Are socially-aware trajectory prediction models really socially-aware?

Saeed Saadatnejad[1,*] Mohammadhossein Bahari[1,*]
Seyed-Mohsen Moosavi-Dezfooli[2] Alexandre Alahi[1]
[1]EPFL [2] ETH Zurich
{saeed.saadatnejad, mohammadhossein.bahari}@epfl.ch

Abstract

Our field has recently witnessed an arms race of neural network-based human trajectory predictors. Yet, adversarial robustness of none of these methods has been carefully studied in order to gain a better understanding of their limitations. In this paper, by introducing socially-attended attack, we assess the social understanding of the prediction models in terms of collision avoidance. Technically, we define collision as a failure mode of the model, and propose a soft-attention mechanism to guide our attack. We demonstrate the strengths of our method on the state-of-the-art trajectory prediction models. Finally, we show that our attack can be employed to increase the social understanding of state-of-the-art models. To access the code and the complete paper (with more details and experiments), visit here: https://s-attack.github.io/

1. Introduction

Understanding the social behavior of humans is a core problem for many autonomous applications, such as social robots [3] or self-driving cars [5, 2]. For a robot to navigate among crowds safely or for an autonomous vehicle to drive in urban areas harmlessly, human behavior anticipation is essential. In particular, dealing with humans makes the problem safety-critical. For instance, a self-driving car's wrong prediction in a crosswalk can put a pedestrian's life in danger. Being a safety-critical problem raises the need for careful assessments of the trajectory prediction methods to mitigate the risks associated with humans. Consequently, the robustness properties of those methods, as one of the important assessment aspects, should be studied.

The pedestrian trajectory prediction problem is to predict future positions of pedestrians given their past positions as inputs. Recently, the problem has received solutions using neural networks. Various models based on Long-Short-Term-Memory networks [1], convolutional neural networks [16], and Generative Adversarial Networks (GAN) [8] are proposed. The core challenge of the problem lies in learning the interactions between people. Therefore, explicit models based on neural networks are designed to tackle the interaction more accurately [1, 22, 9, 15]. Humans' interactions involve different social behaviors such as collision avoidance, walking within a group, and merging from different directions into a specific point. Among all behaviors, collision avoidance, *i.e.,* people choosing a path that avoids collision with others, is one of the key behaviors rarely violated. That is why many previous works consider respecting collision avoidance as the evidence of their model being social [15, 9, 10]. Thus, we consider collision avoidance as an indicator of social behavior of the models.

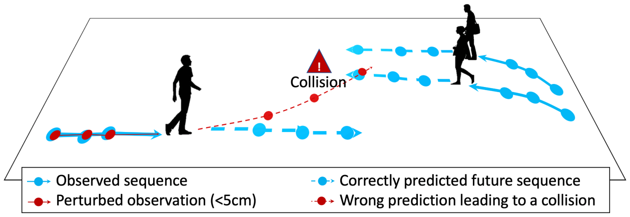

Collision
Observed sequence
Perturbed observation (<5cm)
Correctly predicted future sequence
Wrong prediction leading to a collision

Figure 1. Given the observation trajectories of the agents in the scene, a predictor (here S-LSTM [1]) forecasts the future positions reasonably (blue lines). However, with less than 5 cm perturbation in the observation trajectory (in red), an unacceptable collision is predicted.

We show a conceptually plausible real-world scenario in Figure 1. Given the observed trajectories of humans in the scene, a social predictor forecasts the future positions reasonably without collision. However, by adding a small perturbation of less than 5 cm to the observation trajectory, unexpectedly, a collision between predictions of agents occurs which indicates a non-complete social understanding by the predictors. The trajectories in that figure comes from S-LSTM [1].

In this work, in contrast to the common adversarial attacks which are designed for classifiers [17, 20], we design attack for the trajectory prediction problem which is a mul-

[*] Equal contribution.

timodal regression task. We use adversarial examples to study the collision avoidance behavior of the trajectory prediction models. More specifically, we investigate the worst-case social behavior of a prediction model under small perturbations of the inputs. This study has two primary motivations; (1) it is an evaluation method for the previously-proposed predictors. Our method brings counter-examples in which the models fail in having social behavior, *i.e.,* it cannot avoid collision. (2) leveraging adversarial examples, one can train models with better collision-avoidance. Furthermore, our study highlights practical concerns for employing such models in real-world applications. Notably, it is shown that state-of-the-art localization algorithms give on-average more than 0.2 m errors on human location detection at each frame [7, 4]. While our work focuses on model failures under adversarial settings, it motivates further studies of the model's performance when localization algorithms' error distribution is concerned.

We propose an adversarial attack to fool the trajectory prediction models by causing collision between two agents' predicted trajectories. Hence, the attacker tries to find small perturbations that lead to a collision. The collision can hypothetically happen between any two agents and at any prediction timestep. However, from the attacker's perspective, the choice of the agent and the timestep impacts the final perturbation's size significantly. To address that, we introduce an attention-guided adversarial attack, named *Socially-ATTended ATTack (S-ATTack)*, which learns the best collision points. Our experiments demonstrate that our novel attack can find perturbations that make state-of-the-art trajectory prediction models generate wrong predictions, leading to collisions with small perturbations. Lastly, we introduce an adversarial training scheme to make trajectory prediction models more robust. In particular, we show how our method can improve the models' social understanding in terms of collision avoidance. To the best of our knowledge, this is the first work addressing the adversarial vulnerability and robustness of trajectory prediction models. Our main contributions are summarized as follows:

- We introduce S-ATTack to assess the social understanding of the state-of-the-art trajectory prediction models.

- We demonstrate how to improve the robustness properties of the predictors using our S-ATTack.

2. Method

In this section, we first explain the notations and definitions. Then, we will provide the details of S-ATTack.

2.1. Formulation

2.1.1 Pedestrian trajectory prediction

Pedestrian trajectory prediction addresses a regression task with sequences as inputs and outputs. At any timestep t, the i-th person/agent is represented by his/her xy-coordinates (x_t^i, y_t^i). We denote each agents' observation sequence for T_{obs} timesteps as X^i, a $T_{obs} \times 2$ matrix. Given the observations of all agents in the scene, the trajectory predictor f predicts the next T_{pred} positions of all agents $Y = (Y^1, \ldots, Y^{n+1}) = f(X^1, \ldots, X^{n+1})$. For brevity, we denote the observation sequences of the ego-agent and other agents in the scene as $X = (X^1, \ldots, X^{n+1})$ and without loss of generality, the ego-agent's as X^1. Number of non-ego agents in the scene is denoted as n.

2.1.2 Adversarial examples for trajectory prediction

Equipped with the notations introduced in section 2.1.1, we will provide a definition of adversarial examples for trajectory prediction. In this paper, without loss of generality, we focus on the collisions between the ego-agent and other agents but it can directly be expanded to collisions between any two agents. In addition, we assume the perturbation r is added only to the ego-agent $\hat{X}^1 = X^1 + r$ while the observations of other agents $\{X^j\}_{j \neq 1}$ remain fixed. Therefore, r is a $T_{obs} \times 2$ matrix of adversarial perturbation, the adversarial example is $\hat{X} = (X^1 + r, X^2, \ldots, X^{n+1})$ and the output of the predictor for that example is $\hat{Y} = (\hat{Y}^1, \ldots, \hat{Y}^{n+1}) = f(\hat{X})$. Formally, given a small constant $\epsilon > 0$, a collision distance threshold γ and the maximum of the norm of all rows of a matrix $\|\cdot\|_{\max}$, a socially-attended adversarial example is obtained if:

$$\exists\, r, j \neq 1, t : \|r\|_{\max} \leq \epsilon \Rightarrow: \left\|\hat{Y}_t^j - \hat{Y}_t^1\right\| < \gamma, \qquad (1)$$

In other words, this type of adversarial examples is based on perturbing an observation trajectory so that f predicts the future timesteps with at least a collision occurring between two agents j and 1 in one timestep t. In the next section, we will describe how we obtain r using Socially-attended attack.

2.2. Socially-attended attack (S-ATTack)

Given the perturbation r, and a model f, we define the distance matrix $D(r) \in \mathbb{R}^{n \times T_{pred}}$ as a function of the input perturbation r. It includes the pairwise distance of all non-ego agents from the ego-agent in all prediction timesteps. Let $d_{j,t}$ denote the element at j-th row and t-th column of $D(r)$, i.e., the distance of the agent j from the ego-agent at timestep t of the prediction timesteps. Hence, for a particular r, the distance matrix $D(r)$ can be leveraged to indicate whether a collision has occurred. We now explain a method to find such a perturbation by optimizing a cost function depending on $D(r)$.

Note that sometimes a collision with a farther agent-timestep may require smaller perturbation due to the gradient of the network. To address that, we let the attack attend to the optimal target by itself. We introduce a soft-attention mechanism in which the weights associated to each agent-timestep is assigned by the attack in order to achieve a smaller perturbation. The equation of the soft-attention attack is as follows:

$$\begin{aligned} &\min_{r,W} \mathrm{Tr}\left(W^{\top} \tanh(D(r))\right) + \lambda_r \left\|r\right\|_F - \lambda_w \left\|W\right\|_F , \\ &\text{s.t.} \sum_{j,t} w_{j,t} = 1, \quad w_{j,t} \geq 0, \end{aligned} \tag{2}$$

where $\tanh$ is applied to the entries of $D(r)$ in order to concentrate less on very far agent-timesteps. Besides, W is the attention weight matrix and $w_{j,t}$ is the attention weight for the agent j at timestep t of the prediction timesteps. The size of W is the same as $D(r)$. Also, we discourage uniformity of weights by subtracting the Frobenius norm of W multiplied by a scalar λ_w and we add the regularization on the perturbation with the balancing coefficient λ_r that encourages finding a small perturbation sequence to make a collision.

We use the gradient descent algorithm to optimize for a given input sequence X. The attack stops if a collision occurs. Furthermore, in each iteration, r is projected onto the ℓ_∞ ball with a small radius ϵ around the observation X, similar to the Projected Gradient Descent (PGD) [12]. W is initialized with a uniform distribution. It is progressively updated and puts more weights to the more probable targets for making a collision. Near the convergence point, the best target agent receives a weight value close to 1 while the rest receive 0.

3. Experiments

3.1. Baselines

In order to show the effectiveness of our attack, we conduct our experiments on six well-established trajectory prediction models.

Social-LSTM [1] (**S-LSTM**): where a social pooling method is employed to model interactions based on shared hidden states of LSTM trajectory encoders.

Social-Attention [22] (**S-Att**): where a self-attention block is in charge of learning interactions between agents.

Social-GAN [8] (**S-GAN**): where a max-pooling function is employed to encode neighbourhood information. They leverage a generative adversarial network (GAN) to learn the distribution of trajectories.

Directional-Pooling [10] (**D-Pool**): where the features of each trajectory is learned using the relative positions as well as the relative velocity and then pool the learned features to learn social interactions.

Model	Original	Attacked	
	CR [%] ↓	CR [%] ↓	P-avg [m] ↓
S-LSTM [1]	7.8	89.8	0.031
S-Att [22]	9.4	86.4	0.057
S-GAN [8]	13.9	85.0	0.034
D-Pool [10]	7.3	88.0	0.042
S-STGCNN [15]	16.3	59.1	0.11
PECNet [14]	15.0	64.9	0.071

Table 1. Comparing the performance of different baselines before (Original) and after the attack (Attacked). Horizontal lines separate models with different datasets.

Social-STGCNN [15] (**S-STGCNN**): where graph convolutional neural network is employed to learn the interactions.

PECNet [14] (**PECNet**): where a self-attention based social pooling layer is leveraged with a variational auto-encoder (VAE) network.

3.1.1 Datasets

ETH [18], UCY [11], and WildTrack [6]: these are well-established datasets with human positions in world-coordinates. **SDD [19]**: The Stanford Drone Dataset is a human trajectory prediction dataset in bird's eye view. PECNet is one of the state-of-the-art methods with official published code on this dataset. Hence, we report PECNet performance on this dataset.

3.1.2 Metrics

In the experiments, we report the performances according to the following metrics:

Collision Rate (CR): this metric calculates the percentage of samples in which at least one collision in the predicted trajectories between the ego-agent and its neighbors occurs. Note that we set the distance threshold for indicating a collision γ in eq. (1) equal to 0.2 m .

Perturbation average (P-avg): the average of perturbation sizes at each timestep which is added to the input observation in meters.

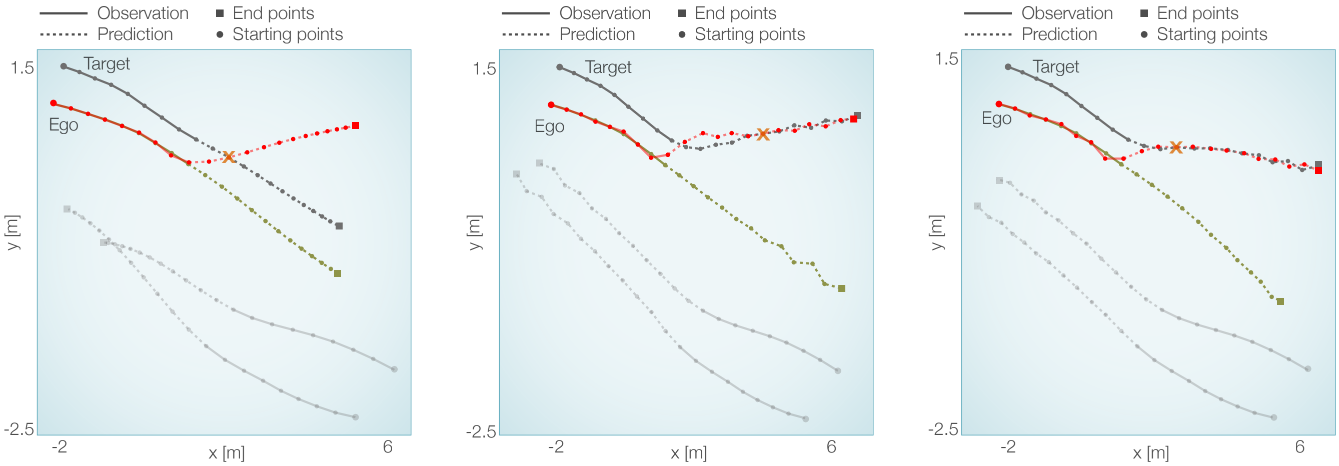
Average / Final Displacement Error (ADE/FDE): the average/final displacement error between the predictions of the model and the ground-truth values.

3.2. Attack results

We first provide the quantitative results of applying S-ATTack to the baselines in Table 1. The results indicate a substantial increase in the collision rate (at least 3 times)

	ADE/FDE [m] ↓	Original		Attacked	
		CR [%] ↓	CR gain [%] ↑	CR [%] ↓	CR gain [%] ↑
D-Pool	0.57 / 1.23	7.3	-	37.3	-
D-Pool w/ rand noise	0.57 / 1.23	7.5	-2.7	36.1	+3.2
D-Pool w/ S-ATTack	0.60 / 1.28	**6.5**	**+10.4**	**14.7**	**+60**

Table 2. Comparing the original model and the fine-tuned model with random-noise data augmentation (D-Pool w/ rand noise) and S-ATTack adversarial examples (D-Pool w/ S-ATTack). ADE, FDE are reported in meters.



(a) S-LSTM (P-avg : 0.009m)

(b) S-Att, (P-avg : 0.028m)

(c) D-Pool, (P-avg : 0.029m)

Figure 2. Comparison of the performance of different models under our attack. The ego-agent is depicted with green before the perturbation and red after it. For brevity, we have not shown the prediction of non-ego agents before the attack. The scale of y axis is enlarged to better show the difference. The orange X denotes the target point. Our attack achieves collisions with adding small perturbations.

across all baselines by adding perturbations with P-avg smaller than 0.11 m. This questions the social behavior of the models in terms of collision avoidance.

Figure 2 visualizes the performance of the three baselines S-LSTM, S-Att, and D-Pool under our attack with the same input. We can observe that almost all models counteract to avoid collisions. While these attempts show that there exists some understanding of the collision avoidance behavior in the prediction models, they are not enough for avoiding a collision.

As D-pool performs better than others in terms of collision avoidance before attack, in the rest of the paper, we conduct our main experiments on it.

3.3. Enhancing the social understanding

We utilize our S-ATTack to improve the collision avoidance of the model. To this end, we employ a similar approach to [13]. We fine-tune the model using a combination of the original training data and the adversarial examples generated by our S-ATTack method. In this experiment, we set the maximum perturbation size ϵ equal to 0.03.

Table 2 indicates that the model's collision avoidance could improve by 11%. Moreover, the collision rate after attack improves by 60% meaning that it is much less vulnerable to the attack. As shown in the table, fine-tuning the model with random noise could not improve the collision avoidance. Therefore, we conclude that our adversarial examples provide useful information to improve the collision avoidance of the model. Note that the prediction error of the model in terms of ADE/FDE is slightly increased. This means that there exists a trade-off between accuracy and robustness. This can be similar to the findings in the previous works on image classifiers [21].

4. Conclusion

In this work, we studied the robustness properties of trajectory prediction models in terms of social understanding under adversarial attack. We introduce our Socially-ATTended ATTack (S-ATTack) to cause collisions in state-of-the-art prediction models with small perturbations. Adversarial training using S-ATTack can not only make the models more robust against adversarial attacks, but also reduce the collision rate and hence, improve their social understanding. This paper reveals common weaknesses of trajectory prediction models opening a window towards their social understandings. As future work, we will use our findings to improve the current neural network-based models.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 1, 3

[2] Mohammadhossein Bahari, Ismail Nejjar, and Alexandre Alahi. Injecting knowledge in data-driven vehicle trajectory predictors. *arXiv preprint arXiv:2103.04854*, 2021. 1

[3] Maren Bennewitz, Wolfram Burgard, and Sebastian Thrun. Learning motion patterns of persons for mobile service robots. In *Proceedings of the IEEE International Conference on Robotics and Automation*, volume 4, pages 3601–3606. IEEE, 2002. 1

[4] Lorenzo Bertoni, Sven Kreiss, Taylor Mordan, and Alexandre Alahi. Monstereo: When monocular and stereo meet at the tail of 3d human localization. In *International Conference on Robotics and Automation (ICRA)*, 2021. 2

[5] Smail Ait Bouhsain, Saeed Saadatnejad, and Alexandre Alahi. Pedestrian intention prediction: A multi-task perspective. *arXiv preprint arXiv:2010.10270*, 2020. 1

[6] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur M. Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. *Prooceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5039, 2018. 3

[7] Wenlong Deng, Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Joint human pose estimation and stereo 3d localization. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2324–2330. IEEE, 2020. 2

[8] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018. 1, 3

[9] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 137–146, 2019. 1

[10] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1, 3

[11] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. *Comput. Graph. Forum*, 26:655–664, 2007. 3

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 4

[14] 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 (ECCV)*, pages 759–776. Springer, 2020. 3

[15] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14424–14432, 2020. 1, 3

[16] Nishant Nikhil and Brendan Tran Morris. Convolutional neural network for trajectory prediction. In *European Conference on Computer Vision (ECCV) Workshops*, 2018. 1

[17] Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi, and Pascal Frossard. Hold me tight! influence of discriminative features on deep network boundaries. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 2935–2946. Curran Associates, Inc., 2020. 1

[18] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *European Conference on Computer Vision (ECCV)*, pages 452–465. Springer, 2010. 3

[19] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European Conference on Computer Vision (ECCV)*, pages 549–565. Springer, 2016. 3

[20] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[21] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv: 1805.12152*, 2018. 4

[22] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *2018 IEEE international Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2018. 1, 3