

## A Appendix / supplemental material

### A.1 Dataset

For the local occupancy prediction task, we adopt the *Occ-ScanNet* dataset [47], which provides per-frame voxelized annotations in a  $60 \times 60 \times 36$  grid, corresponding to a  $4.8 \text{ m} \times 4.8 \text{ m} \times 2.88 \text{ m}$  volume in front of the camera. Each voxel is labeled with one of 12 semantic categories, including 11 foreground classes (e.g., *ceiling*, *floor*, *wall*, etc.) and one for empty space. These dense 3D semantic labels enable supervised training and evaluation of our local occupancy prediction module using monocular RGB inputs. The dataset contains 45,755 training and 19,764 validation frames in total.

To evaluate embodied 3D scene understanding, we utilize the *EmbodiedOcc-ScanNet* dataset [43], which reorganizes scenes from *Occ-ScanNet* to support continuous exploration. It comprises 537 training and 137 validation scenes, each containing 30 camera-pose-aligned RGB frames and their associated voxel-level occupancy annotations in the world coordinate system. This setting allows for temporally consistent updates to a global 3D occupancy map, making it well-suited for assessing progressive prediction under agent-based exploration.

For efficient experimentation and ablation, we also adopt two curated subsets derived from the above datasets: *Occ-ScanNet-mini* and *EmbodiedOcc-ScanNet-mini* [43]. *Occ-ScanNet-mini* includes 5,504 training and 2,376 validation frames sampled from the full *Occ-ScanNet*, preserving the same voxel grid structure. *EmbodiedOcc-ScanNet-mini* consists of 64 training and 16 validation scenes, each with 30 sequential frames and associated poses. This smaller-scale benchmark supports fast evaluation of embodied occupancy models while retaining the core characteristics of global memory construction and online update.

In addition to the above datasets, we further evaluate our method on the *SSCBench-KITTI-360* [24] benchmark to assess its generalization under diverse real-world outdoor scenes. *SSCBench-KITTI-360* extends the *KITTI-360* dataset [26] by providing dense 3D semantic occupancy annotations for autonomous driving scenarios, covering complex suburban environments with high-resolution panoramic imagery and LiDAR data. The benchmark defines a voxelized 3D grid of  $256 \times 256 \times 32$  spanning a  $51.2 \times 51.2 \times 6.4 \text{ m}^3$  volume in front of the ego vehicle. Each voxel is annotated with one of 19 semantic classes, supporting detailed evaluation of semantic scene completion from monocular inputs. This additional experiment demonstrates the broader applicability of our method beyond indoor domains and across diverse driving scenes.

### A.2 On the Connection to Autoregressive Modeling

**Visual Autoregressive (VAR) Modeling in 2D.** Autoregressive modeling is a classical principle where each prediction is explicitly conditioned on previous predictions. In the visual domain, recent Visual Autoregressive (VAR) models [38] have adopted this paradigm by modeling coarse-to-fine dependencies in image synthesis, where finer-resolution content is conditioned on coarser-scale predictions. This structured dependency contrasts with feed-forward prediction approaches that attempt to produce the entire output in a single step without explicitly modeling sequential relationships.

**Autoregression in 3D.** In 3D occupancy prediction, similar challenges arise when modeling fine-scale geometry from sparse or noisy visual cues. Direct voxel-based refinement often suffers from high computational cost and memory usage, while single-scale Gaussian-based models [43, 12] lack a principled mechanism to propagate information across different spatial resolutions. We posit that an autoregressive framework is particularly suited for 3D vision tasks, since fine-scale occupancy details (e.g., thin structures, furniture edges) should be predicted conditional on coarser-level structural priors (e.g., walls, floor layout). This mirrors the 2D VAR paradigm but reformulates it for continuous Gaussian parameter spaces rather than discrete pixel grids.

**Autoregression in Gaussian Parameter Space.** Formally, let  $\mathbf{G}^{(s)} = \{\mathbf{g}_i^{(s)}\}_{i=1}^{N_s}$  denote the set of Gaussians at scale  $s$ , where each Gaussian  $\mathbf{g}_i^{(s)} = \{\mu_i^{(s)}, \lambda_i^{(s)}, q_i^{(s)}, o_i^{(s)}, l_i^{(s)}\}$  is parameterized by its mean, scale, quaternion rotation, opacity, and semantic logits. We define the refinement process in a

hierarchical autoregressive manner as

$$p\left(\mathbf{G}^{(S)} \mid \mathbf{G}^{(1:S-1)}, \hat{\mathbf{G}}^{(1:S)}\right) = \prod_{s=1}^S p\left(\mathbf{G}^{(s)} \mid \mathbf{G}^{(1:s-1)}, \hat{\mathbf{G}}^{(s)}\right), \quad (9)$$

where  $\hat{\mathbf{G}}^{(s)}$  denotes the initial prediction from the scale-specific encoder. Each finer-scale Gaussian set  $\mathbf{G}^{(s)}$  is thus conditioned on all coarser-scale representations  $\mathbf{G}^{(1:s-1)}$  and its own initialization  $\hat{\mathbf{G}}^{(s)}$ , enabling structured, hierarchical propagation of geometric and semantic information across scales.

This formulation directly aligns with the broader definition of autoregression[14]: each finer-scale prediction is explicitly dependent on previously generated coarser-scale values. While classical VAR models operate in discrete image token or patch space, our framework adopts the same principle in the continuous 3D Gaussian parameter domain, thereby bridging autoregressive modeling with efficient 3D scene representation.

### A.3 More Details about the Gaussian Encoder

Our Gaussian encoder follows prior work [12, 43] for the basic initialization of 3D Gaussians, while we introduce novel refinement modules as described in the main paper. For completeness, we summarize the baseline design here.

**Gaussian Initialization.** Following [12, 43], each Gaussian is parameterized as

$$g = \{\mu, \lambda, q, o, l\},$$

where  $\mu \in \mathbb{R}^3$  is the mean position,  $\lambda \in \mathbb{R}^3$  are scale factors,  $q \in \mathbb{R}^4$  is the quaternion rotation,  $o \in \mathbb{R}$  is the opacity, and  $l \in \mathbb{R}^C$  are semantic logits. The initialization uses a depth-aware MLP to estimate per-pixel depth and combine it with image features to produce a structured set of Gaussians inside the camera frustum.

**Self-Encoding Module.** As in [12, 43], Gaussians are projected into a sparse 3D grid, and a 3D sparse convolution module enables interactions among Gaussians. This operation converts the continuous Gaussian set into a format compatible with voxel-style feature processing, producing a sparse 3D grid of Gaussian-centered features.

**Cross-Attention with Image Features.** [12, 43] Each Gaussian samples a set of reference points around its mean (offset by covariance), which are then projected back into the image plane using known intrinsics and extrinsics. The corresponding image features at these positions are aggregated using deformable attention, providing visual cues to update Gaussian descriptors.

### A.4 Implementation details

We follow the dataset configurations established in prior work [47, 43, 12]. All experiments utilize a pretrained EfficientNet [20] as the image encoder. A single monocular RGB image, resized to  $480 \times 640$ , is processed to extract a multi-scale feature pyramid with spatial resolutions downsampled by 50% and 30%. Each scene is represented using Gaussians with an anchor size of 16,200 and a maximum scale of 0.08 m. To ensure domain consistency, the encoder is initialized with weights pretrained on the corresponding dataset. We use the AdamW optimizer with a linear warmup for the first 500 iterations, followed by a cosine learning rate schedule, and a peak learning rate of  $1 \times 10^{-3}$ . All models are trained for 20 epochs on 6 NVIDIA A100 GPUs. Following [43], we perform global occupancy prediction by fine-tuning from models pretrained on local prediction tasks. Specifically, local Gaussian representations are predicted at 0.16 m intervals and aggregated to produce a global view of the scene. Since the encoder operates on a feature pyramid with reduced spatial resolution, the Gaussian anchors and intermediate predictions are proportionally shrunk. The original occupancy label has a resolution of  $60 \times 60 \times 36$ , and we produce predictions at multiple resolutions (e.g., 100%, 50%, 30%) corresponding to the Gaussian refinement hierarchy. The supervision labels are downsampled from the original ground truth to match each prediction scale accordingly.

For the outdoor KITTI-360 experiments, we follow the training setup of [9] to ensure fair comparison with existing baselines. Specifically, input images are resized to  $376 \times 1408$ , and standard augmentations including random horizontal flipping and photometric distortion are applied during training. We adopt a ResNet50 backbone pretrained on ImageNet as the image encoder, consistent with prior work. Models are trained for 30 epochs, while keeping all other configurations identical to our indoor setup—such as the use of AdamW optimizer, cosine learning rate schedule with 500-step linear warmup, and a peak learning rate of  $1 \times 10^{-3}$ .

## A.5 Baseline Methods

We compare our method against both traditional voxel-based approaches and recent Gaussian-based representations for monocular 3D occupancy prediction. Among voxel-based methods, *MonoScene* [3] introduces a FLoSP module to lift 2D features into 3D voxels, and the recent *ISO* [47] method leverages a pretrained depth network and a dual-projection strategy to learn depth-aware voxel features across scales. To capture finer-grained geometry, we further include Gaussian-based baselines: *EmbodiedOcc* [43] models scenes using a persistent 3D Gaussian memory updated with deformable attention, enabling consistent occupancy prediction under embodied exploration; and *SplicingOcc*, which fuses local Gaussian predictions across frames to estimate global scene occupancy. These baselines represent state-of-the-art paradigms in both voxel-centric and object-centric 3D reasoning.

To evaluate the generalization ability of our method in outdoor scenarios, we compare against a range of representative baselines including traditional voxel-based methods such as *LMSCNet* [35] and *SSCNet* [36], as well as recent camera-based 3D semantic occupancy approaches like *VoxFormer* [25], *TPVFormer* [11], and *OccFormer* [50]. We further include Gaussian-based methods such as *GaussianFormer* [12] and *GaussianFormerV2* [9], which leverage sparse deformable primitives for efficient 3D scene representation. Our method consistently achieves superior performance across key semantic classes, demonstrating strong generalization and robust spatial reasoning in complex environments with diverse layouts and occlusions.

## A.6 Additional Experiment

Table 8: **Monocular 3D semantic occupancy prediction results on SSCBench-KITTI-360.** We compare IoU, mIoU, and per-class IoU across 21 semantic categories. **C** denotes models that use only camera (RGB) input, while **L** denotes models that use LiDAR-based depth input.

Method	Input	IoU	mIoU	car	bicycle	motorcycle	truck	other-veh.	person	road	parking	sidewalk	other-grnd	building	fence	vegetation	terrain	pole	traf.-sign	other-struct.	other-object
LMSCNet [35]	L	47.53	13.65	20.91	0	0	0.26	0	0	62.95	13.51	33.51	0.2	43.67	0.33	40.01	26.80	0	0	3.63	0
SSCNet [36]	L	53.58	16.95	31.95	0	0.17	10.29	0.58	0.07	65.7	17.33	41.24	3.22	44.41	6.77	43.72	28.87	0.78	0.75	8.60	0.67
MonoScene [3]	C	37.87	12.31	19.34	0.43	0.58	8.02	2.03	0.86	48.35	11.38	28.13	3.22	32.89	3.53	26.15	16.75	6.92	5.67	4.20	3.09
Voxformer [25]	C	38.76	11.91	17.84	1.16	0.89	4.56	2.06	1.63	47.01	9.67	27.21	2.89	31.18	4.97	28.99	14.69	6.51	6.92	3.79	2.43
TPVFormer [11]	C	40.22	13.64	21.56	1.09	1.37	8.06	2.57	2.38	52.99	11.99	31.07	3.78	34.83	4.80	30.08	17.51	7.46	5.86	5.48	2.70
OccFormer [50]	C	<b>40.27</b>	13.81	<b>22.58</b>	0.66	0.26	9.89	3.82	2.77	<b>54.30</b>	13.44	31.53	3.55	<b>36.42</b>	4.80	<b>31.00</b>	<b>19.51</b>	<b>7.77</b>	<b>8.51</b>	6.95	4.60
GaussianFormer [12]	C	35.38	12.92	18.93	1.02	4.62	<b>18.07</b>	<b>7.59</b>	<b>3.35</b>	45.47	10.89	25.03	<b>5.32</b>	28.44	5.68	29.54	8.62	2.99	2.32	<b>9.51</b>	<b>5.14</b>
GaussianFormerV2 [9]	C	38.37	13.90	21.08	<b>2.55</b>	4.21	12.41	5.73	1.59	54.12	11.04	<b>32.31</b>	3.34	32.01	4.98	28.94	17.33	3.57	5.48	5.88	3.54
<b>Ours</b>	C	39.89	<b>14.58</b>	22.21	1.85	<b>4.88</b>	14.78	5.97	2.03	54.23	<b>15.78</b>	31.89	4.52	32.28	<b>6.12</b>	29.02	18.63	4.02	4.25	6.03	4.03

To assess the generalization of our framework to outdoor environments, we evaluate DFGauss on the SSCBench-KITTI-360 dataset, as shown in Table 8. The depth prediction module is pretrained on SSCBench-KITTI-360 following the same setup as prior work [43]. We compare our method against both traditional grid-based approaches and recent Gaussian-based baselines. DFGauss achieves superior overall performance among Gaussian-based methods, demonstrating its ability to preserve the computational efficiency of sparse Gaussian representations while improving predictive accuracy. Notably, our model exhibits enhanced performance on several fine-grained classes (e.g., *motorcycle*, *fence*, and *parking*), suggesting that multi-scale refinement strengthens the discriminative capacity of Gaussian primitives for small or complex objects.

We further report model statistics of our multi-scale autoregressive Gaussian framework. As shown in Table 9, the latency increases only slightly with additional refinement levels, demonstrating the computational efficiency and scalability of our hierarchical design.

Table 9: Model statistics across hierarchical depths.

Depth	Latency (ms)	Downsampling Ratio (%)
1	133.604	100%
2	133.732	[50%, 100%]
3	133.889	[30%, 50%, 100%]
4	134.101	[10%, 30%, 50%, 100%]
5	134.235	[10%, 30%, 50%, 80%, 100%]

### A.7 Additional Qualitative Evaluation

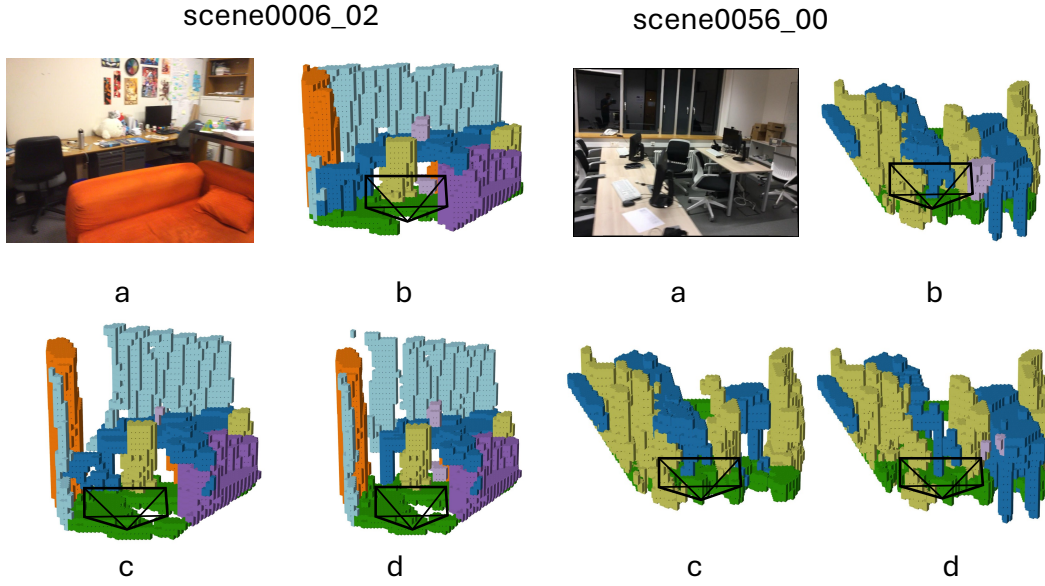


Figure 5: Qualitative results on Occ-ScanNet: (a) Input, (b) GT, (c) EmbodiedOcc, (d) DFGauss.

Additional qualitative results in Figure 5 show that our method produces more fine-grained and complete occupancy predictions compared to the baseline, demonstrating the effectiveness of the proposed multi-scale autoregressive refinement module in DFGauss.

## References

- [1] Mohamed Abdelsamad, Michael Ulrich, Claudius Gläser, and Abhinav Valada. Multi-scale neighborhood occupancy masked autoencoder for self-supervised learning in LiDAR point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [2] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. In *International Conference on Machine Learning*, pages 950–959. PMLR, 2020.
- [3] Anh-Quan Cao and Renaud de Charette. MonoScene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [4] Junliang Chen, Huaiyuan Xu, Yi Wang, and Lap-Pui Chau. OccProphet: Pushing the efficiency frontier of camera-only 4d occupancy forecasting with an observer-forecaster-refiner framework. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
- [5] Liang Chen, Sinan Tan, Zefan Cai, Weichu Xie, Haozhe Zhao, Yichi Zhang, Junyang Lin, Jinze Bai, Tianyu Liu, and Baobao Chang. A spark of vision-language intelligence: 2-dimensional autoregressive transformer for efficient finegrained image generation. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
- [6] Shizhe Chen, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. SUGAR : Pre-training 3d visual representations for robotics. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18049–18060, 2024.
- [7] Zhiwen Fan, Kevin Wang, Kairun Wen, Zehao Zhu, Dejie Xu, and Zhangyang Wang. LightGaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 140138–140158. Curran Associates, Inc., 2024.
- [8] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two stream 3d semantic scene completion. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 416–425, 2019.
- [9] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.
- [10] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. SelfOcc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19946–19956, 2024.
- [11] 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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9223–9232, 2023.
- [12] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. GaussianFormer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 376–393. Springer, 2024.
- [13] Jiwan Hur, Dong-Jae Lee, Gyojin Han, Jaehyun Choi, Yunho Jeon, and Junmo Kim. Unlocking the capabilities of masked generative models for image synthesis via self-guidance. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 130977–130999. Curran Associates, Inc., 2024.
- [14] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2nd edition, 2018. Section 8.3: Autoregressive models.

- [15] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view PointNet for 3d scene understanding. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3995–4003, 2019.
- [16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the Sixth International Conference on Learning Representations (ICLR)*, 2018.
- [17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.
- [18] Wesley Khademi and Fuxin Li. Point-based instance completion with scene constraints. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
- [19] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 353–369. Springer, 2022.
- [20] Brett Koonce. EfficientNet. In *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pages 109–123. Springer, 2021.
- [21] Joo Chan Lee, Daniel Rho, Xiangyu Sun, Jong Hwan Ko, and Eunbyung Park. Compact 3d gaussian representation for radiance field. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21719–21728, 2024.
- [22] Jinke Li, Xiao He, Chonghua Zhou, Xiaoqiang Cheng, Yang Wen, and Dan Zhang. Viewformer: Exploring spatiotemporal modeling for multi-view 3d occupancy perception via view-guided transformers, 2024.
- [23] Xiang Li, Pengfei Li, Yupeng Zheng, Wei Sun, Yan Wang, and yilun chen. Semi-supervised vision-centric 3d occupancy world model for autonomous driving. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [24] Yiming Li, Sihang Li, Xinhao Liu, Moonjun Gong, Kenan Li, Nuo Chen, Zijun Wang, Zhiheng Li, Tao Jiang, Fisher Yu, Yue Wang, Hang Zhao, Zhiding Yu, and Chen Feng. SSCBench: A large-scale 3d semantic scene completion benchmark for autonomous driving. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13333–13340, 2024.
- [25] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9087–9098, 2023.
- [26] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2023.
- [27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017.
- [28] Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 54–71. Springer, 2024.
- [29] Qihao Liu, Zhanpeng Zeng, Ju He, Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Alleviating distortion in image generation via multi-resolution diffusion models and time-dependent layer normalization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 133879–133907. Curran Associates, Inc., 2024.

- [30] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 1999.
- [31] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. OctreeOcc: Efficient and multi-granularity occupancy prediction using octree queries. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 79618–79641. Curran Associates, Inc., 2024.
- [32] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. PanopticFusion: Online volumetric semantic mapping at the level of stuff and things. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212, 2019.
- [33] Felix Petersen, Philipp Krähenbühl, and Vladlen Koltun. Differentiable top-k classification learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.
- [34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017.
- [35] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. LMSCNet: Lightweight multiscale 3d semantic completion. In *Proceedings of the 2020 International Conference on 3D Vision (3DV)*, pages 111–119, 2020.
- [36] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 190–198, 2017.
- [37] Keyu Tian, Yi Jiang, qishuai diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. In *The Eleventh International Conference on Learning Representations*, 2023.
- [38] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 84839–84865. Curran Associates, Inc., 2024.
- [39] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3D: A large-scale 3d occupancy prediction benchmark for autonomous driving. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 64318–64330. Curran Associates, Inc., 2023.
- [40] Guoqing Wang, Zhongdao Wang, Pin Tang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. OccGen: generative multi-modal 3d occupancy prediction for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 95–112. Springer, 2024.
- [41] JiaBao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, JIE YANG, Wei Liu, Qibin Hou, and Ming-Ming Cheng. Opus: Occupancy prediction using a sparse set. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [42] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. SurroundOcc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21729–21740, 2023.
- [43] Yuqi Wu, Wenzhao Zheng, Sicheng Zuo, Yuanhui Huang, Jie Zhou, and Jiwen Lu. EmbodiedOcc: Embodied 3d occupancy prediction for vision-based online scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

- [44] Zhiwen Yan, Weng Fei Low, Yu Chen, and Gim Hee Lee. Multi-scale 3d gaussian splatting for anti-aliased rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20923–20931, 2024.
- [45] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 21875–21911. Curran Associates, Inc., 2024.
- [46] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J. Guibas. GSPN: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3942–3951, 2019.
- [47] Hongxiao Yu, Yuqi Wang, Yuntao Chen, and Zhaoxiang Zhang. Monocular occupancy prediction for scalable indoor scenes. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, Cham, 2024. Springer Nature Switzerland.
- [48] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-Splatting: Alias-free 3d gaussian splatting. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19447–19456, 2024.
- [49] Dingyuan Zhang, Dingkan Liang, Hongcheng Yang, Zhikang Zou, Xiaoqing Ye, Zhe Liu, and Xiang Bai. SAM3D: zero-shot 3d object detection via the segment anything model. *Science China Information Sciences*, 67(4), 2024.
- [50] Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9433–9443, 2023.