Supplementary Material: Dynamic and Static Context-aware LSTM for Multi-agent Motion Prediction

Chaofan Tao$^{1,2[0000-0002-6093-0854]}$, Qinhong Jiang$^{3[0000-0002-5509-7247]}$, Lixin Duan$^{2[0000-0002-0723-4016]}$, and Ping Luo$^{1[0000-0002-6685-7950]}$

[1] The University of Hong Kong, Hong Kong, China
[2] University of Electronic Science and Technology of China, China
{tcftrees,lxduan}@gmail.com
{pluo}@cs.hku.hk
[3] SenseTime, China
{jiangqinhong}@sensetime.com

1 Implementation Details for DSCMP

We conduct experiments on three commonly used dataset (ETH [6], UCY [5], SDD [7]). Both the dataset ETH and UCY involves a single class of agent, pedestrian, in the crowd. For fair comparison with other state-of-the-art methods [1, 2, 10, 8, 4], we adopt the same time duration of observation (3.2s) and prediction (4.8s) in these two datasets. Hence, the results on the datasets ETH and UCY are reported together.

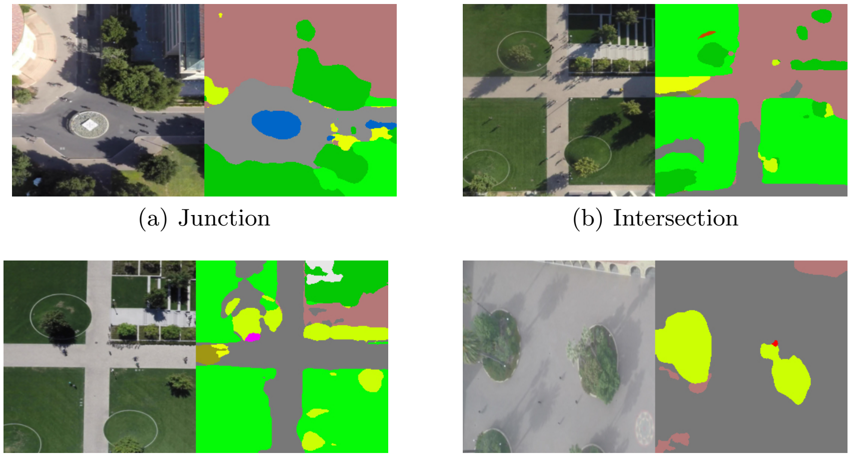
There are two kinds of mode to process the input location in the **prediction phase**. (mode-a) The ground truth at current frame is used as the current input. (mode-b) The predictions generated by the previous frame is used as the current input. During the training phase, we firstly train our model in mode-a for 200 epoches, and then employ mode-b for 150 epoches. During the testing phase, only the mode-b is used since the ground truth in the prediction phase is unavailable.

Instead of mapping the Cartesian coordinates to high dimensional vectors before sending to LSTM like many methods, we directly use the relative coordinates as input since we found it is experimentally efficient. During the prediction phase, we forecast the displacement relative to the previous moment, and then the whole predicted trajectory is generated by adding up the displacement at the last observed location.

Although the labels of scene layout are not provided in the aforementioned datasets, we extract the visual feature by pre-trained PSPNet [11, 9] off-line as prepossessing. The semantic maps are treated as additional input to our proposed model DSCMP. As shown in Fig.1, we show some of the semantic maps extracted in the scene layout.

2 Parameters Comparison and Speed Analysis

In this section, we compare the parameter usage and inference speed with state-of-the-art methods. The inference speed is reported as the testing samples per



(a) Junction

(b) Intersection

(c) Intersection

(d) Parterre

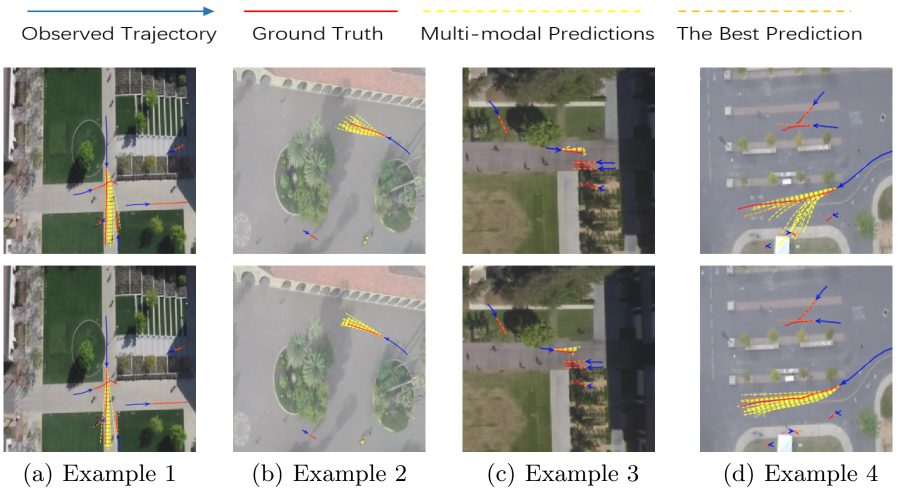
Fig. 1. Example of semantic map extracted in different scene layouts.

Table 1. Quantitative comparisons on the number of parameters and inference speed . ↓ denotes smaller value is better and vice versa. “Rel.” and “Params” is the abbreviation for “relative” and “parameters” respectively. Experiments are conducted on one NVIDIA GeForce GTX 1080 Ti graphics card.

Methods	LSTM [3]	S-LSTM [1]	SGAN [2]	STGAT [4]	**Ours**
Params ↓	14.76k	85.26k	36.39k	44.63k	**24.03k**
Rel. Params ↓	0.17x	1x	0.43x	0.52x	**0.28x**
Speed ↑	125.9	12.27	**42.5**	27.8	30.8
Rel. Speed ↑	10.26x	1x	**3.46x**	2.27x	2.51x

second. We conduct multimodal predictions for 20 times for each sample. As shown in Table 1, long short term memory (LSTM) enjoys small consumption of parameters and high inference speed. However, as discussed in our main paper, LSTM has unsatisfactory performance since it not only ignores the complicated interactions across agents in both spatial dimension and temporal dimension, but also neglects the scene information. S-LSTM [1] is set as unit 1 for comparisons among social-aware methods. The occupancy grids in the S-LSTM cause heavy usage of parameters and long inference time. SGAN [2] makes improvements with a spatial-aware pooling mechanism, which accelerates the inference time. Because STGAT [4] has a multi-head graph attention module and dual LSTM modules in the encoder, it has more consumption of parameters and time compared with SGAN.

In the DSCMP, our proposed queues introduce a few extra parameters in Individual Context Module (ICM) compared with vanilla LSTM cell. Addition-



(a) Example 1 (b) Example 2 (c) Example 3 (d) Example 4

Fig. 2. Top row: multimodal predictions generated by predefined Gaussian noises $\mathcal{N}(0, 1)$. **Bottom row:** multimodal predictions generated by our scene-guided latent variables $\mathcal{N}(\mu, \sigma)$. Zoom in three times for the best view.

ally, the parameters in the Social-aware Context Module (SCM) are shared across spatial dimension and temporal dimension. Therefore, our proposed DSCMP is lightweight in space, which is mobile-friendly for applications like self-driving vehicle and autonomous mobile robot. On the other hand, the inference speed of SGAN is fastest among its counterparts, whereas SGAN does not model the temporal dependencies in social interactions.

Followed by previous methods [2, 8, 4], we treat all agents in the same scene as neighbors for each agent. Although the importance of differnt neighbors are adaptive via Social-aware Context Module (SCM), the computational complexity for pair-wise relations is high. Detecting high-impact neighbors for each agent, and then use the high-impact neighbors only to compute social interactions are supposed to make improvements in speed.

3 Compared with Predefined Gaussian Noise

In order to obtain diverse predictions, most of the existing methods [2, 10, 8, 4] fuse the hidden features in the RNN with vectors sampled from predefined Gaussian noises $\mathcal{N}(0, 1)$. However, the multimodal predictions generated by predefined noises suffer from contextual reasoning. In the main paper, we visualize the multimodal predictions and corresponding probabilistic distributions. In this section, we supplement the comparisons between the performance of predefined Gaussian noises $\mathcal{N}(0, 1)$ and the scene-guided latent variable $\mathcal{N}(\mu, \sigma)$.

As shown in Fig.2, we observe that the multimodal predictions in the top row are randomly distributed with high uncertainty. Some of the predictions even

Table 2. Quantitative comparisons on the new metric Temporal Correlation Coefficient (TCC) in single prediction and multimodal predictions.

Methods	1.0 sec	2.0 sec	3.0 sec	4.0 sec	Avg.
S-LSTM	0.83	0.72	0.63	0.57	0.69
SGAN	0.83	0.71	0.64	0.59	0.69
STGAT	0.87	0.71	0.63	0.58	0.70
Ours (k=1)	**0.88**	**0.73**	**0.66**	**0.61**	**0.72**
S-LSTM	0.86	0.76	0.68	0.63	0.73
SGAN	0.88	0.78	0.70	0.65	0.75
STGAT	0.88	0.78	0.69	0.64	0.75
Ours (k=50)	**0.89**	**0.79**	**0.71**	**0.67**	**0.77**

crosses physically unfeasible areas. For example, in the top row of example 3, some predictions are much shorted than ground truth. Moreover, in the top row of example 4, some of predictions generated by $\mathcal{N}(0,1)$ traverse the parterre and roadblock. It is understandable since the predefined noise does not support for the reasoning about surrounding scene. In contrast, the multimodal predictions by our model are physically plausible with low uncertainty. It verifies that the multimodal predictions benefit from the semantic context of static scene layout further.

4 More Evaluations on the New Metric TCC

In the main paper, we plot the the proposed new metric Temporal Correlation Coefficient (TCC) in multimodal predictions (sample times k= 20) for various methods. Here we provide more evaluations of TCC in both single prediction (sample time $k = 1$) and multimodal predictions (sample times $k = 50$). In the multimodal predictions, we sample the prediction that have lowest ADE with ground truth to compute the TCC. The prediction duration ranges from 1 second to 4 second. As shown in the Table 2, the metric TCC in our method is higher than other state-of-the-art methods regardless of prediction duration and sample times. It shows that the temporal correlation of ground truth is well captured be our method. In addition, the TCC in multimodal predictions (k=50) is consistently higher than the value in single prediction (k=1) for different methods and different prediction duration. It demonstrates that the predictions which are close to the ground truth in Euclidean distance are usually learn the temporal correlation strongly.

5 Description for the Video Demo

In the video demo, we provide the predictions generated by our DSCMP in different scenarios. Blue lines and red lines denote the observed trajectories and ground truth, respectively. The predictions are represented in orange lines. From the video, our DSCMP is capable of making real-time predictions for multiple agents in the crowd, where the spatio-temporal interactions are active and the scene layout is complicated. The future movements for both the short trajectories (*e.g.* pedestrians) and long trajectories (*e.g.* cyclists) are well predicted.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 961–971 (2016)
2. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2255–2264 (2018)
3. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
4. Huang, Y., Bi, H., Li, Z., Mao, T., Wang, Z.: Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6272–6281 (2019)
5. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer graphics forum. vol. 26, pp. 655–664. Wiley Online Library (2007)
6. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 261–268. IEEE (2009)
7. Robicquet, A., Sadeghian, A., Alahi, A., Savarese, S.: Learning social etiquette: Human trajectory understanding in crowded scenes. In: European conference on computer vision. pp. 549–565. Springer (2016)
8. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1349–1358 (2019)
9. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
10. Zhao, T., Xu, Y., Monfort, M., Choi, W., Baker, C., Zhao, Y., Wang, Y., Wu, Y.N.: Multi-agent tensor fusion for contextual trajectory prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12126–12134 (2019)
11. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)