Multimodal Transformer Network for Pedestrian Trajectory Prediction

Ziyi Yin[1], **Ruijin Liu**[1], **Zhiliang Xiong**[2], **Zejian Yuan**[1]

[1]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[2]Shenzhen Forward Innovation Digital Technology Co. Ltd, China
{yzy19980922, lrj466097290}@stu.xjtu.edu.cn, leslie.xiong@forward-innovation.com, yuan.ze.jian@xjtu.edu.cn

Abstract

We consider the problem of forecasting the future locations of pedestrians in an ego-centric view of a moving vehicle. Current CNNs or RNNs are flawed in capturing the high dynamics of motion between pedestrians and the ego-vehicle, and suffer from the massive parameter usages due to the inefficiency of learning long-term temporal dependencies. To address these issues, we propose an efficient multimodal transformer network that aggregates the trajectory and ego-vehicle speed variations at a coarse granularity and that interacts with the optical flow in a fine-grained level to fill the vacancy of highly dynamic motion. Specifically, a coarse-grained fusion stage fuses the information between trajectory and ego-vehicle speed modalities to capture the general temporal consistency. Meanwhile, a fine-grained fusion stage merges the optical flow in the center area and pedestrian area, which compensates the highly dynamic motion of ego-vehicle and target pedestrian. The whole network is only attention-based that can efficiently model long-term sequences for better capturing the temporal variations. Our multimodal transformer is validated on the PIE and JAAD datasets and achieves the state-of-the-art performance with the most light-weight model size. The codes are available at https://github.com/ericyinyzy/MTN_trajectory.

1 Introduction

Pedestrian trajectory prediction anticipates the future bounding boxes of pedestrians in an ego-centric view of a moving vehicle, which is critical for autonomous driving systems to avoid possible collisions. It also benefits various visual research fields such as pedestrians intention estimation [Schneemann and Heinemann, 2016; Rehder *et al.*, 2018; Saleh *et al.*, 2019], video prediction [Wichers *et al.*, 2018; Oliu *et al.*, 2018; Ye *et al.*, 2019; Wu *et al.*, 2020], and pose forecasting [Mangalam *et al.*, 2020; Adeli *et al.*, 2020; Cao *et al.*, 2020]. The task requires different visual modalities to capture the highly dynamic motion information between pedestrians and ego-vehicle, which is hard to reflect in the changes of bounding boxes [Styles *et al.*, 2020]. Additionally, how to model the long-term location dependencies more effectively and implement with fewer parameters also increases the challenges.

Existing approaches have closely studied additional visual modalities, which have significantly improved the performance on pedestrian trajectory prediction tasks compared to those only trajectory-based methods [Alahi *et al.*, 2016; Bhattacharyya *et al.*, 2018]. Some methods [Rasouli *et al.*, 2019; Malla *et al.*, 2020] utilize image sequence to extract a semantic prior for guiding the future trajectory, like crossing intention [Rasouli *et al.*, 2019] or predefined action category [Malla *et al.*, 2020]. The semantic priors can provide general orientation (*e.g.* across or along the sidewalk) of future trajectories whereas it is hard to satisfy the demand for precise locating. Recently, an approach [Makansi *et al.*, 2020] exploits scene segmentation to estimate all possible end-locations of target pedestrian to predict future trajectory. The performance, however, may degenerate because of the low accuracy of end-locations estimation caused by the limited perception perspective and the changing scene from the ego-centric view. A remedy for these drawbacks is to introduce optical flow to extract motion features for compensating the temporal features in the past trajectory [Styles *et al.*, 2019; 2020]. Nevertheless, only using optical flow in the bounding boxes [Styles *et al.*, 2019; 2020] can not effectively compensate the motion from the ego-vehicle. It also sustains the interference from irrelevant motion in the scene.

Apart from that, current RNNs [Rasouli *et al.*, 2019; Dendorfer *et al.*, 2020] or CNNs [Styles *et al.*, 2019; 2020] approaches have been widely applied to relevant tasks and have achieved promising progress. However, CNNs fail to model the long-term dependencies due to the limited receptive field, and RNNs are usually flawed in extracting local sequence patterns [Wang, 2018] which sometimes contain key clues for predicting future. Moreover, in fusion mechanism, most existing networks directly merge the features from different modalities through a simple concatenation. The lack of mining characteristics and relations of distinct modalities makes these approaches hardly capture the interaction between various granular motion features and produce redundant parameters that are limited to deploy on vehicle platforms with less computing resources.

To address such limitations, we propose a Multimodal

Transformer Network (MTN), which integrates the observed trajectory, ego-vehicle speed and optical flows to predict future pedestrian trajectory. Owing to the relations between observed trajectory of target and ego-vehicle speed sequence, a novel coarse-grained fusion stage firstly processes the two modalities to produce a hybrid representation through a co-attentional mechanism. The inspiration comes from vision-language tasks [Lu *et al.*, 2019]. Next, a fine-grained fusion stage integrates the hybrid results of the former stage with the motion representations of pedestrians and ego-vehicle. The latter can provide fine-grained dynamic motion information and is obtained when we process separated patches of the optical flow in the center area and target pedestrians in parallel. This fusion stage can also avoid interference from the motion of irrelevant objects. Finally, MTN outputs future locations of the target in parallel in one time. The whole network is only attention-based, which can efficiently model long-term sequences and better capture the local temporal variations through a coarse-to-fine manner.

The effectiveness of our method is evaluated on the two largest datasets with dense pedestrian bounding box annotations, PIE [Rasouli *et al.*, 2019] and JAAD [Rasouli *et al.*, 2017], under the benchmark of [Rasouli *et al.*, 2019]. Experimental results demonstrate that our method achieves state-of-the-art performance with the fewest parameters.

In summary, the main contributions of this paper can be summarized as follows:

1) The introduction of the center area and target pedestrian optical flow compensates the highly dynamic motion between the ego-vehicle and pedestrians by dividing them into patches and processing in parallel.

2) The proposed MTN integrates multiple modalities at distinct stages according to their granularity to more effectively capture of the highly dynamic motion information. In addition, the MTN takes advantages of attention-based architecture to efficiently model long-range temporal dependencies with much fewer parameters.

2 Method

In this section, we describe the details of our method which include the optical flow representations, the multimodal transformer architecture, and a warm-up training strategy.

2.1 Optical Flow Representations

The center area and target boxes of optical flows imply ego-vehicle and pedestrian motion. Both of them compensate the highly dynamic motion by dividing flows into patches and applying a spatial average pooling on them due to the local smoothness. As Fig. 1 shows, for the t-th frame of the optical flows, a Region Of Proposal (ROI) ϕ^t_{ego} with shape $(2, H_{ego}, W_{ego})$ is cropped at the center. Then, ϕ^t_{ego} is split into M patches with equal area, each patch owns the shape of $(2, \lfloor \frac{H_{ego}}{\sqrt{M}} \rfloor, \lfloor \frac{W_{ego}}{\sqrt{M}} \rfloor)$ and may contain specific motion. Next, the i-th patch $\phi^{t,i}_{ego}$ is operated by a spatial average pooling to generate a vector $\overline{\phi^{t,i}_{ego}} \in \mathbb{R}^{2\times 1}$. After repeating the above operations with a fixed H_{ego} and W_{ego} for each frame of the optical flows, M vectors of each frame are concatenated at the first dimension and results in the motion representation of ego-vehicle $\phi_e \in \mathbb{R}^{2(T-1)\times M}$.

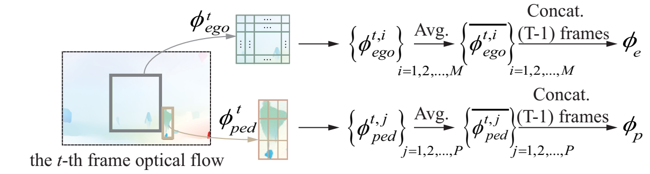


Figure 1: Optical flow representations. At each time step, optical flows from center area and pedestrian area are divided into patches and processed in parallel.

Optical flow is also exploited to compensate the dynamics of pedestrians, like changing direction rapidly. For the t-th frame of optical flows, a ROI ϕ^t_{ped} with shape $(2, H^t_{ped}, W^t_{ped})$ is firstly extracted, where H^t_{ped} and W^t_{ped} are chosen from the bounding box annotation of the target in the frame t. Next, ϕ^t_{ped} is spatially divided into P patches $\left\{\phi^{t,j}_{ped}\right\}_{j=1,2,\dots,P}$, and the motion representation of target pedestrian $\phi_p \in \mathbb{R}^{2(T-1)\times P}$ is obtained after the same processes like ϕ_e. Finally, ϕ_e and ϕ_p incorporate the fine-grained dynamic motion and will be merged by the multimodal transformer network.

2.2 Multimodal Transformer Network

As is shown in Fig. 2, MTN consists of a coarse-grained fusion stage and a fine-grained fusion stage. The former stage merges trajectory and speed sequences by a co-attentional mechanism. The latter stage fuses the former results and representations from optical flows to estimate the future trajectories. Following notions are used: $L_{obs} \in \mathbb{R}^{T\times 4}$ and $S_{obs} \in \mathbb{R}^{T\times 1}$ represent the observed trajectory and the speed sequence, where T is the length of the observation sequence and the 4 dimensions of L_{obs} are defined by top-left coordinate and bottom-right coordinate. ϕ_e and ϕ_p indicate fine-grained motion representations of ego-vehicle and pedestrians as described in section 2.1.

Coarse-grained fusion. Ego-vehicle speed is usually closely related to target trajectory. For example, the trajectory usually changes rapidly when the ego-vehicle is driving at a high speed. Due to such property, the coarse-grained fusion stage combines the observed trajectory with ego-vehicle speed through a co-attentional mechanism and outputs a hybrid representation which contains the relative motion at a coarse granularity. As is illustrated in the top row of Fig. 2, the coarse-grained fusion stage includes two fully connected layers and three blocks that are linked sequentially. Each block consists of a self-attention module, two cross-attention modules and two feed-forward layers. Giving input trajectory L_{obs} and ego-vehicle speed S_{obs}, the coarse-grained fusion stage separately sends them into two independent fully connected layers. For L_{obs}, an initial location representation with shape (T, C) is generated by a linear transformation and adding the positional embeddings like [Vaswani *et al.*, 2017] to provide the order of observed locations. For S_{obs}, a fully connected layer transforms speed sequence into

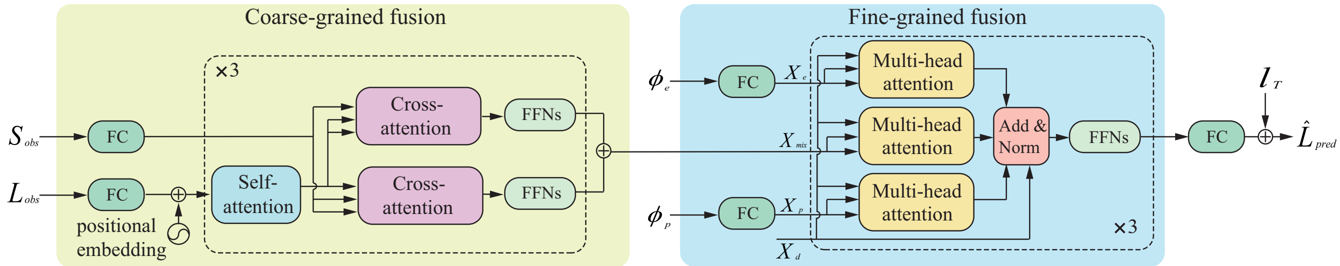


Figure 2: Overall structure of MTN. The MTN is a transformer-based network that consists of two stages to process and interacts modalities at different granularity respectively. The $\oplus$ denotes matrix addition.

a C-dimensional space to capture the overall speed variation patterns by producing a vector with shape $(1, C)$. After that, the initial location representation from L_{obs} is sent into a self-attention module to extract the long-term temporal dependencies. Then two cross-attention modules are utilized to compute cross-correlations between speed and trajectory in a co-attentional mechanism. Specifically, the input of each cross-attention module in Fig. 2 is query, key, and value matrices from top to bottom. Co-attentional mechanism computes query matrices from their own modalities whereas calculates key and value matrices from opposite modalities to perform cross-attention like [Carion *et al.*, 2020]. Next, two intermediate representations that contain the trajectory and speed patterns are generated through separate feed forward layers. After transmitting the intermediate representations into the remaining blocks, the coarse-grained fusion stage finally generates a coarse-grained motion representation $X_{mix} \in \mathbb{R}^{T\times C}$ by adding up the outputs from the last block.

Fine-grained fusion. Fine-grained fusion stage exchanges information between the coarse-grained motion representation X_{mix} and fine-grained motion representations of ego-vehicle ϕ_e and pedestrians ϕ_p which compensate for the lack of highly dynamic motion. Concretely, the fine-grained fusion stage contains three blocks and two fully connected layers. Each block consists of three multi-head attention modules, an add & norm layer and a feed forward layer. Giving X_{mix}, ϕ_e and ϕ_p as inputs, X_e and X_p are firstly generated by projecting ϕ_e and ϕ_p into a C-dimensional space through two independent fully connected layers. The first block expects X_{mix}, X_e, X_p and a trajectory query X_d as inputs, where $X_d \in \mathbb{R}^{N\times C}$ is a sinusoidal embedding since the future locations are fixed chronologically and N is the length of the prediction sequence. Each multi-head attention block takes in charge of interacting X_d with the corresponding representation. Specifically, the input of each multi-head attention module from top to bottom is query, key and value matrices. The attention mechanism mainly extracts the most dependent motion information for the query matrix X_d. For example, it can capture ego-vehicle motion more sensitively and also suppress interference from other factors such as irrelevant motion. After that, the output of each multi-head attention module are added together with X_d,followed by a layer normalization and a feed-forward layer to generate an intermediate representation. Then the intermediate representations are delivered into the left blocks with X_e, X_p and X_{mix} to proceed as before. Finally, the output of the last block is delivered into a fully connected layer, and added by the last observed location l_T to form the trajectory prediction result $\hat{L}_{pred} \in \mathbb{R}^{N\times 4}$.

2.3 Training

At training stage, we adopt mean squared error (MSE) loss function for training our MTN:

$$Loss = \frac{1}{N}\sum_{t=1}^{N} \|\hat{l}_{T+t} - l_{T+t}\|^2, \tag{1}$$

where $\hat{l}_{T+t}$ is the t-th location of $\hat{L}_{pred}$ and l_{T+t} represent corresponding ground truth.

To make training converge more stable and faster, we apply a warm-up training strategy. Specifically, we remove the modules related to ego-vehicle speed in the coarse-grain fusion stage and the modules related to optical flow in the fine-grained fusion stage. The remaining components of MTN are firstly pre-trained by only taking L_{obs} as input for a few epochs. Next, MTN is initialized by the pre-trained model and completes the training process after specified epochs.

3 Experiments

Datasets. We evaluate MTN on Pedestrian Intention Estimation (PIE) [Rasouli *et al.*, 2019] and Joint Attention in Autonomous Driving (JAAD) [Rasouli *et al.*, 2017] datasets. The PIE consists of 1, 842 pedestrian tracks and 909, 480 bounding boxes in 37 videos, recorded by a HD (1080×1920, 30 fps) camera from a front-view in Canada during daytime. It also provides dense frame-wise bounding box annotations and ego-vehicle information. For a fair comparison, we adopt the same kind of ego-vehicle sensor information *e.g.* vehicle speed and train/test splits as in [Rasouli *et al.*, 2019]. The JAAD includes 2, 856 pedestrian tracks and 82, 032 frames in 346 video clips. We apply the same train/test split as in [Rasouli *et al.*, 2019].

Evaluation metrics. The Mean Squared Error (MSE) is the commonly used evaluation metric. MSE computes each time step's average similarity between the predicted bounding box and ground truth. Besides, C_{MSE} and CF_{MSE} are also adopted to evaluate similarity over the spatial location and long-term prediction. C_{MSE} represents the MSE between

		PIE			JAAD		
Method	Para.	MSE	C_{MSE}	CF_{MSE}	MSE	C_{MSE}	CF_{MSE}
B-LSTM	-	855	811	3259	1535	1447	5615
DTP-MOF	11.30	665	566	2373	1158	1014	4143
PIE$_{full}$	3.07	559	520	2162	-	-	-
PIE$_{traj}$	1.24	636	596	2477	1248	1183	4780
STED	13.94	461	415	1871	1044	960	4031
MTN$_{traj}$	**0.11**	581	547	2278	1231	1177	4644
MTN	0.13	**444**	**414**	**1627**	**1005**	**951**	**4010**

Table 1: Quantitative comparison on PIE dataset and JAAD datasets. The number of parameters (Para.) is displayed in M (million).

the center of the predicted bounding box and the ground truth. CF_{MSE} is the C_{MSE} at the last time step. All prediction results are given in pixels. The parameters of different methods also attend in our comparison to evaluate the deployment potential.

Implementation details. Samples of JAAD and PIE are generated following [Rasouli *et al.*, 2019]. For each sample, we employed RAFT [Teed and Deng, 2020] to extract optical flow per frame, and downsample the results by 2 times. The height H_{ego} and width W_{ego} of the center ROI are set to be 160 pixels, and the number of patches M and P are 64 and 9. Each patch owns the same area. The length of observation sequence T is set to be 15 frames (0.5s) and the length of prediction sequence N is 45 frames (1.5s). The number of total training epoch is 80, and ten epochs are used to warm up parts of the MTN as Sec. 2.3 states. The number of batch size is 128 and the Adam optimizer [Kingma and Ba, 2015] is used. All experiments are conducted on a single GTX 2080Ti. Since the JAAD dataset do not provide odometry information and most samples have high visibility, we remove the components related to the ego-vehicle speed and replace the residual term l_T with the locations of linear prediction like [Styles *et al.*, 2019; 2020]. In the following sections, we take B-LSTM [Bhattacharyya *et al.*, 2018], DTP-MOF [Styles *et al.*, 2019], PIE$_{full}$ [Rasouli *et al.*, 2019], PIE$_{traj}$ (the baseline version of PIE$_{full}$ which only takes trajectory as input), and STED [Styles *et al.*, 2020] as the comparative state-of-the-art methods. For DTP-MOF and STED, the length of input and output sequences are changed for a fair comparison. Besides, the original DTP-MOF only considers the centroid of bounding boxes. Thus we change it by training and predicting using bounding boxes. Moreover, we introduce MTN$_{traj}$, a baseline version of MTN which only takes L_{obs} as input. MTN$_{traj}$ is a simple encoder-decoder structure. The encoder is a transformer encoder which contains three blocks. The decoder is also composed of three blocks and each block consists of a cross-attention module and a feed-forward layer. Then the output of the decoder is sent into a fully connected layer and added by l_T to obtain predictions just like MTN.

3.1 Comparisons with State-of-the-art Methods

Tab. 1 shows the results on PIE and JAAD benchmarks. Compared to state-of-the-art optical flow-based method STED, our MTN outperforms it by **17** and **39** MSE on PIE and JAAD respectively with only **107** $\times$ fewer parameters. With the single trajectory modality, our MTN$_{traj}$ also outperforms the PIE$_{traj}$ **55** and **17** MSE with only **11** $\times$ fewer parameters. After introducing optical flows, the MTN further stretches the advantage than PIE$_{full}$, which shows the optical flow is better than the representation of semantic intention for trajectory prediction. Moreover, considering the C_{MSE} and CF_{MSE} which evaluate the locating and long-term modeling ability, our MTN also shows the best performance. Fig. 3 demonstrates detailed comparisons of the MSE with the increasing time-steps. Two pivots are observed: (1) for each dataset, the MSE of our method is kept low at all time steps; (2) more importantly, our method greatly outperforms other approaches in terms of the accuracy of long-term prediction. The visualization of trajectory prediction results is shown in Fig. 4. Our method generates more reasonable predictions under various situations, especially for the dynamic motion of ego-vehicle and pedestrians. This is attributed to (1) optical flow provides motion of pedestrians and vehicles in a fine-grained level, which compensates for the absence highly dynamic information more effectively; (2) the attention mechanism can better capture temporal relations of the sequence in a local-global manner, as we are going to discuss in details in the ablation experiments.

L_{ego}	S_{ego}	M_{ego}	Ped	MSE	C_{MSE}	CF_{MSE}
				537	506	2041
✓				477	445	1835
	✓			465	433	1771
		✓		451	420	1748
			✓	453	422	1790
		✓	✓	**444**	**414**	**1627**

Table 2: Investigation of selecting different areas of optical flow on PIE dataset. L_{ego}, S_{ego} and M_{ego} indicate different areas of the extracted area. Ped refers to extract from the target pedestrian area.

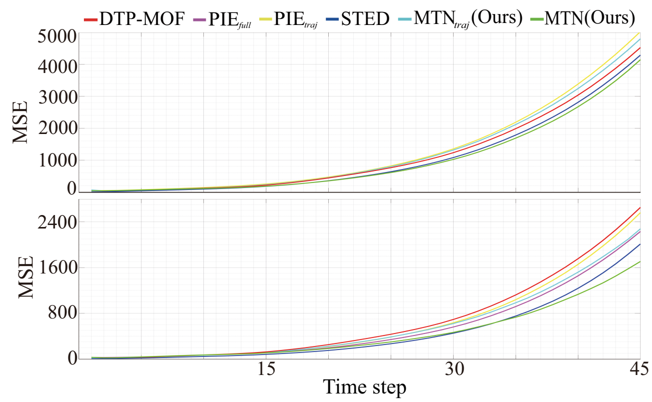


Figure 3: Visualization of MSE variations with increasing time steps on JAAD (top) and PIE (bottom) datasets.

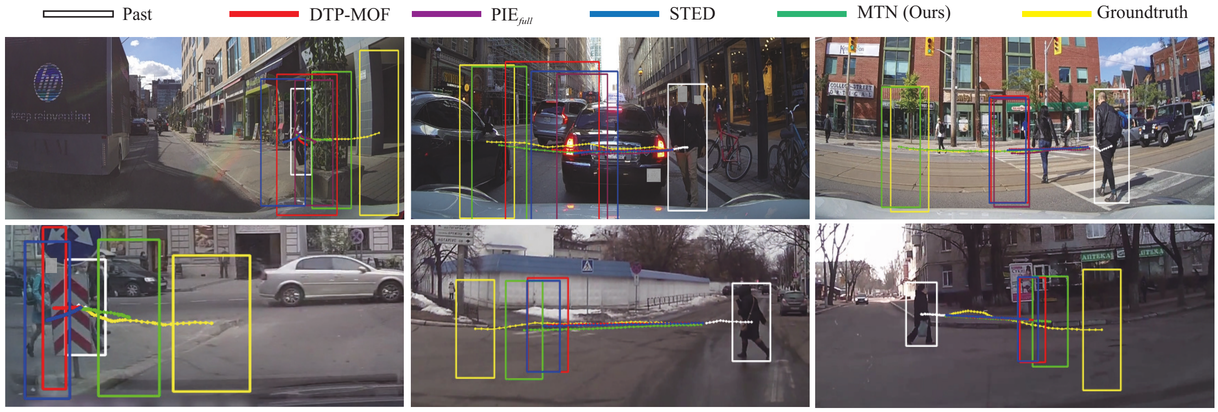
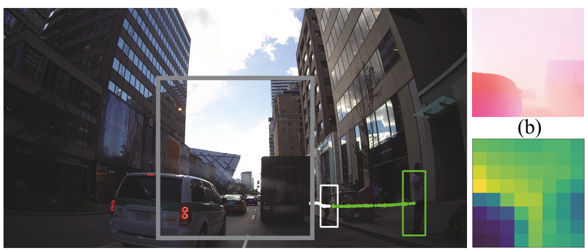


Figure 4: Qualitative results on PIE (top) and JAAD (bottom) datasets. Each white bounding box illustrates the target location of the first frame, and each white line shows the observed trajectory. Other colored boxes represent the final predicted location and colored lines demonstrate the prediction trajectories of different methods. Images are cropped for better visibility.



(a) (b) (c)

Figure 5: Effect of introducing optical flows of ego-vehicle. (a) shows the last observed frame, and the grey, white and green boxes show the interested region, current location and the final predicted location of target pedestrian. (b) illustrates the extracted optical flows from the interested region. (c) shows the attention map learns to block out irrelevant motions and compensate motion caused by ego-vehicle.

Method	MSE	C_{MSE}	CF_{MSE}
Concatenation	567	536	2161
Addition	556	525	2137
Co-attention	**537**	**506**	**2041**

Table 3: Evaluation of different trajectory-speed fusing methods on PIE dataset. Components related to optical flows are removed for more rigor.

3.2 Ablation Study

In this section, we first evaluate the effect of optical flow from disparate regions and different representing models. Then the impact of merging methods in the coarse-grained fusion stage is discussed. After that, we analyze how coarse-grained motion information guides trajectory prediction from intermediate attention maps. Finally, we explore the model complexity from two aspects: (1) number of blocks; (2) selection of embedding size C, and show some failure cases.

Selected area of optical flows. Tab. 2 shows the benefits from different optical flow areas. The use of pedestrian optical flow obtain 84 MSE reduction. Also, the optical flow of ego-vehicle reduces MSE a lot. To evaluate the impact of different height H_{ego} and width W_{ego} of the center ROI, we also set three different sizes (large $L_{ego} = (260 \times 260)$, medium $M_{ego} = (160 \times 160)$, small $S_{ego} = (60 \times 60)$) and the final results show with medium area, the best MSE reduction achieves 86. Fig. 5 visualizes the effect of the fine-grained motion representation ϕ_e. Fig. 5(a) and (b) show the selected area which is fixed to the lower location of image center area and the captured optical flows of ego-vehicle, respectively, which contains the ego-vehicle motion and movement of other objects in the scene, *e.g.* the white van at bottom-left. Fig. 5(c) shows the attention map (the darker the map, the lower the attention) ignores the irrelevant movement of the other vehicles to compensate for the real motion caused by ego-vehicle.

Investigation of optical flow representations. This part explores the different models to represent the fine-grained motion information. A common approach [Styles *et al.*, 2019] is to exploit a CNN to extract motion features from stacked optical flows. Here we use the Resnet-18 to process them and generate the fine-grained motion representations. Following DTP [Styles *et al.*, 2020], Resnet-18 is firstly pre-trained to learn a compensation term of constant velocity assumption , and the parameters of the pre-trained network is fixed when training MTN. Test results on PIE dataset of CNN structure are 460 MSE, 429 C_{MSE} and 1851 CF_{MSE} at the cost of 11.41M parameters. Compared to the best applied results using our method in Tab. 2, CNN produces more parameters without any performance improvement, which appears that a fully connected layer is sufficient to extract the motion features from optical flows.

Fusion methods in the coarse-grained fusion stage. We discuss the performance of different fusion methods between ego-vehicle speed and target trajectory in the coarse-grained fusion stage. In this part, components related to the optical flow representations are not applied. (1) Concatenation.

Blocks	Para.	MSE	C_{MSE}	CF_{MSE}
1	**0.05**	500	470	1823
3	0.13	**444**	**414**	**1627**
5	0.22	474	448	1738

Table 4: Investigation of the number of blocks in MTN. The number of parameters (Para.) is displayed in M (million). We set the embedding size C to 32. The coarse-grained fusion stage and the fine-grained fusion stage contain the same number of blocks.

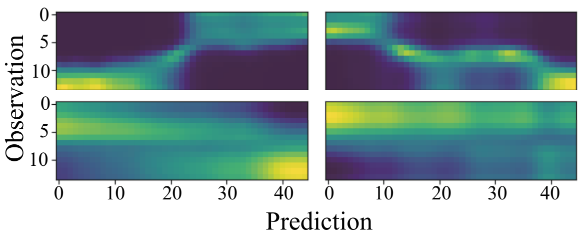


Figure 6: Attention maps between coarse-grained motion representations and the predicted trajectory query in the first (top) and third (bottom) blocks. The left and right columns visualize the maps from the first and fourth attention heads.

Past trajectory L_{obs} is concatenated to ego-vehicle speed S_{obs} after a fully connected layer and a 3-layers transformer encoder, which forms the coarse-grained motion representation $X_{mix} \in \mathbb{R}^{T\times(C+1)}$. There is another fully connected layer in the fine-grained fusion stage to project X_{mix} into a C-dimensional space. Such process is similar to [Rasouli *et al.*, 2019], except we replace future ego-vehicle speed with the observed speed. (2) Addition. Ego-vehicle speed sequence S_{obs} is embedded by a fully-connected layer and then added with the output of the three transformer encoder blocks which process L_{obs}. As is shown in Tab. 3, compared with simple concatenation, passing the vehicle speed through a fully connected layer ameliorates the performance slightly by 11 MSE. Our co-attention method further improves the MSE by 19.

Attention maps between motion representation X_{mix} and trajectory query X_d. Fig. 6 visualizes attention maps to show how fusion mechanism works. The first block (top row) mainly focuses on the local relation between X_{mix} and X_d. In detail, the first and fourth heads separately concentrate on the recent and long ago observation. After subsequent blocks, the last block tends to capture global temporal context to supplement completeness of pedestrian motion by aggregating the whole-time observation.

Number of fusion blocks. To investigate the influence of model complexity, we change the number of blocks in the distinct fusion stages. As Tab. 4 shows, MTN with only one block owns a high prediction error with $0.05M$ parameters. When the number of blocks increases to 3, the prediction error reduces by 56 at the cost of an increase of $0.08M$ parameters. However, the addition of another two blocks raises the MSE by 30, which is caused by the imbalance between the expressive relations between different modalities and the representation capacity of a deeper network.

Embedding size	Para.	MSE	C_{MSE}	CF_{MSE}
16	**0.03**	532	500	2024
32	0.13	444	414	**1627**
64	0.41	**439**	**411**	1688

Table 5: Investigation of different embedding sizes. The number of parameters (Para.) is displayed in M (million). The number of attention heads and dimensionality of inner layers in FFNs are fixed to 4 and 4 $\times$ embedding size.

Figure 7: Failure cases. The failure forecasts are often caused by randomness of 2D trajectory mutations during prediction period. The same color coding (see Fig. 4) is used.

Embedding size C. We also explore the impact of embedding size C. As Tab. 5 illustrates, the improvement of prediction performance is significant (88 MSE reduction) when embedding size C is raised from 16 to 32, but larger embedding size 64 does not bring more meaningful benefits. To obtain the best trade-off between prediction error and computational resource consumption, we set the embedding size C as 32.

Failure cases. Failure cases are shown in Fig. 7, due to randomness of 2D trajectory mutations during prediction period. For example, ego-vehicle is braking in the left situation, or the pedestrian is changing his direction in the right case.

4 Conclusion

In this work, we have developed a multimodal transformer network to predict pedestrian trajectory by introducing optical flows to compensate highly dynamic motion between ego-vehicle and pedestrians. The whole architecture is only-attention-based and consists of two specially stages to process and merge coarse-grained and fine-grained modalities. The coarse-grained fusion stage models the temporal similarity between vehicle speeds and pedestrian trajectory to aggregate a coarse-grained motion representation. The fine-grained fusion stage interacts the fine-grained motion representations, which are extracted from the observed ego-vehicle and pedestrian optical flows, with the former coarse features to compensate the highly dynamic motion. This architecture takes the advantage of attention mechanism to model the long-range dependencies more efficiently than the common convolution and recurrent operations, thus achieving a considerable reduction of overall prediction error. In future work, it would be interesting to employ the semantic understanding of traffic scene to further improve the performance by considering more complex interactions with other objects.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62088102, 61976170, 91648121).

References

[Adeli *et al.*, 2020] Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics Autom. Lett.*, 5(4):6033–6040, 2020.

[Alahi *et al.*, 2016] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Fei-Fei Li, and Silvio Savarese. Social LSTM: human trajectory prediction in crowded spaces. In *CVPR*, pages 961–971, 2016.

[Bhattacharyya *et al.*, 2018] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *CVPR*, pages 4194–4202, 2018.

[Cao *et al.*, 2020] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV (1)*, volume 12346, pages 387–404, 2020.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV (1)*, volume 12346, pages 213–229, 2020.

[Dendorfer *et al.*, 2020] Patrick Dendorfer, Aljosa Osep, and Laura Leal-Taixé. Goal-gan: Multimodal trajectory prediction based on goal position estimation. *CoRR*, abs/2010.01114, 2020.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019.

[Makansi *et al.*, 2020] Osama Makansi, Özgün Çiçek, Kevin Buchicchio, and Thomas Brox. Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior. In *CVPR*, pages 4353–4362, 2020.

[Malla *et al.*, 2020] Srikanth Malla, Behzad Dariush, and Chiho Choi. TITAN: future forecast using action priors. In *CVPR*, pages 11183–11193, 2020.

[Mangalam *et al.*, 2020] Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision. In *WACV*, pages 2773–2782, 2020.

[Oliu *et al.*, 2018] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *ECCV (14)*, volume 11218, pages 745–761, 2018.

[Rasouli *et al.*, 2017] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior. In *ICCV Workshops*, pages 206–213, 2017.

[Rasouli *et al.*, 2019] 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 *ICCV*, pages 6261–6270, 2019.

[Rehder *et al.*, 2018] Eike Rehder, Florian Wirth, Martin Lauer, and Christoph Stiller. Pedestrian prediction by planning using deep neural networks. In *ICRA*, pages 1–5, 2018.

[Saleh *et al.*, 2019] Khaled Saleh, Mohammed Hossny, and Saeid Nahavandi. Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal densenet. In *ICRA*, pages 9704–9710, 2019.

[Schneemann and Heinemann, 2016] Friederike Schneemann and Patrick Heinemann. Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In *IROS*, pages 2243–2248, 2016.

[Styles *et al.*, 2019] Olly Styles, Arun Ross, and Victor Sanchez. Forecasting pedestrian trajectory with machine-annotated training data. In *IV*, pages 716–721, 2019.

[Styles *et al.*, 2020] Olly Styles, Tanaya Guha, and Victor Sanchez. Multiple object forecasting: Predicting future object locations in diverse environments. In *WACV*, pages 679–688, 2020.

[Teed and Deng, 2020] Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV (2)*, volume 12347, pages 402–419, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Wang, 2018] Baoxin Wang. Disconnected recurrent neural networks for text categorization. In *ACL (1)*, pages 2311–2320, 2018.

[Wichers *et al.*, 2018] Nevan Wichers, Ruben Villegas, Dumitru Erhan, and Honglak Lee. Hierarchical long-term video prediction without supervision. In *ICML*, volume 80, pages 6033–6041, 2018.

[Wu *et al.*, 2020] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *CVPR*, pages 5538–5547, 2020.

[Ye *et al.*, 2019] Yufei Ye, Maneesh Singh, Abhinav Gupta, and Shubham Tulsiani. Compositional video prediction. In *ICCV*, pages 10352–10361, 2019.