# Pedestrian Cross Forecasting with Hybrid Feature Fusion

**Meng Dong** MENGYINGYIDAI@GMAIL.COM
*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore*

**Editors:** Berrin Yanıkoğlu and Wray Buntine

## Abstract

Forecasting the crossing intention of pedestrians is an essential task for the safe driving of Autonomous Vehicles (AVs) in the real world. Pedestrians' behaviors are usually influenced by their surroundings in traffic scenes. Recent works based on vision-based neural networks extract key information from images to perform prediction. However, in the driving environment, there exists much critical information, such as the social and scene interaction in the driving area, the location and distance between the ego car and target pedestrian, and the motion of all targets. How properly exploring and utilizing the above implicit interactions will promote the development of Autonomous Vehicles. In this chapter, two novel attributes, the pedestrian's location on the road or sidewalk, and the relative distance from the target pedestrian to the ego-car, which are derived from the semantic map and depth map combined with bounding boxes, are introduced. A hybrid prediction network based on multi-modal is proposed to capture the interactions between all the features and predict pedestrian crossing intention. Evaluated by two public pedestrian crossing datasets, PIE and JAAD, the proposed hybrid framework outperforms the state-of-the-art by about an accuracy of 3%.

**Keywords:** Pedestrian Crossing, Feature Fusion

## 1. Introduction

Pedestrians, as the main participants in traffic roads, easily violate rules and are unpredictable due to the influence and restrictions of the surrounding environment Holländer et al. (2021). Their "stops" and "goes" behaviors are usually safety-critical, especially for road-crossing scenariosSun et al. (2021). Instead of human drivers, Autonomous Vehicles(AVs) could quickly detect and locate pedestrians based on current autonomous systems. Besides, they also could interpret and predict pedestrians' intentions based on the prediction module of Automated Driving Systems (ADS). Some works adopt individual features, such as observed trajectories, motion states, and pose, to forecast future locationsKothari et al. (2021a,b); Liu et al. (2021). These methods have high efficiency when pedestrians move smoothly in regular motion. However, past behaviors and trajectories may not indicate future movements in real traffic environments. Pedestrians may change their directions and velocities suddenly in dynamic surroundings. They may be the front cars, another pedestrian on the left, traffic lights, a repaired road, or sudden heavy rain Kothari et al. (2021a); Liu et al. (2021). Fig. 1 shows a sudden-change case due to the traffic rules and surroundings, the pedestrian does not follow her previous moving direction but changes to another road. Such "incidents" happen regularly as pedestrians keep their eyes and ears open when they are prepared to cross. So, predicting pedestrian crossing intention rather than certain attributes is a multi-modal problem. Recently, many public data sets related to pedestrians of automotive driving Rasouli et al. (2017c,a, 2019); Sun et al. (2020); Zhang et al. (2020) are created and released. These datasets provide rich spatial and behavioral annotations for road users, interaction simulation, and information from multi-sensors. A benchmark PCPA Kotseruba et al. (2021) for pedestrian crossing intention prediction, achieves outstanding accuracy on two public data sets: JAAD Rasouli et al. (2017c,a)and PIE Rasouli et al. (2019), based on multi-model framework incorporated visual features presented as local context and non-visual features including bounding boxes, poses, and ego-car speed. Each feature will be encoded individually. This method employs the multi-modal network to fuse multiple encoded features for final prediction results. However, such a fusion network will bear a heavy computing load when the two features are similar or correlated. To reduce potential redundancy and introduce the interaction between the target pedestrian and the scene, OSU Yang et al. (2022) proposes a spatial-temporal context feature with an attention mechanism based on PCPA. However, the implicit interactionsRasouli et al. (2021a) between labeled features, and other potential features such as distance and location, which will influence the intention, lack consideration in the above methods. Recently, Ham et al. (2022) provided novel fusion strategies by exploring global and local interactions in scenarios. In this paper, we study the problem of intention prediction from the ego-centric view of a moving vehicle by introducing two novel features: distance between the target pedestrian and ego-car, location of the target pedestrian (at road or sidewalk) accompany with present the existing features. Besides, novel fusion strategies are proposed to fully consider correlation or interaction between features. Based on the above analysis, the main contributions of this work are summarized as follows:

- We present a novel hybrid fusion method that utilizes interactions between features based on stacked GRU Rasouli et al. (2020) to predict pedestrian crossing intention.
- Two additional dynamic attributes, relative distances and location in the scene are introduced and evaluated by detected bounding boxes, monocular depth estimation map, and semantic segmentation map. These two additional attributes remove redundant interaction between pedestrians and other road users.
- We evaluate the performance of the proposed method using public datasets, and show that our method achieves stable and better performance over state-of-the-art algorithms.

## 2. Related Work

The problem of pedestrian intention forecasting from image sequences has attracted significant interest recently. As a sub-problem of action prediction, pedestrian crossing intention also raises huge interest in developing Autonomous Vehicles. The aim is to predict whether the target pedestrian crosses the road or not in the field of view of AVs in several future seconds.

There are two main approaches related to pedestrian crossing prediction. Traditional trajectory-based crossing intention prediction methods and hybrid feature fusion are extracted from input data. Based on Mordan et al. (2021), pedestrians have recognized

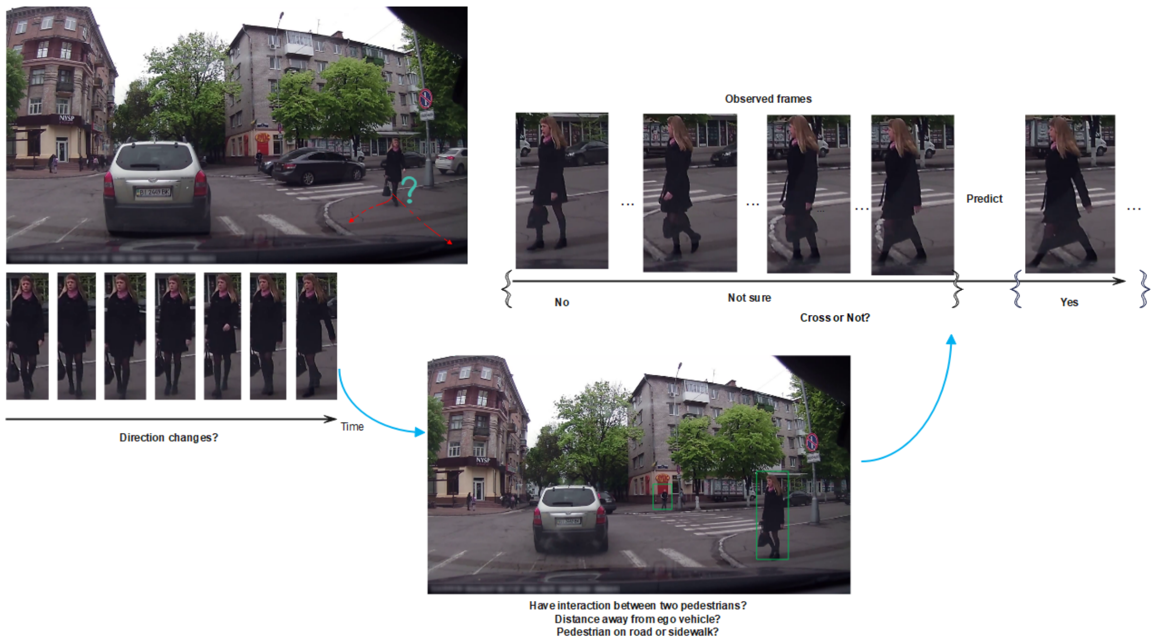


Figure 1: Pedestrian dynamic: change of moving direction; Interaction between pedestrian and other road users; Pedestrian on road or sidewalk?

32 related attributes in traffic scenarios. For the task of crossing intention prediction, researchers usually utilize different features and prediction networks to improve final accuracy. In an early study, JAAD Rasouli et al. (2017b) is created and labeled bounding boxes for all pedestrians, behavior, gender, and age, and contextual tags(weather, time, and street structure). Novel variations of previous individual modal-based methods are proposed to process the datasets. Piccoli et al. (2020) takes the observed motion from bounding boxes as input to a spatiotemporal Densenet to classify the future motion. Besides, pose features usually indicate the direction of future motion, they are extracted from OpenPose Cao et al. (2017a) and adopted in Fang and López (2019) to estimate the future pose of pedestrians. The distance and angle among the joint points are calculated to predict whether pedestrians cross. Recently, feature fusion methods have been explored for this problem. In Kotseruba et al. (2021); Yang et al. (2022); Osman et al. (2022), multiple features, including visual features extracted by CNN and non-visual features (i.e., ego-vehicle speed, pedestrians' pose, and detected bounding box), are fed into gated recurrent units (GRUs) and along with Fully Connected layer for final prediction. Ham et al. (2023) fuses eight input modalities with a systematic combination mechanism to fully explore the global and local features. Similarly, Ham et al. (2022) proposes a novel multi-stream network for pedestrian crossing intention prediction based on 5 inputs. High descriptive features and effective fusion will be critical in intention prediction. Besides the common framework of visual and non-visual modules, transformer-based methods have been widely introduced as the solution for prediction due to the outstanding performance of the transformer Lorenzo

et al. (2021b,a) with non-visual sequential features. Usually, in the driving environment, interactions among road agents have a significant impact on forecasting future behavior. It may exist between ego-vehicles and other road agentsBhattacharyya et al. (2021b) and scenes. So interaction modeling is widely equipped in trajectory prediction and intention estimationBhattacharyya et al. (2021a); Ettinger et al.. Interactions hidden in the traffic scene will vary with time. Semantic segmentation maps Yang et al. (2022) are commonly adopted to model such interactions. Some works introduce graph-based networks to explore the surrounding interactions of target pedestrians Cadena et al. (2022); Rasouli et al. (2021b); Song et al. (2022). However, there usually exist some redundant interactions in a real traffic case. For example, the interaction between a pedestrian and another pedestrian standing at a different crossroad is very limited, even though they are in the same camera view and segmentation from the semantic map. Besides, interactions also exist among features. Taking two features, the location pedestrian being standing and pose direction, for example, it may be probably crossing the road when a pedestrian standing at the road, and his/her head posed towards the road simultaneously. Combining these two features together would generate higher accurate crossing intention compared to individuals. In this paper, we consider such interactions into account for a robust fusion strategy.

## 3. Methods

### 3.1. Formulation

Generally, there are two possible results, crossing and not crossing, in the scenarios of prediction crossing, and it can be solved by classification techniques based on a sequence of observed video frames from a camera mounted in front of the moving ego vehicle. The features adopted in this paper are as follows: (1) Context features surrounding pedestrian $i$:

$$C_{li} = \left\{c_{li}^{t-m}, c_{li}^{t-m+1}, \ldots, c_{li}^{t}\right\} \tag{1}$$

(2) The context features from semantic segmentation mask in frame-level:

$$C_{g} = \left\{c_{g}^{t-m}, c_{g}^{t-m+1}, \ldots, c_{g}^{t}\right\} \tag{2}$$

(3) Ego car' real speed:

$$S_{obs} = \left\{s^{t-m}, s^{t-m+1} \ldots, s^{t}\right\} \tag{3}$$

(4) The location and velocity of target pedestrian $i$ calculated by coordinates of detected 2D bounding box (from top-left to bottom-right) and position changes from the previous frame $t-1$ to frame $t$ :

$$B_{obs} = \left\{b_{i}^{t-m}, b_{i}^{t-m+1}, \ldots, b_{i}^{t}\right\} \tag{4}$$

(5) Distance between ego car and pedestrian $i$, calculated by 2D bounding box and monocular depth estimation:

$$D_{obs} = \left\{d_{i}^{t-m}, d_{i}^{t-m+1}, \ldots, d_{i}^{t}\right\} \tag{5}$$

(6) Location in the scene, position attribute of target pedestrian $i$ where $l_i$ indicate whether the pedestrian is on the road or on the sidewalk:

$$L_{obs} = \left\{l_{i}^{t-m}, l_{i}^{t-m+1}, \ldots, l_{i}^{t}\right\} \tag{6}$$

(7) Pose key points of pedestrian $i$:

$$P_{obs} = \{p_i^{t-m}, p_i^{t-m+1}, \ldots, p_i^t\} \tag{7}$$

To comply with references, we set observation length m = 16, 30 frames per second, the same as the benchmark in Kotseruba et al. (2021).

### 3.2. Architecture

The proposed multi-modal method shown in Fig. 2, illustrates the overall architecture. In Visual modality, local and global context from semantic image sequences are adopted as input of the prediction network. 2D convolution is adopted to extract features and then connected to the GRU module for temporal information extraction. In dynamic modality, relative distance and location attributes are introduced, along with the bounding box, pose key points, and real speed of the ego-car. All these dynamic features will be encoded by Interaction Encoding module. Two sequential encoding mechanisms are introduced to explore the feature interactions. Meanwhile, the estimated speed of pedestrians and the real speed of an ego-car in Dynamic Encoding will also be considered. An attention mechanism is adopted to learn the weights of multi-modalities. The details will be discussed in the following subsections.

### 3.3. Visual Modality

The objects in the view of cameras will affect the decision of target pedestrians. Take the below scenarios as an example, to determine pedestrian will cross or not: (1) The pedestrian is standing at a crossroad, and traffic lights turn to green. At the same time, a vehicle slow-moving along the sidewalk blocks the pedestrian. (2) pedestrian is standing in the middle of the road, other conditions are same as (1). The results may be different due to the location of the target pedestrian. Depending on the actual circumstances, all the possible surroundings will affect the pedestrian feature behaviors. In this work, we model these surroundings and interactions by the local context around the pedestrian and global context in the camera's view. Local context is denoted as $c_{li}^t$, cropped from the original frame with a size of 1.5 times bounding box, and records the changes around pedestrians. The global context, acquired from semantic segmentation maps, is denoted as $C_g$, and represents pixel-level semantic masks, localizing different road users in the image. From this context, all available space in the camera's view can be easily recognized, and the internal interactions can be easily modeled. A DeepLabV3 semantic segmentation model Chen et al. (2017), which is trained on Cityscapes DatasetCordts et al. (2016), will be used to acquire the segmentation masks to select critical objects (e.g. pedestrians, vehicles, sidewalks, street, and road).

A pre-trained VGG19 Simonyan and Zisserman (2014), is adopted to extract features. Images are resized and represented by a 4D array, denoted as [observed frames, rows, cols, channels]. The size of the extracted feature will change from ([512,14,14]) to tensor([16,512]) through the max-pooling layer to the average pooling layer (14x14). A stacked gated recurrent unit (GRU) is adopted for temporal correlation. The interactive information between the local scene and semantic maps is gradually incorporated. In the proposed architecture, GRUs (256 hidden units) are used to generate a tensor size ([16,256]).

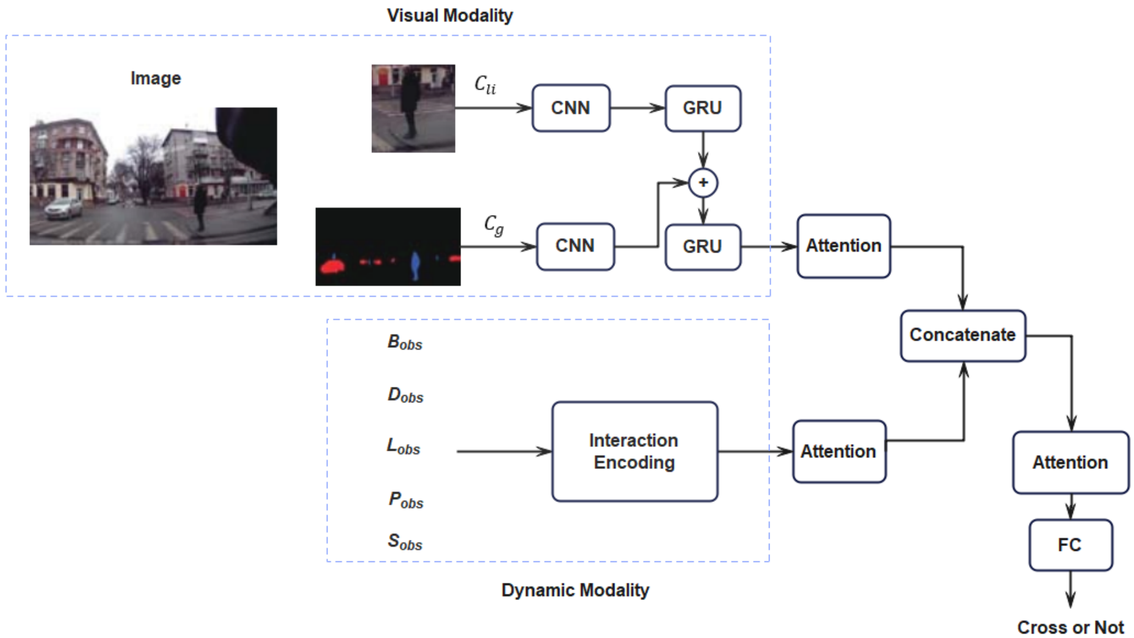


Figure 2: The proposed prediction framework. The input of the model includes: (1) features from visual modality: context features surrounding pedestrian $C_{li}$, semantic segmentation maps $C_g$; (2) features from dynamic modality: relative distance between target pedestrian $D_{obs}$, location of pedestrian in scene $L_{obs}$, pedestrian observed motion in bounding box $B_{obs}$, pose key points $P_{obs}$, and real speed of ego-car $S_{obs}$ are encoded in Interaction Encoding module. The extracted visual and dynamic features will be fed to stacked GRUs. An attention mechanism is adopted to learn the weights of multi-modalities. The final prediction will be output by FC layers.

### 3.4. Dynamic Modality

In crossing scenarios, the pedestrian's motion, location, and distance from the ego-car, as well as the real speed of the ego-car, are the important factors in the estimation of pedestrian crossing behaviors. Generally, pedestrians will remain static when the vehicle moves too fast or too close to the pedestrian. Besides, a pedestrian moving on the road will have a large probability of crossing compared to standing on the sidewalk. Considering the importance of the dynamics features, apart from the existing features, two kinds of novel features have been introduced in this paper: (1) the relative distance from pedestrian to ego-car, (2) scene location indicating the pedestrian's position on the road or sidewalk at the crossing point. Besides, an additional estimated speed of pedestrians is also introduced. The detailed descriptions are as follows:

#### 3.4.1. Pedestrian's Motion and Location in 2-D

Pedestrian location is denoted as $B_i = \{b_i^{t-m}, b_i^{t-m+1}, \ldots, b_i^t\}$. Due to the absence of 3D data, the coordinates ( top-left,bottom-right) of 2D bounding boxes are adopted to estimate the velocity of pedestrians. To formulate the location and velocity, the center points of the detected bounding box, along with the width and height, are calculated and denoted as $P_t = (x_t, y_t, w_t, h_t)$. The $V_t$ represents the position changes from $t-1$ in $\Delta t$:

$$V_t = \frac{P_t - P_{t-1}}{\Delta t} = (\Delta x_t, \Delta y_t, \Delta w_t, \Delta h_t) \tag{8}$$

The novel vectors $B_t = (P_t, V_t)$ of pedestrians consist of position and speed vectors, while t is time steps.

#### 3.4.2. Pedestrian's Relative Distance

Two public datasets, JAAD and PIE, are collected by the wide-angle RGB camera. They don't have real-world coordinates from Lidar or GPS, so there is no distance information in datasets. So, deploying the distance from actual obstacles to the vehicle becomes a challenging problem. Even though real distance can't be acquired, a relative distance could be evaluated to simulate the spatial position relation and scope to a certain extent. Usually, the pixel coordinates of the bounding boxes can somehow depict the distance. The monocular depth estimation approach estimates the distance from each pixel of the obstacle to the camera. This work introduces relative distances derived by bounding boxes and depth maps as novel attributes.

Fig. 3 shows the process of simulating distance. $d_i^t$ denotes the relative distance at time t. $b_i^t[k]$, where k from 0 to 3, denotes top-left to bottom-left coordinates separately in the bounding box.

$$d_i^t = \sum_{i=b_i^t[0]}^{b_i^t[3]} \sum_{j=b_i^t[0]}^{b_i^t[3]} I(i,j) \tag{9}$$

where, $I(i,j)$ depicts the pixel value of monocular depth image.

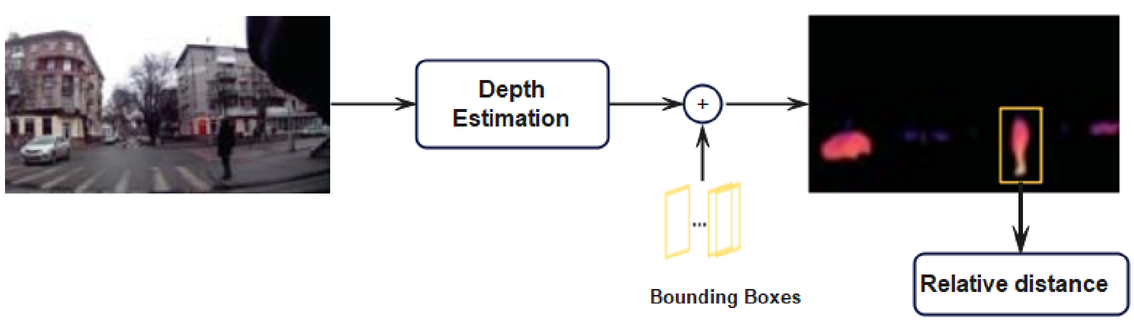


Figure 3: Illustration of the process of relative distance framework

The depth images generated from Godard et al. (2019), combined with the bounding box, will derive the relative distance of pedestrians, this could be achieved by calculating the mean pixel value of the cropped area by bounding box.

### 3.4.3. Pedestrian's Location

Scene location, indicating the pedestrian is standing on the road or sidewalk at a crossing point, will reflect the crossing intention. In this work, we introduce this attribute in the group of dynamic features. Generally, pedestrians on the road will have a higher probability of crossing than standing sidewalks. We simplify the semantic map generated from Chen et al. (2017) into interesting categories, "sidewalk", and "road/street". Fig. 4 shows the process of scene location attribute, denoted $L_i = \{l_i^{t-m}, l_i^{t-m+1}, \ldots, l_i^t\}$ as "road" or "sidewalk", same as in Yang et al. (2022).

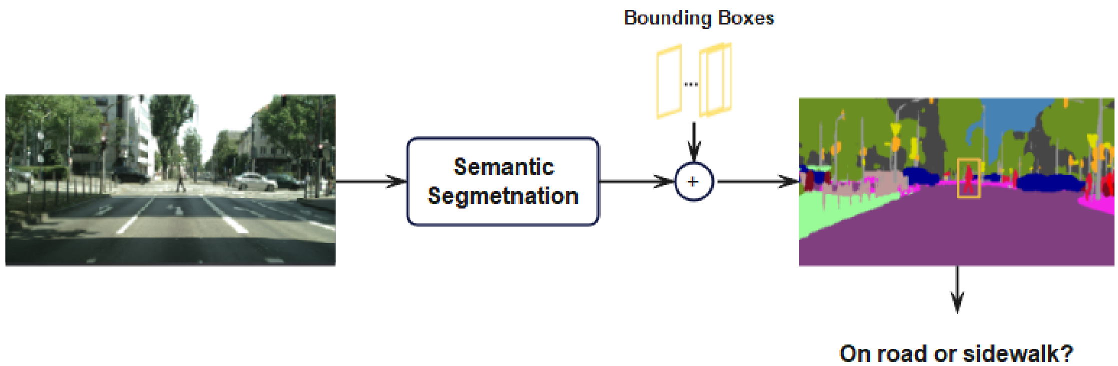


Figure 4: The architecture of scene location attribute segmentation in the input module. All the semantic segmentation is generated by Chen et al. (2017).

### 3.4.4. Real speed of Ego-car

Real speed of ego-car $s^t$ is defined by the ground truth of PIE, while only timestamped behavior labels in JAAD dataset. To process easily, the descriptions provided in JAAD dataset are adopted as the represented speed: "4" vs accelerating, "3" vs decelerating, "2" vs moving fast, "1" vs moving slow, and "0" vs stopped.

### 3.4.5. Pose key points

Similar to Yang et al. (2022), the Pose key points are obtained by applying a pose estimation model on the local context $C_{li}$. JAAD dataset does not provide ground truth of pose key points, a pre-trained OpenPose model Cao et al. (2017b) is adopted to extract pose key points $P_i = \{p_i^{t-m}, p_i^{t-m+1}, \ldots, p_i^t\}$, where $p$ is a 36D vector of 2D coordinates that contain 18 pose joints, i.e.,

$$p_i^{t-m} = \{x_{i1}^{t-m}, y_{i1}^{t-m}, x_{i2}^{t-m}, y_{i2}^{t-m}, \ldots, x_{i18}^{t-m}, y_{i1}^{t-m}\} \tag{10}$$

### 3.4.6. Interaction Encoding

In this paper, two types of interaction encoding are introduced. First is sequential encoding, and second is group encoding. Similar to Kotseruba et al. (2021); Yang et al. (2022), dynamics features will be fed to the neural network sequentially in Fig. 5. However, the interactions between features are not well described. A second group encoding is introduced to explore the specific interactions between features in Fig. 6. Based on the understanding of crossing, the speed of the ego-car with the speed of pedestrians, and the distance between pedestrian and ego-car cooperate for the final decision simultaneously. The standing location and pose information will indicate motion states to a certain extent.

Figure 5: Sequential Interaction Encoding for dynamics features

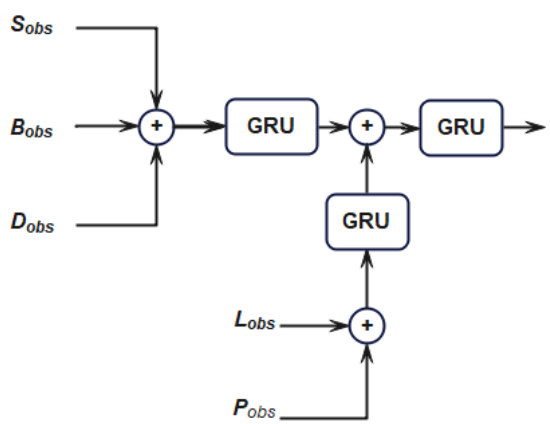


Figure 6: Group Interaction Encoding for dynamics features

#### 3.4.7. Attention module

The attention mechanism learns to put weights on multiple features among feature representations. Only the last frame will be focused. The weight $\alpha$ is as follows:

$$\alpha = \frac{\exp\left(\operatorname{score}\left(h_t, \bar{h}_s\right)\right)}{\sum_{s'} \exp\left(\operatorname{score}\left(h_t, \bar{h}_{s'}\right)\right)}$$

, where $h_t$ and $\bar{h}_s$ represent the last hidden state and each in observed period t. The score $\left(h_t, \bar{h}_s\right) = h_t^{\top} W_a \bar{h}_s$. $W_a$ denotes weight matrix. $c_t = \sum_i \alpha_i \bar{h}_s$ denotes sum of all attention weighted hidden states. A simple concatenation layer is adopted to produce tensor size [16,256]. The final output is denoted as:

$$Y_{\text{attention}} = \tanh\left(W_c\left[c_t; h_t\right]\right) \tag{11}$$

## 4. Experiment

The proposed framework is evaluated on JAAD Rasouli et al. (2017c,a)and PIE Rasouli et al. (2019) datasets. Totally 346 clips for crossing the road in JAAD. Two subsets:

Table 1: Performance of our method with state-of-the-arts on JAAD and PIE datasets

| Models | PIE | | | $JAAD_{beh}$ | | | $JAAD_{all}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | ACC | AUC | F1 | ACC | AUC | F1 | ACC | AUC | F1 |
| PCPAKotseruba et al. (2021) | 0.86 | 0.86 | 0.77 | 0.58 | 0.50 | 0.71 | 0.85 | 0.86 | 0.68 |
| OSUYang et al. (2022) | 0.82 | 0.78 | 0.68 | 0.62 | 0.54 | 0.74 | 0.83 | 0.82 | 0.63 |
| SF-GRURasouli et al. (2020) | 0.87 | 0.85 | 0.78 | 0.53 | 0.53 | 0.59 | 0.84 | 0.84 | 0.65 |
| MCIPHam et al. (2022) | 0.89 | 0.87 | 0.81 | 64 | 0.55 | 0.78 | 0.88 | 0.84 | 0.66 |
| Hybrid-Seq(ours) | 0.90 | 0.87 | 0.81 | 0.65 | 0.56 | 0.79 | 0.88 | 0.82 | 0.68 |
| Hybrid-Group(ours) | 0.91 | 0.89 | 0.81 | 0.67 | 0.61 | 0.79 | 0.90 | 0.84 | 0.69 |

JAADbeh (JAAD behavioral) and JAADall (JAAD all). All pedestrians in JAADall and pedestrians with behaviors are annotated in JAADbeh. All pedestrians in the view are annotated in PIE. Camera internal parameter matrices provided in the dataset correct the image distortion before feeding into the semantic and depth representations. The same configuration as in Kotseruba et al. (2021) is adopted to create a fair benchmark. The overlap of data sampling is set to 0.8, the scale of the context surrounding pedestrians is set to 1.5, the L2 regularization dropout to 0.001, and the dropout is set to 0.5. JAAD is trained for 80 epochs, PIE is trained for 60 epochs set lr as $5x10^-6$. Adam optimizer and binary cross-entropy loss are adopted.

### 4.1. Comparison with State-of-the-art

The comparison with the state-of-the-art on PIE dataset and two JAAD sub-datasets is listed in Table 1. Four benchmarks are adopted in this work to evaluate the proposed framework. C3D is adopted in PCPAKotseruba et al. (2021) to extract spatial-temporal relationships. OSUYang et al. (2022) and SF-GRURasouli et al. (2020) explore different fusion between multiple features. MCIPHam et al. (2022) introduces a segmentation map into non-visual and visual modules to predict crossing intention. Besides, the global and local context information, our methods introduce distance and location as the additional inputs compared to benchmarks. From the results, our hybrid, with sequential and group interactions method achieves the highest accuracy of 90%. The above results show that the distance and location in the scene can provide additional information that could remove redundant correlations between real scenes and pedestrians. Besides, the fusion strategy between dynamic features will slightly impact the performance of our two results on two datasets. The results also reveal dynamic features will interact with each other and group features sequentially may lose some interaction.

#### 4.1.1. Ablation Study

We conduct the ablation study to evaluate the individual features and impact on final prediction results by excluding one feature sequentially. As Table 2 shows, context $C_{li}$ surrounding pedestrian and global semantic context $C_g$ along with pedestrian's observed motion $B$, pedestrian's pose $P$, real speed of ego-car $S$, relative distance $D$, and location $L$, comprise the baseline for fusion of dynamic features. From the results, the features with great impact are the speed of the ego-car, whose accuracy decreased by about 6%, followed

Table 2: Ablation study with individual features

| B | P | S | D | L | $C_{li}$ | $C_g$ | ACC↑ | AUC↑ | F1↑ |
|---|---|---|---|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.91 | 0.89 | 0.81 |
| x | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.89 | 0.88 | 0.79 |
| ✓ | x | ✓ | ✓ | ✓ | ✓ | ✓ | 0.88 | 0.85 | 0.73 |
| ✓ | ✓ | x | ✓ | ✓ | ✓ | ✓ | 0.85 | 0.83 | 0.71 |
| ✓ | ✓ | ✓ | x | ✓ | ✓ | ✓ | 0.89 | 0.88 | 0.80 |
| ✓ | ✓ | ✓ | ✓ | x | ✓ | ✓ | 0.89 | 0.88 | 0.79 |
| ✓ | ✓ | ✓ | ✓ | ✓ | x | ✓ | 0.90 | 0.86 | 0.76 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | 0.89 | 0.89 | 0.79 |
| ✓ | ✓ | ✓ | x | x | ✓ | ✓ | 0.88 | 0.86 | 0.78 |
| ✓ | ✓ | ✓ | x | ✓ | ✓ | ✓ | 0.89 | 0.89 | 0.78 |
| ✓ | ✓ | ✓ | ✓ | x | ✓ | ✓ | 0.88 | 0.87 | 0.78 |

Table 3: Ablation study with different fusion sequences of dynamic information on PIE and JAAD

| Models | PIE | | | $\text{JAAD}_{beh}$ | | | $\text{JAAD}_{all}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | ACC | AUC | F1 | ACC | AUC | F1 | ACC | AUC | F1 |
| S+ B+ P | 0.88 | 0.86 | 0.78 | 0.61 | 0.53 | 0.72 | 0.82 | 0.74 | 0.58 |
| S+ B + P + D | 0.89 | 0.88 | 0.78 | 0.64 | 0.56 | 0.74 | 0.86 | 0.81 | 0.63 |
| S+ B + P + L | 0.88 | 0.87 | 0.76 | 0.62 | 0.53 | 0.73 | 0.83 | 0.73 | 0.57 |
| S+ B + D + L + P | 0.91 | 0.88 | 0.81 | 0.65 | 0.56 | 0.74 | 0.88 | 0.80 | 0.64 |
| B+ D + L + P + S | 0.90 | 0.88 | 0.80 | 0.64 | 0.56 | 0.75 | 0.88 | 0.81 | 0.65 |

by pedestrian motion, relative distance, pose, location, local context, and global context decreased by about 1-2% on the PIE dataset. The complex background information in the global context may contribute less to crossing intention than other features. Besides, there is an accuracy improvement of about 3% with two novel features $D$ and $L$, and individually, about 1-2% improvement. The results prove that the efficiency of the novel two features and the more corresponding features, the better the performance. Furthermore, it depicts that the proposed framework is more concerned with interaction around the pedestrian.

Table 3 shows the performance for the different sequential fusions of dynamic features. With the distance $D$ from pedestrian to ego-car added, the overall accuracy is improved by more than 2%. The only location in scene information $L$, there is still performance improvement, which is slightly lower than distance information. Besides, the sequence with distance, location, and observed motion performs best. It depicts that the proposed framework is concerned more with interaction around the pedestrian.

As Table 4 shows, location in scene information $L$ combines with pedestrian's observed motion $B$, pose kye points $P$ will give an accurate prediction accuracy as these three factors usually work together. With the distance $D$ from pedestrian to ego-car $S$ added, the overall accuracy is improved by more than 2%. Only changing fusion of dynamic information, while visual information keeping local and global sequence, has been proven effective benchmarks in this section. We follow the common group accepted by visual perception. For example,

Table 4: Performance of the proposed method with different fusion sequences of dynamic information on JAAD

| Models | JAAD$_{beh}$ | | | JAAD$_{all}$ | | |
|---|---|---|---|---|---|---|
| | ACC | AUC | F1 | ACC | AUC | F1 |
| $\mathrm{C}_{li} + C_g + B + S + P + D + L$ | 0.64 | 0.55 | 0.74 | 0.88 | 0.80 | 0.64 |
| $\mathrm{C}_{li} + C_g + B + D + P + L + S$ | 0.64 | 0.56 | 0.75 | 0.88 | 0.82 | 0.65 |
| $\mathrm{C}_{li} + C_g + B + L + D + P + S$ | 0.65 | 0.56 | 0.77 | 0.89 | 0.81 | 0.66 |
| $\mathrm{C}_{li} + C_g + D + L + S + B + P$ | 0.64 | 0.54 | 0.75 | 0.88 | 0.81 | 0.65 |

a man standing and facing the road will have a large probability of crossing the road, which means a group of $P$ and $L$. A long-distance $D$ with pedestrian's location $L$, pose $P$, and motion $B$ will have comprehensive prediction results. Therefore, the group order of features will impact the final prediction. This experiment shows that if an interaction exists between features, the correct group of feature sequences will achieve better results.

Figure 7: Samples related to distance and sence location

Furthermore, an ablation study about sequential fusion is conducted. Table 4 shows that location in scene information $L$ combines with pedestrian's observed motion $B$, pose kye points $P$ will give an accurate prediction accuracy as these three factors usually work together. With the distance $D$ from pedestrian to ego-car $S$ added, the overall accuracy is improved by more than 2%. This experiment shows that if there exists an interaction between features, the correct order of feature sequences will achieve better results.

### 4.2. Qualitative Results

Figure 7 displays some samples from the proposed model evaluated on JAAD dataset and PIE dataset. With additional distance and location information, novel interaction with ego-car and surroundings is further explored. Whether a pedestrian stands at crossing points


Figure 8: PCPA Kotseruba et al. (2021) and proposed models Qualitative results.

or on the street has a large probability for future motion. Some complicated samples are shown in Figure 8, which require more information to perform prediction. Besides, changing moving direction suddenly, bad weather conditions (e.g., bad illumination caused by rainy or snowy light), would affect prediction results.

## 5. Conclusion

In this paper, a novel crossing intention prediction framework is proposed. The proposed method explicitly considers the interactive information between surroundings and pedestrians. Two novel interactive features, distance from pedestrian to ego-car and location of pedestrian in the scene, are introduced. The relative distance derived from the monocular depth and semantic segmentation map, respectively, as the complement of provided dynamic features. Results show that more additional dynamic features, both from the visual model and proved by the dataset, will generate obvious results compared to hidden visual information. Two fusion strategies are proposed to explore the feature interactions: sequential and group features. Based on results, real scenarios considering features sequence and group will give better results than solely sequential. Future work can focus on feature fusion improvement around target pedestrians for the robustness of prediction. More stable features will be explored for complicated scenarios, such as sudden changes, occlusion, and bad illumination.

## References

Apratim Bhattacharyya, Daniel Olmeda Reino, Mario Fritz, and Bernt Schiele. Euro-pvi: Pedestrian vehicle interactions in dense urban centers. In *CVPR*. IEEE Computer Society, 2021a.

Apratim Bhattacharyya, Daniel Olmeda Reino, Mario Fritz, and Bernt Schiele. Euro-pvi: Pedestrian vehicle interactions in dense urban centers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6408–6417, 2021b.

Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Pedestrian graph+: A fast pedestrian crossing prediction model based on graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21050–21061, 2022.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017a.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017b.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aur'elien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset.

Zhijie Fang and Antonio M López. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 21(11): 4773–4783, 2019.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

Je-Seok Ham, Kangmin Bae, and Jinyoung Moon. Mcip: Multi-stream network for pedestrian crossing intention prediction. In *European Conference on Computer Vision*, pages 663–679. Springer, 2022.

Je-Seok Ham, Dae Hoe Kim, NamKyo Jung, and Jinyoung Moon. Cipf: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3665–3674, 2023.

Kai Holländer, Mark Colley, Enrico Rukzio, and Andreas Butz. A taxonomy of vulnerable road users for hci based on a systematic literature review. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13, 2021.

Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems*, 2021a.

Parth Kothari, Brian Sifringer, and Alexandre Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15556–15566, 2021b.

Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1258–1268, 2021.

Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15118–15129, 2021.

Javier Lorenzo, Ignacio Parra Alonso, Rubén Izquierdo, Augusto Luis Ballardini, Álvaro Hernández Saz, David Fernández Llorca, and Miguel Ángel Sotelo. Capformer: Pedestrian crossing action prediction using transformer. *Sensors*, 21(17):5694, 2021a.

Javier Lorenzo, Ignacio Parra, and MA Sotelo. Intformer: Predicting pedestrian intention with the aid of the transformer architecture. *arXiv preprint arXiv:2105.08647*, 2021b.

Taylor Mordan, Matthieu Cord, Patrick Pérez, and Alexandre Alahi. Detecting 32 pedestrian attributes for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11823–11835, 2021.

Nada Osman, Enrico Cancelli, Guglielmo Camporese, Pasquale Coscia, and Lamberto Ballan. Early pedestrian intent prediction via features estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3446–3450. IEEE, 2022.

Francesco Piccoli, Rajarathnam Balakrishnan, Maria Jesus Perez, Moraldeepsingh Sachdeo, Carlos Nunez, Matthew Tang, Kajsa Andreasson, Kalle Bjurek, Ria Dass Raj, Ebba Davidsson, et al. Fussi-net: Fusion of spatio-temporal skeletons for intention prediction network. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pages 68–72. IEEE, 2020.

Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 264–269, 2017a.

Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017b.

Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017c.

Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6261–6270, 2019.

Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. *arXiv preprint arXiv:2005.06582*, 2020.

Amir Rasouli, Mohsen Rohani, and Jun Luo. Bifold and semantic reasoning for pedestrian behavior prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15600–15610, 2021a.

Amir Rasouli, Mohsen Rohani, and Jun Luo. Bifold and semantic reasoning for pedestrian behavior prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15600–15610, 2021b.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Xingchen Song, Miao Kang, Sanping Zhou, Jianji Wang, Yishu Mao, and Nanning Zheng. Pedestrian intention prediction based on traffic-aware scene graph model. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9851–9858. IEEE, 2022.

Chen Sun, Zejian Deng, Wenbo Chu, Shen Li, and Dongpu Cao. Acclimatizing the operational design domain for autonomous driving systems. *IEEE Intelligent Transportation Systems Magazine*, 2021.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith Redmill, and Umit Ozguner. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, 2022.

Sibo Zhang, Yuexin Ma, Ruigang Yang, Xin Li, Yanliang Zhu, Deheng Qian, Zetong Yang, Wenjing Zhang, and Yuanpei Liu. Cvpr 2019 wad challenge on trajectory prediction and 3d perception. *arXiv preprint arXiv:2004.05966*, 2020.