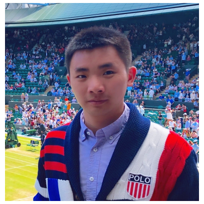
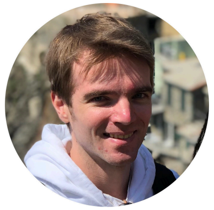
Neural Information
Processing Systems
(NeurIPS) 2019

Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds







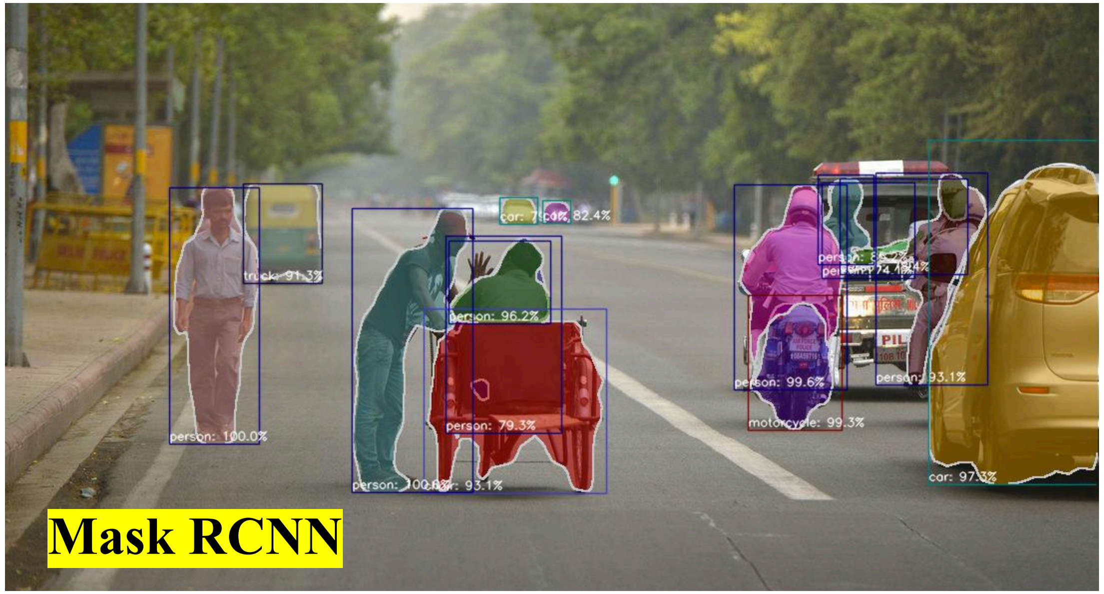
B. Yang, J. Wang, R. Clark, Q. Hu, S. Wang, A. Markham, N. Trigoni

UNIVERSITY OF OXFORD

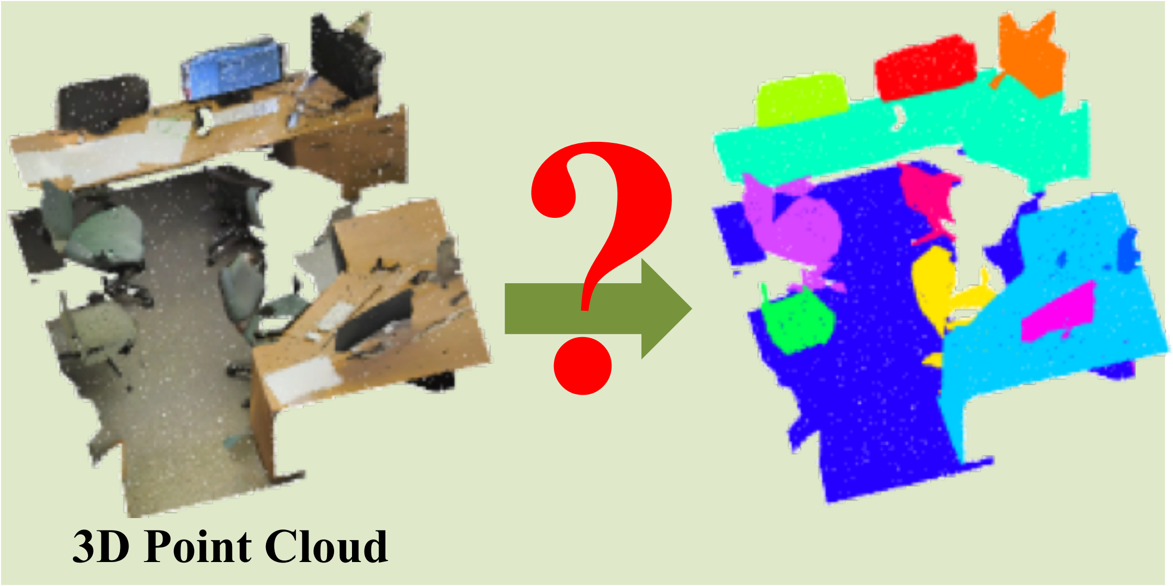
Imperial College London

Background:

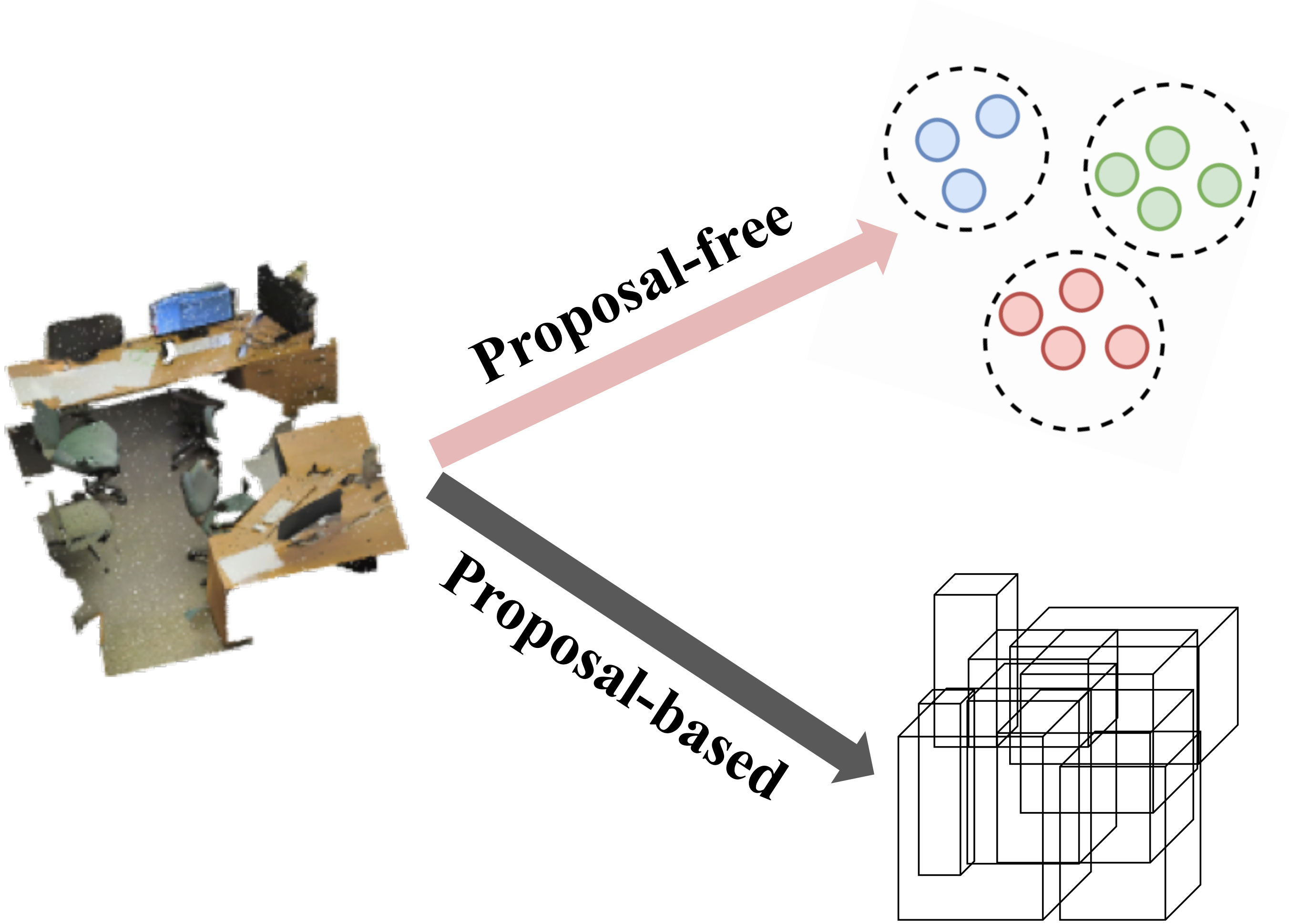
2D Instance Segmentation



3D Instance Segmentation



Background:



Limitations

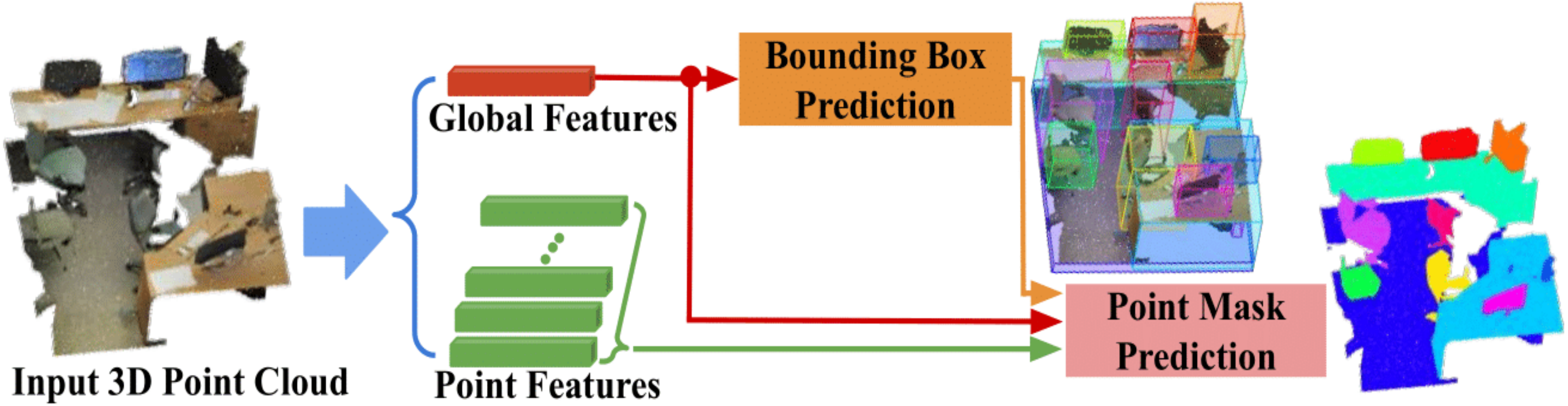
- **Low objectness**
- **Heavy post-processing (grouping)**

SGPN (CVPR'18); ASIS (CVPR'19); JSIS3D (CVPR'19); 3D-BEVIS (GCPR '19); MTML (ICCV'19); MASC (arXiv'19)

- **Two-stage training**
- **Heavy post-processing (NMS)**

3D-SIS (CVPR'19); GSPN (CVPR'19)

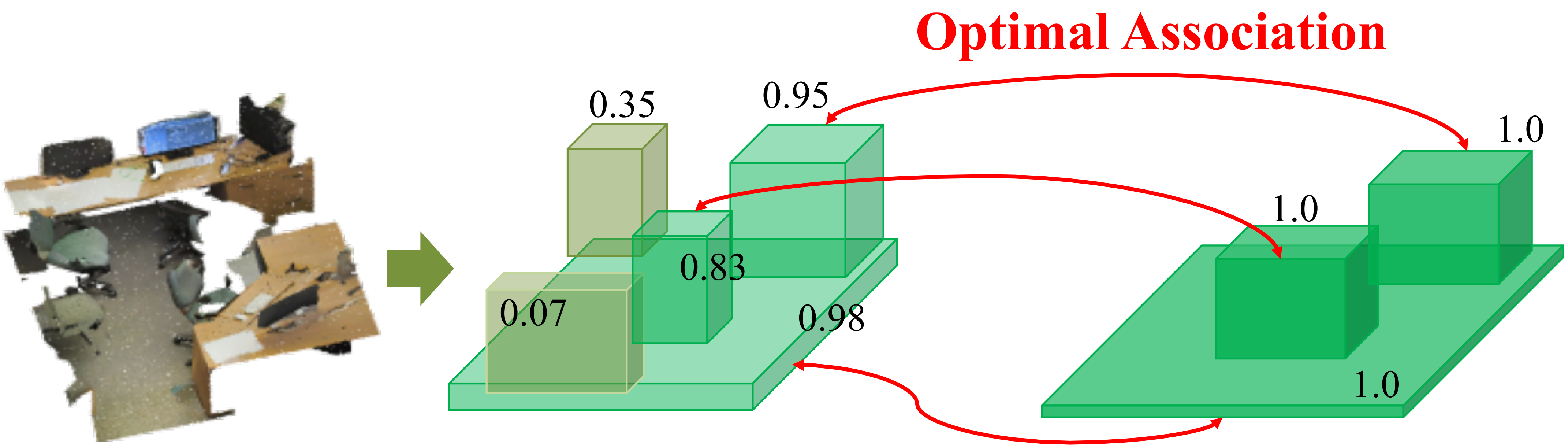
Our Method (3D-BoNet):



Highlights of our pipeline

- Each object is **uniquely detected and segmented**.
- The learnt 3D bounding boxes guarantee **high objectness**.
- It's **end-to-end trainable**, **no post-processing**, **and efficient.**

Our Method (3D-BoNet):



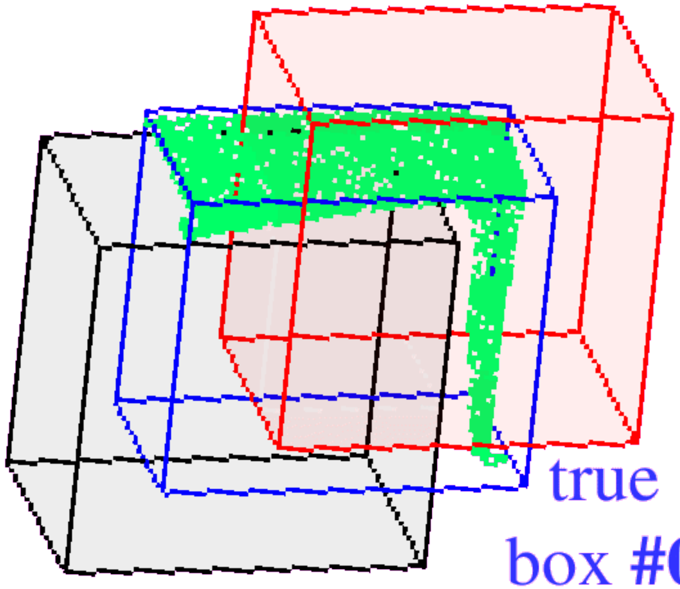
Input
Point Cloud

- Predicted bounding boxes
- Predicted bbox scores

✓ GT bounding boxes
✓ GT bbox scores

$$\boldsymbol{A} = \arg\min_{\boldsymbol{A}} \sum_{i=1}^{H}\sum_{j=1}^{T} \boldsymbol{C}_{i,j}\boldsymbol{A}_{i,j} \quad \text{subject to} \sum_{i=1}^{H}\boldsymbol{A}_{i,j} = 1, \sum_{j=1}^{T}\boldsymbol{A}_{i,j} \leq 1, j \in \{1..T\}, i \in \{1..H\}$$

Our Method (3D-BoNet):



Multiple criteria to match a pred bbox with a GT bbox

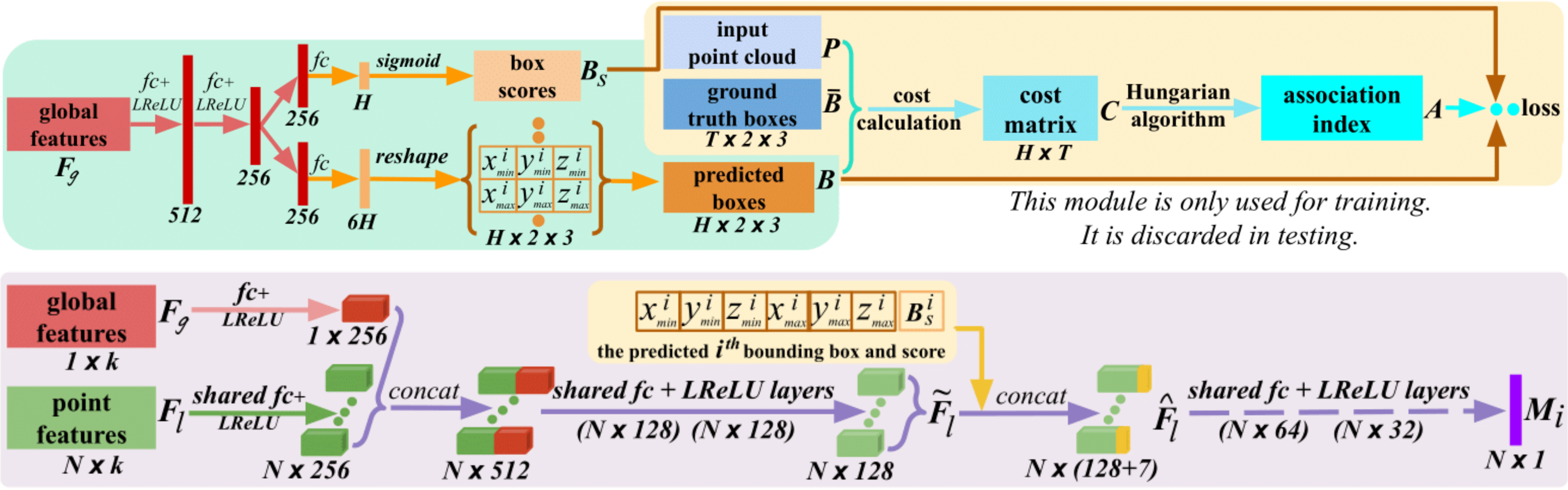
$$C^{ed}_{i,j} = \frac{1}{6}\sum(B_i - \bar{B}_j)^2$$

$$C^{sIoU}_{i,j} = \frac{-\sum_{n=1}^{N}(q_i^n * \bar{q}_j^n)}{\sum_{n=1}^{N} q_i^n + \sum_{n=1}^{N} \bar{q}_j^n - \sum_{n=1}^{N}(q_i^n * \bar{q}_j^n)}$$

$$C^{ces}_{i,j} = -\frac{1}{N}\sum_{n=1}^{N}\left[\bar{q}_j^n \log q_i^n + (1-\bar{q}_j^n)\log(1-q_i^n)\right]$$

$$C_{i,j} = C^{ed}_{i,j} + C^{sIoU}_{i,j} + C^{ces}_{i,j}$$

Our Method (3D-BoNet):



End-to-end training losses

$$\ell_{all} = \ell_{sem} + \ell_{bbox} + \ell_{bbs} + \ell_{pmask}$$

Quantitative Results:

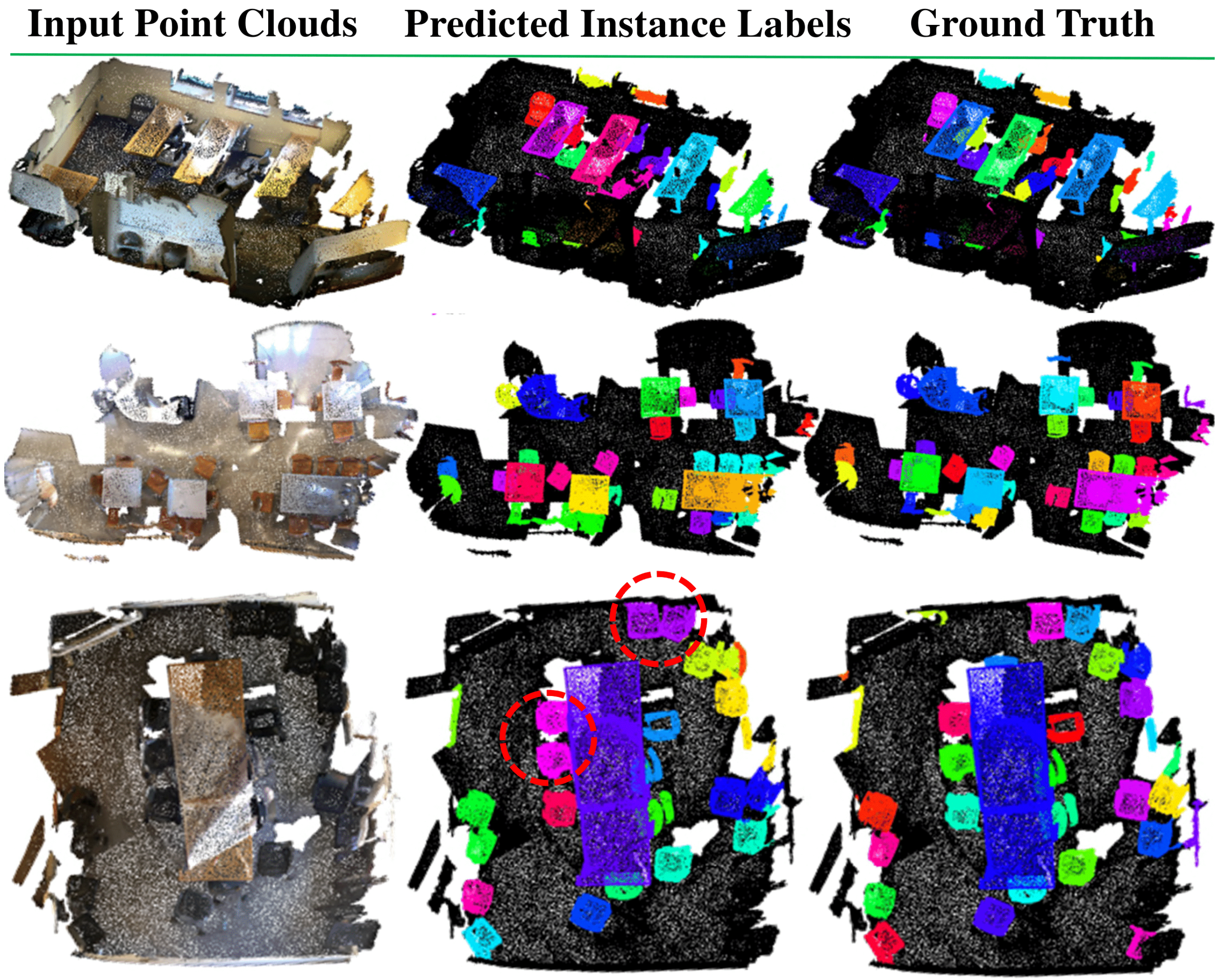
ScanNet Benchmark

Benchmarks ▾ Documentation About Submit

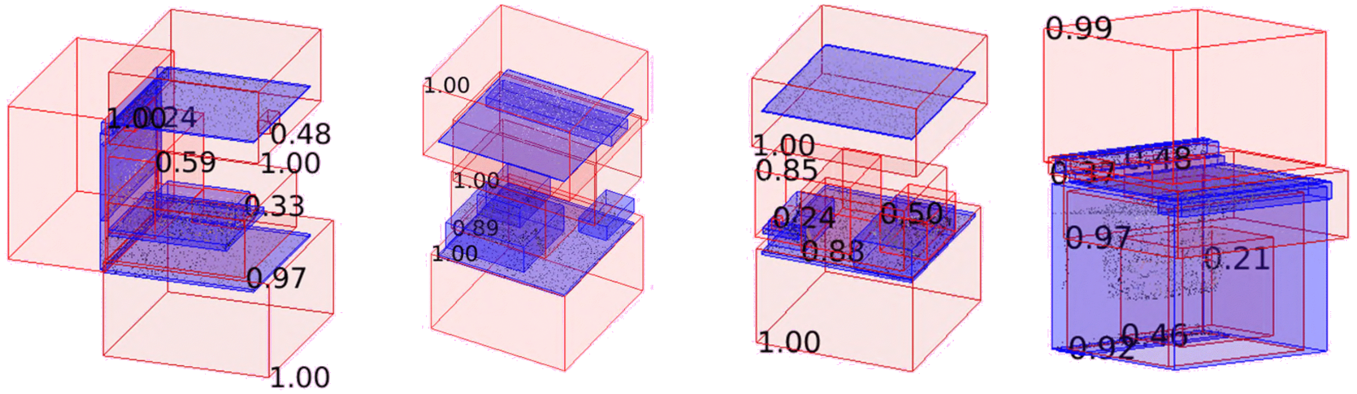
Metric: AP 50% ▾

Method	Info	avg ap 50%	bathtub	bed	bookshelf	cabinet	chair	counter	curtain	desk	door	otherfurniture	picture	refrigerator	shower curtain	sink
3D-BoNet		**0.488** 1	**1.000** 1	0.672 4	0.590 2	0.301 3	0.484 7	0.098 2	**0.620** 1	**0.306** 1	0.341 4	0.259 6	0.125 5	0.434 2	0.796 4	0.402 3
MTML		0.481 2	**1.000** 1	0.666 5	0.377 4	0.272 4	**0.709** 1	0.001 11	0.579 3	0.254 3	0.361 3	0.318 4	0.095 7	0.432 3	**1.000** 1	0.184 6
PanopticFusion-inst		0.478 3	0.667 5	0.712 3	**0.595** 1	0.259 6	0.550 6	0.000 12	0.613 2	0.175 5	0.250 7	**0.434** 1	**0.437** 1	0.411 5	0.857 2	**0.485** 1
Gaku Narita, Takashi Seno, Tomoya Ishikawa, Yohsuke Kaji: PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things. arXiv																
ResNet-backbone		0.459 4	**1.000** 1	**0.737** 1	0.159 10	0.259 5	0.587 4	**0.138** 1	0.475 5	0.217 4	**0.416** 1	0.408 3	0.128 4	0.315 6	0.714 5	0.411 2
MASC	P	0.447 5	0.528 8	0.555 7	0.381 3	**0.382** 1	0.633 2	0.002 9	0.509 4	0.260 2	0.361 2	0.432 2	0.327 2	**0.451** 1	0.571 6	0.367 4
Chen Liu, Yasutaka Furukawa: MASC: Multi-scale Affinity with Sparse Convolution for 3D Instance Segmentation.																
3D-SIS	P	0.382 6	**1.000** 1	0.432 8	0.245 7	0.190 7	0.577 5	0.013 7	0.263 7	0.033 10	0.320 5	0.240 7	0.075 8	0.422 4	0.857 2	0.117 9
Ji Hou, Angela Dai, Matthias Niessner: 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. CVPR 2019																
UNet-backbone		0.319 7	0.667 5	0.715 2	0.233 8	0.189 8	0.479 8	0.008 8	0.218 8	0.067 9	0.201 8	0.173 8	0.107 6	0.123 8	0.438 7	0.150 7
R-PointNet		0.306 8	0.500 9	0.405 9	0.311 5	0.348 2	0.589 3	0.054 3	0.068 10	0.126 6	0.283 6	0.290 5	0.028 9	0.219 7	0.214 10	0.331 5
3D-BEVIS		0.248 9	0.667 5	0.566 6	0.076 11	0.035 12	0.394 9	0.027 5	0.035 11	0.098 7	0.099 10	0.030 11	0.025 10	0.098 9	0.375 8	0.126 8
Cathrin Elich, Francis Engelmann, Jonas Schult, Theodora Kontogianni, Bastian Leibe: 3D-BEVIS: Birds-Eye-View Instance Segmentation.																
Seg-Cluster	P	0.215 10	0.370 10	0.337 11	0.285 6	0.105 9	0.325 10	0.025 6	0.282 6	0.085 8	0.105 9	0.107 9	0.007 12	0.079 10	0.317 9	0.114 10

Qualitative Results:



Intermediate Results:



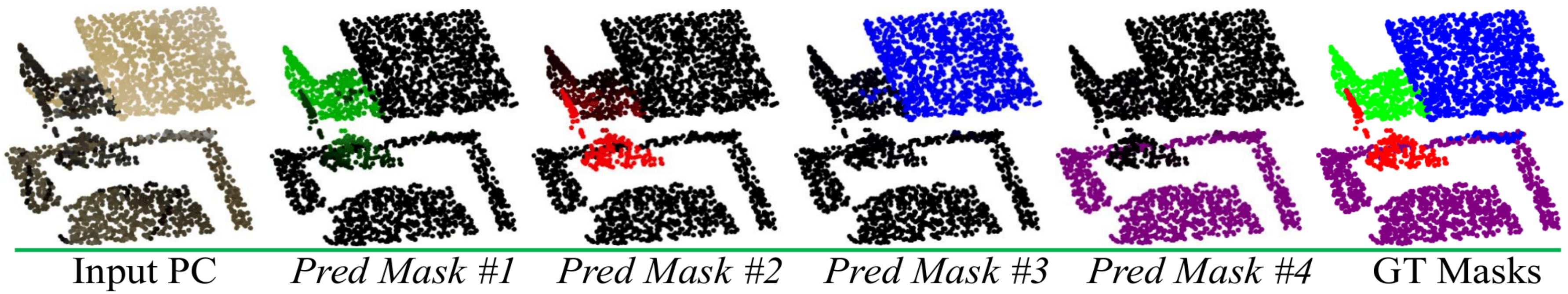


Table 4: Time consumption of different approaches on the validation split (312 scenes) of ScanNet(v2) (seconds).

	SGPN [50]	ASIS [51]	GSPN [58]	3D-SIS [15]	**3D-BoNet(Ours)**
	network(GPU): 650 group merging(CPU): 46562 block merging(CPU): 2221	network(GPU): 650 mean shift(CPU): 53886 block merging(CPU): 2221	network(GPU): 500 point sampling(GPU): 2995 neighbour search(CPU): 468	voxelization, projection, network, etc. (GPU+CPU): 38841	network(GPU): 650 *SCN (GPU parallel): 208* block merging(CPU): 2221
total	49433	56757	3963	38841	**2871**

Thank You