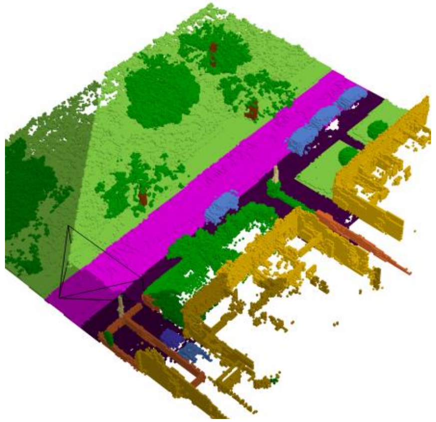
Context and Geometry Aware Voxel Transformer for Semantic Scene Completion

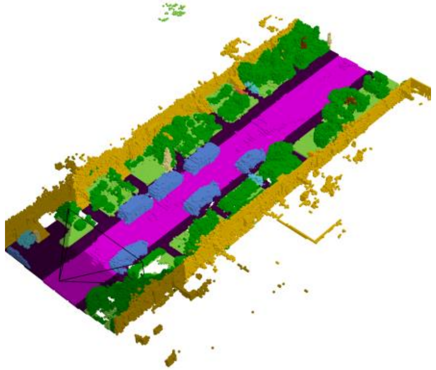
Zhu Yu, Runmin Zhang, Jiacheng Ying,
Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao†, Hui-Liang Shen†

https://github.com/pkqbajng/CGFormer

Background

Semantic Scene Completion



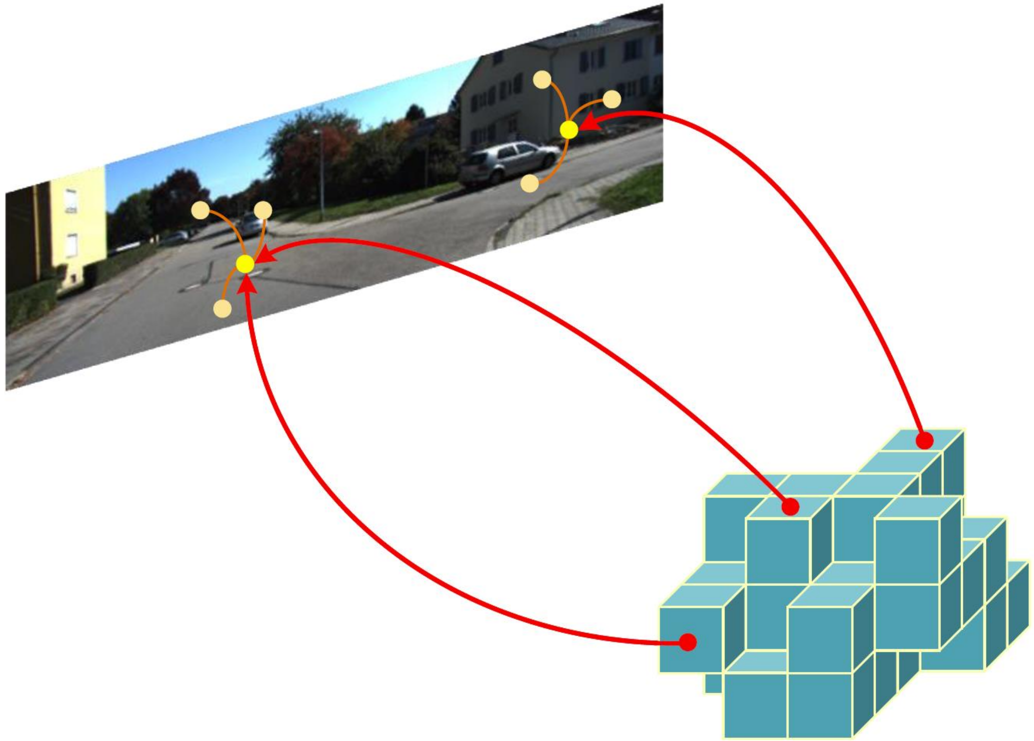


RGB Image

Occupancy

Background

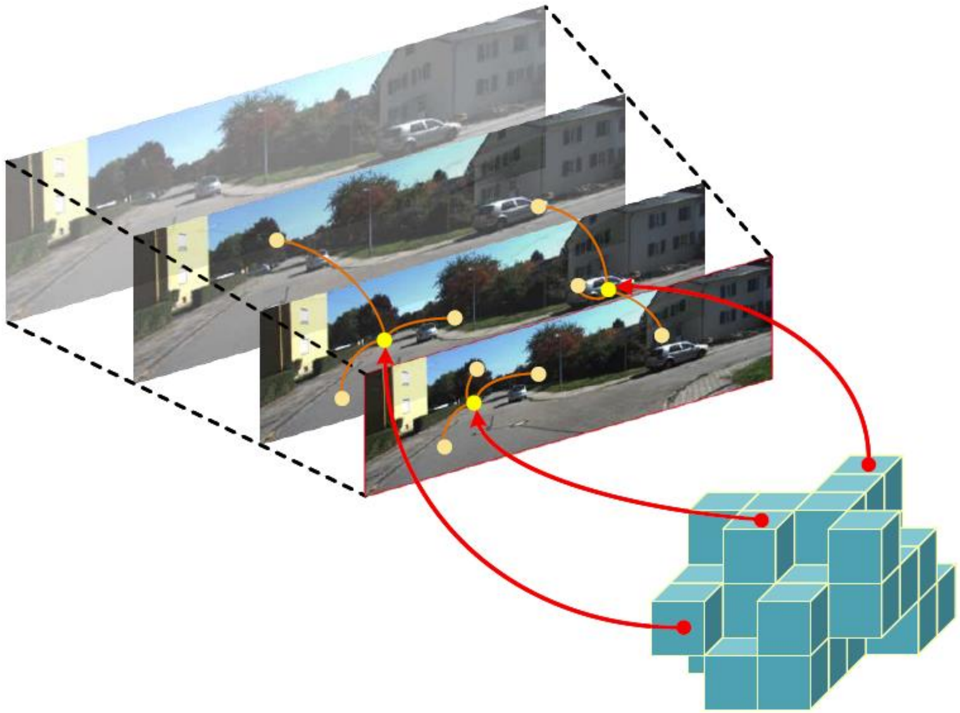
Previous Methods



Context-Independent Queries

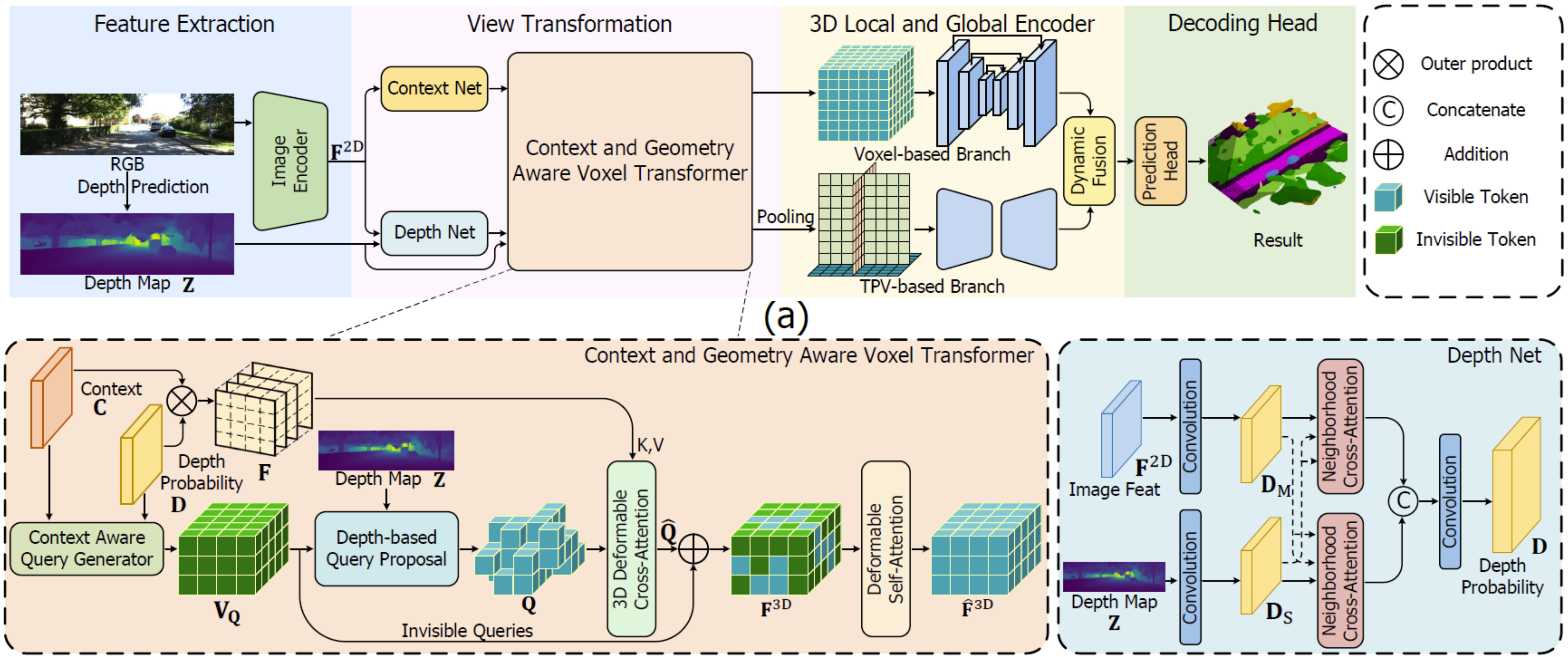
Motivation

Our Context and Geometry Aware Feature Aggregation

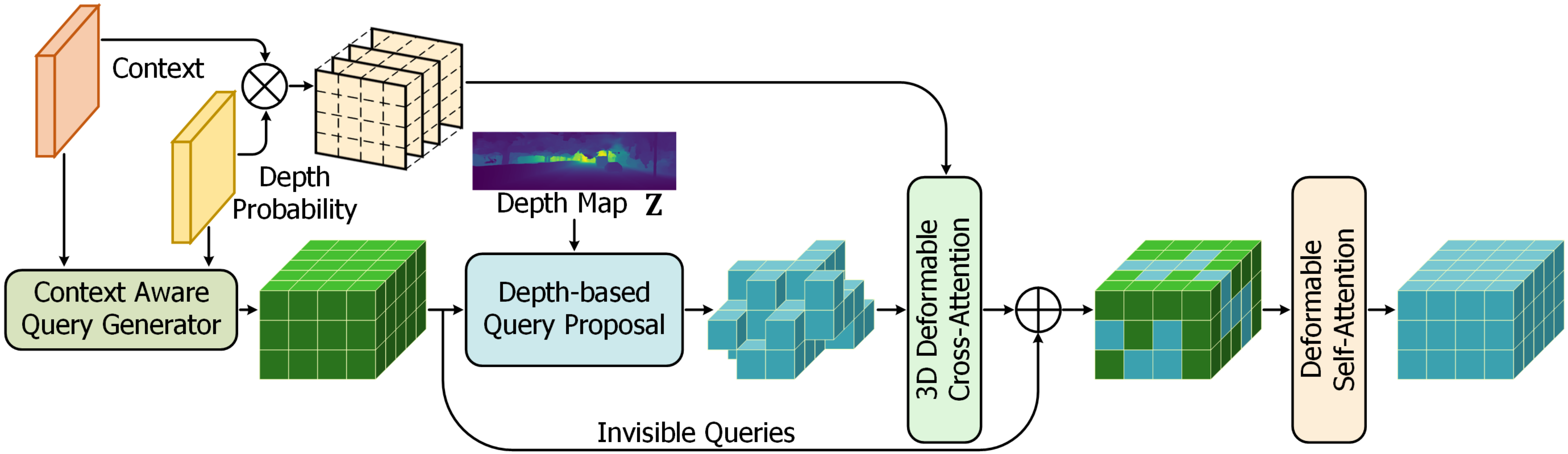


Context-Dependent Queries

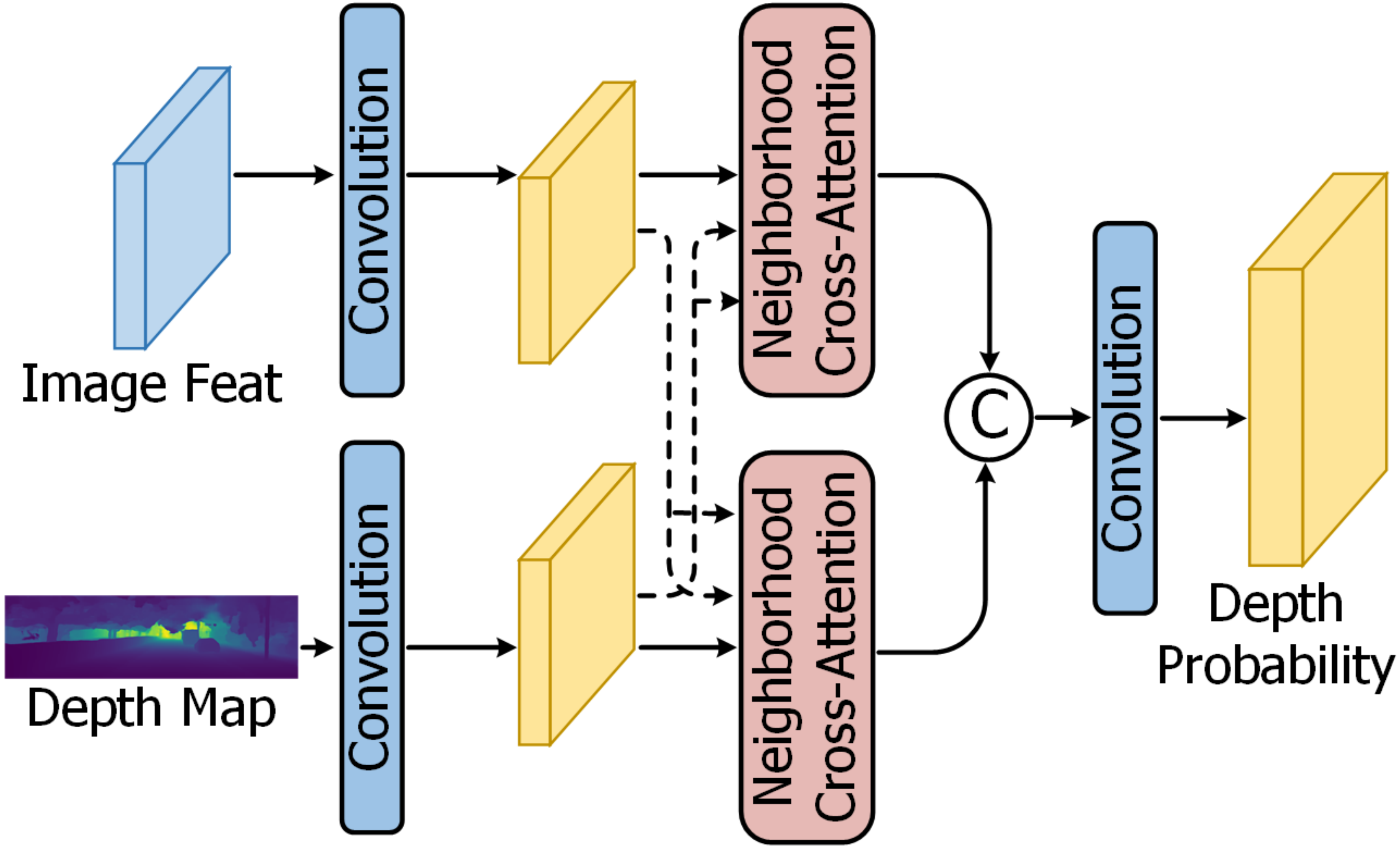
CGFormer: Overall Framework



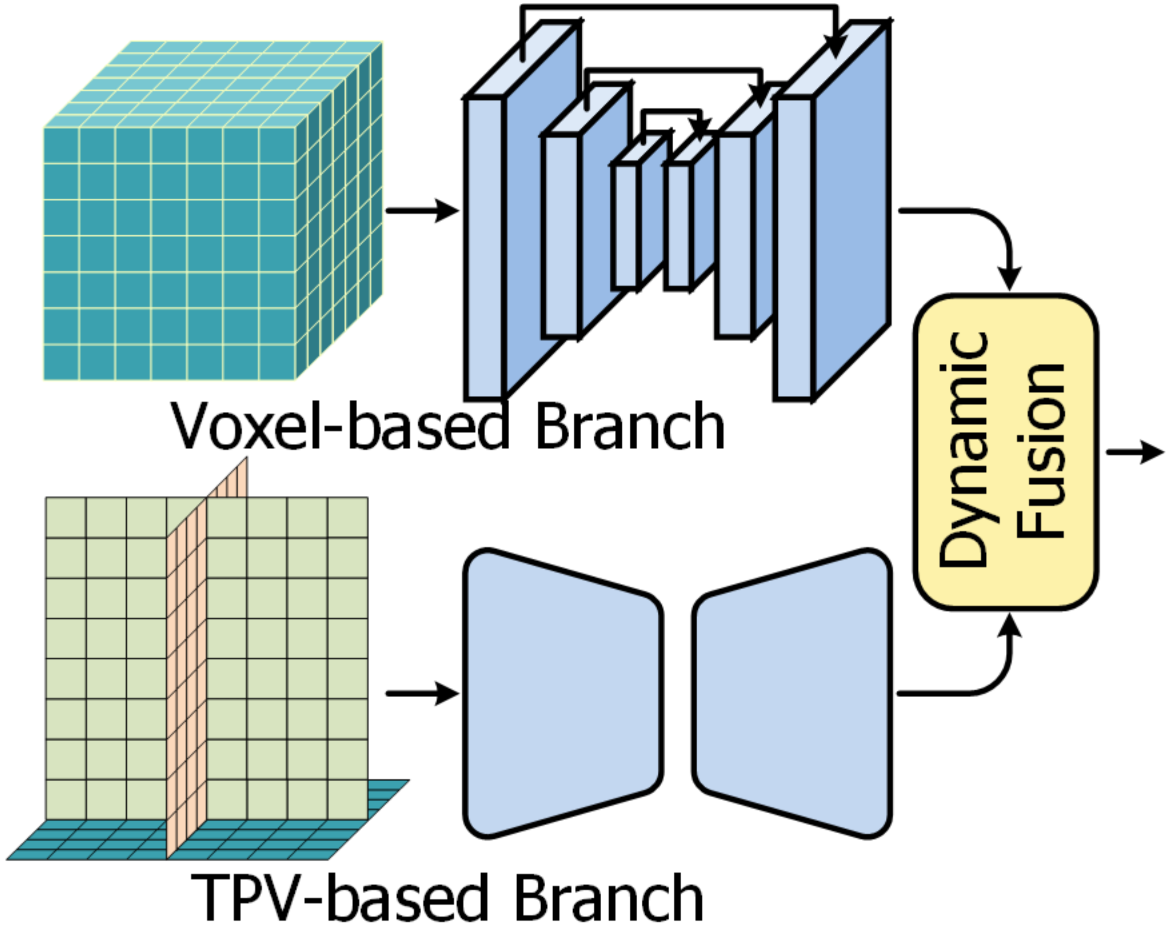
Context and Geometry Aware Voxel Transformer



Depth Refinement Block



3D Local and Global Encoder



Experiments

Quantitative Results on the SemanticKITTI test set

Table 1: Quatitative results on SemanticKITTI [1] test set. * represents the reproduced results in [13, 59]. The best and the second best results are in **bold** and underlined, respectively.

Method	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]	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 [13]	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 [47]	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 [59]	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 [49]	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 [23]	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 [23]	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 [55]	**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 [14]	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 [43]	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 [43]	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 [16]	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 [46]	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 [46]	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 [60]	-	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 [60]	-	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)	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**

Experiments

Quantitative Results on the KITTI-360 test set

Table 2: Quantitative results on SSCBench-KITTI360 test set. The results for counterparts are provided in [22]. 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	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 [40]	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 [38]	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]	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 [13]	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 [59]	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 [23]	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 [49]	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 [55]	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 [14]	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)	**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

Experiments

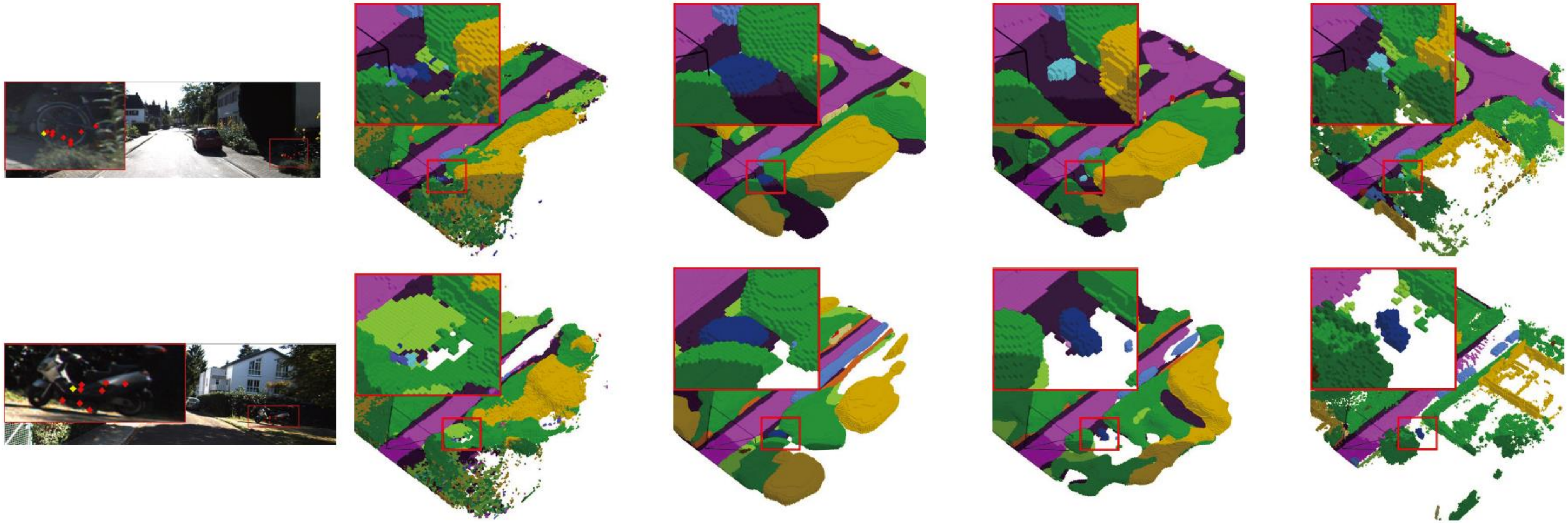
Ablation Studies

Table 3: Ablation study of the architectural components on SemanticKITTI [1] validation set. CGVT: context and geometry aware voxel transformer. LGE: local and global encoder. 3D-DCA: 3D deformable cross attention. CAQG: context aware query generator. LB: local voxel-based branch. $\mathcal{T}_{XY}, \mathcal{T}_{YZ}, \mathcal{T}_{XZ}$: planes of the TPV-based branch. DF: dynamic fusion. There are 32M predefined parameters.

Method	CGVT		LGE					IoU↑	mIoU↑	Params (M)	Memory (M)
	3D-DCA	CAQG	LB	$\mathcal{T}_{XY}$	$\mathcal{T}_{YZ}$	$\mathcal{T}_{XZ}$	DF				
Baseline								37.99	12.71	76.57	13222
(a)	✓							40.14	14.34	86.17	15150
(b)	✓	✓						42.86	15.60	86.19	15488
(c)	✓	✓	✓					44.84	16.41	93.78	17843
(d)	✓	✓	✓	✓	✓	✓		44.63	16.54	122.42	19188
(e)	✓	✓	✓	✓			✓	45.46	16.38	122.12	19024
(f)	✓	✓	✓		✓		✓	45.53	16.74	122.12	18912
(g)	✓	✓	✓			✓	✓	45.71	16.49	122.12	18912
(h)	✓	✓	✓	✓	✓	✓	✓	**45.99**	**16.87**	122.42	19330

Experiments

Visualizations of the sampling points of the context-dependent query

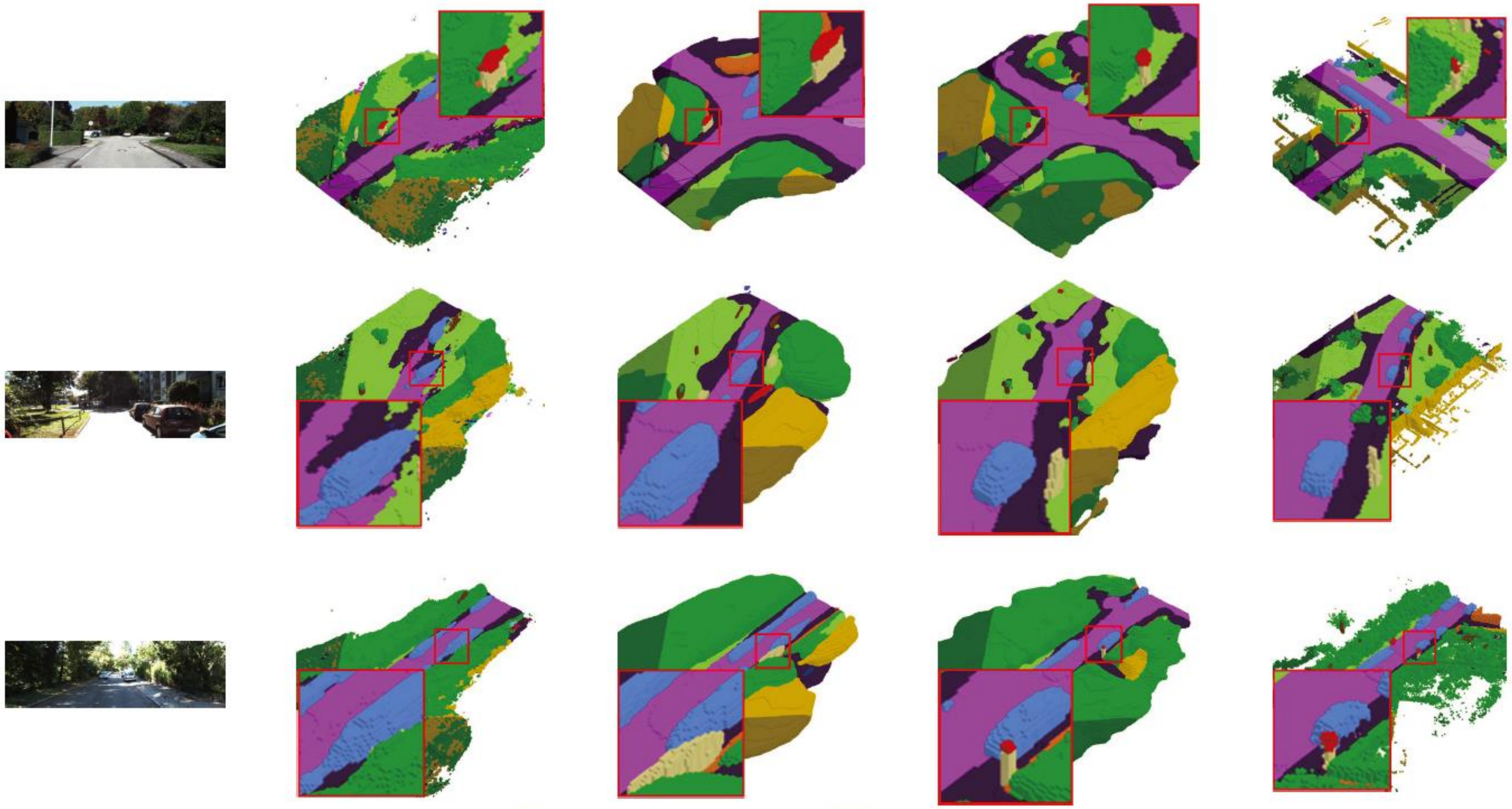


(a) RGB (b) VoxFormer [23] (c) OccFormer [59] (d) CGFormer (ours) (e) Ground Truth

Figure 3: Visualization of the sampling locations for different small objects. The yellow dot represents the query point, while the red dots indicate the locations of the deformable sampling points. The sampling points of the context-dependent query (a) tend to be distributed within the regions of interest. Beneficial from this, CGFormer achieve better performance than previous methods.

Experiments

Qualitative Visualization Results



(a) RGB (b) VoxFormer [23] (c) OccFormer [59] (d) CGFormer (ours) (e) Ground Truth

Conclusions

- We propose a context and geometry aware voxel transformer (CGVT) to improve the performance of semantic scene completion.
- We introduce a simple yet effective depth refinement block to enhance the accuracy of estimated depth probability with only introducing minimal computational burden
- We devise a 3D local and global encoder (LGE) to strengthen the semantic and geometric discriminability of the 3D volume.
- Benefiting from the aforementioned modules, our CGFormer attains state-of-the-art results with a mIoU of 16.63 and an IoU of 44.41 on SemanticKITTI, as well as a mIoU of 20.05 and an IoU of 48.07 on SSCBench-KITTI-360.

Thanks for watching！