姿态非对齐的三维模型分类

丁 博,高 源,范宇飞,何勇军
(哈尔滨理工大学计算机科学与技术学院,黑龙江哈尔滨 150080)

摘 要: 目前的三维模型分类方法均是对初始姿态已经对齐的数据集进行分类,但是在实际应用中,三维模型的姿态是未知的,非对齐的三维模型将导致分类准确率急剧下降. 本文提出了一种新的三维模型分类方法,适用于模型姿态对齐和非对齐两种情况. 该方法采用图卷积神经网络(Graph Convolutional neural Network,GCN)学习视图间的空间关系,将预先设置好的相机位置作为图结构中的顶点,并通过时序特征提取网络以及注意力网络进一步提升GCN的运算效果,从而完成三维模型的分类. 实验表明,该方法在ModelNet10和ModelNet40数据集上进行实验,在三维模型姿态对齐的情况下,分类准确率分别高达99.3%和97.4%,远高于现有方法. 在三维模型姿态非对齐的情况下,也有较高的分类准确率.

关键词: 三维模型分类;三维模型姿态;图卷积神经网络;注意力机制
基金项目: 国家自然科学基金面上项目(No.61673142)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2023)09-2379-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211366

3D Model Classification for Non-Aligned Poses

DING Bo, GAO Yuan, FAN Yu-fei, HE Yong-jun
(*Computer Science and Technology College*, *Harbin University of Science and Technology*, *Harbin*, *Heilongjiang* 150080, *China*)

Abstract: Current 3D model classification methods are validated on the datasets whose initial poses are aligned. However, in practical applications, the poses of 3D models are unknown, resulting in obvious performance degradation of a non-aligned 3D models. A new 3D model classification method which is suitable for both the aligned and non-aligned poses of datasets, is proposed in this paper. This method employs graph convolutional neural network (GCN) to learn the spatial relations between views, and uses the preset camera positions as the vertexes in the graph structure. Moreover, the timing feature extraction network and the attention network are used to further improve the effect of GCN. Experiments on ModelNet10 and ModelNet40 datasets show that the proposed method achieves accuracies of 99.3% and 97.4% under aligned poses of 3D models, which is much higher than other existing methods. On non-aligned poses of 3D models, also has high classification accuracy.

Key words: 3D model classification; 3D model pose; graph convolutional neural network; attention mechanism
Foundation Item(s): National Natural Science Foundation of China (No.61673142)

1 引言

在数字媒体时代背景下,三维模型成为继文本、声音、图像和视频的一种信息的重要载体. 目前,三维模型被广泛的应用于数字娱乐、计算机辅助设计(Computer Aided Design, CAD)、医学诊断和生物信息学等领域[1~3]. 随着三维模型需求的不断增长,三维模型数量激增. 面对数量巨大、种类繁多的三维模型,有效的组织和管理三维模型,探索更有效的三维模型分类和检索方法成为亟待解决的问题.

三维模型根据表示方法不同可以分为三类,分别是基于体素、基于点云和基于视图的方法[4]. 基于体素的三维模型表示方法将三维模型量化后,以像素的形式表示. 将三维模型体素化表示后即可利用深度神经网络进行训练学习. 基于点云的三维模型表示方法将三维数据看成是点的集合,具体来说,将每个点看作一个神经元节点,节点包含点的坐标等信息,然后利用深度神经网络提取点的特征. 基于视图的三维模型表示方法将三维模型通过投影为二维视图集合,再利用深度神经网络提取这些二维视图的特征,然后可以利用

收稿日期:2021-10-09;修回日期:2022-04-23;责任编辑:梅志强

现有的图像处理方法对三维模型进行分类[5]. 同时,使用二维视图进行深度神经网络训练还有另一个优势,即可以使用二维视图描述符的优势以及海量图像数据库(如ImageNet)来对深度神经网络进行预训练,加快其收敛速度.

与传统机器学习方法相比,深度学习让机器自动学习被表征对象的特征描述符,被广泛的应用于计算机视觉领域[6]. 基于视图的三维模型分类方法充分利用了深度学习的优势,取得了不错的分类效果,但仍存在如下问题:(1)三维模型投影为二维视图序列时,相机拍摄具有连续性,但是对三维模型分类时,视图间的这种时序信息经常被忽略;(2)将三维模型表征为二维视图,虽然有利于深度神经网络的训练,但是破坏了三维模型原有的空间结构,丢失了视图间的空间信息;(3)不同视图表示的是三维模型不同视角下的信息,包含的信息量不同,目前的方法忽略了视图间的差异;(4)目前的三维模型分类方法均是针对三维模型姿态对齐的情况,无法在非对齐的三维模型分类上取得良好的效果. 针对以上的问题,本文提出一种既适用于对齐,也适用于非对齐的三维模型分类方法. 本文贡献总结如下:

(1)在特征提取方面,根据预先设置好的空间投影位置将三维模型表征为二维视图序列,采用卷积神经网络(Convolution Neural Network, CNN)和循环神经网络(Recurrent Neural Network,RNN),提取三维模型的静态视觉特征和动态时序特征.

(2)在识别建模方面,考虑到图卷积神经网络在学习空间位置关系中所具有的优势,采用图卷积神经网络获取视图间的空间信息.

(3)引入注意力机制,根据视图的重要性不同对视图加权学习,对三维模型分类更具辨识力的视图赋予更高的权重,挖掘视图间的区分性信息.

(4)本文在三维模型姿态对齐和非对齐的情况下,在公共数据集ModelNet10和ModelNet40上,验证了本文算法的有效性.

2 相关工作

2.1 基于视图的三维模型分类

基于视图的三维模型分类方法分为传统方法和基于代表性视图的方法. 张静等[7]利用卷积神经网络采用主成分分析(Principle Component Analysis, PCA)对三维模型进行预处理后,把三维模型投影成6幅二维图像,然后使用卷积神经网络提取图像特征,得到最后的三维模型描述符. Chen等[8]提出光场描述符(Light Field Descriptor, LFD),该方法首先利用正十二面体将三维模型包围,在正十二面体中取10个非对称点,然后在10个顶点中的每个顶点设置10个不同的光场,最后得到100个视图来表征三维模型. Shi等[9]利用柱体包围三维模型,将每个三维模型转化成1张全景图,然后进行分类任务.

以上方法虽然可以很好的将三维模型表示成二维视图序列,但是忽略了不同视图之间的差异性,大量冗余的视图导致分类效率和准确率低. 王鹏宇等[10]基于视点熵在42个固定相机视角获取的视图中选取12个视点熵最高的代表性视图,减少冗余信息,从而得到最能表征模型的特征,提高了网络的鲁棒性. Zhou等[11]利用长短期记忆(Long Short Term Memory,LSTM)根据多视图上下文自适应选择代表性视图. 汤磊等[12]提出使用K-means算法选择二维代表性视图. Ding等[13]在三维模型周围分布6个视点组,学习6个三维模型分类器. 加权融合多个图像清晰度评价函数,选择1张最具代表性的视图输入到6个分类器中,通过策略融合的方式得出最终的分类结果.

也有学者提出将二维视图集有效融合的方法,进一步提高分类准确率. Liu等[14]提出了一个多视图融合网络,该网络将多个视图特征融合到一个紧凑的描述符中. Huang等[15]提出将卷积神经网络和循环神经网络结合到一起,将视角之间的相关性转化为结构化递归神经网络之间的依赖关系,获取三维模型的结构信息,将视图特征进行融合. Su等[16]提出依靠二维图像分类网络提取多视图特征,然后通过最大池化融合视图特征获得紧凑的特征描述符. Feng等[17]和Wang等[18]通过特征级别的融合策略对视图组上的多视图进行了分组融合. Han等[19]通过分层注意力以及卷积神经网络融合视图的特征. Zhou等[20]利用长短时记忆网络获取视图间的上下文信息对视图特征进行融合. Liu等[21]、Liu[22]等和Liang等[23]进一步将长短时记忆网络和注意力机制结合获取了更具有区分性的三维模型特征描述符.

2.2 图卷积神经网络

传统的卷积神经网络只能处理欧式空间的数据,而现实生活中多数场景如社交网络、交通网络等,都是非欧空间的数据,即图的形式存在. 不同于图像和文本,图中每个节点的局部结构各异,这使得平移不变性不再满足. 图卷积神经网络可以有效的对图上复杂的信息进行建模学习,解决了因平移不变性的缺失而导致的难以在图上定义卷积的问题.

现有的图卷积神经网络(Graph Convolutional Neural Network, GCN)分为谱方法和空间方法两类[24]. Bruna等[25]、Henaff等[26]和Xu等[27]使用谱方法利用图上的卷积定理从谱域定义图卷积,解决图上平移不变性缺失的问题. Wu等[28]、Monti等[29]和Gilmer等[30]采用空间方法从节点域出发,通过定义聚合函数来聚合

每个中心节点和其邻近节点,从节点域学习聚合函数,取得了较好的实验效果.

三维模型存在复杂的空间信息,因此有学者提出将图卷积神经网络应用于三维模型分类中. Lin等[31]跨尺度从点云中提取局部三维特征,定义具有图最大池机制的可学习卷积核,用于三维模型分类. Wei等[32]以视图作为图节点,分层学习和判别三维模型的形状描述符,提出一种基于局部和非局部图卷积的分层网络对三维模型进行分类. Lei等[33]提出利用球形核实现三维点云分类的高效图卷积,量化局部三维空间,学习数据中独特的几何关系. 本文采用图卷积神经网络的空间方法,在空间上根据三维模型的投影方式搭建以多个视图作为图节点的图结构,聚合节点间信息,获取不同视图间复杂的空间信息.

3　三维模型姿态

目前在使用深度神经网络对三维物体进行分类时,采用的训练数据和测试数据都是姿态对齐的. 然而,现实应用中,三维模型的姿态是未知的,对齐姿态训练出来的分类网络在未知姿态输入时分类性能急剧降低. 其原因在于,用于识别的视图在姿态稍微有偏差的情况下,差异很大. 因此,一个有效的三维模型分类方法,不仅需要适应姿态对齐的情况,还需要适应姿态未知的情况. 如图1所示,以ModelNet数据集的bathtub类和person类为例,对比三维模型的姿态. 图1中三维模型的视图均是在前视角下投影得到的.

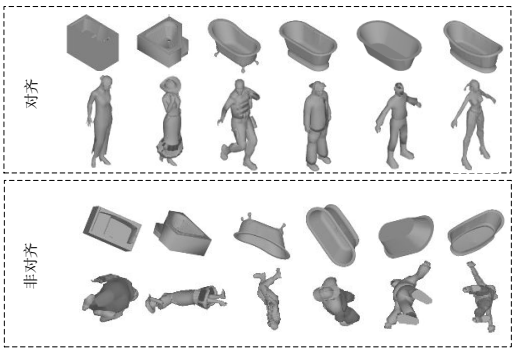


图1　三维模型的姿态

在三维模型姿态对齐的情况下,使用深度神经网络对模型进行分类时,训练视图和测试视图的差异较小,可以取得良好的分类准确率. 然而三维模型在姿态非对齐的情况下,同一视点下,同类别间三维模型视图的差异明显变大. 以person为例,在姿态对齐的情况下,训练和测试时人体的各个部位一一对应,网络可以很轻松的对模型进行预测. 但是在姿态非对齐的情况下,可能会出现使用人的头部进行训练,测试时却变成了四肢,导致网络对三维模型预测难度的增加.

为了获取三维模型非对齐的视图,我们使用开放式图形库(Open Graphics Library,OpenGL)逐行入三维模型的点、线、面数据,载入三维模型. 然后对空间中三维模型的X,Y,Z轴设置随机数,对三维模型进行随机旋转. 然后根据提前设置好的投影位置获取三维模型的二维视图.

4　本文方法

本文提出了一种新的三维模型分类方法,该方法对三维模型姿态对齐和非对齐的情况均可以取得良好的分类效果. 将预先设置好的相机位置作为图结构中的顶点,利用GCN在图结构中良好的特征提取能力来学习视图间的空间信息,并通过时序特征提取网络以及注意力网络进一步提升GCN的运算效果,从而完成三维模型分类.

如图2所示,本文方法由五个主要模块组成,包括:(1)视觉特征提取模块,提取视图的静态视觉特征;(2)时序特征提取模块,提取视图的动态时序特征;(3)图卷积网络模块,获取视图的空间信息;(4)注意力模块,根据视图重要性不同对视图加权学习;(5)分类模块,根据特征描述符对三维模型进行分类.

在学习阶段,根据预先设置好的相机位置,将三维模型表征为多视图序列,然后利用视觉特征提取网络提取静态视觉特征,并引入时序特征提取网络提取动态时序特征. 然后将动态时序特征作为图卷积网络模块以及注意力模块的输入,学习视图的空间信息,并通过注意力模块对视图加权学习最终得到三维模型特征描述符. 最后得到三维模型的分类结果.

4.1　视觉特征提取模块

基于视图的三维模型分类方法首先根据提前设置好的相机位置对三维模型进行投影. 然后将获取到的视图序列输入到视觉特征提取网络中得到三维模型的静态视觉特征.

4.1.1　三维模型投影

目前,三维模型投影为二维视图的算法众多,可以将三维模型随机的放置于球体或者正多面体中进行投影,也可以根据用户实际需求在空间中随机设置相机位置进行投影. 不同的相机位置和相机数量对于三维模型分类结果均有影响,之前已有学者对此展开了研究如:RotationNet[34]和等变多视图网络(Equivariant Multi-view Networks, EMV)[35],分别通过旋转和组卷积的方式,对最佳的视角设置进行研究. 研究发现选用正十二面体这种具有均匀空间分布的视角设置方式,可以取得更好的分类效果. 如图3所示,本文选用正十二面体的投影方法,不指定垂直正向,将虚拟相机放置在围绕该三维模型的正十二面的20个顶点上,得到20张投影视图.

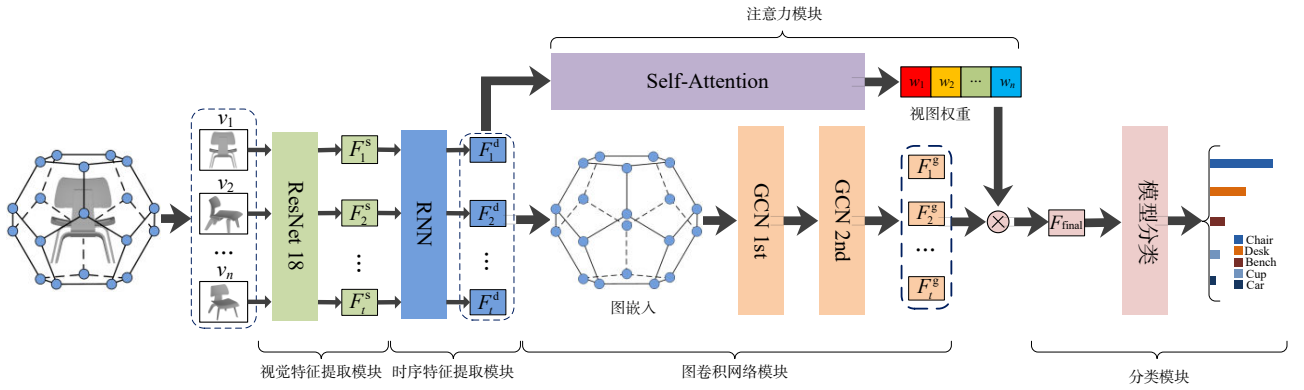


图2 总体框架

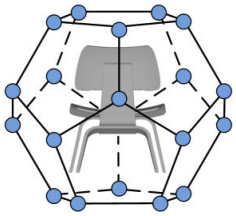


图3 正十二面体的投影方法

4.1.2 静态视觉特征提取

本文采用的ModelNet数据集模型数量较多，训练数据较大，因此本文选用ResNet[36]网络提取视图的静态视觉特征．ResNet的残差结构能在保证网络深度的同时有效避免梯度消失问题．

具体来说，本文选取ResNet18作为视觉特征提取网络，给定一个三维模型S，其对应的二维视图序列为$S(v_i)=(v_1,v_2,\cdots,v_{20}),\ i\in\{1,2,\cdots,20\}$，通过静态视觉特征提取模块学习后可得到该三维模型静态视觉特征$F_t^{\mathrm{s}}=f_{\mathrm{cnn}}(S(v_i)),\ t,i\in\{1,2,\cdots,20\}$.

4.2 时序特征提取模块

静态视觉特征包含的是视图中三维模型的形状、姿态等静态信息．然而相机拍摄具有连续性，因此视图间便具有了一种视角连续变换的动态时序信息，本文使用动态时序特征提取视图间的时序信息．具体来说，相机根据预先设置好的投影方式拍摄获取视图序列$S(v_i)=(v_1,v_2,\cdots,v_{20}),\ i\in\{1,2,\cdots,20\}$．根据相机拍摄的连续性，视图$v_i,\ i\in\{1,2,\cdots,20\}$之间便存在一定的上下文信息，我们称之为视图间的动态时序信息．

本文采用循环神经网络(Recurrent Neural Network，RNN)提取视图间的时序信息，获取具有时序信息的动态时序特征．图4为动态时序特征提取过程示意图，其中F_t^{s}代表视图序列中第t幅视图的静态视觉特征，F_{t-1}^{s}与F_{t+1}^{s}为其相邻视图的静态视觉特征．表达式如下所示：

$$F_t^{\mathrm{d}}=g(V_rZ_t),\ t\in\{1,2,\cdots,20\} \tag{1}$$

$$Z_t=h(U_rF_t^{\mathrm{s}}+W_rZ_{t-1}),\ t\in\{1,2,\cdots,20\} \tag{2}$$

其中，$g(\cdot)$和$h(\cdot)$为激活函数，F_t^{s}代表t时刻的输出，Z_t代表t时刻隐藏层的值，V_r是隐藏层到输出层的参数，U_r是输入层到隐藏层的参数，W_r是每个时间点之间的权重．

本文将视图的静态视觉特征F_t^{s}作为输入，输入到时序特征提取模块中，得到视图的动态时序特征$F_t^{\mathrm{d}}=f_{\mathrm{rnn}}(F_t^{\mathrm{s}})=f_{\mathrm{rnn}}(f_{\mathrm{cnn}}(S(v_i)))$，$t,i\in\{1,2,\cdots,20\}$.

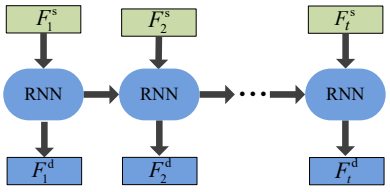


图4 动态时序特征提取

4.3 图卷积网络模块

学习视图间的时序信息和空间信息对于三维模型分类有着十分重要的作用．本文使用RNN对时序信息进行学习，但RNN无法学习到视图的空间信息，学习到的仅为一维的时序信息．为了弥补这一问题，本文使用GCN在空间上对视图进行建模，每一个视图对应于一个图节点，通过邻居节点的信息传递获取多视图的三维空间信息．时序特征提取模块和图卷积模块分别学习了多视图的一维时序信息和三维空间信息，二者缺一不可，共同提升了三维模型的分类准确率．

CNN利用卷积核对特征进行卷积操作，从而提取所需特征．GCN的原理与CNN类似，将欧氏空间的卷积操作推广到非欧空间，通过卷积操作不断更新卷积核的权重参数，聚合节点周围的信息，得到具有空间信息的特征向量，进而完成后续的分类任务[37].

本文的图结构$G=(V,E)$，如图5所示，将预先设置好的相机位置作为图结构中顶点的集合V，视图间的空间距离作为边的集合E，表达式如下所示：

$$V=\{v_i|\forall i\in\{1,2,\cdots,20\}\} \tag{3}$$

$$E=\{e_{ij}|\forall i,j\in\{1,2,\cdots,20\}\} \tag{4}$$

$$e_{ij}=\mathrm{dis}(v_i,v_j),\ i,j\in\{1,2,\cdots,20\} \tag{5}$$

其中,v_i为第i个顶点所对应相机位置的坐标,e_{ij}为顶点间的边,$\mathrm{dis}(\cdot)$表示欧氏距离. 根据顶点间的边e_{ij}找到每个顶点的相邻顶点,表示为邻接矩阵$\boldsymbol{A}$,表达式如下:

$$\boldsymbol{A}=\psi(\min\mathrm{dis}(v_i,v_j);\ \theta_l),\ i,j\in\{1,2,\cdots,20\} \tag{6}$$

其中,$\psi(\cdot)$表示对数据进行索引后通过线性运算得到邻接矩阵$\boldsymbol{A}$,θ_l为线性运算的参数.

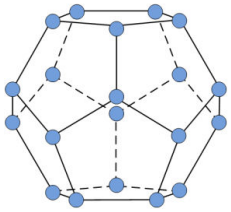


图5　图结构

具体来说,图卷积中顶点V的特征H对应于三维模型的动态时序特征$F_t^{\mathrm{d}},\ t\in\{1,2,\cdots,20\}$. 各个顶点之间的关系组成一个20×20的邻接矩阵$\boldsymbol{A}$,$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$表示图上的拉普拉斯矩阵,其中$\boldsymbol{D}$是度矩阵,归一化后的拉普拉斯矩阵定义为$\boldsymbol{L}=\boldsymbol{I}_n-\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$,其中$\boldsymbol{I}_n\in R^{20\times20}$是单位矩阵,为规范化的邻接矩阵. GCN层与层之间的关系式如下:

$$H^{l+1}=\sigma(\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}H^lW_g^l) \tag{7}$$

其中,H^l为l层顶点的特征,W_g^l为l层的权重,$\sigma(\cdot)$是激活函数.

本文使用两层图卷积神经网络进行训练,将动态时序特征$F_t^{\mathrm{d}},\ t\in\{1,2,\cdots,20\}$通过两层GCN得到输出$F_t^{\mathrm{g}},\ t\in\{1,2,\cdots,20\}$,表达式如下:

$$F_t^{\mathrm{g}}=f_{\mathrm{gcn}}(\boldsymbol{A},F_t^{\mathrm{d}})=\sigma(\boldsymbol{A}\sigma(\boldsymbol{A}F_t^{\mathrm{d}}W_g^0)W_g^1),\ t\in\{1,2,\cdots,20\} \tag{8}$$

其中,W_g为GCN的参数,$\sigma(\cdot)$为激活函数,$f_{\mathrm{gcn}}(\cdot)$为两层GCN运算.

4.4　注意力模块

上述模块聚合了视图的静态视觉特征、动态时序特征和空间位置信息. 然而不同视图对于三维模型分类的重要性是不同的. 因此,本文引入注意力模块,根据重要性不同对视图加权学习,对于三维模型分类更具辨识力的视图赋予更高的权重.

注意力模块由多头自注意力和多层感知器(Multi-Layer Perceptron, MLP)构成,以动态时序特征$F_t^{\mathrm{d}},t\in\{1,2,\cdots,20\}$作为输入,输入到多头注意力块中执行并行运算,输出值连接后通过MLP得到最终值,完成多头注意力运算,表达式如下所示:

$$\mathrm{MultiHead}(F_t^{\mathrm{d}})=\phi(\mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_4),\theta_m) \tag{9}$$

$$w_i=\mathrm{softmax}(\mathrm{MultiHead}(F_t^d)),\ i,t\in\{1,2,\cdots,20\} \tag{10}$$

其中,MultiHead$(\cdot)$为多头注意力运算,Concat$(\cdot)$为连接运算,$\phi(\cdot)$为MLP运算,θ_m是运算参数,head_i为第i头的注意力运算,w_i为权重序列. 多头注意力允许模型关注来自不同特征维度的全局信息,而并行运算有效的提高了实验效率和准确率,本文设置头数为4.

4.5　分类模块

分类模块将注意力模块获取的注意力权重$w_i,\ i\in\{1,2,\cdots,20\}$和图卷积网络模块提取的特征$F_t^{\mathrm{g}},\ t\in\{1,2,\cdots,20\}$进行加权运算后输入到池化层,池化后得到最终的特征描述符F_{final},表达式如下:

$$F_{\mathrm{final}}=\max(F_t^{\mathrm{g}}w_i) \tag{11}$$

将F_{final}输入到全连接层(Fully Connected Layer, FC)中完成三维模型分类任务. 本文的全连接层由两层MLP构成,使用激活函数ReLU增加模型的非线性表达能力,使用交叉熵损失函数训练分类网络,损失函数表达式如下:

$$\mathcal{L}=\mathcal{L}_{\mathrm{c}}(F_{\mathrm{final}},y) \tag{12}$$

其中,y为类别标签,$\mathcal{L}_{\mathrm{c}}$为交叉熵损失.

5　实验与分析

5.1　实验

实验基于Pytorch建立三维模型分类网络框架,在Intel Xeon E5-2678 v3 + RTX 2080的PC机上进行实验. 每个三维模型由包含20个视图的视图序列表征,首先采用ResNet18提取视图的静态视觉特征. 然后将其输出转化成[Batchsize, 20, 512]的形式输入到动态时序特征提取网络,提取视图间时序信息,其中512为特征维度. 接下来将含有时序信息的特征分别输入到图卷积网络和多头自注意力网络,其中图卷积神经网络的层数为2,多头自注意力的头数Heads为4,注意力模块层数为8. 最后将两部分的输出进行加权运算得到最终的模型特征描述符,实现端到端学习的网络结构.

实验中采用随机梯度下降(Stochastic Gradient Descent,SGD)为优化器,其中网络的初始学习率为10^{-2},每10次迭代学习率降低一半. 在第15次迭代的时候将学习率改为10^{-3},衰减率为10^{-3},动量为0.9. 训练时我们使用学习率预热策略[38],学习率在第一轮迭代时从0增加到10^{-2},之后通过余弦函数将学习率从初始值降低到0.

5.2　数据集

实验中采用公共数据集ModelNet10和ModelNet40[39],训练集和测试集比例是4∶1. ModelNet10有10类,共3 991个三维模型,其中测试集有908个三维模型. ModeolNet40有40类,共9 843个三维模型,其中测试集有2 468个三维模型. 本文将实验分为姿态对齐和

非对齐两种情况.

5.3 三维模型姿态对齐实验

为充分证明本文方法在姿态对齐数据集上的有效性,我们在姿态对齐的ModelNet10和ModelNet40数据集上进行实验,并与其他三维模型分类方法进行比较.表1综合对比了基于体素、基于点云和基于视图的算法在ModelNet10和ModelNet40两个数据集上的分类准确率.本文算法在ModelNet10数据集上取得了99.3%的分类准确率,高于排名第二的MVCLN算法3.7%.在ModelNet40数据集上取得了97.4%的分类准确率,高于排名第二的3DRMS算法2.4%,对比结果充分验证了本文方法的有效性.

表1 模型分类准确率对比

算法	输入	数据集	
		ModelNet10	ModelNet40
binVoxNetPlus[40]	体素	92.3%	85.4%
VSL[41]		91.0%	84.5%
G3DNet[42]	点云	93.1%	91.1%
PointNet++[43]		—	90.7%
Nd-Networks[44]		94.0%	91.8%
VA-GCN[45]		—	94.3%
3DRMS[46]	视图	—	95.0%
MVCNN[16]		—	90.1%
LP-3DCNN[47]		94.4%	92.1%
SCFN[21]		94.1%	93.1%
PVR[20]		92.7%	91.6%
MVSG-DNN[11]		94.0%	92.3%
3D2SeqViews[19]		94.6%	91.6%
MVA-CNN[22]		93.0%	92.1%
MVCLN[23]		95.6%	93.4%
本文	视图	**99.3%**	**97.4%**

5.3.1 分类结果分析

图6和图7分别列出了ModelNet10和ModelNet40的混淆矩阵,用于具体分析分类情况.

从图6中可以看出,ModelNet10数据集包括bathtub、bed、chair、desk、dresser、monitor、night_stand、sofa、table和toile 10类三维模型.除了dresser和night_stand这两类存在错分外,其余8类三维模型均取得了接近100%的分类准确率.被错分的三维模型出现在dresser和night_stand中的原因是这两个类别的三维模型极其相似.从图8中可以看出,图8(a)、(c),以及图8(b)、(d)虽然分属dresser类和night_stand类,但是三维模型基本相同,因此极其容易被错分.

从图7中可以看出,除了cup类外,其余类别的三维模型都取得了85%以上的分类准确率.相比ModelNet10准确率有所下降,原因是ModelNet40中存在较多

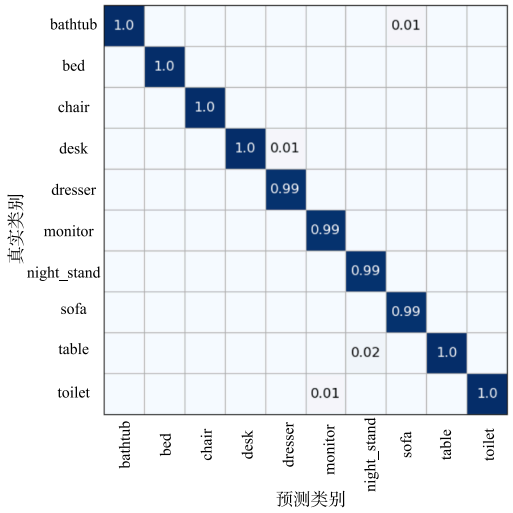


图6 ModelNet10混淆矩阵

的相似三维模型,如:bottle类和cup类、flower_pot类和plant类等.此外ModelNet40中某些类别类内差异很大,如plant类,三维模型是否包含花盆界定不明确.如图9(b)和(c),二者均为plant类,但是图9(b)中的三维模型带有花盆,而图9(c)中的三维模型不带花盆.plant类中包含花盆的三维模型在分类实验中易于与flower_pot类中的三维模型造成混淆.如图9(a)和(b),二者属于不同的类,但是极其相似.

5.4 三维模型姿态非对齐实验

为了验证本文方法在非对齐数据集上的有效性,我们在非对齐的ModelNet10和ModelNet40上进行实验.为进一步研究不同旋转角度对于结果的影响,我们将旋转角度从小到大划分,进行了5组实验,分别为:随机旋转角度(0°~30°)、随机旋转角度(30°~60°)、随机旋转角度(60°~90°)、随机旋转角度(90°~135°)、随机旋转角度(135°~180°),实验结果如表2所示.

从表2可以看出,在非对齐的ModelNet10和ModeNet40数据集上,随着旋转角度的不断增大,三维模型分类准确率有所下降,但并不明显.当随机旋转角度为(0°~30°)时,三维模型的分类准确率最高,分别为94.3%和92.0%.当随机旋转角度为(135°~180°)时,分类准确率最低,分别为93.4%和91.0%.三维模型在随机旋转大角度的分类准确率略低于随机旋转小角度的分类准确率,在ModelNet10和ModelNet40上仅分别降低了0.9%和1.0%,这充分说明了本文提出的方法在三维模型姿态非对齐的情况下具有良好的鲁棒性,可以有效降低因三维模型姿态未知对分类带来的负面影响.

本文方法在三维模型姿态非对齐的情况下,当随机旋转角度(0°~30°)时,在ModelNet10和ModelNet40

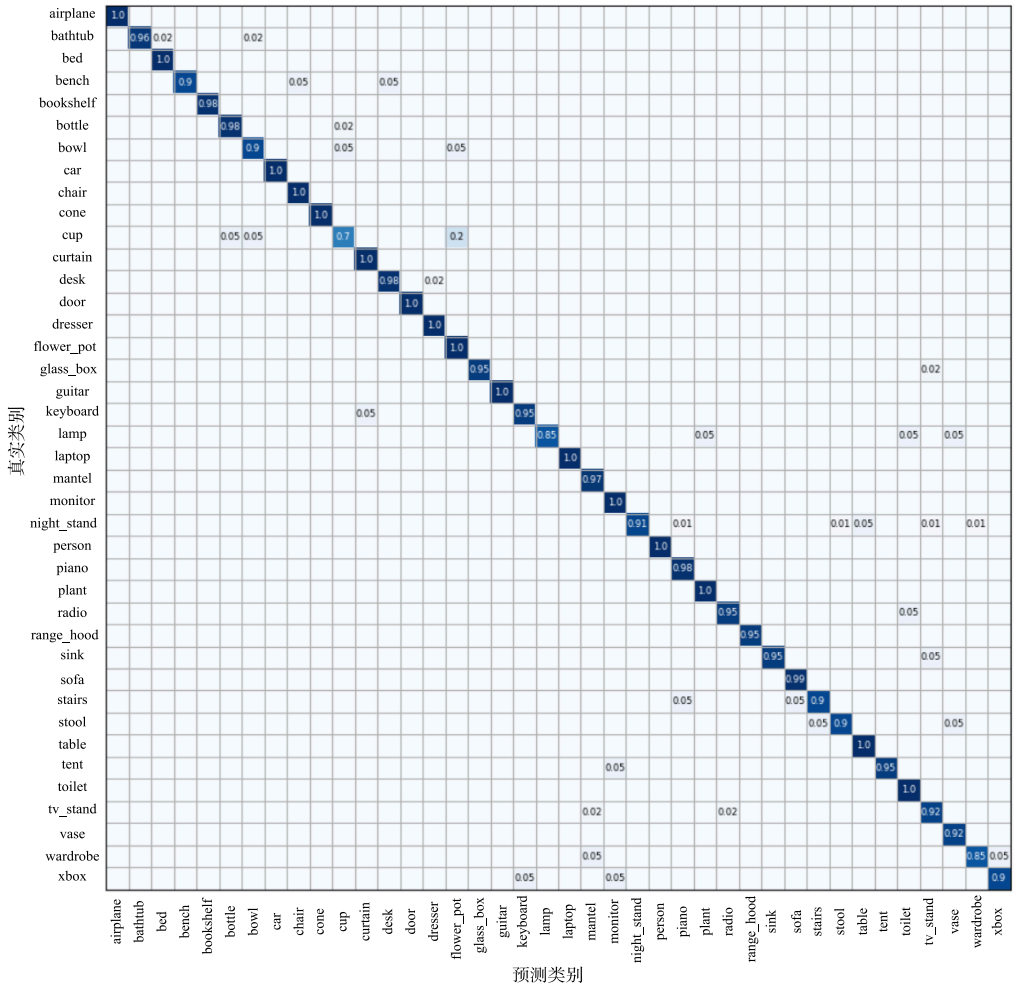
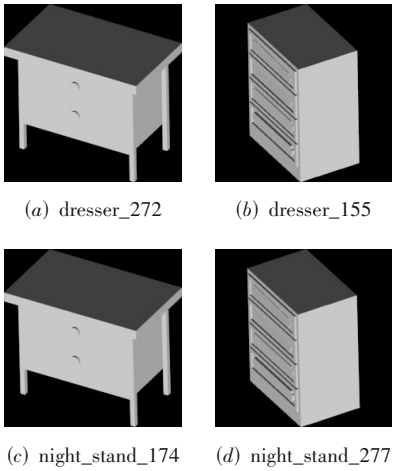


图 7　ModelNet40 混淆矩阵



(a) dresser_272　(b) dresser_155　(c) night_stand_174　(d) night_stand_277

图 8　ModelNet10 中混淆不清的三维模型

(a) flower_pot_02　(b) plant_04　(c) plant_04

图 9　ModelNet40 中混淆不清的三维模型

表 2　三维模型非对齐情况下的分类准确率

非对齐	数据集	
	ModelNet10	ModelNet40
旋转角度(0°~30°)	94.3%	92.0%
旋转角度(30°~60°)	94.3%	91.7%
旋转角度(60°~90°)	93.9%	91.5%
旋转角度(90°~135°)	93.7%	91.1%
旋转角度(135°~180°)	93.4%	91.0%

数据集上分类准确率与对齐相比,下降了 5% 和 5.4%. 虽然有所下降,但仍然保持了一个较高的水平,这说明了我们的方法即适用于三维模型姿态对齐的情况,也适用于姿态非对齐的情况.

5.4.1　姿态非对齐情况下分类准确率对比

在相同的实验条件下,将本文的非对齐数据集迁

移到 MVCNN[16]，GVCNN[17]，RotationNet[35]以及 View-GCN[32]提出的分类网络上进行实验，并与本文方法进行对比．如表3所示，可以看出，在三维模型的非对齐的情况下，MVCNN 在 ModelNet10 上的准确率仅为90.1%，在ModelNet40上的准确率为88.4%，分类准确率最低．而本文方法在两个数据集上的分类准确率分别为94.3%和92.0%，均优于其他四种方法，这说明了本文方法在三维模型姿态非对齐的情况下的有效性．

表3 非对齐下分类准确率对比

方法	旋转角度(0°~30°)	
	ModelNet10	ModelNet40
MVCNN[16]	90.1%	88.4%
GVCNN[17]	91.0%	89.5%
RotationNet[35]	92.9%	91.3%
View-GCN[32]	94.0%	91.8%
本文	**94.3%**	**92.0%**

5.4.2 分类结果分析

图10和图11分别展示了ModelNet10和ModelNet40数据集随机旋转(0°~30°)的混淆矩阵．可以看出，在三维模型姿态非对齐的情况下，仍然有很好的分类准确率．原因在于：(1)本文方法充分提取了视图间的动态时序特征；(2)通过采用图卷积网络模块，有效获取了视图间的空间位置信息．因此，本文方法对于姿态非对齐的三维模型分类是有效的．

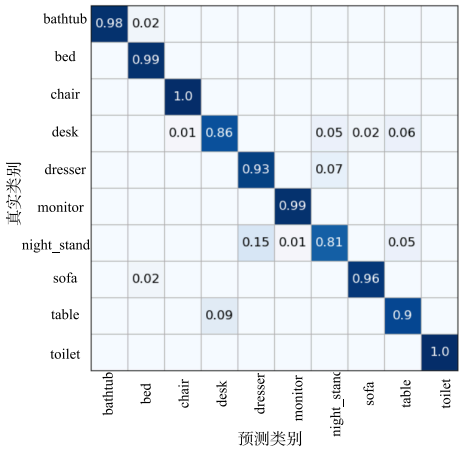


图10 非对齐ModelNet10混淆矩阵

从图10中可以看出，chair和toilet的分类准确率达到了100%，bathtub、bed、dresser、monitor、sofa、table的分类准确率均在90%以上，desk和night_stand表现稍差．其中最差的是night_stand，准确率为81%，有15%的三维模型错分为dresser．从图11中可以看出，非对齐的情况下除了flower_pot、dresser以及night_stand，其余类别的分类准确率均在80%以上，整体分类准确率仍然处于较高水平．flower_pot和plant，night_stand和dresser，正是上文中提到的极容易混淆的类别．

5.5 消融实验

5.5.1 模块消融性实验

本文的三维模型分类网络框架由五个模块构成，分别为：视觉特征提取模块(视觉)、时序特征提取模块(时序)、图卷积网络模块(图卷积)、注意力模块(注意力)和分类模块(分类)．为验证这些模块在提升网络分类性能上的有效性，本文在保证其他条件不变的情况下进行消融实验，实验设置为：(1)视觉+分类；(2)视觉+时序+分类；(3)视觉+图卷积+分类；(4)视觉+时序+图卷积+分类；(5)视觉+时序+图卷积+注意力+分类(本文方法)．在三维模型对齐和非对齐的情况下均进行实验，实验结果如表4所示．

从表4中可以看出，当分类网络包含所有模块时，达到最好的分类效果．在三维模型对齐和非对齐的情况下，视觉+时序+分类相比视觉+分类，分类准确率分别提升了1.8%和1.1%，而视觉+图卷积+分类相比视觉+分类，分类准确率提升了3.4%和2.8%，显然图卷积网络模块对分类效果提升最为明显．可以看出有效的学习视图间的空间关系对三维模型的分类有着至关重要的作用．

表4 不同模块下的分类准确率

算法	对齐	旋转角度(0°~30°)
	ModelNet40	ModelNet40
视觉+分类	91.7%	87.4%
视觉+时序+分类	93.5%	88.5%
视觉+图卷积+分类	95.1%	90.2%
视觉+时序+图卷积+分类	96.2%	91.4%
本文	**97.4%**	**92.0%**

5.5.2 投影方式对比

为评估三维模型的投影方式对分类的影响．本文采用的投影方法与MVCNN[16]提出的12视角以及40视角[12]的投影方法进行了对比．实验结果如表5所示．

从表5中可以看出，在ModelNet10上，无论三维模型是对齐还是非对齐的，20个视角均取得了最好的实验效果．MVCNN将视点全部设置于北纬30°上，相比MVCNN的投影方式，正十二面体的投影方式具有更加均匀的空间分布，可以全方位的表示三维模型．虽然40视角的投影方式同样均匀分布在空间中，但是由于存在冗余视图，导致分类准确率降低．

网络模型的参数量不会随着视角数量增加而改变，然而计算量会增加，ModelNet10以12个视角投影共得到47 892张视图，以20视角投影得到79 820张视图，以40视角投影得到159 640张视图，可以看出40视角

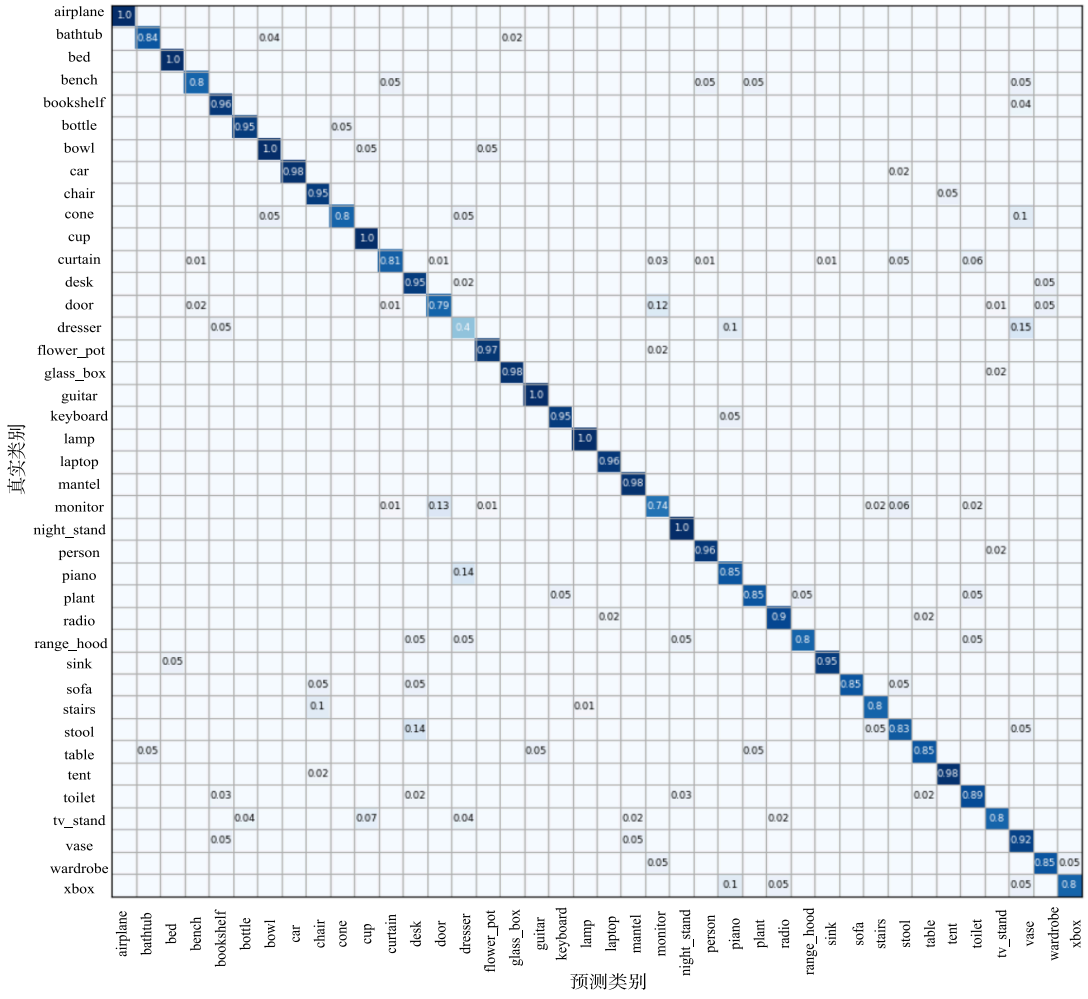


图 11　非对齐 ModelNet40 混淆矩阵

表 5　不同投影方式下的分类准确率

相机数量	ModelNet10		参数量
	旋转角度(0°~30°)	对齐	
12 视角	93.1%	97.1%	37.7 M
20 视角	**94.3%**	**99.3%**	37.7 M
40 视角	93.7%	97.8%	37.7 M

无论是从计算量还是准确率上都明显低于 20 视角,虽然 12 视角的计算量相对较少但是无论是对齐还是非对齐的情况下分类准确率都明显低于 20 视角. 因此通过实验证明,本文所使用的投影方式分类准确率最高并且计算量适中.

5.5.3　注意力模块

表 6 展示了在不同头数和层数下,ModelNet10 随机旋转角度(0°~30°)时的分类准确率和参数量. 通过实验可以看出,在层数相同的情况下,使用多头注意力取得了比单头注意力更好的分类准确率,且参数量不会增长. 此外,头数为 4 和头数为 8 时取得了相同的分类准确率,然而头数为 8 相比头数为 4 有更多的参数量,因此头数足够的情况下再增加头数反而导致实验效率下降. 多头注意力通过并行运算,有助于网络捕捉到更丰富的特征信息. 综合考虑实验准确率和效率,本文设置多头自注意力的头数为 4,层数为 8.

表 6　不同头数和层数注意力块下的分类性能

头数	层数	ModelNet10	参数量
1	12	93.9%	50.3 M
2	12	93.9%	50.3 M
4	12	94.1%	50.3 M
4	8	**94.3%**	**37.7** M
4	6	94.0%	37.4 M
8	12	94.3%	50.3 M

5.5.4　时序特征提取模块

表 7 展示了动态时序特征提取模块选用不同模型对实验效果的影响. 基线是指视觉+图卷积+分类的网络模型,分别对比在动态时序特征提取模块采用 LSTM 或 RNN 的实验结果.

从表 7 中可以看出,动态时序特征提取模块采用

LSTM 或 RNN 的分类准确率均高于基线. LSTM 作为 RNN 的一种变体,在结构中引入了门机制,分别是遗忘门、输入门和输出门,有效的解决了 RNN 由于梯度弥散,导致在序列长度很长时,无法在较后的时间步中按照梯度更新较前时间步的权重矩阵,使得在长时间步过后,网络模型将无法再获取有效的前向序列记忆信息这一问题. 而正因为门机制对输入信息的过滤,对短序列输入来说,会损失一些重要的特征信息,导致实验效果下降. 本文的视图序列属于短序列输入,因此使用没有门机制的 RNN 作为动态时序特征提取网络,减少了视图时序信息的损失,取得了更好的实验效果. 相比 LSTM 在 ModelNet10 和 ModelNet40 数据集上的分类准确率,采用 RNN 使分类准确率提高了 0.3% 和 0.4%.

表 7 不同模型下的分类准确率

深度神经网络模型	旋转角度(0°~30°)	
	ModelNet10	ModelNet40
基线	92.3%	90.2%
LSTM	93.5%	91.0%
RNN	**93.8%**	**91.4%**

6 结论

本文提出了一种适用于三维模型姿态对齐和非对齐两种情况的分类方法. 该方法将三维模型表征为多视图序列,通过视觉特征提取模块、时序特征提取模块、注意力模块和图卷积网络模块,充分融合了多视图序列的静态视觉特征、动态时序特征、空间信息,并根据视图的重要性不同,赋予不同的权重,形成具有代表性的模型特征描述符,实现了对三维模型的分类,取得了良好的实验效果. 实验结果表明,在三维模型姿态对齐的情况下,本文方法均优于现有方法. 在三维模型姿态非对齐的情况下,随着旋转角度的增加,分类准确率均较高并且变化不大,证明了本文方法的鲁棒性和有效性.

参考文献

[1] WANG D, YAO H X, TOMBARI F, et al. Learning descriptors with cube loss for view-based 3-D object retrieval [J]. IEEE Transactions on Multimedia, 2019, 21(8): 2071-2082.

[2] ABDUL R H, YUAN J F, LI B, et al. 2D image-based 3D scene retrieval[C]//Proceedings of the 11th Eurographics Workshop on 3D Object Retrieval. Delft: Eurographics Association, 2018: 37-44.

[3] PHAM Q H, TRAN M K, LI W H, et al. RGB-D object-to-cad retrieval[C]//Proceedings of the 11th Eurographics Workshop on 3D Object Retrieval. Delft: Eurographics Association, 2018: 45-52.

[4] 白静, 司庆龙, 秦飞巍. 基于卷积神经网络和投票机制的三维模型分类与检索[J]. 计算机辅助设计与图形学学报, 2019, 31(2): 303-314.
BAI J, SI Q L, QIN F W. 3D model classification and retrieval based on CNN and voting scheme[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31 (2): 303-314. (in Chinese)

[5] 李海生, 武玉娟, 郑艳萍, 等. 基于深度学习的三维数据分析理解方法研究综述[J]. 计算机学报, 2020, 43(1): 41-63.
LI H S, WU Y J, ZHENG Y P, et al. A survey of 3D data analysis and understanding based on deep learning[J]. Chinese Journal of Computers, 2020, 43(1): 41-63. (in Chinese)

[6] 白静, 周文惠, 拖继文, 等. 时空信息联合嵌入的端到端三维模型草图检索[J]. 计算机辅助设计与图形学学报, 2021, 33(6): 826-836.
BAI J, ZHOU W H, TUO J W, et al. End-to-end sketch-3D model retrieval with spatiotemporal information joint embedding[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(6): 826-836. (in Chinese)

[7] 张静, 曲志坚, 刘晓红. 基于深度学习的三维模型检索研究[J]. 智能计算机与应用, 2019, 9(3): 54-58.
ZHANG J, QU Z J, LIU X H. Research on 3D model retrieval based on deep learning[J]. Intelligent Computer and Applications, 2019, 9(3): 54-58. (in Chinese)

[8] CHEN D Y, TIAN X P, SHEN Y T, et al. On visual similarity based 3D model retrieval[J]. Computer Graphics Forum, 2003, 22(3): 223-232.

[9] SHI B G, BAI S, ZHOU Z C, et al. DeepPano: Deep panoramic representation for 3-D shape recognition[J]. IEEE Signal Processing Letters, 2015, 22(12): 2339-2343.

[10] 王鹏宇, 水盼盼, 余锋根, 等. 基于多视角卷积神经网络的三维模型分类方法[J]. 中国科学: 信息科学, 2019, 49 (4): 436-449.
WANG P Y, SHUI P P, YU F G, et al. 3D shape classification based on convolutional neural networks fusing multi-view information[J]. Scientia Sinica (Informationis), 2019, 49(4): 436-449. (in Chinese)

[11] ZHOU H Y, LIU A N, NIE W Z, et al. Multi-view saliency guided deep neural network for 3-D object retrieval and classification[J]. IEEE Transactions on Multimedia, 2020, 22(6): 1496-1506.

[12] 汤磊, 丁博, 何勇军. 基于卷积神经网络的高效三维模型检索方法[J]. 电子学报, 2021, 49(1): 64-71.

TANG L, DING B, HE Y J. An efficient 3D model retrieval method based on convolutional neural network[J]. Acta Electronica Sinica, 2021, 49(1): 64-71. (in Chinese)

[13] DING B, TANG L, GAO Z, et al. 3D shape classification using a single view[J]. IEEE Access, 8: 200812-200822.

[14] LIU A N, HU N, SONG D, et al. Multi-view hierarchical fusion network for 3D object retrieval and classification [J]. IEEE Access, 2019, 7: 153021-153030.

[15] HUANG X, WANG M T, ZHANG D J, et al. Multi-view fusion with deep learning for 3D shape classification[C]// 2018 International Conference on Audio, Language and Image Processing (ICALIP). Piscataway: IEEE, 2018: 189-194.

[16] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3D shape recognition [C]//2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 945-953.

[17] FENG Y F, ZHANG Z Z, ZHAO X B, et al. GVCNN: Group-view convolutional neural networks for 3D shape recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 264-272.

[18] WANG C, PELILLO M, SIDDIQI K. Dominant set clustering and pooling for multi-view 3D object recognition [EB/OL]. (2019-06-4) [2022-04-19]. https://arxiv.org/abs/1906.01592.

[19] HAN Z Z, LU H L, LIU Z B, et al. 3D2SeqViews: Aggregating sequential views for 3D global feature learning by CNN with hierarchical attention aggregation[J]. IEEE Transactions on Image Processing, 2019, 28(8): 3986-3999.

[20] ZHOU Y, ZENG F Z, QIAN J C, et al. 3D shape classification and retrieval based on polar view[J]. Information Sciences, 2019, 474: 205-220.

[21] LIU A N, GUO F B, ZHOU H Y, et al. Semantic and context information fusion network for view-based 3D model classification and retrieval[J]. IEEE Access, 2020, 8: 155939-155950.

[22] LIU A N, ZHOU H Y, LI M J, et al. 3D model retrieval based on multi-view attentional convolutional neural network[J]. Multimedia Tools and Applications, 2020, 79(7/8): 4699-4711.

[23] LIANG Q, WANG Y X, NIE W Z, et al. MVCLN: Multi-view convolutional LSTM network for cross-media 3D shape recognition[J]. IEEE Access, 2020, 8: 139792-139802.

[24] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述[J]. 计算机学报, 2020, 43(5): 755-780.
XU B B, CEN K T, HUANG J J, et al. A survey on graph convolutional neural network[J]. Chinese Journal of Computers, 2020, 43(5): 755-780. (in Chinese)

[25] BRUNA J, ZAREMBA W, SZLAM A, et al. Spectral networks and locally connected networks on graphs[EB/OL]. (2013-12-21) [2022-04-19]. https://arxiv. org/abs/1312.6203.

[26] HENAFF M, BRUNA J, LECUN Y. Deep convolutional networks on graph-structured data[EB/OL]. (2015-06-16) [2022-04-19]. https://arxiv.org/abs/1506.05163.

[27] XU B B, SHEN H W, CAO Q, et al. Graph wavelet neural network[EB/OL]. (2019-04-12) [2022-04-19]. https://arxiv.org/abs/1904.07785.

[28] WU F, ZHANG T Y, SOUZA A, et al. Simplifying graph convolutional networks[C]//International Conference on Machine Learning. New York: PMLR, 2019: 6861-6871.

[29] MONTI F, BOSCAINI D, MASCI J, et al. Geometric deep learning on graphs and manifolds using mixture model CNNs[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5425-5434.

[30] GILMER J, SCHOENHOLZ S S, RILEY P F, et al. Neural message passing for Quantum chemistry[C]//Proceedings of the 34th International Conference on Machine Learning - Volume 70. Sydney: JMLR 2017: 1263-1272.

[31] LIN Z H, HUANG S Y, WANG Y C F. Convolution in the cloud: Learning deformable kernels in 3D graph convolution networks for point cloud analysis[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1797-1806.

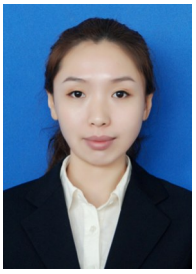
[32] WEI X, YU R X, SUN J. View-GCN: View-based graph convolutional network for 3D shape analysis[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1847-1856.

[33] LEI H, AKHTAR N, MIAN A. Spherical kernel for efficient graph convolution on 3D point clouds[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3664-3680.

[34] KANEZAKI A, MATSUSHITA Y, NISHIDA Y. RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern

Recognition. Piscataway: IEEE, 2018: 5010-5019.
[35] ESTEVES C, XU Y S, ALLEC-BLANCHETTE C, et al. Equivariant multi-view networks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1568-1577.
[36] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.
[37] 王天保, 刘昱, 郭继昌, 等. 图卷积神经网络行人轨迹预测算法[J]. 哈尔滨工业大学学报, 2021, 53(2): 53-60.
WANG T B, LIU Y, GUO J C, et al. Pedestrian trajectory prediction algorithm based on graph convolutional network[J]. Journal of Harbin Institute of Technology, 2021, 53(2): 53-60. (in Chinese)
[38] HE T, ZHANG Z, ZHANG H, et al. Bag of tricks for image classification with convolutional neural networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 558-567.
[39] SHILANE P, MIN P, KAZHDAN M, et al. The princeton shape benchmark[C]//Proceedings Shape Modeling Applications. Piscataway: IEEE, 2004: 167-178.
[40] ZANUTTIGH P, MINTO L. Deep learning for 3D shape classification from multiple depth maps[C]//2017 IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, 2017: 3615-3619.
[41] LIU S K, GILES L, ORORBIA A. Learning a hierarchical latent-variable model of 3D shapes[C]//2018 International Conference on 3D Vision (3DV). Piscataway: IEEE, 2018: 542-551.
[42] DOMINGUEZ M, DHAMDHERE R, PETKAR A, et al. General-purpose deep point cloud feature extractor[C]//2018 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2018: 1972-1981.
[43] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[C]//Conference on Neural Information Processing Systems. Long Beach: NIPS, 2017: 5099-5108.
[44] KLOKOV R, LEMPITSKY V. Escape from cells: Deep kd-networks for the recognition of 3D point cloud models [C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 863-872.
[45] HU H T, WANG F Y, LE H X. VA-GCN: A vector attention graph convolution network for learning on point clouds[EB/OL]. (2021-06-01) [2022-04-19]. https://arxiv.org/abs/2106.00227.
[46] SU J C, GADELHA M, WANG R, et al. A deeper look at 3D shape classifiers[C]//European Conference on Computer Vision. Cham: Springer, 2019: 645-661.
[47] KUMAWAT S, RAMAN S. LP-3DCNN: Unveiling local phase in 3D convolutional neural networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 4898-4907.

作者简介



丁 博 女,1983年出生,河北孟村人. 博士,副教授,硕士生导师. 主要研究方向为三维模型检索、机器学习.
E-mail: dingbo@hrbust.edu.cn

高 源 男,1997年出生,山西吕梁人. 硕士. 主要研究方向为三维模型检索、机器学习.
E-mail: gaoyuan300510@163.com

范宇飞 男,1997年出生,内蒙古包头人. 硕士. 主要研究方向为三维模型检索、机器学习.
E-mail: fyuf520@163.com

何勇军(通讯作者) 男,1980年出生,四川南充人. 博士,教授,博士生导师. 主要研究方向为机器学习、模式识别.
E-mail: holywit@163.com