# PEDESTRIAN CROSS FORECASTING WITH HYBRID FEATURE FUSION

**MENG DONG {mengyingyidai}@gmail.com**

## ABSTRACT

Forecasting the crossing intention of pedestrians is critical for the growth of Autonomous Vehicles (AVs) in the real world. Pedestrians' movements are usually influenced by their surroundings in traffic scenes. Recent works extract explicit and individual information from collected data to perform prediction. However, there still exists much implicit information which is not considered ever, such as interactions between features, location of pedestrians, and distance towards the ego-car. Properly exploring and utilizing the implicit information will promote the prediction of future behaviors. To this end, the surrounding interactions from semantic segmentation and local context, together with two novel introduced attributes: the pedestrian's location at road or sidewalk, and the relative distance from target pedestrian to ego-car are adopted as critical features in this paper. The location and distance attributes are derived from the semantic map and depth map combined with bounding boxes information separately. A hybrid network based on multi-modal, which incorporates interactions between individual features, is proposed to forecast cross or not. Evaluated by two public pedestrian crossing datasets, PIE and JAAD, the proposed features and fusing strategy achieve state-of-the-art performance.

## 1 INTRODUCTION

Pedestrians, as the main participants in traffic roads, are easy to violate rules and be 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); Varytimidis et al. (2018); Alahi et al. (2008); Ridel et al. (2018). 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 necessarily indicate future movements in real traffic environments. Pedestrians may change their directions and velocities suddenly in dynamics 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). Figure 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 always keep their eyes and ears open when they are prepared to cross. So it is a multi-feature problem to predict pedestrian crossing intention. Recently, many public data sets related to pedestrians of automotive driving Rasouli et al. (2017c;a; 2019a); Sun et al. (2020b); 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. (2019a), 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 kind of methods employ the multi-modal network to fusion multiple encoded features for final prediction results. However, such fusion network will bear heavy computing load when the two features are similar or correlated. To reduce potential redundancy and introduce the interaction between the tar-

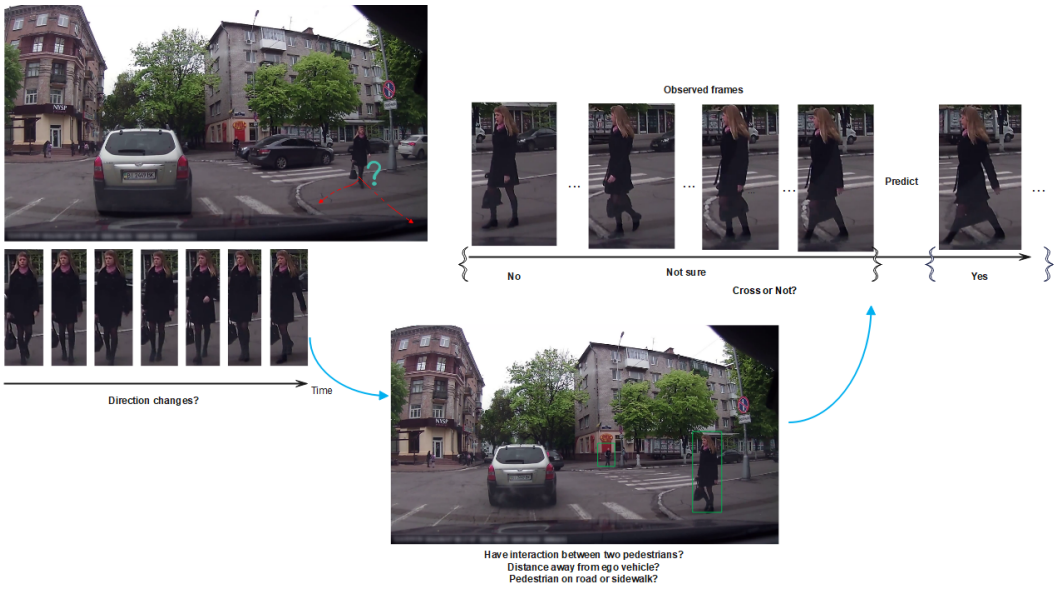


Figure 1: Pedestrian dynamic: change of moving direction; Interaction between pedestrian and other road users; Pedestrian on road or sidewalk?

get pedestrian and the scene, OSU Yang et al. (2022) proposes spatio-temproal context feature with an attention mechanism based on PCPA. However, the implicit interactionsRasouli et al. (2021b) between labeled features, and other potential features such as distance and location, which will influence the intention, are lack of consideration in the above methods. 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 target pedestrian and ego-car, location of target pedestrian (at road or sidewalk) accompany with present the existing features. Besides, novel fusion strategy are proposed to fully consider correlation or interaction between features. 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 are used to remove redundant interaction between pedestrians and other road users.
- We evaluate the performance of the proposed method using public datasets Rasouli et al. (2019b; 2017b), and show that our method achieves stable and better performance over state-of-the-art algorithms.

This paper is organized as follows: In Section 2, we review works on intention prediction, which are the basis of our experiments. In Section 3 we describe the proposed methodology. In Section 4, we describe the experimental results and ablation study. Finally, we summarize the contribution of this work in the concluding section.

## 2 Related Work

The problem of pedestrians intention prediction from image sequences has attracted significant interest recently. As a sub-problem in action prediction Joo et al. (2019); Felsen et al. (2017); Piccoli et al. (2020); Lu et al. (2017); Liang et al. (2017); Oliu et al. (2018); Mahmud et al. (2017); Bhattacharyya et al. (2018); Chen et al. (2018); Lee et al. (2017), pedestrian crossing intention also arises huge interest with the development of Autonomous Vehicles. The aim is to predict whether the target pedestrian crosses the road or not in the field of view of AVs for future several seconds.

### 2.1 PEDESTRIAN CROSSING INTENTION PREDICTION

Based on Mordan et al. (2021), pedestrians have been recognized 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 & López (2019) to estimate the future pose of pedestrians. The distance and angle among the joint points are calculated to predict whether pedestrian cross or not. Recently feature fusion methods have been explored for this problem. In Kotseruba et al. (2021); Yang 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.

### 2.2 DATASETS

Due to the development of automotive driving, the research on traffic prediction (motion prediction, interaction prediction and intention prediction) of ADS has achieved great progress. There have released several public datasets Ma et al. (2018); liu; Rasouli et al. (2017b; 2019a); Bhattacharyya et al. (2021a); Malla et al. (2020); Sun et al. (2020a); Ettinger et al. which consist of multiple data from different sensors, annotations of road users, and information from previous module before prediction. In this paper, two crossing intention datasets are evaluatedRasouli et al. (2017b; 2019a).

### 2.3 INTERACTION MODELING

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 the task of 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. 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 in the same 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.

### 2.4 DISTANCE

Pedestrians' distance to ego-vehicle is a key factor in the task of intention prediction. Only the sensors, such as lidar, stereo camera, and GPS, could generate accurate distance, otherwise, the real distance could not be calculated. The datasets, JAAD and PIE, collected by the RGB camera, lack of such information. Distance evaluation from image sequences also becomes a hot topic. Usually, the pixel coordinates of the bounding boxes can somehow depict the distance. A relative distance could be evaluated to simulate the relation of the spatial position to a certain extent. Recently, the monocular depth estimation approach Rasouli et al. (2020), proposed a pixel-level object distance estimation approach from images and achieves relatively accurate depth predictions. In order to simulate relative distances, a method derived from bounding boxes and depth maps is introduced in this paper.

### 2.5 FEATURE FUSION

Besides visual features, the non-visual attributes such as pedestrian's bounding box, key-points, ego-car motion, and other implicit features are modeled individually in many recent works. However, some features may be correlated in a real driving environment. A proposer strategy that could fuse all these features and explore their intra and inter interactions will affect the final decision. PCPAKotseruba et al. (2021) proposes an attention-based fusing mechanism based on JAAD and PIE to incorporate interactions. This benchmark also be extended in many worksHam et al. (2023); Yang et al. (2022); Rasouli et al. (2021a); Lorenzo et al. (2021); Osman et al. (2022) with different fusion strategies. Our proposed model is also developed on this benchmark, and a hybrid strategy is proposed to consider interactions between encoded features.

## 3 METHODS

### 3.1 FORMULATION

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 $i$:

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

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

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

(3) Ego car' real speed:

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

(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 $t$ :

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

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

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

(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:

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

(7) Pose keypoints of pedestrian $i$:

$$P_{obs} = \left\{p_{i}^{t-m}, p_{i}^{t-m+1}, \ldots, p_{i}^{t}\right\}$$

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

### 3.2 ARCHITECTURE

The proposed multi-modal method shown in Figure 2, illustrates the overall architecture.

In Visual modality, local context 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 GRU module for temporal information encoding. 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 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.

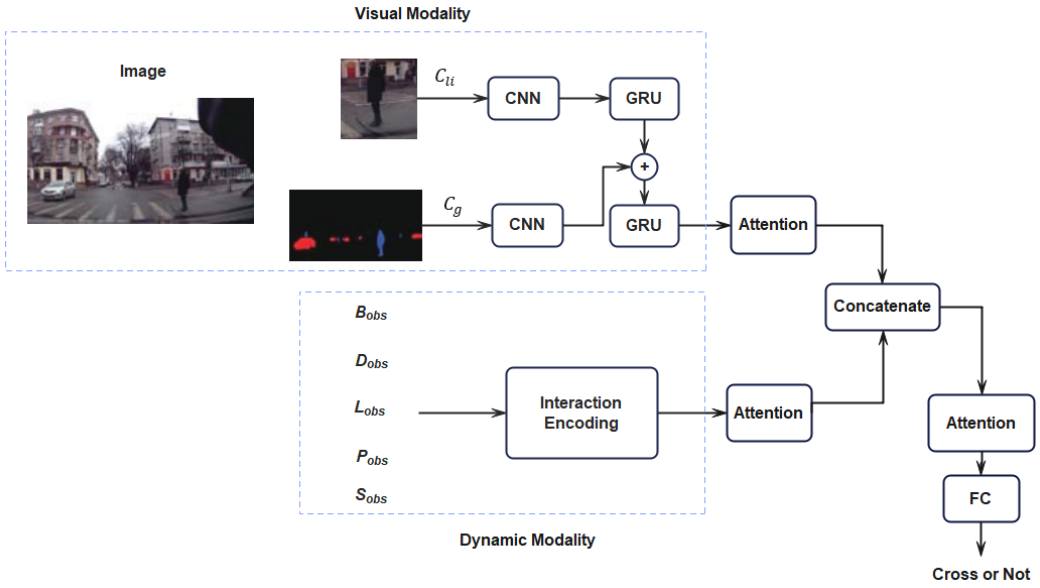


Figure 2: The proposed prediction framework. The input of the model includes: (1) features from visual modality: context features surrounding pedestrian $C_{li}$, semantic segmentation maps $C_g$; (2) features from dynamic modality: relative distance between target pedestrian $D_{obs}$, location of pedestrian 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.

### 3.3 VISUAL MODALITY

The objects in the view of cameras will affect the decision of target pedestrians. Taking the below scenarios for example to determine pedestrian will cross or not: (1) 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 target pedestrian. Depend on the actual circumstances, all the possible surroundings will affect the pedestrian feature behaviors. In this paper, we model these surroundings and interactions by local context around the pedestrian and global context in the view of camera. Local context is denoted as $c_{li}^t$, cropped from 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 camera view can be easily recognized, so the internal interactions in the view can also be easily modeled. A DeepLabV3 semantic segmentation model Chen et al. (2017) which trained on Cityscapes DatasetCordts et al. (2016) will be used to acquire the segmentation masks to select critical objects (e.g. pedestrians, vehicles, sidewalk, street, and road).

A pre-trained VGG19 Simonyan & 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 extracted feature will change from ([512,14,14]) to tensor([16,512]) through max-pooling layer to average pooling layer (14x14). Stacked gated recurrent unit (GRU) is adopted to temporal correlation. The interactive information between the local scene and semantic maps is gradually incorporated. In the proposed architecture, GRUs (256 hidden units) is used to generate a tensor size ([16,256]).

### 3.4 DYNAMIC MODALITY

In crossing scenarios, 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 keep static when the vehicle is moving too fast or too close to pedestrian. Besides, a pedestrian moving on the road will have a large probability to cross compared to standing on 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 pedestrian's position on road or sidewalk at crossing point. Besides, an additional estimated speed of pedestrian is also introduced. The detailed descriptions are as below:

#### 3.4.1 PEDESTRIAN'S MOTION AND LOCATION IN 2-D

Pedestrian location is denoted as $B_i = \left\{b_i^{t-m}, b_i^{t-m+1}, \ldots, b_i^t\right\}$. Due to the absence of 3D data, the coordinates ( top-left,bottom-right) of 2D bounding boxes are adopted to estimate the velocity of pedestrian. In order to formulate the location and velocity, the center points of 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 $\Delta t$:

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

The novel vectors $B_t = (P_t, V_t)$ of pedestrian 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-angles RGB camera. They don't have real world coordinates from Lidar or GPS, so there is no distance information in datasets. So how to deploy 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. Relative distances derived by bounding boxes and depth maps is introduced as a novel attribute in this work.

Figure 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 bounding box.

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

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

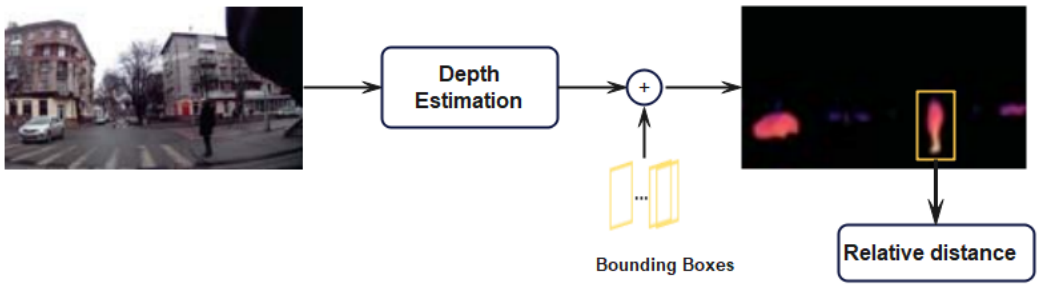


Figure 3: Illustration of the process of relative distance framework

The depth images generated from Godard et al. (2019), combined with the bounding box will derive 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 pedestrian is standing on road or sidewalk at crossing point, will reflect the crossing intention. In this work, we introduce this attribute in the group of dynamic features. Generally, pedestrian on road will higher probability to cross compared to standing side-walk. We simplify semantic map generated from Chen et al. (2017) into interesting categories, "sidewalk", "road/street". Figure 4 shows 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).

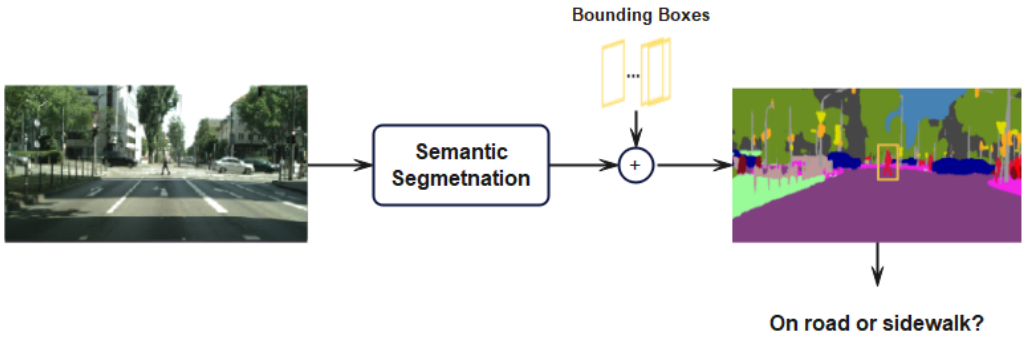


Figure 4: The architecture of scene location attribute segmentation in the input module. All the semantic segmentation is generated by Chen et al. (2017).

#### 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. In order 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 KEYPOINTS

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

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

#### 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 Figure 5. However, the inter interactions between features are not well described. A second group encoding is introduced to explore the specific interactions between features in Figure 6. Based on the understanding of crossing, the speed of ego-car with the speed of pedestrian, and the distance between pedestrian and ego-car cooperate for the final decision simultaneously. The standing location together with pose information will indicate motion states to a certain extent.

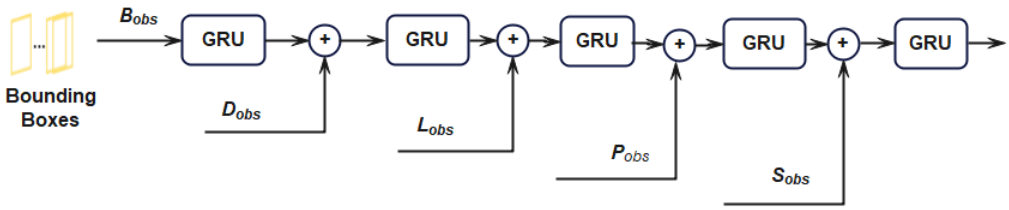


Figure 5: Sequential Interaction Encoding for dynamics features

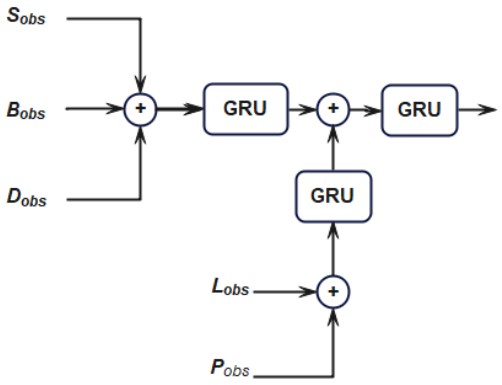


Figure 6: Group Interaction Encoding for dynamics features

#### 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:

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

,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 denoted as:

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

## 4 EXPERIMENTAL

The proposed model is trained on JAAD and PIE datasets. Totally 346 clips for crossing the road in JAAD. Two subsets: JAADbeh (JAAD behavioral) and JAADall (JAAD all). All pedestrians in JAADall are annotated, and pedestrians with behaivors are annotated in JAADbeh. All pedestrians in the view are annotated in PIE. Camera internal parameter matrix provided in the dataset are used to 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 and the scale of context surrounding pedestrians is set to 1.5, set L2 regularization dropout to 0.001, and 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. We use the metrics: accuracy (Acc),F1-score, and area under the ROC curve (AUC) to evaluate model, which is same as in the PCPA. The proposed model is implemented and trained on an Intel Core i9 CPU, 32 GB of RAM, and NVIDIA GeForce RTX 3060, 24GB VRAM.

Table 1: Performance of proposed 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 |
| Proposed-sequential | 0.89 | 0.88 | 0.79 | 0.64 | 0.56 | 0.75 | 0.88 | 0.81 | 0.66 |
| Proposed-group | 0.91 | 0.89 | 0.81 | 0.70 | 0.61 | 0.77 | 0.91 | 0.84 | 0.69 |

### 4.1 Ablation Study

Compared to benchmark Kotseruba et al. (2021); Yang et al. (2022), additional distance and location in scene are introduced in dynamic encoding. Only changing fusion of dynamic information, while visual information keeping local and global sequence, which has been proven effective benchmarks, and the above experiments are all conducted on JAAD dataset. Two kinds of ablation studies are conducted for dynamics feature fusion: First, two interaction encoding schemes, sequential and group, this ablation study is to evaluate the interactions between features. Second, feature order effect, will be conducted on sequential encoding.

### 4.2 Results

#### 4.2.1 Quantitative Results

The proposed approach compared to state-of-the-arts Kotseruba et al. (2021); Yang et al. (2022); Rasouli et al. (2020) on PIE dataset and two JAAD sub-datasets list in Table 1. In the JAADbeh and PIE dataset, the proposed sequential approach slightly better than state-of-the-arts in evaluation of "Acc ,AUC, F1 scores". Accuracy 0.89 is obtained by the proposed approach. The proposed group approach achieves better than the proposed sequential. The above results show that the distance and location in scene can provide additional information which could remove redundant correlations between real scenes and pedestrians. Besides, the group features will explore the interactions among features with better results.

#### 4.2.2 Qualitative Results

Figure 7 displays some samples from the proposed model at group features, 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 or on 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.

#### 4.2.3 Ablation Results

As Table 2 shows, context $C_{li}$ surrounding pedestrian and global semantic context $C_g$ along with pedestrian's observed motion $B$, pose kyepints $P$ and real speed of ego-car $S$, comprise the baseline for the different sequential fusion of dynamic features. With the distance $D$ from pedestrian to ego-car added, the overall accuracy is improved by more than 3%. 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, has the highest performance. It depicts that the proposed framework is concerned more with interaction around the pedestrian.

As Table 3 shows, location in scene information $L$ combines with pedestrian's observed motion $B$, pose kyepints $P$ will give a 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


Figure 7: Samples related to distance and sence location


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

than 2%. These experiment shows that the there exist interaction between features, correct order of feature sequences will achieve better 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 shows that more additional dynamic features both from visual model and proved by dataset, will generate obvious results compared to hidden in visual information. Two fusion strategies are proposed to explore the feature

Table 2: Performance of the proposed method with different features of dynamic information on JAAD

| **Models** | **JAAD**$_{beh}$ | | | **JAAD**$_{all}$ | | |
|---|---|---|---|---|---|---|
| | ACC | AUC | F1 | ACC | AUC | F1 |
| $\mathrm{C}_{li} + C_g + B + S + P$ | 0.61 | 0.53 | 0.72 | 0.82 | 0.74 | 0.58 |
| $\mathrm{C}_{li} + C_g + B + S + P + D$ | 0.64 | 0.56 | 0.74 | 0.86 | 0.81 | 0.63 |
| $\mathrm{C}_{li} + C_g + B + S + P + L$ | 0.62 | 0.53 | 0.73 | 0.83 | 0.73 | 0.57 |
| $\mathrm{C}_{li} + C_g + B + S + P + D + L$ | 0.64 | 0.55 | 0.74 | 0.88 | 0.80 | 0.64 |
| $\mathrm{C}_{li} + C_g + B + P + D + L + S$ | 0.64 | 0.56 | 0.75 | 0.88 | 0.81 | 0.65 |

Table 3: 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 |

interactions: sequential feature and group feature. Based on results, real scenarios based consideration on features sequence and group will give better results than solely sequential. Future work can focus on feature fusion improvement around target pedestrian for the robustness of prediction. More stable features will be explored for complicated scenarios, such as sudden changes , occlusion and bad illumination.

## REFERENCES


Alexandre Alahi, Michel Bierlaire, and Murat Kunt. Object detection and matching with mobile cameras collaborating with fixed cameras. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008*, 2008.

Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4194–4202, 2018.

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*, pp. 6408–6417, 2021b.

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*, pp. 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*, pp. 7291–7299, 2017b.

Lei Chen, Jiwen Lu, Zhanjie Song, and Jie Zhou. Part-activated deep reinforcement learning for action prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 421–436, 2018.

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*, pp. 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.

Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will happen next? forecasting player moves in sports videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 3342–3351, 2017.

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*, pp. 3828–3838, 2019.

Je-Seok Ham, Kangmin Bae, and Jinyoung Moon. Mcip: Multi-stream network for pedestrian crossing intention prediction. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pp. 663–679. Springer, 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*, pp. 1–13, 2021.

Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10873–10883, 2019.

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*, pp. 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*, pp. 1258–1268, 2021.

Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 336–345, 2017.

Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *proceedings of the IEEE international conference on computer vision*, pp. 1744–1752, 2017.

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*, pp. 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, 2021.

Chaochao Lu, Michael Hirsch, and Bernhard Scholkopf. Flexible spatio-temporal networks for video prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6523–6531, 2017.

Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. *arXiv preprint arXiv:1811.02146*, 2018.

Tahmida Mahmud, Mahmudul Hasan, and Amit K Roy-Chowdhury. Joint prediction of activity labels and starting times in untrimmed videos. In *Proceedings of the IEEE International conference on Computer Vision*, pp. 5773–5782, 2017.

Srikanth Malla, Behzad Dariush, and Chiho Choi. Titan: Future forecast using action priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11186–11196, 2020.

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.

Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 716–731, 2018.

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)*, pp. 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*, pp. 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)*, pp. 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*, pp. 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*, pp. 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)*, pp. 6261–6270, 2019a.

Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6262–6271, 2019b.

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 (ICCV)*, pp. 15600–15610, October 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*, pp. 15600–15610, 2021b.

Daniela Ridel, Eike Rehder, Martin Lauer, Christoph Stiller, and Denis Wolf. A literature review on the prediction of pedestrian behavior in urban scenarios. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3105–3112. IEEE, 2018.

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

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, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

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*, pp. 2446–2454, 2020b.

Dimitrios Varytimidis, Fernando Alonso-Fernandez, Boris Durán, and Cristofer Englund. Action and intention recognition of pedestrians in urban traffic. In *14th International Conference on Signal-Image Technology & Internet-Based Systems, SITIS 2018, Las Palmas de Gran Canaria, Spain, November 26-29, 2018*, pp. 676–682. IEEE, 2018. doi: 10.1109/SITIS.2018.00109. URL https://doi.org/10.1109/SITIS.2018.00109.

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.