MPAN: Multi-Part Attention Network for Point Cloud Based 3D Shape Retrieval

ZIRUI LI, JUNYU XU, YUE ZHAO, WENHUI LI, AND WEIZHI NIE, (Member, IEEE)

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing 210007, China

Corresponding authors: Yue Zhao (zhaoxue_tju@163.com), Wenhui Li (liwenhui@tju.edu.cn), and Weizhi Nie (weizhinie@tju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772359, Grant 61572356, Grant 61872267, and Grant 61702471; in part by the Tianjin New Generation Artificial Intelligence Major Program under Grant 19ZXZNGX00110 and Grant 18ZXZNGX00150; and in part by the Elite Scholar Program of Tianjin University under Grant 2019XRX-0035.

ABSTRACT 3D shape retrieval is an important researching field due to its wide applications in computer vision and multimedia fields. With the development of deep learning technology, great progress has been made in recent years and lots of methods have achieved promising 3D shape retrieval results. Due to the effective description of point cloud data on structural information for 3D shapes, lots of methods based on point cloud data format are proposed for better shape representation. However, most of them focus on extracting a global descriptor from the whole 3D shape while the local features and detailed structural information are ignored, which negatively affect the effectiveness of shape descriptors. In addition, these methods also ignore the correlations among different parts of point clouds, which may introduce redundant information to the final shape descriptors. In order to address these issues, we propose a Multi-part attention network (MPAN) for 3D model retrieval based on point cloud. Firstly, we segment a 3D shape into multiple parts by employing a pre-trained PointNet++ segmentation model. After extracting the local features from them, we introduce a novel self-attention mechanism to explore the correlations between different parts. Meanwhile, by considering the structural relevance of them, the redundancy for representing 3D shapes is removed while the effective information is utilized. Finally, informative and discriminative shape descriptors, considering both local features and spatial correlations, are generated for 3D shape retrieval task. To validate the effectiveness of our method, we conduct several experiments on the public 3D shape benchmark, ShapeNetPart dataset. Experimental results and comparisons with state-of-the-art methods demonstrate the superiority of our proposed method.

INDEX TERMS 3D shape retrieval, self-attention, point cloud based method.

I. INTRODUCTION

With the rapid development of computer vision and multimedia technologies, 3D shapes are widely utilized in a surge of fields such as virtual reality, 3D printing and industry designing. In order to manage the massive 3D shapes, exploring effective 3D shape retrieval algorithms is necessary, and it attracts much more research attention in recent years.

Many deep neural network architectures have been introduced to deal with different 3D data representations, such as point clouds, multi-view and volumetric data. Due to the wider source of data acquisition and storage, it is flexible to utilize point clouds data, which pays attention to the positional relations of data points and reflects the structural information of 3D shapes. Therefore, point cloud based methods for 3D model retrieval occupy a large proportion of recent works and show convincing performances.

In point cloud based methods, each 3D shape is represented as a set of 3D points, which could preserve more spatial information and internal local features [1]. However, the point clouds are usually disordered, increasing the difficulty of data processing. In addition, conventional 2D CNN networks are unsuitable to extract global and local features of them. Recently, several networks have been introduced for shape representation of point clouds. Among them, PointNet [2] is a milestone of point cloud based methods, which encodes the points and generates a global point clouds signature by aggregating the point features, considering the permutation
invariance of points. PointNet++ [3], as an improved version of PointNet, introduces a hierarchical neural network, grouping the small neighborhoods into large units for higher-level feature extraction, which improves the discrimination of global features. However, the two fundamental methods PointNet and PointNet++ have their own disadvantages in capturing local features. Concretely, PointNet focuses on learning spatial codes of individual points and aggregating global features, ignoring the local structures induced by the metric. In addition, PointNet++ employs a hierarchical feature extracting architecture for local features extraction and gradually expands the scope, ignoring the correlations of different regions. These shortages have negative impacts on capturing structural information of point clouds, which limits the performance of shape descriptors generation and retrieval results.

Recently, lots of methods are proposed to improve the effectiveness of point cloud descriptors. Dynamic Graph Convolutional Neural Network (DGCNN) [4] employs a fundamental architecture called Edge-Conv layer to capture the local structural features while still maintaining the permutation invariance. Considering the relations between the center point and its neighbors in a constructed local graph, the Edge-Conv architecture could dynamically update the edge features by applying a multi-layer perceptron (MLP) mechanism. The edge features are then utilized to generated point descriptors by employing the pooling operation in each Edge-Conv layer. However, DGCNN mainly focuses on the extraction of local features but ignore the spatial correlations of different shape parts. Shen et al. [5] proposed a novel KCNet to utilize the kernel correlation and graph pooling operation for mining the local features of point clouds, which considers the spatial relationships in point sets at the same time. However, the overlapping information introduces lots of redundancy, which also negatively affects the retrieval performance.

In general, the existing methods have their own disadvantages in capturing the spatial correlations between points and regions, which negatively affects the performance of local feature extraction. In addition, due to the overlapping information in the point cloud representation process, lots of redundant information is introduced. The two main problems should be carefully considered for generating more discriminative shape descriptors.

To alleviate these problems, we propose a novel Multi-part attention network (MPAN) for point cloud based 3D shape retrieval. Firstly, a PointNet++ segmentation model is employed to segment 3D point cloud into different parts, the pre-trained PoinNet++ classification model is then utilized to extract local features of these segmented parts. Secondly, in order to explore the spatial correlations between different segmented parts, we design a novel self-attention network for feature updating. During the training process, effective information for representing 3D shapes is strengthened and the redundancy could be effectively removed. After aggregating all the feature vectors from different parts by max-pooling operation, informative and discriminative 3D shape descriptors are obtained for the classification and retrieval task. The contributions of this paper are summarized as follows:

- We proposed a novel MPAN architecture, which focuses on the correlations among different segmented parts of 3D shape to generate informative and discriminative shape representations.
- We apply the self-attention mechanism to explore the spatial relevance of local features, which can effectively remove the redundancy and make the effective information more focused.
- We validate the superiority of our method on several benchmarks and significant improvements are achieved over the state-of-the-art methods.

The rest of this paper are organized as follows: In Sec.2, we briefly review the related works. We then introduce our network in Sec.3. Experimental results are presented in Sec.4 and we summarize this paper in Sec.5.

II. RELATED WORKS

Recently, the 3D shape retrieval methods are widely divided into two categories: model-based and view-based methods. We briefly review some typical neural networks of each category in the following subsections. Moreover, we shortly introduce the applications of self-attention mechanism.

A. MODEL-BASED METHOD

The model-based methods directly process the raw representations of 3D shapes, including voxel [6]–[8], polygon mesh or surfaces [9]–[11] and point cloud [2], [12], [13]. Concretely, Feng et al. [14] proposed a novel Mesh neural network named MeshNet, which introduce a general architecture with available and effective blocks to capture and aggregate features of polygon faces in 3D shapes. Meanwhile, the complexity and irregularity problems of mesh are effectively solved. Instead of using full-voxel method or full-point-cloud method, Liu et al. [15] proposed a Point-Voxel network that takes the input data as point and applies convolution operation, transforming it to voxel, which greatly reduces the computational complexity and achieves promising experimental result in both efficiency and accuracy. Inspired by Self-Organizing Network (SOM), Li et al. [16] proposed SO-Net which performs dimensional reduction on point clouds and extracts features of SOM nodes, theoretically guaranteeing the invariance of point order. SO-Net explicitly models the spatial distribution of points and provides precise control of the receptive field overlap. Cheraghian et al. [17] proposed a novel 3DCapsule architecture for 3D point cloud representation task. Compared with the traditional Capsule network, a new layer called ComposeCaps is added to in lieu of lost spatial relationships caused by permutation invariance, which improves the performance of 3DCapsule. Yang et al. [18] proposed a novel FoldingNet architecture to solve the point cloud representation task. The FoldingNet utilizes an...
encoder-decoder mechanism to reconstruct point clouds in the unsupervised condition, significantly improving the effectiveness of shape descriptors. Angelina et al. [19] proposed a novel PointNetVLAD architecture, combining the existing Pointnet and NetVLAD for point cloud based feature extraction. The PointNetVLAD network maps each point into a higher dimensional space and learns several cluster centers to compute the contributions of each local feature for global descriptor generation, which improves the discrimination of shape descriptors. Bold et al. [20] employed a novel bidirectional feature match strategy to handle the similarity measurement problem. Based on the nearest neighbor (NN) features of KD-trees, the bidirectional similarities between query and candidates are utilized to calculate the best-buddies similarity (BBS) score, which determines the rank list. The introduction of the bidirectional feature match method significantly improves the retrieval performance.

B. VIEW-BASED METHOD

View-based methods usually project the raw 3D shapes to a set of 2D images from different angles and utilized the extracted visual information to generate shape descriptors. Concretely, Su et al. [21] proposed Multi-view Convolutional Neural Network (MVCNN), which aggregates multi-view descriptors by employing a pooling operation for 3D model recognition and retrieval task.

Han et al. [22] proposed a novel architecture called 3D2SeqViews for 3D global feature learning by aggregating sequential views. It not only focuses on the content information of multiple views but also preserves spatial correlations among views. In addition, the view-level and class-level attention mechanisms are also introduced to comprehensively aggregate content and spatial information, which effectively increases the discrimination of shape descriptors. You et al. [1] proposed a novel architecture PVNet, considering both the point cloud and multi-view data for 3D shape representation. PVNet utilizes the high-level global features from multi-view data and design an attention embedding fusion method to generate the attention-aware features of point cloud models, which strengthen the discrimination of shape descriptors. Similarly, You et al. [23] also proposed a novel multi-modality method PVRNet, which fuses the features from two modalities by exploiting the correlations between point cloud and each individual view by the relevance. This design fuses the two-modality data together and forms integrated shape descriptor, which effectively improve the shape retrieval performance. Jiang et al. [24] proposed a novel multi-loop-view convolutional neural network (MLVCNN) framework, which utilizes the view level, the loop level and the shape level information to conduct informative shape representations, considering the intra-view relationships from different scales. Chen et al. [25] proposed a novel hypergraph based collaborative feature learning scheme for 3D shape retrieval. This method could effectively fuse the descriptors from both the contour and interior region of 3D shapes, and the introduction of Greedy Search (GS) algorithm improves the effectiveness of similarity calculation between the query and the candidates.

C. SELF-ATTENTION MECHANISM

Self-attention mechanism is widely utilized in computer vision and multimedia fields recently and it shows great advantages in exploring the correlations among a set of features. Li et al. [26] proposed a novel architecture named Positional Self-Attention with Co-attention (PSAC) for video question answering. PSAC calculates the response at each position by attending to all positions within the same sequence and then add the descriptors of absolute positions for better representations of sequences. Yang et al. [27] proposed a novel Point Attention Transformers (PATs), using a parameter-efficient Group Shuffle Attention (GSA) to replace the costly Multi-Head Attention mechanism. The input points of PATs are firstly embedded into high-level representations through an Relative Position Embedding (ARPE) module, then they pass through the GSA blocks and down-sampling blocks. They also proposed a Gumbel Subset Sampling (GSS) to replace the Furthest Point Sampling mechanism for a better sampling result.

III. OUR APPROACH

In this section, we detail our approach. The whole framework is shown in Fig. 1 and the architecture consists of three parts: 1) Shape segmentation: we employ the pre-trained PointNet++ segmentation model to segment a 3D shape to multiple parts. Concretely, 3D shapes from a category could be segmented to a fixed number of components. For each part, local features are obtained by utilizing PointNet++ classification model. 2) Self-attention network: we introduce the self-attention mechanism for exploring the correlations among multiple parts. Different weights reflect the importance of each part in the 3D shape representation task. The correlations among multiple parts are effectively explored and utilized, which reduce the redundant information and improve the discrimination of local features. 3) 3D shape retrieval: after aggregating all the local features from multiple parts, informative and discriminative shape descriptors are generated, which could be directly utilized to solve the 3D shape retrieval problem.

A. SHAPE SEGMENTATION

Nowadays, the point cloud data is widely applied, and it could be conveniently achieved by utilizing 3D sensing devices or converted from other data formats. In order to explore the detailed local features of point clouds, we firstly segment a point cloud into a group of parts. Concretely, we utilize the PointNet++ architecture as the segmentation model, which is pre-trained on ShapeNetPart dataset [28], the partial labels are utilized to train the segmentation model. The reason is that the PointNet++ performs better in terms of model size, compared with the existing methods, which could effectively decrease the complexity of the proposed method while maintains an excellent shape segmentation effect at the same
FIGURE 1. The architecture of our proposed MPAN. We firstly segment a 3D shape into multiple parts by utilizing a fine-tuned PointNet++ network. After extracting the local features from different parts based on the PointNet++ classification model, we employ a novel self-attention network to explore the spatial correlations of multiple parts and the redundant information for shape representation could be removed at the same time. Finally, we aggregate all the local features by a max-pooling calculation to generate the informative and discriminative shape descriptors for 3D shape retrieval task.

time. In the segmentation process, 2048 points are uniformly sampled from each shape. Each point is represented as a 6-dimensional vector, which contains x, y, z coordinates and the corresponding normal coordinates.

Then we randomly select some points in different segmented parts and copy them in order to expend the point numbers of each part to 2048. In addition, we employ the PointNet++ classification network to extract CNN features of each part. Different from the original architecture, we remove the last three fully connected layers and utilize the output of feature propagation layer as the local features. Finally, the shape descriptors could be represented by $F_i \in \mathbb{R}^{n \times D}$, where $n$ is the number of segmented parts for each 3D shape, $D$ represents the dimension of local features, where $D = 1024$ in our method.

B. SELF-ATTENTION NETWORK

After the shape segmentation process, the local features for each part of 3D shape are obtained. It is obvious that directly aggregating these local features may introduce lots of redundancy. In addition, the correlations of multiple parts are important for describing the spatial structure of 3D shapes. To address these problems, we propose a novel self-attention network for feature updating, considering the partial correlations and redundancy removing at the same time.

The principle of self-attention is utilizing the self relevance of input feature matrix to strengthen the role of effective information. Based on the traditional self-attention mechanism [29], we connect several self-attention units to gradually increase the proportion of effective information. The calculation process is presented in Eq. 1. Firstly, we initialize three projection matrices $W^Q_i \in \mathbb{R}^{d \times d_k}$, $W^K_i \in \mathbb{R}^{d \times d_k}$ and $W^V_i \in \mathbb{R}^{d \times d_v}$, where $i$ represents the number of self-attention layers, $d_k$ is the dimension of input feature vector and $d_v$ is the dimension of output feature vector after the self-attention operation. $W^Q$, $W^K$ and $W^V$ represent the training weight parameter for query, key and value matrix, respectively. In addition, $F_i$ is the input feature matrix of the $i$-th layer. For each self-attention layer, we multiply the weights by input feature $F$ to obtain the updated self-attention vector. Then we multiply the weighted query matrix $F^i W^Q_i$ by the weighted key matrix $F^i W^K_i$ and apply the softmax operation to obtain the self-relevance matrix, which reflects the correlation parameters between local features. In addition, we scale the dot product by $1/\sqrt{d_k}$.

$$\text{Attention}(X^i) = \text{Softmax}\left(\frac{F^i W^Q_i \times (F^i W^K_i)^T}{\sqrt{d_k}}\right) F^i W^V_i .$$

Note that the self-relevance matrix reflects the contribution of each local feature for describing the whole 3D shape. Then we define a weight parameter $W_F$ to replace the self-relevance matrix, in order to simplify Eq. 1 as the following equation shows:

$$\text{Attention}(X^i) = W_F F^i W^V_i .$$

In order to take the advantages of low-level correlation information among multiple parts, we add the input of current layer to the output feature as the input of next self-attention
layer, which can be presented by the following equation:

\[ F^{i+1} = \text{Attention}(F^i) + F^i. \] (3)

After the calculation of self-attention network, an updated matrix which consists feature vectors from different segmented parts is generated. The proposed scheme can not only remove the redundancy, but enhance the effective information at the same time.

C. 3D SHAPE RETRIEVAL

In order to generate a final shape descriptor of each 3D shape, we apply the max-pooling operation on the column of matrix \( \tilde{F} \) to fuse all the local features, which can be represented by Eq. 4:

\[ F^q(y) = \max_{x \in \mathbb{R}} F^i(x, y) \] (4)

where \( i \) represents the number of self-attention layers, \( n \) is the number of segmented parts, \( y \) is the \( y \)-th element in \( F^i \), the \( F^q \) is the final output feature matrix of self-attention network. \( F^q \) represents the final shape descriptor. Then, a softmax architecture is employed to predict the category label of \( F^q \). We plug the predicted label of \( F^q \) and the corresponding groundtruth label of it into the cross-entropy loss function for model training, which could be represented by the following equation:

\[ \text{Loss} = -\sum_i t_i \log y_i + (1 - t_i) \log (1 - y_i), \] (5)

where \( y_i \) is the predicted category label of \( i \)-th shape descriptor, and \( t_i \) is the groundtruth label of the \( i \)-th shape descriptor.

To improve the performance of shape retrieval, we employ the Mahalanobis metric, directly projecting the shape descriptors to a low-dimensional space, in which the inter-class distances are extended and the intra-class distance is reduced. Concretely, we utilize the large-margin learning algorithm from [30] to learn a weight metric \( W \) projecting high-dimensional shape descriptors \( F^q \in \mathbb{R}^d \) to low-dimensional feature vectors \( WF^q \in \mathbb{R}^p \). Meanwhile, the squared Euclidean distance \( d_W^2(F^q_i, F^q_j) = \|WF^q_i - WF^q_j\|_2^2 \) is employed to compute the distance between shape \( i \) and \( j \). Concretely, \( d_W^2 \) is smaller if \( i \) and \( j \) are from the same category, and larger otherwise. By utilizing the metric learning method, we control the distribution of feature vectors, considering the category information, which makes the projected shape descriptors more discriminative. In addition, the dimension of shape descriptor \( F^q \), before and after dimensionality reduction are 1024 and 128, respectively.

Concretely, the shape retrieval process could be completely explained as follows: Firstly, we compute descriptors of all the shapes, utilizing the proposed method. Then, we employ the Euclidean distance to measure distances between the query and other shapes. Finally, we obtain the sorted rank list based on the calculated distances. Generally, if the retrieved shape is from the same class as the query, they are matched. The retrieval performance could be quantified by different evaluation metrics, which are detailed in the following section.

IV. EXPERIMENTS

A. DATASET

To demonstrate the effectiveness of our proposed method, we design extensive experiments on the ShapeNetPart [28] dataset, which is a subset of ShapeNetCore. Concretely, it contains a large amount of 3D shapes from 16 shape categories and is split into a training set with 13998 shapes and a testing set with 3894 shapes.

B. EVALUATION CRITERIA

In the 3D shape retrieval task, several evaluation metrics are adopted to measure the performance of our proposed method, including Precision-Recall curve (PR-curve), Mean Average Precision (mAP), Nearest Neighbor (NN), the First Tier (FT), the Second Tier (ST), Discounted Cumulative Gain (NDCG) and Average Normalized modified retrieval rank (ANMRR). We further introduce them in detail as follows:

- The Precision-Recall Curve (PR-Curve) is a fundamental method to evaluate the retrieval performance. It plots the change of Accuracy and Recall Rate when the threshold is changed, in order to distinguish the correlation and irrelevance in object retrieval task. The recall and precision could be calculated as follows:

\[ \text{Recall} = \frac{N_c}{N_{all}}, \] (6)

\[ \text{Precision} = \frac{N_{Rall}}{N_{class}}, \] (7)

\[ AR = \sum_{i=1}^{N_{class}} \text{Recall}(i), \] (8)

\[ AP = \sum_{i=1}^{N_{class}} \text{Precision}(i), \] (9)

where \( N_c \) is the number of correctly retrieved objects, \( N_{all} \) is the number of all relevant objects, \( N_{Rall} \) is the number of 3D shape categories.

- The Mean Average Precision (mAP) is calculated based on the area under PR-curve, which presents a numerical result of the retrieval performance.

- Nearest Neighbor (NN): percentage of the closest matches that belong to the same class as the query, which provides an indication of how well a nearest neighbor classifier would perform. And higher scores means better retrieval results.

- The First Tier (FT): the percentage of models in the query’s class that appear within the top K matches, where K depends on the size of query’s class. Specifically, for a class with C members:

\[ K = |C| - 1 \] (10)
The first tier indicates the recall for the smallest $K$ that could possibly include 100% of the shapes in the query class. An ideal matching result gives a score of 100%, again higher values indicate better matches.

- **Second Tier (ST):** the second Tier is calculated based on the size of query’s class, similar as the FT evaluation criteria. And $K = 2 \times (|C| - 1)$, which is a little less stringent compared with the First Tier.
- **The F measure (F):** it is a composite measure of precision and recall for a fixed number of retrieved results. Concretely, F considers the top 20 retrieved objects for each query member. The calculation formula of F is shown as follows:

$$F = (2 \times P_{20} \times R_{20})/(P_{20} + R_{20}),$$

where $P_{20}$ and $R_{20}$ are the precision and recall values of the top 20 retrieval results.

- **The Discounted Cumulative Gain (DCG):** a statistic that weights correct results near the front of the list more than later results in the ranked list. Specifically, the ranked list $R$ is converted to a list $G$, where element $G_i$ has a value 1 if element $R_i$ is in the correct class and 0 otherwise. The DCG is defined as follows:

$$DCG_1 = G_1; DCG_i = DCG_{i-1} + \frac{G_i}{\log_2(i)}, \quad i > 1.$$

Then, the result is divided by the maximum possible DCG to give the final score:

$$DCG = \frac{DCG_k}{1 + \sum_{j=2}^{\lfloor |C|/2 \rfloor} \frac{1}{\log_2(j)}},$$

where $k$ is the number of shapes in the database, $C$ represents size of the class.

- **The Average Normalized Modified Retrieval Rank (ANMRR):** it is a rank-based measure, and it considers the ranking information of relevant objects among the retrieved list. The calculation process of ANMRR is elaborated in [31].

### C. IMPLEMENTATION DETAILS

In the shape segmentation part, we uniformly sample 2048 points from each 3D shape and segment the 3D shape by a pretrained PointNet++ segmentation model based on ShapeNetPart dataset. In the feature extraction part, we randomly choose points from each part and duplicate them in order to expand the point number of each part to 2048. The feature extraction network is based on the PointNet++ classification model, which is pretrained on the ModelNet40 dataset.

Concretely, the training process could be divided into two parts: Firstly, we trained a PointNet++ segmentation model on ShapeNetPart dataset. The segmented parts are then fed into the PointNet++ classification model for feature extraction. Note that the classification model provided by PointNet++ is pre-trained on ModelNet40 dataset, which means that the PointNet++ classification network is also trained based on the same pre-trained model. Then we fed the extracted partial features into the self-attention network for further point cloud representation.

In the training process, we adopt Adam optimizer with a learning rate decay of $10^{-4}$ and the batch size is set to 128. In order to prevent over-fitting, the dropout rate between each layer is set to 0.5. We trained our network for 200 epochs and we chose the best training result to compare with current state-of-the-art methods. We implement our method based on Pytorch [32] and all the experiments are conducted on a server with 2 NVIDIA 1080TI GPU and an Intel(R) Core(TM) I7-9700K CPU.

### D. ABLATION STUDY

In this section, we conduct several experiments to validate the effectiveness of each part in our proposed network, which contains two mainly components: shape segmentation architecture and self-attention network. In our proposed method, we firstly segment 3D shape into several parts, which aims to effectively utilize the local features for a better representation. In order to take the correlations between different parts into consideration, we introduce the self-attention network. Thus, there exists three conditions for MPAN: without segmentation, with segmentation but without self-attention and with both of them.

Concretely, we fine-tune the PointNet++ classification model based on the ShapeNetPart dataset, and the output feature of the last fully-connected layer is taken as the shape descriptor for the 3D shape retrieval process in the condition of without segmentation. In the second case, the 3D shapes are segmented by introducing the fine-tuned PointNet++ segmentation model, then we apply the mean-pooling or max-pooling calculation on the local features from different parts to generate shape descriptors. In the last case, we introduce the self-attention network for feature updating followed by a pooling calculation to obtain shape descriptor. The experimental results on classification accuracy and retrieval mAP are presented in Table 1 and the results on the other evaluation metrics are shown in Table 2. The comparison on PR-Curve is presented in Fig. 2.

| Method                        | Retrieval mAP |
|-------------------------------|---------------|
| Without segmentation          | 79.5%         |
| Without self-attention (max-pooling) | 67.5%         |
| Without self-attention (mean-pooling) | 65.8%         |
| MPAN                          | 88.4%         |

This experiment aims to evaluate the contribution of each part to the overall network. From the results, we have the following observations:

- The proposed network achieves the best experimental result on the ShapeNetPart dataset with the retrieval
### Table 2. The comparison of retrieval performance in terms of several evaluation metrics under different conditions of ablation study.

| Method                        | NN  | FT  | ST  | F   | DCG | ANMRR |
|-------------------------------|-----|-----|-----|-----|-----|-------|
| Without segmentation          | 0.987 | 0.946 | 0.973 | 0.073 | 0.959 | 0.044 |
| Without self-attention (mean-pooling) | 0.967 | 0.721 | 0.868 | 0.069 | 0.759 | 0.254 |
| Without self-attention (max-pooling) | 0.989 | 0.766 | 0.914 | 0.070 | 0.801 | 0.211 |
| MPAN                          | 0.994 | 0.985 | 0.989 | 0.079 | 0.972 | 0.021 |

### Figure 2. The comparison of PR Curves for retrieval under different conditions of ablation study.

mAP of 88.4%, utilizing both the segmentation and self-attention methods. It brings an improvement of 8.9% compared with the condition not utilizing segmentation method.

- From the experimental results, the methods only utilizing the segmentation process performs worse than the other architectures, which could be explained from that the original segmented parts lack the ability to describe local features due to the existence of a large amount of redundant information. In addition, the max-pooling calculation on local feature vectors achieves a better result compared with mean-pooling, demonstrating that max-pooling could keep more effective information to improve shape retrieval performance.

- Compared with the method of only utilizing segmentation process (max-pooling), the case of introducing self-attention network has achieved an improvement of 20.9% in terms of retrieval mAP. This phenomenon demonstrates the effectiveness of self-attention mechanism. The spatial correlations between different segmented parts, which are important in the feature updating process, could be effectively explored and utilized. Moreover, the redundancy could be removed while the effective information is more focused, which improves the discrimination of shape descriptors.

In general, the segmentation process and self-attention network play important roles for 3D shape descriptor generation. The segmentation process makes MPAN pay more attention to the local features, which reflect the detailed characteristics of 3D shapes. The introduction of self-attention network gives MPAN the ability to chose effective information and reduce unnecessary redundancy, which contributes to the discrimination of shape descriptors. The experimental results effectively demonstrate the effectiveness of each component in MPAN.

### E. Sensitivity Analysis on the Number of Self-Attention Layers

As talked above, the self-attention network gives MPAN the ability to select effective information and reduce the redundancy for shape descriptor generation. In this section, we conduct comparative experiments on different numbers of self-attention layers to explore the impact of network depth. Concretely, we select the number from \( \{1, 2, 3, 4, 5\} \) and the performances on retrieval mAP are presented in Table.3.

### Table 3. The comparison of retrieval performance in terms of retrieval mAP on different numbers of self-attention layers.

| Number of self-attention layers | Retrieval mAP |
|---------------------------------|---------------|
| 1                               | 86.1%         |
| 2                               | 86.7%         |
| 3                               | 87.8%         |
| 4                               | 88.4%         |
| 5                               | 87.6%         |
| 6                               | 86.4%         |

Compared with the methods without self-attention mechanism, MPAN achieves significant improvement. The gain of utilizing one self-attention layer is 18.6%, compared with the max-pooling method, which directly demonstrates the effectiveness of self-attention mechanism in terms of redundancy removement and effective information strengthen. As the depth increases, the ability of information selection is strengthened gradually. Since the one-layer self-attention network has made such great progress, the gain from increasing layer number is not significant, but is effective.

In addition, MPAN achieves the best experimental result when the number of self-attention layer is set to 4. There exists an upward trend on retrieval mAP at the beginning, demonstrating that the ability of effective information selection is strengthened as the network depth increases. However, the retrieval performance becomes worse when the number of self-attention layer is greater than 4. This phenomenon could be explained from that the ability of removing redundancy is limited when depth exceeds the reasonable range, which negatively affects the discrimination of shape descriptors. In general, we set the number of self-attention layers to 4 for the better retrieval performance.
F. SENSITIVITY ANALYSIS ON THE NUMBER OF SAMPLING POINTS

The number of sampling points directly impacts the discrimination of shape descriptor. In this section, we conduct comparative experiments on sampling point numbers to validate the robustness of our proposed network. It is necessary since the number of sampling points determines the amount of effective information for representation 3D shapes. Concretely, we sample 512, 1024, 2048, 4096 points on each 3D shape as the input to explore the relationship between sampling point numbers and the retrieval performance of MPAN. The experimental results on several evaluation metrics are presented in Table 4 and the corresponding PR-Curves are shown in Fig. 3.

| Number of Sampling Points | Retrieval mAP |
|---------------------------|---------------|
| 512                       | 83.4%         |
| 1024                      | 86.2%         |
| 2048                      | 88.4%         |
| 4096                      | 87.5%         |

From the experimental results presented in Table 4, we have the following observations:

- When the number of sampling points is set to different numbers, there exists no obvious gap. Concretely, the retrieval mAP reaches 88.4% when 2048 sampling points are utilized, when the number of points is reduced by 50%, the retrieval mAP only decreased by 2.2%. This phenomenon validates that MPAN has robustness to the number of sampling points, which also demonstrates the effectiveness of MPAN for shape representation task.

According to the experimental results on the number of sampling points, we set 2048 as the final point number in the other experiments for a better retrieval performance.

G. SENSITIVITY ANALYSIS ON THE WEIGHT OF DIFFERENT PARTS

After the segmentation process, the point cloud of a 3D shape is divided into multiple parts, which reflect local characteristics. By introducing the self-attention network, the correlations between different parts could be explored and effectively utilized. The weigh parameter $W_F$ mentioned in Eq. 1 reflects the importance of each part to the final shape descriptor. Concretely, the category information could be distinguished by some representative parts, and their corresponding weigh parameters are bigger while the weights of those confused parts are smaller.

In this section, we design the experiment to exploit the contributions of each part to the shape representation process. Here we list some representative examples from several categories to describe the weights of different segmented parts. And the experimental results on the value of $W_F$ are presented in Fig. 4. From the experimental results, we have the following observations:

- The value of $W_F$ directly reflects the importance of different parts to the shape descriptors. Here we take a representative airplane shape as a visualized example. The weights of tail, engine, fuselage, wing in last self-attention layer are 0.206, 0.103, 0.102 and 0.589, respectively. Since the part of wing contains more discriminative information to distinguish the airplane category compared with other parts, the weight of wing is significantly higher than the others after four self-attention layers. The changing weights of different parts demonstrate that our method could effectively capture the effective information for shape representation.

- The introduction of self-attention network gives MPAN the ability to select effective information and remove redundancy. Concretely, the weight of wing increases from 0.197 to 0.366, 0.468 and 0.589, which demonstrates that the ability is strengthened gradually as the number of self-attention layers increases.

In general, the sensitivity analysis on the weight of different parts validate the effectiveness of self-attention network in shape descriptor generation task.
H. COMPARED WITH STATE-OF-ART METHODS

To demonstrate the superiority of our proposed network, we conduct comparative experiments with some typical state-of-the-art methods. We present the experimental results of several representative point cloud based methods, including PointNet [2], PointNet++ [3], PointCNN [33], Kd-Network [34] and DGCNN [4] for comparison.

The experimental results of all the compared methods are shown in Table 5. The corresponding PR-Curves are presented in Fig. 5. As shown in Table 5, our MPAN outperforms all the other methods with the retrieval mAP of 88.4%. Compared with the classic point cloud based methods PointNet and PointNet++, MPAN obtain a gain of 15.0% and 8.9% on retrieval mAP respectively, which also demonstrates the superiority of our proposed method.

TABLE 5. Comparisons of retrieval experimental results on the ShapeNetPart dataset.

| Method          | Data Representation | Retrieval (mAP) |
|-----------------|---------------------|-----------------|
| 1 PointNet [2]  | Point Cloud         | 73.4%           |
| 2 PointNet++ [3]| Point Cloud         | 79.5%           |
| 3 PointCNN [33] | Point Cloud         | 82.5%           |
| 4 DGCNN [4]     | Point Cloud         | 85.4%           |
| 5 Kd-Network [34]| Point Cloud         | 84.6%           |
| 6 MGN           | Point Cloud         | 88.4%           |

The state-of-the-art performance of MPAN can be dedicated to the following reasons:

- Compared with the traditional point cloud based method PointNet and its improvement approach PointNet++, MPAN achieves a better experimental result. PointNet and PointNet++ are typical shape segmentation methods, which focus on the position relations between points and exploit the local features to generate shape descriptors. However, they roughly utilize the intuitive local features and ignore the selection of effective information, which introduces lots of redundancy to the final descriptor.
- Compared with other point cloud based methods, MPAN also represents its superiority. Concretely, the Kd-Network utilize traditional kd-tree architecture to form the computational graph for sharing learnable parameters, which ignore the subtle local features in point clouds. The core of PointCNN is a convolutional operator that weights and permutes input points and features before they are processed by a typical convolution, which only focus on the contributions of each points but ignore the correlations among them.
- Compared with these representative methods, MPAN pays more attention to the detailed local features by introducing the shape segmentation model, which strengthens the discrimination of shape descriptors. In addition, the self-attention network could remove the redundant information while exploring the correlations between different parts for a better representation. These designs improve the retrieval performance of MPAN obviously.
V. CONCLUSION
In this paper, we propose a novel Multi-part Attention Network (MPAN) for 3D shape retrieval task, which utilizes the correlations between local features to strengthen the discrimination of 3D shapes. Concretely, we first employ a fine-tuned PointNet++ segmentation model to divide a 3D shape into different parts. The PointNet++ classification model is utilized for feature extraction. Then, the self-attention network is introduced to explore the correlations between different segmented parts. Meanwhile, the effective information for describing 3D shapes is reserved and strengthened, the redundant informative could be effectively removed. After the feature updating process, we aggregate the local features from different parts to generate the final shape descriptor for the 3D shape retrieval task. To validate the effectiveness of our method, we evaluated MPAN on ShapeNetPart dataset and the experimental results demonstrate the superiority of our approach over the state-of-the-art methods.

REFERENCES
[1] H. You, Y. Feng, R. Ji, and Y. Gao, “PVNet: A joint convolutional network of point cloud and multi-view for 3D shape recognition,” in Proc. ACM Multimedia Conf. Multimedia Conf. (MM), 2018, pp. 1310–1318.
[2] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 652–660.
[3] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep hierarchical feature learning on point sets in a metric space,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5099–5108.
[4] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph CNN for learning on point clouds,” ACM Trans. Graph., vol. 38, no. 5, pp. 1–12, Nov. 2019.
[5] Y. Shen, C. Feng, Y. Yang, and D. Tian, “Mining point cloud local structures by kernel correlation and graph pooling,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 4548–4557.
[6] D. Maturana and S. Scherer, “VoxNet: A 3D convolutional neural network for real-time object recognition,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2015, pp. 922–928.
[7] C. R. Qi, H. Su, M. NieBner, A. Dai, M. Yan, and L. J. Guibas, “Volumetric and multi-view CNNs for object classification on 3D data,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 5648–5656.
[8] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, “3D ShapeNets: A deep representation for volumetric shapes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1912–1920.
[9] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, “MeshCNN: A network with an edge,” ACM Trans. Graph., vol. 38, no. 4, p. 90, 2019.
[10] X. Jie, G. Dai, F. Zhu, E. K. Wong, and Y. Fang, “DeepShape: Deep-learned shape descriptor for 3D shape retrieval,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 7, pp. 1335–1345, Jul. 2017.
[11] Z. Han, Z. Liu, J. Han, C.-M. Vong, S. Bu, and C. L. P. Chen, “Mesh convolutional restricted Boltzmann machines for unsupervised learning of features with structure preservation on 3-D meshes,” IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 10, pp. 2268–2281, Oct. 2017.
[12] M. Dominguez, R. Dhandhere, A. Petkar, S. Jain, S. Sah, and R. Ptscha, “General-purpose deep point cloud feature extractor,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 1972–1981.
[13] A. Garcia-Garcia, F. Gomez-Domos, J. Garcia-Rodriguez, S. Orts-Escolano, M. Cazorla, and J. Arozen-Lopez, “PointNet: A 3D convolutional neural network for real-time object class recognition,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2016, pp. 1578–1584.
[14] Y. Feng, Y. Feng, H. You, X. Zhao, and Y. Gao, “MeshNet: Mesh neural network for 3D shape representation,” in Proc. AAAI Conf. Intell. Agent. (AAAI), vol. 33, 2019, pp. 8279–8286.
JUNYU XU received the M.S. and Ph.D. degrees from the School of Electrical and Information Engineering, Tianjin University. He is currently working with The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing. His research interests include systems engineering, image processing, and machine learning.

YUE ZHAO is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University. His research interests include 3D shape recognition and cross domain learning.

WENHUI LI received the M.S. and Ph.D. degrees from the School of Electrical and Information Engineering, Tianjin University. He was an Intern Student with the Sesame Center, National University of Singapore. His research interests include computer vision, machine learning, and 3D model retrieval.

WEIZHI NIE (Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China. He was a Visiting Scholar with the NExT Center, National University of Singapore, where he was with Prof. T.-S. Chua. He is currently an Assistant Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include computer vision, machine learning, and social networks.

---

* * *