中图法分类号:TP18; TP391.41　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)05-1583-12

论文引用格式:Li Z W, Guo H N, Qi Y P and Yuan D K. 2026. Zooplankton classification based on multiscale feature fusion and attention guidance. Journal of Image and Graphics, 31(5):1583-1594(李忠伟, 郭浩宁, 齐衍萍, 袁德坤. 2026. 多尺度特征融合与注意力引导的浮游动物分类. 中国图象图形学报, 31(5):1583-1594)[DOI:10.11834/jig.250352]

多尺度特征融合与注意力引导的浮游动物分类

李忠伟[1],郭浩宁[2*],齐衍萍[3],袁德坤[1]

1. 中国石油大学(华东)海洋与空间信息学院,青岛　266580; 2. 中国石油大学(华东)青岛软件学院、计算机科学与技术学院,青岛　266580; 3. 国家海洋局北海环境监测中心(自然资源部北海生态中心),青岛　266033

摘　要: **目的**　浮游动物是海洋生态系统的关键群体,其变化反映海洋生态健康并服务于赤潮预警、渔业评估和碳循环研究。传统人工镜检效率低且主观性强,难以满足大规模监测需求,自动化识别因此成为必然趋势。然而,图像背景复杂、目标微小等因素导致现有方法精度与鲁棒性不足,亟需更高辨识能力与更强适应性的模型。**方法**　构建了一种融合多尺度空洞卷积与双重注意力机制的ViT-MDFA(vision Transformer based on multi-scale dilated convolution and dual attention fusion architecture)模型。模型基于ViT(vision Transformer)主干,引入多尺度空洞卷积模块增强局部结构感知能力,加入通道—空间注意力机制突出关键区域表达,采用交替插入策略实现局部增强与全局建模协同优化。模型适用于不同分辨率和背景复杂度的数据样本,并在典型生态监测场景开展评估。**结果**　在WHOI-Plankton、ZooScanNet、Kaggle-Plankton和自建Dec-22等4个浮游动物图像数据集上,该模型的分类准确率分别达到了92.27%、93.34%、96.14%和97.46%,在与其他8种方法的对比中均取得最优结果。消融实验表明,多尺度感知与注意力机制均对性能提升具有显著贡献,联合使用效果最佳。可视化分析显示,该模型的注意力热图更稳定地聚焦于目标关键结构,鲁棒性和收敛效率优于对比方法。**结论**　所提出的ViT-MDFA模型在浮游动物图像识别任务中表现优异,适用于图像质量波动大且背景复杂的海洋生态监测场景。模型结构轻量、模块化强,便于部署于流式细胞仪、边缘节点等平台,为构建智能化、自动化的浮游动物识别系统提供了关键支撑。

关键词: 浮游动物;细粒度图像分类;Vision Transformer;空洞卷积;通道—空间双重注意力

Zooplankton classification based on multiscale feature fusion and attention guidance

Li Zhongwei[1], Guo Haoning[2*], Qi Yanping[3], Yuan Dekun[1]

1. College of Oceanography and Space Information, China University of Petroleum (East China), Qingdao　266580, China; 2. Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao　266580, China; 3. North China Sea Environmental Monitoring Center, State Oceanic Administration, Qingdao　266033, China

Abstract: **Objective** Zooplankton are a key component of the marine ecosystem, the basis of the marine food web, and a sensitive indicator of environmental change. Changes in its population composition and quantity reflect water quality, primary productivity, and ecological balance, which are significant for marine ecological assessment and red tide prediction. The traditional manual microscope inspection method is time consuming, labor consuming, and subjective, hardly meeting

收稿日期:2025-08-03;**修回日期**:2025-11-03;**预印本日期**:2025-11-10

*通信作者:郭浩宁　ghn_20010304@qq.com

基金项目:山东省自然科学基金项目(ZR2024MF037)

Supported by: Natural Science Foundation of Shandong Province, China(ZR2024MF037)

the requirements of large-scale real-time monitoring. With the development of intelligent perception and deep learning, image-based automatic recognition has become an important means of marine ecological research. However, the fine-grained classification of zooplankton still faces challenges, including high similarity between species, significant differences within species, blurred or occluded targets in microscopic images, and considerable scale differences among various species. Therefore, this study proposes a Transformer-based zooplankton recognition method that integrates multiscale features and an attention guidance mechanism. The method aims to enhance the network's ability to perceive and fuse local details and global semantics, thereby improving classification accuracy, model robustness, and interpretability. **Method** In this study, a vision Transformer (ViT) is adopted as the backbone network, and two modules, namely, multiscale dilated convolution (MSDC) and dual attention (DA), are introduced to optimize the network structure. The MSDC module extracts local details and global contour information simultaneously through parallel MSDCs with different dilation rates and expands the receptive field without significantly increasing the computation amount. The DA mechanism consists of channel attention and spatial attention, which are used for weight distribution of feature channels and spatial guidance of significant regions, respectively. Meanwhile, an alternate insertion strategy is adopted, where a DA module is embedded every three layers. This strategy enhances the texture discrimination ability in the shallow layers and the semantic focusing and interclass discrimination abilities in the deep layers. These two modules complement each other: the MSDC module provides multiscale perception capabilities, while the DA module selectively enhances key features, thus improving the expression and discriminative abilities of features. The overall network structure is denoted as ViT-MDFA, and the model is trained and validated on four datasets, namely, WHOI-Plankton, ZooScanNet, Kaggle-Plankton, and the self-built Dec-22. Several ablation experiments are designed to verify the effectiveness of the proposed method. The effectiveness of the MSDC and DA modules is verified through module independence and synergy experiments; the scientificity of the alternate insertion strategy is verified via attention insertion strategy experiments; and the impact of different dilation rate combinations on model performance is verified by conducting different dilation rate combination experiments and scale sensitivity analysis experiments. Finally, the role of the attention mechanism in the image understanding process of the proposed model is explored through visual analysis. **Result** Experimental results show that the classification accuracy and F1-score of the proposed model on the four datasets are higher than those of the existing mainstream convolutional and Transformer-based models. On the Dec-22 dataset, the model achieves an accuracy of 97.4% and an F1-score of 96.7%, which are 3.9% and 4% higher than those of ViT-B/16, respectively. Ablation experiments demonstrate that when only the MSDC module is used, the model accuracy can be improved by approximately 1.5%; when only the DA module is used, the model accuracy can be improved by approximately 0.7%. The combination of these two modules can further enhance the overall performance of the model. The void rate sensitivity experiment shows that the [6, 12, 18] combination exhibits the highest stability across different datasets. In the scale grouping experiment, the small target has the best effect at a low hole rate, the medium target is balanced at a medium hole rate, and the large target has the highest performance at a high hole rate, which verifies the rationality and universality of the [6, 12, 18] configuration. The visualization results show that the attention map generated by the DA module focuses on the main body of zooplankton rather than the background impurities, and the model converges fast and has low variance, significantly improving the interpretability. **Conclusion** The Transformer zooplankton recognition framework, which integrates multiscale hole convolution and DA mechanism, achieves a balance between local detail extraction and global semantic understanding and significantly improves the classification accuracy and generalization ability in complex background. The main contributions of this study are as follows: 1) A new model, ViT-MDFA, is proposed. For the first time, MSDC and channel-spatial DA are embedded into a ViT simultaneously to address the issues of large scale differences, complex backgrounds, and high interclass similarity in zooplankton images. 2) A lightweight collaborative module is designed. The MSDC module pre-enhances multiscale local features, and the DA module is alternately inserted to enhance the focus on key regions. These two modules complement each other, enhancing fine-grained discriminative capabilities and ensuring efficient computation. 3) A performance breakthrough is achieved. The model achieves state-of-the-art accuracy on four mainstream datasets, and ablation experiments and visualization results verify the effectiveness of the modules. This model provides a plug-and-play solution for edge device deployment and intelligent marine ecological monitoring. Future work will explore an adaptive dilation rate learning mechanism

based on gradient optimization to achieve dynamic receptive field adjustment. Model lightweight technology and knowledge distillation strategies will be combined to support the real-time application of marine online monitoring equipment. Multi-modal data (e. g. , environmental parameters and time series data) can be introduced to expand the model's application potential in ecological prediction and biodiversity research.

Key words: zooplankton; fine-grained image classification; vision Transformer; dilated convolution; channel-spatial dual attention

0 引 言

随着“智慧海洋”与“数字海洋”战略推进，信息技术正加速赋能海洋生态监测。浮游动物作为海洋食物链中的关键环节，其种群变化不仅反映水体生态状态，还与渔业资源评估、生物多样性保护、赤潮/藻华预警、生物污染防控及海洋碳汇研究密切相关（Xu 等，2023；Kyathanahally 等，2021）。传统的显微镜人工鉴定方法存在效率低、成本高、主观性强等不足，亟需借助自动化识别技术提升生态监测的智能化水平。

然而，如图 1 和图 2 所示，浮游动物图像存在尺度差异大、边缘模糊、类间相似度高与背景干扰强等挑战，导致现有方法识别精度低、泛化能力弱，难以支撑大规模连续性生态监测任务（Maracani 等，2023）。提升浮游动物图像自动识别的智能化水平，是突破智慧海洋建设关键技术瓶颈的核心需求。

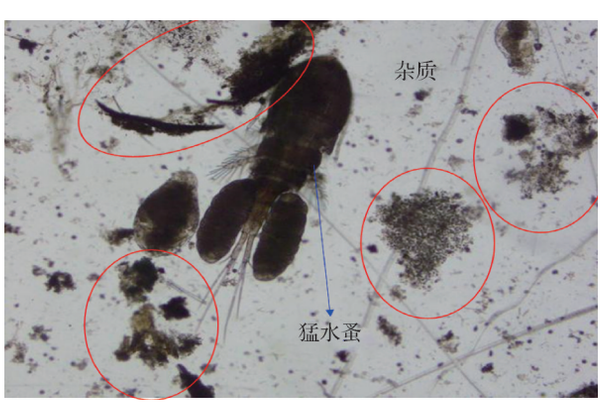
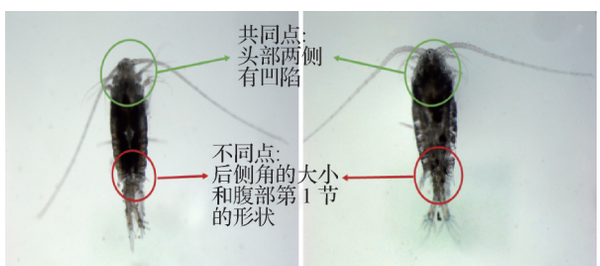


图 1　杂质等冗余信息干扰下的浮游动物图像

Fig. 1　Image of zooplankton under the interference of redundant information such as impurities

深度学习在计算机视觉任务中的广泛应用，为浮游动物显微图像的自动识别提供了可行方案。在众多模型架构中，ViT（vision Transformer）模型（Dosovitskiy 等，2021）依托全局自注意力机制在图像分类中取得突破。然而，ViT 设计之初将图像划分为固定的图像块，破坏了固有的对象结构，减少了输入图像块提供的信息量，导致模型聚焦背景干扰分类（石争浩 等，2023），使其在细粒度图像分类（fine-grained visual classification，FGVC）任务中常面临判别区域模糊、类内差异小和局部特征易丢失等挑战。为此，大量研究对 ViT 结构进行了针对性改进。例如，TPSKG（Transformer with peak suppression and knowledge guidance）（Liu 等，2022）设计了注意力抑制机制，通过弱化高响应特征单元增强次要区域表征，同时引入可配置知识嵌入模块提升特征表达能力；SM-ViT（salient mask-guided vision Transformer）（Demidov 等，2023）引入分层多尺度注意力，增强模型对不同尺度特征的适应性；IELT（internal ensemble learning Transformer）（Xu 等，2023）则结合局部特征增强与全局建模，提升了显微图像的特征表达能力；AA-Trans（attention aggregating Transformer）（Wang 等，2023）通过聚合各层注意力权重，生成更优注意力图。



太腹针胸刺水蚤　　瘦尾胸刺水蚤

图 2　相似物种形态特征比较

Fig. 2　Comparison of morphological characteristics of similar species

ViT 在 FGVC 领域的研究进一步向可解释性与训练目标优化方向延伸。CLE-ViT（contrastive learning encoded Transformer）（Yu 等，2023）通过引入自监督对比学习模块，在特征空间中拉开类间距离、保持类内一致性，显著提高了 FGVC 场景的泛化性能，并在多种细粒度数据集上取得领先表现。INTR（interpretable Transformer）（Paul 等，2024）则提出类

特定查询令牌与交叉注意力机制，使每个类别能够主动定位图像判别区域，从而在保证精度的同时赋予模型较高的可解释性。这些工作表明，结合判别能力增强与透明性提升是改善FGVC模型性能的重要途径。

基于上述背景，本文提出一种融合多尺度感知与注意力协同机制的ViT-MDFA（vision Transformer based on multi-scale dilated convolution and dual attention fusion architecture）模型。与仅依赖自注意力实现全局建模的传统ViT模型不同，本研究在模型输入端引入多尺度空洞卷积（multi-scale dilated convolution，MSDC）模块，使网络在图像块嵌入前即可感知不同尺度的局部纹理与整体形态信息；同时设计通道—空间双重注意力（dual attention，DA）模块，并采用交替插入策略，在不同编码层间实现浅层纹理与深层语义的互补增强。两者的协同作用显著提升了模型对关键判别区域的聚焦能力与整体分类性能。

本文的主要创新点如下：1）提出MSDC模块，通过多空洞率并行卷积与全局上下文融合，实现对浮游动物局部细节与整体轮廓的同步建模，从输入端增强模型的多尺度感知能力；2）设计DA模块，并提出交替插入策略，在保证计算量可控的前提下实现浅层与深层特征的判别协同，有效强化了模型的局部聚焦能力；3）提出融合MSDC与DA的ViT-MDFA模型，在结构层面实现多尺度特征与注意力机制的有机统一，显著提升浮游动物细粒度分类的准确性与泛化能力。

1 相关工作

1.1 细粒度图像分类方法概述

FGVC旨在区分外观高度相似的子类别，例如鸟类（Van Horn等，2015）、花卉犬类（Khosla等，2011）、车辆（Krause等，2013）等。早期方法多依赖手工特征（如尺度不变特征变换（scale-invariant feature transform，SIFT）、方向梯度直方图（histogram of oriented gradients，HOG）及部件检测提取判别性区域，但其特征表达能力有限且依赖复杂预处理与人工标注，难以适应多样化场景。随着深度学习的发展，基于卷积神经网络（convolutional neural network，CNN）的FGVC方法成为主流，利用多分支网络、特征金字塔和注意力机制强化细节特征捕捉。然而，CNN的局部感受野限制了其长程依赖建模能力，导致全局与局部信息协同建模不足，限制了识别效果。赵勋等人（2021）提出一种互补注意力机制，旨在得到高效的特征表示，但受网络结构限制难以得到更加高效的判别性特征。

1.2 Vision Transformer在FGVC中的应用

在FGVC任务中，标准ViT将图像划分为固定大小的图像块，并通过全局自注意力建模长程依赖，但单尺度建模在应对显微图像中尺度变化大、细节稀疏时存在不足。为此，研究者提出多种结构优化方法，其中多尺度特征建模较为常见，如在Patch Embedding前或编码器内部引入多尺度卷积、空洞卷积或特征金字塔结构，以生成不同感受野的令牌表征并融合。RAMS-Trans（recurrent attention multi-scale Transformer）（Hu等，2021）通过递归注意实现多尺度特征生成，SM-ViT（Demidov等，2023）采用跨尺度令牌融合增强尺度适应性，PVT（pyramid vision Transformer）（Wang等，2021）则利用分层结构和递进式令牌降采样降低计算开销。另有方法结合局部与全局注意力，如IELT（Xu等，2023）在全局注意力层间插入局部特征增强模块，AA-Trans（Wang等，2023）通过通道与空间双重注意力强化判别区域响应。针对推理阶段的背景冗余问题，TSVT（token-selective vision Transformer）（Si等，2023）在多层中动态筛选判别性令牌，从而在降低计算量的同时抑制干扰。这些方法在多尺度建模、局部与全局协同及冗余抑制方面显著提升了ViT在FGVC中的表现，但仍面临显存消耗高、参数量增加以及局部与全局平衡难的问题。

1.3 基于训练策略优化的FGVC方法

除结构改进外，另一类方法从训练目标与推理策略入手提升性能。这类方法通常保持主干结构不变，通过优化特征分布或增强判别区域定位来改善效果。CViT-FDRM（CNN-Vision Transformer based fish disease recognition model）（魏立明 等，2023）利用CNN提取图像细粒度特征，采用Transformer模型自注意力机制获取图像全局信息进行并行训练，取得了较好的效果。TransFG（Transformer architecture for fine-grained recognition）（He等，2022）通过扩大混淆特征距离强化判别特征，CLE-ViT（contrastive learning encoded Transformer for ultra-fine-grained visual categorization）（Yu等，2023）在编码器输出后

加入自监督对比学习模块，利用扰动生成的正负样本对构建对比损失，与交叉熵损失联合优化，从而显式拉大类间距离并压缩类内差异，在FGVC场景下显著提升泛化性能。INTR（interpretable Transformer）（Paul等，2024）聚焦可解释性，采用类特定查询令牌与交叉注意力机制，使每个类别在特征图上主动定位判别区域并生成可视化热力图，从而提升透明度并增强对微小目标的识别能力。这类方法从特征分布优化和决策可解释性出发，为FGVC提供了新思路，但在背景复杂、尺度多样的显微图像中，单独依赖这些策略仍不足以全面提升多尺度细节表征能力，需与结构改进结合以发挥更大优势。

综上，现有基于ViT的FGVC方法在全局建模和长程依赖捕捉方面具有优势，但普遍在多尺度局部细节感知上存在不足。本文提出的ViT-MDFA模型在ViT基础上引入多尺度空洞卷积与通道—空间双重注意力机制，实现了输入端的多尺度局部增强与编码器阶段的全局建模协同，有效提升了细粒度识别的准确性与鲁棒性。

2　基于多尺度空洞感知与注意力引导的ViT-MDFA模型

2.1　模型整体结构

针对海洋浮游动物图像中存在的尺度差异大、边缘模糊、类间相似度高与背景干扰强等挑战，本文提出一种融合多尺度感知与双重注意力机制的ViT-MDFA模型，增强在复杂生态图像下的判别和泛化能力。该模型以ViT为主干，在输入阶段引入MSDC模块增强局部结构建模能力；在中后段引入DA模块强化判别区域感知；采用交替插入策略实现局部增强与全局建模的协同优化。模型整体结构图及各模块示意图如图3所示。

模型整体结构图如图3（a）所示。流程包括：1）原始图像经MSDC模块提取并融合多尺度局部语义特征；2）特征送入ViT编码器进行全局建模与长程依赖捕捉；3）在每3个Transformer层后串联1个DA模块，引导模型关注关键区域与显著结构，提升识别性能。该结构兼顾模型性能与轻量化需求，适合集成于生态监测平台的浮游动物智能识别系统。

2.2　多尺度空洞卷积模块（MSDC）

浮游动物图像常存在器官细节模糊、尺度变化剧烈等问题，致使传统卷积层难以提取稳定且具有判别力的局部特征。标准ViT模型虽然具备全局建模能力，但编码初期缺乏细节敏感性。为增强模型的局部感知能力，设计MSDC模块作为前置特征提取单元，通过多路并行的空洞卷积操作，实现对不同空间感受野范围内结构的建模，既能捕捉微小边缘细节，也能覆盖整体形态轮廓。

MSDC模块结构如图3（b）所示，包含5个并行分支。1）第1分支采用1×1标准卷积，保留原始特征；2）第2—4分支分别采用扩张率为6、12和18的3×3空洞卷积，用于提取中等至大感受野下的上下文信息，卷积输出经过批归一化及线性整流函数（rectified linear unit，ReLU）激活；3）第5分支采用全局平均池化提取全图上下文语义，经1×1卷积和上采样操作恢复至原空间尺寸。其中，低空洞率分支捕捉局部纹理细节，中空洞率分支感知中等尺度形态变化，高空洞率分支建模整体轮廓与边界结构，全局分支通过自适应平均池化获取全局背景信息。各分支输出经归一化后在通道维度拼接，并通过1×1卷积融合，得到兼具局部与全局信息的多尺度特征图。

相较于编码器内部嵌入多尺度模块，前置MSDC在令牌生成阶段即丰富了输入特征的尺度多样性，为后续Transformer的注意力建模提供更具判别性的初始表示。同时，空洞卷积在扩展感受野的同时不会增加卷积核参数量，从而在轻量化的条件下提升模型对不同形态与尺寸浮游动物的适应性。

2.3　空洞率设计原理与感受野分析

空洞率直接决定卷积核的有效感受野，从而影响模型对不同尺度目标的特征覆盖范围。设卷积核大小为k，空洞率为r，则感受野近似为

$$R = k + (k - 1) \times (r - 1) \tag{1}$$

当输入图像大小为256×256像素、patch大小为16时，普通3×3卷积的感受野约为3，而空洞率为6、12、18时，其感受野分别约为13、25、37，对应浮游动物图像中从局部纹理（小体型个体）到整体轮廓（大体型个体）的不同尺度范围。

因此，多空洞率并行分支的设计使MSDC模块能够同时捕捉不同尺度的判别信息，天然具备尺度鲁棒性。为了验证不同空洞率组合在不同目标大小样本上的适应性，本文进一步开展了分尺度实验分析（见4.3.3节）。

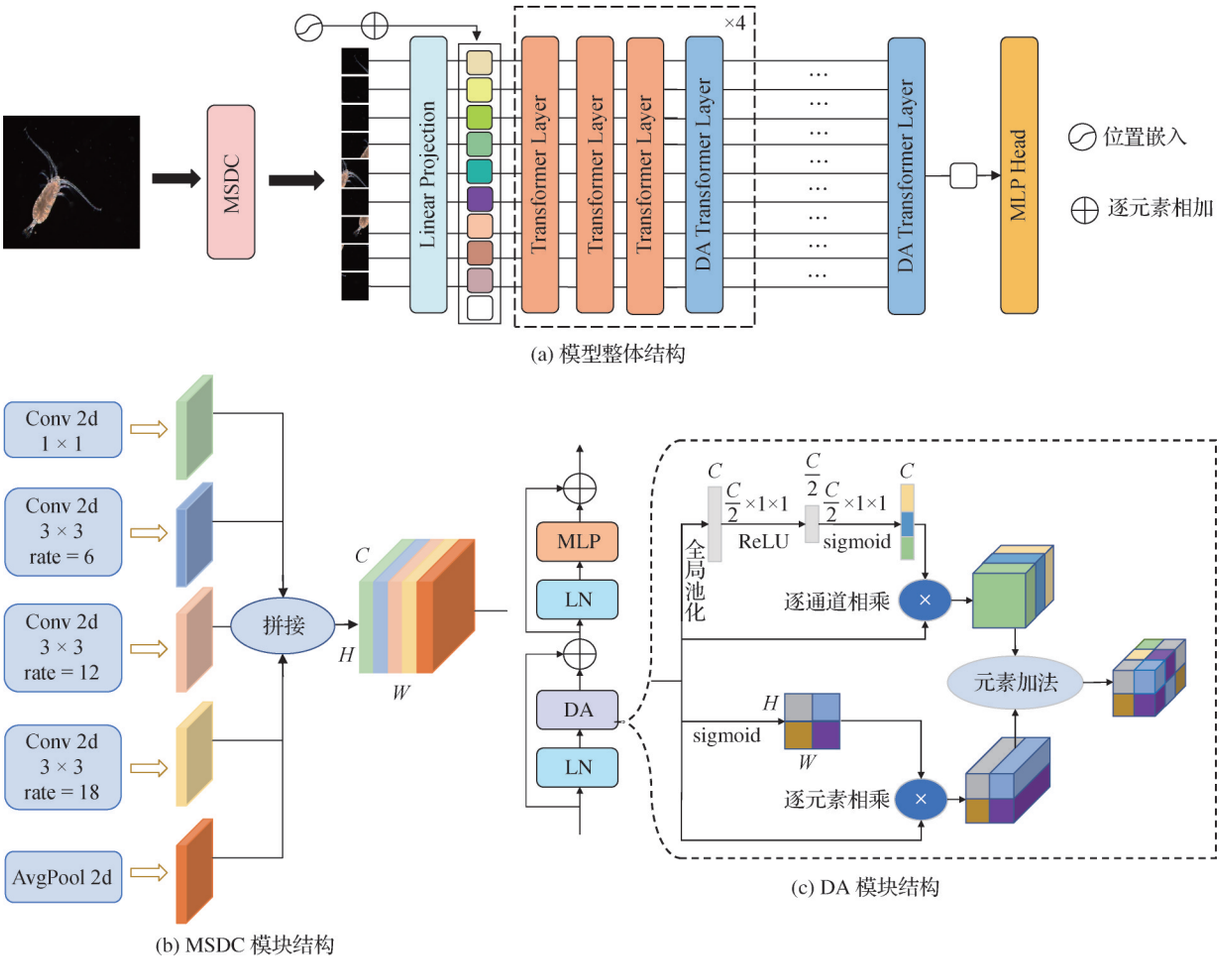


图3　ViT-MDFA整体结构图

Fig. 3　Overall structure diagram of ViT-MDFA

((a) overall structure of ViT-MDFA; (b) MSDC module structure; (c) DA module structure)

2.4　通道—空间双重注意力模块(DA)

生态场景中,浮游动物图像常因运动模糊、水体噪声以及光照波动等因素导致显著性区域不清晰,使模型难以准确聚焦于关键判别结构。在Transformer的多层编码结构中,浅层特征更关注边缘与纹理,深层特征则更偏向全局语义。若在每层均引入相同的注意力增强机制,容易造成特征冗余与梯度耦合。为此,本文提出DA模块,并采用交替插入策略,以实现判别信息的层间协同与动态强化。

DA模块结构如图3(c)所示,包含两个子路径。1)通道注意力路径:通过全局平均池化计算每个通道的统计描述子,经由两层卷积与ReLU激活后,上采样至原图尺寸,形成通道注意力图。原始特征图与通道注意力图逐通道相乘,完成通道增强,突出与类别强相关的特征通道。2)空间注意力路径:直接对输入特征施加卷积运算,生成空间注意力图,反映图像中不同位置的重要性。该注意力图与输入特征逐像素相乘,实现空间增强,引导模型聚焦于关键区域,如触角、尾刺等细小结构。前者通过全局平均池化与多层感知器(multi-layer perceptron,MLP)生成通道重要性权重,从通道维度抑制冗余信息;后者采用3×3卷积获取空间响应图,突出关键区域特征。两者串联可同时在通道和空间层面强化判别区域。在编码器中,采用交替插入策略,每隔3层嵌入一个DA模块,从而在浅层增强纹理判别能力,在深层强化语义聚焦与类间区分度。相较于全层插入方式,交替插入策略在性能提升接近的情况下显著降低了计算开销,并避免了过度注意力叠加导致的特征抑制。

3 数据集与实验设置

3.1 数据集

为验证所提方法的有效性，实验选取4个数据集测试，涵盖3个公开数据集与一个自建细粒度数据集，覆盖不同采集设备、图像分布及类别结构，以全面评估模型的泛化能力。

1）WHOI-Plankton数据集。由美国伍兹霍尔海洋研究所（Woods Hole Oceanographic Institution，WHOI）基于成像流式细胞仪采集，含300万余幅灰度图像、100余类，类别不平衡且质量差异大，实验选取其中50类子集。

2）ZooScanNet数据集。由法国海洋开发研究院（Institut Francais de Recherche Pour I'exploitation de la mer，IFREMER）发布，基于ZooScan光学扫描设备采集，含约1.4万幅尺寸为512 × 512像素的灰度图像，覆盖80类，质量高且分布均衡，适合精度与跨域泛化分析。

3）Kaggle-Plankton数据集。源于Kaggle National Data Science Bowl，含约30万幅图像、121类，由IFCB（Imaging FlowCytobot）采集并经标准化处理，常与WHOI联合用于多源适应性研究。

4）Dec-22数据集。本研究自建，采集自黄渤海海域样本，经徕卡显微镜拍摄，含3 304幅图像、22类，主体清晰、类间相似度高，强调局部细节与微小形态识别能力，难度更高。

3.2 实验设置

实验中统一将图像输入尺寸设置为256 × 256像素，patch大小设置为16。所有模型默认加载在ImageNet-1K数据集上预训练的权重。考虑到各数据集图像尺寸不统一、样本量及细粒度难度存在差异，预处理过程中将灰度图转换为三通道RGB图像，并进行归一化处理。所有实验均在搭载NVIDIA GeForce RTX 3090 GPU的服务器上进行，使用PyTorch框架构建模型并完成训练。各数据集的训练参数设置如表1所示。每组实验均独立运行3次，并报告平均结果，以确保性能评估的稳定性与可靠性。

3.3 评价指标

为全面评估模型在各数据集上的分类性能，同时考虑到浮游动物分类任务中存在类别不平衡问题，本文将准确率和F1分数作为主要评价指标，并额外使用参数量（Params）和浮点运算次数（floating point operations，FLOPs）来衡量模型复杂度。

表1 各数据集训练参数设置

Table 1 Training parameter settings for each dataset

数据集	优化器	学习率	Batch size	训练轮数	权重衰减	Warmup 步数
WHOI-Plankton	AdamW	3.00E−05	64	50	0.010	500
ZooScanNet	AdamW	1.00E−04	64	60	0.005	500
Kaggle-Plankton	AdamW	5.00E−05	64	50	0.010	500
DEC-22	AdamW	1.00E−05	32	100	0.005	300

4 实验结果与分析

4.1 主干模型对比实验

为验证ViT-MDFA模型的有效性，与基于Transformer的细粒度分类模型对比，包括ViT（Vaswani等，2017）、RAMS-Trans（Hu等，2021）、TransFG（He等，2022）、AA-Trans（Wang等，2023）、SM-ViT（Demidov等，2023）、IELT（Xu等，2023）、CLE-ViT（Yu等，2023）和INTR（Paul等，2024）。实验结果如表2所示。

从表2可以看出，基于ViT的方法（INTR、CLE-ViT）相较于标准ViT在细粒度任务上有所改进，但本文提出的ViT-MDFA在所有数据集上仍保持领先。主要原因在于，INTR与CLE-ViT等方法虽然优化了Transformer的判别性，但大多并未在输入特征层显式增强多尺度局部结构感知。而浮游动物显微图像中含大量微小判别结构（触角、刚毛、体节等）与复杂背景噪声，仅依赖自注意力或训练目标调整难以充分强化这些局部细节。

ViT-MDFA通过在ViT前置引入MSDC模块并在编码器中插入DA模块，实现了“输入端多尺度局

表2　各模型在4个浮游动物数据集上的分类性能对比

Table 2　Comparison of classification performance of each model on four zooplankton datasets

/%

模型	WHOI-Plankton		ZooScanNet		Kaggle-Plankton		Dec-22	
	F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率
ViT-B/16	85.48	86.94	82.67	84.73	87.34	89.11	88.94	91.32
TransFG	86.92	87.89	84.25	86.01	89.58	91.13	90.37	92.41
RAMS-Trans	87.33	88.28	85.11	86.94	90.42	92.08	91.78	93.69
AA-Trans	88.69	89.57	86.92	88.80	91.48	93.82	93.86	95.17
SM-ViT	89.14	90.04	88.12	89.87	92.17	94.42	94.71	95.91
IELT	89.87	90.71	89.04	90.68	93.08	95.12	95.24	96.42
CLE-ViT	90.18	91.05	89.42	91.12	93.51	95.04	95.36	96.52
INTR	90.43	91.52	89.97	91.86	94.02	95.48	95.71	96.71
ViT-MDFA(本文)	**91.36**	**92.27**	**91.24**	**93.34**	**94.58**	**96.14**	**96.73**	**97.46**

注:加粗字体表示各列最优结果。

部增强+编码器全局建模”的协同,在细粒度判别与噪声鲁棒性上取得更明显的提升。此外,在少样本数据集(如Dec-22)中,CLE-ViT在部分情形可缓解过拟合,但ViT-MDFA的结构化多尺度增强在此类场景中同样表现稳健。

4.2　与典型浮游动物分类方法的对比实验

为进一步验证ViT-MDFA模型在浮游动物图像分类中的综合表现,选取Kaggle-Plankton和ZooScanNet两个主流数据集,与近年代表性研究方法进行性能对比,涵盖基于CNN与Transformer架构的多种建模方式。

4.2.1　Kaggle-Plankton数据集

表3列出了与Guo等人(2021)、Guo和Guan(2021)、Kyathanahally等人(2021)和Maracani等人(2023)所提方法的性能对比。早期方法,如Guo等人(2021)的准确率仅为77.45%;随着注意力机制与多尺度建模的引入,后续方法有所改进,但在微小结构与跨尺度特征建模上仍有欠缺。ViT-MDFA融合多尺度空洞卷积与双路协同注意力机制,在性能上取得领先,分类准确率达96.14%。

表3　Kaggle-Plankton数据集上的模型性能对比

Table 3　Comparison of model performance on Kaggle-Plankton dataset

方法	准确率/%
Guo等人(2021)	77.45
Guo和Guan(2021)	86.50
Kyathanahally等人(2021)	94.70
Maracani等人(2023)	95.50
ViT-MDFA(本文)	**96.14**

注:加粗字体表示最优结果。

4.2.2　ZooScanNet数据集

表4展示了ViT-MDFA在ZooScanNet数据集上与现有方法的对比情况。如Zheng等人(2017)方法主要采用传统CNN架构,在背景复杂、类间形态差异小的场景中分类准确率较低,仅为88.34%。引入注意力机制和深层特征建模后,Maracani等人(2023)方法表现提升。ViT-MDFA通过“多尺度感知+局部增强”结构进一步提升分类能力,在ZooScanNet上达到93.34%的准确率。

表4　ZooScanNet数据集上的模型性能对比

Table 4　Comparison of model performance on ZooScanNet dataset

方法	准确率/%
Zheng等人(2017)	88.34
Guo和Guan(2021)	86.70
Kyathanahally等人(2021)	89.80
Maracani等人(2023)	92.50
ViT-MDFA(本文)	**93.34**

注:加粗字体表示最优结果。

4.3 消融实验与模块分析

4.3.1 模块独立与协同效应验证

为验证各模块在模型中的作用，对MSDC、DA及其组合进行了消融分析。表5结果显示，单独引入MSDC能显著提升模型在复杂背景样本中的识别性能，说明前置多尺度特征建模有效增强了局部结构判别力；仅引入DA模块同样带来稳定提升，表明通道—空间双重注意力能改善模型对关键区域的聚焦能力。更重要的是，当MSDC与DA同时引入时，模型性能提升幅度超过两者单独使用的叠加效果，说明两模块之间存在显著的协同关系：MSDC与DA的结合实现了多尺度结构建模与注意力聚焦的协同优化。MSDC生成的多尺度结构差异为DA提供了更具可分性的特征分布，使注意力机制能更准确地聚焦于判别区域。

4.3.2 注意力插入策略验证

为验证交替式注意力增强结构“全局建模 + 局部增强”的有效性，在ViT编码器中分别采用3种策略进行对比。1）No-DA：不插入DA模块；2）All-DA：每层Transformer后均插入DA模块；3）Alt-DA：每隔3层插入1次DA模块（第3、6、9、12层之后）。结果如表6所示。Alt-DA策略在保持较低计算复杂度的同时，分类性能优于No-DA，且在性能接近All-DA的前提下将FLOPs从20.7 G降至18.4 G，参数量控制在91.2 M，显示出更优的性能—效率权衡。

表5 MSDC和DA模块对模型分类性能影响比较

Table 5 Comparison of the impact of MSDC and DA modules on model classification performance

/%

模型设置	MSDC	DA	WHOI-Plankton		ZooScanNet		Kaggle-Plankton		Dec-22	
			F1分数	准确率	F1分数	准确率	F1分数	准确率	F1分数	准确率
ViT-B/16 baseline	×	×	88.92	89.81	86.39	88.28	91.23	92.84	94.68	95.41
ViT + MSDC	√	×	90.72	91.64	88.25	90.17	93.56	95.17	96.11	96.94
ViT + DA	×	√	90.11	91.02	87.92	89.66	93.02	94.63	95.89	96.77
ViT + MSDC + DA (本文)	√	√	**91.36**	**92.27**	**91.24**	**93.34**	**94.58**	**96.14**	**96.73**	**97.46**

注：加粗字体表示各列最优结果。“√”表示使用该模块，“×”表示未使用该模块。

表6 不同DA插入策略下模型性能与计算复杂度对比

Table 6 Comparison of model performance and computational complexity under different DA insertion strategies

数据集	策略	F1分数/%	准确率/%	Params/M	FLOPs/G
WHOI-Plankton	No-DA	89.63	90.71	85.8	16.9
	All-DA	**91.75**	**92.68**	103.5	20.7
	Alt-DA	91.36	92.27	91.2	18.4
ZooScanNet	No-DA	88.57	90.12	85.8	16.9
	All-DA	**91.52**	**93.78**	103.5	20.7
	Alt-DA	91.24	93.34	91.2	18.4
Kaggle-Plankton	No-DA	92.03	94.18	85.8	16.9
	All-DA	**95.82**	**96.37**	103.5	20.7
	Alt-DA	94.58	96.14	91.2	18.4
Dec-22	No-DA	94.91	95.87	85.8	16.9
	All-DA	**97.83**	**98.21**	103.5	20.7
	Alt-DA	96.73	97.46	91.2	18.4

注：加粗字体表示各数据集在不同空洞率组合下的最优结果。“Alt-DA”为综合性能与效率的最优插入策略选择。

4.3.3　不同空洞率组合及尺度敏感性分析

为验证不同空洞率配置对模型性能的影响，本文在4个数据集上分别测试3组空洞率组合：[2,4,6]、[4,8,12]和[6,12,18]。整体实验结果如表7所示，[6,12,18]组合取得最高的平均精度与F1值，表明较大的感受野有助于捕获浮游动物的全局形态特征。

然而，不同类别浮游动物在显微图像中的实际尺寸差异显著，仅在整体样本上评估空洞率组合仍显片面。为此，本文进一步根据目标面积占比（目标像素数/图像总像素数）对样本进行尺度划分。结合样本面积分布特征，将数据集划分为小尺度（< 15%）、中尺度（15%～30%）与大尺度（≥ 30%）3类，并在各尺度组中分别评估3组空洞率组合的表现，结果如表8所示。该扩展实验仅针对Dec-22数据集展开，旨在揭示空洞率配置与目标尺度之间的对应关系。

实验结果表明，低空洞率组合[2,4,6]在小尺度组上取得最佳性能（准确率94.54%，F1分数94.20%），说明较小的感受野有助于提取微小个体的细节特征；中空洞率组合[4,8,12]在中尺度组表现最优（准确率95.15%，F1分数94.62%），实现了局部与全局特征的平衡；高空洞率组合[6,12,18]在大尺度组上性能最高（准确率95.91%，F1分数95.36%），表明较大感受野有助于建模整体轮廓结构。

从整体趋势来看，[6,12,18]组合在Dec-22数据集的综合表现最为稳健，与表7中的结论保持一致。统计检验结果表明，在小尺度与大尺度组上，不

表7　MSDC模块不同空洞率组合在各数据集的分类性能表现

Table 7　Classification performance of different void rate combinations of MSDC module on each dataset

数据集	空洞率组合	F1分数/%	准确率/%	Params/M	FLOPs/G
WHOI-Plankton	[2,4,6]	89.17	90.23	90.5	18.2
	[4,8,12]	90.52	91.78	91.0	18.3
	[6,12,18]	**91.36**	**92.27**	91.2	18.4
ZooScanNet	[2,4,6]	88.64	90.41	90.5	18.2
	[4,8,12]	90.27	91.93	91.0	18.3
	[6,12,18]	**91.24**	**93.34**	91.2	18.4
Kaggle-Plankton	[2,4,6]	92.46	94.29	90.5	18.2
	[4,8,12]	93.72	95.21	91.0	18.3
	[6,12,18]	**94.58**	**96.14**	91.2	18.4
Dec-22	[2,4,6]	94.83	95.76	90.5	18.2
	[4,8,12]	95.91	96.79	91.0	18.3
	[6,12,18]	**96.73**	**97.46**	91.2	18.4

注：加粗字体表示在各个数据集上不同空洞率组合下的最优结果。[6,12,18]为MSDC模块的默认空洞率配置。

表8　不同空洞率组合在不同尺度组上的性能对比（Dec-22数据集）

Table 8　Performance of different dilation rate combinations on different scale groups (Dec-22 dataset)

空洞率组合	小尺度（<15%）		中尺度（15%～30%）		大尺度（≥30%）	
	F1分数/%	准确率/%	F1分数/%	准确率/%	F1分数/%	准确率/%
[2, 4, 6]	**94.20**	**94.54**	92.83	93.61	91.95	92.89
[4, 8, 12]	91.50	92.17	**94.62**	**95.15**	93.47	93.96
[6, 12, 18]	90.69	91.38	93.84	94.38	**95.36**	**95.91**

注：加粗字体表示各列最优结果。

同空洞率组合的性能差异具有显著性。

综上，空洞率大小与目标尺度密切相关：小目标更依赖低空洞率以捕获局部细节，而大目标则受益于高空洞率以强化全局表征。考虑到浮游动物图像经过显微裁剪后目标占比普遍较高，固定的[6，12，18]组合在总体性能上表现稳健，适合作为默认配置，而在特定小目标占比高的任务中，可选择较低空洞率组合以进一步提升识别精度。

4.3.4　Grad-CAM 可视化分析

为深入探究模型在图像理解中的关注机制，本文基于 Grad-CAM 方法对 ViT-B/16 与 ViT-MDFA 模型的注意力响应进行了对比可视化。图 4 所示为 4 组具有代表性的浮游动物样本的热力图对比，每组包括原图、ViT-B/16 的注意力图和 ViT-MDFA 的注意力图 3 幅图像。

结果显示，ViT-MDFA 的热力图聚焦更为精准，能显著突出目标区域，并集中注意在附肢、末端触角、刚毛等具有判别性的结构特征上；而 ViT-B/16 的注意力区域则相对分散，存在较多无关背景区域的响应，导致模型对局部判别性结构的学习不足。

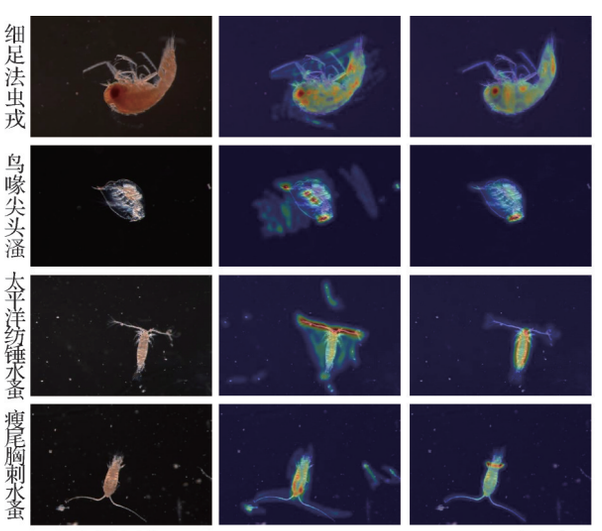


图 4　ViT-B/16 与 ViT-MDFA 对图像的注意力响应区域可视化对比

Fig. 4　Visual comparison of attention response regions of ViT-B/16 and ViT-MDFA to images

5　结　论

本文面向海洋浮游动物图像识别中存在的尺度不一、结构模糊、类间相似与背景复杂等关键挑战，提出一种融合多尺度空洞卷积与双重注意力机制的 ViT-MDFA 模型。通过在 Transformer 输入端引入 MSDC 模块，实现了局部与全局特征的多尺度感知；通过在编码器中交替插入 DA 模块，实现了不同层级间判别信息的动态协同。二者的有机结合不仅提升了模型在复杂背景和多尺度样本下的判别性能，也在语义聚焦与特征表达方面体现出良好的互补性。在 4 个具有代表性的浮游动物图像数据集上开展的实验证明，该方法在分类精度与区域关注能力方面均优于现有主流方法，特别是在图像质量波动大、目标边界不清晰的近岸观测类场景中展现出良好的鲁棒性。模型结构轻量、适配性强，具备部署于流式细胞仪、浮标监测系统等场景的实际应用潜力。

未来工作可进一步探索基于梯度优化的自适应空洞率学习机制，实现动态感受野调节；同时结合模型轻量化与知识蒸馏策略，以支持海洋在线监测设备的实时应用。未来还可引入多模态数据（如环境参数与时间序列分布）以拓展模型在生态预测与生物多样性研究中的应用潜力。

参考文献（References）

Demidov D，Sharif M H，Abdurahimov A，Cholakkal H and Khan F S. 2023. Salient mask-guided vision transformer for fine-grained classification//Proceedings of 2023 International Joint Conference on Computer Vision，Imaging and Computer Graphics Theory and Applications. Lisbon，Portugal：SciTePress-Science and Technology Publications，Lda：27-38［DOI：10.5220/0011611100003417］

Dosovitskiy A，Beyer L，Kolesnikov A，Weissenborn D，Zhai X H，Unterthiner T，et al. 2021. An image is worth 16 × 16 words：Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations Online OpenReview.net［DOI：10.48550/arXiv.2010.11929］

Guo C F，Wei B and Yu K. 2021. Deep transfer learning for biology cross-domain image classification. Journal of Control Science and Engineering，2021：#2518837［DOI：10.1155/2021/2518837］

Guo J and Guan J H. 2021. Classification of marine plankton based on few-shot learning. Arabian Journal for Science and Engineering，46(9)：9253-9262［DOI：10.1007/s13369-021-05786-2］

He J，Chen J N，Liu S，Kortylewski A，Yang C，Bai Y，et al. 2022. TransFG：a transformer architecture for fine-grained recognition//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto，USA：AAAI Press：852-860［DOI：10.1609/aaai.v36i1.19967］

Hu Y Q，Jin X，Zhang Y，Hong H W，Zhang J F，He Y，et al. 2021.

RAMS-Trans: recurrent attention multi-scale transformer for fine-grained image recognition//Proceedings of the 29th ACM International Conference on Multimedia. [s.l.]: ACM: 4239-4248 [DOI: 10.1145/3474085.3475561]

Khosla A, Jayadevaprakash N, Yao B P and Li F F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs//Proceedings of the 1st Workshop on Fine-Grained Visual Categorization Colorado Springs, USA, June 2011.

Krause J, Stark M, Deng J and Li F F. 2013. 3D object representations for fine-grained categorization//Proceedings of 2013 IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE: 554-561 [DOI: 10.1109/iccvw.2013.77]

Kyathanahally S P, Hardeman T, Merz E, Bulas T, Reyes M, Isles P, et al. 2021. Deep learning classification of lake zooplankton. Frontiers in Microbiology, 12: #746297 [DOI: 10.3389/fmicb.2021.746297]

Liu X D, Wang L L and Han X G. 2022. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. Neurocomputing, 492: 137-149 [DOI: 10.1016/j.neucom.2022.04.037]

Maracani A, Pastore V P, Natale L and Rosasco L and Odone F. 2023. In-domain versus out-of-domain transfer learning in plankton image classification. Scientific Reports, 13(1): #10443 [DOI: 10.1038/s41598-023-37627-7]

Paul D, Chowdhury A, Xiong X Q, Chang F J, Carlyn D E, Stevens S, et al. 2024. A simple interpretable transformer for fine-grained image classification and analysis//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net [DOI: 10.48550/arXiv.2311.04157]

Shi Z H, Li C J, Zhou L, Zhang Z J, Wu C W, You Z Z, et al. 2023. Survey on Transformer for image classification. Journal of Image and Graphics, 28(9): 2661-2692 (石争浩, 李成建, 周亮, 张治军, 仵晨伟, 尤珍臻, 等. 2023. Transformer驱动的图像分类研究进展. 中国图象图形学报, 28(9): 2661-2692) [DOI: 10.11834/jig.220799]

Si G Z, Xiao Y, Wei B, Bullock L B, Wang Y Y and Wang X D. 2023. Token-Selective Vision Transformer for fine-grained image recognition of marine organisms. Frontiers in Marine Science, 10: #1174347 [DOI: 10.3389/fmars.2023.1174347]

Van Horn G, Branson S, Farrell R, Haber S, Barry J, Ipeirotis P, et al. 2015. Building a bird recognition app and large scale dataset with citizen scientists: the fine print in fine-grained dataset collection//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 595-604 [DOI: 10.1109/cvpr.2015.7298658]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010 [DOI: 10.5555/3295222.3295349]

Wang Q, Wang J J, Deng H Y, Wu X, Wang Y Z and Hao G F. 2023. AA-trans: core attention aggregating transformer with information entropy selector for fine-grained visual classification. Pattern Recognition, 140: #109547 [DOI: 10.1016/j.patcog.2023.109547]

Wang W H, Xie E Z, Li X, Fan D P, Song K T, Liang D, et al. 2021. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 548-558 [DOI: 10.1109/ICCV48922.2021.00061]

Wei L M, Zhao K, Wang N, Zhang Z Y and Cui H P. 2023. Fine-grained fish disease image recognition algorithm model. Laser and Optoelectronics Progress, 60(16): #1610005. (魏立明, 赵奎, 王宁, 张忠岩, 崔海朋. 2023. 细粒度鱼类疫病图像识别算法模型. 激光与光电子学进展, 60(16): #1610005) [DOI: 10.3788/LOP222630]

Xu Q, Wang J H, Jiang B and Luo B. 2023. Fine-grained visual classification via internal ensemble learning transformer. IEEE Transactions on Multimedia, 25: 9015-9028 [DOI: 10.1109/tmm.2023.3244340]

Yu X H, Wang J and Gao Y S. 2023. CLE-ViT: contrastive learning encoded transformer for ultra-fine-grained visual categorization//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI: 4531-4539 [DOI: 10.24963/ijcai.2023/504]

Zhao X, Wang J B, Li Y, Wang Y P and Miao Z. 2021. Complemented attention method for fine-grained image classification. Journal of Image and Graphics, 26(12): 2860-2869 (赵勋, 王家宝, 李阳, 王亚鹏, 苗壮. 2021. 细粒度图像分类的互补注意力方法. 中国图象图形学报, 26(12): 2860-2869) [DOI: 10.11834/jig.200426]

Zheng H Y, Wang R C, Yu Z B, Wang N, Gu Z R and Zheng B. 2017. Automatic plankton image classification combining multiple view features via multiple kernel learning. BMC Bioinformatics, 18(S6): #570 [DOI: 10.1186/s12859-017-1954-8]

作者简介

李忠伟,男,教授,主要研究方向为数据智能处理及应用。
E-mail:lizhongwei@upc.edu.cn
郭浩宁,男,硕士研究生,主要研究方向为智能信息处理。
E-mail:ghn_2001@qq.com
齐衍萍,女,正高级工程师,主要研究方向为海洋生物多样性调查和海洋生态灾害监测。
E-mail:qiyanping@ncs.mnr.gov.cn
袁德坤,男,博士研究生,主要研究方向为计算机视觉和多模态信息处理。E-mail: ydk_libra0903@163.com