CIPF: Crossing Intention Prediction Network based on Feature Fusion Modules for Improving Pedestrian Safety

Je-Seok Ham[1] Dae Hoe Kim[1*] NamKyo Jung[2*] Jinyoung Moon[1]
[1]Electronics and Telecommunications Research Institute (ETRI), South Korea
[2]Korea University, South Korea
[1]{jsham, dhkim19, jymoon}@etri.re.kr, [2]namkyo0724@gmail.com

Abstract

As the development of autonomous driving technology continues, pedestrian safety is becoming an increasingly important issue. The ability of an autonomous car to accurately predict whether a pedestrian will cross the road is essential for ensuring their safety, as the vehicle can slow down in time or stop to avoid any potential accidents. However, predicting pedestrian behavior is a complex task influenced by various environmental and contextual factors. To deal with this issue, we propose a novel method, Crossing Intention Prediction based on feature Fusion modules (CIPF) that combines eight different input features extracted from both pedestrians and vehicles through three fusion modules using RNN layers and attention mechanisms. We demonstrated state-of-the-art performance of prediction accuracy in the PIE dataset, which is the most widely used for pedestrian crossing intention prediction. We also demonstrated the superiority of the performance of our CIPF network through qualitative and quantitative analysis. In particular, we also performed ablation studies on the verification of the effectiveness of the eight input features, the validity of VGG encoders, and performance comparison of our CIPF over time by adjusting the prediction time.

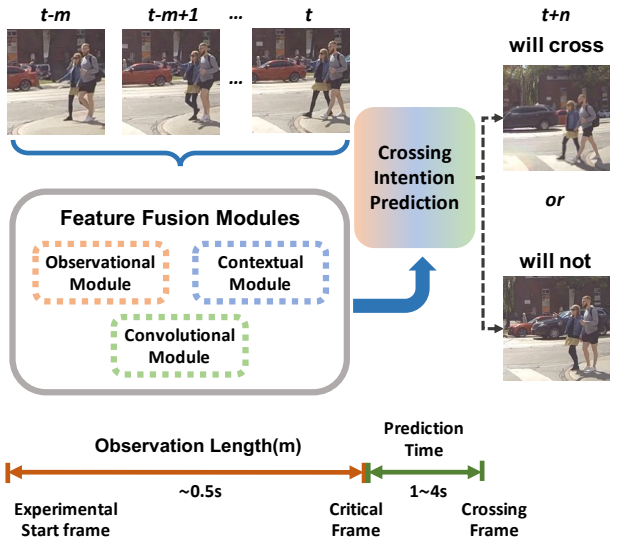

Figure 1. **The concept of the pedestrian crossing intention prediction.** Our proposed model, CIPF, which takes input features observed from *m* frames before to the current time *t* and passes them through three fusion modules - the observational, contextual, and convolutional modules - to extract the prediction results of whether the pedestrian will cross or not cross at future time *t+n*. For prediction, we define three frames: experimental start, critical, and crossing frames.

1. Introduction

The emergence of autonomous driving technology [12] has led to a growing focus on pedestrian safety and transportation convenience. One of the key technologies to achieving these goals is the ability to predict whether pedestrians will cross or not cross. Using the prediction results, the autonomous vehicle can slow down or stop to prevent any accidents related to pedestrians [27, 35]. However, predicting pedestrian behavior is not easy because the intention of humans is unclear, and there are many external factors [3, 41, 42] that influence their behavior, such as interaction with other pedestrians, traffic signs, road congestion, and vehicle speed. These external factors may affect the future actions of pedestrians in the traffic environment. As a result, anticipating crossing intentions in advance is significantly challenging.

Some datasets for predicting pedestrian behavior including videos recorded by on-board cameras equipped in vehicles have been publicized. With datasets, their prediction models have also been provided as a baseline. The baseline models include Standard-TRI Intent Prediction (STIP) [16]

*These authors contributed equally

using graph convolution techniques and Pedestrian Vehicle Interactions in Dense Urban Centers (Euro-PVI) [2] through interactions with vehicles for predicting pedestrian trajectories, and Trajectory Inference using Targeted Action priors Network (TITAN) [22] for anticipating future pedestrian behavior. Among the datasets, the most widely used dataset is the Pedestrian Intention Estimation (PIE) dataset [25] and there have been proposed prediction models based on the PIE dataset. The models depend on which features are used as inputs that are selected from the annotations or derived from visual information in input videos, how to combine them in their network architecture, and whether the attention mechanism is applied or not. As a result, recent studies [21, 32, 40, 43] have focused on how much the performance of prediction accuracy is improved by their own distinct combination method from the diverse feature modalities. However, these approaches are still limited due to the complex factors that influence pedestrian behavior, which can be visually observed in the videos. In addition, there is a lack of neural networks that can fully accommodate the diversity of input features.

To overcome these previous limitations, in this paper, we propose a novel fusion approach using all available input features for anticipating pedestrian crossing intention. Our proposed network, Crossing Intention Prediction based on feature Fusion modules (CIPF), consists of three types of Fusion Modules (FM). FM fuses a total of eight input features, including both non-visual and visual features, as shown in figure 1. Each feature employed a different process based on its attributes. The non-visual features are processed by an observational module, while the visual features including contextual properties are processed by a contextual module composed of a CNN and an RNN. The visual features without contextual properties are processed by a convolutional module composed of a 3D convolutional network and 3D max-pooling, followed by a flatten layer and an FC layer. By efficiently combining multiple feature inputs, our CIPF outperforms the state-of-the-art methods with a prediction accuracy of 91% on the PIE dataset which is the most commonly used for pdestrian behavior research.

The mainstream for predicting pedestrian crossing intention involves collecting data from the pedestrian and vehicles during past observations in specific frames and using the data to anticipate crossing intention at a future point in one or two seconds, as shown in figure 1. Our proposed model also follows this protocol. The experimental start frame is defined as the beginning of the pedestrian observation, and the critical frame is defined as the start of the prediction period. The interval between the experimental start frame and the critical frame is defined as the observational length, and the prediction time is defined as the interval from the critical frame to the crossing frame when the pedestrian takes action to cross or not cross the street.

Our main contributions are summarized as follows:

- We propose a novel feature fusion model, CIPF, which utilizes eight input modalities with a systematic combination mechanism.
- We verify the effectiveness of CIPF by achieving performance outperforming the state-of-the-art methods on the PIE dataset through extensive experiments.
- We provide ablation studies analyzing the effects of eight features, validating the visual encoder, and comparing performance by adjusting the prediction time.
- We introduce the qualitative analysis of prediction results with pedestrian crossing intention.

2. Related Work

To improve pedestrian safety, considerable research has been devoted to developing pedestrian crossing intention prediction. For the pedestrian crossing intention prediction dataset, the JAAD (Joint Attention in Autonomous Driving) dataset [28] was first introduced, providing annotations for behavioral analysis of pedestrians at the point of crossing. And it proposed 2D CNN-based baseline approaches for pedestrian crossing intention prediction. [10] showed adopting human pose-based features with 2D CNN features could improve pedestrian crossing intention prediction. However, the JAAD dataset is limited in that most pedestrian samples have the intention of crossing.

To solve the drawback, [25] introduces a large-scale dataset named PIE (Pedestrian Intention Estimation). To consider the temporal context of crossing behavior, it proposed LSTM network combines visual information and past trajectory information. In [14], the dataset was analyzed by human experiments to identify which visual features correlate to pedestrian crossing intention. The analysis showed that locations of designated crosswalks, orientations of pedestrians, locations with regard to curbs, and whether pedestrians look at the traffic are good predictors of intention. In addition, it demonstrated that considering intention values from human experiments improves the prediction performance of crossing intention with a single RNN model. Further improving prediction performances, [26] studied SF-GRU (Stacked with multilevel fusion GRU), which fuses five different pieces of information, including pedestrian appearance, surrounding context, poses, bounding boxes, and ego-vehicle speed. Many studies also have shown that temporal context is essential for crossing intention prediction that adopts recurrent layers such as LSTM [3, 19, 30]. Further, an attention mechanism [20] is adopted, which can focus on specific parts of input features, thus better for analyzing sequential input [13, 15, 23, 29, 36, 37]. Recently, given the success of

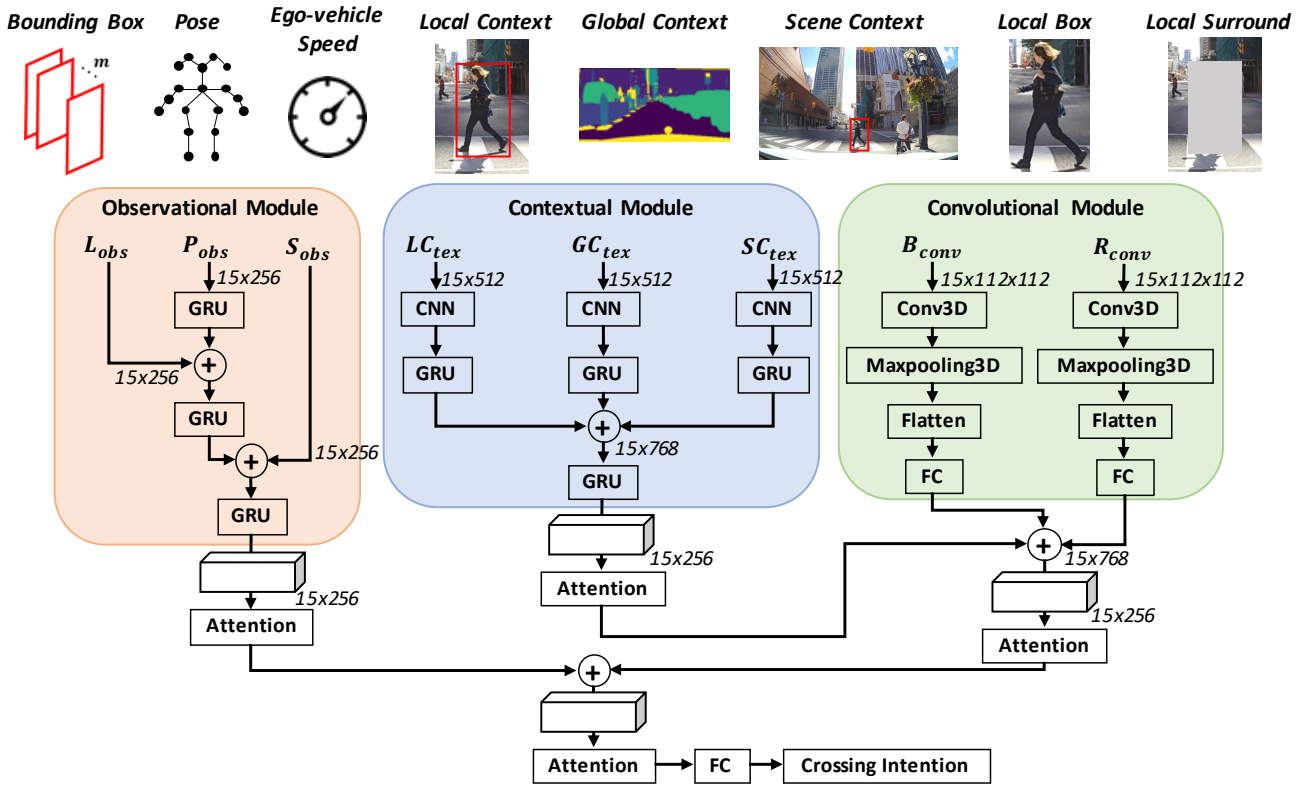

Figure 2. **The overview of the our proposed framework.** CIPF is divided into the three modules for eight input features: Pose(P_{obs}), Bounding Box(L_{obs}), Ego-vehicle Speed(S_{obs}), Local Context(LC_{tex}), Global Context(GC_{tex}), Scene Context(SC_{tex}), Local Box(B_{conv}), Local Surround(R_{conv}). The observational module sequentially stacks pose, bounding box, and ego-vehicle speed features. The contextual module takes in local context, global context, and scene context inputs extracted by CNN encoder. Finally, the convolutional module utilizes C3D encoder for local box and local surround features. Each module is recurrently processed using GRU layer and extracted crossing intention prediction output through an attention module.

transformers in many applications, some researchers have accepted transformer models [17, 18, 33], which benefit from reducing training time compared to LSTM, at the same time, improving the pedestrian crossing intention prediction performances. Instead of extracting visual context by convolution, [16] utilizes graph convolution to infer the spatiotemporal relationships how objects in the scene are related. The method builds a spatiotemporal scene graph and applies it to segmented object instances in video frames. To this end, they devised pedestrian-centric and location-centric graphs to extract rich features from observed frames. In [8], graph convolutional autoencoders are adopted to embed visual features of the pedestrian or scene objects to the graph.

3. CIPF: Crossing Intention Prediction Network based on Feature Fusion Modules

We propose a novel prediction model, Crossing Intention Prediction based on feature Fusion Modules (CIPF), as shown in figure 2. CIPF is a network that predicts pedestrian crossing intention in advance by combining eight input features. CIPF is divided into three modules - the observational module, contextual module, and convolutional module - each of which receives separate inputs and is configured to process input values differently. Ultimately, the outputs from each module are fused to predict crossing intention.

3.1. Model Input Acquisition

3.1.1 Observational Module

There are three inputs used in observational module: pose and bounding box of pedestrian and speed of ego-vehicle. These three features are stacked in GRU layer, starting with the pose feature. The *Pose* feature P_{obs} is defined as:

$$P_{obs} = \{p_i^{t-m}, p_i^{t-m+1}, ..., p_i^t\}, \quad (1)$$

where pose keypoints are generated using Openpose [5, 6, 31, 34], which estimates the pose of a person by capturing 18 keypoints joints from the mouth, neck, shoulders,

elbows, wrists, hips, knees, ankles, eyes, and ears. The 18 keypoints are represented by a 36-dimensional vector, p_i, where each keypoint is composed of 2D coordinates. Observational length, *m*, which means the number of observed frames, is set to 15. The *Bounding Box* feature L_{obs} is defined as:

$$L_{obs} = \{l_i^{t-m}, l_i^{t-m+1}, ..., l_i^t\}, \tag{2}$$

where $l_i = [x_1, y_1, x_2, y_2] \in \mathbb{R}^4$ is a 2D bounding box which means the coordinates determined by top-left ($[x_1, y_1]$) and bottom-right ($[x_2, y_2]$) of each pedestrian. As bounding box consists of 4 coordinates, the dimension of L_{obs} is determined as $m \times 4$. The *Ego-vehicle Speed* is defined as:

$$S_{obs} = \{s_i^{t-m}, s_i^{t-m+1}, ..., s_i^t\}, \tag{3}$$

where s_i refers to the speed of the ego-vehicle, and this feature is already included in the annotations of the PIE dataset. This means that it was generated using the sensor of the ego vehicle at the time of data acquisition. However, in the JAAD dataset, there is no speed information available, and only vehicle action information is provided. Therefore, this feature cannot be used in the JAAD dataset.

3.1.2 Contextual Module

In the context module, the inputs utilized are comprised of three types of contexts: local context, global context, and scene context. Each of these context input features is obtained by extracting the input features using an ImageNet pre-trained VGG19 network as the backbone CNN with maximum pooling layer. Then, each of these features is recursively processed using a GRU. The outputs of the three GRUs are concatenated and combined into a single output, which is then fed into an attention mechanism. The *Local Context* feature LC_{tex} is defined as:

$$LC_{tex} = \{lc_i^{t-m}, lc_i^{t-m+1}, ..., lc_i^t\}, \tag{4}$$

The image surrounding pedestrians including crosswalks, traffic lights, and intersection signs is an essential element for predicting pedestrian behavior contextually. Therefore, lc_i refers to an RGB image of 224×224 pixels cropped at 1.5 times the size of the bounding box of pedestrian. The input feature vector extracted as (*m*, 512) is then processed through a max pooling layer with a 14 x 14 kernel to obtain an (*m*, 256) vector, where *m* represents the observation length. The *Global Context* feature GC_{tex} is defined as:

$$GC_{tex} = \{gc_i^{t-m}, gc_i^{t-m+1}, ..., gc_i^t\}, \tag{5}$$

where gc_i refers to the semantic segmentation values extracted using the pre-trained DeepLabV3 model from the Cityscapes Dataset [39]. By utilizing the extracted semantic map values, global scene or road information can be mainly employed. Similar to local context, the features are extracted as an (*m*, 256) vector and combined with a concatenation operation. The *Scene Context* feature SC_{tex} is defined as:

$$SC_{tex} = \{sc_i^{t-m}, sc_i^{t-m+1}, ..., sc_i^t\}, \tag{6}$$

where sc_i, the scene context, means the entire image, not just the area around the pedestrian. Similar to other context features, the entire image is resized to 224x224 pixels and processed through a (14, 14) kernel with padding by setting their output dimension to equal for fusing with other contextual features.

3.1.3 Convolutional Module

A convolutional module consists of two inputs: local box and local surround. Both inputs are extracted through a Convolutional 3D network (C3D), and the dimension of images is decreased($112 \rightarrow 56 \rightarrow ... \rightarrow 4$) by repeatedly feeding through max-pooling layers. Eventually, the features are arranged through a flatten layer, resulting in a one-dimensional array of data. The *Local Box* feature B_{con} is defined as:

$$B_{conv} = \{b_i^{t-m}, b_i^{t-m+1}, ..., b_i^t\}, \tag{7}$$

where b_i means an image feature that is generated by cropping image the same as the size of a bounding box, padding it, and then resizing the long side of the cropped image to match the desired output size. The remaining parts of the image are zero-padded. The last output of feature is then fed through five C3D networks with max-pooling layer, and then through a flatten layer to adjust its size for concatenation with other features. The *Local Surround* feature R_{con} is defined as:

$$R_{conv} = \{r_i^{t-m}, r_i^{t-m+1}, ..., r_i^t\}, \tag{8}$$

where r_i is an image cropped to 1.5 times the size of the bounding box as local context feature, however, the central area corresponding to the bounding box coordinates is neutral grayed out to keep only the surrounding context of the bounding box. This allows the utilization of information from the pedestrian's surrounding area. C3D is applied in the same way as for the local box feature to concat with b_i.

3.2. Recurrent Module

To consider the temporal context of input features, we used gated recurrent unit (GRU) [9], which is a simpler layer than LSTM [24, 38] Recursion for the equation of GRU, the variables at the j^{th} level of the stack are described as follows,

$$z_j^t = \sigma(x_j^t W_j^{xz} + h_j^{t-1} W_j^{hz} + b_j^z), \tag{9}$$

$$r_j^t = \sigma(x_j^t W_j^{xr} + h_j^{t-1} W_j^{hr} + b_j^r), \tag{10}$$

$$\tilde{h}_j^t = tanh(x_j^t W_j^x + (r_j^t \odot h_j^{t-1}) W_j^h + b), \tag{11}$$

$$h_j^t = (1 - z_j^t) \odot h_j^{t-1} + z_j^t \odot \tilde{h}_j^t, \tag{12}$$

where $\sigma(\cdot)$ is the logistic sigmoid function, x_j^t is the input feature at time step t, W are weights between two units, r_j^t and z_j^t are reset and update gates at time step t, respectively; h_j^{t-1} and h_j^t represent the hidden state at the previous time step and current time step, respectively.

3.3. Attention Module

The attention mechanism [20] is adopted to focus on specific parts of input features, thus better for analyzing input features. The output vector of the attention module is defined as below:

$$\beta_{attention} = \tanh(W_c[h_c : h_m]), \tag{13}$$

$$h_c = \sum_{s^t} \alpha_t h_{s^t}, \tag{14}$$

where W_c is a weight matrix, m is the observation length, h_m is the last hidden state of the encoder, h_c is the sum of all attention weighted hidden states, h_{s^t} is the previous hidden state of the encoder, and α_t is the attention weight vector. The attention weight vector is defined as below:

$$\alpha_t = \frac{\exp(score(h_m, \tilde{h}_s))}{\sum_{s^t=1}^{T} \exp(score(h_m, \tilde{h}_{s^t}))}, \tag{15}$$

where $score(h_m, h_{s^t}) = h_m^T W_p h_s$ is the content-based function and W_p is a trainable weight matrix.

4. Experiment

4.1. PIE dataset

To verify the effectiveness of the proposed method, experiments have been conducted on Pedestrian Intention Estimation (PIE) dataset [25]. The dataset is most widely used for predicting pedestrian crossing intention. The dataset was recorded during the daytime with a dashboard camera in downtown Toronto, Canada. Videos were recorded with HD format (1920 × 1080) 30FPS per 10 minutes, so the total video length of six sets is 6 hours. In addition, the OBD (On-Board Diagnostics) sensor was attached inside the ego-vehicle, measuring speed of vehicle, heading direction, and GPS coordinates. The dataset provides 1,842 tracks of pedestrians who are close to the road. Each track contains annotations of the pedestrian in frame sequences, such as a bounding box and a crossing status, whether the pedestrian is crossing the road or not in the given frame. Related to crossing intention, the tracks are divided into four classes. 1) Pedestrians who intend to cross and actually cross (512 tracks). 2) Pedestrians who intend to cross but do not actually cross (898 tracks). 3) Pedestrians who have no intention of crossing but actually cross (2 tracks). 4) Pedestrians who have no intention of crossing and do not actually cross (430 tracks). In the experiments, tracks were randomly divided into three sets; 48% (880 tracks) of the training set, 39% (719 tracks) of the testing set, and 13% (243 tracks) of the validation set, respectively. Table 1 shows detailed properties of the PIE dataset.

	Property
FPS(Frames per Second)	30
Length of each chunk	10min
Total number of frames	909K
Total number of annotated frames	293K
Number of pedestrians with bounding boxes	739K
Total number of pedestrians	1,842
Crossing intention and do cross	512
Crossing intention and don't cross	898
No crossing intention to cross	430

Table 1. Properties of the PIE dataset

4.2. Evaluation Metrics

To compare the performance of our proposed model with previously developed models, we utilized the five widely used metrics in pedestrian crossing intention prediction field: accuracy (ACC), area under the ROC curve (AUC), precision, recall, and F1 score.

Accuracy measures how accurately the model predicts the binary classification problem of crossing intention, defined as follows,

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{16}$$

where *TP* represents the quantity of true positive samples, *TN* represents the quantity of true negative samples, *FP* represents the quantity of false positive samples, and *FN* represents the quantity of false negative samples, respectively.

F1 score is the harmonic mean of the precision and recall, defined as follows,

$$F1 = \frac{2 \times precision \times recall}{precision + recall}, \tag{17}$$

AUC is the base area of the ROC (Receiver Operation Characteristic) curve defined as follows [4],

$$AUC(f) = \frac{\sum_{t_0 \in \mathcal{D}^0} \sum_{t_1 \in \mathcal{D}^1} \mathbf{1}[f(t_0) < f(t_1)]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|}, \tag{18}$$

Model	Visual Encoder	Features	ACC	AUC	F1	Precision	Recall
SingleRNN [14]	VGG + GRU	P, L, S, B, R	0.76	0.64	0.45	0.63	0.36
MultiRNN [1]	VGG + GRU	P, L, S, B	0.86	0.80	0.73	0.80	0.67
SFRNN [26]	VGG + GRU	P, L, S, B, R	0.83	0.77	0.68	0.72	0.64
PCPA [15]	C3D + GRU	P, L, S, LC	0.86	0.85	0.77	0.70	0.85
	I3D + GRU	P, L, S, LC	0.87	0.86	0.79	0.75	0.84
	VGG + GRU	P, L, S, LC	0.87	0.86	0.77	0.75	0.79
MCIP [11]	C3D + GRU	P, L, S, LC, GC	0.86	0.86	0.78	0.72	0.85
	I3D + GRU	P, L, S, LC, GC	0.85	0.81	0.74	0.76	0.71
	VGG + GRU	P, L, S, LC, GC	0.89	0.87	0.81	0.81	0.81
CIPF(Ours)	VGG + C3D + GRU	$P, L, S, LC, GC, SC, B, R$	**0.91**	**0.89**	**0.84**	0.85	0.83

Table 2. **Prediction performance comparison on PIE dataset.** For each model, the visual encoder and input features are introduced, and their performance is compared. The abbreviation of input feature is as follows. {*P*: Pose, *L*: Bounding Box, *S*: Speed of ego-vehicle, *LC*: Local Context, *GC*: Global Context, *SC*: Scene Context, *B*: Local Box, *R*: Local Surround}

where f is a predictor that returns the probability of crossing intention, $\mathbf{1}[f(t_0) < f(t_1)]$ denotes the indicator function that returns 1 if $f(t_0) < f(t_1)$ otherwise 0, $\mathcal{D}^0$ is a set of negative samples and $\mathcal{D}^1$ is a set of positive samples. A higher AUC indicates that the model is capable of effectively distinguishing between the different classes.

4.3. Experimental Setting

The proposed our novel algorithm was implemented on 8 GPUs with Nvidia RTX A6000 and AMD EPYC 7513 32-core processors with tensorflow environment. CIPF was trained using RMSProp optimizer with a learning rate of 5×10^{-5} during 300 epochs on PIE dataset. Also, we used GRUs with 256 hidden units for observational and contextual modules and C3D networks for convolutional module. We also included a dropout of 0.5 after the attention block for preventing overfitting and added an L2 regularization with 0.0001 to the last fully connected layer.

4.4. Performance Comparison

In this section, we compared the performance results of our novel CIPF network with five other benchmarks for pedestrian crossing intention prediction models. *SingleRNN* [14] is an encoder-decoder structure using RNN, which combines bounding box and ego vehicle speed into a single vector that is fed through the encoder. Then, the last hidden state passes through an FC layer, and it is concatenated with the crossing intent value to become the input of the decoder. *MultiRNN* [1] has the advantage of outputting uncertainty in predictions through Bayesian modeling, allowing for some understanding of incorrect predictions in uncertain situations. Stacked with Multilevel Fusion RNN (*SFRNN*) [26] is a network based on stacked RNNs, where the upper layer RNNs receive the hidden states of the lower layer RNNs as input. Then, the hidden states from the lower layers are combined with other input data and passed to the higher layers. Pedestrian Crossing Prediction with Attention (*PCPA*) [15] extracts local context images using C3D network and applies temporal, modality attention mechanisms. Multi-Stream Network for Pedestrian Crossing Intention Prediction (*MCIP*) [11] divides five inputs, including a segmentation map, into non-visual and visual modules, and applies an attention module to extract crossing intention.

Table 2 shows a performance comparison of the five benchmark models and our proposed model based on visual encoder and input features at a future point in one second. SingleRNN, MultiRNN, and SFRNN used VGG for feature extraction and did not include contextual input. PCPA and MCIP experimented with C3D, I3D [7], and VGG as visual encoders, and did not include convolutional features. Our proposed CIPF received all eight features, where some of the visual features are extracted using VGG, and others are encoded with C3D, achieving the highest prediction accuracy of 91%.

4.5. Qualitative Results

As shown in Figure 3, this is a visualization of pedestrian crossing intention prediction results. The observation started from the *t-15* frame, continued until the *t* frame, and then anticipated the pedestrian crossing intention at time t. The actual behavior of the pedestrian was then compared at the future *t+1* frames. The green boxes in *case 1* indicate predictions of pedestrians with a crossing intention, while the red boxes in other cases indicate predictions of pedestrians without a crossing intention. In particular, *case 3* shows an instance where the initial prediction was a crossing intention, but then correctly predicted a lack of crossing intention. In *case 4*, the model incorrectly predicted that the pedestrian would not cross in the next frame *t+1* even though the pedestrian does cross.

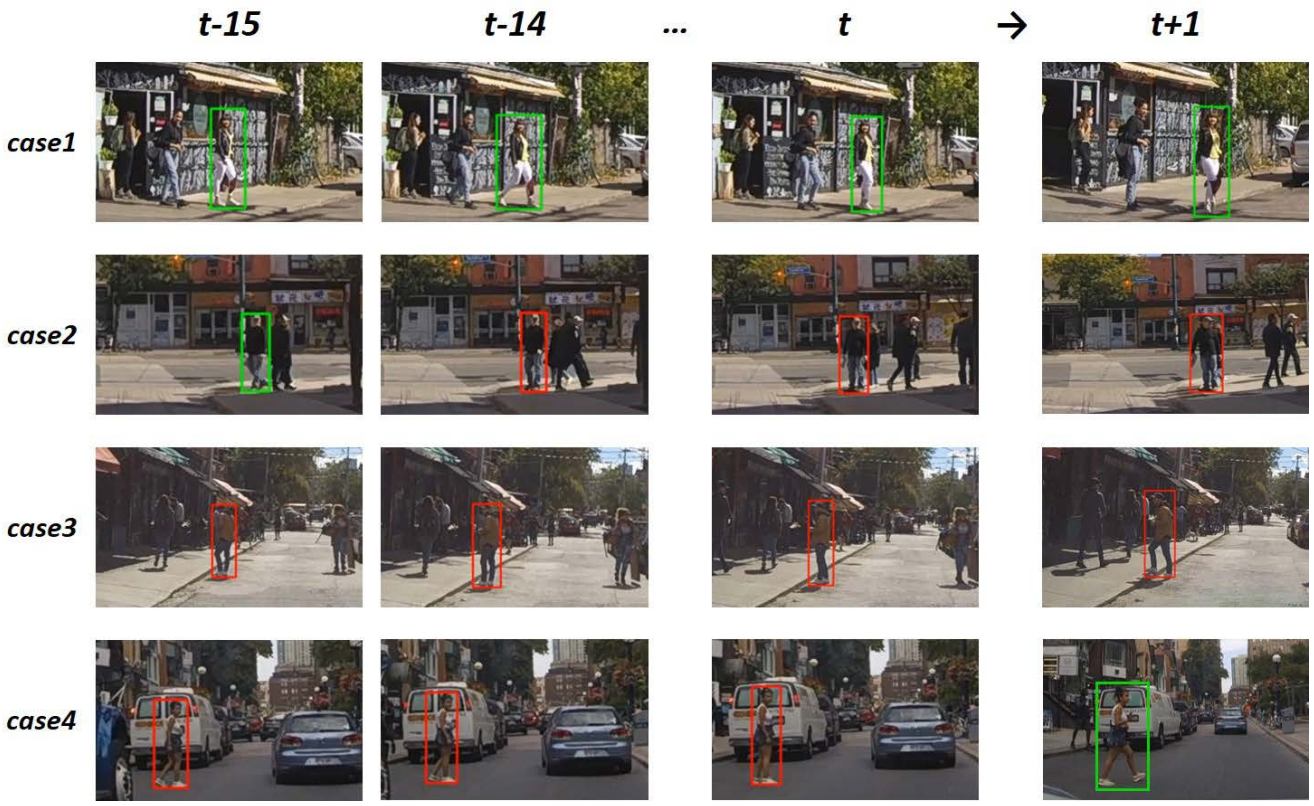

Figure 3. **The qualitative results of CIPF.** This is the predicted result based on the observation from the *t-15* frame to the *t* frame. The green boxes indicate predictions of an intention to cross, while the red boxes indicate predictions of no intention to cross.

4.6. Ablation Studies

4.6.1 Effectiveness of Input Features

To demonstrate the contributions of each eight input features, we examined an ablation study while excluding one of the inputs on the proposed algorithm. As presented in Table 3, the eight input features, the bounding box had the greatest impact on the accuracy, which is 6.3% lower than the baseline performance. Since the bounding box coordinates represent the location of pedestrians, it can be seen that the pedestrian's location plays the most crucial role in improving the accuracy of crossing intention. The feature that had the most negligible impact on the prediction accuracy were scene context and global context only 1% and 1.3% respectively lower than the baseline performance. In the case of scene context, it did not even include the image around the pedestrian, but the whole image, so it did not significantly impact the prediction. Moreover, in the case of global context, since the whole road environment was semantically mapped, not only the target pedestrian but also various noise factors such as trees or buildings were included, which contributed little to improving prediction accuracy. In addition, it was experimentally proven that the order of features that had less impact on the accuracy was local box $<$ local context $<$ local surround $<$ speed $<$ pose. It can be observed that the input features of the observational module have a more significant impact on the accuracy, while those of the contextual module has less impact.

4.6.2 Validity of Visual Encoder

We investigate the performance differences through the type of VGG encoders to change the input features of the convolutional module to be extracted with VGG instead of C3D to prove the effect of C3D. Among eight features, CIPF achieved state-of-the-art performance by extracting contextual features and feeding them through a GRU layer, while using C3D to extract convolutional features and repeatedly perform 3D maxpooling. As shown in table 4, instead of using C3D to extract input features for local box and local surrounds, the VGG was used to extract these features, and the results were compared to those obtained with C3D. *CIPF-BR* processed both convolutional and contextual features in the same way by passing each of the five inputs through a VGG and GRU unit before concatenating them and applying an attention module. *CIPF-B* extracted local surround using C3D and used VGG for the remaining features, while

P	*L*	*S*	*LC*	*GC*	*SC*	*B*	*R*	ACC	AUC	F1	Precision	Recall
✓	✓	✓	✓	✓	✓	✓	✓	0.911	0.888	0.843	0.851	0.834
	✓	✓	✓	✓	✓	✓	✓	0.855-0.056	0.871-0.017	0.782-0.061	0.686-0.165	0.910+0.076
✓		✓	✓	✓	✓	✓	✓	0.848-0.063	0.845-0.043	0.759-0.084	0.693-0.158	0.839+0.005
✓	✓		✓	✓	✓	✓	✓	0.857-0.054	0.836-0.052	0.758-0.085	0.730-0.121	0.789-0.045
✓	✓	✓		✓	✓	✓	✓	0.887-0.024	0.868-0.020	0.806-0.037	0.788-0.063	0.824-0.010
✓	✓	✓	✓		✓	✓	✓	0.898-0.013	0.876-0.012	0.822-0.021	0.820-0.031	0.824-0.010
✓	✓	✓	✓	✓		✓	✓	0.901-0.010	0.884-0.004	0.830-0.013	0.816-0.035	0.844+0.010
✓	✓	✓	✓	✓	✓		✓	0.888-0.023	0.879-0.009	0.814-0.029	0.773-0.078	0.859+0.025
✓	✓	✓	✓	✓	✓	✓		0.859-0.052	0.876-0.012	0.788-0.055	0.692-0.159	0.915+0.081

Table 3. **Ablation studies of CIPF based on the eight input features.** The order of the input features that have the most impact on performance is as follows: Bounding Box(*L*)>Pose(*P*)>Speed of ego-vehicle(*S*)>Local Surround(*R*)>Local Context(*LC*)>Local Box(*B*)>Global Context(*GC*)>Scene Context(*SC*).

Model	VGG	C3D	ACC	AUC	F1
CIPF-BR	LC, GC, SC, B, R		0.90	0.88	0.82
CIPF-B	LC, GC, SC, B	R	0.90	0.89	0.83
CIPF-R	LC, GC, SC, R	B	0.89	0.87	0.81
CIPF	LC, GC, SC	B, R	0.91	0.89	0.84

Table 4. **Performance comparison depending on visual encoder type.** The performance is compared according to whether local box(*B*) and local surround(*R*) are used as C3D input in the visual encoder or not.

	MCIP			CIPF		
time	ACC	AUC	F1	ACC	AUC	F1
4s	0.78	0.74	0.62	0.78	0.74	0.61
3s	0.79	0.77	0.66	0.80	0.80	0.7
2s	0.83	0.81	0.72	0.84	0.83	0.74
1s	0.89	0.87	0.81	0.91	0.89	0.84

Table 5. **Prediction time studies for future point.** For the both two models, MCIP and CIPF, the prediction preformance from 1 second to 4 seconds later are compared with each other.

CIPF-R extracted only local box using C3D. The prediction accuracy of *CIPF-R* was 2% lower than that of CIPF, and both *CIPF-BR* and *CIPF-B* showed a 1% lower accuracy. Therefore, it has been proved that using C3D to extract convolutional features is the most effective approach.

4.6.3 Prediction Time Studies

We evaluated our proposed model at different anticipation times, from 1s to 4s. We compared the prediction accuracy, AUC, and F1 performance of our proposed CIPF and the latest prediction model, MCIP, on the PIE dataset, as shown in Table 5. Both models showed a reduction in performance for all three metrics as the prediction time increases, and the interval with the most significant prediction accuracy drop was from the 1-second later prediction to 2 seconds, particularly for CIPF, which decreased by 7%. The prediction accuracy gradually reduced from the 3 seconds later prediction to 4 seconds interval, and at this interval, the two models decreased by around 1% (MCIP) and 2% (CIPF), respectively. Overall, our proposed CIPF model showed similar or higher accuracy compared to MCIP depending on the different prediction times.

5. Conclusion

In this paper, we introduce a new feature fusion module, CIPF, for predicting pedestrian crossing intention. CIPF is a network that efficiently fuses various inputs by separating modules depending on the properties of each feature, utilizing past visual or non-visual features extracted from pedestrians or vehicles. We achieved state-of-the-art performance in predicting whether a pedestrian will cross at a future point. We also experimented with the impact of eight input features on performance, verified the validity of the visual encoder type, and examined the performance difference when increasing the prediction time. In addition, we demonstrated the pedestrian crossing intention prediction process with qualitative results. By utilizing this network, it is expected to greatly contribute to improving pedestrian safety in the autonomous driving environment by anticipating pedestrian crossing intention in advance.

6. Acknowledgement

This work was supported by IITP grant funded by the Korea government(MSIT) (No.2020-0-00004, Development of Previsional Intelligence based on Long-term Visual Memory Network)

References

[1] 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 (CVPR)*, pages 4194–4202, 2018. 6

[2] 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 (CVPR)*, 2021. 2

[3] Smail Ait Bouhsain, Saeed Saadatnejad, and Alexandre Alahi. Pedestrian intention prediction: A multi-task perspective. *ArXiv preprint arXiv:2010.10270*, 2020. 1, 2

[4] Toon Calders and Szymon Jaroszewicz. Efficient auc optimization for classification. In *Knowledge Discovery in Databases: PKDD 2007*, pages 42–53, 2007. 5

[5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 3

[6] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[8] Tina Chen, Renran Tian, and Zhengming Ding. Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3096–3102, 2021. 3

[9] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. 4

[10] Zhijie Fang and Antonio M. López. Is the pedestrian going to cross? answering by 2d pose estimation. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1271–1276, 2018. 2

[11] 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*, pages 663–679, 2022. 6

[12] Hyunsuk Kim, Woojin Kim, Jungsook Kim, Seung-Jun Lee, Daesub Yoon, Oh-Cheon Kwon, and Cheong Hee Park. Study on the take-over performance of level 3 autonomous vehicles based on subjective driving tendency questionnaires and machine learning methods. *ETRI Journal*, 45(1):75–92, 2023. 1

[13] Kyungdo Kim, Yoon Kyung Lee, Hyemin Ahn, Sowon Hahn, and Songhwai Oh. Pedestrian intention prediction for autonomous driving using a multiple stakeholder perspective model. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2

[14] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1688–1693, 2020. 2, 6

[15] 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 (WACV)*, pages 1258–1268, 2021. 2, 6

[16] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters (RA-L)*, 2020. 1, 3

[17] 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*, 2021. 3

[18] Javier Lorenzo, Ignacio Parra, and MA Sotelo. Intformer: Predicting pedestrian intention with the aid of the transformer architecture. *ArXiv preprint arXiv:2105.08647*, 2021. 3

[19] Javier Lorenzo, Ignacio Parra, Florian Wirth, Christoph Stiller, David Fernández Llorca, and Miguel Angel Sotelo. Rnn-based pedestrian crossing prediction using activity and pose-related features. In *IEEE Intelligent Vehicles Symposium (IV)*, 2020. 2

[20] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015. 2, 5

[21] Jun Ma and Wenhui Rong. Pedestrian crossing intention prediction method based on multi-feature fusion. *World Electric Vehicle Journal*, 13(8):158, 2022. 2

[22] 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 (CVPR)*, 2020. 2

[23] 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, 2022. 2

[24] Ruijie Quan, Linchao Zhu, Yu Wu, and Yi Yang. Holistic lstm for pedestrian trajectory prediction. *IEEE transactions on image processing*, 30:3229–3239, 2021. 4

[25] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6261–6270, 2019. 2, 5

[26] Amir Rasouli, Iuliia Kotseruba, and John Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 49.1–49.13, 2019. 2, 6

[27] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Agreeing to cross: How drivers and pedestrians communicate. In *IEEE Intelligent Vehicles Symposium (IV)*, 2017. 1

[28] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 206–213, 2017. 2

[29] Amir Rasouli, Tiffany Yau, Mohsen Rohani, and Jun Luo. Multi-modal hybrid architecture for pedestrian action prediction. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 91–97, 2022. 2

[30] Haziq Razali, Taylor Mordan, and Alexandre Alahi. Pedestrian intention prediction: A convolutional bottom-up multi-task approach. *Transportation Research Part C: Emerging Technologies*, 2021. 2

[31] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[32] Ankur Singh and Upendra Suddamalla. Multi-input fusion for practical pedestrian intention prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2311, 2021. 2

[33] Ze Sui, Yue Zhou, Xu Zhao, Ao Chen, and Yiyang Ni. Joint intention and trajectory prediction based on transformer. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 3

[34] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[35] Biao Yang, Weiqin Zhan, Pin Wang, Chingyao Chan, Yingfeng Cai, and Nan Wang. Crossing or not? context-based recognition of pedestrian crossing intention in the urban environment. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5338–5349, 2021. 1

[36] 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 (T-IV)*, 2022. 2

[37] Yu Yao, Ella Atkins, Matthew Johnson-Roberson, Ram Vasudevan, and Xiaoxiao Du. Coupling intent and action for pedestrian crossing behavior prediction. In *International Joint Conference on Artificial Intelligence (IJCAI)*, page 1238 – 1244, 2021. 2

[38] Shile Zhang, Mohamed Abdel-Aty, Jinghui Yuan, and Pei Li. Prediction of pedestrian crossing intentions at intersections based on long short-term memory recurrent neural network. *Transportation research record*, 2674(4):57–65, 2020. 4

[39] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[40] Xingchen Zhang, Panagiotis Angeloudis, and Yiannis Demiris. 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, 2022. 2

[41] Junxuan Zhao, Yinfeng Li, Hao Xu, and Hongchao Liu. Probabilistic prediction of pedestrian crossing intention using roadside lidar data. *IEEE Access*, 7:93781–93790, 2019. 1

[42] Shengzhe Zhao, Haopeng Li, Qiuhong Ke, Liangchen Liu, and Rui Zhang. Action-vit: Pedestrian intent prediction in traffic scenes. *IEEE Signal Processing Letters*, 29:324–328, 2021. 1

[43] Xiao Zhou, Hongyu Ren, Tingting Zhang, Xingang Mou, Yi He, and Ching-Yao Chan. Prediction of pedestrian crossing behavior based on surveillance video. *Sensors*, 22(4):1467, 2022. 2