Supplementary Material for Social Aware Multi-Modal Pedestrian Crossing Behavior Prediction

Xiaolin Zhai[1,2[0000−0003−1890−8378]], Zhengxi Hu[1,2[0000−0001−6119−4185]], Dingye Yang[1,2[0000−0002−1039−6817]], Lei Zhou[1,2[0000−0002−1940−8597]], and Jingtai Liu[1,2[0000−0003−2645−5655]]

[1] Institute of Robotics and Automatic Information System, College of Artificial Intelligence, Nankai University
[2] Tianjin Key Laboratory of Intelligent Robotics, Nankai University
{ 2120210410,hzx,1711502}@mail.nankai.edu.cn,zhouleinku@gmail.com, liujt@nankai.edu.cn

To supplement our main submission, we provide additional materials in this document.

1 Additional Experiments

1.1 Ablation Study on Social Aware Encoder

As shown in Table 1, we study the impact of each module in Social Aware Encoder. Spatial-Temporal Heterogeneous Graph boosts the performance of action prediction from 0.15 mAP and 0.23 mAP. On the intention estimation task, this module also improves remarkably by 2.4% $\sim$ 24.6% on multiple metrics. It reveals that spatial-temporal interactions between the pedestrian and surrounding traffic objects contain important relation information. Our proposed STHG comprehensively propagates the relation information on the heterogeneous graph to generate rich relation representations, boosting pedestrian behavior prediction. Appending EVDE Module achieves improvements on pedestrian action prediction. The result verifies our hypothesis that using the future motion plan of the ego-vehicle is beneficial for our model to understand pedestrian actions. PFF module integrates the future hidden states from the decoder, and improves multiple metrics of pedestrian action and intention task.

1.2 Ablation Study on Intention-Directed Decoder

As shown in Table 2, we perform the ablation study on Intention-Directed Decoder to evaluate the effectiveness of the pedestrian intention. We remove the estimated intention from the decoder input and only take the predicted action at the last time step as the input of GRU. Results show an improvement from 0.25 to 0.26 on the action prediction task. It is clear that the pedestrian's intention is important for inferring pedestrian actions.

Table 1. Ablation study of Social Aware Encoder on the PIE dataset using the PIE sampling strategy and the sampling length $T = 30$. We gradually add Spatial-Temporal Heterogeneous Graph (STHG), Ego-Vehicle Dynamics Encoding (EVDE), and Pedestrian Future Fusion (PFF) on the basis of Pedestrian Dynamics Encoding (PDE).

Modules	Action		Intention			
	mAP(dete)	mAP(pred)	Acc	F1	Prec	AUC
PDE (necessary)	0.17	0.15	0.82	0.90	0.84	0.69
PDE+STHG	0.27	0.23	0.84	0.90	**0.92**	0.86
PDE+STHG+EVDE	0.27	0.25	0.84	0.90	**0.92**	**0.87**
PDE+STHG+EVDE+PFF	**0.29**	**0.26**	**0.85**	**0.91**	**0.92**	**0.87**

Table 2. Ablation study of Intention-Directed Decoder on the PIE dataset using PIE sampling strategy and the sampling length $T = 30$.

Modules	Action		Intention			
	mAP(dete)	mAP(pred)	Acc	F1	Prec	AUC
w/o Intention-Directed	0.28	0.25	**0.85**	**0.91**	**0.92**	0.86
Ours	**0.29**	**0.26**	**0.85**	**0.91**	**0.92**	**0.87**

1.3 Run Time

The inference time of our model is $11.4 \pm 2.8ms$ per frame on a single NVIDIA RTX 3090 GPU. The result indicates that our model is fast enough to meet the real-time requirements.