

PedFormer: Pedestrian Behavior Prediction via Cross-Modal Attention Modulation and Gated Multitask Learning

Amir Rasouli^{*1} and Iuliia Kotseruba^{‡2}

Abstract—Predicting pedestrian behavior is a crucial task for intelligent driving systems. Accurate predictions require a deep understanding of various contextual elements that could impact the way pedestrians behave. To address this challenge, we propose a novel framework that relies on different data modalities to predict future trajectories and crossing actions of pedestrians from an egocentric perspective. Specifically, our model utilizes a cross-modal Transformer architecture to capture dependencies between different data types. The output of the Transformer is augmented with representations of interactions between pedestrians and other traffic agents conditioned on the pedestrian and ego-vehicle dynamics that are generated via a semantic attentive interaction module. Lastly, the context encodings are fed into a multi-stream decoder framework using a gated-shared network. We evaluate our algorithm on public pedestrian behavior benchmarks, PIE and JAAD, and show that our model improves state-of-the-art in trajectory and action prediction by up to 22% and 13% respectively on various metrics. The advantages of the proposed components are investigated via extensive ablation studies.

I. INTRODUCTION

Pedestrian behavior prediction is a fundamental task for intelligent driving in urban environments. What makes it challenging is that pedestrians' actions are influenced by the interplay of various environmental factors, such as social interactions and traffic dynamics [1], [2]. Past approaches investigated two different ways of predicting pedestrian behavior: future trajectories [3], [4] and forthcoming actions, e.g. crossing the road [5], [6]. However, recent studies [7], [8] showed that these modes of prediction are complementary.

In this paper, we exploit the relationship between trajectory and action prediction in a multitask learning framework. Our approach relies on a cross-modal attention modulation technique for encoding pedestrians' and scene dynamics using multimodal data. This encoding is augmented with representations of interactions between pedestrians and their surroundings captured by an attentive interaction module that uses semantic scene information as well as pedestrians' and the ego-vehicle dynamics. The combination of interaction representation and modality encodings is fed into a hybrid-gated multi-stream decoder to predict future trajectories and crossing actions. Evaluation on pedestrian behavior benchmark datasets, PIE [4] and JAAD [9], demonstrates that the proposed method significantly improves state-of-the-art performance on both tasks.

^{*}Corresponding author amir.rasouli@huawei.com

[‡] Work done during internship at Huawei, Canada

¹Noah's Ark Laboratory, Huawei, Canada.

²Department of Computer Science and Electrical Engineering, York University, Toronto, Canada

II. RELATED WORKS

A. Pedestrian Behavior Prediction

Human behavior prediction has many applications in different domains [10], [11], [12], [13], [14], [15], [16]. In traffic scene understanding, behavior prediction either takes the form of trajectory or crossing action prediction.

1) *Trajectory Prediction*: Pedestrian trajectory prediction algorithms take two forms. The dominant approach is making predictions from bird's eye view surveillance-like videos recorded with a fixed camera [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27].

Another group of algorithms makes predictions from images captured with a moving monocular camera [28], [29], [30], [4], [31], [32], [33], [34], [35]. In this context, lack of depth and global positions of agents and camera ego-motion make predictions quite challenging. To overcome these difficulties, egocentric models rely on multimodal sources of data to reason about the future trajectories of pedestrians. For instance, the authors of [34] employ a two-stream encoder-decoder architecture. One stream predicts ego-vehicle motion by relying on observed ego-motion and visual context, and the second stream uses encoded observed pedestrian trajectories and predicted ego-motion to forecast future trajectories. Other multi-stream models use pedestrian action, e.g. waiting to cross [30], or intention of performing an action, e.g. crossing the road [4]. In [31], the authors use pedestrian poses in addition to trajectories and ego-motion. Their model has three streams of 1D convolutional layers that encode different data modalities. The outputs of the encoders are combined and fed into a convolutional decoder to generate final predictions.

2) *Action Prediction*: In the driving domain, predicting crossing actions of pedestrians is helpful for assessing risk and anticipating pedestrian trajectories for safe motion planning [36], [37], [38], [39], [9], [5], [6], [40]. To achieve this goal, it is common to use streams of images. In some approaches [39], [37], the images are used in a generative architecture to produce future scenes, which in turn are classified as containing crossing and non-crossing events. Alternatively, in [38] the authors first localize pedestrians and then use a 3D convolutional architecture to classify their actions in the future frames.

Similar to egocentric trajectory prediction algorithms, many crossing action prediction models use multimodal data [5], [6], [40], [9]. For example, the authors of [6] use visual appearance of pedestrians and their surroundings, their locations, poses, and ego-vehicle speed in a hierarchical

architecture. The model in [40] relies on visual features of the scenes, optical flow, and ego-vehicle dynamics each of which are processed with individual recurrent networks, combined and fed to another recurrent network for final prediction. The approach of [36] uses a hybrid architecture in which visual features are encoded using 3D convolutional layers in addition to other modalities, such as poses, trajectories, and ego-speed, that are encoded using recurrent networks. The encodings are combined using self-attention modules before the final prediction.

3) *Multitask Prediction*: Multitask learning has a wide range of applications in various computer vision domains from behavior understanding [41], [42], [43], [44] and object recognition [45], [46], [47], [48] to intelligent driving [7], [49], [50], [51], [52], [53]. These works emphasize the effectiveness of multitask learning for improving performance.

Multitask learning is also widely used in the context of pedestrian behavior prediction [54], [55], [8], [56]. The model in [54] predicts trajectories and head poses of pedestrians simultaneously and shows that correlation between the tasks can result in improved trajectory forecasting. The method in [8] predicts future activities of pedestrians, e.g. interacting with a car, as an auxiliary task to improve trajectory prediction using a recurrent framework.

In an egocentric setting, the method in [7] simultaneously predicts pedestrian crossing actions and trajectories and shows significant improvements on both tasks due to their complementary roles. The proposed bifold approach decodes contextual information using individual and shared decoders whose outputs are averaged to make final prediction. This approach, however, can be computationally inefficient as two sets of predictions take place. In addition, using separate decoders can negatively impact the model's ability to capture influence of prediction tasks on one another.

The method we propose is a multitask learning framework similar to [7] for predicting future pedestrian trajectories and crossing actions from an egocentric perspective. Differently from the past work, we propose a unified decoding approach that uses a gated flow mechanism between shared and individual task decoders. The predictions are made once using the outputs of the task decoders.

B. Multimodal Data

To compensate for the lack of global position information, many egocentric behavior prediction methods rely on multimodal data. The common approach is to process different modalities individually and then collapse them into a single representation by addition or concatenation [4], [8] or using attention to reweigh them prior to merging [36]. Alternatively, in [6], representation are organized hierarchically according to their complexity level (visual at the bottom and ego-speed at the top) and merged gradually, whereas in [57], modality representations are projected into a shared latent space prior to prediction.

Independently encoding different modality representations makes them more susceptible to noise and makes capturing step-wise correlation between them difficult. The model in

[7] addresses these issues with a bifold mechanism which consists of individual encoders and a shared encoder that receives an embedded combination of all input data. The outputs of all encoders are then combined for the final prediction. Although effective, this approach cannot capture the mutual influences between data modalities and does not have any mechanisms to control the weights of different data types in shared representations. To this end, we propose an attention mechanism in which different data modalities are combined pair-wise via multi-head attention units the outputs of which are combined and fed into a Transformer module to produce the final input encoding.

C. Pedestrian Interaction with Environment

A key challenge in multi-agent behavior prediction task is modeling interactions between agents and their surroundings. A common approach for pedestrian behavior prediction is to use pedestrians' positions and determine their effect on each other in the future. This can be done either implicitly by co-processing trajectories within a neighboring region using methods, such as social pooling [27], [22], [23] or more explicitly by determining the importance of interacting trajectories using attention-based [26], [25], [58] or graph architectures [59], [60], [20], [21], [61], [62].

Although position-based techniques have been used in egocentric context [30], [35], their utility is limited if global positions of agents are not available. In the egocentric view, image coordinates of agents can be easily confused. For example, two pedestrians with different heights that are moving next to one another and two pedestrians of the same height but moving at different distances from the ego-vehicle can yield similar trajectories. To this end, the authors of [5] use visual encodings of the agents in addition to their pair-wise spatial locations in the form of union of the bounding boxes surrounding them. The method in [7] removes the dependency on spatial information by relying only on semantic parsing of the scenes. Here, the semantic map of the scene is grouped by object categories, which are processed individually and combined via an attention module. Both approaches have several issues in common. First is the scalability, as the former method requires pair-wise reasoning for all agents and the latter involves visual processing for each pedestrian and surrounding people. Second, neither approach accounts for ego-vehicle motion when reasoning about the interactions. Finally, use of convolutional networks for processing scene information can introduce inductive biases such as location invariance which is not desirable when reasoning about agents' positions. We address these issues by proposing a novel method that is based on visual attention networks in which semantic information about the scene is divided into channels containing objects with shared characteristics. The semantic map is then divided into local patches, which after a series of embedding operations are fed into a multi-head attention network followed by a global attention module that produces a summary given pedestrian and ego-vehicle dynamics.

Contributions: 1) We propose a novel pedestrian behavior

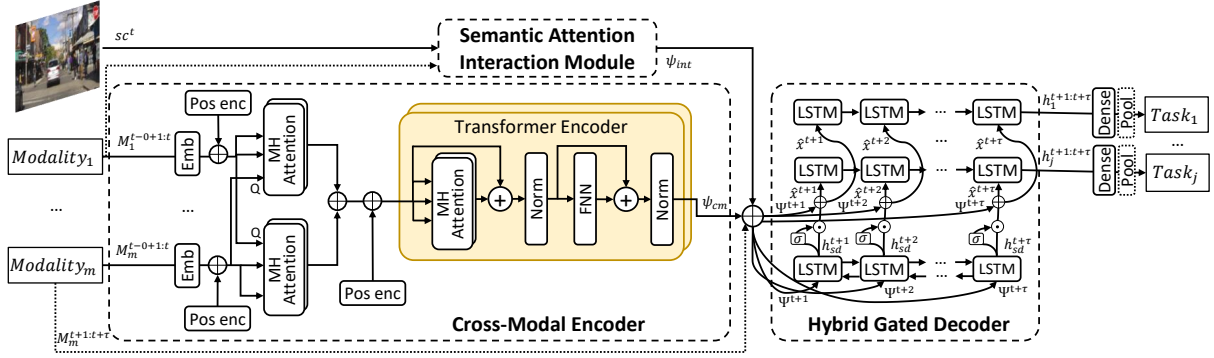


Fig. 1: Proposed architecture comprised of three modules: 1) A *cross-modal encoder* which uses multi-head attention to capture pair-wise dependencies between data modalities, 2) *semantic attention interaction module* (SAIM) that encodes interactions between ego-vehicle, environment, and other agents, and 3) a *hybrid gated decoder* which produces predictions from a combination of modality encodings and interaction representation. \oplus denotes concatenation operation.

prediction model that simultaneously predicts pedestrian trajectory and crossing action in egocentric video sequences. Our model employs a cross-modal attention-based encoding architecture for processing multimodal data, an interaction modeling module that uses visual attention architecture to encode interactions between pedestrians and their surroundings, and a hybrid gated decoder mechanism for processing context representation. 2) We evaluate the model on two publicly available benchmarks datasets and show that it significantly improves state-of-the-art performance on both tasks. 3) We conduct ablative studies to highlight the contributions of different modules to the overall performance.

III. METHOD

A. Problem Formulation

Behavior prediction is formulated as a multi-objective optimization problem. The goal is to learn the following distribution $p(\text{Tasks}|\text{Modalities})$ where $\text{Tasks} = \{l_i^{t+1:t+\tau}, a_i, dl_i^{t+\tau}\}$ correspond to trajectory, future crossing action and discrete location respectively for some pedestrian $1 < i < n$. $\text{Modalities} = \{l_i^{t-o+1:t}, dl_i^{t-o+1:t}, sc^t, em^{t-o+1:t+\tau}\}$ denote i th pedestrian's observed coordinates and discrete location, and scene and ego-vehicle motion respectively. We denote the length of observations as o and prediction as τ .

B. Architecture

The proposed architecture shown in Figure 1 has three modules: *cross-modal encoder* that captures pair-wise dependencies between data modalities via multi-head attention, *semantic attention interaction module* (SAIM) that leverages semantic information to model interactions between pedestrians and their surroundings, and *hybrid gated decoder* that controls flow of information between shared and independent decoders for final prediction. Implementation details for each module are given in the following subsections.

C. Cross-modal Encoder

Input data modalities: *Pedestrian location and velocity.* We use normalized 2D coordinates of pedestrians in the image plane $[(x_1, y_1), (x_2, y_2)]$, corresponding to top-left and

bottom-right corners of the pedestrian bounding boxes. This allows the model to localize the pedestrian with respect to the ego-vehicle and other scene elements. Pedestrian velocity at time t is computed as difference between their 2D positions at time t and $t - 1$.

Discrete locations. Following [7], we transform pedestrian coordinates by dividing the image plane into a grid of $N \times M$ cells with unique class labels. At each time step, the cell whose center is closest to the pedestrian is selected.

Ego-vehicle motion captures both the changes in the scale of pedestrians in egocentric view and impact of the ego-vehicle motion on pedestrian behavior. We use three sources of information to characterize ego-motion: vehicle speed s and vehicle velocity along x and z axes denoted as v_x and v_z , respectively.

Cross-modal attention modulation. It is important to capture correlations between different data modalities to generate representations robust to noise or missing data [63]. Bifold method proposed in [7] processes multimodal data using hard parameter sharing and individually, and combines the results for final representation. However, this method is problematic when pair-wise correlations across different modalities are not similar.

We propose to instead use multi-head attention units, Attn_{MH} . For each of m data modalities we define $m - 1$ attention units as follows,

$$\begin{aligned} \text{Attn}_{\text{MH}}(Q_j, K_m, V_m) &= (\text{head}_1 \oplus \text{head}_2 \dots \text{head}_k)W^O \\ \text{head}_k &= \text{S-Attn}(Q_j W_k^Q, K_m W_k^K, V_m W_k^V) \\ \text{S-Attn}(Q, K, V) &= \text{Softmax}(QK^\top)V, \end{aligned} \quad (1)$$

where K_m and V_m are key and value for m th data modality and Q_j is the query based on j th data modality, where $j \neq m$. S-Attn denotes self-attention network with k heads and \oplus denotes concatenation operation. Prior to entering Attn_{MH} units, data from each modality is fed into an embedding layer followed by concatenation with a positional encoding.

Transformer encoder. Final encoding is generated using a Transformer module [64] which receives the concatenated output of multi-head attention units and a positional encoding as input. The encoder consists of l identical layers connected

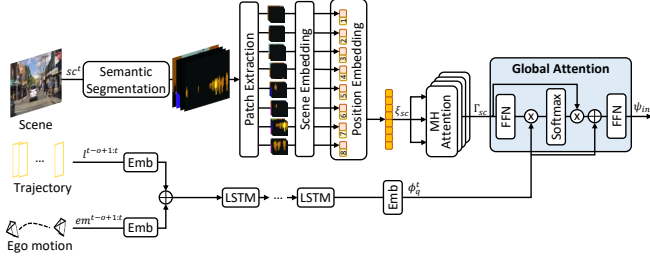


Fig. 2: A diagram of the semantic interaction attention module (SAIM) which takes as input scene image at time t , ego-vehicle motion and pedestrian trajectory.

sequentially, i.e. the output of one layer is the input of the next one. The first module in each layer is self-attention followed by a feed-forward neural network (FFN). The output of each module is added to its input via a residual connection and fed into a normalization layer based on [65],

$$\begin{aligned} \text{Transformer}_1(I) &= \text{Norm}(I' + \text{FFN}(I')) \\ I' &= \text{Norm}(I + \text{Attn}_{\text{MH}}(I)). \end{aligned} \quad (2)$$

D. Semantic Attention Interaction Module (SAIM)

As demonstrated by the authors of [7], visual information can be used to effectively model interactions between pedestrians and their surroundings. Their approach, however, has a number of drawbacks. First, it requires sequences of global scenes grouped by semantic categories, each of which are processed separately using CNNs and recurrent networks. Second, sequences of pedestrians and surrounding people are encoded for each prediction which is computationally expensive. Third, ego-vehicle motion is not accounted for. We address these issues by using an attention-based architecture.

The input to the proposed SAIM module is the RGB image of current scene at time t recorded from an ego-centric perspective. The image is passed to a segmentation algorithm whose output is organized into four channels: persons, bikes/bicyclists, vehicles and static context. The multi-channel map is then split into local patches of size $p_s \times p_s$ producing a total of np patches. The patches are converted into vectors of size λ_p via linear embedding layer. Inspired by [66], the embedded patches are concatenated with positional encodings forming the input to an Attn_{MH} unit that outputs $\Gamma_{sc} \in R^{np \times \lambda_p}$. We encode pedestrian coordinates and ego-motion for the duration of observation using embedding layers and a shared LSTM unit. The output of the LSTM is embedded into vector ϕ_q^t and fed into a global attention module [67] to generate the final interaction encoding ψ_{int} as follows,

$$\begin{aligned} \psi_{int} &= \tanh(W_c[c \oplus \phi_q^t]) \\ c &= \text{softmax}(\phi_q^t W_\Gamma \Gamma_{sc}) \Gamma_{sc} \\ \Gamma_{sc} &= \text{Attn}_{\text{MH}}(\xi_{sc}, \xi_{sc}, \xi_{sc}) \end{aligned} \quad (3)$$

where ξ_{sc} is the output of scene positional embedding.

E. Hybrid Gated Decoder

Decoders. In a multitask learning framework, capturing the correlation between decoders for individual tasks is

essential. [7] used two sets of decoders that independently and jointly produced predictions. However, this method, is computationally inefficient as two sets of predictions are generated and a post-processing stage is needed to combine them. Furthermore, there is no feedback between task-specific decoders and the shared one.

Our proposed hybrid decoder architecture, motivated by [68], also uses independent and shared decoders. Unlike [7], the shared decoder serves as a knowledge-base allowing the individual decoders to benefit from the shared information. To this end, the output of the shared decoder is passed through gate mechanism which controls the flow of information and is formulated as,

$$\begin{aligned} h_j^t &= o_j^t \odot \tanh(c_j^t) \\ c_j^t &= f_j^t \odot c_j^{t-1} + i_j^t \odot \hat{c}_j^t \\ \hat{c}_j^t &= \tanh(W_c \hat{x}^t) + U_c h_j^{t-1} \end{aligned} \quad (4)$$

where i_j^t , f_j^t , and o_j^t are input, forget and output gates of decoder j , c_j^t and h_j^t are its cell state and output at time t . Here, $h_{sd}^t = \overleftarrow{h}_{sd}^t \oplus \overrightarrow{h}_{sd}^t$ is the output of the shared decoder where \leftarrow and \rightarrow denote forward and backward LSTMs. The input \hat{x}^t is given by,

$$\begin{aligned} \hat{x}^t &= \sigma(h_{sd}^t) \odot h_{sd}^t \oplus \Psi^t \\ \Psi^t &= \psi_{cm} \oplus \psi_{int} \oplus em^t. \end{aligned}$$

Prediction. The primary tasks are trajectory and action prediction. Trajectory is predicted as 2D bounding box coordinates of the pedestrians by applying a linear transformation to the hidden states of the trajectory decoder. To predict the probability of crossing action at some time in the future, we perform a step-wise linear transformation at each time step followed by a sigmoid activation. For a unified prediction, we apply a global average pooling to compute the mean prediction across time steps. An auxiliary task, prediction of final discrete location of pedestrians, acts as a regularizer and a bridge between the two primary tasks. Similar to action, we classify the locations of the pedestrian at each time-step and average them over all prediction time steps.

Learning Objectives. Our learning objective is a weighted sum of the objectives for each task,

$$\begin{aligned} L &= \omega_l \sum_{i=1}^n \sum_{j=t}^{t+\tau} \log(\cosh(y_i^j - \hat{y}_i^j)) \\ &- \omega_a \sum_{i=1}^n y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \\ &- \omega_{dl} \sum_{i=1}^n \sum_{dlc} y_{i,dlc} \log(\hat{y}_{i,dlc}) \end{aligned} \quad (5)$$

where ω_l , ω_a and ω_{dl} are weight coefficients for each task determined empirically, dlc is the discrete location classes.

IV. EVALUATION

Implementation. We set the sizes of all input and attention embeddings to 64. For the encoder, we use Attn_{MH} and two transformer layers. In SAIM, semantic maps are generated using [69] pre-trained on the CityScapes dataset [70] and four Attn_{MH} . The maps are split into patches of size 12×12 . The output dimension of the global attention is set to 128.

For LSTM units, we use 128 hidden cells and L2 regularization of 0.0001. We use tanh activation for SAIM and trajectory decoder and softsign for remaining decoders. Shared decoder is implemented as a bidirectional LSTM. Following [7], we divided the image plane into a grid of 18×32 with 60×60 px cells for discrete locations.

Datasets. *Pedestrian Intention Estimation (PIE)* [4] is chosen as the primary dataset for evaluation as it contains over 6 hours of egocentric driving footage along with bounding box annotations for traffic objects, action labels for pedestrians and ego-vehicle sensor information. We use the default data split and sample pedestrians sequences with 50% overlap and time to event between 1 and 2 seconds following [7] resulting in 3980 training sequences out of which 995 are crossing events.

Joint Attention in Autonomous Driving (JAAD) [9] is a pedestrian dataset with short egocentric driving clips of urban driving with bounding box and action annotations for pedestrians. Unlike the PIE dataset, JAAD lacks ego-motion information, therefore we use textual driver action labels, such as *moving slow* and *slowing down*, instead. We use the data split proposed in [71] and sample sequences similar to PIE resulting in 3955 sequences with 807 crossing samples.

Training. The proposed model is trained with the initial learning rate of 10^{-4} on PIE and 5×10^{-5} on JAAD and batch size of 8 using RMSProp [72] for 200 epochs. Learning rate is reduced by a factor of 0.2 based on the validation loss. To compensate for crossing/non-crossing sample imbalance, we use class weights computed using the ratio of the samples. For loss weights we set the following values empirically, $\omega_l = 0.6$ for PIE and $\omega_l = 0.5$ for JAAD with $\omega_a = 1$ and $\omega_g = 1$ same for both datasets.

Metrics. We report the results for 0.5s observation and 1s prediction. For *trajectory prediction*, we use **ADE** and **FDE** metrics computed based on the center coordinates of pedestrian bounding boxes, and **ARB** and **FRB** calculated by RMSE of bounding box coordinates. The latter shows how well the model localizes pedestrians in image plane. All metrics are reported in pixels. For *action prediction*, we use common classification metrics including **accuracy**, Area Under Curve (**AUC**), **F1** and **precision**.

Models. We compare our model to state-of-the-art egocentric behavior prediction methods. For *trajectory prediction* we use the following methods: *Future Person Localization (FPL)* (only predicts center coordinates) [31], *Bayesian LSTM (B-LSTM)* [34], *FOL* [32], and *PIE_{traj}* and *PIE_{full}* introduced in [4]. For *action prediction* the following models are selected: *ATGC* [9], *MM-LSTM* [40], *SF-GRU* [6], *PCPA* [36], and *I3D* [73]. We also compare to state-of-the-art *multitask prediction* model, *BiPed* [7].

A. Behavior Prediction

PIE. According to Table I, our model achieves state-of-the-art performance on all metrics. Trajectory prediction is improved by 22% on localization (bounding box) metrics and 14% on the rest. The most notable improvement on action prediction is on *precision* where a performance boost of 7%

TABLE I: Comparison to SOTA methods. \uparrow and \downarrow mean higher or lower values are better respectively.

Method	Trajectory				Action				
	ADE	FDE	ARB	FRB	Method	Acc	AUC	F1	Prec
PIE									
FOL [32]	73.87	164.53	78.16	143.69	ATGC [9]	0.59	0.55	0.36	0.35
FPL [31]	56.66	132.23	-	-	I3D [73]	0.79	0.75	0.64	0.61
B-LSTM [34]	27.09	66.74	37.41	75.87	MM-LSTM [40]	0.84	0.84	0.75	0.68
PIE _{traj} [4]	21.82	53.63	27.16	55.39	SF-GRU [6]	0.86	0.83	0.75	0.73
PIE _{full} [4]	19.50	45.27	24.40	49.09	PCPA [36]	0.86	0.84	0.76	0.73
BiPed [7]	15.21	35.03	19.62	39.12	BiPed [7]	0.91	0.90	0.85	0.82
PedFormer	13.08	30.35	15.27	32.79	PedFormer	0.93	0.90	0.87	0.89
JAAD									
FOL [32]	61.39	126.97	70.12	129.17	ATGC [9]	0.64	0.60	0.53	0.50
FPL [31]	42.24	86.13	-	-	I3D [73]	0.82	0.75	0.55	0.49
B-LSTM [34]	28.36	70.22	39.14	79.66	MM-LSTM [40]	0.80	0.60	0.40	0.39
PIE _{traj} [4]	23.49	50.18	30.40	57.17	SF-GRU [6]	0.83	0.77	0.58	0.51
PIE _{full} [4]	22.83	49.44	29.52	55.43	PCPA [36]	0.83	0.77	0.57	0.50
BiPed [7]	20.58	46.85	27.98	55.07	BiPed [7]	0.83	0.79	0.60	0.52
PedFormer	17.89	41.63	24.56	48.82	PedFormer	0.93	0.76	0.54	0.65

TABLE II: Performance of different encoding methods. \uparrow and \downarrow mean higher or lower values are better respectively.

Methods	Interaction	ADE \downarrow	FDE \downarrow	ARB \downarrow	FRB \downarrow	F1oU \uparrow
Modality Transformers	\times	15.07	36.28	20.16	40.91	0.103
Shared Transformer	\times	14.88	35.79	19.91	40.03	0.108
PedFormer	\times	13.60	32.06	17.28	35.27	0.111
PedFormer w/o G Att	\checkmark	14.14	33.57	18.95	38.06	0.112
PedFormer w/o Motion	\checkmark	13.51	31.48	17.31	35.22	0.117
PedFormer	\checkmark	13.08	30.35	15.27	32.79	0.122

is gained. Such a boost highlights the effectiveness of our method in capturing the dependency between complementary tasks, in this case using predicted trajectory to distinguish between crossing and non-crossing actions.

JAAD. On JAAD dataset, significant improvement of up to 13% is achieved across different metrics for both trajectory and action prediction. Compared to PIE, improvement on localization metrics is more limited primarily due to the homogeneity of JAAD in which the majority of samples are the pedestrians that are walking on sidewalks towards or away from the camera. On actions, the biggest improvement is achieved on *precision* (as is the case for PIE dataset), while *accuracy* is also improved by up to 10%.

B. Ablation studies

We report the results on the core behavior task of trajectory prediction and elaborate on actions as part of the discussion. We also use a more explicit localization metric, final intersection over union (**F1oU**), to highlight how well the final predicted bounding boxes overlap with the ground truth.

Multimodal encoding. We report on alternative transformer-based encoding schemes, namely *modality transformer* where each modality is passed through independent transformer and then combined at the end, and *shared transformer* in which the input modalities are first combined before entering a shared transformer for encoding. For these variations no interaction module is used and the rest of the model remains the same.

The results of the experiment are shown in Table II. First observation is that, overall, using separate encoders is inferior as the model fails to capture the correlation between different modalities. Simply sharing the encoder, however, does not result in significant improvements. Once pair-wise cross-modal attention scoring is added, a significant boost across all metrics can be observed. The greater improvement over *ARB/FRB* indicates that the model better captures the relation between the scale of bounding boxes and ego-motion which directly affects it.

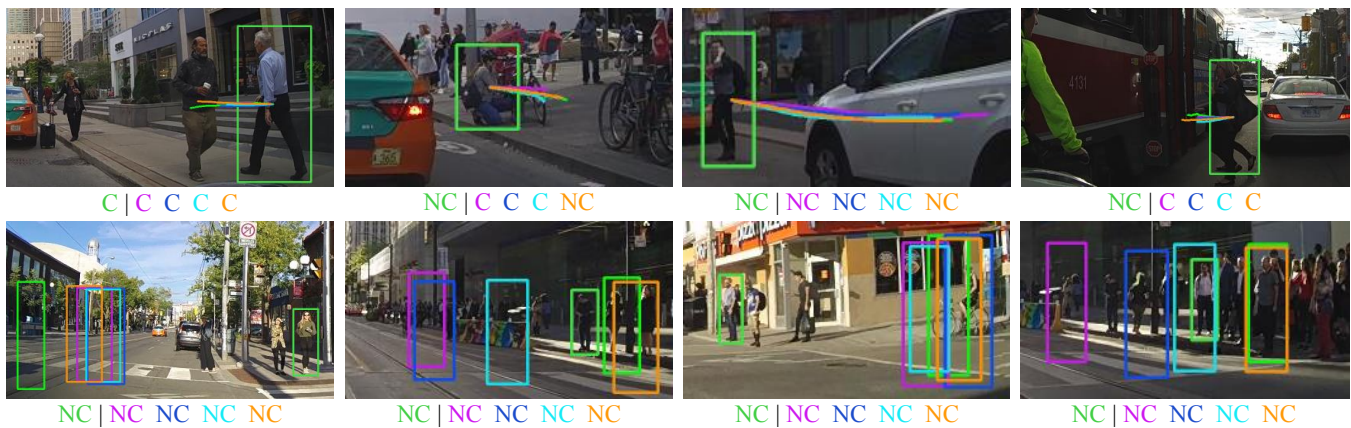


Fig. 3: Qualitative results of PedFormer on PIE. First row shows predicted trajectories and second final predicted bounding boxes. The colors correspond to **ground truth**, **PedFormer (ours)**, **Modality Trans.**, **Task-based decoder** and **PedFormer with no interaction**. The letters indicate predicted crossing (C) or non-crossing (NC) events.

TABLE III: Performance of different decoding mechanisms. \uparrow and \downarrow mean higher or lower values are better respectively.

Method	$ADE \downarrow$	$FDE \downarrow$	$ARB \downarrow$	$FRB \downarrow$	$FIoU \uparrow$
Task-based	16.10	37.79	21.22	42.76	0.100
Shared	15.12	35.44	19.49	39.33	0.100
Bifold	14.83	35.23	18.93	38.04	0.105
Hybrid	13.78	32.89	17.82	36.51	0.110
Gated Hybrid	13.08	30.35	15.27	32.79	0.122

Interaction modeling. As a baseline, we omit the global attention component and replace it with a 1×1 convolutional layer to generate a summary vector over the output of the multi-head attention modules denoted as *PedFormer w/o G. Att.* To highlight the importance of dynamics in interaction modeling, we also remove the motion encoding (*PedFormer w/o Motion*) and use the last element of the multi-head attention output as a reference to generate a summary.

As shown in Table II, simply using an attention framework over scenes not only does not provide any benefits but can also add noise and consequently deteriorate the results. Having a dual attention framework mitigates the issue and results in measurable, albeit limited, improvement. This is due to the fact that although a single image provides information regarding the structure of the scene, on its own it does not capture the movements of the agents. Consequently, once dynamic information is included in the computation of the final representation, a significant improvement is achieved. Overall, using the proposed interaction modeling technique provides up to 12% improvement over trajectory metrics and 3% on action metrics, specifically AUC and $F1$.

Decoding. To study the impact of different decoding techniques on the overall performance of the model we use four variants of decoder architectures while keeping the rest of the model unchanged. The baseline models are *task-based* where each task has a separate decoder and *shared* where a single decoder is used for all tasks. We also use the **bifold** method of [7] and two variations of the proposed *hybrid* method with and without a gating mechanism.

Results in Table III show that using a combined representation, whether bifold or hybrid, leads to better perfor-

mance on all trajectory and action metrics with the biggest improvement achieved on *precision*, which agrees with findings in [7]. Direct feedback mechanism between shared and independent decoders allows hybrid approach to better learn the dependencies between tasks compared to bifold method. Since uncontrolled flow of information into task decoders can introduce noise, proposed gating mechanism allows the task decoders to be selective on the use of shared representation and provides further improvements. Overall, compared to the baseline task-based method, the proposed gated hybrid approach improves the performance by up to 12% on trajectory and 4% on action metrics.

C. Qualitative results

Figure 3 shows qualitative examples of the proposed and baseline methods. PedFormer consistently produces better predictions, both in terms of trajectories and bounding boxes, in various scenarios. It is particularly evident when ego-motion is significant, e.g. when the ego-vehicle is making a right turn or is accelerating (first and second images from the left in bottom row), owing to better encoding of ego-motion and effective use of visual context. In the case of crossing events, sometimes the trajectory can be misleading, e.g. the top right image shows the pedestrian boarding a streetcar but all models classified the event as crossing due to the presence of lateral movement towards the road.

V. CONCLUSION

This paper introduced a novel pedestrian behavior prediction method that simultaneously predicts trajectories and actions. The proposed approach benefits from a cross-modal transformer-based encoder, a scene-based interaction module with attention modulation, and a gated hybrid decoder mechanism for producing predictions. We conducted experiments on public benchmark datasets and showed that our method achieves state-of-the-art performance on both tasks by a wide margin. Furthermore, via ablation studies, we highlighted the contribution of different proposed mechanisms on the overall performance of our method.

REFERENCES

- [1] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2019.
- [2] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *Intelligent Vehicles Symposium (IV)*, 2017.
- [3] L. Shi, L. Wang, C. Long, S. Zhou, M. Zhou, Z. Niu, and G. Hua, "Sgcn: Sparse graph convolution network for pedestrian trajectory prediction," in *CVPR*, 2021.
- [4] 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 *ICCV*, 2019.
- [5] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.
- [6] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked RNNs," in *BMVC*, 2019.
- [7] A. Rasouli, M. Rohani, and J. Luo, "Bifold and semantic reasoning for pedestrian behavior prediction," in *ICCV*, 2021.
- [8] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *CVPR*, 2019.
- [9] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *ICCVW*, 2017.
- [10] M. Liu, S. Tang, Y. Li, and J. Rehg, "Forecasting human object interaction: Joint prediction of motor attention and actions in first person video," in *ECCV*, 2020.
- [11] A. Piergiovanni, A. Angelova, A. Toshev, and M. S. Ryoo, "Adversarial generative grammars for human activity prediction," in *ECCV*, 2020.
- [12] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *CVPR*, 2019.
- [13] T. Yao, M. Wang, B. Ni, H. Wei, and X. Yang, "Multiple granularity group interaction prediction," in *CVPR*, 2018.
- [14] D. Epstein, B. Chen, and C. Vondrick, "Oops! Predicting unintentional action in video," in *CVPR*, 2020.
- [15] M. Qi, J. Qin, Y. Wu, and Y. Yang, "Imitative non-autoregressive modeling for trajectory forecasting and imputation," in *CVPR*, 2020.
- [16] P. Felsen, P. Agrawal, and J. Malik, "What will happen next? Forecasting player moves in sports videos," in *ICCV*, 2017.
- [17] P. Dendorfer, S. Elflein, and L. Leal-Taixe, "Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction," in *ICCV*, 2021.
- [18] N. Shafiee, T. Padir, and E. Elhamifar, "Introvert: Human trajectory prediction via conditional 3d attention," in *CVPR*, 2021.
- [19] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," in *CVPR*, 2020.
- [20] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *CVPR*, 2020.
- [21] J. Sun, Q. Jiang, and C. Lu, "Recursive social behavior graph for trajectory prediction," in *CVPR*, 2020.
- [22] H. Sun, Z. Zhao, and Z. He, "Reciprocal learning networks for human trajectory prediction," in *CVPR*, 2020.
- [23] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *ECCV*, 2020.
- [24] C. Choi and B. Dariush, "Looking to relations for future trajectory forecast," in *ICCV*, 2019.
- [25] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction," in *CVPR*, 2019.
- [26] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *CVPR*, 2019.
- [27] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.
- [28] L. Neumann and A. Vedaldi, "Pedestrian and ego-vehicle trajectory prediction from monocular camera," in *CVPR*, 2021.
- [29] O. Makansi, O. Cicek, K. Buchicchio, and T. Brox, "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior," in *CVPR*, 2020.
- [30] S. Malla, B. Dariush, and C. Choi, "TITAN: Future forecast using action priors," in *CVPR*, 2020.
- [31] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *CVPR*, 2018.
- [32] 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 *ICRA*, 2019.
- [33] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *IROS*, 2019.
- [34] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *CVPR*, 2018.
- [35] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *CVPR*, 2019.
- [36] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *WACV*, 2021.
- [37] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *WACV*, 2020.
- [38] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal DenseNet," in *ICRA*, 2019.
- [39] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *ICRA*, 2019.
- [40] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "VIENA: A driving anticipation dataset," in *ACCV*, 2019.
- [41] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *CVPR*, 2018.
- [42] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic task prioritization for multitask learning," in *ECCV*, 2018.
- [43] G. Hu, L. Liu, Y. Yuan, Z. Yu, Y. Hua, Z. Zhang, F. Shen, L. Shao, T. Hospedales, N. Robertson, and Y. Yang, "Deep multi-task learning to recognise subtle facial expressions of mental states," in *ECCV*, 2018.
- [44] K. Du, X. Lin, Y. Sun, and X. Ma, "CrossInfoNet: Multi-task information sharing based hand pose estimation," in *CVPR*, 2019.
- [45] A. Mallya, D. Davis, and S. Lazebnik, "Piggyback: Adapting a single network to multiple tasks by learning to mask weights," in *ECCV*, 2018.
- [46] Z. Zeng, X. Li, Y. K. Yu, and C.-W. Fu, "Deep floor plan recognition using a multi-task network with room-boundary-guided attention," in *ICCV*, 2019.
- [47] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *ICCV*, 2019.
- [48] K. Hassani and M. Haley, "Unsupervised multi-task feature learning on point clouds," in *ICCV*, 2019.
- [49] S. Casas, W. Luo, and R. Urtasun, "IntentNet: Learning to predict intention from raw sensor data," in *CORL*, 2018.
- [50] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *CVPR*, 2018.
- [51] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3D object detection," in *CVPR*, 2019.
- [52] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *CVPR*, 2019.
- [53] P. Wu, S. Chen, and D. N. Metaxas, "MotionNet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps," in *CVPR*, 2020.
- [54] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani, "MX-LSTM: Mixing tracklets and vislets to jointly forecast trajectories and head poses," in *CVPR*, 2018.
- [55] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "GD-GAN: Generative adversarial networks for trajectory prediction and group detection in crowds," in *ACCV*, 2019.
- [56] Z. Zhang, J. Gao, J. Mao, Y. Liu, D. Anguelov, and C. Li, "STINet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction," in *CVPR*, 2020.

- [57] C. Choi, J. H. Choi, J. Li, and S. Malla, "Shared cross-modal trajectory prediction for autonomous driving," in *CVPR*, 2021.
- [58] S. H. Park, G. Lee, M. Bhat, J. Seo, M. Kang, J. Francis, A. R. Jadhav, P. P. Liang, and L.-P. Morency, "Diverse and admissible trajectory forecasting through multimodal context understanding," in *ECCV*, 2020.
- [59] T. Yau, S. Malekmohammadi, A. Rasouli, P. Lakner, M. Rohani, and J. Luo, "Graph-sim: A graph-based spatiotemporal interaction modelling for pedestrian action prediction," in *ICRA*, 2021.
- [60] J. Li, F. Yang, M. Tomizuka, and C. Choi, "EvolveGraph: Multi-agent trajectory prediction with dynamic relational reasoning," in *NeurIPS*, 2020.
- [61] V. Kosaraju, A. Sadeghian, R. Martin-Martin, I. Reid, H. Rezatofighi, and S. Savarese, "Social-BIGAT: Multimodal trajectory forecasting using Bicycle-GAN and graph attention networks," in *NeurIPS*, 2019.
- [62] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *ICML*, 2018.
- [63] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv:1706.05098*, 2017.
- [64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.
- [65] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.
- [66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.
- [67] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv:1508.04025*, 2015.
- [68] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *IJCAI*, 2016.
- [69] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Re-thinking atrous convolution for semantic image segmentation," *arXiv:1706.05587*, 2017.
- [70] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The CityScapes dataset for semantic urban scene understanding," in *CVPR*, 2016.
- [71] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "It's not all about size: On the role of data properties in pedestrian detection," in *ECCVW*, 2018.
- [72] T. Tieleman and G. Hinton, "Lecture 6.5-RMSProp, coursera: Neural networks for machine learning," Tech. Rep., 2012.
- [73] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *CVPR*, 2017.