Supplementary Material

This is a supplementary material for the paper, *PointMixer: MLP-Mixer for Point Cloud Understanding*. We will further describe the details: efficiency analysis (Sec. A), point receptive fields comparison (Sec. B), limited cardinality issues in token-mixing MLPs (Sec. C), and task-specific training details (Sec. D).

A Efficiency analysis

In this section, we analyze the latency and memory consumption of each point set layer: PointNet++ layer (Eq. (2)), Point Transformer layer (Eq. (3)) and PointMixer layer (Eq. (6) and Eq. (7)). We conduct this ablation study based on the PointMixer network for the 3D shape classification task. For a fair comparison, we **strictly maintain to use the same downsampling layers**. Furthermore, we do not use inter-set mixing layer to keep the other components of the network the same.

We measure mAcc, OA, the average inference time per object, and the peak GPU memory usage of each method on ModelNet40 dataset [80]. Note that we re-implement PointNet++[8] and Point Transformer with the same number of residual blocks as our PointMixer, and train those models on the ModelNet40 dataset [80] with the same training configuration for a fair comparison.

Table 6. A comparison of efficiency.

Layer type	Param. (M)	Memory (MB)	Latency (ms)	mAcc	OA
PointNet++ [56]	**3.3**	**1463**	**13.04**	90.9	93.2
PointTrans [96]	5.3	1473	19.77	90.2	93.1
PointMixer (ours)	3.5	1465	20.72	**91.2**	**93.3**

As shown Table 6, the network with Point Transformer layer [96] consumes the largest amount of GPU memory to infer a 3D object since it calculates a memory-consuming *vector similarity*. On the other hand, our PointMixer layer computes a *scalar score*, denoted by s_j, to aggregate neighbor features, denoted by $\mathbf{x}_j$. It consequently consumes 8MB less GPU memory than Point Transformer layer [96] although both Point Transformer and Point Mixer layers are slower than PointNet++ layer since both of them use the expensive softmax operation. Furthermore, the PointMixer with only intra-set mixing outperforms PointNet++ [56] layer by 0.3 mAcc and 0.1 OA although PointNet++ [56] also requires much less memory than Point Transformer [96]. This result implies that our score-based aggregation can embed local responses more effectively than simple pooling-based aggregation which PointNet++ [56] uses. As a result, our PointMixer layer can encode local relations within a point set both more effectively and efficiently than previous approaches [56,96], along with the other

[8] Since the original implementation of PointNet++ [56] does not use the residual connection, this re-implementation brings performance gain to the model [56].

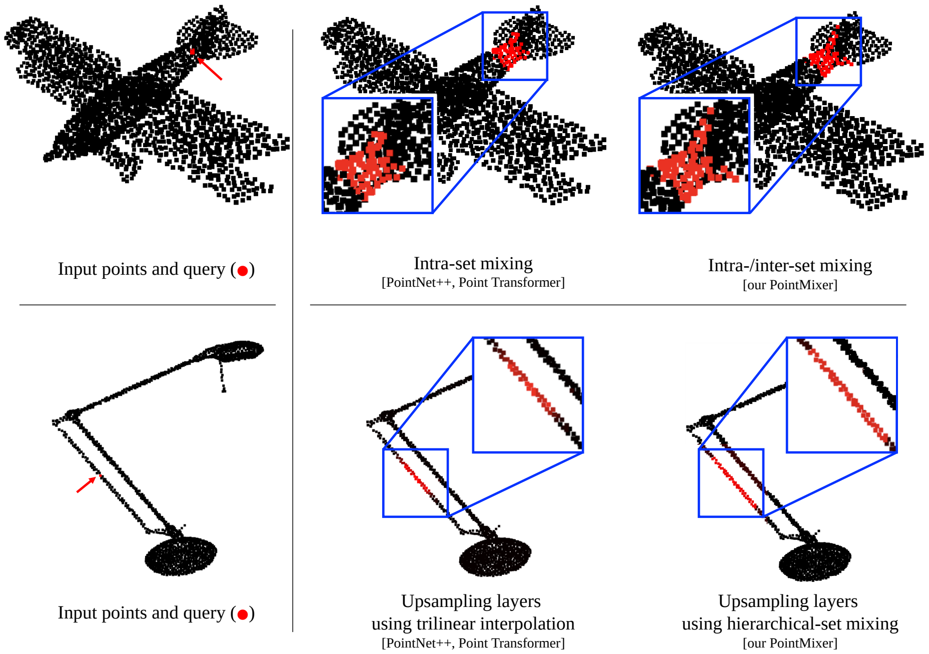
Input points and query (●)
Intra-set mixing
[PointNet++, Point Transformer]
Intra-/inter-set mixing
[our PointMixer]
Input points and query (●)
Upsampling layers
using trilinear interpolation
[PointNet++, Point Transformer]
Upsampling layers
using hierarchical-set mixing
[our PointMixer]

Fig. 8. Point receptive fields. Given a query point (●), we colorize the neighbor points that are affected by the query: intra/inter-set (top) and hier-set (bottom).

strengths that inter-set mixing and hierarchical-mixing layers have, which are already shown in Table 2, Table 3, and Table 5 of the manuscript.

B Point receptive field analysis

Throughout this paper, we emphasize the importance of information sharing among unstructured point clouds. As a universal point set operator, PointMixer layer can function as intra-, inter-, and hierarchical-set mixing while previous studies [56,96,83] only focus on the intra-set mixing operations as shown in Table 1 of the manuscript. For your visual understanding, let us illustrate the receptive fields of our PointMixer layers in different usages as in Fig. 8. Given a query point (●), we visualize the receptive fields of intra-set/inter-set mixing (top row) and upsampling layers (bottom row). In particular, we compare trilinear-based upsampling layers (PointNet++ and Point Transformer) and hierarchical-set mixing from our PointMixer layer. We colorize points as red if the red query point influences these. The PointMixer layer has overall larger receptive fields than previous studies [56,96], which further facilitates point response propagation.

Top row in Fig. 8. As stated in Sec. 3.3 of the manuscript, intra-set mixing is limited to k closest neighbor points. However, combined use of intra-set and inter-set mixing can propagate point responses into variable length of more neighbors from the neighboring point sets $\mathcal{M}_j$, which is consistently supported by Fig. 8.

Bottom row in Fig. 8. Our PointMixer network constitutes a symmetric encoder-decoder network while previous studies do not, as illustrated in Fig. 4 of the manuscript. In particular, since creating a new kNN graph with $k = 16$ is expensive, the trilinear interpolation in upsampling layers of PointNet++ [56] and Point Transformer [96] usually interpolates three nearest neighbor points, which is the much smaller number of neighbors than 16 in the downsampling layer, and limits their receptive fields as well. On the other hand, our PointMixer reuses the kNN graph of the downsampling layer to maximize the receptive fields without additional computational costs. As a result, PointMixer layer can encode point responses in larger contexts than previous approaches [56,96] as shown in Fig. 8. Moreover, these results can be reasons for our superior performance in dense prediction tasks compared to the previous state-of-the-art methods [96,7].

C Limited cardinality issues in token-mixing MLPs

There are two dominant reasons that we remove token-mixing MLPs from our PointMixer layer: limited cardinality and permutation-variant property. In this section, we further describe the technical reasons of the limited cardinality in token-mixing MLPs.

While a given pixel in an image systematically admits eight adjacent pixels, each point can have an arbitrary number of neighbors in a point cloud. In this context, shared MLPs are particularly desirable since they can handle variable input lengths [69]. However, in the vanilla MLP-Mixer layer, token-mixing MLPs limit the process of an arbitrary number of points.

Let us briefly explain the reason. Channel-mixing MLPs require the pre-defined dimensionality in channels for the affine transformation of the input data. In contrast, token-mixing MLPs (Eq. (4) of the manuscript) switch the channel axis and spatial axis, which results in the pre-defined the number of input tokens (*e.g.*, points). Accordingly, we can only take the pre-defined number of points with fixed channel length as an input of token-mixing MLPs. Thus, token-mixing MLPs cannot operate inter-set mixing whose cardinality varies depending on the point cloud distribution.

D Training details

Semantic segmentation We set batch size as 2. Each batch consists of 40K points. We initially set the learning rate as 0.1 and decrease the initial learning rate 10 times smaller at 40, 50 epochs. In total, we train our network for 60 epochs. We use two NVIDIA 1080-Ti GPUs for training. The total training time takes 44 hours.

Point cloud reconstruction We set batch size as 4. The rest of the training conditions are identical to that of semantic segmentation. To train our network, we modify the header layer of the network that we used for semantic segmentation task. Specifically, we change the channel dimensionality of the output as 3 that represents $[x, y, z]$.

Object classification We set batch size as 32. We train our network for 300 epochs and schedule the learning rate using cosine-annealing decay. We use the same SGD optimizer that we used in the semantic segmentation task. In our header network, we utilize MLPs with dropout layers and set the ratio as 0.5.

E Ablation study (rebuttal)

In this section, we provide the more ablation studies requested by the reviewers.

PointMixer vs. previous studies for 3D points. We agree that there are similarities between PointMixer and existing methods (*e.g.*, PointNet++ [56], PointConv [79], and Point Transformer [96]) when it comes to the *intra-set* mixing only. However, in this paper, we aim (1) to improve the network expressiveness via complementary set operations (*intra/inter/hier-set* mixing), (2) to develop a universal set operator, and (3) to design a symmetric network using PointMixer.

We integrate *inter/hier-set* mixing blocks into other existing backbones, and compare those variants with our PointMixer as shown in Table 7. Note that PointNet++ and Point Transformer use max and vector-attention[9] for *intra-set* mixing, respectively. The results show that *inter/hier-set* mixing itself consistently improves the performance of PointNet++ and Point Transformer[10] regardless of the block designs. Interestingly, our mixing scheme (softmax) seems to be more effective than the simple operator (max) as well as the complex layer (vector-attention).

PointMixer as a 3D version of MLP-Mixer. In [68], *"MLP-Mixer contains two types of layers: one with MLPs applied independently to image patches (i.e., "mixing" the per-location features), and one with MLPs applied across patches (i.e., "mixing" spatial information)."*. Similarly, Synthesizer [65] also claims that "*we show that Random Synthesizers are a form of MLP-Mixers. Random Synthesizers apply a weight matrix on the length dimension as a form of projection across the dimension.*". Based on theses concepts, PointMixer can be seen as a form of both MLP-Mixer and Synthesizer through softmax that acts as a projection across the token dimensions, *i.e.*, token-mixing.

Moreover, MLP-Mixer variants[5,23,39,89] also focus on the improved token-mixer. For example, CycleMLP [5] samples pixels in a cyclic style for linear complexity in token-mixing parts. AS-MLP [39] **also removes token-mixing MLPs**, and proposes Axial Shift operations for a better local token communication. From these token-mixing analyses, the direction of PointMixer is aligned

[9] Since vector-attention with an inverse mapping requires many scatter operations, it also heavily consumes GPU memory.

[10] Since there are no available pre-trained weights of Point Transformer on both S3DIS and ModelNet40, we trained the Point Transformer ourselves with the official codes provided by the Point Transformer authors.

Table 7. Semantic segmentation results of existing methods with the proposed inter-/hier-set mixing schemes on the S3DIS dataset. Note that 'max' represents PointNet++ block using maxpool, 'attn' means the vector-attention based Point Transformer block, and 'softmax' implies our PointMixer block. All methods are trained for 30 epoch.

Method	Intra	Inter/Hier	Param. (M)	mIoU (%)
PointNet++	max	✗	2.0	57.3
	max	max	2.3 (↑ 0.3)	62.7 (↑ 5.4)
	max	attn	8.3 (↑ 6.3)	57.8 (↑ 0.5)
	max	softmax	2.7 (↑ 0.7)	66.9 (↑ 9.6)
PointTransformer[10]	attn	✗	7.8	70.0
	attn	max	8.1 (↑ 0.3)	70.2 (↑ 0.2)
	attn	attn	14.1 (↑ 6.3)	70.1 (↑ 0.1)
	attn	softmax	8.5 (↑ 0.7)	70.3 (↑ 0.3)
PointMixer (ours)	softmax	softmax	6.5	71.4

with that of MLP-Mixer variants. Therefore, we respectively argue that PointMixer is a 3D version of MLP-Mixer[11].

Softmax function is not new. Nonetheless, our paper revisits the existing module to emphasize the extended use of kNN graph structure. While previous studies focus on the directional kNN graph (*intra-set* mixing), PointMixer newly notices the 'bi'-directional characteristics of kNN (*inter/hier-set* mixing). Furthermore, to fully utilize this newly-revisited property, the softmax function can be one choice instead of using complex modules. In conclusion, our design choice (replacement token-mixing MLPs with softmax function) is supported by our analysis of permutation-invariant point set operators (Sec. 3.2) across many recent publications [39,90,96] and our extensive experiments (Table 5 and Table 7).

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision (IJCV) **120**(2), 153–168 (2016)
2. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2016)
3. Bello, I.: Lambdanetworks: Modeling long-range interactions without attention. In: International Conference on Learning Representations (2020)
4. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 357–366 (2021)

[11] In this context, PointNet++ also can be seen as one of the simplest versions of MLP-Mixer for point cloud understanding. However, as shown in Table 7, it is outperformed by PointMixer in the set mixing schemes.

5. Chen, S., Xie, E., Ge, C., Liang, D., Luo, P.: Cyclemlp: A mlp-like architecture for dense prediction. In: International Conference on Learning Representations (ICLR) (2022)
6. Choe, J., Joo, K., Imtiaz, T., Kweon, I.S.: Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation. IEEE Robotics and Automation Letters **6**(3), 4672–4679 (2021)
7. Choe, J., Joung, B., Rameau, F., Park, J., Kweon, I.S.: Deep point cloud reconstruction. In: International Conference on Learning Representations (ICLR) (2022)
8. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3075–3084 (2019)
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5828–5839 (2017)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
11. Ding, X., Xia, C., Zhang, X., Chu, X., Han, J., Ding, G.: Repmlp: Re-parameterizing convolutions into fully-connected layers for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: International Conference on Machine Learning. pp. 2286–2296. PMLR (2021)
14. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. CVPR (2018)
15. Graham, B., van der Maaten, L.: Submanifold sparse convolutional networks. arXiv preprint arXiv:1706.01307 (2017)
16. Guo, J., Tang, Y., Han, K., Chen, X., Wu, H., Xu, C., Xu, C., Wang, Y.: Hire-mlp: Vision mlp via hierarchical rearrangement. arXiv preprint arXiv:2108.13341 (2021)
17. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. Computational Visual Media **7**(2), 187–199 (2021)
18. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3d point clouds: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) (2020)
19. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., Wang, Y.: Transformer in transformer. arXiv preprint arXiv:2103.00112 (2021)
20. Handa, A., Whelan, T., McDonald, J., Davison, A.J.: A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In: IEEE Int. Conf. on Robotics and Automation. (ICRA) (2014)
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)

22. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
23. Hou, Q., Jiang, Z., Yuan, L., Cheng, M.M., Yan, S., Feng, J.: Vision permutator: A permutable mlp-like architecture for visual recognition. arXiv preprint arXiv:2106.12368 (2021)
24. Hu, H., Zhang, Z., Xie, Z., Lin, S.: Local relation networks for image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3464–3473 (2019)
25. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11108–11117 (2020)
26. Huang, Q., Wang, W., Neumann, U.: Recurrent slice networks for 3d segmentation of point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
27. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
28. Jiang, L., Zhao, H., Liu, S., Shen, X., Fu, C.W., Jia, J.: Hierarchical point-edge interaction network for point cloud semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10433–10441 (2019)
29. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data (2019)
30. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys (CSUR) (2021)
31. Klokov, R., Lempitsky, V.: Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2017)
32. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) (2017)
33. Komarichev, A., Zhong, Z., Hua, J.: A-cnn: Annularly convolutional neural networks on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7421–7430 (2019)
34. Lee, J., Choy, C., Park, J.: Putting 3d spatially sparse networks on a diet. arXiv preprint arXiv:2112.01316 (2021)
35. Li, J., Hassani, A., Walton, S., Shi, H.: Convmlp: Hierarchical convolutional mlps for vision. arXiv preprint arXiv:2109.04454 (2021)
36. Li, J., Chen, B.M., Lee, G.H.: So-net: Self-organizing network for point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
37. Li, R., Li, X., Heng, P.A., Fu, C.W.: Point cloud upsampling via disentangled refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
38. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. Advances in neural information processing systems **31** (2018)
39. Lian, D., Yu, Z., Sun, X., Gao, S.: AS-MLP: An axial shifted MLP architecture for vision. In: International Conference on Learning Representations (2022)
40. Lin, Y., Yan, Z., Huang, H., Du, D., Liu, L., Cui, S., Han, X.: Fpconv: Learning local flattening for point convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4293–4302 (2020)

41. Liu, H., Dai, Z., So, D., Le, Q.V.: Pay attention to MLPs. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (NeurIPS) (2021)
42. Liu, Y., Fan, B., Meng, G., Lu, J., Xiang, S., Pan, C.: Densepoint: Learning densely contextual representation for efficient point cloud processing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5239–5248 (2019)
43. Liu, Y., Fan, B., Xiang, S., Pan, C.: Relation-shape convolutional neural network for point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8895–8904 (2019)
44. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021)
45. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. Advances in Neural Information Processing Systems (NeurIPS) **32** (2019)
46. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440 (2015)
47. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual MLP framework. In: International Conference on Learning Representations (ICLR) (2022)
48. Mao, J., Wang, X., Li, H.: Interpolated convolutional networks for 3d point cloud understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1578–1587 (2019)
49. Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H., Xu, C.: Voxel transformer for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3164–3173 (2021)
50. Maturana, D., Scherer, S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 922–928. IEEE (2015)
51. Mazur, K., Lempitsky, V.: Cloud transformers: A universal approach to point cloud processing tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10715–10724 (October 2021)
52. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. VISAPP (1) **2**(331-340), 2 (2009)
53. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1520–1528 (2015)
54. Park, C., Jeong, Y., Cho, M., Park, J.: Fast point transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16949–16958 (2022)
55. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 652–660 (2017)
56. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413 (2017)
57. Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R.: 3d graph neural networks for rgbd semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5199–5208 (2017)

58. Qian, G., Hammoud, H., Li, G., Thabet, A., Ghanem, B.: Assanet: An anisotropic separable set abstraction for efficient point cloud representation learning. Advances in Neural Information Processing Systems (NeurIPS) **34** (2021)
59. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
60. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. pp. 68–80 (2019)
61. Simonovsky, M., Komodakis, N.: Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3693–3702 (2017)
62. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
63. Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M.H., Kautz, J.: Splatnet: Sparse lattice networks for point cloud processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2530–2539 (2018)
64. Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., Han, S.: Searching efficient 3d architectures with sparse point-voxel convolution. In: European Conference on Computer Vision (ECCV). pp. 685–702. Springer (2020)
65. Tay, Y., Bahri, D., Metzler, D., Juan, D., Zhao, Z., Zheng, C.: Synthesizer: Rethinking self-attention in transformer models. In: ICML (2021)
66. Tay, Y., Dehghani, M., Bahri, D., Metzler, D.: Efficient transformers: A survey. arXiv preprint arXiv:2009.06732 (2020)
67. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6411–6420 (2019)
68. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems (NeurIPS) **34** (2021)
69. Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., et al.: Resmlp: Feedforward networks for image classification with data-efficient training. arXiv preprint arXiv:2105.03404 (2021)
70. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
71. Trockman, A., Kolter, J.Z.: Patches are all you need? arXiv preprint arXiv:2201.09792 (2022)
72. Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., Shlens, J.: Scaling local self-attention for parameter efficient visual backbones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12894–12904 (2021)
73. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
74. Wang, C., Samari, B., Siddiqi, K.: Local spectral graph convolution for point set feature learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 52–66 (2018)

75. Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J.: Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10296–10305 (2019)
76. Wang, S., Suo, S., Ma, W.C., Pokrovsky, A., Urtasun, R.: Deep parametric continuous convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2589–2597 (2018)
77. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 568–578 (2021)
78. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions of Graphics (ToG) **38**(5), 1–12 (2019)
79. Wu, W., Qi, Z., Fuxin, L.: Pointconv: Deep convolutional networks on 3d point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9621–9630 (2019)
80. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1912–1920 (2015)
81. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems (NeurIPS) **34** (2021)
82. Xu, M., Zhou, Z., Qiao, Y.: Geometry sharing network for 3d point cloud classification and segmentation. In: Association for the Advancement of Artificial Intelligence (AAAI) (2020)
83. Xu, M., Ding, R., Zhao, H., Qi, X.: Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3173–3182 (2021)
84. Xu, Q., Sun, X., Wu, C.Y., Wang, P., Neumann, U.: Grid-gcn for fast and scalable point cloud learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5661–5670 (2020)
85. Yang, Z., Sun, Y., Liu, S., Qi, X., Jia, J.: Cn: Channel normalization for point cloud recognition. In: European Conference on Computer Vision (ECCV). pp. 600–616. Springer (2020)
86. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2018)
87. Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3d shape collections. ACM Transactions on Graphics (ToG) **35**(6), 1–12 (2016)
88. Yu, T., Li, X., Cai, Y., Sun, M., Li, P.: S^2-mlp: Spatial-shift mlp architecture for vision. arXiv preprint arXiv:2106.07477 (2021)
89. Yu, T., Li, X., Cai, Y., Sun, M., Li, P.: S^2-mlpv2: Improved spatial-shift mlp architecture for vision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2022)
90. Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: CVPR (2022)

91. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 558–567 (2021)
92. Zhang, D.J., Li, K., Chen, Y., Wang, Y., Chandra, S., Qiao, Y., Liu, L., Shou, M.Z.: Morphmlp: A self-attention free, mlp-like backbone for image and video. arXiv preprint arXiv:2111.12527 (2021)
93. Zhang, F., Fang, J., Wah, B., Torr, P.: Deep fusionnet for point cloud semantic segmentation. In: European Conference on Computer Vision (ECCV). pp. 644–663. Springer (2020)
94. Zhao, H., Jia, J., Koltun, V.: Exploring self-attention for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10076–10085 (2020)
95. Zhao, H., Jiang, L., Fu, C.W., Jia, J.: Pointweb: Enhancing local neighborhood features for point cloud processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5565–5573 (2019)
96. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16259–16268 (2021)
97. Zhou, Q.Y., Park, J., Koltun, V.: Open3d: A modern library for 3d data processing. arXiv preprint arXiv:1801.09847 (2018)
98. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4490–4499 (2018)