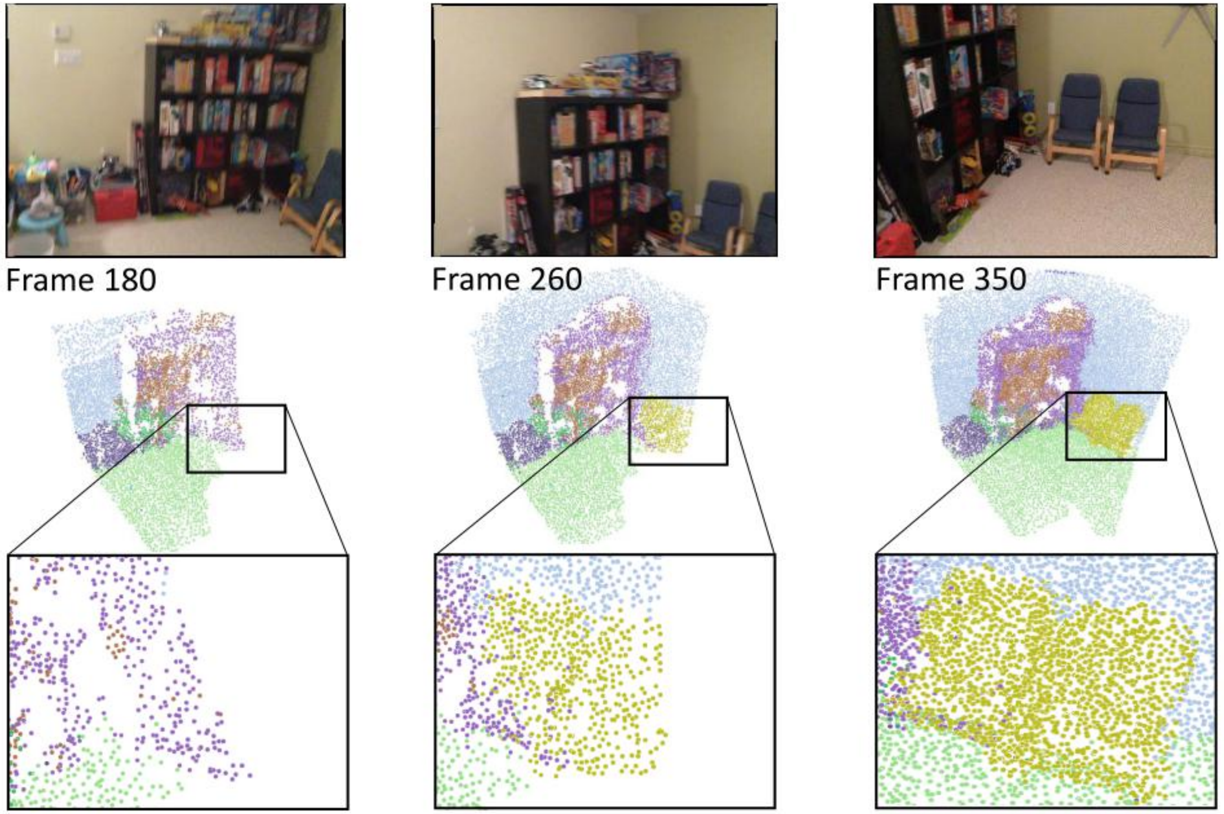
Deep Learning for 3D Point Cloud Analysis



Xu Yan

2020.6.26

香 港 中 文 大 學 (深 圳)
The Chinese University of Hong Kong, Shenzhen

Paper List

- Multi-Path Region Mining For Weakly Supervised 3D Semantic Segmentation on Point Clouds (CVPR2020)
- SESS: Self-Ensembling Semi-Supervised 3D Object Detection (CVPR2020 oral)
- Fusion-Aware Point Convolution for Online Semantic 3D Scene Segmentation (CVPR2020 best paper final list)

Paper List

- Multi-Path Region Mining For Weakly Supervised 3D Semantic Segmentation on Point Clouds (CVPR2020)
- SESS: Self-Ensembling Semi-Supervised 3D Object Detection (CVPR2020 oral)
- Fusion-Aware Point Convolution for Online Semantic 3D Scene Segmentation (CVPR2020 best paper final list)

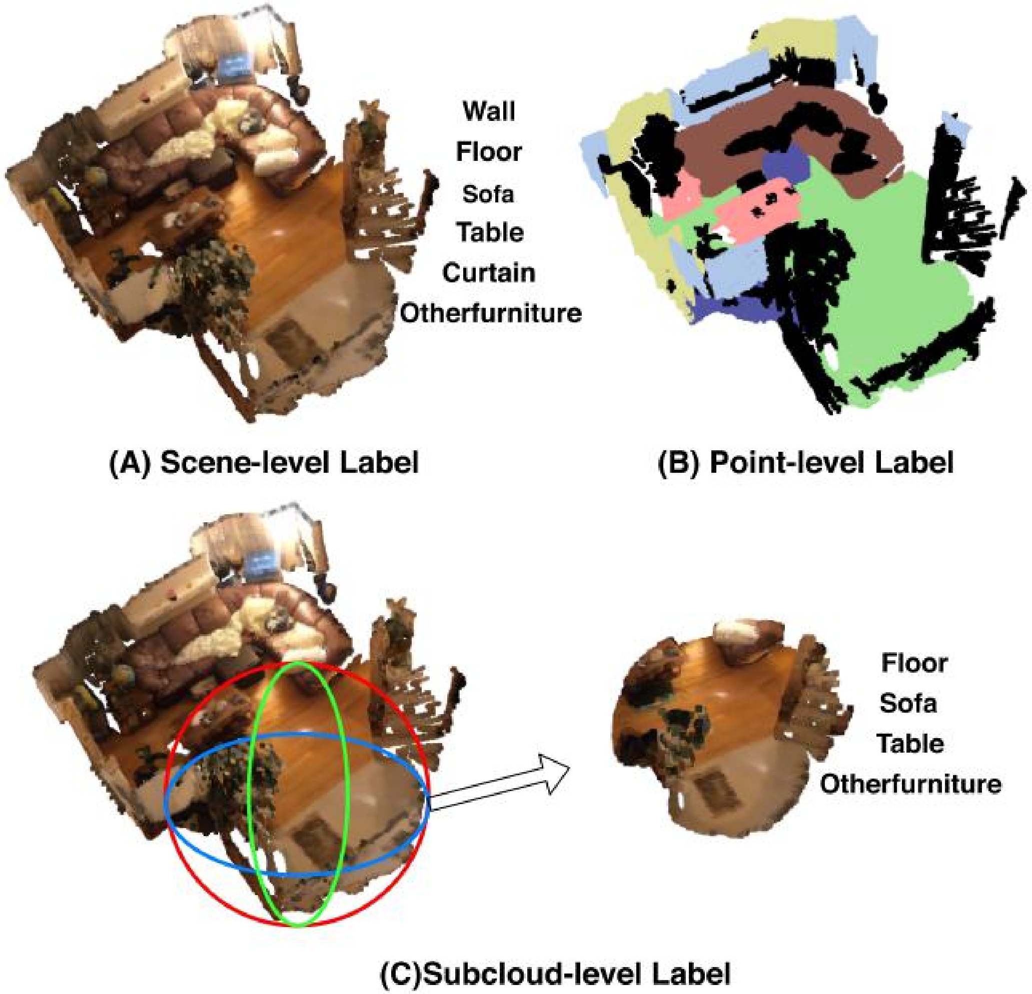
MPRM

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Multi-Path Region Mining For Weakly Supervised 3D Semantic Segmentation on Point Clouds

Jiacheng Wei[1] Guosheng Lin[1]* Kim-Hui Yap[1] Tzu-Yi Hung[2] Lihua Xie[1]
[1]Nanyang Technological University, Singapore [2]Delta Research Center, Singapore
{jiacheng002, gslin, ekhyap, elhxie}@ntu.edu.sg, tzuyi.hung@deltaww.com

MPRM

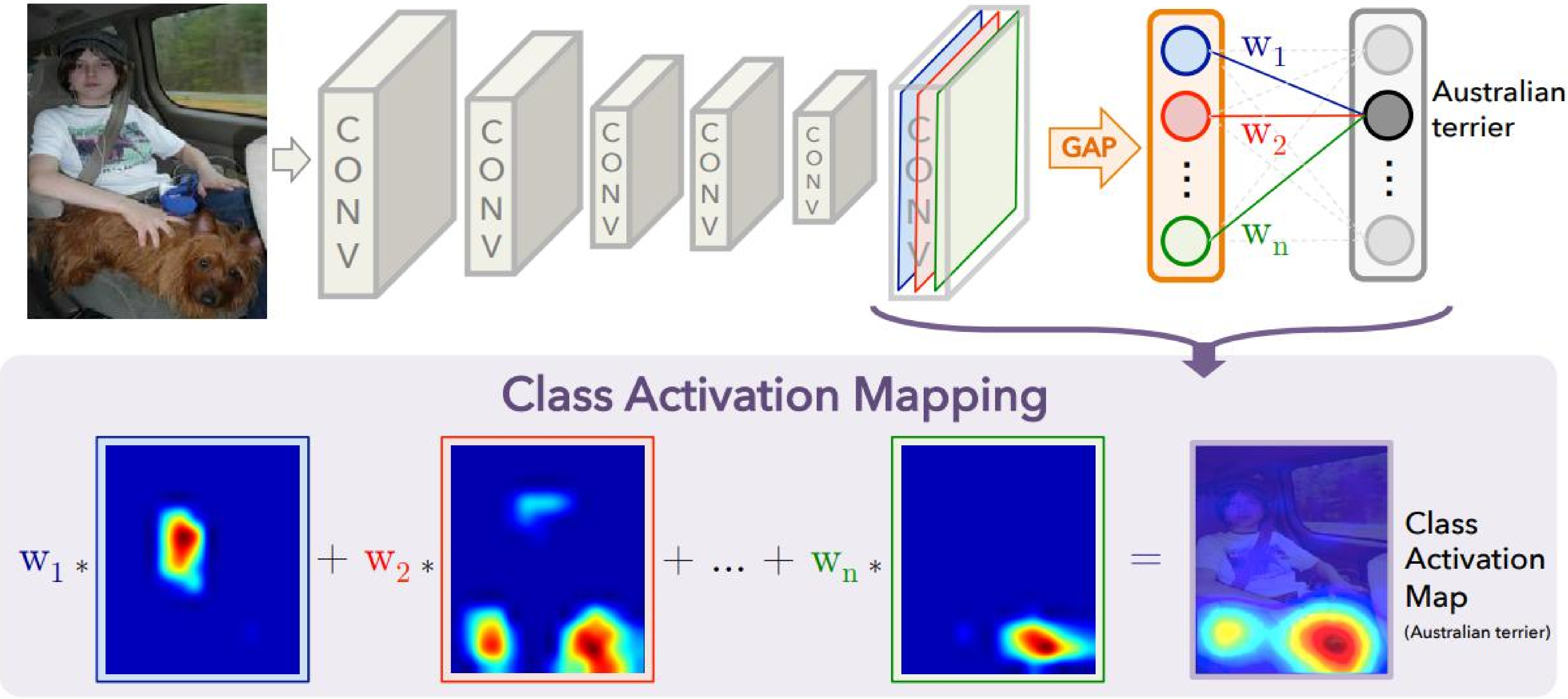


For *ScanNet* dataset

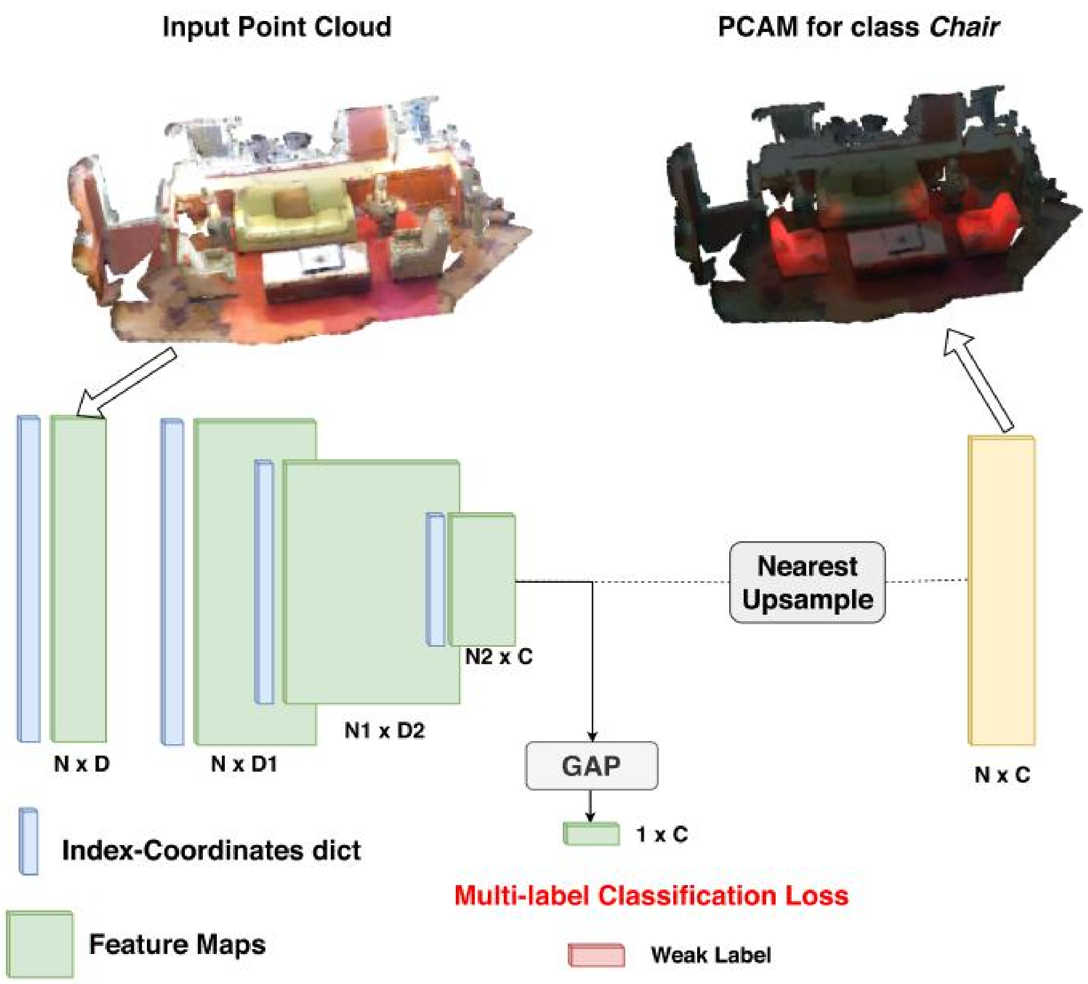
- Only 20 people participated in the collection of 1513 3D scans.
- More than 500 workers participated in the semantic annotation process.
- To ensure the annotation accuracy, each scene was annotated by 2 to 3 participants.
- The median and mean time for annotation per scan is 16.8 min and 22.3 min.
- The estimated annotation time for scene-level labels in a scene is around 15 sec, while the annotation time for subclouds from a scene is lower than 3 min. The average number of subclouds is 18.4 (save 6-70 times the time).

MPRM

Class activation map (**CAM**) in 2D image:



MPRM



- Point class activation map (**PCAM**):

We denote $f_{cam}(p)$ as the PCAM feature vector of point p before the global average pooling layer. For class c, the PCAM $M_c(p)$ for point p can be expressed as:

$$M_c(p) = \mathbf{w}_c^{\intercal} \cdot f_{cam}(p) \cdot \mathbf{y}_c, \quad (2)$$

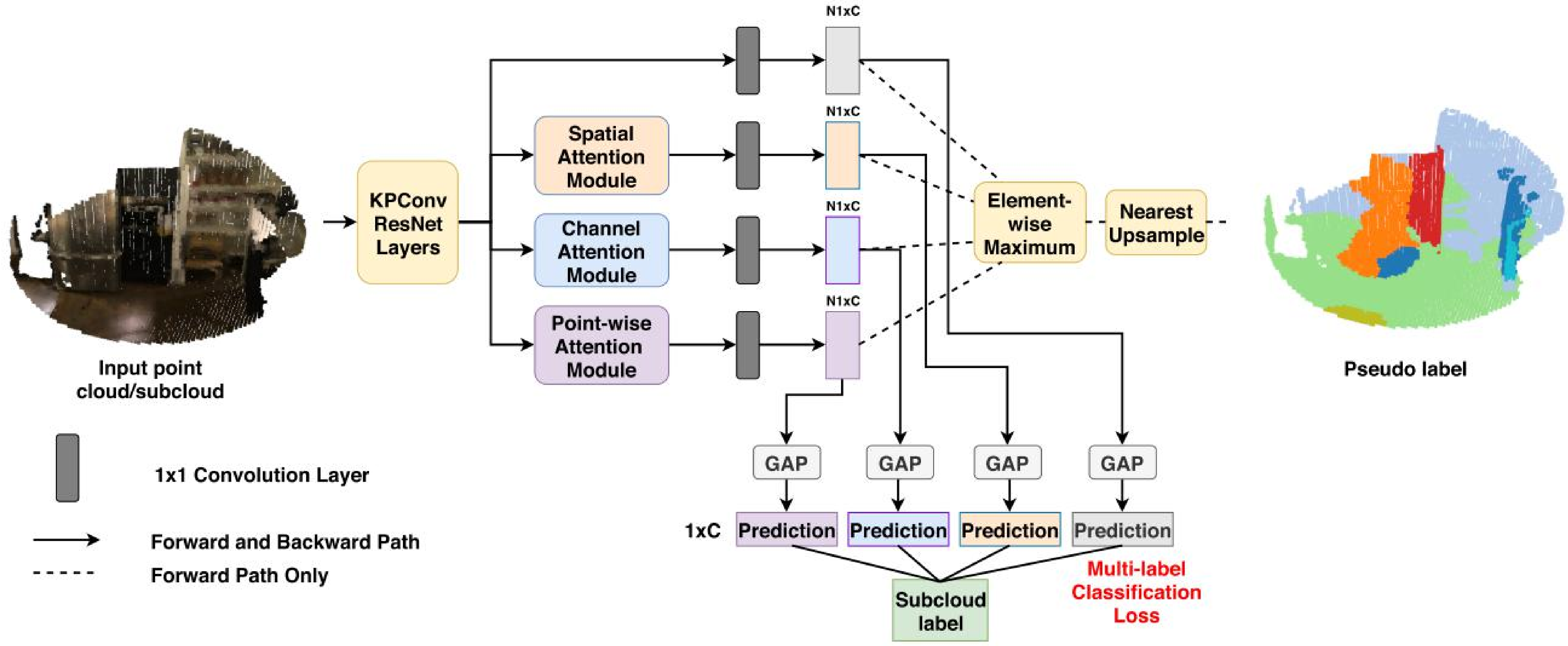
w_c is the classification weights for class c and $y_c \in \{0, 1\}$ indicates the one-hot subcloud ground truth for class c.

- Point-level pseudo masks:

$$argmax(M(p))$$

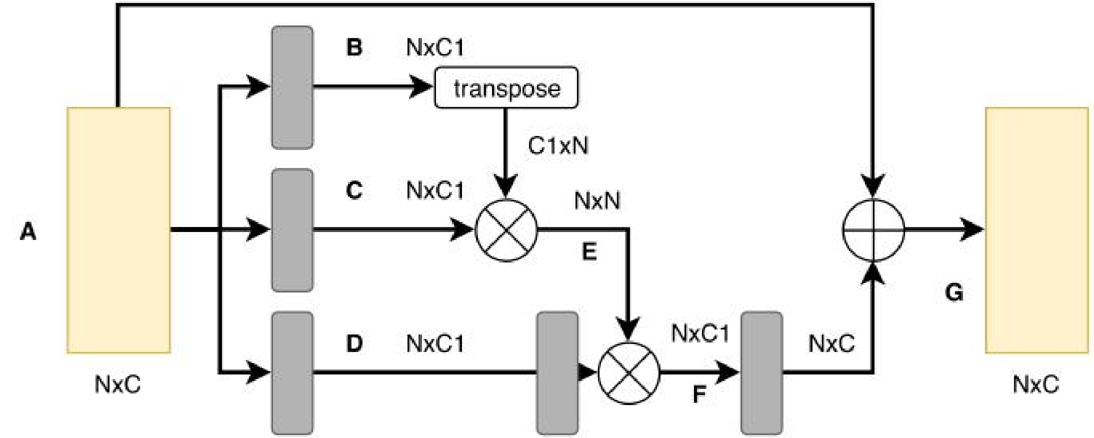
MPRM

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

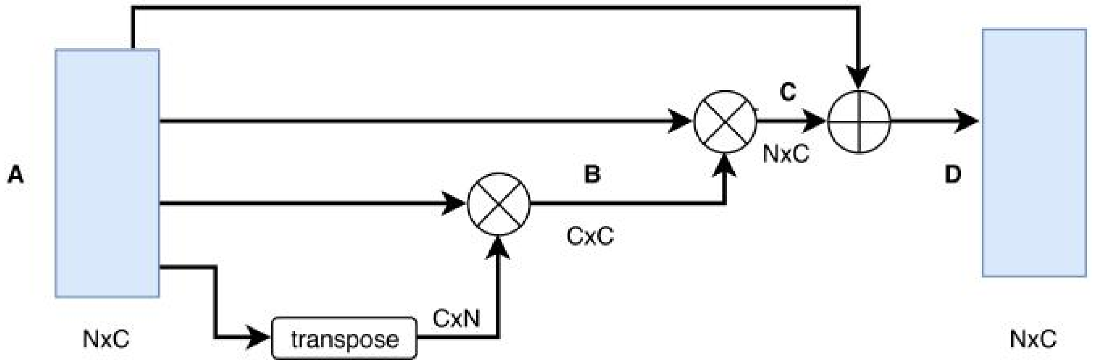


MPRM

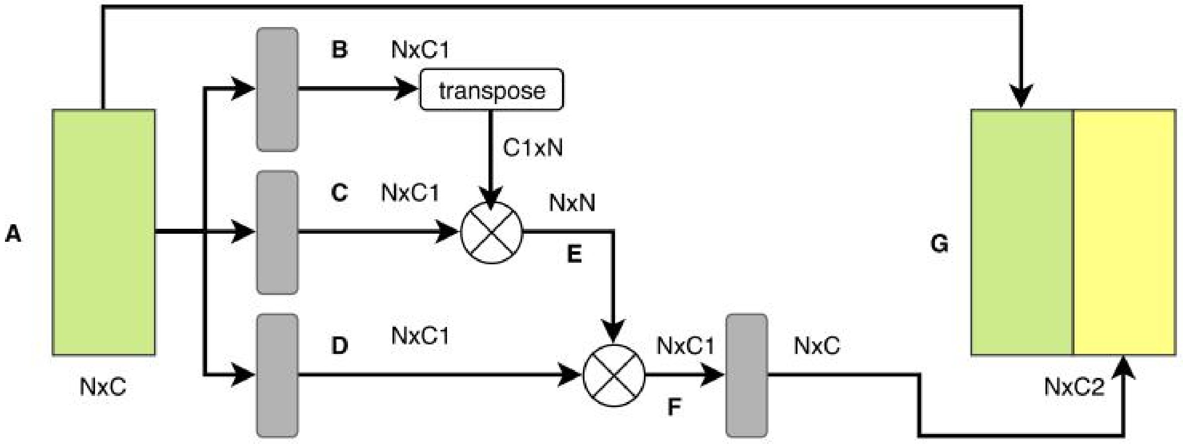
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen



A.Spatial Attention Module



B.Channel Attention Module



C.Point-wise Attention Module

MPRM

Setting	Supervision	wall	floor	cabinet	bed	chair	sofa	table	door	window	B.S.	picture	cnt	desk	curtain	fridge	S.C.	toilet	sink	bathtub	other	mIoU
PCAM(Baseline)	Scene	54.9	48.3	14.1	34.7	32.9	45.3	26.1	0.6	3.3	46.5	0.6	6.0	7.4	26.9	0.0	6.1	22.3	8.2	52.0	6.1	22.1
MPRM(Ours)	Scene	47.3	41.1	10.4	43.2	25.2	43.1	21.5	9.8	12.3	45.0	9.0	13.9	21.1	40.9	1.8	29.4	14.3	9.2	39.9	10.0	24.4
PCAM(Baseline)	Subcloud	59.0	53.8	24.7	64.9	45.7	60.7	42.8	31.5	37.0	55.9	31.0	12.0	39.1	68.7	16.8	49.8	55.2	27.4	59.0	27.7	43.1
MPRM(Ours)	Subcloud	56.1	54.8	32.0	69.6	49.5	67.7	46.6	41.3	44.2	71.5	28.3	21.3	49.2	71.8	38.1	42.8	43.6	20.3	49.0	33.8	46.6
dCRF post-processing:																						
MPRM(Ours)	Subcloud	58.0	57.3	33.2	71.8	50.4	69.8	47.9	42.1	44.9	73.8	28.0	21.5	49.5	72.0	38.8	44.1	42.4	20.0	48.7	34.4	47.4

Table 2. The class-specific segmentation results (mIoU) of pseudo labels on training set generated with different settings and different supervision levels. We only show the dCRF post-processed result for MPRM with subcloud-level supervision since we use this pseudo label to train our final segmentation model. (Here B.S. stands for bookshelf; S.C. stands for shower curtain; cnt stands for counter.)

Fusion	PCAM	SA	CA	PSA	Training	Validation
-	√				44.3	39.3
-		√			44.8	39.4
-			√		44.3	39.3
-				√	44.7	39.5
Max	√	√			46.0	40.3
Max	√		√		45.9	40.0
Max	√			√	45.6	40.4
Max	√	√	√	√	**46.6**	**41.0**
Sum	√	√	√	√	45.9	39.7

Table 3. The mIoU of pseudo labels with different paths and their combinations on training and validation set.

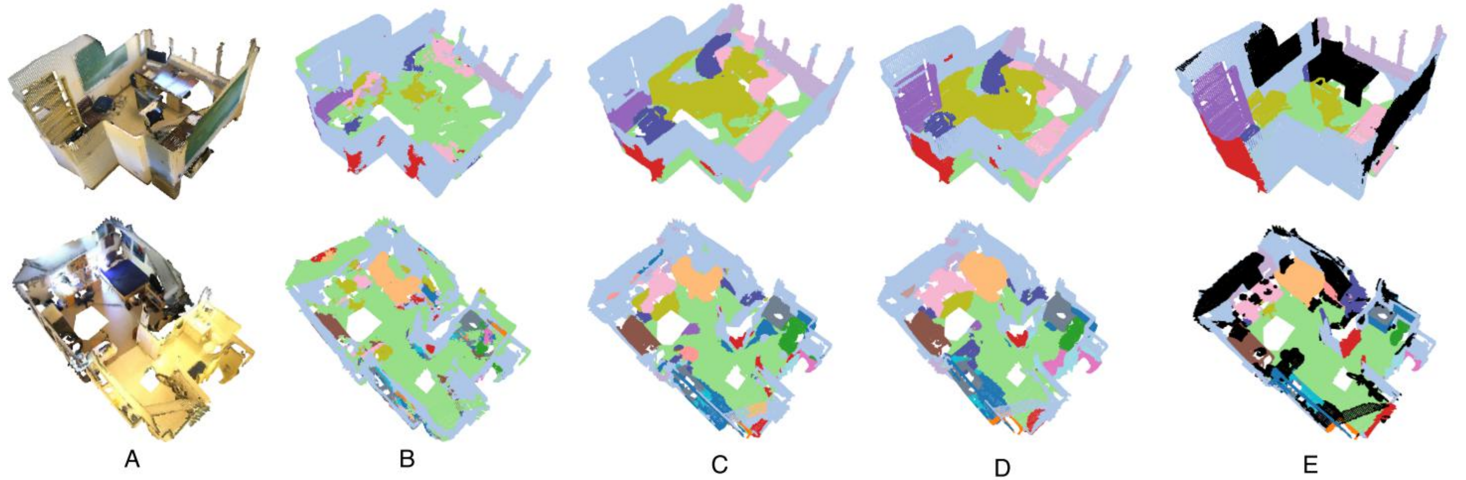


Figure 5. Visualizations of pseudo labels. (A)Input point clouds, (B)PCAMs trained with scene-level labels, (C)PCAMs trained with subcloud-level labels, (D)Multi-path region mining trained with subcloud-level labels, (E)Ground truth.

Paper List

香 港 中 文 大 學（深 圳）
The Chinese University of Hong Kong, Shenzhen

- Multi-Path Region Mining For Weakly Supervised 3D Semantic Segmentation on Point Clouds (CVPR2020)
- SESS: Self-Ensembling Semi-Supervised 3D Object Detection (CVPR2020 oral)
- Fusion-Aware Point Convolution for Online Semantic 3D Scene Segmentation (CVPR2020 best paper final list)

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

SHENZHEN RESEARCH INSTITUTE OF BIG DATA
SRIBD

SESS: Self-Ensembling Semi-Supervised 3D Object Detection

Na Zhao Tat-Seng Chua Gim Hee Lee
Deaprtment of Computer Science, National University of Singapore
{nazhao, chuats, gimhee.lee}@comp.nus.edu.sg

SESS

Semi-supervised learning:

- **Self-training:** Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks (ICML 2013)

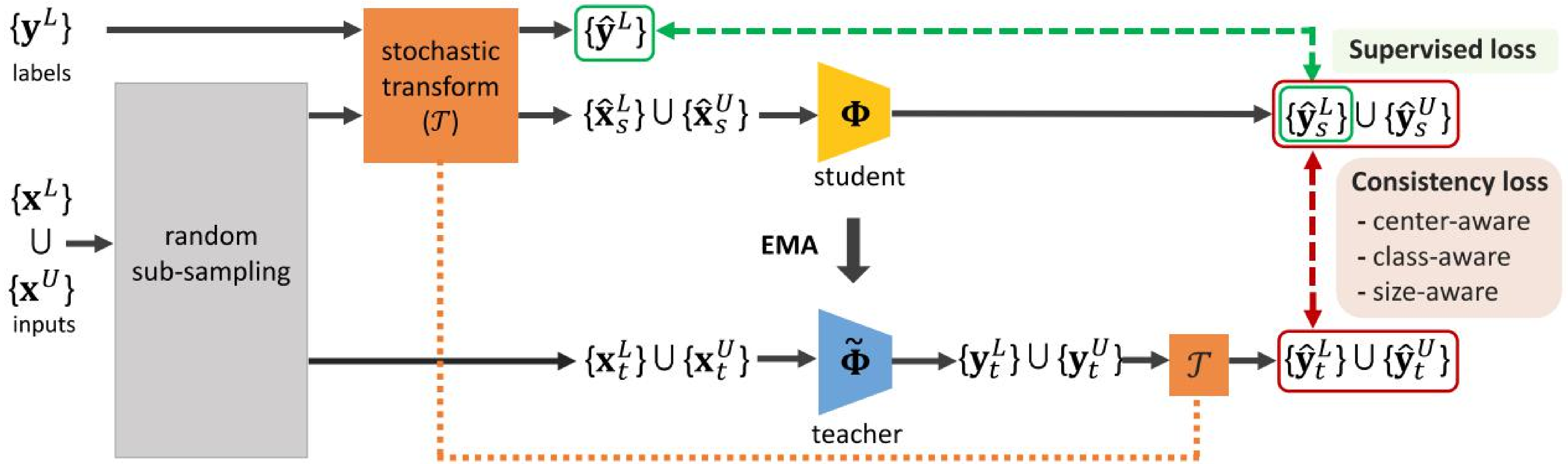
$$L = \sum_{m=1}^{n} \sum_{i=1}^{C} L(y_i^m, f_i^m) + \alpha(t) \sum_{m=1}^{n'} \sum_{i=1}^{C} L({y'}_i^m, {f'}_i^m)$$

- **Temporal ensembling:** Temporal Ensembling for Semi-supervised Learning (ICLR 2017)

- **Self ensembling:** Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results (NIPS 2017)

SESS

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen



SESS

- **Perturbation Scheme:**

  Random Sub-sampling & Stochastic Transform (flipping, rotation and scaling)

- **Consistency Loss:**

  Center-aware consistency loss:
  $$\mathcal{L}_{center} = \frac{\sum_{\hat{c}_s} \|\hat{c}_s - \hat{c}_t^{\mathcal{A}}\|_2 + \sum_{\hat{c}_t} \|\hat{c}_t - \hat{c}_s^{\mathcal{A}}\|_2}{|\hat{C}_s| + |\hat{C}_t|}, \tag{5}$$

  Class-aware consistency loss:
  $$\mathcal{L}_{class} = \frac{1}{|\hat{P}_t|} \sum D_{KL}(\hat{p}_s^{\mathcal{A}} \parallel \hat{p}_t). \tag{6}$$

  Size-aware consistency loss:
  $$\mathcal{L}_{size} = \frac{1}{|\hat{D}_t|} \sum (\hat{d}_s^{\mathcal{A}} - \hat{d}_t)^2. \tag{7}$$

SESS

- **Implementation Details**

(1) VoteNet as backbone

(2) Randomly sub-sampled points is 4,000

(3) Ramp up the coefficient of consistency cost from 0 to its maximum value of 10 during the first 30 epochs.

(4) Pre-train VoteNet with all the available labeled samples. Then initialize the student and teacher networks with the pre-trained weights, and train the student network on both the labeled and unlabeled data by minimizing the supervised loss as well as consistency loss.

SESS

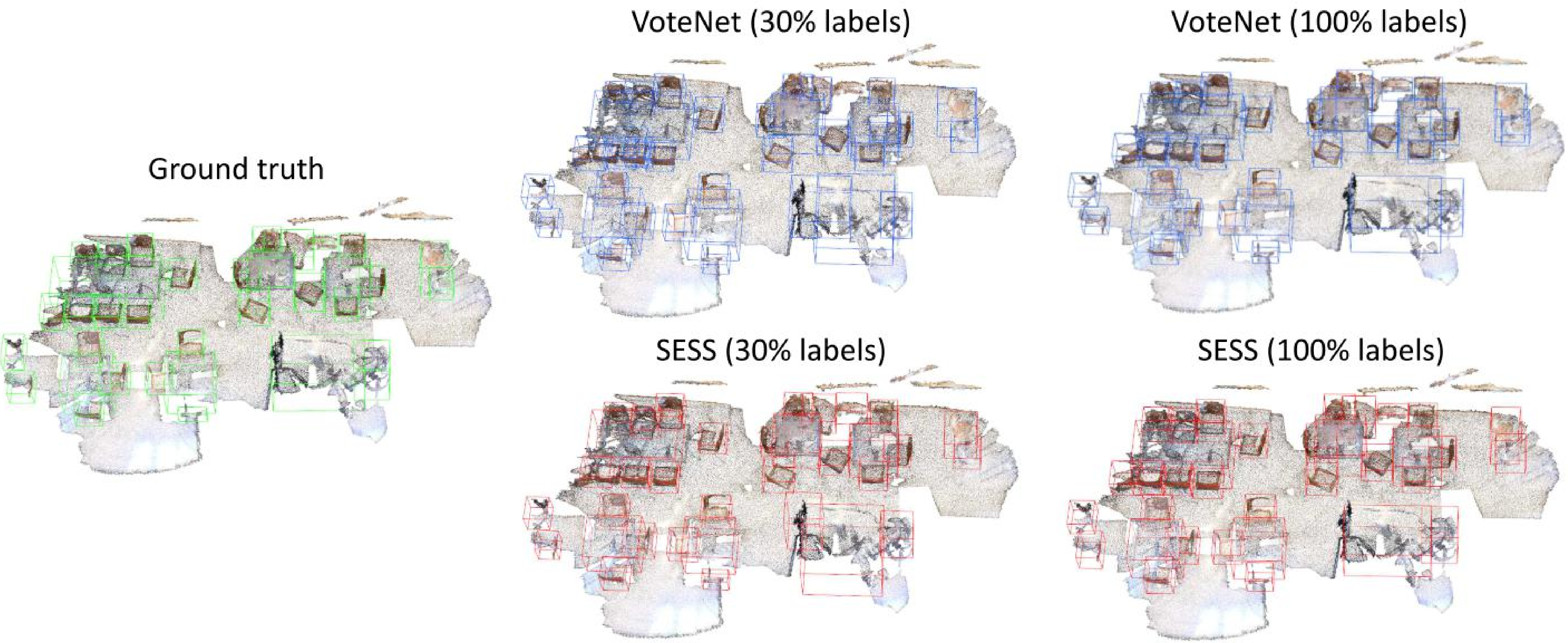
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Table 1: Comparison with VoteNet on SUN RGB-D val set and ScanNetV2 val set with varying ratios of labeled data. mAP@0.25 are reported as mean±standard deviation, based on 3 runs with random sampling. And the improvement (Improv.) is computed based on the mean performances over 3 runs. Note that our SESS is initialized by the VoteNet weights pre-trained on the corresponding labeled data.

Dataset	Model	10%	20%	30%	40%	50%	70%	100%
SUNRGB-D	VoteNet [11]	34.43±1.07	41.13±0.36	47.70±0.17	50.77±0.19	52.5±0.19	56.13±0.18	57.7
	SESS	**42.87±1.01**	**47.87±0.48**	**53.17±0.63**	**54.73±0.26**	**56.37±0.22**	**58.97±0.17**	**61.1**
	Improv.(%)	24.51↑	16.39↑	11.47↑	7.80↑	7.37↑	5.06↑	5.89↑
ScanNetV2	VoteNet [11]	30.97±0.79	41.60±0.46	45.57±0.38	49.2±0.33	52.57±0.07	54.97±0.07	58.6
	SESS	**39.67±0.91**	**47.93±0.39**	**52.20±0.09**	**54.93±0.27**	**57.77±0.41**	**59.20±0.08**	**62.1**
	Improv.(%)	28.09↑	15.22↑	14.55 ↑	11.64 ↑	9.89 ↑	7.70 ↑	5.97↑

SESS

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen



SESS

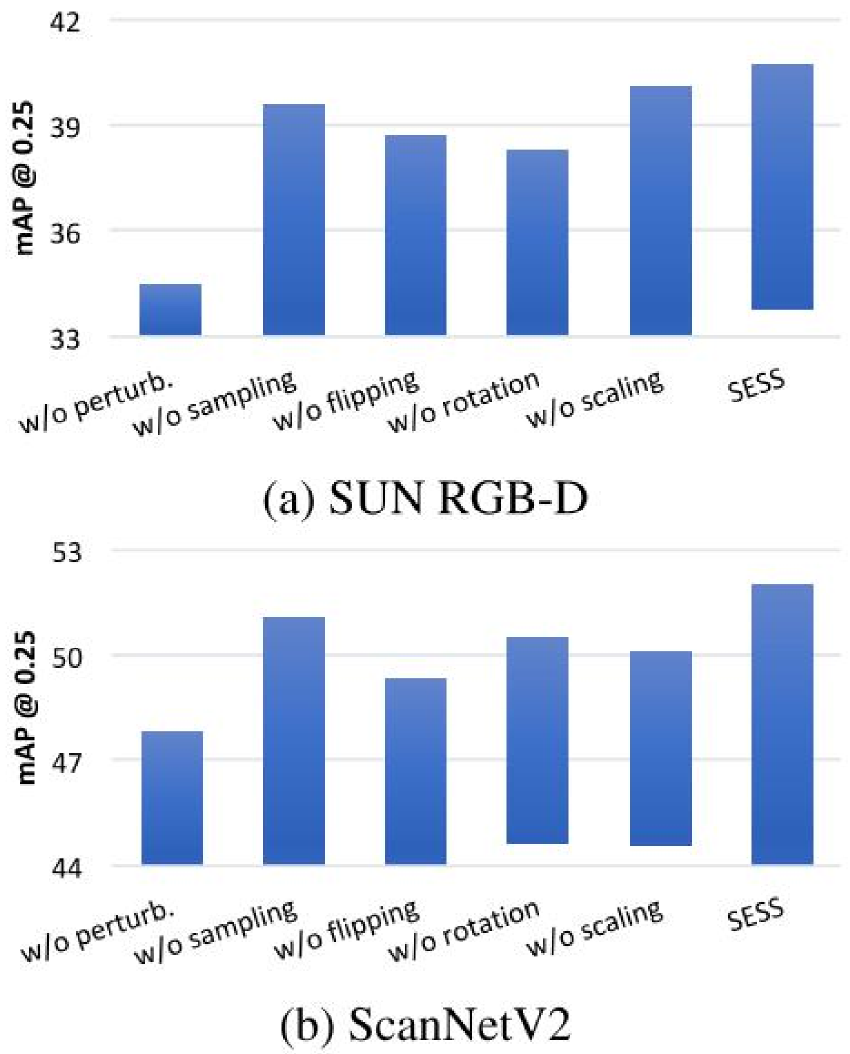


Figure 3: Effects of different perturbations.

Table 4: Ablation study on consistency losses.

center	class	size	SUN RGB-D	ScanNetV2
✓	✗	✗	38.2	50.0
✗	✓	✗	39.2	50.2
✗	✗	✓	38.1	49.2
✗	✓	✓	40.3	50.7
✓	✗	✓	38.9	50.5
✓	✓	✗	40.0	51.5
✓	✓	✓	**40.7**	**52.0**

The training of ablation experiments is conducted on SUN RGB-D with 10% labeled data and ScanNetV2 with of 30% labeled data.

Paper List

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

- Multi-Path Region Mining For Weakly Supervised 3D Semantic Segmentation on Point Clouds (CVPR2020)
- SESS: Self-Ensembling Semi-Supervised 3D Object Detection (CVPR2020 oral)
- Fusion-Aware Point Convolution for Online Semantic 3D Scene Segmentation (CVPR2020 best paper final list)

FusionAwareConv

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Fusion-Aware Point Convolution for Online Semantic 3D Scene Segmentation

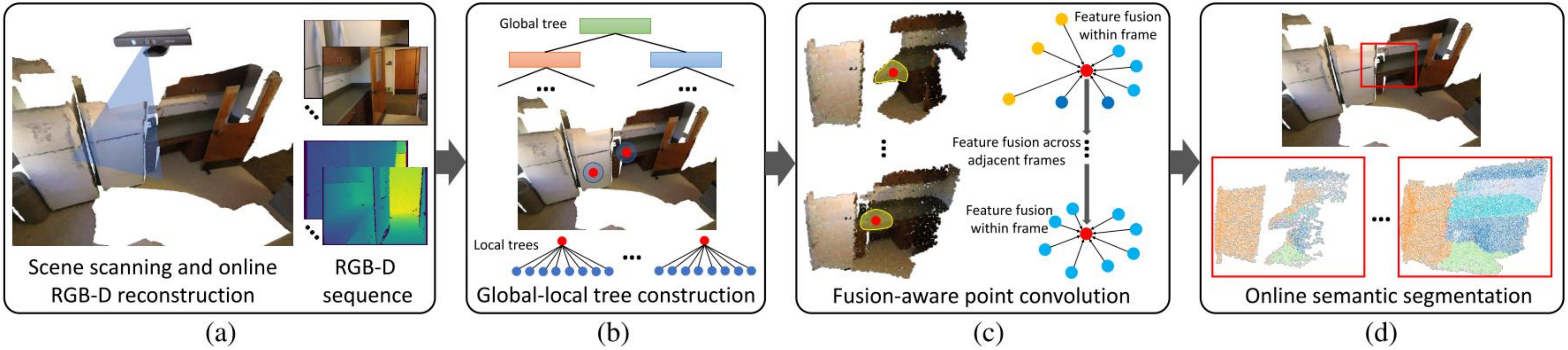
Jiazhao Zhang[1,*] Chenyang Zhu[1,*] Lintao Zheng[1] Kai Xu[1,2†]
[1]National University of Defense Technology [2]SpeedBot Robotics Ltd.

FusionAwareConv

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

FusionAwareConv

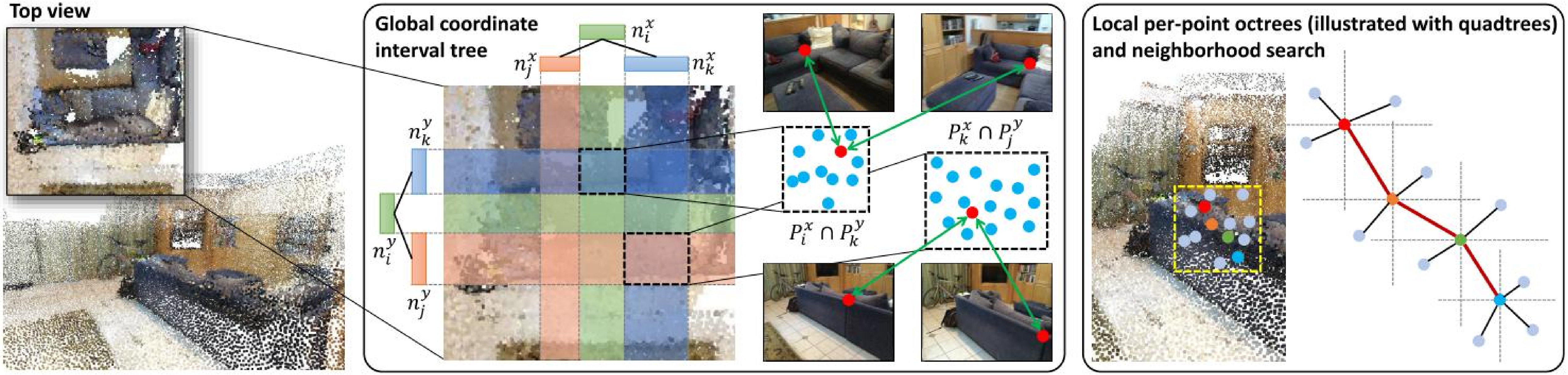
Pipeline



FusionAwareConv

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

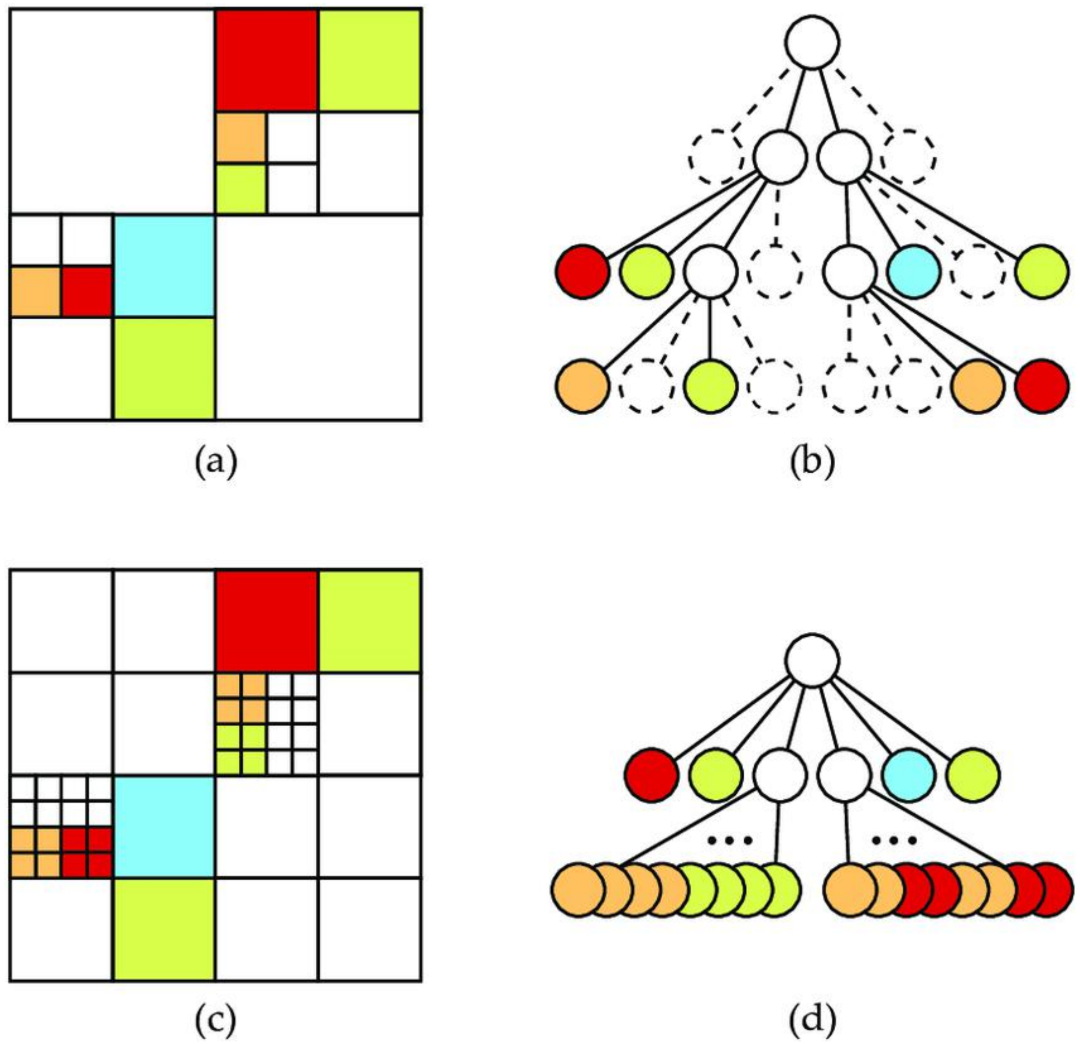
Global-local Tree

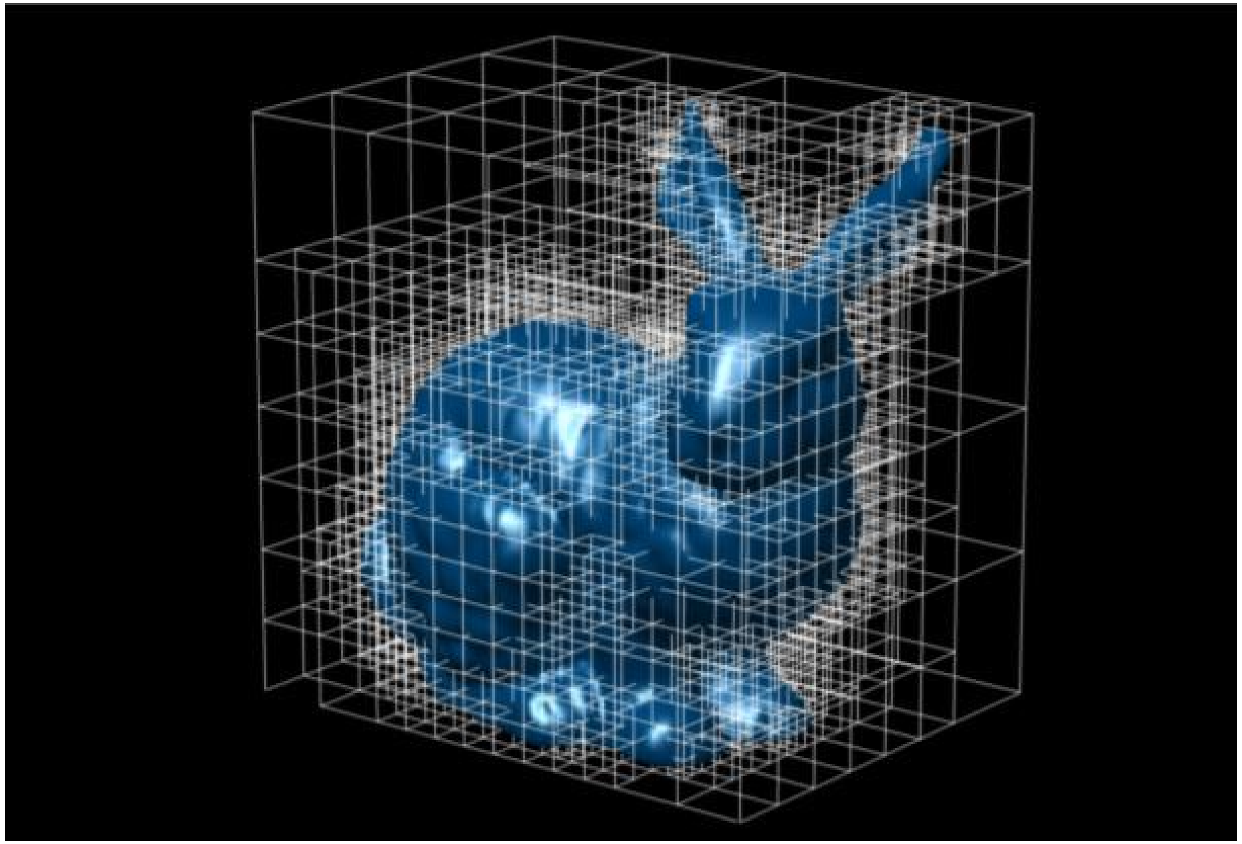


- Global Tree: Coordinate interval trees $x_{\max}(n_l) < x_{\min}(n_p),\ x_{\max}(n_p) < x_{\min}(n_r),$
- Local Tree: Octree

FusionAwareConv

Octree





FusionAwareConv

Fusion-aware Point Convolution

- 3D feature fusion: PointConv:

$$\mathrm{PC}_p(W, F) = \sum_{\Delta p \in \Omega} W(\Delta p) F(p + \Delta p),$$

- 2D-3D feature fusion: Pre-trained FuseNet

$$F^{2D}(c^k) = \mathrm{FuseNet}(f_k, c^k),$$

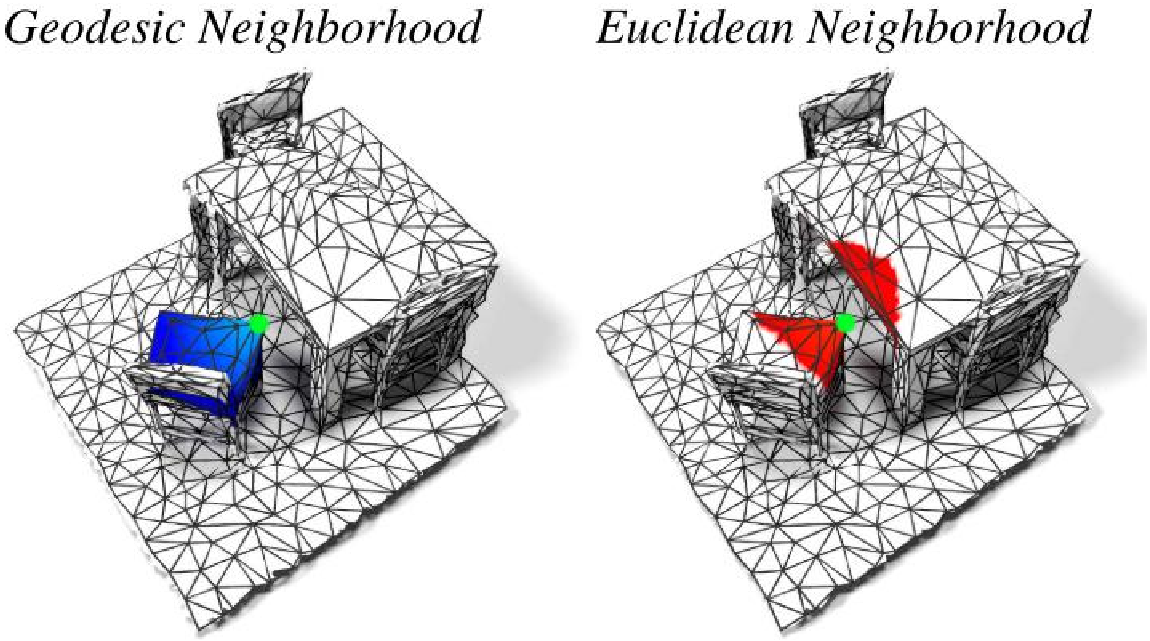
$$F^{\mathrm{2D3D}}(p) = \mathrm{maxpooling}\{F^{2D}(c^k) | c^k \in I(p)\}$$

FusionAwareConv

Fusion-aware Point Convolution

- Octree-induced surface-aware 3D convolution. (Geodesic Neighbors)

$$\mathrm{FPC}_p(W, F^{\mathrm{2D3D}}) = \sum_{\Delta p \in \Omega^n(p)} W(\Delta p) F^{\mathrm{2D3D}}(p + \Delta p).$$



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

FusionAwareConv

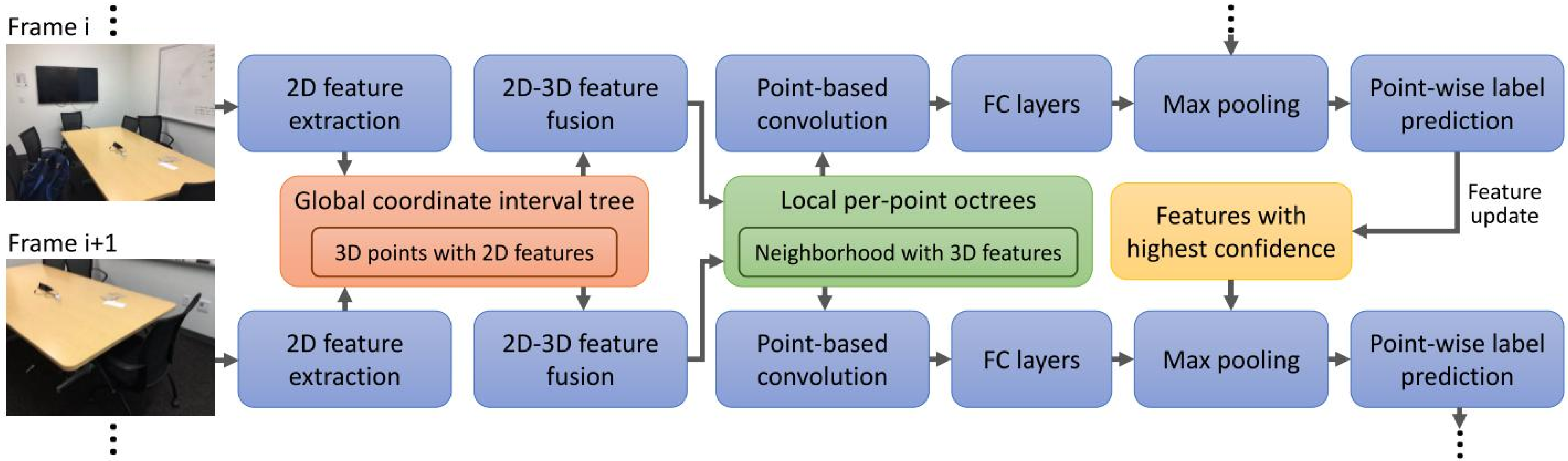
Frame-to-frame feature fusion.

- If prediction has low segmentation uncertainty in the last frame, the current form of feature fusion should be useful in the future prediction.
- Record every uncertainty U(p, i) when processing the frame sequence.

$$F^{\text{fused}}(p) =$$

$$\text{maxpool}\{\text{FPC}_p^{\text{current}}(W, F^{\text{2D3D}}), \underset{\text{FPC}_p^i}{\arg\min}\, U(p, i)(W, F^{\text{2D3D}})\}$$

FusionAwareConv



FusionAwareConv

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Results

- Online methods

Table 1: Accuracy comparison between our method and two state-of-the-art online scene segmentation methods.

Dataset	SemanticFusion [20]	ProgressiveFusion [23]	Ours
ScanNet	0.518	0.566	**0.764**
SceneNN	0.628	0.666	**0.675**

ScanNet: use 1200 sequences for training and the rest 312 for testing

SceneNN: 15 sequences for evaluation

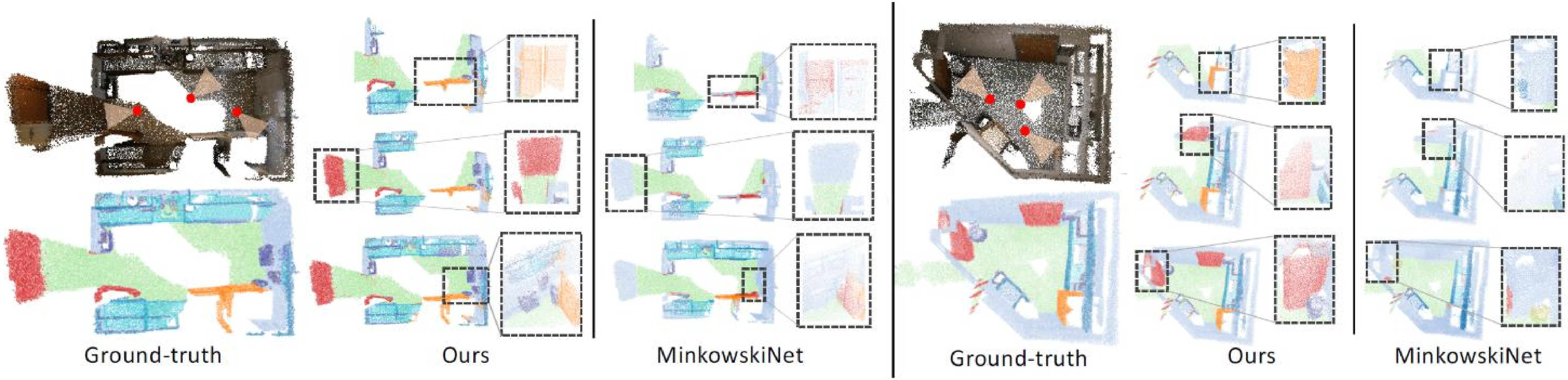
- Offline methods

Table 2: IOU comparison between our method and state-of-the-art offline scene segmentation methods. Our method has the highest mean IOU, outperforming the state-of-the-art methods for nine semantic categories.

model	**mean**	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	bathshade	toilet	sink	bathtub	others
SparseConvNet	0.685	0.828	**0.950**	0.620	0.805	0.894	0.825	0.707	0.633	0.588	0.788	0.252	0.601	0.592	0.681	0.428	0.607	**0.928**	0.596	0.881	0.504
MinkowskiNet	0.715	0.841	0.949	**0.641**	0.806	**0.900**	**0.845**	**0.745**	0.648	0.608	**0.792**	0.289	**0.637**	**0.65**	0.742	0.509	0.690	0.916	**0.689**	0.832	**0.570**
PointConv	0.580	0.741	0.948	0.474	0.672	0.813	0.633	0.651	0.346	0.446	0.713	0.067	0.568	0.525	0.551	0.370	0.520	0.840	0.590	0.750	0.387
Ours	**0.720**	**0.862**	0.924	0.615	**0.848**	0.716	0.804	0.637	**0.680**	**0.698**	0.724	**0.513**	0.617	0.588	**0.764**	**0.734**	**0.696**	0.870	0.681	**0.885**	0.556

FusionAwareConv

- Visualization



- Ablation

2D3D feature	frame-to-frame feature	mean IOU
×	×	0.711
✓	×	0.718
×	✓	0.713
✓	✓	**0.720**

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Thanks for watching!

Xu Yan
2020.6.26