Predicting Pedestrian Intentions in Self-Driving Cars: Leveraging Non-Visual Features and Semantic Mapping

Amin Pakdel[1]*, Behzad Nazari[1] , Saeed Sadri[1]

Abstract-- **Predicting pedestrians' intentions to cross paths with cars, particularly at intersections and crosswalks, is critical for autonomous systems. While recent studies have showcased the effectiveness of deep learning models based on computer vision in this domain, current models often lack the requisite confidence for integration into autonomous systems, leaving several unresolved issues. One of the fundamental challenges in autonomous systems is accurately predicting whether pedestrians intend to cross the path of a self-driving car. Our proposed model addresses this challenge by employing convolutional neural networks to predict pedestrian crossing intentions based on non-visual input data, including body pose, car velocity, and pedestrian bounding box, across sequential video frames. By logically arranging non-visual features in a 2D matrix format and utilizing an RGB semantic map to aid in comprehending and distinguishing fused features, our model achieves improved accuracy in pedestrian crossing intention prediction compared to previous approaches. Evaluation against the criteria of the JAAD database for pedestrian crossing intention prediction demonstrates significant enhancements over prior studies.**

Index Term: **Pedestrian crossing intention detection, Self-driving cars, Body pose keypoints, Convolutional neural network, Semantic map**

I. INTRODUCTION

Accurately forecasting pedestrian behavior is critical for self-driving car safety, especially when people are crossıng junctions or walking in the vehicle's path. In non-autonomous systems, the driver detects a pedestrian's intention to cross or not cross the street and takes the appropriate action. However, with self-driving cars, this involves a broad comprehension of the scene as well as the extraction of crucial components for an intelligent model. To attain high accuracy, the proposed model must comprehend the significance of each extracted feature in the final prediction as well as learn the correlation between features in terms of time and location.

Previous research often relied on single-frame analysis using convolutional neural networks (CNNs), neglecting the temporal continuity between frames [1]. This oversight can significantly reduce the accuracy of pedestrian behavior prediction. A group of methods using the architecture of transformers try to learn long-range dependencies for input characteristics and predict the pedestrian crossing event [19-21]. Another group of methods tries to predict the pedestrian crossing event by combining all the features, such as the pedestrian gesture and bounding box and the speed of the moving vehicle, using attention mechanisms and recurrent networks [15-18]. Some other methods use adversarial generative models to predict the pedestrian crossing event. These methods predict the event by learning the pedestrian movement distribution and its movement patterns [22, 23].

Models utilizing recurrent neural networks (RNNs) and long short-term memory (LSTM) have made strides in pedestrian behavior prediction by accounting for spatial and temporal continuity between frames [2-5]. However, recurrent neural networks have the problem of gradient vanishing in long sequence lengths and will make mistakes in learning long-term sequences [6]. As a result, they will have difficulty learning the dependency of features in the temporal context between frames.

The purpose of this work is to enhance the effectiveness of video-based algorithms in predicting pedestrian crossing intentions at junctions or streets. In other words, predicting whether or not the i-th pedestrian would cross the crossing in 1 to 2 seconds in the future, as detected and recorded by the car's front camera in the preceding m consecutive frames. Fig.1 provides a clear illustration of this technique. In all sequences, there are 15 observed frames for each pedestrian. Also, the timing of the last observed frame is around 1 to 2 seconds (30 to 60 frames) before the commencement of the crossing or not crossing event [4] (as supplied in the JAAD dataset annotation [1]). Based on traffic studies involving pedestrians and self-driving cars, a window of one to two seconds before the occurrence was selected [7]. A certain amount of time needs to be set up for emergency action, even though it is rare that a pedestrian will cross a street or crossroads in less than two seconds [8–9]. Furthermore, because most traffic events are unexpected and human reaction dynamics are complex, it is not practicable to estimate a longer time frame for a pedestrian's movement.

This paper is structured as follows: Section II describes the proposed method, including the input data structure and the design of the model. Section III presents the research results, and Section IV discusses and concludes the findings.

1. Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan8415683111, Iran.

Corresponding author Email: amin.pakdel@alumni.iut.ac.ir

Fig. 1. The operation of predicting pedestrian crossing or not crossing the intersection. According to the sets of observed frames, this operation predicts the event of passing or not passing the pedestrian in the future so that the self-driving vehicle has enough time to react to the behavior of the pedestrian.

II. RESEARCH METHOD

A. Problem Statement

Pedestrian crossing intention prediction is defined as follows: given a sequence of input frames from the vehicle's front view and the vehicle's speed, the proposed model predicts the probability that the target pedestrian (pedestrian i) will cross the road:

$$Predict_i = P(Cross \mid F_1, F_2, ..., F_m, V) \quad (1)$$

Where, F_m represents the non-visual features extracted from frame m and V represents the speed of the moving vehicle. From the input frames, the pedestrian's body key points pose and bounding box coordinates are extracted and fed to the model in separate channels along with the vehicle's speed. The inputs of the proposed model can be categorized as follows:

1. Past 2D location of pedestrian$i$: This input is calculated for m consecutive frames based on the bounding box coordinates [1]. The 2D location is formed from the coordinates of the top-left and bottom-right corners of the bounding box as follows:

$$Location = [x_{tl}, y_{tl}, x_{br}, y_{br}] \quad (2)$$

2. Keypoint pose [2] of pedestrian$i$: This component indicates the pedestrian's motion state in each frame and is indicative of the pedestrian's intention to cross or not cross the intersection. Since the JAAD dataset does not provide pedestrian keypoint poses, we use the pre-trained OpenPose model [10] to extract the pedestrian's keypoints. This is done as follows:

$$Pose = [x_1, y_1, x_2, y_2, \ldots, x_{18}, y_{18}] \quad (3)$$

And is a 36-dimensional vector of the 2D coordinates of 18 pedestrian joints. That is:

$$Pose = [P_1, P_2, \ldots, P_{18}] \quad (4)$$

The structure of the pedestrian keypoint pose is shown in Fig. 2.

3. Speed of the moving vehicle: This component is one of the main factors influencing the decision of pedestrians to cross or not cross the intersection.

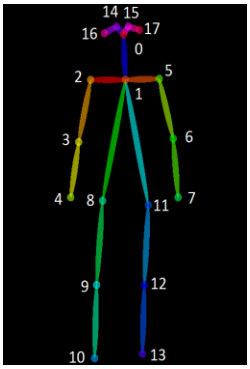

Fig. 2. The structure of 18 gesture points of the key points of the pedestrian body in [11]

The JAAD dataset annotates the speed of the autonomous vehicle qualitatively. That is, the speed component will have a value between 0 and 4 for different states. The details of speed assignment are shown in Table I.

TABLE I
Speed Component Assignment

Speed Status	Autonomous Vehicle Speed
Accelerating (increasing speed)	0
Accelerating (decreasing speed)	1
Moving fast	2
Moving slowly	3
Stopped	4

B. Model Architecture

The proposed model architecture for pedestrian crossing intention prediction is shown in Fig. 3. In this structure, the values of the three components of the moving vehicle's speed, the 2D coordinates of the pedestrian's location (bounding box), and the coordinates of the keypoints of the pedestrian's body pose for $m = 15$ consecutive frames are input to the data module as input data. In the data module, the data preparation process is performed for input to the CNN model. The structure of each of the input components is shown in Table II.

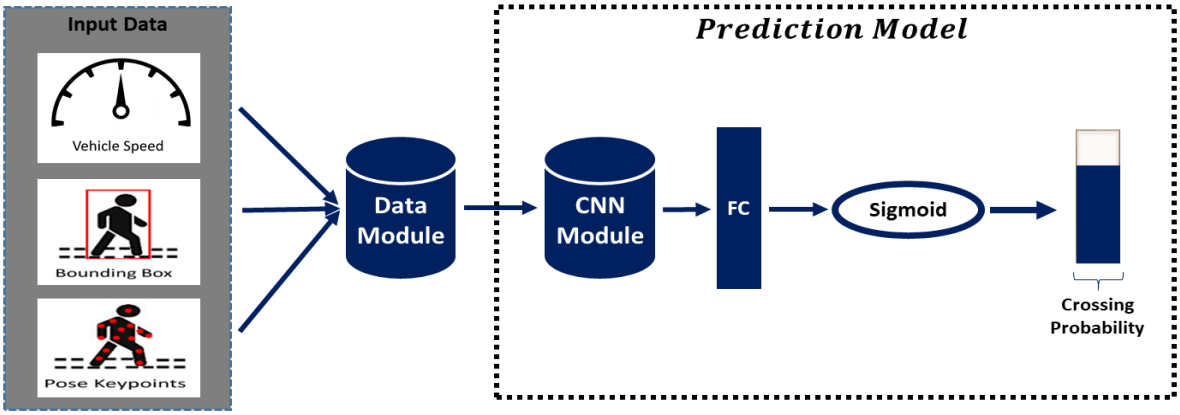

Fig. 3. The structure of the designed model

TABLE II
Structure of Each Input Component

Input Data Type	Data Dimensions for One Frame
Speed of the moving vehicle	1*1
2D coordinates of the pedestrian (Bounding Box)	1*4
Keypoint coordinates of the pedestrian's body pose	2*18

C. Data Module

In this module, the three input data components for each frame are separately arranged in a regular structure in the form of a matrix, as shown in Figure 4-right. In order for the convolutional neural network to understand each of the features placed in the input matrix and the importance of their arrangement, a semantic map like Figure 4-left in the form of a 3D tensor is attached to the feature matrix and fed to the network as input.

D. Semantic Map

The semantic map is an RGB image with dimensions of 4 * 20 * 20 pixels that is attached to the feature matrix and helps the model learn and separate the features and their importance in the final prediction. Considering the convolution of filters on the input data in each convolution layer in a CNN network, the structure of the arrangement of each of the feature components of the input data in the form of a two-dimensional matrix along with an RGB image as a semantic map can be stated as follows:

- The bounding box component indicates the coordinates of the two bottom-right and top-left points of the bounding box, each point having two components x and y.

Since the dimensions of the bounding box data for each frame are in the form of a vector with dimensions of 4 * 1, each of these components is placed in a corner of the input data matrix. The pedestrian's body keypoint pose component consists of 18 keypoints from the joints of the pedestrian's body, each point having two components x and y. Therefore, the pedestrian's body keypoint pose feature is placed in the middle of the input data matrix. On the other hand, since the speed component of the moving vehicle affects the changes in all keypoints of the pedestrian's body and also the coordinates of the pedestrian's bounding box, it must have spatial correlation with all elements of the other two features. Therefore, the outer layer of the input matrix is completely filled with the speed component.

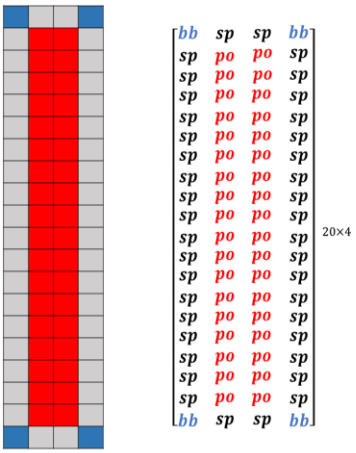

Fig. 4. The designed semantic map for combining non-visual features in each frame on the left and the non-visual feature matrix on the right. In the semantic map, the color blue represents the bounding box component, the color red represents the pedestrian's keypoint pose component, and the color white represents the speed of the moving vehicle component.

Now, for 15 consecutive frames, the input data matrix is generated, and all these matrices and their semantic maps are attached to each other in chronological order. The final input data for each sequence of 15 frames will be a tensor with dimensions of 4 * 60 * 20 (4 channels; each channel width is 20 and each channel length is 60, which is the result of appending 15 frames with a width of 4). The final structure of the semantic map for 15 consecutive frames is shown in Fig. 5.

Since the semantic map of each frame is attached to the input data matrix, the CNN network will also have a complete understanding of the importance of each of the input features, so that it will learn the impact of each feature on the other feature and the temporal texture information of each feature over 15 consecutive frames and the importance of their relationship in the prediction process.

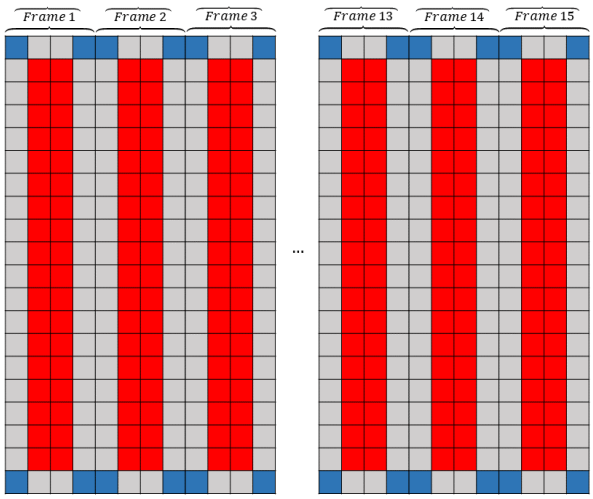

Fig 5: The final structure of the semantic map for 15 consecutive frames.

E. Convolutional Neural Networks (CNNs):

A deep convolutional neural network consists of several convolutional layers. In the first layer of a convolutional neural network, simple visual features such as edges or color spots are usually extracted. Then, in the next layers, the features of the previous layer are combined. Adding more layers extracts higher-level features such as faces, depending on the type of input data and the application of the problem. In other words, each layer in a convolutional neural network acts as a feature extractor module, and the input of each layer, except for the first layer, which is the raw image, is the features extracted from the previous layer. In general, CNNs perform very well whenever the input data structure is important and contains information. Because these networks pay attention to the relationship and correlation between the elements of the input image [12].

In our suggested model, we deploy a novel convolutional neural network structure that does not accept an image as input but rather the extracted non-visual properties, which are organized in a logical matrix and coupled to a semantic map. This function is similar to deleting the first layer of a convolutional neural network and preprocessing the non-visual characteristics before sending them to the second layer of the convolutional neural network. Figure 6 illustrates the construction of the developed CNN model.

III. EXPERIMENTS

A. Evaluation Metrics:

In this paper, the accuracy, AUC, F1 score, precision, and recall metrics are used to evaluate the results more accurately [11].

B. Dataset:

In this paper, the JAAD dataset [1] is used to train and test the proposed model. This dataset consists of two subsets with the following specifications:

1. Behavioral data: This dataset includes pedestrians crossing (495 sequences) or about to cross (191 sequences).
2. All data: This dataset contains additional pedestrians (2100 sequences) with non-interactive actions.

The training, evaluation, and test data splits are provided by the JADD dataset [1]. Approximately 60% of the sequences are assigned to the training dataset, 30% to the test dataset, and 10% to the evaluation dataset. Since this paper only focuses on predicting whether or not a pedestrian will cross, only the behavioral data dataset is considered and evaluated.

In this paper, the JAAD (joint attention for autonomous driving) dataset [1] is used to train and test the proposed model. This dataset is specifically designed to capture a wide range of pedestrian behaviors and interactions with vehicles in urban settings, making it a valuable resource for developing pedestrian intention prediction models.

C. Implementation:

The implementation details of the model are shown in Table III. A dropout layer with a value of 0.5 is also considered in each layer to prevent over fitting of the model.

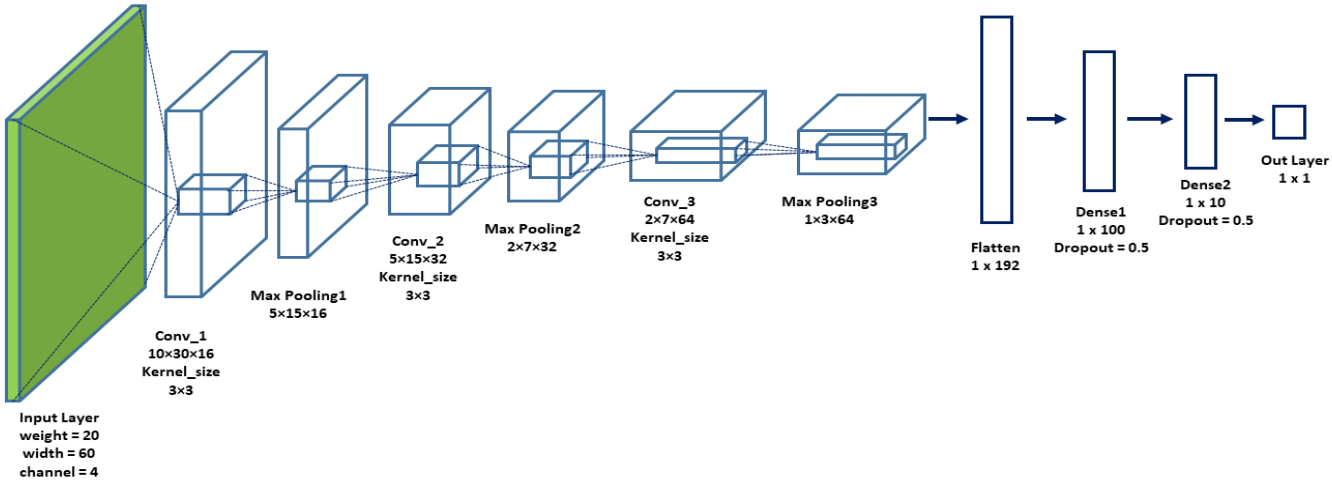

Fig6. The structure of the designed CNN model

TABLE III
Implementation Details of the Designed Model

Parameter Name	Value or Type
Number of epochs	60
Input batch size	64
Cost function	Binary cross-entropy
Optimization algorithm	Adam [14]
Learning rate	0.01

IV. Results

A. Experimental Results

The experimental results are reported in Table IV. These results are compared with the SR [2], SF-GRU [16], PCPA [13], and FusionFeature [15] models.

TABLE IV
Experimental Results on the Behavioral Data Dataset

Model Name	ACC	AUC	F1 score	Precision	Recall	Average
SR	0.59	0.52	0.71	0.64	0.80	0.65
SF-GRU	0.58	**0.56**	0.65	0.68	0.62	0.61
PCPA	0.53	0.53	0.59	0.66	0.53	0.56
Fusion Feature	0.62	0.54	0.74	0.65	0.85	0.68
Semantic Map (our)	**0.64**	0.55	**0.78**	**0.68**	**0.99**	**0.72**

The superior performance of our model can be attributed to the innovative use of the semantic map and the structured feature selection process. The semantic map provides spatial and contextual awareness, enabling the model to better understand the relationships and dependencies between different features, such as pedestrian poses, bounding boxes, and vehicle speed. This holistic understanding is critical for accurately predicting pedestrian intentions, especially in complex and dynamic environments.

Moreover, the structured feature selection ensures that relevant features are highlighted and utilized effectively, improving the model's ability to make accurate predictions. This combination of semantic mapping and feature selection creates a more comprehensive and nuanced representation of the scene, leading to improved performance across various metrics.

B. Computational Cost Analysis

The computational cost of the proposed model is a critical factor for its deployment in real-time self-driving applications. We have conducted a thorough analysis of the model's performance in terms of processing time and memory usage.

TABLE V
Results of Processing Time and Memory Usage

Model	GPU (Tesla T4)	Intel Xeon CPU
SR	602.1 FPS	413.4 FPS
SF-GRU	401.5 FPS	214.7 FPS
PCPA	211.4 FPS	107.6 FPS
Fusion Feature	182.3 FPS	91.3 FPS
Semantic map (ours)	1215.88 FPS	798.84 FPS

Processing Time: The average processing time per frame is approximately 1.2 milliseconds on a high-performance GPU (NVIDIA Tesla T4). This processing time includes semantic map generation and the convolutional neural network (CNN) inference.

GPU Performance:

The Semantic Map model operates at 1215.88 FPS, making it approximately 9 times faster than the Fusion Feature [15] model (182.3 FPS), 8 times faster than the PCPA [13] model (211.4 FPS), and 5 times faster than the SF-GRU [16] model (401.5 FPS). The SR model [2], with a processing time of 602.1 FPS, is also significantly outperformed.

The remarkable processing speed of our proposed model can be attributed to the efficiency of the semantic map and the structured feature selection process. By effectively organizing and prioritizing features, our model minimizes the computational overhead typically associated with complex feature interactions. This streamlined approach ensures rapid inference, which is crucial for real-time applications in self-driving cars.

C. Impact of the Semantic Map

The semantic map significantly enhances the CNN's ability to understand and differentiate between various input features. The ablation study results, presented in Table VI, clearly show the improvement in prediction accuracy when the semantic map is included.

TABLE VI
Results of Ablation Study

Model type	ACC	AUC	F1 score	Precision	Recall	Average
Without Semantic Map	0.52	0.51	0.57	0.57	0.58	0.55
With Semantic Map	0.64	0.55	0.78	0.68	0.99	0.72

The results clearly demonstrate the positive impact of incorporating the semantic map into our model. The significant improvements across all performance metrics underscore the importance of the semantic map in enhancing the model's predictive capabilities. The semantic map helps the model to better understand the spatial and contextual relationships between features, leading to more accurate and reliable predictions. This makes our proposed approach not only faster but also more effective in real-world applications, such as predicting pedestrian intentions in self-driving cars.

D. Performance in Scenarios with Occlusions and Multiple Pedestrians

Scenarios with occlusions and multiple pedestrians present significant challenges for pedestrian intention prediction models. Occlusions can obscure key visual features needed for accurate pose estimation, while multiple pedestrians increase the complexity of the scene, requiring the model to distinguish between individuals and their respective actions.

Handling Occlusions:

Pose Estimation with Occlusions: When a pedestrian is

partially obscured, keypoint identification may be hindered. Our approach uses sequential frames to infer missing keypoints while taking temporal continuity into account. If keypoints are obscured in one frame, the model leverages information from prior and subsequent frames to forecast the missing points.

The semantic map contributes to the model's ability to maintain spatial awareness and contextual knowledge even when specific aspects are obscured, hence boosting its robustness to partial occlusions.

If the degree of blockage is higher than 15 consecutive frames (pedestrian observation time), the critical information for recognizing the pedestrian movement pattern is naturally lost, and the model performs poorly.

Handling Multiple Pedestrians:

Feature Separation: The input data structure, including bounding boxes and keypoint poses, ensures that features corresponding to different pedestrians are processed separately, reducing confusion and improving prediction accuracy. The input data to the model includes the bounding box, the speed of the moving vehicle, and the pose of the pedestrian's body for 15 consecutive frames, joined together in a specific structure and entered into the model. In other words, the desired model performs the prediction operation separately for each pedestrian, and the presence of several pedestrians does not create a problem for the model. It is assumed that the operation of pedestrian detection has been done before the operation of predicting the pedestrian crossing event with high accuracy, and the purpose of this article is not to detect and track pedestrians. Also, the JAAD dataset has prepared input data for the model for each pedestrian separately in the form of a sequence with 15 consecutive frames, solving the problem of multiple pedestrians in the scene at the same time.

Experimental Results:

To evaluate the model's performance in these challenging scenarios, we conducted experiments using a subset of the JAAD dataset specifically annotated for occlusions and scenes with multiple pedestrians.

TABLE VII
Results of Occlusion Scenarios

Model type	ACC	AUC	F1 score	Precision	Recall	Average
With Occlusion Scenarios	**0.63**	0.55	**0.76**	**0.66**	**0.91**	**0.70**
Without Occlusion Scenarios	**0.64**	0.55	**0.78**	**0.68**	**0.99**	**0.72**

Occlusion Scenarios: The model maintains an accuracy of 63% in scenarios with partial occlusions, demonstrating its ability to infer missing information from temporal data.

E. Sensitivity to Keypoint Pose Estimation Errors

The accuracy of pedestrian keypoint pose estimation is crucial for the performance of our proposed model. Since the keypoints are extracted using the OpenPose model, it is important to understand how errors in this estimation process impact the prediction of pedestrian crossing intentions.

Impact of Keypoint Pose Estimation Errors:

To evaluate the sensitivity of the model to pose estimation errors, we conducted a series of experiments where varying levels of Gaussian noise were added to the keypoint coordinates. The performance of the model was then assessed under these conditions.

TABLE VIII
Results of Sensitivity Analysis

Model type	ACC	AUC	F1 score	Precision	Recall	Average
Without Noise	**0.64**	0.55	**0.78**	**0.68**	**0.99**	**0.72**
With Low Noise: Gaussian noise ($\mu = 0$, $\sigma = 1$)	0.60	0.51	0.67	0.59	0.79	0.63
With High Noise ($\mu = 0$, $\sigma = 10$)	0.51	0.50	0.51	0.46	0.56	0.50

The sensitivity analysis demonstrates that the model's performance degrades as the noise level in keypoint pose estimation increases. This highlights the importance of accurate keypoint detection for reliable pedestrian intention prediction. Ensuring high-quality keypoint extraction is essential for maintaining the model's effectiveness in real-world applications. These results underscore the need for robust pose estimation techniques and possible noise reduction methods to enhance the overall performance of the proposed model.

V. DISCUSSION AND CONCLUSION:

In this paper, we proposed a novel method for predicting pedestrian crossing intentions using non-visual features. This method combines non-visual features and employs a semantic map to prepare structured data in the form of a 4-dimensional tensor for the CNN network. By leveraging the feature extraction capabilities of CNNs on data structured similarly to images and videos, the designed model outperforms previous models.

Our experimental results on the JAAD dataset demonstrate that the proposed method achieves superior performance in pedestrian action prediction evaluation metrics compared to existing methods. Additionally, our model exhibits better performance in terms of response speed, surpassing all existing models. The introduction of the semantic map theory, coupled with the new feature selection structure, contributes to the model's robustness and accuracy.

The strengths of our model lie in its innovative feature extraction approach, the effectiveness of the semantic map, its high response speed, and its overall good and acceptable performance. These attributes make our model a valuable contribution to the field of pedestrian intention prediction in self-driving cars.

A. Limitations of the JAAD Database:

While the JAAD dataset is comprehensive in capturing urban pedestrian interactions, it has several limitations that could affect the generalizability of our findings:

- Limited Geographic Diversity: The dataset primarily consists of videos recorded in specific urban environments. This narrow geographic focus may not represent the wide range of conditions self-driving cars will encounter globally, such as in rural areas or different cultural settings.
- Weather and Lighting Conditions: The dataset includes limited variations in weather and lighting conditions. Real-world scenarios often involve challenging conditions such as heavy rain, fog, nighttime driving, and intense sunlight, which are underrepresented in the JAAD dataset.
- Pedestrian Behavior Variability: The behaviors recorded in the dataset might not fully encompass the diversity of pedestrian interactions worldwide. This includes variations due to cultural differences in crossing behavior, jaywalking, and interactions with other road users, which are crucial for developing robust and generalizable pedestrian intention prediction models.

Incorporating Environmental Factors

Environmental factors, such as traffic signals and weather conditions, play a crucial role in pedestrian behavior and decision-making processes. Integrating these factors into our model can provide a more comprehensive understanding of the context in which pedestrian actions occur, thereby improving prediction accuracy.

By acknowledging these limitations, we can better understand the context of our findings and identify areas for future research to enhance the robustness and applicability of pedestrian intention prediction models in diverse real-world scenarios.

B. Generalizability to Real-World Scenarios:

While our model shows improved performance on the JAAD dataset, it is crucial to consider its applicability to real-world scenarios with higher variation. We have expanded the discussion on how our findings might generalize to real-world scenarios, considering the limitations of the dataset. This section also suggests future work to enhance the robustness and applicability of our model.

- Dataset Augmentation: To improve generalizability, future work could augment the training data with synthetic examples that simulate diverse weather, lighting, and geographic conditions.
- Field Testing: Extensive field testing in varied environments is essential to validate the model's performance and make necessary adjustments based on real-world feedback.

C. Model adaptability:

Adaptability of the Model to Predict Other Pedestrian Actions The architecture of our proposed model is designed to be flexible and extensible, making it suitable for predicting a range of pedestrian actions beyond crossing intentions. The key components of the model, such as the convolutional neural network (CNN) and the semantic map, can be adapted to learn and recognize different types of pedestrian behaviors.

- Walking Straight: To predict whether a pedestrian will continue walking straight, additional features such as the pedestrian's trajectory and orientation over time can be incorporated into the model.
- Waiting: Predicting if a pedestrian will wait can involve analyzing stationary periods and body language cues, which can be integrated into the input features.
- Other Actions: The model can be extended to predict other actions, such as turning, stopping, or interacting with other pedestrians, by including relevant contextual information and motion patterns.
- Required Modifications: To adapt the model for these predictions, the following modifications may be necessary:
- Additional Training Data: Collecting and annotating datasets that include various pedestrian actions and their corresponding labels.
- Feature Engineering: Introducing new features that capture different aspects of pedestrian behavior, such as velocity vectors, interaction with the environment, and temporal patterns.
- Model Training: Retraining the model with the extended dataset and features to learn the new action classes.

VI. REFERENCES

[1] Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017). 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* (pp. 206-213).

[2] Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2020, October). Do they want to cross? Understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1688-1693). IEEE

[3] Lorenzo, J., Parra, I., Wirth, F., Stiller, C., Llorca, D. F., & Sotelo, M. A. (2020, October). Rnn-based pedestrian crossing prediction using activity and pose-related features. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1801-1806). IEEE.

[4] Rasouli, A., Kotseruba, I., Kunic, T., & Tsotsos, J. K. (2019). 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* (pp. 6262-6271).

[5] Quan, R., Zhu, L., Wu, Y., & Yang, Y. (2021). Holistic LSTM for pedestrian trajectory prediction. IEEE transactions on image processing, 30, 3229-3239.

[6] Fushishita, N., Tejero-de-Pablos, A., Mukuta, Y., & Harada, T. (2020). Long-term human video generation of multiple futures using poses. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16* (pp. 596-612). Springer International Publishing.

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

[8] Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2017, June). Agreeing to cross: How drivers and pedestrians communicate. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (pp. 264-269). IEEE.

[9] Fang, J., Wang, F., Xue, J., & Chua, T. S. (2024). Behavioral intention prediction in driving scenes: A survey. IEEE Transactions on Intelligent Transportation Systems.

[10] Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7291-7299)

[11] Chen, S., & Demachi, K. (2020). A vision-based approach for ensuring proper use of personal protective equipment (PPE) in decommissioning of Fukushima Daiichi nuclear power station. *Applied Sciences*, *10*(15), 5129.

[12] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458.*

[13] Kotseruba, I., Rasouli, A., & Tsotsos, J. K. (2021). Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1258-1268).

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

[15] Yang, D., Zhang, H., Yurtsever, E., Redmill, K. A., & Özgüner, Ü. (2022). Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, *7*(2), 221-230

[16] Rasouli, A., Kotseruba, I., & Tsotsos, J. K. (2020). Pedestrian action anticipation using contextual feature fusion in stacked rnns. *arXiv preprint arXiv:2005.06582*

[17] Ham, J. S., Kim, D. H., Jung, N., & Moon, J. (2023). 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 (pp. 3666-3675).

[18] Azarmi, M., Rezaei, M., Wang, H., & Glaser, S. (2024). PIP-Net: Pedestrian Intention Prediction in the Wild. arXiv preprint arXiv:2402.12810.

[19] Zhang, Z., Tian, R., & Ding, Z. (2023, June). Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 3, pp. 3534-3542).

[20] Zhou, Y., Tan, G., Zhong, R., Li, Y., & Gou, C. (2023). Pit: Progressive interaction transformer for pedestrian crossing intention prediction. IEEE Transactions on Intelligent Transportation Systems.

[21] Damirchi, H., Greenspan, M., & Etemad, A. (2023, October). Context-aware pedestrian trajectory prediction with multimodal transformer. In 2023 IEEE International Conference on Image Processing (ICIP) (pp. 2535-2539). IEEE.

[22] Li, Y., Zhang, C., Zhou, J., & Zhou, S. (2024). POI-GAN: A Pedestrian Trajectory Prediction Method for Service Scenarios. IEEE Access.

[23] Lv, Z., Huang, X., & Cao, W. (2022). An improved GAN with transformers for pedestrian trajectory prediction models. International Journal of Intelligent Systems, 37(8), 4417-4436.