

# AugTrEP: Scene and Occlusion-Aware Pedestrian Crossing Intention Prediction

Aditya Bhattacharjee, Steven L. Waslander  
University of Toronto Institute for Aerospace Studies  
Toronto, Canada

aditya.bhattacharjee@mail.utoronto.ca,  
steven.waslander@utoronto.ca

**Abstract**—Accurately predicting the crossing behaviour of pedestrians remains a significant challenge due to their complex behavioural dynamics. Although modern transformer-based models have shown promise in being able to accurately capture these dynamics, the crucial role of contextual information, especially under occluded scenarios, has been underexplored. In this work, we demonstrate that additional contextual features, such as crosswalk visibility and traffic light status, can assist in improving prediction performance under degraded conditions, where accurate pedestrian information is not readily available. We propose AugTrEP, inspired by the existing Transformer-based Evidential Prediction (TrEP) network, which uses two transformer encoders with cross-attention to learn pedestrian behaviour by incorporating global traffic context. We evaluate our models against the PIE benchmark and curated test sets simulating the behaviour of real-world perception systems under varying degrees of occlusion. Our analysis reveals a significant improvement in the accuracy, AUC, F1 score, and precision compared to the baseline under degraded input conditions. These findings highlight AugTrEP’s resiliency to disturbances caused by occlusions and emphasize the importance of scene context in accurate behaviour prediction for real-world applicability.

## I. INTRODUCTION

Autonomous vehicles continue to advance at a rapid pace, with the goal of fundamentally revolutionizing how we navigate and interact with our urban environments. Reports forecast fully-autonomous vehicles to become commercially available and viable by 2030, providing huge improvements in road safety and environmental impact compared to existing systems [1].

However, these systems are yet to fully replicate the intelligence of human drivers in navigating dynamic urban landscapes. In particular, understanding the behaviour of Vulnerable Road Users (VRU) such as pedestrians and cyclists still poses significant challenges. In a 2022 report from the U.S. National Highway Traffic Safety Administration, Level 3 to Level 5 driving systems were involved in 130 accidents, of which 11 involved VRUs [2]. Thus, developing autonomous systems that can effectively reason about the current and future behaviour of VRUs is of utmost importance to ensure a safe environment for all agents in a traffic scene.

The crux of this problem is that existing autonomous systems struggle to understand the complex interactions that occur between VRUs and their urban environments; these

systems often fail to account for the changes in VRU behaviour under varying traffic and environmental conditions. Given that driving is a social phenomenon, there exists a social interaction void, as described in [3], in current autonomous systems. Consequently, ambiguous traffic scenarios frequently result in accidents or unpredictable behaviour due to a lack of social understanding [3]. As a result, many autonomous vehicles are purposely designed to drive conservatively and avoid such interactions [4]. In contrast, human drivers are able to effectively incorporate numerous spatial, temporal, and contextual cues to understand these scenarios and make predictions, but with a certain level of uncertainty due to the inherent unpredictability of human behaviour [5].

Pedestrian crossing prediction is one such area of research where a thorough understanding of the scene is critical in making informed predictions. As per [6], a pedestrian’s willingness to cross is influenced by several factors of which the most prominent ones are dynamic factors, social factors, and physical context. Dynamic factors primarily concern a pedestrian’s estimate of their distance to approaching vehicles and their respective velocities, social factors include cultural norms and group sizes, and finally, environmental factors include traffic signal states, crosswalks, weather, and lighting conditions.

Given such a vast wealth of spatial, temporal, and contextual features, many advanced deep-learning models have been developed to leverage this information [5], [7], [8], [9], [10]. A significant portion of the research has been focused on accurate crossing action prediction using conventional CNN and RNN-based models [4]. Recent advances have seen the introduction of self and multi-head attention transformer-based architectures, leveraging their ability to capture long-range dependencies in sequential input data which have shown promising results.

Existing models rely primarily on pedestrian-centric features such as bounding box coordinates, pose keypoints, their local surroundings, etc. to make a prediction on their crossing intention [11]. However, this largely ignores the global environmental factors which can also affect a pedestrian’s willingness to take risks when crossing the street [11]. Furthermore, training is often done under the assumption that the downstream features fed to the models are generated by a perfect perception pipeline. That is, the models are always

given perfect ground truth information during training and testing. In the absence of perfect information, relying solely on context-invariant features may not yield the best performance when deployed on real-world systems. By proposing to incorporate additional temporal and contextual features into existing models, we aim to address the social interaction gap that currently exists, making the models more effective in scenarios where pedestrian features are not readily available.

The main contributions of our work are summarized as follows:

- We demonstrate that additional temporal and contextual information such as the traffic light state, the presence of crosswalks in the scene, and the pedestrian’s motion state improves the performance of several existing models evaluated against a benchmark dataset.
- We propose AugTrEP, an augmented version of the Transformer-based Evidential Prediction (TrEP) network [8] that is occlusion-aware and incorporates additional traffic context information for predicting pedestrian crossing intention.
- We validate the proposed model on an existing public benchmark dataset which achieves state-of-the-art performance on most metrics. In simulated scenarios where pedestrian-centric features are distorted or missing under varying levels of occlusion, AugTrEP suffers minimal degradation in performance, outperforming the baseline TrEP model with notable improvements in accuracy and F1 score.

## II. RELATED WORK

The particular area of research concerning this problem is known as behaviour prediction, which has numerous applications in autonomous driving, sports, and surveillance [12]. It is categorized as either implicit (where the future trajectories and poses are estimated) or explicit (where predictions are made for future actions or events) [12].

*Pedestrian crossing intention prediction* is a sub-problem within this domain where the objective is to predict whether a given pedestrian will cross the street at some point in the future [12]. Datasets such as Joint Attention in Autonomous Driving (JAAD) [13] and Pedestrian Intention Estimation (PIE) [5] have accelerated the ongoing research in this field, with new models consistently exceeding the previous state-of-the-art (SOTA).

Early approaches often employed a combination of 2D or 3D CNNs or RNNs to tackle this problem. In one of the early works, a variant of AlexNet [14] was used on static frames representing the current traffic scene alongside pedestrian behavioural features to predict the crossing action in a given frame [15]. An alternative approach taken by the authors in [16] leveraged Faster R-CNN [17] to detect and track pedestrians in a scene and a pre-trained pose estimation network to generate 2D poses from monocular images. The pose information provided additional insight into the pedestrian’s movement dynamics which were crucial for better predicting pedestrian crossing behaviour [11].

However, these early CNN-based models primarily captured spatial relationships, overlooking the pedestrian’s behavioural patterns embedded in the temporal dimension. To address this, subsequent models incorporated RNN and attention mechanisms to analyze sequential data and capture both temporal and spatial relationships. One of the earliest works [18] adopting this approach used a pre-trained ResNet [19] model for input image feature extraction and a Long Short-Term Memory (LSTM) [20] model for encoding pedestrian behaviour throughout the sequence for improved crossing predictions. Shortly afterwards, Kotseruba et al. proposed a model [12] that fed a cropped image surrounding the pedestrian to a pre-trained C3D [21] network. These features, along with pedestrian bounding-box coordinates, 2D pose keypoints, and ego-vehicle speed were encoded by a Gated Recurrent Unit (GRU) [22] and then fed to an attention block to generate a crossing prediction.

The incorporation of additional context, such as semantic maps of the entire scene [7] as well as traffic light status [23], into existing models demonstrated the effectiveness of environmental factors in predicting crossing intention. The model proposed in [7] utilized a hybrid-fusion strategy and self-attention layers to enhance prediction performance. Similarly, the work in [23] incorporated Monte-Carlo dropout to estimate the epistemic uncertainty to better identify out-of-distribution samples for improved decision-making, significantly enhancing the performance of many older models.

More recently, numerous works have begun employing transformer encoder and decoder networks in their models due to their inherent ability to capture long-range temporal relationships. One notable example is the model proposed in [8], known as Transformer-based Evidential Prediction (TrEP), to predict the pedestrian crossing action. The authors used a compact multi-head transformer encoder and evidential layer to make uncertainty-aware predictions of a pedestrian’s crossing action. Using only the pedestrian’s bounding box and centroid coordinates as well as the ego-vehicle speed, the model was able to achieve SOTA performance on the JAAD and PIE benchmark datasets.

Despite these advances, a gap remains in fully exploiting the contextual and temporal dynamics in urban environments. Our work builds on this foundation, proposing enhancements that further integrate broader contextual information suitable for real-world applicability. This ensures that newer models are much more robust to imperfections in pedestrian-centric features which may occur in occluded or complex driving scenarios.

## III. METHODOLOGY

### A. Overview

It is critical that autonomous systems remain resilient to numerous complex driving scenarios to ensure safe and reliable behaviour. Given the promising results shown by TrEP [8] with minimal scene context information, we evaluate the model’s robustness in situations where pedestrian features are missing or noisy due to occlusions, to simulate the behaviour of a real-world perception pipeline. We

hypothesize that incorporating additional temporal and contextual information such as pedestrian actions, visibility of crosswalks, and traffic light status can significantly enhance the model’s crossing prediction performance.

To achieve this, we modify the PIE benchmark dataset to replicate scenarios with varying levels of occlusion and introduce enhancements to the TrEP network architecture as well as older established models to evaluate the impact of these additional features. We provide a detailed analysis of model performance under these conditions, demonstrating the effectiveness of contextual information for improved model adaptability to real-world scenarios.

### B. Problem Formulation

The pedestrian crossing prediction task is formulated as a binary classification problem, where the goal is to predict whether a pedestrian will begin crossing the street within a future time interval, given a sequence of  $m = 16$  continuous observation frames. This time interval is known as the time-to-event (TTE) which is defined as the time between the last observation frame and the crossing event frame. The crossing event frame is defined as the first frame in which the pedestrian is observed to cross the street, or the last frame in which the pedestrian is observable if no crossing event occurs. The TTE ranges from 1 to 2 seconds (equivalent to 30–60 frames at a 30fps frame rate) which reflect critical reaction times in urban driving contexts. The goal is thus to predict a single crossing action  $z \in \{0: \text{not crossing}, 1: \text{crossing}\}$  for each provided sequence.

## IV. IMPLEMENTATION

### A. Dataset Preparation

The videos in the PIE dataset are split into 6 sets. As per the benchmark [12], videos from *set01*, *set02*, and *set04* are used for training, *set05* and *set06* for validation, and *set03* for testing.

Using the given TTE and observation length, pedestrian tracks with a minimum length of  $l_{min} = m + \max(TTE) = 76$  frames are extracted. These tracks are then sampled with an overlap ratio of 0.6 to increase the number of training samples and introduce slight variations in pedestrian movement. A total of 4770, 1332, and 3816 samples are generated for the training, validation, and testing sets corresponding to approximately a 50 – 10 – 40% split.

The pedestrian features are also normalized by subtracting the features of the first frame from all subsequent frames. This is done to better capture the changes in pedestrian motion and behaviour with respect to the starting frame. Thus, the normalized input features consist of a  $[15 \times f]$  tensor for each sample with  $f$  representing the number of features.

### B. Feature Extraction

The PIE dataset includes numerous ground truth annotations for pedestrians, vehicles, and other objects in the scene [5], which would all be readily estimable in a standard

perception pipeline onboard an autonomous vehicle. From these, the following features are extracted:

- Pedestrian top-left and bottom-right bounding box coordinates,  $\mathbf{b} = [x_l, y_l, x_r, y_r]$ .
- Pedestrian bounding box centroid coordinates,  $\mathbf{c} = [x_c, y_c]$ .
- Ego-vehicle speed,  $v$ .
- Pedestrian action,  $a \in \{0: \text{standing}, 1: \text{walking}\}$ .
- Crosswalk visibility,  $\sigma \in \{0: \text{not visible}, 1: \text{visible}\}$ .
- Occlusion level  $\Omega \in \{0: \text{none}, 1: \text{partial}, 2: \text{full}\}$
- Traffic light status  $\tau \in \{0: \text{undefined}, 1: \text{red}, 2: \text{yellow}, 3: \text{green}\}$

### C. Data Augmentation

To enhance model performance under real-world conditions, we employ several data augmentation techniques, focusing particularly on addressing the challenges that arise due to occlusions. Considering that pedestrian features are the most susceptible to changes in occlusion, we utilize a hybrid augmentation strategy to modify these features based on the pedestrian’s level of occlusion in each frame in a given sequence.

For a given input sequence  $\mathcal{S}_i = \{\mathcal{F}_1, \dots, \mathcal{F}_{m-1}\}$ , we analyze the pedestrian’s occlusion level, denoted as  $\Omega_i$ , at every frame. These levels are categorized as fully visible (0), partially occluded (1), or fully occluded (2), and are provided as ground-truth annotations in the dataset. These values dictate the type of augmentations we apply on pedestrian bounding box  $\mathbf{b}$  and centroid  $\mathbf{c}$  coordinates.

For partially occluded pedestrians, we first transform the relative coordinates back to their absolute form using a pre-determined initial point. Subsequently, to mirror the effects of occlusion on real-world perception pipelines, we randomly apply scale and positional perturbations with probabilities of 0.5 and 0.7 respectively. We enforce maximum scale and positional perturbation tolerances of 5% and 5 pixels for both  $x$  and  $y$  axes respectively. These parameters are chosen since they introduce sufficient disturbances without compromising the original data.

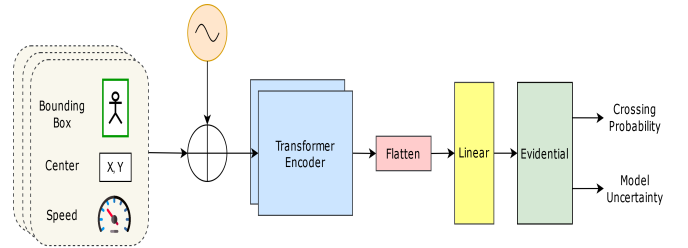


Fig. 1: TrEP Network Architecture

In scenarios concerning fully occluded pedestrians, we opt for a simpler approach. For every frame where the pedestrian is fully occluded, we set the relative bounding box and centroid coordinates to 0, indicating no movement since the last frame in which the pedestrian was visible. This approach aims to simulate worst-case detection failures, providing a

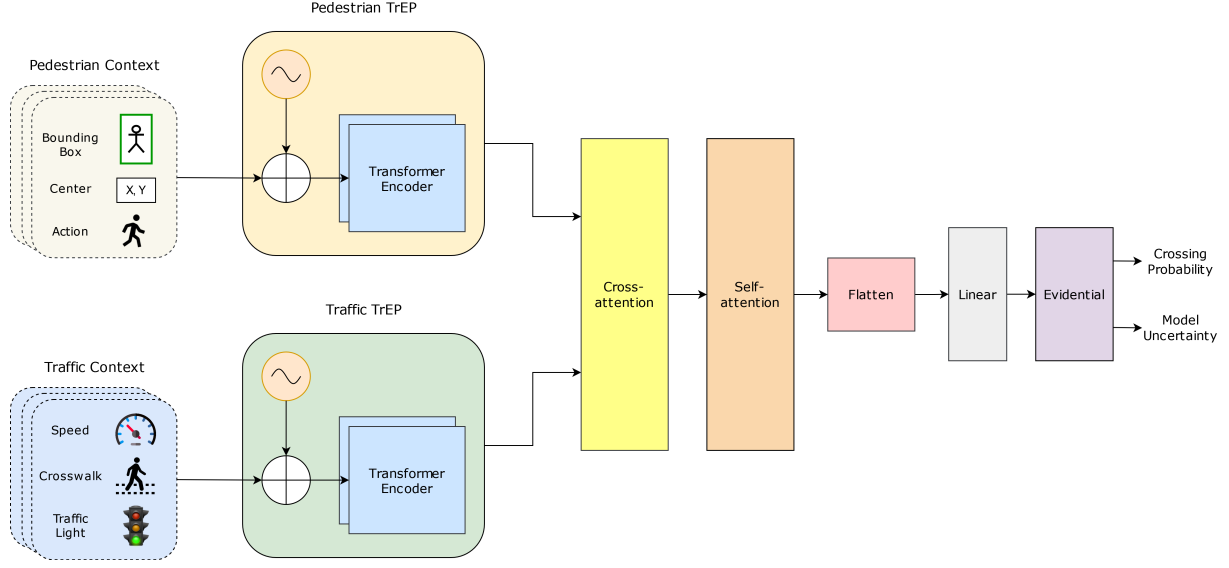


Fig. 2: AugTrEP Network Architecture

baseline for future work that can incorporate prior pedestrian motion models to predict their position.

Furthermore, we prepare two additional test sets based on the existing benchmark to evaluate model robustness to imperfect information. The first one applies the same hybrid augmentation strategy, whereas the other completely removes all relevant pedestrian features if any level of occlusion is observed.

#### D. Baseline Model

1) **Model Architecture:** The original TrEP model is used as the baseline for comparison. The model is built using a positional encoder (to encode temporal relationships between frames in a track) and a 2-layer, 2-headed transformer encoder, as first proposed in [24]. The network architecture for TrEP is shown in Fig. 1.

2) **Model Output:** In contrast with most classification models which output a point probability for each class, TrEP instead uses an evidential layer to transform the model's outputs into parameters of a Dirichlet distribution. Each parameter  $\alpha_i$  represents the belief towards class  $i$  with higher values indicating a stronger belief and less uncertainty. In a binary classification problem, the Dirichlet distribution is parameterized by two  $\alpha$  values. From this, the overall uncertainty  $u$  of the model's predictions for  $C = 2$  classes for sample  $i$  is given as:

$$u = \frac{C}{\sum_{j=1}^C \alpha_{i,j}} = \frac{2}{\alpha_{i,1} + \alpha_{i,2}} \quad (1)$$

The loss function used in TrEP is also used here [8]. For a given sample  $i$ , the loss is given as:

$$\mathcal{L}_i(\Theta) = \sum_{j=1}^2 (z_{i,j} - \mathbf{E}[p_{i,j}])^2 + \mathbf{Var}[p_{i,j}] \quad (2)$$

where  $\Theta$  represents the model parameters,  $p_{i,j}$  refers to the probability distribution of the  $j^{th}$  output class, and  $\mathbf{E}[p_{i,j}]$

and  $\mathbf{Var}[p_{i,j}]$  correspond to the expectation and variance of  $p_{i,j}$  respectively, and are given as:

$$\mathbf{E}[p_{i,j}] = \frac{\alpha_{i,j}}{\alpha_{i,1} + \alpha_{i,2}} \quad (3)$$

$$\mathbf{Var}[p_{i,j}] = \frac{\mathbf{E}[p_{i,j}](1 - \mathbf{E}[p_{i,j}])}{\alpha_{i,1} + \alpha_{i,2} + 1} \quad (4)$$

The overall objective function for  $N$  samples is then given as:

$$\mathcal{L} = \sum_{i=1}^N (\mathcal{L}_i(\Theta) + \lambda \text{KL}[D(\mathbf{p}_i|\mathbf{a}_i)||D(\mathbf{p}_i|\mathbf{1})]) \quad (5)$$

where  $\lambda$  represents the annealing coefficient (set to 10) and  $\text{KL}[\cdot]$  calculates the Kullback-Liebler divergence between two Dirichlet distributions  $D$  with the  $D(\mathbf{p}_i|\mathbf{1})$  term influencing the network to discard classes with minimal evidence.

3) **Training:** To maintain consistency with the original paper, TrEP is trained for 1000 epochs using an Adam optimizer with a learning rate of  $5 \times 10^{-3}$  and a batch size of 32.

#### E. AugTrEP

1) **Model Architecture:** To enhance TrEP's performance in real-world scenarios, we propose AugTrEP, a hybrid architecture inspired by the works presented in [12], [23], and [8] that uses TrEP as its backbone. The modified network architecture is shown in Fig. 2.

In addition to the modified network architecture, we augment the inputs with pedestrian and traffic features. We make a distinction between pedestrian and traffic features as each feature group can be independently used to infer a crossing prediction to a certain extent. The pedestrian  $\mathcal{P} = [\mathbf{b}, \mathbf{c}, a]$  and traffic  $\mathcal{T} = [v, \sigma, \tau]$  feature sets are fed to two separate but identical TrEP networks to independently learn the patterns influencing crossing intention from different perspectives. For a given sample, the shapes of these feature

TABLE I: Model performance on PIE benchmark

Model	Accuracy	AUC	F1	Precision
PCPA	0.85	0.85	0.76	0.69
PCPA + Context	0.89	0.88	0.82	0.77
$\Delta$ Change	4.71%	3.53%	7.89%	11.59%
MaskPCPA	0.86	0.84	0.76	0.72
MaskPCPA + Context	0.89	0.88	0.82	0.80
$\Delta$ Change	3.49%	4.76%	7.89%	11.11%
SFRNN	0.82	0.79	0.69	0.67
SFRNN + Context	0.86	0.84	0.76	0.71
$\Delta$ Change	4.88%	6.33%	10.14%	5.97%
TrEP	<b>0.91</b>	0.94	<b>0.89</b>	<b>0.92</b>
AugTrEP	<b>0.91</b>	<b>0.95</b>	<b>0.89</b>	0.90
$\Delta$ Change	0%	1.06%	0%	-2.17%

vectors are  $[15 \times 7]$  and  $[15 \times 3]$  for the pedestrian and traffic context inputs respectively.

Each TrEP module is configured to have 8 input dimensions, 32 hidden layers, 4 attention heads, and 2 transformer encoder layers. The outputs of each branch are passed to a 4-head cross-attention layer to learn the dynamic patterns in pedestrian behaviour and environmental context that influence the crossing intention. The cross-attention layer produces a query (**Q**), key (**K**), and value (**V**) for a given input feature  $\mathbf{f}_i$  as follows:

$$\mathbf{Q}_i = \mathbf{W}_q \mathbf{f}_i \quad (6)$$

$$\mathbf{K}_i = \mathbf{W}_k \mathbf{f}_i \quad (7)$$

$$\mathbf{V}_i = \mathbf{W}_v \mathbf{f}_i \quad (8)$$

where  $\mathbf{W}_q$ ,  $\mathbf{W}_k$ , and  $\mathbf{W}_v$  are learnable parameters. Cross-attention scores are then computed using queries from one input feature and key-value pairs from another [24]. Given a query  $\mathbf{Q}_i$  from feature  $i$  and key  $\mathbf{K}_j$  and value  $\mathbf{V}_j$  from feature  $j$ , the cross-attention score  $\mathbf{A}_{ij}^h$  for a given head  $h$  is computed as follows:

$$\mathbf{A}_{ij}^h = \text{softmax} \left( \frac{\mathbf{Q}_i \mathbf{K}_j^T}{\sqrt{d_k}} \right) \mathbf{V}_j \quad (9)$$

where  $d_k$  is the dimension of the key vector. The multi-head attention output  $\mathbf{M}$  for  $H$  heads is then simply:

$$\mathbf{M}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{A}^1, \dots, \mathbf{A}^H] \mathbf{W}_0 \quad (10)$$

For this model, we use the pedestrian features as the query and the traffic features as the key-value pairs. This is motivated by the fact that pedestrian behaviour is heavily influenced by social norms [11], which include existing traffic laws. Assuming rational pedestrian behaviour, the environmental context will dictate a pedestrian's risk tolerance to cross and generally constrain their behaviour to predictable actions. Therefore, the output of the cross-attention layer is

simply a  $[15 \times 8]$  tensor which is fed to an external self-attention layer which attempts to learn temporal patterns that indicate a crossing intention. It takes a learnable query vector  $\mathbf{q}$  from the input, which in this case represents a single frame in the sequence. It then calculates the attention score using (9) where the keys and values are the other frames in the sequence. The output of this layer is also a  $[15 \times 8]$  tensor.

The rest of the architecture is identical to the existing TrEP network where the output is flattened, passed through a linear layer with a shape of  $[120 \times 2]$ , and then fed to an evidential layer to produce the crossing prediction and corresponding model uncertainty.

2) **Training:** AugTrEP is trained for 1000 epochs using an Adam optimizer with a learning rate of  $5 \times 10^{-3}$ , and a batch size of 16.

TABLE II: Model performance on PIE benchmark with data augmentations. The PO and FO labels describe the frames with partially and fully occluded pedestrians respectively. The check mark represents the removal of the pedestrian features in these frames during evaluation.

Model	PO	FO	Acc.	AUC	F1	Prec.
TrEP	✗	✗	0.91	0.94	0.89	0.92
TrEP	✗	✓	0.87	0.92	0.86	0.89
TrEP	✓	✓	0.83	0.88	0.83	0.85
TrEP + Data Aug.	✗	✗	0.89	0.95	0.87	0.91
TrEP + Data Aug.	✗	✓	0.88	0.93	0.88	0.89
TrEP + Data Aug.	✓	✓	0.85	0.92	0.84	0.88
AugTrEP	✗	✗	0.91	0.95	0.89	0.90
AugTrEP	✗	✓	0.89	0.95	0.90	0.90
AugTrEP	✓	✓	0.88	0.94	0.88	0.90





Fig. 3: Comparison of Prediction Results from AugTrEP and TrEP in Various Driving Scenarios

TABLE III: AugTrEP ablation study on modified PIE benchmark. The parameters  $a$ ,  $\sigma$ , and  $\tau$  correspond to the pedestrian action, crosswalk visibility, and traffic light status input features respectively, with a check mark indicating the inclusion of the feature during training and evaluation.

Model	$a$	$\sigma$	$\tau$	Acc.	AUC	F1	Prec.
AugTrEP	✓	✓	✓	0.88	0.94	0.88	0.90
AugTrEP	✓	✗	✗	0.84	0.90	0.84	0.85
AugTrEP	✗	✓	✗	0.86	0.92	0.86	0.86
AugTrEP	✗	✗	✓	0.85	0.90	0.86	0.85

## V. EVALUATION

### A. Evaluation Metrics

To compare the performance of each model, the accuracy, area under the receiver operating characteristic (ROC) curve (AUC), F1 score, and precision are evaluated. Each model is trained 3 times and the mean of their performance scores are reported.

### B. Quantitative and Ablative Analysis

To analyze the effectiveness of our features, we use existing RNN-based models such as PCPA [12], MaskPCPA [7], and SFRNN [25] and augment these models with our features (defined as Context) while maintaining the integrity of the original architectures. We then report these results alongside TrEP and AugTrEP. It is important to note that a large portion of the annotated frames in the PIE dataset contain

relevant traffic context features, such as crosswalk visibility  $\sigma$  and traffic light status  $\tau$ . Only about 10.06%, 24.49%, and 5.68% of the total frames in the train, validation, and test sets, respectively, are missing both crosswalk and traffic light features. Thus, there are sufficient samples to highlight the effects of traffic context on model performance against the PIE benchmark.

The results of all the models evaluated against the PIE benchmark are presented in Table I. From the results, it is apparent that incorporating relevant contextual information of the traffic scene has a favourable influence on the model performance as nearly all the models outperform their respective baseline models in most metrics.

We can observe significant gains ranging from 4 – 10% in the accuracy, AUC, F1 score, and precision of PCPA, MaskPCPA, and SFRNN by simply adding information about crosswalks, pedestrian’s actions, and traffic light status which provide additional scene context. However, for AugTrEP, the difference in performance is almost negligible considering that baseline TrEP already performs so well against the benchmark.

To demonstrate the strengths of AugTrEP, we perform an ablation study using the modified test sets prepared earlier which augment the pedestrian features based on occlusion levels. We evaluate the models against three test sets: the original benchmark test set, the test set with fully occluded (FO) frames removed and partially occluded (PO) frames distorted, and the test set with all occluded (FO and PO) frames

removed. The results are summarized in Table II where the check marks describe the frames (FO or PO) whose pedestrian information is removed. Here we can observe that AugTrEP performs better than baseline TrEP when dealing with missing or corrupted pedestrian features. On the test set with all pedestrian features in occluded frames removed, we observe a 6%, 7.48%, 6.12%, and 4.47% improvement in the accuracy, AUC, F1 score, and precision metrics respectively. This signifies that despite missing information about the pedestrian itself, the model is able to rely on contextual information to infer a crossing prediction based on prior knowledge.

We also perform an ablation study analyzing the individual effects of pedestrian action  $a$ , crosswalk visibility  $\sigma$ , and traffic light status  $\tau$  on the model performance, using the test set with all occluded pedestrian features removed. These results are summarized in Table III where a check mark under the feature denotes that it is included during training. From the results, we can see that the model performance is the worst when both the crosswalk visibility  $\sigma$  and traffic light status  $\tau$  are removed. In contrast, the model still performs reasonably well when the pedestrian action  $a$  is removed as long as at least one contextual feature is present, confirming our initial hypothesis. Furthermore, crosswalk visibility plays a critical role in the model's prediction ability, as there is minimal performance degradation when this feature is included. This is likely because crosswalks are not limited to only intersections with traffic lights and are thus applicable to a wider variety of driving scenarios, enhancing the model's ability to generalize. This insight is significant since it shows that only a minimal set of contextual features are required to greatly enhance the model's performance under degraded conditions.

### C. Qualitative Analysis

We present 3 samples with a TTE of 1s from the test set (with FO frames removed) in Fig. 3, highlighting varying levels of pedestrian occlusion. For better visibility in these figures, ground-truth bounding boxes for all frames are drawn in green. We compare the predictions of baseline TrEP and AugTrEP by displaying the crossing intention probability  $P$  and model uncertainty  $U$ , where correct predictions are coloured green and red otherwise.

From the results, it can be observed that AugTrEP produces more accurate and confident predictions due to the additional scene context information. This is evident in Fig. 3a) where AugTrEP correctly predicts the crossing intention of the occluded pedestrian by leveraging the available crosswalk and traffic light information whereas TrEP is unable to do so. In Fig. 3b), both models incorrectly predict a not-crossing intention for the pedestrian standing at the intersection. In this scene, crosswalks are visible and the traffic light status is red but the pedestrian is occluded and relatively far away from the ego-vehicle. The prediction is likely influenced by the vehicle's speed which remains relatively constant throughout the sequence, coming to a stop when the pedestrian is already halfway across the intersection. Finally,

in Fig. 3c) where crosswalk and traffic light information are missing, the two models still correctly predict a crossing action, with AugTrEP producing a more confident result since it has additional knowledge of the pedestrian's motion state.

On a qualitative level, AugTrEP demonstrates a deeper contextual understanding of the environment. The model is able to incorporate subtle environmental and situational cues to better understand pedestrian behaviour, similar to human-like perception. Crosswalks are a critical feature that inform the model of likely pedestrian trajectories and often correlate with a pedestrian's intention to cross depending on their observed state. Similarly, the traffic light status dictates the likely direction of pedestrian movement. By incorporating these features, the model can make predictions with greater certainty under the assumption that existing traffic rules are not violated. This provides the foundation for further context-based analysis where incorporating additional traffic scene information, such as the distance to key roadside features, may result in better estimates. Understanding these complex interactions is critical to enabling safer decision-making processes for autonomous vehicles.

## VI. CONCLUSION

This paper presents a method for integrating additional temporal and contextual information into a transformer-based model for pedestrian crossing prediction. As demonstrated in the results, we significantly improve the model's ability to understand the complex interactions that occur amongst road users under simulated real-world conditions. We propose a new model named AugTrEP, which builds upon the existing TrEP network by incorporating scene context features extracted from the PIE dataset. Furthermore, we also extend older established models with these features, to demonstrate the effectiveness of additional scene context on prediction capability.

Our evaluation of these models against the PIE benchmark reveal the significance of traffic context on enhancing model performance, under both ideal and simulated worst-case scenarios. Through our ablation studies, we highlight the importance of global traffic context in understanding pedestrian behaviour in the absence of reliable pedestrian features.

It is important to acknowledge that these results are presently representative of the model's performance against the PIE benchmark, which is chosen primarily due to the accessibility of relevant traffic context features. Thus, future extensions of this work could focus on expanding the diversity of the training data to capture a wider array of pedestrian behaviours, either through expanded data collection and labeling, or through simulation. This is necessary to enhance the model's robustness and ability to generalize across varying real-world scenarios. Moreover, variations in weather and lighting conditions significantly influence a pedestrian's behaviour [11]. Therefore, the analysis of models' performance under these diverse conditions remains an area well-suited for further research. Finally, to obtain a more

nuanced understanding of the model uncertainty, we wish to concentrate our future efforts on further exploring advanced evaluation metrics for uncertainty estimation, similar to what was proposed in [26].

With this work, we aim to provide a basis for future innovations in the development of resilient and context-aware behavioural prediction models suitable for real-world deployment. We hope that this further motivates research efforts in enhancing the decision-making capabilities of autonomous systems in complex urban environments, ensuring safer and efficient interactions between vehicles and vulnerable agents.

## REFERENCES

- [1] T. Litman, "Autonomous vehicle implementation predictions: Implications for transport planning," tech. rep., Victoria Transport Policy Institute, Oct. 2023.
- [2] National Highway Traffic Safety Administration, "Summary report: Standing general order on crash reporting for automated driving systems," tech. rep., U.S. Department of Transportation, June 2022.
- [3] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2020.
- [4] N. Sharma, C. Dhiman, and S. Indu, "Pedestrian intention prediction for autonomous vehicles: A comprehensive survey," *Neurocomputing*, vol. 508, pp. 120–152, 2022.
- [5] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6261–6270, 2019.
- [6] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Understanding pedestrian behavior in complex traffic scenes," *IEEE Transactions on Intelligent Vehicles*, vol. 3, pp. 61–70, 2018.
- [7] D. Yang, H. Zhang, E. Yurtsever, K. A. Redmill, and U. Ozguner, "Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention," *IEEE Transactions on Intelligent Vehicles*, vol. 7, pp. 221–230, 2021.
- [8] Z. Zhang, R. Tian, and Z. Ding, "Trep: Transformer-based evidential prediction for pedestrian intention with uncertainty," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 3534–3542, Jun. 2023.
- [9] L. Achaji, J. Moreau, T. Fouqueray, F. Aioun, and F. Charpillat, "Is attention to bounding boxes all you need for pedestrian action prediction?," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, pp. 895–902, 2022.
- [10] J. Li, X. Shi, F. Chen, J. C. Stroud, Z. Zhang, T. Lan, J. Mao, J. Kang, K. S. Refaat, W. Yang, E. Je, and C. Li, "Pedestrian crossing action recognition and trajectory prediction with 3d human keypoints," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1463–1470, 2023.
- [11] T. Chen and R. Tian, "A survey on deep-learning methods for pedestrian behavior prediction from the egocentric view," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 1898–1905, 2021.
- [12] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1257–1267, 2021.
- [13] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Joint attention in autonomous driving (jaad)," *ArXiv*, vol. abs/1609.04741, 2016.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.
- [15] A. Rasouli, I. Kotseruba, and J. 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)*, pp. 206–213, 2017.
- [16] Z. Fang and A. M. López, "Is the pedestrian going to cross? answering by 2d pose estimation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, p. 1271–1276, IEEE Press, 2018.
- [17] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.
- [18] J. Lorenzo, I. Parra, F. Wirth, C. Stiller, D. F. Llorca, and M. A. Sotelo, "Rnn-based pedestrian crossing prediction using activity and pose-related features," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1801–1806, IEEE, 2020.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.
- [20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, p. 1735–1780, nov 1997.
- [21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.
- [22] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [23] M. Upreti, J. Ramesh, C. R. Kumar, B. Chakraborty, V. BALISAVIRA, P. Czech, V. Kaiser, and M. Roth, "Uncertainty and traffic light aware pedestrian crossing intention prediction," 2022.
- [24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [25] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked rnns," *ArXiv*, vol. abs/2005.06582, 2020.
- [26] J. Deery, C. W. Lee, and S. L. Waslander, "Propandl: A modular architecture for uncertainty-aware panoptic segmentation," in *2023 20th Conference on Robots and Vision (CRV)*, pp. 137–144, 2023.