Hybrid Transformer-CNN Architecture for Efficient Point Cloud Classification with Attention Mechanisms

Chiyu Wang[1*],Chen Cheng[1], Huahu Xu[1]

[1] School of Big Data, Zhuhai Institute of Science and Technology, 519000, China

Corresponding Email: 2133269990@qq.com

https://doi.org/10.70695/5ck0y871

Abstract

This paper proposes a Transformer-based point cloud classification method. By designing two key modules: a hybrid feature extraction module and a multi-head self-attention module, efficient feature extraction and classification of point cloud data are accomplished. The hybrid feature extraction module integrates the merits of convolutional neural network (CNN) and Transformer architecture, extracts local geometric features through convolution operations. The multi-head self-attention module detects diverse feature correlations in the point cloud through multiple self-attention heads to further enhance feature representation. Experimental results also demonstrate the effectiveness of our method. Our method achieves 91.6% mAcc and 93.3% OA on ModelNet40 dataset; on the ScanObjectNN dataset, it achieves 79.8% mAcc.

Keywords Point Cloud Classifications; Hybrid Feature Extraction; Multi-Head Self-Attention; Transformer; Attention Mechanism.

1 Introduction

In the past decade, point cloud processing has become an important research direction in computer vision and graphics. Point cloud is a three-dimensional data representation that describes the shape and structure of an object through a series of discrete points distributed in three-dimensional space. Point cloud data is widely used in fields such as autonomous driving,robot navigation, virtual reality and augmented reality (Scavarelli et al., 2021; Yurtsever et al., 2020; Zhu & Zhang, 2021). Therefore, the research on efficient processing and classification technology of point clouds has important practical significance and academic value(Dai & Hansen, 2020; Dai et al., 2019, 2020). With the advancement of sensing technology, it has become easier and easier to obtain high-resolution point cloud data. This provides a rich three-dimensional information resource for various applications, but it also raises challenges in data processing and analysis. Point cloud classification technology plays a key role in many practical applications. In addition, in virtual reality and augmented reality applications, point cloud classification helps build more realistic three-dimensional scenes and interactive experiences(Brieden et al.; Wan et al., 2024; C. Wang et al., 2023; C. Wang et al., 2024).

Early point cloud processing methods mainly relied on hand-crafted feature extraction and rule-based classification algorithms(An et al., 2023; X. Wang et al., 2023; Zhang et al., 2024). These methods have achieved certain results in specific applications, but their robustness and generalization capabilities are limited. PointNet(Charles R Qi et al., 2017) is a pioneering work in this field. It directly processes point cloud data. To overcome the limitations of PointNet, researchers have proposed a variety of improved methods, including Point-Net++(Charles Ruizhongtai Qi et al., 2017), and PointCNN(Li et al., 2018). However, existing methods still have shortcomings in handling complex scenes and multi- scale feature fusion. Therefore, we introduce the Transformer(Hao et al., 2024; Vaswani, 2017; G. Wang et al., 2024) architecture, which has powerful sequence modeling and global feature capture capabilities, and provides a new solution for point cloud classification tasks.

In summary, our main contributions are as follows:

1) Innovative architecture design: This paper proposes a hybrid feature extraction module that combines Trans- former and convolutional network.

2) Extensive Experimental Validation: We conduct exten- sive experiments on two representative datasets, Model- Net40 and ScanObjectNN, to verify the robustness and generalization ability of our method.

2 Related Work

2.1 Voxel-based Approaches

This type of method maps point cloud data into voxels by defining a fixed-size grid in 3D space, so that traditional convolution operations can be applied to point clouds. The VoxelNet(Maturana & Scherer, 2015) model is one of the pioneering works in this field. VoxelNet divides point cloud data into fixed-size voxel blocks and uses 3D convolutional neural networks for feature extraction and classification. Zhou et al.(Yan et al., 2018) further optimized the voxelization method and proposed the SECOND model. SECOND processes voxel data by sparse convolution, greatly improving computational efficiency and achieving good per- formance on the KITTI dataset. The introduction of sparse convolution enables SECOND to significantly reduce the consumption of computing resources while maintaining high accuracy. Lang et al.(Lang et al., 2019) proposed the PointPillars method, which divides point cloud data into vertical cylindrical voxel blocks and performs feature aggregation within the cylindrical voxels.

2.2 Point-based Approaches

The PointNet model proposed by Qi et al. directly processes point cloud data and uses symmetric functions, realizing end-to-end point cloud classification and segmentation. The PointNet method pioneered a new idea for directly processing point cloud data, but its limitation is that it cannot fully capture local geometric features. In order to solve the limi- tations of PointNet, Qi et al. further proposed the PointNet++ model. PointNet++ introduced a hierarchical feature extrac- tion module, which greatly improved the classification and segmentation performance by extracting local features step by step. The PointNet++ method proves the effectiveness of introducing hierarchical feature extraction in point cloud data processing. The DGCNN model proposed by Wang et al. effectively captures the local geometric structure of point cloud data by dynamically constructing a local neighborhood graph and performing convolution operations on the graph.

2.3 Transformer-based Approaches

The PCT model proposed by Guo et al.(Guo et al., 2021) applies the Transformer architecture to point cloud processing. Experi- mental results show that PCT has significantly improved per- formance in point cloud classification and segmentation tasks. The PT model proposed by Zhao et al.(Zhao et al., 2021) can efficiently process large-scale point cloud data by introducing a spatial transformation module and a self-attention mechanism. The PT method not only improves the accuracy of point cloud classi- fication, but also performs well in complex scenes, proving its potential in practical applications. The Point-BERT model pro- posed by Yu et al.(Yu et al., 2022) uses a self-supervised learning method to pre- train point cloud data through the BERT architecture, significantly improving the effects of point cloud classifica- tion and segmentation. The Point-BERT method proves the importance of pre-training in point cloud data processing and achieves excellent performance on multiple datasets.

3 Our Method

This paper proposes a Transformer-based point cloud classification method, which achieves efficient point cloud feature extraction and classification by designing two key modules: hybrid feature extraction module and multi-head self-attention module. Our model structure is shown in the Figure 1.

3.1 Hybrid Feature Extraction module(HFE)

The hybrid feature extraction module aims to combine the advantages of convolutional neural networks (CNNs) and Transformer architectures, extracting local geometric features through convolution

operations while capturing global features using the Transformer’s self-attention mechanism. The input of this module is the original point cloud data P = {pi |i = 1, 2,..., N}, where pi ∈ R3 represents the i-th point in the point cloud, and N is the number of points.

First, we extract the local features of the point cloud through a local convolution operation. Let fconv (•) represent the con- volution operation and Xconv represent the local convolution feature, then:

$$X_{conv} = f_{conv}(P) \quad (1)$$

where $X_{conv} \in R^{N \times d}$, d is the dimension of the convolutional feature.

Next, we input the convolutional features into the Trans- former module and extract global features through the self- attention mechanism. Let $X_{input} = X_{conv}$ represent the input features of the Transformer, then the output features of the Transformer X_{trans} can be expressed as:

$$X_{trans} = f_{trans}(X_{input}) \quad (2)$$

where $f_{trans}(\cdot)$ represents the Transformer operation and $X_{trans} \in R^{N \times d}$.

In order to further integrate local and global features, we concatenate and linearly transform the convolutional features and Transformer features to obtain the final mixed feature X_{mix}:

$$X_{mix} = W_{mix}[X_{conv}; X_{trans}] + b_{mix} \quad (3)$$

where [;] represents the feature concatenation operation, $W_{mix} \in R^{2d \times d}$ and $b_{mix} \in R^{d}$ are the parameters of the linear transformation.

3.2 Multi-head Self-Attention module(MSA)

The multi-head self-attention module is used to further enhance the feature representation and capture different feature relationships in the point cloud through multiple self-attention heads. Let Xmix be the input feature and Xsa be the output feature.

First, we transform the input features linearly to obtain the query matrix Q, key matrix K, and value matrix V :

$$Q = WQ\ Xmix, \quad K = WKXmix, \quad V = WV\ Xmix \quad (4)$$

where WQ ,WK , WV ∈ Rd ×d is the weight matrix of the linear transformation.

Next, we calculate the self-attention weight matrix A:

$$A = \text{softmax}(\overline{\sqrt{d}}\ \_) \quad (5)$$

$$Xsa = AV \quad (6)$$

where softmax(·) is the softmax function and $\sqrt{d}$ is the scaling factor.

To capture multiple feature relationships, we introduce MSA. Let h be the number of heads, then the MSA feature Xmsa is:

$$X_{msa} = \text{Concat}(X_{sa}^{1}, X_{sa}^{2}, \ldots, X_{sa}^{h})W_O \quad (7)$$

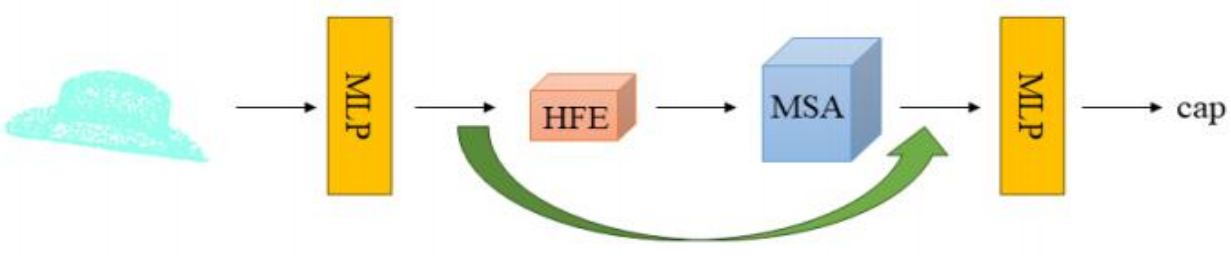

Fig. 1. Diagram of our overall network structure.

Table 1. Comparison results on scanObjectNN and modelnet40.

Model	ModelNet40			ScanObjectNN		Parameters (MB)	FLOPs (GB)
	input	mAcc (%)	OA (%)	input	mAcc (%)		
Other Learning-based Methods							
PointNet	1k	86.2	89.2	1k	63.4	3.47	0.45
PointNet++	5k	-	91.9	1k	75.4	1.48	1.68
CurveNet	1k	90.4	93.1	-	-	2.14	0.30
PointCNN	2k	88.1	92.2	1k	75.1	-	-
Transformer-based Methods							
PointTransformer	1k	89.0	92.8	1k	75.3	21.0	5.05
PCT	1k	90.8	93.2	-	-	2.88	2.32
Our	1k	91.6	93.3	1k	79.8	4.73	1.91

where Concat(·) represents the feature concatenation operation, represents the output feature of the i-th self-attention head, and WO ∈ Rhd ×d is the weight matrix of the linear transformation.

4 Experiment

4.1 Multi-head Self-Attention Module(MSA)

Dataset. The ModelNet40(Wu et al., 2015) dataset has a variety of categories, covering a variety of objects in daily life, such as airplanes, chairs, tables, cups, etc. The dataset provides a large number of 3D object samples, which is suitable for training deep learning models. In addition, each object model is standardized, and all point clouds are normalized to the same scale and position. In our experiments, we preprocess the point cloud data in the ModelNet40 dataset to ensure that each sample contains the same number of points, and train and evaluate the model on the standard training and test splits.

The ScanObjectNN(Uy et al., 2019) dataset is a more complex and chal- lenging point cloud classification dataset released by Nanyang Technological University. This dataset contains scanned point cloud data from real scenes, covering 15 different categories and a total of 2,902 samples. Due to the limitations of scanning technology, point cloud data may have missing parts and irregular shapes, which increases the difficulty of classification tasks. The dataset contains many different types of objects, such as furniture and appliances.

Data preprocessing. To ensure the reliability of the ex- perimental results and the effectiveness of the model, we per- formed corresponding preprocessing steps on the two datasets. First, each point cloud sample was randomly sampled to ensure that each sample contained the same number of points (for example, 1024 points), and all point clouds were normalized to the same scale and center position. Secondly, the point clouds in the ScanObjectNN dataset were denoised and data cleaned.

4.2 Results and Analysis

Table I shows the performance of different methods on these two datasets, including classification accuracy (mAcc and OA), model parameter count (MB), and computational complexity (FLOPs).

On the ModelNet40 dataset, our method achieved 91.6% mAcc and 93.3% OA. This result is significantly better than PointNet (86.2% mAcc and 89.2% OA) and PointNet++ (90.7% mAcc and 91.9% OA). Compared with convolution- based methods such as PointCNN (88.1% mAcc and 92.2% OA) and DGCNN (90.2% mAcc and 92.9% OA), our method also shows advantages in classification accuracy. In addition, our method performs moderately in terms of parameter size (4.73MB) and computational complexity (1.91GFLOPs), re- flecting good efficiency and balance.

On the ScanObjectNN dataset, our method also performs well, achieving 79.8% mAcc. This result is better than most other methods, such as PointNet (63.4% mAcc) and DGCNN (73.6% mAcc). Compared with other Transformer-based meth- ods, such as Point Transformer (75.3% mAcc) and Point- BERT (78.2% mAcc), our method also has a significant advantage in classification accuracy.

4.3 Ablation Experiment

To verify the contribution of the hybrid feature extraction module and multi-head self-attention module proposed by us to the model performance, we conducted an ablation experi- ment on the ModelNet40 dataset. We evaluated the impact on the classification accuracy by removing these two modules respectively. The ablation experiment design is as follows: Complete model (Baseline): includes an HFE module and an MSA module; Remove the hybrid feature extraction module (Model-1): only keep the multi-head self-attention module; Remove the multi-head self-attention module (Model-2): only keep the hybrid feature extraction module.

The Table 2 shows the classification accuracy (mAcc and OA) of different models on the ModelNet40 dataset.

It can be seen that the complete model (Baseline) achieved the highest classification accuracy on the ModelNet40 dataset, with mAcc of 91.6% and OA of 93.3%. When the hybrid feature extraction module (Model-1) was removed, the clas- sification accuracy decreased, with mAcc reduced to 89.4% and OA reduced to 91.2%. This shows that the hybrid feature extraction module plays an important role in capturing local and global features, significantly improving the classification ability of the model. Similarly, when the multi-head self- attention module (Model-2) was removed, the classification accuracy further decreased, with mAcc reduced to 88.7% and OA reduced to 90.8%.

Table 2. Ablantion experiment results on the modelnet40 dataset.

Model	mAcc (%)	OA (%)
Baseline	91.6	93.3
Model-1	89.4	91.2
Model-2	88.7	90.8

5 Conclusion

The Transformer-based point cloud classification method proposed in this paper achieves efficient capture and fusion of features by combining an HFE module and an MSA module. Experimental results also demonstrate the effectiveness of our method. In future work, we plan to further optimize the model structure and explore more types of feature fusion methods in order to achieve better performance on larger and more complex point cloud datasets. At the same time, we will also try to apply this method to other 3D vision tasks, such as point cloud segmentation and object detection, to further verify its generalization ability and practical value in different application scenarios.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. An, Z., Wang, X., T. Johnson, T., Sprinkle, J., & Ma, M. (2023). Runtime monitoring of accidents in driving recordings with multi-type logic in empirical models. International Conference on Runtime Verification,
2. Brieden, A., Cai, Q., Chaimatanan, S., Chen, S., Churchill, A., Couellan, N., Coupe, W. J., Dai, L., De Visscher, I., & de Vries, V. Balakrishnan, Hamsa 101 Bertosio, Florian 130 Blais, Antoine 146.
3. Dai, L., & Hansen, M. (2020). Real-Time Prediction of Runway Occupancy Buffers. 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT),
4. Dai, L., Liu, Y., & Hansen, M. (2019). Modeling go-around occurrence. Proceedings of the Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria,
5. Dai, L., Liu, Y., & Hansen, M. (2020). Predicting go-around occurrence with input-output hidden Markov model. International Conference on Research in Air Transportation,
6. Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R. R., & Hu, S.-M. (2021). Pct: Point cloud transformer. *Computational Visual Media*, *7*, 187-199.
7. Hao, M., Zhang, Z., Li, L., Dong, K., Cheng, L., Tiwari, P., & Ning, X. (2024). Coarse to fine-based image–point cloud fusion network for 3D object detection. *Information Fusion*, *112*, 102551.
8. Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
9. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., & Chen, B. (2018). Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, *31*.
10. Maturana, D., & Scherer, S. (2015). Voxnet: A 3d convolutional neural network for real-time object recognition. 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS),
11. Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition,
12. Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, *30*.
13. Scavarelli, A., Arya, A., & Teather, R. J. (2021). Virtual reality and augmented reality in social learning spaces: a literature review. *Virtual Reality*, *25*(1), 257-277.
14. Uy, M. A., Pham, Q.-H., Hua, B.-S., Nguyen, T., & Yeung, S.-K. (2019). Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. Proceedings of the IEEE/CVF international conference on computer vision,
15. Vaswani, A. (2017). Attention is all you need. *Advances in neural information processing systems*.
16. Wan, Q., Zhang, Z., Jiang, L., Wang, Z., & Zhou, Y. (2024). Image anomaly detection and prediction scheme based on SSA optimized ResNet50-BiGRU model. *arXiv preprint arXiv:2406.13987*.
17. Wang, C., Ning, X., Li, W., Bai, X., & Gao, X. (2023). 3D person re-identification based on global semantic guidance and local feature aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*.
18. Wang, C., Sui, M., Sun, D., Zhang, Z., & Zhou, Y. (2024). Theoretical Analysis of Meta Reinforcement Learning: Generalization Bounds and Convergence Guarantees. *arXiv preprint arXiv:2405.13290*.
19. Wang, G., Yu, L., Tian, S., Zhang, H., Xue, Y., Sang, M., Guo, J., Yu, X., & Si, S. (2024). Pctn: Point cloud data transformation network. *Displays*, *81*, 102610.
20. Wang, X., Liang, A., Sprinkle, J., & Johnson, T. T. (2023). Robustness Verification for Knowledge-Based Logic of Risky Driving Scenes. *arXiv preprint arXiv:2312.16364*.

21. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., & Xiao, J. (2015). 3d shapenets: A deep representation for volumetric shapes. Proceedings of the IEEE conference on computer vision and pattern recognition,
22. Yan, Y., Mao, Y., & Li, B. (2018). Second: Sparsely embedded convolutional detection. *Sensors*, *18*(10), 3337.
23. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., & Lu, J. (2022). Point-bert: Pre-training 3d point cloud transformers with masked point modeling. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,
24. Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020). A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, *8*, 58443-58469.
25. Zhang, H., Wang, C., Yu, L., Tian, S., Ning, X., & Rodrigues, J. (2024). PointGT: A Method for Point-Cloud Classification and Segmentation Based on Local Geometric Transformation. *IEEE Transactions on Multimedia*.
26. Zhao, H., Jiang, L., Jia, J., Torr, P. H., & Koltun, V. (2021). Point transformer. Proceedings of the IEEE/CVF international conference on computer vision,
27. Zhu, K., & Zhang, T. (2021). Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology*, *26*(5), 674-691.
28. Xu Haojun & Bai Jing.(2023).SPS-LCNN: A Significant Point Sampling-based Lightweight Convolutional Neural Network for point cloud processing.Applied Soft Computing Journal
29. Turgut Kaya & Dutagaci Helin.(2023).Local region-learning modules for point cloud classification.Machine Vision and Applications(1),
30. Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic Graph CNN for Learning on Point Clouds.