Learning Scene-Goal-Aware Motion Representation for Trajectory Prediction

Ziyang Ren
rzyrzy@stu.xjtu.edu.cn
Ping Wei*
pingwei@xjtu.edu.cn
Haowen Tang
sdtathw@stu.xjtu.edu.cn
Huan Li
huanli@xjtu.edu.cn
Jin Yang
jinyang@stu.xjtu.edu.cn

National Key Laboratory of
Human-Machine Hybrid Augmented
Intelligence, Institute of Artificial
Intelligence and Robotics, Xi'an
Jiaotong University
Xi'an, China

Abstract

Predicting accurate movement trajectory is a challenging task due to the complexity of human motion patterns and activity scenes. Existing studies focus on extracting motion state information from trajectories but often overlook the representation of future motion trends and interaction with the scene. We present a novel framework called Motion-Scene-Goal Aware Network (MSGANet), which utilizes attention temporal convolutional networks to capture temporal dynamics in motion trajectories. Through self-supervised learning, MSGANet extracts the spatial distribution of motion states. It incorporates multi-scale feature fusion and self-attention mechanisms to extract correlated features between motion states and physical scenes, facilitating inference of goal intentions' spatial distribution. Additionally, MSGANet employs cross-attention mechanisms to enable feature interactions between motion states and goal intentions. By integrating scene semantic aware fusion and aware interaction of goal intentions, it enhances the representation of motion state features for predicting future motion trends. Experiments on ETH-UCY and SDD datasets prove the strength of our method.

1 Introduction

Human motion prediction is crucial for understanding human behavior and can be applied across various domains such as social robots [21] and autonomous driving [5]. Due to the influence of multimodal motion patterns, predicting human movement trajectories is a challenging task. According to social psychology [1], human movement is driven by intention and influenced by interactions with the environment. Therefore, learning the associated representation within motion states regarding scene interactions and intentions is of importance for human movement prediction.

* Corresponding Author

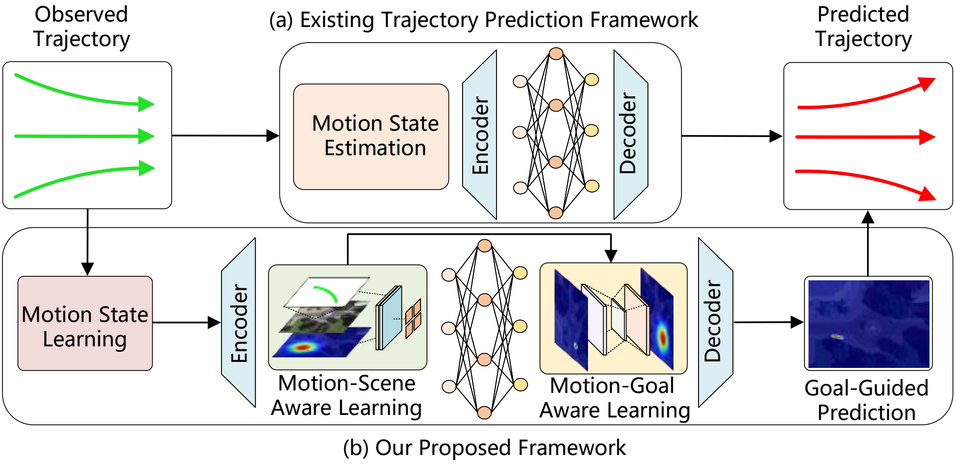

Figure 1: Existing general framework of trajectory prediction and our MSGANet framework. (a) Describing motion trajectories through motion state estimation. (b) Integrating scene information understanding and goal intent derivation to enhance motion state learning.

Previous studies mainly model motion states through two paradigms. The first paradigm, as shown in Figure 1 (a), models motion states directly from raw trajectory sequences [2, 12, 27, 51]. Although it can roughly capture motion states, it fails to uncover many deep characteristics due to the subtle nature of human motion patterns, resulting in information loss. The second paradigm uses high-dimensional space to intricately model motion states [4, 24]. However, it overlooks the interaction among motion states, scene information and motion intentions. This interaction can manifest as repulsion from scene edges and attraction towards expected positions, thereby generating future motion trends. Consequently, it is hard for the paradigm to accurately model motion states under the conditions of interaction between scene and intentions.

To overcome these limitations, we propose the Motion-Scene-Goal Aware Network (MSGANet) model, as illustrated in Figure 1 (b). According to the research [3, 41], intent can shape the kinematics of action, and conversely intent can be inferred from observed behavior. Inspired by this perspective, our method derives pedestrians' destinations as goal intentions by implementing interactions between the learned motion states and the scene semantics. Simultaneously, the goal intention is used to reshape the representation of motion states, guiding the prediction process of future motion trajectories from motion states fused with scene context. Specifically, we utilize the Motion State Attention Temporal Convolutional Network to extract temporal information from motion trajectories. Then the spatial distribution of motion states is obtained based on self-supervised learning with the state log-likelihood loss. Subsequently, we introduce the Motion-Scene Aware Transformer and the Motion-Goal Aware Transformer. The former utilizes the self-attention module to extract aware fusion features of motion state and scene information. The latter combines the cross-attention module to embed goal intentions into the motion states. Furthermore, we propose a reparameterized trajectory log-likelihood loss to precisely converge the goal-guided prediction trajectory distribution. Experiments on ETH [28]-UCY [17] and SDD [29] datasets verify the effectiveness and superiority of our approach.

2 Related Work

Motion state learning. Studying the motion state information in trajectories can effectively enhance the motion representation capability of features. Early approaches utilize latent variables within RNN [2, 27, 39] for description. The subsequent research [26, 38] have employed spatio-temporal graph models to transform the motion velocity into parameters of Gaussian distributions to predict the trend direction of future trajectories. However, these methods are limited by the lower-dimensional trajectories, lacking in-depth analysis of motion state representations. With the rise of self-supervised learning [10, 20], more studies are turning to the use of different feature descriptors [4, 42] to analyze the motion states within trajectories. Elevating the dimension of the motion states can address the issue of information loss caused by sparse trajectory coordinate data. Nonetheless, the motion states extracted by these methods are limited to the observed information. We enhance the representation capability of future motion trends by integrating motion state learning with scene understanding and interaction with goal intent.

Scene interaction and spatial probability estimation. Early approaches [6, 8, 37] enhance models' representation ability of multimodal information by extracting scene features and integrating them with trajectory features. Recent studies have employed cross-attention [25, 50] and conditional attention mechanisms [33] to facilitate interaction between trajectories and scenes. However, these methods for multimodal feature fusion exhibit deficiency in spatial alignment. Therefore, the studies [7, 9, 18, 19, 23] have proposed mapping trajectories onto graphs, aligning them with scenes, and predicting the spatial probability distribution of trajectories. Meanwhile, we promote the perception of global scene information by leveraging self-attention-based motion-scene aware learning.

Goal intention driven approach. Many studies [15, 51] refine the state representation based on interactions with the intentions of other pedestrians. Some subsequent studies have focused on trajectory prediction driven by intention [11, 34, 36, 44, 46, 53], and goal intention playing a prominent role. Some approaches [22, 23, 47, 49] adopt goal intentions as latent variables and learn their distribution through generative models. Then these methods sample multimodal trajectories guided by various goal intentions during inference, further enhancing predictive performance. Our method differs in that we achieve the interaction between motion states and goal intentions through cross-attention-based motion-goal aware learning, then combine it to refine the trajectory prediction process guided by goals.

3 Approach

The proposed MSGANet is shown in Figure 2. Let $x = \{x^t | t = 1, 2, \ldots, t_o\} \in R^{t_o \times 2}$ be the observed trajectory of the pedestrian consisting of t_o frames. $H \in R^{h \times w \times 3}$ is the scene image with a height of h and a width of w. We firstly extract the spatial features of motion states by the Motion State Distribution Learning module (MSDL). Subsequently, based on the Transformer model [40], the spatial features of motion states and scene semantics are fused using the self-attention mechanism in the Motion-Scene Aware Learning module to generate the spatial distribution of goal intention. Ultimately, the Motion-Goal Aware Learning module combines the cross-attention mechanism to learn feature interactions between motion states and goal intentions. And the interaction features are used to guide the prediction of the future trajectory $y = \{y^t | t = 1, 2, \ldots, t_f\} \in R^{t_f \times 2}$ consisting of t_f frames.

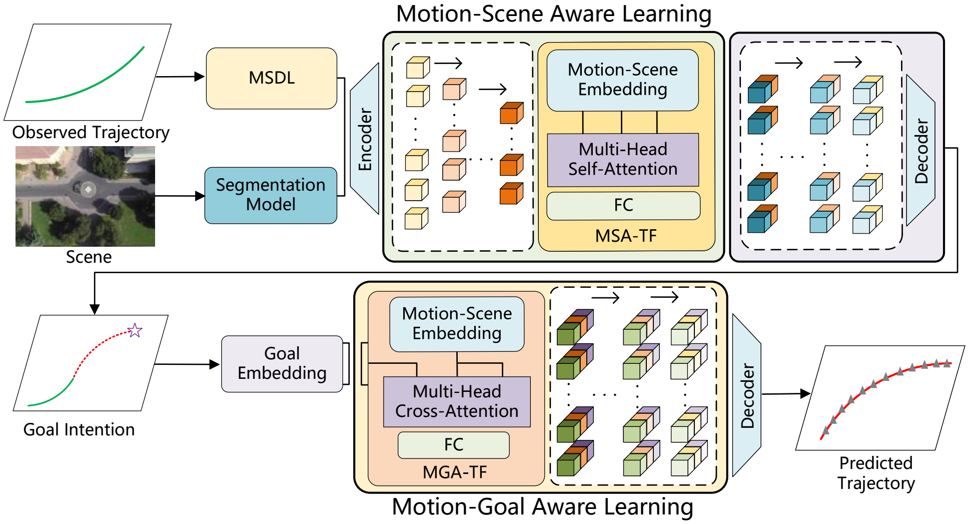

Figure 2: Architecture of the proposed MSGANet.

3.1 Motion State Distribution Learning

To model the spatial distribution of movement trajectories, we conduct motion state distribution learning implemented by Motion State Attention Temporal Convolutional Network (MS-ATCN), which leverages self-attention mechanisms to extract temporal information from the initial features of trajectories. And it employs multilayer temporal convolutional networks to transform the temporal state features into a two-dimensional Gaussian distribution $\{\mu_o \in R^{t_o \times 2}, \sigma_o \in R^{t_o \times 2}, \rho_o \in R^{t_o \times 1}\}$ for representing the motion state distribution. Inspired by the training objective in the study [26], self-supervised learning based on maximizing the state log-likelihood loss L_{sll} is employed to enhance its ability to represent trajectory spatial positions, as shown in Eq. (1). P denotes the Gaussian probability distribution.

$$L_{sll} = -\log(P(x|\mu_o, \sigma_o, \rho_o)). \tag{1}$$

3.2 Motion-Scene Aware Learning

After obtaining the distribution of the motion state, we aim to map it into the scene. By interacting with motion states and the scene, extracting profound representations of motion trends enables the precise derivation of goal intent. Firstly, the distribution is transformed into a motion state spatial probability map H_x through a sequence-to-graph transformation operation. Specifically, each grid in H_x is associated with the probability under the motion state distribution. Then, the normalized motion state spatial map $\widetilde{H}_x$ is obtained through linear normalization, as shown in Eq. (2).

$$\widetilde{H}_x(i) = \frac{H_x(i) - \min H_x}{\max H_x - \min H_x}, i \in \{(0,0),(0,1),\ldots,(h-1,w-1)\}. \tag{2}$$

Meanwhile, the segmentation model [30] is employed to extract semantic information from the scene, which is concatenated with the $\widetilde{H}_x$ along the spatial dimension to obtain the initial motion-scene spatial alignment graph H_{ms}. Drawing inspiration from Y-Net [23] and considering that goal intent encompasses long-term information and maintains a certain distance from the motion state distribution, we adopt the U-Net framework [30] to extract multi-scale motion-scene features. In the core, we introduce the Motion-Scene Aware Transformer (MSA-TF), which extracts the motion-scene embedding features ε_c from the deepest layer feature and then employs the multi-head self-attention module [40] to integrate global information. The final output is the motion-scene aware feature ε_{ms}.

$$\varepsilon_{ms} = \mathrm{fc}(\mathrm{softmax}(\frac{\phi_q(\varepsilon_c)\phi_k(\varepsilon_c)^T}{d_{embed}})\phi_v(\varepsilon_c) + \varepsilon_c), \tag{3}$$

where ϕ_q, ϕ_k, ϕ_v and d_{embed} represent embedding functions and feature dimension, respectively. fc denotes the function of the fully connected layer. Finally, ε_{ms} is fused with multi-scale features through layer-wise upsampling, and decoded to obtain the goal spatial map H_g. The cross-entropy loss with the normal distribution of ground truth goal intentions is used to supervise the estimation of goal intention.

3.3 Motion-Goal Aware Learning

We reshape the deep features of motion states using goal intention to enhance their representation capability for future trend prediction and utilize them for predicting future trajectory distributions. Through layer-wise downsampling, we obtain multi-scale spatial distribution maps of goal intent. In combination with the teacher force strategy [23], the ground truth is used as input during training, while H_g is employed during inference. Similar to Motion-Scene Aware Learning, we introduce the Motion-Goal Aware Transformer (MGA-TF) in the core, which utilizes the multi-head cross-attention module [40] to achieve motion-goal interaction under scene understanding, resulting in the motion-goal aware feature ε_{mg}.

$$\varepsilon_{mg} = \mathrm{fc}(\mathrm{softmax}(\frac{\phi_q'(\varepsilon_g)\phi_k'(\varepsilon_s)^T}{d_{embed}})\phi_v'(\varepsilon_s) + \varepsilon_s), \tag{4}$$

where ε_g and ε_s represent the embedded goal intent feature and the fused motion-scene feature, respectively. $\phi_q', \phi_k', \phi_v'$ denote the embedding functions in cross-attention. Compared to self-attention, cross-attention can address the issue of missing fusion information due to the sparsity of goal intention relative to the motion-scene information. ε_{mg} is layer-wise connected with multi-scale motion-scene features and goal spatial map, and decoded into the spatial distribution of the predicted trajectory H_p. The learning process is supervised by combining cross-entropy loss with the normal distribution of ground truth future trajectories. We employ the softargmax operation in the study [23] to sample the final predicted trajectory. Moreover, to refine the regression process from motion states to the predicted trajectory distribution, we propose the trajectory log-likelihood loss L_{tll} based on reparameterized probability spatial distribution. It is described in Eq. (5),

$$\begin{aligned}
&\widetilde{H} = \mathrm{softmax}(H_p),\ \ \mu_p^t = \sum_{j,k}(j,k) \times \widetilde{H}_p^t(j,k),\ \ \sigma_p^t = \sqrt{\sum_{j,k}((j,k)-\mu_p^t)^2 \times \widetilde{H}_p^t(j,k)}, \\
&\rho_p^t = \sum_{j,k} \frac{(j-\mu_{p,0}^t) \times (k-\mu_{p,1}^t) \times \widetilde{H}_p^t(j,k)}{\sigma_{p,0}^t \times \sigma_{p,1}^t},\ \ t \in \{1,2,\dots,t_f\}, \\
&L_{tll} = -\log(P(y|\mu_p,\sigma_p,\rho_p)),\ \ j \in \{0,1,\dots,h-1\},\ \ k \in \{0,1,\dots,w-1\},
\end{aligned} \tag{5}$$

where $\{\mu_p,\sigma_p,\rho_p\}$ represents the reconstructed parameters. $\{\mu_p^t,\sigma_p^t,\rho_p^t\}$ and $\widetilde{H}_p^t$ respectively denote the values corresponding to the t th index along the temporal dimension of $\{\mu_p,\sigma_p,\rho_p\}$ and $\widetilde{H}_p$. $\mu_{p,l}^t$ and $\sigma_{p,l}^t$ respectively denote the values corresponding to the l th index along the last dimension of μ_p^t and σ_p^t.

4 Experiment

4.1 Experimental Setup

Dataset. We evaluate the prediction performance of the proposed method on Stanford Drone Dataset (SDD) [29] and ETH [28]-UCY [17] datasets, which are widely-used benchmarks for pedestrian trajectory prediction. The SDD dataset employs drones to capture bird's eye view images in 20 different scenes on campus, encompassing over 11,000 pedestrians. ETH-UCY dataset comprises a total of over 1500 pedestrians from five subdatasets, including ETH and HOTEL from the ETH dataset, and UNIV, ZARA1, and ZARA2 from the UCY dataset. We follow the dataset split of research [13, 31], and adopt 2.5 as the sampling frequency to extract multiple samples. The division results in 3.2-seconds ($t_o = 8$) observed trajectories and 4.8-seconds ($t_f = 12$) future trajectories. Additionally, for the ETH-UCY dataset, we employ the leave-one-out evaluation methodology, where four sub-datasets are used for training, and the remaining one is used for testing.
Evaluation Metrics. We evaluate the accuracy of future trajectory prediction and goal intention estimation using the Average Displacement Error (ADE) and the Final Displacement Error (FDE). Considering the randomness in pedestrian trajectories, we follow the research [13, 22, 23] to introduce the best-of-N evaluation protocol, which means 20 trajectories need to be generated and those with the best performance are used for comparison.
Implementation details. Our model is primarily based on the U-net [30] framework incorporating the Transformer [40] model, with feature dimensions are set layer-wise to [32, 32, 64, 64, 64]. The Transformers are all configured with 4 heads and feature dimensions of 256. Additionally, we implement semantic segmentation on scene H using a pre-trained segmentation model provided by research [23], and sample from the spatial probability distribution of goal intentions multiple times to generate 20 trajectories for assessment. The data is augmented spatially by rotation, flipping, scaling and perspective transformation. The method is trained using Adam [14] optimizer for 150 epochs with data batches of size 8. The initial learning rate is set to 0.0001.

4.2 Comparison with Other Methods

We compare our proposed method with 15 other baseline models, where Y-Net [23], Introvert [33], CMATP [50] and our MSGANet introduce the visual features from images or

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**

Table 1: The ADE/FDE metric comparison with other models on the ETH-UCY (meters) and SDD (pixels). 'AVG' represents the average prediction accuracy across the five sub-datasets in ETH-UCY datasets. The lower the metric, the better the prediction performance. The bold font represents the optimal performance within the current dataset.

Method	DESIRE [16]	TNT [52]	PECNet [22]	Y-Net [23]	MSGANet (Ours)
ADE/FDE	19.25/34.05	12.23/21.16	12.79/29.58	11.49/20.23	**11.22/19.58**

Table 2: Comparison results with a sampling quantity of 5 on the SDD dataset.

videos. Table 1 shows the best predictive performance among the 20 generated trajectories. Our method achieves state-of-the-art results on multiple datasets. Compared to the recent approach [50], our method demonstrates an 18.18% improvement in ADE/FDE on the ETH-UCY dataset. Among the subdatasets, ETH and HOTEL primarily consist of long-distance movement trajectories with stable motion patterns. Effective prediction of motion trends under corresponding modalities is achieved through motion state learning, thus exhibiting optimal performance. For datasets such as Zara1 and Zara2, which encompass more modalities, joint aware learning of motion states and goal intentions results in the generation of diverse motion trajectories, with optimal FDE metrics. This also signifies that motion state learning facilitates the inference of multimodal goal intentions.

On the SDD dataset, characterized by richer scene information and multiple motion modalities, our method yields state-of-the-art results. When comparing individual approaches, MSGANet exhibits significant enhancements with respective improvements of 10.16% and 8.29% in ADE/FDE compared to the method [44] without motion state learning. Additionally, compared to the method [35] without goal intention deivation, there is an 8.51% improvement in FDE. By integrating the learning of motion states and goal intentions, we enhance the coherence of their spatial patterns. And coupling cross-attention guides the future trajectory trends of motion states, thereby enhancing the prediction performance of

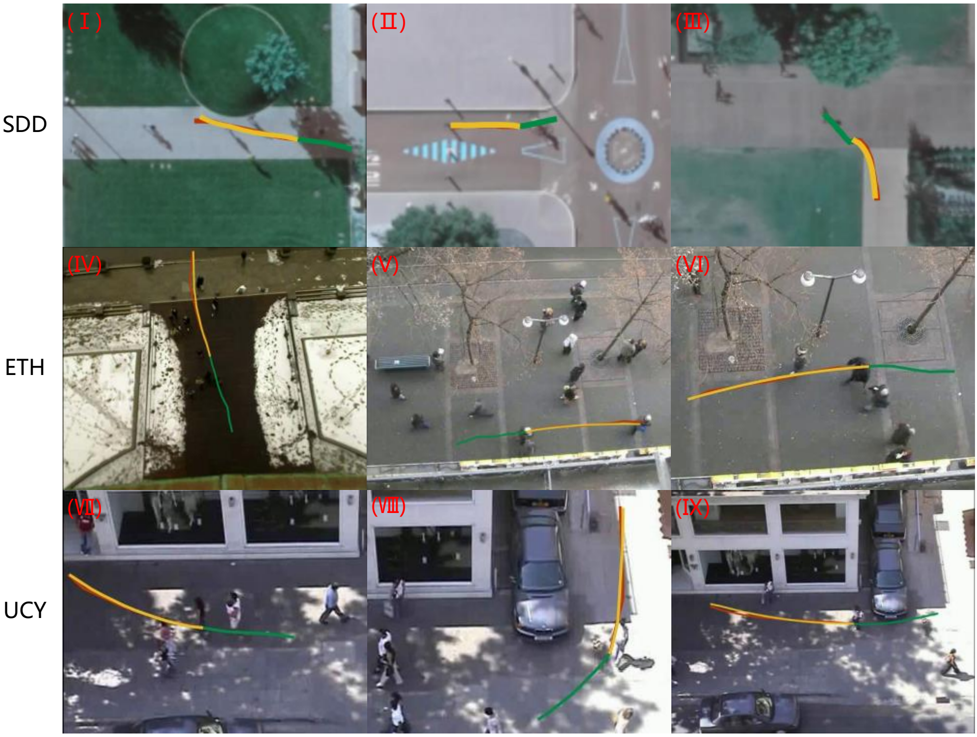

Figure 3: Visualization of the predicted trajectories. The green and red curves denote observed and future trajectories, respectively. The yellow curves represent the predicted trajectories by our method.

future trajectory.

To provide more deterministic results for comparison, we supplement the comparison of the best results obtained under sampling 5 trajectories, as shown in Table 2. Through comparative analysis, our method demonstrates superior performance, particularly in achieving a notable 0.65 reduction in FDE metric compared to the research [23]. Therefore, leveraging the integration of scene semantics and goal intent information enhances the representation capacity of motion state features, enabling effective application in the deterministic trajectory prediction task with fewer samples.

4.3 Visualization Results

We conduct a comprehensive qualitative analysis of the predicted trajectories generated by our proposed method across multiple instances. These instances encompass various pedestrian movements, including straightforward linear motions, directional shifts, and complex turns. The best-performing results among the 20 trajectories predicted by our method are shown in Figure 3, where for clarity the part of the scene images are visualized. Our method exhibits superior predictive capabilities across diverse motion patterns. In the showcase, our approach demonstrates notable proficiency in predicting trajectories involving extended linear movements, as observed in the instances (I, V). It also accurately anticipates regular turns occurring at long distances, as evidenced in the instances (VII, IX). Moreover, when confronted with abrupt turns, as depicted in the instances (II, IV), our method achieves precise predictions through the goal intention estimation. Even in instances involving significant turning maneuvers, such as those illustrated in the instances (III, VIII), it still accurately predicts pedestrians' destinations by combining scene information with motion state learning. Its consistent performance across diverse scenarios demonstrates its ability to capture detailed information in pedestrian motion states and accurately predict destinations. To validate the effectiveness of motion state learning, we exponentiate the learned state space distribution and visualize it, as shown in Figure 4. The motion state distribution not only encompasses the positional information of observed trajectories but also maps the future motion trend to some extent.

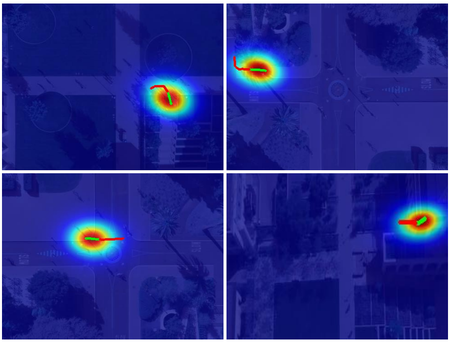
Figure 4: Motion State Distribution.

4.4 Ablation Study

To validate the effectiveness of the proposed modules in the model, we conduct extensive ablation experiments on the SDD dataset. These experiments primarily focus on Motion State Distribution Learning module (MSDL), Motion-Scene Aware Learning module (MSAL), Motion-Scene Aware Transformer (MSA-TF) and Motion-Goal Aware Transformer (MGA-TF). Additionally, we evaluate the effectiveness of the reparameterized trajectory log-likelihood loss L_{tll} used for constraining the predicted trajectory distribution. The results of the ablation study are presented in Table 3, where w/o m denotes the method without module m. It can be seen that each module effectively enhances the predictive performance of our method. The MSDL and MSAL modules improve goal intention derivation, reducing FDE and aiding goal-guided trajectory prediction. MGA-TF and L_{tll} respectively enhance motion-goal interaction and expression of decoded trajectory predictions, effectively improving ADE.

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_{tll}	7.73/11.63
Ours	**7.69/11.61**

Table 3: The results of the ablation study.

Meanwhile, we visualize the prediction results of the methods without MSDL and without MSAL, as shown in Figure 5. The part of the scene images are visualized for clarity. Similarly, the green and red curves indicate observed and ground truth future trajectories. The yellow curve depicts the predicted trajectories by our method. Then the dark blue and purple curves represent trajectories predicted by the methods without MSDL and without MSAL, respectively. Through comparison, it can be seen that the former exhibits deviations in the estimated goal intention due to the absence of motion state representation, leading to directional and displacement biases. The latter lacks scene interaction, resulting in predicted

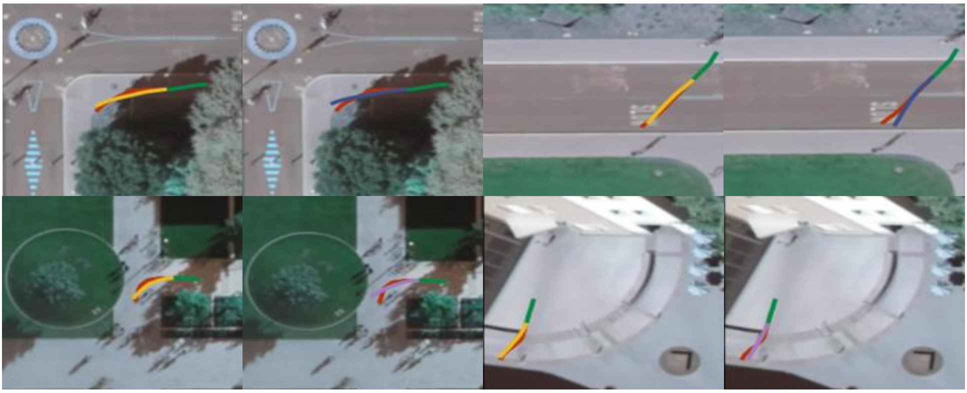

Figure 5: The visualization of ablation study.

trajectories that collide with the edges of the scene. In all instances, our method consistently demonstrates accurate prediction results.

5 Conclusion

This paper proposes the Motion-Scene-Goal Aware Network (MSGANet), a pedestrian trajectory prediction framework that integrates scene understanding and goal intent inference to enhance motion state learning. MSGANet employs temporal convolution to adaptively learn the distribution of the motion state. It combines this with Transformer models based on self-attention to facilitate aware fusion between motion states and scene information, thereby enhancing the derivation of goal intentions. Subsequently, a Transformer model with cross-attention is utilized to facilitate interaction between goal intentions and the motion-scene fusion feature, guiding the trend of trajectory decoding and enhancing the ability of features to represent future motion trends. Extensive experiments on ETH-UCY and SDD datasets validate the superior performance of our model and the effectiveness of each module. Our approach provides reference for researching the relation between human actions and intentions. Future work will further refine the modeling of motion states based on pedestrian psychology and motion physics, aiming to refine the transition process from current motion states to motion intent, while enhancing physical interpretability.

Acknowledgement

This research was supported by the grants National Natural Science Foundation of China (No. U23B2060, 62088102), the Youth Innovation Team of Shaanxi Universities, and the Fundamental Research Funds for the Central Universities.

References

[1] Icek Ajzen. From intentions to actions: A theory of planned behavior. In *Action control: From cognition to behavior*, pages 11–39. 1985.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

[3] Caterina Ansuini, Andrea Cavallo, Cesare Bertone, and Cristina Becchio. Intentions in the brain: the unveiling of mister hyde. *The Neuroscientist*, 21(2):126–135, 2015.

[4] Inhwan Bae, Jean Oh, and Hae-Gon Jeon. Eigentrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *IEEE/CVF International Conference on Computer Vision*, pages 10017–10029, 2023.

[5] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online pomdp planning for autonomous driving in a crowd. In *IEEE International Conference on Robotics and Automation*, pages 454–460, 2015.

[6] Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human trajectory prediction via counterfactual analysis. In *IEEE/CVF International Conference on Computer Vision*, pages 9824–9833, 2021.

[7] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *IEEE/CVF International Conference on Computer Vision*, pages 921–930, 2019.

[8] Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *IEEE/CVF International Conference on Computer Vision*, pages 13158–13167, 2021.

[9] Isht Dwivedi, Srikanth Malla, Behzad Dariush, and Chiho Choi. Ssp: Single shot future trajectory prediction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2211–2218, 2020.

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[11] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. Loki: Long term and key intentions for trajectory prediction. In *IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.

[12] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. In *International Conference on Pattern Recognition*, pages 10335–10342, 2021.

[13] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Parth Kothari, Brian Sifringer, and Alexandre Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15556–15566, 2021.

[16] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.

[17] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664, 2007.

[18] 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 *IEEE/CVF conference on computer vision and pattern recognition*, pages 5725–5734, 2019.

[19] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10508–10518, 2020.

[20] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2023.

[21] Matthias Luber, Johannes A Stork, Gian Diego Tipaldi, and Kai O Arras. People tracking with human motion predictions from social forces. In *IEEE International Conference on Robotics and Automation*, pages 464–469, 2010.

[22] Karttikeya 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 *European Conference on Computer Vision*, pages 759–776, 2020.

[23] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *IEEE/CVF International Conference on Computer Vision*, pages 15233–15242, 2021.

[24] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5517–5526, 2023.

[25] Mancheng Meng, Ziyan Wu, Terrence Chen, Xiran Cai, Xiang Zhou, Fan Yang, and Dinggang Shen. Forecasting human trajectory from scene history. *Advances in Neural Information Processing Systems*, 35:24920–24933, 2022.

[26] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020.

[27] Jeremy Morton, Tim A. Wheeler, and Mykel J. Kochenderfer. Analysis of recurrent neural networks for probabilistic modeling of driver behavior. *IEEE Transactions on Intelligent Transportation Systems*, 18(5):1289–1298, 2017.

[28] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE international Conference on Computer Vision*, pages 261–268, 2009.

[29] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision*, pages 549–565, 2016.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[31] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2019.

[32] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700, 2020.

[33] Nasim Shafiee, Taskin Padir, and Ehsan Elhamifar. Introvert: Human trajectory prediction via conditional 3d attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16810–16820, 2021.

[34] Shreya Sharma, Jigyasa Gupta, Shreshth Tuli, Rohan Paul, et al. Goalnet: Inferring conjunctive goal predicates from human plan demonstrations for robot instruction following. *arXiv preprint arXiv:2205.07081*, 2022.

[35] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *IEEE/CVF International Conference on Computer Vision*, pages 9675–9684, 2023.

[36] Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. In *IEEE/CVF International Conference on Computer Vision*, pages 13250–13259, 2021.

[37] Haowen Tang, Ping Wei, Huan Li, Jiapeng Li, and Nanning Zheng. Relation reasoning for video pedestrian trajectory prediction. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2022.

[38] Haowen Tang, Ping Wei, Jiapeng Li, and Nanning Zheng. Evostgat: Evolving spatiotemporal graph attention networks for pedestrian trajectory prediction. *Neurocomputing*, 491:333–342, 2022.

[39] Chaofan Tao, Qinhong Jiang, Lixin Duan, and Ping Luo. Dynamic and static context-aware lstm for multi-agent motion prediction. In *European Conference on Computer Vision*, pages 547–563, 2020.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[41] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6801–6809, 2018.

[42] 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*, pages 682–700, 2022.

[43] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6507, 2022.

[44] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2022.

[45] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision*, pages 511–528, 2022.

[46] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1352, 2011.

[47] Yu Yao, Ella Atkins, Matthew Johnson-Roberson, Ram Vasudevan, and Xiaoxiao Du. Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robotics and Automation Letters*, 6(2):1463–1470, 2021.

[48] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.

[49] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *European Conference on Computer Vision*, pages 376–394, 2022.

[50] Mayssa ZAIER, Hazem Wannous, Hassen Drira, and Jacques boonaert. Cross-modal attention for accurate pedestrian trajectory prediction. In *British Machine Vision Conference*, 2023.

[51] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019.

[52] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pages 895–904, 2021.

[53] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C Karen Liu, and Leonidas J Guibas. Gimo: Gaze-informed human motion prediction in context. In *European Conference on Computer Vision*, pages 676–694, 2022.