A Appendix / Supplemental Material

In the appendix, we mainly provide implementation details and more experiment results.

A.1 Datasets and Metrics

Datasets. We evaluate our CGFormer on two datasets: SemanticKITTI [1] and SSC-Bench-KITTI-360 [9]. These datasets are derived from the KITTI Odometry [4] and KITTI-360 [11] Benchmarks, respectively. The evaluation focuses on a specific spatial volume: $51.2m$ in front of the car, $25.6m$ to the left and right sides, and $6.4m$ above the car. Voxelization of this volume results in a set of 3D voxel grids with a resolution of $256 \times 256 \times 32$, where each voxel measures $0.2m \times 0.2m \times 0.2m$. SemanticKITTI provides RGB images with dimensions of 1226×370 as inputs, encompassing 20 unique semantic classes (19 semantic classes and 1 free class). The dataset includes 10 sequences for training, 1 sequence for validation, and 11 sequences for testing. SSC-Bench-KITTI-360 [9] offers 7 sequences for training, 1 sequence for validation, and 1 sequence for testing. It contains 19 unique semantic classes (18 semantic classes and 1 free class), with input RGB images having a resolution of 1408×376.

Metrics. Following previous methods [3, 10, 6], we report the intersection over union (IoU) and mean IoU (mIoU) metrics for occupied voxel grids and voxel-wise semantic predictions, respectively. The interplay between IoU and mIoU offers a comprehensive perspective on the model's effectiveness in capturing both geometry and semantic aspects of the scene.

A.2 Implementation Details

Network Structures. Consistent with previous researches [6, 3, 21], we utilize a 2D UNet based on a pretrained EfficientNetB7 [18] as the image backbone. The CGVT generates a 3D feature volume with dimensions of $128 \times 128 \times 16$ and 128 channels. The numbers of deformable attention layers for cross-attention and self-attention are 3 and 2 respectively. We use 8 sampling points around each reference point for the cross and self-attention head. The voxel-based branch of the LGE comprises 3 stages with 2 residual blocks [5] each. SwinT [13] is employed as the 2D backbone in the TPV-based branch. Both are followed by feature pyramid networks (FPNs) [12] to aggregate multi-scale features for dynamic fusion. The final prediction has dimensions of $128 \times 128 \times 16$ and is upsampled to $256 \times 256 \times 32$ through trilinear interpolation to align the resolution with the ground truth.

Training Setup. We train CGFormer for 25 epochs on 4 NVIDIA 4090 GPUs, with a batch size of 4. It approximately consumes 19 GB of GPU memory on each GPU during the training phase. We employ the AdamW [14] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and set the maximum learning rate to 3×10^{-4}. The cosine annealing learning rate strategy is adopted for the learning rate decay, where the cosine warmup strategy is applied for the first 5% iterations.

A.3 Results Using Monocular Inputs

In alignment with previous methods [10, 7], we evaluate the performance of our CGFormer using only a monocular RGB image as input. We replace the depth estimation network with AdaBins [2] and present the results on the semantickitti validation set in the table 2. To better demonstrate the advantage of our CGFormer, we also include the results of VoxFormer, Symphonize, and OccFormer. Compared to the stereo-based methods when using only a monocular image (VoxFormer, Symphonize), CGFormer achieves superior performance in terms of both IoU and mIoU. Furthermore, our method also surpasses OccFormer, the state-of-the-art monocular method.

Table 1: The performance of the CGFormer with more lightweight backbone networks.

Backbone Networks	IoU	mIoU	Parameters	Training Memory
EfficientNetB7, Swin Block	**45.99**	**16.87**	122.42	19330
ResNet50, Swin Block	45.99	16.79	80.46	19558
ResNet50, ResBlock	45.86	16.85	**54.8**	**18726**

1

Table 2: Comparison of the performance using monocular inputs. For stereo-based methods, we replace the MobileStereoNet [16] with Adabins [2].

Method	IoU	mIoU
VoxFormer-S [10]	38.68	10.67
VoxFormer-T [10]	38.08	11.27
Symphonize [7]	38.37	12.20
OccFormer [24]	36.50	13.46
CGFormer (ours)	**41.82**	**14.06**

Table 3: Comparison of training memory and inference time with SOTA methods on the and SemanticKITTI test set. These metrics were measured on the NVIDIA 4090 GPU.

Method	TPVFormer [6]	OccFormer [24]	VoxFormer [10]	Symphonize [7]	StereoScene [8]	CGFormer (ours)
Training Meomry (M)	18564	18080	18725	17757	19000	19330
Inference Time (ms)	207	199	204	216	258	205
IoU	34.25	34.53	42.95	42.19	43.34	**44.41**
mIoU	11.26	12.20	12.20	15.04	15.36	**16.63**

A.4 Reults with More Lightweight Backbone Networks

We reanalyze the components of CGFormer, finding that replacing EfficientNetB7, used as the image backbone, and the Swin blocks, used in the TPV branch backbone, with more lightweight ResNet50 and residual blocks, respectively, can significantly reduce the number of parameters of our network. Besides, we also remove the predefined parameters as we find it doesn't influence the final performance. The results on the semantickitti validation set are presented in the Table 1. Compared to the original architecture, CGFormer maintains stable performance regardless of the backbone networks used for the image encoder and TPV branch encoder, underscoring its effectiveness, robustness, and potential.

A.5 Additional Quantitative Results

For more comprehensive comparison, we list the results with input modality and image backbones in Table 4 and Table 5. Table 6 presents the comparison results of CGFormer with the state-of-the-art methods on the SemanticKITTI validation set. CGFormer outperforms all other methods in terms of both IoU and mIoU. Additionally, it ranks either first or second on most of the classes, demonstrating consistent performance across various semantic categories, as indicated in previous tables.

A.6 Computational Cost

In Table 3, we display the training memory and inference time of CGFormer, along with those of the comparison methods. Additionally, the table includes the corresponding IoU and mIoU metrics for comprehensive comparison. As shown in the table, CGFormer achieves the best performance in terms of both IoU and mIoU, with comparable training memory and inference time.

A.7 Additional Qualitative Results

We offer additional visualization results in Fig.2 and Fig.3. These examples are randomly selected from the SemanticKITTI [1] validation set.

A.8 Failure Cases

We provide two failure cases in Fig. 1.

A.9 Limitations

While CGFormer exhibits strong performance on benchmarks, but the accuracy on most of the categories (*e.g.*, person, bicyclist, other vehicle) is unsatisfactory. Improving the performance on these instances could be beneficial for the downstream application tasks. Furthermore, there is a need to explore designing depth estimation networks under multi-view scenarios to extend the geometry-aware view transformation to these scenes. Despite these limitations, we are confident that CGFormer will contribute to advancing the field of 3D perception.

2

Table 4: Quatitative results on SemanticKITTI [1] test set. * represents the reproduced results in [6, 24]. The best and the second best results are in **bold** and underlined, respectively. Our CGFormer outperforms temporal stereo-based (Stereo-T) methods or those methods with larger image backbones in terms of IoU and mIoU.

Method	Input	Image Backbone	IoU	mIoU	road	sidewalk	parking	other-grnd.	building	car	truck	bicycle	motorcycle	other-veh.	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traf.-sign
MonoScene* [3]	Mono	EfficientNetB7	34.16	11.08	54.70	27.10	24.80	5.70	14.40	18.80	3.30	0.50	0.70	4.40	14.90	2.40	19.50	1.00	1.40	0.40	11.10	3.30	2.10
TPVFormer [6]	Mono	EfficientNetB7	34.25	11.26	55.10	27.20	27.40	6.50	14.80	19.20	3.70	1.00	0.50	2.30	13.90	2.60	20.40	1.10	2.40	0.30	11.00	2.90	1.50
SurroundOcc [21]	Mono	EfficientNetB7	34.72	11.86	56.90	28.30	30.20	6.80	15.20	20.60	1.40	1.60	1.20	4.40	14.90	3.40	19.30	1.40	2.00	0.10	11.30	3.90	2.40
OccFormer [24]	Mono	EfficientNetB7	34.53	12.32	55.90	30.30	31.50	6.50	15.70	21.60	1.20	1.50	1.70	3.20	16.80	3.90	21.30	2.20	1.10	0.20	11.90	3.80	3.70
IAMSSC [22]	Mono	ResNet50	43.74	12.37	54.00	25.50	24.70	6.90	19.20	21.30	3.80	1.10	0.60	3.90	22.70	5.80	19.40	1.50	2.90	0.50	11.90	5.30	4.10
VoxFormer-S [10]	Stereo	ResNet50	42.95	12.20	53.90	25.30	21.10	5.60	19.80	20.80	3.50	1.00	0.70	3.70	22.40	7.50	21.30	1.40	2.60	0.20	11.10	5.10	4.90
VoxFormer-T [10]	Stereo-T	ResNet50	43.21	13.41	54.10	26.90	25.10	7.30	23.50	21.70	3.60	1.90	1.60	4.10	24.40	8.10	24.20	1.60	1.10	0.00	13.10	6.60	5.70
DepthSSC [23]	Stereo	ResNet50	**44.58**	13.11	55.64	27.25	25.72	5.78	20.46	21.94	3.74	1.35	0.98	4.17	23.37	7.64	21.56	1.34	2.79	0.28	12.94	5.87	6.23
Symphonize [7]	Stereo	MaskDINO	42.19	15.04	58.40	29.30	26.90	11.70	24.70	23.60	3.20	3.60	**2.60**	5.60	24.20	10.00	23.10	**3.20**	1.90	**2.00**	16.10	7.70	8.00
HASSC-S [19]	Stereo	ResNet50	43.40	13.34	54.60	27.70	23.80	6.20	21.10	22.80	4.70	1.60	1.00	3.90	23.80	8.50	23.30	1.60	4.00	0.30	13.10	5.80	5.50
HASSC-T [19]	Stereo-T	ResNet50	42.87	14.38	55.30	29.60	25.90	11.30	23.10	23.00	2.90	1.90	1.50	4.90	24.80	9.80	26.50	1.40	3.00	0.00	14.30	7.00	7.10
StereoScene [8]	Stereo	EfficientNetB7	43.34	15.36	61.90	31.20	30.70	10.70	24.20	22.80	2.80	3.40	2.40	6.10	23.80	8.40	27.00	2.90	2.20	0.50	16.50	7.00	7.20
H2GFormer-S [20]	Stereo	ResNet50	44.20	13.72	56.40	28.60	26.50	4.90	22.80	23.40	4.80	0.80	0.90	4.10	24.60	9.10	23.80	1.20	2.50	0.10	13.30	6.40	6.30
H2GFormer-T [20]	Stereo-T	ResNet50	43.52	14.60	57.90	30.40	30.00	6.90	24.00	23.70	5.20	0.60	1.20	5.00	25.20	10.70	25.80	1.10	0.10	0.00	14.60	7.50	9.30
MonoOcc-S [25]	Stereo	ResNet50	-	13.80	55.20	27.80	25.10	9.70	21.40	23.20	5.20	2.20	1.50	5.40	24.00	8.70	23.00	1.70	2.00	0.20	13.40	5.80	6.40
MonoOcc-L [25]	Stereo	InternImage-XL	-	15.63	59.10	30.90	27.10	9.80	22.90	23.90	7.20	4.50	2.40	7.70	25.00	9.80	26.10	2.80	4.70	0.60	16.90	7.30	8.40
CGFormer (ours)	Stereo	EfficientNetB7	44.41	**16.63**	**64.30**	**34.20**	**34.10**	**12.10**	**25.80**	**26.10**	4.30	3.70	1.30	2.70	24.50	**11.20**	**29.30**	1.70	3.60	0.40	**18.70**	**8.70**	**9.30**

Table 5: Quantitative results on SSCBench-KITTI360 test set. The results for counterparts are provided in [9]. The best and the second best results for all camera-based methods are in **bold** and underlined, respectively. The best results from the LiDAR-based methods are in red.

Method	Input	Image Backbone	IoU	mIoU	car	bicycle	motorcycle	truck	other-veh.	person	road	parking	sidewalk	other-grnd.	building	fence	vegetation	terrain	pole	traf.-sign	other-struct.	other-obj.
LiDAR-based methods																						
SSCNet [17]	LiDAR	-	53.58	16.95	31.95	0.00	0.17	10.29	0.00	0.07	65.70	17.33	41.24	3.22	44.41	6.77	43.72	28.87	0.78	0.75	8.69	0.67
LMSCNet [15]	LiDAR	-	47.35	13.65	20.91	0.00	0.00	0.26	0.58	0.00	62.95	13.51	33.51	0.20	43.67	0.33	40.01	26.80	0.00	0.00	3.63	0.00
Camera-based methods																						
MonoScene [3]	Mono	EfficientNetB7	37.87	12.31	19.34	0.43	0.58	8.02	2.03	0.86	48.35	11.38	28.13	3.32	32.89	3.53	26.15	16.75	6.92	5.67	4.20	3.09
TPVFormer [6]	Mono	EfficientNetB7	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	17.52	7.46	5.86	5.48	2.70
OccFormer [24]	Mono	EfficientNetB7	40.27	13.81	22.58	0.66	0.26	9.89	3.82	2.77	54.30	13.44	31.53	3.55	36.42	4.80	31.00	19.51	7.77	8.51	6.95	4.60
VoxFormer [10]	Stereo	ResNet50	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
IAMSSC [22]	Stereo	ResNet50	41.80	12.97	18.53	2.45	1.76	5.12	3.92	3.09	47.55	10.56	28.35	4.12	31.53	6.28	29.17	15.24	8.29	7.01	6.35	4.19
DepthSSC [23]	Stereo	ResNet50	40.85	14.28	21.90	2.36	4.30	11.51	4.56	2.92	50.88	12.89	30.27	2.49	37.33	5.22	29.61	21.59	5.97	7.71	5.24	3.51
Symphonies [7]	Stereo	MaskDINO	44.12	18.58	**30.02**	1.85	**5.90**	**25.07**	**12.06**	**8.20**	54.94	13.83	32.76	**6.93**	35.11	**8.58**	38.33	11.52	14.01	9.57	**14.44**	**11.28**
CGFormer (ours)	Stereo	EfficientNetB7	**48.07**	**20.05**	29.85	3.42	3.96	17.59	6.79	6.63	**63.85**	**17.15**	**40.72**	5.53	**42.73**	8.22	**38.80**	**24.94**	**16.24**	**17.45**	10.18	6.77

Table 6: Quatitative results on SemanticKITTI [1] validation set. * represents the reproduced results in [6, 24, 23]. The best and the second best results are in **bold** and underlined, respectively.

Method	Input	Image Backbone	IoU	mIoU	road	sidewalk	parking	other-grnd.	building	car	truck	bicycle	motorcycle	other-veh.	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traf.-sign
MonoScene* [3]	Mono	EfficientNetB7	36.86	11.08	56.52	26.72	14.27	0.46	14.09	23.26	6.98	0.61	0.45	1.48	17.89	2.81	29.64	1.86	1.20	0.00	5.84	4.14	2.25
TPVFormer [6]	Mono	EfficientNetB7	35.61	11.36	56.50	25.87	20.60	0.85	13.88	23.81	8.08	0.36	0.05	4.35	16.92	2.26	30.38	0.51	0.89	0.00	5.94	3.14	1.52
OccFormer [24]	Mono	EfficientNetB7	36.50	13.46	**58.85**	26.88	19.61	0.31	14.40	25.09	**25.53**	0.81	1.19	8.52	19.63	3.93	32.62	2.78	2.82	0.00	5.61	4.26	2.86
IAMSSC [22]	Mono	ResNet50	44.29	12.45	54.55	25.85	16.02	0.70	17.38	26.26	8.74	0.64	0.15	5.06	24.63	4.95	30.13	1.32	3.46	0.01	6.86	6.35	3.56
VoxFormer-S [10]	Stereo	ResNet50	44.02	12.35	54.76	26.35	15.50	0.70	17.65	25.79	5.63	0.59	0.51	3.77	24.39	5.08	29.96	1.78	3.32	0.00	7.64	7.11	4.18
VoxFormer-T [10]	Stereo-T	ResNet50	44.15	13.35	53.57	26.52	19.69	0.42	19.54	26.54	7.26	1.28	0.56	7.81	26.10	6.10	33.06	1.93	1.97	0.00	7.31	9.15	4.94
DepthSSC [23]	Stereo	ResNet50	45.84	13.28	55.38	27.04	18.76	0.92	19.23	25.94	6.02	0.35	1.16	7.50	26.37	4.52	30.19	2.58	**6.32**	0.00	8.46	7.42	4.09
Symphonize [7]	Stereo	MaskDINO	41.92	14.89	56.37	27.58	15.28	0.95	21.64	**28.68**	20.44	2.54	2.82	**13.89**	25.72	6.60	30.87	**3.52**	2.24	0.00	8.40	9.57	5.76
HASSC-S [19]	Stereo	ResNet50	44.82	13.48	57.05	28.25	15.90	1.04	19.05	27.23	9.91	0.92	0.86	5.61	25.48	6.15	32.94	2.80	4.71	0.00	6.58	7.68	4.05
HASSC-T [19]	Stereo-T	ResNet50	44.58	14.74	55.30	29.60	**25.90**	**11.30**	23.10	23.00	2.90	1.90	1.50	4.90	24.80	9.80	26.50	1.40	3.00	0.00	**14.30**	7.00	7.10
H2GFormer-S [20]	Stereo	ResNet50	44.57	13.73	56.08	29.12	17.83	0.45	19.74	28.21	10.00	0.50	0.47	7.39	26.25	6.80	34.42	1.54	2.88	0.00	7.24	7.88	4.68
H2GFormer-T [20]	Stereo-T	ResNet50	44.69	14.29	57.00	29.37	21.74	0.34	20.51	28.21	6.80	0.95	0.91	9.32	27.44	7.80	36.26	1.15	0.10	0.00	7.98	9.88	5.81
CGFormer (ours)	Stereo	EfficientNetB7	45.99	**16.87**	**65.51**	**32.31**	20.82	0.16	**23.52**	**34.32**	19.44	**4.61**	2.71	7.67	26.93	8.83	**39.54**	2.38	4.08	0.00	**9.20**	**10.67**	**7.84**

(a) RGB   (b) Ground Truth   (c) CGFormer (ours)

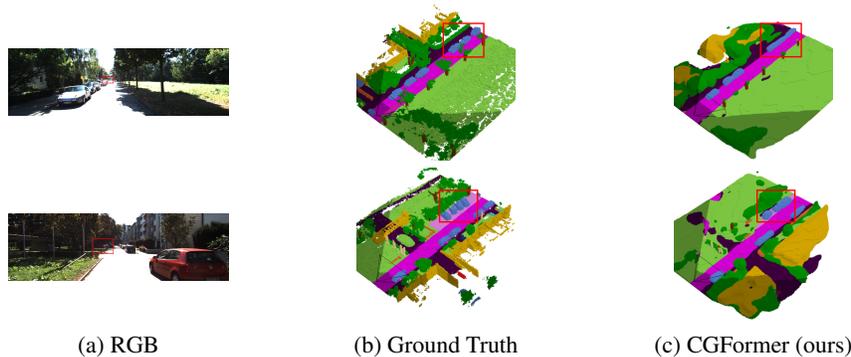



Figure 1: Failure cases.

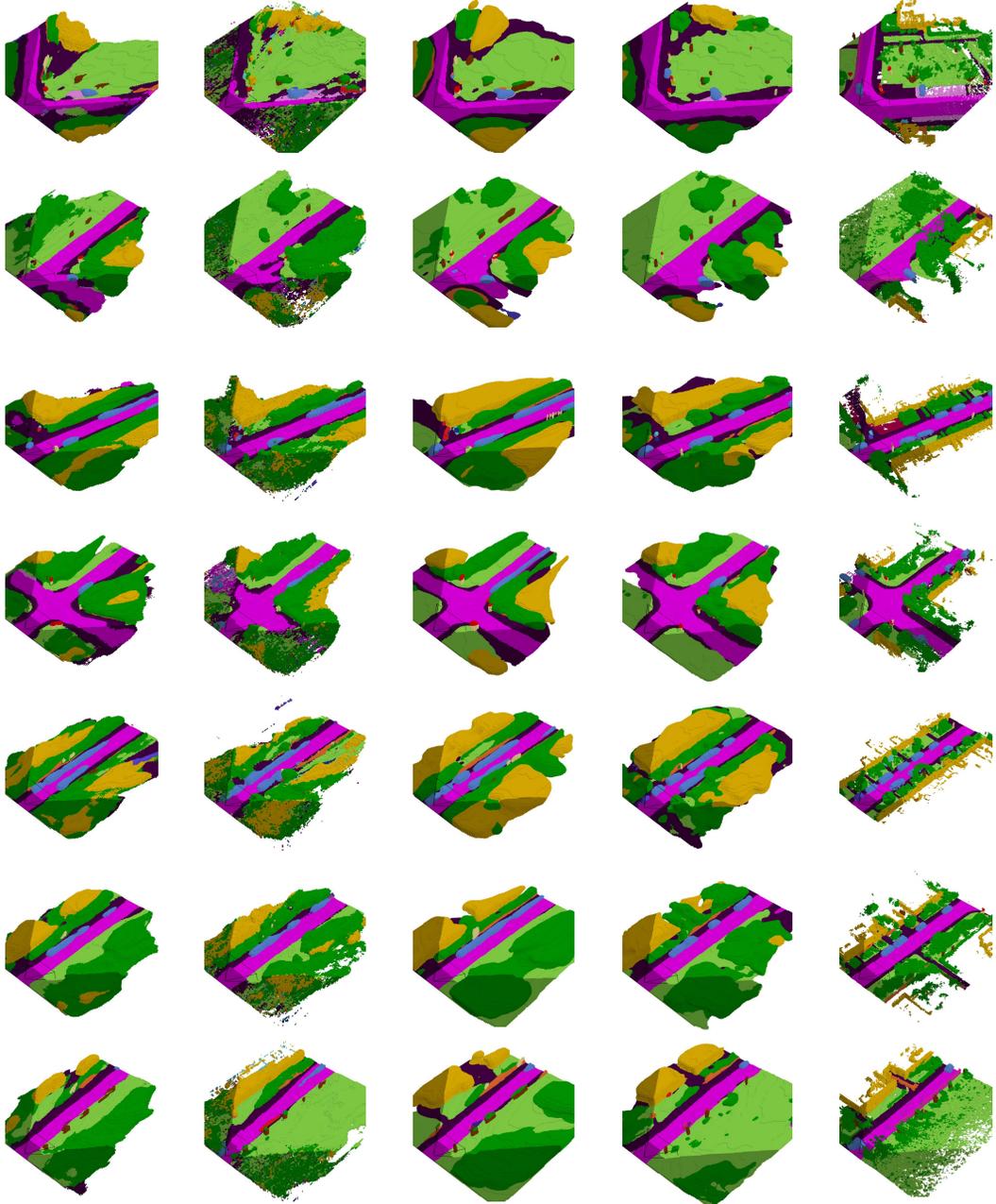
3

(a) MonoScene [3]   (b) VoxFormer [10]   (c) OccFormer [24]   (d) CGFormer (ours)   (e) Ground Truth

Figure 2: More qualitative comparison results on the SemanticKITTI [1] validation set.

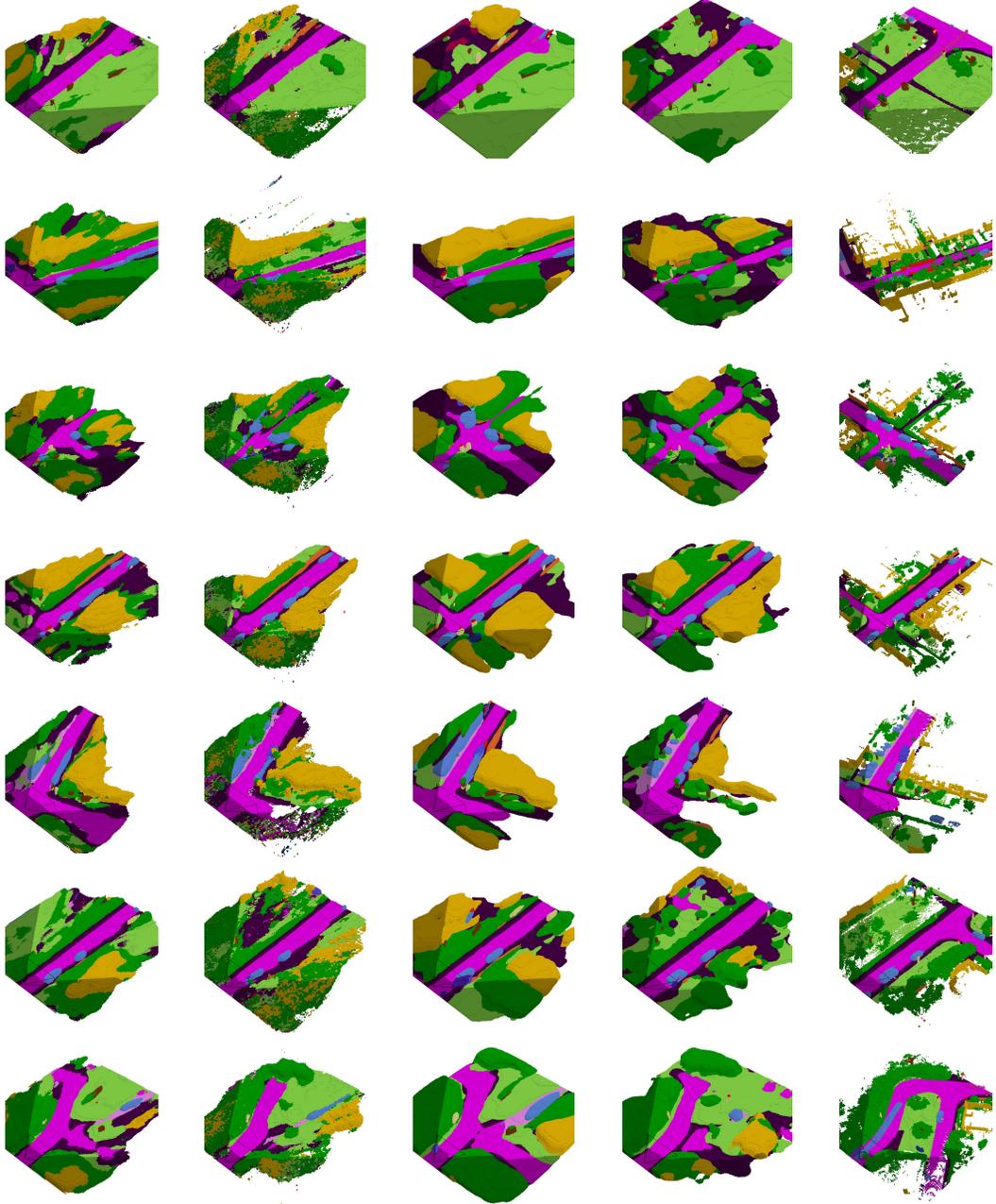
4

(a) MonoScene [3] (b) VoxFormer [10] (c) OccFormer [24] (d) CGFormer (ours) (e) Ground Truth

Figure 3: More qualitative comparison results on the SemanticKITTI [1] validation set.

5

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 1, 2, 3, 4, 5

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 1, 2

[3] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3981–3991, 2022. 1, 3, 4, 5

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] 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. 1, 2, 3

[7] Haoyi Jiang, Tianheng Cheng, Naiyu Gao, Haoyang Zhang, Wenyu Liu, and Xinggang Wang. Symphonize 3d semantic scene completion with contextual instance queries. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 3

[8] Bohan Li, Yasheng Sun, Xin Jin, Wenjun Zeng, Zheng Zhu, Xiaoefeng Wang, Yunpeng Zhang, James Okae, Hang Xiao, and Dalong Du. Stereoscene: Bev-assisted stereo matching empowers 3d semantic scene completion. *arXiv preprint arXiv:2303.13959*, 2023. 2, 3

[9] 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. *arXiv preprint arXiv:2306.09001*, 2023. 1, 3

[10] Yiming Li, Zhiding Yu, Christopher B. Choy, Chaowei Xiao, José M. Álvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 1, 2, 3, 4, 5

[11] 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*, pages 3292–3310, 2022. 1

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 1

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[15] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *Proceedings of the International Conference on 3D Vision*, pages 111–119, 2020. 3

[16] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2417–2426, 2022. 2

[17] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas A. Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 190–198, 2017. 3

[18] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114, 2019. 1

[19] Song Wang, Jiawei Yu, Wentong Li, Wenyu Liu, Xiaolu Liu, Junbo Chen, and Jianke Zhu. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. *arXiv preprint arXiv:2404.11958*, 2024. 3

6

[20] Yu Wang and Chao Tong. H2gformer: Horizontal-to-global voxel transformer for 3d semantic scene completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5722–5730, 2024. 3

[21] 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*, pages 21729–21740, 2023. 1, 3

[22] Haihong Xiao, Hongbin Xu, Wenxiong Kang, and Yuqiong Li. Instance-aware monocular 3d semantic scene completion. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 3

[23] Jiawei Yao and Jusheng Zhang. Depthssc: Depth-spatial alignment and dynamic voxel resolution for monocular 3d semantic scene completion. *arXiv preprint arXiv:2311.17084*, 2023. 3

[24] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9433–9443, 2023. 2, 3, 4, 5

[25] Yupeng Zheng, Xiang Li, Pengfei Li, Yuhang Zheng, Bu Jin, Chengliang Zhong, Xiaoxiao Long, Hao Zhao, and Qichao Zhang. Monoocc: Digging into monocular semantic occupancy prediction. *arXiv preprint arXiv:2403.08766*, 2024. 3

7