计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(8):166−175 [doi: 10.15888/j.cnki.csa.009571]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

结合动态自适应调制和结构关系学习的细粒度图像分类①

王衍根, 陈　飞, 陈　权

(福州大学 计算机与大数据学院, 福州 350108)
通信作者: 陈　飞, E-mail: chenfei314@fzu.edu.cn

摘　要: 由于细粒度图像类间差异小, 类内差异大的特点, 因此细粒度图像分类任务关键在于寻找类别间细微差异. 最近, 基于 Vision Transformer 的网络大多侧重挖掘图像最显著判别区域特征. 这存在两个问题: 首先, 网络忽略从其他判别区域挖掘分类线索, 容易混淆相似类别; 其次, 忽略了图像的结构关系, 导致提取的类别特征不准确. 为解决上述问题, 本文提出动态自适应调制和结构关系学习两个模块, 通过动态自适应调制模块迫使网络寻找多个判别区域, 再利用结构关系学习模块构建判别区域间结构关系; 最后利用图卷积网络融合语义信息和结构信息得出预测分类结果. 所提出的方法在 CUB-200-2011 数据集和 NA-Birds 数据集上测试准确率分别达到 92.9% 和 93.0%, 优于现有最先进网络.

关键词: 细粒度图像分类; Vision Transformer (ViT); 动态自适应调制; 结构关系学习; 图卷积网络

引用格式: 王衍根,陈飞,陈权.结合动态自适应调制和结构关系学习的细粒度图像分类.计算机系统应用,2024,33(8):166–175. http://www.c-s-a.org.cn/1003-3254/9571.html

Fine Grained Image Classification Combining Dynamic Adaptive Modulation and Structural Relationship Learning

WANG Yan-Gen, CHEN Fei, CHEN Quan

(College of Computer and Data Science, Fuzhou University, Fuzhou 35108, China)

Abstract: Due to the small inter-class differences and large intra-class differences of fine-grained images, the key to fine-grained image classification tasks is to find subtle differences between categories. Recently, Vision Transformer-based networks mostly focus on mining the most prominent discriminative region features in images. There are two problems with this. Firstly, the network ignores mining classification clues from other discriminative regions, which can easily confuse similar categories. secondly, the structural relationships of images are ignored, resulting in inaccurate extraction of category features. To solve the above problems, this study proposes two modules: dynamic adaptive modulation and structural relationship learning. The dynamic adaptive modulation module forces the network to search for multiple discriminative regions, and then the structural relationship learning module is used to construct structural relationships between discriminative regions. Finally, the graph convolutional network is used to fuse semantic and structural information to obtain predicted classification results. The proposed method achieves testing accuracy of 92.9% and 93.0% on the CUB-200-2011 dataset and NA-Birds dataset, respectively, which is superior to existing state-of-the-art networks.

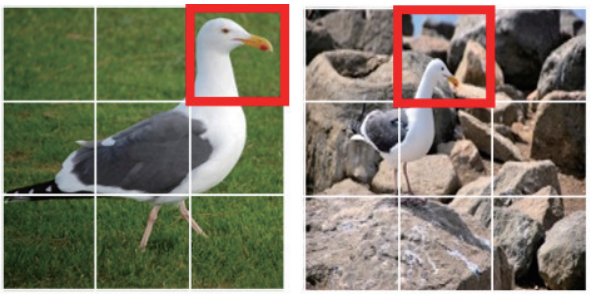
Key words: fine grained image classification; Vision Transformer (ViT); dynamic adaptive modulation; structural relationship learning; graph convolutional network (GCN)

① 基金项目: 国家自然科学基金 (61771141); 福建省自然科学基金 (2021J01620)
收稿时间: 2024-01-27; 修改时间: 2024-02-29; 采用时间: 2024-03-11; csa 在线出版时间: 2024-06-28
CNKI 网络首发时间: 2024-07-01

1 引言

细粒度分类是计算机视觉中极具挑战性的问题之一, 需要区分从属于同一大类的子类别, 例如不同种类的鸟类[1]、狗[2]、飞机[3]、车辆[4]等. 由于光线、遮挡、视角和姿态影响, 子类别物体类间差异小, 类内差异大. 因此, 现有的细粒度分类方法致力于识别不同类别视觉特征中的细微差异, 如纹理、颜色、形状等. 而这些特征通常存在于各个局部区域中, 因此, 细粒度分类的关键之处就是从图像中寻找这些具有辨别力的局部区域. 最初的细粒度方法通过将物体分割成多个组成部分[5,6], 例如将鸟类分为头、翅膀、脚等, 来比较寻找判别区域. 然而这样的方法需要大量人工标注, 有时甚至需要行业专家介入, 成本过高. 为解决这一问题, 近年来研究人员纷纷引入弱监督方法[7–9], 不再利用预先标注锚框定位, 而是利用类激活图[10,11], 通过特征图中高响应区域来定位判别区域.

近些年, Vision Transformer (ViT)[12]在计算机视觉领域多个任务均达到最先进水平, 从而细粒度分类任务上提出许多基于 ViT 网络的方法[13,14], 这些方法主要通过设置硬阈值筛选图像令牌, 通过设定一个固定阈值, 将特征值大于阈值的令牌视为重要令牌从而定位判别区域. 然而, 硬阈值筛选方法有两个问题: 一是使网络倾向专注于图像最显著区域, 忽略其他局部区域差异, 但如图 1 所示, 加粗方框内为加州海鸥和西美鸥的最显著区域, 可以看出加州海鸥和西美鸥的最显著区域非常相似, 若仅依赖此作为判别区域容易混淆这两个类别; 二是缺乏对对象整体结构关系的学习, 欠缺对判别区域间的依赖关系的挖掘, 图 1 中鸟类形态结构上大致可以划分为头、颈部、腹部、翅膀、脚、尾翼等部位, 其中头部与颈部相连, 而脚则与腹部相连, 头与尾翼距离最远等.



(a) 加州海鸥 (b) 西美鸥

图 1 加州海鸥和西美鸥对比

为解决上述问题, 本文提出了结合动态自适应调制模 (dynamic adaptive modulation, DAM) 和结构关系学习模块 (structural relationship learning, SRL) 的细粒度分类网络 (DAMSRL). 为提取准确的类别特征首先需要定位图像对象各部件区域, 从聚类角度出发, 分割图像对象各部件相当于将每个像素分派到各个类簇, 每个类簇即代表一类部件预测掩码. 因此, 本文提出动态自适应调制模块, 通过 K-means 聚类算法将每个图像块分配到不同的类簇中, 用每个类簇代表一类部件区域预测; 然后, 由于主干网络提取的令牌特征图表示了网络的注意力分布, 所以为使网络定位到更多的部件, 避免网络仅关注到最显著的判别区域, 提出在每个类簇通过高斯拟合生成软阈值调制掩码, 调制网络注意力分布, 从而迫使网络关注到多个判别区域. 之后, 结构关系学习模块先是通过一个简单的采样网络过滤对构建结构关系不重要的信息, 再通过令牌特征图的注意力权重来计算初始边权, 接着过滤次要边权, 最后通过两个线性变换层来挖掘判别区域的依赖关系, 从而构建整体的结构关系. 最后, 通过图卷积网络融合全部令牌特征的语义信息和结构信息, 使网络提取到更准确的类别特征.

本文主要贡献包括: 1) 提出动态自适应调制模块, 用聚类方法模拟图像分割, 从而可以根据对象大小自适应分割出各个部件区域, 生成部件区域预测, 再通过调制令牌注意力分布, 使网络关注到更多判别区域. 2) 提出结构关系学习模块, 挖掘判别区域间依赖关系, 从而构建对象整体结构关系, 提高了所提取的类别特征的表达能力. 3) 所提出方法在 CUB-200-2011 数据集和 NA-Birds 数据集上性能优于现有最先进方法.

2 相关工作

2.1 基于卷积神经网络的细粒度分类

早期细粒度分类工作大多基于卷积神经网络 (CNN), 主要可以分为两类: 基于部件定位方法和基于特征编码方法. 基于部件定位方法最初直接使用人工标注的部件注释, 如 Huang 等人[5]提出从标注部位提取特征以增强网络识别能力; Wei 等人[6]提出结合目标检测方法, 先定位部件区域再进行细粒度分类. 然而, 部件标注成本昂贵, 因此后来方法采用弱监督方法定位部件. Yang 等人[9]提出利用多尺度特征图来生成候选区域, 再设计过滤器来定位具有辨别力的部件; Ge 等人[8]提

出利用弱监督目标检测和实例分割方法提取粗糙的对象实例, 再从中捕获部件. 这些方法虽无需部件注释, 但需要额外添加目标检测网络, 随之为进一步简化网络, 产生了基于注意力机制的定位方法. Zheng 等人[15]提出利用三线性注意力建模通道间关系生成对应注意力图, 再利用采样器以高分辨率突出关键部位; Rao 等人[16]提出引入反事实干预评估网络注意力质量, 促使网络学习更有效的注意力. 与部件定位不同, 特征编码方法旨在学习更全面的类别特征. Chen 等人[17]通过将图像分块打乱促使网络学习更多判别性的局部细节; Song 等人[18]通过抑制图像最显著区域迫使网络挖掘其他潜在部件信息.

2.2 基于 ViT 的细粒度分类

由于 CNN 更擅长捕获图像局部特征, 欠缺提取图像全局特征能力, 而 ViT 能很好解决这一问题, 因而近年来提出了许多基于 ViT 的细粒度分类方法. He 等人[14]首次在细粒度识别任务验证了 ViT 网络有效性, 并提出结合所有层注意力筛选得到最具辨别力令牌. Zhang 等人[19]提出利用两个 ViT 网络, 通过第 1 个 ViT 网络挖掘令牌间上下文关系, 然后裁剪关键区域, 其中在选择关键区域时采用了硬阈值筛选方法, 将大于阈值的区域认定为关键判别区域, 再利用第 2 个 ViT 网络提取类别特征. Wang 等人[13]提出相互关注权重方法重新计算令牌重要性, 同样采用了硬阈值筛选的方法, 设定阈值为权重矩阵均值, 再选择大于阈值的令牌视为重要令牌, 然后聚合各层重要令牌特征来提高类别特征表示能力. Zhao 等人[20]提出使用多个 Transformer 块来提取全局特征和局部特征. Liu 等人[21]提出结合知识图谱方法建立可学习知识嵌入集, 为获取更准确的图像表示添加入知识库, 采用峰值抑制模块抑制网络对高响应令牌关注. Yu 等人[22]提出将像素特征与具体 query 目标的交互注意力建模重新定义为一个聚类过程, 由此提出 kMaX-DeepLab 网络来进行图像分割, 取得良好效果. 由于硬阈值筛选方导致网络仅关注到最显著的区域, 因此受上述论文启发, 本文提出用动态自适应调制方法用聚类方法来将图像块定位到不同部件区域, 同时用高斯函数重新调制注意力分布, 使网络关注到更多的判别区域; 其次, 上述工作忽略了部件区域的结构关系, 忽略挖掘判别区域间的依赖关系, 而结构关系对类别判断有重要意义, 因此需进一步构建部件结构关系, 从而提取更准确的类别特征.

2.3 基于图神经网络的细粒度分类

图神经网络原来主要用于处理图结构数据, 近年来通过从图像数据中构造节点来建立图神经网络, 在图像领域各项任务取得良好效果, 例如 Kosman 等人[23]通过将超像素视作图形节点, 将视频特征编码为图特征, 实现视频理解任务. 而在细粒度分类领域, 通常做法是将局部区域视为节点, 例如 Wang 等人[24]利用图神经网络学习区域间相关性提高了网络判别能力, Zhao 等人[25]提出基于图的关系发现方法来探索不同语义特征间的内在关系. 这些方法大大提高了网络细粒度识别精度, 因此, 本文提出将令牌特征视为节点特征, 将构建的结构关系作为邻接矩阵, 建立图卷积神经网络[26], 融合所有令牌特征的语义信息和结构信息, 进而提取到更准确的类别特征, 提高网络细粒度分类精度.

3 提出方法

本节介绍了所提出的结合动态自适应调制和结构关系学习模块网络架构, 如图 2 所示, 主要由主干网络、动态自适应调制模块以及结构关系学习模块 3 部分组成, 第 1 个虚线框为动态自适应调制模块, 用颜色标注调制注意力分布后网络关注到的判别区域, 红色为最显著区域; 第 2 个虚线框为结构关系学习模块, 将判别区域视为节点, 连接线表示判别区域间依赖关系, 从而构建整体结构关系. 具体来说, 先将输入图像划分为标准的网格区域, 每个网格用一个令牌表示; 再将令牌序列输入带有特征金字塔网络 (feature pyramid network, FPN[27]) 的 Swin Transformer[28]来提取令牌特征. 接着, 将提取到的最后一层的令牌特征输入动态自适应调制模块: 动态自适应调制模块主要由 K-means 聚类算法以及软阈值调制两部分组成, 通过动态自适应调制模块可以定位到对象各个部件区域, 并使模型关注到更多具有辨别力的部件区域, 更全面地学习到类别间的细微差异. 再接着, 结构关系学习模块先通过一个简单采样网络过滤令牌特征中与结构关系无关的信息, 再在判别区域间建立全连接关系, 然后过滤次要边权, 最后通过两层线性变换层来挖掘判别区域间的依赖关系, 从而构建对象整体结构关系. 最后, 将提炼后的令牌特征与原始令牌特征建立残差连接, 将新的令牌特征作为节点特征, 构建的整体结构关系作为邻接矩阵, 从而建立图卷积神经网络, 再通过两层图卷积神经网络融合所有令牌特征信息给出分类预测结果. 下面我们将对这两个模块做全面详细地描述.

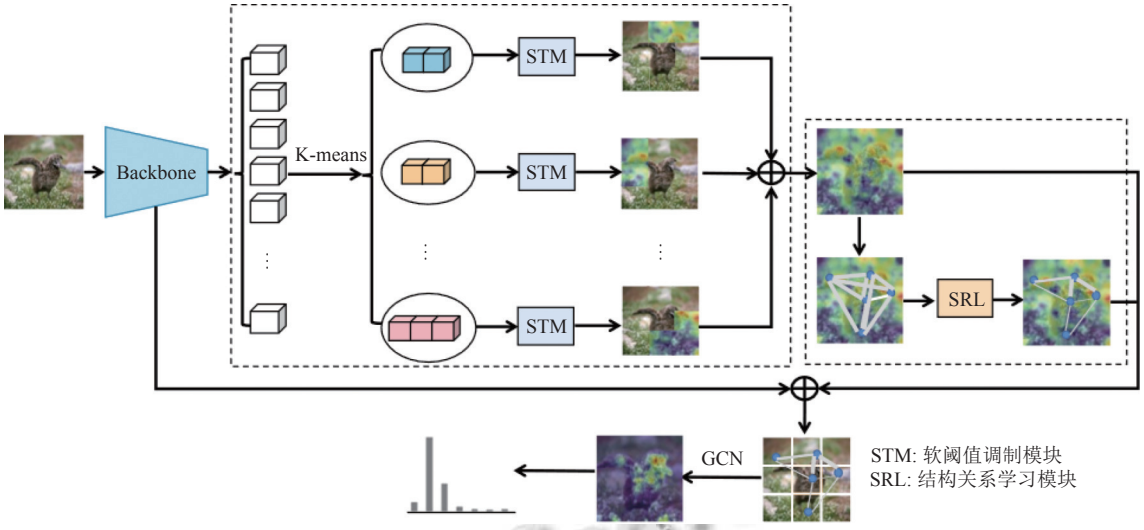


图 2 DAMSRL 结构示意图

3.1 动态自适应调制

主干网络提取的令牌特征包含了图像对象各个部件区域, 之前工作通常采用图像分割方法将各个部件区域分割出来, 或通过设定阈值, 选取大于阈值的令牌特征视为定位到的判别区域. 图像分割方法虽能准确定位对象各个部件区域但需有昂贵的人工标注帮助, 而硬阈值加权方法容易导致网络仅关注到最显著区域, 忽略了其他细微的类别间差异, 从而增加相似类别间的区分难度. 因此, 本文提出动态自适应调制方法来选择令牌特征并调制令牌特征的注意力分布, 从而既能定位图像对象的各个部件区域, 又能迫使网络关注到更多判别区域, 从而增强网络细粒度分类能力.

动态自适应调制模块主要由 K-means 聚类和软阈值调制两部分组成. 首先, 采用 K-means 聚类算法将令牌特征分派到多个类簇, 用每个类簇的分派结果代表一个部件区域的定位预测结果. 具体来说, 先将主干网络提取到的最后一层特征图$\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N]\in\boldsymbol{R}^{N\times H\times W}$, 其中$N$表示组成特征图的令牌特征数量, H表示特征图的高度, W表示特征图的宽度, 将特征图作为 K-means 聚类算法的输入, 记聚类簇数为K, K为大于零的常量, 由多次实验结果确定, 记第i个聚类中心为C_i, 为了便于计算, 将聚类中心点C_i扩充为矩阵$\boldsymbol{C}_i\in\boldsymbol{R}^{H\times W}$, 则每个类簇内令牌特征$\boldsymbol{X}_j$与聚类中心距离$\boldsymbol{C}_i$采用欧氏距离计算公式为:

$$D\left(\boldsymbol{X}_j,\boldsymbol{C}_i\right)=\left\|\boldsymbol{X}_j-\boldsymbol{C}_i\right\|_F \tag{1}$$

其中, $\|\cdot\|_F$表示 F 范数, 聚类迭代至聚类中心不变或迭代次数达到设定最大次数 10 次则停止迭代. 与图像分割方法相比, 通过聚类方法, 可以自适应对象大小, 定位对象的各个部件区域, 且无需昂贵的人工标注.

接着, 逐一对每个类簇利用高斯函数进行拟合, 生成每个令牌特征的权重调制掩码, 来重塑类簇内令牌特征注意力分布, 迫使网络关注到更多具有辨别力局部区域, 在各个类簇内寻找相似类别间细微的差异. 具体来说, 如图 3 所示, 记第i个类簇内令牌特征图大小为$\boldsymbol{R}^{N_i\times H\times W}$, 对每个类簇内令牌特征图分别进行通道注意力和空间注意力调制. 在进行通道注意力调制时, 先将令牌特征图通过最大值池化到$\boldsymbol{R}^{N_i\times 1\times 1}$, 再用池化后令牌特征图权值$t_j$来计算均值$\mu_i$和标准差$\sigma_i$, 则令牌特征图通道注意力调制的高斯权值为:

$$b_j\left(t_j\,|\,\mu_i,\sigma_i\right)=\frac{1}{\sigma_i\sqrt{2\pi}}\mathrm{e}^{-\frac{(t_j-\mu_i)^2}{2\sigma_i^2}} \tag{2}$$

再将高斯权值扩展维度生成作用于令牌特征图的通道注意力调制掩码$\boldsymbol{B}_j\in\boldsymbol{R}^{N_i\times H\times W}$; 接着对于令牌特征图的空间注意力调制, 先将令牌特征图通过全局平均池化到$\boldsymbol{R}^{1\times H\times W}$, 再采用上述通道注意力调制掩码相同的计算方法来计算得到相应的空间注意力调制掩码$\boldsymbol{B}'_j\in\boldsymbol{R}^{N_i\times H\times W}$. 最后将令牌特征图的通道注意力调制掩码和空间注意力调制掩码通过矩阵乘法得到最终的令牌特征图调制掩码. 由于令牌特征图是主干网络通过一系列的自注意力和交互注意模块来对特征进行聚合后得到的, 令牌特征值表示了网络注意力分布, 则将权值掩码作用于令牌特征上调整了令牌特征值, 也即

表示生成的权值掩码调整了网络注意力分布, 使得网络可以关注到更多具有辨别力的局部区域.

接着, 将令牌特征分为两类: 一类主要包含前景信息称之为前景类, 而另一类主要包含背景信息称之为背景类. 为过滤掉背景信息等干扰信息, 使用可学习参数α, $\alpha\in(0,1)$, 计算调制注意力后特征图$\boldsymbol{X}'$的均值$\bar{x}$, 再将大于均值的令牌特征视为前景类生成对应的软阈值调制掩码, 同时将置为零的令牌特征保留下来生成背景类掩码, 则新的软阈值调制掩码$\boldsymbol{M}$和背景类掩码$\bar{\boldsymbol{M}}$为:

$$\boldsymbol{M}(i,j)=\begin{cases}\boldsymbol{X}'(i,j), & \boldsymbol{X}'(i,j)>\alpha\bar{x}\\ 0, & \boldsymbol{X}'(i,j)\leqslant\alpha\bar{x}\end{cases} \tag{3}$$

$$\bar{\boldsymbol{M}}(i,j)=\begin{cases}0, & \boldsymbol{M}(i,j)\neq 0\\ 1, & \boldsymbol{M}(i,j)=0\end{cases} \tag{4}$$

生成的背景类掩码用于划分背景类令牌特征和前景类令牌特征来计算分类损失. 而生成的软阈值调制掩码与硬阈值筛选方法相比, 添加了可学习参数帮助网络更好地过滤干扰信息, 且保留了调制后注意力分布, 从而既避免网络仅关注最显著的局部区域, 又能避免关注到背景区域等干扰信息.

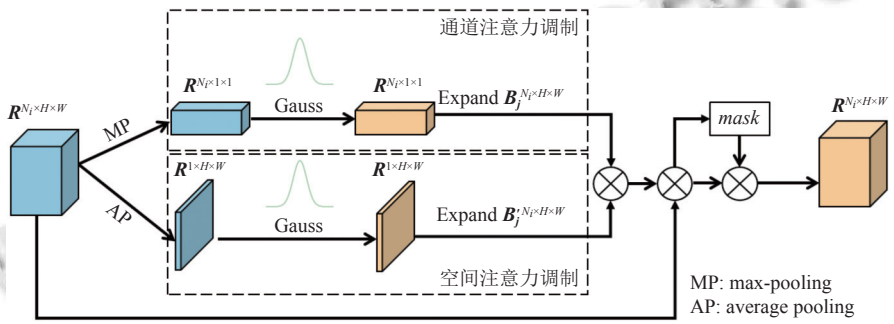


图 3 STM 模块示意图

3.2 结构关系学习

经过动态自适应调制模块后, 网络从令牌中捕获了多个判别区域, 但仍缺失对对象整体结构关系的挖掘, 忽略了判别区域间的空间关系, 而这对细粒度分类的精度非常重要, 因此本文提出结构关系学习模块来构建对象结构关系, 从而将判别区域的上下文信息整合进特征图中.

为提高构建的结构关系准确性, 本文先采用一个由卷积和池化组成的简单采样网络来过滤掉对结构关系构建无用的干扰信息如颜色纹理等. 如图 4 所示, 将特征图$\boldsymbol{X}'\in\boldsymbol{R}^{N\times H\times W}$通过一个$2\times2$的最大池化层下采样到$\boldsymbol{Y}\in\boldsymbol{R}^{N\times H/2\times W/2}$, 缩小特征图尺寸; 再通过一个个$1\times1$的卷积层减少令牌数量到$\boldsymbol{Y}\in\boldsymbol{R}^{S\times H/2\times W/2}$, 其中$S^2=N$, N为完全平方数. 然后, 再通过双线性插值方法将特征图上采样到$\boldsymbol{Y}\in\boldsymbol{R}^{S\times H\times W}$, 最后通过一个$1\times1$的卷积层并变形回令牌特征图$\boldsymbol{Y}'\in\boldsymbol{R}^{N\times H\times W}$. 通过这个采样网络, 过滤掉特征图中对结构关系构建不重要的信息, 有利于接下来构建整体结构关系和挖掘部件区域间的空间依赖关系.

接着, 对于过滤不重要信息后的特征图, 将每个令牌特征视为一个节点特征, 然后再基于特征图的注意力权重计算得到初始结构权重矩阵. 具体来说, 将采用后令牌特征图变形为$\boldsymbol{Y}'\in\boldsymbol{R}^{N\times HW}$和$(\boldsymbol{Y}')^{\mathrm{T}}\in\boldsymbol{R}^{HW\times N}$, 则初始权重矩阵计算公式为:

$$\boldsymbol{A}_{\mathrm{adj}}=\boldsymbol{Y}'\times(\boldsymbol{Y}')^{\mathrm{T}} \tag{5}$$

然后, 计算初始权重矩阵的均值, 将低于均值的边权重置为 0, 以过滤部件间不重要的关系:

$$\boldsymbol{A}_{\mathrm{adj}}(i,j)=\begin{cases}\boldsymbol{A}_{\mathrm{adj}}(i,j), & \boldsymbol{A}_{\mathrm{adj}}(i,j)>\bar{A}_{\mathrm{adj}}\\ 0, & \boldsymbol{A}_{\mathrm{adj}}(i,j)\leqslant\bar{A}_{\mathrm{adj}}\end{cases} \tag{6}$$

再接着, 为进一步挖掘部件间依赖关系, 将初始权重矩阵与令牌特征图相乘, 从而将结构信息融合进令牌特征图中. 对新的令牌特征图, 再利用线性变换层来动态学习节点边权关系, 从而得到新邻接矩阵来更准确表示图像结构关系. 通过两层独立的线性变换层得到新的结构关系可以表示为:

$$\boldsymbol{A}=\psi\left(\boldsymbol{A}_{\mathrm{adj}}\times\boldsymbol{Y}'\times\boldsymbol{W}_1\right)\psi\left(\boldsymbol{A}_{\mathrm{adj}}\times\boldsymbol{Y}'\times\boldsymbol{W}_2\right)^{\mathrm{T}} \tag{7}$$

其中, $\boldsymbol{W}_1$、$\boldsymbol{W}_2$表示线性变换参数, $\psi(\cdot)$表示激活函数, 本文采用tanh函数作为激活函数. 最后将权重矩阵$\boldsymbol{A}$进行归一化:

$$\boldsymbol{A}'=\frac{\boldsymbol{A}_{i,j}-\min(\boldsymbol{A})}{\max(\boldsymbol{A})-\min(\boldsymbol{A})} \tag{8}$$

归一化后得到构建的图像结构关系矩阵$\boldsymbol{A}'$.

接着, 通过将原始令牌特征与动态自适应调制模块调制后的令牌特征建立残差连接, 再将前述的结构关系矩阵作为邻接矩阵, 将令牌特征视为节点特征, 建立图卷积网络, 从而利用图卷积网络聚合令牌特征的语义信息和结构信息, 增强令牌特征的表达能力. 通过两层图卷积网络获得更精确的类别特征可以表示为:

$$\boldsymbol{Z}'=\psi\left(\boldsymbol{A}'\times\psi\left(\boldsymbol{A}'\times\boldsymbol{Z}\times\boldsymbol{W}_{g1}\right)\times\boldsymbol{W}_{g2}\right) \tag{9}$$

其中, $\boldsymbol{Z}$表示残差连接后的特征图, $\boldsymbol{W}_{g1}$、$\boldsymbol{W}_{g2}$表示两层图卷积网络的线性参数. 总之, 通过结构关系学习模块可以将对象的结构信息, 也即关键判别区域的空间构成信息融合进最终提取的类别特征中, 使得网络不仅可以学习到类别对象的外观特征, 还能学习对象的结构特征, 以提高网络细粒度分类的精度.

最后, 通过分类器给出最终的分类预测结果:

$$\boldsymbol{P}=Softmax(\boldsymbol{W}(\text{Avgpool}(\boldsymbol{Z}'))+\boldsymbol{C}) \tag{10}$$

其中, $\boldsymbol{W}$、$\boldsymbol{C}$为分类器参数, Avgpool为池化函数, 将每个令牌特征从$\boldsymbol{R}^{H\times W}$池化到$\boldsymbol{R}^{1\times 1}$. 通过添加本文提出的全部模块, 网络细粒度分类的预测结果准确率得到显著提高.

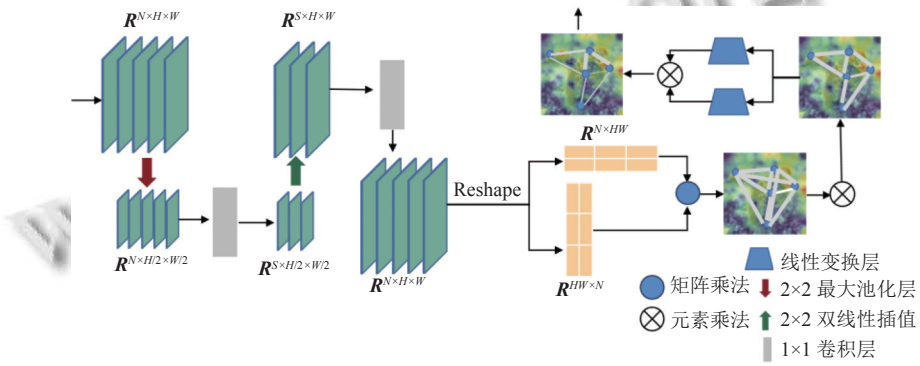


图 4 SRL 模块示意图

3.3 损失函数

网络目标是给出图像细粒度分类预测结果, 仅使用类别标签作为监督, 不使用其他人工标注. 对 FPN 网络各层特征图, 采用对令牌特征做平均池化再用*Softmax*函数得到预测概率, 利用交叉熵函数来计算各层分类预测损失, 记分类器参数为$\boldsymbol{W}_{li}$, 总层数为n, 预测类别结果为y_{li}:

$$\boldsymbol{P}_{li}=Softmax(\boldsymbol{W}_{li}(\text{Avgpool}(\boldsymbol{Z}'))) \tag{11}$$

$$L_l=-\sum_{i=1}^{n}y_{li}\log(\boldsymbol{P}_{li}) \tag{12}$$

其中, L_l表示所有层分类预测损失. 接着, 对划分得到前景类令牌特征采用计算层分类损失的同样方法计算损失L_f, 而对背景类令牌特征$\boldsymbol{X}_b'$采用通过全连接层分类器后用tanh函数激活获得分类预测结果, 再利用均方误差 (MSE) 计算损失, 记类别数量为V, 则用背景类令牌特征来预测图像类别的损失L_b计算公式为:

$$\boldsymbol{P}_b=\tanh\left(\boldsymbol{W}_b\boldsymbol{X}_b'\right) \tag{13}$$

$$L_b=\sum_{i=1}^{V}(P_{b,i}+1)^2 \tag{14}$$

其中, $\boldsymbol{W}_b$表示分类器参数, $P_{b,i}$表示第 i 类的预测概率, 最后通过图卷积网络融合所有信息, 对图分类结果也采用交叉熵计算损失L_g.

最后, 汇总所有损失:

$$L=\lambda_1L_l+\lambda_2L_f+\lambda_3L_b+\lambda_4L_g \tag{15}$$

其中, $\lambda_1,\lambda_2,\lambda_3,\lambda_4$为平衡参数, 本文中设置为$\lambda_1$=0.3, λ_2=1.0, λ_3=5.0, λ_4=1.0.

4 实验过程与结果分析

在本节中, 将本文提出的 DAMSRL 模型在经典细粒度分类数据集 CUB-200-2011 和 NA-Birds 上进行分类准确率评估, 并说明实验超参数的设置. 然后, 将本文所提出的 DAMSRL 模型测试准确率与当前的一些最先进的方法进行比较. 最后, 通过消融实验分析一些影响识别精度的因素并可视化结果.

4.1 数据集和实验细节

CUB-200-2011 数据集和 NA-Birds 数据集是两个经典的鸟类分类数据集. CUB-200-2011 数据集共有 200 个鸟类类别, 包括 5 994 张训练图像和 5 794 张测

试数据. 每个类别包含大约 30 个训练数据. NA-Birds 有 555 种鸟类、23 929 张训练图像和 24 633 张测试图像. 在两个数据集上评估时都仅使用类别标签作为监督.

本文使用 Swin Transformer 作为主干网络, 输入彩色图像大小为384×384; 数据增强方法如下: 将输入图像缩放至510×510大小再裁剪至384×384大小; 在训练阶段, 采用随机裁剪, 水平翻转和随机擦除执行数据增强; 而在测试阶段, 仅使用中心裁剪执行数据增强. 训练时, 聚类类簇设为 5 类, 最大迭代次数为 10 次, 学习率设置为 0.000 5, 学习率衰减使用余弦衰减, 衰减初始权重设置为 0.000 3; 并且使用随机梯度下降 (SGD) 作为优化器, 批大小为 16, 总共训练 100 个 epoch, 划分参数 t 初始值为 0.5. 所有实验均在单块 Nvidia GeForce RTX 3 090 显卡上完成, 并使用 PyTorch 为主要框架.

4.2 与最先进方法比较

在表 1 中, 将本文提出的 DAMSRL 方法与 CUB-200-2011 数据集上最先进的方法进行了比较. 主干网络使用的是在 ImageNet22k 上预训练的 Swin-L 模型.

表 1 在 CUB-200-2011 数据集上与其他方法的对比 (%)

方法	主干网络	Top-1精度
AT-CNN[29]	InceptionV3	87.8
PMG[30]	ResNet-50	89.6
DeepFVE[31]	InceptionV3	91.0
CAMF[32]	Swin-B	91.2
TPSKG[21]	ViT-B_16	91.3
AFTrans[19]	ViT-B_16	91.5
TransFG[14]	ViT-B_16	91.7
CAP[33]	Xception	91.8
SR-GNN[34]	Xception	91.9
MetaFormer[35]	MetaFormer-1	92.3
本文	Swin-L	92.9

如表 1 所示, 提出的 DAMSRL 的 Top-1 精度可以达到 92.9%, 优于其他 10 种现有方法. 在原先的方法中, MetaFormer 使用了额外的文本信息作为输入, 通过改进 ViT 的令牌混合器实现了很大的性能提升; SR-GNN 方法关注到全局结构关系对细粒度分类有重要的意义, 采用上下文注意力机制细化关系特征从而构建对象不同部位的结构提升了细粒度分类精度. 而在表 2 中, 将本文提出的 DAMSRL 方法与 NA-Birds 数据集上最先进的方法进行了比较, 提出的方法的 Top-1 精度可以达到 93.0%, 同样优于其他方法. 原有的先进方法如 API-Net 通过成对交互注意力网络来捕捉输入的图像对间细微差异从而学习细粒度分类线索; CS-Parts 通过初始预测结果的反向传播来学习修正判别区域; TransIFC 设计了一种不变线索感知特征变换器来捕捉鸟类细粒度分类中的不变特征.

表 2 在 NA-Birds 数据集上与其他方法的对比 (%)

方法	主干网络	Top-1精度
API-Net[36]	DenseNet-161	88.1
CS-Parts[37]	InceptionV3	88.5
TPSKG[21]	ViT-B_16	90.1
DeepFVE[31]	InceptionV3	90.3
TransFG[14]	ViT-B_16	90.8
TransIFC[38]	Swin-B	90.9
CAP[33]	Xception	91.0
SR-GNN[34]	Xception	91.2
MetaFormer[35]	MetaFormer-1	92.7
本文	Swin-L	93.0

与上述方法相比, 本文既关注到结构信息对细粒度分类至关重要, 采用结构关系学习模块来构建不同部件区域的结构关系, 同时还关注到网络容易仅从最显著区域寻找分类线索, 导致网络容易混淆相似类别, 因此提出利用动态自适应调制模块在实现部件区域预测的同时, 迫使网络从更多判别区域中寻找分类线索, 大大降低网络区分相似类别的难度. 因此, 在 CUB-200-2011 数据集和 NA-Birds 数据集上本文提出的方法均取得最高的细粒度分类精度.

4.3 消融实验及可视化

首先, 为研究聚类数量对网络的细粒度分类准确率影响, 如表 3 所示, 在 CUB-200-2011 数据集上, 采用 Swin-L 作为主干网络, 并保持其余参数不变的前提下, 通过设置不同的聚类数量, 进行多次实验. 从实验结果可以看出当聚类数设置为 5 时, 网络的准确率最高. 因此, 本文聚类数量设置为 5 类, 也即定位到的部件区域数量为 5 个.

表 3 在 CUB-200-2011 数据集测试不同聚类数量对准确率影响 (%)

聚类数量	准确率
5	92.9
6	92.6
7	92.3

然后, 为更好理解 DAMSRL 中提出的每个模块的影响, 分别将 FPN、DAM 和 SRL 模块添加到主干网络上. 首先, 使用 Swin-L 作为测试主干网络, 使用 CUB-200-2011 数据集作为测试集. 如表 4 所示, 在数据集

上 Swin-L 的原始准确率为 91.9%, 之后逐步添加各个模块, 在添加了 FPN 后准确率略有提高, 提高了 0.1%, 在此基础上再添加 DAM 模块, 准确率再提高了 0.3%; 表 4 最后一行显示添加了 DAMSRL 模块将主干网络准确率提高约 1%, 证明了添加模块的有效性.

表 4 在 CUB-200-2011 数据集上添加不同模块准确率对比 (%)

添加模块	准确率
Backbone	91.9
+FPN	92.0
+FPN+DAM	92.3
+FPN+DAM+SRL	92.9

进一步利用 Grad-CAM[11]方法生成热力图, 如图 5 所示, 在 CUB-200-2011 数据集上选择狐色雀鹀和尼尔森沙鹀作为例子, 可视化添加不同模块时模型响应区域变化, 用红色和绿色标注高响应区域, 图 5 中从左往右各列分别为 (a) 原始图片, (b) 仅使用主干网络时响应分布情况, (c) 添加了特征金字塔后网络响应情况, (d) 添加了特征金字塔和动态自适应调制模块后网络响应情况, (e) 添加了全部模块后网络响应情况. 对比各响应情况, 在仅使用主干网络时候, 几乎全图都分布有高响应区域, 表明原始主干网络令牌特征混淆了前景和背景信息, 将背景信息也作为了细粒度分类线索, 而这显然不利于细粒度分类. 在添加 FPN 网络后, 高响应区域集中在很小的局部区域, 与原始网络相比, 这种改进使得网络降低了对背景的关注, 但容易导致一是网络关注到的错误判别区域, 二是网络集中关注区域过小不足以区分相似类别. 与仅添加特征金字塔时高响应区域分布相比, 添加了动态自适应调制模块后, 由于网络可以通过聚类方法定位到多个部件区域, 并利用高斯调制来重塑注意力分布, 这使得网络高响应区域分散分布在鸟类对象各个部件区域, 表明此时网络关注到更多局部具有辨别力的区域, 但仍然缺失对判别区域空间关系的挖掘, 从而容易将遮挡鸟类部件的树枝也当做分类线索. 最后, 将所有模块添加到主干网络上, 此时高响应区域集中于鸟类各部件上, 且通过添加了结构关系学习模块增强了对干扰信息的分辨能力, 让网络能区分树枝等干扰信息, 同时又能发现更多的分类线索, 充分挖掘对象部件区域的细微差异.

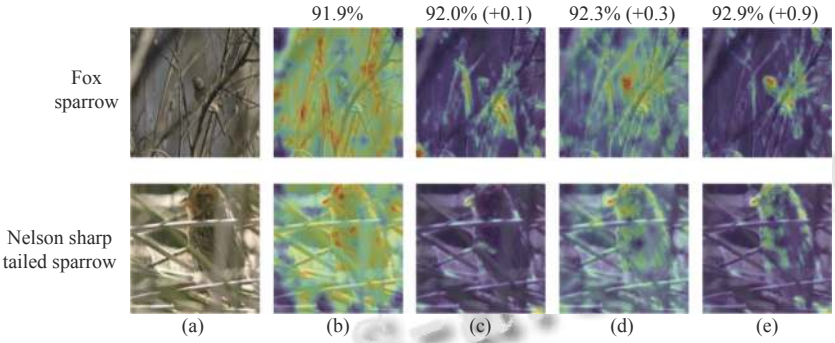


图 5 在 CUB-200-2011 数据集上可视化结果

在这个例子中, 由于密集的树枝遮挡, 只有添加了全部模块的网络正确分类了两种鸟类, 表明添加模块后, 即使存在严重遮挡, 网络依然能够具有良好的区分前景与背景能力, 并从中寻找到了鸟类多个局部判别区域, 并通过构建结构关系来融合多个局部的有效信息给出更准确的预测结果.

5 结论

在本文中, 我们提出了具有动态自适应调制模块和结构关系学习模块的 DAMSRL 网络. 所提出的模块能使网络关注到适当规模的局部判别区域, 有效区分前景和背景信息, 并将结构信息融合进所提取的类别特征, 从而促使网络更好区分相似类别. 我们在细粒度视觉分类任务上的实验表明, DAMSRL 网络显著提高了准确性, 并且在 CUB-200-2011 和 NA-Birds 基准数据集上优于最先进的方法. 未来工作可以基于 DAMSRL 网络探索如何更好定位小目标的判别区域. 总的来说, 所提出的 DAMSRL 网络在两个数据集上均可以达到接近 93% 的 Top-1 精度, 为提高细粒度视觉分类任务的性能提供了一个有效的解决方案.

参考文献

1 Wah C, Branson S, Welinder P, *et al*. The caltech-UCSD Birds-200-2011 dataset. Technical report, Pasadena:

California Institute of Technology, 2011.
2 Khosla A, Jayadevaprakash N, Yao B, *et al*. Novel dataset for fine-grained image categorization: Stanford dogs. Proceedings of the 2011 CVPR Workshop on Fine-grained Visual Categorization (FGVC): Vol. 2. 2011.
3 Maji S, Rahtu E, Kannala J, *et al*. Fine-grained visual classification of aircraft. arXiv:1306.5151, 2013.
4 Krause J, Stark M, Deng J, *et al*. 3D object representations for fine-grained categorization. Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops. Sydney: IEEE, 2013. 554–561.
5 Huang SL, Xu Z, Tao DC, *et al*. Part-stacked CNN for fine-grained visual categorization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1173–1182.
6 Wei XS, Xie CW, Wu JX, *et al*. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. Pattern Recognition, 2018, 76: 704–714. [doi: 10.1016/j.patcog.2017.10.002]
7 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.
8 Ge WF, Lin XR, Yu YZ. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3029–3038.
9 Yang SK, Liu S, Yang C, *et al*. Re-rank coarse classification with local region enhanced features for fine-grained image recognition. arXiv:2102.09875, 2021.
10 Zhou BL, Khosla A, Lapedriza A, *et al*. Learning deep features for discriminative localization. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2921–2929.
11 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 618–626.
12 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
13 Wang J, Yu XH, Gao YS. Feature fusion Vision Transformer for fine-grained visual categorization. Proceedings of the 32nd British Machine Vision Conference. BMVA Press, 2021. 170.
14 He J, Chen JN, Liu S, *et al*. TransFG: A Transformer architecture for fine-grained recognition. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI Press, 2022. 852–860.
15 Zheng HL, Fu JL, Zha ZJ, *et al*. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5007–5016.
16 Rao YM, Chen GY, Lu JW, *et al*. Counterfactual attention learning for fine-grained visual categorization and re-identification. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 1005–1014.
17 Chen Y, Bai YL, Zhang W, *et al*. Destruction and construction learning for fine-grained image recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5152–5161.
18 Song JW, Yang RY. Feature boosting, suppression, and diversification for fine-grained visual classification. Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen: IEEE, 2021. 1–8.
19 Zhang Y, Cao J, Zhang L, *et al*. A free lunch from ViT: Adaptive attention multi-scale fusion Transformer for fine-grained visual recognition. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 3234–3238.
20 Zhao YF, Li J, Chen XW, *et al*. Part-guided relational Transformers for fine-grained visual recognition. IEEE Transactions on Image Processing, 2021, 30: 9470–9481. [doi: 10.1109/TIP.2021.3126490]
21 Liu XD, Wang LL, Han XG. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. Neurocomputing, 2022, 492: 137–149. [doi: 10.1016/j.neucom.2022.04.037]
22 Yu QH, Wang HY, Qiao SY, *et al*. K-means mask Transformer. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 288–307.
23 Kosman E, Di Castro D. GraphVid: It only takes a few nodes to understand a video. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 195–212.
24 Wang ZH, Wang SJ, Li HJ, *et al*. Graph-propagation based correlation learning for weakly supervised fine-grained image classification. Proceedings of the 34th AAAI

Conference on Artificial Intelligence. New York: AAAI Press, 2020. 12289–12296.

25 Zhao YF, Yan K, Huang FY, *et al*. Graph-based high-order relation discovery for fine-grained recognition. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 15074–15083.

26 Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview.net, 2017.

27 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.

28 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical Vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.

29 李文书, 王志骁, 李绅皓, 等. 基于注意力机制的弱监督细粒度图像分类. 计算机系统应用, 2021, 30(10): 232–239. [doi: 10.15888/j.cnki.csa.008141]

30 Du RY, Chang DL, Bhunia AK, *et al*. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 153–168.

31 Korsch D, Bodesheim P, Denzler J. End-to-end learning of fisher vector encodings for part features in fine-grained recognition. Proceedings of the 43rd DAGM German Conference on Pattern Recognition. Bonn: Springer, 2021. 142–158.

32 Miao Z, Zhao X, Wang JB, *et al*. Complemental attention multi-feature fusion network for fine-grained classification. IEEE Signal Processing Letters, 2021, 28: 1983–1987. [doi: 10.1109/LSP.2021.3114622]

33 Behera A, Wharton Z, Hewage PRPG, *et al*. Context-aware attentional pooling (CAP) for fine-grained visual classification. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI Press, 2021. 929–937.

34 Bera A, Wharton Z, Liu YH, *et al*. SR-GNN: Spatial relation-aware graph neural network for fine-grained image categorization. IEEE Transactions on Image Processing, 2022, 31: 6017–6031. [doi: 10.1109/TIP.2022.3205215]

35 Diao QS, Jiang Y, Wen B, *et al*. MetaFormer: A unified meta framework for fine-grained recognition. arXiv:2203.02751, 2022.

36 Zhuang PQ, Wang Y, Qiao Y. Learning attentive pairwise interaction for fine-grained classification. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020. 13130–13137.

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

38 Liu H, Zhang C, Deng YJ, *et al*. TransIFC: Invariant cues-aware feature concentration learning for efficient fine-grained bird image classification. IEEE Transactions on Multimedia, 2023. (published online). [doi: 10.1109/TMM.2023.3238548]

(校对责编: 孙君艳)