OccTransformer: Improving BEVFormer for 3D camera-only occupancy prediction

Jian Liu[1] Sipeng Zhang[2] Chuixin Kong[3] Wenyuan Zhang[8] Yuhang Wu[7] Yikang Ding[1]
Borun Xu[3] Ruibo Ming[4] Donglai Wei[5] Haotian Yao[4] Xiaoming Zhang[6]
Jianming Hu[4] Lihui Peng[4] Xianming Liu[1*]
[1]HIT [2]ZJU [3]UESTC [4]THU [5]FDU [6]SWJTU [7]HUST [8]XJTU

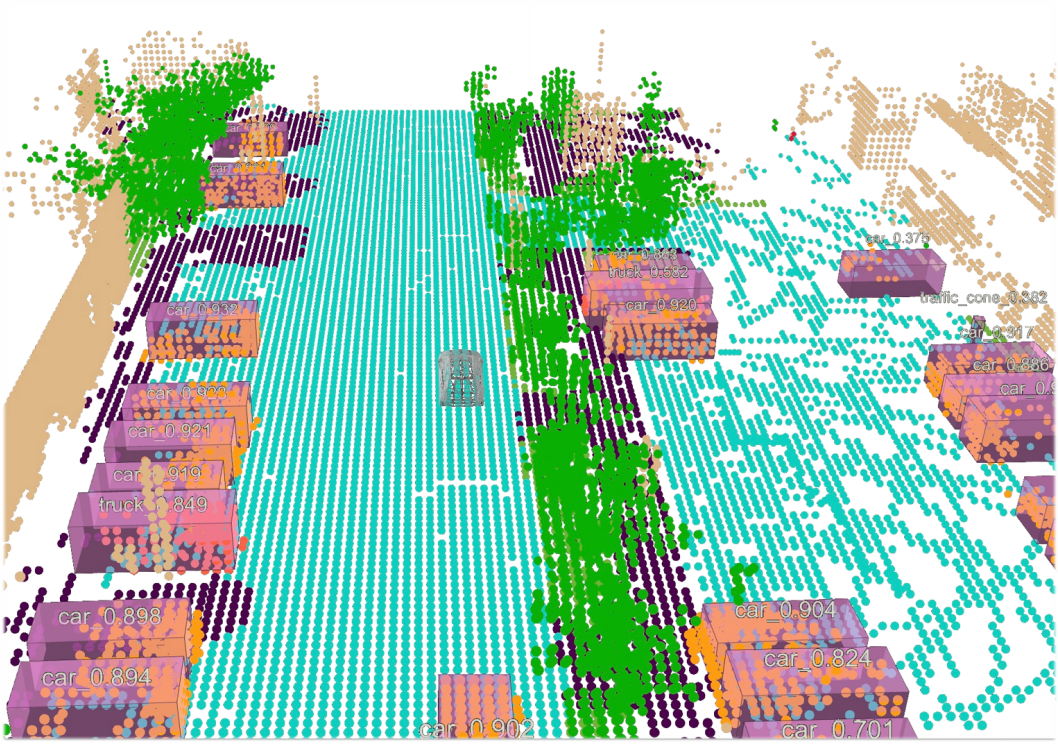

Figure 1: Occ results on the testing set of nuScences Dataset 3D Occupancy prediction track. BBox is 3D detection results.

Abstract

This technical report presents our solution, "occTransformer," for the 3D occupancy prediction track in the autonomous driving challenge at CVPR 2023. Our method builds upon the strong baseline, BEVFormer [1], and improves its performance through several simple yet effective techniques. Firstly, we employed data augmentation to increase the diversity of the training data and improve the model's generalization ability. Secondly, we used a strong image backbone to extract more informative features from the input data. Thirdly, we incorporated a 3D Unet Head to better capture the spatial information of the scene. Fourthly, we added more loss functions to better optimize the model. Additionally, we used an ensemble approach with the occ model BevDet [2] and surroundOcc [3] to further improve the performance. Most importantly, we integrated 3D detection model StreamPETR [4] to enhance the model's ability to detect objects in the scene. Using these methods, our solution achieved 49.23 miou on the 3D occupancy prediction track in the autonomous driving challenge.

1. Introduction

The nuScenes image-based 3D occupancy prediction challenge at CVPR 2023, held from May 12th to 12th, 2023, is the largest and most exciting competition for challenging perception tasks in autonomous driving. This challenge

* Corresponding author: Xianming Liu (csxm@hit.edu.cn)

was endorsed by Mobileye at CES 2023 and Tesla AI Day 2022, highlighting the importance of advancing perception technology for autonomous driving. The objective is to predict the current occupancy state and semantics of each voxel grid in the scene, given images from multiple cameras.

To tackle this challenge, our approach involves several steps to improve the accuracy and robustness of the model. Firstly, we enhance the baseline model by proposing a new model called occTransformer. Secondly, we use ensemble methods with multiple occupancy models to further improve the model's performance. Finally, we experiment with using a detection model to convert occupancy predictions and improve the model's mIoU for dynamic objects. By taking these steps, we aim to create a more effective and reliable model for 3D occupancy prediction.

2. Methods

3D Occupancy Prediction aims to predict a 3D semantic scene with multi-camera images as input. The occupancy prediction is represented as dense cubic features $V \in \mathbb{R}^{H \times W \times Z \times C}$, where H, W, Z are the spatial resolution of the voxel space, C denotes the semantic label of each voxel.

In this section, we present our solution to the 3D occupancy prediction challenge. We first introduce our data augmentation, followed by a detailed explanation of our models. Finally, we describe the ensemble strategy we used to achieve our results.

2.1. data augmentation

Input and Data Augmentation. To encourage the model to rely on more local features, which can help in training more robust and resilient models, we apply the cutout augmentation [5] on the input multi-camera images $I = \{I^1.I^2, ...I^N\}$. Cutout is a widely used image data augmentation that randomly selects certain regions within an input image and masks them by setting their pixel values to zero (or some fixed value). And we do not use any other augmentations in our method.

2.2. Model

We use BEVFormer [1] as our baseline model, and several significant improvements were proposed during the challenge, which are introduced as follow.

2D to 3D. In this study, we aimed to improve the 3D perception tasks by replacing the baseline's bev generation module with different methods, including LSS [6], FLosp [7], and FLosp Depth [8]. The lifting of 2D to 3D is a crucial aspect of these tasks. However, our experiments, as shown in Table 1, revealed that the 2D-3D module proposed by BEVFormer [1] is the most effective method. It is important to note that all experiments were conducted without camera mask.

Image Backbone. Given a set of surrounding multi-camera images, the image backbone(e.g. ResNet-101 [9]) extracts multi-scale image features. To obtain more detailed visual clues, we use Swin-L [10], InternImage-XL [11], convnextv2-L [12] as image backbone, respectively.

3D Occupancy Head. As shown in Figure 2, after lifting image feature from 2D to 3D, we can obtain the bev queries $F_{bev} \in \mathbb{R}^{H \times W \times C}$, where C is embedding dims of each bev query. Then a two-layer MLP is utilized to decode 3D voxel feature $F_{3D} \in \mathbb{R}^{H \times W \times Z}$ from bev queries F_{bev}. Further more, following MLP, we use 3d UNet to get fine-grained 3D voxel feature. Specifically, 3d Unet module strengthens the spatial representation by fusing multi-scale voxel feature. In practice, we set the downscale ration = $\{2, 4, 8\}$. The output of 3d Unet F_{unet} is used as occupancy predictor input.

Loss. The baseline model uses cross-entropy loss to supervise the occupancy prediction. We find that only using cross-entropy loss will result in ambiguous occupancy boundaries. To tackle this problem, we add dice loss [13] when training. Thus, the overall loss is formulated as:

$$L = \lambda_{ce} L_{ce} + \lambda_{dice} L_{dice}$$

2.3. Ensemble

In this section, we describe our ensemble strategy. We found that weighting the probabilities from different models is more effective than taking the maximum probability or using a voting approach to combine the predictions from different models. Therefore, we used a weighted approach to combine the probabilities from different models.

Pretrain Backbone Ensemble. We first use different image backbone to train several improved version of BEVFormer models.

Occ Model Ensemble. In addition to our own ensemble strategy, we also reproduced other methods for ensembling, including BEVDet [2], SurroundOcc [3], and VoxFormer [14].

DET Ensemble. We found that 3D object detection models sometimes perform better than occ models on dynamic classes, so we decided to incorporate the detection model into our approach. To do this, we used StreamPETR [4] to generate bbox frames and convert them into 3D bbox occ results. The process involved setting a threshold for each class based on the score and selecting high-confidence boxes. Then, we generated point clouds with a spacing of t within each box and checked whether each point was inside the box. After voxelizing the points, we assigned the corresponding semantic label to the voxels inside the box. When multiple semantic labels were predicted for a voxel, we selected the one with the highest score. Finally, we obtained the occ result and performed a prob average ensemble with the previous occ model.

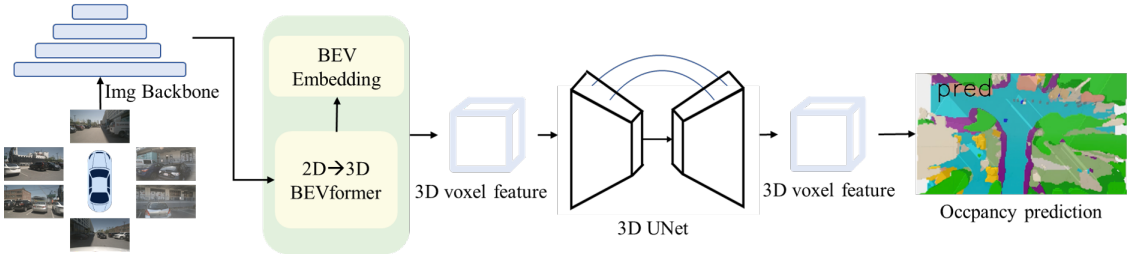

Figure 2: The occTransformer framework involves the use of the bevformer method, which converts 2D images from six different viewpoints into 3D features. These 2D features are first extracted and then aggregated into a bird's eye view (BEV) embedding. A simple decoder is then used to generate 3D voxel features, which are further enriched using a 3D U-Net head. The final output of the framework is a 3D occupancy prediction.

2d to 3d method	miou
BevFormer [1]	23.464
Flosp Depth [8]	23.18
Flosp [7]	22.038
Lss [6]	22.773

Table 1: 2d to 3d method w/o camera mask

id	Method	image backbone	mIOU
a	Bevformer	res101 [9]	40.6
b	Bevformer	swin-L [10]	42.9
c	Bevformer	convnextv2-L [12]	44.0
d	Bevformer	InternImage-XL [11]	43.3
e	Bevdet	swin-B [10]	43.1
f	SurroundOcc	InternImage-B [11]	40.7
g	VoxFormer	res101 [9]	40.7
h	ensemble b+c+d	-	46.28
i	ensemble h+e+f+g	-	48.01
j	ensemble i+det	-	48.91

Table 2: Ensemble results.

cutout	unet3d	dice loss	history frames	mIOU
✓	✓	✓	✓	**38.1035**
-	✓	✓	✓	37.7701
✓	-	✓	✓	37.5975
✓	✓	-	✓	34.5992
✓	✓	✓	-	37.4897

Table 3: Ablation study of our proposed improvements

3. Experiments

3.1. Dataset and Evaluation

The nuScenes dataset [15] is a large-scale dataset specifically designed for autonomous driving research and has been fine-tuned for the 3D occupancy prediction competition. It consists of over 34,000 samples, including 28,130 samples for training, 6,019 samples for validation, and 6,008 samples for testing. The dataset includes data from six cameras and has a voxel size of 0.4m. The range of the dataset is from -40m to 40m in the x and y directions and from -1m to 5.4m in the z. The volume size is [200, 200, 16]. The dataset contains 18 classes, with classes 0 to 16 defined the same as in the nuScenes-lidarseg dataset. The label 17 category represents voxels that are not occupied by anything, which is named as *free*. The ground truth labels of occupancy are derived from accumulative LiDAR scans with human annotations.

During training, both [mask lidar] and [mask camera] masks are optional, and participants are not required to predict the masks. However, during evaluation, only [mask camera] is used, and we use the provided camera mask as default.

3.2. Implementation Details

We use AdamW2 [16] optimizer with cosine annealing policy. We set learning rate max to 2×10^{-4}, with 0.01 weight decay. The model is trained on 8 NVIDIA V100 GPUs with 24 epochs. We use R101-DCN as backbone during the exploration stage, and finally use Swin-L [10], ConvNextv2-L [12], InternImage-XL [11] to build stronger models.

Unless we explicitly indicate, all models used for ablation study, including Table 2 and Table 3, are trained on training data, tested on validation data. In the final submission, we first train 24 epochs on training data and finetune 12 epochs on the whole trainval data.

3.3. Ablation Study

Table 3 shows the ablation results of our improvements on baseline. We find that the dice loss significantly boost the performance of occupancy.

Method	mIou	others	barrier	bicycle	bus	car	construction vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	driveable surface	other flat	sidewalk	terrain	manmade	vegetation
NVOCC	54.19	28.95	57.98	46.40	52.36	63.07	35.68	48.81	42.98	41.75	60.82	49.56	87.29	58.29	65.93	63.30	64.28	53.76
42dot	52.45	27.80	56.28	42.62	50.27	61.01	35.41	47.97	38.90	40.29	56.66	47.03	86.96	57.48	63.64	62.53	63.00	53.74
UniOcc	51.27	26.94	56.17	39.55	49.40	60.42	35.51	44.77	42.96	38.45	59.33	45.90	83.90	53.53	59.45	56.58	63.82	54.98
occ-heiheihei	49.36	28.43	54.49	39.04	45.45	59.15	32.05	43.46	36.33	40.72	51.67	43.73	84.97	57.03	61.38	56.95	57.95	46.26
occTransformer(Ours)	49.23	26.91	53.57	39.53	47.56	59.54	32.59	44.34	37.36	37.28	54.81	44.70	84.61	55.15	60.34	56.35	57.14	45.04

Table 4: Top five submissions of 3d occupancy prediction track

4. Results

In the 3D occupancy prediction track, we improved the BEVFormer model and effectively integrated the results of existing occupancy models. Finally, we utilized a detection model to improve the mIoU for dynamic objects. The top five final results are shown in Table 4

References

[1] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 2022. 1, 2, 3

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

[3] 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

[4] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. *arXiv preprint arXiv:2303.11926*, 2023. 1, 2

[5] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[6] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2, 3

[7] 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

[8] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv:2302.13540*, 2023. 2, 3

[9] 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. 2, 3

[10] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[11] 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. 2, 3

[12] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023. 2, 3

[13] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 2

[14] 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. 2

[15] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *arXiv preprint arXiv:2109.03805*, 2021. 3

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