

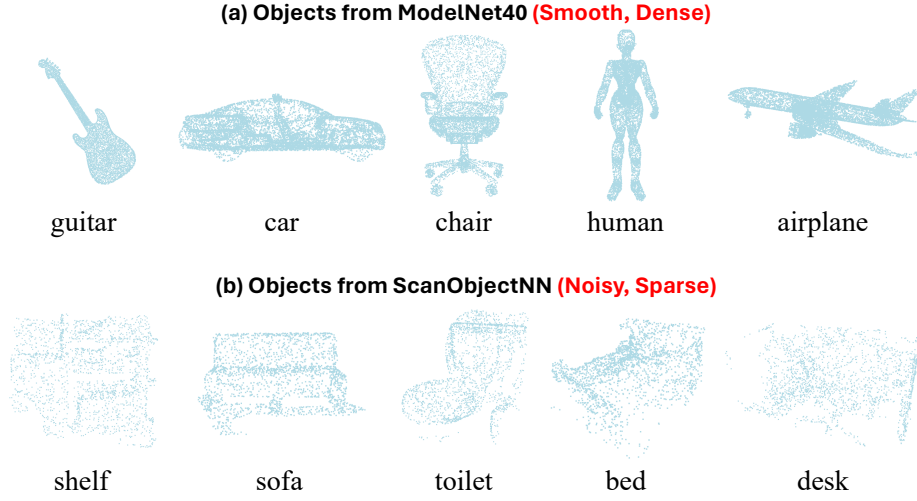


Figure 1: Visual comparison between objects from ModelNet40 and ScanObjectNN.



Figure 2: Examples of multi-view depth maps and their corresponding depth-aligned images.

Table 1: Comparison of 3D open-vocabulary object detection methods in the same setting as OV-3DET on ScanNet. 'Mean' represents the average precision across all 20 categories.

Methods	Mean	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink
OV-3DET	18.02	57.29	42.26	27.06	31.50	8.21	14.17	2.98	5.56	23.00	31.60
CoDA	19.32	68.09	44.04	28.72	44.57	3.41	20.23	5.32	0.03	27.95	45.26
OpenDlign (Ours)	19.27	58.30	41.95	34.13	32.25	12.49	18.12	3.89	4.15	25.41	34.53
Methods		bathtub	refrigerator	desk	nightstand	counter	door	curtain	box	lamp	bag
OV-3DET		56.28	10.99	19.27	0.77	0.31	9.59	10.53	3.78	2.11	2.71
CoDA		50.51	6.55	12.42	15.15	0.68	7.95	0.01	2.94	0.51	2.02
OpenDlign (Ours)		58.20	8.74	17.85	5.49	1.53	11.46	6.82	5.03	1.54	3.17