated in Fig. 5. The vehicle state prediction rules are listed in equations (1) to (6) in the figure. We assume that the vehicle maintains its steering angle and speed, i.e., equation (1) and (2). Among all the states, the position (x, y) and heading ψ are considered directly observable, while the other three states are hidden (highlighted in red in Fig. 5). The parameter vector $\mathbf{b}$ and the vehicle dimension are also estimated iteratively using the model fitting results in a similar way, assuming they do not change among images. The EKF approximation works well since the model fitting uncertainty is generally small and the vehicle motion between two adjacent frames is also small. Finally, the estimated states of all vehicles on all video images are exported as the vehicle trajectory data in Fig. 2.

E. Implementation Details

We built a prototype system that implements the proposed framework with a few small improvements. First, for some scenes, a piecewise flat ground model was used to better capture the uneven ground surface. The added cost is that more point correspondences are required to be annotated at carefully chosen places. Second, we augmented the camera recalibration step to a sparse monocular Simultaneous Localization and Mapping (SLAM) pipeline with key frame selection to improve the robustness. Third, we implemented a backup vehicle tracking and localization pipeline using the instance segmentation masks of vehicles, which is similar to [6]. When the keypoint detector misses a vehicle but the mask detector detects it, this backup pipeline works. Two additional estimators were implemented for the backup pipeline. When the vehicle heading can be obtained, an EKF with a point-mass and no-side-slip kinematic model is used. If the vehicle heading cannot be obtained, the Kalman Filter estimator in [6] is used.

IV. Empirical Evaluation

We conducted several experiments to evaluate the proposed framework and our prototype implementation. First, we evaluate the vehicle detection and tracking performance with three video tracks taken from three different scenes. The results are shown in TABLE II, most of the metrics are from [56]. Here, "#Veh" is the number of vehicles in the video track. "IDE" is the number of vehicles with tracking ID errors. "MT" is the number of vehicles that are tracked for over 80% of the time (i.e., "mostly tracked"). "ML" is the number of vehicles that are tracked for less than 20% of the time (i.e., "mostly lost"). "VFP" is the number of non-vehicle objects that are wrongly tracked as vehicles (i.e., "vehicle false positive"). A vehicle is considered "tracked" if it is either tracked by the proposed pipeline (using keypoints) or the backup pipeline (using masks). We only track vehicles on the traversable ground area labeled on the map, and we only assign a tracking ID to a vehicle if it can be detected and associated for at least five consecutive video images. We intentionally set a more strict score threshold for the keypoint detector so that there are less false positive and more false negatives (as shown in the "FN (kp)" column in TABLE II). In most cases, these false negatives can be handled by the backup pipeline with slight loss of vehicle localization accuracy. Overall, our prototype can track most of the vehicles correctly. Qualitative results and visualizations are available online in our GitHub repository [57].

TABLE III: Model fitting evaluation results.

Metric	x (m)	y (m)	ψ (°)	L (m)	W (m)	H (m)
Avg Error	0.092	0.084	0.891	0.075	0.044	0.099
Std Dev	0.113	0.090	1.055	0.108	0.047	0.116

Next, we quantitatively evaluate the model fitting performance. As shown in Fig. 6 (a), we parked a test vehicle in an empty lot and flew the drone at 85 meters. We moved the drone in a way such that the test vehicle can be seen in different places in the camera field of view (FOV), which is illustrated as the dashed yellow trajectory in the figure. Four large ArUco markers [58] were placed on the ground to facilitate camera recalibration. Three different test vehicles with known dimensions are used (a sedan, a hatchback, and an SUV), and six video tracks are collected (20 minutes in total). For each video track, we marked the four contact points of the wheels and the ground to derive the ground truth vehicle pose. The evaluation results are shown in TABLE III, averaging over all images from all video tracks. In this table, the first two data columns represent the position error in the vehicle's longitudinal direction (x) and lateral direction (y). The third data column means the error of heading angles (ψ) in degrees. The last three data columns are vehicle dimension errors (i.e., length, width, height) in meters. Additionally, in Fig. 6 (b), we show the average localization error across the camera FOV. These results show that the vehicle pose and shape can be captured precisely.

After that, we quantitatively evaluate the performance of vehicle localization and speed estimation. As shown in Fig. 6 (c), we drove two test vehicles equipped with differential GPS devices in an intersection and compared the GPS data with our results obtain from a drone at 120 meters. The differential GPS has a localization accuracy of about 2 cm, and its measurements are used as references. We drove each test vehicle across the intersection 24 times in various

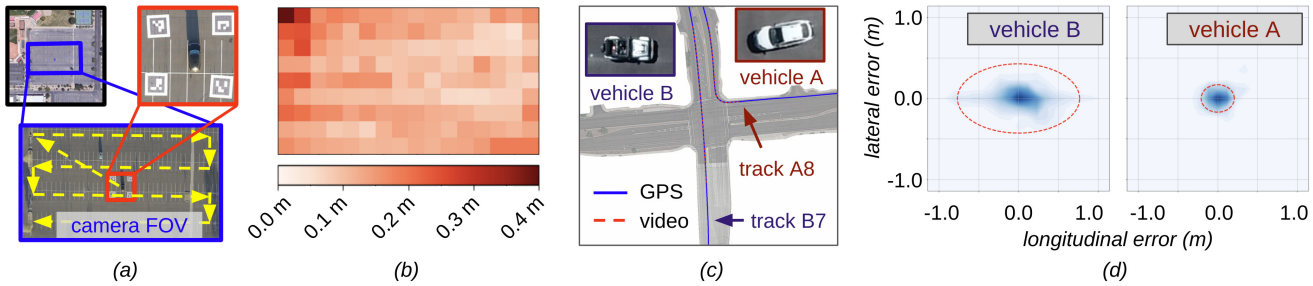


Fig. 6: Empirical evaluation: (a) experiment setting for model fitting evaluation, (b) vehicle position error in the camera FOV, (c) experiment setting for vehicle tracking and localization, and (d) vehicle location error in the ego vehicle frame.

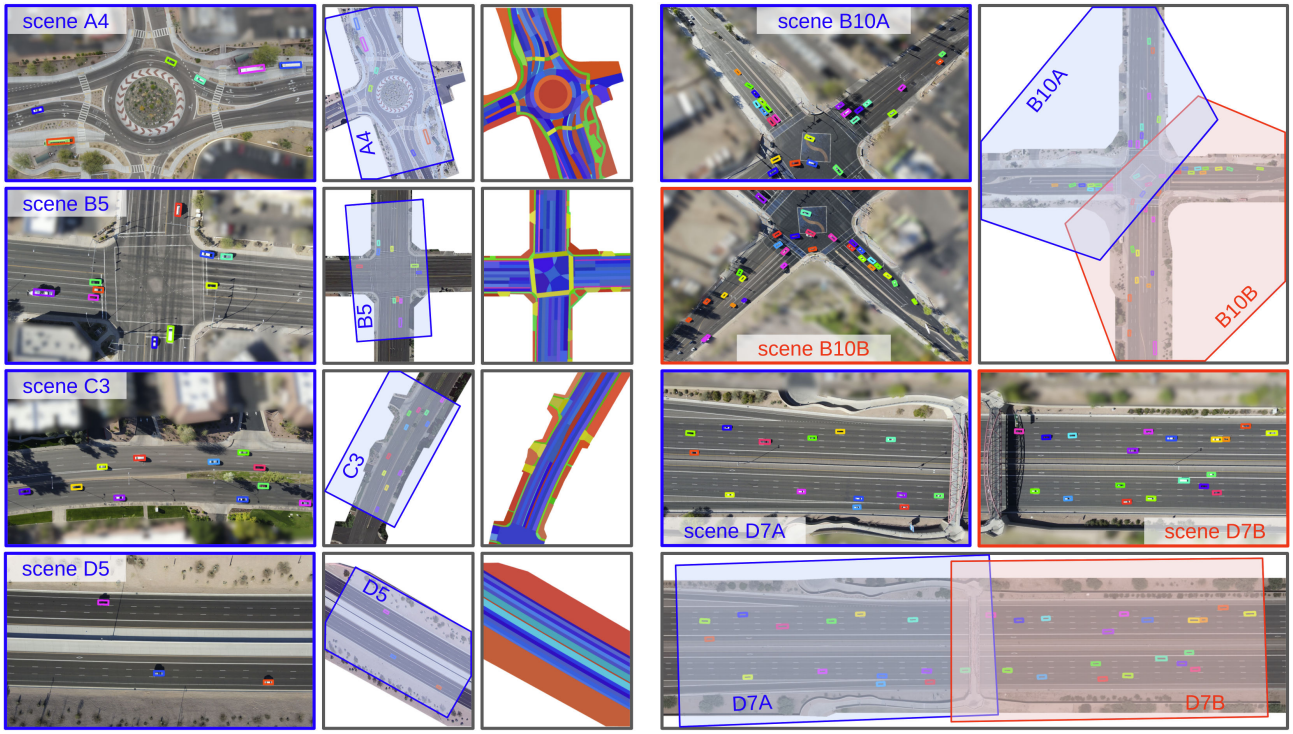


Fig. 7: Examples from the CAROM Air dataset.

directions. Two example trajectories are shown in Fig. 6 (c). With our prototype system, the keypoints of vehicle A can be reliably detected all the time, and the proposed pipeline is used. In contrast, the keypoints of vehicle B can only be detected occasionally, and the backup pipeline is used most of the time. The average location difference between our method and the reference is 0.10 m and 0.26 m for vehicles A and B, respectively. The average speed estimation difference is 0.22 m/s and 0.36 m/s for vehicles A and B, respectively. Additionally, the distribution of location differences in the vehicle's reference frame is shown in Fig. 6 (d). In this figure, the red ovals show the approximated two-sigma range, i.e., about 95% of the measurement differences are inside the ovals. These results indicate that our framework can localize vehicles accurately. We believe that the primary sources of errors are as follows: (a) camera lens distortion, (b) inaccurate drone camera pose estimation in recalibration, (c) ground flatness, and (d) keypoint detection errors. In some cases, under strong sunlight, the detector can also make mistakes with featureless black vehicles, vehicle shadows, and the specular reflection on the vehicle surface. Besides, sometimes vehicles with similar shapes are misclassified into the wrong types, *e.g.*, sedan to coupe, SUV to minivan, etc.

V. The CAROM Air Dataset

We constructed a vehicle trajectory dataset from about 100 hours of drone videos in 50 different traffic scenes covering a variety of traffic patterns, including roundabouts, intersections, local road segments, and highway segments. For a few scenes, we flew two drones simultaneously to cover larger areas, and we manually synchronize the videos using a flashlight visible from both drones. Besides the vehicle motion data, we also segmented the map at the lane level and annotated the type of these segmented areas, e.g., vehicle driving lanes, curb areas, sidewalks, crosswalks, buffer areas, etc. Examples are shown in Fig. 7. More details about the dataset content and the data format are available online [57].

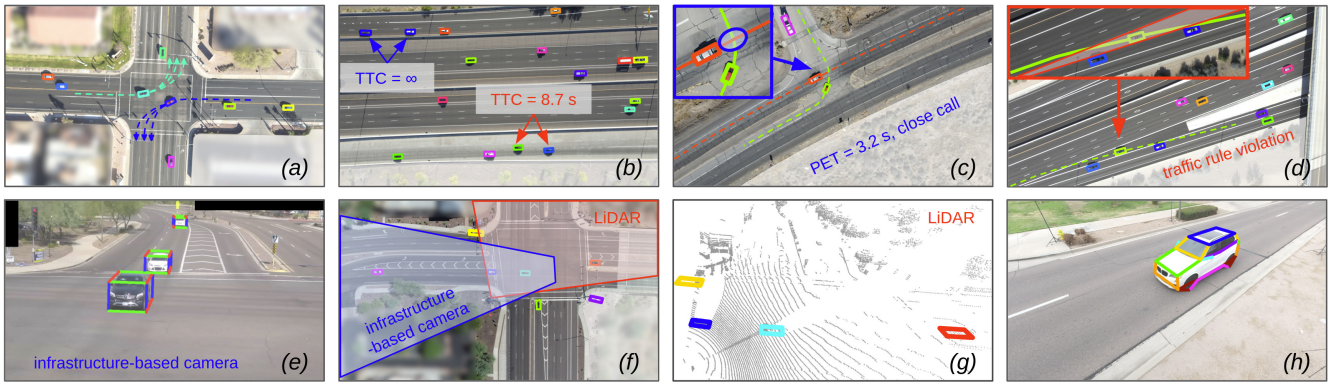


Fig. 8: Example applications of the CAROM Air framework.

VI. Applications

In this section, we demonstrate five different applications enabled by our framework and dataset.

(1) Fine-grained traffic counting. Traffic counting and statistical analysis are crucial for traffic management. Our framework can automate the counting and analysis at the lane level by utilizing a semantically segmented map. Each vehicle's trajectory can be converted into a list of map segments traversed by the vehicle, and a program can count vehicles that follow a specified pattern. For example, in Fig. 8 (a), on the southbound (blue trajectories), we observed that the percentages of left-turning vehicles that leave the intersection in the leftmost lane, middle lane, and the rightmost lane are 45%, 45%, 10%. On the northbound (cyan trajectories), the percentages are 23%, 55%, and 22%. Similarly, in Fig. 8 (b), we can obtain the speed of vehicles on each lane using the segmented map, which shows that 54% of the vehicles on the leftmost lane of the highway segment in both directions exceed the speed limit.

(2) Driving safety analysis. Various assessment metrics have been proposed to objectively evaluate driving safety [2][3][5]. For example, in Fig. 8 (b), utilizing our vehicle trajectory data and the segmented map of a traffic scene, we can compute the Time-To-Collision (TTC) metric [59] for any pair of adjacent vehicles in the same lane. Similarly, in Fig. 8 (c), given a pair of intersecting vehicle trajectories and the area of encroachment (shown as the blue-shaded circle), we can compute the Post Encroachment Time (PET) metric [60]. A low TTC or PET generally indicates unsafe driving behavior. Moreover, we can also "re-simulate" the motion of vehicles using our data, and then probe the safety envelope by changing the physical properties of the vehicle [61].

(3) Traffic incident detection. Researchers spend hundreds of hours studying traffic data, which is laborious and costly. With our framework, an automated program can search through the dataset and detect incidents of interest. For example, in Fig. 8 (d), a vehicle drives through an area separating the main lanes on the highway and the ramp (shown as the red-shaded area). This is a traffic rule violation. We can check whether each vehicle trajectory passes through that area on the segmented map in our dataset to detect incidents of this type. Similarly, in Fig. 8 (c), we can detect a close call incident if the PET is lower than a threshold or an aggressive driving incident if the acceleration of a vehicle is higher than a threshold.

(4) Reference measurement and labeling. In order to deploy cameras and LiDARs on road infrastructure to monitor traffic, effective neural network models are needed to detect and track vehicles. However, it is expensive to construct labeled datasets to train these models, especially if it is required to label vehicle 3D bounding boxes manually. With accurate cross-sensor calibration, vehicle trajectory data generated from our framework can be used as labels for the data obtained from other sensors or as reference measurements to evaluate the performance of other traffic monitoring systems [6][8]. For example, in Fig. 8 (f), we show the vehicle localization results on the aerial video. These results are projected onto an image obtained from an infrastructure-based camera in Fig. 8 (e). They are also shown in the 3D space together with the point cloud obtained from an infrastructure-based LiDAR in Fig. 8 (g).

(5) Generalization to roadside perspectives. We can also apply our keypoint-based vehicle localization method to videos from non-aerial perspectives. An example is shown in Fig. 8 (h). However, the keypoint detector must be trained with data from the same perspective. If some keypoints are not observable, *e.g.*, when a vehicle moves toward the camera or when it is partially occluded by another vehicle, more robust regularization is required for the model fitting step.

VII. Conclusions

This paper presents CAROM Air, a keypoint-based vehicle localization and traffic scene reconstruction framework using aerial videos recorded by drones. Our framework achieves decimeter-level localization accuracy and enables many practical downstream traffic analysis applications. Still, it has certain limitations, such as short flight time, restricted fly zones in cities, potential risks of drone crashes, etc. The drone camera also has a limited dynamic range, and the detector can produce errors on certain vehicles that appear infrequently in our training data (e.g., motorcycles, trucks, and trailers). With further development, we hope it can serve as a flexible method for road traffic analysis and eventually help improve road safety and transportation efficiency.

REFERENCES

[1] "Institute of Automated Mobility (IAM)," https://https://www.azcommerce.com/iam.

[2] J. Wishart, S. Como, M. Elli, B. Russo, J. Weast, N. Altekar, E. James, and Y. Chen, "Driving Safety Performance Assessment Metrics for ADS-equipped Vehicles," *SAE Technical Paper*, vol. 2, no. 2020-01-1206, 2020.

[3] V. C. Jammula, J. Wishart, and Y. Yang, "Evaluation of Operational Safety Assessment (OSA) Metrics for Automated Vehicles Using Real-World Data," Tech. Rep., 2022.

[4] N. Kidambi, J. Wishart, M. Elli, and S. Como, "Sensitivity of Automated Vehicle Operational Safety Assessment (OSA) Metrics to Measurement and Parameter Uncertainty," Tech. Rep., 2022.

[5] S. Das, P. Rath, D. Lu, T. Smith, J. Wishart, and H. Yu, "Comparison of Infrastructure-and Onboard Vehicle-Based Sensor Systems in Measuring Safety Metrics," SAE Technical Paper, Tech. Rep., 2023.

[6] D. Lu, V. C. Jammula, S. Como, J. Wishart, Y. Chen, and Y. Yang, "CAROM - Vehicle Localization and Traffic Scene Reconstruction from Monocular Cameras on Road Infrastructures," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 725–11 731.

[7] N. Altekar, S. Como, D. Lu, J. Wishart, D. Bruyere, F. Saleem, and K. L. Head, "Infrastructure-Based Sensor Data Capture Systems for Measurement of Operational Safety Assessment (OSA) Metrics," *SAE Technical Papers*, no. 2021, 2021.

[8] A. Srinivasan, Y. Mahartayasa, V. C. Jammula, D. Lu, S. Como, J. Wishart, Y. Yang, and H. Yu, "Infrastructure-Based LiDAR Monitoring for Assessing Automated Driving Safety," Tech. Rep., 2022.

[9] F. Nex and F. Remondino, "UAV for 3D Mapping Applications: a review," *Applied Geomatics*, vol. 6, no. 1, pp. 1–15, 2014.

[10] E. V. Butilă and R. G. Boboc, "Urban Traffic Monitoring and Analysis Using Unmanned Aerial Vehicles (UAVs): A Systematic Literature Review," *Remote Sensing*, vol. 14, no. 3, p. 620, 2022.

[11] R. Guirado, J.-C. Padró, A. Zoroa, J. Olivert, A. Bukva, and P. Cavestany, "Stratotrans: Unmanned Aerial System (UAS) 4G Communication Framework Applied on the Monitoring of Road Traffic and Linear Infrastructure," *Drones*, vol. 5, no. 1, p. 10, 2021.

[12] H. Gupta and O. P. Verma, "Monitoring and Surveillance of Urban Road Traffic using Low Altitude Drone Images: a Deep Learning Approach," *Multimedia Tools and Applications*, vol. 81, no. 14, pp. 19 683–19 703, 2022.

[13] C. Christodoulou and P. Kolios, "Optimized Tour Planning for Drone-based Urban Traffic Monitoring," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. IEEE, 2020, pp. 1–5.

[14] E. Barmpounakis and N. Geroliminis, "On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment," *Transportation research part C: emerging technologies*, vol. 111, pp. 50–71, 2020.

[15] H. Niu, N. Gonzalez-Prelcic, and R. W. Heath, "A UAV-based Traffic Monitoring System," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE, 2018, pp. 1–5.

[16] J. Lee, Z. Zhong, K. Kim, B. Dimitrijevic, B. Du, and S. Gutesa, "Examining the Applicability of Small Quadcopter Drone for Traffic Surveillance and Roadway Incident Monitoring," in *Transportation Research Board 94th Annual Meeting*, no. 15-4184, 2015, p. 15.

[17] N. A. Khan, N. Jhanjhi, S. N. Brohi, R. S. A. Usmani, and A. Nayyar, "Smart Traffic Monitoring System using Unmanned Aerial Vehicles (UAVs)," *Computer Communications*, vol. 157, pp. 434–443, 2020.

[18] R. Krajewski, L. Vater, M. Klimke, T. Moers, J. Bock, and L. Eckstein, "Drone-based Generation of Sensor Reference and Training Data for Highly Automated Vehicles," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 3067–3074.

[19] F. Outay, H. A. Mengash, and M. Adnan, "Applications of Unmanned Aerial Vehicle (UAV) in Road Safety, Traffic and Highway Infrastructure Management: Recent Advances and Challenges," *Transportation Research Part A: Policy and Practice*, vol. 141, pp. 116–129, 2020.

[20] A. Gupta, T. Afrin, E. Scully, and N. Yodo, "Advances of UAVs toward Future Transportation: The State-of-the-Art, Challenges, and Opportunities," *Future Transportation*, vol. 1, no. 2, pp. 326–350, 2021.

[21] S. Srivastava, S. Narayan, and S. Mittal, "A Survey of Deep Learning Techniques for Vehicle Detection from UAV Images," *Journal of Systems Architecture*, vol. 117, p. 102152, 2021.

[22] R. Krajewski, J. Bock, L. Kloeker, and L. 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)*, 2018, pp. 2118–2125.

[23] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein, "The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1929–1934.

[24] R. Krajewski, T. Moers, J. Bock, L. Vater, and L. Eckstein, "The rounD Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–6.

[25] T. Moers, L. Vater, R. Krajewski, J. Bock, A. Zlocki, and L. Eckstein, "The exiD Dataset: A Real-World Trajectory Dataset of Highly Interactive Highway Scenarios in Germany," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 958–964.

[26] A. Breuer, J.-A. Termöhlen, S. Homoceanu, and T. Fingscheidt, "openDD: A Large-Scale Roundabout Drone Dataset," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–6.

[27] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle *et al.*, "INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps," *arXiv preprint arXiv:1910.03088*, 2019.

[28] O. Zheng, M. Abdel-Aty, L. Yue, A. Abdelraouf, Z. Wang, and N. Mahmoud, "CitySim: A Drone-Based Vehicle Trajectory Dataset for Safety Oriented Research and Digital Twins," *arXiv preprint arXiv:2208.11036*, 2022.

[29] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. 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*, 2020, pp. 2446–2454.

[30] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3D Tracking and Forecasting with Rich Maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8748–8757.

[31] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes *et al.*, "Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[32] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The Apolloscape Dataset for Autonomous Driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 954–960.

[33] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The KITTI Dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[34] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, and J.-N. 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*, 2019, pp. 8797–8806.

[35] A. Krämmer, C. Schöller, D. Gulati, V. Lakshminarasimhan, F. Kurz, D. Rosenbaum, C. Lenz, and A. Knoll, "Providentia - a Large-Scale Sensor System for the Assistance of Autonomous Vehicles and ITS Evaluation," *arXiv preprint arXiv:1906.06789*, 2019.

[36] Z. Zou, R. Zhang, S. Shen, G. Pandey, P. Chakravarty, A. Parchami, and H. X. Liu, "Real-Time Full-Stack Traffic Scene Perception for Autonomous Driving with Roadside Cameras," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 890–896.

[37] F. Poggenhans, J.-H. Pauls, J. Janosovits, S. Orf, M. Naumann, F. Kuhnt, and M. Mayr, "Lanelet2: A High-Definition Map Framework for the Future of Automated Driving," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 1672–1679.

[38] "ASAM OpenDRIVE," https://www.asam.net/standards/detail/opendrive/.

[39] S. Tulsiani and J. Malik, "Viewpoints and Keypoints," in *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1510–1519.
[40] L. Yu, X. Zhi, J. Hu, S. Jiang, W. Zhang, and W. Chen, "Small-Sized Vehicle Detection in Remote Sensing Image Based on Keypoint Detection," *Remote Sensing*, vol. 13, no. 21, p. 4442, 2021.
[41] M. Abdel-Aty, Y. Wu, O. Zheng, and J. Yuan, "Using Closed-Circuit Television Cameras to Analyze Traffic Safety at Intersections based on Vehicle Key Points Detection," *Accident Analysis & Prevention*, vol. 176, p. 106794, 2022.
[42] P. Khorramshahi, A. Kumar, N. Peri, S. S. Rambhatla, J.-C. Chen, and R. Chellappa, "A Dual-Path Model with Adaptive Attention for Vehicle Re-Identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6132–6141.
[43] A. Simoni, A. D'Eusanio, S. Pini, G. Borghi, and R. Vezzani, "Improving Car Model Classification Through Vehicle Keypoint Localization," in *16th International Conference on Computer Vision Theory and Applications*, vol. 5, 2021, pp. 354–361.
[44] W. Yang, Z. Li, C. Wang, and J. Li, "A multi-task Faster R-CNN Method for 3D Vehicle Detection based on a Single Image," *Applied Soft Computing*, vol. 95, p. 106533, 2020.
[45] Z. Hu, Y. Xu, R. S. P. Raj, X. Cheng, L. Sun, and L. Wu, "Vehicle Re-Identification based on Keypoint Segmentation of Original Image," *Applied Intelligence*, pp. 1–17, 2022.
[46] Z. Liu, D. Zhou, F. Lu, J. Fang, and L. Zhang, "AutoShape: Real-Time Shape-Aware Monocular 3D Object Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 641–15 650.
[47] P. Li, H. Zhao, P. Liu, and F. Cao, "RTM3D: Real-Time Monocular 3D Detection from Object Keypoints for Autonomous Driving," in *European Conference on Computer Vision*. Springer, 2020, pp. 644–660.
[48] J. K. Murthy, G. S. Krishna, F. Chhaya, and K. M. Krishna, "Reconstructing Vehicles from a Single Image: Shape Priors for Road Scene Understanding," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 724–731.
[49] S. Kreiss, L. Bertoni, and A. Alahi, "OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association," *IEEE Transactions on Intelligent Transportation Systems*, 2021.
[50] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.
[51] C. Harris, M. Stephens *et al.*, "A Combined Corner and Edge Detector," in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.
[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.
[53] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active Shape Models - Their Training and Application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.
[54] J. J. Moré, "The Levenberg-Marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.
[55] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: a Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[56] K. Bernardin, A. Elbs, and R. Stiefelhagen, "Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment," in *Sixth IEEE International Workshop on Visual Surveillance, in conjunction with ECCV*, vol. 90, no. 91. Citeseer, 2006.
[57] "CAROM - CARs On the Map," https://github.com/duolu/CAROM.
[58] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic Generation and Detection of Highly Reliable Fiducial Markers under Occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.
[59] R. Van Der Horst and J. Hogema, "Time-To-Collision and Collision Avoidance Systems," 1993.
[60] W. Qi, W. Wang, B. Shen, and J. Wu, "A Modified Post Encroachment Time Model of Urban Road Merging Area based on Lane-Change Characteristics," *IEEE Access*, vol. 8, pp. 72 835–72 846, 2020.
[61] "Evaluation of Operational Safety Assessment (OSA) metrics for automated vehicles in simulation, author=Elli, M and Wishart, Jeffrey and Como, Steven and Dhakshinamoorthy, Siddhaarthan and Weast, Jack, journal=SAE Technical Paper, pages=01–0868, year=2021."