Investigating Attention Mechanism in 3D Point Cloud Object Detection

Shi Qiu *,1,2, Yunfan Wu *,1, Saeed Anwar 1,2,3, Chongyi Li 4
1 The Australian National University; 2 Data61-CSIRO, Australia;
3 University of Technology Sydney; 4 Nanyang Technological University

3DV 2021
International Conference on 3D Vision
Online, December 1-3, 2021

Background

Point Cloud Object Detection:

◇ The aim is to predict the class label and 3D bounding box for each object of the given point cloud scene.

◇ The standard approaches for 3D object detection can be categorized into two streams: the region proposal-based and single-shot methods.

◇ Due to 3D data's sparsity and unorderedness, specially designed networks and modules are needed to process the point clouds.

◇ However, it is unclear how attention modules would affect the performance of 3D point cloud object detection and what sort of attention modules could fit with the inherent properties of 3D data.

Contributions:

◇ We push the VoteNet pipeline towards better performance by integrating attention mechanisms into it.

◇ We comprehensively evaluate the performances of *ten* recent attention modules on *SUN RGB-D* and *ScanNetV2* datasets.

◇ We summarize the effects and characters of different attention modules and provide novel insights to facilitate the understanding of the attention mechanism for 3D point cloud object detection.

Attention Structures

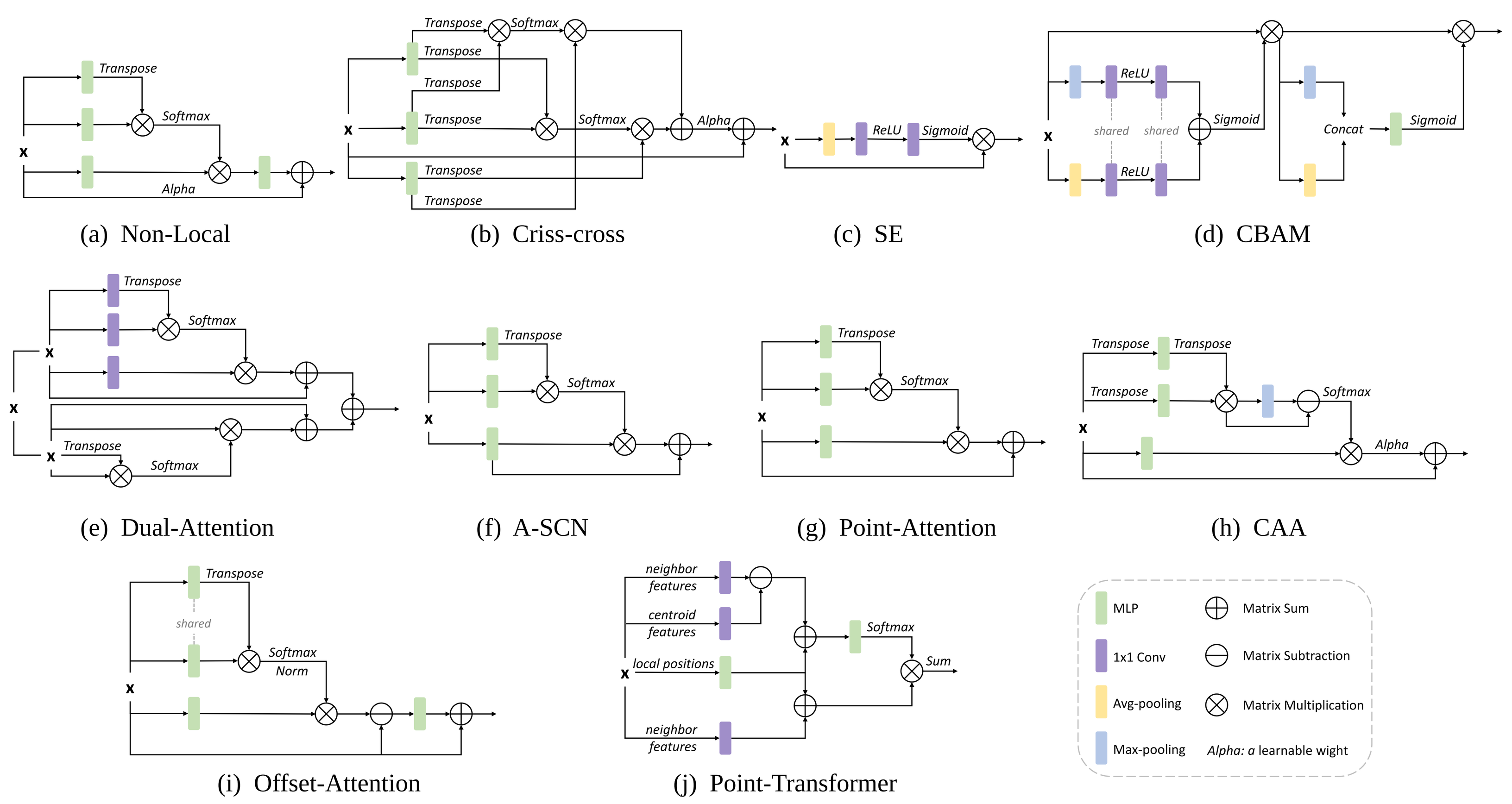
2D Attentions:

◇ Non-Local ◇ Criss-cross ◇ SE ◇ CBAM ◇ Dual-Attention

3D Attentions:

◇ A-SCN ◇ Point-Attention ◇ CAA ◇ Offset-Attention
◇ Point-Transformer

Detailed Structures:

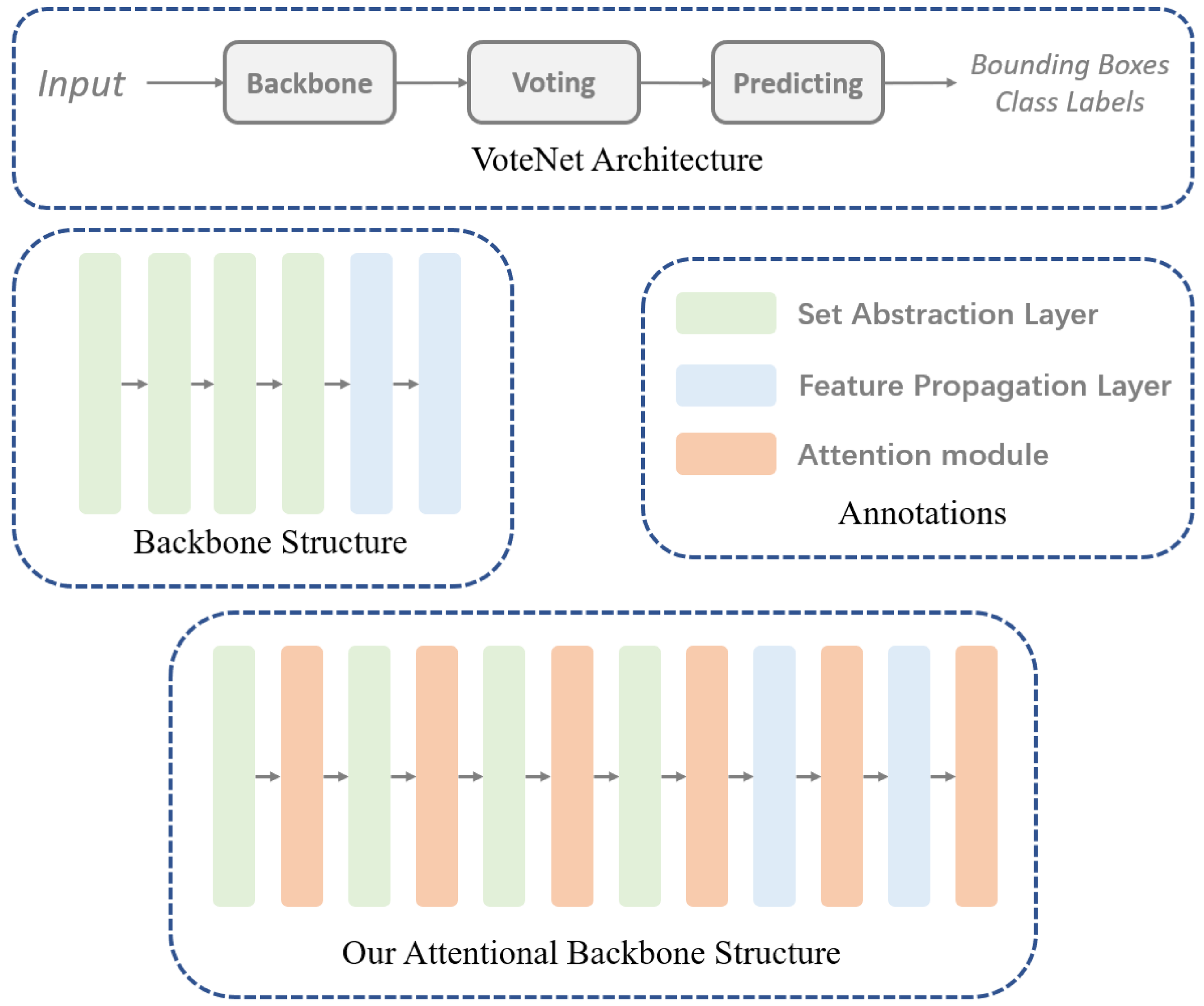


(a) Non-Local (b) Criss-cross (c) SE (d) CBAM (e) Dual-Attention (f) A-SCN (g) Point-Attention (h) CAA (i) Offset-Attention (j) Point-Transformer

Attentional Object Detection

VoteNet Pipeline: consists of a backbone that learns features, a voting module estimating the object centers, as well as a predicting module regressing the bounding boxes and class labels.

Attentional Backbone: the attention module is placed after each encoder and decoder of the backbone.



VoteNet Architecture

Backbone Structure

Annotations

Our Attentional Backbone Structure

Experiments

Experimental Results:

	SUN RGB-D Dataset				*ScanNetV2 Dataset*			
Method	mAP@0.25	AR@0.25	mAP@0.5	AR@0.5	mAP@0.25	AR@0.25	mAP@0.5	AR@0.5
VoteNet	57.5	86.1	33.1	51.1	57.3	80.0	33.7	49.9
Non-local	58.3	86.0	31.4	49.7	57.5	80.9	34.6	49.5
Criss-cross	56.2	84.9	33.1	50.0	56.7	79.1	33.8	49.2
SE	**59.6**	**86.8**	34.5	52.1	58.6	80.3	35.8	51.4
CBAM	59.1	86.3	**34.9**	**53.1**	58.7	**81.0**	37.1	52.5
Dual-attn	50.6	82.9	24.4	42.1	54.7	80.2	30.2	47.2
A-SCN	55.6	84.0	30.1	48.2	56.5	80.6	33.1	48.7
Point-attn	56.4	84.3	32.2	49.7	54.7	79.4	30.8	46.7
CAA	58.8	85.9	33.3	51.4	57.6	80.6	35.1	50.4
Point-trans	58.5	85.8	34.3	51.3	**59.1**	80.3	**38.0**	**53.5**
Offset-attn	55.7	84.6	30.6	48.2	58.0	79.9	36.0	50.4

Model Complexity:

Method	**model size** (MB)	**training time** (s/epoch)	**inference time** (s/epoch)	**# parameters** ($\times 10^3$/attention*)
VoteNet	11.0	43.8	35.0	-
Non-local	13.0	48.2	35.9	8.5
Criss-cross	16.0	54.6	35.2	20.8
SE	11.9	44.2	35.1	4.1
CBAM	11.5	45.7	36.4	4.1
Dual-attn	15.9	50.6	36.7	21.0
A-SCN	16.0	48.5	35.9	20.8
Point-attn	16.0	48.6	35.6	20.8
CAA	34.7	47.2	36.7	106.6
Point-trans	25.8	88.1	38.7	100.1
Offset-attn	19.5	50.1	35.3	35.6

Insights

◇ The self-attention modules: (i) its fashion needs high computational resources, and (ii) the effectiveness of point-wise long-range dependencies is relatively limited as such an operation may cause some redundancies in representing the large-scale 3D data.

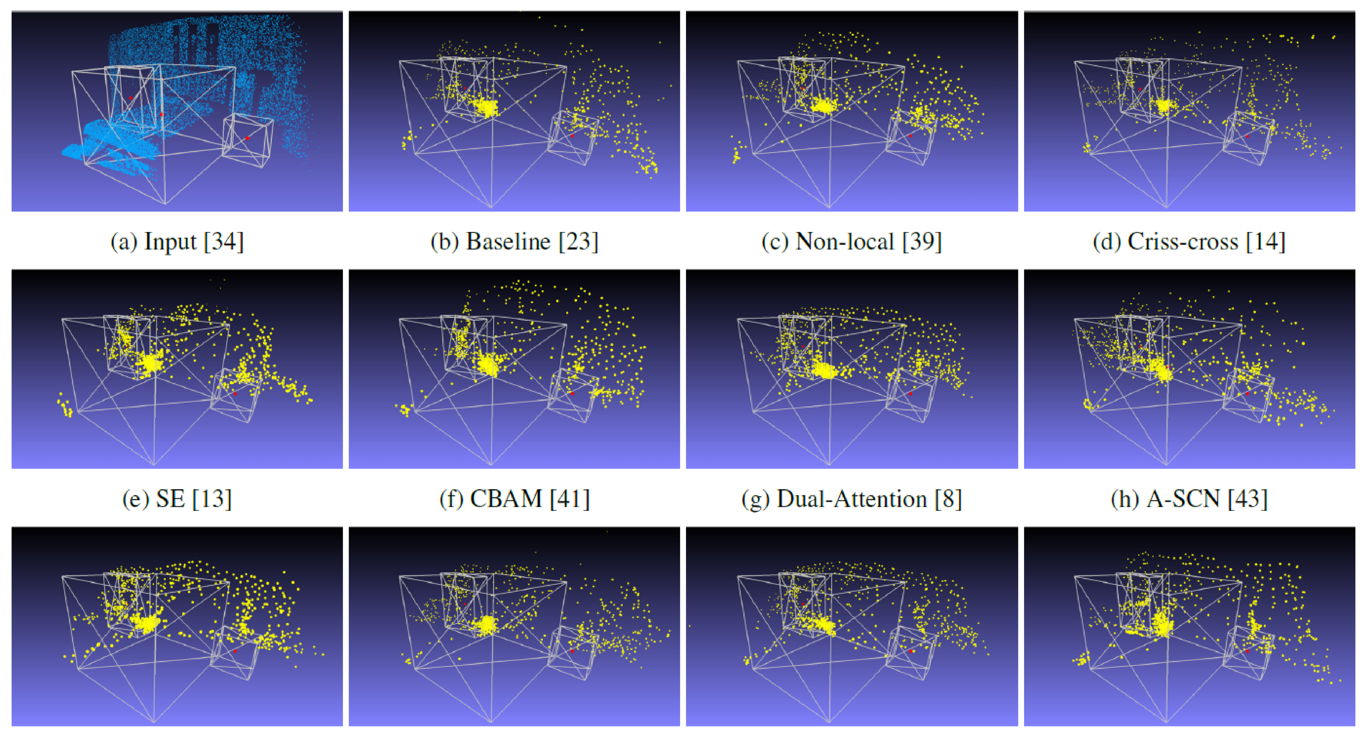
◇ The compact attention structures like SE and CBAM enable the effectiveness and efficiency of 3D point cloud feature refinement. This is achieved by capturing the global perception in feature space.

◇ Comparing the spatial-attention with the channel-attention modules: the channel-related information is more important when embedded into the attention modules for point cloud feature representations.

◇ As reflected from the Point Transformer's results, incorporating more local context could better represent the complex point cloud scenes, thus leading to better 3D point cloud object detection performance.

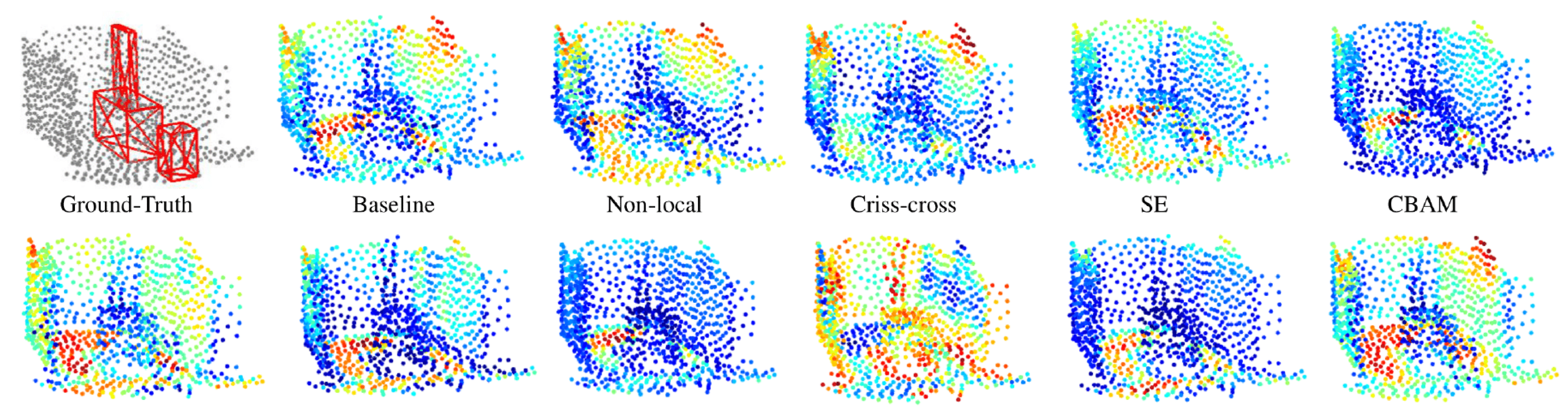
Visualization

Votes Visualization:
The bounding boxes (ground-truths) are drawn in white frames, where the generated votes (yellow points) are expected to be around the centroids (red points) of detected objects as many as possible.



(a) Input [34] (b) Baseline [23] (c) Non-local [39] (d) Criss-cross [14] (e) SE [13] (f) CBAM [41] (g) Dual-Attention [8] (h) A-SCN [43] (i) Point-Attention [7] (j) CAA [28] (k) Offset-Attention [10] (l) Point-Transformer [51]

Feature Visualization:
The features are learned from different attentional backbones, where the channels of feature map are normalized and averaged to be illustrated in a heat map view.



Ground-Truth Baseline Non-local Criss-cross SE CBAM Dual-Attention A-SCN Point-Attention CAA Offset-Attention Point-Transformer