Supplementary Materials

A Sensitivity test on the length of consecutive masks.

The consecutive mask length is an important hyper-parameter for controlling the amount of sequence information. Figure 1 shows the performance of using different masking percentages in masked cross-sequence to sequence pretext. The consecutive masking length of input ranges from 2 (25% of masking percentage) to 5 frames (62.5% of masking percentage) within a 8-frame observable target sequence. Social-SSL achieves the best performance on downstream trajectory prediction task when using 4 consecutive frames to mask the target sequence, which is 50% of the observable target sequence. This study shows a balanced masking length that leverage the target sequence within encoder and decoder input achieve better results.

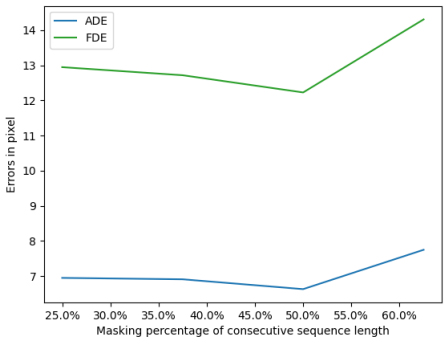


Fig. 1. Different masking percentage of consecutive sequence on SDD dataset.

B Sensitivity test on distance thresholds in closeness pretext.

In order to determine a universal distance threshold in closeness pretext for all the scenes to be used for distinguishing the "social" or "non-social" agents via closeness prediction, we first re-scale all the locations of each scene into $[0, 1]^2$ to ensure that the prediction value falls within this range. Table 1 compares the results of using different distance thresholds, which suggests that setting the distance threshold to 0.15 leads to a better result. Therefore, the threshold is set to 0.15 for all the experiments.

Distance Threshold	ETH	HOTEL	UNIV	ZARA1	ZARA2	SDD
0.1	0.71/1.42	0.25/0.46	0.56/1.07	0.45/0.87	0.36/0.68	7.02/12.87
0.15	0.69/1.37	**0.24/0.44**	**0.51/0.93**	**0.42/0.84**	**0.34/0.67**	**6.63/12.23**
0.2	**0.68/1.36**	0.27/0.50	0.52/1.01	0.45/0.86	0.37/0.69	7.25/13.54

Table 1. Results of Social-SSL with different distance thresholds on ETH/UCY and SDD datasets.

C Classification results in social-related pretext tasks.

Table 2 shows the percentage of occurrences for different interaction types and closeness types on ETH/UCY and SDD datasets. The F1 scores for interaction type prediction are 89% and 87% on ETH/UCY and SDD datasets, respectively. Moreover, the F1 scores for closeness prediction are 99% on both ETH/UCY and SDD datasets. The results indicate that the model indeed learns from the proposed self-supervised pretexts.

	Interaction Types			Closeness Types	
	Leaving	Neutral	Closing	Social	Non-social
ETH/UCY	34.9%	8.6%	56.3%	58.1%	41.9%
SDD	39.8%	14.1%	45.9%	35.0%	65.0%

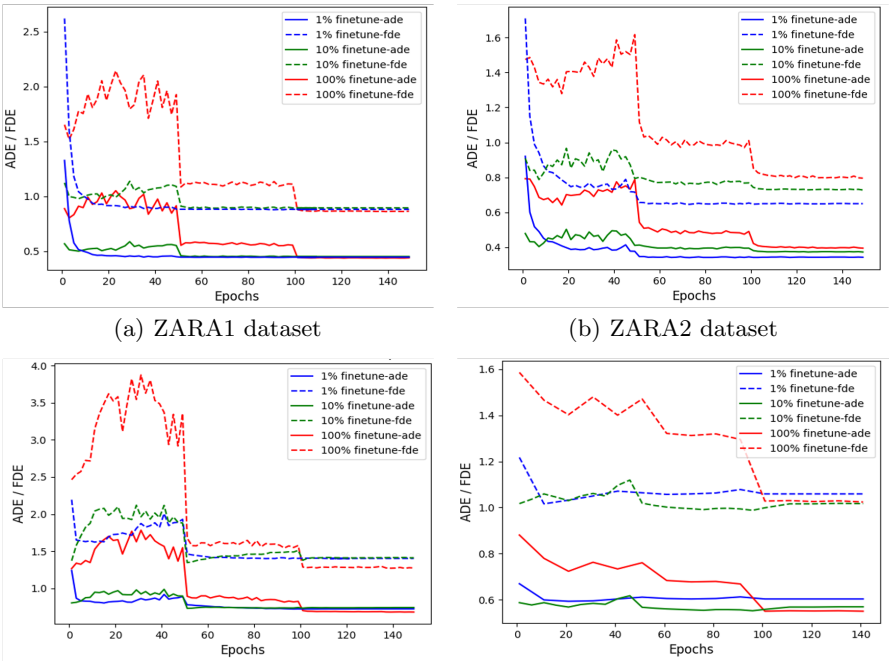
Table 2. Percentage of occurrences for different interaction types and closeness types.

D Different percentages of data for finetuning.

Figure 2 demonstrates the performance of the proposed method with different percentages of data used for finetuning on 4 datasets under the leave-one-out setting. Specifically, we randomly select 1% and 10% of data from the leave-one-out dataset (100%) as the alternative target to fine-tune, and compare the results under the same learning rate setting. The results indicate that using 100% dataset to fine-tune requires a greater number of epochs for reaching good results. In contrast, randomly choosing 1% of data provides easier cues for the model with prior knowledge to follow. It is worth noting that the time cost of each epoch increases linearly with the proportion of the data used for fine-tuning. Therefore, using 1% of data for fine-tuning results in both less epochs and less training time for each epoch. This is because the proposed pretext tasks have managed to capture the context closely related to the trajectory prediction task.

E More visualization cases.

Speed controlling. Figure 3 visualizes the results of Social-SSL with Social-STGCNN and STGAT. In the top left region of Figure 3, the agent should gradually slow down based on its past sequence. However, Social-STGCNN and STGAT both predict that the agent continues to move. This could be due to Social-STGCNN and STGAT were not able to learn specific social interaction features between agents, such as interaction type, to guide their trajectories. As such, the prediction may depend on all the social agents in the scene. In contrast, Social-SSL predicts a steady path without being affected by other social agents due to the cross-sequence to sequence pretext, leveraging the information of intra- and inter-relation to achieve results that better match the ground truth.



(a) ZARA1 dataset (b) ZARA2 dataset

(c) ETH dataset (d) UNIV dataset

Fig. 2. Evaluation results using different amounts of data for fine-tuning. We set the initial learning rate as 3e-6 and the decay rate by 0.1 for every 50 epochs to better observe the convergence phenomenon.

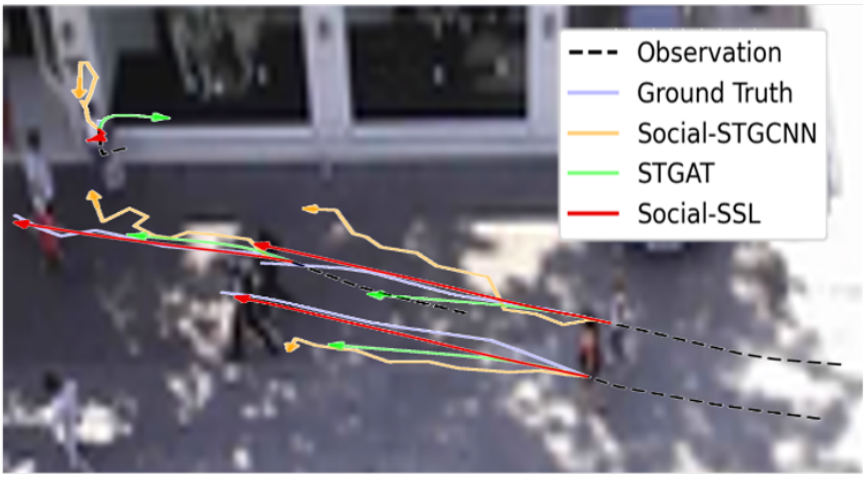


Fig. 3. An illustrative example on ZARA1 dataset.

Subsequence learning. Figure 4 shows that agent A and agent B go for route 2 and route 1 separately. This may be a challenging case for the model since the locations of past trajectories of the two agents are nearly the same. If we take a closer look at both their past trajectories, we can observe a slight change in direction for agent B in the middle of the past trajectory. Interestingly, our masked cross-sequence to sequence pretext, which enforces the model to learn the subsequence for each target agent, seems to capture this slight change in direction. Despite the historical trajectories of both agents being so similar, the trajectory prediction results of agent A and agent B still performs accurately.

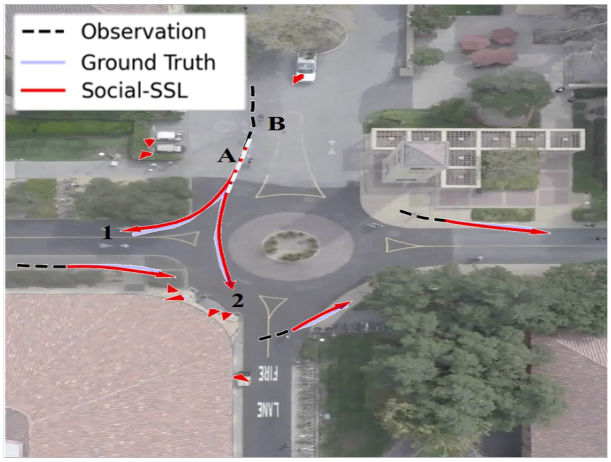


Fig. 4. An illustrative example of deterministic result on SDD dataset. The observation of agent A is colored in white to better distinguish the difference to agent B.