

语义增强图像-文本预训练模型的零样本三维模型分类

丁博^① 张立宝^① 秦健^① 何勇军^{*②}

^①(哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080)

^②(哈尔滨工业大学计算学部 哈尔滨 150006)

摘要: 目前, 基于对比学习的图像-文本预训练模型(CLIP)在零样本3维模型分类任务上表现出了巨大潜力, 然而3维模型和文本之间存在巨大的模态鸿沟, 影响了分类准确率的进一步提高。针对以上问题, 该文提出一种语义增强CLIP的零样本3维模型分类方法。该方法首先将3维模型表示成多视图; 然后为了增强零样本学习对未知类别的识别能力, 通过视觉语言生成模型获得每张视图及其类别的语义描述性文本, 并将其作为视图和类别提示文本之间的语义桥梁, 语义描述性文本采用图像字幕和视觉问答两种方式获取; 最后微调语义编码器将语义描述性文本具化为类别的语义描述, 其拥有丰富的语义信息和较好的可解释性, 有效减小了视图和类别提示文本的语义鸿沟。实验表明, 该文方法在ModelNet10和ModelNet40数据集上的分类性能优于现有的零样本分类方法。

关键词: 3维模型分类; 零样本; 基于对比学习的图像-文本预训练模型; 语义描述性文本

中图分类号: TN911.7; TP391.4

文献标识码: A

文章编号: 1009-5896(2024)08-3314-10

DOI: 10.11999/JEIT231161

Zero-shot 3D Shape Classification Based on Semantic-enhanced Language-Image Pre-training Model

DING Bo^① ZHANG Libao^① QIN Jian^① HE Yongjun^{*②}

^①(School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

^②(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150006, China)

Abstract: Currently, the Contrastive Language-Image Pre-training (CLIP) has shown great potential in zero-shot 3D shape classification. However, there is a large modality gap between 3D shapes and texts, which limits further improvement of classification accuracy. To address the problem, a zero-shot 3D shape classification method based on semantic-enhanced CLIP is proposed in this paper. Firstly, 3D shapes are represented as views. Then, in order to improve recognition ability of unknown categories in zero-shot learning, the semantic descriptive text of each view and its corresponding category are obtained through a visual language generative model, and it is used as the semantic bridge between views and category prompt texts. The semantic descriptive texts are obtained through image captioning and visual question answering. Finally, the finely-adjusted semantic encoder is used to concretize the semantic descriptive texts to the semantic descriptions of each category, which have rich semantic information and strong interpretability, and effectively reduce the semantic gap between views and category prompt texts. Experiments show that our method outperforms existing zero-shot classification methods on the ModelNet10 and ModelNet40 datasets.

Key words: 3D shape classification; Zero-shot; Contrastive Language-Image Pre-training (CLIP); Semantic descriptive text

收稿日期: 2023-10-26; 改回日期: 2024-03-16; 网络出版: 2024-03-26

*通信作者: 何勇军 heyongjun@hit.edu.cn

基金项目: 国家自然科学基金(61673142), 黑龙江省自然科学基金(LH2022F029, JQ2019F002)

Foundation Items: The National Natural Science Foundation of China (61673142), The Natural Science Foundation of Heilongjiang Province (LH2022F029, JQ2019F002)

1 引言

随着深度学习的不断发展, 闭集3维模型分类达到了较高水平。这类方法假定训练数据中的类别涵盖应用场景中的所有类别, 因而只能分类训练集中出现过的类别样本, 无法分类未出现的类别样本。然而现实应用中, 新的类别不断涌现, 这种分类方法受到了严重局限。一方面, 这种方法需要不停地收集大量新类别数据进行人工标注后重新训练分类器, 这将带来巨大的人力物力消耗; 另一方面, 在一些领域, 新类别的数据非常稀少, 难以搜集到足够的数据用于标注, 这严重制约着分类性能的进一步提升^[1]。针对这些问题, 研究者们提出了零样本学习 (Zero-Shot Learning, ZSL)^[2,3], 通过使用可见类样本的视觉特征和语义表示(比如属性、类别)训练模型, 然后将不可见类的语义表示作为桥梁, 使模型具有识别不可见类样本的能力^[4]。零样本学习可以极大地缓解分类任务中样本获取困难、模型需要重新训练和人工标注成本高的问题, 对分类技术落地应用起到极大的促进作用。

目前, 零样本学习已经成功应用于图像识别^[5]、自然语言处理^[6]等领域, 并且不断扩展和改进。例如, 基于生成式网络的方法^[7-9], 可以更好地模拟可见类和不可见类之间的数据分布差异; 同时, 基于多模态信息学习的方法^[10-12]可以将不同类型信息结合起来提高零样本分类的鲁棒性。

基于对比学习的图像-文本预训练模型 (Contrastive Language-Image Pre-training, CLIP)^[10]通过比较图像和文本之间的相似度学习对应关系。该模型由两个分支组成, 一个用于处理图像输入, 另一个用于处理文本输入。首先, 图像和文本被编码为向量, 并在共享嵌入空间中对齐。然后, 使用对比损失函数来最小化同类样本对之间的距离并最大化不同类样本对之间的距离, 以学习有意义的嵌入表示。这种预训练方法可以用于视觉-文本检索^[13]、视觉问答^[14,15]和图像分割^[16-18]等任务, 能够提高模型的性能和泛化能力。然而, 由于图像和文本在表达方式、特征空间、语义层面等多个方面存在很大差异。因此, 图像和文本之间存在着巨大的模态鸿沟^[19], 将它们进行有效地对齐是一个非常具有挑战性的问题。

针对以上问题, 本文提出语义增强CLIP实现图像和文本对齐的方法, 使它们更加具有语义一致性。本文将视图、文本和语义知识编码为向量, 并在共享空间中使视图和文本分别通过语义知识对齐, 最终达到视图和文本精准对齐的目的。这种方法可以缩小模态间的语义鸿沟, 提高图像和文本之间的对齐效果。

语义知识是指对不同模态下同一事物的语义描述, 其可以同时接近图像域和文本域, 为图像和文本提供更多相似的语义信息。其核心优势在于能够将文本描述的语义信息与图像中的视觉信息精准对齐, 从而使CLIP更好地理解图像内容。如图1所示, 为了充分发挥语义知识的优势, 本文将语义描述性文本作为语义知识文本。其通过文本的方式描述图像中的视觉信息, 使图像中丰富的语义信息拥有了强大的可解释性, 提高了模型对未知类别图像的识别能力。

具体来说, 首先将3维模型用多张视图表示, 然后通过视觉语言生成模型获得包含对每张视图及其类别详细信息的语义描述性文本。之后借助CLIP实现视图与类别提示文本的预对齐, 然后微调语义编码器对齐视图与语义描述性文本, 最终实现类别提示文本与3维模型的精准对齐。语义编码器和文本编码器共享参数, 微调语义编码器的过程把语义描述性文本具化为对3维模型类别的语义描述, 增强了语义描述性文本与3维模型各局部对齐的能力, 减小了类别提示文本与3维模型之间的语义鸿沟。

本文的主要贡献为: (1)提出一种语义增强CLIP的零样本3维模型分类方法, 通过微调语义编码器增强了文本与3维模型各局部对齐的能力, 显著提高了3维模型的零样本分类性能; (2)提出使用视觉语言生成模型生成语义描述性文本的方法, 分别通过图像字幕和视觉问答两种方式, 为视图中丰富的语义知识提供了强大的可解释性; (3)本文通过在ModelNet10^[20]和ModelNet40^[20]数据集上进行实验, 证实了本文方法在零样本和少样本的3维模型分类任务上具有最优的分类结果。

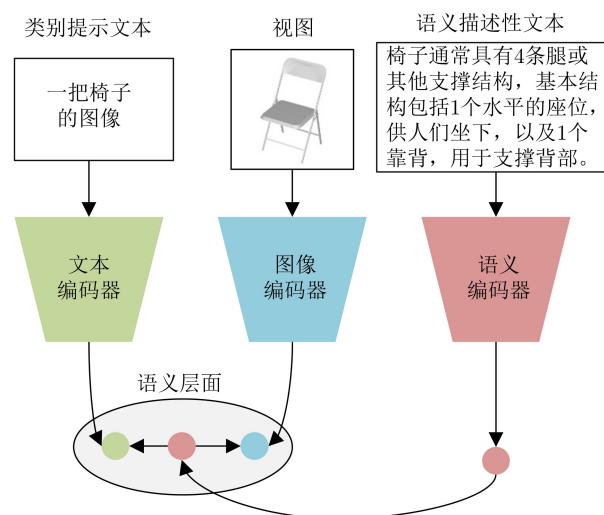


图1 语义增强CLIP的视图和类别提示文本对齐

2 相关工作

2.1 零样本3维模型分类

Cheraghian等人^[21-23]在零样本3维模型分类方面提出了3维点云的零样本学习方法、缓解3维零样本学习中枢纽点问题的方法和基于直推式零样本学习的3维点云分类方法,并将它们封装进一个全新的零样本3维点云方法^[24]中。以上方法均是利用已知类样本的点云表征及其词向量对未知类别进行分类,开创了零样本3维模型分类方法。近年来,CLIP在零样本图像分类上取得了良好的效果,因此有研究者将CLIP应用到零样本3维模型分类方法中,Zhang等人^[25]提出了基于CLIP的3维点云理解(Point cloud understanding by CLIP, PointCLIP)模型,PointCLIP首先将3维点云投影成多个深度图,然后利用CLIP的预训练图像编码器提取深度图特征,同时将类别名称通过CLIP预先训练的文本编码器提取文本特征。但是PointCLIP的性能受到深度图和图像之间的域差异以及深度分布的多样性限制。为了解决这一问题,基于图像-深度图预训练CLIP的点云分类方法(transfer CLIP to Point cloud classification with image-depth pre-training, CLIP2Point)^[26]将跨模态学习与模态内学习相结合训练了一个深度图编码器。在分类时,冻结CLIP的图像编码器,使用深度图编码器提取深度图特征,该方法缓解了深度图和图像间的模型差异。用于3维理解的图像-文本-点云一致性表征学习方法(learning Unified representation of Language, Image and Point cloud for 3D understanding, ULIP)^[27]构建了一个图像、文本和点云3种模态的统一嵌入空间,该方法利用大规模图像-文本对预训练的视觉语言模型,并将3维点云编码器的特征空间与预先对齐的视觉-文本特征空间对齐,大幅提高了3维模型的识别能力。与之相似的是,基于提示文本微调的3维识别方法(CLIP Goes 3D, CG3D)^[28]同样使用3元组形式确保同一类别的3维模型特征和图像特征之间以及3维模型特征和文本特征之间存在相似性,从而使点云编码器获得零样本识别的能力。另外,PointCLIP V2^[29]在PointCLIP的基础之上,通过利用更先进的投影算法和更详细的3维模型描述,显著提高了零样本3维模型分类准确率。本文采用语义增强CLIP解决图像和文本的语义鸿沟问题,通过在语义层面为图像和文本提供更多相似的语义信息,使图像和文本对齐更具有一致性,从而有效提高3维模型的零样本分类性能。

2.2 提示工程

生成式预训练Transformer模型(Generative

Pre-trained Transformer, GPT-3)^[30]通过利用自然语言提示显著提升了少样本学习的性能,催生了提示工程(prompt engineering)^[31]领域的研究。该领域寻求在给定任务和语言模型的情况下构建最佳提示,现有方法通常手动构建或利用反向传播学习得到^[32]。CLIP是手动构建提示模版的代表性工作,通过使用提示模版“A photo of a {label}.”将预先训练好的多模态知识迁移到下游的零样本图像预测任务中,展现了卓越的图像理解能力。在此基础上,Zhang等人^[33]提出了从图像和文本中进行视觉表征对比学习方法(Contrastive Visual Representation learning from Text, ConVIRT),通过无监督对比学习最大化图像与文本对之间的一致性来提高视觉表示。然而,手动构建的提示文本往往缺乏充足的语义信息,因此有研究人员使用反向传播学习最佳提示^[34-36]。上下文优化方法(Context Optimization, CoOp)^[37]是一种简单的、基于可学习提示文本的方法,通过使用可学习向量对提示的上下文单词进行建模,同时保持整个预训练参数不变,可以使CLIP适应下游图像识别任务。但是CoOp存在一个问题,即所学的上下文无法推广到更广泛的未见类别,因此Zhou等人^[38]提出了条件上下文优化方法(Conditional Context Optimization, Co-CoOp)通过进一步学习轻量级神经网络来为每个图像生成一种输入条件令牌,从而产生对类别偏移不太敏感的动态提示。然而,这种在连续空间中优化的方法缺乏可解释性,导致在零样本等任务上泛化性能较差。对此,本文采用视觉语言生成模型获取图像的语义描述性文本,既拥有充足的语义信息,也增加了语义信息的可解释性。

3 本文方法

如图2所示,本文所提语义增强CLIP的零样本3维模型分类方法步骤如下:(1)3维模型的多视图表示;(2)语义描述性文本获取;(3)训练语义编码器;(4)零样本3维模型分类。本文首先根据设置好的虚拟相机为每个3维模型投影出12张视图,然后如图2(a)所示,通过视觉语言生成模型一个用于视觉语言理解和生成的统一多模态预训练框架(unified multi-modal Pre-training framework for both vision-Language Understanding and Generation, mPLUG)^[39]获得包含对每张视图及其类别详细信息的语义描述性文本,该文本将作为微调语义编码器的输入,把视图中的视觉信息与类别提示文本描述的语义信息结合起来,从而让CLIP学会更好地理解视图的内容。训练阶段如图2(b)所示,首先借助CLIP实现视图与类别提示文本的预对齐,然后冻

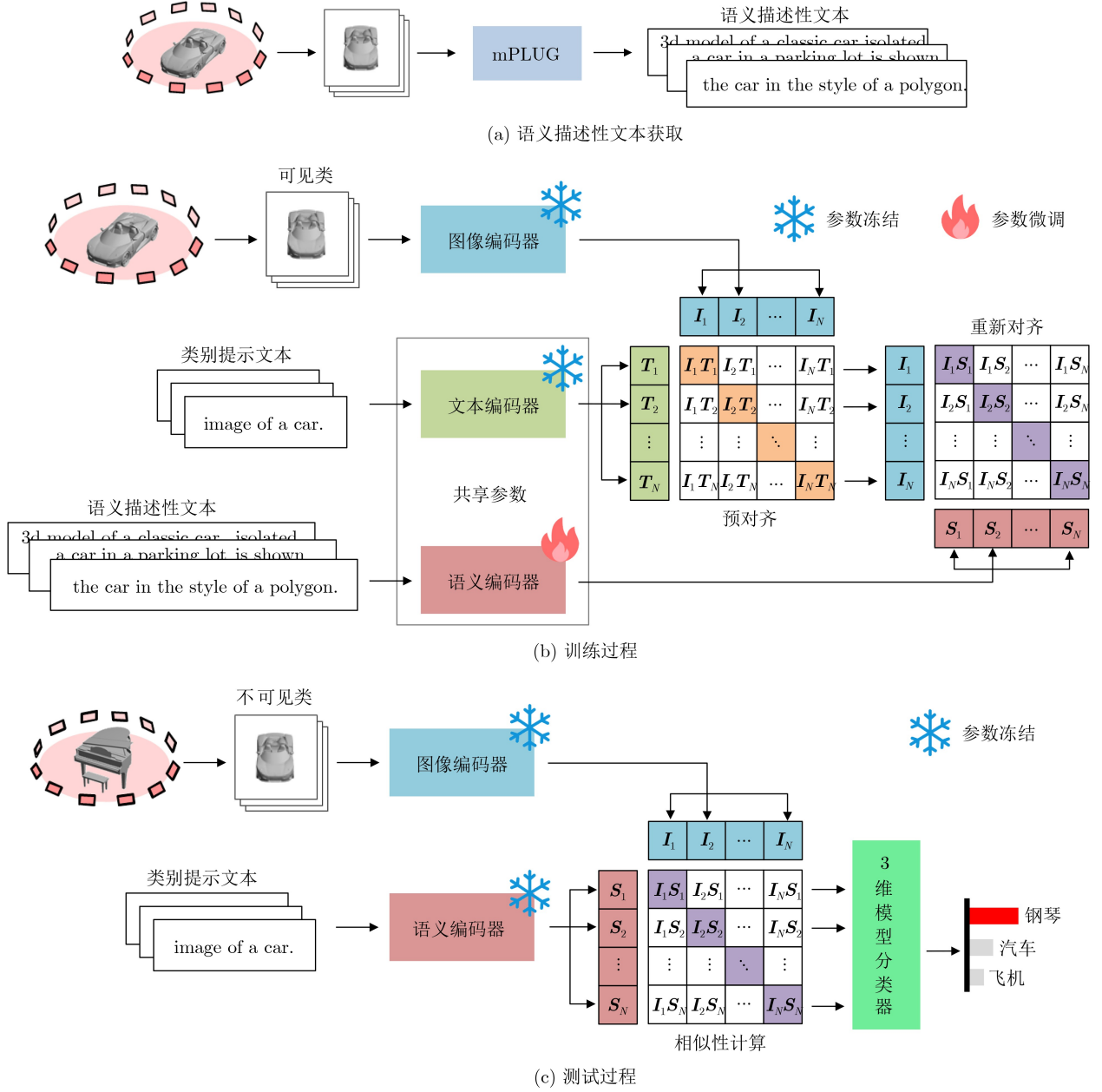


图2 总体框架

结图像编码器，接下来输入语义描述性文本训练语义编码器，使语义描述性文本具化为对3维模型类别的语义描述，从而实现视图与类别提示文本的重新对齐。分类阶段如图2(c)所示，输入不可见类3维模型的多视图和所有类别的提示文本，计算视图特征与语义特征的相似性，最后通过3维模型分类器获得零样本3维模型分类结果。

3.1 CLIP

CLIP不依赖于预先确定的对象类别，利用文本进行计算机视觉训练。CLIP通过在 4×10^8 个图像-文本对的大规模数据集上预训练，能够将图像与开放词汇表中的任何语义概念对齐，以进行零样本分类^[25]。

具体地，CLIP由两个分支组成，一个是图像编码器，另一个是文本编码器。其中，图像编码器有两种架构，分别是被广泛采用的残差神经网络(Residual neural Network, ResNet)^[40]和最新被提出的视觉Transformer(Vision Transformer, ViT)^[41]；文本编码器是一个拥有63M参数的12层Transformer^[42]。训练时，给定 N 个图像-文本对，CLIP通过联合图像编码器和文本编码器最大化 N 个对齐的图像和文本的余弦相似性，同时最小化 $N^2 - N$ 个未对齐的图像和文本的余弦相似性。预测过程可以表示为

$$F_I = \text{ImageEncoder}(I) \quad (1)$$

$$\mathbf{F}_T = \text{TextEncoder}(\mathbf{T}) \quad (2)$$

$$\mathbf{E}_I = L_2\text{Norm}(\mathbf{F}_I) \quad (3)$$

$$\mathbf{E}_T = L_2\text{Norm}(\mathbf{F}_T) \quad (4)$$

$$\text{Logits} = \mathbf{E}_I \mathbf{E}_T^T \quad (5)$$

其中, $\mathbf{I}$ 和 $\mathbf{T}$ 分别表示预对齐的图像和文本, $\mathbf{F}_I$ 和 $\mathbf{F}_T$ 分别表示编码后的图像和文本特征, $\mathbf{E}_I$ 和 $\mathbf{E}_T$ 分别表示 L_2 归一化后的图像和文本特征, Logits表示图像和文本的预测概率。

3.2 语义增强CLIP

区别于传统的有监督方法, CLIP是从图像和原生文本中获取广泛的监督信息。受此启发, 本文将3维模型的每张视图及其类别详细信息的语义描述性文本视为语义增强文本, 然后将其作为额外的监督信息获得图像中更有意义的嵌入表示。首先通过固定位置的相机得到3维模型的多视图表示, 然后通过视觉语言生成模型mPLUG获取每张视图的语义描述性文本, 最后用这些语义描述性文本微调语义编码器。

3.2.1 3维模型的多视图表示

多视图卷积神经网络 (Multi-View Convolutional Neural Network, MVCNN)^[43]作为基于多视图的3维模型分类方法具有里程碑意义。本文中与MVCNN使用的投影方法和渲染方法相同, 即将3维模型映射为12张视图。具体地, 将12个虚拟相机在北纬30°位置围绕3维模型均匀放置, 最终得到12张大小为224×224的视图。

3.2.2 语义描述性文本

PointCLIP采用手动制作的提示模版作为文本编码器的输入, 该编码器缺乏对投影图像和类别信息的特定描述^[29]。PointCLIP V2通过大规模语言模型GPT-3生成的文本提示作为编码器输入, 然而文本提示生成只能接收文本命令, 编码器仍然缺乏来自投影图像的特定描述。因此, 考虑到视觉语言生成模型特有的图像理解优势, 本文利用mPLUG为CLIP生成具有充足语义知识的文本提示。mPLUG是一种具备高效跨模态理解能力的视觉语言模型, 它在图像字幕和视觉问答等下游任务中具有最先进的性能。因此, 本文将3维模型投影得到的每张视图作为mPLUG的输入获取语义描述性文本。本文使用两种方式获取对视图及其类别详细信息的语义描述性文本:

(1)图像字幕(image captioning): 给定一张图像, mPLUG为其生成合适且流畅的字幕文本。例如: 输入图3(a); 输出“Aerial view of a military aircraft isolated on white background.”。

(2)视觉问答(visual question answering): 给定一张图像并提出问题, mPLUG为其生成问题答案。例如: 输入图3(b)并提问“What is the {person} gray image looks like?”; 输出“Construction worker”。

3.2.3 语义编码器训练

在CLIP中, Radford等人^[10]注意到通过为每个任务定制提示文本, 可以显著提高零样本任务的性能。对此, 本文为3维模型的每张视图定制了12条提示文本, 并将每条提示文本与类别标签和通过视觉语言生成模型生成的文本组合成语义描述性文本, 然后输入语义编码器进行训练。在训练语义编码器时, 每条语义描述性文本被分别输入语义编码器, 这有助于模型更好地理解每个文本的语义信息, 从而提高整体性能。通过使用具有详细语义内容的语义描述性文本微调语义编码器, 可以减小类别提示文本与3维模型之间的语义鸿沟, 实现类别提示文本与3维模型的精准对齐。本文为两种文本生成方式分别定制了12条提示文本:

(1)图像字幕(image captioning): “an image of a {CLASS}, {CAPTION}.”, “a gray photo of a {CLASS}, {CAPTION}.”, “a rendering of a {CLASS}, {CAPTION}.”, “art of a {CLASS}, {CAPTION}.”, “a model of a {CLASS}, {CAPTION}.”, “a projection of a {CLASS}, {CAPTION}.”, “a view of a {CLASS}, {CAPTION}.”, “a drawing of a {CLASS}, {CAPTION}.”, “an oblique drawing of a {CLASS}, {CAPTION}.”, “a side view of a {CLASS}, {CAPTION}.”, “a 3D shape of a {CLASS}, {CAPTION}.”, “an object of a {CLASS}, {CAPTION}.”。

(2)视觉问答(visual question answering): “an image of a {CLASS} like a {ANSWER}.”, “a gray photo of a {CLASS} like a {ANSWER}.”, “a rendering of a {CLASS} like a {ANSWER}.”, “art of a {CLASS} like a {ANSWER}.”, “a model of a {CLASS} like a {ANSWER}.”, “a projection of a {CLASS} like a {ANSWER}.”, “a view of a {CLASS} like a {ANSWER}.”。



(a) 图像字幕输入示例



(b) 视觉问答输入示例

图 3 输入图像

{ANSWER}.”, “a drawing of a {CLASS} like a {ANSWER}.”, “an oblique drawing of a {CLASS} like a {ANSWER}.”, “a side view of a {CLASS} like a {ANSWER}.”, “a 3D shape of a {CLASS} like a {ANSWER}.”, “an object of a {CLASS} like a {ANSWER}.”。

其中, 将类别标签放置在“{CLASS}”位置, 将图像字幕生成的字幕文本放置在“{CAPTION}”位置, 将视觉问答生成的答案文本放置在“{ANSWER}”位置。

3.3 基于投票策略的零样本3维模型分类

在预测时, CLIP通过余弦相似性计算每张图像与文本的预测概率。对于由12张视图表示的3维模型, 本文融合来自视图与定制类别提示文本“image of a {CLASS}.”之间的余弦相似性, 实现3维模型分类。本文采用投票的策略进行分类, 以概率投票法^[44]零样本3维模型分类准确率作为验证指标。3维模型分类结果如下。

$$u = \underset{i=1}{\operatorname{argmax}}^k T_{\text{vote}}(\omega_i | M) \rightarrow M \in \omega_u \quad (6)$$

$$T_{\text{vote}}(\omega_i | M) = \sum_{l=1}^n p_{il} \quad (7)$$

其中, M 表示3维模型, ω_i 表示类别 i , k 为类别数目, p_{il} 表示视图 v_l 属于某个类别 i 的概率, $T_{\text{vote}}(\omega_i | M)$ 是类别 i 所获视图投票值, n 为单个3维模型的视图数量。

4 实验

本节首先介绍所提出方法的实施细节, 然后通过消融实验证明方法的有效性, 最后展示本文方法在零样本3维模型分类任务上的性能。

4.1 实施细节

实验在ModelNet10和ModelNet40数据集上进行评估。ModelNet10由10个类的4 899个3维模型组成, ModelNet40由40个类的12 311个3维模型组成。在本文中, 每个3维模型被投影为12张视图, 每张视图是大小为 224×224 的灰度图像。因此, ModelNet10由58 788($4\ 899 \times 12$)张视图组成, ModelNet40由147 732($12\ 311 \times 12$)张视图组成。首先将3维模型投影出12张视图, 然后通过视觉语言生成模型mPLUG为每张视图生成语义描述性文本。接下来冻结图像编码器, 将生成的语义描述性文本在语义编码器上微调20个回合, 其中语义编码器与文本编码器共享参数, 使用Adam优化器, 学习率和权重衰减率分别为 10^{-7} 和 10^{-4} 。在此期间, 本文测试了不同图像编码器引导语义编码器的训

练, 当ViT-B\16作为图像编码器时获得了最佳性能。测试时, 计算每张视图与定制文本之间的余弦相似性, 视觉特征来自图像编码器, 文本特征来自语义编码器。不同的定制文本所得到的分类准确率不同, 本文测试并分析了在常用定制文本下的分类准确率。

4.2 消融实验

在表1中, 使用“image of a {CLASS}.”作为测试的类别提示文本, 并对所提出的方法进行消融实验, 涉及来自两种方式生成的语义描述性文本的实验结果。如表1所示, 当使用CLIP在ModelNet40上进行零样本视图分类时, 准确率为52.34%, 如果使用图像字幕或视觉问答生成方法生成的语义描述性文本微调语义编码器, 零样本视图分类性能分别提高了7.64%和10.60%。同样地, 通过概率投票法在ModelNet40上使用CLIP进行零样本3维模型分类时的准确率为51.62%, 使用图像字幕生成的语义描述性文本可以将准确率提升至67.59%, 使用视觉问答生成方法生成的语义描述性文本可以达到最佳性能, 此时分类准确率为69.73%。以上结果表明通过语义描述性文本微调语义编码器, 可以有效地提高类别提示文本与3维模型的各局部重新对齐的效果, 从而提升3维模型的零样本分类性能。

4.3 分类准确率对比

在表2中, 将本文提出的零样本3维模型分类方法与其他方法进行比较。早期的一些方法需要围绕图像编码器进行额外的工作, 如PointCLIP和CLIP2Point, 为了适应3维点云的深度图, 均使用可见类样本微调了图像编码器; 后来, PointCLIP V2通过使用将点云体素化再映射的方法, 本质是通过提升图像质量使其接近图像编码器的域, 该方法虽然无需对图像编码器进行微调, 但映射前的工

表 1 在ModelNet40上的消融实验(%)

方法	视图	3维模型
CLIP	52.34	51.62
CLIP+图像字幕	59.98	67.59
CLIP+视觉问答	62.94	69.73

表 2 在ModelNet10和ModelNet40数据集上零样本分类准确率(%)

方法	ModelNet10	ModelNet40
CLIP2Point	66.63	49.38
PointCLIP	30.23	23.78
PointCLIP V2	73.13	64.22
CLIP+图像字幕	79.63	67.59
CLIP+视觉问答	79.74	69.73

作十分复杂。本方法直接使用视图进行零样本分类,大大降低了工作的复杂性,而且分类准确率优于现有方法。本文使用“image of a {CLASS}.”作为本次测试的类别提示文本,通过仅使用图像字幕或视觉问答生成方法生成的语义描述性文本微调语义编码器,在ModelNet10数据集上进行零样本3维模型分类时准确率分别达到了79.63%和79.74%;在ModelNet40数据集上使用CLIP和图像字幕生成方法时,零样本分类准确率为67.59%。当使用CLIP和视觉问答生成方法时,本文方法的零样本分类准确率为69.73%,相较于当前最卓越的方法PointCLIP V2准确率提高了5.51%,证明了本文方法的先进性。

在少样本实验中,本文同样使用类别提示文本“image of a {CLASS}.”用于测试,并对比训练样本数量为2-shot的PointNet^[45], PointNet++^[46], SimpleView^[47], CurveNet^[48]和PointCLIP等方法。首先,选取ModelNet40训练集中的前2个样本用于少样本训练。然后冻结语义编码器后微调图像编码器,训练50个回合,实验结果如表3所示。从表3中可以看出,本文所提出的少样本分类方法优于现有方法。具体地,通过使用图像字幕或视觉问答生成方法微调的语义编码器引导图像编码器的训练,本文方法在ModelNet10数据集上的少样本3维模型分类准确率分别达到了86.01%和86.89%;在ModelNet40数据集上使用图像字幕生成方法时少样本分类准确率为76.74%,相较于最先进的PointCLIP方法提升了8.43%,进一步使用视觉问答生成方法时少样本分类准确率高达78.48%,达到了最优性能。

表3 在ModelNet10和ModelNet40上数据集上2-shot
分类准确率(%)

方法	ModelNet10	ModelNet40
PointNet	—	33.10
SimpleView	—	36.43
CurveNet	—	56.56
PointNet++	—	56.93
PointCLIP	—	68.31
CLIP+图像字幕	86.01	76.74
CLIP+视觉问答	86.89	78.48

表4 图像字幕生成方法在类别提示文本上的准确率(%)

类别提示文本	ModelNet10	ModelNet40
“image of a {CLASS}.”	79.63	67.59
“a view of a {CLASS}.”	82.60	67.95
“a 3D shape view of a {CLASS}.”	83.15	70.91
“a 3D shape of a {CLASS}.”	83.15	70.66

表4和表5分别展示了图像字幕生成方法和视觉问答生成方法在4种定制类别提示文本下的测试准确率。4种定制类别提示文本来源于常用语句模版与常用词汇的组合。在ModelNet40数据集上,当使用“image of a {CLASS}.”作为类别提示文本时,视觉问答生成方法下的零样本3维模型分类准确率为69.73%,将“image”替换为“a view”后提升至70.91%。由于投影的视图来自3维模型,因此进一步添加“3D shape”可以将分类准确率提升至72.57%。以上趋势在图像字幕生成方法下也同样存在。但是,如果删除“a 3D shape view of a {CLASS}.”中的“view”一词,图像字幕生成方法下的准确率受到了影响,而在视觉问答生成方法下准确率可以进一步提高至72.89%。这是因为通过图像字幕生成的语义描述性文本未将注意力完全聚焦在图像内容中,而是额外关注了图像内容以外的其他内容;而通过视觉问答生成的语义描述性文本受到问题的引导会聚焦关注图像内容,因此“view”一词的出现带来了负面作用,所以删除“view”会提高分类准确率。

表6展示了利用不同主干网络的图像编码器取得的分类准确率。不同的图像编码器会对语义编码器的训练产生不同的引导效果,进而影响分类准确率。本方法将ViT-B\16作为图像编码器的主干网络实现了在ModelNet40数据集上最高的分类准确率,达到69.73%。该准确率相较于ResNet作为图像编码器的主干网络时有较大提升,这是因为ViT在大规模图像数据集上能更好地捕捉到图像特征,并且使用自然语言文本作为监督信号时,可以更好地实现图像与文本的匹配。另外,如表6所示,尽管使用不同的图像编码器产生了不同的分类准确率,但使用视觉问答生成方法时的分类准确率普遍高于使用图像字幕生成方法时的分类准确率,这进一步证明了视觉问答生成方法的优越性和有效性。

表5 视觉问答生成方法在类别提示文本上的准确率(%)

类别提示文本	ModelNet10	ModelNet40
“image of a {CLASS}.”	79.74	69.73
“a view of a {CLASS}.”	80.73	70.91
“a 3D shape view of a {CLASS}.”	81.28	72.57
“a 3D shape of a {CLASS}.”	83.15	72.89

表6 不同的图像编码器上的准确率(%)

方法	RN50	RN101	ViT-B\16	ViT-B\32
CLIP+图像字幕	26.10	28.90	67.59	60.05
CLIP+视觉问答	28.89	30.33	69.73	65.80

4.4 讨论

实验结果显示, 通过使用图像字幕和视觉问答生成的语义描述性文本微调语义编码器, 可以有效地提高类别提示文本与3维模型的各局部重新对齐的效果, 从而提升零样本3维模型分类准确率。实

验结果表明视觉问答的效果优于图像字幕, 这是因为视觉问答基于问题引导, 能够回答出具有丰富语义知识的答案, 而图像字幕的生成受到图像内容的限制, 无法提供详细的语义信息。不同3维模型的文本提示生成结果示例如表7所示。

表 7 生成的文本提示示例

						
图像字幕	flying	goblet	plastic	straight line	tin can	wooden chair
视觉问答	3D model of a fighter jet isolated on white background	a tall glass of milk on a white background	a white trash can with a lid on a white background	a long line of gray stairs on a white background	a stack of metal cans with plants growing out of them	a 3D model of a beach chair

5 结论

本文提出一种语义增强CLIP的零样本3维模型分类方法, 本方法首先将3维模型表示成多视图, 并通过视觉语言生成模型生成视图的语义描述性文本; 然后借助CLIP实现视图与类别提示文本的预对齐; 最后输入语义描述性文本微调语义编码器, 目的是把语义描述性文本具化为对3维模型类别的语义描述, 从而减小类别提示文本与3维模型之间的语义鸿沟, 提高CLIP对3维模型语义知识的理解能力。本文提出的方法有效提高了零样本3维模型分类的准确率, 同时该方法还适用于其他的视觉-文本领域, 具有广泛的应用前景。

参 考 文 献

- [1] 赵鹏, 汪纯燕, 张思颖, 等. 一种基于融合重构的子空间学习的零样本图像分类方法[J]. 计算机学报, 2021, 44(2): 409–421. doi: [10.11897/SP.J.1016.2021.00409](#).
ZHAO Peng, WANG Chunyan, ZHANG Siying, et al. A zero-shot image classification method based on subspace learning with the fusion of reconstruction[J]. *Chinese Journal of Computers*, 2021, 44(2): 409–421. doi: [10.11897/SP.J.1016.2021.00409](#).
- [2] FU Yanwei, XIANG Tao, JIANG Yugang, et al. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content[J]. *IEEE Signal Processing Magazine*, 2018, 35(1): 112–125. doi: [10.1109/MSP.2017.2763441](#).
- [3] ZHAI Xiaohua, WANG Xiao, MUSTAFA B, et al. LiT: Zero-shot transfer with locked-image text tuning[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 18102–18112. doi: [10.1109/CVPR52688.2022.01759](#).
- [4] JI Zhong, YAN Jiangtao, WANG Qiang, et al. Triple discriminator generative adversarial network for zero-shot image classification[J]. *Science China Information Sciences*, 2021, 64(2): 120101. doi: [10.1007/s11432-020-3032-8](#).
- [5] LIU Yang, ZHOU Lei, BAI Xiao, et al. Goal-oriented gaze estimation for zero-shot learning[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 3793–3802. doi: [10.1109/CVPR46437.2021.00379](#).
- [6] CHOI H, KIM J, JOE S, et al. Analyzing zero-shot cross-lingual transfer in supervised NLP tasks[C]. The 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021: 9608–9613. doi: [10.1109/ICPR48806.2021.9412570](#).
- [7] GAO Rui, HOU Xingsong, QIN Jie, et al. Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning[J]. *IEEE Transactions on Image Processing*, 2020, 29: 3665–3680. doi: [10.1109/TIP.2020.2964429](#).
- [8] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]. The 38th International Conference on Machine Learning, 2021: 8821–8831.
- [9] NICHOL A Q, DHARIWAL P, RAMESH A, et al. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models[C]. The 39th International Conference on Machine Learning, Baltimore, USA, 2022: 16784–16804.
- [10] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]. The 38th International Conference on Machine Learning, 2021: 8748–8763.
- [11] CHENG Ruizhe, WU Bichen, ZHANG Peizhao, et al. Data-efficient language-supervised zero-shot learning with self-distillation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Nashville, USA, 2021: 3119–3124. doi: [10.1109/CVPRW53098.2021.00348](#).

- [12] Doshi K, Garg A, Uzcent B, *et al.* A multimodal benchmark and improved architecture for zero shot learning[C]. The IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 2021–2030.
- [13] LUO Huaishao, JI Lei, ZHONG Ming, *et al.* CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning[J]. *Neurocomputing*, 2022, 508: 293–304. doi: [10.1016/j.neucom.2022.07.028](https://doi.org/10.1016/j.neucom.2022.07.028).
- [14] CHEN Long, ZHENG Yuhang, and XIAO Jun. Rethinking data augmentation for robust visual question answering[C]. The 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 95–112. doi: [10.1007/978-3-031-20059-5_6](https://doi.org/10.1007/978-3-031-20059-5_6).
- [15] CHO J, YOON S, KALE A, *et al.* Fine-grained image captioning with CLIP reward[C]. Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, United States, 2022: 517–527. doi: [10.18653/v1/2022.findings-naacl.39](https://doi.org/10.18653/v1/2022.findings-naacl.39).
- [16] LIANG Feng, WU Bichen, DAI Xiaoliang, *et al.* Open-vocabulary semantic segmentation with mask-adapted clip[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 7061–7070. doi: [10.1109/CVPR52729.2023.00682](https://doi.org/10.1109/CVPR52729.2023.00682).
- [17] XIE Jinheng, HOU Xianxu, YE Kai, *et al.* CLIMS: Cross language image matching for weakly supervised semantic segmentation[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 4473–4482. doi: [10.1109/CVPR52688.2022.00444](https://doi.org/10.1109/CVPR52688.2022.00444).
- [18] ZHOU Chong, LOY C C, and DAI Bo. Extract free dense labels from CLIP[C]. The 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 696–712. doi: [10.1007/978-3-031-19815-1_40](https://doi.org/10.1007/978-3-031-19815-1_40).
- [19] WANG Fengyun, PAN Jinshan, XU Shoukun, *et al.* Learning discriminative cross-modality features for RGB-D saliency detection[J]. *IEEE Transactions on Image Processing*, 2022, 31: 1285–1297. doi: [10.1109/TIP.2022.3140606](https://doi.org/10.1109/TIP.2022.3140606).
- [20] WU Zhirong, SONG Shuran, KHOSLA A, *et al.* 3D shapeNets: A deep representation for volumetric shapes[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1912–1920. doi: [10.1109/CVPR.2015.7298801](https://doi.org/10.1109/CVPR.2015.7298801).
- [21] CHERAGHIAN A, RAHMAN S, and PETERSSON L. Zero-shot learning of 3D point cloud objects[C]. The 16th International Conference on Machine Vision Applications, Tokyo, Japan, 2019. DOI: [10.23919/MVA.2019.8758063](https://doi.org/10.23919/MVA.2019.8758063).
- [22] CHERAGHIAN A, RAHMAN S, CAMPBELL D, *et al.* Mitigating the hubness problem for zero-shot learning of 3D objects[C]. The 30th British Machine Vision Conference, Cardiff, UK, 2019. doi: [10.48550/arXiv.1907.06371](https://doi.org/10.48550/arXiv.1907.06371).
- [23] CHERAGHIAN A, RAHMAN S, CAMPBELL D, *et al.* Transductive zero-shot learning for 3D point cloud classification[C]. 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass, USA, 2020: 912–922. doi: [10.1109/WACV45572.2020.9093545](https://doi.org/10.1109/WACV45572.2020.9093545).
- [24] CHERAGHIAN A, RAHMAN S, CHOWDHURY T F, *et al.* Zero-shot learning on 3D point cloud objects and beyond[J]. *International Journal of Computer Vision*, 2022, 130(10): 2364–2384. doi: [10.1007/s11263-022-01650-4](https://doi.org/10.1007/s11263-022-01650-4).
- [25] ZHANG Renrui, GUO Ziyu, ZHANG Wei, *et al.* PointCLIP: Point cloud understanding by CLIP[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 8552–8562. doi: [10.1109/CVPR52688.2022.00836](https://doi.org/10.1109/CVPR52688.2022.00836).
- [26] HUANG Tianyu, DONG Bowen, YANG Yunhan, *et al.* CLIP2Point: Transfer CLIP to point cloud classification with image-depth pre-training[C]. 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 2023. doi: [10.1109/ICCV51070.2023.02025](https://doi.org/10.1109/ICCV51070.2023.02025).
- [27] XUE Le, GAO Mingfei, XING Chen, *et al.* ULIP: Learning unified representation of language, image and point cloud for 3D understanding[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023. doi: [10.1109/CVPR52729.2023.00120](https://doi.org/10.1109/CVPR52729.2023.00120).
- [28] HEGDE D, VALANARASU J M J, and PATEL V M. CLIP goes 3D: Leveraging prompt tuning for language grounded 3D recognition[C]. 2023 IEEE/CVF International Conference on Computer Vision Workshops, Paris, France, 2023. doi: [10.1109/ICCVW60793.2023.00217](https://doi.org/10.1109/ICCVW60793.2023.00217).
- [29] ZHU Xiangyang, ZHANG Renrui, HE Bowei, *et al.* PointCLIP V2: Prompting CLIP and GPT for powerful 3D open-world learning[J]. arXiv: 2211.11682, 2022. doi: [10.48550/arXiv.2211.11682](https://doi.org/10.48550/arXiv.2211.11682).
- [30] BROWN T B, MANN B, RYDER N, *et al.* Language models are few-shot learners[C]. The 34th International Conference on Neural Information Processing Systems, Vancouver, Canada, 2020.
- [31] GAO Tianyu, FISCH A, and CHEN Danqi. Making pre-trained language models better few-shot learners[C]. The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 3816–3830. doi: [10.18653/v1/2021.acl-long.295](https://doi.org/10.18653/v1/2021.acl-long.295).
- [32] SORENSEN T, ROBINSON J, RYTTEING C, *et al.* An information-theoretic approach to prompt engineering without ground truth labels[C]. The 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 2022: 819–862. doi: [10.18653/v1/2022.acl-long.60](https://doi.org/10.18653/v1/2022.acl-long.60).
- [33] ZHANG Yuhao, JIANG Hang, MIURA Y, *et al.* Contrastive learning of medical visual representations from paired images and text[C]. 2022 Machine Learning for

- Healthcare Conference, Durham, USA, 2022: 2–25. doi: [10.48550/arXiv.2010.00747](https://doi.org/10.48550/arXiv.2010.00747).
- [34] LESTER B, AL-ROUF R, and CONSTANT N. The power of scale for parameter-efficient prompt tuning[C]. The 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 2021: 3045–3059. doi: [10.18653/v1/2021.emnlp-main.243](https://doi.org/10.18653/v1/2021.emnlp-main.243).
- [35] LI X L and LIANG P. Prefix-Tuning: Optimizing continuous prompts for generation[C]. The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021: 4582–4597. doi: [10.18653/v1/2021.acl-long.353](https://doi.org/10.18653/v1/2021.acl-long.353).
- [36] LIU Xiao, ZHENG Yanan, DU Zhengxiao, *et al.* GPT understands, too[J]. *AI Open*, 2023. doi: [10.1016/j.aiopen.2023.08.012](https://doi.org/10.1016/j.aiopen.2023.08.012).
- [37] ZHOU Kaiyang, YANG Jingkan, LOY C C, *et al.* Learning to prompt for vision-language models[J]. *International Journal of Computer Vision*, 2022, 130(9): 2337–2348. doi: [10.1007/s11263-022-01653-1](https://doi.org/10.1007/s11263-022-01653-1).
- [38] ZHOU Kaiyang, YANG Jingkan, LOY C C, *et al.* Conditional prompt learning for vision-language models[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 16816–16825. doi: [10.1109/CVPR52688.2022.01631](https://doi.org/10.1109/CVPR52688.2022.01631).
- [39] LI Chenliang, XU Haiyang, TIAN Junfeng, *et al.* mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections[C]. The 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 2022. doi: [10.18653/v1/2022.emnlp-main.488](https://doi.org/10.18653/v1/2022.emnlp-main.488).
- [40] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al.* Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [41] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale[C]. The 9th International Conference on Learning Representations, 2021.
- [42] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017. doi: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762).
- [43] SU Hang, MAJI S, KALOGERAKIS E, *et al.* Multi-view convolutional neural networks for 3D shape recognition[C]. 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 945–953. doi: [10.1109/ICCV.2015.114](https://doi.org/10.1109/ICCV.2015.114).
- [44] 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. doi: [10.3724/SP.J.1089.2019.17160](https://doi.org/10.3724/SP.J.1089.2019.17160).
BAI Jing, SI Qinglong, and QIN Feiwei. 3D model classification and retrieval based on CNN and voting scheme[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2019, 31(2): 303–314. doi: [10.3724/SP.J.1089.2019.17160](https://doi.org/10.3724/SP.J.1089.2019.17160).
- [45] QI CHARLES R, SU Hao, KAICHUN M, *et al.* PointNet: Deep learning on point sets for 3D classification and segmentation[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 652–660. doi: [10.1109/CVPR.2017.16](https://doi.org/10.1109/CVPR.2017.16).
- [46] QI CHARLES R, YI Li, SU Hao, *et al.* PointNet++: Deep hierarchical feature learning on point sets in a metric space[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017.
- [47] GOYAL A, LAW H, LIU Bowei, *et al.* Revisiting point cloud shape classification with a simple and effective baseline[C]. The 38th International Conference on Machine Learning, 2021: 3809–3820.
- [48] XIANG Tiange, ZHANG Chaoyi, SONG Yang, *et al.* Walk in the cloud: Learning curves for point clouds shape analysis[C]. 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 895–904. doi: [10.1109/ICCV48922.2021.00095](https://doi.org/10.1109/ICCV48922.2021.00095).
- 丁 博: 女, 博士, 副教授, 硕士生导师, CCF会员(H0875M), 研究领域为计算机图形学、计算机视觉。
张立宝: 男, 硕士生, 研究领域为计算机图形学、计算机视觉。
秦 健: 男, 博士生, 研究领域为计算机图形学、计算机视觉。
何勇军: 男, 博士, 教授, 博士生导师, 研究领域为计算机图形学、计算机视觉、人工智能。

责任编辑: 余 蓉