LangOcc: Supplementary Material

Simon Boeder
Robert Bosch GmbH
simon.boeder@de.bosch.com

Fabian Gigengack
Robert Bosch GmbH
fabian.gigengack@de.bosch.com

Benjamin Risse
University of Münster
b.risse@uni-muenster.de

A. Details: Zero-shot Semantic Occupancy Estimation

We provide detailed results for each class in the Occ3D-nuScenes dataset in Tab. A.1. We outperform SelfOcc and OccNeRF in terms of geometric reconstruction and semantic segmentation performance already with the full embedding space, by just using the feature distillation loss. When using a predefined vocabulary (cf. Appendix C) to train the proposed dimensionality reducer, we further increase the performance on this task. We want to highlight that LangOcc can segment small, dynamic and vulnerable objects like bicycle, pedestrians and motorcycle much better than the other vision-only methods, which is reflected in the IoU score of these classes. These classes are usually more difficult to segment, but are also more important than background objects for downstream planners. On the other hand, our model seems to have difficulties in segmenting the sidewalk from the driveable street accurately. We think this comes from the fact that the text prompts we use to detect the sidewalk and the driveable surface are very similar in the CLIP space, leading to confusions. As mentioned in the paper, methods trained with just images are still far behind approaches using annotated LiDAR and voxel labels in terms of performance, but are much more scalable and are independent of expensive data acquisition.

B. Additional Qualitative Results

In Fig. B.1 we present a comparison between semantic occupancy estimations of LangOcc (Reduced) and the ground truth voxel labels of Occ3D-nuScenes on different scenes, depicted from a third-person view of the ego vehicle. Figure B.2 provides additional qualitative results from a birds-eye-view perspective. Despite the model has never seen any semantic or voxel labels, and the lack of explicit supervision for depth or geometry, the model can estimate the scene geometry well, and can detect most semantic features in the scenes. The model can even generalize to areas behind occluding objects to a certain extend. However, it is clearly visible that the semantic predictions are still fairly noisy, which likely is a result of the sometimes ambiguous vision-language features. We further provide videos in the *videos* directory of the supplementary material. Each video shows the input images, the predicted zero-shot semantic occupancy and the ground truth labels of a whole scene.

Figure B.3 illustrates feature maps of our model when rendering the estimated voxel features back to the image space, in comparison to the ground truth feature maps extracted via MaskCLIP. The features are reduced to three dimensions using PCA for visualisation (each scene individually). Note that the rendered feature maps are not the output of the model, but LangOcc outputs vision-language features in 3D voxel space. The estimated feature maps are generated by volume rendering our predictions into the input cameras. These feature maps are essentially the input into the loss function, which can be backpropagated through the whole model, as the volume rendering is fully differentiable. One can see that the rendered feature maps retain the expressiveness of the original vision-language aligned feature maps, even in 3D space. We therefore inherit all the vision-language capabilities of the original feature space, enabling *open vocabulary occupancy*.

C. Vocabulary

We present the vocabulary that we use for zero-shot semantic occupancy estimation on Occ3D-nuScenes and to train the *Reducer* in Tab. C.2. For each class in the Occ3D-nuScenes benchmark, we define a set of text prompts that describe that category. As described in the paper, during inference, we compare the estimated voxel features with each text embedding of the vocabulary, and assign each voxel the label belonging to the prompt with the highest similarity score. To train the *Reducer*, we concatenate all text prompts to form a single dataset of text embeddings, on which the *Reducer* is trained.

D. Source Code

To reproduce the results, we provide our source code at https://github.com/boschresearch/LangOcc . Follow the instructions in the *README.md* file to install the repository and run trainings and validation.

Table A.1. **Semantic occupancy estimation performance on the Occ3D-nuScenes.** Performance is measured in %IoU, best performing per column and category in **bold**, second best in *italics*.

Method	Mode	IoU	mIoU	others	barrier	bicycle	bus	car	cons. vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	driv. surf.	other flat	sidewalk	terrain	manmade	vegetation
OccFormer [7]	3D	-	21.93	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
TPVFormer [3]	3D	-	27.83	7.2	38.9	13.7	40.8	**45.9**	17.2	20.0	18.8	14.3	26.7	34.2	55.6	35.5	37.6	30.7	19.4	16.8
CTF-Occ [5]	3D	-	28.53	8.1	39.3	20.6	38.3	42.2	16.9	24.5	22.7	21.0	23.0	31.1	53.3	33.8	38.	33.2	20.8	18.0
TPVFormer [3]	L	17.20	13.57	0.0	14.8	9.4	21.3	16.8	14.5	13.8	11.2	5.3	16.1	19.7	10.8	9.4	9.5	11.2	16.5	17.0
RenderOcc [4]	L	-	23.93	5.7	27.6	14.4	19.9	20.6	12.0	12.4	12.1	14.3	20.8	18.9	68.8	33.4	42.0	43.9	17.4	22.6
OccFlowNet [1]	L	-	26.14	3.2	28.8	*22.2*	28.0	21.7	17.2	19.6	11.0	18.0	24.1	22.0	67.3	28.7	40.0	41.0	26.2	25.6
SelfOcc [2]	C	45.01	9.30	0.0	0.2	0.7	5.5	12.5	0.0	*0.8*	2.1	0.0	0.0	8.3	**55.5**	0.0	**26.3**	*26.6*	14.2	5.6
OccNeRF [6]	C	-	10.13	0.0	0.8	0.8	5.1	12.5	**3.5**	0.2	3.1	1.8	0.5	3.9	*52.6*	0.0	*20.8*	24.8	*18.4*	13.2
LangOcc (Full)	C	*51.59*	*10.71*	0.0	*2.7*	*7.2*	*5.8*	*13.9*	*0.5*	**10.8**	**6.4**	*8.7*	*3.2*	**11.0**	42.1	*1.6*	12.5	**27.2**	14.1	*14.5*
LangOcc (Reduced)	C	**51.76**	**11.84**	0.0	**3.1**	**9.0**	**6.3**	**14.2**	0.4	**10.8**	*6.2*	**9.0**	**3.8**	*10.7*	43.7	**2.23**	9.5	26.4	**19.6**	**26.4**

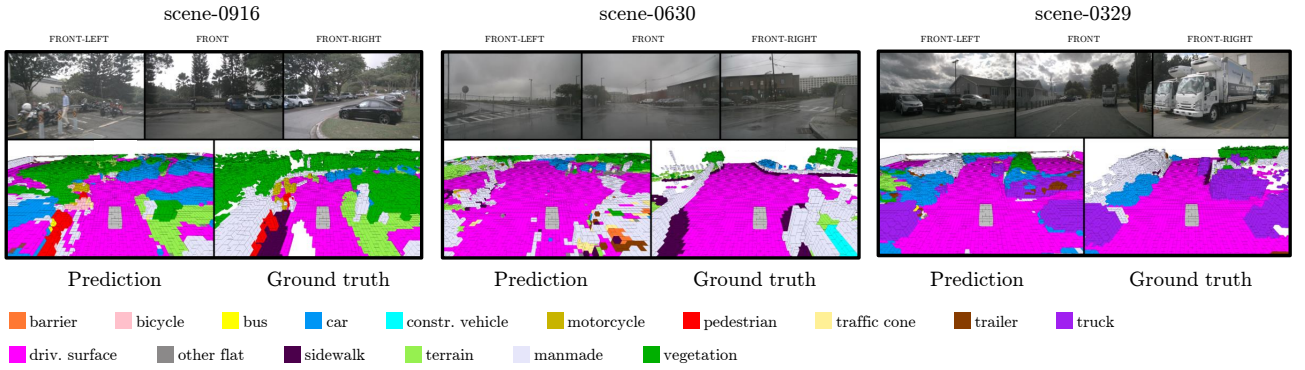

Figure B.1. **Qualitative results showing zero-shot semantic occupancy estimations.**

References

[1] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Occflownet: Towards self-supervised occupancy estimation via differentiable rendering and occupancy flow. *arXiv preprint arXiv:2402.12792*, 2024. 2

[2] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. *arXiv preprint arXiv:2311.12754*, 2023. 2

[3] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2023. 2

[4] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. *arXiv preprint arXiv:2309.09502*, 2023. 2

[5] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023. 2

[6] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*, 2023. 2

[7] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023. 2

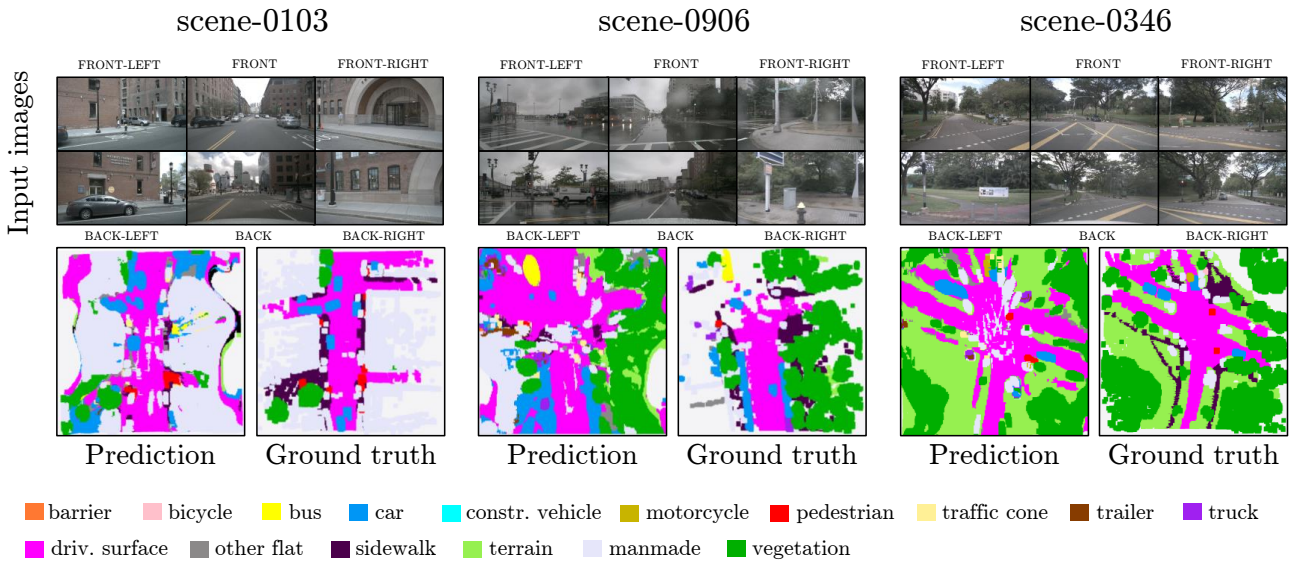

Figure B.2. **Qualitative results showing zero-shot semantic occupancy estimations.**

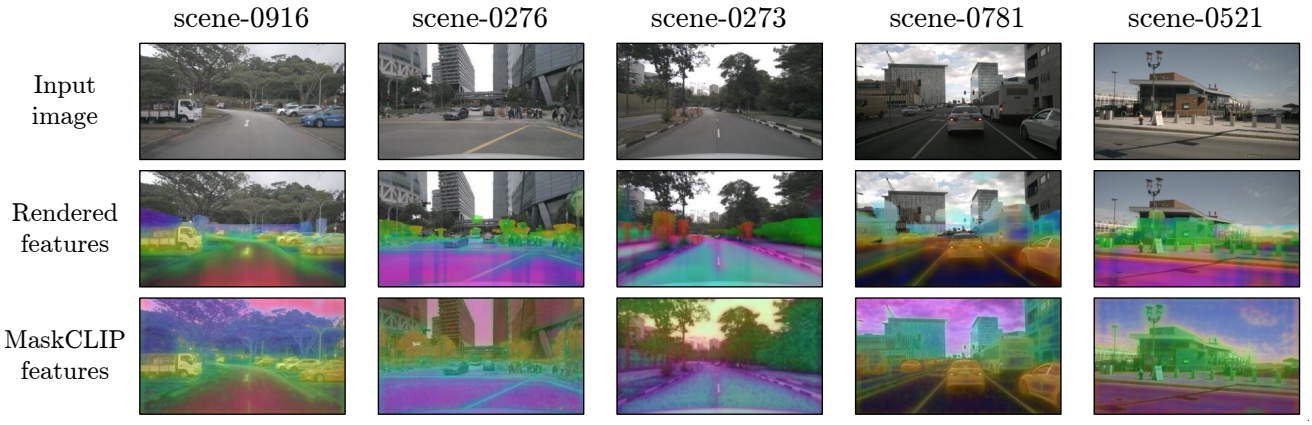

Figure B.3. **Qualitative results depicting rendered estimated 3D features and ground truth features in 2D image space.** As is visible, given just the input image, our model can replicate the original CLIP embeddings accurately. However, our model estimates them in full 3D space and still keeps the full expressiveness of the latent space.

Table C.2. **Vocabulary used for zero-shot semantic occupancy estimation and to train the *Reducer*.**

Class	Prompts
'car'	'Vehicle designed primarily for personal use.', 'car', 'vehicle', 'sedan', 'hatch-back', 'wagon', 'van', 'mini-van', 'SUV', 'jeep'
'truck'	'Vehicle primarily designed to haul cargo.', 'pick-up', 'lorry', 'truck', 'semi-tractor'
'trailer'	'trailer', 'truck trailer', 'car trailer', 'bike trailer'
'bus'	'Rigid bus', 'Bendy bus'
'construction_vehicle'	'Vehicle designed for construction.', 'crane'
'bicycle'	'Bicycle'
'motorcycle'	'motorcycle', 'vespa', 'scooter'
'pedestrian'	'Adult.', 'Child.', 'Construction worker', 'Police officer.'
'traffic_cone'	'traffic cone.'
'barrier'	'Temporary road barrier to redirect traffic.', 'concrete barrier', 'metal barrier', 'water barrier'
'driveable_surface'	'Paved surface that a car can drive.', 'Unpaved surface that a car can drive.'
'other_flat'	'traffic island', 'delimiter', 'rail track', 'small stairs', 'lake', 'river'
'sidewalk'	'sidewalk', 'pedestrian walkway', 'bike path'
'terrain'	'grass', 'rolling hill', 'soil', 'sand', 'gravel'
'manmade'	'man-made structure', 'building', 'wall', 'guard rail', 'fence', 'pole', 'drainage', 'hydrant','flag', 'banner', 'street sign', 'electric circuit box', 'traffic light', 'parking meter','stairs'
'vegetation'	'bushes', 'bush', 'plants', 'plant', 'potted plant', 'tree', 'trees'
'background'	'Any lidar return that does not correspond to a physical object, such as dust, vapor, noise, fog, raindrops, smoke and reflections.','sky'