

# EFFOcc: A Minimal Baseline for Efficient Fusion-based 3D Occupancy Network

Yining Shi<sup>1,3</sup> Student member, IEEE, Kun Jiang<sup>1†</sup>, Ke Wang<sup>2</sup>, Kangan Qian<sup>1</sup>, Yunlong Wang<sup>1</sup>, Jiusi Li<sup>1</sup>, Tuopu Wen<sup>1</sup>, Mengmeng Yang<sup>1</sup>, Yiliang Xu<sup>4</sup>, Diange Yang<sup>1†</sup>

**Abstract**—3D occupancy prediction (Occ) is a rapidly rising challenging perception task in the field of autonomous driving which represents the driving scene as uniformly partitioned 3D voxel grids with semantics. Compared to 3D object detection, grid perception has great advantage of better recognizing irregularly shaped, unknown category, or partially occluded general objects. However, existing 3D occupancy networks (occnets) are both computationally heavy and label-hungry. In terms of model complexity, occnets are commonly composed of heavy Conv3D modules or transformers on the voxel level. In terms of label annotations requirements, occnets are supervised with large-scale expensive dense voxel labels. Model and data inefficiency, caused by excessive network parameters and label annotations requirement, severely hinder the onboard deployment of occnets. This paper proposes an efficient 3d occupancy network (EFFOcc), that targets the minimal network complexity and label requirement while achieving state-of-the-art accuracy. EFFOcc only uses simple 2D operators, and improves Occ accuracy to the state-of-the-art on multiple large-scale benchmarks: Occ3D-nuScenes, Occ3D-Waymo, and OpenOccupancy-nuScenes. On Occ3D-nuScenes benchmark, EFFOcc has only 18.4M parameters, and achieves 50.46 in terms of mean IoU (mIoU), to our knowledge, it is the occnet with minimal parameters compared with related occnets. Moreover, we propose a two-stage active learning strategy to reduce the requirements of labelled data. Active EFFOcc trained with 6% labelled voxels achieves 47.19 mIoU, which is 95.7% fully supervised performance. The proposed EFFOcc also supports improved vision-only occupancy prediction with the aid of region-decomposed distillation. Code and demo videos will be available at <https://github.com/synsin0/EFFOcc>.

**Index Terms**—Autonomous driving, 3D occupancy prediction, Multi-sensor fusion, Knowledge distillation, Active learning

## I. INTRODUCTION

Autonomous perception requires a comprehensive understanding of the environment. Common object-centric pipelines, which consist of detection, tracking and prediction, represents obstacles as bounding boxes. They are difficult to deal with extra-long, irregularly-shaped objects. Recent years witness a revival of occupancy grids on autonomous perception. Tesla has pioneered the extension of occupancy grid map (OGM) to an occupancy network (occnet). Tesla's FSD perception uses deep learning techniques to project visual features into 3D voxels and decode a variety of information such as occupancy, semantics, and motion flow [1]. Following this trend, new 3d occupancy benchmarks [2], [3] are built upon large-scale public datasets. These benchmarks formulate the task as semantic

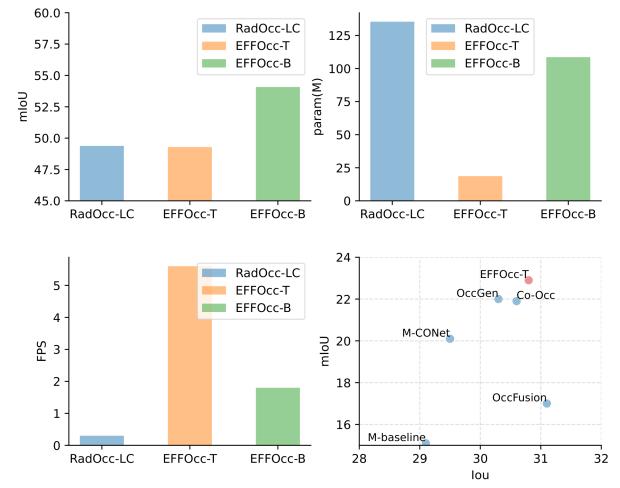


Fig. 1. Comparison of two variants of EFFOcc with other state-of-the-art occnets on Occ3D-nuScenes and OpenOccupancy-nuScenes benchmarks. EFFOcc-T uses ResNet18 as image backbone while EFFOcc-B and RadOcc-LC uses Swin-B as image backbone. We achieve better accuracy (top left), less parameters (top right), faster speed (bottom left), than RadOcc-LC. On OpenOccupancy-nuScenes dataset, EFFOcc-T model beats other fusion occnets in terms of geometric IoU and semantic mean IoU.

segmentation of foreground objects and background stuff on 3D voxel grids.

Despite recent success of high performance occnets, they bring a large amount of computational load which refrains real-time onboard deployment. Reduction of model complexity is one of the major directions of improving occnets [4]–[7]. However, existing computationally-efficient occnets focus on vision networks, while lightweight LiDAR-camera fusion occnets are rarely explored.

To this end, we introduce EFFOcc, an efficient 3D occupancy network with much faster inference speed and state-of-the-art performance. We expect EFFOcc to be able to train in an acceptable duration with commonly used GPU devices (2080ti) and will be deployed in real-time latency. Our motivation starts from the fact that LiDAR point cloud is naturally suitable for geometry reconstruction while a lightweight vision branch is readily enough to compensate for semantic recognition capability. We design a lightweight fusion network, EFFOcc, and use detection task as beneficial pretraining to achieve state-of-the-art performance. The comparison between EFFOcc and other fusion occnets is shown in Fig. 1.

Moreover, we conduct efficient active learning to figure out what is the minimal data requirement of occnet training. As

<sup>1</sup> School of Vehicle and Mobility, Tsinghua University, <sup>2</sup> Kargobot, <sup>3</sup> DiDi Chuxing, <sup>4</sup> Zongmu Technology. This work was done during Yining Shi's internship at DiDi Chuxing. <sup>†</sup>: Corresponding authors: Kun Jiang, Diange Yang (jiangkun@mail.tsinghua.edu.cn, ydg@mail.tsinghua.edu.cn).

the label generation process requires a lot of pre-processing, such as aggregation, matching, occlusion elimination, we believe that reducing the need of number of labelled voxels can significantly reduce the cost. We propose a two-stage active learning method for label-efficient adaptive annotation. First, select a certain proportion of high-value frames from the unlabeled pool as candidates, and then find high-value voxels from high-value frames as the final selections for active learning.

In summary, our contributions are listed as follows:

- We provide a simple baseline, EFFOcc, for fusion-based 3D occupancy prediction and lift EFFOcc to the state-of-the-art with lightweight design and proper pretraining techniques. We provide a occupancy-oriented distillation baseline to distill real-time vision-only occupancy network with fusion-based teacher model and gets competitive performance compared to other real-time vision occnets.
- We propose a novel two-stage active learning tailored on annotation cost reduction of 3D occupancy labels that succeeded in reducing annotation costs and computational cost for training occnets.
- We validate our models on three public benchmarks on two large-scale datasets, nuScenes and Waymo Open Dataset, and demonstrates their effectiveness.

## II. RELATED WORKS

### A. 3D Occupancy Prediction

3D occupancy prediction originates from the task of semantic scene completion (SSC) [8]. Occ3D [2] release benchmarks on large-scale surround-view datasets nuScenes and Waymo, and OccNet [9] extends the challenge to 3d occupancy and flow. Occupancy networks are generally developed from the extension of the vision bird's-eye view (BEV) networks (e.g. BEVDet [10] and BEVFormer [11]) and the semantic segmentation networks [12]. SurroundOcc [13] and OpenOccupancy [3] extracts multi-scale 3D voxel features with transformer merge them through deconvolutional upsampling. Multi-scale features proves better segmentation performance. TPVFormer [14] and PointOcc [15] extends BEV formulation to a triple-view feature transformation. PanoSSC [16], Symphonies [17] and OccFormer [18] proposes different mask-based transformer head for panoptic occupancy segmentation. FB-Occ [19] proposes a forward BEV-pooling style view transformation module and a backward BEVformer-style transformation and wins 1st on the occupancy prediction challenge.

### B. Computationally-efficient Occupancy Network

3D occupancy networks usually bring a huge amount of computation. Different methods are proposed in an effort to reduce the amount of 3D voxel calculations. PanoOcc [20] replaces 3D Conv operators with sparse conv at each layer while predicting the occupancy rate of nonempty voxels and deleting predicted empty voxels to maintain sparsity. FlashOcc [4] and FastOcc [5] proposes efficient channel-to-height devoid of complex 3D convolution computation. SparseOcc [6] proposes

fully sparse model to exploit geometry sparsity and sparse instance queries to fit object sparsity with mask transformers. SparseOcc [21] removes empty voxels after the geometry-based view transformation and uses spconv operators after that. Moreover, a sparse latent diffuser is proposed to diffuse empty voxels adjacent to occupied voxels. They achieve a remarkable 74.9& reduction of FLOPs.

### C. Efficient Learning for Autonomous Perception

The scope of efficient learning includes knowledge distillation, active learning, semi-supervised learning, and other techniques. These technologies are widely studied in 2D domain, but relatively fewer are studied in the 3D domain. Knowledge distillation is widely applied on vision BEV detection learning from teacher model (e.g. LiDAR-based or fusion-based detector) [22], [23]. However, occupancy task is more challenging compared to detection task as occupancy prediction has more severe class imbalance, not only class imbalance between foreground objects, but also background stuff elements. As a result, current occupancy networks suffer from low accuracy for foreground obstacles. To distill between voxel features, RadOcc [24] applies neural rendering to image plane as auxiliary supervision for distillation on voxel features.

Another label-efficient approach which effectively enhance training performance and reduces label costs is active learning, or dataset distillation, which is already explored in LiDAR-based 3D detection [25] and LiDAR segmentation [26], [27]. Annotator [27] proposes a voxel-centric active learning paradigm which actively labels points insides certain voxels where voxel confusion degree (VCD) is high. They demonstrate competitive performance with model trained with actively selected labels with a portion of less than 1% labelled minimum unit.

## III. METHODOLOGY

### A. Architecture

The architecture of EFFOcc is depicted in Fig. 2. Our overall goal is to pursue the minimization of the network and the lowest cost of training, from a model-centric and data-centric perspective, respectively. In section III-B, we first propose lightweight EFFOcc-LiDAR and EFFOcc-Fusion with only 2D operators and further lift their performance to the state-of-the-art with proper detection pretraining. In section III-C, using the fusion-based EFFOcc as teacher model and occupancy results as regions of interest, region-decomposed distillation improves the performance of vision-only EFFOcc-C. In section III-D, we propose a two-stage active leaning based on maximum entropy frame and voxel selection to explore extreme compression of labels without a significant drop in occupancy accuracy.

### B. EFFOcc Fusion Network

Our design goal is to achieve similar accuracy with the minimal possible network parameters. We start from the voxel-level dense fusion method introduced in OpenOccupancy [3]. It consists of a visual branch, a point cloud branch, an

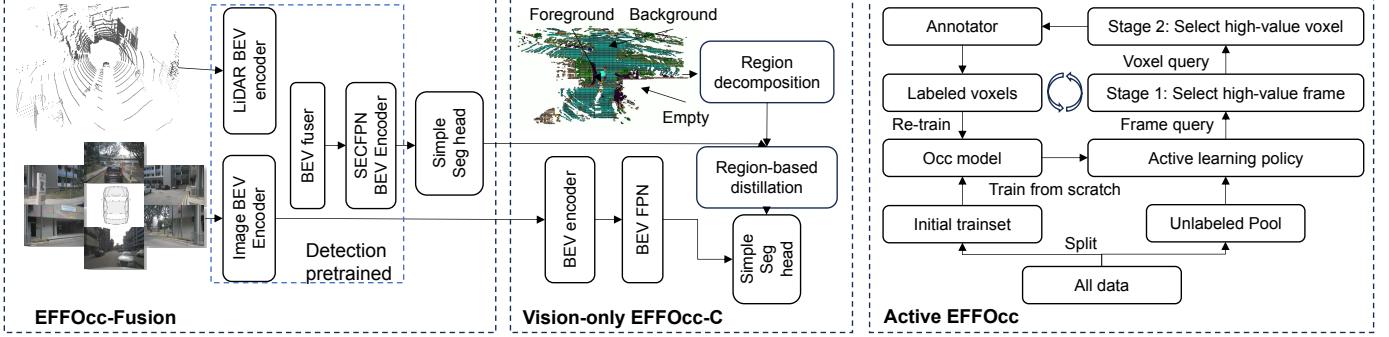


Fig. 2. EFOcc architecture consist of three parts: LiDAR-camera fusion network EFFOCC-Fusion, distillation enhanced vision-only network EFFOCC-C and active learning for EFFOCC trained on limited labels.

adaptive voxel fusion module supported by 3D convolution operators, and a multi-scale segmentation head with coarse-to-fine query. We replace each module with a lightweight version without losing performance accuracy. We remove 3D CNN on OpenOccupancy LiDAR branch, replace all Occ pooling and Conv3D Occ encoder with BEVpoolv2 [28] and Conv2D BEV encoder on OpenOccupancy vision branch, and replace voxel fusion layer with BEV fusion layer. Moreover, the model only uses single-scale feature map and single-stage coarse prediction. The comparison between our network and voxel-level dense fusion method is shown in Fig. 3.

For pointcloud branch, we use mean feature encoding as preprocessing and Spconv8x with downsample stride 8 as the LiDAR encoder. Then the sparse 3D features are splatted to BEV features. For image branches, We use a image encoder and adopt BEVpoolv2 [28] as view projector to accelerate transformation from perspective view to BEV. We adopt simple conv2d-based operator as fusion layer. After the fusion, we enter the BEV encoder stage, and the BEV encoder network structure is the same SECONd FPN as the 2D encoder of point cloud detection, instead of resnet18, which benefits more from detection pre-training and uses less parameters. The occupancy head consists of two Conv2d layers and height channel is detached from feature channel for the final 3D output.

We training the lightweight fusion model from scratch performs around  $-3.0$  mIoU inferior with latest fusion-based RadOcc [24] teacher model, which has very much similar structure with OpenOccupancy [3]. We find that devoid of the 3D CNN after sparse convolution(Spconv) naturally reduces the performance. We find that one promising approach to mitigate the gap is to pretrain model parts with detection tasks. Based on this observation, we adjust the network to be closer to well-established LiDAR detection network. Experiments show that if the model is loaded from the checkpoint of the corresponding detection network, EFOCC achieves similar performance with more complex occnets.

**Training strategies.** We initialize from checkpoints of DAL [29] instead of random initialization cause we find the detection pretraining improve foreground segmentation accuracy.

We use existing losses from prior works. They are cross-entropy loss  $\mathcal{L}_{ce}$ , lovasz-softmax loss  $\mathcal{L}_{ls}$  [30], affinity loss  $\mathcal{L}_{geo}^{seg}$  and  $\mathcal{L}_{seg}^{geo}$  [31]. For Occ3D-waymo case, we use OHEM [32] loss. The lovasz-softmax loss and affinity loss consume

more GPU memory, and improve greatly on OpenOccupancy benchmark, but helps less ( $< 1.0$  in term of mIoU) on Occ3D benchmark. For most experiments on Occ3D benchmark unless especially mentioned, we only use cross-entropy loss to save GPU memory at the training stage. The total loss  $\mathcal{L}_{total}$  is the weighted sum of each loss,

$$\mathcal{L}_{total} = w_{ce} \cdot \mathcal{L}_{ce} + w_{ls} \cdot \mathcal{L}_{ls} + w_{geo} \cdot \mathcal{L}_{geo}^{seg} + w_{seg} \cdot \mathcal{L}_{seg}^{geo} \quad (1)$$

We set all weights  $w_{ce} = w_{ls} = w_{geo} = w_{seg} = 1.0$ .

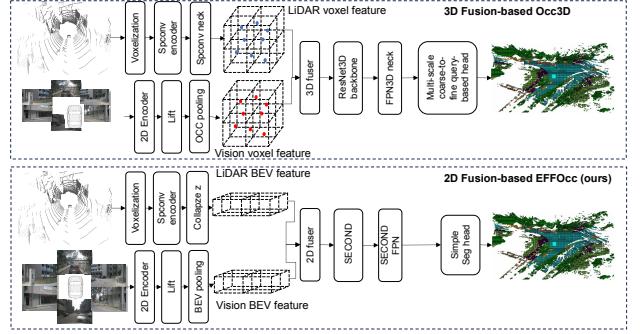


Fig. 3. Network details of EFOCC framework compared to dense fusion occnets [3], [24]. Our lightweight design replaces voxel feature with BEV feature, occ pooling to BEV pooling, ResNet3D backbone to SECONd backbone, complex prediction head to a simple Conv2D head.

### C. Occupancy-oriented Distillation

We try to apply the trained fusion model to improve the performance of the vision-only model. Existing BEV distillation methods are usually designed for detection methods, mainly for multi-stage (feature-level, box-level, etc) knowledge transfer of detected boxes. To start with, we first conduct a naive distillation which simply performs full-space feature alignment between BEV features generated from fusion-based teacher model and vision-based student model, but fails to improve accuracy. One possible reason is that the occupancy network needs to deal with the foreground, background and empty surroundings at the same time, and faces a severer unbalanced semantic distribution. Our statistics on BEV feature maps finds that less than 1% pillars are with foreground objects, around 40% pillars are with background, while the rest pillars are all empty. We design the distillation strategy to focus more

on foreground voxels. We decompose the full BEV space to three sub-regions as above. We force the student to focus more on foreground and background regions by defining a region weight map  $W$ :

$$W_{i,j} = \begin{cases} 4e - 5, & \text{if } (i, j) \in \text{Foreground} \\ 2e - 6, & \text{if } (i, j) \in \text{Background} \\ 4e - 7, & \text{if } (i, j) \in \text{Empty} \end{cases} \quad (2)$$

In Eq. 2,  $i, j$  denote the coordinate indexes on a BEV feature map.

The distillation loss between BEV feature from teacher  $F^t$  and student  $F^s$  are:

$$L_{\text{distill}} = \sum_{c=1}^C \sum_{i=1}^H \sum_{j=1}^W W_{i,j} \left( F_{c,i,j}^{t(n)} - \tilde{F}_{c,i,j}^{s(n)} \right)^2 \quad (3)$$

The vision-only student network is trained with the sum of distillation loss and classification loss.

#### D. Two-stage Active Learning

We hope active learning approach helps to save the annotation cost of occupancy labels and the computation cost of occupancy frames while achieving similar performance with 100% label trained model.

The active learning process is depicted in the left part of Fig. 2. Firstly, We evenly collect samples with fixed skip intervals and form the initialized labelled training set. The other unlabelled samples form a unlabelled pool. For model inference on each sample, the model outputs class probabilities on each voxel. On the first stage, we actively choose high-value frames given frame budget mainly for the sake of reducing computing costs. On the second stage, we actively choose high-value voxels mainly for every chosen labelled frame for the sake of reducing annotation costs.

Evaluating the value of new sample data without annotation is the core issue of active learning. Many prior works characterize value by prediction uncertainty: If the prediction results demonstrates highly uncertainty, Annotating these samples bring benefits to a new-round training. Otherwise, samples of which predictions of low uncertainty are regarded as redundant training resources. This paper thinks occupancy as a segmentation task rather than a localization task, which is the classification of every voxel grid. So we choose ENTROPY [33] as main active selection policy. ENTROPY is an uncertainty-based approach that targets the semantic classification head of the occupancy predictor. We use maximum entropy as the primary criteria for active selection.

We pass all samples from the unlabeled pool to the occnet and extract the predicted voxel occupancy  $O$  with shape  $[H, W, L, C]$  where  $H, W, L$  are the height, width and length of voxels and  $C$  is the number of semantic categories. The mean entropy  $H_i$  of the  $i$ -th frame sample  $O_i$  is calculated as,

$$H_i = \frac{1}{N} \sum_{n=1}^N -(O_i) \cdot \log(O_i) \quad (4)$$

where  $N$  is the number of voxels,  $N = H \times W \times L$ .

For the first stage, we get mean entropy for every unlabelled frame, sort them in descending order and select the budget limit number of samples and record the selected frames. This active selection cycles may last for multiple times. For the second stage, we sort entropy of all voxels, group them by semantic category and retain budget limit number of voxels as labelled subsets.

## IV. EXPERIMENTS

### A. Datasets and Metrics

We validate our model on three popular occupancy benchmarks: Occ3D-nuScenes [2], Occ3D-Waymo [2], OpenOccupancy-nuScenes [3]. The detailed introductions of each dataset and benchmark are available in the Appendix. The primary metric of all benchmarks for 3D occupancy prediction is mean IoU from average of all semantic categories.

### B. Implementation Details

**Data Pre-processing.** For both training and inference phrases, we first load multi-view images with camera parameters, then apply normalization, padding, and multi-scale flipping to each input image for image augmentation. We aggregate multi-sweep LiDAR point clouds and conduct random flipping on point clouds and voxel labels for BEV augmentation. We don't use any test-time augmentation techniques.

**Training and Inference.** We build our upon the MMDetection3D version 1.0.0rc4 [34]. Most experiments are trained on 8 2080TI GPUs for 24 epochs in a total batch size of 16, with training time less than 24 hours. We use AdamW optimizer with a learning rate of 0.0001 and weight decay 0.01. We use exponential moving average (EMA) hook for better accuracy. We set batch size as 1 in the inference stage.

### C. Comparison with State-of-the-art

For comparison on different model scales, we provide three model variants: EFFOcc-L is the LiDAR-only model with sparse conv as LiDAR backbone; EFFOcc-C is the camera-only model which is the student model distilled from the fusion-based teacher model. The model design is very much similar to FlashOcc [4]; EFFOcc-T and EFFOcc-B is the tiny and base scale version of fusion-based model, mainly differing in the scale of image backbones.

*1) Results on Occ3d-nuScenes:* Results on Occ3d-nuScenes validation set is shown in Tab. I. The baselines are state-of-the-art vision models and the most recent teacher model of RadOcc, RadOcc-LC in the table. Our LiDAR-model EFFOcc-L achieves similar mIoU compared to state-of-the-art vision-based models on top of Swin-base backbone. Our EFFOcc-T<sup>A</sup> (trained for 24 epochs) achieves +4.16 gain compared to LiDAR model and is only -0.09 worse than RadOcc-LC. If we train longer to 48 epochs, the model EFFOcc-T<sup>B</sup> improves to +0.21 mIoU better than RadOcc-LC. If EFFOcc-T<sup>C</sup> is trained with all loss, it has a slight improvement to 50.46mIoU. We further train EFFOcc-base which has similar parameters (the same Swin-base image

TABLE I

**3D OCCUPANCY PREDICTION PERFORMANCE ON THE OCC3D-NUSCENES VALIDATION SET.** † DENOTES THE PERFORMANCE REPRODUCED BY OFFICIAL CODES. \* MEANS THE RESULTS PROVIDED BY ORIGINAL PAPER. WE REPORT THREE VARIANTS OF EFFOCC-T WHICH HAVE SAME STRUCTURES. EFFOCC<sup>A</sup> IS TRAINED ONLY WITH CE LOSS WITH 2X LEARNING SCHEDULE (24 EPOCHS). EFFOCC<sup>B</sup> IS TRAINED WITH CE LOSS WITH 4X LEARNING SCHEDULE (48 EPOCHS). EFFOCC<sup>C</sup> IS TRAINED WITH ALL FOUR LOSSES WITH 2X LEARNING SCHEDULE (24 EPOCHS). EFFOCC-BASE IS TRAINED WITH CE LOSS WITH 2X LEARNING SCHEDULE (24 EPOCHS).

Method	Modality	Backbone	mIoU	Others	barrier	bicycle	bus	car	cons. veh	motorcycle	pedestrian	traffic cone	trailer	truck	dri. sur	other flat	sidewalk	terrain	mammal	vegetation
Performances on nuScenes Validation Set																				
CTF-Occ	C	R101	28.53	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
BEVFormer	C	R101	39.24	10.13	47.91	24.90	47.57	54.52	20.23	28.85	28.02	25.73	33.03	38.56	81.98	40.65	50.93	53.02	43.86	37.15
PanoOcc	C	R101	42.13	11.67	50.48	29.64	49.44	55.52	23.29	33.26	30.55	30.99	34.43	42.57	83.31	44.23	54.40	56.04	45.94	40.40
BEVDet†	C	Swin-B	42.02	12.15	49.63	25.10	52.02	54.46	27.87	27.99	28.94	27.23	36.43	42.22	82.31	43.29	54.62	57.90	48.61	43.55
RadOcc-C*	C	Swin-B	46.06	9.78	54.93	20.44	55.24	59.62	30.48	28.94	44.66	28.04	45.69	48.05	81.41	39.80	52.78	56.16	64.45	62.64
RadOcc-LC*	LC	Swin-B	49.38	10.93	58.23	25.01	57.89	62.85	34.04	33.45	50.07	32.05	48.87	52.11	82.9	42.73	55.27	58.34	68.64	66.01
EFFOCC-L(ours)	L	-	45.13	7.70	49.55	17.93	55.46	60.26	29.11	27.57	51.93	30.15	42.12	47.25	77.88	33.10	48.88	54.18	68.21	65.88
EFFOCC-T <sup>A</sup> (ours)	LC	R18	49.29	10.57	56.16	21.73	58.68	63.16	31.98	37.71	55.4	36.15	45.87	50.81	81.02	39.07	53.08	57.15	70.41	68.90
EFFOCC-T <sup>B</sup> (ours)	LC	R18	49.59	10.76	56.64	23.39	58.45	63.53	32.05	38.61	55.93	36.93	45.91	50.8	81.08	39.24	53.23	57.17	70.29	68.97
EFFOCC-T <sup>C</sup> (ours)	LC	R18	50.46	14.34	57.22	40.82	57.60	61.99	34.93	50.18	55.92	42.9	40.05	50.09	77.84	38.6	47.78	54.9	67.36	65.31
EFFOCC-Base(ours)	LC	Swin-B	<b>54.08</b>	15.74	60.98	36.21	62.24	66.42	38.68	43.88	52.12	42.40	50.29	56.08	84.92	48.00	58.60	61.99	71.29	69.48

TABLE II

**3D OCCUPANCY PREDICTION PERFORMANCE ON THE OCC3D-WAYMO VALIDATION SET.** † DENOTES THE PERFORMANCE REPRODUCED BY OFFICIAL CODES. \* MEANS THE RESULTS PROVIDED BY OCC3D [2].

Method	Modality	Backbone	mIoU	general object	vehicle	bicyclist	pedestrian	sign	traffic light	car	construction cone	bicycle	motorcycle	building	vegetation	tree truck	road	sidewalk
Training with 20% training data for 8 epochs																		
BEVDet*	C	R101	9.88	0.13	13.06	2.17	10.15	7.80	5.85	4.62	0.94	1.49	0.0	7.27	10.06	2.35	48.15	34.12
BEVFormer*	C	R101	15.62	2.59	25.76	13.87	4.11	14.23	3.35	8.41	7.54	3.45	0.0	18.46	16.21	6.87	67.72	41.68
CTF-Occ*	C	R101	18.73	6.26	28.09	14.66	8.22	15.44	10.53	11.78	13.62	16.45	0.65	18.63	17.3	8.29	67.99	42.98
EFFOCC-C(ours)	C	R50	19.20	5.90	26.80	16.41	7.18	12.98	8.50	10.84	9.55	4.22	0.00	23.16	22.19	7.89	76.70	55.64
EFFOCC-L(ours)	L	-	41.62	4.28	66.61	51.97	34.00	30.82	30.23	45.39	27.16	12.88	0.00	65.35	61.52	41.26	81.93	70.86
EFFOCC-T(ours)	LC	R18	<b>43.52</b>	10.04	65.05	54.74	35.85	39.57	30.23	46.76	32.08	18.07	0.03	62.53	60.78	43.41	83.26	70.42
Training with 100% training data for 24 epochs																		
EFFOCC-T(ours)	LC	R18	49.59	13.99	69.5	57.76	45.64	47.50	34.48	51.44	38.49	40.02	1.65	69.11	64.02	47.52	86.34	76.32
EFFOCC-L(ours)	L	-	<b>50.35</b>	10.34	75.97	63.90	46.35	50.14	33.19	55.50	30.93	27.58	0.00	74.10	72.99	50.11	86.83	77.37

backbone) with baselines and gets a remarkable +4.70 mIoU gain compared with RadOcc-LC.

We detail the network component, image configurations, parameters and running speed in Tab. III. In summary, our model is a computationally-efficient occupancy network which has 18x acceleration under the same mIoU precision and +3 mIoU gain under the same backbone compared with the state-of-the-art.

2) *Results on Occ3d-Waymo:* Results on Occ3d-waymo validation set is shown in Tab. II. As cameras in waymo dataset do not have the coverage of 360 view, we evaluate

voxels visible to cameras. For fast validation on smaller scale data, We follow the practice of Occ3D [2] and trains every model with 20% training data for 8 epochs. The vision model baselines are from Occ3D [2]. Our vision-only model has a +0.47mIoU gain compared to CTF-Occ [2]. LiDAR model achieves almost twice the mIoU precision as compared to vision-based methods, mainly because that Waymo's LiDARs are significantly stronger than nuScenes's 32-beam LiDAR. Adding a light image branch to LiDAR model achieves a reasonable increase of +1.90 mIoU. However, if the model

TABLE III

COMPLEXITY AND RUNTIME ANALYSIS OF OUR EFFOCC AND RADOCC-LC. RADOCC STATISTICS ARE REPORTED FROM THE ORIGINAL PAPER.

Model	Modality	PC backbone	Img backbone	Image Size	Epochs	FPS	Param	mIoU
BEVDetOcc	C	-	Swin-B	512x1408	24	1.0	125.91M	44.10
RadOcc-LC	LC	SECOND+3DCNN	Swin-B	512x1408	-	0.3	135.39M	49.38
EFFOcc-T(Ours)	LC	SECOND	R18	256x704	24	<b>5.6</b>	<b>18.66M</b>	49.29
EFFOcc-B(Ours)	LC	SECOND	Swin-B	512x1408	24	1.8	108.59M	<b>54.08</b>

TABLE IV

3D OCCUPANCY PREDICTION PERFORMANCE OF REAL-TIME VISION-ONLY MODELS ON THE OCC3D-NUSCENES [2] DATASET. “8F” MEANS FUSING TEMPORAL INFORMATION FROM 7+1 FRAMES.

Method	Backbone	Input Size	mIoU	Others	barrier	bicycle	bus	car	cons. veh	motorcycle	pedestrian	traffic cone	trailer	truck	dri. sur	other flat	sidewalk	terrain	mammade	vegetation
MonoScene [31]	R101	1600×900	6.1	1.8	7.2	4.3	4.9	9.4	5.7	4.0	3.0	5.9	4.5	7.2	14.9	6.3	7.9	7.4	1.0	7.7
OccFormer [18]	R101	1600×900	21.9	5.9	30.3	12.3	34.4	39.2	14.4	16.5	17.2	9.3	13.9	26.4	51.0	31.0	34.7	22.7	6.8	7.0
BEVFormer [11]	R101	1600×900	26.9	5.9	37.8	17.9	40.4	42.4	7.4	23.9	21.8	21.0	22.4	30.7	55.4	28.4	36.0	28.1	20.0	17.7
CTF-Occ [2]	R101	1600×900	28.5	8.1	39.3	20.6	38.3	42.2	16.9	24.5	22.7	21.1	23.0	31.1	53.3	33.8	38.0	33.2	20.8	18.0
TPVFormer [14]	R101	1600×900	27.8	7.2	38.9	13.7	40.8	45.9	17.2	20.0	18.9	14.3	26.7	34.2	55.7	35.5	37.6	30.7	19.4	16.8
SparseOcc [6] (1f)	R50	704×256	27.0	8.8	33.2	17.1	34.4	41.0	16.1	19.2	20.8	21.0	18.4	27.9	62.4	31.0	39.2	35.1	17.5	16.8
SparseOcc [6] (8f)	R50	704×256	30.9	10.6	39.2	20.2	32.9	43.3	19.4	23.8	23.4	29.3	21.4	29.3	67.7	36.3	44.6	40.9	22.0	21.9
BEVDetOcc [10] (1f)	R50	704×256	31.64	6.65	36.97	8.33	38.69	44.46	15.21	13.67	16.39	15.27	27.11	31.04	78.7	36.45	48.27	51.68	36.82	32.09
FlashOcc [4] (1f)	R50	704×256	32.08	6.74	37.65	10.26	39.55	44.36	14.88	13.4	15.79	15.38	27.44	31.73	78.82	37.98	48.7	52.5	37.89	32.24
EFFOcc-C(ours)	R50	704×256	<b>33.43</b>	7.31	40.7	13.05	39.2	46.43	20.85	16.98	18.45	18.55	27.79	33.2	78.42	37.62	47.59	52.18	38.1	31.83

is saturated trained for 24 epochs on 100% data(150k training samples in total), the LiDAR-only model outperforms fusion model by 0.76 mIoU. The abnormal phenomenon of accuracy decrease with image branch may be due to incomplete coverage of vision and conflict between LiDAR and vision, which we leave for future research.

3) *Results on OpenOccupancy-nuScenes:* Results on OpenOccupancy-nuScenes validation set is shown in Tab. V. Our model is more lightweight than others cause we use image backbone ResNet-18 and image size  $256 \times 704$ , but they use ResNet-50 and image size  $896 \times 1600$ . Compared with other LiDAR-camera fusion occnets, we achieve the best semantic mIoU of 22.9 and the best geometric IoU of 30.8. By comparing the task design of this benchmark with Occ3D-nuScenes, our method also demonstrates equally good performance under larger perception range and finer grid resolution.

4) *Results on Distilled Vision Occupancy Network:* Results of vision-only occupancy network are shown in Tab. IV. EFFOcc-C uses ResNet-50 [35] as image backbone and inputs single-frame image with size  $704 \times 256$  and uses ResNet50 as input. In this setting, our distilled model works 1.35 mIoU better than latest state-of-the-arts FlashOcc [4].

#### D. Ablation Study

We conduct ablative experiments to demonstrate the efficacy of each component in our network design and learning strategies. Without further notice, all experiments are conducted on EFFOcc-T with ResNet18 as image backbone.

1) *Effect of Pretraining Strategy:* We provide model details in Tab. VI and report their mIoUs respectively. RadOcc LiDAR and fusion-based version are our baseline with dense Conv3D

operators. If model is initialized randomly, we observe an obvious gap of 3.45 mIoU between Conv2D-based EFFOcc-L and RadOcc-L. Similarly, EFFOcc-T is 3.41 mIoU inferior compared with RadOcc-LC. However, after our detection pretraining, our EFFOcc-L and EFFOcc-T is only 0.97 and 0.09 mIoU away from our baselines.

2) *Effect of Active Learning Strategies:* We provide active learning details in Tab. VII and report their mIoUs repetitively. We conducted three series of active learning experiments, using only the first stage (frame selection), using only the second stage (voxel selection in selected frames), and a two-stage protocol. For stage one only, the accuracy of model trained with active labels is significantly enhanced, and the accuracy is higher when the number of frames is the same (e.g., the performance of initial 5% plus active 5% is higher than initial 10% by 2.49mIoU). For stage two only, active selection of extremely few voxels (1% or 0.1%) on all available frames achieves competitive results ( $-2.1$  for 1% data and  $-4.26$  for 0.1% data) compared to model trained with 100% frames and 100% voxel labels. The two-stage experiments are trade-offs between computation costs which mainly depend on stage one and annotated costs which mainly depend on stage two. A good balance between annotation, computation and accuracy is achieved with 50% frames, 1% voxels per frame and 44.51mIoU. For the frequency of active selection, it is recommended to use a high frequency to avoid overfitting when the budget is low, and a low frequency to fully learn when the budget is rich.

## V. CONCLUSION

This paper mainly discusses the minimal requirements of training a high performance occupancy network baseline with

TABLE V

**3D SEMANTIC OCCUPANCY PREDICTION RESULTS ON NUSCENES-OCCUPANCY VALIDATION SET.** WE REPORT THE GEOMETRIC METRIC IoU, SEMANTIC METRIC mIoU, AND THE IOU FOR EACH SEMANTIC CLASS. THE C, D, AND L DENOTES CAMERA, DEPTH, AND LiDAR, RESPECTIVELY. **BOLD** REPRESENTS THE BEST SCORE.

Method	Modality	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [31]	C	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3	7.9	7.4	10.0	7.6
TPVFormer [14]	C	15.3	7.8	9.3	4.1	11.3	10.1	5.2	4.3	5.9	5.3	6.8	6.5	13.6	9.0	8.3	8.0	9.2	8.2
3DSketch [36]	C&D	25.6	10.7	12.0	5.1	10.7	12.4	6.5	4.0	5.0	6.3	8.0	7.2	21.8	14.8	13.0	11.8	12.0	21.2
AICNet [37]	C&D	23.8	10.6	11.5	4.0	11.8	12.3	5.1	3.8	6.2	6.0	8.2	7.5	24.1	13.0	12.8	11.5	11.6	20.2
LMSNet [38]	L	27.3	11.5	12.4	4.2	12.8	12.1	6.2	4.7	6.2	6.3	8.8	7.2	24.2	12.3	16.6	14.1	13.9	22.2
JS3C-Net [39]	L	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3	14.9	16.2	14.0	24.9
OccFusion [40]	C&L	31.1	17.0	15.9	15.1	15.8	18.2	15.0	17.8	17.0	10.4	10.5	15.7	26.0	19.4	19.3	18.2	17.0	21.2
M-baseline [3]	C&L	29.1	15.1	14.3	12.0	15.2	14.9	13.7	15.0	13.1	9.0	10.0	14.5	23.2	17.5	16.1	17.2	15.3	19.5
M-CONet [3]	C&L	29.5	20.1	23.3	13.3	21.2	24.3	15.3	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
Co-Occ [41]	C&L	30.6	21.9	26.5	16.8	22.3	27.0	10.1	20.9	20.7	14.5	16.4	21.6	36.9	23.5	25.5	23.7	20.5	23.5
OccGen [42]	C&L	30.3	22.0	24.9	16.4	22.5	26.1	14.0	20.1	21.6	14.6	17.4	21.9	35.8	24.5	24.7	24.0	20.5	23.5
EFFOcc-T(ours)	C&L	<b>30.8</b>	<b>22.9</b>	28.1	16.7	22.1	27.3	13.0	24.8	36.2	22.6	16.8	21.6	29.4	13.9	18.2	20.6	26.5	28.8

TABLE VI

ABLATION STUDY OF PRETRAINING STRATEGY ON LiDAR-BASED AND FUSION-BASED OCCUPANCY NETWORKS.

Network	Modality	Pretrain	mIoU
RadOcc-L	L	-	46.01
EFFOcc-L	L	None	42.56
EFFOcc-L	L	Det	45.13
RadOcc-LC	LC	-	49.38
EFFOcc-T	LC	None	45.97
EFFOcc-T	LC	Det	49.29

joint input of point clouds and multi-view cameras. While on-par with or surpassing the performance of existing occnets on two large-scale public datasets, we significantly reduce training costs and equipment requirements, and enhance usability. Furthermore, we design a distillation strategy so that the fusion network can assist in enhancing the accuracy of the vision-only lightweight occupancy network. We also propose a two-stage active learning strategy that greatly reduces training cost and annotation cost while maintaining competitive performance. For future works, we will investigate active occnets under self-supervised conditions. We will also add TensorRT support for EFFOcc to further improve inference speed.

#### ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (U22A20104, 52372414, 52102464), Beijing Natural Science Foundation (L231008), and Young Elite Scientist Sponsorship Program By BAST(BYESS2022153). This work was also sponsored by Tsinghua University-DiDi Joint Research Center for Future

TABLE VII

ABLATION STUDY OF ACTIVE STRATEGY ON DIFFERENT PORTIONS OF ANNOTATED LABELS. ALL EXPERIMENTS ARE CONDUCTED ON FUSION-BASED TINY MODEL EFFOCC-T. 'INIT FR.' AND 'NEW FR.' IS THE INITIAL TRAINING FRAMES PORTION AND STAGE 1 NEWLY SELECTED FRAMES. 'LABEL.VO.' IS THE PORTION OF LABELLED VOXELS PER FRAME. 'FREQ' DENOTES ACTIVE SELECTION FREQUENCY, I.E. THE EPOCH INTERVAL OF ACTIVE SELECTIONS.

Init Fr.	Stage 1 New Fr.	Freq	Stage 2		mIoU
			Label.Vo.	Freq	
100%	0%	-	100%	-	49.29
5%	0%	-	100%	-	33.42
5%	5%	6	100%	-	38.09(+4.67)
5%	5%	1	100%	-	39.03(+5.61)
10%	0%	-	100%	-	36.54
10%	10%	1	100%	-	42.03(+5.49)
20%	0%	-	100%	-	41.69
20%	10%	1	100%	-	43.38(+1.69)
5%	95%	-	1%	10	46.83
5%	95%	-	1%	1	<b>47.19</b>
5%	95%	-	0.1%	10	45.03
5%	95%	-	0.1%	1	44.48
5%	5%	1	1%	10	38.77
5%	15%	1	1%	10	40.70
5%	45%	1	1%	10	<b>44.51</b>
10%	10%	1	1%	10	40.63
20%	10%	1	1%	10	41.27

Mobility and Tsinghua University-Zongmu Technology Joint Research Center.

#### REFERENCES

- [1] Y. Shi, K. Jiang, J. Li, J. Wen, Z. Qian, M. Yang, K. Wang, and D. Yang, "Grid-centric traffic scenario perception for autonomous driving: A comprehensive review," *arXiv preprint arXiv:2303.01212*, 2023.
- [2] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

- [3] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, “Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception,” 2023.
- [4] Z. Yu, C. Shu, J. Deng, K. Lu, Z. Liu, J. Yu, D. Yang, H. Li, and Y. Chen, “Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin,” *arXiv preprint arXiv:2311.12058*, 2023.
- [5] J. Hou, X. Li, W. Guan, G. Zhang, D. Feng, Y. Du, X. Xue, and J. Pu, “Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird’s-eye view and perspective view,” 2024.
- [6] H. Liu, H. Wang, Y. Chen, Z. Yang, J. Zeng, L. Chen, and L. Wang, “Fully sparse 3d panoptic occupancy prediction,” *arXiv preprint arXiv:2312.17118*, 2023.
- [7] P. Tang, Z. Wang, G. Wang, J. Zheng, X. Ren, B. Feng, and C. Ma, “Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction,” *arXiv preprint arXiv:2404.09502*, 2024.
- [8] 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, pp. 9297–9307.
- [9] A. Ahmed, Z. Xiaoyang, M. H. Tunio, M. H. Butt, S. A. Shah, Y. Chengxiao, F. A. Pirzado, and A. Aziz, “Ocnet: Improving imbalanced multi-centred ovarian cancer subtype classification in whole slide images,” in *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. IEEE, 2023, pp. 1–8.
- [10] 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.
- [11] 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, pp. 1–18.
- [12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Giridhar, “Masked-attention mask transformer for universal image segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022.
- [13] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, “Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 729–21 740.
- [14] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, “Tri-perspective view for vision-based 3d semantic occupancy prediction,” *arXiv preprint arXiv:2302.07817*, 2023.
- [15] S. Zuo, W. Zheng, Y. Huang, J. Zhou, and J. Lu, “Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction,” *arXiv preprint arXiv:2308.16896*, 2023.
- [16] Y. Shi, J. Li, K. Jiang, K. Wang, Y. Wang, M. Yang, and D. Yang, “Panosc: Exploring monocular panoptic 3d scene reconstruction for autonomous driving,” in *2024 International Conference on 3D Vision (3DV)*, 2024.
- [17] H. Jiang, T. Cheng, N. Gao, H. Zhang, W. Liu, and X. Wang, “Symphonize 3d semantic scene completion with contextual instance queries,” *arXiv preprint arXiv:2306.15670*, 2023.
- [18] Y. Zhang, Z. Zhu, and D. Du, “Ocformer: Dual-path transformer for vision-based 3d semantic occupancy prediction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9433–9443.
- [19] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, and J. M. Alvarez, “Fb-bev: Bev representation from forward-backward view transformations,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 6919–6928.
- [20] Y. Wang, Y. Chen, X. Liao, L. Fan, and Z. Zhang, “Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation,” *arXiv preprint arXiv:2306.10013*, 2023.
- [21] H. Liu, H. Wang, Y. Chen, Z. Yang, J. Zeng, L. Chen, and L. Wang, “Fully sparse 3d panoptic occupancy prediction,” *arXiv preprint arXiv:2312.17118*, 2023.
- [22] S. Zhou, W. Liu, C. Hu, S. Zhou, and C. Ma, “Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird’s-eye view,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [23] Z. Wang, D. Li, C. Luo, C. Xie, and X. Yang, “Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8637–8646.
- [24] H. Zhang, X. Yan, D. Bai, J. Gao, P. Wang, B. Liu, S. Cui, and Z. Li, “Radocc: Learning cross-modality occupancy knowledge through rendering assisted distillation,” *arXiv preprint arXiv:2312.11829*, 2023.
- [25] Y. Luo, Z. Chen, Z. Wang, X. Yu, Z. Huang, and M. Baktashmotagh, “Exploring active 3d object detection from a generalization perspective,” *arXiv preprint arXiv:2301.09249*, 2023.
- [26] M. Liu, Y. Zhou, C. R. Qi, B. Gong, H. Su, and D. Anguelov, “Less: Label-efficient semantic segmentation for lidar point clouds,” in *European conference on computer vision*. Springer, 2022, pp. 70–89.
- [27] B. Xie, S. Li, Q. Guo, C. Liu, and X. Cheng, “Annotator: A generic active learning baseline for lidar semantic segmentation,” *Advances in Neural Information Processing Systems*, vol. 36, 2023.
- [28] J. Huang and G. Huang, “Bevpoolv2: A cutting-edge implementation of bevdet toward deployment,” *arXiv preprint arXiv:2211.17111*, 2022.
- [29] J. Huang, Y. Ye, Z. Liang, Y. Shan, and D. Du, “Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection,” *arXiv preprint arXiv:2311.07152*, 2023.
- [30] M. Berman, A. R. Triki, and M. B. Blaschko, “The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4413–4421.
- [31] 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, pp. 3991–4001.
- [32] A. Shrivastava, A. Gupta, and R. Girshick, “Training region-based object detectors with online hard example mining,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 761–769.
- [33] D. Wang and Y. Shang, “A new active labeling method for deep learning,” in *2014 International Joint Conference on Neural Networks (IJCNN)*, 2014, pp. 112–119.
- [34] M. Contributors, “MMDetection3D: OpenMMLab next-generation platform for general 3D object detection,” <https://github.com/open-mmlab/mmdetection3d>, 2020.
- [35] 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, pp. 770–778.
- [36] 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, pp. 4193–4202.
- [37] 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, pp. 3351–3359.
- [38] 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, pp. 111–119.
- [39] 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*, vol. 35, no. 4, 2021, pp. 3101–3109.
- [40] J. Zhang and Y. Ding, “Occfusion: Depth estimation free multi-sensor fusion for 3d occupancy prediction,” *arXiv preprint arXiv:2403.05329*, 2024.
- [41] J. Pan, Z. Wang, and L. Wang, “Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction,” *arXiv preprint arXiv:2404.04561*, 2024.
- [42] G. Wang, Z. Wang, P. Tang, J. Zheng, X. Ren, B. Feng, and C. Ma, “Ocugen: Generative multi-modal 3d occupancy prediction for autonomous driving,” 2024.
- [43] 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, pp. 11 621–11 631.
- [44] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

## VI. APPENDIX

### A. Definition of 3D Occupancy Prediction Task

The occupancy prediction task aims at a semantic 3D occupancy grid with fixed given perception range and resolution. The sensor information is from surround-view cameras and 360 view LiDAR. Each voxel grid is represented by multiple attributes, including occupancy and semantics.

### B. Dataset and Benchmarks Introduction

**Occ3D-nuScenes** is built upon large-scale public available nuScenes dataset [43], which is a critical resource for advancing 3D object detection and occupancy technologies. It consists of 700 training scenes, 150 scenes for validation and 150 scenes for test, each meticulously annotated at a key frame rate of 2Hz. The sensor configuration of the ego vehicle is 6 ring cameras with resolution  $1600 \times 900$  and one 32-beam LiDAR on the top roof. There are 17 categories of semantics in occ3d-nuScenes including general object.

**Occ3D-Waymo** is built upon large-scale public available Waymo Open Dataset [44]. The dataset comprises 1,000 sequences for trainval split, among which 798 sequences are allocated for the training set and the remaining 202 sequences are designated for validation. Ground-truth labels are annotated at 10Hz, which is five times more than nuScenes samples. The sensor configuration of the ego vehicle is 5 ring cameras (The rear of the vehicle is not visible by any camera) with resolution  $1920 \times 1280$  or  $1920 \times 1080$  and five LiDARs (equivalent to high-beam LiDAR) on the top roof. There are 15 categories of semantics in occ3d-waymo including general object.

For both datasets, the Occ3D splits the surrounding world into 3D voxel grids with the resolution of  $[200, 200, 16]$ . The perception range is  $[-40m, -40m, -5m, 40m, 40m, 3m]$ . For both datasets, we use a voxel size of  $0.4m$  meters for voxelization following the prior works.

**OpenOccupancy-nuScenes** is also built on nuScenes [43] dataset and shares the same sensor suite with Occ3D-nuScenes. This benchmark is more challenging than Occ3D-nuScenes in that it requires larger perception range  $[-51.2m, -51.2m, -5m, 51.2m, 51.2m, 3m]$  and finer resolution  $[512, 512, 40]$  and the voxel size is  $0.2m$ . A slight difference regarding semantic categories is that OpenOccupancy ignore general object and only has 16 semantic categories.

**Metrics for 3D occupancy prediction.** We adopt the official evaluation protocol from CVPR2023 Occupancy Prediction Challenge<sup>1</sup>. We evaluate Intersection-over-Union (IoU) metric over categories and mean IoU. OpenOccupancy also evaluates geometric IoU of occupied voxels as one of the primary metrics. Let  $C$  be the number of classes, where  $TP_c$ ,  $FP_c$  and  $FN_c$  correspond to the number of true positive, false positive, and false negative predictions for class  $c_i$ .

$$mIoU = \frac{1}{C} \sum_{c=1}^C \frac{TP_c}{TP_c + FP_c + FN_c} \quad (5)$$

### C. Visualizations

We visualize 3d occupancy prediction on Occ3D-nuScenes benchmark validation set based on the visualization tools provided by FlashOcc [4]. For the sake of rendering speed, we only plot predictions that are within camera visibility mask of each frame.

We visualize 3d occupancy prediction performance on nuScenes validation set as demo videos on <https://github.com/synsin0/EFFOcc>.

<sup>1</sup>[https://opendrivelab.com/challenge2023/#3d\\_occupancy\\_prediction](https://opendrivelab.com/challenge2023/#3d_occupancy_prediction)

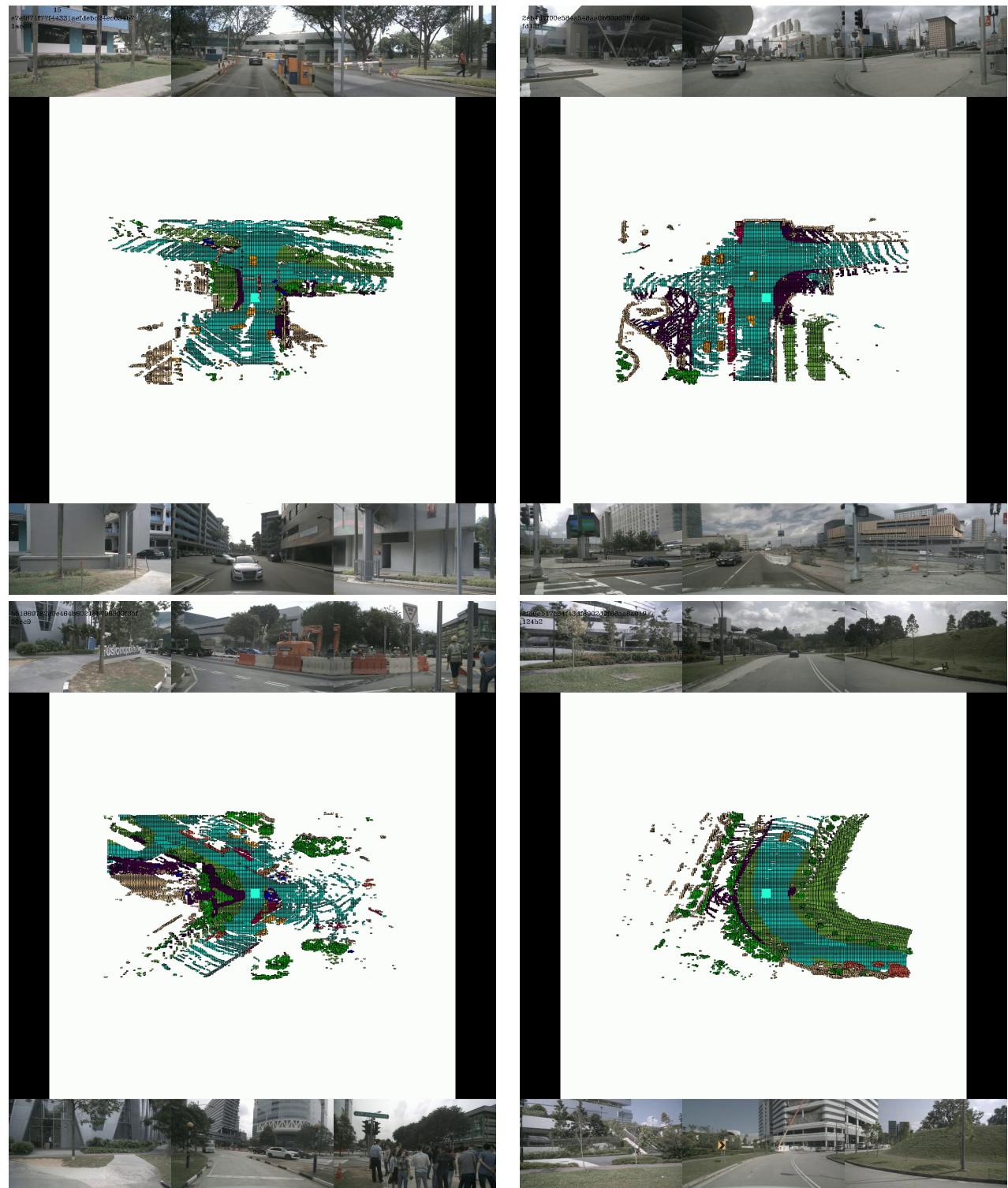


Fig. 4. Visualizations of occupancy prediction results on Occ3D-nuScenes validation set.