Addressing the 3D Occupancy Prediction Challenge: Towards Real-world Application in Varying Camera Settings

Yuichi Inoue
Turing Inc.
y.inoue@turing-motors.com

Daiki Shiotsuka
Turing Inc.

Abstract

This report introduces the solution for the 3D Occupancy Prediction track in Autonomous Driving Challenge from Turing Inc. We improved the performance of a robust BEVDet baseline by incorporating an InternImage backbone known for its effectiveness in 3D recognition tasks. To enhance training efficiency, we initially utilized a limited amount of temporal information and increased the length of the temporal sequence. Furthermore, we evaluated the capabilities of our trained model by applying it to our data, obtaining satisfactory results. This report provides insights into our efforts and achievements in the challenge, showcasing the advances made in the vision of autonomous driving computers.

1. Introduction

The 3D occupancy prediction challenge is a new challenge designed for the nuScenes dataset [1], which contains voxel-formatted data with semantic labels. At CVPR 2023, a challenge was held specifically for the prediction of 3D occupancy using only camera input. In the camera-only 3D occupancy prediction track, the challenge requires the algorithm to predict voxel class labels using only camera data.

We incorporated the powerful image encoder, InternImage [6], into the BEVDet architecture [4, 5]. Furthermore, we found that training the model with a short temporal sequence and then fine-tuning it with a longer temporal sequence led to significant mIoU score improvements within a few epochs.

Additionally, we conducted inference on data collected by ourselves using the model trained on the challenge's dataset. Despite utilizing a different number of cameras from the nuScenes dataset, we concatenated and cropped our images to match those of nuScenes. This adjustment allowed us to obtain reasonable results without additional training, demonstrating the effectiveness and versatility of our approach.

Through this report, our goal is to share the methodology, findings, and achievements of our solution in the 3D occupancy prediction challenge. Our report contributes to the advancement of computer vision techniques in the field of autonomous driving and holds potential for real-world applications.

2. Solution

In this section, we present the details of our solution. We introduced the BEVDet as our baseline and made modifications to improve its score. Specifically, we replaced the image encoders and incorporated more temporal information. Finally, we introduced an ensemble method to use the aggregated results for the final evaluation.

2.1. Baseline architectures

BEVDet is a method composed of four components: an image view encoder, a view transformer, a BEV encoder, and a task-specific head. The image view encoder is tasked with transforming input images into high-level features using a backbone for feature extraction and a neck for feature fusion. Common structures such as ResNet and SwinTransformer are used in this process. The view transformer then takes these features and predicts depth, allowing the conversion of the image view to a Bird's Eye View (BEV). This process involves point-cloud rendering and the application of pooling operations to generate the BEV features.

Subsequently, the BEV encoder module refines these features, mirroring the structure of the image view encoder but focusing on detecting crucial aspects such as scale, orientation, and velocity within the BEV space. The task-specific head, built on the BEV feature, plays a pivotal role in 3D occupancy prediction. In our solution, we utilize 3D convolution layers in the task-specific head to predict the class of each voxel.

We have incorporated temporal information into the BEVDet framework. Using the sensor information from

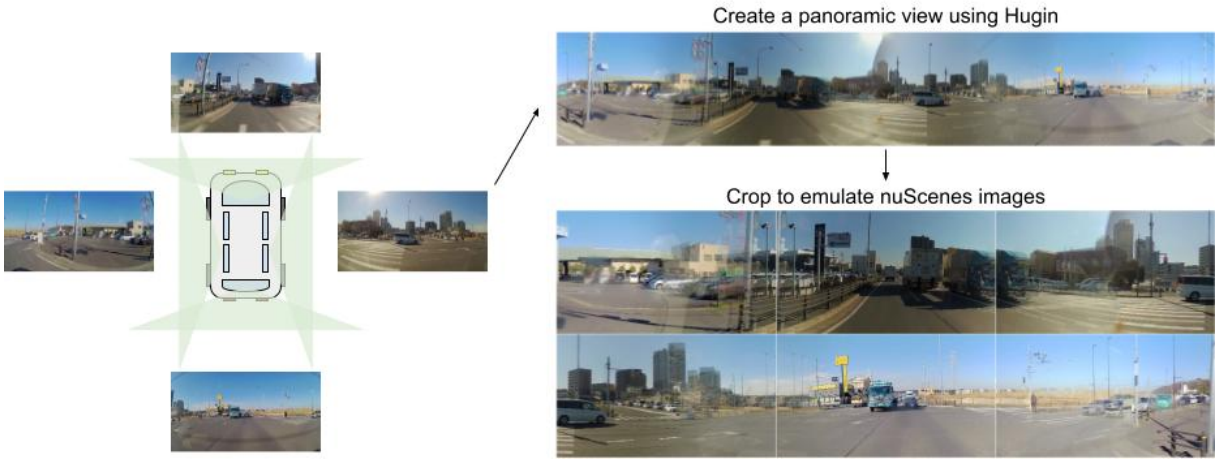

Figure 1. Our data were collected using 4 cameras with a 140° angle-of-view, covering the front, left, right, and rear of the vehicle. The footage was fused into a panoramic view using the Hugin [2], and this view was then cropped to simulate a six-camera setup, aligned with the desired field of view for evaluation purposes.

the ego car, we shifted the BEV features from the previous frame to align with the current BEV features. This method enables us to exploit temporal cues more effectively, enhancing the model's ability to track and predict object motion over time.

2.2. Improving baseline models

InternImage backbone

In our solution, we leverage InternImage, a CNN-based foundation model adept at managing large-scale parameters and data, employing an extended variant of the flexible convolution, Deformable Convolution v3 (DCNv3). The essential component of this model is a dynamic sparse convolution with a standard 3x3 window size. The sampling offsets of this operator are adaptable by dynamically learning suitable receptive fields from the provided data. DCNv3 bridges the gap between regular convolutions and Multi-Head Self-Attention, allowing for long-range dependencies and adaptive spatial aggregation. Furthermore, DCNv3 amends some limitations of regular convolutions, maintaining the inductive bias, thus making the model more efficient with fewer training data and shorter training time. Importantly, DCNv3, due to its sparse sampling, is computationally and memory efficient, and only requires a 3x3 kernel to learn long-range dependencies.

Fine-tuning technique for temporal data

To improve training efficiency, we initially trained our model using temporal data with a sequence length of 2. Subsequently, by fine-tuning the weights obtained from this initial training, we efficiently trained the model with an extended sequence length of 8 for better temporal context learning. This two-step approach optimized the learning process, allowing the model to effectively capture temporal dependencies and achieve improved performance while maintaining computational efficiency.

Ensemble

To enhance the performance of our baseline models, we used two ensemble techniques: the traditional model averaging ensemble and the Snapshot ensemble [3]. Ensemble methods have been widely recognized for their ability to improve model robustness and accuracy by combining multiple models. However, training multiple deep networks for model averaging can be computationally expensive. To address this challenge, we adopted the Snapshot ensemble approach, which involves ensembling different checkpoints of models trained under the same conditions.

3. Experiments

3.1. Dataset and Evaluation

Dataset

The dataset used for the challenge of 3D occupancy prediction is derived from the nuScenes dataset. It consists of images captured from six cameras, and the objective is to predict the occupancy state and semantic labels for each voxel grid in the scene. The occupancy and semantic labels are determined using accumulated LiDAR scans with human annotations. Each voxel grid is classified as either "free" or "occupied", and, in the case of occupied voxels, participants are required to predict their semantic class. Additionally, the dataset provides a binary mask that indicates whether each voxel grid is observed or unobserved in the current camera view. During the evaluation, only the ob-

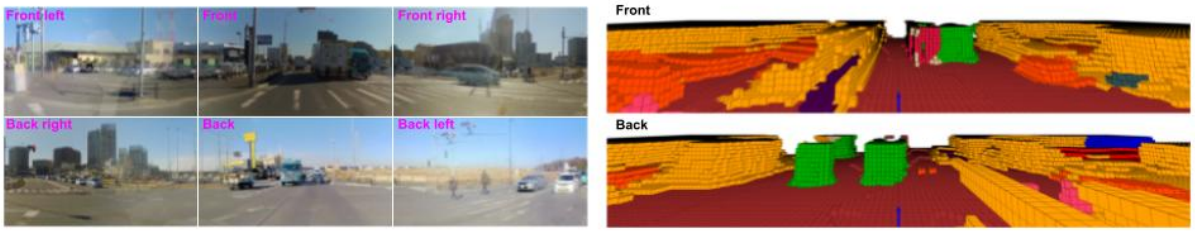

Figure 2. Preliminary evaluation results on our collected data demonstrate the robustness and applicability of our models. Despite variations in camera numbers, panoramic synthesis and appropriate cropping enable the successful utilization of diverse imaging conditions. Further validation is required to consolidate these findings.

served voxels are considered, whereas the unobserved voxels are excluded from the evaluation process.

To evaluate the robustness of our models, we performed inference on uniquely collected data, as shown in Fig. 1. Data were obtained from four cameras, each with a 140° angle of view, positioned at the front, left, right, and rear of the vehicle, thereby capturing a comprehensive 360° range. A Honda NBox served as the vehicle of use, and the footage was recorded using a Garmin drive recorder. To align with the six-camera system of nuScenes, we initially fused the four-direction camera data into a panoramic view. This panoramic view was created using Hugin [2], ensuring a seamless blend at the joining points through weighted merging. Following this, the panoramic view was intentionally cropped to emulate a six-camera setup, thus aligning with the desired field of view.

Evaluation metrics

The evaluation metric used in the 3D occupancy prediction challenge is the mean intersection-over-union (mIoU) calculated over all classes. The mIoU is computed as the average of the IoU scores for each class, where the IoU is the ratio of the intersection to the union between the predicted and ground truth voxels. The mIoU provides a comprehensive measure of the model's accuracy in predicting both the occupancy state and semantic labels of the voxel grids. A higher mIoU score indicates better performance in the challenge.

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{1}$$

where C represents the number of classes, TP_c denotes the number of true positive predictions for class C, FP_c represents the number of false positive predictions, and FN_c corresponds to the number of false negative predictions.

3.2. Implementation Details

In our implementation, we utilized pre-trained weights from the COCO dataset as the initial weights for the InternImage model. AdamW optimizer with an initial learning rate of 2e-4 and weight decay of 0.01 was used. The total batch size was set to 16, using 8 NVIDIA A100 40GB GPUs. To preprocess the images, we resized them from 900x1600 to 512x1408 and applied data augmentation techniques such as resizing, rotation, and flipping. The training process spanned 24 epochs for a sequence length of 2 and 6 epochs for a sequence length of 8. For the Snapshot ensemble, we selected checkpoints with high validation scores. No external data were used in our final results.

ID	Seq	epoch	mIoU
1	2	23	42.9
2	2	24	43.0
3	8	3	45.1
4	8	4	44.1
1+2			43.8
3+4			45.6
1+2+3+4			46.2

Table 1. The results for the validation set are presented. InternImage Base was used as backbone. The term *Seq* represents the number of past information points employed. By employing Seq=2 checkpoints for fine-tuning, higher mIoU scores are achieved with a reduced number of epochs for Seq=8. The Snapshot ensemble technique demonstrates enhanced mIoU performance.

3.3. Results

As shown in Tab. 1, the results of the Snapshot ensemble demonstrate a significant improvement in the mIoU scores compared to those of individual models. Additionally, our findings indicate that training with a shorter sequence length initially and then increasing it leads to achieving high scores with fewer epochs. The final test results were obtained by ensembling 7 models with varying backbone sizes and sequence lengths. The ensemble achieved an mIoU score of **47.84** on the validation set and an mIoU score of **47.36** on the test set, securing the **18th** position in the challenge.

To assess the robustness and applicability of our models, we performed qualitative evaluations on the data we

collected. As shown in Fig. 2, despite the different number of cameras, we achieved satisfactory results through panoramic synthesis and appropriate cropping to match the field of view. This suggests the potential for utilizing such data even under varying imaging conditions. However, it is important to note that these are preliminary results, and further validation is required to solidify these findings.

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, 2020. 1

[2] Pablo d'Angelo. Radio metric alignment and vignetting calibration. 2007. 2, 3

[3] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free, 2017. 2

[4] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection, 2022. 1

[5] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view, 2022. 1

[6] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions, 2023. 1