ISBN 978-989-9121-69-0
70th ISTANBUL World Congress on Artificial Intelligence: Applications, Challenges & Impacts (WCAIACI-26)
Istanbul (Turkey) June 1-3, 2026

Integrating Pose, Scene Context, and Ego Dynamics for Robust Pedestrian Crossing Prediction

Andree Honoree Iradukunda[1] and John Olorunfemi Olaifa[1]

[1]Department of Computer Engineering, İstanbul Okan Üniversitesi, Tuzla Kampüsü, 34959 Akfırat, Tuzla, İstanbul, Türkiye

***Abstract*:** *Pedestrian intention prediction is critical for safe and reliable autonomous driving systems, particularly in complex urban environments characterized with sudden and unpredictable pedestrian-road interaction. In this study, we propose a hybrid deep learning framework for binary pedestrian intention classification (crossing vs. non-crossing) using the Joint Attention in Autonomous Driving (JAAD) dataset. The model integrates multi-modal contextual cues, including local visual features, global scene context, pedestrian bounding boxes, human pose keypoints, and ego-vehicle speed, to capture both spatial and behavioural dynamics influencing pedestrian decisions. Results from the study demonstrate consistently strong predictive performance in tests (0.87 accuracy, 0.94 AUC), and training to validation AUC score of 0.99 and 0.94 respectively. These findings suggest the significance of combining pedestrian motion dynamics with scene perception for prediction robustness enhancement in autonomous vehicles. The study framework is designed to be extensible, enabling modeling of more complex pedestrian behaviours beyond binary crossing decisions.*

Keywords: *Pedestrian Intention Prediction, Autonomous Vehicles, Multi-Modal.*

1. Introduction

Autonomous vehicles (AVs) amongst other modern technologies have highly benefitted from advancements in the field of image processing [1], [2]. Since they must operate safely in dynamic and unpredictable environments where interactions with pedestrians present one of its major challenges, a critical capability therefore is the accuracy of pedestrian intention prediction, particularly the ability to anticipate whether a pedestrian will cross the road or not [3]. Early and accurate prediction enables vehicles to make proactive decisions, reducing collision risks and improving overall traffic efficiency [4], [5].

In realistic traffic conditions, crossing intention of pedestrians are not influenced by a single factor, but rather by a combination of factors and observations [6]. While pedestrian-centric factors such as bounding boxes and pose are crucial for the outcome of a crossing prediction, spatial-centric factors such as ego vehicle speed, traffic sign and lights have equally been demonstrated to hold equally significant information, and as such, the synthesis of several parameters and their relationships is indeed crucial for deployable prediction. Even though spatial information is gathered in a noise-dense environment, identifying and filtering features that shape pedestrian behaviour presents a challenge [7].

The JAAD Dataset has become a benchmark for studying pedestrian behaviour in urban environments, providing rich annotations of pedestrian actions, contextual cues, and driver-pedestrian interactions [4], [8], [9], [10]. Prior research using JAAD has explored a variety of approaches, including appearance-based methods, trajectory modelling, and temporal sequence learning. While these methods have achieved reasonable performance, many rely on limited modalities, often focusing on either visual cues (e.g., bounding boxes or pose) or motion dynamics in isolation [11]. However, pedestrian intention is inherently multi-factorial, influenced not

only by individual motion and posture but also by environmental context and the state of the ego vehicle. For instance, a pedestrian's pose may indicate readiness to cross, whereas, this intention is modulated by traffic conditions, road layout, and vehicle speed. Existing approaches often fail to fully capture these interdependencies and intersection of influence, leading to reduced robustness in complex or ambiguous scenarios.

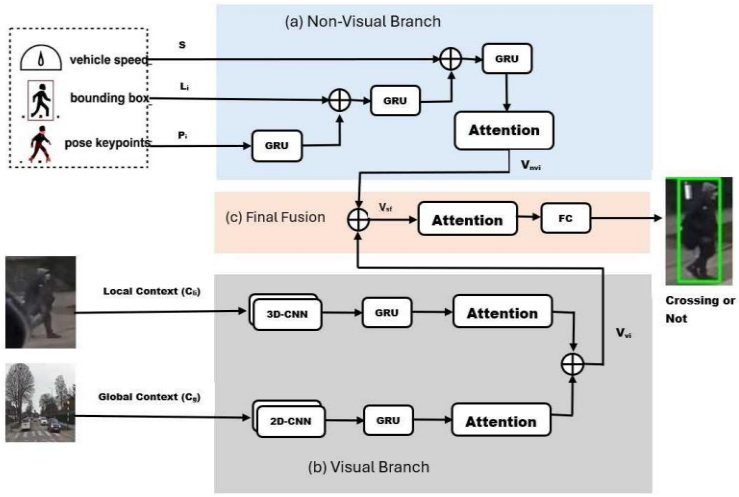

Fig. 1: Internal representation of the proposed pedestrian crossing behaviour prediction model.

To address these limitations, we propose a hybrid multi-modal framework (Fig. 1) that integrates local visual features, global scene context, pedestrian bounding boxes, human pose keypoints, and ego-vehicle speed for binary intention classification (crossing vs. non-crossing). By jointly modelling these heterogeneous signals, the proposed method captures both behavioural cues and interaction dynamics, enabling more reliable prediction of pedestrian intent. Unlike prior works that emphasize single or dual modalities, our approach highlights the importance of comprehensive feature fusion, resulting in balanced performance across classes and improved resilience to ambiguous cases. Our paper contributes the following:

- We propose a multi-modal system that employs the joint use of human pose keypoints, global scene context, pedestrian bounding-box trajectory, local visual features, and ego-vehicle speed.
- We adopted a hybrid architecture of 2D, and 3D Convolution Neural Networks CNNs with Gated Recurrent Unit GRU that can be easily generalized, and is applicable to AV implementation requirements.

The rest of the article is organized as follows. Section 2 contains a review of existing work on pedestrian's crossing intention prediction. Section 3 describes the implementation details and method of the study, including a thorough description of the hybrid deep learning model. We describe the experimental setup and results in section 4. Conclusion and future work are presented in section 5 to conclude the study.

2. Related Works

Pedestrian intention prediction has gained significant attention in recent years, particularly in the context of autonomous driving where anticipating pedestrian behaviour is essential for safe navigation. Existing approaches can be broadly categorized into context-based methods, trajectory-based models, pose-driven approaches, and multi-modal fusion techniques.

2.1. Context-Based Intention Prediction

Early work on pedestrian intention prediction largely relied on contextual and visual cues derived from the JAAD Dataset, introduced in [8]. These methods utilize features such as bounding boxes, pedestrian location, and scene attributes (e.g., traffic signals, crosswalk presence) to infer crossing behaviour. While effective in

structured environments, such approaches often fail to capture fine-grained behavioural cues, as they treat pedestrians as coarse objects rather than dynamic agents.

Subsequent extensions, including the work of Rasouli in [12] and [13], incorporate richer annotations and temporal information. However, even these models primarily emphasize scene context and temporal progression, with limited integration of detailed human motion or ego-vehicle dynamics.

2.2. Trajectory-Based Models

Trajectory prediction methods focus on modeling pedestrian motion patterns over time, often using recurrent neural networks such as LSTMs. A notable example is [14], which captures interactions among multiple agents in crowded environments. A trajectory prediction framework was also proposed in [15], in which pedestrian motion intentions and social interaction with immediate surroundings are modelled to forecast future trajectory. A pedestrian crossing knowledge graph with Bayesian inference was constructed in [16] for intention estimation, which in turn are used for trajectory prediction using a Bi-LSTM-Attention trajectory prediction network.

Generally, these approaches demonstrate strong performance in predicting future positions but are inherently limited in early intention prediction, as they depend heavily on observable motion history. In the context of JAAD, trajectory-based methods typically leverage past pedestrian movement to infer crossing decisions. However, such approaches struggle in static or pre-motion scenarios, where pedestrians exhibit minimal movement before initiating a crossing. This limitation highlights the need for incorporating additional behavioural and contextual cues beyond trajectory alone.

2.3. Pose-Based Approaches

Human pose estimation has emerged as a valuable signal for understanding pedestrian intent, as body posture and orientation often provide early indicators of action [17], [18], [19]. A pedestrian crossing intention prediction framework that leverages pose estimation and body posture analysis was proposed in [18] to identify the willingness of pedestrians to cross. Graph-based approaches like Skeleton-Based Action Recognition with Spatial Temporal Graph Convolutional Networks have been employed to capture spatial-temporal relationships between keypoints as done by the authors in [20] and [21]. The model proposed in [20] used skeletal joint dynamics in conjunction with temporal fusion to identify the intentions of pedestrians.

While pose-based methods improve sensitivity to subtle behavioural cues, they often operate in isolation, without considering environmental context or vehicle interaction. As a result, their performance may degrade in real-world scenarios where intention is influenced by external factors such as traffic flow and road conditions

2.4. Multi-Modal Fusion Methods

Recent work has explored multi-modal approaches that combine visual, contextual, and temporal features. Attention-based frameworks, such as Multi-Modal Pedestrian Intention Prediction Using Attention Mechanisms [22], attempt to dynamically weigh different inputs to improve prediction accuracy. Other studies focus on modelling interactions between pedestrians and their surroundings.

Despite these advancements, most multi-modal methods remain limited in two key aspects. First, they often exclude ego-vehicle dynamics, which play a crucial role in shaping pedestrian decisions. Second, the integration of modalities is frequently shallow or loosely coupled, reducing the ability to capture complex interdependencies between behavioural, contextual, and motion cues.

2.5. Research Contributions

It is evident that most existing approaches either rely on single modalities (e.g., trajectory or pose), or employ partial multi-modal fusion without fully capturing interaction dynamics. To meet this need, the proposed approach introduces a comprehensive hybrid framework that jointly integrates local visual features, global scene context, pedestrian bounding boxes, human pose keypoints, and ego-vehicle speed. This unified representation

enables the model to capture both intrinsic behavioural signals and extrinsic interaction factors, addressing key limitations in prior work.

3. Method

3.1. Overview of the Proposed Framework

We propose a hierarchical hybrid deep learning framework for pedestrian behaviour prediction. The proposed methodological framework follows a multi-stage pipeline that begins with dataset selection and preprocessing, followed by spatiotemporal feature extraction, hybrid model construction and rigorous evaluation. First, video sequences from the JAAD dataset, are preprocessed to generate synchronized multimodal inputs including pedestrian-centered visual clips (local context), semantic scene context (global context), pose key-points, bounding boxes and ego vehicle speed. Next, to balance spatiotemporal representation quality and computational efficiency, the proposed methodological framework employs 3D CNN for local context extraction capturing pedestrian motion and appearance cues while using 2D CNN for global semantic context. These features together with Gated Recurrent Unit (GRU) and attention mechanism can make a robust framework to model evolution of pedestrian behaviours.

The vision branch of the system as described in Fig. 1 represents the fusion of visual features (local and global contexts). 3D Convolutional Neural Networks was used to capture motion cues from local context. 2D Convolutional Neural Networks was used to capture the global pedestrian surroundings from global context input. These CNNs were encoded by GRUs then passed into attention. The outputs of these two features are concatenated to provide the vector of visual inputs. The non-vision branch represents the fusion of non-visual features (ego-vehicle speed, pedestrian bounding box and pedestrian pose key points) which was encoded by GRU and the output was fed into attention to get the vector of non-visual inputs. Finally, the outputs of these two branches were fused and passed into attention then fed into fully connected layers (FC) for final prediction.

The hybrid approach combines the strengths of CNNs (spatial representation learning) and RNNs (temporal reasoning), thereby addressing the limitations of single-modality models and some of multi-models designed to capture only pedestrian intention, action or future trajectory. The ultimate objective is to learn interpretable and generalizable pedestrian behaviour representations that improve intention prediction from predicting whether a pedestrian intends to cross or not to the prediction of irregular behaviours during crossing with better accuracy across diverse environments. This hybrid design preserves temporal without overwhelming memory resources. The methodological steps aligned with research objectives are as follows:

- Develop a data preprocessing pipeline for multimodal features for nuanced behaviours capturing during crossing. This is implemented via dataset processing and annotation synchronization.
- Design a hybrid model combining visual and non-visual features. Achieved through 2D and 3D CNNs plus GRU with attention fusion.
- Build a chaining hybrid framework that operates effectively in both structured and unstructured urban environments and evaluate predictive performance across multiple behaviours. Validated using training vs validation plots, ROC-AUC, confusion matrices, and class-wise F1-scores.

The framework integrates multi-modal inputs, combining visual and non-visual signals to capture the complex and context-dependent nature of pedestrian decision-making. This design enables the model to jointly reason about what the pedestrian is doing and how the surrounding environment and vehicle dynamics influence that behaviour.

3.2. Dataset Processing

We selected the robust JAAD dataset for the study since it is publicly available, covers a wide demographic for generalization, and has been rigorously tested in similar studies. The JAAD dataset provides videos from which frames can extracted for further processing. We processed the video files into 128 x 128 pixels image frames for computational efficiency, and temporal consistency was established by smoothing all behaviours

using sliding windows of 16 consecutive frames to avoid noisy one-frame annotations. Trajectory based verification was also conducted to confirm motion direction (lateral, stopped or reversed) by evaluating JAAD's bounding box coordinates. DeepLabV3-based semantic segmentation maps were generated to capture environmental context. (e.g: sidewalks, roads, vehicles). A pre-trained OpenPose model was used to extract 18-keypoint human poses (target pedestrian pose key points), and ego-vehicle speed was parsed from JAAD annotations. Label merging was done by assigning each 16-frame clip a label based on the dominant behaviour within the last four frames. For example, if majority frames reflected to crossing, the label for the clip was set to crossing. Manual validation was also conducted by visually classifying random samples from each behaviour class to ensure semantic accuracy.

Annotations were synchronized using official JAAD XML metadata, ensuring temporal alignment between video frames, bounding boxes, pedestrian IDs, and contextual attributes. Label smoothing was applied for sequences where behaviours transitioned gradually between categories.

4. Experiments

To efficiently train the proposed hybrid multimodal model, coordinated optimization strategies ensuring convergence across the dual-branch architecture was necessary. The first branch of the hybrid model processed the temporal features (speed, bounding boxes, pose) while the complementary branch evaluated visual (3D CNN + 2D CNN) features. Because these submodules learn at different abstraction levels, a unified training framework was adopted using TensorFlow/Keras with mixed-precision optimization on an NVIDIA GPU. 70% of the obtained frames were used for the model training, while 15% were used during hyper-parameter tuning during validation. The remaining 15% were used for the model testing. Pedestrian identities were mutually exclusive across the splits to prevent identity leakage and maintain generalization integrity.

4.1. Experiment results

To validate the effectiveness of the proposed model, it was benchmarked against several state-of-the-art and classical models trained to predict pedestrian crossing intention over same dataset. See Table I for the baseline model listing.

TABLE I: Results Benchmarking

Baseline models	Description
PCPA [8]	Attention-based pedestrian crossing prediction using scene and pose features.
VGG-GRU [23]	Combines VGG16 features with recurrent temporal modeling.
Self-Attention Weather-Robust model (SAWR) [24]	Incorporates adaptive attention for varying environmental conditions.

Using standard performance evaluation metrics such as Accuracy, Area under Curve (AUC), Precision, Recall, and F1 Score, we compared the results of the proposed model with the models in Table I.

TABLE II: Performance Comparison of Model results with state of the art, and classical models.

Metric	PCPA	VGG-GRU	SAWR	**Proposed**
Accuracy	0.76	0.83	0.83	**0.87**
AUC	0.79	0.82	0.80	**0.94**
Precision	0.41	0.51	0.47	**0.87**
Recall	0.83	0.81	0.82	**0.87**
F1 Score	0.55	0.63	0.62	**0.87**

The proposed model exhibits convincing performance improvements when compared to the existing ones. Considering the best performances amongst the existing models metric-wise, the proposed model outperforms the tied performance of VGG-GRU and SAWR by 0.04 in accuracy, outperforms VGG-GRU by 0.12 in area under curve, it again exceeds the performance of VGG-GRU in precision by 0.36, it exceeds the recall performance of PCPA by 0.04, and finally when compared to VGG-GRU again, it outperforms it by 0.24. Comparative results of the experimental outcomes are provided in Table II, while percentage difference in the performance over the metrics can be seen in Fig. 2.

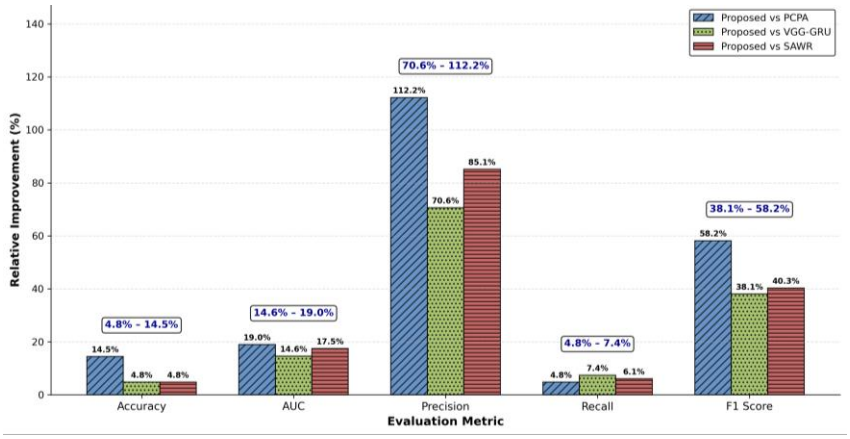

Fig. 2: Percentage improvement demonstrated by the proposed model against existing results, with ranges of improvement per metric.

5. Conclusiosn and Future Work

In this paper, we have introduced a hybrid multi-model pedestrian intention prediction model that incorporates several visual and non-visual features. The results (4.8%-14.5% improvement in Accuracy, 14.6%-19.0% improvement in AUC, 70.6%-112.2% improvement in Precision, 4.8% to 7.4% improvement in Recall, and 38.1-58.2% improvement in F1 score) have demonstrated that the proposed study is indeed capable of accurately predicting the crossing intention of vulnerable road users, thus promoting safe integration of autonomous vehicles in complicated urban environment. Several avenues for expanding the coverage of the work still exist and a few have been identified. Currently, we are extending the work to predict jockeying and renegade intentions of pedestrians in structured and unstructured crossing environments, which are more fine-grained pedestrian behaviours. Other natural extensions to the work include investigating the effectiveness of the algorithm in adverse weather with and without lidar and radar integration, evaluation of the computational efficiency of the hybrid algorithm, and vehicle deployment of the algorithm using available Open Innovation Autonomous Vehicle (OPINA) facility.

6. References

[1] K. Elgazzar *et al.*, "Revisiting the internet of things: New trends, opportunities and grand challenges," *Front. Internet Things*, 2022, doi: 10.3389/friot.2022.1073780.

[2] C. G. Keller and D. M. Gavrila, "Will the pedestrian cross? A study on pedestrian path prediction," *IEEE Trans. Intell. Transp. Syst.*, 2014, doi: 10.1109/TITS.2013.2280766.

[3] C. Luetge, "The German Ethics Code for Automated and Connected Driving," 2017. doi: 10.1007/s13347-017-0284-0.

[4] J.-S. Ham *et al.*, "Omnipredict: GPT-4o enhanced multi-modal pedestrian crossing intention prediction," in *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024.

[5] T. W. Wang and S. H. Lai, "Pedestrian Crossing Intention Prediction with Multi-Modal Transformer-Based Model," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023*, 2023. doi: 10.1109/APSIPAASC58517.2023.10317161.

[6] I.-H. Kao and C.-Y. Chan, “Impact of posture and social features on pedestrian road-crossing trajectory prediction,” *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–16, 2022.

https://doi.org/10.3389/friot.2022.1073780

[7] S. Zhang, X. Chen, W. Xu, L. Yang, and J. Yang, “Evidential Multimodal Fusion Network for Trusted Pedestrian Crossing Intent Prediction,” *IEEE Trans. Comput. Soc. Syst.*, 2025, doi: 10.1109/TCSS.2025.3576113.

[8] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, “Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior,” in *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, 2018. doi: 10.1109/ICCVW.2017.33.

[9] Y. He, Y. Sun, Y. Cai, C. Yuan, J. Shen, and L. Tian, “Multi-modal Pedestrian Trajectory Prediction based on Pedestrian Intention for Intelligent Vehicle,” *KSII Trans. Internet Inf. Syst.*, 2024, doi: 10.3837/tiis.2024.06.008.

[10] X. Zhai, Z. Hu, D. Yang, L. Zhou, and J. Liu, “Social Aware Multi-modal Pedestrian Crossing Behavior Prediction,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2023. doi: 10.1007/978-3-031-26316-3_17.

[11] S. Gazzeh, L. Lo Presti, A. Douik, and M. La Cascia, “Context-ped: multi-modal context fusion for pedestrian crossing intention prediction,” *Mach. Vis. Appl.*, 2025, doi: 10.1007/s00138-025-01751-3.

[12] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, “PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019. doi: 10.1109/ICCV.2019.00636.

[13] A. Kalatian and B. Farooq, “A context-aware pedestrian trajectory prediction framework for automated vehicles,” *Transp. Res. Part C Emerg. Technol.*, 2022, doi: 10.1016/j.trc.2021.103453.

[14] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human trajectory prediction in crowded spaces,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. doi: 10.1109/CVPR.2016.110.

[15] K. Chen, X. Song, and X. Ren, “Modeling social interaction and intention for pedestrian trajectory prediction,” *Phys. A Stat. Mech. its Appl.*, 2021, doi: 10.1016/j.physa.2021.125790.

[16] J. Zhou, X. Bai, W. Fu, B. Ning, and R. Li, “Pedestrian intention estimation and trajectory prediction based on data and knowledge-driven method,” *IET Intell. Transp. Syst.*, 2024, doi: 10.1049/itr2.12453.

[17] R. Quintero, I. Parra, D. F. Llorca, and M. A. Sotelo, “Pedestrian path prediction based on body language and action classification,” in *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, 2014. doi: 10.1109/ITSC.2014.6957768.

[18] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, “Pedestrian Crossing Intention Prediction at Red-Light Using Pose Estimation,” *IEEE Trans. Intell. Transp. Syst.*, 2022, doi: 10.1109/TITS.2021.3074829.

[19] J. Ma and W. Rong, “Pedestrian Crossing Intention Prediction Method Based on Multi-Feature Fusion,” *World Electr. Veh. J.*, 2022, doi: 10.3390/wevj13080158.

[20] F. Piccoli *et al.*, “FuSSI-Net: Fusion of Spatio-temporal Skeletons for Intention Prediction Network,” in *Conference Record - Asilomar Conference on Signals, Systems and Computers*, 2020. doi: 10.1109/IEEECONF51394.2020.9443552.

[21] X. Zhang, P. Angeloudis, and Y. Demiris, “ST CrossingPose: A Spatial-Temporal Graph Convolutional Network for Skeleton-Based Pedestrian Crossing Intention Prediction,” *IEEE Trans. Intell. Transp. Syst.*, 2022, doi: 10.1109/TITS.2022.3177367.

[22] T.-W. Wang and S.-H. Lai, “Multi-Modal Pedestrian Crossing Intention Prediction with Transformer-Based Model,” *APSIPA Trans. Signal Inf. Process.*, 2024, doi: 10.1561/116.20240019.

[23] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and Ü. Özguner, “Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention,” *IEEE Trans. Intell. Veh.*, 2022, doi: 10.1109/TIV.2022.3162719.

[24] A. Elgazwy, K. Elgazzar, and A. Khamis, “Predicting Pedestrian Crossing Intentions in Adverse Weather With Self-Attention Models,” *IEEE Trans. Intell. Transp. Syst.*, 2025, doi: 10.1109/TITS.2024.3524117.'