PIT: Progressive Interaction Transformer for Pedestrian Intention Prediction

Yuchen Zhou , Jie Zhu , Yaokun Li , Chao Gou*
School of Intelligent Systems Engineering, Shenzhen Campus of Sun Yat-sen University
{zhouych37, zhuj95, liyk58}@mail2.sysu.edu.cn, gouchao@mail.sysu.edu.cn

Abstract

For autonomous driving, one of the major challenges is to predict pedestrian crossing intention in ego-view. Pedestrian intention depends not only on their own goals, but also on stimulation of surrounding traffic elements. Most previous methods focus on representation learning of spatio-temporal features for detected pedestrians, instead of trying to capture the dynamic interaction relationship between traffic elements as human drivers do. In this work, inspired by neuroscience that humans' understanding of natural vision is obtained through progressive stimulation, we propose a framework termed as Progressive Interaction Transformer (PIT) for pedestrian crossing intention prediction. In particular, local pedestrian, global environment, and ego-vehicle motion are encoded as input simultaneously. Through the introduced temporal fusion block between Transformer layers and self-attention mechanism, the dynamic interaction relationship between pedestrian, ego-vehicle, and environment is jointly and progressively modeled. Hence, PIT can progressively process temporal information and capture the dynamic interaction relationship to predict the pedestrian intention more like human drivers. Experimental results demonstrate that PIT achieves significantly higher performance compared with other state-of-the-arts.

1 Introduction

With the development of artificial intelligence, autonomous driving has made significant progress in recent years. However, there are still some challenges to achieving high-level autonomous driving in complex urban scenarios. One of the major challenges is that vehicles need to analyze and understand the intentions of other traffic participants just like human drivers do.

In particular, compared with vehicles, pedestrians' intention is more irregular and more difficult to predict. In addition to themselves goal-oriented, pedestrians' crossing intention depends on various environmental factors, such as social interactions, and traffic dynamics [Rasouli and Tsotsos, 2019]. Social interactions abundantly exist between pedestrians and other traffic elements like environment and ego-vehicle. For instance, pedestrians are more likely to choose not to cross for a while when vehicles are moving faster or closer to them [Schmidt and Faerber, 2009], or when the pedestrian crossing light is changing to red. Hence, how to capture the interaction relationship between pedestrians and other traffic elements in dynamic scenes becomes the key point of pedestrian crossing intention prediction algorithms.

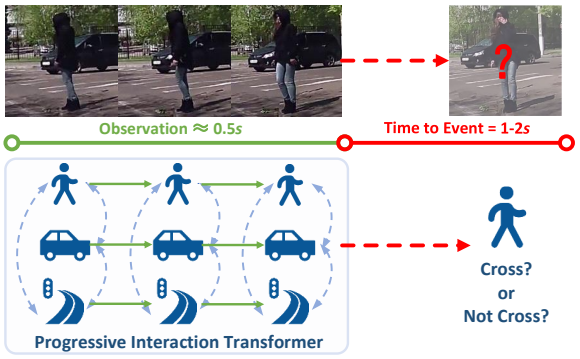

Figure 1: The dynamic interaction relationship between pedestrian, ego-vehicle, and environment is progressively modeled to predict the pedestrian intention. In this task, taking the observation of about $0.5s$ as input to predict pedestrian intention after $1-2s$, which can give autonomous vehicles sufficient time to react to pedestrian behavior. We term the time as Time-to-event(TTE).

Recently, most researches on pedestrian crossing intention prediction focus on multi-branch based models [Yang *et al.*, 2021; Kotseruba *et al.*, 2021]. They often use different branches to process information from different sources independently, and finally concatenate all information to predict. However, it can not model the potential interaction relationship well. Graph convolution-based model [Liu *et al.*, 2020] may be a better solution, but there will be some redundancy in the process of scene graph building. Therefore, the field needs a more efficient and concise framework. Transformer [Dosovitskiy *et al.*, 2020], which models the rela-

*Chao Gou is the corresponding author.

tionship between image patches through the powerful self-attention mechanism is a good choice. Different from standard Transformer, we regard traffic elements as tokens, and use self-attention mechanism to model the potential interaction relationship between them.

On the other hand, there have been some studies trying to propose methods for autonomous driving combined with human cognitive mechanisms [Xia *et al.*, 2020; Plebe *et al.*, 2021; Gou *et al.*, 2022]. The neuroscience theory points out that one of the mechanisms of humans' understanding of natural vision is obtained through stimulation of the brain [Turner *et al.*, 2019]. After inputting stimulus, a message composed of an appropriately timed periodic train of pulse packets will be progressively amplified, and eventually will be strong enough to be propagated to the receiver neuronal network [Chang *et al.*, 2018; Hahn *et al.*, 2019]. **In general, humans understand natural vision through a progressive process of brain stimulation.** Similarly, in driving scenarios, when human drivers detect the pedestrian in front of vehicle (stimulation), drivers will observe the pedestrian for a while to progressive reason his intention and react accordingly.

Inspired by Transformer and neuroscience, we propose Progressive Interaction Transformer(PIT), which can progressively process information and capture the dynamic interaction relationship between traffic elements more like human drivers, as shown in Figure 1. Specifically, according to classical traffic theory, we consider three traffic elements: pedestrian, ego-vehicle, and environment (also known as road). These elements are progressively fed into the transformer layer as tokens, and the dynamic interaction relationship between them is jointly and progressively modeled through the introduced temporal fusion block and self-attention mechanism. In summary, the contributions of our work can be concluded to three-fold:

- We propose a novel framework, Progressive Interaction Transformer(PIT), which can progressively capture the dynamic interaction relationship between pedestrian, ego-vehicle, and environment more like human drivers.
- We introduce a simple but effective Temporal Fusion Block to model long-range temporal dependencies progressively in PIT.
- Experimental results show that the proposed approach outperforms existing methods on pedestrian crossing intention prediction task.

2 Related Work

2.1 Pedestrian Intention Prediction

Human intention can be inferred by focusing on his past and current behavior, including their dynamics, current activity, and context. In autonomous driving, understanding and predicting the intention of other road participants is one of the major challenges. There have been some work to predict road participants' intentions, including the intentions of the drivers [Gebert *et al.*, 2019; Li *et al.*, 2020], other drivers [Li *et al.*, 2016; Mylavarapu *et al.*, 2020], and pedestrians [Rasouli *et al.*, 2019a; Razali *et al.*, 2021]. In this paper, we mainly discuss and study pedestrian intention prediction.

Pedestrian intention prediction can be divided into two categories: based on trajectory [Rehder *et al.*, 2018; Yu *et al.*, 2020], and based on action and context [Liu *et al.*, 2020; Rasouli *et al.*, 2021].The trajectory-based methods focus on observing pedestrians from the top view of traffic scenarios. The action- and context-based methods are to observe pedestrians from the view of ego-vehicles, which are more practical for autonomous vehicle systems. With the release of JAAD dataset [Rasouli *et al.*, 2017], pedestrian crossing intention prediction research has gained more attention. Based on pedestrian image sequences, [Saleh *et al.*, 2019] used 3D Convolution-based architectures to predict pedestrian crossing. Further, [Singh and Suddamalla, 2021] combined pedestrian pose features as input. Graph Convolution-based models [Liu *et al.*, 2020; Chen *et al.*, 2021] were proposed to build the spatio-temporal relationship in driving scene for reasoning pedestrian crossing. In [Chaabane *et al.*, 2020], they predicted future scene representations, and fed them into a classifier to predict whether a crossing event based on generative networks. Recently, [Kotseruba *et al.*, 2021] proposed a benchmark to evaluate the performance of models.

Besides, some multi-branch architecture based methods were proposed to fuse more key information about pedestrian intention. In [Rasouli *et al.*, 2019b], they proposed a stacked GRU network to fuse different modalities, e.g. pedestrian appearance, poses, surrounding context, and bounding boxes. [Yang *et al.*, 2022] fused sequences of images, semantic segmentation masks, and ego-vehicle information using attention mechanisms and a stack of recurrent neural networks. [Lorenzo *et al.*, 2021] proposed a multi-branch architecture based on RubiksNet [Fan *et al.*, 2020] and Transformer to fuse pedestrian appearance, bounding box coordinates, pose keypoints, and ego-vehicle spend information. In this paper, we expect to fuse information from different sources through a simple and single structure, and it can better learn the interaction relationship between them.

2.2 Video Transformer

Transformer [Vaswani *et al.*, 2017], as an attention-based structure, has first shown great advantage in the natural language processing field. Inspired by this, transformer has gradually been migrated to computer vision tasks, and plays an essential role in the field. For instance, ViT [Dosovitskiy *et al.*, 2020] is a pure transformer structure for image classification that converts a single image to tokens. Not only that, transformer is also widely used in other computer vision tasks, such as object detection [Carion *et al.*, 2020; Zhu *et al.*, 2020], semantic segmentation [Zheng *et al.*, 2021; Xie *et al.*, 2021], and image enhancement [Zhang *et al.*, 2021b], etc.

Nonetheless, the above methods focus on learning spatial context without temporal dependencies that are important for video understanding. Most recently, some works [Arnab *et al.*, 2021; Zhang *et al.*, 2021a] tried to obtain temporal dependencies based on transformer and achieved state-of-the-art results on popular video understanding datasets. Most of them extract spatio-temporal tokens from entire input video. However, the memory usage and computational requirements still are big issues, and inputting the entire video also makes them

impossible to directly apply to real-time scenarios, such as autonomous driving. In our work, inspired by neuroscience, we don't process the entire video sequences at once, but progressively establish spatio-temporal dependencies through self-attention mechanism and temporal fusion blocks.

3 Method

In this section, we first formulate the problem, introduce our model inputs based on traffic theory, and briefly review the vision transformer architecture as preliminary. Then, we describe our work in detail, which through **Interaction Transformer** and **Temporal Fusion Block**, the dynamic interaction relationship between pedestrian, ego-vehicle, and environment is jointly and progressively modeled. An overall pipeline of the framework is illustrated in Figure 2. All in all, PIT can progressively process input information and learn the dynamic interaction relationship between traffic elements to predict pedestrian crossing intention.

3.1 Preliminary

Problem Formulation

We formulate pedestrian crossing intention prediction as a binary classification task. As shown in Figure1, given 16 video frames($\approx$ 0.5s) from the front view of ego-vehicle and the corresponding ego-vehicle motion information, the goal is to predict whether the target pedestrian will cross the road. Note that we follow the settings given by [Kotseruba *et al.*, 2021]: the last frame of observation is between 1 and 2s prior to the crossing event start. It is more challenging and practical than previous works that use complete videos to predict pedestrian crossing. Specifically, we term the prediction horizon as Time-to-event (TTE).

Generally, classical traffic theory divides traffic into three key elements: person, vehicle, and environment (also known as road). In this work, we hope to model the dynamic interaction relationship between the three traffic elements. Hence, as follow, three inputs are set in our proposed framework to correspond to the three elements:

- **Pedestrian:** the bounding box coordinates C_i^t and the corresponding image information I_i^t of pedestrian i.

$$p_i^t = \{C_i^t, I_i^t\} \tag{1}$$

$$P_i = \{p_i^{t_0}, p_i^{t_1}, p_i^{t_2}, \cdots, p_i^{t_{15}}\} \tag{2}$$

- **Ego-vehicle:** ego-vehicle speed or motion behavior.

$$V_i = \{v_i^{t_0}, v_i^{t_1}, v_i^{t_2}, \cdots, v_i^{t_{15}}\} \tag{3}$$

- **Environment:** observed video frames from ego-vehicle's front view.

$$E_i = \{e_i^{t_0}, e_i^{t_1}, e_i^{t_2}, \cdots, e_i^{t_{15}}\} \tag{4}$$

Vision Transformer Architecture

In this subsection, we briefly review the architecture of vision transformer [Dosovitskiy *et al.*, 2020]. The standard vision transformer is an encoder structure that stacked multiple transformer layers. Further, a complete transformer layer consists of a Multi-Head Self-Attention sub-layer (MSA) followed by a Mutli-Layer Perceptron sub-layer (MLP). Residual connection and layer normalization (LN) operations are employed around each of the two sub-layers.

Among them, MSA sub-layer consists of h parallel self-attention heads to jointly attend to information from different representation subspaces at different positions. The inputs of each attention head are query $Q_i = QW_i^Q$, key $K_i = KW_i^K$ and value $V_i = VW_i^V$, where W_i^Q, W_i^Q and W_i^V are learnable weight parameters and d is the dimension of Q_i, K_i and V_i. A single-head self-attention head (SA) is computed by:

$$SA_i(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d}})V_i \tag{5}$$

All self-attention heads are concatenated and multiplied by the learnable parameter W^O for the output of MSA sub-layer:

$$MSA(Q, K, V) = [SA_1, SA_2, , SA_h]W^O \tag{6}$$

Last, with residual connection operation, the output of MSA sub-layer fed into the MLP sub-layer for additional processing. Specifically, the MLP sub-layer contains two layers with a GELU non-linearity.

$$y = MLP(MSA(x)) + MSA(x) \tag{7}$$

3.2 The Network Structure

Figure 2(a) shows the overview of the proposed network. First, for frame $t-1$, we convert the traffic elements information into tokens by the corresponding **Pedestrian, Ego-Vehicle Motion, and Environment Encoders**, and feed them to **Interaction Transformer Layer** $t-1$, which can obtain the interaction relationship among tokens using self-attention mechanism. Then, for frame t, using the same encoders to process information at t, and fuse tokens at $t-1$ and t by our **Temporal Fusion Block**, and feed them to **Interaction Transformer Layer** t. Similarly, the process will be repeated several times until the last frame T of observation. Last, we feed the class token at T with dynamic interaction information to a MLP for pedestrian crossing intention prediction.

Pedestrian Encoder

As shown in Figure 2(b), Pedestrian Encoder is divided into two parts: pedestrian feature extractor and coordinate embedding. In pedestrian feature extractor, the pedestrian image I_i^t provided by the original dataset is resized and fed to a convolutional network to obtain a high-level visual feature $X_i^t \in \mathbb{R}^{w \times h \times d}$. Then, we flatten and map the feature to $1 \times D$ dimensions with trainable linear projection, which is termed as pedestrian embedding $E_{P_i^t}$.

The coordinate embedding is designed to inform PIT the global location of pedestrian bounding box C_i^t. Specifically, C_i^t is a 4-d vector, as $(x_{LT}, y_{LT}, x_{RB}, y_{RB})$, where (x_{LT}, y_{LT}) and (x_{RB}, y_{RB}) denote the top-left and bottom-right corner coordinate of bounding box respectively. We also map C_i^t to $1 \times D$ dimensions as coordinate embedding $E_{c_i^t}$. In the end, pedestrian embedding $E_{p_i^t}$ is added to coordinate

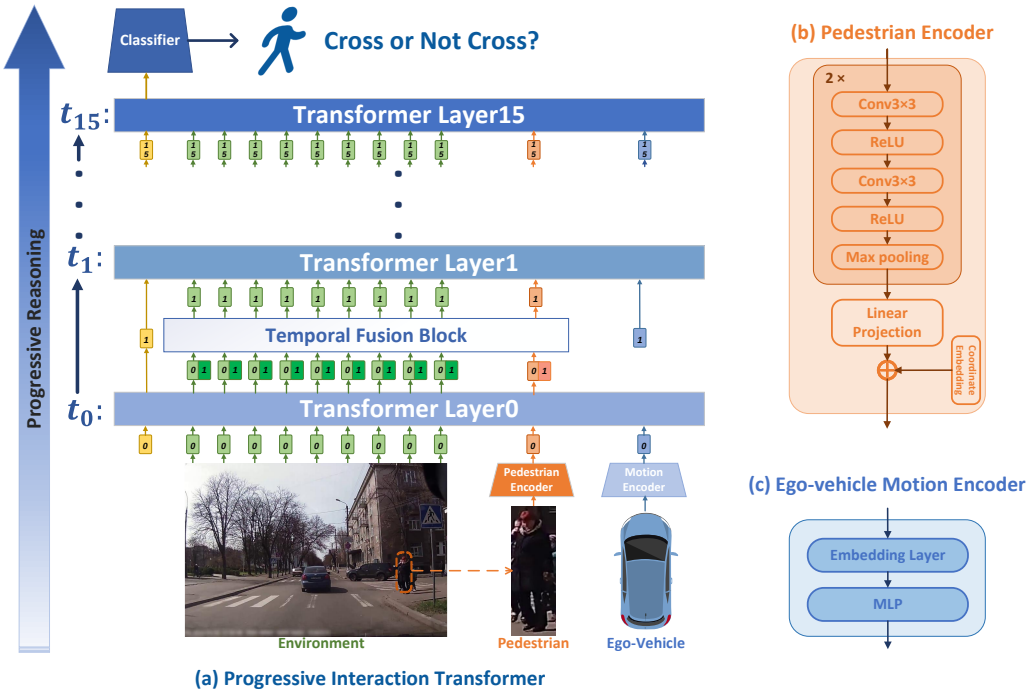

Figure 2: Overview of our proposed framework: local pedestrian, global environment, and ego-vehicle motion of the same frame are encoded as input simultaneously. Each video frame corresponds to a special **Interaction Transformer Layer** to extract the spatial context (Sec.3.2). Between each two video frames, they have a **Temporal Fusion Block** to update tokens to capture temporal dependencies (Sec.3.3). Through progressive input information and stacking of Interaction Transformer Layer and Temporal Fusion Block, the dynamic interaction relationship between pedestrian, ego-vehicle, and environment is jointly and progressively modeled.

embedding $E_{c_i^t}$ to obtain positional information relative to the whole picture.

$$E_{p_i^t} = E_{p_i^t} + E_{c_i^t} \tag{8}$$

Ego-Vehicle Motion Encoder

The ego-vehicle's behavior is critical to the decision-making of pedestrians, hence we introduce ego-vehicle information V_i into our proposed network. In JAAD dataset, five ego-vehicle motion behaviors are labeled per frame: stopping, accelerating, decelerating, moving slow, and moving fast. As shown in Figure 2(c), we encode them and generate the ego-vehicle motion embedding $E_{v_i^t} \in 1 \times D$ to feed it to Transformer layer t.

Environment Encoder

We regard the front view of ego-vehicle as environment in driving scenarios. Similarly to standard vision transformer, we interpret an observed video frame as a sequence of patches. Specifically, we reshape the observation $e_i^t \in \mathbb{R}^{H\times W\times C}$ into sequence of flattened patches and map to $E_{e_i^t} \in \mathbb{R}^{(P^2\cdot C)\times D}$ with a trainable linear projection, where (P, P) is the size of each image patch,and $N = \frac{HW}{P^2}$ is the number of patches.

Interaction Transformer

Transformer can learn token dependencies and encode contextual information from the input through self-attention mechanism. In PIT, we convert pedestrian, ego-vehicle and environment features into tokens.

Specifically, for time t_0, we concatenate class token CLS, environment embedding $E_{e_i^0}$, pedestrian embedding $E_{p_i^0}$, and ego-vehicle embedding $E_{v_i^0}$. Then, we add the position embedding E_{pos} to retain positional information of tokens for Z_0. Then, we feed Z_0 to a standard transformer layer with Multi-Head Attention (MSA) and Multi-Layer Perceptron (MLP) to extract the potential interaction relationship among tokens at t_0 :

$$Z_0 = [CLS; E_{e_i^0}; E_{p_i^0}; E_{v_i^0}] + E_{pos}, \tag{9}$$

$$Z_0^{'} = MSA(LN(Z_0)) + Z_0, \tag{10}$$

$$Z_0^{''} = MLP(LN(Z_0^{'})) + Z_0^{'}. \tag{11}$$

For time t_1-t_{15}, we design a variety of temporal fusion blocks to update tokens representing different traffic elements to obtain temporal dependencies. The specific method design is detailed in Subsection 3.3. For $t-1$, the output of interaction transformer layer $t-1$ is:

$$Z_{t-1}^{''} = [CLS_{t-1}^{''}; E_{e_i^{t-1}}^{''}; E_{p_i^{t-1}}^{''}; E_{v_i^{t-1}}^{''}] \tag{12}$$

Then, for t, we extract environment embedding $E_{e_i^t}$, pedestrian embedding $E_{p_i^t}$, and ego-vehicle embedding $E_{v_i^t}$

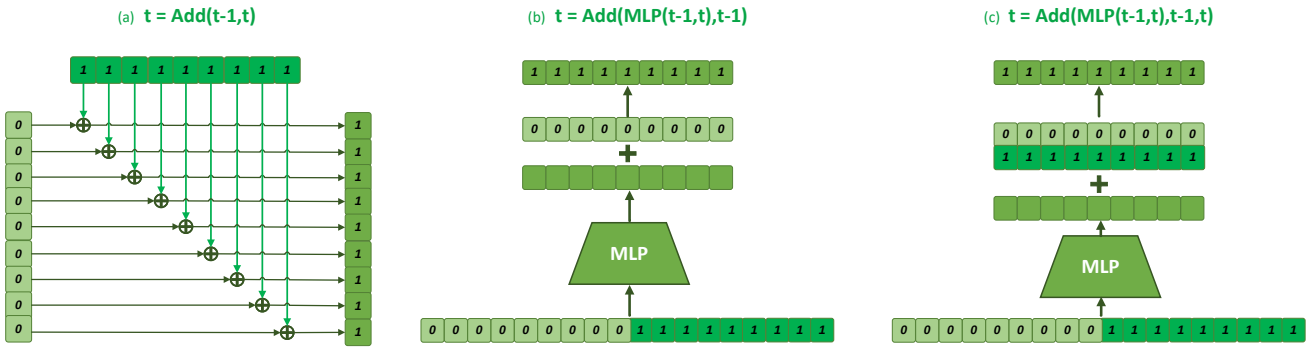

Figure 3: Various types of Temporal Fusion Blocks: **(a)Direct Addition:** directly adds tokens $t-1$ through Interaction Transformer to corresponding tokens at t. **(b)Based on MLP:** Tokens $t-1$ and tokens t are fused by MLP. **(c)Based on Addition and MLP:** Combining from (a) and (b), tokens $t-1$ and tokens t are fused by addition and MLP.

through corresponding encoders. For extracting temporal dependencies, we use them to update the corresponding tokens based on temporal fusion blocks(TFB):

$$E_{e_i^t} = TFB(E^{''}_{e_i^{t-1}}, E_{e_i^t}) \tag{13}$$

$$E_{p_i^t} = TFB(E^{''}_{p_i^{t-1}}, E_{p_i^t}) \tag{14}$$

Note that ego-vehicle embedding don't use TFB, but directly replace $E^{''}_{v_i^{t-1}}$ with $E_{v_i^t}$. For interaction transformer layer t:

$$Z_t = [CLS^{''}_{t-1}; E_{e_i^t}; E_{p_i^t}; E_{v_i^t}], \tag{15}$$

$$Z^{'}_t = MSA(LN(Z_t)) + Z_t, \tag{16}$$

$$Z^{''}_t = MLP(LN(Z^{'}_t)) + Z^{'}_t. \tag{17}$$

Similarly, the process(Eq.13-17) will be repeated several times until the last frame of observation t_{15}. Last, we feed the class token at t_{15} with progressive interaction information to a MLP for pedestrian crossing intention prediction:

$$y = MLP(LN(CLS^{''}_{t_{15}})) \tag{18}$$

3.3 Temporal Fusion Block

In this work, we expect a simple way to update tokens to progressively establish temporal dependencies, just as humans do with the progressive reception of dynamic information. As shown in Figure 3, we design three Temporal Fusion Blocks to fuse the corresponding tokens feature at $t-1$ and t: (a)**Direct Addition**, (b)**Based on MLP**, and (c)**Based on Addition and MLP**. We insert a Temporal Fusion Block between two Interaction Transformer layers.

Note that we only use Temporal Fusion Block to update environment and pedestrian information without ego-vehicle information. This is because there are two different information mechanisms here. Specifically, from the view of human drivers, they *progressively receive* dynamic information of environment and pedestrians, and *react* to it to control ego-vehicle. We believe that ego-vehicle information does not need to be progressively updated in PIT.

Block(a): Direct Addition

Block(a) fuses temporal information by direct addition, which adds neither extra parameter nor computational complexity. It is the simplest method, but already works well (Details are in Section 4). Between two transformer layers, directly add updated corresponding embeddings at t and tokens that have extracted spatial context and interaction relationship through transformer at $t-1$.

$$E_{e_i^t} = E_{e_i^t} + E^{''}_{e_i^{t-1}} \tag{19}$$

$$E_{p_i^t} = E_{p_i^t} + E^{''}_{p_i^{t-1}} \tag{20}$$

Block(b): Based on MLP

In block(b), we concatenate tokens of corresponding traffic elements at $t-1$ and t, and use a MLP to fuse information and compress back to the original dimension $1 \times D$. Then, for making information propagation smooth, we correspondingly add tokens at $t-1$.

$$E_{e_i^t} = MLP(E_{e_i^t}, E^{''}_{e_i^{t-1}}) + E^{''}_{e_i^{t-1}} \tag{21}$$

$$E_{p_i^t} = MLP(E_{p_i^t}, E^{''}_{p_i^{t-1}}) + E^{''}_{p_i^{t-1}} \tag{22}$$

Block(c): Based on Addition and MLP

Block(c) is combined with blocks(a) and (b). In this block, not only direct addition is required, but also need to fuse temporal information through MLP.

$$E_{e_i^t} = MLP(E_{e_i^t}, E^{''}_{e_i^{t-1}}) + E^{''}_{e_i^{t-1}} + E_{e_i^t} \tag{23}$$

$$E_{p_i^t} = MLP(E_{p_i^t}, E^{''}_{p_i^{t-1}}) + E^{''}_{p_i^{t-1}} + E_{p_i^t} \tag{24}$$

4 Evaluation

In the section, we conduct exhaustive comparison experiments on JAAD dataset to evaluate the performance of PIT. Experimental results show that PIT outperforms current state-of-the-art methods. Further, we also design extensive ablation experiments to explore more possibilities of PIT.

Models	Model Variants	ACC	AUC	F1	P	R
VGG16	/	0.59	0.52	0.71	0.63	0.82
ResNet50	/	0.46	0.45	0.54	0.58	0.51
ATGC	AlexNet	0.48	0.41	0.62	0.58	0.66
SingleRNN-LSTM	LSTM	0.51	0.48	0.61	0.63	0.59
SingleRNN-GRU	GRU	0.58	0.54	0.67	0.67	0.68
MultiRNN	GRU	0.61	0.50	0.74	0.64	0.86
StackedRNN	GRU	0.60	0.60	0.66	**0.73**	0.61
SFRNN	GRU	0.51	0.45	0.63	0.61	0.64
C3D	3D CNN	0.61	0.51	0.75	0.63	0.91
I3D-RGB	3D CNN	0.62	0.56	0.73	0.68	0.79
I3D-Optical flow	3D CNN	0.62	0.51	0.75	0.65	0.88
PCPA	3D CNN+RNN+Attention	0.58	0.50	0.71	/	/
TrouSPI-Net	GRU+Attention	0.64	0.56	0.76	0.66	0.91
FFSTP	GRU+Attention	0.62	0.54	0.74	0.65	0.85
IntFormer	Transformer	0.59	0.54	0.69	/	/
PIT-Block(a)	Transformer	<u>0.67</u>	**0.69**	0.77	<u>0.71</u>	0.84
PIT-Block(b)	Transformer	0.64	0.55	<u>0.78</u>	0.65	<u>0.96</u>
PIT-Block(c)	Transformer	**0.69**	<u>0.67</u>	**0.81**	0.69	**0.97**

Table 1: Performance Comparison with State-of-the-Art Methods. **Bold** is best and <u>underline</u> is second best.

4.1 Datasets

We evaluate our framework on JAAD [Rasouli *et al.*, 2017] that is one of the most widely used pedestrian datasets. To create fair experiment settings, we reimplement the benchmark [Kotseruba *et al.*, 2021] as dataset configuration. Joint Attention in Autonomous Driving (JAAD) dataset contains 346 short clips of pedestrians prior to crossing events filmed. In JAAD, each frame includes bounding box information and contextual annotations for the scenes.

4.2 Implementation Details

In our method, we set the dimension of each transformer layer to 1024, the dimension of MLP to 2048, the number of attention head to 16. We train the model with Adam optimizer and crossentropy loss. we set the learning rate starts from $8e^{-5}$ and decays by 0.2 rate every 5 epochs. We set the batch size to 10. Considering the significant imbalance of dataset, we follow the benchmark [Kotseruba *et al.*, 2021] to apply class weights that are inversely proportional to the percentage of samples of each class.

4.3 Comparison with State-of-The-Art methods

We compare ours with 11 state-of-the-art methods, including ATGC [Rasouli *et al.*, 2017], SingleRNN [Kotseruba *et al.*, 2020], MultiRNN [Bhattacharyya *et al.*, 2018], StackedRNN [Bhattacharyya *et al.*, 2018], SFRNN [Yang *et al.*, 2021], C3D [Tran *et al.*, 2015], I3D [Carreira and Zisserman, 2017], PCPA [Kotseruba *et al.*, 2021], TrouSPI-Net [Gesnouin *et al.*, 2021], IntFormer [Lorenzo *et al.*, 2021], and FFSTP [Yang *et al.*, 2022]. The evaluation metrics include accuracy, AUC, F1 score, precision, and recall that are widely used in related work. The results are shown in Table1.

Effect of Our Method

Table 1 shows that PIT is superior to the others. Specifically, PIT using temporal fusion block(a) outperforms all other methods in the three most important evaluating indicators of accuracy, AUC, and F1 score. For PIT using temporal fusion block(c), it achieves the best results in accuracy, AUC, F1 score, and recall. For PIT using temporal fusion block(b), it also outperforms most other methods. From above, PIT is more effective than the state-of-the-art methods on the pedestrian crossing intention prediction task.

Effect of Proposed Temporal Fusion Block

From Table 1, block(c) combined with block(a) and block(b) achieves the best effect. Although block(a) is the simplest one without adding extra parameter, it also achieves a good effect. Compared with the three proposed temporal fusion blocks, direct addition may be more cost-effective in establishing temporal dependencies than using MLP. All in all, we find that our proposed temporal fusion block can effectively fuse temporal information and establish temporal dependencies in Transformer structure.

4.4 Ablation Study

Inspired by neuroscience, PIT can progressively process input information(called stimulation in neuroscience) and understand the dynamic interaction relationship to predict pedestrian intention like human drivers. In this subsection, we conduct extensive ablation experiments to verify the effectiveness of fusing different traffic element information. Meanwhile, we also explore a variety of ways to reduce observation(stimulation) lengths to achieve a more efficient reasoning process.

Are all traffic element information necessary?

In standard PIT, we consider the interaction relationship with pedestrian, ego-vehicle, and environment. But do all traffic elements need to be considered? We conduct two additional ablation experiments as follows: 1) Consider only the

interaction between pedestrian and environment without ego-vehicle. 2) Consider only the interaction between pedestrian and ego-vehicle without environment.

Input Element	ACC	AUC	F1	P	R
P_i, E_i	0.64	0.57	0.74	0.70	0.79
P_i, V_i	0.64	0.60	0.76	0.68	**0.86**
P_i, V_i, E_i	**0.67**	**0.69**	**0.77**	**0.71**	0.84

Table 2: Ablation experiments for inupt element information.

Table 2 shows the comparison of different traffic element inputs on JAAD dataset. The standard PIT considering all traffic elements achieves the best results. It shows that both ego-vehicle and environment contain important cues that influence pedestrians' decision-making. For better performance, it is necessary to consider all traffic element information.

How much effect will reducing sampling rate have?

In standard PIT, we follow [Kotseruba *et al.*, 2021] with using the observation length of about $0.5s$ and the sampling rate of 1. In other words, use 16 consecutive frames as input. A question worth considering is whether to reduce the sampling rate to achieve a more lightweight model. Based on this, we try to adjust the sampling rate to 2 and 4 and conduct ablation experiments. Note that although the number of use frames is reduced, the observation range is still $0.5s$ without reduction.

Samling Rate	Use Frames	ACC	AUC	F1	P	R
4	4	0.62	0.61	0.74	0.68	0.81
2	8	0.63	0.60	0.74	0.68	0.81
1	16	**0.67**	**0.69**	**0.77**	**0.71**	**0.84**

Table 3: Ablation experiments for sampling rate.

Table 3 shows that reducing the sampling rate will reduce accuracy compared to standard PIT, which may be due to weakening the temporal dependencies and compression of reasoning processes. However, compared with Tables 1 and 3, PIT with lower sampling rates can still outperform most state-of-the-art methods with higher sampling rates. It shows that PIT's powerful ability to extract dynamic interaction relationships.

How much effect will reducing observation length have?

Different from previous experiments that reduce sampling rate to compress reasoning processes, we explore the effect of reducing observation length on model performance in this problem. In particular, we reduce the observation length to 12 frames ($t_4 - t_{15}$), 8 frames ($t_8 - t_{15}$), and 4 frames ($t_{12} - t_{15}$). Note that the observation length of standard PIT is 16 frames ($t_0 - t_{15}$).

Table 4 shows the effect of reducing observation length on prediction results. Generally, the prediction results decrease slightly with the decrease of observation length. Compared with Table 3 and 4, it can be found that under the same number of use frames, observation close to the last frame can have better prediction accuracy. It is consistent with intuition that the closer to crossing event, the more visual information clues will be revealed.

Observation	Use Frames	ACC	AUC	F1	P	R
$t_{12} - t_{15}$	4	0.64	0.59	0.74	0.69	0.80
$t_8 - t_{15}$	8	0.65	0.58	0.75	0.70	0.81
$t_4 - t_{15}$	12	0.65	0.60	**0.78**	0.67	**0.93**
$t_0 - t_{15}$	16	**0.67**	**0.69**	0.77	**0.71**	0.84

Table 4: Ablation experiments for observation length.

5 Conclusion

In this paper, we propose Progressive Interaction Transformer(PIT) for pedestrian crossing intention prediction. Our method relies on the introduced temporal fusion block and self-attention mechanism to progressively model the dynamic interaction relationship between pedestrian, ego-vehicle, and environment. Based on this, our method can progressively reason pedestrian intention more like human drivers. Experimental results show that the proposed approach outperforms existing methods on pedestrian crossing intention task. In addition, we further explore the effect of reducing input elements, reducing sampling rate, and reducing observation length on model performance through ablation studies. All in all, PIT is a promising model to be applied to other computer vision and robotics tasks, such as human-object interaction detection, video understanding, etc.

Acknowledgments

This work is supported in part by Shenzhen Science and Technology Program under Grant RCBS20200714114920272, and the Key Research and Development Program of Guangzhou under Grant 202007050002.

References

[Arnab *et al.*, 2021] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, October 2021.

[Bhattacharyya *et al.*, 2018] 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*, pages 4194–4202, 2018.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[Carreira and Zisserman, 2017] 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.

[Chaabane *et al.*, 2020] Mohamed Chaabane, Ameni Trabelsi, Nathaniel Blanchard, and Ross Beveridge. Looking ahead: Anticipating pedestrians crossing with future frames prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2297–2306, 2020.

[Chang *et al.*, 2018] Wei-Chih Chang, Jan Kudlacek, Jaroslav Hlinka, Jan Chvojka, Michal Hadrava, Vojtech Kumpost, Andrew D Powell, Radek Janca, Matias I Maturana, Philippa J Karoly, et al. Loss of neuronal network resilience precedes seizures and determines the ictogenic nature of interictal synaptic perturbations. *Nature neuroscience*, 21(12):1742–1752, 2018.

[Chen *et al.*, 2021] Tina Chen, Renran Tian, and Zhengming Ding. Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3103–3109, 2021.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fan *et al.*, 2020] Linxi Fan, Shyamal Buch, Guanzhi Wang, Ryan Cao, Yuke Zhu, Juan Carlos Niebles, and Li Fei-Fei. Rubiksnet: Learnable 3d-shift for efficient video action recognition. In *European Conference on Computer Vision*, pages 505–521. Springer, 2020.

[Gebert *et al.*, 2019] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 969–974. IEEE, 2019.

[Gesnouin *et al.*, 2021] Joseph Gesnouin, Steve Pechberti, Bogdan Stanciulescu, and Fabien Moutarde. Trouspi-net: Spatio-temporal attention on parallel atrous convolutions and u-grus for skeletal pedestrian crossing prediction. *arXiv preprint arXiv:2109.00953*, 2021.

[Gou *et al.*, 2022] Chao Gou, Yuchen Zhou, and Dan Li. Driver attention prediction based on convolution and transformers. *The Journal of Supercomputing*, pages 1–17, 2022.

[Hahn *et al.*, 2019] Gerald Hahn, Adrian Ponce-Alvarez, Gustavo Deco, Ad Aertsen, and Arvind Kumar. Portraits of communication in neuronal networks. *Nature Reviews Neuroscience*, 20(2):117–127, 2019.

[Kotseruba *et al.*, 2020] 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. IEEE, 2020.

[Kotseruba *et al.*, 2021] 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.

[Li *et al.*, 2016] Bo Li, Tianfu Wu, Caiming Xiong, and Song-Chun Zhu. Recognizing car fluents from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3803–3812, 2016.

[Li *et al.*, 2020] Chengxi Li, Yue Meng, Stanley H Chan, and Yi-Ting Chen. Learning 3d-aware egocentric spatial-temporal interaction via graph convolutional networks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8418–8424. IEEE, 2020.

[Liu *et al.*, 2020] 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*, 5(2):3485–3492, 2020.

[Lorenzo *et al.*, 2021] J Lorenzo, I Parra, and MA Sotelo. Intformer: Predicting pedestrian intention with the aid of the transformer architecture. *arXiv preprint arXiv:2105.08647*, 2021.

[Mylavarapu *et al.*, 2020] Sravan Mylavarapu, Mahtab Sandhu, Priyesh Vijayan, K Madhava Krishna, Balaraman Ravindran, and Anoop Namboodiri. Understanding dynamic scenes using graph convolution networks. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8279–8286. IEEE, 2020.

[Plebe *et al.*, 2021] Alice Plebe, Julian FP Kooij, Gastone Pietro Rosati Papini, and Mauro Da Lio. Occupancy grid mapping with cognitive plausibility for autonomous driving applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2934–2941, 2021.

[Rasouli and Tsotsos, 2019] Amir Rasouli and John K Tsotsos. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE transactions on intelligent transportation systems*, 21(3):900–918, 2019.

[Rasouli *et al.*, 2017] 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, 2017.

[Rasouli *et al.*, 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*, pages 6262–6271, 2019.

[Rasouli *et al.*, 2019b] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. In *BMVC*, 2019.

[Rasouli *et al.*, 2021] 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, 2021.

[Razali *et al.*, 2021] Haziq Razali, Taylor Mordan, and Alexandre Alahi. Pedestrian intention prediction: A convolutional bottom-up multi-task approach. *Transportation research part C: emerging technologies*, 130:103259, 2021.

[Rehder *et al.*, 2018] Eike Rehder, Florian Wirth, Martin Lauer, and Christoph Stiller. Pedestrian prediction by planning using deep neural networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5903–5908. IEEE, 2018.

[Saleh *et al.*, 2019] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9704–9710. IEEE, 2019.

[Schmidt and Faerber, 2009] Sarah Schmidt and Berthold Faerber. Pedestrians at the kerb–recognising the action intentions of humans. *Transportation research part F: traffic psychology and behaviour*, 12(4):300–310, 2009.

[Singh and Suddamalla, 2021] 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.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[Turner *et al.*, 2019] Maxwell H Turner, Luis Gonzalo Sanchez Giraldo, Odelia Schwartz, and Fred Rieke. Stimulus-and goal-oriented frameworks for understanding natural vision. *Nature neuroscience*, 22(1):15–24, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Xia *et al.*, 2020] Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, Teresa Canas-Bajo, and David Whitney. Periphery-fovea multi-resolution driving model guided by human attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1767–1775, 2020.

[Xie *et al.*, 2021] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

[Yang *et al.*, 2021] Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith Redmill, and Ümit Özgüner. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *arXiv preprint arXiv:2104.05485*, 2021.

[Yang *et al.*, 2022] 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.

[Yu *et al.*, 2020] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European Conference on Computer Vision*, pages 507–523. Springer, 2020.

[Zhang *et al.*, 2021a] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021.

[Zhang *et al.*, 2021b] Zhaoyang Zhang, Yitong Jiang, Jun Jiang, Xiaogang Wang, Ping Luo, and Jinwei Gu. Star: A structure-aware lightweight transformer for real-time image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4106–4115, 2021.

[Zheng *et al.*, 2021] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.