

VRU-CIPI: Crossing Intention Prediction at Intersections for Improving Vulnerable Road Users Safety

Ahmed S. Abdelrahman Mohamed Abdel-Aty Quoc Dai Tran
 University of Central Florida, United States
 {ahmed.abdelrahman, m.aty, quocdai.tran}@ucf.edu

Abstract

Understanding and predicting human behavior in-the-wild, particularly at urban intersections, remains crucial for enhancing interaction safety between road users. Among the most critical behaviors are crossing intentions of Vulnerable Road Users (VRUs), where misinterpretation may result in dangerous conflicts with oncoming vehicles. In this work, we propose the VRU-CIPI framework with a sequential attention-based model designed to predict VRU crossing intentions at intersections. VRU-CIPI employs Gated Recurrent Unit (GRU) to capture temporal dynamics in VRU movements, combined with a multi-head Transformer self-attention mechanism to encode contextual and spatial dependencies critical for predicting crossing direction. Evaluated on UCF-VRU dataset, our proposed achieves state-of-the-art performance with an accuracy of 96.45% and achieving real-time inference speed reaching 33 frames per second. Furthermore, by integrating with Infrastructure-to-Vehicles (I2V) communication, our approach can proactively enhance intersection safety through timely activation of crossing signals and providing early warnings to connected vehicles, ensuring smoother and safer interactions for all road users.

1. Introduction

Understanding and predicting human behavior in dynamic urban environments represents a critical challenge at the intersection of object detection, pose estimation, and activity recognition [35]. This capability is particularly essential for autonomous driving systems and smart traffic management, where accurately interpreting the intentions of VRUs can prevent potentially fatal collisions. Among the most challenging scenarios are urban intersections, where misinterpreting a pedestrian or cyclist's crossing intentions may result in dangerous conflicts with vehicles [52]. In the United States, traffic crashes resulted in 6,516 pedestrian fatalities in 2020, a 3.9% increase from 6272 pedestrian fatalities

in 2019, and an estimated 54,769 pedestrians were injured [29]. In 2020, the number of bicyclist fatalities reached 938, 26% of which occurred at intersections. This represented a 9% increase in bicyclist fatalities, up from 859 in 2019 [23]. Pedestrian fatalities represent 17% of all traffic fatalities, with an increase of 53% in 2018 compared to 2009 [4]. Florida ranks third in terms of pedestrian fatalities per capita in the U.S., and the Orlando-Kissimmee-Sanford area is fifth among the deadliest metropolitan regions, with 656 pedestrian deaths recorded between 2008 and 2017 [2].

Human behavior prediction at intersections represents a complex activity recognition problem that requires sophisticated spatio-temporal modeling of body movements, poses, and contextual factors. Significant advancements have been made in monitoring their activity [1, 33, 37] to improve their safety [19, 39]. Despite these advances, understanding vulnerable road user (VRU) behavior at intersections remains challenging, as VRUs frequently exhibit risky actions such as neglecting to activate pedestrian signals or crossing during prohibited phases [7, 40]. Such behaviors substantially increase the likelihood of dangerous encounters with vehicles, leading to severe conflicts and collisions [5, 9]. VRU violations and frequent disregard for compliance emerge as leading factors contributing to VRU conflicts at intersections. Furthermore, it is imperative not only to enhance VRU safety but also to simultaneously maintain or enhance pedestrian signal performance. Improving signal performance while ensuring VRU safety is driven by the urgent need to balance the critical aspects of urban mobility, the safety of VRUs and the efficiency of vehicular traffic flow. In modern urban centers, intersections are more than mere crossroads; they are pivotal points where the safety of VRUs such as pedestrians and cyclists intersects with the smooth functioning of traffic [3, 28]. Traditional traffic signal systems often struggle with the challenge of maintaining VRU safety at intersections without impeding vehicular flow, leading to a compromise on either safety or efficiency. The aim of this research work is to provide safe and ample crossing experiences for VRUs while reducing unnecessary vehicle delays caused by false calls during the pedestrian

phase. This not only enhances the safety and experience of VRUs but also contributes to a more fluid and less congested traffic system [8, 10, 15, 18, 46].

Many studies have been addressing few of these issues; however, there are three main research gaps in these studies, which are intended to be addressed in this work, as follows:

- At a typical intersection, there are multiple waiting areas that each serve two crosswalks; however, most of the current studies focus only on deciding whether the VRU will cross or not, without specifying which crosswalk it is intended to go through [18, 47].
- The lack of real-time processing in many of the proposed systems that address predicting VRU crossing intention [45, 47, 48]. Achieving real-time processing stands as a critical element in the development of a practical and intelligent system suitable for deployment at intersections.
- Most of the datasets used are typically addresses pedestrians only missing many other VRU types such as cyclists and e-scooter riders and misses nighttime instances and raining weather conditions.

In this work, we address these limitations by introducing VRU-CIPI, a novel real-time framework that leverages multiple input features including pose estimation to understand their behavior combined with temporal modeling for precise crossing direction prediction. Our approach employs GRU to capture the temporal dynamics of VRU movements, integrated with a multi-head self-attention Transformer mechanism to model the spatial relationships between pose keypoints and contextual information. Deploying VRU-CIPI at intersections can aid in proactively activate corresponding pedestrian phases based on the predicted crossing directions improving road users safety.

2. Related Work

2.1. Crossing Intention Prediction from Vehicle Camera View

Pedestrian crossing intention prediction is crucial for safe navigation of autonomous vehicles, prompting extensive research using diverse deep learning architectures. Current models primarily fall into CNN-based, RNN-based, GCN-based, transformer-based, and fusion-based methods, each uniquely leveraging different modalities, features, and temporal-spatial dynamics. The most widely used datasets for onboard vehicle camera-based pedestrian intention prediction are JAAD [31] and PIE [32]. Early pedestrian intention prediction models primarily relied on only convolutional neural network (CNN) to extract visual features from image data. [30] employed VGG16 and ResNet50 backbones to process pedestrian image crops and global scene images, integrating high-level visual features with pedestrian intent labels. However, CNN-based approaches largely overlooked temporal dependencies, limiting their

ability to model sequential motion patterns necessary for accurate crossing intention prediction. To address temporal modeling, recurrent neural network (RNN), including LSTM and GRU, were introduced. [22] incorporated multi-modal features such as bounding box coordinates, vehicle speed, and pedestrian intention data to enhance prediction performance. [44] extended this by integrating segmentation maps, improving environmental awareness. However, these models faced computational inefficiencies and struggled to capture geometric relationships within pedestrian movement. PIP-Net [6] refined this approach by introducing categorical depth and multi-camera inputs, effectively predicting pedestrian intentions up to four seconds in advance on the PIE dataset. Graph Convolutional Network (GCN) emerged as a compelling alternative, particularly for modeling spatial relationships within pedestrian trajectories. Unlike CNN and RNN, GCN excel in preserving the structural properties of pedestrian movement. [11] pioneered a GCN-based approach that extracted pose-based skeletal features, while [49] introduced a spatial-temporal GCN to capture dynamic motion dependencies. [26] further refined this method with spatial, temporal, and channel attention mechanisms. However, pure GCN-based models often lacked environmental and vehicle motion awareness, prompting [12] to integrate pedestrian image crops, skeleton data, speed, and segmentation maps with an attention mechanism for improved feature fusion.

The rise of Transformers introduced significant improvements in long-range dependency modeling. [54] developed a Transformer framework incorporating pedestrian image crops, skeletal features, global scene images, bounding boxes, and ego-vehicle velocity. By leveraging self-attention and temporal fusion blocks, this model effectively captured dynamic spatiotemporal interactions. [50] addressed AI uncertainty in challenging scenarios by integrating bounding box features with deep evidential learning. Hybrid fusion-based models have demonstrated superior predictive performance by integrating multiple feature extraction techniques [16]. [43] combined GCN and RNN, utilizing GRU for processing bounding boxes and vehicle speed while applying GCN to skeleton data. CIPF (Crossing Intention Prediction using Feature Fusion Modules) [17] introduced three specialized fusion modules, observational, contextual, and convolutional, leveraging RNN layers and attention mechanisms to process eight diverse pedestrian and vehicle input features. This model set a new benchmark on the PIE dataset, showcasing significant improvements in predictive accuracy. Complementarily, the Dual-STGAT architecture [24] addresses real-time considerations and efficient feature fusion. Its Pedestrian Module adopts a spatio-temporal graph attention network for movement cues, while a Scene Module aggregates visual, semantic, and motion inputs, employing self-attention and

Efficient Channel Attention (ECA) to reduce redundancy. Evaluations on the PIE dataset confirm its capacity for high accuracy with minimal latency, highlighting the practicality of graph-based fusion strategies for real-world autonomous driving.

2.2. Crossing Intention Prediction at Intersections

Predicting pedestrian crossing intentions at intersections is essential for proactively preventing accidents and improving overall traffic safety. Recent research has primarily focused on off-board sensing methods due to their broader field of view and stable observation capabilities, offering significant advantages over vehicle-mounted (on-board) systems. Off-board systems typically employ stationary cameras, such as CCTV cameras to monitor pedestrian movements widely [1]. These approaches leverage computer vision and machine learning techniques to analyze various visual and motion-related features, including pedestrian posture, gait, head orientation, trajectory data, and interactions with vehicles and infrastructure [47, 51, 53]. Several studies have demonstrated the efficacy of video-based analysis combined with RNN model for crossing prediction tasks. For instance, [47] employed stationary surveillance cameras and applied computer vision tracking in combination with an Long Short-Term Memory (LSTM) neural network, achieving better accuracy in predicting pedestrian intentions at signalized crosswalks. Similarly, [42] integrated pedestrian-vehicle interaction data from cameras and LiDAR sensors at unsignalized intersections. They extracted key temporal and spatial features, such as pedestrian speed, vehicle distance, and time-to-collision, to train an LSTM-based model, which yielded high predictive accuracy. To enhance predictive performance further, [53] introduced a cross-stacked GRU network designed explicitly for off-board surveillance video. Their model fused multiple contextual inputs, including pedestrian posture, local environmental cues, and global contextual information, addressing limitations inherent to vehicle-mounted camera systems and providing robust predictions of pedestrian crossing behavior.

Beyond RNN-based methods, others have explored probabilistic and graphical models to capture the complexity of pedestrian crossing dynamics. [51] implemented a roadside LiDAR-based predictive system, utilizing a probabilistic Naïve Bayes variant to forecast pedestrian crossing intentions at unsignalized intersections and mid-block crossings. Their approach provided reliable predictions within a critical 0.5–3-second window, proving particularly effective in proactive pedestrian safety interventions. [55] utilized a Hidden Markov Model, integrating social interaction forces with high-dimensional pedestrian features such as posture, speed, and spatial relationships. This approach achieved higher prediction accuracy compared to conven-

tional classifiers like Bi-LSTM, SVM, and Random Forest. Graph-based modeling techniques have also emerged as promising methods for pedestrian intention prediction. [52] proposed a graph-centric learning framework leveraging surveillance video data. Their model involved constructing a pedestrian-centric environment graph consisting of appearance and motion sub-graphs, processed through GCN-based model. Combined with pedestrian-specific trajectory and skeletal features analyzed via an LSTM decoder, this integrated approach effectively captured spatiotemporal relationships, achieving state-of-the-art performance in early pedestrian crossing prediction tasks. [48] utilized pose estimation model to extract more information about pedestrian body movements by analyzing CCTV-derived trajectories and pose estimation (keypoint detection). Machine learning models to forecast crossing intentions effectively, highlighting the optimal prediction performance within a horizon of 2 seconds. To support diverse pedestrian safety needs, [45] proposed the Vision Enhanced Nonmotorized Users Services (VENUS), an intelligent infrastructure node integrating real-time sensing, computer vision, and edge artificial intelligence. VENUS provides directional crossing requests and mobility information for pedestrians and cyclists, including persons with disabilities, thereby proactively addressing safety and accessibility simultaneously.

3. Methodology

3.1. VRU Crossing Prediction and Crossing Monitoring Framework

The proposed framework for monitoring and predicting the crossing intentions of VRUs consists of four main stages as visualized in Figure 1 starting from the input video stream from cameras to crossing prediction of VRUs at intersections as detailed below.

3.1.1. First Stage | ZoI Declaration

Frames are extracted sequentially from the streamed video, and specific zones of interest (ZoI) are defined. These zones include the waiting area, the start-crossing zone, and the crossing zone. The start-crossing zone is introduced because some VRUs begin moving toward the crosswalk before the pedestrian signal is activated. However, their presence in this zone indicates a high likelihood of crossing, regardless of the traffic signal status. By tracking VRU trajectories, their positions within these defined zones can be identified [20]. A key feature extracted for each VRU is body and face pose, which is determined using a pose estimation model. However, some cameras are positioned far from the waiting area, reducing the accuracy of pose estimation. To address this issue, the waiting area region is cropped from the full image and then processed using the pose estimation model to extract VRU body keypoints.

VRU Crossing Intention Prediction at Intersection (VRU-CIPI) Framework

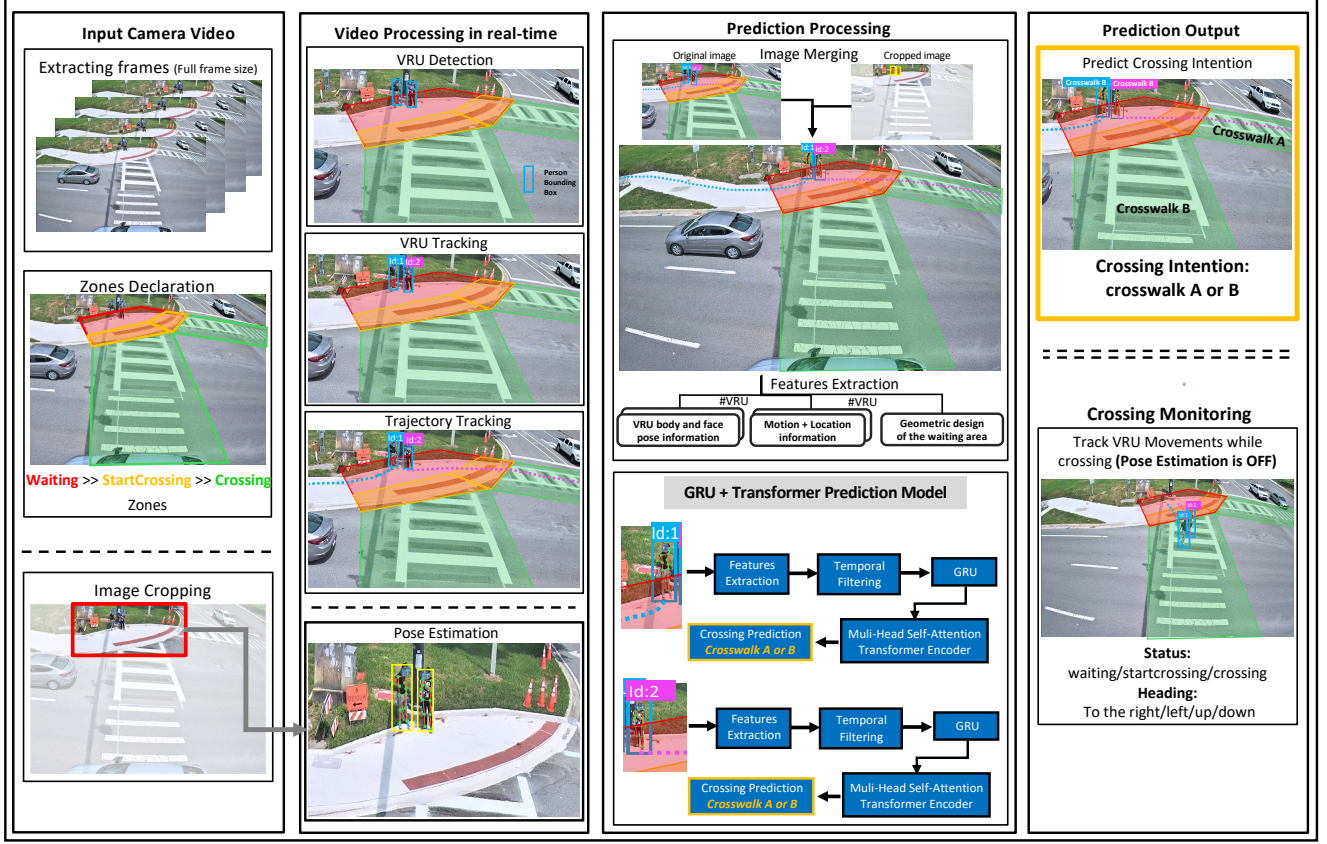


Figure 1. VRU-CIPI framework for VRU monitoring and crossing prediction at intersections. It integrates detection, tracking, and pose estimation models to extract individual features while considering the geometric design of the waiting area for accurate prediction.

3.1.2. Second Stage | Video Processing and Feature Extraction

The primary goal is to accurately predict the crossing direction of VRUs by detecting, tracking, and estimating their poses at intersections. YOLOv8m and YOLOv8m-pose [21] are employed for both VRU detection and pose estimation, respectively. VRU movements are tracked frame-by-frame using OC-SORT [13], enabling the calculation of their walking speed and heading direction using positional information extracted across multiple frames. Additionally, pose estimation identifies keypoints of VRUs and calculates angles and distances between body joints, including the body pose angle, face pose angle, and the distance between the shoulder points. Since the geometric features of waiting areas remain constant within each area, these spatial attributes complement the motion and pose features. Distinctly separated waiting areas for each crosswalk simplify crossing predictions, whereas small waiting areas serving multiple crosswalks complicate predictions because a VRU's position alone does not clearly indicate their intended direction. Consequently, accurate pose esti-

mation features become critical, especially in constrained waiting areas. The extracted VRU features, walking speed, heading direction, and pose angles are integrated to robustly determine crossing intentions, with particular emphasis on pose estimation in compact waiting zones.

3.1.3. Third Stage | Crossing Prediction

Since the detection and tracking models are applied to the full image while the pose estimation model operates on the cropped image, merging these outputs is necessary to assign all extracted features to the correct person ID. The two images are combined, and the pose estimation data is linked to the original person ID by matching locations in the merged image. Once merged, motion and pose estimation features are assigned to each person ID. Features are extracted every 10 frames, corresponding to a 0.5-second time interval. The input window size is 2.5 seconds (50 frames), with a step size of 0.5 seconds (10 frames) for accurate predictions as analyzed by [38]. To ensure smooth readings, features extracted from every 10-frame segment are subjected to temporal filtering through averaging. The filtered features are then grouped into sequences of five and passed to the VRU-

CIPI model. The VRU-CIPI model is based on GRU with multi-head self-attention Transformer, to predict the crossing behavior of each VRU.

3.1.4. Fourth Stage | Crossing Monitoring

The final stage outputs the crossing predictions for each VRU while continuously monitoring their movements. This proactive method enables the activation of the appropriate signal phase without depending on potentially inaccurate push-button activations. It also helps reduce violations by VRUs who cross without pressing the push button. Once a VRU enters the crosswalk zone, detection and tracking continue to monitor their crossing. The pose estimation model is deactivated at this stage, as it is only required for crossing prediction. This approach ensures efficient real-time processing of the entire framework.

3.2. VRU-CIPI: VRU Crossing Intention Prediction at Intersections

Figure 2 shows the VRU-CIPI model architecture for crossing prediction. The model processes features through filtering and concatenation before sequential and contextual analysis is performed using a GRU [14] and a multi-head self-attention transformer encoder [34, 41]. The GRU captures temporal dependencies within the input sequence $X = \{x_1, x_2, \dots, x_T\}$, where $x_t \in \mathbb{R}^d$, by updating its hidden state h_t at each time step. The GRU uses an update gate z_t and a reset gate r_t to regulate the flow of information:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (1)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (2)$$

where W_z, W_r, U_z, U_r are learnable weight matrices, b_z, b_r are bias terms, and σ is the sigmoid activation function. The reset gate r_t modulates the influence of previous states on the candidate hidden state, defined as:

$$\tilde{h}_t = \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \quad (3)$$

The final hidden state h_t is updated by balancing the contributions of the previous hidden state and the candidate hidden state through the update gate:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (4)$$

The sequence of hidden states $H = \{h_1, h_2, \dots, h_T\}$ produced by the GRU is forwarded to a Transformer encoder for contextual feature extraction. The Transformer encoder uses a scaled dot-product attention mechanism to capture relationships across the sequence. For the GRU outputs, the query (Q), key (K), and value (V) matrices are computed as:

$$Q = HW_Q, \quad K = HW_K, \quad V = HW_V \quad (5)$$

where W_Q, W_K, W_V are learnable weights. The attention scores are then calculated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

The encoder of the VRU-CIPI model uses two attention heads to enhance its ability to capture diverse dependencies. The outputs of the two heads are concatenated and transformed using W_O , a projection matrix to combine the heads, as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2)W_O \quad (7)$$

The output from the transformer encoder is passed through fully connected layers and a sigmoid activation function to predict VRU crossing behavior. By linking the Multi-Head Self-Attention outputs to the GRU, the architecture captures both contextual relationships across features and temporal dependencies within the sequence.

4. Experiments

4.1. UCF-VRU Dataset

This research focused on two major four-leg intersections in Florida, United States, with significant VRUs activity to develop a comprehensive dataset. These locations were chosen due to their proximity to prominent shopping areas and educational institutions, ensuring representation of a diverse spectrum of VRU categories. At each intersection, the study aimed to identify optimal camera positions to capture VRU movements at waiting areas and crosswalks comprehensively. Major intersections with a significant number of VRUs were specifically selected to evaluate the reliability of the crossing prediction model and its potential scalability to other locations. A total of 270 hours of video footage was collected using eight cameras (four per intersection), capturing diverse VRU activities under varying conditions, including daytime and nighttime, as well as clear and rainy weather. This dataset includes 962 VRUs categorized into pedestrians, non-motorized (Non-Mot) VRUs such as cyclists and scooter riders, and electric mobility (E-Mob) VRUs such as e-scooters, and e-wheelchairs. Nighttime data was particularly crucial for evaluating the model's performance during periods of higher pedestrian accident risks [27, 36]. Table 1 provides a summary of the data collection, highlighting variations across different times of the day and weather conditions.

4.2. Evaluation Metrics

To assess and compare the performance of the proposed model in predicting whether a VRU intends to cross at

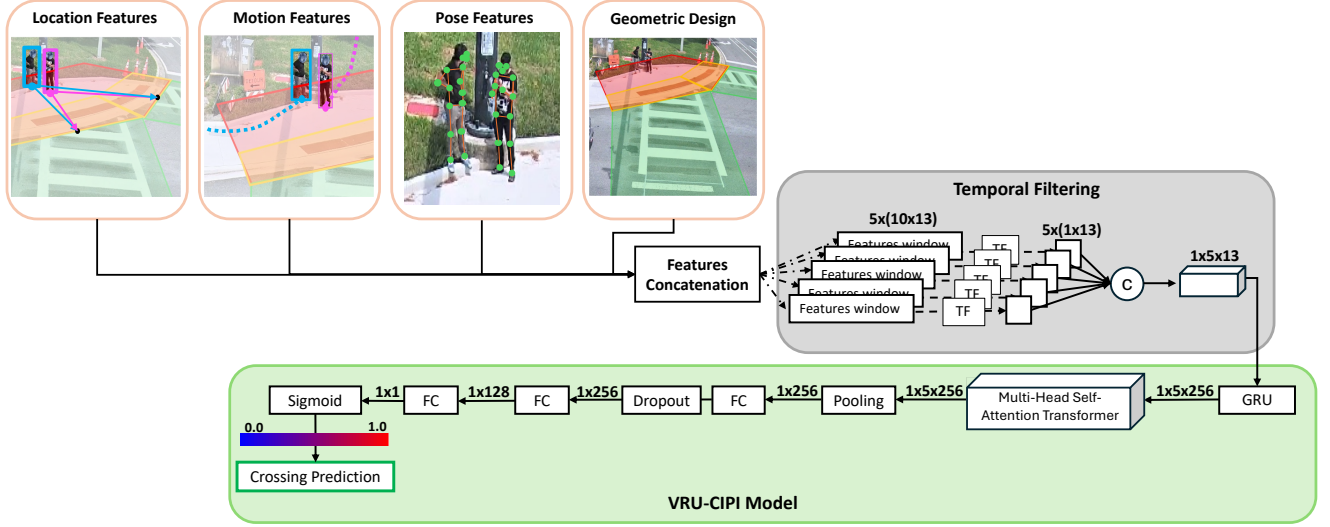


Figure 2. **VRU-CIPI Model Architecture for Crossing Intention Prediction.** Features are filtered, concatenated, and then passed to GRU to capture temporal dependencies, while multi-head self-attention transformer encodes contextual relationships. The final prediction is made through fully connected layers with a sigmoid activation function. TF: temporal filtering. C: concatenation. FC: Fully Connected Layer.

Table 1. **Summary of UCF-VRU dataset.** It includes three primary VRU categories and under various conditions, including nighttime and rainy weather scenarios.

	Ped	Non-Mot	E-Mob	Day	Night	Clear	Rain
Inter. 1	332	156	70	497	59	511	45
Inter. 2	241	144	19	346	60	350	56
Total	573	300	89	845	119	861	101
	59.6%	31.2%	9.2%	87.6%	12.4%	89.5%	10.5%

Crosswalk A or Crosswalk B, we utilized five widely adopted metrics in the pedestrian crossing intention prediction literature: accuracy (ACC), precision, recall, and F1-score. Accuracy measures how accurately the model predicts the binary classification problem of crossing intention, defined as follows,

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \quad (8)$$

where TP represents the quantity of true positive samples, TN represents the quantity of true negative samples, FP represents the quantity of false positive samples, and FN represents the quantity of false negative samples, respectively.

Precision measures the proportion of correct positive predictions among all positive predictions, defined as follows,

$$Precision = \frac{TP}{TP + FP}. \quad (9)$$

Recall measures the proportion of correct positive pre-

dictions among all actual positive cases, defined as follows,

$$Recall = \frac{TP}{TP + FN}. \quad (10)$$

F1-score is the harmonic mean of precision and recall, defined as follows,

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (11)$$

4.3. Experimental Setting

VRU-CIPI model was implemented on Nvidia RTX 3080 Ti and AMD Ryzen threadripper pro 3955wx 16-cores processor with pytorch environment. VRU-CIPI was trained on the UCF-VRU dataset for 250 epochs using the AdamW optimizer with an initial learning rate of 2.5×10^{-4} . The learning rate was dynamically adjusted using a *ReduceLROnPlateau* scheduler, reducing it by half whenever validation loss plateaued for two consecutive epochs. The model architecture consists of a GRU with 256 hidden units and two layers, integrated with a multi-head self-attention transformer encoder for capturing contextual relationships. Dropout was applied at a rate of 0.5 within the GRU, Transformer layers, and fully connected layers to reduce overfitting. Additionally, L2 regularization (weight decay) of 1×10^{-4} was utilized. To ensure stable gradients and improve training convergence, gradient clipping was employed with a maximum gradient norm of 1.0.

4.4. VRU-CIPI Results

The performance of the proposed VRU-CIPI model was comprehensively evaluated using the UCF-VRU dataset,

a diverse dataset with various types of VRUs, including pedestrians, non-motorized VRUs such as cyclists, and electric mobility users like e-bikes and scooters, collected under daytime and nighttime as well as clear and rainy weather conditions. This variety ensures the robustness and generalizability of the model across real-world intersection scenarios. To the best of our knowledge, no prior relevant datasets captured from CCTV camera views include nighttime and rainy conditions. Table 2 compares the performance of VRU-CIPI models with different numbers of attention heads. The proposed two-head VRU-CIPI achieved the highest accuracy (96.45%), precision (96.38%), recall (96.68%), and F1-score (96.53%). The single-head model exhibited slightly lower accuracy (96.27%, a decrease of 0.18%) but higher recall (97.63%, an increase of 0.95%). The four-head model attained an accuracy of 96.28% (0.17% lower than the two-head), with slightly reduced performance in other metrics. These results indicate that the two-head attention model optimally balances complexity and prediction accuracy. The entire framework, from reading video streams to crossing prediction, achieves real-time processing with an inference speed of 33 frames per second (FPS).

Table 2. **Performance of VRU-CIPI on UCF-VRU dataset.** Comparison of prediction results across different numbers of attention heads: single-head (VRU-CIPI-1), two-head (VRU-CIPI-2), and four-head (VRU-CIPI-4)

Model	Accuracy	Precision	Recall	F1-Score
VRU-CIPI-1	0.9627	0.9518	0.9763	0.9639
VRU-CIPI-2	0.9645	0.9638	0.9668	0.9653
VRU-CIPI-4	0.9628	0.9592	0.9682	0.9637

Figure 3 presents qualitative results for VRU crossing intention prediction at intersections. Each case shows the crossing predictions of each VRU while waiting and after crossing, where the number of crossings per direction is visualized in the top left box. **Case a**, presents the crossing prediction of a VRU riding a skateboard where while the VRU based on the location features the VRU is nearer to crosswalk B, however, by analyzing other features, especially pose features, the prediction is correctly identified to be Crosswalk A. In **Case b**, several pedestrians are present with different crossing intentions, where VRU-CIPI is applied for predicting the crossing prediction for each VRU based on each aggregated features. **Case c** shows the crossing prediction of a cyclist at night and while it is raining, where the crossing prediction during such adverse weather conditions is crucial since the visibility of drivers decreases during such conditions. Such predictions enable timely warnings to vehicles, significantly improving VRU safety at intersections and other road junctions.

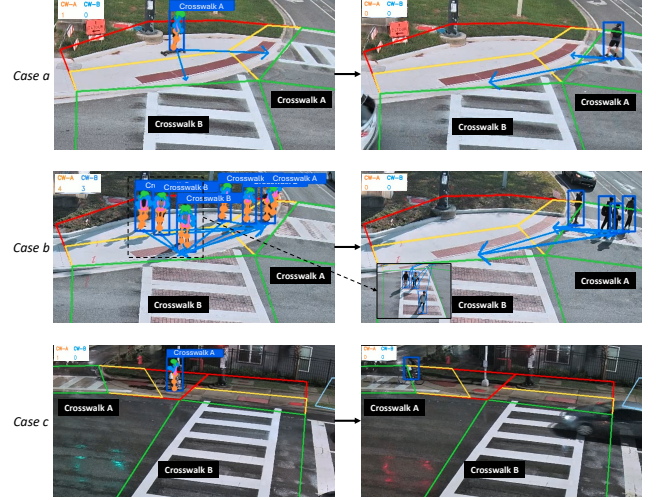


Figure 3. **The qualitative results of VRU-CIPI.** Examples illustrating model predictions for different VRU types and under different weather and lighting conditions. Cases include: (a) a scooter rider during clear daytime conditions; (b) a scenario with seven pedestrians; and (c) a cyclist during nighttime with rainy weather.

When compared to previous research (Table 3), the proposed two-head attention VRU-CIPI model demonstrates superior performance across three defined categories. The comparison table is structured as follows: the first section presents models predicting VRU crossing intentions from a vehicle viewpoint, with the highest accuracy of 85% achieved by PIP-Net [6], where results reported are on PIE dataset and time-to-event is 2 seconds. The second section includes studies employing CCTV camera views to predict cross/no-cross behaviors, where the highest reported accuracy is 95.05% by [42], however, their dataset size is only includes 250 pedestrians which is relatively small. Finally, the third section is also from CCTV camera view but on models predicting crossing intentions for multiple crosswalks, which are particularly important at intersections. In this critical area, the proposed VRU-CIPI model achieves an accuracy of 96.45%, substantially surpassing the previous best model by [45], which reported 90.24%. Furthermore, VRU-CIPI framework maintains real-time inference speed (33 FPS), unlike previous frameworks. [45] achieved only 2 FPS for their entire framework, and while [52] claimed 32 FPS for the prediction model alone, they stated that their complete framework could not sustain real-time processing. [24] reported a 5.01 ms latency for prediction model inference only, whereas our model achieves a significantly lower inference latency of 0.78 ms.

Utilizing CCTV cameras to predict VRU crossing intentions offers distinct advantages in intersection scenarios, providing more accurate, reliable, and comprehensive observations. Furthermore, coupling these predictions

with (Vehicles-to-Everything) V2X and I2V communication technologies enables vehicles to receive precise, real-time alerts about the crossing direction of VRUs, significantly improving intersection safety for all road users. It is important to recognize the limitations when directly comparing our framework to others. Due to differences in datasets and the unavailability of publicly shared models of CCTV view models, the evaluation focuses primarily on overall prediction model performance and dataset diversity. Additionally, our approach relies on fixed-camera setups and known intersection layouts to identify waiting areas, making it less generalizable compared to methods using vehicle-mounted cameras.

Table 3. **Comparison with existing person crossing intention prediction.** For each model, the visual encoder and input features are introduced, and their performance is compared. GAN: Graph Attention Network. GCN: Graph Convolutional Network.

Reference	Year	Model Type	VRU Types	Accuracy
MCIP [16]	2022	GRU+Attention	1	83%
CIPF [17]	2023	GRU+Attention	1	84%
PIP-Net [6]	2024	GRU+Attention	1	85%
Dual-STGAT [24]	2025	GAN+LSTM	1	83%
Zhang et al. [47]	2020	LSTM	1	91.60%
Liang et al. [25]	2023	CNN	2	84.96%
Zhou et al. [52]	2023	GCN+LSTM	1	75.7%
Xiong et al. [42]	2024	LSTM	1	95.05%
VENUS [45]	2022	CNN	6	90.24%
VRU-CIPI (ours)	2025	GRU+Transformer	5	96.45%

4.5. Ablation Study

An ablation study was conducted to evaluate the incremental contribution of four feature sets Location (L), Motion (M), Geometric design (G), and Pose estimation (P) to the VRU-CIPI model performance (Table 4). Starting with location features alone, the model achieved a baseline accuracy of 92.72%. Adding motion features resulted in a modest accuracy increase of 0.26% (92.98%) and improved recall (+1.09%), highlighting better temporal capture of VRU dynamics. Incorporating geometric design features further boosted accuracy by 0.30% (93.28%) and significantly increased precision (+1.16%), enhancing prediction in complex intersections. Most notably, adding pose estimation features led to a substantial accuracy improvement of 3.17%, reaching the highest accuracy (96.45%), precision (96.38%), recall (96.68%), and F1-score (96.53%). This significant gain underscores the critical role of pose estimation features such as body orientation and face looking angle in accurately identifying crossing intentions, particularly when position alone is ambiguous.

Table 4. **Ablation study of the VRU-CIPI model with incremental feature combinations.** Incorporating all features, particularly pose estimation, significantly improves prediction accuracy. Location (L), Motion (M), Geometric design (G), and Pose estimation (P).

Features				Accuracy	Precision	Recall	F1-Score
L	M	G	P				
✓				0.9272	0.9379	0.9179	0.9278
✓	✓			0.9298	0.9333	0.9288	0.9310
✓	✓	✓		0.9328	0.9449	0.9291	0.9369
✓	✓	✓	✓	0.9645	0.9638	0.9668	0.9653

5. Conclusion

Analyzing human behavior to predict their crossing intentions at urban intersections, remains crucial for enhancing interaction safety between road users. In this paper, we introduced VRU-CIPI, a prediction framework based on GRU with multi-head self-attention, designed to predict the crossing intentions of VRUs including pedestrians, cyclists, and other VRU categories at intersections. The proposed model leverages various input features including location, motion trajectories, human pose estimations, and geometric attributes of intersection waiting areas. We gathered the UCF-VRU dataset, capturing diverse VRU behaviors under different conditions, including daytime and nighttime, as well as clear and rainy weather scenarios. VRU-CIPI achieved state-of-the-art crossing prediction accuracy of 96.45%, operating effectively in real-time at 33 FPS for the whole framework. By integrating the proposed method with V2X and I2V communication to connected vehicles, our approach proactively improves intersection safety by promptly activating pedestrian crossing signals while simultaneously supporting autonomous driving through early warnings of the crossing events.

References

- [1] Ahmed S Abdelrahman, Mohamed Abdel-Aty, and Dongdong Wang. Video-to-text pedestrian monitoring (vtpm): Leveraging large language models for privacy-preserve pedestrian activity monitoring at intersections. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 366–375, 2025. 1, 3
- [2] Smart Growth America. *Dangerous by design 2016*. Smart Growth America, 2017. 1
- [3] Roja Ezzati Amini, Kui Yang, and Constantinos Antoniou. Development of a conflict risk evaluation model to assess pedestrian safety in interaction with vehicles. *Accident Analysis & Prevention*, 175:106773, 2022. 1
- [4] Governors Highway Safety Association. Pedestrian traffic fatalities by state: 2019 preliminary data, 2020. 1
- [5] Jarvis Autey, Tarek Sayed, and Mohamed H Zaki. Safety

- evaluation of right-turn smart channels using automated traffic conflict analysis. *Accident Analysis & Prevention*, 45: 120–130, 2012. 1
- [6] Mohsen Azarmi, Mahdi Rezaei, He Wang, and Sebastien Glaser. Pip-net: Pedestrian intention prediction in the wild, 2024. 2, 7, 8
- [7] Kate Bradbury, Jaedi Stevens, Linda Ng Boyle, and Scott Rutherford. To go or not to go: Pedestrian behavior at intersections with standard pedestrian call buttons. *Transportation research record*, 2299(1):174–179, 2012. 1
- [8] Marilyne Brosseau, Sohail Zangenehpour, Nicolas Saunier, and Luis Miranda-Moreno. The impact of waiting time and other factors on dangerous pedestrian crossings and violations at signalized intersections: A case study in montreal. *Transportation research part F: traffic psychology and behaviour*, 21:159–172, 2013. 2
- [9] Thomas Skallebæk Buch and Søren Underlien Jensen. Incidents between straight-ahead cyclists and right-turning motor vehicles at signalised junctions. *Accident Analysis & Prevention*, 105:44–51, 2017. 1
- [10] Marcin Budzynski, Kazimierz Jamroz, and Tomasz Mackun. Pedestrian safety in road traffic in poland. In *IOP Conference Series: Materials Science and Engineering*, page 042064. IOP Publishing, 2017. 2
- [11] Pablo Rodrigo Gantier Cadena, Ming Yang, Yeqiang Qian, and Chunxiang Wang. Pedestrian graph: Pedestrian crossing prediction based on 2d pose estimation and graph convolutional networks. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2000–2005, 2019. 2
- [12] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Pedestrian graph+: A fast pedestrian crossing prediction model based on graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21050–21061, 2022. 2
- [13] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirrodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9686–9696, 2023. 4
- [14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 5
- [15] Hongwei Guo, Wuhong Wang, Weiwei Guo, Xiaobei Jiang, and Heiner Bubb. Reliability analysis of pedestrian safety crossing in urban traffic environment. *Safety Science*, 50(4): 968–973, 2012. 2
- [16] Je-Seok Ham, Kangmin Bae, and Jinyoung Moon. Mcip: Multi-stream network for pedestrian crossing intention prediction. In *European Conference on Computer Vision*, pages 663–679. Springer, 2022. 2, 8
- [17] Je-Seok Ham, Dae Hoe Kim, NamKyo Jung, and Jinyoung Moon. Cipf: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3666–3675, 2023. 2, 8
- [18] Faizanul Haque and Farhan Ahmad Kidwai. Modeling pedestrian behavior at urban signalised intersections using statistical-ann hybrid approach—case study of new delhi. *Case studies on transport policy*, 13:101038, 2023. 2
- [19] Alireza Jafari and Yen-Chen Liu. Pedestrians’ safety using projected time-to-collision to electric scooters. *Nature communications*, 15(1):5701, 2024. 1
- [20] Yuntae Jeon, Dai Quoc Tran, Minsoo Park, and Seunghee Park. Leveraging future trajectory prediction for multi-camera people tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5399–5408, 2023. 3
- [21] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 4
- [22] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1688–1693, 2020. 2
- [23] Motorcycle Helmet Use Laws. Traffic safety facts. 2022. 1
- [24] Jing Lian, Yiyang Luo, Xuecheng Wang, Linhui Li, Ge Guo, Weiwei Ren, and Tao Zhang. Dual-stgat: Dual spatio-temporal graph attention networks with feature fusion for pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 2, 7, 8
- [25] Eric Liang and Mark Stamp. Predicting pedestrian crosswalk behavior using convolutional neural networks. *Traffic injury prevention*, 24(4):338–343, 2023. 8
- [26] Yancheng Ling, Qi Zhang, Xiaoxiong Weng, and Zhenliang Ma. Stma-gcnpedcross: Skeleton based spatial-temporal graph convolution networks with multiple attentions for fast pedestrian crossing intention prediction. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 500–506, 2023. 2
- [27] Paul Lutkevich, Don McLean, and Joseph Cheung. Fhwa lighting handbook. 2012. 5
- [28] Wanjing Ma, Philip Kofi Alimo, Ling Wang, and Mohamed Abdel-Aty. Mapping pedestrian safety studies between 2010 and 2021: A scientometric analysis. *Accident Analysis & Prevention*, 174:106744, 2022. 1
- [29] National Highway Traffic Safety Administration. Pedestrian safety, 2022. 1
- [30] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 206–213, 2017. 2
- [31] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 206–213, 2017. 2
- [32] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6262–6271, 2019. 2

- [33] Amir Rasouli, Mohsen Rohani, and Jun Luo. Bifold and semantic reasoning for pedestrian behavior prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15600–15610, 2021. 1
- [34] Selim Reza, Marta Campos Ferreira, José JM Machado, and João Manuel RS Tavares. A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. *Expert Systems with Applications*, 202:117275, 2022. 5
- [35] Friederike Schneemann and Patrick Heinemann. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2243–2248. IEEE, 2016. 1
- [36] Leni Schwendinger et al. Lighting for pedestrians: Essential considerations for complete streets planning and design. *Institute of Transportation Engineers. ITE Journal*, 93(9):24–29, 2023. 5
- [37] Neha Sharma, Chhavi Dhiman, and Seema Indu. Pedestrian intention prediction for autonomous vehicles: A comprehensive survey. *Neurocomputing*, 508:120–152, 2022. 1
- [38] Neha Sharma, Chhavi Dhiman, and Sreedevi Indu. Predicting pedestrian intentions with multimodal intentformer: A co-learning approach. *Pattern Recognition*, 161:111205, 2025. 4
- [39] Dai Quoc Tran, Yuntae Jeon, Armstrong Aboah, Jinyeong Bak, Minsoo Park, and Seunghye Park. Leveraging semisupervised learning for domain adaptation: Enhancing safety at construction sites through long-tailed object detection. *Journal of Construction Engineering and Management*, 151(1):04024190, 2025. 1
- [40] Ron Van Houten, Ralph Ellis, Jose Sanda, and Jin-Lee Kim. Pedestrian push-button confirmation increases call button usage and compliance. *Transportation research record*, 1982(1):99–103, 2006. 1
- [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5
- [42] Shuai Xiong, Weiwei Meng, Ruiqi Luo, and Jianing Gao. Research on the prediction model of pedestrian crossing intention at intersections based on lstm. In *2024 4th International Conference on Neural Networks, Information and Communication (NNICE)*, pages 982–985. IEEE, 2024. 3, 7, 8
- [43] Biao Yang, Junrui Zhu, Chuan Hu, Zhitao Yu, Hongyu Hu, and Rongrong Ni. Faster pedestrian crossing intention prediction based on efficient fusion of diverse intention influencing factors. *IEEE Transactions on Transportation Electrification*, 10(4):9071–9087, 2024. 2
- [44] Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith A. Redmill, and Ümit Özgüner. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, 7(2):221–230, 2022. 2
- [45] Hao Frank Yang, Yifan Ling, Cole Kopca, Sam Ricord, and Yinhai Wang. Cooperative traffic signal assistance system for non-motorized users and disabilities empowered by computer vision and edge artificial intelligence. *Transportation research part C: emerging technologies*, 145:103896, 2022. 2, 3, 7, 8
- [46] Shile Zhang, Mohamed Abdel-Aty, Qing Cai, Pei Li, and Jorge Ugan. Prediction of pedestrian-vehicle conflicts at signalized intersections based on long short-term memory neural network. *Accident Analysis & Prevention*, 148:105799, 2020. 2
- [47] Shile Zhang, Mohamed Abdel-Aty, Jinghui Yuan, and Pei Li. Prediction of pedestrian crossing intentions at intersections based on long short-term memory recurrent neural network. *Transportation research record*, 2674(4):57–65, 2020. 2, 3, 8
- [48] Shile Zhang, Mohamed Abdel-Aty, Yina Wu, and Ou Zheng. Pedestrian crossing intention prediction at red-light using pose estimation. *IEEE transactions on intelligent transportation systems*, 23(3):2331–2339, 2021. 2, 3
- [49] Xingchen Zhang, Panagiotis Angeloudis, and Yiannis Demiris. St crossingpose: A spatial-temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):20773–20782, 2022. 2
- [50] Zhengming Zhang, Renran Tian, and Zhengming Ding. Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3534–3542, 2023. 2
- [51] Junxuan Zhao, Yinfeng Li, Hao Xu, and Hongchao Liu. Probabilistic prediction of pedestrian crossing intention using roadside lidar data. *IEEE Access*, 7:93781–93790, 2019. 3
- [52] Wei Zhou, Yuqing Liu, Lei Zhao, Sixuan Xu, and Chen Wang. Pedestrian crossing intention prediction from surveillance videos for over-the-horizon safety warning. *IEEE Transactions on Intelligent Transportation Systems*, 25(2):1394–1407, 2023. 1, 3, 7, 8
- [53] Xiao Zhou, Hongyu Ren, Tingting Zhang, Xingang Mou, Yi He, and Ching-Yao Chan. Prediction of pedestrian crossing behavior based on surveillance video. *Sensors*, 22(4):1467, 2022. 3
- [54] Yuchen Zhou, Guang Tan, Rui Zhong, Yaokun Li, and Chao Gou. Pit: Progressive interaction transformer for pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):14213–14225, 2023. 2
- [55] Zhuping Zhou, Yang Liu, Bowen Liu, Molan Ouyang, and Ruiyao Tang. Pedestrian crossing intention prediction model considering social interaction between multi-pedestrians and multi-vehicles. *Transportation research record*, 2678(5):80–101, 2024. 3