STrajNet: Occupancy Flow Prediction via Multi-modal Swin Transformer

Haochen Liu, Zhiyu Huang, Chen Lv
School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore
haochen002@e.ntu.edu.sg, zhiyu001@e.ntu.edu.sg, lyuchen@ntu.edu.sg

Abstract

*Making an accurate prediction of occupancy and flow is essential to enable better safety and interaction for autonomous vehicles under complex traffic scenarios. This work proposes **STrajNet**: a multi-modal Swin Transformer-based framework for effective scene occupancy and flow predictions. We employ Swin Transformer to encode the image and interaction-aware motion representations and propose a cross-attention module to inject motion awareness into grid cells across different time steps. Flow and occupancy predictions are then decoded through temporal-sharing Pyramid decoders. The proposed method shows competitive prediction accuracy and other evaluation metrics in the Waymo Open Dataset benchmark.*

1. Introduction

Accurately forecasting the future motion of multiple traffic participants (agents) is one of the most challenging tasks for autonomous driving. A novel and effective representation for motion prediction is the recently-proposed occupancy flow fields [5], which consist of future occupancy grid maps warped by backward motion flow, constructing a spatial-temporal grid set accompanied by the corresponding flow. Prediction of occupancy flow fields captures rich distributions of traffic participants' future motion with uncertainties, maintaining the track-and-trace ability for every participant through the predicted flow. It also helps promote safety and efficiency for perception and planning. Importantly, predicting the occupancy flow fields requires a global reception field of the driving scene, as well as a mechanism able to incorporate the motion of each traffic participant into the occupancy grid. However, the challenge is how to handle the diverse traffic agents and traffic elements in the driving scene, and how to model the underlying interactions among all agents.

To tackle these challenges, we propose a multi-modal Transformer-based prediction framework. The recent prevalence of Transformer-based structures in computer vision and motion prediction tasks is primarily due to its ability in graph-like interaction modeling through multi-head attention with global fields, as well as its superior computational efficiency. For instance, [8] explicitly models interactions among agents and map segments across time steps by unified Transformer. Graph attention-based interaction modeling among agents [6] and edges [7] also facilitates the capabilities of motion forecasting. [9] utilized a Transformer network to incorporate both spatial and temporal interactions for flow estimation. In our work, a carefully designed Swin Transformer [4] encoder is proposed to capture the interactions among image patches of historical occupancy, backward flow, and dense road map; and vectorized historical motion trajectories are separately encoded by a trajectory Transformer considering interaction awareness. To better associate occupancy grid cells with corresponding agent trajectories to capture their motion tendencies, we propose a cross-attention module that queries the encoded interaction trajectories of each grid cell across future steps.

2. Method

2.1. Problem formulation

Occupancy flow fields prediction entails a multi-task objective $\hat{\mathbf{Y}}$ that simultaneously predicts future keyframes of observed occupancy $\hat{O}_k^b$, occluded occupancy $\hat{O}_k^c$, and corresponding backward flow $\hat{F}_k$ at different timesteps $k \in [1, T_f]$ ($T_f = 8$ the number of future frames sampled at 1Hz), given the past and current states of traffic agents and environmental context. More specifically, an occupancy grid can be regarded as a binary single-channel image $\hat{O}_k^b, \hat{O}_k^c \in R^{H \times W \times 1}$, and the backward flow as a two-channel image-like tensor for motion shifting along x and y-axis: $\hat{F}_k \in R^{H \times W \times 2}$. In our framework, the state input $\mathbf{X}$ consists of multiple modalities (detailed in Sec. 2.2): historical temporal frames of vehicle-only occupancy $O_t, t \in [-T_h, 0]$ ($T_h = 10$ the number of historical frames sampled at 10Hz), dense road map $\mathcal{M}$, one frame of historical flow F_h, as well as vectorized trajectories of n agents $\mathcal{S} = \{S_1, S_2, \cdots, S_n\}$. S_i represents a sequence of motion states ($s_t^i \in S_i$) over historical and current timesteps $t \in [-T_h, 0]$. Mathematically, the problem is formulated as:

$$\mathbf{X} = [\{O_t | t \in [-T_h, 0]\}; \mathcal{M}; \mathcal{S}; F_h],$$
$$\hat{\mathbf{Y}} = \{(\hat{O}_k^b, \hat{O}_k^c, \hat{F}_k) | k \in [1, T_f]\}. \quad (1)$$

2.2. Data processing

We employ the Waymo Open Motion dataset [1] in the experiment, which consists of over 500,000 samples covering diverse driving scenarios and dynamical interactions among traffic agents including vehicles, cyclists, and pedestrians. The historical timesteps are sampled at 10Hz for the past one second, and the objective is to predict keyframes over future eight seconds at 1Hz.

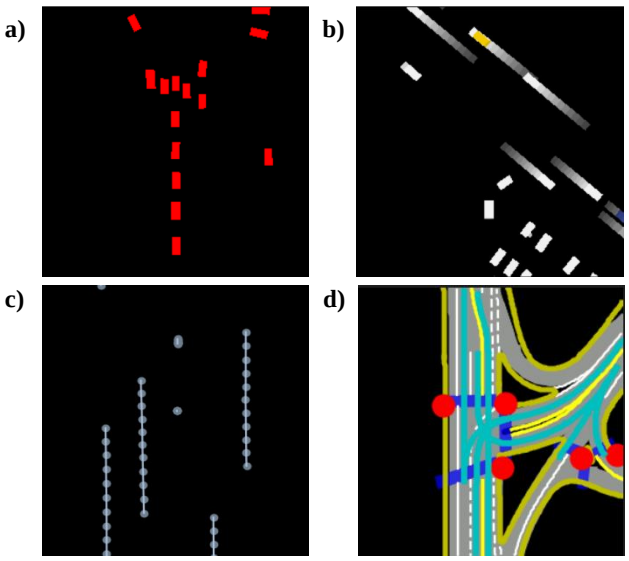

Figure 1. Multi-modal input representations: a) historical vehicle occupancy grid map; b) historical backward flow; c) vectorized historical motion trajectories; d) dense road map.

To construct the multi-modal input, as shown in Fig. 1, all of the vectorized data points are firstly normalized according to the ego vehicle's current status. For image-based inputs, the **vehicle historical and current occupancy** O_t is processed and stacked by the standard pipeline in [5]. The **environmental context** of the road network and traffic light states $\mathcal{M}$ are densely rasterized as an RGB image. Different types of road segments are associated with a color map and rendered according to their sizes. For instance, yellow represents road border lines, blue for crosswalks, and grey for drivable lanes. Current states of traffic lights are plotted as colored circles (red, yellow, and green) centered by their corresponding locations. **Historical backward flow** F_h is a single-frame flow field constructed from the (d_x, d_y) displacement of agents in the occupancy grid between the two timesteps $-T_h, 0$. The size of the image-based input is 256×256, encoding the driving scene with an area of $80m \times 80m$, which is the same as the prediction output.

For vectorized inputs of **historical motion trajectories** $\mathcal{S}$, we collected all the agents currently inside the grid. The agents in $\mathcal{S}$ are sorted by their distances to the current ego agent and only $n = 64$ agents are kept. For each timestep, the motion state of an agent $s_t^i \in S_i$ is represented in a tuple format: $s_t^i = (x, y, v_x, v_y, \theta)$, meaning normalized locations, speed and yaw angle. We also maintained the one-hot encoding of traffic participant's type (vehicle, cyclist, or pedestrian) $\mathbf{Tp}_i$ for each agent $\mathcal{S}_i$. The vector inputs are organized into a 3D-tensor with shape $64 \times 11 \times 5$ (agents, timesteps, features) and 64×3 for agent types.

2.3. STrajNet

As illustrated in Fig. 2, the overall structure of STrajNet is an end-to-end multi-modal framework comprising 4 fundamental modules for occupancy and flow prediction across future timesteps.

1) Swin Transformer-based Image Encoder: For image-based representations, the visual features of various scales are encoded through Swin Transformer-based encoder. A separated Swin Transformer block is designed for historical flow F_h for information "shortcut" directly to future flow prediction heads. All of the vehicle occupancy O_t, dense RGB map $\mathcal{M}$ and flow F_h are initially embedded and down-sampled into $H \times W \times C$ by separated 4×4 convolution kernels with a stride of 4. We follow the Swin-Transformer (Tiny) [4] and set the embedded dimension as $C = 96, H, W = 64$. More specifically, each Swin-Transformer module is a two-layer Transformer with both global window self-attention (WMSA) and shifted window attention (SWMSA). It enables global and intersected attention-based interaction modeling for visual features. Each attention module is multi-head attention with relative positional encoding bias $\mathbf{B}$:

$$\mathrm{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\mathrm{head}_1 \,||\, \cdots \,||\mathrm{head}_\mathbf{h})W^O,$$
$$\mathrm{head}_i = \mathrm{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d_k} + \mathbf{B})\mathbf{V}, \quad (2)$$

where the head numbers are $\mathbf{h} = [3, 6, 12]$ as the module goes deeper, and d_k is the dimension of the key token. The image encoder outputs sets of visual features in varying scales: $64 \times 64 \times 96$, $32 \times 32 \times 192$, and $16 \times 16 \times 384$.

2) Trajectory Encoder: Trajectories features are encoded considering interaction awareness. Historical motion features for each agent are firstly aggregated across time by 4-head MSA with global max-pooling, concatenated with its embedded agent type through the dense layer. Next, a 6-head self-attention layer (Fig. 3(b)) with residual connection is introduced to build an interaction graph among all currently present agents. The latent dimensions are kept the same as 384 for all layers.

3) Trajectory-aware Cross-attention: The purpose is to associate each grid cell with trajectory information of presented agents, so that the information is more directed and

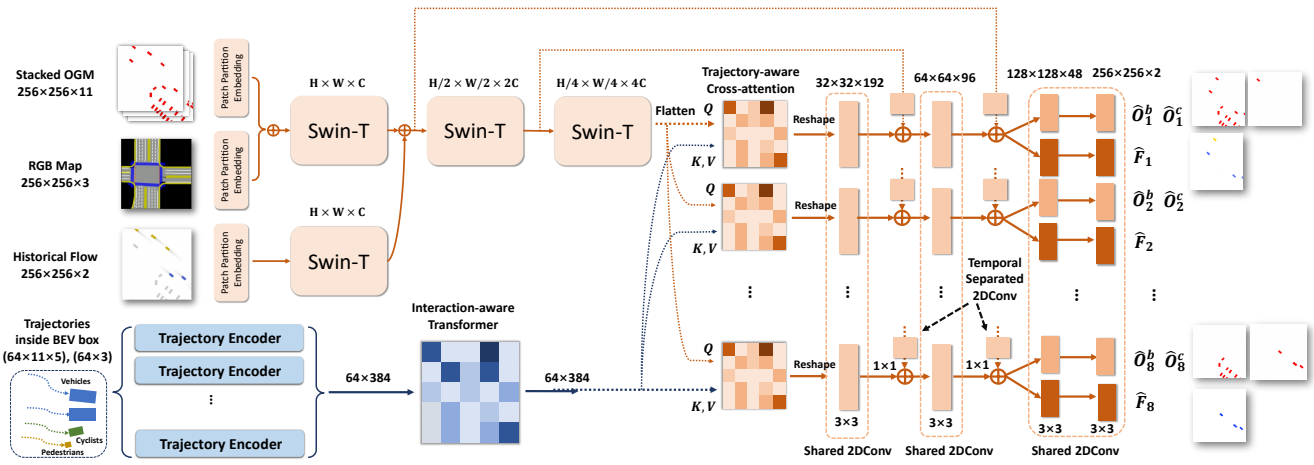

Figure 2. An overview of STrajNet. Multi-modal inputs are separately encoded by Swin-Transformer encoders and trajectory encoders; trajectory-based interaction awareness of each grid cell is encoded by cross-attention and decoded by a shared feature pyramid decoder to predict occupancy and flow for each timestep.

no longer constrained by patches or nearby features. As shown in Fig. 3(c), the highest level visual features are flattened by each pixel to $H/4 \cdot W/4 \times 384 = 256 \times 384$, and used as queries. Keys and values are the output trajectory features from the trajectory encoder. To query different trajectories interactions for different future timesteps, we implement 8 separated cross-attention Transformer modules, each with 3 heads. The outputs are then reshaped back to image shape with the same latent dimension: $16 \times 16 \times 384$.

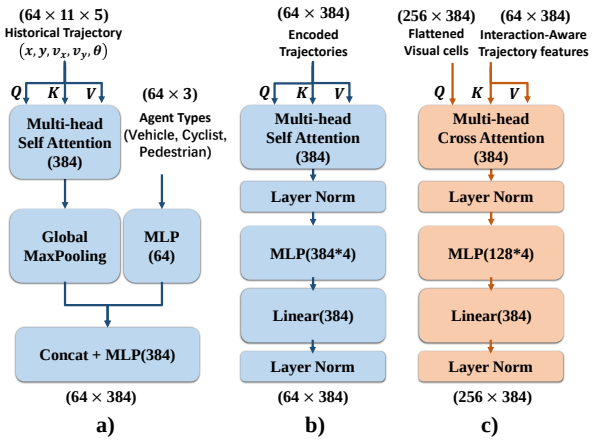

Figure 3. Structures of a) Trajectory Encoder; b) Interaction-aware Transformer; c) Trajectory-aware Cross-attention.

4) Pyramid Decoder: A feature pyramid network [2] decoder is utilized to decode the occupancy and flow with residual connections of visual features from the image encoder and trajectory fusion. 2D-CNNs (kernel size 3×3) are shared across future time steps, and separated 2D-CNNs (kernel size 1×1) are used to process the information from the residual path. We select the dimensions of pyramid decoder as $[192, 128, 96, 48, 2]$. We also design the split decoding heads for occupancy and flow to enable varying projections in receiving shared features and direct information paths. The output of each occupancy head would be two dimensions of observed $\hat{O}_k^b$ and occluded $\hat{O}_k^c$ occupancy grids, while the output of the flow head is also 2 dimensions with d_x and d_y for the future flow $\hat{F}_k$.

2.4. Objective functions

To achieve better performance for joint predictions of occupancy and flow, we modify the objective functions in [5], and we follow the binary probabilistic modeling of the occluded and observed occupancy. However, as the ground-truth samples are imbalanced heavily towards zero (unoccupied area), we alleviated this issue by replacing the cross-entropy term with focal loss [3]:

$$\begin{aligned} \mathrm{FL}\,(y,p) = &- y\alpha\,(1-p)^{\gamma}\log\,(p) \\ &- (1-y)(1-\alpha)p^{\gamma}\log\,(1-p)\,, \end{aligned} \tag{3}$$

where the parameters are $\alpha = 0.25, \gamma = 2$.and the modified losses for $\hat{O}_k^b$ and $\hat{O}_k^c$ become:

$$\mathcal{L}_{obs} = \sum_{k=1}^{T_f}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}\mathrm{FL}\left(O_k^b(x,y), \hat{O}_k^b(x,y)\right), \tag{4}$$

$$\mathcal{L}_{occ} = \sum_{k=1}^{T_f}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}\mathrm{FL}\left(O_k^c(x,y), \hat{O}_k^c(x,y)\right). \tag{5}$$

The flow loss remains L1 distance weighted by ground-truth occupancy grids averaged by valid time-steps:

$$\mathcal{L}_F = \sum_{k=1}^{T_f}\sum_{x=0}^{w-1}\sum_{y=0}^{h-1}\left\|F_k(x,y) - \hat{F}_k(x,y)\right\|_1 O_k(x,y). \tag{6}$$

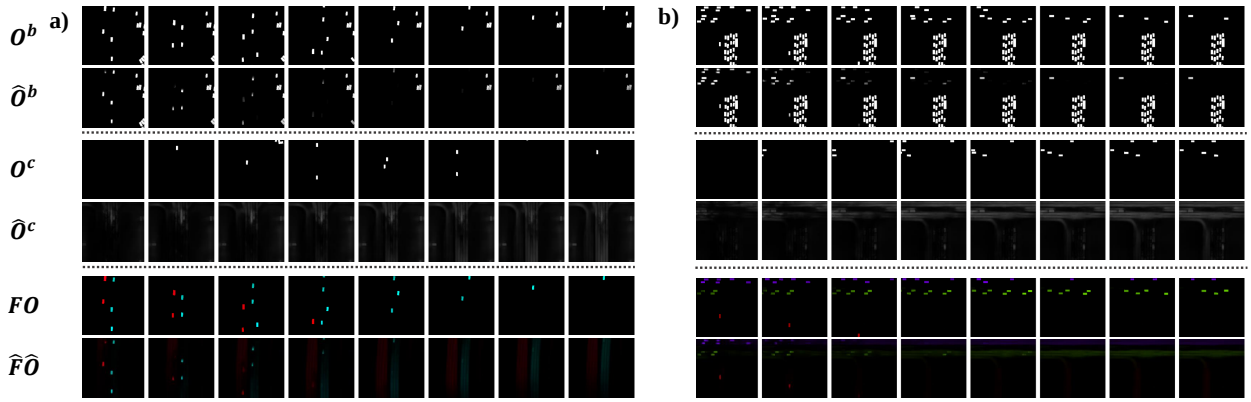

Figure 4. Visualization of validation results from two samples: a) sample A b) sample B. The dotted lines separate ground-truth and the prediction results of observed occupancy $O^b, \hat{O}^b$; occluded occupancy $O^c, \hat{O}^c$; and the occupancy flow $FO, \hat{F}\hat{O}$, ($O = O^b + O^c; \hat{O} = \hat{O}^b + \hat{O}^c$).

For flow-warped occupancy loss $\mathcal{L}_W$ defined in [5], we replace the predicted occupancy $\mathcal{W}_k\hat{O}_k$ with the ground truth one $\mathcal{W}_k O_k$ for better performance:

$$\mathcal{L}_W = \sum_{k=1}^{T_f} \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} \mathrm{FL}\left(O_k(x,y), \mathcal{W}_k(x,y) O_k(x,y)\right), \quad (7)$$

The final multi-task objective sums up the following loss terms averaged by the height, width and timestep of the output $h, w = 256; T_f = 8$:

$$\mathcal{L} = \frac{1}{hwT_f}\left(1000\mathcal{L}_{obs} + 1000\mathcal{L}_{occ} + 1000\mathcal{L}_w + \mathcal{L}_F\right). \quad (8)$$

2.5. Implementation details

We choose GELU as the activation function in all encoders and ELU in the pyramid decoder. To mitigate overfitting, dropout is added after each MLP layer and also in the image encoder, all with a dropout rate of 0.1. Due to the numerous size of data inputs and predictions, we use a distributed training strategy on 4 Tesla V100 with a total batch size of 16. Adam optimizer is used with an initial learning rate of 1e-4, and the learning rate decays by a factor of 50% every 3 epochs. The total training epochs are set to 15.

3. Results

Qualitative results in comparison with the ground-truth are displayed in Fig. 4. The results reveal high accuracy in predicting observed occupancy together with the flow. Still, the performance of occluded occupancy needs further improvements. Quantitative results are given in Table 1. Flow errors are minimized well by separated prediction heads. Improved performance is shown in terms of prediction accuracy, especially soft-IOU which reveals high accuracy in predicted occupied locations of all vehicles.

Table 1. Testing results of occupancy and flow prediction.

Testing Metrics						
obs-AUC	obs-IOU	occ-AUC	occ-IOU	flow-EPE	FG-AUC	FG-IOU
0.7514	0.4818	0.161	0.0183	3.5867	0.7772	0.5551

References

[1] Scott Ettinger et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 2

[2] Tsung-Yi Lin et al. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[3] Tsung-Yi Lin et al. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[4] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 2

[5] Reza Mahjourian et al. Occupancy flow fields for motion forecasting in autonomous driving. *IEEE Robotics and Automation Letters*, 7(2):5639–5646, 2022. 1, 2, 3, 4

[6] Xiaoyu Mo et al. Graph and recurrent neural network-based vehicle trajectory prediction for highway driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1934–1939. IEEE, 2021. 1

[7] Xiaoyu Mo et al. Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 1

[8] Ngiama et al. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *International Conference on Learning Representations*, 2021. 1

[9] Mingxing Xu et al. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020. 1