Supplemental Material -
Joint Prediction for Kinematic Trajectories in Vehicle-Pedestrian-Mixed Scenes

Huikun Bi[1,2] Zhong Fang[1] Tianlu Mao[1] Zhaoqi Wang[1] Zhigang Deng[2]*
[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Houston
{bihuikun,fangzhong,ltm,zqwang}@ict.ac.cn, zdeng4@uh.edu

In this supplemental material, we present more details of model implementations and the proposed vehicle-pedestrian-mixed dataset. More statistical analysis and experiments are also provided here.

1. Implementation Details

In this work, the size of the orientation vector is set to 75 pixels in the BJI data and 50 pixels in the TJI data. We embed the input position of a pedestrian as a 64 dimensional vector. Each boundary vertex is embedded into a 64 dimensional vector (Eq. 2) first, and the trajectory of a vehicle is finally embedded into 64 dimensional vectors (Eq. 3). The dimension of the hidden states used in LSTMs for pedestrians and vehicles is fixed to 128. Additionally, the dimension of the embedding function on $H_t^{(vp,j)}$, $H_t^{(vv,j)}$, $H_t^{(pp,i)}$, and $H_t^{(pv,i)}$ is 64. N_o for the grids of both VO and PO is set to 64. The size of the neighborhood is set to 16. The initial learning rate for pedestrians and vehicles is set to 0.005 and 0.01, respectively. The learning process adopts RMS-prop[3] to update the network iteratively with a batch size of 16 for 100 epochs. The implementation is built on the Tensorflow platform[1].

2. More Details of the Proposed Dataset

A new vehicle-pedestrian-mixed dataset is proposed in our work. The initial video dataset was acquired with a DJI Mavic Pro drone, which hovered on the intersections and took video from a top-down view.

In order to obtain the accurate trajectories of vehicles and pedestrians from the acquired video data, we employed a visual tracking algorithm called Multi-Domain Convolutional Neural Networks (MDNet)[6] to track the locations of vehicles and pedestrians in each video frame. We used the center of the bounding box of each vehicle/pedestrian as the location of the vehicle/pedestrian. After the automated video tracking, we further manually checked and corrected any miss-tracked places for every frame to ensure the data quality. Although there is an embedded stable gyroscope to ensure the quality of the video taken by the drone, the inevitable effects of wind and slight fluctuations of the drone make the video slightly rotated and translated. To remedy this issue, we detected the SURF features[2] and transformed the annotated positions in the original video into their corresponding positions in a rectified coordinate system. In order to obtain the orientations of the vehicles, we first used Gaussian smoothing to remove potential tracking noise and high-frequency information in the trajectories. For a vehicle v^j, we computed the trajectory tangent at t as the orientation of v^j.

Our in-house built dataset has several advantages over existing similar datasets. First, our dataset contains the trajectories of a large number of vehicles and pedestrians in vehicle-pedestrian-mixed scenes. The agents were well classified and capture complex interactions among vehicles and pedestrians. Second, our dataset with well-annotated trajectories has been checked and manually corrected frame by frame. In total, 6405 pedestrians and 6478 vehicles in two scenarios were captured from the top-down view. Moreover, there are 23498 and 8000 annotated frames in two scenarios, respectively.

3. More Experiment Qualitative Analysis

We further show more quantitative and qualitative analysis here.

3.1. More Quantitative Analysis

Because SGAN performs better than other methods on predicting homogeneous trajectories in the prior experiments, we choose SGAN [4] as a baseline. We further compare our method with TrafficPredict [5], specifically designed to predict trajectories for heterogeneous traffic-

*Corresponding Author

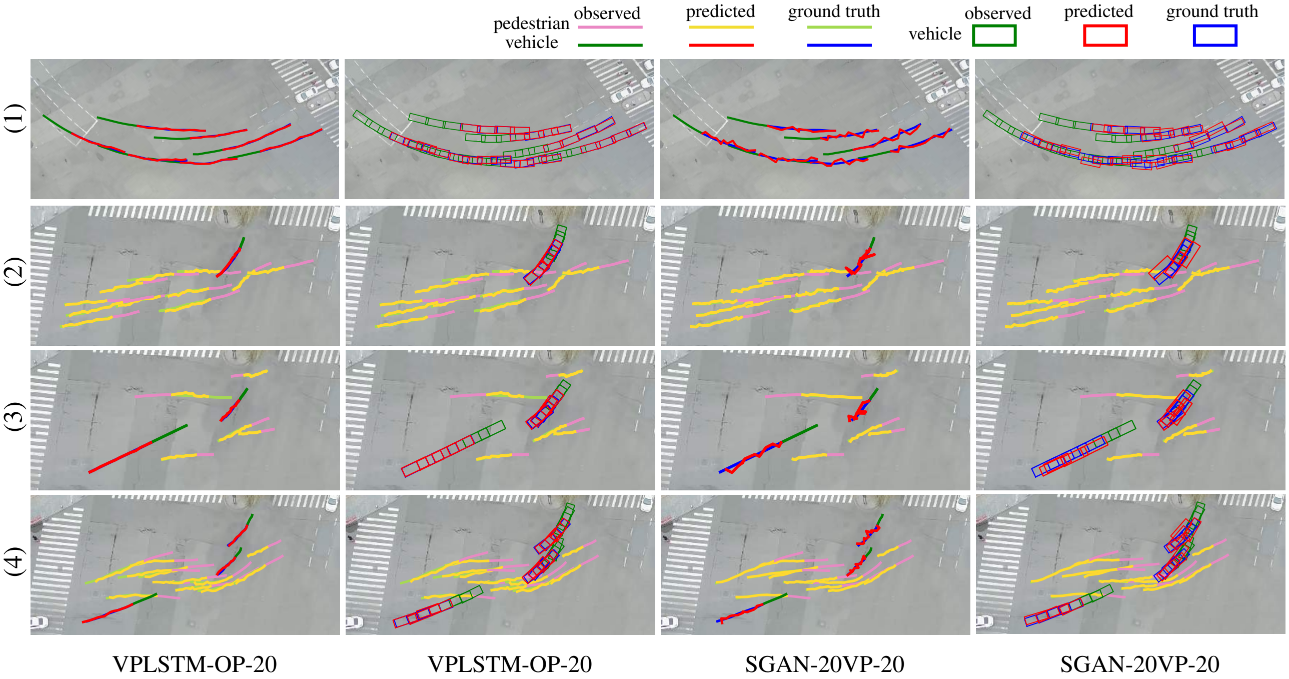

Figure 1. Four examples of the predicted trajectories compared with the ground truth and SGAN-20VP-20. The left and center-right columns show the position trajectories of both vehicles and pedestrians predicted by VPLSTM-OP-20 and SGAN-20VP-20, respectively. The kinematic trajectories of vehicles illustrated in the center-left and right columns are represented by OBB. Here $T_{\text{obs}} = 8$ and $T_{\text{pred}} = 12$. In order to clearly illustrate kinematic trajectories, we sample trajectories and show vehicles at $t = 3, 6, 9, 12, 15, 18$.

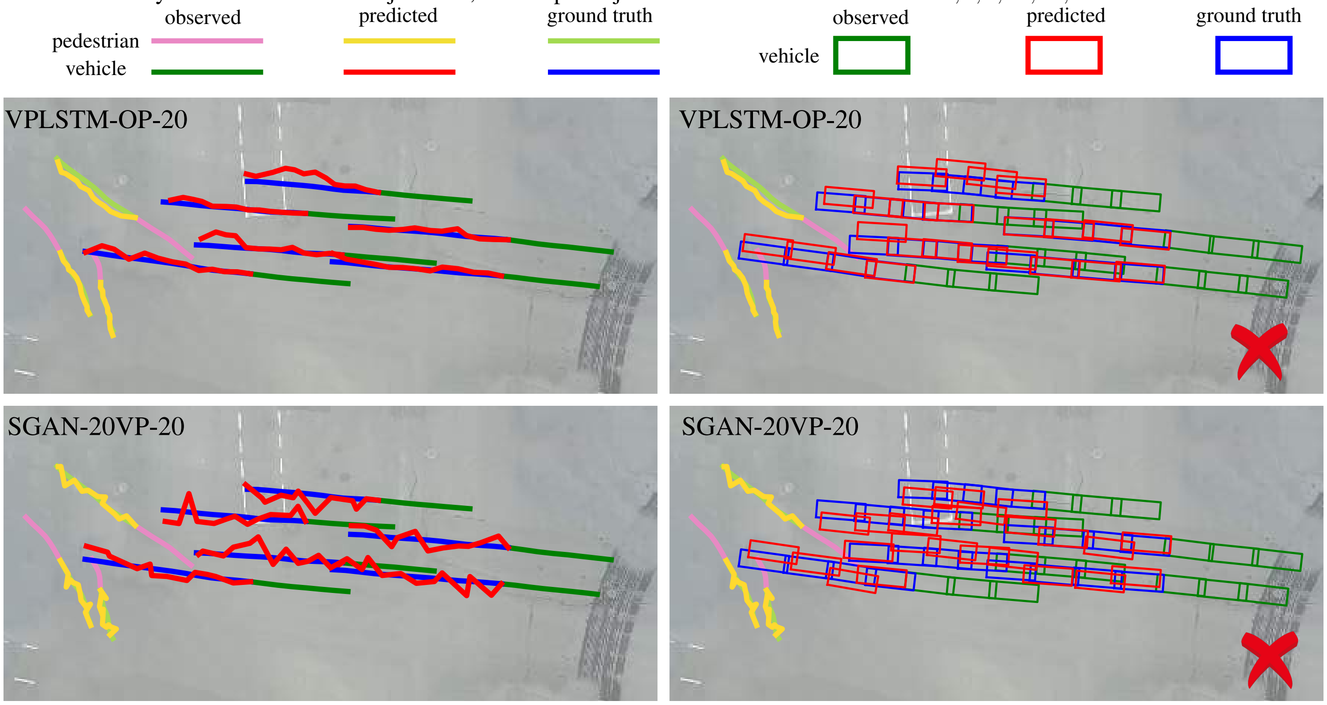

Figure 2. Failed cases of the predicted trajectories compared with the ground truth and SGAN-20VP-20. Here $T_{\text{obs}} = 8$ and $T_{\text{pred}} = 12$. The trajectories of vehicles are shown at $t = 3, 6, 9, 12, 15, 18$.

agents, on dataset Apollo [5]. As shown in Table 1, most results of our method outperform state-of-the-art competitors.

3.2. More Qualitative Analysis

More qualitative evaluation results are shown in Fig.1. As illustrated in example (1) and (4), the turning vehi-

Metric	**Data-set**	**Agent**	**TrafficPredict** [5]	**SGAN** [4]		**VP-LSTM**	
				1VP-1	20VP-20	O-20	OP-20
ADE	**Apollo**	**Vehicle**	21.72 / 30.43	3.19 / 6.85	2.03 / 4.20	2.07 / 2.33	**1.90 / 2.09**
		Pedestrian	12.20 / 15.22	2.29 / 5.09	**1.58** / 3.38	2.18 / 2.60	2.28 / **2.39**
		Average	20.35 / 28.64	2.91 / 6.37	**1.89** / 3.97	2.10 / 2.39	2.01 / **2.14**
FDE	**Apollo**	**Vehicle**	39.54 / 50.41	8.82 / 13.91	5.70 / 8.48	2.92 / 3.34	**2.66 / 3.07**
		Pedestrian	21.24 / 24.38	5.97 / 9.58	4.29 / 6.31	3.27 / 4.03	**3.17 / 3.35**
		Average	37.72 / 48.45	7.93 / 12.72	5.26 / 7.88	3.02 / 3.51	**2.80 / 3.14**
$ADE_{\mathbf{O}}$	**Apollo**	**vehicle**	23.23 / 31.81	3.87 / 7.88	2.63 / 5.11	2.57 / 2.86	**2.45 / 2.66**
$FDE_{\mathbf{O}}$	**Apollo**	**vehicle**	41.08 / 51.52	10.03 / 15.14	6.72 / 9.51	3.44 / 3.92	**3.24 / 3.64**

Table 1. Quantitative results for the predicted positions and orientations on Apollo. Error metrics are reported in meters.

cles limited with kinematics will slow down to avoid collisions with other vehicles and pedestrians. Although SGAN-20VP-20 can predict reasonable trajectories (see the center-right column) for vehicles that drive as particles without size. However, the predicted kinematic trajectories of vehicles are unsmooth due to the estimated orientations. The trajectory of a pedestrian predicted in example (3) is visibly away from the ground truth. However, considering the oncoming vehicle in the front of the pedestrian, our predicted trajectory of the pedestrian tends to slow down to avoid implicit collisions.

Failed cases: We also show a failed case in Fig.2. Compared with the ground truth, the following vehicles tend to keep space to avoid collisions with neighboring vehicles. In the same situation, the kinematic trajectories predicted by SGAN-20VP-20 could cause obvious collisions.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems. `https://www.tensorflow.org/`, 2015. Software available from tensorflow.org.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer, 2006.

[3] Yann Dauphin, Harm De Vries, and Yoshua Bengio. Equilibrated adaptive learning rates for non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1504–1512, 2015.

[4] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6120–6127, 2019.

[6] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.