Multi-Scale Occ: 4th Place Solution for CVPR 2023 3D Occupancy Prediction Challenge

Yangyang Ding* Luying Huang* Jiachen Zhong
SAIC AI Lab
{dingyangyang01, huangluying, zhongjiachen}@saicmotor.com

Abstract

In this report, we present the 4th place solution for CVPR 2023 3D occupancy prediction challenge. We propose a simple method called Multi-Scale Occ for occupancy prediction based on lift-splat-shoot framework, which introduces multi-scale image features for generating better multi-scale 3D voxel features with temporal fusion of multiple past frames. Post-processing including model ensemble, test-time augmentation, and class-wise thresh are adopted to further boost the final performance. As shown on the leaderboard, our proposed occupancy prediction method ranks the 4th place with 49.36 mIoU.

1. Introduction

Recently, 3D occupancy prediction has attracted extensive attention as it is crucial for autonomous driving systems to understand geometric and semantic information in 3D scenes. 3D occupancy is more accurate and suitable for describing objects in arbitrary shape and undefined classes at a fine-grained level. 3D occupancy prediction may be performed using various modality input (e.g. LiDAR, Radar, Image). In the challenge, we focus on predicting 3D occupancy from pure multi-camera images which still remains as a challenging problem.

Based on previous perception tasks including 3D object detection and semantic map segmentation, current multi-camera 3D perception methods mainly lies in two types: 1) lift-splat-shoot(LSS)-based [3, 5, 16], lifting 2D image features to plausible 3D volume space via implicit or explicit depth estimation. 2) Transformer-based [8, 9, 20], which define 3D queries in 3D volume space and use transformer attention mechanism to query corresponding 2D image features. We build our method based on LSS-style framework, improved with long-term temporal stereo matching and multi-scale 3D features fusion to learn spatial and temporal details simultaneously. Besides, we use decoupled head to perform occupancy and semantic prediction separately in order to relieve the extreme imbalanced distribution between unoccupied voxel grids and those grids occupied by semantic classes.

*Equal Contribution.

2. Method

The overall architecture of our approach is shown in Fig. 1. Given N camera images with T timestamps, we first use a 2D image encoder to extract M scale features (Sec. 2.1). Image features are then lifted to a 3D voxel feature followed by a long-term temporal feature aggregation of past frames independently on each scale to construct multi-scale 3D representation of the current frame (Sec. 2.2). To fuse multi-scale 3D features thoroughly, we use a lightweight 3D UNet [1] to integrate local and global geometric and semantic information (Sec. 2.3). We perform occupancy and semantic prediction separately on the largest resolution with 2 decoupled heads. Multi-scale supervision is used during training to facilitate the convergence (Sec. 2.4). Finally, model ensemble, test-time augmentation, and class-wise thresh are applied to further improve the performance (Sec. 2.5).

2.1. 2D Backbone

For pure vision based 3D perception, 2D image backbone plays a crucial role in feature extraction from image. We use 2D backbone with a FPN [11] to generate multi-scale 2D features. Specifically, given N camera images with T timestamps, the 2D backbone firstly extract three scales features for each image in strides of $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$. Then, a simple FPN is applied to fuse features from different receptive fields and output three level features with the same channels for building 3D representation later.

In order to obtain better performance and strong generalization ability, we use Dual-InternImage-B which are composed by two InternImage-B [19] connected via DHLC (Dense Higher-Level Composition) method [10] as the backbone of our main model. Moreover, we use a single Swin-B [13] to train an extra simpler model for ensemble.

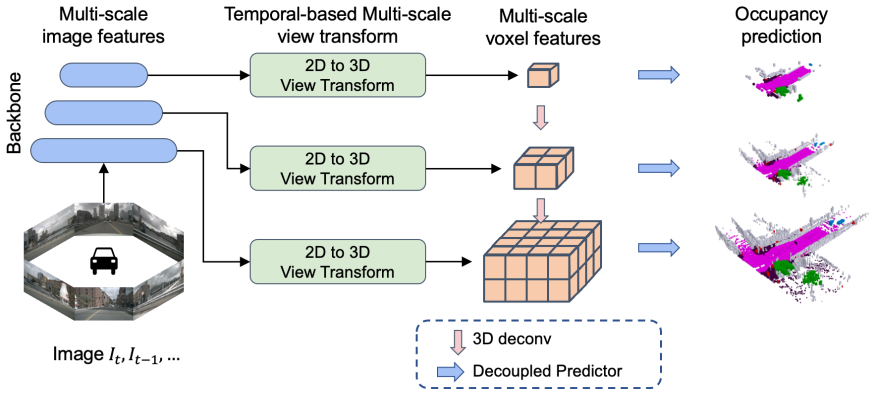

Figure 1. The architecture of our proposed Multi-scale Occ framework for 3D occupancy prediction.

2.2. Long-term Temporal Stereo Matching

To project multi-view and multi-timestamp 2D features to 3D space, we apply LSS view transformer [16] with BEVPoolv2 [4] to create 3D voxel features at each scale. In LSS-based mechanism, accurate depth estimation is important to performance [7] and temporal stereo style methods are very helpful in improving model depth prediction ability [6, 15, 21]. Therefore, following [15], we use the total of T = 9 timestamps(1 current frame + 8 past frames) to create cost volumes to enhance depth estimation. Specifically, we create total of 8 frames cost volumes using image features at the stride of $\frac{1}{4}$ between each two adjacent frames. Then those cost volumes are rescaled to the strides of $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ and concatenated with image features at each corresponding scale to predict the depth. Notice that, the earliest frame is dropped after building cost volumes, thus, 8 frames of 3D voxel features are lifted in total. We use ego-motion to align previous frames to current frame following BEVDet4D [3] and concatenate them along channel dimension. Finally, we feed the concatenated temporal aligned feature to ResNet and LSS-FPN following [5] but replacing the 2D convolution with 3D to further fuse temporal information. Notice that, as shown in 1, we project 3 scales of image features independently to construct 3 different scales of 3D voxel features. Besides depth cost volume are constructed at $\frac{1}{4}$ stride, the temporal fusion happens also individually in different scales of 3D voxel features.

2.3. Multi-scale 3D Features Fusion

In computer vision tasks, multi-scale features fusion and supervision often achieve better results [11, 18]. Recent work [20] indicates that constructing 3D representation at multi-scale independently improves performance in occupancy prediction. Inspired by the idea, as described in Sec. 2.2, we construct 3 different scales of 3D voxel features ($50 \times 50 \times 4$, $100 \times 100 \times 8$, and $200 \times 200 \times 16$) individually from corresponding scales of 2D image features ($\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$). Different from [20] which does not utilize temporal information, we perform extra long-term temporal fusion individually at each scale of 3D voxel features. After individually building 3 scales of 3D voxel features, we use a 3D U-Net [1] to further fuse multi-scale 3D features. Supervision is applied at each scale.

2.4. Decoupled Prediction Head

Class-imbalanced problem naturally exists in real-world vision tasks which also occurs in this challenge. The class of free, which indicates the occupancy status of voxel grid, occurs much more frequent than other semantic classes (free class takes around 96% in training set). To tackle this problem, we decouple the occupancy prediction head and semantic prediction head. Specifically, given the 3D feature volume V_i generated at each scale $i = 0, 1, 2$, two parallel prediction heads with different output channels are applied to predict whether each voxel is occupied or not (1 class), and which semantic label (16 classes) it belongs to. Each head is a simple 2-layer MLP with Softplus activation.

For occupancy prediction head, we use binary cross-entropy loss L_i^{occ}. For semantic prediction head, we adopt multi-class focal loss L_i^{sem} [12]. Besides, we only compute the loss of those voxel grids which can be observed in the current camera view by using the binary voxel mask $mask_i^{cam}$ during training:

$$L_i^{occ} = BCE(V_i, GT_i^{occ}) * mask_i^{cam}, \quad (1)$$

$$L_i^{sem} = FL(V_i, GT_i^{sem}) * mask_i^{cam}, \quad (2)$$

where GT_i^{occ} is the geometric ground truth with $\{0, 1\}$ label (0 for unoccupied and 1 for occupied). GT_i^{sem} is the semantic ground truth with 16 classes. A lower resolution of GT_i^{sem} is obtained from majority vote pooling [17] of the full size ground truth and GT_i^{occ} is obtained via max pooling. For better handling data imbalance, we also re-weight each class loss with the inverse of the class-frequency as

in [17], which is applied to both occupancy head and semantic head individually. Finally, our model is trained by minimizing the following objective:

$$L_i = L_i^{occ} + L_i^{sem} + L_i^{depth}, \tag{3}$$

$$L_{total} = \sum_{i=0}^{2} \alpha_i L_i, \tag{4}$$

We rescale the loss on i-th scale via $\alpha_i = \frac{1}{2^i}$ to make sure larger resolution prediction plays more important role in training.

2.5. Post-process

In order to further enhance the performance of our model, we adopt model ensemble and test-time augmentation. We train one extra simpler model with single Swin-B [13] as backbone without multi-scale supervision. We apply test-time augmentation separately to both models. We use image horizontal flip, 3D space horizontal and vertical flips as augmentation. Each model inferences 8 times, so that we can obtain 16 occupancy predictions and 16 semantic predictions for each frame. Due to performance difference between the two models, we weight Dual-InternImage-B model with coefficient 0.55 and Swin-B model with 0.45. We fuse the prediction results to obtain final results P_{occ} and P_{sem} using the following formulation:

$$P_{occ} = 0.45 \sum_{i=1}^{16} P_{occ1_i} + 0.55 \sum_{i=1}^{16} P_{occ2_i} \tag{5}$$

$$P_{sem} = argmax(0.45 \sum_{i=1}^{16} P_{sem1_i} + 0.55 \sum_{i=1}^{16} P_{sem2_i}) \tag{6}$$

where P_{occ1_i} and P_{sem1_i} denote as the i-th test-time augmentation prediction probability of the Swin-B model, P_{occ2_i} and P_{sem2_i} stand for the prediction of Dual-InternImage-B model.

To obtain the ultimate prediction, we select threshold for each category. If the occupancy prediction for a voxel grid is below the threshold, it is considered as unoccupied. The threshold for each category are shown in Table 1.

Class	Threshold
Others	0.92
Barrier	0.94
Bicycle	0.94
Bus	0.94
Car	0.93
Construction Vehicle	0.93
Motorcycle	0.91
Pedestrian	0.91
Traffic Cone	0.91
Trailer	0.93
Truck	0.93
Driveable Surface	0.96
Other Flat	0.95
Sidewalk	0.95
Terrain	0.95
Manmade	0.93
Vegetation	0.92

Table 1. Threshold for each class during post-process.

3. Experiments

3.1. Experimental Setup

Dataset. The challenge dataset contains 28130 frames for training, 6019 for validation, and 6008 for testing respectively. Each frame contains 6 views of camera images with 1600×900 resolution.

Architecture. We initialize the two InternImage-B backbones with the same weight from the official repository [19] which is trained for COCO object detection and instance segmentation task via Mask-RCNN [2] method. Swin-B backbone is initialized with the weight from BEVDet4D [3] official repository which is trained for nuScenes 3D object detection task. The input image resolution is resized to 512×1408 during training and inference.

Training Details. We use AdamW optimizer [14] with a constant learning rate 2e-4 through training and apply Exponential Moving Average(EMA) strategy with average factor 0.999 to update our model. Our Swin-B model are trained for 24 epochs with weight decay 0.01 on 14 Tesla V100 GPUs. And our Dual-InternImage-B model are totally trained for 31 epochs, in which the first 6 epochs is trained with weight decay 0.05 on 24 Tesla V100 GPUs and the last 25 epochs is trained with weight decay 0.01 on 31 Tesla V100 GPUs. We keep batch size on single GPU equal to 2 for all training. Both two models are trained on training and validation set for leaderboard submission.

3.2. Main Results

Our three submission results are presented in 2, each submission improves the performance compared with previous one. Due to the limitation of computation resources, we are unable to perform comprehensive ablation study of different settings. We may only roughly summarize that stronger image backbone, multi-scale 3D voxel features, model ensemble, and test-time augmentation contribute positively to the performance of this task. Our best submission ranks the 4th place with 49.36 mIoU on the leaderboard.

Method	mIoU
Swin-B + Prev8	46.55
Dual-InternImage-B + Prev8 + Multi-Scale	48.00
Model Ensemble + TTA	49.36

Table 2. Submission results on the test set.

4. Conclusion

In this technical report, we present our Multi-Scale Occ solution for 3D occupancy prediction. In order to achieve better prediction results, we construct multi-scale 3D voxel features with long-term temporal information fusion individually at each scale. 3D U-Net and decoupled heads are applied to perform fusion and prediction of multiple scales. In addition, we conduct model ensemble, test-time augmentation, and class-wise thresh selection to enhance the performance. Our overall 3D occupancy prediction framework achieves the 4th place in the CVPR 2023 3D occupancy prediction challenge.

References

[1] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 1, 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 3

[3] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1, 2, 3

[4] Junjie Huang and Guan Huang. Bevpoolv2: A cutting-edge implementation of bevdet toward deployment. *arXiv preprint arXiv:2211.17111*, 2022. 2

[5] Junjie Huang, Guan Huang, Zheng Zhu, Ye Yun, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2

[6] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo, 2022. 2

[7] 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. *arXiv preprint arXiv:2206.10092*, 2022. 2

[8] 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 (CVPR)*, 2023. 1

[9] 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. *arXiv preprint arXiv:2203.17270*, 2022. 1

[10] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnet: A composite backbone network architecture for object detection. *IEEE Transactions on Image Processing*, 31:6893–6906, 2022. 1

[11] 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. 1, 2

[12] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. 1, 3

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[15] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. 2023. 2

[16] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*, 2020. 1, 2

[17] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion, 2020. 2, 3

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[19] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022. 1, 3

[20] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2303.09551*, 2023. 1, 2

[21] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2