2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Vision-Based Panoptic Occupancy Prediction in Urban Environments

Rodrigo Marcuzzi Lucas Nunes Elias Marks Xingguang Zhong Jens Behley Cyrill Stachniss

***Abstract*— Understanding the surrounding scene geometrically and semantically is a key requirement for autonomously navigating systems. Vision-based 3D panoptic occupancy prediction aims to provide a 3D representation of the surroundings including semantic meaning and identifying individual objects such as traffic participants in the context of urban navigation. The majority of vision-based approaches to occupancy prediction require 3D voxel labels or segmented LiDAR scans as supervision signal. While other vision-based approaches use only a few consecutive images for supervision, these approaches typically do not provide instance-level information, which is crucial for achieving a holistic understanding of the scene. In this paper, we propose a novel method for 3D panoptic occupancy prediction that relies solely on image data for both training and inference. We use bundle adjustment to align all available images in the training set to obtain depth information. We further use a pre-trained open-vocabulary image model to obtain panoptic segmentation of the RGB images and generate occupancy pseudo labels to directly optimize for the 3D panoptic occupancy prediction task. Furthermore, we use a 3D foundation model to obtain depth predictions for individual images to add dynamic objects into the pseudo labels. Without any manual or LiDAR-based annotations, our approach outputs occupancy, semantic class, and instance ID for each 3D voxel in the full voxel grid. We achieve state-of-the-art results on 3D semantic occupancy prediction among label-free methods, and we propose the first method for 3D panoptic occupancy without any LiDAR supervision.**

I. Introduction

One of the key requirements for safe outdoor navigation is scene understanding, which includes 3D geometric and semantic information of the scene. To achieve such 3D understanding, several perception systems rely on 3D sensors like LiDAR. Besides that, vision-centric approaches to scene understanding aim to tackle scene understanding by relying only on RGB cameras. 3D occupancy prediction is a recent vision-centric task that aims to represent the 3D geometry of the surroundings and provide semantic and instance information using only a setup of cameras covering the surroundings of the vehicle.

State-of-the-art methods for 3D panoptic occupancy prediction [21], [35], [38] assume the availability of 3D voxel labels for supervision, which are challenging and expensive to obtain. This is also the case for a large number of methods that focus on 3D semantic occupancy prediction [14], [31], [41], also relying on 3D voxel labels. Other methods [9], [26] rely on more costly 3D sensors like LiDAR and use segmented scans for training. To cope with the sparsity of the scans, they aggregate multiple point clouds for better supervision.

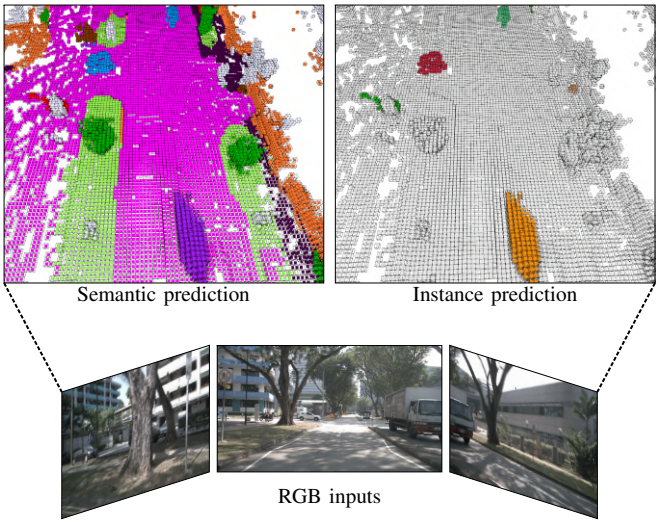

Fig. 1: Our approach allows us to predict, from a set of RGB images, the occupancy of the surrounding scene along with the semantic classes and instances. We show with different colors the semantic classes on the left and the instance IDs on the right. See Fig. 6 for the legend of class colors.

To avoid relying on 3D sensors, recent research [9], [10], [13], [39] focuses on using only RGB images for training. They use a few consecutive images from the same camera to enforce multi-view consistency and train a depth estimator. This means that they only provide information up to the first surface (and thus not for occluded areas), and they supervise for proxy tasks such as volume rendering [25]. A recent method [24] proposes to leverage the available training images and generate semantic occupancy pseudo labels for explicit supervision. This method, however, only provides semantic information and does not consider moving objects in the pseudo label generation.

In this paper, we tackle the task of 3D panoptic occupancy prediction using only RGB images for training, and also want to deal with dynamic objects. We provide explicit supervision while relying only on image information to predict the geometry of the whole scene alongside its semantic and instance segmentation, as shown in Fig. 1.

The main contribution of this paper is a novel method to

All authors are with the Center for Robotics, University of Bonn, Germany. Cyrill Stachniss is additionally with the Department of Engineering Science at the University of Oxford, UK, and with the Lamarr Institute for Machine Learning and Artificial Intelligence, Germany.

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob, and by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG). We thank for the access to the Marvin cluster of the University of Bonn.

generate panoptic occupancy pseudo labels relying on image data and foundation models, which we can use to supervise neural networks for 3D panoptic occupancy. We leverage all images from the training set and run bundle adjustment to estimate camera poses and compute depth images. However, in contrast to prior work [24], we use a foundation model and leverage the *panoptic* segmentation of each RGB image and combine them with the depth images to generate sparse *panoptic* occupancy pseudo labels, which we can use as explicit supervision. Furthermore, we additionally use a 3D foundation model to obtain dense depth predictions and combine them with the panoptic predictions to detect and add dynamic objects into the pseudo labels. This way, our model is able to predict not only semantics but can also predict instances and reconstruct dynamic objects in the scene. To the best of our knowledge, we provide the first method for 3D panoptic occupancy prediction trained without LiDAR supervision, and achieve state-of-the-art performance for vision-based 3D semantic occupancy prediction.

In summary, we make three key claims: (i) we present the first image-based approach for 3D panoptic occupancy prediction without 3D data for supervision, (ii) our approach achieves state-of-the-art performance on 3D semantic occupancy prediction among methods using only images for training, and (iii) our approach is able to predict dynamic objects in the scene.

These claims are backed up by the paper and our experimental evaluation. The implementation of our approach and the generated pseudo labels are available at https://github.com/PRBonn/SfmPanOcc.

II. Related Work

3D Semantic Occupancy Prediction aims to provide geometric and semantic information of the scene by dividing it into a voxel grid and predicting the state of each voxel. A similar task named 3D semantic scene completion was first introduced as a benchmark in the SemanticKITTI [1] dataset using a LIDAR and several works [17], [43] focus on completing the scene using LiDAR information. Different datasets for this task [18], [31] provide 3D semantic occupancy ground-truth by aggregating LiDAR scans, where the input used for the task is the captured surrounding views.

Most methods [14], [31], [41] follow a common configuration that includes an image feature extractor, a lifting operation to project features to 3D and a final 3D neural network to obtain occupancy and semantics. State-of-the-art approaches [14], [31], [37], [41] rely on manually annotated 3D voxel ground-truth for supervision. Recent works [3], [9], [26], [27] rely on segmented LiDAR scans, which they project into 2D to provide supervision.

To drop the requirement for 3D sensors and manual annotations, other approaches [9], [10], [13], [39] rely only on unlabeled RGB images and use a few consecutive frames and optimize for a proxy task such as volume rendering [25] to implicitly learn occupancy. To predict also semantics, these methods rely on foundation models [29] to semantically segment the RGB images and use the predictions as supervision during training. Recently, Marcuzzi et al. [24] proposed to use bundle adjustment to obtain scale-aware depth images and combine them with image semantics to generate semantic occupancy pseudo labels. We follow this approach to obtain pseudo labels using only RGB images, but also exploit instance information, as well as moving objects.

3D Panoptic Occupancy Prediction aims to provide not only geometric and semantic information but also to identify individual instances. Wang et al. [35] propose to jointly perform occupancy prediction, semantic segmentation, and object detection based only on RGB images. They learn occupancy in a coarse-to-fine manner and use voxel queries to perform object detection. Liu et al. [21] propose a sparse architecture that iteratively prunes empty voxels to predict occupancy in a sparse way. In a following step, they perform semantic and instance segmentation of the occupied voxels using sparse queries and a transformer decoder. Chen et al. [6] tackle simultaneously panoptic occupancy and tracking by predicting consistent instance IDs over time. Yu et al. [38] first predict occupancy leveraging Flash-Occ [37] for its efficiency, and predict instance offsets and heatmaps, and use clustering to group voxels into instances in a bottom-up fashion. We follow a similar bottom-up approach and use three heads to predict occupancy values, semantic classes, and instance offsets. To obtain instances, we shift voxel centers using the predicted offsets and group voxels using a clustering algorithm. Previous methods [37], [38] use voxel-level ground-truth for supervision, while our approach relies only on RGB images.

Depth Estimation aims to provide dense per-pixel depth predictions. While early approaches [20], [42] rely on dense depth annotations, more recent works [22], [30], [32] follow a self-supervised training paradigm. To adapt to multi-camera systems, research focused on predicting depth for these surrounding views [11], [33], [36]. SurroundDepth [36] uses structure-from-motion to generate sparse pseudo depths that are scale-aware to pretrain their model. We also use bundle adjustment, but use all available images for a scene to generate sparse depth images.

Recently, 3D foundation models [16], [34], [40] have demonstrated impressive performance. They take two input images and output dense 3D point clouds, confidence maps, depth images, and camera poses in a feed-forward manner. We use such a 3D foundation model [34] to obtain dense depth predictions and use them to detect and add dynamic objects to our pseudo labels.

III. Our Approach to 3D Panoptic Occupancy Prediction

A. Overview

The goal is to predict a semantic class and instance ID for each voxel in a dense 3D voxel grid $\tilde{\boldsymbol{O}} \in \mathbb{Z}^{S_x \times S_y \times S_z \times 2}$, where S_x, S_y, S_z is the volume size in x, y, and z direction, and we predict two values for each voxel. One value corresponds to the semantic class $c \in \{1, \dots, C\}$, where C is the number of semantic classes (including the "empty" class). This value indicates the voxel's occupancy and its semantic

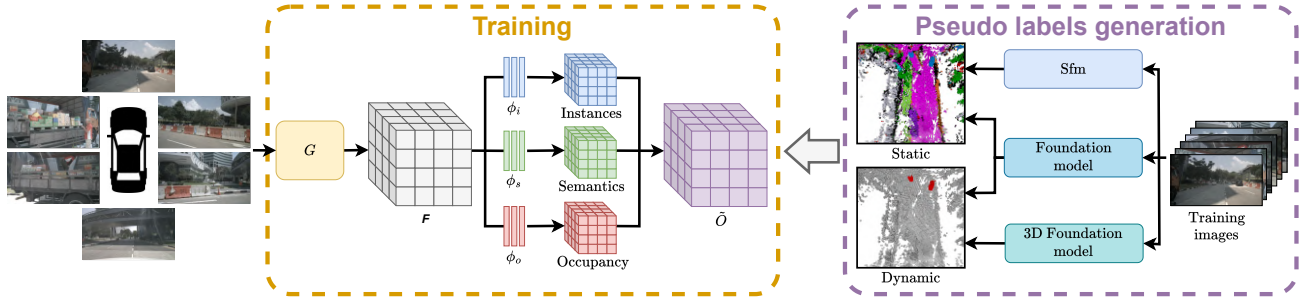

Fig. 2: Overview of our method. At each timestep, the multi-camera setup captures a set of input RGB images from which we extract 3D voxel features $\boldsymbol{F}$. Using three MLP heads ϕ_o, ϕ_s, and ϕ_i we predict an occupancy value, semantic class, and instance ID for each voxel, which we fuse to obtain the final voxel grid $\tilde{\boldsymbol{O}}$. We use all available training images and use structure-from-motion (sfm) and volume reconstruction to generate occupancy pseudo labels for the static scene. We leverage a 3D foundation model to add dynamic objects to the pseudo labels. We use a foundation model to segment the images and add the semantic class and instance ID to each voxel. We explicitly supervise for panoptic occupancy prediction while relying only on camera data.

class, which can belong to both, "stuff" or "things" [15]. The other predicted value is the instance ID, only valid for voxels with a predicted "thing" class. This way, we predict the geometry, semantics, and instances of the surrounding scene. Fig. 2 shows the overview of our approach.

We use as input RGB images $\mathcal{I} = \{I_{1,t}, \dots, I_{N,t}\}$ captured at timestep t by a sensor setup consisting of N cameras. We employ a network G to extract 3D voxel features $\boldsymbol{F} \in \mathbb{R}^{S_x \times S_y \times S_z \times D}$ from the input images $\mathcal{I}$, where D is the feature dimension.

For each voxel, we predict occupancy probabilities, semantic logits, and offsets to the instance center by applying three different MLP heads ϕ_s, ϕ_o, and ϕ_i over the voxel features $\boldsymbol{F}$. We threshold the occupancy probabilities to obtain occupancy values for each voxel and use $\arg\max$ over the logits to obtain semantic classes for the occupied voxels. To obtain instance IDs, we follow a bottom-up approach by filtering voxels with predicted "stuff" class, shifting the 3D coordinates of the remaining voxels using the predicted offsets and using a clustering algorithm [7] to group voxels into instances. Note that the proposed method can use different networks G to extract 2D features and project them to 3D via attention [14] or via depth prediction [26].

To train the approach, we obtain pseudo labels for each scene using all the images captured by the N cameras at all M timesteps. We first use bundle adjustment to align the $N \cdot M$ images and then generate depth images (Sec. III-B), with which we perform volume reconstruction and obtain occupancy pseudo labels. We use a foundation model [29] to segment the 2D images and obtain semantic classes and instance IDs. We utilize this information to generate *panoptic* occupancy pseudo labels (Sec. III-C). Additionally, we use a 3D foundation model to obtain dense depth images and leverage them to detect and add dynamic objects into our pseudo labels (Sec. III-D).

B. Depth Image Generation

To generate our occupancy pseudo labels, we first need to compute depth images. We use all RGB images captured by the N cameras at all M timesteps for a given scene and use photogrametric scene reconstruction, including bundle adjustment to align these RGB images $\{\{I_{1,1}, \dots, I_{N,1}\}, \dots, \{I_{1,M}, \dots, I_{N,M}\}\}$. We assume rigidly attached cameras and constant intrinsic parameters $\boldsymbol{K}_i$ of each camera i. With the obtained camera poses for each camera at each given timestep in the scene, we use multi-view stereo reconstruction [8], [28] to generate a sparse point cloud that we project into each camera i using the intrinsic parameters $\boldsymbol{K}_i$ to obtain depth images. We generate a triangle mesh and use it for depth filtering. Given the extrinsic parameters provided by the dataset, this allows us to obtain, for each camera i at timestep t, a scale-aware sparse depth image $\boldsymbol{D}^f_{i,t}$.

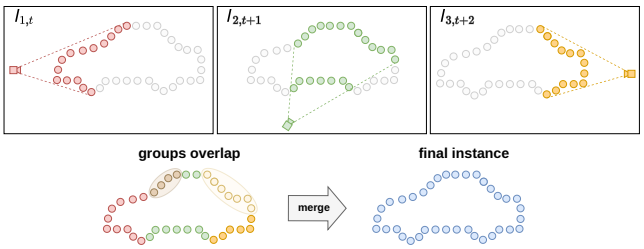

Fig. 3: Instance merging process. We project the depth images from the different cameras and assign instance IDs to the projected points (top). Because IDs are not consistent across cameras, a single 3D instance has points with multiple IDs. We merge all groups of points based on their overlap and the distance between their centroids to obtain an instance with a single ID (bottom).

C. Panoptic Occupancy Pseudo Labels

We use the sequence of images $\{I_{i,1}, \dots, I_{i,M}\}$ recorded by each camera i as the input to a foundation model. Similar to previous works [24], [39], we choose Grounded SAM [29] to obtain the semantic and instance segmentation for each image, where the instance IDs are consistent over time. This allows us to obtain panoptic segmentation for the M images captured by each camera. We project the depth images to 3D using the camera parameters $\boldsymbol{K}_i$ to obtain point clouds and assign to each point, the semantic class and an instance ID from the panoptic segmentation prediction, as shown in Fig. 3 (top). We do the same for the N cameras and aggregate all the point clouds from all $N \cdot M$ images in the sequence to obtain a single point cloud $\mathcal{P} = \{\boldsymbol{p}_1, \dots, \boldsymbol{p}_J\}$

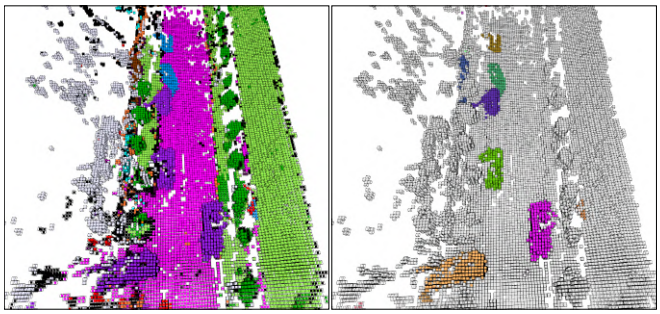

Fig. 4: Generated panoptic occupancy pseudo labels. We show the semantic classes for each voxel with different colors on the left. On the right, we show with colors the different instance IDs.

where each point $\boldsymbol{p}_j = (\boldsymbol{x}_j, l_j)$ consists of 3D coordinates $\boldsymbol{x}_j$ and instance ID l_j and J is the total number of points.

The instance IDs of points belonging to the same instance might be different for points projected from different cameras, as shown in Fig. 3 (bottom). To obtain a single instance ID, we merge groups of points based on their overlap. Given a group of points with the same instance ID $\mathcal{P}_l = \{\boldsymbol{p}_j \in \mathcal{P} \mid l_j = l\}$, we count for each point $\boldsymbol{p}_j \in \mathcal{P}_l$ the number of neighbors (nn) in the other group:

$$\text{nn}(\mathcal{P}_l, \mathcal{P}_{l'}) = \left|\left\{\boldsymbol{p}_j \in \mathcal{P}_l \mid \exists\, \boldsymbol{p}_k \in \mathcal{P}_{l'} \text{ s.t. } \|\boldsymbol{x}_j - \boldsymbol{x}_k\|_2 < \tau_n\right\}\right|. \quad (1)$$

We define the symmetric overlap as:

$$\text{overlap}(\mathcal{P}_l, \mathcal{P}_{l'}) = \frac{\text{nn}(\mathcal{P}_l, \mathcal{P}_{l'}) + \text{nn}(\mathcal{P}'_l, \mathcal{P}_l)}{|\mathcal{P}_l| + |\mathcal{P}_{l'}|}. \quad (2)$$

We merge groups $\mathcal{P}_l$ and $\mathcal{P}_{l'}$ if the symmetric overlap between them is larger than a threshold τ_o. This allows us to remap the instance IDs in each image to be consistent across all cameras and timesteps in the scene.

To generate pseudo labels, we build a global voxel grid and perform a similar operation to occupancy mapping [12]. We perform a ray-casting operation to determine the voxels along each ray of each camera and use a 3D variant of the Bresenham's algorithm [4] to step through the voxel grid from the camera center to the endpoint of the ray. We set all traversed voxels as "empty" and the voxel at the end of the ray as "occupied". We accumulate in a histogram the different semantic classes and instance IDs for each voxel and obtain the final values via majority voting. As a result, we obtain panoptic occupancy pseudo labels from images, as shown in Fig. 4, relying on a foundation model.

D. Pseudo Labels for Dynamic Objects

Given the static scene assumption made in the bundle adjustment step, the depth images $\boldsymbol{D}^f_{i,t}$ do not contain depth for pixels belonging to dynamic objects due to the outlier rejection, and therefore, dynamic objects are not present in the pseudo labels.

To predict depth also for moving objects, we use the 3D foundation model MonST3R [40] for depth prediction, and input the sequences of images $\{I_{i,1}, \ldots, I_{i,M}\}$ captured by each camera i. The output are dense up-to-scale depth predictions $\boldsymbol{D}^p_{i,t}$ and confidence maps $\boldsymbol{C}^p_{i,t}$. These depth images $\boldsymbol{D}^p_{i,t}$ include artifacts and are noisier than $\boldsymbol{D}^f_{i,t}$ due to the low resolution of the predictions. However, they contain depth for the dynamic objects, as shown in Fig. 5, which we can use to add moving objects into our pseudo labels.

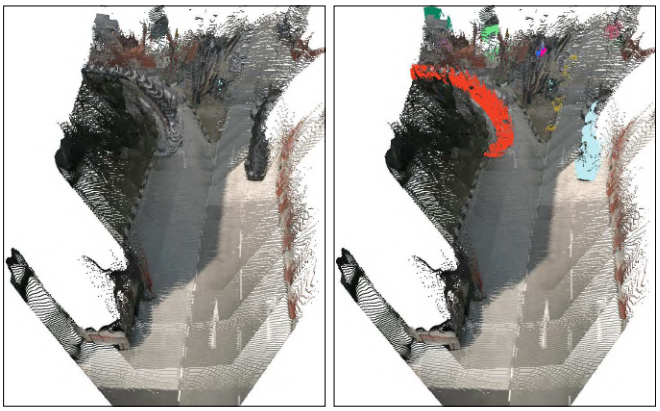

Fig. 5: Depth predictions using the 3D foundation model and dynamic objects detection. We show the projected point clouds from the depth predictions for a sequence of images. On the left, we show the aggregated point cloud of the scene from above. On the right, we further highlight the detected dynamic objects.

We scale the depth images $\boldsymbol{D}^p_{i,t}$ by estimating a scale factor s^* and scale bias b^* using the depth from structure-from-motion $\boldsymbol{D}^f_{i,t}$ given by:

$$s^*, b^* = \arg\min_{s,b} \boldsymbol{M}^p_{i,t} \cdot \boldsymbol{M}^f_{i,t} \| \boldsymbol{D}^p_{i,t} - s\boldsymbol{D}^f_{i,t} + b \|_2, \quad (3)$$

where $\boldsymbol{M}^p_{i,t} = \mathbb{I}[\boldsymbol{C}^p_{i,t} > \tau_c]$ is the binary mask of pixels with confidence larger than τ_c, and $\boldsymbol{M}^f_{i,t} = \mathbb{I}[\boldsymbol{D}^f_{i,t} > 0]$ is the valid sparse depth mask. In this setup, Eq. (3) represents a least squares problem that can be solved in closed form.

For each input image, we obtain the scaled depth prediction $\boldsymbol{D}^{p'}_{i,t} = s^* \boldsymbol{D}^p_{i,t} + b^*$ and use the panoptic segmentation to filter out "stuff" pixels. We project $\boldsymbol{D}^{p'}_{i,t}$ into a point cloud where each point has an instance ID. For each group of points with the same instance ID $\mathcal{P}_l = \{\boldsymbol{p}_j \in \mathcal{P} \mid l_j = l\}$, we compute its centroid $\boldsymbol{c}_{l,t} \in \mathbb{R}^3$ at each time step t as the mean of the 3D point coordinates.

After processing all M timesteps, we classify instances as moving if the centroid displacement exceeds a threshold τ_m:

$$\max_t \|\boldsymbol{c}_{l,t} - \boldsymbol{c}_{l,t-1}\|_2 > \tau_m. \quad (4)$$

This way, we detect moving objects in each individual image and use this information together with the scaled depth prediction $\boldsymbol{D}^{p'}_{i,t}$ to add dynamic objects to the panoptic occupancy pseudo labels at each individual timestep as described in Sec. III-C. With this procedure, we add dynamic objects into our generated pseudo labels using image data and relying on a 3D foundation model for depth prediction.

IV. Experimental Evaluation

The main focus of this work is a method for 3D panoptic occupancy prediction, trained without any LiDAR supervision. We generate sparse pseudo labels using all training images and a foundation model to add semantics and instances. With a 3D foundation model, we detect and add dynamics to the pseudo labels. We present our experiments to evaluate our

method and support our key claims: (i) we present the first image-based approach for 3D panoptic occupancy prediction without 3D data for supervision, (ii) our approach achieves state-of-the-art performance on 3D semantic occupancy prediction among methods using only images for training, and (iii) our approach is able to predict dynamic objects in the scene.

A. Implementation Details

The voxel grid has a size $S_x = 200, S_y = 200, S_z = 16$ with voxel size $\{0.4 \times 0.4 \times 0.4\}\, m^3$. We use BEVStereo [19] as the network G to extract the 3D voxel features F and only add the occupancy, semantic, and instance heads. The occupancy and semantic heads ϕ_o and ϕ_s consist of a single linear layer and softplus activation function. The instance head ϕ_i consists of a stack of 3D convolutions and ReLU activation functions followed by a linear layer. We use AdamW [23] optimizer with learning rate of 10^{-4} and train for 12 epochs with batch size 8. We use the cross entropy loss between predictions and pseudo labels for occupancy and semantics, and L2 loss for the offset predictions. We conduct all experiments on 8 NVIDIA A40 GPUs.

We use Grounded SAM [29] to obtain semantic maps and consistent instance IDs for the whole sequence. We follow OccNerf [39] and prompt the model with synonyms of the class names of the nuScenes dataset [5]. We use $\tau_n = 0.1\, m$ to count neighbors, and $\tau_o = 0.1$ for the overlap in the label generation. We obtain dense depth predictions for the dynamic objects using MonST3R [40] with the default parameters. We use $\tau_c = 0.5$ to filter out points with low confidence and $\tau_m = 1\, m$ to detect moving objects.

B. Experimental Setup

To evaluate our model, we use the Occ3D-nuScenes dataset [31], which builds on top of nuScenes [5] and provides 3D voxel labels for semantic occupancy prediction. nuScenes consists of 1000 driving scenes captured by six cameras covering the complete 360° of the vehicle. The occupancy ground-truth covers a range of $[-40, 40]\, m$ in x and y direction and $[-1, 5.4]\, m$ in z direction with a voxel size of $\{0.4 \times 0.4 \times 0.4\}\, m^3$ and containing the 16 semantic classes of nuScenes plus an extra "empty" class. We obtain voxel-level panoptic labels, following SparseOcc [21] to combine the bounding boxes provided by nuScenes and the voxel-level semantic labels. We evaluate the occupancy performance using IoU and consider also the semantics by computing the mIoU across all classes. For panoptic segmentation performance, we use voxel-level panoptic quality (PQ) [15]. Given that Occ3D-nuScenes only provides labels for two sets: training and validation sets of nuScenes. We use the split suggested in [24] with 60 scenes separated from the training set to use as our validation set. This allows us to run experiments in 60 scenes and evaluate performance in 150 scenes, therefore using different subsets of data.

C. Pseudo Labels

We generate pseudo labels to train our approach for panoptic occupancy prediction and therefore use only images from the training set of nuScenes. For some scenes, bundle adjustment fails to align the images due to several reasons, like too dark scenes, scenes with too many dynamic objects or sequences where the vehicle does not move enough. In total, we generate depth images and pseudo labels for 584 out of the 700 scenes in the training set. We produce 127,458 depth images with which we generate 21,231 sparse panoptic occupancy pseudo labels with the same voxel grid and range as Occ3D-nuScenes. Our generated labels contain the 17 classes included in Occ3D-nuScenes and an extra "uncertain" class for pixels for which no semantic class was given by the pre-trained open-vocabulary segmentation model.

Method	Supervision	IoU [%]	mIoU [%]	PQ [%]
SparseOcc [21]	3D	37.7	30.9	13.0
Panoptic-FlashOcc [38]	3D	43.3	**31.6**	**15.8**
Ours	C	**58.2**	**17.9**	**11.5**

TABLE I: 3D *panoptic* occupancy prediction performance on occ3D-nuScenes. In the column mode: 3D are methods trained with occupancy ground truth labels, and C trained only with cameras.

D. 3D Panoptic Occupancy Prediction Performance

In the first experiment, we evaluate our approach for 3D panoptic occupancy prediction and compare it with previous approaches in Tab. I. Given that we propose the first method that does not rely on LiDAR supervision, we compare it against existing state-of-the-art methods that rely on ground-truth voxel labels, namely SparseOcc [21] and Panoptic-FlashOcc [38]. Our approach achieves 58.2% IoU, better than previous methods that even rely on voxel labels. Regarding the semantic segmentation performance, our approach achieves 17.9% compared with 31.6% of the state-of-the-art approach, but without manual annotations for semantic classes. In terms of panoptic quality, our approach achieves 11.5% against the 15.8% of the best performing method, see Tab. I. Overall, our method shows strong performance given the fact that it is not trained with ground truth depth, occupancy, instances, or semantic classes. We observe that, when training with pseudo labels, the model seems to concentrate on the reconstruction in the visible area, which leads to better IoU compared to other methods.

E. 3D Semantic Occupancy Prediction Performance

In the second experiment, we evaluate the performance of our approach in the task of 3D semantic occupancy prediction. We show the results in Tab. II, where we compare against methods that supervise using ground-truth 3D voxel labels (3D), LiDAR (L), and only camera data (C). Our approach outperforms previous approaches that rely only on camera data, both in terms of IoU and mIoU, even outperforming a method using LiDAR for supervision. While relying only on images, our approach is able to better learn the geometry of the scene, depicted by the high IoU, compared to the previous camera-based methods, but has a lower performance in terms of mIoU compared with methods that use ground-truth semantics. We believe that the reason for this drop in performance might be the wrong semantic classes predicted by the foundation model, and later assigned

Method	Supervision	GT Sem.	IoU [%]	mIoU [%]
OccFormer [41]	3D	✓	-	21.9
TPVFormer [14]	3D	✓	-	27.8
CTF-Occ [31]	3D	✓	-	**28.5**
TPVFormer [14]	L	✓	17.2	13.6
RenderOcc [26]	L	✓	**45.9**	23.9
OccFlowNet [3]	L	✓	-	**26.1**
SimpleOcc [10]	C		-	7.1
SelfOcc [13]	C		45.0	9.3
OccNeRF [39]	C		45.0	9.5
GaussianOcc [9]	C		-	9.9
LangOcc [2]	C		51.8	11.8
SfmOcc [24]	C		57.7	17.7
Ours	C		**59.2**	**18.4**

TABLE II: 3D *semantic* occupancy prediction performance on occ3D-nuScenes. In the column supervision: 3D are methods trained with occupancy ground truth labels, L trained with LiDAR supervision, and C trained only with cameras. GT Sem. indicates the usage of ground-truth semantic labels for supervision.

Method	mIoU	bicycle	bus	car	cons. veh.	motorcycle	pedestrian	trailer	truck
SelfOcc [13]	10.5	0.1	6.6	13.2	0.0	0.4	2.4	0.0	7.7
GaussianOcc [9]	11.0	3.8	14.6	17.2	0.8	2.9	10.1	0.14	10.6
SfmOcc [24]	19.6	**8.2**	14.8	20.9	**7.1**	12.0	12.2	**1.6**	**15.9**
Ours	**20.3**	7.2	**25.5**	**23.6**	6.2	**15.0**	**13.3**	0.8	15.0

TABLE III: 3D semantic occupancy prediction performance on scenes with many dynamic objects in Occ3D-nuScenes.

to our pseudo labels. Despite the small improvement in semantic segmentation, the model also predicts instances, which enables a more holistic scene understanding.

F. Performance in Scenes with Dynamic Objects

In the third experiment, we evaluate the performance of our method in scenes with many dynamic objects. In our method, we generate our occupancy pseudo labels in two steps. First, we use bundle adjustment assuming a static scene and do not reconstruct dynamic objects. Second, we use a 3D foundation model to obtain dense depth predictions and add dynamic objects to the pseudo labels. To show the advantage of our approach, we evaluate the semantic occupancy prediction performance of image-based methods on 30 scenes with many dynamic objects, as done by SfmOcc. We show the results in Tab. III. Previous methods either supervise using a few consecutive images [9], [13] or all the available images in the scene [24], relying on multi-view geometry. Therefore, their supervision does not consider dynamic objects. Our approach, on the other hand, includes dynamic objects and outperforms all previous methods in terms of mIoU. Compared with SfmOcc, which provides labels only for static objects, our approach improves the performance for commonly seen semantic classes such as bus, car, and pedestrian, but slightly decreases the performance for the classes that are easier to confuse, such as construction vehicles, trailers, and trucks. We argue that the problem lies in the semantic classes predicted by the foundation model. For the static parts, we obtain the class via majority voting over the semantic maps for multiple images and even from different cameras. In contrast, we rely on a single image and its semantic map to add moving objects, which is more susceptible to wrong semantic predictions.

#	Depth images	IoU [%]	mIoU [%]
A	Sfm	**68.3**	20.5
B	MonST3R (static)	51.4	11.0
C	MonST3R (static) + dynamics	52.7	11.4
D	Sfm + dynamics	66.5	**21.2**

TABLE IV: Comparison of generated labels using different depth images from (A) bundle adjustment, (B) MonST3R considering only the static parts, (C) MonST3R static scene and adding dynamics, and (D) static part from bundle adjustment and dynamics

#	Labels	IoU [%]	mIoU [%]	PQ
E	Pseudo labels	**66.5**	21.2	15.0
F	Ground truth	61.4	**48.9**	**33.4**

TABLE V: Performance comparison when training with ground truth labels vs. our generated pseudo labels on the validation set.

G. Ablation Studies

In this section, we conduct experiments to evaluate how different strategies to generate pseudo labels impact the performance. We compare to training with ground truth labels, and also evaluate the generated pseudo labels against the ground truth labels. We evaluate our model in the 60 scenes that we separate as the validation set in Tab. IV and Tab. V.

1) Different Depth Images: We train our model with labels generated using different depth images and show the semantic occupancy prediction performance in Tab. IV on the validation set.

In setup [A], we use the pseudo labels generated using the depth images from bundle adjustment only. The achieved performance is 68.3% IoU and 20.5% mIoU. In our approach, we use a 3D foundation model to obtain dense depth predictions that are up-to-scale. As another option, we scale them with the depth images from bundle adjustment as described in Sec. III-D and generate occupancy pseudo labels. However, using these depth images would generate a trace of moving objects as shown in Fig. 5. Therefore, in setup [B], we first detect dynamic objects as described in Sec. III-D and remove them to consider only the static background. The achieved performance is 51.4% IoU and 11.0% mIoU, which shows that although the 3D foundation model provides denser depth predictions, the quality of the bundle adjustment images is better suited to generate pseudo labels. In setup [C], we add the dynamic objects into the pseudo labels of setup [B] and increase the IoU and mIoU by 1.3 and 0.4 percent points respectively, which shows the advantage of adding moving objects. Finally, in setup [D], we combine the static scene from bundle adjustment and the dynamic objects from the 3D foundation model prediction to generate pseudo labels. The performance is 66.5% IoU and 21.2%mIoU, which shows how combining the best of both methods to generate pseudo labels leads to better learning the semantics of the scene. This setup shows a drop in IoU compared to [A], likely because the model is now optimized for an extra objective.

2) Comparison with Ground Truth Labels for Training: Here, we compare the performance of our model trained with

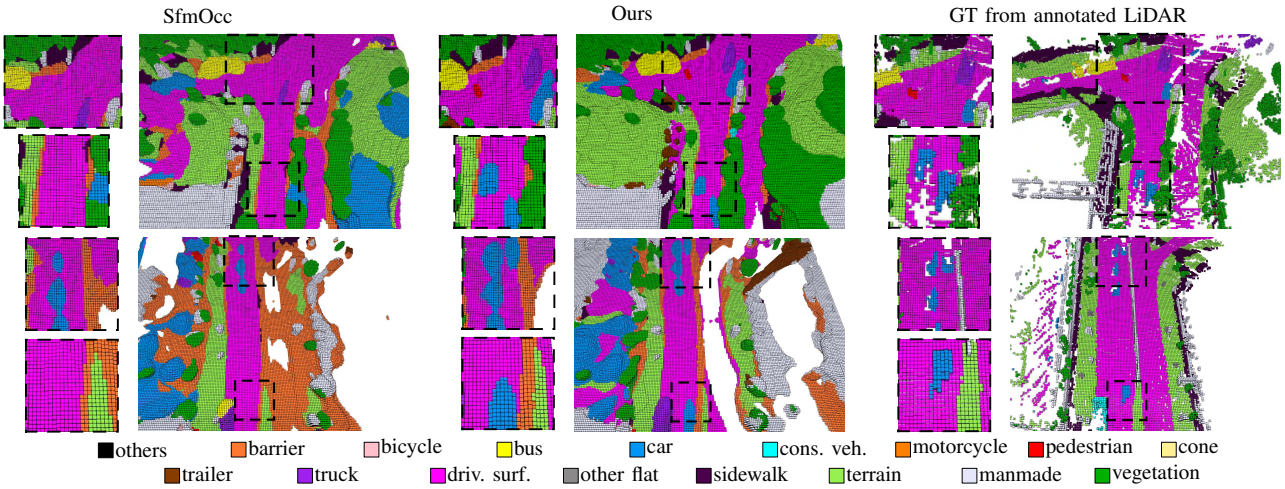

Fig. 6: Qualitative results of our approach against SfmOcc [24], the best previous approach using only images. We highlight areas containing dynamic objects. SfmOcc does not include dynamic objects in the supervision, and it is able to predict only a few of them. In comparison, our approach is better able to predict dynamic objects, showing the advantage of our method to generate pseudo labels.

ground truth labels vs. our generated pseudo labels and show the performance in Tab. V. The model trained with pseudo labels [E] achieves a better occupancy prediction in terms of IoU. However, the difference in performance of 27.7 and 18.4 percent points in mIoU and PQ in [F] shows that despite relying only on image information, the quality of our pseudo labels is not enough to close the gap with methods trained with ground truth labels.

3) Evaluation of the Pseudo Labels: To show the quality of our generated panoptic occupancy pseudo labels, we compute the IoU, mIoU, and PQ to compare our pseudo labels with the ground truth labels for the training set of Occ3D-nuScenes. Our pseudo labels achieve 51.8% IoU, 15.6% mIoU, and 9.1% PQ. This shows the limitations in the generated pseudo labels, mainly in the semantics, due to the segmentation from the foundation model.

H. Qualitative Results

We qualitatively evaluate how our approach can reconstruct dynamic objects and detect individual instances. In Fig. 6, we compare the results of our approach to SfmOcc [24], the closest method to ours. Due to the supervision of voxels in the whole scene, our approach can reconstruct the whole shape of objects, including occluded voxels. Different from SfmOcc, our method also predicts dynamic objects in the scene, as highlighted in the images. Furthermore, we also identify individual instances, see Fig. 7. We do not use ground-truth semantics but rely on a foundation model; therefore, our method sometimes predicts wrong semantics, which can lead to wrong instance predictions.

In summary, our experiments show that relying only on images and foundation models, our method is the first to perform 3D panoptic occupancy estimation, being able to reconstruct the scene and predict semantic classes and instances. Furthermore, our approach achieves state-of-the-art performance on semantic occupancy prediction among methods using only camera data for supervision. With the inclusion of dynamic objects in our pseudo labels, our model is able to predict not only static but also moving objects.

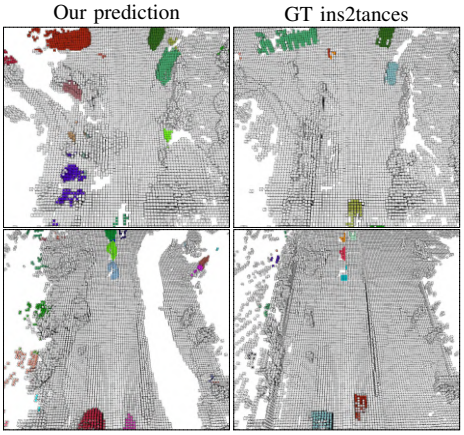

Fig. 7: Instance prediction results. We compare the ground truth and our instance predictions. Although our approach is able to differentiate between different instances, wrong semantic class predictions can lead to wrong instance predictions.

V. Conclusion

In this paper, we presented a novel approach to generate pseudo labels for 3D panoptic occupancy prediction relying on camera data and foundation models for image segmentation and depth prediction. This yields, to the best of our knowledge, the first model for 3D panoptic occupancy prediction trained without any LiDAR supervision.

We evaluated our approach and provided comparisons with existing techniques while supporting all claims made in this paper. Our model achieves state-of-the-art performance in 3D semantic occupancy prediction among methods that rely only on images for supervision and is able to predict dynamic objects in the scene. Our experiments suggest that we can generate pseudo labels by leveraging all the available images and adding dynamic objects using a 3D foundation model. We can use those labels to train our model for 3D panoptic occupancy prediction and estimate occupancy, semantics, and instances, also for dynamic objects in the scene.

REFERENCES

[1] 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 *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2019.

[2] S. Boeder, F. Gigengack, and B. Risse. LangOcc: Open Vocabulary Occupancy Estimation via Volume Rendering. In *Proc. of the Intl. Conf. on 3D Vision (3DV)*, 2025.

[3] S. Boeder and B. Risse. OccFlowNet: Occupancy Estimation via Differentiable Rendering and Occupancy Flow. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2025.

[4] J.E. Bresenham. Algorithm for computer control of a digital plotter. In *Seminal graphics: pioneering efforts that shaped the field*, pages 1–6. 1998.

[5] 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 *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[6] Z. Chen, K. Li, X. Yang, T. Jiang, Y. Li, and H. Zhao. Trackocc: Camera-based 4d panoptic occupancy tracking. *arXiv preprint*, arXiv:2503.08471, 2025.

[7] D. Comaniciu and P. Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(5):603–619, 2002.

[8] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proc. of the IEEE Intl. Conf. on Computer Vision (ICCV)*, 2015.

[9] W. Gan, F. Liu, H. Xu, N. Mo, and N. Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. *arXiv preprint*, arXiv:2408.11447, 2024.

[10] W. Gan, N. Mo, H. Xu, and N. Yokoya. A simple attempt for 3d occupancy estimation in autonomous driving. *arXiv preprint*, arXiv:2303.10076, 2023.

[11] V. Guizilini, I. Vasiljevic, R. Ambrus, G. Shakhnarovich, and A. Gaidon. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters (RA-L)*, 7(2):5397–5404, 2022.

[12] A. Hornung, K. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard. OctoMap: An Efficient Probabilistic 3D Mapping Framework Based on Octrees. *Autonomous Robots*, 34(3):189–206, 2013.

[13] Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu. SelfOcc: Self-Supervised Vision-Based 3D Occupancy Prediction. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu. Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[15] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár. Panoptic Segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] V. Leroy, Y. Cabon, and J. Revaud. Grounding Image Matching in 3D with MASt3R. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2024.

[17] P. Li, R. Zhao, Y. Shi, H. Zhao, J. Yuan, G. Zhou, and Y. Zhang. LODE Locally Conditioned Eikonal Implicit Scene Completion from Sparse LiDAR. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2023.

[18] Y. Li, S. Li, X. Liu, M. Gong, K. Li, N. Chen, Z. Wang, Z. Li, T. Jiang, F. Yu, et al. Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2024.

[19] 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 *Proc. of the Conf. on Advancements of Artificial Intelligence (AAAI)*, 2023.

[20] F. Liu, C. Shen, G. Lin, and I. Reid. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(10):2024–2039, 2016.

[21] H. Liu, Y. Chen, H. Wang, Z. Yang, T. Li, J. Zeng, L. Chen, H. Li, and L. Wang. Fully sparse 3d occupancy prediction. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2024.

[22] J. Liu, L. Kong, J. Yang, and W. Liu. Towards better data exploitation in self-supervised monocular depth estimation. *IEEE Robotics and Automation Letters (RA-L)*, 9(1):763–770, 2023.

[23] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2019.

[24] R. Marcuzzi, L. Nunes, E. Marks, L. Wiesmann, T. Läbe, J. Behley, and C. Stachniss. SfmOcc: Vision-Based 3D Semantic Occupancy Prediction in Urban Environments. *IEEE Robotics and Automation Letters (RA-L)*, 10(5):5074–5081, 2025.

[25] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2020.

[26] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, B. Wang, H. Xie, L. Liu, and S. Zhang. RenderOcc Vision-Centric 3D Occupancy Prediction with 2D Rendering Supervision. In *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2024.

[27] M. Pan, L. Liu, J. Liu, P. Huang, L. Wang, S. Zhang, S. Xu, Z. Lai, and K. Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint*, arxiv:2306.09117, 2023.

[28] T. Pock, L. Zebedin, and H. Bischof. Tgv-fusion. In *Rainbow of Computer Science*, pages 245–258. 2011.

[29] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint*, arXiv:2401.14159, 2024.

[30] L. Song, D. Shi, J. Xia, Q. Ouyang, Z. Qiao, S. Jin, and S. Yang. Spatial-aware dynamic lightweight self-supervised monocular depth estimation. *IEEE Robotics and Automation Letters (RA-L)*, 9(1):883–890, 2023.

[31] 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. In *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2024.

[32] C. Wang, J. Miguel Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] K. Wang, C. Liu, Z. Liu, F. Xiao, Y. An, X. Zhao, and S. Shen. Multi-view depth estimation by using adaptive point graph to fuse single-view depth probabilities. *IEEE Robotics and Automation Letters (RA-L)*, 9(7):6400–6407, 2024.

[34] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. DUst3R: Geometric 3D Vision Made Easy. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[35] Y. Wang, Y. Chen, X. Liao, L. Fan, and Z. Zhang. PanoOcc: Unified Occupancy Representation for Camera-based 3D Panoptic Segmentation. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[36] 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 *Proc. of the Conf. on Robot Learning (CoRL)*, 2023.

[37] 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.

[38] Z. Yu, C. Shu, Q. Sun, Y. Bian, X. Wei, J. Yu, Z. Liu, D. Yang, H. Li, and Y. Chen. Panoptic-flashocc: An efficient baseline to marry semantic occupancy with panoptic via instance center. *arXiv preprint*, arXiv:2406.10527, 2024.

[39] C. Zhang, J. Yan, Y. Wei, J. Li, L. Liu, Y. Tang, Y. Duan, and J. Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint*, arXiv:2312.09243, 2023.

[40] J. Zhang, C. Herrmann, J. Hur, V. Jampani, T. Darrell, F. Cole, D. Sun, and M.H. Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In *Proc. of the Intl. Conf. on Learning Representations (ICLR)*, 2025.

[41] Y. Zhang, Z. Zhu, and D. Du. OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2023.

[42] Z. Zhang, C. Xu, J. Yang, J. Gao, and Z. Cui. Progressive hard-mining network for monocular depth estimation. *IEEE Trans. on Image Processing*, 27(8):3691–3702, 2018.

[43] H. Zou, X. Yang, T. Huang, C. Zhang, Y. Liu, w. li, F. Wen, and H. Zhang. Up-To-Down Network Fusing Multi-Scale Context for 3D Semantic Scene Completion. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2021.