



SocialVAE: Human Trajectory Prediction using Timewise Latents – Supplemental material

Pei Xu^{1,2}, Jean-Bernard Hayet³, and Ioannis Karamouzas¹

¹ Clemson University, South Carolina, USA

² Roblox

³ CIMAT, A.C., México

peix@clemson.edu, jbhayet@cimat.mx, ioannis@clemson.edu

<https://motion-lab.github.io/SocialVAE>

1 Evaluation Metrics

Below are the computation details of the evaluation metrics:

- *Average Displacement Error* (ADE), the Euclidean distance between a prediction trajectory $\{\mathbf{x}_i^t\}$ and the GT value $\{\hat{\mathbf{x}}_i^t\}$ averaged over all prediction frames for $t = T + 1, \dots, T + H$:

$$\text{ADE}(\{\mathbf{x}_i^t\}, \{\hat{\mathbf{x}}_i^t\}) = \frac{1}{H} \sum_{t=T+1}^{T+H} \|\mathbf{x}_i^t - \hat{\mathbf{x}}_i^t\|. \quad (1)$$

- *Final Displacement Error* (FDE), the Euclidean distance between the predicted position in the final frame and the corresponding GT value:

$$\text{FDE}(\{\mathbf{x}_i^t\}, \{\hat{\mathbf{x}}_i^t\}) = \|\mathbf{x}_i^{T+H} - \hat{\mathbf{x}}_i^{T+H}\|. \quad (2)$$

- *Negative Log Likelihood* (NLL), the negative logarithm of the value of the predictive PDF at GT trajectories. The predictive distribution is obtained by Gaussian kernel density estimation from 2,000 samples. For simplicity, distributions at each time step are estimated independently and we use the joint distributions to compute PDF values.

2 Social Features

SocialVAE employs three social features for attention computation as shown in Fig. S1. Given an agent i at time step t and its neighbor j , these features are:

- the Euclidean distance between agents i and j , i.e. $\|\mathbf{p}_{ji}^t\|$ where $\mathbf{p}_{ji}^t = \mathbf{x}_j^t - \mathbf{x}_i^t$;
- the cosine value of the bearing angle from agent i to neighbor j , i.e. $\cos(\mathbf{p}_{ji}^t, \mathbf{d}_i^t)$;

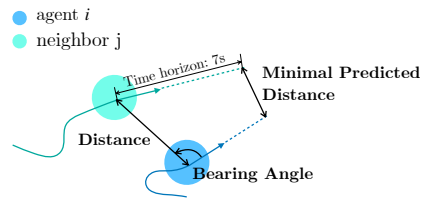


Fig. S1. Demonstration of social features used for attention computation.

- the minimal predicted distance [2] from agent i to j within a time horizon h (7s by default), i.e., $\|\mathbf{p}_{ji}^t + \min(\tau, h)\mathbf{v}_{ji}^t\|$, where $\mathbf{v}_{ji}^t = (\mathbf{d}_j^t - \mathbf{d}_i^t)/\Delta t$, $\tau = -(\mathbf{p}_{ji}^t \cdot \mathbf{v}_{ji}^t)/\|\mathbf{v}_{ji}^t\|^2$, and Δt is the sampling interval between two frames.

3 Data Acquisition of SportVU NBA Dataset

To test our approach on scenarios with complex and intensive human-human interactions, we have extracted two sub-datasets from the SportVU basketball movement dataset [3,1] focusing on games from the 2015-2016 NBA regular season:

- **Rebounding dataset.** This dataset focuses on scenes involving a missed shot with players moving to grab the rebound. The dataset contains a number of interesting interactions, including players boxing out their opponents to allow a team member to grab a rebound, players moving toward the basket, and players starting to run on the other side of the half court for offensive or defensive purposes.
- **Scoring dataset.** This dataset focuses on scenes involving a team scoring a basket. The resulting dataset contains a rich set of player-player interactions, both cooperative and adversarial, including highly non-linear player motions, set plays employed by different teams, and different offensive and defensive schemes.

Table S1. Statistical information on SportVU NBA Datasets.

	Scoring	Rebounding
# of Training/Testing Scenes	2,979/744	3,754/938
Avg. Play Duration (s)	11.82	2.94
# of Trajectories (20-frame)	2,958,480	257,230
Avg. Trajectory Length (m)	4.55	3.87

We refer to Table S1 for detailed characteristics of the two datasets. For each dataset, scenes are randomly split into testing and training sets using a 1:4 ratio. The original data were recorded at 25 FPS with a time interval of 0.04s between frames. In consideration that basketball players move much faster than normal pedestrians, we downsample the data to the time interval of 0.12s (instead of 0.04s that we use on ETH/UCY and SDD benchmarks). We employ the same network structure that we have used for the ETH/UCY and SDD benchmarks, and do 12-frame predictions for players (excluding the ball) based on 8-frame observations. This leads to training and testing trajectories having 20 frames, with the average length around 4m, as reported in Table S1. The neighborhood radius is set such that the whole arena is covered, which means that all the players and the ball are taken into account during observation encoding.

4 Additional Results on SDD

Table S2. ADE/FDE in meters on SDD. The reported numbers are the mean value of the best-of-20 predictions.

	Trajectron++	BiTraP	SGNet-ED		SocialVAE	SocialVAE+FPC
SDD	0.34/0.58	0.32/0.57	0.33/0.58		0.30/0.50	0.27/0.39

5 Sensitivity Analysis on FPC

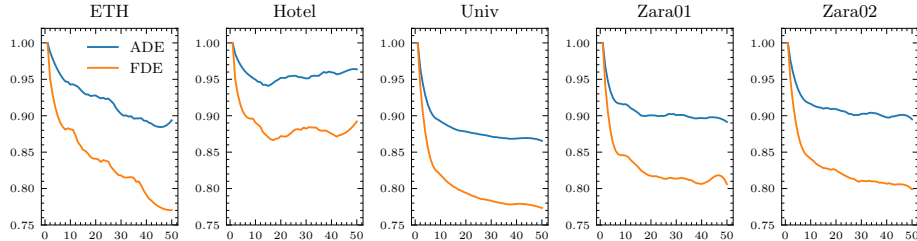


Fig. S2. Performance of FPC with respect to different sampling rates (1-50). All values are normalized by that of sampling rate 1 (no FPC).

Figure S2 plots the ADE and FDE values when FPC is applied with varying sampling rate on the ETH/UCY benchmark. As shown in the figure, the errors decrease roughly as the sampling rate increases. Typically, FPC can lead to a significant improvement about 10% on ADE and 18% on FDE within a sampling rate around 20. Further increasing the sampling rate can only bring about a 2% extra improvement (with the exception of a 5% FDE improvement on ETH), at the cost though of higher running time.

6 Latent Space Analysis

To show that our model can learn a structured embedding of the observed trajectories, we plot the latent variable distributions in Fig. S3. To do so, we run a model pre-trained using the ETH/UCY datasets on 15 different 8-frame observations, which are the combinations of five distinct trajectory headings and three distinct speeds. For each observation, we draw 150 samples of the latent variables from the prior at the first time step of prediction, i.e. \mathbf{z}_i^{T+1} . As it can be seen, our model can clearly distinguish observations with semantically different features.

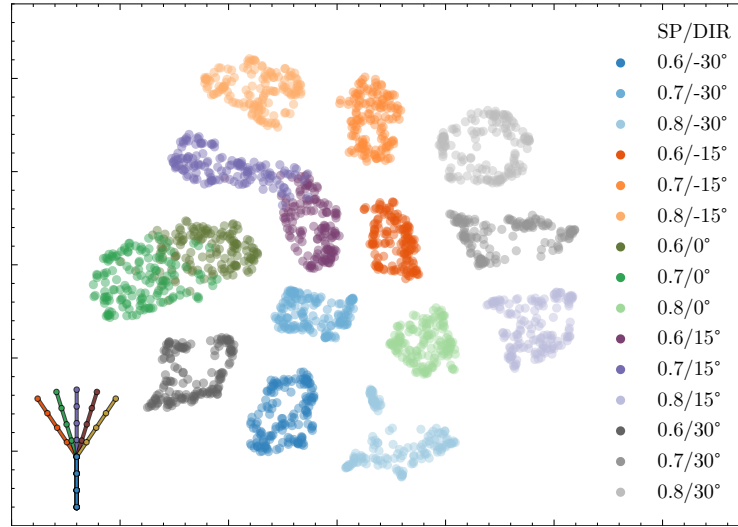


Fig. S3. t-SNE visualization of latent variable distributions given varying observation trajectories with different speeds (SP) and turn directions (DIR). The left bottom corner gives an example of the observed trajectories with different turn directions at the 5th frame from -30° to 30° . For each of the five trajectory shapes, we consider observations with three constant speeds from 0.6m to 0.8m. This gives us a combination of 15 observations, as shown in the legend.

References

1. Makansi, O., Kügelgen, J.V., Locatello, F., Gehler, P.V., Janzing, D., Brox, T., Schölkopf, B.: You mostly walk alone: Analyzing feature attribution in trajectory prediction. In: International Conference on Learning Representations (2022)
2. Olivier, A.H., Marin, A., Crétual, A., Pettré, J.: Minimal predicted distance: A common metric for collision avoidance during pairwise interactions between walkers. *Gait & Posture* **36**(3), 399–404 (2012)
3. Yue, Y., Lucey, P., Carr, P., Bialkowski, A., Matthews, I.: Learning fine-grained spatial models for dynamic sports play prediction. In: IEEE International Conference on Data Mining. pp. 670–679 (2014)