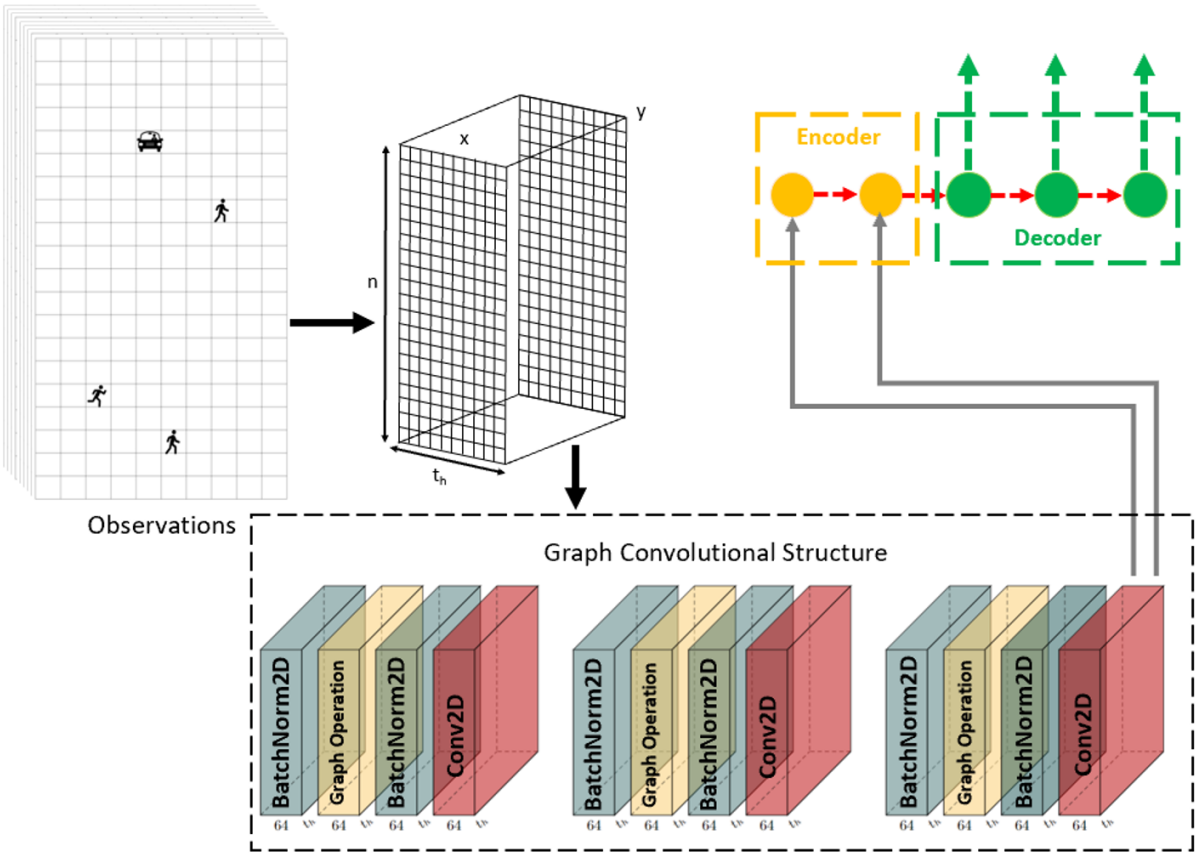
CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Final Report #61
June 2022



Translation of driver-pedestrian behavioral models at semi-controlled crosswalks into a quantitative framework for practical self-driving vehicle applications

Yunchang Zhang
Jon D. Fricker

PURDUE UNIVERSITY®

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Report No. 61
Report Date: June 2022
Project Start Date: 01/01/2020
Project End Date: 12/31/2021

Translation of Driver-Pedestrian Behavioral Models at Semi-Controlled Crosswalks into a Quantitative Framework for Practical Self-Driving Vehicle Applications

Yunchang Zhang
Graduate Researcher

Jon. D. Fricker
Professor

Purdue University

UMTRI | The University of Akron | Illinois Center for Transportation | Purdue University | Washtenaw Community College | Central State University

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

ACKNOWLEDGEMENT AND DISCLAIMER

Funding for this research was provided by the Center for Connected and Automated Transportation under Grant No. 69A3551747105 of the U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology (OST-R), University Transportation Centers Program. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

Suggested APA Format Citation:

Zhang, Y., and Fricker, J.D. (2022). Translation of Driver-Pedestrian Behavioral Models at Semi-Controlled Crosswalks into a Quantitative Framework for Practical Self-Driving Vehicle Applications, Technical Report Nr. 61, Center for Connected and Automated Transportation, Purdue University, West Lafayette, IN.

Contact Information

Samuel Labi
3000 Kent Ave., West Lafayette, IN
Phone: 7654945926
Email: labi@purdue.edu

Jon. D. Fricker
550 Stadium Mall Dr.
W. Lafayette, IN
Phone: (765) 494-2205
Email: fricker@purdue.edu

CCAT
University of Michigan Transportation Research Institute
2901 Baxter Road
Ann Arbor, MI 48152

uumtri-ccat@umich.edu
(734) 763-2498
www.ccat.umtri.umich.edu

UMTRI | The University of Akron | Illinois Center for Transportation | Purdue University | Washtenaw Community College | Central State University

CENTER FOR CONNECTED AND AUTOMATED TRANSPORTATION

Technical Report Documentation Page

1. Report No. CCAT Report #61	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Pedestrian-Vehicle Interaction in a CAV Environment: Explanatory Metrics		**5. Report Date** June 2022 **6. Performing Organization Code** N/A
7. Author(s) Yunchang Zhang, Jon. D. Fricker		**8. Performing Organization Report No.** N/A
9. Performing Organization Name and Address Center for Connected and Automated Transportation Purdue University, 550 Stadium Mall Drive, W. Lafayette, IN 47907; and University of Michigan Ann Arbor, 2901 Baxter Road, Ann Arbor, MI 48109		**10. Work Unit No.** N/A **11. Contract or Grant No.** Contract No. 69A3551747105
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final Report. 01/01/2020 - 12/31/2021 **14. Sponsoring Agency Code** OST-R

15. Supplementary Notes
Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.

16. Abstract
The research described in this report was motivated by frequent questions from users of several crosswalks near a university campus. At each crosswalk was a sign indicating that motorists should yield to pedestrians in the crosswalk. The notion that this message was not being interpreted uniformly was a concern at locations where heterogeneous road users (pedestrians, cyclists, and motorists) were interacting. Instead of trying to impose a single interpretation on users of each crosswalk, it was decided to observe and analyze interactions between users of the crosswalk.

Several hours of video were recorded of pedestrians and motorists "negotiating" the right of way at the crosswalk. Because these crossing locations were marked but not signalized, they were called "semi-controlled crosswalks". Recently, computer vision (CV) algorithms have been extensively used in road users' detection and tracking at an unparalleled spatial-temporal scale. In this study, CV algorithms have been applied to convert the video recordings into a large-scale spatial-temporal trajectory dataset including 800 pedestrians and cyclists interacting with more than 500 vehicles. Utilizing the trajectory dataset, a spatial-temporal graph convolutional network-based sequence to sequence (ST-GCN-Seq2Seq) algorithm has been developed to forecast heterogeneous road users' trajectories and behavior in real time.

To demonstrate the model's performance, the proposed ST-GCN-Seq2Seq was compared with with state-of-the-art human motion prediction models. Comparison results and case studies confirmed that the ST-GCN-Seq2Seq model can accurately predict future movements and interactions of heterogeneous road users. Combining CV and ST-GCN-Seq2Seq algorithms can help both design an intelligent tracking system and achieve a form of "smart" interaction at semi-controlled crosswalks for heterogeneous road users.

17. Key Words Pedestrian crossings, Pedestrian-motorist interaction, Pedestrian wait behavior	**18. Distribution Statement** No restrictions.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 36 pages	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Background and Problem Statement

Transportation systems, consisting of motor vehicles, non-motorized modes, and pedestrians, exist to efficiently move people and goods. At some points, road users of different types interact in shared spaces. In the case of motor vehicles and pedestrians, when a pedestrian crossing is a marked crossing without stop signs or signals (see Figure 1), pedestrian-motorist interactions rely on the parties cooperating and invoking "social rules" to establish priority for use at the site. Pedestrians and motorists plan their trajectories to avoid collisions with conflicting road users, as they interact at the semi-controlled crosswalks (Fricker and Zhang, 2019). For example,

1. Pedestrian-motorist interaction (PMI) Case 1: pedestrians will cross immediately if an approaching vehicle is far away from the crosswalk (Zhang et al., 2020); and
2. PMI Case 2: pedestrians will stop and let vehicles go first if a vehicle is too close to yield to the subject pedestrian in the crosswalk (Zhang et al., 2020).

Nevertheless, "social rules" can be ambiguous due to the lack of communication between road users. There are two cases of special interest:

3. PMI Case 3: if the interacted motorist is neither too far from the crosswalk nor too close (40 feet to 50 feet), the "negotiation" between a pedestrian and a motorist will be more complicated (the pedestrian does not cross, and the motorist yields) and result in some amount of delay (Zhang et al., 2020); and
4. PMI Case 4: if the subject pedestrian steps into the crosswalk, and the interacted driver does not yield to the pedestrian, it would be a dangerous situation (Zhang and Fricker, 2021a).

Figure 1 R1-6 Sign at Semi-Controlled Crosswalk

Instead of four PMI cases, pedestrian-pedestrian interactions (PPIs) and hybrid interactions occur at semi-controlled crosswalks:

1. Pedestrian-pedestrian interaction (PPI) Case 1. pedestrians will respect personal spaces and keep safe distances from other pedestrians (Helbing and Molnar, 1995); and
2. PMI and PPI Case 2: pedestrians will include "safety in numbers" if one or more other pedestrians are present on the crosswalk or in the curb areas (Zhang and Fricker, 2021b).

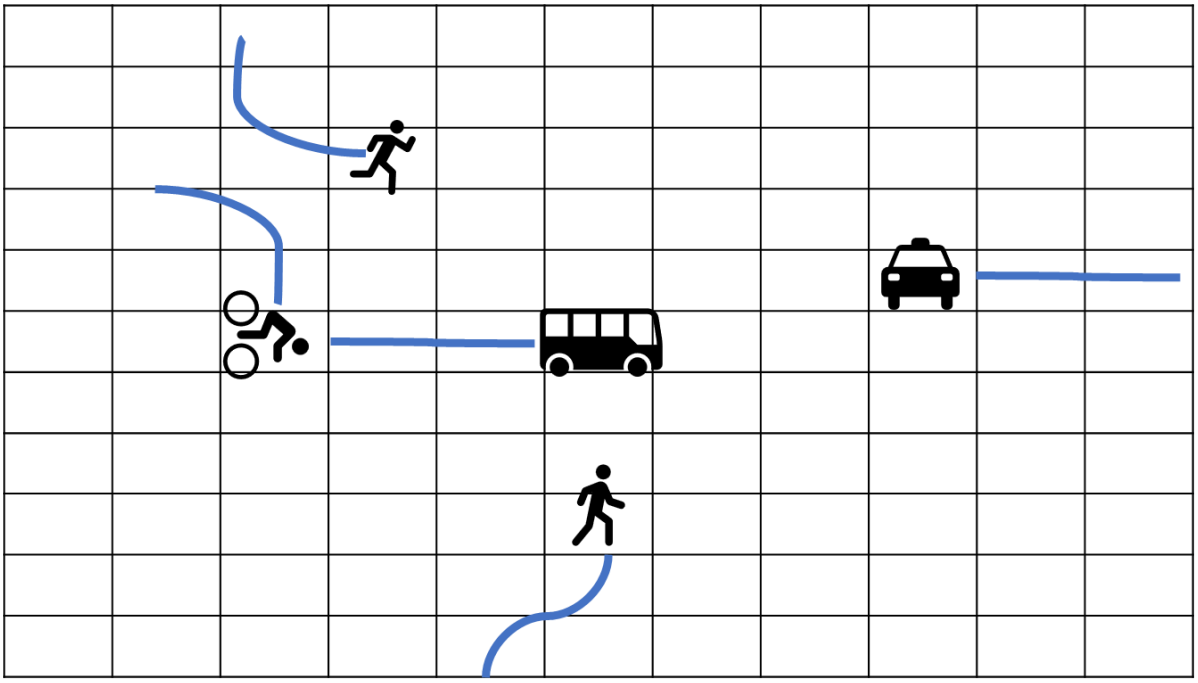
Semi-controlled crossing locations are complex traffic environments where heterogeneous road users (pedestrians, cyclists, and drivers) follow the widely adopted "social rules" (like the four PMI cases and two PPI cases shown before) while interacting with other road users. The "social rules" guide pedestrians and motorists to plan their trajectories and avoid collisions with conflicting road users, as they interact near crosswalk areas. See PMI (Case 1 and Case 2) and PPI (Case 1). "Social rules" are correlated with complicated factors (pedestrian characteristics, vehicle dynamics and environmental factors) that have been proved to significantly influence on road users' decisions (Rasouli and Tsotsos, 2019). However, there is some uncertainty as to how the widely adopted "social rules" will be followed.

Trajectory forecasting is complex. Observations indicate that road users are involved in non-verbal (gesture, pose, etc.) communications over multiple time steps as they interact in shared spaces:

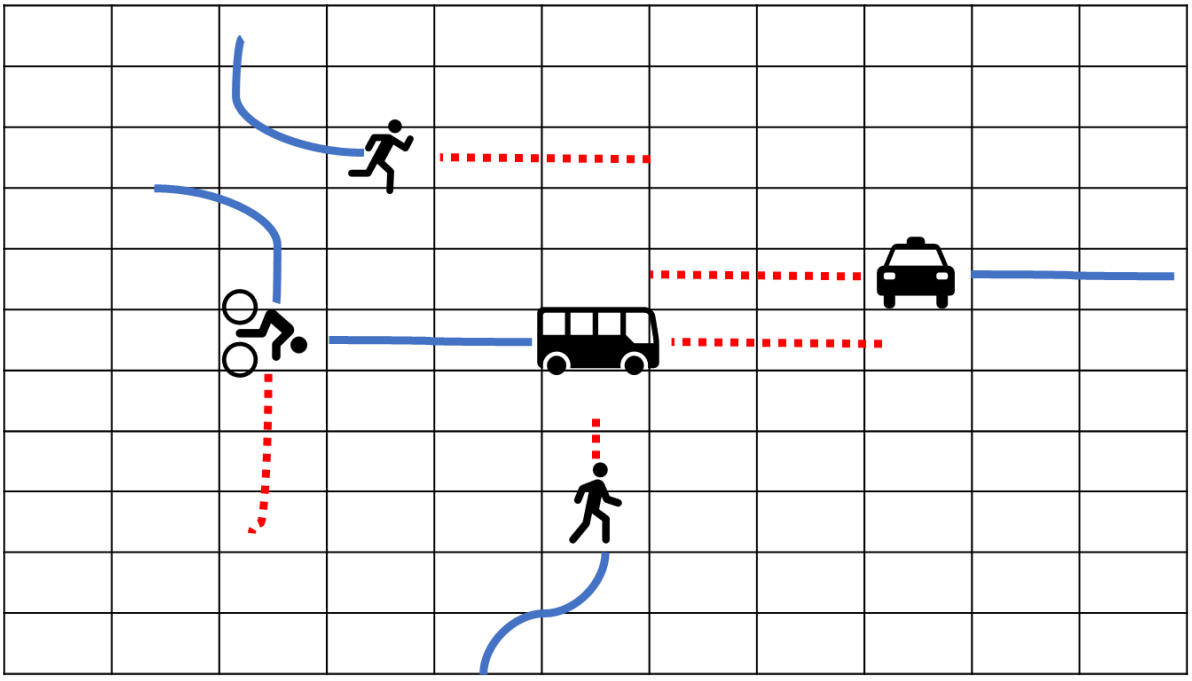
1. If there is a vehicle approaching the crosswalk, a pedestrian is likely to
    a. enter the curb area,
    b. wait for the response of the driver, and
    c. step into the crosswalk if the driver finally decelerates or stops.
2. If there is a pedestrian waiting in the curb area, a driver is likely to
    a. decelerate to give a signal to the pedestrian,
    b. wait for the response of the pedestrian, and
    c. accelerate to leave the crosswalk area if the pedestrian waves to the driver.

To address multiple time interactions and hybrid interactions between heterogeneous road users, the second learning algorithm is introduced as a spatial-temporal graph convolutional network-based sequence to sequence (ST-GCN-Seq2Seq) model. ST-GCN-Seq2Seq predicts road users' future trajectories on the basis of observed trajectories and interactions between road users. Figure 2 reveals how ST-GCN-Seq2Seq works:

1. A frame from video recordings at time *t* is extracted. See Figure 2, and five heterogeneous road users (two pedestrians, one cyclist, and two drivers) are observed.
2. The trajectory dataset is accessed, and the most recent three-second trajectories of heterogeneous road users (blue-solid lines in Figure 2a) are extracted as observed trajectories.
3. ST-GCN-Seq2Seq utilizes the observed trajectories and encodes interactions between heterogeneous road users as input to predict future three-second trajectories of heterogeneous road users (red dotted lines in Figure 2b).



a. Observed Trajectories



b. Predicted Trajectories

Figure 2 A Visualization of Motion Prediction

2. DATA COLLECTION

The objective of this study is to propose an observing-tracking-learning framework that can be generally used in the design of an intelligent tracking system and achieve a form of smart interaction at "smart" crosswalks:

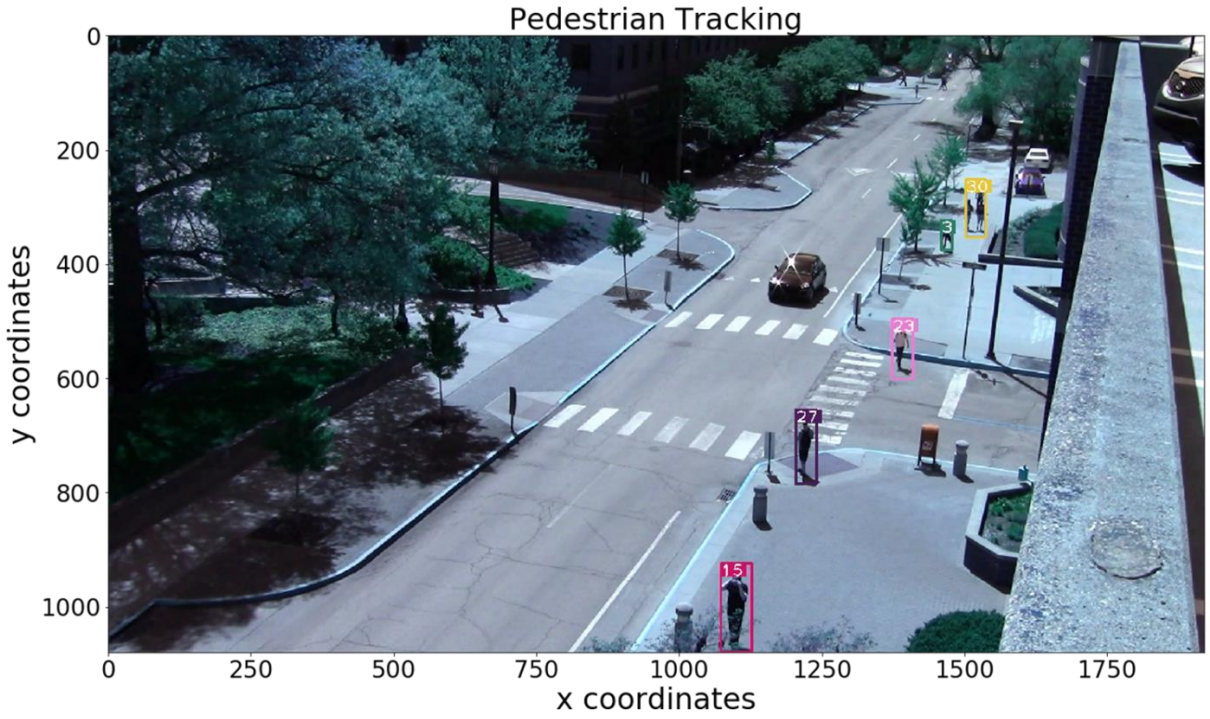
- Module 1: the behavior of pedestrians and motorists interacting in real street-crossings has been documented by hours of video,
- Module 2: computer vision-based techniques have been applied to detect road users (pedestrians, cyclists, and motorists) and track their positions, and
- Module 3: a ST-GCN-Seq2Seq model has been developed predict future trajectories and interactions of heterogeneous road users.

2.1. Data Collection

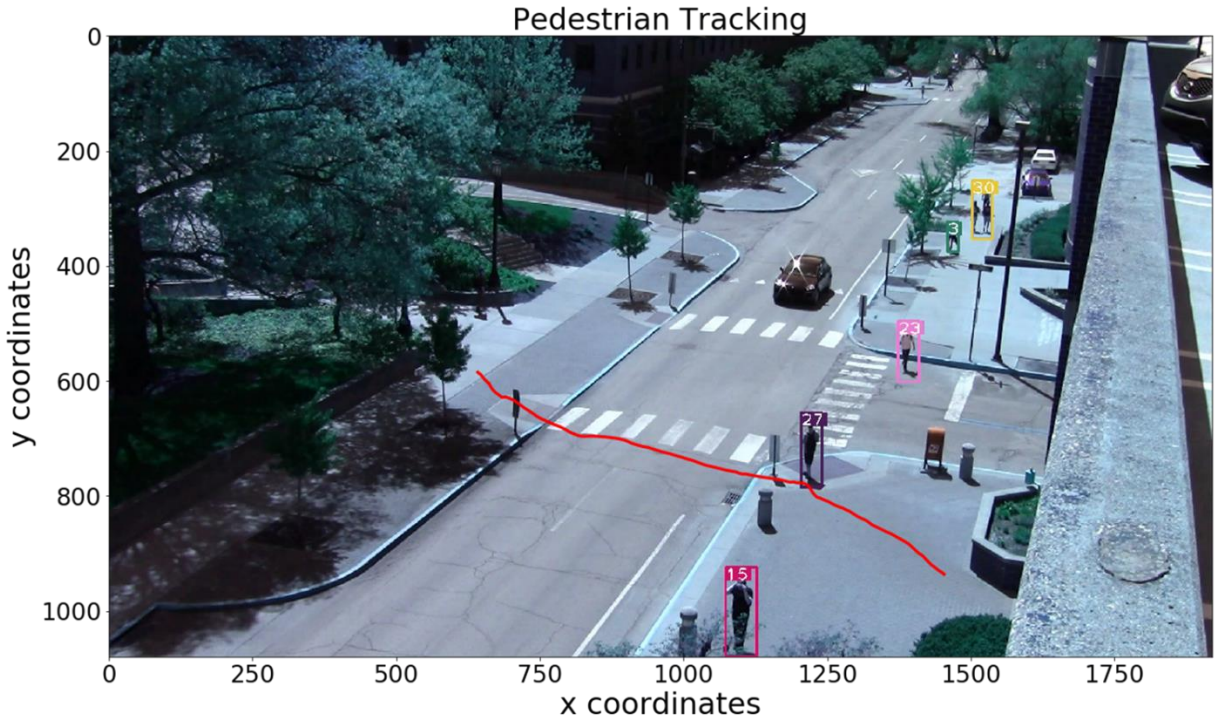
The dataset for this study was collected at a semi-controlled crossing location on the Purdue University campus. The marked crosswalk is "controlled" by "yield to pedestrian" signs. See Figure 1. The sketches of the target crosswalk are shown in Figure 3(a) and Figure 3(b). Video recordings were made in Spring 2017 when University Street was a one-way northbound street. The street had two 10-ft wide lanes (plus a 4-foot bicycle lane) with a speed limit of 25 mph. Video recordings were made at four different 40-minute periods (Zhang and Fricker, 2021).

2.2. Object Detection and Tracking

Yolo-V3 and deep-sort algorithms have been applied in the multi-object (pedestrian, cyclist, and vehicle) detection and tracking. An example of pedestrian detection and tracking has been shown in Figure 3(a) and Figure 3(b). The red solid line in Figure 3(b) represents the extracted trajectory for the subject pedestrian No. 27. The original Yolo-V3 and deep-sort algorithms can be found in the GitHub repository (https://github.com/ZQPei/deep_sort_pytorch). A modified version of the algorithms used in the study site can be found in our GitHub repository (https://github.com/YZhang-Genghis).



(a) An Example of the Pedestrian Detection Module



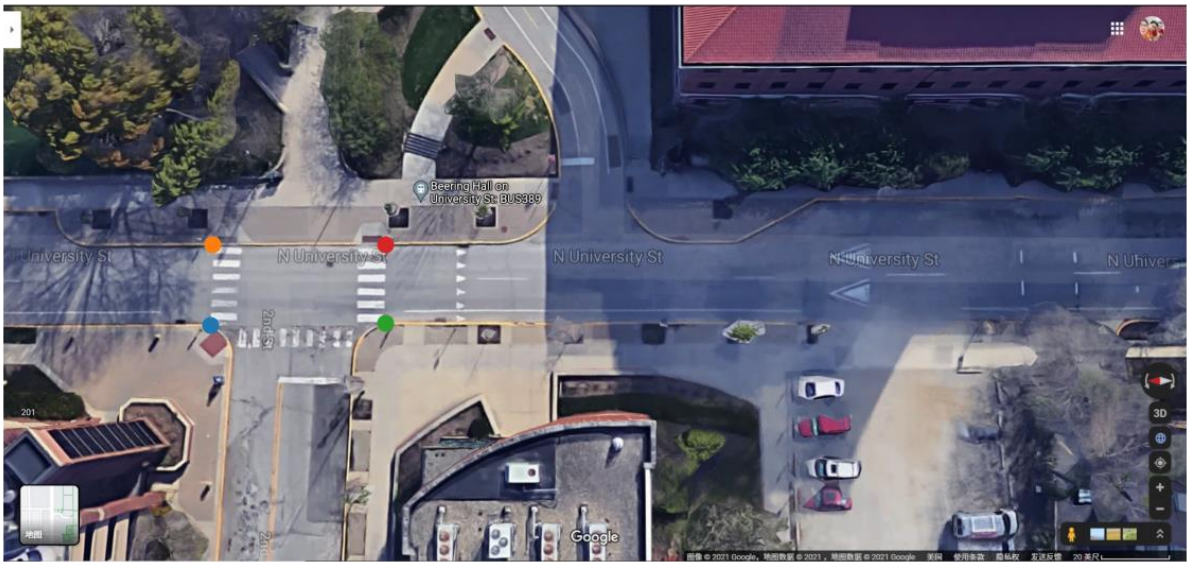
(b) An Example of the Pedestrian Tracking Module

Figure 3 Semi-Controlled Crosswalk

2.3. Homography

After extracting pedestrian, cyclist, and vehicle trajectories, we hope to project the dataset of trajectories from the camera view into the Google Map view to obtain the precise latitude and longitude maneuvers of agents. Consider two images of the intersection shown in Figure 4(a) and Figure 4(b), the four corresponding points in four different colors – red, green, orange, and blue dots represent the same physical points in the two images.

(a) Camera View of the Study Site



(b) Google Map View of the Study Site

Figure 4 Homography

The four colored points in Figure 4 (a) can be projected onto the corresponding points in Figure 4 (b) using the Homography matrix. Considering a two-dimensional point (x_1, y_1) in Figure 4(a) and the corresponding two-dimensional point (x_2, y_2) in Figure 4(b), a Homography is a transformation (a 3×3 matrix) that maps the (x_1, y_1) to the corresponding point (x_2, y_2):

$$\begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = H \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} \quad (1)$$

The Homography matrix *H* can be estimated using the *findHomography* function in OpenCV (https://docs.opencv.org/3.4/d7/dff/tutorial_feature_homography.html). A script of the Homographic transformation of the given dataset can be found in our GitHub repository (https://github.com/YZhang-Genghis). The transformed pedestrian trajectories and vehicle trajectories are shown in Figure **5**. W extracted a total of 812 vehicle trajectories and 511 pedestrian trajectories.

(a) Pedestrian Trajectories (b) Vehicle Trajectories

Figure 5 Transformed Road Users' Trajectories

3. MOTION PREDICTION

The motion prediction problem is to estimate future trajectories (from time t+1 to time t_f) of road users who appear at the time stamp t, based on their observed trajectories (from time t-t_h+1 to time t) and interactions.

3.1. Trajectory

The definition of the trajectory for the subject road user i can be formulated as a time sequence: $Tr_i\left(x_t, y_t\right) \in \{\Re^2\}$, where $\left(x_t, y_t\right) \in \Re^2$ represents the spatial coordinates of the subject road user's position at time t.

3.2. Input – Observed Trajectory

The input of the model is the observed trajectory of the subject road user i over t_h time steps (t_h = 3 seconds): $X_i\left(Ob\_Tr_i\right) = \left[\left(x^{(i)}_{t-t_h+1}, y^{(i)}_{t-t_h+1}\right), \left(x^{(i)}_{t-t_h+2}, y^{(i)}_{t-t_h+2}\right), \ldots, \left(x^{(i)}_{t}, y^{(i)}_{t}\right)\right]$.

3.3. Prediction – Future Trajectory

The output of the model is the future trajectory of the subject road user i over t_f time steps (t_f = 3 seconds): $Y_i\left(Fut\_Tr_i\right) = \left[\left(x^{(i)}_{t+1}, y^{(i)}_{t+1}\right), \left(x^{(i)}_{t+2}, y^{(i)}_{t+2}\right), \ldots, \left(x^{(i)}_{t+t_f}, y^{(i)}_{t+t_f}\right)\right]$.

4. NETWORK ARCHITECTURE

The Spatial-Temporal Graph-Based Seq2Seq model structure is shown in Figure 6.

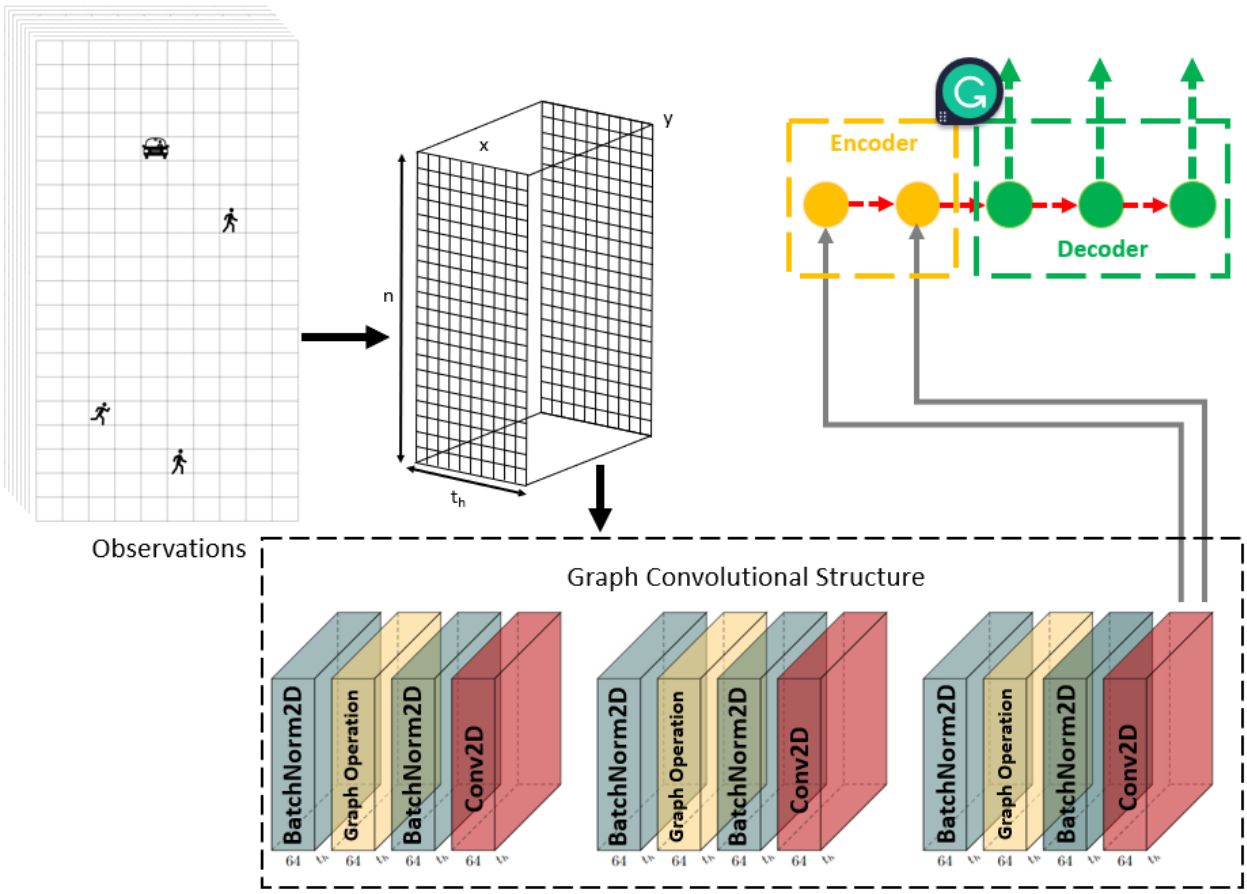


Figure 6 Spatial-Temporal Graph-Based Seq2Seq Model Structure

4.1. Input

The dataset can be represented as $t \times C \times P \times V$:

- t represents the total amount of timeframes.
- C represents the number of features (category of road users, longitudinal coordinate, latitudinal coordinate, and head direction).
- $P = t_h + t_f$ represents the temporal domain with observed 3 seconds and predicted 3 seconds.

- $V = 20$ represents the maximum number of road users showing in one frame. The value of V can change based on the types of data.

The entire dataset is split into subsets (mini batches). Each mini batch can be represented as a tensor with a size of ($n \times C \times t_h \times V$), where n represents the number of mini-batches, C represents the vector of 2-dimensional spatial coordinates $\left(x_t, y_t \right) \in \mathfrak{R}^2$ and additional features such as category of road users and head direction, t_h represents the number of observed time frames, and $V = 20$ represents the maximum number of road users showing in one frame. See Figure 6.

4.2. Spatial-Temporal Graph Construction

In urban traffic settings, the subject road user's movements/behavior are significantly influenced by the nearby agents (see the four PMI cases mentioned in the Introduction). To handle the inter-dependencies between the trajectory of the subject road user and the trajectories of surrounding road users, we propose a social network graph -- an undirected graph $G = \{V, E\}$, where the nodes (V) represent road users and the edges (E) represent interactions between road users.

- Each node v_{it} in the node set V, represents a road user i appearing in a time frame t. Then, the node set can be constructed as $V = \left\{ v_{it} \mid i = 1, \ldots, n;\ t = 1, \ldots, t_h \right\}$.
    - n is the total number of road users observed at time frame t; and
    - the feature vector ($F\left(v_{i,t}\right)$) of the node v_{it} consists of the spatial coordinate $\left(x_t^{(i)},\ y_t^{(i)} \right)$ of road user i at a time t.
- The edge set E consists of two parts:
    - $E_S = \left\{ \left(v_{it},\ v_{jt} \right) \mid Dist\left(i, j\right) \le D \right\}$: the set of edges describes the spatial interactions between node v_{it} and node v_{jt}.
        - $Dist\left(i, j\right)$ represents the Euclidean distance between the road user i and the road user j.
        - D is a threshold value that represents the spatial closeness between the road user i and the road user j in one frame. In this report, we first choose a large D value as D = 380 feet (116 meters).
            - If the spatial distance between node v_{it} and node v_{it} is less than D at time t, the pair of nodes $\left(v_{it},\ v_{jt} \right)$ is included in the edge set E_S .

- A large *D* value represents that every pair of - nodes $\left(v_{it},\ v_{jt}\right)$ shown in the same time frame *t* will be included in the edge set E_S regardless of the spatial distance between road user *i* and road user *j*

- $E_T = \left\{\left(v_{it},\ v_{i(t+1)}\right)\right\}$: the set of edges describes the temporal difference only for road user *i* between time *t* and *t*+1. All pairs of edges in E_T for road user *i* represents the road user *i's* trajectory.

- To better demonstrate the interaction between road users at a single frame *t*, an adjacency matrix $A=\left[I, A_s\right]$ is proposed:
  - I indicates the self-connection of the subject road user in temporal space; and
  - A_S indicates whether pairs of nodes $\left(v_{it},\ v_{jt}\right)$ are in the edge set E_S :

$$A_S[i][j] = \begin{cases} 1 & \text{if } \left(v_{it},\ v_{jt}\right) \in E_S \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

- Both I and A_S are $n \times n$ square matrices, where *n* is equal to the number of road users appearing at each time frame.

4.3. Spatial Temporal Graph Convolutional Network (ST-GCN) Module

The graph convolutional module consists of three parts:

1. BatchNorm2D: batch normalization is a technique for training deep neural networks that standardizes the inputs to a layer for each mini batch (Ioffe and Szegedy, 2015). For each input ($n \times C \times t_h \times V$), BatchNorm2D operation normalizes the input using Equation (3).

$$y = \frac{X - E[X]}{Var[X] + \varepsilon} \times \gamma + \beta \quad \textbf{(3)}$$

where:

- ε is a value added to the denominator for numerical stability to avoid that the denominator is zero.
- γ, β are learnable parameter vectors of size *C*.
- The Batch Normalization is done over the *C* dimension, computing statistics on (n, t_h, V) slices.

2. Graph operations: graph operations are known as layer-wise propagations across graph convolutional networks (GCN). Layers of GCN are proposed to capture the spatial interactions of all road users appearing at time *t*. The layer-wise propagations across GCN can be implemented with the following equation (Kipf and Welling 2016):

$$f_{g,\,out} = \sigma\left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} f_{g,\,in} W\right) \quad \textbf{(4)}$$

where:

- $A = A_S + I$ is the adjacency matrix we defined in the Spatial-Temporal Graph Construction section.
- $D^{ii} = \sum_j A_{ij} + \alpha$, D represents the degree matrix of *A* (Kipf and Welling, 2016), and $\alpha = 0.001$ is a small number to avoid empty rows in A_{ij}.
- $\sigma(.)$ denotes the activation function such as $ReLU(.) = max(0, .)$.
- *W* denotes the layer-specific learnable weight matrix.

3. Conv2D: temporal graph convolutional (TCN) layers are proposed to capture the temporal dependencies between consecutive spatial positions of a road user. The output of a GCN layer ($f_{g,\,out}$) is normalized by one BatchNorm2D layer and fed into the TCN layer as $f_{t,\,in}$. A kernel with the size of 1 × 5 is applied in each 2-D convolutional layer with appropriate paddings and strides to move along the temporal axis (t_h) shown in Figure 6. For an input $f_{t,\,in}$, the output of each TCN layer can be represented as $f_{t,\,out}$.

4.4. Seq2Seq

The graph convolutional module is followed by the Seq2Seq module. The Seq2Seq framework is an encoder-decoder network.

The encoder takes the sequence of output of graph convolutional module (length of t_h) - $O_{t'} = \left[O_{t-t_h+1}, O_{t-t_h+2}, \ldots, O_{t-1}, O_t\right]$, feeds it to the embedding layer, derives the series of hidden states $h_{t'} = \left[h_{t-t_h+1}, h_{t-t_h+2}, \ldots, h_{t-1}, h_t\right]$ (for example, a Gated Recurrent Unit (GRU) encoder will calculate the hidden states as h_t = EncoderGRU($e(O_t)$, h_{t-1}), where *e* denotes the embedding operation), and generates the context vector $z = h_t$. The context vector *z* will be fed to the decoder for the future trajectory prediction.

The decoder first calculates the series of hidden states $s_{t'} = \left[s_{t+1}, s_{t+2}, \cdots, s_{t+t_f-1}, s_{t+t_h} \right]$. For example, a GRU decoder will calculate the hidden states as s_t = DecoderGRU($d(y_t)$, s_{t-1}, z). The target value for the next time stamp is calculated as $y_{t+1} = f(d(y_t), s_t, z)$, where f is a linear layer.

Recall that the prediction $Y_i\left(Fut\_Tr_i\right) = \left[\left(x_{t+1}^{(i)}, y_{t+1}^{(i)}\right), \left(x_{t+2}^{(i)}, y_{t+2}^{(i)}\right), \ \ldots, \left(x_{t+t_f}^{(i)}, y_{t+t_f}^{(i)}\right) \right]$ represents the future spatial coordinates of the subject road user's position over t_f time steps. At each time step, the decoder is to predict the two-dimensional spatial coordinate - $\left(x_{t+i}, y_{t+i}\right) \in \Re^2$ where $i \in \left[1, t_f\right]$. The predicted coordinate (x or y) is activated by a *tanh* function within the scale (-1, 1), which will be further re-scaled into real coordinates.

5. EXPERIMENTS AND RESULTS

5.1. Implementation Details

ST-GCN shares the same weights (weight matrix *W* in Equation 4) on different nodes, it is important to keep the scale of input data consistent. In this study, we normalize the spatial coordinates within the range (-1, 1). The ST-GCN-Seq2Seq model is composed of 3 unites of ST-GCN. All three ST-GCN unites have 64 channels for output. And we randomly dropout the features with the probability of 0.5 in each ST-GCN unit to avoid overfitting.

Smooth L1 Loss is chosen as the criterion. For a batch size of *N*, smooth L1 loss can be represented as:

$$l(R,P) = L = \{l_1, l_2, \ldots, l_N\}^T$$

$$l_n = \begin{cases} 0.5(R_n - P_n)^2 / beta & if\ |R_n - P_n| < beta \\ |R_n - P_n| - 0.5 \times beta & otherwise \end{cases} \quad \textbf{(5)}$$

$$l(R,P) = mean(L) \quad \textbf{(6)}$$

beta = 1 is chosen as the default parameter.

Adam (Kingma and Ba, 2014) is used as the optimization algorithm to train the model with a learning rate of 0.01. The learning rate is decayed by 0.1 every 10 iterations.

5.2. Evaluation Metrics

The root mean squared error (RMSE) of the predicted trajectories in the future (3-second horizons) will be reported. The RMSE can be calculated as:

$$RMSE = \sqrt{\frac{1}{N} \times \frac{1}{T} \times \frac{1}{t_f} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{k=1}^{t_f} \left(x_{t+k}^{(i)} - \hat{x}_{t+k}^{(i)}\right)^2 + \left(y_{t+k}^{(i)} - \hat{y}_{t+k}^{(i)}\right)^2} \quad \textbf{(7)}$$

where,

$x_{t+k}^{(i)}$ denotes the real longitudinal coordinate of road user *i* at time *t*+*k*.

$\hat{x}_{t+k}^{(i)}$ denotes the predicted longitudinal coordinate of road user *i* at time *t*+*k*.

$y_{t+k}^{(i)}$ denotes the real latitudinal coordinate of road user *i* at time *t*+*k*.

$\hat{y}_{t+k}^{(i)}$ denotes the predicted latitudinal coordinate of road user *i* at time *t*+*k*.

5.3. Comparison Methods

To confirm the model performance, we compared the proposed model with state-of-the-art (SOTA) models. The SOTA models are chosen based on two metrics:

1. The SOTA method should be applicable in pedestrian, cyclist, or vehicle motion prediction.
2. The SOTA method should be open-access and has been evaluated/validated by other studies.

Accordingly, five SOTA models are chosen as baselines:

1. Social-Force (SF) model: a physics-based model developed by Helbing and Molnar (1995). It is widely used to simulate pedestrian dynamics in an urban traffic environment.
2. Convolutional Social Pooling LSTM (ConvSP-LSTM): the ConvSP-LSTM (Deo et al., 2018) adopted a convolutional social pooling mechanism in the LSTM encoder-decoder model, addressing vehicle-vehicle interactions using the Next Generation Simulation (NGSIM) datasets. A visualization of the proposed ConvSP is shown.
3. Social-LSTM: Social-LSTM has the same model structure as the ConvSP LSTM model except for the convolutional social pooling module (Alahi et al., 2016). Instead, a fully connected pooling (FCSP) module has been adopted to address the interactions between road users. The difference between ConvSP and FCSP is similar to the difference between the Convolutional Neural Network (CNN) and fully connected neural network in image recognition (LeCun and Bengio. 1995).
4. Seq2Seq: it is a classical sequence to sequence model that has been widely applied in the area of natural language processing. LSTM Encoder-Decoder can be considered as the same model as the ConvSP-LSTM model without the convolutional social pooling module.

5.4. Model Results

We split the complete dataset into training and validation sets. The validation set contains the last 20% trajectories in the dataset. We reported the RMSE values of the validation set. All units are in meters. Lower MSE values indicate superior model performance (bold-face values in Table **1**).

Table 1 Model Comparisons

Evaluation Metric	Prediction Horizon (s)	SF	Seq2Seq	Social-LSTM	ConvSP-LSTM	ST-GCN-Seq2Seq
RMSE (m)	0.5	0.406	0.672	0.256	0.267	**0.196**
	1	0.661	1.082	0.322	0.343	**0.234**
	1.5	0.783	1.287	0.359	0.385	**0.272**
	2	1.022	1.704	0.447	0.481	**0.341**
	2.5	1.140	1.916	0.498	0.538	**0.377**
	3	1.371	2.348	0.638	0.706	**0.452**

It is worth noting that the SF model is used for pedestrian trajectory prediction only because it is widely used for pedestrian simulation. The SF model has rarely been used in the simulation of vehicle motions. In addition, the accuracy of pedestrian trajectory predictions is higher than vehicle trajectory predictions (less RMSE) by the other models. Consequently, it is reasonable to conclude that the SF model results are not as good as ST-GCN-Seq2Seq.

6. CASE STUDIES

Case studies were conducted based on the validation set of data.

6.1. Pedestrian-Vehicle Interaction Scenario

Figure 7 shows the trajectory prediction results considering the pedestrian-vehicle interaction. Recall that there is a stop sign on 2nd St., which means that vehicles on 2nd St. have to stop before the stop line. As shown in the Figure 7, the blue solid lines indicate the ground truth future trajectories (3s) for pedestrians and vehicles, and the red dash lines represent the predicted future trajectories (3s) for pedestrians and vehicles.

The subject pedestrian saw a vehicle approaching and jaywalked without stopping in the curb area because if the subject vehicle is too far away, the normal pedestrian decision is to cross immediately.

The subject vehicle was keeping a constant speed. The driver did not have to slow down or stop to avoid a conflict with the subject pedestrian.

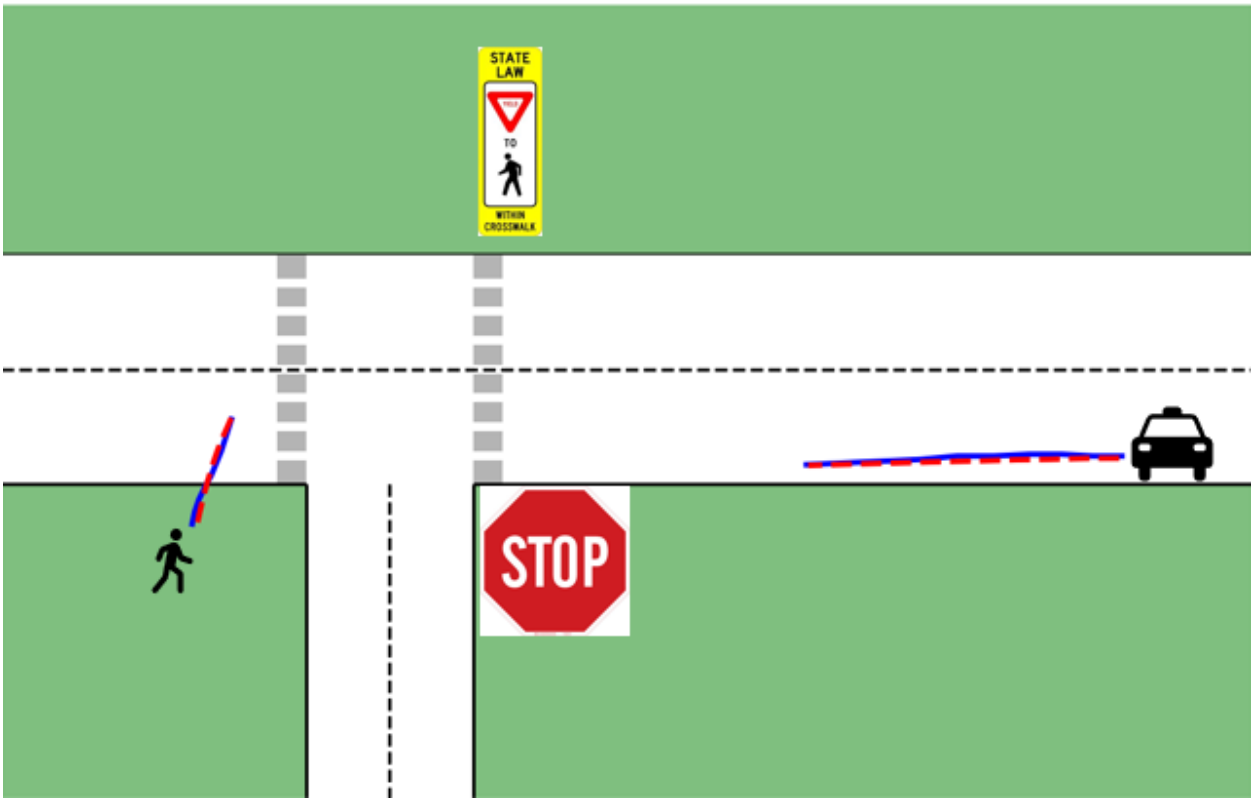


Figure 7 Trajectory Predictions in Pedestrian-Vehicle Interaction Scenario

6.2. Hybrid Interactions Scenario 1

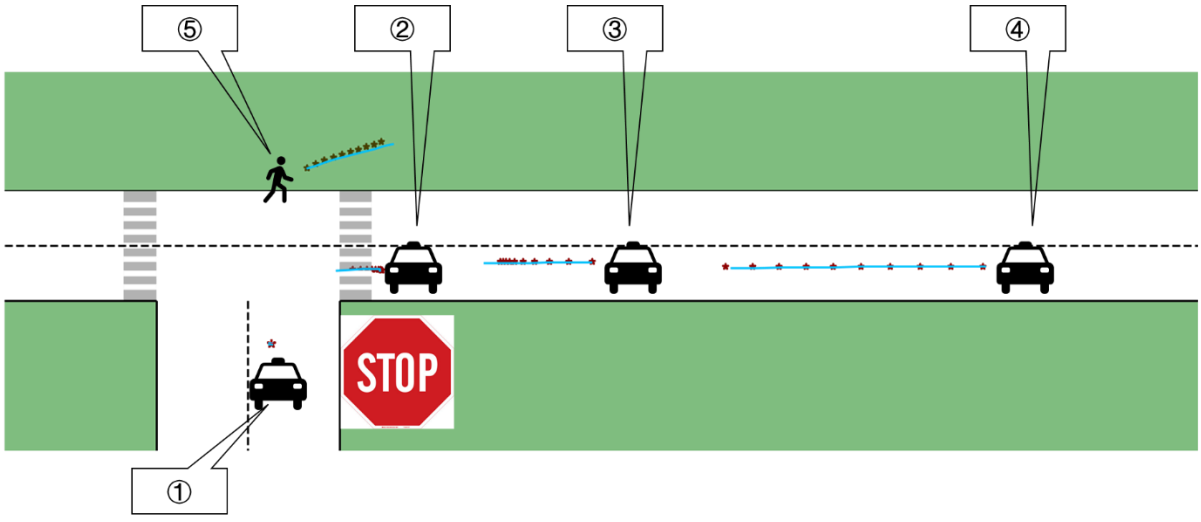
For a more complex traffic environment, Figure 19a indicates the ST-GCN-Seq2Seq can reasonably predict road users' future trajectories. Recall that the blue solid lines indicate the ground truth future trajectories (3s) for road users, and the red scatter plots represent the predicted future trajectories (3s) for road users:

Vehicle No. 1 on 2nd Street had to stop and wait for the approaching vehicle on University St. The red dots in Figure 8b demonstrate that the model successfully predicts the "waiting" behavior of the vehicle on 2nd Street.

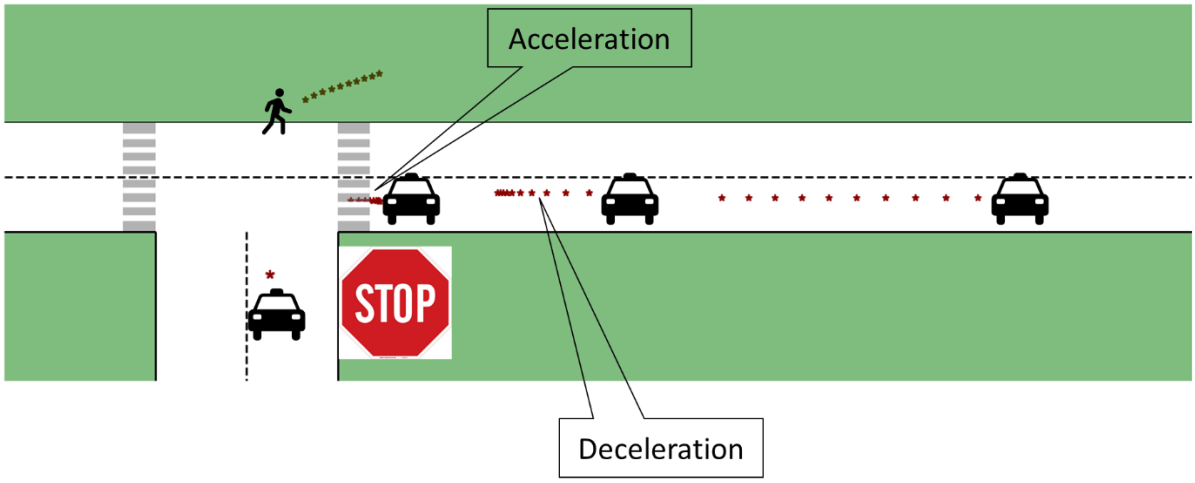
Pedestrian No. 5 just finishes crossing, and vehicle No. 2 on University Street yields to the subject pedestrian. Pedestrian No. 5 leaves the crosswalk. Vehicle No. 2 accelerates to leave the crossing area. The ST-GCN-Seq2Seq reasonably captures the acceleration behavior of Vehicle No. 2 (the gap between consecutive red points is increasing in Figure 8b) but does not precisely represent the magnitude of the acceleration.

Vehicle No. 3 is following Vehicle No. 2. Due to the yielding behavior of Vehicle No. 2, the ST-GCN-Seq2Seq reasonably predicts the "deceleration" behavior of Vehicle No. 3 (the gap between consecutive red points is increasing) to avoid a rear-end collision. See the red dots for Vehicle No. 3 in Figure 8b.

Vehicle No. 4 is following Vehicle No. 3. The ST-GCN-Seq2Seq accurately predicts that Vehicle No.4 will keep a constant speed, because there is a significant distance between Vehicle No. 3 and Vehicle No. 4. See Figure 8b.



a. Future Trajectories and Predicted Trajectories



b. Predicted Trajectories

Figure 8 Trajectory Predictions in Hybrid Interactions Scenario 1

6.3. Hybrid Interactions Scenario 2

Figure 9 indicates that the ST-GCN-Seq2Seq can reasonably predict road users' future trajectories in another complex traffic environment.

Vehicle No. 5 on University Street yields to the subject pedestrians on the crosswalk or in the curb areas. The ST-GCN-Seq2Seq reasonably captures the yielding behavior of Vehicle No. 5 (see red points of Vehicle No. 5 in Figure 9).

The ST-GCN-Seq2Seq precisely predicts the future movements of Pedestrians No. 3 and No. 4 (see red scatter plots of Pedestrians No. 3 and No. 4).

The prediction results for Pedestrian No. 2 are interesting. The Pedestrian No. 2 has conflicts with Pedestrians No. 3 and No. 4. Pedestrian No. 2 actually takes "evasive" behavior to avoid a collision with conflicting pedestrians (see the blue solid line of Pedestrian No. 2 in Figure 9). The ST-GCN-Seq2Seq successfully predicts the "evasive" behavior of Pedestrian No. 2. See the red dots for Pedestrian No. 2 in Figure 9.

However, we should report that the predicted trajectory of Pedestrian No. 1 is reasonable but not accurate. Before stepping into the crosswalk, Pedestrian No. 1 hesitates and "negotiates" with Vehicle No. 5 while in the curb area (see the blue solid line of Pedestrian No. 1 in Figure 9). But the ST-GCN-Seq2Seq directly predicts that Pedestrian No. 1 will step into the crosswalk immediately and accelerate (the gap between consecutive red points is increasing). In this case, complementary information using visual (head, facial expressions, and gaze direction) and map-based cues can be captured in Module 2 to improve the accuracy of predictions.

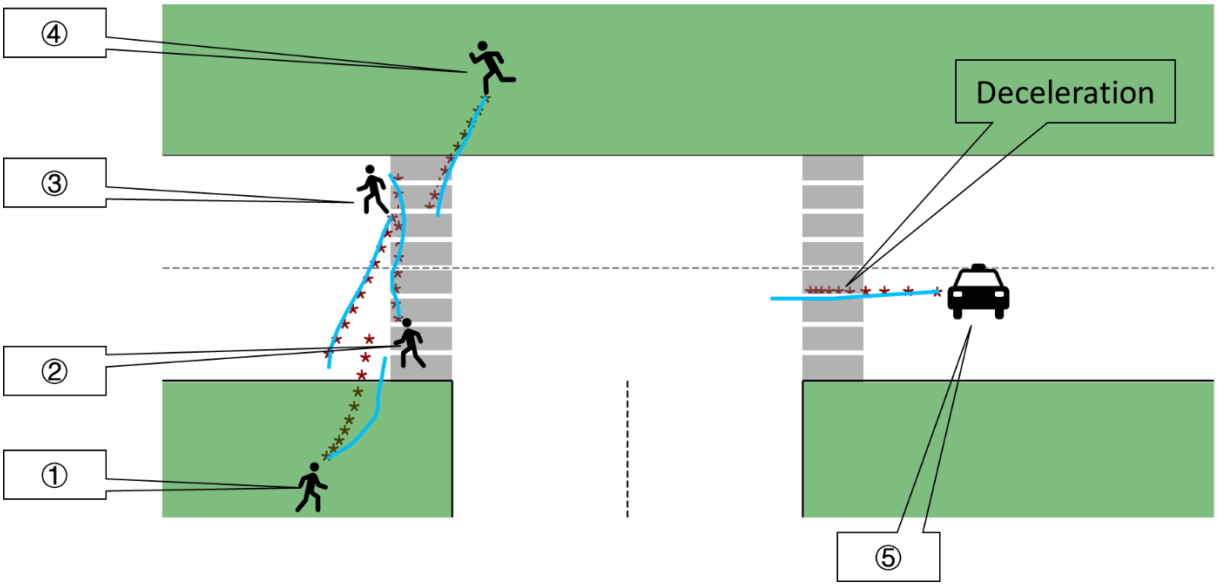


Figure 9 Trajectory Predictions in Hybrid Interactions Scenario 2

7. CONCLUSION

This report presents an observing-tracking-learning framework:

1. Module 1: the behavior of pedestrians and motorists interacting in real street-crossings has been documented by hours of video,
2. Module 2: computer vision-based techniques (Yolo-V3 and DeepSort Algorithms) have been applied to detect road users and track their positions, and Homography has been conducted to transfer the observed coordinates into real coordinates, and
3. Module 3: a spatial temporal graph convolutional network-based sequence to sequence (ST-GCN-Seq2Seq) model has been developed to learn observed road users' movements and predict their future trajectories and interactions between heterogeneous road users.

7.1. Contributions

7.1.1. Open-Sourced Trajectory Dataset

On-site cameras in Module 1 provided bird-eye view of the intersection. Module 2 generates a large-scale spatial-temporal trajectory dataset from over three hours of videos. The current dataset includes 500,000 frames/instances of spatial-temporal positions of heterogeneous road users (pedestrians, cyclists, and vehicles). This includes more than 500 pedestrians and cyclists interacting with more than 600 vehicles.

1. This dataset will be larger than KITTI (Geiger et al., 2013) and ApolloScape (Ma et al., 2019) which have been widely used in trajectory predictions for heterogeneous road users.
2. The bird-eye view provides more interaction scenarios between heterogeneous road users (four PMIs in the Introduction) than BDD100K (Yu et al., 2018) and Argoverse (Chang et al., 2019) collected from naturalistic driving data.
    - Naturalistic driving studies collect recordings of driving information from cameras inside multiple vehicles, which only provides interaction scenarios between the subject motorist and other road users (one-to-many). Our datasets provide many-to-many interaction scenarios.
    - Recordings in naturalistic driving studies can only offer front-view and cannot provide adequate information about surrounding environment.
3. Video recordings from another intersection are being processed by Module 2. After data cleaning, the dataset with more than 1 million frames will be open-sourced. Miovision cameras deployed in intersections of West Lafayette will be a perfect complement to the dataset.

Similar to the applications of KITTI, ApolloScape, BDD100K, and Argoverse, the open-sourced dataset can be used in planning, prediction and simulation tasks.

7.1.2. Prediction Task and Implications

Module 3 offers a hands-on approach ST-GCN-Seq2Seq) to predict movements and behavior of heterogeneous road users. Experiment results indicate that the proposed ST-GCN-Seq2Seq model outperforms the state-of-the-art models in predicting movements of road users near crosswalks. Three case studies have been conducted to demonstrate the robustness of the proposed ST-GCN-Seq2Seq model that accurately predicts the future movements and interactions between heterogeneous road users.

But how do Module 2 and Module 3 help the design of an intelligent tracking system at smart crosswalks? A three-step strategy suggests itself:

1. The appropriate sensor can be deployed to capture the spatial-temporal coordinates of each road user. In this research, an on-site camera is enough. Emerging technologies such as Miovision (https://miovision.com/) will be more helpful.
2. A computer or smartphone application incorporated with Module 2 and Module 3 will perform the detection, tracking, and prediction.
3. Pedestrians and cyclists who download the smartphone application can be notified of real-time future trajectory predictions of surrounding road users (pedestrians, cyclists, and motorists). Vehicle-to-infrastructure technology can also share the prediction results with drivers.

How does Module 3 help achieve a form of "smart interaction" in practice? Trajectory predictions can inform road users of the surrounding environment and other road users' decisions in real time. Several examples are enumerated:

1. In Figure 7, because a low vehicle speed can be inferred from the predicted vehicle trajectory, the pedestrian will cross without hesitation, and the vehicle can keep a constant speed and move across the crosswalk.
2. In Figure 8, if the "deceleration" behavior of Vehicle No. 3 is accurately predicted and the information is sent to Vehicle No. 4, Vehicle No. 4 can prepare the deceleration in advance to avoid an abrupt brake or a full stop.
3. In Figure 9, if the yield behavior of Vehicle No. 5 is accurately predicted and the information is sent to pedestrians, all pedestrians will cross without hesitation or stopping. The predicted yield behavior will help reduce the probability of confusion between pedestrian and motorist (pedestrian and vehicle yield at the same time) and the probability of conflict between pedestrian and motorist (pedestrian crosses and vehicle does not yield).

7.2. Future Directions

As more data are collected and fed into the observing-tracking-learning framework, the model results are expected to be improved. In addition, complementary information using visual (such as facial expressions and gaze estimations) and map-based cues can be captured in Module 2 to improve the accuracy of future motion prediction (Module 3).

8. OUTPUTS, OUTCOMES, AND IMPACTS

8.1. Research Outputs

8.1.1. Synopsis of Project

On-site cameras provided bird-eye view of the intersection. Module 2 generates a large-scale spatial-temporal trajectory dataset from more than three hours of videos. The current dataset includes 500,000 frames/instances of spatial-temporal positions of heterogeneous road users (pedestrians, cyclists, and vehicles). More than 500 pedestrians and cyclists interacting with more than 600 vehicles are included.

Module 3 offers a hands-on approach ST-GCN-Seq2Seq) to predict movements and behavior of heterogeneous road users. Experiment results indicate that the proposed ST-GCN-Seq2Seq model outperforms the state-of-the-art models in predicting movements of road users near crosswalks. Three case studies have been conducted to demonstrate the robustness of the proposed ST-GCN-Seq2Seq model that accurately predicts the future movements and interactions between heterogeneous road users.

8.1.2. List of Publications

Zhang, Y., Fricker, J. (2022). "Forecasting the Motion and Behavior of Heterogenous Road Users at Crosswalks: A Spatial-Temporal Graph-Based LSTM Approach". Under Review by Transportation Research Part C.

Zhang, Y., Fricker, J. (2022). "CrosswalkTrajectory: A Large-scale Spatial-Temporal Trajectory Dataset for Heterogeneous Road Users Behavior Prediction". Pre-print. URL: https://github.com/YZhang-Genghis/XwalkTrajectory.

8.1.3. List of Presentations

Zhang, Y., & Fricker, J. (2022). Making Crosswalks Smarter: Using Sensors and Learning Algorithms to Safeguard Heterogenous Road Users. In 2022 Global Symposium on Connected and Automated Vehicles and Infrastructure, April 14, 2022, Ann Arbor, Michigan.

8.1.4. List of Outcomes and Highlights

This section will emphasize list of outcomes and highlights:

- CV algorithms have been applied to convert the video recordings into a large-scale spatial-temporal trajectory dataset including 800 pedestrians and cyclists interacting with more than 500 vehicles.
- Utilizing the trajectory dataset, a spatial-temporal graph convolutional network-based sequence to sequence (ST-GCN-Seq2Seq) algorithm has been developed to forecast heterogeneous road users' trajectories and behavior in real time.

- Combining CV and ST-GCN-Seq2Seq algorithms can help both design an intelligent tracking system and achieve a form of "smart" interaction at semi-controlled crosswalks for heterogeneous road users.

8.1.5. List of Impacts

But how does this research help the design of an intelligent tracking system at smart crosswalks? A three-step strategy suggests itself:

1. The appropriate sensor can be deployed to capture the spatial-temporal coordinates of each road user. In this research, an on-site camera is enough. Emerging technologies such as Miovision (https://miovision.com/) will be more helpful.
2. A computer or smartphone application incorporated with Module 2 and Module 3 will perform the detection, tracking, and prediction.
3. Pedestrians and cyclists who download the smartphone application can be notified of real-time future trajectory predictions of surrounding road users (pedestrians, cyclists, and motorists). Vehicle-to-infrastructure technology can also share the prediction results with drivers.

This study improves the operation and safety of semi-controlled crosswalks by developing a database and identifying factors that affect pedestrian and motorist behavior.

1. This information will be used to test the impact of new technologies on crosswalk safety and performance.
2. A coupling project with INDOT is a perfect complement to this study, in that it offers opportunities to apply a variety of designs and control methods to other types of crossing locations.

LIST OF REFERENCES

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE conference on computer vision and pattern recognition, 961-971.

Chang, M. F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., ... & Hays, J. (2019). Argoverse: 3d tracking and forecasting with rich maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8748-8757.

Deo, N., & Trivedi, M. M. (2018). Convolutional social pooling for vehicle trajectory prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1468-1476.

Fricker, J. D., & Zhang, Y. (2019). Modeling pedestrian and motorist interaction at semi-controlled crosswalks: the effects of a change from one-way to two-way street operation. Transportation Research Record 2673(11), 433-446.

Geiger, A., Lenz, P., Stiller, C., & Unreason, R. (2013). Vision meets robotics: The kitten dataset. The International Journal of Robotics Research, 32(11), 1231-1237.

Helbing, D., & Molnar, P. (1995). Social force model for pedestrian dynamics. Physical review E, 51(5), 4282.

Offer, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, 448-456. PMLR.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.

Ma, Y., Zhu, X., Zhang, S., Yang, R., Wang, W., & Manocha, D. (2019, July). Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In Proceedings of the AAAI Conference on Artificial Intelligence 33(1), 6120-6127.

Rasouli, A., & Tsotsos, J. K. (2019). Autonomous vehicles that interact with pedestrians: A survey of theory and practice. IEEE transactions on intelligent transportation systems, 21(3), 900-918.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2636-2645.

Zhang, Y., Qiao, Y., & Fricker, J. D. (2020). Investigating Pedestrian Waiting Time at Semi-Controlled Crossing Locations: Application of Multi-State Models for Recurrent Events Analysis. *Accident Analysis & Prevention* 137, 105437.

Zhang, Y., & Fricker, J. D. (2020). Multi-State Semi-Markov Modeling of Recurrent Events: Estimating Driver Waiting Time at Semi-Controlled Crosswalks. *Analytic Methods in Accident Research*, 100131.

Zhang, Y., & Fricker, J. D (2021). Investigating temporal variations in pedestrian crossing behavior at semi-controlled crosswalks: A Bayesian multilevel modeling approach. *Transportation Research Part F: Traffic Psychology and Behaviour*, 76, 92-108.

Zhang, Y., & Fricker, J. D. (2021). Incorporating conflict risks in pedestrian-motorist interactions: A game theoretical approach. Accident Analysis & Prevention, 159, 106254.

Zhang, Y. (2022). Making Crosswalks Smarter: Using Sensors and Learning Algorithms to Safeguard Heterogeneous Road Users, Doctoral dissertation, Purdue University, W. Lafayette, IN.