UNIVERSITY OF CALIFORNIA SAN DIEGO

Predicting Pedestrian Crossing Intention

A thesis submitted in partial satisfaction of the
requirements for the degree Master of Science

in

Electrical Engineering
(Intelligent Systems, Robotics, and Control)

by

Afnan Alofi

Committee in charge:

Professor Mohan M. Trivedi, Chair
Professor Edward J. Wang
Professor Xiaolong Wang

2023

The thesis of Afnan Alofi is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Professor Mohan M. Trivedi for his support as the chair of my committee. I want to extend my deepest gratitude to my family, whose unwavering support and belief in my abilities have been the cornerstone of my journey. To my parents, whose love and sacrifices have shaped the person I am today, I owe a debt of gratitude that words can scarcely express. Your endless encouragement and wisdom have been my guiding lights in moments of doubt and challenge.

ABSTRACT OF THE THESIS

Predicting Pedestrian Crossing Intention

by

Afnan Alofi

Master of Science in Electrical Engineering
(Intelligent Systems, Robotics, and Control)

University of California San Diego, 2023

Professor Mohan M. Trivedi, Chair

Autonomous vehicles face significant challenges in understanding pedestrian behavior, particularly in urban environments. The system must recognize pedestrians' intentions and anticipate their actions to achieve intelligent driving. This paper focuses on predicting pedestrian crossings, aiming to enable oncoming vehicles to react in a timely manner. We investigate the effectiveness of various input modalities for pedestrian crossing prediction, including human poses, bounding boxes and ego vehicle speed features. We propose a novel lightweight architecture based on LSTM and attention to accurately identifying crossing pedestrians. Our methods evaluated on two widely used public datasets for pedestrian behavior, PIE and JAAD datasets,

and our algorithm achieved a state-of-the-art performance in both datasets.

Chapter 1

Introduction

Autonomous vehicles, or self-driving cars, significantly advance transportation technology. These autonomous vehicles are designed to operate without human intervention, using a combination of sensors, algorithms, and artificial intelligence systems to navigate roads and make driving decisions. Autonomous vehicle development has the potential to revolutionize transportation by improving road safety, reducing congestion, and increasing accessibility. With the ability to perceive and interpret their surroundings, autonomous vehicles can adapt to changing road conditions and interact with other vehicles, pedestrians, and infrastructure.

Having the capacity to comprehend the surrounding surroundings and predict the intentions of other road users is of utmost importance in mitigating traffic fatalities. Autonomous vehicles play a vital role in ensuring pedestrian safety, and preventing collisions with pedestrians is paramount. Current methods for pedestrian collision prevention primarily involve integrating visual pedestrian detectors with Automatic Emergency Braking (AEB) systems, which can trigger warnings and apply brakes as a pedestrian enters the vehicle's path. However, these pedestrian-detection-based systems face limitations, including reduced effectiveness in low-light conditions and when pedestrians are occluded and challenges in handling complex scenarios and adverse weather conditions. These limitations highlight the need for advancements in sensor technology and algorithms to improve pedestrian detection and response capabilities in autonomous vehicles.

Based on the 2020 report by the National Highway Traffic Safety Administration [1], there were 6,516 pedestrian fatalities and 54,769 pedestrian injuries resulting from traffic accidents in the United States. Despite vehicle and road safety advancements over the past two decades, there has been a significant 42% increase in pedestrian fatalities on public roads from 2000 to 2020. Most pedestrian deaths (80%) occurred in urban areas, with 75% happening on open roads rather than at intersections. The report highlighted that the leading cause of pedestrian fatalities (50%) was the failure of drivers to yield the right of way.

In light of the alarming statistics and the need for improved pedestrian safety, this project aims to develop a predictive model to anticipate whether a pedestrian is likely to cross the street. By analyzing various visual and contextual cues, the model will assess the intention of pedestrians and provide early indications of potential crossing behavior. The objective is to leverage machine learning algorithms, such as utilizing a combination of vision and non-vision features, to accurately predict pedestrian intentions, especially scenarios where they might cross the street unexpectedly, such as crossing in the middle of the street instead of using designated crosswalks as shown in figure 1.1(d). Another challenging scenario for predicting pedestrian crossing behavior occurs when pedestrians are standing on a crosswalk but display hesitation in crossing, as depicted in (a) and (b) of Figures1.1. Additionally, predicting their behavior becomes even more complex when they suddenly change their crossing direction, as shown in Figures 1.1(c).

By building this predictive model, we aim to contribute to the prevention of pedestrian accidents and provide valuable insights for developing advanced driver assistance systems (ADAS). The proposed solution can improve road safety by enabling vehicles to anticipate and respond effectively to pedestrian behavior, reducing the risk of collisions and ultimately saving lives.

1.1 contributions

In summary, the contributions of this work are three folds as follows:

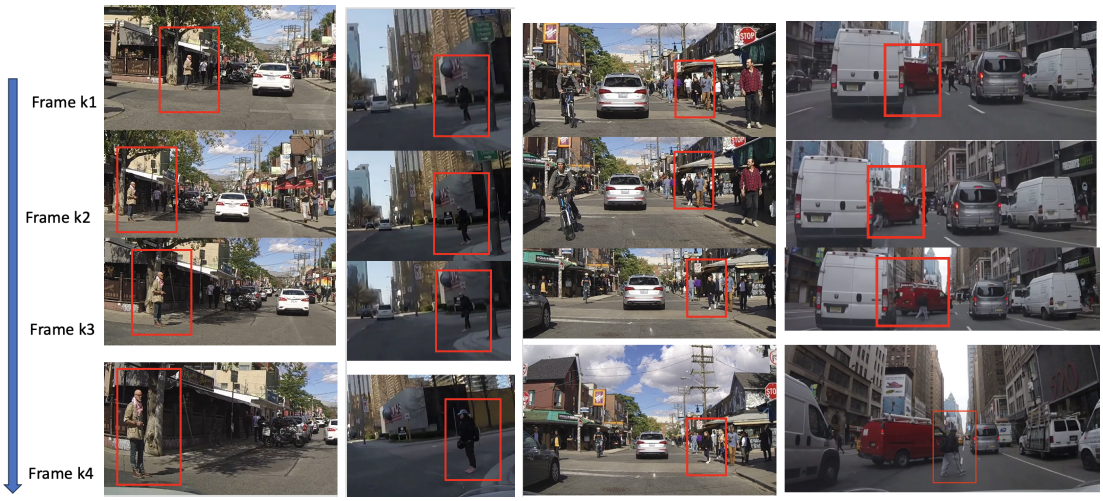

Figure 1.1. (a) Pedestrian have the intention to cross but did not crossing,PIE dataset. (b) Pedestrian have the intention to cross but did not crossing,JAAD dataset. (c) Pedestrian crossing the street unexpectedly,PIE dataset. (d) Pedestrians crossing the street from the middle, JAAD dataset.

1. A new framework is proposed, utilizing non-visual features, to accurately predict a pedestrian's intention to cross the street

2. Through comprehensive ablation studies, different features are systematically evaluated as inputs to the model, aiming to identify the optimal set of features.

3. The effectiveness and efficiency of the proposed method are demonstrated by evaluating its performance on two widely used pedestrian datasets, namely JAAD dataset [2] and PIE dataset [3].

Chapter 2

Related work

The Related Work section provides an overview of prior research in understanding pedestrian behavior and predicting their actions in the context of autonomous vehicles. The investigation into predicting pedestrian crossing actions can be systematically classified into four distinct categories, each delineated by its approach: historical data, utilization of multimodal data models, and application of unimodal data models.

2.1 History

In our previous work, ”Pedestrian Behavior Maps for Safety Advisories: CHAMP Framework and Real-World Data Analysis” [4], we addressed these issues by developing an online, map-based pedestrian detection aggregation system. The system leverages repeated passes of locations to learn common pedestrian zones and overcome challenges like dark lighting or pedestrian occlusion. Through careful data collection and annotation in La Jolla, CA, we demonstrated the system's ability to learn pedestrian zones and generate advisory notices when a vehicle approaches a pedestrian in challenging conditions.

2.2 Multimodal data

In recent years, numerous crossing-action prediction models have employed multimodal data. These models leverage diverse features to comprehensively understand intricate real-world

scenarios, encompassing elements such as the visual appearance of pedestrians and their surroundings, their precise locations, poses, and the speed of the ego-vehicle. For instance, in the paper "Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention" [5], a novel neural network architecture is proposed. The model incorporates CNN modules, RNN modules, attention modules, and a unique feature fusion approach. It combines visual features, such as local and global context, with non-visual features, such as pose keypoints, bounding boxes, and vehicle speed. The paper achieved state-of-the-art performance on the JAAD dataset, demonstrating the effectiveness of their approach.

Another example is found in the paper titled "Early intention prediction of pedestrians using contextual attention-based LSTM" [6]. The proposed model consists of two layers of LSTM and one layer of attention mechanism as its core components. The model first detects and tracks pedestrians and then utilizes visual, contextual, and motion feature extraction to predict their crossing behavior through an attention-based LSTM (CA-LSTM). Visual features are extracted by applying two convolution layers and one average pooling layer to the pedestrian area cropped from the image based on bounding boxes. Similarly, contextual features are extracted, with the only difference being the expansion of the pedestrian's bounding box at a specific scale. Motion features encompass the pedestrian's velocity and walking angle, with the angle range determined by the direction relative to the horizontal axis. The paper demonstrates state-of-the-art performance in predicting future pedestrian crossing behavior, as evaluated on the JAAD dataset.

However, it is essential to highlight that a previous study achieved outstanding performance by utilizing non-visual features. In the research paper titled "Feature Selection and Multi-task Learning for Pedestrian Crossing Prediction" [7], the proposed model incorporated three features: vehicle speed, bounding box, and pose. While the body pose provided in the JAAD and PIE datasets was extracted using the OpenPose method, the study observed limitations in OpenPose's performance on the image data within these datasets. The challenges stemmed from most pedestrians being located at a considerable distance, resulting in small image crops

and out-of-focus image regions, which hindered accurate pose estimation. To overcome this issue, the researchers adopted the HRNet [8] method by training on the BDD100K dataset [9]. This strategic choice substantially enhanced the precision of detected poses within the JAAD and PIE datasets. Consequently, the pedestrian crossing prediction accuracy significantly improved, achieving a state-of-the-art performance level of 91% accuracy.

The paper "Benchmark for Evaluating Pedestrian Action Prediction" [10] has made a notable contribution to the field by introducing a novel evaluation protocol designed explicitly for benchmarking pedestrian action prediction algorithms. The authors recognized the importance of establishing a standardized framework to enable fair comparisons among different prediction models. They conducted thorough evaluations using two publicly available datasets, PIE and JAAD, considering various data properties such as time-to-event, occlusion, and scale. However, they found it challenging to attribute the difficulty of the samples to specific data properties, and they observed inconsistency in model agreement. Furthermore, based on the benchmark results, the authors proposed a groundbreaking hybrid model that combines recurrent and 3D convolutional approaches with temporal and modality attention mechanisms. This novel model achieved state-of-the-art performance on both the PIE and JAAD datasets. The meticulous annotation of ground truth trajectories and action labels in the benchmark dataset allowed for the precise evaluation of prediction models. The benchmark dataset and evaluation framework presented in this paper have gained widespread adoption within the research community, serving as a fundamental resource for evaluating and comparing various pedestrian action prediction algorithms.

The authors of the paper titled "Pedestrian Graph +: A Fast Pedestrian Crossing Prediction Model Based on Graph Convolutional Networks" [11] introduce the Pedestrian Graph + model, an advancement over their earlier Pedestrian Graph model, designed to predict pedestrian crossing actions in urban environments using a Graph Convolutional Network. By incorporating two convolutional modules into the new model, the researchers provide supplementary context information such as cropped images, segmentation maps, and ego-vehicle velocity data to en-

hance the accuracy of predictions. Notably, the Pedestrian Graph + model is more efficient than other state-of-the-art models, offering equivalent accuracy while exhibiting a faster inference time of 6 ms and minimal memory consumption. The model's performance is validated on the Joint Attention in Autonomous Driving (JAAD) and Pedestrian Intention Estimation (PIE) datasets, achieving accuracies of 86% and 89%, respectively.

2.3 Unimodal data

Conversely, within the scope of unimodal data, certain studies have successfully harnessed individual features to yield impressive outcomes. For example, the paper titled "Is attention to bounding boxes all you need for pedestrian action prediction?" [12], where the researchers introduce a framework that employs multiple variations of Transformer models to predict pedestrian street-crossing decisions based on their initiated trajectory dynamics. The study reveals that the framework surpasses previous state-of-the-art results by solely considering bounding boxes as input features. Notably, on the PIE dataset, the framework achieves a prediction accuracy 91% and an F1-score of 83%.

An additional intriguing study, titled "Anticipating Pedestrian Crossing Intentions through Head Gestures: Leveraging Head Pose Estimation," [13]. This research utilizes advanced techniques, including Head Pose Estimation, to predict pedestrians' intentions to cross the road. The methodology employs the YOLOv3 algorithm to detect human heads within groups of pedestrians precisely. Subsequently, the WHENet model is utilized to estimate the head's pose accurately. To finalize the decision-making process, the researchers employ a K-Nearest Neighbor classifier to determine whether a pedestrian will engage in crossing or not. This approach achieves a remarkable accuracy of 97.2% when tested on a subset of the jaad dataset. This subset comprises 18 distinct sample videos featuring pedestrians characterized by unique head gestures.

Table 2.1. Related Work

study	Methodology	Features	PIE			JAAD			Important findings
			acc ↑	***auc*** ↑	***f1*** ↑	***acc*** ↑	***auc*** ↑	***f1*** ↑	
Kotseruba [10] 2021	RNN with attention modules	bounding box Speed,pose local context	0.87	0.86	0.77	0.85	0.86	0.68	Establishes benchmark evaluating pedestrian for action prediction models .
Perdana MI [13] 2021	K-Nearest Neighbor	Head Pose Estimation	-	-	-	0.97	-	-	Using head pose to predicted Pedestrian Crossing Intention but just part of jaad dataset , just 18 videos.
Yang D [5] 2022	RNN modules and attention modules	bounding box, pose, global context	-	-	-	0.83	0.82	0.63	combines visual features with non-visual features
Schörkhuber [7] 2022	RNN-based encoder-decoder	bounding box Speed PoesHRNet	0.91	0.93	0.82	0.90	0.95	0.76	- Multi-task learning improvement over all performance, jointly trains to predict crossing ,trajectory and location. - Pose from HRNet improvement performance.
Cadena PR [11] 2022	Graph Convolution Network and parallel RNN	Speed pose local context	0.89	0.90	0.81	0.86	0.88	0.65	fast pedestrian crossing prediction model based on graph convolutional networks
Achaji L, [12] 2022	Transformer	bounding box	0.91	0.91	0.83	-	-	-	Using solely input feature bounding boxes surpasses previous state-of-the-art result
Lian J [6] 2023	Series LSTM and attention modules	Visual feature Contextual feature pedestrian's velocity and walking angle	-	-	-	0.89	-	0.75	using pedestrian's velocity and walking angle

Chapter 3

Datasets and Features

3.1 Datasets

3.1.1 Pedestrian Intention Estimation (PIE) dataset

The Pedestrian Intention Estimation (PIE) dataset is a valuable and extensive resource for studying pedestrian behavior in traffic. It provides a comprehensive understanding of pedestrians' actions and intentions through over 6 hours of recorded footage and accurate vehicle information. With rich annotations for pedestrians, vehicles, and infrastructure, spanning more than 300,000 labeled video frames, the dataset offers detailed insights into various aspects of pedestrian behavior. The pedestrian counts within the dataset reveal that 519 pedestrians successfully crossed, 894 intended to cross but did not, and 429 showed no intention to cross. These counts provide significant insights for studying pedestrian behavior in diverse traffic situations. [3].

3.1.2 Joint Attention in Autonomous Driving (JAAD) dataset

The JAAD (Joint Attention in Autonomous Driving) dataset is a comprehensive resource for studying pedestrian and driver behaviors in autonomous driving scenarios. It consists of 346 short video clips, providing a richly annotated collection from over 240 hours of driving footage. The dataset includes spatial annotations with bounding boxes for 2,793 pedestrians and

Table 3.1. Comparing the PIE Dataset with the JAAD Dataset

Dataset	**PIE**	**JAAD**
Released year	2019	2017
Total number of frames	909,480	82,032
Video Duration	over 6 hours	over 240 hours
Number of pedestrians with behavior annot.	1842	686
Number of pedestrians who cross the street	519	495
Number of pedestrians who do not cross the street	1323	191
Geographic Scope	Toronto, Canada	North America and Eastern Europe
Ego-vehicle sensor information	yes	No
Nighttime Cases	None	4 Videos
Pedestrian adult and young	1640	574
Pedestrian child	17	16
Pedestrian senior	185	96

behavioral annotations specifying their actions. With a variety of scenarios captured in North America and Europe, the JAAD dataset offers valuable insights into pedestrian behavior during street crossings, including 495 instances of pedestrians crossing and 191 instances of pedestrians not crossing.[2].The comparison between the PIE dataset and the JAAD dataset is presented in Table 3.1.

3.2 Features

3.2.1 Pose:

The Human Body Poses (Pose), sourced from the JAAD and PIE datasets, were estimated using the OpenPose technique. These poses were the subject of investigation in the research paper titled ”Feature Selection and Multi-task Learning for Pedestrian Crossing Prediction” [7]. Initially, the findings indicated limitations in using the OpenPose method to extract body poses from the JAAD and PIE datasets. However, a more robust performance was achieved by integrating the HRNet method into the OpenPose approach. In our methodology, we harnessed the capabilities of HRNet [8] to enhance the accuracy of human pose estimation. The procedure encompassed the cropping of the video frame to align with the respective bounding box of pedestrian annotations. Subsequently, the HRNet was employed to execute the pose estimation, and the resulting coordinates of estimated keypoints were normalized in relation to the dimensions of the video frame. This normalization process confined the values within the range of $[-1, 1] \subset \mathbb{R}$.

3.2.2 Speed of Ego-Vehicle:

The ego vehicle Speed (Speed), it is incorporated as factual data in the PIE and JAAD datasets. While the vehicle speed in the PIE dataset is recorded using multiple sensors and is presented in kilometers per hour (km/h), the JAAD dataset lacks such quantitative measurements. Instead, it employs numerical labels ranging from 0 to 4, signifying stationary, slow movement, rapid movement, deceleration, and acceleration, respectively.

3.2.3 Bounding Box:

Bounding Box Sequences (Box) are utilized to capture the positional and dimensional information of pedestrians. These boxes are defined by their upper-left and lower-right corner coordinates, namely (x1, y1) and (x2, y2).

Chapter 4

Pedestrian Prediction Modules

4.1 Predicting Pedestrian Crossing Intention model (PPCI)

The overall architecture is shown in Figure 4.1. It consists of RNN modules, and a novel way of fusing different features.

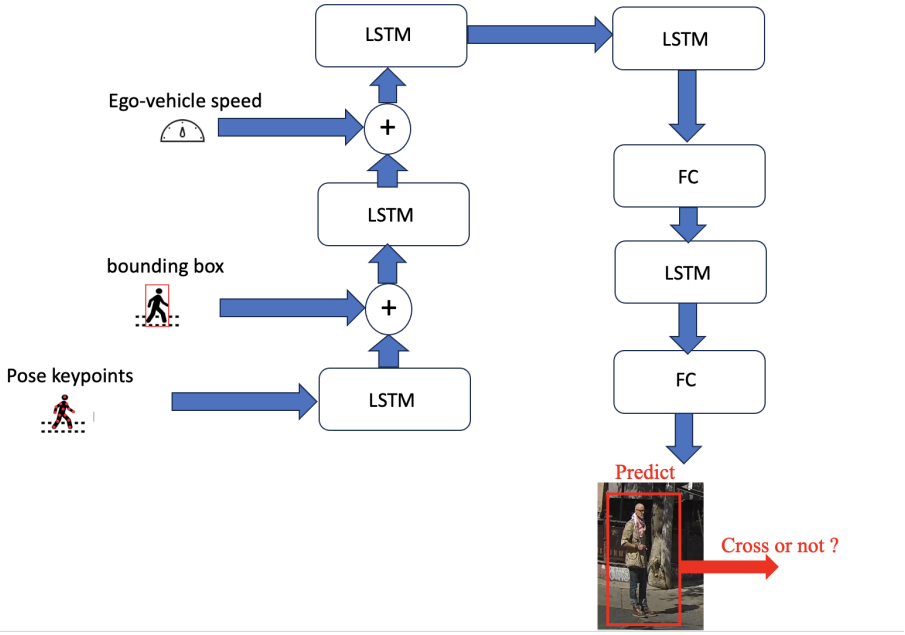

Figure 4.1. The Predicting Pedestrian Crossing Intention model (PPCI) architecture

RNN module. We use Long Short-Term Memory (LSTM) [14] to build the RNN module. The applied LSTMs have 256 hidden units. LSTMs are a special kind of RNNs that are equipped with memory cells and gates. These structures enable them to maintain information

in their memory for extended periods. Each hidden units corresponds to a memory cell in the LSTM, allowing the network to store and retrieve information efficiently over time.

LSTMs possess a unique cell structure that distinguishes them from traditional RNNs. They introduce three critical gates: the input gate (i_t), forget gate (f_t), and output gate (o_t), along with a cell state (C_t).

Given an input vector x_t and a previous hidden state h_{t-1}, the LSTM updates are as follows:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{4.1}$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{4.2}$$

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{4.3}$$

$$\tilde{C}_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \tag{4.4}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{4.5}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{4.6}$$

Where:

- σ is the sigmoid activation function.
- tanh is the hyperbolic tangent activation function.
- W and b are the weight matrices and bias terms for the respective gates.

These gates control the flow of information, making sure the network retains or forgets data as needed. The main advantage of LSTMs over standard RNNs is their ability to remember long-term dependencies, thanks to the cell state and these gating mechanisms.

Hybrid fusion. We applied a hierarchically way of fusing the features from different sources. The RNN module branch fuses three non-visual features (bounding boxes, pose key

points, and vehicle speed). They are hierarchically fused according to their complexity and level of abstraction. This is illustrated in Figure . First, sequential pedestrian poses key points are fed to an RNN-based encoder. Second, the output of the first stage is concatenated with a bound box and fed to a new RNN-based encoder. Last, the output of the second stage is concatenated with ego-vehicle speed S and fed to a final RNN-based encoder. The output will fed to double LSTM then to fully-connection (FC) , followed with LSTM. Finally, the output fed into a fully-connection (FC) layer to obtain the final predicted action.

Training. We train the model with Adam optimizer [15], binary crossentropy loss and batch size set to 8. We train for 20 epochs on PIE dataset with learning rate set to 1×10^{-4} and reduce it after every epoch with a decay rate of 0.20 and for 90 epochs on JAAD dataset with learning rate 1×10^{-4} and reduce it after every epoch with a decay rate of 0.95.

4.2 Predicting Pedestrian Crossing Intention with Attention Mechanisms Model (PPCI_att)

The overall architecture is shown in Figure 4.2. It consists of RNN modules, attention modules, and a novel way of fusing different features.

RNN module. We use Long Short-Term Memory (LSTM) [14] to build the RNN module. The applied LSTMs have 256 hidden units. see section 4.1 RNN module.

Attention module. Attention module [16], by selectively focusing on parts of features, is used for better memorizing sequential sources. Sequential features are represented as hidden states $h_s = \{h_1, h_2, ..., h_t\}$. The attention weight is computed as shown in Equation 4.7 .

$$\alpha_{ts} = \frac{\exp(\text{score}(h_t, h_s))}{\sum_{s'=1}^{S} \exp(\text{score}(h_t, h_{s'}))} \qquad \text{[Attention weights]} \qquad (4.7)$$

where $\text{score}(h_t, h_s) = h_t^T W_s h_s$ and W_s is a weight matrix. Such attention weight trades off the end hidden state h_t with each previous source hidden state h_s. The output vector of the

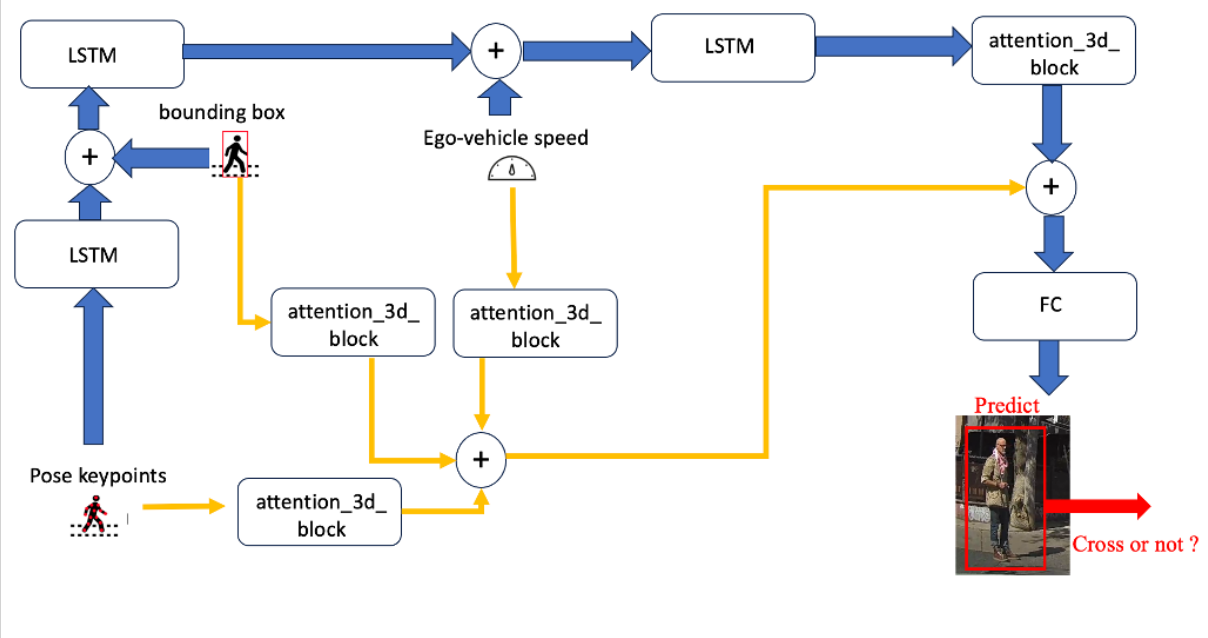

Figure 4.2. The Predicting Pedestrian Crossing Intention model (PPCI) architecture

attention module is produced as shown in Equation 4.9

$$c_t = \sum_s \alpha_{ts} h_s \qquad \text{[Context vector]} \qquad (4.8)$$

$$a_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t]) \qquad \text{[Attention vector]} \qquad (4.9)$$

where W_c is a weight matrix, and c_t is the sum of all attention weighted hidden states as shown in Equation 4.8 . The output of the attention module in our work is a feature tensor.

Hybrid fusion. We applied a hybrid way of fusing the features from different sources. The strategy is shown in Figure 4.2. The proposed architecture has two branches, one for RNN module features and one for Attention module features. The RNN module branch fuses three non-visual features (bounding boxes, pose key points, and vehicle speed). They are hierarchically fused according to their complexity and level of abstraction. This is illustrated in Figure 4.2 . First, sequential pedestrian poses key points are fed to an LSTM. Second, the output of the first stage is concatenated with a bound box and fed to a new LSTM. Last, the output of the

second stage is concatenated with ego-vehicle speed S and fed to a final LSTM , then fed into the Attention module block.

The Attention module's feature branch combines three distinct non-visual components: bounding boxes, pose key points, and vehicle speed, which are the same features used in the RNN module branch. These features are individually fed into the Attention module, and their outputs are subsequently merged by concatenating them together.

Finally, the output of RNN module branch and Attention module branch are concatenated together and fed into a fully-connection (FC) layer to obtain the final predicted action.

Training. We train the model with Adam optimizer [15], binary crossentropy loss, and batch size set to 8. We train for 20 epochs on the PIE dataset with a learning rate set to 1×10^{-4} and reduce it after every epoch with a decay rate of 0.20 and for 40 epochs on JAAD dataset with learning rate 5×10^{-4}.

Chapter 5

Experimental Analysis

5.1 Results

The evaluation is performed with the recently proposed Benchmark for Evaluating Pedestrian Action Prediction [10]. The Benchmark for Evaluating Pedestrian Action Prediction integrates the datasets PIE (Pedestrian Intention Estimation) and JAAD (Joint Attention in Autonomous Driving) into a common evaluation framework. We compare Predicting Pedestrian Crossing Intention model (PPCI) results with state-of-the-art and recently published methods for pedestrian crossing prediction [7] ,[10] and [12] which use the PIE, JAAD or both datasets for evaluation.

5.1.1 Predicting Pedestrian Crossing Intention model (PPCI)

The performance of the Predicting Pedestrian Crossing Intention model (PPCI) is presented in Table 5.1 on the PIE and JAAD datasets. The table includes the following evaluation metrics: accuracy (acc), the area under the curve (AUC), and F1 score (f1). The results indicate that the ”Predicting Pedestrian Crossing Intention” model achieves state-of-the-art performance on the JAAD Beh dataset with an accuracy of 65%, an improvement of 2% over the previous state-of-the-art.

Table 5.1. Predicting Pedestrian Crossing Intention model (PPCI) result.

Model	features	PIE			Jaad beh			Jaad all		
		ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
PPCI (our)	$pose_{HRNet}$, bound box speed	0.89	0.87	0.81	0.65	0.58	0.75	0.83	0.76	0.58
MTL [7]	bound box Pose (HRNet pertrain) speed	0.91	0.93	0.82	0.63	0.65	0.77	0.90	0.95	0.76
TED [12]	bound box	0.91	0.91	0.83	-	-	-	-	-	-
PCPA [10]	pose, box, speed, local_context	0.87	0.86	0.77	0.58	0.50	0.71	0.85	0.86	0.68

5.1.2 Predicting Pedestrian Crossing Intention with Attention Mechanisms Model (PPCI_att)

The performance of the Predicting Pedestrian Crossing Intention model (PPCI) is presented in Table 5.2 on the PIE and JAAD datasets. The table includes the following evaluation metrics: accuracy (acc), the area under the curve (AUC), and F1 score (f1). The results indicate that the Predicting Pedestrian Crossing Intention with Attention Mechanisms model achieves state-of-the-art performance on both the PIE and JAAD Beh datasets, excelling in accuracy and F1 score metrics. Accuracy measures the proportion of correctly identified predictions out of the total predictions. On the other hand, the The F1 score offers insights into the balance between the model's precision (its ability to avoid false positives) and its recall (its capacity to correctly identify true positives). On the PIE dataset, the model showcases an accuracy of 91% coupled with an F1 score of 84%, marking a notable improvement of 1% in F1 compared to the preceding best results. When evaluated on the JAAD Beh dataset, the model demonstrates an accuracy of 67% and an F1 score of 77%. Impressively, this denotes an enhancement in accuracy by 4% over the previous state-of-the-art. Moreover, these results highlight the advancements the "Predicting Pedestrian Crossing Intention with Attention Mechanisms" model has made over its predecessor, the "Predicting Pedestrian Crossing Intention" model. Including attention

mechanisms has contributed to this significant leap in performance.

Table 5.2. Predicting Pedestrian Crossing Intention with Attention Mechanisms Model (PPCI_att) result.

Model	features	PIE			Jaad beh			Jaad all		
		ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
PPCI_att (our)	$pose_{HRNet}$, bound box speed	0.91	0.89	0.84	0.67	0.60	0.77	0.81	0.78	0.75
PPCI (our)	$pose_{HRNet}$, bound box speed	0.89	0.87	0.81	0.65	0.58	0.75	0.83	0.76	0.58
MTL [7]	bound box Pose (HRNet pertrain) speed	0.91	0.93	0.82	0.63	0.65	0.77	0.90	0.95	0.76
TED [12]	bound box	0.91	0.91	0.83	-	-	-	-	-	-
PCPA [10]	pose, box, speed, local context	0.87	0.86	0.77	0.58	0.50	0.71	0.85	0.86	0.68

5.2 An Ablation Study

5.2.1 Feature Selection

Feature selection is critical in developing an efficient and accurate Pedestrian Crossing Prediction model. As shown in Table 5.3, different feature combinations yield varying accuracy, AUC, and F1 scores for the PIE, Jaad beh, and Jaad all datasets. When combining Pose Hrent, bound box, and speed as features, the model demonstrates the highest accuracy (ACC) of 0.91 on the PIE dataset, suggesting that this combination is particularly effective for this specific dataset. This combination also yields promising results for the Jaad beh and Jaad all datasets. Interestingly, when the pose (Open pose) was estimated by using the Open pose method and combined with box and speed, there was a slight decrease in performance metrics across all datasets. This indicates that while ”Open pose” might be a popular pose estimation method, it

might not be the most optimized for pedestrian crossing prediction in the given context.

Table 5.3. Feature Selection for Predicting Pedestrian Crossing Intention with Attention Mechanisms Model (PPCI_att)

Features	PIE			Jaad beh			Jaad all		
	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
$pose_{HRNet}$, bound box speed	**0.91**	**0.89**	**0.84**	**0.67**	**0.60**	**0.77**	**0.81**	**0.78**	**0.75**
$pose_{Openpose}$, bound box, speed	0.86	0.86	0.77	0.64	0.58	0.75	0.79	0.77	0.58
$pose_{HRNet}$, bound box	0.88	0.86	0.79	0.62	0.53	0.74	0.79	0.78	0.56
$pose_{HRNet}$	0.80	0.76	0.65	0.52	0.45	0.65	0.74	0.75	0.50
bound box	0.82	0.82	0.72	0.62	0.54	0.73	0.78	0.76	0.54

5.2.2 Fusion Strategies

An ablation study was also conducted to compare different strategies for fusing features. We tried different fusion strategies, including later fusion and early fusion, to be compared with the proposed hybrid fusion strategy, as shown in the table 5.4. According to the results detailed in Table 5.4, the hybrid-fusion approach—when fusing features such as $pose_{HRNet}$ as estimated by HRNet, bounding box, and speed—outperformed the other methods for the PIE dataset, achieving an accuracy of 0.91, AUC of 0.89, and an F1 score of 0.84. Comparatively, early fusion rendered slightly inferior results with an accuracy of 0.89, while the later-fusion strategy clocked in with 0.88 for the same dataset. The hybrid fusion strategy in amalgamating features effectively paves the way for a more precise prediction of pedestrian intentions across various datasets.

5.2.3 layer ablation study

In another ablation study, we examined the impact of removing particular layers from the Predicting Pedestrian Crossing Intention with attention mechanisms model (PPCI_att) to observe the resulting changes in performance. We began by assessing the hierarchy model without any

Table 5.4. Comparing Feature Fusion Strategies for Predicting Pedestrian Crossing Intention with Attention Mechanisms Model (PPCI_att)

Fusion Approach	Features	PIE			Jaad beh			Jaad all		
		ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
hybrid-fusion (OUR)	$pose_{HRNet}$, bound box speed	0.91	0.89	0.84	0.67	0.60	0.77	0.81	0.78	0.75
early fusion	$pose_{HRNet}$, bound box speed	0.89	0.87	0.80	0.59	0.51	0.71	0.78	0.77	0.54
later fusion	$pose_{HRNet}$, bound box speed	0.88	0.87	0.80	0.58	0.54	0.67	0.77	0.76	0.54

attention layers to establish a baseline, as shown in Figure 5.1. Subsequently, we incorporated a late attention layer at the end of the hierarchy model, as shown in Figure 5.2.

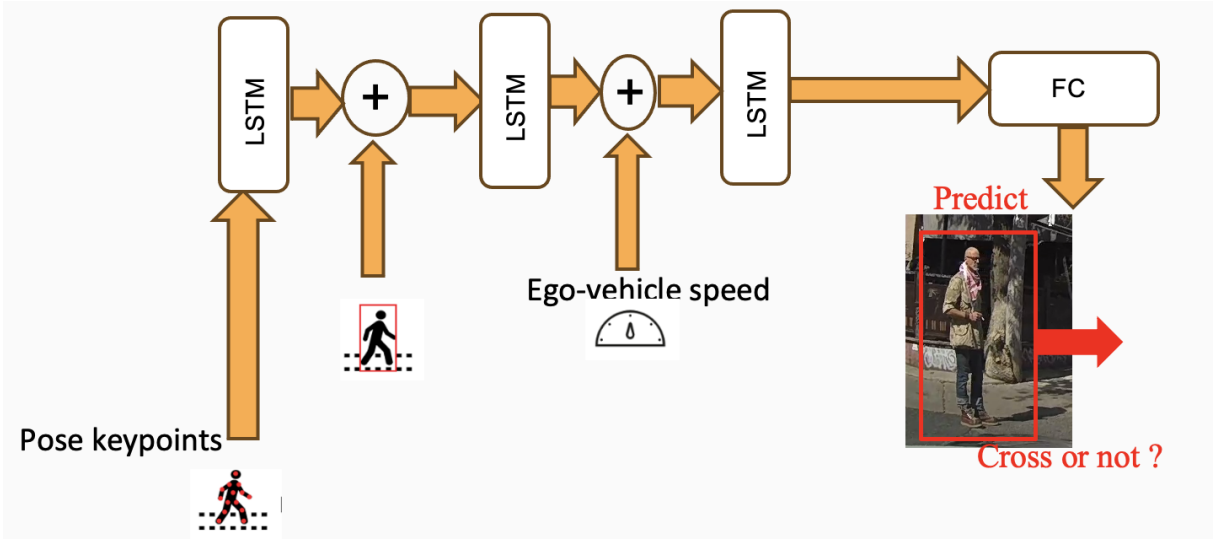

Figure 5.1. The hierarchy model architecture.

Our architecture is distinctly structured into two branches: one that processes features using an LSTM module and another that handles features through an Attention module, as shown in Figure 4.2. The results are shown in table 5.5, our model demonstrated a marked increase in accuracy, AUC (Area Under the Curve), and F1 score across both the JAAD behavioral (Jaad beh) and JAAD all datasets and the PIE dataset. This indicates that the attention mechanisms

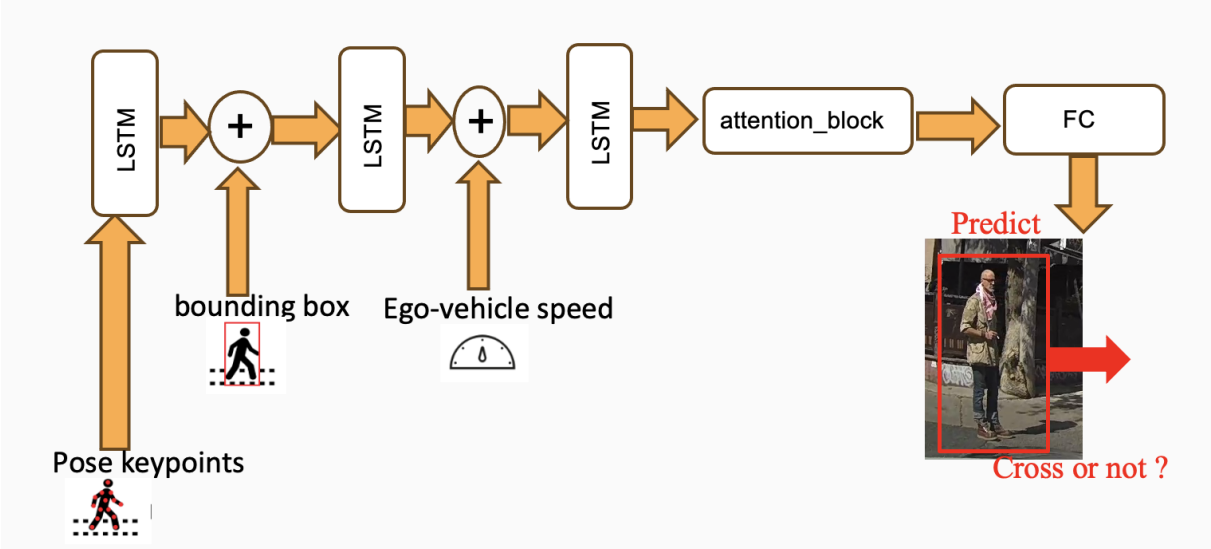

Figure 5.2. The hierarchy + late attention model architecture.

employed in our model significantly enhance its ability to predict pedestrian crossing intentions, ultimately making it the best-performing architecture in our study.

Table 5.5. layer ablation study.

Model Architecture Approach	**Features**	**PIE**			**Jaad beh**			**Jaad all**		
		ACC↑	**AUC↑**	**F1↑**	**ACC↑**	**AUC↑**	**F1↑**	**ACC↑**	**AUC↑**	**F1↑**
hierarchy	$pose_{HRNet}$, box, speed	0.89	0.86	0.80	0.62	0.55	0.74	0.80	0.76	0.56
hierarchy + late attention	$pose_{HRNet}$, box, speed	0.89	0.87	0.80	0.63	0.56	0.74	0.81	0.76	0.57
PPCI_att (OUR)	$pose_{HRNet}$, box, speed	0.91	0.89	0.84	0.67	0.60	0.77	0.81	0.78	0.75

5.3 Qualitative results

We show qualitative results for the best performing model Predicting Pedestrian Crossing Intention with an attention mechanisms model (PPCI_att).

This first case from the Jaad dataset is shown in Figure 5.3, which illustrates two pedestrians crossing the street. Our model PPCI_att was able to predict it correctly.

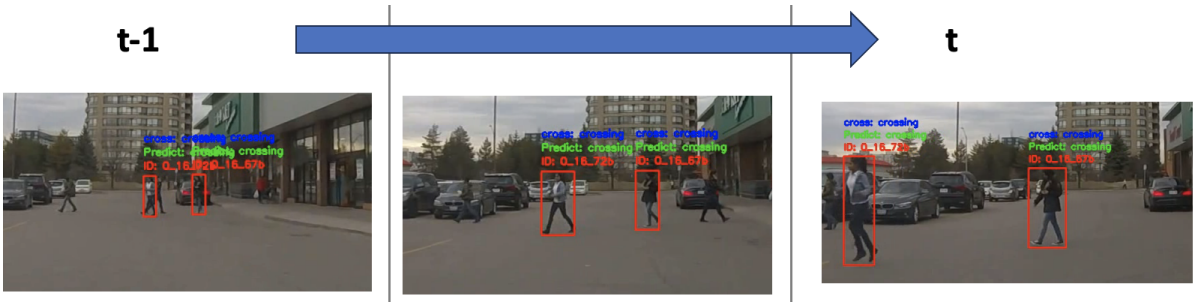

Figure 5.3. Two pedestrians crossing the street from the Jaad dataset.

Similarly, Figure 5.4 from the PIE dataset illustrates another instance of pedestrians crossing the street, where our PPCI_att model's prediction was also accurate.

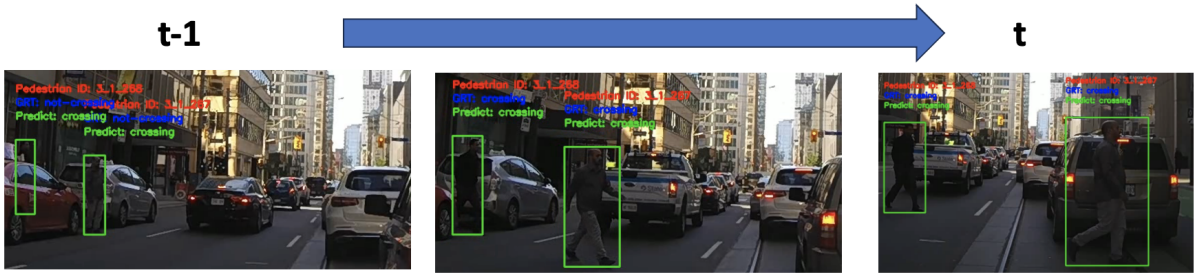

Figure 5.4. Two pedestrians crossing the street from the PIE dataset.

Another Figure 5.5 from the PIE dataset illustrates instance of pedestrians not crossing the street and our model PPCI_att was able to predict it correctly.

Figure 5.5. Two pedestrians not crossing the street from the PIE dataset.

Figure 5.6 illustrates a scenario where two pedestrians cross the street from PIE dataset. Our algorithm successfully predicted the crossing intentions of one pedestrian but failed with the other. We observed that the pedestrian our model PPCI_att failed to predict momentarily exits

the frame, rendering them invisible. This temporary absence likely impacted our model's ability to accurately predict their crossing intention.

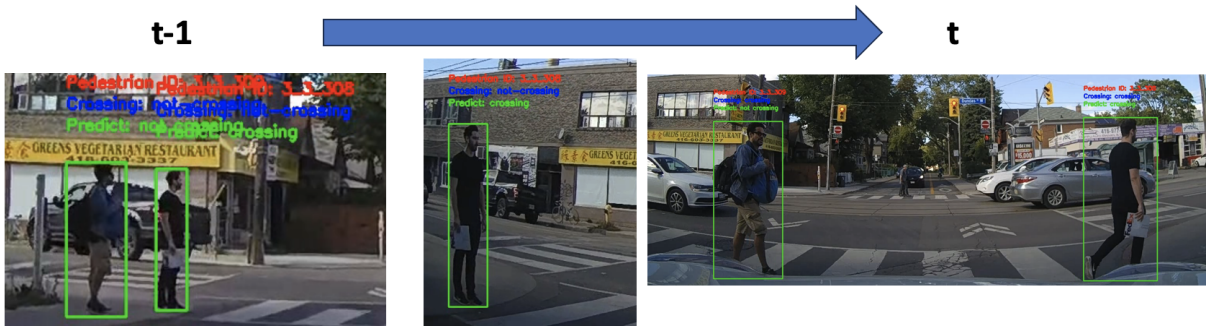

Figure 5.6. Two pedestrians crossing the street from the PIE dataset.

Chapter 6

CONCLUSION

In this work, we propose a novelty Predicting Pedestrian Crossing Intention with an attention mechanisms model (PPCI_att) based on LSTM and attention mechanism to predict pedestrian attention intention to cross the street or not by using non-visual feature pose boundary box and vehicle speed. We Evaluate the model on two widely-used pedestrian datasets: the Joint Attention in Autonomous Driving (JAAD) dataset and The Pedestrian Intention Estimation (PIE) dataset. Predicting Pedestrian Crossing Intention with the attention mechanisms model (PPCI_att) achieved state-of-the-art results on both the PIE and JAAD datasets. We show different Ablation studies, study the effects of different input features. The second ablation study is about the features of Fusion.

Finally, we showed qualitative results and failed cases. Future work can focus on improving our model by using feature fusion with more information sources can be explored, e.g., using trajectory.

Bibliography

[1] "National highway traffic safety administration (nhtsa). traffic safety facts 2020." https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813375., accessed: 2023-1-26.

[2] A. Rasouli, I. Kotseruba, and J. 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*, 2017, pp. 206–213.

[3] A. Rasouli, I. Kotseruba, T. Kunic, and J. 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*, 2019, pp. 6262–6271.

[4] R. Greer, S. Desai, L. Rakla, A. Gopalkrishnan, A. Alofi, and M. Trivedi, "Pedestrian behavior maps for safety advisories: Champ framework and real-world data analysis," *arXiv preprint arXiv:2305.04506*, 2023.

[5] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özgüner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 221–230, 2022.

[6] J. Lian, F. Yu, L. Li, and Y. Zhou, "Early intention prediction of pedestrians using contextual attention-based lstm," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 14 713–14 729, 2023.

[7] D. Schörkhuber, M. Pröll, and M. Gelautz, "Feature selection and multi-task learning for pedestrian crossing prediction," in *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2022, pp. 439–444.

[8] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[9] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[10] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1258–1268.

[11] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "Pedestrian graph+: A fast pedestrian crossing prediction model based on graph convolutional networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21 050–21 061, 2022.

[12] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillet, "Is attention to bounding boxes all you need for pedestrian action prediction?" in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 895–902.

[13] M. I. Perdana, W. Anggraeni, H. A. Sidharta, E. M. Yuniarno, and M. H. Purnomo, "Early warning pedestrian crossing intention from its head gesture using head pose estimation," in *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE, 2021, pp. 402–407.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.