

RoadTrack: Realtime Tracking of Road Agents in Dense and Heterogeneous Environments

Rohan Chandra¹, Uttaran Bhattacharya¹, Tanmay Randhavane², Aniket Bera², and Dinesh Manocha¹

¹University of Maryland, ²University of North Carolina

Supplementary Material at <https://gamma.umd.edu/ad/roadtrack>

Abstract—We present a realtime tracking algorithm, RoadTrack, to track heterogeneous road-agents in dense traffic videos. Our approach is designed for dense traffic scenarios that consist of different road-agents such as pedestrians, two-wheelers, cars, buses, etc. sharing the road. We use the tracking-by-detection approach where we track a road-agent by matching the appearance or bounding box region in the current frame with the predicted bounding box region propagated from the previous frame. RoadTrack uses a novel motion model called the Simultaneous Collision Avoidance and Interaction (SimCAI) model to predict the motion of road-agents by modeling collision avoidance and interactions between the road-agents for the next frame. We demonstrate the advantage of RoadTrack on a dataset of dense traffic videos and observe an accuracy of 75.8% on this dataset, outperforming prior state-of-the-art tracking algorithms by at least 5.2%. RoadTrack operates in realtime at approximately 30 fps and is at least 4× faster than prior tracking algorithms on standard tracking datasets.

I. INTRODUCTION

Tracking of road-agents on a highway or an urban road is an important problem in autonomous driving [1], [2] and related areas such as trajectory prediction [3], [4], [5]. These road-agents may correspond to large or small cars, buses, bicycles, rickshaws, pedestrians, moving carts, etc. Different agents have different shapes, move at varying speeds, and their underlying dynamics constraints govern their trajectories. Furthermore, the traffic patterns or behaviors can vary considerably between highway traffic, sparse urban traffic, and dense urban traffic with a variety of such heterogeneous agents, *e.g.*, in Figure 1. The traffic density can be defined based on the number of distinct road-agents captured in a single frame of the video or the number of agents per unit length of the roadway.

Given a traffic video, the tracking problem corresponds to computing the consistency in the temporal and spatial identity of all agents in the video sequence. Recent developments in autonomous driving and large-scale deployment of high-resolution cameras for surveillance has generated interest in the development of accurate tracking algorithms, especially in dense scenarios with a large number of heterogeneous agents. The complexity of tracking increases in dense scenarios as different types of road-agents come in close proximity and interact with each other. Examples of such interactions include passengers boarding or deboarding buses, bicyclists riding alongside cars and so on. Such traffic scenarios arise frequently in densely populated metropolitan cities.

There is extensive prior work on tracking objects and road-agents [6], [7]. But they are mostly designed and used in scenarios with sparse or lower density of road-agents. Such methods are unable to perform tracking in dense traffic due to occlusions and other challenges. In this paper, we mainly focus on developing efficient algorithms for dense traffic scenarios with heterogeneous interactions.

Recently, techniques based on deep learning are widely used for object detection and tracking. In order to solve



Fig. 1: We highlight the performance of our tracking algorithm, RoadTrack, in this urban video. This frame consists of 27 road-agents, including pedestrians, two-wheel scooters, three-wheel rickshaws, cars, and bicycles. RoadTrack can track agents with 75.8% accuracy at approximately 30 fps on an TitanXp GPU. We observe average improvement of at least 5.2% in MOTA accuracy and 4× in fps over prior methods.

the tracking problem in dense and heterogeneous traffic scenarios, we require a motion model that can account for interactions among heterogeneous agents and the high density in which these agents move. We adopt the tracking-by-detection paradigm, which is a two-step process of object detection and state prediction using the motion model. The first step, object detection, is performed to generate vectorized representations, called features, for each road-agent that facilitate identity association across frames. The second step is to predict the state (position and velocity) for the next frame using a motion model.

Main Contributions: We present a realtime tracking algorithm, called RoadTrack, to track heterogeneous road-agents in dense videos. RoadTrack uses a new motion model to represent the motion of different road-agents by simultaneously accounting for collision avoidance and pairwise interactions. We show it is better suited for dense and heterogeneous traffic scenes in comparison to linear constant velocity, non-linear, and learning-based motion models. We name this motion model, SimCAI (“Simultaneous Collision Avoidance and Interaction (SimCAI)”).

RoadTrack makes no assumption regarding camera motion and camera view. For example, we show our algorithm can track road-agents in heavy traffic captured from both front view and top view cameras that can be either stationary or moving. We further do not make assumptions for lighting conditions and can even track road-agents during night-time with glare from oncoming traffic (see supplementary video).

Main Benefits: The advantages of using RoadTrack are summarized below:

- 1) **Accuracy:** On a dense traffic dataset, RoadTrack is state-of-the-art with an absolute accuracy of 75.8%. This is an increase of 5.2% over the next best method. This is equivalent to a rank difference of 42 with the next best method on the current state-of-the-art tracking

benchmark dataset [8].

- 2) **Speed:** Our method demonstrates realtime performance at approximately 30 fps on dense traffic scenes containing up to 100 agents per frame as well as standard tracking datasets. All results were obtained on a TITAN Xp GPU with 8 cores of CPU at 3.6 Ghz frequency. On the MOT benchmark, RoadTrack is at least $4\times$ faster than SOTA methods, and on the dense traffic dataset, it is comparable to the fastest SOTA method.

II. RELATED WORK

A. Pedestrian and Vehicle Tracking

There is extensive work on pedestrian tracking [9], [10]. Bruce et al. [11] and Gong et al. [12] predict pedestrians' motions by estimating their destinations. Liao et al. [13] compute a Voronoi graph from the environment and predict the pedestrian motion along the edges. Mehran et al. [14] apply the social force model to detect anomalous pedestrian behaviors from videos. Pellegrini et al. [15] use an energy function to build a goal-directed short-term collision-avoidance motion model. Bera et al. [16], [17] use reciprocal velocity obstacles and hybrid motion models to improve the accuracy. All these methods are specifically designed for tracking pedestrian movement.

Vehicle tracking has been studied in computer vision, robotics, and intelligent transportation. Some of the earlier techniques are based on using cameras [18] and laser range finders [19]. The authors of [20] model dynamic and geometric properties of the tracked vehicles and estimate their positions using a stereo rig mounted on a mobile platform. Ess et al. [21] present an approach to detect and track vehicles in highly dynamic environments. Multiple cameras have also been used to perform tracking all surrounding vehicles [22], [23]. Moras et al. [24] use an occupancy grid framework to manage different sources of uncertainty for more efficient vehicle tracking; Wojke et al. [25] use LiDAR for moving vehicle detection and tracking in unstructured environments. Finally, [26] uses a feature-based approach to track the vehicles under varying lighting conditions. Most of these methods focus on vehicle tracking and do not take into account interactions with other road-agents such as pedestrians, two-wheelers, rickshaws etc. in dense urban environments. For an up-to-date review of tracking-by-detection algorithms, we refer the reader to methods submitted to the MOT benchmark [8].

B. Motion Models for Tracking

There is substantial work on tracking multiple objects and use of motion models to improve the accuracy [27], [28], [29], [30], [31], [32]. Kim et al. [28] perform Multiple Hypotheses Tracking (MHT) [33] by using an effective online classifier for efficient branch pruning. The constant velocity linear motion model has been used to join fragmented pedestrian tracks caused by occlusion [30]. However, dense traffic often cause road-agents to perform complex maneuvers to avoid collisions that are often non-linear. Hence, linear motion models do not work well in dense scenes.

RVO [34] is a non-linear motion model that has been used for pedestrian tracking in dense crowd videos. However, RVO does not take into account agents interacting with one another. An extension to RVO, called AutoRVO [2] includes dynamic constraints between road-agents. However, AutoRVO is based on CTMAT [35] representations of road-agents that cannot be translated to front-view scenes. Other

non-linear motion models that have been used for tracking include social forces [36], LTA [15], and ATTR [37]. However, these are mainly designed for tracking pedestrians. Social Forces, in particular, holds resemblance to our proposed motion model, SimCAI, in that it models the attraction and repulsion between agents (pedestrians only) through the concept of potential energy functions. With the recent rise in popularity of deep learning, recurrent neural networks such as LSTMs have been used as motion models for tracking [38], [39]. We compare SimCAI with both learning- and non-learning-based motion models in this paper.

III. ROADTRACK: OVERVIEW

In this section, we present the RoadTrack algorithm that combines Mask R-CNN object segmentation with SimCAI. Informally, the tracking problem is stated as follows: Given a video, we want to assign an ID to all road-agents in all frames. This is formally equivalent to solving the following sub-problem at each time-step (or frame): At current time t , given the ID labels of all road-agents in the frame, assign labels for road-agents in the next frame (time $t + 1$).

We start by using Mask R-CNN to implicitly perform pixel-wise segmentation of the road-agents. This generates a set of segmented boxes [31]. For each detected road-agent, h_j , generated using Mask R-CNN, we extract their corresponding features, f_{h_j} , using the deep learning-based feature extraction architecture proposed in [31]. We do not use the provided pre-trained models and instead, fine-tune the existing feature extraction network on traffic datasets to learn meaningful features pertaining to traffic. We discuss the fine-tuned hyperparameters in the supplementary material.

Next, we predict the next state (state consists of spatial coordinates (p_i) and velocities (v_i)) for each road-agent for the next time-step using SimCAI. This step is the main contribution of this work and is described in detail in Section IV. This step results in another set of segmented boxes for each road-agent at time $t + 1$.

Finally, we use these sets of segmented boxes to compute features using a Convolutional Neural Network [40]. The features generated are compared using association algorithms [41] to compute the ID of each agent in the next frame. The features are matched in two ways: the Cosine metric and the IoU overlap [42]. The Cosine metric is computed using the following optimization problem:

$$\min_{h_j} (l(f_{p_i}, f_{h_j}) | p_i \in \mathcal{P}, h_j \in \mathcal{H}_i). \quad (1)$$

where \mathcal{H}_i is the subset of all detected road-agents in the current frame that are within a circular region around agent p_i that have not been matched to a predicted agent. The IoU overlap metric is used in conjunction with the cosine metric. This metric builds a cost matrix Σ to measure the amount of overlap of each predicted bounding box with all nearby detection bounding box candidates. $\Sigma(i, j)$ stores the IoU overlap of the bounding box of p_i with that of h_j and is calculated as:

$$\Sigma(i, j) = \frac{\mathbb{B}_{p_i} \cap \mathbb{B}_{h_j}}{\mathbb{B}_{p_i} \cup \mathbb{B}_{h_j}}, h_j \in \mathcal{H}_i.$$

If we denote the cosine and the IOU overlap metrics by C and I , respectively, then the combined cost function value is obtained through,

$$\text{Combined Cost} = \lambda_1 C + \lambda_2 I, \lambda_1 + \lambda_2 = 1, \quad (2)$$

where λ_1, λ_2 are constants representing the weights for the individual metric costs. Matching a detection to a predicted measurement with maximum overlap thus becomes a max-weight matching problem and we solve it efficiently using the Hungarian algorithm [41]. The ID of the road-agent at time t is assigned to that road-agent at time $t + 1$ whose appearance is most closely associated to the road-agent at time t .

IV. SIMCAI: SIMULTANEOUS COLLISION AVOIDANCE AND INTERACTIONS

One of the major challenges with tracking heterogeneous road-agents in dense traffic is that road-agents such as cars, buses, bicycles, road-agents, etc. have different sizes, geometric shape, maneuverability, behavior, and dynamics. This often leads to complex inter-agent interactions that have not been taken into account by prior multi-object trackers. Furthermore, road-agents in high-density scenarios are in close-proximity to one another or are almost colliding. So we need an efficient approach for predicting the next state of a road-agent by modeling the collisions and interactions. We thus present SimCAI, that takes into account both,

- Reciprocal collision avoidance [34] with car-like kinematic constraints for trajectory prediction and collision avoidance.
- Heterogeneous road-agent interaction between pedestrians, two-wheelers, rickshaws, buses, cars and so on.

All the notations used in the paper are provided in Table I of full version of this text [43].

A. Velocity Prediction by Modeling Collision Avoidance

Reciprocal Velocity Obstacles (RVO) [34] extends Velocity Obstacles motion model by modeling collision avoidance behavior for multiple engaging agents. RVO can be applied to pedestrians in a crowd and we modify it to work with bounding boxes as our algorithm conforms to the tracking-by-detection paradigm.

We represent each agent as, $\Psi_t = [u, v, \dot{u}, \dot{v}, v_{\text{pref}}]$, where u, v, \dot{u}, \dot{v} , and v_{pref} represent the top left corner of the bounding box, their velocities, and the preferred velocity of the agent in the absence of obstacles respectively. v_{pref} is computed internally by RVO.

The computation of the new state, Ψ_{t+1} , is expressed as an optimization problem. For each agent, RVO computes a feasible region where it can move without collision. This region is defined according to the RVO collision avoidance constraints (or ORCA constraints [34]). If the ORCA constraints forbid an agent's preferred velocity, that agent chooses the velocity closest to its preferred velocity that lies in the feasible region, as given by the following optimization:

$$v_{\text{new}} = \arg \min_{v \notin \text{ORCA}} \|v - v_{\text{pref}}\| \quad (3)$$

The velocity, v_{new} , is then used to calculate the new position of a road-agent.

The difference in shapes, sizes, and aspect ratios of road-agents motivate the need to use appearance-based features. In order to combine object detection with RVO, we modify the state vector, Ψ_t , to include bounding box information by setting the position to the centers of the bounding boxes. Thus, $u = \frac{u+w}{2}$ and $v = \frac{v+h}{2}$, where w, h denote the width and height, respectively, of the corresponding bounding box.

Finally, the original RVO models the motion of agents seen from a top-view. Therefore, to account for front-view

traffic as well as top-view, we use the modification proposed by the authors of [31] that allow RVO to model the motion of road-agents in front-view traffic scenes.

B. Velocity Prediction by Modeling Road-Agent Interactions

In a traffic scenario, interactions can occur between different types of road-agents: vehicle-vehicle, pedestrian-pedestrian, vehicle-pedestrian, bicycle-pedestrian, etc. In this section, we present a formulation to model such interactions. Our input is an RGB video captured from a camera with known camera parameters. By using the camera center as the origin, we transform pixel coordinates to scene coordinates for the computations that follow in this section.

1) **Intent of Interaction:** The idea of using spatial regions to characterize agent behavior was proposed in [44]. The authors introduced the notion of “public” and “social” regions, that are of the form of concentric circles. We show a quadrant of these regions in Figure 2, where the yellow area is the social region and the orange area is the public region. Based on this work, Satake et al. [45] proposed a model of approach behavior with which a robot can interact with humans. At the public distance the robot is allowed to approach the human to interact with them, and at the social distance, interaction occurs. In SimCAI, we have set the public and social distances heuristically.

We say that a road-agent, p_i , intends to interact with another agent, p_k , when p_i is within the social distance of p_k for some minimum time τ . When two road-agents intend to interact, they move towards each other and come in close proximity.

2) **Ability to Interact:** Even when two road-agents want to interact, their movements could be restricted in dense traffic. We determine the ability to interact (Figure 2(right)) as follows.

Each agent has a personal space, which we define as a circular region ζ of radius ρ , centered around p_k . Given a road-agent p_i , the slope of its v_{pref} is $\tan \theta$. θ is the angle with the horizontal defined in the world coordinate system. In dense traffic, each agent, p_i has a limited space in which they can steer, or turn. This space is the feasible region determined by the ORCA constraints described in the previous section. We define a 2D cone, γ , of angle ϕ as the ORCA region in which the agent can steer. ϕ is thus the steering angle of the agent. We denote the extreme rays of the cone as r_1 and r_2 . $\perp_{\mathbb{G}_2}$ denotes the smallest perpendicular distance between any two geometric structures, say, \mathbb{G}_1 and \mathbb{G}_2 . These parameters are fixed for different agent types and are not learned from data.

If p_i has intended to interact with p_k , the projected cone of p_i , defined by extending r_1 and r_2 , is directed towards p_k . Then, in order for interaction to take place, it is sufficient to check for either one of two conditions to be true:

- 1) Condition Ω_1 : Intersection of ζ with either r_1 or r_2 (if either ray intersects, then the entire cone intersects ζ).
- 2) Condition Ω_2 : $\zeta \subset \gamma$ (if ζ lies in the interior of the cone, see Figure 2).

For these conditions to hold, we require that the cone does not intersect or contain any $p_j \in \mathcal{P}, j \neq i$. We now make these equations more explicit.

We parametrize r_1, r_2 by their slopes $\tan \delta$, where $\delta = \theta_i + \phi_i$ if $\perp_{\zeta}^{r_1} \geq \perp_{\zeta}^{r_2}$, else $\delta = \theta_i - \phi_i$. The resulting equation of r_1 (or r_2) is $(Y - v_i) = \tan \delta (X - u_i)$ and the equation of ζ is $(X - u_k)^2 + (Y - v_k)^2 = \rho^2$. Solving both equations

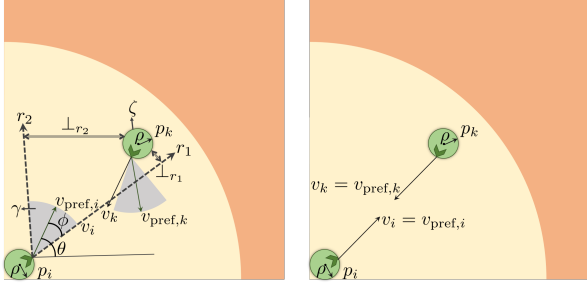


Fig. 2: Inner yellow circle denotes the social distance and the outer orange area denotes the public region. At time $t \geq \tau$ and using IV-B.1, p_i intends to interact with p_k . Then using IV-B.2 (left), p_i determines its ability to interact with p_k . We observe that γ (grey cone) of p_i contains ζ of p_k (green circle around p_k). Thus p_i can interact with p_k . Using IV-B.3 (right), p_i and p_k align their preferred velocities toward each other.

simultaneously, we obtain an equation Ω_1 . Intersection occurs if the discriminant of $\Omega_1 \geq 0$. This provides us with the first condition necessary for the occurrence of an interaction between p_i and p_k .

Next, we observe that if ζ lies in the interior of γ , then p_k lies on the opposite sides of r_1 and r_2 which is modeled by the following equation:

$$\Omega_2 \equiv r_1(p_k) \cdot r_2(p_k) \leq 0 \quad (4)$$

Solving Equation 4 further provides us with the second condition for the occurrence of an interaction between p_i and p_k , where $\Omega_1, \Omega_2 : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$.

3) **Interaction:** If either Ω_1 or Ω_2 is true, then road-agents p_i, p_k will move towards each other to interact at time $t \geq \tau$. When this happens, we assume that p_i and p_k align their current velocities towards each other. Thus, $v_{\text{new}} = v_{\text{pref}}$. The time taken for the two road-agents to be meet or converge with each other is given by $t = \frac{\|p_i - p_k\|_2}{\|v_i - v_k\|_2}$. If two road-agents are overlapping (based on the values of Ω_1 and Ω_2), we model them as a new agent with radius 2ϵ .

Our approach can be extended to model multiple interactions. Currently, we restrict an interaction to take place between 2 road-agents. Therefore, in the case of multiple possible interactions with an agent, p_k , we form a set $\mathcal{Q} \subseteq \mathcal{P}$, where \mathcal{Q} is the set of all road-agents p_ω , that are intending to interact with p_k . We determine the road-agent that will interact with p_k as the road-agent that minimizes the distance between p_k and p_ω after a fixed time-step, Δt . Thus, $p_\omega = \arg \min_w \|(p_\omega + v_\omega \Delta t) - p_k\|, p_\omega \in \mathcal{Q}$. road-agents that are not interacting avoid each other and continue moving towards their destination.

C. Analysis

We analyze the accuracy and runtime performance of SimCAI in traffic scenarios with increasing density and heterogeneity.

Accuracy Analysis: We analytically show the advantage of SimCAI over other motion models such as Social Forces [36], RVO [34], [46], and constant velocity [40].

We denote the multiple object tracking accuracy, $MOTA$ of a system using a particular motion model as $MOTA^{\text{model}}$ and define it as $MOTA^{\text{model}} = \sum_c MOTA_c + \sum_i MOTA_i$ where c and i denote an agent whose motion is being modeled using collision avoidance and interaction, and $MOTA_c$ and $MOTA_i$ denote their individual accuracies, respectively.



Fig. 3: Qualitative analysis of RoadTrack on the TRAF dataset at night time consisting of cars, 2-wheelers, 3-wheelers, and trucks. Frames are chosen with a gap of 2 seconds (~ 60 frames). For visual clarity, each road-agent is associated with a unique ID number. The ID is displayed in orange. Note the consistencies in the ID, for example, the 3-wheeler (1), car (2), and 2-wheeler (3).

Let n represent the number of total road-agents in a video, then we have $n = n_c + n_i$, where n_c, n_i correspond to the number of agents that are avoiding collisions and are interacting, respectively.

Increasing n would increase the number of road-agents whose motion is modeled through collision avoidance or heterogeneous interaction formulations. Linear models do not account for either formulation. Standard RVO only accounts for collision avoidance. SimCAI models both. Therefore, we rationalize that,

$$\begin{aligned} MOTA_c^{\text{linear}} &\leq MOTA_c^{\text{RVO}} \approx MOTA_c^{\text{SimCAI}} \\ MOTA_i^{\text{linear}} &\leq MOTA_i^{\text{RVO}} \leq MOTA_i^{\text{SimCAI}} \\ \Rightarrow MOTA^{\text{linear}} &\leq MOTA^{\text{RVO}} \leq MOTA^{\text{SimCAI}} \end{aligned}$$

We validate the analysis presented here in Section V-D.

Runtime Analysis: At approximately 30 fps, we achieve a minimum speed-up of approximately $4\times$, and upto approximately $30\times$, over state-of-the-art methods on the MOT dataset (Table II). The selection of state-of-the-art methods is done in Section V-B. The state-of-the-art use RNNs to model the motion of road-agents [38], [39], while we use the modified RVO formulation. We exploit the geometrical formulation of SimCAI to state and prove the following theorem:

Theorem IV.1. Given $\mathcal{P} = \{p_i | 1 \leq i \leq n\}$, that represents a set of n road-agents in a traffic scene that may assume any shape, size, and agent-type, if state $p_i \in \{\text{stationary, collision avoiding, interacting}\}$, $\forall i \in n$, then SimCAI can track the n road-agents in $\mathcal{O}(n_c + \omega n_i)$, where $\omega \ll n_i$.

Proof. RVO is based on linear programming and can perform tracking with a proven runtime complexity of $\mathcal{O}(n)$ [34]. Now, if we assume that agents always assume one of the following states: stationary, avoiding collision, or interacting, then we have $n = n_c + n_i$, where n_c, n_i correspond to the number of agents in collision avoidance states and interacting states, respectively. We ignore stationary road-agents. Following the formulation in Section IV-B, for each interacting road-agent, SimCAI predicts a new velocity by solving a linear optimization problem over ω road-agents. Thus, the runtime complexity of SimCAI is $\mathcal{O}(n_c + \omega n_i)$, where $\omega \ll n_i$. ■

Our high fps is a consequence of our linear runtime complexity and we validate our theoretical claims in Section V. We further hypothesize that prior deep learning-based methods [38], [39] are less optimal in terms of runtime due to the intensive computation requirements by deep neural networks [47], [48]. For example, ResNet [49] needs more than 25 MB for storing the computed model in memory, and more than 4 billion float point operations (FLOPs) to process a single image of size 224×224 [47].

We would like to clarify that by realtime performance, we refer to the realtime computation of the tracking algorithm

only. We do not consider the computation time of Mask R-CNN. This is standard practice by tracking-by-detection algorithms [39] that only contribute to the tracking component, similar to this work. We therefore compare with realtime tracking algorithms.

V. EXPERIMENTS

A. Datasets

We highlight the performance of RoadTrack through extensive experiments on different traffic datasets.

(Dense) TRAF Dataset: We use the TRAF traffic dataset [3] that consists of a set of 60 video sequences that contain dense traffic with highly heterogeneous agents with front and top-down viewpoints, stationary and moving camera motions, and during both day and night. These videos are of highway and urban traffic in high population countries like China and India. Most importantly, ground truth annotations consisting of 2-D bounding box coordinates and agent types are provided with the dataset. The key aspects of this dataset are the high density and the heterogeneity.

(Sparse) MOT & KITTI-16 Datasets: There are now several popular open-source tracking benchmarks available on which researchers can test and compare the performance of tracking algorithms. The current state-of-the-art benchmark is the MOT benchmark [8], which contains a mix of pedestrians and traffic sequences. However, the MOT benchmark is a general tracking benchmark dataset. Therefore, we additionally conduct experiments exclusively on the KITTI-16 traffic sequence [50]. It should be noted that the KITTI-16 sequence is sparse, consisting of mostly cars, and does not contain road-agent interactions.

B. Evaluation Methods

Due to the open-source nature of the MOT benchmark, there are a large number of methods available in the MOT benchmark (80 and 87 on 2D MOT15 and MOT16, respectively). To demonstrate the superiority of RoadTrack, it is therefore sufficient to select state-of-the-art methods from all the methods, and compare RoadTrack against this set of methods. We define a state-of-the-art method as one that satisfies all of the following criteria simultaneously:

- 1) Higher Average Rank: The MOT benchmark assigns an “average rank” to each method. The average rank ([51], page 390) of a tracking algorithm is computed by averaging over all the metrics. This metric effectively ranks the overall performance of a tracking algorithm by taking all the metrics into account simultaneously. We select competitor methods that have a *higher (better)* average rank than ours.
- 2) Published Work: Many of the tracking methods submitted on the MOT benchmark are anonymous. We therefore select methods that are published in peer-reviewed conferences and journals.
- 3) Online Tracking: RoadTrack performs tracking using only the information from the previous frame and assumes no knowledge of future frames, thus making it an online tracking method. Therefore, we compare RoadTrack with top-performing online methods for fair comparison.
- 4) Realtime Performance: RoadTrack has realtime performance and performs tracking at up to approximately 30 fps (see Tables II,III). Note that by realtime performance, we refer to the realtime computation of

the tracking algorithm only. We do not consider the computation time of Mask R-CNN. This is conventionally accepted by tracking-by-detection methods that optimize only the tracking component. We therefore compare with algorithms that also compute tracking in realtime.

The methods that satisfy these criteria are listed in Tables II and III. For evaluation on the TRAF dataset, however, one additional criterion is required: the availability of open-sourced code. There are only two methods (Table I) that satisfy all of the above criteria.

We point out that in selecting methods to compare with for each dataset according to the above criteria, all tables need not have the same selection of methods. For example, the methods in Tables II,III do not have open-sourced code.

C. Evaluation Metrics

We use standard tracking metrics defined in [52]. We compare the overall accuracy (MOTA) which is computed using the formula: $MOTA = \frac{1 - (FN + FP + IDS)}{GT}$ where FN, FP, IDS, and GT correspond to the number of false negatives, false positives, ID switches, and ground truth agents, respectively. Additionally, we report the number of mostly tracked (MT) and mostly lost (ML) agents as well as the precision of the detector (MOTP), as per their provided definitions in [52]. In accordance with the strict annotation protocol adopted by the MOT benchmark, we do not count stationary agents such as parked vehicles in our formulation. Detected objects such as traffic signals are thus considered false positives.

Dataset	Tracker	FPS \uparrow	MT(%) \uparrow	ML(%) \downarrow	IDS \downarrow	FN \downarrow	MOTP(%) \uparrow	MOTA(%) \uparrow
TRAF1	MOTDT	37.9	0	98.2	15 (<0.1%)	18,764 (33.0%)	63.3	67.0
	MDP	9.3	0	98.2	21 (<0.1%)	18,667 (32.8%)	60.1	67.1
	RoadTrack	43.9	0	95.6	163 (0.3%)	17,953 (31.6%)	58.8	68.1
TRAF2	MOTDT	41.6	0	98.8	17 (<0.1%)	18,201 (32.7%)	60.3	67.3
	MDP	20.9	0	100.0	7 (<0.1%)	18,105 (32.5%)	59.6	67.5
	RoadTrack	12.3	0	92.3	55 (0.1%)	17,202 (30.9%)	60.8	69.0
TRAF3	MOTDT	50.7	3.3	67.1	64 (<0.1%)	34,883 (27.0%)	69.6	72.9
	MDP	51.8	0	100.0	0 (0.0%)	43,057 (33.3%)	69.2	66.7
	RoadTrack	36.6	32.2	40.0	62 (<0.1%)	19,521 (15.1%)	70.1	84.8
TRAF4	MOTDT	36.6	1.2	76.3	123 (0.1%)	54,349 (29.0%)	65.3	70.9
	MDP	9.0	1.2	87.2	16 (<0.1%)	59,097 (31.3%)	66.2	68.7
	RoadTrack	40.6	6.0	54.6	266 (0.1%)	47,444 (25.1%)	65.1	74.7
TRAF5	MOTDT	36.0	0.7	75.9	221 (0.2%)	33,774 (28.9%)	63.2	70.9
	MDP	22.5	0	98.4	6 (<0.1%)	38,091 (32.6%)	64.9	67.3
	RoadTrack	41.4	1.5	55.7	299 (0.3%)	24,860 (21.3%)	63.1	78.4
TRAF6	MOTDT	33.0	0	87.5	161 (0.1%)	58,212 (29.4%)	63.3	70.5
	MDP	4.3	0	99.3	0 (0.0%)	65,687 (33.2%)	68.6	66.8
	RoadTrack	14.6	0.7	67.8	283 (0.1%)	52,017 (26.3%)	62.8	73.6
Summary	MOTDT	34.7	0.9	83.6	601 (0.1%)	218,683 (29.3%)	65.5	70.6
	MDP	10.1	0.2	97.0	50 (<0.1%)	242,704 (32.6%)	65.3	67.4
	RoadTrack	31.6	7.0	66.9	1128 (0.2%)	178,997 (24.0%)	65.7	75.8

TABLE I: Evaluation on the TRAF dataset with MOTDT [53] and MDP [54]. MOTDT is currently the best *online* tracker on the MOT benchmark with open-sourced code. Bold is best. Arrows (\uparrow, \downarrow) indicate the direction of better performance. **Observation:** RoadTrack improves the accuracy (MOTA) over the state-of-the-art by 5.2% and precision (MOTP) by 0.2%.

D. Analysis of Results & Discussion

On Dense Datasets: We provide results on the TRAF dataset using RoadTrack and demonstrate a state-of-the-art average MOTA of 75.8% (Table I). The aim of this experiment is to highlight the advantage of our overall tracking algorithm in dense and heterogeneous traffic. We compare RoadTrack with methods (selected according to criteria established in Section V-B) on the dense TRAF dataset in Table I. MOTDT [53] and MDP [54] are the only state-of-the-art methods with available open-source code. All methods are evaluated using a common set of detections obtained using Mask R-CNN. Compared to these methods, we improve upon MOTA by 5.2% on absolute. This is roughly equivalent to a rank difference of 46 on the MOT benchmark.

Tracker	FPS \uparrow	MT(%) \uparrow	ML(%) \downarrow	IDS \downarrow	FN \downarrow	MOTP(%) \uparrow	MOTA(%) \uparrow
KITTI-16	AP_HWDPL_p [55]	6.7	17.6	11.8	18	831	72.6
	RAR_15_pub [39]	5.4	0.0	17.6	18	809	41.2
	AMIR15 [38]	1.9	11.8	11.8	18	714	50.4
	HybridDAT [56]	4.6	5.9	17.6	10	706	46.3
	AM [57]	0.5	5.9	17.6	19	805	40.6
	RoadTrack	28.9	29.4	11.7	15	668	12.2

TABLE II: Evaluation on the KITTI-16 dataset from the MOT benchmark with *online methods* that have an average rank higher than ours. RoadTrack is at least approximately $4\times$ faster than prior methods. While we do not outperform on the MOTA metric, we still achieve the highest MT, ML, FN, and MOTP. We analyze our MOTA performance in Section V-D. Bold is best. Arrows (\uparrow, \downarrow) indicate the direction of better performance. The values for all methods correspond to the KITTI-16 sequence specifically, and not the entire 2D MOT15 dataset.

Tracker	FPS \uparrow	MT(%) \uparrow	ML(%) \downarrow	IDS \downarrow	FN \downarrow	MOTP(%) \uparrow	MOTA(%) \uparrow
2D MOT15	AMIR15 [38]	1.9	15.8	26.8	1026	29,397	37.6
	HybridDAT [56]	4.6	11.4	42.2	358	31,140	35.0
	AM [57]	0.5	11.4	43.4	348	34,848	34.3
	AP_HWDPL_p [55]	6.7	8.7	37.4	586	33,203	38.5
	RoadTrack	28.9	18.6	32.7	429	27,499	20.0
MOT16	EAMTT_pub [58]	11.8	7.9	49.1	965	102,452	38.8
	RAR16pub [39]	0.9	13.2	41.9	648	91,173	45.9
	STAM16 [57]	0.2	14.6	43.6	473	91,117	46.0
	MOTDT [53]	20.6	15.2	38.3	792	85,431	47.6
	AMIR [38]	1.0	14.0	41.6	774	92,856	47.2
	RoadTrack	18.8	20.3	36.1	722	78,413	40.9

TABLE III: Evaluation on the full MOT benchmark. The full MOT dataset is sparse and is not a traffic-based dataset. RoadTrack is at least approximately $4\times$ faster than previous methods. While we do not outperform on the MOTA metric, we still achieve the highest MT, ML (MOT16), FN, and MOTP(MOT15). We analyze our MOTA performance in Section V-D. Bold is best. Arrows (\uparrow, \downarrow) indicate the direction of better performance.

Motion Model	FPS \uparrow	MT(%) \uparrow	ML(%) \downarrow	IDS \downarrow	FN \downarrow	MOTP(%) \uparrow	MOTA(%) \uparrow
Const. Vel	30	0.0	100	11	247,738(33.3%)	66.3	66.7
SF	30	0.1	98.6	147	246,528 (33.1%)	63.8	66.3
RVO	30	0.0	100	38	247,675 (33.2%)	63.8	66.9
SimCAI	30	7.0	66.9	1128	178,997 (24.0%)	65.7	75.8

TABLE IV: Ablation experiments to show the advantage of SimCAI. We replace SimCAI with a constant velocity (Const Lin Vel) [40], Social Forces (SF) [36], and RVO motion model (RVO)[34]. The rest of the method is identical to the original method. All variations operate at similar fps of approximately 30 fps. Bold is best. Arrows (\uparrow, \downarrow) indicate the direction of better performance.

MOTDT is currently the fastest method (according to the selection criteria of Section V-B) on the MOT16 benchmark. Our approach operates at realtime speeds upto approximately 30 fps and is comparable with MOTDT (Table I). Our realtime performance results from the runtime analysis from Section IV-C and theorem IV.1.

Note that we observe an abnormally high number of identity switches compared to other methods; however, this is because prior methods mostly fail to maintain an agent’s track for more than 20% of their total visible time (near 100% ML). Not being able to track road-agents for most of the time excludes those agents as possible candidates for IDS, thereby resulting in lower IDS for prior methods. Interestingly, the low IDS score for prior methods also contributes to their reasonably high MOTA score, despite near-failure to track agents in dense traffic.

On Standard Benchmarks: In the interest of completeness and thorough evaluation, we also evaluate RoadTrack on sparser tracking datasets and present results on both traffic-only datasets (KITTI-16) in Table II as well as datasets containing only pedestrians (MOT) in Table III. RoadTrack’s main advantage is SimCAI, which is based on modeling collision avoidance and interactions. In the absence of one or both, we do not expect it demonstrate superior performance over prior methods on the sparse KITTI-16 and MOT datasets. While not conclusive, we believe our low MOTA score on the 2D MOT15 and KITTI-16 may also be

attributed to a high number of detections that are incorrectly classified as false positives. For instance, road-agents that are too distant to be manually labeled are not annotated in the ground truth sequence. We observed this to be true for the methods we compared with as well. Therefore, we exclude FP from the calculation of MOTA for all methods in the interest of fair evaluation.

We note, however, that RoadTrack is least $4\times$ faster on the KITTI-16 and 2D MOT15 datasets at approximately 30 fps (Tables II,III). To explain the speed-up, we refer to theorem IV.1 and the runtime analysis presented in Section IV-C. We specially point to the $15\times$ and $5\times$ speed-up over learning-based tracking methods, [38], [39] in Table II which we attribute the linear time computation of SimCAI as opposed to the intensive computation required by deep learning models.

Ablation Experiments: We highlight the advantages of SimCAI through ablation experiments in Table IV. The aim of these experiments is to isolate the benefit of SimCAI. We compare with the following variations of RoadTrack in which we replace our novel motion model SimCAI with standard and state-of-the-art motion models, while keeping the rest of the system untouched:

- Constant Linear Velocity (Const Lin Vel). We replace SimCAI with a constant velocity linear motion model [40].
- Social Forces (SF). We replace SimCAI with the Social Forces motion model [36].
- Reciprocal Velocity Obstacles (RVO) [34]. We replace SimCAI with the RVO motion model.

We compare SimCAI with other motion models (Constant linear velocity, Social Forces, and RVO) on the dense TRAF dataset. These experiments were performed by *only* replacing SimCAI with other motion models, keeping the rest of the system unchanged. We observe that SimCAI outperforms the motion models by at least 8.9% on absolute on MOTA. All the variations used in the ablation experiments operated at the same fps of approximately 30 fps. Additionally, we experimentally verify the analysis of Section IV-C by observing that $MOTA^{\text{linear}} \leq MOTA^{\text{RVO}} \leq MOTA^{\text{SimCAI}}$. Once again, we point to our high IDS in Table IV, compared to the IDS of other motion models. As mentioned previously, this is due to the near-failure of other motion models (near 100% ML) to track road agents in dense traffic. Not being able to track a road-agent excludes them as a IDS candidate.

VI. LIMITATIONS AND FUTURE WORK

There are many avenues of future work for our presented work. Currently, many parameters in our algorithm such as the radii for the social and public regions, steering angles, and cone angles, are heuristically chosen for optimum performance. It would be more efficient to learn these parameters instead, using data driven and machine learning techniques. Furthermore, the results from tracking road-agents can be directly used to further research in related areas such as trajectory prediction. With the increased popularity of deep-learning and improved tracking methods, deep learning techniques can be employed for predicting the future motion of road-agents in dense and heterogeneous traffic.

VII. ACKNOWLEDGEMENTS

This work was supported in part by ARO Grants W911NF1910069 and W911NF1910315, Semiconductor Research Corporation (SRC), and Intel.

REFERENCES

- [1] Alex Teichman and Sebastian Thrun. Practical object recognition in autonomous driving and beyond. In *Advanced Robotics and its Social Impacts*, pages 35–38. IEEE, 2011.
- [2] Yuexin Ma, Dinesh Manocha, and Wenping Wang. Autorvo: Local navigation with dynamic constraints in dense heterogeneous traffic. *arXiv preprint arXiv:1804.02915*, 2018.
- [3] Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Traffic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [4] Rohan Chandra, Uttaran Bhattacharya, Christian Roncal, Aniket Bera, and Dinesh Manocha. Robusttp: End-to-end trajectory prediction for heterogeneous road-agents in dense traffic with noisy sensor inputs. *arXiv preprint arXiv:1907.08752*, 2019.
- [5] Rohan Chandra, Tianrui Guan, Srujan Panuganti, Trisha Mittal, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Forecasting trajectory and behavior of road-agents using spectral clustering in graph-lstms. *arXiv preprint arXiv:1912.01118*, 2019.
- [6] Sayanan Sivaraman and Mohan Manubhai Trivedi. Integrated lane and vehicle detection, localization, and tracking: A synergistic approach. *IEEE Transactions on Intelligent Transportation Systems*, 14(2):906–917, 2013.
- [7] Ricardo Guerrero-Gómez-Olmedo, Roberto J López-Sastre, Saturnino Maldonado-Bascón, and Antonio Fernández-Caballero. Vehicle tracking by simultaneous detection and viewpoint estimation. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 306–316. Springer, 2013.
- [8] 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.
- [9] Jinshi Cui, Hongbin Zha, Huijing Zhao, and Ryosuke Shibasaki. Tracking multiple people using laser and vision. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2116–2121. IEEE, 2005.
- [10] Louis Kratz and Ko Nishino. Tracking pedestrians using local spatiotemporal motion patterns in extremely crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (99):11, 2011.
- [11] Allison Bruce and Geoffrey Gordon. Better motion prediction for people-tracking. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, 2004.
- [12] Haifeng Gong, Jack Sim, Maxim Likhachev, and Jianbo Shi. Multi-hypothesis motion planning for visual object tracking. In *2011 International Conference on Computer Vision*, pages 619–626. IEEE, 2011.
- [13] Lin Liao, Dieter Fox, Jeffrey Hightower, Henry Kautz, and Dirk Schulz. Voronoi tracking: Location estimation using sparse and noisy sensor data. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2003.
- [14] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 935–942, 2009.
- [15] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You’ll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.
- [16] Aniket Bera and Dinesh Manocha. REACH: Realtime crowd tracking using a hybrid motion model. *Proc. of the International Conference on Robotics and Automation (ICRA)*, 2015.
- [17] Aniket Bera, Sujeong Kim, Tanmay Randhavane, Srihari Pratapa, and Dinesh Manocha. Glmp-realtime pedestrian path prediction using global and local movement patterns. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5528–5535. IEEE, 2016.
- [18] Frank Dellaert and Chuck Thorpe. Robust car tracking using kalman filtering and bayesian templates. In *Conference on intelligent transportation systems*, volume 1, 1997.
- [19] Daniel Streller, K Furstenberg, and Klaus Dietmayer. Vehicle and object models for robust tracking in traffic scenes using laser range images. In *Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on*, pages 118–123. IEEE, 2002.
- [20] Anna Petrovskaya and Sebastian Thrun. Model based vehicle detection and tracking for autonomous urban driving. *Autonomous Robots*, 26(2-3):123–139, 2009.
- [21] Andreas Ess, Konrad Schindler, Bastian Leibe, and Luc Van Gool. Object detection and tracking for autonomous navigation in dynamic environments. *The International Journal of Robotics Research*, 29(14):1707–1725, 2010.
- [22] Akshay Rangesh and Mohan M Trivedi. No blind spots: Full-surround multi-object tracking for autonomous vehicles using cameras & lidars. *arXiv preprint arXiv:1802.08755*, 2018.
- [23] Michael Darms, Paul Rybski, and Chris Urmson. Classification and tracking of dynamic objects with multiple sensors for autonomous driving in urban environments. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 1197–1202. IEEE, 2008.
- [24] Julien Moras, Véronique Cherfaoui, and Philippe Bonnifait. Credibilist occupancy grids for vehicle perception in dynamic environments. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 84–89. IEEE, 2011.
- [25] Nicolai Wojke and Marcel Häselich. Moving vehicle detection and tracking in unstructured environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3082–3087. IEEE, 2012.
- [26] Benjamin Coifman, David Beymer, Philip McLauchlan, and Jitendra Malik. A real-time computer vision system for vehicle tracking and traffic surveillance. *Transportation Research Part C: Emerging Technologies*, 6(4):271–288, 1998.
- [27] Matthias Luber, Johannes A Stork, Gian Diego Tipaldi, and Kai O Arras. People tracking with human motion predictions from social forces. In *2010 IEEE International Conference on Robotics and Automation*, pages 464–469. IEEE, 2010.
- [28] Chanh Kim, Fuxin Li, Arridhana Ciptadi, and James M Rehg. Multiple hypothesis tracking revisited. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4696–4704, 2015.
- [29] Jiahui Chen, Hao Sheng, Yang Zhang, and Zhang Xiong. Enhancing detection model for multiple hypothesis tracking. In *Conf. on Computer Vision and Pattern Recognition Workshops*, pages 2143–2152, 2017.
- [30] Hao Sheng, Li Hao, Jiahui Chen, Yang Zhang, and Wei Ke. Robust real effective matching model for multi-target tracking. In *Pacific Rim Conference on Multimedia*, pages 233–243. Springer, 2017.
- [31] Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Densepeds: Pedestrian tracking in dense crowds using front-view and sparse features. *arXiv preprint arXiv:1906.10313*, 2019.
- [32] Aniket Bera, Nico Galoppo, Dillon Sharlet, Adam Lake, and Dinesh Manocha. Adapt: real-time adaptive pedestrian tracking for crowded scenes. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 1801–1808. IEEE, 2014.
- [33] Donald Reid et al. An algorithm for tracking multiple targets. *IEEE transactions on Automatic Control*, 24(6):843–854, 1979.
- [34] Jur Van Den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics research*, pages 3–19. Springer, 2011.
- [35] Yuexin Ma, Dinesh Manocha, and Wenping Wang. Efficient reciprocal collision avoidance between heterogeneous agents using ctmat. In *International Conference on Autonomous Agents and Multiagent Systems*, 2018.
- [36] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.
- [37] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1345–1352. IEEE, 2011.
- [38] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 300–311, 2017.
- [39] Kuan Fang, Yu Xiang, Xiaocheng Li, and Silvio Savarese. Recurrent autoregressive networks for online multi-object tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 466–475. IEEE, 2018.
- [40] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple Online and Realtime Tracking with a Deep Association Metric. *arXiv preprint arXiv:1703.07402*, March 2017.
- [41] Harold W Kuhn. The hungarian method for the assignment problem. In *50 Years of Integer Programming 1958-2008*, pages 29–47. Springer, 2010.
- [42] Michael Levandowsky and David Winter. Distance between sets. *Nature*, 234(5323):34, 1971.
- [43] Rohan Chandra, Uttaran Bhattacharya, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. Roadtrack: Tracking road agents in dense and heterogeneous environments. *arXiv preprint arXiv:1906.10712*, 2019.
- [44] Edward Twitchell Hall. *The hidden dimension*, volume 609. Garden City, NY: Doubleday, 1966.
- [45] Satoru Satake, Takayuki Kanda, Dylan F Glas, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. How to approach humans?: strategies for social robots to initiate interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 109–116. ACM, 2009.
- [46] Aniket Bera and Dinesh Manocha. Realtime multilevel crowd tracking using reciprocal velocity obstacles. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4164–4169. IEEE, 2014.

- [47] Qing Zhang, Mengru Zhang, Mengdi Wang, Wanchen Sui, Chen Meng, Jun Yang, Weidan Kong, Xiaoyuan Cui, and Wei Lin. Efficient deep learning inference based on model compression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.
- [48] Meiqi Wang, Zhisheng Wang, Jinming Lu, Jun Lin, and Zhongfeng Wang. E-Istm: An efficient hardware architecture for long short-term memory. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2019.
- [49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [50] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [51] Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors. *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11205 of *Lecture Notes in Computer Science*. Springer, 2018.
- [52] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (1):1, 2013.
- [53] Chen Long, Ai Haizhou, Zhuang Zijie, and Shang Chong. Real-time multiple people tracking with deeply learned candidate selection and person re-identification. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2018.
- [54] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Proceedings of the IEEE international conference on computer vision*, pages 4705–4713, 2015.
- [55] Long Chen, Haizhou Ai, Chong Shang, Zijie Zhuang, and Bo Bai. Online multi-object tracking with convolutional neural networks. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 645–649. IEEE, 2017.
- [56] Min Yang, Yuwei Wu, and Yunde Jia. A hybrid data association framework for robust online multi-object tracking. *arXiv preprint arXiv:1703.10764*, 2017.
- [57] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4846–4855, 2017.
- [58] Ricardo Sanchez-Matilla, Fabio Poiesi, and Andrea Cavallaro. Online multi-target tracking with strong and weak detections. In *European Conference on Computer Vision*, pages 84–99. Springer, 2016.