Spatiotemporal Decoupling for Efficient Vision-Based Occupancy Forecasting Supplementary Material

Jingyi Xu[1*] Xieyuanli Chen[2*] Junyi Ma[3] Jiawei Huang[4] Jintao Xu[4] Yue Wang[5] Ling Pei[1†]
[1]Shanghai Key Laboratory of Navigation and Location Based Services, Shanghai Jiao Tong University
[2]College of Intelligence Science and Technology, National University of Defense Technology
[3]Department of Automation, Shanghai Jiao Tong University [4]HAOMO.AI
[5]State Key Laboratory of Industrial Control and Technology, Zhejiang University

1. Additional Ablation Results

In this section, we present additional ablation results for occupancy forecasting (OCF) mentioned in Sec. 4 of the main text. In Sec. 1.1, we ablate temporal decoupling for EfficientOCF. In Sec. 1.2, we then present the improvement of our proposed adaptive dual pooling (ADP) strategy for 3D-2D transformation in EfficientOCF. Next, in Sec. 1.3, we report performance changes with different flow formats. Finally, in Sec. 1.4, we evaluate the contributions of multiple head combinations.

1.1. Ablation on Temporal Decoupling

We first study the proposed temporal decoupling by presenting supplementary OCF performance results of EfficientOCF over varying time horizons, as shown in Tab. 1. The experimental results here are extensions of Tab. 3 of the main text. As can be seen, our proposed temporal decoupling consistently enhances the baseline EfficientOCF‡'s performance of all time horizons after using temporal refinement. Moreover, we observe that on the nuScenes-Occupancy dataset [6], the performance improvement becomes more significant with longer time horizons. In contrast, on the nuScenes dataset [1], the performance gains diminishment as the timestep increases. This observation suggests that our proposed temporal decoupling enables EfficientOCF to focus better on the temporal dynamics of movable objects, thereby significantly enhancing its ability to forecast fine-grained occupancy states. In Fig. 1, we further visualize the comparison of TP, FP, and FN results at continuous timesteps before and after instance-aware refinement. The two visualized cases show that our proposed temporal decoupling helps to increase TP and decrease FP predictions. Besides, it removes the predictions of non-existing movable objects, as shown in the blue circles in the lower case of Fig. 1.

1.2. Ablation on Adaptive Dual Pooling

Here we provide an ablation study on adaptive dual pooling in different time horizons. This can be regarded as an extension of Tab. 4 of the main text. As shown in Tab. 2, the forecasting performance of EfficientOCF is comprehensively improved in all IoU metrics by ADP against the single pooling strategy, including the average pooling and max pooling. In particular, ADP improves C-IoU$_c$ prominently compared to other metrics.

1.3. Ablation on Flow Formats

We further examine the impact of flow formats on OCF accuracy. The baseline model is constructed by substituting the backward centripetal flow in EfficientOCF with vanilla backward flow [3, 5]. As shown in Tab.2, the backward centripetal flow yields superior occupancy forecasting performance, particularly in our proposed metrics, C-IoU$_c$ and C-IoU$_f$. These results indicate that backward centripetal flow is more robust to significant flow prediction errors, aligning with key findings in previous studies [2, 4].

1.4. Ablation on Segmentation/Flow/Height Heads

In Tab. 3, we compare the OCF performance of EfficientOCF models with different combinations of prediction heads. The 3D OCF performance is not reported for the baseline with only the segmentation head, as it lacks height estimation to lift the predicted BEV occupancy into 3D space. The best performance is achieved when all heads are utilized across all time horizons. Notably, introducing the flow head results in a more significant performance improvement than the height head. This highlights that predicting future flow effectively captures the sequential motion of movable objects, thereby enhancing occupancy estimation at both present and future timesteps.

*Equal contribution
†Corresponding author
This work was funded by Science and Technology Commission of Shanghai Municipality (24DZ3101300, 24TS1402600, 24TS1402800).

Table 1. Ablation on temporal decoupling in different time horizons

Approach	nuScenes										nuScenes & nuScenes-Occupancy									
	IoU_c (2D)	IoU_f (2D)				IoU_c (3D)	IoU_f (3D)				IoU_c (3D)	IoU_f (3D)				C-IoU_c (3D)	C-IoU_f (3D)			
		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s
EfficientOCF‡	39.44	38.40	37.52	36.71	35.89	32.19	31.63	30.82	30.45	30.08	19.98	18.61	18.21	17.79	17.31	45.52	45.12	44.35	43.65	42.90
EfficientOCF	**39.93**	**38.80**	**37.85**	**36.98**	**36.15**	**35.60**	**34.81**	**34.07**	**33.37**	**32.73**	**21.28**	**20.21**	**19.86**	**19.46**	**19.02**	**47.53**	**47.63**	**46.91**	**46.24**	**45.57**
Improvement	0.49	0.40	0.33	0.27	0.26	3.41	3.18	3.25	2.92	2.65	1.30	1.60	1.65	1.67	1.71	2.01	2.51	2.59	2.59	2.67

Table 2. Ablation on adaptive dual pooling and flow formats in different time horizons

Approach	nuScenes										nuScenes & nuScenes-Occupancy									
	IoU_c (2D)	IoU_f (2D)				IoU_c (3D)	IoU_f (3D)				IoU_c (3D)	IoU_f (3D)				C-IoU_c (3D)	C-IoU_f (3D)			
		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s
Average Pooling	38.04	37.11	36.32	35.55	34.79	33.80	33.32	32.70	32.10	31.53	19.29	18.00	17.65	17.24	16.77	44.02	43.72	43.08	42.43	41.71
Max Pooling	38.84	37.91	37.11	36.33	35.53	34.52	34.05	33.44	32.84	32.23	19.42	18.13	17.77	17.36	16.90	44.69	44.42	43.74	43.07	42.33
ADP	**39.93**	**38.80**	**37.85**	**36.98**	**36.15**	**35.60**	**34.81**	**34.07**	**33.37**	**32.73**	**21.28**	**20.21**	**19.86**	**19.46**	**19.02**	**47.53**	**47.63**	**46.91**	**46.24**	**45.57**
	IoU_c (2D)	IoU_f (2D)				IoU_c (3D)	IoU_f (3D)				IoU_c (3D)	IoU_f (3D)				C-IoU_c (3D)	C-IoU_f (3D)			
		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s
Backward	37.64	36.68	35.87	35.10	34.34	33.32	32.79	32.17	31.57	31.00	19.32	18.04	17.67	17.26	16.80	43.39	43.29	42.69	41.93	41.22
Backward Centripetal	**39.93**	**38.80**	**37.85**	**36.98**	**36.15**	**35.60**	**34.81**	**34.07**	**33.37**	**32.73**	**21.28**	**20.21**	**19.86**	**19.46**	**19.02**	**47.53**	**47.63**	**46.91**	**46.24**	**45.57**
Improvement	2.29	2.12	1.98	1.88	1.81	2.28	2.02	1.90	1.80	1.73	1.96	2.17	2.19	2.20	2.22	3.96	4.14	4.22	4.31	4.35

2. Visualization of OCF Results

In this section, we present additional visualizations of fine-grained OCF results on the nuScenes-Occupancy dataset, comparing our proposed EfficientOCF with the SOTA baseline, OCFNet [4]. As shown in Fig. 2, EfficientOCF consistently forecasts more accurate future occupancy states and captures more precise shapes of movable objects than OCFNet. In the first column of Fig. 2, we highlight the occupancy of a moving vehicle within the blue box, where EfficientOCF predicts more detailed and accurate contours. The second and third columns demonstrate that EfficientOCF exhibits superior velocity awareness for movable objects compared to OCFNet. Additionally, the fourth and fifth columns reveal that OCFNet produces more false detections for invalid objects than our proposed EfficientOCF. Notably, the fifth column also highlights missing annotations in nuScenes-Occupancy compared to the original nuScenes labels (bottom right of the ground-truth subfigure), likely due to LiDAR sparsity. This underscores the necessity of our new metrics, C-IoU, for evaluating fine-grained occupancy forecasting.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[2] Peizheng Li, Shuxiao Ding, Xieyuanli Chen, Niklas Hanselmann, Marius Cordts, and Juergen Gall. Powerbev: a powerful yet lightweight framework for instance prediction in bird's-eye view. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1080–1088, 2023. 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

[4] Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21486–21495, 2024. 1, 2

[5] 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

[6] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023. 1, 3

Table 3. Ablation on segmentation/flow/height heads in different time horizons

Seg. Head	Height Head	Flow Head	nuScenes									nuScenes & nuScenes-Occupancy								
			IoU_c (2D)	IoU_f (2D)				IoU_c (3D)	IoU_f (3D)				IoU_c (3D)	IoU_f (3D)				C-IoU_c (3D)	C-IoU_f (3D)			
				0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s		0.5s	1.0s	1.5s	2.0s
✓			38.48	37.63	36.92	35.84	35.26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
✓	✓		38.55	37.67	36.88	36.10	35.33	34.16	33.61	33.01	32.43	31.87	19.41	18.15	17.80	17.37	16.89	44.09	43.89	43.24	42.56	41.83
✓	✓	✓	**39.93**	**38.80**	**37.85**	**36.98**	**36.15**	**35.60**	**34.81**	**34.07**	**33.37**	**32.73**	**21.28**	**20.21**	**19.86**	**19.46**	**19.02**	**47.53**	**47.63**	**46.91**	**46.24**	**45.57**

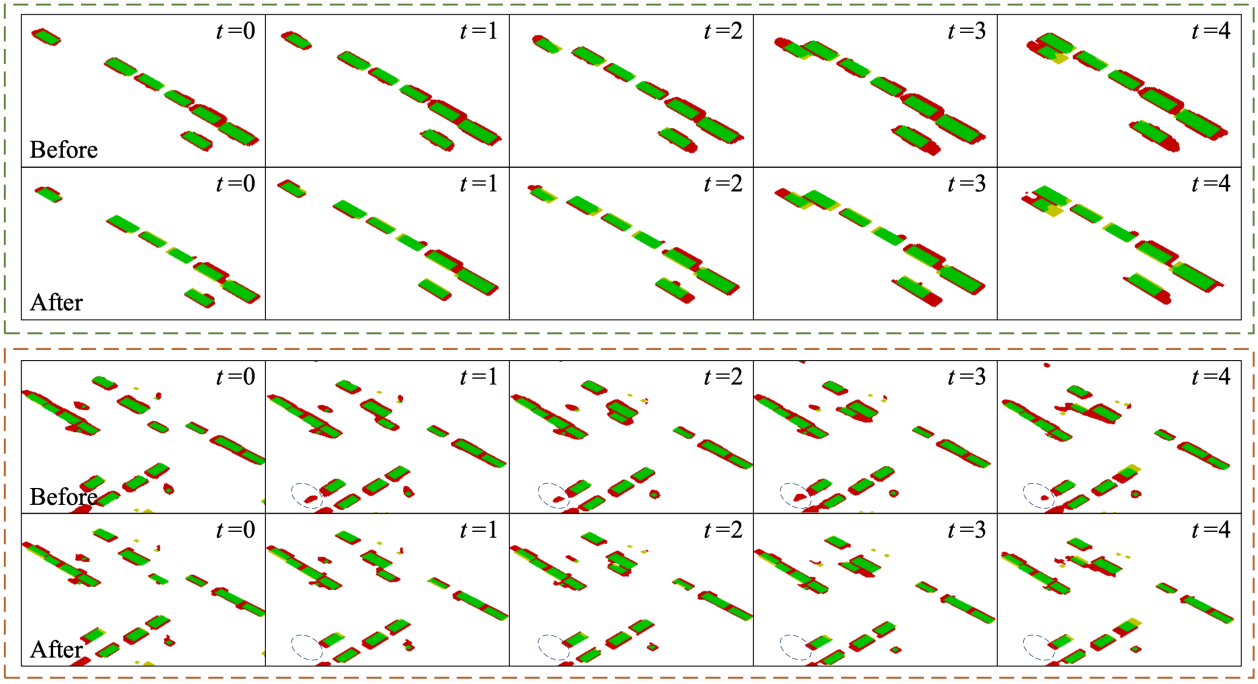

Figure 1. Visualization of TP (green), FP (red), and FN (yellow) results against ground truth at continuous timesteps before and after instance-aware refinement.

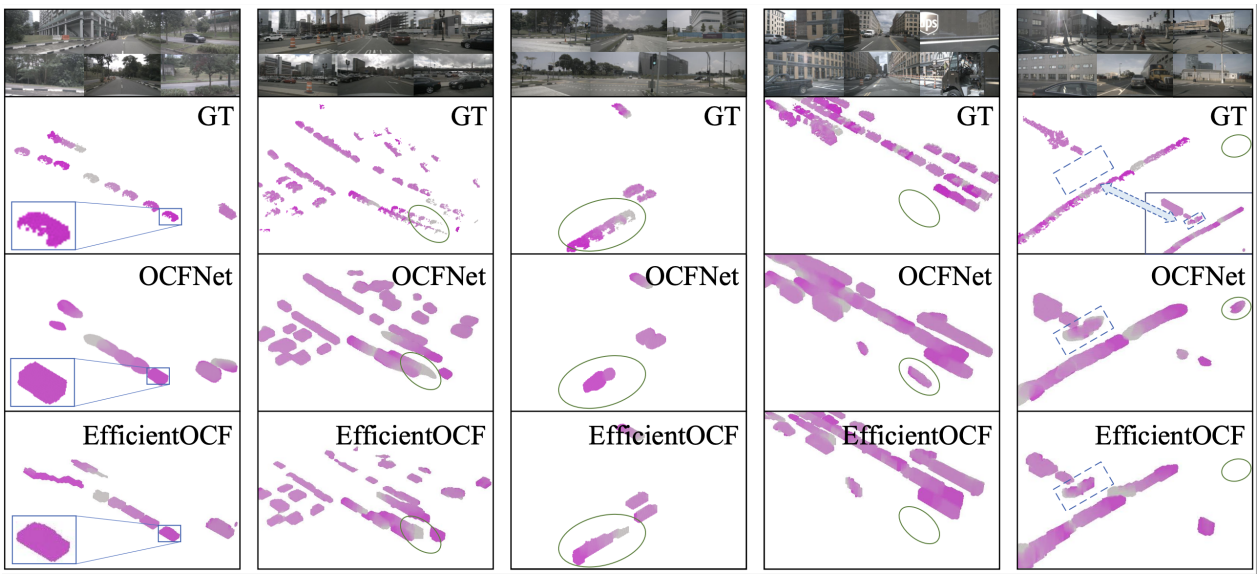

Figure 2. Visualization of OCF results of EfficientOCF, OCFNet, and ground truth from the nuScenes-Occupancy dataset [6]. The occupancy forecasting results and ground-truth labels from timesteps 0 to N_f are assigned colors from dark to light.