CascadeFlow: 3D Occupancy and Flow Prediction with Cascaded Sparsity Sampling Refinement Framework

Zhimin Liao, Ping Wei
Xi'an Jiaotong University

Abstract

In this technical report, we present our solution, named CascadeFlow, for the Vision-Centric 3D Occupancy and Flow Prediction track of Autonomous Grand Challenge at CVPR 2024. Our proposed solution, CascadeFlow, builds upon CascadeOcc, a cascade sparsity sampling refinement framework for vision-based occupancy prediction. Based on CascadeOcc, and considering the scale-invariant and temporal coherence of scene flow, we propose a flow embedding that learns the relationships between adjacent occupancy in a 3D scene and refines them stage by stage within the cascade framework. Thanks to the sparse refinement and cascade design, our model achieves high performance: 5th place in RayIoU with a score of 37.8, 2nd place in MAVE with a score of 0.31, and 5th place in occupancy score with a score of 40.89. This is accomplished without the need for test-time augmentation (TTA), post-processing, or model ensembling, and even with a lightweight setting requiring only 7.3 GB of memory for training.

1. Introduction

Due to the ability of 3D occupancy representation to offer a more fine-grained depiction of 3D scenes compared to 3D bounding boxes, 3D occupancy prediction plays an important role in autonomous vehicles. Additionally, the capability to forecast future environmental conditions is essential for advanced collision avoidance and trajectory optimization methods. Therefore, the challenge of occupancy and flow prediction is crucial for performing downstream tasks safely and reliably in autonomous driving applications.

Our solution builds upon CascadeOcc, which differs from other sparse methods [9, 16, 19, 23] that utilize a one-off decision to select non-empty voxels using a set threshold or top-k method. CascadeOcc employs a cascade design to refine proposal voxels, combined with a probability sampling method to select seed voxels. It sparsely refines coarse features in a coarse-to-fine manner, with deeper stages providing a more comprehensive perception of the 3D scene. We extend CascadeOcc for the flow prediction task and propose CascadeFlow. Considering the scale-invariant and temporal coherence of scene flow, we propose a flow embedding that learns the relationships between adjacent frames and employs a cascade refinement design to refine the flow embedding in multiple stages.

In this challenge, we aims to propose a resource-efficient method to tackle this challenge rather than simply enlarge the scale of model to achieve better performance. Our method requires 7.3G of memory for training, and achieves 5th place in RayIoU with a score of 37.8, 2nd place in MAVE with a score of 0.31, and 5th place[1] in occupancy score with a score of 40.89.

2. Our Solution

2.1. Model Design

Our solution, CascadeFlow, builds upon a 3D occupancy prediction method termed CascadeOcc. Here, we provide a brief introduction to facilitate a better understanding of CascadeFlow. Due to the intrinsic properties of density and redundancy in 3D space, simply using dense methods [10, 21, 22, 25] to process 3D features is not appropriate. Generally speaking, CascadeOcc primarily follows the forward-backward projection paradigm [11, 12], as shown in Fig. 1. It refines the coarse voxel features generated by forward projection methods [5–8] using a sparse technique. Our forward projection structure is based on the BEVStereo [7] method. CascadeOcc introduces a sparse refinement module that employs successive Transformer [10] layers to enhance the non-empty voxel features. This is combined with occlusion-aware spatial cross-attention (OA-SCA) and a probability sampling method to address the cumulative errors found in previous sparse methods [9, 14, 16, 19]. Additionally, to improve the geometry of features created by the Sparse Decoder, it utilizes volume rendering techniques [18] to provide depth supervision during training.

For CascadeFlow, considering the scale-invariant and temporal coherence of scene flow, we utilize the cascade design of the Sparse Decoder to predict the scene flow in a coarse-to-fine manner. Additionally, we propose using flow

[1] The rankings are captured up to the time when the report is submitted.

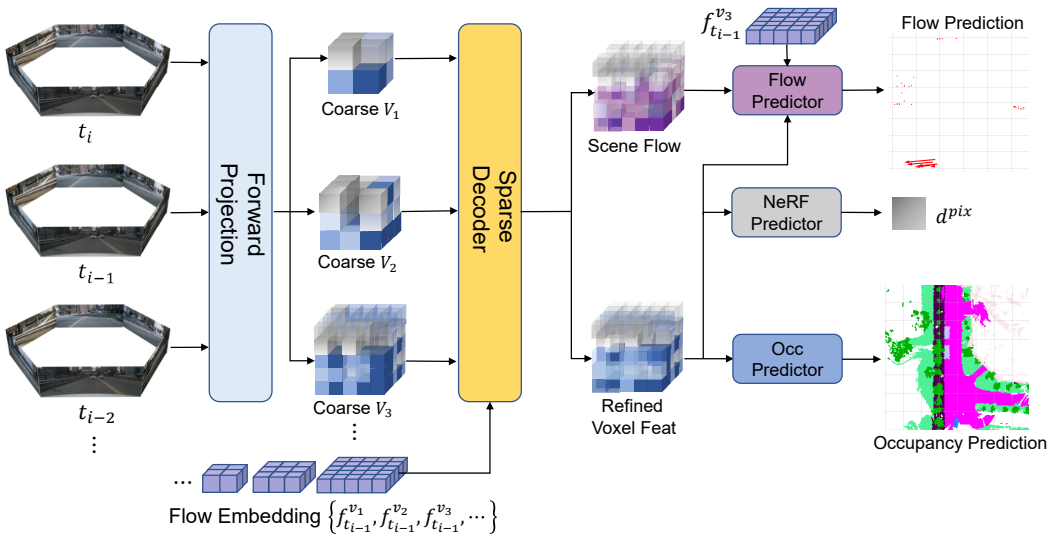

Figure 1. Model framework. The Sparse Refinement module utilizes a sparse method to refine coarse voxel features and employs the flow embedding from the previous frame to produce the scene flow at each stage in a residual design.

embedding to represent the relationship between adjacent frames. Instead of directly estimating the scene flow, we adopt a residual flow learning structure to refine the scene flow estimation.

We first utilize the forward projection method to produce multi-scale coarse voxel features and then employ the Sparse Decoder to progressively refine these voxel features using a sparse approach. As depicted in Fig. 2, the Sparse Decoder uses successive Transformer layers to refine the coarse voxel features and scene flow. The Transformer layers mainly contain self-recursive occupancy and flow predictors to refine the proposed voxel features and the scene flow. The flow predictor follows a residual design to refine the scene flow layer by layer, as depicted in Fig. 3. It utilizes the flow embedding to produce the residual scene flow, and the output scene flow is computed by adding the flow from the last layer to the residual scene flow. The occupancy predictor generates non-empty proposal voxels and refines these non-empty proposal voxels layer by layer. Additionally, it uses probability sampling to select seed voxels for refinement. The OA-SCA (Occlusion-Aware Spatial Cross-Attention) then refines these seed voxels using the proposed occlusion weight, which includes depth consistency [11] and the occupied weight produced by the occupancy predictor. This approach addresses projection errors caused by inaccurate depth estimation. At the end of each stage, the refined BEV (Bird's Eye View) features and scene flow are upsampled to serve as the input for the next stage. Furthermore, the scene flow is combined with the last frame's flow embedding to generate the current frame's flow embedding, which serves as the input for the next frame.

With the cascaded refinement design, CascadeFlow obtains the required resolution voxel features and scene flow at the end of the Sparse Decoder. We use a simple MLP (Multi-Layer Perceptron) as the occupancy predictor and a similarly structured flow predictor, as described above, to predict the final scene flow. Additionally, to incorporate geometry information into the backward projection method, we use the NeRF [20] technique to supervise the depth information. The NeRF predictor employs a simple MLP to generate the density volume.

2.2. Training Loss

To train the model, in each Sparse Decoder stage, we use the distance-aware Focal loss L_{focal} inspired by FB-Occ [11], Lovasz loss L_{lov}, and affinity losses L_{scal}^{geo} and L_{scal}^{sem} from MonoScene [2] to supervise the occupancy prediction. For the scene flow prediction, we utilize the weighted-L1 loss, where the weights are set to 0.1 for zero-speed areas. In each stage, we control the weight using a factor defined as w, which is computed by $\frac{1}{2^{N-i}}$, where N is the number of stages and i is the current stage index. We utilize cross-entropy loss to supervise the depth in the forward projection method and SILog loss [3] for NeRF supervision.

3. Experiments

3.1. Datasets and Metrics

Dataset. The challenge dataset, nuScenes OpenOcc [21], is built based on the existing nuScenes dataset [1]. Each scene is paired with six surround-view images. For this challenge, each sample is defined within a spatial range of $[-40m, -40m, -1m, 40m, 40m, 5.4m]$ and the voxel resolution is $200 \times 200 \times 16$. The 3D voxel semantics include 17 classes, with flow labels assigned only to the foreground classes, such as cars and pedestrians.

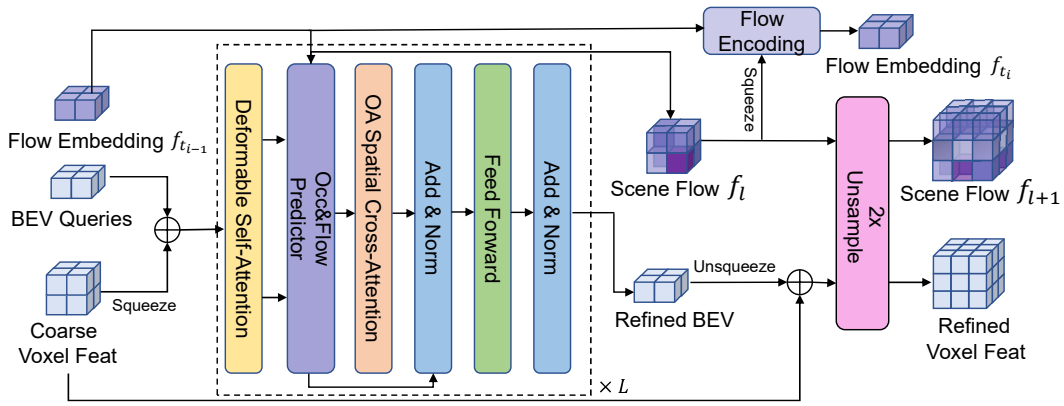

Figure 2. Sparse Decoder. The Sparse Decoder takes the coarse voxel features produced by the forward projection method and refines them using successive Transformer layers. It utilizes the flow embedding produced by the previous frame with a residual flow predictor to progressively refine the scene flow. At the each stage of Sparse Decoder, it unsample $2\times$ to server as the next stage input.

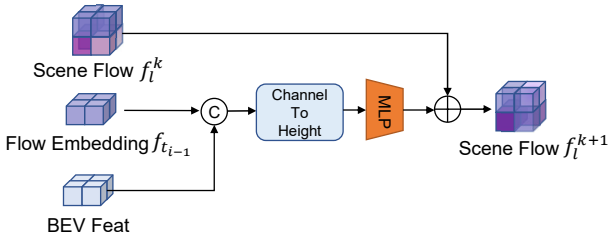

Figure 3. The architecture of Flow Predictor in Sparse Decoder. We utilize the method introduced by FlashOcc [24] to convert BEV features into voxel features.

Metrics. The ranking for this challenge is determined by the occupancy score, which consists of two parts: Ray-based mIoU [15] and absolute velocity error for occupancy flow. The RayIoU is computed at three distance thresholds: 1, 2, and 4 meters. For flow prediction, it measures the velocity errors for a set of true positives, and the absolute velocity error (AVE) is calculated only for the foreground classes, such as cars, trucks, and bicycles. The final occupancy score is defined as a weighted sum of the mean RayIoU (MIoU) and the mean AVE (MAVE):

$$Score = MIoU * 0.9 + max(1 - MAVE, 0.0) * 0.1 \quad (1)$$

3.2. Implementation Details

We crop the source image to a size of 256×704. We employ common data augmentation strategies similar to those used in BEVDetOcc [6], including flipping and rotation in both image and 3D space. For the image backbone, we use ResNet-50 [4] and employ BEVStereo [7] as the forward projection method. Voxel pooling is applied to create voxel features with a resolution of $100 \times 100 \times 8$. We then utilize ResNet3D [4] and FPN3D [13] to produce multi-scale voxel features ranging from $25 \times 25 \times 2$ to $100 \times 100 \times 8$. The Sparse Decoder is composed of three stages, each containing 3 layers of Sparse Refinement modules. Additionally, for NeRF supervision, we emit 1000 rays per image plane and uniformly sample 96 points along each ray.

We use a global batch size of 16 across 8 NVIDIA RTX 4090 GPUs. The AdamW optimizer [17] is employed with a learning rate of 1×10^{-4}. A linear warm-up is applied during the first 200 iterations. We utilize 8 temporal frames and follow BEVDet4D [5] to fuse different frame features.

3.3. Ablations

We use a lightweight of BEVDetOcc [6] as our baseline, refereed as Version A, which utilize the ResNet-50 [4] and FPN [13] as the image encoder, and output the coarse voxel feature with a resolution of $100 \times 100 \times 8$. A simple MLP is employed as the occupancy prediction head. For the flow prediction, we assume the scene is static, setting all values to zero. For Version B, we add the utilize the same backward refinement as FB-Occ [11]. For Version C, we replace the dense backward refinement of bevformer [10] with our proposed Sparse Decoder. For Version D, we add the NeRF supervsion into the training phase. For Version E, we extend the original single-stage model to multi-stage which refine the coarse voxle features stage by stage. For Version F, we incorporate information from 8 temporal frames into the model, similar to BEVDet4D [5]. For Version G, we integrate the flow predictor into the framework. The ablation study results are shown in Table 1.

3.4. Main Results

We utilize a lightweight model with a ResNet-50 image backbone and an input image size of 256×704. This configuration requires only 7.3 GB of memory for training and achieves a RayIoU of 37.8 and an MAVE of 0.31. These results demonstrate the potential of our proposed method. The results are shown in Table 2.

Method	$RayIoU_{1m}$(%)	$RayIoU_{2m}$(%)	$RayIoU_{4m}$(%)	RayIoU(%)	MAVE	OccScore
Version A	26.8	34.1	38.8	33.2	1.73	29.88
Version B	28.1	34.8	39.8	34.2	1.73	30.81
Version C	29.3	36.2	40.6	35.3	1.73	31.77
Version D	29.7	36.9	41.5	36.0	1.73	32.40
Version E	33.0	39.3	43.1	38.5	1.73	34.62
Version F	**34.2**	**41.2**	**45.7**	**40.3**	1.73	36.30
Version G	33.5	40.3	45.0	39.6	**0.47**	**40.94**

Table 1. 3D occupancy and flow prediction performance of different setting on the nuScenes OpenOcc [21] val set.

Bacbkone	InputSize	Set	Memory(G)	$RayIoU_{1m}$(%)	$RayIoU_{2m}$(%)	$RayIoU_{4m}$(%)	RayIoU(%)	MAVE	OccScore
R50	256×704	val	7.3	33.5	40.6	45.3	39.6	0.47	40.94
R50	256×704	test	7.3	32.4	38.5	42.3	37.8	0.31	40.89

Table 2. Main Results on the nuScenes Openocc [21] val and test set.

4. Conclusions

We propose CascadeFlow, a multi-stage sparse refinement method designed to efficiently process coarse voxel features and scene flow. Without using a heavy model configuration or relying on test-time augmentation, model ensembles, or post-processing, our model achieves impressive performance in this challenge.

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 (CVPR)*, 2020. 2

[2] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 2

[3] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Cornell University - arXiv,Cornell University - arXiv*, 2014. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[5] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1, 3

[6] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 3

[7] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1486–1494, 2023. 1, 3

[8] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 1

[9] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 1

[10] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 3

[11] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 1, 2, 3

[12] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6919–6928, 2023. 1

[13] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[14] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 1

[15] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023. 3

[16] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia

Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023. 1

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Learning,Learning*, 2017. 3

[18] N. Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, page 99–108, 1995. 1

[19] Jianbiao Mei, Yu Yang, Mengmeng Wang, Junyu Zhu, Xiangrui Zhao, Jongwon Ra, Laijian Li, and Yong Liu. Camera-based 3d semantic scene completion with sparse guidance network. *arXiv preprint arXiv:2312.05752*, 2023. 1

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[21] Chonghao Sima, Wenwen Tong, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as occupancy. 2023. 1, 2, 4

[22] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1

[23] Junkai Xu, Liang Peng, Haoran Cheng, Linxuan Xia, Qi Zhou, Dan Deng, Wei Qian, Wenxiao Wang, and Deng Cai. Regulating intermediate 3d features for vision-centric autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6306–6314, 2024. 1

[24] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 3

[25] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 1