融合目标定位与异构局部交互学习的细粒度图像分类

陈 权[1] 陈 飞[1] 王衍根[1] 程 航[2] 王美清[2]

摘 要 由于细粒度图像之间存在小的类间方差和大的类内差异, 现有分类算法仅仅聚焦于单张图像的显著局部特征的提取与表示学习, 忽视了多张图像之间局部的异构语义判别信息, 较难关注到区分不同类别的微小细节, 导致学习到的特征缺乏足够区分度. 本文提出了一种渐进式网络以弱监督的方式学习图像不同粒度层级的信息. 首先, 构建一个注意力累计目标定位模块 (Attention accumulation object localization module, AAOLM), 在单张图像上从不同的训练轮次和特征提取阶段对注意力信息进行语义目标集成定位. 其次, 设计一个多张图像异构局部交互图模块 (Heterogeneous local interactive graph module, HLIGM), 提取每张图像的显著性局部区域特征, 在类别标签引导下构建多张图像的局部区域特征之间的图网络, 聚合局部特征增强表示的判别力. 最后, 利用知识蒸馏将异构局部交互图模块产生的优化信息反馈给主干网络, 从而能够直接提取具有较强区分度的特征, 避免了在测试阶段建图的计算开销. 通过在多个数据集上进行的实验, 证明了提出方法的有效性, 能够提高细粒度分类的精度.

关键词 深度学习, 细粒度图像分类, 弱监督目标定位, 图神经网络, 知识蒸馏

引用格式 陈权, 陈飞, 王衍根, 程航, 王美清. 融合目标定位与异构局部交互学习的细粒度图像分类. 自动化学报, 2024, **50**(11): 2219−2230

DOI 10.16383/j.aas.c230507 **CSTR** 32138.14.j.aas.c230507

Fine-grained Image Classification by Integrating Object Localization and Heterogeneous Local Interactive Learning

CHEN Quan[1] CHEN Fei[1] WANG Yan-Gen[1] CHENG Hang[2] WANG Mei-Qing[2]

Abstract Due to the existence of small inter-class differences and large intra-class variance among fine-grained images, the existing classification algorithms only focus on the extraction and representation learning of salient local features of a single image, ignoring the local heterogeneous semantic discrimination information between multiple images, difficult to pay attention to the subtle details that distinguish different categories, resulting in the lack of sufficient discrimination of the learned features. This paper proposes a progressive network to learn the information of different granularity levels of the image in a weakly supervised manner. First, attention accumulation object localization module (AAOLM) is constructed to perform semantic target integration localization on attention information from different training epochs and feature extraction stages on a single image. Second, a multi-image heterogeneous local interactive graph module (HLIGM) is designed to construct a graph network and aggregate information between the local region features of multiple images under the guidance of the category label after extracting the salient local region features of each image to enhance the discriminative power of the representation. Finally, the optimization information generated by HLIGM is fed back to the backbone by using knowledge distillation so that it can directly extract features with strong discrimination, avoiding the computational overhead of building the graph in the test phase. Through experiments on multiple data sets, it proves the effectiveness of the proposed method, which can improve the fine-grained classification accuracy.

Key words Deep learning, fine-grained image classification, weakly supervised object localization, graph neural network, knowledge distillation

Citation Chen Quan, Chen Fei, Wang Yan-Gen, Cheng Hang, Wang Mei-Qing. Fine-grained image classification by integrating object localization and heterogeneous local interactive learning. *Acta Automatica Sinica*, 2024, **50**(11): 2219−2230

细粒度图像分类是计算机视觉和模式识别领域一项长久并且极具挑战的研究课题, 在现实世界中有着广泛的应用. 不同于普通的图像分类, 细粒度图像分类旨在对粗粒度的大类别进行更加细致的子类划分, 由于不同类别的目标外观相似, 类间差异

收稿日期 2023-08-16 录用日期 2024-03-07
Manuscript received August 16, 2023; accepted March 7, 2024
国家自然科学基金 (61771141, 62172098), 福建省自然科学基金 (2021J01620) 资助
Supported by National Natural Science Foundation of China (61771141, 62172098) and Natural Science Foundation of Fujian Province (2021J01620)
本文责任编委 张向荣
Recommended by Associate Editor ZHANG Xiang-Rong
1. 福州大学计算机与大数据学院 福州 350108 2. 福州大学数学与统计学院 福州 350108

1. College of Computer and Data Science, Fuzhou University, Fuzhou 350108 2. School of Mathematics and Statistics, Fuzhou University, Fuzhou 350108

只存在于显著性即具有判别力的局部部位的细微不同, 并且同一类别不同的图像可能因为目标姿态、光照、背景等干扰有着巨大的方差, 使得细粒度的图像分类更加具有挑战性.

在研究[1]中发现, 许多方法通常从两个方面设计解决细粒度分类的问题:"更有区分度的表征学习"[2−5]和"定位目标特征显著的部分"[6−9]. 文献 [10−11] 学习局部显著特征并将它们直接进行拼接来提高特征的判别力, 却没有考虑到局部之间其实存在着一定的关系. 一些研究[12−13]开始对局部之间的相关信息进行学习, 但是这些研究本质上都只是在单个图像上通过类别损失监督对特征提取网络进行优化, 如图 1 所示 (实线部分), 仅关注于特征单独的分类特性, 网络较难注意到区别于其他类别的局部的细微处, 忽略了特征空间整体的聚簇特性——相同类别的特征内聚、不同类别的特征疏离的程度. 由于细粒度数据集的特点, 数据在空间的分布往往会更加离散, 如果不考虑不同图像局部区域之间存在的语义联系, 可能会遗漏很多本该关注的信息, 导致无法学到有区分度的表示. 除此之外, 由于在图像中前景目标所处的位置和尺度大小不尽相同, 直接在原图中使用预先设计的固定大小的锚框采样目标显著的局部部件并不是一个好的选择. 如图 2 所示, 采样的锚框无法对不同大小的目标都很好地适配, 对于小目标的图像, 大的锚框会包含更多的背景噪声, 网络更难准确地挖掘到关键的局部部位. 同时, 在文献 [14] 中发现, 一些无关的背景信息会被用于识别, 这都会影响到模型整体的性能. 因此, 应该将不同图像的目标调整到相同尺度的大小再进行采样.

本文设计了一种弱监督学习的渐进式框架来解决细粒度分类的这些问题, 包含注意力累计目标定位模块 (Attention accumulation object localization module, AAOLM) 和异构局部交互图模块 (Heterogeneous local interactive graph module, HLIGM). 注意力累计目标定位模块通过提取到的特征计算出对应的注意力图来定位目标, 裁剪得到细节信息更精细的目标图像, 再采样各个显著的局部部件. 然而, 由于模型随着训练会渐渐只聚焦于图像的某个局部区域, 为了定位到一个结构更加完整的目标获取到更多有用的信息, 在文献 [15] 中发现, 在训练的不同轮次, 模型的关注点会在语义对象的不同区域变换并且它们之间是互补的, 因此, 在训练的每个轮次会使用先前轮次生成的注意力图来更新当前计算的注意力图的各个响应值, 从而保存高响应区域, 同时, 更进一步对不同高层特征生成的注意力图进行集成. 对于提出的异构局部交互图模块, 细粒度数据集图像由于目标只存在局部的细微不同, 依靠单个图像学习到的特征较难判别, 因此, 在图中使用了来自不同图像的局部特征作为节点构建一个完整的图. 如图 1 所示, 在图中产生了两种不同类型的节点对构成的边, 分别为连接相同类别的局部区域构成的正样本对的正对边和连接不同类别的局部区域构成的负样本对的负对边. 为了同时学习不同类型的局部关系, 受对比学习的启发, 本

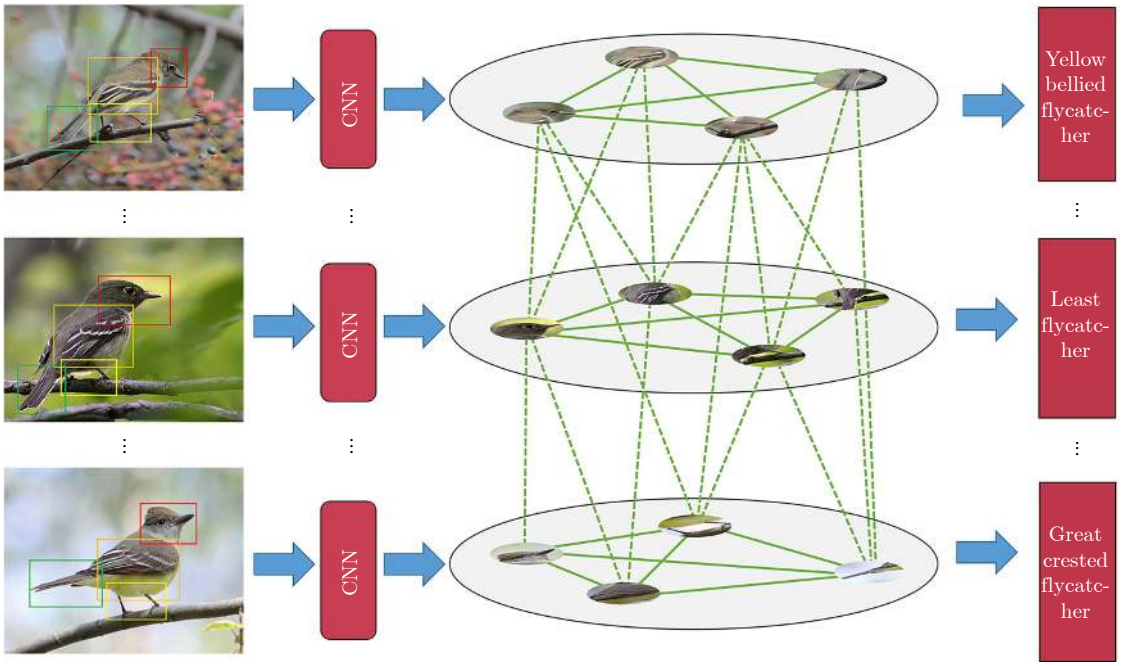
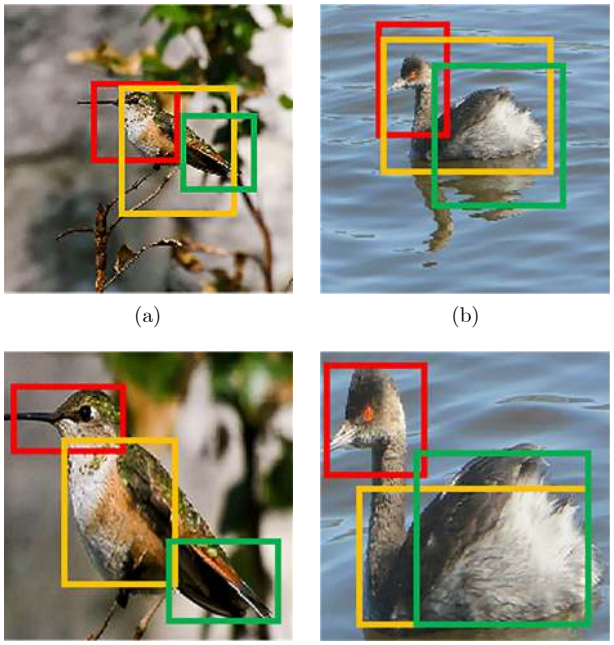


图 1 异构局部交互图模块说明图

Fig. 1 Illustration of HLIGM



(a) (b) (c) (d)

图 2 原始图像和目标图像上的部件采样对比 ((a)、(b) 使用固定大小的锚框直接在原图中采样有用的目标局部部件, 没有很好地区分开不同的部件并且包含了更多无关的背景信息; (c)、(d) 展示了定位到目标后放大到一定的尺度再进行部件采样的效果)

Fig. 2 Comparison of patch sampling between original images and object images ((a)、(b) show that using the fixed-size anchor directly samples useful local patches of the object in the original image, which does not distinguish different patches well and contains more irrelevant background information; (c)、(d) show the effect of patch sampling that it is zoomed in to a certain scale after the object is located)

文使用了一个注意力正则化损失来约束不同类型的边计算的权重, 在类别间形成对比, 增强相同类别而弱化不同类别局部间的语义关系来正确地描述局部关系, 让局部特征学习到对其类别更加显著的信息, 从而取得更好的聚簇特性. 除此之外, 本文还通过知识蒸馏将异构局部交互图模块学习到的优化信息反馈给特征提取网络, 让网络能够直接提取具有区分度的图像特征表示.

总结起来, 本文主要贡献如下: 1) 提出了注意力累计目标定位模块, 在单张图像上能够从不同的训练轮次和特征提取阶段对注意力信息进行语义目标集成定位, 从而排除无关背景噪声的干扰. 2) 提出构建异构局部交互图, 学习多张图像局部部件之间存在的语义联系并且针对图中不同类型的边进行相应的损失约束, 从而能够增强特征表示的判别力. 3) 建立了一个用于图像细粒度分类的多流网络, 能够学习从粗到细不同粒度的特征, 有效地结合图像的全局信息和局部信息, 在多个不同的数据集上和许多同类型方法进行对比, 该方法能够取得更好的表现.

1 相关工作

1.1 基于定位−识别的细粒度分类方法

基于定位−识别的细粒度分类方法聚焦于寻找能够区分目标的显著性局部部件. 递归注意力卷积神经网络 (Recurrent attention convolutional neural network, RA-CNN)[6] 训练了一个额外的子网络来预测图像中目标区域的三元组坐标, 通过一种递归学习的策略获取到不同粒度的特征信息. 文献 [16] 使用注意力机制在不同粒度等级的标签监督下学习从目标到局部的信息, 并且进行跨层级的特征融合学习. 特定粒度专家混合卷积神经网络 (Mixture of granularity-specific experts convolutional neural network, MGE-CNN)[17] 设计了多个专家网络, 基于类激活图 (Class activation mapping, CAM) 以一种逐渐增强的策略在网络间实现知识的传递, 让模型逐渐从目标聚焦到判别性局部区域. 文献 [18] 使用梯度加权类激活图 (Gradient-weighted class activation mapping, Grad-CAM), 不需要额外的训练从图像中检测出目标, 随后再进一步对关键的区域进行采样和学习. 但是, 这些方法往往容易受到图像背景的干扰并且主要依赖于单张图像的特征提取与学习, 忽略了多张图像之间的相似目标的局部判别信息.

1.2 基于图关系网络的细粒度分类方法

图神经网络 (Graph neural network, GNN) 利用深度学习技术为非欧式结构数据提供了有效的处理范式, 它通过平滑邻居节点的消息传递和聚合机制, 在图表示的非欧式结构数据方面展现了强大的处理能力. 图神经网络在许多领域得到了应用, 例如目标检测、关系推理等, 但是在细粒度图像分类领域, 仍未被充分探索. 基于关系学习的图传播 (Graph-propagation based correlation learning, GCL) 网络[12] 设计了一个图传播子网络寻找显著性的局部区域, 并且使用图卷积网络 (Graph convolutional network, GCN) 学习区域特征向量之间的内部语义相关性. 文献 [13] 使用图神经网络对子类的显著性区域间的语义关系进行建模从而学习重要的属性. 但是, 以上两种方法都受限于只使用了单个图像少量的局部区域建图, 对于细粒度图像分类的语义关系建模, 可能导致信息在图网络传递和聚合时产生次优解. 为了有效地捕捉到目标间细微的变化, 对齐增强网络 (Alignment enhancement network, AENet)[19] 通过自注意力和 GNN 聚合最相关图像

区域的上下文感知特征以及它们在区分细粒度类别中的重要性. 在基于图的高阶关系发现 (Graph-based high-order relation discovery, GHRD) 网络[20]中, 使用了注意力机制对特征不同通道的语义信息进行建模, 并且引入 GCN 对特征进行降维, 通过分组学习的策略来学习具有判别力的特征.

2 方法

人类在识别图像时其视觉注意力会呈现一种从粗粒度层级到细粒度层级的变化, 人们首先关注目标的整体结构, 然后会注意到目标各个部位的细节特点从而判断其所属类别. 例如, 对于鸟类的识别, 首先根据目标的形状判断出是鸟类, 再通过鸟的头部、背部、尾部等识别具体是什么品种的鸟类. 因此, 应该先对原图像进行目标定位, 再在目标图像上采样有判别力的局部部件, 最后学习各个部件的细节信息.

本文提出了一种不依靠目标边界框和部位标注, 仅使用图像类别标签监督的端到端训练的模型用于细粒度图像分类任务, 网络的整体结构如图 3 所示. 该模型协同整合了三个阶段, 包括全局流的目标定位阶段、目标流的显著性局部部件采样阶段和部件流的异构局部交互图特征学习阶段. 首先通过一个注意力累计目标定位模块从原图中定位出要识别的目标对象, 通过裁剪获得目标图像, 再进入下一个阶段采样出目标的关键局部部件, 最后在通过主干网络对各个局部区域进行特征提取的基础上, 融合多张图像构建异构局部交互图模块, 再进一步对它们之间的内在语义关系进行学习. 在这里, 以 ResNet-50[21] 作为实验的骨干网络.

2.1 注意力累计目标定位

对于一张输入图像, 首先将它输入全局流主干网络中提取到特征图 $F \in \mathbf{R}^{C\times H\times W}$, 其中 H, W 和 C 分别代表特征图的高, 宽和通道数. 通过一个类似通道注意力的方式对特征图 F 进行聚合, 将 F 的通道数减少为单通道, 如以下公式所示:

$$\alpha_c = \frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W} F_c(i,\ j) \tag{1}$$

$$A = \frac{1}{C}\sum_{c}\alpha_c \otimes F_c \tag{2}$$

其中 $\otimes$ 指按位相乘, 通过以上操作得到注意力图 $A \in$

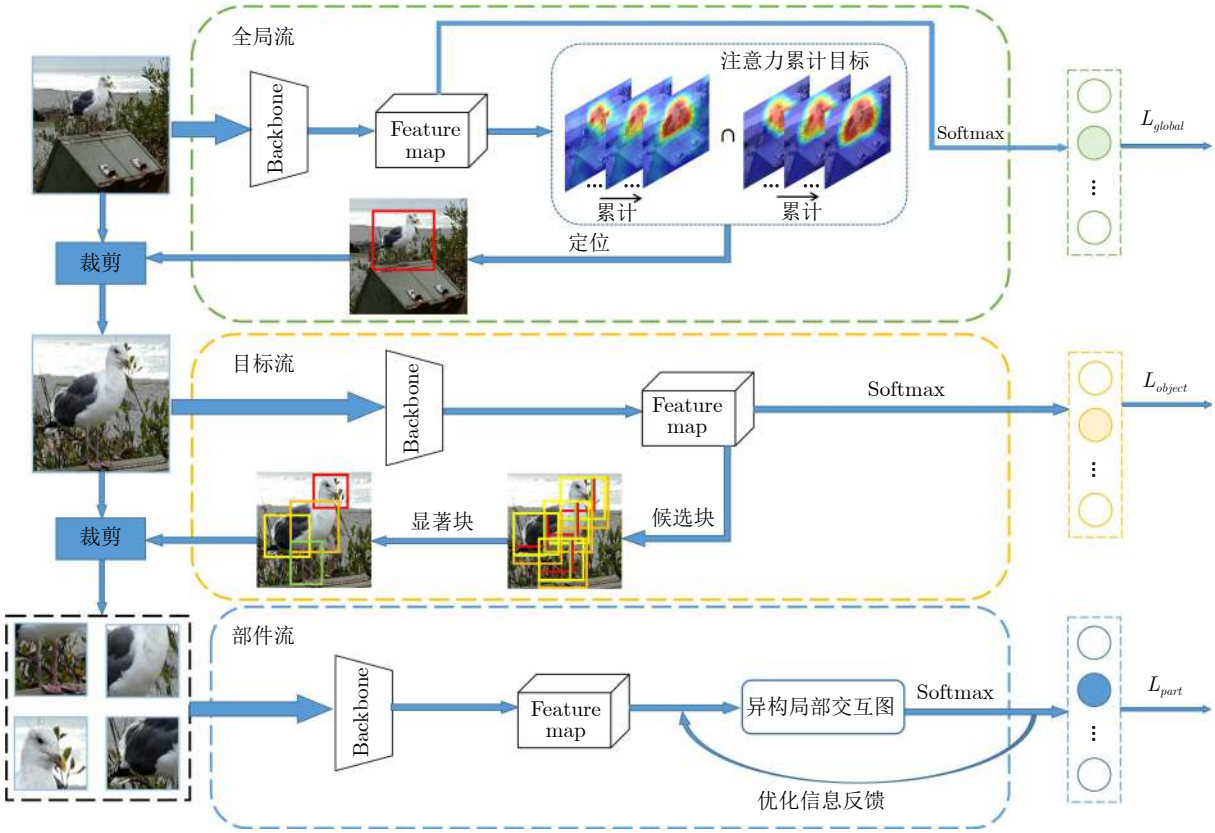


图 3 模型的基本框架图

Fig. 3 The basic framework diagram of the model

$\mathbf{R}^{1\times H\times W}$ 并执行一个归一化操作:

$$A = \frac{A}{\max(|A|)} \tag{3}$$

A 中高响应值区域表示模型的主要关注点, 对应所要识别的目标所在区域, 可以借此从注意力图 A 中定位出目标对象. 但是就像经典的 CAM[22] 方法所面临的问题, 弱监督目标定位的注意力图会随着训练轮次的增加渐渐地只关注一部分很小的高响应值区域而丧失了目标的完整性, 因此, 在这里通过建立一个注意力累计图 U 用来保存各个训练轮次关注的高响应区域, 从而能够定位到一个更加完整的目标.

首先, 使用第一个轮次生成的注意力图 A_1 对 U 进行初始化, 之后当训练每进入一个轮次, 使用当前轮次生成的注意力图 A_k 对上一轮状态的注意力累计图 U_{k-1} 进行更新:

$$U_k = \max(U_{k-1},\ A_k) \tag{4}$$

其中, $\max(\cdot,\cdot)$ 代表对两个输入项的各个元素值进行比较取最大操作. 将注意力累计图 A 的均值作为阈值来划分图中各个位置的点, 得到一个 0 和 1 的二值掩码图 $\widetilde{U}$, 通过计算出包围 $\widetilde{U}$ 中最大 1 值连通区域的最小边界框确定目标所在位置. 由于文献 [23] 和文献 [24] 的方法从集成多层网络的结果中受益, 受此启发, 在本文中, 对 ResNet-50 的最后一个卷积块 $Conv_{5c}$ 和其上一个卷积块 $Conv_{5b}$ 输出的特征分别计算其二值掩码图取交集确定最终的目标定位结果.

2.2　显著性局部部件采样

为了从目标图像中获得显著性的局部部件, 本文采取类似区域生成网络 (Region proposal network, RPN)[25] 的做法, 使用预先定义的多个不同尺度和比例的 anchor 进行采样. 不同的是, 在目标流中网络使用和全局流相同的方法生成图像特征的注意力图来计算各个 anchor 分块的得分, 而非直接提取到的多尺度特征图:

$$\bar{s}_p = \frac{1}{H_p \times W_p} \sum_{x=0}^{W_p-1} \sum_{y=0}^{H_p-1} A_p(x,\ y) \tag{5}$$

$\bar{s}_p$ 分块的得分反映了对应局部部件的显著程度, 其中, H_p 和 W_p 分别是 anchor 在注意力图上生成的分块的高和宽, A_p 为分块在相应区域的局部注意力图. 按照分值 $\bar{s}_p$ 的大小选择出更加显著的局部部件, 而为了尽可能采样到目标不同的部件, 避免块之间的重叠带来的信息冗余, 同时使用了非极大值抑制 (Non-maximum suppresion, NMS) 筛选出固定数量的块作为最终采样到的局部部件.

2.3　异构局部交互图

细粒度图像分类的困难在于不同类别目标之间外观相似, 区别只在于局部的细微差异, 而大多研究只关注于单张图像内的显著性局部部件和它们局部内的关系, 没有考虑到和其他图像间的局部关系, 无法学习到一个有区分度的特征表示. 在这里提出了一个异构局部交互图模块对不同类型的局部区域的内在语义关系进行学习来增强特征判别力, 其结构如图 4 所示.

给定一个批次的不同图像集合 $X = \{x_i\}$, 其中 $i = 1, \cdots, N$ 为批次中的图像数量. 在采样到显著性局部部件后, 将它们输入到部件流主干网络中进行特征提取, 执行全局平均池化操作获取到长度为 d_f 的局部区域特征向量集合 $F = \{f_p^i \in \mathbf{R}^{d_f}\}$, 其中 $p = 1, \cdots, M$, f_p^i 表示第 i 张图像的第 p 个局部特征. 将局部特征集合 F 作为节点集构建一个完整的异构局部交互图 $G = (F,\ E)$, E 表示图的边集, 其中存在着两种不同类型的边, 一种为节点所对应的局部区域来自于同一类别图像的正对边, 另一种则是来自不同类别图像的负对边. 异构局部交互图中的边值决定了节点之间的关系, 通过一个可学习的线性前馈层 g 来计算图中 E 的边值:

$$E_{pq}^{ij} = \text{LeakyReLU}\left(g(W_c f_p^i \oplus W_c f_q^j)\right) \tag{6}$$

其中 $W_c \in d_h \times d_f$ 是一个参数矩阵, 先将输入的节点通过线性变换为长度为 d_h 的特征向量, 再进行后续的计算, $\oplus$ 表示在通道维度上连接, E_{pq}^{ij} 反映了第 i 张图像的第 p 个局部特征 f_p^i 和第 j 张图像的第 q 个局部特征 f_q^j 之间的关联强度. 将图结构定义为一个全连接图, 这意味着任意两个节点之间都会计算它们的关联值, 由于我们所构建的图包含了一个 mini-batch 中不同图像的局部区域表示, 因此, 能够捕获到一个更完整的局部关系. 为了让不同节点之间的关联值能够更好地进行比较, 对 E 按行进行 softmax 归一化:

$$\delta_{pq}^{ij} = \frac{\exp(E_{pq}^{ij})}{\sum\limits_{n=1}^{N}\sum\limits_{m=1}^{M}\exp(E_{pm}^{in})} \tag{7}$$

得到一个 $NM \times NM$ 的注意力权重矩阵, δ_{pq}^{ij} 反映了一个批次中第 j 张图像的第 q 个局部区域对第 i 张图像的第 p 个局部区域的相对重要性. 为了整合局部内和局部间的关系, 根据该权重对节点表示进行线性组合来更新原特征:

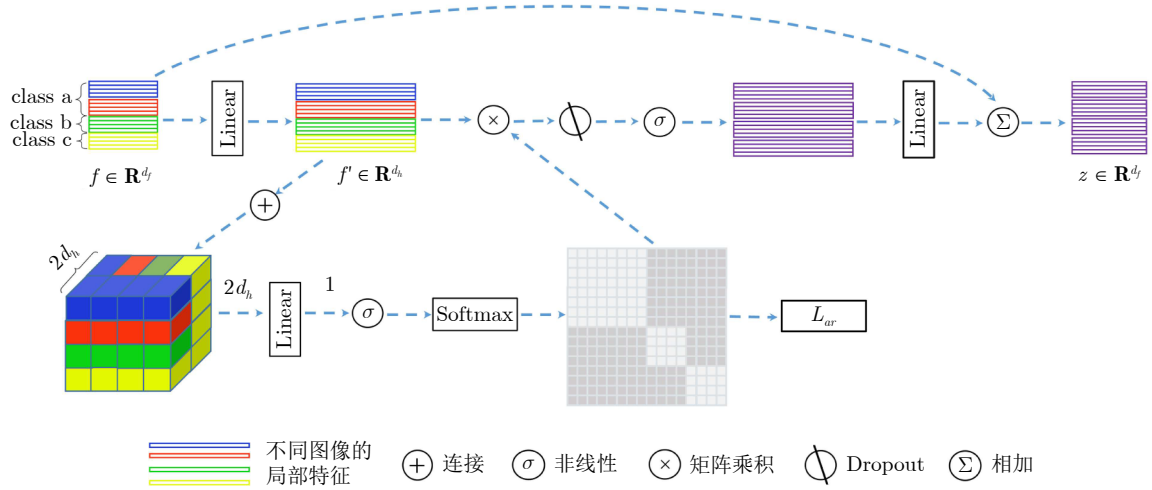


图 4 异构局部交互图模块结构

Fig. 4 The structure of HLIGM

$$z_p^i = f_p^i + W_r\sigma\left(\sum_{f_q^j \in F}\delta_{pq}^{ij}W_c f_q^j\right) \tag{8}$$

其中 $W_r \in d_f \times d_h$ 是一个参数矩阵, σ 表示非线性激活函数 (Exponential linear unit, ELU), 在这里同时加入了残差学习来更新节点特征.

由于图中的局部节点间具有相同或者不同的类别, 如果两个节点对应的局部区域属于相同类别, 比起属于不同类别的节点应该具有更强的关联关系. 相应地, 在图中应体现为具有更大的注意力权重值. 因此, 为了准确描述这些关系, 借鉴对比学习的思想, 设计了一个注意力正则化损失对图中相同类别节点的正对边和不同类别节点的负对边的注意力权重值进行不同的约束计算, 形成对比.

$$L_{ar} = \sum_{i=1}^{N}\sum_{p=1}^{M}\sum_{j=1}^{N}\sum_{q=1}^{M}L_{bce}\left(\delta_{pq}^{ij},\ \tau\right) \tag{9}$$

$L_{bce}(\cdot,\cdot)$ 为二值交叉熵损失, 如果节点 f_p^i 和 f_q^j 属于相同的类别, 则 τ 等于 1, 否则 τ 等于 0. 通过这样的一个注意力正则化损失, 不同类别节点间的注意力权重被减小, 而相同类别节点间的注意力权重则被相对地放大, 从而能够引导图正确学习不同局部区域之间的关系. 同时, 正负样本对的局部关系形成对比能够有效地突出特征中对类别更加显著的信息, 达到提高特征空间的聚簇特性的目的, 网络能够学习到一个更加具有区分度的特征表示.

2.4 损失函数

本文使用三个不同的分支流, 包括全局流、目标流和部件流构建成一个统一的网络, 在多任务损失 L_{total} 作为目标函数的监督下对模型进行优化, 从而学习不同尺度层级的图像信息. L_{total} 具体如下所示:

$$L_{total} = L_{cls} + \lambda_1 L_{ar} + \lambda_2 L_{distill} \tag{10}$$

其中包含基础的分类损失 L_{cls}, 注意力正则化损失 L_{ar} 和蒸馏损失 $L_{distill}$, λ_1 和 λ_2 为平衡这些损失的超参. L_{cls} 对三个分支输出的特征都分别计算交叉熵损失:

$$\begin{aligned} L_{cls} = L_{global} + L_{object} + L_{part} = \\ -\sum_{i=1}^{N} y_i \log\left(K\left(F_g^i\right)\right) - \sum_{i=1}^{N} y_i \log\left(K\left(F_o^i\right)\right) - \\ \sum_{i=1}^{N}\sum_{p=1}^{M} y_i \log\left(K\left(z_p^i\right)\right) \end{aligned} \tag{11}$$

其中包含全局流分类损失 L_{global},目标流分类损失 L_{object} 和部件流分类损失 L_{part}, y_i 为输入图像 x_i 的真实类别标签, K 表示一个全连接层的分类器, F_g 为全局流主干网络提取的原图像特征, F_o 为目标流主干网络提取的目标图像特征. 为了让特征提取网络能够直接学习到异构局部交互图模块的优化信息从而直接产生优化的特征表示, 通过 $L_{distill}$ 在特征提取网络和模块之间进行知识蒸馏, 公式具体如下:

$$\begin{aligned} L_{distill} = \sum_{i=1}^{N}\sum_{p=1}^{M}(1-\beta)\mathrm{CrossEntropy}\left(K\left(f_p^i\right),\ y_i\right) + \\ \beta\mathrm{KL}\left(K\left(f_p^i\right),\ K\left(z_p^i\right)\right) \end{aligned} \tag{12}$$

式中的第一项为从部件流主干网络直接提取的局部特征的分类交叉熵损失, KL 表示 Kullback-Leibler 散度, 使用它来约束初始特征和更新后的节点特征之间的预测分布差异, β 则是一个平衡因子.

在测试阶段, 舍弃异构局部交互图模块, 直接利用主干网络提取的特征进行预测, 最终的预测结果 $\hat{Y}$ 为:

$$\hat{Y} = K(F_g) + K(F_o) + \sum_{p=1}^{M} K(f_p) \tag{13}$$

3　实验

3.1　实验设置

在实验中, 调整原始图像尺寸大小为 448×448 像素, 再输入到全局流网络, 将裁剪后的目标图像的尺寸调整为和原始图像相同大小, 而所有的局部部件图像则调整为 224×224 像素. 在对目标图像进行区域采样时, 设置了 3 类不同等级大小共 8 种 anchor, 分别为: $\{[4 \times 4,\ 3 \times 5],\ [6 \times 6,\ 5 \times 7],\ [8 \times 8,\ 6 \times 10,\ 7 \times 9,\ 7 \times 10]\}$. 根据经验和多次的实验, 总的采样数 M 设置为 7, 其中最小等级的 anchor 采样 2 个, 次级采样 3 个, 最大的采样 2 个. 除此之外, 在实验中发现目标函数的超参 λ_1 的设置对结果影响比较轻微, 而 λ_2 则会造成较大影响, 最终选定 λ_1 和 λ_2 分别为 0.01 和 1, 蒸馏损失的平衡因子 β 为 0.7. 采用随机梯度下降 (Stochastic gradient descent, SGD) 作为优化器, 优化器的动量设为 0.9, 学习率初始化为 0.001 并且训练每经过 60 个 epoch 便乘以 0.1, 总共训练 200 轮, 权重衰减设置为 0.000 1, 最小批次数为 6.

3.2　CUB 数据集

表 1 展示了在 CUB-200-2011 数据集[26] 上模型和所对比方法取得的实验结果. CUB-200-2011 数据集是一个鸟类数据集, 它含有 11 788 张来自 200 种不同鸟类的图像. 本文所提出的方法在该数据集上取得了 90.2% 的准确率, 超过了其他各种不同类型的方法, 体现出了良好的性能. 对比于受欢迎的同为基于部件的方法[27−28], 本文的方法实现了大幅的提升. 即使对比于启发继承网络 (Heuristic successor network, HSnet)[29] 和 Mask-CNN (Mask convolutional neural network)[30] 使用了额外的目标或者部件的边界框标注信息, 本文仅仅依靠类别标签信息仍取得了最优的准确率. 元学习细粒度网络 (Meta-learning fine-grained network, Meta-FGNet)[31] 使用元学习的方法进行训练, 除了目标任务网络外, 还使用额外的辅助数据训练一个网络, 起到目标网络训练的一个正则化的效果, 而在本文中仅使用有限的训练数据. GCL 和 GHRD 同样引入 GNN 对局部区域之间的关系进行建模, 但是由于其有限的图节点个数, 无法捕捉到图像完整的局部关系, 并不能取得理想的效果. 在这里, 本文也和基于端到端特征编码的方法——深度双线性变换 (Deep bilinear transformation, DBT)[32] 和渐进多粒度 (Progressive multi-granularity, PMG)[33] 进行了对比, 取得了更好的表现. AENet 设计了双层级对齐框架, 在 ResNet-50 作为骨干网络的情况下参数量达到了 95 M, 分类精度则达到了 87.6%; MGE-CNN 使用了专家网络混合模型学习先验信息, 同样在选用 ResNet-50 作为骨干网络的情况下参数量达到了 104 M, 取得了 88.5% 的准确率. 虽然本文参考了 RA-CNN 和 GCL, 为了对各个尺度层级的图像信息能够进行更好的适应, 在各个分支流之间不进行参数的共享并且训练需要一定的运算量, 但是参数量也只有 75 M 并且比对比的其他方法取得了更高的准确率, 在推理速度方面也有着 20 FPS (Frames per second).

表 1　在 CUB-200-2011 数据集上的对比实验结果, Anno./DATA 表示是否使用了额外的标注信息或者辅助数据

Table 1　The comparative experimental results on CUB-200-2011 dataset, and Anno./DATA indicates whether additional labeling information or auxiliary data is used

方法	主干网络	Anno./DATA	Accuracy (%)
RA-CNN[6]	VGG-19	—	85.3
HSnet[29]	Inception	Anno.	87.5
PART[27]	ResNet-50	—	89.6
Mask-CNN[30]	VGG-16	Anno.	87.3
S3N[28]	ResNet-50	—	88.5
NTSN[46]	ResNet-50	—	87.5
ACNet[47]	ResNet-50	—	88.1
GDSMP-Net[48]	ResNet-101	—	88.1
MetaFGNet[31]	ResNet-50	DATA	87.6
DCL[37]	ResNet-50	—	88.6
DBT[32]	ResNet-101	—	88.1
GCL[12]	ResNet-50	—	88.3
AENet[19]	ResNet-101	—	88.6
MGE-CNN[17]	ResNet-101	—	89.4
GHRD[20]	ResNet-50	—	89.6
PMG[33]	ResNet-50	—	89.9
Ours	textNet-50	—	**90.2**
Ours	ResNet-101	—	**90.5**

3.3　NA Birds 数据集

NA Birds 数据集[34] 和 CUB-200-2011 数据集一样都是关于鸟类的数据集, 但是 NA Birds 相较 CUB-200-2011 是一个图片数量更庞大、种类更丰

富的数据集, 包含了 555 种不同种类, 48 562 张的北美鸟类图像. 许多需要复杂操作的方法在这个数量维度的数据集上并不容易进行实验, 使得在该数据集上的细粒度图像分类任务更具挑战性. 表 2 中罗列了几个不同的方法和本文所提出的方法在该数据集上所取得的分类精度, 可以看到, 本文模型达到了 89.5% 的分类准确率, 在表 2 所列的所有方法中取得了最优的表现, 体现了模型在跨数据集尺度和类别数上的鲁棒性, 对比于 MGE-CNN 使用更深的网络结构来提取特征, 本方法使用更小的网络但取得了更高的分类准确率. FixSENet-154 (Fix resolution discrepancy with SENet-154)[35] 使用不同的分辨率策略取得不错的表现, 而本方法则更进一步使用图像中不同尺度的信息取得了更好的效果.

表 2 在 NA Birds 数据集上的对比实验结果
Table 2 The comparative experimental results on NA Birds dataset

方法	主干网络	Anno./DATA	Accuracy (%)
DSTL[49]	Inception-v3	—	87.9
MaxEnt[50]	DenseNet-161	—	83.0
PMG[33]	ResNet-50	—	87.9
MGE-CNN[17]	ResNet-101	—	88.6
CS-Part[51]	ResNet-50	—	88.5
API-NET[52]	ResNet-101	—	88.1
FixSENet-154[35]	SENet-154	—	89.2
GHRD[20]	ResNet-50	—	88.0
Ours	ResNet-50	—	**89.5**
Ours	ResNet-101	—	**89.9**

3.4 StanfordCars 数据集

本文还在 StanfordCars 数据集[36] 上进行了实验, StanfordCars 数据集中包含了 16 185 张来自 196 种不同类型汽车的图像, 该数据集中的目标不存在 CUB-200-2011 数据集中的目标姿态变换, 具有更规则的结构信息. 实验结果如表 3 所示, 本文所提出的方法取得了 95.1% 的准确率, 实现了最优的效果. 比起使用图神经网络的方法 GCL 容易受到背景的影响, 本文的方法取得了更好的表现. 破坏重建学习 (Destruction and construction learning, DCL)[37] 使用区域混淆的机制训练一个端到端特征编码的网络, 但是不可避免地会引入背景噪声扰动, 本文的方法考虑到了这一点, 实现了更高的精度. TransFG (Transformer architecture for fine-grained recognition)[38] 基于最近流行的 transfomer 使用自注意力机制选择图像的显著性区域并计算它们之间的关系, 而本文的方法仅使用 ResNet-50 作为主干网络便取得比之高 0.3% 的准确率.

表 3 在 StanfordCars 数据集上的对比实验结果
Table 3 The comparative experimental results on StanfordCars dataset

方法	主干网络	Anno./DATA	Accuracy (%)
RA-CNN[6]	VGG-19	—	92.5
PSA-CNN[53]	VGG-19	Anno.	92.6
HSnet[29]	Inception	Anno.	93.9
ACNet[47]	ResNet-50	—	94.6
S3N[28]	ResNet-50	—	94.7
NTSN[46]	ResNet-50	—	93.9
DCL[37]	ResNet-50	—	94.5
GCL[12]	ResNet-50	—	94.0
AENet[19]	ResNet-101	—	93.7
MGE-CNN[17]	ResNet-101	—	93.9
API-NET[52]	ResNet-101	—	94.9
SDNs[54]	ResNet-101	—	94.6
M2B[55]	ResNet-50	—	94.7
TransFG[36]	ViT-B 16	—	94.8
Ours	ResNet-50	—	**95.1**
Ours	ResNet-101	—	**95.5**

3.5 FGVC-Aircraft 数据集

FGVC-Aircraft 数据集[39] 包含有 10 000 张来自 100 种不同类型航空飞机的图像, 和 StanfordCars 数据集一样, FGVC-Aircraft 数据集中的目标也具有固定的形状. 在该数据集上的分类精度如表 4 所示, 所提出的方法取得了 94.6% 的准确率, 在 FGVC-Aircraft 数据集上比其他方法同样取得了更好的表现, 即使相较于利用额外标注信息的多粒度卷积神经网络 (Multiple granularity convolutional neural network, MG-CNN)[40], 本文的方法依然有着极大的提升. 比同样采用图神经网络的 GCL 和 GHRD, 所提出的方法有着明显的提升.

3.6 消融实验

本文在 CUB-200-2011 数据集上进行了消融实验来验证所提出模块的有效性, 包含 AAOLM 和 HLIGM, 具体结果如表 5 所示. 在本文中使用 ResNet-50 进行特征提取, 并且将它的分类精度作为基准线 (Base line, BL), 其分类准确率为 84.5%. 在采样显著性部件 (Discriminate part, DP) 后, 准确率比基准提升了 0.5%, 达到了 85%. 当引入注意力累计目标定位模块后再采样显著性部件, 网络的分类精度提升到 89.3%. 因此, 可以发现, 无关的背景噪

表 4　在 FGVC-Aircraft 数据集上的对比实验结果
Table 4　The comparative experimental results on FGVC-Aircraft dataset

方法	主干网络	Anno./DATA	Accuracy (%)
DTRG[56]	ResNet-50	—	94.1
MG-CNN[40]	VGG-19	Anno.	83.0
ACNet[47]	ResNet-50	—	92.4
S3N[28]	ResNet-50	—	92.8
NTSN[46]	ResNet-50	—	91.4
DCL[37]	ResNet-50	—	93.0
DBT[32]	ResNet-101	—	91.6
GCL[12]	ResNet-50	—	93.2
AENet[19]	ResNet-101	—	93.8
API-NET[52]	ResNet-101	—	93.4
GHRD[20]	ResNet-50	—	94.3
M2B[55]	ResNet-50	—	93.3
PMG[33]	ResNet-50	—	94.1
Ours	ResNet-50	—	**94.6**
Ours	ResNet-101	—	**94.8**

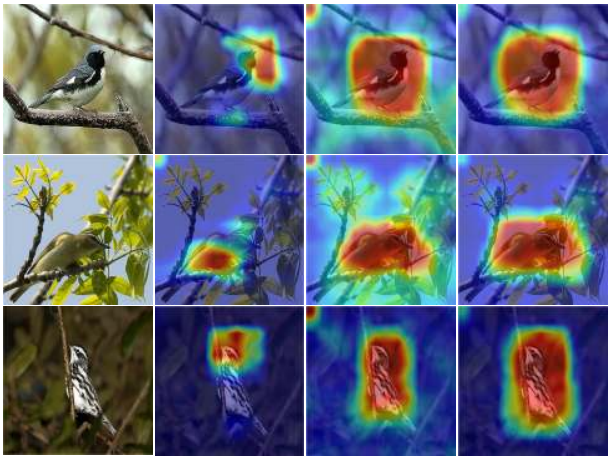
表 5　在 CUB-200-2011 数据集上的消融实验结果
Table 5　Ablation experimental results on CUB-200-2011 dataset

方法	Accuracy (%)
BL	84.5
BL+DP	85.0
BL+DP+ HLIGM	88.4
BL+DP+AAOLM	89.3
BL+DP+AAOLM+HLIGM	90.2

声会损害网络的整体性能, 模型无法精确地采样到显著性的局部区域, 而注意力累计目标定位模块能够提高模型的表现和增强模型的鲁棒性. 当引入异构局部交互图模块显式地学习区域之间的语义关系时, 准确率达到了 88.4%, 比单纯的特征提取网络和显著性区域采样组合 (BL+DP) 提升了 3.4%, 说明该模块能够很好地增强特征的表示能力. 为了更进一步对模块的有效性进行探究, 还实验了去除注意力正则化损失, 发现模型的分类准确率降到了 89.7%, 下降了 0.5%, 说明了通过类别对比的方式引导不同图像间局部关系学习的重要性. 同时, 还实验了只在单张图像上构建同构图的方法, 发现模型最终的分类准确率只达到 89.8%.

3.7　可视化及泛化性分析

为了直观地感受 AAOLM 的定位效果, 在图 5 中可视化了经典的 CAM 生成的热力图和通过 AAOLM



(a)　(b)　(c)　(d)

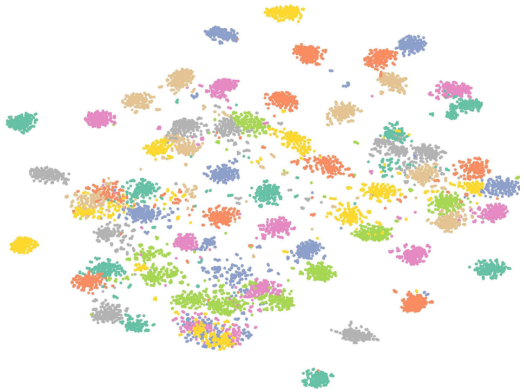
图 5　使用 CAM 和本文 AAOLM 的峰值响应图的可视化结果 ((a) 原始图像; (b) CAM 生成的热力图; (c) AAOLM 在 ResNet-50 的 $Conv_{5b}$ 卷积块输出特征上的注意力图; (d) AAOLM 在 ResNet-50 的 $Conv_{5c}$ 卷积块输出特征上的注意力图)

Fig. 5　Visualization results of peak response maps using CAM and AAOLM in this paper ((a) Original image; (b) Heat map generated by CAM; (c) Attention map of $Conv_{5b}$ convolution block of ResNet-50 by AAOLM; (d) The attention map of the $Conv_{5c}$ convolution block of ResNet-50 by AAOLM)

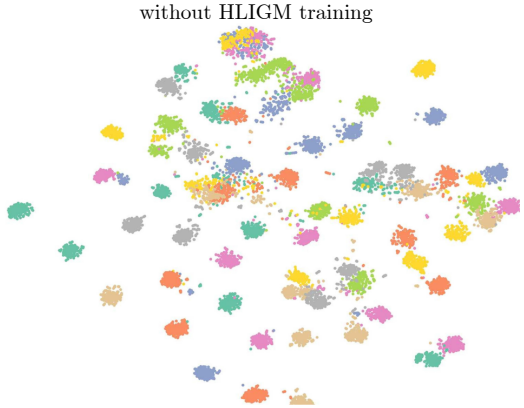
生成的注意力图, 从中可以看到所提出的 AAOLM 能够定位到一个更加完整的目标.

为了更深入探究异构局部交互图模块对模型判别性的影响, 本文使用 t 分布–随机近邻嵌入 (t-distributed stochastic neighbor embedding, t-SNE)[41] 对部件流主干网络提取到的局部部件图像的特征进行可视化来更直观地判断类间的分散度和类内的聚合度, 实验选用了来自 CUB-200-2011 测试数据集的图像. 在图 6(a) 展示了移除异构局部交互图模块训练模型后, 从部件流主干网络中直接提取到的特征的可视化效果; 而图 6(b) 则是部件流主干网络学习异构局部交互图模块反馈的优化信息后提取到的特征的可视化效果. 可以清楚地看到, 当引入异构局部交互图模块后, 特征空间中代表各个类别的聚簇大体上变得更加紧凑并且相互之间会更加疏远, 不同聚簇间形成了更加清晰的决策边界, 这说明了网络提取到的特征会突出对类别更显著的信息, 区分度得到了明显的提高.

为了对模型的泛化性进行分析, 还选择在 SUN-397[42] 数据集上进行了实验, 该数据集是一个复杂的户外场景识别数据集, 不同于前述的细粒度数据集, 该数据集图像大多没有明确要识别的目标对象, 其实验结果如表 6 所示, 本文所提出的方法在该数



(a) 不带异构局部交互图模块训练的部件流主干网络输出特征的可视化结果
(a) Visualization results of the output features of the part stream backbone network without HLIGM training



(b) 学习异构局部交互图模块反馈的信息后部件流网络输出特征的可视化结果
(b) Visualization results of the output features of the part stream network after learning feedback information from the HLIGM

图 6 通过 t-SNE 可视化部件流主干网络输出特征的聚类分布, 在 CUB-200-2011 测试数据集上对比异构局部交互图模块对判别性的影响

Fig. 6 The clustering distribution of the output features of the part stream backbone network is visualized through t-SNE, comparing the impact of HLIGM on discriminative performance on the CUB-200-2011 test dataset

表 6 泛化性实验分析对比结果 (%)
Table 6 Comparison results of generalization experiment analysis (%)

方法	SUN397	FGVC-Aircraft	StanfordCars
SimCLR[43]	63.9	—	—
BYOL[44]	63.7	—	—
WSL[45]	**67.9**	53.9	72.3
Ours	66.0	**94.6**	**95.1**

据集上仍取得了 66% 的分类精度. 与之对比, 一些经典的算法例如 SimCLR (Simple framework for contrastive learning of visual representations)[43] 采取对比学习的方式进行表征学习取得了 63.9% 的准确率. BYOL (Bootstrap your own latent)[44] 依赖于两个神经网络, 即在线网络和目标网络, 它们相互作用和相互学习, 取得了 63.7% 的准确率. WSL (Exploring the limits of weakly supervised pretraining)[45] 采用了弱监督学习的方式进行图像特征的端到端编码, 取得了 67.9% 的效果, 虽然略高于本文的方法, 但是其使用了 ResNeXt-101 作为骨干网络而本文则用了更小的 ResNet-50. 另外本文所提方法主要针对有明确的目标对象且不同类别目标外观较为相似的数据集图像, 往往有较为复杂的背景, 会对识别造成影响, 对此, 该方法采取定位关键目标来排除无关背景的影响, 再采样图像中有判别力的局部区域进行特征的增强表示学习. 并且 WSL 在 FGVC-Aircraft 和 StanfordCars 这两个数据集上的表现不如本文所提方法, 分别只取得了 53.9% 和 72.3% 的准确率, 而本方法则分别达到了 94.6% 和 95.1% 的准确率. 因此, 本文所提方法仍具有一定的泛化能力.

4 结论

本文构建了一种有效的多流弱监督学习网络, 在不需要额外的边界框或者部件标注情况下用于图像细粒度分类任务. 为了结合图像的全局和局部信息, 采取一种从粗粒度到细粒度的结构, 通过注意力累计目标定位模块有效地从原图像中定位目标, 再对目标图像进行显著性区域采样获取到部件图像. 本文利用图像局部部件之间存在的语义关系, 设计了一个多张图像输入的异构局部交互图模块, 基于一种对比学习的思想, 对局部正样本对和负样本对之间的关系进行相应的约束从而让特征学习到对类别更加显著的信息, 解决细粒度图像的目标外观相似的问题.

References

1 Luo Jian-Hao, Wu Jian-Xin. A survey on fine-grained image categorization using deep convolutional features. *Acta Automatica Sinica*, 2017, **43**(8): 1306−1318
(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)

2 Chen Jun-Ying, Chen Ying. Saliency enhanced hierarchical bilinear pooling for fine-grained classification. *Journal of Computer-Aided Design & Computer Graphics*, 2021, **33**(2): 241−249
(陈珺莹, 陈莹. 基于显著增强分层双线性池化网络的细粒度图像分类. 计算机辅助设计与图形学学报, 2021, **33**(2): 241−249)

3 Liu D C, Zhao L J, Wang Y, Kato J. Learn from each other to classify better: Cross-layer mutual attention learning for fine-grained visual classification. *Pattern Recognition*, 2023, **140**: Article No. 109550

4 Song Y, Sebe N, Wang W. On the eigenvalues of global covariance pooling for fine-grained visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(3): 3554−3566

5 Chou P Y, Kao Y Y, Lin C H. Fine-grained visual classification with high-temperature refinement and background suppression. arXiv preprint arXiv: 2303.06442, 2023.

6 Fu J L, Zheng H L, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 4476−4484

7 Nie X, Chai B S, Wang L Y, Liao Q Y, Xu M. Learning enhanced features and inferring twice for fine-grained image classification. *Multimedia Tools and Applications*, 2023, **82**(10): 14799−14813

8 Zheng S J, Wang G C, Yuan Y J, Huang S Q. Fine-grained image classification based on TinyVit object location and graph convolution network. *Journal of Visual Communication and Image Representation*, 2024, **100**: Article No. 104120

9 Hu X B, Zhu S N, Peng T L. Hierarchical attention vision transformer for fine-grained visual classification. *Journal of Visual Communication and Image Representation*, 2023, **91**: Article No. 103755

10 Zheng H L, Fu J L, Mei T, Luo J B. Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 5219−5227

11 He X T, Peng Y X, Zhao J J. Fine-grained discriminative localization via saliency-guided faster R-CNN. In: Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: ACM, 2017. 627−635

12 Wang Z H, Wang S J, Li H J, Dou Z, Li J J. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 12289−12296

13 Wang S J, Wang Z H, Li H J, Ouyang W L. Category-specific semantic coherency learning for fine-grained image recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM, 2020. 174−183

14 Li K P, Wu Z Y, Peng K C, Ernst J, Fu Y. Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 9215−9223

15 Jiang P T, Han L H, Hou Q B, Cheng M M, Wei Y C. Online attention accumulation for weakly supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(10): 7062−7077

16 Liu Y, Zhou L, Zhang P C, Bai X, Gu L, Yu X H, et al. Where to focus: Investigating hierarchical attention relationship for fine-grained visual classification. In: Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 57−73

17 Zhang L B, Huang S L, Liu W, Tao D C. Learning a mixture of granularity-specific experts for fine-grained categorization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 8330−8339

18 Chen W J, Ran S, Wang T, Cao L H. Learning how to zoom in: Weakly supervised ROI-based-DAM for fine-grained visual classification. In: Proceedings of the 30th International Conference on Artificial Neural Networks. Bratislava, Slovakia: Springer, 2021. 118−130

19 Hu Y T, Liu X H, Zhang B C, Han J G, Cao X B. Alignment enhancement network for fine-grained visual categorization. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2021, **17**(1s): Article No. 12

20 Zhao Y F, Yan K, Huang F Y, Li J. Graph-based high-order relation discovery for fine-grained recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 15074−15083

21 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 770−778

22 Zhou B L, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 2921−2929

23 Wei X S, Luo J H, Wu J X, Zhou Z H. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 2017, **26**(6): 2868−2881

24 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(4): 640−651

25 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137−1149

26 Wah C, Branson S, Welinder P, Perona P, Belongie S. The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology, Pasadena, CA, USA, 2011.

27 Zhao Y F, Li J, Chen X W, Tian Y H. Part-guided relational transformers for fine-grained visual recognition. *IEEE Transactions on Image Processing*, 2021, **30**: 9470−9481

28 Ding Y, Zhou Y Z, Zhu Y, Ye Q X, Jiao J B. Selective sparse sampling for fine-grained image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019. 6598−6607

29 Lam M, Mahasseni B, Todorovic S. Fine-grained recognition as HSnet search for informative image parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 6497−6506

30 Wei X S, Xie C W, Wu J X, Shen C H. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, 2018, **76**: 704−714

31 Zhang Y B, Tang H, Jia K. Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 241−256

32 Zheng H L, Fu J L, Zha Z J, Luo J B. Learning deep bilinear transformation for fine-grained image representation. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM, 2019. Article No. 385

33 Du R Y, Xie J Y, Ma Z Y, Chang D L, Song Y Z, Guo J. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(12): 9521−9535

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

35 Touvron H, Vedaldi A, Douze M, Jegou H. Fixing the train-test resolution discrepancy. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM, 2019. Article No. 741

36 Krause J, Stark M, Deng J, Li F F. 3D object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE, 2013. 554−561

37 Chen Y, Bai Y L, Zhang W, Mei T. Destruction and construction learning for fine-grained image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 5152−5161

38 He J, Chen J N, Liu S, Kortylewski A, Yang C, Bai Y T, et al. TransFG: A transformer architecture for fine-grained recognition. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2022. 852−860

39 Maji S, Rahtu E, Kannala J, Blaschko M, Vedaldi A. Fine-grained visual classification of aircraft. arXiv preprint arXiv: 1306.5151, 2013.

40 Wang D Q, Shen Z Q, Shao J, Zhang W, Xue X Y, Zhang Z. Multiple granularity descriptors for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 2399−2406

41 Van Der Maaten L. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research*, 2014, **15**(1): 3221−3245

42 Xiao J X, Hays J, Ehinger K A, Oliva A, Torralba A. Sun database: Large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 3485−3492

43 Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: ACM, 2020. 1597−1607

44 Grill J B, Strub F, Altché F, Tallec C, Richemond P H, Buchatskaya E, et al. Bootstrap your own latent a new approach to self-supervised learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: ACM, 2020. Article No. 1786

45 Mahajan D, Girshick R, Ramanathan V, He K M, Paluri M, Li Y X, et al. Exploring the limits of weakly supervised pretraining. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 185−201

46 Yang Z, Luo T G, Wang D, Hu Z Q, Gao J, Wang L W. Learning to navigate for fine-grained classification. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 420−435

47 Ji R Y, Wen L Y, Zhang L B, Du D W, Wu Y J, Zhao C, et al. Attention convolutional binary neural tree for fine-grained visual categorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 10465−10474

48 Ke X, Cai Y H, Chen B T, Liu H, Guo W Z. Granularity-aware distillation and structure modeling region proposal network for fine-grained image classification. *Pattern Recognition*, 2023, **137**: Article No. 109305

49 Cui Y, Song Y, Sun C, Howard A, Belongie S. Large scale fine-grained categorization and domain-specific transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4109−4118

50 Dubey A, Gupta O, Raskar R, Naik N. Maximum entropy fine-grained classification. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: ACM, 2018. 635−645

51 Korsch D, Bodesheim P, Denzler J. Classification-specific parts for improving fine-grained visual categorization. In: Proceedings of the 41st DAGM German Conference on Pattern Recognition. Dortmund, Germany: Springer, 2019. 62−75

52 Zhuang P Q, Wang Y L, Qiao Y. Learning attentive pairwise interaction for fine-grained classification. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 13130−13137

53 Krause J, Jin H L, Yang J C, Li F F. Fine-grained recognition without part annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 5546−5555

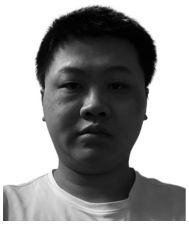
54 Zhang L B, Huang S L, Liu W. Learning sequentially diversified representations for fine-grained categorization. *Pattern Recognition*, 2022, **121**: Article No. 108219

55 Liang Y Z, Zhu L C, Wang X H, Yang Y. Penalizing the hard example but not too much: A strong baseline for fine-grained visual classification. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(5): 7048−7059

56 Liu K J, Chen K, Jia K. Convolutional fine-grained classification with self-supervised target relation regularization. *IEEE Transactions on Image Processing*, 2022, **31**: 5570−5584

陈 权 福州大学计算机与大数据学院硕士研究生. 主要研究方向为计算机视觉.
E-mail: justchenquan@gmail.com
(**CHEN Quan** Master student at the College of Computer and Data Science, Fuzhou University. His main research interest is computer vision.)

陈 飞 福州大学计算机与大数据学院副教授. 主要研究方向为计算机视觉, 机器学习和图信号处理. 本文通信作者.
E-mail: chenfei314@fzu.edu.cn
(**CHEN Fei** Associate professor at the College of Computer and Data Science, Fuzhou University. His research interest covers computer vision, machine learning and graph signal processing. Corresponding author of this paper.)



王衍根 福州大学计算机与大数据学院硕士研究生. 主要研究方向为计算机视觉.
E-mail: lCRZakHCfh237@hotmail.com
(**WANG Yan-Gen** Master student at the College of Computer and Data Science, Fuzhou University. His main research interest is computer vision.)

程 航 福州大学数学与统计学院教授. 主要研究方向为机器学习和多媒体信息安全.
E-mail: hcheng@fzu.edu.cn
(**CHENG Hang** Professor at School of Mathematics and Statistics, Fuzhou University. His research interest covers machine learning and multimedia information security.)

王美清 福州大学数学与统计学院教授. 主要研究方向为图像处理和数值计算. E-mail: mqwang@fzu.edu.cn
(**WANG Mei-Qing** Professor at School of Mathematics and Statistics, Fuzhou University. Her research interest covers image processing and numerical calculation.)