PVCLN: Point-View Complementary Learning Network for 3D Shape Recognition

SHANLIN SUN¹, YUN LI¹, MINJIE REN², GUO LI¹, AND XING YAO¹
¹College of Electronic Information and Automation, Guilin University of Aerospace Technology, Guilin 541004, China
²School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

Corresponding authors: Yun Li (liyun@guat.edu.cn) and Minjie Ren (renminjie@tju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61861014; in part by the Guangxi Nature Science Fund under Grant 2015GXNSFAA139298 and Grant 2016GXNSFAA380226; in part by the Guangxi University High Level Innovation Team and Outstanding Scholar Program; in part by the Guangxi Thousand Bones Project; in part by the Guangxi Science and Technology Project under Grant AC16380094, Grant AA17204086, and Grant 1598008-29; in part by the Guangxi Nature Science Fund Key Project under Grant 2016 GXNSFDA380031; and in part by the Guangxi University Science Research Project under Grant ZD 2014146.

ABSTRACT

As an important topic in computer vision and multimedia analysis, 3D shape recognition has attracted much research attention in recent years. For point cloud data and multiview data, various approaches have been proposed with remarkable performance. However, few works simultaneously employ the point cloud data and multiview data to represent 3D shapes, which is complementary and beneficial in our consideration. Moreover, existing multimodal approaches mainly focus on the multimodal fusion strategy or on exploring the relation between them. However, the intra-modality characteristic information and inter-modality complementary information are ignored in these methods. In this paper, we tackle the above limitations by introducing a novel Point-View Complementary Learning Network (PVCLN) to explore the potential of both the complementary information and characteristic information for 3D shape recognition. Inspired by the successful application of graph neural networks in capturing relations between features, we introduce a novel multimodal fusion strategy. Concretely, we first separately extract the visual feature from multiview data and structural feature from point cloud data. We then project the visual feature and structural feature into the same feature space to learn the complementary information between two modalities by modeling the inter-modality affinities. The characteristic information in each modality is also preserved by considering the intra-modality affinities. The intra-modality and inter-modality affinities compensate for the lacking characteristic information and enhance the complementary information in the feature learning process. Finally, the updated visual and structural features are further combined to achieve a unified representation for a 3D shape. We conduct extensive experiments to validate the superiority of the overall network and the effectiveness of each component. The proposed method is evaluated on the ModelNet40 dataset and the experimental results demonstrate that our framework achieves competitive performance in the 3D shape recognition task.

INDEX TERMS

3D shape recognition, point cloud, multiview, complementary learning.

I. INTRODUCTION

With a wide range of applications from virtual reality to medical imaging, 3D data recognition and analysis is undoubtedly a fundamental and intriguing area in multimedia and computer vision. Owing to the development of both hardware and deep neural network technologies, tremendous improvements have been achieved in 3D shape recognition with different representations. To deal with the volumetric data, several works [1], [2] utilize 3D convolutional neural networks (CNNs). However, the performance of volumetric data based methods is constrained due to the high computation cost and the sparse structure of volumetric data. Compared with the volumetric data based methods, view based and point cloud-based models achieve more promising performance since the input data are multiple views from camera arrays and point clouds from sensors and LiDARs, respectively, which have rich information and can be easily acquired.
Therefore, multiview data and point cloud data are becoming increasingly popular in dealing with the 3D shape recognition problem.

In multiview based methods, each 3D shape is represented by multiple views generated by the cameras with different angles. The well-established convolutional neural networks (CNNs), such as AlexNet [3], VGG [4], GoogleNet [5], ResNet [6], and DenseNet [7] can be directly applied to extract the visual features from these views to boost the performance. However, the view extraction process of multiview based methods determines the performance and is influenced by the camera angles, and the structural information is inevitably discarded during this process. The multiview based methods concentrate more on the visual information of 3D shapes.

In point cloud-based methods, each 3D shape is represented by a group of points with 3D coordinates captured by the 3D sensing devices, which can better preserve the original 3D spatial information. Although point cloud data are easily acquired and have rich information, it is difficult for a conventional CNN to exploit the spatial information since the point cloud data are disordered and irregular, which further makes the extracted feature ineffective.

Considering the pros and cons of the view-based models and point cloud-based models, it is natural to think that simultaneously employing visual features from the multiview data and structural features from the point cloud data may obtain a better representation for 3D shape. Existing works [8], [9] proposed by You et al. mainly focus on the multimodal fusion strategy and explore the effect of view features with different angles in the feature fusion stage, while the intra-modality characteristic information and inter-modality complementary information are ignored in these methods. In our consideration, because the view based and point cloud-based deep models have achieved promising performance, the visual information from multiview data and spatial information from point cloud data should be preserved as much as possible in the feature learning process, saving the characteristic information of each modality. Moreover, because the multiview data and point cloud data are used to describe the same 3D shape but focus on different aspects, the relation between the extracted features from two kinds of data and complementary information should also be considered in the feature learning process, making the final shape representation more discriminative. Therefore, in the case of simultaneously exploiting the multiview data and point cloud data, it is essential to explore the potential of both the intra-modality characteristic information and inter-modality complementary information to boost the 3D shape recognition performance.

In this paper, we propose a point-view complementary learning network, termed ‘PVCLN’, which can take both the complementary information between two kinds of data and characteristic information from each type of data into consideration for the 3D shape recognition problem. This approach models the affinities between intra-modality and inter-modality samples and utilizes them to propagate information. Every sample accepts the information from its inter-modality and intra-modality neighbors while sharing its own information with them. This scheme can enhance each modality’s characteristic information and explore the correlation between two modalities, thus improving the representation ability. The experimental results show that our framework can outperform not only existing point cloud-based or view-based methods but also multimodal fusion methods. The experimental results confirm that the proposed algorithm achieves a powerful 3D shape representation.

The main contributions of this paper can be summarized as follows:

- We propose a novel point-view complementary learning network (PVCLN) that simultaneously employs visual features from multiview data and structural features from point cloud data. Unlike existing multimodal methods, we jointly consider the characteristic information of each modality and the complementary information from two modalities in the feature learning process.
- We introduce a feature learning method by modeling the inter-modality and intra-modality affinity. The affinities propagate information among and across modalities according to near neighbors, which can effectively utilize the characteristic information and complementary information to obtain discriminative 3D shape descriptors.
- We evaluate the effectiveness of the proposed method on the ModelNet40 dataset. The experimental results demonstrate that the proposed network can achieve competitive performance over state-of-the-art methods.

The rest of this paper is organized as follows. In Section II, we introduce related works. The proposed approach is detailed in Section III. In Section IV, we show relevant experimental results. Insightful analysis of the results is also given in this section. Finally, we conclude this paper and present discussions of future work in Section V.

**II. RELATED WORKS**

With the proliferation of deep learning, 3D shape recognition has attracted much research attention and lots of deep models have been proposed. Since there is a vast amount of...
literature for 3D shape recognition problem, we mainly focus on reviewing related view-based methods, point cloud-based methods and multimodal-based methods to highlight the novelty of our method.

A. POINT CLOUD-BASED METHODS

For point cloud-based models, the input data are a set of points sampled from the surface of the 3D shape. Although the point cloud preserves more completed structure information, the irregular and unstructured data prevent the usage of the conventional 2D CNN, which constrains the performance to some extent. As a pioneering approach, PointNet [10] first introduces a deep neural network to process disordered point cloud data. To be invariant to the permutation, a spatial transform block and asymmetric function max-pooling are applied. PointNet++ is proposed in [11] to employ a PointNet module in local point sets, and then, the local features are aggregated in a hierarchical way. Klokov and Lempitsky [12] proposed Kd networks, which can handle unstructured point clouds and gather the features by the subdivision of points on Kd-trees. Recently, DGCNN [13] was proposed by Wang et al. to better exploit local structure information of 3D point clouds. To both maintain permutation invariance and capture the local geometric features of the point cloud, the authors proposed a new neural network module named EdgeConv. Liu et al. [14] proposed a novel Point2Sequence architecture to capture the correlations between different areas for feature learning, which takes full advantage of contextual information and increases the discrimination of descriptors. The Point2Sequence is the first RNN-based model for local feature capturing and it shows superiority over other point cloud-based methods. Xu et al. [15] proposed GS-Net for point cloud classification and segmentation. Unlike previous works, the GS-Net aggregates features in both Euclidean space and Eigenvalue space. Moreover, the Eigen-Graph is exploited to calculate the structure tensor in order to measure local geometric properties of input points, enabling the network to identify points with similar local structures but located distantly in Euclidean space.

B. VIEW-BASED METHODS

In view-based methods, input data are the views taken by the camera arrays from different angles of the 3D shape. Compared with other data representations, such as voxels and point clouds, these views are more easily acquired. Since 3D shapes can be simply represented by a set of views, the view-based methods have been gaining increasing traction in recent years. Moreover, the mature models, such as VGG [4], GoogleNet [5] and ResNet [6], can be directly utilized in the view feature extraction process of the deep learning-based methods. Aiming to achieve more discerning 3D shape descriptors, GVCNN is proposed by Feng et al. [16] to capture the hierarchical correlation among the multiple views, obtaining competitive performance on the ModelNet40 dataset. Similar to GVCNN, MLVCNN is proposed by Jiang et al. [17] to explore the intrinsic hierarchical associations among the extracted views for the 3D shape recognition. In the framework of MLVCNN, different loop directions are introduced to extract views, and then, the proposed view-loop-shape architecture can hierarchically capture the view level, the loop level, and the shape level information to represent the 3D shape. MLVCNN also achieves superior performance in both classification and retrieval tasks. Inspired by the successful application of n-gram models in the natural language processing field, He et al. [18] proposed the VNN to effectively aggregate all the view features via the utilization of the n-gram mechanism. The VNN captures the spatial information across multiple views by dividing the view sequence into a set of visual n-grams. Promising performance has been achieved by VNN on several benchmark datasets. Han et al. [19] proposed a novel 3D2SeqViews architecture to extract the global features of 3D shape by aggregating sequential views. In addition, the hierarchical attention mechanism is introduced to explore the semantical information of the sequential views, which significantly improves the discrimination of shape descriptors. Feng et al. [20] utilized the Hypergraph Neural Network for the 3D shape classification task. The proposed hypergraph architecture could efficiently explore the high-order correlations between different shapes with the help of hyperedge convolution operation. The novel graph-based network shows the superiority of 3D shape representation.

C. MULTIMODAL-BASED METHODS

Compared with the 2D image, 3D shape usually has more complex structure information, which makes it difficult to describe the 3D shape completely. So the multimodal fusion is necessary and beneficial if implemented in an effective way [21]. In [22], FusionNet jointly combines volumetric data and view data to learn a unified feature representation. MMJN [23] proposed by Nie et al. considers the correlation between different modalities to extract the robust feature vector, employing point cloud data, multiview data, and panorama data. Regarding the fusion of view data and point cloud data, PVNet [8] uses the global view feature to guide the local feature extraction of point cloud. PVRNet [9] introduces a relation score module to investigate the relation between the point cloud and views, and to adequately fuse the view and point cloud features. Our work differs greatly from these approaches by not only exploiting the
relationships between point cloud data and multiview data but also preserving the characteristic information from each modality during the feature learning process.

III. OUR APPROACH

In this section, we give a detailed introduction to our proposed method. In general, the whole architecture can be roughly divided into three parts: (1) Feature extraction: we take the point clouds and 2D rendering views as the shape representations of PVCLN, and then employ the fine-tuned convolutional neural networks separately for feature extraction in each branch. (2) Inter- and intra-modality affinity modeling: we introduce an embedding attention fusion mechanism to incorporate the extracted features from the point cloud and multiview branches, which provides an effective measurement for the contribution and significance of the features from different modalities. (3) Characteristic and complementary information learning: we consider the modality-specific characteristics and cross-modality interactions to generate discriminative shape descriptors for 3D shape classification and retrieval tasks. Fig. 3 illustrates the detailed flowchart of the framework.

A. FEATURE EXTRACTION

1) POINT CLOUD

In the point cloud branch, we employ a set of F-dimensional points $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^n$ for 3D shape representation, where $n$ equals 3. Each point is represented by 3D coordinates. Then we introduce the DGCNN [13] architecture to extract the structural features of the input point cloud. Concretely, DGCNN includes a 3D spatial transform network, several EdgeConv layers, and a max-pooling layer for feature extraction. Note that the framework is compatible with various models for point cloud feature extraction.

2) MULTIVIEW

In the multiview branch, each 3D shape can be represented by a group of rendered views captured by a predefined camera array. The placement of virtual cameras follows the approach mentioned in [24]. Subsequently, we feed these views to the multiview Convolutional Network for feature extraction. After aggregating all the visual features by a view-pooling operation, a global shape descriptor in the multiview branch is generated.

In general, the calculation of the two-stream feature extraction process can be represented as follows:

$$ S^P = \text{Conv}^P X^P, $$
$$ S^V = \text{Conv}^V X^V, $$

(1)

where $S^P$ and $S^V$ represent the extracted feature vectors of the point cloud input $X^P$ and multiview input $X^V$, respectively. $\text{Conv}^P$ and $\text{Conv}^V$ are the aforementioned employed CNN architectures.

To preserve the characteristic information from each modality, we add classification loss $L_C$ in each modality branch:

$$ L_C^P = \mathbb{E}_{d,p}[−\log(p(y_{i}^P|\tilde{S}_i^P))], $$
$$ L_C^V = \mathbb{E}_{d,v}[−\log(p(y_{i}^V|\tilde{S}_i^V))], $$
$$ L_C = L_C^P + L_C^V, $$

(2)

where $p(y_{i}^P)$ and $p(y_{i}^V)$ represent the probabilities of belonging to the ground-truth class $y_{i}^P$ and $y_{i}^V$ for the input point cloud feature $S_i^P$ and multiview descriptor $S_i^V$, respectively.
B. INTER-MODALITY AND INTRA-MODALITY AFFINITY MODELING

To leverage both the modality-specific characteristics and inter-modality correlations, we introduce the intra- and inter-modality affinity measurement to separately represent the similarity between samples from different modalities following [25]. Here, the obtained affinities are used as a guide to the joint feature learning process. Specifically, we first project the features from two modalities into the same space. Concretely, the features of two modalities, \( S^P \) and \( S^V \) pass through the embedding network, which is simplified as a fully connected (FC) layer to project the global features. The obtained features from point cloud modality and multiview modality are denoted as \( P^F \) and \( V^F \), respectively. Note that \( P^F \) and \( V^F \) have the same dimension after passing through the embedding network. In our consideration, \( P^F \) and \( V^F \) both represent the same 3D shapes in a different way and they are both the global feature of the 3D shapes. Therefore, they have complementary information. We model this information as the inter-modality affinities. To shorten the distance between \( P^F \) and \( V^F \), the loss \( L_D \) is introduced to facilitate the inter-modality affinity generation as follows:

\[
L_D = ||P^F - V^F||_2 \quad (3)
\]

Subsequently, we take the features in each modality to compute the intra-modality affinity and the features from different modalities for inter-modality affinity generation. The computation process is presented as the following equation:

\[
A_{ij}^{m,m'} = d(F_i^m, F_j^{m'}), \quad A_{ij}^{m,m'} = d(F_i^m, F_j^{m'}), \quad (4)
\]

where \( A_{ij}^{m,m'} \) is the intra-modality affinity between the \( i \)-th and \( j \)-th samples, and both of them belong to the \( m \) modality. The inter-modality affinity is represented as \( A_{ij}^{m,m'} \). The affinities could be seen as the similarity between the features, which means that the larger \( A_{ij}^{m,m'} \) is, the more similar \( F_i^m \) and \( F_j^{m'} \) are. The \( d(x, y) \) in our work is defined as follows:

\[
d(x, y) = 1 - 0.5 \cdot \frac{x}{\|x\|} - \frac{y}{\|y\|}. \quad (5)
\]

The intra-modality and inter-modality affinity represent the relation between each sample with others of both the same and different modalities. We define the final affinity matrix as:

\[
A = \begin{bmatrix}
\Gamma(A^{P,P} \cdot k) & \Gamma(A^{P,V} \cdot k) \\
\Gamma(A^{V,P} \cdot k) & \Gamma(A^{V,V} \cdot k)
\end{bmatrix},
\quad (6)
\]

where \( \Gamma(\bullet, k) \) is a function that retains the top-k values for each row of a matrix and sets the other values to zero.

C. CHARACTERISTIC AND COMPLEMENTARY INFORMATION LEARNING

The affinity matrix represents the similarity across sample intra- and inter-modalities, which is utilized for the joint feature learning process in the proposed network. We concatenate point cloud and multiview features in the row dimension before the learning process, which means that each row stores the feature of a sample as follows:

\[
F = \left[ F^P \right] V^F. \quad (7)
\]

Inspired by the success of graph convolutional network (GCN) approaches in capturing correlation between the features in many fields, we follow GCN methods to propagate the inter-modality characteristic information and inter-modality complementary information. With the calculation \( d_{ij} = \sum_d A_{ij} \), the diagonal matrix \( D \) is obtained based on the affinity matrix \( A \). The padded features are first propagated with the near neighbor structure \( (D^{-1/2}AD^{-1/2})F) \). Then the features are fused by a learnable nonlinear transformation. Since the affinity matrix contains the sample correlations from intra- and inter-modalities, the propagated features include both the modality-specific characteristics and the cross-modality relevance. Finally, the propagated features \( \tilde{F} \) are calculated by the following equation:

\[
\tilde{F} = \sigma(D^{-1/2}AD^{-1/2}FW), \quad (8)
\]

where \( \sigma \) is the activation function, which is a ReLU in our work, and \( W \) represents the learnable parameters. These propagated features are fed into a feature learning stage to optimize the whole training process. The transferred features \( Z \) are denoted as:

\[
Z = MaxPool \left\{ \left[ \frac{Z^P}{Z^V} \right] \right\} = MaxPool \left\{ \text{Feat}(\tilde{F}) \right\}. \quad (9)
\]

We fuse the transferred point cloud features \( Z^P \) and multiview features \( Z^V \) via the max-pooling operation. The final obtained feature is used as the shape descriptor for the shape classification task, which is followed by simple multilayer perceptrons (MLPs) and softmax function to produce the classification result. In the joint feature learning process, we utilize the classification loss for optimization:

\[
L_F(Z^m) = E_{i, m}[- \log p(y^m_i | Z^m_i)], \quad (10)
\]

where \( p(y^m_i | z^m_i) \) is the predicted probability of belonging to the ground-truth class \( y^m_i \) for the input shape \( i \) and \( Z^m(m \in \{P, V\}) \) indicates the feature vector of the \( i \)-th shape from \( m \) modality. Therefore, the overall loss function for the PVCLN is the combination of the \( L_C \) in the feature extraction process, \( L_D \) in the complementary information learning process and the final classification loss \( L_F \):

\[
L_{all} = \beta_1 L_C + \beta_2 L_D + \beta_3 L_F, \quad (11)
\]

where \( \beta_1, \beta_2 \) and \( \beta_3 \) are hyperparameters to weight the corresponding stage in PVCLN.

IV. EXPERIMENTS

In this section, the experimental results of PVCLN and related analysis are presented. We first provide a comparison between the experimental results with the recent effective methods to demonstrate the superiority of our method on
the classification and retrieval tasks. Next, an ablation experiment is conducted to further investigate the influence of different components on the performance of the proposed network. Finally, we investigate the effect of different number of views and points of PVCLN toward the classification performance.

A. DATASET
We use the ModelNet40 dataset to evaluate the performance of the proposed method on the 3D model recognition task. As a subset of ModelNet, ModelNet40 contains 40 categories with 12,311 CAD models, consisting of 9843 and 2468 models in the training subset and test subset, respectively. The 3D shapes are represented as polygonal meshes in the ModelNet40 dataset. We follow the MVCNN [24] to render the models into multiple views. Concretely, we place 12 virtual cameras pointed toward the centroid of the shape and elevated 30 degrees from the ground plane. Therefore, there are 12 views extracted from each shape for the multiview data. The point cloud data are sampled from the surface of each model as in [10].

B. IMPLEMENTATION DETAILS
In our PVCLN, the pretrained MVCNN and DGCNN are employed for view feature extraction and point cloud feature extraction, respectively. Note that, any view-based and point cloud-based models could be used to extract the global point cloud feature and view feature. The embedding network is simplified as a FC layer with satisfactory performance on projecting global features. The projected global features from two modalities are used to compute the inter-modality affinities and intra-modality affinities. The hyperparameters $\beta_1$, $\beta_2$ and $\beta_3$ in Eq.11 are set to 0.5, 0.1 and 1.0, respectively. PVCLN is trained in an end-to-end fashion. Following [8], we adopt the alternative optimization strategy to update our framework. In particular, we first freeze the parameters of multiview branch and only update our point cloud branch for some epochs, and then all the parameters are updated together for some epochs. All experiments are conducted on the PyTorch platform and we test our PVCLN on an Intel(R) Core(TM) i7-9700K CPU @ 3.60GHz CPU system, with 32 G RAM and a GeForce GTX1080 Ti GPU, with 12 G RAM. The training process is accelerated via the CUDA instruction set on the GPU. The learning rate is set as 0.0001, and the batch size is set as 16. We choose the best result to compare with current state-of-the-art methods.

C. COMPARISON WITH STATE-OF-THE-ART METHODS ON THE MODELNET40 DATASET
To demonstrate the effectiveness of PVCLN, 3D shape classification and retrieval experiments are conducted on the ModelNet40 dataset. The training and testing sets are predefined in the ModelNet40, and we directly utilize the constructed database in the following experiments. Various models based on different representations are used as comparison methods in experiments, including volumetric-based models (3D ShapeNets [2], VoxNet [26], VRN [27] and MVCNN-MultiRes [1]), handcrafted descriptors for multiview data (LFD [28]), deep learning models for multiview data (MVCNN [24] and GVCNN [16]), point cloud-based models (PointNet [10], PointNet++ [11], Kd-Network [12], DGCNN [13] and GS-Net [15]) and multimodal methods (FusionNet [22], PVNet [8] and PVRNet [9]). The comparison results are presented in Table 1.

From the experimental results in Table 1, we find that the proposed PVCLN outperforms all the comparison methods, with a classification accuracy of 93.8% and retrieval mAP of 90.9% on the ModelNet40 dataset. In addition, we make the following observations:

- Among the view-based methods, PVCLN achieves the best classification and retrieval results. The MVCNN is a typical view-based method for shape representation. It employs the view-pooling mechanism for multiview feature fusion, which ignores the correlations between views. Compared with the MVCNN method with GoogleNet as the backbone, the proposed PVCLN exhibits gains of 1.6% and 7.9% in terms of overall accuracy and retrieval mAP, respectively. GVCNN is an optimization method for MVCNN. It introduces the grouping module to group the views according to their content information. As a result, the views are grouped into different clusters with different weights, which increases the discrimination of shape descriptors. Compared with the GVCNN, the proposed PVCLN pays more attention to the cross-modality interaction. The view-based characteristics are also considered, which brings an improvement of 1.2% and 5.2% on the classification task and retrieval task, respectively.

- Among the point cloud-based methods, PVCLN outperforms not only the baseline point cloud method (“DGCNN” by a large margin but also the state-of-the-art point cloud method “GS-Net” by approximately 0.9% in overall accuracy. The DGCNN utilizes the EdgeConv operation to capture the local geometric features of point clouds, which increases the discrimination of point cloud descriptors. Additionally, the “GS-Net” utilizes the Eigen-Graph to capture the geometric features of points. Moreover, these features are invariant to rotation and translation, improving the robustness and effectiveness of the GS-Net. PointNet, PointNet++, and KD-Network are traditional point cloud-based methods to explore the spatial relations between points. The superior performance is attributed to the exploitation of the complementary information between two modalities.

- Among the multimodal based methods, “PVRNet”, which is an optimization of PVNet, mainly focuses on the relation exploration between the extracted views and point clouds, and a novel fusion strategy is also introduced to boost the performance. Different from PVRNet, PVCLN pays more attention to enhance the characteristic information of two modalities, while extracting the complementary information.
TABLE 1. Comparisons of classification and retrieval experimental results on the ModelNet40 dataset.

| Method                          | Data Representation | Classification (Overall Accuracy) | Retrieval (mAP) |
|---------------------------------|---------------------|-----------------------------------|-----------------|
| (1) 3D ShapeNets [2]            | Volumetric          | 77.3%                             | 49.2%           |
| (2) VoxNet [26]                 | Volumetric          | 83.0%                             | -               |
| (3) VRN [27]                    | Volumetric          | 91.3%                             | -               |
| (4) MVCNN-MultiRes [1]          | Volumetric          | 91.4%                             | -               |
| (5) LFD [28]                    | 10 Views            | 75.5%                             | 40.9%           |
| (6) MVCNN (AlexNet) [24]        | 12 Views            | 89.9%                             | 80.2%           |
| (7) MVCNN (GoogleNet) [24]      | 12 Views            | 92.2%                             | 84.0%           |
| (8) GVCNN (GoogleNet) [16]      | 8 Views             | 93.1%                             | 84.5%           |
| (9) GVCNN (GoogleNet)           | 12 Views            | 92.6%                             | 85.7%           |
| (10) MVCNN with TCL [29]        | 12 Views            | -                                 | 88.0%           |
| (11) PointNet [10]              | Point Cloud         | 89.2%                             | -               |
| (12) PointNet++ [11]            | Point Cloud         | 90.7%                             | -               |
| (13) KD-Network [12]            | Point Cloud         | 91.8%                             | -               |
| (14) PointCNN [30]              | Point Cloud         | 91.8%                             | -               |
| (15) DGCNN [13]                 | Point Cloud         | 92.2%                             | 81.6%           |
| (16) GS-Net [15]                | Point Cloud         | 92.9%                             | -               |
| (17) FusionNet [22]             | Volumetric and 20/60 Views | 90.8%                       | -               |
| (18) PVNet [8]                  | Point Cloud and 12 Views | 93.2%                           | 89.5%           |
| (19) PVRNet [9]                 | Point Cloud and 12 Views | 93.6%                         | 90.5%           |
| (20) PVCLN                      | Point Cloud and 12 Views | 93.8%                         | 90.9%           |

TABLE 2. Ablation study on PVCLN.

|               | Point Cloud Branch | multiview Branch | Late Fusion | Complementary Learning | $L_C$ | $L_D$ | Overall Accuracy |
|---------------|--------------------|------------------|------------|------------------------|-------|-------|-----------------|
| 1             | ✓                  | -                | -          | -                      | -     | -     | 92.2%           |
| 2             | -                  | ✓                | -          | -                      | -     | -     | 89.9%           |
| 3             | ✓                  | ✓                | ✓          | ✓                      | -     | -     | 92.62%          |
| 4             | ✓                  | ✓                | ✓          | ✓                      | ✓     | -     | 93.25%          |
| 5             | ✓                  | ✓                | ✓          | ✓                      | ✓     | ✓     | 93.41%          |
| 6             | ✓                  | ✓                | ✓          | ✓                      | ✓     | ✓     | 93.77%          |

between two modalities by modeling the intra-modality and inter-modality affinities. Therefore, the performance of PVCLN also exceeds that of PVRNet. Fig. 4 provides the precision-recall curves of the comparison methods and PVCLN. It is shown that our PVCLN outperforms all other compared methods with the mAP of 90.9%. These experimental results validate the promising discriminative capacity of our method for 3D shape recognition.

D. ABLATION STUDY

We ablate our method in this section and evaluate the performance of each component separately to see the influence. The results are presented in Table 2. The influences on the utilization of multimodal information are shown in the 1st 3rd row. “Point-Cloud Branch” and “multiview Branch” denote the employed feature extractor for point cloud data and multiview data, respectively. The “Late Fusion” means the global features from the point cloud branch and multiview branch are concatenated together for shape representation. Compared to the “Point Cloud Model” and “multiview Model”, the late fusion strategy leads to increases of approximately 0.42% and 2.72% in overall accuracy, respectively, which demonstrates that 3D shape recognition benefits from the multimodal information. The influences on the components of the complementary learning network are shown in
S. Sun et al.: PVCLN: Point-View Complementary Learning Network for 3D Shape Recognition

FIGURE 4. Precision-recall curves of comparison methods and the proposed network on the ModelNet40 dataset.

rows 4–6. It can be seen that our PVCLN even without the two introduced loss functions, also outperforms the late fusion method by about 0.63% overall accuracy, which indicates that our framework can better fuse multimodal information to achieve a more discriminative shape descriptor. The loss function $L_C$ is introduced to enhance the characteristic information of each modality for the complementary learning in the next stage. We can see that the utilization of $L_C$ improves the performance of our PVCLN from 93.25% to 93.41%, which validates the effectiveness of enhancing the characteristic information. To better model the inter-modality affinities, $L_D$ is used to shorten the feature distance between two modalities and improves the overall accuracy by 0.36%. This experiment demonstrates the reasonableness of the proposed PVCLN, showing that the complementary information between modalities can indeed contribute to the shape recognition in our network.

E. SENSITIVITY ANALYSIS ON THE NUMBER OF VIEWS
Since the number of extracted view images directly affects the discrimination of shape descriptors in the multiview branch, we conduct comparative experiments to select the optimal view number. In the view rendering process, we set up the virtual cameras around the z-axis with different intervals to control the number of taken views. Concretely, the interval angle $\theta$ is set to $\{90^\circ, 60^\circ, 45^\circ, 36^\circ, 30^\circ, 18^\circ\}$, separately, which means there are $\{4, 6, 8, 10, 12, 20\}$ views are generated for shape representation.

First, we keep the number of points in the point cloud data branch as constant and vary the view number for comparison. The classification results are presented in Table 3. From the experimental results, we make the following observations:

- As the number of taken views increases, the performance of our proposed network continues improving until the view number reaches 12. Since multiple views can reflect the characteristics and visual details from different angles, the introduction of more effective information significantly improves the discrimination of shape descriptors, which deserves better classification performance;
- When the number of projected views increases over 12, redundant information for representing the 3D shape is introduced, which leads to worsened classification performance.

In general, increasing the view number negatively affects the discrimination of multiview shape descriptors. However, the contribution of view number is limited since excessive views introduce visual redundancy in the shape descriptor generation process. Compared with other numbers of projected views, we achieve the best classification result when $\theta$ is set to 12, and we select it as the optimal view number in the experiments.

F. SENSITIVITY ANALYSIS ON THE NUMBER OF POINTS
The number of points in the point clouds directly determines the amount of effective information and structural details as shown in Fig. 5, which further influences the discrimination of shape descriptors in the point cloud branch. We conduct comparative experiments on the number of points to quantify its impact. Concretely, we retain the multiview data as constant and vary the number of points from $\{128, 256, 384, 512, 640, 768, 1024\}$. The experimental results of the overall accuracy are shown in Fig. 6 with the comparison of our point cloud branch backbone DGCNN.

From the experimental results, we can first observe that with an increasing number of input points, the classification performance improves better due to the introduction of descriptive information. The missing data lead to the defects of structural representation. Second, the obvious drop in the number of points does not have a significant impact on the experimental results, and even as the number of input points is decreased to 256, our framework still achieves a good performance. The architectural stability comes from the compensation of multiview data, which reduces the negative influence of missing data. In general, the interaction of multimodality data benefits the modality-specific feature learning, which also improves the robustness of our network.

### TABLE 3. Performance(%) on ModelNet40 with different view numbers.

| Methods | Number of Views | Overall Accuracy |
|---------|-----------------|------------------|
| MVCNN   | 4               | 84.6             |
|         | 8               | 89.0             |
|         | 10              | 89.3             |
|         | 12              | 89.9             |
| GVCNN   | 4               | 90.3             |
|         | 8               | 92.1             |
|         | 10              | 92.4             |
|         | 12              | 92.6             |
| PVCLN   | 4               | 91.8             |
|         | 8               | 93.3             |
|         | 10              | 93.7             |
|         | 12              | 93.8             |

3458

VOLUME 9, 2021
characteristic information but also enhances the overall discrimination of 3D shapes. The experimental results on the public ModelNet40 dataset have demonstrated the effectiveness of the proposed network, which means that the complementary information between the modalities is crucial for multimodal methods in 3D shape recognition. Related experimental results also showed that the proposed method achieves robust and discriminative representations for 3D shapes. In future work, there are multiple venues for us to study. One natural question that we would like to ask is whether the proposed framework performs well on other multimodal tasks. For the multimodal 3D shape recognition task, we plan to further explore effective fusion strategy and capture cross-modal interactions by graph neural networks.

V. CONCLUSION
There exist several challenges in improving the classification and retrieval performances. First, most of the methods, which utilize the characteristics of a single modality, are uninformative and incomplete for shape representation. Exploring more effective approaches utilizing multimodality information could be one research direction. In addition, for multimodality fusion methods, determining how to explore the cross-modality relations and contain the modality-specific characteristics is a key problem. To address these limitations, we have presented a novel multimodal network PVCLN based on complementary information learning between the point cloud data and multi-view data. Moreover, PVCLN utilizes the characteristic information from each modality, which is ignored by conventional multimodal methods. Concretely, the characteristic information and complementary information are modeled as intra-modality affinities and inter-modality affinities, respectively. Inspired by the success of GCN methods, PVCLN propagates information among and across modalities by introducing the adjacency matrices, which not only compensates for the lacking

REFERENCES

[1] C. R. Qi, H. Su, M. NieBner, A. Dai, M. Yan, and L. J. Guibas, “Volumetric and multi-view CNNs for object classification on 3D data,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 5648–5656.
[2] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, “3D ShapeNets: A deep representation for volumetric shapes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1912–1920.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556
[5] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.
[6] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.
[8] H. You, Y. Feng, R. Ji, and Y. Gao, “PVNet: A joint convolutional network of point cloud and multi-view for 3D shape recognition,” in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 1310–1318.
[9] H. You, Y. Feng, X. Zhao, C. Zou, R. Ji, and Y. Gao, “PVRNet: Point-view relation neural network for 3D shape recognition,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 9119–9126.
[10] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 652–660.
[11] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep hierarchical feature learning on point sets in a metric space,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5099–5108.
[12] R. Klokov and V. Lempitsky, “Escape from cells: Deep kd-networks for the recognition of 3D point cloud models,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 863–872.
[13] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph CNN for learning on point clouds,” ACM Trans. Graph., vol. 38, no. 5, p. 146, 2019.
[14] X. Liu, Z. Han, Y.-S. Liu, and M. Zwicker, “Point2sequence: Learning the shape representation of 3D point clouds with an attention-based sequence to sequence network,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8778–8785.
[15] M. Xu, Z. Zhou, and Y. Qiao, “Geometry sharing network for 3D point cloud classification and segmentation,” in Proc. AAAI, 2020, pp. 12500–12507.
[16] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao, “GVCNN: Group-view convolutional neural networks for 3D shape recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 264–272.
[17] J. Jiang, D. Yao, Z. Chen, X. Zhao, and Y. Gao, “MLVCNN: Multi-loop-view convolutional neural network for 3D shape retrieval,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8513–8520.
[18] X. He, T. Huang, S. Bai, and X. Bai, “View N-Gram network for 3D object retrieval,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7515–7524.

[19] Z. Han, H. Lu, Z. Liu, C.-M. Vong, Y.-S. Liu, M. Zwicker, J. Han, and C.-L. P. Chen, “3DSeqViews: Aggregating sequential views for 3D global feature learning by CNN with hierarchical attention aggregation,” IEEE Trans. Image Process., vol. 28, no. 8, pp. 3986–3999, Aug. 2019.

[20] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, “Hypergraph neural networks,” in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 3558–3565.

[21] P. Gao, H. Li, S. Li, P. Lu, Y. Li, S. C. Hoi, and X. Wang, “Question-guided hybrid convolution for visual question answering,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 469–485.

[22] V. Hegde and R. Zadeh, “Fusionnet: 3D object classification using multiple data representations,” 2016, arXiv:1607.05695 [Online]. Available: https://arxiv.org/abs/1607.05695

[23] W. Nie, Q. Liang, A.-A. Liu, Z. Mao, and Y. Li, “MMIN: Multi-modal joint networks for 3D shape recognition,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 908–916.

[24] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, “Multi-view convolutional neural networks for 3D shape recognition,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 945–953.

[25] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, “Cross-modality person re-identification with shared-specific feature transfer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13379–13389.

[26] D. Maturana and S. Scherer, “VoxNet: A 3D convolutional neural network for real-time object recognition,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2015, pp. 922–928.

[27] A. Brock, T. Lim, J. M. Ritchie, and N. J. Weston, “Generative and discriminative voxel modeling with convolutional neural networks,” in Proc. Neural Inf. Process. Conf. 3D Deep Learn., 2016, pp. 1–9.

[28] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, “On visual similarity based 3D model retrieval,” in Computer Graphics Forum, vol. 22, no. 3, Hoboken, NJ, USA: Wiley, 2003, pp. 223–232.

[29] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, “Triplet-center loss for multi-view 3D object retrieval,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1945–1954.

[30] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, “PointCNN: Convolution on X-transformed points,” in Proc. Adv. Neural Inf. Process. Syst., 2018, pp. 820–830.

SHANLIN SUN received the B.S. degree in communication and information and the M.S. and Ph.D. degrees in information and communications engineering from Tianjin University, China, in 1997, 2006, and 2009, respectively. He is currently a Professor with the Guilin University of Aerospace Technology. His research interest includes image processing.

YUN LI received the B.S. degree in electrical engineering and automation from Central South University, China, in 2002, the M.S. degree in communication and information systems from Sun Yat-sen University, China, in 2008, and the Ph.D. degree in information and communications engineering from Tianjin University, China, in 2018. She is currently a Professor with the Guilin University of Aerospace Technology. Her research interest includes underwater wireless sensor networks.

MINJIE REN is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University. Her research interests include 3D shape recognition and cross domain learning.

GUO LI received the M.S. degree in software engineering from Xidian University, China, in 2019. He is currently an Engineer with the Guilin University of Aerospace Technology. His research interest includes underwater wireless sensor networks.

XING YAO received the B.S. degree in detection guidance and control technology from the North University of China, in 2012, and the M.S. degree in arms engineering from the Beijing Institute of Technology, China, in 2015. He is currently a Teacher with the Guilin University of Aerospace Technology. His research interests include application of UAV and underwater acoustic communication.

* * *