

TSC-NET: PREDICTION OF PEDESTRIAN TRAJECTORIES BY TRAJECTORY-SCENE-CELL CLASSIFICATION (SUPPLEMENTAL MATERIAL)

Bo Hu, Tat-Jen Cham

College of Computing and Data Science

Nanyang Technological University

50 Nanyang Ave, Block N4, Singapore

hubo0005@e.ntu.edu.sg, astjcham@ntu.edu.sg

A OVERVIEW

We first introduce the implementation details of our framework in Section B. Then the supplementary experiment results are reported in Section C, followed by more visualizations in Section D.

B IMPLEMENTATION DETAILS

B.1 TSC FEATURE EMBEDDING

In the trajectory-scene-cell (TSC) feature embedding stage, three types of TSC features are obtained: trajectory cell feature, step scene cell feature, and goal scene cell feature. Every feature type contains a scene feature embedding from a cropped scene. A trajectory cell feature contains a scene embedding from a $u \times u$ area. A step scene cell feature contains a scene embedding of one position from a $lr \times lr$ area. Similarly, a goal scene cell feature contains a scene embedding of one position from a $LR \times LR$ area. As both trajectory cell and step scene cell only requires local scene embedding, we set $u = lr$ for simplification.

l and L determine how many cells are classified during prediction the goal/step. Empirically, we set $l = 3$ and $L = 15$ for all experiments (different choices of L are explored in ablation study). The selection of r and R depends on the dataset: LR should cover all goal position of trajectories in the training set and lr should cover all next step positions of trajectories in the training set. Therefore, the selection of r and R are listed in the Table 1.

Table 1: Selection of l , r , L , and R in different datasets with different prediction settings

Dataset	Prediction Setting	l	r	lr	L	R	LR
SDD	Short-term	3	20	60	15	80	1200
ETH-UCY	Short-term	3	20	60	15	80	1200
SDD	Long-term	3	60	180	15	200	3000
inD	Long-term	3	24	72	15	100	1500

The TSC feature embedding stage includes one coordinate embedding network $f(\cdot)$ and three scene embedding networks $g^{(traj)}(\cdot)$, $g^{(step)}(\cdot)$, and $g^{(goal)}(\cdot)$. The coordinate embedding network $f(\cdot)$ is a 3-layer MLP. Three scene embedding networks $g^{(traj)}(\cdot)$, $g^{(step)}(\cdot)$, and $g^{(goal)}(\cdot)$ are CNNs. The CNNs consist of 7 convolutional layers and 2-3 max-pooling layers. Different max-pooling layers are applied depends on different r and R . Since the original size of scene is too large and the most useful information is the semantic label instead of detailed appearance texture, the scene is down-sampled by a factor equals to 4 before cropped and sent to the scene embedding CNNs. After the TSC feature embedding, the dimension of each cell $D = 256$, where 128-dim coordinate feature embedding and 128-dim scene feature embedding.

B.2 GOAL PREDICTION AND TRAJECTORY COMPLETION

For all attention based encoders: history encoder, goal encoder, and step encoder, the number of self-attention and cross-attention layers $K = 2$. For both CVAE decoder and step decoder, 3-layer cell-level MLPs are applied, where features of different cells are decoded independently. In the CVAE structure, ground truth embedding layer and CVAE encoder layer are also 3-layer cell-level MLPs. During training, the condition is concatenated with embedded ground truth before sending to CVAE encoder. The output of CVAE is a 64-dim latent variable for every cell, which is sent to CVAE decoder, as well as used for KLd loss computing. When computing the reconstruction loss for both goal and each step, the confidence reconstruction loss is computed for all cells, while the offset reconstruction loss is only computed for the cells having the ground truth goal/step.

C ADDITIONAL EXPERIMENTAL RESULTS

For the short-term prediction, our TSC-Net is compared with more existing methods, as Shown in Table 2.

Table 2: Additional comparison between our method and existing methods on ETH-UCY dataset and SDD dataset. “ADE/FDE” are reported.

Methods	ETH	HOTEL	UNIV	ZARA1	ZARA2	Average	SDD
P2TIRL Deo & Trivedi (2020)	—	—	—	—	—	—	10.97/12.40
CF-VAE Bhattacharyya et al. (2019)	—	—	—	—	—	—	12.60/22.30
SimAug Liang et al. (2020)	—	—	—	—	—	—	10.27/19.71
SIT Su et al. (2021)	0.38/0.88	0.11/0.21	0.20/0.46	0.16/0.37	0.12/0.27	0.19/0.44	—
Social-BiGAT Kosaraju et al. (2019)	0.69/1.29	0.49/1.01	0.55/1.32	0.30/0.62	0.36/0.75	0.48/1.00	—
Next Liang et al. (2019)	0.73/1.65	0.30/0.59	0.60/1.27	0.38/0.81	0.31/0.68	0.46/1.00	—
STAR Yu et al. (2020)	0.36/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53	—
Singular Bae et al. (2024)	0.35/0.42	0.13/0.19	0.25/0.44	0.19/0.32	0.15/0.25	0.21/0.32	—
Causal-STGCNN Chen et al. (2021)	0.64/1.00	0.38/0.45	0.49/0.81	0.34/0.53	0.32/0.49	0.43/0.66	—
CGNS Li et al. (2019)	0.62/1.40	0.70/0.93	0.48/1.22	0.32/0.59	0.35/0.71	0.49/0.97	15.60/28.20
PECNet Mangalam et al. (2020)	0.54/0.87	0.18/0.24	0.35/0.60	0.22/0.39	0.17/0.30	0.29/0.48	9.96 /15.88
LB-EBM Pang et al. (2021)	0.30/0.52	0.13/0.20	0.27/0.52	0.20/0.37	0.15/0.29	0.21/0.38	8.87 /15.61
PCCSNET Sun et al. (2021)	0.28/0.54	0.11/0.19	0.29/0.60	0.21/0.44	0.15/0.34	0.21/0.42	8.62 /16.16
CSCNet Xia et al. (2022)	0.51/1.05	0.22/0.42	0.36/0.81	0.31/0.68	0.47/1.02	0.37/0.79	14.63/26.91
TSC-Net (Ours)	0.32/ 0.39	0.12/ 0.19	0.25/ 0.46	0.17/ 0.30	0.15/ 0.26	0.20/ 0.32	6.44/9.97

The visualizations in our main paper show that compared to heatmap based method Y-Net Mangalam et al. (2021), our method predict more accurate trajectories when the speed is irregular in the history and prediction. To further demonstrate the capability of predicting trajectories with irregular speed, velocity difference between history and future V_{diff} is defined. First, one step velocity between two neighboring frames is defined as $v_t = ||p_{t+1} - p_t||_2$. Then the history mean velocity is defined as the average velocity of all historical frames (from frame 1 to τ). Similarly, The future mean velocity is the average velocity of all frames in the future (from frame $\tau + 1$ to T). Thus, the velocity difference V_{diff} is computed by absolute difference between mean velocity of history and future, which is

$$V_{diff} = \left| \frac{\sum_{t=1}^{\tau} v_t}{\tau} - \frac{\sum_{t=\tau+1}^T v_t}{T - \tau} \right|. \quad (1)$$

For SDD dataset with long-term prediction setting, samples are sorted by their velocity difference, and ADE/FDE with several velocity different thresholds are reported, where the results are shown in Table 3. It can be observed that compared to heatmap based method Y-Net Mangalam et al. (2021), our method significantly achieves much better FDE, especially with the largest V_{diff} .

NBA SportVU dataset Linou et al. (2024) is evaluated to test the trajectory prediction in sports scenario. This dataset contains trajectories of 10 players and a basketball for prediction, where the 5 frames (2 seconds) are observed and 10 frames (4 seconds) or prediction. The results are shown as follow.

Table 3: “ADE/FDE” with different V_{diff} thresholds in SDD dataset with long-term prediction setting.

Top (%) Samples with Largest V_{diff}	50%	40%	30%	20%	10%
Y-Net Mangalam et al. (2021)	61.82 / 89.69	66.98 / 98.17	74.93 / 109.49	85.56 / 127.88	100.10 / 162.50
TSC-Net (Ours)	66.16 / 82.84	66.07 / 83.50	72.10 / 85.54	81.29 / 94.23	85.41 / 101.86

Table 4: Trajectory prediction in MBA dataset.

Methods	MemoNet Xu et al. (2022b)	V^2 -Net Wong et al. (2022)	GroupNet Xu et al. (2022a)	E- V^2 -Net-SC Wong et al. (2024)	Ours
ADE/FDE	1.25 / 1.47	1.28 / 1.68	1.13 / 1.69	1.18 / 1.46	1.24 / 1.50

Our method results in slightly larger ADE and FDE when comparing to other methods. One possible reason is that trajectories in the NBA dataset are highly influenced by player-player interactions, as players need to react to each other’s movements, while the scene has little influence. Our method is designed to address feature alignment between the scene and trajectory, and its capability could be limited when scenes are excluded.

D VISUALIZATIONS

Visualizations of predicted goal distributions for short-term and long-term predictions on SDD are shown in Figure 1 and Figure 2 respectively. In short-term prediction setting, although trajectories are relative smooth, our method tends to generate a goal distribution covering larger area than Y-Net, unless for the short and straight trajectories such as the third row in Figure 1. In the long-term prediction, our framework shows better capability than Y-Net for predicting the goal for the trajectories with sudden change of velocity. Visualizations for long-term prediction also demonstrate that TSC feature embedding has comparable capability of learning the relationship between scene and trajectory to the heatmap based method Y-Net. For example, the sampled goals distributed along the cross sidewalk in the second and third rows in Figure 2.

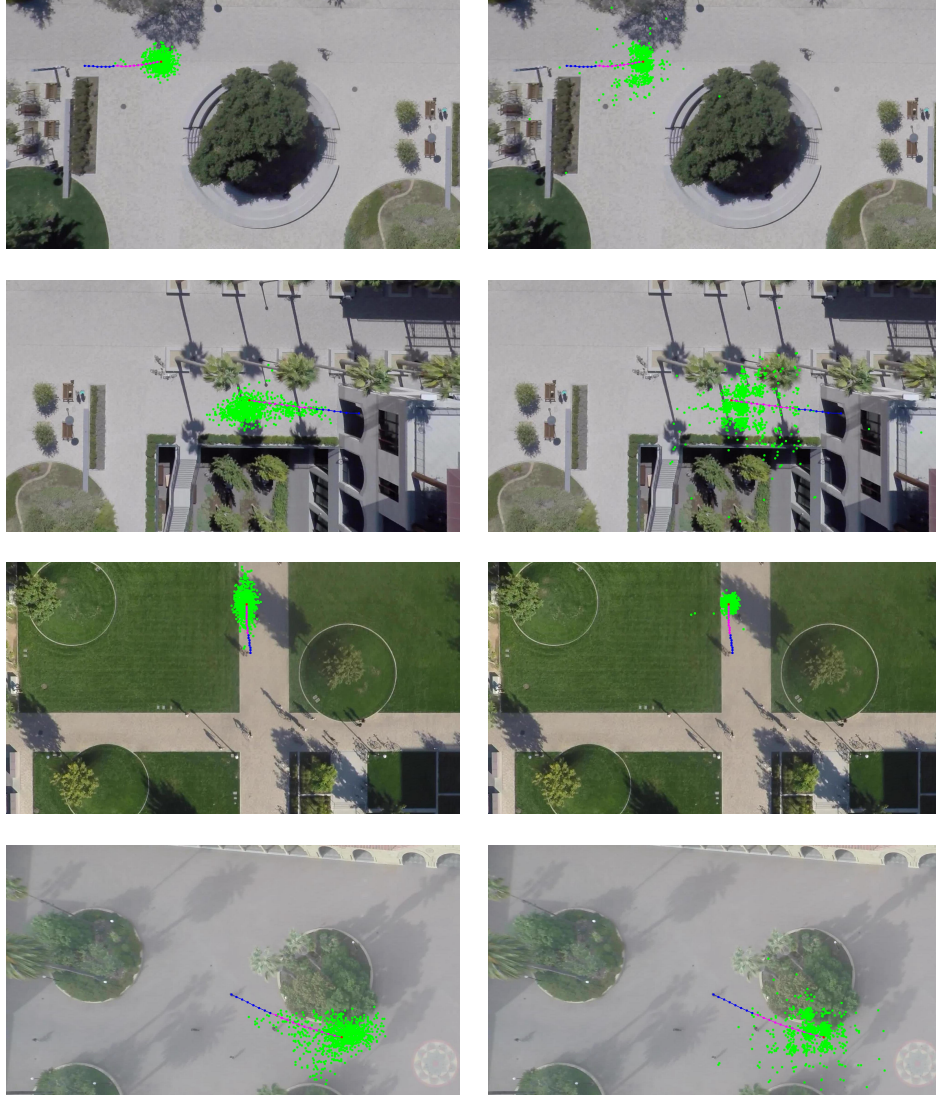


Figure 1: Visualization of trajectory prediction results comparison between Y-Net and our method on SDD dataset with short-term setting. The first column show results from Y-Net, while the second column show results from our method. Blue and magenta curves: observed part and future part of ground truth trajectory. Red and green dots: the ground truth goal and predicted goals.

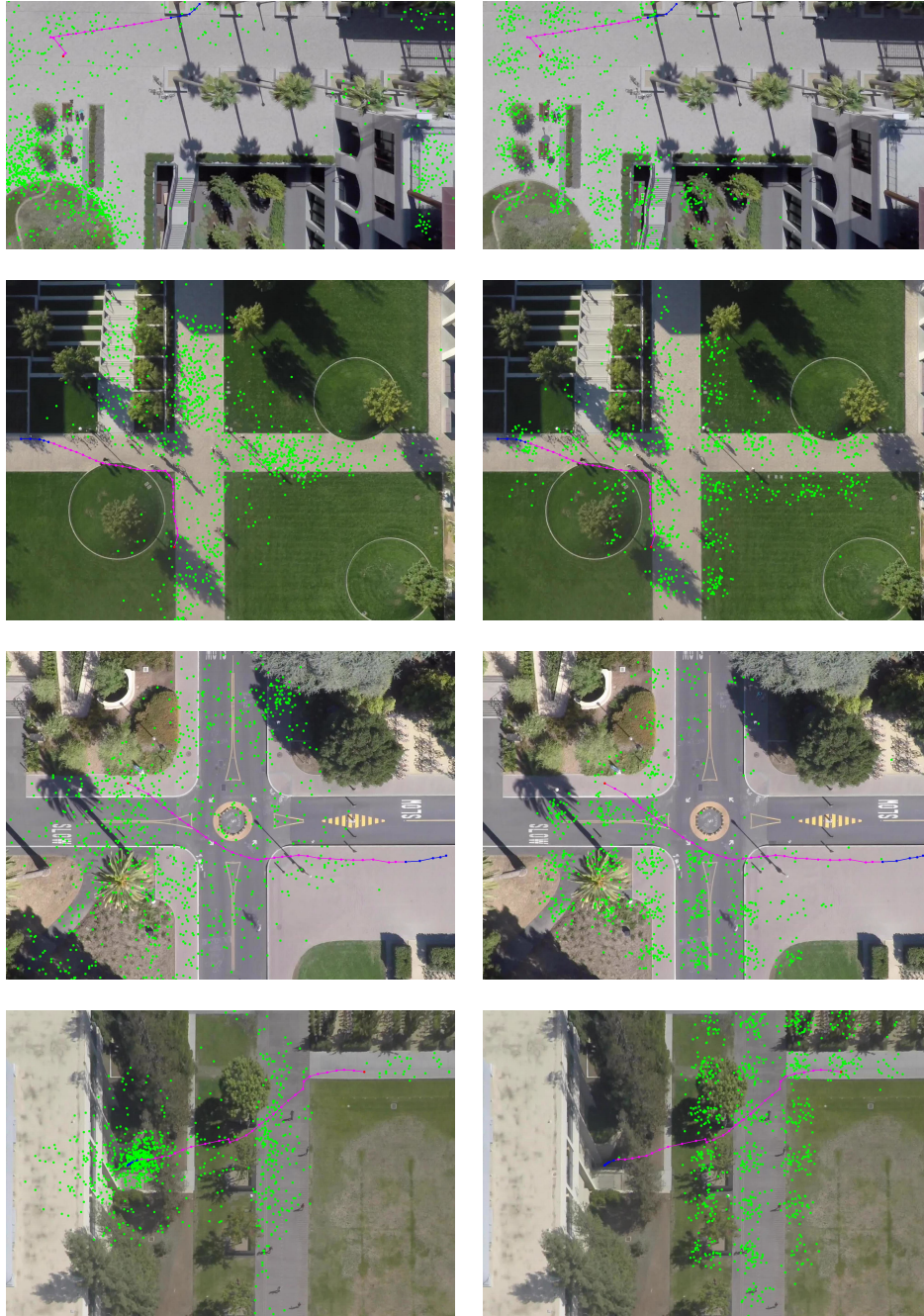


Figure 2: Visualization of trajectory prediction results comparison between Y-Net and our method on SDD dataset with long-term setting. The first column show results from Y-Net, while the second column show results from our method. Blue and magenta curves: observed part and future part of ground truth trajectory. Red and green dots: the ground truth goal and predicted goals.

REFERENCES

- Inhwan Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17890–17901, 2024.
- Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele, and Christoph-Nikolas Straehle. Conditional flow variational autoencoders for structured sequence prediction. In *4th Workshop on Bayesian Deep Learning*. bayesiandeeplearning.org, 2019.
- Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human trajectory prediction via counterfactual analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9824–9833, 2021.
- Nachiket Deo and Mohan M Trivedi. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735*, 2020.
- Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32, 2019.
- Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional generative neural system for probabilistic trajectory prediction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6150–6156. IEEE, 2019.
- Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5725–5734, 2019.
- Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from simulation for trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pp. 275–292. Springer, 2020.
- Kostya Linou, Dzmitryi Linou, and Martijn de Boer. Nba player movements. <https://github.com/linouk23/NBA-Player-Movements>, 2024.
- Kartikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 759–776. Springer, 2020.
- Kartikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15233–15242, 2021.
- Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. Trajectory prediction with latent belief energy-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11814–11824, 2021.
- Tong Su, Yu Meng, and Yan Xu. Pedestrian trajectory prediction via spatial interaction transformer network. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*, pp. 154–159. IEEE, 2021.
- Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13250–13259, 2021.
- Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, Yibo Yang, and Xinge You. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *European Conference on Computer Vision*, pp. 682–700. Springer, 2022.
- Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19005–19015, 2024.

- Beihao Xia, Conghao Wong, Qinmu Peng, Wei Yuan, and Xinge You. Cscnet: Contextual semantic consistency network for trajectory prediction in crowded spaces. *Pattern Recognition*, 126: 108552, 2022.
- Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6498–6507, 2022a.
- Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6488–6497, 2022b.
- Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 507–523. Springer, 2020.