JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Tsinghua University

LONG RANGE POOLING FOR 3D LARGE-SCALE SCENE UNDERSTANDING

XIANG-LI LI[1] MENG-HAO GUO[1] TAI-JIANG MU[1] RALPH R. MARTIN[2] SHI-MIN HU[1]

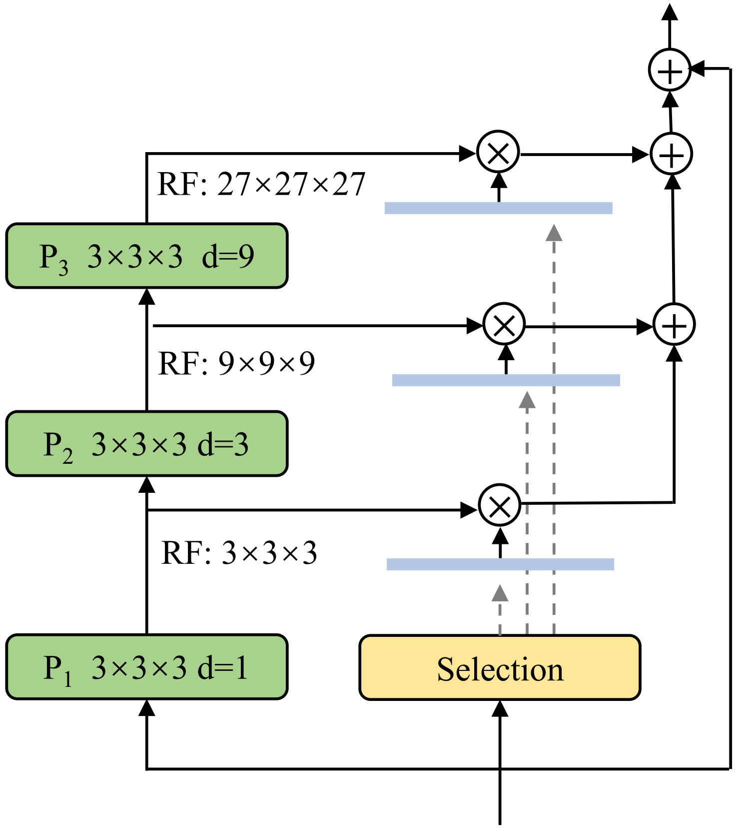
[1]BNRIST, TSINGHUA UNIVERSITY [2]CARDIFF UNIVERSITY

WED-AM-198

QUICK PREVIEW

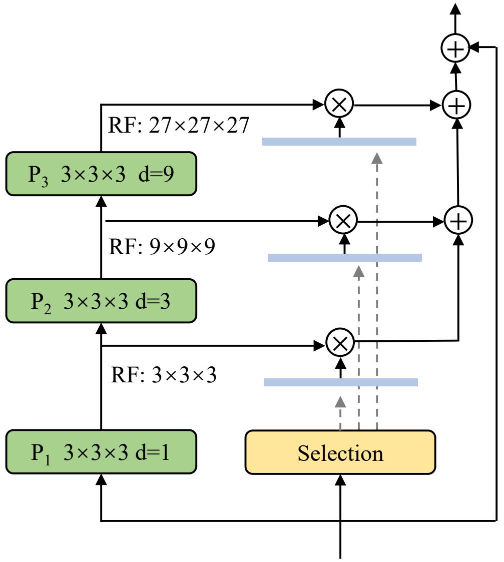
QUICK PREVIEW

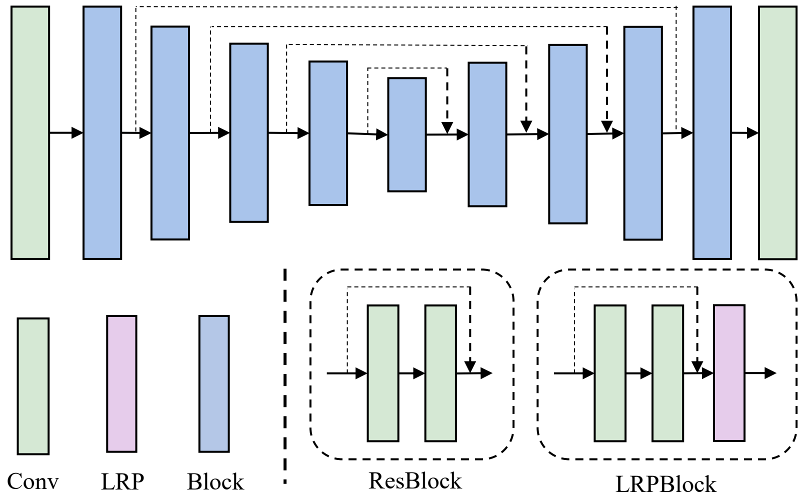
- We propose a simple and effective module, the long range pooling (LRP) module by using dilation max pooling.
  - ➢ Providing a network with a large adaptive receptive field.
  - ➢ Introducing more operations with greater non-linearity.



QUICK PREVIEW

- LRPNet is constructed by straightforwardly incorporating the LRP module after each block of the baseline model.
  - LRPBlock is constructed by adding the LRP module after the ResBlock.
  - The ResBlock is replaced with the LRPBlock module.





QUICK PREVIEW

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

- LRPNet achieves superior 3D segmentation results on various large-scale 3D scene benchmarks.

Method	Params (M)	Runtime (ms)	mIoU (%)
SparseConvNet	30.1	173.5	69.3
MinkowskiNet	37.9	166.1	72.2
Fast Point Transformer	37.9	341.4	72.0
Stratified Transformer	18.8	1149.9	74.3
Baseline	8.1	38.1	71.2
LRPNet (Ours)	8.5	67.9	75.0

Method	Input	Val	Test
PointNet++	point	53.5	55.7
3DMV [26]	point	-	48.4
PointCNN	point	-	45.8
PointConv	point	61.0	66.6
JointPointBased	point	69.2	63.4
PointASNL	point	63.5	66.6
RandLA-Net	point	-	64.5
KPConv	point	69.2	68.6
PointTransformer	point	70.6	-
SparseConvNet	voxel	69.3	72.5
MinkowskiNet	voxel	72.2	73.6
LargeKernel3D	voxel	73.2	73.9
Fast Point Transformer	voxel	72.1	-
Stratified Transformer	point	74.3	73.7
LRPNet (ours)	voxel	**75.0**	**74.2**

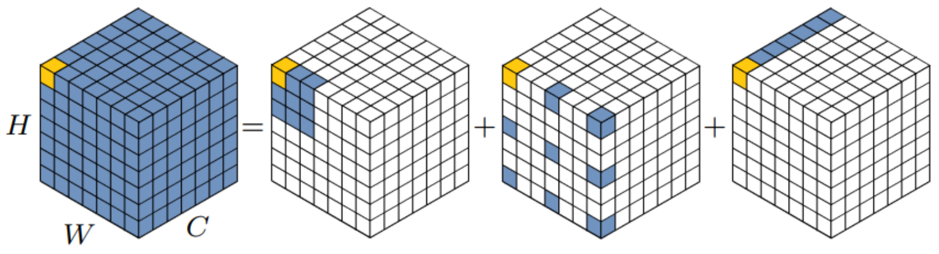
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

INTRODUCTION

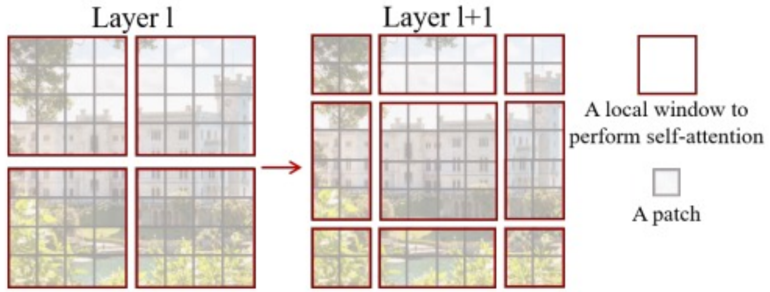
INTRODUCTION

Background The success of transformers and large-kernel cnns

- A large receptive field.
- Operations with greater non-linearity.



Decomposition of large-kernel convolution in VAN[1]



The shifted window approach in swin transformer[2]

[1] Guo, Meng-Hao, et al. "Visual attention network." *arXiv preprint arXiv:2202.09741* (2022).
[2] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

INTRODUCTION

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation How to apply large kernel design in 3D?

Large-Kernel 3D convolution ❯ Cubic increase in parameters and computational complexity

⬇

- ➢ Employing dilation operations to decompose the large kernel.
- ➢ Replacing convolution with pooling operations.

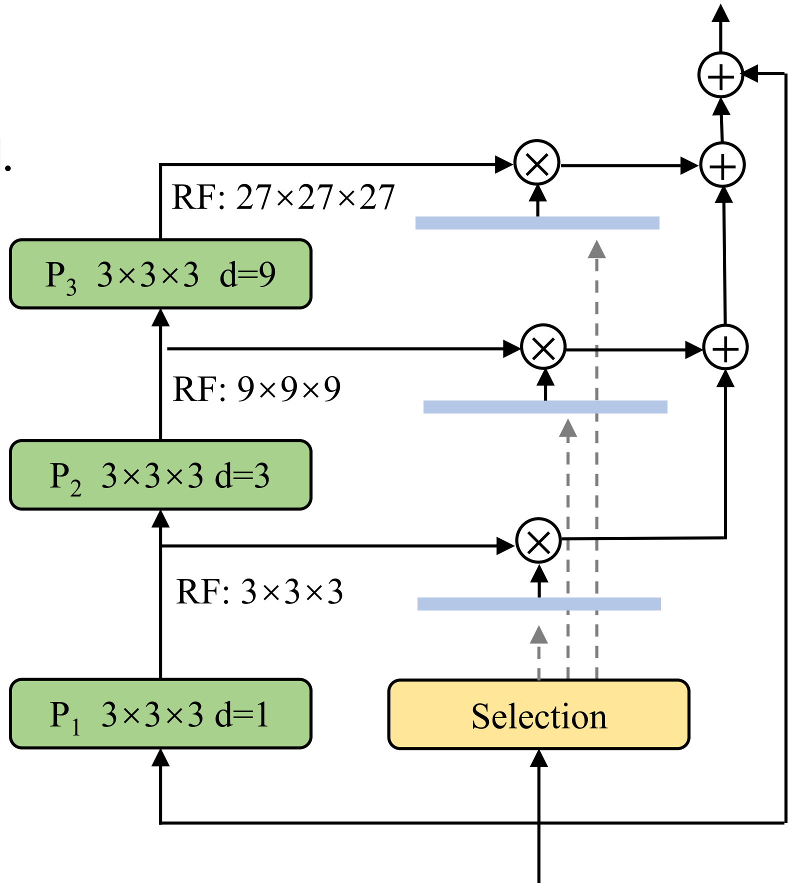
METHOD

METHOD

LRP Module

- Dilation max pooling is employed to expand the receptive field.
- Selection module produces selection weights for each voxel.

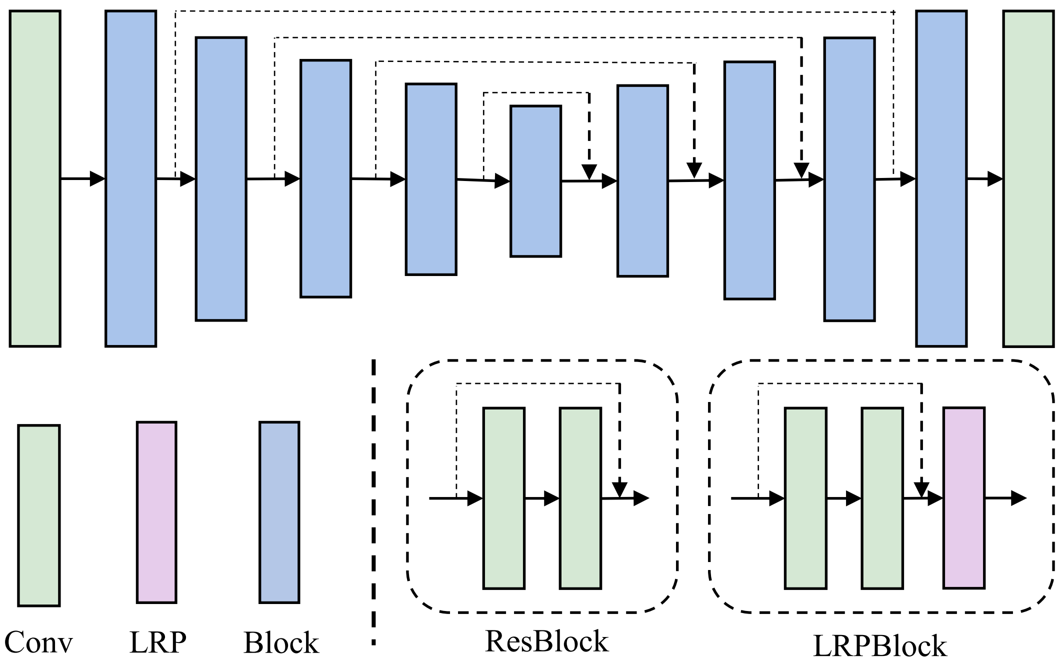
$$S = (S_1, S_2, S_3) = \mathrm{linear}(x),$$
$$P_1 = p(x, 1), P_2 = p(P_1, 3), P_3 = p(P_2, 9),$$
$$Output = S_1 P_1 + S_2 P_2 + S_3 P_3,$$



METHOD

LRPNet

- LRPBlock is constructed by adding the LRP module after the ResBlock.
- We use the sparse convolutional U-Net as our baseline, consisting of four stages.
- We replace the ResBlock with LRPBlock to build LRPNet.



JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

EXPERIMENTS

EXPERIMENTS

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Comparisons

Experiments on ScanNet and S3DIS

Method	Input	Val	Test
PointNet++	point	53.5	55.7
3DMV [26]	point	-	48.4
PointCNN	point	-	45.8
PointConv	point	61.0	66.6
JointPointBased	point	69.2	63.4
PointASNL	point	63.5	66.6
RandLA-Net	point	-	64.5
KPConv	point	69.2	68.6
PointTransformer	point	70.6	-
SparseConvNet	voxel	69.3	72.5
MinkowskiNet	voxel	72.2	73.6
LargeKernel3D	voxel	73.2	73.9
Fast Point Transformer	voxel	72.1	-
Stratified Transformer	point	74.3	73.7
LRPNet (ours)	voxel	**75.0**	**74.2**

Table 1. mIoU (%) scores for various methods on the ScanNet v2 3D semantic benchmark, for validation and test sets. The best number is in boldface. “-” means the number is unavailable.

Method	Input	OA	mAcc	mIoU
PointNet	point	-	49.0	41.1
SegCloud	point	-	57.4	48.9
TangentConv	point	-	62.2	52.6
PointCNN	point	85.9	63.9	57.3
HPEIN	point	87.2	68.3	61.9
GACNet	point	87.8	-	62.9
PAT	point	-	70.8	60.1
ParamConv	point	-	67.0	58.3
SPGraph	point	86.4	66.5	58.0
PCT	point	-	67.7	61.3
SegGCN	point	88.2	70.4	63.6
PAConv	point	-	-	66.6
KPConv	point	-	72.8	67.1
MinkowskiNet	voxel	-	71.7	65.4
Fast Point Transformer	voxel	-	77.3	70.1
PointTransformer	point	90.8	76.5	70.4
Stratified Transformer	point	**91.5**	**78.1**	**72.0**
LRPNet (ours)	voxel	90.8	74.9	69.1

Table 2. Several scores (%) for various methods on the S3DIS segmentation benchmark. The best number is in boldface. “-” means the number is unavailable

EXPERIMENTS

Ablation Study on LRP module

Method	Params (M)	Runtime (ms)	mIoU (%)
SparseConvNet	30.1	173.5	69.3
MinkowskiNet	37.9	166.1	72.2
Fast Point Transformer	37.9	341.4	72.0
Stratified Transformer	18.8	1149.9	74.3
Baseline	8.1	38.1	71.2
LRPNet (Ours)	8.5	67.9	75.0

Table 3. Number of network parameters and speed for various models on the ScanNet v2 validation set.

Baseline	MaxPool	Dilation	Selection	Params (M)	Runtime (ms)	mIoU
✓				8.1	38.1	71.2
✓	✓			8.1	74.2	72.6
✓	✓	✓		8.1	65.0	73.7
✓	✓		✓	8.5	76.4	73.1
✓	✓	✓	✓	8.5	67.9	75.0

Table 4. Ablation study on LRP module. Baseline: U-Net described in section 3.3. MaxPool: The max pooling used in LRP. Dilation: Dilation used in max pooling (default for LRP is 1, 3, 9). Selection: the selection module in LRP. Params: parameters of the network. Runtime: the average time for inferencing one scene. mIoU: the segmentation accuracy metric (%)

EXPERIMENTS

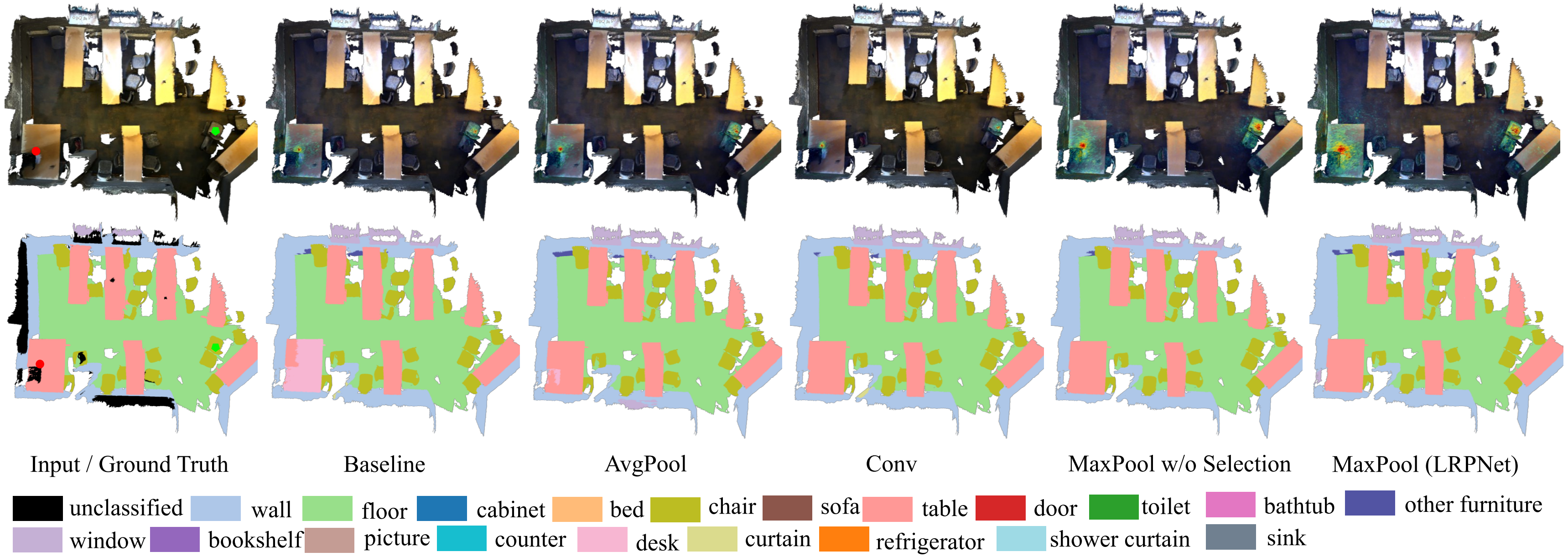
Ablation

Study on non-linearity and receptive field

Method	Range	Params (M)	Runtime (ms)	mIoU (%)
MaxPool	[×3]	8.2	66.0	72.7
MaxPool	[×9]	8.2	64.8	72.8
MaxPool	[×27]	8.2	66.1	73.9
MaxPool	[×9, ×27]	8.4	66.0	74.0
MaxPool	[×3, ×9, ×27]	8.5	67.9	75.0
AvgPool	[×27]	8.2	58.2	73.0
AvgPool	[×9, ×27]	8.4	58.2	73.2
AvgPool	[×3, ×9, ×27]	8.5	60.1	73.9
Conv	[×27]	17.4	65.5	72.0
Conv	[×9, ×27]	17.5	66.2	73.3
Conv	[×3, ×9, ×27]	17.6	66.8	73.8

EXPERIMENTS

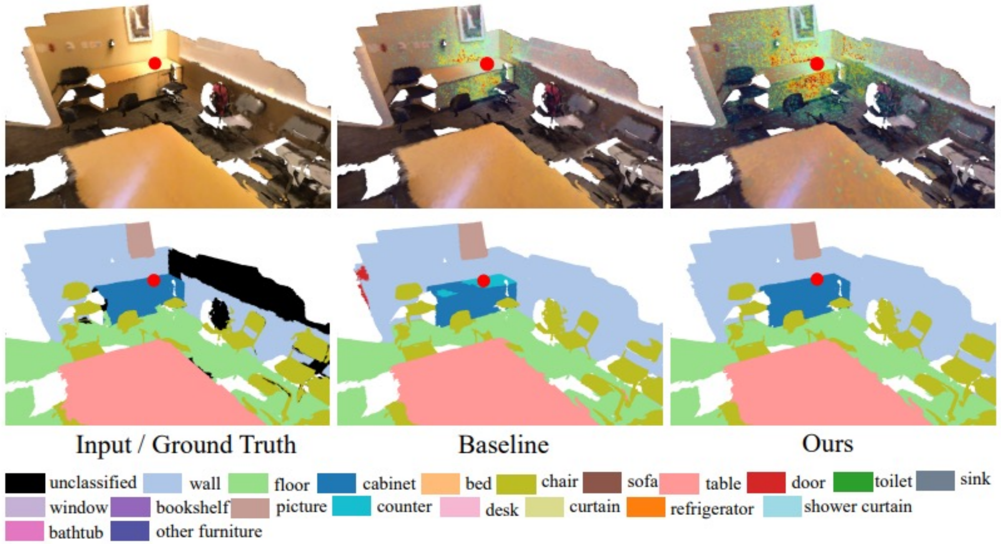
Qualitative Comparisons of ERF



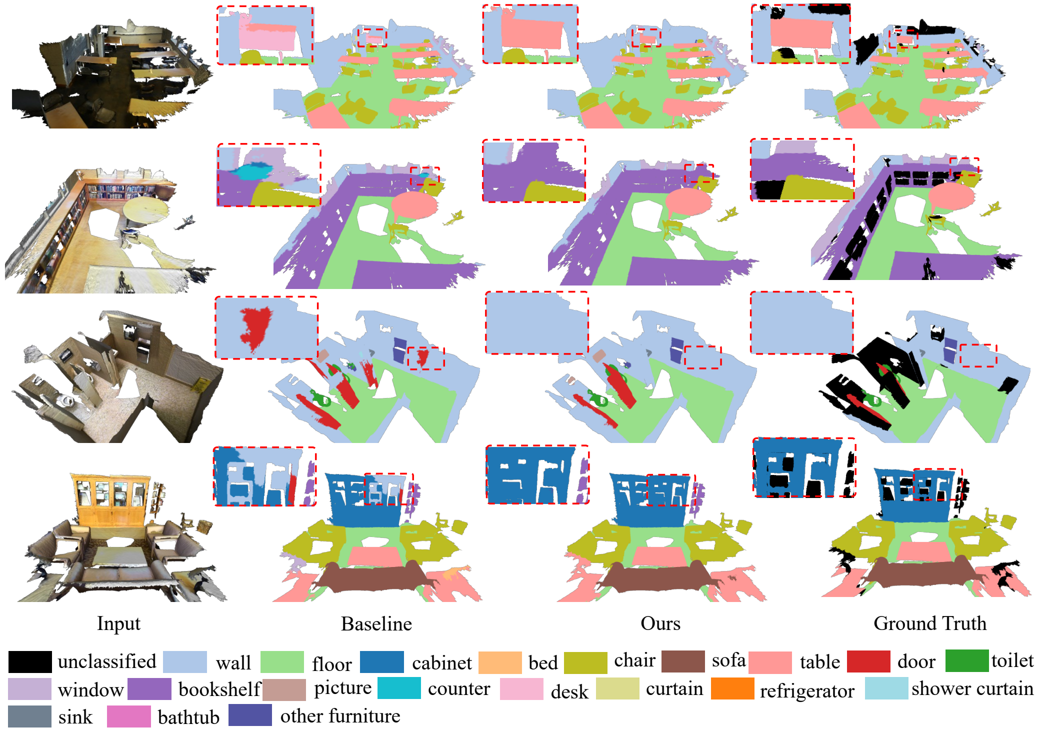
Red Dot and Green Pentagon: two different positions of interest (marked only in the input). First Row: the ERFs of different methods. Second Row: the ground truth and predictions. From Left to Right: input, baseline, average pooling, convolution, max pooling without selection, max pooling (LRPNet).

EXPERIMENTS

Qualitative Analysis



Left: the input and ground truth. Middle: the ERF and results of. Right: the ERF and results of LRPNet



Red dotted boxes highlight differences between our results and the baseline results

CONCLUSION

CONCLUSION

- We proposed a simple and effective module, the **long range pooling (LRP)** module, which provides a network with a large adaptive receptive field without a large number of parameters.
- Our experiments indicate that a **larger receptive field** and aggregation operations with **greater non-linearity** enhance the capacity of a sparse convolution network.
- We constructed LRPNet, a simple sparse convolution network using the LRP module, which achieves superior 3D segmentation results on various large-scale 3D scene benchmarks

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

清華大學 Tsinghua University

THANK YOU !