An Innovative n-Shifted Sigmoid Channel and Spatial Attention Module for Efficient 3D Scene Object Detection

Shengzhi Du *, Desire Mulindwa Burume, Qingxue Liu

Posted Date: 29 May 2024

doi: 10.20944/preprints202405.1941.v1

Keywords: attention mechanism; Hough voting; point clouds; activation function

Article

An Innovative n-Shifted Sigmoid Channel and Spatial Attention Module for Efficient 3D Scene Object Detection

Shengzhi Du [1,*], Desire Burume Mulindwa [1] and Qingxue Liu [2]

[1] Department of Electrical Engineering, Tshwane University of Technology
[2] School of Mechanical and Electrical Engineering, Kunming University
* Correspondence: dushengzhi@gmail.com

Abstract: Recently, attention mechanisms have developed into an important tool for performance improvement of deep neural networks. In computer vision, attention mechanisms are generally divided into two main branches: spatial and channel attention. Both attention categories have their own advantages. The fusion of both attentions achieves higher performance, on the cost of the computational load. This paper introduces an innovative and lighter n-shifted sigmoid channel and spatial attention (CSA) module to reduce the computational cost and to improve the 3D scene relevant features selection. To validate the proposed attention module, 3D scene object detection in the deep Hough voting point sets is considered as the testing application. The proposed attention module with its piecewise n-shifted sigmoid activation function improves the network's learning and generalization capacity which effectively predict bounding box parameters directly from 3D scenes and detect objects more accurately. This advantage is achieved by selectively attending to more relevant features of the input data. When used in the deep Hough voting point sets, the proposed attention module outperforms state-of-the-art 3D detection methods on the sizable SUNRGBD dataset. Experiments conducted showed an increase of 12.02 mean accuracy precision (mAP) when compared to the celebrated VoteNet (without attention). It also got 9.92 mAP higher compared to the MLVCNet, and 10.32 mAP higher than the Point Transformer. The proposed model not only decreases the sigmoid vanishing gradient problem but also brings out valuable features by fusing channel-wise and spatial information while improving accuracy results in 3D object detection.

Keywords: attention mechanism; Hough voting; point clouds; activation function

1. Introduction

Attention mechanisms have been attracting increasing attention in research communities since they focus on key features while suppressing redundant ones [1–3]. Latest studies have demonstrated that correctly integrating attention mechanisms into convolution blocks substantially enhance the performance of a wide range of computer vision tasks such as image classification, object detection, instance segmentation, etc.

In computer vision, the attention mechanisms are divided into two main types: channel attention and spatial attention. Recent studies such as GCNet [1] and CBAM [10] have combined both channel attention and spatial attention to achieve significant improvement in object detection [16]. However, these models commonly suffered from either substantial computational burdens or converging challenges. Despite that, other researchers were able to simplify the structure for both channel and spatial attention like ECA-Net [7] which makes the process of computing channel weights in SE block much easier by using a 1-D convolution. SGE [8] groups the channels dimension into several sub-features to symbolize different semantics and implements a spatial module to every feature group through a feature vectors scale over all locations with an attention mask.

 doi:10.20944/preprints202405.1941.v1

The main question is to find out if it is possible to fuse different attention modules in a lighter but more efficient manner. ShuffleNet v2 [9] is the first to attempt an answer to the question because it can efficiently construct a multi-branch structure and process various divisions in parallel. Subsequently, several convolution layers are adopted to capture a higher-level representation of the input. Then, the two divisions are concatenated to render the number of channels and the number of input similar. Finally, the "channel shuffle" operator (defined in [10]) is adopted to allow for information communication between the two divisions. Moreover, SGE [8] introduces a grouping strategy to improve calculation speed, which divides the input feature map into groups following channel dimensions.

This paper proposes to introduce a novel n-shifted channel and spatial attention module to be used in the well celebrated Votenet [12]. To our knowledge this will be the first time a deep Hough voting model for object detection will be used in conjunction with any attention module to improve detection accuracy while selectively attending to more relevant features of the input data.

For both spatial and channel attentions, the authors use an n-shifted sigmoid gating approach as an activation function for best function approximation and faster convergence. This allows a better modelling of the interconnections between the channels, the preservation of meaningful features while subduing less beneficial features.

This paper introduces a n-shifted sigmoid channel and spatial attention module to deduce 3D bounding boxes of the objects in the scene and suggest object proposals from a point cloud focused 3D object detection. This model is based on recent advances in 3D deep learning models for point clouds and is inspired by both an innovative channel and spatial attention module and the generalized Hough voting process [13]. Table 1 describes the advantages of the proposed model when compared to the existing ones on feature extraction, attention benefits, predictions, and accuracy. In the proposed model, PointNet++ [14], a point cloud deep learning model, alleviates the need to convert point clouds to regular structures.

Table 1. Comparison of benefits of n shifted channel and spatial attention in VoteNet to the other methods in SOTA in 3D object detection.

n-sigmoid Channel and Spatial Attention VoteNet	Other related Networks
Enhanced feature refinement	Standard feature extraction
Improved attention focus	Traditional attention
More robust predictions	Moderate detection accuracy
Higher detection accuracy	Decent object localization and limited refinement

As the point cloud generated by depth sensors only captures surfaces of objects, 3D object centers are likely to be in empty space, far away from any point. As a result, point based networks have difficulty aggregating scene context in the vicinity of object centers. To increase the capacity of the network to recognize objects in the 3D scene, an n-shifted sigmoid channel and spatial attention module is added to the deep learning model.

In essence, this paper proposes to endow a deep point sets Hough voting network with a combined n-shifted piecewise sigmoid gating mechanism. By voting, the new method essentially generates new points that lie close to objects centers, which can be grouped and aggregated to generate box proposals.

Specifically, after passing the input point cloud through a backbone network, the authors sample a set of seed points and generate votes from their features. Votes are targeted to reach object centers. As a result, vote clusters emerge near object centers and in turn can be aggregated through an efficient learned module to generate box proposals. The result is a powerful 3D object detector that is purely geometric and can be applied directly to point clouds. The authors validate this approach on the SUN RGBD dataset [15].

In summary, the contributions of this work are:

 doi:10.20944/preprints202405.1941.v1

1. To introduce an efficient n-shifted piecewise sigmoid channel and spatial 3D attention module to improve the network's learning and generalization capability while reducing the inherent sigmoid vanishing gradient problem.
2. To demonstrate that the proposed attention module can just be plugged into existing models and boost their performance.
3. To validate that performance is greatly improved in the deep point sets Hough voting process using the SUNRGB-D dataset by inserting the proposed lightweight attention module.

The rest of this paper is arranged as follows. In Section 2, the related works are discussed. In Section 3, the proposed methodology is presented, where the n-sigmoid CSA deep Hough voting network is thoroughly described for a more accurate 3D objects detection. In Section 4, the implementation details are discussed followed by the experiments and their results in Section 5. Section 6 offers the conclusions and future works.

2. Related Works

Attention mechanisms. The attention has been exploried extensively since it predisposes the distribution of the most informative features while suppressing the less useful ones. Squeeze-and-Excitation (SE) [16] developed channel-wise relationships using two FC layers. ECA-Net [7] implemented a 1-D convolution filter to produce channel weights and meaningfully decreased the SE model complexity. Zhu et al. [17] proposed the non-local module that computes the correlation matrix between each spatial point in an extensive attention map. CBAM [18], GCNet [1], and SGE [8] fused the channel attention and spatial attention in series, while DANet [2] adaptively incorporated local features with their global dependencies.

Feature grouping. Attention mechanisms allow a model to concentrate on specific parts of the input data, while feature grouping aggregates all the pertinent features to extract and exhibit more information. Together, attention mechanism and feature grouping enable the extraction of meaningful patterns and relationships within complex data.

The transformer [19] architecture employed self-attention to establish dependencies between different positions in a sequence, enabling the model to focus on relevant context while processing the input. This attention mechanism, when combined with feature grouping strategies, helps in capturing long-range dependencies and contextual information, leading to improved performance in various natural language processing tasks. Bi-LS-AttM [20] showcased the application of attention mechanism for image captioning, emphasizing how attention facilitates the grouping of relevant image features for generating descriptive captions. These papers emphasize the key role of attention mechanisms in leading feature grouping strategies while enhancing performance in various computer vision tasks.

Activation functions. Activation functions cause nonlinearity in neural networks [21]. Conventional activation functions (such as sigmoid and tanh) are continuous and differentiable, but the sigmoid has only a positive value, while the tanh has a negative one [22]. Most enhancements of the sigmoid function generally focus on varying the slope of the sigmoid or shifting the original sigmoid, as opposed to the new proposed sigmoid that is a piecewise log-shifted function in a finite input–output space. The sigmoid activation function is often used in feed-forward neural networks (FFNN) [23] to introduce nonlinearity. To accelerate network convergence, CNNs mostly use the hyperbolic tangent as an activation function. One of the latest advances in activation functions is the non-negative rectified linear unit (ReLU), where the identity map in the positive portion solved the gradient vanishing challenge [22].

3. The Proposed n-Sigmoid Channel Spatial Attention

3.1. 3D n-Sigmoid Channel and Spatial Attention Mechanism

Attention mechanisms, which enable a neural network to accurately focus on all the relevant elements of the input, have become an essential component to improve the performance of deep neural networks.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2024 doi:10.20944/preprints202405.1941.v1

In this paper, the authors propose a lighter but more efficient n-shifted sigmoid channel and spatial attention module to improve computational overhead where both spatial and channel attentions are combined and to enhance 3D scene relevant/important features selection (Figure 1). This module is integrated into the VoteNet architecture to improve the model's ability to handle complex 3D object detection.

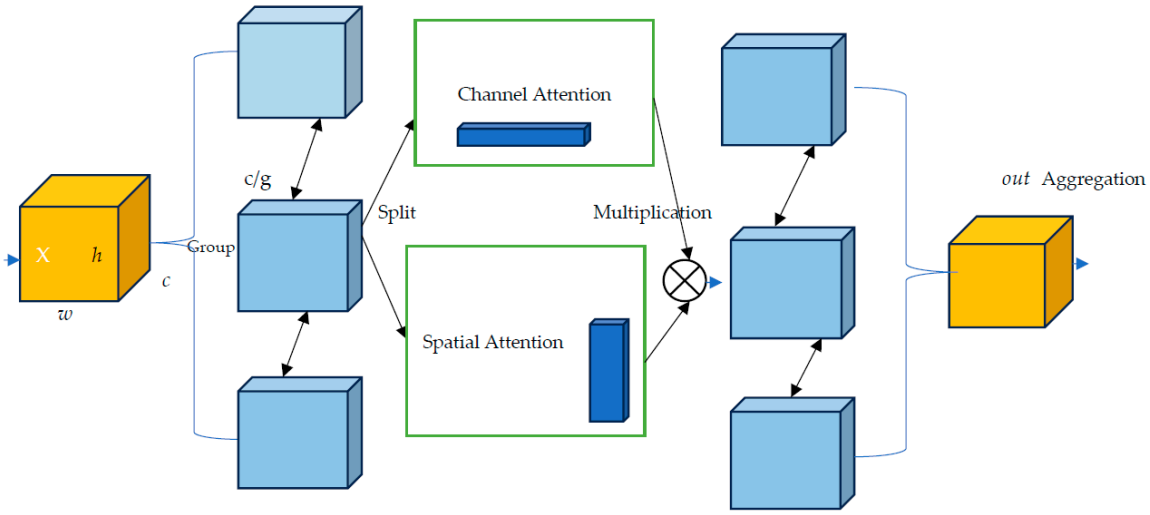

Figure 1. An overview of the proposed attention module. The module adopts a channel split to process the features of each group in parallel. For the channel attention branch, both the average and max pooling generate improve channel-wise statistics and use a pair of parameters to scale the channel vector. For spatial attention branch, the module adopts group norm to generate spatial-wise statistics and use a compact feature. The two branches are then multiplied together to emphasize regions where both channel and spatial attentions are high, potentially focusing more on salient features.

In Figure 1, the multiplication is an elementwise product between channel and spatial attention. The n-shifted sigmoid CSA does split the channel to allow parallel processing of each group sub-features. For the channel attention branch, it uses a pair of parameters to scale and shift the channel vector. For spatial attention branch, it adopts a group normalization to generate spatial-wise statistics. The two branches are then multiplied together elementwise before all sub-features are aggregated.

The new n-sigmoid CSA layer strategically combines channel and spatial attention mechanisms to enable the network to focus on crucial features while preserving spatial information. Specifically, the CSA layer is designed to carry significant improvements to the feature learning process, which plays a pivotal role in accurately predicting and localizing 3D objects in point clouds [24].

Contrary to common usage in image related CNNs, here the attention module is not repeatedly placed after each encoder (Set Abstraction) and decoder (Feature Propagation) of the VoteNet backbone but rather once only after the backbone that learns the features and before the voting module that estimate the object centers.

This innovative procedure is able to improve the accuracy score due to several reasons, including (1) upgraded discriminative features where the integration of the CSA module enhances the discriminative power of the features used in the Hough voting, (2) a context-aware voting where the model adapts its voting strategy based on the learned context, (3) an adaptive attention where the model dynamically adjusts the importance of different votes based on the spatial and channel-wise context.

The CSA layer operates by dividing the feature map $x \in R^{cxhxw}$ where c, h, w are channel number, spatial height, and width. n-shifted sigmoid CSA divides x into g groups according to the channel dimension. At the start of each attention unit, the input of X is split into two branches, namely the channel attention branch and the spatial attention branch (employed in [12]) and, at the end, their their concatenation by elementwise product has an improvement on the accuracy results.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2024 doi:10.20944/preprints202405.1941.v1

The channel attention branch employs average pooling to capture the essence of the input features across different channels (Figure 2).

Beyond the previous works, the authors propose that max-pooling be also simultaneously used to gather another important clue about distinctive object features to infer finer channel-wise attention. Thus, both average-pooled and max-pooled features are used concurrently. Using both features greatly improves the networks representation power more than using each independently.

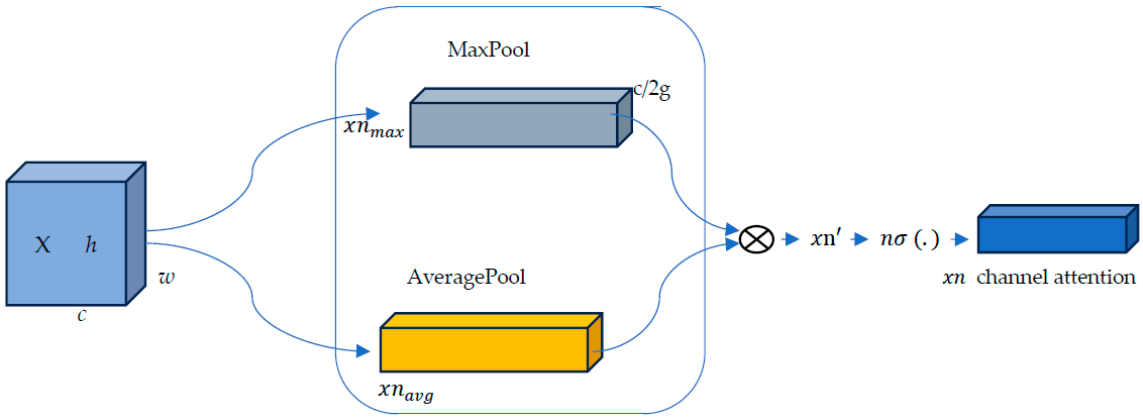

Figure 2. Diagram of the Channel attention sub-module. The channel sub-module uses max-pooling and average pooling where the outputs are multiplied before the n-shifted sigmoid is added together with a pair of parameters to scale the channel vector.

In Figure 2, the channel sub-module utilizes both max-pooling and average-pooling outputs.

The enhanced application for the channel attention branch (shown in Figure 3) is to first use global averaging pooling (Eq. 1) together with the global maximum pooling (Eq. 3) to generate channel-wise statistics as $s \in R^{C/2G \times 1 \times 1}$.

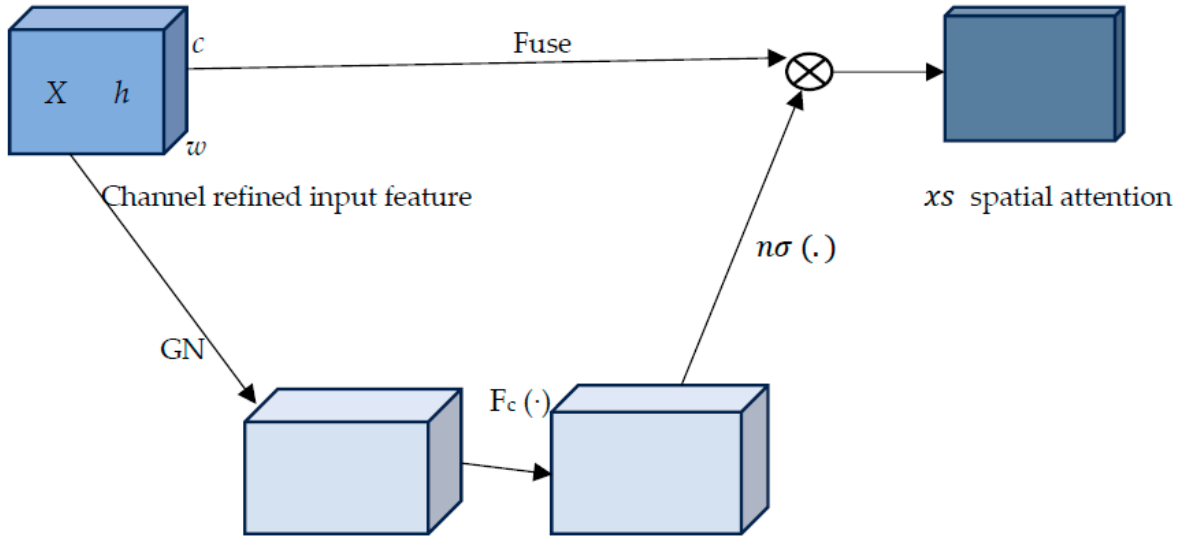

Figure 3. Diagram of the Spatial attention overview. Group normalization (GN) is adopted to generate spatial-wise statistics before a compact feature $F_c(.)$ is created.

This operation consists in:

1. Average Pooling Branch:
   - Average pooling operation

$$xn_{avg} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} xn(i,j) \qquad (1)$$

   - n-shifted sigmoid activation:

 doi:10.20944/preprints202405.1941.v1

$$xn_{avg} = n\sigma\,(xn_{avg}) \quad (2)$$

where $n\sigma$ is the n-shifted activation function

2. Max Pooling Branch:

- Max pooling operation:

$$xn_{avg} = \max i,j(\,xn_{i,j}) \quad (3)$$

- n-shifted sigmoid activation:

$$xn_{max} = n\sigma\,(xn_{max}) \quad (4)$$

where $n\sigma$ is the n-shifted activation function

In both pooling branches, a compact feature is created to enable guidance for precise and adaptive selection. This is achieved by a gating mechanism using n-shifted sigmoid activation [22] on both the average pooling (Eq. 2) and the max pooling (Eq. 4) operations.

The final output of the channel attention (Eq. 5-6) can be obtained by multiplying both averaged and max_pooled tensors and adding the n-shifted sigmoid:

$$xn' = [xn_{avg} \otimes xn_{max}] \quad (5)$$

$$xn = n\sigma\,(W_1\,.\,(xn') + b_1)\,.\,xn' \quad (6)$$

where $W_1 \in R^{c/2g \times 1 \times 1}$, $b_1 \in R^{C/2g \times 1 \times 1}$ are parameters used to scale xn'

This customized non-linear activation function, the n-shifted sigmoid, is tailored to accentuate relevant features while suppressing noise and irrelevant information.

On the other hand, the **spatial attention branch** focuses on "where" most scene informatiton lies, and is complementary to channel attention. At the onstart, a group normalization (GN) [26] is used over x to obtain spatial-wise statistics. F_c are the compact feature generated from the spatial branch. F_c (·) does improve xs. The final output of spatial attention is:

$$xs = n\sigma\,(W_2\,.\,GN(xs) + b_2\,)\,.\,xs \quad (6)$$

where $W_2 \in R^{c/2g \times 1 \times 1}$, $b_2 \in R^{C/2g \times 1 \times 1}$ and are parameters of shape $R^{C/2g \times 1 \times 1}$ $n\sigma$ is the n-sigmoid activation (Eq. 6).

So, channel attention utilizes group normalization GN to process the features, followed by the application of the same n-sigmoid activation function to refine the spatial information. The n-shifted sigmoid activation function exhibits a distinctive behavior that allows for controlled feature enhancement based on a learned scaling factor.

It effectively emphasizes significant features while dampening the impact of less important ones, thereby enabling the model to focus on relevant information critical for accurate 3D object detection. This controlled non-linearity facilitates the n-shifted sigmoid CSA layer in adaptively reshaping the feature space, leading to a more discriminative and informative representation for subsequent stages of the network.

The last but very important step is to multiply both channel and spatial attention (Eq. 7) that will act as a gating mechanism where each channel's importance is modulated by its spatial relevance. Here the channel and spatial attentions are multiplied element-wise to preserve the feature representations influenced by both the the channel and spatial attention mechanisms. This provides a different form of modulation where multiplication is expected to emphasize regions where both channel and spatial attentions are high, potentially focusing more on salient features.

$$out = [xn_{avg} * xn_{max}] \quad (7)$$

where xn is the channel attention and xs is the spatial attention.

Python code of the proposed n-shifted sigmoid CSA

```
import torch
import torch.nn as nn f
rom torch.nn.parameter import Parameter
```

 doi:10.20944/preprints202405.1941.v1


```
class csa_layer(nn.Module):
"""Constructs a Channel Spatial Group module.
Args: k_size: Adaptive selection of kernel size """

def __init__ (self, channel, groups=64):
super (csa_layer, self). __init__ ()
self.groups = groups
self.avg_pool = nn. AdaptiveAvgPool3d (1)
self.max_pool = nn. AdaptiveMaxPool3d (1
)
self.cweight_avg = Parameter(torch.zeros(1, channel // (4 * groups), 1, 1))
self.cbias_avg = Parameter(torch.ones(1, channel // (4 * groups), 1, 1))
self.cweight_max = Parameter(torch.zeros(1, channel // (4 * groups), 1, 1))
self.cbias_max = Parameter(torch.ones(1, channel // (4 * groups), 1, 1))
self.sweight = Parameter(torch.zeros(1, channel // (2 * groups), 1, 1))
self.sbias = Parameter(torch.ones(1, channel // (2 * groups), 1, 1))
self.sigmoid = nn.Sigmoid() self.gn = nn.GroupNorm(channel // (2 * groups), channel // (2 *
groups))

def forward (self, x):
b, c, h, w = x.shape
x = x.reshape(b * self.groups, -1, h, w)
x_0, x_1 = x.chunk(2, dim=1)

# Channel attention
# Average pooling branch
xn_avg = self.avg_pool(x_0)
xn_avg = self.cweight_avg * xn_avg + self.cbias_avg
xn_avg = x_0 * self.sigmoid(xn_avg)

# Max pooling branch
xn_max = self.max_pool(x_0)
xn_max = self.cweight_max * xn_max + self.cbias_max
xn_max = x_0 * self.sigmoid(xn_max)

# Concatenate average and max pooled tensors
xn = torch.cat ([xn_avg, xn_max], dim=1)

# Spatial attention
xs = self.gn(x_1)
xs = self.sweight * xs + self.sbias
xs = self.sigmoid(xs)
# Multiply channel and spatial attentions
out = xn * xs

return out
```

The idea of introducing this new n-shifted sigmoid CSA attention module present several innovations that include:

1. Combination of Average and Max Pooling: Several attention mechanisms normally use either average pooling or max pooling to aggregate channel-wise data. Merging both average and max

 doi:10.20944/preprints202405.1941.v1

pooling allows the model to capture various aspects of the feature maps, thereby enhancing the capacity to attend to relevant features.

2. Multiplying Channel and Spatial: While addition is a common operation for combining attention mechanisms, multiplying channel and spatial attentions together can provide a totally different type of data regarding the scene in question. Multiplication will lay emphasis on regions where both channel and spatial attentions are important, thus focusing more on salient features.
3. The use of n-shifted sigmoid activation as a gating mechanism in the attention module.
4. The flexibility that permits the model to learn various types of interactions between spatial and channel attentions. Furthermore, the exploitation of trainable parameters from both max and average pooling as well as from channel and spatial attention, empowers the model to adaptively learn.

3.2. 3D n-Sigmoid CSA Module Integrated in the Point Cloud Learning Using Hough Voting

The integration of the CSA layer into the VoteNet model (Figure 2) aims at proving the ability of the n-sigmoid attention module to better capture intricate patterns and complex relationships within the point clouds [29].

In the 3D n-sigmoid CSA layer's forward function, a global average pooling is used as a 3D Adaptive Average Pooling operation, allowing the extraction of relevant features from the 3D spatial scene [27]. In this paper, the channel attention structure implies the utilization of the n_sigmoid activation function which enhances feature discrimination and learning adaptability within the volumetric feature space. Additionally, the spatial attention mechanism employs a Group Normalization operation tailored for 3D data, promoting effective feature normalization, and enhancing feature representations within the spatial domain [28]. All these steps contribute to the resulting improved feature discrimination and subsequent feature fusion.

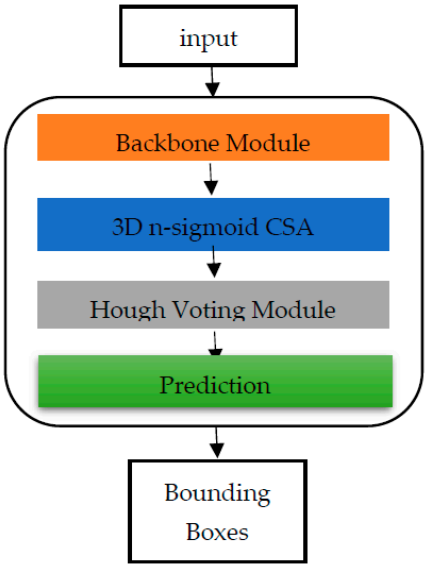

Figure 4. An overview of the n-shifted sigmoid based VoteNet [10] pipeline.

The proposed n-sigmoid attention mechanism primarily operates within the feature extraction while positively affecting the voting stages. Essentially, it acts on the features extracted from point clouds data enabling the network to focus on relevant information and suppress irrelevant and noisy elements during voting process (Figure 5)

This, in turn, enables the model to generate more accurate and reliable predictions, improving both the localization and classification of 3D objects. The CSA layer effectively enhances the feature extraction process [30], improving the model's capacity to discern subtle details and patterns that are essential for robust 3D object detection.

 doi:10.20944/preprints202405.1941.v1

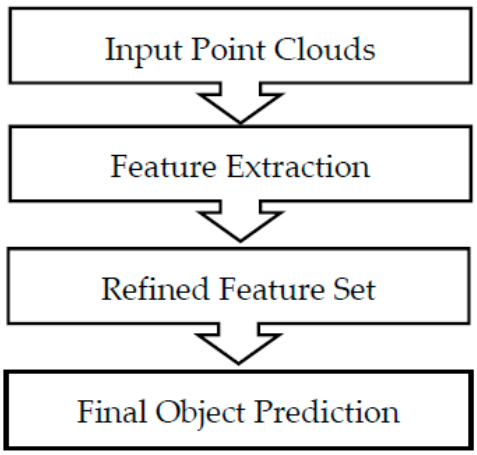

Figure 5. Simplified illustration of the 3D n-sigmoid CSA operation in the Hough Voting point cloud learning set.

This improvement contributes significantly to the overall performance and reliability of the VoteNet model, positioning it as a formidable solution for intricate 3D object analysis tasks in point clouds. By leveraging the enhanced channel and spatial attention mechanisms of the CSA layer, the modified VoteNet demonstrates improved discriminative capabilities, leading to superior object detection performance while the integration of the n_sigmoid activation function [22] augments the network's nonlinear processing allowing for a more refined emphasis on critical features while suppressing noise and irrelevant information.

In essence, the modified n-sigmoid CSA VoteNet architecture represents a significant advancement in 3D object detection methodologies, showcasing enhanced accuracy, robustness, and adaptability in challenging real-world scenarios.

4. Implementation Details

To validate the n-sigmoid CSA, the authors have used the VoteNet network for object detection with input of 3D point cloud of N points from indoor scenes. PointNet++ [12] is the backbone feature learning network. The entire n-sigmoid CSA based VoteNet is trained end-to-end from scratch with an Adam optimizer, batch size of 8 and an initial learning rate of 0.001. The learning rate is set to decrease by 10x after 80 epochs and to decrease by another 10x after 120 epochs.

Training the model to convergence on one Tesla T4 GPU takes around 30 hours on SUNRGBD when uninterrupted. Same as the original VoteNet, the proposed n-sigmoid Channel Spatial Attention VoteNet (n-sigmoid CSA-VoteNet) can collect 3D point clouds of a scene and generate proposals in one forward pass. The proposals are post-processed by a 3D NMS module with an IoU threshold of 0.25.

5. Experiments

In this section, the authors first compare the proposed n-sigmoid Channel and Spatial Attention (n-sigmoid CSA) with current state-of-the-art methods on several evaluation metrics that involve efficiency, accuracy…. After that, an ablation study is provided to understand the importance of the proposed attention mechanism on the process of voting point clouds and demonstrate the proposed method's advantages in its efficiency and accuracy in the model complexity study.

5.1. 3D n-Sigmoid CSA Module Integrated in the Point Cloud Learning Using Hough Voting

Dataset. SUNRGBD [15] is a single-view dataset for 3D scenes. It is made of 5,285 training and 5,050 testing RGB-D images in the dataset, where each object is indeed annotated with a bounding box and is part of 35 semantic classes. The authors only use the 3D coordinates as input and report on the overall metrics of mean average precision (mAP) and average recall (AR).

 doi:10.20944/preprints202405.1941.v1

Methods in comparison. The procedure to assert the validity of the proposed attention module consists of two steps. First it is to compare the n-sigmoid Channel and Spatial Attention in VoteNet (CSA-VoteNet) with a wide range of state-of-the-art methods (as shown in Table 2). Then, the authors compare the n-sigmoid CSA-VoteNet with other Attention models including the VoteNet itself.

All attention modules presented in Table 3A are slight modifications of the VoteNet official implementation where different 3D attention modules are added. Table 3B exhibits the results on the recall metric using SUNRGBD dataset. Even though CAA [31] and Point Transformer [32] had shown good performances assessed under either precision-related or recall-related metrics, the proposed n-sigmoid CSA-VoteNet achieves the best overall result **(69.72 mAP**) among all the attention modules. Most importantly, the advantages of the proposed n-sigmoid CSA-VoteNet stems from two significant facts: on the one hand, it assigns different relevance weights to different elements of the input data during the voting process; on the other hand, the attention map enables the network to emphasize important features and suppress irrelevant or noisy ones.

Table 2. 3D n-sigmoid CSA-VoteNet object detection on SUNRGBD compared with other state-of-the-art methods. (IoU threshold = 0.25).

Methods	Input	Bathtub	bed	Book shelf	chair	desk	dresser	night-stand	sofa	table	toilet	mAP
DSS [33]	GEO + RGB	44.2	78.8	11.9	61.2	20.5	6.4	15.4	53.5	50.3	78.9	42.1
COG [34]	GEO + RGB	58.3	63.7	31.8	62.2	45.2	15.5	27.4	51.0	51.3	70.1	47.6
2D-driven [35]	GEO + RGB	43.5	64.5	31.4	48.3	27.9	25.9	41.9	50.4	37.0	80.4	45.1
F-PointNet [36]	GEO + RGB	43.5	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	90.9	54.0
VoteNet [12]	GEO only	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.5
MLCVNet [37]	GEO only	79.2	83.0	31.9	75.8	26.5	31.3	61.5	66.3	50.4	89.1	59.8
DeMF [38]	GEO + RGB	79.5	87.0	44.1	80.7	33.8	**46.4**	66.3	**72.5**	52.8	92.7	65.6
CSA-VoteNet (ours)	GEO only	80.9	88.1	**49.6**	**83.8**	**49.7**	45.2	**72.8**	72.4	**59.8**	**94.9**	**69.72**

Table 3. Comparing n-sigmoid CSA-VoteNet with other attention mechanism baseline using VoteNet on SUNRGBD on both Average Precision and Recall.

A: The results of *Average Precision* on *SUNRGBD* [15] dataset of SA_VoteNet compared to other attention mechanisms. (IoU threshold = 0.25)

Methods	bed	table	sofa	chair	toilet	desk	dresser	night-stand	book-shelf	bath-tub	mAP
VoteNet [12]	83.3	49.8	64.1	74.1	89.3	23.8	26.4	60.7	30.9	72.8	57.7
A-SCN [39]	81.8	48.9	63.8	74.0	88.3	24.5	26.7	57.5	24.9	65.4	55.6
Point-attention [40]	84.4	49.0	61.9	73.8	87.4	25.7	24.6	56.0	28.2	73.1	56.4
CAA [31]	83.7	50.2	63.4	74.9	89.7	25.7	30.6	64.7	27.5	77.6	58.8
Point-transformer [32]	83.9	50.4	63.7	75.2	86.6	26.3	28.1	62.5	35.8	72.2	58.5
Offset-attention [41]	82.8	49.8	60.5	73.0	86.5	23.6	27.1	56.5	25.6	71.2	55.7
CSA-VoteNet (ours)	**88.1**	**59.8**	**72.4**	**83.8**	**94.9**	**49.7**	**45.2**	**72.8**	**49.6**	**80.9**	**69.72**

B: The results of *Recall* on *SUN RGB-D* [15] dataset of SA-VoteNet compared to other attention mechanisms. (IoU threshold = 0.25)

 doi:10.20944/preprints202405.1941.v1

Methods	bed	table	sofa	chair	toilet	desk	dresser	night-stand	book-shelf	bath-tub	AR
VoteNet [12]	95.2	85.5	89.5	86.7	97.4	78.8	81.0	87.8	68.6	90.4	86.1
A-SCN [39]	94.1	83.3	88.4	**87.3**	96.7	78.8	77.3	85.4	67.6	80.8	84.0
Point-attention [40]	94.8	83.6	88.9	86.3	95.4	78.7	78.2	88.2	62.5	86.5	84.3
CAA [31]	94.1	**84.7**	89.7	86.8	97.4	**79.3**	80.6	**89.8**	65.9	90.4	85.9
Point-transformer [32]	93.4	84.5	89.4	86.1	94.7	77.4	80.6	89.4	71.9	90.4	85.8
Offset-attention [41]	94.1	83.5	87.8	86.1	97.4	78.9	78.2	88.2	64.9	86.5	84.6
CSA-VoteNet (ours)	**95.5**	82.9	**91.6**	**87.3**	**97.6**	77.6	**83.2**	88.4	**73.2**	**91.0**	**86.8**

5.2. 3D n-Sigmoid CSA Module Integrated in the Point

This ablation study does provided an understanding of the importance of the proposed attention mechanism. Its use in the process of Hough voting on point clouds establish the proposed method's benefits, effectiveness and accuracy in the model complexity study. Table 4 shows the improvement obtained in the VoteNet pipeline when the n-shifted sigmoid CSA is added when compared to other state-of-the-art networks (both with or without attention).

Table 4. Comparing CSA Hough Voting with attention mechanism baseline (IoU threshold at 0.25 and 0.5) on SUNRGBD dataset.

Methods	mAP @ 0.25	mAP @ 0.5
Methods without attention		
H3DNet [43]	60.0	39.0
LGR-Net [44]	62.2	-
HGNet [45]	61.6	-
SPOT [46]	60.4	36.3
Feng [47]	59.2	-
MLCVNey [37]	59.2	-
VENet [46]	62.5	39.2
DeMF [38]	65.6	45.4
CAGroup3D [49]	66.8	50.2
TR3D+FF [50]	69.4	53.4
Point-GCC+TR3D+FF [52]	**69.7**	**54.0**
Methods with attention		
VoteNet [12]	57.7	41.3
ImVoteNet 40]	-	43.4
CSA-VoteNet (Ours)	**69.72**	**54.17**

 doi:10.20944/preprints202405.1941.v1

When comparing with the no-attention module in 3D object detection, the proposed network still emerges best since it outperforms all the existing methods notably with 69.72 mAP @ IoU 0.25 and 54.17 @ IoU 0.5.

Results summarized in Table 4 show that SA Hough Voting outperforms all previous methods (by at least 4.12 mAP on the DeMF [36]) using the SUNRGBD dataset. Also, a per-category evaluation for SUNRGBD is provided. In Table 3 (A and B), the proposed n-sigmoid CSA-VoteNet demonstrated superior results when compared to other attention mechanisms baseline using VoteNet on SUNRGBD dataset. Just to roundup the whole research, Table 4 establishes a comparison with any other no-attention mechanism in the 3D object detection domain. The proposed method tremendous improvements when only the geometric input (point clouds) is used.

Advantage of using both average and max pooling techniques in the n-shifted sigmoid CSA. The authors have also performed experiments to confirm the advantage of using both pooling methods as opposed to using only either the average or the max pooling operation in the proposed attention module. Table 5 presents the results obtained that show that accuracy is neatly improved when using both average and max pooling operations together while using max pooling operation yields better results than the average pooling results.

Table 5. Model results on using concatenated average pooling or either of the pooling operations.

Methods	mAP @ 0.25	mAP @ 0.5
VoteNet without attention	57.7	41.3
n-sigmoid CSA VoteNet with both avg and max pooling	**69.72**	**54.17**
n-sigmoid CSA VoteNet with max pooling only	69.11	53.67
n-sigmoid CSA VoteNet with avg pooling only	68.84	53.22

The training results obtained show that the improvement in the accuracy score, when using both the average and max pooling operation, is due to context-aware voting where the model adapts its voting strategy based on the learned context. In essence, in this proposed adaptive attention module, the model dynamically adjusts the importance of different votes based on the spatial and channel-wise setting.

Advantage of using the n-shifted sigmoid instead of the traditional sigmoid. The authors have also performed experiments to confirm the benefit of using the n-shifted sigmoid as opposed to using the traditional sigmoid as a gating mechanism in the proposed attention module. Table 6 presents the results obtained that show that accuracy is neatly improved when using the n-shifted sigmoid activation function. The authors believed that the improvement in accuracy is due to the n-shifted sigmoid activation function ability to improve feature discrimination and to enhance spatial and channel attention.

Table 6. Model results on using n-shifted sigmoid or traditional sigmoid.

Methods	mAP @ 0.25	mAP @ 0.5
VoteNet without attention	57.7	41.3
n-shifted sigmoid CSA VoteNet	**69.72**	**54.17**
traditional sigmoid CSA VoteNet	69.21	53.96
p-sigmoid CSA VoteNet	69.32	53.92

5.3. Discussion

In this experiment analysis, the advantages of using the attention mechanism are described in the Hough voting system.

The integration of an n-shifted sigmoid CSA mechanism within VoteNet does provide key benefits that enhance the network's performance and capabilities in various ways such as:

 doi:10.20944/preprints202405.1941.v1

An improved relevance weighting. By incorporating this attention mechanism, VoteNet does assign different relevance weights to different elements of the input data during the voting process. This allows the network to focus on critical features and downplay less relevant ones, leading to more accurate and precise decisions.

To verify that the n-shifted sigmoid CSA does improve relevance weighing of scene objects, the authors compare the performance of a model with and without the proposed n-shifted sigmoid CSA mechanism.

Figure 6 shows that accuracy significantly increases after the n-shifted sigmoid CSA is added to the VoteNet meaning that the relevance weighting is greatly enhanced. It is observable that the accuracy increase is very remarkable for some elements like the bookshelf, the desk. This could be due to specific factors in the dataset such as proximity, cluttering…The objects detected (bounding boxes) in the scene result from the voting strategy based on this relevance weighting which allows for salient objects to be voted for.

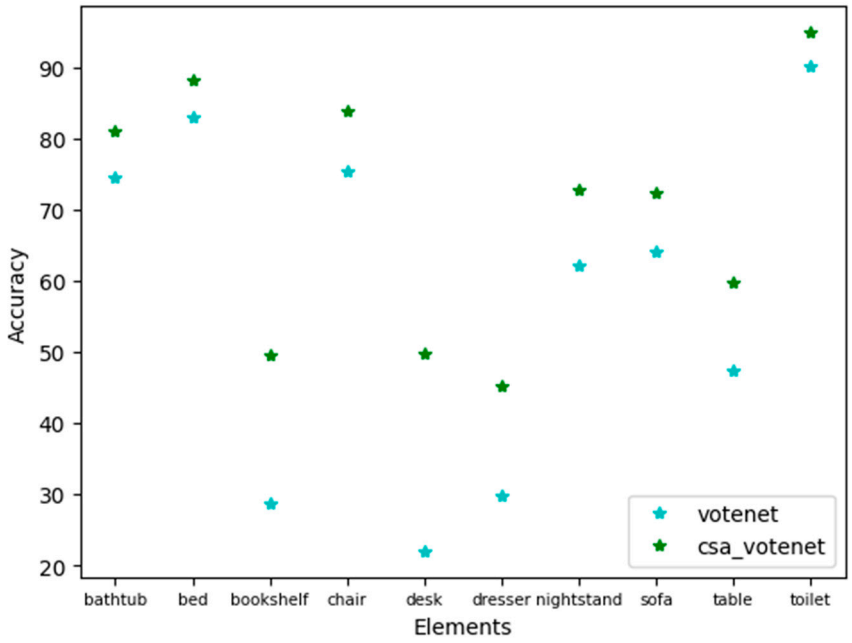

Figure 6. Accuracy results of the n-shifted sigmoid CSA VoteNet (with attention) and the VoteNet (without attention) on each element.

An enhanced feature representation. Attention mechanisms enable the network to emphasize important features and suppress irrelevant or noisy ones, facilitating the extraction of more informative and discriminative feature representations.

To verify that the n-shifted sigmoid CSA does enhance feature representation of scene objects, the authors compare the performance of first, the proposed n-shifted sigmoid CSA model with channel and spatial attention concatenation and secondly, the proposed n-shifted sigmoid CSA model with channel and spatial attention multiplication instead.

Here the channel and spatial attentions are multiplied elementwise to preserve the feature representations influenced by both the channel and spatial attention mechanisms. This provides a different form of modulation where multiplication is expected to emphasize regions where both channel and spatial attentions are high, potentially focusing more on salient features.

Figure 7 shows that accuracy significantly increases when multiplication is used instead of mere concatenation. Thereby demonstrating that feature representation of scene objects does emphasize important features and neglect others.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 May 2024 doi:10.20944/preprints202405.1941.v1

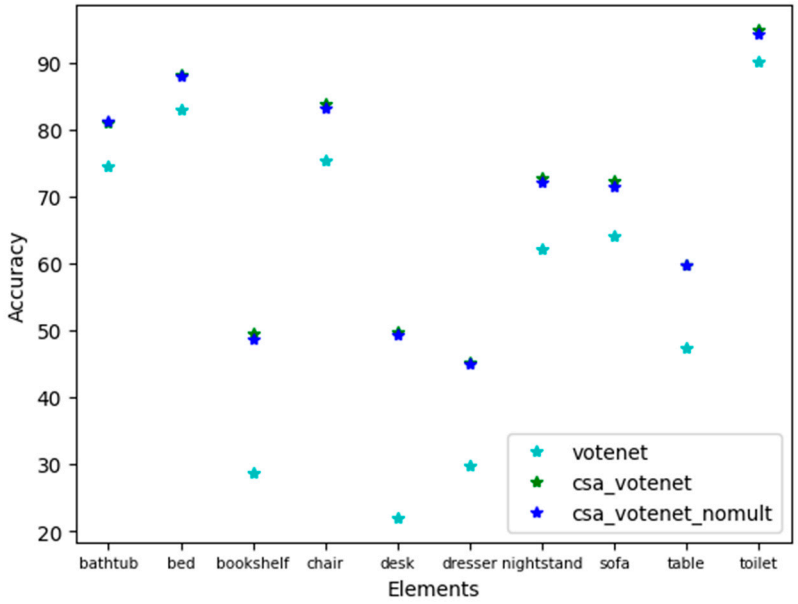

Figure 7. Accuracy results of the n-shifted sigmoid CSA VoteNet (with concatenation) and the n-shifted sigmoid CSA VoteNet (with multiplication) on each element. The mAP @ 0.25 accuracy is 69.29 for the n-shifted sigmoid CSA VoteNet (with concatenation).

In summary, all these benefits collectively lead to more accurate, robust, and context-aware 3D object detection and localization, making n-shifted sigmoid CSA-VoteNet more effective in handling diverse real-world scenarios.

5.4. Model Complexity

Table 7 presents a set of reference data (model size, inference time…) regarding the proposed model's complexity by comparing different object detection modules. For the SUNRGBD dataset, the model sizes of the network are also included to show its impact on complexity and speed especially when the Shuffle Attention has been used because of its ability to reduce the computational overhead. Although some networks such as F-PointNet [34] does achieve relatively higher performances regarding speed, it also requires more computational resources such as longer training time or larger memory consumption.

Table 7. Model size and processing time on SUNRGBD dataset.

Methods	Model size	Inference time (seconds/epoch)	Training time (seconds/epoch)
F-PointNet [34]	47.0 MB	**0.09**	-
3D-SIS [48]	19.7 MB	-	-
H3DNet [39]	-	-	**42.0**
VENet [44]	-	0.10	85
VoteNet [12]	**11.2 MB**	0.16	45.8
CSA-VoteNet (ours)	13.6 MB	0.25	123.2

Despite the model's light size increase, it is easy to notice that the model still performs acceptably when it comes to speed compared to the original VoteNet despite reducing its performance with a result of 123.2 seconds in training time.

 doi:10.20944/preprints202405.1941.v1

6. Conclusions

In this paper, the authors proposed a novel and effective n-shifted sigmoid Channel and Spatial Attention module that not only reduces computational overhead but also enhance the 3D scene relevant features selection of 3D convolutional neural networks. Specifically, it improves the seed points feature representation to effectively predict bounding box parameters directly from 3D scenes and detect objects more accurately.

The new attention mechanism is placed just before the voting module to improve the accuracy score since it improves the discriminative features, provides more context-aware decisions before the voting process, focuses on adaptive attention to dynamically adjusts the importance of different votes based on the spatial and channel-wise context.

The proposed method achieved state-of-the-art detection accuracy on the SUNRGBD dataset with only geometric information given, demonstrating the effectiveness of the proposed approach in the Deep Hough voting network. Experimental results have shown that the proposed n-shifted sigmoid Channel and Spatial Attention is an extremely light plug-and-play module, that is able to significantly improve the performance of numerous deep CNN architectures.

For future research, the focus will be to implement a 3D instance segmentation using the n-shifted sigmoid CSA-VoteNet by using methods like non maximum clustering of the points clouds inside the bounding boxes for instance.

Author Contributions: Conceptualization, D. Burume; methodology, D. Burume; software, D. Burume and S. Du.; validation, S. Du., and D. Burume.; formal analysis, Q. Liu; investigation, D. Burume.; resources, D. Burume; data curation, D. Burume.; writing—original draft preparation, D. Burume; writing—review and editing, S. Du and Q. Liu; supervision, S. Du and Q. Liu; project administration, S. Du; funding acquisition, S. Du and Q. Liu All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This research was funded by the NATIONAL RESEARCH FOUNDATION OF SOUTH AFRICA (Grant Numbers SRUG2203291049 and 145975), KUNMING UNIVERSITY FOUNDATION (No. YJL2205), and the FOUNDATION OF YUNNAN PROVINCE SCIENCE AND TECHNOLOGY DEPARTMENT (No. 202305AO350007)

Data Availability Statement: The dataset used for this study are openly available in at http: https://rgbd.cs.princeton.edu.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yan, J; Zhu, R.; Chen, B.; Xu, H. and Zhu, X. Channel and Spatial Attention Fusion Module for Detection. *2023*, https://doi.org/10.21203/rs.3.rs-2804607/v1.
2. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang and H. Lu. Dual attention network for scene segmentation. 2019. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, pp. 3146 –3154.
3. Wang, J., Mo, W., Wu, Y., Xu, X., Li, Y., Ye, J. and Lai, X. 2022. Combined Channel Attention and Spatial Attention Module Network for Chinese Herbal Slices Automated Recognition; Publisher: Sec. Brain Imaging Methods, volume 16, https://doi.org/10.3389/fnins.2022.920820.
4. Liu, M; Fang, W; Ma, X; Xu, W; Xiong, N. and Ding, Y. 2021. Channel Pruning Guided by Spatial and Channel Attention for DNNs in Intelligent Edge Computing. arXiv:2011.03891v2.
5. Zhu, Z; Xu, M; Bai, S; Huang, T. and Bai, X. 2019. Asymmetric non-local neural networks for semantic segmentation. CoRR. Vol. abs/1908.07678.
6. Zhu, Y; Liang, Y; Tang, K; and Ouchi, K. 2022. SC-NET: Spatial and Channel Attention Mechanism for Enhancement in Face Recognition. *5th International Conference on Information and Computer Technologies (ICICT)*, New York, NY, USA, pp. 166-172, doi: 10.1109/ICICT55905.2022.00036.
7. Wang, Q; Wu, B; Zhu, P; Li, P; Zuo, W. and Hu, Q. 2020. Eca-net: Efficient channel attention for deep convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2020, Seattle, WA, USA, pp. 11531–11539, IEEE.

8. Li, X; Hu, X; and Yang, J. 2019. Spatial group wise enhance: Improving semantic feature learning in convolutional networks. CoRR. Vol.abs/1905.09646.
9. Ma, N; Zhang, X; Zheng, H; and Sun, J. 2018. Shufflenet V2: practical guidelines for efficient CNN architecture design. In *Computer Vision- ECCV 2018- 15th European Conference,* Munich, Germany, September 8-14, 2018, Proceedings, Part XIV, 2018, pp. 122–138.
10. Zhang, X; Zhou, X; Lin, M; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. 2018, pp. 6848–6856, IEEE Computer Society.
11. Zhang, Q; Yang, Y. 2021. SA-Net: Shuffle Attention For Deep Convolutional Neural Networks. arXiv:2102.00240v1 [cs.CV].
12. Qi, C. R.; Litany, O.; He K. and Guibas. L. 2019. Deep Hough Voting for 3D Object Detection in Point Clouds. In *IEEE/CVF International Conference on Computer Vision (ICCV),* Seoul, Korea (South), pp. 9276-9285, doi: 10.1109/ICCV.2019.00937.
13. A. Fernández, J. Umpiérrez, and J. R. Alonso. 2023. Generalized Hough transform for 3D object recognition and visualization in integral imaging. In *J. Opt. Soc. Am.* A 40, C37-C45.
14. Qi, C. R.; Yi, L.; Su, H. and Guibas. L. J. 2017. Pointnet++: Deep hierarchical feature learning on pointsets in a metric space. arXivpreprintarXiv:1706.02413.
15. Song, S; Lichtenberg, S.P.; and Xiao, J. 2015. Sunrgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576.
16. Chappa, R. T. N. V. S.; and El-Sharkawy, M. 2020. Squeeze-and-Excitation SqueezeNext: An Efficient DNN for Hardware Deployment. In *10th Annual Computing and Communication Workshop and Conference (CCWC)*, 0691–0697. https://doi.org/10.1109/CCWC47524.2020.9031119.
17. Zhu, H.; Xie, C.; Fei, Y.; Tao, H. 2021. Attention Mechanisms in CNN-Based Single Image Super-Resolution: A Brief Review and a New Perspective. *Electronics*, *10*, 1187. https://doi.org/10.3390/electronics10101187.
18. Woo, S; Park, J; Lee, J; and Kweon, I.S. 2018. CBAM: convolutional block attention module. In *Computer Vision-ECCV 2018 -15th European Conference,* Munich, Germany, Proceedings, Part VII, pp. 3–19.
19. Vaswani, A.; Shazeer, N; Parmar, N; Uszkoreit, J; Jones, L.; Gomez, V; Kaiser, L; Polosukhin, L. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL], https://doi.org/10.48550/arXiv.1706.03762.
20. Xie, T.; Ding, W.; Zhang, J.; Wan, X.; Wang, J. 2023. Bi-LS-AttM: A Bidirectional LSTM and Attention Mechanism Model for Improving Image Captioning. *Appl. Sci.*, *13*, 7916. https://doi.org/10.3390/app13137916.
21. Zhao, Z.; Feng, F.; Tingting, H. 2022. FNNS: An Effective Feedforward Neural Network Scheme with Random Weights for Processing Large-Scale Datasets. Appl. Sci. 2022, 12, 12478.
22. Mulindwa, D.B., Du, S. 2023. "An n-Sigmoid Activation Function to Improve the Squeeze-and-Excitation for 2D and 3D Deep Networks. In Electronics, 12, 911. https:// doi.org/10.3390/electronics12040911.
23. Rane, C; Tyagi, K; and Manry, M. 2023. Optimizing Performance of feedforward and convolutional neural networks through dynamic activation functions. arXiv:2308.05724v1 [cs.LG]
24. Zhang, R.; Wu, Y.; Jin, W.; Meng, X. 2023. "Deep-Learning-Based Point Cloud Semantic Segmentation: A Survey". *Electronics*, , 3642. https://doi.org/10.3390/electronics12173642.
25. Zhang, C.; Xu, F.; Wu, C. and Xu, C. 2022. "A lightweight multi-dimension dynamic convolutional network for real-time semantic segmentation", Front Neurorobot. 16:1075520. Doi: 10. 3389/fnbot.2022.1075520
26. Wu, Y.; and He, K. 2018. Group normalization. In *ComputerVision-ECCV2018-15th European Conference*, Munich, Germany, Proceedings, Part XIII, pp. 3–19.
27. Liu, D.; Han, G.; Liu, P.; Yang, H.; Chen, D.; Li, Q.; Wu, J.; Wang, Y. 2022. A Discriminative Spectral-Spatial-Semantic Feature Network Based on Shuffle and Frequency Attention Mechanisms for Hyperspectral Image Classification. *Remote Sens.*, *14*, 2678. https://doi.org/10.3390/rs14112678.
28. Guo, M; Xu, T.; Liu, J. et al. 2022. Attention mechanisms in computer vision: A survey". Computational Visual Media, https://doi.org/10.1007/s41095-022-0271-y, Vol. 8, No. 3, 331–368.
29. Tliba, M.; Chetouani, A.; Valenzise, G. and F. Dufaux. 2023. Quality Evaluation of Point Clouds: A Novel No-Reference Approach Using Transformer-Based Architecture. arXiv:2303.08634v1 [cs.CV]
30. Kong, J.; Wang, H.; Yang, C.; Jin, X.; Zuo, M.; Zhang, X. 2022. A Spatial Feature-Enhanced Attention Neural Network with High-Order Pooling Representation for Application in Pest and Disease Recognition. *Agriculture*, *12*, 500. https://doi.org/10.3390/agriculture12040500.

 doi:10.20944/preprints202405.1941.v1

31. Qiu, S.; Anwar, S.; and Barnes, N. 2021. Geometric back projection network for point cloud classification. IEEE Transactions on Multimedia.
32. Zhao, H. ; Jiang, L.; Jia, J.; Torr, P. and Koltun, V. 2020. Point transformer. arXiv preprint arXiv:2012.09164.
33. Songand, S. and Xiao, J. 2016. Deep sliding shapes for amodal 3dobject detection in rgb-d images. In Proceedings of the IEEE *Conference on Computer Vision and Pattern Recognition*, pages 808-816.
34. Ren, Z.; and Sudderth, E. B. 2016. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1525–1533.
35. Lahoud, J.; and Ghanem, B. 2017. 2d-driven 3d object detection in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4622–4630.
36. Qi, C. R.; Liu, W.; Wu, C.; Su, H. and Guibas. L. J. 2018. Frustum pointnets for 3d object detection from rgbd data. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927.
37. Wang, Z.; Xie, Q.; Wei, M.; Long, K. and Wang, J. 2022. Multi-feature Fusion VoteNet for 3D Object Detection. ACM Trans. Multimedia Comput. Commun. Appl. 18, 1, Article 6, 17 pages.
38. Yang, H.; Shi, C.; Chen, Y.; Wang, L. 2022. Boosting 3D Object Detection via Object-Focused Image Fusion. arXiv.2207.10589[cs.CV].
39. Xie, S.; Liu, S.; Chen, Z.; and Tu, Z. 2018. Attentional shape contextnet for point cloud recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages 4606–4615.
40. Feng, M.; Zhang, L.; Lin, X.; Gilani, S. Z.; and Mian, A.. 2020. Point attention network for semantic segmentation of 3d point clouds. Pattern Recognition, 107:107446.
41. Guo, M.; Cai, J.; Liu, Z.; Mu, T.; Martin, R. R.; and Hu, S. 2021. Pct: Point cloud transformer. Computational Visual Media, 7: pages 187–199.
42. Qi, C. R.; Chen, X.; Litany, O; Guibas. L. J. 2020. ImVoteNet: Boosting 3D Object Detection in Point Clouds with Image Votes. arXiv:2001.10692 [cs.CV].
43. Zhang, Z.; Sun, B.; Yang, H.; and Huang, Q. 2020. H3dnet: 3dobject detection using hybrid geometric primitives. In *IEEE/CVF European Conference on Computer Vision (ECCV),* pp. 311-329.
44. Li, J.; and Feng, J. 2020. Local grid rendering networks for 3D object detection in point clouds. arXiv preprint arXiv:2007.02099.
45. Chen, J.; Lei, B.; Song, Q.; Ying, H.; Chen, D. Z.; and Wu, J. 2020. A hierarchical graph network for 3D object detection on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 392-401.
46. Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. ScanNet: Richly annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839.
47. Du, H.; Li, L.; Liu, B.; and Vasconcelos, N. 2020. SPOT: Selective point cloud voting for better proposal in point cloud object detection. In ECCV: 16th European Conference, Glasgow, UK, Proceedings, Part XI, Pages 230–247, https://doi.org/10.1007/978-3-030-58621-8_14.
48. Xie, Q.; Lai, Y.; Wu, J.; Wang, Z.; Lu, D.; Wei, M.; and Wang, J. 2021. VENet: Voting Enhancement Network for 3D Object Detection. ICCV 2021.
49. Wang, H.; Ding, L.; Dong, S.; Shi, S.; Li, A.; Li, J.; Li, Z.; Wang, L. 2022. CAGroup3D: Class-Aware Grouping for 3D Object Detection on Point Clouds. arXiv:2210.04264 [cs.CV].
50. Rukhovich, D.; Vorontsova, A.; Konushin. A. 2022. TR3D: Towards Real-Time Indoor 3D Object Detection. DOI: 10.1109/CVPR52688.2022.00118. *Conference: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*
51. Fan, G.; Qi, Z.; Shi, W.; and Ma, K. 2023. Point-GCC: Universal Self-supervised 3D Scene Pre-training via Geometry-Color Contrast. arXiv:2305.19623v2 [cs.CV].
52. Hou, J.; Dai, A.; Nießner, M. 2019. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. arXiv:1812.07003 [cs.CV].