Learning Scene-Goal-Aware Motion Representation for Trajectory Prediction

Ziyang Ren, Ping Wei, Haowen Tang, Huan Li, Jin Yang

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Introduction

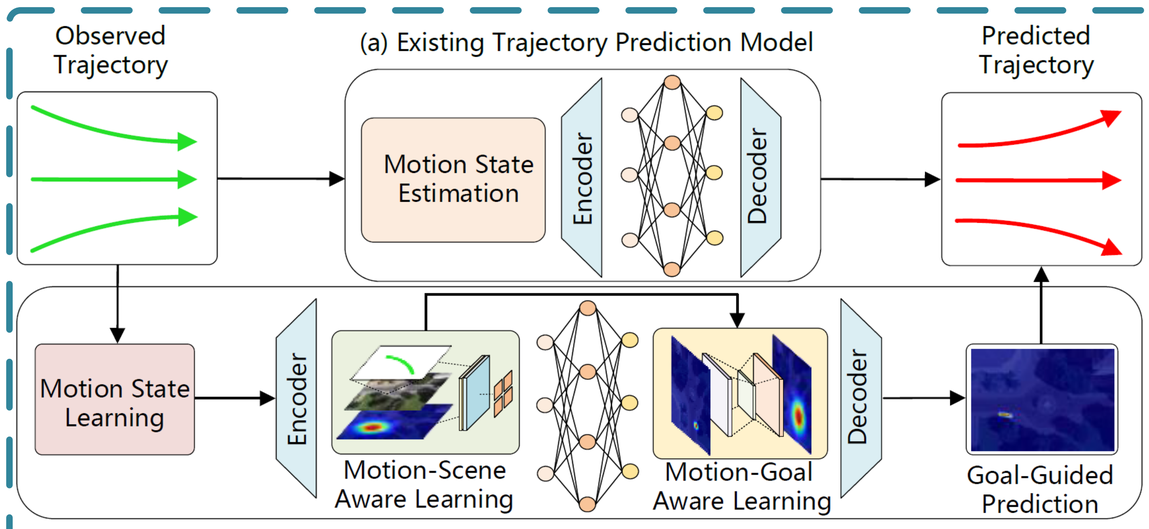
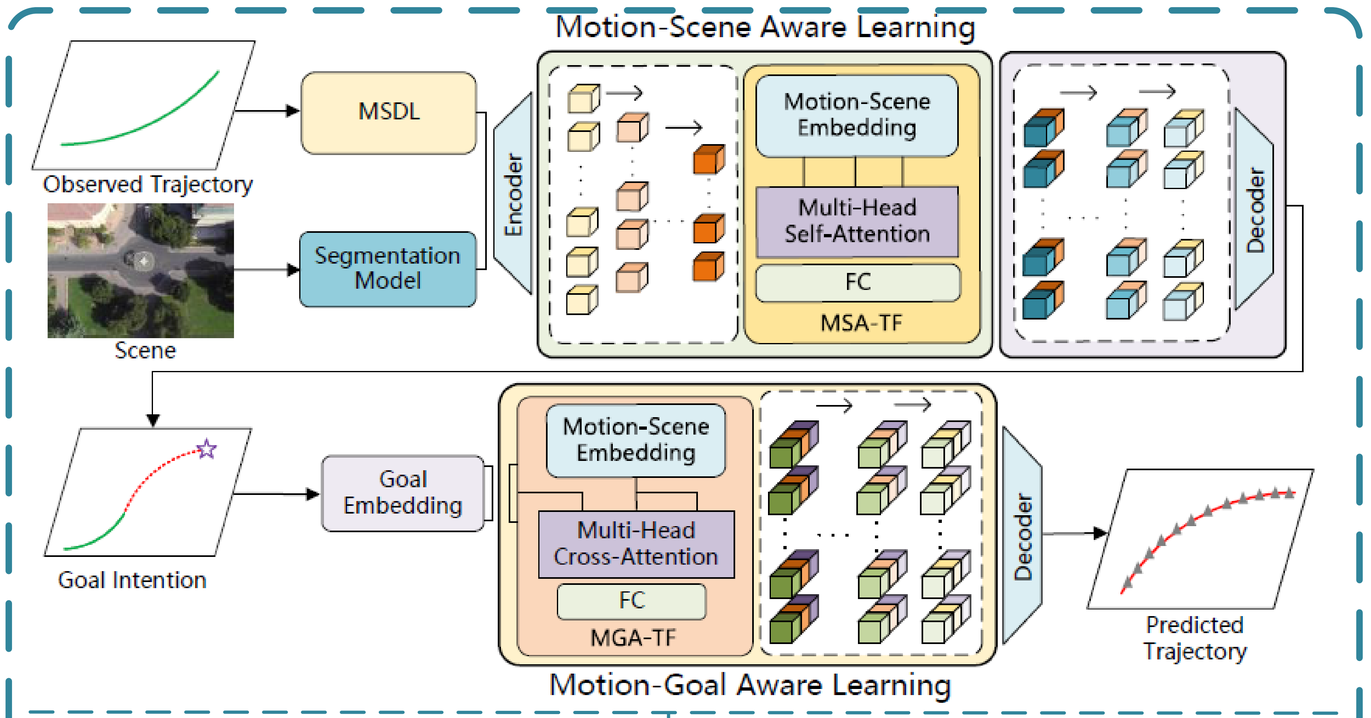


Figure 1: Existing general framework of trajectory prediction and our MSGANet framework. (a) Describing motion trajectories through motion state estimation. (b) Integrating scene information understanding and goal intent derivation to enhance motion state learning.

Highlights of this work

- Extracting motion state representations through the combination of temporal attention convolutional networks and self-supervised losses based on state log-likelihood maximization.
- Enhancing the interaction between the motion state and scene semantics by fusing motion states with scene semantic features through the Motion-Scene Aware Transformer. We utilize the fused motion states incorporating scene semantics to infer the distribution of goal intentions.
- We design the Motion-Goal Aware Transformer. By calculating the correlation between motion states and goals, it guides the transition of motion states towards future motion trends.

Methodology & Network Architecture



•1. Motion State Distribution Learning

Maximizing the log-likelihood function of motion trajectories under the learned motion state probability distribution.

$$L_{sll} = -\log(P(x|\mu_o, \sigma_o, \rho_o)).$$

•2. Motion-Scene Aware Learning

Integrating Motion State Spatial Distribution with Scene Segmentation Features based on Self-Attention for deriving goal intent.

$$\varepsilon_{ms} = \mathrm{fc}(\mathrm{softmax}(\frac{\phi_q(\varepsilon_c)\phi_k(\varepsilon_c)^T}{d_{embed}})\phi_v(\varepsilon_c) + \varepsilon_c),$$

•3. Motion-Goal Aware Learning

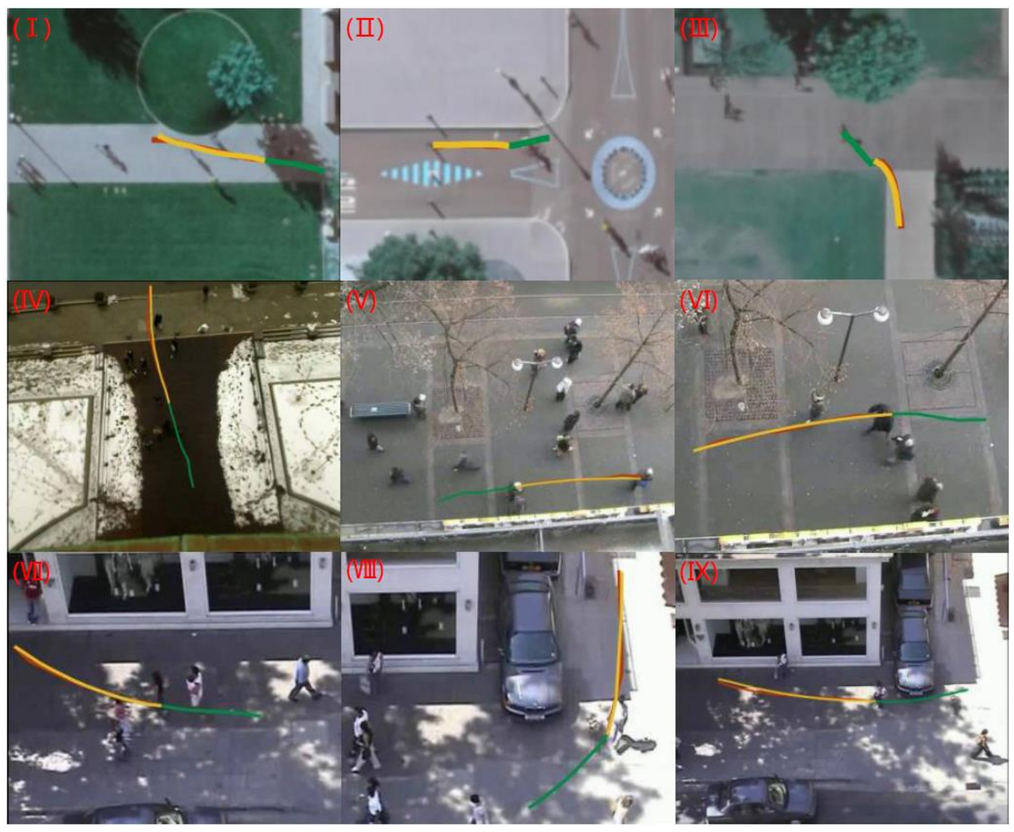
Using cross attention to facilitate the interaction between the distribution of movement states and goal intentions, thereby guiding the decoding process of future trajectories.

$$\varepsilon_{mg} = \mathrm{fc}(\mathrm{softmax}(\frac{\phi'_q(\varepsilon_g)\phi'_k(\varepsilon_s)^T}{d_{embed}})\phi'_v(\varepsilon_s) + \varepsilon_s)$$

Results

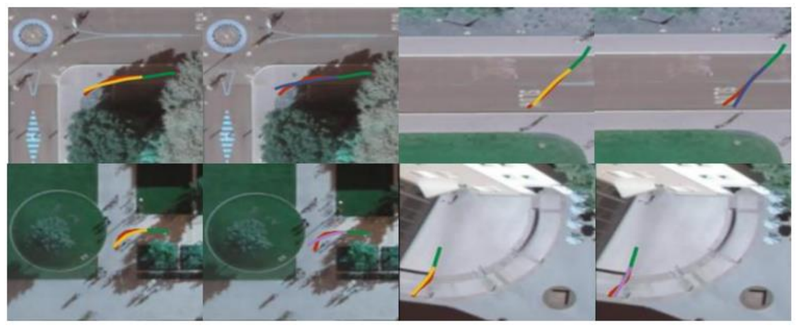
Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG	SDD
Social-GAN [13]	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.84	0.61/1.21	27.23/41.44
Trajectron++ [32]	0.61/1.02	0.19/0.28	0.30/0.54	0.24/0.42	0.18/0.32	0.30/0.51	19.30/32.70
Social-STGCNN [26]	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75	N/A
PECNet [22]	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
Transformer-TF [12]	0.61/1.12	0.18/0.30	0.35/0.65	0.22/0.38	0.17/0.32	0.31/0.55	N/A
BiTraP-NP [47]	0.37/0.69	0.12/0.21	**0.17**/0.37	**0.13**/0.29	**0.10**/0.21	**0.18**/0.35	N/A
AgentFormer [48]	0.45/0.75	0.14/0.22	0.25/0.45	0.18/0.30	0.14/0.24	0.23/0.39	N/A
Y-Net [23]	0.28/0.33	**0.10/0.14**	0.24/0.41	0.17/**0.27**	0.13/0.22	**0.18/0.27**	7.85/11.85
Introvert [33]	0.42/0.70	0.11/0.17	0.20/**0.32**	0.16/**0.27**	0.16/0.25	0.21/0.34	N/A
GroupNet [43]	0.46/0.73	0.15/0.25	0.26/0.49	0.21/0.39	0.17/0.33	0.25/0.44	9.31/16.11
MemoNet [44]	0.40/0.61	0.11/0.17	0.24/0.43	0.18/0.32	0.14/0.24	0.21/0.35	8.56/12.66
Social-VAE [45]	0.41/0.58	0.13/0.19	0.21/0.36	0.17/0.29	0.13/0.22	0.21/0.33	8.10/11.72
LED [24]	0.39/0.58	0.11/0.17	0.26/0.43	0.18/0.26	0.13/0.22	0.21/0.33	8.48/11.66
TUTR [35]	0.40/0.61	0.11/0.18	0.23/0.42	0.18/0.34	0.13/0.25	0.21/0.36	7.76/12.69
CMATP [50]	0.32/0.51	0.11/0.16	0.37/0.52	0.19/**0.27**	0.14/0.21	0.22/0.33	N/A
MSGANet (Ours)	**0.27 /0.31**	**0.10/0.14**	0.24/0.44	0.17/**0.27**	0.13/**0.20**	**0.18/0.27**	**7.69/11.61**

Our method outperforms the recent approach by 18.18% in ADE/FDE on the ETHUCY dataset, showcasing superior performance in predicting diverse motion trajectories by jointly learning motion states and goal intentions across different subdatasets.



Our approach demonstrates notable proficiency in predicting trajectories involving extended linear movements (I, V), regular turns occurring at long distances(VII, IX) and significant turning maneuvers (II,III, VIII).

Ablation



Method	Performance ADE/FDE
Ours w/o MSDL	7.78/11.72
Ours w/o MSAL	7.83/11.81
Ours w/o MSA-TF	7.81/11.77
Ours w/o MGA-TF	7.76/11.64
Ours w/o L_{rll}	7.73/11.63
Ours	**7.69/11.61**

The absence of motion state learning results in biases in estimating goal intentions, leading to directional deviations. MSAL facilitates the interaction between scene information and motion states, preventing collisions between predicted trajectories and scene boundaries.

Conclusion

This paper introduces MSGANet, a pedestrian trajectory prediction framework that integrates scene comprehension and goal intent inference to enhance motion state learning. MSGANet uses temporal convolution to learn motion state distribution and combines it with Transformer models for effective fusion of motion states and scene details, improving goal intention derivation. Additionally, a Transformer model with cross-attention facilitates interaction between goal intentions and the motion-scene fusion feature, enhancing trajectory decoding and feature representation of future motion trends.