软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(5):2208−2219 [doi: 10.13328/j.cnki.jos.007026]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于多模态关系建模的三维形状识别方法*

陈浩楠[1], 朱映映[1], 赵骏骐[1], 田 奇[2]

[1](深圳大学 计算机与软件学院, 广东 深圳 518060)

[2](华为技术有限公司, 广东 深圳 518172)

通信作者: 朱映映, E-mail: zhuyy@szu.edu.cn

摘 要: 为了充分利用点云和多视图两种模态数据之间的局部空间关系以进一步提高三维形状识别精度, 提出一个基于多模态关系的三维形状识别网络, 首先设计多模态关系模块 (multimodal relation module, MRM), 该模块可以提取任意一个点云的局部特征和一个多视图的局部特征之间的关系信息, 以得到对应的关系特征. 然后, 采用由最大池化和广义平均池化组成的级联池化对关系特征张量进行处理, 得到全局关系特征. 多模态关系模块分为两种类型, 分别输出点-视图关系特征和视图-点关系特征. 提出的门控模块采用自注意力机制来发现特征内部的关联信息, 从而将聚合得到的全局特征进行加权来实现对冗余信息的抑制. 详尽的实验表明多模态关系模块可以使网络获得更优的表征能力; 门控模块可以让最终的全局特征更具判别力, 提升检索任务的性能. 所提网络在三维形状识别标准数据集 ModelNet40 和 ModelNet10 上分别取得了 93.8% 和 95.0% 的分类准确率以及 90.5% 和 93.4% 的平均检索精度, 在同类工作中处于先进水平.

关键词: 三维形状识别; 关系建模; 多模态学习

中图法分类号: TP391

中文引用格式: 陈浩楠, 朱映映, 赵骏骐, 田奇. 基于多模态关系建模的三维形状识别方法. 软件学报, 2024, 35(5): 2208–2219. http://www.jos.org.cn/1000-9825/7026.htm

英文引用格式: Chen HN, Zhu YY, Zhao JQ, Tian Q. 3D Shape Recognition Based on Multimodal Relation Modeling. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2208–2219 (in Chinese). http://www.jos.org.cn/1000-9825/7026.htm

3D Shape Recognition Based on Multimodal Relation Modeling

CHEN Hao-Nan[1], ZHU Ying-Ying[1], ZHAO Jun-Qi[1], TIAN Qi[2]

[1](College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China)

[2](Huawei Technologies Co. Ltd., Shenzhen 518172, China)

Abstract: To make full use of the local spatial relation between point cloud and multi-view data to further improve the accuracy of 3D shape recognition, a 3D shape recognition network based on multimodal relation is proposed. Firstly, a multimodal relation module (MRM) is designed, which can extract the relation information between the local features of any point cloud and that of any multi-view to obtain the corresponding relation features. Then, a cascade pooling consisting of maximum pooling and generalized mean pooling is applied to process the relation tensor and obtain the global relation feature. There are two types of multimodal relation modules, which output the point-view relation feature and the view-point relation feature, respectively. The proposed gating module adopts a self-attention mechanism to find the relation information within the features so that the aggregated global features can be weighted to achieve the suppression of redundant information. Extensive experiments show that the MRM can make the network obtain stronger representational ability; the gating module can allow the final global feature more discriminative and boost the performance of the retrieval task. The proposed network

* 基金项目: 国家自然科学基金 (62072318); 广东省自然科学基金 (2021A1515012014); 深圳市科技研发资金重点项目 B 类 (20220810142553001)
本文由“多模态协同感知与融合技术”专题特约编辑孙立峰教授、宋新航副研究员、蒋树强教授、王莉莉教授、申恒涛教授推荐.
收稿时间: 2023-04-10; 修改时间: 2023-06-08; 采用时间: 2023-08-23; jos 在线出版时间: 2023-09-11
CNKI 网络首发时间: 2023-12-29

achieves 93.8% and 95.0% classification accuracy, as well as 90.5% and 93.4% average retrieval precision on two standard 3D shape recognition datasets (ModelNet40 and ModelNet10k), respectively, which outperforms the existing works.

Key words: 3D shape recognition; relation modeling; multimodal learning

1 引 言

近些年来, 随着三维深度学习的发展, 三维形状识别领域涌现出大量的方法, 这些方法根据采用的数据表征类型, 主要可以分为以下 3 类: (1) 基于体素的方法. 这些方法会将原始点云数据转化成 3D 体素, 然后应用三维卷积来识别形状, 采用此类方法的工作包括 VoxNet[1], 3D ShapeNets[2]等. 基于体素的方法虽然将不规则的点云转化成了规则的体素, 但是它的缺点也是显而易见的, 即三维卷积带来的高额计算时间和内存开销. (2) 基于多视图的方法, 如 MVCNN[3], GVCNN[4]等. 借鉴了二维卷积神经网络 (convolutional neural network, CNN) 的成功策略, 在这些方法中, 每个单独的视图均采用 CNN 进行特征提取. 在 MVCNN 中, 作者使用一个类似 VGG 的网络分别对每个视角图像进行特征提取, 然后通过 view-pooling 层将所有视角图像的特征聚合, 得到全局特征. 尽管这些方法巧妙地用二维卷积实现了三维数据的特征提取, 但是由于视角的局限性, 多视图数据表征往往难以全面捕捉三维物体完整的结构信息. (3) 基于点云的方法. 作为点云深度学习的先驱, PointNet[5]在特征提取时采用多层感知机 (MLP) 来处理点云的无序性, 随后的 PointNet++[6]在 PointNet 的基础上采用多尺度特征提取的方式来获取点云的局部结构信息. 此外, 如 DGCNN[7], SO-Net[8]等工作也在点云形状识别领域做出了贡献. 然而, 由于点云的稀疏性和无序性, 点云的特征提取过程面临着巨大的困难, 这一困难限制了点云在三维形状表征方面的能力.

上述方法都是基于单一数据表征进行特征提取, 进而实现识别三维物体的形状. 但是, 不同的数据表征具有各自的特点, 如果能够在三维形状识别任务中有效利用多种数据表征的优点并实现他们的互补, 那么模型的性能可能会得到进一步提升. 因此, 一些研究采用多模态学习的方式, 利用两种或更多模态的数据表征作为模型的输入数据, 以此提升模型的性能. 当前, 基于多模态数据的三维形状识别方法大多采用多视图和点云这两种类型的数据表征, 其原因主要有以下两点: (1) 多视图和点云可以直接从设备中获取原始数据. 多视图可以通过 RGB 相机从不同角度拍摄获得, 而点云则可以依靠激光雷达 (LiDAR) 等三维传感器中收集. 相比之下, 体素通常需要通过点云数据转化得到, 并且其内存和计算开销要大于多视图和点云. (2) 多视图和点云存在互补性. 多视图基于图像, 能够捕获到三维物体的颜色、纹理等特征, 但由于视角的限制, 多视图无法完整反映出三维物体的结构. 而点云作为三维数据, 包含了丰富的物体结构信息, 但由于其稀疏性, 缺少如物体的纹理等细节特征. 因此, 在三维形状识别应用中, 结合多视图和点云数据能够弥补两种数据表征各自的不足, 以最大程度地丰富三维物体的特征表现.

现有的基于多模态数据的方法通常会探索并提取不同模态数据之间的对应关系, 并将各模态的数据特征进行融合. 例如, PVNet[9]和 PVRNet[10]分别对多视图全局特征与点云局部特征之间的关系, 以及每个多视图特征与点云全局特征之间的对应关系进行了研究. 然而, 这两项工作并未充分利用两种数据局部特征之间的关系. 如图 1 所示, 对于飞机的头部结构, 红圈标注的几个视图显然能够更多的包含该结构中点云所需的特征, 而其他的视图受限于角度, 很难获取到飞机头部结构的信息. 也即, 对于一个点云局部特征, 只有部分视图的特征与其存在较强的空间关系; 同样, 对于一个视图特征, 也只有部分点云局部特征与其存在较强的空间关系. 这种对应关系的挖掘和利用, 将对多模态数据的特征提取和融合带来新的视角和提升.

为了充分利用不同模态数据的局部空间关系, 本文提出了一个多模态关系模块 (MRM) 用于提取多视图和点云两种模态数据之间的对应关系. 该模块能够提取点云局部特征与多视图特征之间的关系信息, 以得到关系特征张量, 然后通过级联池化将这些关系特征聚合为全局关系特征. 在所提网络中, 本文使用两个多模态关系模块, 分别提取点-视图关系特征与视图-点关系特征. 此外, 为了有效地融合点云全局特征、多视图全局特征以及两种关系特征, 以生成全局特征, 本文设计了一个门控模块. 该模块采用自注意力机制来寻找特征内部的关联信息, 从而将聚合得到的全局特征进行加权, 实现对冗余信息的抑制.

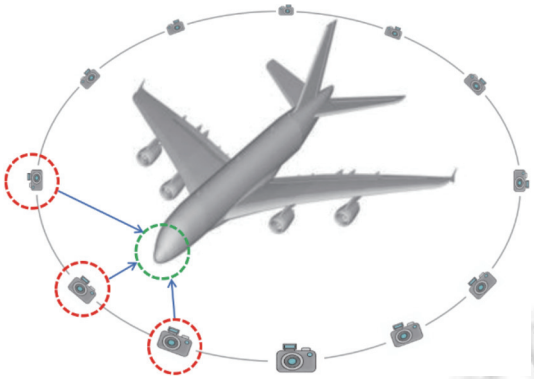

图 1 点云与多视图的空间对应关系示意图

2 相关工作

三维物体的形状识别是计算机视觉领域中的一个基础且重要的研究课题, 选择何种类型的数据作为三维物体的表征, 则是其中的一个关键问题. 在形状识别研究中, 有效的三维数据表征方式能够直接影响到模型的性能和效率, 为后续的特征提取、模式识别等步骤提供基础和依据.

Su 等人提出了名为 MVCNN[3]的神经网络, 该网络基于多视图 (multi-view) 实现三维形状识别. 多视图是指对一个三维物体从多个视角进行拍摄, 得到多张图片. 这样的一组图像可以视为是一个三维物体的表示. MVCNN 使用类似 VGG-M[11]的二维卷积网络对 M 个视角的图像进行特征提取, 特征提取阶段网络共有 M 个分支, 每个分支的卷积网络参数共享, 最后输出 M 个特征. 接着他们使用一个 view-pooing 网络层对着 M 个特征按元素维度实现最大池化, 聚合特征. 最后通过另一个 CNN 得到一个紧凑的形状特征. Feng 等人[4]发现有些视图之间的较为相似, 而有些视图之间的差异较大, 因此直接将视图的特征进行池化会导致特征冗余. 于是他们在 MVCNN 的基础上提出了 GVCNN 网络结构, GVCNN 在特征提取后为每个视图特征计算区分度. 之后对特征分组, 将区分度相近的特征分至同一组, 最后分别对组内特征和组间特征进行池化来聚合特征. 这些工作巧妙地利用了多视图的特点, 将二维卷积网络的成功迁移到了三维形状识别任务中, 得到了良好的准确率. 然而, 多视图由于受到摄像机角度的限制, 不可避免地会丢失部分结构信息, 并且二维图片本身难以展现出三维结构信息.

与多视图不同, 体积表征 (volumetric representation) 将三维空间划分栅格, 保留三维物体占据的栅格, 最后得到的数据由许多体素 (voxel) 组成. Maturana 等人[1]提出利用 3D CNN 对体积表征进行处理, 他们首先对数百个 3D CNN 网络架构进行了简单分类实验, 发现大多数高性能模型的参数量不到 200 万. 基于以上先验, 他们设计出 VoxNet 网络结构, 包含两个卷积层和两个全连接层. VoxNet 在 3 种不同的 3D 数据源中取得了优秀的性能. 然而, 体素的稀疏结构和高额的计算开销限制了网络的性能, 使得基于体积表示的方法通常限于低分辨率的网格.

此外, 点云也是一种重要的三维数据表征. 由于点云数据的稀疏性和无序性, 基于点云的深度学习方法研究曾经进展缓慢[12]. 直到 2017 年, Charles 等人[5]首次将深度学习应用于点云的分类和分割任务中, 提出了名为 PointNet 的网络架构. 相比于过去手工提取点云特征的方法, PointNet 直接对点云进行处理, 在形状识别任务上取得了当时最先进的性能. 然而, PointNet 的缺点在于直接独立地处理每一个点的特征, 忽略了邻域点之间的几何关系. 后续的工作大多关注于局部特征提取, 比如 PoinNet++[6]借鉴 CNN 中的多层感受野思想, 通过下采样加上聚合局部区域特征的方式来扩大感受野; MSP-Net[13]采用多尺度局部区域划分算法聚合多尺度信息. 而 DGCNN[7]则采用动态图卷积来实现局部特征提取.

以上工作都是选择基于一种数据表征来实现三维物体的形状识别, 然而, 每种数据表征存在各自的优缺点. 比如多视图中的图像包含三维物体的颜色、纹理特征, 但是缺少几何结构信息; 而点云虽然包含丰富的几何结构, 但由于稀疏性难以体现出物体的纹理. 由于不同模态的数据存在互补效果, 因此一些工作使用多模态数据作为输入. PVNet[9]

开创性地将多视图和点云两种数据作为输入来实现三维物体的形状识别. PVNet 使用 CNN 对多视图进行特征提取, 并使用 DGCNN 中的边卷积 (EdgeConv) 来对点云进行特征提取. 在将多视图特征进行池化后, PVNet 采用注意力机制来衡量全局多视图特征与点云局部特征之间的相关性, 最后将两种特征拼接, 输入到全连接层得到全局特征. 与之相反, PVRNet[10]进一步探究了多视图特征与点云全局特征之间的相关性, 然后根据注意力得分将点云特征分别与单视图特征与多视图特征融合. 因此, 基于多模态数据的三维形状识别已被证明有效, 是一个亟待研究的方向.

尽管现有的研究已经开始探究两种模态数据特征之间的关系, 但它们主要是提取一种模态的全局特征与另一种模态的局部特征之间的关系信息, 未充分挖掘不同模态的数据之间的局部空间关系. 并且, 现有的方法大多采用晚期融合策略, 即分别对两种模态的数据进行独立的特征提取, 然后对提取到的两种全局特征进行拼接和融合. 此外, 不同模态特征对全局描述符的贡献也值得进一步探索.

3 基于多模态关系建模的网络结构

图 2 展示了本文所提网络的整体结构. 在特征提取阶段, 网络采用双分支结构, 分别针对输入的点云特征和多视图特征进行处理. 对于多视图的特征提取, 本文借鉴 MVCNN[3]的思想, 采用一个二维卷积网络 (本文采用 AlexNet[14]) 对每一个视图进行特征提取, 所有视图的特征提取共享该卷积网络的参数, 最后从每个视图中提取出一个特征向量. 对于点云的特征提取, 本文利用点云深度学习的经典网络 DGCNN[7]来获取点云的局部特征.

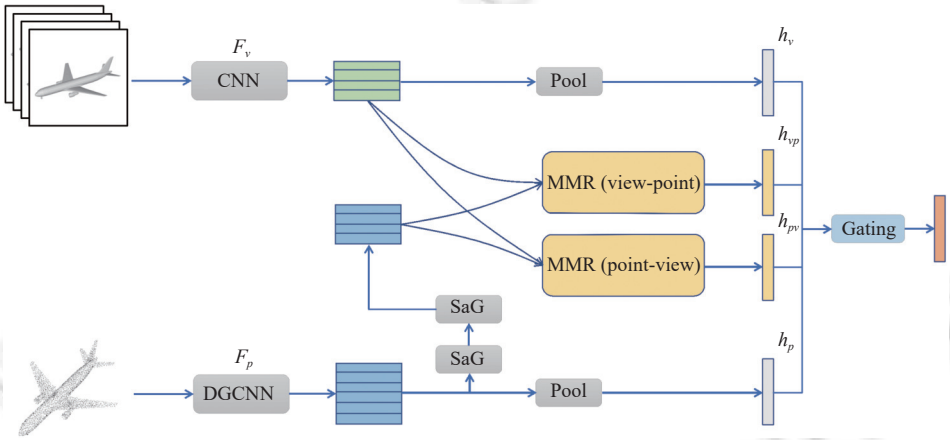

图 2 网络整体结构

在提取两种模态数据的局部空间关系之前, 本文先对点云数据进行下采样来减少点的数量, 这样做的原因有两点: (1) 降低计算开销. (2) 避免提取到冗余的关系. 本文使用采样和分组 (sampling and grouping, SaG) 的方式来实现下采样. SaG 首先使用最远点采样 (furthest point sampling, FPS) 算法选采样 N'个点. 然后以这 N'个点为中心, 使用球查询 (ball query) 算法为每个中心点找到 k 个邻居节点. 最后采用 MLP 来对 k 个邻居节点的特征进行特征提取, 并且使用最大池化聚合局部空间信息:

$$f_p = MaxPooling(MLP(f[f_{pN}^1, f_{pN}^2, \ldots, f_{pN}^k]^{\mathrm{T}})) \tag{1}$$

之后本文使用多模态关系模块 (MRM) 来建模点云和多视图之间的关系信息. 网络中设计了两个 MRM 模块, 其中一个模块使用视图-点 (view-point) 分支, 另一个模块采用点-视图 (point-view) 分支, 分别对应提取点-视图关系特征以及视图-点关系特征. 最后, 本文分别将点云局部特征和多视图局部特征进行最大池化, 然后将两种数据的全局特征和两种关系特征进行拼接, 通过全连接层 (fully connected layers, FC) 融合, 最后输入到门控模块中进行特征加权.

3.1 多模态关系模块

鉴于点云与多视图数据的局部特征存在的空间对应关系, 如果能提取并有效利用这种关系信息, 那么网络的

特征表达能力将得到提升. 现有的工作已尝试探索两种模态数据的关系, 例如, PVNet[9]设计一个注意力融合模块来探索多视图全局特征与点云的局部特征的关系; PVRNet[10]将点云全局特征与不同数量的视图特征进行融合, 并基于关系得分建模点云全局特征和每个视图特征的关系. 这些方法都是基于一种模态数据的全局特征与另一种模态数据的局部特征来建模关系, 但没有充分利用两种模态数据的局部特征之间的关系.

为了能够有效提取点云局部特征与多视图局部特征之间的空间关系信息, 本文设计了多模态关系模块 (multi-modal relation module, MRM), 模块结构如图 3. 对于输入的点云特征 $F_p=[f_{p_1},f_{p_2},\dots,f_{p_N}]$ 以及多视图特征 $F_v=[f_{v_1},f_{v_2},\dots,f_{v_M}]$, 本文分别将点云特征 $F_p\in\mathbb{R}^{N\times D_p}$ 复制 M 次, 扩展得到张量 $F_p\in\mathbb{R}^{N\times M\times D_p}$; 将多视图特征 $F_v\in\mathbb{R}^{N\times D_p}$ 复制 N 次, 扩展得到张量 $F_v\in\mathbb{R}^{N\times M\times D_p}$. 随后将两个特征张量沿着通道维度进行拼接, 使用多层感知机对拼接特征进行提取得到关系特征, 计算过程如下:

$$f_{r_{i,j}}=MLP\left(f_{p_i}\oplus f_{v_j}\right) \tag{2}$$

其中, f_{p_i} 表示第 i 个点的特征, f_{v_i} 表示第 j 个视图的特征, $f_{r_{i,j}}$ 代表这两个局部特征之间的关系特征. $\oplus$ 表示拼接操作, MLP 代表多层感知机.

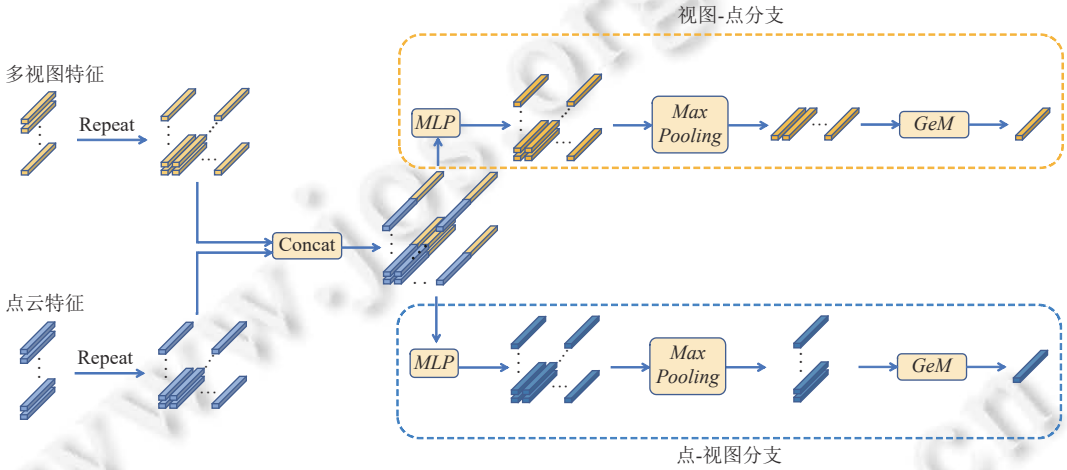

图 3 多模态关系模块

在获得点云和多视图任意两个局部特征的关系信息后, 需要将这些关系特征聚合为全局关系特征, 本文采用级联池化对关系特征张量进行处理. 由图 3 可以看到, 多模态关系模块有两种类型的分支, 分别以多视图特征为基准生成视图-点关系特征, 以及以点云特征为基准生成点-视图关系特征. 以点-视图关系分支为例, 对于关系特征张量 $F_r\in\mathbb{R}^{N\times M\times D_p}$, 本文首先对 M 维进行处理. 这是因为对于点云中的一个点, M 个关系特征分别代表其与 M 个视图之间的关系信息. 由于空间位置的限制, 一个点无法与所有视图都存在较强的空间对应关系, 因此这 M 个关系特征中存在冗余和干扰信息. 为了减少无用信息的干扰, 本文首先使用最大池化 (*MaxPooling*) 对关系特征张量 F_r 进行处理, 得到 $N\times D_r$ 维的关系特征矩阵. 最大池化保留了每个通道的最大响应值, 使得关系信息中的冗余信息被排除. 第 2 次池化用于对关系特征矩阵进行聚合, 即对点云中 N 个点的关系特征进行聚合.在这一过程中, 采用最大池化会导致边缘信息被忽略, 而使用平均池化又会导致聚合到的特征携带冗余信息, 所以本文采用广义平均池化 (generalized-mean pooling, *GeM*)[15], 其计算公式如下:

$$h_{pv}=[h_1\dots h_k\dots h_{D_p}]^{\mathrm{T}},\quad h_k=\left(\frac{1}{|\mathcal{X}_k|}\sum\nolimits_{x_i\in\mathcal{X}_k}x_i^p\right)^{\frac{1}{p}} \tag{3}$$

其中, $\mathcal{X}$ 代表输入的特征矩阵. p 是池化的超参数, 当 $p=1$ 时, *GeM* 池化等于平均池化. 当 $p\to\infty$ 时, *GeM* 池化等于最大池化. p 可以是手动设置的固定数值, 也可以是可训练的参数, 通过反向传播自动调整. 在本文的模块中, 本文将 p 设为可训练的参数, 让网络自行在最大池化和平均池化中进行权衡. 经过由最大池化和 *GeM* 池化组成的级

联池化后, 最终的点-视图关系特征如下:

$$h_{pv} = GeM(MaxPooling(F_r)) \tag{4}$$

视图-点关系分支的流程和点-视图关系分支的流程类似, 不同之处在于视图-点关系分支先对每个视图的 N 个关系特征采用最大池化, 然后对 M 个视图的特征采用 GeM 池化, 最后得到视图-点关系特征 h_{vp} .

3.2 门控模块

在经过对点云和多视图数据的特征提取以及多模态关系模块的计算后, 网络中共有 4 个特征, 分别是: 点云全局特征 h_p , 多视图全局特征 h_v , 点-视图关系特征 h_{pv} , 视图-点关系特征 h_{vp} . 为了将这 4 个特征融合得到全局描述符, 本文首先将 4 个特征进行拼接. 然而由于拼接后的向量维度过高, 本文还使用一个全连接层 (FC) 来将高维向量映射到低维空间, 得到一个低维的全局描述符, 过程如下:

$$h_g = FC(h_p \oplus h_v \oplus h_{pv} \oplus h_{vp}) \tag{5}$$

该全局描述符中包含了从点云和多视图提取到的信息, 这些信息中存在内部关联, 同时也包含一定的冗余信息. 为了找到特征内部的依赖关系, 抑制冗余信息, 本文使用门控机制将不同特征的信息自适应地加权, 计算过程如下:

$$h'_g = \sigma(Wh_g + b) \circ h_g \tag{6}$$

其中, σ 表示激活函数, 这里本文采用了 Sigmoid 激活函数. W 代表一个可学习的参数矩阵, 而 b 代表偏置. $\circ$ 代表逐元素乘积. 门控模块的整体结构如图 4 所示, 对于聚合后的特征, 模块采用自注意力的方式, 将特征本身作为输入来计算出权重向量, 然后使用该权重向量对特征加权. 门控模块捕获了全局特征内部的关联性, 从而输出更精确的结果.

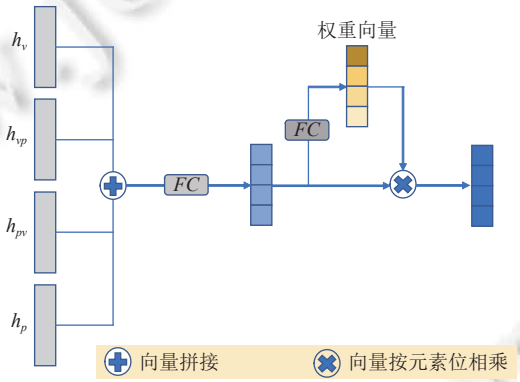

图 4 门控模块

4 实验分析

本文选用两个主流的三维形状识别数据集 (ModelNet40 数据集[2], ModelNet10 数据集[2]) 来对模型的性能进行评估. ModelNet40[2] 是一个大型三维 CAD 模型数据集. 该数据集共有 12 311 个三维物体数据, 可用于点云分类以及检索任务. ModelNet10[2] 是 ModelNet40 数据集[2]的子集, 它包含 10 个流行类别的对象, 共计有 4 899 个三维物体, 并且采用人工方式对齐了这 10 类三维模型的方向. 本文分别在分类和检索两个任务上验证模型的性能. 在分类任务中, 网络的最后一层输出分类分数, 本文使用总体准确率 (overall accuracy, OA) 作为评价指标. 在检索任务中, 本文将网络倒数第 2 层的输出作为全局特征来进行检索, 本文使用平均检索精度 (mean average precision, mAP) 作为评价指标.

4.1 实验设置与训练策略

模型输入的数据分别是具有 1 024 个点的点云数据以及 12 个视图的多视图数据. 本文网络的多视图特征提取

分支采用 MVCNN 的方法, 使用一个 AlexNet 网络对所有视图进行特征提取, 每个视图得到一个 4 096 维的特征, 然后使用一个全连接层将特征降维至 1 024 维. 在点云特征提取中, 本文采用 DGCNN 中的 EdgeConv, EdgeConv 中的邻居数量设为 20. 对于这两个特征提取分支, 本文都使用了预训练模型以加快训练速度.

本文使用深度学习框架 PyTorch 在一张显存为 12 GB 的 NVIDIA GTX TITAN X 显卡上训练所提出的网络. 本文采用一种特殊的学习策略来训练网络. 在前 12 个 epoch 中, 将两个特征提取分支的参数固定, 使得反向传播只调整网络其余部分 (包括 MRM 模块、门控模块等) 的参数. 在经过 12 个 epoch 的训练之后, 网络才对所有的参数进行更新. 选取该训练策略是因为网络的特征提取分支经过预训练, 容易达到收敛, 而在特征提取之后的网络结构的参数是随机初始化, 因此需要更多的训练次数来进行优化. 本文使用交叉熵损失函数来进行训练, 设置训练批次大小为 20. 使用 SDG 算法进行优化, 学习率设为 0.01, 权重衰减设为 0.000 1, 动量为 0.9.

4.2 与其他先进方法的比较

将本文提出的网络与近年来的三维形状识别任务中的先进工作进行了比较, 这些方法包括: 基于体素的方法 (如 3D ShapeNets[2]、VoxNet[1]、VRN[16]), 基于多视图的方法 (如 MVCNN[3]、GVCNN[4]、SeqViews[17]、3D2SeqViews[18]), 基于点云的方法 (PointNet[5], PointNet++[6], DGCNN[7], PCT[19]). 为了公平比较, 本文还选择了基于多模态数据的形状识别方法 (如 FusionNet[20] , PVNet[9], PVRNet[10]) 来进行对比.

比较结果如表 1 所示, Vx, Mv 和 Pt 分别代表模型输入的数据为体素、多视图和点云. 可以看到本文的方法在 ModelNet40 数据集上的分类任务中, 其总体准确率 (OA) 达到了 93.8%, 性能超越了所有的对比方法. 而在检索任务中, 本文方法的平均检索精度 (mAP) 达到了 90.5%. 在对比的 14 个方法中, 仅次于单模态方法 3D2SeqViews[18]. 而在采用多模态的方法中达到最佳. 在特征提取阶段, 本模型分别采用了 MVCNN 和 DGCNN 的模块结构, 因此这两个方法可以被视为基准方法. 相比于 MVCNN 和 DGCNN, 本文的模型在 OA 评价指标上分别提升了 3.9% 和 1.6%, 在 mAP 评价指标上更有明显的提升, 分别达到了 10.3% 和 8.9%. 这说明使用多模态数据相比于只使用单模态数据能显著提升模型的性能, 在三维形状识别上获得更优的表现.

表 1 与其他方法在分类和检索的性能比较 (%)

方法	输入	ModelNet40		ModelNet10	
		OA	mAP	OA	mAP
ShapeNets[2]	Vx	77.30	49.20	83.50	68.30
VoxNet[1]	Vx	83.00	—	92.00	—
VRN[16]	Vx	91.30	—	93.60	—
MVCNN[3]	Mv	89.90	80.20	—	—
GVCNN[4]	Mv	93.10	84.50	—	—
SeqViews[17]	Mv	93.40	89.10	94.80	91.40
3D2SeqViews[18]	Mv	93.40	**90.80**	94.70	92.10
PointNet[5]	Pt	89.20	—	—	—
PointNet++[6]	Pt	90.70	—	—	—
DGCNN[7]	Pt	92.20	81.60	—	—
PCT[19]	Pt	93.40	—	—	—
FusionNet[20]	Vx, Mv	90.80	—	93.10	—
PVNet[9]	Pt, Mv	93.20	89.50	—	—
PVRNet[10]	Pt, Mv	93.60	90.50	—	—
Ours	Pt, Mv	**93.80**	90.50	**95.00**	**93.40**

本文对比了基于点云和基于多视图的前沿方法, 这些方法主要聚焦特征提取的改进. 例如, 基于点云的 PCT[19] 采用 Transformer[21]结构来捕捉点云中的局部上下文, 而基于多视图的神经网络 3D2SeqViews[19]聚合视图内容信息和视图之间的顺序空间性. 相比之下, 尽管本文模型虽然在特征提取阶段采用了相对简单、轻量的结构 (AlexNet

和 EdgeConv), 但是根据实验结果, 本文的模型性能仍具备竞争力.

并且, 本文还对比了其他基于多模态数据的方法, 如 PVNet[9]和 PVRNet[10]. 它们均为基于点云和多视图数据的方法, 其中 PVRNet 在性能上略优于 PVNet. 实验结果表明, 本文方法在 OA 分类指标上超越了这两项工作, 在 mAP 检索指标上与 PVRNet 相当, 优于 PVNet. 这足以说明相比现有的基于多模态数据的方法, 本文提出的方法能够更好地利用不同模态数据间的空间对应关系.

此外, 本文还在 ModelNet10 数据集上进行了测试. 由表 1 可以看到, 本文的模型在 OA 评价指标和 mAP 评价指标上均超越了所有的对比方法, 这进一步证实了所提方法的有效性. 相较于该数据集上的最先进方法 3D2Seq-Views, 所提模型在 OA 和 mAP 指标上分别提升了 0.3% 和 1.3%.

4.3 消融实验

为了充分验证模型中各个模块的有效性, 本文在 ModelNet40 数据集上进行了消融实验. 实验结果如表 2 所示, 其中 Multi-View Branch 和 Point Cloud Branch 分别代表只使用多视图全局特征和只使用点云全局特征 (即直接使用 MVCNN 和 DGCNN 的输出). Late Fusion 表示直接将点云特征和多视图特征进行拼接, 然后通过 *MLP* 层融合. 可以看到这种简单的融合方式在分类任务上的性能提升有限, 相较于 Point Cloud Branch 仅略微提升了 0.3%. 本文还测试了单个 MRM 模块在网络中的效用, Point-View 和 View-Point 分别表示将 MRM 模块输出的点-视图关系特征和视图-点关系特征分别单独的融合到全局特征中, 相对于 Late Fusion, 它们在 OA 指标上性能分别提升了 0.7% 和 0.5%. 另外, 网络使用两个不同分支的 MRM 模块 (即 both) 在性能表现上更优, 这说明点-视图关系特征和视图-点关系特征能够共同提升模型的性能. 最后, 通过加入门控模块 (gating), 网络达到了最佳性能. 结果显示, 门控模块对检索任务的性能指标 mAP 有更大的提升. 分析认为, 这是由于门控模块对用于检索的全局特征进行加权, 可以突出有价值的信息, 抑制干扰信息, 从而进一步增强特征的判别力.

表 2 在 ModelNet40 数据集上的消融实验结果 (%)

模型	OA	mAP
Multi-View Branch (MVCNN)	89.90	80.20
Point Cloud Branch (DGCNN)	92.20	81.60
Late Fusion	92.50	89.10
MRM (Point-View)	93.20	89.90
MRM (View-Point)	93.00	89.70
MRM (both)	93.60	89.90
MRM (both)+Gating	**93.80**	**90.50**

4.4 鲁棒性实验

为了评估本模型的鲁棒性, 本文探索了模型在面对数据缺失情况下的性能表现. 在标准情况下, 本模型使用的数据是具有 1 024 个点的点云数据和具有 12 个视图的多视图数据. 在进行鲁棒性测试中, 本文会刻意减少多视图和点云数据的数量, 从而破坏了数据的完整性, 引入数据缺失的情况.

(1) 多视图数据的缺失. 在固定点云中点的数量为 1 024 的情况下, 本文分别将视图的数量设为 4、8、10、12 进行实验. 多视图的视角是围绕在三维物体一周的, 对于 4、8、10、12 视图, 每个视图的角度差分别为 90°、45°、36°、30°. 这意味着视图数量越少, 每个视图之间的角度差越大, 因此信息丢失也就越严重. 实验结果如表 3 所示, 其中 AAC 评价指标是类平均准确率 (accuracy average class), 代表分类任务中每个类别准确率的均值. 从表 3 中可以看到, 在视图数量从 12 下降到 8 时, MVCNN 和本文的模型在性能上都只有轻微的下降. 然而, 当视图数量降至 4 时, MVCNN 的性能下降明显, 与 12 个视图的情况相比, 其在总体准确率 (OA) 上下降了 5.3%. 相反, 本文的模型在 OA 指标上仅下降了 1.3%.

(2) 点云数据的缺失. 为了模拟点云数据的缺失, 本文对原始点云数据进行下采样来减少点的数量, 即在视图数量固定为 12 的情况下, 将点的数量分别设为 128、256、384、512、768 和 1 024. 随着点的数量从 1 024 减少

到 128, 点云的结构在视觉上越来越难以分辨, 这对模型的识别能力造成了艰巨的挑战. 图 5 显示了测试结果, 可以观察到 DGCNN 对于点的缺失比较敏感: 当点的数量为 256 时, 模型 OA 指标上的性能仅为约 50%. 当点的数量降至 128 时, DGCNN 分类能力几乎完全丧失. 然而, 本文的模型在点的数量减少的情况下依然能够维持较好的结果, 即使点的数量为 128 时, 仍有接近 80% 的总体准确率.

表 3　在 ModelNet40 数据集上视图数量对模型性能的影响

模型	视图数量	OA (%)	AAC (%)
MVCNN	4	84.60	82.50
	8	89.00	87.20
	10	89.30	87.40
	12	89.90	87.60
Ours	4	92.50	90.60
	8	93.40	91.00
	10	93.70	91.50
	12	**93.80**	**91.70**

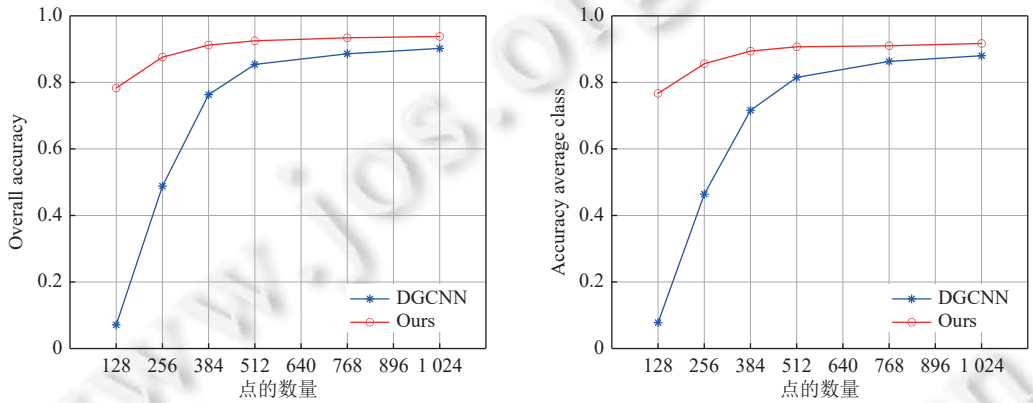

图 5　ModelNet40 数据集上点的数量对模型性能的影响

这些实验结果充分验证了本文模型的鲁棒性: 即使在数据存在缺失的情况下, 所提的模型仍能保持较高的性能. 本文认为多模态数据的输入是模型能展现出良好鲁棒性的主要原因, 即使其中一种数据存在缺失, 模型也能从另一种模态的数据中获取补充信息. 此外, 多模态关系模块也增强了模型的鲁棒性, 在面对数据缺失情况, 模型可以依赖多模态关系模块提取的关系特征, 从而确保网络具有充足的信息来完成分类与检索任务.

4.5　级联池化策略对多模态关系模块性能的影响

在多模态关系模块中, 本文使用了级联池化来聚合关系特征张量, 输出最终的关系特征. 本节探索了不同池化策略对模块性能的影响. 实验结果如图 6 所示, 其中 M、S、G 分别代表最大池化、求和池化和广义平均池化. 本实验中测试了多种池化策略的组合, 例如 M+G 代表先使用最大池化, 再使用广义平均池化. 根据图 6 中的结果, 可以得到以下结论.

(1) 在第 1 次池化过程中, 使用最大池化要优于求和池化和广义平均池化. 实验结果表明, 使用求和池化会导致性能有较大幅下降, 这是因为第一次池化是用于聚合每个视图与所有点 (或者每个点与所有视图) 的关系特征, 由于空间位置的限制, 这些关系特征存在较多的冗余信息, 而求和池化会将所有的信息视作同等贡献来进行聚合, 导致聚合的特征中包含大量干扰信息. 此外, 尽管广义平均池化虽然在指数参数 p 趋于无穷时等同于最大池化, 但是在网络训练中反向传播过程中, 难以将参数 p 调整为非常大的数值. 因此, 与最大池化相比, 广义平均池化提取高响应信息的能力较为不足.

(2) 在第 2 次池化过程中, 使用广义平均池化要优于最大池化和求和池化. 第 2 次池化的目标是将所有视图

(点) 的关系特征进行聚合, 以得到最终的输出. 在这个过程中, 求和池化会融合所有的关系信息, 但没有考虑突出重要信息. 而最大池化只保留高响应的信息, 忽视了信息的全面性. 而广义平均池化能够通过自适应的调整参数 p, 综合最大池化和平均池化, 使得该池化过程能兼顾提取全面信息和突出重要信息. 因此, 在第 2 次池化过程中广义平均池化相较于其他两个池化策略, 有更好的表现.

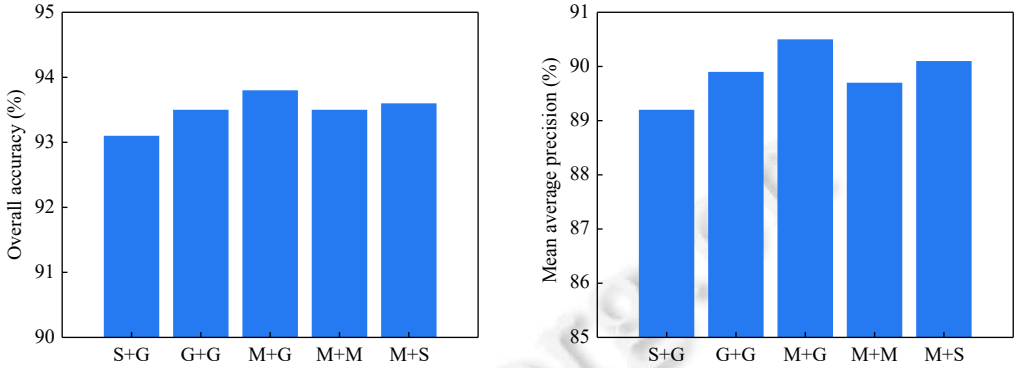

图 6 在 ModelNet40 数据集上池化策略对性能的影响

根据以上的实验结果与分析, 本文最终在多模态关系模块中选择了先使用最大池化, 再使用广义平均池化的级联池化策略.

4.6 模型复杂度分析

在本节中, 对本文提出的模型进行了复杂度分析, 以评估其效率和性能. 为了进行公平地比较, 本文选择了基于多模态数据的形状识别的开源方法, 包括 FusionNet[20]、P2P-HorNet[22]和 RECON[23], 并在广泛使用的 ModelNet40 数据集上进行了实验对比, 如表 4 所示, 其中#Param 表示模型的推理参数量, #FLOPs 表示模型的浮点运算量, PT 表示是否在预训练阶段使用了教师模型. 从表 4 与表 1 可以看出, 相较于同样采用 CNN 架构的 FusionNet, 本文提出的模型参数量仅为 FusionNet 的 1/3, 而在 ModelNet40 和 ModelNet10 两个数据集上, 分类任务综合评价指标 OA 分别提升了 3% 和 1.9%. 尽管相对于基于 Transformer 架构的 P2P-HorNet 与 RECON, 本文模型的 OA 略低, 但本文无需使用教师模型, 模型的推理参数量以及浮点运算量 (FLOPs) 也优于 P2P-HorNet 与 RECON.

表 4 模型的复杂度比较

方法	骨干网	PT	#Param (M)	#FLOPs (G)	OA (%)
					ModelNet40
FusionNet[20]	CNN	×	118.5	—	90.8
P2P-HorNet[22]	Transformer	√	—	34.6	94.0
RECON[23]	Transformer	√	43.6	5.3	94.7
Ours	CNN	×	36.7	4.9	93.8

4.7 可视化实验

本文对 ModelNet40 数据集上的检索结果进行了可视化, 实验结果如图 7 (以三维形状的视图作为示例). 第 1 列为查询数据, 其余 4 列展示的是与查询数据特征的欧氏距离最小的前 4 个结果. 需要强调的是, 本文的模型是在分类任务上进行训练, 并直接用于检索测试. 虽然可以采用检索标签对模型进行重训练或者采用度量学习方法对模型进行微调, 以此增强模型的检索性能, 但本文认为在分类任务中训练得到的模型已有能力生成具有判别力的特征. 图 7 表明所提网络能准确地检索出查询数据的同类对象, 这证明了模型能够有效地将类内特征之间的距离减小, 类间特征之间的距离增大, 从而得到具有判别力的全局特征.

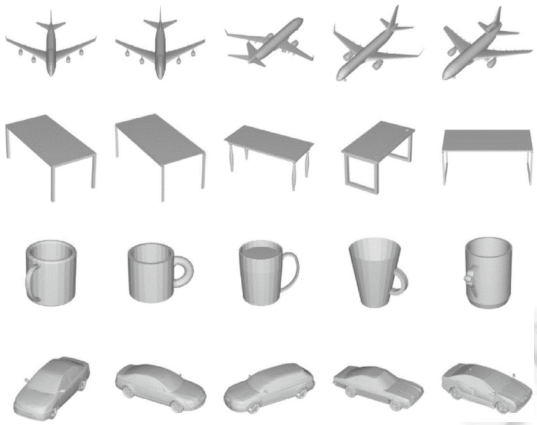

图 7　在 ModelNet40 数据集上检索结果可视化图

5 总　结

本文首先对现有的三维形状识别工作进行了论述, 然后介绍了基于关系建模的多模态三维形状识别方法. 本文设计了一个多模态关系模块 (MRM) 用于建模点云和多视图局部特征之间的空间关系. 在此基础上, 本文提出了一个基于多视图和点云数据的三维形状识别网络, 该网络分别对点云和多视图进行特征提取, 然后对于两种模态的局部特征, 使用两个多模态关系模块分别提取出点-视图关系特征和视图-点关系特征. 并且, 网络能将多种特征进行融合, 利用本文提出的门控模块加权融合输出特征, 以抑制冗余信息. 为了验证所提方法的有效性, 本文在 ModelNet40 和 ModelNet10 两个知名的三维形状识别数据集上进行了广泛的实验. 对比基于多视图、体素、点云和基于多模态数据的主流方法, 本文方法在两个数据集上的分类和检索任务中都取得了具有有竞争力的结果. 此外, 本文还通过消融实验、鲁棒性测试、模型复杂度、可视化等实验实现了对所提模型的深入分析.

References:

[1] Maturana D, Scherer S. VoxNet: A 3D convolutional neural network for real-time object recognition. In: Proc. of the 2015 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Hamburg: IEEE, 2015. 922–928. [doi: 10.1109/IROS.2015.7353481]

[2] Wu ZR, Song SR, Khosla A, Yu F, Zhang LG, Tang XO, Xiao JX. 3D ShapeNets: A deep representation for volumetric shapes. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1912–1920. [doi: 10.1109/CVPR.2015.7298801]

[3] Su H, Maji S, Kalogerakis E, Learned-Miller E. Multi-view convolutional neural networks for 3D shape recognition. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 945–953. [doi: 10.1109/ICCV.2015.114]

[4] Feng YF, Zhang ZZ, Zhao XB, Ji RR, Gao Y. GVCNN: Group-view convolutional neural networks for 3D shape recognition. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 264–272. [doi: 10.1109/CVPR.2018.00035]

[5] Charles RQ, Su H, Kaichun M, Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 77–85. [doi: 10.1109/CVPR.2017.16]

[6] Qi CR, Yi L, Su H, Guibas LJ. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5105–5114.

[7] Wang Y, Sun YB, Liu ZW, Sarma SE, Bronstein MM, Solomon JM. Dynamic graph CNN for learning on point clouds. ACM Trans. on Graphics, 2019, 38(5): 146. [doi: 10.1145/3326362]

[8] Li JX, Chen BM, Lee GH. SO-Net: Self-organizing network for point cloud analysis. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9397–9406. [doi: 10.1109/CVPR.2018.00979]

[9] You HX, Feng YF, Ji RR, Gao Y. PVNet: A joint convolutional network of point cloud and multi-view for 3D shape recognition. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: ACM, 2018. 1310–1318. [doi: 10.1145/3240508.3240702]

[10] You HX, Feng YF, Zhao XB, Zou CQ, Ji RR, Gao Y. PVRNet: Point-view relation neural network for 3D shape recognition. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence and the 31st Innovative Applications of Artificial Intelligence Conf. and the 9th AAAI Symp. on Educational Advances in Artificial Intelligence. Honolulu: AAAI, 2019. 1119. [doi: 10.1609/aaai.v33i01.33019119]

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[12] Wang CS, Wang H, Ning X, Tian SW, Li WJ. 3D point cloud classification method based on dynamic coverage of local area. Ruan Jian Xue Bao/Journal of Software, 2023, 34(4): 1962–1976 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6683.htm [doi: 10.13328/j.cnki.jos.006683]

[13] Bai J, Xu HJ. MSP-Net: Multi-scale point cloud classification network. Journal of Computer-aided Design & Computer Graphics, 2019, 31(11): 1917–1924 (in Chinese with English abstract). [doi: 10.3724/SP.J.1089.2019.17903]

[14] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the 25th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

[15] Radenović F, Tolias G, Chum O. Fine-tuning CNN image retrieval with no human annotation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1655–1668. [doi: 10.1109/TPAMI.2018.2846566]

[16] Brock A, Lim T, Ritchie JM, Weston N. Generative and discriminative voxel modeling with convolutional neural networks. arXiv:1608.04236, 2016.

[17] Han ZZ, Shang MY, Liu ZB, Vong CM, Liu YS, Zwicker M, Han JW, Chen CLP. SeqViews2SeqLabels: Learning 3D global features via aggregating sequential views by RNN with attention. IEEE Trans. on Image Processing, 2019, 28(2): 658–672. [doi: 10.1109/TIP.2018.2868426]

[18] Han ZZ, Lu HL, Liu ZB, Vong CM, Liu YS, Zwicker M, Han JW, Chen CLP. 3D2SeqViews: Aggregating sequential views for 3D global feature learning by CNN with hierarchical attention aggregation. IEEE Trans. on Image Processing, 2019, 28(8): 3986–3999. [doi: 10.1109/TIP.2019.2904460]

[19] Guo MH, Cai JX, Liu ZN, Mu TJ, Martin RR, Hu SM. PCT: Point cloud transformer. Computational Visual Media, 2021, 7(2): 187–199. [doi: 10.1007/s41095-021-0229-5]

[20] Hegde V, Zadeh R. FusionNet: 3D object classification using multiple data representations. arXiv:1607.05695, 2016.

[21] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017: 6000–6010.

[22] Wang ZY, Yu XM, Rao YM, Zhou J, Lu JW. P2P: Tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. In: Proc. of the 36th Conf. on Neural Information Processing Systems. New Orleans, 2022. 1–15.

[23] Qi ZK, Dong RP, Fan GF, Ge Z, Zhang XY, Ma KS, Yi L. Contrast with reconstruct: Contrastive 3D representation learning guided by generative pretraining. In: Proc. of the 40th Int'l Conf. on Machine Learning. Honolulu: JMLR.org, 2023. 1171.

附中文参考文献:

[12] 王昌硕, 王含, 宁欣, 田生伟, 李卫军. 基于局部区域动态覆盖的 3D 点云分类方法. 软件学报, 2023, 34(4): 1962–1976. http://www.jos.org.cn/1000-9825/6683.htm [doi: 10.13328/j.cnki.jos.006683]

[13] 白静, 徐浩钧. MSP-Net: 多尺度点云分类网络. 计算机辅助设计与图形学学报, 2019, 31(11): 1917–1924. [doi: 10.3724/SP.J.1089.2019.17903]

陈浩楠(1997－), 男, 硕士, 主要研究领域为点云识别, 点云检索.

赵骏骐(1996－), 男, 硕士, 主要研究领域为目标检测, 目标识别.

朱映映(1976－), 女, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为计算机视觉, 多媒体内容分析, 机器学习.

田奇(1970－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为计算机视觉, 多媒体信息检索.