Holistic LSTM for Pedestrian Trajectory Prediction

Ruijie Quan, Linchao Zhu, *Member, IEEE*,
Yu Wu, *Student Member, IEEE*, and Yi Yang, *Senior Member, IEEE*

***Abstract*—Accurate predictions of future pedestrian trajectory could prevent a considerable number of traffic injuries and improve pedestrian safety. It involves multiple sources of information and real-time interactions, *e.g.*, vehicle speed and ego-motion, pedestrian intention and historical locations. Existing methods directly apply a simple concatenation operation to combine multiple cues while their dynamics over time are less studied. In this paper, we propose a novel Long Short-Term Memory (LSTM), namely Holistic LSTM, to incorporate multiple sources of information from pedestrians and vehicles adaptively. Different from LSTM, our Holistic LSTM considers mutual interactions and explores intrinsic relations among multiple cues. First, we introduce extra memory cells to improve the transferability of LSTMs in modeling future variations. These extra memory cells include a speed cell to explicitly model vehicle speed dynamics, an intention cell to dynamically analyze pedestrian crossing intentions and a correlation cell to exploit correlations among temporal frames. These three individual cells uncover the future movement of vehicles, pedestrians and global scenes. Second, we propose a gated shifting operation to learn the movement of pedestrians. The intention of crossing the road or not would significantly affect pedestrian's spatial locations. To this end, global scene dynamics and pedestrian intention information are leveraged to model the spatial shifts. Third, we integrate the speed variations to the output gate and dynamically reweight the output channels via the scaling of vehicle speed. The movement of the vehicle would alter the scale of the predicted pedestrian bounding box: as the vehicle gets closer to the pedestrian, the bounding box is enlarging. Our rescaling process captures the relative movement and updates the size of pedestrian bounding boxes accordingly. Experiments conducted on three pedestrian trajectory forecasting benchmarks show that our Holistic LSTM achieves state-of-the-art performance.**

***Index Terms*—Pedestrian Trajectory Prediction, Long Short-Term Memory, Holistic LSTM, Pedestrian Intention.**

I. Introduction

RECENTLY, significant progress has been made in autonomous driving systems [1], [2]. With the rapid development of computer vision, autonomous driving systems are successful in detecting and recognizing roads [3], cars [4], [5], pedestrians [6], [7], [8], etc. It is also essential to anticipate the near future to guarantee the safety of pedestrians, especially for the future trajectories of pedestrians at crossings. Pedestrian trajectory prediction can prevent a considerable number of traffic injuries as it enables more reaction time to take action. Despite its significance, less attention has been paid, and it is challenging to anticipate the trajectories of pedestrians since they exhibit highly variable behavior patterns [9].

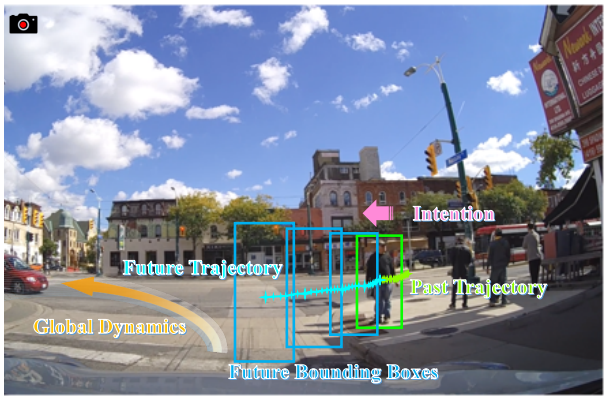


Fig. 1: Illustration of pedestrian trajectory prediction on an on-board camera setting. "Global Dynamics" indicates the motion of view. The ground truth bounding box of the pedestrian intending to cross the street is shown in green and the predicted future bounding boxes are in blue. We aim to predict the pedestrian's future trajectory based on his trajectory history, crossing intention and the global scene dynamics.

In this paper, we study the problem of forecasting pedestrian's future trajectory in a first-person view, which depends on the visual observations of the pedestrian's motion history. The trajectory forecasting is compounded by the presence of the relative motion and social interactions between humans and vehicles. An example of this task is illustrated in Figure. 1. The pedestrian in the green box has been standing by the road in the past few seconds. Our target is to predict his trajectory in next few seconds given the observations. Information such as the movement of vehicles, pedestrian crossing intention and global scene dynamics in observations contribute to the accurate trajectory prediction since there are complex interactions and intrinsic relations between vehicles and pedestrians.

Most existing research works on pedestrian trajectory prediction focus on a bird-eye view instead of a first-person view. They cannot be applied in an autonomous driving setting. A recent work [10] forecasts the future trajectory from an on-board camera view, but it only concerns the simple motion learning of pedestrians and vehicle dynamics. These motions are fused with other cues like speed information via a straightforward concatenation operation. With the concatenated information, it leverages a standard Long Short-Term Memory (LSTM) [11] to directly make predictions for pedestrian future trajectory. Although the typical LSTM has been proved to be efficient and successful in trajectory prediction tasks [12], [13], it contains only a single memory cell which limits the representation

This work was done when R. Quan and Y. Wu interned at Baidu Research.
R. Quan and Y. Wu are with Baidu Research, Beijing, China and the ReLER Lab, Australian Artificial Intelligence Institute, University of Technology Sydney, Ultimo, NSW 2007, Australia. E-mail: ruijie.quan@student.uts.edu.au; yu.wu-3@student.uts.edu.au. L. Zhu and Y. Yang are with the ReLER Lab, Australian Artificial Intelligence Institute, University of Technology Sydney, Ultimo, NSW 2007, Australia. E-mail: linchao.zhu@uts.edu.au; yi.yang@uts.edu.au. Linchao Zhu is the corresponding author.

capability. Therefore, it is not able to handle complex motion information, such as interactions between pedestrians and vehicles or the environment. As a result, this simple framework overlooks the intrinsic relations among different cues, *e.g.*, vehicle speed, pedestrian intention, and the distance between vehicle and pedestrian. However, these intrinsic relations and mutual interactions are critical in reasoning and predicting the pedestrian's future trajectory.

We propose a novel Holistic Long Short-Term Memory (Holistic LSTM) to introduce interactions among different information cues and adaptively incorporate multiple sources of motion information from pedestrians, vehicles and global scenes. In our Holistic LSTM unit, memory cells model the inter-related dynamics of pedestrian intention, vehicle speed and global scene dynamics among the temporal frame sequences. Excluding a typical memory cell that maintains the state of input data, a pedestrian intention cell is introduced to analyze pedestrian intentions that change over time, as intentions of crossing the road directly affect pedestrian's future trajectory. Besides, we propose a speed cell to model vehicle speed dynamics as there are complex interactions between the vehicle and the target pedestrian. For example, a car at high speed would change the man's mind to cross the road, resulting in changes in his future trajectory. Moreover, we leverage a correlation cell to learn the motion dynamics of global scenes. The correlation, which has a strong relation to the changes in pedestrian's trajectory, has both direction and magnitude, whereas ego-speed cue only has magnitude. Besides, past motion dynamics are able to help forecast future dynamics, thereby further improving the accuracy of the trajectory prediction. Multiple extra memory cells significantly improve the ability to model complex information such as interactions and motion dynamics that they all help for future trajectory predictions.

Second, we propose a gated shifting operation to learn the movement of pedestrians by dynamically incorporating the pedestrian intention information and global dynamics. The intention of crossing the road or not will affect the spatial locations of the pedestrian. For example, the pedestrian in Figure. 1 has a firm intention to cross, which means he would change his locations to the middle of the road in the next few seconds. Our gated shifting operation enables to adjust the spatial locations based on these two essential cues. In the gated shifting operation, we introduce an intention gate to control the intensity of intention information. It takes the hidden state of the previous step and current step data as input and outputs a vector to extract the relevant information from the cell state. The extracted useful intention information along with global scene dynamics are incorporated into the outputs to improve the precision of predictions further.

Third, we dynamically rescale the output gate of Holistic LSTM to exploit the changing of vehicle speed, which aims to model the changes in the scale of pedestrian's bounding boxes. The movement of vehicles inevitably alters the distance between the pedestrian and the vehicle. Therefore the vehicle speed further influences the scale of the predicted pedestrian bounding box. We integrate the speed variations into LSTM's output gate, which reweights the output channels towards accurate bounding box rescaling. Extensive experiments conducted on three benchmarks of pedestrian trajectory prediction task show that our Holistic LSTM achieves state-of-the-art performance. Our contributions are summarized as follows:

- Targeting the challenging pedestrian trajectory prediction task, we propose to leverage intrinsic relations and mutual interactions among information cues from both pedestrians and vehicles.
- We propose the Holistic LSTM to introduce information interactions. Three additional memory cells are designed explicitly to model future variations.
- We propose a gated shifting operation to learn the spatial movement of the pedestrian and then dynamically rescale the output by the vehicle speed variations. Both operations are well-designed to explore intrinsic relations among multiple cues.
- Extensive experimental results show that our Holistic LSTM achieves state-of-the-art performance on three benchmarks.

II. Related Work

Pedestrian Detection and Tracking. Pedestrian detection and tracking are the foundation of pedestrian trajectory prediction. The classical detection task first applies a sliding window on the input image to extract features at candidate regions, then classifies regions containing the target object. Recently, state-of-the-art detection performance is achieved by deep CNN [14], [15], [16]. As for pedestrian tracking, single-person tracking is considered as person re-identification [17], [18] problem while multi-person tracking methods are used to track multiple person in a crowded scene [19], [20].

Pedestrian Trajectory Prediction. There are relatively fewer work that focus on pedestrian behavior prediction from a moving vehicle perspective. On the contrary, pedestrian trajectory prediction has been studied extensively in a surveillance setting from a birds-eye view [21], [8], [22], [23]. These works have simulated bird's eye views by projecting egocentric video frames onto the ground plane, however road irregularities and some other distortions would make these projections become incorrect, and then prevent accurate position prediction of pedestrians. Recent works [24], [25], [12] forecast the future pedestrian trajectory in 3D space. However, the 3D coordinates are difficult to obtain in real world scenarios, as it requires expensive stereo cameras and LIDAR equipment and there would be a lot of noise data in the obtained 3D maps. The most recent works such as Peek Into The Future (PIF) [8] and State-Refinement LSTM (SR-LSTM) [21] extends [26] with visual features and new pooling mechanisms to improve the prediction precision. It is noticeable that SR-LSTM weighs the contribution of each pedestrian to others via a weighting mechanism. It is similar to the idea in Social-BiGAT [27] which uses an attention mechanism to weigh the contribution of the recurrent states that represent the trajectories of pedestrians. As for existing few highly related work [28], it does not consider interactions between ego-vehicles and pedestrians. However, our method depends on first-person videos captured by an on-board camera and considers the complex interactions

between vehicles and pedestrians. Concretely, we take three critical information cues: vehicle speed, pedestrian intention and global correlations into account.

RNNs for Trajectory Prediction. Recurrent Neural Networks (RNN) and its variant structures such as LSTM [11] and Gated Recurrent Units (GRU) [29] is widely used in various tasks including speech recognition [30], language translation [31], image captioning [32], [33], action recognition [34], [35], [36], [37], [38], [39], [40], [41] and pedestrian trajectory prediction [26], [42], [43], [44], [45], [8]. Among them, Differential RNN [39] adopts a differential gating mechanism for the LSTM network to extract the derivatives of internal state (DoS). Then it leverages the derived Dos to learn salient dynamic motions from successive skeleton data. However, we leverage a correlation cell which learns motion dynamics from the optical flow extracted from the observed sequences. Besides, we propose a gated shifting operation to learn the motion of pedestrians by dynamically incorporating the pedestrian intention information and global dynamics. Part-aware LSTM [40] separates the memory cell of the LSTM into part-based sub-cells which learns the long-term context representations individually for each body part. As for our work, we introduce three extra memory cells which uncover future movement of vehicles, pedestrians and global scenes, respectively. ST-LSTM [41] explores the hidden states of action-related information in both spatial and temporal domains concurrently, and it exploits a trust gate mechanism to handle the noisy input data. In Holistic LSTM, intention states, speed states and correlation states are dynamically incorporated and all updated in a recurrent way. Alahi et al. [26] propose a social LSTM which uses a LSTM network architecture and a social pooling layer to leverages spatial information of nearby pedestrians. Therefore, it can model interactions among the scene. Sun et al. [12] adopt a sequence-to-sequence LSTM encoder-decoder architecture to predict pedestrian position and direction angle. Gupta et al. [13] leverage a recurrent based generative adversarial network which consists of a LSTM-based encoder-decoder generator and LSTM-based discriminator to generate and predict the future pedestrian trajectory. Xue et al. [43] introduce a hierarchical LSTM model to leverage the scene structure for predicting the future trajectory, which incorporates observed trajectory, social neighbourhood and global scene features extracted from CNN. However, most of these trajectory prediction work do not focus on an on-board camera setting and these used information like social interactions may not be available in driving assistance system.

Autonomous and Assisted Driving. Research for vehicle odometry prediction or autonomous driving dates back to ALVINN[46]. This work predicts the driving direction the car should follow through the designed neural network. More recently, diverse datasets are proposed for autonomous driving. Senthil et al. [47] proposes a fish-eye automotive dataset, WoodScape, which comprises of four surround view cameras and nine tasks. Song et al. [2] provides a large-scale database suitable for 3D car instance understanding named ApolloCar3D. Hong et al. [48] predicts the future states of vehicles by learning complex interactions into a unified representation. Since great progress has been made in autonomous driving, it is important to anticipate the near future to guarantee the safety of pedestrians by predicting their future trajectory.

III. Methodology

Our target is to forecast the future pedestrian trajectory based on observed frames over a video. We first formulate this problem in Sec. III-A. Then we introduce the proposed Holistic LSTM in Sec. III-B, which leverages three critical information cues to model the complex interactions. At last, we show that Holistic LSTM can be readily incorporated into the framework for trajectory prediction in Sec. III-C.

A. Formulations

Given the bounding boxes of a pedestrian in the past m frames, our goal is to predict its bounding boxes in the future n frames. Formally, the bounding box of the i-th pedestrian at time step t can be described by the top-left (tl) and bottom-right (br) pixel coordinates: $b_i^t = \{(x_{tl}, y_{tl}), (x_{br}, y_{br})\}$, where x_{tl} and y_{tl} indicate the position of the top-left corner while x_{br} and y_{br} are the bottom-right one. Our objective is to learn the distribution $p(B_f|B_{obs})$ in an optimization process of future trajectory prediction, where $B_{obs} = \{b_i^{t-m}, b_i^{t-m+1}, ..., b_i^t\}$ are the observed sequences of pedestrians, $B_f = \{b_i^{t+1}, b_i^{t+2}, ..., b_i^{t+n}\}$ are the future sequences of pedestrians. However, it is not easy to directly predict the future trajectory B_f based solely on B_{obs} as future trajectory is inevitably uncertain even with the same historical trajectory. Additional information cues, *e.g.*, pedestrian intention, vehicle speed and global scene dynamics are essential to the accuracy of future trajectory prediction. We introduce how we obtain the above information cues in the following:

Intention Estimation. We denote the intention of a pedestrian as $int \in \{0, 1\}$, which indicates the probability of the pedestrian crossing the street. Pedestrian intention gives a hint about the pedestrian's future movement, which helps predict the future trajectory. Following [10], [49], [50], we adopt an encoder-decoder architecture for pedestrian intention estimation, where ConvLSTM [51] is the encoder and the standard LSTM is the decoder. The encoder receives a sequence of square cropped images around the pedestrians over frames. The decoder takes the output of the encoder together with the position coordinates of the pedestrian. The final output of the decoder is a binary value, where 1 indicates the pedestrian intends to cross the street and 0 means not. We set the intention information cue as the mean of pedestrian intentions among the observed frames.

Ego-Speed Estimation. Most of first-view benchmarks lack ego-speed information which is effective for pedestrian trajectory prediction. We estimate ego-speed by using deep neural network that leverages optical flow and monocular depth information. They are extracted from images captured by a camera mounted on the moving car. Since the magnitude of optical flow is highly correlated with the moving speed of the observer, the closer objects are to the observer the faster they appear to be moving. In the pipeline of ego-speed estimation, we introduce an optical flow algorithm (PWC-Net [52]) and a depth estimation algorithm (MonoDepth [53]).

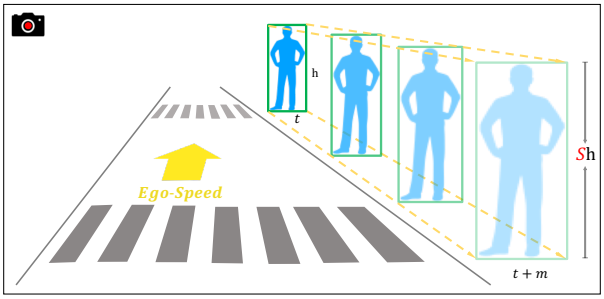


Fig. 2: We simulate the scale changing problem in this figure: a pedestrian is standing by the road from time t to $t+m$. The closer the pedestrian is to the vehicle, the larger his scale seems to change. The height of the pedestrian's bounding box changes from h to Sh.

The speed estimation pipeline is shown as follows. First, we run the optical flow and depth estimation algorithm on image frames. Then, the quotient $OF/DISP$ is regarded as the scaled speed estimates, where OF denotes the mean of the magnitude of optical flow vectors and $DISP$ indicates the mean of the disparity values. We set thresholds: OF>1.0 and $DISP$>0.01 to obtain valid pixels following [54]. After that, we concatenate the scaled speed estimates over frames of a video. The aggregated vectors are then temporally smoothed to get the estimated speed by using a 1D convolution. Finally, we use a scaling factor to convert the estimated speed to the real-world domain and obtain the estimated vehicle speed.

Correlations of Global Scenes. The motion of the global scenes induces additional apparent visual movements of pedestrians and nearby objects, which also is an important cue to the pedestrian's future locations. We view the correlation as a pattern of relative motion in the scene. Following FlowNet [55] whose correlation layer computes correlations over video frames, we define the correlation of two patches centered at x_{i-1} in the first frame b_i^{t-1} and x_i in the second frame b_i^t as

$$Cor(x_{i-1}, x_i) = \sum_{\mu \in [-k,k] \times [-k,k]} \langle f_{i-1}(x_{i-1}+\mu), f_i(x_i+\mu) \rangle \tag{1}$$

For each location x_{i-1} we compute correlations $Cor(x_{i-1}, x_i)$ only in a neighborhood of size $D := 2d+1$ by limiting the range of x_i and we set $d = 10$ as [55]. Then we obtain the correlation of size $(w \times h \times D^2)$. And we apply an average pooling operation to downsample it along $H \times W$ sides. Finally, we concatenate the preprocessed correlations over frames in a video as the correlation cue.

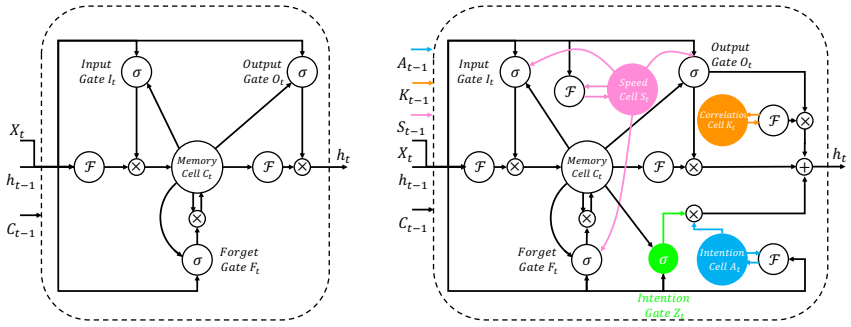
Pedestrian Scale Information. As shown in Figure. 2, vision-based displacement measurement in first-person videos is not consistent with the displacement measurement in physical domain. For example, if a person walks towards the vehicle at a constant speed, we would anticipate him to continue this constant speed in subsequent frames. However, visual displacements shown in the camera change faster as the closer pedestrians are to the vehicles, the larger their scale seems to change. More intuitively, the movement of vehicle induces the changes of the distance between the target pedestrian and the vehicle, therefore it would affect the scale of the predicted pedestrian bounding box. Motivated by these observations, we propose two effective approaches to model the pedestrian scale information. First, we learn both spatial movements and scale information of the target pedestrian jointly. Formally, we add the scale information (h_t) into each location b_i^t, and $x_t = ((b_i^t)^\mathsf{T}, h_t)^\mathsf{T}$. Then, the network input is $X_{in} = (x_{t-m}, x_{t-m+1}, ..., x_t)$, and the output is $X_{out} = \bar{x}_{t+1}, \bar{x}_{t+2}, ..., \bar{x}_{t+n})$, where $\bar{x}_{t+1} = ((b_i^{t+1} - b_i^t)^\mathsf{T}, h_{t+1})^\mathsf{T}$. Second, we dynamically rescale the output gate to exploit the changes of vehicle speed, so as to model the changes in the scale of pedestrian's bounding boxes. Specifically, the vehicle speed is obtained by an on-board diagnostics (OBD) sensor or the ego-speed estimation module.

B. Holistic LSTM

Information cues mentioned above are all highly correlated with the prediction of pedestrian's future trajectory. To leverage the intrinsic interactions among these different cues, we propose Holistic LSTM which introduces extra memory cells, a gated shifting operation and dynamically rescales the output gate according to the changes of vehicle speed.

The architecture of the typical LSTM and the proposed Holistic LSTM are illustrated in Figure 3. The typical LSTM contains an input gate $\mathcal{I}_t$, a forget gate $\mathcal{F}_t$, an output gate $\mathcal{O}_t$, a memory cell $\mathcal{C}_t$, an output response $\mathcal{H}_t$. Besides these, our Holistic LSTM involves a speed cell $\mathcal{S}_t$, an intention gate $\mathcal{Z}_t$, an intention cell $\mathcal{A}_t$ and a correlation cell $\mathcal{K}_t$. In the two LSTMs, the memory cell maintains its state over time, the input gate and forget gate govern the information flow into and out of the memory cell. The output gate controls how much information from the memory cell is passed to the output. Whereas the proposed speed cell, intention gate, intention cell and correlation cell help to handle complex motion information. In our proposed gated shifting operation, pedestrian intention and global scene dynamics are leveraged to perceive the movement of the pedestrian. Mathematically, we use $\mathcal{O}_t \cdot \mathcal{K}_t$ and $\mathcal{Z}_t \cdot \mathcal{A}_t$ to dynamically change the output $\mathcal{H}_t$. Moreover, Holistic LSTM learns the changes in the scale of pedestrian's bounding boxes by exploiting the changing of vehicle speed. And we encode $\mathcal{S}_t$ into $\mathcal{O}_t$ for dynamically rescaling the output gate. Instead of adopting a simple concatenation operation on those helpful cues, we propose Holistic LSTM which is beneficial to making good use of the multiple information cues for pedestrian future trajectory prediction. To better understand how we leverage these important cues and designed operations, we introduce the details of these extra memory cells and gates in Holistic LSTM in the following.

Speed Cell. We observed that the movement of vehicle changes the distance between the target pedestrian and the vehicle. It implicitly indicates that the vehicle speed further influences the scale of pedestrian bounding box. Motivated by this observation, a speed cell is designed to learn the relation between vehicle speed and the change in the scale of pedestrian bounding box. The initial input of the speed cell is a concatenation of all vehicle speeds in observed frames, $\mathcal{S}_0 \in \mathbb{R}_{N \times m}$, named dynamic speed units. At each



(a) LSTM (b) Holistic LSTM

Fig. 3: (a). The structure of a standard LSTM neuron. (b) The structure of our proposed Holistic LSTM neuron which has a speed cell $\mathcal{S}_t$, an intention cell $\mathcal{A}_t$, a correlation cell $\mathcal{K}_t$ and an intention gate $\mathcal{Z}_t$ additionally, where σ is sigmoid activation function and $\mathcal{F}$ is softsign activation function.

time step, a new speed state will be calculated by $\mathcal{S}_t = softsign(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{S}_{t-1} + b_s)$. Then the speed item is incorporated into the output gate $\mathcal{O}_t = \sigma(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{C}_t + W\mathcal{S}_t + b_o)$ to dynamically rescale the output responses (changes in bounding box scale). Finally, the updated output $\mathcal{H}_t$ and speed state $\mathcal{S}_t$ will be received by the Holistic LSTM neuron of the next step.

Intention Gate. The pedestrian intention for crossing the street will influence his spatial movements in physical domain. For example, a pedestrian with a strong intention to cross the street will change his bounding box locations, even observed from a stationary vehicle. However, the target pedestrian's physical locations will not change if he has no intention ($int = 0$) to cross (stop by the road). We propose an intention gate to govern how much the location of the pedestrian will change according to different pedestrian intentions. We use a sigmoid function as the activation to achieve the intention gate, $\mathcal{Z}_t = \sigma(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{C}_t + b_z)$. The input intention value is the output of the applied intention estimation module. An intention gate is developed to make Holistic LSTM robust to noisy intention data as intention results are predicted by the intention module which consists of uncertainties. In this way, it can analyze the reliability of the intention data at each spatiotemporal step and give better insight to the network about when to update, forget, or remember the contents of the intention memory cell, the representation of long-term location movements information.

Intention Cell. As mentioned above, the pedestrian crossing intention has impacts on his spatiotemporal movements. Besides, there are complex interactions between pedestrian and vehicle, e.g. the pedestrian would change his intention for crossing the road if the coming vehicle speeds up. Therefore, we propose an intention cell to maintain the dynamic intention states rather than using a fixed intention value in previous methods. The initial input ($\mathcal{A}_0 \in \mathbb{R}_{N\times 1}$) of the intention cell are the results of the pedestrian intention estimation model. Then the intention state will be updated by $\mathcal{A}_t = softsign(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{A}_{t-1} + b_a)$ at every time step and taken as input for the intention cell in next LSTM neuron. Each intention state will also be incorporated into the output responses through the gated shifting operation.

Correlation Cell. Correlation cell aims to learn motion dynamics and directions of global scenes in the observed sequences, which help to predict the future pedestrian trajectory. The input of correlation cell in Holistic LSTM is the correlation cue ($\mathcal{K}_0 \in \mathbb{R}_{N\times 441}$) preprocessed in Sec. III-A. We use an activation operation to update the correlation state. Like other memory cells, each updated correlation result would be encoded into the hidden state of current step to further influence other states.

Recurrent State. The speed variations will influence the LSTM's output gate and the output will change speed states in next LSTM neuron, just as the mutual interactions between the vehicle and the pedestrian. Mathematically, the proposed Holistic LSTM is formulated as:

$$\mathcal{I}_t = \sigma(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{C}_{t-1} + W\mathcal{S}_{t-1} + b_i) \quad (2)$$
$$\mathcal{F}_t = \sigma(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{C}_{t-1} + W\mathcal{S}_{t-1} + b_f) \quad (3)$$
$$\mathcal{K}_t = softsign(W\mathcal{K}_{t-1} + b_k) \quad (4)$$
$$\mathcal{C}_t = \mathcal{F}_t\mathcal{C}_{t-1} + \mathcal{I}_t softsign(W\mathcal{X}_t + W\mathcal{H}_{t-1} + b_c) \quad (5)$$
$$\mathcal{S}_t = softsign(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{S}_{t-1} + b_s) \quad (6)$$
$$\mathcal{O}_t = \sigma(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{C}_t + W\mathcal{S}_t + b_o) \quad (7)$$
$$\mathcal{Z}_t = \sigma(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{C}_t + b_z) \quad (8)$$
$$\mathcal{A}_t = softsign(W\mathcal{X}_t + W\mathcal{H}_{t-1} + W\mathcal{A}_{t-1} + b_a) \quad (9)$$
$$\mathcal{H}_t = \mathcal{O}_t softsign(\mathcal{C}_t) + \mathcal{O}_t\mathcal{K}_t + \mathcal{Z}_t\mathcal{A}_t, \quad (10)$$

where σ is the sigmoid function, weight matrices are denoted as $\mathcal{W}$ and bias vectors denoted as b. The gated shift operation includes a dot product between the intention gate $\mathcal{Z}_t$ and intention cell state $\mathcal{A}_t$, another dot product between the output gate $\mathcal{O}_t$ and correlation cell state $\mathcal{K}_t$. Ego-speed cell state $\mathcal{S}_t$ dynamically reweights the output channels which can be beneficial to a more accurate location prediction of the pedestrian bounding boxes. In the proposed LSTM network, the speed cell, intention cell and correlation cell would update its state at every time step. Then the speed cell will affect the output responses by changing the output gate to reweight the output channels, which towards accurate bounding box rescaling. However, the intention cell and correlation cell directly alters the output which is in line with the reality that pedestrian spatial movements depend on both the intention value of

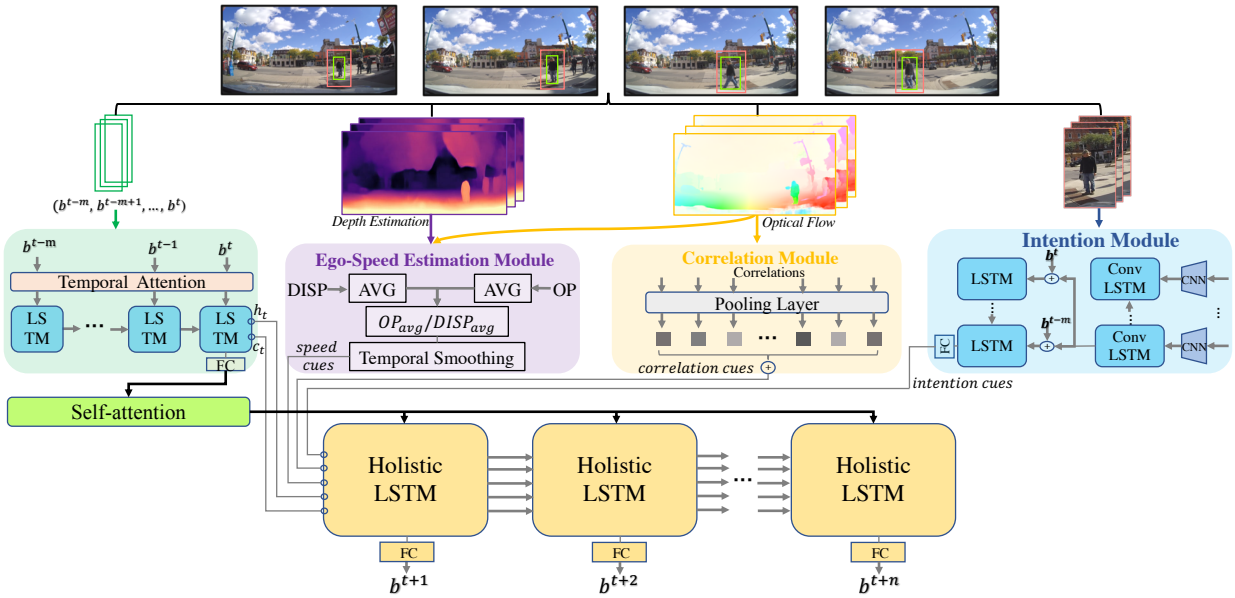


Fig. 4: Framework of the pedestrian trajectory prediction method equipped with the Holistic LSTM, which consists of four modules: pedestrian intention module, ego-speed estimation module, correlation module and trajectory prediction module. The intention module (blue) is applied to predict pedestrian crossing intentions by taking a series of cropped images around the target pedestrian as input. The ego-speed module (purple) receives both depth estimation and optical flow of frames then estimates ego-speed cues. The correlation module (yellow) computes correlations over frames. The encoder model (green) of the prediction network processes location information of pedestrian bounding boxes. Finally, Holistic LSTM as the decoder model receives the output of encoder model, predicted intentions, correlations and ego-speeds to predict the future locations.

crossing the street and the motion dynamics of global scenes. In a word, the influence of speed and intention information can be seen as the changes in the scale of pedestrian bounding box and the movement of spatial locations respectively, which can be shown in Figure. 2. The updated intention states, speed states and correlation states will be taken as the input of next LSTM neuron. Besides, the output responses of each Holistic LSTM layer will also be taken as the input of next LSTM layer to further influence the future gates and cells. In this way, all states can be updated on a recurrent way. Therefore, a comprehensive representation incorporating these critical information cues for pedestrian trajectory prediction can be learned by Holistic LSTM.

C. Holistic LSTM for Trajectory Prediction

Our framework for pedestrian trajectory prediction can be seen in Figure 4. It contains four necessary modules: pedestrian intention module, ego-speed estimation module, global correlation module and pedestrian trajectory module. Our proposed Holistic LSTM is applied in pedestrian trajectory prediction module. As shown in the figure, we apply a RNN encoder-decoder network where Holistic LSTM is equipped in the decoder architecture. The inputs to the encoder are the observed bounding box locations of the pedestrian. A temporal attention module applied to the encoder inputs aims to find the most relevant frames in the observed data. Then the output of the encoder is processed by a fully connected network and a self-attention module. The former aims to reduce the data dimension and the latter is to find the most related information. For the Holistic LSTM network in the decoder model, it receives six inputs $\{\mathcal{C}_{t-1}, \mathcal{S}_{t-1}, \mathcal{H}_{t-1}, \mathcal{X}_{t-1}, \mathcal{A}_{t-1}, \mathcal{K}_{t-1}\}$ at $t-th$ time step, which is described in III-B. It will update each memory state at every time step after learning the interactions among these information cues. The Holistic LSTM is followed by fully connected network and the final output are the prediction results of future pedestrian locations.

IV. Experiments

A. Datasets

JAAD [25]. This dataset is designed to study the behavior of traffic participants including pedestrians and vehicles. It contains videos captured with a front-view camera under various scenes, weathers and lighting conditions. It consists of 346 high-resolution video clips (5-15s) with annotations and 82,032 frames extracted from 240 hours driving videos. We use the same train/test split as in [10].

Pedestrian Intention Estimation (PIE) [10]. PIE is a recently proposed dataset that includes over 6 hours of video footage of pedestrians at various types of crosswalks, it is also captured by a on-board camera. It is the only dataset that simultaneously contains pedestrian intentions and ego-vehicle information (e.g. ego-speed) so far. Besides, the data provides bounding boxes for traffic objects, pedestrian attributes and road boundaries which are necessary for perception and visual reasoning. There are 1,842 pedestrians samples and 293K annotated frames. It is divided into train, test and validation sets with the ratio of 50%, 40% and 10% respectively.

S-KITTI. KITTI [1] is a challenging benchmark for the tasks of stereo, optical flow, visual odometry/SLAM and 3D

Method	PIE					JAAD				
	MSE			C_{MSE}	CF_{MSE}	MSE			C_{MSE}	CF_{MSE}
	0.5s	1s	1.5s	1.5s	1.5s	0.5s	1s	1.5s	1.5s	1.5s
Linear [10]	123	477	1365	950	3983	223	857	2303	1565	6111
LSTM [10]	172	330	911	837	3352	289	569	1558	1473	5766
B-LSTM [56]	101	296	855	811	3259	159	539	1535	1447	5615
PIE$_{traj}$ [10]	62	186	559	520	2162	110	399	1248	1183	4780
Ours	**56**	**167**	**507**	**466**	**1917**	**105**	**389**	**1177**	**1116**	**4493**

S-KITTI	$MSE_{1.5s}$	$C_{MSE_{1.5s}}$
Baseline$_{bbox}$	628	557
Baseline$_{concat}$	599	544
PIE$_{traj}$	574	528
Ours	**525**	**470**

TABLE I: **Left**: Prediction errors over multiple future time steps of different methods. MSE is calculated over bounding box coordinates in pixels. C_{MSE} and CF_{MSE} are the $MSEs$ calculated over the center of the bounding boxes for all the predicted frames and the last time step respectively. In these data, 1s represents 30 future frames in the video. As JAAD dataset does not contain necessary vehicle ego-speed information, we obtain ego-speed cues by applying the ego-speed estimation module. **Right**: Results on the S-KITTI dataset. Baseline$_{bbox}$ indicates that only to leverage observed location information for future trajectory prediction. Baseline$_{concat}$ concatenates different information cues (estimated ego-speed, predicted intention and calculated correlation) into the input.

Input	Method	MSE	C_{MSE}	CF_{MSE}
$loc + int$	PIE$_{traj}$ [10]	611	570	2414
	Holistic LSTM	**596**	**557**	**2346**
$loc + speed$	PIE$_{traj}$ [10]	572	535	2204
	Holistic LSTM	**554**	**516**	**2131**
$loc + int + speed$	PIE$_{traj}$ [10]	559	520	2162
	Holistic LSTM	**539**	**501**	**2057**

TABLE II: Comparisons with state-of-the-art method PIE$_{traj}$ with various input. loc, int and $speed$ stand for location, intention and vehicle speed in PIE. To have a fair comparison, we do not leverage scale information like PIE$_{traj}$ and we also report 1.5s prediction on the three metrics.

object detection. We manually select a part of data including crossing pedestrians as S-KITTI (Small-KITTI). S-KITTI has 69 persons which make up 712 sequences of pedestrian trajectory data. In the experiments of S-KITTI, we keep the same setting (0.5s observation, 1.5s prediction) as in the other two benchmarks and we only test on S-KITTI using model pretrained on PIE because it is a small amount of data.

Evaluation Metrics. To evaluate the performance of our Holistic LSTM and compare with other methods, we report the following metrics: MSE over bounding box coordinates[56], C_{MSE} which is the average MSE of the center of the bounding boxes on the predicted sequences, F_{MSE} stands for the MSE of the last time $(t+n)$, CF_{MSE} indicates the C_{MSE} of the last time $(t+n)$. All results of the pedestrian bounding box predictions are in pixels. We conduct each experiment for five times and use the mean results as our final results in the table, since the prediction task consists of uncertainties.

B. Implementation

Our framework is implemented by Keras[1] and PaddlePaddle[2]. In the training phrase, we set batch size as 64 and total epoch as 100. We use EarlyStopping and ReduceLROnPlateau as the learning schedulers in which we set min_delta to 0.1, $factor$ to 0.2, $patience$ as 10 in EarlyStopping and 5 in ReduceLROnPlateau. A learning rate is initially set to 0.001. The number of units in LSTMs is set to 256. As our target is to design a new LSTM network for learning a better incorporation among various information cues, we do not train pedestrian intention estimation model and just adopt the model pretrained in [10]. In our experiments, models are trained and tested on 0.5s observation, then predict future trajectories over 0.5s, 1.0s and 1.5s. As both JAAD and S-KITTI benchmarks lack necessary vehicle speed information, we use the speed estimation module to obtain the estimated speed of each frame.

[1]https://github.com/fchollet/keras

[2]https://github.com/PaddlePaddle/Paddle

C. Comparison to State-of-the-art

We compare our proposed method with state-of-the-art methods in Table. I. As shown in the table, our Holistic LSTM outperforms all other existing methods on three benchmarks. On the PIE dataset, we improve the performance by up to 10% than PIE$_{traj}$ on $MSE_{0.5s}$ (from 62 to 56). Also, Holistic LSTM is better than (by about 10%) PIE$_{traj}$ in a longer time prediction on $MSE_{1.5s}$ (from 559 to 507). On the JAAD dataset, we improve the performance from 4780 to 4493 on the metric $CF_{MSE_{1.5s}}$. This indicates that Holistic LSTM exhibits stronger ability on both short and long time predictions. Due to the small scale of the S-KITTI dataset, we directly apply the model trained on the PIE dataset without finetuning for evaluation. In this case, Holistic LSTM can achieve a better performance than PIE$_{traj}$ (574 vs. 525) which indicates that our method has a strong generalization ability on different data. To have a more comprehensive comparison to the state-of-the-art method PIE$_{traj}$, we conduct experiments by using different input information as shown in Table. II, and we do not encode pedestrian scale information into the Holistic LSTM network for a fair comparison. We outperform PIE$_{traj}$ on $MSE_{1.5s}$ by 15, 18, 20 with $loc + int$, $loc + speed$, $loc + int + speed$ input information respectively. This indicates that our method is able to handle different information cues and comprehensively exceed the state-of-the-art method PIE$_{traj}$. Improvements on JAAD are slightly less significant than on PIE because JAAD has a large amount of data and its vehicle speed information are estimated by the learned model. Holistic LSTM reduces 5 prediction errors (4.5%) in $MSE_{0.5s}$, 71 prediction errors (5.7%) in $MSE_{1.5s}$ and 287 (6.0%) in $CF_{MSE_{1.5s}}$, it also proves that our method performs better in a long time prediction.

Input	Scale	MSE			F_{MSE}	C_{MSE}			CF_{MSE}
		0.5s	1s	1.5s	1.5s	0.5s	1s	1.5s	1.5s
$loc + int$	×	56	187	596	2455	40	164	557	2346
	√	57	178	555	2273	40	154	516	2169
$loc + speed$	×	61	181	554	2232	43	158	516	2131
	√	60	176	529	2129	43	152	491	2028
$loc + int + speed$	×	60	181	539	2159	43	156	501	2057
	√	58	173	517	2068	41	150	481	1975

TABLE III: Comparison results of encoding pedestrian scale information or not in Holistic LSTM with different inputs.

Input	Method	MSE			C_{MSE}			CF_{MSE}
		0.5s	1s	1.5s	0.5s	1s	1.5s	1.5s
int	Concatenation Baseline	60	186	576	43	162	536	2243
	Holistic LSTM	**57**	**178**	**555**	**40**	**154**	**516**	**2169**
$speed$	Concatenation Baseline	63	183	547	45	159	508	2086
	Holistic LSTM	**60**	**176**	**529**	**43**	**152**	**491**	**2028**
$int + speed$	Concatenation Baseline	63	183	545	45	158	505	2074
	Holistic LSTM	**58**	**173**	**517**	**41**	**150**	**481**	**1975**
$int + speed + correlation$	Concatenation Baseline	60	179	536	43	152	494	2033
	Holistic LSTM	**56**	**167**	**507**	**38**	**139**	**466**	**1917**

TABLE IV: Comparisons between 'Concatenation Baselines' and 'Holistic LSTMs'. Models of 'Concatenate Baseline' concatenate different information cues as input data and apply a typical LSTM network to encode these cues. However, our proposed Holistic LSTM network introduces extra memory cells and novel operations to leverage all the cues.

D. Ablation Studies

We conduct ablation studies on the PIE dataset to see how each of components: pedestrian scale information, pedestrian intention cues, ego-vehicle speed cues, global correlation cues and the proposed LSTM architecture contributes overall prediction performance.

Scale Information. To investigate the effect of scale information, we conduct experiments with/without scale information and the results are shown in Table. III. We observe that encoding scale information helps the performance most for the model with $loc + int$ as input, which reduces mse from 596 to 555 on $MSE_{1.5s}$ and 557 to 516 on $C_{MSE_{1.5s}}$. However, it reduces less mse on $MSE_{1.5s}$ by 25 (from 554 to 529) on the input $loc + speed$, this may be because ego-speed information is also leveraged to perceive the changes of pedestrian's scale information. When the input to the model turns to $loc + int + speed$, it also reduces 22 and 20 on $MSE_{1.5s}$ and $C_{MSE_{1.5s}}$ respectively. All the results in Table. III prove that scale information is critical for pedestrian trajectory prediction.

Different information cues. Three information cues are leveraged to construct four kinds of input for Holistic LSTM. The experiment results with the four different input are shown in Table. IV. With only the ego-speed information as input, Holistic LSTM gets 529 on MSE-1.5s which is better than only using the intention information cue. However, the model with intention information as input performs slightly better in 0.5s prediction, which demonstrates that pedestrian intention information may contribute more in a short time prediction whereas vehicle ego-speed information helps more in a long time prediction. Also Holistic LSTM performs better when the proposed operations are applied to incorporate both ego-speed and intention information denoted as Holistic LSTM$_{int+speed}$. Finally, Holistic LSTM achieves 507 MSE by incorporating an additional correlation information cue. All the results illustrate that the proposed extra memory cells and operations are effective for the pedestrian trajectory prediction.

Holistic LSTM In Table IV, 'Concatenate Baselines' adopts a traditional LSTM network and directly concatenates the input information as decoder input, whereas Holistic LSTM learns to dynamically incorporate the input information cues. Both models of Baselines and models of Holistic LSTMs encode pedestrian's scale information into the input data. With four kinds of information cues as input, Holistic LSTMs outperform Baselines by a margin of about 6% on all evaluation metrics. Concretely, Holistic LSTM achieves the best performance by encoding these four cues, it reduces 29 on $MSE_{1.5s}$ and 28 on $C_{MSE_{1.5s}}$ than the Concatenation Baseline.

E. Visualization

To better understand experimental results and the improvements to the state-of-the-art method, we display several of our prediction results in Figure. 5. From theses visualized results, our proposed Holistic LSTM outperforms the state-of-the-art method PIE$_{traj}$ in many different scenarios. Our Holistic LSTM can handle the changes in pedestrian scale well as shown in 1st and 3rd row in Figure. 5: a person is crossing the road from the front of the vehicle while the vehicle moves at high speed. Although the pedestrian has spatial movements as well as changes in the scale of his bounding box, Holistic LSTM exhibits much better prediction results than PIE$_{traj}$.

A few failure cases are shown in Figure. 6 to demonstrate some possible aspects can be improved in future work. One failure case is shown in the first row: when the vehicle stops at the crossing and a pedestrian walks towards the car, Holistic LSTM $_{int}$ demonstrates a stronger prediction ability than Holistic LSTM $_{int+speed+correlation}$. In such scenario, pedestrian's intention plays a greater role while global dynamics

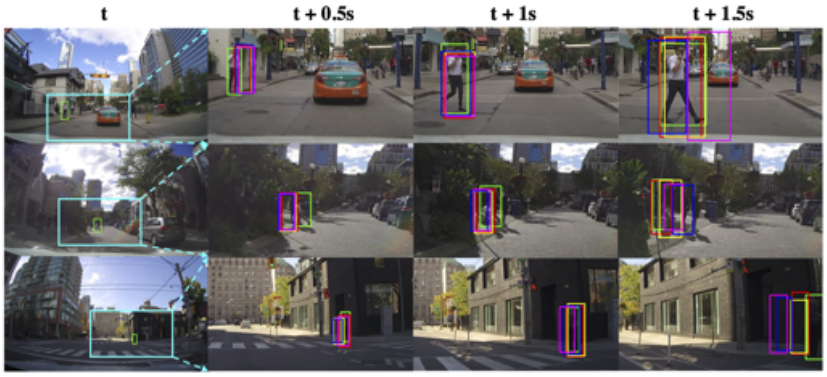


Fig. 5: Visualization examples of pedestrian trajectory predictions. The first column consists of one observed video frame with the green ground truth bounding box. Column 2∼4 represent the prediction results for the future 0.5s, 1.0s and 1.5s respectively. And bounding boxes with different color correspond to different model: PIE$_{traj}$ (purple), Holistic LSTM$_{int}$ (blue), Holistic LSTM$_{speed}$ (yellow) and Holistic LSTM$_{int+speed+correlation}$ (red).

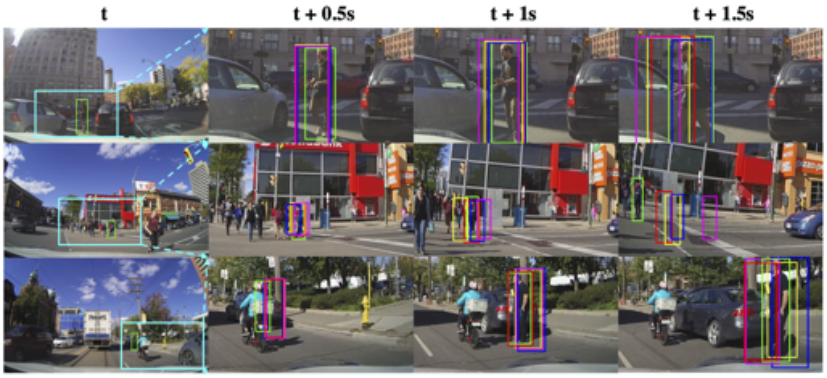


Fig. 6: Visualization of failure modes in pedestrian trajectory predictions.

extracted from noisy backgrounds might hinder the prediction results. Another failure case is shown in the second row, the vehicle is turning right at the crossing and all methods acquire bad prediction results. However, Holistic LSTM still performs best whereas PIE$_{traj}$ performs much worse than all Holistic LSTM models, which indicates that the proposed Holistic LSTM is good at sensing the changes in directions of the scenes. In the third row, Holistic LSTM does not obtain an optimal result that might due to the interference brought from the cyclist. The cyclist not only affects the mutual interactions between the vehicle and the pedestrian to a certain extent but also potentially changes their trajectories. Observed from these visualization results, we can find that different information cues help Holistic LSTM to handle different scenarios, and it is necessary to leverage such a well-designed LSTM to dynamically incorporate these useful cues.

F. Limitation

Since depth estimation and optical flow technologies have limitations in night-time scenes and adverse weather conditions, we could incorporate some image restoration methods (e.g., deraining, dehazing, desnowing and low-light enhancement) to improve the applicability of these technologies, or take radar signal information into consideration in future work. Moreover, given that the existing datasets do not provide night-time scenes and adverse weather conditions for pedestrian trajectory prediction, we would leave the studies of these scenes in future work.

V. Conclusions

In this paper, we propose a novel Holistic LSTM network for pedestrian trajectory prediction, which analyses the location movements of the target pedestrian, vehicle speed information, motion dynamics of global views and pedestrian intentions for crossing the street together at every time step. Extra memory cells: speed cell, intention cell and correlation cell are proposed in Holistic LSTM to improve the ability of LSTMs in modeling future dynamic variations. And a novel gated shifting operation is introduced to dynamically incorporate the pedestrian intention and global correlation information, which mainly governs the spatial movement of the pedestrian. Moreover, we explore to rescale the output of Holistic LSTM dynamically according to the vehicle speed variations, which results in more accurate predictions of pedestrian's bounding boxes. In experiments, the proposed Holistic LSTM achieves state-of-the-art performance on three first-view benchmarks of pedestrian trajectory prediction.

References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, 2013.

[2] X. Song, P. Wang, D. Zhou, R. Zhu, C. Guan, Y. Dai, H. Su, H. Li, and R. Yang, "Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving," in *CVPR*, 2019.

[3] R. Satzoda and M. Trivedi, "Vision-based lane analysis: Exploration of issues and approaches for embedded realization," in *CVPR-W*, 2013.

[4] X. Du, M. H. Ang, and D. Rus, "Car detection for autonomous vehicle: Lidar and vision fusion approach through deep learning framework," in *IROS*, 2017.

[5] W. Chu, Y. Liu, C. Shen, D. Cai, and X.-S. Hua, "Multi-task vehicle detection with region-of-interest voting," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 432–441, 2017.

[6] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *CVPR*, 2019.

[7] C. Zhou, M. Yang, and J. Yuan, "Discriminative feature transformation for occluded pedestrian detection," in *ICCV*, 2019.

[8] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *CVPR*, 2019.

[9] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *IV*, 2017.

[10] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *ICCV*, 2019.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[12] L. Sun, Z. Yan, S. M. Mellado, M. Hanheide, and T. Duckett, "3dof pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data," in *ICRA*, 2018.

[13] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.

[14] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting," in *ECCV*, 2018.

[15] J. Noh, S. Lee, B. Kim, and G. Kim, "Improving occlusion and hard negative handling for single-stage pedestrian detectors," in *CVPR*, 2018.

[16] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3820–3834, 2020.

[17] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[18] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[19] S. Tang, B. Andres, M. Andriluka, and B. Schiele, "Multi-person tracking by multicut and deep matching," in *ECCV*, 2016.

[20] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "Arttrack: Articulated multi-person tracking in the wild," in *CVPR*, 2017.

[21] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction," in *CVPR*, 2019.

[22] J. Liang, L. Jiang, and A. Hauptmann, "Simaug: Learning robust representations from 3d simulation for pedestrian trajectory prediction in unseen cameras," *arXiv preprint arXiv:2004.02022*, 2020.

[23] M. Pfeiffer, G. Paolo, H. Sommer, J. Nieto, R. Siegwart, and C. Cadena, "A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments," in *ICRA*, 2018.

[24] C. G. Keller, C. Hermes, and D. M. Gavrila, "Will the pedestrian cross? probabilistic path prediction based on learned motion features," in *JPRS*, 2011.

[25] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (jaad)," *arXiv preprint arXiv:1609.04741*, 2016.

[26] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *CVPR*, 2016.

[27] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," in *NeurIPS*, 2019.

[28] V. Karasev, A. Ayvaci, B. Heisele, and S. Soatto, "Intent-aware long-term prediction of pedestrian motion," in *ICRA*, 2016.

[29] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[30] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, 2014.

[31] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *NeurIPS*, 2015.

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[33] Z. Zhao, Z. Zhang, S. Xiao, Z. Xiao, X. Yan, J. Yu, D. Cai, and F. Wu, "Long-form video question answering via dynamic hierarchical reinforced networks," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5939–5952, 2019.

[34] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based lstm networks for 3d action recognition and detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3459–3471, 2018.

[35] M. Lu, Z. Li, Y. Wang, and G. Pan, "Deep attention network for egocentric action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3703–3713, 2019.

[36] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1347–1360, 2018.

[37] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3275–3287, 2018.

[38] Y. Wu, L. Zhu, X. Wang, Y. Yang, and F. Wu, "Learning to anticipate egocentric actions by imagination," *IEEE Transactions on Image Processing*, vol. 30, pp. 1143–1152, 2021.

[39] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *ICCV*, 2015, pp. 4041–4049.

[40] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016, pp. 1010–1019.

[41] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal lstm network with trust gates," *TPAMI*, vol. 40, no. 12, pp. 3007–3021, 2017.

[42] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, "Desire: Distant future prediction in dynamic scenes with interacting agents," in *CVPR*, 2017.

[43] H. Xue, D. Q. Huynh, and M. Reynolds, "Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction," in *WACV*, 2018.

[44] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *ICLR*, 2019.

[45] J. Liang, L. Jiang, K. Murphy, T. Yu, and A. Hauptmann, "The garden of forking paths: Towards multi-future trajectory prediction," in *CVPR*, 2020.

[46] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," in *NeurIPS*, 1989.

[47] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricar, S. Milz, M. Simon, K. Amende, C. Witt, H. Rashed, S. Chennupati, S. Nayak, S. Mansoor, X. Perrotton, and P. Perez, "Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving," in *ICCV*, 2019.

[48] J. Hong, B. Sapp, and J. Philbin, "Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions," in *CVPR*, 2019.

[49] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NeurIPS*, 2014.

[50] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

[51] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NeurIPS*, 2015.

[52] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *CVPR*, 2018.

[53] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019.

[54] R.-A. Rill, "Speed estimation evaluation on the kitti benchmark based on motion and monocular depth information," *arXiv preprint arXiv:1907.06989*, 2019.

[55] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *ICCV*, 2015.

[56] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *CVPR*, 2018.