Pedestrian Crossing Intention Prediction at Red-Light Using Pose Estimation

Shile Zhang, Mohamed Abdel-Aty, *Member, IEEE*, Yina Wu, and Ou Zheng

***Abstract*—Pedestrians' red-light crossing can present a threat to traffic safety. Among all the existing work related to pedestrian's red-light crossing, there are few studies using trajectory data in time sequence. This paper uses pose estimation (keypoint detection) to generate pedestrians' variables from CCTV videos. Four machine learning models are used to predict pedestrians' crossing intention at intersections' red-light. The best model achieves an accuracy of 0.920 and AUC value of 0.849, with data from three intersections. Different prediction horizons (up to 4 sec) are used. With longer prediction horizons, the sample size gets smaller, which partially leads to worse model performance. However, the performance with prediction horizon up to 2 sec is still good (AUC value as 0.841). It is found that keypoint variables such as the angles between ankle and knee (left side) and elbow and shoulder (right side) are important. This model can be further implemented in the Infrastructure-to-Vehicle (I2V) applications and thus prevent accidents due to pedestrians' red-light crossing by issuing warnings to drivers.**

***Index Terms*—Pedestrian crossing intention, red-light crossing, pose estimation, artificial intelligence (AI).**

I. Introduction

PEDESTRIANS are regarded as the most vulnerable road users. According to the World Health Organization (WHO), 1.35 million fatalities were caused due to road crash annually. Among the total fatalities, 23% were pedestrians' fatalities [1]. In the U.S., the number of pedestrians' death increased by 54% during the 10-year period from 2010 to 2019, from 4,280 deaths in 2010 to 6,590 deaths in 2019 [2]. For pedestrian-related crashes, pedestrian's unexpected crossing behavior such as suddenly walking out from the designated crosswalk/sidewalk can be one of the causal factors, especially the red-light crossings at signalized intersections. Based on NHTSA Fatality Analysis and Reporting System (FARS), pedestrians' red-light crossings can cause hundreds of fatalities annually. And the number of fatalities has been growing in recent years [3]. Numerous studies have been carried out on the prediction of pedestrian's crossing intention, however, there are few studies about pedestrian's crossing intention during the red-light signal phase at signalized intersection, which is a special case that can be more critical.

This paper uses CCTV (closed-circuit television) videos with pose estimation (keypoint detection) technique to extract the key landmarks on pedestrians' bodies. While CCTVs are not the ideal cameras for advanced computer vision application such as pose estimation, substantial benefits would be made with them to improve pedestrian safety. Besides, they are cost-efficient due to the wide coverage. Four machine learning models, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GBM), and Extreme Gradient Boosting (XGBoost) based on the generated variables are used. The dependent variable is divided into three classes, standing, walking (normal) (i.e., starting to cross during pedestrian signal phase), and walking (red-light) (starting to cross during red-light signal phase). The best model achieves an AUC value of 0.849. Different prediction horizons are taken into consideration as well.

Compared with traditional studies, this paper uses pedestrians' trajectory data to predict pedestrian's red-light crossing intention. This is an application of Artificial Intelligence (AI) in transportation safety. With the development of Infrastructure-to-Vehicle (I2V) technologies, the established model can be used to warn drivers of unexpected crossing pedestrians. It can also be used for signal timing optimization at signalized intersections.

II. Literature Review

A. Pedestrians' Crossing Intention Prediction

Pedestrians' crossing intention prediction was typically conducted in the same context with trajectory prediction. Among all the sensors, Wi-Fi and Bluetooth were usually used for indoor localization, while camera and LiDAR were used more in the road environment [4]–[6]. The related studies are summarized in Table I. It can be found that most studies used cameras to predict pedestrians' crossing intention or trajectories [4], [5], [7]. From the perspective of modeling methods, three types of methods were mainly used in the literature, including parametric models such as Kalman Filter (KF) and Gaussian Process Dynamical Models (GPDMs), machine learning models such as SVM, and deep learning models such as long short-term memory (LSTM) [4], [8]–[10]. From the perspective of the predicting objectives, the output data were trajectories or crossing/non-crossing intentions [11]–[14]. It was found that pedestrians could change their motions abruptly, or could stop at any time. Quintero, *et al.* [15] used GPDMs and naïve-Bayes classifiers to predict pedestrians'

Manuscript received December 3, 2020; revised February 15, 2021; accepted April 17, 2021. This work was sponsored by Advanced Transportation and Congestion Management Technologies Deployment (ATCMTD) grant from Federal Highway Administration (FHWA). The authors also acknowledge Florida Department of Transportation (FDOT) and the project manager Mr. Jeremy Dilmore, P.E. All results and opinions are those of the authors only and do not reflect the opinions or positions of the sponsors. The Associate Editor for this article was T.-H. Kim. *(Corresponding author: Shile Zhang.)*

The authors are with the Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816 USA (e-mail: shirleyzhang@knights.ucf.edu).

Digital Object Identifier 10.1109/TITS.2021.3074829

TABLE I
LITERATURE ON PEDESTRIANS' INTENTION PREDICTION

Title	Sensor	Method	Objective/output
Bonnin, et al. [17], 2014	Camera	Context model tree	Crossing intention (crossing/not crossing)
Kooij, et al. [18], 2014	Camera	Neural network	Trajectory
Ferguson, et al. [19], 2015	Lidar	Gaussian process mixture model	Crossing intention (crossing/not crossing), trajectory
Völz, et al. [20], 2015	Lidar	Machine learning	Crossing intention (crossing/not crossing)
Goldhammer, et al. [21], 2015	Camera	Neural network	Trajectory
Quintero, et al. [15], 2015	Camera	Gaussian Process Dynamical Models (GPDM)	Crossing intention (crossing/not crossing), trajectory
Hashimoto, et al. [22], 2015	Camera	Dynamic Bayesian Network (DBN)	Crossing intention (crossing/not crossing)
Bock, et al. [23], 2017	Camera	Neural network	Trajectory
Rehder, et al. [24], 2018	Camera	Neural network (LSTM)	Trajectory, goal prediction
Saleh, et al. [12], 2018	Camera	Neural network (LSTM)	Behavior (bending in/ starting/ crossing/ stopping)
Mínguez, et al. [13], 2019	Camera	Gaussian process dynamical models (GPDMs)	Behavior (walking/ standing/ starting/ stopping)
Abughalieh and Alawneh [25], 2020	Camera	Neural network	Moving direction, distance to vehicle
Goldhammer, et al. [26], 2020	Camera	KF, machine learning	Trajectory/ behavior (waiting/ starting/ moving/ stopping)

trajectories and crossing intentions. However, trajectories of more than four seconds were used to predict the next second. The prediction horizon up to 2.5 sec were regarded as short-term prediction. The research gap was to find a robust way with less previous moving profiles as input [16]. Besides, as a more critical case, pedestrians' crossing intention at red-light signals was not emphasized.

It should be noted that it's usually complicated to define pedestrians' crossing intention. Most of the traditional studies defined pedestrians' crossing intention as binary categories, crossing/not crossing. To better define crossing intentions, some studies classified the pedestrians' intention into several categories such as walking, standing, starting, stopping, etc. [11]–[14], [27]. Hariyono and Jo [11] used observers' ratings to label the levels of pedestrians' intention. In most of the cases, the pedestrians were labeled with certain categories such as crossing(1)/not crossing(0). Other categories between 0 and 1 were caused by some of the pedestrians' behaviors, such as turning heads to watch for vehicles. Rasouli, *et al.* [28] collected a data set labeling pedestrians' behaviors across various countries under different lighting conditions. Most of the behavioral patterns found are the sequences of "standing, looking, and crossing", or "moving, looking, and crossing".

B. Human Pose Estimation

Traditional studies learned pedestrians' trajectories for predicting future states. However, it was found that merely trajectories of pedestrians and vehicles were not sufficient [29], [30]. Body languages such as leg movements or turning of body were indispensable among all the factors used for predicting pedestrians' crossing intention. And there were controversial conclusions about whether pedestrian's gaze or head orientation were important [31]–[33].

The development of pose estimation (keypoint detection) could better help recognize pedestrians' states [34]. The pose estimation technique were used to detect the key points on human body. Pavllo, *et al.* [35] first applied a convolutional neural network on keypoint data generated from video. Luvizon, *et al.* [36] used pose estimation to conduct activity recognition. [37] used videos to recognize drivers that were distracted by phones while driving. Face detection, hand detection, as well as pose estimation of the upper body were used. Moreover, pose estimation offered a robust and effective way to estimate pedestrians' crossing intention Ghori, *et al.* [33] used a long short-term memory (LSTM) model to predict pedestrians' and bicyclists' crossing intention. A Bayesian inference function was used to predict the probabilities of five categories of behaviors (crossing, stopping, starting, etc.). Konrad, *et al.* [38] used a sequence of poses to extract variables such as lengths, angles, rotation rates, and linear accelerations formed by pedestrians' joints. The kinematic variables of pedestrians were found to be reliable and accurate enough compared with an inertial measurement unit (IMU).

C. Pedestrians' Red-Light Crossing Behavior

To investigate pedestrians' red-light crossing intention, behavioral models such as the theory of planned behavior (TPB) model and statistical models were used [22], [39]–[41]. It was found that pedestrians' characteristics, such as age, gender, grouping behavior, pedestrian volume, and safety awareness were significant factors [42], [43]. More emphasis should be placed on integrating pedestrians' characteristics into the analysis. Besides, the pedestrians' red-light crossing intention increased with longer waiting time, especially during the last few seconds before crossing [42], [44].

Compared with traditional traffic lights, countdown displays can significantly improve pedestrians' signal compliance [45]. However, countdown displays are often installed near schools or busy intersections. As the studied locations in this paper are mostly located on major arterials in suburban area, the traditional traffic signals are installed. Pedestrians who cross at red-light have potential conflicts with the high-speed vehicular traffic. Besides, the push-button operation at intersections can help to separate the vehicular traffic and pedestrians, thus improving pedestrian safety [46], [47].

Despite the existing work on modeling of pedestrians' red-light crossing intention, few work used sequential data [48]. With the development of Infrastructure-to-Vehicle (I2V) technology, the prediction of the pedestrians' red-light crossing intention can be integrated with other communication media to warn drivers.

This study presents the prediction of pedestrians' red-light crossing intention. Pose estimation is used to generate pedestrians' variables from videos. Four machine learning models, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GBM), and Extreme Gradient Boosting (XGBoost), are used to predict pedestrians' red-light crossing intentions from 1 sec up to 4 sec ahead. The best model achieves a recall value of 0.757 on the walking (red-light) class and an overall AUC value of 0.849.The model performance is still good when the model is used for predicting pedestrians' red-light crossing intentions 2 sec ahead, with the AUC value as 0.841. This work can be applied in the I2V environment to better warn drivers.

III. Data Collection

The videos used in this study are from three signalized intersections located in Seminole County, Florida. All the videos are collected using CCTV (closed-circuit television) cameras during 8:00-19:00 on five sunny workdays in October and November 2019. All the intersections are four-lane by two-lane intersections to ensure the performance of the pose estimation model. The detailed information is listed in Table II. A total of 150-hour of videos are processed with 182 pedestrians collected as valid samples. The pedestrians' trajectories before crossing (in waiting zone) are extracted. Another data source is ATSPM (Automated Traffic Signal Performance Measures) signal timing data to label pedestrians who cross at the red-light [49].

TABLE II
Locations (Data Collection)

Intersection	Road width (major/minor)	Vehicle volume (daily, major /minor road)	Vehicle approach speed (major road)
US 17-92@3rd St	61 ft/20 ft	24023/9519	28 mph
US 17-92@13th St	62 ft/36 ft	24251/4769	27 mph
SR 46@Park Dr	63 ft/39 ft	9959/5864	32 mph

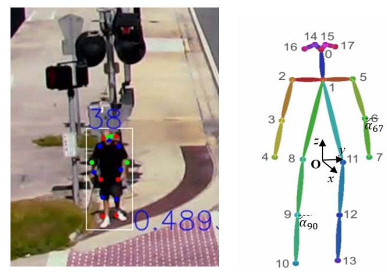


Fig. 1. Pedestrian keypoint detection and transformation.

A. Video Processing

1) Pose Estimation and Object Tracking: The main objective of pose estimation is to derive a representation of the pedestrian's skeleton from each frame of video. Eighteen key points on the human body are generated, as shown in Fig. 1 (a), which mainly include nose, eyes, shoulders, elbows, wrists, hips, knees, ankles, etc. [50], [51]. The object tracking model is used to follow the movement of each road user, and generate the trajectory of the pedestrian [52]. The blue number on the top left corner of the bounding box is the pedestrian's tracking ID number. The number on the bottom right corner is the confidence level of pose estimation model.

As shown in Fig. 1 (b), the coordinates of keypoint are first normalized and transformed to the egocentric coordinates on the pedestrian's body. The origin of this egocentric coordinate system is located at the middle point between left hip and right hip, with three orthogonal axes. Previous studies have found that variables from human bodies such as angles formed by joints were related to their acceleration [38]. Thus, some of the angles generated between joints (such as angle between left wrist and left elbow α_{67}, angle α_{90} between right ankle and right knee, etc.) are extracted as input variables for further modeling. Facial variables such as angle between nose and eye are also extracted as they can potentially reflect head orientation.

2) Perspective Transformation: The middle bottom point on the pedestrian's bounding box (in Fig. 1 (b)) can be used as the reference point to locate the pedestrian. To create a mapping between image coordinates and world coordinates, perspective transformation is conducted. As shown in Equation (1), suppose the point on a video frame is $(u, v, 1)$, and the respective world coordinate is $(X, Y, 1)$, a homograph matrix $\boldsymbol{h}$ is used to convert the coordinates from image plane to world plane. $\boldsymbol{h}$ matrix contains nine values in total, from h_1 to h_9. This is a typical Perspective-n-Point problem, which is to calibrate the camera given n points on the image plane and their corresponding projections on the world plane [53]. A linear least squares method is used [54], [55]. As shown in Equation (2), the image coordinates (u_i, v_i) and the world coordinates (X_i, Y_i) (GPS coordinates in decimal degrees) are used to form matrix A. Each pair of points forms two rows of matrix A. Singular value decomposition (SVD) method is used to derive the solution by minimizing the value of $\|A\boldsymbol{h}\|$ with $h_9 = 1$. After obtaining $\boldsymbol{h}$ matrix and the inverse matrix of h, all generated image coordinates can be transformed to world coordinates. The walking speed is calculated using haversine formula, which is the distance traveled by a pedestrian between two timestamps t_1 and t_2 (Equation (3)).

$$\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \boldsymbol{h} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} = \begin{pmatrix} h_1 & h_4 & h_7 \\ h_2 & h_5 & h_8 \\ h_3 & h_6 & h_9 \end{pmatrix} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} \quad (1)$$

$$A * \boldsymbol{h} = \begin{bmatrix} 0 & 0 & 0 & -X_1 & -Y_1 & 1 & v_1 X_1 & v_1 Y_1 & v_1 \\ X_1 & Y_1 & 1 & 0 & 0 & 0 & -u_1 X_1 & -u_1 Y_1 & -u_1 \\ 0 & 0 & 0 & -X_2 & -Y_2 & 1 & v_2 X_2 & v_2 Y_2 & v_2 \\ X_2 & Y_2 & 1 & 0 & 0 & 0 & -u_2 X_2 & -u_2 Y_2 & -u_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \\ h_9 \end{bmatrix} = 0 \quad (2)$$

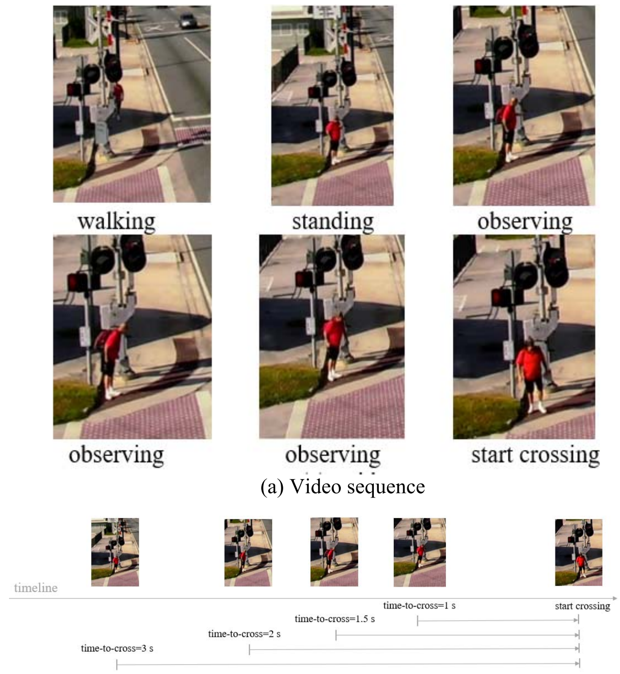
$$Wallking\ speed = \frac{haversine((X_{t1}, Y_{t1}), (X_{t2}, Y_{t2}))}{(t_2 - t_1)} \quad (3)$$

B. Input Variables Overview

Using pose estimation, the angles between some of the key joints are generated. Besides, pedestrians' walking directions, waiting time (time elapsed after the pedestrian reaches the waiting zone), walking speed, and whether pedestrian presses the pushbutton (to activate pedestrian signal phase), are also used as input variables. Some external variables are also included. The hourly temperature data are from National Oceanic Atmospheric Administration (NOAA). Total vehicle volume and right-turn vehicle volume at the current signal cycle, and green time of the vehicle signal phase on pedestrian's conflicting direction are extracted from ATSPM. An overview of all input variables is listed in Table III.

C. Pedestrians' Crossing Intention Labeling

Previous studies found that pedestrians' red-light intention increased when waiting time increased [42], [44]. Thus, the last few moments are an important research target when a pedestrian approaches the road, stops at the curb (waiting zone), and finally starts crossing at red-light. Fig. 2 (a) shows a sequence of video frames. The time-to-cross has been previously used in the related work as the time difference between each frame and the frame when the pedestrian starts crossing [33], [56]. Time-to-cross equals zero means that the pedestrian starts to cross. As the time-to-cross gets closer to zero (shown in Fig. 2 (b)), the pedestrian behaves more



(a) Video sequence

(b) Red-light crossing intention on the timeline

Fig. 2. The red-light crossing intention of a pedestrian over time.

and more impatiently while looking around and watching for approaching traffic. Meanwhile, his crossing intention becomes clearer over time.

On average, the time intervals pedestrians spent on observing the surrounding environment are between 1 sec and 2 sec, which are around 1.32 sec for adults and 1.45 sec for the elderly and children [28]. This time interval is important for the decision-making of the crossing/not-crossing behavior. So, the last 1 sec to 2 sec before crossing can be important for the prediction of pedestrian's red-light crossing behavior.

In this study, the dependent variable is pedestrians' crossing intention. The labeling procedure is shown in Fig. 3. The CCTV videos are first processed using pose estimation and object tracking techniques. The frame rate of CCTV videos is 30 frames per second (fps). The samples in every 0.5 sec are then smoothed and aggregated into one sample to remove noise. The samples in the waiting zones are labeled with three classes, standing, walking normally (for pedestrians who cross during pedestrian signals), and walking at red-light (for pedestrians who cross at red-light). Basically, the first class is from the video frames when the pedestrians stand still, and the other two classes are from the video frames when the pedestrians start to cross (last 1 sec - 2 sec before time-to-cross=0). The labels are validated through manual checks to ensure accuracy.

For prediction purpose, we suppose the driver will yield to pedestrians after capturing the pedestrians' crossing intentions after the reaction time 1 sec [57]. In this case, vehicles travel at 20 mph will have a stopping distance of 40 ft. The dependent variable is shifted 1 sec ahead of time (Fig. 4). This is regarded as the prediction horizon. The generated data set is later split into training set and test set for further modeling.

TABLE III
INPUT VARIABLE OVERVIEW

Description	Mean	Standard deviation	Minimum	Maximum	Unit
Walking direction	3.154	1.659	0.079	6.259	Rad
Walking speed	1.418	1.042	0.026	3.091	Ft/s
Pushing button	0.725	0.446	0.000	1.000	-
Waiting time	10.485	10.001	0.500	55.667	Sec
Angle α (ear & eye, left)	1.362	0.792	7.5e-04	3.135	Rad
Angle α (ear & eye, right)	0.521	0.270	2e-04	3.141	Rad
Angle α (nose & eye, left)	1.374	0.870	3e-04	3.134	Rad
Angle α (nose & eye, right)	0.397	0.214	9.70e-05	3.091	Rad
Angle α (elbow & shoulder, left)	0.583	0.340	1e-03	3.116	Rad
Angle α (elbow & shoulder, right)	0.692	0.458	1e-03	3.140	Rad
Angle α(wrist & elbow, left)	0.649	0.406	4e-04	3.139	Rad
Angle α (wrist & elbow, right)	0.629	0.487	1e-04	3.129	Rad
Angle α (ankle & knee, left)	0.801	0.592	8e-04	3.132	Rad
Angle α (ankle & knee, right)	0.756	0.576	7.49e-05	3.122	Rad
Vehicle volume (current cycle)	72.000	16.102	20.000	120.000	-
Vehicle green time (current cycle)	45.208	21.575	0.007	84.256	Sec
Vehicle counts (right-turn, current cycle)	5.000	3.060	0.000	11.000	-
Temperature	83.847	3.337	68.000	89.000	Fahrenheit

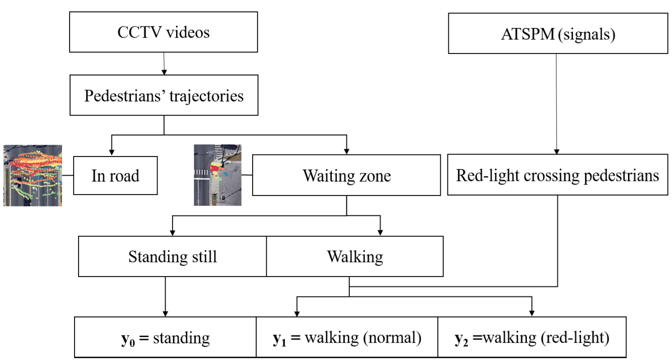


Fig. 3. Procedure of labeling the dependent variable.

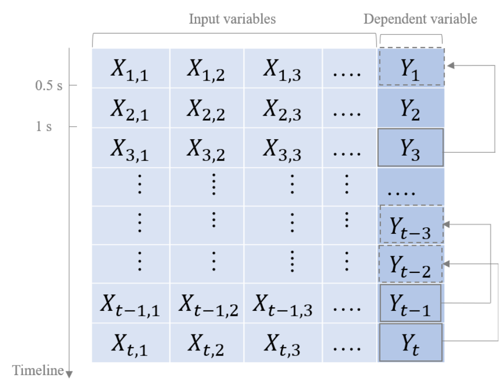


Fig. 4. Shifting the dependent variable (1 s) ahead.

IV. EXPERIMENT AND RESULTS

Four machine learning models, SVM, RF, GBM, and XGBoost are established to predict pedestrians' red-light crossing intentions. The models' hyper-parameters are tuned to reach the best performance.

A. Evaluating Metrics

The evaluating metrics, such as precision, recall, F1 score, accuracy, and AUC are illustrated as below.

(1) Precision: the proportion of correctly classified samples among classified positive samples, as shown in Equation (4).

(2) Recall: or sensitivity, the proportion of correctly classified samples among actual positive samples, as shown in Equation (5).

(3) F1 score: a weighted average of precision and recall, as shown in Equation (6).

(3) Accuracy: the proportion of correctly classified samples among all the samples, as shown in Equation (7).

(4) AUC: area under the ROC curve. The ROC (Receiver Operating Characteristic) curve is used as a comprehensive metric to evaluate the model's performance. This curve plots two parameters, recall and false alarm rate (FAR), at different classification thresholds. The AUC value, which ranges from 0.5 to 1, is the area under the ROC curve. For an imbalanced data set, the AUC value is more reliable than accuracy.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

B. Experiment Results

Among all the 182 pedestrians collected from CCTV data, 61 pedestrians start to cross the road during the red-light

TABLE IV
CLASSIFICATION REPORT FOR FOUR USED MODELS (ON TEST DATA SET)

Model	Class	Precision	Recall	F1 score	Accuracy	AUC
SVM	Walking (red-light)	0.714	0.400	0.513	0.838	0.668
	Macro average	0.797	0.562	0.659		
RF	**Walking (red-light)**	**0.800**	**0.757**	**0.778**	**0.920**	**0.849**
	Macro average	**0.859**	**0.818**	**0.837**		
GBM	Walking (red-light)	0.800	0.673	0.731	0.903	0.818
	Macro average	0.846	0.751	0.796		
XGBT	Walking (red-light)	0.818	0.667	0.735	0.912	0.836
	Macro average	0.862	0.777	0.817		

*Marked in **bold**: the best model; macro average: average value of the metrics over three classes.

signals. With the sampling time window as 0.5 sec, there are 2,375 data samples collected, with the number of samples between the three classes is 1,725: 407: 243. Eighty percent of the samples are used as the training data set, and twenty percent of the samples are used as the test data set. Synthetic Minority Over-Sampling Technique (SMOTE) is used to balance the numbers of samples in three classes in the training data set, to make all three categories balanced [58], [59].

In this study, the dependent variable is divided into three classes, standing, walking (normal), and walking (red-light). The last class is the most critical class. So, the model's performance of this class should be put more emphasis on. Meanwhile, the average value of metrics over three classes, which is usually called macro average value, is also calculated. The modeling results of the four models with prediction horizon as 1 sec are listed in Table IV.

The best model is determined to be RF. The recall value for walking (red-light) class is 0.757, which means the model can recognize 75.7% of the samples (video frames) in which the pedestrians start walking at red-light. Meanwhile, the precision value is 0.800. It also achieves the best performance over three classes compared with the other models. Overall, RF achieved an accuracy of 0.920 and an AUC value of 0.849 over the test data set.

Confusion matrix is usually used to check the overall performance of the model, and identify the specific errors affecting each class. The confusion matrix of the RF model on the test data set is shown in Table V. Most of the samples in each class are classified correctly, denoting the model has a good performance.

The variable importance plot with the top fifteen important variables is shown in Fig. 5. It can be found that walking speed, waiting time, green time (vehicle signal phase), pushing button behavior play important roles for predicting pedestrians' red-light crossing intention. Besides, the angles between knee and ankle (on the left side) also play an important role. Facial variables are also found to be important, such as the angle between left ear and left eye. This may be related

TABLE V
CONFUSION MATRIX FROM RF MODEL (TEST DATA SET)

		Predicted class		
		Standing	Walking (normal)	Walking (red-light)
Actual class	Standing	**362**	7	5
	Walking (normal)	15	**46**	2
	Walking (red-light)	7	2	**28**

*Marked in **bold**: number of correctly classified samples.

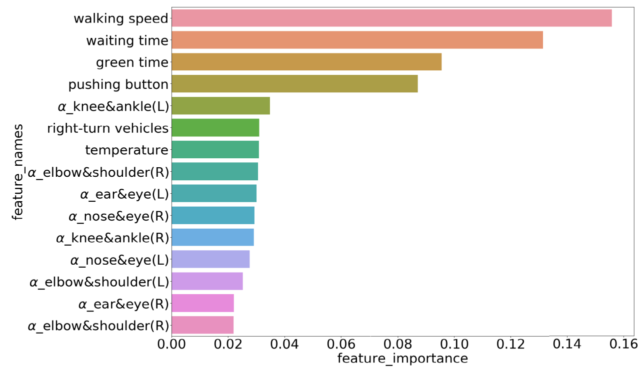


Fig. 5. Variable importance (top 15) from RF model.

TABLE VI
CLASSIFICATION REPORT WITH DIFFERENT VALUES OF PREDICTION HORIZON

Prediction horizon	Class	Precision	Recall	F1 score	Accuracy	AUC
1 sec	Walking (red-light)	0.800	0.757	0.778	0.920	0.849
	Macro average	0.859	0.818	0.837		
2 sec	Walking (red-light)	0.838	0.623	0.715	0.920	0.841
	Macro average	0.813	0.797	0.797		
3 sec	Walking (red-light)	0.750	0.316	0.444	0.889	0.719
	Macro average	0.795	0.619	0.674		
4 sec	Walking (red-light)	0.625	0.417	0.500	0.925	0.688
	Macro average	0.769	0.715	0.735		

*Macro average: average value of the metrics calculated over three classes.

to head orientation. As the data are collected in Florida, where extremely hot weather is usually present at noon, the temperature is also an influencing factor.

Given higher speed limits, there is a need to use longer prediction horizon to build the model. Thus, the other values, 2 sec, 3 sec, and 4 sec, are also taken into consideration. The experiment results are shown in Table VI. When the prediction horizon is 2 sec, the model still maintains an AUC value of 0.841. With the prediction horizon increases to up to 4 sec, the sample size keeps shrinking. So, the model's performance on walking (red-light) class (the most minority class) gets worse, resulting in low values of the evaluating metrics. Besides, the macro average values of evaluating metrics show an overall tendency of decreasing. For an imbalanced data set, the AUC can better reflect model performance than

TABLE VII
CLASSIFICATION REPORT FOR GENERAL CASE (ON TEST DATA SET)

Model	Class	Precision	Recall	F1 score	Accuracy	AUC
SVM	Walking	0.717	0.589	0.647	0.862	0.763
	Macro average	0.805	0.763	0.781		
RF	**Walking**	**0.883**	**0.817**	**0.849**	**0.924**	**0.889**
	Macro average	**0.909**	**0.855**	**0.878**		
GBM	Walking	0.869	0.736	0.797	0.921	0.886
	Macro average	0.900	0.853	0.874		
XGBT	Walking	0.797	0.708	0.750	0.900	0.830
	Macro average	0.861	0.830	0.844		

*Marked in **bold**: the best model; macro average: average value of the metrics calculated over three classes.

accuracy. The AUC shows a steady decreasing tendency with the prediction horizon increases.

C. Comparison Study

For comparison, a more general model is also established. Without including the information of the red-light signals, pedestrians' crossing intentions are divided into standing and walking. As shown in Table VII, the experiment results on the test data set is shown as below. It can be found that the best model achieves the recall value of 0.817 on the walking class, and an AUC value of 0.889.

V. CONCLUSION AND DISCUSSION

This paper uses video data to predict pedestrians' red-light crossing intentions at the signalized intersections with pose estimation and various machine learning models. The highlights of the study mainly include:

(1) The pose estimation technique is used to capture the variables of the pedestrians' bodies, such as angles formed by some of the key joints (wrist, elbows, etc.) and facial landmarks (nose, eyes, and ears) over time.

(2) Upon labeling the dependent variable, pedestrians' red-light crossing intention, the study takes into consideration both mobility (standing/walking) and pedestrians' red-light crossings.

(3) Four machine learning models are used to predict the pedestrians' red-light crossing intentions with multiple prediction horizons. The best model achieves an AUC value of 0.849.

Through the established models, there are a few points to be marked on pedestrians' crossing intention prediction. The walking speed is the top important variable to reflect pedestrians' crossing intentions. The other variables such as button pushing and waiting time may be related to red-light violations. The leg movement denoting by the angle between knee and ankle is an important variable. Compared with the body part, the facial landmarks also reveal early signs of starting walking, which can be related to head orientation.

With respect to different prediction horizons, though the evaluating metrics on walking (red-light) class fluctuate, the model still shows a fairly good performance over all target classes. Overall, the AUC value decreases as prediction horizon increases. When the prediction horizon is 2 sec the model's performance is still good, with the recall value as 0.623 (on walking (red-light) class and AUC value as 0.841.

A more generic model with the dependent variable labeled as standing/walking is also established for comparison. It can be found that the model's performance gets improved, with the AUC value as 0.889. If the signal timing data is not available, then this model can be used instead for warning approaching vehicles, especially the right-turn vehicles.

The limitation of this study is that data from merely three intersections are used with similar geometric design (four-lane by two-lane intersections). But as the CCTV cameras are installed at different locations with different angles, the model successfully deals with the heterogeneity of the generated data set. The study sheds light on the application of pose estimation for studying pedestrian safety. The variables automatically generated from pose estimation can better predict pedestrians' red-light crossing intention compared with only mobility variables such as position and speed [48]. Future work can be conducted to implement the proposed model in field test.

REFERENCES

[1] WHO. *Global Status Report on Road Safety 2018*. Accessed: Jul. 22, 2020. Online]. Available: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

[2] Governors Highway Safety Association. *Pedestrian Traffic Fatalities by State: 2019 Preliminary Data*. Accessed: Jan. 20, 2021. [Online]. Available: https://www.ghsa.org/sites/default/files/2020-02/GHSA-Pedestrian-Spotlight-FINAL-rev2.pdf

[3] Federal Highway Administration. *Intersection Safety: Red Light Running*. Accessed: Jan. 31, 2021. [Online]. Available: https://safety.fhwa.dot.gov/intersection/conventional/signalized/rlr/

[4] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? A study on pedestrian path prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 494–506, Apr. 2014, doi: 10.1109/TITS.2013.2280766.

[5] D. Ellis, E. Sommerlade, and I. Reid, "Modelling pedestrian trajectory patterns with Gaussian processes," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops, ICCV Workshops*, Sep. 2009, pp. 1229–1234, doi: 10.1109/ICCVW.2009.5457470.

[6] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Modeling pedestrians' near-accident events at signalized intersections using gated recurrent unit (GRU)," *Accident Anal. Prevention*, vol. 148, Dec. 2020, Art. no. 105844, doi: 10.1016/j.aap.2020.105844.

[7] D. Ka, D. Lee, S. Kim, and H. Yeo, "Study on the framework of intersection pedestrian collision warning system considering pedestrian characteristics," *Transp. Res. Rec.*, vol. 2673, no. 5, pp. 747–758, 2019, doi: 10.1177/0361198119838519.

[8] M. Goldhammer, M. Gerhard, S. Zernetsch, K. Doll, and U. Brunsmann, "Early prediction of a pedestrian's trajectory at intersections," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2013, pp. 237–242.

[9] A. T. Schulz and R. Stiefelhagen, "A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 173–178.

[10] E. Rehder and H. Kloeden, "Goal-directed pedestrian prediction," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 50–58.

[11] J. Hariyono and K.-H. Jo, "Pedestrian action recognition using motion type classification," in *Proc. IEEE 2nd Int. Conf. Cybern. (CYBCONF)*, Jun. 2015, pp. 129–132.

[12] K. Saleh, M. Hossny, and S. Nahavandi, "Intent prediction of pedestrians via motion trajectories using stacked recurrent neural networks," *IEEE Trans. Intell. Veh.*, vol. 3, no. 4, pp. 414–424, Dec. 2018.

[13] R. Q. Mínguez, I. P. Alonso, D. Fernandez-Llorca, and M. A. Sotelo, "Pedestrian path, pose, and intention prediction through Gaussian process dynamical models and pedestrian activity recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1803–1814, May 2019.

[14] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6262–6271.

[15] R. Quintero, I. Parra, D. F. Llorca, and M. A. Sotelo, "Pedestrian intention and pose prediction through dynamical models and behaviour classification," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 83–88.

[16] D. Ridel, E. Rehder, M. Lauer, C. Stiller, and D. Wolf, "A literature review on the prediction of pedestrian behavior in urban scenarios," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3105–3112.

[17] S. Bonnin, T. H. Weisswange, F. Kummert, and J. Schmuedderich, "Pedestrian crossing prediction using multiple context-based models," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 378–385.

[18] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, "Context-based pedestrian path prediction," in *Computer Vision—ECCV*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 618–633.

[19] S. Ferguson, B. Luders, R. C. Grande, and J. P. How, "Real-time predictive modeling and robust avoidance of pedestrians with uncertain, changing intentions," in *Proc. Sel. Contrib. 11th Int. Workshop Algorithmic Found. Robot.*, H. L. Akin, N. M. Amato, V. Isler, and A. F. van der Stappen Eds. Cham, Switzerland: Springer, 2015, pp. 161–177.

[20] B. Völz, H. Mielenz, G. Agamennoni, and R. Siegwart, "Feature relevance estimation for learning pedestrian behavior at crosswalks," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 854–860, doi: 10.1109/ITSC.2015.144.

[21] M. Goldhammer, S. Köhler, K. Doll, and B. Sick, "Camera based pedestrian path prediction by means of polynomial least-squares approximation and multilayer perceptron neural networks," in *Proc. SAI Intell. Syst. Conf. (IntelliSys)*, Nov. 2015, pp. 390–399, doi: 10.1109/IntelliSys.2015.7361171.

[22] Y. Hashimoto, G. Yanlei, L.-T. Hsu, and K. Shunsuke, "A probabilistic model for the estimation of pedestrian crossing behavior at signalized intersections," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 1520–1526, doi: 10.1109/ITSC.2015.248.

[23] J. Bock, T. Beemelmanns, M. Klösges, and J. Kotte, "Self-learning trajectory prediction with recurrent neural networks at intelligent intersections," in *Proc. 3rd Int. Conf. Vehicle Technol. Intell. Transp. Syst.*, 2017, pp. 346–351.

[24] E. Rehder, F. Wirth, M. Lauer, and C. Stiller, "Pedestrian prediction by planning using deep neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 5903–5908, doi: 10.1109/ICRA.2018.8460203.

[25] K. M. Abughalieh and S. G. Alawneh, "Predicting pedestrian intention to cross the road," *IEEE Access*, vol. 8, pp. 72558–72569, 2020, doi: 10.1109/ACCESS.2020.2987777.

[26] M. Goldhammer, S. Köhler, S. Zernetsch, K. Doll, B. Sick, and K. Dietmayer, "Intentions of vulnerable road users—Detection and forecasting by means of machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 3035–3045, Jul. 2020, doi: 10.1109/TITS.2019.2923319.

[27] F. Schneemann and P. Heinemann, "Context-based detection of pedestrian crossing intention for autonomous driving in urban environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 2243–2248.

[28] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Understanding pedestrian behavior in complex traffic scenes," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 1, pp. 61–70, Mar. 2018, doi: 10.1109/TIV.2017.2788193.

[29] S. Schmidt and B. Färber, "Pedestrians at the kerb—Recognising the action intentions of humans," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 12, no. 4, pp. 300–310, Jul. 2009.

[30] P. Li, M. Abdel-Aty, Q. Cai, and Z. Islam, "A deep learning approach to detect real-time vehicle maneuvers based on smartphone sensors," *IEEE Trans. Intell. Transp. Syst.*, early access, Oct. 28, 2020, doi: 10.1109/TITS.2020.3032055.

[31] A. T. Schulz and R. Stiefelhagen, "Pedestrian intention recognition using latent-dynamic conditional random fields," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2015, pp. 622–627.

[32] Z. Fang and A. M. López, "Intention recognition of pedestrians and cyclists by 2D pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4773–4783, Nov. 2020.

[33] O. Ghori *et al.*, "Learning to forecast pedestrian intention from pose dynamics," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1277–1284.

[34] Z. Fang and A. M. Lopez, "Is the pedestrian going to cross? Answering by 2D pose estimation," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1271–1276.

[35] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7753–7762.

[36] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5137–5146.

[37] J. W. Elings, "Driver handheld cell phone usage detection," M.S. thesis, Dept. Inf. Comput. Sci., Utrecht Univ., Utrecht, The Netherlands, 2018.

[38] S. G. Konrad, M. Shan, F. R. Masson, S. Worrall, and E. Nebot, "Pedestrian dynamic and kinematic information obtained from vision sensors," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1299–1305.

[39] I. Ajzen, "The theory of planned behavior," *Org. Behav. Hum. Decis. Process.*, vol. 50, no. 2, pp. 179–211, 1991, doi: 10.1016/0749-5978(91)90020-T.

[40] D. Evans and P. Norman, "Understanding pedestrians' road crossing decisions: An application of the theory of planned behaviour," *Health Educ. Res.*, vol. 13, no. 4, pp. 481–489, Dec. 1998, doi: 10.1093/her/13.4.481-a.

[41] M. Brosseau, S. Zangenehpour, N. Saunier, and L. Miranda-Moreno, "The impact of waiting time and other factors on dangerous pedestrian crossings and violations at signalized intersections: A case study in Montreal," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 21, pp. 159–172, 2013/11/01/2013, doi: 10.1016/j.trf.2013.09.010.

[42] H. Guo, Z. Gao, X. Yang, and X. Jiang, "Modeling pedestrian violation behavior at signalized crosswalks in China: A hazards-based duration approach," *Traffic Injury Prevention*, vol. 12, no. 1, pp. 96–103, Jan. 2011, doi: 10.1080/15389588.2010.518652.

[43] M. M. Hamed, "Analysis of pedestrians' behavior at pedestrian crossings," *Saf. Sci.*, vol. 38, no. 1, pp. 63–82, 2001, doi: 10.1016/S0925-7535(00)00058-8.

[44] O. Keegan and M. O'Mahony, "Modifying pedestrian behaviour," *Transp. Res. A, Policy Pract.*, vol. 37, no. 10, pp. 889–901, 2003, doi: 10.1016/S0965-8564(03)00061-2.

[45] K. Lipovac, M. Vujanic, B. Maric, and M. Nesic, "The influence of a pedestrian countdown display on pedestrian behavior at signalized pedestrian crossings," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 20, pp. 121–134, Sep. 2013, doi: 10.1016/j.trf.2013.07.002.

[46] V. P. Sisiopiku and D. Akin, "Pedestrian behaviors at and perceptions towards various pedestrian facilities: An examination based on observation and survey data," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 6, no. 4, pp. 249–274, Dec. 2003, doi: 10.1016/j.trf.2003.06.001.

[47] K. Bradbury, J. Stevens, L. N. Boyle, and S. Rutherford, "To go or not to go: Pedestrian behavior at intersections with standard pedestrian call buttons," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2299, no. 1, pp. 174–179, Jan. 2012, doi: 10.3141/2299-19.

[48] S. Zhang, M. Abdel-Aty, J. Yuan, and P. Li, "Prediction of pedestrian crossing intentions at intersections based on long short-term memory recurrent neural network," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2674, no. 4, pp. 57–65, Apr. 2020, doi: 10.1177/0361198120912422.

[49] Florida Department of Transportation. *Automated Traffic Signal Performance Measures*. Accessed: Jul. 23, 2020. [Online]. Available: https://atspm.cflsmartroads.com/ATSPM/

[50] CMU-Perceptual-Computing-Lab. *Openpose*. Accessed: Jul. 27, 2020. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/output.md

[51] Y. W. O. Zheng. *OpenCV C++ Implementation of Drone View Car Tracker*. Accessed: Jul. 23, 2020. [Online]. Available: https://github.com/ozheng1993/HighwayDroneVideoCarTracker

[52] X. Zhou, V. Koltun, and P. Krähenbähl, "Tracking objects as points," 2020, *arXiv:2004.01177*. [Online]. Available: http://arxiv.org/abs/2004.01177

[53] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981, doi: 10.1145/358669.358692.

[54] J. Špaňhel, V. Bartl, R. Juránek, and A. Herout, "Vehicle re-identification and multi-camera tracking in challenging city-scale environment," presented at the Proc. CVPR Workshops, Jan. 2019.

[55] Z. Tang *et al.*, "CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8797–8806.

[56] N. Schneider and D. M. Gavrila, "Pedestrian path prediction with recursive Bayesian filters: A comparative study," in *Proc. German Conf. Pattern Recognit.* Berlin, Germany: Springer, 2013, pp. 174–183.

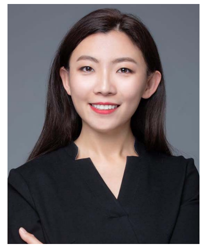
[57] H. Obeid, H. Abkarian, M. Abou-Zeid, and I. Kaysi, "Analyzing driver-pedestrian interaction in a mixed-street environment using a driving simulator," *Accident Anal. Prevention*, vol. 108, pp. 56–65, Nov. 2017, doi: 10.1016/j.aap.2017.08.005.

[58] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[59] P. Li, M. Abdel-Aty, and J. Yuan, "Real-time crash risk prediction on arterials based on LSTM-CNN," *Accident Anal. Prevention*, vol. 135, Feb. 2020, Art. no. 105371, doi: 10.1016/j.aap.2019.105371.

Shile Zhang is currently pursuing the Ph.D. degree with the University of Central Florida (UCF). She has worked on several funded projects in the context of connected vehicle and traffic safety. She has several publications in journals, including *Accident Analysis and Prevention* and *Transportation Research Record*. Her research interests include pedestrian safety, surrogate safety measures (SSMs), and computer vision applications.

Mohamed Abdel-Aty (Member, IEEE) is currently a Pegasus Professor and the Chair of the Department of Civil, Environmental, and Construction Engineering, University of Central Florida (UCF), Orlando, FL, USA. He has managed more than 70 research projects of $20. He has published more than 650 articles, 350 in journals (As of October 2020, Google Scholar citations are 19 800 H-Index 74, Science direct papers downloaded more than 360 000 times). His main expertise and research interests include ITS, traffic safety, simulation, CAV, and active traffic management. He received nine best paper awards from ASCE, TRB, and WCTR.



Yina Wu received the Ph.D. degree in civil engineering (transportation) from the University of Central Florida (UCF). She is currently a Research Associate Professor with the Department of Civil, Environmental, and Construction Engineering, UCF. She is the PI or Co-PI for six projects that are related to traffic safety. She has published 20 peer-reviewed articles related to real-time prediction, connected and autonomous vehicles (CAVs), human/behavioral factors, computer vision technologies, and other aspects of traffic safety in top journals, including *Transportation Research Part C*, *Accident Analysis and Prevention*, IEEE, *Journal of Intelligent Transportation Systems*, and *Transportation Research Record*. Her research interests include driver behavior, surrogate safety measure, computer vision, crash risk analysis, simulation, human factors, and connected vehicles. She serves as a Committee Member for the TRB road weather committee.

Ou Zheng received the master's degree in transportation (smart cities) and computer vision. He has strong engineering professional with his master's degree. He is currently a Computer Vision Research Engineer with the Department of Civil, Environmental, and Construction Engineering, University of Central Florida (UCF). He is also an experienced Algorithm Engineer with a demonstrated working experience from traffic safety-related research. He has been instrumental in many major projects with the UCF SST Team, including the USDOT award-winning safety system, UCF SST A.R.C.I.S computer vision system, and FDOT ATTAIN smartphone applications in connected vehicles. He is skilled in Python, C++, Distributed Systems, Deep Learning, and Computer Vision.