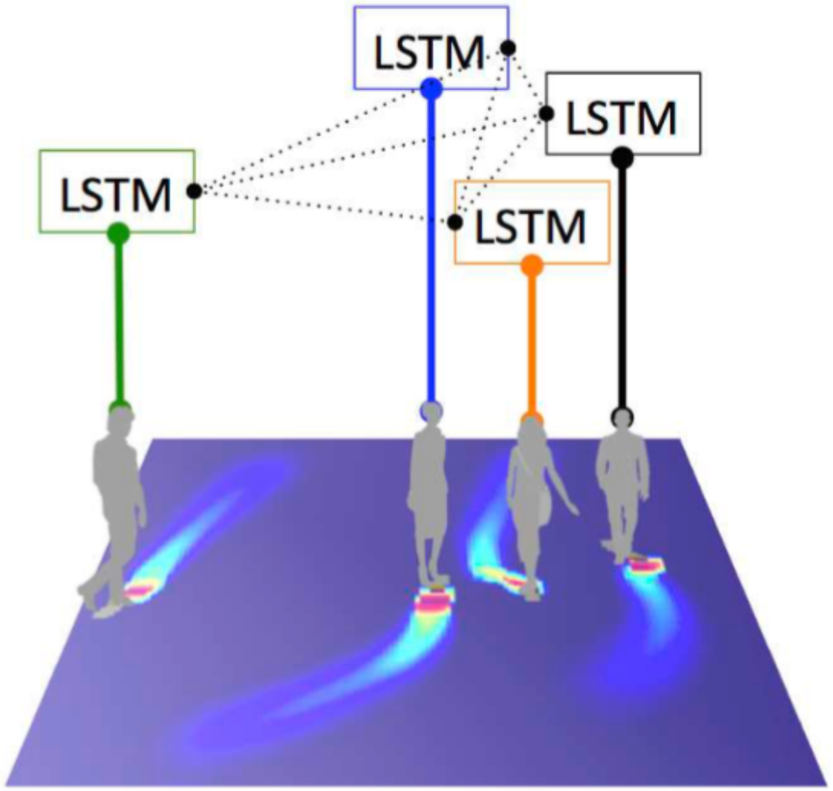
Paper Reading

Social LSTM: Human Trajectory Prediction in Crowded Spaces



Alexandre Alahi*, Kratarth Goel*, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, Silvio Savarese.
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. (equal contribution)*
(Spotlight oral)

*** Supported by funding from the Stanford AI Lab-Toyota Center for Artificial Intelligence Research and the ONR sparse grant**

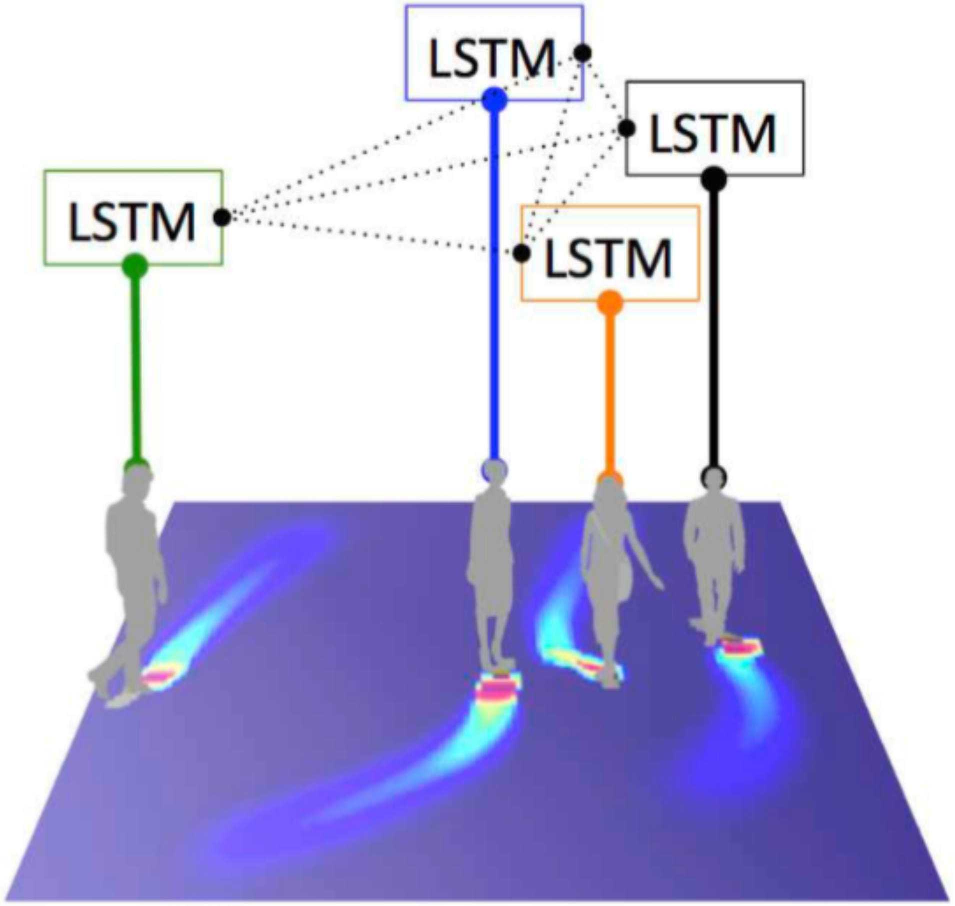
Contents

1. Introduction
2. Related work
3. Proposed method
4. Experiments
5. Conclusions

Introduction

Background

Pedestrians follow different trajectories to avoid obstacles and accommodate fellow pedestrians. Any autonomous vehicle navigating such a scene should be able to foresee the future positions of pedestrians and accordingly adjust its path to avoid collisions.



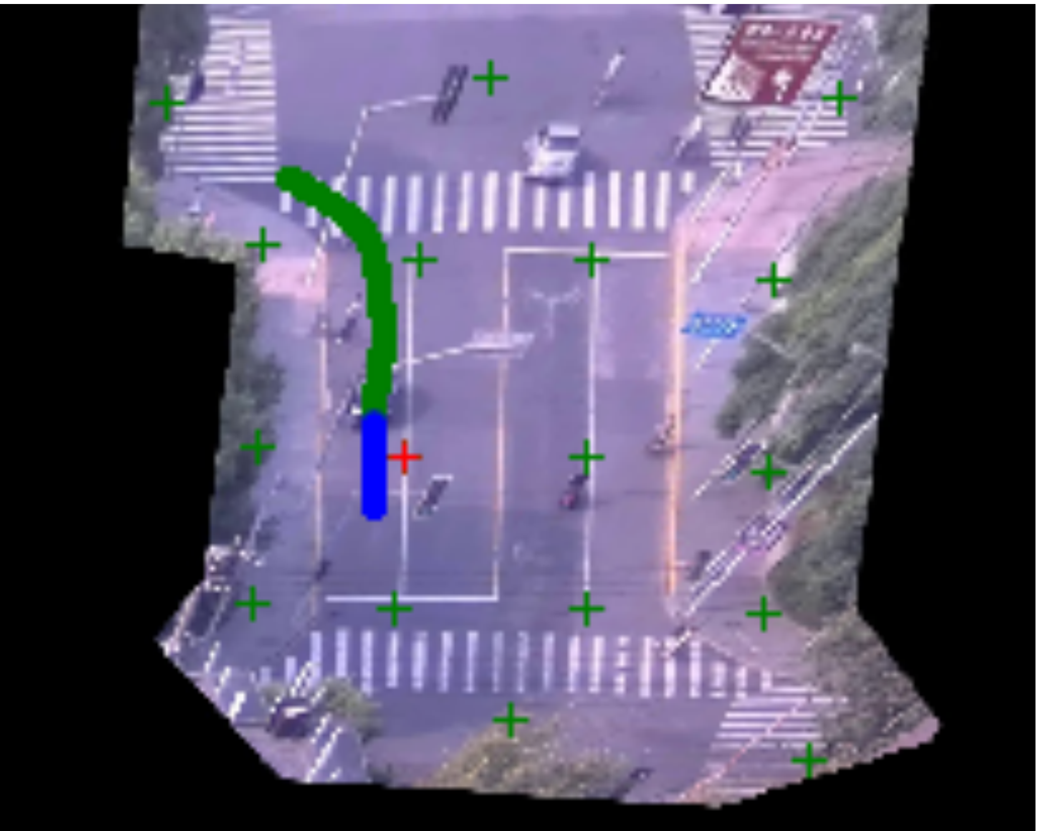
Introduction

Problem Description

This problem of trajectory prediction can be viewed as a sequence generation task, where we are interested in predicting the future trajectory of people based on their past positions.

Contribution

The goal of this paper is to predict the motion dynamics in crowded scenes - This is, however, a challenging task as the motion of each person is typically affected by their neighbors.



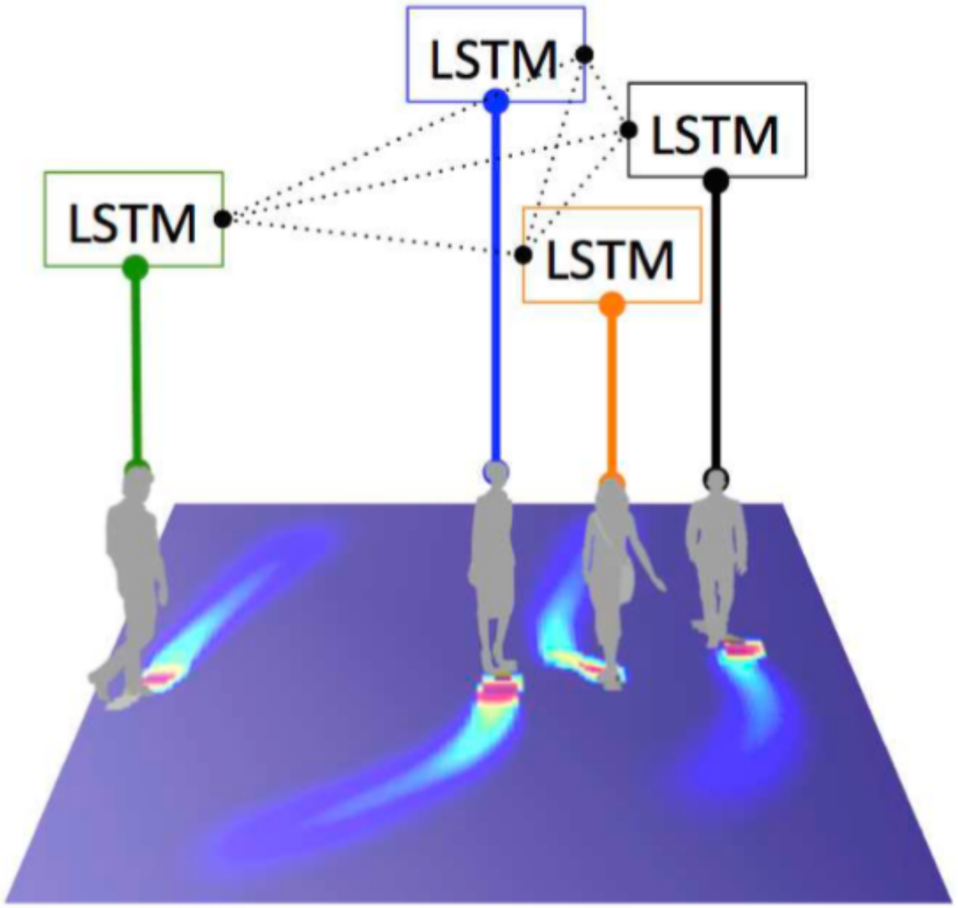
Introduction

Challenging

This requires understanding the complex and often subtle interactions that take place between people in crowded spaces.

Proposed method

We propose a new model which we call ”Social” LSTM (Social- LSTM) which can jointly predict the paths of all the people in a scene by taking into account the common sense rules and social conventions that humans typically utilize as they navigate in shared environments. The predicted distribution of their future trajectories is shown in the heat-map



Introduction

Pioneering works

Kitani et. al. [1] have demonstrated that the inferred knowledge about the semantics of the static environment (e.g., *location of sidewalks, extension of grass areas*, etc) helps predict the trajectory of pedestrians in future instants more accurately than a model which ignores the scene information.

Previous works by [2, 3, 4] have also proposed ways to model human-human interactions (often called ”social forces”) to increase robustness and accuracy in multi-target tracking problems.

[1] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In Computer Vision–ECCV 2012, pages 201–214. Springer, 2012.

[2] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. Physical review E, 51(5):4282, 1995.

[3] H. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. 2013.

[4] S. Pellegrini, A. Ess, and L. Van Gool. Improving data as- sociation by joint modeling of pedestrian trajectories and groupings. InComputerVision–ECCV2010, pages452–465. Springer, 2010.

Introduction

Motivation

Most of previous works are limited by the following two assumptions.

i) They use hand-crafted functions to model ”interactions” for specific settings rather than inferring them in a data-driven fashion. This results in favoring models that capture simple interactions (e.g. repulsion/attractions) and might fail to generalize for more complex crowded settings.

ii) They focus on modeling interactions among people in close proximity to each other (to avoid immediate collisions). However, they do not anticipate interactions that could occur in the more distant future.

Related work

Human-human interactions

Most of these models provide hand-crafted energy potentials based on relative distances and rules for specific scenes. In contrast, we propose a method to learn human- human interactions in a more generic data-driven fashion.

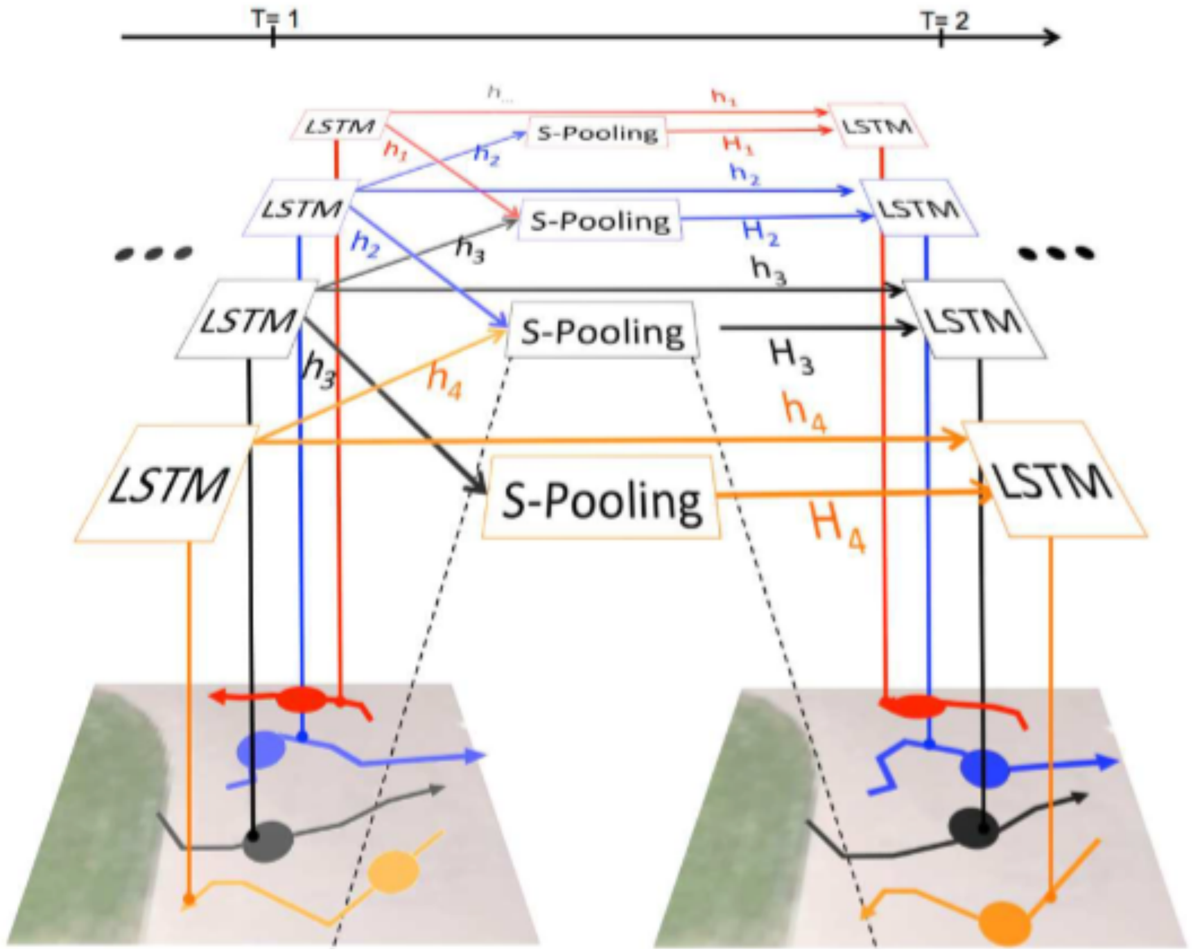
Activity forecasting

RNN models for sequence prediction

[1] Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
[2] Bojarski, Mariusz, Philip Yeres, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Lawrence Jackel, and Urs Muller. "Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car." *arXiv preprint arXiv: 1704.07911* (2017).

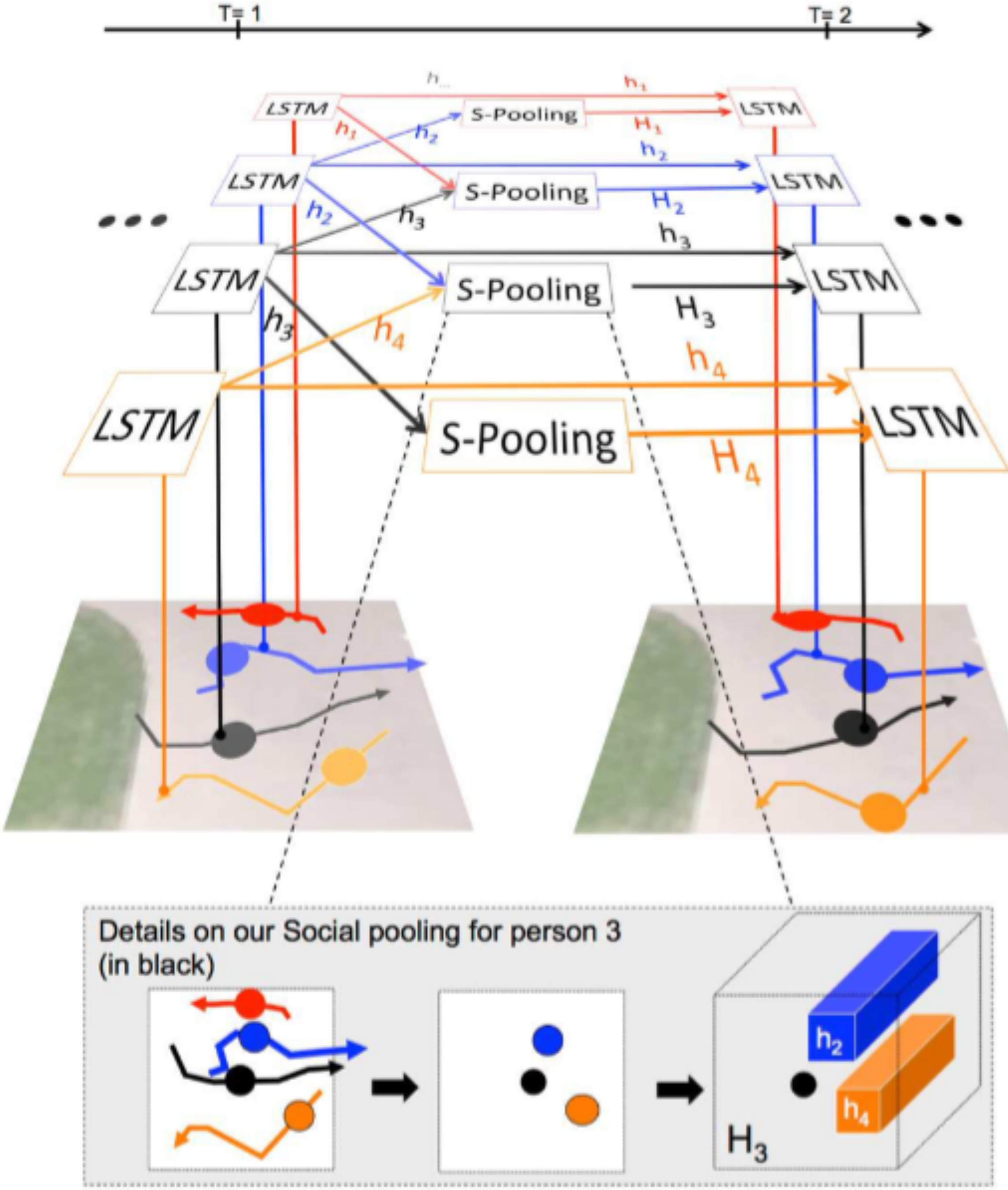
Proposed method

A model which can account for the behavior of other people within a large neighborhood, while predicting a person's path.



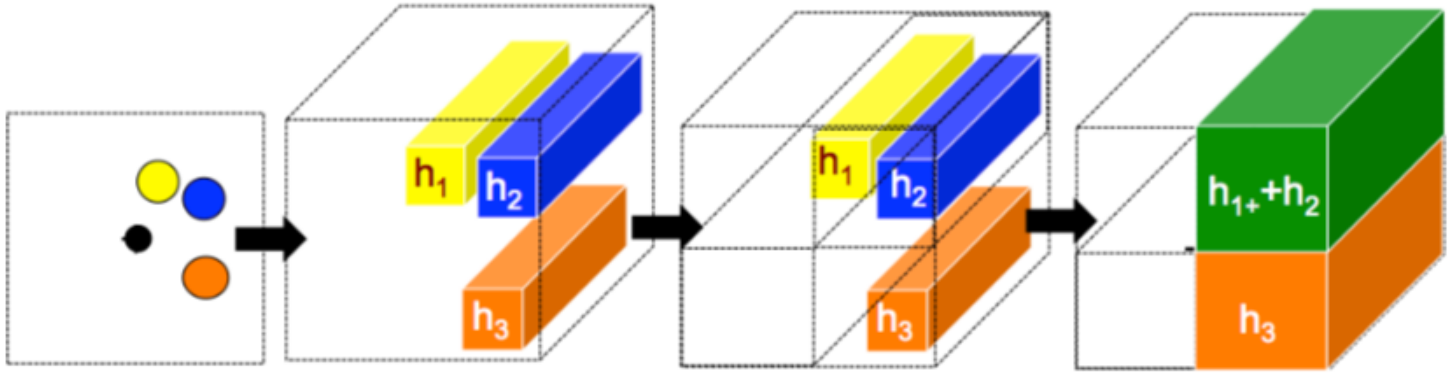
Proposed method

Social pooling of hidden states



Details on our Social pooling for person 3 (in black)

$$H_t^i(m,n,:) = \sum_{j\in\mathcal{N}_i} \mathbf{1}_{mn}[x_t^j - x_t^i, y_t^j - y_t^i] h_{t-1}^j$$



Proposed method

Position estimation

$$L^i(W_e, W_l, W_p) \quad = \quad - \sum_{t=T_{obs}+1}^{T_{pred}} \log\left(\mathbb{P}(x_t^i, y_t^i | \sigma_t^i, \mu_t^i, \rho_t^i)\right)$$

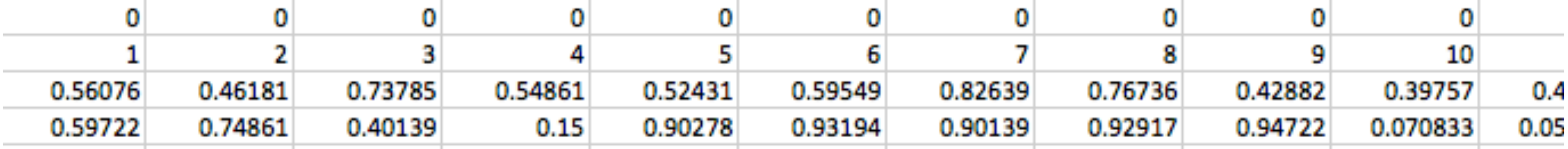
Occupancy map pooling (referred to as O-LSTM in the experiments)

$$O_t^i(m,n) = \sum_{j \in \mathcal{N}_i} \mathbf{1}_{mn}[x_t^j - x_t^i, y_t^j - y_t^i]$$

Experiments

Datasets

We present experiments on two publicly available human-trajectory datasets: *ETH* and *UCY*.



0	0	0	0	0	0	0	0	0	0	
1	2	3	4	5	6	7	8	9	10	
0.56076	0.46181	0.73785	0.54861	0.52431	0.59549	0.82639	0.76736	0.42882	0.39757	0.4
0.59722	0.74861	0.40139	0.15	0.90278	0.93194	0.90139	0.92917	0.94722	0.070833	0.05

Metrics

Average displacement error - The mean square error (MSE) over all estimated points of a trajectory and the true points.

Final displacement error - The distance between the predicted final destination and the true final destination at end of the prediction period

Average non-linear displacement error - The is the MSE at the non-linear regions of a trajectory.

Experiments

Validation

Leave-one-out approach

Prediction

We observe a trajectory for 3.2secs and predict their paths for the next 4.8secs. At a frame rate of 0.4, this corresponds to observing 8 frames and predicting for the next 12 frames.

Experiments

Comparison

- *Linear model (**Lin.**)* We use an off-the-shelf Kalman filter to extrapolate trajectories with assumption of linear acceleration.
- *Collision avoidance (**LTA**)*. We report the results of a simplified version of the Social Force [1] model which only uses the collision avoidance energy, commonly referred to as linear trajectory avoidance.
- *Social force (**SF**)*. We use the implementation of the Social Force model from [1] where several factors such as group affinity and predicted destinations have been modeled.
- *Iterative Gaussian Process (**IGP**)*. We use the imple- mentation of the IGP from [2]. Unlike the other base- lines, IGP also uses additional information about the final destination of a person.

[1] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1345–1352. IEEE, 2011.

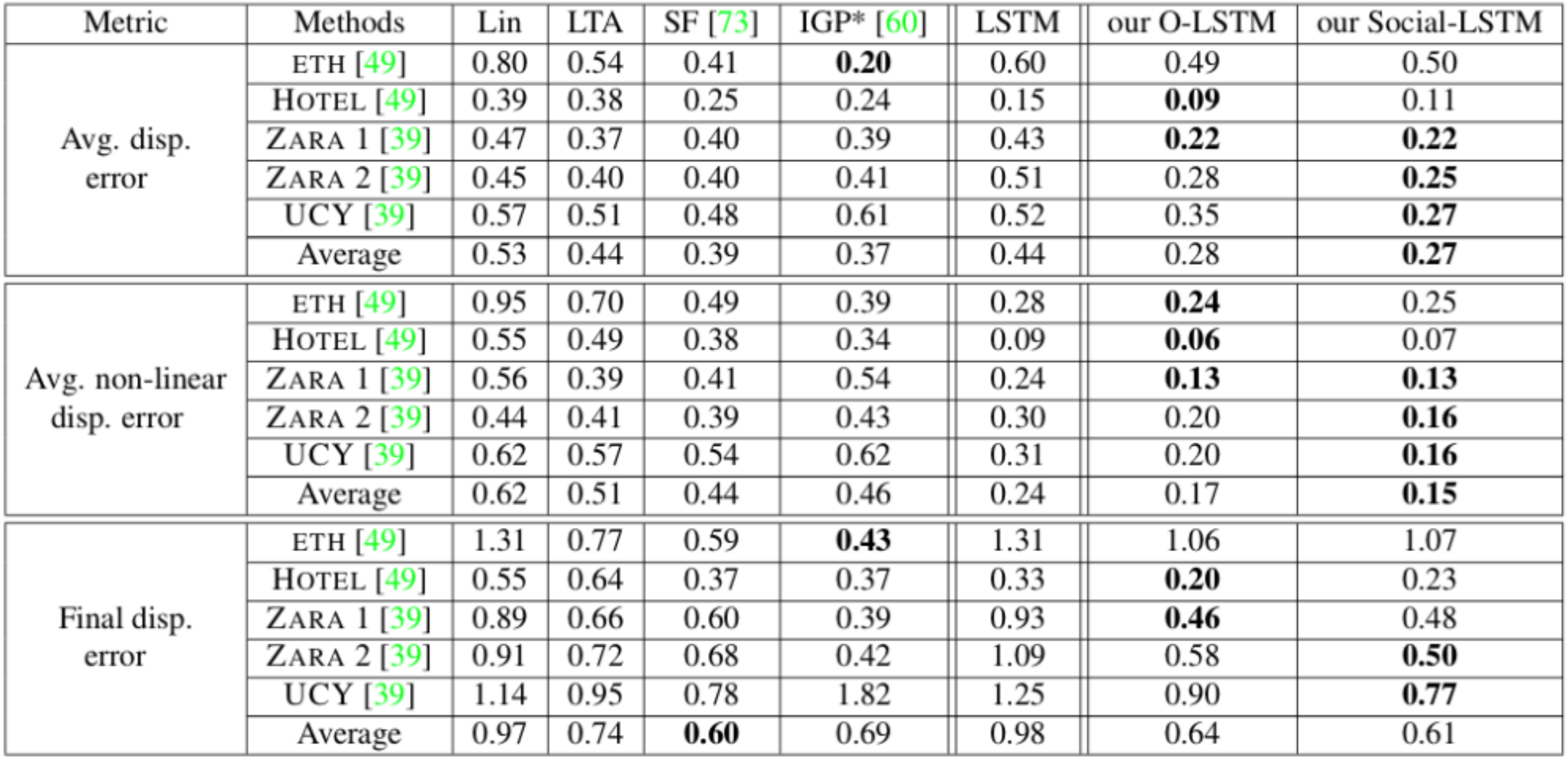
[2] P.Trautman, J.Ma, R.M.Murray, andA.Krause. Robotnav- igation in dense human crowds: the case for cooperation. In Robotics and Automation (ICRA), 2013 IEEE International Conference on, pages 2153–2160. IEEE, 2013.

Experiments

Comparison

- *Our Vanilla LSTM (**LSTM**)*. This is a simplified setting of our model where we remove the “Social” pooling layers and treat all the trajectories to be independent of each other.
- *Our LSTM with occupancy maps (**O-LSTM**)*. We show the performance of a simplified version of our model. As a reminder, the model only pools the coordinates of the neighbors at every time- instance.

Quantitative Evaluation



Metric	Methods	Lin	LTA	SF [73]	IGP* [60]	LSTM	our O-LSTM	our Social-LSTM
Avg. disp. error	ETH [49]	0.80	0.54	0.41	**0.20**	0.60	0.49	0.50
	Hotel [49]	0.39	0.38	0.25	0.24	0.15	**0.09**	0.11
	Zara 1 [39]	0.47	0.37	0.40	0.39	0.43	**0.22**	**0.22**
	Zara 2 [39]	0.45	0.40	0.40	0.41	0.51	0.28	**0.25**
	UCY [39]	0.57	0.51	0.48	0.61	0.52	0.35	**0.27**
	Average	0.53	0.44	0.39	0.37	0.44	0.28	**0.27**
Avg. non-linear disp. error	ETH [49]	0.95	0.70	0.49	0.39	0.28	**0.24**	0.25
	Hotel [49]	0.55	0.49	0.38	0.34	0.09	**0.06**	0.07
	Zara 1 [39]	0.56	0.39	0.41	0.54	0.24	**0.13**	**0.13**
	Zara 2 [39]	0.44	0.41	0.39	0.43	0.30	0.20	**0.16**
	UCY [39]	0.62	0.57	0.54	0.62	0.31	0.20	**0.16**
	Average	0.62	0.51	0.44	0.46	0.24	0.17	**0.15**
Final disp. error	ETH [49]	1.31	0.77	0.59	**0.43**	1.31	1.06	1.07
	Hotel [49]	0.55	0.64	0.37	0.37	0.33	**0.20**	0.23
	Zara 1 [39]	0.89	0.66	0.60	0.39	0.93	**0.46**	0.48
	Zara 2 [39]	0.91	0.72	0.68	0.42	1.09	0.58	**0.50**
	UCY [39]	1.14	0.95	0.78	1.82	1.25	0.90	**0.77**
	Average	0.97	0.74	**0.60**	0.69	0.98	0.64	0.61

Experiments

Conclusions of Quantitative Evaluation

The naive **linear model** produces high prediction errors, which are more pronounced around non-linear regions as seen from the average non-linear displacement error.

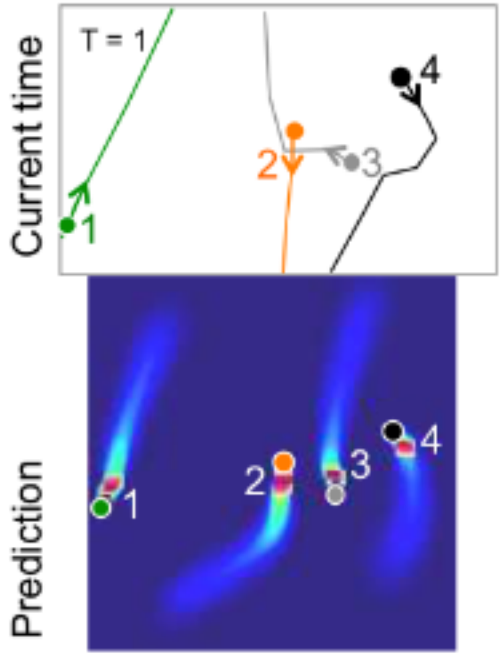
The **vanilla LSTM** outperforms this linear baseline since it can extrapolate non-linear curves as shown in Graves et al.a[1].

However, this simple LSTM is noticeably worse than the **Social Force** and **IGP** models which explicitly model human-human interactions. This shows the need to account for such interactions.

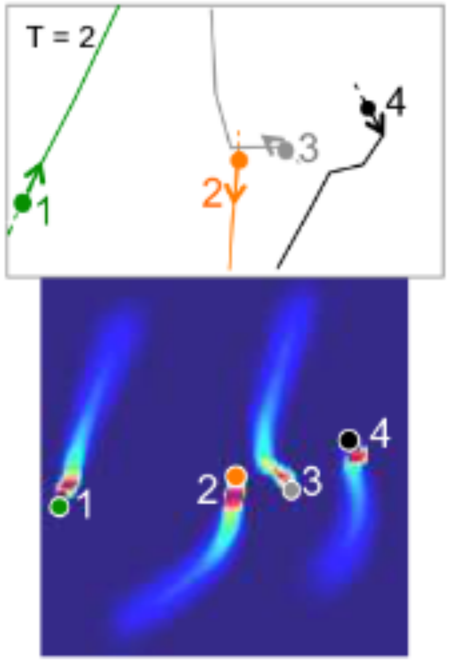
[1] A. Graves. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850, 2013.

Experiments

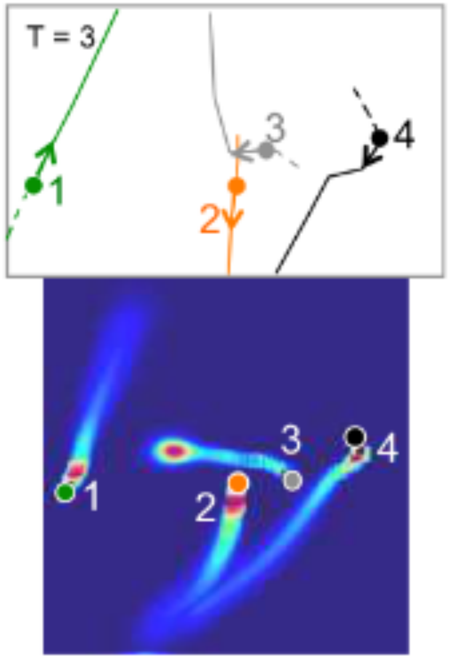
Analyzing the predicted paths



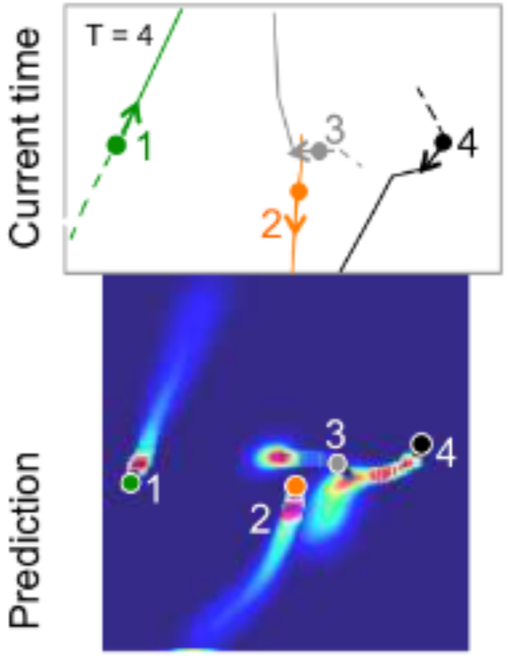
- 1 will go straight, and
- 2,3,4 will interact



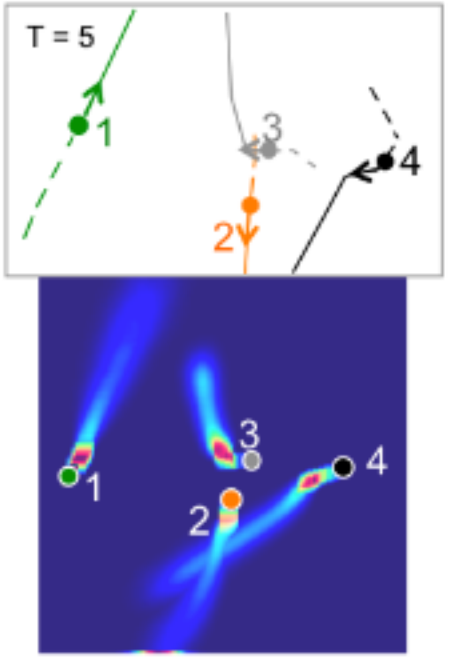
- 3 will turn to avoid 1,
- 4 is turning around 3



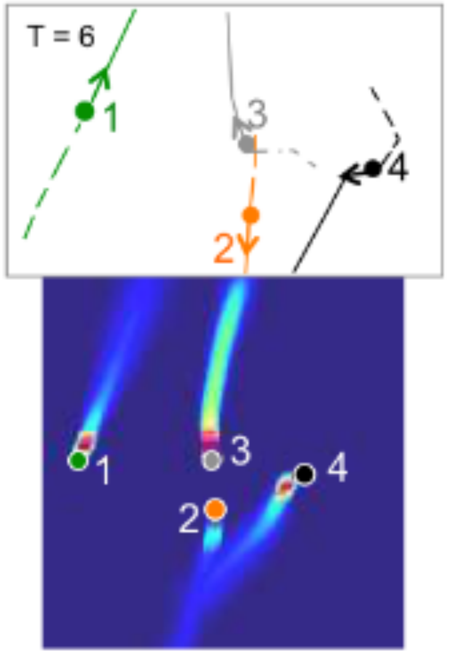
- 3 will stop in front 1,
- 4 updated the turn around 3



- 4 will decelerate and turn around 3



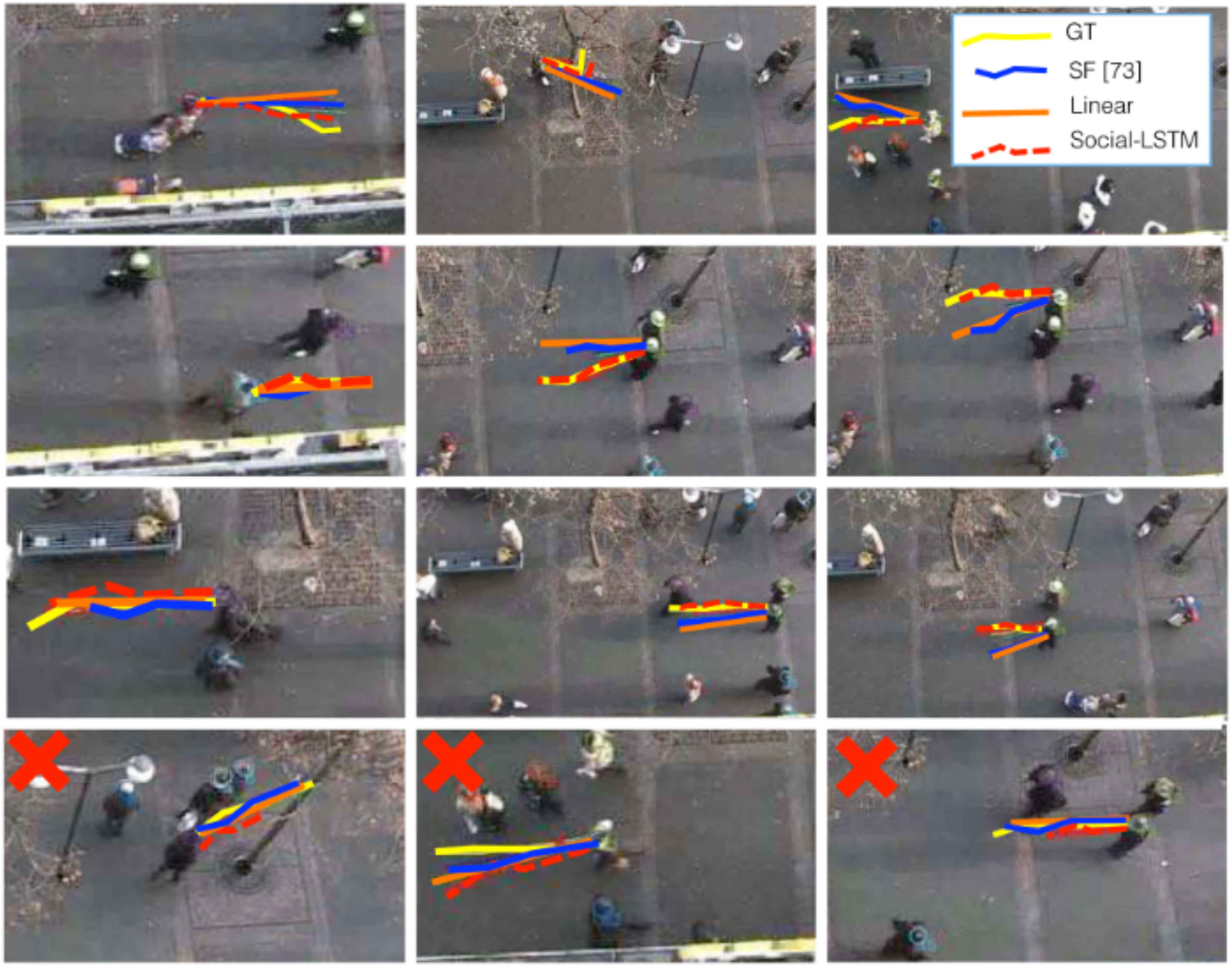
- 3 will turn towards his/her destination



- 1,2,3 and 4 will have linear path

Experiments

Analyzing the predicted paths



Conclusions

We have presented a LSTM-based model that can jointly reason across **multiple individuals** to predict human trajectories in a scene.

We use one LSTM for each trajectory and share the information between the LSTMs through the introduction of a new **Social pooling layer**.

Note that our “Social” pooling layer does not introduce any additional parameters.