# HGNET: A Hierarchical Feature Guided Network
# for Occupancy Flow Field Prediction

Zhan Chen[1]    Chen Tang[1]    Lu Xiong[1]

[1]Tongji University

{zhan_chen, chen_tang, xiong_lu}@tongji.edu.cn

## Abstract

*Predicting the motion of multiple traffic participants has always been one of the most challenging tasks in autonomous driving. The recently proposed occupancy flow field prediction method has shown to be a more effective and scalable representation compared to general trajectory prediction methods. However, in complex multi-agent traffic scenarios, it remains difficult to model the interactions among various factors and the dependencies among prediction outputs at different time steps. In view of this, we propose a transformer-based hierarchical feature guided network (HGNET), which can efficiently extract features of agents and map information from visual and vectorized inputs, modeling multimodal interaction relationships. Second, we design the Feature-Guided Attention (FGAT) module to leverage the potential guiding effects between different prediction targets, thereby improving prediction accuracy. Additionally, to enhance the temporal consistency and causal relationships of the predictions, we propose a Time Series Memory framework to learn the conditional distribution models of the prediction outputs at future time steps from multivariate time series. The results demonstrate that our model exhibits competitive performance, which ranks 3rd in the 2024 Waymo Occupancy and Flow Prediction Challenge.*

## 1. Introduction

Predicting the motion of multiple traffic participants has consistently been a significant challenge in autonomous driving technology. An accurate and robust prediction module must effectively handle a wide range of traffic scenarios and participant behaviors. Additionally, it is crucial to account for the potential interactions among different traffic participants, as simplistic predictions can result in unrealistic and contradictory outputs. Leveraging the robust capabilities of deep learning, recently proposed occupancy flow field prediction method offers an enhanced and more
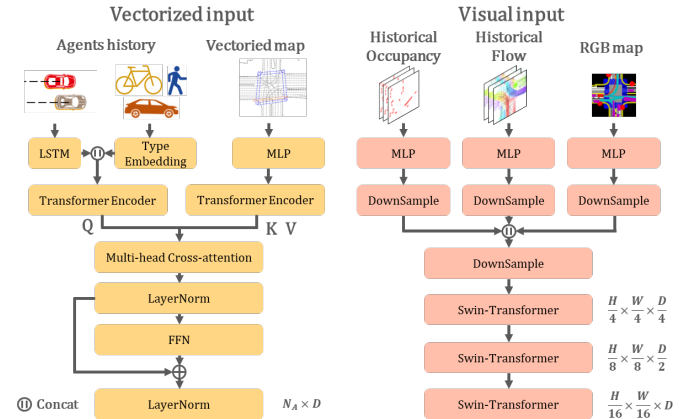


Figure 1. Transformer-based encoder for multimodal inputs.

efficient representation for multimodal predictions in multi-agent scenarios [4].

However, current prediction techniques for occupancy flow field face several significant challenges. Firstly, while predictions for visible agent motion typically perform well within conventional tasks, forecasts for occluded obstacles often exhibit suboptimal performance. Additionally, there is a notable absence of suitable inference structures in network design, with many methods relying on high-dimensional abstract features at the network's front end to generate the final prediction targets [2, 3]. Secondly, there is a lack of integration and correlation among the predictions for future flow and the future occupancy predictions for both visible and occluded obstacles. Effectively leveraging these correlations could substantially enhance the prediction performance of each component. Thirdly, for sequential prediction tasks, modeling the relationships between outputs at different time steps is essential for improving prediction accuracy.

In this technical report, we propose a hierarchical feature guided network for predicting occupancy flow field, along with several specialized structural designs to efficiently extract key features for forecasting the behaviors of

multiple agents in complex traffic scenarios characterized by strong interaction relationships. Firstly, we employ a transformer-based encoder to extract visual and vectorized historical information, as well as map information, serving as input context tokens. Secondly, with the proposed Feature-Guided Attention (FGAT) module, we introduce a flexible and hierarchical framework that fully exploits the intrinsic relationships among flow, visible, and occluded agents' occupancy grid, thereby efficiently extracting correlated features. Thirdly, to extract the temporal relationships within prediction results over the forecast horizon and enhance the continuity and correlation among features, we have designed a Time Series Memory framework to capture and store temporal information. It should be noted that our proposed transformer-based prediction framework exhibits significant scalability and flexibility while ensuring superior predictive performance. Experiments on the Waymo Open Motion Dataset [1] demonstrate that HGNET can accurately forecast trajectories in the form of occupancy flow field at the scene level.

## 2. Approach

In this section, we provide a detailed introduction to the HGNET framework. First, we briefly introduce the network's input and encoder, with the overall architecture illustrated in Figure 1. Next, we describe the proposed decoder, along with several specialized structures designed specifically for the prediction tasks. Finally, we explain the training objectives used to optimize the prediction model for joint occupancy flow field (OFF).

### 2.1. Multi-modal Context Tokens Encoding

To maximize the utilization of available multimodal input information, we employ two types of input information, including vectorized input and visual input. Vectorized input consists of the historical trajectory state sequences of $N_A$ agents within the scene over the past $T_h$ time steps, containing information including position, velocity, heading angle, and agent type. We also introduce vectorized map information $\mathcal{M}_{vec}$ to the system. Finally, the positional attributes of all agents and map elements are transformed into the local coordinate system of the ego vehicle. For visual input related to the prediction task of occupancy flow, we establish a historical occupancy grid along with the backward flow field between time steps $t = -T_h$ and $t = 0$. Additionally, following the approach in [3], we introduce an RGB visualization representation $\mathcal{M}_{vis}$ of the map network to thoroughly incorporate essential map information like traffic light signals.

As shown in Fig. 1, to consider the interaction among all elements within the traffic scenario, the historical states of all agents are firstly encoded by LSTM networks for all traffic agents, and concatenate it to agent's type embedding out-

put. Then a two-layer self-attention Transformer encoder is applied to model the agent-agent interaction. The vectorized map waypoints are effectively encoded by a MLP layer as the latent feature, followed by a self-attention Transformer encoder. To capture relationships and dependencies between agents and map, we employ a cross-attention Transformer as agent-map interaction encoder, utilizing agent's interaction feature as query ($\mathbf{Q}$) and map feature encoded from vectorized map as key and value ($\mathbf{K}, \mathbf{V}$). Without loss of generality, we let all latent features have $D$ hidden dimensions. Therefore, the vectorized tokens have the shapes of $[N_A, D]$. For visual features, the original inputs are initially encoded by three MLP layers, then down sampled separately. We concatenate them with each other as a whole visual feature and feed it into the Swin-Transformer-based encoder. Each Swin-Transformer module comprises a two-layer Transformer equipped with both window self-attention and shifted window self-attention. This configuration facilitates comprehensive interaction modeling for visual features through global and intersected attention mechanisms. Additionally, each attention module incorporates multi-head attention with relative positional bias. The outputs from three stages of Swin-Transformer blocks are aggregated into a list $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ with shapes of $[\frac{H}{4}, \frac{W}{4}, \frac{D}{4}], [\frac{H}{8}, \frac{W}{8}, \frac{D}{2}], [\frac{H}{16}, \frac{W}{16}, D]$ respectively, and serve as the final output of visual features.

### 2.2. Hierarchical Feature Guided Decoder

To organize the prediction inference sequence of various prediction targets more systematically and fully leverage the guiding role of different features, we design the structure of the hierarchical feature guided decoder as shown in Fig. 2. We choose flow as the first prediction and utilize its high-dimensional features to inform subsequent prediction tasks for it represents the changes in occupancy grids between adjacent timesteps. Though occluded occupancy cannot be directly inferred, the relevant features can be effectively extracted using visible information and historical data [5]. Thus, we predict occluded occupancy as the last prediction target, merging the features of both flow and observed occupancy as guiding features. For each prediction pathway, we first encode the corresponding inputs using a similar method as described before, obtaining a feature list with the same shape as the visual features (where the original features of occluded occupancy are derived from visible occupancy and flow). Subsequently, the encoded features pass through a self-attention layer and are fed into our proposed FGAT module as the query.

**Feature-Guided Attention module.** We designed the FGAT module to amplify the query with corresponding features guided by learnable offsets generated from the guiding feature. Within the hierarchical network architecture, the FGAT module aggregates various features from future
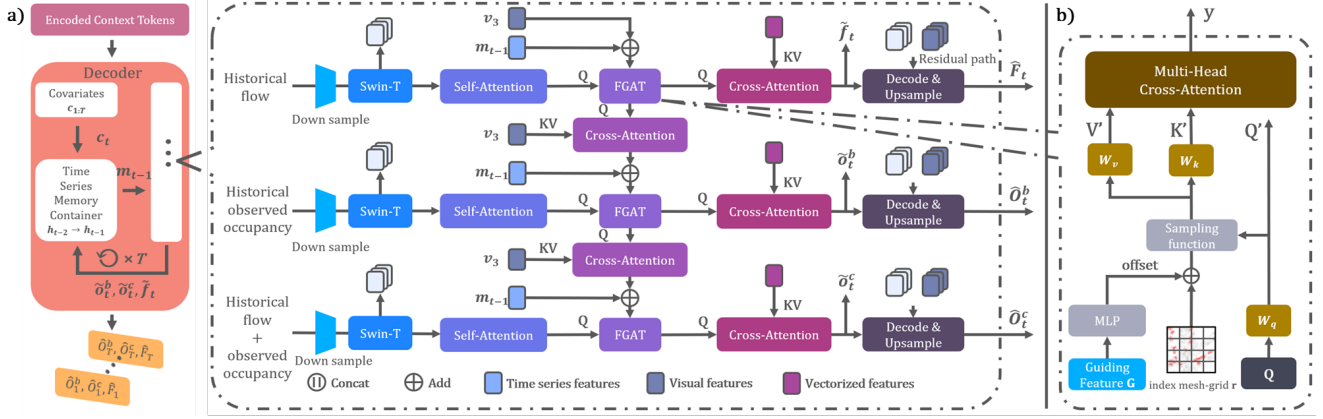
Figure 2. a) Framework of the decoding pipeline. b) Structure of the Feature-Guided Attention module.

timesteps. Particularly, except for the top-level FGAT module, all guiding features are first input into a cross-attention module as queries (with visual feature $\mathbf{v}_3$ as keys and values) then added with time series feature $\mathbf{m}_{t-1}$, before entering the FGAT module. Given the encoded historical feature as query, guiding feature $\mathbf{Q}, \mathbf{G} \in \mathbb{R}^{H/16 \times W/16 \times C}$, and a uniform index mesh-grid of points $\mathbf{r} \in \mathbb{R}^{H/16 \times W/16 \times 2}$ as the references, the offsets $\Delta \mathbf{r}$ for reference points are generated from the guiding feature by a MLP layer along with a *tanh* layer:

$$\mathbf{Q}' = \mathbf{Q}W_q, \mathbf{K}' = \mathbf{x}W_k, \mathbf{V}' = \mathbf{x}W_v,$$
$$\mathbf{x} = f_\phi(\mathbf{Q}'; \mathbf{r} + \Delta \mathbf{r}), \Delta \mathbf{r} = \mathtt{tanh}(\mathtt{MLP}(\mathbf{G})), \quad (1)$$

where $\mathbf{K}'$ and $\mathbf{V}'$ represent the feature-guided key and value embeddings, and we use a bilinear interpolation as $f_\phi(\cdot; \cdot)$:

$$f_\phi(\mathbf{G}; \mathbf{R}) = \sum_{(x,y)} g(\mathbf{R}_x, x)g(\mathbf{R}_y, y)\mathbf{G}[y, x, :], \quad (2)$$

where $g(i, j) = \mathtt{max}(0, 1 - |i - j|)$ and $(x, y)$ represents every point location of $G \in \mathbb{R}^{H/16 \times W/16 \times D}$. Finally we perform a multi-head cross-attention on $\mathbf{Q}', \mathbf{K}', \mathbf{V}'$ with relative positional bias $B$, the projection matrice $W_o$ and the dimension of the key token $d$,

$$\mathtt{MHCA}(\mathbf{Q}', \mathbf{K}', \mathbf{V}') = (h_i || ... || h_{\mathbf{M}})W_o,$$
$$h_i = \mathtt{softmax}(\mathbf{Q}'\mathbf{K}'^\top / \sqrt{d} + B)\mathbf{V}', \quad (3)$$

**Time Series Memory Framework.** To improve the accuracy of temporal feature prediction results, we adopt this framework to learn a model of the conditional distribution of future time steps of a multivariate time series given the historical features and covariates as:

$$q(\mathbf{y}_{t_0:T} | \mathbf{y}_{1:t_0-1}, \mathbf{c}_{1:T}) = \prod_{t=t_0}^{T} q(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \mathbf{c}_t), \quad (4)$$

where $t_0$ denotes the current prediction time step. We use the embeddings of future time steps as covariates $\mathbf{c}_{1:T}$. To model the temporal dynamics via the updated hidden state $h_{t-1}$, we employ three multi-layer GRU networks to encode the time series sequence up to time step $t - 1$, given the covariates of the next time step $\mathbf{c}_t$:

$$\mathbf{m}_{t-1}, \mathbf{h}_{t-1} = \mathbf{GRU}(\mathtt{concat}(\mathbf{y}_{t-1}, \mathbf{c}_t), \mathbf{h}_{t-2}), \quad (5)$$

where $\mathbf{h}_0 = \mathbf{0}$ and $\mathbf{m}_{t-1}$ is the output of GRU network. In three prediction heads, $\mathbf{y}_{t-1}$ represents $\tilde{\mathbf{f}}_{t-1}, \tilde{\mathbf{o}}_{t-1}^b, \tilde{\mathbf{o}}_{t-1}^c$ respectively. They are the outputs of the cross-attention module, where the output of the FGAT module serves as the query, the encoded vector features are used as the key and value. By dynamically updating the hidden states, the information from previous time steps is preserved and fused for predicting features at the next time step.

Finally, we decode the flow and occupancy from the feature tensors using feature pyramid network (FPN), which consists of multi-layer 2D-CNNs and upsampling layers, along with additional 2D-CNNs employed to process the features in the residual paths.

### 2.3. Training Objectives

For the occupancy loss $\mathcal{L}_{occ}$, we utilize the focal loss and the cross-entropy loss for the observed and occluded occupancy regression. Similar to [4], smooth L1 loss is applied as flow loss $\mathcal{L}_f$ to supervise the flow prediction. The final multi-task training objective sum up the loss terms scaled by the size of the grid map (with height $h$ and width $w$) and length of timesteps of the output:

$$\mathcal{L} = \frac{1}{hwT}(100\mathcal{L}_{occ} + \mathcal{L}_f) \quad (6)$$

| Evaluation Metrics | Observed Occupancy | | Observed Occupancy | | Flow | Flow-grounded Occupancy | |
|---|---|---|---|---|---|---|---|
| **Method** | **AUC ↑** | **Soft-IoU ↑** | **AUC ↑** | **Soft-IoU ↑** | **EPE ↓** | **AUC ↑** | **Soft-IoU ↑** |
| DOPP | **0.797** | 0.343 | **0.194** | 0.024 | **2.957** | **0.803** | **0.516** |
| STNet | 0.755 | 0.230 | 0.166 | 0.018 | 3.378 | 0.756 | 0.443 |
| **Ours** | 0.733 | **0.421** | 0.166 | **0.039** | 3.670 | 0.740 | 0.450 |

Table 1. Summary of the testing performance on the Waymo occupancy and flow prediction benchmark.

| FG-AT | time series | Observed AUC↑ | Occluded AUC↑ | Flow EPE↓ | FG AUC↑ |
|---|---|---|---|---|---|
| ✗ | ✗ | 0.713 | 0.131 | 3.905 | 0.719 |
| ✓ | ✗ | 0.721 | 0.154 | 3.724 | 0.733 |
| ✓ | ✓ | **0.742** | **0.158** | **3.561** | **0.743** |

Table 2. Ablation study on FGAT module and time series memory framework.

## 3. Experiments

### 3.1. Implementation Details

The hidden feature dimension is 256, We choose GELU as the activation function in all encoders and RELU in the decoder. Dropout is followed after every MLP layer, all with a dropout rate of 0.1. we use a distributed training strategy on 2 Nvidia RTX 6000 Ada GPUs with a total batch size of 16. The training process lasts 16 epochs. We use the Adam optimizer for training with the initial learing rate of 1e-4, and the learning rate decays by a factor of 50% every 2 epochs.

### 3.2. Quantitative Results

The performance of HGNET on the the Waymo occupancy and flow prediction benchmark is shown in Tab. 1, where we can see that our proposed approach outperforms other method on some metrics, and overall, it demonstrates good performance across other metrics and exhibits a certain level of competitiveness.

### 3.3. Ablation Study

We conduct an ablation study to investigate the infuences of key modules in our proposed method, i.e., FGAT module and time series memory framework. We conducted experiments with two ablation variants of our model: one excluding the time series memory framework (i.e., without updates to the hidden states of the time series and the fusion of time series features) and another excluding the FGAT module (substituting it with a standard cross-attention module). As shown in Tab. 2, the performance metrics on the Waymo occupancy flow validation set exhibit a decline across all metrics for the ablated models. These results substantiate the efficacy of our proposed framework in enhancing pre-diction accuracy.

## 4. Conclusion

We propose HGNET, a hierarchical multi-modal feature guided framework for joint multi-agent occupancy flow field prediction. Leveraging the proposed Feature-Guided Attention module for feature guidance and an effective Time Series Memory framework for temporal feature extraction, our model achieves accurate multi-agent motion prediction in the form of occupancy flow fields. Experimental results demonstrate that our method achieves competitive performance on the Waymo occupancy and flow prediction benchmark.

## References

[1] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, 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] Yihan Hu, Wenxin Shao, Bo Jiang, Jiajie Chen, Siqi Chai, Zhening Yang, Jingyu Qian, Helong Zhou, and Qiang Liu. Hope: Hierarchical spatial-temporal network for occupancy flow prediction. *arXiv preprint arXiv:2206.10118*, 2022. 1

[3] Haochen Liu, Zhiyu Huang, and Chen Lv. Multi-modal hierarchical transformer for occupancy flow field prediction in autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1449–1455. IEEE, 2023. 1, 2

[4] Reza Mahjourian, Jinkyu Kim, Yuning Chai, Mingxing Tan, Ben Sapp, and Dragomir Anguelov. Occupancy flow fields for motion forecasting in autonomous driving. *IEEE Robotics and Automation Letters*, 7(2):5639–5646, 2022. 1, 3

[5] Hao Shao, Letian Wang, Ruobing Chen, Steven L Waslander, Hongsheng Li, and Yu Liu. Reasonnet: End-to-end driving with temporal and global reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13723–13733, 2023. 2