24^{th} *Computer Vision Winter Workshop*
Friedrich Fraundorfer, Peter M. Roth, Fabian Schenk (eds.)
Stift Vorau, Austria, February 6–8, 2019
DOI: 10.3217/978-3-85125-652-9-01

Situation-Aware Pedestrian Trajectory Prediction with Spatio-Temporal Attention Model

Sirin Haddad Meiqing Wu He Wei
Siew Kei Lam
Nanyang Technological University (NTU)
50 Nanyang Ave, Singapore
{siri0005,wei005}@e.ntu.edu.sg {meiqingwu,assklam}@ntu.edu.sg

Abstract.

Pedestrian trajectory prediction is essential for collision avoidance in autonomous driving and robot navigation. However, predicting a pedestrian's trajectory in crowded environments is non-trivial as it is influenced by other pedestrians' motion and static structures that are present in the scene. Such human-human and human-space interactions lead to non-linearities in the trajectories. In this paper, we present a new spatio-temporal graph based Long Short-Term Memory (LSTM) network for predicting pedestrian trajectory in crowded environments, which takes into account the interaction with static (physical objects) and dynamic (other pedestrians) elements in the scene. Our results are based on two widely-used datasets to demonstrate that the proposed method outperforms the state-of-the-art approaches in human trajectory prediction. In particular, our method leads to a reduction in Average Displacement Error (ADE) and Final Displacement Error (FDE) of up to 55% and 61% respectively over state-of-the-art approaches.

1. Introduction

The provision to estimate future trajectories of pedestrians and predicting the possibility of collisions can prevent accidents in autonomous driving and robot navigation. However, pedestrian trajectory prediction in crowded environments is a challenging task as human navigation decisions are influenced by their interactions with other traffic participants and the static physical objects. In particular, humans navigate in a situation-aware manner by avoiding collisions with static objects and other pedestrians in the space surrounding them, based on common social rules. As such, prediction models must take into account the interactions of both static and dynamic elements in the environment in order to accurately predict the pedestrians' motion paths. Figure 1 shows a real scenario that requires awareness of the lamp post presence in order to make realistic prediction about the pedestrian trajectory who will avoid walking into paths leading to the obstacles area.

Figure 1: Estimating pedestrian trajectory given the surrounding environment physical structure in a real-life scenario. In the figure above, a lamp post lies in the direction of traversal of the pedestrian of interest. Thus it is essential to capture the existence of static obstacle and understand how they will navigate around it.

Previous works that addressed human motion prediction focused on modeling human-human and human-space interaction separately. [1, 2, 3, 4, 5] account for scene static configuration such as obstacles and scene structures for improving human trajectory predictions in the presence of dynamic objects. However, these works mainly target constrained environments with low crowd density.

Recently, the work in [6] presented a deep convolutional network that models the impact of scene static elements on the pedestrian motion. However, they relied on complex tools comprising convolution layers and multiple feature maps for modeling knowledge about the scene. Recurrent neural networks in [7, 8, 9] tackled pedestrian trajectory prediction on challenging datasets of outdoor scenes [10]. Nevertheless, these approaches only modeled the social interaction among pedestrians without taking into account the surrounding static context. Social Attention [7] encapsulated the social interactions along the spatial and temporal domains by adopting spatio-temporal graph architecture. Their model considered the social interaction as a global event occurring between each and every pedestrian using their velocity to state their influence on each other. In contrast, Social LSTM [8] only accounted for the influences within a fixed-size local neighborhood.

In our work, we propose an enhancement to the models in [7, 11] and improve the modeling of multiple trajectories correlations over space-time dimensions using the 2D locations of the the static and dynamic elements. In particular, the proposed model overcomes the limitation of Social LSTM [8] which only accounted for the influence of other pedestrians within a local neighborhood, while at the same time being cognizant of the static obstacles at close proximity. This concern was not present in Social Attention [7]. Our intuition for this model is that while a pedestrian's trajectory can be affected by the dynamic motions of other pedestrians at a distance, the decision to avoid static objects is usually made when the pedestrian is close to the object. Thus, we manage to reduce the graph complexity and achieve more stable predictions by dynamically incorporating the static elements in the graph structure only when they potentially pose an impact on pedestrian trajectory.

Our main contributions are as follows: (1) we present a spatio-temporal graph that explicitly captures the global interaction of all the pedestrians in the scene and the local interaction with the static objects, and (2) we propose a new spatio-temporal attention mechanism for each pedestrian trajectory. This mechanism takes into account the local interaction among pedestrians and objects. Our spatio-temporal mechanism is inspired by the work of [12] which casts the attention methods [13] for sequence learning tasks on graphs. Experimental results on two widely-used datasets demonstrate that our method achieves significant quantitative and qualitative improvements over state-of-the-art methods for pedestrian trajectory prediction.

2. Related Work

In this section, we present a summary of research on pedestrians trajectory prediction. The literature branches into two main trends regarding context inclusion: local context and global context. Additionally, the existing works unfolds into two other branches in terms of distinguishing multiple objects influence: attention-based and uniformly-based approaches.

Local context Versus Global context. It is obvious from the previous introduction that the modern trajectory prediction approaches [14, 15, 8, 3, 16, 17] resorted to a limited spatial extent of the surrounding context as they observed the interactions occurring within short distance from the pedestrian included, while [7, 9] were globally-based as they considered all the pedestrians in the scene even those who are far away from each other.

According to local context methods, observing the interaction for a short duration once pedestrians are close enough to each other, gives limited understanding of the social interaction. While including the social interactions on a global scene scale, enables the model to better understand how the interaction evolves between a pair of pedestrians based on the velocity effect that the model inherently grasps upon capturing the change in the spatial distances along time.

Attention-based Versus Uniformly-based approaches. Pedestrians navigating in urban environments influence each other and very often are influenced by the obstacles around them, thus it is essential for predicting multiple pedestrians trajectories to recognize the importance of various sources impact on a pedestrian and pay attention to the more influential ones. Applying attention in sequence learning tasks has proved its effectiveness in the overall algorithm performance and in pedestrian trajectory prediction methods it helped drawing more plausible trajectories.

The variational encoder-decoder methods, such as, Social GAN [9] took the global neighborhood around pedestrian but it evaluates all pedestrians in a uniform manner, by assigning equal importance values to them. Existing RNN approaches [7] applied soft attention mechanism to assign different impor-

tance weights to multiple pedestrians based on their velocities. While [16] applied hard attention to assign weights based on pedestrians distance, they also introduced additional soft attention to evaluate the interaction salience in a scene region. So, their trajectory prediction drew conclusions about which region was more likely for a pedestrian to navigate through. In our work, we are rather interested in microscopic prediction of the interaction between pedestrians and a specific fixed obstacle, hence, we use the soft attention mechanism [12] to evaluate the social interactions only.

Graph-Structured Networks. Real-life applications generate complicated forms of information in which they are best represented through graph structures compared to other rigid hierarchical and end-to-end organizations. Variational Encoder-Decoder methods[9, 18, 19], have the advantage of generating a variety of results, however, they are not capable of providing a factorized and explicit high-level representation of the environment components. Graph Neural Network [20] advanced the application of graph-structured data in neural networks in environments that naturally contain highly interrelated behaviors, such as: social media, molecular biology, etc. Outdoor pedestrians navigation typically induces a spatio-temporal nature due to alterations that happen in pedestrian motion trajectory and the complex interactions with different objects. Therefore, modeling a rich interactive context requires a scalable graph-based structuring of the elements and factorize their relationships in a principled way. Neural relational networks [14, 15], attempted to predict the interactions among multiple moving object using physical motion semantics, however, they did not account for realistic scenarios such as urban environments, which makes these networks better fitting for object linear motion in free space.

Recurrent Neural Networks. Recently, Recurrent Neural Networks (RNN) have shown notable success in modeling data sequences and time-varying patterns. They organize in a recursively unfolded structure, which makes them a perfect choice for temporal analysis and sequence learning tasks, such as machine translation and human motion forecasting [21, 22, 23, 24, 25]. Tree-structured RNN [24], illustrated spatio-temporal network organization analogous to [11]. However, their spatio-temporal architecture was designed around a skeletal-based human activity prediction such that, all the units had fixed dependencies and belong to one cohesive movement. This prior assumption does not fit with highly dynamic contexts such as crowd motion.

Few models [11, 26, 27] structured RNN units based on graph topology that explicitly represented elements and their interactions semantics. In our paper, we extend the generic spatio-temporal graph used in [11] in a hybrid manner, by combining globally-based human-human interaction with locality-based human-space interaction, in addition to using attention mechanism to distinctively model social interactions.

3. Approach

3.1. Problem Definition

Given a set of static objects O, and a set of pedestrians V and their trajectories $X^t_{v_i}$ observed at time-steps t = 1,...,T_{obs}, our model predicts the future locations $\hat{X}^t_{v_i}$ at t = $T_{obs}+1, ..., T_{pred}$ time-steps, with regards to potential influence of any obstacles presented in the scene, such that T_{obs} = 8, $v_i \in V$, T_{pred} = 12.

3.2. Model Architecture

The spatio-temporal graph is a dynamic structure that evolves temporally and spatially, due to the motion state of the pedestrians and changes in the scene (e.g. as the elements in the scene increase/decrease). Figure 2 shows the corresponding representation of crowd subjects in spatio-temporal graphs $G = (V, \Sigma_S, \Sigma_T)$, comprising three key components: nodes set $V*$, spatial edges set Σ_S and temporal edges set Σ_T, where nodes represent the dynamic and the static element (e.g. pedestrians and static objects), spatial edges represent the relationship between two nodes to indicate the interaction between them. Temporal edges link the same pedestrian node over successive time-steps and thus connect the graph when it is unrolled over time.

Figure 2a illustrates the dynamic structure with an arbitrary crowd at two consecutive time-steps. At (t=1), there are four pedestrians. At (t=2), a new pedestrian (5) enters the scene. Notice that by (t=2), pedestrian (2) enters the vicinity of the red obstacle, where they appear to pass through the dashed circular boundary. Figure 2b shows the corresponding spatio-temporal graph representation, which evolves dynamically over the spatial and temporal domain. This is evident when the graph unfolds at (t=2), where a new node is introduced for pedestrian (5) and all

pedestrian nodes are connected by undirected edges to model the mutual interaction. This creates 2(N-1) spatial edges between pedestrian nodes at every time-step, where N is the number of pedestrians. In contrast, only a single directed edge is pointing from the obstacle node to the corresponding pedestrian node to depict the influence posed by the static obstacle on pedestrian (2).

The components of graph G are replaced with the corresponding LSTM components, *temporal edgeLSTM, spatial edgeLSTM, nodeLSTM*. The relationship between two nodes is characterized by their relative coordinates, where $x_{v_2v_3}$ is the spatial distance between nodes v_2 and v_3, and $x_{v_2v_2}$ is location of node v that changes over time.

Eq. (1) defines *spatial edgeLSTM* embedding function ϕ that takes as input: $x^t_{v_2.}$, all the relative spatial distances between node x_{v_2} and its neighbors (e.g. including $x_{v_2v_3}$), embedding weight matrix W_s .

$$e^t_{v_2.} = \phi(x^t_{v_2.}; W_s) \tag{1}$$

The *spatial edgeLSTM*s take the embedded input feature along with previous spatial hidden states from all related nodes $h^{t-1}_{v_2.}$ and transform them using normally initialized weight matrix W^{lstm}_s. The output hidden states vector $h^t_{v_2.}$ is shown in Eq. (2).

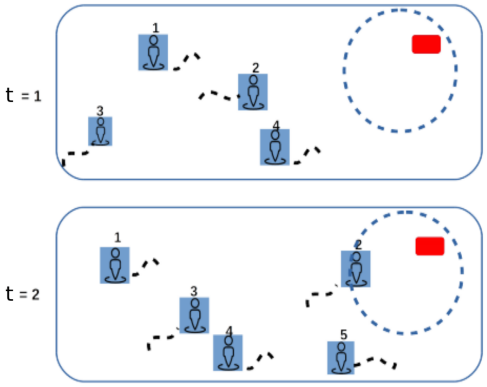
$$h^t_{v_2.} = LSTM(h^{t-1}_{v_2.}, e^t_{v_2.}, W^{lstm}_s) \tag{2}$$

Eq. (3) defines *temporal edgeLSTM* embedding function ϕ that takes as input: the temporal location of pedestrian node $x^t_{v_2v_2}$, embedding weight matrix W_t.

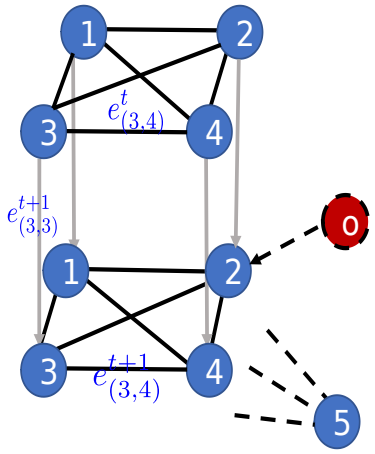
$$e^t_{v_2v_2} = \phi(x^t_{v_2v_2}; W_t) \tag{3}$$

Eq. (4) defines the LSTM cell and its inputs: previous temporal hidden state $h^{t-1}_{v_2v_2}$, embedded input feature $e^t_{v_2v_2}$ from Eq. (3) and normally initialized weight matrix W^{lstm}_t for transforming these inputs into the current hidden state $h^t_{v_2v_2}$.

$$h^t_{v_2v_2} = LSTM(h^{t-1}_{v_2v_2}, e^t_{v_2v_2}, W^{lstm}_t) \tag{4}$$



(a) Crowded environment displayed over 2 time-steps.



(b) Crowd mapping to abstract spatio-temporal graph unrolled through two time-steps

Figure 2: Crowd mapping to Spatio-temporal Graph. (a) A static obstacle is drawn as red rectangle surrounded by a virtual circle which indicates its neighborhood boundaries. (b) The Blue nodes represent pedestrians 1,2,3,4,5 and the red dashed node represents obstacle o such that $o \in O$. Directed downward lines indicate temporal edges linking the same node over time-steps and undirected lines are two-way spatial edges connecting pedestrian nodes. A directed edge is pointing from Obstacle node to pedestrian node to indicate obstacle influence on pedestrian. For the sake of clarity, we use dashed links from node (5) to indicate the remaining spatial edges. (Best viewed in color).

3.3. Spatio-Temporal Attention Module

Given the success of attentional mechanisms in sequence-based prediction of natural language processing applications, this work adopts the concept of attention-based generative algorithms [13]. We propose a variation of Multi-Head method, a soft attention based on two simple operations, i.e. concatenation and averaging across all edge feature vectors for

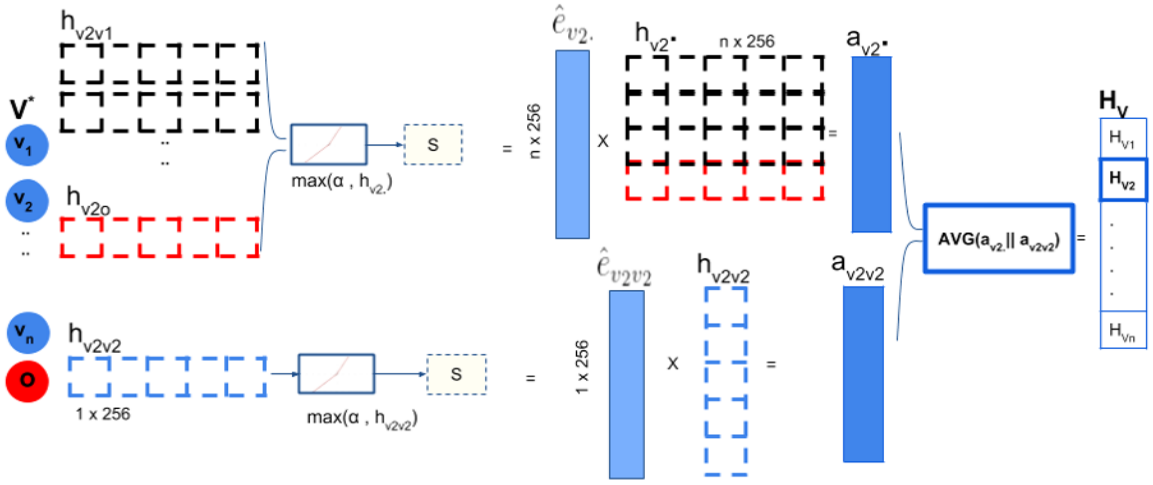


Figure 3: Multi-node attention mechanism pipeline for pedestrian node v_2 at time-step $t = 2$. On the left-most side, it shows nodes set $V^* = \{V,O\}$. The black dashed vectors store $h^{t+1}_{v2.}$, the hidden states of the spatial edges related to node v_2. The red dashed vector stores h^{t+1}_{v2o}, the hidden state of spatial edge between node v_2 and obstacle o. The blue dashed vector stores h^{t+1}_{v2v2}, temporal edge hidden state of node $v2$. These hidden states are then passed into PReLU and Softmax (S) activations to generate new embeddings $\hat{e}$. The concrete blue vectors store spatial hidden states $\hat{e}^{t+1}_{v2.}$ and temporal hidden state $\hat{e}^{t+1}_{v_2v_2}$. Multiplying the new embeddings vector by their hidden states array results in attention coefficients vectors a, where $a^{t+1}_{v2.}$ is the spatial attention coefficients vector and $a^{t+1}_{v_2v_2}$ is the temporal attention coefficients vector.

each node. In [13], the input comprises fixed number of words with fixed positions, and the Multi-Head attention works by stacking multiple attention layers (heads) in which each layer makes mappings between words in two sentences. We use a simple attention mechanism, i.e. Multi-Node attention, which only has a single layer that jointly pays attention to the features from spatial and temporal domains and store the attention coefficients into single vector for node v_2 trajectory at each time-step. To illustrate this, Figure 3 exemplifies attention on pedestrian (2) and its neighbors at time (t=2). Neighboring *edgeLSTM*s states are transformed before concatenation using the embedding function in Eq. (5) and Eq. (6), which is a composite of Parametric ReLU and softmax. This combined activation ensures that hidden states remain within a small range of [-1,1] which will be mapped once again at the sampling stage to a range of normalized outputs range of [0,1].

$$\hat{e}^t_{v_2.} = softmax(PReLU(h^t_{v_2.})) \quad (5)$$

$$\hat{e}^t_{v_2v_2} = softmax(PReLU(h^t_{v_2v_2})) \quad (6)$$

The Parametric ReLU as illustrated in Eq (7), is the generalized ReLU function as it ties the leak parameter α as a network learnable parameter. Employing such activation function with an adaptive leak parameters, allows a slightly different span of the negative hidden states along training batches. This has proved its benefit for the model prediction performance.

$$PReLU(h) = max(0,h) + \alpha * min(0,h); \quad (7)$$
$$\alpha = 0.2$$

The product of embedding vectors $\hat{e}$ with the original hidden states results in attention weights (also called coefficients). Eq. (8) and Eq. (9) shows the spatial attention coefficients $a^t_{v_2.}$ and temporal attention coefficients $a^t_{v_2v_2}$, respectively.

$$a^t_{v_2.} = \hat{e}^t_{v_2.} h^t_{v_2.} \quad (8)$$

$$a^t_{v_2v_2} = \hat{e}^t_{v_2v_2} h^t_{v_2v_2} \quad (9)$$

Eventually, these coefficients will be concatenated and averaged to generate the final weighted hidden states vector ${H_{v_2}}^t$ as shown by Eq. (10):

$$H^t_{v_2} = \frac{\sum_v^N (a^t_{v_2v_2} || a^t_{v_2.})}{N}; \; N = |\, a^t_{v_2v_2} || \, a^t_{v_2.} \,| \quad (10)$$

Comparing the Multi-Head attention with the single head multiplicative attention (scaled dot-Product), it turns out that the scaled dot-Product

gives a compact representation of all incoming hidden states and it serves a similar objective to the linear pooling mechanism in [8] due to the highly variable-sized environment. However, it diminishes the expressive power lost upon compressing feature vectors size.

While Multi-Head attention averages across the spatial and temporal attention coefficients without compressing their depth. Hence, we realized that retaining the vectors depth provides sufficient feature representation for learning the influence of pedestrians on each other.

The pedestrian location coordinates $x^t_{v_2}$ are passed through an embedding layer ϕ as in Eq. (11) before its taken as input into *nodeLSTM*:

$$e^t_{v_2} = \phi(x^t_{v_2}; W_{embed}) \tag{11}$$

Finally, the output vector ${H_{v_2}}^t$ is concatenated with previous hidden state $h^{t-1}_{v_2}$, and which are then passed to *nodeLSTM* v_2, along with transformation weight matrix W^{lstm} to generate current hidden state $h^t_{v_2}$.

$$h^t_{v_2} = LSTM(e^t_{v_2}, concat(h^{t-1}_{v_2}, {H_{v_2}}^t, e^t_{v_2}), W^{lstm}) \tag{12}$$

The future location of pedestrian is sampled from a bivariate normal distribution *N* as in Eq. (14). For estimating the Mean μ, variance σ and correlation ρ we apply a linear transformation layer in Eq. (13) W_{out} to transform $h^t_{v_2}$ into the estimated parameters.

$$(\mu^{t+1}_{v_2}, \sigma^{t+1}_{v_2}, \rho^{t+1}_{v_2}) = W_{out} h^t_{v_2} \tag{13}$$

$$(x^{t+1}_{v_2}, y^{t+1}_{v_2}) \sim N\,(\mu^{t+1}_{v_2}, \sigma^{t+1}_{v_2}, \rho^{t+1}_{v_2}) \tag{14}$$

4. Experimental Results

4.1. Datasets and Metrics

Our evaluation is based on two widely-used datasets, ETH Walking Pedestrians (EWAP) [28], UCY Students and Zara [29]. In total, the datasets consist of five videos taken from outdoor surveillance cameras. The datasets contain 2206 human trajectories, exhibiting different traits that range between straight linear and curvilinear motion splines. From our observations, ETH scenes consist of more straight trajectories with few social interactions as the video captures people motion at the university entrance, while UCY scenes display more scenarios pertaining to human-space interactions. For example, the UCY-ZARA datasets include pedestrians bending at the shop entrance, while UCY-University scenes have more social interactions among standing groups. Furthermore, these cases in particular, increase the unpredictability of an individual path unless social and spatial contexts are taken into account. In our experiments, two benchmark metrics are used, i.e. Averaged Displacement Error (*ADE*) and Final Displacement Error (*FDE*) of the TrajNet challenge [10], for measuring Euclidean deviations (*in meters*) between predicted trajectory and actual trajectory.

Averaged Displacement Error: The mean average *l2 distances* between predicted trajectory coordinates $(\hat{x}, \hat{y})$ and true trajectory (x, y) for all time-steps $i = (1, .., n)$ over N pedestrian trajectories in the scene.

$$ADE = \frac{\sum_{j=1}^{N} \frac{\sum_{i=1}^{n} \sqrt{(\hat{x}_i^j - x_i^j)^2 + (\hat{y}_i^j - y_i^j)^2}}{n}}{N} \tag{15}$$

Final Displacement Error: The average *l2 distance* between the final predict step $(\hat{x}_n, \hat{y}_n)$ and the true step (x_n, y_n) over j pedestrians trajectory, where $j = (1, ..., N)$.

$$FDE = \frac{\sum_{j=1}^{N} \sqrt{(\hat{x}_n^j - x_n^j)^2 + (\hat{y}_n^j - y_n^j)^2}}{N} \tag{16}$$

4.2. Ablation Study

We have performed an ablation study by dropping the scaled-dot attention module from Social Attention and restoring back original settings of Structural-RNN, to study the usefulness of dot-Product attention model. The comparison between the quantitative results of both baselines with our method, shows that the scaled dot-Product performance is lower than the Multi-Node mechanism performance for the 5 datasets in Table 1. On the other hand, the optimal choice of the human-obstacle connectivity threshold $\lambda = 0.5$ parameter, was determined empirically, based on the objective of lowering the Euclidean errors for both evaluation metrics.

4.3. Training Setup

We accumulated trajectory data for every pedestrian with *skip_rate* = 10 frames to avoid overfitting the minimal changes in pedestrian trajectory. Each LSTM cell is of 256 depth. We transform data into normalized interpolated pixel coordinates within

range [0,1]. In batch processing, we fixate the batch size *batch_size* = 24, observation length T_{obs} = 8 time-steps (3.2 seconds), prediction length T_{pred} = 12 time-steps (4.8 seconds) and epochs *epoch_num* = 100. After several hyper-parameter tunings, learning rate is set as *lr* = 0.001 and optimizer algorithm is Adam. Activation function in attention layer is Parametric ReLU, initialized to negative slope α = 0.20 and fractionally degraded throughout the training process. The training objective is to minimize the negative log-likelihood loss of the i^{th} trajectory from time-step T_{obs+1} to T_{pred}:

$$L_i = \quad - \sum_{t=T_{obs}+1}^{T_{pred}} log(P(x_i^t, y_i^t | \sigma_i^t, \mu_i^t, \rho_i^t)) \quad (17)$$

4.4. Quantitative Results

As illustrated in Table 1, we set up experiments to evaluate our proposed models, H-H and H-H-O, which stand for Human-Human and Human-Human-Obstacle respectively. The table has two segments, the first 4 rows evaluate our model with graph-based baselines: Social Attention and Structural-RNN, while the next 3 rows evaluate our model with state-of-the-art models: Social-LSTM and Social GAN (SGAN). Our attention mechanism for graphs improved prediction for human-human interaction and human-obstacle interaction over the other graph-based baselines: Social Attention and Structural-RNN. This is observed from the average errors under column (AVG) in Table 1. Comparing H-H-O with Social Attention, H-H-O achieves 55% in the average of ADE and 61% in the average of FDE in all datasets. As Social GAN and Social-LSTM display the best trajectories produced by their models, we extracted the average of minimum errors pertaining only to the best predicted trajectories in H-H-O model. It can be observed that the minimum FDE is considerably lower than minimum FDE generated by SGAN model and Social LSTM, due to our model awareness of surrounding context. This has made predictions to be plausible and compliant with the environmental constraint. The Social GAN work shows several versions of their model, so we selected their best model version which is SGAN-20V for our comparison. The most significant improvement is realized when comparing our model with Social GAN model, under the Hotel set with 93% reduction in FDE. Furthermore, the Hotel scene contains more static elements such as trees and lamp posts as indicated in Figure 2. The second best improvement is realized when comparing our model with Social-LSTM model under the ETH set with 89% reduction in FDE. The ETH dataset consists of a set of tightly coupled trajectories due to the crowd at the university entrance. This is a busy contextual point where pedestrians are mostly concerned about avoiding collisions with each others at the entrance site. Additionally, our model performance yields 69% reduction on FDE metric in Ucy-University, which proves that embedding information about physical structure of the scene and busy interaction points, refines the model understanding of pedestrian navigation in crowded sites and reduced the prediction errors in FDE, as our model was more capable of predicting the final step on a pedestrian trajectory. From the previous table, it is noticeable that the ADE and FDE exhibit small discrepancies due to the accumulative nature in prediction errors. If the predicted path was entirely approximate to the ground-truth, the final predicted point will not have large error, but if the prediction was increasingly deviating along the ground-truth, this can impact the final point errors. This supports our quantitative results as being consistent and realistic.

4.5. Qualitative Results

In this section, we qualitatively evaluate model predictions in Hotel and ZARA sets. Figure 4 displays predicted paths from our models. We have spotted interesting cases for pedestrian moving near static objects, and compared both of our models outputs, Social Attention and Social LSTM with ground-truth trajectory. Notice the Human-Human model prediction for pedestrian walking near the bench in Figure 4a. The ground-truth shows that pedestrian is avoiding the bench, while Human-Human model spline achieves lower displacement than the baseline splines, those fail at evading the bench area. This case is correctly predicted in our Human-Human-Obstacle model as illustrated in Figure 4b. Additionally, Figure 4d shows that Social Attention and Social LSTM predicts plausible paths that pedestrian might have chosen, however, it is not compliant with pedestrian surrounding objects. Thus, with the aid of obstacle awareness, our model understands pattern of collision avoidance with any static subject in their way.

Figure 4c plots trajectories from H-H-O model where pedestrians are bending toward the shop en-

Table 1: Prediction errors ADE/FDE (in meters). Our results are averaged over 30 sampled sequences of 12-steps length for every set under our method. For baselines errors, Social Attention results are obtained upon re-training their model, while Structural-RNN results are obtained upon manual implementation of their architecture in PyTorch.

Method	ETH	HOTEL	ZARA1	ZARA2	UNIV	AVG
Structural-RNN	2.72/4.60	0.85/1.35	1.05/2.20	1.60/3.50	1.45/3.00	1.53/2.93
Social Attention	3.60/4.70	0.79/1.44	1.30/2.66	0.95/2.05	1.00/2.14	1.53/3.52
H-H	**1.19/2.00**	**0.39**/0.96	0.55/1.56	0.58/1.50	0.74/1.89	**0.69**/1.58
H-H-O	1.24/2.35	0.48/**0.80**	**0.51/1.15**	**0.56/1.13**	**0.69/1.45**	0.70/**1.38**
Social LSTM	1.09/2.35	0.79/1.76	0.47/1.00	0.56/1.17	0.67/1.40	0.72/1.54
SGAN-20V	**0.81**/1.52	0.72/1.61	**0.34**/0.69	**0.42**/0.84	0.60/1.26	**0.58**/1.18
Minimum H-H-O	0.96/**0.16**	**0.35/0.11**	0.57/**0.30**	0.58/**0.33**	**0.53/0.38**	0.60/**0.26**

trance, and our model generates splines that approximate the curvy ground-truth trajectory, as the model learns the motion pattern at the entrance point.

In some situations, the predictions do not perfectly match the ground-truth path, although the deviations are quite small. This situation also applies for the baseline models. Upon extensive visual comparisons for all frames in all datasets, we confirmed that the erroneous results and deviations of the proposed method are much fewer than those found in the baselines plots. Quantitatively, Euclidean deviations at the path endings have been reduced by up to 61%, which identifies the improvements that we highlighted earlier.

5. Conclusion

In this paper we have presented a new spatio-temporal graph that operates on the local and global contexts around pedestrian, for predicting their trajectory in outdoor environments. For an accurate modeling of human-human interactions and human-space interactions, we employ a simplified version of Multi-Head attention mechanism for accumulating the influence from spatial and temporal subspaces. Our attention mechanism consistently demonstrated improved prediction results over baseline methods, for groups as well as individual non-linear trajectories.

References

[1] H. Kretzschmar, M. Kuderer, and W. Burgard, "Learning to predict trajectories of cooperatively navigating agents," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4015–4020. 1

[2] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016. 1

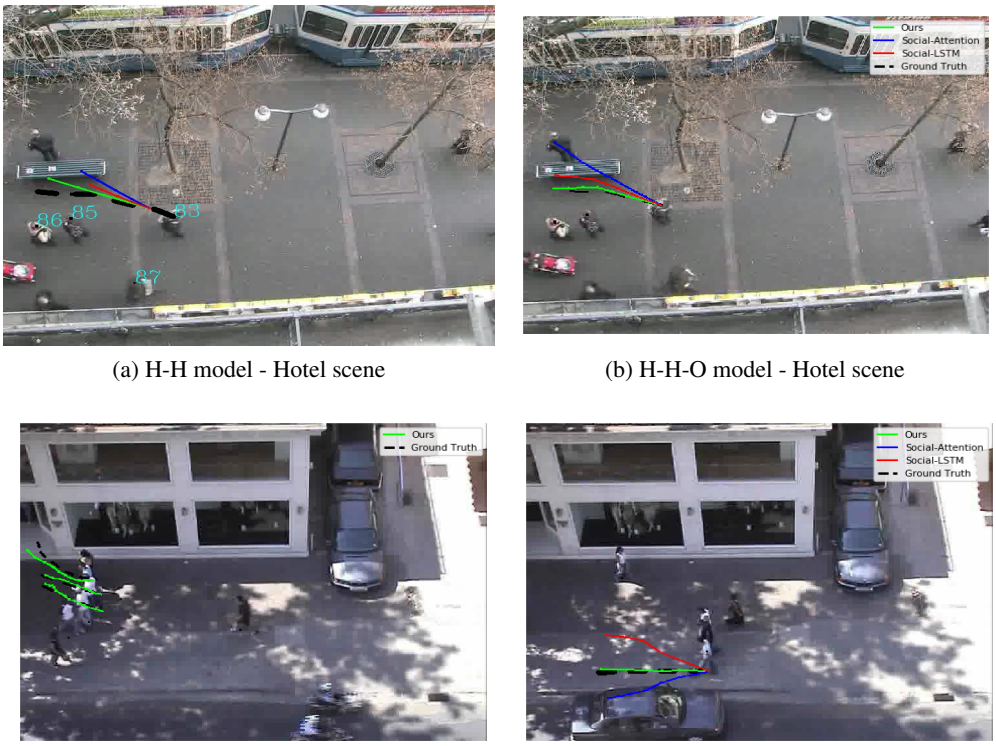
[3] F. Bartoli, G. Lisanti, L. Ballan, and A. Del Bimbo, "Context-aware trajectory prediction," *arXiv preprint arXiv:1705.02503*, 2017. 1, 2

[4] D. Ellis, E. Sommerlade, and I. Reid, "Modelling pedestrian trajectory patterns with gaussian processes," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1229–1234. 1

[5] K. Kim, D. Lee, and I. Essa, "Gaussian process regression flow for analysis of motion trajectories," in *Computer vision (ICCV), 2011 IEEE international conference on*. IEEE, 2011, pp. 1164–1171. 1

[6] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *European conference on computer vision*. Springer, 2016, pp. 549–565. 2

[7] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds,"



(a) H-H model - Hotel scene

(b) H-H-O model - Hotel scene

(c) H-H-O model - ZARA scene

(d) H-H-O model - ZARA scene

Figure 4: Visualization results for Hotel and Zara sets.

in *Proceedings of the International Conference on Robotics and Automation (ICRA) 2018*, May 2018. 2

[8] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 961–971. 2, 6

[9] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, no. CONF, 2018. 2, 3

[10] A. Sadeghian, V. Kosaraju, A. Gupta, S. Savarese, and A. Alahi, "Trajnet: Towards a benchmark for human trajectory prediction," *arXiv preprint*, 2018. 2, 6

[11] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5308–5317. 2, 3

[12] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *stat*, vol. 1050, p. 20, 2017. 2, 3

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010. 2, 4, 5

[14] S. van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber, "Relational neural expectation maximization: Unsupervised discovery of objects and their interactions," *arXiv preprint arXiv:1802.10353*, 2018. 2, 3

[15] P. Battaglia, R. Pascanu, M. Lai, D. J. Rezende *et al.*, "Interaction networks for learning about objects, relations and physics," in *Advances in neural information processing systems*, 2016, pp. 4502–4510. 2, 3

[16] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection," *Neural networks*, vol. 108, pp. 466–478, 2018. 2, 3

[17] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995. 2

[18] D. Varshneya and G. Srinivasaraghavan, "Human trajectory prediction using spatially aware deep attention models," *arXiv preprint arXiv:1705.09436*, 2017. 3

[19] 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," 2017. 3

[20] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009. 3

[21] 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. 3

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014. 3

[23] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data." in *AAAI*, vol. 1, no. 2, 2017, p. 7. 3

[24] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–833. 3

[25] H. Xue, D. Q. Huynh, and M. Reynolds, "Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1186–1194. 3

[26] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *European Conference on Computer Vision*. Springer, 2016, pp. 125–143. 3

[27] Y. Yuan, X. Liang, X. Wang, D. Y. Yeung, and A. Gupta, "Temporal dynamic graph lstm for action-driven video object detection," *arXiv preprint arXiv:1708.00666*, 2017. 3

[28] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 261–268. 6

[29] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer Graphics Forum*, vol. 26, no. 3. Wiley Online Library, 2007, pp. 655–664. 6