3D Model Retrieval Algorithm Based on Attention and Multi-view Fusion

Ziqi,Shi[1]

752188882@qq.com

Ziyang,Quan[1]

879849874@qq.com

Jingshan,Shi[1]

2842903046@qq.com

Zhuyan,Guo[1]

1550709422@qq.com

Mandun,Zhang[1,2]

zhangmandun@scse.hebut.edu.cn

Zhidong,Xiao[3]

zxiao@bournemouth.ac.uk

1.School of Artificial Intelligence, Hebei University of Technology

2.Tianjin International Joint Center for virtual reality and visual computing

3.Faculty of Media and Communication, Bournemouth University

With the rapid development of computer vision, 3D data is increasing rapidly. How to retrieve similar model from a large number of models has become a hot research topic. However, in order to meet people's demand, the retrieval accuracy need to be further improved. In terms of multi-view 3D model retrieval, how to effectively learn the information between views is the key to improving performance. In this paper, we propose a novel 3D model retrieval algorithm based on attention and multi-view fusion. Specifically, we mainly constructed two modules. First, dynamic attentive graph learning module is used to learn the intrinsic relationship between view blocks; Then we propose the Attention-NetVlad algorithm, which combines the channel attention algorithm and the NetVlad algorithm. It learns the information between feature channels to enhance the feature expression ability firstly, then uses the NetVlad algorithm to fuse multiple view features into a global feature according to the clustering information.Finally the global feature is used as the only feature of the model to retrieve according to Euclidean distance. In comparison with other state-of-the-art methods by utilizing ModelNet10 and ModelNet40 the proposed method has demonstrated significant improvement for retrieval mAP. Our experiments also demonstrate the effectiveness of the modules in the algorithm.

Additional Keywords and Phrases: 3D model retrieval, Convolutional Neural Network, Attention, Feature fusion,Multiview

1 INTRODUCTION

3D model retrieval, which aims to retrieve most similar model according to 3D model geometry, has important applications in industries that use 3D models, e.g., industrial product design, virtual reality, and 3D games. Accurate and efficient 3D model retrieval algorithm has drawn much attention recently as the number of 3D models increases rapidly due to the emergence of low-cost 3D acquisition equipment.

The current 3D model retrieval work can be generally categorized into two paradigms: model-based retrieval [1,2,3]and view-based retrieval[4,5].Model-based retrieval can better retain the original data information and spatial geometric features by extracting 3D data features of 3D models, e.g., polygon mesh, voxel mesh[6], point cloud and implicit surface. However, in reality, due to the unavailability of opensource 3D feature datasets, it is difficult to utilize 3D data to represent the model directly, and it also limits the related research. In addition, due to the high dimensional features, using 3D feature descriptors to characterize the high-level features of a 3D model can easily lead to overfitting. To overcome the concerns, view-based methods use a set of 2D pictures to represent the 3D model. A set of 2D views of a 3D model from different angles are obtained by placing virtual cameras at different viewpoints, which can reduce the matching between 3D models to a 2D level. Querying the searched model by matching the similarity of the view can greatly avoid the overfitting problem. Moreover, more public 2D image classification datasets can be used to pretrain the network, so that the network can optimize parameters in advance.

In recent years, view-based methods are the main stream for 3D model retrieval. MVCNN[7] proposed to render 12 different angle views of the model from different viewpoints, and generate the feature descriptor of each model through the view pool layer, which revealed good performance in the classification and retrieval of the model. However, the mean average precision of 3D model retrieval need to be further improved.

In response to this urgent problem, this paper proposes a novel 3D model retrieval algorithm based on attention and multi-view fusion. First, a set of 2D views of the 3D model are rendered through a virtual camera, and the 2D views are input to the ResNet network to extract local features. The network performs excellently in multi-view image field. Then,we iuput the local features into dynamic attentive graph learning module and the Attention-NetVlad module to obtain global feature. Finally realize the similarity measurement through the Euclidean distance between the model features, so as to complete the model retrieval work and improve the purpose of model retrieval accuracy.

To summarize, the three main contributions of this work are as follows:

(1) To the best of our knowledge, it is the first research to explicitly use algorithm related to graph attention network to solve the precision problem in 3D model retrieval. Aiming at the task of 3D model retrieval precision, An algorithm called dynamic graph attentive learning(DAGL) is used to learn the correlations between features.

(2) Attention-NetVlad are proposed. Combining channel attention and NetVlad algorithm to effectively fuse information between multiple views. Channel attention learns the information between different channels of the same view, NetVlad algorithm fuse different view features into a global feature.

(3) The experiment result has shown a statistically significant improvement compared with other advanced algorithms in the 3D model shape benchmark dataset ModelNet10 and ModelNet40

The remainder of the paper is structured as follows. In Section 2, the related work is briefly reviewed, followed by a complete introduction to various parts of the proposed algorithm in Section 3. In Section 4, the specific configuration of the experiment and the comparison of the experiment results with other algorithms are presented. Finally, the conclusion of the paper is presented in Section 5.

2 RELATED WORK

Here, we first present several different methods for 3D model retrieval, then introduce the graph attention network(GAT) related to our algorithm.

2.1 3D model retrieval Methods

Model-based Methods Wang [8] et al. proposed a voxel-based convolutional neural network (NormalNet) for the retrieval of 3D models. The network uses the normal vector of the surface as input, and uses a reflection-convolution-connection (RCC) module to implement the convolutional layer. It extracts distinguishable features for 3D vision tasks and significantly reducing the number of parameters. The combination of a network with normal vectors and voxels as inputs further improves the performance of NormalNet. Fuyura and Ohbuchi[9] proposed a new deep neural network for 3DMR, called deep local feature aggregation network (DLAN), which uses a single deep structure to extract rotation-invariant 3D local features and aggregate them to generate a 3D model feature that is not affected by 3D rotation. Kumawat [10]et al. proposed an algorithm (Unveiling Local Phase in 3D Convolution Nerual Networks, LP-3D) in 2019, in which the Rectified Local Phase Volume block(ReLPV) effectively replaced the standard 3D convolutional layer solves the problem of large calculation amount, memory intensive, and easy overfitting, and enhances the feature learning ability.

View-based Methods Feng [11] et al. proposed a group view-based convolutional neural network (GVCNN). Specifically, an extended CNN is used to extract a view-level descriptor, and then a grouping module is introduced to estimate each view. On this basis, all views can be divided into different groups according to their degree of discrimination. Finally, according to the discriminative weight of group-level descriptors, they are combined into shape-level descriptors to measure the model similarity. Nie[12] et al. proposed a new multilayer deep network (MSCNN). First, extract multiple rendered images from a 3D object and combine them into a representative view. Use the upper network to capture the language information of the view, and use the lower network to learn the low score features. In combination, a feature learning model with local and global information of the 3D object is generated. Sun[13] et al.proposed the DRCNN network and constructed a new network layer called Dynamic Routing Layer (DRL) to fuse the features of each view more efficiently. Nie[14] et al. proposed a deep attention network (DAN) by improving the multi-layer deep self-attention. Zhao[15] et al. proposed SVHAN, which learns the relationship between features hierarchically and selectively fuses information from different features.Although the view-based retrieval algorithm has achieved good results, it also has some inherent problems Achieving effective query in a large number of model views is still a problem to be considered and solved.

2.2 Graph Attention Networks

Graph Attention Network (GAT) [16] is derived from graph convolutional neural network[17] and performs attention learning on data with graph structure. The basic idea is: for each node, calculate its similarity with all its adjacent nodes, and then update the representation of the current node with the weighted sum of the adjacent node features.

The graph attention network uses the attention mechanism to perform aggregation operations on the neighbors of nodes, and adaptively assigns weights to different neighbors, which has stronger expressive ability than graph convolutional networks. The following work in this paper will also use the improved graph attention network —DAGL[18]as an integral part of the model.

3 METHODS

Inspired by MVCNN, the proposed network is established, which adds the DAGL and Attention-NetVlad algorithm to the MVCNN network. The network architecture is shown in Figure 1. The network can be divided into three modules: (1)View Feature Extraction (2) Dynamic Attentive Graph Learning.(3) Attention-NetVlad module.

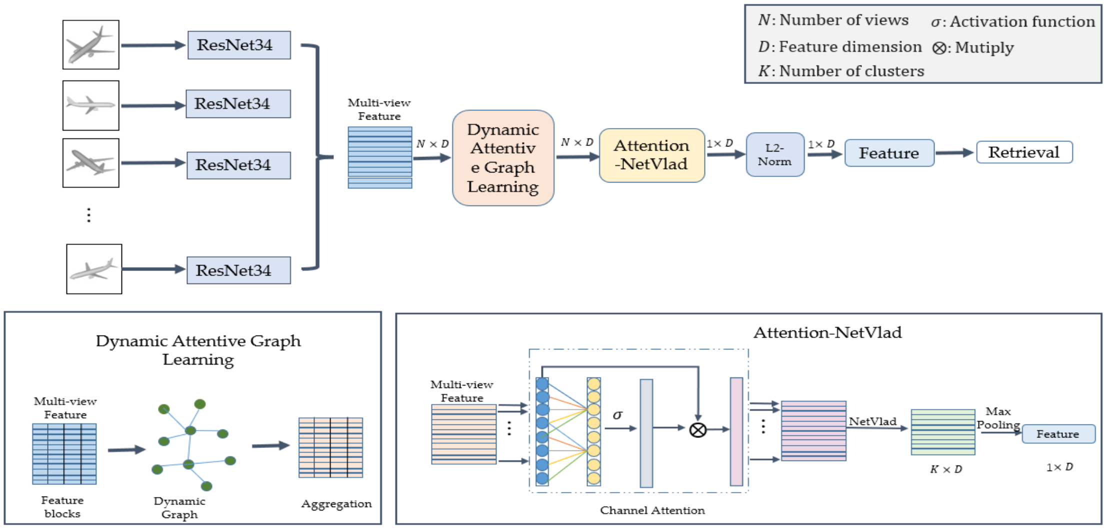

Figure 1: Network structure of our algorithm

3.1 View Feature Extraction

In this paper, a multi-view representation is chosen which is derived from the traditional classic view selection method proposed in MVCNN by applying the 12-view rendering method as shown in Figure 2. A specific 3D mesh model V is described as $V(M) = \{v_l, 1 \leq l \leq n\}$, where n is the number of views.

As shown in Figure 2, 12 virtual cameras at 30 ° intervals along the circumference of the unit ball are evenly placed, and the cameras are perpendicular to the line between the center of the ball and the camera. These views are evenly distributed in different viewpoints of the 3D model, so they have strong complementarity and low correlation with each other. Therefore, the multi-view representation obtained on this basis can constitute a relatively complete description of the 3D model.

The basic features are extracted by ResNet34[19] as shown in Figure 1.

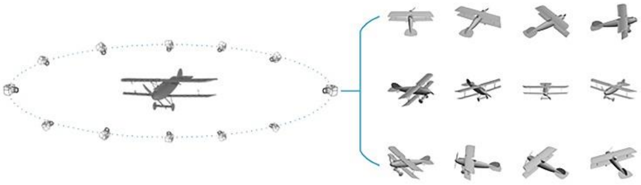

Figure 2:Render the model to get 12 views

3.2 Dynamic Attentive Graph Learning

In the MVCNN algorithm, only view pooling is performed between multiple views, and the relationship between them is ignored . Dynamic Attentive Graph Learning (DAGL)[18] was originally an algorithm for image restoration[18]. It is improved on the graph attention network. In order to obtain the connection between views and enhance the feature discriminative ability, we added the DAGL module to the algorithm.

First, the features obtained from ResNet34 are $N \times D$, where N represents N views, and we divide each view feature into K blocks, with a total of $L = N \times K$ image blocks. The biggest difference from GAT is that GAT fixes the edges in advance. While in DAGL module, image blocks represent nodes, and edges are dynamically established between nodes with high similarity.

In order to construct edges between nodes, DAGL dynamically selects neighbor nodes for each node according to the similarity principle. Firstly, we use the dot product to calculate the similarity value between nodes, and generate a similarity matrix M, where M_{ik} represents the similarity between node g_i and g_k. Secondly, calculate the average of each row of the matrix M, then the average of the i -th row represents the similar average of node i with other nodes. Finally, for the i -th row, the node whose similarity value is greater than the average is the neighbor of the i-th node. So, we name this threshold matrix T, The formula for calculating T_i is as follows:

$$T_i = \frac{1}{N}\sum_{k=1}^{N} M_{i,k} \quad (1)$$

In order to enhance the adaptability, we add the affine transformation of nodes to calculate threshold matrix $T_i{}'$:

$$T_i{}' = \phi_1(g_i)T_i + \phi_2(g_i) = \gamma T_i + \beta \quad (2)$$

Where, $g_i \in R^d$, $d = D/K$, $\phi_1$and ϕ_2 are fully connected layers whose weights are not shared, with a size of $d \times 1$.Then, two affine transformation parameters(γ, β) are obtained. DAGL uses the ReLU function to keep edges whose similarity is greater than a threshold, and remove edges whose similarity is less than the threshold, as shown in the formula(3):

$$A_i = ReLU(M_{i,:} - T_i{}') \quad (3)$$

$$g_i{}' = \sum_{j \in N_i} softmax(A_{ij})\, g_i \quad (4)$$

Where $A \in R^{L*L}$ is the adjacency matrix in which A_{ij} represents the similarity weight between node i and node j. If the similarity is less than the threshold, A_{ij} is equal to zero. Finally, DAGL uses the softmax function to normalize the matrix A, and performs feature aggregation according to adjacent nodes N_i as shown in the formula(4). After the above process, the feature blocks are spliced into the previous spatial dimension and get the new multi-view feature f_i.

3.3 Attention-NetVlad

In order to further improve the expressiveness of features, we combine channel attention with the NetVlad[20] algorithm. Efficient Channel Attention (ECA)[21][21] can learn the inter-channel information of view features, so that the feature can focus on its adjacent K channel features and further improve the retrieval performance. The calculation process of the channel attention weight ω_i is shown in the formula(5):

$$\omega_i = \sigma\left(\sum_{j=1}^{k} \alpha^j f_i^j\right) \quad (5)$$

$$z_i = f_i * \omega_i \quad (6)$$

where k is obtained by adaptive calculation, α is shared parameter, f_i^j represents the j -th neighbor of the i -th channel feature, $i = 1,2,\dots,D$, $j = 1,2,\dots,k$, D represents feature channel dimension. σ is the activation function. Further, the channel attention weights are multiplied by the original features to obtain the updated features z_i.

NetVlad is an algorithm that fuses local features into global features in the image field. It calculates the residual weighted sum of local features and class center features as a global special diagnosis.It can fuse valid information together, remove redundant information, and further improve the ability to distinguish features.Given local feature z_i, the feature dimension is N, the K cluster centers c_k. and finally the global feature of $K \times D$ -dimensiona is obtained. This process can be as Equation (7).

$$V(j,k) = \sum_{i=1}^{N} a_k(z_i)(z_i(j) - c_k(j)) \quad (7)$$

Among them, $z_i(j)$ and $c_k(j)$ represent the j-th eigenvalue of the i -th local descriptor and the k -th cluster center. $a_k(z_i)$ is the weight of the i -th local feature belonging to the k -th cluster. In order to make the algorithm backpropagation, NetVlad uses soft assignment to obtain $a_k(z_i)$, the formula is calculated as follows:

$$a_k(z_i) = \frac{e^{w_k^T z_i + b_k}}{\sum_k e^{w_k^T z_i + b_k}} \quad (8)$$

where parameters W, b and $c_k(j)$ in the formula are trainable. Through NetVlad, the output view features are $K \times D$.We further obtain concise and discriminative global features through max pooling, and the final feature dimension is $1 \times D$

4 EXPERIMENTS

4.1 Experimental configuration

Our experiments use the ResNet34 network structure,and the optimizer is Adam with a weight decay of 0.001.The initial learning rate is 5e-5. It is reduced to 0.01 every 10 epochs. Each experiment is trained for 60 epochs with softmax loss. All experiments were run on GPU 1650 and CPU AMD epyc 7543.

This paper verifies the performance of the algorithm on two benchmark datasets.

ModelNet10 contains 4899 CAD models in 10 categories, a total of 3991 models in the training set, and 908 models in the test set.

ModelNet40[22] contains 12311 3D models well annotated with multi-category labels in 40 categories, where 9843 models are used for training and 2468 models are used for testing.

4.2 Comparison on ModelNet40

To verify the performance of the method, further extensive experiments are conducted to compare with other state-of-the-art methods.We have selected a variety of well-known algorithms, including SPH[23], LFD[24] and 3DshapeNets model-based algorithms multimodal algorithms of PVNet[25], PVRNet[26] and MMFN[27], and view-based methods such as DAN etc.The comparison works with other well-known methods are shown in

Table 1, which the proposed method has shown good performance in retrieval. On ModelNet40, the retrieval mAP reaches 91.1%, which is better than other algorithms.In general, the algorithm in this paper has achieved significant gains compared with others.

Table 1: Algorithm performance comparison on ModelNet40

Method	Train Config		Number of views	Retrieval (mAP)
	Pre train	Fine tune		
SPH[23]	-	-	-	33.3%
LFD[24]	-	-	-	40.9%
3DshapeNets[6]	ModelNet40	ModeNet40	Voxels	49.2%
MVCNN[7]	ImageNet 1K	ModeNet40	12	80.2%
CNN+LSTM[28]	ImageNet 1K	ModeNet40	12	84.3%
GVCNN[11]	ImageNet 1K	ModeNet40	12	85.7%
Triplet-Center loss[29]	ImageNet 1K	ModeNet40	12	88.0%
SeqViews2SeqLabels[30]	-	ModeNet40	12	89.09%
PVNet[25]		ModeNet40	Multimodality	89.5%
PVRNet[26]		ModeNet40	Multimodality	90.5%
MMFN[27]		ModeNet40	Multimodality	90.3%
DAN[14]	ImageNet 1K	ModeNet40	12	90.4%
Ours	ImageNet 1K	ModeNet40	12	**91.1%**

4.3 Comparison on ModelNet10

Similarly, in order to further verify the performance of the experiments, we conduct comparative experiments on the subclass dataset ModelNet10 of ModelNet40. We also selected some methods, including traditional methods SPH, LFD, etc., as well as advanced methods SVHAN, DAN, etc. In Table 2 we find that our mAP achieves 93.6%, an increase of nearly 25.4% over 3DshapeNets, an increase of nearly 1% over the state-of-the-art algorithm SVHAN, and an increase of 1.3% over DAN. In general, our algorithm makes retrieval performance has been further improved.

Table 2: Algorithm performance comparison on ModelNet10

Method	Train Config		Number of views	Retrieval (mAP)
	Pre train	Fine tune		
SPH[23]	-	-	-	44.05%
LFD[24]	-	-	-	49.82%
3DshapeNets[6]	ModelNet10	ModeNet10	Voxels	68.26%
SeqViews2SeqLabels[30]	-	ModeNet10	12	91.4%
SVHAN[15]	ImageNet 1K	ModeNet10	12	92.7%
DAN[14]	ImageNet 1K	ModeNet10	12	92.3
Ours	ImageNet 1K	ModeNet10	12	**93.6%**

4.4 Analysis of results

To demonstrate the effectiveness of each module of our algorithm, we conduct a series of ablation experiments on ModelNet40, as shown in Table 3. In the first row, according to the MVCNN algorithm, after using ResNet34 as the backbone network to extract features, the features are fused according to the max pooling method, and the retrieved mAP is 82.7%. In our work, it is found that L2 normalization of features before retrieval can reduce the influence of larger and smaller values in the feature on distance calculation, and effectively improve the performance, with mAP of 87%, as shown in row 2.

Table 3: Ablation experiments for the proposed network on ModelNet40

Mode	ResNet18	ResNet34	DAGL	Attention-NetVlad	L2-norm	Retrieval (mAP)
1	-	✓	-	-	-	82.7%
2	-	✓	-	-	✓	87%
3	-	✓	✓	-	✓	89.2%
4	-	✓	-	✓	✓	89.8%
5	-	✓	✓	✓	✓	91.1%
6	-	✓	✓	NetVlad	✓	90.6%
7	✓	-	✓	✓	✓	90.4%

Effectiveness of DAGL Dynamic attentive graph learning inherits the idea of the graph attention network, and dynamically selects relevant features, which can effectively learn the relationship between views. As shown in Table 3, adding the DAGL module on the basis of the second row, the result reaches 89.2%, an improvement of 2.2%. Therefore, it can be proved that the DAGL algorithm is effective for model retrieval.

Effectiveness of Attention-NetVlad In our experiment,we propose the Attention-NetVlad algorithm, which fuses view channel information and inter-view information to output a discriminative global information. Adding the Attention-NetVlad module on the basis of the second row of Table 3, the retrieval mAP reaches 89.8%, with a 2.8% increase, as shown in the fourth row. Finally, our algorithm uses ResNet34 as the backbone network to extract features and performs DAGL and Attention-NetVlad.The final experimental result reaches 91.1%. In addition, in order to reflect the effect of channel attention on our algorithm, as shown in the sixth row of Table 3, we only use NetVlad, and compared with the fifth row, the retrieval mAP is reduced by 0.5% . It can be seen that the channel attention can improve the retrieval performance.

Effectiveness of backbone network . As shown in rows 5 and 7 of Table 3, we can see that selecting the ResNet34 network to extract basic features improves the retrieval mAP by 0.7% compared to ResNet18. So we choose ResNet34 as the backbone network for extracting basic features

The PR curves (Precision-Recall, PR) and category mAP of several algorithms are obtained, as shown in Figure 3 and Figure 4.

4.5 Analysis of Parameter

Table 4: Parameter experiments on ModelNet40

K	Retrieval (mAP)
2	90.8%
4	91.1%
8	90.2%

In DAGL, each view feature needs to be divided into K blocks. In order to evaluate the influence of the value of K on the retrieval performance, experiments are carried out when K is 2, 4, and 8. Table 4 shows the changes of mAP retrieved by the algorithm in this paper when K takes different values. We conduct experiments on the ModelNet40 dataset. From the results in Table 3, the retrieval performance is the highest when K is equal to 4. Therefore, our experimental parameter K is set to 4.

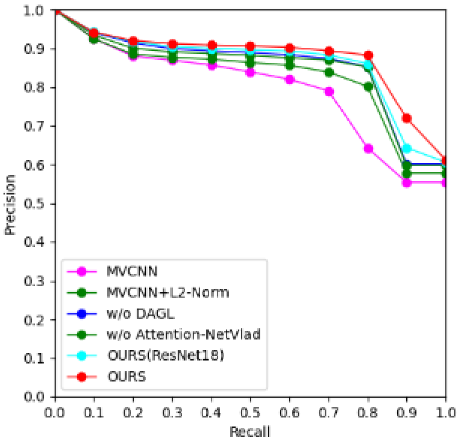

Figure 3: Comparison on PR curves

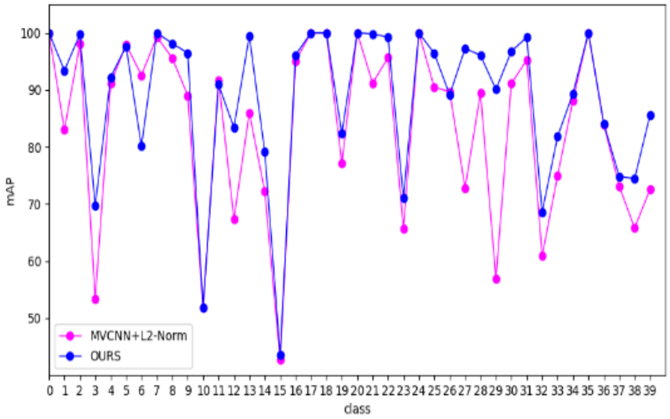

Figure 4: Comparison on category mAP

5 CONCLUSION

In this study, a 3D model retrieval algorithm based on attention and multi-view fusion is proposed. The dynamic attentive graph learning algorithm is mainly used for inter-view relationship learning to enhance the discriminative ability of features. In addition, the Attention-NetVlad algorithm is proposed, which combines channel attention and NetVlad algorithm to fuse multi-view local features into an effective global feature. Finally, the feature is calculated according to the Euclidean distance. Due to the advantages of the algorithm, we show excellent performance in 3D model retrieval work. Evaluations on ModelNet10 and ModelNet40 datasets are provided to show superior results in retrieval precision compared to other methods.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of Hebei Province (F2019202054).

REFERENCES

[1] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo and Leonidas J. Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'17).IEEE, Honolulu, HI, USA. 77-85.https://doi.org/10.1109/CVPR.2017.16

[2] Ali Cheraghian and Lars Petersson.2019. 3DCapsule: Extending the Capsule Architecture to Classify 3D Point Clouds. In Proceedings of the Winter Conference on Applications of Computer (WACV'19). Waikoloa Village, HI, USA,1194-1202. https://doi.org/10.1109/WACV.2019.00132

[3] Hongsen Liu, Yang Cong, Chenguang Yang and Yandong Tang. 2019. Efficient 3D object recognition via geometric information preservation.Pattern Recognit 92, (2019),135-145. https://doi.org/10.1016/j.patcog.2019.03.025

[4] Jianwen Jiang , Di Bao ,Ziqiang Chen , Xibin Zhao and Yue Gao. 2019. MLVCNN: Multi-loop-view convolutional neural network for 3D shape retrieval. In Proceedings of the Conference on Artificial Intelligence (AAAI2019). AAAI, Honolulu,Hawaii, USA,8513-8520. https://doi.org/10.1609/aaai.v33i01.33018513

[5] Heyu Zhou,An-An Liu, Weizhi Nie and Jie Nie. 2020. Multi-view saliency guided deep neural network for 3-d object retrieval and classification. IEEE Trans Multim 22,6 (2020) 1496–1506.https://doi.org/10.1109/TMM.2019.2943740

[6] Zhirong Wu , Shuran Song , Aditya Khosla , Fisher Yu , Linguang Zhang, Xiaoou Tang and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the Conference on Computer Vision and Pattern Recognition.(CVPR'15) IEEE.Boston, MA, USA ,1912–1920. https://doi.org/10.1109/CVPR.2015.7298801

[7] Hang Su, Subhransu Maji, Evangelos Kalogerakis and Erik G. Learned-Miller. 2015 Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the International Conference on Computer Vision(ICCV'15). IEEE. Santiago, Chile, 945-953 https://doi.org/10.1109/ICCV.2015.114

[8] Cheng Wang, Ming Cheng, Ferdous Sohel, Mohammed Bennamoun andJonathan Li. 2019. Normalnet: A voxel-based cnn for 3d object classification and retrieval. Neurocomputing 323,JAN.5 (2019),139–147. https://doi.org/10.1016/j.neucom.2018.09.075

[9] Takahiko Furuya, Ryutarou Ohbuchi. 2016. Deep aggregation of local 3d geometric features for 3d model retrieval.In Proceedings of the British Machine Vision Conference(BMVC'16).York, UK

[10] Sudhakar Kumawat and Shanmuganathan Raman. 2019. Lp-3dcnn: Unveiling local phase in 3d convolutional neural networks.In Proceedings of the Conference on Computer Vision and Pattern Recognition(CVPR'19). IEEE, Long Beach, CA, USA, 4898–4907

[11] Yifan Feng ,Zizhao Zhang ,,Xibin Zhao , Rongrong Ji and Yue Gao. 2018. Gvcnn: Group view convolutional neural networks for 3d shape recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition(CVPR'18).IEEE, Salt Lake City, UT, USA, 264–272

[12] Weizhi Nie, Shu Xiang and An-an Liu. 2018. Multi-scale cnns for 3d model retrieval. Multim Tools Appl 77,17 (2018), 22953–22963. https://doi.org/10.1007/s11042-018-5641-1

[13] Kai Sun , Jiangshe Zhang, Junmin Liu, Ruixuan Yu and Zengjie Song. 2021. DRCNN: Dynamic routing convolutional neural network for multi-view 3D object recognition. IEEE Trans. Image Process. 30 (2021), 868-877. https://doi.org/10.1109/TIP.2020.3039378

[14] Weizhi Nie, Yue Zhao, Dan Song and Yue Gao. 2021 Dan: deep-attention network for 3d shape recognition. IEEE Trans. Image Process, 30 (2021), 4371-4383. https://doi.org/10.1109/TIP.2021.3071687

[15] Yue Zhao, Weizhi Nie, An-An Liu, Zan Gao and Yuting Su. 2021. Svhan: Sequential view based hierarchical attention network for 3d shape recognition. In Proceedings of the Conference on Multimedia (MM'2021). ACM, China, 2130-2138. https://doi.org/10.1145/3474085.3475371

[16] Petar Velickovic ,Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio and Yoshua Bengio. 2017. Graph Attention Networks.CoRR. abs/1710.10903. (2017) http://arxiv.org/abs/1710.10903

[17] Joan Bruna, Wojciech Zaremba , Arthur Szlam and Yann LeCun. 2014. Spectral Networks and Locally Connected Networks on Graphs. In Proceedings of the International Conference on Learning Representations Computer Science (ICLR'14). Banff, AB, Canada. http://arxiv.org/abs/1312.6203

[18] Chong Mou, Jian Zhang and Zhuoyuan Wu. 2021. Dynamic Attentive Graph Learning for Image Restoration. In Proceedings of the International Conference on Computer Vision (ICCV'21). IEEE, Montreal, QC, Canada, 4308—4317. https://doi.org/10.1109/ICCV48922.2021.00429

[19] Kaiming He , Xiangyu Zhang , Shaoqing Ren and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'16) IEEE. Las Vegas, NV, USA,770-778 https://doi.org/10.1109/CVPR.2016.90

[20] Relja Arandjelovic , Petr Gronat , Akihiko Torii, Tomas Pajdla and Josef Sivic. 2016. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'16). IEEE Las Vegas, NV, USA, 5297—5307. https://doi.org/10.1109/CVPR.2016.572

[21] Qilong Wang, Banggu Wu , Pengfei Zhu , Peihua Li , Wangmeng Zuo and Qinghua. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'20) IEEE Seattle, WA, USA 11531—11539

[22] P.University. 2015. ModelNet40 Retrieved November 20, 2020 from http://modelnet.cs.princeton. edu/

[23] Michael M. Kazhdan ,Thomas A. Funkhouser and Szymon Rusinkiewicz.2003. Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. In Proceedings of the Eurographics Symposium on Geometry Processing.ACM, Aachen, Germany,156-164. https://doi.org/10.2312/SGP/SGP03/156-165

[24] Ding-Yun Chen, Xiao-Pei Tian , Yu-Te Shen and Ming Ouhyoung. 2003. On visual similarity based 3d model retrieval. Comput Graph Forum 22, 3 (2003), 223-232. https://doi.org/10.1111/1467-8659.00669

[25] Haoxuan You, Yifan Feng, Rongrong Ji and Yue Gao. 2018. PVNet: A Joint Convolutional Network of Point Cloud and Multi-View for 3D Shape Recognition. In Proceedings of the Multimedia Conference (MM'18). ACM, Seoul, Republic of Korea, 1310-1318 https://doi.org/10.1145/3240508.3240702

[26] Haoxuan You, Yifan Feng, Xibin Zhao, Changqing Zou , Rongrong Ji and Yue Gao. 2019 PVRNet: Point-View Relation Neural Network for 3D Shape Recognition . In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'19). Honolulu, Hawaii,USA.9119-9126. https://doi.org/10.1609/aaai.v33i01.33019119

[27] Weizhi Nie, Qi Liang, Yixin Wang, dXing Wei and Yuting Su. 2021. MMFN: Multimodal Information Fusion Networks for 3D Model Classification and Retrieval. ACM Trans. Multim. Comput. Commun. Appl.16,4 (2021) 131:1-131:22. https://doi.org/10.1145/3410439

[28] Chao Ma, Yulan Guo , Jungang Yang and Wei An. 2019. Learning Multi-View Representation With LSTM for 3-D Shape Recognition and Retrieval. IEEE Trans. Multim.21,5 (2019), 1169—1182. https://doi.org/10.1109/TMM.2018.2875512

[29] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai and Xiang Bai. 2018. Triplet-Center Loss for Multi-View 3D Object Retrieval. In Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'18). IEEE, Salt Lake City, UT, USA,1945-1954

[30] Zhizhong Han, Mingyang Shang , Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicke, Junwei Han and C. L. Philip Chen. 2019. SeqViews2SeqLabels: Learning 3D Global Features via Aggregating Sequential Views by RNN With Attention. IEEE Trans. Image Process, 28,2(2019) 658-672. https://doi.org/10.1109/TIP.2018.2868426