2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
September 27 - October 1, 2021. Prague, Czech Republic

2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 978-1-6654-1714-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IROS51168.2021.9636241

Joint Intention and Trajectory Prediction Based on Transformer

Ze Sui, Yue Zhou, Xu Zhao, Ao Chen, Yiyang Ni

***Abstract*— Although autonomous driving technology has made tremendous progress in recent years, it is still challenging to predict the intentions and trajectories of pedestrians. The state-of-the-art methods suffer from two problems. (1) Existing works consider these two tasks separately, ignoring the connection between them. (2) The selection and integration of inputs for these tasks are not well designed. In this paper, these two tasks are taken into consideration in a unified model. In this way, the information provided by the labels of each other is shared, improving the performance of both tasks. Besides, in addition to the bounding boxes and speeds, orientation and road semantic segmentation features are taken into consideration to show the potential intention and road context of the pedestrian. And all the inputs are weighted by an attention module before integration. Meanwhile, a Transformer encoder is applied in our method to extract the temporal information from the fused feature sequence. Our method outperforms all previous models for both trajectory prediction and intention prediction tasks on the JAAD dataset and PIE dataset.**

I. Introduction

The past decade has witnessed the rapid development of autonomous driving techniques. The capability to predict the intentions and trajectories of vulnerable road users (VRUs), especially pedestrians, will significantly improve the safety of autonomous systems. However, even though the present autonomous systems are competent to detect road users well, there are still challenges predicting the intentions and trajectories of pedestrians. Three reasons limit the ability of autonomous driving systems to predict the intentions and trajectories of pedestrians. (1) The definition of intention in previous works brings misunderstanding to the system, as the label of pedestrian does not vary with time. (2) Intention prediction and trajectory prediction are regarded as two separate tasks, and the connection between them is neglected. (3) The potential intention of movement and the environment around pedestrians are not well provided by the input, and diverse inputs are simply concatenated during information fusion.

In most previous works, pedestrians' intention is defined as whether the pedestrians will cross the road in front of the vehicle [9], [20], [24]. That definition gives every pedestrian a fixed intention label, which does not vary with time and distance. The intention label will be marked as "crossing" all the time, even long before crossing or after crossing, as long as the pedestrian has crossed the road. That may lead to confusion for autonomous driving. So, in our definition, we only concern about whether the pedestrians will cross in the next fixed period. As shown in Fig. 1, the target will be labeled as a positive sample only when crossing occurs in the forecast period.

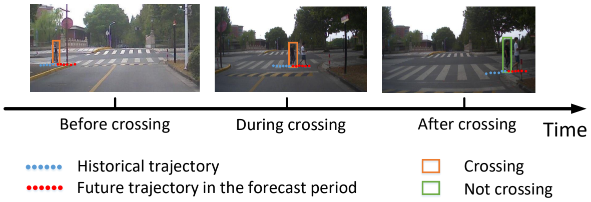

Fig. 1. The intention prediction task in our work. The target will be labeled as a positive sample only when crossing takes place in the forecast period.

Based on the above definition, the future trajectories and intentions are highly related but irreplaceable for two reasons. (1) Using future trajectories as supervision can help the model learn the potential intentions of the pedestrians, while the information from future crossing intention can help the model rectify the predicted trajectory. (2) Future trajectory provides more information for autonomous driving as it contains every location of a pedestrian, while pedestrian crossing intention gives more direct clues for driving decision. Therefore, a new framework is needed to exploit the relationship between the two tasks to improve performance.

In terms of input information selection, most previous methods only input historical trajectory information for prediction [10], [18]. Although a small number of pedestrian information (key point detection results [9], appearance features [19], etc.) are added, the information selected is relatively rough. The contexts are not directly related to the pedestrians in [20], and the appearance features are extracted through a pre-trained classification network in [19], which have low relevance to tasks. Multi-inputs are simply concatenated during multi-information fusion, which is challenging to extract pedestrians' intentions and trajectories.

In this paper, we propose a novel multi-task framework for intentions and trajectories prediction. We select the pedestrian's historical trajectory, speed, orientation, and the road semantic segmentation as the input of our framework. The orientation sequences can well reflect the potential movement intention, and the road semantic segmentation around the pedestrians shows the context interaction of the pedestrians. To union these clues effectively, we introduce an attention module to help the network pay attention to the critical parts in the fusion information. In addition, we utilize a Transformer encoder to extract temporal information. The temporal self-attention module in Transformer offers strong

Ze Sui, Yue Zhou, and Xu Zhao are with the School of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University. Emails: {sjtusz, zhaoxu, zhouyue, llykevin, niyiyang}@sjtu.edu.cn.

competence to extract sequential information. Besides, instead of processing the embedding sequence step by step like LSTM [12], Transformer processes the whole embedding sequence, which makes it more suitable for parallel computing.

The contributions of this work are summarized as follows.

- We propose a novel multi-task joint training framework to realize pedestrian trajectory prediction and intention prediction.
- We build a novel multi-input network based on an attention module and the Transformer encoder.
- The performance of trajectory prediction and intention prediction in our proposed network reaches state-of-the-art on the PIE dataset and JAAD dataset.

II. Related Work

Related work can be primarily categorized as three components: the definition of behavior prediction, the selection of input information, and the method for sequential modeling.

Behavior prediction for pedestrians. Understanding pedestrians' intentions and predicting their behavior in traffic scenes is an important task for autonomous driving. Early research on pedestrians mainly focused on the trajectories of pedestrians [17], [27], [28], as trajectories are easy to obtain and can intuitively reflect the motion of pedestrians. However, it is difficult to achieve a deep understanding of pedestrians' behavior simply by trajectories. In order to solve this problem, JAAD [20] release the first pedestrian intention prediction dataset to characterize pedestrians' crossing intentions, and many works are conducted on this dataset [7], [11], [16], [22], [24], [24]. However, the prediction of pedestrian's intentions is simply defined as whether the crossing occurs or not regardless of time. Subsequently, [19] publics a new data set, which supplements the label of pedestrian crossing tendency, but still fails to consider the relationship between crossing and time. Therefore, in this paper, we define the intention prediction problem as whether the pedestrian will cross in the next fixed period. At the same time, as a supplement to intention information, trajectory prediction tasks for pedestrians are also integrated into our framework, which can help improve the performance of intention prediction and provide more behavioral clues.

Input information selection. In order to predict the intention of pedestrians, previous works try to add different input information. [4] is one of the earliest works for intention analysis of pedestrians, and it takes the location and speed of pedestrians as input. As motion is insufficient to reflect the status of pedestrians, [6], [20], [21] use context information for intention prediction, including traffic lights, lanes, the width of the road, and so on. As a supplement, [9] judges the crossing intention of pedestrians by their keypoint features, and [19] proposes an intention prediction model taking trajectories and appearance as input. However, in these works, the context is extracted from the whole image rather than the region around the pedestrians, and the appearance features are extracted through classification networks, which are indirectly related to the task. In this work, we introduce the road semantic segmentation around pedestrians and the pedestrian's orientation as the input of the network, to provide more relevant information for the intentions and trajectories prediction task.

Sequential modeling. For the temporal information extraction from the input sequence, different methods have been proposed in different works. Early works model sequential information by Gaussian processes [25] and Bayesian nonparametric reachability trees [3]. With the emergence of Recurrent Neural Network (RNN) [29] networks, RNN-based networks have also been applied to the trajectory prediction [1], [14], [15]. Because of the problem of vanishing gradient in the RNN network, the LSTM network is proposed by [12]. Due to the powerful performance of LSTM, many works based on LSTM or the variants of LSTM are proposed to achieve sequential modeling [2], [5], [27], [30]. However, RNN-based and LSTM-based models need to input and process the sequence one by one, which is not only inefficient but also difficult to obtain global information. The proposal of Transformer [26] solves this problem well. The Transformer encoder can extract all input information at once through the position encoding and self-attention module, with strong global understanding capabilities and suitability for parallel computing. Therefore, in our work, we apply a Transformer encoder to extract the sequential information of pedestrians' feature sequences.

III. Proposed Method

A. Algorithm framework

The target of our algorithm is to predict pedestrians' trajectories and intentions for autonomous driving. We set a fixed time T_O as the observation period, a fixed time T_F as the forecast period. The inputs of the network are extracted from all frames during T_O, and the target outputs are pedestrians' trajectories and intentions during T_F. The framework of our proposed method is shown in Fig. 2. The framework consists of 6 modules, including the input module, feature extraction module, attention module, temporal module, and prediction module. The bounding box sequence and image sequence are sampled in observation period T_O, with the same sequence length s. Multi-inputs including orientation, trajectories, speed, and road semantic segmentation are obtained in the input module. The outputs of the input module are embedded as feature sequences with the same dimension. After feature extraction, attention weights for all feature sequences are calculated in the attention module, and the weights are multiplied element-wised with the feature sequences output by the feature extraction module. By concatenating all weighted feature sequences, fused feature sequences are generated in the feature fusion module. Furthermore, the fused feature sequences are fed into a Transformer encoder to extract the temporal information. In the end, an MLP network outputs the trajectories and intentions of the pedestrians.

B. Input module

Data preprocessing is implemented at the input module. To represent pedestrian's movement and the context, in addition

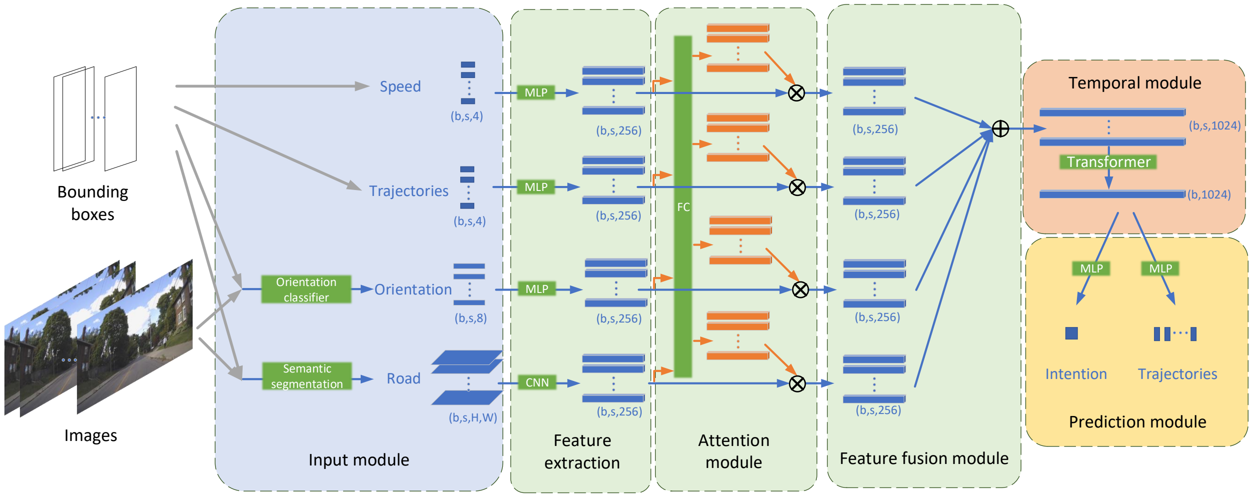

Fig. 2. The proposed intention estimation and trajectory prediction framework. The system receives a sequence of images and bounding boxes. The input module extracts the road segmentation and calculates the pedestrians' orientation. These cues are extracted as 256-D feature vectors separately by MLPs in the feature extraction module. After that, attention weights are calculated by Fully-connected networks in the attention module. In the feature fusion module, different features are concatenated after element-wise multiplication with their attention. The temporal module is a Transformer network that extracts the sequential information during the observation period. In the end, trajectories and intentions are predicted in the prediction module by a fully-connected network. In the diagram, b means the batch size during training, and s means the sequence length of the observation period, $\oplus$ means concatenation operation, and $\otimes$ means element-wise multiplication.

to taking the bounding boxes and pedestrians' speed as inputs, the orientation and road semantic segmentation is also taken into consideration.

Historical trajectories and speeds. The trajectories and speeds can easily be obtained by the bounding boxes and ids from most trajectories and intention datasets. In this work, we use the coordinates of bounding boxes sequence as every pedestrian's historical trajectories, and the speed is obtained by subtracting the bounding box coordinates of the previous frame from the next frame's coordinates.

Pedestrian orientation. Pedestrian's orientation sequence can well reflect the potential movement intention, which is very important for both trajectory prediction and intention prediction tasks. However, most trajectory prediction and intention prediction datasets do not contain the annotations of orientation. As a result, we divide pedestrian's orientation into eight categories and build an eight classification classifier to get this information. The classifier is a CNN network, with ResNet50 pre-trained on ImageNet as its backbone, finetuned in the CityPersons Dataset [31]. With the well-trained classifier, pedestrian's orientation is available in our model.

Road segmentation. One of the most important contexts for crossing intention prediction is the location at which the pedestrian stands. It is hard to judge the pedestrian's specific position in the environment by relying solely on the bounding boxes, which will lead to misjudgment of the crossing intention. To get the road semantic segmentation, we use the ERFNet [23] trained in the CityScape dataset [8]. In addition, though there are 19 categories in CityScape, we simplify the semantic segmentation to 3 categories: “sidewalk”, “road”, and “others”. The road segmentation around the pedestrian is cropped as the input of the trajectory and intention prediction network. As shown in Fig. 4, the region of interest is centering at the bottom of the original orange boxes, twice the width and half the height of the origin box.

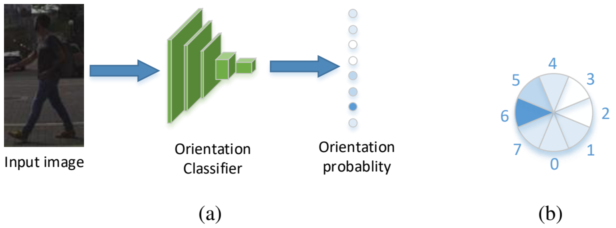

Fig. 3. Illustration of (a) pedestrian orientation classifier architecture and (b) the division of the eight orientations. The orientation of the pedestrian in (a) is “left”, corresponding to the category “6” in (b). Other orientations are also shown in (b).

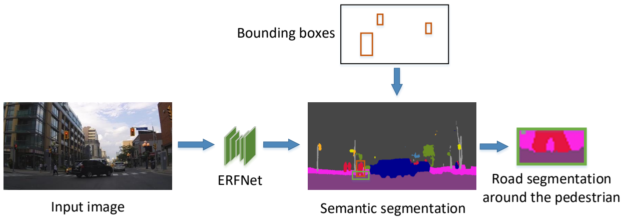

Fig. 4. The architecture of road segmentation in the input module. All the semantic segmentation is generated by a well-trained ERFNet. Orange bounding boxes refer to the original bounding box, and green boxes refer to the edited boxes.

To reduce the training time, all the above information is processed in advance to avoid repeated calculations during processing.

C. Feature extraction and fusion

Different kinds of inputs need embedding and fusion after preprocess. To facilitate the fusion of different types of information, we extract every input at every frame as a 256-dimension vector by networks. Road segmentation feature is extracted by a CNN network, and others are extracted by fully-connected networks.

In order to enable the neural network to pay attention to more critical information, we introduce an attention mechanism module before feature fusion. As shown in Fig. 2, attention weights are calculated by the attention module, with all the feature vectors as input. After that, all the features will take element-wise multiplication with corresponding attention weight. The formulation is written as,

$$G_{ori} = \Phi_{ori}(F_{ori}, F_{traj}, F_{speed}, F_{road}) \otimes F_{ori}, \quad (1)$$

$$G_{traj} = \Phi_{traj}(F_{ori}, F_{traj}, F_{speed}, F_{road}) \otimes F_{traj}, \quad (2)$$

$$G_{speed} = \Phi_{speed}(F_{ori}, F_{traj}, F_{speed}, F_{road}) \otimes F_{speed}, \quad (3)$$

$$G_{road} = \Phi_{road}(F_{ori}, F_{traj}, F_{speed}, F_{road}) \otimes F_{road}, \quad (4)$$

where F denotes the features extracted by feature extraction module; $\Phi(\cdot)$ denotes an MLP network or a CNN network, which is used to calculate the attention weights of the features; G denotes the weighted features; $\otimes$ denotes element-wise multiplication operation. The output of the attention module will be concatenated directly in the feature fusion module.

D. Transformer encoder

The temporal module is implemented by a Transformer encoder. Transformer networks are first proposed in [26] for Nature Language Processing (NLP). Unlike LSTM, Transformer networks do not need to process embedding sequence step by step; instead, it processes the whole embedding sequence. The positional encoding module enables the Transformer to extract temporal information with the embeddings fed at once, which makes it suitable for parallel computing. The Multi-head self-attention module in the Transformer enhances the temporal relation of inputs. So the Transformer encoder is suitable for the sequential modeling in our method. The architecture of the Transformer encoder is shown in Fig. 5.

Positional Encoding: The Transformer encoder encodes time t for each embedding from the feature sequence by positional encoding. A positional encoding mask PE is added to the embeddings in this module. PE is calculated as follow:

$$PE_{t,i} = \begin{cases} sin(\frac{t}{10000^{\frac{i}{D}}}) & i = 2k \\ cos(\frac{t}{10000^{\frac{i-1}{D}}}) & i = 2k+1 \end{cases}, \quad (5)$$

where i denotes the i^{th} dimension of the embedding, D denotes the dimension of embeddings.

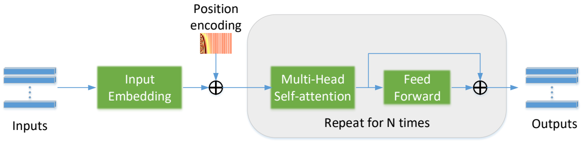

Fig. 5. The architecture of the Transformer encoder.

Self-Attention: The inputs are embedded into three matrices: Q (Query), K (Key), and V (Value). K and V are used to calculate the correlation among the input sequence. An attention matrix is calculated through scale and softmax layer. By multiplying V with the attention matrix, every part of the outputs is merged with the temporal information from other parts in the sequence. The formulation is shown below:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad (6)$$

where d_k denotes the dimension of K.

E. Multi-task joint training

The target of our algorithm is to predict pedestrians' intentions and trajectories during the forecast period. So two different outputs need optimization: intentions and trajectories.

Intention prediction. Intention prediction is a binary classification task. As the number of positive samples is not as large as the number of negative samples, we balance the positive samples and negative samples by re-sampling the positive samples to make the P/N ratio equals to 1:1. Binary Cross Entropy (BCE) loss is chosen as the loss function of the intention prediction task,

$$\mathcal{L}_{int}(x, y) = -y \cdot log(f(x)) - (1-y)log(1-f(x)), \quad (7)$$

where x denotes the input of the network, $f(x)$ denotes the predicted pedestrian intention, y denotes the ground truth of pedestrian intention.

Trajectories prediction. Trajectories prediction can be modeled as a regression task. The most intuitive method is to directly calculate the Mean Square Error (MSE) loss between the coordinates of ground truth trajectories and the coordinates of the predicted trajectories,

$$\mathcal{L}_{traj} = \sum_{t=1}^{T}(C_t - \hat{C}_t)^2, \quad (8)$$

where T denotes the sequence length to predict, C_t denotes the t^{th} coordinates of the ground truth location, $\hat{C}_t$ denotes coordinates of the t^{th} predicted location. However, in this way, the outputs of the network may be harder to reach convergence even though the coordinates have been normalized, as the values of target trajectories vary a lot from sample to sample. So we designed a new loss function to predict the speed of pedestrians,

$$\mathcal{L}_{traj} = \sum_{t=1}^{T}(S_t - \hat{S}_t)^2, \quad (9)$$

where

$$S_t = C_t - C_{t-1}, \hat{S}_t = \hat{C}_t - \hat{C}_{t-1}. \quad (10)$$

To avoid the accumulation of the speed error, the final loss function of trajectories prediction is written as,

$$\mathcal{L}_{traj} = \sum_{t=1}^{T}(C_t - (\sum_{i=1}^{t}\hat{S}_i + C_0))^2, \tag{11}$$

where $\hat{S}_i$ is the i^{th} predicted speed directly output by the network.

Automatic loss balance for joint training. The learning objective of each prediction head in the prediction module could be modeled as a multi-task learning problem. The total loss function could be written as,

$$\mathcal{L}_{total} = \omega_{traj}\mathcal{L}_{traj} + \omega_{int}\mathcal{L}_{int}, \tag{12}$$

where ω denotes the weight of each task. To balance the loss weights automatically, we adopt an adaptive strategy proposed in [13] using task-dependent uncertainty. The final loss function could be written as,

$$\mathcal{L}_{total} = e^{-\alpha}\mathcal{L}_{traj} + e^{-\beta}\mathcal{L}_{int} + \alpha + \beta, \tag{13}$$

where α and β is the task-dependent uncertainty, which could be adjusted adaptively during training. We refer readers to [13] for more details.

IV. Experiments

In this section, we evaluate our method by comparing our results with other state-of-the-art methods on two datasets of intentions and trajectories prediction. Besides, we conduct several ablation studies to discuss the influence of different inputs, different temporal models, multi-task joint training, and the attention module.

A. Datasets and evaluation metrics

Joint Attention Autonomous Driving (JAAD) Dataset. JAAD [20] is a dataset for studying joint attention in the context of autonomous driving published in 2017. It contains 346 short video clips of 5-10 seconds. Bounding boxes with trajectory ids and crossing tags are provided for most pedestrians, making this dataset suitable for pedestrian intentions and trajectories prediction evaluation. In addition, the dataset contains traffic information for each frame.

Pedestrian Intention Estimation (PIE) Dataset. PIE [19] is a new dataset for studying pedestrian behavior in traffic proposed in 2019. PIE contains 56 video clips of 4-10 minutes. There are much more frames, pedestrians, and longer trajectories than the JAAD dataset. Pedestrians' bounding boxes, trajectory ids, and behavioral annotations, including crossing state, are also included in the dataset. There are over 300K labeled video frames with 1842 pedestrian samples making this the largest publicly available dataset for evaluating the performance of our work.

Evaluation metrics. To evaluate the performance of different methods, we use Average Precision (AP), Accuracy (ACC) for intention prediction, and Average Displacement Error (ADE), Final Displacement Error (FDE) for trajectory prediction. AP is the average of precisions at different threshold values between 0 and 1 in the precision-recall curve. ADE means the root mean square error (RMSE) of all the predicted positions and real positions during the prediction period. FDE denotes the RMSE distance between the final predicted positions at the end of the predicted trajectory and the corresponding actual location. The units of both ADE and FDE are pixels.

B. Implementation details

Input and outputs. The input sequence is a 15-frame image sequence with the pedestrians' bounding boxes, and the target for prediction is the crossing intention and the trajectories in the next 15 frames. The input and output trajectories are defined as the bounding box sequences in this work.

Networks. In the input module, the orientation classifier is a CNN network with ResNet-50 as its backbone trained on the CityPerson dataset, and the semantic segmentation network is ERFNet trained on the CityScape dataset. The MLP networks in the feature extraction module are all 2-layer fully-connected networks, with Relu and BatchNormalization layers, and the output size of the hidden layer is 1024. The CNN network in the feature extraction module for road feature embedding consists of 2 CNN layers with kernel size 3 and 1 fully-connected layer. The FC networks in the attention module is a fully-connected network activated by a Sigmoid layer. In the temporal module, the number of layers N in the Transformer encoder is set to 3, and d_{model}, d_K, and d_V are all set to 512. The MLP networks in the prediction module are both 3-layer fully-connected networks. The activation function of the output layer for intention prediction is the Sigmoid function, and the Relu activation function is applied in the rest of the module.

C. Comparison with previous methods

Since most previous works defined pedestrian intention prediction in different ways, we choose PIE [19] whose code is available and SRR [16] whose definition is similar to our work for comparison. Due to the different definitions and pipelines for intentions and trajectories prediction, we run the training and evaluating code provided by PIE using the same processed data and label for a fair comparison. The training epoch for both PIE's code and ours is set to 100. The results on the PIE dataset and JAAD dataset are shown in TABLE I, and values are all evaluated in the test set. Since SRR did not conduct trajectory prediction experiments and their code is unavailable, only the results of intention prediction on the JAAD dataset from the original paper are shown in the table. All the other results are obtained by running the same experiment three times and taking the average value. As is shown in TABLE I, our method has better performance than previous works in both intention prediction and trajectories prediction tasks.

D. Ablation study

In order to evaluate different options of our network, we launched a series of ablation experiments. To make it easier for evaluation, all results in the tables in section IV-D are

TABLE I

Comparison with previous methods. Arrows after metrics indicate that the higher (↑) or lower (↓) value is better.

Method	JAAD Dataset			PIE Dataset		
	Intention	Trajectory		Intention	Trajectory	
	AP↑	ADE↓	FDE↓	AP↑	ADE↓	FDE↓
SRR [16]	79.28	-	-	-	-	-
PIE [19]	56.90	22.01	36.51	81.76	11.74	19.39
ours	**84.48**	**17.73**	**31.34**	**91.66**	**11.08**	**19.28**

TABLE II

Comparison of different temporal models and different number of layers

temporal network	AP ↑	ACC ↑	ADE ↓	FDE ↓
LSTM 1 layer	92.95	81.80	12.26	22.26
LSTM 3 layers	92.97	81.56	12.41	22.24
Transformer 1 layer	93.95	82.35	12.46	22.68
Transformer 2 layers	93.51	81.50	11.84	21.54
Transformer 3 layers	**94.25**	**82.46**	**11.82**	21.47
Transformer 6 layers	93.74	81.50	11.97	**21.21**

tested on the PIE validation set, different from the results in Table I which are tested on the test set.

Comparison of different temporal models. We first compare the Transformer encoder with the Long Short-Term Memory (LSTM) networks. We replace the 3-layer Transformer encoder with LSTM or Transformer encoder with other numbers of layers. As shown in TABLE II, the Transformer encoder outperforms the LSTM networks. Besides, with more layers, the Transformer encoder performs better on trajectories prediction, while 3 layers are best for intention prediction. The reason is that the deeper network has a more robust understanding but is harder for training. In this paper, we choose a 3-layer Transformer encoder as the temporal module.

Comparison of multi-task and single-task. The results of joint training and individual training of the two tasks are shown in Table III. The first two rows of the table are the results of separate training for intention prediction and trajectory prediction, and the third row is the result of multi-task joint training between the two tasks. The performance of multi-task is better than both task training individually. Multi-task training enables intention prediction and trajectory prediction to share weights and the extra information brought by each other's labels.

Ablation study on the attention module. To evaluate the effect of the attention module, we conduct an ablation experiment that removes the attention or replaces the sigmoid activation function with the softmax activation function. Results are shown in TABLE IV. When the attention module is removed, performance drops on both intention prediction and trajectory prediction tasks. Besides, as the dimension of the attention matrix is 1024, using Softmax as the activation function will make most of the attention weights close to 0, and little information is passed to the next layer.

TABLE III

Comparison of multi-task and single-task

Training tasks	AP ↑	ACC ↑	ADE ↓	FDE ↓
intention prediction	93.57	81.62	-	-
trajectory prediction	-	-	11.92	21.67
intention+trajectory prediction	**94.25**	**82.46**	**11.82**	**21.47**

TABLE IV

Ablation study on the attention module

Ways for attention module	AP ↑	ACC ↑	ADE ↓	FDE ↓
No attention	92.68	82.11	12.10	21.88
Softmax attention	50.01	50.01	44.68	84.06
Sigmoid attention	**94.25**	**82.46**	**11.82**	**21.47**

TABLE V

Comparison of different inputs

Inputs	AP ↑	ACC ↑	ADE ↓	FDE ↓
bboxes+speed	93.07	81.86	**11.49**	**20.69**
bboxes+speed+orientation	93.10	81.87	11.62	20.83
bboxes+speed+road	94.40	82.35	11.69	20.96
bboxes+speed+orientation+road	**94.25**	**82.46**	11.82	21.47

Comparison of different inputs. We analyze the impact of different inputs on the performance, and the results are shown in TABLE V. As we expect, more inputs bring better performance in terms of intention prediction. Although the adoption of orientation and road semantic segmentation reduces trajectory prediction performance, this can be explained by two reasons. 1. The correlation between road semantic segmentation and trajectory prediction is weaker than intention prediction, and the semantic segmentation obtained through ERFNet also brings errors 2. The bounding boxes and speed as input are ground truth, which covers the precise direction of movement, while the pedestrian orientation obtained by networks will bring errors instead. In real applications, accurate movement cannot be obtained, and the predicted pedestrian orientation and road semantic segmentation can be employed as a supplement to the potential movement direction.

V. Conclusion

In this paper, we propose a novel multi-input and multi-task framework that allows intention prediction and trajectories prediction to be learned in a shared model. We extract orientation and road features and combine them with the speed and trajectory features by an attention module, enabling the network to pay more attention to what matters. Besides, we introduce a Transformer encoder to extract the temporal information. Moreover, our method outperforms the previous method on both PIE and JAAD datasets. As a direction for future work, better models could be exploited to improve orientation prediction and road segmentation performance, then improving the performance of trajectories and intention prediction.

References

[1] Abdulrahman Al-Molegi, Mohammed Jabreel, and Baraq Ghaleb. Stf-rnn: Space time features-based recurrent neural network for predicting people next location. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE, 2016.

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[3] Georges Aoude, Joshua Joseph, Nicholas Roy, and Jonathan How. Mobile agent trajectory prediction using bayesian nonparametric reachability trees. In *Infotech Aerospace 2011*, page 1512. 2011.

[4] Tirthankar Bandyopadhyay, Chong Zhuang Jie, David Hsu, Marcelo H Ang, Daniela Rus, and Emilio Frazzoli. Intention-aware pedestrian avoidance. In *Experimental Robotics*, pages 963–977. Springer, 2013.

[5] Niccoló Bisagno, Bo Zhang, and Nicola Conci. Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[6] Sarah Bonnin, Thomas H Weisswange, Franz Kummert, and Jens Schmüdderich. Pedestrian crossing prediction using multiple context-based models. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 378–385. IEEE, 2014.

[7] Pablo Rodrigo Gantier Cadena, Ming Yang, Yeqiang Qian, and Chunxiang Wang. Pedestrian graph: Pedestrian crossing prediction based on 2d pose estimation and graph convolutional networks. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2000–2005. IEEE, 2019.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] Zhijie Fang and Antonio M López. Is the pedestrian going to cross? answering by 2d pose estimation. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1271–1276. IEEE, 2018.

[10] Francesco Giuliari, Irtiza Hasan, Marco Cristani, and Fabio Galasso. Transformer networks for trajectory forecasting. *arXiv preprint arXiv:2003.08111*, 2020.

[11] Pratik Gujjar and Richard Vaughan. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2097–2103. IEEE, 2019.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[14] Aida Khosroshahi, Eshed Ohn-Bar, and Mohan Manubhai Trivedi. Surround vehicles trajectory analysis with recurrent neural networks. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2267–2272. IEEE, 2016.

[15] ByeoungDo Kim, Chang Mook Kang, Jaekyum Kim, Seung Hi Lee, Chung Choo Chung, and Jun Won Choi. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 399–404. IEEE, 2017.

[16] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[17] Wei-Chiu Ma, De-An Huang, Namhoon Lee, and Kris M Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 774–782, 2017.

[18] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *European Conference on Computer Vision*, pages 759–776. Springer, 2020.

[19] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019.

[20] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 206–213, 2017.

[21] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Understanding pedestrian behavior in complex traffic scenes. *IEEE Transactions on Intelligent Vehicles*, 3(1):61–70, 2017.

[22] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Towards social autonomous vehicles: Understanding pedestrian-driver interactions. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 729–734. IEEE, 2018.

[23] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. Efficient convnet for real-time semantic segmentation. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1789–1794. IEEE, 2017.

[24] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9704–9710. IEEE, 2019.

[25] Meng Keat Christopher Tay and Christian Laugier. Modelling smooth paths using gaussian processes. In *Field and Service Robotics*, pages 381–390. Springer, 2008.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[27] Hao Xue, Du Q Huynh, and Mark Reynolds. Bi-prediction: pedestrian trajectory prediction based on bidirectional lstm classification. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2017.

[28] Hao Xue, Du Q Huynh, and Mark Reynolds. Pedestrian trajectory prediction using a social pyramid. In *Pacific Rim International Conference on Artificial Intelligence*, pages 439–453. Springer, 2019.

[29] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[30] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12085–12094, 2019.

[31] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017.