Enhanced 3D Point Cloud Object Detection with Iterative Sampling and Clustering Algorithms

Shane Ward and Hossein Malekmohamadi
Institute of Artificial Intelligence, De Montfort University, The Gateway, Leicester, U.K.

Keywords: mAP– Mean Average Precision, RANSAC – Random Sampling and Consensus, DBSCAN – Density-based Spatial Clustering of Applications with Noise, BIRCH – Balanced Iterative Reducing and Clustering using Hierarchies, OPTICS – Ordering Points to Identify the Clustering Structure, MLVCNet – Multi-level Context VoteNet.

Abstract: Existing state-of-the-art object detection networks for 3D point clouds provide bounding box results directly from 3D data, without reliance on 2D detection methods. While state-of-the-art accuracy and mAP (mean-average precision) results are achieved by GroupFree3D, MLCVNet and VoteNet methods for the SUN RGB-D and ScanNet V2 datasets, challenges remain in translating these methods across multiple datasets for a variety of applications. These challenges arise due to the irregularity, sparsity and noise present in point clouds which hinder object detection networks from extracting accurate features and bounding box results. In this paper, we extend existing state-of-the-art 3D point cloud object detection methods to include filtering of outlier data via iterative sampling and accentuate feature learning via clustering algorithms. Specifically, the use of RANSAC allows for the removal of outlier points from the dataset scenes and the integration of DBSCAN, K-means, BIRCH and OPTICS clustering algorithms allows the detection networks to optimise the extraction of object features. We demonstrate a mean average precision improvement for some classes of the SUN RGB-D validation dataset through the use of iterative sampling against current state-of-the-art methods while demonstrating a consistent object accuracy of above 99.1%. The results of this paper demonstrate that combining iterative sampling with current state-of-the-art 3D point cloud object detection methods can improve accuracy and performance while reducing the computational size.

1 INTRODUCTION

For common point cloud object detection applications involving scene understanding, the accuracy and performance of the method relies heavily on pre-processing of the input data prior to training the object detection neural network. In complex real-world applications, the scene and objects to be inspected are susceptible to large amounts of outlier points and noise which results in reduced accuracy and performance. This also results in suboptimal use of computational power on input data points which provide misleading information of the objects in the scene. Recent works related to neural networks for 3D object detection, specifically using point cloud input, have yielded promising results for various applications. It has also been demonstrated that the use of purely geometric data with existing state-of-the-art neural networks such as VoteNet (Qi et. al, 2019), MLCVNet (Xie et. al, 2020) and GroupFree3D (Liu et. al, 2021) can produce superior results compared to methods which utilize 2D detectors and approximate 3D bounding box proposals based on 3D region networks. Methods heavily influenced by 2D detectors become computationally expensive for deducing 3D proposals for complex scene understanding and applications where speed is critical.

The PointNet (Qi et. al, 2017) architecture was the catalyst for the development of this new set of deep learning methods with the objective of directly processing point cloud data to tackle classification, segmentation, and object detection tasks. Prior to this work, most 3D object detection methods performed operations on 2D and 2.5D data to infer or project detection algorithms onto 3D space such as Shape-based 3D matching or by transforming the 3D point cloud data from irregular point clouds to regular 3D voxel grids with methods based on VoxelNet. The PointNet architecture was improved in terms of capturing local structures in metric space, addressed by PointNet++.

Ward, S. and Malekmohamadi, H.
Enhanced 3D Point Cloud Object Detection with Iterative Sampling and Clustering Algorithms.
DOI: 10.5220/0010910600003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 4: VISAPP*, pages 674-681
ISBN: 978-989-758-555-5; ISSN: 2184-4321

The PointNet++ architecture is a direct extension of PointNet using additional sampling and grouping in conjunction with PointNet. This improved the earlier method by using a hierarchical network utilizing sampling and grouping layers which in turn improved the model's ability to classify and segment in metric space. The current state-of-the-art works for 3D point cloud object detection all utilize a PointNet++ backbone with additional network architectures for each such as deep hough voting (Qi et. al, 2019), multi-level context attention (Xie et. al, 2020) and transformer-based attention (Liu et. al, 2021).

In this paper, we build on the existing state-of-the-art 3D point cloud object detection methods by demonstrating the importance of iterative sampling and clustering algorithms to achieve both fast and accurate 3D bounding box proposals. We propose enhanced versions of the current state-of-the-art methods by integrating a RANSAC iterative sampling method and combining this with multiple clustering algorithms to serve a wide variety of applications (DBSCAN, K-Means, BIRCH, OPTICS). The iterative sampling method provides a customisable filter for the raw input point cloud data to separate outliers and the various clustering algorithms allow for the early extraction of features prior to neural network training. For fair comparison we run our enhanced VoteNet, MLCVNet and GroupFree3D methods on two common benchmark indoor 3D datasets, SUN-RGBD and ScanNet. The objective of this work is to present the following contributions:

1. Propose a novel iterative sampling and clustering framework for 3D point cloud object detection and can be applied to a wide variety of applications. We demonstrate increased efficiency, accuracy and speed through our pre-processing framework.

2. Enhanced VoteNet, MLCVNet and GroupFree3D methods achieving state-of-the-art results through:

- Integration of customisable iterative sampling method for the filtering of outlier points.
- Integration and comparison of four customisable clustering methods to allow for early feature extraction in training phase.

3. Considerations for deployment of state-of-the-art 3D object detection methods in real-world applications where efficiency, accuracy and speed are paramount.

2 BACKGROUND

In recent times, there have been many contributions to the state-of-the-art methods for 3D object detection on various input data. In this section, we review the methods most relevant to this work and specifically for methods with point cloud input data.

PointNet. The PointNet architecture as previously stated was a large breakthrough in the direct processing of raw point cloud data to achieve results without the use of 2D detectors. There are advantages to this method such as the processing time and ability to process low numbers of data points but disadvantages such as poor accuracy and the disconnect between the data representation and the actual world scene, make the method unusable for 3D point cloud object detection in applications where dense scene understanding is a requirement. The PointNet architecture provided an end-to-end network for the classification, part segmentation and semantic segmentation of raw point cloud data. The method which uses sampling of point sets, is an alternative to 3D voxelization which approximates errors for applications where high accuracy is required. This work demonstrated that with a basic architecture reasonable results are achieved. For testing robustness, it was shown that with 50% of points missing from an input set via random sampling, the accuracy only dropped by 2.4% and 3.8%. Also, the method demonstrated robustness to outlier points, achieving greater than 80% accuracy even when 20% of points are outliers. PointNet was the first of its kind in demonstrating computational cost efficiency which is an important factor in industrial applications. PointNet is capable of processing greater than 1M points/second with 1080X GPU showing great potential for real-time applications but the method did not capture local structures in metric space.

PointNet++. The shortcomings of the PointNet architecture in terms of capturing local structures in metric space were quickly addressed with PointNet++. The architecture is a direct extension of PointNet using additional sampling and grouping in conjunction with PointNet. This improved the earlier method by using a hierarchical network utilizing sampling and grouping layers which in turn improved the models ability to classify and segment in metric space. The performance of the PointNet++ method on the ModelNet40 dataset outperformed Subvolume (voxel method), MVCNN (image method) and the earlier PointNet method (Point clouds) with an accuracy of 91.9%. The paper acknowledged that further work in improving inference speed (especially for MSG and MRG layers) was a future option. It is also noted that CNN based methods do not apply to unordered point sets (point cloud data) and that the method can scale well.

VoteNet. Perhaps the biggest breakthrough related to this work was the introduction of the Deep Hough

Voting network for object detection, also known as VoteNet. The method of this paper, utilizes a PointNet++ backbone for feature learning and couples this with Deep Hough voting in order to sample, group and propose classification. The VoteNet method utilizes 3D bounding boxes and depends solely on geometric information. As previously stated, VoteNet does not make use of RGB or Depth images similar to other methods which supports the theory that state-of-the-art object detection methods may be developed from the processing of raw point clouds i.e., this is an end-to-end method. In summary, the VoteNet method learns to vote to object centroids directly from raw point clouds and aggregates votes through their features and local geometry to generate high-quality detection proposals using only point cloud input, outperforming other methods where depth and colour images are also used.

MLCVNet. The objective of the MLCVNet (multi-level context VoteNet) method is to recognize 3D objects correlatively, building on the state-of-the-art VoteNet. This method utilizes a self-attention mechanism and multi-scale feature fusion to model the multi-level contextual information and propose three sub-modules. The testing performed by the authors of this paper proves that the contextual sub-modules improve the accuracy and performance of 3D object detection. The results of the MLCVNet architecture described in the MLCVNet paper can be described as state-of-the-art. On the ScanNet v2 dataset, the MLCVNet method outperformed VoteNet and 3DSIS methods for all categories of the dataset in terms of mAP. Also, the qualitative results of 3D object detection on the SUN-RGBD dataset demonstrate state-of-the-art results. The ground truth bounding boxes were compared to the results of mainly the VoteNet and MLCVNet networks.

GroupFree3D. At the time of undertaking this work, the most recent state-of-the-art neural network method for performing object detection on point cloud data is the GroupFree3D method. The method computes the feature of an object from all points in the scene point cloud through the help of an attention mechanism where the contribution of each point is automatically learned during the training phase. GroupFree3D proposes an attention mechanism which utilises a Transformer decoder allowing for all points in the input point cloud to be used during training. Implemented on the benchmark SUN RGB-D and ScanNet v2 datasets, the method obtained state-of-the-art mAP results of 69.1 @ 0.25 and 52.8 @ 0.50. The authors for this work also executed ablation studies on sampling strategy which demonstrated improvements on the initial results. The objective of this work is to build on recent state-of-the-art developments through implementation and evaluation of enhanced versions of the identified current state-of-the-art end to end 3D object detection methods on a benchmark 3D point cloud dataset.

3 METHODOLOGY

We present a framework for performing iterative sampling and clustering of point cloud data for 3D object detection methods. The desired outcome of combining iterative sampling and clustering methods results is to reduce the number of points in the input point cloud. As a result of iterative sampling, the input point cloud will have outlier points filtered which improves the neural network's ability to accurately detect objects in a dense or noisy scene. Adding clustering methods in combination with sampling will allow for early extraction of key features and the identification of point clusters, the building blocks of each object present in a scene. We recognise that a wide variety of applications may be served by such a framework for 3D point cloud object detection and we therefore include several clustering algorithm options in the framework to cater for this.

3.1 Iterative Sampling

Iterative sampling algorithms have existed for decades and have proven to be powerful tools in the pre-processing and filtering of input data prior to neural network training. Perhaps the most common and effective iterative sampling method is the Random sample consensus (RANSAC) which estimates parameters of a mathematical model from a set of observed data that contains outliers. A basic assumption is that the data consists of inlier data points whose distribution can be described by a model, and outliers which are data points which do not fit the model. These outlier points in point cloud dada, can result in incorrect detection approximations about the interpretation of the point set.

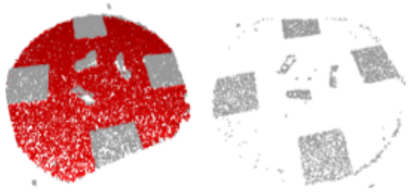


Figure 1: Segmentation of inlier and outlier points using RANSAC method on industrial MVTec ITODD dataset.

This outlier detection method applies to a wide range of data science applications, but in this context applies to dense point clouds for real-world applications. The removal of points in input point clouds which provide no contextual information of objects to be detected will reduce bounding box detection inaccuracies and result in increased computational efficiency. The issue of computational size in training neural networks with point cloud data remains one of the most prevalent and RANSAC allows for significant reductions in non-contextual points in the input data. The most relevant purpose of of the RANSAC method is to provide a robust method for the segmentation and removal of planes from point cloud scenes which is important in many applications where base planes are present with the objects to be inspected on top of the base plane.

3.2 Clustering

Similar to iterative sampling methods, clustering algorithms allow data points to be grouped into clusters in an unsupervised manner. However clustering methods allow for further subdivision of point sets into several groups as opposed to just inlier and outlier groups with the RANSAC method. Multiple clustering algorithms exist and are widely used in data science applications. Relevant to this work on clustering point cloud data for 3D object detection, we have included four of the most common algorithms as options within our proposed framework for training neural networks.

3.2.1 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a density-based clustering algorithm. DBSCAN is one of the most common clustering algorithms and most cited in scientific literature, hence our selection for our proposed framework. Given a set of points in space, DBSCAN groups together points that are closely packed together i.e. points with multiple nearest neighbors. DBSCAN marks points that lie alone in low-density regions as outliers i.e., points whose nearest neighbors are too far away.

3.2.2 K-means

K-means clustering is another popular unsupervised clustering algorithm, which aims to group a number of observations n into a target number of clusters k. For each observation, it belongs to the cluster with the nearest mean (cluster centroid), which serves as a prototype of the cluster. The overall effect of minimizing the averages of the squared distances between the data points in the same point set. The pseudo-code for the K-means clustering algorithm is described in Fig. 4 below.

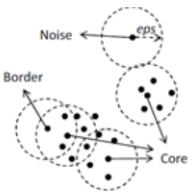


Figure 2: DBSCAN clustering algorithm diagram.

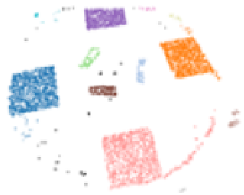
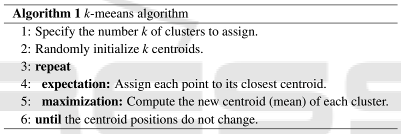


Figure 3: DBSCAN clusters example using industrial MVTec ITODD dataset.



```
Algorithm 1 k-meeans algorithm
Specify the number k of clusters to assign.
Randomly initialize k centroids.
repeat
    expectation: Assign each point to its closest centroid.
    maximization: Compute the new centroid (mean) of each cluster.
until the centroid positions do not change.
```

Figure 4: Pseudo-code for K-means clustering algorithm.

The K-means clustering method provides a useful alternative to DBSCAN which focuses on the centroid centres of clusters. This method aligns with the overall desired outcome of 3D point cloud object detection and pointwise networks due to the use of centroid centres. The size of the clusters must be set and for this, the mean average size of each object class is used.

3.2.3 BIRCH

Balanced iterative reducing and clustering using hierarchies (BIRCH) is another commonly used unsupervised data science algorithm used to perform hierarchical clustering over particularly large datasets. The main advantage of the BIRCH clustering algorithm is its ability to incrementally cluster multi-dimensional metric data points in a given point set to produce the best quality clustering for a given set of memory and time resource constraints. As a result of the efficiency of the BIRCH clustering algorithm we implement BIRCH as another option in the proposed framework. BIRCH has been successfully implemented in several related works for the clustering of multi-dimensional point sets.

3.2.4 OPTICS

Ordering points to identify the clustering structure (OPTICS) is the final commonly used clustering algorithm implemented in the framework for enhancing 3D point cloud object detection. The OPTICS clustering algorithm is also used for finding density-based clusters in spatial data. The principle of OPTICS is similar to DBSCAN, however it addresses the main DBSCAN weakness: the detection of meaningful clusters in data of varying density. In order to achieve this, each point in the point set is ordered such that the spatially closest points are neighbors in the ordered structure. A special distance is also stored for each point that represents the density that must be accepted for a cluster so that both points belong to the same cluster.

4 RESULTS AND DISCUSSION

In order to evaluate the performance of our iterative sampling and clustering framework, we first integrate it to the current state-of-the-art VoteNet, MLCVNet and GroupFree3D 3D point cloud object detection methods. We demonstrate the ability of the iterative sampling to separate outlier points and reduce the size of the input point cloud while all relevant data points remain. We also demonstrate and compare the ability of each clustering algorithm to enhance feature extraction in each of te state-of-the-art methods using the benchmark SUN RGB-D and ScanNet V2 datasets with PointNet++ backbone for fair comparison. All experiments for the purposes of this paper were run utilizing the same setup for a fair comparison also. The workstation consists of an Intel i9-10900 processor (2.8GHz) and Nvidia GeForce RTX 2060 GPU. The workstation is running Ubuntu 20.04 and we use a python 3.7 anaconda environment to install all required packages, including PyTorch 1.1 and Cuda 10.1.

4.1 Evaluation of Iterative Sampling

Enhanced VoteNet. For the implementation of this method, we follow the provided instructions of the VoteNet paper. This includes the use of a PointNet++ backbone with 4 set abstraction layers and 2 feature propagation layers for a fair comparison and the use of the common benchmark SUN RGB-D training and validation datasets. The integration of the iterative sampling framework method includes the modifying the SUN RGB-D detection dataset class to include the option to run VoteNet with RANSAC iterative sampling. To achieve the results described in Tables 1 and 2 below, we use 20,000 points as the input for each point cloud scene. We run 400 epochs with a batch size of 8 and a learning rate of 0.001. A key point to note is that for training we use only geometric information and no image data for a fair comparison against methods utilising image data.

Table 1: Mean average precision mAP @ 0.25 Enhanced VoteNet comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 1.

Method	bath	bed	bookshelf	chair	desk
VoteNet	74.4	**93.0**	28.8	75.3	22.0
MLVCNet	79.2	85.8	31.9	75.8	26.5
GroupFree3D	80.0	87.8	**32.5**	**79.4**	32.6
Ours RANSAC	**80.4**	87.4	30.2	63.2	**96.0**
Ours DBSCAN	61.8	65.4	21.3	48.6	63.5
Ours BIRCH	66.1	69.3	20.7	47.4	62.6
Ours KMeans	75.4	73.7	23.2	51.3	66.8
Ours OPTICS	64.3	66.5	21.3	44.7	59.3

Table 2: Mean average precision mAP @ 0.25 Enhanced VoteNet comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 2.

Method	dresser	nightstand	sofa	table	toilet
VoteNet	29.8	62.2	64.0	47.3	90.1
MLVCNet	31.3	61.5	66.3	50.4	89.1
GroupFree3D	**36.0**	**66.7**	**70.0**	**53.8**	**91.1**
Ours RANSAC	24.1	63.6	66.4	45.2	69.1
Ours DBSCAN	17.9	36.6	31.9	36.5	80.3
Ours BIRCH	18.1	42.2	32.5	39.9	84.4
Ours KMeans	21.3	49.4	34.3	43.5	**91.1**
Ours OPTICS	18.8	43.4	31.0	40.3	79.6

Table 3: Mean average precision mAP @ 0.5 Enhanced VoteNet comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 1.

Method	bath	bed	bookshelf	chair	desk
VoteNet	45.4	53.4	6.8	56.5	5.9
GroupFree3D	**64.0**	**67.1**	12.4	**62.6**	14.5
Ours RANSAC	37.7	18.6	**14.2**	37.4	**56.4**
Ours DBSCAN	42.1	19.7	6.9	19.3	5.8
Ours BIRCH	50.3	17.2	4.1	20.2	7.5
Ours KMeans	57.1	35.4	7.0	23.3	11.4
Ours OPTICS	52.2	17.7	6.3	18.2	6.9

Table 4: Mean average precision mAP @ 0.5 Enhanced VoteNet comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 2.

Method	dresser	nightstand	sofa	table	toilet
VoteNet	12.0	38.6	49.1	21.3	68.5
GroupFree3D	**21.9**	**49.8**	**58.2**	**29.2**	**72.2**
Ours RANSAC	13.6	34.7	14.4	16.0	61.3
Ours DBSCAN	6.4	17.0	7.2	23.5	63.6
Ours BIRCH	5.6	19.6	7.8	28.4	52.5
Ours KMeans	7.3	23.8	9.1	22.6	64.7
Ours OPTICS	6.1	21.0	8.1	28.9	50.7

The integration of the RANSAC iterative sampling method to remove outlier points yielded promising results across the bath and desk class at mAP @ 0.25 and bookshelf and desk at mAP @ 0.5 improving on the current state-of-the-art GroupFree3D methods shown in Tables 1 and 2 above, however the result proved inconsistent across all classes.

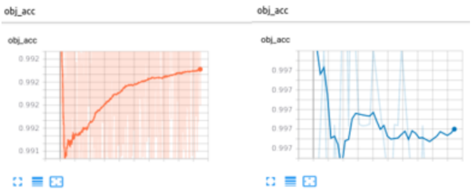


Figure 5: Average Object Accuracy of 99.27% and 99.84% on SUN RGB-D training and validation datasets.

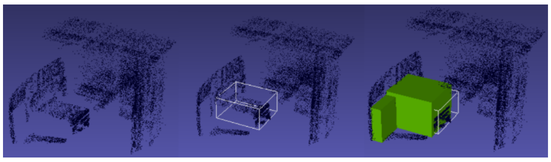


Figure 6: Input point cloud scene (20k points, Enhanced VoteNet bounding box prediction, Ground truth comparison vs prediction.

Enhanced MLCVNet. For the implementation of this method, we follow the provided instructions of the MLCVNet paper. This includes the use of a PointNet++ backbone with 4 set abstraction layers, 2 feature propagation layers and three sub-modules (patch-patch context, object-object context and global-scene context) to support a multi-level context attention mechanism for a fair comparison and the use of the common benchmark SUN RGB-D training and validation datasets. The integration of the iterative sampling framework method includes the modifying the SUN RGB-D detection dataset class to include the option to run MLCVNet with RANSAC iterative sampling. To achieve the results described in Tables 3 and 4 below, we again use 20,000 points as the input for each point cloud scene. We run 400 epochs with a batch size of 8 and a learning rate of 0.001 due to time constraints. Additionally, A key point to note is that for training we use only geometric information and no image data for a fair comparison against methods utilising image data.

Table 5: Mean average precision mAP @ 0.25 Enhanced MLCVNet comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 1.

Method	bath	bed	bookshelf	chair	desk
VoteNet	74.4	**93.0**	28.8	75.3	22.0
MLVCNet	79.2	85.8	31.9	75.8	26.5
GroupFree3D	**80.0**	87.8	**32.5**	**79.4**	**32.6**
Ours RANSAC	76.7	79.6	15.6	60.9	9.8
Ours DBSCAN	23.8	74.0	5.6	55.4	7.6
Ours BIRCH	26.5	76.7	4.6	56.4	8.5
Ours KMeans	31.2	76.7	4.6	56.4	8.5
Ours OPTICS	25.6	79.2	12.3	43.9	6.2

Enhanced GroupFree3D. For the implementation of this method, we follow the provided instructions of the GroupFree3D paper. This includes the use of a PointNet++ backbone with 4 set abstraction layers and 2 feature propagation layers and transformer decoder module to support a multi-head attention mechanism for iterative object feature extraction and box prediction for a fair comparison and the use of the common benchmark SUN RGB-D training and validation datasets. The integration of the iterative sampling framework method includes the modifying the SUN RGB-D detection dataset class to include the option to run GroupFree3D with RANSAC iterative sampling. To achieve the results described in Tables 3 and 4 below, we again use 20,000 points as the input for each point cloud scene. We run 400 epochs with a batch size of 8 and a learning rate of 0.001 due to time constraints. A key point to note is that for training we use only geometric information and no image data for a fair comparison against methods utilising image data.

Table 6: Mean average precision mAP @ 0.25 Enhanced MLCVNet comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 2.

Method	dresser	nightstand	sofa	table	toilet
VoteNet	29.8	62.2	64.0	47.3	90.1
MLVCNet	31.3	61.5	66.3	50.4	89.1
GroupFree3D	**36.0**	**66.7**	**70.0**	**53.8**	91.1
Ours RANSAC	12.2	35.4	27.4	39.2	**95.5**
Ours DBSCAN	9.2	23.2	28.0	34.3	86.1
Ours BIRCH	8.6	13.1	21.7	24.0	72.8
Ours KMeans	9.6	26.5	32.1	36.1	87.3
Ours OPTICS	7.8	19.0	25.3	26.1	77.7

Table 7: Mean average precision mAP @ 0.5 Enhanced MLCVNet comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 1.

Method	bath	bed	bookshelf	chair	desk
VoteNet	45.4	53.4	6.8	56.5	5.9
GroupFree3D	**64.0**	**67.1**	**12.4**	**62.6**	14.5
Ours RANSAC	16.6	32.2	10.0	36.2	3.9
Ours DBSCAN	14.5	29.1	4.7	23.2	4.4
Ours BIRCH	13.9	36.1	2.6	17.5	3.4
Ours KMeans	27.8	31.2	7.1	23.5	**29.1**
Ours OPTICS	16.5	33.4	3.8	15.9	8.7

4.2 System Performance

As demonstrated by the experimental results performed for this work, there is significant potential to further enhance existing state-of-the-art 3D point cloud object detection methods with the use of iterative sampling and clustering methods. Our proposed framework demonstrates improvements on the state-of-the-art VoteNet and MLCVNet methods for 2 classes in each evaluation run. Due to time constraints, our experimental works on clustering methods and the use of the ScanNet V1 dataset was omitted from this version of the paper. We demonstrate the success of the evaluated RANSAC iterative sampling method on the SUN RGB-D validation dataset.

For Enhanced VoteNet, we improve on the state-of-the-art mAP results for the bath and desk classes @ 0.25 with +0.4 and +64.4 respectively. For Enhanced VoteNet with mAP @ 0.5 we improve on the state-of-

Table 8: Mean average precision mAP @ 0.5 Enhanced MLCVNet comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 2.

Method	dresser	nightstand	sofa	table	toilet
VoteNet	12.0	38.6	49.1	21.3	68.5
GroupFree3D	**21.9**	**49.8**	**58.2**	**29.2**	**72.2**
Ours RANSAC	7.3	7.0	11.6	8.4	56.2
Ours DBSCAN	5.3	32.8	20.3	13.7	39.9
Ours BIRCH	7.1	37.6	42.5	17.9	11.2
Ours KMeans	8.0	39.6	21.0	14.3	45.1
Ours OPTICS	9.8	40.4	36.8	11.9	28.5

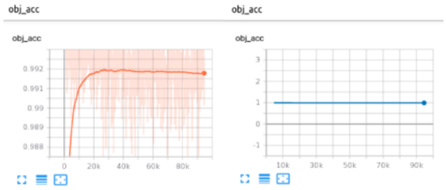


Figure 7: Average Object Accuracy of 99.18% and 99.09% on SUN RGB-D training and validation datasets.

the-art mAP results for the bookshelf and desk classes with +1.8 and +41.9. We also demonstrate an object accuracy of 99.27% during training and 99.84% during testing.

For Enhanced MLCVNet, we improve on the state-of-the-art mAP results again for the desk classes @ 0.25 with +25.8. For Enhanced MLCVNet with mAP @ 0.5 we did not achieve any improvements on the state-of-the-art mAP results for any classes in the validation dataset. We also demonstrate an object accuracy of 99.18% during training and 99.09% during testing.

For Enhanced GroupFree 3D, we do not improve on the state-of-the-art mAP results @ 0.25 or @ 0.5. Due to time constraints the number of epochs for training this model was reduced. We do however demonstrate an object accuracy of 99.23% during training and 99.11% during testing. Overall, it is clear from the experimental results that the addition of the iterative sampling method to each of the current state-of-the-art methods can achieve improved results due to the filtering of outlier points. However, it is also clear that this is inconsistent across all object classes in the SUN RGB-D dataset and will require future works to fine tune and improve results on other classes to yield improved results.

5 CONCLUSIONS

In this paper, we propose an iterative sampling and clustering framework to enhance 3D point cloud object detection. For iterative sampling we utilize the popular RANSAC algorithm which allows for the filtering out outlier points in the input point cloud. For clustering, we utilize the DBSCAN, K-means, BIRCH and OPTICS algorithms which are widely used for data pre-processing techniques. We evaluate our framework by integrating to the current state-of-the-art VoteNet, MLCVNet and GroupFree3D methods which boast the fastest, most accurate and highest performing results across the benchmark SUN RGB-D and ScanNet V2 point cloud datasets.

Through the experimental results demonstrated in this paper, the RANSAC iterative sampling method

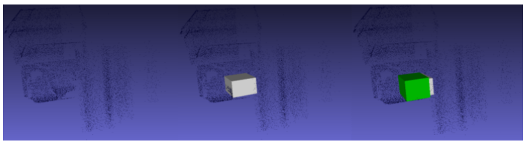


Figure 8: Input point cloud scene (20k points, Enhanced MLCVNet bounding box prediction, Ground truth comparison vs prediction.

Table 9: Mean average precision mAP @ 0.25 Enhanced GroupFree3D comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 1.

Method	bath	bed	bookshelf	chair	desk
VoteNet	74.4	**93.0**	28.8	75.3	22.0
MLVCNet	79.2	85.8	31.9	75.8	26.5
GroupFree3D	**80.0**	87.8	**32.5**	**79.4**	32.6
Ours RANSAC	30.8	38.9	10.0	35.1	**58.4**
Ours DBSCAN	29.7	36.6	21.8	33.7	37.2
Ours BIRCH	34.3	44.9	24.7	39.1	40.3
Ours KMeans	31.3	34.6	20.3	35.2	38.1
Ours OPTICS	33.4	43.5	23.1	37.8	39.0

Table 10: Mean average precision mAP @ 0.25 Enhanced GroupFree3D comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 2.

Method	dresser	nightstand	sofa	table	toilet
VoteNet	29.8	62.2	64.0	47.3	90.1
MLVCNet	31.3	61.5	66.3	50.4	89.1
GroupFree3D	**36.0**	**66.7**	**70.0**	**53.8**	**91.1**
Ours RANSAC	13.2	11.3	17.9	53.6	49.6
Ours DBSCAN	16.6	22.9	23.5	49.7	50.4
Ours BIRCH	19.8	26.3	19.2	51.0	63.8
Ours KMeans	28.5	32.6	28.1	52.0	51.3
Ours OPTICS	21.8	23.7	19.0	50.5	62.1

Table 11: Mean average precision mAP @ 0.5 Enhanced GroupFree3D comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 1.

Method	bath	bed	bookshelf	chair	desk
VoteNet	45.4	53.4	6.8	56.5	5.9
GroupFree3D	**64.0**	**67.1**	12.4	**62.6**	14.5
Ours RANSAC	38.6	41.3	10.2	51.7	12.9
Ours DBSCAN	32.0	42.4	13.8	37.6	36.4
Ours BIRCH	30.8	38.9	20.0	35.1	**58.4**
Ours KMeans	34.3	44.9	**24.7**	39.1	40.3
Ours OPTICS	33.1	40.2	16.4	38.0	41.4

Table 12: Mean average precision mAP @ 0.5 Enhanced GroupFree3D comparison against current state-of-the-art methods on SUN RGB-D v1 validation set - Part 2.

Method	dresser	nightstand	sofa	table	toilet
VoteNet	12.0	38.6	49.1	21.3	68.5
GroupFree3D	**21.9**	**49.8**	**58.2**	29.2	**72.2**
Ours RANSAC	16.3	21.4	18.6	43.3	51.2
Ours DBSCAN	14.1	20.9	22.3	47.8	55.7
Ours BIRCH	13.2	11.3	17.9	**53.6**	49.6
Ours KMeans	19.8	26.3	29.2	51.0	63.8
Ours OPTICS	20.2	18.7	21.9	29.0	54.7

can be a useful addition to enhance current state-of-the-art 3D point cloud object detection methods, as shown with the improvements made on the state-of-the-art mean average precision values @ 0.25 and @ 0.5 for some classes. However, along with this, the experimental results proved that the iterative sampling method caused inconsistency across all classes. This indicates the limitations of utilizing ou unsupervised iterative sampling and clustering framework on a dataset of varying classes and object shapes/sizes demonstrating this may be best suited to applications with primitive shapes or similar point cloud scenes. In future works, we plan to further extend and fine tune the framework to achieve superior results on other common benchmark datasets.

The results show that Enhanced VoteNet and Enhanced MLCVNet achieved high object accuracy results for both training and testing on the benchmark SUN RGB-D dataset with all runs yielding object accuracy results greater than 99.1% which is promising. The objective of this work is to evaluate the above dataset and methods using key considerations of industrial applications which has not been previously done for raw point cloud object detection methods. VoteNet and MLCVNet, which were implemented on the 3D point cloud dataset, show promising results in terms of accuracy, computation, and real-time capability for industrial applications. However, one additional consideration which needs further evaluation is the process of updating the models for new object classes, changes in ambient conditions or infrastructure in an industrial setting as this is important in modern real-world applications.

REFERENCES

Saifullahi Aminu Bello, Shangshu Yu, and Cheng Wang. Review: Deep learning on 3d point clouds. 2020.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. 2020.

B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. pages 510–517, 2015.

A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. 2015.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. 2017.

Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd - a dataset for 3d object recognition in industry. Oct 2017.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. 2013.

Timo Hackel, N. Savinov, L. Ladicky, Jan D. Wegner, K. Schindler, and M. Pollefeys. Semantic3d.net: A new large-scale point cloud classification benchmark. volume IV-1-W1, pages 91–98, 2017.

S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab. Modelbased training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. 2012.

T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis. T-less: An rgb-d ˇ dataset for 6d pose estimation of texture-less objects. 2017.

R. Larsen, H. Aanaes, and S. Gudmundsson. Fusion of stereo vision and time-of-flight imaging for improved 3d estimation. volume 1, pages 1–9, 2019.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. 2018.

Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. 2019.

D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. pages 922–928, 2015.

Charles R. Oi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. Apr 2017.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in metric space. 2017.

Charles R. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. 2019.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. pages 567–576, 2015.

Gusi Te, Wei Hu, Zongming Guo, and Amin Zheng. Rgcnn: Regularized graph cnn for point cloud segmentation. 2018.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. Jun 2019.

Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. 2015.

Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. MLCVNet: Multi-level contextvotenetfor 3d object detection. 2020.

Ze Liu, Zheng Zhang, Yue Cao, Han Hu, Xin Tong. Group-Free 3D Object Detection via Transformers. 2021.