Semantic Segmentation & Object Detection in Point Clouds

Jiantao Gao

2020.8.20

Paper List

- JSENet: Joint Semantic Segmentation and Edge Detection Network for 3D Point Clouds (ECCV 2020)
- H3DNet: 3D Object Detection Using Hybrid Geometric Primitives (ECCV 2020)
- Weakly Supervised 3D Object Detection from Lidar Point Cloud (ECCV 2020)

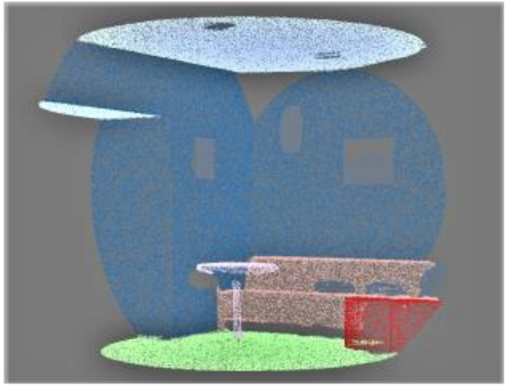
JSENet: Joint Semantic Segmentation and Edge Detection Network for 3D Point Clouds

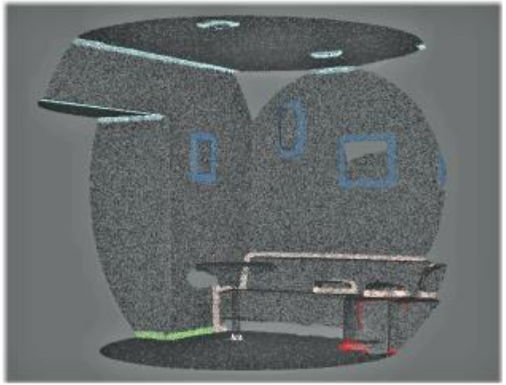
Zeyu HU[1[0000−0003−3585−7381]], Mingmin Zhen[1[0000−0002−8180−1023]], Xuyang BAI[1[0000−0002−7414−0319]], Hongbo Fu[2[0000−0002−0284−726X]], and Chiew-lan Tai[1[0000−0002−1486−1974]]

[1] Hong Kong University of Science and Technology
{zhuam,mzhen,xbaiad,taicl}@cse.ust.hk
[2] City University of Hong Kong
hongbofu@cityu.edu.hk

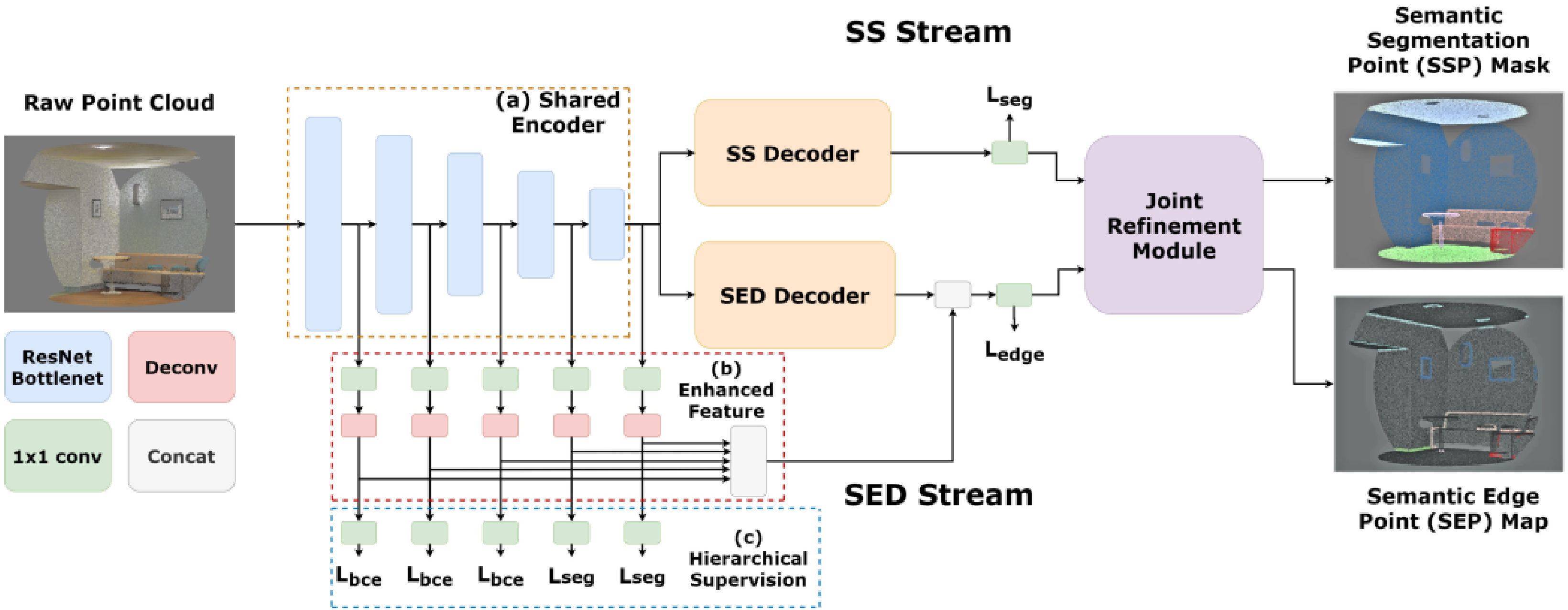
Motivation:

The Semantic segmentation (SS) and the semantic edge detection (SED) tasks can be seen as two dual problems with even interchangeable outputs in an ideal case.





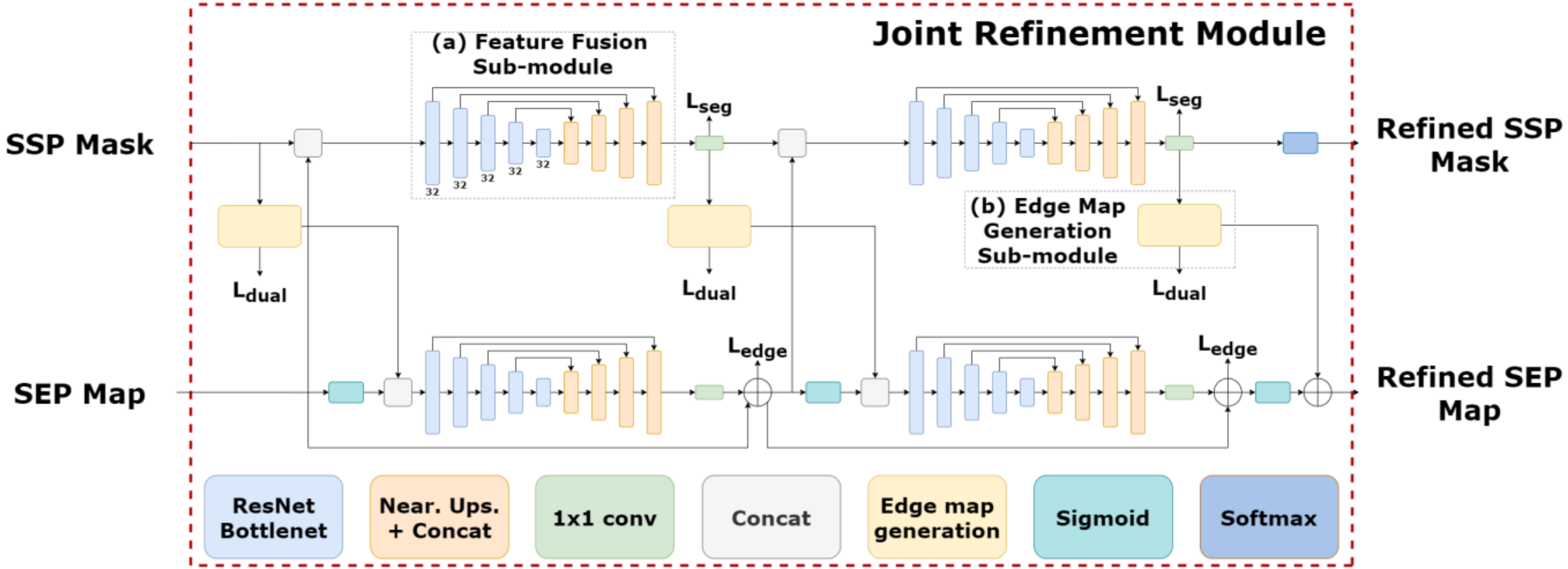
Pipeline:



Two streams of networks with a shared feature encoder

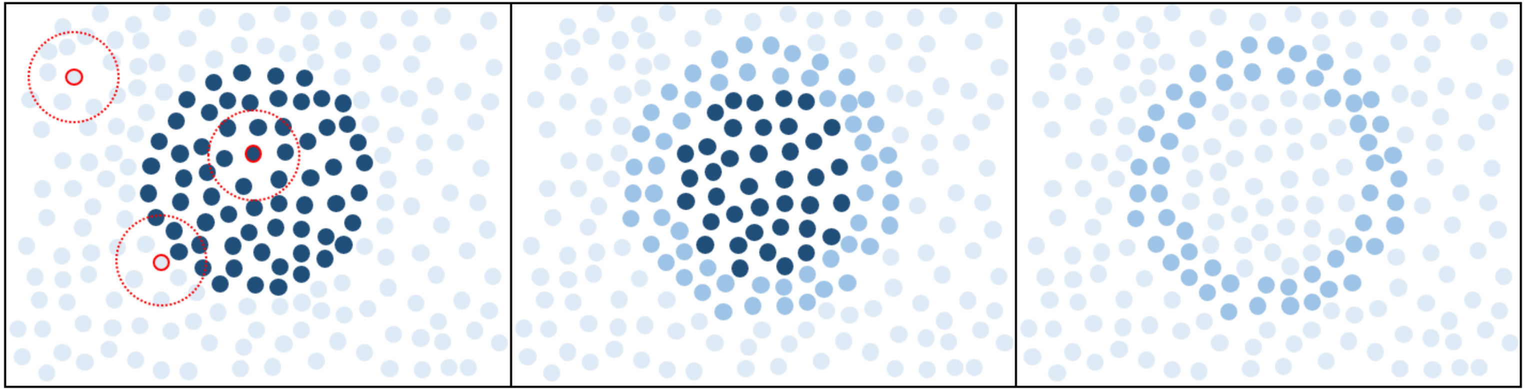
Joint refinement module

Joint refinement module:



Edge map generation sub-module:

Edge map generation sub-module converts an SSP mask to SEP mask (edge activation point maps)

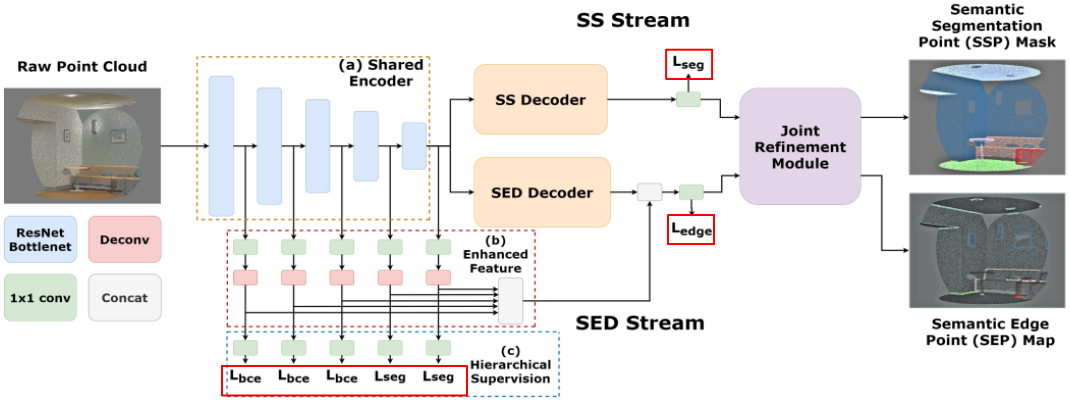


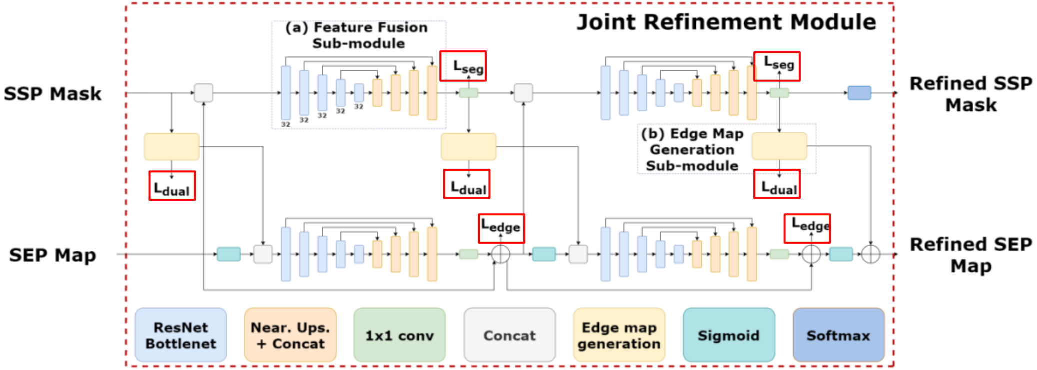
$$a_i = col_i(|M * Softmax(s) - Softmax(s)|)$$

The points nearer to the predicted boundaries will have larger activation values

Joint Multi-task Learning:

$$\boxed{L_{total} = \lambda_0 L_{seg} + \lambda_1 L_{edge} + \lambda_2 L_{bce} + \lambda_3 L_{dual}}$$





Experiments:

Semantic segmentation results on S3DIS:

Table 3: mIoU scores (%) of semantic segmentation task.

Method	S3DIS	ScanNet
TangentConv [54]	52.6	43.8
RNN Fusion [55]	53.4	-
SPGraph [56]	58.0	-
FCPN [57]	-	44.7
PointCNN [3]	57.3	45.8
ParamConv [58]	58.3	-
PanopticFusion [59]	-	52.9
TextureNet [60]	-	56.6
SPH3D-GCN [61]	59.5	61.0
HPEIN [37]	61.9	61.8
MCCNN [62]	-	63.3
MVPNet [4]	62.4	64.1
PointConv [42]	-	66.6
KPConv rigid [21]	65.4	68.6
KPConv deform [21]	67.1	68.4
SparseConvNet [29]	-	72.5
MinkowskiNet [30]	65.4	**73.6**
JSENet (ours)	**67.7**	69.9

Experiments:

Semantic edge detection results on S3DIS and ScanNet:

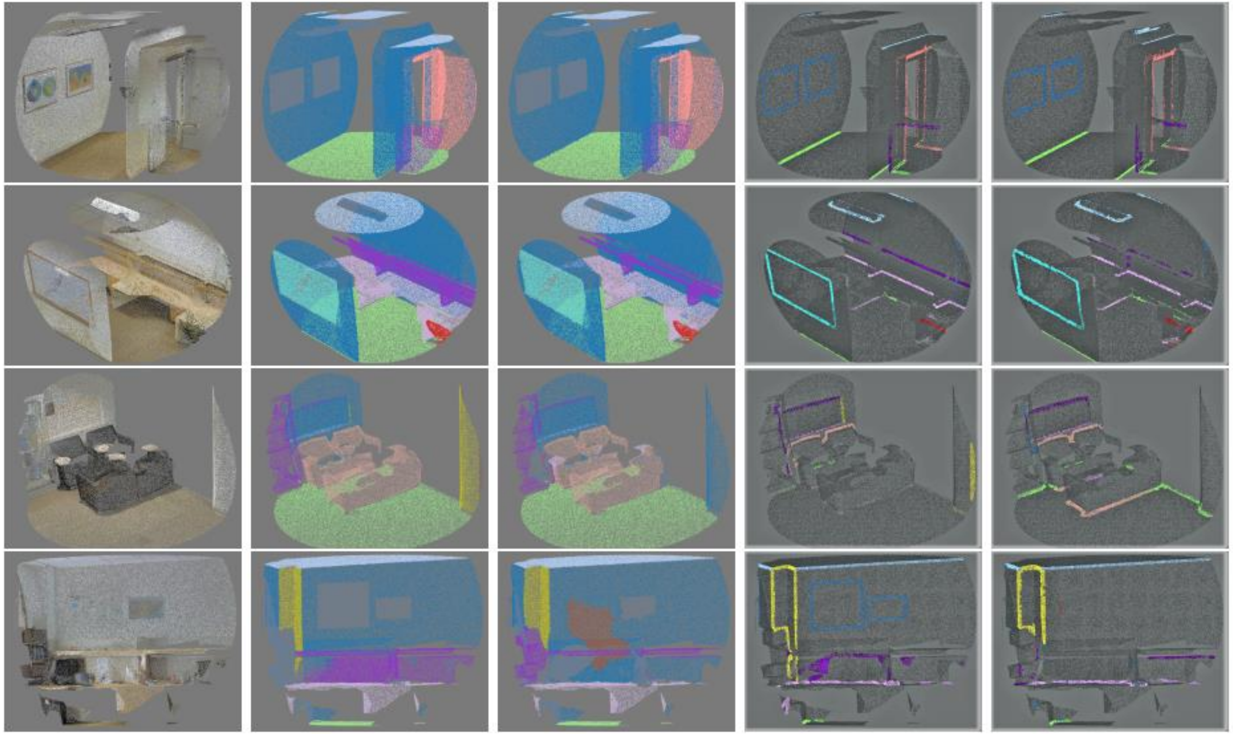
Table 4: MF (ODS) scores (%) of semantic edge detection on S3DIS Area-5.

Method	mean	ceil.	floor	wall	beam	col.	wind.	door	chair	table	book.	sofa	board	clut.
CASENet [8]	27.1	**46.5**	**49.0**	33.3	0.2	21.9	12.6	22.6	36.9	33.6	21.8	25.1	22.6	26.1
KPConv [21]	29.4	43.7	41.8	36.4	0.2	23.6	**13.4**	29.7	**39.8**	**37.3**	26.6	29.4	29.2	31.3
JSENet (ours)	**31.0**	44.5	43.2	**38.8**	**0.2**	**24.1**	13.2	**36.7**	37.7	36.3	**29.1**	**34.0**	**33.3**	**32.4**

Table 5: MF (ODS) scores (%) of semantic edge detection on ScanNet val set.

Method	mean	bath	bed	bksf	cab	chair	cntr	curt	desk	door	floor	othr	pic	ref	show	sink	sofa	tab	toil	wall	wind
CASENet [8]	32.3	38.2	55.9	29.9	36.0	36.0	36.8	28.1	28.5	19.1	26.6	25.1	32.2	28.6	26.6	23.9	19.0	45.7	27.1	54.6	27.4
KPConv [21]	34.8	40.5	**55.9**	33.6	**38.5**	39.3	38.0	32.9	31.2	22.0	25.1	28.5	36.2	30.8	30.9	22.7	22.8	46.5	33.3	**55.2**	32.3
JSENet (ours)	**37.3**	**43.8**	55.8	**35.9**	38.2	**41.0**	**40.8**	**34.5**	**35.9**	**25.5**	**28.7**	**29.5**	**37.3**	**36.2**	**31.7**	**28.1**	**28.3**	**48.5**	**35.6**	53.2	**37.8**

Experiments:



Point Cloud GT-SSP Mask Pred-SSP Mask GT-SEP Map Pred-SEP Map

Fig. 5: Qualitative results on S3DIS Area-5. For better visualization, we thickened all the semantic edges.

Ablation Study on S3DIS:

Table 1: Ablation experiments of network structures on S3DIS Area-5. **SEDS**: semantic edge detection stream; **EFE**: enhanced feature extraction; **HS**: hierarchical supervision; **SSS**: semantic segmentation stream; **JRM**: joint refinement module. The results in some cells (with '-') are not available, since the corresponding models perform either SS or SED.

0	SEDS	EFE	HS	SSS	JRM	mIoU (%)	mMF (ODS)(%)
1	✓	✓	✓	✓	✓	67.7	31.0
2	✓	✓	✓	✓		66.2	30.5
3				✓		64.7	-
4	✓	✓	✓			-	30.2
5	✓	✓				-	29.9
6	✓					-	29.4

Ablation Study on S3DIS:

Table 2: (a) Comparison of different supervision choices for SED. (b) Effects of the dual semantic edge loss in terms of boundary quality (F-score).

(a)

Method	mMF (ODS) (%)
L_{bce} for all five layers	30.1
L_{seg} for all five layers	30.1
No hierarchical supervision	29.9
L_{bce} for first three, L_{seg} for last two	30.2

(b)

Method	F-score (%)
JSENet w/o dual loss	22.7
JSENet	23.1

H3DNet: 3D Object Detection Using Hybrid Geometric Primitives

Zaiwei Zhang[1], Bo Sun[⋆1], Haitao Yang[⋆1], and Qixing Huang[1]

The University of Texas at Austin, Austin, Texas, USA, 78710

Motivation:

Hybrid Geometric Primitives: BB centers, BB face centers, and BB edge centers

Advantages:

1. The hybrid set of geometric primitives not only provides more accurate signals for object detection than using a single type of geometric primitives, but it also provides an overcomplete set of constraints on the resulting 3D layout;
2. The hybrid set of geometric primitives can make the model tolerate outliers in the predicted geometric primitives better.

Introduce hybrid geometric primitives to the 3D object detection ?

Pipeline:

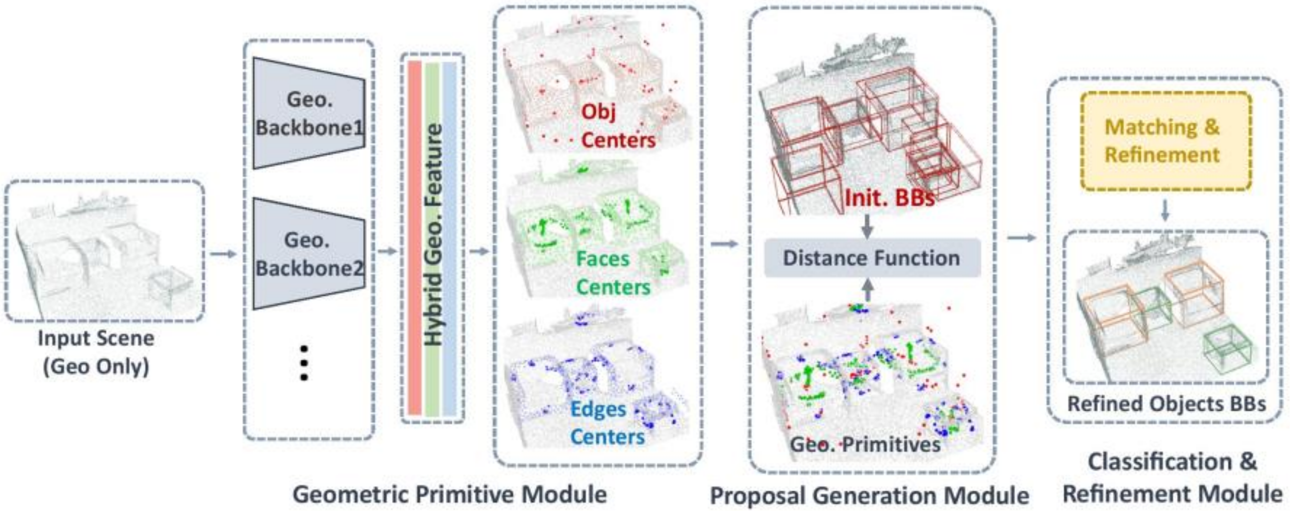
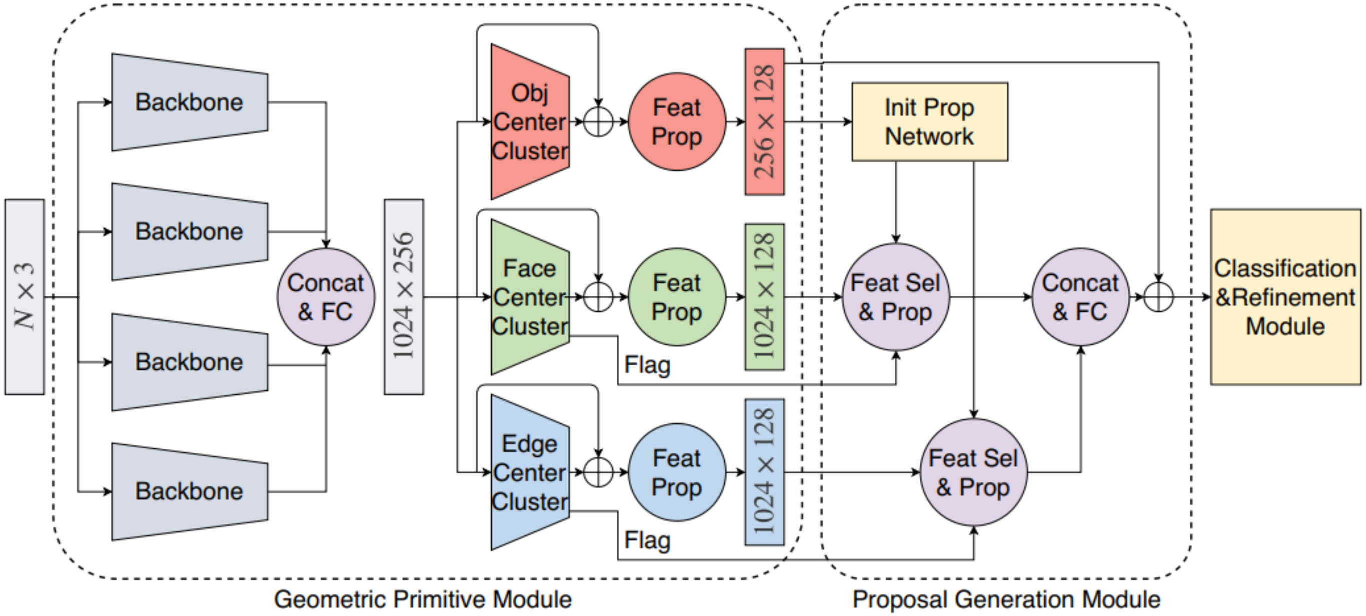
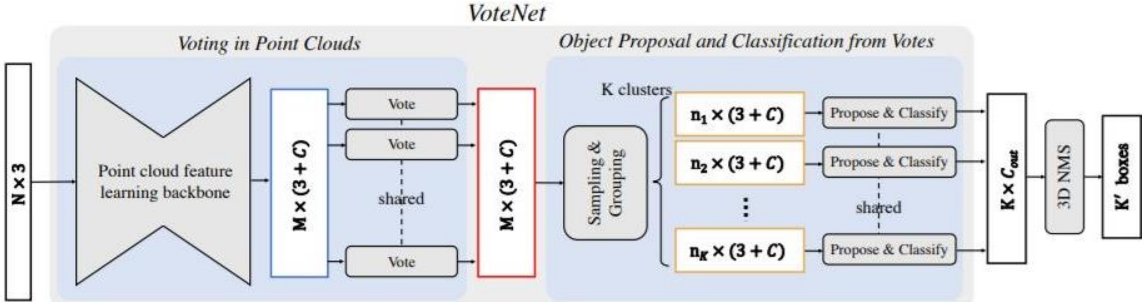


Fig. 2: H3DNet consists of three modules. The first module computes a dense descriptor and predicts three geometric primitives, namely, BB centers, BB face centers, and BB edge centers. The second module converts geometric primitives into object proposals. The third module classifies object proposals and refines the detected objects.

Pipeline:

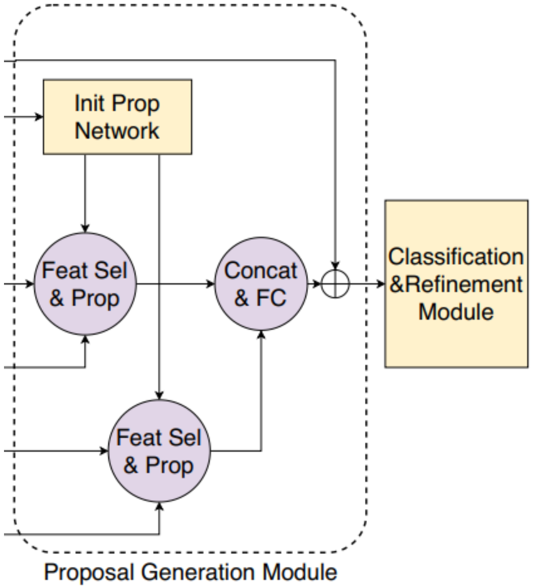


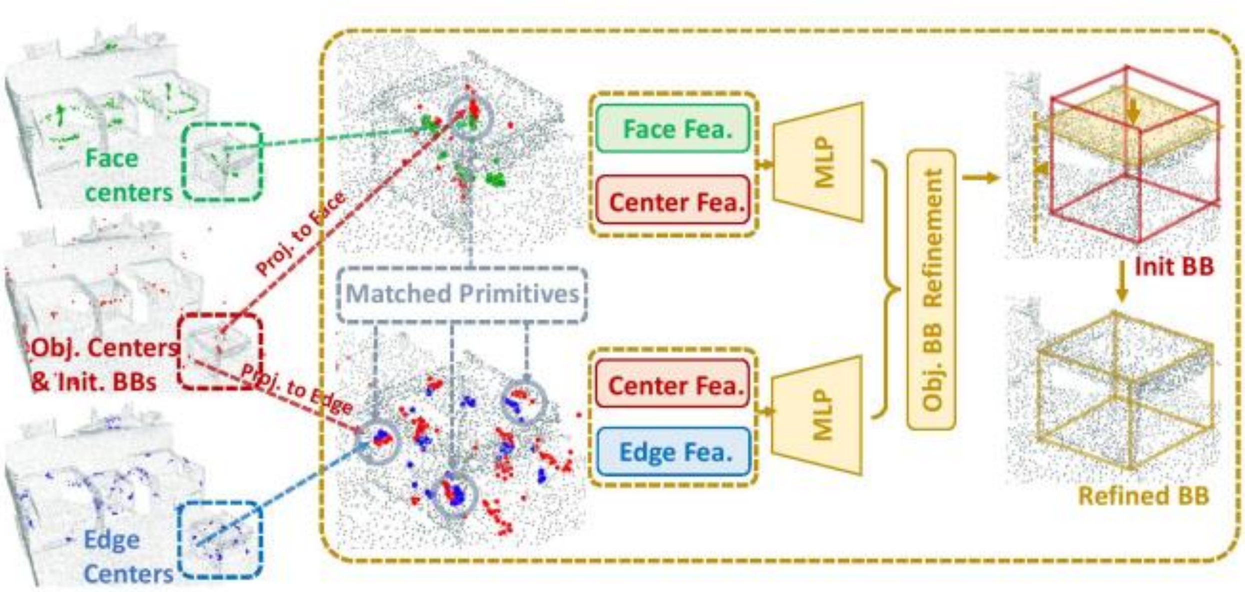
$N \times 3$
Backbone
Backbone
Backbone
Backbone
Concat & FC
1024×256
Obj Center Cluster
Face Center Cluster
Edge Center Cluster
Feat Prop
Feat Prop
Feat Prop
256×128
1024×128
1024×128
Flag
Flag
Init Prop Network
Feat Sel & Prop
Feat Sel & Prop
Concat & FC
Classification &Refinement Module
Geometric Primitive Module
Proposal Generation Module



VoteNet
Voting in Point Clouds
Object Proposal and Classification from Votes
$N \times 3$
Point cloud feature learning backbone
$M \times (3 + C)$
Vote
Vote
shared
Vote
$M \times (3 + C)$
K clusters
Sampling & Grouping
$n_1 \times (3 + C)$
$n_2 \times (3 + C)$
$n_K \times (3 + C)$
Propose & Classify
Propose & Classify
shared
Propose & Classify
$K \times C_{out}$
3D NMS
K' boxes

Classification & Refinement Module:





Experiments:

ScanNet v2 and SUN RGB-D:

Table 2: **Left:** 3D object detection results on ScanNetV2 val set. **Right:** results on SUN RGB-D V1 val set. We show mean of average precision (mAP) across all semantic classes with 3D IoU threshold 0.25 and 0.5.

	Input	mAP@0.25	mAP@0.5
DSS[39]	Geo + RGB	15.2	6.8
F-PointNet[29]	Geo + RGB	19.8	10.8
GSPN[51]	Geo + RGB	30.6	17.7
3D-SIS [9]	Geo + 5 views	40.2	22.5
VoteNet [28]	Geo only	58.7	33.5
Ours	Geo only	**67.2**	**48.1**
w\o refine	Geo only	60.2	37.3

	Input	mAP@0.25	mAP@0.5
DSS[39]	Geo + RGB	42.1	-
COG[32]	Geo + RGB	47.6	-
2D-driven[17]	Geo + RGB	45.1	-
F-PointNet[29]	Geo + RGB	54.0	-
VoteNet [28]	Geo only	57.7	32.9
Ours	Geo only	**60.1**	**39.0**
w\o refine	Geo only	58.5	34.2

Experiments:

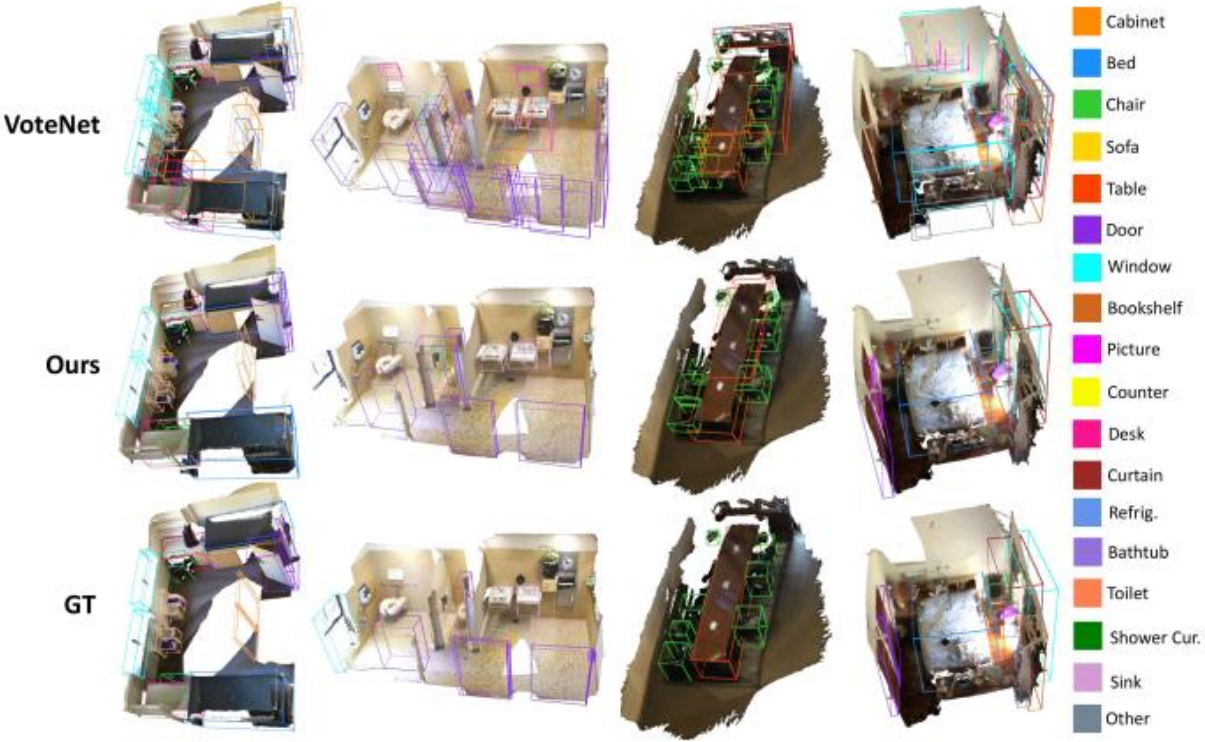


Fig. 5: Qualitative baseline comparisons on ScanNet V2.

Experiments:

Fig. 6: Qualitative baseline comparisons on SUN RGB-D.

Ablation Study on ScanNet v2:

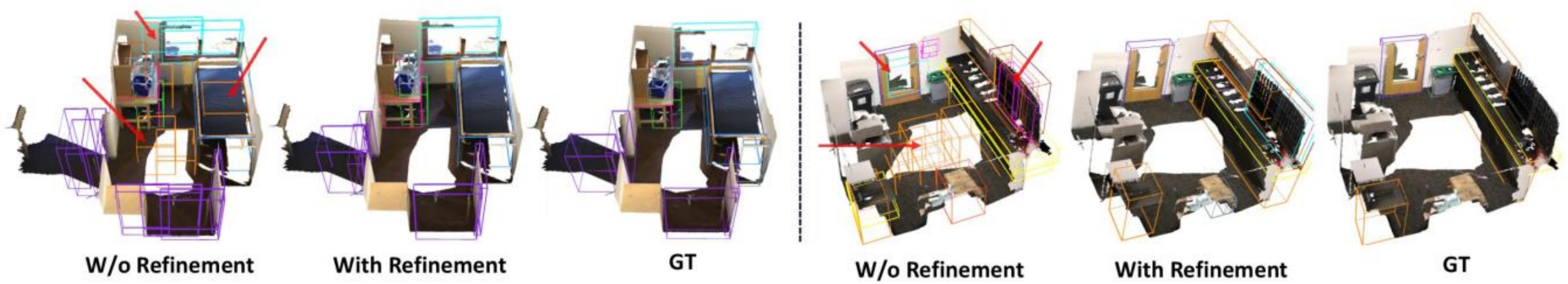


Fig. 4: Effect of geometric primitive matching and refinement.

Ablation Study on ScanNet v2:

Table 3: Quantitative results without refining predicted center, size, semantic or object existence score for ScanNet, and without refining predicted angle for SUN RGB-D and differences compared with refining all.

	mAP@0.25		mAP@0.5	
w\o center	66.9	-0.3	46.3	-1.8
w\o size	65.4	-1.8	44.2	-3.9
w\o semantic	66.2	-1.0	47.3	-0.8
w\o existence	65.2	-1.8	45.1	-3.0
w\o angle	58.6	-1.5	36.6	-2.4

Table 4: Quantitative comparisons between different number of descriptor computation towers, among our approach and VoteNet, for ScanNet and SUN RGB-D.

	# of Towers	mAP@0.25	mAP@0.5
Ours	1	64.4	43.4
	2	65.4	46.2
	3	66.0	47.7
	4	67.2	48.3
Vote	4 (Scan)	60.11	37.12
	4 (SUN)	57.5	32.1

Weakly Supervised 3D Object Detection from Lidar Point Cloud

Qinghao Meng[1], ✉Wenguan Wang[2], Tianfei Zhou[3], Jianbing Shen[3], Luc Van Gool[2], and Dengxin Dai[2]

[1]School of Computer Science, Beijing Institute of Technology
[2]ETH Zurich [3]Inception Institute of Artificial Intelligence
https://github.com/hlesmqh/WS3D

Data Annotation Strategy for Our Weak Supervision:

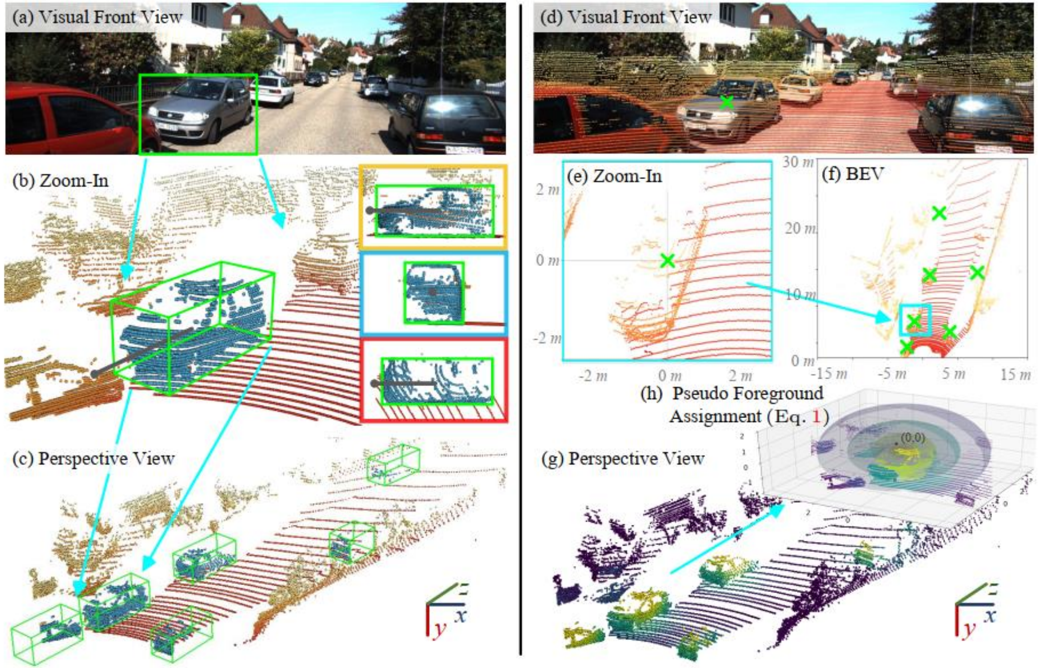


Fig. 3. (a-c): Precise annotations require extensive labeling efforts (see §3). **(d-f)**: Our weak supervision is simply obtained by clicking object centers (denoted by ✖) on BEV maps (see §3). **(g-h)**: Our pseudo groundtruths for fore-/background segmentation (yellower indicates higher foreground score; see §4.1).

Pseudo Ground-truth Generation:

A labeled vehicle center point ***o*** on Bev: (x_o, z_o)

Setting its height as: $y_o = 0$

Pseudo Ground-truth Generation:

$$f^p = \max_{o\in\mathcal{O}}(\iota(p,o)), \quad \text{where} \quad \iota(p,o) = \begin{cases} 1 & \text{if } d(p,o) \leq 0.7, \\ \frac{1}{\kappa}\mathcal{N}(d(p,o)) & \text{if } d(p,o) > 0.7. \end{cases}$$

$$d(p,o) = [(x_p - x_o)^2 + \tfrac{1}{2}(y_p - y_o)^2 + (z_p - z_o)^2]^{\frac{1}{2}}$$

Pipeline:

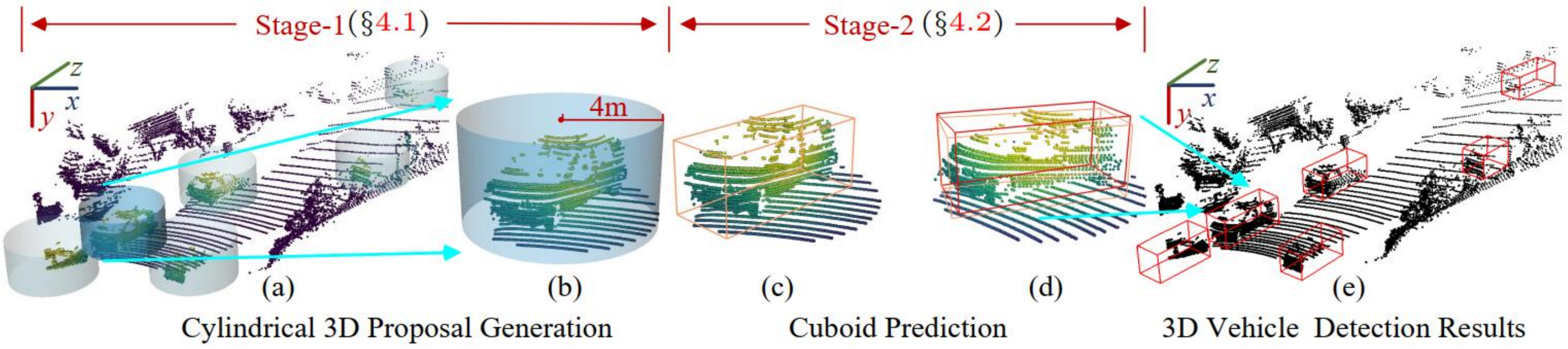
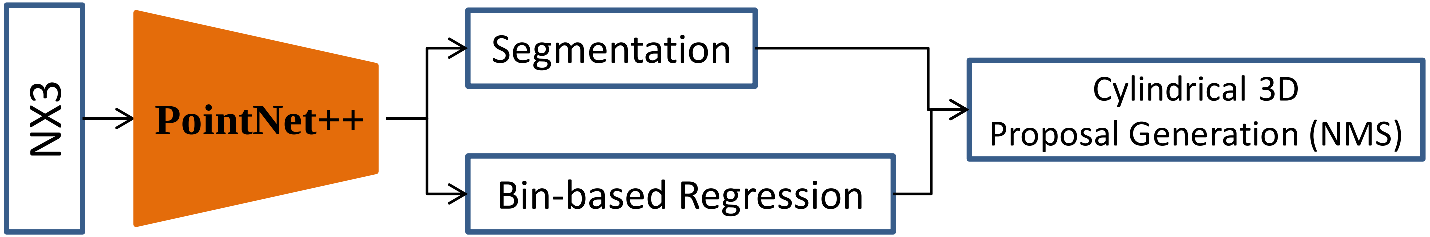


Fig. 5. Our 3D object detection pipeline (§4). (a-b) Cylindrical 3D proposal generation results from Stage-1 (§4.1). Yellower colors correspond to higher foreground probabilities. (c-d) Cuboid prediction in Stage-2 (§4.2). (e) Our final results.

Stage-1:



Pipeline:

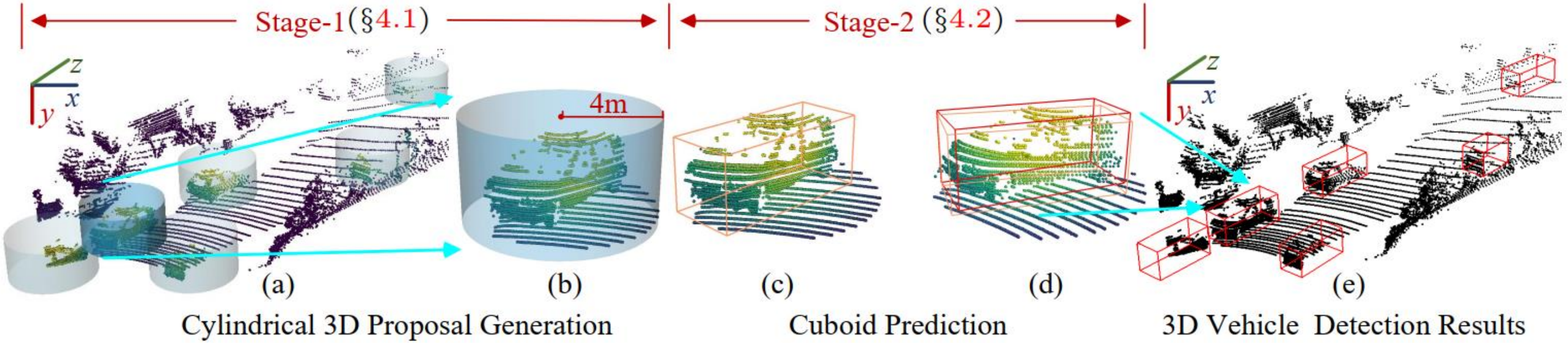
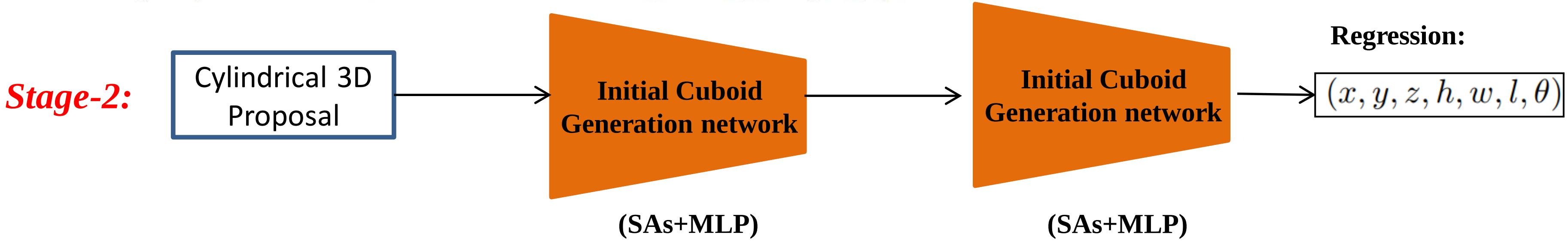


Fig. 5. Our 3D object detection pipeline (§4). (a-b) Cylindrical 3D proposal generation results from Stage-1 (§4.1). Yellower colors correspond to higher foreground probabilities. (c-d) Cuboid prediction in Stage-2 (§4.2). (e) Our final results.



Experiments:

Kitti:

Table 1. Evaluation results on KITTI val set (Car). See §5.2 for details.

Learning Paradigm	Detector	Modality	BEV@0.7			3D Box@0.7		
			Easy	Moderate	Hard	Easy	Moderate	Hard
Trained with the whole KITTI **train** set: 3,712 precisely labeled scenes with 15,654 vehicle instances								
Fully supervised	VeloFCN [11]	LiDAR	40.14	32.08	30.47	15.20	13.66	15.98
	PIXOR [5]	LiDAR	86.79	80.75	76.60	-	-	-
	VoxelNet [17]	LiDAR	89.60	84.81	78.57	81.97	65.46	62.85
	SECOND [16]	LiDAR	89.96	87.07	79.66	87.43	76.48	69.10
	PointRCNN [13]	LiDAR	-	-	-	88.45	77.67	76.30
	PointPillars [2]	LiDAR	89.64	86.46	84.22	85.31	76.07	69.76
	Fast PointR-CNN [6]	LiDAR	90.12	88.10	86.24	89.12	79.00	77.48
	STD [18]	LiDAR	90.50	88.50	88.10	89.70	79.80	79.30
Trained with a part of KITTI **train** set: 500 precisely labeled scenes with 2,176 vehicle instances								
Fully supervised	PointRCNN [13]	LiDAR	87.21	77.10	76.63	79.88	65.50	64.93
	PointPillars [2]	LiDAR	86.27	77.13	75.91	72.36	60.75	55.88
Trained with a part of KITTI **train** set: 125 precisely labeled scenes with 550 vehicle instances								
Fully supervised	PointRCNN [13]	LiDAR	85.09	74.35	67.68	67.54	54.91	51.96
	PointPillars [2]	LiDAR	85.76	75.30	73.29	65.51	51.45	45.53
Trained with a part of KITTI **train** set: 500 weakly labeled scenes with 534 precisely annotated instances								
Weakly supervised	**Ours**	LiDAR	88.56	84.99	84.74	84.04	75.10	73.29

Experiments:

Kitti:

Table 2. Evaluation results on KITTI test set (Car). See §5.2 for details.

Learning Paradigm	Detector	Modality	BEV@0.7			3D Box@0.7		
			Easy	Moderate	Hard	Easy	Moderate	Hard
Trained with the whole KITTI `train` set: 3, 712 precisely annotated scenes with 15, 654 vehicle instances								
Fully supervised	PIXOR [5]	LiDAR	87.25	81.92	76.01	-	-	-
	VoxelNet [17]	LiDAR	89.35	79.26	77.39	77.47	65.11	57.73
	SECOND [16]	LiDAR	88.07	79.37	77.95	83.13	73.66	66.20
	PointRCNN [13]	LiDAR	89.47	85.68	79.10	85.94	75.76	68.32
	PointPillars [2]	LiDAR	88.35	86.10	79.83	79.05	74.99	68.30
	Fast PointR-CNN [6]	LiDAR	88.03	86.10	78.17	84.28	75.73	67.39
	STD [18]	LiDAR	94.74	89.19	86.42	87.95	79.71	75.09
Trained with a part of KITTI `train` set: 500 weakly labeled scenes + 534 precisely annotated instances								
Weakly supervised	**Ours**	LiDAR	90.11	84.02	76.97	80.15	69.64	63.71

Experiments:

Kitti:

Fig. 6. Qualitative results of 3D object detection (Car) on KITTI `val` set (§5.2). Detected 3D bounding boxes are shown in yellow; images are used only for visualization.

Experiments:

Kitti:

Table 3. Evaluation results on KITTI `val` set (Pedestrian). See §5.2 for details.

Learning Paradigm	Detector	Modality	BEV@0.5			3D Box@0.5		
			Easy	Moderate	Hard	Easy	Moderate	Hard
Trained with: 951 precisely labeled scenes with 2,257 pedestrian instances								
Fully supervised	PointPillars [2]	LiDAR	71.97	67.84	62.41	66.73	61.06	56.50
	PointRCNN [13]	LiDAR	68.89	63.54	57.63	63.70	69.43	58.13
	Part-A^2 [40]	LiDAR	-	-	-	70.73	64.13	57.45
	VoxelNet [17]	LiDAR	70.76	62.73	55.05	-	-	-
	STD [18]	LiDAR	75.90	69.90	66.00	73.90	66.60	62.90
Trained with: 951 weakly labeled scenes with 515 pedestrian instances								
Weakly supervised	**Ours**	LiDAR	74.79	70.17	66.75	74.65	69.96	66.49

Experiments:

Performance as An Annotation Tool:

Table 5. Comparison of annotation quality on KITTI val set (see §5.4).

Learning Paradigm	Method	Mode	Speed (sec./inst.)	BEV@0.5 Easy	BEV@0.5 Moderate	BEV@0.5 Hard	3D Box@0.5 Easy	3D Box@0.5 Moderate	3D Box@0.5 Hard
Trained with the whole KITTI **train** set: 3,712 well-labeled scenes with 15,654 vehicle instances									
Fully Supervised	[15]	Active	3.8	-	-	-	-	-	88.33
Trained with KITTI **train+val**: 7,481 scenes (implicitly using 2D instance segmentation annotations)									
Fully-Supervised	[14]	Auto	8.0	80.70	63.36	52.47	63.39	44.79	37.47
Trained with a part of KITTI **train** set: 500 weakly labeled scenes + 534 precisely annotated instances									
Weakly Supervised	**Ours**	Auto	0.1	96.33	89.01	88.52	95.85	89.14	88.32
		Active	2.6	99.99	99.92	99.90	99.87	90.78	90.14

Table 6. Performance of PointRCNN [13] and PointPillars [2] when trained using different annotations sources. Results are reported on KITTI val set (§5.4).

Detector	Annotation Source	BEV@0.7 Easy	BEV@0.7 Moderate	BEV@0.7 Hard	3D Box@0.7 Easy	3D Box@0.7 Moderate	3D Box@0.7 Hard
PointRCNN [13]	Manual	90.21	87.89	85.51	88.45	77.67	76.30
	Automatic (ours)	88.02	85.75	84.27	83.22	74.54	73.29
	Active (ours)	88.64	85.41	84.94	84.21	76.08	74.91
PointPillars [2]	Manual	89.64	86.46	84.22	85.31	76.07	69.76
	Automatic (ours)	88.55	85.62	83.84	84.79	74.18	68.52
	Active (ours)	88.94	85.88	83.86	84.53	75.03	68.63

Thanks