SocialCVAE: Predicting Pedestrian Trajectory via Interaction Conditioned Latents

Supplemental materials for Submission 6512

S1 Energy-based Interaction Anticipating

Time to collision. We provide a detailed explanation of how to solve the quadratic function $\tau_{ij}^{t+1} = \arg\min_{\tau} \|\mathbf{x}_{ij}^{t} + \widetilde{\mathbf{v}}_{ij}^{t+1} \cdot \tau\|_2$ to obtain τ_{ij}^{t+1}, which denotes the time taken for pedestrian i to collide with a neighbor j (time to collision, i.e., when the predicted distance is 0). We predict collisions by assuming a linear extrapolation of the positions of the pedestrian and the neighbor based on the pedestrian's coarse preferred velocity and the neighbor's current velocity. Then determine whether there is a time τ^* at which the two future linear trajectories intersect, i.e.,

$$\|\mathbf{x}_{ij}^{t} + \widetilde{\mathbf{v}}_{ij}^{t+1} \cdot \tau^*\|_2 = 0, \tag{1}$$

by rearranging, we obtain the solution

$$\tau^* = -\frac{\|\mathbf{x}_{ij}^{t}\|^2}{\mathbf{x}_{ij}^{t} \cdot \widetilde{\mathbf{v}}_{ij}^{t+1}}. \tag{2}$$

If the solution τ^* doesn't exist (i.e., $\mathbf{x}_{ij}^{t} \cdot \widetilde{\mathbf{v}}_{ij}^{t+1} = 0$) or is negative, $\tau_{ij} = \infty$, which means the pedestrian and the neighbor will not collide in the future. Otherwise, $\tau_{ij} = \tau^*$.

S2 Evaluation

S2.1 Details of Datasets

The ETH-UCY dataset [3, 1] includes pedestrians' trajectories in 5 scenes (ETH, HOTEL UNIV, ZARA1, and ZARA2), including more than 1500 pedestrians and thousands of nonlinear trajectories with various interactions. The trajectories are recorded in the world coordinates (i.e., meters). We consider the pedestrian-pedestrian, pedestrian-static obstacle (e.g., buildings) interactions for ETH-UCY. We follow the leave-one-out strategy [2, 5] for training and evaluation, i.e., training our model on four sub-datasets and testing it on the remaining one. Following the common practice [2, 5], the raw trajectories are segmented into 8-second trajectory segments with time step $\Delta t = 0.4$ seconds, we train the model to predict the future 4.8 seconds (12 frames) based on the observed 3.2 seconds (8 frames). Since our model works in the image coordinate space (i.e., pixel space), we project the world coordinates in ETH-UCY into the pixel using the homography matrices provided in Y-net [2] for training and testing, and then project the pixel results to meters for calculating quantitative metrics.

The SDD dataset [4] contains videos of a university campus with six classes of traffic agents with rich interactions, including about 185,000 interactions between different agents and approximately 40,000 interactions between the agent and the environment. The trajectories are recorded in pixels. More complex interactions are considered in SDD, including pedestrian-pedestrian, pedestrian-static obstacle (e.g., buildings), and pedestrian-dynamic obstacle (e.g., vehicles, bicycles) interactions. We extract the trajectories like that of ETH-UCY to train the model for predicting the future $4.8s$ (12 frames) based on the observed $3.2s$ (8 frames).

S2.2 Qualitative Comparisons

Predicted trajectories. As shown in Fig. S2, we provide more visualization results of the predicted trajectories and compare them with that of the state-of-the-art method NSP-SFM [5]. Results show that our method predicts future trajectories closer to the ground truth than NSP-SFM. The predicted results of NSP-SFM (i.e., results exhibited in the first row) show strong determinism in reaching a sampled final goal driven by a goal-attraction model, resulting in the predicted trajectories deviating from the ground-truth final positions. In contrast, our method employs an interaction-conditioned CVAE model for learning socially reasonable human motion uncertainty, thus achieving better prediction performance.

Multimodal prediction. As shown in Fig. S2, we provide more visualization results of the multiple predicted trajectories and compare them with that of the state-of-the-art method NSP-SFM [5]. The better results of our method exhibited in the second row demonstrate that, by conditioning on the socially explainable interaction energy map, our method learns better human motion uncertainty than NSP-SFM which doesn't consider social interactions in human motion randomness learning.

S2.3 Experiment Setup

Hyperparamters. The hyperparameters utilized in our experiments are shown in Table S1. The same set of hyperparameters was employed for both the ETH-UCY and SDD datasets.

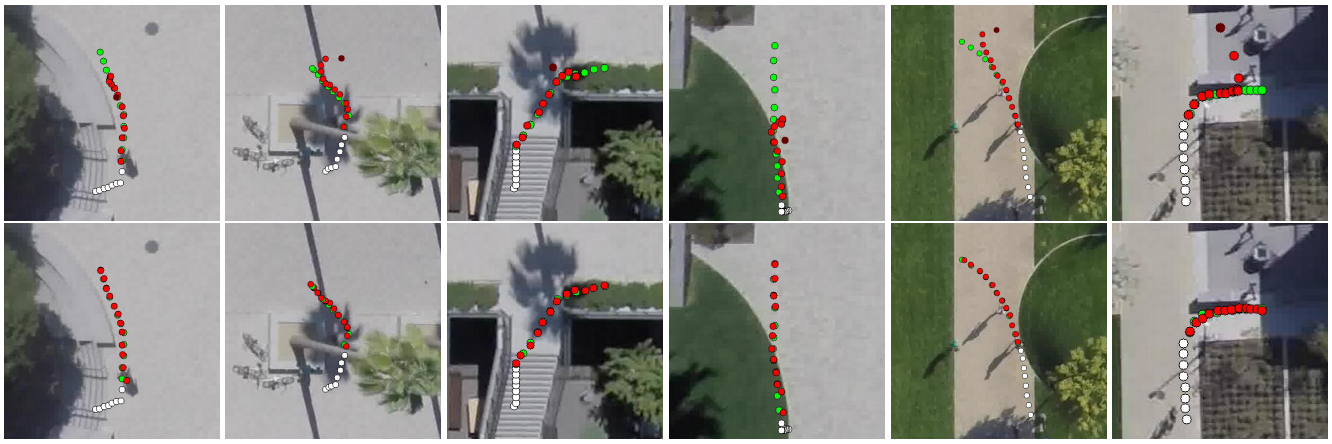

Figure S1: Supplementary materials for visualization comparisons with NSP-SFM. The results of NSP-SFM and our method are shown in the first and second rows, respectively. The visualized trajectories are the best predictions sampled from 20 trials. The white, green, and red dots represent the observed, ground-truth, and predicted trajectories respectively. The purple dots in the visualization results of NSP-SFM (i.e., the results in the first row) represent the sampled goals.

Figure S2: Supplementary materials for visualization comparisons of multiple predicted trajectories with NSP-SFM. The results of NSP-SFM and our method are shown in the first and second rows, respectively. The white and green lines are the observed and ground-truth trajectories. The yellow lines for each pedestrian are the 20 predicted trajectories.

Table S1: The hyperparameters utilized for implementing SocialCVAE.

$\Delta\theta$	k_θ	Δr (pixel)	k_r	d_s (pixel)
$\frac{1}{6}\pi$	4	5	2	5
L (pixel)	R_{ped} (pixel)	λ_1	λ_2	λ_3
100	5	1	0.5	0.5

Learnable parameters. We also show the learned weights of different interaction energies for ETH-UCY and SDD in Table S2.3. It is worth noting that all learnable weights were initialized as 1 prior to training the model. For simplicity, the weight of the interaction energy with pedestrian neighbors was not trained and remained at the value 1 in all experiments.

Table S2: The learned weights of different interaction energies. w_s and w_d respectively represent the learned weight of interaction energy with the static and dynamic obstacles.

Learnable Parameter	w_s	w_d
ETH	1.0229	-
HOTEL	1.0081	-
UNIV	1.2104	-
ZARA1	1.0497	-
ZARA2	1.1429	-
SDD	1.0644	1.1887

Sub-network architecture. We also provide the detailed network architectures of the sub-networks employed in our experiments in Table S2.3. The ReLU activation function is used in our network for the non-linearity of the network. The network configurations for both ETH-UCY and SDD

datasets were identical.

Table S3: The architecture of the sub-networks employed in our experiments.

Sub-network		Network Architecture
Section 3.2	LSTM	$[4, 256, 64]$
	MLP_1	$[32, 1024, 512, 64]$
	MLP_2	$[2, 1024, 512, 64]$
	MLP_3	$[64, 1024, 512, 2]$
	Linear (query encoding)	$[64, 64]$
	Linear (key encoding)	$[64, 64]$
Section 3.4	E_{mot}	$[16, 512, 256, 16]$
	E_{map}	$[10000, 1024, 512, 256, 32]$
	E_{res}	$[2, 8, 16, 16]$
	E_{latent}	$[64, 8, 50, 32]$
	D_{latent}	$[64, 1024, 512, 1024, 2]$

References

[1] Lerner, A.; Chrysanthou, Y.; and Lischinski, D. 2007. Crowds by example. In *Computer Graphics Forum*, volume 26, 655–664.

[2] Mangalam, K.; An, Y.; Girase, H.; and Malik, J. 2021. From goals, waypoints & paths to long term human trajectory forecasting. In *International Conference on Computer Vision (ICCV)*, 15233–15242.

[3] Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *International Conference on Computer Vision (ICCV)*, 261–268.

[4] Robicquet, A.; Sadeghian, A.; Alahi, A.; and Savarese, S. 2016. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision (ECCV)*, 549–565.

[5] Yue, J.; Manocha, D.; and Wang, H. 2022. Human trajectory prediction via neural social physics. In *European Conference on Computer Vision (ECCV)*, 376–394.