中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)03-0811-13

论文引用格式:Wu H, Hu L C, Yang Y, Jie B and Luo Y L. 2025. Multiview consistent and complementary information fusion method for 3D model classification. Journal of Image and Graphics, 30(3):0811-0823(吴晗, 胡良臣, 杨影, 接标, 罗永龙. 2025. 融合多视图一致和互补信息的深度3D模型分类. 中国图象图形学报, 30(3):0811-0823)[DOI:10.11834/jig.240060]

融合多视图一致和互补信息的深度3D模型分类

吴晗,胡良臣*,杨影,接标,罗永龙

1. 安徽师范大学计算机与信息学院,芜湖　241002; 2. 工业智能数据安全安徽省重点实验室,芜湖　241002

摘　要:**目的**　基于深度学习的方法在3D模型分类任务中取得了先进的性能,此类方法需要提取三维模型不同数据表示的特征,例如使用深度学习模型提取多视图特征并将其组合成单一而紧凑的形状描述符。然而,这些方法只考虑了多视图之间的一致信息,而忽视了视图与视图之间存在的差异信息。为了解决这一问题,提出了新的特征网络学习3D模型多视图数据表示的一致信息和互补信息,并将其有效融合,以充分利用多视图数据的特征,提高3D模型分类准确率。**方法**　该方法首先在残差网络的残差结构中引入空洞卷积,扩大卷积操作的感受野。随后,对网络结构调整以进行多视图特征提取。然后,通过设计的视图分类网络划分一致信息和互补信息,充分利用每个视图。为了处理这两类不同的信息,引入了一种结合注意力机制的学习融合策略,将两类特征视图融合,从而得到形状级描述符,实现可靠的3D模型分类。**结果**　该模型的有效性在ModelNet数据集的两个子集上得到验证。在基于ModelNet40数据集的所有对比方法中具有最好的性能表现。为了对比不同的特征提取网络,设置单分类任务实现性能验证,本文方法在分类准确度和平均损失方面表现最好。相较于基准方法—多视图卷积神经网络(multi-view convolutional neural network, MVCNN),在不同视图数下本文方法的性能最高提升了3.6%,整体分类准确度提高了5.43%。实验结果表明,相较于现有相关方法,本文方法展现出一定的优越性。**结论**　本文提出的一种多视图信息融合的深度3D模型分类网络,深度融合多视图的一致信息和互补信息,在3D模型分类任务中获得明显的效果。

关键词:多视图学习;3D模型分类;一致性与互补性;改进残差网络;视图融合

Multiview consistent and complementary information fusion method for 3D model classification

Wu Han, Hu Liangchen*, Yang Ying, Jie Biao, Luo Yonglong

1. School of Computer and Information, Anhui Normal University, Wuhu　241002, China;

2. Anhui Provincial Key Laboratory of Industrial Intelligence Data Security, Wuhu　241002, China

Abstract: **Objective**　3D model classification holds considerable promise across diverse applications, including autonomous driving, game design, and 3D printing. With the rapid advancement of deep learning, numerous deep neural networks have been investigated for 3D model classification. Among these approaches, view-based methods consistently out-

收稿日期:2024-02-06;**修回日期**:2024-09-04;**预印本日期**:2024-09-11

*通信作者:胡良臣　cslc.hu@ahnu.edu.cn

基金项目:国家自然科学基金项目(62306010, 62272006);安徽省高校自然科学研究项目(2022AH040028);安徽师范大学高峰和奖补学科建设项目(2023GFXK171);安徽省自然科学基金项目(2408085MF169)

Supported by: National Natural Science Foundation of China (62306010, 62272006); Anhui University Natural Science Research Project (2022AH040028); Anhui Normal University High-Peak and Incentive Discipline Construction Project (2023GFXK171); Natural Science Foundation of Anhui Province, China (2408085MF169)

perform voxel mesh and 3D point cloud-based methods. The view-based method captures multiple 2D perspectives from various angles of 3D objects to represent their 3D information. This approach closely aligns with human visual processing, transforming 3D problems into manageable 2D tasks that can be solved using standard convolutional neural networks (CNNs). In contrast, voxel-based and point cloud-based methods primarily focus on the spatial characteristics of 3D models, necessitating the generation of substantial datasets. The view-based method visually transforms 3D challenges into 2D tasks by obtaining multiple 2D views from different angles of 3D models, mirroring human approaches and facilitating resolution through conventional CNNs. The utilization of CNNs typically involves employing established models such as the visual geometry group (VGG) network, the inception network (GoogLeNet), and the residual network (ResNet) to derive a view representation of 3D models. Methods such as MVCNN and the group-view convolutional neural network leverage pretrained network weights to obtain view descriptors from multiple perspectives. However, these approaches often neglect the complementary information between views, an aspect crucial for shaping the final descriptor. As shape descriptors are integral to the recognition task, acquiring 3D model shape descriptors through view descriptors remains a fundamental challenge for achieving optimal 3D model classification. Recent studies, including MVCNN and the dynamic routing convolutional neural network (DRCNN), employ a view pooling scheme to generate discriminative descriptors from feature representations of multiple views, marking crucial milestones in 3D model classification with notable performance improvements. Notably, existing methods inadequately exploit the view characteristics among multiple perspectives, severely limiting the efficacy of shape descriptors. On the one hand, the inherent differences in the two-dimensional views projected from three-dimensional objects constitute complementary information that enhances the generation of the final shape descriptor. On the other hand, each 2D view can, to some extent, represent its corresponding 3D object, indicating consistent features between views. Integrating these consistent features enhances the accuracy of recognition tasks. Consequently, learning complementary information between views and integrating it with consistent information emerges as a critical aspect for advancing 3D model classification. **Method** Addressing this challenge, our paper introduces a network model that integrates complementary and consistent information derived from multiple views, thereby enhancing the overall comprehensiveness of the information. Specifically, the model aims to fuse association information between views for 3D object classification. Initially, an enhanced residual network is employed to extract feature representations from multiple views, resulting in view descriptors. Subsequently, a pre-classification network, combined with an attention mechanism and a weight learning strategy, is utilized to merge these view descriptors and generate shape descriptors. Multiscale dilated convolution is introduced following ordinary convolution within the residual module during one-way network propagation in a single view to improve the residual structure of the ResNet model. This improvement extends the receptive field of the convolution operation, facilitating the extraction of complementary information. Additionally, a pre-classification module is proposed to assess the recognition degree of each view based on shape characteristics. Using this information, the views are categorized into complementary and consistent groups. A subset of both types of views is combined into feature views, each possessing two key characteristics. These feature views are then input into an attention network to strengthen consistency and complementarity. Subsequently, a learnable weight fusion module is applied to the two feature views, weighting and merging them to generate shape descriptors. Finally, the overall network structure is refined and the pre-classification and attention layers are strategically positioned to ensure optimal performance for the proposed methodology. **Result** This study conducted a series of experiments to validate the effectiveness of the proposed model on the ModelNet10 and ModelNet40 datasets. Initially, an ablation experiment was performed to compare the positions of module insertion, feature extraction networks, and the number of views. Experimental results indicate that tightly coupling the pre-classification module with the attention module and placing them in the second or third layer of the residual network yields superior final classification accuracy compared to inserting them in other layers. The method introduced in this study demonstrates higher average single-class classification accuracy than models such as ResNet50 with an equivalent number of training iterations. Additionally, this method exhibits lower average losses and more robust convergence of losses in terms of loss reduction. The performance of the model is further evaluated across varying numbers of views. Regardless of the number of views, the method consistently outperforms or matches MVCNN and DRCNN. As the number of views increases from 3 to 6 to 12, the proposed model and the DRCNN model exhibit a continuous improvement in accuracy. Finally, when compared to 15 classic multiview 3D

object classification algorithms, the proposed model achieved an average instance accuracy of 97.27% on ModelNet10 using the designed feature extraction network with the same 12 views. However, this model falls slightly behind DRCNN, possibly due to limited data in the ModelNet10 dataset, which may lead to overfitting and reduced classification accuracy. In ModelNet40, the model achieved an average instance accuracy of 96.32%. Comparisons with ResNet50 and VGG-M also highlight the advantages of the model in classification accuracy, confirming its capability to more effectively extract information between views on the backbone network than other methods. **Conclusion** This study presents a robust deep 3D object classification method that leverages multiview information fusion to integrate consistency and complementarity information among views through adjustments in the network structure. The experimental findings substantiate the favorable performance of the proposed model.

Key words: multi-view learning; 3D model classification; consistency and complementarity; improving residual network; view fusion

0 引　言

近年来，计算机辅助设计与三维建模技术等领域得到了快速发展，3D模型已经成为继图像与声音信息后不可或缺的数据类型，是信息的重要载体之一。3D模型在虚拟现实、无人驾驶、工业设计和分子生物学等领域具有广泛的应用，数据量呈指数级增长。3D模型分类是诸多应用的研究基础，对构建完整的数字几何研究具有重要的意义。

目前，伴随着深度学习的蓬勃发展，点云数据、体素数据、视图数据作为3D模型的典型数据表现形式，与深度学习结合后都体现出相应的优势。基于体素的方法识别3D模型为规则的3D网格，考虑其具有的空间特性，Brock等人(2016)使用基于体素的变分自编码器和集成学习来处理3D物体分类，并取得了一定效果。Malik等人(2017)结合基于截断符号距离函数和3D体素化深度图进行三维手型和姿态估计任务。基于点云的方法是将三维形状数据看做空间中无序点的集合进行学习。Charles等人(2017)提出一种处理点云集合的方法PointNet，通过独立计算点云特征，并使用“最大池化”此类的阶不变函数聚合。此后，许多方法探索局部结构以获得更好的三维物体表征，如Gao等人(2021)提出了融合体素描述符和形状分布描述符的3D模型分类算法，Gao等人(2023)根据变换网络在二维计算机视觉任务的启发，提出了一种学习局部特征和点云之间相关性的局部特征变换网络。武斌等人(2024)针对三维点云分类，提出空间结构卷积并结合注意力机制学习关联局部和全局特征表示，有效提取并保留更具代表性和差异性的特征信息。

利用体素和点云来描述3D模型，所产生的数据规模较大，并且主要利用3D模型空间特性。而基于视图的方法从三维物体的不同角度获取多个2D视图来表示3D模型，这种方法从视觉上将3D问题变成2D问题，Su等人(2015)提出多视图卷积神经网络(multi-view convolutional neural network，MVCNN)方法，首先使用卷积神经网络提取每个视图的特征，然后使用Max-pooling来聚合不同视图的特征。一些后续工作提出了不同的策略来为视图分配权重，以执行视图特定特征的加权平均池化。Feng等人(2018)提出了视图组卷积神经网络(group-view convolutional neural network，GVCNN)，引入了一种分层形状描述框架。该框架包含视图、组和形状描述符，通过分组学习形状描述符，修改了视图选择机制。Kanezaki等人(2021)则提出了RotationNet模型，除了提取视图特征，还将视点标签作为潜在变量，并联合估计其姿态和对象类别。Yu等人(2018)将协调双线性池化作为一个层，构成多视图协调双线性网络(multi-view harmonized bilinear network，MHBN)。同样地，为了选择具有代表性的视图，Sun等人(2021)提出了动态路由卷积神经网络(dynamic routing convolutional neural network，DRCNN)。该方法构建了一种新的动态路由层，通过动态选择特征视图，以更有效地融合各个视图的特征。此外，Xu等人(2021)提出对应感知表示(correspondence-aware representation，CAR)模块，并构建CAR-Net在感知特征时考虑视点位置与视图内位置之间的空间关系。Liu等人(2022)提出基于视图过滤的多视图聚合框架，采用基于投票的视图过滤策略选择具有代表性的视图。Song等人(2024)将基于对比学习的多模态模型CLIP(contrastive language-image pre-training)融

入多视图三维模型识别中，提出MV-CLIP（multi-view CLIP）方法，通过视图选择和分层提示来提高预训练的语言图像模型在三维模型识别方面的置信度。

基于视图的深度学习模型在其视图特征提取网络上也具有不同的策略。在MVCNN中，使用了Simonyan和Zisserman（2015）提出的VGG（Visual Geometry Group）网络作为骨干网络进行特征提取，在训练集中对所有视图进行微调。其后，Su等人（2018）利用单个视图对骨干网络进行训练，再将通过训练后的骨干网络用于多个视图的特征提取，并且对比了VGG、GoogLeNet（Szegedy等，2015）和ResNet（residual network）（He等，2016）等骨干网络的性能，证明了微调和预训练可以显著提高网络性能。除了使用VGG网络作为特征提取网络，GVCNN采用GoogLeNet作为骨干网络；在RotationNet中，使用AlexNet（Krizhevsky等，2012）作为骨干网络进行特征提取；而Wei等人（2020）提出的基于视图的图卷积网络（view-based graph convolutional network，View-GCN）利用混合的多视图图像对预训练的ResNet18网络进行微调，将不同视图的特征矢量化为视图特征，并在View-GCN的基础上提出具有局部注意图卷积操作和用于图粗化的旋转鲁棒视图采样操作的View-GCN++（Wei等，2023）方法。此外，为了解决多视图CNN模型不能建模不同视图像素之间通信的问题，Chen等人（2021）利用VisionTransformer本身具有的特性，学习多视图像素之间的信息；Li等人（2023）则提出一种CNN与Transformer混合的特征提取网络，提取每个视图的多尺度信息并捕获不同多尺度信息的相关性。

目前，学习3D模型特征的方法普遍采用视图池层及其优化策略进行视图选择。同时，在获取特征级别的视图表示时，使用在大规模数据集上进行预训练的深度学习模型来进行3D模型的浅层特征提取。这些策略主要通过学习视图数据间的一致信息进行分类，但却忽略了视图之间存在的互补信息。事实上，不同视图具有代表同一物体的不同特征，而这些互补的特征对于提高形状描述符的性能至关重要。在现有方法中，多个视图之间存在的视图特性并没有得到充分利用，这显著地限制了形状描述符的性能。一方面，由三维物体投影得到的二维视图之间存在一定的差异性，即互补信息，这些信息可以增强最终形状描述符所具有的物体特征。另一方面；每个二维视图在一定程度上能够代表其对应的三维物体，即视图间存在一致特征。深度融合这些一致性信息和互补性信息可以显著提高识别任务的精度。因此，关键在于如何学习视图之间的互补信息，并将其与一致信息融合，这是非常重要的问题。

为解决上述问题，本文提出了一种深度3D模型分类算法，专注于多视图信息的融合，以便通过关联信息进行分类。本文的主要贡献包括：1）与当前基于视图的全局特征池化方法不同，本文采用了先分类再融合的策略，通过这种方式获得具有互补和一致信息的形状描述符，以解决三维形状识别问题。通过引入一个可学习的小型权重迭代网络，能够更有效地融合多视图信息，提供更具判别性的3D形状表示。2）针对视图特征提取网络，本文对残差结构进行了调整，并重新设计了网络结构。在此基础上，添加了预分类模块，用于获取视图的一致信息和互补信息。随后，借助注意力机制增强了这两类信息传递到后续网络，从而更精准地融合了两类视图。实验证明，本文方法在ModelNet40数据集上的分类效果优于其他方法。

1 本文方法

基于视图的3D模型分类算法基本模型结构如图1所示，主要由3个部分组成，包括视图描述符提取层、生成形状描述符的视图池层和三维形状分类层。图中FCN（first convolutional network）表示第1层卷积神经网络，FC（full connection）表示全连接层，$m@224 \times 224 \times 3$表示输入图像的尺寸，$m$代表输入图像。视图描述符提取层采用在大规模数据上预训练的卷积神经网络作为特征提取网络，基本结构是卷积层的深度叠加对输入图像进行特征提取，在最终分类层之前利用视图池操作对多个视图进行选择或融合，得到最终的形状描述符用以得到分类结果。卷积神经网络在图像处理任务中具有明显优势，并且不断地产生更多改进模型。

对于三维模型分类任务，现有的多视图模型存在一些不足。首先，多视图数据本身存在其特有的属性，利用预训练的网络对输入图像进行处理，没有充分考虑视图间存在的一致性和互补性，导致3D模

型的某些特征信息丢失，最终降低模型的性能。其次，现有的多视图模型选择将输入视图完全提取后，再进行视图池操作，此操作可以保证使一个输入视图的特征得到深度提取，但是针对其余输入视图的特征提取使模型计算量增大，并且缺少对多视图间信息的利用。另外，虽然视图池在多视图数据的利用方面，可以选择或融合最具优势的视图，但是视图池也仅仅采用最大池化或分组选择的方式选择形状级描述符，同样不能避免资源的浪费和视图信息的丢失。最后，部分形状级描述符的生成方式采用直接相加的策略，不可避免地在多个视图充分提取后，视图与视图间会产生噪声数据相互干扰，可能导致分类任务的精度不理想。

针对以上问题，本文在传统ResNet模型的残差模块中添加空洞卷积用以增强模型对视图信息的感知，添加用于互补性和一致性分类的预分类模块和用于融合视图特征的可迭代权重融合模块。整体模型在现有的3D模型分类算法的基础上调整了结构，在视图特征学习过程中进行视图属性的分类，再通过注意力机制网络突出视图属性，最后融合得到形状级描述符用以分类。具体结构如图2所示，图中标注表明了不同层次所对应内容，预分类模块（pre-classified）是用于区分一致性视图和互补性视图的可移动模块。

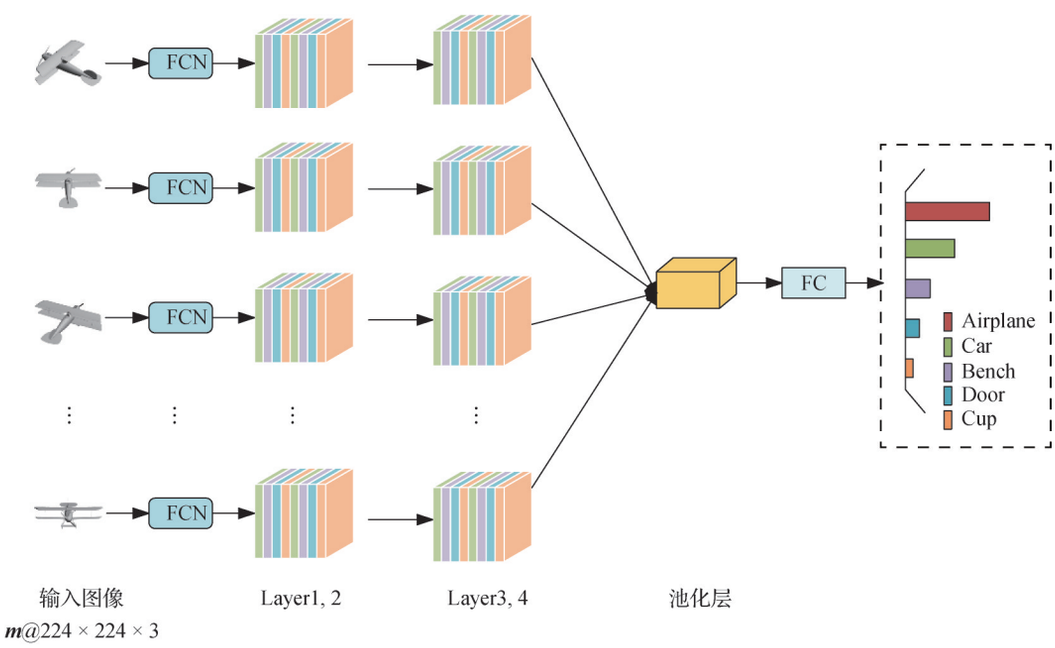


图1　基于视图的3D模型分类网络

Fig. 1　View based 3D model classification network

1.1　改进残差结构

与普通图像数据集相比，3D模型数据集具有典型的一致性和互补性，其数据集通常由一类三维物体的一组（12个）视图组成。使用VGG、GoogLeNet和ResNet等经典模型作为骨干网络在训练中容易过拟合，泛化效果不佳。常见的解决过拟合的方法主要包括正则化和调整网络结构等。为了解决过拟合并提高模型对3D模型数据的学习能力，本文对经典的残差网络ResNet进行修改，在其残差结构中添加多尺寸空洞卷积，并根据训练过程调整网络的结构和深度。此外，空洞卷积虽然能够提升卷积操作的感受野，但是单一使用空洞卷积并不能有效保证特征信息的传递，并且多次使用空洞卷积还会导致信息丢失，因此本文将空洞卷积和普通卷积并联使用，能够有效提高特征学习能力。

改进的残差部分如图3所示，图3（a）表示原残差网络中的残差结构，图3（b）由空洞率为d的空洞卷积与普通卷积并联替换原有的普通卷积，构成修改后的残差结构。此外，空洞卷积在网络叠加深度时，保持同样的空洞率会降低感受野，因此，依据特征视图的尺寸设置空洞率，在初始层设置空洞率为3，之后随着叠加深度逐渐减小为1。不同的空洞率设置可以在提取特征时，确保输入图像的每一层在提升感受野的同时不会造成信息的丢失，以此进行

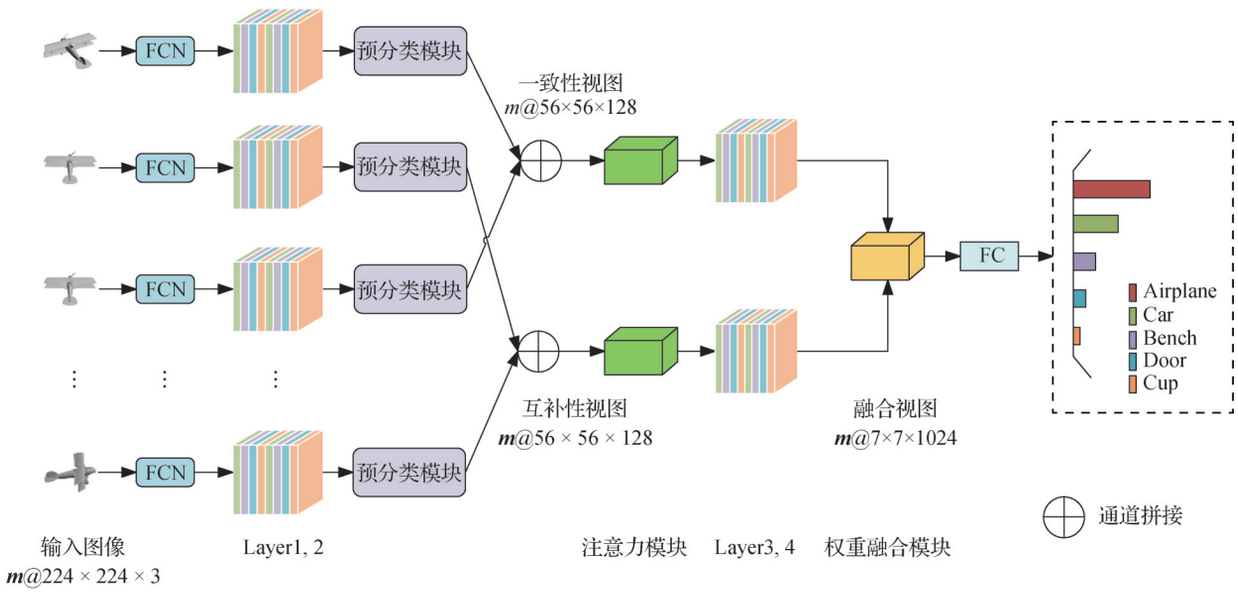


图2　融合多视图一致和互补信息的3D模型分类网络

Fig. 2　Multi-view consistent and complementary information fusion network for 3D model classification

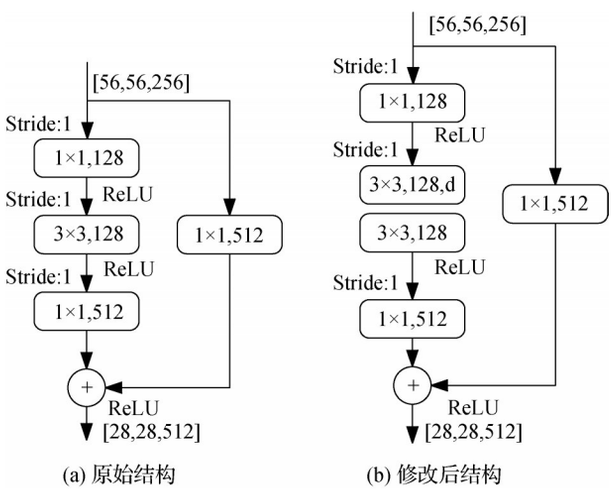


图3　残差结构对比

Fig. 3　Residual structure comparison

((a) original structure; (b) improved structure)

多尺度空洞卷积操作并提高模型的泛化能力和准确性。

1.2　视图分类模块

在基于视图的三维形状识别任务中，由于视图间存在一致性和互补性信息，因此如何对一致性信息进行准确提取，并且保证互补性信息有效利用是本文主要关注的问题。原始方法使用收集的所有视图，然而，由于处理所有视图需要高度的计算，并且并非每个视图都有助于识别，因此一些方法引入了代表性视图选择机制，或者下一个视图选择机制。本文设计一种分类机制将视图划分为两类，区别于其他方法，既避免了完全处理所有视图，也将视图的所有特征全部凸显，具体的分类划分方法如下：

首先设计一种预分类网络，判断通过特征提取网络得到视图描述符所具有的视图属性。预分类模块利用当前视图描述符的浅层表示，计算得到一组代表3D模型的分数，根据分数对区分视图具有的不同属性，最终获得两组代表性视图。本文采用通道拼接方法，将两组代表性视图在通道上依次拼接得到一致性视图和互补性视图。

定义函数$\xi(\cdot)$来量化预分类网络的判别。首先，给定一组视图$\boldsymbol{S}=\{\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_m\}$，该单元的输出为$\boldsymbol{O}=\{\boldsymbol{O}_1,\boldsymbol{O}_2,\cdots,\boldsymbol{O}_m\}$，其中$\boldsymbol{S}_m$表示向特征提取网络输入的单一视图，$\boldsymbol{O}_m$表示从特征提取网络输出的单一视图，$m$表示输入输出视图的个数，由此可得

$$x(\boldsymbol{S}_i)=f_{\text{sigmoid}}\left(\log\left(L_1(\boldsymbol{O}_i)\right)\right) \tag{1}$$

式中，L_1表示预分类网络，log函数确保输入到sigmoid函数的值具有离散性。当输入值小于−5或大于5时，其输出将趋近于0与1。因此，为了方便计算，将输出结果计算对数值，再计算sigmoid值，保证每个视图的判别分数都在0～1之间。在得到每个视图的判别分数后，定义判别分数大于0.5的视图具有一致属性、小于0.5的视图具有互补属性，将这两类属性视图分别融合成一致性特征视图和互补性特征视图。这种视图分类选择方案的优点是，可以动态地选择每个视图对最终描述符的影响，同时

可以最大化考虑视图间的一致信息，即用于代表三维形状的信息，也利用了单个视图可能表示三维形状的特异性信息。预分类模块作为一个插件可以在特征提取网络的任意残差层后使用，考虑到本文的分类策略紧跟预分类方法，尝试将残差网络输入通道数由[64，128，256，512]提升到[96，192，384，768]，保证数据在残差网络中的传递达到预计的学习效果。

这些视图描述符已经学习到对应视图具有的互补信息，同时使用预分类网络也利用了视图具有的浅层一致信息。但是这些一致信息所代表的目标类别信息并没有在网络中进行深度特征学习。为了更好地获得多视图间的一致信息，将两个代表性视图输入到注意力模块和深层特征提取网络加深对特征的表征能力。

1.3 学习融合模块

在基于视图的三维形状识别任务中，如何获得融合视图用于最终形状表达是决定模型最终性能的关键一步。现有的部分方法采用视图池机制选择一个具有最大特征信息的视图，也有方法采用强化学习的思想，将单个视图输入网络学习后，再依次输入其余视图进行学习。本文采用一种应用于残差结构跳跃连接中的融合机制，通过设计两个分支网络联合学习输入特征的权重，通过权重进行融合获得最终的形状描述符。

利用预分类网络对多个视图进行分类选择得到两类特征视图后，在默认情况下，假设一致性特征视图$\boldsymbol{X} \in \mathbf{R}^{C\times H\times W}$是具有更大接受域的特征映射，互补性特征视图$\boldsymbol{Y} \in \mathbf{R}^{C\times H\times W}$作为另一个特征映射，最终得到两个特征映射集合$\boldsymbol{X}$和$\boldsymbol{Y}$。融合学习模块包含两个分支，深化对特征视图的学习并学习两类特征视图的权重，最终根据学习的权重融合两类视图。两个分支的信息感知网络由残差网络的第3、4层先进行深度特征提取，同时将特征映射$\boldsymbol{X}$和$\boldsymbol{Y}$输入权重学习模块学习权重，最终利用式(2)融合两个特征映射，具体为

$$\boldsymbol{y} = \boldsymbol{X}(L_2(\boldsymbol{X}+\boldsymbol{Y}) + \boldsymbol{Y}(1 - L_2(\boldsymbol{X}+\boldsymbol{Y})) \quad (2)$$

式中，$\boldsymbol{y}$表示融合视图，即形状级描述符，$\boldsymbol{X}$和$\boldsymbol{Y}$分别表示输入的两类特征图像集合，L_2表示可迭代学习的权重网络，主要由全局信息感知网络、局部信息感知网络以及归一化函数构成。值得注意的是，在注意力融合模型中，由于输入图像融合权重由0到1之间的实数组成，这使得该网络能够在$\boldsymbol{X}$和$\boldsymbol{Y}$之间做选择和加权平均。此外，考虑到初始的融合质量对最终的融合权重产生的影响，在本文新残差结构的第4层之前设置分支网络，将两类视图进行最后一次特征提取，然后分别计算相应的权重进行融合得到形状级描述符。最后，在网络的最后一层设计一个分类层，得到最终的分类结果。

除此之外，学习融合模块考虑将两类视图先深度特征提取后再学习权重，可能造成权重不稳定。考虑将特征视图进行初始融合后，再进行特征提取和权重学习，然而初始的融合质量对最终的融合权重产生深远的影响。由于这依然是一个特征融合问题，一种直观的方法是使用一个注意力模块来融合特征。将这种两阶段的方法称为可学习权重融合，则初始积分$\boldsymbol{X}+\boldsymbol{Y}$可以重新表述为

$$\boldsymbol{X}+\boldsymbol{Y} = \boldsymbol{X}(L_2(\boldsymbol{X}+\boldsymbol{Y}) + \boldsymbol{Y}(1 - L_2(\boldsymbol{X}+\boldsymbol{Y})) \quad (3)$$

最终依据可学习权重融合策略，将最优视图映射与其余视图映射输入到可迭代权重融合网络，让一致和互补信息感知网络学习融合权重，获得最终的形状描述符。在网络的最后一层设计一个分类层，得到最终的分类结果。

2 实 验

2.1 数据集和实验环境

实验数据集采用Princeton ModelNet项目(Wu等，2015)提供的数据集，该项目为计算机视觉、计算机图形学、机器人和认知科学的研究人员提供的全面、干净的物体三维CAD(computer-aided design)模型集合，涉及622个对象类别的127 915个CAD模型。本文采用该项目具有代表性的两个子数据集，分别是ModelNet10和ModelNet40，ModelNet10由4 899个3D模型组成，共有10个类别；ModelNet40由来自40个流行类别的12 311个对象组成。这些数据集被随机划分为训练集，测试集和验证集，其中ModelNet10的3 991个对象作为训练样本，908个作为测试样本，从训练样本中选择20%作为验证样本；同样地，在ModelNet40中选择9 843个对象用于训练，训练样本的20%作为验证样本，2 468个对象用于测试，以确保样本在各个集合中均匀分布。

所有数据经过一致的预处理流程，如图4所示，

将单个三维CAD模型在空间坐标系摆正，沿着竖直方向与水平面的60°夹角线设置12个相机，全面覆盖三维模型的四周，依次投影获得12个二维图像。

图4　投影方式与样本示例

Fig. 4　Projection method and sample examples

在一台GPU为NVIDIA RTX A5000的设备上使用PyTorch进行相关实验。由于GPU显存限制，将批处理大小设置为16，并应用Adam优化算法，在100个epoch中初始学习率设置为0.000 01，权重衰减设置为0.000 01。本文中，考虑到调整了网络结构，制定一种单分类任务实验，将批处理大小设置为64，在100个epoch中初始学习率设置为0.01，其余与上述设置一致。

2.2　实验设置

对于通用三维形状分类任务，采用本文设计的网络模型验证模型有效性，并根据使用到的每一个数据集进行预训练。将三维形状识别任务作为单分类任务验证不同特征提取网络时，不添加预分类模块进行实验，并对比本文的骨干网络在低通道数版本和高通道数版本的性能。在进行三维识别任务时，本文将预分类模块插入特征提取网络的第2层和第3层残差结构之间，并添加一个注意力模块与之结合使用，其余与MVCNN-new类似，将特征融合层放在互补信息提取层的倒数第2层之后。所有的对比方法在实验设置以及网络参数明确的情况下，都使用本文的实验设备进行了新的实验获得结果，保证基础条件的一致性。

此外，考虑到预分类模块在网络中插入的位置也会影响最终的结果，本文对比了预分类模块在不同层之间插入的性能。所有的模型都使用Adam优化器进行训练。在实验中，视图的数量总是设置为12，特别的任务会根据实际情况指定。学习率设置分为两种情况，在互补特征提取层进行预训练中设置为0.01，在进行3D分类任务中设置为1E−5，所有的过程中全部使用交叉熵损失。对于仿射变化中矩阵的大小，本文采用与ResNet同样大小为12 × 1的变化矩阵。

2.3　评价指标

为了评价三维形状分类的准确度（accuracy，ACC）性能，本文采用平均实例准确度和平均分类准确度作为衡量指标。平均实例准确度是正确分类的对象数量与对象总数的比率；而分类准确度是所有类别中实例准确度的平均值。分类准确度是客观的，因为不同类别的测试对象数量不均匀，而实例准确度是直观的，因为它反映了错误分类的对象数量。在消融实验中，进行10次重复实验，将实验的每10轮的平均分类准确度取平均值进行对比。

平均实例准确度和平均分类准确度的计算分别为

$$f_{\text{instance accuracy}}=\frac{\sum_{i=1}^{C}TP_i+TN_i}{\sum_{i=1}^{C}P_i+N_i} \tag{4}$$

$$f_{\text{classification accuracy}}=\frac{1}{C}\frac{\sum_{i=1}^{C}TP_i+TN_i}{\sum_{i=1}^{C}P_i+N_i} \tag{5}$$

式中，TP_i，TN_i，P_i，N_i分别是对应于第i个类别的真阳性、真阴性、阳性和阴性结果的样本数量，C是样本中类别的总数。

3　实验结果分析

3.1　与其他先进方法的对比

针对三维形状分类任务，将本文提出的多视图信息融合的深度3D模型分类方法与各种方法进行对比，以验证本文模型的有效性。当使用ModelNet10或ModeNet40数据集时，选择3DShapeNets（Wu等，2015）、PointNet（Charles等，2017）、MVCNN（Su等，2015）、MHBN（Yu等，2018）、RotationNet（Kanezaki等，2021）、SeqViews2SeqLabels（Han等，2019）、GVCNN（Feng等，2018）、MLVCNN（multi-loop-view CNN）（Jiang等，2019）、MVCNN-new（Su等，2018）、DRCNN（Sun等，2021）、BayAda（bayesian ada-Boost）（Gao等，2021）、View-GCN（Wei等，2020）、MVSAN（multi-view softpool attention network）（Wang等，2022）和LFT-Net（local feature transformer network）（Gao等，2023）进行比较。由于不同方法采用3D模型的不同数据表现形式，表中列举了每个方法使用的数据输入，Voxel表示体素数据、Points表示点

云数据、View表示多视图数据、Dep表示深度图数据，View+Dep、Points+View和View+Voxel表示综合两种模态数据。

实验结果如表1所示。可以看出，所提出的多视图信息融合的深度3D模型分类模型显著优于基于点云和基于体素的深度学习模型。在基于视图的模型中，本文模型也取得了优异性能。具体地说，同样使用12个视图的情况下，本文模型使用优化的特征提取网络在ModelNet10中获得了97.27%的平均实例准确率，弱于DRCNN方法。造成这种问题的原因是ModelNet10数据集规模较小，通过空洞卷积增强感受野学习的同时产生了过拟合问题，导致分类精度下降。在ModelNet40中，本文方法的平均分类准确度和平均实例准确度分别达到了95.53%和96.92%。由于ResNet和VGG在多种三维模型分类方法中用于特征提取网络，因此这两个方法可以作为特征提取网络的基准方法。相比于使用ResNet和VGG进行特征提取，本文在残差结构中添加空洞卷积并调整神经元数量，在两种性能指标上同样有一定的优势。这表明了本文模型在骨干网络上比其他方法能够更有效地提取视图间的信息。需要注意的是，由于部分方法在其原文中使用的视图数比本文要多，如MLVCNN的视图数为8×3，View-GCN的视图数为20，因此这些方法与原文中的实验结果并不相同。此外，当前使用多模态数据的方法逐渐流行，本文方法与融合视图和体素的BayAda方法以及融合视图和点云的LFT-Net方法相比，同样取得领先优势。

表1　在ModelNet40和ModelNet10数据集上不同方法的性能对比

Table 1　Performance comparison of different methods on the ModelNet40 and ModelNet10 datasets

/%

方法	输入	视图数	特征提取网络	ModelNet10		ModelNet40	
				平均分类准确度	平均实例准确度	平均分类准确度	平均实例准确度
3D ShapeNets	Voxel	–	–	83.54	–	77.32	–
PointNet	Points	–	–	–	–	86.2	89.2
MVCNN	View	80	VGG-M	–	–	90.1	–
MHBN	View+Dep	6	VGG-M	95	95	93.1	94.7
RotationNet	View	12	VGG-M	–	–	–	94.68
SeqViews2SeqLabels	View	12	VGG-19	94.8	94.82	91.12	93.31
GVCNN	View	12	GoogleNet	–	–	–	92.6
MLVCNN	View	12	ResNet	–	–	–	93.07
MVCNN-new	View	12	ResNet	–	–	94	95.5
DRCNN	View	12	ResNet	**99.3**	**99.34**	94.86	95.84
View-GCN	View	12	ResNet	–	–	94.46	95.29
LFT-Net	Points+View	–	–	–	–	89.7	93.2
BayAda	Voxel+View	–	–	–	90.2	–	–
MVSAN	Views	12	–	98.37	98.57	95.31	96.8
本文	View	12	ResNet	95.3	95.72	93.92	95.14
本文	View	12	–	96.03	97.27	**95.53**	**96.92**

注：加粗字体表示各列最优结果。“–”表示该数据在相关文献中缺失。

3.2　消融实验

3.2.1　预分类模块位置对比

预分类模块是本模型中将全部视图分类的重要手段，因此其插入在残差网络的位置对最终的三维识别任务十分重要。考虑到残差网络很多层之间可以插入模块，并且浅层的卷积网络并没有很好地提

取特征，因此从第1层残差结构之后开始插入验证。预分类模块的设计全部采用同样的标准，通过一个卷积层将通道数调整为1，再使用全连接层将通道数调整为36计算输出，分类后输入后续模块。

如图5所示，将预分类模块插入第2层或第3层之后，最终任务精度比其余层要好。插入第1层后，可能由于一个残差结构并没有充分提取互补特征，导致在预分类后得到的两个特征视图在融合时相互干扰。插入第4层之后，由于特征得到充分提取，导致出现过拟合。因此，选择将预分类模块应用于在第2层之后，并添加一个注意力模块，进一步提高特征提取能力。

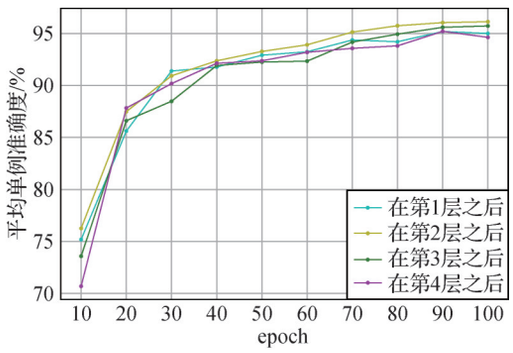


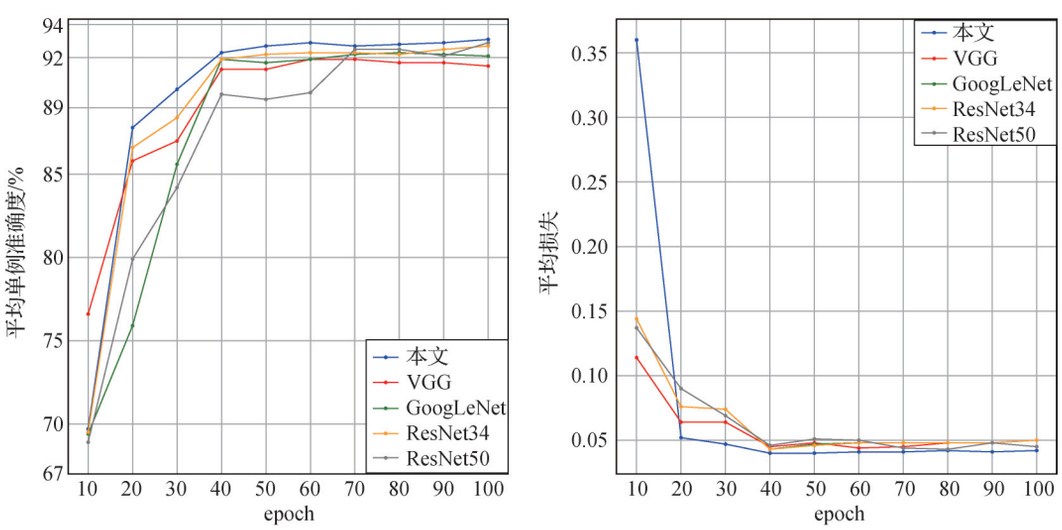
图5　预分类模块在不同位置的实验结果

Fig. 5　Experimental results of the pre classification module at different positions

3.2.2　特征提取网络

在特征提取方面，本文将改进的残差网络用于特征提取，该网络是针对多视图任务修改的，并没有在大规模数据集ImageNet（Deng等，2009）上进行预训练。为了突出学习视图间互补信息的优势，本文与现有方法中的骨干网络进行对比，如VGG、ResNet和GoogLeNet等。这些骨干网络在一些方法中只用做特征提取，并没有形成最终的分类结果，而3D模型转化成多视图后，单个视图对应一个三维模型即可看做一个具体的任务。因此，将本文的特征提取网络和各个骨干网络添加一个相同的分类层，构成一个完整的卷积神经网络，在ModelNet40数据集上进行单分类任务，与现有的骨干网络进行对比，验证所提模型的有效性。

如图6所示，本文方法在同样的训练次数下，平均单类分类准确度比ResNet50等方法具有更高的分类精度。同时，在损失下降方面，本文方法具有更低的平均损失，并且损失的收敛性较强，而其他方法在损失达到最低点后出现了一定的过拟合现象。



（a）分类准确度　　（b）平均损失

图6　特征提取网络的实验结果

Fig. 6　Comparison experiment results of feature extraction network ((a) classification accuracy; (b) average loss)

此外，针对特征提取网络的单分类任务并不能完全说明本文方法的优越性，因此，进一步将不同特征提取网络作为基准方法MVCNN的骨干网络进行3D模型分类实验。具体而言，将GoogLeNet、ResNet34、ResNet50和本文方法分别替换掉MVCNN中原有的VGG网络。同时，考虑到本文方法在利用预分类与注意力融合策略获得形状级描述符，而使用以上骨干网络提取特征得到形状级描述符方式各

不同，因此尝试按照本文的思想将不同骨干网络的MVCNN获取形状级描述符的方式统一，进行实验并对比结果。

在ModelNet40数据集得到平均分类准确度结果如表2所示，表中MVMPL（multi view max pooling lyear）表示利用MVCNN中最大池化生成形状级描述符的思想，VPCAF（view pre classification and attention fusion）表示利用本文视图分类与注意力融合的思想。从实验结果可以看出，用做MVCNN的骨干网络时，本文特征提取网络方法具有最好的性能表现，比基于VGG的基准MVCNN方法提高了4.3%；在结合本文提出的形状级描述符获取方式中，本文修改的特征提取网络在分类准确度上比ResNet50方法获得了1.07%的领先。另外，本文的视图分类和注意力融合思想与各种特征提取网络结合用于分类的结果都优于使用MVCNN中使用最大池化层的方法，从侧面验证了视图分类与注意力融合模块的有效性。

表2　不同网络在ModelNet40数据集的分类准确度对比

Table 2　Comparison of classification accuracy in different networks on ModelNet40 dataset

/%

方法	ACC^{MVMPL}	ACC^{VPCAF}
VGG-11	90.10	91.73
GoogLeNet	91.82	92.62
ResNet34	92.86	93.58
ResNet50	93.07	94.25
本文	**94.40**	**95.32**

注：加粗字体表示各列最优结果。

3.3　视图数量的影响

实验评估了视图数量对本文方法在ModelNet40数据集上的平均实例精度。本文选择了MVCNN和DRCNN的结果，如表3所示。表中展示了MVCNN和DRCNN在不同视图数量上的准确度，其中MVCNN的结果是使用ResNet作为骨干网络重新实验获得的。

从表3可以看出，无论视图数量如何，本文方法均优于MVCNN，略优于DRCNN。此外，在MVCNN中，视图数量从3增加到6时，精度变得更高，视图数量由6增加到12时，精度略有下降。视图数量从3到6到12时，本文方法和DRCNN模型的精度持续增长。不同的是，本文模型在处理高视图数时，采用了分类学习再融合的方式，却取得了与DRCNN相近的结果，可能是因为本文方法可以比MVCNN和DRCNN利用更多的多视图信息，这进一步体现了本文模型的优越性。

表3　不同视图数的平均分类准确度对比

Table 3　Classification accuracy comparison with different views

/%

方法	视图数 = 3	视图数 = 6	视图数 = 12
MVCNN	91.33	92.01	91.53
DRCNN	93.03	93.76	94.86
本文	**93.27**	**94.31**	**95.13**

注：加粗字体表示各列最优结果。

3.4　时间和空间复杂度

表4对比了本文方法与两种经典3D模型分类方法的时间和空间复杂度。通过记录单个输入样本的浮点运算数量（giga floating-point operations per second，GFLOPs）和前向传播时间，本文方法对比MVCNN和ViewGCN在浮点运算数上并没有优势，这是在特征提取网络中添加空洞卷积和调整每层神经网络的神经元数量导致的。这同样导致了在其余两个指标上，与现有方法相比并没有太多优势，但是低成本的平台同样可以进行重复实验。

表4　时间和空间要求

Table 4　Time and memory requirements

方法	GFLOPs	时间/ms	参数量/M
MVCNN	43.72	39.89	**11.20**
ViewGCN	44.19	**26.06**	23.56
本文	**41.23**	29.68	15.79

注：加粗字体表示各列最优结果。

4　结　论

基于视图的3D模型分类一方面要考虑视图的获取方式，另一方面需要考虑视图与视图之间存在的关联属性。本文提出一种基于多视图信息融合的深度3D模型识别网络。首先，通过在残差结构中嵌入空洞卷积，增大卷积的感受野，增强网络获取特征

信息的能力;其次,通过重新设计特征提取网络的结构,解决传统网络对视图一致信息和互补信息的获取问题,并嵌入空间和通道注意力模块,以放大并保存目标特征信息;此外,设计预分类网络根据视图信息获取一致性和互补性分类视图,并采用可迭代的融合模块,融合两种性质的视图得到最终的形状级描述符,以充分利用视图间存在的关联属性。最后在ModelNet40和ModelNet10数据集上进行训练和测试,并与14种现有的经典方法进行对比,这些方法基于各类深度学习模型或基于数学结构特征,验证了模型的有效性。实验结果表明,所提网络模型明显优于对比的经典方法,分类精度高,能够在多种领域中应用。虽然该模型旨在解决如何利用多视图之间的关联属性问题,即通过预定策略划分一致性视图和互补性视图,但随着深度学习的发展,在未来可以扩展到采用强化学习的思想进行多视图分类研究。此外,作为一种信息深度结合方法,本文采用的视图融合方法对其他需要融合的任务具有参考意义。

参考文献(References)

Brock A, Lim T, Ritchie J M and Weston N. 2016. Generative and discriminative voxel modeling with convolutional neural networks [EB/OL]. [2024-02-06]. https://arxiv.org/pdf/1608.04236.pdf

Charles R Q, Su H, Kaichun M and Guibas L J. 2017. PointNet: deep learning on point sets for 3d classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Chen S, Yu T and Li P. 2021. MVT: multi-view vision transformer for 3D object recognition [EB/OL]. [2024-02-06]. https://arxiv.org/pdf/2110.13083.pdf

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]

Feng Y F, Zhang Z Z, Zhao X B, Ji R R and Gao Y. 2018. GVCNN: group-view convolutional neural networks for 3D shape recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 264-272 [DOI: 10.1109/CVPR.2018.00035]

Gao X Y, Li K P, Zhang C X and Yu B. 2021. 3D model classification based on Bayesian classifier with AdaBoost. Discrete Dynamics in Nature and Society, 2021: #2154762 [DOI: 10.1155/2021/2154762]

Gao Y B, Liu X B, Li J, Fang Z J, Jiang X Y and Huq K M S. 2023. LFT-Net: local feature transformer network for point clouds analysis. IEEE Transactions on Intelligent Transportation Systems, 24(2): 2158-2168 [DOI: 10.1109/TITS.2022.3140355]

Han Z Z, Shang M Y, Liu Z B, Vong C M, Liu Y S, Zwicker M, Han J W and Chen C L P. 2019. SeqViews2SeqLabels: learning 3D global features via aggregating sequential views by RNN with attention. IEEE Transactions on Image Processing, 28(2): 658-672 [DOI: 10.1109/TIP.2018.2868426]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Jiang J W, Bao D, Chen Z Q, Zhao X B and Gao Y. 2019. MLVCNN: multi-loop-view convolutional neural network for 3D shape retrieval//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 8513-8520 [DOI: 10.1609/AAAI.V33I01.33018513]

Kanezaki A, Matsushita Y and Nishida Y. 2021. RotationNet for joint object categorization and unsupervised pose estimation from multi-view images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1): 269-283 [DOI: 10.1109/TPAMI.2019.2922640]

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Li J, Liu Z, Li L, Lin J Q, Yao J and Tu J M. 2023. Multi-view convolutional vision transformer for 3D object recognition. Journal of Visual Communication and Image Representation, 95: #103906 [DOI: 10.1016/J.JVCIR.2023.103906]

Liu Z H, Zhang Y H, Gao J and Wang S R. 2022. VFMVAC: view-filtering-based multi-view aggregating convolution for 3D shape recognition and retrieval. Pattern Recognition, 129: #108774 [DOI: 10.1016/J.PATCOG.2022.108774]

Malik J, Shimada S, Elhayek A, Ali S A, Theobalt C, Golyanik V and Stricker D. 2022. HandVoxNet++: 3D hand shape and pose estimation using voxel-based neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 8962-8974 [DOI: 10.1109/TPAMI.2021.3122874]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2024-02-06]. https://arxiv.org/pdf/1409.1556.pdf

Song D, Fu X W, Liu N, Nie W Z, Li W H, Wang L J, Yang Y and Liu A A. 2024. MV-CLIP: multi-view CLIP for Zero-shot 3D shape recognition [EB/OL]. [2024-02-06]. https://arxiv.org/pdf/2311.18402.pdf

Su H, Maji S, Kalogerakis E and Learned-Miller E. 2015. Multi-view

convolutional neural networks for 3D shape recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 945-953 [DOI: 10.1109/ICCV.2015.114]

Su J C, Gadelha M, Wang R and Maji S. 2018. A deeper look at 3D shape classifiers [EB/OL]. [2024-02-06]. https://arxiv.org/pdf/1809.02560.pdf

Sun K, Zhang J S, Liu J M, Yu R X and Song Z J. 2021. DRCNN: dynamic routing convolutional neural network for multi-view 3D object recognition. IEEE Transactions on Image Processing, 30: 868-877 [DOI: 10.1109/TIP.2020.3039378]

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Wang W J, Wang X L, Chen G and Zhou H R. 2022. Multi-view SoftPool attention convolutional networks for 3D model classification. Frontiers in Neurorobotics, 16: #1029968 [DOI: 10.3389/FNBOT.2022.1029968]

Wei X, Yu R X and Sun J. 2020. View-GCN: view-based graph convolutional network for 3D shape analysis//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1847-1856 [DOI: 10.1109/CVPR42600.2020.00192]

Wei X, Yu R X and Sun J. 2023. Learning view-based graph convolutional network for multi-view 3D shape analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6): 7525-7541 [DOI: 10.1109/TPAMI.2022.3221785]

Wu B, Liu Y A and Zhao J. 2024. Classification network for 3D point cloud based on spatial structure convolution and attention mechanism. Journal of Image and Graphics, 29(2): 520-532 (武斌, 刘溢安, 赵洁. 2024. 结合空间结构卷积和注意力机制的三维点云分类网络. 中国图象图形学报, 29(2): 520-532) [DOI: 10.11834/jig.230137]

Wu Z R, Song S R, Khosla A, Yu F, Zhang L G, Tang X O and Xiao J X. 2015. 3D ShapeNets: a deep representation for volumetric shapes//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1912-1920 [DOI: 10.1109/CVPR.2015.7298801]

Xu Y, Zheng C D, Xu R T, Quan Y H and Ling H B. 2021. Multi-view 3D shape recognition via correspondence-aware deep learning. IEEE Transactions on Image Processing, 30: 5299-5312 [DOI: 10.1109/TIP.2021.3082310]

Yu T, Meng J J and Yuan J S. 2018. Multi-view harmonized bilinear network for 3D object recognition//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 186-194 [DOI: 10.1109/CVPR.2018.00027]

作者简介

吴晗，男，硕士研究生，主要研究方向为深度学习和多视图学习。E-mail：wu2ha1n@163.com

胡良臣，通信作者，男，讲师，硕士生导师，主要研究方向为模式识别与机器学习、图像处理与计算机视觉。E-mail：cslc.hu@ahnu.edu.cn

杨影，女，本科生，主要研究方向为模式识别与机器学习、图像处理与计算机视觉。E-mail：yangying_16@163.com

接标，男，教授，博士生导师，主要研究方向为机器学习、数据挖掘和图像分析。E-mail：jbiao@ahnu.edu.cn

罗永龙，男，教授，博士生导师，主要研究方向为数据安全与隐私保护、空间数据处理。E-mail：ylluo@ustc.edu.cn