Temporal-contextual Event Learning for Pedestrian Crossing Intent Prediction

Hongbin Liang[1,2] *, Hezhe Qiao[3] *, Wei Huang[4], Qizhou Wang[5], Mingsheng Shang[1], and Lin Chen[1] (✉)

[1] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China
[2] Chongqing School, University of Chinese Academy of Sciences, Chongqing 400714, China
{lianghongbin,msshang,chenlin}@cigit.ac.cn
[3] Singapore Management University, 178902, Singapore
hezheqiao.2022@phdcs.smu.edu.sg
[4] Beijing University of Posts and Telecommunications, 100876, China
huangwei@bupt.edu.cn
[5] Monash University, 3800, Australia
qizhou.wang@monash.edu

Abstract. Ensuring the safety of vulnerable road users through accurate prediction of pedestrian crossing intention (PCI) plays a crucial role in the context of autonomous and assisted driving. Analyzing the set of observation video frames in ego-view has been widely used in most PCI prediction methods to forecast the cross intent. However, they struggle to capture the critical events related to pedestrian behaviour along the temporal dimension due to the high redundancy of the video frames, which results in the sub-optimal performance of PCI prediction. Our research addresses the challenge by introducing a novel approach called Temporal-contextual Event Learning (TCL). The TCL is composed of the Temporal Merging Module (TMM), which aims to manage the redundancy by clustering the observed video frames into multiple key temporal events. Then, the Contextual Attention Block (CAB) is employed to adaptively aggregate multiple event features along with visual and non-visual data. By synthesizing the temporal feature extraction and contextual attention on the key information across the critical events, TCL can learn expressive representation for the PCI prediction. Extensive experiments are carried out on three widely adopted datasets, including PIE, JAAD-beh, and JAAD-all. The results show that TCL substantially surpasses the state-of-the-art methods. Our code can be accessed at https://github.com/dadaguailhb/TCL.

Keywords: Crossing Intent Prediction· Temporal Event Learning · Contextual Attention Mechanism.

* Equal contribution

1 Introduction

In the rapidly advancing field of autonomous and assisted driving [11,14], the paramount concern is human safety. Facilitating effective interactions between vehicles and vulnerable road users is essential. Within this context, accurately identifying pedestrian crossing intention (PCI) is especially vital. In reality, pedestrian behavior is affected by various factors [26,7], such as traffic signs, vehicle speeds, and the conduct of other traffic participants, making the accurate prediction of pedestrian behavior a challenging task. Over the past few years, a variety of RNN-based methods for PCI have been introduced, which analyze both visual and non-visual input data, achieving improved performance in PCI prediction [12,13,28,7]. Given that RNNs have a limited capacity to retain global information, often leading to information loss when processing temporal data, some works utilize transformer-based architectures to more effectively capture long-range dependencies and enhance overall performance [14,32,30]. Although RNN-based and transformer-based PCI prediction methods have achieved a certain level of success, they typically extract features from all the observed video frames to forecast a pedestrian's intention to cross the road. [3,18]. This approach leads to high redundancy since the adjacent frames often exhibit significant similarity. The high redundancy of temporal sequence data makes the model struggle to capture essential information, resulting in sub-optimal performance.

To deal with the challenges above, we propose a new network, called temporal-contextual event learning (TCL), illustrated in Fig. 1, which differs from the prior approaches like RNNs and transformer-based approaches that analyze the whole observed video frames in PCI prediction. We first propose the Temporal Merging Module (TMM), which employs event clustering to categorize the observed video frames into multiple key events according to the behavior changes of pedestrians, such as standing, walking, or turning around. Furthermore, we introduce a contextual attention block (CAB) to explore the relation of multiple critical events, aggregating the critical features at both the event level and the data level. Finally, by leveraging the TMM and CAB, the critical features can be effectively captured in the expressive representation learning of pedestrians, enhancing the accuracy of PCI prediction. Our primary contributions in this work are:

- We first introduce the temporal merging module to effectively identify the critical information by clustering the temporal frames into multiple critical events.
- We then develop a contextual attention block to adaptively aggregate the critical features from the key events along with visual and non-visual data.
- Conducting thorough studies on three prominent datasets, we illustrate that the proposed TCL significantly surpasses existing state-of-the-art approaches.

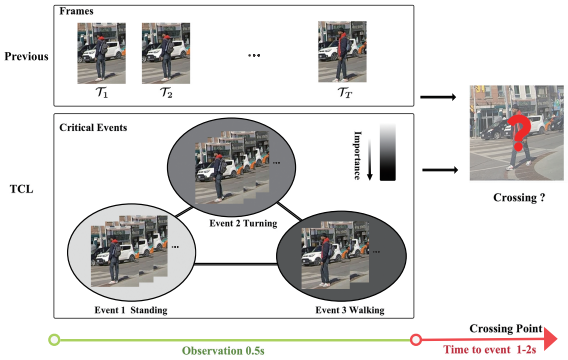

Fig. 1: TCL focuses on the critical events identification and adaptively critical features extraction across the events for PCI prediction, contrasting the previous methods that analyze the whole observed video frames.

2 Related Work

2.1 RNN-based Cross Intent Detection

Due to the strong capability of modeling temporal dependencies and relationships between elements, RNN has been widely used for analyzing the video frame sequence [20,16]. Bhattacharyya et al. [1] introduce RNNs to sequence 2D features, thus integrating temporal data into the analysis. Kotseruba et al. [13] further employ the attention mechanism [17] to refine these models, enabling a focused analysis of key temporal and spatial details. Over time, the complexity and variety of input features used in RNNs have significantly evolved to enhance the performance of PCI prediction [13,19,28,29]. More non-visual features have been incorporated into the PCI prediction model, such as pedestrian bounding boxes with pose keypoints [13,6], traffic objects [29,12], and contextual segments [28]. Ham et al. [8] propose CIPF, which utilizes eight different input modalities. Yang et al. [27] utilize graph convolutional neural networks (GCN) to analyze pedestrian poses and deploy RNN to examine temporal sequences. Additionally, advancements in feature extraction technologies, such as C3D [13], have enabled direct joint analysis of spatial and temporal aspects using video data.

2.2 Transformer-based Cross Intent Detection

RNNs or C3D, while effective in certain tasks, have a limited ability to retain global information[31], typically remembering only the content from recent sequences, which leads to information loss when processing temporal data. In contrast, vision transformer (ViT) architecture networks are capable of building long-range dependencies between images, offering a clear overview of the global context for video recognition. Recently, some ViT-based methods have been introduced to improve the PCI prediction and obtained convincing results [14,15].

Zhang et al. [30] introduce a transformer-based evidential prediction method aimed at uncertainty-aware estimation of pedestrian intentions. Zhou et al. [32] employ stacked transformer layers where each layer corresponds to a time step. Nonetheless, although transformer-based models provide a robust overview of the global context for video recognition, the high similarity between video frames can lead to dispersed attention when the attention mechanism is applied to all frames.

To address this limitation, we built upon video-ViT, using TMM to segment the frame sequence into key events. Then, we applied the contextual attention mechanism to aggregate the critical features adaptively. The proposed TCL can significantly focus on the important information reflected in the data as noticeable dynamic changes.

3 Methodology

3.1 Problem Definition

Considering a series of observed video frames$\{\mathcal{T}_1, \mathcal{T}_2, ..\mathcal{T}_T\}$ where n is the number of observed frames. The goal of cross-intent detection is to design a model $f : \{\mathcal{T}_1, \mathcal{T}_2, ..\mathcal{T}_T\} \rightarrow \mathbb{R}$ to predict the probability of the pedestrian's action.

Due to the significant impact of dynamic environments on pedestrian motion [10], numerous explicit features are widely considered in research. These features can be broadly categorized into visual data and non-visual data.

1) **Visual data.** The visual data, denoted as $\{I_1, I_2, \ldots, I_T\}$, mainly consists of the i-th pedestrian's local context image, which refers to a square zone delineated by an expanded pedestrian bounding box ("bbox"), including both the pedestrian and contextual elements such as ground, curb, crosswalks, etc.

2) **Non-visual data.** The Non-visual data, denoted as $\{I'_1, I'_2, \ldots, I'_T\}$, include the bounding box, pose keypoints of i-th pedestrian, and the traffic objects in the scene which are described as the following.

Bounding Boxes: Represented by $[x_1, y_1, x_2, y_2]$, the position coordinates of a target pedestrian indicate that (x_1, y_1) are the top-left and (x_2, y_2) are the bottom-right coordinates of the pedestrian's bounding box.

Pose keypoints: The movements of target pedestrians by extracting pedestrian pose keypoints. Following the previous studies [6,7,8], we apply a pre-trained OpenPose model [2] to obtain body keypoints denoted as $P = \{p_i^1, p_i^2, \ldots, p_i^T\}$ for pedestrian. Each keypoint P consists of a 36-dimensional vector, which includes the 2D coordinates for 18 different pose joints.

Traffic objects: The traffic object refers to the features which can capture specific aspects of the driving environment, including the Traffic Neighbor Feature (f_{tn}), the Traffic Light Feature (f_{tl}), the Traffic Sign Feature (f_{ts}), the Crosswalk Feature (f_c), the Station Feature (f_s) and the Ego Motion Feature (f_e).

3.2 Network Structure

Fig. 2 illustrates the overall architecture of the proposed model, which comprises three essential components: the feature encoder, the temporal merging module,

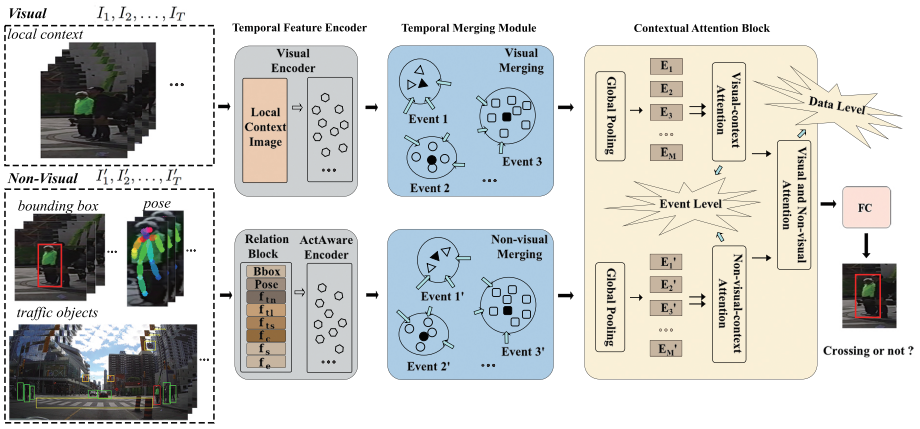

Fig. 2: Overview of TCL. It consists of the Temporal Feature Encoder for visual feature extraction, the Temporal Merging Module for frame clustering, and the Contextual Attention Block for the temporal event feature fusion at both event level and data level.

and the contextual attention block. The feature encoder consists of the temporal feature encoder implemented by the Visual Encoder and the ActAware Encoder, which aim to extract the visual and non-visual features, respectively. The temporal merging module clusters the frame into multiple events to exploit the relation of frames, making the model focus on the key dynamic changes in critical events. The contextual attention block employs the attention mechanism to fuse the contextual information and improve the temporal feature extraction at both the event level and the data level.

3.3 Temporal Feature Encoder

The temporal feature encoder includes both the Visual Encoder and Non-visual Encoder, namely the Actaware Encoder, which is combined with the Relation Block. These encoders help the model learn the representation from the visual and non-visual data, respectively. Then, the visual and non-visual features are fed into the temporal merging modules.

Visual Encoder. A pre-trained video-ViT network was used as the Visual Encoder to extract the visual feature [24]. Following previous studies, [4,9,24], the local context images are resized into a resolution 224×224. The ViT-B/16-based model, pre-trained on Kinetics-400, is chosen as the backbone due to its high efficiency and accuracy.

Non-Visual Encoder. The Non-Visual data, including the bounding box, pose key points, and traffic objects, are the input of the Non-Visual Encoder consisting of the Relation Block and ActAware Encoder.

1) Relation Block. With the Relation Block, the bounding box and each traffic object feature were initially scaled to a uniform size using a corresponding Fully

Connected (FC) layer. Subsequently, these features were concatenated together with pose keypoints, resulting in the new feature $f_r \in \mathbb{R}^{T\times(d_1\times(n+1)+d_2)}$, where T is the observed length, d_1 is the embedding dimension of each traffic object, n represents the number of categories for traffic object features and d_2 represents the dimension of pose keypoints.

2) ActAware Encoder. The ActAware Encoder is implemented based on transformer architecture, which was originally introduced for the tasks in natural language processing [25]. It receives the relation block's output as input and utilizes a transformer-based network with a specification similar to ViT-small [24] to learn the representation of non-visual tokens relevant to pedestrian intent. Finally, it yields the new feature vector of size $f_a \in \mathbb{R}^{T\times d}$, where T represents the observed length and d represents the feature dimension at each time step.

3.4 Temporal Merging Module

After feature extraction by the Temporal Feature Encoder, we obtained visual and non-visual features with a length of T. Subsequently, these T features will be clustered utilizing the TMM.

Although video-ViT is a powerful global attention mechanism for the long series of action predictions, it struggles to prioritize the critical information across the observed frames. The high redundancy significantly hinders the PCI. Therefore, we propose a temporal merging module (TMM) to deal with this concern and make the model focus on the key events. It first categorizes the observed frame into different events by employing a density peaks clustering algorithm [5] based on k-nearest neighbors. Clustering is performed on both visual and non-visual data. Here we take the clustering on visual features as an example, starting with the T time-step features $\mathcal{I} = \{I_t\}_{t=1}^{T}$ derived from the Visual Encoder, we initially compute the local density ρ_t for each I_t based on its K-nearest neighbors, which is formulated as:

$$\rho_t = \exp\big(-\frac{1}{K}\sum_{I_k\in \mathrm{KNN}(I_t,\mathcal{I})}\|I_k - I_t\|^2\big), \tag{1}$$

where $\mathrm{KNN}(I_t, \mathcal{I})$ represents the K-nearest neighbors(excluding itself) of I_t in $\mathcal{I}$. We also need to calculate δ_t of I_t representing the shortest distance from point I_t to any other point that has higher density, which are defined as follows:

$$\delta_t = \begin{cases} \min\limits_{m:\rho_m>\rho_t}\|I_m - I_t\|^2, & \text{if}\exists m \text{ s.t. } \rho_m > \rho_t. \\ \max\limits_{m}\|I_m - I_t\|^2, & \text{otherwise.} \end{cases} \tag{2}$$

The center of each cluster is selected from points with relatively high ρ_t and high δ_t. Then we allocate other points to their closest cluster center and obtain the clusters $\mathcal{E} = \{E_1, E_2, \ldots, E_M\}$, representing the multiple events, where M denotes the number of events. Ultimately, the non-visual events $\mathcal{E}' = \{E'_1, E'_2, \ldots, E'_M\}$ can be obtained using the same principles.

3.5 Contextual Attention Block (CAB)

After obtaining the visual and non-visual event features, we need to fuse the features between events and integrate the features from the visual and non-visual branches using the CAB.

In order to make the model focus on the critical feature after temporal feature extraction, we further employ the CAB to adaptively aggregate across event features, which selectively emphasizes specific aspects of features. Contextual attention is applied on both the event level and the data level. Here, we still take the visual feature $\mathcal{E} = \{E_1, E_2, \ldots, E_M\}$ as an example, the attention weights for each event is defined as follows:

$$\alpha_i = \frac{\exp(score(E_i, E_j))}{\sum_{j=1}^{M} \exp(score(E_j, E_i))}, \tag{3}$$

where $score(E_i, E_j) = {E_i}^{\top} \mathbf{W}_c^1 E_j$ represents similarity between evener E_i and E_j, $\mathbf{W}_c^1$ is a trainable weight matrix. Combined with the weighted sum of features from all preceding time steps, denoted as E_p. It is noted that the same principles are applied to non-visual features. The aggregated features of all events E_p are the new features of the events.

$$F = tanh(\mathbf{W}_p^1[E_p; E_M]), E_p = \sum_m \alpha_m E_m, \tag{4}$$

where $\mathbf{W}_p^1$ is the trainable parameters mapping the representation into a new space.

Apart from utilizing the attention mechanism at the data level by adaptively fusing the features from the visual and non-visual branches. Then, we obtain the overall visual and non-visual event features, denoted as F and F', respectively. Correspondingly, the attention weight of both features is calculated as follows:

$$\lambda_1 = \frac{\exp(score(F', F))}{\exp(score(F', F)) + \exp(score(F', F'))}, \tag{5}$$

$$\lambda_2 = \frac{\exp(score(F', F'))}{\exp(score(F', F)) + \exp(score(F', F'))}, \tag{6}$$

where $score(F', F) = {F'}^{\top} \mathbf{W}_c^2 F$ with the trainable matrix $\mathbf{W}_c^2$. The ultimate integration of visual and non-visual features F_p is computed as:

$$F_p = \lambda_1 F + \lambda_2 F', \tag{7}$$

where $\mathbf{W}_p^2$ is a trainable matrix. The aggregated features are subsequently processed by a fully connected (FC) layer to predict the intention to cross.

$$P = Sigmoid(\mathbf{W}_p^2[F_p; F']), \tag{8}$$

where $P = \{p_1, p_2, ..., p_N\}$ is the output of the model representing predicted probabilities of crossing.

Table 1: Comparison of properties between PIE, and JAAD Datasets

	PIE [20]	JAAD [21]
Number of annotated frames	293K	75K
Number of pedestrians	1.8K	2.8K
Number of pedestrians with behavior annotation	1.8K	686
Number of pedestrians bboxes	740K	391K
Average pedestrian track length	401	140
Ego-vehicle sensor information	yes	no

3.6 Loss Function

we employ binary cross-entropy loss in our PCI prediction task to quantify the discrepancy between the actual labels and the predicted probabilities.

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)], \tag{9}$$

where N denotes the sample size, y_i represents the ground truth label of the i-th sample, and p_i represents the predicted likelihood that the i-th sample belongs to the positive class.

4 Experiments

In this part, the proposed TCL method for predicting pedestrian crossing intent is evaluated for its effectiveness by conducting experiments that compare it against leading state-of-the-art methods.

4.1 Datasets and Experimental settings

Datasets. Following the previous studies [28,29], evaluation of the proposed model and competing methods is conducted on the two large public naturalistic traffic video datasets, Pedestrian Intent Estimation (PIE) [20] and Joint Attention in Autonomous Driving (JAAD) [21]. The PIE dataset, comprising six hours of driving footage captured by an onboard camera, includes 1,842 pedestrians annotated with 2-D bounding boxes and behavioural tags at 30Hz. Ego-vehicle velocity was obtained using gyroscope measurements collected by the camera. The JAAD dataset comprises two subsets: JAAD Behavioral Data (JAAD-beh) and JAAD All Data (JAAD-all). JAAD-beh includes 495 samples of pedestrians crossing and 191 samples of pedestrians intending to cross., while JAAD-all encompasses an extra 2,100 samples of pedestrians performing non-crossing actions. Please refer to Tab. 1 for further details on dataset statistics.
Competing Methods. To verify the performance of TCL, we compared it against state-of-the-art PCI prediction models. MultiRNN [1] utilizes distinct

RNN streams to independently process each feature type. SingleRNN [12] identifies key visual factors, including crosswalk location and pedestrian orientation, and integrates them into a single RNN model. SF-GRU [23] utilizes stacked GRUs to receive the hidden states from the GRUs in the lower layers as their input. PCPA [13] utilizes a C3D network to capture local context images and employs GRUs to analyze critical elements like poses. CAPformer [14] employs two distinct transformer-based encoders to extract features from video sequences and other modalities, respectively. The Coupling Intent and Action (Coupling)[29] model utilizes stacked GRUs to handle temporal sequences, incorporating both current actions and predictive outcomes to enhance pedestrian crossing prediction. The Predicting model [28] extracts features from pedestrian context images and utilizes Hybrid fusion to integrate features from the two branches effectively. MCIP [7] categorizes five inputs, including a segmentation map, into visual and non-visual modules, utilizing an attention mechanism to discern crossing intentions. PIT [32] considered local pedestrian, the global environment, and the movement of ego-vehicle simultaneously. Hybrid-Group [3] introduces a novel hybrid fusion method and incorporates two additional dynamic attributes. TREP [30] proposed an innovative transformer-based model to capture temporal correlations and address uncertainty. PFRN [18] develops a novel RNN architecture that combines spatial and temporal feature fusion.

Evaluation Metric. PCI is essentially a binary classification problem, determining whether a pedestrian will cross or not based on the observed frames. We employ several metrics for evaluation, including accuracy, Area Under the Curve (AUC), F1 score, precision, and recall, which are widely recognized and commonly employed in related research [22,28,13,29].

Implementation Detail. TCL is implemented in Pytorch 1.11.1 with Python 3.8. In TCL we utilized a resized 224×224 resolution as the input image size for the video-ViT backbone, which was initially pre-trained on Kinetics-400 [24]. In the Relation Block, the embedding dimension of each traffic object is 32. Besides, we employ the ActAware Encoder, a transformer-based network, to process non-visual information. This network is designed with an embedding dimension of 384, 6 attention heads, and 12 layers in depth.

Following the previous studies, for each target pedestrian, we sample observation data, ensuring that the final observed frame is captured within a 1 to 2-second interval (or 30 to 60 frames) before the commencement of the crossing event, as specified in the dataset's annotations. For all models, the number of observation frames is 16. Sample overlap ratios are determined to be 0.6 for the PIE and JAAD datasets. Considering the training samples, which span approximately 0.5 seconds, the number of clusters is set as $M = 3$ and used KNN for the event merging process. All models were trained with the RMSProp optimizer with a learning rate of 10^{-5}, and L2 regularization with $\lambda = 1 \times 10^{-3}$.

The TCL model, comprising approximately 100 million parameters, was trained on the JAAD-all dataset using a cluster of eight NVIDIA 1080 Ti GPUs, setting the batch size to 2. Each training epoch took approximately 12 minutes,

Table 2: The experiment results on the JAAD-beh dataset. Each metric's top score is emphasized in boldface, while the runner-up results are underlined.

Methods	Encoder	AUC	ACC	F1	Prec	Rec
MultiRNN(2018) [1]	VGG + LSTM	0.50	–	0.74	–	–
SingleRNN(2020) [12]	VGG + GRU	0.52	0.59	0.71	0.64	0.80
PCPA(2021) [13]	C3D + GRU	0.50	0.58	0.71	–	–
CAPformer(2021) [14]	Transformer	<u>0.55</u>	–	0.74	–	–
Predicting(2022) [28]	VGG + GRU	0.54	0.62	0.74	<u>0.65</u>	<u>0.85</u>
MCIP(2022) [7]	VGG + GRU	<u>0.55</u>	0.64	0.78	–	–
Hybrid-Group(2024) [3]	VGG + GRU	**0.61**	<u>0.67</u>	<u>0.79</u>	–	–
TCL (Ours)	Transformer	**0.61**	**0.74**	**0.85**	**0.75**	**0.98**

while inference on the JAAD-all validation set required around 2 minutes per epoch.

4.2 Comparison with the state-of-the-art methods

The experimental results for JAAD-beh, JAAD-all, and PIE are shown in Tab. 2, Tab. 3, and Tab. 4, respectively. We evaluated TCL's performance by contrasting it with some other benchmark models for PCI, where different models are utilized as their encoder.

As shown in Tab. 2, TCL demonstrates exceptional performance on the JAAD-beh dataset, achieving a very high recall value of 0.98 and the best overall accuracy (ACC) of 0.74 when compared with other baseline models. It is worth highlighting that higher recall values are crucial for ensuring safety in autonomous driving. Given that JAAD-beh is particularly focused on pedestrians with the intent of crossing the road, where they exhibit more significant behavioral changes, with the TMM, the division of pedestrian postures and contextual information into distinct events becomes more manageable, enabling TCL to capture these variations more effectively. In this way, TCL can focus on crucial information among a myriad of redundant data, enabling accurate predictions of pedestrian intentions. Similar results can be found in Tab. 3. On the JAAD-all dataset, TCL achieves top-tier performance with an AUC of 0.92 and competitive accuracy. Moreover, it attains the highest F1 score of 0.71, reflecting a balanced trade-off between Precision and Recall—both at 0.68 and 0.74, respectively. This is primarily due to TCL's increased focus on critical events, underscoring the essential role of TMM in TCL.

As shown in Tab. 4, TCL outperforms the most competing models on the PIE dataset, yielding the highest AUC (0.88), F1 score (0.92), and Recall (0.96). The main reason is that, by merging key events, TCL can capture critical information without causing attention dispersion. These results indicate the effectiveness of TCL in identifying PCI accurately while minimizing false negatives, which is crucial for safety in autonomous driving applications.

Table 3: The experiment results on the JAAD-all dataset. Each metric's top score is emphasized in boldface, while the runner-up results are underlined.

Methods	Encoder	AUC	ACC	F1	Prec	Rec
MultiRNN(2018) [1]	VGG + LSTM	0.79	–	0.58	–	–
SingleRNN(2020) [12]	VGG + GRU	0.76	0.79	0.54	0.44	0.71
PCPA(2021) [13]	C3D + GRU	0.86	0.85	0.68	–	–
Coupling(2021) [29]	VGG + GRU	**0.92**	0.87	0.70	0.66	0.74
CAPformer(2021) [14]	Transformer	0.70	–	0.51	–	–
Predicting(2022) [28]	VGG + GRU	0.82	0.83	0.63	0.51	**0.81**
MCIP(2022) [7]	VGG + GRU	0.84	**0.88**	0.66	–	–
PIT(2023) [32]	Transformer	0.89	0.87	0.67	0.58	0.80
TREP(2023) [30]	Transformer	0.86	**0.88**	0.61	**0.70**	0.54
TCL (Ours)	Transformer	**0.92**	0.87	**0.71**	0.68	0.74

Table 4: The experiment results on the PIE dataset. Each metric's top score is emphasized in boldface, while the runner-up results are underlined.

Methods	Encoder	AUC	ACC	F1	Prec	Rec
MultiRNN(2018) [1]	VGG + LSTM	0.80	0.83	0.71	0.69	0.73
SingleRNN(2020) [12]	VGG + GRU	0.64	0.76	0.45	0.63	0.36
SF-GRU(2020) [23]	VGG + GRU	0.83	0.84	0.72	0.66	0.80
PCPA(2021) [13]	C3D + GRU	0.86	0.87	0.77	–	–
Coupling(2021) [29]	VGG + GRU	**0.88**	0.84	0.90	**0.96**	0.84
CAPformer(2021) [14]	Transformer	0.85	–	0.78	–	–
MCIP(2022) [7]	VGG + GRU	0.87	0.89	0.81	–	–
PFRN(2024) [18]	VGG + GRU	0.85	**0.90**	0.77	0.81	0.74
TCL(Ours)	Transformer	**0.88**	0.87	**0.92**	0.89	**0.96**

4.3 Ablation Study

In this section, we analyze the efficiency of the essential components, including non-visual input features, TMM, and CAB within TCL on the PIE dataset. Specifically, we sequentially disabled each module, while maintaining the other two modules active, resulting in three reduced TCL configurations (*i.e.*without non-visual input features, without TMM, and without CAB). This allows the isolation and evaluation of each module's individual contribution to the overall system performance. We report the performance of these three TCL variants in comparison with the default TCL model in Tab. 5.

It is evident that the inclusion of non-visual input features significantly enhances detection performance across all metrics, as demonstrated by the substantially improved performance of the TCL model and its variants that leverage these features compared to the variants without them. The integration of non-visual input features into the TMM and CAB modules results in a notable improvement in the overall performance of the default TCL framework compared to its two variants, which also include non-visual inputs. This confirms that dynamic environmental changes have a significant impact on pedestrians' motion.

Table 5: Ablation studies on the proposed model using the PIE dataset evaluate the impact of non-visual input features, TMM, and CAB. Each metric's top score is emphasized in boldface, while the runner-up results are underlined.

NV	TMM	CAB	AUC	ACC	F1	Precision	Recall
	✓	✓	0.80	0.84	0.90	<u>0.88</u>	0.93
✓		✓	<u>0.86</u>	<u>0.85</u>	<u>0.91</u>	<u>0.88</u>	0.93
✓	✓		0.84	<u>0.85</u>	<u>0.91</u>	0.86	**0.97**
✓	✓	✓	**0.88**	**0.87**	**0.92**	**0.89**	<u>0.96</u>

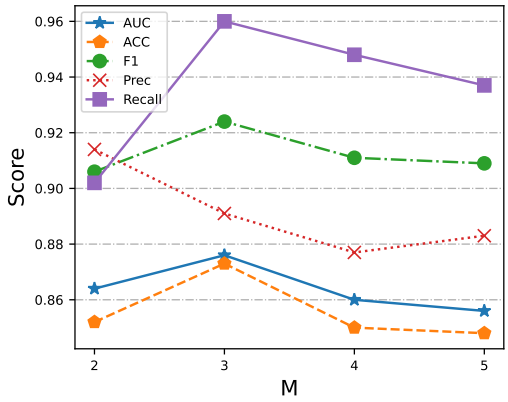

Fig. 3: Performance of TCL w.r.t the cluster number M in TMM.

Additionally, in the CAB, we initially allocate attention to different events and then adaptively aggregate the event feature along with visual and non-visual data. This allows TCL to further focus on the key features that are most crucial for aligning intentions during these dynamic changes. Besides, in the TMM, event clustering ensures that the distinctions between events are significantly greater than those between individual frames. The pronounced variation in pedestrian posture and environmental information enables TCL to avoid attention dispersion due to data redundancy, allowing it to focus more effectively on the dynamic changes related to pedestrians' action. Using the TMM alone achieves the highest recall at 0.97, showcasing TMM's proficiency in detecting samples where pedestrians intend to cross.

4.4 Effectiveness of the Cluster Number

To analyze the impact of the number of clusters M on TCL's performance, we evaluate TCL's performance by varying M from 2 to 5 and report the results in Fig. 3 across all the metrics above. For all metrics except for precision, TCL's performance initially displays an increasing trend, achieving peak performance

when M is set to 3, and then declines. Although TCL yields higher precision when M is set to 2, the recall is significantly lower than when M is set to 3, resulting in less optimal overall performance. Therefore, for this task, setting M to 3 enables the TMM to segment events most effectively, leading to the best overall performance. On the other hand, setting M higher than necessary may also hinder the effectiveness of TCL, as indicated by the declining trend. This is because, within a relatively short time frame, increasing the number of events diminishes their distinctiveness, making critical event identification less meaningful and resulting in a sharp decline in performance.

5 Conclusion

In this paper, we introduce a new PCI method named TCL, which synthesizes temporal feature extraction and contextual attention to merge both visual and non-visual temporal features for PCI. The proposed TCL is implemented by the TMM and the CAB. The TMM merges temporal features into limited key events, and the CAB employs the attention mechanism to fuse contextual features at both the event level and the data level. This allows the TCL to focus on critical dynamic changes, which is important for PCI. Extensive experiments demonstrate the superiority of our TCL compared to state-of-the-art approaches across several datasets. Specifically, the proposed TCL achieves the highest recall values, i.e., 0.96 and 0.98 on the PIE and JAAD-beh datasets, respectively. Overall, this research establishes a new benchmark in pedestrian crossing prediction and contributes to the advancement of autonomous driving systems, promoting safer and more reliable navigation.

While the TCL model holds promise for future applications like vehicle behavior prediction, its performance is contingent upon the quality and diversity of the training data. This can be a significant obstacle in real-world settings. To address this, future research should explore expanding the model's robustness by considering incorporating pedestrian trajectory information and domain adaptation capabilities and by testing more datasets under different environmental conditions.

Acknowledgments. This work is supported in part by the Chongqing Key Project of Technological Innovation and Application (No. CSTB2023TIAD-STX0015, CSTB2023TIAD-STX0031), and by the Key Cooperation Project of Chongqing Municipal Education Commission (HZ2021008).

References

1. Bhattacharyya, A., Fritz, M., Schiele, B.: 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)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: 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 (2017)

3. Dong, M.: Pedestrian cross forecasting with hybrid feature fusion. In: Asian Conference on Machine Learning. pp. 327–342. PMLR (2024)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Du, M., Ding, S., Jia, H.: Study on density peaks clustering based on k-nearest neighbors and principal component analysis. Knowledge-Based Systems **99**, 135–145 (2016)
6. Fang, Z., López, A.M.: Is the pedestrian going to cross? answering by 2d pose estimation. In: 2018 IEEE intelligent vehicles symposium (IV). pp. 1271–1276. IEEE (2018)
7. Ham, J.S., Bae, K., Moon, J.: Mcip: Multi-stream network for pedestrian crossing intention prediction. In: European Conference on Computer Vision. pp. 663–679. Springer (2022)
8. Ham, J.S., Kim, D.H., Jung, N., Moon, J.: Cipf: Crossing intention prediction network based on feature fusion modules for improving pedestrian safety. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3666–3675 (2023)
9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
10. Karasev, V., Ayvaci, A., Heisele, B., Soatto, S.: Intent-aware long-term prediction of pedestrian motion. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 2543–2549. IEEE (2016)
11. Khan, M.A., Sayed, H.E., Malik, S., Zia, T., Khan, J., Alkaabi, N., Ignatious, H.: Level-5 autonomous driving—are we there yet? a review of research literature. ACM Computing Surveys (CSUR) **55**(2), 1–38 (2022)
12. Kotseruba, I., Rasouli, A., Tsotsos, J.K.: Do they want to cross? understanding pedestrian intention for behavior prediction. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 1688–1693. IEEE (2020)
13. Kotseruba, Iuliia and Rasouli, Amir and Tsotsos, John K: Benchmark for evaluating pedestrian action prediction. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1258–1268 (2021)
14. Lorenzo, J., Alonso, I.P., Izquierdo, R., Ballardini, A.L., Saz, Á.H., Llorca, D.F., Sotelo, M.Á.: Capformer: Pedestrian crossing action prediction using transformer. Sensors **21**(17), 5694 (2021)
15. Lorenzo, J., Parra, I., Sotelo, M.: Intformer: Predicting pedestrian intention with the aid of the transformer architecture. arXiv preprint arXiv:2105.08647 (2021)
16. Lorenzo, J., Parra, I., Wirth, F., Stiller, C., Llorca, D.F., Sotelo, M.A.: Rnn-based pedestrian crossing prediction using activity and pose-related features. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 1801–1806. IEEE (2020)
17. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025 (2015)
18. Lv, N., Huang, Y., Zhang, H., Wu, F.: Pedestrian crossing prediction with pathwise feature fusion and stacked gate recurrent unit. IEEE Sensors Letters (2024)
19. Osman, N., Cancelli, E., Camporese, G., Coscia, P., Ballan, L.: Early pedestrian intent prediction via features estimation. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3446–3450. IEEE (2022)

20. Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K.: 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 (2019)
21. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: 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 (2017)
22. Rasouli, A., Yau, T., Rohani, M., Luo, J.: Multi-modal hybrid architecture for pedestrian action prediction. In: 2022 IEEE intelligent Vehicles symposium (IV). pp. 91–97. IEEE (2022)
23. Rasouli, Amir and Kotseruba, Iuliia and Tsotsos, John K: Pedestrian action anticipation using contextual feature fusion in stacked rnns. arXiv preprint arXiv:2005.06582 (2020)
24. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In: Advances in Neural Information Processing Systems (2022)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
26. Yang, B., Zhan, W., Wang, P., Chan, C., Cai, Y., Wang, N.: 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)
27. Yang, B., Zhu, J., Hu, C., Yu, Z., Hu, H., Ni, R.: Faster pedestrian crossing intention prediction based on efficient fusion of diverse intention influencing factors. IEEE Transactions on Transportation Electrification pp. 1–1 (2024). `https://doi.org/10.1109/TTE.2024.3360966`
28. Yang, D., Zhang, H., Yurtsever, E., Redmill, K.A., Özgüner, Ü.: Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. IEEE Transactions on Intelligent Vehicles **7**(2), 221–230 (2022)
29. Yao, Y., Atkins, E., Roberson, M.J., Vasudevan, R., Du, X.: Coupling intent and action for pedestrian crossing behavior prediction. arXiv preprint arXiv:2105.04133 (2021)
30. Zhang, Z., Tian, R., Ding, Z.: Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3534–3542 (2023)
31. Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., Parmar, M.: A review of convolutional neural networks in computer vision. Artificial Intelligence Review **57**(4), 99 (2024)
32. Zhou, Y., Tan, G., Zhong, R., Li, Y., Gou, C.: Pit: Progressive interaction transformer for pedestrian crossing intention prediction. IEEE Transactions on Intelligent Transportation Systems (2023)