Pedestrian Safety by Intent Prediction: A Lightweight LSTM-Attention Architecture and Experimental Evaluations with Real-World Datasets

Afnan Alofi
aalofi@ucsd.edu

Ross Greer
regreer@ucsd.edu

Akshay Gopalkrishnan
agopalkr@ucsd.edu

Mohan Trivedi
mtrivedi@ucsd.edu

***Abstract*—Autonomous vehicles face significant challenges in understanding pedestrian behavior, particularly in urban environments. In such settings, the system must recognize pedestrian intentions and anticipate their actions to achieve safe and intelligent driving. This paper focuses on predicting pedestrian crossings, enabling oncoming vehicles to react to pedestrians in a traffic scene in a timely manner. We investigate the effectiveness of various input features 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 identify crossing pedestrians. Our methods are evaluated on two widely used public datasets for pedestrian behavior, PIE and JAAD datasets, and our algorithm achieves a state-of-the-art performance in both datasets by reaching a prediction accuracy of 91% and an F1-score of 84% on the PIE dataset and an accuracy of 67% and an F1-score of 77% on the JAAD Behavior (BEH) split. We make our code available at https://github.com/afnan29alofi/Pedestrian-Intent-Prediction.git**

***Index Terms*—machine learning, pedestrian intention, autonomous vehicles, crossing safety, intent prediction**

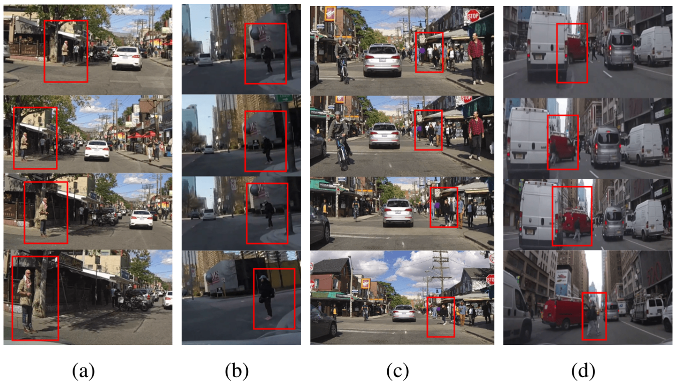

Fig. 1. (a) PIE Dataset: The pedestrian has the intention to cross but did not actively cross the street (b) Jaaad Dataset: The pedestrian has the intention to cross but did not actively cross the street (c) PIE Dataset: The Pedestrian unexpectedly crosses the street (d) Jaad Dataset: A pedestrians crossing in the middle of the road

I. Introduction

In 2020, the National Highway Traffic Safety Administration reported 6,516 pedestrian fatalities and 54,769 pedestrian injuries resulting from traffic accidents in the United States [1]. Despite vehicle and road safety advancements over the past two decades, there has been a 42% increase in pedestrian fatalities on public roads from 2000 to 2020. 80% of pedestrian deaths 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, we 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. We leverage deep learning techniques to accurately predict pedestrian intentions for any situation, especially potentially harmful scenarios where they might cross the street unexpectedly, such as crossing in the middle of the street instead of using designated crosswalks [2] as shown in Figure 1d. Another challenging scenario for predicting pedestrian crossing behavior occurs when pedestrians are standing on a crosswalk but display hesitation in crossing, as depicted in Figures 1a and 1b. Figure 1c highlights another challenge prediction scenario when a pedestrian suddenly changes their crossing direction.

By building this predictive model, our research contributes to the development of automated driver assistance systems for prevention of pedestrian accidents. 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. In addition, this pedestrian intention prediction model can be integrated with separate models to predict driver takeover time [3] to have an accurate understanding of when to alert drivers if a pedestrian is in danger. In summary, the contributions of this work are threefold:

1) We propose a model, PPCI_{att}, which uses non-visual features to accurately predict a pedestrian's intention to cross the street.

All authors are associated with the Laboratory for Intelligent & Safe Automobiles (LISA) at the University of California, San Diego in La Jolla, USA.

TABLE I
COMPARING THE PIE AND JAAD DATASETS

Dataset	PIE	JAAD
Year released	2019	2017
Total number of frames	909,480	82,032
Video duration	over 6 hours	over 240 hours
Pedestrians with behavior annot.	1842	686
Number of pedestrians crossing	519	495
Number of pedestrians not crossing	1323	191
Geographic Scope	Toronto, Canada	North America & Eastern Europe
Ego-vehicle sensor	yes	No
Nighttime Cases	None	4 Videos
Pedestrian adult and young	1640	574
Pedestrian child	17	16
Pedestrian senior	185	96

2) We conduct extensive ablation studies, focusing on fusion strategies, layer analysis, and feature selection, to identify the most effective configuration for optimal performance.
3) Finally, we evaluate the model performance on two widely used pedestrian datasets: the JAAD dataset [4] and PIE dataset [5].

II. Related Research

Though general awareness of the presence of pedestrians in urban driving environments may help drivers to stay advised and alert to possible crossing [6], [7], observing and predicting the motion of individual pedestrians remains an important safety challenge.

A. *Pedestrian Intention Datasets*

The Pedestrian Intention Estimation (PIE) dataset [5] and the Joint Attention in Autonomous Driving (JAAD) dataset [4] are two widely utilized public datasets for analyzing pedestrian behavior. Both the PIE and JAAD datasets serve as essential resources in the study of pedestrian dynamics, offering behavior annotations for 2,337 pedestrians and approximately one million frames in total. A full comparison between the PIE dataset and the JAAD dataset is presented in Table I. The PIE dataset includes 519 pedestrian crossings and 1,323 non-crossings. The JAAD dataset includes 495 crossings and 191 non-crossings. JAAD is split into JAAD BEH, focused on crossing behaviors, and JAAD All, which includes 2,100 including pedestrians not involved in crossing activities. There are various features of these datasets that we specifically use in our research, similar to the methods discussed in Table II:

- **Pose:** The human pose detections from the JAAD and PIE datasets were estimated using the OpenPose technique; these poses were refined in later research to provide more precise labels. In our methodology, we harnessed the capabilities of HRNet [13] to enhance the accuracy of human pose estimation.
- **Speed of Ego-Vehicle:** In PIE, vehicle speed is measured in kilometers per hour using multiple sensors, while JAAD uses numerical labels (0-4) to indicate vehicle states like stationary, slow, fast, decelerating, and accelerating.
- **Bounding Box:** PIE and JAAD datasets use sequences to define pedestrians' position and size, marked by coordinates of the upper-left (x_1, y_1) and lower-right (x_2, y_2) corners of the bounding box.

We detail the JAAD All split performance in Table II to introduce pedestrian behavior prediction. Our research primarily focuses on the JAAD BEH split, emphasizing pedestrians near crosswalks and excluding distant ones, aligning with our study on potential crossings.

B. *Which Visual Details Are Important?*

Lian et al. [8] proposed a pedestrian intention prediction using a contextual attention-based LSTM. This proposed model uses two LSTM layers and an attention mechanism for predicting pedestrian crossing behavior. It detects and tracks pedestrians by extracting visual, contextual, and motion features. Visual features come from convolution layers and pooling applied to pedestrian areas in images, while contextual features are similar but with expanded bounding boxes. Motion features used in this approach include velocity and the walking angle.

However, strong performance for pedestrian intention prediction can also be achieved using low-level features from images, without extracting, learning, or using fine visual details that most human drivers would consider informative (eye contact, face direction, etc.). Schörkhuber et al. [9] design a pedestrian intention model that incorporates three coarse features: vehicle speed, pedestrian bounding box, and pedestrian pose. While the body pose provided in the JAAD and PIE datasets were originally extracted using the OpenPose method, the study observed limitations in OpenPose model performance on the image data within these datasets. The pose detection challenges stemmed from most pedestrians being located at a considerable distance, resulting in small image crops and out-of-focus image regions, hindering accurate pose estimation. To overcome this issue, the researchers adopted HRNet [13] for body pose detection, training on the BDD100K dataset [14] and generating more accurate poses for the pedestrians of JAAD and PIE. This strategic choice substantially enhanced the precision of detected poses within the datasets, and consequently, the pedestrian crossing prediction accuracy significantly improved, achieving a state-of-the-art performance level of 91% accuracy. This serves to highlight that the quality with which data is labeled serves as a significant factor in the performance of models trained on such data, sometimes having even more impact than the choice of input features.

Kotseruba et al. [10] introduced 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 PIE and JAAD, considering various data properties such as time-to-event, occlusion, and scale.

TABLE II
PROGRESSION OF DATA-DRIVEN APPROACHES TO PEDESTRIAN INTENT ESTIMATION

Methodology	Features	PIE			JAAD - All			Important Findings
		acc ↑	*auc* ↑	*f1* ↑	*acc* ↑	*auc* ↑	*f1* ↑	
Series LSTM and Attention Modules [8]	Visual + Contextual Features, Pedestrian Velocity, Walking Angle	-	-	-	0.89	-	0.75	Pedestrian's velocity and walking angle are useful features.
RNN-Based Encoder-Decoder [9]	Bounding Box, Speed, PoseHRNet	0.91	0.93	0.82	0.90	0.95	0.76	Multi-task learning to predict crossing, trajectory, and location improves all performance; pose from HRNet improves performance.
RNN with Attention Modules [10]	Bounding Box, Speed, Pose, Local Context [10]	0.87	0.86	0.77	0.85	0.86	0.68	Establishes a benchmark for evaluating pedestrian action prediction models.
Graph Convolution Network and Parallel RNN [11]	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.
Transformer [12]	Bounding Box	0.91	0.91	0.83	-	-	-	Achieves superior results using only input feature bounding boxes.
LSTM + Self Attention Modules with Hybrid Fusion (Ours)	Pose, Bounding Box, Speed of Ego-Vehicle	**0.91**	0.89	**0.84**	0.81	0.78	0.75	Hybrid fusion with LSTM and Transformer Modules achieves best performance amongst other fusion schemes.

The authors proposed a hybrid model that combines recurrent and 3D convolutional approaches with temporal and modality attention mechanisms, achieving 87% and 85% classification accuracy on the PIE and JAAD datasets.

Another approach which uses various learned and raw features from data is presented by Cadena et al. [11], who created the Pedestrian Graph+ model to predict pedestrian crossings in urban areas using a Graph Convolutional Network. It includes two convolutional modules that use images, segmentation maps, and vehicle velocity for more accurate predictions. The Pedestrian Graph+ model stands out for its efficiency, offering fast inference (6 ms) and low memory usage while maintaining accuracy. It achieved 86% and 89% accuracy on the JAAD and PIE datasets, respectively.

Conversely to the above approaches which use many features of the dataset, certain studies have successfully harnessed individual features to yield impressive outcomes. For example, Achaji et al. [12] 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 91% prediction accuracy and an F1-score of 83%.

Unlike the methodologies listed in the table, our model is engineered to be exceptionally lightweight as it utilizes non-visual input features, while still maintaining robustness and accuracy. It is uniquely designed to adapt to a variable number of inputs, which allows for greater flexibility in deployment across different scenarios. Moreover, our use of HRNet for pose estimation sets us apart, as we do not require pre-training on the BDD100K dataset—a key distinction from the approach of [9]. Additionally, we did not employ multi-task learning, which typically involves joint training to predict crossing, trajectory, and location. By focusing directly on the core task of intent estimation, our model is not only efficient but also more suitable for real-time applications where computational resources are limited, without the need to process extensive visual and trajectory data.

III. METHODS

A. *Input Features and Output*

We define three input feature groups $\{p_t, b_t, s_t\}$ for each time t from total clip duration T:

1) Pose keypoints, defined as:

$$p_t = \{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}$$

where n is the number of extracted pose keypoints for an individual frame.

2) Bounding boxes, defined as:

$$b_t = \{(x_{top}, y_{top}), (x_{bottom}, y_{bottom})\}$$

which provide the pixel boundary corners of the box surrounding the pedestrian.

3) Ego vehicle speeds, defined as s_t, which may be a value (PIE) in kilometers per hour, or numerical labels (0-4) to indicate vehicle states like stationary, slow, fast, decelerating, and accelerating (JAAD).

The output of the model is defined for each clip as I_T, a binary value for the pedestrian intention, representing whether they intend to cross or do not intend to cross.

B. *Enhanced Feature Extraction*

To first focus on pedestrians in the scene, we would crop the video frame to align with the respective bounding box of pedestrian annotations. Subsequently, 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 $[-1, 1] \in \mathbb{R}$.

C. *Predicting Pedestrian Crossing Intention with Attention Mechanisms Model (PPCI_{att})*

The overall architecture for the PPCI_{att} model is shown in Figure 2. There are three main modules that make up the PPCI_{att} model:

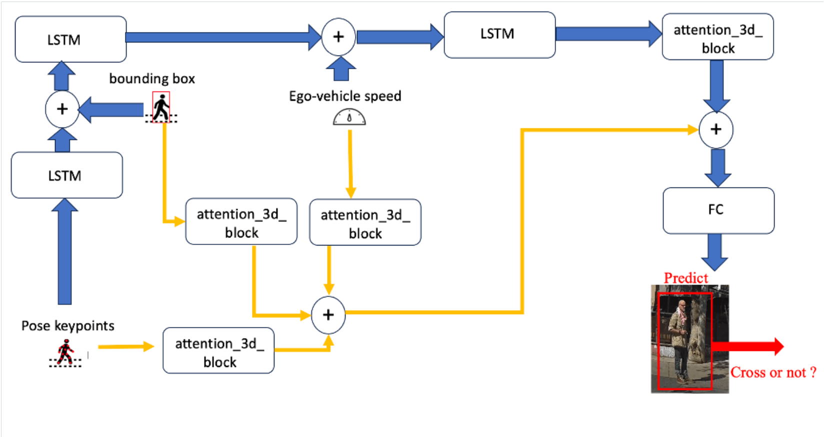

Fig. 2. The Predicting Pedestrian Crossing Intention with Attention Mechanisms Model (PPCI$_{\text{att}}$) Architecture has two branches, hierarchically RNN module features and later Attention module features. The RNN branches sequentially fused pose keypoints, a pedestrian bounding box, and the ego vehicle speed in that order, while the attention module fuses all features in one step. Then, the output from the final steps of the attention and LSTM module are concatenated together and sent through a

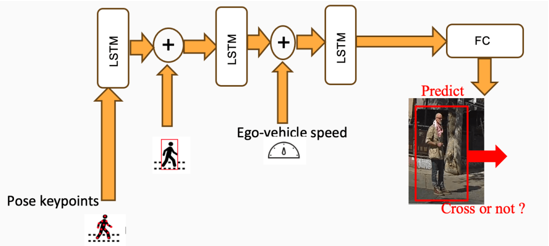

Fig. 3. The hierarchy model architecture used for our LSTM layers. The pose keypoints, pedestrian bounding box, and ego-vehicle speed are sequentially passed into different LSTM layers. The output of one layer is then concatenated with the input feature for the current layer.

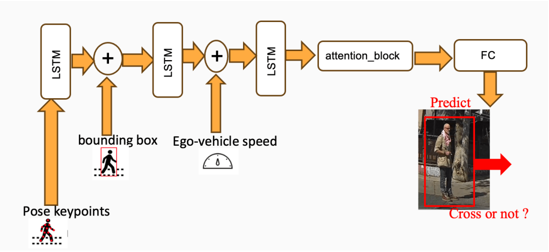

Fig. 4. The hierarchy + single attention layer model architecture. This model follows the same structure as Figure 3, except for the addition of an attention block before the fully-connected layer.

1) LSTM module: We use a Long Short-Term Memory (LSTM) network [15] to build the RNN module. The applied LSTMs have 256 hidden units. LSTMs are a helpful building block for our model since they help us learn long-term temporal dependencies within the features we use.

2) Attention module: Attention [16] helps learn dependencies amongst parts of features and is used to improve long-dependency learning from sequential sources. In our implementation, sequential features are represented as hidden states $h_s = \{h_1, h_2, ..., h_t\}$. The attention weight is computed as:

$$\alpha_{ts} = \frac{\exp(\text{score}(h_t, h_s))}{\sum_{s'=1}^{S} \exp(\text{score}(h_t, h_{s'}))} \tag{1}$$

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 attention module is produced as:

$$c_t = \sum_s \alpha_{ts} h_s \tag{2}$$

$$a_t = f(c_t, h_t) = \tanh(W_c[c_t; h_t]) \tag{3}$$

where W_c is a weight matrix, and c_t is the sum of all attention weighted hidden states. We use the output of the attention module as a feature tensor. Data-driven models which use attention have been very successful in a variety of safe autonomous driving tasks, including sign and light detection [17], [18] and trajectory prediction [19], [20]. By integrating attention into this model, we expect that the model will be able to selectively learn from the most relevant features in a given context.

3) Hybrid fusion: We fuse the features from different sources using a hybrid strategy, shown in Figure 2. The proposed architecture has two branches, one for the RNN module features and one for the 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. First, sequential pedestrian pose key points are fed into an LSTM. Then, the output from the first stage is concatenated with the vehicle bounding box features and then sent as input into a new LSTM. Last, the output of the second stage is concatenated with ego-vehicle speed S and fed to a final LSTM, whose output is then fed into the Attention module block.

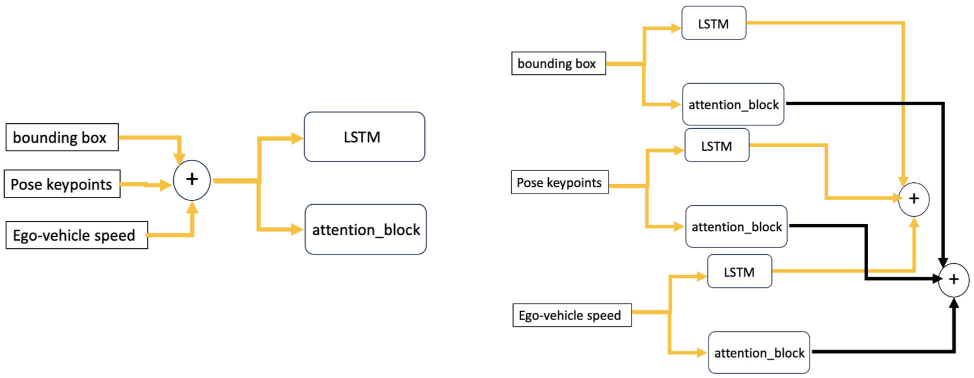

Fig. 5. The Early Fusion strategy (left) combines input features prior to neural network processing, while the Late Fusion strategy (right) combines the neural network-extracted features. Both strategies continue into a fully-connected layer for intent classification.

TABLE III
PREDICTING PEDESTRIAN CROSSING INTENTION WITH ATTENTION MECHANISMS MODEL ($\mathrm{PPCI_{ATT}}$) RESULTS

Model	**features**	**PIE**			**JAAD BEH**			**JAAD all**		
		ACC↑	**AUC↑**	**F1↑**	**ACC↑**	**AUC↑**	**F1↑**	**ACC↑**	**AUC↑**	**F1↑**
$\mathrm{PPCI_{att}}$ (ours)	$\mathrm{pose_{HRNet}}$, bounding box, speed	**0.91**	0.89	**0.84**	**0.67**	0.60	**0.77**	0.81	0.78	0.75
MTL [9]	bounding box, Pose (HRNet pretrained), speed	**0.91**	**0.93**	0.82	0.63	**0.65**	0.77	**0.90**	**0.95**	**0.76**
TED [12]	bounding box	0.91	0.91	0.83	-	-	-	-	-	-
PCPA [10]	pose, bounding box, speed, local context	0.87	0.86	0.77	0.58	0.50	0.71	0.85	0.86	0.68

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 concatenation.

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 of the pedestrian.

D. Training

We utilized the recently proposed Benchmark for Evaluating Pedestrian Action Prediction [10]. This benchmark merges the PIE and JAAD datasets into a common evaluation framework. Within this framework, the sample overlap standardized at 0.6 for the PIE dataset and 0.8 for the JAAD dataset and the observation period for all models is standardized at 16 frames ,the final frame observed occurs within a 1 to 2 second window (equivalent to 30 to 60 frames) before the initiation of the crossing event. We train the model with Adam optimizer [21], binary cross entropy 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}.

IV. ABLATION STUDY

A. Feature Selection

Feature selection is critical for developing an efficient and accurate Pedestrian Crossing Prediction model. We perform ablation studies to compare various input combinations of features such as bounding box, speed, and two types of pose estimates (from HRNet and OpenPose).

B. Fusion Strategies

Fusion strategies are techniques used to integrate multiple sources of data or features to enhance the predictive performance of models. We conduct ablation studies to investigate the efficacy of various feature fusion strategies for combining our selected features of pose, bounding box, and vehicle speed. We define three fusion strategies:

1) Late Fusion, in which each feature is passed to an individual LSTM and attention block, then the corresponding outputs are fused in the fully-connected layer for inference.
2) Early Fusion, in which each of the features are concatenated before entering the LSTM and attention blocks.
3) Hybrid Fusion, in which features are joined to the LSTM one at a time (that is, pose features are passed to LSTM, then bounding box features are concatenated to the LSTM-features of pose, then ego vehicle features are concatenated to the joint pose-bounding-box LSTM features). This strategy is illustrated in Figure 2.

We compare our proposed hybrid fusion approach, which combines elements of both early and late fusion, to these traditional strategies. Early fusion involves integrating data at the beginning of the processing pipeline, while late fusion combines the outputs of independent models towards the end of the process, are shown in Figure 5.

TABLE IV
FEATURE SELECTION FOR PREDICTING PEDESTRIAN CROSSING INTENTION WITH ATTENTION MECHANISMS MODEL (PPCI$_{\text{ATT}}$)

Features	PIE			JAAD BEH			JAAD all		
	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑
pose$_{\text{HRNet}}$, bounding box, speed	**0.91**	**0.89**	**0.84**	**0.67**	**0.60**	**0.77**	**0.81**	**0.78**	**0.75**
pose$_{\text{OpenPose}}$, bounding box, speed	0.86	0.86	0.77	0.64	0.58	0.75	0.79	0.77	0.58
pose$_{\text{HRNet}}$, bounding box	0.88	0.86	0.79	0.62	0.53	0.74	0.79	0.78	0.56
pose$_{\text{HRNet}}$	0.80	0.76	0.65	0.52	0.45	0.65	0.74	0.75	0.50
bounding box	0.82	0.82	0.72	0.62	0.54	0.73	0.78	0.76	0.54

TABLE V
COMPARING FEATURE FUSION STRATEGIES FOR PREDICTING PEDESTRIAN CROSSING INTENTION WITH ATTENTION MECHANISMS MODEL (PPCI$_{\text{ATT}}$)

Fusion Approach	**Features**	**PIE**			**JAAD BEH**			**JAAD all**		
		ACC↑	**AUC↑**	**F1↑**	**ACC↑**	**AUC↑**	**F1↑**	**ACC↑**	**AUC↑**	**F1↑**
Hybrid-Fusion	pose$_{\text{HRNet}}$, bounding box, speed	**0.91**	**0.89**	**0.84**	**0.67**	**0.60**	**0.77**	**0.81**	**0.78**	**0.75**
Early Fusion	pose$_{\text{HRNet}}$, bounding box, speed	0.89	0.87	0.80	0.59	0.51	0.71	0.78	0.77	0.54
Late Fusion	pose$_{\text{HRNet}}$, bounding box, speed	0.88	0.87	0.80	0.58	0.54	0.67	0.77	0.76	0.54

TABLE VI
LAYER ABLATION STUDY.

Model Architecture	**Features**	**PIE**			**JAAD BEH**			**JAAD all**		
		ACC↑	**AUC↑**	**F1↑**	**ACC↑**	**AUC↑**	**F1↑**	**ACC↑**	**AUC↑**	**F1↑**
Hierarchy	pose$_{\text{HRNet}}$, bounding box, speed	0.89	0.86	0.80	0.62	0.55	0.74	0.80	0.76	0.56
Hierarchy + Single Attention Layer	pose$_{\text{HRNet}}$, bounding box, speed	0.89	0.87	0.80	0.63	0.56	0.74	**0.81**	0.76	0.57
PPCI$_{\text{att}}$	pose$_{\text{HRNet}}$, bounding box, speed	**0.91**	**0.89**	**0.84**	**0.67**	**0.60**	**0.77**	**0.81**	**0.78**	**0.75**

C. Layer Ablation Study

In another ablation study, we examined the impact of removing particular layers from the Predicting Pedestrian Crossing Intention with Attention Mechanisms (PPCI$_{\text{att}}$) model to observe the resulting changes in performance. We first use the hierarchical model architecture without any attention layers, as shown in Figure 3. Second, we use the model with the single attention layer incorporated at the end of the hierarchy model, as shown in Figure 4. Both of these variations lack the benefit of PPCI$_{\text{att}}$'s attention blocks directly applied to the input features, so we expect these to perform with less accuracy than PPCI$_{\text{att}}$.

V. RESULTS

We compare the Predicting Pedestrian Crossing Intention with Attention Mechanisms Model (PPCI$_{\text{att}}$) results with state-of-the-art and recently published methods for pedestrian crossing prediction [9], [10], and [12] which use PIE, JAAD, or both datasets for evaluation. This comparison is presented in Table III. The table includes the following standard binary classification evaluation metrics: accuracy (acc), the area under the curve (AUC), and F1 score.Accuracy measures the proportion of correctly identified predictions out of the total predictions. On the other hand, 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).These metrics are frequently used by most related works.

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 Behavior datasets, excelling in accuracy and F1 score metrics.

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 Behavior dataset, the model demonstrates an accuracy of 67% and a F1 score of 77%, denoting an enhancement in accuracy by 4% over the previous state-of-the-art. We note that Schörkhuber et al. [9] outperform PPCI$_{\text{att}}$ on the JAAD all dataset. While we both use the same features, pretraining HRNet on the BDD100k may have helped them have improve performance on the JAAD all dataset.

Moreover, these results highlight the advancements the "Predicting Pedestrian Crossing Intention with Attention Mechanisms" model has made over its non-attention predecessor, suggesting that including attention mechanisms has contributed to this significant leap in performance.

We show some qualitative results for the best performing model PPCI$_{\text{att}}$, which can be seen in Figures 7-9. All of these examples are taken from the JAAD and PIE datasets. In these examples, we see accurate predictions on pedestrians regardless of location, occlusion, and number of pedestrians in the scene. One interesting scenario in which PPCI$_{\text{att}}$ made an error is shown in Figure 9. In this scenario, one of the pedestrian momentarily exits the frame, which likely impacted the model to accurately predict their crossing intention. Solving scenarios

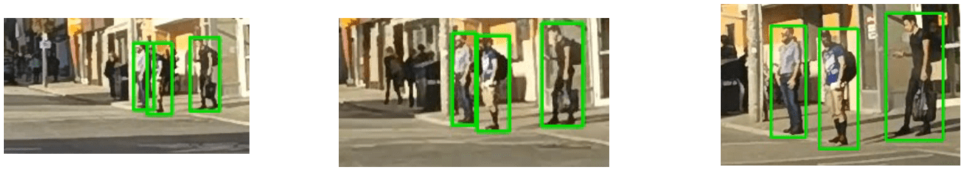

Fig. 6. Three pedestrians who will not be crossing the street from the PIE dataset. Although these pedestrians are near a road crossing, the model correctly predicts they will not be crossing the street.

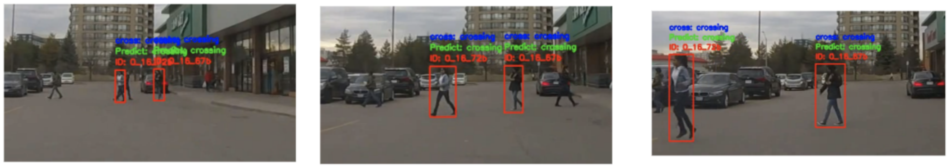

Fig. 7. Two pedestrians crossing the street from the JAAD dataset. For both pedestrians, the model makes the correct prediction.

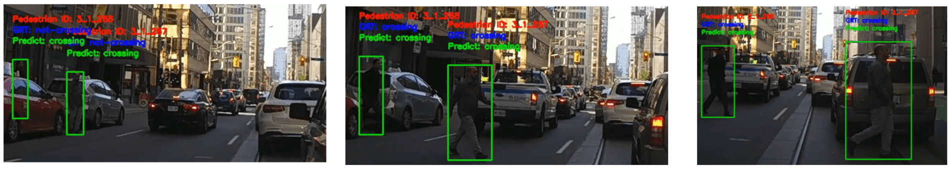

Fig. 8. Two pedestrians crossing the street from the PIE dataset. Although one of the pedestrians is partially occluded in these scene, the model still makes a correct intention prediction for this pedestrian.

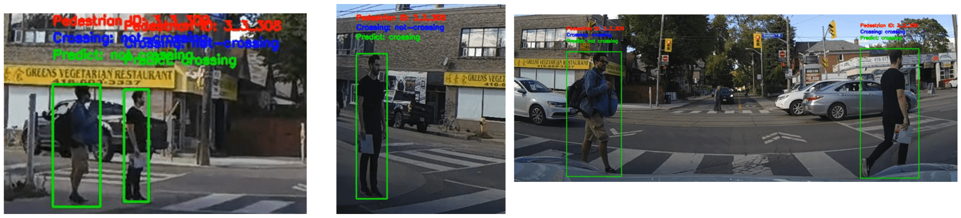

Fig. 9. Two pedestrians crossing the street from the PIE dataset. In this scenario, one of the pedestrians momentarily leaves the frame, causing an incorrect intention prediction for this model. Maintaining accurate understanding of objects when tracking fails is a challenge for pedestrian intent prediction models.

like so might require longer temporal windows of tolerance for tracking out-of-view pedestrians, which will allow for a longer accurate contextual window for each pedestrian in the scene.

A. Ablation Study Results

Feature Selection As shown in Table IV, the model achieves the highest accuracy of 91% on the PIE dataset when combining Pose HRNet, bounding box, and speed features, indicating effectiveness for this dataset and showing promising results for the JAAD BEH and JAAD all datasets. The performance gain of the HRNet features over OpenPose features illustrates the importance of a strong pose detection step prior to using the pose in downstream models. Further, in general, the more features made available to the model, the stronger the observed performance.

Fusion Strategies The hybrid-fusion method, integrating features like pose$_{\text{HRNet}}$, bounding box, and speed, was the most effective for the dataset as shown in the table V, outperforming both early fusion and later fusion. This indicates that hybrid fusion is superior for accurately predicting pedestrian intentions across various datasets. In other words, it is helpful for a model to selectively learn from relevant features coming from each input modality, rather than selecting from the composite group of features across modalities all at once.

Layer Ablation Study The results are shown in Table VI, our model demonstrated a marked increase in accuracy, AUC, and F1 score across both the JAAD behavioral (JAAD BEH) and JAAD all datasets and the PIE dataset. This indicates that the attention mechanisms employed in our model significantly enhance its ability to predict pedestrian crossing intentions, ultimately making it the best-performing architecture in our study.

B. Limitations and Future Research

We identify the following limitations and recommendations toward future research:

1) Adopting a one-hot categorical encoding scheme for the speed feature on the JAAD dataset may allow for better optimization in the learning algorithm.
2) Integration of additional informative features, in particular trajectory features [22], may further improve performance.
3) Integration of visual features, encoded using convolutional architectures, may help to provide a more detailed set of relevant features which are not fully expressed in their non-visual counterparts (e.g. fine-grained head direction versus overall body pose).

VI. Concluding Remarks

In this work, we propose a novel "Predicting Pedestrian Crossing Intention with Attention Mechanisms" model (PPCI_{att}), which uses non-visual features such as human pose keypoints, pedestrian bounding boxes, and vehicle speed to predict a pedestrian's intention to cross the street. The underlying structure is composed of LSTM and attention mechanisms. 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, showing that PPCI_{att} achieves state-of-the-art results on select metrics over both the PIE and JAAD datasets. In our ablation studies, we show the impact of particular features and fusion techniques on model performance. Finally, we show qualitative results and failed cases for the PPCI_{att} model on both pedestrian datasets.

In summary, we find the use of non-visual features to be highly informative toward the task of pedestrian intent prediction. We find that the accuracy of these extracted features, in particular pose, is highly impactful to model performance, highlighting the importance of accurate data annotation and feature extraction for intention prediction tasks [23]. We also demonstrate the performance gain enabled by the integration of attention mechanisms into the analysis of temporal features by LSTM networks. We identify open challenges to further push the state-of-the-art in pedestrian intention prediction higher in accuracy, and through our ablative study identify relevant model design and feature choices to assist in this important safety task.

References

[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] R. Greer and M. Trivedi, "From pedestrian detection to crosswalk estimation: An em algorithm and analysis on diverse datasets," *International Conference on Machine Learning, Workshop on Safe Learning for Autonomous Driving*, 2022.

[3] R. Greer, N. Deo, A. Rangesh, M. Trivedi, and P. Gunaratne, "Safe control transitions: Machine vision based observable readiness index and data-driven takeover time prediction," in *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, no. 23-0331, 2023.

[4] 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.

[5] 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.

[6] 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," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–8.

[7] R. Greer, L. Rakla, S. Desai, A. Alofi, A. Gopalkrishnan, and M. Trivedi, "Champ: Crowdsourced, history-based advisory of mapped pedestrians for safer driver assistance systems," *arXiv preprint arXiv:2301.05842*, 2023.

[8] 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.

[9] 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.

[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] 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.

[14] 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.

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

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

[17] R. Greer, A. Gopalkrishnan, N. Deo, A. Rangesh, and M. Trivedi, "Salient sign detection in safe autonomous driving: Ai which reasons over full visual context," in *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, no. 23-0333, 2023.

[18] R. Greer, A. Gopalkrishnan, J. Landgren, L. Rakla, A. Gopalan, and M. Trivedi, "Robust traffic light detection using salience-sensitive loss: Computational framework and evaluations," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–7.

[19] K. Messaoud, N. Deo, M. M. Trivedi, and F. Nashashibi, "Trajectory prediction for autonomous driving based on multi-head attention with joint agent-map representation," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 165–170.

[20] R. Greer, N. Deo, and M. Trivedi, "Trajectory prediction in autonomous driving with a lane heading auxiliary loss," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4907–4914, 2021.

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

[22] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund, "Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations," in *2015 IEEE intelligent vehicles symposium (IV)*. IEEE, 2015, pp. 330–335.

[23] R. Greer, A. Gopalkrishnan, M. Keskar, and M. M. Trivedi, "Patterns of vehicle lights: Addressing complexities of camera-based vehicle light datasets and metrics," *Pattern Recognition Letters*, vol. 178, pp. 209–215, 2024.