**Proceedings of 10th IOE Graduate Conference**
*Peer Reviewed*
**ISSN: 2350-8914 (Online), 2350-8906 (Print)**
**Year: 2021 Month: October Volume: 10**

# Pedestrian Movement Prediction Using Encoder-Decoder Model

Bikram Acharya [a], Diwakar Raj Pant [b]

[a, b] *Institute of Engineering, Pulchowk Campus*
**Corresponding Email**: [a] 075mscsk004.bikram@pcampus.edu.np, [b] drpant@ioe.edu.np

**Abstract**

Pedestrian trajectory prediction in crowded space with multi-agent are extensively researched with possibility of being automated using learned models. The key task is to accurately encode observation sequence, model long-term dependencies from the past trajectories and forecast potential trajectories and reducing the task complexity to a manageable subset from which we can learn social impact from other pedestrians, scene limits, and multi-modal possibilities of expected routes and generalize to challenging scenarios and even output unacceptable solutions. This paper presents effective use of hard negatives samples with contrastive learning to preserve motion representation, which captures desirable generalization properties, very little computational overhead and improved the quality of visual representations in socially aware pedestrian trajectory prediction. The data set used was ETH-UCY, comprising of total 5 different sets ETH, Hotel, Univ, Zara1 and Zara2 with ADE,FDE and Collision Avoidance Metric as metrics for performance. The result shows that proposed methodology with hard negative sampling has better collision avoidance with values 0.3, 0.56 and 0.08 in Hotel,Univ and Zara1 dataset respectively. However, state-of-art Social-NCE[1] shows better average FDE for all dataset i.e 0.381(Social-NCE)<0.47(Our).

**Keywords**

Pedestrian Trajectory, Contrastive Learning, Motion Representation, Spatio-temporal Encoding, Multi-agent System, Trajectron++

## 1. Introduction

With the development in deep learning techniques, RNN and LSTM networks have been commonly applied to time sequence data for a variety of issues, including speech recognition, language processing, and machine learning. Similarly, literature shows ample research on extraction of features from human trajectories [2, 3], simulation of human-human/space social relationships [4]. Exploring contrastive learning for data augmentation in this study with primary focus on hard negative sampling, we learn motion representation and train encoder-decoder architecture.

Different neural network models for learning socially-aware motion, representations has been extensively used and demonstrated their utility for human trajectory forecasting [2, 3] in crowded environments. Models built around covariance shift did not contain enough scenes with complex situations while models making use of interactive data collections, such as expert queries and interaction with the environment,vare inexpensive but are infeasible for forecasting problems. Approach suggested by [1] uses prior knowledge about socially unfavorable events and exploit learning in a robust neural motion model. However, such learning techniques uses both positive and negative samples, which significantly increase batch sizes and computational overhead.An effective approach may be to use hard negatives sampling with user controlled hardness which captures desirable generalization properties, very little computational overhead and improved the quality of visual representations, as seen on previous works on image dataset, presented by works in [5]. However, such comprehensive confrontation is still lacking in pedestrian trajectory prediction.

## 2. Related Literature

With time, pedestrian trajectory prediction has shifting from physics-based models to data-driven models based on deep learning. [6] pioneered one of the first approaches to pedestrian behavior modeling, known as the Social Forces Model using handcrafted features that reflect various powers that work on the pedestrian.This work highlighted three forces: acceleration toward the target velocity of motion, repulsive effect, and attractive effect. Agent-based modeling[7], including the social force model, has been used to predict human behavioral behaviors. [8] modelled human interaction behavior using strong priors in a discrete decision system.Further, Crowd simulations makes use of motion models and [9] used agent-based approaches for this purpose. The method models each person uniquely, and in order to produce practical simulations, a thorough understanding of various agents is needed.

Physics-based pedestrian behavior simulation has improved over time, with the advent of sophisticated strategies such as BRVO[10], which draws on Reciprocal Velocity Obstacle **RVO!** (**RVO!**)[11] and the Ideal Reciprocal Collision Avoidance(ORCA)[12]. These physics-based models, on the other hand, are constrained by the fact that they use hand-crafted functions, and therefore can only describe a subset of all possible behaviors.

As of 2016, forecasting potential trajectories use evidence, a data-driven approach that entails understanding how people walk by training a machine learning algorithm for real-world pedestrian trajectories. Data-driven methods can explicitly extrapolate the laws and nuances of human walking behavior that would be difficult to formalize from data. Learning how people walk solely from observable trajectories necessitates three key components: a machine learning algorithm with sufficient representation capacity, an efficient optimization strategy, and a sufficient amount of real-world evidence. deep learning models for pedestrian trajectory prediction in the literature depend primarily on the use of Recurrent Neural Networks (RNN), specifically Long Short-Term Memory (LSTM)[13] cells. The Social LSTM[2] model was one of the first to use such an approach, pioneering the use of deep learning in pedestrian trajectory prediction. To model social interaction, social knowledge is depicted as a grid of nearby pedestrians. For modelling social interaction, social knowledge is depicted as a grid of nearby pedestrians. Extraction of features from human trajectories [2, 3], simulation of human-human/space social relationships [4], and understanding the mutual activities of heterogeneous social actors [14] have been the subject of most of the recent existing trajectory prediction study.

## 3. Methodology

This research work, in addition to the encoder decoder model implementation, also integrates hard negative sampling and motion representation based on socially aware interaction which is well suited for trajectory predictions, and thus helps to avoid possible hazards. The block diagram in Figure 1 is the model pipeline for encoder decoder architecture, which given their previous motion states, predicts future trajectories of pedestrians in a scenes using encoded motion representations based on shared social information. Sequence encoder is used to encode time sequence trajectory position of primary and secondary agents and interaction encoder makes use of socially aware contrastive learning based motion encoding to capture interaction between pedestrians and establish social representation among them. Using decoder, future predictions can thus be made.

### 3.1 Dataset Description

This research work makes use of ETH[15] and UCY[16] dataset, which represents the real world coordinates i.e pedestrian are annotated by their position in meters with origin in an arbitrary point of world. ETH consists two scenes namely ETH and HOTEl, taken from bird's eyes view, where every frame contains annotations of pedestrian's position for every 0.4 sec and a total of 750 different pedestrian.

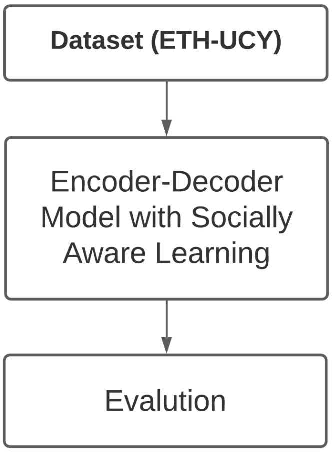

**Figure 1:** Methodology

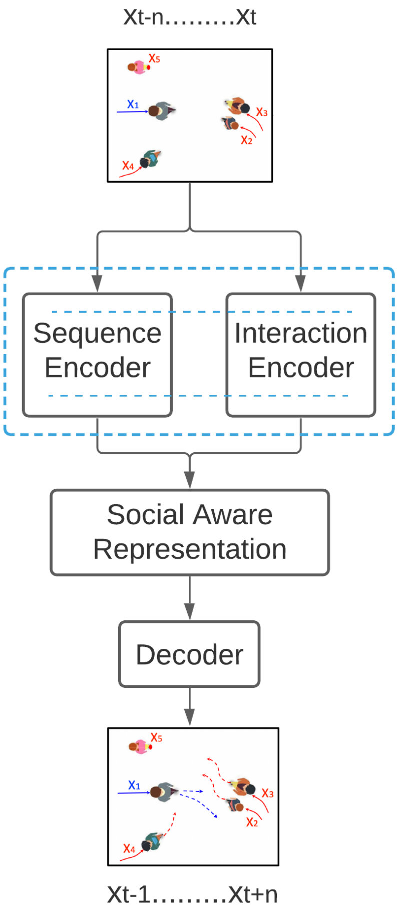

**Figure 2:** Model Pipeline

Similarly, UCY contains three scenes(Zara1,Zara2 and Univ) with 900 different pedestrians and their trajectories from bird's eyes view.

## 3.2 Encoder-Decoder Architecture

The Encoder-Decoder architecture has become a reliable and commonly used tool for neural machine translation (NMT) and sequence-to-sequence (seq2seq) prediction in general.

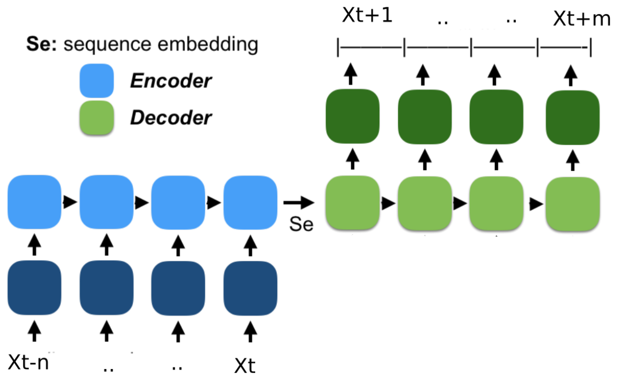

**Figure 3:** Encoder-Decoder Architecture

The Figure 3 show the encoder decoder architecture which consists of 3 parts: encoder, intermediate (sequence embedding) vector and decoder.

**Encoder** A stack of multiple recurrent units (LSTM or GRU cells for improved performance), each of which accepts a single element from the input list, gathers information for that element, and propagates it forward. The hidden states $h_i$ are computed using the formula:

$$\mathbf{h_t} = \mathbf{f}(\mathbf{W}^{(\mathbf{hh})}\mathbf{h_{t-1}} + \mathbf{W}^{(\mathbf{hx})}\mathbf{x_t}) \tag{1}$$

This is the model's final hidden state as created by the encoder.This vector attempts to encapsulate all input element's information in order to assist the decoder in making correct predictions. It serves as the model's original hidden state and input to the decoder.

**Sequence Encoding**
Since the model is trained and tested using trajectron++, the input trajectory sequence is as shown in Figure 1, dataset block. Sequence Encoder encodes temporal information.

**Interaction Encoder**
This work makes use of non grid based interaction model, which captures the social interactions in a grid-free manner, thus the spatial information is preserved.

**Decoder** A stack of periodic units, each of which predicts an output $y_t$ at a time phase t. Each recurrent unit accepts a hidden state from the previous unit and generates both an output and its own hidden state. The formula is used to compute every hidden state $h_i$:

$$\mathbf{h_t} = \mathbf{f}(\mathbf{W}^{(\mathbf{hh})}\mathbf{h_{t-1}}) \tag{2}$$

## 3.3 Contrastive Representation Learning

This research work makes use of contrastive learning, a technique to learn an embedding space using similarity measures and selection of hard negative samples to approximate viable neighbourhood relationship. Learning a parametric function that maps raw data into a feature space to extract abstract and usable knowledge for downstream tasks is characteristic of representation learning[17]. To train an encoder, recent contrastive learning methods often use the concept of noise contrastive estimation in an embedding space, namely the InfoNCE loss[18] given by equation below:

$$\mathscr{L}_{\mathbf{NCE}} = -\mathbf{log}\frac{\mathbf{exp}(\mathbf{sim}(\mathbf{q},\mathbf{k}^{+})/\tau)}{\sum_{\mathbf{n=0}}^{\mathbf{N}}\mathbf{exp}(\mathbf{sim}(\mathbf{q},\mathbf{k_n})/\tau)} \qquad (3)$$

where the encoded query $q$ is brought close to one positive key $k_0 = k^+$ and pushed apart from $N$ negative keys $k_1, \ldots, k_N$, $\tau$ is a temperature hyperparameter, and $sim(u,v) = u^T v/(||u||||v||)$ is the cosine similarity between two feature vectors.

### 3.4 Evaluation Metrics

The evaluation of models that suggest a single future mode for a given past observation is referred to as unimodal evaluation. In the unimodal context, the most widely used metrics for human trajectory forecasting are as follows:

1. **Average Displacement Error (ADE)**
   This metric, like the one used in [19], calculates overall predicted time steps average $L_2$ distance between ground truth and model prediction

2. **Final Displacement Error (FDE)**
   At the completion of the forecast cycle, the distance between the predicted final destination $T_{pred}$ and the ground truth destination.

3. **Collision Avoidance**
   This metric shows whether or not the expected model trajectories intersect, indicating whether or not the model knows the concept of collision avoidance.

## 4. Result and Analysis

Figure 4 shows pedestrian present in frame 395 of dataset.Different color and line codes are done to represent past, prediction and ground truth trajectories. This frame was used for training the model. The frame consists of group and other category pedestrians.

Similarly, Figure 5 is the figurative representation of frame 108 and contains multiple ground truth because those trajectories are sampled for multi modal trajectories. This frame consists of single pedestrian.

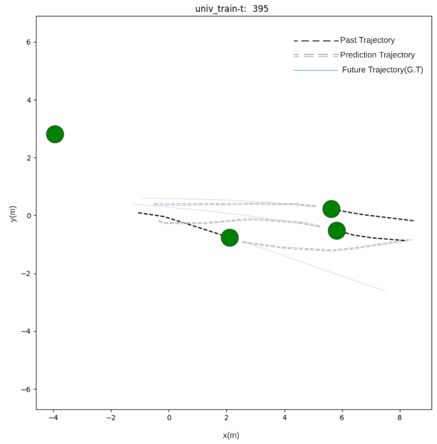

**Figure 4:** Univ: Trajectory for Train

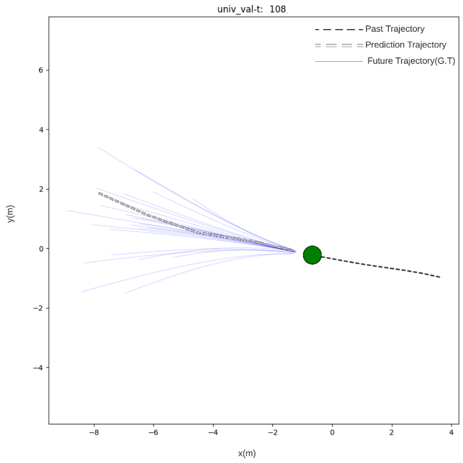

**Figure 5:** Univ: Trajectory for Test

Similar result was observed for multi pedestrian for both training and testing frames of Univ as presented in Graphics 6 and Figure 7.

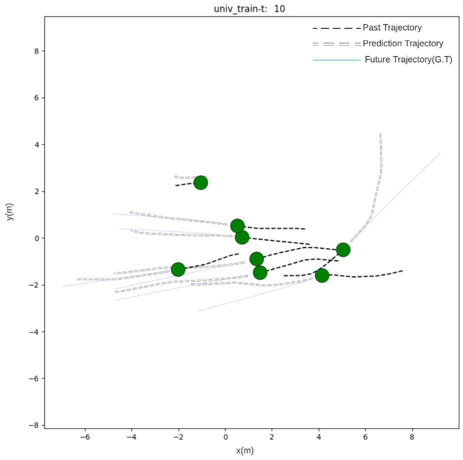

**Figure 6:** Univ: Multi Pedestrian Trajectory for Train

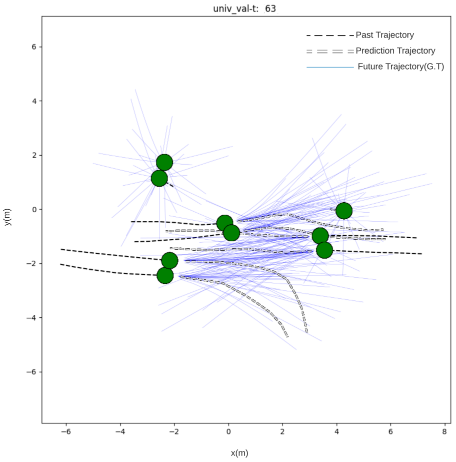

**Figure 7:** Univ: Multi Pedestrian Trajectory for Test

Also, similar observation can be seen for Zara dataset as shown in Figure 8, Figure 9, Figure 10 and Figure 11.

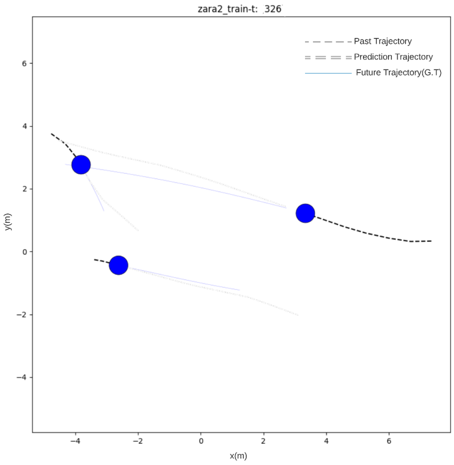

**Figure 8:** Zara1: Trajectory for Train Scene 326

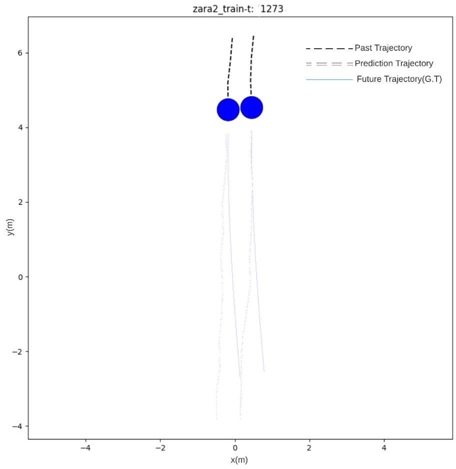

**Figure 9:** Zara1: Trajectory for Test Scene 1273

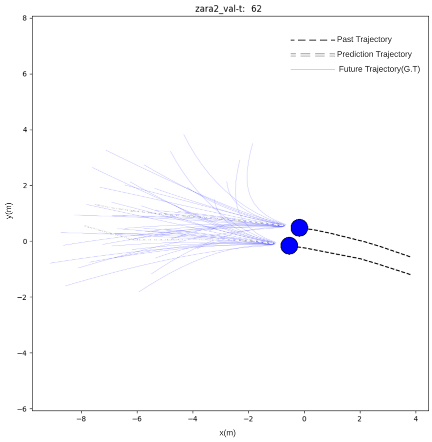

**Figure 11:** Zara1: Multi Pedestrian Trajectory for Test Scene 62

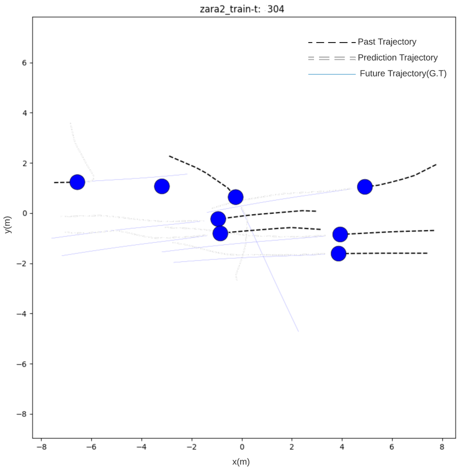

**Figure 10:** Zara1: Multi Pedestrian Trajectory for Train Scene 302

The evaluation of proposed architecture was done using ADE, FDE and Collison Avoidance metrics described above. The change in ADE,FDE and Collision Avoidance is shown using curve in Figure [12,13,14,15,16] .

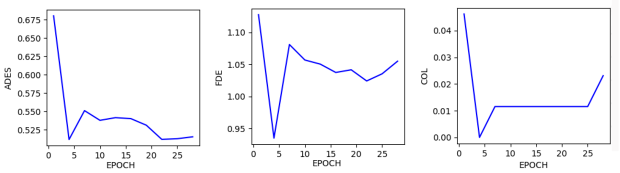

**Figure 12:** ADE, FDE and Collision for ETH Dataset

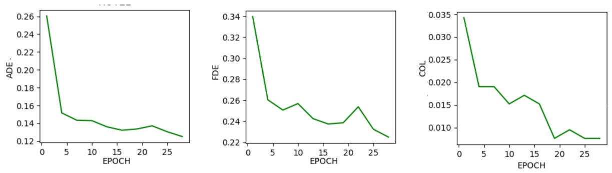

**Figure 13:** ADE, FDE and Collision for HOTEL Dataset

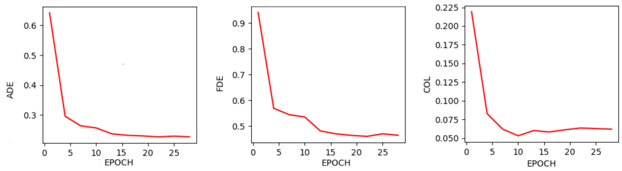

**Figure 14:** ADE, FDE and Collision for UNIV Dataset

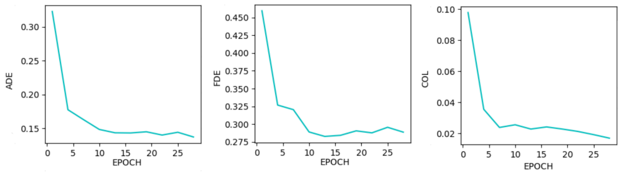

**Figure 15:** ADE, FDE and Collision for ZARA1 Dataset

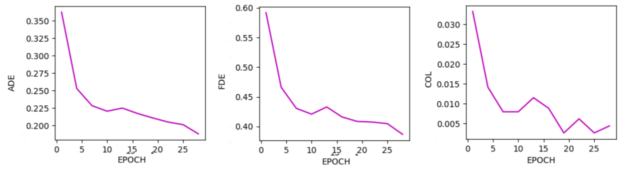

**Figure 16:** ADE, FDE and Collision for ZARA2 Dataset

The decreasing nature of all metrics curve suggests that the model is learning.

**Table 1:** Comparison With Social-NCE

| Dataset | Social-NCE [1] | | Our | |
|---|---|---|---|---|
| | FDE | COL | FDE | COL |
| ETH | **0.71** | **0.00** | 1.055 | 0.23 |
| Hotel | **0.177** | 0.38 | 0.225 | **0.3** |
| Univ | **0.435** | 3.08 | 0.465 | **0.56** |
| Zara1 | 0.330 | 0.18 | **0.32** | **0.08** |
| Zara2 | **0.255** | 0.99 | 0.289 | 1.70 |

The metrics presented in Table 1 is for 30 epoch and shows that proposed hard negative sampling performs better collision avoidance than Social-NCE model in all datasets except ETH , however Social-NCE shows better ADE and FDE for all dataset.

## 5. Conclusion

From the results as presented in Section 4, we can observer better collision avoidance performance by the implemented model. This paper proposed a methodology to effectively sample data using hard negative sampling as data augmentation technique. An enhanced method that derives motion representation with no changes to the primary task, thus this method can be easily implemented along with any deep learning models. As the result shows, there is a significant improvement in collision avoidance using the hard negative sampling approach, as measured by values of 0.3 in the Hotel dataset and 0.56 in the Univ dataset as well as 0.08 in the Zara1 dataset.

## References

[1] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations, 2020.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016.

[3] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks, 2018.

[4] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast, 2019.

[5] Yannis Kalantidis, Mert Bülent Sariyildiz, Noé Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *CoRR*, abs/2010.01028, 2020.

[6] Dirk Helbing and Péter Molnár. Social force model for pedestrian dynamics. *Physical Review E*, 51(5):4282–4286, May 1995.

[7] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287, 2002.

[8] Gianluca Antonini, Santiago Venegas, Michel Bierlaire, and Jean-Philippe Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69, 08 2006.

[9] John Funge, Xiaoyuan Tu, and Demetri Terzopoulos. Cognitive modeling: Knowledge, reasoning and planning for intelligent characters. In *Proceedings of*

*the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 29–38, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[10] Sujeong Kim, Stephen Guy, Wenxi Liu, Rynson Lau, Ming Lin, and Dinesh Manocha. Predicting pedestrian trajectories using velocity-space reasoning. *The International Journal of Robotics Research*, 34, 01 2014.

[11] Jur van den Berg, Stephen J. Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In Cdric Pradalier, Roland Siegwart, and Gerhard Hirzinger, editors, *Robotics Research*, pages 3–19, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[12] Javier Alonso-Mora, Andreas Breitenmoser, Martin Rufli, Paul Beardsley, and Roland Siegwart. *Optimal Reciprocal Collision Avoidance for Multiple Non-Holonomic Robots*, pages 203–216. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[14] T. Leung and G. Medioni. Visual navigation aid for the blind in dynamic environments. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 579–586, 2014.

[15] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th International Conference on Computer Vision*, pages 261–268, 2009.

[16] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. *Computer Graphics Forum*, 26(3):655–664, 2007.

[17] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[19] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective, 2021.