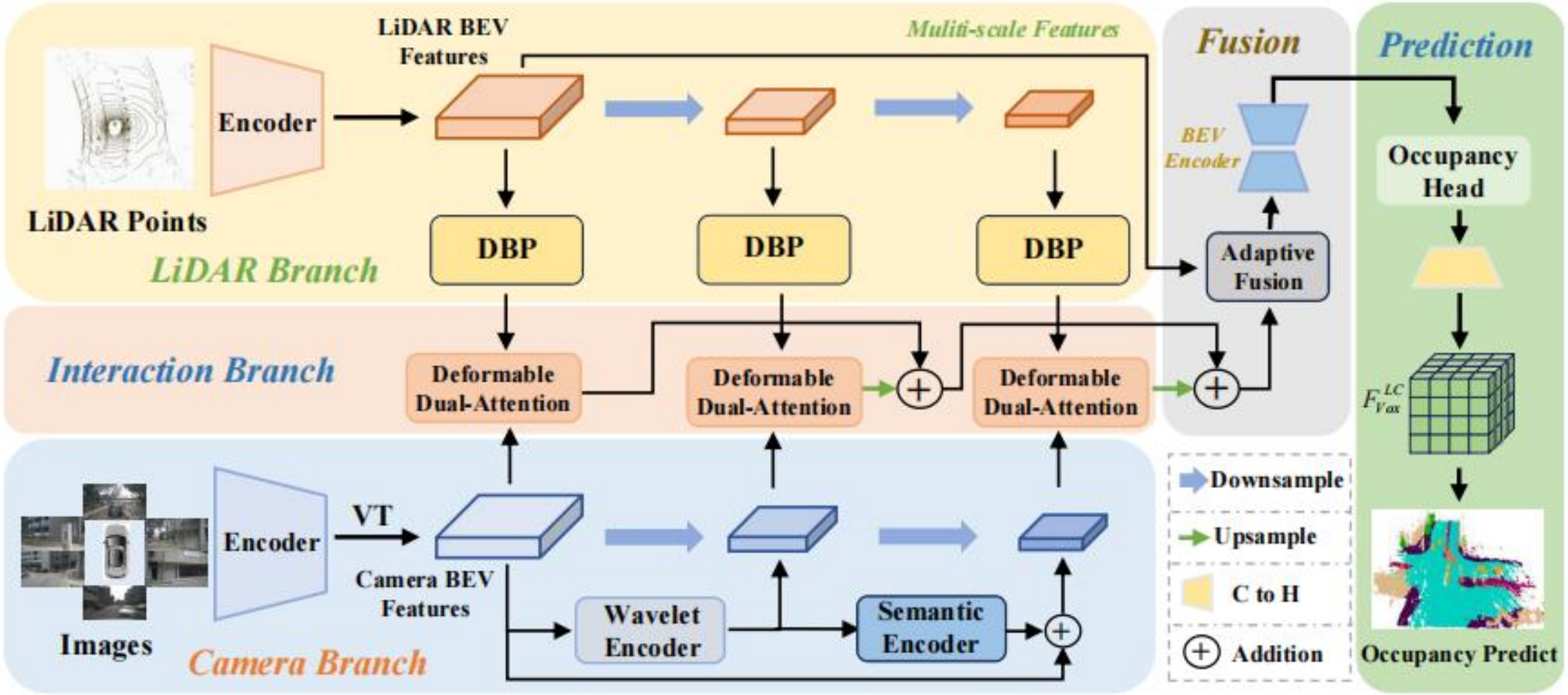
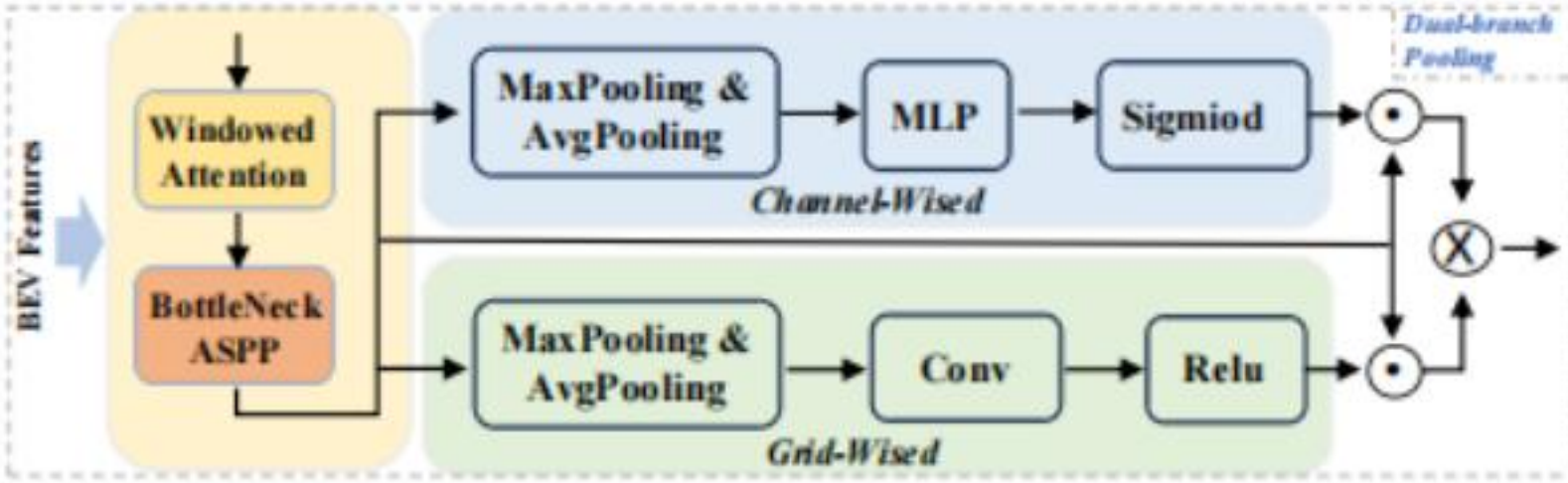
RIOcc: Efficient Cross-Modal Fusion Transformer with Collaborative Feature Refinement for 3D Semantic Occupancy Prediction

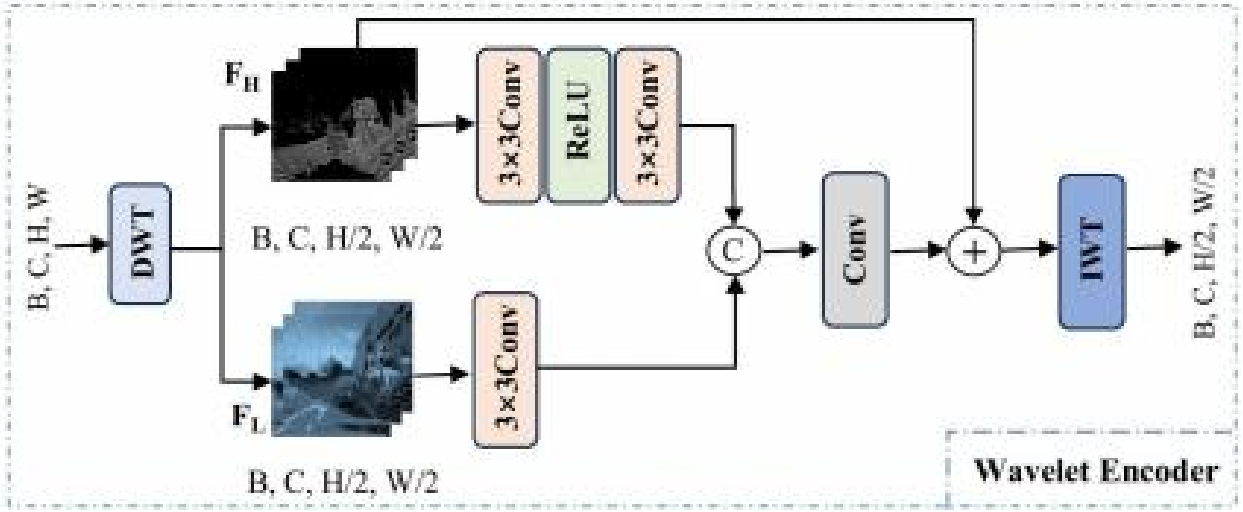
The overall framework of RIOcc.



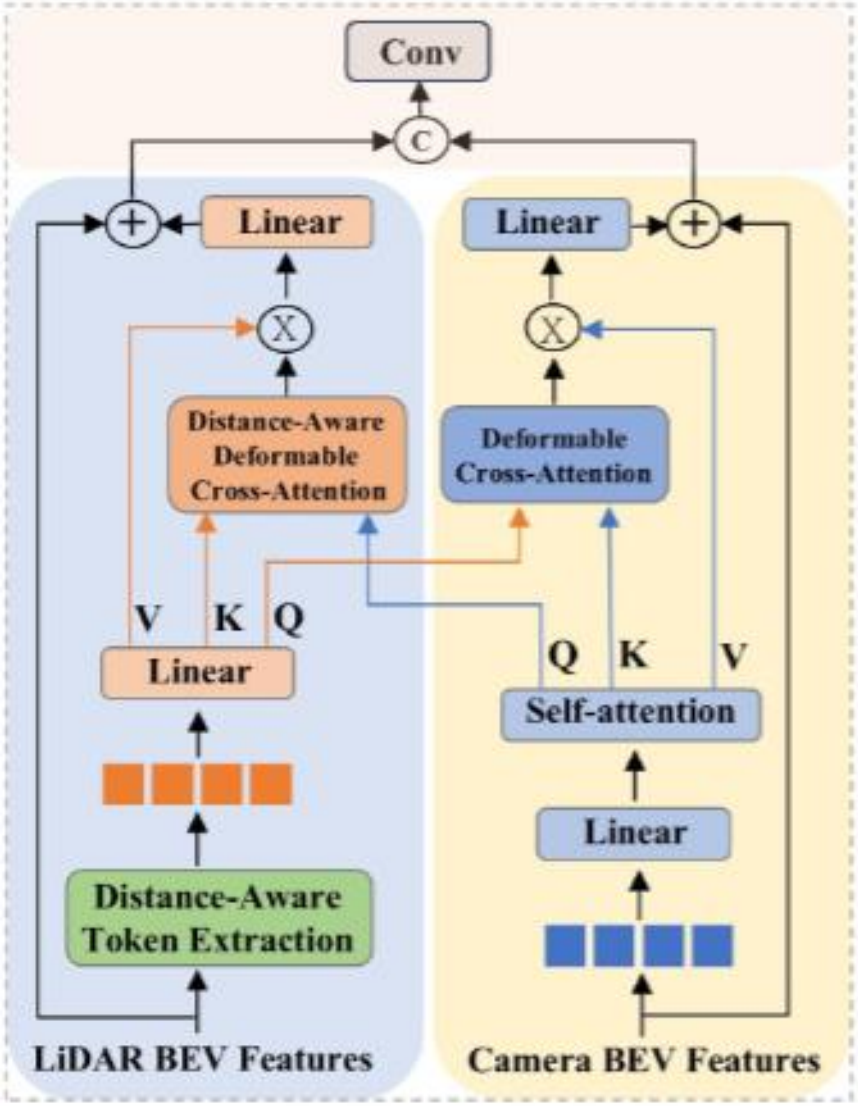
. The schema of Dual-branch Pooling (DBP).



Detailed structure diagram of the wavelet encoder.



Overview of Deformable Dual-Attention (DDA)

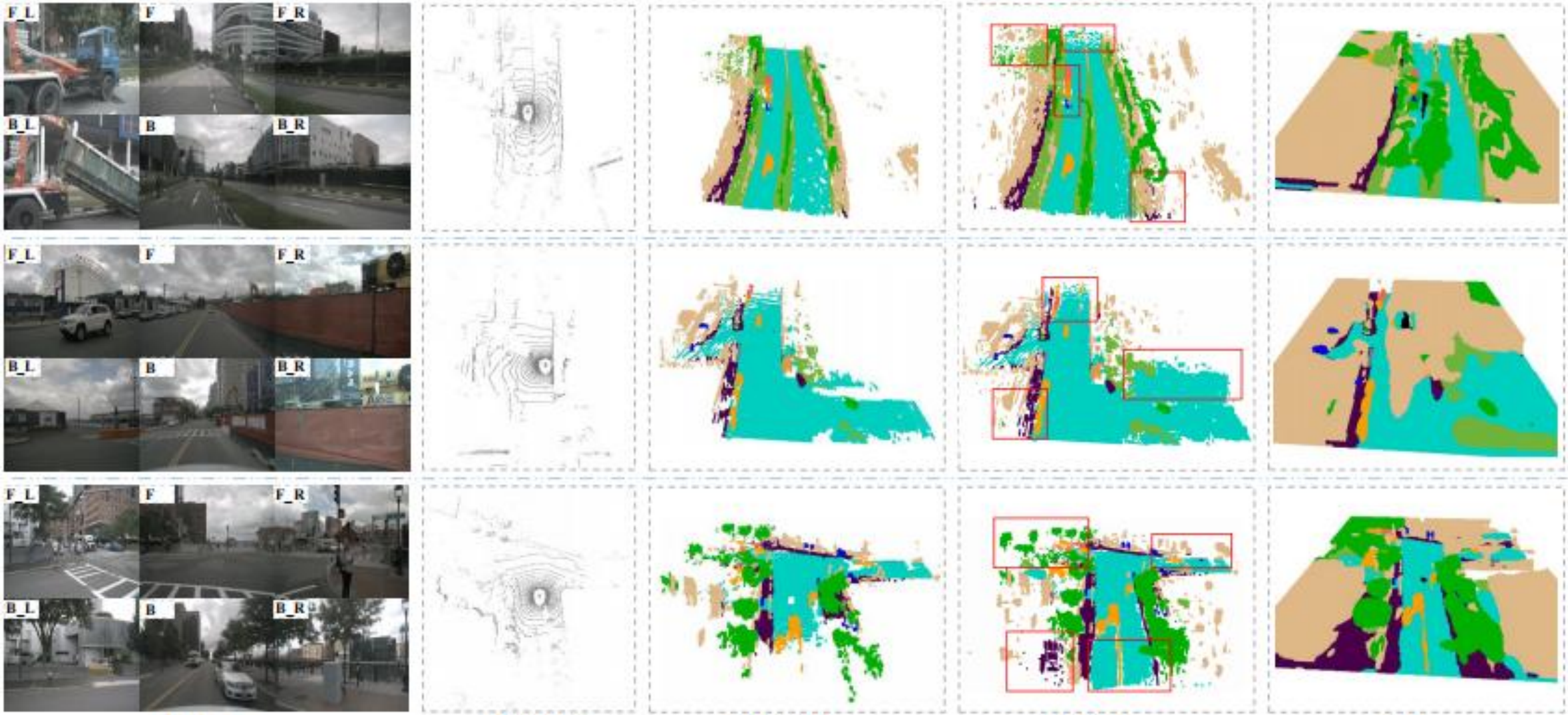


3D Occupancy prediction performance on the Occ3D-nuScenes dataset.

Method	Modality	Resolution	Image Backbone	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
LangOcc [3]	C	256×704	R50	11.84	0.00	3.10	90.00	6.30	14.20	0.40	10.80	6.20	9.00	3.80	10.70	43.70	2.23	9.50	26.40	19.60	26.40
FB-Occ [24]	C	256×704	R50	23.12	0.04	37.15	16.81	34.17	38.22	13.41	16.97	19.69	18.94	11.65	21.94	55.94	26.98	29.65	26.92	10.24	14.33
UniVision* [11]	C	256×704	R50	37.50	11.00	44.70	23.10	43.00	50.50	21.60	24.90	26.90	25.70	30.70	35.80	79.80	41.40	49.10	53.80	40.30	34.70
GEOcc* [43]	C	256×704	R50	43.64	**14.29**	51.27	31.11	46.13	55.09	29.12	30.46	30.99	35.47	35.20	41.82	84.00	47.00	55.52	59.50	50.03	44.82
TEOcc* [26]	C&R	256×704	R50	42.90	10.82	50.33	24.28	48.99	57.32	29.38	24.41	30.14	28.46	36.46	43.01	83.96	43.09	56.00	59.34	54.18	49.16
EFFOcc* [38]	C&L	256×704	R18	49.29	10.57	56.16	21.73	58.68	63.16	31.98	37.71	55.40	36.15	45.87	50.81	81.02	39.07	53.08	57.15	70.41	68.90
OccFormer [58]	C	900×1600	R101	21.93	5.94	30.29	12.32	34.40	19.17	14.44	16.45	17.22	9.27	13.90	26.34	50.99	30.96	34.66	22.73	6.76	6.97
RenderOcc [32]	C	900×1600	R101	26.11	4.84	31.72	10.72	27.67	36.45	13.87	18.20	17.67	17.84	21.19	23.25	63.20	36.42	46.21	44.26	19.58	20.72
TPVFormer [14]	C	900×1600	R101	28.34	6.67	39.20	14.24	41.54	46.98	19.21	22.64	17.87	14.54	30.20	35.51	56.18	33.65	35.69	31.61	19.97	16.12
CTF-Occ [45]	C	928×1600	R101-DCN	28.53	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.00
PanoOcc* [48]	C	640×1600	R101-DCN	42.13	11.67	50.48	29.64	49.44	55.52	23.29	33.26	30.55	30.99	34.43	42.57	83.31	44.23	54.40	56.04	45.94	40.40
OctreeOcc* [28]	C	900×1600	R101-DCN	44.02	11.96	51.70	29.93	53.52	56.77	30.83	33.17	30.65	29.99	37.76	43.87	83.17	44.52	55.45	58.86	49.52	46.33
OccFusion* [30]	C&L	900×1600	R101	46.79	11.65	47.81	32.07	57.27	57.51	31.80	40.11	47.35	33.74	45.81	50.35	78.79	37.17	44.36	53.36	61.18	63.20
OccFusion* [56]	C&L	900×1600	R101	48.74	12.35	51.77	33.01	54.56	57.65	33.99	43.03	48.35	35.54	41.22	48.55	83.00	**44.65**	57.13	60.01	62.46	61.25
RadOcc* [54]	C&L	512×1408	Swin-B	49.38	10.93	58.23	25.01	57.89	62.85	34.04	33.45	50.07	32.05	48.87	52.11	82.90	42.73	55.27	58.34	68.64	66.01
RIOcc* (Ours)	C&L	256×704	R50	**54.21**	11.82	**59.73**	**36.98**	**62.21**	**68.72**	**36.45**	**47.45**	**58.25**	**44.20**	**49.92**	**54.39**	**85.10**	44.60	**59.67**	**61.77**	**70.51**	**69.56**

3D Occupancy prediction performance on nuScenes-Occupancy validation set

Method	Modality	Resolution	Image Backbone	LiDAR Backbone	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [5]	C	900×1600	R101-DCN	-	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3	7.9	7.4	10.0	7.6
C-CONet [47]	C	900×1600	R50	-	20.1	12.8	13.2	8.1	15.4	17.2	6.3	11.2	10.0	8.3	4.7	12.1	31.4	18.8	18.7	16.3	4.8	8.2
SparseOcc [44]	C	704×256	R50	-	21.8	14.1	16.1	9.3	15.1	18.6	7.3	9.4	11.2	9.4	7.2	13.0	31.8	21.7	20.7	18.8	6.1	10.6
LMSCNet [35]	L	-	-	VoxelNet	27.3	11.5	12.4	4.2	12.8	12.1	6.2	4.7	6.2	6.3	8.8	7.2	24.2	12.3	16.6	14.1	13.9	22.2
JS3C-Net [50]	L	-	-	VoxelNet	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3	14.9	16.2	14.0	24.9
L-CONet [47]	L	-	-	VoxelNet	30.9	15.8	17.5	5.2	13.3	18.1	7.8	5.4	9.6	5.6	13.2	13.6	34.9	21.5	22.4	21.7	19.2	23.5
PointOcc [60]	L	-	-	VoxelNet	34.1	23.9	24.9	19.0	20.9	25.7	13.4	**25.6**	30.6	17.9	16.7	21.2	36.5	**25.6**	**25.7**	24.9	24.8	**29.0**
M-CONet [47]	C&L	900×1600	R50	VoxelNet	29.5	20.1	23.3	13.3	21.2	24.3	15.3	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
Co-Occ [31]	C&L	900×1600	R101	VoxelNet	30.6	21.9	26.5	16.8	22.3	27.0	10.1	20.9	20.7	14.5	16.4	21.6	36.9	23.5	25.5	23.7	20.5	23.5
OccGen [46]	C&L	896×1600	R50	VoxelNet	30.3	22.0	24.9	16.4	22.5	26.1	14.0	20.1	21.6	14.6	17.4	21.9	35.8	24.5	24.7	24.0	20.5	23.5
OccFusion [56]	C&L	900×1600	R101	VoxelNet	32.4	22.4	25.3	17.0	22.5	25.9	16.5	22.4	24.0	16.1	16.0	22.1	35.6	22.1	24.0	23.9	21.3	24.0
EFFOcc [38]	C&L	256×704	R18	VoxelNet	30.8	22.9	28.1	16.7	22.1	27.3	13.0	24.8	**36.2**	**22.6**	16.8	21.6	29.4	13.9	18.2	20.6	26.5	28.8
OccMamba [20]	C&L	900×1600	R50	VoxelNet	33.7	25.1	29.6	**20.2**	25.7	28.5	16.7	25.0	23.2	19.9	20.3	24.5	36.1	25.3	25.1	24.8	**27.7**	28.9
RIOcc (Ours)	C&L	256×704	R50	VoxelNet	**35.4**	**25.9**	**30.2**	19.8	**25.8**	**28.7**	**18.3**	24.8	31.8	21.8	**20.5**	**24.9**	**37.2**	24.5	25.5	**24.9**	27.0	28.8


F_L
F
F_R
B_L
B
B_R
Images
LiDAR
M-CONet
RIOcc
GT

Thank you for listening!