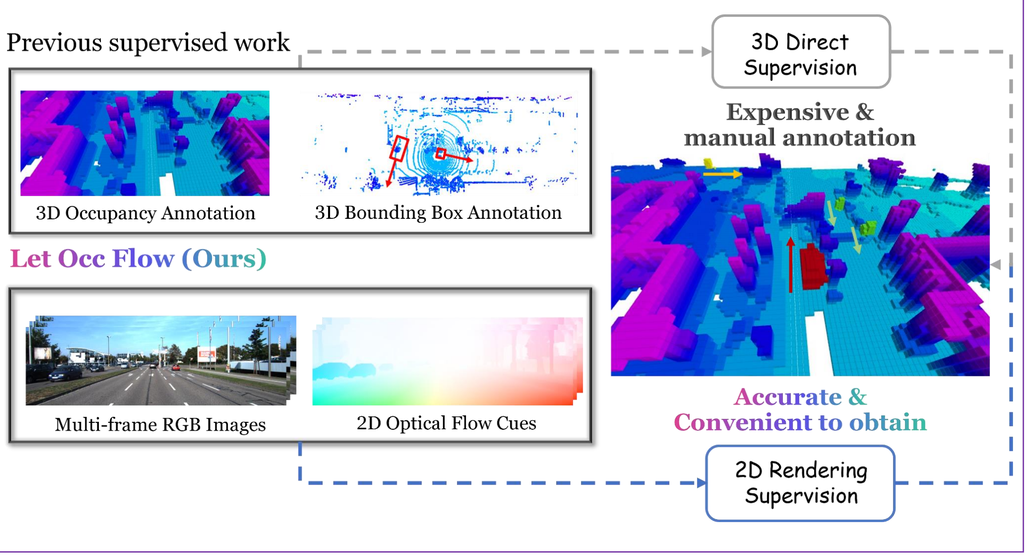
Let Occ Flow: Self-Supervised 3D Occupancy Flow Prediction

Yili Liu[1*], Linzhan Mou[1*], Xuan Yu[1], Chenrui Han[1], Sitong Mao[2], Rong Xiong[1], Yue Wang[1]

Introduction

- Accurate perception of the surrounding environment is crucial in autonomous driving and robotics for downstream planning and action decisions.
- Existing occupancy flow prediction methods rely heavily on **detailed 3D occupancy and occupancy flow annotations,** which complicates their scalability to extensive training datasets.
- Recent 2D Perception Models have shown superior zero-shot generalization capability in comparison to 3D models.
- Our approach **bridges 2D perception cues to enable self-supervised occupancy flow prediction through differentiable rendering.**

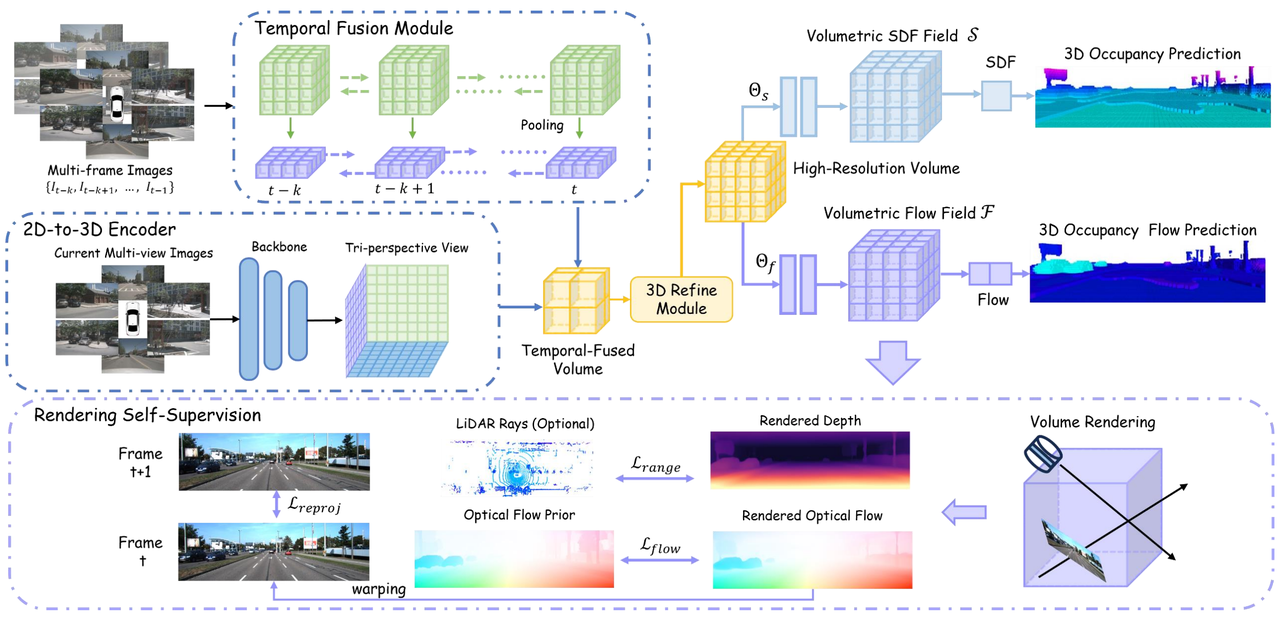


Contribution

- We proposed Let Occ Flow, **the first self-supervised** method for jointly predicting 3D occupancy and Occupancy Flow, by integrating 2D optical flow cues into geometry and motion optimization.
- We designed a novel **attention-based temporal fusion** module for efficient temporal interaction. Furthermore, we proposed a **flow-oriented optimization strategy** to mitigate the training instability and sample imbalance problem.

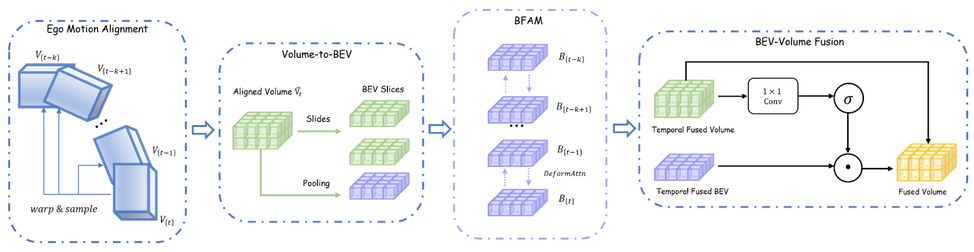
Method

Pipeline



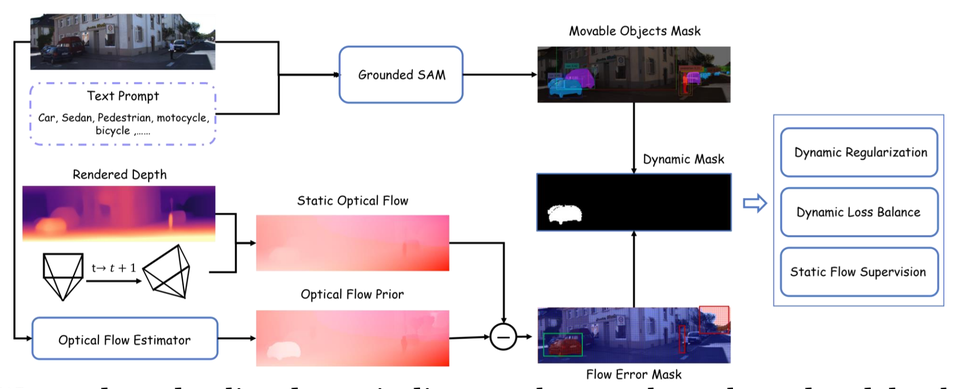
Our **Let Occ Flow** introduced a **vision-based** pipeline for 3D occupancy and occupancy flow prediction. We utilized **Tri-perspective View(TPV)** to extract 3D Volume Features from a temporal sequence of multi-view camera inputs. We then construct the surrounding scene into volumetric SDF and flow field and perform joint occupancy and occupancy flow learning utilizing reprojection consistency, optical flow cues, and optional sparse LiDAR ray supervision via **differentiable volume rendering**.

Attention-Based Temporal Fusion



Our temporal fusion module enhanced the scene representation through ego-motion alignment in voxel space and a Backward-Forward Attention Module (BFAM) in BEV space.

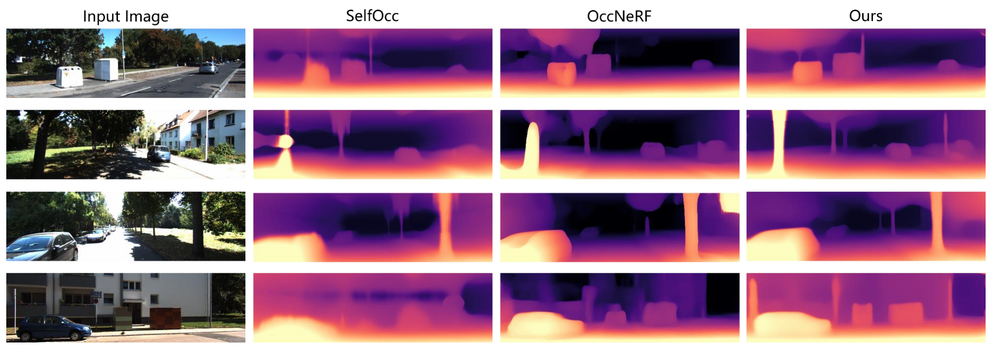
Dynamic Disentanglement Strategy



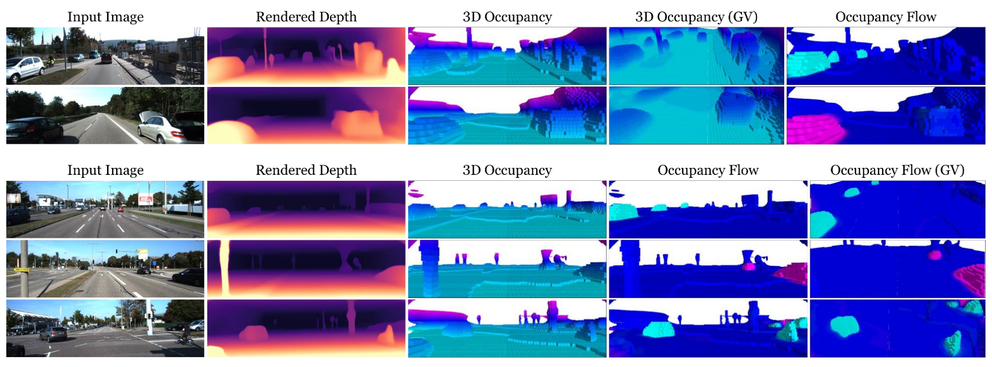
We conducted online dynamic disentanglement through rendered depth, Grounded-SAM and optical flow cues for regularization.

Result

Depth Estimation on Semantic KITTI

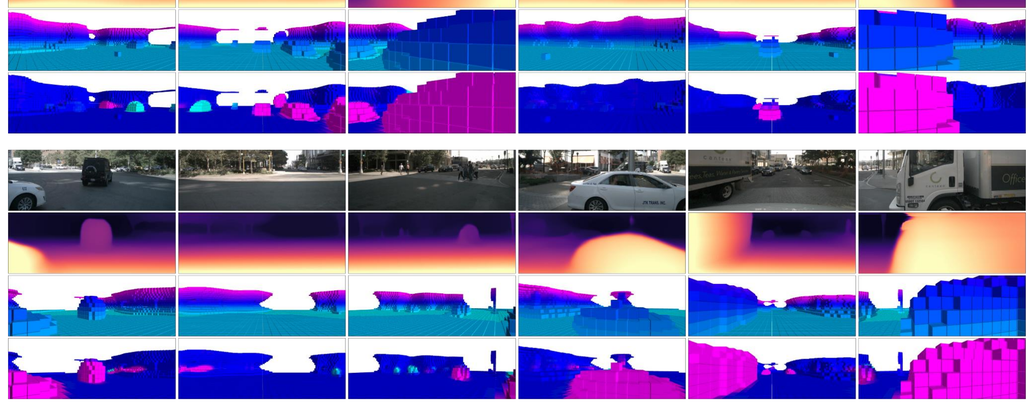


Occupancy and Occupancy Flow Prediction on KITTI-MOT



Occupancy and Occupancy Flow Prediction on nuScenes

Method	Supervision	3D Occupancy & Occupancy Flow RayIoU$_{1m, 2m, 4m}$↑			RayIoU↑	mAVE↓
OccNet [8]	3D	**29.28**	**39.68**	50.02	39.66	1.61
OccNeRF-C* [5]	C	9.93	19.06	35.84	21.61	1.53
Ours-C	C	17.49	28.52	44.33	30.12	**1.42**
RenderOcc* [7]	L	20.27	32.68	49.92	36.67	1.63
OccNeRF-L* [5]	C+L	16.62	29.25	49.17	31.68	1.59
Ours-L	C+L	25.49	39.66	**56.30**	**40.48**	1.45



Limitation & Future Work

Limitation

- Although we use temporal sequence input to better exploit the historical information, our model cannot completely handle **the occlusion problem** due to the inherent rendering-based limitation.
- The accuracy of occupancy flow prediction relies on the quality of optical flow cues.
- Our occupancy flow prediction does not explicitly enforce **consistency within instances**.

Future Work

- Subsequent research could investigate **long-term occupancy modeling** and solutions to leverage the temporal sequence supervision to scale up the visible range of perspective.
- Future work may explore to integrate **instance perception** into occupancy flow prediction.

Contact

About Me

Yili Liu
ylliu01@zju.edu.cn
Zhejiang University

3D Perception on Autonomous Driving & Driving Scene Reconstruction

Project Homepage