Trajectory Prediction in Autonomous Driving System

First Author
2021320331 Boyoung Kim 2021320301 Seonga Choi 2022320112 Kyehong Park
2022320124 Sangjun Song

Abstract

This paper focuses on trajectory prediction for autonomous driving, addressing the challenge of forecasting the paths of multiple interacting agents. We propose an advanced framework based on Generative Adversarial Networks (GANs), which leverages both historical path data and contextual scene information through image analysis. Our method integrates physical and social attention mechanisms to accurately predict future trajectories by highlighting crucial elements in the environment and interactions among agents. This integrated-attention approach enables our model to generate realistic and diverse future paths. Additionally, our method uses top-view images to understand the environment more precisely. Our experimental results demonstrate state-of-the-art performance across various trajectory forecasting benchmarks, showcasing the effectiveness of combining physical and social insights for trajectory prediction in autonomous driving scenarios.

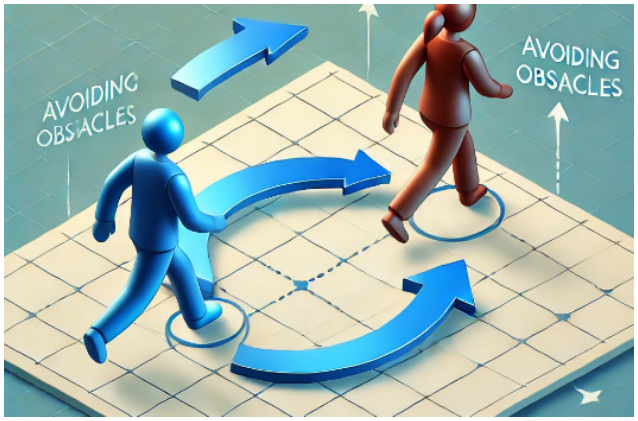


Figure 1. Illustration of two pedestrians avoiding each other with social interaction.

1. Introduction

Trajectory prediction in human-centric environments is a challenging task due to the complex and dynamic nature of human behavior. Two primary factors contribute to this difficulty:

1. **Influence of Human Interaction**
   Human interaction plays a significant role in trajectory prediction. People do not make movement decisions in isolation; rather, they are influenced by the presence and actions of others around them. This leads to a multitude of possible trajectories depending on the specific context and interactions with other individuals. For instance, in a crowded environment, a person may alter their path to avoid collisions or to maintain personal space, resulting in varied and unpredictable trajectories.
2. **Adherence to Social Norms**
   Humans are social beings who adhere to social norms and conventions. These social norms often dictate behavior that deviates from purely physical constraints. For example, while a certain path may be physically accessible, individuals may avoid it due to social conventions or etiquette. This adherence to social norms means that not all physically feasible paths are utilized, adding another layer of complexity to trajectory prediction.

To address these challenges, social acceptability has been recently revisited with data-driven techniques based on Recurrent Neural Networks (RNNs). These techniques leverage the ability of RNNs to model temporal dependencies in sequential data, making them well-suited for predicting human trajectories. Specifically, models such as Social-LSTM utilize Long Short-Term Memory (LSTM) networks to capture social interactions among pedestrians, enhancing prediction accuracy by considering both individual and group behaviors.

Furthermore, Generative Adversarial Networks (GANs) have been employed to generate realistic and diverse future trajectories, modeling the uncertainty and variability inherent in human motion. The use of GANs allows for the creation of multiple plausible future paths, which is critical for applications in autonomous driving and robotics.

To improve the computational efficiency and speed of trajectory prediction models, Gated Recurrent Units (GRUs) have been used as an alternative to LSTMs. GRUs simplify the network architecture while maintaining performance, enabling faster training and inference times.

In this paper, we propose an advanced framework that combines these approaches, integrating physical and social attention mechanisms to accurately predict future trajectories. By leveraging top-view images to gain a comprehensive understanding of the environment, our model achieves state-of-the-art performance across various trajectory forecasting benchmarks. This highlights the effectiveness of integrating physical and social insights for trajectory prediction in autonomous driving scenarios.

2. Related Work

Research in the field of human behavior forecasting can be broadly categorized into two groups: predicting human-space interactions and predicting human-human interactions. Human-space interaction research focuses on how individuals navigate through environments, taking into account static obstacles and dynamic elements of the scene. Key contributions in this area have identified and modeled scene-specific motion patterns to predict individual trajectories based on environmental cues.

In contrast, predicting human-human interactions involves modeling the dynamic and often complex interactions between pedestrians. This area of research is crucial for understanding how individuals influence each other's paths in shared spaces. Our work primarily addresses this second category, focusing on the intricate behaviors arising from social interactions.

2.1. RNNs for Sequence Prediction

Recurrent Neural Networks (RNNs) are a versatile class of models that extend traditional feedforward neural networks to handle sequential data. They have been successfully applied in various domains such as speech recognition, machine translation, and image captioning.[8] However, RNNs often face challenges in capturing high-level spatio-temporal structures due to their inherent limitations in handling long-range dependencies and complex interactions.

To overcome these challenges, several approaches have been developed.Introduced a social pooling layer that models the influence of nearby pedestrians on an individual's trajectory.[1][4][6] This approach enhances the RNN's ability to capture social interactions by pooling information from surrounding agents. Similarly, other methods have employed multiple networks to capture different aspects of interaction dynamics, thus improving the prediction accuracy in crowded scenes.

2.2. GRUs for Improved Efficiency

Gated Recurrent Units (GRUs)[3] have been proposed as an efficient alternative to traditional RNNs and Long Short-Term Memory (LSTM) networks.[2] GRUs simplify the network architecture by reducing the number of parameters while maintaining performance, which leads to faster training and inference times. This makes GRUs particularly suitable for real-time applications where computational efficiency is crucial. By incorporating GRUs, our approach aims to achieve high accuracy in trajectory prediction while ensuring the model operates efficiently in real-time scenarios.

2.3. Generative Modeling

Generative models, such as variational autoencoders (VAEs)[7], are designed to learn the underlying distribution of the training data and generate new samples that resemble the input data. VAEs achieve this by maximizing the lower bound of the data likelihood. An alternative approach, Generative Adversarial Networks (GANs)[5], formulates the training process as a game between a generator, which creates synthetic data, and a discriminator, which differentiates between real and generated data.

GANs have demonstrated remarkable success in various tasks, including super-resolution, image-to-image translation, and image synthesis. These models excel in producing diverse and realistic outputs, making them well-suited for applications that require generating multiple plausible outcomes. Despite their potential, the application of GANs to sequence generation problems, such as natural language processing, has been limited due to the challenges associated with non-differentiable sampling processes.

Our work leverages the strengths of both RNNs and GANs to develop a robust framework for predicting human-human interactions in trajectory forecasting. By integrating physical and social attention[9][10] mechanisms, our model achieves state-of-the-art performance across multiple benchmarks, demonstrating the effectiveness of combining these advanced techniques for trajectory prediction in autonomous driving scenarios.

3. Method

3.1. Model Architecture

Our model improves upon the S-GAN by incorporating background-contexts and modifying the pooling module.

LSTM to GRU. We replaced the LSTM used in both the Generator and Discriminator encoders with GRU. This change was made to address the slow training speed of the original S-GAN, as GRUs generally provide faster computation and require fewer resources compared to LSTMs.

Incorporating Background Context. One of the significant shortcomings of SGAN was its limited context. Information such as obstacles and traffic signals, which are cru-

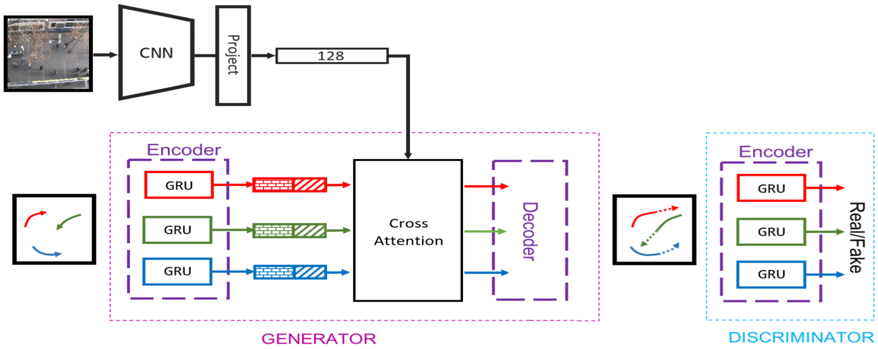


Figure 2. Our model consists of Image module, Cross-Attention module, and Encoder-Decoder GAN architecture. The generator is designed to process both trajectories and images. Specifically, the Cross-Attention outputs integrated vector for each person by taking feature vectors derived from images and trajectory vectors. This integrated approach enhances the model's ability to generate more accurate and contextually relevant outputs.

cial for realistic pedestrian trajectory prediction are omitted. To address this, we introduced an image-based module:

- Each pedestrian dataset consists of 8 frames for observation. The last image of these frames is used to extract the feature vector.
- We used a pretrained EfficientNet to extract feature vectors from the images. These feature vectors were then transformed into 128-dimensional vectors to align with the trajectory vectors.

Cross Attention. The Pooling Module, originally used to consider interactions between pedestrians, has been replaced with an Attention mechanism. The Pooling Module, while effective, adds significant computational overhead and increases training time. By using Cross Attention, we maintain the core functionality of interaction consideration while reducing the computational burden.

- The key and value are derived from each pedestrian's trajectories, which are encoded using a Encoder. This allows us to capture the temporal dependencies and movement patterns inherent in pedestrian trajectories.
- The query is extracted from feature vectors obtained from images. This method leverages the rich contextual information from images, enhancing the model's ability to understand and predict interactions in dynamic scenes.

With Attetion mechanism, the social contexts of pedestrians and environmental contexts are both considered. By integrating these elements, we achieve a more computationally efficient model that still retains the functionality of Pooling Module.

3.2. Data

For this project, we utilized the Stanford Drone Dataset(figure 3), which includes videos and trajectory data captured by drones over specific areas at Stanford University. The original dataset used by Social GAN, namely ETH/UCY(figure 4), lacks a sufficient number of scenes featuring both cars and pedestrians, as illustrated in the right image. Consequently, we selected the ”deathCircle” subset from the Stanford Drone Dataset, which is more appropriate for autonomous driving tasks due to its top-down view encompassing both vehicles and pedestrians. The structure of the Stanford Drone Dataset is depicted in the figure 5. We restructured this dataset to conform to the format of the ETH/UCY dataset(table 1) used in the original Social GAN study. Additionally, we converted the videos from the Stanford Drone Dataset into images for our analysis.

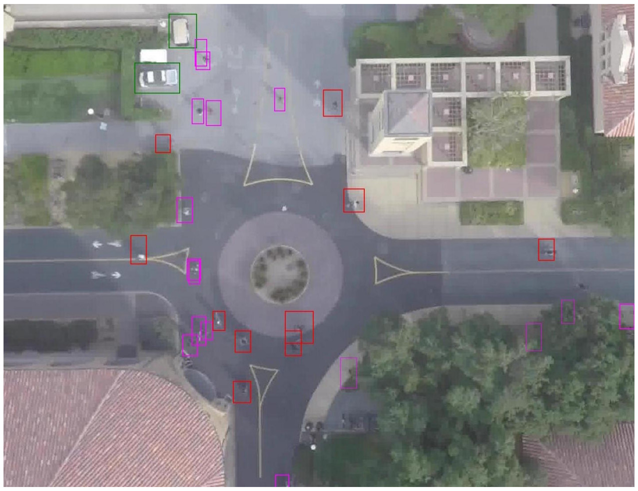
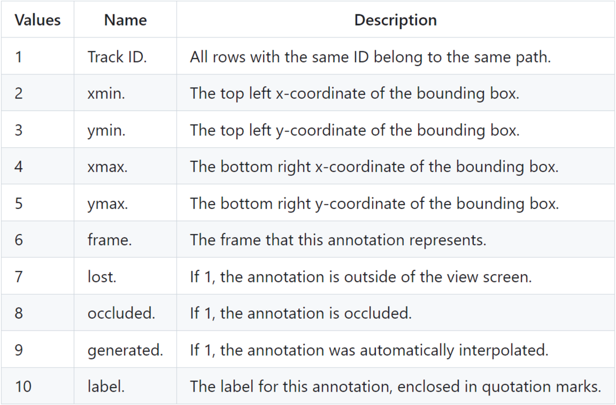


Figure 3. Stanford Drone Dataset

Figure 4. ETH/UCY Dataset



Values	Name	Description
1	Track ID.	All rows with the same ID belong to the same path.
2	xmin.	The top left x-coordinate of the bounding box.
3	ymin.	The top left y-coordinate of the bounding box.
4	xmax.	The bottom right x-coordinate of the bounding box.
5	ymax.	The bottom right y-coordinate of the bounding box.
6	frame.	The frame that this annotation represents.
7	lost.	If 1, the annotation is outside of the view screen.
8	occluded.	If 1, the annotation is occluded.
9	generated.	If 1, the annotation was automatically interpolated.
10	label.	The label for this annotation, enclosed in quotation marks.

Figure 5. Structure of the Stanford Drone Dataset

ETH/UCY	**Stanford Drone Dataset**	**Description**
Frame Number	6 (frame)	Frame number
Pedestrian ID	1 (track id)	Unique ID for each pedestrian
x Coordinate	Mean of 2 (xmin) and 4 (xmax)	Average x-coordinate of the bounding box
y Coordinate	Mean of 3 (ymin) and 5 (ymax)	Average y-coordinate of the bounding box

Table 1. Reformatted ETH/UCY Dataset Structure using Stanford Drone Dataset

3.3. Evaluation Metrics

For trajectory evaluation, we utilized two primary metrics: Average Displacement Error (ADE) and Final Displacement Error (FDE).

Average Displacement Error (ADE) calculates the mean difference between the predicted trajectory and the ground truth trajectory at each time step. Mathematically, ADE is defined as:

$$\text{ADE} = \frac{1}{T}\sum_{t=1}^{T}\|\hat{y}_t - y_t\| \qquad (1)$$

where T is the number of time steps, $\hat{y}_t$ is the predicted position at time step t, and y_t is the ground truth position at time step t.

Final Displacement Error (FDE) measures the difference between the predicted trajectory and the ground truth trajectory only at the final time step. It is given by:

$$\text{FDE} = \|\hat{y}_T - y_T\| \qquad (2)$$

where T is the final time step, $\hat{y}_T$ is the predicted position at the final time step, and y_T is the ground truth position at the final time step.

Our evaluation method follows the approach used by Social GAN. We predict the trajectory of pedestrians from time steps 9 to 16 and from time steps 9 to 20 based on their trajectory up to time step 8. We then measure the ADE and FDE for these two prediction intervals to assess the accuracy of our trajectory predictions.

4. Results

4.1. Training Time

We calculated the average training time when the total iteration of each model reached 2,000 times.

Model	Average Training Time
SGAN	47.44s
Ours	7.61s

Table 2. It summarizes the average training time per epoch for SGAN and ours.

Our model demonstrates a significant reduction in training time, achieving an approximately 85% decrease compared to SGAN. This efficiency gain can be attributed to the replacement of the LSTM with GRU and the use of the Attention mechanism instead of the Pooling Module.

4.2. Comparison with S-GAN

The table below displays the ADE/FDE results for each model. While including the values from the original SGAN paper for reference, we focus on the outcomes obtained by the SGAN model trained under our environments.

Our model shows improved ADE and FDE values compared to the S-GAN, although it falls slightly short of the original paper's reported values. This indicates that our model is capable of better performance, with further refinement of the training environment.

	SGAN (Original)	SGAN	Our Model
ETH	0.79 / 1.13	4.11 / 7.52	1.62 / 2.85
SDD	-	3.85 / 6.96	0.77 / 1.38

Table 3. ADE/FDE values comparison for SGAN and our model.

5. Conclusion

5.1. Analysis

This project successfully addresses the limitations of Social GAN (SGAN) by introducing several key improvements. By integrating a Convolutional Neural Network (CNN), the model effectively incorporates spatial information, significantly enhancing obstacle detection and environmental awareness. Replacing LSTM with GRU and removing the pooling layer in favor of Cross-attention resulted in a substantial reduction in computational complexity, making the model 3-5 times faster than the original SGAN. Through rigorous analysis and architectural modifications, this project demonstrates SGAN's enhanced performance in complex real-world scenarios. Improved data preprocessing, including converting the SDD to ETH format and normalizing coordinates, further boosted model accuracy. The contributions of this project extend beyond addressing SGAN's limitations. The innovative integration of CNN, GRU, and Cross-attention provides a valuable framework for future research in social interaction modeling. Additionally, the improved data preprocessing techniques offer practical solutions for optimizing model performance in real-world applications. Overall, this project successfully enhances the robustness, flexibility, and efficiency of SGAN, making it a more valuable tool for understanding and predicting social interactions in dynamic environments.

5.2. Limitation

This project also has some limitations. First, the GAN model may experience instability during training, requiring careful hyperparameter tuning to achieve optimal performance. The project's performance could be further enhanced with more extensive hyperparameter optimization.

Additionally, the addition of CNN increases model complexity, leading to longer training times and higher computational resource consumption. This could be a limiting factor for real-time applications or those with limited resources.

Moreover, due to time constraints, training was limited to the DeathCircle dataset, chosen for its suitability to the model's goals. This limitation makes it difficult to guarantee the model's generalizability to other datasets without further evaluation and potentially additional training.

Future work could focus on addressing these limitations by exploring more efficient training methods for GANs, optimizing the model architecture to reduce complexity, and conducting more extensive evaluations on a wider range of datasets to ensure broader applicability.

5.3. Future Work

Based on the identified limitations and findings of this study, we propose several promising future research directions:

1) Enhancing GAN Stability: Exploring various techniques like Wasserstein GAN (WGAN) or Gradient Penalty could significantly improve the stability of GAN training, leading to more robust and reliable model performance.

2) Alternative Network Architectures: Investigating the performance of Transformer-based models or other neural network structures besides GRU could reveal alternative approaches for further improving social interaction modeling and trajectory prediction.

3) Evaluating Generalization Performance: Conducting comprehensive evaluations on publicly available Trajectory Prediction datasets such as ETH and UCY would provide valuable insights into the model's generalization capabilities and robustness across diverse scenarios.

4) Leveraging State-of-the-Art Techniques: Referencing recent advancements like Social-BiGAT and Trajectron++ could inspire further model improvements by incorporating the latest technologies and methodologies in social interaction modeling.

5) Optimizing Model Efficiency: Exploring techniques like Pruning or Knowledge Distillation could offer effective solutions for reducing model complexity and enhancing computational efficiency without sacrificing performance.

By pursuing these research directions, we anticipate significant advancements in social interaction modeling, leading to more accurate, reliable, and efficient trajectory prediction models for real-world applications.

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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 2

[2] Roberto Cahuantzi, Xinye Chen, and Stefan Güttel. *A Comparison of LSTM and GRU Networks for Learning Symbolic Sequences*, page 771–785. Springer Nature Switzerland, 2023. 2

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. 2

[4] Nachiket Deo and Mohan M. Trivedi. Convolutional social pooling for vehicle trajectory prediction, 2018. 2

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2

[6] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks, 2018. 2

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 2

[8] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404: 132306, 2020. 2

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2

[10] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds, 2018. 2