A Unified Environmental Network for Pedestrian Trajectory Prediction

Yuchao Su[1], Yuanman Li[1*], Wei Wang[2], Jiantao Zhou[3], Xia Li[1]

[1]Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University

[2]Department of Engineering, Shenzhen MSU-BIT University

[3]Department of Computer and Information Science, University of Macau

yuchaosu@126.com, {yuanmanli, lixia}@szu.edu.cn, ehomewang@ieee.org, jtzhou@um.edu.mo

Abstract

Accurately predicting pedestrian movements in complex environments is challenging due to social interactions, scene constraints, and pedestrians' multimodal behaviors. Sequential models like long short-term memory fail to effectively integrate scene features to make predicted trajectories comply with scene constraints due to disparate feature modalities of scene and trajectory. Though existing convolution neural network (CNN) models can extract scene features, they are ineffective in mapping these features into scene constraints for pedestrians and struggle to model pedestrian interactions due to the loss of target pedestrian information. To address these issues, we propose a unified environmental network based on CNN for pedestrian trajectory prediction. We introduce a polar-based method to reflect the distance and direction relationship between any position in the environment and the target pedestrian. This enables us to simultaneously model scene constraints and pedestrian social interactions in the form of feature maps. Additionally, we capture essential local features in the feature map, characterizing potential multimodal movements of pedestrians at each time step to prevent redundant predicted trajectories. We verify the performance of our proposed model on four trajectory prediction datasets, encompassing both short-term and long-term predictions. The experimental results demonstrate the superiority of our approach over existing methods.

Introduction

Understanding the motion behaviors of pedestrians plays an important role in many applications. However, accurately predicting the trajectories of pedestrians is rather challenging due to environmental impacts and diverse motions of pedestrians. First, a pedestrian's motion is impacted by surrounding scene elements and pedestrians. Second, the motion shows high multimodal properties, i.e., a past trajectory can correspond to multiple paths leading to a destination.

Motions of pedestrians are always impacted by the surrounding environment, and environmental impacts can be classified as static and dynamic impacts. Static impacts are presented by scene constraints, which are unchanged for a long-term period. Dynamic impacts refer to social interactions between pedestrians. Some methods (Sadeghian et al. 2019; Yuan et al. 2021) have separately characterized scene impacts by CNN models and social interactions by sequential models, and then both scene and interaction features are flattened into a vector to show environmental impacts. These methods suffer from two inherent limitations: First, due to modality differences, scene constraints and social interactions are separately modeled. However, scene elements can affect social interactions. For instance, when two pedestrians are very close with a wall between them, modeling their social interactions without considering scene elements fails to accurately reflect their interactions. Second, CNN-extracted scene features describe compact global semantic information. However, most of the scene constraints imposed on pedestrians are local, primarily related to the scene around pedestrians. Directly interacting the compact global semantic scene features with trajectory features is challenging due to the absence of spatial correspondences, making it difficult to reflect the impacts of the local scene on pedestrians.

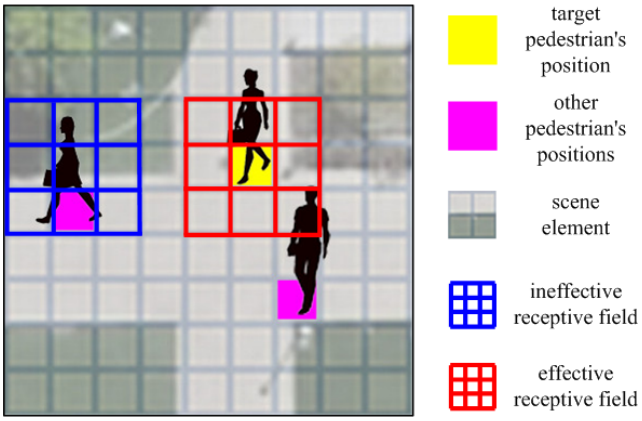

Figure 1: The shortcoming of CNN in trajectory prediction. It is difficult to characterize impacts from scene elements and other pedestrians to the target pedestrian, because CNN doesn't always learn the target pedestrian's position.

Recently, the work (Mangalam et al. 2021) employs a CNN model to extract both trajectory and scene features for trajectory prediction. However, the model cannot present social interactions between pedestrians, and struggles to map scene features into scene constraints, as shown in Fig. 1. In

*Corresponding author.

(a) crowded predicted goals (b) tedious predicted motions

Figure 2: The deficiency of goal-guided strategies. Goal-guided methods obtain a tedious movement behavior at non-goal time steps. The stars and diamonds indicate the future and predicted destinations, respectively. The purple curves mean the predicted paths. The yellow and brown curves represent segments of future trajectories that are being approached and not being approached, respectively.

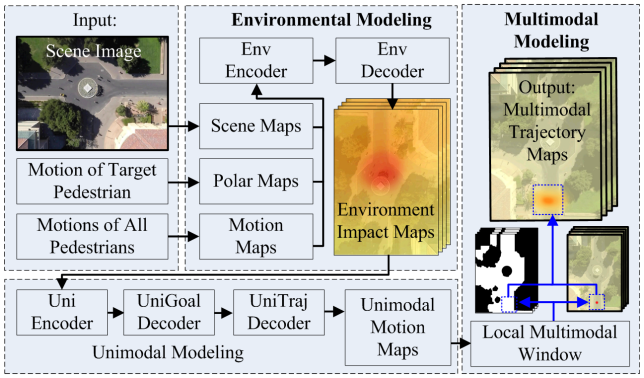

Figure 3: The framework of UEN.

Fig. 1, the receptive field (the red grid) of CNN can correctly construct spatial relationships due to the known position of the target pedestrian. When the receptive field (the blue grid) doesn't capture the target pedestrian, CNN only extracts features but cannot effectively depict the spatial relationships between these features and the target pedestrian.

It is difficult to generate reasonable multimodal trajectories due to diverse motions of pedestrians. Most methods proposed recently (Gu, Sun, and Zhao 2021; Mangalam et al. 2020) adopt goal-guided strategies to obtain multiple goals and then each goal guides a unimodal trajectory. However, goal-guided methods probably produce a tedious movement behavior in non-goal time steps. As shown in Fig. 2(a), the predicted destinations are crowded so the predicted trajectories deviate from the future trajectory. In Fig. 2(b), the predicted positions at many time steps fail to approach the actual ones due to losses of diverse motion modes at non-goal time steps. Accordingly, multimodal motions at each time step should be considered to effectively characterize the potential movement patterns of pedestrians.

We propose a unified environmental network to address the above issues. Our contributions are summarized below:

- We propose a polar-based method to reflect the distance and direction relationship between any position in an environment and a target pedestrian. This approach allows us to simultaneously model scene constraints and social interactions in feature maps, enabling us to accurately construct environmental impacts on a target pedestrian.
- We propose a local multimodal window to capture essential local features for modeling multimodal trajectories in feature maps at each time step. This method can generate dissimilar positions in local scenes to depict diverse and reasonable motions of pedestrians.
- Compared with ten models on four datasets, our model achieves the best-predicted results in most of the scenes.

Related Works

Modeling of Social Interactions: Some works (Mangalam et al. 2020) adopted social pooling mechanisms to model social interactions. To intuitively depict the interactions, graph-based methods have been employed in many recent works (An et al. 2022; Lv et al. 2023; Wang et al. 2023; Tang et al. 2022; Zhou et al. 2022). However, existing methods characterize social interactions by sequential features and have a difficulty to present scene constraints.

Extraction of Scene Features: CNNs (Zhang et al. 2021) were widely used in image processing. Most researchers adopted CNNs to extract scene features from scene images (Chen et al. 2021; Yuan et al. 2021). However, scene features were usually flattened as a vector to integrate trajectory features, and the pixel alignment and meaningful spatial signals were destroyed. Recently, (Mangalam et al. 2021) employed a CNN model to extract scene features and predict future motions in feature maps. However, as mentioned in Fig. 1, these methods extract scene features but struggle to accurately construct spatial relations for a target pedestrian.

Multimodal Motion Prediction: Generative models (Chen et al. 2022) were adopted in many tasks to produce diverse audio, text, and images. Generative adversarial networks (Liang et al. 2021) and conditional variational autoencoders (CVAE) (Xie et al. 2022) were widely used to generate multimodal positions. Very recently, flow-based models (Liang et al. 2023) and diffusion models (Mao et al. 2023) were also applied in trajectory prediction tasks. Many methods proposed recently adopted goal-guided strategies and presented remarkable performance improvements (Li et al. 2023; Duan et al. 2022; Xu et al. 2022), but they ignored diverse motions in non-goal time steps, as shown in Fig. 2.

Unified Environmental Network

We propose **U**nified **E**nvironmental **N**etwork based on CNN for trajectory prediction, called UEN, as shown in Fig. 3.

Definition of Trajectory Prediction

Define a sequence of a pedestrian's true locations as $[l_1, ..., l_h, l_{h+1}, ..., l_{h+p}]$, where h and p represent the number of past and future time steps, respectively, and l_t is a position at the t-th time step. The objective of models is to reduce errors between $[l_{h+1}, ..., l_{h+p}]$ and predicted ones.

Environmental Modeling

To obtain environmental impacts on pedestrians, firstly, we secure scene features and past motion behaviors in the same

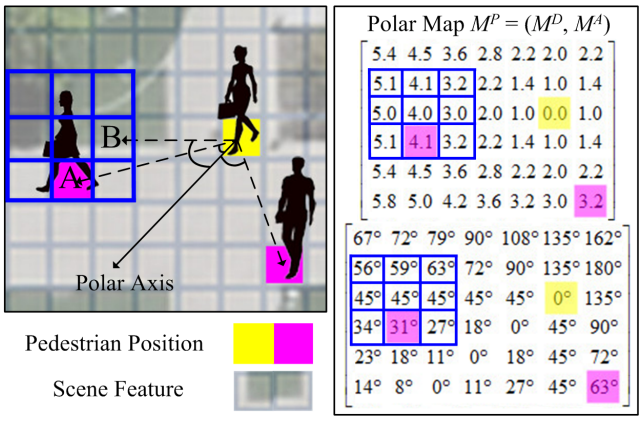

Figure 4: The proposed polar-based method. The polar map can characterize the spatial relationship between the target pedestrian and any position in the environment, even when the receptive field (the blue grid) of CNN doesn't learn the target pedestrian's position.

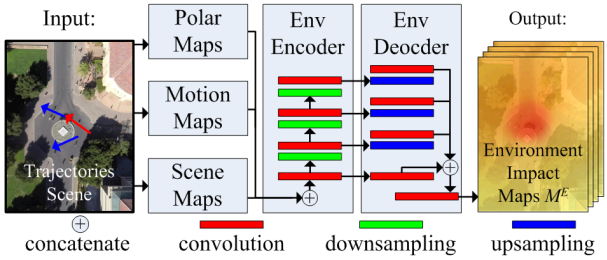

Figure 5: Environmental Modeling.

feature modality. Then, we propose a polar-based method to construct the spatial relationship. Last, a CNN model is designed to simultaneously present the dynamic and static environmental impacts on a target pedestrian.

Scene Maps: A semantic segmentation model, which is similar to U-net (Ronneberger, Fischer, and Brox 2015), is pretrained to obtain semantic features as the scene features, as executed in (Mangalam et al. 2021). The semantic classes include pavements, roads, buildings, lands, trees, and others. Thus, the output size is $W \times H \times 6$, where W and H indicate the width and height of the input image, respectively.

Motion Maps: For making motions of pedestrians in a scene have the same modality as scene feature maps, We transform all pedestrians' past positions into heatmaps:

$$H_t^i(u, l_t^i) = 2||u - l_t^i||/max||v - l_t^i||, v \in I \quad (1)$$

$H_t^i(u, l_t^i)$ indicates the heatmap of the i-th pedestrian at the t-th time step, where l_t^i is the pedestrian's position, u represents the coordinate of any position in a scene, and I is the position coordinate set of the scene, as employed in (Mangalam et al. 2021). However, CNNs cannot handle a variable number of channels due to the dynamic quantity of pedestrians. To solve this problem, we integrate the heatmaps of all pedestrians at the t-th time step into a feature map:

$$M_t = min2D(\{H_t^i | i = 1, ..., k\}), \quad (2)$$

where k means the number of pedestrians, and $min2D$ picks up the minimal value of each position in a scene from all heatmaps. By this method, all pedestrians' motions can be expressed and there is the same modality for trajectory and scene features, which avoids flattening scene features as a vector and the destruction of meaningful spatial information.

Polar Maps: By the above methods, we can effectively capture the past motion behaviors of pedestrians and scene features in feature maps. Now, we introduce a polar-based method to model the spatial relationship between the target pedestrian and any position within the environment, for overcoming the issue mentioned in Fig. 1.

A polar map $M_t^P = (M_t^D, M_t^A)$ at the t-th past time step characterizes the distance and direction relationship between a target pedestrian and any position in an environment, by

$$M_t^D(u, l_t) = \frac{||u - l_t||}{\sqrt{W + H}}, \quad (3)$$

$$M_t^A(u, l_t, l_t + 1) = \frac{(l_{t+1} - l_t)(u - l_t)^T}{||(l_{t+1} - l_t)|| \times ||(u - l_t)||}. \quad (4)$$

M_t^D and M_t^A respectively depict the distance and direction relationship from a target pedestrian to any position u in an environment, where l_t and $l_t + 1$ indicate the positions of a target pedestrian at current and next time steps, respectively. We give a visual case in Fig. 4 for understanding the polar-based method, where M^D and M^A exclude the subscript and their values are represented by Euclidean distances and angle degrees. The polar axis indicates the motion direction of the target pedestrian and her current position is highlighted by the yellow cube. Fig. 4 demonstrates that the CNN with our method has the ability to learn the distance and direction relationship between the target pedestrian and the pedestrian A, as well as the scene element B, even when CNN doesn't capture the position of the target pedestrian. In other words, the proposed polar-based method can construct the distance and direction relationship between any position in the environment and the target pedestrian.

Environmental Impact Maps: To obtain dynamic and static impacts on a target pedestrian, we adopt a CNN model to simultaneously model scene constraints and social interactions, as shown in Fig. 5. The polar, motion, and scene maps are inputted into Env Encoder to extract dynamic and static environmental features with different scales. By Env Decoder, these features with different scales are integrated to achieve comprehensive impacts of the environment.

By Environmental Modeling, the social interactions and scene constraints can be simultaneously modeled in feature maps. This overcomes the issue that existing models fail to effectively integrate both trajectory and scene features. Moreover, the polar-based method can accurately characterize the spatial relationship between a target pedestrian and any position in an environment, which effectively presents environmental impacts on the motion of a target pedestrian.

Unimodal Modeling

Before predicting multimodal trajectories, we model unimodal future behaviors of the target pedestrian to get a deterministic motion, as shown in Unimodal Modeling of Fig. 3.

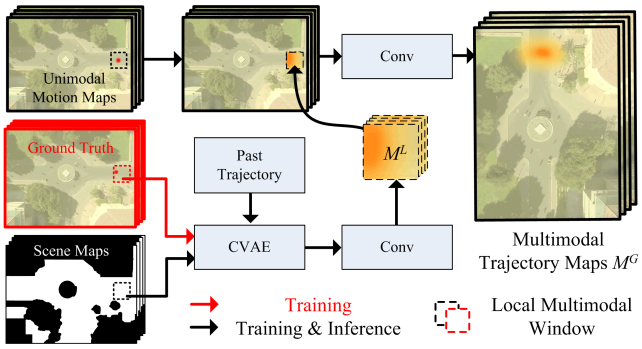

Figure 6: Multimodal Modeling.

The encoder and decoders follow the same structure as described in (Mangalam et al. 2021). The processes are below. Step 1: Concatenate the environment impact maps and target pedestrian's heatmaps and then input the joint features into Uni Encoder to get multiscale features. Step 2: Feed the multiscale features into UniGoal Decoder to obtain a goal map with the size $H \times W \times 1$ for a potential destination. Step 3: Concatenate the multiscale features and the goal map and then input the joint features into UniTraj Decoder to secure the unimodal motion maps U with the size $H \times W \times p$.

Next, we use U to characterize multimodal motions at each time step by Multimodal Modeling.

Multimodal Modeling

Many local features in feature maps cannot effectively characterize meaningful motions and future trajectories are mainly impacted by local scene elements. To generate reasonable future motions in feature maps, we propose a local multimodal window to capture essential local features for modeling multimodal positions at each time step.

Local Multimodal Window: We firstly decide the centers C of the local multimodal windows in all future time steps:

$$C = Softmax2D(U), \tag{5}$$

where $C = [C_1, ..., C_p]$, and $Softmax2D$ captures the two-dimensional coordinate C_i $(i = 1, ..p)$ with the maximal value in each channel (time step) from U. The local multimodal window Win gets the local motion features LU with the size $z \times z \times p$ to present reasonable motion behaviors in local regions:

$$LU = Win(U|C, z). \tag{6}$$

In the future time steps, other pedestrians' dynamic behaviors are unknown but the scene still impacts the future motion of the target pedestrian. Thus, we also use the window to capture local scene features LS from the scene maps S:

$$LS = concat(Win(S|C_1, z), ..., Win(S|C_p, z)). \tag{7}$$

Training: We adopt a CVAE model to characterize diverse motion positions at each time step, as shown in Fig. 6. The true future trajectory is transformed to the heatmaps $\hat{M}^G$ and the local multimodal window obtains the local future features LG for modeling future motion modes:

$$LG = Win(\hat{M}^G|C, z). \tag{8}$$

After that, the past trajectory of the target pedestrian, the local scene features, and the local future features are inputted into the CVAE to obtain local multimodal maps M^L with the size $z \times z \times p$ for modeling diverse motion modes, where the CVAE is similar to that in (Mangalam et al. 2020). Subsequently, we overlap M^L into the unimodal motion maps to get global feature maps, according to the captured regions of the local multimodal windows in (6), i.e., $LU = M^L$. Last, the global feature maps go through a convolution operator to achieve the multimodal trajectory maps M^G.

We adopt the binary cross entropy (BCE) and Kullback–Leibler divergence (KLD) to train our model:

$$\begin{aligned} loss = &BCE(M^G, \hat{M}^G) + BCE(M_p^G, \hat{M}_p^G) \\ &+KLD(\mathcal{N}(\mu, \sigma), \mathcal{N}(0, \sigma')), \end{aligned} \tag{9}$$

where M_p^G and $\hat{M}_p^G$ respectively mean the multimodal trajectory map and true map in the goal time step. KLD makes the CVAE approximate the true probability distribution.

By the proposed local multimodal window, our model can capture essential local feature maps to effectively characterize reasonable and diverse motions of a target pedestrian in feature maps at all future time steps.

Inference: Due to the unknown probability distribution of future trajectories, a probability distribution $\mathcal{N}(0, \sigma')$ is adopted to replace the true probability distribution in the inference stage, as executed in (Mangalam et al. 2020). Positions of multiple predicted goals are obtained by the clustering centers based on K-means, as executed in (Mangalam et al. 2021). If the distance of two goals is smaller than r, we will use Multimodal Modeling to regenerate them, to avoid the holding of crowded predicted goals. Multimodal Modeling utilizes essential local features to effectively characterize reasonable and diverse motions at all time steps, which can overcome the issues mentioned in Fig. 2.

Experiments

We compare our model with ten algorithms on four datasets.

Experimental Details

We evaluate our proposed model on four pedestrian trajectory prediction datasets, i.e., ETH (Pellegrini et al. 2009), UCY (Lerner, Chrysanthou, and Lischinski 2007), Intersection Drone Dataset (inD) (Bock et al. 2020), and Stanford Drone Dataset (SDD) (Robicquet et al. 2016). Each dataset is used for a short-term prediction and a long-term prediction. The process of all data follows the implementation in (Mangalam et al. 2021). For the short-term prediction, the number h of the past time steps is set to 8, 8, 8, and 14 for ETH, UCY, SDD, and inD, respectively, and the number p of the future step times is set to 12, 12, 12, 21, respectively. For the long-term prediction, h is set to 4, 4, 5, and 5, respectively, and p is respectively set to 16, 16, 30, and 30.

We compare UEN with 7 SOTA and 3 latest models: TPNSTA (Li et al. 2022), AgentFormer (Yuan et al. 2021), VDRGCN (Su et al. 2022), Trajectron++ (Salzmann et al. 2020), BiTraP (Yao et al. 2021), PECNet (Mangalam et al. 2020), Y-net (Mangalam et al. 2021), LED(Mao et al. 2023), SICNet (Dong et al. 2023), and MSRL (Wu et al. 2023).

	TPNSTA	AgentFormer	VDRGCN	Trajectron++	BiTraP	PECNet	Y-net	UEN
eth_s	0.55/0.91	0.45/0.75	0.62/0.81	0.43/0.86	0.37/0.69	0.54/0.87	**0.28**/0.33	**0.28/0.31**
hotel_s	0.23/0.40	0.14/0.22	0.27/0.37	0.12/0.19	0.12/0.21	0.18/0.24	**0.10**/0.14	**0.10/0.13**
univ_s	0.52/1.10	0.25/0.45	0.38/0.58	0.22/0.43	**0.17/0.37**	0.35/0.60	0.24/0.41	0.24/0.40
zara1_s	0.34/0.70	0.18/0.30	0.29/0.42	0.17/0.32	**0.13**/0.29	0.22/0.39	0.17/0.27	0.17/**0.26**
zara2_s	0.26/0.55	0.14/0.24	0.21/0.32	0.12/0.25	**0.10/0.21**	0.17/0.30	0.13/0.22	0.13/**0.21**
AVG_s	0.37/0.73	0.23/0.39	0.35/0.50	0.21/0.41	**0.18**/0.35	0.29/0.48	**0.18**/0.27	**0.18/0.26**
eth_l	0.76/1.33	0.63/1.05	0.71/1.23	0.88/1.66	0.65/1.01	1.21/2.05	0.50/0.70	**0.46/0.52**
hotel_l	0.27/0.52	0.20/0.34	0.40/0.63	0.29/0.52	0.19/0.31	0.51/0.59	0.21/0.36	**0.15/0.22**
univ_l	0.84/1.76	0.42/0.80	0.52/0.89	0.47/0.89	**0.35**/0.66	1.14/1.78	0.37/0.62	**0.35/0.59**
zara1_l	0.55/1.10	0.32/0.59	0.39/0.60	0.30/0.48	0.33/0.65	0.61/1.12	0.29/0.50	**0.28/0.42**
zara2_l	0.38/0.78	0.24/0.44	0.27/0.45	0.23/0.40	0.22/0.44	0.60/1.04	0.21/0.35	**0.20/0.31**
AVG_l	0.56/1.10	0.36/0.64	0.46/0.76	0.43/0.79	0.35/0.61	0.81/1.32	0.32/0.51	**0.29/0.41**

Table 1: $mADE$/$mFDE$ on the five scenes of ETH and UCY.

	LED	SICNet	MSRL	UEN
eth_s	0.39/0.58	**0.27**/0.45	0.28/0.47	0.28/**0.31**
hotel_s	0.11/0.17	0.11/0.16	0.14/0.22	**0.10/0.13**
univ_s	0.26/0.43	0.26/0.46	**0.24**/0.43	**0.24/0.40**
zara1_s	0.18/0.26	0.19/0.33	0.17/0.30	**0.17/0.26**
zara2_s	0.13/0.22	0.13/0.26	0.14/0.23	**0.13/0.21**
AVG_s	0.21/0.33	0.19/0.33	0.19/0.33	**0.18/0.26**
SDD_s	8.5/11.7	8.4/13.7	8.2/13.4	**7.3/10.4**

Table 2: Compared with latest models on $mADE$/$mFDE$.

Following the prior work (Mangalam et al. 2021), we use two error metrics to evaluate the performance. Minimal Average Displacement Error ($mADE$) calculates the minimal average Euclidean distance between multimodal trajectories and the true one. Minimal Final Displacement Error ($mFDE$) computes the minimal Euclidean distance between multimodal positions and the true one at the goal time step. A smaller $mADE$ or $mFDE$ means a better result.

All used convolution kernels are set to 3×3 and $Relu$ is employed as the activation function. In Fig. 5, Env Encoder and Env Decoder contain 5 convolution layers. z is set to 32. r and σ' are set to 0.5 and 4 for the short-term prediction, and 2 and 16 for the long-term prediction. The training epoch is set to 500 with a batch size of 4, and Adam optimizer with a learning rate of $1e-4$ is used. The proposed model runs in the PyTorch framework with an NVIDIA RTX 3090.

Each compared model generates 20 multimodal predicted trajectories for a past trajectory. In Tables 1-4, the subscripts s and l represent the short- and long-term prediction, respectively, where the best results are shown by **bold** fonts. In Figs. 7, 8, 11, the red, green, and blue circles indicate the past, future, and predicted positions, where the predicted trajectory is one of the multimodal predicted trajectories, which is closest to the future trajectory. While in Figs. 9-10, the blue curves indicate the predicted multimodal trajectories.

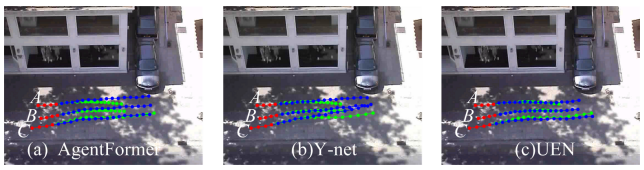

Figure 7: Influences of social interactions on future motions.

Experimental Results on ETH and UCY

The scene complexity of the scenes eth and hotel in ETH is low, social interactions between pedestrians don't obviously impact future motions due to a small number of pedestrians in the scenes, and the accuracy of prediction is primarily influenced by multimodal motion prediction. According to Table 1, irrespective of the short- or long-term prediction, BiTraP, Y-net, and UEN always obtain the top three ranks, for eth and hotel. Compared with the SOTA model Y-net on ETH in the short-term prediction, our model is slightly better than Y-net in terms of $mFDE$, due to the avoidance of generating similar goals. While in the long-term prediction, our method is totally superior to Y-net on ETH, because multimodal motion behaviors in the long-term prediction are obvious and the advantage of modeling diverse motions at non-goal time steps can be presented.

In three scenes univ, zara1, and zara2 of UCY, BiTraP, Y-net, and UEN show high prediction performance. For the short-term prediction on UCY, UEN is not as good as BiTraP and has the same results as those of Y-net, in terms of $mADE$, because multimodal motion behaviors are not obvious in the short-term prediction. Based on $mADE$ and $mFDE$ on univ, zara1, and zara2, we conclude that our method is slightly worse than BiTraP but is comprehensively better than Y-net in the short-term prediction. In the long-term prediction, UEN is obviously better than all its rivals, because pedestrians in the long-term prediction present high multimodal motion behaviors, and UEN can effectively characterize diverse movement patterns at all time steps.

Also, we compare UEN with the three latest models LED,

	TPNSTA	AgentFormer	VDRGCN	Trajectron++	BiTraP	PECNet	Y-net	UEN
SDD_s	13.7/20.2	8.5/12.0	11.1/18.3	8.8/15.2	10.4/18.7	10.0/15.9	7.9/11.9	**7.3/10.4**
SDD_l	76.7/125.8	52.5/71.2	72.9/119.7	53.6/77.3	56.6/80.2	72.2/118.1	47.9/66.7	**42.8/54.6**
inD_s	16.6/23.0	13.1/17.5	15.3/21.3	13.0/20.8	15.4/28.3	14.0/20.5	11.7/15.2	**10.7/13.6**
inD_l	22.1/35.8	20.2/28.6	21.8/32.3	24.0/37.7	23.4/40.9	20.3/33.0	15.0/21.1	**13.9/15.2**

Table 3: $mADE/mFDE$ on SDD and inD.

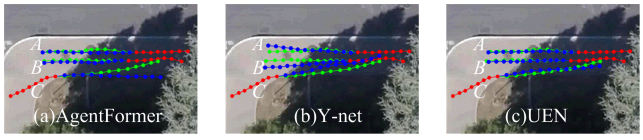

Figure 8: Impacts of social interaction and scene constraints.

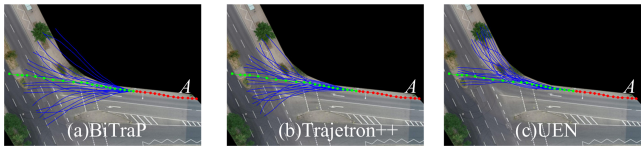

Figure 9: Multimodal predictions based on scene elements.

SICNet, and MSRL, on the short-term prediction. The three models use different baselines to generate multimodal prediction trajectories. LED enhances a diffusion model to produce varied pedestrian behaviors. SICNet incorporates a CVAE module with a sequential model to generate multimodal movements. MSRL utilizes a memory-based framework to pick up potential motion modes. The results are shown in Table 2, where the results of the three models are cited from their papers. We can find that UEN gets the lowest predicted errors in all the scenes. Considering the average $mADE$, our model obtains the relative performance improvements of 14%, 5%, and 5%, when compared with the latest models. For the average $mFDE$, UEN gets 21% relative performance improvement. Our model can achieve reasonable and diverse positions by the local multimodal windows at all time steps, and the experimental results demonstrate that our method is superior to the latest models for depicting diverse motions of pedestrians.

Social interactions between pedestrians also impact future motion behaviors. We provide a visual case in Fig. 7 to show how social interactions influence pedestrians' motions, where the pedestrians A and B are a movement group and C is overtaking the group. It is easy to find that the group's predicted trajectories achieved by our method are closer to the true paths. Considering the impacts of C on others, Y-net fails to avoid a collision between the pedestrians B and C due to ignoring social interactions. AgentFormer seems to incorrectly serve the three pedestrians as a group, and the predicted trajectories show a highly collaborative behavior. Differently, our proposed model can keep the group motion and a secure distance between the group and C, because the direction and distance relationship can simultaneously be characterized by the proposed polar-based method.

Experimental Results on SDD

Different from ETH and UCY, scenes in SSD include a number of scenario elements, and pedestrians' future motions are simultaneously impacted by scene constraints and social interactions. Compared with PECNet, BiTraP, VDRGCN, and TPNSTA, which ignore scene information, UEN has obviously lower prediction errors on the short- and long-term prediction. UEN still shows a better predicted performance when compared with the algorithms with scene features (Y-net, AgentFormer, and Trajectron++), because of the unified modeling of dynamic and static impacts of environments. Y-net ignores the modeling of social interactions, making it difficult to reflect surrounding pedestrians' impacts on a target pedestrian. Compared with Trajectron++ and AgentFormer, whose trajectory and scene features are presented in different modalities, our method can simultaneously characterize social interactions and scene influences in feature maps to comprehensively environmental impacts on a target pedestrian. Also, Table 2 shows the predicted results of the three latest models and UEN on the short-term prediction of SDD, where the results are cited from their papers. Note that none of these three models extract scene features to characterize scene impacts on a target pedestrian. Compared with the three models, UEN obtains the relative performance improvements of 14%/11%, 13%/24%, and 11%/22%, in terms of $mADE/mFDE$. In complex scenes of SDD, our method can effectively model environmental impacts on a target pedestrian by reflecting the distance and direction relationship between the target pedestrian and any position in an environment. The experimental results demonstrate that our method is superior to the SOTA and latest models.

We provide a visual case in Fig. 8 to show how social interactions and scene constraints impact the predicted accuracy, where the pedestrians A and B are a motion group and C is walking in the negative motion direction of the group. In Fig. 8(b), there is a collision risk for B and C, and the group behavior seems to have broken off due to the loss of social interactions in Y-net. In Fig. 8(a), A and B have a good group behavior, but C gradually deviates from the sidewalk. AgentFormer characterizes social interactions based on sequence features and scene constraints cannot be effectively presented, because of different modalities of the trajectory and scene features. Differently, UEN simultaneously models scene constraints and social interactions in feature maps and gets reasonable motions to meet scene constraints.

	MM	EM	UEN
eth_l	0.48/0.55	0.50/0.58	**0.46/0.52**
hotel_l	0.17/0.24	0.17/0.25	**0.15/0.22**
univ_l	0.37/0.60	0.37/0.61	**0.35/0.59**
zara1_l	0.29/0.46	0.29/0.47	**0.28/0.42**
zara2_l	0.21/0.32	0.22/0.34	**0.20/0.31**
inD_l	14.9/19.2	14.3/16.0	**13.9/15.2**
SDD_l	47.1/61.5	43.2/54.8	**42.8/54.6**

Table 4: Results of ablation study on $mADE/mFDE$.

Experimental Results on inD

All scenes in inD are captured in intersections, and the scene information significantly influences motions. The models Trajectron++, AgentFormer, Y-net, and UEN, which can extract scene features, always get the top four ranks in the short-term prediction. AgentFormer and Trajectron++ characterize past motions based on sequential models and fail to effectively present scene constraints, due to different modalities of the trajectory and scene features. Compared with the SOTA model Y-net, our algorithm obtains 9%/11% and 7%/28% relative performance improvement, in terms of $mADEmFDE$ on the short- and long-term predictions. Our method can effectively map image features into scene constraints by the polar-based method, and static environment impacts can be reflected in motions of pedestrians.

We provide a case in Fig. 9 to present multimodal trajectory predictions based on scene constraints, where the pedestrian A is crossing the road. Although BiTraP can generate diverse trajectories, some trajectories are out of the scene or in unreasonable regions because of the absence of scene information. Trajectorn++ predicts many future positions on the road due to the ineffective expression of scene constraints on the predicted trajectories. In Fig. 9(c), our method can generate reasonable trajectories according to the scene constraints because our method can model motion behaviors and scene constraints in the same feature modality.

Ablation Experiment

To better understand the contributions of Multimodal Modeling (MM) and Environmental Modeling (EM), we evaluate MM, EM, and UEN on the long-term predictions in Table 4, where MM is UEN without Environmental Modeling and its input has only a target pedestrian's heatmaps, and EM is UEN without Multimodal Modeling and gets multiple goals to predict trajectories as executed in (Mangalam et al. 2021).

For the results on ETH and UCY in Table 4, EM consistently performs the worst. The prediction of multimodal motions significantly impacts the performance in the five scenes with low complexity, because future motions of pedestrians are not easily influenced by the scenes. While on inD and SDD, the scenes are high complexity and social interactions are diverse, and EM is superior to MM on these two datasets.

Fig. 10 gives a case to show the superiority of Multimodal Modeling, where the pedestrian A is walking straight and then turns right. Note that the grass is walkable. It is difficult

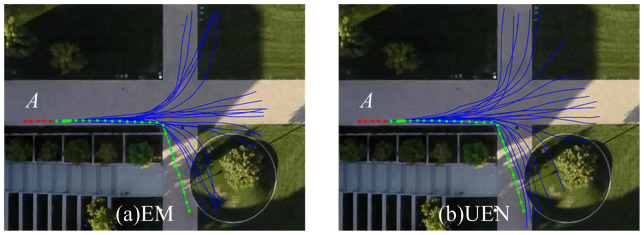

Figure 10: A visual comparison between EM and UEN.

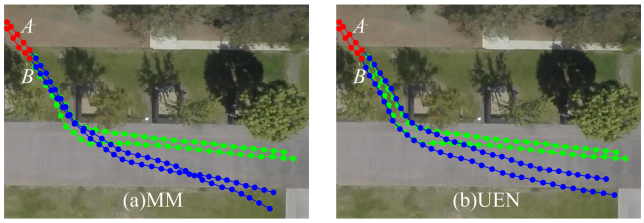

Figure 11: A visual comparison between MM and UEN.

to forecast the future trajectory by a few past positions. Thus, more diverse predicted trajectories have a higher probability of approximating the future trajectory. Compared with EM, UEN can avoid the generation of crowded predicted goals. Moreover, the predicted positions at non-goal time steps can be independently produced to ensure the predicted diversity.

We provide a visual case in Fig. 11 to show the importance of Environmental Modeling on reflecting the impacts of social interactions, where the pedestrians A and B are a motion group and are passing through the dirt road and lawn. It is easy to find that the predicted trajectories of MM are chaotic and there is a collision between A and B. In Fig. 11(b), a secure distance between A and B and the group motion can be maintained by the proposed polar-based method.

Conclusion

Motions of pedestrians significantly suffer from environmental impacts. However, sequential models fail to effectively integrate scene features to make predicted trajectories comply with scene constraints due to different modalities of scene and trajectory features. Although existing CNN models can extract scene features, they are ineffective in mapping scene features into scene constraints and struggle to model social interactions due to the loss of target pedestrian information. This work proposes a unified environmental network to address these problems. We propose a polar-based method to reflect the distance and direction relationship between a target pedestrian and any position in an environment, and scene constraints and social interactions can be simultaneously modeled in feature maps to comprehensively present environmental impacts. For achieving reasonable and diverse future trajectories, we capture essential local features in feature maps to characterize potential multimodal motions at each time step. Compared with ten trajectory prediction algorithms on four datasets, the experimental results have shown the superiority of our model.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62001304 and Grant 52102400; in part by the Key project of Shenzhen Science and Technology Plan under Grant 20220810180617001. 3.

References

An, J.; Liu, W.; Liu, Q.; Guo, L.; Ren, P.; and Li, T. 2022. DGInet: Dynamic graph and interaction-aware convolutional network for vehicle trajectory prediction. *Neural Networks*, 151: 336–348.

Bock, J.; Krajewski, R.; Moers, T.; Runde, S.; Vater, L.; and Eckstein, L. 2020. The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections. In *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, 1929–1934.

Chen, B.; Tan, W.; Wang, Y.; and Zhao, G. 2022. Distinguishing Between Natural and GAN-Generated Face Images by Combining Global and Local Features. *Chinese Journal of Electronics*, 31(1): 59–67.

Chen, G.; Li, J.; Zhou, N.; Ren, L.; and Lu, J. 2021. Personalized Trajectory Prediction via Distribution Discrimination. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 15580–15589.

Dong, Y.; Wang, L.; Zhou, S.; and Hua, G. 2023. Sparse Instance Conditioned Multimodal Trajectory Prediction. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 9763–9772.

Duan, J.; Wang, L.; Long, C.; Zhou, S.; Zheng, F.; Shi, L.; and Hua, G. 2022. Complementary Attention Gated Network for Pedestrian Trajectory Prediction. In *Proceedings of AAAI Conference on Artificial Intelligence*, 542–550.

Gu, J.; Sun, C.; and Zhao, H. 2021. DenseTNT: End-to-End Trajectory Prediction From Dense Goal Sets. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 15303–15312.

Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. *Computer graphics forum*, 26(3): 655–664.

Li, Y.; Liang, R.; Wei, W.; Wang, W.; Zhou, J.; and Li, X. 2022. Temporal Pyramid Network With Spatial-Temporal Attention for Pedestrian Trajectory Prediction. *IEEE Transactions on Network Science and Engineering*, 9(3): 1006–1019.

Li, Y.; Xie, C.; Liang, R.; Du, J.; Zhou, J.; and Li, X. 2023. A Synchronous Bi-Directional Framework With Temporally Dependent Interaction Modeling for Pedestrian Trajectory Prediction. *IEEE Transactions on Network Science and Engineering*, 1–14.

Liang, R.; Li, Y.; Li, X.; Tang, Y.; Zhou, J.; and Zou, W. 2021. Temporal Pyramid Network for Pedestrian Trajectory Prediction with Multi-Supervision. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2029–2037.

Liang, R.; Li, Y.; Zhou, J.; and Li, X. 2023. STGlow: A Flow-Based Generative Framework With Dual-Graphormer for Pedestrian Trajectory Prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Lv, P.; Wang, W.; Wang, Y.; Zhang, Y.; Xu, M.; and Xu, C. 2023. SSAGCN: Social Soft Attention Graph Convolution Network for Pedestrian Trajectory Prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Mangalam, K.; An, Y.; Girase, H.; and Malik, J. 2021. From Goals, Waypoints & Paths to Long Term Human Trajectory Forecasting. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 15233–15242.

Mangalam, K.; Girase, H.; Agarwal, S.; Lee, K.-H.; Adeli, E.; Malik, J.; and Gaidon, A. 2020. It Is Not the Journey But the Destination: Endpoint Conditioned Trajectory Prediction. In *Proceedings of European Conference on Computer Vision*, 759–776.

Mao, W.; Xu, C.; Zhu, Q.; Chen, S.; and Wang, Y. 2023. Leapfrog Diffusion Model for Stochastic Trajectory Prediction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5517–5526.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 261–268.

Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes. In *Proceedings of European Conference on Computer Vision*, 549–565.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and Savarese, S. 2019. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1349–1358.

Salzmann, T.; Ivanovic, B.; Chakravarty, P.; and Pavone, M. 2020. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In *Proceedings of European Conference on Computer Vision*, 683–700.

Su, Y.; Du, J.; Li, Y.; Li, X.; Liang, R.; Hua, Z.; and Zhou, J. 2022. Trajectory Forecasting Based on Prior-Aware Directed Graph Convolutional Neural Network. *IEEE Transactions on Intelligent Transportation Systems*, 23(9): 16773–16785.

Tang, H.; Wei, P.; Li, J.; and Zheng, N. 2022. EvoSTGAT: Evolving spatiotemporal graph attention networks for pedestrian trajectory prediction. *Neurocomputing*, 491: 333–342.

Wang, R.; Song, X.; Hu, Z.; and Cui, Y. 2023. Spatio-Temporal Interaction Aware and Trajectory Distribution Aware Graph Convolution Network for Pedestrian Multimodal Trajectory Prediction. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–11.

Wu, Y.; Wang, L.; Zhou, S.; Duan, J.; Hua, G.; and Tang, W. 2023. Multi-Stream Representation Learning for Pedestrian Trajectory Prediction. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2875–2882.

Xie, C.; Li, Y.; Liang, R.; Dong, L.; and Li, X. 2022. Synchronous Bi-Directional Pedestrian Trajectory Prediction with Error Compensation. In *Proceedings of Asian Conference on Computer Vision*, 2796–2812.

Xu, C.; Mao, W.; Zhang, W.; and Chen, S. 2022. Remember Intentions: Retrospective-Memory-based Trajectory Prediction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6478–6487.

Yao, Y.; Atkins, E.; Johnson-Roberson, M.; Vasudevan, R.; and Du, X. 2021. BiTraP: Bi-Directional Pedestrian Trajectory Prediction With Multi-Modal Goal Estimation. *IEEE Robotics and Automation Letters*, 6(2): 1463–1470.

Yuan, Y.; Weng, X.; Ou, Y.; and Kitani, K. M. 2021. AgentFormer: Agent-Aware Transformers for Socio-Temporal Multi-Agent Forecasting. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, 9813–9823.

Zhang, Z.; Wang, B.; Yu, Z.; and Li, Z. 2021. Dilated Convolutional Pixels Affinity Network for Weakly Supervised Semantic Segmentation. *Chinese Journal of Electronics*, 30(6): 1120–1130.

Zhou, L.; Zhao, Y.; Yang, D.; Liu, J.; and Beijbom, O. 2022. GCHGAT: pedestrian trajectory prediction using group constrained hierarchical graph attention networks. *Applied Intelligence*, 52: 11434–11447.