E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)05-1450-16

论文引用格式:Yan H, Bai J and Zheng H. 2025. Consistency constraint guided network for zero-shot 3D classification. Journal of Image and Graphics, 30(5):1450-1465(晏浩, 白静, 郑虎. 2025. 一致性约束引导的零样本三维模型分类网络. 中国图象图形学报, 30(5):1450-1465)[DOI:10.11834/jig.240397]

一致性约束引导的零样本三维模型分类网络

晏浩[1],白静[1,2*],郑虎[1]

1. 北方民族大学计算机科学与工程学院,银川　750021; 2. 国家民委图像图形智能处理实验室,银川　750021

摘　要:**目的**　零样本三维模型分类任务自提出起,始终面临大规模数据集与高质量语义信息的短缺问题。为应对这些问题,现有方法引入二维图像领域中蕴含丰富的数据集和语义信息的大规模预训练模型,这些方法基于语言—图像对比学习预训练网络,取得了一定的零样本分类效果。但是,现有方法对三维信息捕捉不全,无法充分利用来自三维领域的知识,针对这一问题,提出一致性约束引导的零样本三维模型分类网络。**方法**　一方面,在保留来自预训练网络中的全部二维知识的同时,通过视图一致性学习三维数据的特征,从视图层面将三维信息增补至二维视图特征中;另一方面,通过掩码一致性约束引导网络通过自监督增强网络对三维模型的整体性学习,提高网络泛化性能;同时,提出同类一致性约束引导的非互斥损失,确保网络在小规模数据集训练中学习方向的正确性与能力的泛化性。**结果**　在ZS3D(zero-shot for 3D dataset)、ModelNet10和Shrec2015(shape retrieval 2015)3个数据集上进行零样本分类,分别取得70.1%、57.8%和12.2%的分类精度,与当前最优方法相比,分别取得9.2%、22.8%和2.3%的性能提升;在ScanObjectNN的3个子集OBJ_ONLY(object only)、OBJ_BG(object and background)及PB_T50_RS(object augmented rot scale)上,本文方法也取得了具有竞争力的分类准确率,分别是32.4%、28.9%和19.3%。**结论**　相较于完全依赖预训练模型能力的方法,本文方法在充分利用语言—图像预训练网络的基础上,将三维模型领域的知识引入网络,并提升网络泛化能力,使零样本分类结果更加准确。

关键词:三维模型分类;零样本学习;自监督学习;图像文本预训练;视觉语言多模态

Consistency constraint guided network for zero-shot 3D classification

Yan Hao[1], Bai Jing[1,2*], Zheng Hu[1]

1. School of Computer Science and Engineering, North Minzu University, Yinchuan　750021, China;

2. The Key Laboratory of Images and Graphics Intelligent Processing of State Ethnic Affairs Commission, Yinchuan　750021, China

Abstract: **Objective** Deep learning has made remarkable progress in 3D model classification. However, most existing classification methods rely on supervised learning, which limits their capability to recognize only the model categories seen during training. With the development of computer-aided design and LiDAR sensor technologies, an increasing number of novel 3D model classes are emerging, presenting a challenge: how to effectively identify model classes that were not encountered during training. Zero-shot learning has been proposed to address this challenge. However, this approach faces a major limitation due to the shortage of large-scale datasets with high-quality semantic information. To overcome this

收稿日期:2024-09-29;**修回日期**:2024-10-16;**预印本日期**:2024-10-23

*通信作者:白静　baijing@nun.edu.cn

基金项目:国家自然科学基金项目(62162001);宁夏自然科学基金项目(2022AAC02041);宁夏优秀人才支持计划项目;北方民族大学研究生创新项目(YCX23160)

Supported by: National Natural Science Foundation of China (62162001); Natural Science Foundation of Ningxia, China (2022AAC02041); Ningxia Excellent Talent Program; Graduate Innovation Project of North Minzu University (YCX23160)

issue, many existing methods introduce large-scale pre-trained models with rich semantic information from 2D image domains, such as the contrastive language-image pre-training (CLIP) network. While these methods project 3D models to 2D space to meet the input requirements of CLIP visual encoder and achieve reasonable results, they do not fully capture the 3D information from the datasets and fail to leverage the knowledge inherent to the 3D domain. A straightforward approach to addressing this limitation is to adopt the learning strategy of multiview convolutional neural networks, which involves fine-tuning the CLIP visual encoder and optimizing its network parameters using a 3D model dataset. The goal is to leverage the advantages of 2D data annotation while incorporating the inherent characteristics of 3D models. However, this strategy does not yield effective results for CLIP. The fine-tuned network tends to overfit the training set, causing it to gradually forget much of the valuable 2D knowledge during the tuning process. Therefore, this strategy is not feasible. This paper proposes a consistency constraint guided network (CCG-Net) for zero-shot 3D model classification to overcome these problems. **Method** CCG-Net aims to leverage the strengths of 2D and 3D domains while mitigating the issues of overfitting and knowledge forgetting. CCG-Net comprises fixed and dynamic parts. The fixed part of the network employs a frozen CLIP model to learn cross-modal information from large-scale 2D visual and semantic data. Stopping the backpropagation in this part forces the network to focus on preserving 2D information. In contrast, the dynamic part is a learnable encoder designed for extracting global features from 3D models, with a strong emphasis on acquiring 3D knowledge. A view consistency constraint is applied in the dynamic part to guide the extraction of 3D features. This design ensures that the 2D knowledge from the pre-trained model is fully preserved, while also allowing the network to learn new information from 3D data. The information from two modalities is then effectively fused into comprehensive 3D model features, which are used for classification. Mask consistency constraints are introduced to enhance the extraction of features for 3D data and improve the robustness of the 3D encoding process. This constraint guides the network in enhancing its capability to learn the 3D model through self-supervised learning. The specific approach involves employing different masking methods to obtain a diverse set of mask features. Once these features are generated, the next step is to constrain their consistency. The network can effectively learn and integrate the essential characteristics from the masked data by ensuring the consistency of these mask features, finally enhancing model robustness and accuracy. Additionally, the pre-trained network employs a mutual exclusion loss, which assumes a mutual exclusion relationship between the labels to be classified. However, this network is unsuitable for the zero-shot task of tuning on a small-scale dataset. A non-mutual exclusion loss, guided by the homogeneity consistency constraints, is also proposed to address this issue, ensuring the accuracy of the learning direction and the network's capability to generalize its learning during training on a small-scale dataset. **Result** Three different consistency constraint schemes work collaboratively within the network to optimize its parameters, effectively preventing overfitting during fine-tuning on 3D data. This approach enhances the reliability and generalization of feature extraction, ultimately enhancing zero-shot classification performance. Quantitatively, on the ZS3D dataset, the proposed method achieves a classification accuracy of 70.1%, marking a substantial 9.2% improvement over the current best results, achieved by discriminative feature-guided zero-shot learning of 3D model classification (DFG-ZS3D). Additionally, this method demonstrates improvements on the dataset proposed by Cheraghian, achieving classification accuracies of 57.8%, 19.9%, and 12.2% on the ModelNet10, McGill, and Shrec 2015 subsets, respectively. These results correspond to improvements of 22.8%, 3.3%, and 2.3% over the state-of-the-art methods. The ScanObjectNN dataset, which comprises 3D models obtained from real-world scans rather than synthetic data, further validates the effectiveness of CCG-Net. On this dataset, CCG-Net attains the highest performance across its three subsets, with classification accuracies of 32.4%, 28.9%, and 19.3% on the OBJ_ONLY (Object only), OBJ_BG (Object and background), and PB_T50_RS (Object augmented rot scale) subsets, respectively. The performance improvement on real-world datasets further validates the generalization capability of the proposed method. Additionally, ablation experiments confirm the effectiveness of the three consistency constraints. Finally, qualitative analysis results of the confusion matrix demonstrate that the network can avoid overfitting to a certain extent. However, this analysis also reveals shortcomings in the capability of the network to extract discriminative features, providing a perspective for future research. **Conclusion** Compared to methods that rely solely on pre-trained models, the proposed approach in this paper leverages the strengths of language-image pre-trained network while incorporating knowledge from the 3D modeling domain through view consistency constraint. This method improves the robustness and general-

ization capability of the network by designing self-supervised enhancement under mask consistency constraint and refining the homogeneity consistency constraint loss function. Therefore, this method achieves accurate improvement for zero-shot 3D model classification.

Key words: 3D model classification; zero-shot learning; self-supervised learning; image-text pre-training; visual-language multimodality

0 引 言

三维模型分类作为三维模型应用的一项核心任务，是工业制造、自动驾驶和虚拟现实等诸多领域的研究基础(白静 等，2021；龙霄潇 等，2021)。得益于深度学习的广泛应用，深度神经网络在该任务上取得了令人瞩目的成果。然而，现有的方法大多基于监督学习，因此其识别能力主要局限于训练过程中已见过的模型类别。随着计算机辅助设计和三维传感器技术的不断进步，各行业面临着越来越多的新颖三维模型类别。这引发了一个新的挑战：如何有效识别那些在训练过程中未曾遭遇的模型类别。

零样本学习(zero-shot learning，ZSL)为应对这一挑战提供了有力工具(冯耀功 等，2021)。针对三维点云的零样本分类任务，Cheraghian 等人(2019)提出首个零样本分类模型。相较于传统的三维模型分类器，该方法在对未知类别进行分类时表现出更高的准确性。此后的一些研究工作也取得了性能的增益。然而，与二维领域相比，三维模型的零样本学习仍然面临诸多挑战，特别是：1)三维模型标注数据集的有限可用性(Zhang 等，2022)。二维图像分类通常可以受益于大规模标注图像集上的预训练骨干网络。然而，在三维模型领域，缺乏类似规模的数据集，这种数据的稀缺性影响了网络对未知类别的泛化能力。2)在文本和三维模型之间建立无缝的语义对齐仍是一个挑战(Cheraghian 等，2019)。虽然词向量可以捕捉可见类和未见类之间的语义信息，但文本语义与三维模型特征之间的直接关联仍然不明确。相比之下，面向二维图像的零样本学习任务可以通过属性(如“有爪子”或“是黑色的”)描述对应图像，从而在文本线索和视觉内容之间建立更明确的连接。

创建大规模的数据集并标注良好对齐的属性是耗时且昂贵的。为了应对这些挑战，白静等人(2022)将基于二维域大规模数据集训练的图像编码器应用于三维模型零样本分类中，提出零样本分类模型 ZS3D-Net，取得了不错的效果，展示了间接应用二维数据拓展三维零样本学习任务的优势。此后，PointCLIP(Zhang 等，2022)将同时蕴含大规模二维数据信息和良好文本标注的对比视觉—语言预训练网络(contrastive language-image pre-training，CLIP)(Radford 等，2021)应用于三维模型零样本分类，取得了突破性进展，验证了利用 CLIP 大模型完成三维模型分类任务的潜力。

具体而言，PointCLIP 将三维模型投影为多视图，通过固定参数的 CLIP 对视图进行编码得到多视图特征，随后为每个视图手工设置权重并拼接为三维特征描述符，最后将三维特征描述符与标签语义特征进行相似度评价，完成分类。通过分析发现，PointCLIP 没有针对三维模型的训练，优势是其可以完整保留 CLIP 中蕴含的大规模数据集和与之对应的丰富语义信息；但是另一方面，PointCLIP 不具备对三维数据的可学习性，造成三维知识的匮乏，这势必会影响网络对三维模型的编码能力，从而影响分类效果。为了克服这一问题，一个简单且直接的思路是借鉴多视图卷积神经网络(multi-view convolutional neural network，MVCNN)的学习策略(Su 等，2015)，即引入训练好的 CLIP 对视图进行编码，并使用三维模型数据集对 CLIP 的网络参数调优，以兼具二维数据标注优势与三维模型的自适应性。然而，如图 1 所示，这种策略无法对 CLIP 奏效，经过“调优”后的网络分类性能较 PointCLIP 显著下降。分析可知，大量未见类的二维知识在调优过程中会被逐渐遗忘；同时由于训练集规模的制约，导致网络对有限三维模

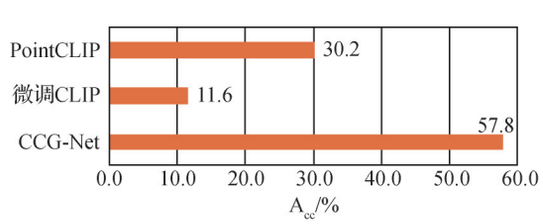

图 1 不同方法在 ModelNet10 测试集上零样本分类精度

Fig. 1 Zero-shot classification accuracy of different methods on ModelNet10 testing set

型内可见类的过拟合;两者相互作用,严重影响了网络在零样本学习任务上的分类性能。

基于以上分析,本文提出一致性约束引导的零样本三维模型分类网络(consistency constraint guided network of zero-shot 3D classification, CCG-Net)。这是一种基于CLIP的多视图表示学习,用于零样本三维分类。CCG-Net包括1个固定部分和1个动态可学习部分。固定部分利用CLIP从大规模二维图像与语义信息中学习跨模态信息,侧重二维信息的保存;可变部分由一致性约束引导的全局视图编码器组成,强调三维知识的习得。两者结合以适应三维模型数据与零样本学习任务的特性。总体来看,本文的贡献包括以下3点:1)提出视图一致性约束引导的零样本深度学习网络框架。构建与CLIP视图编码器并行的可学习网络(全局视图编码网络)取代直接微调CLIP参数策略,首先对三维模型用CLIP(固定参数)提取更具泛化性的二维视图特征,同时用全局视图编码网络提取专注三维知识的视图特征,然后将其增补到CLIP输出的视图特征中,促进三维数据和二维预训练知识之间的互补,使网络在保持二维知识的前提下针对三维数据集的调优,将CLIP从二维图像扩展到三维模型。2)设计掩码一致性约束引导的自监督增强模块。对每一个三维模型进行多次不同的掩码操作,进而利用全局视图编码器完成三维模型编码,再将其增补至CLIP视图特征前以自监督的方式约束不同掩码特征的一致性,进一步提升全局视图编码器的鲁棒性,以获取更高质量的全局视图特征。3)构建同类一致性约束引导的非互斥损失。CLIP采用的“softmax + 交叉熵损失”假定了待分类标签之间的互斥关系,不适用于在小规模数据集上调优的零样本任务。本文基于“sigmoid + 二元交叉熵损失”,构建非互斥损失确保同类数据的一致性和异类数据的互斥性,进而改善零样本分类任务性能。

在包含5个测试集的3个公开基准上进行实验,结果验证了CCG-Net在零样本三维模型分类任务上的有效性。

1 相关工作

1.1 三维模型分类

基于不同表征形式,现有的三维模型分类方法可以分为基于多视图和基于点云的方法。

基于多视图的方法将三维模型渲染或投影到二维图像作为输入。面向三维模型识别的多视图卷积神经网络(Su等,2015)作为先驱,使用在二维数据集上预训练的卷积神经网络(convolutional neural network, CNN)作为特征提取器,并通过视图池化聚合紧凑的特征描述符完成分类任务并取得好的分类结果。在此基础上,许多改进工作相继提出,如权衡视图序列对的贡献(Johns等,2016)、引入长短期记忆机制(Ma等,2019)以及图神经网络(Wei等,2020)等。上述研究侧重视图关系的理解与融合,很少讨论对预训练特征提取器改进研究。

基于点云的方法直接在原始点云上进行处理。PointNet(Qi等,2017a)首先尝试使用多层感知机对点进行编码,并使用最大池化操作实现排列不变性。在此基础上,分层点云(Qi等,2017b)、轻量化网络(白静 等,2019a)和多尺度点云(白静 等,2019b)等改进被提出;2023年以后对三维点云判别性特征的提取受到关注,相关工作展示了其在不同粒度点云分类任务的重要性(白静 等,2023)。

上述方法在三维模型分类任务中取得了较好的效果,但是这些方法都基于监督学习,训练和测试都在具有相同标签的三维模型上,无法对未见类进行正确分类。

1.2 零样本分类

零样本分类在三维模型任务上的工作相对比较少,本节首先简单介绍在二维图像领域中的零样本分类方法,然后再讨论零样本三维模型分类任务。

1.2.1 零样本图像分类

二维图像的零样本分类近年来取得了显著的进步。早期的方法主要关注如何学习一个嵌入模型,通过桥接视觉和语义领域,将语义知识从已见类别转移到未见类别。由于属性与视觉内容高度相关,人工智能生成内容(刘安安 等,2024)被引入这一领域,生成对抗网络(generative adversarial network, GAN)(Goodfellow等,2020)和变分自编码器(variational autoencoder, VAE)(Kingma 和 Welling, 2014)等生成模型被集成到ZSL方法中,通过条件生成的方式为未见类别生成样本来解决枢纽性(hubness)和领域转移(domain shift)等问题(Li等,2019;Vyas等,2020)。

上述方法通常采用ImageNet预训练的卷积神经

网络作为骨干网络，改进主要在特征提取之后进行。CLIP（Radford等，2021）提出新的思路，通过互联网采集4亿个图像—文本对构建数据集，以此预训练视觉和语义特征编码，通过对比学习来建立语言和视觉之间的关联，进而完成零样本分类任务。

1.2.2 零样本三维模型分类

三维模型的零样本分类工作相对较少。Cheraghian等（2019）提出首个基于点云的方法ZSLPC（zero-shot learning of 3D point cloud），使用点云网络PointNet（Qi等，2017a）作为特征提取网络提取三维模型的全局特征，以词向量作为辅助信息构建可见类与未见类之间的关联，完成对未见类的识别。同年，该团队提出通过缓解枢纽性问题的改进方法MHPC（mitigating the hubness problem for ZSL of 3D point cloud）（Cheraghian等，2019）。为了引入生成神经网络以解决域间差异问题，Hao等人（2023）提出CGRL（contrastive generative network with recursive loop）以扩大类间距离、缩小类内差距；Abdullah等人（2024）结合VAE与GAN两种网络的优势，针对零样本三维模型分类提出VAE-GAN3D，并在生成网络的基础上，利用二维视觉语义作为辅助信息。白静等人（2022）基于多视图表征方式提出ZS3D-Net（zero-shot classification network for 3D models），采用MVCNN为基础架构提取全局特征，间接利用其中的二维预训练知识，增强了网络的泛化性能并进一步缓解了枢纽型问题，实现了更高的准确性。然而，这一方法在有效捕捉二维视图之间及视图与语义之间的关联关系方面仍然存在不足。范有福等人（2024）提出基于双线性注意力池化的判别特征引导方法（discriminative feature-guided zero-shot learning of 3D model classification，DFG-ZS3D），采用语义判别损失和内容感知损失联合监督，从语义到内容共同约束真实视觉特征和伪视觉特征的对齐，解决了这个问题。

随着CLIP的提出，PointCLIP（Zhang等，2022）将其首次应用于零样本三维模型分类任务，该方法将三维模型投影为多视图，通过CLIP对视图进行编码得到多视图特征，随后为每个视图手工设置权重并拼接为三维特征描述符，最后将三维特征描述符与标签语义特征进行相似度评价，完成分类。随后，PointCLIPv2（Zhu等，2023）对投影算法进行改进使其更适应CLIP原始图像域，提升了零样本分类准确度，CALIP（CLIP with attention）（Guo等，2023）在视图特征与语义特征间引入无参注意力机制，也取得了效果提升。上述方法均无需使用三维模型对编码器进行训练，这一设计简化了网络，也展示了大规模预训练模型在跨模态推理中的潜力，启发了“三维视觉—语言”推理技术（雷印杰 等，2024），但是限制了网络的性能。

2 本文方法

2.1 网络框架

通过分析可知，将CLIP迁移到零样本三维模型分类任务的关键在于两种模态间信息的相互补充，即保留二维预训练网络中既有知识的同时，学习三维模型特有信息。直接在三维数据集上微调CLIP网络会造成二维知识的遗忘，导致网络对有限三维模型内可见类的过拟合，影响网络在零样本学习任务上的分类准确性。为此，本文设计了一种一致性约束引导的零样本分类网络CCG-Net，如图2所示。

任务目标：三维模型的零样本分类涉及将三维模型数据划分为两个不同的子集，分别表示为$\boldsymbol{X}^s$和$\boldsymbol{X}^u$，其中s代表“已见”，u代表“未见”，对应的标签集合分别是$\boldsymbol{Y}^s$和$\boldsymbol{Y}^u$，且有$\boldsymbol{Y}^s \cap \boldsymbol{Y}^u = \varnothing$。零样本分类网络在“已见”集$\boldsymbol{D}^s = \{\boldsymbol{X}^s, \boldsymbol{Y}^s\}$上进行训练，在“未见”集$\boldsymbol{D}^u = \{\boldsymbol{X}^u, \boldsymbol{Y}^u\}$上进行测试，以评估模型的性能和有效性。

在训练阶段，将训练集$\boldsymbol{D}^s$中的三维模型与其对应的标签$\{\boldsymbol{X}^s, \boldsymbol{Y}^s\}$成对输入CLIP编码网络。其中，三维模型$\boldsymbol{X}^s$被输入CLIP的视图编码器，提取其视图特征向量$\boldsymbol{F}^s_{\text{Visual}} \in \mathbf{R}^{N^s \times V \times P}$，$N^s$表示模型数量，$V$表示每一个模型的视图数，$P$表示视图特征维度；标签$\boldsymbol{Y}^s$被向量化后送入CLIP语义编码器，提取其语义特征向量$\boldsymbol{F}^s_{\text{Semantic}} \in \mathbf{R}^{N^s \times P}$。利用CLIP所蕴含的二维知识，建立了三维视觉与语义信息之间的初步关联。同时，将结合掩码操作并联合掩码一致性约束（图2（a））和同类一致性约束（图2（b））构建可学习的三维模型编码器，提取三维模型的全局特征$\boldsymbol{F}^s_{\text{Global}} \in \mathbf{R}^{N^s \times V' \times P}$（$V'$表示全局视图数），用以弥补CLIP中三维知识的缺失（图2（c））。

在测试阶段，输入待分类三维模型集合$x^u \in \boldsymbol{X}^u$以及未知类标签集合$\boldsymbol{Y}^u$。三维模型x^u被同时输入CLIP视图编码器与全局视图编码器，分别得到包含

三维模型中的二维视觉信息的视图特征向量 $\boldsymbol{F}^{u}_{\text{Visual}} \in \mathbf{R}^{N^u \times V \times P}$ 和补充三维信息的全局视图特征向量 $\boldsymbol{F}^{u}_{\text{Global}} \in \mathbf{R}^{N^u \times V' \times P}$，按视图拼接后送入线性层得到所有待分类三维模型的最终特征向量集 $\boldsymbol{F}^{u}_{\text{3D}} \in \mathbf{R}^{N^u \times P}$；所有未知类标签集合 Y^u 在进行向量化后，输入 CLIP 语义编码器，得到标签语义特征向量集合 $\boldsymbol{F}^{u}_{\text{Semantic}} \in \mathbf{R}^{K^u \times P}$，$K^u$ 为未知类标签个数；然后在特征空间计算三维特征向量集合与语义特征向量集合之间的余弦相似度，得到矩阵 $\boldsymbol{M}^u \in \mathbf{R}^{N^u \times K^u} = \boldsymbol{F}^{u}_{\text{3D}} \otimes (\boldsymbol{F}^{u}_{\text{Semantic}})^{\text{T}}$，$\otimes$ 代表矩阵乘法。由相似度矩阵可得 N^u 个三维模型各自在 K^u 个标签上的 logits，将与三维特征向量相似度最大的语义特征向量所属标签作为每一个三维模型的分类结果，记为 $result^u \in \mathbf{R}^{N^u} = \text{argmax}(\boldsymbol{M}^u, 1)$。

下面将具体介绍 CCG-Net 的各个模块以及损失函数。为了简化表达，后文中不再特别区分训练集数据或测试集数据，省略上标 s 和 u。

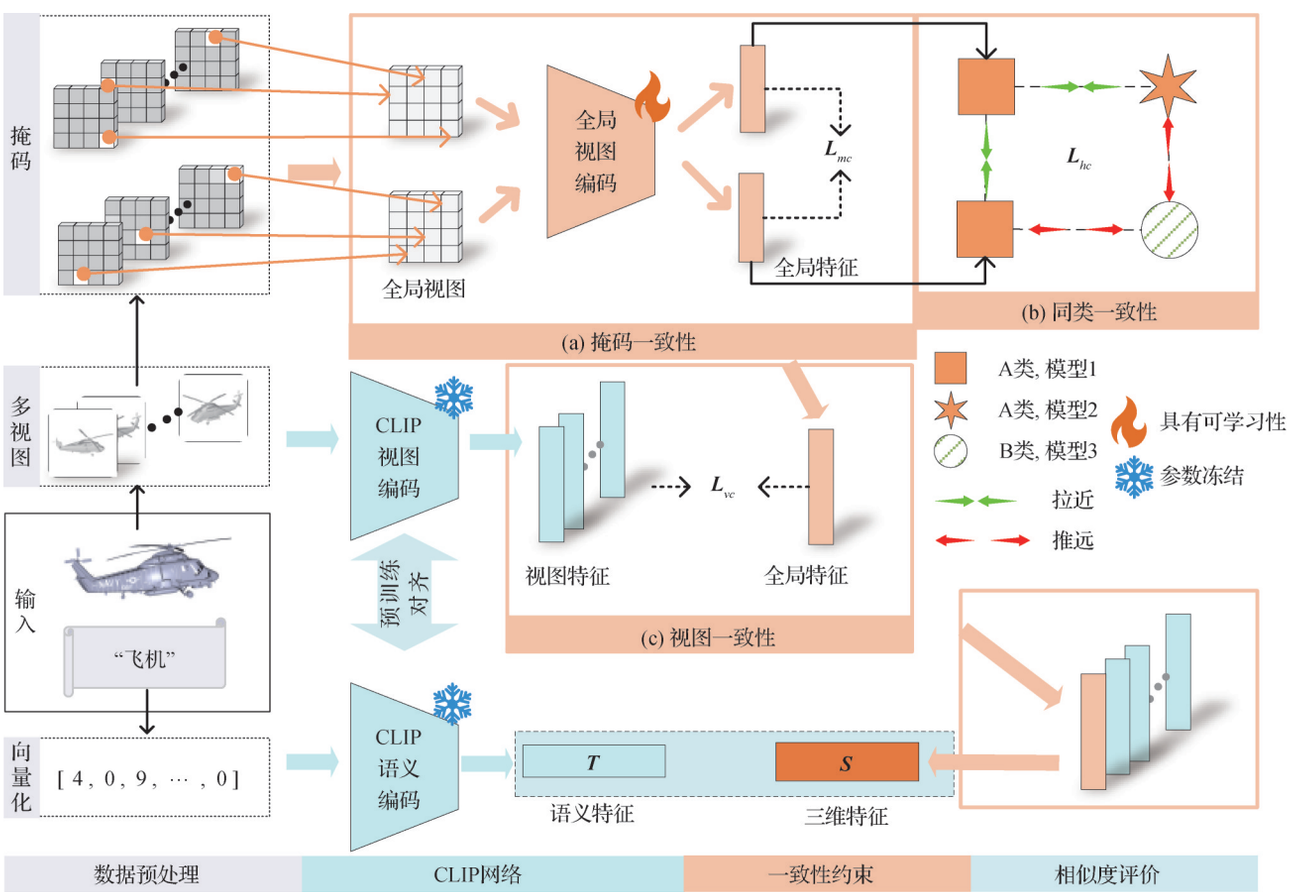

图 2　CCG-Net 整体框架

Fig. 2　Architecture of CCG-Net

2.2　基于视图一致性的零样本分类网络

零样本学习在可见类数据上训练，在未见类上测试。因此，直接使用三维数据集微调二维预训练网络（如 MVCNN）仅能优化网络对可见类的损失，在这一过程中，训练集中未见过的视觉概念会被网络“遗忘”，这样的调整通常会导致模型过拟合，进而损害在未见类的分类效果。为了应对这一问题，本文提出在视图层面对预训练网络进行调优，构建基于视图一致性的三维模型编码器，为 CLIP 输出的多视图特征补充三维知识，以适应零样本三维模型分类任务。具体步骤如下：

1）二维信息保持。三维模型以多视图为输入，将它们输入到 CLIP 视觉编码器 $Enc_{\text{Visual}}(\cdot)$ 中，以获取与语义信息隐式关联的视图特征。在训练过程中固定 CLIP 的原始参数，保留全部来自二维预训练网络的知识。整个过程可以表示为三维模型 $x \in \mathbf{R}^{V \times H \times W \times C}$ 到多视图特征 $\boldsymbol{f}_{\text{Visual}} \in \mathbf{R}^{V \times P}$ 的映射。

2）三维多视图掩码。多视图包含大量冗余信

息，且大量视觉信息已经通过CLIP视图编码器提取并保留。为了去除视图间冗余，高效捕获三维信息，引入了多视图掩码操作。具体步骤为，随机以(0，1)初始化掩码矩阵$\boldsymbol{M}' \in \mathbf{R}^{V\times 4\times 4}$，随后上采样至$\boldsymbol{M} \in \mathbf{R}^{V\times H\times W\times C}$对各个视图进行掩码，对掩码后的视图进行拼接并重新整理得到一个全局视图，记做$\boldsymbol{x}_M \in \mathbf{R}^{1\times H\times W\times C}$。

3)全局视图特征编码。掩码后的视图通过全局视图编码网络$Enc_{\text{Global}}(\cdot)$进行特征提取。全局视图编码网络由全局视图特征编码器和一个投影头组成，前者将$\boldsymbol{x}_M$映射到一个特征嵌入空间，后者负责将特征嵌入投影至与CLIP编码特征空间相同的维度，得到全局视图特征$\boldsymbol{f}_{\text{Global}} \in \mathbf{R}^{1\times P}$。

4)视图一致性约束。维度匹配后的全局视图特征与CLIP视觉特征计算KL(Kullback-Leibler)散度，记做L_{vc}并将其作为损失函数以约束全局视图编码网络的学习，具体为

$$L_{\text{vc}} = \sum_{i=1}^{N} f_{\text{Visual}}^{i} \times \log\left(\frac{\boldsymbol{f}_{\text{Visual}}}{\boldsymbol{f}_{\text{Global}}}\right) \tag{1}$$

通过以上步骤，网络从三维数据中学习额外的全局视图特征，并将其增补至CLIP视图特征集合中。通过全局视图特征，使网络的输出具备三维信息，也最大程度地保留了二维预训练知识，避免了直接微调策略导致的二维信息遗忘与训练集上过拟合现象。

2.3 基于掩码一致性的三维自监督增强

在2.2节中，采用掩码操作将多视图压缩为单个全局视图，减少了三维模型多视图的冗余性，同时提高了网络的泛化能力。但是可以观察到，仅从单一的全局视图中提取的特征，忽略了多视图表征的整体性，网络缺少对三维模型整体以及多视图间信息的理解能力。基于以上分析，本文提出掩码一致性约束：使用不同的掩码操作获取更具多样性的全局视图并提取特征，执行对比学习；约束同模型全局特征间的一致性，实现自监督，增强网络鲁棒性并进一步提高泛化性能。具体步骤如下：

1)多样化全局视图构造。有别于2.2节中一个三维模型仅构造一个全局视图，为了使网络关注多视图间的关联关系，对于任意三维模型，构造不同掩码矩阵以获得不同的全局视图。以两个全局视图为例，记$\boldsymbol{x}_M^j$和$\boldsymbol{x}_M^k$，其中，$1 \leqslant j \neq k \leqslant V$。

2)全局视图编码。同一个三维模型x经过两种不同的掩码操作，分别得到两个不同的全局视图$\boldsymbol{x}_M^j$和$\boldsymbol{x}_M^k$，随后经过一组相应数量、参数共享的全局视图编码器，得到全局视图特征$\boldsymbol{f}_{\text{Global}}^j$和$\boldsymbol{f}_{\text{Global}}^k$。

3)掩码一致性约束。视每一个三维模型实例的不同全局特征互为正样本，其他实例的全局特征作为负样本，采用对比损失L_{mc}，最大化正样本相似度，使三维模型内部不同掩码操作后的全局特征保持一致；同时最小化与同批次中其他模型的全局特征的相似度。

2.4 基于同类一致性的非互斥对比损失

在掩码一致性约束中，采用的对比损失源自CLIP。在CLIP训练过程中，图像与文本成对输入编码网络，输出的特征互为标签并进行相似度评价，每一个视觉特征与不同文本特征间的相似度经过softmax运算，计算结果作为该视觉特征对不同标签(实际上是文本特征)的后验概率；反之文本特征对视觉特征亦然。最后使用交叉熵损失(cross entropy，CE)，约束成对特征间的后验概率使其最大。softmax假设待分类标签之间存在互斥关系，在计算后验概率时，对数的相对强度(即对数比率)才是最重要的。当一个训练批次内存在重复标签时，该假设会受到破坏。然而，在大规模数据集预训练网络(即CLIP)中，此类情况被认为是低概率事件。此外，“softmax + CE”这一组合及其改进工作也被广泛应用在分类、分割等视觉理解任务中，待分类标签间的互斥假设在该类任务中被认为是可以接受的。

但在零样本三维模型分类任务中，由于三维模型数据集规模较小，在对网络进行调优时，每一个批次内出现存在重复类的概率会显著增加。在这种情况下，使用互斥损失会导致约束对象错误。同时，零样本任务会将网络用于分类未见类对象，未见类的类别范围与训练中的可见类不同，这将导致未见类的对数与其他未见类别的对数校准不佳。

这个问题反映在掩码一致性约束中，表现为损失函数会降低具有相同标签的不同实例间的特征相似度。为解决这些问题，建议在训练时使用非互斥损失，具体而言，使用sigmoid和二元交叉熵(binary cross entropy，BCE)损失，以避免相同类别的不同实例特征发生互斥，确保训练方向正确。

将同一训练批次内的两个不同全局特征记做$\boldsymbol{z}^j$和$\boldsymbol{z}^k$，在掩码一致性约束中采用的对比损失只能约束：当且仅当两者来自同一模型使它们之间相似度s

(z^j, z^k)最大，并通过 softmax 操作被动约束来自不同模型的z^j和z^k之间相似度最小，即使它们属于同一类别。该约束并不合理，当z^j和z^k来自同类不同实例时相似度也应最大。为实现这一目标，本文将相同类别不同实例的全局特征相互视为正样本，来自不同类别的全局特征则为负样本。每一组损失采用加权平均即得同类一致性损失L_{hc}，具体为

$$L_{\mathrm{hc}} = \sum_{j=1}^{2N}\left(\frac{1}{N_{\mathrm{pos}}}\sum_{k=1}^{2N} 1_{jk}^{\mathrm{pos}} l_{jk} + \frac{1}{N_{\mathrm{neg}}}\sum_{k=1}^{2N} 1_{jk}^{\mathrm{neg}} l_{jk}\right) \tag{2}$$

式中，将掩码一致性对比损失L_{mc}中对一组相似度取 softmax 操作改为分别取 sigmoid（记为σ）操作，使用 BCE 计算该组损失l_{jk}，具体为

$$\begin{aligned} l_{jk} = &-y_{jk} \times \log\sigma(s(z^j, z^k)) - \\ &(1 - y_{jk}) \times \log\sigma((1 - s(z^j, z^k))) \end{aligned} \tag{3}$$

综合视图一致性约束L_{vc}与同类一致性约束损失L_{hc}，网络 CCG-Net 的整体损失定义为

$$L_{\mathrm{total}} = \alpha L_{\mathrm{hc}} + \beta L_{\mathrm{vc}} \tag{4}$$

式中，α与β为权重，旨在调节不同分量在最终算式中所占的比重。本文中，α与β设置为 1。

3 实验与分析

3.1 数据集

零样本三维模型分类任务中公开使用的数据集包括 ZS3D（zero-shot classification network for 3D models）数据集、Cheraghian 数据集和 ScanObjectNN 数据集。

ZS3D 数据集（白静 等，2022）是专为零样本三维模型分类任务设计的数据集，以 Shrec2014 和 Shrec2015（Lian 等，2015）为数据源构建，其训练集包含来自 33 个类的 1 493 个模型，测试集包含来自 8 个类的 184 个模型，共计 41 个类、1 677 个非刚性三维模型。

Cheraghian 数据集（Cheraghian 等，2019）基于室内合成三维数据集 ModelNet40、ModelNet10（Wu 等，2015）与小型非刚性建模数据集 McGill（Siddiqi 等，2008）、Shrec2015（Lian 等，2015）4 个数据集二次划分。训练集采用 ModelNet40，但不包括与 ModelNet10 重合的 10 个类别，即 30 个类 5 976 个三维模型。测试集采用 ModelNet10、McGill 和 Shrec2015，其中 McGill 包含 14 个类别的 301 个模型，Shrec2015 包含 30 个类别的 720 个模型。

ScanObjectNN 数据集（Uy 等，2019）是一个真实的三维点云分类数据集，其三维模型来自真实世界室内场景扫描而非 CAD 合成。该数据集包含来自 15 个类别的 2 902 个三维模型实例。同时，根据对原始数据的预处理策略，该数据集包含多种变体，本文选取了过滤背景点的 OBJ_ONLY（object only）、包含背景点的 OBJ_BG（object and background）以及经过数据增强的 PB_T50_RS（object augmented rot scale）3 个变体。

3.2 实验配置与评价指标

本文实验在配备了 Nvidia Tesla V100-SXM GPU 的 Ubuntu 18.04 系统上进行。开发环境为 PyTorch1.7.0+CUDA10.1 +Python3.8。

与之前的研究一致，本文采用 12 视图表征三维模型，每个视图为 3 × 256 × 256 的图像。对于 CLIP 网络，使用 ResNet50 提取视图特征，并使用简单的模板[*classname*]构造语义提示；全局视图特征编码网络采用 ResNet18。网络在训练时使用 AdamW 优化器，学习率为10^{-3}，权重衰减为1×10^{-2}，迭代 100 次，batchsize 为 32。

实验包含定量实验与定性实验；定量实验包括在 ZS3D 与 Cheraghian 数据集上的对比实验，以及在 Cheraghian 数据集中 ModelNet10 上的消融实验，两组实验均采用 top-1 准确率作为衡量效果的指标，简记为 Acc；定性实验包括混淆矩阵和样本分析。

3.3 在 ZS3D 数据集上的实验对比与分析

相较于二维图像，面向三维模型的零样本三维模型分类方法较少。对于 ZS3D 数据集，仅有 ZS3D-Net 和 DFG-ZS3D，这两种方法均采用多视图表征三维模型，且视图数为 12，与本文方法一致。除此之外，本文选取了一部分面向二维图像的经典方法作为补充，包括基于嵌入的方法 SJE（structured joint embeddings）（Akata 等，2015）、SAE（semantic autoencoder）（Kodirov 等，2017）和基于生成的方法 LisGAN（leveraging the invariant side for GAN）（Li 等，2019）、LsrGAN（leverages the semantic relationship for GAN）（Vyas 等，2020），为适应二维图像方法的要求并保证信息完整性，输入采用多视图拼接后形成的大尺寸图像。

如表 1 所示，本文方法 CCG-Net 在 ZS3D 数据集中取得了最优的性能。具体对比及其分析如下：

表1　在ZS3D数据集上对比实验

Table 1　Comparison experiment on ZS3D dataset

表征形式	方法	Acc/%
二维图像	SJE(Akata等,2015)	32.2
	SAE(Kodirov等,2017)	18.8
	LisGAN(Li等,2019)	36.8
	LirGAN(Vyas等,2020)	42.6
三维模型	ZS3D-Net(白静 等,2022)	58.6
	DFG-ZS3D(范有福 等,2024)	<u>60.9</u>
	CCG-Net(本文)	**70.1**

注:加粗、下划线字体分别表示最优、次优结果。

1)二维图像领域的零样本学习经典方法,其在处理三维模型时仍能取得一定效果,展示出二维信息的有效性。

2)对比不同模态间的方法,专门针对三维模型的零样本分类网络具有更好的表现。尤其是与LirGAN相比,CCG-Net取得了超过27%的性能提升。经分析认为尽管二维信息是有用的,但是面向二维图像的方法不能捕捉三维模型多视图之间的信息,面向三维模型的方法通过捕捉视图间的关系取得了性能提升。

3)对比零样本三维模型分类方法,DFG-ZS3D相较于ZS3D-Net取得2.3%的提升;本文方法CCG-Net相较于ZS3D-Net提升了11.5%的显著增益,相较于DFG-ZS3D也取得了9.2%的性能增益。这是因为受限于三维模型零样本数据集规模偏小、泛化性能不够等问题,DFG-ZS3D这种基于预训练二维网络的方法存在信息遗忘的缺陷,较难取得高的性能提升;本文方法引入大规模二维图像预训练网络,并保留其参数不变,能够更加充分地获取二维图像数据集及其良好标注带来的性能增益,为三维模型的零样本分类提供了良好的解决思路。

3.4　在Cheraghian数据集上的实验对比与分析

为进一步验证本文方法CCG-Net的普适性,选取Cheraghian数据集进行综合对比。如表2所示,对比方法包括3类,涵盖了经典方法及先进方法。其中,基于二维图像的方法包括f-CLSWGAN(feature generating networks for ZSL)(Xian等,2018)和CADAVAE(cross and distribution aligned VAE)(Schönfeld等,2019),并沿用3.3节中的输入设定;基于三维点云的方法以三维点云为输入并提取点云特征提取,进而完成分类任务,包括Cheraghian提出的ZSLPC(Cheraghian等,2019)、MHPC(Cheraghian等,2019)、CGRL(Hao等,2023)与VAE-GAN3D(Abdulla等,2024);基于三维多视图的方法需要将三维模型预处理为多视图表征之后再进行后续操作,包括PointCLIP(Zhang等,2022)、PointCLIPv2(Zhu等,2023)、ZS3D-Net(白静 等,2022)和DFG-ZS3D(范有福 等,2024)。

横向对比表2中相同方法在不同测试集上的实验结果可以发现:在以ModelNet40为训练集时,采用ModelNet10(MN10)为测试集时整体准确率较好,采用McGill(MG14)和Shrec2015(SH15)为测试集时准确率较低。主要有两方面原因,其一是零样本学习的前提是未知的测试类和已知的训练类存在语义关联性,而Ali数据集整体关联性较弱,其二是训练集ModelNet40主要包括刚性三维模型,而测试集McGill和Sherc2015对应非刚性三维模型,域间差异大导致网络性能不佳。

纵向对比表2中相同测试集上不同方法的分类准确率可以发现:1)基于二维图像的f-CLSWGAN、CADA-VAE与面向三维点云的ZSLPC性能相当,精度较低。这是因为基于图像的方法对三维模型表征能力弱,基于点云的方法缺少泛化性强的主干网络,这些问题限制了对应方法的有效性;2)本文方法CCG-Net在ModelNet10、McGill和Shrec2015三个测试集上分别取得了57.8%、19.9%和12.2%的准确率,位居第1,且远超其他方法。这是因为CCG-Net有效综合了高泛化性二维预训练网络CLIP的二维知识,并提取补充三维模型的全局信息,形成二维共性知识和三维个性知识的优势互补,因而取得了最佳分类性能。

3.5　在ScanObjectNN数据集的实验对比与分析

为提高模型的可信度,并探索本文CCG-Net在真实场景中的泛化性表现,加入ScanObjectNN数据集进行对比,如表3所示。选择基于CLIP的PointCLIP(Zhang等,2022)与PointCLIPv2(Zhu等,2023)以及基于原始点云的CGRL(Hao等,2023)与VAE-GAN3D(Abdulla等,2024)作为对比方法。其中,基于CLIP的方法根据视觉编码器的不同又分为ResNet50与ViT-B/16(vision transformer base patch 16)两种,基于原始点云的方法选择各自分类精度最

表 2　在 Cheraghian 数据集上对比实验

Table 2　Comparison experiment on Cheraghian's dataset

表征形式	方法	不同测试集准确率 Acc/%		
		MN10	MG14	SH15
二维图像	f-CLSWGAN(Xian 等,2018)	20.7	10.2	5.2
	CADAVAE(Schönfeld 等,2019)	23	10.7	6.2
三维点云	ZSLPC(Cheraghian 等,2019)	23	10.7	5.2
	MHPC(Cheraghian 等,2019)	33.9	12.5	6.2
	CGRL(Hao 等,2023)	35.3	**19.9**	×
	VAE-GAN3D(Abdulla 等,2024)	37.4	×	×
三维多视图	PointCLIP(Zhang 等,2022)	30.2	×	×
	ZS3D-Net(白静 等,2022)	30	15.1	6.7
	PointCLIPv2(Zhu 等,2023)	<u>35</u>	×	×
	DFG-ZS3D(范有福 等,2024)	31.9	<u>16.6</u>	<u>9.9</u>
	CCG-Net(本文)	**57.8**	**19.9**	**12.2**

注:加粗、下划线字体表示各列最优、次优结果。"×"表示暂无数据。

好的三维视觉编码器,分别是 PointNet 和 PointConv。

横向对比表 3 所示各数据集变体可以发现,过滤背景点后的数据集 OBJ_ONLY 难度最低,各方法表现良好;包含背景点的 OBJ_BG 和经过数据增强的 PB_T50_RS 难度较大,各方法表现不尽相同。

纵向对比基于 CLIP 的方法发现,使用 ResNet50 作为视觉编码器时,PointCLIP 与 PointCLIPv2 取得了较低的精度。考虑到 PointCLIP 与 PointCLIPv2 具有不可学习性,模型性能完全依赖 CLIP 自身的能力,尝试将视觉编码器替换为编码能力更强的 ViT-B/16 编码器,由于编码器能力的提升,两个模型均取得了性能提升,但是仍然远低于使用 ResNet50 的 CCG-Net。

进一步对比 PointCLIP 与 PointCLIPv2,可以观察到 PointCLIPv2 针对投影算法的改进并不适合所有数据集。在面向经过数据增强的变体数据集 PB_T50_RS 时,无论使用何种视觉编码器(ResNet50 或 ViT-B/16),都表现出不同程度的性能下降情况。

与基于原始点云的方法相比,CCG-Net 取得了最好效果,VAE-GAN 取得次优。CGRL 采用词向量作为辅助信息,VAE-GAN3D 提出二维辅助信息,辅助信息都只针对单一模态,而 CCG-Net 基于 CLIP 跨模态预训练网络,展示出跨模态信息的潜力。

表 3　在 ScanObjectNN 数据集上对比实验

Table 3　Comparison experiment on ScanObjectNN dataset

/%

方法	视觉编码器	OBJ_ONLY	OBJ_BG	PB_T50_RS
PointCLIP(2022)	ResNet50	10.5	6.8	7.3
	ViT-B/16	15.2	12.7	<u>15.4</u>
PointCLIPv2(2023)	ResNet50	9.8	7.2	7.1
	ViT-B/16	18.9	<u>15.1</u>	11.4
CGRL(2023)	PointNet	14.0	×	×
VAE-GAN3D(2024)	PointConv	<u>26.7</u>	×	×
CCG-Net(本文)	ResNet50	**32.4**	**28.9**	**19.3**

注:加粗、下划线字体分别表示各列最优、次优结果。"×"表示暂无数据。

3.6 计算开销对比

如表4所示，本节对CCG-Net与同样基于CLIP的两个方法PointCLIP（Zhang等，2022）与PointCLIPv2（Zhu等，2023）在计算开销上进行对比，对比指标涵盖了计算量GFLOPs（gigabyte floating point operations）与参数量Params。由于PointCLIP与PointCLIPv2两个方法直接使用CLIP模型完成分类任务，并未增加额外的神经网络模块，因此网络计算开销取决于CLIP编码器自身的计算需求，即采用相同基网时，其计算开销一致。为此，如表4所示，本节将两种方法合并表征为PointCLIP/v2；同时，考虑到CLIP编码器由视觉编码器（vision encoder，VE）与语义编码器（semantic encoder，SE）两部分组成，为了进行更加细致的对比分析，表中分别计算了两部分的计算开销和参数量。在视觉编码器部分，以ResNet50为基准，加入计算开销的增长比率。

表4 计算开销对比

Table 4 Comparison of computational costs

方法	GFLOPs		Params/M	
	VE	SE	VE	SE
PointCLIP/v2 w/ResNet50	12.75	5.96	38.72	63.69
PointCLIP/v2 w/ViT-B/16	35.13 (+176%)	5.59	86.19 (+123%)	63.43
CCG-Net（本文）	14.39 (**+13**%)	5.96	50.43 (**+30**%)	63.69

注：括号内数据表示计算开销的增长率，加粗字体为计算开销的增长幅度更小，即更优。

分析表4所展示的计算开销可得：1）语义编码器对整体计算开销的影响较小。表4中无论采用什么方法，语义编码器的开销始终维持较小的变化；2）PointCLIP/v2性能的提升依赖开销更大的视觉编码器。结合表3可知，使用ViT-B/16编码器取代ResNet50能够使PointCLIP/v2得到性能提升，然而ViT-B/16编码器的GFLOPs与Params两项开销分别是ResNet50的2.76倍与2.23倍；3）CCG-Net只增加较小的计算开销就取得了显著的性能提升。相较于使用ResNet50视觉编码器的PointCLIP/v2，CCG-Net的GFLOPs与Params两项开销只增加了13%和30%，均低于ViT-B/16的计算开销。结合表3对比实验结果可知，CCG-Net通过增加可学习模块，将三维知识融入CLIP视觉特征，能够在仅增加较少的计算开销的前提下取得更高的性能提升。

3.7 消融实验

消融实验中，选取Cheraghian数据集中的ModelNet10作为测试集，对不同全局视图数量和CCG-Net的不同模块的效果进行对比分析。

3.7.1 不同全局视图数对网络的影响与分析

掩码一致性约束要求对同一个三维模型同时进行多次掩码操作，得到不同的全局视图。图3展示了不同数量的全局视图对分类结果的影响。由图3可以看出，1）当全局视图数量为1时，分类精度为43.3%，相对较低。这是因为单一的全局视图无法构建掩码一致性约束，影响网络对三维模型鲁棒特征的提取；2）当全局视图数量从1增加到2或4时，分类准确率得到改善，分别为57.8%和56.6%。这种提升既来自更多全局视图补充的三维知识，也来自掩码一致性与同类一致性的约束；掩码一致性约束可以增强视图间的交互，关注到更多三维模型视图间的关联关系；同类一致性约束可以改善学习目标，进而提升网络对未见类的泛化性。3）当进一步增加全局视图数量至6的时候，分类准确度降至37.4%；全局视图数量为12时，分类准确度仅有21.9%。性能降低的主要原因是过多的全局视图会生成更多的全局特征，调优的力度过强；同时，随着全局视图不断增加，即与三维模型原始视图数量一致时，网络逐渐丧失去冗余的功能，大量冗余的信息作为全局特征，不同模态间的知识都不能得到有效利用，使经过微调后的三维特征质量降低。

上述实验结果表明，通过掩码操作得到少量的全局视图中已经包含不同视图间信息，通过自监督网络实现了对所有视图之间关联关系的约束（即掩码一致性约束），并且这种约束是以一种具有泛化性的方式完成的（即掩码与同类一致性约束），此时网络足以具备对未知类的识别能力；过多的全局视图会引入大量冗余信息，其特征被增补至二维视图特征后将损害网络的性能。无特殊说明的情况下，本文实验中全局视图数量取2。

3.7.2 不同模块对网络的影响与分析

本实验旨在对比CCG-Net各模块对网络性能的影响。为了保证对比公平性，采用ResNet50提取各视图初始特征，并使用简单的[*classname*]作为语义

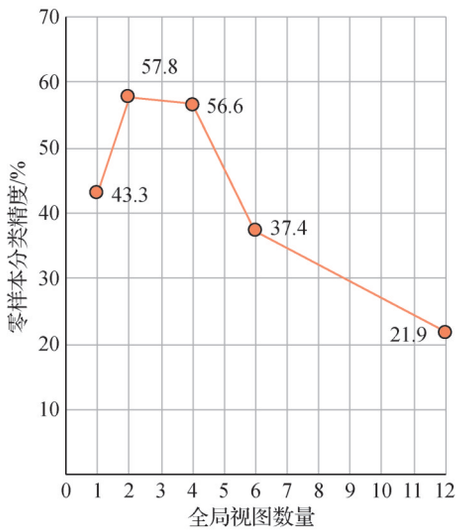

图 3　不同全局视图数的消融实验

Fig. 3　Ablation on different number of global views

提示，在此基础上使用 PointCLIPv2 作为基础网络（baseline）。随后采用 MVCNN 的训练策略，解除 CLIP 参数的固定，使用三维模型数据集对其进行微调（fine tune，FT）。接下来采用视图一致性约束，提取全局视图特征对 CLIP 进行调优（view consistency，VC），取代直接对 CLIP 参数进行微调；最后陆续加入掩码一致性约束的三维自监督增强网络（mask consistency，MC）以及同类一致性约束的损失函数（homo consistency，HC）进行消融对比。CCG-Net 不同模块的消融实验结果如图 4 所示。

由图 4 的消融实验结果可以得到以下结论：1）尝试直接在三维模型数据集上训练 CLIP 以对其参数进行微调（FT），网络的分类性能降低了 23.4%。这是因为预训练网络中的二维知识随着训练过程的推进被逐渐遗忘，较小规模的数据集又不足以支持网络学习到 CLIP 那样强大的泛化性能，与此同时习得的三维知识都集中在训练集类别上，进而导致在测试集上的性能降低。2）使用视图一致性约束（VC），网络的分类性能提升了 8.3%。这表明采用全局特征的形式对 CLIP 输出视图特征进行特征增补可以克服直接微调的缺陷，实现网络在保持二维知识的前提下对三维数据集上的调优。同时，相较于将所有视图重复进行编码，通过掩码操作获得的全局视图再进行编码，更有助于在减少多视图中的冗余提升网络的泛化能力。3）进一步加入掩码一致性约束（MC），保留 CLIP 所提出的对比损失（具有互斥性的对比损失），网络性能得到进一步提高，较基础网络相比涨幅达 20.1%。这一结果表明在掩码一致性约束的引导下，全局视图特征编码网络可以更好地捕捉视图之间的关联，加强对三维模型多视图表征的整体性理解，提取更具鲁棒性的特征，进而促进三维模型的零样本学习。4）当采用同类一致性损失（HC）替代 CLIP 所提出的互斥对比损失时，得到本文所构建的网络 CCG-Net，此时取得最高精度 57.8%，证明非互斥的学习有利于网络泛化性提升。

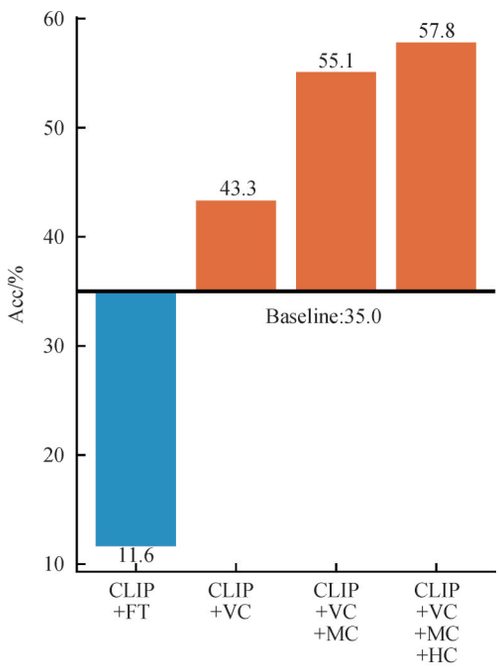

图 4　CCG-Net 不同模块的消融实验

Fig. 4　Ablation of different modules of CCG-Net

3.8　可视化结果与分析

本节使用混淆矩阵对 CCG-Net 的分类结果进行可视化展示，并对比直接微调 CLIP 参数的方法（CLIP + FT）。数据集选取 ZS3D 数据集与 Cheraghian 数据集，其中后者采用 ModelNet10 与 McGill 作为测试集。可视化结果如图 5 所示。

纵向对比相同方法下不同数据集的混淆矩阵可以观察到：对于 CLIP + FT 的方法，网络更倾向于将三维模型都分为某一类或两类。在 ZS3D 数据集（图 5（a1））上，所有三维模型基本都被归为 giraffe 类与 centaur 类。这一现象在 ModelNet10 数据集（图 5（a2））与 McGill 数据集（图 5（a3））上更为明显。前者将所有三维模型归为 sofa 类，后者将所有模型归为 hand 类。分析这一现象产生的原因，网络在依据训

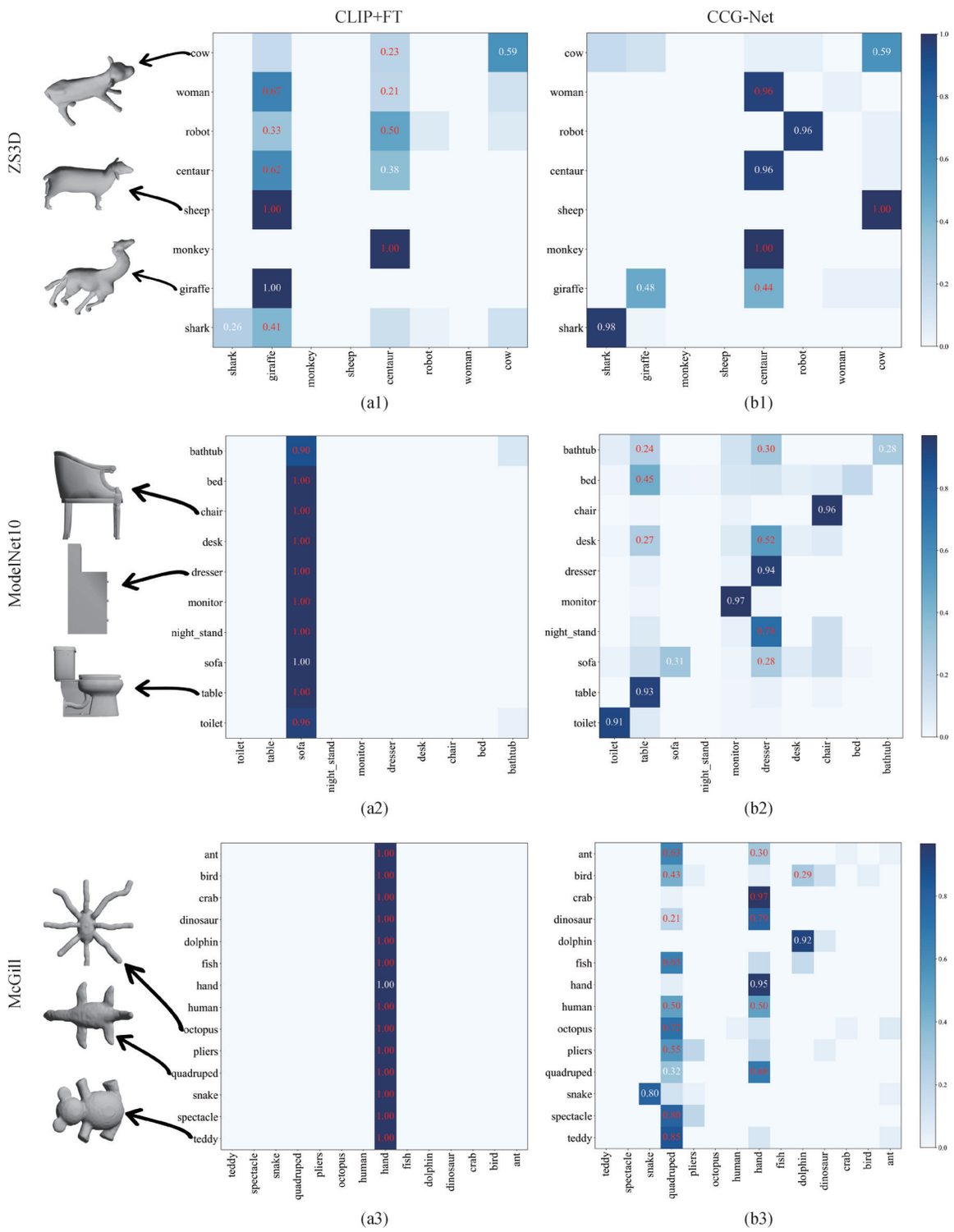

图5　在ZS3D、ModelNet10和McGill测试集上的不同方法混淆矩阵结果对比

Fig. 5　Comparison of the confusion matrix results of different methods on ZS3D, ModelNet10 and McGill testing set

练集进行参数更新时，二维知识被遗忘且泛化性能丢失，此时对测试集三维模型进行分类时，网络无法泛化到未见类当中，而同一数据集通常处于相近域内，网络倾向于将相近域中所有三维模型都归为一类；ZS3D的数据集由不同数据集构成，域间存在一定差异，因此没有另外两个数据集那么明显的倾向。

横向对比图5中相同数据集下不同方法的结果，可以观察到：1）对于ZS3D数据集，本文方法的分类精度提升源自robot类、centaur类和shark类三者的提升；sheep类原本被错分至giraffe类（图5(a1)），

本文方法中被错分至cow类(图5(b1)),从视觉特征与生物分类学(同属牛科)的角度,应当认可这是更“正确的”错误。2)对于ModelNet10测试集,toilet、table、monitor、dresser和chair类的正确分类是性能提升的关键;同时,分类结果中很少有严重甚至完全错分到某一类的情况。这一方面归功于本文方法的提升,但也应当承认这是数据集的特性使然,尽管在训练过程中网络没有见过ModelNet10中的类,但是作为同一个数据集,较低的域间差异降低了网络迁移至未见类的难度。3)对于McGill测试集,可以看到性能提升主要来自snake类与dolphin类。尽管训练集中全部都是室内物品,但是通过一致性约束引导的网络能够在三维模型和语义信息间建立有效关联,进而借助CLIP泛化能力完成分类。进一步分析错误案例,与(图5(a3))中方法将所有模型分类至hand类不同,本文方法(图5(b3))中有一部分被错分到quadruped类,诸如teddy类、spectacle类和octopus类等。根据三维模型视图可以观察到,这些类都具有足或形似足的部位,表明网络初步具备了从视图特征中捕捉局部细节的能力,但是有待进一步发掘。

4 结 论

本文提出一种一致性约束引导的零样本三维模型分类网络(CCG-Net)。提出视图一致性约束引导的调优方法,将来自三维模型的知识以全局视图特征的形式融入CLIP视图特征,取代直接微调CLIP参数的策略,促进三维数据和二维预训练知识之间的互补;设计了掩码一致性约束引导的自监督增强模块,对全局视图特征以自监督的方式约束不同掩码视图特征的一致性,进一步提升网络泛化性与鲁棒性;构建了同类一致性约束引导的非互斥损失,确保同一批次内同类不同实例之间能够取得正确的相似度约束,有效提升网络泛化性,更适用于在小规模数据集调优的零样本任务。通过在ZS3D和Cheraghian数据集上的实验和可视化,充分验证了本文方法的有效性和先进性。根据实验中表现出的域间差异对网络从可见类迁移至未见类难度的影响,以及局部判别性能力对分类效果的提升潜力,未来将考虑如何提高网络跨域适应能力以及局部细节判别能力,进一步推进三维模型零样本学习研究。

参考文献(References)

Abdullah M T, Rahman S, Rahman S and Islam M F. 2024. VAE-GAN3D: leveraging image-based semantics for 3D zero-shot recognition. Image and Vision Computing, 147: #105049 [DOI: 10.1016/j.imavis.2024.105049]

Akata Z, Reed S, Walter D, Honglak Lee N and Schiele B. 2015. Evaluation of output embeddings for fine-grained image classification//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE Computer Society Press: 2927-2936 [DOI: 10.1109/cvpr.2015.7298911]

Bai J, Shao H H, Ji H and Wu R S. 2023. An end-to-end fine-grained classification network for 3D point clouds. Journal of Computer-Aided Design and Computer Graphics, 35(1): 128-134 (白静, 邵会会, 姬卉, 武如嵩. 2023. 面向三维点云的端到端细粒度分类网络. 计算机辅助设计与图形学学报, 35(1): 128-134) [DOI: 10.3724/SP.J.1089.2023.19283]

Bai J, Si Q L and Qin F W. 2019. Lightweight real-time point cloud classification network LightPointNet. Journal of Computer-Aided Design and Computer Graphics, 31(4): 612-621 (白静, 司庆龙, 秦飞巍. 2019a. 轻量级实时点云分类网络LightPointNet. 计算机辅助设计与图形学学报, 31(4): 612-621) [DOI: 10.3724/SP.J.1089.2019.17328]

Bai J and Xu H J. 2019. MSP-Net: multi-scale point cloud classification network. Journal of Computer-Aided Design and Computer Graphics, 31(11): 1917-1924 (白静, 徐浩钧. 2019b. MSP-Net: 多尺度点云分类网络. 计算机辅助设计与图形学学报, 31(11): 1917-1924) [DOI: 10.3724/SP.J.1089.2019.17903]

Bai J, Yuan T and Fan Y F. 2022. ZS3D-Net: zero-shot classification network for 3D models. Journal of Computer-Aided Design and Computer Graphics, 34(7): 1118-1126 (白静, 袁涛, 范有福. 2022. ZS3D-Net: 面向三维模型的零样本分类网络. 计算机辅助设计与图形学学报, 34(7): 1118-1126) [DOI: 10.3724/SP.J.1089.2022.19173]

Bai J, Zhou W H, Tuo J W and Qin F W. 2021. End-to-end sketch-3D model retrieval with spatiotemporal information joint embedding. Journal of Computer-Aided Design and Computer Graphics, 33(6): 826-836 (白静, 周文惠, 拖继文, 秦飞巍. 2021. 时空信息联合嵌入的端到端三维模型草图检索. 计算机辅助设计与图形学学报, 33(6): 826-836) [DOI: 10.3724/SP.J.1089.2021.18574]

Cheraghian A, Rahman S, Campbell D and Petersson L. 2019. Mitigating the hubness problem for zero-shot learning of 3D objects//Proceedings of the 30th British Machine Vision Conference. Cardiff, UK: BMVA: #41

Fan Y F, Bai J, Shao H H and Peng B. 2024. Discriminative feature-

guided zero-shot learning of 3D model classification. Journal of Computer-Aided Design and Computer Graphics, 36(2): 223-235 (范有福, 白静, 邵会会, 彭斌. 2024. 判别性特征引导的零样本三维模型分类算法. 计算机辅助设计与图形学学报, 36(2): 223-235) [DOI: 10.3724/SP.J.1089.2024.19715]

Feng Y G, Yu J, Sang J T and Yang P B. 2021. Survey on knowledge-based zero-shot visual recognition. Journal of Software, 32(2): 370-405 (冯耀功, 于剑, 桑基韬, 杨朋波. 2021. 基于知识的零样本视觉识别综述. 软件学报, 32(2): 370-405) [DOI: 10.13328/j.cnki.jos.006146]

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2020. Generative adversarial networks. Communications of the ACM, 63(11): 139-144 [DOI: 10.1145/342262]

Guo Z Y, Zhang R R, Qiu L T, Ma X Z, Miao X P, He X M and Cui B. 2023. CALIP: zero-shot enhancement of clip with parameter-free attention//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press: 746-754 [DOI: 10.1609/aaai.v37i1.25152]

Hao Y, Su Y K, Lin G S, Su H J and Wu Q Y. 2023. Contrastive generative network with recursive-loop for 3D point cloud generalized zero-shot classification. Pattern Recognition, 144: #109843 [DOI: 10.1016/j.patcog.2023.109843]

Johns E, Leutenegger S and Davison A J. 2016. Pairwise decomposition of image sequences for active multi-view recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 3813-3822 [DOI: 10.1109/cvpr.2016.414]

Kingma D P and Welling M. 2014. Auto-encoding variational Bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR: 14-16 [DOI: 10.48550/arXiv.1312.6114]

Kodirov E, Xiang T and Gong S G. 2017. Semantic autoencoder for zero-shot learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE Computer Society Press, 4447-4456 [DOI: 10.1109/cvpr.2017.473]

Lei Y J, Xu K, Guo Y L, Yang X, Wu Y W, Hu W, Yang J Q and Wang H Y. 2024. Comprehensive survey on 3D visual-language understanding techniques. Journal of Image and Graphics, 29(6): 1747-1764 (雷印杰, 徐凯, 郭裕兰, 杨鑫, 武玉伟, 胡玮, 杨佳琪, 汪汉云. 2024. "三维视觉—语言"推理技术的前沿研究与最新趋势. 中国图象图形学报, 29(6): 1747-1764) [DOI: 10.11834/jig.240029]

Li J J, Jin M M, Lu K, Ding Z M, Zhu L and Huang Z. 2019. Leveraging the invariant side of generative zero-shot learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7394-7403 [DOI: 10.1109/cvpr.2019.00758]

Lian Z, Zhang J, Choi S, ElNaghy H, El-Sana J, Furuya T, Giachetti A, Guler R A, Lai L, Li C, Li H, Limberger F A, Martin R, Nakanishi R U, Neto A P, Nonato L G, Ohbuchi R, Pevzner K, Pickup D, Rosin P, Sharf A, Sun L, Sun X, Tari S, Unal G and Wilson R C. 2015. Non-rigid 3D shape retrieval//Proceedings of 2015 Eurographics Workshop on 3D Object Retrieval. Goslar, Germany: Eurographics Association, 107-120 [DOI: 10.2312/3dor.20151064]

Liu A A, Su Y T, Wang L J, Li B, Qian Z X, Zhang W M, Zhou L N, Zhang X P, Zhang Y D, Huang J W and Yu N H. 2024. Review on the progress of the AIGC visual content generation and traceability. Journal of Image and Graphics, 29(6): 1535-1554 (刘安安, 苏育挺, 王岚君, 李斌, 钱振兴, 张卫明, 周琳娜, 张新鹏, 张勇东, 黄继武, 俞能海. 2024. AIGC视觉内容生成与溯源研究进展. 中国图象图形学报, 29(6): 1535-1554) [DOI: 10.11834/jig.240003]

Long X X, Cheng X J, Zhu H, Zhang P J, Liu H M, Li J, Zheng L T, Hu Q Y, Liu H, Cao X, Yang R G, Wu Y H, Zhang G F, Liu Y B, Xu K, Guo Y L and Chen B Q. 2021. Recent progress in 3D vision. Journal of Image and Graphics, 26(6): 1389-1428 (龙霄潇, 程新景, 朱昊, 张朋举, 刘浩敏, 李俊, 郑林涛, 胡庆拥, 刘浩, 曹汛, 杨睿刚, 吴毅红, 章国锋, 刘烨斌, 徐凯, 郭裕兰, 陈宝权. 2021. 三维视觉前沿进展. 中国图象图形学报, 26(6): 1389-1428) [DOI: 10.11834/jig.210043].

Ma C, Guo Y L, Yang J G and An W. 2019. Learning multi-view representation with LSTM for 3-D shape recognition and retrieval. IEEE Transactions on Multimedia, 21(5): 1169-1182 [DOI: 10.1109/TMM.2018.2875512]

Qi C R, Su H, Mo K C and Guibas L J. 2017a. 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 Computer Society Press: 77-85 [DOI: 10.1109/cvpr.2017.16]

Qi C R, Yi L, Su H and Guibas L J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5105-511 [DOI: 10.5555/3295222.3295263]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Schönfeld E, Ebrahimi S, Sinha S, Darrell T and Akata Z. 2019. Generalized zero- and few-shot learning via aligned variational autoencoders//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE Computer Society Press: 8239-8247 [DOI: 10.1109/cvpr.2019.00844]

Siddiqi K, Zhang J, Macrini D, Shokoufandeh A, Bouix S and Dickinson S. 2008. Retrieving articulated 3-D models using medial sur-

faces. Machine Vision and Applications, 19(4): 261-275 [DOI: 10.1007/s00138-007-0097-8]

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 Computer Society: 945-953 [DOI: 10.1109/iccv.2015.114]

Uy M A, Pham Q H, Hua B S, Nguyen T and Yeung S K. 2019. Revisiting point cloud classification: a new benchmark dataset and classification model on real-world data//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE Computer Society: 1588-1597 [DOI: 10.1109/iccv.2019.00167]

Vyas M R, Venkateswara H and Panchanathan S. 2020. Leveraging seen and unseen semantic relationships for generative zero-shot learning//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 70-86 [DOI: 10.1007/978-3-030-58577-8_5]

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]

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 Computer Society: 1912-1920 [DOI: 10.1109/cvpr.2015.7298801]

Xian Y Q, Lorenz T, Schiele B and Akata Z. 2018. Feature generating networks for zero-shot learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE Computer Society Press: 5542-5551 [DOI: 10.1109/cvpr.2018.00581]

Zhang R R, Guo Z Y, Zhang W, Li K C, Miao X P, Cui B, Qiao Y, Gao P and Li H S. 2022. PointCLIP: point cloud understanding by CLIP//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE Computer Society: 8542-8552 [DOI: 10.1109/cvpr52688.2022.00836]

Zhu X Y, Zhang R R, He B W, Guo Z Y, Zeng Z Y, Qin Z P, Zhang S H and Gao P. 2023. PointCLIP v2: prompting CLIP and GPT for powerful 3D open-world learning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE Computer Society: 2639-2650 [DOI: 10.1109/iccv51070.2023.00249]

作者简介

晏浩，男，硕士研究生，主要研究方向为零样本学习。
E-mail: yanhao@stu.nmu.edu.cn
白静，通信作者，女，教授，主要研究方向为计算机辅助设计与图形学、机器学习。E-mail: baijing@nun.edu.cn
郑虎，男，硕士研究生，主要研究方向为图像处理与计算机视觉。E-mail: zhenghu@stu.nmu.edu.cn