Computing and Informatics, Vol. 40, 2021, 1217–1243, doi: 10.31577/cai_2021_6_1217

PVRAR: POINT-VIEW RELATION NEURAL NETWORK EMBEDDED WITH BOTH ATTENTION MECHANISM AND RADON TRANSFORM FOR 3D SHAPE RECOGNITION

Jie Zhou, Ziping Ma, Jinlin Ma

North Minzu University
750021 Yinchuan, China
e-mail: maziping@tom.com

Abstract. Owing to the favorable performance of deep neural networks for 3D shape recognition, an increasing number of researchers are interested in designing novel 3D shape descriptors. However, the relationship between multiple views and point clouds needs to be further elucidated. We propose a multimodal method that combines the features of point clouds and multiple views, i.e., point-view relation neural network embedded with both attention mechanism and Radon transform, to obtain better descriptors. First, a two-dimensional linear Radon transform is performed to investigate linear and color features in multiple views, and the features are used as the input of our network to enable significant distinctions between different views. Moreover, a convolutional block attention module is adopted to enhance the features of point clouds and hence improve the expression ability of feature descriptors. The effectiveness of the proposed method is verified using ModelNet40 and ModelNet10 datasets. Experimental results show that our method can effectively improve the capability of feature extraction and expression as well as achieve state-of-the-art performance on two well-known 3D datasets.

Keywords: Multimodal fusion, 2D linear Radon transform, attention mechanism, 3D shape recognition

Mathematics Subject Classification 2010: 68T10

1 INTRODUCTION

Owing to the development of computer technology, three-dimensional (3D) shape recognition has become a popular research topic in fields of medical diagnosis, 3D printing, medical imaging, and digital entertainment. Meanwhile, as the number of 3D models increases, the difficulty in recognizing and retrieving 3D models is increasing as well. Three-dimensional models with different shapes and volumes typically have different functions and categories. Some of the same types of 3D models exhibit different shapes owing to esthetic and practical factors. Consequently, understanding and identifying 3D models has gradually become critical to some application fields.

Recently, convolutional neural networks (CNNs) have achieved satisfactory results in 3D shape recognition. Although 3D CNNs can directly process regular voxel data [1, 2], the performance of neural networks is generally limited by the sparse degree of voxels and computational cost. More recently, owing to the rapid development of relevant technologies, point clouds and multi-view that are used to express 3D models have become easier to achieve, thereby resulting in a significant amount of results pertaining to these two types of data. Multi-view-based methods primarily involve perspective selection and multi-view feature fusion. In some methods [3, 4, 5, 6], a 3D model is represented by different two-dimensional (2D) views, which reduces the demand for computer memory and facilitates deep neural network processing. However, self-occlusion may result in deviations in local information. It is well known that point clouds, as another type of 3D data, represent 3D models in the form of points and contain more stereoscopic information than 2D views. In methods based on point clouds [7, 8, 9], the number of sampling points affects the amount of information. In addition, the irregularity and disorder of point clouds cause more processing difficulties.

To balance the advantages and disadvantages of point clouds and multiple views, You designed a convolutional neural network (PVNet) [10] that explored the complementary relationship between point clouds and multi-view. Subsequently, the PVRNet [11] was proposed to integrate the features of point clouds and multiple views further. Compared with the PVNet [10], it focused on investigating the relationship between point clouds and different views. Moreover, Peng et al. [12] proposed an intermodality attention enhancement module and a view-context attention fusion module to investigate discriminative shape features. Yang and Dang [13] proposed the modules of point-view attention fusion and point-view-point attention fusion to facilitate the fusion process of point clouds and multi-view and discard redundant view features adaptively. Most methods based on multiple views for 3D model recognition represent each 3D model in a 2D form, which inevitably causes stereoscopic information loss. Researchers [14, 15, 16] have recently investigated the contribution of the Radon transform in feature representation, and the significance of the attention mechanism [17, 18, 19, 20] was investigated extensively in the computer vision domain. Our goal is to achieve favorable recognition performance by combining the Radon transform and a neural

network to extract features with stronger expression ability in multimodal methods.

The main contributions of this study are as follows:

1. We propose a new deep convolutional neural network, named point-view relation neural network embedded with both attention mechanism and Radon transform (PVRAR), for 3D shape recognition. In our method, a 2D linear Radon transform is performed to process multiple views, and the corresponding results are input to the module for view feature extraction. To improve the expression ability of features in the point-view fusion module, we enhance the features of point clouds using a convolutional block attention module (CBAM) and a residual connection.
2. We verify the effectiveness of our method on well-known 3D shape datasets (ModelNet10 and ModelNet40). Our method can effectively extract the features of point clouds and multiple views and obtain the relation between point clouds and multiple views to combine the features of multimodal data to obtain global features, as well as classify and retrieve 3D models. Compared with existing methods, our method achieves better performance.

The remainder of this paper is organized as follows: First, we present studies related to 3D shape recognition. Next, we introduce the proposed method comprehensively. Then, we present our experimental results and analyses. Finally, a summary of our findings is provided.

2 RELATED STUDIES

In this section, we review some representative studies, which can be classified into two categories: 3D model recognition methods based on deep learning and studies associated with the 2D Radon transform.

2.1 Three-Dimensional Model Recognition Methods Based on Deep Learning

Currently, the forms to represent 3D models primarily include multiple views, point clouds, meshes, and voxels. We review the studies associated with our proposed method. Those studies can be classified into three categories of methods: methods based on

1. point clouds and multiple views,
2. multiple views, and
3. point clouds.

Methods based on point clouds and multiple views. In 3D shape recognition, multimodal methods refer to the combined information of several different

modes to implement information complementarity by capturing the internal correlation between different modes. In the PVNet [10], an embedding attention fusion module was adopted to project global view features into the subspace of point cloud features to investigate the intrinsic correlation and discriminability of views and point clouds. However, in most cases, the relationship between the point cloud and view is not well represented. Hence, You et al. proposed a relation score module to learn the correlation between point clouds and multiple views in the PVRNet [11]. Similarly, Peng et al. [12] employed a two-stage attention fusion network to gradually refine and combine the features of point clouds and multiple views to improve the retrieval performance by 0.9 % as compared with that of the PVRNet [11]. To address redundant information between adjacent images, Zhao et al. [21] applied a soft attention fusion module to generate attention weights for different views and used a residual connection to process point and view features to enhance the final features. To fully utilize the correlation between point clouds and multiple views, Yang and Dang [13] proposed point-view attention modules to facilitate the fusion of point clouds and views and discard redundant view features adaptively. Their method is better than the PVNet [10] for 3D shape recognition using ModelNet40. However, in their method, "bonsai" was mistaken for "plants", which indicates that the ability to distinguish key features must be improved.

Methods based on multiple views. Two-dimensional multiple views are typically used to represent 3D models. A Siamese network [22] was designed to capture and integrate the corresponding features of RGB views and binary images for 3D model retrieval. Li et al. [6] applied a differentiable renderer [23] to generate viewpoints that can be optimized to yield more information renderings for neural networks. It is noteworthy that the discrimination of views should be quantified accurately because the information of views differ significantly. Hence, a hierarchical view-group-shape architecture [4] was designed to extract shape-level descriptors based on the content discrimination of each view, which demonstrated better performance compared with the view integration strategy used in an MVCNN. Xu et al. [24] constructed multiple 2D-CNNs and a novel multi-view loss fusion strategy, where several discriminative and informative views were selected to identify 3D models. In addition, to address the limitation of feature discriminability of aggregated multi-view using the pooling operation [3], Ma et al. [25] used a CNN to extract low-level features of views and then aggregated features via long short-term memory and a sequence voting layer. Similarly, Han et al. [26] proposed a modified CNN with hierarchical attention aggregation to effectively focus on the content information within a view and the spatial relationship between multiple views. However, this method is not applicable to unordered views.

Methods based on point clouds. Point clouds represent 3D models in the form of points, and each point contains coordinates and other information. The dynamic graph CNN (DGCNN) [27] achieved favorable results in point cloud

recognition. In this method, the edge convolutional layer can obtain the local neighborhood information. In fact, their local neighborhoods may be extremely similar, thereby preventing the receipt of valuable edge information. Hence, Zhang et al. [28] applied the k-nearest neighbor and multilayer perceptron (MLP) with shared parameters to extract the local feature from the central point and its neighbors, and then added shortcuts to link the hierarchical features to obtain informative edge vectors. To further promote correlation learning between different areas, an attention-based encoder-decoder network [29] was introduced to capture the fine-grained contextual information of local regions for point cloud processing in shape classification. Moreover, in point cloud-based retrieval for place recognition, the matching problem of global descriptors is important, based on which PointNetVLAD [30], which combines PointNet [7] and NetVLAD [31], is designed to extract the global descriptor from scanned 3D point clouds. Although PointNetVLAD is more efficient in obtaining global features than PointNet, it seldom explores the positive effects of local features on global features. In a later study, Zhang and Xiao [32] added a context-aware attention mechanism to the global feature extraction structure and applied the attention map to the NetVLAD layer.

2.2 Work Associated with 2D Radon Transform

Radon transform is a well-established feature extraction method, specifically in the computer vision domain. Recently, several researchers have proposed projection-based descriptors based on the Radon transform. Tizhoosh applied the Radon transform to support vector machines (SVMs) for medical image classification [14]. A Radon-based neural network [16] consists of Radon projections and encoding projections, and the classification results of this method are obtained via a shallow MLP to learn medical images. More recently, Sriram et al. [33] used the Radon transform to extract the features of images and then stored them as a compact histogram. The classification process was expressed using a histogram intersection algorithm with an SVM classifier. In fact, the method might result in the loss of staining information containing chemical significance in histopathology owing to graying. Similarly, Eltaieb et al. [34] applied the Radon transform and singular value decomposition with an SVM classifier to achieve modulation format identification based on lower decibel levels than the literature [35]. In a study [36], a new Radon cumulative distribution transform classifier was designed with few iterations and hyperparameter adjustment. Eventually, the method obtained a lower cost and less algorithm complexity than the Resnet [37].

Because Radon transform can be used to retain useful information regarding stripes, Ding et al. [38] adopted the Radon transform and grayscale transform enhancement method to realize image denoising. The Laplacian of a Gaussian distribution can be calculated after the Radon transform to restore multidirectional motion-blurred objects in images. Hence it can be extended for different nonuniform motions in the object [39]. However, the performance of this method is

affected by the detection angles. Additionally, the Radon transform can be applied for wake detection, in which ship-centered-masked images can be utilized to restrict the search area between two sine curves to reduce the computational time [40]. Instead of using the method above, the Radon transform was used to enhance the linear features of synthetic SAR images [41]. Additionally, owing to the ability of Radon transform to detect circular objects, it is combined with a residual CNN to segregate small extracellular vesicles [42]. This method can accurately process highly heterogeneous images but cannot segment large vesicles effectively.

The studies above suggest that the fusion of multiple views and point clouds facilitates 3D model recognition. Therefore, we attempted to devise a more effective method to identify 3D models by combining point clouds and multiple views. The 2D Radon transform is effective in extracting image features; it can highlight the linear feature of an image and compensate for the loss of stereoscopic information in 2D images. Moreover, attention mechanisms can enhance key information in datasets. Therefore, in our method, the 2D linear Radon transform and CBAM [43] were used to extract features of 3D shapes to obtain more expressive shape features.

3 PVRAR FOR 3D SHAPE RECOGNITION

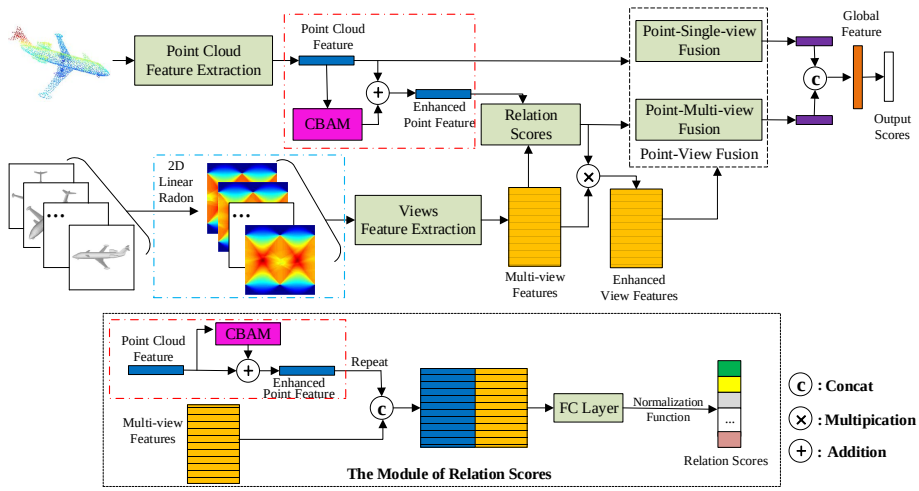


Figure 1. Architecture of PVRAR

We herein propose a multimodal fusion framework, PVRAR, for 3D shape recognition. The architecture of the PVRAR is illustrated in Figure 1. In the branch for extracting point cloud features, the module of point cloud feature extraction was realized via the DGCNN [27]. The CBAM [43] was used to process the output from

the point cloud feature extraction module and then enhanced point features were obtained via a residual connection. In the branch for extracting multi-view features, the module for view feature extraction was realized using the MVCNN [3]. Two-dimensional images with more color and linear information were obtained via 2D linear Radon transform, and the images were used as the input of the view feature extraction module. In the branch for point-view fusion, relation scores were calculated based on enhanced point features and multi-view features, and then the view features were enhanced using the relation scores. Subsequently, we employed two strategies to combine the features of point clouds and multiple views, i.e., point-single-view fusion and point-multi-view fusion. Next, global features were constructed, and the class of each 3D model was predicted. In the PVRAR, a cross-entropy loss function and a stochastic gradient descent with a momentum of 0.9 were used. For the retrieval task, we used the 256-dimensional feature prior to the last full connection layer to represent the shape, and the Euclidean distance was applied to measure the similarity between 3D models.

3.1 Branch for Extracting Point Cloud Features

Each 3D point cloud model composed of 1024 points was input to the point cloud feature extraction module implemented using the DGCNN [27].

The convolutional block attention module (CBAM) [43] is an attention mechanism that focuses on exploiting both spatial-wise and channel-wise attention to improve the ability of feature expression. Hence, we incorporated the CBAM (as shown in Figure 2) into our method to enhance the representation of point features. The areas highlighted by the dashed red boxes shown in Figure 1 illustrate the method to enhance the features of the point clouds via a residual connection. The enhanced point features and multi-view features were used to calculate the relation scores.

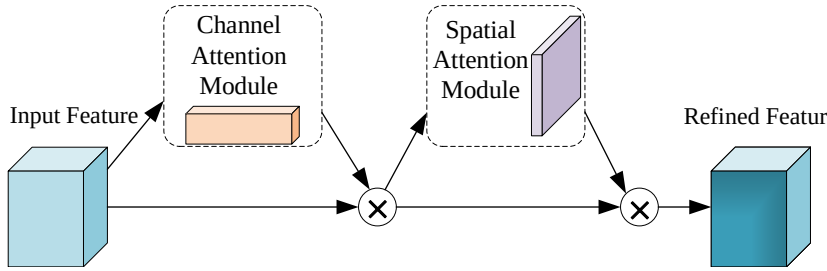


Figure 2. CBAM architecture [43]

The CBAM not only identifies the location of focus, but also improves the representation of key information. It includes two sequential submodules: channel attention and spatial attention. For a feature map $I \in \mathbb{R}^{C\times H\times W}$ as input, the CBAM sequentially infers a one-dimensional channel attention map, $M_C \in \mathbb{R}^{C\times 1\times 1}$, and

a 2D spatial attention map, $M_S \in \mathbb{R}^{1\times H\times W}$, as illustrated in Figure 2. The overall attention process can be expressed as shown in Equation (1).

$$I' = M_C(I) \otimes I, I'' = M_S(I') \otimes I', \tag{1}$$

where symbol $\otimes$ denotes element-wise multiplication. During multiplication, Channel attention values are broadcasted along the spatial dimension, and Spatial attention values are broadcasted along the channel dimension. More details regarding the CBAM are available in the literature [43].

3.2 Branch for Extracting View Features

We employed 12 views of each 3D shape as our original view data; the views were captured using cameras at 30° intervals and then further processed via 2D linear Radon transform to obtain the corresponding views as the input of the view feature extraction module. Subsequently, the extraction module was initialized by the pretrained MVCNN [3], where the AlexNet [44] was used as the basic network. In some multimodal methods [10, 11], multi-view data are generated directly from off-type files in ModelNet10 and ModelNet40 datasets, thereby resulting in insufficient color information. In fact, the color and brightness information of 2D images is critical as it affects the feature extraction ability of neural networks. Therefore, 2D linear Radon transform was adopted to generate 2D images with bright colors to improve the expression ability of the view features. Moreover, 2D views inevitably lose the stereoscopic information of 3D models, whereas 2D linear Radon transform can extract the linear features of 2D views, thereby enriching the contour information of the 3D model in each 2D image. The area highlighted by the dashed blue box in Figure 1 represents the multi-view processed by the 2D linear Radon transform. Next, we briefly review the 2D linear Radon transform for 2D images. More details are available in [45].

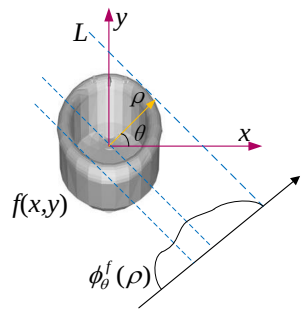


Figure 3. Two-dimensional linear Radon transform [45]

The 2D linear Radon transform (as shown in Figure 3) of a signal $f(x, y)$ is a set of integrals of $f(x, y)$ along the dashed lines, where the dashed lines can be

conveniently parameterized using the offset ρ and orientation θ. A 2D linear Radon transform is expressed as shown in Equation (2)

$$\phi_\theta^f(\rho) = \iint f(x,y)\delta(\rho - xcos(\theta) - ysin(\theta))\,\mathrm{d}x\,\mathrm{d}y \tag{2}$$

where ρ is the vertical distance from the origin to the dashed line L, θ is the angle between the positive-half axis of the x-axis and the normal line. The vertical line is perpendicular to the dashed line L, and $\delta(\cdot)$ is the Dirac delta function. The function $\phi_\theta^f(\rho)$ for a fixed θ can be regarded as a projection of the signal f along the direction orthogonal to orientation θ.

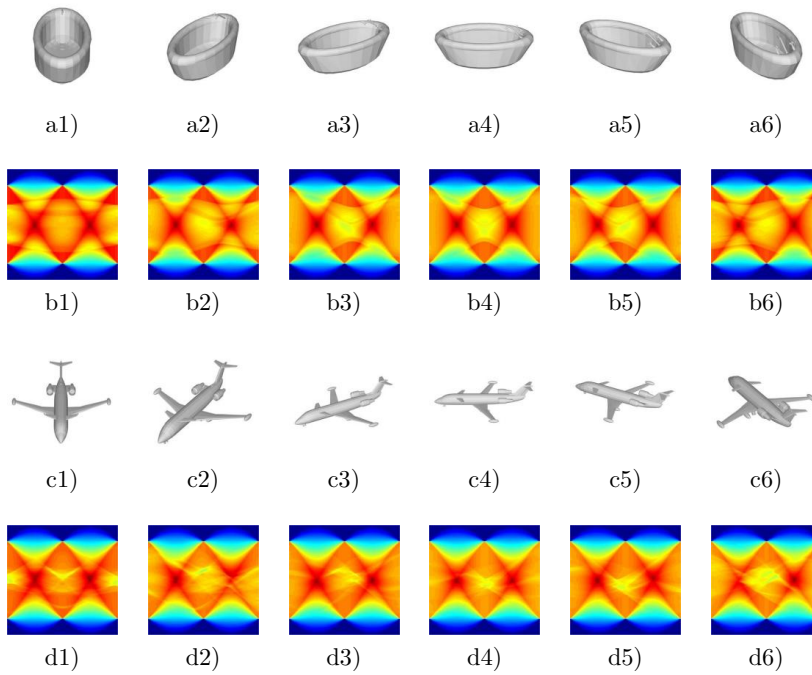


Table 1. Original views and the corresponding images after 2D linear Radon transform

The 2D linear Radon transform is used to calculate the projection of an image on straight lines in different directions. The mapped matrix obtained by the 2D linear Radon transform can accurately reflect the geometric information of an image. Therefore, we used the 2D linear Radon transform to process multiple views and saved the mapped matrix as colorful 2D images to represent the change in linear features using different colors. For a visual illustration, we present the multiple views of a bathtub, an airplane, and the corresponding images after 2D linear Radon transform, as shown in Table 1. The first row shows six different views a1)–a6) of a bathtub, the second row shows the images b1)–b6) after a 2D linear Radon

transform of the views presented in a1)–a6), respectively, the third row shows six different views c1)–c6) of an airplane, and the fourth row shows the images d1)–d6) after a 2D linear Radon transform of the views presented in c1)–c6), respectively. For example, a1) represents the 2D view of a bathtub, and b1) is the corresponding image obtained after a 2D linear Radon transform of a1). After performing a 2D linear Radon transform, a line in an image domain is mapped to a point in a transform domain; therefore, the bright or dark linear features in the image are transformed into peak or valley points in the transform domain. By performing a 2D linear Radon transform, the linear features in view are highlighted. Moreover, the color information of the multi-view becomes more apparent. We used the corresponding image from each view processed via 2D linear Radon transform as the input of our network.

3.3 Branch for Point-View Fusion

The core task in multimodal fusion is to effectively combine different types of data. Because each view represents a section of a 3D shape, the contribution of each view is different in terms of shape representation. When combined with point cloud features, different view features should be treated differently to ensure that the multimodal features are effectively combined. Inspired by the PVRNet [11], for each 3D model, we used Equation (3) to calculate the relation scores of the enhanced point cloud feature and the i^{th} view. The structure highlighted by the dashed gray box at the bottom of Figure 1 represents the module of relation scores.

$$RS_i(p, V) = \xi(g_\theta(p, v_i)) \tag{3}$$

where p is the point cloud feature, and $V = \{v_1, v_2, \ldots, v_n\}$ denotes n extracted view features from a 3D model. The function g_θ is a simple MLP that provides the relations between the point cloud feature and each view feature. ξ is a sigmoid function with normalization. For each view, the output is a relation score ranging from 0 to 1 that represents the significance of the correlation between different views and the point cloud.

We used the relation scores to obtain the enhanced view features v_i' via a residual connection, as shown in Equation (4).

$$v_i' = v_i * (1 + RS_i(p, V)). \tag{4}$$

To fully utilize the local features in each view, we used two point-view fusion strategies. Point-single-view fusion is shown in Figure 4, where the point cloud feature is concatenated with each view feature using the "Concat" function. By processing the FC (full connection layer) and a max pooling layer, a key feature representing the input is obtained and then is named as "SF". In the point-multi-view fusion strategy (Figure 5), the relation scores decide the view feature to be included. We organized the views based on the relation scores corresponding to

each view (the higher the relation score, the higher is the ranking position of the corresponding view), and then constructed four point-view feature sets, as shown in Figure 5. Subsequently, an average pooling layer was used to process the output of the four FC layers. We denote the output feature map obtained using the point-multi-view fusion module as "MF". Subsequently, the SF and MF were concatenated using the "Concat" function. The final feature representing the 3D model was proceeded by two FC layers and a softmax function to generate the classification result.

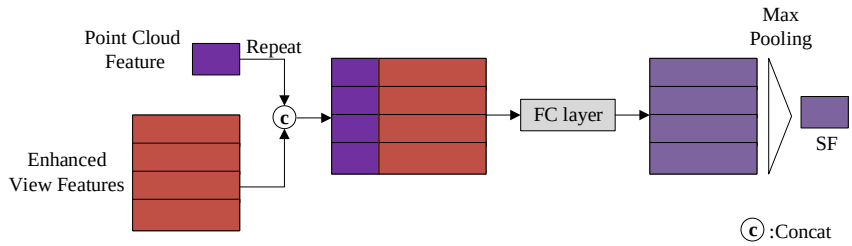


Figure 4. Architecture of point-single-view fusion [11]

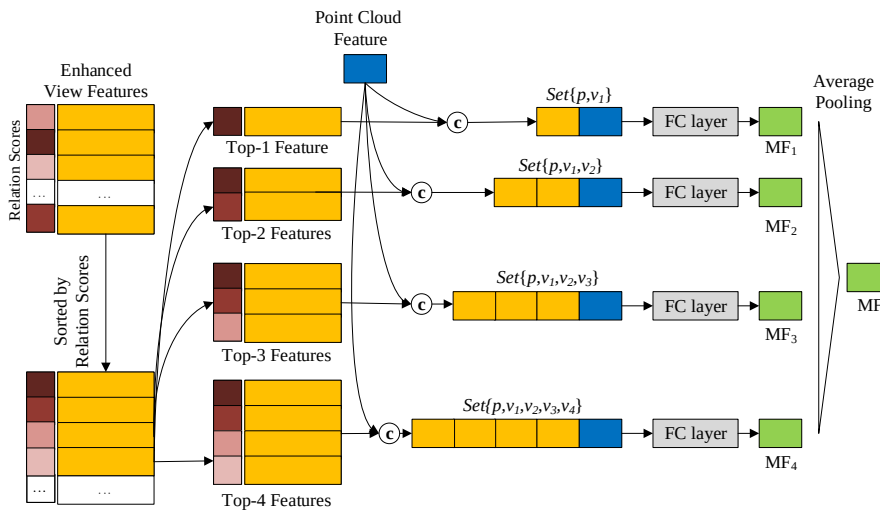


Figure 5. Architecture of point-multi-view fusion [11]

Moreover, in the point-view fusion strategy of the PVRAR, one type of data was the point cloud features extracted using the DGCNN, and the other type of data was enhanced view features that were indirectly guided by the enhanced point cloud features. This strategy was used to improve the ability of data representation in the modules of point-single-view and point-multi-view fusion.

3.4 Training Strategy

We employed a two-stage approach to train our network so that network training stability was ensured. Firstly, we saved the network weights corresponding to the best performance of the views and point cloud feature extraction modules. Secondly, we loaded the checkpoints saved in the first step into the network structure of the PVRAR during training. In the first 10 epochs, the network weights were fixed in the feature extraction modules of the multiple views and point clouds. Hence, we trained only the other network sections. After the first 10 epochs, all network weights were updated simultaneously to achieve favorable performance gradually. Finally, the network weights with the best recognition performance in our method were saved.

4 EXPERIMENTAL RESULTS AND ANALYSES

To verify the effectiveness and robustness of the PVRAR, a series of experiments was performed on standard datasets. The experimental environment was Ubuntu 18.04.3 LTS and the Nvidia TITAN V graphics card was used. Additionally, Python version 3.6.8 was used, and PyTorch was used as the machine learning framework. The number of points and views of each 3D model were set to 1 024 and 12, respectively.

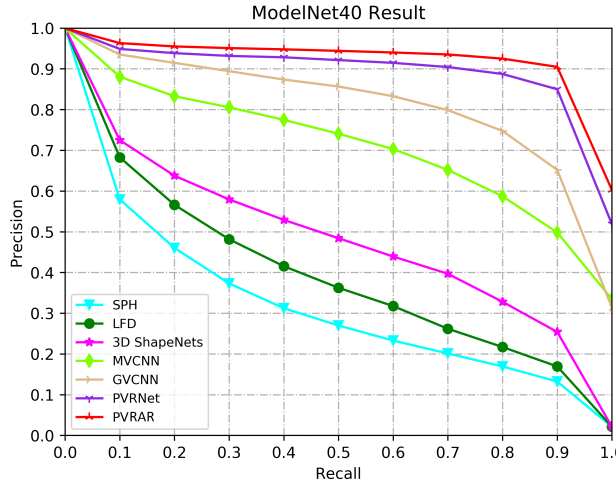
4.1 Experimental Datasets

The ModelNet10 and ModelNet40 datasets [1] are the most typically used datasets for 3D shape classification, and their details are as follows:

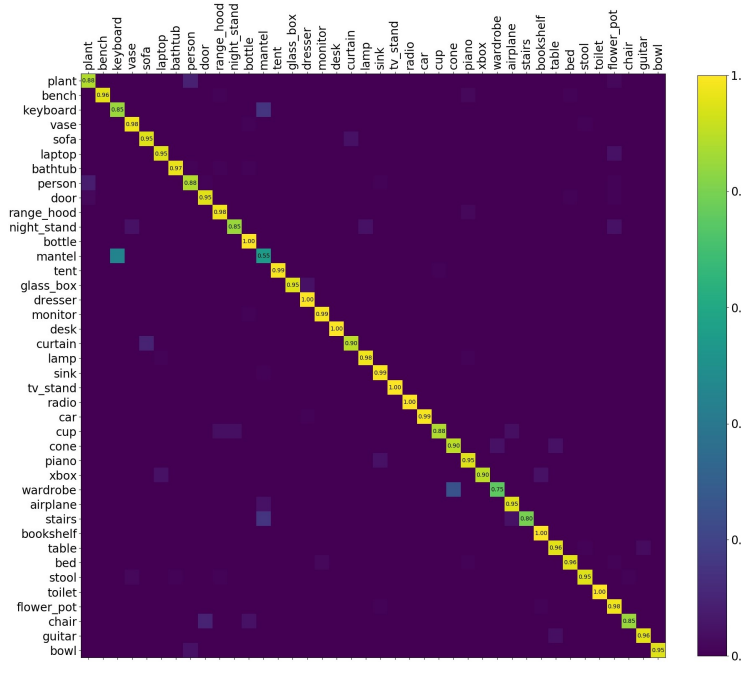
1. ModelNet10, an orientation-aligned dataset including 4 899 3D models from 10 categories, in which there are 3 991 models for training and 908 models for testing.
2. ModelNet40, a dataset including 12 311 3D models from 40 categories, in which there are 9 843 models for training and 2 468 models for testing. Compared with ModelNet10, the 3D models in ModelNet40 are not orientation-aligned.

4.2 Three-Dimensional Shape Classification and Retrieval

In this section, we present the experimental results of the PVRAR for classification and retrieval tasks. We compared the PVRAR with some representative methods, including point-based methods (PointNet [7], PointNet++ [8], 3DmFV [46], DGCNN [27], and LDGCNN [28]), multi-view-based methods (MVCNN [3], GVCNN [4], methods in the literature [25], and 3D2SeqViews [26]), and multimodal methods (PVNet [10], PVRNet [11], MANet [21], and PVFNet [13]). The classification and retrieval performances of each method are represented by the overall accuracy and mean average precision (mAP), respectively.



a) PR curves of different methods for 3D model retrieval on ModelNet40



b) Confusion matrix of PVRAR on ModelNet40 classification

Figure 6. Precision-recall curves and confusion matrix on ModelNet40 dataset

Method	Data Representation	ModelNet40		ModelNet10	
		Classification (overall accuracy)	Retrieval (mAP)	Classification (overall accuracy)	Retrieval (mAP)
MVCNN [3]	12 Views	89.9 %	80.2 %	–	–
GVCNN [4]	8 Views	93.1 %	84.5 %	–	–
Ma et al. [25]	12 Views	91.1 %	84.3 %	95.3 %	93.2 %
3D2SeqViews [26]	12 Views	93.4 %	90.8 %	94.7 %	92.1 %
PointNet [7]	Point Coluds	89.2 %	–	–	–
PointNet++ [8]	Point Coluds	90.7 %	–	–	–
3DmFV [46]	Point Coluds	91.4 %	–	95.2 %	–
DGCNN [27]	Point Coluds	92.2 %	81.6 %	–	–
LDGCNN [28]	Point Coluds	92.9 %	–	–	–
PVNet [10]	Point Coluds and 12 Views	93.2 %	89.5 %	–	–
PVRNet [11]	Point Coluds and 12 Views	93.6 %	90.5 %	–	–
MANet [21]	Point Coluds and 12 Views	93.4 %	90.1 %	–	–
PVFNet [13]	Point Coluds and 12 Views	94.1 %	90.8 %	95.0 %	91.7 %
PVRAR (ours)	Point Coluds and 12 Views	95.3 %	93.2 %	94.6 %	93.1 %

Table 2. Classification and retrieval results on ModelNet40 and ModelNet10 datasets

A comparison of the experimental results is presented in Table 2. The following can be inferred from Table 2:

1. Compared with other methods, the overall accuracy/mAP of the PVRAR were 95.3 %/93.2 % and 94.6 %/93.1 % on the ModelNet40 and ModelNet10 datasets, respectively, which demonstrates that the PVRAR achieved better performance on ModelNet40.
2. Compared with multi-view-based methods, the PVRAR indicated an improvement of 5.4 % and 13.0 % in terms of the overall accuracy and mAP over the MVCNN on ModelNet40, respectively. In particular, the PVRAR outperformed 3D2SeqViews, which is a more recently developed method, by 1.9 % and 2.4 % on ModelNet40 in terms of the overall accuracy and mAP, respectively.
3. In the classification and retrieval results based on point clouds, the PVRAR indicated a significant improvement by 3.1 % and 11.6 % in terms of the overall accuracy and mAP, respectively, compared with the DGCNN on ModelNet40.
4. For the multimodal methods based on point clouds and multiple views, the PVRAR demonstrated better performances, except that the overall accuracy of the PVRAR was slightly lower than that of the PVFNet on ModelNet10.

Although the PVRNet utilized a point-view fusion strategy similar to that adopted by the PVRAR, the latter outperformed the PVRNet by 1.7 % and 2.7 % in terms of the overall accuracy and mAP, respectively, on ModelNet40.

The precision-recall (PR) curves of the different methods on ModelNet40 are shown in Figure 6 a). As shown, the PVRAR achieved better performance in terms of 3D model retrieval compared with the other methods. In particular, the curve of PVRAR was higher than that of the PVRNet, indicating that our method is superior to the PVRNet.

Then, we calculated the classification distribution of each class in the ModelNet40 dataset and plotted the confusion matrix (as shown in Figure 6 b)) to further analyze the capability of our method in extracting fine-grained geometric features. The closer the value in the confusion matrix is to 1, the better is the classification performance. As shown in Figure 6 b), the values of bottle, dresser, desk, tv_stand, radio, bookshelf, and toilet in the confusion matrix were 1.00 for all the classes. In addition, the values of some classes (vase, range_hood, tent, monitor, lamp, sink, car, and flower_pot) in the matrix were greater than or equal to 0.98, which was desirable. Whereas most mantels were mistakenly predicted as the keyboard class, resulting in the value of the mantel class was 0.55. The misjudgment occurred because the multi-view of a section from both the mantels and keyboards were similar, which resulted in an indistinguishable difference between the features extracted using our method.

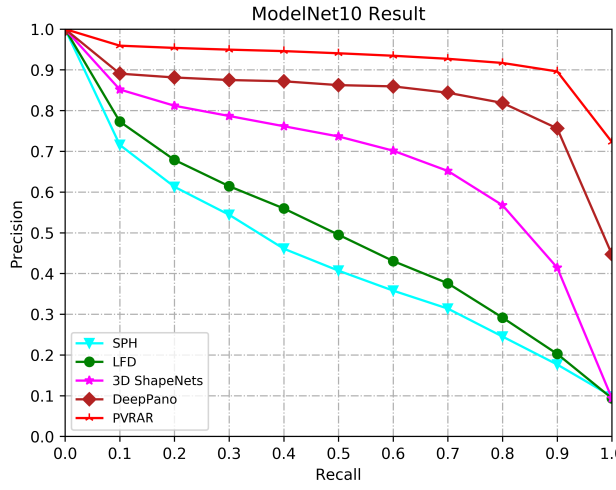
In addition, to evaluate the robustness of our method, we performed a series of experiments on another well-known 3D model dataset. On ModelNet10, the precision-recall curves of different methods representing retrieval performance are shown in Figure 7 a), and the classification distribution of each class is shown in Figure 7 b). As presented, compared with the SPH [47], LFD [48], 3D ShapeNets [1], and DeepPano [49], our method demonstrated the best retrieval performance on the ModelNet10 dataset. Meanwhile, as shown by the confusion matrix, the values for sofa, bathtub, monitor, bed, toilet, and chair class exceeded 0.95.

4.3 Ablation Studies

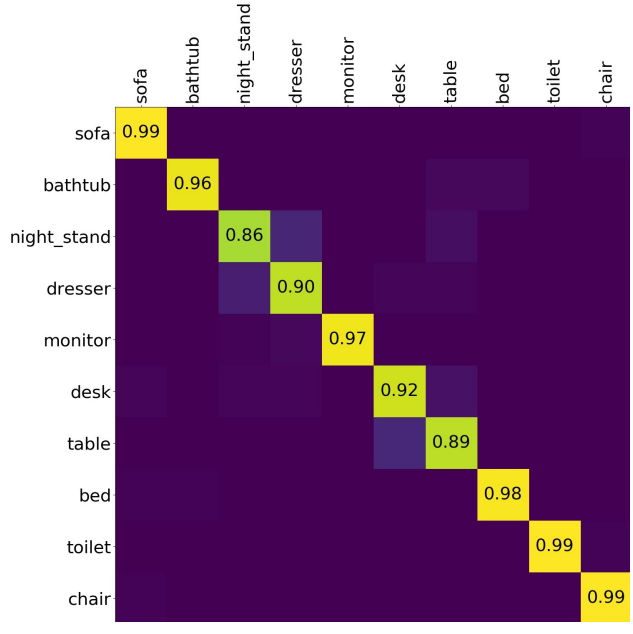
We designed a series of ablation experiments to analyze the effects of the PVRAR. The ablation experiments were conducted using the ModelNet40 dataset.

4.3.1 Ablation Experiments of 2D Linear Radon Transform

To verify the effectiveness of implementing the 2D linear Radon transform in our method, we evaluated the classification and retrieval performances in different combinations. Because our aim was to investigate the contribution of the 2D linear Radon transform to our method, for a fair comparison, we did not use CBAM in the modules evaluated. The classification performance was evaluated based on the mean class accuracy and overall accuracy, whereas their retrieval performances were measured based on the mAP.



a) Precision-recall curves of different methods



b) Confusion matrix of PVRAR

Figure 7. Precision-recall curves and confusion matrix on ModelNet10 dataset

The detailed experimental results are listed in Table 3. In the table, "R & ..." indicates the view feature extraction model where the result of 2D linear Radon transform was used as the data input. "Multi-View" denotes the view feature extraction model adopted, where the MVCNN with the AlexNet was employed. "MFusion" shows that only the point-multi-view fusion module was used. In the MFusion module, we used four point-view feature sets. "SFusion" implies that we applied only the point-single-view fusion module. "SFusion & MFusion" implies that we employed both the point-single-view and point-multi-view fusion modules.

As shown in Table 3, the performances of the Multi-View, SFusion, and SFusion & MFusion modules improved when 2D linear Radon transform was applied. Meanwhile, we observed the following:

1. In general, the performances of all modules with 2D linear Radon transform were superior to the corresponding modules without Radon transform. In particular, the R & Multi-View module achieved the highest improvements of 3.85 % and 1.59 % in terms of the mean class accuracy and overall accuracy, respectively.
2. Furthermore, among the modules with 2D linear Radon transform, the R & SFusion & MFusion module outperformed not only all the modules without 2D linear Radon transform significantly, but also other modules with 2D linear Radon transform, particularly for the Multi-View module, where significant increases of 5.88 % and 5.28 % were recorded in terms of the mean class accuracy and overall accuracy, respectively. In other words, the R & SFusion & MFusion module performed the best, which implies that the 2D linear Radon transform should be combined with both SFusion and MFusion modules to achieve favorable results.
3. Among the modules with Radon transform, the R & MFusion module improved by 0.83 %, 0.56 %, and 1.84 % in terms of mean class accuracy, overall accuracy, and mAP, respectively, compared with SFusion & MFusion. By contrast, these three indicators of the R & SFusion module were 1.35 %, 1.45 %, and 2.26 % higher than those of the SFusion & MFusion module, respectively. Hence, it is clear that the 2D linear Radon transform is crucial for extracting features and can improve the performance of point-view fusion modules (SFusion and MFusion).
4. By comparing between R & MFusion and R & SFusion, we discovered that the point-single-view fusion module outperformed the point-multi-view fusion module. This indicates that the point-single-view fusion module contributed more significantly than the point-multi-view fusion module in our method when the 2D linear Radon transform was applied.

In conclusion, the 2D linear Radon transform improved the performance of 3D model recognition. In this study, the 2D linear Radon transform was adopted for each of the view-related modules and favorable performances were achieved because the transform enhanced the linearity of the views. Furthermore, we saved the

mapped matrices from 2D linear Radon transform as colorful 2D images as data input, which was able to provide more color information to our network. Consequently, the feature extraction ability of our network improved significantly.

	Module	Classification		Retrieval
		Mean Class Accuracy	Overall Accuracy	mAP
Without Radon	Multi-View	87.60 %	89.90 %	–
	SFusion	90.93 %	92.84 %	–
	SFusion & MFusion	91.64 %	93.61 %	90.52 %
With Radon	R & Multi-View	91.45 %	91.49 %	–
	R & MFusion	92.47 %	94.17 %	92.36 %
	R & SFusion	92.99 %	95.06 %	92.78 %
	R & SFusion & MFusion	93.48 %	95.18 %	93.15 %

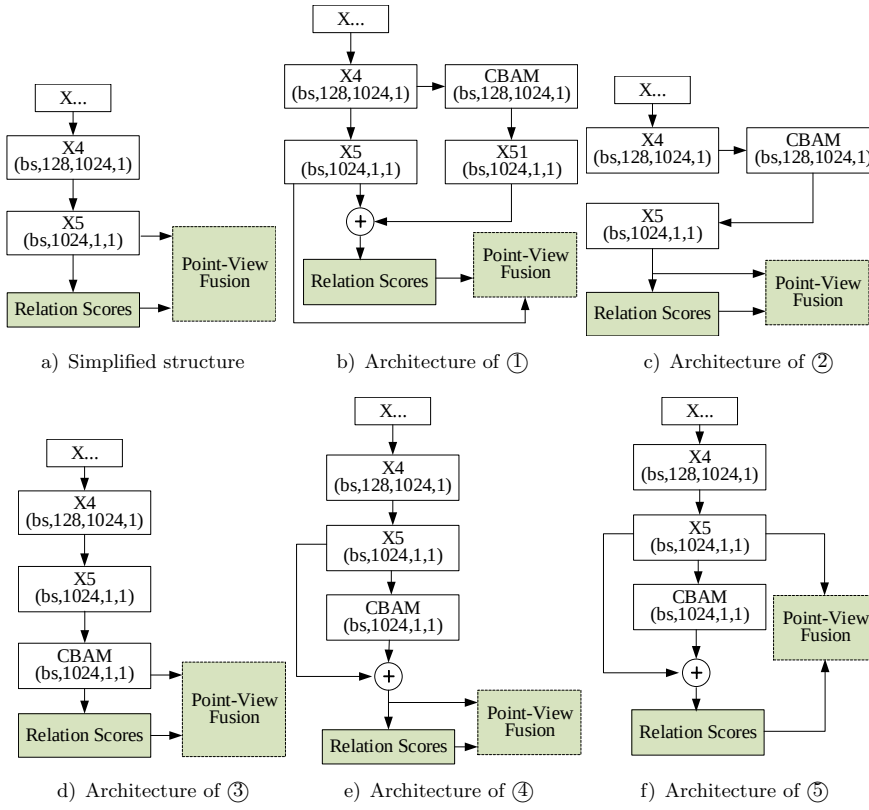
Table 3. Effectiveness of different modules without CBAM

4.3.2 Ablation Experiments of CBAM

To verify the contribution of the CBAM to our method, we conducted a series of experiments. As presented in Section 4.3.1, the R & SFusion & MFusion module performed the best, as shown in Table 3. Therefore, we incorporated the CBAM into the R & SFusion & MFusion module. The different methods of incorporating the CBAM are shown in Figure 8.

In our network, the CBAM was used to extract the point cloud features. To facilitate the ablation experiments, we simplified the point cloud feature extraction module, as shown in Figure 8 a). Because the CBAM was adopted to process the outputs of X4 and X5, we present the size of the feature mapping for the outputs of X4 and X5. Additionally, "bs" represents the batch size, e.g., "(bs, 128, 1 024, 1)" denotes the size of the feature mapping of bs "$128 \times 1\,024 \times 1$". We incorporated the CBAM at different locations of the network structure. In structures ①–⑤ shown in b-f in Figure 8, "$\oplus$" represents the addition of different point cloud feature mappings, and the direction indicated by the arrow represents the direction of data flow.

We designed experiments A and B to identify the best method to incorporate the CBAM into our network, as well as the best combination of point and view features. The experimental results are listed in Table 4. In experiment A, the R & SFusion module was used in five different methods to incorporate the CBAM, as shown in Figure 8, to identify the best method to incorporate the CBAM. The R & SFusion module was used because "R & SFusion" outperformed "R & MFusion", as indicated in Table 3. In sub-experiment A, only the point-single-view fusion module was used to combine the features of point clouds and multiple views. Based on the results of experiment A, experiment B was conducted to identify the best methods for point-view fusion and CBAM incorporation.



a) Simplified structure b) Architecture of ① c) Architecture of ②

d) Architecture of ③ e) Architecture of ④ f) Architecture of ⑤

Figure 8. Different methods to incorporate CBAM

Meanwhile, the following can be inferred from Table 4:

1. In sub-experiment A, R & SFusion & ⑤ recorded the highest overall accuracy score. Moreover, the scores of R & SFusion & ④ were the highest, except for the overall accuracy. In the architectures of ④ and ⑤, the feature map output from the CBAM was added to the map output from X5, which increased the weight of the key information in point features and enhanced the point features. When the features were combined in the point-single-view fusion module, the view features were enhanced by the relation scores calculated using the enhanced point features. The results from sub-experiment A show that the methods used to incorporate the CBAM in ④ and ⑤ yielded better performances. Therefore, the methods of ④ and ⑤ were the most suitable for incorporating the CBAM in experiment B.
2. In sub-experiment B, we used the methods of ④ and ⑤ to evaluate the 3D model recognition performance based on the R & SFusion & MFusion module.

As shown by the experimental results listed in Table 4, compared with the R & SFusion & MFusion module, the classification and retrieval performances of the R & SFusion & MFusion & ④ module did not improve. By contrast, the R & SFusion & MFusion & ⑤ module indicated improvements in the values of the mean class accuracy, overall accuracy, and mAP by 0.11 %, 0.08 %, and 0.03 %, respectively. This indicates that when both the point-single-view and point-multi-view fusion modules are employed, the structure of ⑤ can effectively improve the network performance. In structure ⑤ in the point-view fusion module, the view features are indirectly processed by the CBAM, and the point cloud features are directly extracted from the DGCNN.

Moreover, R & SFusion & MFusion & ⑤ outperformed R & SFusion & ⑤ which indicates the effectiveness of the point-multi-view fusion module. Using the architecture of ⑤, we increased the representation forms of the data involved in the fusion to improve the ability of feature expression, so that the network can potentially learn the maximum amount of data when performing predictions.

Based on the experimental results above, it is concluded that the best approach to incorporate the CBAM into our method is by using ⑤, and that the best fusion module is R & SFusion & MFusion. Our method can enrich features and yield greater expression ability from the network.

	Module	Classification		Retrieval
		Mean Class Accuracy	Overall Accuracy	mAP
A	R & SFusion	92.99 %	95.06 %	92.78 %
	R & SFusion & ①	92.34 %	94.33 %	92.50 %
	R & SFusion & ②	92.58 %	94.41 %	92.51 %
	R & SFusion & ③	92.82 %	94.45 %	92.81 %
	R & SFusion & ④	93.47 %	95.02 %	92.88 %
	R & SFusion & ⑤	93.29 %	95.14 %	92.69 %
B	R & SFusion & MFusion	93.48 %	95.18 %	93.15 %
	R & SFusion & MFusion & ④	93.03 %	94.85 %	93.10 %
	R & SFusion & MFusion & ⑤	93.59 %	95.26 %	93.18 %

Table 4. The experimental results

4.3.3 Ablation Experiments of Point-Multi-View Fusion Module

The effects of the 2D linear Radon transform and CBAM can be understood based on the experimental results presented in Sections 4.3.1 and 4.3.2. In this section, we discuss the effect of the number of views involved in the point-multi-view fusion module on performance. In Table 5, "SFusion & MFusion ($1+\cdots+n$ views)" implies that both point-single-view fusion and point-multi-view fusion were employed with a certain number of views in n sets, where each set includes point clouds and n views.

We discuss the effect of the number of views involved in the point-multi-view fusion module on the performance of the PVRAR.

As shown in Table 5, "SFusion & MFusion ($1 + 2 + 3 + 4$ views)" with Radon transform and ⑤ in Figure 8 f) demonstrated the best performances, indicating that the fusion of the point cloud features with the four sets organized by the relation scores can achieve the best state for our method. In other cases, the combined features indicated information shortage and redundancy, separately. The information shortage leads to insufficient network learning ability, and the redundancy of information leads to the learning of some interference information. Therefore, we used "SFusion & MFusion (1+2+3+4 views)" with Radon transform and ⑤ in Figure 8 f) as our 3D model recognition method, which is represented as the PVRAR.

	Module	Classification		Retrieval
		Mean Class Accuracy	Overall Accuracy	mAP
Without Radon transform and ⑤ in Figure 8 f)	SFusion & MFusion ($1 + 2$ views)	91.35 %	92.99 %	–
	SFusion & MFusion ($1 + 2 + 3$ views)	91.43 %	93.47 %	–
	SFusion & MFusion ($1 + 2 + 3 + 4$ views)	91.64 %	93.61 %	90.52 %
With Radon transform and ⑤ in Figure 8 f)	SFusion & MFusion ($1 + 2$ views)	93.09 %	94.77 %	91.90 %
	SFusion & MFusion ($1 + 2 + 3$ views)	92.81 %	94.94 %	92.11 %
	SFusion & MFusion ($1 + 2 + 3 + 4$ views)	93.59 %	95.26 %	93.18 %
	SFusion & MFusion ($1 + 2 + 3 + 4 + 5$ views)	93.06 %	95.05 %	92.28 %
	SFusion & MFusion ($1 + 2 + 3 + 4 + 5 + 6$ views)	92.57 %	94.73 %	91.70 %

Table 5. Ablation experiments based on number of views in point-multi-view fusion

5 CONCLUSION

Herein, a point-view relation neural network incorporated with both the attention mechanism and Radon transform, abbreviated PVRAR, was proposed to combine the features of point clouds and multiple views for 3D shape recognition. In the PVRAR, a two-stage approach was used to train the network, which ensured the stability of network training. In particular, we applied the 2D linear Radon transform to process multiple views of each 3D model to enhance the linear and color information. Subsequently, we incorporated the CBAM into our network structure to obtain enhanced point cloud features, which facilitated the calculation of relation

scores and the fusion of multimodal features. Our experimental results indicated that the overall accuracy and mAP of our network were 95.3 % and 93.2 %, respectively, on ModelNet40, and 94.6 % and 93.1 %, respectively, on ModelNet10. These results implied that our method was less affected by the size of the datasets, and that the PVRAR can consider the internal relationship between point clouds and multiple views, as well as obtain feature descriptors with greater expression ability. However, owing to the small number of 3D models in the ModelNet10 dataset, our network could not be trained effectively, thereby resulting in substandard performances occasionally. Therefore, in the future, we will continue to identify more efficient methods for extracting geometric features as well as evaluate our method based on more 3D shape datasets.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is jointly supported by the National Natural Science Foundation of China (No. 61462002), Major special projects of North Minzu University (No. ZDZX-2001801), First class discipline construction of Ningxia's University (mathematics discipline) (No. NXYLXK2017B09) and the Special Project of North Minzu University (No. FWNX21). We would like to thank Editage (`www.editage.cn`) for English language editing.

REFERENCES

[1] Wu, Z.—Song, S.—Khosla, A.—Yu, F.—Zhang, L.—Tang, X.—Xiao, J.: 3D ShapeNets: A Deep Representation for Volumetric Shapes. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1912–1920, doi: 10.1109/CVPR.2015.7298801.

[2] Maturana, D.—Scherer, S.: VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 922–928, doi: 10.1109/IROS.2015.7353481.

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

[4] Feng, Y.—Zhang, Z.—Zhao, X.—Ji, R.—Gao, Y.: GVCNN: Group-View Convolutional Neural Networks for 3D Shape Recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 264–272, doi: 10.1109/CVPR.2018.00035.

[5] Wei, X.—Yu, R.—Sun, J.: View-GCN: View-Based Graph Convolutional Network for 3D Shape Analysis. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1847–1856, doi: 10.1109/CVPR42600.2020.00192.

[6] Li, L.—Zhu, S.—Fu, H.—Tan, P.—Tai, C. L.: End-to-End Learning Local Multi-View Descriptors for 3D Point Clouds. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1916–1925, doi: 10.1109/CVPR42600.2020.00199.

[7] Qi, C. R.—Su, H.—Mo, K.—Guibas, L. J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 77–85, doi: 10.1109/CVPR.2017.16.

[8] Qi, C. R.—Yi, L.—Su, H.—Guibas, L. J.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In: Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.): Advances in Neural Information Processing Systems 30 (NIPS 2017), 2017, pp. 5099–5108.

[9] Rao, Y.—Lu, J.—Zhou, J.: Global-Local Bidirectional Reasoning for Unsupervised Representation Learning of 3D Point Clouds. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5375–5384, doi: 10.1109/CVPR42600.2020.00542.

[10] You, H.—Feng, Y.—Ji, R.—Gao, Y.: PVNet: A Joint Convolutional Network of Point Cloud and Multi-View for 3D Shape Recognition. Proceedings of the 26th ACM International Conference on Multimedia (MM '18), 2018, pp. 1310–1318, doi: 10.1145/3240508.3240702.

[11] You, H.—Feng, Y.—Zhao, X.—Zou, C.—Ji, R.—Gao, Y.: PVRNet: Point-View Relation Neural Network for 3D Shape Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, No. 1, pp. 9119–9126, doi: 10.1609/aaai.v33i01.33019119.

[12] Peng, B.—Yu, Z.—Lei, J.—Song, J.: Attention-Guided Fusion Network of Point Cloud and Multiple Views for 3D Shape Recognition. 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), 2020, pp. 185–188, doi: 10.1109/VCIP49819.2020.9301813.

[13] Yang, J.—Dang, J.: PVFNet: Point-View Fusion Network for 3D Shape Recognition. In: Li, G., Shen, H., Yuan, Y., Wang, X., Liu, H., Zhao, X. (Eds.): Knowledge Science, Engineering and Management (KSEM 2020). Springer, Cham, Lecture Notes in Computer Science, Vol. 12274, 2020, pp. 291–303, doi: 10.1007/978-3-030-55130-8_26.

[14] Tizhoosh, H. R.—Babaie, M.: Representing Medical Images with Encoded Local Projections. IEEE Transactions on Biomedical Engineering, Vol. 65, 2018, No. 10, pp. 2267–2277, doi: 10.1109/TBME.2018.2791567.

[15] Tizhoosh, H. R.—Mitcheltree, C.—Zhu, S.—Dutta, S.: Barcodes for Medical Image Retrieval Using Autoencoded Radon Transform. 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 3150–3155, doi: 10.1109/ICPR.2016.7900119.

[16] Sriram, A.—Kalra, S.—Tizhoosh, H. R.: Projectron – A Shallow and Interpretable Network for Classifying Medical Images. 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–9, doi: 10.1109/IJCNN.2019.8851758.

[17] Luo, Z.—Liu, D.—Li, J.—Chen, Y.—Xiao, Z. et al.: Learning Sequential Slice Representation with an Attention-Embedding Network for 3D Shape Recognition and Retrieval in MLS Point Clouds. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 161, 2020, pp. 147–163, doi: 10.1016/j.isprsjprs.2020.01.003.

[18] Jiang, G.—Jiang, X.—Fang, Z.—Chen, S.: An Efficient Attention Module for 3D Convolutional Neural Networks in Action Recognition. Applied Intelligence, Vol. 51, 2021, No. 10, pp. 7043–7057, doi: 10.1007/s10489-021-02195-8.

[19] Zhang, Y.—Li, H.—Du, J.—Qin, J.—Wang, T.—Chen, Y.—Liu, B.—Gao, W.—Ma, G.—Lei, B.: 3D Multi-Attention Guided Multi-Task Learning Network for Automatic Gastric Tumor Segmentation and Lymph Node Classification. IEEE Transactions on Medical Imaging, Vol. 40, 2021, No. 6, pp. 1618–1631, doi: 10.1109/TMI.2021.3062902.

[20] Li, W.—Qin, S.—Li, F.—Wang, L.: MAD-UNet: A Deep U-Shaped Network Combined with an Attention Mechanism for Pancreas Segmentation in CT Images. Medical Physics, Vol. 48, 2020, No. 1, pp. 329–341, doi: 10.1002/mp.14617.

[21] Zhao, Y.—Jiao, J.—Li, N.—Deng, Z.: MANet: Multimodal Attention Network Based Point-View Fusion for 3D Shape Recognition. 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 134–141, doi: 10.1109/ICPR48806.2021.9413135.

[22] Li, H.—Zheng, Y.—Cao, J.—Cai, Q.: Multi-View-Based Siamese Convolutional Neural Network for 3D Object Retrieval. Computers and Electrical Engineering, Vol. 78, 2019, pp. 11–21, doi: 10.1016/j.compeleceng.2019.06.022.

[23] Liu, S.—Li, T.—Chen, W.—Li, T.—Li, H.: Soft Rasterizer: A Differentiable Renderer for Image-Based 3D Reasoning. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7707–7716, doi: 10.1109/ICCV.2019.00780.

[24] Xu, J.—Zhang, X.—Li, W.—Liu, X.—Han, J.: Joint Multi-View 2D Convolutional Neural Networks for 3D Object Classification. Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence (IJCAI-PRICAI-20), 2020, pp. 3202–3208, doi: 10.24963/ijcai.2020/443.

[25] Ma, C.—Guo, Y.—Yang, J.—An, W.: Learning Multi-View Representation with LSTM for 3-D Shape Recognition and Retrieval. IEEE Transactions on Multimedia, Vol. 21, 2019, No. 5, pp. 1169–1182, doi: 10.1109/TMM.2018.2875512.

[26] Han, Z.—Lu, H.—Liu, Z.—Vong, C. M.—Liu, Y. S.—Zwicker, M.—Han, J.—Chen, C. L. P.: 3D2SeqViews: Aggregating Sequential Views for 3D Global Feature Learning by CNN with Hierarchical Attention Aggregation. IEEE Transactions on Image Processing, Vol. 28, 2019, No. 8, pp. 3986–3999, doi: 10.1109/TIP.2019.2904460.

[27] Phan, A. V.—Nguyen, M. L.—Nguyen, Y. L. H.—Bui, L. T.: DGCNN: A Convolutional Neural Network over Large-Scale Labeled Graphs. Neural Networks,

Vol. 108, 2018, pp. 533–543, doi: 10.1016/j.neunet.2018.09.001.

[28] Zhang, K.—Hao, M.—Wang, J.—de Silva, C. W.—Fu, C.: Linked Dynamic Graph CNN: Learning on Point Cloud Via Linking Hierarchical Features. 2019, arXiv: 1904.10014.

[29] Liu, X.—Han, Z.—Liu, Y.—Zwicker, M.: Point2Sequence: Learning the Shape Representation of 3D Point Clouds with an Attention-Based Sequence to Sequence Network. Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, No. 1, pp. 8778–8785, doi: 10.1609/aaai.v33i01.33018778.

[30] Uy, M. A.—Lee, G. H.: PointNetVLAD: Deep Point Cloud Based Retrieval for Large-Scale Place Recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4470–4479, doi: 10.1109/CVPR.2018.00470.

[31] Arandjelović, R.—Gronat, P.—Torii, A.—Pajdla, T.—Sivic, J.: NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, 2018, No. 6, pp. 1437–1451, doi: 10.1109/TPAMI.2017.2711011.

[32] Zhang, W.—Xiao, C.: PCAN: 3D Attention Map Learning Using Contextual Information for Point Cloud Based Retrieval. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12428–12437, doi: 10.1109/CVPR.2019.01272.

[33] Sriram, A.—Kalra, S.—Babaie, M.—Kieffer, B.—Al Drobi, W.—Rahnamayan, S.—Kashani, H.—Tizhoosh, H. R.: Forming Local Intersections of Projections for Classifying and Searching Histopathology Images. In: Michalowski, M., Moskovitch, R. (Eds.): Artificial Intelligence in Medicine (AIME 2020). Springer, Cham, Lecture Notes in Computer Science, Vol. 12299, 2020, pp. 227–237, doi: 10.1007/978-3-030-59137-3_21.

[34] Eltaieb, R. A.—Farghal, A. E. A.—Ahmed, H. H.—Saif, W. S.—Ragheb, A.—Alshebeili, S. A.—Shalaby, H. M. H.—Abd El-Samie, F. E.: Efficient Classification of Optical Modulation Formats Based on Singular Value Decomposition and Radon Transformation. Journal of Lightwave Technology, Vol. 38, 2020, No. 3, pp. 619–631, doi: 10.1109/JLT.2019.2947154.

[35] Jiang, L.—Yan, L.—Yi, A.—Pan, Y.—Bo, T.—Hao, M.—Pan, W.—Luo, B.: Blind Density-Peak-Based Modulation Format Identification for Elastic Optical Networks. Journal of Lightwave Technology, Vol. 36, 2018, No. 14, pp. 2850–2858, doi: 10.1109/JLT.2018.2827118.

[36] Shifat-E-Rabbi, M.—Yin, X.—Rubaiyat, A. H. M.—Li, S.—Kolouri, S.—Aldroubi, A.—Nichols, J. M.—Rohde, G. K.: Radon Cumulative Distribution Transform Subspace Modeling for Image Classification. 2020, arXiv: 2004.03669.

[37] He, K.—Zhang, X.—Ren, S.—Sun, J.: Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[38] Ding, C.—Tang, L.—Cao, L.—Shao, X.—Wang, W.—Deng, S.: Preprocessing of Multi-Line Structured Light Image Based on Radon Transformation and Gray-Scale Transformation. Multimedia Tools and Applications, Vol. 80, 2021, No. 5, pp. 7529–7546, doi: 10.1007/s11042-019-08031-z.

[39] Kapuriya, B. R.—Pradhan, D.—Sharma, R.: Detection and Restoration of Multi-Directional Motion Blurred Objects. Signal Image and Video Processing, Vol. 13, 2019, No. 2, pp. 1001–1010, doi: 10.1007/s11760-019-01438-z.

[40] Graziano, M. D.—Grasso, M.—D'Errico, M.: Performance Analysis of Ship Wake Detection on Sentinel-1 SAR Images. Remote Sensing, Vol. 9, 2017, No. 11, Art. No. 1107, doi: 10.3390/rs9111107.

[41] Yang, T.—Karakuş, O.—Achim, A.: Detection of Ship Wakes in SAR Imagery Using Cauchy Regularisation. 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 3473–3477, doi: 10.1109/ICIP40778.2020.9190920.

[42] Gómez de Mariscal, E.—Maška, M.—Kotrbová, A.—Pospíchalová, V.—Matula, P.—Muñoz-Barrutia, A.: Deep-Learning-Based Segmentation of Small Extracellular Vesicles in Transmission Electron Microscopy Images. Scientific Reports, Vol. 9, 2019, Art. No. 13211, doi: 10.1038/s41598-019-49431-3.

[43] Woo, S.—Park, J.—Lee, J. Y.—Kweon, I. S.: CBAM: Convolutional Block Attention Module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.): Computer Vision – ECCV 2018. Springer, Cham, Lecture Notes in Computer Science, Vol. 11211, 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

[44] Krizhevsky, A.—Sutskever, I.—Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q. (Eds.): Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012, pp. 1097–1105.

[45] Resnick, J.: The Radon Transforms and Some of Its Applications. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 33, 1985, No. 1, pp. 338–339, doi: 10.1109/TASSP.1985.1164533.

[46] Ben-Shabat, Y.—Lindenbaum, M.—Fischer, A.: 3DmFV: Three-Dimensional Point Cloud Classification in Real-Time Using Convolutional Neural Networks. IEEE Robotics and Automation Letters, Vol. 3, 2018, No. 4, pp. 3145–3152, doi: 10.1109/LRA.2018.2850061.

[47] Kazhdan, M.—Funkhouser, T.—Rusinkiewicz, S.: Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. Eurographics Symposium on Geometry Processing (SGP '03), 2003, pp. 156–164.

[48] Chen, D. Y.—Tian, X. P.—Shen, Y. T.—Ming, O.: On Visual Similarity Based 3D Model Retrieval. Computer Graphics Forum, Vol. 22, 2003, No. 3, pp. 223–232, doi: 10.1111/1467-8659.00669.

[49] Shi, B.—Bai, S.—Zhou, Z.—Bai, X.: DeepPano: Deep Panoramic Representation for 3-D Shape Recognition. IEEE Signal Processing Letters, Vol. 22, 2015, No. 12, pp. 2339–2343, doi: 10.1109/LSP.2015.2480802.

Jie Zhou received her B.Sc. degree in computer science and technology from the Jining Medical University, Jining, China in 2018 and her M.Sc. degree in computer technology from the North Minzu University, Yinchuan, China in 2021. Her research interests are 3D models retrieval, computer vision.

Ziping Ma received her B.Sc. degree from the Computer Institute, North Minzu University, Yinchuan, China in 2003 and her M.Sc. degree from the Institute of Mathematics and Computer, Ningxia University, Yinchuan, China in 2006. She received her Ph.D. degree in information science and technology from the Northwest University, Xi'an, China in 2013. She is currently Associate Professor in the School of Mathematics and Information Science at North Minzu University. Her main research interests are image processing, and 3D models retrieval.

Jinlin Ma received his B.Sc. degree from the North Minzu University, Yinchuan, China in 1999 and his M.Sc. degree from the Ningxia University, Yinchuan, China in 2009. He received his Ph.D. degree in information science and technology from the Northwest University, Xi'an, China in 2011. He is currently Vice Director and Associate Professor of the School of Computer Science and Engineering, North Minzu University. His main research interests are artificial intelligence, image processing, and computer vision.