Dense Point Diffusion for 3D Object Detection

Xu Liu[1]* Jiayan Cao[2] Qianqian Bi[2] Jian Wang[3] Boxin Shi[4] Yichen Wei[2]†
[1]The University of Tokyo [2]Megvii Research Shanghai [3]Snap Inc [4]Peking University

Abstract

The backbone network adopted in state-of-the-art 3D object detectors lacks a good balance between high point resolution and large receptive field, both of which are desirable for object detection on point clouds. This work proposes Dense Point Diffusion module, a novel backbone network that solves these issues. It adopts dilated point convolution as a building block to enlarge the receptive field and retain the point resolution at the same time. Further, a number of such layers are densely connected, giving rise to large receptive field and multi-scale feature fusion, which are effective for object detection task. Comprehensive experiments verify the efficacy of our approach. The source code[1] has been released to facilitate the reproduction of the results.

1. Introduction

3D object detection on point clouds [23, 5, 7] is crucial for applications in autonomous driving, robotics, and virtual reality. The problem is challenging because the point clouds are typically sparse and have irregular structures. To deal with these issues, a category of previous methods [29, 25, 12, 22] firstly convert the point clouds into regular grid-like data structures via a quantization process, such as 3D voxels, and then apply conventional 3D or 2D CNN based detectors. Such methods are easy to apply but suffer from information loss in the quantization process. More recent methods [17, 21, 27] directly work on the irregular point clouds. They adopt PointNet/PointNet++ [19, 20] as backbone networks to extract point-wise features. Without information loss caused by quantization, such methods perform better and lead the state of the art.

However, these methods suffer from the insufficiency of the backbone network. PointNet++ [20], which is not developed for object detection at first, as shown in Figure 1 (a), consists of a series of down-sampling and up-sampling layers, mimicking the design of modern 2D image-based deep CNNs. Down-sampling is via Set Abstraction (SA) operator. The receptive field of points is enlarged, but the point resolution is reduced. It could not retain high point resolution as well as large receptive field, both of which are desirable for object detection. Up-sampling is via Feature Propagation (FP) operator. Point resolution is increased but the new point features are recovered by coarse interpolation, which will bring sub-optimal results to 3D object detection, as shown in Figure 1 (b).

This work proposes a novel backbone network, illustrated in Figure 1 (c), which solves the above issues. Instead of using conventional SA-FP structure, Dilated Point Convolution (DPC) is adopted to enlarge the receptive field, without losing point resolution. Further, a number of such layers with increasing dilation rates are densely connected, giving rise to quickly increasing receptive field size, multi-scale feature fusion, and dense feature sampling, as shown in Figure 1 (d). All these factors are beneficial for 3D object detection on the sparse and irregular point cloud. We call the resulting module *Dense Point Diffusion*. We note that, the ingredients of dilated convolution [28, 1], dense layer connection [9, 26] and multi-scale feature fusion [2, 3, 24] are well known and effective techniques for 2D object detection and segmentation. This work, for the first time, adapts them for 3D object detection.

Comprehensive experiments and ablation studies on two prevailing 3D object detection benchmarks (SUN RGB-D [23] and ScanNet[5]) verify the efficacy of proposed approach. Our method surpasses the conventional PointNet++ [20] across the board, indicating it to be a compelling alternative to the conventional PointNet++ [20].

2. Related Works

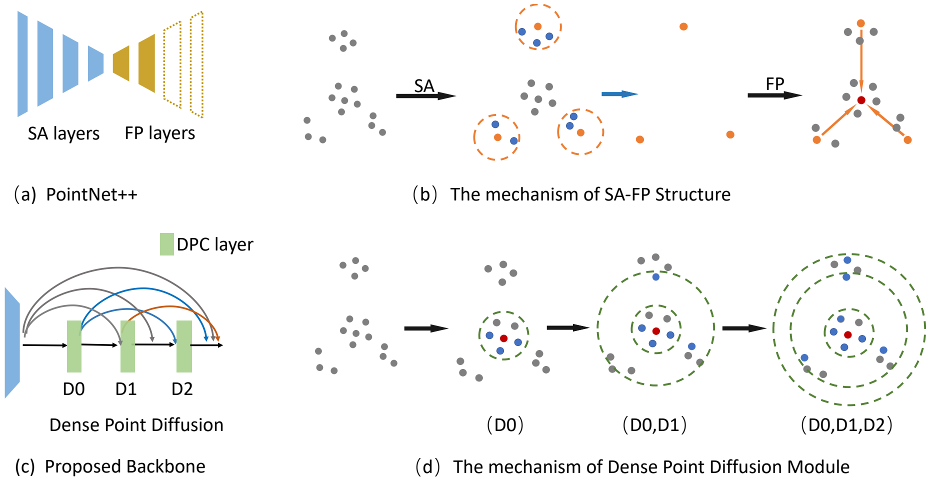
The methods for 3D object detection on point clouds by deep learning could be roughly grouped into two categories, based on whether point clouds are converted into other representations or not.

With conversion. Directly working on raw point clouds is difficult because the data is scattered and unordered. Early works usually convert the irregular point cloud data into regular data either by multi-view projection/rendering

*This work was done in Megvii Research Shanghai

†Corresponding Author. Email: weiyichen@megvii.com

[1] https://github.com/AsahiLiu/PointDetectron



(a) PointNet++

(b) The mechanism of SA-FP Structure

(c) Proposed Backbone

(d) The mechanism of Dense Point Diffusion Module

Figure 1. Comparison of backbone of PointNet++ [20] (a) and the proposed method (c). PointNet++ uses a series of down-sampling (via SA operator) and up-sampling (via FP operator) layers (a). (b) illustrates a typical data-flow via a pair of SA and FP operators: the orange points are chosen by SA layer according to the Farthest Point Sampling strategy to enlarge the receptive field. Then in the FP layer, the feature at the red point will be coarsely interpolated from the three orange points. The design couldn't retain high point resolution as well as large receptive field. Both are desirable for object detection. Unlike PointNet++, the proposed module (c) adapts Dilated Point Convolution (DPC), which samples the points with the fixed stride, to enlarge the receptive field and keep point resolution, as introduced in Section 4.1 and 4.2 later. (d) shows that the proposed module can yield large receptive field (represented by expanding dashed circles) and dense point sampling (represented by blue points).

or voxelization. In [4, 11, 14], the raw input of point cloud is firstly projected into bird's eye view images (BEV) and then 2D convolution neural network is applied to generate BEV features. These methods yield learned features from both BEV and RGB images to refine 3D bounding boxes but suffer from coarse feature representation during projection.

For voxel-based methods [29, 25, 12], they firstly encode the raw input of point cloud into grid-like representation and then apply conventional operators (e.g. 2D or 3D convolution in [29]) to extract feature maps. However, successive usage of voxelization may inevitably lead to information loss during the quantization process, as described in Section 1. Moreover, the increasing demands on higher resolution would also exacerbate the dilemma between the sparsity of point clouds and fine-grained features, since the resolution of voxel-grids grows in cubic. To remedy this, Second [25] utilizes sparse convolution operator [15] to boost the real-time performance. While PointPillar [12] designs pillar-shaped structures to further balance the trade-off between computational cost and detection performance, the overall performance, however, is barely satisfactory due to the fundamental information loss.

Without conversion. In point cloud classification task, pioneering work PointNet [19] proposes novel network design with permutation-free function property, where the order of input data does not influence output, to work on irregular point cloud data directly. Coming after it, PointNet++ [20] proposes the "encoder-decoder" framework: the Set Abstraction (SA) operators can be regarded as encoder to extract the features and sub-sample the input points hierarchically, and Feature Propagation (FP) is deployed to recover the feature back to original size.Following them, object detection methods via deep networks [17, 21, 27] that directly process the point cloud were proposed, by borrowing operators or backbone design from [20]. However, it is neglected that the structures for object classification could be unsuitable for the object detection task. Backbone design of PointNet++ [20] enlarges receptive field but loses resolution. In the next section, we will analyze the drawbacks of this design in details.

In this paper, we adapt dilated point convolution to object detection task and design a dedicated backbone where both large receptive field and high point resolution could be kept. Even though similar idea [6, 13] has been adopted for point cloud segmentation, we are the first to utilize this operator for the challenging 3D object detection task. Additionally, we have analyzed how to construct the *Point Convolution-based* backbone module with this operator in depth, which is crucial for 3D object detection. The method in Densepoint [16] can be regarded as a special case of our module when all of the dilation rates in our module are set as 1. In contrast, our method employs denser sampling as well as denser scale feature aggregation on point sets , which can be leveraged to tackle the challenges posed by the sparsely and irregularly distributed point clouds. Moreover, we have abandoned the use of the flawed FP operator in our design, which was regarded as a crucial part in [16, 20].

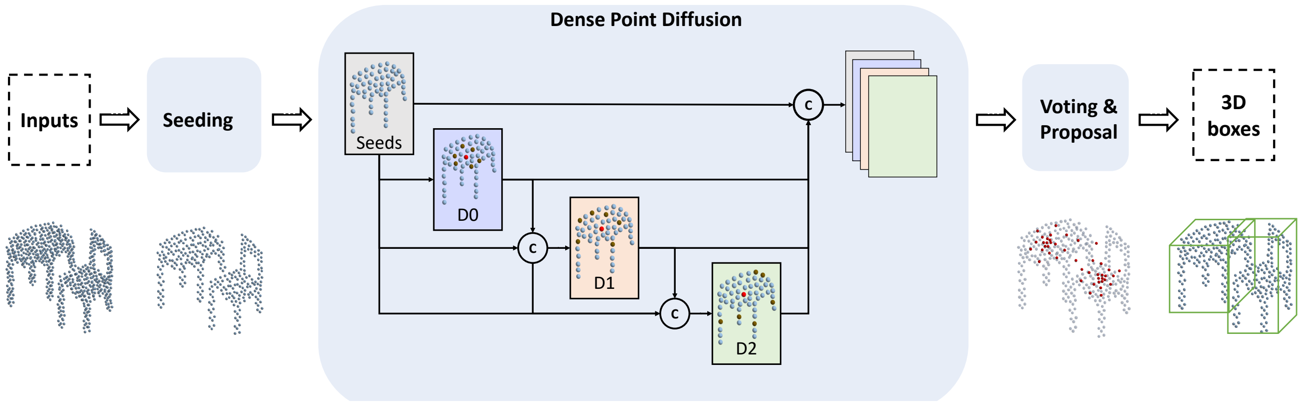


Figure 2. The framework of proposed 3D detector. Given the input point cloud, the seeding process will down-sample the points by Farthest Point Sampling method. The seeds are then fed to the Dense Point Diffusion module. In this module, feature maps from all previous layers are concatenated before being fed into the next stage of the Dilated Point Convolution. The concatenation operation is represented by "c" in the diagram. After generating point-wise feature via the proposed backbone, voting and proposal module is applied to vote and cluster the seed points and then predict the 3D bonding boxes.

3. Drawbacks of Existing Backbone Design

PointNet++ [20] is widely used by the SOTA 3D detectors [17, 21, 27] as backbone network. The backbone is illustrated in Figure 1 (a). Multiple SA (Set Abstraction) layers are introduced to enlarge the receptive field, with the point number or point resolution reduced to half sequentially. To make up for the loss of point resolution, the operator of FP (Feature Propagation) is introduced to recover the features from lower resolution.

SA module picks a subset of points from the input point clouds via farthest point sampling (FPS). Then features of sampled points are grouped and further encoded by a standard PointNet layer. For the FP layer, the features of missing points are recovered by interpolation from their nearest neighbours. The up-sampled features will be concatenated with features from corresponding SA layers and pass through the MLP (Multi-Layer Perception) layers for feature propagation. Such procedure will repeat until the original point resolution is fully recovered.

The limitations of these operators are illustrated in Figure 1 (b). It is sub-optimal for SA operator to reduce the point resolution, since the information of the points will be reduced by FPS (Farthest Point Sampling) sub-sampling method. The FP operator, a coarse up-sampling operator , interpolating the point features with a limited number of point candidates, is deployed to recover back to the original point resolution. Since point clouds are usually sparsely and irregularly distributed, the points chosen for interpolation may be loosely-correlated with the contextual feature of interpolated points. This will bring unfavorable effects to the object detection.

As a result, these choices would lead to unsatisfying results for 3D object detection. Actually, this is reported from [17] and illustrated in Figure 6 that contrary to the common belief, the increasing on the number of layers in PointNet++ [17] does not necessarily improve the performance of the 3D detection, thus inhibiting the performance for 3D detection.

In summary, this problem arises from the inherent drawbacks of PoinNet++ backbone: SA layer enlarges the receptive fields at the cost of point resolution. Then in order to recover back to its original resolution, features will be coarsely up-sampled, inevitably introducing noise to feature aggregation. We will introduce our method Dense Point Diffusion in the next section to address this issue.

4. Method

We introduce our Dense Point Diffusion module for 3D point cloud object detection. At first, we introduce the operator of Dilated Point Convolution in Section 4.1, which can preserve the point resolution and enlarge receptive field at the same time. Then in Section 4.2, we discuss the rationales to design our module of "Dense Point Diffusion" with this operator. Finally, in Section 4.3, this module is applied to VoteNet [17], the current SOTA algorithm for 3D object detection, which is illustrated in Figure 2. The proposed module can boost the performance of the detector significantly and we will discuss it in the next section.

4.1. Dilated Point Convolution

The conventional point convolution can be formulated in Equation 1. For brevity, we dub x_i as coordinate of the point and f_i as its corresponding feature. The convolution of feature f at the point x with kernel g is given by

$$(f * g)(x) = \sum_{x_i \in N_x} g(x_i - x) f_i \tag{1}$$

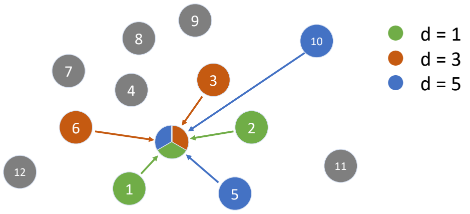


Figure 3. Dilated Point Convolution. To enlarge the receptive field effectively, we choose the d, $2d$,..., kd-th nearest points, and d is referred to as the dilation rate. Let $k = 2$. The figure shows the choice of points with different dilation rate 1, 3, or 5.

where $N_x = \{\|x_i - x\| \leq r\}$ is the set of the points chosen by neighbourhood within certain radius r. In practice, the points in N_x are sampled by k-NN method.

Large receptive fields are essential to detect big objects. The previous methods mentioned in Section 3 use the operator of SA to increase the receptive field. However, the point number or point resolution will be reduced in this procedure. It is essential to preserve point resolution to achieve high performance on 3D detection tasks.

We enlarge the receptive field by choosing the nearest d-th, $2d$-th,..., kd-th candidates within the radius of neighborhood and denote such operator as *Dilated Point Convolution*, which is illustrated in Figure 3. The fixed stride d is referred as the dilation rate of the Dilated Point Convolution. It should be noted that the receptive field of the point sets could be extended while the point resolution is preserved at the same time. Based on the operator of Dilated Point Convolution, we can obtain high quality contextual features at different scales, which will be valuable for high performance 3D detection tasks. We further propose an efficient 3D detection Backbone based on this operator.

4.2. Dense Point Diffusion Module

In this section, we mainly discuss how to build a suitable module based on the operator of Dilated Point Convolution, which can effectively aggregate the multi-scale contextual features for 3D object detection.

A straightforward way to aggregate the multi-scale contextual feature is to concatenate features at different dilation rates, which is similar to the method of Deeplab [1]. Multi-scale features are computed by applying a set of dilated convolution operators with various dilation rates in parallel and then concatenated to form the features. We dub this method as Point Diffusion module because multiple dilated operators will expand the range of the points to different locations, which is analogous to the diffusion process of a point. The Point Diffusion module is expressed by

$$y = \mathcal{H}_{N,D_0}(x) + \mathcal{H}_{N,D_1}(x) + ... + \mathcal{H}_{N,D_l}(x) \quad (2)$$

where $H_{N,D}(x)$ is the point convolution of N points with dilation rate D, y is the output feature, and $D_0, D_1, ...D_l$ stand for different dilation rates that operate at multiple scales. We use '+' to stand for the concatenation operation.

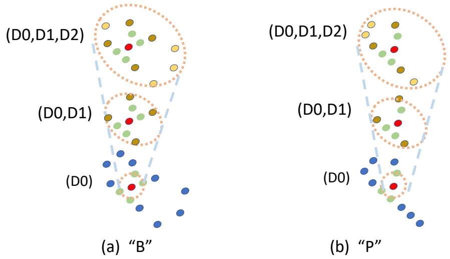


Figure 4. Dilated point convolution on the stacked feature maps with different dilation rate brings dense sampling of the points, thus distinguishing the sparse point set with confusing appearance like "B" and "P" in this case.

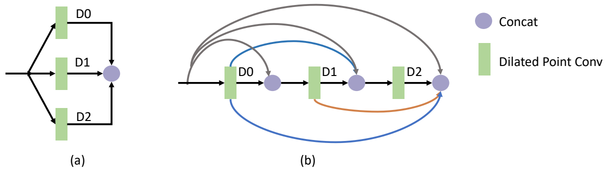


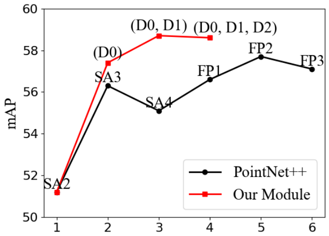
Figure 5. Comparison between Point Diffusion (a) and Dense Point Diffusion (b) structures.

It can be inferred from this expression, to cater for varying sizes of the objects in the scenarios, this method requires a wide range of dilation rates to cover the entire scene, resulting in tremendous use of dilated convolution operators.

However, the intensive use of dilated point convolution is computationally expensive. We observe that in this design, the feature generated by the dilated point convolution is only utilized once, which is inefficient and thus costly. Therefore, an important issue has been aroused. Can we find a way to represent multi-scale features efficiently by utilizing only a limited choice of dilated point convolution operators and then reuse the dilated convolution feature maps? This inspires us to propose the method of "Dense Point Diffusion" in the following.

It is known in 2D CNN, convolution on concatenated CNN features can yield larger receptive field size. Similarly, assuming the input point clouds are uniformly distributed in the space, point convolution on the stacked feature maps with two receptive field $\mathcal{R}_1$ and $\mathcal{R}_2$ yields a larger receptive field, which is approximately written as $\mathcal{R} = \mathcal{R}_1 + \mathcal{R}_2$. If the dilation rate of this operator is d, the receptive field size is then expanded by $d\times$, which can be expressed as $\mathcal{R} \approx d \times (\mathcal{R}_1 + \mathcal{R}_2)$. The visualization illustrated in Figure 4 shows that this method can also densely aggregate the points at different scales, which is beneficial for handling the challenges posed by sparse point clouds.

Consequently, to further exploit the potentials of Dilated Point Convolution for object detection, we propose a novel Dense Point Diffusion module illustrated in Figure 2. Given



PointNet++	SA2	SA3	SA4	FP1	FP2	FP3
Point Resolution	1024	512	256	512	1024	2048
mAP	51.2	56.3	55.1	56.6	**57.7**	57.1
Our Module	SA2	(D0)	(D0,D1)	(D0,D1,D2)	_	_
Point Resolution	1024	1024	1024	1024	_	_
mAP	51.2	57.4	**58.7**	58.6	_	_

Figure 6. The results show that in contrast with the PointNet++, the performance of our module is better and can be generally improved when more stages are incorporated (except for the third stage when performance get saturated). *Left*: The plot shows the tendency of performance of PointNet++ and our module w.r.t. the seeding layers, evaluated with mAP@0.25 on SUN RGB-D. *Right*: The table shows the performance of PointNet++ (quoted from Table 8 of VoteNet [17]) and our module w.r.t. the seeding layers, evaluated with mAP@0.25 on SUN RGB-D.

the input feature, this module progressively concatenates features from the previous stages and feeds them to the next stage of Dilated Point Convolution, generating features with larger receptive field. In the parallel direction, the concatenation operator aggregates features inclusively from each of the previous layers, covering a wider range of scales. The concatenated features are set as the output of the module and are formulated as following,

$$y_l = \mathcal{H}_{N,D_l}[y_{l-1}, y_{l-2}, y_{l-3}, ..., y_0] \tag{3}$$

In this equation, the input of the current stage is the concatenation of all its previous feature maps. D_l is the dilation rate and y_l is the output feature at l-th stage. We arrange the dilation rate in ascending order so that the size of receptive field and range of feature scales are gradually increased. Empirically, we choose $D_0 = 3$, $D_1 = 6$, $D_2 = 12$.

The difference in structure between Point Diffusion and Dense Point Diffusion module is illustrated in Figure 5. In comparison, densely-connected version not only has larger receptive field but also can cover wider range of scales with the same configuration of Dilated Point Convolution, making it a preferable choice for the challenging 3D object detection task.

4.3. Architecture of Detector

To verify the efficacy of our module, we experiment on VoteNet [17] for 3D object detection. As illustrated in Figure 2, we replace the original backbone with our module Dense Point Diffusion. The architecture of detector consists of backbone, voting and proposal modules.

The backbone network takes the raw point cloud as input and generates feature maps of points. Its structure can be further split into two parts: seeding layer and Dense Point Diffusion module. Seeds refer to a subset of points chosen from the raw point cloud for voting process, which is also defined in [17]. The seed features will then be processed by our method Dense Point Diffusion and fed into the voting module to generate the votes. The Votes will be grouped into clusters and processed by proposal module to predict the 3D boxes. For fair comparison, the configuration of voting and proposal module is the same as in [17].

5. Experiments

In this section, we conduct a complete series of ablation studies to verify the efficacy of the proposed backbone. The results on two standard benchmark are also reported, which surpass the conventional PointNet++ across the board, indicating our module Dense Point Diffusion a compelling alternative for PointNet++.

5.1. Dataset

SUN RGB-D [23] for 3D indoor scene understanding consists of around 10k RGB-D images annotated with 64,595 oriented 3D bounding boxes for nearly 40 object categories. In our experiment, following [17], we split the training/testing set and report 3D detection performance on the 10 most common categories.

ScanNet [5] provides a wider range of indoor scenes with more densely scanned objects compared with the SUN RGB-D dataset. We use the 1205 scans and 312 scans for training and testing, respectively. Vertices from meshes are sampled as the input point clouds. Following the ground-truth annotation mentioned in [17], we predict axis-aligned 3D bounding boxes in these scenarios.

In experiments, we follow the same protocol in [17] and use the metrics, mean average precision (mAP) at IoU threshold 0.25, for evaluation.

5.2. Implementation Details

The Seeding Layers are indispensable in our design. If these intermediate layers are totally removed, the model capacity will be reduced and the performance will be compromised. To make up for the loss in model capacity, more dilated operators performing on original point resolution need to be introduced, which will be computationally expensive.

Table 1. Ablation studies w.r.t. the seeding layer and the setting of the Dense Point Diffusion module.

	Method	mAP@0.25		
		SUN RGB-D V1	SUN RGB-D V2	ScanNet V2
(a)	SA2 - (SA3 - SA4 - FP1 - FP2) [29]	57.7	59.2	58.6
(b)	SA2	51.2	54.0	51.2
(c)	SA2 - Dense Point Diffusion (3)	57.4	59.2	57.2
(d)	SA2 - Dense Point Diffusion (3, 6)	**58.7**	59.8	58.9
(e)	SA2 - Dense Point Diffusion (3, 6, 12)	58.6	59.6	**59.6**
(f)	SA3	56.1	58.0	54.4
(g)	SA3 - Dense Point Diffusion (3)	57.6	59.8	55.7
(h)	SA3 - Dense Point Diffusion (3, 6)	58.4	**60.3**	56.1
(i)	SA3 - Dense Point Diffusion (3, 6, 12)	58.2	60.1	56.2

Table 2. Additional experiments to justify the design of dense aggregation and the choice of dilation rates in the proposed modules.

	Method	mAP@0.25		
		SUN RGB-D V1	SUN RGB-D V2	ScanNet V2
(a)	SA2 - Cascaded (3,6,12)	50.4	52.4	42.7
(b)	SA2 - Point Diffusion (3, 6, 12)	58.0	**59.6**	54.9
(c)	SA2 - Dense Point Diffusion (3, 3, 3)	57.9	59.2	58.1
(d)	SA2 - Dense Point Diffusion (3, 6, 12)	**58.6**	**59.6**	**59.6**

Therefore, the seeding layers at initial stage should be preserved for the trade-off between model capacity, computational cost as well as point resolution. With reference to the results shown in Figure 6 (right), quoted from Table 8 of VoteNet [17], we choose seed points from SA2 and SA3 layers because they can meet the requirements for point resolution and model capacity. The configuration of these SA layers (SA1, SA2, SA3) are the same as in VoteNet [17] for fair comparisons.

Point Diffusion module The operators of Dilated Point Convolution in the module have increasing receptive radius r of 0.8, 1.2 and 1.8 and ascending dilation rate 3, 6 and 12 accordingly. The value of k for k-NN search is set to be 32, which is the same with SA2 layer in VoteNet. The number of points to be sampled from are scaled according to the dilation rate, which is 96, 192 and 384 respectively. In the experiments, we use the term “SA2 - Dense Point Diffusion $(a,\ b,\ c)$” or “SA2 - Point Diffusion $(a,\ b,\ c)$” to represent the configuration of our module, where SA2 denotes the seeding layer, and the list of $(a,\ b,\ c)$ represents the dilation rates to be a, b and c respectively.

Miscellaneous Our framework is borrowed from VoteNet [17] and we keep the configuration unchanged for a fair comparison with VoteNet. We adopt the same loss function and loss weight with VoteNet. We also choose the same vote aggregation method, Vote factor in our design. Additionally, we deploy the FPS (Farthest Point Sampling) method to choose the cluster for the votes and incorporate the “height feature” into the framework, which is the same as in the VoteNet.

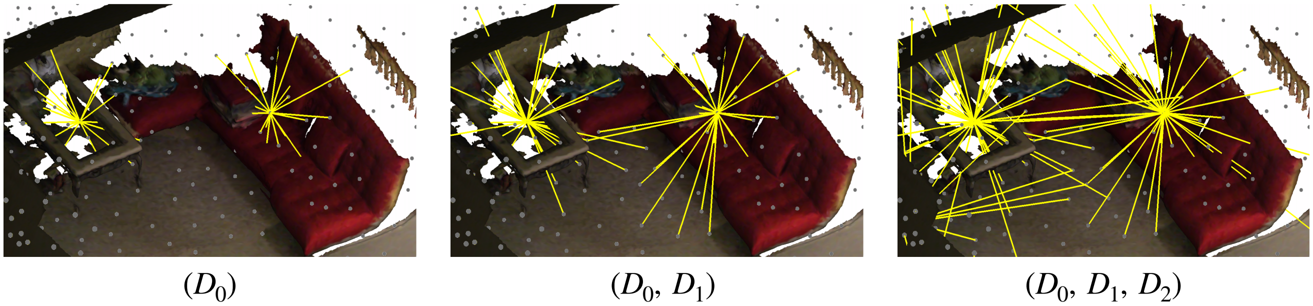
5.3. Training and Inference

The process of data augmentation is the same as VoteNet [17] to make a fair comparison. We adopt Adam Optimizer [10] with an initial learning rate of 0.001. Learning rate is then scheduled to be decayed by the factor of 0.1 after 80 epochs and another 0.1 after 120 epochs. There are 180 epochs in total. The whole model is trained on a single Nvidia GTX-1080 Ti GPU. In the inference stage, points of the entire scenes are taken as the input. With a single feed-forward path, the region proposals are generated by the network and further post-processed by 3D NMS (Non-maximum Suppression) method to generate the 3D bounding boxes.

5.4. Ablation Studies

Going Deeper The results in Table 1 shows the tendency of performance w.r.t. the number of stages in our module. The results are basically able to verify the hypothesis that increasing stage of the module contributes to better performance. Compared with the performance of SA2 baseline (i.e., Table 1 (b, c)), the initial stage of 3× dilated convolution helps to achieve the boost of 6.2 mAP and 6.0 mAP in performance for SUN RGB-D V1 and ScanNet V2 respectively. With extra two more stages involved, the improvement can be up to 7.7 mAP for ScanNet V2, which is illustrated in Table 1 (b, d).

Efficacy of Dense Feature Aggregation To verify the effi-



(D_0) (D_0, D_1) (D_0, D_1, D_2)

Figure 7. Visualization of enlarged receptive field (represented by the extended rays from the dots) and denser point sampling (represented by denser rays) with increased stage of our module.

Table 3. Comparison with the state-of-the-art approaches on SUN RGB-D V1 benchmark.

Method	Input	bathtub	bed	book shelf	chair	desk	dresser	nightstand	sofa	table	toilet	mAP
F-PointNet [18]	Geo+RGB	43.3	81.1	**33.3**	64.2	24.7	**32.0**	58.1	61.1	**51.1**	**90.9**	54.0
VoteNet [17]	Geo Only	**74.4**	83.0	28.8	**75.3**	22.0	29.8	62.2	64.0	47.3	90.1	57.7
Ours	Geo Only	71.9	**86.3**	30.5	74.1	**26.3**	30.3	**63.1**	**65.2**	49.2	90.3	**58.7**

cacy of "Dense Aggregation" in our module, we remove all the concatenation operators in our design and degenerate it into the cascaded form. The comparison of performance listed in Table 2 (a) and (d) shows that on the benchmark of SUN RGB-D V1, the performance of the cascaded version is 8.2 mAP lower than that of the Dense Point Diffusion module. Although the structure of cascaded dilated point convolution can help to increase the receptive field, large receptive field alone is not enough. The overall performance is limited because the range of the feature scales is limitted. Therefore, the method to effectively aggregate feature from multiple scales is the key to better performance.

To show the advantage of our Dense Point Diffusion for feature aggregation, we also make comparison between Dense Point Diffusion and the ordinary version, Point Diffusion, which is shown in Table 2 (b) and (d). The two methods share the same configuration of input seeding layers and dilation rates. The results show that the Dense Point Diffusion method outperforms the Point Diffusion counterpart (except for the case of SUN-RGBD V2).

To summarize, Dense Point Diffusion module combines the merit of the parallel pathways to densely aggregate the multi-scale features and the cascaded structure of Dilated Point Convolution to gain larger receptive fields, thus showing advantages over other modules for 3D object detection tasks in complex scenes.

Choice of Dilation Rates As mentioned in Section 4, we arrange the dilation rate in ascending order to yield the receptive field from small to large. To verify this arrangement, we set all the dilation rates uniformly to 3. The results in Table 2 (c), (d) show that method with uniform value of dilation rate is inferior to ours (i.e., 57.9 to 58.6 in SUN RGB-D V1 and 58.1 to 59.6 in ScanNet V2), thus justifying our choice of the dilation rates.

Dense Sampling on Points Dense Point Diffusion has the merit of sampling point sets densely and being invariant to density distribution, which is crucial for aggregating features at various scales. In Figure 7, we show several examples of how the proposed module densely samples the neighboring points at different stages. With more stages introduced in the module, the receptive field is expanded steadily and more points are aggregated for better feature representation.

5.5. Main Results

We experiment on two prevailing benchmarks for 3D detection on point clouds, SUN RGB-D [23] and ScanNet [5], as illustrated in Table 1. Experiments of the proposed module on both datasets have shown promising results.

Results on SUN RGB-D We start with the SA3 seeding layer. The results in Table 1 show that the overall performance will improve when more stages of Dilated Point Convolution operators are introduced, reaching its peak performance 58.4 mAP and 60.3 mAP on the benchmarks of SUN RGB-D V1 and V2, respectively.

We also perform experiments on the basis of the SA2 feature, which is 51.2 mAP and 54.0 mAP on SUN RGB-D V1 and V2. With the increasing stage of dilated convolution operators incorporated, Dense Point Diffusion Module achieves performance of **58.7 mAP**, which is **1 mAP** over the performance of the VoteNet [17]. The details can be found in Table 3.

Table 4. Comparison with state-of-the-art baselines on ScanNetV2, evaluated with mAP@0.25.

Method	cab	bed	chair	sofa	table	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
3DSIS Geo [8]	12.75	63.14	65.98	46.33	26.91	7.95	2.79	2.3	0.00	6.92	33.34	2.47	10.42	12.17	74.51	22.87	58.66	7.05	25.36
VoteNet [17]	36.27	87.92	**88.71**	**89.62**	58.77	47.32	38.10	44.62	7.83	**56.13**	**71.69**	**47.23**	45.37	57.13	94.94	**54.70**	**92.11**	37.20	58.65
Ours	**38.42**	**89.14**	86.73	89.45	**62.70**	**48.62**	**38.14**	**49.86**	**8.03**	51.27	63.11	47.21	**59.08**	**69.28**	**95.78**	51.01	89.36	**39.42**	**59.65**

Table 5. Comparison with state-of-the-art baselines on ScanNetV2, evaluated with mAP@0.5.

	cab	bed	chair	sofa	table	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn	mAP
3DSIS Geo [8]	5.06	42.19	50.11	31.75	15.12	1.38	0.00	1.44	0.00	0.00	13.66	0.00	2.63	3.00	56.75	8.68	28.52	2.55	14.60
VoteNet [17]	8.07	76.06	**67.23**	68.82	42.36	**15.34**	6.43	28.00	1.25	9.52	**37.52**	11.55	27.80	9.96	**86.53**	16.76	78.87	11.69	33.54
Ours	**9.07**	**83.53**	64.47	**71.02**	**44.19**	13.85	**9.25**	**32.98**	**9.88**	**17.76**	29.69	**17.00**	**37.80**	**13.93**	81.51	**20.92**	**80.40**	**14.37**	**35.71**

It should also be noted that the performance of the Dense Point Diffusion module get saturated when the dilation rate reaches 12 and the performance slightly dropped with a margin from 0.1 to 0.2 in mAP. This is due to the fact that the scene of this benchmark is not complicated and the point resolution of the scene is limited. Thus we choose to deploy the module on more challenging scenarios of ScanNet [5] to validate the efficacy of our method.

Results on ScanNet The ablation studies on the SA3 feature are also illustrated in Table 1. Similar to SUN RGB-D dataset, the overall performance improves when getting more stages of Dilated Point Convolution involved. The performance of the Dense Point Diffusion is steadily better than the parallel counterpart of Point Diffusion Module. While the Dense Point Diffusion boosts the performance of SA3 baseline with a large margin of 1.8 mAP in Table 1, it is still inferior to the performance of [17], which is 58.6 mAP on this benchmark because its point resolution is limited.

To achieve higher performance on ScanNet [5], it is essential to retain point resolution higher than SA3 layer. The SA2 operates on 1024 points, which is the same with VoteNet [17], so we adopt it as the input seeding layer. Similar to our previous conclusions, the performance of Dense Point Diffusion is better than that of the Point Diffusion counterpart. Contrary to the simple scenes in SUN RGB-D [23], the performance still increases with large dilation rate of 12 introduced. Compared with peak performance reported from VoteNet [17], our method exceed it with a margin of **1.0 mAP** in mAP@0.25, and an even larger margin of **2.1 mAP** on a stricter metrics of mAP@0.5. The details can be referred in Table 4 and Table 5.

Model Size and Run-time Analysis Results from Table 6 keep the track of the model size for different approaches. We firstly analyze the cases with SA2 seed layer, which has the same point resolution with VoteNet [17]. The model size of the proposed method with the setting of SA2 - Dense Point Diffusion(3, 6) is 2× smaller than VoteNet. Furthermore, for SA3 layer, even though an extra SA layer is introduced in contrast with the “SA2 - Dense Point Diffusion(3, 6)”, the model size of “SA3 - Dense Point Diffusion(3, 6)” is still much smaller than VoteNet.

Table 6. Model size and averaged inference time of the proposed methods.

Method	Model size	Inference Time	
		SUN RGB-D	ScanNet
F-PointNet[18]	47.0MB	0.09s	-
3D-SIS [8]	19.7MB	-	2.85s
VoteNet (FP2)[17]	11.2MB	0.08s	**0.10s**
SA2 - Dense Point Diffusion (3, 6)	**4.5MB**	0.10s	0.14s
SA3 - Dense Point Diffusion (3, 6)	7.2MB	**0.07s**	**0.10s**

We also investigate the run-time performance of the proposed approaches by measuring the average inference time per scan, which is also adopted by the VoteNet [17]. The inference time of “SA3 - Dense Point Diffusion (3, 6)” is comparable with VoteNet in time consumption (i.e., 0.07s to 0.08s on SUN-RGBD V1 and 0.10s to 0.10s on ScanNet shown in Table 6) while the performance has outperformed the VoteNet with a large margin (i.e., 58.4mAP to 57.7 mAP on SUN RGB-D v1 in Table 1). It should also be noted that it is not a fair comparison for the proposed module because “SA3” seeding layer contains only 512 point, much less than the VoteNet. But the performance of our method is better.

These experiments show that the proposed backbone is advantageous over the conventional PointNet++ backbone in terms of accuracy and model size.

6. Conclusion

In this paper, we resolve the demands of enlarging receptive field and preserving point information for objects in 3D space via the operator of Dilated Point Convolution. Based on this operator, we propose Dense Point Diffusion module that can efficiently extend fields of view and densely sample and aggregate point features from various scales without sacrificing the point resolution, which is critical for 3D detection. The proposed method has been proved to be a compelling alternative to the conventional PointNet++.

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1, 4

[2] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[4] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 2

[5] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 5, 7, 8

[6] F. Engelmann, T. Kontogianni, and B. Leibe. Dilated point convolutions: On the receptive field of point convolutions. *arXiv preprint arXiv:1907.12046*, 2019. 2

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[8] J. Hou, A. Dai, and M. Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019. 8

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[11] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018. 2

[12] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019. 1, 2

[13] G. Li, M. Muller, A. Thabet, and B. Ghanem. Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9267–9276, 2019. 2

[14] M. Liang, B. Yang, S. Wang, and R. Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018. 2

[15] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Pensky. Sparse convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 806–814, 2015. 2

[16] Y. Liu, B. Fan, G. Meng, J. Lu, S. Xiang, and C. Pan. Densepoint: Learning densely contextual representation for efficient point cloud processing. *arXiv preprint arXiv:1909.03669*, 2019. 2

[17] C. R. Qi, O. Litany, K. He, and L. J. Guibas. Deep hough voting for 3d object detection in point clouds. *arXiv preprint arXiv:1904.09664*, 2019. 1, 2, 3, 5, 6, 7, 8

[18] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018. 7, 8

[19] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 1, 2

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 1, 2, 3

[21] S. Shi, X. Wang, and H. Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019. 1, 2, 3

[22] S. Shi, Z. Wang, X. Wang, and H. Li. Part-aˆ 2 net: 3d part-aware and aggregation neural network for object detection from point cloud. *arXiv preprint arXiv:1907.03670*, 2019. 1

[23] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1, 5, 7, 8

[24] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019. 1

[25] Y. Yan, Y. Mao, and B. Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2

[26] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3684–3692, 2018. 1

[27] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia. Std: Sparse-to-dense 3d object detector for point cloud. *arXiv preprint arXiv:1907.10471*, 2019. 1, 2, 3

[28] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 1

[29] Y. Zhou and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018. 1, 2, 6