How Can I See My Future? FvTraj: Using First-person View for Pedestrian Trajectory Prediction

Huikun Bi[1,2], Ruisi Zhang[3], Tianlu Mao[1,2], Zhigang Deng[4], and Zhaoqi Wang[1,2]

[1] Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences, Beijing, China
[3] University of Utah, Salt Lake City, USA
[4] University of Houston, Houston, USA
{bihuikun,ltm,zqwang}@ict.ac.cn, ruisi.zhang@utah.edu, zdeng4@uh.edu

Abstract. This work presents a novel **F**irst-person **V**iew based **Traj**ectory predicting model (FvTraj) to estimate the future trajectories of pedestrians in a scene given their observed trajectories and the corresponding first-person view images. First, we render first-person view images using our in-house built **F**irst-person **V**iew **Sim**ulator (FvSim), given the ground-level 2D trajectories. Then, based on multi-head attention mechanisms, we design a social-aware attention module to model social interactions between pedestrians, and a view-aware attention module to capture the relations between historical motion states and visual features from the first-person view images. Our results show the dynamic scene contexts with ego-motions captured by first-person view images via FvSim are valuable and effective for trajectory prediction. Using this simulated first-person view images, our well structured FvTraj model achieves state-of-the-art performance.

Keywords: deep learning, human behavior, trajectory prediction, crowd simulation, multi-head attention.

1 Introduction

Pedestrian trajectory prediction has attracted increasing attention of researchers in computer vision community due to its various potential applications including robotic navigation, autonomous driving, and anomaly detection [21, 5]. It is often necessary to consider all three major inherent properties of pedestrian trajectory prediction: social interactions, multimodality, and scene contexts. The first two properties have been well considered in the state-of-the-art frameworks [1, 13, 18, 17]. Scene contexts are particularly essential yet challenging for modern studies, since they contain both the stationary obstacles (e.g., buildings, trees) and dynamic objects (e.g., moving pedestrians). Recently, researchers have started to exploit scene contexts for pedestrian trajectory prediction [37, 47, 9, 22], using

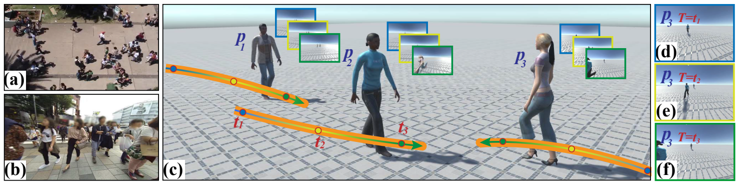

Fig. 1. (a) A top-down view image [26] contains large-scale scene contexts. (b) A first-person view image from the First-Person Locomotion (FPL) dataset [47] contains dynamic scene contexts (i.e., moving pedestrians). (c) A simulated scenario using FvSim. (d), (e), and (f) are the first-person view images of the pedestrian p_3 in (c) at step t_1, t_2, and t_3, respectively. Note, the first-person view images are individual-specific and are not shared among pedestrians in a scene.

either top-down view images (Fig. 1(a)) or first-person view images (Fig. 1(b)). However, both methods suffer from the following limitations.

Top-down view images are easy to access and applicable to anomaly detection, but provide limited dynamic scene contexts. Prior works [37, 9, 22] introduced the top-down view images (Fig. 1(a)), which are shared between pedestrians, to mainly capture the stationary obstacles. But it is difficult to capture each individual pedestrian's detailed dynamic information (e.g., poses, ego-motions, visual occlusion information) due to the pixel-precise in a top-down view image.

First-person view images provide detailed dynamic scene contexts with ego-motions, but is difficult to access for each pedestrian in a scene. They are applicable to various applications like blind navigation [27, 19], robotic navigation [36], and autonomous driving [8, 33, 39]. They (Fig. 1(b)) can well capture moving pedestrians by observing the ego-motion of each pedestrian (i.e., camera wearer), the pedestrian's visual perspective effect on the neighbors, and pedestrians' detailed poses [47]. Obtaining comprehensive and accurate scene contexts requires the first-person view images from each pedestrian in a scene, since the images from a single pedestrian can only provide partial scene contexts and the relative position of each pedestrian. In reality, to do so we need to mount at least one camera per pedestrian in a scene, which is expensive, time-consuming, and sometimes infeasible.

To overcome these limitations, we first build an in-house simulator *FvSim* using Unity to generate a dynamic virtual environment and render the corresponding first-person view images for each pedestrian based on the observed trajectories. This environment is proportional to the real-world environment using SI units, which can be generalized to any dataset collected in real world using unit conversions and some linear transformations between coordinate systems. Unlike physically collecting images required a large number of camera-wearing robots moving in a given scene, our FvSim requires zero physical cameras (i.e., low-cost) to provide desired information of given pedestrians, which is ideal for capturing dynamic scene contexts. We also use FvSim to evaluate the effectiveness and importance of first-person view information for some trajectory-prediction tasks.

We then propose *FvTraj*, a model to predict future trajectories of pedestrians, by considering two given inputs: the observed trajectory of each pedestrian in a scene, and the pedestrians' corresponding first-person view images simulated by FvSim. Without any preconceived scene contexts (e.g., top-down view information), FvTraj considers trajectory prediction holistically by taking into account historical motion patterns, social interactions, and self dynamic scene contexts with ego-motions. Through experimental comparisons with various state-of-the-art models, we show that FvTraj can achieve better performance.

The main contributions of this work can be summarized as: (1) To address the problem of hardware limitation commonly faced in the pedestrian trajectory prediction task, we develop FvSim, a trajectory simulator that is capable of providing multi-view information in the scene. We show the first-person view images via FvSim could be valuable and effective for trajectory prediction. (2) We develop FvTraj, a novel architecture to predict future trajectories of pedestrians based on historical trajectories and the corresponding first-person view images of each pedestrian in a scene. Our FvTraj uses social-aware attention and view-aware attention based on a multi-attention mechanism, which captures both social behaviors and visual features including ego-motions.

2 Related Work

In this section, we will mainly focus on the reviewing of recent related efforts on pedestrian trajectory prediction [1, 13, 18, 37, 9, 29, 7].

Social Interaction Schemes. Prior works had successfully presented that hand-crafted features of pedestrians are essential for modeling social interactions [16, 3, 24, 31, 45, 32, 44]. Recently, some works [1, 13, 14] modeled complex human-human interactions using DNN-based methods, which adopt social pooling schemes to describe the social behaviors and assign equal importance of neighboring pedestrians. Attention-based models [43, 37, 2, 49] intentionally select useful information from neighboring pedestrians based on the relative locations and motion correlations. Furthermore, by adopting a graph to describe the pedestrians in a scene, a graph attention model (GAT) was proposed to model social interactions in order to generate realistic pedestrian trajectories [17, 22].

Semantic Scene Contexts. The physical scene around pedestrians is important for trajectory prediction, because visually stationary or dynamic obstacles (e.g., buildings, trees, and moving pedestrians) generally influence pedestrians' trajectories. Lee et al. [25] built a scene context fusion unit to encode semantic context information in dynamic scenes. Sadeghian et al. [38] proposed the Car-net that uses single-source and multi-source attention mechanisms to visualize fine-grained semantic elements of navigation scenes. Sadeghian et al. [37] proposed Sophie that could produce plausible social trajectories using pre-trained CNN to extract the visual features. Choi et al. [9] visually extract spatiotemporal features of static road structures, road topology, and road appearance. Liang et al. [28] proposed a person interaction module to encode both the nearby scene of a person, as well as the explicit geometric relations and the surrounding object types in the scene. Kosaraju et al. [22] proposed Social-BiGAT that applies soft attention to capture physical features in the scene context.

First-person View. In some applications like autonomous driving and robotic navigation, the most naturally accessible visual input for trajectory prediction is the first-person view [40, 6, 47, 29]. Yagi et al. [47] proposed a method to predict the future location of a person seen in a first-person video based on the ego-motions of the video, poses, scales, and locations of the person. Yao et at. [48] proposed an unsupervised approach for traffic accident detection in first-person videos. Lai et at. [23] proposed a new collision avoidance system for the first-person viewpoint, to show the trajectory of people and to predict the future location. Ma et al. [29] proposed Trafficpredict to predict the trajectories of heterogeneous agents based on a proposed first-person view driving dataset. Of the particular interest in the field of autonomous driving, there is a variety of driving datasets recorded in the first-person view (i.e., collected by the cameras rigidly attached to vehicles), which could be potentially used to train a model to predict the trajectories of heterogeneous road users [12, 30, 35].

3 Methodology

As aforementioned, our proposed *FvTraj* model can output the future trajectories of pedestrians in a scene, given their previous motion states and the corresponding first-person view images simulated by the *FvSim* simulator.

3.1 Problem Formulation

Trajectory prediction for pedestrians can be formally defined as the problem of predicting the future trajectory of any focus pedestrian in a scene, given the pedestrian's previous states and the scene information. We consider the previous states of a pedestrian $p^i (i \in [1, N])$ in a N-pedestrians scene as a two-dimensional (2D) position $X_t^i = (x_t^i, y_t^i)$ within an observation period from time step $t = 1$ to $t = T_{\text{obs}}$. We denote the trajectory of p^i in a period from $t = T_{\text{start}}$ to $t = T_{\text{end}}$ as $X^i_{t \in [T_{\text{start}}, T_{\text{end}}]}$. In our case, the scene information is described as the first-person view image I_t^i of the pedestrian p^i within the same observation period, which is denoted as $I^i_{t \in [1, T_{\text{obs}}]}$. Each focus pedestrian in the scene does not share their first-person view images with others. Given the above two input variables, $X^i_{t \in [1, T_{\text{obs}}]}$ and $I^i_{t \in [1, T_{\text{obs}}]}$ $(i \in [1, N])$, the goal of our model is to output the 2D position of each pedestrian in the scene within the prediction period from $t = T_{\text{obs}} + 1$ to $t = T_{\text{pred}}$, which is denoted as $X^i_{t \in [T_{\text{obs}}+1, T_{\text{pred}}]}$.

Although the first-person view images are accessible in some applications, it is difficult to access the first-person view images for all pedestrians in the scene due to practical cost and technical difficulty. This could be formally defined as the problem of simulating the first-person view images for each pedestrian in a scene, given their observed states.

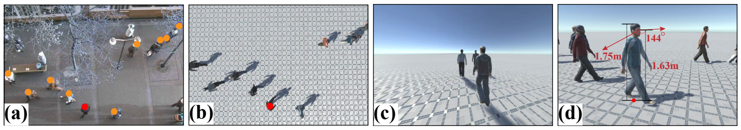

Fig. 2. (a) A frame in HOTEL dataset [34]. The orange dots represent the labeled pedestrians. (b) The corresponding simulated scenario for (a) using a top-down view. (c) The focus pedestrian's first-person view image (i.e., red dot in (a)). (d) The camera settings for pedestrians in FvSim.

3.2 Model Overview

FvSim–using Unity–extends given 2D pedestrian trajectories into a 3D simulated scene, from which we can obtain multi-view information, especially first-person views for each pedestrian (Fig. 2(c)). FvSim is proportional to the real-world environment using SI units. The input of FvSim is ground-level 2D trajectory data from a given dataset, which is either presented in or converted to our coordinate system defined in our simulated environment. We prepare 27 3D human models with walking behavior embedded. It enables FvSim to randomly assign a prepared human model to each pedestrian from a given dataset. Since the body, head, and gaze orientations are necessary required information for FvSim but not accessible from the original datasets (e.g. ETH [34] and UCY [26]), we assume they are aligned with the focus pedestrian's forward direction (i.e., the direction of the computed velocity using 2D trajectories). FvSim assumes the height of each pedestrian is 1.75 m, and the first-person view is provided via a camera with a 144° wide-angle [14] and an optical axis parallel to the ground plane, which is rigidly mounted on each pedestrian's head 1.63 m above the ground (Fig. 2(d)).

As illustrated in Fig. 3, FvTraj is composed of five modules: (1) a Traj-Encoder (Section 3.3), a trajectory encoder that captures historical motion patterns of each pedestrian; (2) a View-Encoder (Section 3.4), an encoder module that extracts visual features from the simulated first-person view image sequence; (3) a Social-aware attention module (Section 3.5) that builds relations with other socially interacted pedestrians in the scene; (4) a View-aware attention module (Section 3.6) that captures the latent relations between motions and visual features (i.e., extracted from the first-person view images with ego-motion information) using an attention mechanism; (5) a Traj-Decoder (Section 3.7) that generates multimodal pedestrian trajectories given all observed information including pedestrians' historical trajectories, social interactions with other pedestrians in the scene, and the dynamic scene contexts from the first-person view images.

3.3 Trajectory Encoder

We build Traj-Encoder, an encoder for any given pedestrian $p^i (i \in [1, N])$ in a scene to capture the historical motion patterns. Given the observed trajectory $X^i_{t \in [1, T_{\text{obs}}]}$, we calculate its relative displacements $\Delta X^i_t = X^i_t - X^i_{t-1}$. Then, ΔX^i_t

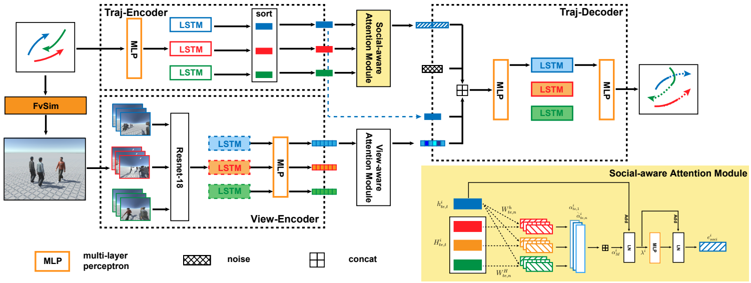

Fig. 3. Pipeline overview of the FvTraj. Given the pedestrian trajectories in the observation period, we use FvSim to simulate the corresponding crowd scenario and render the first-person view images for each pedestrian. The Traj-Encoder and the View-Encoder are used to extract latent representations for observed pedestrian trajectories, and the first-person view images, respectively. Their outputs are fed into the follow-up social-aware attention and view-aware attention modules to capture social behaviors and visual view-aware features based on multi-head attention mechanisms. The final multimodal trajectories are produced by the Traj-Decoder. Social-aware attention module (with yellow background). It captures the latent social interactions between the focus pedestrian p^i and the other pedestrians. In this module, we use a multi-head attention to calculate the scaled dot product attentions over all other pedestrians except p^i for each of the four heads (i.e., $n = 4$). Followed by an MLP, the social-aware attention module finally returns the social interaction representations e^i_{soci}.

is embedded into a high dimensional space using a multi-layer perceptron (MLP) and then fed into an LSTM (TE-LSTM) as follows:

$$h^i_{te,t} = \text{TE-LSTM}\left(h^i_{te,t-1}, \text{MLP}(\mathit{\Delta} X^i_t, W^{emb}_{rel}); W_{te}\right), \tag{1}$$

where $h^i_{te,t}$ is the hidden states of TE-LSTM, which carries latent representations of historical motion states of p^i, W^{emb}_{rel} denotes the embedding weights of MLP, and W_{te} denotes the LSTM weights in the trajectory encoder TE-LSTM. In our model, all the pedestrians in the scene share the same parameter values in TE-LSTM.

3.4 View Image Encoder

We build View-Encoder, an encoder for any given first-person view image I^i_t of the corresponding pedestrian p^i. The simulated first-person view images with the original size of 768×1024 are resized to 36×48 for FvTraj. We use a ResNet-18 model [15] pre-trained on ImageNet [10] and fine tune the model to extract visual features, which denoted as V^i. We then pass these visual features V^i into an LSTM (IE-LSTM) as $h^i_{ie,t} = \text{IE-LSTM}\left(h^i_{ie,t-1}, V^i_t; W_{ie}\right)$, where $h^i_{ie,t}$ denotes the hidden states of IE-LSTM, W_{ie} denotes the LSTM weights in the IE-LSTM. Then we feed $h^i_{ie,t}$ to an MLP with embedding weights W^{emb}_{ie} to get the visual feature $\hat{h}^i_{ie,t}$.

3.5 Social-aware Attention Module

Since the pedestrians in a scene often socially interact with each other, modeling social interactions among the pedestrians is important to the realism of real-world crowds, besides the purpose of collision avoidance. We build a social-aware attention module (Fig. 3) based on a multi-head attention mechanism [42] to learn latent social interactions between a focus pedestrian and all other pedestrians in the scene. Inspired by the REFER module [20], which can learn latent relationships between a given question and a dialog history in the visual dialog task and reach the state-of-the-art performance, we design a similar structure for this module.

Similar to the prior works [11, 37, 2], we sort the order of the pedestrians other than the focus pedestrian based on their relative distances between the focus pedestrian and themselves. We denote the concatenated hidden states (i.e., which are calculated in the trajectory encoder) of these sorted pedestrians as $H^i_{te,t}$, which carry the latent representations of historical motion patterns.

To capture how the sorted pedestrians influence the future trajectories of the focus pedestrian, we use the scaled dot product attention [42] to obtain the interactions between the focus pedestrian p^i and the others as follows:

$$\alpha^i_{te,n} = Attn((h^i_{te,t}W^h_{te,n}), (H^i_{te,t}W^H_{te,n})), \quad Attn(a, b) = \text{softmax}(\frac{ab^T}{\sqrt{d_{te}}})b, \quad (2)$$

where $W^h_{te,n}$ and $W^H_{te,n}$ are the linear weights to transform the hidden states into d_{te} dimensions, respectively.

To stabilize the learning process, we operate a multi-head attention mechanism [42] by calculating the attention n times with distinct $W^h_{te,n}$ and $W^H_{te,n}$ using Eq. 2, yielding $\alpha^i_{te,1}, ..., \alpha^i_{te,n}$. The multi-head representations are concatenated as α^i_M, followed by a linear function as $\alpha^i_M = \alpha^i_{te,1} \oplus ... \oplus \alpha^i_{te,n}$, where $\oplus$ is a concatenation operation. Note that α^i_M is then passed into another linear function with weights W^M_{te}. To add their hidden states $h^i_{te,t}$, we apply a residual connection [15] and employ layer normalization (LN) [4] as $\lambda^i = \text{LN}(\alpha^i_M W^M_{te} + h^i_{te,t})$.

To obtain the social interaction representations e^i_{soci} for the pedestrian p^i in the scene, we again adopt an MLP with weights W_{soci}, followed by the other residual connection and LN as $e^i_{soci} = \text{LN}(\text{MLP}(\lambda^i, W_{soci}) + \lambda^i)$.

3.6 View-aware Attention Module

We also build a view-aware attention module to extract visual features from the first-person view images in the observation period, which adopts the module structure from the previous social-aware attention module (Section 3.5). Similarly, we exploit the multi-head attention mechanism to concatenate information, followed by residual connection and LN to obtain the relationships between a given latent motion pattern representation and the historical latent visual features extracted from the first-person view images, denoted as e^i_{view}. Note that only the structure of the social-aware attention module and that of the view-aware attention module are the same, the parameter values in the two modules are not shared and could be different.

The input of first-person view information is not shared between pedestrians in opposition to the historical trajectory information. Considering the visually dynamic and continuous scene context of pedestrians, we denote $\hat{h}^i_{ie,t}$ with the latent representations of visual features of p^i from view image encoder at $t \in [1, T_{\text{obs}}]$ as H^i_{ie}. The shape of H^i_{ie} is $T_{\text{obs}} \times d_{ie}$, where d_{ie} is the dimension of $\hat{h}^i_{ie,t}$. Based on the hidden states $h^i_{te,t}$ of p^i in the trajectory encoder, we can obtain the scaled dot product attention as $\alpha^i_{ie,n} = Attn((h^i_{te,t}\hat{W}^h_{ie,n}), (H^i_{ie}W^H_{ie,n}))$, where $\hat{W}^h_{ie,n}$ and $W^H_{ie,n}$ are linear weights to transform the hidden states into d_{ie} dimensions, respectively.

3.7 Trajectory Decoder

We build the Traj-Decoder, a trajectory decoder that generates future trajectories for each pedestrian in a scene. To mimic the actual motions of pedestrians, we consider their major inherent properties: multimodality, self historical motion patterns, social interactions with other pedestrians, and scene contexts.

Traj-Decoder utilizes an LSTM decoder (TD-LSTM), inspired by the previous works [13, 37, 17] that exploit a noise vector z sampled from a multivariate normal distribution to produce multimodal future trajectories. We use the concatenation of four components: (1) the latent representation of the motion patterns in the observation period from the last step of LSTM trajectory encoder $h^i_{te,T_{\text{obs}}}$, (2) the embedding of social interactions between the focus pedestrian and the other pedestrians e^i_{soci}, (3) the captured view-aware representation e^i_{view} from the first-person view images with scene contexts, and (4) the sampled noise vector z. The output of this concatenation is then passed through an MLP with weights W^{emb}_{td} to initialize the hidden states of the LSTM decoder. Based on the Seq2seq framework [41], the latter process can be represented as:

$$h^i_{td,T_{\text{obs}}+1} = \text{MLP}(h^i_{te,T_{\text{obs}}} \oplus e^i_{soci} \oplus e^i_{view} \oplus z, W^{emb}_{td}), \tag{3}$$

where $h^i_{td,T_{\text{obs}}+1}$ denotes the initialized hidden states of TD-LSTM.

The recursion equation of the Traj-Decoder for p^i in the prediction period is:

$$h^i_{td,t} = \text{TD-LSTM}(h^i_{td,t-1}, \text{MLP}(\Delta\hat{X}^i_t, W^{emb}_{rel}); W_{td}), \tag{4}$$

where $\Delta\hat{X}^i_t$ is the relative positions based on the predicted results at the last step. Note that $\Delta\hat{X}^i_t$ at the first step $T_{\text{obs}} + 1$ of the prediction period is the same as the last input of TE-LSTM at step T_{obs}. The MLP with weights W^{emb}_{rel} shares the parameters with MLP in Eq. 1. W_{td} and $h^i_{td,t}$ are the LSTM weights and hidden states in TD-LSTM, respectively.

Lastly, we pass the hidden states $h^i_{td,t}$ in TE-LSTM into another MLP with weights W_d one at a time to calculate the relative positions $\Delta\hat{X}^i_t$ in the prediction period, and we obtain the predicted positions based on $\Delta\hat{X}^i_t$ and the last 2D positions, represented as $\Delta\hat{X}^i_t = \text{MLP}(h^i_{td,t}, W_d)$, and $\hat{X}^i_t = \Delta\hat{X}^i_t + \hat{X}^i_{t-1}$.

3.8 Training and Implementation Details

Losses. The entire network is trained end-to-end by minimizing the L2 loss ($L = ||\Delta X_t^i - \Delta \hat{X}_t^i||_2$), which is the difference between the predicted trajectories in the prediction period and the ground-truth trajectories [13, 37, 2, 17]. Based on a noise vector z, FvTraj can produce multimodal trajectories. We adopt a similar training process, following the variety loss in the previous works [13, 37, 2, 17]. For each training step, we generate k possible trajectories according to the randomly sampled z, and then choose the best result as the prediction.

Implementation Details. In the Traj-Encoder, the 2D position of each pedestrian is embedded into a vector of 32 dimensions, and followed by LSTMs with 64 hidden states. In the View-Encoder, the first-person view images at each step in the observation period are processed into 1000 dimensions using ResNet-18, and followed by LSTMs with 128 hidden states. The output of the LSTMs in View-Encoder is further processed by a two-layer MLP ($128 \times 64 \times 64$) with ReLU activation functions. In the social-aware attention module, the number of multi-head attention is $n = 4$. $h_{te,t}^i$ and $H_{te,t}^i$ are projected into 16 dimensions. The MLP for λ^i comprises 2-layer 1D convolution operations with ReLU activation functions. In the view-aware attention module, the parameters have the same dimensions as those in the social-aware attention module. In the Traj-Decoder, we add a 32 dimension noise vector. The concatenation of $h_{te,T_{\text{obs}}}^i \oplus e_{soci}^i \oplus e_{view}^i \oplus z$ is fed into a 3-layer MLP ($224 \times 192 \times 128 \times 64$), with ReLU functions and batch normalizations. The hidden states of the LSTM in the Traj-Decoder is fixed to 64 dimensions. The $h_{td,t}^i$ with 64 dimensions will finally transformed into 2D relative positions. The initial learning rate is set to 0.001 and decayed into 0.0001 after 20 epochs. The learning process adopts Adam optimizer to iteratively update the network with a batch size 8 for 500 epochs.

4 Experiment Results

We compared FvTraj with state-of-the-art pedestrian trajectory prediction models, and presented quantitative and qualitative evaluation in this section. We used two relevant and publicly accessible datasets: ETH [34] and UCY [26]. The ETH dataset comprises two distinct scenes: ETH and HOTEL, and the UCY dataset comprises three distinct scenes: ZARA1, ZARA2, and UCY. We used the data preprocessing method proposed in S-GAN [13], and the corrected ETH-Univ frame rate presented in the work [49]. Following a similar approach as in the prior works [13, 37, 22], we used a leave-one-out method to use four scenes as the training data and the remaining one scene as the test data. In our experiments, the pedestrian trajectories for the initial eight steps (i.e., the observation period on a timescale of 3.2 s) are given, and we aim to predict trajectories for the next 12 steps (i.e., the prediction period on a timescale of 4.8 s).

Baselines. We compared FvTraj to four state-of-the-art pedestrian trajectory prediction models, including two models without scene contexts: Social-GAN (**S-GAN**) [13] and a spatial-temporal graph attention network (**STGAT**)

Table 1. Quantitative results for the predicted positions. We use ADE and FDE in meters to evaluate the task of predicting the trajectories within a period of 12 steps (4.8 s), given the previous observed 8 steps (3.2 s). The lower evaluation is the better.

	Without Scene Contexts		With Scene Contexts		Ours			
Dataset	**S-GAN** [13] 20-20	**STGAT** [17] 20-20	**Sophie** [37] 20-20	**Bi-GAT** [22] 20-20	**FvTraj** 1-1	**FvTraj-noSocial** 5-20	**FvTraj-noView** 5-20	**FvTraj** 5-20
ETH	0.87 / 1.62	0.65 / 1.12	0.70 / 1.43	0.69 / 1.29	0.62 / 1.23	0.60 / 1.22	0.58 / 1.21	**0.56 / 1.14**
HOTEL	0.67 / 1.37	0.35 / 0.66	0.76 / 1.67	0.49 / 1.01	0.53 / 1.10	0.34 / 0.70	0.42 / 0.89	**0.28 / 0.55**
UNIV	0.76 / 1.52	**0.52 / 1.10**	0.54 / 1.24	0.55 / 1.32	0.57 / 1.19	0.55 / 1.16	0.56 / 1.16	**0.52 / 1.12**
ZARA1	0.35 / 0.68	0.34 / 0.69	**0.30 / 0.63**	**0.30 / 0.62**	0.42 / 0.89	0.39 / 0.80	0.37 / 0.78	0.37 / 0.78
ZARA2	0.42 / 0.84	**0.29 / 0.60**	0.38 / 0.78	0.36 / 0.75	0.38 / 0.79	0.35 / 0.69	0.33 / 0.67	0.32 / 0.68
Average	0.61 / 1.21	0.43 / 0.83	0.54 / 1.15	0.48 / 1.00	0.50 / 1.04	0.45 / 0.91	0.45 / 0.94	**0.41 / 0.85**

[17], and the other two with scene contexts: the social GAN with attention networks (**Sophie**) [37] and the Bicycle-GAN with graph attention networks (**Social-BiGAT**) [22]. The results of STGAT are obtained by our implementations of the ADE and FDE metrics and evaluation of the trained models that are released by the authors. The results of S-GAN, Sophie, and Social-BiGAT are obtained from the original papers [13, 37, 22].

Evaluation Metrics. Inspired by prior works [13, 37, 22], we chose Average Displacement Error (**ADE**) and Final Displacement Error (**FDE**) as the evaluation metrics. ADE is the average Euclidean distance error between the predicted result and the ground truth over the whole sequence. FDE is the Euclidean distance error at the last step between the predicted result and the ground truth.

Ablation Study. We performed an ablation study using various control settings to evaluate the contribution of each major component of our model. FvTraj is our final model with all the components; FvTraj-noSocial is a version of our model without the social-aware attention module; FvTraj-noView is a version of our model without both the view image encoder and the view-aware attention module. The model with N-K variety loss represents that the model with the lowest ADE and FDE selected from K randomly sampled trajectories after N times training, which is similar to the prior works [13, 37, 22].

4.1 Quantitative Evaluation

The quantitative comparison results between our model to the baseline models are reported in Table. 1, which include ADE and FDE for the predicted trajectories within the prediction period of 12 steps given the observed eight steps.

For ETH, HOTEL, and UNIV, the baseline models with scene contexts (i.e., Sophie and Bi-GAT) outperformed S-GAN but not STGAT. FvTraj outperformed both S-GAN and STGAT, except the performance of FvTraj and STGAT on UNIV are similar. These results suggest that the contribution of the top-down view images used in the baselines is not as obvious as that of the first-person view images used in FvTraj. ETH, HOTEL, and UNIV can be characterized as a spacious environment with few stationary obstacles such that the main obstacles in the scene are the moving pedestrians. It is our conjecture that the success of FvTraj in these scenes is due to that the first-person view images can better capture the detailed motion of each pedestrian, especially the ego-motions.

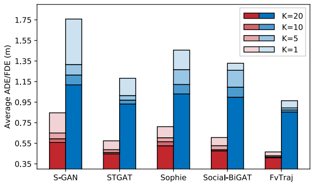

Fig. 4. Quantitative results for the predicted positions demonstrate the effect of the variety loss. Reducing K results in a higher average ADE/FDE across all five scenes, and less change means better generalization. Note, we use $N = 5$ for FvTraj, and $N = 20$ for S-GAN, STGAT, Sophie, and Social-BiGAT.

For ZARA1, the baseline models with scene contexts (i.e., Sophie and Bi-GAT) outperformed both S-GAN and STGAT, but FvTraj outperformed neither S-GAN nor STGAT. For ZARA2, both Sophie and Bi-GAT outperformed S-GAN but not STGAT, so does the FvTraj model. The results seem to suggest that the contribution of the top-down view images might be more than that of the first-person view images used in the FvTraj model. ZARA1 and ZARA2 can be characterized as the environment with some large-scale stationary obstacles such that the pedestrians' motions would be limited. It is possible that the performance of the FvTraj model is no better than the baseline models in these cases, due to the lacking of stationary scene contexts in our simulated first-person view images. By considering the performance of baseline models, it is reasonable to believe the performance of FvTraj can be further improved by introducing the simulated stationary obstacles in our FvSim.

In terms of our proposed module-based architecture, we found that incorporating both the social-aware attention module and the view-aware attention module can significantly improve the performances on ETH, HOTEL, and UNIV. However, we cannot find noticeable differences between FvTraj-noView and FvTraj for ZARA1 and ZARA2. This might be caused by the same reason described above, which is the lacking of stationary scene contexts in our simulated images.

The baseline models were evaluated using 20-20 variety loss, and FvTraj was evaluated using 5-20 variety loss. We choose K for FvTraj to be consistent with the four baseline models. We chose the reduced $N = 5$ due to the computational complexity of FvTraj, which contains the computationally intensive architecture of networks and the pre-processing procedure of first-person view images sequence for each pedestrian in the scene. It is reasonable to believe the performance of FvTraj can be further improved if we increase $N = 5$ to $N = 20$.

Figure 4 shows the effect of varying K from $K = 20$ to $K = 1$ when evaluating the generalization of each model. Although we found that the increase of K generally leads to better accuracy in terms of ADE and FDE for all the five models, the effect of varying K on the performance of FvTraj is not significant compared to the four baseline models. The average ADE and FDE of our model with 5-1 various loss are 0.47 and 0.96, respectively. When K is increased to 20, the average ADE and FDE of our model decrease to 0.41 and 0.85, respectively,

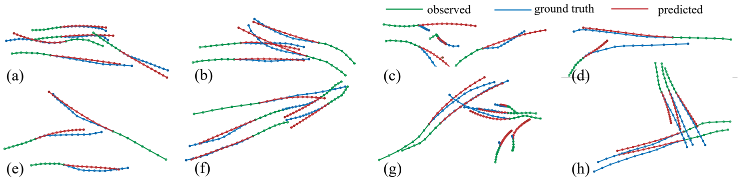

Fig. 5. Visual comparisons between the ground truth and the predicted trajectories by FvTraj across eight scenes. Each scene shows at least one of the scenarios among pedestrians: individual following, group following, meeting, and collision avoiding.

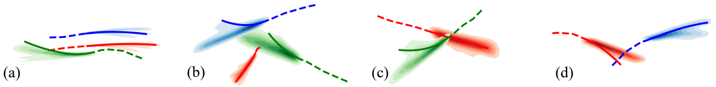

Fig. 6. Visualization of the predicted trajectory distributions ($K = 20$) and the final trajectories. The trajectories in the observation period and the prediction period are illustrated as solid and dash lines, respectively.

which leads to a performance increase of 13.7% and 12.9%, respectively. This result indicates across all five scenes, on average, drawing more samples from our model does not cause a significant increase in accuracy. Therefore, our FvTraj is more robust and better generalized than all the baseline models.

4.2 Qualitative Evaluation

To better evaluate the performance of FvTraj, we visualize the predicted trajectories (Fig. 5) across eight scenes given the observed trajectories, compared to the ground truth. We are aware that the multimodality of pedestrians might be caused by scene contexts, self intentions, destinations, etc. Although the predicted trajectories in Fig. 5(c), (d), (g), and (h) seems do not to agree to the ground truth, they might still be reasonable and safe for the pedestrians. These predicted trajectories are more conservative in terms of safety, especially in meeting scenes to avoid potential and future collisions. This is of particular interest for some specific applications such as robotic navigation and blind navigation.

Fig. 6 shows the predicted trajectory distribution in various scenes. Figs. 6 (a) and (b) describe two meeting scenes with three pedestrians; Figs. 6(c) and (d) describe two meeting scenes with two pedestrians. We observe that: (1) the directions of the potential trajectories of a single pedestrian could be far apart (i.e., the pedestrian colored in green in Fig. 6(c), (2) the trajectories of neighboring pedestrians have been well considered, (3) pedestrian collision is unlikely to occur due to the inexistence of overlapping among the predicted trajectory distributions at any step, and (4) the variance of the predicted trajectory distribution is reduced with the increased probability of collisions occurring between pedestrians. Since these observed scenarios are likely to happen in real world, which suggests that FvTraj can well capture the fundamental factors including multimodality, social interactions, and scene contexts.

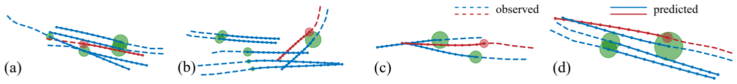

Fig. 7. Visualization of the predicted attention weights by FvTraj. Here, we visualize the average attention weights (green circles) of the four head attentions used in the social-aware attention module at T_{obs}. Note, the green circles' radii are proportional to the attention weights, the red circles represent the position of the focus pedestrian at T_{obs}, the red trajectories represent the focus pedestrian whose attention weights are predicted, and the blue trajectories represent the other pedestrians in the scene.

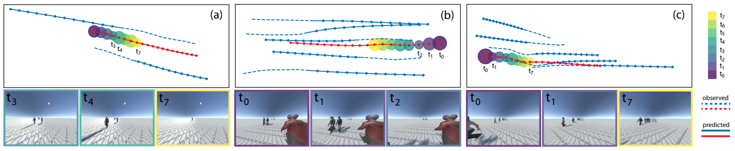

Fig. 8. Visualization of learned attention weights by our view-aware attention module of 8 steps in observation period and simulated first-person view images by FvSim for 3 corresponding steps. Each scene shows at least one of the scenarios among pedestrians: meeting, following, and collision avoiding. The colored circles' radii are proportional to the attention weights, the red trajectories represent the focus pedestrian whose attention weights are predicted, and the blue trajectories show the other pedestrians.

We visualize the learned attention weights (Fig. 7) using the social-aware attention module. We observe that social module assigns higher attention weights to the pedestrians: (1) who have relative small Euclidean distances from the focus pedestrians, (2) who move toward the focus pedestrians, and (3) whose observed trajectories are close to the focus pedestrians' observed trajectory. These observations implicitly address our safety concerns, which implies the social-aware attention module can well capture the social interactions within a scene.

The Contribution of First-person View Information. Table 1 shows the comparison between our full-model FvTraj and FvTraj-noView, which is a model without first-person view information. Adding first-person view information to FvTraj leads to performance increases of 9.8% and 10.6% for average ADE and FDE, respectively. Fig. 8 shows simulated first-person view images can well capture ego-motions for the focus pedestrians using learned attention weights by view-aware attention module. Ego-motions (i.e, the focus pedestrian's visual perspective effect on the neighbors and moving intentions) are important in trajectory prediction, which is difficult to capture using third-person view images or social-aware module (focusing on capturing historical motion and social patterns learned from numerical inputs). Combining the view-aware and the social-aware modules, FvTraj can well capture all these important features.

Failure Cases. Figure 9 shows the visual comparisons between the ground truth and the predicted trajectory using FvTraj for ZARA1. Although the differences between our predicted trajectories and the ground truth are not sig-

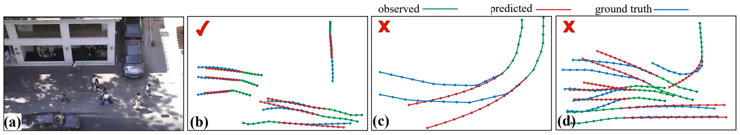

Fig. 9. Visualization of the predicted pedestrian trajectories using FvTraj for ZARA1. FvTraj is not optimized for the scene packed with large-scaled stationary obstacles, such as (a) ZARA1, shown in the top-down view. Here, (b) FvTraj can successfully predict the trajectories, (c) and (d) but sometimes may fail, especially for the cases that the pedestrians' intention changed dramatically to avoid the obstacles.

nificantly noticeable for most of the cases (Fig. 9(b)), the differences in some specific cases (Fig. 9(c) and (d)) are noticeable. We observe that the effect of the stationary obstacles (e.g., buildings, parked vehicles) on pedestrians' trajectories in ZARA1 cannot be neglected. It seems that pedestrians intentionally maintain a relatively large distance from the stationary obstacles, which are not well captured in FvTraj. These qualitative results for ZARA1 are consistent with the qualitative results described in Section 4.1. These results suggest more work is required to understand the relationship between the stationary obstacles and the dynamic scene contexts, which motivates us to develop an advanced FvSim in our future work. Although current FvSim without any scene context may cause failure cases, simulation without scene contexts are universal and can be applicable to any scenes without any scene-related constraint.

5 Conclusion

This work presents a novel first-person view based trajectory prediction model, FvTraj. To obtain the first-person view information in an efficient way, we develop a simulator, FvSim, to generate a 3D simulated scenario with multi-view information including the first-person view, given the observed 2D trajectories. FvTraj takes into account historical motion patterns of pedestrians, social interactions, and the first-person view scene contexts, based on multi-head attention mechanisms to predict realistic and plausible trajectories. Our experimental results suggested that: (1) the first-person view information successfully introduces detailed dynamic scene contexts with ego-motions, (2) FvTraj is well structured for the pedestrian trajectory prediction task, and (3) FvTraj achieves state-of-the-art performance via comparisons with baseline models.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103000, 2017YFC0804900, and 2018YFB1700905, in part by the National Natural Science Foundation of China under Grant 61532002, 61972379, and 61702482. Zhigang Deng was in part supported by US NSF grant IIS-1524782.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2016)
2. Amirian, J., Hayet, J.B., Pettré, J.: Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In: Proc. IEEE Conf. Computer Vision Pattern Recognition Workshops (CVPRW) (2019)
3. Antonini, G., Bierlaire, M., Weber, M.: Discrete choice models of pedestrian walking behavior. Transp. Res. Part B: Methodological **40**(8), 667–687 (2006)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
5. Bagautdinov, T., Alahi, A., Fleuret, F., Fua, P., Savarese, S.: Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). pp. 4315–4324 (2017)
6. Bertasius, G., Chan, A., Shi, J.: Egocentric basketball motion planning from a single first-person image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
7. Bi, H., Fang, Z., Mao, T., Wang, Z., Deng, Z.: Joint prediction for kinematic trajectories in vehicle-pedestrian-mixed scenes. In: Proc. IEEE Int. Conf. Computer Vision (ICCV) (2019)
8. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
9. Choi, C., Dariush, B.: Looking to relations for future trajectory forecast. In: Proc. IEEE Int. Conf. Computer Vision (ICCV) (2019)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, F.F.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA (2009)
11. Felsen, P., Lucey, P., Ganguly, S.: Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders. In: The European Conference on Computer Vision (ECCV) (September 2018)
12. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)
13. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2018)
14. Hasan, I., Setti, F., Tsesmelis, T., Del Bue, A., Galasso, F., Cristani, M.: Mx-lstm: Mixing tracklets and vislets to jointly forecast trajectories and head poses. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
16. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Physical Rev. E **51**(5), 4282 (1995)
17. Huang, Y., Bi, H., Li, Z., Mao, T., Wang, Z.: Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In: Proc. IEEE Int. Conf. Computer Vision (ICCV) (2019)

18. Ivanovic, B., Pavone, M.: The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: Proc. IEEE Int. Conf. Computer Vision (ICCV) (2019)
19. Johnson, L.A., Higgins, C.M.: A navigation aid for the blind using tactile-visual sensory substitution. In: Proc. Int. Conf. IEEE Eng. Medicine Biol. Soc. pp. 6289–6292 (2006)
20. Kang, G., Lim, J., Zhang, B.: Dual attention networks for visual reference resolution in visual dialog. CoRR **abs/1902.09368** (2019), http://arxiv.org/abs/1902.09368
21. Kantorovitch, J., Väre, J., Pehkonen, V., Laikari, A., Seppälä, H.: An assistive household robot–doing more than just cleaning. J. Assistive Technol. **8**(2), 64–76 (2014)
22. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, S.H., Savarese, S.: Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. arXiv preprint arXiv:1907.03395 (2019)
23. Lai, G.Y., Chen, K.H., Liang, B.J.: People trajectory forecasting and collision avoidance in first-person viewpoint. In: 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW). pp. 1–2. IEEE (2018)
24. Lee, J.G., Han, J., Whang, K.Y.: Trajectory clustering: a partition-and-group framework. In: Proc. ACM SIGMOD Int. Conf. Manage. of Data. pp. 593–604 (2007)
25. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H.S., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2017)
26. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Proc. Computer Graphics Forum. vol. 26, pp. 655–664 (2007)
27. Leung, T.S., Medioni, G.: Visual navigation aid for the blind in dynamic environments. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops. pp. 565–572 (2014)
28. Liang, J., Jiang, L., Carlos Niebles, J., Hauptmann, A.G., Fei-Fei, L.: Peeking into the future: Predicting future person activities and locations in videos. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2019)
29. Ma, Y., Zhu, X., Zhang, S., Yang, R., Wang, W., Manocha, D.: Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In: Proc. AAAI Conf. Artif. Intell. pp. 6120–6127 (2019)
30. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017)
31. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). pp. 935–942 (2009)
32. Morris, B., Trivedi, M.: Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). pp. 312–319 (2009)
33. Patil, A., Malla, S., Gang, H., Chen, Y.: The H3D dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. CoRR **abs/1903.01568** (2019), http://arxiv.org/abs/1903.01568
34. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: Proc. IEEE Int. Conf. Computer Vision (ICCV). pp. 261–268 (2009)

35. Ramanishka, V., Chen, Y.T., Misu, T., Saenko, K.: Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7699–7707 (2018)
36. Rios-Martinez, J., Spalanzani, A., Laugier, C.: From proxemics theory to socially-aware navigation: A survey. Int. J. Soc. Robot. **7**(2), 137–153 (2015)
37. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2019)
38. Sadeghian, A., Legros, F., Voisin, M., Vesel, R., Alahi, A., Savarese, S.: Car-net: Clairvoyant attentive recurrent network. In: Proc. Eur. Conf. Computer Vision (ECCV). pp. 151–167 (2018)
39. Song, X., Wang, P., Zhou, D., Zhu, R., Guan, C., Dai, Y., Su, H., Li, H., Yang, R.: Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. CoRR **abs/1811.12222** (2018), http://arxiv.org/abs/1811.12222
40. Soo Park, H., Hwang, J.J., Niu, Y., Shi, J.: Egocentric future localization. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
41. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. CoRR **abs/1409.3215** (2014), http://arxiv.org/abs/1409.3215
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762
43. Vemula, A., Muelling, K., Oh, J.: Social attention: Modeling attention in human crowds. In: Proc. IEEE Int. Conf. Robot. and Automat. (ICRA). pp. 1–7 (2018)
44. Wang, X., Ma, K.T., Ng, G.W., Grimson, W.E.L.: Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. Int. J. Computer Vision **95**(3), 287–312 (2011)
45. Wang, X., Ma, X., Grimson, W.E.L.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Trans. Pattern Anal. and Mach. Intell. **31**(3), 539–555 (2009)
46. Xu, Y., Piao, Z., Gao, S.: Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2018)
47. Yagi, T., Mangalam, K., Yonetani, R., Sato, Y.: Future person localization in first-person videos. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2018)
48. Yao, Y., Xu, M., Wang, Y., Crandall, D.J., Atkins, E.M.: Unsupervised traffic accident detection in first-person videos. CoRR **abs/1903.00618** (2019), http://arxiv.org/abs/1903.00618
49. Zhang, P., Ouyang, W., Zhang, P., Xue, J., Zheng, N.: Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2019)