软件学报 ISSN 1000-9825, CODEN RUXUEW
2026,37(5):2286−2308 [doi: 10.13328/j.cnki.jos.007491] [CSTR: 32375.14.jos.007491]

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

基于视觉 Transformer 的双视图融合细粒度图像识别*

唐 昊, 李泽超, 蒋 鑫, 唐金辉

(南京理工大学 计算机科学与工程学院, 江苏 南京 210094)

通信作者: 李泽超, E-mail: zechao.li@njust.edu.cn

摘 要: 随着计算机视觉技术的不断进步, 细粒度图像识别在众多应用领域中发挥着重要作用. 与传统的粗粒度图像识别不同, 细粒度图像识别着重于在同一大类别下对具有细微视觉差异的子类别进行精确划分, 因此该任务更具有挑战性. 近年来, 视觉 Transformer 以其在全局上下文信息建模方面的出色表现而被广泛应用于图像识别领域. 然而, 当应用于细粒度图像识别任务时, 视觉 Transformer 在处理细节特征和背景噪声方面却存在一定的局限性. 针对上述问题, 提出一种基于视觉 Transformer 的双视图融合识别框架, 有效融合细粒度图像的全局视图与局部视图以提升识别准确率. 该框架设计了一个基于注意力融合的冗余信息过滤模块, 在编码器内部通过层级注意力权重的融合筛选图像块特征, 以优化全局视图的分类标记嵌入. 同时, 还设计了一个基于注意力阈值的关键区域定位模块, 通过自适应阈值策略动态选定并放大全局视图中的关键区域, 形成细致的局部视图以供再次分析. 此外, 所提出的局部区域特征自适应增强模块进一步增强了对局部细节的关注, 有效提升了细粒度特征的辨识能力. 为优化此双视图融合框架, 提出了基于双视图相似度的对比损失函数和基于双视图置信度的自适应推理策略, 旨在增强视觉 Transformer 模型输出的全局与局部特征辨识性, 同时有效节约计算资源并缩短推理时间. 在 CUB-200-2011、Stanford Dogs、NABirds 和 iNaturalist2017 这 4 个公共数据集上的实验结果表明, 该方法相较于传统视觉 Transformer 模型在识别准确率上实现了显著提升, 展示了其在细粒度图像识别任务中的有效性和优越性.

关键词: 视觉 Transformer; 细粒度图像识别; 注意力机制; 对比学习; 特征表示; 数据增强

中图法分类号: TP391

中文引用格式: 唐昊, 李泽超, 蒋鑫, 唐金辉. 基于视觉Transformer的双视图融合细粒度图像识别. 软件学报, 2026, 37(5): 2286–2308. http://www.jos.org.cn/1000-9825/7491.htm

英文引用格式: Tang H, Li ZC, Jiang X, Tang JH. Dual-view Fusion for Fine-grained Image Recognition with Vision Transformer. Ruan Jian Xue Bao/Journal of Software, 2026, 37(5): 2286–2308 (in Chinese). http://www.jos.org.cn/1000-9825/7491.htm

Dual-view Fusion for Fine-grained Image Recognition with Vision Transformer

TANG Hao, LI Ze-Chao, JIANG Xin, TANG Jin-Hui

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: With the continuous advancement of computer vision technology, fine-grained image recognition plays a crucial role across various application domains. Unlike traditional coarse-grained image recognition, fine-grained image recognition aims to precisely distinguish subcategories with subtle visual differences within the same major category, making this task particularly challenging. In recent years, the vision Transformer has gained widespread adoption in image recognition due to its exceptional performance in modeling global contextual information. However, the vision Transformer exhibits certain limitations when applied to fine-grained image recognition, particularly in processing detailed features and mitigating background noise. To address these issues, this study proposes a dual-view recognition framework based on the vision Transformer. This framework effectively integrates global and local views to enhance recognition accuracy. In this framework, an attention-based fusion module is designed to filter redundant information and optimize the classification token embedding of global views by merging and filtering patch features through hierarchical attention weights within the encoder. In addition, an attention threshold-based key region localization module is introduced. This module dynamically selects and

* 基金项目: 国家自然科学基金 (62425603, U21B2043); 江苏省基础研究计划攀登项目 (BK20240011)
收稿时间: 2024-10-15; 修改时间: 2025-03-11; 采用时间: 2025-06-12; jos 在线出版时间: 2025-10-29
CNKI 网络首发时间: 2025-10-31

magnifies key patches in the global view using an adaptive threshold strategy, forming detailed local views for further analysis. Furthermore, an adaptive enhancement module for local region features is proposed to strengthen the focus on local details, thus enhancing the recognition capability of fine-grained features. To optimize the dual-view framework, a contrastive loss function based on dual-view similarity and an adaptive inference strategy based on dual-view confidence are proposed. These strategies aim to enhance the global and local feature discriminability of the vision Transformer model while efficiently saving computational resources and shortening inference time. Experimental results on the CUB-200-2011, Stanford Dogs, NABirds, and iNaturalist2017 public datasets demonstrate that the proposed method achieves significant improvements in recognition accuracy compared to the traditional vision Transformer model. These results validate the proposed method's effectiveness and superiority in fine-grained image recognition tasks.

Key words: vision Transformer (ViT); fine-grained image recognition; attention mechanism; contrastive learning; feature representation; data augmentation

随着计算机视觉技术[1]的迅猛发展, 图像识别已成为众多应用领域的核心技术, 应用范围从日常生活的图像搜索和社交媒体标签, 扩展至更复杂的应用如医学图像分析和生物物种识别. 虽然常规图像识别已取得显著进步, 但细粒度图像识别[2,3]依然是一个颇具挑战性的研究领域. 不同于传统的粗粒度图像识别, 细粒度图像识别更加注重在基础图像分类之上进行更精细的子类划分, 因此它在实际应用中有着广泛的需求, 相关的研究领域包括生物多样性监测[4]、智慧交通[5]、智能零售[6]等. 然而, 正如图 1 所示, 细粒度图像识别面临着诸多复杂且具有挑战性的问题, 主要包括: (1) 类内差异较大, 类间差异微小: 细粒度图像识别的一个核心特征是目标类别间的差异通常非常细微, 而每个类别内部的差异则相对较大; (2) 标注难度大: 由于细粒度特征往往需要专业知识来识别, 获取高质量的注释数据变得异常困难和昂贵; (3) 数据稀缺: 细粒度图像识别常面临着标注数据不足的问题, 这限制了深度学习模型的性能和泛化能力. 鉴于传统的粗粒度图像识别模型难以应对复杂的细粒度图像识别任务, 因此, 实现低成本、高效率的细粒度图像识别对学术界和工业界具有极其重要的意义.

图 1 细粒度图像识别任务示意图

随着深度神经网络的快速发展, 通过卷积神经网络 (convolutional neural network, CNN)[7]提取的特征相较于传统手工特征展现出更强的表达能力, 使得基于 CNN 的细粒度图像识别方法[8−10]日益成熟. 由于目标的关键局部部分具有细微的差异, 加之 CNN 在全局特征学习方面存在局限性, 研究者们开始努力提升模型捕捉细微差异的能力, 进而更准确地定位关键部位并丰富特征表示. 当前的研究主要聚焦于两个方向: (1) 通过计算高阶信息来构建更具辨识性的特征表征, 以适应复杂的细粒度图像识别任务; (2) 通过采用注意力机制进行弱监督学习, 专注于具有区分度的局部区域, 并从这些区域中提取局部细粒度特征进行识别. 尽管这些方法能够通过类别标签学习到具有区分性的细粒度特征映射, 但它们普遍无法充分关注图像中某一局部区域内的细微差异. 此外, 在这些方法中,

定位模块的设计和训练过程都较为复杂. 同时, 由于细粒度图像受到姿态、光照、背景遮挡等多种因素的影响, 基于 CNN 的方法在提取深层特征时容易受到非特征区域噪声的干扰, 导致识别效果不佳.

近年来, 具有强大全局关系建模能力的视觉 Transformer (vision Transformer, ViT)[11]成功地将 Transformer[12]结构应用于视觉识别与检索领域[13], 展现出其在捕捉全局和局部特征方面的优势, 并在传统识别任务上的性能超越了主流的卷积神经网络. ViT 通过将图像切分成固定大小的块, 依次将这些图像块输入至 Transformer 编码器, 并通过级联的自注意力模块提取图像块序列间的交互信息, 从而捕获全局上下文特征, 在大规模图像分类任务中展示了巨大潜力. 然而, 当 ViT 被应用于细粒度图像识别任务时, 研究者们[14−16]发现其存在一些局限性: (1) 冗余信息处理不足: 自注意力机制平等地处理所有图像块, 无法有效去除背景噪声和冗余信息, 从而影响分类标记特征嵌入的表示能力, 进而降低识别准确性; (2) 细节信息的丢失: 由于图像分块的固定大小和序列长度的固定性, 这限制了感受野的有效扩展, 重要的细粒度特征可能会被忽略或部分丢失; (3) 上下文关联不足: 分类标记嵌入与图像块特征嵌入之间的关联建模不足, 容易忽视局部特征间的依赖关系, 无法捕捉细粒度特征的关键差异. 此外, 尽管 ViT 克服了 CNN 无法捕获长距离依赖方面的不足, 其归纳偏置能力相对较弱. 在细粒度图像识别任务中, 局部区分性特征的表征对识别效果至关重要, 因此, 如何在 ViT 模型中有效地融合全局上下文信息和局部辨识性特征, 成为提升细粒度图像识别准确率的关键挑战.

针对这些挑战及现有基于 ViT 方法的局限性, 本文提出了一种高效利用 ViT 模型的细粒度图像识别方法, 并探索了融合全局视图和局部视图的双视图识别范式. 为了提高对背景冗余信息的处理能力, 本文设计了一个基于注意力融合的冗余信息过滤模块, 该模块在 ViT 编码器内融合层级间的注意力权重, 筛选出含有丰富语义信息的图像块特征嵌入, 从而显著提升全局分类标记嵌入的辨识性. 同时, 为了增强对关键局部区域的识别能力, 本文引入了一个基于注意力阈值的关键区域定位模块, 该模块能够根据图像内容复杂度自适应地定位并放大重要的前景区域, 提供更具辨识性的局部视图. 这一过程不仅有效减少了由背景带来的噪声干扰, 而且强化了局部关键区域的特征表达. 进一步地, 针对基于局部视图的细粒度识别, 本文提出了一个局部区域特征自适应增强模块, 该模块通过给图像块特征嵌入赋予相关性权重, 增强了对局部辨识性特征的关注. 如图 2 所示, 上述 3 个模块被整合到一个双视图融合识别框架中, 以实现不同层次特征信息与预测结果的融合. 为进一步优化所提出的双视图融合识别框架, 本文还设计了基于双视图相似度的对比损失函数和基于双视图置信度的自适应推理策略, 前者使 ViT 模型输出的全局与局部特征更具辨识性, 而后者则进一步节约计算资源和缩短推理时间.

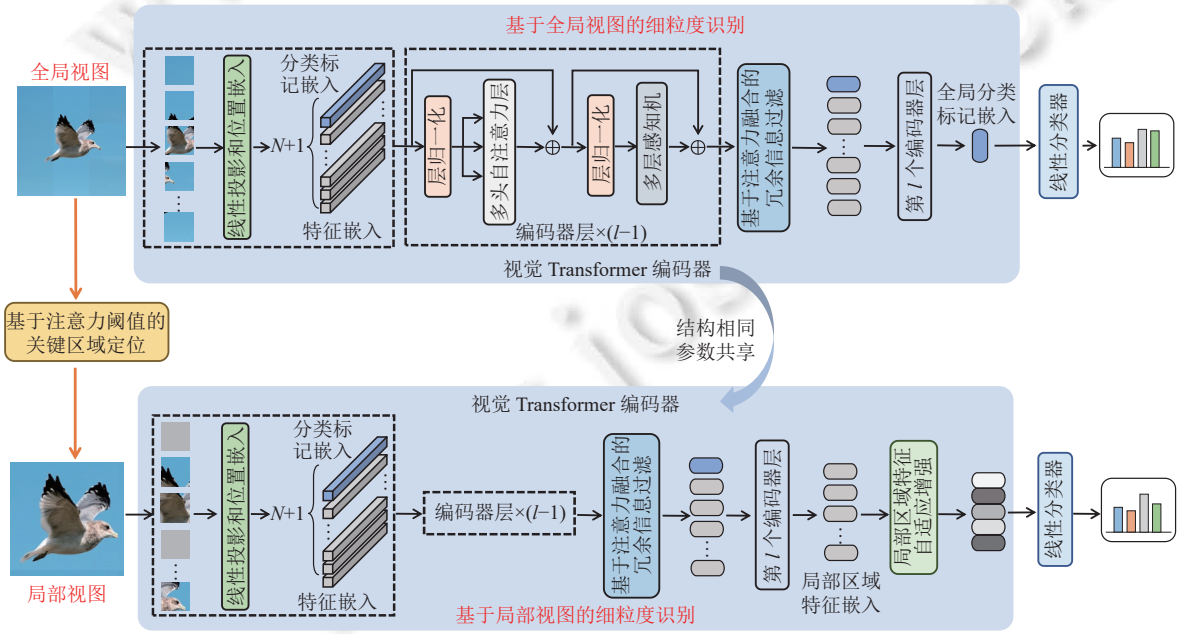

图 2 基于全局视图和局部视图的细粒度图像双视图融合识别框架

本文提出的方法在 CUB-200-2011[17]、Stanford Dogs[18]、NABirds[19]和 iNaturalist2017[20]这 4 个常用的细粒度图像识别数据集上进行了全面的对比实验. 实验结果表明, 相比于传统的单视图 ViT 模型, 本文提出的基于 ViT 的双视图识别范式在识别性能上实现了显著的提升. 此外, 与当前最先进的基于 ViT 模型的细粒度图像识别方法相比, 本文提出的双视图融合细粒度识别框架展现出更加卓越的性能. 这一性能提升归功于本文对 ViT 编码器的高效利用以及精心设计的推理策略. 这些结果不仅证明了本文方法的有效性, 也展示了其在细粒度图像识别领域的应用潜力.

1 相关工作

随着深度学习技术的不断发展, 目前的细粒度图像识别方法主要基于卷积神经网络 (CNN) 和视觉 Transformer (ViT) 这两种模型框架. 其中, 基于 CNN 的方法已相对成熟, 这类方法进一步分为基于强监督学习和基于弱监督学习的两大研究方向. 相比之下, 基于 ViT 的细粒度图像识别方法尚处于起步阶段.

1.1 基于卷积神经网络的细粒度图像识别

1.1.1 基于强监督学习的算法

基于强监督学习的细粒度图像识别算法[21−24]主要依赖训练数据集中的额外人工标注信息, 这些信息帮助检测目标物体的边框, 并定位关键局部区域. 进一步地, 通过主干网络提取特征, 并进行特征调整和融合等操作, 最终通过训练分类器实现了较好的识别效果. 具有代表性的研究包括 Part-based R-CNN[23]、Pose-normalized CNN[24]等. 例如, Zhang 等人[23]提出的 Part-RCNN 算法, 通过目标检测模型对细粒度图像的局部区域进行检测, 并对这些候选区域进行特征提取, 以便 SVM 分类器的训练. 虽然这种依赖于对象或部位级别标签的强监督方法取得了一定成果, 但其对人工标注信息的高度依赖性使得方法既耗时又成本高昂, 这在一定程度上限制了其实际应用性.

1.1.2 基于弱监督学习的算法

基于弱监督学习的细粒度图像识别算法不依赖于详细的部位标注, 仅通过图像的标签信息就能捕获全局和局部特征. 特别值得一提的是, 注意力机制的引入极大地提高了弱监督图像识别的性能, 使其成为近年来的研究热点. 当前主流的弱监督模型主要集中于基于区域定位[25−27]和基于特征编码[9,28,29]的方法. 基于区域定位的方法通常利用注意力机制和特征聚类等技术定位关键局部区域, 从而提取更有辨识力的视觉特征. 例如, Fu 等人[25]设计的循环注意卷积神经网络 (RA-CNN) 使用注意力区域网络来定位图像中的目标物体区域, 进行裁剪放大后送入下一层网络以获取物体部位级别的图像, 并在最后一层网络中融合全局和局部特征进行预测分类. 另一方面, 基于特征编码的方法则通过计算高阶信息来捕获丰富的关键特征表示, 例如, Yu 等人[28]利用层级双线性结构将图像特征和注意力特征编码为高阶特征向量, 以显著提升网络的特征表达能力.

总体来看, 卷积神经网络 (CNN) 作为主干网络在捕捉图像局部区域内细微差异方面存在局限性. 因此, 基于 CNN 的细粒度图像识别方法通常通过多次利用骨干网络来提取区域特征, 或通过定位最具辨识力的局部区域来提升捕获细节差异的能力. 然而, 这些方法往往使训练过程复杂化, 并增加了识别过程的复杂度. 鉴于目前基于 CNN 的细粒度图像识别方法正面临性能瓶颈, 本文将视觉 Transformer 引入此领域, 并进一步改进框架结构以提高其在细粒度图像识别任务中的效率和准确度.

1.2 基于视觉 Transformer 的细粒度图像识别

视觉 Transformer (ViT)[11]主要由嵌入层和编码器层构成, 其核心是通过多头自注意力机制来捕获全局上下文特征. 作为首个采用纯 Transformer 结构的视觉识别模型, ViT 的架构也被研究人员尝试应用于细粒度图像分类任务, 并探索如何解决 ViT 在数据高效训练和特征提取方面的潜在问题. He 等人[14]提出了 TransFG, 这是第 1 个基于 ViT 的细粒度图像识别框架. 其核心思想是利用多头自注意力机制筛选出具有辨识力的图像块, 然后将这些图像块特征嵌入与分类标记嵌入一起作为编码器最后一层的输入, 以进行后续预测分类. 此方法通过处理与目标物体区域高度相关的图像块间的内部关系, 有效提高了识别细微差异的能力. 然而, TransFG 无法生成多尺度的细粒度识别特征, 为了解决这一问题, Zhang 等人[30]提出了自适应注意力多尺度融合模型 AFTrans, 该模型利用注意力

权重自适应地筛选出重要的图像块, 从而提取更加鲁棒的多尺度特征. 尽管如此, 这些方法往往忽视了判别区域间的相互依赖性和目标的整体结构信息, 这些信息对于模型的决策至关重要. 为此, Sun 等人[15]提出了 SIM-Trans 模型, 通过挖掘关键区域的空间上下文关系, 加强了对外观和结构信息的表示学习. 与基于 CNN 的方法相比, 基于 ViT 的方法在细粒度图像识别任务中获得了更高的准确率.

当前基于 ViT 的研究重点主要集中于如何优化编码器以挖掘含有丰富语义信息的图像块, 并利用自注意力机制将这些图像块的信息有效整合到分类标记嵌入中, 以此显著提升图像的识别性能. 然而, 现有的方法主要依赖于分类标记嵌入来实现基于全局视图的图像识别, 而对细粒度图像中局部视图的重要性及多粒度特征表示的丰富性尚未给予充分的考虑. 因此, 本文的研究目标是探索在不改变 Transformer 编码器结构的前提下, 如何自适应地结合细粒度图像的全局和局部视图, 以实现基于多角度视图的高效识别推理. 本文提出的方法致力于深入挖掘细粒度图像内在的丰富信息, 旨在同时提高识别的准确率和效率.

2 视觉 Transformer 的架构介绍

2.1 基础结构

正如前文图 2 所示, 视觉 Transformer (ViT)[11]是由多个编码器层堆叠而成的. 每一层中, 包括了多头自注意力层 (multi-head self-attention layer, MHSA)、残差连接、层归一化 (layer norm, LN) 和多层感知机 (multi-layer perceptron, MLP). ViT 通过这些层的前向传播, 逐步细化特征, 从而提取图像的全局特征.

对于分辨率为 $H\times W$ 的图像 x, ViT 首先将其划分为 N 个尺寸为 $P\times P$ 的不重叠图像块序列 $x_p\in\mathbb{R}^{N\times(P\times P\times C)}$, 其中 $N=HW/P^2$, C 为通道数. 接着, 通过可学习的线性映射向量 $E\in\mathbb{R}^{(P^2\times C)\times D}$, 每个图像块被映射到 D 维空间. 此外, ViT 在图像块序列前附加一个维度相同的分类标记嵌入 $x^{\text{class}}\in\mathbb{R}^{1\times D}$, 以集成全局特征表示. 随后, 通过位置编码 $E_{\text{pos}}\in\mathbb{R}^{(N+1)\times D}$ 为序列中的每个特征向量赋予位置信息, 形成 ViT 编码器的首层输入 $Z_0\in\mathbb{R}^{(N+1)\times D}$, 如公式 (1) 所示:

$$Z_0=\left[x_0^{\text{class}};x_0^1E;x_0^2E;\ldots;x_0^NE\right]+E_{\text{pos}} \tag{1}$$

图像输入序列 Z_0 在 ViT 的编码器层中经过层归一化 (LN) 进行标准化后, 再经由每一层内部的多头自注意力 (MHSA) 层和多层感知机 (MLP) 模块进行特征的提取和细化. ViT 中第 l 层的输出结果如公式 (2) 所示:

$$\begin{cases} Z_l'=MHSA(LN(Z_{l-1}))+Z_{l-1},\ l=1,\ldots,L \\ Z_l=MLP(LN(Z_l'))+Z_l',\ l=1,\ldots,L \end{cases} \tag{2}$$

其中, Z_l' 和 Z_l 分别为经过第 l 层 MHSA 和 MLP 模块处理后的图像序列输出, $LN(\cdot)$ 表示层归一化操作. 通过多层编码器的堆叠和前向传播, ViT 逐步提取出更为精细且有效的特征, 并将其集成到分类标记嵌入中. 最终, 编码器末层输出的分类标记嵌入被用作图像的全局特征, 输入至分类器中完成识别预测.

多头自注意力机制是上述过程的关键, 它允许模型在处理每个图像块时, 同时考虑图像中其他部分的信息. 具体来说, MHSA 通过多个独立的“注意力头”并行处理输入数据, 每个头学习到不同的表示子空间. 在每个头中, 自注意力机制通过计算输入序列中所有图像块之间的相关性来实现. 这一过程涉及以下步骤: 首先, 每个图像块被转换为查询 (query, Q)、键 (key, K) 和值 (value, V) 的表示形式, 即 $Q=XW^Q$, $K=XW^K$ 和 $V=XW^V$. 其中, X 代表输入的特征表示, W^Q、W^K 和 W^V 是相应的线性变换矩阵. 随后, 自注意力机制按照公式 (3) 计算:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{3}$$

其中, d_k 表示键向量的维度, 分母中的 $\sqrt{d_k}$ 用于实现点积的缩放, 防止梯度消失. $Softmax$ 函数应用于每一行, 以计算不同图像块间的注意力权重. 上述计算在多个头上并行执行, 每个头都有独自的 W^Q、W^K 和 W^V 权重矩阵. 最终, 所有头的输出经拼接后, 通过另一线性变换得到最终输出:

$$MHSA(Q,K,V)=Concat(H_1,H_2,\ldots,H_i)W^O \tag{4}$$

其中, 每个头 H_i 是基于公式 (3) 中自注意力机制的输出, W^O 为输出变换的线性矩阵, $Concat(\cdot)$ 表示特征拼接操作. 通过这种方式, ViT 模型能够有效地捕捉图像中不同区域间的复杂关系, 为图像识别提供丰富和精确的特征表示.

2.2 局限性分析

通过第 2.1 节对视觉 Transformer 核心架构的详细分析, 可以发现几个结构性的缺陷: (1) 缺乏空间归纳偏置: 模型依赖于图像块的位置编码与注意力权重来建模空间关系, 缺少类似于卷积操作的局部归纳偏置; (2) 感受野固定: 固定的图像块尺寸限制了关键特征的捕捉, 尤其在区分细微差异时显得不足; (3) 注意力均等化: 尤其在区分细微差异时显得不足; (4) 全局优先: 尽管分类标记嵌入在提取全局上下文信息方面表现出色, 它却往往无法精准捕捉关键的局部特征差异. 当视觉 Transformer 在处理具有复杂背景的细粒度图像识别任务时, 这些内在的结构性缺陷突显了模型在处理需要背景噪声时的局限性, 具体表现为: (1) 分类标记嵌入稀释: 由于视觉 Transformer 均等地处理每个图像块, 大量背景块的特征可能被错误地引入分类标记嵌入中, 影响全局特征的准确表达; (2) 特征块处理的均等化: 视觉 Transformer 无法区分含有目标特征的关键图像块与背景噪声块, 这降低了模型对目标特征的关注, 增加了背景噪声对目标对象的干扰; (3) 类别间视觉相似性误导: 在背景特征与目标物体视觉属性相似的情况下, 视觉 Transformer 可能会将背景的某些部分错误地解释为目标类别的特征, 特别是在不同类别的图像具有相似背景的情况下. 为了克服这些限制, 本研究提出了一系列改进策略, 旨在增强模型对背景噪声的鲁棒性, 优化对关键特征的捕捉能力.

3 本文方法

本文提出了一个双视图融合细粒度图像识别框架, 如前文图 2 所示. 该框架不仅整合了标准的 ViT 编码器, 还集成了 3 个专为细粒度识别设计的模块: 一个即插即用的基于注意力融合的冗余信息过滤模块、一个无参数的基于注意力阈值的关键区域定位模块, 以及一个局部区域特征自适应增强模块. 在识别过程中, 框架首先使用 ViT 编码器对输入样本进行基于全局视图的初步分析, 然后通过关键区域定位模块识别并提取出前景区域, 再将这些区域重新送回编码器进行基于局部视图的细粒度识别. 最终, 框架综合全局视图和局部视图的预测结果, 形成最终的识别判断. 本框架的独特之处在于它从全局和局部两个互补的视角高效地利用结构相同且共享参数的 ViT 编码器, 可根据输入图像的识别难度动态选择最高效的推理路径.

3.1 基于注意力融合的冗余信息过滤

3.1.1 冗余信息过滤的动机

在细粒度图像识别中, 冗余信息指的是那些对分类决策贡献较小或无关的图像特征, 如背景元素和非关键视觉噪声. 传统 ViT 编码器通过多头自注意力机制捕获丰富的全局上下文信息, 细化这些全局特征并聚集到最后一层的分类标记嵌入中, 但这可能忽略图像中关键的局部区域信息. 此外, 自注意力机制的全局性质有时会错误地将注意力分配给与目标类别无关的图像区域, 如背景树木或建筑. 因此, 如何确保模型聚焦于含有类别决定性特征的图像区域, 是提高 ViT 模型在细粒度图像识别任务中的关键挑战之一. 近期, 基于掩码图像建模 (masked image modeling, MIM) 的无监督学习方法[31,32]的成功实践表明, 在 ViT 中大约只需 25% 的图像块就能有效表征整张图片的语义信息. 这一发现启发本文在 ViT 编码器的深层处进行冗余信息过滤, 筛选出较为重要的图像块特征嵌入, 以增强分类标记嵌入的辨识性.

3.1.2 层级注意力的融合

He 等人[14]指出, 在 ViT 架构中, 编码器的不同层级学习到的局部特征存在差异. 基于这一发现, 本文为具有 l 层编码器的 ViT 模型设计了一个即插即用的冗余信息过滤模块. 该模块巧妙地融合了不同层级间的注意力权重, 充分挖掘编码器浅层的细节信息和深层的语义信息. 它自适应地从第 $l-1$ 层编码器的输出中筛选出具有辨识力的图像块特征嵌入, 并将其作为第 l 层的输入. 这种方法不仅精细化了最终的分类标记嵌入, 而且有效地进行了冗余信息的过滤, 从而提升了整体模型的识别能力和效率. 如图 3 所示, 对于输入图像 x, 设 ViT 编码器第 $l-1$ 层的输出为 $Z_{l-1}=[x_{l-1}^{\text{class}};Z_{l-1}^{1};Z_{l-1}^{2};\ldots;Z_{l-1}^{N}]$, 相应的多头自注意力层会根据公式 (3) 中的点积注意力机制计算出的注意力分数矩阵 $A_{l-1}\in\mathbb{R}^{S\times(N+1)\times(N+1)}$. 在此, S 与 N 分别表示注意力头的数目和图像块特征嵌入的数目. 为了保留局部信息的完整性并减少计算复杂度, 通过累加操作融合 S 个注意力头的权重, 可得到该层的综合注意力分数矩阵 $\hat{A}_{l-1}\in$

$\mathbb{R}^{(N+1)\times(N+1)}$. 显而易见, $\hat{A}_{l-1}$ 中的第 1 行展示了在第 $l-1$ 层编码器中, 分类标记嵌入 x_{l-1}^{class} 与自身以及 N 个图像块特征嵌入间的语义关联程度. 如果图像块特征嵌入与分类标记嵌入的语义相似度较高, 意味着该图像块携带的视觉信息对于模型输出正确预测类别更为重要. 为了避免仅依赖单一层次的注意力权重而忽略其他层级具有辨识性的局部信息, 通过累加前 $l-1$ 层的注意力分数矩阵可得到图像块特征嵌入的综合注意力分数序列 $A_{\text{final}}=\sum_{i=1}^{l-1}\hat{A}_i[0,1:]$.

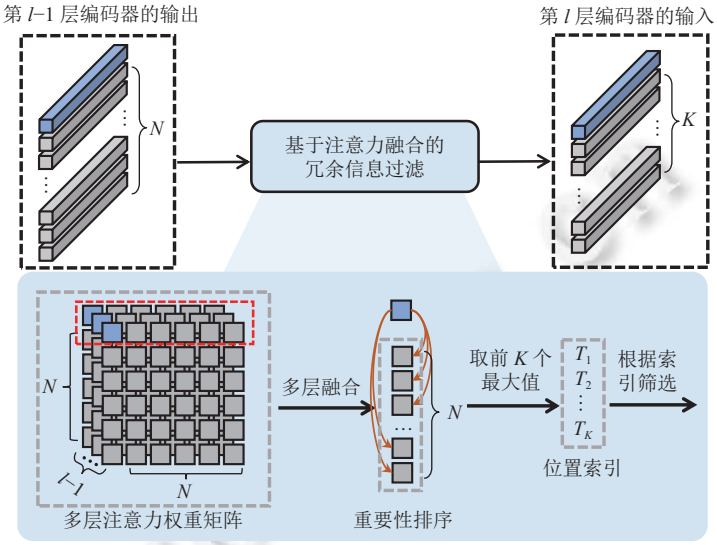

图 3 基于注意力融合的冗余信息过滤模块流程展示图

3.1.3 冗余信息的过滤

为了更有效地过滤掉冗余噪声信息, 我们对 A_{final} 中的注意力分数进行降序排序, 选择前 K 个最大值对应的位置索引 $[T_1,T_2,\ldots,T_K]$. 接着, 根据这些索引从 Z_{l-1} 中挑选出与分类标记嵌入关联度较高的图像块特征嵌入与第 $l-1$ 层的分类标记嵌入 x_{l-1}^{class}, 组成新的序列 Z_{l-1}^{select}, 输入至编码器的最后一层. 该过程可用公式 (5) 表示:

$$Z_{l-1}^{\text{select}}=[x_{l-1}^{\text{class}};Z_{l-1}^{T_1};Z_{l-1}^{T_2};\ldots;Z_{l-1}^{T_K}] \tag{5}$$

与当前大多数基于 ViT 的细粒度识别方法类似, 只需将编码器最后一层输出的分类标记嵌入 x_l^{class} 作为全局特征输入至由一个简单的全连接层构成的线性分类器 FC_1 中, 即可完成基于全局视图的细粒度识别. 预测的类别概率 P_1 由公式 (6) 计算得出:

$$P_1=FC_1\left(x_l^{\text{class}}\right) \tag{6}$$

其中, $FC_1(\cdot)$ 表示应用在 x_l^{class} 上的全连接层, 其输出为各类别的预测概率. 本文所提出的基于注意力融合的冗余信息过滤是一种无参操作, 仅通过对编码器最后一层的输入进行微小改动, 就可以有效去除从背景区域提取的无效特征. 这种方法不仅融合了不同层级间的局部辨识性信息, 还避免了最终输出的全局分类标记嵌入存在的冗余问题, 提升对大多数简单样本的辨识能力.

3.2 基于注意力阈值的关键区域定位

3.2.1 关键区域定位的动机

在大多数自然图像中, 复杂的背景往往包含大量噪声信息. 尽管 ViT 模型通过将图像分割为序列化的图像块来提供了新颖的视觉建模方法, 但很多图像块中可能仅包含背景信息或部分前景对象, 从而可能干扰 ViT 捕获关键区域的注意力信息, 影响对细粒度类别的识别. 然而, 人类在识别类间差异微小的细粒度类别时, 通常会优先关注关键对象或区域, 并忽略不重要的背景信息. 受到这种高效视觉处理策略[33,34]的启发, 本文提出了一种基于注意力阈值的关键区域定位模块. 该模块通过对输入图像的注意力激活值进行分析, 自适应地定位和放大关键区域, 从而构成新的局部视图, 使得模型能够更加近距离地观察细粒度目标. 特别是, 当放大关键区域时, 局部视图的细微

特征, 如纹理、形状和颜色变化, 将变得更加明显, 这对于区分仅具有微妙差异的细粒度子类别至关重要.

3.2.2 注意力阈值的生成

考虑到自然图像中目标的视角和大小经常发生变化, 直接采用固定的注意力响应阈值在目标区域定位中往往效果不理想. 基于此, 本文提出一种基于注意力累积分布的自适应阈值生成策略, 用于动态调整图像前景区域的定位, 同时抑制对背景的响应. 在第 3.1.2 节中构建层级注意力机制时, 通过叠加编码器不同层的注意力分数矩阵 $\hat{A}_i$, 即可生成综合注意力分数序列 $A_{\text{final}} \in \mathbb{R}^{1\times N}$. 该序列有效表征了局部区域的特征响应. 如图 4 所示, 为精准筛选关键区域, 本文引入了累积分布阈值方法以确定注意力阈值 τ. 具体地, 首先对 A_{final} 中的所有元素按注意力的大小进行排序, 然后计算每个位置 i 上的累计注意力值 $G(i)$:

$$G(i)=\sum_{j=0}^{i} A_{\text{final}}[j] \tag{7}$$

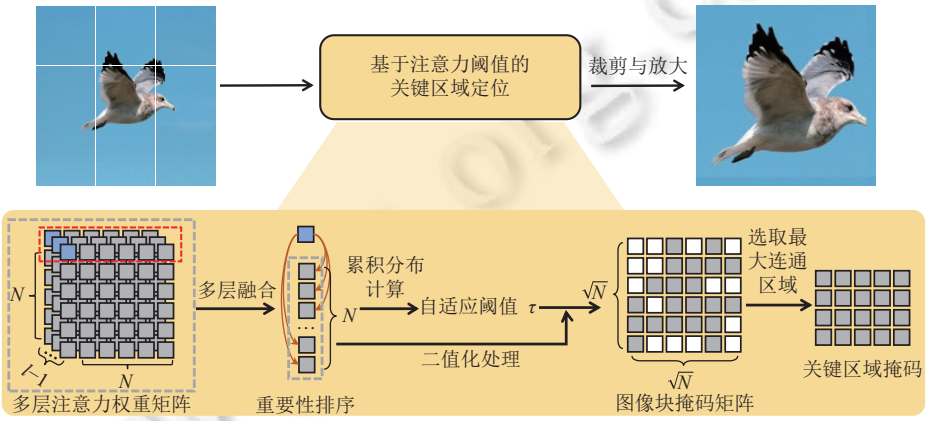

图 4 基于注意力阈值的关键区域定位模块流程展示图

累积注意力值 $G(i)$ 表示从排序后的列表中位置 0–i 的总和. 为确定阈值 τ, 此处设定了一个预定义的阈值参数 u, 当 $G(i)$ 的值首次大于或等于 u 时, 对应的 $A_{\text{final}}[i]$ 即为整张图片所需的注意力阈值 τ, 即 $\tau = A_{\text{final}}[i]$. 这种自适应阈值生成策略使得累积分布阈值 u 的取值可以控制通过层级注意力机制从图像块序列中筛选出的关键区域比例. 在训练和推理阶段, 根据图像内容和复杂度, 以及累积分布阈值 u, 本文方法可灵活确定 A_{final} 中的自适应阈值 τ, 有效地提升了模型对不同尺度目标的识别能力.

3.2.3 关键区域的定位

自适应阈值生成策略允许模型更精准高效地定位并关注图像中的关键区域. 在处理不同复杂度的细粒度图像时, 根据所生成的自适应阈值动态地对图像进行注意力裁剪和放大, 从而形成细粒度图像的局部视图. 具体过程如图 4 所示, 首先根据图像块序列的位置信息, 将综合注意力分数序列 A_{final} 重塑为二维矩阵形状. 然后, 通过对矩阵中各元素 $A_{\text{final}}(i,j)$ 进行二值化处理, 生成图像块掩码矩阵 $I \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$, 具体操作定义如下:

$$I(i,j)=\begin{cases}1, & \text{if } A_{\text{final}}(i,j)>\tau\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

其中, 大于自适应阈值 τ 的元素值设为 1, 代表判别性关键区域; 小于或等于阈值的元素值设为 0, 表示背景噪声. 在生成图像块掩码矩阵 I 后, 具体定位最小边界框的步骤如下: (1) 首先将掩码矩阵 I 中所有值为 1 的元素对应的位置索引提取出来, 得到坐标索引集合 S; (2) 根据集合 S 中各元素的坐标信息, 分别求解横坐标和纵坐标的最小值和最大值, 分别记作 $x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$; (3) 最终将上述坐标作为边界框的 4 个顶点, 从而确定出覆盖掩码矩阵中大多数重要区域的最小边界框, 即矩形区域 $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$. 通过这种方法确定的边界框能够高效地覆盖图像中注意力最高、最具辨识性的区域, 同时适当忽略某些分散的低注意力区域, 从而有效地抑制背景噪声的干扰. 这种设计确保了边界框的紧凑性, 集中捕捉了高权重的区域特征, 有助于提升模型的识别性能和计算效率. 最后,

利用边界框坐标, 通过适当的比例放大, 以与原始输入图像的尺寸保持一致, 并裁剪出相应的图像区域, 最终形成细粒度图像的局部视图 X, 作为新的输入重新送入 ViT 编码器中进行分析.

3.3 局部区域特征自适应增强

在 ViT 的编码器中, 虽然通过冗余信息过滤模块可以增强分类标记嵌入表达关键特征的能力, 但基于分类标记嵌入的细粒度识别主要通过自注意力机制捕获全局上下文信息, 在计算图像块特征嵌入注意力的过程中容易忽略图像局部区域信息. 当通过关键区域定位生成图像的局部视图后, 辨识性区域在特征表示中的比重会增加. 因此, 本文设计了一种局部区域特征自适应增强模块, 不再依赖基于分类标记嵌入的全局视图识别范式, 而是通过显式地建模出局部区域之间的相互依赖性, 以提升对局部区域特征嵌入的有效利用, 进一步增强模型对局部视图的关注度.

值得注意的是, 局部视图 X 被重新输入至 ViT 编码器前, 首先对局部视图的图像块序列进行了基于随机擦除的数据增强. 具体而言, 本文随机生成与图像块数量一致的、值在 0–1 之间的随机噪声向量, 随后对这些噪声进行排序, 并选择噪声值最小的 1/10 的图像块进行擦除, 即将这部分图像块的像素值设置为 0. 通过随机擦除图像块序列中约 1/10 的图像块, 迫使 ViT 模型关注更多局部细节信息, 而不是过度拟合于某些显著特征, 从而提高模型的泛化能力. 经过基于注意力融合的冗余信息过滤后, 最后一层编码器的输出可以表示为 $Z_l=[x_l^{\text{class}};Z_l^1;Z_l^2;\ldots;Z_l^{K_{\text{local}}}]$, 其中 $Z_l^{K_{\text{local}}}\in\mathbb{R}^{1\times D}$, K_{local} 代表过滤后的局部区域特征嵌入个数, D 为向量维度. 移除分类标记嵌入 x_l^{class} 后, 便可得到局部区域特征嵌入序列 $Z_l^{\text{local}}=[Z_l^1;Z_l^2;\ldots;Z_l^{K_{\text{local}}}]$. 如图 5 所示, 为了让模型自行学习不同局部区域的权重而无需过多人工干预, 本文首先将所有局部区域特征嵌入串联拼接成一组输入向量 $Z_l^{\text{cat}}\in\mathbb{R}^{1\times(K\times D)}$, 然后通过一个注意力模块捕获局部区域特征嵌入的相关性, 即局部上下文信息. 该注意力模块由两个全连接层和两个激活函数层组成, 输出与局部区域特征嵌入数量相等的一组特征权值 s_l, 用于刻画 Z_l^{local} 中 K_{local} 个局部区域特征嵌入的权重. 具体过程可以表示为:

$$s_l=\sigma(W_4\delta(W_3Z_l^{\text{cat}})) \tag{9}$$

其中, $W_3\in\mathbb{R}^{K_{\text{local}}\times(K_{\text{local}}\times D)}$ 与 $W_4\in\mathbb{R}^{K_{\text{local}}\times K_{\text{local}}}$ 是两个全连接层的线性变换矩阵, $\delta(\cdot)$ 代表 ReLU[35]非线性激活函数层, $\sigma(\cdot)$ 代表 Sigmoid 激活函数层. 最后, 将得到的权值 s_l 与对应输入向量加权相乘, 进而得到一系列增强后的局部区域特征嵌入 $Z_l^{\text{aug}}=\left[Z_l^1s_l^1;Z_l^2s_l^2;\ldots;Z_l^{K_{\text{local}}}s_l^{K_{\text{local}}}\right]$. 这种处理方式既抑制了不相关信息的表达, 又降低了注意力计算量.

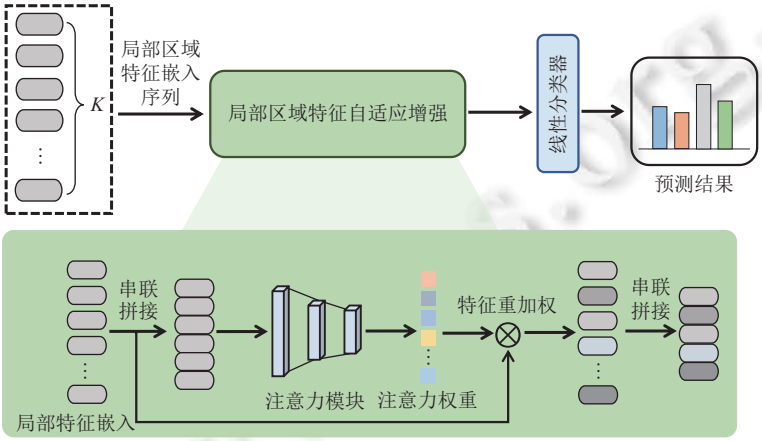

图 5 局部区域特征自适应增强模块结构展示图

在基于局部视图的细粒度识别过程中, 为了进行最终的分类预测, 首先需要将 Z_l^{aug} 中的所有局部区域特征嵌入进行串联拼接, 形成一个综合的特征向量 x_l^{aug}, 该过程可以通过公式 (10) 表示:

$$x_l^{\text{aug}}=Concat(Z_l^{\text{aug}}) \tag{10}$$

其中, $Concat(\cdot)$ 表示特征串联拼接操作. 得到串联拼接后的特征向量 x_l^{aug} 后, 再将其输入至一个由全连接层构成的线性分类器 FC_2 中便可完成分类预测. 预测的类别概率 P_2 由公式 (11) 计算得出:

$$P_2 = FC_2(x_l^{\text{aug}}) \tag{11}$$

其中, $FC_2(\cdot)$ 表示应用在 x_l^{aug} 上的全连接层, 其输出为各类别的预测概率. 本文所提出的局部区域特征自适应增强模块通过为局部视图中的图像块特征嵌入赋予相关性权重系数, 加强了对局部关键区域的关注, 使模型具备更强的辨识能力, 特别是对于那些较难辨识的细粒度样本.

3.4 模型框架的训练与推理

如图 2 所示, 本文设计的双视图融合识别框架将前文介绍的 3 个专为 ViT 定制的模块进行综合应用, 实现了基于全局视图和局部视图的细粒度识别. 需要特别指出的是, 该模型在训练和测试阶段的流程有所差异. 在训练阶段, 对所有训练样本同时进行基于全局视图和基于局部视图的两阶段细粒度识别, 以此来驱动模型提取更具区分性的全局和局部特征. 在测试阶段, 本文设计了一种基于双视图置信度的自适应推理策略, 根据样本识别难易程度的判断, 自适应地选择单视图或双视图的识别方式, 以实现对大规模细粒度图像的高效识别.

3.4.1 基于双视图相似度的对比损失函数

针对细粒度图像类间差异小、类内差异大的特点, 以及 Transformer 模型归纳能力的局限, 本文对损失函数进行了多角度优化. 由于模型训练阶段通过基于自适应阈值的关键区域定位模块引入了图像的局部视图, 因此本文对于图像的双视图均采用分类任务常用的交叉熵 (cross-entropy) 损失来计算分类损失. 具体计算公式如下:

$$L_{\text{cls}}^1 = L_{\text{ce}}\left(Softmax\left(W_1 x_l^{\text{class}} + b_1\right), y\right),\ L_{\text{cls}}^2 = L_{\text{ce}}(Softmax(W_2 x_l^{\text{aug}} + b_2), y) \tag{12}$$

其中, W_1、W_2 和 b_1、b_2 分别表示两个分类器 FC_1 和 FC_2 的权重矩阵和偏置向量, $Softmax(\cdot)$ 函数将全连接层的输出转换为类别概率分布, y 代表输入样本的真实类别标签.

细粒度图像有类间混淆性高且类内差异性较大的特性, 交叉熵损失函数虽能优化 ViT 模型捕获显著的类间差异, 但在捕获类间细微差异和类内差异方面能力有限. 鉴于此, 本文还设计了一种基于双视图相似度的对比损失函数, 用于驱动模型在增大不同子类特征差异的同时减小相同子类特征差异. 对于一个大小为 B 的训练批次, 首先通过基于自适应阈值的关键区域定位模块生成图片的局部视图, 即可将该批次中训练数据的总量扩充至 $2B$. 然后, 在获取所有全局视图的分类标记嵌入 x_l^{class} 与局部视图的局部区域特征嵌入 x_l^{aug} 后计算所有样本对的平均对比损失, 具体过程如下:

$$L_{\text{con}} = \frac{1}{(2B)^2}\sum_{i=1}^{2B}\left[\sum_{\substack{i=1\\ y_i=y_j}}^{2B}[1-\cos(z_i,z_j)] + \sum_{\substack{j=1\\ y_i\neq y_j}}^{2B}\max\{[\cos(z_i,z_j)-\alpha],0\}\right] \tag{13}$$

其中, z_i 为第 i 个图像经过 ViT 编码器映射生成的全局分类标记嵌入或局部区域特征嵌入, $\cos(z_i,z_j)$ 表示 z_i 和 z_j 的余弦相似度, 当其大于超参数 α 时才会在对比损失中起作用. 为了防止重复计算, 对于每个输入样本的全局分类标记嵌入及其局部区域特征嵌入, 本文只计算一次与其他样本特征嵌入的相似度.

综合来看, 通过分别计算全局视图与局部视图的交叉熵损失和对比损失, 并将其融合, 帮助模型区分不同子类差异, 同时归并相同子类差异. 因为基于全局视图和局部视图分别计算出的交叉熵损失是利用同一批次的输入样本得到的, 所以在融合过程中, 将二者之和视为总损失的一部分, 基于双视图相似度的对比损失视为另一部分. 最终, 模型的总体损失函数为:

$$L_{\text{total}} = 0.5\times L_{\text{cls}}^1 + 0.5\times L_{\text{cls}}^2 + L_{\text{con}} \tag{14}$$

3.4.2 基于双视图置信度的自适应推理策略

人类视觉系统在处理视觉信息时, 通常先快速捕获全局信息, 如整体形状和场景概貌, 帮助迅速构建对环境的基本理解. 在面对视觉信息不清晰或存在不确定性的情况下, 人类往往会采取额外的步骤以澄清或确认, 如改变视角、聚焦细节或结合其他感官信息, 进一步关注局部细节. 这种视觉感知机制可以根据需求在全局视图和局部视

图处理之间灵活切换, 有效地分配注意力以优化效率和准确性. 受此启发, 本文提出了一种基于双视图置信度的自适应推理策略, 如图 6 所示, 该策略旨在使所提出的双视图融合识别框架更适应细粒度图像分类任务, 在保持高精度的同时提升推理效率. 本文首先对输入样本执行基于全局视图的细粒度识别, 获取初步预测结果及其置信度. 若置信度高于设定阈值, 模型将当前预测视为易识别样本, 以此作为最终结果, 停止进一步推理. 若置信度低于阈值, 模型则判定样本为难识别, 随即激活基于注意力阈值的关键区域定位模块, 对输入样本进行基于注意力激活的裁剪, 并将放大后的视图重新输入编码器进行基于局部视图的细粒度识别, 旨在对难以识别图像的关键部分或特定特征进行更细致的分析和重新分类. 最终, 两次预测结果结合形成最终的预测判断.

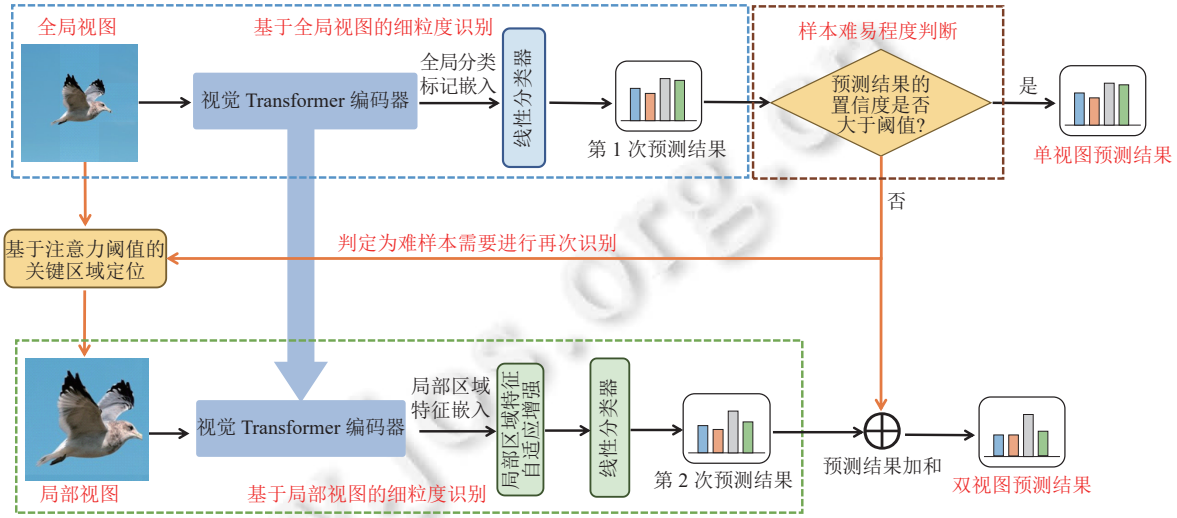

图 6 基于双视图置信度的自适应推理策略示意图

4 实验与评估

本节将在 4 个广泛使用的细粒度图像识别数据集上进行对比实验, 来验证本文所提方法的优越性. 此外, 本节还进行了一系列的消融实验, 不仅验证了每个模块对于整体性能的贡献, 而且还涵盖了模型的参数敏感性分析与可视化分析.

4.1 数据集

本文实验在 4 个极具挑战性的细粒度图像识别数据集上进行: CUB-200-2011[17], Stanford Dogs[18], NABirds[19] 和 iNaturalist2017[20].

CUB-200-2011 数据集由加州理工学院和加州大学圣迭戈分校共同发布, 专门用于研究鸟类图像的细粒度分类问题. 这个数据集涵盖了 200 种鸟类, 每种约有 60 张图片, 总共 11 788 张. 它被官方划分为 5 994 张训练图片和 5 794 张测试图片.

Stanford Dogs 数据集出自斯坦福大学, 专注于狗的品种识别. 包含 120 个犬种, 总计约 20 580 张图片, 其中 12 000 张用于训练, 8 580 张用于测试. 该数据集的特点是包含了各种背景和姿态的图像, 这对训练和测试能在复杂环境下识别狗品种的算法非常有帮助.

NABirds 数据集由康奈尔大学发布, 是一个大型的北美鸟类识别数据集. 它包含了 555 个子类别的北美鸟类, 共 48 562 张图像, 分为 23 929 张训练图像和 24 633 张测试图像. 与 CUB-200-2011 相比, NABirds 提供了更多种类和更多数量的图片, 适合需要大规模数据集进行训练和测试的研究.

iNaturalist2017 数据集来源于真实的生物多样性观察平台 iNaturalist, 涵盖了包括植物、动物、真菌等在内的 13 个大类, 共计 5 089 个细粒度类别. 其中, 训练集和测试集分别包含 579 184 和 95 986 张图像, 整体规模远超

其他常用细粒度识别数据集, 具有背景复杂、拍摄角度多变、类别数量庞大以及类别间长尾分布明显等特点.

4.2 实验细节

本文选择了含有 12 层编码器的 ViT 基础模型 ViT-B_16[11]作为基准网络, 并采用加载来自 ImageNet 预训练的参数进行模型初始化, 以便确保性能对比的公平性. 遵循细粒度图像识别任务的标准数据增强策略, 所有输入图像首先被统一调整到 600×600 分辨率. 训练阶段, 采用随机水平翻转作为数据增广手段, 通过随机裁剪得到 448×448 分辨率的图像作为网络输入. 测试阶段, 则仅采用中心裁剪方式, 以最大程度保留样本信息. 实验中, 所有图像均被划分成 784 个 16×16 大小的图像块. 训练过程中, 双视图相似度对比损失函数的超参数 α 被设定为 0.3, 并使用随机梯度下降法 (SGD) 进行网络训练和优化, 动量 (momentum) 设置为 0.9, 批次大小设为 32. 鉴于 Stanford Dogs 数据集的图片内容复杂度与其他 3 个数据集有所不同, 在该数据集上的训练将学习率初始化为 0.003, 而在其他 3 个数据集上则将学习率初始化为 0.03, 并采用余弦退火策略控制学习率下降. 测试过程中, 自适应推理策略中的置信度阈值在 4 个数据集上均被设定为 0.8. 本文所有实验都是在配备有 NVIDIA GeForce RTX 3090 GPU (4 张)、Intel Xeon Gold 6330 CPU @ 2.00 GHz 和 256 GB RAM 的高性能服务器上进行. 所使用的操作系统为 Ubuntu 20.04 LTS, 实验中所有的算法均在 Python 3.7 环境下开发, 依赖深度学习框架 PyTorch 1.10 和 CUDA Toolkit 11.2.

4.3 对比方法定量分析

为了全面验证本文方法的优越性, 本文选取了近年来细粒度图像识别领域的主流方法进行了对比分析. 在 CUB-200-2011、Stanford Dogs、NABirds 和 iNaturalist2017 这 4 个广泛使用的公开数据集上的对比结果展现在表 1 中, 其中加粗字体的数据表示在相应指标上取得的最佳结果. 这些结果不仅验证了本文方法的有效性, 还展示了其在不同数据集上的泛化能力.

表 1 不同弱监督细粒度图像识别方法在 CUB-200-2011、Stanford Dogs、NABirds 和 iNaturalist2017 数据集上的识别准确率对比 (%)

数据集	对比方法	基础网络	Top-1 准确率	数据集	对比方法	基础网络	Top-1 准确率
CUB-200-2011	MSHQP[8]	ResNet-50	—	Stanford Dogs	MSHQP[8]	ResNet-50	90.4
	ViT[11]	ViT-B_16	90.2		ViT[11]	ViT-B_16	90.2
	TransFG[14]	ViT-B_16	91.1		TransFG[14]	ViT-B_16	90.5
	SIM-Trans[15]	ViT-B_16	91.3		SIM-Trans[15]	ViT-B_16	—
	IELT[16]	ViT-B_16	90.9		IELT[16]	ViT-B_16	90.6
	API-Net[29]	ResNet-101	—		API-Net[29]	ResNet-101	90.3
	ResNet[36]	ResNet-50	84.5		ResNet[36]	ResNet-50	82.7
	MA-CNN[37]	ResNet-50	86.5		MA-CNN[37]	ResNet-50	—
	NTS-Net[38]	ResNet-50	87.5		NTS-Net[38]	ResNet-50	87.5
	Cross-X[39]	ResNet-50	87.7		Cross-X[39]	ResNet-50	88.9
	DCL[40]	ResNet-50	87.8		DCL[40]	ResNet-50	—
	MRDMN[41]	ResNet-50	88.8		MRDMN[41]	ResNet-50	89.1
	FDL[42]	DenseNet-161	89.1		FDL[42]	DenseNet-161	—
	PRIS[43]	ResNet-101	90.0		PRIS[43]	ResNet-101	90.7
	CAL[44]	ResNet-101	90.6		CAL[44]	ResNet-101	88.7
	AA-Trans[45]	ViT-B_16	91.4		AA-Trans[45]	ViT-B_16	90.8
	ACC-ViT[46]	ViT-B_16	91.8		ACC-ViT[46]	ViT-B_16	91.3
	MP-FGVC[47]	ViT-B_16	91.8		MP-FGVC[47]	ViT-B_16	91.0
	MpT-Trans[48]	ViT-B_16	92.0		MpT-Trans[48]	ViT-B_16	91.4
	CGL[49]	ViT-B_16	92.0		CGL[49]	ViT-B_16	—
	本文方法	ViT-B_16	**92.1**		本文方法	ViT-B_16	**91.7**

表 1 不同弱监督细粒度图像识别方法在 CUB-200-2011、Stanford Dogs、NABirds 和 iNaturalist2017 数据集上的识别准确率对比 (%)(续)

数据集	对比方法	基础网络	Top-1 准确率	数据集	对比方法	基础网络	Top-1 准确率
NABirds	ViT[11]	ViT-B_16	89.9	iNaturalist2017	ViT[11]	ViT-B_16	68.7
	TransFG[14]	ViT-B_16	90.5		TransFG[14]	ViT-B_16	—
	SIM-Trans[15]	ViT-B_16	90.3		SIM-Trans[15]	ViT-B_16	69.9
	IELT[16]	ViT-B_16	90.1		IELT[16]	ViT-B_16	—
	API-Net[29]	DenseNet-161	88.1		API-Net[29]	ResNet-101	—
	Cross-X[39]	ResNet-50	86.4		Cross-X[39]	ResNet-50	—
	PRIS[43]	ResNet-101	88.4		PRIS[43]	ResNet-101	—
	AA-Trans[45]	ViT-B_16	90.2		AA-Trans[45]	ViT-B_16	—
	ACC-ViT[46]	ViT-B_16	91.3		ACC-ViT[46]	ViT-B_16	70.2
	MP-FGVC[47]	ViT-B_16	91.0		MP-FGVC[47]	ViT-B_16	—
	MpT-Trans[48]	ViT-B_16	91.3		MpT-Trans[48]	ViT-B_16	—
	PAIRS[50]	ResNet-50	87.9		PAIRS[50]	ResNet-50	—
	GHRD[51]	ResNet-50	88.0		GHRD[51]	ResNet-50	—
	DSTL[52]	Inception-v3	87.9		DSTL[52]	Inception-v3	—
	MGE-CNN[53]	SENet-154	88.6		MGE-CNN[53]	SENet-154	—
	SSN[54]	ResNet-101	—		SSN[54]	ResNet-101	65.2
	IARG[55]	ResNet-101	—		IARG[55]	ResNet-101	66.8
	TASN[56]	ResNet-50	—		TASN[56]	ResNet-50	68.2
	SGRN[57]	ResNet-50	—		SGRN[57]	ResNet-50	73.6
	HI2R[58]	ViT-B_16	—		HI2R[58]	ViT-B_16	73.9
	SwinFG[59]	Swin-B	—		SwinFG[59]	Swin-B	73.0
	本文方法	ViT-B_16	**91.9**		本文方法	ViT-B_16	**74.1**

在 CUB-200-2011 数据集上的实验结果表明, 仅使用基础 ViT 模型[11]即超越了多数基于传统 CNN 的方法, 达到了 90.2% 的识别准确率. 而当采用本文提出的双视图融合识别框架后, 准确率更是实现了显著的提升, 相较于原始 ViT 模型进一步提升了 1.9%. 这证明了本文所设计模块的有效性以及框架设计的合理性. 在 Stanford Dogs 数据集上, 本文的模型相比于原始 ViT 模型实现了 1.5% 的准确率提升, 与目前最佳的 MpT-Trans 算法[48]相比还提升了 0.3%. 特别地, Stanford Dogs 数据集中包含大量相似的背景区域, 这使得大多数单视图方法具有挑战性. 然而, 本文方法中的编码器内部冗余信息过滤模块和外部关键区域定位模块有效地抑制了背景噪声, 显著提升了模型在此类复杂环境中的判别能力. 在 NABirds 数据集上的实验结果更是印证了本文方法的卓越性能. 本文方法不仅大幅度超越了其他算法, 而且在识别准确率上相比目前表现最佳的 ACC-ViT 模型[46]高出 0.6%, 该模型还额外地引入了图卷积网络对目标的结构信息进行建模. 尽管有些方法采用了更深层的卷积网络如 ResNet-101, 其准确率仍低于本文的方法约 1.0%. 值得注意的是, 本文提出的双视图融合识别框架通过精细的调整和优化处理流程, 未引入额外复杂结构, 便实现了性能的提升. 此外, 在更具挑战性的 iNaturalist2017 数据集上, 本文方法同样表现优异, 取得了 74.1% 的识别准确率. 相较于采用相同 ViT-B_16 基础网络的 HI2R 方法[58], 本文方法提高了 0.2%. 更值得关注的是, 与其他主流方法相比, 本文方法表现出明显优势, 例如超越了基于 Swin Transformer 架构的 SwinFG 模型[59] 1.1%, 同时显著领先于传统 ResNet 架构的 SGRN 方法[57] 0.5%. 这一结果充分表明, 本文提出的双视图融合策略不仅能够高效提取细粒度特征, 而且在大型复杂自然环境数据集中具备强大的泛化能力. 综上, 本文所提出的基于 ViT 的双视图融合框架在多个数据集上都展现了显著的优越性.

4.4 消融实验

4.4.1 双视图识别范式的有效性与高效性分析

为验证双视图识别范式在细粒度图像识别任务中的有效性, 本节在 CUB-200-2011、Stanford Dogs 和 NABirds

这 3 个数据集上对预测结果进行了一系列的消融实验. 消融实验结果展示在表 2 中, 其目的是探索基于全局视图的识别范式与基于局部视图的识别范式在模型最终性能上的独立和联合影响. 实验结果表明, 在这 3 个数据集上, 虽然仅使用基于全局视图的识别或仅使用基于局部视图的识别就已经能够获得相当高的识别准确率, 但当将双视图识别范式应用时, 识别准确率得到了更为显著的提升. 此外, 对两种不同视图的识别范式在各数据集上的表现进行比较后发现, 不同数据集的特性可能对基于全局视图或局部视图识别的效果产生不同影响. 例如, 在 Stanford Dogs 数据集上, 基于局部视图的识别效果优于基于全局视图的识别, 这可能与该数据集中图片的背景噪声较大有关. 这一发现提示我们, 在设计细粒度图像识别模型时, 需要根据数据集特性合理融合全局视图与局部视图的上下文信息.

表 2 CUB-200-2011、Stanford Dogs 及 NABirds 数据集上双视图识别范式的消融实验研究 (%)

数据集	全局视图	局部视图	Top-1 准确率
CUB-200-2011	√	—	91.7
	—	√	91.6
	√	√	**92.1**
Stanford Dogs	√	—	90.3
	—	√	91.3
	√	√	**91.7**
NABirds	√	—	90.6
	—	√	90.9
	√	√	**91.9**

为进一步分析本文提出的双视图识别范式的高效性, 本节还在 CUB-200-2011 数据集上对比了本文方法 (包括单视图和双视图形式) 与基线模型 ViT[11]及代表性单视图方法 TransFG[14]的浮点运算次数 (FLOPs) 和 GPU 内存占用情况. 相关实验结果如表 3 所示, 相较于传统单视图 ViT 模型[11]与 TransFG 模型[14], 本文提出的单视图方法在识别准确率方面分别提高了 1.5% 和 1.2%, 而计算复杂度 (以浮点运算次数表示) 与 TransFG 模型基本持平 (约 62 GFLOPs), GPU 内存占用仅增加 0.14 GB. 进一步采用本文的双视图融合识别方法后, 识别准确率又提升了 0.4%, 达到了 92.1%, 显示了明显的性能优势; 而 GPU 内存占用仅略微增加至 0.47 GB, 基本未造成额外负担. 尽管双视图方法的浮点运算次数增加到单视图的约 2 倍 (达到 124.51 GFLOPs), 但鉴于本文所设计的自适应推理策略可根据样本难易程度动态选择是否启用双视图推理, 绝大多数简单样本仍只需单视图即可完成识别任务, 因此实际应用中, 双视图识别仅针对少部分难识别样本执行, 有效控制了整体计算开销. 这表明本文提出的双视图识别范式在性能与效率之间实现了良好平衡.

表 3 在 CUB-200-2011 数据集上单视图识别方法与双视图识别方法的高效性对比

对比方法	准确率 (%)	浮点运算次数 (GFLOPs)	GPU内存占用 (GB)
ViT[11] (单视图)	90.2	67.46	0.32
TransFG[14] (单视图)	90.5	61.97	0.32
本文方法 (单视图)	91.7	62.24	0.46
本文方法 (双视图)	92.1	124.51	0.47

4.4.2 定制化模块的有效性分析

本节深入分析了双视图融合识别框架中各个模块的有效性. 为此, 本文在 CUB-200-2011、Stanford Dogs 和 NABirds 这 3 个数据集上执行了详尽的消融实验, 目的是分析 3 个专门为细粒度图像识别任务设计的模块 (即冗余信息过滤、关键区域定位和局部特征增强) 对模型性能的具体影响, 其结果已展示于表 4 中. 基线模型 (即传统的 ViT 模型) 在 3 个数据集上的表现仅为基础水平. 当引入冗余信息过滤模块后, 模型性能在所有数据集上都实现了显著提升, 尤其是在 CUB-200-2011 和 Stanford Dogs 数据集上, 这证明了冗余信息过滤对于提升模型识别关

键特征的能力的重要性. 而后, 单独加入关键区域定位模块进行的局部视图实验也展示了性能提升, 虽然提升幅度略低于加入冗余信息过滤. 这表明通过定位图像中的关键区域来生成局部视图, 能有效增强模型对局部特征的捕捉能力. 更重要的是, 将关键区域定位与冗余信息过滤结合使用时, 模型性能得到了进一步的提升, 证明了这两个模块的互补性. 此外, 引入局部特征增强模块后, 模型在所有 3 个数据集上均实现了更高的识别准确率, 凸显了局部视图在细粒度识别任务中的重要性. 最终, 将所有模块整合至双视图融合识别框架中 (即同时考虑全局和局部视图), 本文的模型在所有数据集上都取得了最佳性能, CUB-200-2011、Stanford Dogs 和 NABirds 数据集的准确率分别提升至 92.1%、91.7% 和 91.9%. 这种全面的性能提升充分证实了双视图融合识别框架中 3 个定制模块的有效性和创新性.

表 4　双视图识别框架中不同模块在 CUB-200-2011、Stanford Dogs 及 NABirds 数据集上的消融实验研究 (%)

方法	识别视图	CUB-200-2011	Stanford Dogs	NABirds
基线模型	全局	90.2	90.2	89.9
基线模型+冗余信息过滤	全局	91.0	91.0	90.2
基线模型+关键区域定位	局部	90.9	90.8	90.3
基线模型+关键区域定位+冗余信息过滤	局部	91.2	91.1	90.4
基线模型+关键区域定位+冗余信息过滤+局部特征增强	局部	91.6	91.3	90.9
基线模型+关键区域定位+冗余信息过滤+局部特征增强	全局+局部	**92.1**	**91.7**	**91.9**

4.4.3　双视图损失函数的有效性分析

为验证本文提出的基于双视图相似度的对比损失函数的有效性, 本节在 CUB-200-2011、Stanford Dogs 和 NABirds 这 3 个数据集上, 将其与传统的交叉熵损失函数及单视图对比损失函数进行了对比实验. 相关的实验结果呈现在图 7 中. 实验表明, 仅使用交叉熵损失函数时, 模型在 3 个数据集上的准确率分别为 91.7%、91.4% 和 91.6%. 引入单视图对比损失函数后, 模型准确率略有提升. 然而, 当采用基于双视图相似度的对比损失函数后, 模型在 CUB-200-2011、Stanford Dogs 和 NABirds 数据集上的准确率分别显著提升至 92.1%、91.7% 和 91.9%, 这一结果明显优于其他两种损失函数的表现. 此外, 双视图对比损失函数的设计充分考虑了不同子类别间视图特征的差异性和同一子类别内视图特征的一致性. 这种策略有效地放大了类间特征表示的差异性, 同时减少了类内不同视图间的差异性. 这样的方法解决了细粒度识别任务中类间差异较小和类内差异较大的挑战, 不仅提升了识别准确率, 而且增强了模型的泛化能力. 综上所述, 通过对比实验, 本文验证了双视图对比损失函数在细粒度图像识别任务中的有效性. 这种方法不仅提升了模型的识别准确率, 而且增强了双视图融合识别框架对复杂图像中细节特征的学习能力, 为多视图框架的进一步研究提供了新的思路和方法.

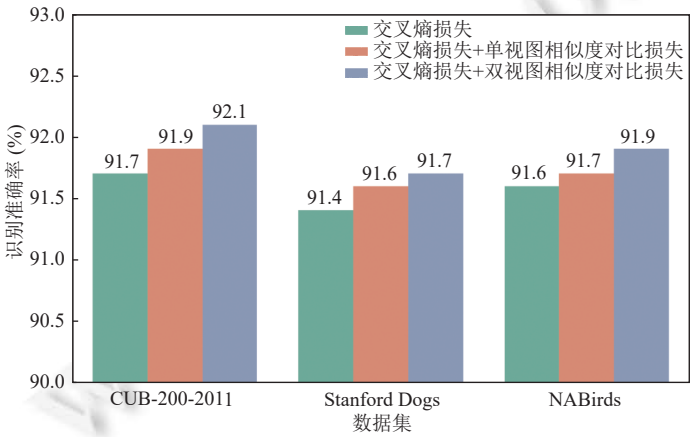

图 7　在 CUB-200-2011、Stanford Dogs 和 NABirds 数据集上对比损失函数的消融实验研究

为进一步分析公式 (14) 中两个视图分类损失函数权重参数对模型性能的影响, 本节在 CUB-200-2011 数据集上进行了详细的消融实验研究. 具体而言, 我们固定全局视图损失函数的权重为 $1-\alpha$, 变化局部视图损失函数的权重 α, 探讨不同 α 值对模型识别准确率的影响. 实验结果如图 8 所示, 当局部视图损失函数权重较小时, 模型无法有效捕获局部视图的细节特征, 导致准确率较低. 随着 α 逐渐增大, 模型对局部视图特征的关注程度逐渐提高, 识别准确率显著提升, 并在 $\alpha=0.5$ 时达到最高 (92.1%). 然而, 当 α 继续增大, 模型可能过度依赖局部视图特征, 忽略全局视图的信息补充作用, 导致识别准确率逐渐降低. 综上所述, 适度平衡全局视图与局部视图损失函数的权重至关重要. 本实验结果验证了所提出的双视图损失函数设计的合理性和有效性, 最优权重比例 (α=0.5) 能够使模型在全局与局部视图之间达到良好的平衡, 从而充分利用双视图融合框架的优势, 有效提高细粒度图像识别的性能.

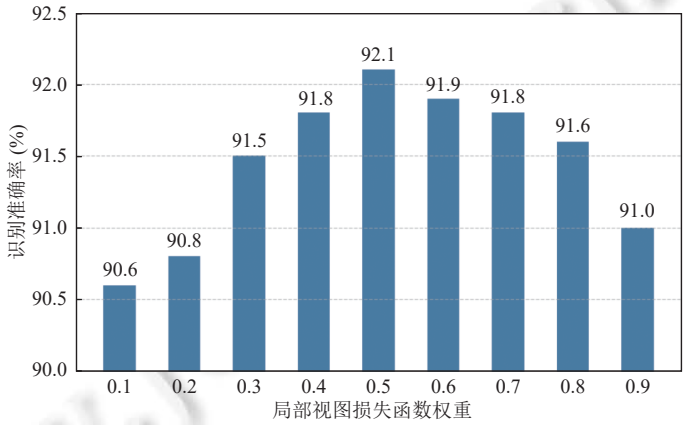

图 8 在 CUB-200-2011 数据集上局部视图损失函数权重的消融实验研究

4.4.4 自适应推理策略的有效性分析

为展现自适应推理策略在双视图融合识别框架中的高效作用, 本节在 CUB-200-2011、Stanford Dogs 和 NABirds 这 3 个数据集上进行了一系列实验, 比较了是否采用基于双视图置信度的自适应推理策略对识别准确率和推理时间的影响. 为了确保推理时间测量的准确性与可靠性, 本节实验采用 PyTorch 1.10 框架的内置时间记录工具进行时间测量. 为稳定 GPU 性能, 实验开始前先以 64 张图像为一个批次, 连续运行 50 个批次进行 GPU 预热. 预热完成后, 以同样的批次大小计算整个数据集的总推理时间. 为保证结果的一致性和准确性, 所有数据集的推理时间测量均重复进行 3 次, 最终结果取 3 次实验的时间平均值. 实验结果如表 5 所示, 在保持准确率不变的同时, 自适应推理策略显著降低了推理时间. 在 CUB-200-2011 数据集上从 142.1 s 减少至 76.6 s, 在 Stanford Dogs 从 209.3 s 减少至 136.1 s, 在 NABirds 从 600.8 s 减少至 359.7 s. 这一显著的推理时间优化证明了自适应推理策略在提高效率方面的显著效用. 自适应推理策略的关键在于基于全局视图识别结果置信度动态评估样本难易程度, 从而在不牺牲准确率的情况下有效减少推理过程中的计算负担. 这一实验结果不仅验证了该策略的有效性, 而且表明它可以作为一种有效手段将本文的双视图融合识别框架应用于大规模数据集上.

表 5 不同推理策略在 CUB-200-2011、Stanford Dogs 和 NABirds 数据集上的性能对比

数据集	自适应推理	准确率 (%)	时间 (s)
CUB-200-2011	—	92.1	142.1
	√	92.1	**76.6**
Stanford Dogs	—	91.7	209.3
	√	91.7	**136.1**
NABirds	—	91.9	600.8
	√	91.9	**359.7**

4.5 参数敏感性分析

4.5.1 累积分布阈值 u 的对比实验分析

本节深入研究了累积分布阈值 u 在基于注意力阈值的关键区域定位模块中的作用, 并在 CUB-200-2011、Stanford Dogs 和 NABirds 这 3 个数据集上进行了对比实验分析, 以揭示阈值 u 对识别准确率的影响, 如图 9 所示. 在 CUB-200-2011 数据集上, 实验结果显示, 当阈值 u 设置为 0.7 时, 识别准确率达到最高 (92.1%), 而在 Stanford Dogs 数据集上, 阈值 u 为 0.9 时达到最高准确率 (91.7%), 在 NABirds 数据集上则是阈值 u 为 0.6 时准确率最高 (91.9%). 这表明不同数据集的最优阈值存在显著差异, 可能与数据集中目标对象的大小、比例及背景比例的差异有关. 进一步分析发现, 当阈值 u 设定过低时, 定位的关键区域过小, 导致局部视图无法覆盖对象的所有重要特征, 从而影响性能. 反之, 阈值 u 过高时, 会导致过多背景信息被包含在局部视图中, 引入的噪声也会降低模型的识别准确率. 因此, 合理设定累积分布阈值 u 对于平衡关键特征的提取和背景噪声的抑制至关重要, 这是提升模型识别性能的关键. 总之, 累积分布阈值 u 的合理设置对于增强模型的识别能力具有显著影响. 本节的实验结果不仅证实了阈值 u 在关键区域定位中的重要作用, 而且为不同特性的数据集提供了阈值选择的参考. 通过这种方法, 可以更精确地捕捉目标对象的关键特征, 从而在细粒度图像识别任务中实现更高的准确率.

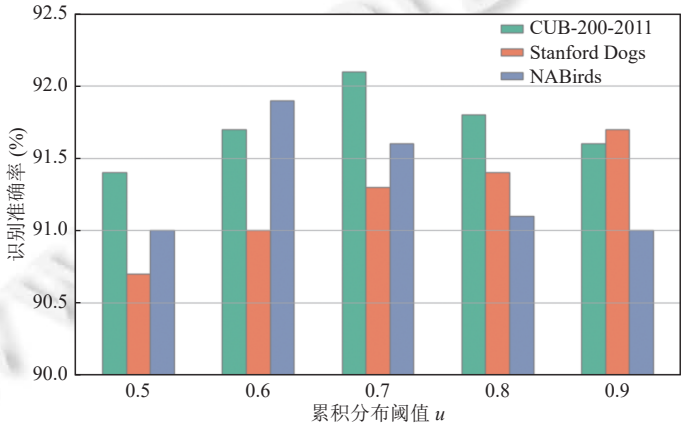

图 9 在 CUB-200-2011、Stanford Dogs 和 NABirds 数据集上累积分布阈值 u 取不同值时的识别准确率对比

4.5.2 冗余信息过滤阈值 K 的影响

本节着重探讨了基于注意力融合的冗余信息过滤模块中阈值 K 对细粒度子类别辨识性差异特征提取能力的影响. 为此, 本文在 CUB-200-2011、Stanford Dogs 和 NABirds 这 3 个数据集上进行了一系列消融实验, 旨在详细分析不同阈值 K 对模型性能的影响, 实验结果呈现在后文图 10. 从实验数据可以看出, 随着冗余信息过滤阈值 K 的增加, 模型的识别准确率呈现上升趋势, 直至阈值 K 达到 50 时, 在 3 个数据集上均实现了最高准确率 (CUB-200-2011 为 92.1%, Stanford Dogs 为 91.7%, NABirds 为 91.9%). 这一结果表明适度的信息过滤有助于提取对细粒度子类别识别更为关键的特征. 然而, 当阈值 K 继续增加至 75 及以上时, 模型性能出现下降, 这可能是因为较高的阈值不但筛选出部分重要的局部区域特征, 同时也保留了过多的背景信息, 这对细粒度类别识别无益. 这一发现揭示了在 ViT 的视觉编码器中, 适当的冗余信息过滤至关重要. 过低的阈值 K 可能导致重要特征的丢失, 而过高的阈值则可能引入干扰信息, 影响模型对细微差异的捕捉能力. 因此, 精心选择冗余信息过滤阈值 K 对于优化 ViT 模型性能、增强其对细粒度类别识别精度具有重要意义.

4.5.3 自适应推理策略中置信度阈值的影响

本节通过在 CUB-200-2011、Stanford Dogs 和 NABirds 这 3 个数据集上进行的参数敏感性实验, 评估了自适应推理策略中置信度阈值对识别准确率及推理效率的影响. 实验结果如图 11 所示. 在 NABirds 数据集上, 随着置信度阈值从 0.5 增至 1.0, 识别准确率从 91.3% 逐步提升至 91.9%, 而推理时间则从 316.4 s 增至 600.8 s. 可以观察

到, 尽管置信度阈值的提高带来了准确率的轻微提升, 但推理时间显著增加. 特别是当阈值从 0.9 增至 1.0 时, 推理时间的增加并未带来准确率的进一步提升, 这暗示了阈值设置为 1.0 时的低效性. 对于 CUB-200-2011 数据集, 识别准确率在置信度阈值 0.8 时达到 92.1%, 随后保持稳定, 但推理时间持续增加, 尤其是阈值为 1.0 时, 推理时间几乎翻倍, 达到 141.3 s. 在 Stanford Dogs 数据集上, 置信度阈值对识别准确率的影响较稳定, 保持在 91.6%–91.7% 之间, 而推理时间则从 117.5 s 逐步上升至 209.3 s, 尤其在阈值 0.9–1.0 之间的增长并未带来显著的准确率提升. 综合 3 个数据集上的实验结果, 本文选择 0.8 作为统一的置信度阈值, 为不同特性的数据集提供一个稳健的推理策略, 实现了识别性能与推理效率的最佳平衡. 这一决策旨在不牺牲识别准确性的同时, 避免过度增加推理时间. 本节的实验研究不仅展示了自适应推理策略在提升细粒度图像识别任务效率方面的有效性, 还证明了通过对不同数据集的置信度阈值进行精细调整, 可以实现识别性能与推理效率的最佳折中.

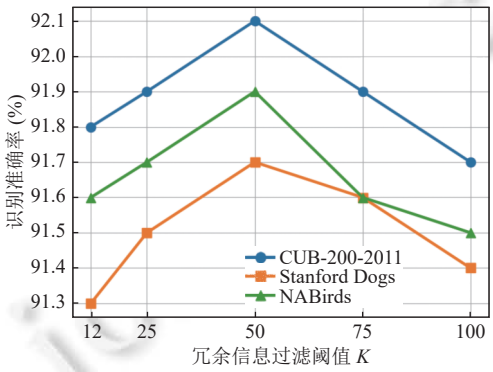

图 10　在 CUB-200-2011、Stanford Dogs 和 NABirds 数据集上冗余信息过滤阈值 K 对识别准确率的影响

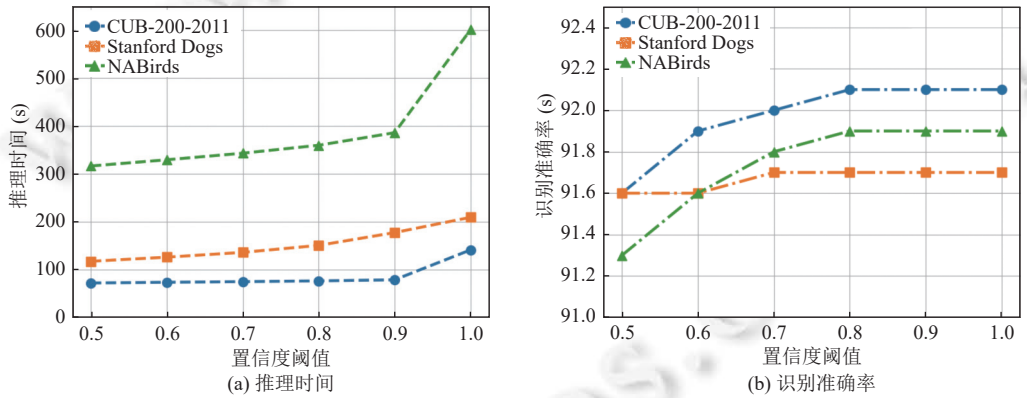

图 11　在 CUB-200-2011、Stanford Dogs 和 NABirds 数据集上自适应推理策略中置信度阈值的参数敏感性分析

4.6　可视化分析

4.6.1　冗余信息过滤效果的可视化

本节通过一系列可视化实验直观地展示了基于注意力融合的冗余信息过滤模块的效果. 图 12(a) 清晰展示了模型是如何从全局视图和局部视图的图像块中筛选出与分类标记嵌入关联度较高的区域特征. 通过观察所提供的可视化图像, 可以看到模型对某些关键区域展现出更高的关注度. 这些区域在可视化图像中以高亮像素呈现, 突显了模型识别过程中的焦点区域. 在全局视图中, 尽管背景复杂, 模型依然能够准确聚焦于鸟类所在区域, 这证明了注意力融合机制在识别关键信息方面的有效性. 在局部视图中, 模型对鸟类的细节特征, 如头部和翅膀等, 展现出更精细的关注, 高亮区域准确展示了模型在进行局部特征分析时的精确性. 这进一步证实了本文提出的冗余信息

过滤模块在聚焦细粒度特征上的效力. 此外, 通过比较全局视图和局部视图中的关注区域, 可以推断出模型在全局视图中已初步识别出关键特征, 并在局部视图中进一步强化这些特征的学习. 这种从全局到局部的递进式注意力关注机制有效缓解了背景噪声带来的干扰, 确保了关键特征被模型有效捕捉并加以强化.

(a) 所提出模型中基于注意力融合的冗余信息过滤示例　(b) 所提出模型中基于注意力阈值的关键区域定位示例　(c) 所提出模型在处理具有不同背景图片时的注意力分布可视化图

图 12　本文方法在 CUB-200-2011 数据集上的可视化分析

4.6.2　关键区域定位效果的可视化

本节通过可视化分析成功地展示了基于注意力阈值的关键区域定位模块在细粒度图像识别中的应用效果. 如图 12(b) 所示的可视化结果清楚地揭示了模型如何在全局视图中通过自适应阈值识别并突出关键区域, 以及如何将这些区域转化为局部视图进行深入分析. 可以看出, 在全局视图 (第 1 列) 的基础上, 模型生成了图像块掩码矩阵 (第 2 列). 在这些掩码矩阵中, 较亮的像素块表示高于自适应阈值的区域, 被模型识别为包含关键信息和对分类判别至关重要的部分. 较暗的区域则表示注意力分数较低, 通常与图像背景或不重要的前景相关联. 第 1 列展示了模型处理的原始全局图像, 这些图像包含了丰富的背景信息, 可能为细粒度类别识别带来干扰. 第 2 列中通过自适应阈值计算得到的图像块掩码矩阵清晰指出了模型所关注的区域, 这些较亮区域的边界与背景噪声形成鲜明对比, 突显了模型认为重要的图像块. 第 3 列的局部视图则展示了经过裁剪和放大的关键区域, 对区分相似子类别至关重要. 这些可视化结果清晰地表明, 本文提出的关键区域定位模块有效地模仿了人类的视觉处理机制, 即在复杂场景中优先关注对任务有决定性影响的视觉信息. 通过自适应阈值的引入, 该模块能够灵活适应不同图像的特性和复杂度, 准确地定位关键视觉特征, 并有效地抑制背景噪声.

4.7　局限性与未来工作方向

4.7.1　模型的泛化能力

本文所提出的双视图融合识别框架在细粒度数据集如 CUB-200-2011、Stanford Dogs、NABirds 和 iNaturalist2017 上展现出了出色的识别性能. 然而, 尽管在这些数据集上取得了高准确率, 但它们的图像背景通常具有较高的一致性和简化性, 未能充分反映自然环境中的复杂性. 在自然环境下, 多变的光照条件、复杂多样的背景噪声以及动态变化的观察视角可能会挑战模型的泛化能力, 导致识别准确率下降. 这主要归因于模型未能有效区分图像的前景与复杂背景噪声. 图 12(c) 清楚地展示了这一问题. 在图中第 1 行, 可以观察到模型在处理简单背景时, 注意力集中在图像的关键前景区域; 而在第 2 行与第 3 行中, 当背景变得复杂时, 模型的注意力分布变得高度分散, 导致模型难以有效识别和聚焦于主要目标. 为了进一步提升模型在实际应用中的泛化能力和实用性, 未来工作计划采取以下几种策略: (1) 背景抑制技术: 深入研究并开发新的算法以更好地区分和抑制背景噪声, 例如通过改进模型的注意力机制, 增强其对图像关键区域的聚焦能力, 而非背景; (2) 多环境数据训练: 在训练阶段引入具有多种背景的图像数据, 使模型能够适应并在多样化的环境中维持性能稳定. 这包括使用生成对抗网络生成的具

有挑战性背景的训练样本, 以提高模型的适应性和鲁棒性; (3) 域适应技术: 利用域适应方法来缩小训练数据与真实应用场景之间的差异, 尤其是针对背景复杂度和样本分布的不同, 以增强模型对复杂环境的适应能力.

4.7.2 模型的可解释性

尽管本文提出的双视图细粒度识别框架在多个细粒度图像识别任务上表现出色, 但其决策过程的透明度与可解释性仍存在一定局限. 当前模型基于深度视觉 Transformer 架构, 该架构能够通过分析大量图像数据捕捉复杂的视觉特征. 然而, Transformer 内部的决策机制相对封闭, 使得模型的决策原理难以被直观理解. 图 12(c) 展示了模型在识别过程中的注意力分布, 尽管背景复杂且目标物种间的差异微小, 模型依然能够准确识别并聚焦于关键特征. 虽然能够观察到模型在处理图像时对特定区域的聚焦, 但这些区域的选择以及它们在模型分类决策中的具体语义贡献尚缺乏明确的解释. 为了克服这一局限性, 未来的研究需致力于提升模型的可解释性, 进一步揭示模型内部的决策逻辑. 一个可行的策略是引入先进的可解释性框架或算法, 如集成梯度或相关传播算法. 这些方法能够揭示 Transformer 层间信息流动的具体模式及其对最终决策的影响, 从而使模型决策过程更加透明. 通过引入这些可解释性方法, 可以定量地评估图像各区域在决策中的重要性, 并详细解释这些区域如何影响整体的识别结果.

5 总 结

本文提出了一种基于视觉 Transformer (ViT) 的细粒度图像识别框架, 该框架采用了双视图识别范式. 通过有效整合全局视图和局部视图, 显著地克服了 ViT 在细粒度图像识别任务中的局限性, 尤其是在处理细节特征和背景噪声方面. 在 4 个标准的细粒度图像识别数据集上进行的广泛实验结果证明了该双视图识别框架的有效性, 并展现了其在解决细粒度图像识别领域实际挑战中的应用潜力, 为未来相关研究领域提供了新的视角和方法.

References

[1] Zhang S, Gong YH, Wang JJ. The development of deep convolution neural network and its applications on computer vision. Chinese Journal of Computers, 2019, 42(3): 453–482 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2019.00453]

[2] Luo JH, Wu JX. A survey on fine-grained image categorization using deep convolutional features. Acta Automatica Sinica, 2017, 43(8): 1306–1318 (in Chinese with English abstract). [doi: 10.16383/j.aas.2017.c160425]

[3] Wei XS, Song YZ, Aodha OM, Wu JX, Peng YX, Tang JH, Yang J, Belongie S. Fine-grained image analysis with deep learning: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(12): 8927–8948. [doi: 10.1109/TPAMI.2021.3126648]

[4] Choudhury S, Laina I, Rupprecht C, Vedaldi A. The curious layperson: Fine-grained image recognition without expert labels. Int'l Journal of Computer Vision, 2024, 132(2): 537–554. [doi: 10.1007/s11263-023-01885-9]

[5] Qi L, Yu PZ, Gao Y. Research on weak-supervised person re-identification. Ruan Jian Xue Bao/Journal of Software, 2020, 31(9): 2883–2902 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6083.htm [doi: 10.13328/j.cnki.jos.006083]

[6] Wei XS, Cui Q, Yang L, Wang P, Liu LQ, Yang J. RPC: A large-scale and fine-grained retail product checkout dataset. Science China Information Sciences, 2022, 65(9): 197101. [doi: 10.1007/s11432-022-3513-y]

[7] Dou H, Zhang LM, Han F, Shen FR, Zhao J. Survey on convolutional neural network interpretability. Ruan Jian Xue Bao/Journal of Software, 2024, 35(1): 159–184 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6758.htm [doi: 10.13328/j.cnki.jos.006758]

[8] Tan M, Yuan F, Yu J, Wang GJ, Gu XL. Fine-grained image classification via multi-scale selective hierarchical biquadratic pooling. ACM Trans. on Multimedia Computing, Communications, and Applications, 2022, 18(1s): 31. [doi: 10.1145/3492221]

[9] Chang DL, Ding YF, Xie JY, Bhunia AK, Li XX, Ma ZY, Wu M, Guo J, Song YZ. The devil is in the channels: Mutual-channel loss for fine-grained image classification. IEEE Trans. on Image Processing, 2020, 29: 4683–4695. [doi: 10.1109/TIP.2020.2973812]

[10] He XT, Peng YX. Unsupervised fine-grained video categorization via adaptation learning across domains and modalities. Ruan Jian Xue Bao/Journal of Software, 2021, 32(11): 3482–3495 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6058.htm [doi: 10.13328/j.cnki.jos.006058]

[11] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021. 1–21.

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

[13] Yin ZZ, Li BH, Wang M, He RL, Wu WL, Wang HF. Partial multimodal hashing based on fine-grained feature fusion. Ruan Jian Xue Bao/Journal of Software, 2024, 35(3): 1074–1089 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7076.htm [doi: 10.13328/j.cnki.jos.007076]

[14] He J, Chen JN, Liu S, Kortylewski A, Yang C, Bai YT, Wang CH. TransFG: A Transformer architecture for fine-grained recognition. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI Press, 2022. 852–860. [doi: 10.1609/aaai.v36i1.19967]

[15] Sun HB, He XT, Peng YX. SIM-Trans: Structure information modeling Transformer for fine-grained visual categorization. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisbon: ACM, 2022. 5853–5861. [doi: 10.1145/3503161.3548308]

[16] Xu Q, Wang JH, Jiang B, Luo B. Fine-grained visual classification via internal ensemble learning Transformer. IEEE Trans. on Multimedia, 2023, 25: 9015–9028. [doi: 10.1109/TMM.2023.3244340]

[17] Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD birds-200-2011 dataset. 2011. https://gwern.net/doc/ai/dataset/2011-wah.pdf

[18] Khosla A, Jayadevaprakash N, Yao B, Li FF. Novel dataset for fine-grained image categorization: Stanford Dogs. In: Proc. of the 1st Workshop on Fine-grained Visual Categorization. Colorado Springs: CVPR, 2011.

[19] van Horn G, Branson S, Farrell R, Haber S, Barry J, Ipeirotis P, Perona P, Belongie S. Building a bird recognition APP and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 595–604. [doi: 10.1109/CVPR.2015.7298658]

[20] van Horn G, Aodha OM, Song Y, Cui Y, Sun C, Shepard A, Adam H, Perona P, Belongie S. The iNaturalist species classification and detection dataset. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City, 2018. 8769–8778. [doi: 10.1109/CVPR.2018.00914]

[21] Gavves E, Fernando B, Snoek CGM, Smeulders AWM, Tuytelaars T. Fine-grained categorization by alignments. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney, 2013. 1713–1720. [doi: 10.1109/ICCV.2013.215]

[22] Branson S, van Horn G, Wah C, Perona P, Belongie S. The ignorant led by the blind: A hybrid human-machine vision system for fine-grained categorization. Int'l Journal of Computer Vision, 2014, 108(1): 3–29. [doi: 10.1007/s11263-014-0698-4]

[23] Zhang N, Donahue J, Girshick R, Darrell T. Part-based R-CNNs for fine-grained category detection. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 834–849. [doi: 10.1007/978-3-319-10590-1_54]

[24] Xiao TJ, Xu YC, Yang KY, Zhang JX, Peng YX, Zhang Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 842–850. [doi: 10.1109/CVPR.2015.7298685]

[25] Fu JL, Zheng HL, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4476–4484. [doi: 10.1109/CVPR.2017.476]

[26] Ding Y, Zhou YZ, Zhu Y, Ye QX, Jiao JB. Selective sparse sampling for fine-grained image recognition. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 6598–6607. [doi: 10.1109/ICCV.2019.00670]

[27] Ding YF, Ma ZY, Wen SG, Xie JY, Chang DL, Si ZW, Wu M, Ling HB. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification. IEEE Trans. on Image Processing, 2021, 30: 2826–2836. [doi: 10.1109/TIP.2021.3055617]

[28] Yu CJ, Zhao XY, Zheng Q, Zhang P, You XG. Hierarchical bilinear pooling for fine-grained visual recognition. In: Proc. of the 15th European Conf on Computer Vision. Munich: Springer, 2018. 595–610. [doi: 10.1007/978-3-030-01270-0_35]

[29] Zhuang PQ, Wang YL, Qiao Y. Learning attentive pairwise interaction for fine-grained classification. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI Press, 2020. 13130–13137. [doi: 10.1609/aaai.v34i07.7016]

[30] Zhang Y, Cao J, Zhang L, Liu XC, Wang ZY, Ling F, Chen WQ. A free lunch from ViT: Adaptive attention multi-scale fusion Transformer for fine-grained visual recognition. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Singapore, Singapore: IEEE, 2022. 3234–3238. [doi: 10.1109/ICASSP43922.2022.9747591]

[31] He KM, Chen XL, Xie SN, Li YH, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 15979–15988. [doi: 10.1109/CVPR52688.2022.01553]

[32] Xie ZD, Zhang Z, Cao Y, Lin YT, Bao JM, Yao ZL, Dai Q, Hu H. SimMIM: A simple framework for masked image modeling. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 9653–9663. [doi: 10.1109/CVPR52688.2022.00943]

[33] Tang H, Yuan CC, Li ZC, Tang JH. Learning attention-guided pyramidal features for few-shot fine-grained recognition. Pattern Recognition, 2022, 130: 108792. [doi: 10.1016/j.patcog.2022.108792]

[34] Zha ZC, Tang H, Sun YL, Tang JH. Boosting few-shot fine-grained recognition with background suppression and foreground alignment. IEEE Trans. on Circuits and Systems for Video Technology, 2023, 33(8): 3947–3961. [doi: 10.1109/TCSVT.2023.3236636]

[35] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proc. of the 27th Int'l Conf. on Machine Learning. Haifa: Omnipress, 2010. 807–814.

[36] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[37] Zheng HL, Fu JL, Mei T, Luo JB. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 5209–5217. [doi: 10.1109/ICCV.2017.557]

[38] Yang Z, Luo TG, Wang D, Hu ZQ, Gao J, Wang LW. Learning to navigate for fine-grained classification. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 420–435. [doi: 10.1007/978-3-030-01264-9_26]

[39] Luo W, Yang XT, Mo XJ, Lu YH, Davis L, Li J, Yang J, Lim SN. Cross-X learning for fine-grained visual categorization. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 8241–8250. [doi: 10.1109/ICCV.2019.00833]

[40] Chen Y, Bai YL, Zhang W, Mei T. Destruction and construction learning for fine-grained image recognition. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5152–5161. [doi: 10.1109/CVPR.2019.00530]

[41] Xu KR, Lai R, Gu L, Li YS. Multiresolution discriminative mixup network for fine-grained visual categorization. IEEE Trans. on Neural Networks and Learning Systems, 2023, 34(7): 3488–3500. [doi: 10.1109/TNNLS.2021.3112768]

[42] Liu CB, Xie HT, Zha ZJ, Ma LF, Yu LY, Zhang YD. Filtration and distillation: Enhancing region attention for fine-grained visual categorization. In: Proc. of the 34th AAAI Conf on Artificial Intelligence. New York: AAAI Press, 2020. 11555–11562. [DOI: 10.1609/aaai.v34i07.6822]

[43] Du RY, Xie JY, Ma ZY, Chang DL, Song YZ, Guo J. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(12): 9521–9535. [doi: 10.1109/TPAMI.2021.3126668]

[44] Rao YM, Chen GY, Lu JW, Zhou J. Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 1005–1014. [doi: 10.1109/ICCV48922.2021.00106]

[45] Wang Q, Wang JJ, Deng HY, Wu X, Wang YZ, Hao GF. AA-Trans: Core attention aggregating Transformer with information entropy selector for fine-grained visual classification. Pattern Recognition, 2023, 140: 109547. [doi: 10.1016/j.patcog.2023.109547]

[46] Zhang ZC, Chen ZD, Wang YX, Luo X, Xu XS. A vision Transformer for fine-grained classification by reducing noise and enhancing discriminative information. Pattern Recognition, 2024, 145: 109979. [doi: 10.1016/j.patcog.2023.109979]

[47] Jiang X, Tang H, Gao JY, Du XY, He SF, Li ZC. Delving into multimodal prompting for fine-grained visual classification. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2024. 2570–2578. [doi: 10.1609/aaai.v38i3.28034]

[48] Wang CM, Fu HY, Ma HD. Multi-part token Transformer with dual contrastive learning for fine-grained image classification. In: Proc. of the 31st ACM Int'l Conf. on Multimedia. Ottawa: ACM, 2023. 7648–7656. [doi: 10.1145/3581783.3612303]

[49] Bi Q, Zhou BC, Ji W, Xia GS. Universal fine-grained visual categorization by concept guided learning. IEEE Trans. on Image Processing, 2025, 34: 394–409. [doi: 10.1109/TIP.2024.3523802]

[50] Guo P, Farrell R. Aligned to the object, not to the image: A unified pose-aligned representation for fine-grained recognition. In: Proc. of the 2019 IEEE Winter Conf. on Applications of Computer Vision. Waikoloa: IEEE, 2019. 1876–1885. [doi: 10.1109/WACV.2019.00204]

[51] Zhao YF, Yan K, Huang FY, Li J. Graph-based high-order relation discovery for fine-grained recognition. In: Proc. of 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15074–15083. [doi: 10.1109/CVPR46437.2021.01483]

[52] Cui Y, Song Y, Sun C, Howard A, Belongie S. Large scale fine-grained categorization and domain-specific transfer learning. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4109–4118. [doi: 10.1109/CVPR.2018.00432]

[53] Zhang LB, Huang SL, Liu W, Tao DC. Learning a mixture of granularity-specific experts for fine-grained categorization. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 8330–8339. [doi: 10.1109/ICCV.2019.00842]

[54] Recasens A, Kellnhofer P, Stent S, Matusik W, Torralba A. Learning to zoom: A saliency-based sampling layer for neural networks. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 52–67. [doi: 10.1007/978-3-030-01240-3_4]

[55] Huang ZX, Li Y. Interpretable and accurate fine-grained recognition via region grouping. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8659–8669. [doi: 10.1109/CVPR42600.2020.00869]

[56] Zheng HL, Fu JL, Zha ZJ, Luo JB. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5007–5016. [doi: 10.1109/CVPR.2019.00515]

[57] Wang SJ, Wang ZH, Li HJ, Chang JL, Ouyang WL, Tian Q. Accurate fine-grained object recognition with structure-driven relation graph

networks. Int'l Journal of Computer Vision, 2024, 132(1): 137–160. [doi: 10.1007/s11263-023-01873-z]

[58] Chen L, Wang QC, Li ZH, Yin YL. Hypergraph-guided intra- and inter-category relation modeling for fine-grained visual recognition. In: Proc. of the 32nd ACM Int'l Conf. on Multimedia. Melbourne: ACM, 2024. 8043–8052. [doi: 10.1145/3664647.3680589]

[59] Ma ZP, Wu XY, Chu AZ, Huang L, Wei ZQ. SwinFG: A fine-grained recognition scheme based on swin Transformer. Expert Systems with Applications, 2024, 244: 123021. [doi: 10.1016/j.eswa.2023.123021]

附中文参考文献

[1] 张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用. 计算机学报, 2019, 42(3): 453–482. [doi: 10.11897/SP.J.1016.2019.00453]

[2] 罗建豪, 吴建鑫. 基于深度卷积特征的细粒度图像分类研究综述. 自动化学报, 2017, 43(8): 1306–1318. [doi: 10.16383/j.aas.2017.c160425]

[5] 祁磊, 于沛泽, 高阳. 弱监督场景下的行人重识别研究综述. 软件学报, 2020, 31(9): 2883–2902. http://www.jos.org.cn/1000-9825/6083.htm [doi: 10.13328/j.cnki.jos.006083]

[7] 窦慧, 张凌茗, 韩峰, 申富饶, 赵健. 卷积神经网络的可解释性研究综述. 软件学报, 2024, 35(1): 159–184. http://www.jos.org.cn/1000-9825/6758.htm [doi: 10.13328/j.cnki.jos.006758]

[10] 何相腾, 彭宇新. 跨域和跨模态适应学习的无监督细粒度视频分类. 软件学报, 2021, 32(11): 3482–3495. http://www.jos.org.cn/1000-9825/6058.htm [doi: 10.13328/j.cnki.jos.006058]

[13] 殷崭祚, 李博涵, 王萌, 黄瑞龙, 吴文隆, 王昊奋. 基于细粒度特征融合的部分多模态哈希. 软件学报, 2024, 35(3): 1074–1089. http://www.jos.org.cn/1000-9825/7076.htm [doi: 10.13328/j.cnki.jos.007076]

作者简介

唐昊, 博士, 主要研究领域为机器学习, 计算机视觉, 多模态学习.

李泽超, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为多媒体分析与检索, 计算机视觉, 人工智能.

蒋鑫, 博士生, 主要研究领域为机器学习, 计算机视觉, 多模态学习.

唐金辉, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为多媒体分析与检索, 社交媒体分析, 计算机视觉, 人工智能.