SelfOcc: Self-Supervised Vision-Based 3D Occupancy Prediction Supplementary Material

Yuanhui Huang* Wenzhao Zheng* Borui Zhang Jie Zhou Jiwen Lu†
Beijing National Research Center for Information Science and Technology, China
Department of Automation, Tsinghua University, China
{huangyh22,zhang-br21}@mails.tsinghua.edu.cn; wenzhao.zheng@outlook.com;
{jzhou,lujiwen}@tsinghua.edu.cn

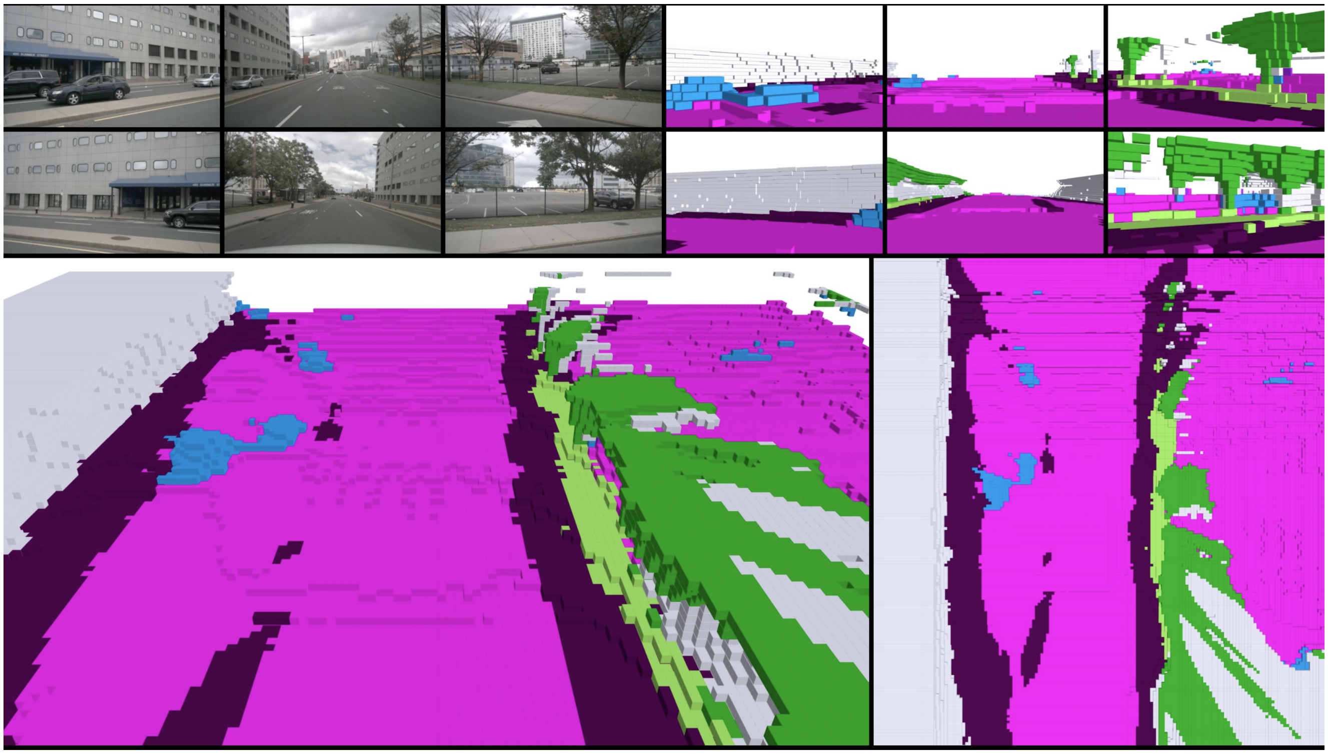

Figure 1. **Visualizations of the proposed SelfOcc method for 3D semantic occupancy prediction on the nuScenes validation set.** We show the six input surrounding images in the upper left and the predicted semantic occupancy from the corresponding views. The lower parts demonstrate the predicted results in the global view (left) and bird's eye view (right).

A. Dataset Details

The nuScenes [2] dataset consists of 1000 sequences of various driving scenes under different weather and lighting conditions, which are officially split into 700/150/150 sequences for training, validation and testing. Each sequence lasts 20 seconds with LiDAR point cloud and RGB images collected by 6 surround cameras, and the keyframes are annotated at 2Hz. In addition, the Occ3D-nuScenes [12] dataset provides 3D semantic occupancy annotations with a resolution of 200x200x16 for 18 classes, covering an area of 80/80/6.4 meters around the ego car in the x/y/z-axis.

The KITTI-2015 [5] dataset holds stereo images from two forward-facing cameras and LiDAR point cloud. Following [6], we use the Eigen split [4] and remove static frames from the training set, which results in 39,810 frames for training and 4,424 for validation.

The SemanticKITTI [1] dataset is based on the odometry subset of the KITTI-2015 [5] dataset and provides voxelized lidar scans for 22 sequences with a resolution of 256x256x32. Each voxel has a side length of 0.2m and is labeled with one of the 21 classes (19 semantic, 1 free and 1 unknown). In our experiments, we only use the images from cam2 and follow the official split of the dataset, i.e. 10, 1 and 11 sequences for training, validation, and test.

[*] Equal contribution. †Corresponding author.

B. Additional Implementation Details

2D segmentor for semantic prediction. For 3D semantic occupancy prediction on nuScenes [2], we leverage the tiny version of the open-vocabulary 2D segmentor OpenSeeD [16] trained on COCO2017 [10] and Objects365v1 [11] to directly predict semantic segmentation maps for supervision. Note that although OpenSeeD ignores some classes due to rarity (e.g. construction vehicle) or semantic-ambiguity (e.g. others and other flat), we still consider all classes when calculating mIoU.

Geometric settings. We use TPV [7] or BEV [8] uniformly divided to represent a cuboid area, i.e. [80, 80, 6.4] meters around the ego car for nuScenes [2] and [51.2, 51.2, 6.4] meters in front of the ego car for SemanticKITTI [1] and KITTI-2015 [5]. The resolution for a single TPV/BEV grid cell is 0.4 meters for nuScenes and 0.2 meters for SemanticKITTI and KITTI-2015, respectively. For depth prediction, we calculate metrics for depth values in the range of [0.1, 80] meters following [6, 13]. And we evaluate depth prediction at 1:2 resolution against the raw image.

Training settings. The resolution of input image is 384×800 for nuScenes, 370×1220 for SemanticKITTI following [3] and 320×1024 for KITTI-2015 following [6, 17]. For the loss weights, we set $\lambda_c = \lambda_e = \lambda_H = 0.1$, $\lambda_s = 0.001$ if present, and the weights for the edge L_{edg} and the semantic L_{sem} losses are 0.01 and 0.1, respectively, if applied. We train our models on 8 RTX-3090 GPUs with 24GB memory. Experiments on SemanticKITTI [1] and KITTI-2015 [5] take less than one day, while experiments on nuScenes [2] finish within two days.

C. Mathematical Derivation

In this section, we further discuss the advantage of our proposed MVS-embedded depth optimization over the traditional reprojection loss with mathematical derivations.

As in Eq. (8), the reprojection loss can be formulated as

$$L_{rpj}(\mathbf{x}, \mathbf{I}_t, \mathbf{I}_s; \boldsymbol{\theta}) = \big\|\mathbf{I}_t(\mathbf{x}) - \mathbf{I}_s(\hat{\mathbf{x}}(\boldsymbol{\theta}))\big\|. \quad (1)$$

Then we further expand $\mathbf{I}_s(\hat{\mathbf{x}}(\boldsymbol{\theta}))$ according to the definition of bilinear interpolation to get

$$L_{rpj} = \Big\|\mathbf{I}_t(\mathbf{x}) - \sum_{i,j\in\{0,1\}} w_{ij}(\boldsymbol{\theta})\mathbf{I}_s\big[\lfloor\hat{\mathbf{x}}+(i,j)\rfloor\big]\Big\|, \quad (2)$$

where $\lfloor\cdot\rfloor$, $\lfloor\hat{\mathbf{x}}+(i,j)\rfloor$ and $\mathbf{I}_s[\cdot]$ denote the floor operation, the adjacent corner pixels of $\hat{\mathbf{x}}$ and the indexing operation, respectively. In addition, $w_{ij}(\boldsymbol{\theta})$ is the normalized interpolation weight of the ijth adjacent corner pixel. Note that once $\hat{\mathbf{x}}$ is calculated according to perspective transformation, $\mathbf{I}_s[\lfloor\hat{\mathbf{x}}+(i,j)\rfloor]$ is fixed and not differentiable with respect to $\boldsymbol{\theta}$. Therefore, the receptive field of the optimization problem in (2) is limited to only four adjacent corner pixels involved in bilinear interpolation, which has an adverse effect on the efficiency and stability of depth learning. Moreover, the summation operation in (2) is inside the norm bracket, which could lead to coupling of the adjacent corner pixels and local minima.

$$L_{mvs} = \sum_{m=1}^{M} w_m(\boldsymbol{\theta})\|\mathbf{I}_t(\mathbf{x}) - \mathbf{I}_x(\pi(\mathbf{x}, d_m, \boldsymbol{\Pi}))\|. \quad (3)$$

In contrast, our MVS-embedded depth optimization in (3) moves the summation outside the dissimilarity metric, and effectively enlarges the receptive field by incorporating multiple depth candidates d_m along the ray.

D. Additional Experiments

D.1. Analysis on Unbounded Regions

Unbounded regions are one key factor for autonomous driving applications, which refer to the regions outside the boundary of 3D representations. However, we find that these regions have negligible influence on the optimization of SelfOcc. In Tab 1, we calculate the averaged distance (Dist2Bnd) between the predicted 3D locations of unbounded points and the boundary of TPV cuboid with the model for depth estimation on nuScenes. We identify LiDAR points outside the TPV cuboid as unbounded points and retrieve their corresponding pixels and rendered depths for inverse projection. Unbounded regions are predicted close (∼ 1m) to the boundary of TPV cuboid, thus having little impact on the correctness of SelfOcc. We attribute it to the noisy nature of gradients from unbounded regions which are neutralized across samples. On the other hand, simple nonlinear mapping could extend the representation range of TPV to the same as methods conditioned on image-level features (80m). We pad our original TPV planes with a small padding size to represent the interval between 51.2m and 80m, and the result is also reported in Tab 1.

Tab 1. **Analysis on unbounded pixels.**

Method	Dist2Bnd (m)	Abs Rel↓	Sq Rel↓	RMSE↓	$\delta 1$↑
SelfOcc	1.021	0.263	3.650	**7.266**	0.716
SelfOcc-Pad	-	**0.242**	**3.454**	7.454	**0.718**

D.2. Analysis on Color Supervision

We provide further analysis on whether and how to apply color supervision in the surround-view setting and conduct additional experiments on nuScenes. The first question is sampled radiance [14] or MLP-predicted radiance. Since the nuScenes dataset is more challenging than the KITTI dataset considering differences among surround cameras, we think it is necessary to use the MLP-predicted radiance to account for the viewpoint dependency of color. The second question is whether to apply color supervision. Our

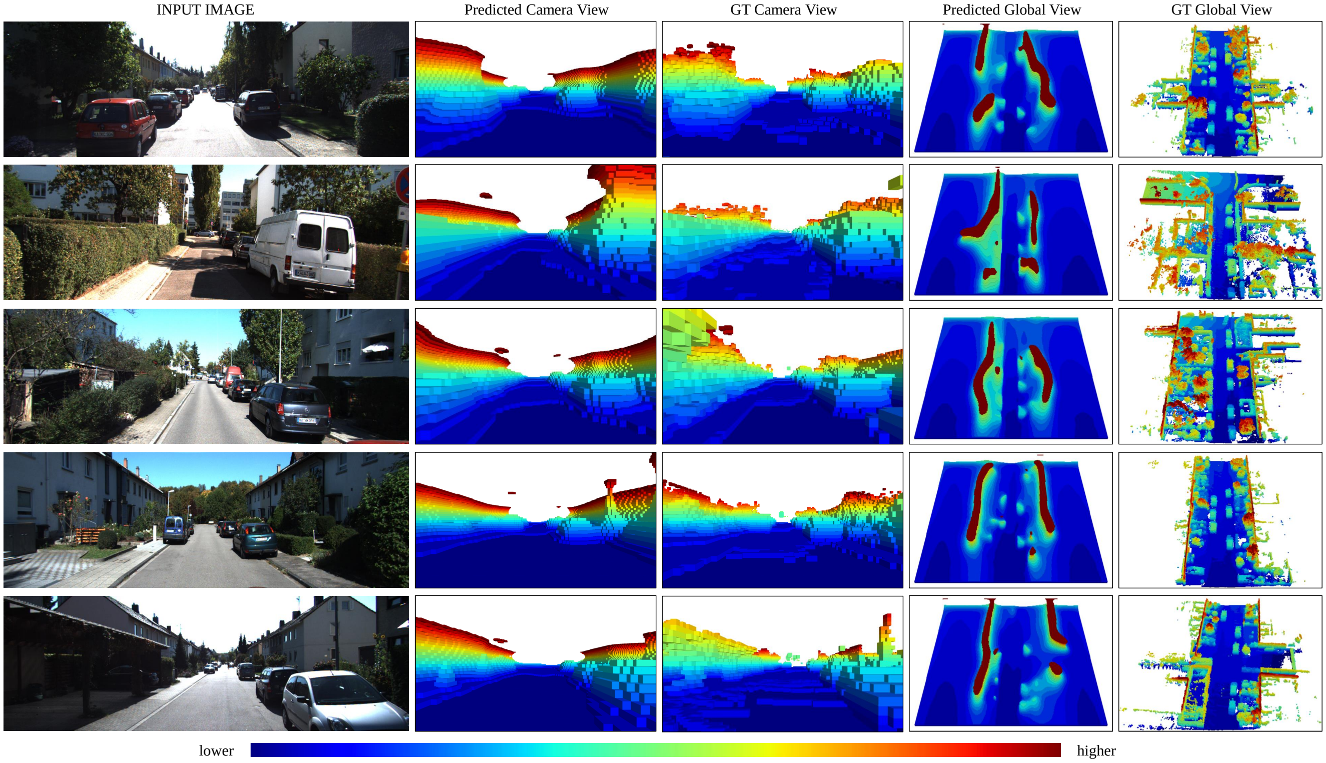

Figure 2. **Visualizations of 3D occupancy prediction on the SemanticKITTI [1] validation set.**

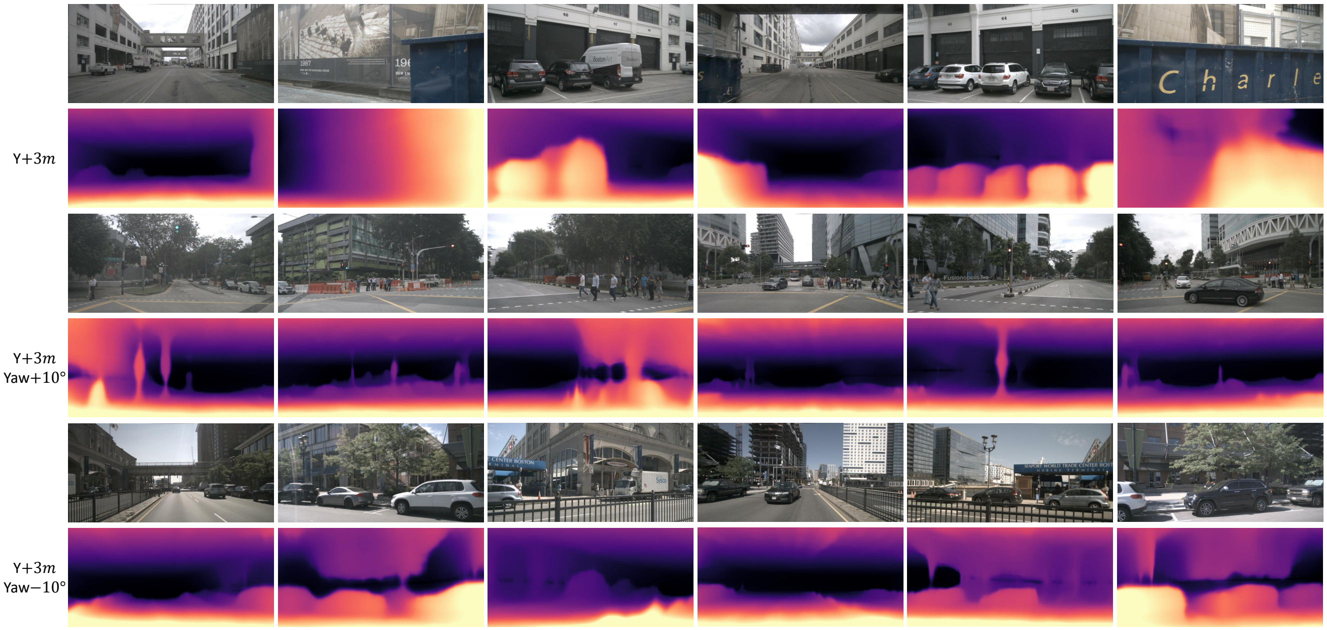

Figure 3. **Visualizations of novel depth synthesis on the nuScenes validation set.**

motivation is to incorporate texture clues with color supervision, since depth optimization based on photometric loss can be very noisy in real-world outdoor scenarios. According to Tab 2 and Table 7, color supervision benefits occupancy prediction and novel depth synthesis on both datasets.

Tab 2. **Analysis on color supervision.**

Method	Occ. IoU↑	Occ. mIoU↑	N.D. Abs Rel↓	N.D. RMSE↓
SelfOcc	**43.38**	**7.97**	**0.4003**	**8.460**
SelfOcc w/o L_{rgb}	38.06	7.27	0.4013	8.474

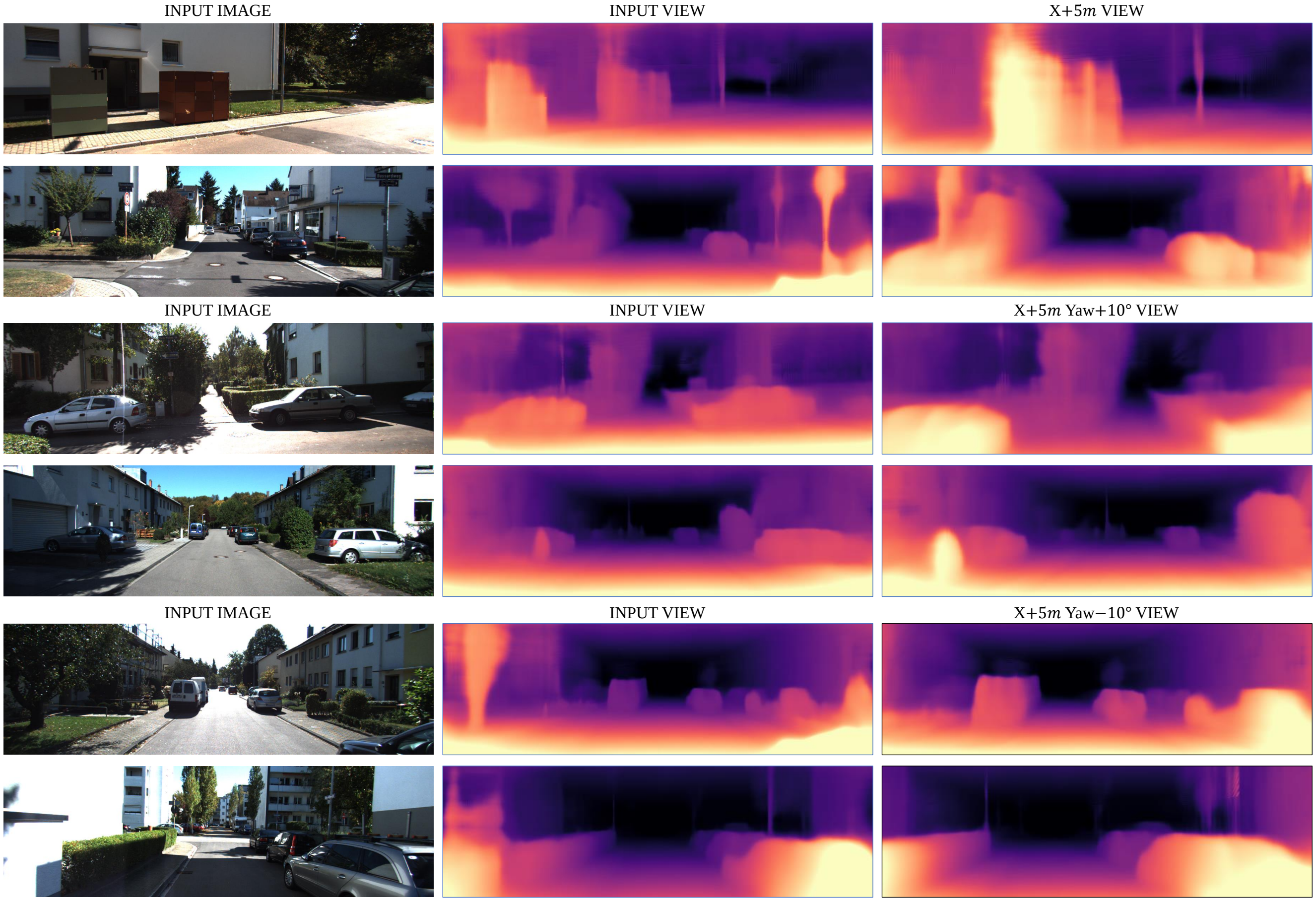

Figure 4. **Visualizations of novel depth synthesis on the SemanticKITTI validation set.**

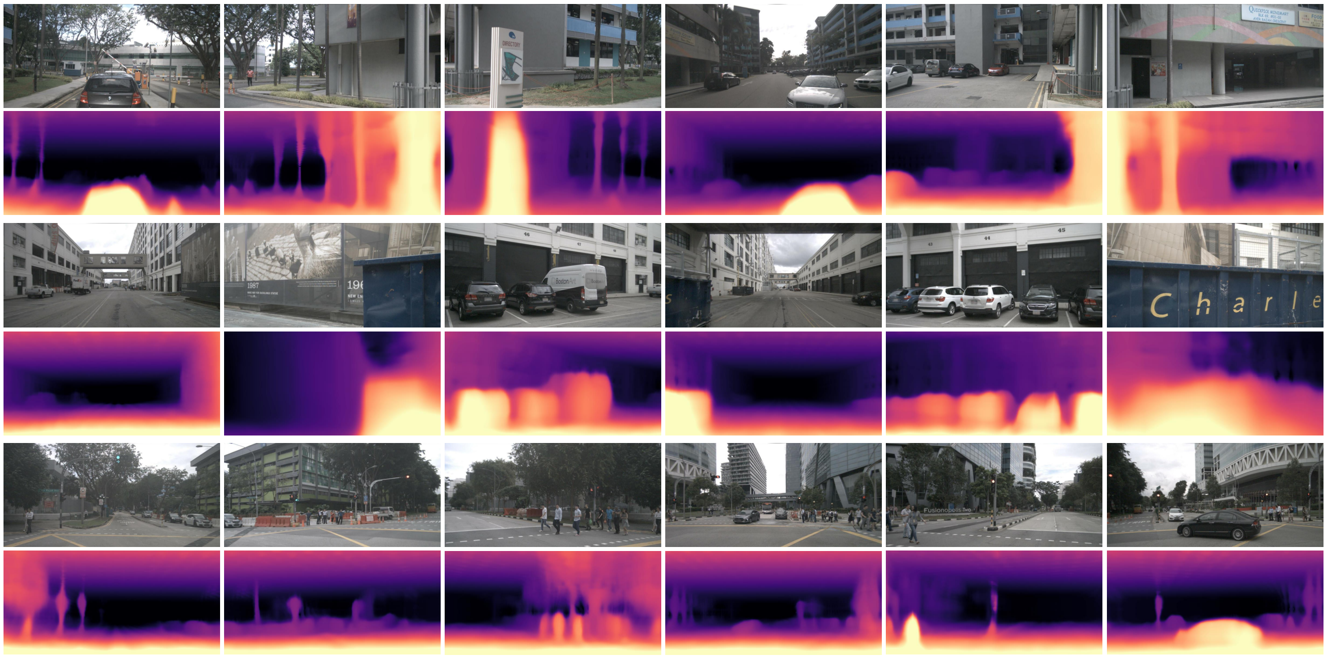

Figure 5. **Visualizations of surrounding depth prediction on the nuScenes validation set.**

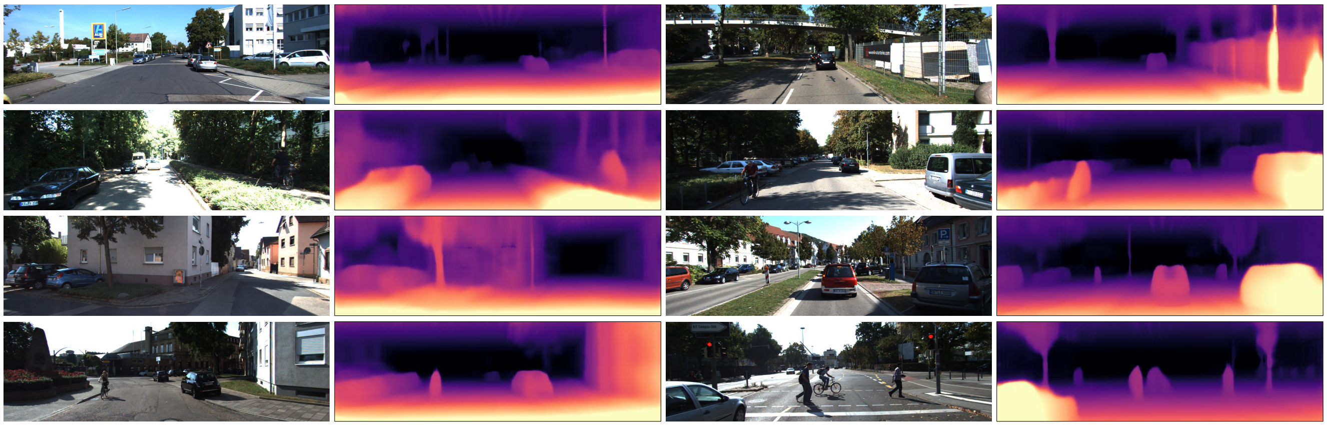

Figure 6. **Visualizations of depth estimation on the KITTI-2015 [5] test split.**

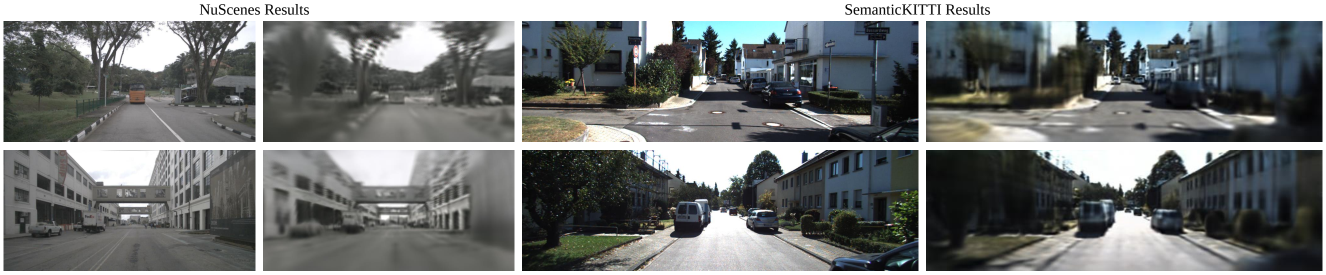

Figure 7. **Visualizations of novel view synthesis on nuScenes [2] and SemanticKITTI [1].** We use the models for novel depth synthesis to synthesize novel views since these two tasks are similar. SelfOcc suffers from the blurring effect.

D.3. Application for Pretraining

In Tab 3, we use different number of labeled scenes from Occ3D-nuScenes to finetune two versions of TPVFormer with ImageNet-pretrained and SelfOcc-pretrained weights, respectively. SelfOcc-pretrained models outperform ImageNet-pretrained counterparts under all settings, which demonstrates the potential of SelfOcc as an effective pretraining method.

Tab 3. **Performance of SelfOcc for pretraining.**

Pretrain	100% (IoU / mIoU)	50% (IoU / mIoU)	25% (IoU / mIoU)
ImageNet	63.29 / 29.98	59.81 / 27.24	55.70 / 23.00
SelfOcc	**64.26 / 31.53**	**61.22 / 28.99**	**58.08 / 25.55**

E. Visualizations

E.1. 3D Occupancy Prediction

Figure 1 shows a sampled image from the video demos [1] for 3D geometric and semantic occupancy prediction on nuScenes [2] validation set. The demos show that SelfOcc can successfully infer semantic and geometric occupancy even for occluded areas. Figure 2 shows the visualizations for 3D occupancy prediction on the SemanticKITTI [1] validation set, in which SelfOcc predicts accurate shapes and sizes of cars without any occupancy shadows.

E.2. Novel Depth Synthesis

Figure 3 and 4 show the visualization results of novel depth synthesis on the nuScenes [2] validation set and SemanticKITTI [1] validation set, respectively. Y+3m (X+5m) means moving +3 (+5) meters along the y-axis (x-axis) of the LiDAR coordinate. Yaw+$10°$/$-10°$ means turning left/right for $10°$. SelfOcc trained with temporal supervision can predict 3D structures beyond the visible surface, thus generating high-quality novel depth views.

E.3. Depth Estimation

Figure 5 and 6 shows the visualizations for depth estimation on the nuScenes [2] validation set and KITTI-2015 [5] test split, respectively. In addition to vehicles, our method successfully predicts sharp and accurate depth even for thin poles, moving pedestrians and cyclists.

F. Limitations and Future Work

Although we use color supervision during training to better exploit texture priors of RGB images, our model cannot synthesize high-quality novel views, suffering from the blurring effect as shown in Figure 7, which is a long-standing problem in the field of generalizable NeRFs [3, 9, 15]. In addition, although SelfOcc can predict accurate occupancy and depth for moving objects just as well as static ones, we do not include specific designs for motion. We

[1] https://huang-yh.github.io/SelfOcc.

think the model might generalize the knowledge it learns from static elements to non-static ones. Therefore, high-quality novel view synthesis and motion awareness could be potential focuses of future work.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, pages 9297–9307, 2019. 1, 2, 3, 5

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1, 2, 5

[3] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9387–9398, 2023. 2, 5

[4] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, pages 2650–2658, 2015. 1

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 1, 2, 5

[6] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 1, 2

[7] 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

[8] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 2

[9] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yichang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 5

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2

[11] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. 2

[12] 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. 1

[13] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *CoRL*, pages 539–549. PMLR, 2023. 2

[14] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9076–9086, 2023. 2

[15] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 5

[16] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *ICCV*, pages 1020–1031, 2023. 2

[17] Kaichen Zhou, Lanqing Hong, Changhao Chen, Hang Xu, Chaoqiang Ye, Qingyong Hu, and Zhenguo Li. Devnet: Self-supervised monocular depth learning via density volume construction. In *ECCV*, pages 125–142. Springer, 2022. 2