# *DOPP*: Differentiable Integrated Occupancy Flow Field and Motion Prediction

Haochen Liu, Zhiyu Huang, Wenhui Huang, Haohan Yang
Xiaoyu Mo, Hongyang Gao, Chen Lv
Nanyang Technological University, Singapore
haochen002@e.ntu.edu.sg, zhiyu001@e.ntu.edu.sg, wenhui001@e.ntu.edu.sg,
haohan.yang@ntu.edu.sg,xiaoyu.mo@ntu.edu.sg,hgao006@e.ntu.edu.sg,lyuchen@ntu.edu.sg

## Abstract

*Consistently forecasting the accurate future states of surrounding traffic participants is a paramount role among the autonomous driving system. The occupancy flow field prediction offers a scalable and effective framework for jointly forecasting the future motions of multiple agents within a scene. However, significant challenges persist in accurately modeling future consistency between prediction patterns, as well as traceability for each rasterized agent. This work introduce* ***DOPP****: a differentiable prediction system integrating occupancy flow field with future motion states. With query-informed modular co-design insights upon our preceding work [7], we devised MS-OccFormer module, which achieves multi-stage alignment per occupancy flow field forecasting with consistent awareness from agent-wise marginal motion predictions. Additionally, we developed an integrated learning paradigm in consistently update the all of the prediction objectives. Through our proposed method, we achieved competitive prediction accuracy and displacements for occupancy and flow predictions, demonstrating impressive flow-traced performance and ranked* $1^{st}$ *in the 2024 WOMD leaderboard.*

## 1. Introduction

Forecasting the accurate and social-consistent status of targeted traffic participants represents one of the most significant challenges in the realm of autonomous driving. [1]. Notably, the hereditary heterogeneous agents and scalable interactive behaviors among driving scene underscores the challenging tasks for prediction. To adequately tackle this challenge, existing approaches are primarily twofold predicting occupancy field probabilities, or conducting multi-agent trajectory predictions anchored per agent instance. The former spotlights scalable predictions allowing arbitrary surrounding agents' forecasting [9–11], while the latter generates tractable trajectories without spatial limits [5]. However, relying on a single prediction format inevitably presents inherent limitations. Expressly,occupancy-based approaches achieve accurate joint predictions but lose agent-specific tractability. This can result in temporal conflicts and the omission of critical agents. Conversely, motion prediction models face inconsistencies and exponential computational costs when modeling joint interaction patterns among multiple agents

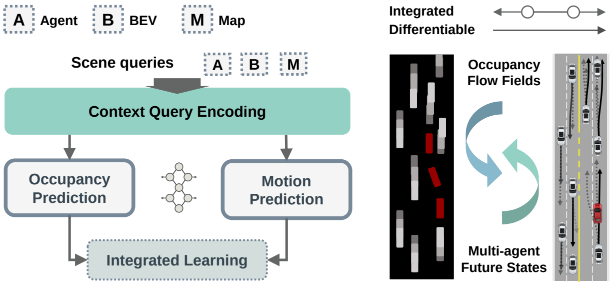

Figure 1. **Overview:** We propose an integrated network which couples the prediction task of occupancy flow field with multi-agent marginal forecasting. The method functions a partial variant as our preceding system [7] and delivers compliant predictions.

These characteristics underscore the complementary nature of prediction representations. Multiple trials have been witnessed in the community, presenting similar ideas integrating both formats as an efficient multi-task paradigm [6] or end-to-end pipelines [3]. However, current approaches falls short in addressing consistency among different predictions, demonstrating potential conflicts and missing agents for occupancy field. This may further leads to non-strategic planning behaviors and safety concerns. The shortfall motivates a co-design which enables a mutual information guidance under an integrated model between occupancy flow field and motion prediction for multiple agents.

In this study, we aimed to address these challenges through an differentiable integration predicting occupancy flow field with multi-agent motion states, termed as **DOPP**. It functions as a variant upon the prediction module of our preceding system [7]. At its core, we launch an integrated

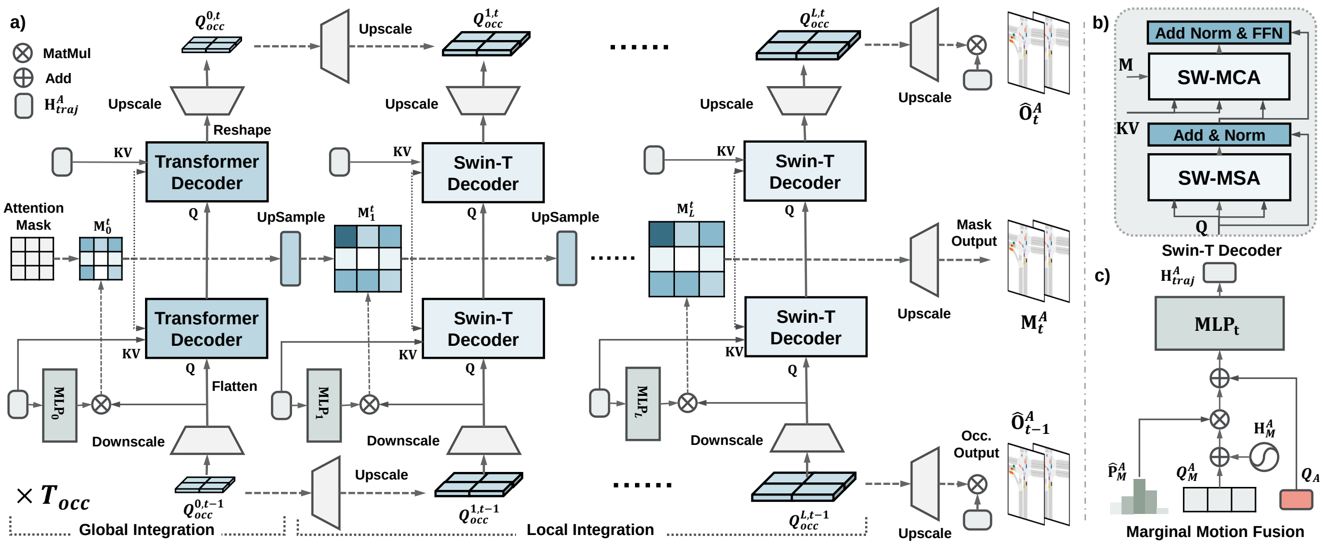

Figure 2. Cascade integrated decoding framework of Ms-OccFormer in DOPP a) A single block of multi-scale agent-conditioned occupancy predictor. Joint occupancy and backward flow $\hat{\mathbf{O}}, \hat{\mathbf{F}}$ are consistently integrated with marginal prediction features $\mathbf{H}^A_{traj}$, through global interactions and local refinements, and guided by iteratively updated learnable attention mask $\mathbf{M}$; b) A Swin-T decoder for local interactions through shifted-window cross attention; c) Agent-wise fusion for marginal prediction features.

pipeline performing occupancy and flow field prediction that aligns and refines consistently with marginal motion prediction. To elaborate further, as depicted in Figure 1, our proposed framework primarily focuses on the decoding stage integrating both tasks. With encoded scene features upon our previous work [9], we propose MS-OccFormer, a cascade decoder network to perform marginal-conditioned occupancy prediction. The proposed model fuses motion predictive features while forecasting step-wise occupancy flow features, enabling mutual predictive awareness. The motion prediction decoder are inherited from another previous work [5]. The proposed paradigm demonstrates state-of-the-art accuracy in occupancy and flow field prediction, thanks to compliant integration from the motion prediction.

## 2. Method

### 2.1. Formulation and Encoding

The occupancy flow challenge is formulated as multi-task objectives predicting future frames of observed occupancy $\mathbf{O}_{1:T} \in \mathbb{R}^{T\times H\times W}$, occluded occupancy $\mathbf{O}^{\text{occ}}_{1:T} \in \mathbb{R}^{T\times H\times W}$, and backward flow $\mathbf{F}_{1:T} \in \mathbb{R}^{T\times H\times W\times 2}$ simultaneously [10]. To model the global interactions between scene elements with BEV perceptions, we inherit from our previous work [9] to gather and encode separate visual and vectorized scene inputs. Then, visual features are encoded under similar structure of [9] as BEV features $Q_B \in \mathbb{R}^{H\times W\times D}$. Meanwhile, agent $Q_A \in \mathbb{R}^{N_A\times D}$ and map $Q_{Map} \in \mathbb{R}^{N_M\times D}$ features from respective encoders [8] are concatenated and encoded by stack of 4 Transformer encoders. Encoded features $\mathbf{X} = \{Q_B, Q_{Map}, Q_A\}$ are then served as input for DOPP integrating dual prediction objectives.

### 2.2. Ms-OccFormer

To further tackle the consistency challenges for occupancy given encoded scene features $\mathbf{X}$, we propose MS-OccFormer spotlights twp aspects, i.e. Agent-conditioned occupancy that defines tractable predictions for occupancy, and Multi-scale prediction-wise integration that deals with the interactive alignments for occupancy with different granularity. Illustrated in Fig. 2, MS-OccFormer utilizes a cascaded pipeline to rollout the future horizon $T$ , decoding per-step occupancy flow field prediction. It is based on $L$ levels fusion of previous-step occupancy features and motion prediction features through a motion predictor [5].

#### 2.2.1 Predictive Queries

We leverage occupancy queries $Q_{occ} \in \mathbb{R}^{H\times W\times D}$ in multi-scale aggregating for positional and BEV features: $Q_{occ} = \mathrm{MLP}([\mathrm{PE}(I_B); Q_B])$. Positional grids $I_B \in \mathbb{R}^{H\times W\times 2}$ are encoded using sinusoidal $\mathrm{PE}(\cdot)$ and transformed by multi-layer perceptron (MLP). We further downsample $Q_{occ}$ under $L$ levels of query sets $\{Q^{l,0}_{occ} \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times D}\}^0_{l=L}$ to recurrently query multi-scale interactions.

To fully extract the interactive marginal prediction features, we conduct an agent-wise fusion (see Fig. 2c) that leverages the multi-modal prediction $\hat{\mathbf{Y}}$ outputs from motion predictor [5]. Motion prediction features $Q^A_M \in \mathbb{R}^{N_A\times M\times D}$ are received from the decoding features in motion predictor. $Q^A_M$ is then fused with marginal features

$\mathbf{H}^A_M = \max_{T_p} \mathrm{MLP}(\mathrm{PE}(\hat{\mathbf{y}}))$, where $\hat{y} \in \mathbb{R}^{N_A \times M \times T_p \times 2}$ denotes the predicted trajectories. The fused features are projected by each horizon:

$$\mathbf{H}^A_{traj} = \mathrm{MLP}_{1:T}(\hat{\mathbf{p}}(Q^A_M + \mathbf{H}^A_M) + Q_A), \qquad (1)$$

where $\mathbf{H}^A_{traj} \in \mathbb{R}^{T \times N_A \times D}$ denotes the marginal features.

### 2.2.2 Agent-conditioned Occupancy

The primary challenge in formulating $\mathbf{O}_{1:T} \in \mathbb{R}^{T \times H \times W}$ lies in the intractability with motion predictions that cause joint inconsistencies. Inspired by instance-level occupancy $\mathbf{O}^A_{1:T} \in \mathbb{R}^{T \times H \times W \times N_A}$ [3], we propose the marginal-conditioned occupancy prediction task. This models the consistent joint occupancy $p(\mathbf{O}^A_{1:T} | \mathbf{Y}^A_M, \mathbf{X})$ over agent-wise motion predictions. To associate uncertainty and mutual interactions, given final joint decoding features $Q^L_{occ}$ and marginal features $\mathbf{H}^A_{traj}$, the agent-conditioned occupancy will be eventually modeled by dot products:

$$\mathbf{O}^A_{1:T} = \sigma(Q^L_{occ} \cdot \mathrm{MLP}(\mathbf{H}^A_{traj})^T), \qquad (2)$$

where $\sigma$ denotes Sigmoid for per-grid probabilities. The original task can be then transformed back $\hat{\mathbf{O}}_{1:T} = \max_A \hat{\mathbf{O}}^A_{1:T}$ for the observed occupancy prediction. Predicted flow $\hat{\mathbf{F}}_{1:T}$ and occluded occupancy $\hat{\mathbf{O}}^{\mathrm{occ}}_{1:T}$ are concurrently decoded from $Q^{t,L}_{occ}$ for each future step $t \in [1, T]$.

### 2.2.3 Multi-scale Prediction-wise Integration

This aims to iteratively align multi-scale interaction features between hybrid prediction in decoding $\mathbf{O}^A_{1:T}$.In Fig. 2a, multi-scale succeeded occupancy features $\{Q^{l,t-1}_{occ}\}^L_{l=1}$ query aligned marginal features by attentions at different granularities from two-stage Transformer decoders.

The global integration stage leverages the vanilla Transformer decoders to perform per-grid interactions from flattened high-level joint features $Q^{L,t-1}_{occ}$ with marginal ones. Subsequently, with the upscaling of occupancy features $\{Q^{l,t-1}_{occ}\}^{L-1}_{l=1}$, the local integration stage focuses on capturing consistency from partial joint behaviors with marginal features. This motivates us to design shift-window multi-head cross-attention (SW-MCA) [7]. As depicted in Fig. 2b, we employ the rolling process to simultaneously capture local interactions under shifted windows attention.

To ensure interactive consistency across multi-scale integration, we devise a learnable attention mask $\mathbf{M}^l_{1:T} \in \mathbb{R}^{T \times \frac{H}{2^{L-l}} \times \frac{W}{2^{L-l}} \times N_A}$ for Transformer decoder that iteratively refines upon interaction results from the previous scale. This aligns the attention modeling based on the previous results Shown in Fig. 2a, for each level, the attention mask gets updated with agent-conditioned occupancy on the current scale level: $\hat{\mathbf{M}}^l = \sigma(Q^l_{occ} \cdot \mathrm{MLP}_l(\mathbf{H}^A_{traj})^T)$. The attention masks are then iteratively updated following:

$$\mathbf{M}^l = \lambda_m \, \mathrm{Upsample}(\mathbf{M}^{l-1}) + (1 - \lambda_m)\hat{\mathbf{M}}^l, \qquad (3)$$

where $\lambda_m = 0.5$ is the update factor. In general, given Transformer decoder at certain stage as Trans, the prediction-wise integration under $l$ of $t$ is defined as:

$$Q^{l,t}_{occ} = \mathrm{Trans}(q = Q^{l,t-1}_{occ}, k, v = \mathbf{H}^{A,t}_{traj}, m = \mathbf{M}^l_t). \quad (4)$$

Output joint occupancy features $Q^{L,t}_{occ}$ will be eventually fused via Equ. 2 for conditioned occupancy predictions $\hat{\mathbf{O}}^A_{1:T}$ and transformed back to observed occupancy $\hat{\mathbf{O}}_{1:T}$. $Q^{L,t}_{occ}$ also concurrently decode flow $\hat{\mathbf{F}}_{1:T}$ and occluded occupancy $\hat{\mathbf{O}}^{\mathrm{occ}}_{1:T}$. Both occupancy forms are added, and warped by predicted flow for the final output.

## 2.3. Integrated Learning

We formulate a multi-task learning paradigm for the proposed network. For precise prediction of marginal-conditioned observed occupancy $\hat{\mathbf{O}}^A_{1:T}$ and the occluded ones $\mathbf{O}^{\mathrm{occ}}_{1:T}$ in MS-OccFormer, we employ a combination of top-k BCE loss and Dice loss [3] jointly for $\hat{\mathbf{O}}^A_{1:T}$ and $\mathbf{M}_{1:T}$, for balanced predictions of occupancy probabilities: $\mathcal{L}_{occ} = \mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{Dice}}\mathcal{L}_{\mathrm{Dice}}$, with $\lambda_{\mathrm{Dice}} = 5$. The backward flow regression are updated by the L1 loss as [9]. All objectives learn jointly following our previous work [9].

## 2.4. Implementation details

The challenge is defined by predicting $T = 8$ frames of future occupancy, each rasterized by $80 \times 80m^2$ driving scene with $H, W = 256$. Encoded scene comprises $N_A = 33$ agent and $N_M = 100$ map features with $D = 256$. Motion predictor outputs $M = 6$ modals of predictions with $T_p = 80$. The integration level is set to $L = 3$ in Ms-OccFormer decoder. The DOPP framework is trained from scratch by WOMD [2] training set without augmentations or post-processing. We choose ReLU as the activation function, dropout is added after each layer with a dropout rate of 0.1. We use a distributed training strategy on 4 Tesla A100 with a total batch size of 16. AdamW optimizer is used with cosine annealing strategy, initializing the learning rate as 1e-4. The total training epochs are set to 20.

# 3. Results

Table 1 presents the quantitative results in comparison to other methods on the 2024 Waymo Occupancy and Flow Prediction Leaderboard. The proposed method marks exceptional accuracy (AUC) for all occupancy predictions, presenting +2.4% observed, +8.9% occluded, and +2.3% flow-traced occupancy in comparing to previous state-of-the-art approaches [9]. The compliant integration further

Table 1. Testing performance on the Waymo Occupancy and Flow Prediction Leaderboard

| Evaluation Metrics | Observed Occupancy | | Occluded Occupancy | | Flow | Occupancy Flow Field | |
|---|---|---|---|---|---|---|---|
| **Model** | **AUC ↑** | **Soft-IOU ↑** | **AUC ↑** | **Soft-IOU ↑** | **EPE ↓** | **FT-AUC ↑** | **FT-Soft-IOU ↑** |
| STrajNet [9] | 0.778 | 0.491 | 0.178 | **0.045** | 3.204 | 0.785 | 0.531 |
| VectorFlow [4] | 0.755 | 0.488 | 0.174 | 0.045 | 3.583 | 0.767 | 0.531 |
| OFMPNet [11] | 0.769 | **0.502** | 0.165 | 0.042 | 3.587 | 0.761 | **0.538** |
| **DOPP (Ours)** | **0.797** | 0.343 | **0.194** | 0.024 | **2.957** | **0.803** | 0.515 |

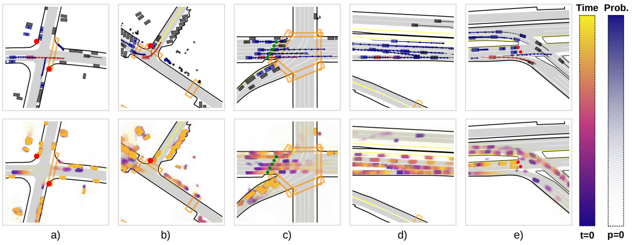

Figure 3. Qualitative results of WOMD scenarios: a) Crossing an unsignalized intersection with an incoming vehicle; b) Unprotected left-turn with a potential take-over vehicle; c) Cruising a five-point intersection with heavy traffic; d) Merging in a highway; and e) Right-turn with heading cyclist and low-speed vehicle.

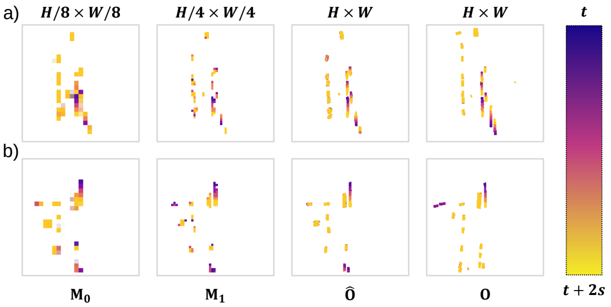

Figure 4. Qualitative ablations of short term multi-scale attention masks $\mathbf{M}_{1:T}$ compared with occupancy predictions $\hat{\mathbf{O}}_{1:T}$ and the ground-truth $\mathbf{O}_{1:T}$. Aligned results from multiple granularities under scenarios a) and b) reveal the validity of multi-scale prediction-wise integration design in MS-OccFormer.

result in a reduction of 8.5% flow prediction error. Qualitative performance under interactive scenarios (Fig. 3) further corroborates the scene compliance of proposed prediction system. For the key design of multi-scale predictions integration in Ms-OccFormer, qualitative (see Fig. 4) ablations have presented notable results from attention mask update, as well as local integration for prediction consistency.

## References

[1] Long Chen, Yuchen Li, Chao Huang, Bai Li, Yang Xing, Daxin Tian, Li Li, Zhongxu Hu, Xiaoxiang Na, Zixuan Li, et al. Milestones in autonomous driving and intelligent vehicles: Survey of surveys. *IEEE Transactions on Intelligent Vehicles*, 2022. 1

[2] 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. 3

[3] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1, 3

[4] Xin Huang, Xiaoyu Tian, Junru Gu, Qiao Sun, and Hang Zhao. Vectorflow: Combining images and vectors for traffic occupancy and flow prediction. *arXiv preprint arXiv:2208.04530*, 2022. 4

[5] Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. *arXiv preprint arXiv:2303.05760*, 2023. 1, 2

[6] Jinkyu Kim, Reza Mahjourian, Scott Ettinger, Mayank Bansal, Brandyn White, Ben Sapp, and Dragomir Anguelov. Stopnet: Scalable trajectory and occupancy prediction for urban autonomous driving. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8957–8963. IEEE, 2022. 1

[7] Haochen Liu, Zhiyu Huang, Wenhui Huang, Haohan Yang, Xiaoyu Mo, and Chen Lv. Hybrid-prediction integrated planning for autonomous driving. *arXiv preprint arXiv:2402.02426*, 2024. 1, 3

[8] Haochen Liu, Zhiyu Huang, and Chen Lv. Strajnet: Occupancy flow prediction via multi-modal swin transformer. *arXiv preprint arXiv:2208.00394*, 2022. 2

[9] 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, 3, 4

[10] 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, 2

[11] Youshaa Murhij and Dmitry Yudin. Ofmpnet: Deep end-to-end model for occupancy and flow prediction in urban environment. *Neurocomputing*, 586:127649, 2024. 1, 4