Going Deeper with Lean Point Networks

Eric-Tuan Le[1]   Iasonas Kokkinos[1,2]   Niloy J. Mitra[1,3]

[1]University College London   [2]Ariel AI   [3]Adobe Research

Abstract

In this work we introduce Lean Point Networks (LPNs) to train deeper and more accurate point processing networks by relying on three novel point processing blocks that improve memory consumption, inference time, and accuracy: a convolution-type block for point sets that blends neighborhood information in a memory-efficient manner; a crosslink block that efficiently shares information across low- and high-resolution processing branches; and a multi-resolution point cloud processing block for faster diffusion of information. By combining these blocks, we design wider and deeper point-based architectures. We report systematic accuracy and memory consumption improvements on multiple publicly available segmentation tasks by using our generic modules as drop-in replacements for the blocks of multiple architectures (PointNet++, DGCNN, SpiderNet, PointCNN). Code and pretrained models are publicly available at geometry.cs.ucl.ac.uk/projects/2020/deepleanpn/.

1. Introduction

Geometry processing has started profiting from applying deep learning to graphics and 3D shape analysis [25, 33, 6, 3], delivering networks that guarantee desirable properties of point cloud processing, such as permutation-invariance and quantization-free representation [28, 31, 32]. Despite these advances, the memory requirements of a majority of the point processing architectures still impede breakthroughs similar to those made in computer vision.

Directly working with unstructured representations of 3D data (i.e., not residing on a spatial grid), necessitates re-inventing, for geometry processing, the functionality of basic image processing blocks, such as convolution operations, information exchange pathways, and multi-resolution data flow. Taking a Lagrangian perspective, in order to avoid quantization or multi-view methods, the pioneering PointNet architecture [24] demonstrated how to directly work on point cloud datasets by first lifting individual points to higher dimensional features (via a shared MLP) and then performing permutation-invariant local pooling. Point-

Net++ [25], by considering local point patches, groups information within a neighborhood based on Euclidean distance and then applies PointNet to the individual groups. This design choice requires explicitly duplicating local memory information among neighboring points, and potentially compromises performance as the networks go deeper.
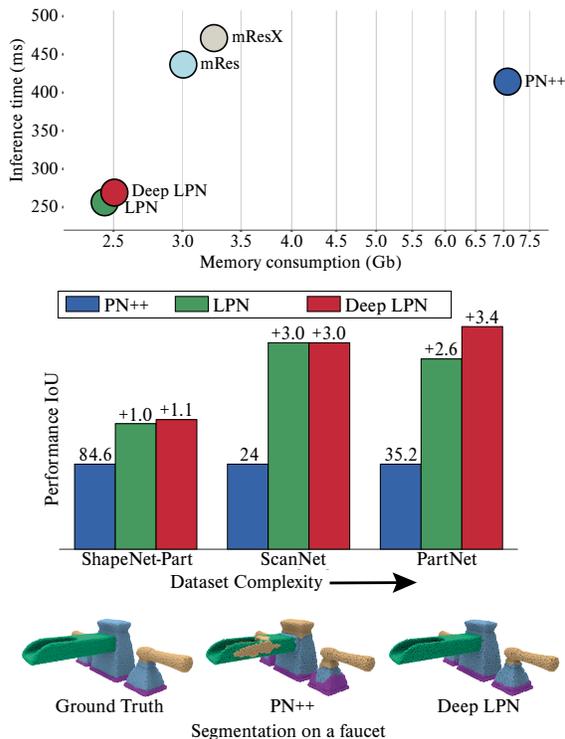

Figure 1. Lean Point Networks (LPNs) can achieve higher point cloud segmentation accuracy while operating at substantially lower memory and inference time. (Top) Memory footprint and inference speed of LPN variants introduced in this work compared to the PointNet++ (PN++) baseline. (Middle) Improvements in accuracy for three segmentation benchmarks of increasing complexity. On the –most complex– PartNet dataset our deep network outperforms the shallow PointNet++ baseline by 3.4%, yielding a 9.7% relative increase. (Bottom) Part Segmentation by PointNet++ and Deep LPN.

In particular, the memory and computational demands of point network blocks (e.g., PointNet++ and many follow-up architectures) can affect both training speed and, more crucially, inference time. One of the main bottlenecks for such point networks is their memory-intensive nature, as detailed in Sec. 3.1. Specifically, the PointNet++ architecture and its variants replicate point neighborhood information, letting every node carry in its feature vector information about all of its neighborhood. This results in significant memory overhead, and limits the number of layers, features and feature compositions one can compute.

In this work, we enhance such point processing networks that replicate local neighborhood information by introducing a set of modules that improve memory footprint and accuracy, without compromising on inference speed. We call the resulting architectures *Lean Point Networks*, to highlight their lightweight memory budget. We build on the decreased memory budget to go deeper with point networks. As has been witnessed repeatedly in the image domain [12, 14, 39], we show that going deep also increases the prediction accuracy of point networks.

We start in Sec. 3.2 by replacing the grouping operation used in point cloud processing networks with a low-memory alternative that is the point cloud processing counterpart of efficient image processing implementations of convolution. The resulting *'point convolution block'* – defined slightly differently from convolution used in classical signal processing – is 67% more memory-efficient and 41% faster than its PointNet++ counterpart, while exhibiting favorable training properties due to more effective mixing of information across neighborhoods.

We then turn in Sec. 3.3 to improving the information flow across layers and scales within point networks through three techniques: a *multi-resolution* variant for multi-scale network which still delivers the multi-scale context but at a reduced memory and computational cost, *residual links*, and a new *cross-link block* that broadcasts multi-scale information across the network branches. By combining these blocks, we are able to successfully train deeper point networks that allow us to leverage upon larger datasets.

In Sec. 4 we validate our method on the ShapeNet-Part, ScanNet and PartNet segmentation benchmarks, reporting systematic improvements over the PointNet++ baseline. As shown in Fig. 1, when combined these contributions deliver multifold reductions in memory consumption while improving performance, allowing us in a second stage to train increasingly wide and deep networks. On PartNet, the most complex dataset, our deep architecture achieves a 9.7% relative increase in IoU while decreasing memory footprint by 57% and inference time by 47%.

Having ablated our design choices on the PointNet++ baseline, in Sec. 4.3 we turn to confirming the generic nature of our blocks. We extend the scope of our experiments to three additional networks, (i) DGCNN [33], (ii) Spider-CNN [35], and (iii) PointCNN [21] and report systematic improvements in memory efficiency and performance.

2. Related Work

Learning in Point Clouds. Learning-based approaches have recently attracted significant attention in the context of Geometric Data Analysis, with several methods proposed specifically to handle point cloud data, including Point-Net [24] and several extensions such as PointNet++ [25] and Dynamic Graph CNNs [33] for shape segmentation and classification, PCPNet [10] for normal and curvature estimation, P2P-Net [36] and PU-Net [38] for cross-domain point cloud transformation. Alternatively, kernel-based methods [2, 13, 22, 30] have also been proposed with impressive performance results. Although many alternatives to PointNet have been proposed [27, 20, 21, 13, 40] to achieve higher performance, the simplicity and effectiveness of PointNet and its extension PointNet++ make it popular for many other tasks [37].

Taking PointNet++ as our starting point, our work facilitates the transfer of network design techniques developed in computer vision to point cloud processing. In particular, significant accuracy improvements have been obtained with respect to the original AlexNet network [18] by engineering the scale of the filtering operations [41, 26], the structure of the computational blocks [29, 34], and the network's width and depth [12, 39]. A catalyst for experimenting with a larger space of network architecture, however, is the reduction of memory consumption - this motivated us to design lean alternatives to point processing networks. Notably, [42] introduce a new operator to improve point cloud network efficiency, but only focus on increasing the convergence speed by tuning the receptive field. [19] has investigated how residual/dense connections and dilated convolution could help mitigate vanishing gradient observed for deep graph convolution networks but without solving memory limitations, [13] use Monte Carlo estimators to estimate local convolution kernels. By contrast our work explicitly tackles the memory problem with the objective of training deeper/wider networks and shows that there are clear improvements over strong baselines.

Memory-Efficient Networks. The memory complexity of the standard back-propagation implementation grows linearly with network's depth as backprop requires retaining in memory all of the intermediate activations computed during the forward pass, since they are required for the gradient computation in the backward pass.

Several methods bypass this problem by trading off speed with memory. Checkpointing techniques [5, 9] use anchor points to free up intermediate computation results, and re-compute them in the backward pass. This is 1.5x slower during training, since one performs effectively two

forward passes rather than just one. More importantly, applying this technique is easy for chain-structured graphs, e.g., recursive networks [9] but is not as easy for general Directed Acyclic Graphs, such as U-Nets, or multi-scale networks like PointNet++. One needs to manually identify the graph components, making it cumbersome to experiment with diverse variations of architectures.

Reversible Residual Networks (RevNets) [8] limit the computational block to come in a particular, invertible form of residual network. This is also 1.5x slower during training, but alleviates the need for anchor points altogether. Unfortunately, it is unclear what is the point cloud counterpart of invertible blocks.

We propose generic blocks to reduce the memory footprint inspired from multi-resolution processing and efficient implementations of the convolution operation in computer vision. As we show in Sec. 4.3, our blocks can be used as drop-in replacements in generic point processing architectures (PointNet++, DGCNN, SpiderNet, PointCNN) without any additional network design effort.

3. Method

We start with a brief introduction of the PointNet++ network, which serves as an example point network baseline. We then introduce our modules and explain how they decrease memory footprint and improve information flow.

3.1. PointNet and PointNet++ Architectures

PointNet++ [25] has been proposed as a method of augmenting the basic PointNet architecture with a grouping operation. As shown in Fig. 4(a), grouping is implemented through a 'neighborhood lookup', where each point $\mathbf{p}_i$ looks up its k-nearest neighbors and stacks them to get a point set, say $P_{N_k}^i$. If each point comes with a D-dimensional feature vector, the result of this process is a tensor $T = \begin{bmatrix} \mathbf{v} & \mathbf{v}_{[.,1]} & \ldots & \mathbf{v}_{[.,K]} \end{bmatrix}$ of size $N \times D \times (K+1)$. Within the PointNet modules, every vector of this matrix is processed separately by a Multi-Layer-Perceptron that implements a function $\mathrm{MLP} : R^D \to R^{D'}$, while at a later point a max-pooling operation over the K neighbors of every point delivers a slim, $N \times D'$ matrix.

When training a network every layer constructs and retains a matrix like T in memory, so that it can be used in the backward pass to update the MLP parameters, and send gradients to earlier layers. While demonstrated to be extremely effective, this design of PointNet++ has two main shortcomings: first, because of explicitly carrying around k-nearest neighbor information for each point, the network layers are memory intensive; and second, being reliant on PointNet, it also delays transmission of global information until the last, max-pooling stage where the results of decoupled MLPs are combined. Many subsequent variants of PointNet++ suffer

from similar memory and information flow limitations. As we describe now, these shortcomings are alleviated, by our convolution-type point processing layer.

3.2. convPN: a convolution-type PointNet layer

We propose a novel convolution-type PointNet layer (convPN), that is inspired from efficient implementations of convolution. Standard convolution operates in two steps: (i) neighborhood exposure and (ii) matrix multiplication. Our convPN block follows similar steps, as shown in Fig. 2, but the weight matrix blocks treating different neighbors are tied, so as to ensure permutation invariance.

In more detail, standard 2D image convolution amounts to forming a K^2 tensor in memory when performing $K \times K$ filtering and then implementing a convolution as matrix multiplication. This amounts to the `im2col` operation used to implement convolutions with General Matrix-Matrix Multiplication (GEMM) [15]. In point clouds the nearest neighbor information provides us with the counterpart to the $K \times K$ neighborhood.

Based on this observation we propose to use the strategy used in memory-efficient implementations of image convolutions for deep learning: we free the memory from a layer as soon as the forward pass computes its output, rather than maintaining the matrix in memory. In the backward pass we reconstruct the matrix *on the fly* from the outputs of the previous layer. We perform the required gradient computations and then return the GPU memory resources.

As shown in Fig. 3, we gain further efficiency by replacing the MLPs of PointNet++ by a sequence of SLP-Pooling modules. This allows us to further reduce memory consumption, saving the layer activations only through the pooled features while at the same time increasing the frequency at which neighbors share information.

As detailed in the Supplemental Material, a careful implementation of our convolution-type architecture shortens, on average, the time spent for the forward pass and the backward pass by 41% and 68%, respectively, while resulting in a drastic reduction in memory consumption.

For a network with L layers, the memory consumption of the baseline PointNet++ layer grows as $L \times (N \times D \times K)$, while in our case memory consumption grows as $L \times (N \times D) + (N \times D \times K)$, where the term, $L \times (N \times D)$ accounts for the memory required to store the layer activations, while the second term $N \times D \times K$ is the per-layer memory consumption of a single neighborhood convolution layer. As L grows larger, this results in a K-fold drop, shown on Fig. 5. This reduction opens up the possibility of learning much deeper networks, since memory demands now grow substantially more slowly in depth. With minor, dataset-dependent, fluctuations, the memory footprint of our convolution type architecture is on average 67% lower than the PointNet++ baseline, while doubling the number of layers
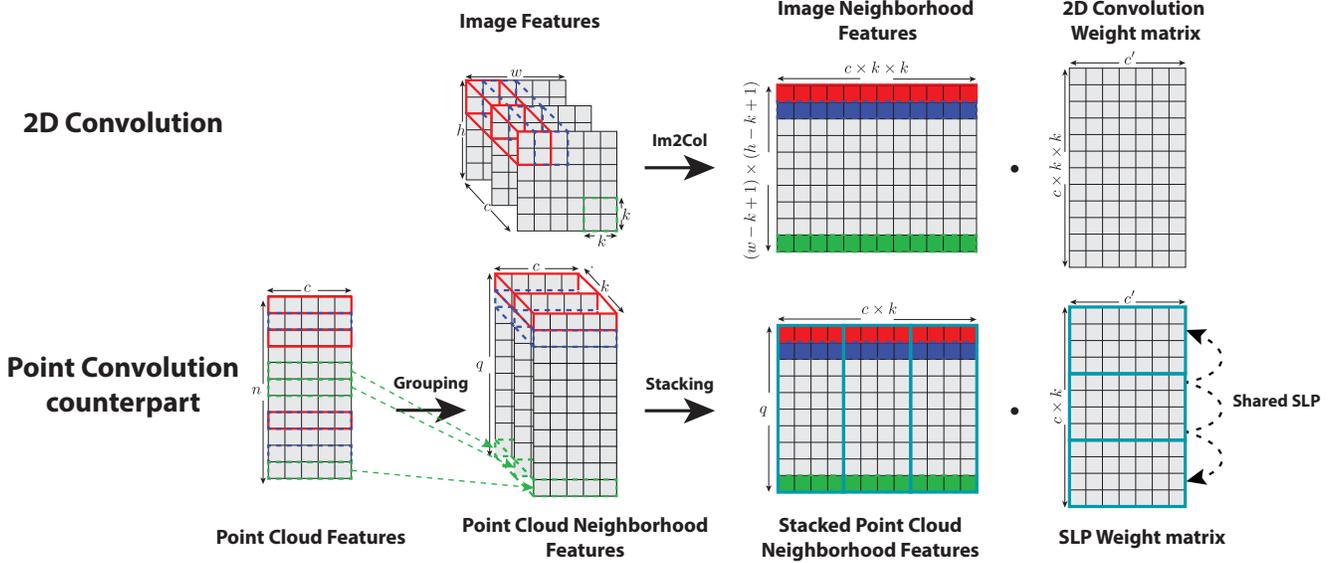
Figure 2. Analogy between 2D convolution and its point cloud counterpart. In both cases, the layer operates in two steps: (i) neighborhood exposure and (ii) matrix multiplication. 2D image convolution amounts to forming a K^2 tensor in memory when performing $K \times K$ filtering and then implementing a convolution as matrix multiplication.
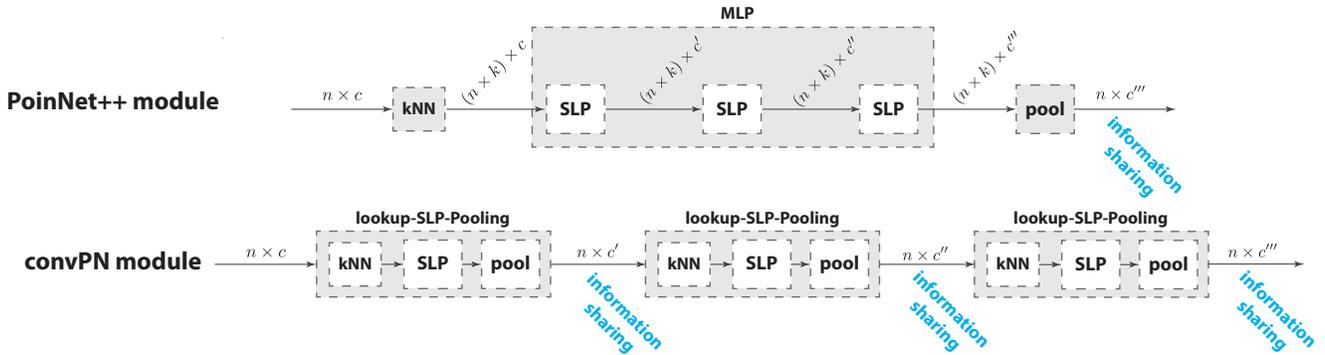

Figure 3. Comparison of a PointNet++ module with our convPN module. The convPN module replaces the MLP with its pooling layer by a sequence of SLP-Pooling modules which has two benefits (i) memory savings as the layer activations are saved only through the pooled features and (ii) better information flow as it increases the frequency at which neighbors share information.

comes with a memory overhead of 2.7%.

3.3. Improving Information Flow

We now turn to methods for efficient information propagation through point networks. As has been repeatedly shown in computer vision, this can drastically impact the behavior of the network during training. Our experiments indicate that this is also the case for point processing.

(a) Multi-Resolution vs Multi-Scale Processing.

Shape features can benefit from both local, fine-grained information and global, semantic-level context; their fusion can easily boost the discriminative power of the resulting features. For this we propose to extract neighborhoods of fixed size in downsampled versions of the original point cloud. In the coordinates of the original point cloud this amounts to increasing the effective grouping area, but it now comes with a much smaller memory budget. We observe a 58% decrease in memory footprint on average on the three tested datasets. Please refer to in Supplemental for an illustration of the difference between both types of processing.

(b) Residual Links.
We use the standard Residual Network architecture [12], which helps to train deep networks reliably. Residual networks change the network's connectivity to improve gradient flow during training: identity connections provide early network layers with access to undistorted versions of the loss gradient, effectively mitigating the vanishing gradient problem. As our results in Sec. 4 show, this allows us to train deeper networks.

Network Layers

(a) PointNet++ (PN++)

(b) mRes

(c) mResX

**(d) convPN**

(e) Stack of convPN layers
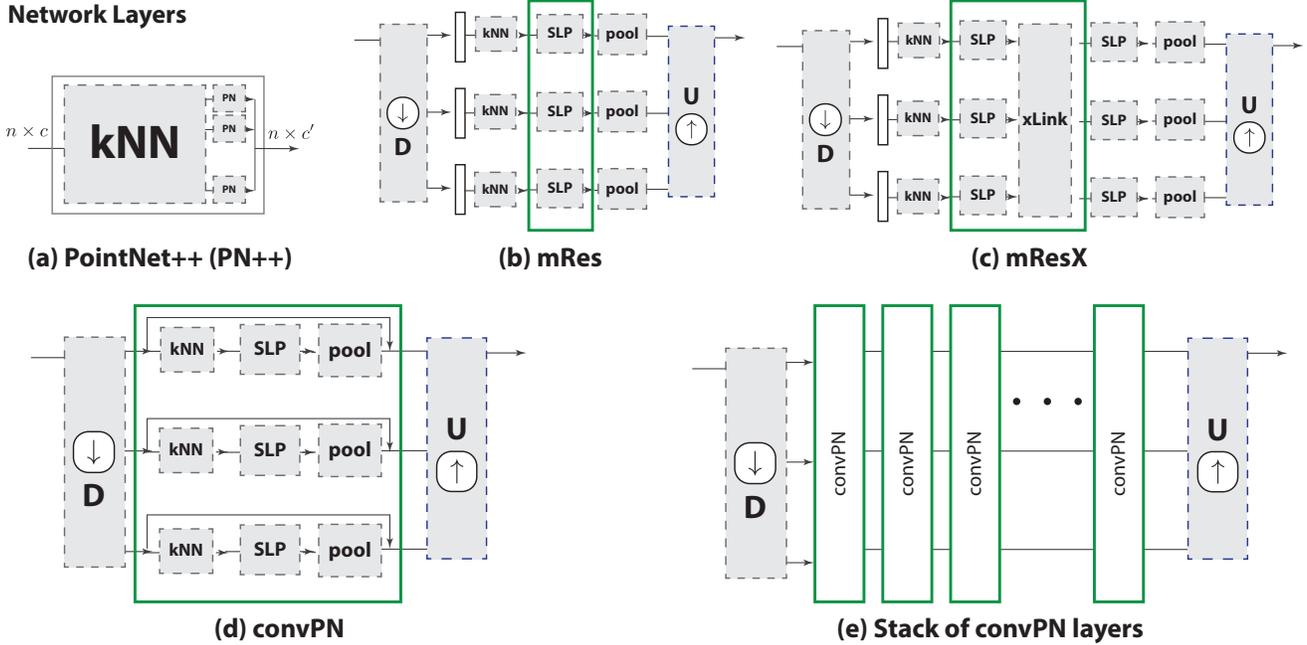
Figure 4. The standard PN++ layer in (a) amounts to the composition of a k-Nearest Neighbor (kNN)-based lookup and a PointNet element. In (b) we propose to combine parallel PointNet++ blocks in a multi-resolution architecture (D and U stand for down- and up-sampling operations), using multiple Single Layer Perceptrons (SLPs) and in (c) allow information to flow across branches of different resolutions through a cross-link element ('xLink'). In (d) we propose to turn the lookup-SLP-pooling cascade into a low-memory counterpart by removing the kNN elements from memory once computed; we also introduce residual links, improving the gradient flow. In (e) we stack the green box in (d) to grow in depth and build our deep architecture.
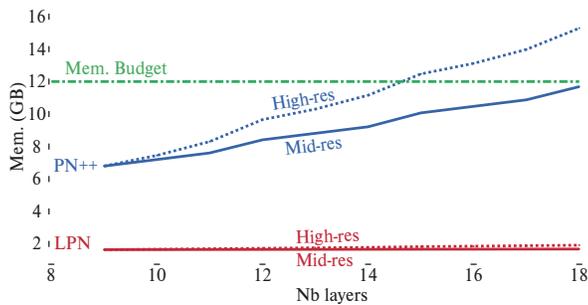

Figure 5. Evolution of memory consumption as the number of layers increases for PointNet++ and LPN (convolution block counterpart) on ShapeNet-Part. Doubling the number of layers for LPN results only in an increase in memory by +2.3% and +16.8% for mid- and high- resolution respectively, which favorably compares to the +72% and +125% increases for PointNet++.

(c) Cross Links. Further, we use Cross-Resolution Links to better propagate information in the network during training. We draw inspiration from the Multi-Grid Networks [16], Multiresolution Tree Networks [7], Hypercolumns [11]; and allow layers that reside in different resolution branches to communicate with each other, thereby exchanging low-, mid-, and high-resolution infor-
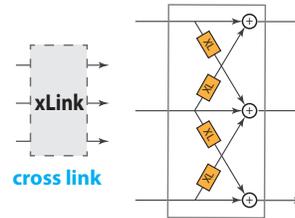
Figure 6. Cross-link module to connect across resolutions.

mation throughout the network processing, rather than fusing multi-resolution information at the end of each block.

Cross-links broadcast information across resolutions as shown in Fig. 6. Note that unlike [7], an MLP transforms the output of one branch to the right output dimensionality so that it can be combined with the output of another branch. Each resolution branch can focus on its own representation and the MLPs will be in charge of making the translation between them. Taking in particular the case of a high-resolution branch communicating its outputs to a mid-resolution branch, we have $N \times D^H$ feature vectors at the output of a lookup-SLP-pooling block cascade, which need to be communicated to the $N/2 \times D^M$ vectors of the mid-resolution branch. We first downsample the points, going

from N to $N/2$ points, and then use an MLP that transforms the vectors to the target dimensionality. Conversely, when going from low- to higher dimensions we first transform the points to the right dimensionality and then upsample them. We have experimented with both concatenating and summing multi-resolution features and have observed that summation behaves systematically better in terms of both training speed and test performance.

4. Evaluation

We start by defining our tasks and metrics and then turn to validating our two main contributions to model accuracy, namely better network training through improved network flow in Sec. 4.1, and going deeper through memory-efficient processing in Sec. 4.2. We then turn to validating the merit of our modules when used in tandem with a broad array of state-of-the-art architectures in Sec. 4.3, and finally provide a thorough ablation of the impact of our contributions on aspects complementary to accuracy, namely parameter size, memory, and efficiency in Sec. 4.4.

Dataset and evaluation measures. Our modules can easily be applied to any point cloud related tasks, such as classification, however, we focus here on evaluating our modules on the point cloud segmentation task on three different datasets as it is a more challenging task. The datasets consist of either 3D CAD models or real-world scans. We quantify the complexity of each dataset based on (i) the number of training samples, (ii) the homogeneity of the samples and (iii) the granularity of the segmentation task. Note that a network trained on a bigger and diverse dataset would be less prone to overfitting - as such we can draw more informative conclusions from more complex datasets. We order the datasets by increasing complexity: ShapeNet-Part [4], ScanNet [6] and PartNet [23] for fine-grained segmentation. By its size (24,506 samples) and its granularity (251 labeled parts), PartNet is the most complex dataset.

To have a fair comparison (memory, speed, accuracy), we re-implemented all the models in Pytorch and consistently compared the vanilla network architectures and our memory-efficient version. We report the performance of networks using their last saved checkpoint (i.e. when training converges), instead of the common (but clearly flawed) practice of using the checkpoint that yields best performance on the test set. These two factors can lead to small differences from the originally reported performances.

We use two different metrics to report the Intersection over Union (IoU): (i) the mean Intersection over Union (mIoU) and (ii) the part Intersection over Union (pIoU). Please refer to Supplemental for further details.

4.1. Effect of improved information flow

We report the performance of our variations for Point-Net++ on the Shapenet-Part, ScanNet and PartNet datasets

Table 1. Performance of our modules compared to PointNet++ baseline. The impact of our modules becomes most prominent as the dataset complexity grows. On PartNet our Deep LPN network increases pIoU by 9.7% over PointNet++, outperforming its shallow counterpart by +2.1%.

	ShapeNet-Part (13,998 samp.)	ScanNet (1,201 samp.)		PartNet (17,119 samp.)
	mIoU (%)	Vox. Acc. (%)	pIoU (%)	pIoU (%)
PN++	84.60 (+0.0%)	80.5 (+0.0%)	24 (+0.0%)	35.2 (+0.0%)
mRes	85.47 (+1.0%)	79.4 (-1.4%)	22 (-8.3%)	37.2 (+5.7%)
mResX	85.42 (+1.0%)	79.5 (-1.2%)	22 (-8.3%)	37.5 (+6.5%)
LPN	85.65 (+1.2%)	**83.2 (+3.4%)**	**27 (+12.5%)**	37.8 (+7.4%)
Deep LPN	**85.66 (+1.3%)**	82.2 (+2.1%)	**27 (+12.5%)**	**38.6 (+9.7%)**

(Table 1). Our lean and deep architectures can be easily deployed on large and complex datasets. Hence, for Part-Net, we choose to train on the full dataset all at once on a segmentation task across the 17 classes instead of having to train a separate network for each category as in [23].

Our architectures substantially improve the memory efficiency of the PointNet++ baseline while also delivering an increase in performance for more complex datasets (see Fig. 1). Indeed, as the data complexity grows, having efficient information flow has a larger influence on the network performance. On PartNet, the spread between our architectures and the vanilla PointNet++ becomes significantly high: our multiresolution (mRes) network increases relative performance by +5.7% over PointNet++ and this gain reaches +6.5% with cross-links (mResX). Our convolution-type network (LPN) outperforms other architectures when dataset complexity increases (+3.4% on ScanNet and +7.4% on PartNet) by more efficiently mixing information across neighbours.

4.2. Improvement of accuracy by going deeper

The memory savings introduced in Sec. 3.2 give the opportunity to design deeper networks. Naively increasing network depth can harm performance [12]. Instead, we use residual connections to improve convergence for our deep network. The architecture, detailed in the Supplemental material, consists in doubling the number of layers in the encoding part. While keeping the impact on efficiency very small (+6.3% on inference time on average and +3.6% on memory consumption at most compared to the shallow LPN), the performance improved (see Table 1). On PartNet, this margin reaches +2.1% over the shallow LPN and +9.7% over the vanilla PointNet++. Note the low growth of memory as a function of depth, shown in Fig. 5. As shown in Fig. 7, having a deeper network improves the segmentation quality at the boundaries between parts. In contrast, naively increasing the number of encoding layers from 9 to 15 in PointNet++ leads only to a small increase in performance IoU from 84.60% to 84.66%.

In Table 2 we compare against Deep GCN [19] in terms

Table 2. Performance of our deepConPN network compared to Deep GCN (ResGCN-28) and related methods on S3DIS based on a 6-fold validation process. The difference in performance observed on each class can be explained by the different approaches networks have for point convolution. Our deep network clearly outperforms PointNet++ baseline by a spread of +6.8% for mIoU. We achieve similar performance compared to Deep GCN while relying on a weaker baseline (PointNet++ against DGCNN)

Method	OA	mIOU	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
MS+CU	79.2	47.8	88.6	**95.8**	67.3	36.9	24.9	48.6	52.3	51.9	45.1	10.6	36.8	24.7	37.5
G+RCU	81.1	49.7	90.3	92.1	67.9	44.7	24.2	52.3	51.2	58.1	47.4	6.9	39.0	30.0	41.9
3DRNN+CF	**86.9**	56.3	92.9	93.8	73.1	42.5	25.9	47.6	59.2	60.4	66.7	24.8	**57.0**	36.7	51.6
DGCNN	84.1	56.1	-	-	-	-	-	-	-	-	-	-	-	-	-
ResGCN-28	85.9	**60.0**	**93.1**	95.3	**78.2**	33.9	**37.4**	**56.1**	**68.2**	64.9	61.0	**34.6**	51.5	51.1	54.4
PointNet	78.5	47.6	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
PointNet++	-	53.2	90.2	91.7	73.1	42.7	21.2	49.7	42.3	62.7	59.0	19.6	45.8	48.2	45.6
Deep LPN	85.7	**60.0**	91.0	95.6	76.1	**50.3**	25.9	55.1	56.8	**66.3**	**74.3**	25.8	54.0	**52.3**	**55.3**
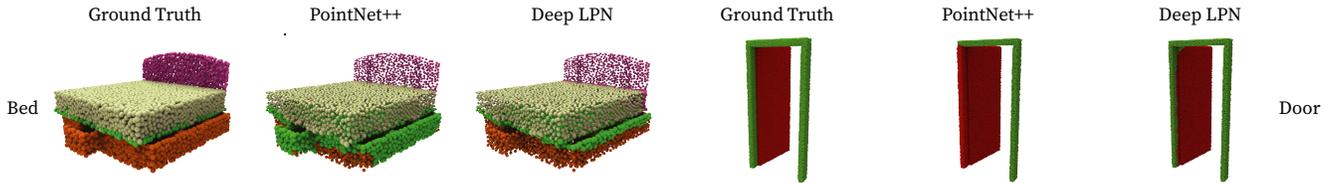
Figure 7. Segmentation prediction for both PointNet++ and Deep LPN networks compared to the ground truth. While PointNet++ struggles to detect accurately the boundaries between different parts, ours performs a much finer segmentation in those frontier areas.

of the overall accuracy and the mean IoU on the S3DIS dataset [1] by following the same 6-fold evaluation process. We attain similar performance to Deep GCN, while relying on our generic memory-efficient network blocks and while based on a weaker baseline compared to Deep GCN (i.e., DGCNN). Moreover, as Deep GCN is not designed to tackle the efficiency issue in Point Networks, our network wins on all counts and successfully reduces both the memory (more than 74%) and the speed (-48% and -89% for the inference and the backward speed respectively). As we show in the following section, these blocks come with the advantage of being applicable to many other point processing networks.

4.3. Evaluation on more architectures

We have introduced building blocks for point processing networks based on two key ideas, (i) a memory efficient convolution and (ii) improved information flow. Our blocks make it really efficient to capture, process and diffuse information in a point neighbourhood. Diffusing information across neighborhood is the main behavior that most networks, if not all, share. We validate the generality of the proposed modular blocks in the context of other point-based learning setups, as shown in Table 3. Each of our macro-blocks can be stacked together, extended into a deeper block by duplicating the green boxes (see Figure 4) or even be modified by changing one of its components by another. We test our framework on three additional networks among the recent approaches, (i) Dynamic Graph CNN [33], (ii) PointCNN [21] and (iii) SpiderCNN [35]. These networks involve a diverse set of point convolution approaches; these

experiments allow us to assess the generic nature of our modular blocks, and their value as drop-in replacements for existing layers.

All three of the networks make extensive use of memory which is a bottleneck to depth. We implant our modules directly in the original networks, making, when needed, some approximations from the initial architecture (see Supplemental). We report the performance of each network with our lean counterpart on two metrics: (i) memory footprint and (ii) accuracy in Table 3. Our lean counterparts consistently improve both the accuracy (from +0.4% up to +8.0%) and the memory consumption (from -19% up to -69%).

Our modular blocks can thus be applied to a wide range of state-of-the-art networks and improve significantly their memory consumption while having a positive impact on performance.

4.4. Ablation study

In this section we report our extensive experiments to assess the importance of each block of our network architectures. Our lean structure allows us to adjust the network architectures by increasing its complexity, either by (i) adding extra connections or by (ii) increasing the depth. We analyze our networks along four axes: (i) the performance (IoU or accuracy) (Table 1), (ii) the memory footprint, (iii) the inference time and (iv) the backward time. Our main experimental findings regarding network efficiency are reported in Table 4 and ablate the impact of our proposed design choices for point processing networks.

Memory-efficient Convolutions: As described in

Table 3. Performance of our blocks on three different architectures (DGCNN, PointCNN and SpiderCNN) on three datasets using two different metrics: (i) memory consumption in Gb and (ii) performance in % (mIoU for ShapeNet-Part, Vox. Acc. for ScanNet and pIoU for PartNet). Our lean counterparts improve significantly both the performance (up to +8.0%) and the memory consumption (up to -69%).

		DGCNN			PointCNN			SCNN		
		ShapeNet-P	ScanNet	PartNet	ShapeNet-P	ScanNet	PartNet	ShapeNet-P	ScanNet	PartNet
Mem.	Vanilla	2.62 (+0%)	7.03 (+0%)	9.50 (+0%)	4.54 (+0%)	5.18 (+0%)	6.83 (+0%)	1.09 (+0%)	4.33 (+0%)	5.21 (+0%)
	Lean	**0.81 (-69%)**	**3.99 (-43%)**	**5.77 (-39%)**	**1.98 (-56%)**	**3.93 (-24%)**	**5.55 (-19%)**	**0.79 (-28%)**	**3.25 (-25%)**	**3.33 (-36%)**
Perf.	Vanilla	82.59 (+0.0%)	74.5 (+0.0%)	20.5 (+0.0%)	83.60 (+0.0%)	77.2 (+0.0%)	25.0 (+0.0%)	79.86 (+0.0%)	72.9 (+0.0%)	17.9 (+0.0%)
	Lean	**83.32 (+0.9%)**	**75.0 (+0.7%)**	**21.9 (+6.8%)**	**84.45 (+1.0%)**	**80.1 (+3.8%)**	**27.0 (+8.0%)**	**81.61 (+2.2%)**	**73.2 (+0.4%)**	**18.4 (+2.8%)**

Table 4. Efficiency of our network architectures measured with a batch size of 8 samples on a Nvidia GTX 2080Ti GPU. All of our lean architectures allow to save a substantial amount of memory on GPU wrt. the PointNet++ baseline from 58% with mRes to a 67% decrease with LPN. This latter convolution-type architecture wins on all counts, decreasing both inference time (-41%) and the length of backward pass (-68%) by a large spread. Starting from this architecture, the marginal cost of going deep is extremely low: doubling the number of layers in the encoding part of the network increases inference time by 6.3% on average and the memory consumption by only 3.6% at most compared to LPN. Please refer to Supplemental for absolute values.

	Parameters (M)			Memory Footprint (Gb)			Inference Time (ms)			Length Backward pass (ms)		
	ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet
PointNet++	+0.0%	+0.0%	+0.0%	+0.0%	+0.0%	+0.0%	+0.0%	+0.0%	+0.0%	+0.0%	+0.0%	+0.0%
mRes	-17.0%	-17.6%	-16.6%	-69.3%	-56.5%	-47.6%	+14.8%	+59.2%	-19.4%	-68.8%	**-53.8%**	-63.2%
mResX	-10.6%	-10.7%	-10.1%	-65.0%	-53.2%	-46.3%	+28.2%	+60.9%	-12.5%	-29.5%	+0.0%	-25.4%
LPN	+13.8%	+13.4%	+12.6%	-75.7%	**-66.6%**	**-57.9%**	**-45.6%**	**-30.3%**	**-47.9%**	**-82.7%**	-42.3%	**-78.9%**
Deep LPN	**+54.3%**	**+54.0%**	**+50.8%**	**-79.1%**	-65.4%	-57.0%	-40.4%	-25.6%	-46.5%	-78.6%	-11.5%	-72.4%

Sec. 3.2, our leanest architeture is equivalent to constraining each PointNet unit to be composed of a single layer network, and turning its operation into a memory-efficient block by removing intermediate activations from memory. In order to get a network of similar size, multiple such units are stacked to reach the same number of layers as the original network. Our convolution-type network wins on all counts, both on performance and efficiency. Indeed, the IoU is increased by 3.4% on ScanNet and 7.4% on PartNet compared to PointNet++ baseline. Regarding its efficiency, the memory footprint is decreased by 67% on average while decreasing both inference time (-41%) and the length of the backward pass (-68%). These improvements in speed can be seen as the consequence of processing most computations on flattened tensors and thus reducing drastically the complexity compared to PointNet++ baseline.

Multi-Resolution: Processing different resolutions at the same stage of a network has been shown to perform well in shallow networks. Indeed, mixing information at different resolutions helps to capture complex features early in the network. We adopt that approach to design our mRes architecture. Switching from a PointNet++ architecture to a multi-resolution setting increases the IoU by 1.0% on ShapeNet-Part and 5.7% on PartNet. More crucially, this increase in performance come with more efficiency. Although the inference time is longer (18% longer on average) due to the extra downsampling and upsampling operations, the architecture is much leaner and reduces memory footprint by 58%. Training is quicker though due to a 62% faster backward pass.

Cross-links: Information streams at different resolutions are processed separately and can be seen as complementary. To leverage this synergy, the network is provided with additional links connecting neighborhood resolutions. We experiment on the impact of those cross-resolution links to check their effect on the optimization. At the price of a small impact on memory efficiency (+8% wrt. mRes) and speed (+7% on inference time wrt. mRes), the performance can be improved on PartNet, the most complex dataset, with these extra-links by 0.8%.

5. Conclusion

We have introduced new generic building blocks for point processing networks, that exhibit favorable memory, computation, and optimization properties when compared to the current counterparts of state-of-the-art point processing networks. Based on PointNet++, our lean architecture LPN wins on all counts, memory efficiency (-67% wrt. PointNet++) and speed (-41% and -68% on inference time and length of backward pass). Its deep counterpart has a marginal cost in terms of efficiency and achieves the best IoU on PartNet (+9.7% over PointNet++). Those generic blocks exhibit similar performance on all of the additionally tested architectures producing significantly leaner networks (up to -69%) and increase in IoU (up to +8.0%). Based on our experiments, we anticipate that adding these components to the armament of the deep geometry processing community will allow researchers to train the next generation of point processing networks by leveraging upon the advent of larger shape datasets [23, 17].

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[2] Matan Atzmon, Haggai Maron, and Yaron Lipman. Point convolutional neural networks by extension operators. *ACM Trans. Graph.*, 37(4):71:1–71:12, July 2018.

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nie√üner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. 09 2017.

[4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[5] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *CoRR*, abs/1604.06174, 2016.

[6] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[7] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018.

[8] Aidan N. Gomez, Mengye Ren, Raquel Urtasun, and Roger B. Grosse. The reversible residual network: Backpropagation without storing activations. *CoRR*, abs/1707.04585, 2017.

[9] Audrunas Gruslys, Remi Munos, Ivo Danihelka, Marc Lanctot, and Alex Graves. Memory-efficient backpropagation through time. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4125–4133. Curran Associates, Inc., 2016.

[10] Paul Guerrero, Yanir Kleiman, Maks Ovsjanikov, and Niloy J. Mitra. PCPNet: Learning local shape properties from raw point clouds. *CGF*, 37(2):75–85, 2018.

[11] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. *CoRR*, abs/1411.5752, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016.

[13] P. Hermosilla, T. Ritschel, P-P Vazquez, A. Vinacua, and T. Ropinski. Monte carlo convolution for learning on non-uniformly sampled point clouds. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2018)*, 37(6), 2018.

[14] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[15] Yangqing Jia. *Learning Semantic Image Representations at a Large Scale*. PhD thesis, University of California, Berkeley, USA, 2014.

[16] Tsung-Wei Ke, Michael Maire, and Stella X. Yu. Neural multigrid. *CoRR*, abs/1611.07661, 2016.

[17] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. ABC: A big CAD model dataset for geometric deep learning. *CoRR*, abs/1812.06216, 2018.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2013.

[19] Guohao Li, Matthias Müller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns?, 2019.

[20] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. pages 9397–9406, 2018.

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

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

[23] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. *CoRR*, abs/1812.02713, 2018.

[24] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, 1(2):4, 2017.

[25] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5099–5108, 2017.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[27] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. pages 2530–2539, 2018.

[28] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. ICCV*, 2015.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[30] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *CoRR*, 2019.

[31] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-CNN: Octree-based Convolutional Neu-

ral Networks for 3D Shape Analysis. *ACM Transactions on Graphics (SIGGRAPH)*, 36(4), 2017.

[32] Peng-Shuai Wang, Chun-Yu Sun, Yang Liu, and Xin Tong. Adaptive O-CNN: A Patch-based Deep Representation of 3D Shapes. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 37(6), 2018.

[33] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *arXiv preprint arXiv:1801.07829*, 2018.

[34] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016.

[35] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018.

[36] Kangxue Yin, Hui Huang, Daniel Cohen-Or, and Hao Zhang. P2p-net: Bidirectional point displacement net for shape transform. *ACM TOG*, 37(4):152:1–152:13, July 2018.

[37] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Ec-net: an edge-aware point set consolidation network. pages 386–402, 2018.

[38] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *CVPR*, 2018.

[39] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[40] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. 2017.

[41] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. 2014.

[42] Zhiyuan Zhang, Binh-Son Hua, and Sai-Kit Yeung. Shell-net: Efficient point cloud convolutional neural networks using concentric shells statistics. In *International Conference on Computer Vision (ICCV)*, 2019.

Supplementary Material - Going Deeper with Lean Point Networks

Eric-Tuan Le[1]   Iasonas Kokkinos[1,2]   Niloy J. Mitra[1,3]

[1]University College London   [2]Ariel AI   [3]Adobe Research

1. Details on evaluation results

1.1. Datasets

We evaluate our networks on the point cloud segmentation task on three different datasets, ordered by increasing complexity:

- ShapeNet-Part [1]: CAD models of 16 different object categories composed of 50 labeled parts. The dataset provides 13, 998 samples for training and 2, 874 samples for evaluation. Point segmentation performance is assessed using the mean point Intersection over Union (mIoU).

- ScanNet [2]: Scans of real 3D scenes (scanned and reconstructed indoor scenes) composed of 21 semantic parts. The dataset provides 1, 201 samples for training and 312 samples for evaluation. We follow the same protocol as in [5] and report both the voxel accuracy and the part Intersection over Union (pIoU).

- PartNet [4]: Large collection of CAD models of 17 object categories composed of 251 labeled parts. The dataset provides 17, 119 samples for training, 2, 492 for validation and 4, 895 for evaluation. The dataset provides a benchmark for three different tasks: fine-grained semantic segmentation, hierarchical semantic segmentation and instance segmentation. We report on the first task to evaluate our networks on a more challenging segmentation task using the same part Intersection over Union (pIoU) as in ScanNet.

1.2. Evaluation metrics

To report our segmentation results, we use two versions of the Intersection over Union metric:

- mIoU: To get the per sample mean-IoU, the IoU is first computed for each part belonging to the given object category, whether or not the part is in the sample. Then, those values are averaged across the parts. If a part is neither predicted nor in the ground truth, the IoU of the part is set to 1 to avoid this indefinite form.

The mIoU obtained for each sample is then averaged to get the final score as,

$$\text{mIoU} = \frac{1}{n_{\text{samples}}} \sum_{s \in \text{samples}} \frac{1}{n_{\text{parts}}^{\text{cat(s)}}} \sum_{p^i \in \mathcal{P}_{\text{cat(s)}}} \text{IoU}_s(p^i)$$

with n_{samples} the number of samples in the dataset, cat(s), $n_{\text{parts}}^{\text{cat(s)}}$ and $\mathcal{P}_{\text{cat(s)}}$ the object category where s belongs, the number of parts in this category and the sets of its parts respectively. $\text{IoU}_s(p^i)$ is the IoU of part p^i in sample s.

- pIoU: The part-IoU is computed differently. The IoU per part is first computed over the whole dataset and then, the values obtained are averaged across the parts as,

$$\text{pIoU} = \frac{1}{n_{\text{parts}}} \sum_{p \in \text{parts}} \frac{\sum_{s \in \text{samples}} \text{I}_s(p^i)}{\sum_{s \in \text{samples}} \text{U}_s(p^i)}$$

with n_{parts} the number of parts in the dataset, $\text{I}_s(p^i)$ and $\text{U}_s(p^i)$ the intersection and union for samples s on part p^i respectively.

To take into account the randomness of point cloud sampling when performing coarsening, we use the average of 'N' forward passes to decide on the final segmentation during evaluation when relevant.

Table 1. Summary of the impact of our module implants in five different networks on ShapeNet-Part. The impact is measured by four metrics: (i) memory footprint, (ii) IoU, (iii) inference time and (iv) backward time. With all tested architectures, our lean modules decrease the memory footprint while allowing small improvements in terms of IoU. The impact on inference time depends on the choice of the network but can range from positive impact to a small slowdown.

	Memory	IoU	Inference	Backward
PN++	-76%	+1.2%	-46%	-83%
DGCNN	-69%	+0.9%	-22%	-49%
SCNN	-28%	+2.2%	+27%	+193%
PointCNN	-56%	+1.0%	+35%	+71%

Table 2. Per-category performance mIoU on ShapeNet-Part (Top) and pIoU on PartNet (Bottom) based on a training on each whole dataset all at once. On ShapeNet-Part, all of our network architectures outperform PointNet++ baseline by at least +1.0%. Our deep architecture still improves the performance of its shallower counterpart by a small margin of +0.1%. On PartNet, the fine details of the segmentation and the high number of points to process make the training much more complex than previous datasets. PointNet++, here, fails to capture enough features to segment the objects properly. Our different architectures outperform PointNet++ with a spread of at least +2.0% (+5.7% increase). With this more complex dataset, deeper networks become significantly better: our Deep LPN network achieves to increase pIoU by +9.7% over PointNet++ baseline, outperforming its shallow counterpart by +2.1%.

	Tot./Av.	Aero	Bag	Cap	Car	Chair	Ear	Guitar	Knife	Lamp	Laptop	Motor	Mug	Pistol	Rocket	Skate	Table
No. Samples	13998	2349	62	44	740	3053	55	628	312	1261	367	151	146	234	54	121	4421
PN++	84.60	82.7	76.8	84.4	78.7	90.5	72.3	90.5	86.3	82.9	**96.0**	72.4	94.3	80.5	**62.8**	76.3	81.2
mRes	85.47	**83.7**	77.1	85.4	79.6	**91.2**	73.4	91.6	**88.1**	84.1	95.6	75.1	95.1	81.4	59.7	76.9	82.1
mResX	85.42	83.1	77.0	84.8	79.7	91.0	67.8	91.5	88.0	84.1	95.7	74.6	**95.4**	82.4	57.1	77.0	82.3
LPN	85.65	83.3	77.2	**87.8**	80.6	91.1	72.0	**91.8**	**88.1**	84.6	95.8	**75.8**	95.1	**83.6**	60.7	75.0	82.4
Deep LPN	**85.66**	82.8	**79.2**	82.7	**80.9**	91.1	**75.4**	91.6	**88.1**	**84.9**	95.3	73.1	95.1	83.3	61.6	**77.7**	**82.6**

	Tot./Av.	Bed	Bott	Chair	Clock	Dish	Disp	Door	Ear	Fauc	Knife	Lamp	Micro	Frid	Storage	Table	Trash	Vase
No. samples	17119	133	315	4489	406	111	633	149	147	435	221	1554	133	136	1588	5707	221	741
PN++	35.2	30.1	32.0	39.5	30.3	29.1	81.4	31.4	35.4	46.6	37.1	25.1	31.5	32.6	40.5	34.9	33.0	56.3
mRes	37.2	29.6	32.7	40.0	34.3	29.9	80.2	**35.0**	**50.0**	**56.5**	41.0	26.5	**33.9**	**35.1**	41.0	35.4	35.3	57.7
mResX	37.5	32.0	37.9	40.4	30.2	31.8	80.9	34.0	43.0	54.3	42.6	26.8	33.1	31.8	**41.2**	**36.5**	40.8	57.2
LPN	37.8	**33.2**	40.7	40.8	**35.8**	31.9	81.2	33.6	48.4	54.3	41.8	26.8	31.0	32.2	40.6	35.4	41.1	57.2
Deep LPN	**38.6**	29.5	**42.1**	**41.8**	34.7	**33.2**	**81.6**	34.8	49.6	53.0	**44.8**	**28.4**	33.5	32.3	41.1	36.3	**43.1**	**57.8**

1.3. Summary of the impact of our module

We experiment on four different networks that all exhibit diverse approach to point operation: (i) PointNet++ [6], (ii) Dynamic Graph CNN [7], (iii) SpiderCNN [8], (iv) PointCNN [3] . As detailed in Table 1, our lean blocks, being modular and generic, can not only increase the memory efficiency of that wide range of networks but can as well improve their accuracy. The effect of our blocks on inference time does vary with the type of network, from a positive impact to a small slowdown.

1.4. Detailed results from the paper

The following section provides more details on the evaluation experiments introduced in the paper. We present the per-class IoU on both ShapeNet-Part and PartNet datasets in Table 2 for each of the PointNet++ based architecture. Due to the high number of points per sample and the level of details of the segmentation, PartNet can be seen as much more complex than ShapeNet-Part.

On PartNet, the spread between an architecture with an improved information flow and a vanilla one becomes significant. Our PointNet++ based networks perform consistently better than the original architecture on each of the PartNet classes. Increasing the depth of the network allows to achieve a higher accuracy on the most complex classes such as Chairs or Lamps composed of 38 and 40 different part categories respectively. Our deep architecture is also able to better capture the boundaries between parts and thus to predict the right labels very close from part edges. When a sample is itself composed of many parts, having a deep

Table 3. Per-class IoU on PartNet when training a separate network for each category, evaluated for three different architectures for *Chairs* and *Tables* (60% of the whole dataset). Our lean networks achieve here similar performance as their vanilla counterpart while delivering significant savings in memory.

		Chair	Table
DGCNN	Vanilla	29.2 (+0.0%)	22.5 (+0.0%)
	Lean	24.2 (-17.1%)	**28.9 (+28.4%)**
SCNN	Vanilla	30.8 (+0.0%)	**21.3 (+0.0%)**
	Lean	**31.1 (+1.0%)**	21.2 (-0.5%)
PointCNN	Vanilla	40.4 (+0.0%)	32.1 (+0.0%)
	Lean	**41.4 (+2.5%)**	**33.1 (+3.1%)**

Table 4. Memory and speed efficiency of our deep network Deep LPN with respect to two different implementations of DeepGCNs. Our network wins on all counts and successfully reduces the memory (- 75%) and increases the speed (- 48% and - 89% for the inference and the backward time respectively).

	Memory (Gb)	Inference Time (ms)	Backward Time (ms)
DeepGCN (Dense)	8.56	664	1088
DeepGCN (Sparse)	10.00	1520	445
Deep LPN	**2.18 (-75%)**	**345 (-48%)**	**67 (-85%)**

architecture is a significant advantage.

As additional reference, we provide on Table 3 the performance of our lean blocks applied to three architectures when training one network per-object category on PartNet, trained on *Chairs* and *Tables* as they represent 60% of the dataset.

For reference, we provide as well the absolute values for the efficiency of the previous networks measured by three different metrics on Table 4 and Table 5: (i) memory foot-

Table 5. Efficiency of our network architectures measured with a batch size of 8 samples or less on a Nvidia GTX 2080Ti GPU. All of our lean architectures allow to save a substantial amount of memory on GPU wrt. the PointNet++ baseline from 58% with mRes to a 67% decrease with LPN. This latter convolution-type architecture wins on all counts, decreasing both inference time (-41%) and the length of backward pass (-68%) by a large spread. Starting form this architecture, the marginal cost of going deep is extremely low: doubling the number of layers in the encoding part of the network increases inference time by 6.3% on average and the memory consumption by only 3.6% at most compared to LPN). When used in conjunction with other base architectures, similar memory savings are achieved by our blocks with low impact on inference time.

	Efficiency (%)			Memory Footprint (Gb)			Inference Time (ms)			Length Backward pass (ms)		
	ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet
PointNet++	84.60	80.5	35.2	6.80	6.73	7.69	344	238	666	173	26	185
mRes	85.47	79.4	37.2	2.09	2.93	4.03	395	379	537	54	**12**	68
mResX	85.42	79.5	37.5	2.38	3.15	4.13	441	383	583	122	26	138
LPN	85.65	**83.2**	37.8	1.65	**2.25**	**3.24**	**187**	**166**	**347**	**30**	15	**39**
Deep LPN	**85.66**	82.2	**38.6**	**1.42**	2.33	3.31	205	177	356	37	23	51

		Efficiency (%)			Memory Footprint (Gb)			Inference Time (ms)			Length Backward pass (ms)		
		ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet	ShapeNet-Part	ScanNet	PartNet
DGCNN	Vanilla	82.59	74.5	20.5	2.62	7.03	9.50	41	194	216	41	82	104
	Lean	**83.32**	**75.0**	**21.9**	**0.81**	**3.99**	**5.77**	**32**	**158**	**168**	**21**	**45**	**57**
SCNN	Vanilla	79.86	72.9	17.9	1.09	4.33	5.21	**22**	**279**	**142**	**45**	**99**	**249**
	Lean	**81.61**	**73.2**	**18.4**	**0.79**	**3.25**	**3.33**	28	281	150	132	443	637
PointCNN	Vanilla	83.60	77.2	25.0	4.54	5.18	6.83	**189**	**229**	**228**	**109**	**71**	**77**
	Lean	**84.45**	**80.1**	**27.0**	**1.98**	**3.93**	**5.55**	256	278	263	186	225	208

print, (ii) inference time and (iii) length of backward pass. Our lean architectures consistently reduce the memory consumption of their vanilla couterparts while having a very low impact on inference time. When compared to Deep-GCNs, our Deep LPN architecture wins on all counts by achieving the same performance while requiring less memory (-75%) and shorter inference (-48%) and backward (-89%) time.

2. Design of our architectures

In this section, we provide more details about how we design our lean architectures to ensure reproducible results for the following architectures, (i) PointNet++ [6], (ii) Dynamic Graph CNN [7], (iii) SpiderCNN [8], (iv) PointCNN [3] . We implement each networks in Pytorch following the original code in Tensorflow and we implant our blocks directly within those networks.

2.1. PointNet++ based architectures

To keep things simple and concise in this section, we adopt the following notations:

- S(n): Sampling layer of n points;

- rNN(r): query-ball of radius r;

- MaxP: Max Pooling along the neighborhood axis;

- $\oplus$: Multi-resolution combination;

- Lin(s): Linear unit of s neurons;

- Drop(p): Dropout layer with a probability p to zero a neuron.

Inside our architectures, every downsampling module is itself based on FPS to decrease the resolution of the input point cloud. To get back to the original resolution, upsampling layers proceed to linear interpolation (Interp) in the spatial space using the $K_u = 3$ closest neighbors. To generate multiple resolutions of the same input point cloud, a downsampling ratio of 2 is used for every additional resolution.

2.1.1 PointNet++

In all our experiments, we choose to report the performance of the multi-scale PointNet++ (MSG PN++) as it is reported to beat its alternative versions in the original paper on all tasks. We code our own implementation of PointNet++ in Pytorch and choose the same parameters as in the original code.

For segmentation task, the architecture is designed as follow:

Encoding1:
$$S(512) \rightarrow \begin{bmatrix} rNN(.1) \rightarrow mLP([32, 32, 64]) \rightarrow MaxP \\ rNN(.2) \rightarrow mLP([64, 64, 128]) \rightarrow MaxP \\ rNN(.4) \rightarrow mLP([64, 96, 128]) \rightarrow MaxP \end{bmatrix} \oplus$$

Encoding2:
$$S(128) \rightarrow \begin{bmatrix} rNN(.2) \rightarrow mLP([64, 64, 128]) \rightarrow MaxP \\ rNN(.4) \rightarrow mLP([128, 128, 256]) \rightarrow MaxP \\ rNN(.8) \rightarrow mLP([128, 128, 256]) \rightarrow MaxP \end{bmatrix} \oplus$$

Encoding3:
$S(1) \rightarrow mLP([256, 512, 1024]) \rightarrow MaxP$
Decoding1: Interp(3)$\rightarrow mLP([256, 256])$
Decoding2: Interp(3)$\rightarrow mLP([256, 128])$
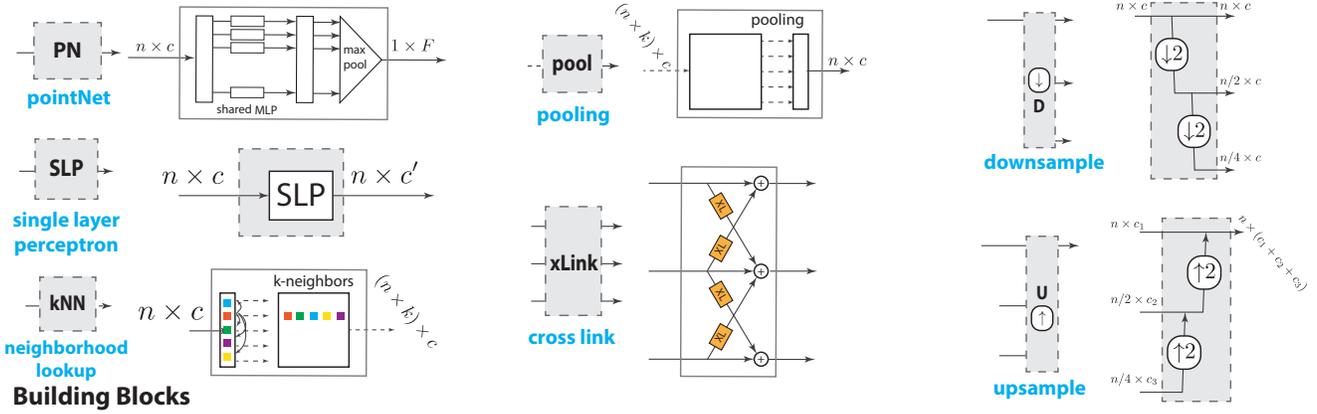
Figure 1. Elementary building blocks for point processing. Apart from standard neighborhood lookup, pooling and SLP layers, we introduce cross-link layers across scales, and propose multi-resolution up/down sampling blocks for point processing. PointNet module combines a stack of shared SLP (forming an MLP) to lift individual points and then performs permutation-invariant local pooling.

Decoding3: $\text{Interp}(3) \to \text{mLP}([128, 128])$

Classification: $\text{Lin}(512) \to \text{Drop}(.7) \to \text{Lin}(\text{nb}_{\text{classes}})$

We omit here skiplinks for sake of clarity: they connect encoding and decoding modules at the same scale level.

2.1.2 mRes

The mRes architecture consists in changing the way the sampling is done in the network to get a multi-resolution approach (see Fig. 2). We provide the details only for the encoding part of the network as we keep the decoding part unchanged from PointNet++.

Encoding1:

$$\begin{bmatrix} S(512) \to \text{rNN}(.1) \to \text{mLP}([32, 32, 64]) \to \text{MaxP} \\ S(256) \to \text{rNN}(.2) \to \text{mLP}([64, 64, 128]) \to \text{MaxP} \\ S(128) \to \text{rNN}(.4) \to \text{mLP}([64, 96, 128]) \to \text{MaxP} \end{bmatrix} \oplus$$

Encoding2:

$$\begin{bmatrix} S(128) \to \text{rNN}(.2) \to \text{mLP}([64, 64, 128]) \to \text{MaxP} \\ S(96) \to \text{rNN}(.4) \to \text{mLP}([128, 128, 256]) \to \text{MaxP} \\ S(64) \to \text{rNN}(.8) \to \text{mLP}([128, 128, 256] \to \text{MaxP} \end{bmatrix} \oplus$$

Encoding3:

$S(1) \to \text{mLP}([256, 512, 1024]) \to \text{MaxP}$

Starting from this architecture, we add Xlink connections between each layer of each mLP to get our mResX architecture. A Xlink connection connects two neighboring resolutions to merge information at different granularity. On each link, we use a sampling module (either down-sampling or upsampling) to match the input to the target resolution. We use two alternatives for feature combination: (i) concatenation, (ii) summation. In the later case, we add an additional sLP on each Xlink to map the input feature dimension to the target. To keep this process as lean as possible, we position the SLP at the coarser resolution, i.e. before the upsampling module or after the downsampling module.
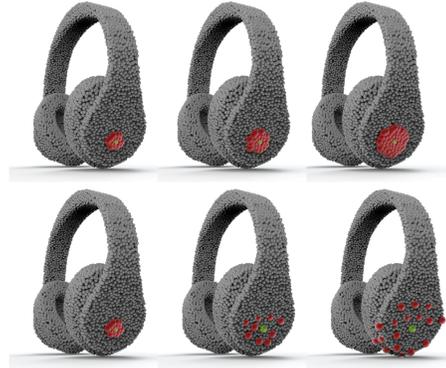
Figure 2. Comparison of multi-scale processing (top) with multi-resolution processing (down): multi-resolution processing allows us to process larger-scale areas while not increasing memory consumption, making it easier to elicit global context information.

2.1.3 LPN

Our convPN module can be seen as a point counterpart of 2D image convolution. To do so, the convPN module replaces the MLP with its pooling layer by a sequence of SLP-Pooling modules.

To simplify the writing, we adopt the additional notations:

- *Sampling* block $S([s_1, s_2, .., s_n]^T)$ where we make a sampling of s_i points on each resolution i. When only one resolution is available as input, the block $S([., s_1, s_2, ..., s_{n-1}]^T)$ will sequentially downsample the input point cloud by $s_1, s_2, ..$ points to create the desired number of resolutions.

- *Convolution* block $C([r_1, r_2, ..., r_n]^T)$ is composed itself of three operations for each resolution i: neighbor-

hood lookup to select the r_iNN for each points, an sLP layer of the same size as its input and a max-pooling.

- *Transition* block $T([t_1, t_2, ..., t_n]^T)$ whose main role is to change the channel dimension of the input to the one of the convolution block. An sLP of ouput dimension t_i will be apply to the resolution i.

Residual connections are noted as *.

Encoding1:
$$S \begin{bmatrix} . \\ 512 \\ 256 \end{bmatrix} \rightarrow T \begin{bmatrix} 32 \\ 64 \\ 64 \end{bmatrix} \rightarrow C^* \begin{bmatrix} .1 \\ .2 \\ .4 \end{bmatrix} \rightarrow T \begin{bmatrix} 32 \\ 64 \\ 96 \end{bmatrix} \rightarrow$$
$$C^* \begin{bmatrix} .1 \\ .2 \\ .4 \end{bmatrix} \rightarrow T \begin{bmatrix} 64 \\ 128 \\ 128 \end{bmatrix} \rightarrow C^* \begin{bmatrix} .1 \\ .2 \\ .4 \end{bmatrix} \rightarrow S \begin{bmatrix} 512 \\ 256 \\ 128 \end{bmatrix} \rightarrow \bigoplus$$

Encoding2:
$$S \begin{bmatrix} . \\ 128 \\ 96 \end{bmatrix} \rightarrow T \begin{bmatrix} 64 \\ 128 \\ 128 \end{bmatrix} \rightarrow C^* \begin{bmatrix} .2 \\ .4 \\ .8 \end{bmatrix} \rightarrow C^* \begin{bmatrix} .2 \\ .4 \\ .8 \end{bmatrix} \rightarrow$$
$$T \begin{bmatrix} 128 \\ 256 \\ 256 \end{bmatrix} \rightarrow C^* \begin{bmatrix} .2 \\ .4 \\ .8 \end{bmatrix} \rightarrow S \begin{bmatrix} 128 \\ 96 \\ 64 \end{bmatrix} \rightarrow \bigoplus$$

Encoding3:
$S(1) \rightarrow mLP([256, 512, 1024]) \rightarrow MaxP$

Note here that there is no *Transition* block between the first two C blocks in the Encoding2 part. This is because those two *Convolution* blocks work on the same feature dimension.

We also add Xlinks inside each of the C blocks. In this architecture, the features passing through the Xlinks are combined by summation and follow the same design as for mResX.

In the case of SLPs, using the on-the-fly re-computation of the neighborhood features tensor has a significant positive impact on both the forward and backward pass by means of a simple adjustment. Instead of applying the SLP on the neighborhood features tensor, we can first apply the SLP on the flat feature tensor and then reconstruct the neighborhood just before the max-pooling layer (Algorithm 1). The same can be used for the backward pass (see Algorithm 2).

2.1.4 Deep LPN

Our deep architecture builds on LPN to design a deeper architecture. For our experiments, we double the size of the encoding part by repeating each convolution block twice. For each encoding segment, we position the sampling block after the third convolution block, so that the first half of the convolution blocks are processing a higher resolution point cloud and the other half a coarser version.

Algorithm 1: Low-memory grouping - Forward pass

Data: Input features tensor $\mathcal{T}_f$ ($N \times R^D$), input spatial tensor $\mathcal{T}_s$ ($N \times R^3$) and indices of each point's neighborhood for lookup operation $\mathcal{L}$ ($N \times K$)

Result: Output feature tensor $\mathcal{T}_f^o$ ($N \times R^{D'}$)

1 **begin**
    /* Lifting each point/feature to $R^{D'}$ */
2     $\mathcal{T}_{f'} \longleftarrow \mathbf{SLP}_f(\mathcal{T}_f)$
3     $\mathcal{T}_{s'} \longleftarrow \mathbf{SLP}_s(\mathcal{T}_s)$
    /* Neighborhood features
    $(N \times R^{D'} \rightarrow N \times R^{D'} \times (K+1))$ */
4     $\mathcal{T}_{f'}^K \longleftarrow \mathbf{IndexLookup}(\mathcal{T}_{f'}, \mathcal{T}_{s'}, \mathcal{L})$
    /* Neighborhood pooling
    $(N \times R^{D'} \times (K+1) \rightarrow N \times R^{D'})$ */
5     $\mathcal{T}_{f'}^o \longleftarrow \mathbf{MaxPooling}(\mathcal{T}_{f'}^K)$
6     $\mathbf{FreeMemory}(\mathcal{T}_{s'}, \mathcal{T}_{f'}, \mathcal{T}_{f'}^K)$
7   **return** $\mathcal{T}_{f'}^o$
8 **end**

Algorithm 2: Low-memory grouping - Backward pass

Data: Input features tensor $\mathcal{T}_f$ ($N \times R^D$), input spatial tensor $\mathcal{T}_s$ ($N \times R^3$), gradient of the output $\mathcal{G}_{out}$ and indices of each point's neighborhood for lookup $\mathcal{L}$ ($N \times K$)

Result: Gradient of the input $\mathcal{G}_{in}$ and gradient of the weights $\mathcal{G}_w$

1 **begin**
    /* Gradient Max Pooling
    $(N \times R^{D'} \rightarrow N \times R^{D'} \times (K+1))$ */
2     $\mathcal{G}_{out}^{mp} \longleftarrow \mathbf{BackwardMaxPooling}(\mathcal{G}_{out})$
    /* Flattening features
    $(N \times R^{D'} \times (K+1) \rightarrow N \times R^{D'})$ */
3     $\mathcal{G}_{out}^{fl} \longleftarrow \mathbf{InverseIndexLookup}(\mathcal{G}_{out}^{mp}, \mathcal{L})$
    /* Gradient wrt. input/weight */
4     $\mathcal{G}_w, \mathcal{G}_{in} \longleftarrow \mathbf{BackwardSLP}(\mathcal{T}_f, \mathcal{T}_s, \mathcal{G}_{out}^{fl})$
5     $\mathbf{FreeMemory}(\mathcal{T}_f, \mathcal{T}_s, \mathcal{G}_{out}, \mathcal{G}_{out}^{mp}, \mathcal{G}_{out}^{fl})$
6   **return** $(\mathcal{G}_{in}, \mathcal{G}_w)$
7 **end**

2.2. DGCNN based architecture

Starting from the authors' exact implementation, we swap each edge-conv layer, implemented as an MLP, by a sequence of single resolution convPN blocks. This set of convPN blocks replicates the sequence of layers used to design the MLPs in the original implementation.

To allow the use of residual links, a transition block is placed before each edge-conv layer to match the dimension of both ends of the residual links.

2.3. SpiderCNN based architecture

A SpiderConv block can be seen as a bilinear operator on the input features and on a non-linear transformation of

the input points. This non-linear transformation consists of changing the space where the points live in.

In the original architecture, an SLP is first applied to the transformed points to compute the points' Taylor expansion. Then, each output vector is multiplied by its corresponding feature. Finally a convolution is applied on the product. Therefore, the neighborhood features can be built *on-the-fly* within the block and deleted once the outputs are obtained. We thus modify the backward pass to reconstruct the needed tensors when needed for gradient computation.

2.4. PointCNN based architecture

For PointCNN, we modify the χ-conv operator to avoid having to store the neighborhood features tensors for the backward pass. To do so, we make several approximations from the original architecture.

We replace the first MLP used to lift the points by a sequence of convPN blocks. Thus, instead of learning a feature representation per neighbor, we retain only a global feature vector per representative point.

We change as well the first fully connected layer used to learn the χ-transformation matrix. This new layer now reconstructs the neighborhood features *on-the-fly* from its inputs and deletes it from memory as soon as its output is computed. During the backward pass, the neighborhood features tensor is easily rebuilt to get the required gradients.

We implement the same trick for the convolution operator applied to the transformed features. We further augment this layer with the task of applying the χ-transformation to the neighborhood features once grouped.

Finally, we place transition blocks between each χ-conv layer to enable residual links.

2.5. Implementation details

In all our experiments , we process the dataset to have the same number of points N for each sample. To reach a given number of points, input pointclouds are downsampled using the furthest point sampling (FPS) algorithm or randomly upsampled.

We keep the exact same parameters as the original networks evaluated regarding most of parameters.

To regularize the network, we interleave a dropout layer between the last fully connected layers and parameterize it to zero 70% of the input neurons. Finally, we add a weight decay of 5e-4 to the loss for all our experiments.

All networks are trained using the Adam optimizer to minimize the cross-entropy loss. The running average coefficients for Adam are set to 0.9 and 0.999 for the gradient and its square, respectively.

References

[1] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Mano-

lis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1

[3] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. 2018. 2, 3

[4] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. *CoRR*, abs/1812.02713, 2018. 1

[5] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, 1(2):4, 2017. 1

[6] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NIPS*, pages 5099–5108, 2017. 2, 3

[7] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *arXiv preprint arXiv:1801.07829*, 2018. 2, 3

[8] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 87–102, 2018. 2, 3