Joint Prediction for Kinematic Trajectories in Vehicle-Pedestrian-Mixed Scenes

Huikun Bi[1,2] Zhong Fang[1] Tianlu Mao[1] Zhaoqi Wang[1] Zhigang Deng[2*]
[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Houston
{bihuikun,fangzhong,ltm,zqwang}@ict.ac.cn, zdeng4@uh.edu

Abstract

Trajectory prediction for objects is challenging and critical for various applications (e.g., autonomous driving, and anomaly detection). Most of the existing methods focus on homogeneous pedestrian trajectories prediction, where pedestrians are treated as particles without size. However, they fall short of handling crowded vehicle-pedestrian-mixed scenes directly since vehicles, limited with kinematics in reality, should be treated as rigid, non-particle objects ideally. In this paper, we tackle this problem using separate LSTMs for heterogeneous vehicles and pedestrians. Specifically, we use an oriented bounding box to represent each vehicle, calculated based on its position and orientation, to denote its kinematic trajectories. We then propose a framework called VP-LSTM to predict the kinematic trajectories of both vehicles and pedestrians simultaneously. In order to evaluate our model, a large dataset containing the trajectories of both vehicles and pedestrians in vehicle-pedestrian-mixed scenes is specially built. Through comparisons between our method with state-of-the-art approaches, we show the effectiveness and advantages of our method on kinematic trajectories prediction in vehicle-pedestrian-mixed scenes.

1. Introduction

Trajectory prediction is a challenging and essential task due to its broad applications in the computer vision field, including the navigation of autonomous driving, anomaly detection, and behavior understanding. Trajectory prediction for pedestrians has been extensively studied in recent years [34, 28, 5, 2, 9, 32, 29]. By encoding human-human interactions in a complex environment, these methods can predict future trajectories based on historical and surrounding human behaviors. Many methods have also been proposed to predict vehicle trajectories based on the states of surrounding vehicles [18, 14, 6].

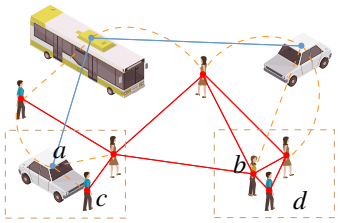


Figure 1. Illustration of various interactions in a vehicle-pedestrian-mixed scene. The vehicle-vehicle, human-human, and vehicle-human interactions are separately represented with solid blue lines, solid red lines, and orange dash lines. The vehicle *a* and pedestrian *b* in gray dash box have similar interactions with surrounding pedestrians. *b* walks freely to avoid collisions with *d*. However, the vehicle *a*, limited with kinematics, stops to avoid collisions with *c*.

All the above methods predict the trajectories of homogeneous traffic agents, namely, the whole scene with only pedestrians or only vehicles. Furthermore, in these methods, each agent is treated as a particle, with the same motion pattern. However, such naive simplifications are not suitable in common *vehicle-pedestrian-mixed* scenes, where vehicles and pedestrians have different sizes and motion patterns. As shown in Fig. 1, the interactions among different traffic agents in a vehicle-pedestrian-mixed scene include human-human, human-vehicle, and vehicle-vehicle interactions. The pedestrians with free movement are treated as particles, while the vehicles should be treated as rigid, non-particle objects ideally due to their sizes. Besides, only trajectories represented with positions are predicted in existing methods for traffic agents, which is insufficient to describe the accurate trajectories of heterogeneous vehicles in vehicle-pedestrian-mixed scenes. Different orientations along which vehicles drive forward in Fig. 1, will result in different interactions with surrounding agents. Besides, the kinematic motion of vehicles has been seldom considered yet in existing trajectory prediction literature. Therefore, predicting the accurate kinematic trajectories of heterogeneous vehicles, treated as rigid non-particle

*Corresponding Author

objects, as well as the pedestrian trajectories separately in vehicle-pedestrian-mixed scenes is of importance and generally considered as a widely open problem.

In this work, we treat a vehicle as a rigid non-particle object and use an *oriented bounding box (OBB)* to describe its detailed trajectory. Besides, we use the orientation of OBB to denote the driving-forward direction of a vehicle. Vehicles with the same position but different orientations will cause different interactions with surrounding agents. We further propose a Vehicle-Pedestrian LSTM (called VP-LSTM) to predict the trajectories of both pedestrians and vehicles simultaneously. The kinematic trajectories of vehicles can be learned and predicted based on their positions and orientations. All the aforementioned three types of interactions (vehicle-vehicle, human-human, and vehicle-human), are considered in our model. Through many experiments and comparisons with existing methods, we show the advantages of VP-LSTM on a large-scale, mixed traffic dataset that includes the trajectories of both vehicles and pedestrians.

The main contributions of this work include: (i) we propose a novel multi-task learning architecture, VP-LSTM, to jointly predict kinematic trajectories of both vehicles and pedestrians in vehicle-pedestrian-mixed scenes, where vehicles and pedestrians are treated as rigid bodies and particles, respectively. Thanks to the size information of heterogeneous vehicles, we exploit OBBs to represent vehicles and predict their positions and orientations. Because of the different trajectory definitions of vehicles and pedestrians, we adopt different methods to optimize the separate d-variate Gaussian distributions ($d = 4$ for vehicles and $d = 2$ for pedestrians). (ii) We introduce a large-scale and high-quality dataset containing the trajectories of both heterogeneous vehicles and pedestrians in two scenarios (BJI and TJI) under different traffic densities. The dataset is available at http://vr.ict.ac.cn/vp-lstm.

2. Related Work

Human Trajectory Prediction. Based on how features are selected, existing human trajectory prediction methods can be roughly divided into hand-crafted [11, 4, 17, 23, 31, 24, 30], and DNN-based. In general, hand-crafted features based methods are inefficient and only can generate limited results.

Recently, DNN-based methods have demonstrated superior performances due to the intrinsic encoding of complex human-human interactions in the network. Alahi et al. [2] proposed a social-LSTM model to predict the trajectories of pedestrians. Varshneya et al. [28] proposed a sequence-to-sequence model coupled with a soft attention mechanism to learn the motion patterns of dynamic objects. Bartoli et al. [5] adopted a "context-aware" LSTM model to predict human motion in crowded space. The DNN-based methods were also extended based on various attention mechanisms [7, 29]. Gupta et al. [9] used generative adversarial networks with a pooling module to predict the social pedestrians' motion. The CIDNN model [32] mapped the location to high dimensional feature space and used the inner product to encode crowd interactions. The joint prediction of trajectories with head poses and activities for pedestrians were respectively proposed [10, 21]. All these approaches well encoded human-human interactions with DNN models and could better predict human trajectories based on historical trajectory sequences and interactions.

Vehicle Trajectory Prediction. Based on different hypothesis levels, the task of vehicle trajectory prediction can be divided into the following categories [19]: physics-based, maneuver-based, and interaction-aware models. The Gaussian process regression flow [15] and the Bayesian nonparametric approach [13] ignore the interactions among objects in the scene. Vehicle trajectories can be predicted based on semantic scene understanding and optimal control theory [16]. Lee et al. [18] proposed DESIRE to predict future distances for interacting agents in dynamic scenes. Kim et al. proposed an LSTM-based probabilistic prediction approach [14] by building an occupancy grid map. Deo et al. built a convolutional social pooling network [6] to predict vehicle trajectories on highways. All the above methods focused on the macro behaviors of vehicles by treating vehicles as particles, but they fell short of characterizing the potential interactions among heterogeneous vehicles and pedestrians. Ma et al. proposed an LSTM-based algorithm, TrafficPredict, to predict trajectories for heterogeneous traffic agents [22]. But the kinematics of vehicles was ignored.

Human and Vehicle Trajectory Datasets. Quite a few human trajectory datasets have been built for the analysis of crowd behavior [20, 25, 35, 3, 27, 33]. A widely-known traffic dataset, including detailed vehicle trajectories and high-quality video, is the Next Generation Simulation (NGSIM) program [1]. Although the precise locations of vehicles are recorded, only vehicle-vehicle interaction behaviors are insufficient to describe vehicle-pedestrian-mixed scenes, especially in crowded space. Ma et al. used Apollo acquisition car to collect a trajectory dataset of heterogeneous traffic agents [22]. However, the available online portion of the Apollo dataset contains much noise that may be caused by LiDAR.

3. Our Method

Our goal is to predict the *kinematic* trajectories for *all heterogeneous* agents in vehicle-pedestrian-mixed scenes *jointly* and *simultaneously*. We present the details of the proposed VP-LSTM model in this section.

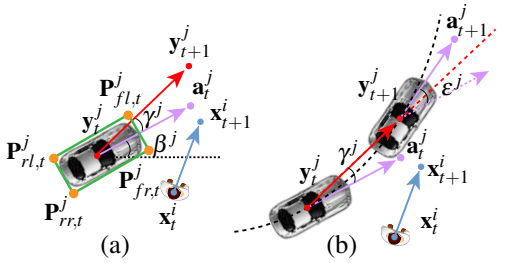


Figure 2. Illustration of the used symbols and terms. The velocity and orientation of vehicle v^j at t are respectively illustrated with a red arrow and a purple arrow. The blue arrow is the velocity of pedestrian p^i. (a) The green rectangle is OBB for v^j, whose four vertices are denoted as $\mathbf{P}_t^j = \{\mathbf{P}_{fl,t}^j, \mathbf{P}_{fr,t}^j, \mathbf{P}_{rr,t}^j, \mathbf{P}_{rl,t}^j\}$. β^j indicates the pan angle of the orientation from the X-axis. γ^j is the angle between the orientation and the velocity. (b) The black dash line is the motion path of vehicle v^j and ε^j denotes the angle between the orientations at two adjacent steps.

3.1. Formulation

We assume there are a total of N pedestrians and M vehicles, respectively in a vehicle-pedestrian-mixed scene. For a pedestrian p^i ($i \in [1,N]$), his/her trajectory at step t is represented by position $\mathbf{x}_t^i = (x,y)_t^i$. The input/output trajectory of a pedestrian is a sequence formed of consecutive positions.

Thanks to the size information of vehicles, we treat vehicles as rigid bodies, represented with OBBs. The input trajectory of a vehicle is represented by a temporal sequence of the four vertices on the OBB. As illustrated in Fig. 2(a), for a vehicle $v^j (j \in [1,M])$, its input trajectory at step t is represented by $\mathbf{P}_t^j = \{\mathbf{P}_{fl,t}^j, \mathbf{P}_{fr,t}^j, \mathbf{P}_{rr,t}^j, \mathbf{P}_{rl,t}^j\}$. Here $\mathbf{P}_{*,t}^j = (x_*, y_*)_t^i, * \in (fl, fr, rr, rl)$.

Due to the geometric constraints among the OBB vertices, we do not take the positions of the four vertices on an OBB as the output trajectory of the vehicle. Here we exploit the positions and orientations, represented by $\mathbf{y}_t^j = (x,y)_t^j$ and $\mathbf{a}_t^j = (\alpha_x, \alpha_y)_t^j$ respectively, as the output trajectory of v^j at step t. Inspired by the previous work of [10], in order to ensure the continuity of the orientations, we choose a vector representation, instead of the angular representation, to denote the orientation. $\mathbf{a}_t^j$ is the anchor point of the vector originating from $\mathbf{y}_t^j$, towards v^j oriented.

In this work, in order to jointly predict the trajectories of both vehicles and pedestrians, we feed the kinematic trajectory sequences of both pedestrians ($\mathbf{x}_t^i$) and vehicles ($\mathbf{P}_t^j$) in an observation period from step $t = 1$ to $t = T_{\text{obs}}$ as the input. Then, the positions of pedestrians, and both the positions ($\mathbf{y}_t^j$) and orientations ($\mathbf{a}_t^j$) of vehicles in the prediction period from step $t = 1$ to $t = T_{\text{pred}}$ can be predicted simultaneously.

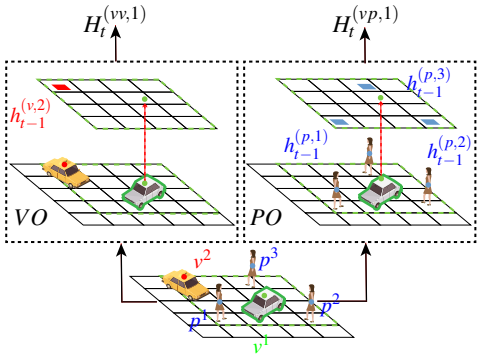


Figure 3. Illustration of mixed social pooling. For any agent involved in a vehicle-pedestrian-mixed scene (here we take vehicle v^1 as an example), the hidden states of its neighbors are separately pooled on VO and PO. The interactions from scene for v^1 are captured with $H_t^{(vp,1)}$ and $H_t^{(vv,1)}$.

3.2. Pedestrian and Vehicle Models

For any pedestrian p^i and any vehicle v^j, we first use separate embedding functions $\phi(\cdot)$ with ReLU nonlinearity to embed $\mathbf{x}_t^i$, $\mathbf{P}_t^j$ as follows:

$$\begin{aligned} e_t^{(x,i)} &= \phi(\mathbf{x}_t^i, W_x) \\ e_t^{(P_*,j)} &= \phi(\mathbf{P}_{*,t}^j, W_{P_*}), * \in (fl, fr, rr, rl) \\ e_t^{(\mathbf{P},j)} &= \phi(e_t^{(P_f,j)}, e_t^{(P_{fr},j)}, e_t^{(P_r,j)}, e_t^{(P_r,j)}, W_{\mathbf{P}}). \end{aligned} \tag{1}$$

Here W_x, W_{P_*}, and $W_{\mathbf{P}}$ are the embedding weights.

Mixed Social Pooling. The social pooling mechanism proposed in [2] and developed in [10, 9] can capture the motion dynamics of pedestrians in crowded space. We adopt this pooling scheme in our network to collect the latent motion representations of vehicles and pedestrians in the neighborhood. We use a similar grid of $N_o \times N_o$ cells in [2], called *occupancy map*, which is centered at the position of a pedestrian or vehicle. N_o denotes the size of the neighborhood. The positions of all the neighbors, including pedestrians and vehicles, are pooled on the occupancy map.

The hidden states of p^i and v^j, denoted as $h_t^{(p,i)}$ and $h_t^{(v,j)}$ respectively, carry their latent representations. Through the occupancy map, pedestrians and vehicles share the latent representations with hidden states. As shown in Fig. 3, the occupancy map VO and PO are built respectively for both vehicles and pedestrians. The pooling occurs on vehicle v^j involved in vehicle-pedestrian-mixed scenes as follows:

$$H_t^{(vp,j)}(m,n,:) = \sum_{k \in PO_{t-1}^j} h_{t-1}^{(p,k)}, \quad H_t^{(vv,j)}(m,n,:) = \sum_{l \in VO_{t-1}^j} h_{t-1}^{(v,l)}. \tag{2}$$

where $h_{t-1}^{(p,k)}$ is the hidden state of the pedestrians who are included into the PO of vehicle v^j; similarly, $h_{t-1}^{(v,l)}$ is the hidden state of the vehicles that are included into the VO of vehicle v^j, and m and n denote the indices of the $N_o \times N_o$ grid. So $H_t^{(vp,j)}$ and $H_t^{(vv,j)}$ carry the vehicle-human interac-

tions and vehicle-vehicle interactions respectively for vehicle v^j. As for pedestrian p^i, the human-human interactions and human-vehicle interactions are defined in a similar way, denoted as $H_t^{(pp,i)}$ and $H_t^{(pv,i)}$, respectively.

After mixed social pooling, separate embedding functions $\phi(\cdot)$ with ReLU nonlinearity are used to embed the heterogeneous interactions for v^j as follows:

$$e_t^{(vp,j)} = \phi(H_t^{(vp,j)}, W_H^{vp}), \quad e_t^{(vv,j)} = \phi(H_t^{(vv,j)}, W_H^{vv}). \quad (3)$$

Here W_H^{vp} and W_H^{vv} denote the corresponding embedding weights for vehicle v^j. $e_t^{(pp,i)}$ and $e_t^{(pv,i)}$ for pedestrian p^i are defined similarly with embedding parameters W_H^{pp} and W_H^{pv}.

Recursion for VP-LSTM. Finally, the recursion equations for pedestrian p^i and vehicle v^j are as follows:

$$\begin{aligned} h_t^{(p,i)} &= LSTM(h_{t-1}^{(p,i)}, e_t^{(x,i)}, e_t^{(pp,i)}, e_t^{(pv,i)}, W_{LSTM}^p) \\ h_t^{(v,j)} &= LSTM(h_{t-1}^{(v,j)}, e_t^{(\mathbf{P},j)}, e_t^{(vp,j)}, e_t^{(vv,j)}, W_{LSTM}^v) \end{aligned} \quad (4)$$

Here, W_{LSTM}^p and W_{LSTM}^v are respective LSTM weights for pedestrians and vehicles.

3.3. VP-LSTM Optimization

As a multi-task problem, we adopt different optimization methods for respective modules. The entire network is trained end-to-end by minimizing respective objectives of vehicles and pedestrians in the scene.

Optimization for Pedestrians. VP-LSTM estimates separate d-variate conditional distributions for pedestrians and vehicles, respectively. For pedestrians, we create a bivariate Gaussian distribution ($d = 2$) to predict the position $\hat{\mathbf{x}}_t^i = (\hat{x}, \hat{y})_t^i$. Following the work of [8], the distribution is parameterized by the mean $\mu_t^{(p,i)} = (\mu_x, \mu_y)_t^{(p,i)}$ and the covariance matrix $\Sigma_t^{(p,i)}$. Specifically, for a bivariate Gaussian distribution, $\Sigma_t^{(p,i)}$ can be obtained by optimizing the standard deviation $\sigma_t^{(p,i)} = (\sigma_x, \sigma_y)_t^{(p,i)}$ and the correlation coefficient $\rho_t^{(p,i)}$ [8].

The parameters of the module of pedestrians in VP-LSTM can be learned by minimizing a negative log-Likelihood loss as follows:

$$[\mu_t^{(p,i)}, \sigma_t^{(p,i)}, \rho_t^{(p,i)}] = W_O^p h_{t-1}^{(p,i)} \quad (5)$$

$$L^{(p,i)}(W_x, W_H^{pp}, W_H^{pv}, W_{LSTM}^p, W_O^p) = -\sum_{t=T_{\text{obs}}+1}^{T_{\text{pred}}} log(\mathbb{P}(\mathbf{x}_t^i | \mu_t^{(p,i)}, \sigma_t^{(p,i)}, \rho_t^{(p,i)})), \quad (6)$$

where $L^{(p,i)}$ is for the trajectory of the pedestrian p^i.

Optimization for Vehicles. Different from pedestrians, we use a four dimensional Gaussian multivariate distribution (d=4) to predict the position $\hat{\mathbf{y}}_t^j = (\hat{x}, \hat{y})_t^j$ and orientation $\hat{\mathbf{a}}_t^j = (\hat{\alpha}_x, \hat{\alpha}_y)_t^j$ of vehicles. Also, the distribution is parameterized by the mean $\mu_t^{(v,j)} = (\mu_x, \mu_y, \mu_{\alpha_x}, \mu_{\alpha_y})_t^{(v,j)}$ and the covariance matrix $\Sigma_t^{(v,j)}$. The previous work [10] studied a higher dimensional problem of the optimization of Gaussian parameters. For a higher dimensional problem, pairwise correlation terms cannot be optimized and used to build a covariance matrix. Its main reasons include: (i) the optimization process for each correlation term is independent; and (ii) multiple variables need to satisfy the positive-definiteness constraint [26].

Following the work of [10], we adopt the Cholesky factorization to optimize parameters for vehicles. With the Cholesky factorization $\Sigma_t^{(v,j)} = \mathbf{L}^T\mathbf{L}$, we first exponentiate the diagonal values for $\mathbf{L}$ to make it unique. Then, we use $\Sigma_t^{(v,j)} = \mathbf{L}^T\mathbf{L}$ to obtain the covariance matrix $\Sigma_t^{(v,j)}$. Here $\mathbf{L}$ is a 4×4 upper triangular matrix. The optimization process of a four dimensional Gaussian multivariate distribution can be transformed to search for ten scalar values in $\mathbf{L}$ and four mean parameters, namely, $\mu_t^{(v,j)} = (\mu_x, \mu_y, \mu_{\alpha_x}, \mu_{\alpha_y})_t^{(v,j)}$.

We denote ten vectorized scalar values in the upper triangular matrix $\mathbf{L}$ at t for v^j as $\theta_{L_t}^{(v,j)}$. The parameters of the vehicles' module can be learned by minimizing a negative log-Likelihood loss:

$$[\mu_t^{(v,j)}, \theta_{L_t}^{(v,j)}] = W_O^v h_{t-1}^{(v,j)} \quad (7)$$

$$L^{(v,j)}(W_{P_{fl}}, W_{P_{fr}}, W_{P_{rr}}, W_{P_{rl}}, W_{\mathbf{P}}, W_H^{vp}, W_H^{vv}, W_{LSTM}^v, W_O^v) = -\sum_{t=T_{\text{obs}}+1}^{T_{\text{pred}}} log(\mathbb{P}(\mathbf{y}_t^j, \mathbf{a}_t^j | \mu_t^{(v,j)}, \theta_{L_t}^{(v,j)})) \quad (8)$$

Here, $L^{(v,j)}$ is for vehicle v^j. In order to avoid over-fitting, we also add a l_2 regularization term onto the trajectory loss of pedestrians (Eq. 6) and vehicles (Eq. 8), respectively.

3.4. Displacements Prediction

Our model can simultaneously predict the future position $\hat{\mathbf{x}}_t^i = (\hat{x}, \hat{y})_t^i$ for pedestrian p^i, and both the position $\hat{\mathbf{y}}_t^j = (\hat{x}, \hat{y})_t^j$ and orientation $\hat{\mathbf{a}}_t^j = (\hat{\alpha}_x, \hat{\alpha}_y)_t^j$ for vehicle v^j. Through the occupancy maps respectively for pedestrians and vehicles, frame-by-frame heterogeneous interactions are pooled. The predicted kinematic trajectories of pedestrians and vehicles at t are respectively given by:

$$\begin{aligned} (\hat{x}, \hat{y})_t^i &\sim \mathcal{N}(\mu_t^{(p,i)}, \sigma_t^{(p,i)}, \rho_t^{(p,i)}) \\ (\hat{x}, \hat{y}, \hat{\alpha}_x, \hat{\alpha}_y)_t^j &\sim \mathcal{N}(\mu_t^{(v,j)}, \theta_{L_t}^{(v,j)}) \end{aligned} \quad (9)$$

Based on the sampled position $\hat{\mathbf{y}}_t^j = (\hat{x}, \hat{y})_t^j$ and orientation $\hat{\mathbf{a}}_t^j = (\hat{\alpha}_x, \hat{\alpha}_y)_t^j$ with Eq. 9, the input vertices of OBB, $\hat{\mathbf{P}}_{*,t+1}^j, * \in (fl, fr, rr, rl)$, at $t+1$ are given by:

$$\cos\hat{\beta}^j = \frac{(\hat{\mathbf{a}}_t^j - \hat{\mathbf{y}}_t^j) \cdot e_{\mathbf{x}}}{||\hat{\mathbf{a}}_t^j - \hat{\mathbf{y}}_t^j||} \quad (10)$$

$$\hat{\mathbf{P}}_{*,t+1}^j = \mathbf{P}_*^O \begin{bmatrix} \cos\hat{\beta}^j & \sin\hat{\beta}^j \\ -\sin\hat{\beta}^j & \cos\hat{\beta}^j \end{bmatrix}^T + \hat{\mathbf{y}}_t^j. \quad (11)$$

$e_{\mathbf{x}}$ is the unit vector along X-axis. $\mathbf{P}_*^O = \{\mathbf{P}_{fl}^O, \mathbf{P}_{fr}^O, \mathbf{P}_{rr}^O, \mathbf{P}_{rl}^O\}$ denotes the OBB centered at the coordinate origin and is

Table 1. The specifications of our dataset.

Property		Scenario I	Scenario II
Dataset name		BJI	TJI
City		Beijing	Tianjin
Latitude		40.219049N	39.120511N
Longitude		116.220789E	117.173421E
Traffic density		Low	High
Height of drone (meter)		74	121
Resolution (pixel)		3840×2160	3840×2160
Total video duration		39'58"	22'01"
Frame rate (fps)		30	30
Annotated frame number		23498	8000
Annotated frame rate (fps)		10	6
Annotated pedestrian number	Walking	1336	690
	Bike & Motor	1689	2690
	Total	3025	3380
Average pedestrian number per frame		29	46
Max pedestrian number per frame		67	105
Annotated vehicle number	Auto	2581	3523
	Bus & Truck	82	170
	Articulated bus	92	30
	Total	2755	3723
Average vehicle number per frame		19	34
Max vehicle number per frame		33	63

oriented the direction of positive X-axis, which is determined by the width w and length l of OBB.

As analyzed in [9], trajectory prediction is a multi-modal problem by nature, where each sampling produces one of multiple possible future trajectories. Apart from the variety loss function designed in [9], k_v acceptable kinematic trajectories for vehicles can be obtained by randomly sampling from the distribution $\mathcal{N}(\mu_t^{(v,j)}, \theta_{L_t}^{(v,j)})$ in our work. k_p possible trajectories for a pedestrian can be generated in a similar way by sampling $\mathcal{N}(\mu_t^{(p,i)}, \sigma_t^{(p,i)}, \rho_t^{(p,i)})$. The optimal predictions for vehicles and pedestrians at t can be chosen with $L_v = \min_{k_v} ||\hat{\mathbf{x}}_t^i(k_v) - \mathbf{x}_t^i||$ and $L_p = \min_{k_p} ||(\hat{\mathbf{y}}_t^j, \hat{\alpha}_t^j)(k_p) - (\mathbf{y}_t^j, \alpha_t^j)||$, respectively.

4. Vehicle-Pedestrian-Mixed Dataset

Existing human trajectory datasets [20, 25, 35, 3, 27, 33] only focus on homogeneous pedestrians. On the other hand, the existing vehicle trajectory dataset NGSIM [1] only captures the motion of vehicles. For this work, we specifically build a new vehicle-pedestrian-mixed dataset, which is designed for the trajectory analysis of vehicles and pedestrians in vehicle-pedestrian-mixed scenes. The original video data was acquired with a drone from a top-down view. We chose two traffic scenarios, where large heterogeneous vehicles and pedestrians pass through under different traffic densities. The trajectories in the two scenarios (called **BJI** and **TJI**, respectively) are carefully annotated, including 6405 pedestrians and 6478 vehicles (Fig. 4). Details of the dataset are summarized in Table 1.

Statistical Analysis. As aforementioned, the moving direction of a pedestrian (treated as a particle) can be simplified as its velocity. For vehicles, we calculated their γ in Fig. 2(b) of all the trajectory sequences in BJI. Then we show γ (in degrees) in ascending order in Fig. 5 (the solid orange line and axis). Note that only γ in the range [0,5] are reported, which is satisfied with vehicle kinematics. Those

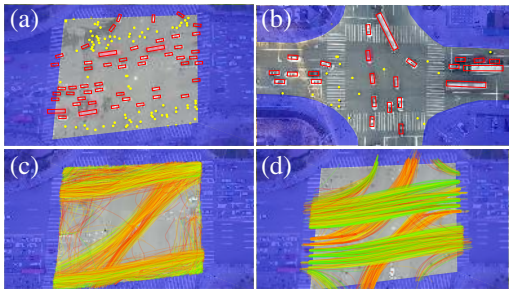


Figure 4. (a)(b) show annotated heterogeneous vehicles and pedestrians of one frame in TJI under a high traffic density and BJI under a low traffic density. (c)(d) separately show the examples of pedestrian and vehicle trajectories in TJI.

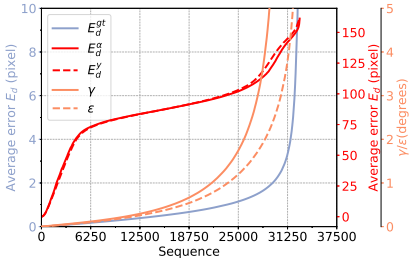


Figure 5. Analysis of BJI data. Solid and dashed orange curves separately represent the angles of γ and ε (Fig. 2(b)). The other three curves analyze the error E_d (Eq. 12).

cases with γ over 5 degrees, caused by noise, were omitted. This shows that vehicles, including size information, are significantly different from pedestrians, and the directions of velocities cannot represent the orientations of the vehicles directly.

The orientation of vehicle v^j at t (Fig. 2(b)) is the trajectory tangent at t. In order to obtain the orientation, besides the historical trajectory information, the positions $\mathbf{y}^j$ at several subsequent steps are also needed. Therefore, it is infeasible to obtain accurate orientations in the forecasting phase. We also plot ε^j, namely, the angle between the orientations at two consecutive steps, for all the trajectory sequences of the vehicles in BJI (the dashed orange line and axis). As shown in Fig. 5, a small ε indicates a small turning angle of a vehicle between two consecutive steps. Intuitively, we can approximately use its known orientation at $t-1$ to estimate its orientation at t.

In order to evaluate the relationship between the velocity and the orientation of a vehicle, we define the following error for v^j:

$$E_d = ||\mathbf{y}_t^j - \mathbf{y}_{t-1}^j|| - ||\mathbf{P}_{fm,t}^j - \mathbf{P}_{fm,t-1}^j||. \quad (12)$$

$\mathbf{P}_{fm}^j = \frac{1}{2}(\mathbf{P}_{fl}^j + \mathbf{P}_{fr}^j)$ is the midpoint of the front side, which also corresponds to $\mathbf{a}^j$. We plot the results (denoted as E_d^{gt}) in Fig. 5 (the solid blue line and axis). As seen from this figure, the errors between the displacements of $\mathbf{y}^j$ and the average displacements of $\mathbf{a}^j$, denoted by $\mathbf{P}_{fm}^j$, in two consecutive steps are small and consistent. However, we use

Metric	V-LSTM [2]	S-LSTM [2]	SGAN[9]		VP-LSTM (Ours)			
			1VP-1	20VP-20	OP-1	P-20	O-20	OP-20
ADE	34.01 / 40.00	11.73 / 15.16	4.80 / 7.05	4.56 / 6.52	7.72 / 10.46	5.74 / 7.95	3.90 / 5.67	**2.19 / 2.99**
FDE	43.87 / 52.64	20.03 / 23.64	9.17 / 11.07	9.13 / 10.99	13.02 / 18.21	10.27 / 14.34	7.12 / 10.31	**3.70 / 5.20**
ADE_O	33.89 / 39.87	12.68 / 15.75	5.69 / 8.15	5.26 / 7.52	8.27 / 10.88	6.76 / 8.90	4.93 / 6.60	**3.29 / 4.00**
FDE_O	43.77 / 52.50	22.59 / 24.95	11.18 / 13.33	11.07 / 13.10	13.45 / 18.43	11.39 / 15.33	8.19 / 11.19	**4.88 / 6.22**

Table 2. Quantitative results for the predicted positions and orientations of vehicles in NGSIM. Metrics ADE, FDE, ADE_O and FDE_O for $T_{pred} = 8$ and $T_{pred} = 12$ (8/12) are reported in feet. Our method consistently outperforms the state-of-the-art methods (lower is better).

Metric	Dataset	Agent	V-LSTM [2]	S-LSTM [2]	SGAN[9]		VP-LSTM (Ours)			
					1VP-1	20VP-20	OP-1	P-20	O-20	OP-20
ADE	BJI	Vehicle	66.69 / 85.48	29.05 / 51.41	23.05 / 29.74	20.21 / 24.65	51.33 / 79.40	22.94 / 34.50	17.25 / 27.12	**16.38 / 24.33**
		Pedestrian	34.70 / 48.91	25.26 / 46.89	19.73 / 26.06	17.82 / 20.33	23.25 / 32.29	4.84 / 6.23	**4.79 / 6.21**	4.92 / 6.39
		Average	44.13 / 62.09	26.88 / 48.49	21.52 / 28.66	18.64 / 22.53	32.21 / 47.33	10.61 / 15.26	8.77 / 12.89	**8.58 / 12.72**
ADE	TJI	Vehicle	142.17 / 185.93	46.17 / 85.52	40.20 / 59.30	26.82 / 39.86	64.40 / 96.91	56.21 / 89.54	26.43 / 47.66	**22.38 / 29.79**
		Pedestrian	115.44 / 135.97	41.19 / 75.55	21.81 / 23.67	19.81 / 25.89	37.19 / 50.91	8.32 / 10.55	9.89 / 12.54	**7.42 / 9.12**
		Average	125.27 / 154.31	43.13 / 79.22	31.30 / 46.57	24.42 / 34.73	48.67 / 70.34	24.79 / 36.48	16.86 / 27.37	**13.43 / 17.32**
FDE	BJI	Vehicle	114.11 / 152.91	61.49 / 126.03	39.41 / 46.02	38.36 / 44.68	94.91 / 153.96	42.87 / 65.47	34.16 / 54.20	**31.27 / 43.60**
		Pedestrian	54.64 / 81.72	56.93 / 111.05	32.90 / 41.37	32.57 / 40.52	38.82 / 56.63	7.40 / 10.15	**7.28 / 10.09**	7.55 / 10.44
		Average	72.17 / 107.38	58.54 / 116.47	37.88 / 44.75	35.00 / 42.62	56.71 / 87.71	18.71 / 27.81	15.86 / 24.18	**15.11 / 23.47**
FDE	TJI	Vehicle	215.94 / 303.54	103.67 / 203.90	50.62 / 59.46	48.22 / 56.95	114.79 / 181.80	109.64 / 176.37	56.59 / 102.88	**35.38 / 49.31**
		Pedestrian	156.29 / 192.92	92.55 / 177.10	40.23 / 50.93	39.98 / 49.31	62.03 / 88.97	12.47 / 16.56	15.07 / 19.92	**10.53 / 13.90**
		Average	178.21 / 233.52	96.91 / 186.97	46.55 / 57.89	43.42 / 55.43	84.29 / 128.17	45.89 / 69.04	32.59 / 54.95	**20.51 / 27.95**
ADE_O	BJI	vehicle	65.51 / 83.78	42.55 / 65.70	34.41 / 41.59	27.56 / 33.47	58.54 / 83.68	33.38 / 42.92	27.28 / 34.95	**26.65 / 32.49**
	TJI	vehicle	140.52 / 183.60	50.35 / 88.44	50.75 / 56.79	29.69 / 38.65	68.11 / 99.83	60.31 / 93.17	32.87 / 53.51	**26.15 / 33.69**
FDE_O	BJI	vehicle	112.01 / 149.81	76.18 / 135.49	46.28 / 51.01	43.61 / 49.59	98.74 / 153.40	53.19 / 72.42	43.26 / 60.13	**40.61 / 48.02**
	TJI	vehicle	213.04 / 299.39	105.94 / 203.24	64.69 / 79.57	50.94 / 64.49	117.21 / 182.70	112.84 / 178.57	61.89 / 107.04	**38.93 / 52.73**

Table 3. Quantitative results for the predicted positions and orientations of separate traffic agents based on our dataset (BJI and TJI). Each error metric for $T_{pred} = 8$ and $T_{pred} = 12$ (8/12) is reported in pixels. The top four rows show the errors ADE and FDE for the predicted positions of the heterogeneous traffic agents. The bottom rows show the errors ADE_O and FDE_O for the predicted orientations of vehicles denoted as OBB. Our method consistently outperforms the state-of-the-art methods (lower is better).

the orientation $\mathbf{a}_{t-1}^j$ to approximate $\mathbf{a}_t^j$, and then obtain E_d^α using Eq. 11 and Eq. 12.

In addition, we also calculated E_d^y with Eq. 11 and Eq. 12, where $\mathbf{a}_t^j$ is estimated based on the direction of the velocity $(\overrightarrow{\mathbf{y}_t^j - \mathbf{y}_{t-1}^j})$ at $t-1$. As shown in Fig. 5 (solid/dashed red lines and axis), although γ and ε are small, the direct estimation of $\mathbf{a}_t^j$ from $\mathbf{a}_{t-1}^j$ and $\overrightarrow{\mathbf{y}_t^j - \mathbf{y}_{t-1}^j}$ will cause a larger error if we treat vehicles as rigid objects. Therefore, it is necessary to predict the orientation together with position simultaneously in order to obtain more accurate kinematic trajectories of vehicles in crowded space. We also use the Pearson circular correlation coefficient [12] to measure the correlations between $||\mathbf{y}_t - \mathbf{y}_{t-1}||$ and $||\mathbf{a}_t - \mathbf{a}_{t-1}||$; the obtained correlation is 0.97. This indicates that the faster a vehicle moves, the more significant change of $\mathbf{a}$ will have. The above statistical analysis inspired us to build VP-LSTM to learn separate trajectories for pedestrians and vehicles. Specifically, the orientations and positions are jointly learned for vehicles in order to obtain accurate kinematic trajectories.

5. Experiments and Evaluation

We compared VP-LSTM with state-of-the-art human trajectory prediction methods and present quantitative and qualitative evaluation results in this section.

5.1. Quantitative Evaluation

We compared VP-LSTM with state-of-the-art human trajectory prediction methods including Vanilla LSTM (**V-LSTM**) [2], Social LSTM (**S-LSTM**) [2], two Social-Gan variants (**SGAN-PV** and **SGAN-PV-20**) [9]. To be consistent with the evaluation protocol in [9], the length of the observation period T_{obs} was set to 8, and we separately predicted the trajectories with two different lengths T_{pred} (i.e., 8 and 12).

Evaluation Metrics. Inspired by the previous work [25], we chose **ADE** (Average Displacement Error: the average Euclidean distance error between the predicted result and ground truth over the whole sequence) and **FDE** (Final Displacement Error: the Euclidean distance error at the last step between the predicted result and ground truth) as the evaluation metrics to evaluate the prediction results on the positions of both vehicles and pedestrians.

In order to evaluate the orientations predicted for the kinematic trajectories of vehicles, we define $\mathbf{ADE_O}$ and $\mathbf{FDE_O}$ to describe the errors between predicted orientations and ground truth, as follows:

$$\mathrm{ADE_O} = \frac{\sum_{j=1}^{M} \sum_{t=T_{obs}+1}^{T_{obs}+T_{pred}} ||\hat{\mathbf{P}}_{fm,t}^j - \mathbf{P}_{fm,t}^j||}{MT_{pred}}. \quad (13)$$

$\mathrm{FDE_O}$ is the average Euclidean distance error for $t = T_{obs} + T_{pred}$ for $\mathbf{P}_{fm}^j (j \in [1, M])$, and $\hat{\mathbf{P}}_{fm}^j$ is the predicted midpoint of the front of OBB oriented. Since all of V-LSTM, S-LSTM, SGAN-PV, and SGAN-PV-20 treat each vehicle as a particle and only predict its positions, we use its orientations at $t-1$ to approximate its orientations at t in these methods, based on the statistical analysis in Sec. 4.

We also performed an ablation study with different control settings to evaluate the contribution of each part of the network. Our methods are referred to as VPLSTM-OP-N. Here O denotes that the vehicles in scenes are treated as

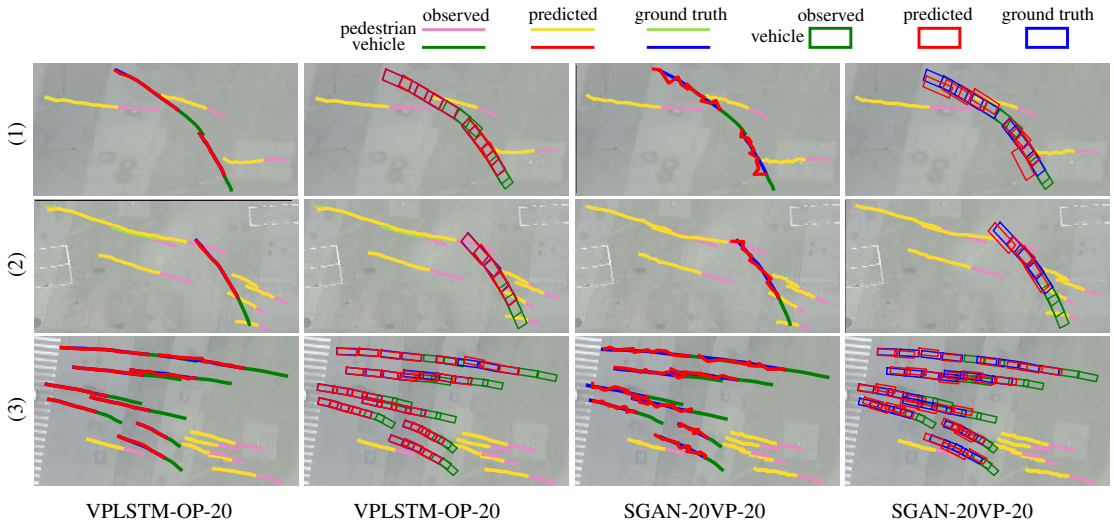


Figure 6. Three examples of the predicted trajectories compared with the ground truth and SGAN-20VP-20. The left and center-right column show the position trajectories of both vehicles and pedestrians predicted by VPLSTM-OP-20 and SGAN-20VP-20, respectively. The kinematic trajectories of vehicles illustrated in the center-left and right column are represented by OBB. Here $T_{\text{obs}} = 8$ and $T_{\text{pred}} = 12$. In order to clearly illustrate kinematic trajectories, we sample trajectories and show vehicles at $t = 3, 6, 9, 12, 15, 18$. More examples are provided in the supplementary material.

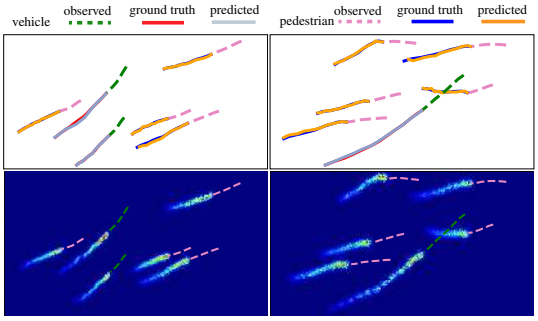
OBB, whose orientations and positions are predicted jointly, and P signifies the mixed social pooling are adopted in the model. Assume L_v and L_p denote the optimal predictions of vehicles and pedestrians in the test phase, respectively. We randomly sample N times from each learned d-variate Gaussian distributions. Note that the vehicles without predicted orientations in VPLSTM-P-N are treated as particles, and we employ the same optimization as for pedestrians.

Scenes with Vehicles only. In order to evaluate the kinematic trajectory definition for vehicles, we test our model on a publicly available dataset, NGSIM [1], that includes the trajectories of heterogeneous vehicles with different sizes.

As shown in Table 2, V-LSTM had the highest error on each metric in terms of position prediction and orientation estimation, since it cannot capture the interactions among vehicles. All of S-LSTM, SGAN-1VP-1, and SGAN-20VP-20 can capture the interactions among vehicles and perform better than V-LSTM. Due to the variety loss adopted in SGAN-20VP-20, which encourages the network to produce diverse samples, the SGAN-20VP-20 obtained better performance than other comparative methods (except ours). However, these models treat each vehicle as a particle and ignore the heterogeneous size information of various vehicles. VPLSTM-OP-20 significantly outperformed VPLSTM-P-20 due to the kinematic feature of vehicles. In a nutshell, by exploiting the OBB of each vehicle, the size information, including positions and orientations, are implicitly encoded, which helps to predict more accurate kinematic trajectories for vehicles.

Vehicle-Pedestrian-Mixed Scenes. We further evaluated our model on our vehicle-pedestrian-mixed trajectories dataset. Because V-LSTM, S-LSTM, and SGAN are built for predicting human trajectories, and each agent in these methods does not distinguish the difference between vehicles and pedestrians. We adopted the assumption in these methods, where each agent, treated as a particle, has the same motion pattern.

We compared the accuracy of position prediction using ADE and FDE for vehicles and pedestrians in Table 3. Because V-LSTM can predict the trajectory of any traffic agent based on its historical trajectory, ignoring interactions, it had the highest error in terms of the prediction of both positions and orientations. Although both S-LSTM and SGAN capture the interactions among heterogeneous traffic agents and perform better than V-LSTM, these models treat each agent as a particle and cannot distinguish the difference among various interactions. The pedestrians and vehicles with different motion patterns share the same LSTM in SGAN, which produced higher ADE and FDE errors on the position prediction, especially for pedestrians. Besides, it is apparent that the position ADE and FDE errors of pedestrians predicted by SGAN, VPLSTM-P-20, and VPLSTM-O-20 are small, but the position errors for vehicles are higher due to the assumption that vehicles are treated as particles and share the same LSTM with pedestrians. VPLSTM-OP-1 obtained high errors as the predicted trajectories can be any of the multiple possible trajectories from the learned distributions. SGAN-20VP-20, VPLSTM-O-20, and VPLSTM-OP-20 predicted the close position results for vehicles in the BJI dataset due to its low traffic density. By using OBB to encode size and orientation information for the vehicle and adopting separate LSTMs to



(1) (2)

Figure 7. Two examples of comparison between our model and ground truth (top row). For each traffic agent, we visualize its probability distribution at one time-step in predicted duration (bottom).

represent traffic agents, our VPLSTM-OP-20 captures all possible scenarios and produces the best trajectories from various samples for both vehicles and pedestrians. Please refer to supplemental material for more results.

With the defined metrics $\text{ADE}_{\mathbf{O}}$ and $\text{FDE}_{\mathbf{O}}$, the orientations predicted for vehicles were also compared with the estimated orientations in the two bottom rows of Table 3. The predicted results of VPLSTM-OP-20 are significantly better than the results estimated based on the last orientation. This shows that our model can predict more accurate and stable kinematic trajectories for vehicles, including both positions and orientations.

Summary. The main findings from our quantitative experiments include: 1) In vehicle-pedestrian-mixed scenes, separately encoding behaviors and interactions among both vehicles and pedestrians are indeed necessary; 2) vehicles, limited with kinematics, are different from pedestrians, and they should be treated as rigid bodies to predict more accurate kinematic trajectories, which also contributes to the prediction of pedestrian trajectories; 3) predicting the orientations of vehicles is indeed necessary, and the predicted orientations improve the performance than the approximation of the orientations using straightforward geometric methods. Our VP-LSTM model can predict the trajectories of both pedestrians and vehicles simultaneously. Not only the positions but also the orientations of vehicles can be more accurately predicted in crowded scenes.

5.2. Qualitative Evaluation

We show some qualitative evaluation results on our dataset in Fig. 6. As discussed above, the motions of both pedestrians and vehicles are influenced by the neighboring traffic agents. We report the kinematic trajectories of vehicles, represented with OBB, which are computed based on the predicted positions and orientations. As shown in Fig. 6, pedestrians' trajectories were predicted by both VPLSTM-OP-20 and SGAN-20VP-20, which indicates the trajectories of pedestrians (as particles) are more straightforward than vehicles (limited with kinematics). Although vehicles, treated as particles in SGAN-20VP-20, also can forecast more comparatively accurate trajectories (as shown in the center-right column). However, the kinematic trajectories estimated with the last orientation will still result in more significant errors. Our model can learn different patterns for vehicles and pedestrians involved in various interactions, including following, merging, and avoiding (see Fig. 6). When vehicles drive in a group mixed with pedestrians, the trajectories predicted by our VP-LSTM model is close to the ground truth (see the examples (2)(3) in Fig. 6).

We visualize the separate probability distributions of vehicles and pedestrians predicted with our model at one time-step in Fig. 7. In the two examples, we observe that the vehicles capture various heterogeneous interactions to avoid collisions. Pedestrians in crowded spaces also find their ways to keep moving forward. Hence, different distributions for different traffic agents, learned by our model, make the prediction suitable in widely existing, vehicle-pedestrian-mixed scenes.

6. Conclusion

In this work, to tackle the problem of trajectory prediction for vehicles and pedestrians simultaneously in crowded vehicle-pedestrian-mixed scenes, we build a carefully-annotated dataset under different traffic densities, and further proposed a VP-LSTM framework to predict the trajectories of both vehicles and pedestrians jointly. The trajectories of vehicles in crowded spaces are limited to kinematics and involved with pedestrians. These vehicles, treated as rigid and defined with more accurate representations (OBBs), can be computed based on the predicted positions and orientations. To generate the kinematic trajectories for different traffic agents, we adopt different optimizations for vehicles and pedestrians. We also report the performance of our work compared with various state-of-the-art methods. Further research on a more accurate trajectory prediction of vehicles involved with pedestrians and limited with kinematics can improve the efficiency and accuracy of autonomous driving.

7. Acknowledgement

This work is in part supported by the National Key Research and Development Program of China (2017YFC0804900, 2017YFB1002600), the National Natural Science Foundation of China (61532002, 61702482), the 13th Five-Year Common Technology pre Research Program (41402050301-170441402065), the Science and Technology Mobilization Program of Dongguan (KZ2017-06). Zhigang Deng is in part supported by US NSF grant IIS-1524782.

References

[1] Next generation simulation fact sheet. https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm, 2018.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. Socially-aware large-scale crowd forecasting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[4] Gianluca Antonini, Michel Bierlaire, and Mats Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006.

[5] Federico Bartoli, Giuseppe Lisanti, Lamberto Ballan, and Alberto Del Bimbo. Context-aware trajectory prediction. In *International Conference on Pattern Recognition (ICPR)*, pages 1941–1946. IEEE, 2018.

[6] Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1468–1476, 2018.

[7] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural Networks*, 108:466–478, 2018.

[8] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[9] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Alessio Del Bue, Fabio Galasso, and Marco Cristani. Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[12] S.R. Jammalamadaka and A. Sengupta. *Topics in Circular Statistics*. Series on multivariate analysis. World Scientific, 2001.

[13] Joshua Joseph, Finale Doshi-Velez, Albert S Huang, and Nicholas Roy. A bayesian nonparametric approach to modeling motion patterns. *Autonomous Robots*, 31(4):383, 2011.

[14] ByeoungDo Kim, Chang Mook Kang, Jaekyum Kim, Seung Hi Lee, Chung Choo Chung, and Jun Won Choi. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pages 399–404. IEEE, 2017.

[15] Kihwan Kim, Dongryeol Lee, and Irfan Essa. Gaussian process regression flow for analysis of motion trajectories. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1164–1171. IEEE, 2011.

[16] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012.

[17] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering: a partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 593–604. ACM, 2007.

[18] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[19] Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *Robomech Journal*, 1(1):1, 2014.

[20] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[21] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G. Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6120–6127, 2019.

[23] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 935–942. IEEE, 2009.

[24] Brendan Morris and Mohan Trivedi. Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 312–319. IEEE, 2009.

[25] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 261–268. IEEE, 2009.

[26] Mohsen Pourahmadi. Covariance estimation: The glm and regularization perspectives. *Statistical Science*, pages 369–387, 2011.

[27] Jing Shao, Chen Change Loy, and Xiaogang Wang. Scene-independent group profiling in crowd. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[28] Daksh Varshneya and G Srinivasaraghavan. Human trajectory prediction using spatially aware deep attention models. *arXiv preprint arXiv:1705.09436*, 2017.

[29] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018.

[30] Xiaogang Wang, Keng Teck Ma, Gee-Wah Ng, and W Eric L Grimson. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International Journal of Computer Vision*, 95(3):287–312, 2011.

[31] Xiaogang Wang, Xiaoxu Ma, and W Eric L Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on pattern analysis and machine intelligence*, 31(3):539–555, 2009.

[32] Yanyu Xu, Zhixin Piao, and Shenghua Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[33] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[34] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Pedestrian behavior understanding and prediction with deep neural networks. In *European Conference on Computer Vision*, pages 263–279. Springer, 2016.

[35] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3441–3448. IEEE, 2011.