TrEP: Transformer-Based Evidential Prediction for Pedestrian Intention with Uncertainty

Zhengming Zhang[1], **Renran Tian**[2], **Zhengming Ding**[3]
[1] School of Industrial Engineering, Purdue University, West Lafayette, Indiana, USA
[2] Department of Computer Information Technology, Indiana University Purdue University Indianapolis, Indiana, USA
[3] Department of Computer Science, Tulane University, New Orleans, Louisiana, USA
zhan3988@purdue.edu, rtian@iupui.edu, zding1@tulane.edu

Abstract

With rapid development in hardware (sensors and processors) and AI algorithms, automated driving techniques have entered the public's daily life and achieved great success in supporting human driving performance. However, due to the high contextual variations and temporal dynamics in pedestrian behaviors, the interaction between autonomous-driving cars and pedestrians remains challenging, impeding the development of fully autonomous driving systems. This paper focuses on predicting pedestrian intention with a novel transformer-based evidential prediction (**TrEP**) algorithm. We develop a transformer module towards the temporal correlations among the input features within pedestrian video sequences and a deep evidential learning model to capture the AI uncertainty under scene complexities. Experimental results on three popular pedestrian intent benchmarks have verified the effectiveness of our proposed model over the state-of-the-art. The algorithm performance can be further boosted by controlling the uncertainty level. We systematically compare human disagreements with AI uncertainty to further evaluate AI performance in confusing scenes. The code is released at https://github.com/zzmonlyyou/TrEP.git.

Introduction

With the rapid progress in AI technologies, the numerous successes in intelligent transportation systems have made autonomous driving promising (Liu et al. 2022; Tang et al. 2023; Cui et al. 2022; Liu et al. 2021; Zeng et al. 2021). These transformative technologies have the potential to fundamentally change daily life for everyone and create vast social and individual benefits (Litman 2017). Mercedes has recently begun selling their Level 3 self-driving system (defined by SAE International as Conditional Driving Automation) on their S-Class, marking a significant milestone as higher-level automated driving techniques enter ordinary people's lives. However, while Level 2 and 3 automated systems can drive autonomously under human supervision and within the Operational Design Domain (ODD), the main challenge for fully automated cars to safely and efficiently drive in urban settings remains interactions with pedestrians (Domeyer, Lee, and Toyoda 2020; Herman et al. 2021; Zhang, Tian, and Duffy 2023).

A large number of studies have concentrated on modeling and predicting pedestrian behaviors, with various deep learning techniques and benchmark data sets constructed in the past few years (Chen and Tian 2021). In general, trajectory prediction represents the majority of pedestrian behavior modeling efforts. While traditional algorithms predict trajectory in a fixed bird eye view from mainly surveillance cameras (Xu, Piao, and Gao 2018; Zhang et al. 2019; Shi et al. 2021; Ma, Karimpour, and Wu 2020; Liu et al. 2020b,c), many recent studies focus on the moving egocentric view in front of the vehicles to better serve the needs of automated driving (Rasouli et al. 2019; Rasouli, Rohani, and Luo 2021; Yagi et al. 2018; Chen, Tian, and Ding 2021). Although recent trajectory prediction algorithms have achieved improved accuracies, inherent limitations prevent satisfying prediction accuracy in longer prediction horizon (Herman et al. 2021), including the behavior temporal dynamics, uncertainty related to the scene complexity, and accumulated position prediction errors over time steps. Such limitations restrict the common trajectory prediction horizon to about 1-2 seconds.

A Limited trajectory prediction horizon may be sufficient for automatic braking features focusing on last-second braking to improve safety, but cannot support efficient motion planning for higher-level automatic cars to interact with pedestrians smoothly. Some studies have shown that human drivers need at least 3 seconds of prediction horizon to plan driving behaviors during pedestrian interactions (Herman et al. 2021; Zhang et al. 2022a, 2021b; Pang, Guo, and Zhuang 2022), indicating a similar requirement for automatic driving algorithms. Also, in the case to detect the out-of-ODD event and start the transition from automatic control to manual driving, drivers need up to 20 seconds to fully control the car given a sudden automatic driving failure (Eriksson and Stanton 2017; Merat et al. 2014), which poses high requirements of pedestrian behavior prediction horizon as well to ensure driving safety.

Solutions are needed to address the limitations of pedestrian trajectory prediction. Besides some task-specific probabilistic behavior prediction metrics such as In-ROI Sensitivity (IRS) (Herman et al. 2021), many studies started to focus on pedestrian intention prediction. The goal of intention prediction is to help identify crossing pedestrians (Fang and López 2018), anticipating crossing timing (Zhang et al.

2021a), and improving trajectory and action prediction with intentions as guidance and boundaries (Rasouli et al. 2019; Yao et al. 2021; Wang et al. 2022; Jing et al. 2022; Zhang et al. 2022c; Ding et al. 2018; Zhang et al. 2022b).

Although pedestrian intention prediction appears to be a promising and vital research direction, there are significant challenges of uncertainty that are neglected in the current research frontier. Many studies have found uncertainty and disagreement among human annotators (Wu et al. 2022; Ji et al. 2021). In human annotations (Rasouli et al. 2019; Chen et al. 2021), significant disagreements among human drivers in estimating pedestrian crossing intentions were observed. Given the same driving scenes, two studies reported that human drivers not only disagreed on pedestrian crossing intentions at the same pre-determined critical frames but also tended to estimate crossing/non-crossing at different timings. These phenomena reflect the uncertainty of understanding complex pedestrian-crossing driving scenes, which are highly dynamic, non-deterministic, and context-dependent. We argue that without modeling uncertainty in intention prediction, deep learning algorithms will struggle to achieve higher accuracies and predictability.

Our Contributions. We propose a transformer-based evidential prediction (**TrEP**) algorithm for uncertainty-aware estimation of pedestrian intentions. Taking ego-centric pedestrian encountering scene videos as input, the algorithm automatically learns the evidence[1] towards different intention categories from the motion information of the car and targeted pedestrian. Trained evidence distributions in the high-dimensional spatial-temporal-mixed feature space are then employed for intention prediction and uncertainty estimation. The study has achieved four main contributions:

- The proposed TrEP is able to capture more temporal correlation and be aware of pedestrian intention uncertainty so that it outperforms the state-of-the-art algorithms on three benchmark datasets with large margins.
- Strong negative relationship has been observed consistently between the uncertainty levels and algorithm prediction accuracies, with the uncertainty-aware prediction helping to secure high-level accuracy reliably by filtering out the cases with higher uncertainty levels.
- The uncertainty associated with pedestrian intention estimation results improves the predictability and trustworthiness of the algorithm behavior, which can significantly enhance human-AI coordinated automatic driving.
- Our data-driven pedestrian intention estimation uncertainty learned by the model is comparable with corresponding human disagreements in certain situations, although human annotation disagreement levels are not included during the model training.

Related Works

Pedestrian intention prediction received a lot of attention in recent years to facilitate the interactions between autonomous cars and vulnerable road users. Based on some pioneering benchmark datasets on pedestrian intention, like the JAAD (Rasouli, Kotseruba, and Tsotsos 2017) and PIE (Rasouli et al. 2019), a lot of pedestrian intention prediction algorithms have been proposed (Rasouli, Kotseruba, and Tsotsos 2017; Liu et al. 2020a; Kotseruba, Rasouli, and Tsotsos 2020). One early work used a CNN to extract features from a static frame of driving scenes to predict pedestrian intention(Rasouli, Kotseruba, and Tsotsos 2017). In another study, Fang et al. used a pre-trained pose estimation network to estimate pedestrian pose and then predict the crossing intention (Fang and López 2018). More recently, a graph convolution network was trained to model the pedestrian pose along with visual features for intention prediction (Chen, Tian, and Ding 2021). Although different deep learning structures have been implemented in the domain, most present studies consider the input as a sequence of frames and the output as a single probability of crossing (Rasouli et al. 2019; Gujjar and Vaughan 2019; Liu et al. 2020a; Rasouli, Kotseruba, and Tsotsos 2020).

In a comparison study (Kotseruba, Rasouli, and Tsotsos 2021), results show that both 3D convolution networks and two stream networks are capable of dealing with the temporal visual information (Simonyan and Zisserman 2014; Tran et al. 2015; Carreira and Zisserman 2017). One proposed a network fusing the temporal-spatial features from a 3D CNN along with the bounding box coordinates and vehicle speed predictions (Kotseruba, Rasouli, and Tsotsos 2021). A recent work utilized the self-attention mechanism to capture the spatial-temporal feature and fused it with semantic segmented context (Yang et al. 2022). Similarly, (Rasouli et al. 2022) adopted an attention mechanism to fuse the multi-modal features that achieved state-of-the-art performance.

Most previous studies have utilized the RNN-based encoder-decoder framework to develop their models. While RNN-variants, such as LSTM, have incorporated mechanisms to capture the temporal relationships across the frame series (Qu et al. 2020), the pedestrian intention prediction domain has not fully explored other modern techniques, such as transformer-based sequential models. The latter option has the potential to capture longer temporal patterns.

More importantly, none of the existing algorithms has adopted pedestrian intention estimation uncertainty as inputs or outputs of their models. In particular, we want to emphasize that existing algorithms rely solely on the accuracy or F1 scores to evaluate intention prediction performance, ignoring the facts that the ground-truth labels in benchmark datasets (Rasouli et al. 2019; Chen et al. 2021) contain inherent uncertainties. The disagreement levels among human annotators shall be considered when developing and evaluating corresponding algorithms.

Our Proposed Method

Preliminary & Motivation

The goal of intention prediction is to determine if the interested pedestrian is crossing or not given the raw input (Chen et al. 2021; Rasouli et al. 2019). Thus, it can be formulated as a binary classification $I \in \{0, 1\}$, where 1 indicates crossing and vice versa. Given a sequence of ego-centric frames

[1]A higher-order coding scheme for scene features following the Dirichlet distribution, with details described in later sections.

$\{s_1, s_2, ..s_l\}$ with length l, there is a bounding box represented by a quaternion, b_i, in each frame i annotating the same pedestrian. Each quaternion contains the 2D coordinates for the upper-left and bottom-right points of the bounding box. In addition to the bounding box and visual information, each frame comes with an action annotation of the ego-vehicle, a_i, such as speed for PIE and driver action for JAAD (Rasouli, Kotseruba, and Tsotsos 2017).

However, existing methods ignore the pedestrian intents to be conflicting in terms of various drivers, defined as human disagreement. It is intuitive to capture such intent uncertainty so that AI can mimic human cognition. To achieve this, evidential learning models second-order probabilities and uncertainty (Sensoy, Kaplan, and Kandemir 2018; Amini et al. 2020; Bao, Yu, and Kong 2021), instead of modeling the probability assignment of a given sample. In other words, a Dirichlet distribution parameterized over evidence represents the density of each such probability assignment, where the predicted evidence (parameters of the Dirichlet distribution) is the model output. In this sense, we could consider the uncertainty as a variance estimation of the Dirichlet distribution.

Framework Overview

Compared to (Chen et al. 2021; Rasouli et al. 2019), our model is more compact. Since our objective is to devise an intention prediction model using purely tabular data, we discarded the visual information $\{s_1, s_2, ..s_l\}$. The input is a sequence of tabular information including bounding box and ego-vehicle action. The first type is to simply consider the quaternion b_i a type of feature. The second one is inspired by the SORT tracking algorithm (Bewley et al. 2016). We first calculate the center of the bounding box denoted as a tuple c_i and then compute the area and ratio between the length and width of the bounding box (a_i, r_i) for each frame i. The overall structure of our model is shown in Figure 1.

Base Model Firstly, we concatenate all the input features b_i, c_i, a_i, r_i at each frame i to get the feature x_i. y_i is its corresponding ground-truth intent label. RNN-based encoder-decoder captures the temporal correlation through model parameters (like memories), transformer-based model (Vaswani et al. 2017a; Han et al. 2021; Xu et al. 2021; Yi and Qu 2022; Wu et al. 2023) design attention modules to capture all the possible relationships. In other words, for a trained model, the attention mechanism relies on the data itself explicitly to capture the temporal correlation, while LSTM/RNN memorizes the temporal information implicitly through model parameters.

Thus, a shared feed-forward layer is used to extend the feature dimension for a later transformer layer with multi-head attention (where the output is f_i.) Before the sequence fusion, we include a positional encoder to add temporal information g_i. The positional encoder injects some information about the relative or absolute position of the frames in the sequence. The positional encodings are summed with the inputs of the transformer, $k_i = x_i + g_i$. We use sine and cosine functions of different frequencies to encode the temporal order. The later layers until the last

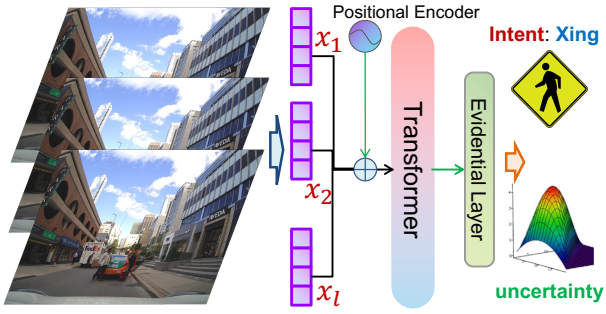

Figure 1: Overview of the proposed model, where transformer module aims to capture temporal correlation explicitly from $\{x_1, x_2, \cdots, x_l\}$ while the evidential layer is to generate the model uncertainty u, providing one metric to reject the model prediction.

feed-forward layer could be considered a simple version of the transformer encoder (Vaswani et al. 2017b). The transformer encoder includes blocks of one residual connected self-attention layer and one fully connected layer with layer normalization. It transformed each input vector according to self-attention. Then, we simply flatten the output for each component at frame i to merge the sequence to get final embeddings $\mathbf{f}$ as the following formulation: $\mathbf{f} = \mathbf{Flatten}(\mathbf{Transformer}(k_1, k_2, \ldots, k_l))$.

After that, we deploy the softmax as the activation function on concatenated feature $\mathbf{f}$, and apply the cross-entropy loss function as follows:

$$\mathcal{L} = -\sum_{j=1}^{N} y_j \log\Big(\mathbf{Softmax}(\mathbf{f}_j)\Big). \tag{1}$$

Uncertainty-Aware Evidential Learning In our **Base Model** for intent prediction, the softmax function is used to predict intent assignment probabilities. However, it provides only a point estimate for the intent probabilities of a sample and does not provide the associated uncertainty for this prediction. On the other hand, Dirichlet distributions can be used to model a probability distribution for the class probabilities (Sensoy, Kaplan, and Kandemir 2018; Sensoy et al. 2020). Therefore, we can use the variance estimates of the Dirichlet distribution to calculate the model uncertainties.

We replaced our last activation function (softmax) with a rectified linear unit (ReLU) to ascertain non-negative outputs. The outputs are no longer modeled as the probability of the classes. Instead, they are considered the evidence for the predicted Dirichlet distribution. In addition, we did not use the cross-entropy loss, since the training goal is not to maximize the likelihood of the model parameters given samples. By decomposing the loss function in Equation (2), the first part aims to achieve the goals of minimizing the prediction error while reducing the variance of the Dirichlet experiment generated by the model, specifically for each sample in the training set (Sensoy, Kaplan, and Kandemir 2018).

Given any sample i, the evidence e_i refers to the output, and the uncertainty estimation loss is modeled as

$$\mathcal{L}_i(\Theta) = \sum_{j=1}^{K} (y_{ij} - \mathbf{E}[p_{ij}])^2 + \mathbf{Var}(p_{ij}), \tag{2}$$

where y_{ij} is the j-th element of ground-truth label y_i, and p_{ij} is the j-th element of the probabilistic prediction p_i referring to a simplex for probability assignments for sample i. $\mathbf{E}[\cdot]$, and $\mathbf{Var}(\cdot)$ are the operators for the expectation and variance over the Dirichlet distribution. Note p_i is not the output of the model but a random vector following the Dirichlet distribution. To estimate the p_j, we use the following equation $\mathbf{E}[p_{ij}] = \frac{\alpha_{ij}}{S}$, where $S = \sum_{j=1}^{K} \alpha_{ij} = \sum_{j=1}^{K}(e_{ij} + 1)$. $\mathbf{Var}[p_{ij}] = \frac{\mathbf{E}[p_{ij}](1-\mathbf{E}[p_{ij}])}{(S+1)}$. K indicates the number of classes and Θ represent the model parameters.

In this sense, the overall objective function integrated with Kullback–Leibler (KL) divergence is formulated as:

$$\mathcal{L} = \sum_{i=1}^{N} \Big(\mathcal{L}_i(\Theta) + \lambda \mathrm{KL}\big(D(p_i|\alpha_i)||D(p_i|\mathbf{1})\big)\Big), \quad (3)$$

where N indicates the number of samples and $D(p_i|\mathbf{1})$ means a uniform prior if there is no evidence for the assignment. $\mathrm{KL}(\cdot)$ denotes the operators for the KL divergence, which aims to regularize our predictive distribution by penalizing those divergences from the "I do not know" state that do not contribute to data fit. the λ is the trade-off parameter (by default we set it to 10.). The model uncertainty u is computed as $u = \frac{K}{S}$.

Experiment

Dataset

In our experiments, three intention/action prediction benchmarks are explored, which are JAAD (Rasouli, Kotseruba, and Tsotsos 2017), PIE (Rasouli et al. 2019), and PSI (Chen et al. 2021). To our best knowledge, those three benchmarks are the most representative datasets regarding the intent prediction task. Specifically, JAAD and PSI are collected in a similar sense where both of them consist of recorded dashcam video clips. In contrast, PIE is collected in a continuum fashion, where the entire dataset is from a 4-hour drive in Toronto downtown. PIE and JAAD have a similar number of annotated pedestrians ($> 1k$), while PSI is smaller-scale.

Furthermore, both PIE and JAAD datasets utilized a similar annotation pipeline, wherein each pedestrian was labeled with crossing action and crowdsourcing-labeled intention labels, and we used the crossing labels as a substitute for estimating pedestrian intention. However, the PSI dataset identified weaknesses in the above-mentioned approach and adopted an intention segmentation methodology to tackle the issue of intention dynamics. Specifically, PSI annotated the crossing intention of each pedestrian for every frame.

Evaluation and Metrics

To compare with the existing works fairly, we applied the evaluation protocol for both PIE and JAAD datasets (Kotseruba, Rasouli, and Tsotsos 2021). In short, we sampled clips at least one second before the appearance of the crossing action and predict the crossing intention. We set the overlap ratio as 0.5 for both datasets. In total, PIE has 3,980 training sequences 995 of which are crossing cases. On the other hand, JAAD has 3,955 training sequences including 805 crossing cases. For the ego-vehicle action annotation, JAAD offered the driver's behaviors while PIE recorded the speed of the ego-vehicle. We calculated the F1 score, accuracy, the area under the receiver operating characteristic (AUC), and precision to comprehensively evaluate the model performance.

Because of the different annotations between PSI and JAAD/PIE, we followed the original PSI for the task setup (Chen et al. 2021). We sampled the clips with an overlap ratio of 0.8 across the whole video as long as the pedestrian appears. Differently from one intention label for each pedestrian in PIE/JAAD, the annotated pedestrians in PSI have a crossing intention label for each frame. The prediction task is to assign the crossing intention at the 16th frame given 15 frames as input. Note that PSI does not provide any kind of ego-vehicle action annotations. There are 6,262 training sequences with 3,927 crossing cases. For the convenience of comparing with the others, we reported accuracy, F1 score, and balanced accuracy for the models trained on PSI.

Implementation Details

Due to the different annotations and feature engineering for the model on each dataset, the input dimensions ($b \times t \times f_d$) are slightly different (where b refers to batch size (we set $b = 64$), t and f_d refers to the size of time span and feature dimension). We projected the input features dimensions f_d to 8 dimensions in the first linear layer. The fully connected layers in the transformer projected the 8 dimensions to 16. There is one layer of multi-head attention (2 heads) for PIE and PSI and two layers for JAAD. The dropout rates are set to 0.1. All the models are trained by Adam optimizer with a learning rate of 5e-3 for 2,000 epochs.

Comparison Results

Results on PIE/JAAD The benchmark results for models trained on PIE and JAAD are shown in Table **??**, where we compare with ATGC (Rasouli, Kotseruba, and Tsotsos 2017), I3D (Carreira and Zisserman 2017), MM-LSTM (Aliakbarian et al. 2018), SF-GRU (Rasouli, Kotseruba, and Tsotsos 2020), PCPA (Kotseruba, Rasouli, and Tsotsos 2021), MMHA (Rasouli et al. 2022), and BiPed (Rasouli, Rohani, and Luo 2021). ATGC is the only model with a static input (one frame of the sequence). I3D is a well-known 3D convolution network for video action recognition. PCPA, MMHA, and BiPed used multi-modality as their input, where the main difference is the incorporation of the fusion methods. Our proposed models outperformed all the existing models on both benchmark datasets. BiPed (Rasouli, Rohani, and Luo 2021) performed close to our proposed model on the PIE dataset, which adopts multi-modality sources as input. Our model simply used bounding box information along with ego-vehicle actions. However, our model dominated the others on JAAD datasets for all metrics. Especially, the AUC score increased by 9%. On both datasets, the performance of our base and evidential models are similar, which demonstrates the feasibility of evidential deep learning.

Results on PSI Table **??** listed the performance for all models trained on the PSI dataset, where we compare with

	PIE				JAAD			
Model\Metric	Accuracy	AUC	F1	Precision	Accuracy	AUC	F1	Precision
ATGC	0.59	0.55	0.36	0.35	0.64	0.60	0.53	0.50
I3D	0.79	0.75	0.64	0.61	0.82	0.75	0.55	0.49
MM-LSTM	0.84	0.84	0.75	0.68	0.80	0.77	0.58	0.51
SF-GRU	0.86	0.83	0.75	0.73	0.83	0.77	0.58	0.51
PCPA	0.86	0.84	0.76	0.73	0.83	0.77	0.57	0.50
MMHA	0.89	0.88	0.81	0.77	0.84	0.80	0.62	0.54
BiPed	0.91	0.90	0.85	0.82	0.83	0.79	0.60	0.52
Ours	0.91	0.93	0.85	0.84	0.87	**0.88**	0.63	0.63
Ours ($u = 1$)	0.92	**0.94**	0.85	0.88	0.88	0.86	0.61	0.70
Ours ($u = 0.6$; PIE = 96%, JAAD = 89%)	**0.93**	**0.94**	**0.87**	**0.89**	**0.91**	0.86	**0.69**	**0.71**

Table 1: Performance of the proposed models and the other existing models on the JAAD and PIE datasets. u refers to the uncertainty threshold, where we reject the predictions with higher uncertainties. When $u = 1$, all the samples are included. When $u = 0.6$, 96% of the PIE dataset and 89% of the JAAD dataset are included.

Model\Metric	Accuracy	Balanced Accuracy	F1
VR-GCN	0.74	0.61	0.64
PIE-Intention	0.69	0.58	0.79
PSI-Intention	0.76	0.67	0.66
Ours	0.83	0.75	0.88
Ours ($u = 1$)	0.82	0.75	0.87
Ours ($u = 0.6$; 75%)	**0.85**	**0.77**	**0.90**

Table 2: Performance of the proposed models and the other existing models on the PSI dataset.u refers to the uncertainty threshold, where we reject the predictions with higher uncertainties. When $u = 1$, all the samples are included. When $u = 0.6$, 75% of the PSI dataset is included.

VR-GCN (Chen, Tian, and Ding 2021), PIE-Intention (Rasouli et al. 2019) and PSI-Intention (Chen et al. 2021). VR-GCN used graph neural networks to model the pedestrian poses, whereas PSI-intention is based on multi-task learning (reasoning, trajectory, and intention). Similar to the JAAD and PIE datasets, our proposed model performed better than the existing works. The F1 scores are boosted by 12%, and the accuracy is increased by 7%. Again, the evidential and base models performed comparably on the PSI dataset.

Ablation Study

In the ablation study, all the reported results are based on the base model, because we did not find any significant difference between the base and evidential models. We test models with different combinations of features. Also, we trained models with and without positional encoder. All the results are shown in Table 3. Besides the bounding box feature, we found that the center coordinates of the bounding box are a very useful feature, which boosted at least 8% of the accuracy for all datasets. On the other hand, the ratio between the width and length of the bounding box and the bounding box area is helpful in PIE while decreasing the performance in PSI. One reason might be the number of sample sequences. The sequences sampled from the PSI are nearly twice the sequences from PIE. At last, our results proved that the use of a positional encoder increased the performance in all datasets because it allows the model to capture the temporal changes.

	PIE		JAAD		PSI	
Model\Metric	Acc	F1	Acc	F1	Acc	F1
Bbox+Action	0.80	0.72	0.79	0.58	0.72	0.69
Bbox+Action +Center	**0.91**	**0.85**	**0.87**	0.63	0.80	0.85
Bbox+Action +Center+Ratio	0.89	0.81	0.86	**0.65**	**0.83**	**0.88**
No Pos. Encoder	0.90	0.83	0.85	0.61	0.81	0.87

Table 3: Performance for each variation of the base model on three datasets. “bbox” refers to bounding box coordinates. “action” refers to ego-vehicle action. ”center” refers to the coordinates of the bounding box center. “ratio” refers to the bounding box area and the ratio between length and width.

Uncertainty Analysis

Intuitively, we hypothesize that the samples with higher uncertainty generated by the evidential model have lower scores using the metrics because the evidence in our framework is a measure of the amount of support collected from data in favor of a sample to be classified into a certain class. In other words, our evidential mode is not confident in the prediction when having a high uncertainty score.

The test samples were grouped based on predicted uncertainty, as depicted by the blue bars in Fig. 2, with each bar representing the proportion of samples falling within the corresponding uncertainty range. For the top three graphs in Fig. 2, the uncertainty value refers to the range that the uncertainty that is less than the value and 0.1 larger than the

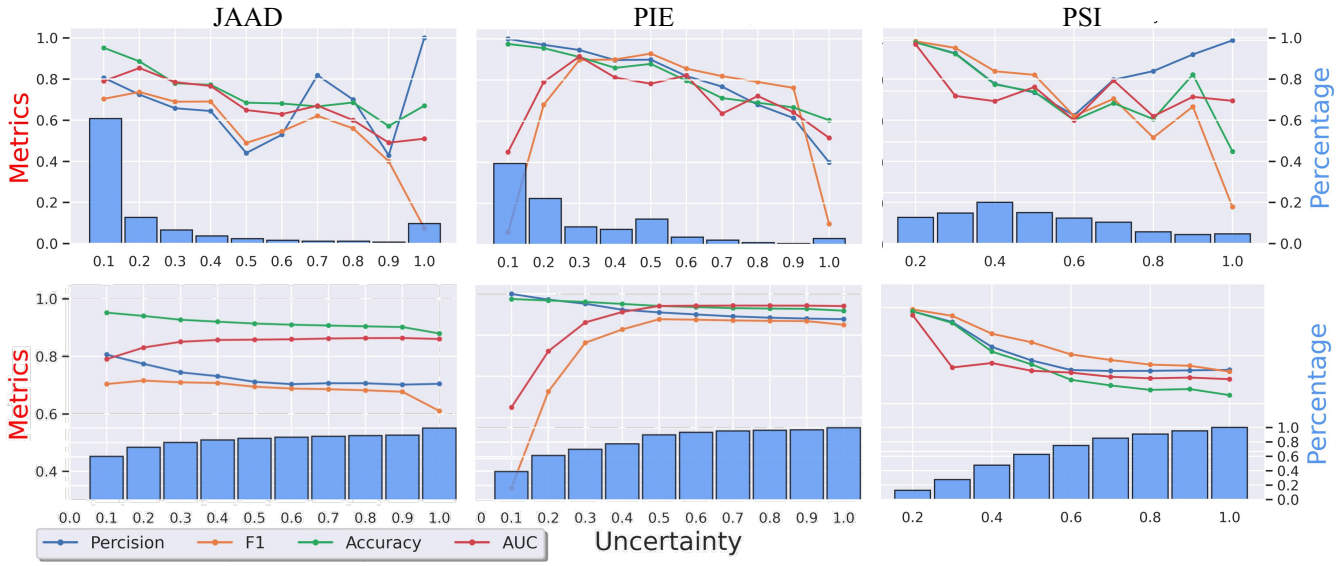

Figure 2: AI Uncertainty vs. Metrics on three datasets. The bars in the top three graphs are the proportion of the samples grouped by the uncertainty score (where 0.1 refers to the samples with uncertainty greater than 0 and less than 0.1). Each colored curve denotes the performance using a specific metric. The graphs at the bottom are the corresponding cumulative version.

value. For the bottom three graphs, the range is simply less than the corresponding value. We found that the uncertainty distributions in JAAD and PIE datasets have long right tails while the distribution in PSI is right-tailed.

As the uncertainty values increased for all three datasets, most metrics exhibited a decrease. In JAAD and PSI, the precision score reached 1 when the uncertainty was 1, as the model predicted all samples as "crossing," resulting in a low F1 score. In the case of the PIE dataset, it might seem like the model performs worse when the uncertainty is low, given the low F1 score and AUC on the left side. However, we observed that the accuracy was very high on the left side, and the low scores were due to the small number of "not crossing" samples. The cumulative graphs showed more stability and gradual decrease as the uncertainty increased. These findings support our hypothesis that "the models perform better on samples with lower uncertainties".

Disagreement Analysis

Since the PIE and PSI datasets provide the distribution of the annotators' decisions. For example, we know the number of annotators reporting the given pedestrian is crossing and vice versa. We use the **entropy** to measure disagreement among the annotators. When all annotators have the same intention estimation, the entropy is zero. On the other hand, entropy is one when the predictions are grouped into half and half. We present Fig. 3 using a similar fashion with Fig. 2 while the samples are grouped by the human disagreement scores. In addition, the green bars indicate the average uncertainty values for the corresponding groups.

Since both figures indicated a trend of decreasing performance with larger human disagreement scores, we concluded that our model performed worse on the human conflicting cases. We calculated the correlation coefficient between human disagreement and model uncertainty to testify whether the predicted uncertainties represent human disagreements. We found a weak negative correlation (correlation = -0.17, p-value < 0.001) and a strong positive correlation (correlation = 0.60, p-value < 0.001) for the PIE and PSI datasets, respectively. One possible explanation is that the intention segmentation (annotation in PSI) gives each frame a crowd-sourced label delivering more information to the model and allowing the model to capture patterns similar to the human. In contrast, PIE provides each pedestrian with one fixed label across the whole time span might supervise the model to ignore some discriminative features.

Case Study

In Fig. 4, we select some interesting cases from the test set of PIE datasets 4 to qualitatively analyze our proposed model. The qualitative analyses for the other two datasets are in the supplemental materials. The first two top figures are in the same scene, where a group of pedestrians were crossing in front of the ego vehicle to get on the bus. Though the ground-truth label from PIE was "not crossing" for both pedestrians, our model predicted "crossing". Admittedly, the trajectories of the pedestrians in the first two figures are very like the crossing case, and those could be considered as "crossing" in some sense. Moreover, these cases are very rare where the training set does not have similar cases to supervise the model's learning for this specific case. The third figure at the top and the first figure at the bottom are cases where the model successfully predicts the crossing intention with low uncertainty. The prior one is a typical situation for the "not crossing" case. However, the previous bounding boxes of the later case demonstrated large lateral movements. The model

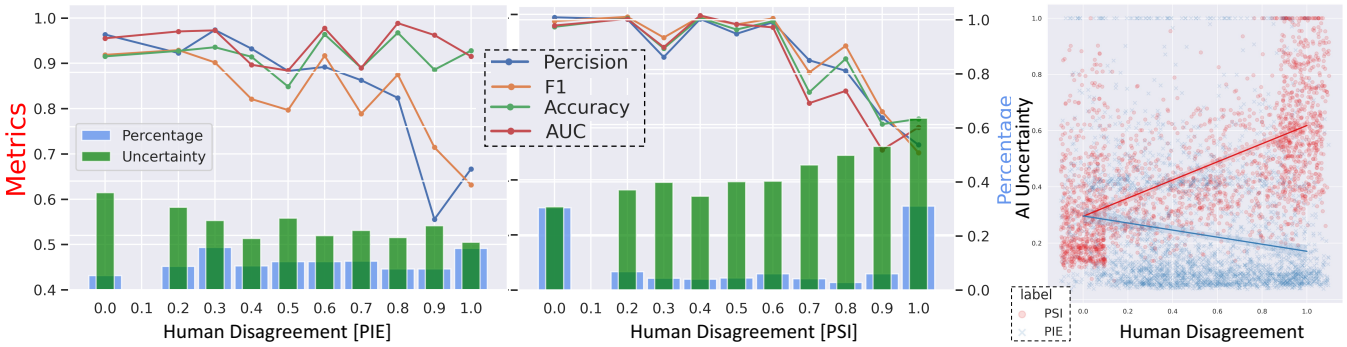

Figure 3: Human disagreement vs. Metrics on PIE and PSI datasets [Left two sub-figures], where blue bars are the proportion of the samples grouped by the uncertainty score (where 0.1 refers to the samples with uncertainty greater than 0 and less than 0.1) and green bars are the corresponding average uncertainty values. Each colored curve denotes the performance using a specific metric. The rightmost sub-figure shows the correlation between AI uncertainty and human disagreement.

Figure 4: Case study on PIE dataset, where U denotes the uncertainty, P means probability and I represents the decision (X is crossing, while the crossed X is not crossing. The red bounding boxes indicate the cases where our model predicts wrongly, while the green bounding boxes are the correct cases. The yellow bounding boxes are those in the previous 3 frames.

recognized the cause of the large lateral movement as from the ego vehicle's turning and predicted it correctly.

The last two figures in fig. 4 are also from the same scenario, where an adult takes a child to cross the street. Our model is extremely unsure about the crossing intention of the child (uncertainty = 1) and gave a useless prediction ("crossing" = "not crossing" = 0.5). It might be because of the deficiency of the samples of children because the model predicts the intention of the adult correctly with less uncertainty. Moreover, we examined the video clip and found that the pedestrians were negotiating with the ego vehicle where both sides did not carry a firm intention. However, the annotations in the PIE dataset do not show this intention dynamics. Overall, we believe the model performance is limited by the diversity of the training sample.

Conclusion

In this paper, we proposed a novel transformer-based evidential prediction (TrEP) algorithm for pedestrian intentions, aiming to capture the temporal correlation and model the AI uncertainty. We did comprehensive evaluations using three popular datasets for both existing and our proposed models. Our model outperformed all the existing works on all three datasets. Moreover, we utilized the crowd-sourced annotations in PIE and PSI to represent human disagreement and compared human disagreement with AI uncertainty. We found that our model shared the same uncertainty pattern with various human annotators provided in the PSI dataset.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant No.2145565.

References

Aliakbarian, M. S.; Saleh, F. S.; Salzmann, M.; Fernando, B.; Petersson, L.; and Andersson, L. 2018. VIENA2: A Driving Anticipation Dataset. In *Asian Conference on Computer Vision*, 449–466. Springer.

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33: 14927–14937.

Bao, W.; Yu, Q.; and Kong, Y. 2021. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13349–13358.

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE International Conference on Image Processing (ICIP)*, 3464–3468.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chen, T.; and Tian, R. 2021. A survey on deep-learning methods for pedestrian behavior prediction from the egocentric view. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1898–1905. IEEE.

Chen, T.; Tian, R.; Chen, Y.; Domeyer, J.; Toyoda, H.; Sherony, R.; Jing, T.; and Ding, Z. 2021. PSI: A Pedestrian Behavior Dataset for Socially Intelligent Autonomous Car. *arXiv preprint arXiv:2112.02604*.

Chen, T.; Tian, R.; and Ding, Z. 2021. Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3103–3109.

Cui, Y.; Cao, Z.; Xie, Y.; Jiang, X.; Tao, F.; Chen, Y. V.; Li, L.; and Liu, D. 2022. Dg-labeler and dgl-mots dataset: Boost the autonomous driving perception. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 58–67.

Ding, Y.; Harirchi, F.; Yong, S. Z.; Jacobsen, E.; and Ozay, N. 2018. Optimal input design for affine model discrimination with applications in intention-aware vehicles. In *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS)*, 297–307. IEEE.

Domeyer, J. E.; Lee, J. D.; and Toyoda, H. 2020. Vehicle automation–Other road user communication and coordination: Theory and mechanisms. *IEEE Access*, 8: 19860–19872.

Eriksson, A.; and Stanton, N. A. 2017. Takeover time in highly automated vehicles: noncritical transitions to and from manual control. *Human factors*, 59(4): 689–705.

Fang, Z.; and López, A. M. 2018. Is the pedestrian going to cross? answering by 2d pose estimation. In *2018 IEEE intelligent vehicles symposium (IV)*, 1271–1276. IEEE.

Gujjar, P.; and Vaughan, R. 2019. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, 2097–2103. IEEE.

Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; and Wang, Y. 2021. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34.

Herman, M.; Wagner, J.; Prabhakaran, V.; Möser, N.; Ziesche, H.; Ahmed, W.; Bürkle, L.; Kloppenburg, E.; and Gläser, C. 2021. Pedestrian Behavior Prediction for Automated Driving: Requirements, Metrics, and Relevant Features. *IEEE Transactions on Intelligent Transportation Systems*.

Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; and Zheng, Y. 2021. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12341–12351.

Jing, T.; Xia, H.; Tian, R.; Ding, H.; Luo, X.; Domeyer, J.; Sherony, R.; and Ding, Z. 2022. Inaction: Interpretable action decision making for autonomous driving. In *European Conference on Computer Vision*, 370–387. Springer.

Kotseruba, I.; Rasouli, A.; and Tsotsos, J. K. 2020. Do they want to cross? understanding pedestrian intention for behavior prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, 1688–1693. IEEE.

Kotseruba, I.; Rasouli, A.; and Tsotsos, J. K. 2021. Benchmark for evaluating pedestrian action prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1258–1268.

Litman, T. 2017. *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute Victoria, BC, Canada.

Liu, B.; Adeli, E.; Cao, Z.; Lee, K.-H.; Shenoi, A.; Gaidon, A.; and Niebles, J. C. 2020a. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2): 3485–3492.

Liu, D.; Cui, Y.; Chen, Y.; Zhang, J.; and Fan, B. 2020b. Video object detection for autonomous driving: Motion-aid feature calibration. *Neurocomputing*, 409: 1–11.

Liu, D.; Cui, Y.; Guo, X.; Ding, W.; Yang, B.; and Chen, Y. 2021. Visual localization for autonomous driving: Mapping the accurate location in the city maze. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 3170–3177. IEEE.

Liu, X.; Masoud, N.; Zhu, Q.; and Khojandi, A. 2022. A markov decision process framework to incorporate network-level data in motion planning for connected and automated vehicles. *Transportation Research Part C: Emerging Technologies*, 136: 103550.

Liu, X.; Zhao, G.; Masoud, N.; and Zhu, Q. 2020c. Trajectory planning for connected and automated vehicles: Cruising, lane changing, and platooning. *arXiv preprint arXiv:2001.08620*.

Ma, X.; Karimpour, A.; and Wu, Y.-J. 2020. Statistical evaluation of data requirement for ramp metering performance assessment. *Transportation Research Part A: Policy and Practice*, 141: 248–261.

Merat, N.; Jamson, A. H.; Lai, F. C.; Daly, M.; and Carsten, O. M. 2014. Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation research part F: traffic psychology and behaviour*, 27: 274–282.

Pang, Y.; Guo, Z.; and Zhuang, B. 2022. ProspectNet: Weighted Conditional Attention for Future Interaction Modeling in Behavior Prediction. *arXiv preprint arXiv:2208.13848*.

Qu, X.; Mei, Q.; Liu, P.; and Hickey, T. 2020. Using EEG to distinguish between writing and typing for the same cognitive task. In *Brain Function Assessment in Learning: Second International Conference, BFAL 2020, Heraklion, Crete, Greece, October 9–11, 2020, Proceedings 2*, 66–74. Springer.

Rasouli, A.; Kotseruba, I.; Kunic, T.; and Tsotsos, J. K. 2019. 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 (ICCV)*.

Rasouli, A.; Kotseruba, I.; and Tsotsos, J. K. 2017. 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*, 206–213.

Rasouli, A.; Kotseruba, I.; and Tsotsos, J. K. 2020. Pedestrian action anticipation using contextual feature fusion in stacked rnns. *arXiv preprint arXiv:2005.06582*.

Rasouli, A.; Rohani, M.; and Luo, J. 2021. Bifold and Semantic Reasoning for Pedestrian Behavior Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15600–15610.

Rasouli, A.; Yau, T.; Rohani, M.; and Luo, J. 2022. Multi-Modal Hybrid Architecture for Pedestrian Action Prediction. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, 91–97.

Sensoy, M.; Kaplan, L.; Cerutti, F.; and Saleki, M. 2020. Uncertainty-aware deep classifiers using generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5620–5627.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Zhou, M.; Niu, Z.; and Hua, G. 2021. SGCN: Sparse Graph Convolution Network for Pedestrian Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8994–9003.

Simonyan, K.; and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27.

Tang, Y.; Song, S.; Gui, S.; Chao, W.; Cheng, C.; and Qin, R. 2023. Active and Low-Cost Hyperspectral Imaging for the Spectral Analysis of a Low-Light Environment. *Sensors*, 23(3): 1437.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017b. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, C.; Wang, Y.; Xu, M.; and Crandall, D. J. 2022. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters*, 7(2): 2716–2723.

Wu, J.; Fang, H.; Shang, F.; Wang, Z.; Yang, D.; Zhou, W.; Yang, Y.; and Xu, Y. 2022. Learning self-calibrated optic disc and cup segmentation from multi-rater annotations. *arXiv preprint arXiv:2206.05092*.

Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; and Xu, Y. 2023. MedSegDiff-V2: Diffusion based Medical Image Segmentation with Transformer. *arXiv preprint arXiv:2301.11798*.

Xu, T.; Chen, W.; Pichao, W.; Wang, F.; Li, H.; and Jin, R. 2021. CDTrans: Cross-domain Transformer for Unsupervised Domain Adaptation. In *International Conference on Learning Representations*.

Xu, Y.; Piao, Z.; and Gao, S. 2018. Encoding Crowd Interaction With Deep Neural Network for Pedestrian Trajectory Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yagi, T.; Mangalam, K.; Yonetani, R.; and Sato, Y. 2018. Future Person Localization in First-Person Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, D.; Zhang, H.; Yurtsever, E.; Redmill, K.; and Ozguner, U. 2022. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*.

Yao, Y.; Atkins, E.; Johnson-Roberson, M.; Vasudevan, R.; and Du, X. 2021. BiTraP: Bi-Directional Pedestrian Trajectory Prediction With Multi-Modal Goal Estimation. *IEEE Robotics and Automation Letters*, 6(2): 1463–1470.

Yi, L.; and Qu, X. 2022. Attention-Based CNN Capturing EEG Recording's Average Voltage and Local Change. In *Artificial Intelligence in HCI: 3rd International Conference, AI-HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings*, 448–459. Springer.

Zeng, Z.; Zhao, W.; Qian, P.; Zhou, Y.; Zhao, Z.; Chen, C.; and Guan, C. 2021. Robust Traffic Prediction From Spatial–Temporal Data Based on Conditional Distribution Learning. *IEEE Transactions on Cybernetics*, 52(12): 13458–13471.

Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; and Zheng, N. 2019. SR-LSTM: State Refinement for LSTM Towards Pedestrian Trajectory Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, S.; Abdel-Aty, M.; Wu, Y.; and Zheng, O. 2021a. Pedestrian crossing intention prediction at red-light using pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 23(3): 2331–2339.

Zhang, Z.; Shen, D.; Tian, R.; Li, L.; Chen, Y.; Sturdevant, J.; and Cox, E. 2021b. Implementation and Performance Evaluation of In-vehicle Highway Back-of-Queue Alerting System Using the Driving Simulator. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 1753–1759.

Zhang, Z.; Tian, R.; and Duffy, V. G. 2023. *Trust in Automated Vehicle: A Meta-Analysis*, 221–234. Cham: Springer International Publishing. ISBN 978-3-031-10784-9.

Zhang, Z.; Tian, R.; Duffy, V. G.; and Li, L. 2022a. The Comfort of the Soft-Safety Driver Alerts: Measurements and Evaluation. *International Journal of Human–Computer Interaction*, 0(0): 1–11.

Zhang, Z.; Tian, R.; Elahi, F. M.; Luo, X.; Domeyer, J.; and Sherony, R. 2022b. Modeling Pedestrian Situated Intent in Dynamic Driving Scenes from the Driver's Perspective. *Available at SSRN 4281923*.

Zhang, Z.; Tian, R.; Sherony, R.; Domeyer, J.; and Ding, Z. 2022c. Attention-Based Interrelation Modeling for Explainable Automated Driving. *IEEE Transactions on Intelligent Vehicles*.