MiLO: Multi-task Learning with Localization Ambiguity Suppression for Occupancy Prediction
CVPR 2023 Occupancy Challenge Report

Thang Vu Jung-Hee Kim Myeongjin Kim Seokwoo Jung Seong-Gyun Jeong
42dot Inc.
{firstname}.{lastname}@42dot.ai

Abstract

*We present **M**ult**i**-task Learning with **L**ocalization Ambiguity Suppression for **O**ccupancy Prediction (MiLO) as our solution for camera-based 3D Occupancy Prediction Challenge at CVPR 2023. The proposed MiLO is unique in two important aspects: (1) varying-depth multi-task learning to incorporate perspective semantic prediction, depth estimation, and occupancy prediction for more robust representations; and (2) localization ambiguity suppression to adaptively suppress low-confident localization in camera-based system with respect to object class and distance. In addition, our method employs several techniques to boost the performance. Our final model achieves 52.45 points mIoU without using external data and wins 2nd place in CVPR 2023 3D Occupancy Prediction Challenge.*

1. Introduction

3D scene understanding plays an important role in autonomous driving. The 3D Occupancy Prediction Challenge (3DOPC) at CVPR 2023 provides the first 3D Occupancy Benchmark for Scene Perception in Autonomous Driving. The task is to jointly estimate the occupancy state and semantic label of every voxel in the scene from multi-view images. Compared to LiDAR-based systems, camera-based systems offer lower cost and higher resolution with the ability to capture color information. However, camera-based occupancy is generally challenging, requiring robust representation and accurate localization.

Our solution is built upon the BEVDet4D-Occ [5] baseline. Compared to the baseline, the proposed MiLO is unique in two important aspects. First, MiLO incorporates perspective semantic prediction, depth estimation, and occupancy prediction for more robust representations. Each task is performed at a different network module to attain varying-depth gradient flow and thus ease the deep network training. In particular, the task of perspective semantic, depth, and occupancy predictions are performed at the backbone, view transformer, and occupancy head, respectively. Second, MiLO relies on localization ambiguity suppression to refine the occupancy localization results. 3D localization in camera-based systems is generally more challenging compared to that in LiDAR-based systems since the presence of objects can be identified by the LiDAR. We propose a method that suppresses voxels associated with low-confident localization based on network's prediction score. To further boost the performance, we employ several well-known techniques including class-balanced losses, customized architecture, atrous spatial pyramid pooling 3D, high-resolution input, long-term temporal, stronger backbone, test-time augmentation, pseudo labeling, and model ensemble. Our final model achieves 52.45 points mIoU without using external data and wins 2nd place in CVPR 2023 3DOPC.

2. Baseline Method

Our baseline is BEVDet4D-Occ [5], which is an extension of BEVDet4D by replacing the detection head with the occupancy head. BEVDet4D-Occ consists of 3 main components, which are the image encoder, view transformer, and occupancy encoder. Image encoder composes of a pre-trained ResNet [4] and feature pyramid network (FPN) [7] to obtain features from multi-view images. Noted that only the finest features of FPN are used as the input of the next component. A view transformer projects 2D image features to 3D volumetric features guided by estimated depth [6,11]. Volume features from previous frames are spatially aligned with the coordinate of the current timestamps and concatenated to fused temporal features. Concatenated features are fed to the occupancy encoder [4, 7, 11] to directly predict occupancy. The cross-entropy and class-wise binary cross-entropy are used as occupancy and depth losses, respectively. We refer readers to [5] for further details of the baseline.

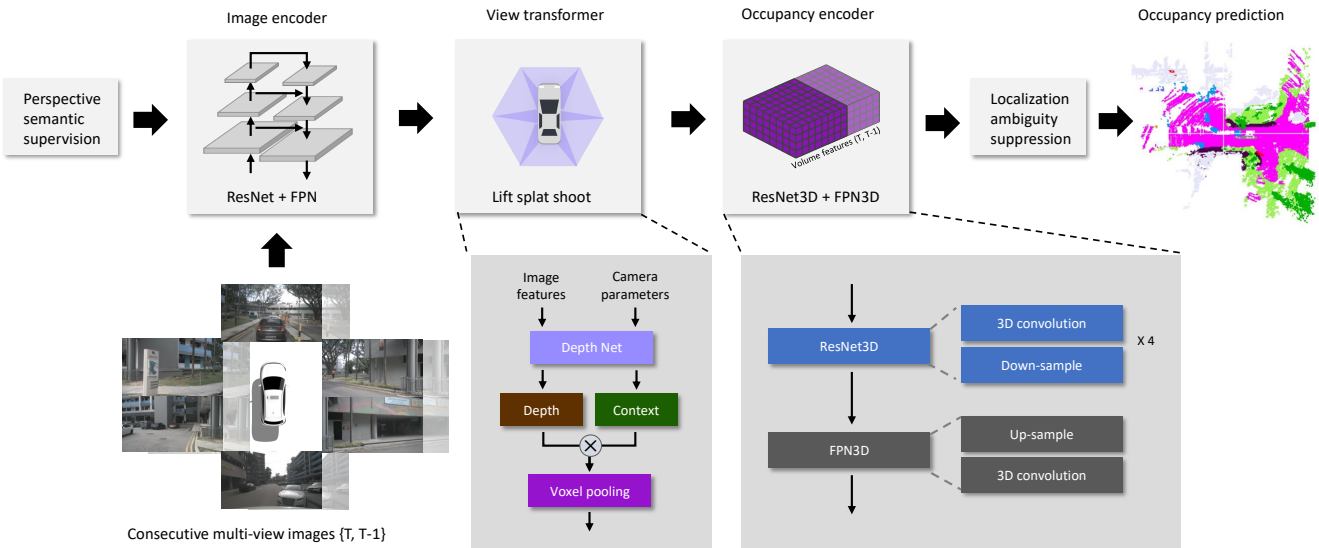

Figure 1. **Overview of MiLO**. Multi-view images are fed into an image encoder to extract 2D image features. Perspective semantic supervision provides the image encoder a short path to the supervision signals. A view transformer [6] transforms 2D features to 3D space guided by estimated depth. To attain temporal information, the volume features from current and previous frames are aligned and concatenated. Then, an occupancy encoder predicts voxel-wise occupancy. Localization ambiguity suppression refines the occupancy predictions to obtain final results.

3. Proposed Method

In this section, we provide a description of our approaches for training and inference, respectively. Figure 1 presents an overview of the proposed method.

3.1. Training

We tailor network architecture of the baseline and employ class-balanced loss function. We leverage additional supervision obtained from sparse LiDAR and pseudo labels to ensure consistent 2D-3D semantic information.

Image Encoder. Using stronger image feature extractors is a simple way to improve prediction accuracy. We employ two modern architectures, which are InternImage-XL [12] and Swin-L [9]. We use publicly available pre-trained weights[1] on ImageNet and COCO datasets.

Occupancy Encoder. We customize the network architecture for higher accuracy with small extra overhead. We increase the number of 3D ResNet3D stages and corresponding FPN3D levels from 3 to 4 and balance the number of base channels.

Atrous Spatial Pyramid Pooling for 3D (ASPP-3D). Occupancy prediction is a dense prediction task, which requires the incorporation of large context. We extend ASPP [1] and propose ASPP-3D to probe upcoming 3D features with filters at multiple sampling rates and fields-of-views.

Class-balanced Losses. The nuScenes dataset is heavily imbalanced. For instance, the portions of common classes (*e.g.*, `driveable_space` and `free`) are roughly $10^4\times$ larger than those of rare classes (*e.g.*, `bicycle` and `motorcycle`). To mitigate this problem, we utilize weighted cross-entropy, and dice loss to supervise the occupancy prediction. Following class balanced loss [2], the weights of imbalanced classes are obtained using the number of voxels for each class as the number of samples with the $\beta = 0.9$. To alleviate computation issues due to a large number of voxel samples, we normalize the number of samples across different classes. Finally, the final multi-task loss is the combination of weighted cross-entropy loss L_{wce}, dice loss L_{dice}, perspective semantic loss L_{sem}, and depth loss L_{depth} with $\lambda_{\{\dots\}}$ being the loss weights:

$$L_{\text{total}} = \lambda_{\text{wce}} L_{\text{wce}} + \lambda_{\text{dice}} L_{\text{dice}} + \lambda_{\text{sem}} L_{\text{sem}} + \lambda_{\text{depth}} L_{\text{depth}} \quad (1)$$

Perspective Semantic Supervision. Since the network consists of complex components in different domains (2D, 2D-to-3D, 3D) leading to potential optimization difficulties. To cope with this problem, we propose varying-depth multi-task learning where each task is performed at a different depth level of the network. In particular, perspective semantic supervision is employed to additionally supervise

[1] https://github.com/microsoft/Swin-Transformer and https://github.com/OpenGVLab/InternImage

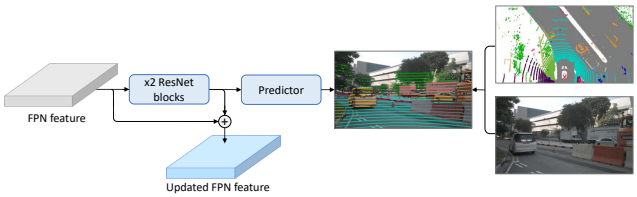

Figure 2. Perspective Semantic Supervision. The finest FPN features are fed into two ResNet blocks followed by a predictor to obtain 2D semantic maps. Semantic features right before the predictor is used to refine the FPN features.

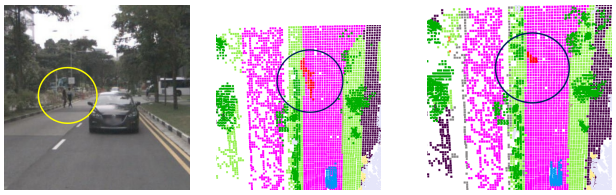

Figure 3. Localization ambiguity in `pedestrian`.

the backbone such that perspective semantic, depth, and occupancy predictions are performed at the backbone, view transformer, and occupancy head respectively. The advantages of this approach are two-fold: (1) it creates a short path to supervision signal for each component, which is shown to be effective in prior work [8, 13, 14], and (2) it provides ensuing components prior for more accurate prediction.

Regarding perspective semantic supervision, since image-level semantic annotations are not available for nuScenes dataset, we project LiDAR points to multi-view images and employ LiDAR semantic annotations to construct corresponding semantic images, as shown in Figure 2. Since LiDAR points are sparse, we create a mask to provide supervision signals to the image locations that the point projection hits only.

High-Resolution Input. The baseline is trained with image resolution of 704×256, we increase the resolution to 1600×640 in our final model.

Long-Term Temporal. The baseline method uses 2 frames to extract temporal information. We increase the number of frames to 5 for more long-term temporal.

Pseudo Labeling. We utilize the inference results as pseudo labels showing confidence score higher than a threshold of 0.9. To avoid over-fitting on pseudo data, we also learn a pseudo camera mask on test data and filter out unobserved locations via pseudo camera mask. Then, we

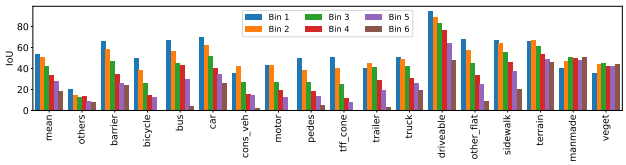

Figure 4. Results on validation set with different object classes and distances. Higher bin indices indicate more distant objects.

fine-tune the model with the combination of the pseudo and training data for 1 epoch.

3.2. Inference

We employ ensembles of two models that rely on different image feature extractors. In this work, we propose techniques for test time augmentation and suppression of localization ambiguities.

Ensemble. To capture different modalities of image features, we employ an ensemble of two backbone architectures which are InternImage-XL and Swin-L.

Test-Time Augmentation. We propose test-time augmentation on images and BEV features including image vertical flipping, image multi-scale of $[0.96, 1.0, 1.06]$, and BEV flipping along the x and y axes. Due to resource constraints, we select randomly 6 augmentation configurations for each testing sample.

Localization Ambiguity Suppression. Camera-based approach for 3D scene understanding is highly challenging. In LiDAR-based systems, one can physically determine the presence of surrounding objects from particles detected. In contrast, methods for camera-based systems require 3D object localization in addition to object recognition. However, 3D localization from images is often ambiguous. Figure 3 shows that the localization quality of the prediction for `pedestrian` is unsatisfactory in comparison with the ground truth. We further analyze the prediction accuracy with respect to different object classes and distances. Intuitively, distant and small objects are difficult to precisely localize. We divide the scenes into multiple spatial bins (6 in our experiments) in BEV polar coordinates based on the distances from points to the ego car, and then evaluate per-class IoU on each bin in Figure 4. As expected, the IoU drops quickly as the distance increases in most of the classes. For rare small classes such as `bicycle` and `motorcycle`, the mIoU is almost 0 for distant objects (bin 6). To cope with the localization ambiguity problem, we propose to adaptively suppress voxels with low-confident localization by assigning them to the background (*i.e.*, `free` class). For simplicity, we assume that the localization confidence correlates with the network prediction score, which is widely

Table 1. Comparison between the competition baseline and our method on the test set.

Method	mIoU	others	barrier	bicycle	bus	car	cons_veh	motor	pedes	tff_cone	trailer	truck	driveable	other_flat	sidewalk	terrain	manmade	veget
Competition BL	23.70	10.24	36.77	11.70	29.87	38.92	10.29	22.05	16.21	14.69	27.44	23.13	48.19	33.10	29.80	17.64	19.01	13.75
MiLO (ours)	**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**
Improvement	**28.75**	**17.56**	**19.51**	**30.92**	**20.40**	**22.09**	**25.12**	**25.92**	**22.69**	**25.60**	**29.22**	**23.90**	**38.77**	**24.38**	**33.84**	**44.89**	**43.99**	**39.99**

Table 2. Ablation studies on validation set using ResNet-50.

exp_id	Baseline	Semantic	Loss	Arch	ASPP	High-Res	Longterm	Pseudo	TTA	Localization	mIoU
1	✓										33.50
2	✓	✓									34.92
3	✓	✓	✓								39.53
4	✓	✓	✓	✓							40.49
5	✓	✓	✓	✓	✓						41.13
6	✓	✓	✓	✓	✓	✓					42.71
7	✓	✓	✓	✓	✓	✓	✓				43.95
8	✓	✓	✓	✓	✓	✓	✓	✓			45.05
9	✓	✓	✓	✓	✓	✓	✓	✓	✓		46.02
10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	46.76

used in other tasks such as object detection. During inference, we search for the optimal per-class per-bin confidence score threshold on the validation set. If the confidence score of a voxel is lower than the the threshold of corresponding class and spatial bin, the voxel is re-assigned to `free` class.

4. Experiments

In this section, we explain dataset, implementation details, and ablation studies. Table 1 shows the final results.

Dataset and Evaluation Metric. The nuScenes dataset consists of 700 scenes for training, 150 for validation, and 150 for testing. During training, two data modalities are utilized in our method which are multi-view images and LiDAR points. Specifically, LiDAR coordinates and semantic labels are used as extra supervision signals. During inference, only multi-view images serve as the input. No external data is utilized in our method. The standard mean Intersection over Union (mIoU) is employed for evaluation.

Implementation Details. The final model utilizes InternImage-XL [12] and Swin-L [9] as the backbones. Multi-level features from the backbone are fused via LSS-FPN [11]. Multi-task loss-weights are set to $\lambda_{\text{wce}} = 1$, $\lambda_{\text{dice}} = 0.3$, $\lambda_{\text{sem}} = 0.1$, and $\lambda_{\text{depth}} = 0.05$. The number of contiguous frames to compute temporal information is set to 5. The multi-view images are cropped into the size of 1600×640 pixels. The images are augmented by vertical flipping, random scaling in the range of [0.94, 1.11], and random rotation in the range of [-5.4°, 5.4°]. In addition, BEV features are also augmented via flipping.

The network is trained with a batch size of 32 on 16 GPUs (two samples per GPU) for 36 epochs using AdamW [10] optimizer. Learning rate and weight decay are set to $1e^{-4}$ and $1e^{-2}$, respectively. We adopt Exponential Moving Average (EMA) [3] for updating model weights.

Ablation Analysis. To quickly validate the efficacy of the proposed components, we perform ablation analysis using ResNet-50 [4] backbone with 24 training epochs. Table 2 reports the ablation results, where proposed components are gradually integrated. The baseline is BEVDet4D with ResNet-50 backbone. Components provide additive performance improvements. Overall the proposed components attain 13.26 points mIoU improvement over the baseline.

5. Conclusions

We proposed the MiLO that is ranked in second row in 3D Occupancy Prediction Challenge at CVPR2023, and reported its recipe for training and inference. Specifically, we introduced multi-task learning framework for classifying and localizing semantic information on voxels. We also proposed a post-processing method suppressing low-confident localization based on network's prediction score. Our final model achieves 52.45 points mIoU, which is 28.75 points higher than the competition baseline.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017. 2

[2] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2

[3] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv:2107.08430*, 2021. 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 4

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

[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. *arXiv*, 2022. 1, 2

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[8] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 3

[9] 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 *ICCV*, 2021. 2, 4

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv*, 2017. 4

[11] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020. 1, 4

[12] 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. In *CVPR*, 2023. 2, 4

[13] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *CVPR*, 2023. 3

[14] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 3