2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

MonoOcc: Digging into Monocular Semantic Occupancy Prediction

Yupeng Zheng[1,2*], Xiang Li[3*], Pengfei Li[3], Yuhang Zheng[3],
Bu Jin[1,2], Chengliang Zhong[3], Xiaoxiao Long[4†], Hao Zhao[3], and Qichao Zhang[1,2†✉]

***Abstract*— Monocular Semantic Occupancy Prediction aims to infer the complete 3D geometry and semantic information of scenes from only 2D images. It has garnered significant attention, particularly due to its potential to enhance the 3D perception of autonomous vehicles. However, existing methods rely on a complex cascaded framework with relatively limited information to restore 3D scenes, including a dependency on supervision solely on the whole network's output, single-frame input, and the utilization of a small backbone. These challenges, in turn, hinder the optimization of the framework and yield inferior prediction results, particularly concerning smaller and long-tailed objects. To address these issues, we propose MonoOcc. In particular, we (i) improve the monocular occupancy prediction framework by proposing an auxiliary semantic loss as supervision to the shallow layers of the framework and an image-conditioned cross-attention module to refine voxel features with visual clues, and (ii) employ a distillation module that transfers temporal information and richer knowledge from a larger image backbone to the monocular semantic occupancy prediction framework with low cost of hardware. With these advantages, our method yields state-of-the-art performance on the camera-based SemanticKITTI Scene Completion benchmark. Codes and models can be accessed at https://github.com/ucaszyp/MonoOcc.**

I. Introduction

3D scene understanding serves as a foundation for autonomous driving systems, exerting a direct influence on downstream tasks such as planning, navigation, VR [1], [2], map construction [3], [4]. The past years have witnessed the rapid development and significant impact of lidar-based algorithms [5]–[9] in outdoor 3D scene understanding. Nevertheless, they are often considered expensive in terms of hardware for autonomous vehicles. Consequently, monocular scene understanding [10]–[15] has garnered considerable attention from the robotics community due to its cost-efficiency and rich visual information. A popular topic in this domain is Semantic Occupancy Prediction, also denoted as Semantic Scene Completion (SSC) [16]. Its objective is to predict the semantic occupancy of each voxel throughout the entire scene, encompassing both visible and occluded regions, while relying solely on monocular observations.

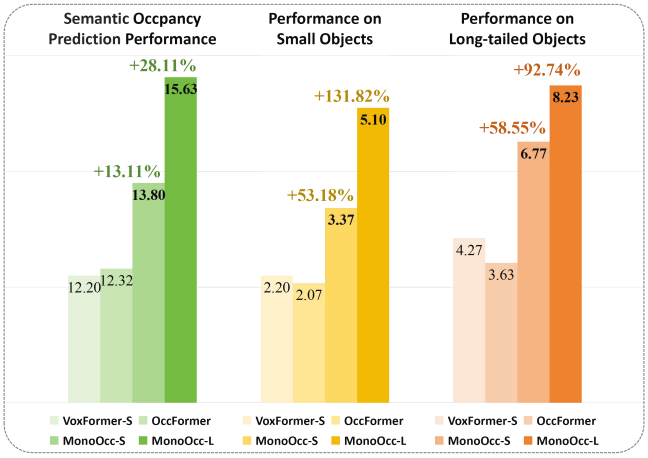

Fig. 1. Quantitative results of semantic occupancy prediction on SemanticKITTI [16] test set compared with the state-of-the-art VoxFormer-S [17] and OccFormer [18]. Note that our method outperforms the latter methods in the SSC mIoU, while also achieving a significant boost on both small objects (*bicycle*, *motorcycle*, *traffic-sign*) and long-tailed objects (*truck*, *other-vehicle*, *other-ground*). Compared with VoxFormer-S, the relative percentage increase of our method on average performance, small objects and long-tailed objects are denoted by **green**, **yellow** and **orange**, respectively.

The existing cutting-edge method, VoxFormer [17], proposes a sparse-to-dense architecture, which aggregates 2D features to voxel space with depth-based queries and completes the entire scene conditioned on the feature of queries. It should be emphasized that the inaccuracies of depth estimation affect the accuracy of the query, leading to an increase in the difficulty of subsequent completion and semantic parsing. Besides, all of the existing methods such as [10], [17], [19] rely solely on the supervision of 3D ground truth to train the deep cascaded architecture, including the 2D image backbone, 2D-3D view transformer, and 3D completion network, bringing challenges to the optimization of heterogeneous sub-modules. Additionally, these methods only utilize visual information from a single frame input, resulting in poor performance, particularly causing failures in predictions for small objects and long-tailed objects (see Fig.1).

To this end, we first propose an image-conditioned cross-attention module, aiming to refine inaccurate voxel features brought by depth estimation with the extra information from image features and introduce an auxiliary semantic loss as supervision to the shallow layers of the small framework, facilitating more efficient optimization. Secondly, we employ a large pre-trained image backbone instead of small models trained on benchmark datasets as the former arts

[1]The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, {zhangqichao2014, zhengyupeng2022}@ia.ac.cn
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China,
[3]Institute for AI Industry Research (AIR), Tsinghua University, China,
[4]Department of Computer Science, the University of Hong Kong.
∗ Equal contribution.
† Project leader.
✉ Corresponding to zhangqichao2014@ia.ac.cn

used to assist in SSC. Recent research (e.g., [20], [21]) has demonstrated that large image backbones can significantly enhance the adaptability and generality of 2D image semantic segmentation. However, how to efficiently utilize these models in the SSC task has not been explored. Considering the efficiency and resource constraints in real-world applications, we propose to use model distillation to get more compact yet efficient models that can approximate the behavior of larger models. As shown in Fig. 2, we denote the larger model as a privileged branch, inspired by the idea of privileged learning which is widely recognized in the robotics community [22], [23]. This branch is designed to take temporal image frames as inputs, thus mitigating uncertainty in occluded areas. As illustrated in Fig. 1, the comparison between SOTAs' SSC results and ours on SemanticKITTI [16] test set demonstrates that our method achieves a significant gain on general, small and long-tailed objects.

For easy reference, we summarize our contributions below.

- We propose an image-conditioned cross-attention module and semantic auxiliary loss to improve the performance of Monocular SSC.
- We propose a privileged branch with pre-training a larger backbone and employing a cross-view transformer to acquire more visual cues from temporal frames.
- We propose a distillation module to transfer knowledge from the privileged branch to the monocular branch.
- We achieve SOTA performance on SemanticKITTI benchmark [16] and release our codes and models.

II. Related Works

Camera-based 3D Perception. In recent years, there has been a growing interest in camera-based 3D sensing techniques [24]–[27], primarily driven by the advantages of richer visual information, ease of deployment, and cost-effectiveness offered by cameras. Recent research in camera-based 3D perception focuses on constructing BEV feature representations and subsequently performing various downstream tasks in the BEV space. The Lift-Splat-Shoot (LSS) method [28], along with its subsequent advancements [26], [29], serves as the archetypal technique for forward projection. LSS projects image features into 3D space and aggregates them into the BEV space, incorporating depth uncertainty through predicted pixel-wise depth distributions. BEVFormer [24] represents one of the backward projection methods, utilizing deformable attention-based spatiotemporal transformers to construct BEV queries and aggregate corresponding 2D features from multiple frames into the BEV space. Given that 3D occupancy representation contains richer spatial information compared to BEV representation, it plays a crucial role in the perception and planning abilities of self-driving cars. Consequently, there is a noticeable shift towards employing camera-based solutions in 3D Semantic Occupancy Prediction.

Camera-based 3D Semantic Occupancy Prediction. After the introduction of SemanticKITTI dataset [16], an abundance of outdoor Single-View 3D Semantic Occupancy Prediction (synonymous with Semantic Scene Completion) techniques have emerged. MonoScene [10] is the pioneering method for monocular semantic occupancy prediction, which proposes 2D-3D feature projections along the line of sight to generate voxel features and utilizes the 3D UNet to process the volumetric data. TPVFormer [30] introduces a simple yet efficient tri-perspective view representation as an alternative to the BEV representation, enabling the capture of 3D structural intricacies. OccFormer [18] devises dual-path transformer blocks comprising local and global transformers to decompose the 3D processing. VoxFormer [17] replaces BEV queries with depth-based proposal queries to aggregate features from images and introduces an MAE-like design to achieve dense occupancy completion with sparse queries.

III. Method

An overall framework of MonoOcc is illustrated in Fig. 2

We briefly describe the sparse-to-dense monocular 3D semantic occupancy prediction pipeline of a baseline method in section III-A. Two innovations, including an image-conditioned cross attention and a 2D semantic auxiliary loss, are proposed to improve the current framework in section III-B. To promote the performance of small objects and long-tailed objects, we further propose a privileged branch by pre-training a larger image backbone and introducing a cross view transformer to enhance temporal view features in section III-C. Finally, we propose a distillation module to transfer the knowledge from the privileged branch to the monocular branch, making a trade-off between performance and efficiency in section III-D.

A. *Sparse-to-dense Monocular 3D Semantic Occupancy Prediction*

Image Feature Extractor. To extract 2D feature maps $F_t^{2D} \in \mathbb{R}^{d\times h\times w}$ from corresponding RGB images I_t, an image feature extractor Φ_f is constructed by employing ResNet-50 [31] as backbone and FPN [32] as neck, where d and (h, w) represent the dimension and resolution of the image feature, respectively. Later we will leverage a stronger image feature extractor pre-trained on a bunch of diverse autonomous driving datasets.

Depth-based Query. Following VoxFormer [17], we generate a total of N_d queries Q_d based on the depth map predicted by a pre-trained depth network. Specifically, we utilize pixel-wise depth to unproject pixels into 3D space, and then obtain initial occupancy by voxelizing these points. Afterward, we acquire a tractable number of basically reasonable initial queries Q_d by correcting initial occupancy with an occupancy prediction network (LMSCNet [33]).

Voxel Feature Generator. Following VoxFormer [17], the process of generating voxel features $\hat{F}_S^{3D} \in \mathbb{R}^{x\times y\times z\times d}$ with resolution (x, y, z) can be divided into two steps:

1) We acquire $O_S^{3D} \in \mathbb{R}^{N_d\times d}$, the feature of visible regions, by utilizing Q_d to aggregate 2D feature F_t^{2D} into 3D space with deformable cross-attention mechanism (DCA) [34]:

$$O_S^{3D} = \mathrm{DCA}\left(Q_d, F_t^{2D}\right). \tag{1}$$

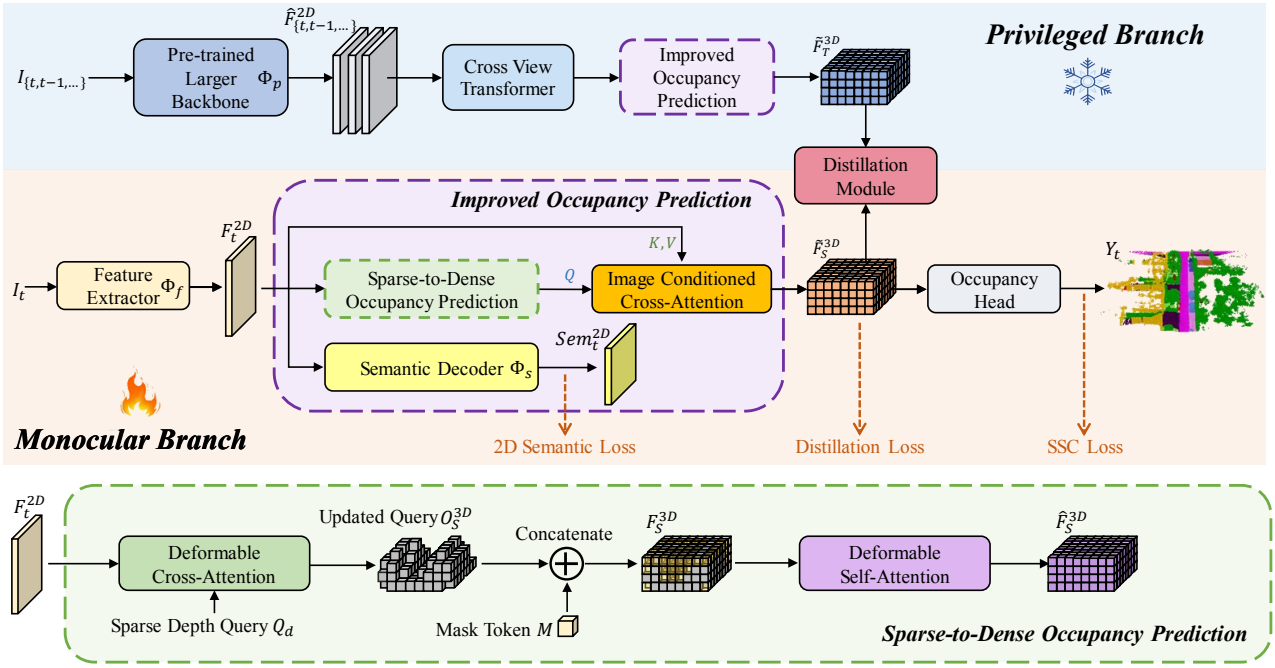

Fig. 2. The architecture of our proposed framework (see section III for details)

For a set of temporal view features $F^{2D}_{\{t,t-1,...\}} \in \mathbb{R}^{N\times d\times h\times w}$ with a valid quantity of N, O^{3D}_T is obtained by an average of aggregated feature:

$$O^{3D}_T = \frac{1}{N}\mathrm{DCA}\left(Q_d, F^{2D}_{\{t,t-1,...\}}\right). \tag{2}$$

2) We acquire initial voxel features $F^{3D}_S \in \mathbb{R}^{x\times y\times z\times d}$ of the whole scene by filling the occluded regions with mask token $M \in \mathbb{R}^d$ and then update F^{3D}_S to $\hat{F}^{3D}_S$ with deformable self-attention mechanism (DSA) [34]:

$$\hat{F}^{3D}_S = \mathrm{DSA}\left(F^{3D}_S, F^{3D}_S\right). \tag{3}$$

Semantic Voxel Map. The predicted semantic voxel map $Y_t \in \mathbb{R}^{X\times Y\times Z\times C}$ is obtained by up-sampling and linear projection of $\hat{F}^{3D}_S$, where (X, Y, Z) denotes the resolution of the 3D volume, C represents the number of classes, including non-occupied.

B. Improved Architecture for Monocular Semantic Occupancy Prediction

Image-Conditioned Cross-Attention. VoxFormer [17] treats semantic occupancy prediction as a generative task. The MAE-like transformer hallucinates the occluded scene conditioned on the visible scene. Mathematically, given the features of the visible region O^{3D}_S and the initial voxel feature F^{3D}_S, the refined features of the entire scene $\hat{F}^{3D}_S$ can be acquired by:

$$\hat{F}^{3D}_S = \mathrm{Complete}\left(F^{3D}_S | O^{3D}_S\right), \tag{4}$$

where $\mathrm{Complete}(\cdot|\cdot)$ means complete the former conditioned on the latter. Reviewing the generation of O^{3D}_S, the inaccuracy of depth estimation introduces inaccurate geometric information, bringing uncertainty to the completion of the entire scene. We believe that the features from the input image can help correct the inaccuracies as they can provide semantic clues. Thus, we complete the scene conditioned on O^{3D}_S and F^{2D}_t:

$$\tilde{F}^{3D}_S = \mathrm{Complete}\left(F^{3D}_S | O^{3D}_S, F^{2D}_t\right). \tag{5}$$

This can be naturally achieved through the deformable cross-attention mechanism. Specifically, we treat the $\hat{F}^{3D}_S$ as the query, F^{2D}_t as the key and value, and leverage deformable cross attention to obtain the corrected image-conditioned voxel features from the refined feature:

$$\tilde{F}^{3D}_S = \mathrm{DCA}\left(\hat{F}^{3D}_S, F^{2D}_t\right). \tag{6}$$

2D Semantic Auxiliary Loss. The occupancy prediction network is a long cascaded framework with components of different domains including a 2D feature extractor, 2D-3D cross-attention, 3D completion self-attention, and occupancy head, which increases difficulties of optimization. To address this problem, we propose a 2D auxiliary semantic loss as deep supervision to the feature extractor. It provides a shorter path for backpropagation, enabling better optimization of the feature extractor, which serves as the source of features for the entire framework.

Regarding the implementation of 2D semantic loss, we first employ a semantic decoder Φ_s composed of convolutional layers and a fully connected layer to predict semantic map Sem^{2D}_t from image feature F^{2D}_t:

$$Sem^{2D}_t = \Phi_s\left(F^{2D}_t\right). \tag{7}$$

Then we project point clouds with semantic labels to corresponding images to generate sparse ground truth. Finally, we employ cross-entropy loss $\mathcal{L}_{sem}$ between ground truth and Sem^{2D}_t to directly optimize the feature extractor.

C. Privileged Branch

Pre-training Scaled-Up Feature Extractor. Increasing the size of the model is an effective strategy to improve

the accuracy of dense prediction tasks. However, due to the limited training samples in SemanticKITTI [16] which only contains about 12K images, using a larger image backbone would lead to overfitting. Moreover, SemanticKITTI not only lacks dense semantic labels but also contains multiple long-tailed classes such as *other-vehicle* (0.20%), *truck* (0.16%), and *other-ground* (0.56%) which only have very limited label points as supervision, making it ineffective to train a larger backbone.

To cope with these two problems, we pre-train the larger image backbone with more data of the driving scenario. And we employ the InternImage-XL [20] loading the pre-trained model as the visual backbone Φ_p for the privilege branch to extract 2D features from multiple frames of images:

$$\hat{F}^{2D}_{\{t,t-1,...\}} = \Phi_p\left(I^{2D}_{\{t,t-1,...\}}\right), \tag{8}$$

where t represents the t-th frame image arranged in chronological order. More details about pre-training are elaborated in section IV-B.

Cross View Transformer. Previous work [17], [35], [36] has demonstrated that temporal information boosts downstream 3D scene perception. When aggregating 2D features into 3D with deformable cross-attention, only the centroid of voxels are projected to the image as reference points. Limited by the voxel resolution, the deviation between the real position of objects in 3D space and the voxel centroid can affect the occupancy prediction of a single frame, which can be alleviated by involving more viewpoints. To acquire as much visual information as possible, we adopt multiple frames as the inputs of the privileged branch.

To further enhance the features of temporal views, we introduce the Cross View Transformer (CVT) [37] to integrate knowledge across multi-views, which is proved to be effective on dense prediction tasks such as depth estimation [37], optical flow estimation [38], [39], and map-view semantic segmentation [40]. In particular, we first add positional encoding to the independently extracted temporal features. Then, we input pairs of adjacent features in chronological order into the CVT for feature enhancement. The enhanced features are lifted to the voxel space through the deformable cross-attention mechanism (DCA).

D. Distillation Module

In the previous subsection, we adopt a temporal view as input and scale up the image backbone to acquire visual clues as rich as possible. However, the usage of multiple frames as input and the scaling up of the 2D backbone significantly increase computational costs, affecting deployment in autonomous driving systems. To address this, inspired by privilege learning [23], we propose a distillation module, composed of a privileged teacher branch and a monocular student branch. The module aims to transfer the knowledge from the privileged branch, which has richer clues of temporal information and a larger backbone prior, to the monocular branch, resulting in performance improvement for the monocular branch. Specifically, inspired by Frustums Proportion Loss from MonoScene [10], we use Kullback-Leibler Divergence as the loss function to provide the cues from teacher to student:

$$\mathcal{L}_{distill} = \mathrm{KL}(\tilde{F}^{3D}_{T} || \tilde{F}^{3D}_{S}). \tag{9}$$

E. Training Loss

Reviewing the training process, we employ multiple loss functions to supervise varying depths of the network.

- For the feature extractor, we introduce loss $\mathcal{L}_{sem}$.
- For the completion network, we propose temporal distillation loss $\mathcal{L}_{distill}$.
- For the final output semantic grid map, we utilize Loss functions $\mathcal{L}_{ssc}$, $\mathcal{L}^{sem}_{scal}$, and $\mathcal{L}^{geo}_{scal}$ from MonoScene [10].

The total loss function can be represented as follows:

$$\begin{aligned}\mathcal{L} = \lambda_1\mathcal{L}_{sem} + \lambda_2\mathcal{L}_{distill} + \lambda_3\mathcal{L}_{ssc} \\ + \lambda_4\mathcal{L}^{sem}_{scal} + \lambda_5\mathcal{L}^{geo}_{scal},\end{aligned} \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, and λ_5 are hyper-parameters.

IV. Experiments

A. Dataset

We evaluate our method on the SemanticKITTI dataset [16], which provides dense semantic occupancy annotations of all lidar scans from the KITTI Odometry Benchmark [41]. Each lidar scan of SemanticKITTI covers a range of $[0 \sim 51.2\mathrm{m}, -25.6 \sim 25.6\mathrm{m}, -2 \sim 4.4\mathrm{m}]$ ahead of the ego car. The ground-truth semantic occupancy is represented as $256 \times 256 \times 32$ 3D voxel grids through voxelizing aggregated lidar scans with 0.2m resolution. Each voxel is annotated with 20 classes (19 semantic classes and 1 free). The official split for train, validation, and test sets is employed. We report our main result (Table I) on the test set and do ablation studies (Table II, III) on the validation set.

Evaluation metrics. Following common practices, we report the mean intersection over union (mIoU) of 19 semantic classes for the Semantic Occupancy Prediction task.

B. Implementation Details

We provide two versions of MonoOcc, namely MonoOcc-S and MonoOcc-L. As shown in Fig. 2, MonoOcc-S employs ResNet50 as the image backbone of the monocular branch, while MonoOcc-L replaces the ResNet50 with our pre-trained larger backbone. For training, we set the hyper-parameters as follows: $\lambda_1 = 4.0$, $\lambda_2 = 3.0$, $\lambda_3 = 2.0$, $\lambda_4 = 1.0$ and $\lambda_5 = 0.5$. We train MonoOcc-S on 4 GeForce 3090 GPUs and MonoOcc-L on 4 A100 GPUs for 20 epochs.

To pre-train the larger backbone, we choose the InternImage-XL [20](350M parameters) as our backbone, and we process approximately 200K training data from open-source autonomous driving datasets including Mapillary Vistas [45], KITTI-360 [46], BDD100K [47], Cityscapes [48], and nuImages [49]. Based on the open-source pre-trained model of InternImage-XL, we first train on the Mapillary Vistas dataset, which includes 124 semantic categories with detailed annotations of road elements and long-tailed objects, effectively enhancing the model's understanding of the

TABLE I

SEMANTIC SCENE COMPLETION RESULTS ON SEMANTICKITTI [16] TEST SET.

Method	Pub	Input	Semantic Occupancy Prediction																				mIoU
			road (15.30%)	sidewalk (11.13%)	parking (1.12%)	other-ground (0.56%)	building (14.1%)	car (3.92%)	truck (0.16%)	bicycle (0.03%)	motorcycle (0.03%)	other-vehicle (0.20%)	vegetation (39.3%)	trunk (0.51%)	terrain (9.17%)	person (0.07%)	bicyclist (0.07%)	motorcyclist (0.05%)	fence (3.90%)	pole (0.29%)	traffic-sign (0.08%)	
LMSCNet* [33]	3DV 2020	Camera	46.70	19.50	13.50	3.10	10.30	14.30	0.30	0.00	0.00	0.00	10.80	0.00	10.40	0.00	0.00	0.00	5.40	0.00	0.00	7.07
3DSketch* [42]	CVPR 2020	Camera	37.70	19.80	0.00	0.00	12.10	17.10	0.00	0.00	0.00	0.00	12.10	0.00	16.10	0.00	0.00	0.00	3.40	0.00	0.00	6.23
AICNet* [43]	CVPR 2020	Camera	39.30	18.30	19.80	1.60	9.60	15.30	0.70	0.00	0.00	0.00	9.60	1.90	13.50	0.00	0.00	0.00	5.00	0.10	0.00	7.09
JS3C-Net* [44]	AAAI 2021	Camera	47.30	21.70	19.90	2.80	12.70	20.10	0.80	0.00	0.00	4.10	14.20	3.10	12.40	0.00	0.20	0.20	8.70	1.90	0.30	8.97
MonoScene [10]	CVPR 2022	Camera	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10	11.08
TPVFormer [30]	CVPR 2023	Camera	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50	11.26
VoxFormer-S [17]	CVPR 2023	Camera	53.90	25.30	21.10	5.60	19.80	20.80	3.50	1.00	0.70	3.70	22.40	7.50	21.30	1.40	2.60	0.00	11.10	5.10	4.90	12.20
VoxFormer-T† [17]	CVPR 2023	Camera	54.10	26.90	25.10	7.30	**23.50**	21.70	3.60	1.90	1.60	4.10	24.40	8.10	24.20	1.60	1.10	0.00	13.10	6.60	5.70	13.41
OccFormer [18]	ICCV 2023	Camera	55.90	30.30	**31.50**	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70	12.32
SurroundOcc [19]	ICCV 2023	Camera	56.90	28.30	30.20	6.80	15.20	20.60	1.40	1.60	1.20	4.40	14.90	3.40	19.30	1.40	2.00	0.10	11.30	3.90	2.40	11.86
MonoOcc-S(Ours)	—	Camera	55.20	27.80	25.10	9.70	21.40	23.20	5.20	2.20	1.50	5.40	24.00	8.70	23.00	1.70	2.00	0.20	13.40	5.80	6.40	13.80
MonoOcc-L(Ours)	—	Camera	**59.10**	**30.90**	27.10	**9.80**	22.90	**23.90**	**7.20**	**4.50**	**2.40**	**7.70**	**25.00**	**9.80**	**26.10**	**2.80**	**4.70**	**0.60**	**16.90**	**7.30**	**8.40**	**15.63**

* represents the results adapted for RGB inputs and reported in MonoScene [10].
† represents the result with temporal inputs.
The best and second-best performances are represented by **bold** and underline respectively.

TABLE II

ABLATION STUDY ON THE EFFECTIVENESS OF EACH PROPOSED COMPONENT ON IMPROVED MONOCULAR BRANCH (VALIDATION SET).

row	2D Semantic Auxiliary Loss	Distillation Module	Image Conditioned Cross Attn.	Train MEM	mIoU↑
1	×	×	×	16G	12.35
2	✓	×	×	16G	13.08
3	×	✓(L)	×	16G	12.88
4	×	×	✓	18G	12.70
5	✓	✓(L)	×	16G	13.26
6	✓	×	✓	18G	13.14
7	×	✓(L)	✓	18G	13.35
8	✓	✓(S)	✓	18G	13.80
9	✓	✓(L)	✓	18G	**14.01**

TABLE III

ABLATION STUDY ON THE EFFECTIVENESS OF SCALING UP AND PRE-TRAINING (VALIDATION SET).

Scaling-up	Pre-training	Test MEM	mIoU↑
×	×	8G	12.35
✓	×	12G	14.09
✓	✓	12G	14.43

driving scenario. Then, we semantically align the KITTI-360, BDD100K, Cityscapes, and nuImages datasets, further training on 19 common road elements specific to driving scenarios. Finally, We pre-train the backbone on 8 A100 GPUs for a total of 20k iterations.

C. *Quantitative and Qualitative Results*

In this section, we compare our method with competitive baselines on the test split of SemanticKITTI in Table I and demonstrate qualitative results in Fig. 3. Table I shows the comparison results with some methods adapted for RGB input such as LMSCNet [33], JS3C-Net [44] and other competitive camera-based semantic occupancy prediction methods such as TPVFormer [30] VoxFormer [17], and OccFormer [18] on SemanticKITTI. Overall, our method achieves a significant improvement over the other baselines in nearly all classes and sets new SOTA. To be specific, MonoOcc-S and MonoOcc-L achieve a remarkable boost of 1.60 mIoU and 3.43 mIoU, respectively, compared to the baseline method VoxFormer [17]. For the sake of fairness, in the following, we compare MonoOcc-S with VoxFormer-S to analyze the advantages of our methods. Delving into the qualitative results, we find that our algorithm achieves the expected performance improvement on long-tailed objects and small objects.

Our Method performs better on long-tailed objects. As shown in Table I, MonoOcc-S shows a significant improvement in predicting long-tailed objects compared with VoxFormer-S, such as the *other-ground* (0.56%, 5.60 → 9.70), *other-vehicle* (0.20%, 3.70 → 5.40) and *truck* (0.16%, 3.50 → 5.20).

Our Method performs better on small objects. As shown in Table I, MonoOcc-S demonstrates a significant boost in predicting small objects compared with VoxFormer-S, such as the *bicycle* (1.00 → 2.20), *motorcycle* (0.70 → 1.50) and *traffic-sign* (4.90 → 6.40).

Qualitative Results. To demonstrate the performance of our algorithm more intuitively, we also provide the qualitative visualization of the predicted semantic occupancy in Fig. 3. The four rows demonstrate the superiority of our method on **long-tailed objects**, **small objects** and **road segmentation**. The red box in the first row shows that VoxFormer-S cannot estimate the instance of *other-vehicle*. The orange box in the second row, third row, and last row show that our method performs better on small objects like *person*, *bicycle* and *pole*, respectively. To further demonstrate the effectiveness of our method, we show the impressive prediction result of *road*, a crucial category for autonomous vehicles to estimate the drivable area, within blue box in the second row and last row.

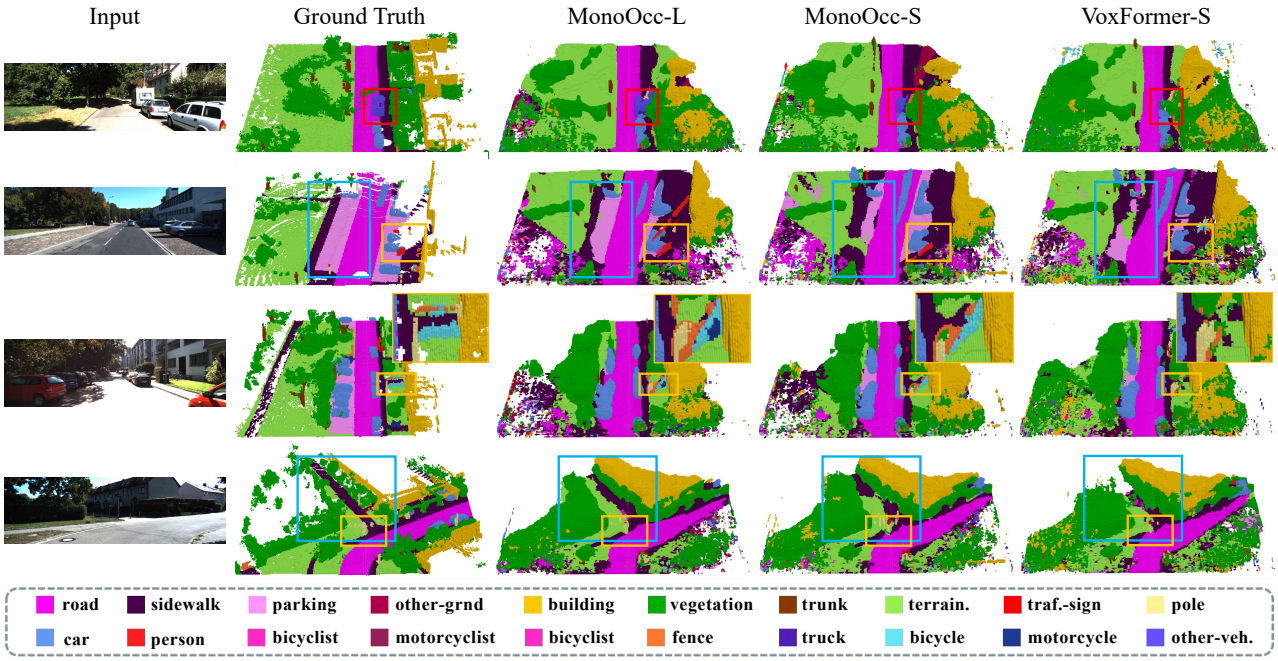

Fig. 3. Qualitative results of our method and VoxFormer on SemanticKITTI dataset

D. *Ablation Study*

We provide ablations on SemanticKITTI for the designs of each proposed component. Table II demonstrates the effectiveness of each component individually and in combination with other components on the single frame branch. In the column of the Distillation Module, (L) or (S) means distilling the temporal branch with the large or small backbone to the single frame branch.

2D Semantic Auxiliary Loss: The effectiveness of 2D Semantic Auxiliary Loss is shown in row 2 of Table II. It exceeds VoxFormer-S by 0.73 mIoU and scarcely increases GPU memory cost during training. Since auxiliary loss makes it possible to optimize the feature extractor in a shorter path, the performance is largely promoted.

Image Conditioned Cross-Attention: Row 4 of Table II shows the advantage of the image-conditioned cross-attention. The performance of the framework is improved by 0.35 mIoU with minimal extra memory cost (about 2G). Row 5 and row 9 demonstrate that the cross-attention significantly improves the performance of the single frame branch combined with the other two components, by introducing visual cues for completing occluded regions.

Scaling-up And Pre-training: Table III shows the positive impacts of scaling-up and pre-training. According to the comparison between row 1 and row 2, it is clear that increasing the parameters of the backbone can significantly improve the performance of occupancy prediction. While the results in row 2 and row 3 prove that pre-training the backbone network on driving datasets further improves the performance of occupancy prediction.

Distillation Module: Row 3, 5 and 7 of Table II demonstrate the effectiveness of the distillation module. Thanks to the transfer of knowledge from the privileged branch, the performance of the single-frame branch significantly increases. In addition, by comparing row 6 with row 8 and row 6 with row 9 of TableII, it is verified that knowledge from both multiple frames and large models can be introduced to the single-frame branch through the distillation module. The comparison between row 8 and row 9 of the table shows that scaling up the backbone of the privileged branch can also enhance the performance of the single-frame branch through the distillation module.

Table III shows the necessity of the distillation module. Using a larger backbone results in a significant increase in GPU memory usage during test time (8G $\rightarrow$ 12G), while the distillation module can transfer richer knowledge into the single-frame branch at a low cost.

V. Conclusion

In this paper, we present MonoOcc, a high-performance and efficient framework for monocular semantic occupancy prediction. We propose a semantic auxiliary loss and an image-conditioned cross-attention module, improving the existing 3D semantic occupancy prediction method. By proposing a distillation module to transfer temporal information and richer knowledge to the monocular branch from a privileged branch, we increase the performance of the framework especially on small and long-tailed objects, while striking a balance between performance and efficiency. Benefiting from these improvements, MonoOcc achieves SOTA performance on SemanticKITTI benchmark.

VI. Aknowledgement

This work is supported by the National Natural Science Foundation of China (NSFC) under Grants No.62173324 and the CAS for Grand Challenges under Grants 104GJHZ2022013GC.

REFERENCES

[1] Z. Yan, X. Li, K. Wang, S. Chen, J. Li, and J. Yang, "Distortion and uncertainty aware loss for panoramic depth completion," in *International Conference on Machine Learning*. PMLR, 2023, pp. 39 099–39 109.

[2] Z. Yan, X. Li, K. Wang, Z. Zhang, J. Li, and J. Yang, "Multi-modal masked pre-training for monocular panoramic depth completion," in *European Conference on Computer Vision*. Springer, 2022, pp. 378–395.

[3] L. Wang, H. Ye, Q. Wang, Y. Gao, C. Xu, and F. Gao, "Learning-based 3d occupancy prediction for autonomous navigation in occluded environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[4] M. Popovic, F. Thomas, S. Papatheodorou, N. Funk, T. Vidal-Calleja, and S. Leutenegger, "Volumetric occupancy mapping with probabilistic depth completion for robotic navigation," in *IEEE Robotics and Automation Letters*, 2021, pp. 5072–5079.

[5] P. Li, R. Zhao, Y. Shi, H. Zhao, J. Yuan, G. Zhou, and Y.-Q. Zhang, "Lode: Locally conditioned eikonal implicit scene completion from sparse lidar," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8269–8276, 2023.

[6] Z. Xia, Y. Liu, X. Li, X. Zhu, Y. Ma, Y. Li, Y. Hou, and Y. Qiao, "Scpnet: Semantic scene completion on point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[7] Y. Chen, H. Li, R. Gao, and D. Zhao, "Boost 3-d object detection via point clouds segmentation and fused 3-d giou-l1 loss," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 762–773, 2020.

[8] Y. Chen, D. Zhao, L. Lv, and Q. Zhang, "Multi-task learning for dangerous object detection in autonomous driving," *Information Sciences*, vol. 432, pp. 559–571, 2018.

[9] H. Li, Y. Chen, Q. Zhang, and D. Zhao, "Bifnet: Bidirectional fusion network for road segmentation," *IEEE transactions on cybernetics*, vol. 52, no. 9, pp. 8617–8628, 2021.

[10] A.-Q. Cao and R. de Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[11] Y. Zheng, C. Zhong, P. Li, H. Gao, Y. Zheng, B. Jin, L. Wang, H. Zhao, G. Zhou, Q. Zhang, and D. Zhao, "Steps: Joint self-supervised nighttime image enhancement and depth estimation," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[12] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[13] B. Jin, X. Liu, Y. Zheng, P. Li, H. Zhao, T. Zhang, Y. Zheng, G. Zhou, and J. Liu, "Adapt: Action-aware driving caption transformer," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[14] Z. Yan, K. Wang, X. Li, Z. Zhang, J. Li, and J. Yang, "Desnet: Decomposed scale-consistent network for unsupervised depth completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 3109–3117.

[15] ——, "Rignet: Repetitive image guided network for depth completion," in *European Conference on Computer Vision*. Springer, 2022, pp. 214–230.

[16] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[17] Y. Li, Z. Yu, C. Choy, C. Xiao, J. M. Alvarez, S. Fidler, C. Feng, and A. 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*, 2023.

[18] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," *arXiv preprint arXiv:2304.05316*, 2023.

[19] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," *arXiv preprint arXiv:2303.09551*, 2023.

[20] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[21] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[22] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*. PMLR, 2020.

[23] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[24] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022.

[25] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[26] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[27] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *2022 International conference on robotics and automation (ICRA)*, 2022.

[28] J. Philion and S. 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*. Springer, 2020.

[29] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[30] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[33] L. Roldao, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3d semantic completion," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020.

[34] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*, 2020.

[35] Y. Chen, D. Zhao, H. Li, D. Li, and P. Guo, "A temporal-based deep learning method for multiple objects detection in autonomous driving," in *2018 international joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–6.

[36] X. Zhao, Y. Chen, J. Guo, and D. Zhao, "A spatial-temporal attention model for human trajectory prediction." *IEEE CAA J. Autom. Sinica*, vol. 7, no. 4, pp. 965–974, 2020.

[37] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, Y. Rao, G. Huang, J. Lu, and J. Zhou, "Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Conference on Robot Learning*, 2023.

[38] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao, "Gmflow: Learning optical flow via global matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[39] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[40] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[41] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*, 2012.

[42] X. Chen, K.-Y. Lin, C. Qian, G. Zeng, and H. Li, "3d sketch-aware semantic scene completion via semi-supervised structure prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[43] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[44] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[45] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, 2017.

[46] Y. Liao, J. Xie, and A. Geiger, "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, 2022.

[47] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[48] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[49] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.