



Pedestrian Crossing Intention Prediction: An AI-Powered System for Pedestrian Crossing Intention Prediction Using Onboard Camera Vision and Deep Learning Methods

P. Swathika^{1*} , B. Vinothkumar² 

¹ Department of Artificial Intelligence and Data Science, Mepco Schlenk Engineering College (Autonomous), Sivakasi 626005, India

² Department of Computer Applications, Mepco Schlenk Engineering College (Autonomous), Sivakasi 626005, India

Corresponding Author Email: swathikap@mepcoeng.ac.in

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.430430>

ABSTRACT

Received: 12 June 2026

Revised: 16 August 2026

Accepted: 24 August 2026

Available online: 31 August 2026

Keywords:

cross-modal transformer, intention prediction, onboard camera, pedestrian behavior, pedestrian safety, spatial-temporal graph convolutional network, traffic environment

Ensuring pedestrian safety is important in autonomous driving and intelligent traffic monitoring systems for collision prevention in urban environments. This work proposes a hybrid deep learning system for real-time pedestrian crossing intention prediction (PCIP) using onboard camera perception. Addressing the limitations of prior works that rely solely on bounding boxes or skeletal poses, our core novelty lies in integrating pedestrian-vehicle distance as an explicit interaction cue alongside bounding box localization, vehicle velocity, and environmental visual features. The proposed architecture combines a Spatio-Temporal Graph Convolutional Network (ST-GCN) for modeling spatial-environmental topologies, a feed-forward multilayer perceptron (FF-MLP) for velocity/distance feature alignment, a Cross-Modal Transformer Network to capture long-range inter-modality dependencies, and a Bottleneck Feature Fusion (BFF) module to condense discriminative features. Evaluated on the benchmark PIE dataset and real-world urban video sequences, the framework achieves an accuracy of 0.93, precision of 0.87, and F1-score of 0.88, demonstrating superior predictive capabilities in complex, multi-pedestrian, and unsignalized traffic scenarios.

1. INTRODUCTION

Pedestrian safety is a critical concern in urban traffic management, as pedestrian-related traffic accidents contribute significantly to road fatalities worldwide. The rapid increase in vehicular traffic, driver inattention, and unpredictable pedestrian behavior make pedestrians particularly vulnerable, especially in unsignalized and densely populated urban environments. Accelerating vehicles, limited visibility, and uncertain pedestrian movements can increase the risk of pedestrian-vehicle collisions. Therefore, intelligent road safety systems capable of anticipating pedestrian crossing intentions are essential for providing sufficient reaction time to human drivers and autonomous vehicles. Conventional pedestrian safety mechanisms, including traffic signals, crosswalks, warning signs, and braking systems, primarily rely on predefined rules or reactive responses. Such mechanisms may be insufficient when pedestrian behavior is uncertain or changes rapidly. In this context, pedestrian action recognition and pedestrian crossing-intention prediction represent two related but distinct tasks. Action recognition determines the current state of pedestrian movement, such as whether a pedestrian is crossing at time t , whereas crossing-intention prediction aims to determine whether the pedestrian is likely to cross the road at a future time step. Existing action-recognition approaches have been investigated for understanding pedestrian behavior in traffic scenes [1]. However, recognizing the current action may provide limited

reaction time for a vehicle approaching a pedestrian. Predicting the intention before the crossing action occurs can therefore provide additional time for preventive decision-making.

Existing pedestrian intention prediction methods have explored different types of pedestrian representations. Skeleton- and keypoint-based approaches use human body configurations and temporal posture changes to infer pedestrian behavior [2]. Trajectory-based approaches model the temporal evolution of pedestrian movement and have demonstrated the importance of historical motion patterns for predicting future pedestrian behavior [3]. However, trajectory-based prediction generally requires sufficient observations of pedestrian motion over time, while skeleton-based approaches depend on reliable pose estimation. Both representations primarily describe the pedestrian and may not sufficiently capture the broader spatial relationship between the pedestrian and surrounding traffic elements.

Environmental context is also an important factor in pedestrian crossing decisions. The interaction among pedestrians, vehicles, traffic signals, crosswalks, and other scene elements can provide valuable contextual information for intention prediction. Graph-based approaches have therefore been investigated to represent spatial and temporal relationships among traffic participants and environmental elements. For example, Scene-STGCN incorporates environmental elements into a spatiotemporal graph representation for Pedestrian Intention Estimation (PIE) [4].

Such graph-based representations can preserve the structural relationships among objects in a traffic scene and provide richer contextual information than pedestrian-centric representations alone.

Another important factor in pedestrian–vehicle interaction is the relative proximity between the pedestrian and an approaching vehicle. Although existing approaches have incorporated pedestrian trajectories, bounding boxes, skeletal information, scene context, and vehicle motion, pedestrian–vehicle distance has not been consistently modeled as an explicit complementary feature. Similarly, vehicle velocity provides important information about the dynamic state of the approaching vehicle and the potential risk associated with a pedestrian crossing. Combining environmental context with pedestrian–vehicle proximity and vehicle motion can therefore provide a more comprehensive representation of the interaction between pedestrians and vehicles.

Considering these challenges, this work aims to predict pedestrian crossing intention by jointly analyzing the surrounding traffic environment, pedestrian information, pedestrian–vehicle distance, and ego-vehicle velocity. The proposed framework is designed for both signalized and unsignalized crossing scenarios and does not depend exclusively on predefined traffic elements or marked crossing locations. This makes the framework applicable to relatively unstructured traffic environments where pedestrian behavior can be uncertain. To achieve this objective, we employ a Spatio-Temporal Graph Convolutional Network (ST-GCN) to extract spatial and temporal information from the local traffic environment [4]. A feed-forward multilayer perceptron (FF-MLP) is used to encode the pedestrian–vehicle distance and ego-vehicle velocity features. Subsequently, a Cross-Modal Transformer is employed to model relationships between the environmental representation and the velocity–distance representation. Transformer-based cross-modal modeling has demonstrated the capability to learn relationships among heterogeneous pedestrian and vehicle features [5]. In the proposed framework, this mechanism enables the model to jointly reason about environmental context, pedestrian motion, vehicle motion, and pedestrian–vehicle proximity.

The main contributions of this work are summarized as follows:

Multimodal environmental and vehicle-interaction representation: Unlike approaches that primarily rely on pedestrian speed, skeletal information, trajectory, or bounding-box features, the proposed framework explicitly incorporates local environmental information together with pedestrian–vehicle distance and ego-vehicle velocity to improve pedestrian crossing-intention prediction.

Explicit pedestrian–vehicle proximity modeling: Pedestrian–vehicle distance is incorporated as an independent complementary feature to represent the spatial proximity between the pedestrian and the approaching vehicle. This provides interaction information that is not directly available from pedestrian-centric features alone.

Hybrid spatiotemporal and cross-modal architecture: The proposed framework integrates ST-GCN-based environmental feature extraction, FF-MLP-based velocity–distance feature encoding, and cross-modal Transformer-based feature fusion to jointly model spatial, temporal, and vehicle–pedestrian interaction information.

Bottleneck feature fusion (BFF) and multi-task learning (MTL): A BFF module is employed to obtain a compact representation from the extracted features, followed by MTL

for pedestrian crossing-intention classification and bounding-box regression. To clearly distinguish the complementary roles of the input representations, the proposed framework considers the following information:

Bounding box: Represents the pedestrian's image-space location, scale, and spatial extent.

Skeleton pose: Represents body configuration, posture, and fine-grained motion cues.

Environmental features: Represent the spatial and contextual relationships among pedestrians, vehicles, traffic signals, crosswalks, and other relevant scene elements.

Vehicle velocity: Represents the kinematic state of the ego vehicle.

Pedestrian–vehicle distance: Provides an explicit proximity cue describing the spatial relationship between the pedestrian and the approaching vehicle.

While bounding boxes and skeletal keypoints primarily describe pedestrian-centric characteristics, they do not directly represent the physical proximity between a pedestrian and an approaching vehicle. Similarly, environmental representations describe the surrounding scene but may not explicitly encode the dynamic risk associated with vehicle motion. Integrating pedestrian–vehicle distance with environmental graph representations and vehicle velocity therefore provides a more comprehensive interaction state. The cross-modal Transformer can subsequently learn the relationships among these heterogeneous features for improved pedestrian crossing-intention prediction.

The remainder of this paper is organized as follows. Section 2 presents a critical review of existing pedestrian crossing-intention prediction approaches. Section 3 describes the proposed methodology, including environmental feature extraction, velocity–distance feature extraction, cross-modal feature fusion, BFF, and MTL. Section 4 presents the experimental setup, evaluation metrics, comparative results, behavioral analysis, convergence analysis, and real-time evaluation. Section 5 concludes the paper and discusses future research directions.

2. RELATED WORK

The pedestrian crossing intention prediction (PCIP) framework plays a crucial role in enhancing pedestrian safety within intelligent traffic management systems. This PCIP framework aims to be proactive in predicting whether a pedestrian has an intention to cross the road or not in future seconds. In past years, various models have been proposed, utilizing different types of input data. The existing systems are categorised based on the type of input representation and modeling strategy adopted as follows.

2.1 Trajectory-based approaches

Trajectory-based approaches primarily model the temporal evolution of pedestrian movement and interactions with surrounding agents. Yang et al. [6] proposed an Interaction-Aware LSTM (IA-LSTM) that assigns different importance to pedestrians involved in interaction events, thereby improving the modeling of individual pedestrian motion. Rasouli [7] introduced a hierarchical fusion strategy that integrates pedestrian trajectory, ego-vehicle motion, and environmental context at different levels. Quan et al. [8] incorporated pedestrian motion, vehicle speed, and global scene

information using a Holistic LSTM architecture to account for vehicle-motion variations and pedestrian behavior. Alahi et al. [9] proposed Social-LSTM, which uses social pooling to model interactions among nearby pedestrians for trajectory prediction. These approaches demonstrate that temporal motion and interaction information are important for modeling pedestrian behavior. However, trajectory-based representations mainly describe historical movement and may provide limited structural and semantic information about the surrounding traffic environment. Although some methods incorporate ego-vehicle motion, explicit pedestrian-vehicle proximity is generally not modeled as an independent feature. This limitation motivates the integration of distance information with environmental and vehicle-motion features in the proposed framework.

2.2 Pose and skeletal-based approaches

Pose- and skeleton-based approaches represent pedestrian behavior using body keypoints, joint configurations, posture, and their temporal evolution. Kitchat et al. [10] proposed PedCross, which combines pose estimation with Random Forest and LSTM-based modeling to predict pedestrian crossing behavior. Yang et al. [11] developed a dual-channel pedestrian crossing-intention architecture that integrates spatiotemporal skeleton information with scene-interaction features. Gesnouin et al. [12] proposed TrouSPI-Net, a spatio-temporal attention-based deep learning model for pedestrian crossing intention prediction using skeletal motion. Ahmed et al. [13] proposed a multi-scale deep learning approach that exploits 3D skeletal joint information to capture spatio-temporal features for pedestrian intent prediction. While Ling et al. [14] proposed a skeleton-based spatiotemporal graph convolutional network with multiple attention mechanisms for fast pedestrian crossing intention prediction. Razali et al. [15] proposed a convolutional bottom-up multi-task approach for pedestrian intention prediction, jointly learning pedestrian-related features to improve crossing intention prediction. Zhang et al. [16] proposed a pedestrian crossing intention prediction approach based on pose estimation to predict pedestrians' crossing behavior at red-light intersections. Previous studies have explored pedestrian intention prediction using probabilistic models [17], 2D skeletal pose sequences and deep learning [18], multi-feature fusion of pedestrian and environmental information [19], and monocular 2D pose estimation for pedestrian and cyclist intention recognition [20].

The major strength of pose-based approaches is their ability to capture detailed posture and fine-grained motion cues that may not be available from bounding-box representations alone. However, their effectiveness depends on the reliability of pose estimation and can be affected by occlusion, viewpoint variation, illumination, and image resolution. More importantly, skeletal representations primarily describe the pedestrian's internal motion and do not directly encode the spatial relationship between a pedestrian and an approaching vehicle. Therefore, pose information can benefit from complementary environmental and vehicle-related features.

2.3 Graph-based approaches

Graph-based approaches model pedestrians, vehicles, and environmental elements as interconnected entities and are particularly suitable for representing spatial and temporal

interactions. Zhou et al. [21] proposed a graph-based framework for pedestrian crossing-intention prediction that jointly models pedestrian states and environmental dynamics using appearance and motion information. Ling et al. [22] introduced PedAST-GCN, which combines pedestrian pose keypoints, bounding boxes, and vehicle speed with spatial-temporal and modality attention mechanisms. Zhang et al. [23] proposed STCrossingPose, which employs ST-GCN to model spatial-temporal relationships in pedestrian skeleton sequences. Cadena et al. [24] developed Pedestrian Graph+, integrating pedestrian images, segmentation maps, and ego-vehicle velocity to provide contextual information for crossing prediction. Sofianos et al. [25] investigated a space-time-separable graph convolutional architecture for efficient spatial and temporal modeling, while Liu et al. [26] explored spatiotemporal relationship reasoning for pedestrian intent prediction.

2.4 Transformer-based approaches

Transformer architectures have recently been adopted for pedestrian crossing-intention prediction because of their ability to model long-range dependencies and interactions among heterogeneous modalities. Lorenzo et al. [27] developed CAPFormer, a Transformer-based architecture for pedestrian crossing-action prediction that exploits visual and contextual information. Transformer-based models have also been explored extensively for long-sequence temporal modeling [28].

The main advantage of Transformer-based approaches is their ability to learn relationships among heterogeneous features through attention mechanisms without relying exclusively on recurrent sequential processing. However, their effectiveness depends strongly on the quality and diversity of the input representations. Multimodal Transformer architectures can also increase computational requirements, particularly when multiple high-dimensional visual and temporal features are processed simultaneously. More importantly, existing approaches generally combine visual, pose, bounding-box, or ego-vehicle-motion information, whereas pedestrian-vehicle distance is not consistently incorporated as an explicit complementary feature. The proposed framework addresses this limitation by providing the Transformer with environmental representations together with explicitly modeled distance and vehicle-velocity features.

2.5 Other approaches

Several studies have explored multimodal architectures that combine pedestrian, vehicle, visual, and environmental information. Ham et al. [29] proposed a crossing-intention prediction framework that fuses multiple pedestrian and vehicle features using dedicated feature-fusion modules and attention-based recurrent modeling. Yao et al. [30] formulated pedestrian intention as an internal state and incorporated environmental information through relationship modeling. Ham et al. [31] developed a multi-stream architecture that separately processes pose, bounding-box, ego-vehicle speed, and contextual information before multimodal fusion.

Other approaches have investigated visual reasoning and multimodal interaction. Chen et al. [32] employed graph convolutional networks to perform visual reasoning for pedestrian crossing-intention prediction, while Singh and Suddamalla [33] proposed a multi-input fusion framework for

practical pedestrian intention prediction. Kotseruba et al. [34] introduced a benchmark for evaluating pedestrian action prediction, providing a basis for comparing different pedestrian behavior prediction approaches. Zhang et al. [35] investigated contextual feature fusion using stacked recurrent networks for pedestrian action anticipation. Piccoli et al. [36] explored the fusion of spatiotemporal skeletal features for pedestrian intention prediction.

Vision-based and probabilistic approaches have also been investigated. Saleh et al. [37] proposed a monocular RGB-based real-time pedestrian-intention prediction framework using spatiotemporal DenseNet, YOLOv3, and SORT. Gujjar and Vaughan [38] investigated anticipatory pedestrian action classification using predicted video representations. Kooij et al. [39] explored context-based pedestrian path prediction using a Dynamic Bayesian Network; Schulz and Stiefelwagen [40] proposed a latent-dynamic conditional random field model for recognizing pedestrian intentions from temporal behavioral patterns. Zhou et al. [41] developed the SEU_PML dataset and YOLO SOD detector for robust traffic participant detection in complex urban mixed-traffic environments.

Overall, existing studies demonstrate that trajectory, pose, visual, graph, and vehicle-motion representations provide complementary information for pedestrian behavior prediction. However, most existing approaches either focus on a particular representation or combine multiple modalities without explicitly modeling pedestrian–vehicle proximity as a dedicated feature. Graph-based approaches are effective in capturing spatial relationships among environmental elements, whereas Transformer-based methods provide effective multimodal feature interaction. Nevertheless, the joint integration of environmental graph representations, pedestrian–vehicle distance, and ego-vehicle velocity remains insufficiently explored.

In contrast, the proposed framework integrates environmental and pedestrian representations using ST-GCN, explicitly models pedestrian–vehicle distance and ego-vehicle

velocity using an FF-MLP, and subsequently combines these heterogeneous representations through a Cross-Modal Transformer and BFF module. This design aims to jointly capture scene context, pedestrian motion, vehicle dynamics, and pedestrian–vehicle proximity, thereby providing a more comprehensive representation for pedestrian crossing-intention prediction.

3. PROPOSED METHODOLOGY

In a real-time scenario, PCIP is a complex task that utilizes both the environmental scene and the distance/velocity of the ego vehicle.

3.1 Problem formulation

In this work, we formulate the pedestrian crossing intention as a maximization problem with the optimal solution as $p(C^{t+n}|LS, PB, D, V)$, where we focus on finding the crossing intention $C^{t+n} \in \{0, 1\}$ (i.e., 0 as not crossing, 1 as crossing) of pedestrian i in a future time step $t+n$. Here, the prediction depends on the local scene LS in the pedestrian environment as $\{ls^1, ls^2, ls^3 \dots ls^n\}$, the bounding box of the pedestrian PB as $\{pb^1, pb^2, pb^3 \dots pb^n\}$, distance D between the ego vehicle and the pedestrian as $\{d, d, d \dots d\}$ and the velocity V of the ego vehicle as $\{v, v, v \dots v\}$.

3.2 Architecture

We proposed a hybrid deep learning architecture model for processing the features (LS, PB, D, V) to provide an optimized and accurate crossing intention. Figure 1 elucidates the entire proposed architecture that is separated into three phases as follows:

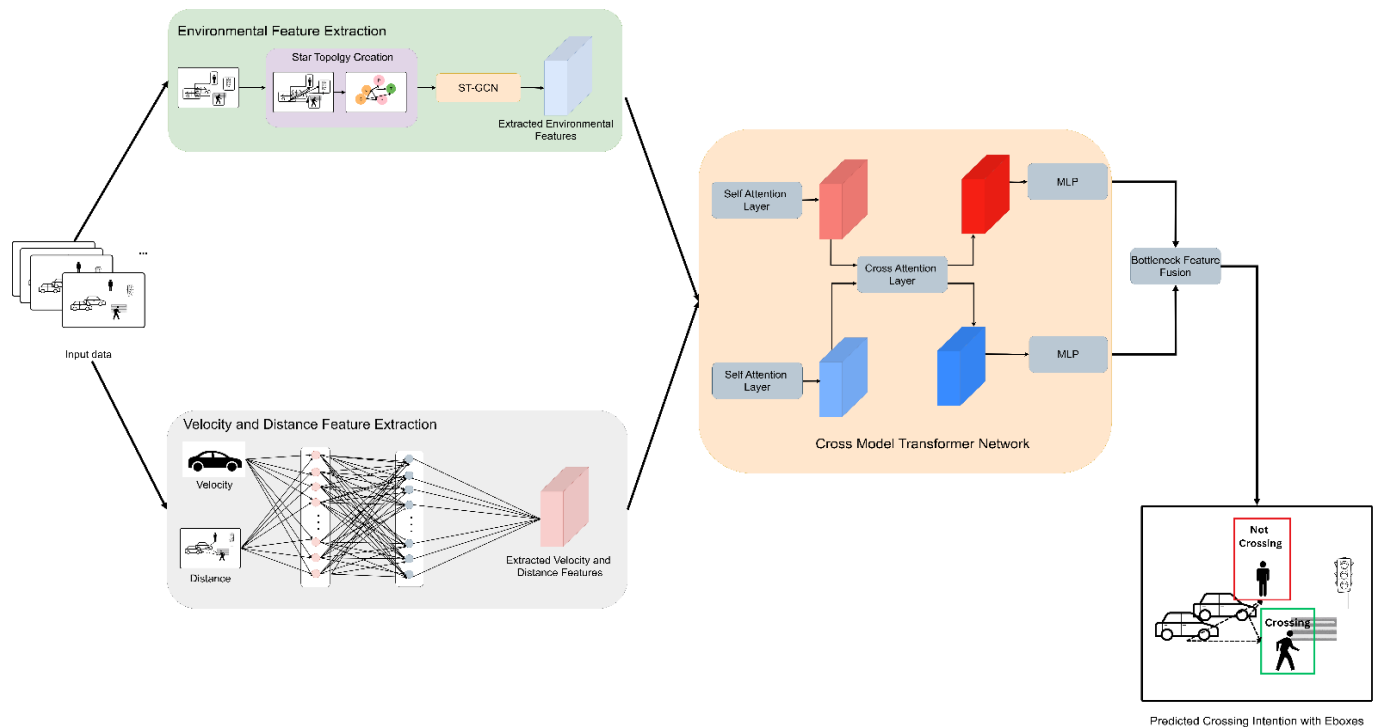


Figure 1. System architecture of the proposed Pedestrian Crossing Intention Prediction (PCIP) framework, which leverages hybrid deep learning approaches to predict pedestrian crossing intention

Environmental Feature Extractor: In this phase, ST-GCN is used to extract the local scene LS and the bounding box PB of the pedestrian, which provides the context of the surrounding environment.

Distance and Velocity Feature Extractor: This part of the work processes both the distance between the pedestrian and vehicle as well as the velocity of the vehicle through FF-MLP.

Cross-Model Transformer Network: Finally, this network processes both the extracted environment features (LS , PB) and the extracted distance/velocity (D , V) features through the attention layers.

3.3 Environmental feature extractor

Environmental features provide the contextual information of the scene that influences a pedestrian's decision to cross the road. In this part, we use the ST-GCN for extracting and processing the local scene elements (i.e., traffic signals, zebra crossing) and the bounding boxes of the pedestrian, which are used to get the key structure of the pedestrian, such as visual features.

3.3.1 Data preprocessing

In this phase, data preprocessing involves processing video frames that are captured from an onboard camera of a vehicle. It includes object detection and star topology creation, which are further provided to the ST-GCN model for extracting the environment features.

- Object Detection: For processing the video frames, object detection is the first part that localizes and detects the pedestrians, vehicles, and road elements in each frame. We use the YOLOv8 model [42] for this process that provide the bounding box of the scene elements, especially the pedestrian's bounding box which is used to extract the visual features and contextual features (i.e., appearance and bounding box scaling).

- Star Topology Creation: We use star topology which is a network structure where a single central node is connected to all other nodes and the surrounding nodes do not connect with each other, for preventing the redundant connections among the detected elements of the environment. The Star Topology Graph $G = (V, E)$ clarifies that the ego-vehicle is the central node (v_0) to avoid redundant $O(N^2)$ inter-pedestrian edges. In star topology, the approaching vehicle that is detected in the local scene acts as a central node, while other detected elements like pedestrians, other vehicles, and traffic signals act as surrounding nodes to process all contextual information efficiently in the local scene.

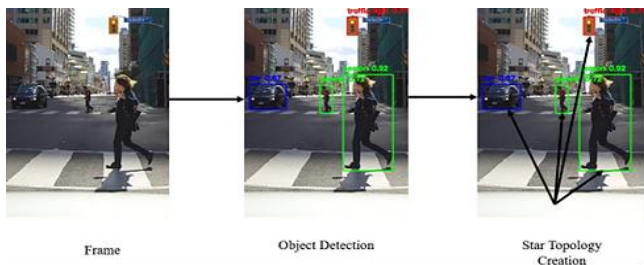


Figure 2. Visualization of the data preprocessing pipeline, from object detection to star topology creation

Figure 2 represents a visualization of the data preprocessing flow, starting from object detection to the creation of a star topology graph which connects the approaching car with

detected objects, including two pedestrians, one other vehicle, and one traffic signal within a frame. It doesn't connect the two pedestrians to each other to focus only on vehicle-pedestrian interaction.

3.3.2 Spatio-Temporal Graph Convolutional Network

ST-GCN is a deep learning model that captures both spatial and temporal relations of data. It is applied for sequence-based problems by incorporating the feature of temporal dynamics, which is extended from the traditional Graph Convolutional Networks (GCNs). Additionally, it preserves the correlations between connected items with the help of graph convolutional layers. It understands the motion patterns and forecasts feature behaviors of the pedestrian. ST-GCN is split into two convolutional layers as follows.

- Spatial Convolutional Layer: Extracts spatial information related to the positional and structural details of the elements in the environment, including their location, size and orientation of the local scene elements.

- Temporal Convolutional Layer: Maintains the dynamic behavioral changes of these elements which affect the pedestrian behavior over each time step, such as their motion, velocity, and trajectory across each subsequent frame.

By incorporating these layers, ST-GCN is able to analyze not only the static position of elements but also their dynamic characteristics, which are related to each other. ST-GCN overcomes traditional methods like CNNs or RNNs by combining these two layers. Because CNNs act as a spatial feature extractor, they work with grid-based data (2D images) and have trouble capturing associations between nodes, while RNNs process sequence-based data for temporal modelling; they have a struggle to understand long-term dependencies and are limited by vanishing gradient problems. The proposed ST-GCN model is driven by the Scene ST-GCN network approach presented in the work [4]. This model processes three inputs simultaneously: visual image tensor, adjacency matrix and boundary box within a frame.

- Visual image tensor: Achieves the local scene (LS) information which is a numerical representation of a video frame at a time step in an N-dimensional array.

- Adjacency matrix: Adapts the star topology of a video frame, where the central node is connected to all other nodes, representing pedestrians, vehicles, and traffic elements within the network.

- Bounding box: Localizes the pedestrians in the local scene and tracks their movement across the frames to understand their motion patterns in a consequent time. This feature is useful to estimate depth, predict the intention to cross, and analyze interactions with nearby vehicles in real-time.

For reducing the dimensionality of visual features, VGG16 is used as the feature extractor. The pre-trained VGG16 model takes each detected element of the scene and extracts high-level feature representations, discarding redundant or unnecessary details. Using a convolutional operation that passes through layers of VGG16, the network minimizes the appearance characteristics of objects to a lower-dimensional feature. The ST-GCN layers are then employed to extract both motion and contextual information, effectively modeling the interactions between the various entities in the scene. To account for temporal dependencies, an LSTM temporal encoder is used to track actions over time, ensuring the model captures dynamic changes in the scene. The final prediction, made by a fully connected network (FCN), classifies whether a pedestrian is crossing or not crossing, providing the decision

output based on the learned features and temporal context.

Algorithm 1. Environmental perception using Spatio-Temporal Graph Convolutional Network (ST-GCN)

Input:

XML data containing:

- Pedestrians (Pt)
- Vehicles (Vt)
- Traffic Lights (TLt)
- Crosswalks (CWt)
- Frame ID (Ft)

Output:

$y \leftarrow$ Pedestrian behavior classification (0 = Normal, 1 = Crossing)

$fenv \leftarrow$ Environmental feature representation

Procedure:

- 1: for each frame ft in XML data do
 - 2: Extract objects:
Pt, Vt, TLt, CWt
 - 3: Construct a star-topology graph
Sgraph = (N, E)
 - 4: Create graph nodes
 $N \leftarrow \{Pt, Vt, TLt, CWt\}$
 - 5: Create graph edges based on
spatial relationships, bounding-box overlap,
and contextual interactions
 - 6: Initialize adjacency matrix A
 - 7: for every pair of nodes (i, j) do
 - 8: if Edge(i, j) exists then
 - 9: $A(i, j) \leftarrow 1$
 - 10: else
 - 11: $A(i, j) \leftarrow 0$
 - 12: end if
 - 13: end for
 - 14: Generate node feature embeddings
 $e \leftarrow \text{Concatenate}(Pt, Vt, TLt, CWt)$
 - 15: Combine graph structure
 $e \leftarrow e \oplus A$
 - 16: Feed graph into ST-GCN
 - 17: $[y, fenv] \leftarrow \text{STGCN}(e)$
 - 18: Store y and fenv
 - 19: end for
 - 20: Return y, fenv
-

3.4 Distance and velocity feature extractor

The velocity and distance characteristics are necessary to compute the dynamic interaction between vehicles and pedestrians. In this phase, we use the feed-forward multilayer perceptron (FF-MLP) to process both vehicle velocity and pedestrian-vehicle distance in an efficient manner. The workflow of this phase is described below, detailing each step involved in the processing of the extracted features to ensure an accurate intention prediction.

3.4.1 Distance calculation

We determine the distance between pedestrians and vehicles as the novel key of our work, which is necessary to understand the intentions of the crossing. This distance is calculated through the Haversine distance between the approaching car and the nearest objects, including pedestrians, other vehicles, traffic lights using the detected bounding boxes from the YOLOv8 model. The Haversine distance is described

in Eqs. (1)-(3). In addition, it also provides an evaluation of the relative velocity of vehicles approaching a pedestrian who is very close to the vehicle.

$$a = \sin^2\left(\frac{\Delta\theta}{2}\right) + \cos(\theta_1)\cos(\theta_2)\sin^2\left(\frac{\Delta\lambda}{2}\right) \quad (1)$$

$$c = 2\text{atan2}(\sqrt{a}, \sqrt{1-a}) \quad (2)$$

$$d = R \cdot c \quad (3)$$

where, θ_1, θ_2 are Latitudes of Pedestrian and Vehicle (in radians); λ_1, λ_2 are Longitudes of Pedestrian and Vehicle (in radians) and R is Radius of the Earth (approximately 6371 km or 3958.8 miles)

3.4.2 Velocity calculation

The PIE dataset offers rich ego-vehicle motion and spatial orientation data, which is vital for modeling actual urban driving scenarios and analyzing pedestrian behavior in context. The frames in the dataset are annotated with high-resolution data obtained from the vehicle's On-Board Diagnostics (OBD) system and external positioning systems. The vehicle's speed is directly measured from the internal OBD sensors, and GPS-speed, derived from satellite geolocation information. In addition to linear motion, the dataset records the ego vehicle's orientation through heading angle, pitch, and roll measurements, providing a comprehensive representation of the vehicle's motion and orientation. These attributes define the directional direction and angular offsets of the vehicle, which are required to model the vehicle's relative position and orientation in dynamic traffic situations. This set of annotations enables the development of multi-modal models that combine vision and motion signals to improve the robustness of PCIP.

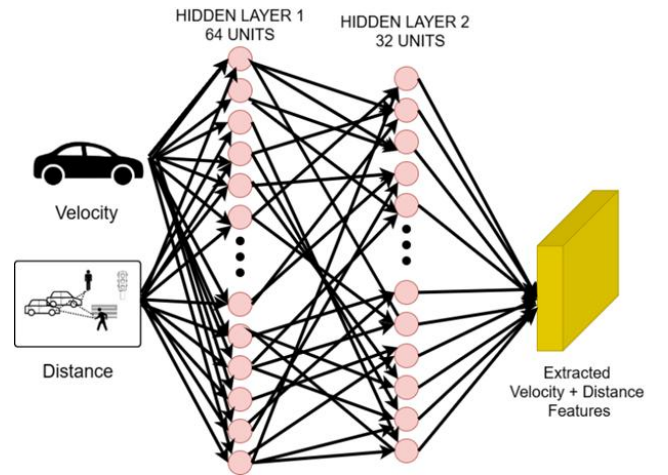


Figure 3. Proposed feed-forward multilayer perceptron (FF-MLP) model structure to extract the velocity and vehicle-pedestrian distance features

3.4.3 Feed-forward multilayer perceptron

The proposed framework uses a FF-MLP to process the velocity and distance features extracted from the video frames. The MLP applies non-linear transformations to map these features into a more meaningful and compact representation. As shown in Figure 3, the velocity-distance feature vector is fed into a multi-layer perceptron (MLP) consisting of two

hidden layers with 64 and 32 units. The first hidden layer with 64 units captures high-level motion representations, while the second layer with 32 units refines the learned features, ensuring robust encoding of pedestrian movement patterns. The MLP helps reduce the dimensionality of the extracted features to keep only crucial information related to pedestrian behavior. The system leverages this MLP structure for filtering, refining, and transforming velocity-distance features into the structured format required to complement the spatial and temporal representations learned by ST-GCN.

Algorithm 2. Velocity and distance feature extraction through multi-layer perceptron (MLP)

Input:

Bounding Box Data:

$$B = \{b1, b2, \dots, bt\}$$

Vehicle Data:

$$V = \{v1, v2, \dots, vt\}$$

Distance Data:

$$D = \{d1, d2, \dots, dt\}$$

Weight Matrices:

$$W = \{W1, W2, W3\}$$

Output:

$fvd \leftarrow$ Vehicle–Distance feature representation

Procedure:

- 1: for each frame ft do
 - 2: Extract bounding-box features
 $H1 \leftarrow \text{MLP}(B, W1)$
 - 3: Extract vehicle motion features
 $H2 \leftarrow \text{MLP}(V, W2)$
 - 4: Extract distance features
 $H3 \leftarrow \text{MLP}(D, W3)$
 - 5: Fuse extracted features
 $fvd \leftarrow H1 \oplus H2 \oplus H3$
 - 6: Store fvd
 - 7: end for
 - 8: Return fvd
-

3.5 Cross-modal transformer network

In this phase, we adapt the transformer model as employed in Chen et al. [5], which is used to combine environmental features (LS) and motion features (PB, D, V) for better prediction of pedestrian crossing intention. This transformer model employs self-attention and cross-attention mechanisms, which embed the relevant information into the transformer attention layers and integrate that information to improve the overall ability of the system to understand pedestrian behavior.

3.5.1 Self-attention layer

The self-attention layer captures long-range dependencies associated with each feature modality (i.e., environmental and velocity/distance features). With reference to the environmental feature stream, self-attention helps relate different objects, such as pedestrians, vehicles, and traffic elements, within a given frame to refine spatial characteristics of the local scene (LS). In velocity-distance feature processing, self-attention tracks trends of movement over time, ensuring that pedestrian dynamics are accurately represented. By dynamically weighting the importance of various features using self-attention, only the most relevant spatial and motion features are weighted towards the final prediction.

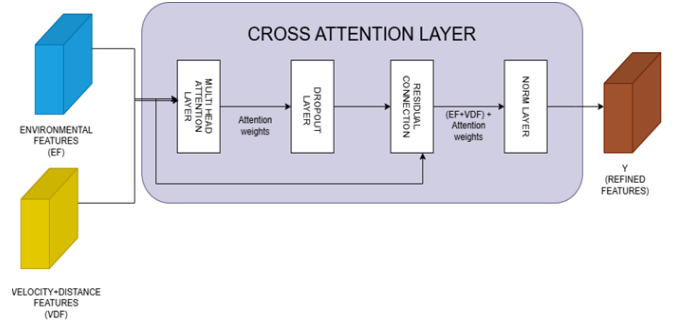


Figure 4. Architecture of the cross-attention layer used in the cross-modal transformer network

3.5.2 Cross attention layer

The Cross Attention Layer helps the transformer network correlate spatial information with pedestrian velocity and distance features, ensuring appropriate mapping from the environmental context to pedestrian motion. This mechanism increases the understanding of complicated pedestrian-vehicle interactions, dynamically weighing the contribution of each modality according to the context. Cross-attention ensures spatial and motion cues are preserved in the fused representation, feeding forward toward better PCIPs.

As illustrated in Figure 4, the cross-attention layer takes the Environmental Features as Query(Q) and Velocity-Distance Features as Key(K) and Value(V). The attention mechanism measures how much a particular environmental feature needs to focus on speed and distance. Then the computed attention scores are weighted and integrated into the environmental context by weighting speed and distance information.

3.6 Bottleneck Feature Fusion

BFF module is introduced to merge two potentially redundant feature representations, bounding box $bbox$ and $Pred_m$, into a more concise and informative representation. This combination is essential to enhance PCIP by reducing the dimensionality of the feature space while preserving important information. The process starts by expanding the feature dimensions of both using a MLP. This conversion yields two expanded features, $bbox$ and $Pred_m$, which combine the original features into a larger feature space of dimension f_{id} . The expanded features are then divided into two parts: one part, $bbox$, maintains the original dimension f_{id} , while the other part represents a compressed version of the original features with a smaller dimension f_{dc} . These compressed parts serve as the bottleneck of the original features. Their combination is done by simple addition, yielding the condensed representation. The resulting feature f_i is then combined with the original parts to yield improved features, which are then utilized for the subsequent stages of the model.

In order to address the temporal dynamics, an attention mechanism is added. This enables the model to assign varying weights to varying time steps in the sequence to provide more emphasis on the most critical temporal features in crossing intention prediction. Therefore, the BFF module realizes a significant reduction in computational cost for feature processing without sacrificing critical information for effective prediction. In order to improve the PCIP method, the environmental and velocity-distance characteristics retrieved are combined using the BFF technique. The BFF module is used to fuse redundant temporal-spatial feature representations into a compact vector without losing discriminative

information for crossing intention recognition. The fusion is subsequently followed by a temporal attention mechanism that provides importance weights to different frames. For temporal dynamics extraction, an attention mechanism is introduced. This allows the model to provide different weights for different time steps in the sequence, thus providing importance to the most important temporal features for crossing intention prediction. In general, the BFF module reduces the computational expense of feature processing considerably but maintains critical information for effective prediction.

3.7 Multi-task learning

The PCIP framework concurrently refines the bounding box localization and predicts pedestrian crossing intention using MTL. This enhances robustness and accuracy. By distributing the model's representation-sharing capabilities among related tasks, MTL improves generalization and accelerates the learning process. The model is designed to provide two significant predictions: the task of anticipating a pedestrian's intention to cross or not is called pedestrian intention prediction, and better bounding box regression enhances spatial localization of pedestrians with the help of motion and context cues. The network enhances classification and localization performance by leveraging common features from training these tasks together.

3.8 Implementation details

The proposed PCIP framework was implemented using PyTorch and consists of an ST-GCN-based environmental feature extractor, an FF-MLP-based velocity and distance feature extractor, a Cross-Modal Transformer, and a BFF module. YOLOv8 pretrained on the MS COCO dataset was used for pedestrian, vehicle, and traffic-element detection. VGG16 was selected as the visual feature extractor based on the comparative evaluation with ResNet50. The FF-MLP consists of two hidden layers with 64 and 32 units for encoding the velocity and distance features. The environmental and velocity-distance representations are subsequently processed through the self-attention and cross-attention mechanisms of the Cross-Modal Transformer. The fused representation is passed to the BFF module and subsequently to the multi-task prediction layer for pedestrian crossing-intention classification and bounding-box regression.

The model was trained using the PIE dataset with a video-level split of 70% for training, 10% for validation, and 20% for testing. The Adam optimizer was used with an initial learning rate of 10^{-4} and a batch size of 32. The multi-task objective combines binary cross-entropy loss for crossing-intention classification with Smooth L1 loss for bounding-box regression. All experiments were conducted using PyTorch on an NVIDIA GeForce RTX 3090 GPU with an Intel Core i5 processor. The model was trained for 50 epochs, and the validation data were used to monitor model convergence and generalization.

Algorithm 3. Cross-modal transformer network

Input:

Scene context feature matrix	fenv
Vehicle-distance feature representation	fvd
Positional encoding	Epos
Intra-modal tokens	Uintra(fenv)
Inter-modal tokens	Uinter(fvd)

Output:

Pedestrian intention label	y
Predicted bounding box	bbox = (x, y, w, h)

Procedure:

```

1: Generate intra-modal tokens
Uintra ← Tokenize(fenv)
2: Generate inter-modal tokens
Uinter ← Tokenize(fvd)
3: // Intra-modal encoding
4: for each token m in Uintra do
5: Hm1 ← Uintra(m) + Epos
6: αm ← Softmax(QmKTm)
7: Gm ← αmVm
8: R̄m1 ← MLP(Concat(G1, G2, ..., Gm))
9: Rm1 ← MHSA(R̄m1 + Hm1)
10: end for
11: fbbox ← IMSA(Rm1)
12: // Inter-modal encoding
13: for each token m in Uinter do
14: Hm2 ← Uinter(m) + Epos
15: αm ← Softmax(QmKTm)
16: Gm ← αmVm
17: R̄m2 ← MLP(Concat(G1, G2, ..., Gm))
18: Rm2 ← MHSA(R̄m2 + Hm2)
19: end for
20: fpred ← IMSA(Rm2)
21: Predm ← CMSA(
  LN(
    MLP((fbbox, r0),
          (fpred, r1))
  )
)
22: Bf ← BFF(Predm)
23: [bbox, y] ← MultiTaskLearning(Bf)
24: Return [bbox, y]

```

4. EXPERIMENTAL RESULTS

4.1 Dataset settings

To evaluate the performance of the proposed PCIP framework, experiments were conducted using the PIE dataset, while the SEU_PML dataset was used for pre-training the spatial graph representation. The PIE dataset provides urban driving video sequences captured from an onboard vehicle camera, together with pedestrian annotations and ego-vehicle motion information, including vehicle speed and orientation-related measurements. The dataset therefore provides complementary visual and vehicle-motion information required for pedestrian crossing-intention prediction. The PIE dataset was divided at the video level into 70% training, 10% validation, and 20% testing subsets. A video-level split was adopted to ensure that frames from the same video sequence did not occur in both training and testing subsets, thereby reducing the possibility of information leakage and providing a more reliable evaluation of generalization to unseen video sequences. The SEU_PML dataset was used specifically to pre-train the spatial graph representation used in the environmental feature extraction stage. The pre-trained spatial representation was subsequently integrated into the proposed PCIP framework, which was

trained and evaluated using the PIE dataset. The proposed framework was implemented using PyTorch. The model was trained using the Adam optimizer with an initial learning rate of 10^{-4} and a batch size of 32. The multi-task training objective combines binary cross-entropy loss for pedestrian crossing-intention classification and Smooth L1 loss for bounding-box regression. All experiments were conducted using an NVIDIA GeForce RTX 3090 GPU and an Intel Core i5 processor.

4.1.1 SEU_PML dataset

The SEU_PML dataset [43] is an open-source traffic-participant dataset containing surveillance images collected from road environments. It contains 6,588 images organized into four supercategories and thirteen subcategories, providing visual information about different traffic participants and road elements. In this work, the SEU_PML dataset was used for pre-training the spatial graph representation of the environmental feature extractor. The dataset provides diverse traffic participants and environmental elements that can be represented as graph nodes, enabling the spatial graph module to learn relationships among objects before being integrated into the pedestrian crossing-intention prediction framework. Figure 5 shows sample surveillance-camera images from the SEU_PML dataset showing traffic participants and road-scene elements.



Figure 5. Sample surveillance camera images captured at a road junction from the SEU_PML dataset

4.1.2 Pedestrian Intention Estimation dataset

The PIE dataset [3] contains more than six hours of urban driving video recorded using an onboard vehicle camera. The dataset provides pedestrian annotations together with vehicle-related information, making it suitable for studying pedestrian behavior and crossing intention from the vehicle's perspective. The video sequences were converted into individual frames using FFmpeg at a resolution of 1920×1080 pixels. To reduce unnecessary processing, frames were retained primarily from temporal segments containing detected pedestrians. Consequently, portions of the videos without relevant pedestrian activity were excluded from the processing pipeline. The PIE dataset was subsequently divided into training, validation, and testing subsets using a video-level split of 70%, 10%, and 20%, respectively. This procedure ensures that the evaluation sequences are independent of the sequences used for model training. The two datasets therefore serve different purposes in the proposed framework:

SEU_PML is used for pre-training the spatial graph representation, whereas PIE is used for training and evaluating the complete PCIP framework.

The SEU_PML dataset contains 6588 images as 4 supercategories and 13 subcategories, with high-resolution photos as shown in Figure 5. It supports all modes of traffic and is best suited for surveillance tasks such as vehicle and pedestrian detection through graph models.

4.2 Evaluation metrics

The performance of the proposed framework was evaluated from two perspectives. First, the reconstruction capability of the spatial graph pre-training module was evaluated using reconstruction loss. Second, the final pedestrian crossing-intention prediction performance was evaluated using accuracy, precision, and F1-score.

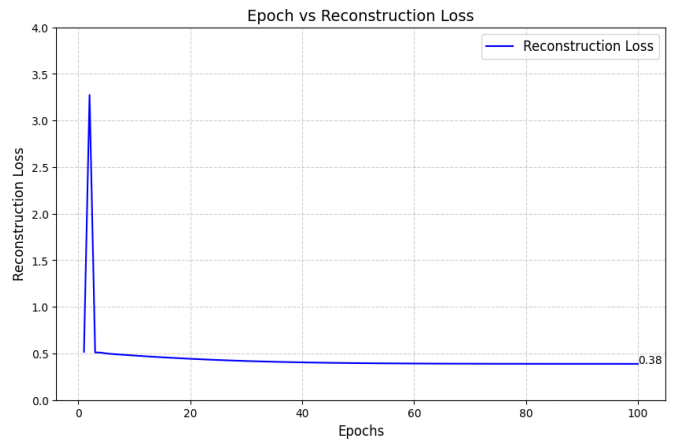


Figure 6. The plot illustrates the change in reconstruction loss over training epochs for the Graph Convolutional Network (GCN) Autoencoder

Reconstruction Loss: Measures the ability of GCN layer to accurately reconstruct the adjacency matrix and node features, ensuring spatial and feature-based dependencies are preserved. The reconstruction loss is calculated using the equation:

$$L_{\{recons\}} = ||\bar{A} - A|| \quad (4)$$

where, $\bar{A}$ represents the reconstructed adjacency matrix and A represents the original adjacency matrix.

The proposed graph autoencoder achieves a reconstruction loss of 0.38, as shown in Figure 6. The decreasing trend in reconstruction loss indicates that the graph representation gradually converges during training and is able to preserve the spatial relationships represented by the input graph. Accuracy, precision, and F1-score are used to evaluate the proposed model as a classification problem.

4.2.1 Accuracy

Accuracy provides a general measure of the overall correctness of the model. It is the ratio of the total number of correct classifications to the total number of classifications. The formula for accuracy is given by,

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

4.2.2 Precision

Precision, also known as positive-predictive value (PPV), is the ratio of correctly predicted positive observations to the total predicted positives. It focuses on the accuracy of positive predictions. The formula for precision is given by,

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

4.2.3 F1-score

The F1-score is the harmonic mean of precision and recall. It provides a balance between the two metrics and is especially useful when there is an uneven class distribution. The formula for F1-score is given by,

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

As denoted in the Eqs. (5)-(7), TP denotes True Positives, TN denotes True Negatives, FP denotes False Positives, and FN denotes False Negatives.

4.3 Comparison of visual feature extractors

Visual feature extractors constitute the basis for comprehending scene context, as they convert raw image inputs into meaningful representations, which contain spatial and semantic information required for downstream processing tasks. We use VGG16 and ResNet50 to extract the visual features in this work. For best performance, these models are also examined in terms of how they affect MTL. The impact of hyperparameters such as learning rate, batch size, and weight decay on convergence and generalization is also examined in the work.

Table 1 shows the network parameters used to train the ResNet50 and VGG16 models for the proposed model to achieve training accuracy efficiently. As shown in Figure 7, ResNet50 achieves a training accuracy of 73% and ultimately saturates at 50 steps, which indicates poor learning capability in this work. On the other hand, VGG16 has consistently shown an upward trend and achieves more than 90% training accuracy after 100 steps, thereby making it a better fit for the task of visual feature extraction.

Table 1. Network parameters of the feature extractor for the trained model

Models	Learning Rate	Optimizer	Batch Size	Epochs
ResNet50	0.01	Adam	16	100
VGG16	0.001	Adam	16	100

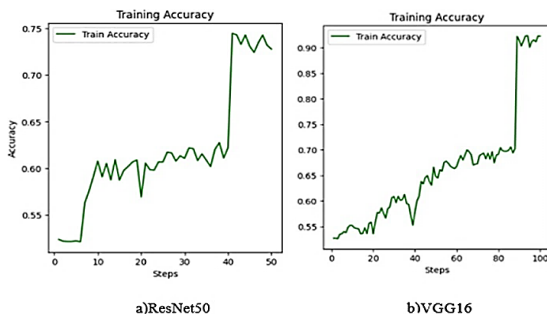


Figure 7. Accuracy across steps for trained visual feature extractor models

4.4 Performance comparison

The results in Table 2 indicate that the performance of pedestrian crossing-intention prediction depends strongly on the type and combination of input features. The trajectory-based model using visual-flow information provides limited performance because it primarily captures pedestrian motion without explicitly modeling the surrounding traffic context. In contrast, graph-based approaches achieve improved performance by representing spatial relationships among pedestrians and environmental elements.

Table 2. Comparative evaluation of the proposed PCIP model and existing models using different input modalities

Model	Input Features	Accuracy	F1 Score	Precision
LSTM	Vis	0.76	0.57	0.45
WaveNet	Vis, BB	0.85	0.42	0.63
ST-GCN	LS, BB	0.90	0.82	0.87
PedGraph+ [24]	Vis, SK, I	0.89	0.81	0.83
PedAST-GCN [22]	SK, V, BB	0.90	0.80	0.826
IntFormer [28]	BB, V, SK, I	0.89	0.81	–
PedCMT [5]	BB, V, I	0.93	0.87	–
PCIP	BB, LS, V, D	0.93	0.88	0.87

Note: PCIP = pedestrian crossing intention prediction; *V* = Visual Features; *BB* = Bounding Box; *SK* = Skeletal Keypoints; *LS* = Local Scene; *I* = Image; *V* = Velocity; *D* = Distance.

The performance of ST-GCN and PedAST-GCN demonstrates the benefit of incorporating spatial and structural relationships into pedestrian behavior modeling. Transformer-based approaches such as PedCMT further improve prediction by learning dependencies among heterogeneous input features. However, these approaches do not explicitly combine pedestrian–vehicle proximity with environmental graph representations in the same manner as the proposed framework. The proposed PCIP framework achieves an accuracy of 0.93, an F1-score of 0.88, and a precision of 0.87. The improvement can be attributed to the complementary information provided by the different feature branches. The ST-GCN captures spatial and temporal relationships among pedestrians, vehicles, and environmental elements, while the distance and velocity branch provides explicit information about the dynamic relationship between the pedestrian and the approaching vehicle. The Cross-Modal Transformer further integrates these heterogeneous representations through self-attention and cross-attention, allowing the model to learn relationships between environmental context and vehicle–pedestrian motion. The BFF module provides a compact representation by reducing redundant information before the final prediction.

Therefore, the performance improvement is not attributed to a single component but to the complementary interaction among environmental context, pedestrian information, vehicle velocity, pedestrian–vehicle distance, and multimodal feature fusion. These results demonstrate the advantage of combining graph-based spatial-temporal modeling with cross-modal attention for pedestrian crossing-intention prediction.

The baseline models compared in Table 2 utilize varying input feature combinations (pure trajectory coordinates, skeletal keypoints, or visual flow). The benchmark

comparison in Table 2 evaluates the overall predictive capability of the system architectures rather than assuming identical input modalities. The superior performance of our hybrid model (Accuracy: 0.93, F1: 0.88) highlights the predictive advantage of integrating explicit metric Haversine distance (D) and vehicle velocity (V) alongside ST-GCN spatial graph topologies and Cross-Modal Transformer fusion.

4.5 Behavioral analysis of pedestrian crossing intention prediction

To further examine the behavioral prediction capability of the proposed PCIP framework, qualitative analysis was performed under single-pedestrian and multi-pedestrian scenarios. These scenarios were selected to examine how the proposed model responds to different pedestrian behaviors and surrounding traffic conditions.

Single Pedestrian Environment: A scene where only one pedestrian is present and detected in the frame. As shown in Figure 8, three scenarios demonstrate 10 seconds of predicted pedestrian behavior, generated by the proposed model, in a single- pedestrian environment. In Scenario 1, the pedestrian shows an intention to cross the road after 5 seconds. At time $t = 0$, the pedestrian arrives at the crossing point, and by $t = 3$, there is a noticeable change in posture indicating the intention to initiate crossing. In Scenario 2, the pedestrian begins crossing at the zebra crossing and continues to do so throughout the 10-second interval. In Scenario 3, although the pedestrian is detected, no intention to cross the road is observed during the entire 10-second duration.

Multiple Pedestrian Environment: A scene in which two or more pedestrians interact or appear simultaneously within the same frame.

As shown in Figure 9, two scenarios from a multi-pedestrian environment are illustrated. In Scenario 1, two pedestrians are detected in the frame at time $t = 0$. By $t = 3$, one pedestrian changes posture and begins crossing, while the second pedestrian starts to exhibit signs of intention. By $t = 5$, the proposed model successfully recognizes both pedestrians as having the intention to cross the road. In Scenario 2, multiple pedestrians are detected in the frame, and among them, those showing an intention to cross are correctly identified by the proposed model. Based on these behavioral analyses, the proposed model accurately predicts pedestrian intentions in both isolated and crowded areas, across signalized and unsignalized environments.

Figure 10 illustrates the temporal correspondence between the predicted and ground-truth crossing intentions. The predicted sequence follows the major transitions in pedestrian behavior, including the transition from non-crossing to crossing. This qualitative comparison provides additional evidence that the proposed framework can capture temporal changes in pedestrian crossing intention. The qualitative results demonstrate that the proposed framework can handle both single- and multi-pedestrian scenes by jointly considering pedestrian behavior, environmental context, and vehicle-related information. In multi-pedestrian scenarios, the graph representation provides an explicit mechanism for modeling relationships between the approaching vehicle and individual scene elements. This enables the model to distinguish pedestrian-specific interactions within the same traffic scene.

However, challenging conditions such as severe occlusion, high pedestrian density, ambiguous behavioral cues, abrupt vehicle movements, and variations in illumination may affect

prediction performance. Since the proposed framework relies on object detection and subsequent graph construction, errors in pedestrian or vehicle localization can propagate to the ST-GCN and subsequent multimodal feature-extraction stages. These observations highlight the need for further evaluation under more diverse and challenging traffic conditions.



Figure 8. Single pedestrian crossing scenes showing frame-by-frame temporal predictions generated by the proposed PCIP model on the PIE dataset over a 10-second activity duration

Note: PCIP = pedestrian crossing intention prediction; PIE = Pedestrian Intention Estimation.

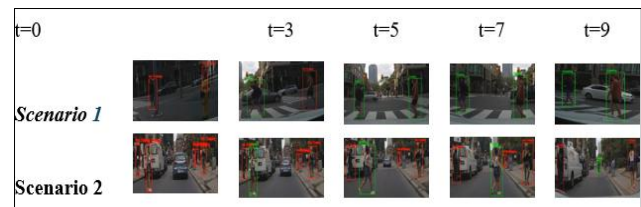


Figure 9. Multi pedestrian crossing scenes showing frame-by-frame temporal predictions generated by the proposed PCIP model on the PIE dataset over a 10-second activity duration

Note: PCIP = pedestrian crossing intention prediction; PIE = Pedestrian Intention Estimation.

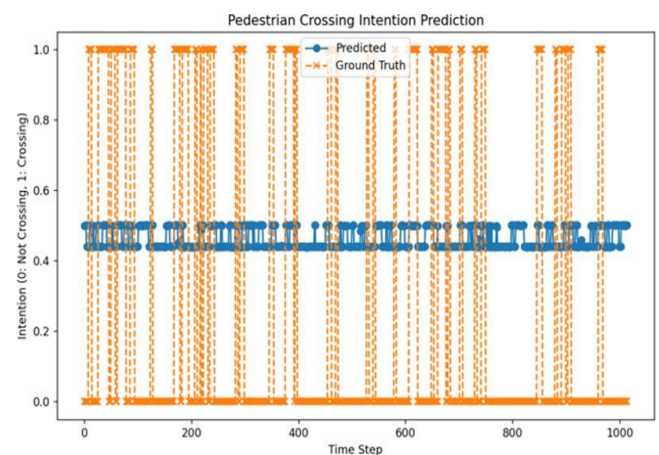


Figure 10. Comparison of predicted and ground truth pedestrian crossing intentions across temporal frames

4.6 Convergence analysis of pedestrian crossing intention prediction

Convergence analysis is essential to understand the training stability and efficiency of the proposed model. This indicates how well the proposed model minimizes the loss function and how it adapts to the training data over epochs. Figure 10 shows

that the frame-wise evaluation confirms that predictions align closely with actual pedestrian movements. This step ensures that the model is reliable in detecting pedestrian intent frame by frame. The training dynamics of the proposed model over a period of 50 epochs are shown in Figure 11. In the Full Loss plot, training and validation losses move quite smoothly and consistently downwards along epochs, indicating good optimization and a generalized model without overfitting. Conversely, the Classification Loss plot shows variation from training classification loss to validation classification loss. While training classification loss declines theoretically with minor oscillations, validation loss is more varied, showing occasional spikes. Despite this, validation loss trends downward to the training loss, indicating good consistency by the proposed model in terms of classification performance on unseen data. These trends confirm that the model converges effectively and maintains a balance between fitting the training data and generalizing to the validation set.

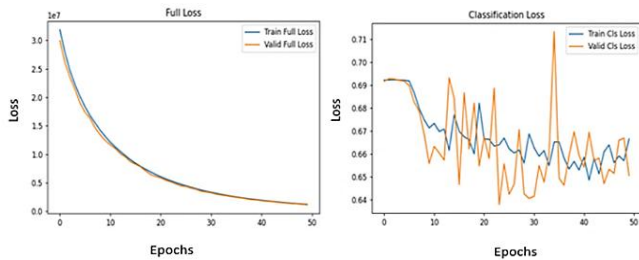


Figure 11. Comparison of classification loss and overall loss for evaluating model generalization capability

4.7 Evaluation on unconstrained video

To examine the practical applicability of the proposed PCIP framework outside the benchmark dataset, a qualitative evaluation was conducted using an unconstrained road-scene video recorded at Elgin Road, Kolkata. The video contains pedestrian movements and road-crossing activities under natural traffic conditions. Since the video does not provide ground-truth pedestrian intention annotations, this evaluation is presented as a qualitative assessment rather than a quantitative performance evaluation.

Figure 12 presents two representative scenarios from the unconstrained video. In Scenario 1, multiple pedestrians are detected simultaneously, with three pedestrians predicted as crossing and one pedestrian predicted as non-crossing. In Scenario 2, the model predicts a crossing intention for one pedestrian, while another pedestrian remains relatively stationary near a vehicle and is classified as non-crossing or uncertain. These demonstrate the ability of the proposed framework to generate frame-level pedestrian crossing-intention predictions under unconstrained traffic conditions.



(a) Scenario 1



(b) Scenario 2

Figure 12. Multi-pedestrian crossing scenarios predicted by the proposed model

Note: The top row shows (a) Scenario 1 with multiple pedestrians exhibiting both crossing and non-crossing intentions, and (b) Scenario 2 with one pedestrian crossing and another with uncertain behavior.

However, because the external video does not contain ground-truth intention annotations, the results cannot be used to calculate quantitative accuracy, precision, or F1-score. Therefore, this experiment is intended only to demonstrate the practical applicability and qualitative behavior of the proposed framework in an unseen traffic environment.

5. CONCLUSION

In this work, we proposed a hybrid deep learning framework for PCIP that integrates spatial, temporal, and vehicle-related information for pedestrian behavior prediction in dynamic traffic environments. The proposed framework combines local environmental representations and pedestrian bounding-box information extracted using ST-GCN with vehicle velocity and pedestrian-vehicle distance features processed through an FF-MLP. These heterogeneous representations are subsequently integrated using a Cross-Modal Transformer and BFF module to capture spatial-temporal dependencies and vehicle-pedestrian interactions. Experimental evaluation on the PIE dataset demonstrated that the proposed framework achieves an accuracy of 0.93, an F1-score of 0.88, and a precision of 0.87. The behavioral analysis further demonstrated the capability of the framework to generate crossing-intention predictions in both single- and multi-pedestrian scenarios. In addition, qualitative evaluation using an unconstrained road-scene video indicated that the framework can generate pedestrian crossing-intention predictions outside the benchmark dataset. However, since the external video does not contain ground-truth intention annotations, these results are considered qualitative and are not used for quantitative performance evaluation. The results indicate that combining environmental graph representations with pedestrian-vehicle distance and vehicle velocity provides complementary information for pedestrian crossing-intention prediction. The explicit representation of pedestrian-vehicle proximity, together with spatial-temporal environmental features and cross-modal feature fusion, contributes to the overall predictive performance of the proposed framework.

Future work will focus on (i) evaluating the generalization of the framework across diverse environments, adverse weather conditions such as rain and fog, and nighttime traffic scenarios; (ii) improving computational efficiency through knowledge distillation, model compression, and quantization for deployment on resource-constrained automotive platforms; and (iii) incorporating Vehicle-to-Everything

(V2X) communication to provide additional information from connected vehicles and infrastructure, thereby enabling more comprehensive multi-vehicle and environmental awareness.

ACKNOWLEDGEMENTS

The authors would like to extend their appreciation to Mepco Schlenk Engineering College for offering the necessary infrastructure and support for this research. The authors also recognize the utilization of publicly available datasets, including the annotation files.

REFERENCES

- [1] Xu, F.Y., Xu, F., Xie, J.C., Pun, C.M., Lu, H.M., Gao, H. (2022). Action recognition framework in traffic scene for autonomous driving system. *IEEE Transactions on Intelligent Transportation Systems*, 23(11): 22301-22311. <https://doi.org/10.1109/tits.2021.3135251>
- [2] Li, J., Shi, X., Chen, F., et al. (2023). Pedestrian crossing action recognition and trajectory prediction with 3D human keypoints. In 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, pp. 1463-1470. <https://doi.org/10.1109/icra48891.2023.10160273>
- [3] 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, pp. 6261-6270. <https://doi.org/10.1109/iccv.2019.00636>
- [4] Naik, A.Y., Bighashdel, A., Jancura, P., Dubbelman, G. (2022). Scene spatio-temporal graph convolutional network for pedestrian intention estimation. In 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, pp. 874-881. <https://doi.org/10.1109/iv51971.2022.9827231>
- [5] Chen, X., Zhang, S., Li, J., Yang, J. (2024). Pedestrian crossing intention prediction based on cross-modal transformer and uncertainty-aware multi-task learning for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 25(9): 12538-12549. <https://doi.org/10.1109/tits.2024.3386689>
- [6] Yang, J., Chen, Y., Du, S., Chen, B., Principe, J.C. (2024). IA-LSTM: Interaction-aware LSTM for pedestrian trajectory prediction. *IEEE Transactions on Cybernetics*, 54(7): 3904-3917. <https://doi.org/10.1109/tcyb.2024.3359237>
- [7] Rasouli, A. (2024). A novel benchmarking paradigm and a scale- and motion-aware model for egocentric pedestrian trajectory prediction. In 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, pp. 5630-5636. <https://doi.org/10.1109/icra57147.2024.10610614>
- [8] Quan, R., Zhu, L., Wu, Y., Yang, Y. (2021). Holistic LSTM for pedestrian trajectory prediction. *IEEE Transactions on Image Processing*, 30: 3229-3239. <https://doi.org/10.1109/tip.2021.3058599>
- [9] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 961-971. <https://doi.org/10.1109/cvpr.2016.110>
- [10] Kitchat, K., Chiu, Y.L., Lin, Y.C., et al. (2024). PedCross: Pedestrian crossing prediction for auto-driving bus. *IEEE Transactions on Intelligent Transportation Systems*, 25(8): 8730-8740. <https://doi.org/10.1109/TITS.2024.3379421>
- [11] Yang, B., Wei, Z., Hu, H., Wang, R., Yang, C., Ni, R. (2023). DPCIAN: A novel dual-channel pedestrian crossing intention anticipation network. *IEEE Transactions on Intelligent Transportation Systems*, 25(6): 6023-6034. <https://doi.org/10.1109/TITS.2023.3333328>
- [12] Gesnouin, J., Pechberti, S., Stanculescu, B., Moutarde, F. (2021). TrouSPI-Net: Spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, pp. 01-07. <https://doi.org/10.1109/FG52635.2021.9666989>
- [13] Ahmed, S., Al Bazi, A., Saha, C., Rajbhandari, S., Huda, M.N. (2023). Multi-scale pedestrian intent prediction using 3D joint information as spatio-temporal representation. *Expert Systems with Applications*, 225: 120077. <https://doi.org/10.1016/j.eswa.2023.120077>
- [14] Ling, Y., Zhang, Q., Weng, X., Ma, Z. (2023). STMA-GCN_PedCross: Skeleton based spatial-temporal graph convolution networks with multiple attentions for fast pedestrian crossing intention prediction. In 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, pp. 500-506. <https://doi.org/10.1109/ITSC57777.2023.10421893>
- [15] Razali, H., Mordan, T., Alahi, A. (2021). Pedestrian intention prediction: A convolutional bottom-up multi-task approach. *Transportation Research Part C: Emerging Technologies*, 130: 103259. <https://doi.org/10.1016/j.trc.2021.103259>
- [16] Zhang, S., Abdel-Aty, M., Wu, Y., Zheng, O. (2021). Pedestrian crossing intention prediction at red-light using pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 23(3): 2331-2339. <https://doi.org/10.1109/TITS.2021.3074829>
- [17] Hashimoto, Y., Gu, Y., Hsu, L.T., Iryo-Asano, M., Kamijo, S. (2016). A probabilistic model of pedestrian crossing behavior at signalized intersections for connected vehicles. *Transportation Research part C: Emerging technologies*, 71: 164-181. <https://doi.org/10.1016/j.trc.2016.07.011>
- [18] Gesnouin, J., Pechberti, S., Bresson, G., Stanculescu, B., Moutarde, F. (2020). Predicting intentions of pedestrians from 2D skeletal pose sequences with a representation-focused multi-branch deep learning network. *Algorithms*, 13(12): 331. <https://doi.org/10.3390/a13120331>
- [19] Ma, J., Rong, W. (2022). Pedestrian crossing intention prediction method based on multi-feature fusion. *World Electric Vehicle Journal*, 13(8): 158. <https://doi.org/10.3390/wevj13080158>
- [20] Fang, Z., López, A.M. (2019). Intention recognition of pedestrians and cyclists by 2D pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 21(11): 4773-4783. <https://doi.org/10.1109/TITS.2019.2946642>
- [21] Zhou, W., Liu, Y., Zhao, L., Xu, S., Wang, C. (2023).

- Pedestrian crossing intention prediction from surveillance videos for over-the-horizon safety warning. *IEEE Transactions on Intelligent Transportation Systems*, 25(2): 1394-1407. <https://doi.org/10.1109/TITS.2023.3314051>
- [22] Ling, Y., Ma, Z., Zhang, Q., Xie, B., Weng, X. (2024). PedAST-GCN: Fast pedestrian crossing intention prediction using spatial-temporal attention graph convolution networks. *IEEE Transactions on Intelligent Transportation Systems*, 25(10): 13277-13290. <https://doi.org/10.1109/TITS.2024.3398252>
- [23] Zhang, X., Angeloudis, P., Demiris, Y. (2022). St crossingpose: A spatial-temporal graph convolutional network for skeleton-based pedestrian crossing intention prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(11): 20773-20782. <https://doi.org/10.1109/TITS.2022.3177367>
- [24] Cadena, P.R.G., Qian, Y., Wang, C., Yang, M. (2022). Pedestrian graph+: A fast pedestrian crossing prediction model based on graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(11): 21050-21061. <https://doi.org/10.1109/TITS.2022.3173537>
- [25] Sofianos, T., Sampieri, A., Franco, L., Galasso, F. (2021). Space-time-separable graph convolutional network for pose forecasting. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 11189-11198. <https://doi.org/10.1109/iccv48922.2021.01102>
- [26] Liu, B., Adeli, E., Cao, Z., et al. (2020). Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2): 3485-3492. <https://doi.org/10.1109/lra.2020.2976305>
- [27] Lorenzo, J., Alonso, I.P., Izquierdo, R., et al. (2021). CAPformer: Pedestrian crossing action prediction using transformer. *Sensors*, 21(17): 5694. <https://doi.org/10.3390/s21175694>
- [28] Zhou, H., Zhang, S., Peng, J., et al. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *AAAI*, 35(12): 11106-11115. <https://doi.org/10.1609/aaai.v35i12.17325>
- [29] 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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, pp. 3666-3675. <https://doi.org/10.1109/CVPRW59228.2023.00374>
- [30] Yao, Y., Atkins, E., Johnson-Roberson, M., Vasudevan, R., Du, X. (2021). Coupling intent and action for pedestrian crossing behavior prediction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1238-1244. <https://doi.org/10.24963/ijcai.2021/171>
- [31] Ham, J.S., Bae, K., Moon, J. (2022). MCIP: Multi-stream network for pedestrian crossing intention prediction. In *European Conference on Computer Vision*, Tel Aviv, Israel, pp. 663-679. https://doi.org/10.1007/978-3-031-25056-9_42
- [32] Chen, T., Tian, R., Ding, Z. (2021). Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, pp. 3096-3102. <https://doi.org/10.1109/ICCVW54120.2021.00345>
- [33] Singh, A., Suddamalla, U. (2021). Multi-input fusion for practical pedestrian intention prediction. In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, pp. 2304-2311. <https://doi.org/10.1109/ICCVW54120.2021.00260>
- [34] Kotseruba, I., Rasouli, A., Tsotsos, J.K. (2021). Benchmark for evaluating pedestrian action prediction. In 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, pp. 1257-1267. <https://doi.org/10.1109/wacv48630.2021.00130>
- [35] Zhang, X., Wang, X., Zhang, W., Wang, Y., Liu, X., Wei, D. (2024). Multi-attention network for pedestrian intention prediction based on spatio-temporal feature fusion. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 238(13): 4202-4215. <https://doi.org/10.1177/09544070231190522>
- [36] Piccoli, F., Balakrishnan, R., Perez, M.J., et al. (2020). FuSSI-Net: Fusion of spatio-temporal skeletons for intention prediction network. In 2020 54th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, pp. 68-72. <https://doi.org/10.1109/IEEECONF51394.2020.9443552>
- [37] Saleh, K., Hossny, M., Nahavandi, S. (2019). Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal DenseNet. In 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, pp. 9704-9710. <https://doi.org/10.1109/icra.2019.8793991>
- [38] Gujjar, P., Vaughan, R. (2019). Classifying pedestrian actions in advance using predicted video of urban driving scenes. In 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, pp. 2097-2103. <https://doi.org/10.1109/ICRA.2019.8794278>
- [39] Kooij, J.F.P., Flohr, F., Pool, E.A.I., Gavrila, D.M. (2018). Context-based path prediction for targets with switching dynamics. *International Journal of Computer Vision*, 127(3): 239-262. <https://doi.org/10.1007/s11263-018-1104-4>
- [40] Schulz, A.T., Stiefelhagen, R. (2015). Pedestrian intention recognition using latent-dynamic conditional random fields. In 2015 IEEE Intelligent Vehicles Symposium (IV), pp. 622-627. <https://doi.org/10.1109/ivs.2015.7225754>
- [41] Zhou, W., Wang, C., Xia, J., Qian, Z., Wu, Y. (2023). Monitoring-based traffic participant detection in urban mixed traffic: A novel dataset and a tailored detector. *IEEE Transactions on Intelligent Transportation Systems*, 25(1): 189-202. <https://doi.org/10.1109/TITS.2023.3304288>
- [42] Lei, S., Yi, H., Sarmiento, J.S. (2024). Synchronous end-to-end vehicle pedestrian detection algorithm based on improved YOLOv8 in complex scenarios. *Sensors*, 24(18): 6116. <https://doi.org/10.3390/s24186116>