Decouple Ego-View Motions for Predicting Pedestrian Trajectory and Intention

Zhengming Zhang, *Student Member, IEEE*, Zhengming Ding, *Member, IEEE*, and Renran Tian, *Member, IEEE*

***Abstract*— Pedestrian trajectory prediction is a critical component of autonomous driving in urban environments, allowing vehicles to anticipate pedestrian movements and facilitate safer interactions. While egocentric-view-based algorithms can reduce the sensing and computation burdens of 3D scene reconstruction, accurately predicting pedestrian trajectories and interpreting their intentions from this perspective requires a better understanding of the coupled vehicle (camera) and pedestrian motions, which has not been adequately addressed by existing models. In this paper, we present a novel egocentric pedestrian trajectory prediction approach that uses a two-tower structure and multi-modal inputs. One tower, the vehicle module, receives only the initial pedestrian position and ego-vehicle actions and speed, while the other, the pedestrian module, receives additional prior pedestrian trajectory and visual features. Our proposed action-aware loss function allows the two-tower model to decompose pedestrian trajectory predictions into two parts, caused by ego-vehicle movement and pedestrian movement, respectively, even when only trained on combined ego-view motions. This decomposition increases model flexibility and provides a better estimation of pedestrian actions and intentions, enhancing overall performance. Experiments on three publicly available benchmark datasets show that our proposed model outperforms all existing algorithms in ego-view pedestrian trajectory prediction accuracy.**

***Index Terms*— Pedestrian trajectory prediction, scene understanding, automated driving, pedestrian intention.**

I. Introduction

AUTONOMOUS driving, also known as driverless or self-driving cars, is a rapidly growing field that has the potential to revolutionize transportation. Artificial Intelligence (AI) plays a critical role in enabling vehicles to navigate roads safely and make decisions without human input [1]. AI-based autonomous driving systems involve various subsystems like perception, control, communication, and others [2], all of which have seen significant advances [3]. In the area of perception systems, for example, computer vision AI techniques have been developed to allow autonomous vehicles to detect and identify other road users with improved performance [4], [5], [6], [7], [8], [9]. Accurate perception and understanding of the environment are critical foundations for safe and efficient driving decisions.

With the significant advancements in AI technologies in recent years, autonomous driving systems can now make more sophisticated decisions in complex and dynamic driving environments [10]. Machine learning and other AI techniques allow vehicles to learn from data and improve their performance over time, obtaining the capabilities of real-time decision-making, handling edge cases and unexpected situations, and ensuring safety. By learning from human driving behaviors, autonomous vehicles can analyze sensor data and make decisions about how to navigate the road appropriately, such as changing lanes, slowing down, or stopping [11], [12]. In order to deploy autonomous cars safely and reliably in natural road situations, the AI systems need to be robust enough to handle the uncertainty, unpredictability, and complexity of real-world scenarios [2].

Despite the advancements in autonomous vehicles, there remain many challenges that must be addressed. Interacting with vulnerable road users, such as pedestrians and bicyclists, is one of the current challenges facing the development of fully autonomous driving technology in urban settings [13], [14]. These road users are more susceptible to injury in accidents [15]. According to NHTSA (National Highway Traffic Safety Administration) 2021 data [16], fatalities and injuries of vulnerable road users (primarily human road users outside vehicles) kept increasing in the past decade, and reached the highest numbers in 40 years. There were 7,388 pedestrians killed and 60,577 injured in 2021, highlighting the need for better protection technologies. As the most common vulnerable road users, pedestrians can appear at different road locations during urban driving, move in dynamic ways with sudden changes, and may not always follow traffic laws. These characteristics make it challenging for autonomous vehicles to anticipate their behaviors and plan interaction strategies [17], [18]. As a result, autonomous vehicle road testing reports show that 80% to 90% of automation disengagements (failure of autonomous driving functions) occur in urban settings, especially with the presence of pedestrians [19], [20].

Thus, in the context of autonomous driving perception, the prediction of pedestrian trajectories is critical to ensuring the safety of both pedestrians and vehicle occupants. Accurate

Manuscript received 5 May 2023; revised 19 February 2024, 9 July 2024, and 30 July 2024; accepted 30 July 2024. Date of publication 26 August 2024; date of current version 30 August 2024. This work was supported by the National Science Foundation under Grant 2145565. The associate editor coordinating the review of this article and approving it for publication was Dr. Xiaolin Hu. *(Corresponding author: Renran Tian.)*

Zhengming Zhang is with the School of Industrial Engineering, Purdue University, West Lafayette, IN 47907 USA.

Zhengming Ding is with the Department of Computer Science, Tulane University, New Orleans, LA 70118 USA.

Renran Tian is with the Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695 USA (e-mail: rtian2@ncsu.edu).

Digital Object Identifier 10.1109/TIP.2024.3445734

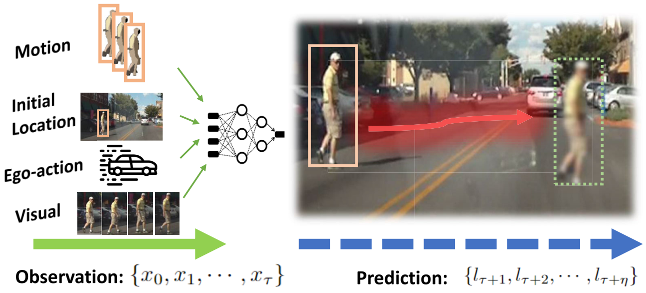
Motion
Initial Location
Ego-action
Visual
Observation: $\{x_0, x_1, \cdots, x_\tau\}$
Prediction: $\{l_{\tau+1}, l_{\tau+2}, \cdots, l_{\tau+\eta}\}$

Fig. 1. Illustration of pedestrian trajectory prediction from egocentric view, where the left-side is the multi-modality observation and the right-side is the trajectory prediction.

prediction of pedestrian trajectories can help autonomous vehicles make better driving decisions, improve the safety of human-driven vehicles by providing advanced driver warnings, and smooth the transition from automatic control to manual control by early detection of challenging situations [21].

Traditionally, bird's-eye view pedestrian trajectory prediction is a task that aims to predict the trajectory of a pedestrian from surveillance cameras, typically in an urban environment. This type of prediction is important for traffic management and urban planning [9], [22], [23], [24], [25], [26]. Ego-view pedestrian trajectory prediction, as shown in Figure 1, on the other hand, is a task that aims to predict the trajectory of a pedestrian from the perspective of a moving vehicle based on the observations of prior actions/behaviors of the pedestrian and the ego-vehicle, as well as other important scene features [27]. This type of prediction is crucial for safe navigation and motion planning in autonomous driving.

In contrast to the bird's-eye view trajectory, ego-view prediction needs to consider the movement of the ego-vehicle in addition to pedestrian movements, as the pedestrian position changes captured in the scene camera are affected by the two motions combined. When the car is stopped or moving at low speed, the captured pedestrian movements may be primarily caused by pedestrian motions; whereas when the car is running at high speed, the movement of the on-board camera may contribute greatly to the captured pedestrian position changes. The situation can become much more complicated when the car is turning, as small camera angle changes can cause significant pedestrian position changes in the captured view when there is a longer distance between the two. Current models in ego-view pedestrian trajectory prediction barely consider the contributions of the two separate motions explicitly, which limits the model's flexibility to fit different interaction situations with varying motion patterns.

Aside from jeopardized prediction accuracy, the intertwined motion prediction also makes it difficult to interpret pedestrian intentions and estimate their future moving directions, worsening the already-critical interpretability issues of adopting deep learning models in autonomous driving [28]. When the ego-view pedestrian trajectory prediction model only outputs pedestrian position changes in the captured camera view, such translations may be caused by different combinations of possible car and pedestrian motions at different distances. For example, a position change towards the center of the camera screen may be caused by the car turning or a change in distance instead of actual pedestrian crossing actions. Thus, it may be impossible or misleading to infer pedestrian intentions and future moving directions accurately.

Decoupling the pedestrian and car motions may be a viable solution to address the above-mentioned limitations for ego-view pedestrian trajectory prediction. Although 3D scene reconstruction-based or bird's-eye view-based trajectory prediction algorithms can achieve similar results [29], [30], [31], [32], these algorithms rely on significantly increased sensing requirements and higher computational capacity, leading to system complexity and reliability concerns.

In this paper, we present a novel approach for ego-view pedestrian trajectory prediction that utilizes a two-module structure. The first module, the vehicle module, models the trajectory prediction from the perspective of the ego-vehicle, while the second module, the pedestrian module, considers the perspective of the pedestrian. Incorporating our innovative loss function allows the model to decompose the trajectory prediction into two separate parts. The contributions of the proposed model are summarized in three points:

- First, we separate the contributions of vehicle motions and pedestrian motions when predicting the ego-view pedestrian trajectories, without requiring separate data inputs. This design increases the model's flexibility to fit different interaction situations where the two objects contribute differently to the final relative motions, without significantly increasing computational and system complexity.
- Second, the proposed model outperforms existing ego-view pedestrian trajectory prediction algorithms by a significant margin on three publicly available benchmark datasets, particularly with an improvement of over 15% on the JAAD dataset.
- Finally, this motion decomposition explains the trajectory prediction, giving insights into how the overall prediction is influenced by the ego-vehicle and pedestrian motions, and providing better inferences of pedestrian intentions and future moving directions.

II. Related Works

In the area of automated driving, pedestrian behavior prediction models focus on different outputs like trajectories, actions, and intentions. Trajectory prediction usually provides greater spatio-temporal granularity, which can directly support driving motion planning and control strategies. Action and intention predictions can provide longer-duration estimations of future pedestrian behaviors to guide trajectory prediction and support higher-level vehicle decisions and path planning. There are a large number of studies published on all of these tasks.

A. Pedestrian Intention/Action Prediction

Pedestrian intention prediction and action prediction are closely related, but they differ in terms of their level of detail. Intention prediction focuses on predicting the ultimate destination of a pedestrian, whereas action prediction is more detailed and aims to forecast the specific intermediate actions that a pedestrian will take. Some studies even use action labels as a proxy for pedestrian intentions [33], [34]. A key similarity between these two tasks is that both aim to understand and predict pedestrian behavior, which is crucial for improving decision-making in autonomous driving and intelligent transportation systems. The PIE [35], JAAD [36], and PSI [37] datasets are well-known benchmarks for pedestrian behavior prediction in the field of autonomous driving and intelligent transportation systems. Many existing models [38], [39], [40], [41], [42] have been evaluated on these datasets, making them standard benchmarks for comparing and evaluating the performance of different approaches.

Kotseruba et al. [40] demonstrated in their work that models incorporating trajectory information outperform video-based action recognition methods for pedestrian action prediction. Some studies [35], [39], [42] found that incorporating multiple modalities and learning multiple tasks can greatly improve performance in action prediction. Furthermore, Chen et al. [37] leveraged a graph convolutional neural network to incorporate pedestrian posture information and achieved improved performance on intention prediction. Additionally, Zhang et al. [41] employed a transformer-based evidential learning approach, achieving state-of-the-art results on both action and intention prediction, while also including uncertainty estimation in their predictions.

Although significant progress has been made for predicting pedestrian intentions and actions, such outputs are not sufficient to support vehicle motion planning and ensure safety without pedestrian trajectory information. Also, intention and action datasets require large amounts of labor-intensive manual annotations for algorithm training and evaluation, which are difficult to scale up. Another primary concern is the subjectivity involved in annotating intentions and actions, which can introduce inconsistencies and biases due to different interpretations by various annotators. Moreover, many of these datasets are derived from controlled environments or specific geographical locations, resulting in a lack of real-world diversity. This limitation is critical as it can adversely affect the generalizability of the models trained on such datasets.

B. Ego-View Pedestrian Trajectory Prediction

The use of multiple input modalities in ego-view pedestrian trajectory prediction has become a common practice, where information about the pedestrian's behavior and surrounding environment is simultaneously utilized [38], [39], [43]. Traditionally, Recurrent Neural Networks (RNNs) are commonly used to model pedestrian trajectories for the sequence-to-sequence task [35], [38], [44], [45]. A method proposed by Yao et al. [46] estimates pedestrian goals in both forward and backward directions, decoding multi-modal future trajectories. Another method [47] also utilizes a recurrent structure by setting stepwise goals to guide the trajectory prediction. To further improve the accuracy of trajectory prediction in egocentric view, we believe that it is important to separately process vehicle and pedestrian trajectories through a two-tower structure and an action-aware loss function, which will be introduced in detail in Chapter III.

Recent works have adopted transformer-based architectures for modality fusion, such as Yin et al. two-stage transformer approach, to improve the fusion of modalities [48]. Incorporating additional information, such as reachability priors, as in Makansi et al. work [43], or intention prediction [35], [49], has been proposed to further improve trajectory prediction. Multi-task learning has also been found to be effective, with many works incorporating final destination [38] or action/intention prediction as additional tasks [39].

Although previous methods have attempted to use transformer-based techniques for feature extraction and multi-modal fusion, none have applied transformer networks for sequential modeling. Additionally, existing pedestrian behavior prediction algorithms usually treat the observed trajectory as a single entity, neglecting that an egocentric trajectory comprises two distinct components: the pedestrian's movement and the movement of the ego-vehicle.

In contrast, our approach employs a transformer-based architecture for both feature extraction and fusion, and sequence-to-sequence modeling. More importantly, the proposed model recognizes the distinct components that make up an ego-view pedestrian trajectory. We introduce a two-module architecture with an innovative action-aware loss function that enables the model to separate the trajectory components, enhancing accuracy and interpretability.

III. The Proposed Method

A. Problem Formulation

Pedestrian trajectory prediction is typically framed as a supervised learning problem, where the aim is to learn a mapping function from past observations to the pedestrian's future trajectory. Notably, this problem is inherently temporal, so capturing temporal dependencies is a crucial factor for accurate predictions.

Formally, let $\mathcal{X} = \{x_0, x_1, \cdots, x_\tau\}$ be a sequence of observations of the pedestrian from the starting time to time τ. Each observation x_i is a quadruplet, composed of four modalities of information: visual information v_i, pedestrian initial location l_0, pedestrian motion m_i, and ego-vehicle action a_i. The visual information is the images captured by an egocentric camera. The pedestrian initial location l_0 is represented by a pixel-wise bounding box location at the first frame x_0 denoted as $\{(u_0^{TL}, v_0^{TL}), (u_0^{BR}, v_0^{BR})\}$, where TL and BR represent the top-left and bottom right point of the bounding box. The pedestrian motion m_i is computed by subtracting the pedestrian location l_i at time i from the initial pedestrian location l_0. Thus, the observation x_i can also be represented as a quadruplet $\{\mathbf{v}, l_0, \mathbf{m}, \mathbf{a}\}$, where $\mathbf{v}$ is the visual sequence, l is the initial pedestrian location, $\mathbf{m}$ is the motion sequence, and $\mathbf{a}$ is the ego-vehicle action sequence. The pedestrian and vehicle trajectories are mutually dependent, as the following Equation 1:

$$l_{i+1:}^{vehicle} \sim \pi(.|l_{:i}^{vehicle}, l_{:i}^{ped}), \tag{1}$$

where the future trajectory of vehicle $l_{i+1:}^{vehicle}$ follows a distribution $\pi(.)$ which is conditioned on the past vehicle trajectory $l_{:i}^{vehicle}$ and pedestrian trajectory $l_{:i}^{ped}$. The goal

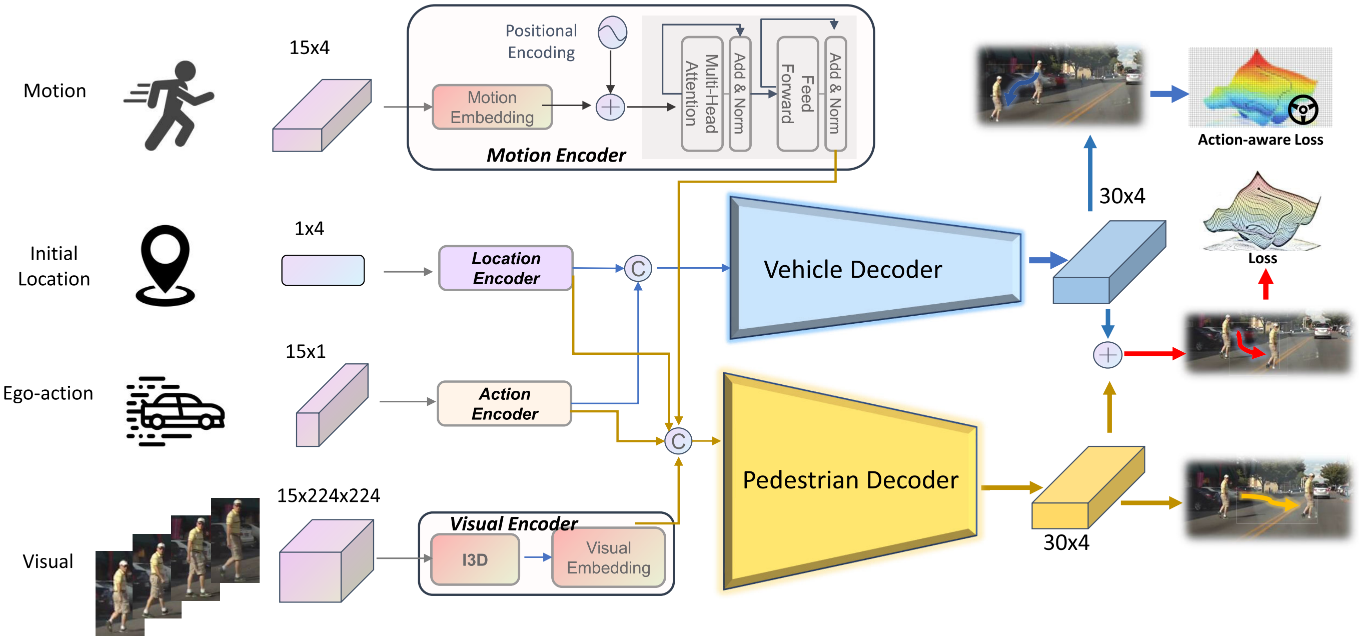

Fig. 2. The overall model architecture consists of four different types of input modalities: pedestrian bounding box motion, the initial bounding box location, ego-vehicle action, and video clips of the pedestrian. These modalities are independently encoded using a transformer encoder, two multi-layer perceptrons (MLPs), and a 3D convolutional backbone network. The location and ego-action embeddings are then concatenated and fed into the vehicle decoder, while all four embeddings are concatenated and fed into the pedestrian decoder. The vehicle-caused trajectories and overall trajectories are trained using an action-aware loss and a regular regression loss, respectively.

is to predict a sequence of egocentric pedestrian locations $\mathcal{L}_\eta = \{l_{\tau+1}, l_{\tau+2}, \cdots, l_{\tau+\eta}\}$ given the sequence of observations $\{x_0, x_1, \cdots, x_\tau\}$, where η is the prediction duration.

B. Framework Overview

Our proposed model predicts the pedestrian's future trajectory in a single pass. The overall architecture of the model is illustrated in Figure 2. Three sequential modalities of information are encoded using two different temporal-aware methods. A pre-trained 3D convolutional neural network is used to encode the visual information, while the motion and ego-vehicle actions are encoded using a transformer encoder. The initial location (bounding box coordinates of the starting point), as a static feature, is transformed through a feed-forward layer. Following feature encoding, the model adopts a two-tower structure, comprising a vehicle module and a pedestrian module, with a late fusion strategy.

Both the vehicle and pedestrian modules are constructed using feed-forward layers. The vehicle module utilizes the initial location and ego-vehicle action features to predict the trajectory changes caused by the vehicle's movement. In contrast, the pedestrian module is more complex and handles a wider range of information. It incorporates all four modalities of information, including visual features, initial location features, motion features, and ego-vehicle action features, to predict the trajectory caused by the pedestrian's movement, and to compensate for the prediction made by the vehicle module. The final prediction is the sum of the outputs of the vehicle and pedestrian modules.

C. Multi-Modal Feature Encoding

Our approach utilizes three distinct feature encoders designed to handle different modalities of features.

The first encoder is a pre-trained 3D convolutional network, which extracts pedestrian motion features and traffic context information. We leverage the effectiveness of 3D CNNs in action recognition tasks, where they have demonstrated the ability to capture both spatial and temporal information from video data [50], [51]. In our approach, the 3D CNN is pre-trained on a large dataset of video data to learn a representation of pedestrian motion and traffic context information. This is a crucial component of our method, as 3D CNNs are capable of learning spatio-temporal features by applying convolutional filters in both the spatial and temporal dimensions.

One of the advantages of using a 3D CNN is that it can effectively capture the temporal information of the pedestrian's motion without the need for time-consuming optical flow feature extraction. Optical flow is a technique used to estimate the movement of pixels between consecutive frames, which can be computationally expensive. By using a 3D CNN, we can bypass this step and still capture the temporal information of the pedestrian's motion.

Another advantage of using a 3D CNN is that it can learn spatiotemporal features that are invariant to viewpoint changes. The 3D CNN can learn features that are robust to changes in the camera perspective, which is crucial for the egocentric viewpoint of our approach. This makes our method more robust to occlusions and changes in viewpoint which are common in dash cam footage.

The second type of feature encoder we use in this work is a transformer encoder [52]. The transformer architecture has been a rising star in the field of sequence-to-sequence processing and has been shown to be superior to recurrent neural networks in many applications, particularly in natural language processing [53]. In our study, we apply the transformer encoder to capture the temporal relationship of location changes. The motion embedding represents the change of the pedestrian's bounding box, which is processed through a transformer encoder to capture the dynamic temporal characteristics of pedestrian behavior.

Transformer encoders consist of multi-head self-attention mechanisms, which allow the model to weigh different parts of the input sequence according to their importance. In our case, the transformer encoder uses self-attention to weigh the importance of different location and ego-vehicle action changes in the input sequence, allowing it to learn a more comprehensive representation of the temporal relationship.

Additionally, the transformer encoder is able to handle longer input sequences than traditional RNNs which makes it a great choice for capturing the temporal relationship of longer trajectory predictions. The transformer's architecture allows the model to process the entire sequence at once, which is beneficial for the task of trajectory prediction, where the whole sequence is needed to make the prediction.

In addition to the two temporal-aware feature encoders we previously discussed, we also utilize a feed-forward layer to transform the pedestrian's initial location feature and the ego vehicle's action into a representation that is more suitable for our model. This feed-forward layer is a simple yet effective method for converting the initial location feature, which is a static feature, into a form that can be effectively utilized by the model in its prediction. We also found that, even though the ego vehicle's action is sequential data, it is not as dynamic in the granularity of trajectory prediction. In other words, the ego vehicle's action such as speed does not change much within the time span of trajectory prediction, so we model it without using a temporal-aware feature encoder.

D. Two-Tower Structure

The two-tower structure is a neural network architecture consisting of two separate components, or "towers", that work together to achieve a specific task [54], [55]. In our case, the two-tower structure is used to predict pedestrian trajectories from an egocentric viewpoint using a dash cam. The first tower, the vehicle module, predicts the egocentric pedestrian trajectory caused by the ego vehicle's movement; and the second tower, the pedestrian module, predicts the trajectory caused by the pedestrian's movement. The two-tower structure allows for the separation of the prediction into two parts, which is a natural way of thinking about trajectory changes in an egocentric view and provides better explanations for the result.

To make each tower work in this natural way, we employed two techniques. The first technique is the use of different feature modalities as inputs. For the vehicle module, the ego-view trajectory caused by the vehicle only requires the ego-vehicle action and the pedestrian's initial location to be estimated. Since the camera is fixed on the ego-car, these features will predict the movements of the camera and the changes of camera angles, which result in corresponding egocentric view pedestrian position changes. The pedestrian module needs additional information such as pedestrian motion and visual features to generate a comprehensive scene representation and predict the trajectory changes caused by the pedestrian's movement.

Additionally, we devised an innovative loss function to ensure that the model decomposes the trajectory into two parts as intended. This is the second technique we used, and it will be discussed in the next subsection.

Our proposed model combines the outputs of the vehicle and pedestrian modules by point-wise summation to form the final prediction. The two-tower structure enables the use of specific combinations of features as input, allowing the model to make the most of the strengths of each feature modality and to follow a natural decomposition. This leads to a more accurate prediction of the pedestrian's trajectory.

E. Action-Aware Loss

To further ensure that each tower captures its responsibility for ego-vehicle-caused and pedestrian-caused trajectory components, we decompose the overall trajectory change at any time i in egocentric view as shown in the following Equation (2):

$$l_i = l_i^{vehicle} + l_i^{ped}, \tag{2}$$

where $l_i^{vehicle}$ represents the trajectory change induced by the movement of the ego-vehicle, l_i^{ped} denotes the trajectory change resulting from the actions of the target pedestrian, and l_i is the trajectory as observed from an egocentric viewpoint.

We propose an innovative loss function based on the assumption that vehicle-caused ego-view pedestrian movement dominates pedestrian-caused movement when the ego-vehicle is traveling at high speeds. As a result, the trajectory caused by the movements of pedestrians shrinks towards zero as the vehicle's speed increases with the following objective function:

$$\lim_{\text{speed}\to\sigma} \frac{l_i^{vehicle}}{l_i} = 1, \tag{3}$$

where σ is a large number. Thus, in addition to the commonly-used loss, we have added an auxiliary task to account for this observation. The overall objective function consists of two parts shown as:

$$\mathbf{loss} = \mathbf{RMSE}(\mathbf{l}_\eta, \hat{\mathbf{l}}_\eta) + \omega(\mathbf{a})\mathbf{RMSE}(\mathbf{l}_\eta, \hat{\mathbf{l}}_\eta^{Vehicle}), \tag{4}$$

where $\omega(\cdot)$ is a monotonically increasing function of the ego vehicle's speed, and $\mathbf{a}$ can be a variable related to the vehicle's speed, acceleration, or speed changing actions.

The first term in Equation (4) focuses on the final outputs, $\hat{\mathbf{l}}_\eta$, as the combination of the vehicle and pedestrian modules. Using a Root Mean Square Error loss (RMSE), shown in Equation (5), this commonly used loss minimizes the difference between the target and overall predicted trajectory:

$$\mathbf{RMSE}(\mathbf{l}_\eta, \hat{\mathbf{l}}_\eta) = \sqrt{\frac{1}{\eta}\sum_{i=\tau}^{\tau+\eta}(l_i - \hat{l}_i)^2}. \tag{5}$$

The second term of the loss function (Equation (4)) is an ego-action weighted Root Mean Square Error (Action-RMSE). This loss focuses on the differences between the output of the vehicle module, $\hat{\mathbf{l}}_{\eta}^{Vehicle}$, and the target ground-truth, $\mathbf{l}_{\eta}$, and is weighted by the ego vehicle's speed or accelerating/decelerating actions. The Action-RMSE penalizes the overall loss more as the vehicle speed increases, meaning that the vehicle-caused ego-view trajectory change dominates the pedestrian-caused changes. At lower vehicle speeds, the weighting function, $\omega(\cdot)$, will reduce the effects of the second term, so the overall model is optimized mainly based on the first term in Equation (4).

With a good selection of $\omega(\cdot)$, this loss function is designed to train the vehicle module to capture the pedestrian position changes caused by the vehicle's movement while optimizing the overall prediction. Moreover, as a consequence of the trade-off between the overall errors (which are based on a sum of outputs from the pedestrian module and the vehicle module) and the errors from the vehicle module, the pedestrian module is primarily trained to predict the contributions of pedestrian movements to the ego-view pedestrian position changes. Intuitively, the model is expected to perform well in both lower and higher vehicle speed situations as the two modules focus on separate contributions:

- When the vehicle speed is low, such as when the car is stopped, the vehicle module with ego-motion features as inputs does not have much information to predict pedestrian position changes. Consequently, the entire model relies primarily on the pedestrian module to predict pedestrian trajectories and prioritize minimizing the first term in Equation (4). Since the car's contribution is limited, the pedestrian module also mainly predicts actual pedestrian motions. In this scenario, the vehicle module plays a passive role in compensating for the portion of pedestrian trajectories that the pedestrian module cannot predict.
- When the vehicle speed is high, the vehicle module will be more penalized and will also have more information from the ego-action features to predict observed pedestrian trajectories. As the second term in Equation (4) gains more weight, the model will rely more on the vehicle module in this scenario, and the pedestrian module becomes more passive to compensate for the portion of pedestrian trajectories that cannot be predicted using only vehicle action data. We can assume that the pedestrian module is mainly predicting actual pedestrian movements in this scenario as well.

Note that although there is no direct loss function supervising the output of the pedestrian module, the rationale for this design choice stems from our model's framework. We base our approach on the understanding that the trajectory change of the pedestrian, as seen from the egocentric perspective, is influenced by the movements of both the ego vehicle and the pedestrian. Consequently, even in the absence of a direct loss function for pedestrian movement, the pedestrian output is indirectly shaped as intended. This is achieved as long as the two other loss functions effectively penalize both the combined trajectory changes and the component attributed solely to the vehicle. We have tried different functions for $\omega(\cdot)$ and will report the results in later sections. The loss function presented is for a single sample, and the overall objective function is obtained by summing up the loss over all samples.

TABLE I
SUMMARY OF PEDESTRIAN DATASETS

Dataset	Year	Frames	Number of Pedestrians
JAAD	2016	75K	2800
PIE	2019	293K	1800
PSI	2021	25K	100

IV. Experiment

A. Datasets

We conducted experiments to evaluate the effectiveness of our model using three widely-used benchmark datasets, namely JAAD, PIE, and PSI. All of these datasets are prepared for pedestrian trajectory prediction tasks from an egocentric view with captured dash cam video clips. JAAD and PSI comprise dash cam video clips, while PIE was collected from a 6-hour drive in downtown Toronto. While PIE and JAAD contain more pedestrians, PSI has more interactive cases. The number of unique pedestrians are shown in Table I.

To ensure a fair comparison with existing methods on JAAD and PIE datasets, we adopted the time-to-event setup in our evaluation as [38]. We clipped the pedestrian tracks up to the crossing event frames and sampled sequences with a 50% overlap and with time to event between 1 to 2 seconds (30 to 60 frames), as suggested in [8]. The prediction length was set at 45 frames ($\eta = 45$) given 16 frames of observation ($\tau = 16$). In the PIE dataset, we used 3,980 training sequences, 995 of which were crossing cases. On the other hand, the JAAD dataset had 3,955 training sequences, including 805 crossing cases. For the ego-vehicle action, the JAAD dataset provided the driver's behaviors, while the PIE dataset recorded the speed of the ego-vehicle. Therefore, the ego vehicle action sequence is defined as the vehicle's speed in km/h in the PIE dataset, while in the JAAD dataset, it refers to the driver's actions. We re-encoded the driver's behaviors in JAAD as 0-stop, 1-slow down, 2-maintain speed, and 3-speed up, capturing partial speed information while keeping it simple.

For the PSI dataset, we followed the original task setup since it had different annotations compared to JAAD and PIE. We sampled the clips with an overlap ratio of 0.8 across the entire video as long as the pedestrian appeared. The PSI dataset did not provide any ego-vehicle action annotations. The dataset contained 6,262 training sequences with 3,927 crossing cases.

B. Implementation

The proposed model architecture employs a powerful pre-trained Inflated 3D convolution (I3D) [51] on the Kinetics 400 dataset [56] for action recognition, as a means of converting the visual input into a tensor of 2048 × 4 × 8 × 8. This tensor is then fed through a two-layer convolutional neural network with batch normalization and a 3D adaptive average pool for further processing. The model also features

TABLE II
COMPARISON OF NUMBER OF PARAMETERS, AND INFERENCE TIME

Method	# of Paras. (in millions)	Inference Time (ms)
PIE_full	3.04	25.7
PSI	8.22	-
PEvT	0.26	6.7
MTN	0.13	3.9
Ours	3.61	27.8

a transformer encoder with a one-head attention layer and an embedding size of 16, as well as a location and action encoder with embedding sizes of 16 and 32, respectively.

The vehicle and pedestrian decoders, which are responsible for generating the predicted trajectories, are implemented as two-layer feed-forward networks. Additionally, the action-aware weight function is designed as a power function with a normalized vehicle speed, $\mathbf{v}$, and a hyperparameter $p = 1$, denoted as the formulation $\omega(\mathbf{v}) = \mathbf{v}^p$.

To train the model, the Adam optimizer is employed with a learning rate of 3e-3 for 1000 epochs and a batch size of 128. During the first 50 epochs, the pedestrian module is kept frozen and only the parameters along the action-aware loss are updated. This allows the model to initially learn the action-aware trajectory prediction without interfering with the pedestrian module's pre-trained weights. Overall, this architecture and training procedure provides an efficient and effective means of accurately predicting trajectories for complex action sequences.

C. Model Efficiency and Scalability

Excluding the fixed backbone of I3D, our model consists of approximately 4 million parameters, which is comparable to other state-of-the-art models in this domain (Table II). This parameter efficiency ensures that our model can be trained on standard hardware without excessive resource demands. We conducted experiments to measure the average inference time per frame on two NVIDIA RTX Titan GPUs. Our model processes each sequence in approximately 30 milliseconds, making it suitable for real-time applications in autonomous driving systems.

D. Evaluation Metrics

In our evaluation, we adopt two widely-used metrics, Average Displacement Error (ADE) and Final Displacement Error (FDE), to assess the model performance, consistent with prior work [37], [38]. ADE quantifies the average distance between the predicted and ground truth trajectories for all pedestrians over the entire prediction horizon, defined as Equation (6):

$$\text{ADE} = \frac{1}{N \times \eta} \sum_{i=1}^{N} \sum_{j=t+1}^{t+\eta} ||l_j^i - \hat{l}_j^i||_2, \tag{6}$$

where N is the total number of pedestrians, t is the observation time step, η is the prediction horizon, l_j^i is the ground truth location of pedestrian i at time step j, and $\hat{l}_j^i$ is the corresponding predicted location.

Similarly, FDE measures the L2 norm between the predicted and ground truth trajectories for all pedestrians at the final time step and is defined as Equation (7):

$$\text{FDE} = \frac{1}{N} \sum_{i=1}^{N} ||l_{t+\eta}^i - \hat{l}_{t+\eta}^i||_2. \tag{7}$$

Both ADE and FDE are computed based on the center coordinates of the bounding boxes. To further evaluate the accuracy of bounding box predictions, we report the Average Root Mean Squared Error (ARB) and Final Root Mean Squared Error (FRB) for bounding box coordinates. Specifically, ARB measures the average distance between the predicted and ground truth bounding box coordinates over the entire prediction horizon, while FRB only considers the final time step. Both ARB and FRB are reported in pixels, based on a prediction length of 30 frames (equivalent to 1 second), which is consistent with prior work.

E. Comparisons With State-of-the-Art Methods

We evaluated our method against eight existing approaches, including Future Person Localization (FPL) [27], Bayesian LSTM (B-LSTM) [57], FOL [58], and two variations of the methods in the PIE dataset [35], PEvT [44], multimodal transformer networks (MTN) [48], as well as the current state-of-the-art method, BiPeds [38]. To ensure a fair comparison, we used the same experimental settings as in BiPeds.

The experiment results are listed in Table III. In our experiments, our method outperformed all existing approaches on the PIE and JAAD datasets. On the PIE dataset, our method achieved an ADE of 17.41, which is approximately 1 lower than that of BiPeds. On the JAAD dataset, the improvement was more substantial, with ADE and FDE decreased by more than 3 and 7 respectively, representing an improvement of 15% and 20%. Therefore, our proposed method has achieved state-of-the-art performance on both PIE and JAAD datasets when compared to six existing methods.

In order to ensure a fair comparison with existing methods on the PSI dataset, we adopted the specific sampling method used in the dataset for evaluating our proposed model. Our model achieved a significant improvement in comparison to the existing PSI models (Table IV), with an ADE reduction of more than 25%. Overall, our evaluation results on the PIE, JAAD, and PSI datasets indicate that the proposed model is able to effectively predict ego-view pedestrian trajectories.

F. Ablation Study

In this section, we conducted an ablation study to comprehensively evaluate the impact of various modules and loss functions on the performance of our proposed model in Table V. In our model, both the vehicle and pedestrian modules are capable of independently producing outputs offering an alternative perspective where each module functions as a standalone predictive model. Our analysis revealed that the vehicle module alone performed poorly due to its limited information input. On the other hand, the pedestrian module, which utilizes a late fusion model with multiple modalities as

TABLE III

PERFORMANCE OF THE PROPOSED MODELS AND THE OTHER EXISTING MODELS ON THE JAAD AND PIE DATASETS

Dataset	PIE				JAAD			
Method\Metric	ADE ↓	FDE ↓	ARB ↓	FRB ↓	ADE ↓	FDE ↓	ARB ↓	FRB ↓
FOL	73.87	164.53	78.16	143.49	61.39	126.97	70.12	129.17
FPL	56.66	132.23	-	-	42.24	86.13	-	-
B-LSTM	27.09	66.74	37.41	75.87	28.36	70.22	39.14	79.66
PIE_traj	21.82	53.63	27.16	55.39	23.49	50.18	30.40	57.17
PIE_full	19.50	45.27	24.40	49.09	22.83	49.44	29.52	55.43
PEvT	19.15	45.98	25.14	50.43	21.08	49.08	29.02	55.10
MTN	18.89	45.50	24.77	49.74	20.90	48.55	28.54	54.19
BiPed	18.44	45.07	24.81	50.64	21.13	48.88	29.98	56.52
Ours	**17.41**	**44.92**	**23.16**	**49.18**	**17.92**	**41.33**	**25.74**	**52.84**

TABLE IV

PERFORMANCE OF THE PROPOSED MODELS AND THE OTHER EXISTING MODELS ON THE PSI DATASETS

Method\Metric	ADE ↓	FDE ↓	ARB ↓	FRB ↓
LSTM	39.87	66.56	43.07	69.34
PIE	35.39	61.50	37.45	63.40
PSI	31.07	52.03	35.03	55.08
Ours	**22.34**	**46.63**	**24.72**	**48.77**

TABLE V

PERFORMANCE FOR EACH VARIATION OF THE BASE MODEL ON THE PIE DATASET. "VEHICLE/PEDESTRIAN MODULE" REFERS TO MODELS USING A SINGLE MODULE. "RMSE" REFERS TO THE MODEL USING THE RMSE LOSS. "ACT-RMSE" REFERS TO MODELS WITH A COMBINATION OF RMSE AND ACTION-RMSE LOSS, WHERE P REFERS TO THE POWER OF THE $\omega(a)$. "FULL" REFERS TO THE COMPLETE MODEL

Models\Metric	ADE ↓	FDE ↓	ARB ↓	FRB ↓
Vehicle Module	47.82	79.14	49.01	75.62
Pedestrian Module	19.21	46.25	24.87	47.88
RMSE	18.19	45.38	23.96	50.17
Act-RMSE (p = 2)	17.56	**44.81**	23.41	49.50
Act-RMSE(p = $\frac{1}{2}$)	18.27	45.69	24.34	50.28
Full (p = 1)	**17.41**	44.92	**23.16**	**49.18**

input, performed significantly better than most existing methods, highlighting the importance of incorporating multi-modal information in behavior prediction.

Additionally, we explored three different loss function variations. Our findings revealed that simply applying the Root Mean Square Error (RMSE) to the final output was sufficient to slightly surpass the performance of the current state-of-the-art methods. This highlights the robustness and efficacy of our model's structure, even without resorting to specialized loss functions. In our experiments, we also tested two variations of a power function ($\omega(\mathbf{v}) = \mathbf{v}^p$), with a power of 2 (Act-RMSE (p = 2)) denoting a convex function and a power of $\frac{1}{2}$ (Act-RMSE (p = $\frac{1}{2}$)) denoting a concave function. These variations were designed to test different hypotheses regarding how speed influences the reduction in vehicle-induced movements. Our results indicated that treating the speed reduction as linear (power of 1) yielded the best performance, while the quadratic approach (power of 2) also showed promise, outperforming the linear assumption in terms of Final Displacement Error (FDE). Consequently, we adopted a power of 1 for our comprehensive model (Full (p = 1)).

Finally, the ablation study showed that the two-tower structure outperformed the pedestrian module alone by approximately 10% in ADE, proving the importance of incorporating both the vehicle and pedestrian modules.

TABLE VI

MODEL PERFORMANCE ACROSS DIFFERENT PREDICTION HORIZONS ON THE PIE DATASET. THE METRICS ARE EVALUATED FOR PREDICTION HORIZONS OF 1 SECOND, 2 SECONDS, AND 3 SECONDS

Horizon\Metric	ADE ↓	FDE ↓	ARB ↓	FRB ↓
1 second	5.11	12.26	5.60	14.37
2 seconds	11.58	26.98	13.01	31.45
3 seconds	17.41	44.92	23.16	49.18

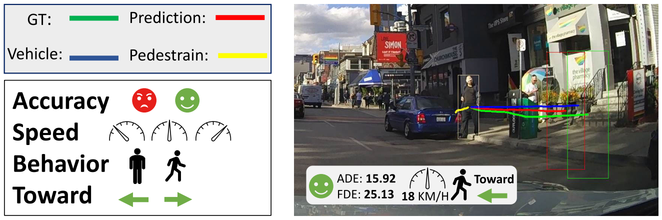

Fig. 3. One sample of demonstrations. The target pedestrian is bounded with a yellow bounding box, where the green line represents the ground truth trajectory, the red line represents the combined predicted trajectory, the blue line represents the predicted trajectory from the vehicle module, and the yellow line represents the prediction from the pedestrian module. The ground truth and predicted bounding boxes are depicted in green and red, respectively. Note that the images are cropped and magnified for better visualization. The icons on the top of each image indicate the displacement error (a smiling icon represents a low error, while a frowning icon represents a high error), the ego-vehicle speed, and the pedestrian's walking or standing status and their facing direction (red arrow indicates a predicted wrong direction).

G. Prediction Horizon

Table VI illustrates our model's performance on the PIE dataset across different prediction horizons (1, 2, and 3 seconds), evaluated using metrics such as ADE, FDE, ARB, and FRB. As the prediction horizon extends, there is an increase in all error metrics, reflecting the inherent uncertainties in

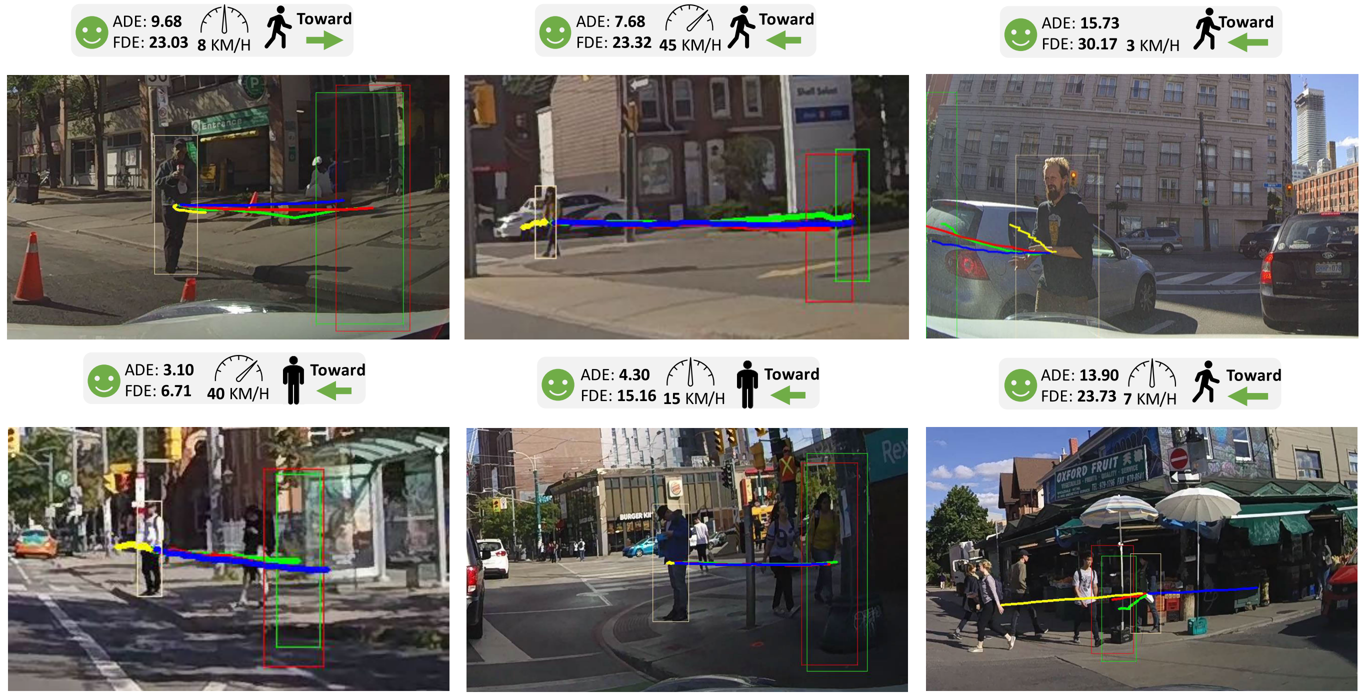

Fig. 4. The cases with correct decompositions on the PIE dataset. The target pedestrian is bounded with a yellow bounding box, where the green line represents the ground truth trajectory, the red line represents the combined predicted trajectory, the blue line represents the predicted trajectory from the vehicle module, and the yellow line represents the prediction from the pedestrian module. The ground truth and predicted bounding boxes are depicted in green and red.

long-term pedestrian trajectory prediction, which is consistent with the findings in related studies. Such challenges highlight the contributions of the proposed model in predicting pedestrian moving directions along with trajectories to better support driving decision-making ahead of time.

H. Case Study

In this section, our proposed model's performance was evaluated using representative examples following the format illustrated in Figure 3. The target pedestrian is enclosed within a yellow bounding box, while the ground truth trajectory is denoted by the green line, the combined predicted trajectory is denoted by the red line, the predicted trajectory from the vehicle module is depicted by the blue line, and the prediction from the pedestrian module is represented by the yellow line. The ground truth and predicted bounding boxes of the last prediction are shown in green and red, respectively. Please note that the images have been cropped and magnified for improved visualization. Additionally, the icons located at the top of each image indicate the displacement error (a smiling icon denotes a low error, while a frowning icon represents a high error), the speed of the ego-vehicle, and the pedestrian's status of either walking or standing, as well as their facial direction (a red arrow indicates a predicted incorrect direction).

As mentioned in the section on action-aware loss, one contribution of the proposed model is the ability to decompose the egocentric trajectory prediction into pedestrian-caused and vehicle-caused components. It is essential to assess whether the decomposed trajectories are logical or not. Therefore, we manually label the pedestrian's action (walking or standing) and facing direction (left or right) to demonstrate the validity of the predicted pedestrian-caused trajectory. We expect the predicted actual pedestrian movements (yellow lines in Figures 4 to 6) to align with the labeled action and facing direction. The yellow line should be shorter if the pedestrian is not walking, and the direction of the line should point towards the facing direction.

In the first demonstration (Figure 4), the model's accuracy in predicting the center and final bounding boxes is excellent. The majority of the predicted trajectories and bounding boxes overlap with the ground-truth, indicating a high degree of trajectory prediction accuracy. Additionally, the yellow line (predicted trajectory caused by actual pedestrian movements) has the same direction as the pedestrian facing direction, and the length aligns with the pedestrian actions as well.

While the combined predictions in Figure 5 are still accurate, the decomposition of the pedestrian and vehicle movements is not satisfactory when the yellow lines point toward the wrong facing directions. For all these cases, the pedestrian did not walk and the car was moving relatively fast. The lack of actual pedestrian movements creates challenges for the pedestrian module to accurately predict pedestrian trajectories. Although all the yellow lines in these cases are quite short, which align with the standing action, the results show that the direction of the yellow line may be more random and the model lacks the capability to detect pedestrian facing directions when the pedestrian does not move at all.

The last demonstrations (Figure 6) show some poorer outputs from the model, with higher values of ADE and FDE. These errors highlight the complexities and difficulties of pedestrian trajectory prediction and will be further investigated in our future research.

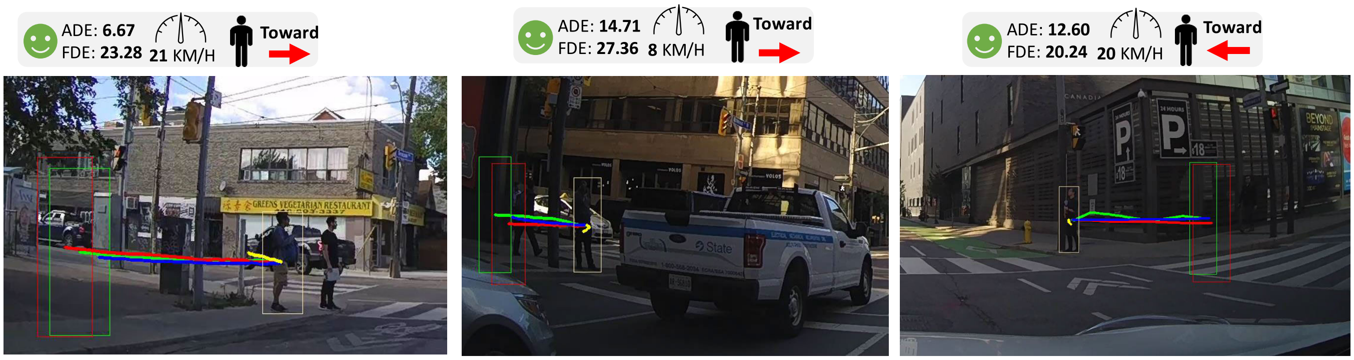

Fig. 5. The cases with incorrect decompositions on the PIE dataset. The target pedestrian is bounded with a yellow box, where the green line represents the ground truth trajectory, the red line represents the combined predicted trajectory, the blue line represents the predicted trajectory from the vehicle module, and the yellow line represents the prediction from the pedestrian module. The ground truth and predicted bounding boxes are depicted in green and red.

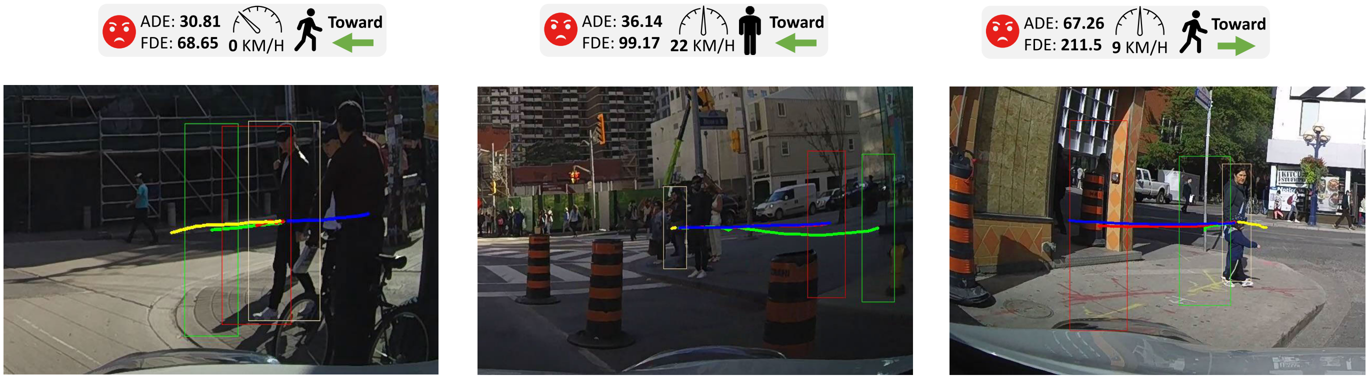

Fig. 6. Inaccurate predictions on the PIE dataset. The target pedestrian is bounded with a yellow box, where the green line represents the ground truth trajectory, the red line represents the combined predicted trajectory, the blue line represents the predicted trajectory from the vehicle module, and the yellow line represents the prediction from the pedestrian module. The ground truth and predicted bounding boxes are depicted in green and red.

The proposed model is demonstrated to have state-of-the-art performance in predicting pedestrian trajectories from an egocentric view and can interpret the predicted trajectories by separating the contributions from vehicle and pedestrian movements, as shown in the case study. To the best of our knowledge, the proposed model is the first to decouple the ego-view pedestrian motions without requiring additional inputs and labels in the area of pedestrian behavior prediction and autonomous driving. Such capability can aid in estimating actual pedestrian actions and predicting their intentions to cross the street.

V. Conclusion

In conclusion, this paper proposes a novel approach to pedestrian trajectory prediction from an egocentric viewpoint using a two-tower structural multi-modality model. The proposed action-aware loss function allows the model to decompose trajectory prediction into two parts, improving performance and providing an explanation for prediction results based on the contributions of pedestrian and vehicle. This explanation helps the model to estimate actual pedestrian actions and facing directions without the need for additional inputs or 3D scene reconstruction computations.

The results on three publicly available benchmark datasets demonstrate that our model outperforms existing algorithms in trajectory prediction and is the first to decouple predicted ego-view pedestrian trajectories. Overall, this work contributes to the development of safer autonomous driving systems by improving the accuracy and interpretability of pedestrian trajectory prediction. Our future work will focus on ensuring the safety of implementing pedestrian behavior prediction algorithms in driving motion planning by monitoring algorithm uncertainties and better handling tail cases.

References

[1] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4316–4336, Jul. 2021.

[2] P. Hang, C. Lv, C. Huang, J. Cai, Z. Hu, and Y. Xing, "An integrated framework of decision making and motion planning for autonomous vehicles considering social behaviors," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14458–14469, Dec. 2020.

[3] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[4] B. Wu, A. Wan, F. Iandola, P. H. Jin, and K. Keutzer, "SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2017, pp. 129–137.

[5] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2446–2454.

[6] X. Du and K. K. Tan, "Comprehensive and practical vision system for self-driving vehicle lane-level localization," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2075–2088, May 2016.

[7] Y. Cai, L. Dai, H. Wang, and Z. Li, "Multi-target pan-class intrinsic relevance driven model for improving semantic segmentation in autonomous driving," *IEEE Trans. Image Process.*, vol. 30, pp. 9069–9084, 2021.

[8] C. Zhu and Y. Peng, "A boosted multi-task model for pedestrian detection with occlusion handling," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5619–5629, Dec. 2015.

[9] Y. Li, X.-Y. Lu, J. Wang, and K. Li, "Pedestrian trajectory prediction combining probabilistic reasoning and sequence learning," *IEEE Trans. Intell. Vehicles*, vol. 5, no. 3, pp. 461–474, Sep. 2020.

[10] S. Aradi, "Survey of deep reinforcement learning for motion planning of autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 740–759, Feb. 2022.

[11] Z. Zhang, R. Tian, R. Sherony, J. Domeyer, and Z. Ding, "Attention-based interrelation modeling for explainable automated driving," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 2, pp. 1564–1573, Feb. 2023.

[12] T. Jing et al., "InAction: Interpretable action decision making for autonomous driving," in *Computer Vision—ECCV 2022* (Lecture Notes in Computer Science), vol. 13698, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-031-19839-7_22.

[13] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 900–918, Mar. 2020.

[14] K. Wang et al., "The adaptability and challenges of autonomous vehicles to pedestrians in urban China," *Accident Anal. Prevention*, vol. 145, Sep. 2020, Art. no. 105692.

[15] A. Constant and E. Lagarde, "Protecting vulnerable road users from injury," *PLoS Med.*, vol. 7, no. 3, Mar. 2010, Art. no. e1000228.

[16] T. Stewart, "Overview of motor vehicle traffic crashes in 2021," Nat. Highway Traffic Saf. Admin., Washington, DC, USA, Tech. Rep. DOT HS 813 435, 2023. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813435

[17] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Understanding pedestrian behavior in complex traffic scenes," *IEEE Trans. Intell. Vehicles*, vol. 3, no. 1, pp. 61–70, Mar. 2018.

[18] *Pedestrian Safety: A Road Safety Manual for Decision-Makers and Practitioners*, World Health Organization, Geneva, Switzerland, 2013.

[19] A. M. Boggs, B. Wali, and A. J. Khattak, "Exploratory analysis of automated vehicle crashes in california: A text analytics & hierarchical Bayesian heterogeneity-based approach," *Accident Anal. Prevention*, vol. 135, Feb. 2020, Art. no. 105354.

[20] S. Dadvar and M. M. Ahmed, "California autonomous vehicle crashes: Explanatory data analysis and classification tree," Transp. Res. Board, Washington, DC, USA, Tech. Rep. TRBAM-21-04068, 2021. [Online]. Available: https://annualmeeting.mytrb.org/OnlineProgram/Details/15795

[21] G. Wiegand, M. Eiband, M. Haubelt, and H. Hussmann, "'I'd like an explanation for that!' Exploring reactions to unexpected autonomous driving," in *Proc. 22nd Int. Conf. Human-Comput. Interact. Mobile Devices Services*, 2020, pp. 1–11.

[22] P. Dendorfer, S. Elflein, and L. Leal-Taixé, "MG-GAN: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13158–13167.

[23] N. Shafiee, T. Padir, and E. Elhamifar, "Introvert: Human trajectory prediction via conditional 3D attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16815–16825.

[24] C. Choi and B. Dariush, "Looking to relations for future trajectory forecast," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 921–930.

[25] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[26] X. Song et al., "Pedestrian trajectory prediction based on deep convolutional LSTM network," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3285–3302, Jun. 2021.

[27] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7593–7602.

[28] R. Häuslschmid, M. von Bülow, B. Pfleging, and A. Butz, "SupportingTrust in autonomous driving," in *Proc. 22nd Int. Conf. Intell. User Interfaces*, Mar. 2017, pp. 319–329.

[29] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," 2022, *arXiv:2207.05844*.

[30] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3D object detection via color-embedded 3D reconstruction for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6851–6860.

[31] T. Roddick, B. Biggs, D. O. Reino, and R. Cipolla, "On the road to large-scale 3D monocular scene reconstruction using deep implicit functions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2875–2884.

[32] B. Varadarajan et al., "MultiPath++: Efficient information fusion and trajectory aggregation for behavior prediction," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 7814–7821.

[33] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Pedestrian crossing intention prediction at red-light using pose estimation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2331–2339, Mar. 2022.

[34] T. Fu, L. Miranda-Moreno, and N. Saunier, "A novel framework to evaluate pedestrian safety at non-signalized locations," *Accident Anal. Prevention*, vol. 111, pp. 23–33, Feb. 2018.

[35] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6262–6271.

[36] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 206–213.

[37] T. Chen et al., "PSI: A pedestrian behavior dataset for socially intelligent autonomous car," 2021, *arXiv:2112.02604*.

[38] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15580–15590.

[39] A. Rasouli, T. Yau, M. Rohani, and J. Luo, "Multi-modal hybrid architecture for pedestrian action prediction," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 91–97.

[40] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1257–1267.

[41] Z. Zhang, R. Tian, and Z. Ding, "TrEP: Transformer-based evidential prediction for pedestrian intention with uncertainty," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2023, pp. 3534–3542.

[42] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "Coupling intent and action for pedestrian crossing behavior prediction," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1238–1244, doi: 10.24963/ijcai.2021/171.

[43] O. Makansi, Ö. Çiçek, K. Buchicchio, and T. Brox, "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4354–4363.

[44] L. Neumann and A. Vedaldi, "Pedestrian and ego-vehicle trajectory prediction from monocular camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10199–10207.

[45] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic LSTM for pedestrian trajectory prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 3229–3239, 2021.

[46] Y. Yao, E. Atkins, M. Johnson-Roberson, R. Vasudevan, and X. Du, "BiTraP: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1463–1470, Apr. 2021.

[47] C. Wang, Y. Wang, M. Xu, and D. J. Crandall, "Stepwise goal-driven networks for trajectory prediction," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2716–2723, Apr. 2022.

[48] Z. Yin, R. Liu, Z. Xiong, and Z. Yuan, "Multimodal transformer networks for pedestrian trajectory prediction," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1259–1265.

[49] H. Girase et al., "LOKI: Long term and key intentions for trajectory prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9783–9792.

[50] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2015, pp. 4489–4497.

[51] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[52] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[53] T. Brown et al., “Language models are few-shot learners,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[54] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[55] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Computer Vision—ECCV 2016* (Lecture Notes in Computer Science), vol. 9912, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham, Switzerland: Springer, 2016, pp. 483–499, doi: 10.1007/978-3-319-46484-8_29.

[56] W. Kay et al., “The kinetics human action video dataset,” 2017, *arXiv:1705.06950*.

[57] A. Bhattacharyya, M. Fritz, and B. Schiele, “Long-term on-board prediction of people in traffic scenes under uncertainty,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4194–4202.

[58] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, “Egocentric vision-based future vehicle localization for intelligent driving assistance systems,” in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9711–9717.

Zhengming Zhang (Student Member, IEEE) received the M.S. degree in statistics from the University of California at San Diego (UCSD). He is currently pursuing the Ph.D. degree in industrial engineering with Purdue University, West Lafayette, IN, USA. His research interests include human–computer interaction, human factors, and deep learning for intelligent transportation systems.

Zhengming Ding (Member, IEEE) received the B.Eng. degree in information security and the M.Eng. degree in computer software and theory from the University of Electronic Science and Technology of China (UESTC), China, in 2010 and 2013, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Northeastern University, USA, in 2018. He has been a Faculty Member with the Department of Computer Science, Tulane University, since 2021. His research interests include transfer learning, multi-view learning, and deep learning. He received the National Institute of Justice Fellowship from 2016 to 2018. He was a recipient of the Best Paper Award (SPIE 2016), the Best Paper Candidate (ACM MM 2017), and the Best Paper Finalist (CVPR 2022). He is currently an Associate Editor of IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology, and the *Journal of Electronic Imaging* (JEI).

Renran Tian (Member, IEEE) received the B.S. and M.S. degrees from Tsinghua University, China, in 2002 and 2005, respectively, and the Ph.D. degree in industrial engineering from Purdue University in 2013. He is currently an Assistant Professor of industrial and systems engineering with North Carolina State University, Raleigh, NC, USA. His research interests include human-centered computing, human–AI teaming, artificial intelligence, cognitive psychology, and autonomous driving. He was a recipient of the NSF CAREER Award from the Human-Centered Computing Program in 2022.