POINTMLP++

Friedrich Dang
TU Munich
friedrich.dang@tum.de

Gorkem Guzeler
TU Munich
gorkem.guezeler@tum.de

Begum Altunbas
TU Munich
begum.altunbas@tum.de

Han Keceli
TU Munich
han.keceli@tum.de

Abstract

In recent years, deep learning architectures tailored for point cloud processing have gained significant attention due to their effectiveness in various computer vision tasks. Among these architectures, PointMLP [1] stands out as a notable variant, leveraging PointNet++ [3] while integrating residual MLP layers [9] for enhanced feature representation. In this study, we introduce PointMLP++, an extension of PointMLP, which incorporates geometric affine modules to normalize local point distributions, thereby refining its ability to capture geometric features effectively.

Additionally, we leverage the Stanford Large-Scale 3D Indoor Spaces (S3DIS) [8] benchmark as a new evaluation metric for scene segmentation task, providing a comprehensive assessment platform. Through several experimentation on the S3DIS benchmark, we evaluate the impact of our modifications, including enhanced point sampling techniques, integration of different geometric affine modules, and refined network architectures, with a particular focus on incorporating attention mechanisms. Our comparative analysis puts forward the strengths and weaknesses of each module, shedding light on their respective contributions towards scene segmentation. This work not only extends the capabilities of PointMLP but also offers valuable insights for researchers aiming to improve point cloud analysis methodologies for real-world applications. Our code is released here: https://github.com/friedang/pointMLP-project

1. Introduction

The field of point cloud analysis has witnessed a significant interest from both academic and industrial domains in recent years [2, 4, 5]. Unlike traditional 2D images characterized by structured pixel grids, point clouds represent unordered and irregular sets of points. Moreover, the inherent

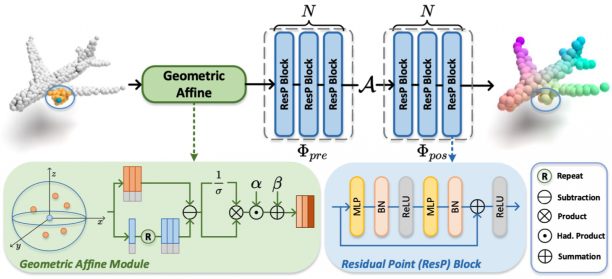

Figure 1. General pipeline introduced by [1].

sparsity and noise present in point clouds further complicate analysis tasks. However, the advent of neural network-based approaches has revolutionized point cloud analysis, leveraging significant advancements in various applications including 3D shape classification [2], semantic segmentation [7], and object detection [4].

Recent research efforts have primarily focused on leveraging local geometric information through techniques such as convolution [6], graph-based methods [6], and attention mechanisms [7]. While these methods have demonstrated promising results, they often rely on complex local feature extraction mechanisms. However, the computational overhead and memory requirements associated with these methods hinder their scalability and efficiency, particularly in real-world scenarios. Consequently, simpler approaches like PointNet and PointNet++, as well as voxel-based methods, remain prevalent in practical applications, with advanced methods being less commonly adopted.

Our model named PointMLP++, aims to have a balance between efficiency and effectiveness inspired by PointMLP [1]. Our contributions in this study extend beyond architectural design. We use a novel benchmark for scene segmentation task, namely the Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset, enabling comprehensive evaluation of our model's performance across diverse domains. Fur-

thermore, we explore various enhancements including different point sampling methods such as Poisson Disk Sampling + Furthest Point Sampling, implementations of geometric affine modules including layer normalization and non-linear normalization, and integration of self-attention and multi-attention mechanisms between feature maps. By testing these modifications on the new benchmark and comparing results, we aim to ensure the robustness and efficacy of these modules in diverse contexts.

In summary, our contributions include not only the development of PointMLP++ for point cloud analysis but also the adaptation and evaluation of diverse modules on a new benchmark, shedding light on their effectiveness and robustness across different domains.

2. Modifications

2.1. Dataset

2.1.1 Original Dataset: ShapeNet Parts

The ShapeNet Parts dataset is a widely used benchmark for 3D object understanding tasks. It consists of 16 object classes, each annotated with 50 part labels. The dataset is formatted with point clouds, normals, part labels, and target labels.

2.1.2 New Dataset: S3DIS

We introduce the S3DIS dataset as a benchmark for testing our model and modifications in the context of 3D indoor space understanding. S3DIS comprises 13 object classes found within 6 different indoor areas, including offices, conference rooms, and lounges. The dataset is structured with point clouds, RGB information, and object labels.

Both datasets serve as valuable resources for evaluating the performance and generalization capabilities of our proposed model.

2.2. Point sampling

The primary sampling method employed in this paper is the Farthest Point Sampling (FPS) algorithm. Here, the index 's' denotes the stage, and for each sampled point, 'K' neighbors are selected. These neighbors are then aggregated using max-pooling to effectively capture local structures.

$$g_i = A\left(\Phi(f_{i,j}) \,|\, j = 1, \ldots, K\right) \tag{1}$$

In our quest to enhance the exploration of point-sampling strategies, we conducted experiments using alternative methods, including random sampling. Moreover, we pursued a combined approach that capitalizes on the synergy between Poisson Disk Sampling and Farthest Point Sampling. In this integrated strategy, we leverage the initial centroids obtained through FPS and seamlessly integrate them with Poisson Disk Sampling. The quantitative results of these experiments are presented in the following sections.

2.3. Geometric affine module

The geometric affine module (GAM) is introduced by PointMLP serves as a lightweight yet effective mechanism for transforming local point features into a more standardized distribution, thereby enhancing the model's performance and generalization capabilities. Recognizing the limitations of deep MLP structures in capturing the diverse geometric structures present in local regions often sparse and irregular, leveraging the geometric affine module helped to overcome these challenges.

The original transformation process involves adjusting the local neighbor points using learnable parameters, specifically a vector a and B. This adjustment is carried out through a normalization operation, where the original geometric properties are preserved while aligning the features to a more standardized distribution.

2.4. Attention

In Figure 2 multiple stages of PointMLP's encoder can be seen. From each stage, features are extracted with an MLP and aggregated. In contrast to the aggregation function in the local groups, the aggregation function for the global context is only a max pooling layer that has no learnable parameters. The function just shrinks the dimensionality of the features to a common size so that it can be processed by one MLP. In the end, it needs to be up-sampled again to match with the decoder's dimensionality.

At this point, we thought it would be beneficial to use also a learnable aggregation function for the global context that does not reduce dimensionality in such a way that features need to be up-sampled again. This is why we used first self-attention and then multi-head self-attention as an aggregation function, followed by an MLP to bring all features to the same dimensionality for global and color context. This modification led to our biggest model version with a capacity of 95 Million learnable parameters, while the non-modified PointMLP had 16 Million parameters.

3. Experiments

3.1. Dataset Adaptation

To adapt our model to the S3DIS dataset, we made several implementation changes, including preprocessing the raw data, incorporating a new dataloader, and adjusting the model layers.

3.1.1 Preprocessing Raw Data

The raw data from the S3DIS dataset is turned into numpy files which combine annotations and data. We used Point-

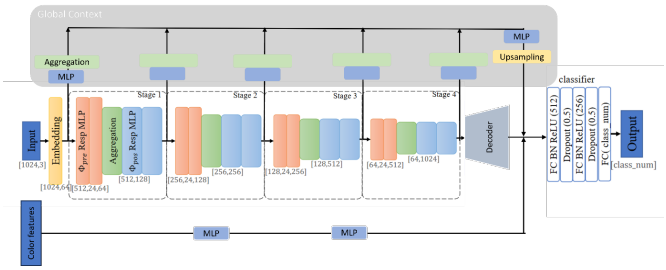

Figure 2. Shows the original architecture without modification for PointMLP for segmentation. While the aggregation function in stages includes the learnable GAM model, the aggregation function within the global context is only an adaptive max-pooling layer, reducing the input dimensionality from [batch size, 64, features] to [batch size, 64, 1], requiring up-sampling to match the decoder's output dimensionality of [batch size, 128, 2048] for concatenation.

Net++ code [3] for this conversion. It is important to note that due to limited storage space and computational resources, we had to use a subsample of the original S3DIS dataset which includes 50 scenes. 39 scenes are used for training and 11 for testing.

3.1.2 Including a New Dataloader

A new dataloader was used for the S3DIS dataset as it has different attributes than ShapeNet. It is responsible for loading batches of preprocessed data during training, and testing phases. A single data instance includes points, normalized points, RGB values. At every iteration, number of points loaded to the model is chosen as 2048.

3.1.3 Adjusting Model Layers

Modifications were made to the model architecture to better suit the characteristics of the S3DIS dataset. This included adjusting the number of input channels to accommodate RGB information, modifying the output layer to match the number of object classes in the dataset, and fine-tuning intermediate layers to capture relevant features for indoor scene understanding.

3.2. Point sampling

In our exploration of point sampling strategies, we employed various methods. Initially, we utilized random sampling, followed by experimentation with Poisson Disk Sampling in combination with farthest point sampling. Surprisingly, our results indicated that random sampling outperformed the combined approach of Poisson Disk Sampling and Farthest Point Sampling. While the reasons for this superior performance are not explicitly detailed in this context, they may be attributed to factors such as data distribution characteristics or the sensitivity of the model to specific sample patterns.

Furthermore, we attempted an experiment with the Lloyd method coupled with farthest point sampling, aiming to refine point set relaxation by iteratively updating the point set based on farthest point sampling. However, due to time constraints, we opted not to train the model with this configuration, as it proved to be computationally intensive and time-consuming.

3.3. Geometric affine module

In our exploration of enhancing model performance, we initially considered layer-based normalization and non-linear normalization methods as potential solutions instead of leveraging the geometric affine module (GAM). However, several factors led us to ultimately choose GAM as the preferred approach.

Firstly, while layer-based normalization methods like batch normalization have been widely adopted in deep learning architectures, they may not effectively capture the nuanced geometric properties of local point features. These methods primarily focus on standardizing activations within each layer, overlooking the specific geometric transformations required at the point level. Similarly, non-linear normalization techniques, such as ReLu activation function may introduce additional complexities without adequately addressing the underlying geometric irregularities present in local regions. These methods often lack the adaptability needed to handle diverse geometric structures across different points.

Furthermore, our experimentation with layer-based and non-linear normalization methods revealed that they did not consistently outperform the default approach, particularly when applied to our proposed dataset S3DIS.

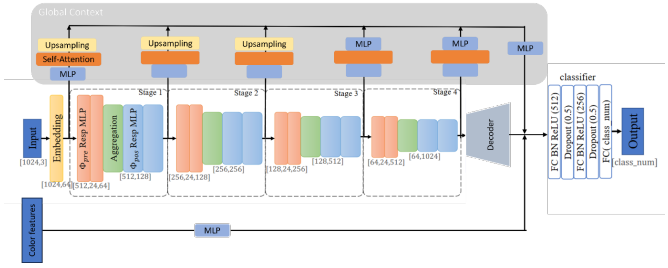

Figure 3. Shows the final version of our modified PointMLP with a single self-attention layer as aggregation function, followed by up-sampling for the first three extractions and MLPs for the last two extractions for global context.

3.4. Attention

During testing, it was noticed that using eight instead of one self-attention head for global context and adding mul-

tihead attention for color context do not result in significant performance improvements for our training and testing setup. Using only a single self-attention head for global context as modification, further model pruning tests were conducted to lower the number of model parameters while keeping performance high. Pruning without significant performance decrease was realized on the color branch (using only a single MLP) and global feature branch (exchanging MLPs with upsampling) for the first 3 extractions. As additionally using random sampling led to a performance decrease, and the final version of our model with 22 Million learnable parameters can be seen in Figure 3. Other tested pruned model versions include using only a single attention aggregation function for all extractions, and using no MLP for the last extractions. All other pruning attempts led to a performance decrease, highlighting the importance of the last extractors' features.

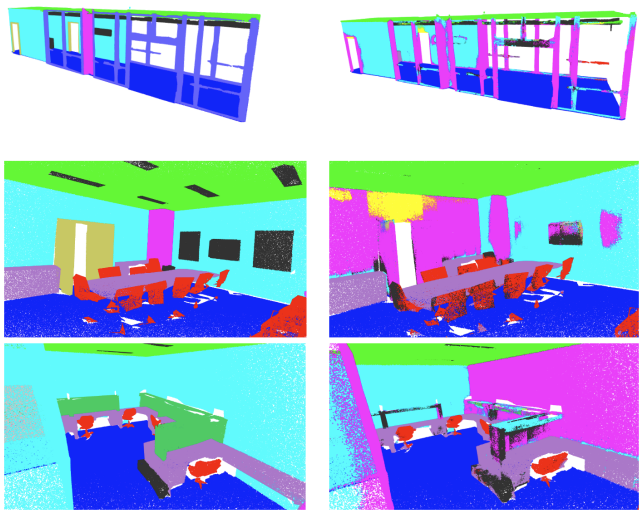

Figure 4. Qualitative results. Left column indicates ground truth segmentation data, right column shows our model's prediction.

3.5. Test results on S3DIS

In Table 1, we present our latest test results obtained after approximately 50 epochs. At the top of the figure, we showcase the performance of PointMLP with no modifications. Notably, the introduction of Attention modifications demonstrates significant improvements, with a consistent enhancement ranging from 3% to 5% across all categories. This highlights the effectiveness of incorporating Attention mechanisms in our model. In Figure 4, we highlight our qualitative results. Due to limited computational resources, we conducted brief training iterations, focusing solely on a subset of the S3DIS dataset. Despite these constraints, our model exhibited strong performance, achieving an overall accuracy of 65%. It's crucial to highlight that our model encounters challenges in effectively distinguishing underrepresented classes, such as the distinction between "wall" and "column."

Modification	Accuracy	Instance mIOU	Class mIOU
None	0.67	0.65	0.52
Random Sampling	0.68	0.67	0.53
Poisson + FPS Sampling	0.68	0.64	0.51
GAM Non-Linear Normalization	0.61	0.59	0.47
Self-Attention with global context	0.70	0.69	0.56
Multihead Self Attention with color & global context	0.70	0.69	0.57
Final, pruned version with Self Attention & Up-Sampling	0.69	0.69	0.57

Table 1. Test results on S3DIS

4. Conclusion

In summary, our paper introduces impactful improvements, showcasing the superiority of Self Attention with global context and Multihead Self Attention with color and global context modifications over the original implementation. Surprisingly, a random point sampling method slightly outperforms the traditional farthest point sampling method. Additionally, our investigation into layer-based and non-linear normalization methods reveals inconsistent performance, particularly on the S3DIS dataset. These findings prompt a reconsideration of established methodologies and inspire further exploration in the field of point cloud analysis.

References

[1] Ma, X. et al. (2022) Rethinking network design and local geometry in point cloud: A simple residual MLP framework, Proceedings of the International Conference on Learning Representations (ICLR), 2022. 1

[2] Qi, Charles R., et al. "PointNet: Deep learning on point sets for 3D classification and segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. 1

[3] Qi, Charles R., Li, Hao, Su, Leonidas J. Guibas. "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 1, 3

[4] Shi, Boxin, et al. ”PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. 1

[5] Xu, Danfei, et al. ”Grid R-CNN.” Proceedings of the European Conference on Computer Vision. 2020. 1

[6] Li, Yifan, et al. ”DeepGCNs: Making GCNs Go as Deep as CNNs.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. 1

[7] Guo, Yunhe, et al. ”RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021. 1

[8] Armeni, Iro, et al. ”3D semantic parsing of large-scale indoor spaces.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. 1

[9] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. ”Deep residual learning for image recognition.” In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016. 1

5. Appendix

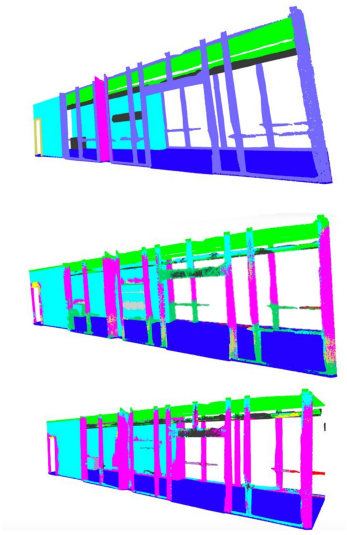

Figure 5. Comparison of Qualitative results. First row indicates ground truth segmentation. Second row shows raw pointMLP performance on S3DIS and third row represents our attention added pointMLP segmentation. It is observed that our modified pointMLP model enhances the segmentation continuity of object parts.