MMJN: Multi-Modal Joint Networks for 3D Shape Recognition

Weizhi Nie*
Tianjin University
truman.nie@gmail.com

Qi Liang
Tianjin University
tjuliangqi@tju.edu.cn

An-An Liu*
Tianjin University
anan0422@gmail.com

Zhendong Mao
University of Science and Technology of China
maozhengdong2008@gmail.com

Yangyang Li
National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC), CAEIT
liyangyang@cetc.com.cn

ABSTRACT

3D shape recognition has attracted wide research attention in the field of multimedia and computer vision. With the recent advance of deep learning, various deep models with different representations have achieved the state-of-the-art performances. Among them, many modalities are proposed to represent 3D model, such as point cloud, multi-view, and PANORAMA-view. Based on these representations, many corresponding deep models have shown significant performances on 3D shape recognition. However, few work considers utilizing the fusion information of multiple modalities for 3D shape recognition. Since different modalities represent the same 3D model, they should guide each other to get a better feature representation. In this paper, we propose a novel multi-modal joint network (MMJN) for 3D shape recognition, which can consider the correlation between different modalities to extract the robust feature vector. Specifically, we propose a novel correlation loss which can utilize the correlation between different features extracted by different modality networks to increase the robustness of feature representation. Finally, we utilize the late fusion method to fuse multi-modal information for 3D shape representation and recognition. Here, we define the weight of different modalities based on the statistic method and utilize the advantages of different modalities to generate more robust feature. We evaluated the proposed method on the ModelNet40 dataset for 3D shape classification and retrieval tasks. Experimental results and comparisons with the state-of-the-art methods demonstrate the superiority of our approach.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision; 3D imaging**; • **Information systems** → Information retrieval.

KEYWORDS

3D Shape Recognition, Multi-View, Multiple Modalities, PANORAMA-View

ACM Reference Format:
Weizhi Nie, Qi Liang, An-An Liu*, Zhendong Mao, and Yangyang Li. 2019. MMJN: Multi-Modal Joint Networks for 3D Shape Recognition. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351009

*Corresponding Author: Weizhi Nie and An-An Liu, Email:truman.nie@gmail.com, anan0422@gmail.com

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351009

1 INTRODUCTION

With the development of digitization techniques and computer vision, 3D models are widely used in our daily life, such as computer-aided design, medical diagnoses, bioinformatics, 3D printing, medical imaging and digital entertainment. How to automatically recognize 3D shapes has attracted much attentions in recent years. With the development of advanced sensors, various modalities have been employed to represent 3D models, such as multi-view, point cloud, sketch image and PANORAMA image. Thus, it is natural and reasonable to utilize different approaches to learn the representation of 3D models based on multi-modal information.

MVCNN [34] extracts a collection of 2D view images by rendering the 3D model, and combines information from multiple views of a 3D shape into a single and compact shape descriptor. PointNet [9] uses density occupancy grids representations for the 3D point cloud data, and PointNet++ [27] recursively put it into a hierarchical neural network to get a representation of 3D shape. The PANORAMA representation is consecutively extracted by posing normalized 3D models using the SYMPAN method. The panoramic views consist of 3-channel images, containing the Spatial Distribution Map, the Normals' Deviation Map and the magnitude of the Normals' Devation Map Gradient Image [30, 31]. The sketch modality utilizes the sketch information to represent a 3D model. [38] proposes a novel network to extract the sketch information for 3D model representation. Sketch information can effectively handle the shape changes due to scale changes. However, all of these approaches only focus on the single modality of 3D data and ignore the correlation among them.

3D models can be represented by different modalities, and various approaches have tried to learn 3D representation using these modalities. Since these features represent the same 3D model, it is intuitive that these features have strong correlation. Therefore,

they can guide each other in the training step for more robust representation. In this paper, we propose a novel Multi-Modal Joint Network (MMJN) for 3D shape recognition. First, we extract the data of three modalities for each 3D model. Specifically, we utilize point cloud and multi-view data to represent the structure and visual information of the 3D model, respectively. Meanwhile, we utilize PANORAMA to represent the surface information of 3D model. Second, we utilize PointNet, MVCNN and PANORAMA-MVCNN to learn the feature vectors of 3D model, respectively. We design a novel correlation loss, which can effectively mitigate the distribution discrepancy across different modalities, guide the feature learning and increase the learning rate in training step. Finally, we propose an effective fusing approach to fuse the features of different modalities for final recognition.

The contributions of this paper include:

- We propose a novel Multi-Modal Joint Netwrok (MMJN) for 3D shape recognition. To the best of our knowledge, we are the first to consider multi-modal information fusion for 3D shape recognition;
- We propose a novel correlation loss to mitigate the distribution discrepancy across different modalities. This loss can effectively increase the learning rate and improve the robustness of feature representation learned by different networks;
- We propose an effective feature fusion method, which can define the weights of different networks to effectively utilize the advantage of multi-modal features for final classification;
- The popular dataset is used to validate the performances by the proposed method, and several classic methods are used for comparison. The final experiment demonstrates the superiority of our approach.

The rest of the paper is organized as follows. In Section II, we review the related work. Then, we introduce the detail of our methods in Section III. We will detail how to fuse different models by aggregation function with different weights in Section III. The experimental settings, results, and analysis are introduced in Section IV. Section V introduces the implementation details of our work. In section VI, we will conclude the paper.

2 RELATED WORKS

The number of different methods[5, 8, 11, 15, 33, 36, 37, 43] of 3D shape recognition has exploded in recent years. The researchers designed different convolutional neural networks, taking the preprocessed 3D data like voxels, image projections, raw point clouds, and graphs transformed from source data as input. Most 3D shape recognition methods are validated on the ModelNet10 and ModelNet40 datasets. In this section, the methods of learning 3D features by deep learning models are reviewed.

2.1 Mesh-based methods.

3D mesh composed of vertices which are connected by edges is an important raw representation for 3D shapes. There have been some studies on learning features from 3D meshes directly from raw 3D representations using deep learning models. Richard Socher et al. [32] proposed a model combining convolution and recursive neural networks (CNN and RNN) which is introduced for learning the features of RGBD images and classifying the corresponding 3-D shapes. To learn 3D local features, Han et al. [13] proposed a method that learns unsupervised 3D local features, and the features are expressed by the surface patterns which capture the common geometry and structure among the huge number of 3-D local regions. In order to learn global features, a novel deep learning model, mesh convolutional restricted Boltzmann machines (MCRBMs), is proposed for unsupervised feature learning for 3-D meshes by Han et al. [12]. They also proposed a deep context learner [14], a deep neural network with a novel model structure which encodes not only the discriminative information among local regions but also the one among global shapes. Feng et al. [7] proposed MeshNet which uses face-unit and feature splitting. In this way, MeshNet is able to solve the complexity and irregularity problem of mesh and conduct 3D shape representation well.

2.2 Volume-based methods.

In order to convolve 3D model just like any other tensor [9, 17, 27, 29, 35], many works based on voxelized shapes have been done. These methods are constrained by their resolution owing to data sparsity and costly computation of 3D convolution. Wu et al. [35] proposed 3D ShapeNets to learn global features from voxelized 3D shapes based on convolutional restricted boltzmann machine. In order to deal with problems of the additional computational complexity (volumetric domain) and data sparsity, Qi et al. [25] proposed two distinct network architectures of volumetric CNNs. PointNet [9] first proposed a method using deep neural networks to directly process point clouds, whereas the local features are ignored. In this regard , Qi et al. [27] proposed a method learning to extract point features and balance multiple feature scales in an end-to-end fashion. Klokov et al. [20] proposed a new architecture works with unstructured point clouds to avoid poor scaling behavior.

2.3 View-based methods.

The first view based 3D descriptor is Lighting Field Descriptor [3], and the similarity of 2D features of their corresponding two view sets is employed to measure the similarity between two 3D shapes. Similarly, GIFT [1] measures Hausdorff distance between their corresponding view sets. Recently Su et al. [34] proposed a multi-view convolutional neural network, which generates multiple 2D projection features learned by CNN within an end-to-end trainable fashion. In order to exploit the structural information in views of 3D shape, DeepPano [31] was proposed to learn features from PANORAMA views using CNN. Sfikas [30] proposed a method capturing PANORAMA views feature aiming at continuity of 3D shapes and minimizing data preprocessing via the construction of an augmented image representation. Zhang et al. [42] propose an inductive multi-hypergraph learning algorithm, which targets on learning an optimal projection for the multi-modal training data and geting the projection matrices and the optimal multi-hypergraph combination weights simultaneously.

2.4 Multi-modal Fusion Methods.

For 3D shapes, whether mesh-based methods, volume-based methods, or view-based methods can describe 3D shapes well separately. So we naturally think of using the fusion method to take advantage of each modality. Hegde et al. [16] proposed new Volumetric CNN

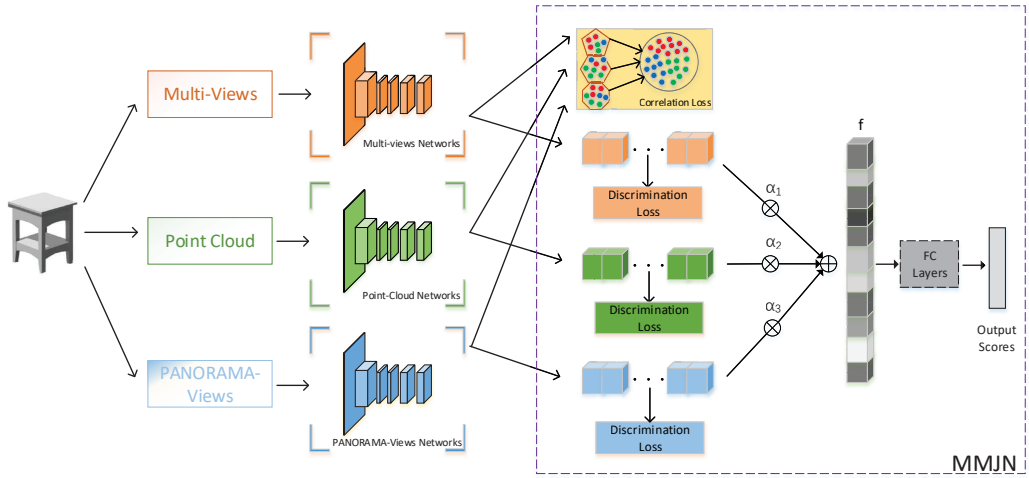

Figure 1: Our MMJN framework is composed of 4 parts: point cloud network, multi-view network, PANORAMA-view network and feature fusion part. Point cloud network: The classic PointNet structure is employed. This network takes n points with 3-dimensional coordinates as input. Then in spatial transform 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. Multi-view network: The structure of MVCNN is employed, and the view pooling layer conducts max pooling across all views. The PANORAMA-view network: It also utilizes the structure of MVCNN. However, we retrain the parameter of MVCNN based on PANORAMA view data. The classification fusion part: based on the feature vectors produced by the above three networks, this fusion part defines the weight of different modality features by statistic experiment and utilizes the advantage of different modality features for a better classification result.

(V-CNN) architectures to generate features learned from the two representations. In order to jointly classify object proposals and do oriented 3D box regression, Chen et al.[4] designed a region-based fusion network to effectively combine features from multiple views and point cloud data from LIDAR. In [10] they explored the fusion of RGB, depth maps and ranging for 2D pedestrian detection used in autonomous driving. Poria et al.[24] proposed attention based networks for improving both context learning and dynamic feature fusion to achieve attentive multi-modal fusion. Similarly, You et al. employed PVNet[39] to model the intrinsic correlation and discriminability of different structure features from the point cloud data using high-level features from the multi-view data. In addition, You et al.[40] proposed PVRNet, a novel multi-modal fusion network which takes full advantage of the relationship between point cloud and views.

3 OUR APPROACH

Figure 1 shows the framework of our work, which mainly includes three steps: 1) Multi-modal 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) Multi-modal joint network learning: it is used to extract the features of 3D model based on different modalities. Here, we propose a correlation loss to make these networks share the feature information, increase the final learning rate, and improve the robustness of feature; 3) Feature fusion: we propose an effective feature fusion method to utilize the advantages of different modality networks for a more robust feature of 3D model. In the next part, we will detail these three steps.

3.1 Data Processing

Multi-View (MV modality): The NPCA method [23] is used to normalize each 3D model. Then, the visual tool developed by OpenGL is utilized to extract a set of views from each 3D model like a human observer. These views wrap around the model and are extracted every 30 degrees around the Z axis. We can extract 12 views to represent the visual and structure information of the 3D model. These views can also be seen as a sequence of images.

Point Cloud (PC modality): We convert the PLY models into Point Cloud Data (PCD) clouds by Meshlab [28]. Because the size of the models in the dataset is not uniform, the mesh density of the model surface is also different. In order to get more dense point cloud data, we perform mesh subdivision on the loaded model by adding triangles, and indirectly increase the number of points. Here we use the butterfly subdivision algorithm [6], and we export the point cloud data for each model with 1024 points.

PANORAMA View (PV modality): The panoramic view of 3D model is proposed by [31]. Compared with a normal projection view, the panoramic view uses a 2D image to represent the structural information of a 3D model. We project the surface of the 3D model onto the side surface of the cylinder. The radius of the cylinder is set to the maximum distance between the model surface and the centroid, and the height is set to twice the radius. For example, we extract the panorama views of the Z-axis consist of a set of

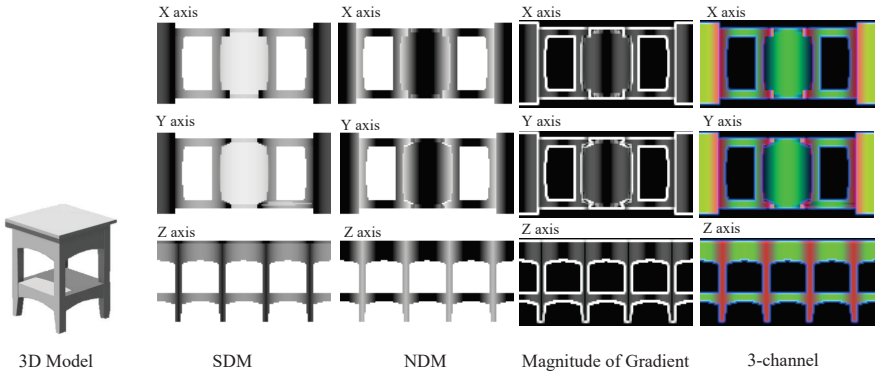

Figure 2: The 3D model and the SDM, NDM, Magnitude of Gradient and 3-channel stacked images on three axes.

points s(φ,y) where $\varphi \in [0, 2\pi]$ is the angle in xy plane and y$\in [0, H]$ is sampled at rates 2B and B. In this paper, we set $B = 128$. s(φ,y) represents the different characteristics on 3D model's surface, which are the position of the model's surface in 3D space (Spatial Distribution Map or SDM) and the orientation of the model's surface, (Normals' Deviation Map or NDM). We further generate the gradient image (Magnitude of Gradient) from NDM view. Finally, one 3-channel image (3-channel) is computed by combining the above three images. At last, we can get 12 different views like Figure.2 from X, Y and Z axes for each 3D model.

3.2 MMJN: Multi-modality Joint Networks

Based on the multi-modal 3D data, we propose a novel multi-modal joint network. Figure. 1 shows the detailed framework of our proposed method. The proposed framework consists of three networks, one for 3D point cloud feature extraction, utilizing the popular PointNet to learn the transformation function, one for the 3D multi-view feature extraction, utilizing the classic MVCNN to learn the feature extraction function, and one for 3D PANORAMA View feature extraction, also utilizing the MVCNN structure to handle this problem. In this paper, we consider that the features based on different modalities should be similar because they represent the same 3D model. Thus, in order to demonstrate the assumption, we suppose the three networks have the same dimension of $fc \in \mathbb{R}^{1024}$ as the feature vector of 3D model. The proposed method trains these three deep neural networks simultaneously with proposed loss including the traditional discrimination loss for each domain and the correlation loss for cross-modal.

In the traditional training step of single modality network, the discrimination loss aims at minimizing the intra-class distance of the extracted features and maximizing the inter-class distance of the extracted features to a large margin within each modality. The definition of discrimination loss is followed as:

$$\mathcal{L}_d = -\sum_{i=1}^{u}\sum_{j=1}^{K} y_{ij} \log p_{ij}(\beta_j|\beta_1, \beta, ..., \beta_K) \tag{1}$$

This is the softmax loss. y_{ij} is the real label of sample i from category j. K is the total number of all categories. With the softmax layer, the probability prediction of β_j is defined based on the modality feature f as below:

$$p_j(\beta_j|\beta_1, \beta, ..., \beta_K) = \frac{e^{\beta_j}}{\sum_{n=1}^{K} e^{\beta_n}} \tag{2}$$

The discrimination loss has been utilized in many application and also gets excellent results in many classic classification problems. In this work, for each network, we fist introduce the traditional discrimination loss to guarantee the final performance of transformer function. Meanwhile, we also introduce the correlation loss to guide each other in training step. Thus we can increase the final learning speed in training step and improve the robustness of final feature vector. The correlation loss is followed as:

$$L_c(M_i, M_n) = \| \xi(f_{M_i}) - \xi(f_{M_n}) \|_2 \tag{3}$$

where f represents the feature vector extracted by different modality networks, M represents the modality data whose subscript can be $1, 2, 3$ in this work, and $\xi = sigmoid(log(abs(\cdot)))$ is a normalization function. Here, we apply the 2 norms as the distance metric between two different feature vectors to measure the correlation. The value of the correlation loss should be smaller and smaller in the learning step. It means that these features guide each other and utilize the advantage of different modalities' feature in training step to obtain a more robust feature vector. Based on the design of correlation loss, the final loss function of different modality networks is followed as:

$$L_{M_1} = L_{d,M_1} + L_c(M_1, M_2) + L_c(M_1, M_3); \tag{4}$$

where L^{d,M_1} is the discrimination loss based on modality M_1 network. $L_c(M_1, M_2)$ and $L_c(M_1, M_3)$ represent the correlation loss with modality M_2 and modality M_3 respectively. Finally, we optimize these three networks through back-propagation with stochastic gradient descent. [18].

3.3 Multi-modal Information Fusion

According to the joint learning of different modalities, we can get three feature extraction models based on different 3D modalities data. These features should have small distance or similar position in the feature space. In this work, we employ the weighted fusion

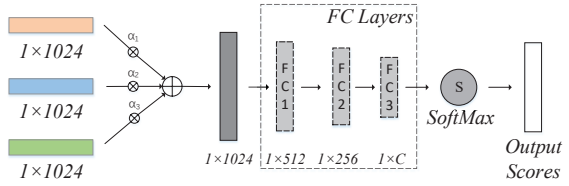

Figure 3: The figure shows the process of fusing the features, and then we fed the features into the fully-connected layers to make the classification of 3D modalities. First, we put the features into the 1×512, 1×256 and 1×C fully-conneted layers. The C represent the categories of the dataset. We used the ModelNet40 dataset here, so we set C to 40. Then we connect a softmax layer to get the probability that the object belongs to C classes.

method to fuse these three feature vectors. The framework of this method is shown in Figure.3. The detail is shown in Equation.5.

$$f = \sum_{i=1}^{3} \alpha_i \xi(f_{M_i}); \quad \sum_{i=1}^{3} \alpha_i = 1; \tag{5}$$

where f represents the feature vector extracted by PointNet, multi-view MVCNN and PANORAMA-MVCNN respectively based on different modalities of 3D model. α_i is the weight of modality feature in order to balance the multi-view feature, point cloud feature and PANORAMA feature. The fused feature is also processed by softmax to get the class label. The related experiment is shown in Section.4.2.

4 EXPERIMENT

4.1 Dataset

In order to evaluate the performance of our proposed method of classification, we made extensive use of a well-known dataset named ModelNet [35] which consists of two versions and they are publicly available for download: ModelNet10 and ModelNet40. ModelNet10 comprises 4899 CAD models split into 10 categories. The training and testing subsets 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. They are specially clean since the models that do not belong to the specified categories were manually deleted. Especially, ModelNet10 models are pose normalized in terms of translation and rotation, and ModelNet40 models are not pose normalized.

4.2 Experiment on The Effectiveness of Correlation Loss

In this paper, we introduce the correlation loss in the global loss function. The goal of the design is to make these features guide each other in training step for more robust representation. In order to demonstrate the performance of the correlation loss, we compare the convergence trend of discrimination loss under the action of correlation loss with that of the discrimination loss in traditional single-modal network. The corresponding experimental results are shown in Figure.4. Here, we only show the convergence trend in 100 epoch, as these three networks have an obvious convergence trend in the first 100 epoch. From these figures, we can find that the discrimination loss converges quickly when we add the correlation loss in the global loss function. Meanwhile, the final classification result also outperforms the results of tradition networks which only have discrimination loss. The related results are shown in Table.1. This experiment demonstrates the reasonableness and effectiveness of our approach.

4.3 Comparison on the Combinations of Different Modality Networks

In this work, we propose a novel feature fusion method to fuse the multi-modal information extracted by these different modality networks. The goal of this design is to utilize the advantages of different modality networks to get more accurate classification result and more robust feature representation. In order to demonstrate the performance of this approach, we compare the classification results of single modality network with the combinations of different modality networks. The corresponding experimental results are shown in Table.1. From this table, we can find that the combination of different modality networks brings a significant improvement in performance compared with single modality network. Here, for ModelNet40, MV+PC brings a 4.05% and 1.58% improvement over MV and PC respectively. MV+PV brings a 2.15% and 6.28% improvement over MV and PV respectively. PC+PV brings a 0.25% and 6.85% improvement over PC and PV respectively. Finally, MV+PC+PV brings a 5.23%, 2.76% and 9.36% improvement over each single modality respectively. We can find that the PC network brings the biggest improvement under different conditions. Meanwhile, the single modality network PC also gets the best classification results compared with other single modality network. There are reasons to think that point cloud data represents more information of 3D modality. This experimental result also demonstrates that different modalities have different contributions for the final classification results. We will discuss this problem in the next section.

Table 1: Comparisons of classification accuracy with different modalities combination on ModelNet10 and ModelNet40

Method	Classification Accuracy	
	ModelNet10	ModelNet40
MV	89.01%	87.23%
PC	93.28%	89.70%
PV	87.33%	83.10%
MV + PC	92.29%	91.28%
MV + PV	90.41%	89.38%
PC + PV	93.61%	89.95%
MV + PC + PV	93.83%	92.46%

4.4 Experiment on the Multi-modal Information Fusion

According to the above sections, we draw the conclusion that different modalities should have different weights in the final information

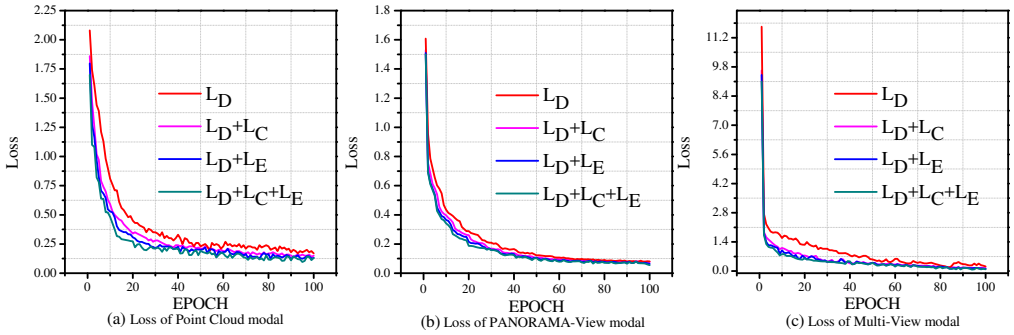

Figure 4: Experiment on classifier loss of different modalities. In order to compare the effects on discrimination loss on each branch after adding the correlation loss, we calculated the mean discrimination loss on each epoch. (a) compares the effects on Point Cloud modality, (b) compares the effects on PANORAMA-View modality and (c) compares the effects on Multi-View modality. $L_D + L_C$ represents the discrimination loss after adding the correlation loss. L_D represents the discrimination loss without the correlation loss. This figure significantly shows that the discrimination loss drops faster after adding the correlation loss.

fusion in order to utilize the advantage of each modality. In order to define the weights of different modality networks, we sample different values of fusion parameters to find the best values. The related experiment on ModelNet40 dataset is shown in Figure 5. Finally, we set the parameter $\alpha_1 = 0.7$, $\alpha_2 = 0.2$ and $\alpha_3 = 0.1$ as the weights of Point Cloud, Multi-view and PANORAMA-view respectively in fusion equation 5. From this experiment, we can find that the modality PC has the biggest weight. It also demonstrates our assumption in the above section. Meanwhile, the optimized parameters bring 1.36% improvement over the un-optimized condition, which also demonstrates the effectiveness of the proposed method.

Meanwhile, We also evaluate the performance of the proposed method using different combinations of the components and demonstrate the performance of our fusion method in Table.2. In this table,"P" denotes only the PointNet of our method is used for 3D model representation and classification. "MV" denotes the multi-view MVCNN is used for 3D model representation and classification. "PV" denotes the PANORMAN views are used for 3D model representation and classification. "Late Fusion" denotes that the three modality networks' classification results are fused. "Ours" denotes our fusion method. As shown in the comparison, we observe that our method obviously outperforms the other comparison methods in mean class accuracy and global class accuracy. This condition also demonstrates that our fusion method can effectively utilize the advantage of each modality data to achieve the best performance. The visual confusion matrix on ModelNet40 is shown in Figure 7.

4.5 Comparison with State-of-the-art methods on ModelNet40

To validate the efficiency of the proposed MMJN, 3D shape classification experiments have been conducted on the Princeton ModelNet dataset [35]. 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 testing split setting as in [35].

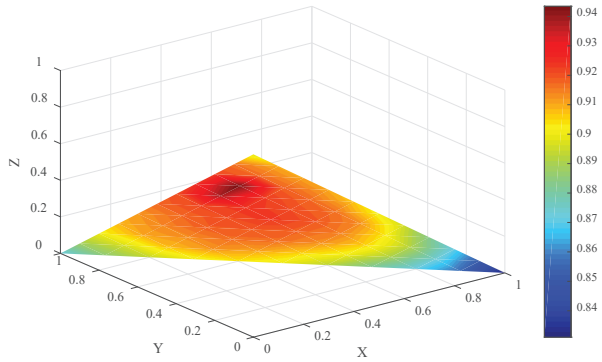

Figure 5: Experiment on Fusion Parameters. The x, y and z axes represent the weight of PANORAMA-View modality, Multi-View modality and Point Cloud modality respectively. The color bar represents the accuracy of classification with the weight, where red illustrates the highest accuracy and blue indicates the lowest accuracy.

Table 2: Comparisons of different feature fusion methods on Classification (ModelNet40)

Methods	Mean Class Accuracy	Overall Class Accuracy
MV	85.68%	87.23%
PC	87.27%	89.70%
PV	82.30%	83.10%
Late Fusion	90.53%	92.46%
Ours	**92.24%**	**93.82%**

In experiments, we have compared the proposed MMJN with various models based on different representations, including volumetric based models [35], hand-craft descriptors for multi-view

Table 3: Comparisons of classification accuracy and retrieval mAP on ModelNet40

Method	Train Config		Data Representation	Classification	Retrieval
	Pre train	Fine tune	#Number of Views	(Overall Accuracy)	(mAP)
(1)SPH[19]	-	-	-	68.2%	33.3%
(2)LFD[3]	-	-	-	75.5%	40.9%
(3)3D ShapeNets[35]	ModelNet40	ModelNet40	Volumetric	77.3%	49.2%
(4)VoxNet[22]	ModelNet40	ModelNet40	Volumetric	83.0%	-
(5)VRN[2]	ModelNet40	ModelNet40	Volumetric	91.3%	-
(6)MVCNN-MultiRes[26]	-	ModelNet40	Volumetric	91.4%	-
(7)MVCNN,12×[34]	ImageNet1K	ModelNet40	12 Views	89.9%	70.1%
(8)MVCNN,metric,12×[34]	ImageNet1K	ModelNet40	12 Views	89.5%	80.2%
(9)MVCNN,80×[34]	ImageNet1K	ModelNet40	80 Views	90.1%	70.4%
(10)MVCNN,metric,80×[34]	ImageNet1K	ModelNet40	80 Views	90.1%	79.5%
(11)PointNet[9]	-	ModelNet40	Point Cloud	89.2%	-
(12)PointNet++[27]	-	ModelNet40	Point Cloud	90.7%	-
(13)KD-Network[20]	-	ModelNet40	Point Cloud	91.8%	-
(14)PointCNN[21]	-	ModelNet40	Point Cloud	91.8%	-
(15)DGCNN[41]	-	ModelNet40	Point Cloud	92.2%	-
(16)PANORAMA-NN[31]	-	ModelNet40	PANORAMA-Views	90.7%	83.4%
(17)PVNet[39]	ImageNet1K	ModelNet40	Point Cloud and Multi-Views	93.2%	89.5%
(18)MMJN(Our)	**ImageNet1K & ModelNet40**	**ModelNet40**	**Point Cloud & 12 Views & PANORAMA-Views**	**93.8%**	**89.8%**

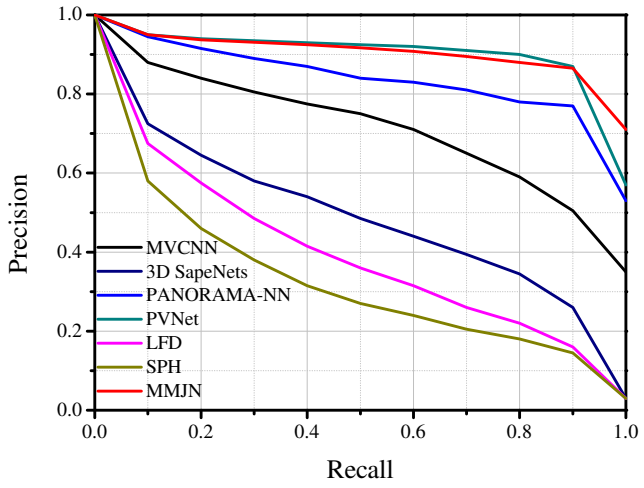

Figure 6: Precision-recall curves for our MMJN and other methods on the task of shape retrieval on the ModelNet40 dataset.

data [3, 19], deep learning models for multi-view data [26, 34], deep learning models for PANORAMA-Views [31] and point cloud based models [9, 20, 21, 27, 41].

In Tab.3, the classification results of all compared methods are provided. As shown in the results, our proposed MMJN can achieve the best performance with the classification accuracy of 93.8%. Compared with the MVCNN using GoogLeNet, our MMJN wins by 1.0% more gains on the classification tasks. For point cloud based models, our MMJN also outperforms the state-of-the-art point cloud based model DGCNN by 1.0% in terms of classification accuracy.

In the retrieval task, we apply the fusion feature f in equation.5 as the feature vector of 3D model. The Euclidean distance is used to compute the similarity between two different 3D models. The precision-recall curves for retrieval of all compared methods are demonstrated in Fig.6. From the retrieval results, our approach achieves an exciting state-of-the-art performance of 89.8%, which efficiently demonstrates the effective of our MMJN in the 3D shape retrieval task.

The exciting performance of our proposed MMJN can be explained from the following reasons. First, the correlation loss can jointly utilize the advantage of different modalities to guide the parameter learning and increase the feature learning rate in training step. Second, the proposed multi-modal feature fusion method can expand the advantage of different modalities in the final classification and retrieval problem. The related experiment also demonstrates the superiority of our approach.

5 IMPLEMENTATION

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 network. For Multi-View network, 12 views for each object are fed into network. The parameters of CNN in multi-view network are initialized by the pre-trained ModelNet model. For PANAORAMA-View network, 12 views are fed into the network that same to Multi-View network, whereas the parameters aren't initialized. We pre-train the model on our dataset, and find the best model to initialize the parameters. The learning rate we set is 0.0001. All the experiments are conducted on two

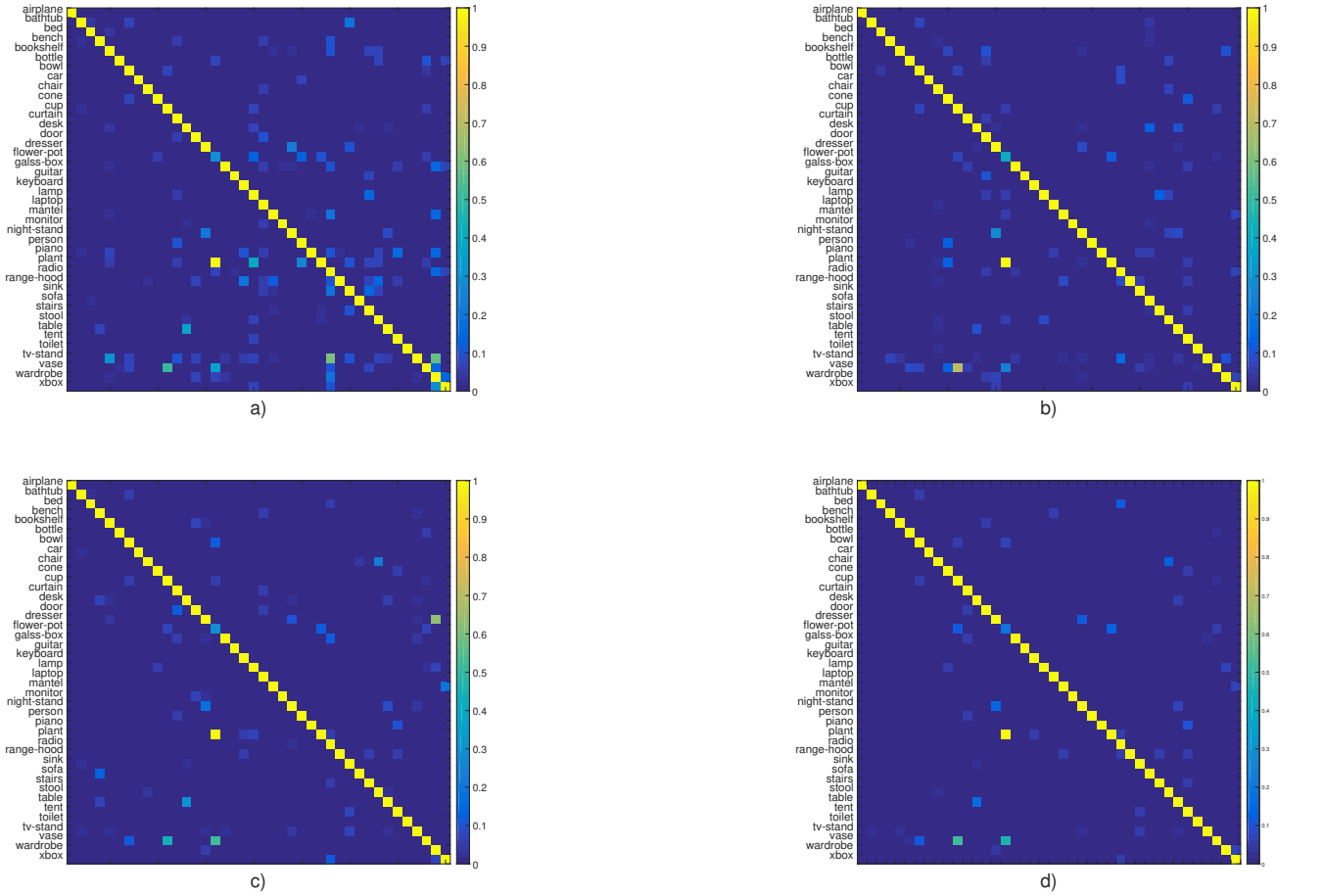

Figure 7: Confusion matrix for the 3D models of the ModelNet-40 dataset classes. Figure a) is the Confusion matrix's color map of PV modality; Figure b) is the Confusion matrix's color map of MV modality; Figure c) is the Confusion matrix's color map of PC modality; Figure d) is the Confusion matrix's color map of late fusion modality. Yellow indicates the highest percentage of the model's predict labels, while blue indicates the lowest percentage of the model's predict labels.

NVIDIA 1080Ti GPUs. Our framework is trained in an end-to-end fashion.

6 CONCLUSION

In this paper, we propose a novel jointly networks: MMJN, which can jointly employ different modality data for 3D shape classification and retrieval. In our framework, the correlation loss is introduced to employ the advantage of different modality networks to guide each other for the feature learning. It can increase the learning rate and also improve the performance of each single modality network. Then, we propose a novel multi-modal feature fusion method, which defines the different weights for each modality feature. It can expand the advantage of each modality network to get more robust representation for each 3D model. 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.

7 ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (61772359, 61572356, 61872267, 61502477), the grant of 2019 Tianjin New Generation Artificial Intelligence Major Program, the grant of Tianjin New Generation Artificial Intelligence Major Program (18ZXZNGX00150), the grant of Elite Scholar Program of Tianjin University (2019XRX-0035), and the Director Foundation Project of National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC). The Open Project Program of the State Key Lab of CAD & CG, Zhejiang University (Grant No.A1907).

REFERENCES

[1] S. Bai, X. Bai, Z. Zhou, Z. Zhang, Q. Tian, and L. J. Latecki. 2017. GIFT: Towards Scalable 3D Shape Retrieval. *IEEE Transactions on Multimedia* 19, 6 (June 2017), 1257–1271. https://doi.org/10.1109/TMM.2017.2652071

[2] Andrew Brock, Theodore Lim, J. M. Ritchie, and Nick Weston. 2016. Generative and Discriminative Voxel Modeling with Convolutional Neural Networks. *Computer Science* (2016).

[3] Ding Yun Chen, Xiao Pei Tian, Yu Te Shen, and Ouhyoung Ming. 2010. On Visual Similarity Based 3D Model Retrieval. *Computer Graphics Forum* 22, 3 (2010), 223–232.

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2016. Multi-View 3D Object Detection Network for Autonomous Driving. *CoRR* abs/1611.07759 (2016). arXiv:1611.07759 http://arxiv.org/abs/1611.07759

[5] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C. Kanjirathinkal, and Mohan Kankanhalli. 2018. MMALFM: Explainable Recommendation by Leveraging Reviews and Images. 37, 2 (2018).

[6] Nira Dyn, David Levine, and John A. Gregory. 1990. A Butterfly Subdivision Scheme for Surface Interpolation with Tension Control. *ACM Transaction on Graphics* 9 (04 1990), 160–. https://doi.org/10.1145/78956.78958

[7] Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. 2018. MeshNet: Mesh Neural Network for 3D Shape Representation. *CoRR* abs/1811.11424 (2018). arXiv:1811.11424 http://arxiv.org/abs/1811.11424

[8] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. 2018. GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] A. Garcia-Garcia, F. Gomez-Donoso, J. Garcia-Rodriguez, S. Orts-Escolano, M. Cazorla, and J. Azorin-Lopez. 2016. PointNet: A 3D Convolutional Neural Network for real-time object class recognition. In *2016 International Joint Conference on Neural Networks (IJCNN)*. 1578–1584. https://doi.org/10.1109/IJCNN.2016.7727386

[10] A. GonzÃąlez, D. VÃązquez, A. M. LÃşpez, and J. Amores. 2017. On-Board Object Detection: Multicue, Multimodal, and Multiview Random Forest of Local Experts. *IEEE Transactions on Cybernetics* 47, 11 (Nov 2017), 3980–3990. https://doi.org/10.1109/TCYB.2016.2593940

[11] H. Guo, J. Wang, Y. Gao, J. Li, and H. Lu. 2016. Multi-View 3D Object Retrieval With Deep Embedding Network. *IEEE Transactions on Image Processing* 25, 12 (Dec 2016), 5526–5537. https://doi.org/10.1109/TIP.2016.2609814

[12] Z. Han, Z. Liu, J. Han, C. M. Vong, S. Bu, and C. L. Chen. 2017. Mesh Convolutional Restricted Boltzmann Machines for Unsupervised Learning of Features With Structure Preservation on 3-D Meshes. *IEEE Transactions on Neural Networks Learning Systems* 28, 10 (2017), 2268–2281.

[13] Zhizhong Han, Zhenbao Liu, Junwei Han, Chi Man Vong, Shuhui Bu, and C. L. Philip Chen. 2017. Unsupervised Learning of 3-D Local Features From Raw Voxels Based on a Novel Permutation Voxelization Strategy. *IEEE Transactions on Cybernetics* PP, 99 (2017), 1–14.

[14] Z. Han, Z. Liu, C. M. Vong, Y. S. Liua, S. Bu, J. Han, and Clp Chen. 2018. Deep Spatiality: Unsupervised Learning of Spatially-Enhanced Global and Local 3D Features by Deep Neural Network with Coupled Softmax. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society* PP, 99 (2018), 1–1.

[15] Z. Han, M. Shang, Z. Liu, C. Vong, Y. Liu, M. Zwicker, J. Han, and C. L. P. Chen. 2019. SeqViews2SeqLabels: Learning 3D Global Features via Aggregating Sequential Views by RNN With Attention. *IEEE Transactions on Image Processing* 28, 2 (Feb 2019), 658–672. https://doi.org/10.1109/TIP.2018.2868426

[16] Vishakh Hegde and Reza Zadeh. 2016. FusionNet: 3D Object Classification Using Multiple Data Representations. (2016).

[17] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. 2018. Pointwise Convolutional Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Rie Johnson and Tong Zhang. 2013. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 315–323. http://papers.nips.cc/paper/4937-accelerating-stochastic-gradient-descent-using-predictive-variance-reduction.pdf

[19] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. 2003. Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. In *Symposium on Geometry Processing*.

[20] Roman Klokov and Victor S. Lempitsky. 2017. Escape from Cells: Deep Kd-Networks for The Recognition of 3D Point Cloud Models. *CoRR* abs/1704.01222 (2017). arXiv:1704.01222 http://arxiv.org/abs/1704.01222

[21] Yangyan Li, Rui Bu, Mingchao Sun, and Baoquan Chen. 2018. PointCNN. *CoRR* abs/1801.07791 (2018). arXiv:1801.07791 http://arxiv.org/abs/1801.07791

[22] D. Maturana and S. Scherer. 2015. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 922–928. https://doi.org/10.1109/IROS.2015.7353481

[23] Panagiotis Papadakis, Ioannis Pratikakis, Stavros Perantonis, and Theoharis Theoharis. 2007. Efficient 3D shape matching and retrieval using a concrete radialized spherical projection representation. *Pattern Recognition* 40, 9 (2007), 2437 – 2452. https://doi.org/10.1016/j.patcog.2006.12.026

[24] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. Morency. 2017. Multi-level Multiple Attentions for Contextual Multimodal Sentiment Analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*. 1033–1038. https://doi.org/10.1109/ICDM.2017.134

[25] Charles R. Qi, Su Hao, Matthias Niessner, Angela Dai, and Leonidas J. Guibas. 2016. Volumetric and Multi-View CNNs for Object Classification on 3D Data. (2016).

[26] Charles R. Qi, Su Hao, Matthias Niessner, Angela Dai, and Leonidas J. Guibas. 2016. Volumetric and Multi-View CNNs for Object Classification on 3D Data. (2016).

[27] Charles R. Qi, Yi Li, Su Hao, and Leonidas J. Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. (2017).

[28] Guido Ranzuglia, Marco Callieri, Matteo Dellepiane, Paolo Cignoni, and Roberto Scopigno. 2013. MeshLab as a complete tool for the integration of photos and color with high resolution 3D geometry data. In *CAA 2012 Conference Proceedings*. Pallas Publications - Amsterdam University Press (AUP), 406–416. http://vcg.isti.cnr.it/Publications/2013/RCDCS13

[29] Charles Ruizhongtai Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas Guibas. 2016. Volumetric and Multi-View CNNs for Object Classification on 3D Data. (04 2016).

[30] Konstantinos Sfikas, Ioannis Pratikakis, and Theoharis Theoharis. 2018. Ensemble of PANORAMA-based convolutional neural networks for 3D model classification and retrieval. *Computers Graphics* 71 (2018), 208 – 218. https://doi.org/10.1016/j.cag.2017.12.001

[31] Konstantinos Sfikas, Theoharis Theoharis, and Ioannis Pratikakis. 2017. Exploiting the PANORAMA Representation for Convolutional Neural Network Classification and Retrieval. In *Eurographics Workshop on 3D Object Retrieval*, Ioannis Pratikakis, Florent Dupont, and Maks Ovsjanikov (Eds.). The Eurographics Association. https://doi.org/10.2312/3dor.20171045

[32] Richard Socher, Brody Huval, Bharath Putta Bath, Christopher D. Manning, and Andrew Y. Ng. 2012. Convolutional-Recursive Deep Learning for 3D Object Classification.. In *NIPS*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, LÃĺon Bottou, and Kilian Q. Weinberger (Eds.). 665–673. http://dblp.uni-trier.de/db/conf/nips/nips2012.html#SocherHBMN12

[33] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. 2018. SPLATNet: Sparse Lattice Networks for Point Cloud Processing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[34] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. 2015. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 945–953. https://doi.org/10.1109/ICCV.2015.114

[35] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, Jianxiong Xiao, Zhirong Wu, Shuran Song, and Aditya Khosla. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision Pattern Recognition*.

[36] Sun Xiao, Zhenguang Liu, Yuxing Hu, Luming Zhang, and Roger Zimmermann. 2018. Perceptual multi-channel visual feature fusion for scene categorization. *Information Sciences* 429 (2018), 37–48.

[37] J. Xie, G. Dai, F. Zhu, E. K. Wong, and Y. Fang. 2017. DeepShape: Deep-Learned Shape Descriptor for 3D Shape Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 7 (July 2017), 1335–1345. https://doi.org/10.1109/TPAMI.2016.2596722

[38] Gang-Joon Yoon and Sang Min Yoon. 2017. Sketch-based 3D object recognition from locally optimized sparse features. *Neurocomputing* 267 (2017), 556 – 563. https://doi.org/10.1016/j.neucom.2017.06.034

[39] 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. *CoRR* abs/1808.07659 (2018). arXiv:1808.07659 http://arxiv.org/abs/1808.07659

[40] Haoxuan You, Yifan Feng, Xibin Zhao, Changqing Zou, Rongrong Ji, and Yue Gao. 2018. PVRNet: Point-View Relation Neural Network for 3D Shape Recognition. *CoRR* abs/1812.00333 (2018). arXiv:1812.00333 http://arxiv.org/abs/1812.00333

[41] Wang Yue, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. 2018. Dynamic Graph CNN for Learning on Point Clouds. (2018).

[42] Z. Zhang, H. Lin, X. Zhao, R. Ji, and Y. Gao. 2018. Inductive Multi-Hypergraph Learning and Its Application on View-Based 3D Object Classification. *IEEE Transactions on Image Processing* 27, 12 (Dec 2018), 5957–5968. https://doi.org/10.1109/TIP.2018.2862625

[43] Lei Zhu, Zi Huang, Zhihui Li, Liang Xie, and Heng Tao Shen. 2019. Exploring Auxiliary Context: Discrete Semantic Transfer Hashing for Scalable Image Retrieval. *CoRR* abs/1904.11207 (2019). arXiv:1904.11207