Multimedia Tools and Applications (2021) 80:16173–16184
https://doi.org/10.1007/s11042-019-08552-7

3D shape recognition based on multi-modal information fusion

Qi Liang[1] · **Mengmeng Xiao**[1] · **Dan Song**[1]

Received: 29 April 2019 / Revised: 3 September 2019 / Accepted: 27 November 2019 /
Published online: 23 January 2020

Abstract
The classification and retrieval of 3D models have been widely used in the field of multimedia and computer vision. With the rapid development of computer graphics, different algorithms corresponding to different representations of 3D models have achieved the best performance. The advances in deep learning also encourage various deep models for 3D feature representation. For multi-view, point cloud, and PANORAMA-view, different models have shown significant performance on 3D shape classification. However, There's not a way to consider utilizing the fusion information of multi-modal for 3D shape classification. In our opinion, We propose a novel multi-modal information fusion method for 3D shape classification, which can fully utilize the advantage of different modal to predict the label of class. More specifically, the proposed can effectively fuse more modal information. it is easy to utilize in other similar applications. We have evaluated our framework on the popular dataset ModelNet40 for the classification task on 3D shape. Series experimental results and comparisons with state-of-the-art methods demonstrate the validity of our approach.

Keywords 3D shape · Classification · Multi-view · Multi-modal

1 Introduction

With the rapid development of science in recent years, 3D technology has been widely used in industrial design, medical industry, architectural design, aerospace, automotive manufacturing, education, film and television animation, and other fields. The categories and numbers of 3D models are increasing year by year, and the classification of 3D models has become a popular trend in terms of computer vision. The 3D model adds structural information compared to 2D images. In the development of 3D model classification, multi-view, point cloud, PANORAMA, mesh, voxel, etc. are all 3D shape representation methods.

Multi-view is an object that is photographed at a fixed angular interval by tilting the virtual camera down 30 degrees. In other words, Multi-view is a set of series of 2D views. Su et al. [27] input multiple views into a series of convolutional neural networks to obtain

✉ Mengmeng Xiao
xmm_minnie@163.com

[1] The School of Electrical and Information Engineering, Tianjin University, Tianjin, China

descriptors corresponding to the model. Using this descriptor to classify and retrieve 3D models, it achieves good results on popular data sets. But this descriptor only contains the visual information of the model. The point cloud is a uniform point-like processing of the 3D model. PointNet inputs the original point cloud into the network to maximize the spatial characteristics of the point cloud, while the data volume is small and the 3D model can be efficiently classified [20]. Sfikas et al. [25] used PANORAMA to enter the convolutional neural networks to effectively classify 3D models. PANORAMA consists of a series of images that represent the vision and structure information. Mesh data of 3D shapes is a collection of vertices, edges, and faces. A 3D model can obtain several meshes, and the mesh data has complexity and irregularity [7]. Voxel is the smallest unit of 3D model segmentation. Similar to the pixel concept of 2D images, it is very complex and is not the popular method.

All of the above methods are classified models by single modal so that none of these methods make good use of the structural and visual information of the 3D model. It is a natural thought that these feature vectors have the same or similar parts. Thus, these feature vectors contain different modal information, we can take advantage of each other in the prediction step to robust classification. In order to demonstrate our idea, we propose a novel Multi-Modal classification Network for 3D shape classification in this work, we use multi-view, point cloud, and PANORAMA to represent the visual, structural, and surface information of the model. Then use MVCNN, PointNet, and PANORAMA-MVCNN to get the predict-scores of the 3D model classification. Finally, we propose an effective fusing approach to fuse the classification results of different models for the final 3D model classification with different weights.

The main contributions of this paper are summarized in the following two aspects:

- We propose a new multi-modal classification network, which uses predict-scores fusion and different weighting coefficients to obtain more accuracy classification results.
- The popular dataset is used to demonstrate the performance of the proposed method. Several classic methods are used for comparison. The final experiment also demonstrates the superiority of our approach.

The rest of the paper is structured as follows. Section 2 introduces some of the work related to 3D classification. Section 3 illustrates the details of the multi-modal classification network. Section 4 gives a lot of experimental results and discussion. Section 5 describes the relevant details at work. Finally, Section 6 draws conclusions.

2 Related work

3D shape recognition can be divided into four parts according to the different modalities, mesh-based methods, volume-based methods, view-based methods, and multi-modal fusion methods.

- **Mesh-based methods**: Mesh data of 3D shapes consist of vertices, edges, and faces, and the mesh has a stronger 3D shape description capability than other data.

  Socher et al. [26] use a model of combined convolutional neural network and recursive neural network to learn the characteristics of RGB-D images for classification. Novotny et al. [18] use a joint view decomposition meshes to align objects by detecting unsupervised ways of moving. Hubeli and Gross[10] design a semi-automatic framework to extract surface meshes features that require users to manually enter parameters

and operators. Kokkinos et al. [15] solve the direction ambiguity problem by constructing a shape context (ISC) meta-descriptor based on the development within the 3D shape. Feng et al. [7] propose a MeshNet based on face-unit and feature splitting to solve the complexity and irregularity of traditional meshes.

- **Volume-based methods**: Representing 3D models in voxel and point cloud first, the convolution operation can be performed like a two-dimensional image.

  Wu et al. [30] use the convolution depth belief network to represent the geometric 3D shape as the probability distribution of the binary variables on the 3D voxel grid and realize the active recognition object through view planning. Brock et al. [1] train voxel variational autoencoders to provide the possibility of voxel representations in model classifications. Qi et al. [20] design a network that directly applied to point cloud data, which has a good effect on 3D model classification. However, PointNet cannot capture the local structure, which results in low accuracy in identifying fine-grained patterns and poor ability to generalize complex scenes. Qi et al. [22] propose PointNet++ that learns local features by processing a set of points sampled in the metric space in a hierarchical manner.
- **View-based methods**: Compared with 3D data features, computer vision processing 2D images is more mature and faster, so it is necessary to reduce the dimensionality of 3D models. The earliest work based on view recognition 3D model was that Murase and Nayar [17] get a large number of images by the object changing posture and lighting conditions to form the certificate space, and classify the 3D model by matching the appearance. Kanezaki et al. [11] design otationNet based on convolutional neural network, which uses partial multi-views for reasoning. The viewpoints train object data sets in an unsupervised learning manner, select perspectives, estimate categories and poses. Su et al. [27] propose multi-view CNN (MVCNN), which uses different perspective images as raw data and inputs novel CNN framework training to obtain high-accuracy shape descriptors. Sfikas et al. [25] input model PANORAMA into the convolutional neural network for 3D model classification based on the fisher vector. Schneider et al. [24] provide different benchmarks for different sketches, perform classification-driven analysis, and extract the semantic features of the sketch.
- **Multimodal Fusion Methods**: Each 3D classification method has a good performance, and different methods are multiplied by different weights to achieve mutual compensation.

  Chen et al. [3] use a 3D point cloud to generate 3D candidate frames in autonomous driving scenarios, and then multi-view features obtained by fusing Frac-talNet and Deeply-Fused Net. Gonzalez et al. [9] use the detection to obtain RGB spectral map and depth image fusion to improve the accuracy of identifying two-dimensional objects. Enzweiler et al. [6] detect pedestrians by blending features, clues, and improve classification goals. You et al. [32] propose PVnet. It combines point cloud and multi-view data to compensate each other.

3 Method

Figure 1 shows the framework of our work, which mainly includes three steps: 1) Multimodal data generation: we utilize OpenGL to extract visual and PANORAMA information and employ Point cloud to extract point cloud information for each 3D model; 2) Multimodal network learning: it is used to get the predict-scores of 3D model based on different modalities. we trained the network on the single modal independently, and we get the best

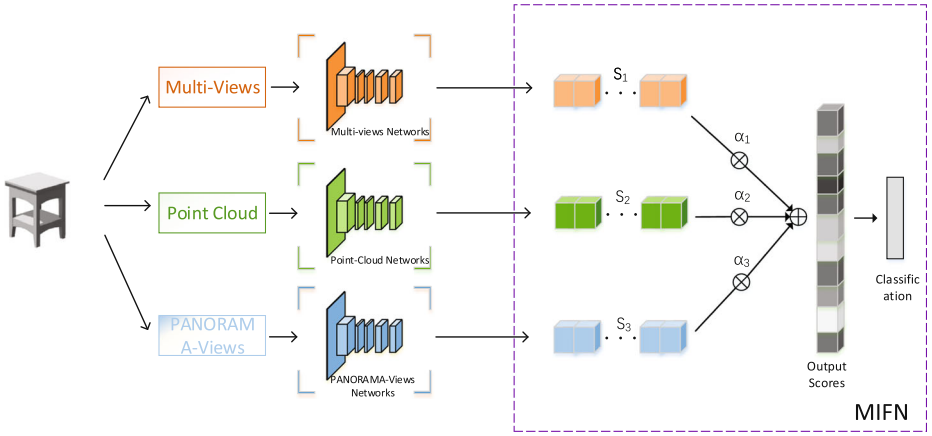

Fig. 1 Our MIFN framework is composed of 4 parts: multi-view network, point cloud network, PANORAMA-view network and predict-score fusion part. Multi-view network: The structure of MVCNN is employed, the view pooling layer that conducts max-pooling across all views. Point cloud network: The classic PointNet structure is employed. This network takes n points with 3-dimensional coordinates as input. Then in spatial transform net, a 3×3 matrix is learned to align the input points to a canonical space. For EdgeConv, it extracts the local patches of each point by their k-nearest neighborhoods and computes edge features for each point by applying a 1×1 convolution with output channels M', and then generates the tensor after pooling among neighboring edge features. PANORAMA-view network: It also utilizes the structure of MVCNN. However, we retrain the parameter of MVCNN based on PANORAMA view data. The predict-score fusion part: based on the predict-score produced by the above three networks, this fusion part defines the weight of different modal predict-score by statistic experiment and utilize the advantage of different modal predict-score for a better classification result

performance on each modal; 3) Classification Fusion: we propose an effective classification fusion method to utilize the advantages of different classifiers for a more accurate classification result of the 3D model. In the next part, we will detail these three steps

3.1 Data processing

Multi-view(MV Modal) Since the size and angle of the 3D model data are not uniform, we first use the NPCA[19] to calibrate the 3D model. Then place a virtual camera every 30 degrees from the Z-axis around the 3D model. The lens points to the 3D model centroid and tilts down 30 degrees. Finally, get twelve views of the three-dimensional model by OpenGL visualization tools.

Point cloud(PC modal) The point cloud data is obtained by meshing the surface of the 3D model and using the centroid of the grid to represent the mesh. Due to the different volume of the 3D model, we need to process the surface of each 3D model, which need to be subdivision to get more mesh. Here we use the butterfly subdivision [5] algorithm to get more points. At last, we obtain 1024 points of each 3D model and then convert the PLY model into point cloud data.

PANORAMA-view(PV modal) PANORAMA view is a set of 2D images that contains the surface information of a 3D model. When we get the PANORAMA view, we need to project the surface of the 3D model onto the surface of the cylinder with the centroid of the 3D model as the origin. The radius R of this cylinder is the three times the maximum distance

from the surface of the 3D model to the axis of the cylinder, and the hight of this cylinder is $2R$. Taking the Z-axis as an example, we use a series of point sets s(φ,z) to represent the projected data, where φ is the angle of the point of the 3D shape, and z is the height of the point. According to this point set, we can get a panorama of four different data for each axis. 1)the position of the model's surface in 3D Space(SDM). 2)the orientation of the model's surface(NDM). 3)the gradient map of NDM. 4)three-channel graphics consist of the above three kinds of graphics. So for each model, we can get 12 views for three-axis projections like Fig. 2.

3.2 MIFN:Multi-modal classification network

In order to utilize the advantages of different modalities of 3D shapes, we design the Multi-modal Information Fusion Network(MIFN), which uses a new strategy in the prediction part. Comparing with the traditional method, instead of using the prediction score of single modal, we refer to all modal information of the 3D shape. To make the final prediction correctly represent the 3D shape while maxing the precision of the prediction, we process each model according to the method of data preprocessing in Section 3.1. Each modal products its own modal-level prediction using its own network, and a consensus function is designed to aggregate these modal-level predictions into the final prediction scores of different classes named model-level prediction. This model-level prediction score is more reliable and informative than the original modal-level prediction since it aggregates three modality prediction results and gets more credible results. However, we train the single-modal network and update the parameters separately. In this way, the single modal network can get the best performance.

For a 3D shape M, we can get K modalities data $P_1, P_2, \cdots, P_K$ after preprocessing. Then, these modalities data are feed into their own networks $M_1, M_2, \cdots .M_3$. At last, Multi-modal Information Fusion Network (MIFN) aggregate the prediction scores of different modalities as follows:

$$MIFN(P_1, P_2, \cdots, P_K) = H(G(M(P_1 : W_1), M(P_2 : W_2), \cdots M(P_K : W_K))) \quad (1)$$

Here, W_K represents the parameters of M_k which is updating by feeding single modal data P_K into it and produce the modal-level prediction score. G is the aggregate function that combines the outputs from M_k to aggregate the modal-level prediction scores. Based on this

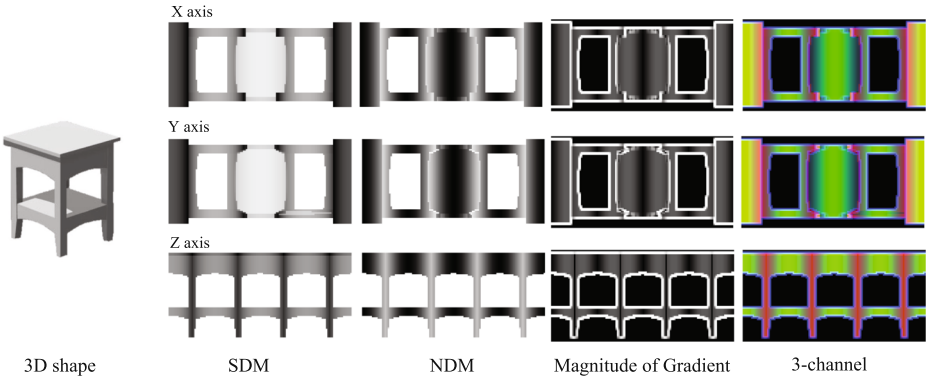

Fig. 2 The PANORAMA views of 3D model on three axis which consist of SDM, NDM gradient map of NDM and 3-channel images

aggregate result, we use Softmax function H predicts the probability of each class for the 3D shape. In our MIFN, the aggregate function is of great importance. It should retain useful information as soon as possible. Meanwhile, it should be able to treat each modality differently, since different modalities have different characteristics and have different advantages in category prediction. So we want this function can aggregate modality information, not a function that happens to a model. We will provide the details on function G in the next subsection.

3.3 Multi-modal information fusion

Through the above analysis, we know that aggregate functions are the most important part of our framework(MIFN). In this section, we will introduce our idea of designing aggregate functions. According to the different modal learning independently, we can get three modal-level prediction scores based on different modality data. These scores represent the probability that a single modality predicts a model as a class, but these are based on features learned by a single modal. Obviously, the reliability of a single modal prediction is not very high, we can make a simple addition on the predict-scores as equation (2). But in this way, we ignore the difference between multi-modal, so we use a linear weighted average method to fuse these modal-level prediction scores as equation (3). In this work, we employ the weighted fusion method to fuse the three modal-level prediction scores. The framework of this method is shown in Fig. 1. The detail is shown in (1).

$$S = \sum_{i=1}^{K} M(P_K : W_K) \tag{2}$$

$$S = \sum_{i=1}^{K} \alpha_i M(P_K : W_K) \tag{3}$$

where $M(P_K : W_K)$ represents the modal-level prediction score produced by MVCNN, PointNet and PANORAMA-MVCNN respectively based on different modalities of 3D model. α_i is the weight of different modal-level prediction scores in order to balance the Multi-view, Point cloud and PANORAMA-View. The fusion score is also processed by softmax to get the class label. The related experimental is shown in Section 4.2.

4 Experiment

4.1 Datasets

One well-known dataset was used to prove our ideas. It is ModelNet.ModelNet consists of two versions of this dataset and they are publicly available for download: ModelNet10 and ModelNet40. ModelNet10 comprises 4899 CAD models split into 10 categories. The training and testing subsets of ModelNet10 consist of 3991 and 908 models. ModelNet40 comprises 12,311 CAD models split into 40 categories. The training and testing subsets of ModelNet40 consist of 9843 and 2468 models. The ModelNet dataset is manually filtered to remove 3D models that do not belong to the specified category, but in particular, the pose in terms of translation and rotation of ModelNet10 dataset is normalized, and ModelNet40 dataset does not.

4.2 Comparison the combinations of different modal networks

In this work, we propose a novel predict-score fusion method to fuse the multi-modal information extracted by these different modal networks. The goal of this design is to utilize the advantages of different modal networks to get more accuracy classification results. In order to demonstrate the performance of this approach, we compare the classification results of a single modal network with the combinations of different modal networks. The corresponding experimental results are shown in Table 1. From this table, we can find that the combination of different modal networks brings a significant improvement in performance compared with a single modal network. Here, MV+PC brings a 4% and 1.5% improvement over MV and PC respectively. MV+PV brings a 2% and 6% improvement over MV and PC respectively. PC+PV brings a 0.25% and 6.5% improvement over PC and PV respectively. Finally, MV+PC+PV brings a 5%, 3% and 9% improvement over each single modal respectively. We can find that the PC network brings the biggest improvement under different conditions. Meanwhile, the single modal network PC also gets the best classification results compared with another single modal network. There are reasons to think that the point cloud data represents more information on 3D modal.

After the analysis of Section 3.3, we know that in order to let the different modalities play their respective advantages and maximize their advantages, we should take weights on different modalities and maximize the advantages of the modality. For example, in the previous experiment, we know that the point cloud works best. Naturally, we think that the power of the point cloud is a bit more important, and related experiments have proved this. From Fig. 3 we can see that when the weights of Point Cloud, Multi-view and PANORAMA-view is set to the parameter $\alpha_1 = 0.7$, $\alpha_2 = 0.2$ and $\alpha_3 = 0.1$, we can get the best results. Comparing with the method of directly averaging without increasing the weight, we get a gain of 0.54% after increasing the weight, and we can find that the modality PC has the biggest weight. This experiment result demonstrates the effectiveness of the proposed method.

4.3 Comparison to state-of-the-art methods on ModelNet-40

To validate the efficiency of the proposed MIFN, 3D shape classification experiments have been conducted on the Princeton ModelNet dataset [31]. Totally, 127,915 3D CAD models from 662 categories are included in the ModelNet dataset. ModelNet40, a common-used subset of ModelNet, containing 12,311 shapes from 40 common categories, is applied in our experiments. We follow the same training and test split setting as in [31].

Table 1 Comparisons of different Model Classification accuracy on ModelNet10 and ModelNet40

Method	Classification accuracy	
	ModelNet10	ModelNet40
MV	88.53%	86.93%
PC	91.24%	89.23%
PV	86.57%	82.69%
MV + PC	91.59%	90.83%
MV + PV	89.69%	88.85%
PC + PV	92.39%	89.54%
MV + PC + PV	92.97%	91.86%

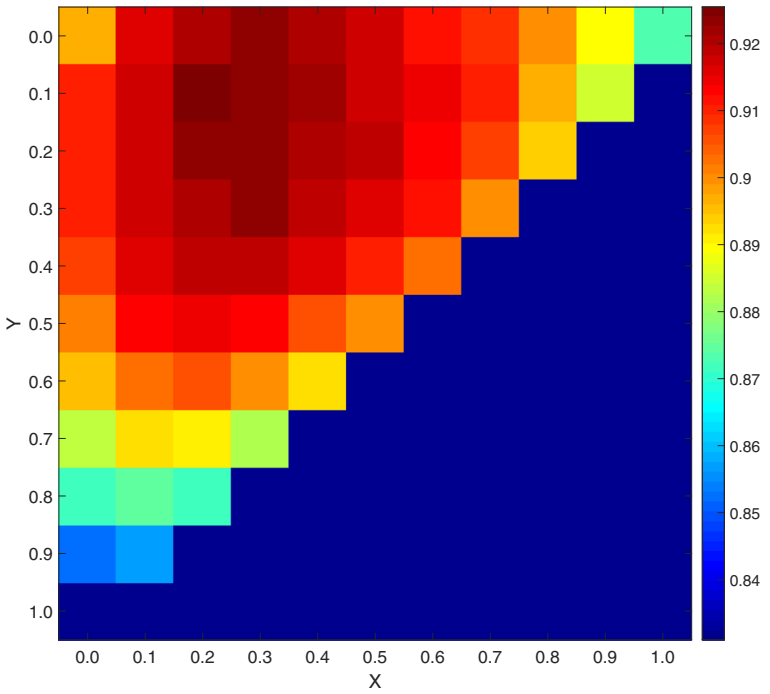

Fig. 3 The X axis represents the weight of the PANORAMA-view modality, the Y axis represents the weight of the Multi-view modality, the weight of the point cloud is determined by $1 - x - y$, and the colorbar represents the classification accuracy of the 3D model on ModelNet40

In experiments, we have compared the proposed MIFN with various models based on different representations, including volumetric based models [31], hand-craft descriptors for multi-view data [4, 12], deep learning models for multi-view data [21, 28], deep learning models for PANORAMA-Views [29] and point cloud based models [8, 13, 16, 23, 33].

In Table 2, the classification results of all compared methods are provided. As shown in the results, our proposed MIFN can achieve the best performance with the classification accuracy of 92.4%. Compared with the MVCNN using GoogLeNet, our MIFN achieves the state-of-art accuracy on the classification tasks. For point cloud based models, our MIFN also achieves the state-of-the-art point cloud based model DGCNN in terms of classification accuracy.

5 Implementation details

Our framework contains point cloud network, multi-view network, and PANORAMA-View network. For point cloud network, 1,024 raw points for each object are fed into the network. For Multi-View network, 12 views for each object are fed into the network. The parameters of CNN in multi-view network are initialized by the pre-trained MVCNN model. For PANAORAMA-View network, 12 views are fed into the network that same to Multi-View network, differently, the parameters aren't initialized. We pre-train the model on our dataset and find the best model to initialize the parameters. At last, we fuse the modal-level predict-score to generate model-level predict-score, we can use these scores to predict the class of the model.

Table 2 Comparisons of classification accuracy on ModelNet40

Method	Train config		Data representation	Classification
	Pre train	Fine tune	#Number of views	Accuracy
(1) SPH[12]	–	–	–	68.2%
(2) LFD[4]	–	–	–	75.5%
(3) 3D ShapeNets[31]	ModelNet40	ModelNet40	Volumetric	77.3%
(4) VoxNet[14]	ModelNet40	ModelNet40	Volumetric	83.0%
(5) VRN[2]	ModelNet40	ModelNet40	Volumetric	91.3%
(6) MVCNN-MultiRes[21]	–	ModelNet40	Volumetric	91.4%
(7) MVCNN,12×[28]	ImageNet1K	ModelNet40	12 Views	89.9%
(8) MVCNN,metric,12×[28]	ImageNet1K	ModelNet40	12 Views	89.5%
(9) MVCNN,80×[28]	ImageNet1K	ModelNet40	80 Views	90.1%
(10) MVCNN,metric,80×[28]	ImageNet1K	ModelNet40	80 Views	90.1%
(11) PointNet[8]	–	ModelNet40	Point Cloud	89.2%
(12) PointNet++[23]	–	ModelNet40	Point Cloud	90.7%
(13) KD-Network[13]	–	ModelNet40	Point Cloud	91.8%
(14) PointCNN[16]	–	ModelNet40	Point Cloud	91.8%
(15) DGCNN[33]	-	ModelNet40	Point Cloud	92.2%
(16) PANORAMA-NN[29]	–	ModelNet40	PANORAMA-Views	90.7%
(17) MIFN(Our)	**ImageNet1K & ModelNet40**	**ModelNet40**	**Point Cloud & 12 Views & PANORAMA-Views**	**92.4%**

6 Conclusion

In this paper, we propose a novel modal fusion network: MIFN, which can employ different modal data for 3D shape classification. In our framework, the model-level predict-scores is introduced to employ the advantage of different modal networks to predict the label of class. More specifically, the proposed can effectively fuse more modal information. it is easy to utilize in other similar applications. The effectiveness of our proposed framework has been demonstrated by experimental results and comparisons with the state-of-the-art models on the ModelNet dataset. We have also investigated the effectiveness of different components of our model to demonstrate the robustness of our framework.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (61472275, 61170239, 61303208, 61502337).

References

1. Brock A, Lim T, Ritchie JM, Weston N (2016) Generative and discriminative voxel modeling with convolutional neural networks. arXiv:1608.04236
2. Brock A, Lim T, Ritchie JM, Weston N (2016) Generative and discriminative voxel modeling with convolutional neural networks, Computer Science

3. Chen X, Ma H, Wan J, Li B, Xia T (2016) Multi-view 3d object detection network for autonomous driving. arXiv:1611.07759
4. Chen DY, Tian XP, Shen YT, Ming O (2010) On visual similarity based 3d model retrieval. Computer Graphics Forum 22(3):223–232
5. Dyn N, Levine D, Gregory JA (1990) A butterfly subdivision scheme for surface interpolation with tension control. ACM Transaction on Graphics 9:160–. https://doi.org/10.1145/78956.78958
6. Enzweiler M, Gavrila DM (2011) A multilevel mixture-of-experts framework for pedestrian classification. IEEE Trans Image Process 20(10):2967–2979. https://doi.org/10.1109/TIP.2011.2142006
7. Feng Y, Feng Y, You H, Zhao X, Gao Y (2018) Meshnet: Mesh neural network for 3d shape representation. arXiv:1811.11424
8. Garcia-Garcia A, Gomez-Donoso F, Garcia-Rodriguez J, Orts-Escolano S, Cazorla M, Azorin-Lopez J (2016) Pointnet: a 3d convolutional neural network for real-time object class recognition. In: 2016 International joint conference on neural networks (IJCNN), pp 1578–1584. https://doi.org/10.1109/IJCNN.2016.7727386
9. González A, Vázquez D., López A. M., Amores J (2017) On-board object detection: Multicue, multimodal, and multiview random forest of local experts. IEEE Trans Cyber 47(11):3980–3990. https://doi.org/10.1109/TCYB.2016.2593940
10. Hubeli A, Gross M (2001) Multiresolution feature extraction from unstructured meshes, vol 1. https://doi.org/10.1109/VISUAL.2001.964523
11. Kanezaki A (2016) Rotationnet: Learning object classification using unsupervised viewpoint estimation. arXiv:1603.06208
12. Kazhdan M, Funkhouser T, Rusinkiewicz S (2003) Rotation invariant spherical harmonic representation of 3D shape descriptors. In: Symposium on geometry processing
13. Klokov R, Lempitsky VS (2017) Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. arXiv:1704.01222
14. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems, vol 25. https://doi.org/10.1145/3065386
15. Kokkinos I, Bronstein M, Litman R, Bronstein A (2012) Intrinsic shape context descriptors for deformable shapes, pp 159–166. https://doi.org/10.1109/CVPR.2012.6247671
16. Li Y, Bu R, Sun M, Chen B (2018) Pointcnn. arXiv:1801.07791
17. Murase H, Nayar SK (1995) Visual learning and recognition of 3-d objects from appearance. Int J Comput Vis 14(1):5–24. https://doi.org/10.1007/BF01421486
18. Novotný D., Larlus D, Vedaldi A (2017) Learning 3d object categories by looking around them. arXiv:1705.03951
19. Papadakis P, Pratikakis I, Perantonis S, Theoharis T (2007) Efficient 3d shape matching and retrieval using a concrete radialized spherical projection representation. Pattern Recogn 40(9):2437–2452. https://doi.org/10.1016/j.patcog.2006.12.026
20. Qi CR, Su H, Mo K, Guibas LJ (2016) Pointnet: Deep learning on point sets for 3d classification and segmentation. arXiv:1612.00593
21. Qi CR, Hao S, Niessner M, Dai A, Guibas LJ (2016) Volumetric and multi-view cnns for object classification on 3d data
22. Qi CR, Yi L, Su H, Guibas LJ (2017) Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv:1706.02413
23. Qi CR, Li Y, Hao S, Guibas LJ (2017) Pointnet++: Deep hierarchical feature learning on point sets in a metric space
24. Schneider RG, Tuytelaars T (2014) Sketch classification and classification-driven analysis using fisher vectors. ACM Trans Graph 33(6):174:1–174:9. https://doi.org/10.1145/2661229.2661231
25. Sfikas K, Pratikakis I, Theoharis T (2018) Ensemble of panorama-based convolutional neural networks for 3d model classification and retrieval. Computers & Graphics 71:208–218. https://doi.org/10.1016/j.cag.2017.12.001
26. Socher R, Huval B, Bath BP, Manning CD, Ng AY (2012) Convolutional-recursive deep learning for 3d object classification, pp 665–673
27. Su H, Maji S, Kalogerakis E, Learned-Miller EG (2015) Multi-view convolutional neural networks for 3d shape recognition. arXiv:1505.00880
28. Su H, Maji S, Kalogerakis E, Learned-Miller E (2015) Multi-view convolutional neural networks for 3d shape recognition. In: 2015 IEEE international conference on computer vision (ICCV), pp 945–953. https://doi.org/10.1109/ICCV.2015.114
29. Sfikas K, Theoharis T, Pratikakis I (2017) Exploiting the PANORAMA representation for convolutional neural network classification and retrieval. In: Pratikakis I, Dupont F, Ovsjanikov M (eds) Eurographics workshop on 3D object retrieval, The Eurographics association. https://doi.org/10.2312/3dor.20171045

30. Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J (2015) 3d shapenets: A deep representation for volumetric shapes, pp 1912–1920. https://doi.org/10.1109/CVPR.2015.7298801
31. Wu Z, Song S, Khosla A, Yu F, Zhang L, Tang X, Xiao J, Wu Z, Song S, Khosla A (2015) 3d shapenets A deep representation for volumetric shapes. In: IEEE conference on computer vision & pattern recognition
32. You H, Feng Y, Ji R, Gao Y (2018) Pvnet: a joint convolutional network of point cloud and multi-view for 3d shape recognition. arXiv:1808.07659
33. Yue W, Sun Y, Liu Z, Sarma SE, Bronstein MM, Solomon JM (2018) Dynamic graph cnn for learning on point clouds

Qi Liang received the B.S. degrees in electronic science and technology from Taiyuan University of Technology. His research interests include 3D shape recognition, cross domain learning and Data mining.

Mengmeng Xiao received the B.S. degrees in electronic engineering from Tianjin University of China. Her research interests include 3D shape recognition and cross domain learning.

Dan Song received the Ph.D. degree in computer science and technology from Zhejiang University of China. Her research interests include computer graphics, computer vision, 3D human body reconstruction and virtual fitting.