分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220920112

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于深度学习的细粒度图像分类方法与研究

论 文 作 者　丁伯宇
指 导 教 师　徐晓峰(副教授)
学科（专业）　软件工程
研 究 方 向　计算机视觉

论文提交日期：2025 年 5 月 9 日

分类号：TP391　　单位代码：10363
密　级：公开　　学　　号：2220920112

题　目　基于深度学习的细粒度图像分类方法与研究

题　目　FINE-GRAINED IMAGE CLASSIFICATION METHOD AND RESEARCH BASED ON DEEPLEARNING

学生姓名：丁伯宇
指导教师：徐晓峰(副教授)
专　　业：软件工程
研究方向：计算机视觉

论文答辩日期：2025 年 5 月 10 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：丁伯多

日期：2025 年 5 月 9 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位 保密 □，在 ____ 年解密后适用本版权书。

论文属于 不保密 ☑。

学位论文作者签名：丁伯多　　指导教师签名：徐晓峰

日期：2025年5月9日　　日期：2025 年5月9 日

基于深度学习的细粒度图像分类方法研究

摘 要

细粒度图像分类（Fine Grained Image Classification, FGIC）是在区分出图像基本类别的基础上，再次划分出对应的子类，因此与传统的图像分类比较，对于图像的鉴别性特征选择提出了更高的要求。细粒度图像分类不仅存在类间差异小、类内差异大的问题，还存在光照、视角、遮挡、背景等干扰因素，使模型分类更具有挑战性。FGIC 任务常采用基于多阶段裁剪的方法，虽然这类方法能够捕捉到重要的区域，却难以对这些区域的重要性做出客观的评估。目前，现有的 FGIC 方法大多利用注意机制提取高响应区域的视觉特征作为鉴别性特征，然而其中的视觉特征也包含粗粒度信息（非鉴别性特征），这不仅会提升模型的计算量，还会干扰后续阶段的鉴别性特征提取。针对上述问题，本文的主要工作如下：

（1）针对模型识别区域容易受到背景干扰的问题，本文提出了基于多阶段特征提取的细粒度图像分类方法。该模型由两个模块组成，即可调整定位模块和多角度注意模块。在可调整定位模块中，首先定位目标感兴趣的区域，并通过背景遮掩对感兴趣区域进行调整，得到调整后的裁剪区域。然后，对调整后的区域进行聚集，定位潜在的鉴别性特征。此外，对于多角度注意模块，逐步最大化原始注意图与随机选择的注意图之间的差异。因此，该方法相较于其它方法具有更好的泛化性。

（2）传统基于多阶段的细粒度图像分类的方法虽然能够捕捉到重要的区域，却难以对这些区域的重要性做出客观的评估。针对这一问题，本文提出了基于树形分支选举的细粒度图像分类模型。首先，该模型在初始阶段模型通过最大化与粗粒度信息的差异，能够将关注的区域引导到目标上，并为后续阶段提供更有效的指导。随后阶段以前一阶段为基准，对高响应区域进行裁剪，得到新的图像并逐渐发现鉴别性特征。通过决策树集成的多个分支路由模块对决策树的子分支选举，不断增加鉴别性特征所在阶段的贡献。最后，

汇总来自叶节点的预测以获得最终决策。实验结果表明，该方法相比于其它方法具有更好的表现。

（3）针对高响应区域中包含粗粒度信息（非鉴别性特征）问题，本文提出了基于多尺度特征提取的 Transformer 细粒度图像分类。具体而言，该方法包含过滤聚合模块和低温冻结模块，以选择有效的高响应区域并学习更丰富的判别特征。一方面，为了选择有效的高响应区域，过滤聚合模块首先根据特征图之间的相关性过滤掉无用的低响应区域。然后，从真正的高响应区域聚合特征进一步扩大了差异。另一方面，所提出的低温冻结模块能够学习更丰富的鉴别特征。可变化的温度函数鼓励模型探索更多的鉴别性区域并将弱相关特征冻结在不同尺度的特征图中，提取合适的鉴别特征。实验结果表明，该方法能够去除粗粒度信息的同时，发现潜在的鉴别性区域，以提高模型的分类能力。

关键词：细粒度图像分类, 多尺度特征提取, 特征增强, 注意力机制

Fine-grained Image Classification Method and Research Based on Deep Learning

ABSTRACT

Fine-grained Image Classification (FGIC) classifies the basic categories of images and then divides the corresponding subcategories. Compared with the traditional image classification, it puts forward higher requirements for the discriminative feature selection of images. Fine-grained image classification not only has the problem of small inter-class differences and large intra-class differences, but also has interference factors such as illumination, view Angle, occlusion, background, which makes model classification more challenging. FGIC tasks often use methods based on multi-stage pruning. Although these methods can capture important regions, it is difficult to make an objective evaluation of the importance of these regions. At present, most of the existing FGIC methods use the attention mechanism to extract the visual features of high-response regions as discriminative features. However, the visual features also contain coarse-grained information (non-discriminative features), which will not only increase the computational complexity of the model, but also interfere with the discriminative feature extraction in the subsequent stage. Aiming at the above problems, the main work of this paper is as follows:

(1) Aiming at the problem that the recognition area of the model is susceptible to background interference, this paper proposes a fine-grained image classification method based on multi-stage feature extraction. The model consists of two modules, the positioning module and the multi-angle attention module. In the adjustable positioning module, the region of interest of the target is located at first, and the region of interest is adjusted by background masking to

obtain the adjusted cropping region. Then, the adjusted regions are aggregated to locate potential discriminative features. In addition, for the multi-view attention module, the difference between the original attention map and the randomly selected attention map is gradually maximized. Therefore, this method has better generalization than other methods.

(2) Although the traditional multi-stage fine-grained image classification method can capture the important regions, it is difficult to make an objective assessment of the importance of these regions. To solve this problem, this paper proposes a fine-grained image classification model based on tree branch election. Firstly, by maximizing the difference with coarse-grained information in the initial stage, the model can guide the concerned area to the target, and provide more effective guidance for the subsequent stages. The subsequent stage is based on the previous stage, cropping the high response region, obtaining new images and gradually discovering discriminative features. The multiple branch routing modules of the decision tree ensemble elected the sub-branches of the decision tree, and continuously increased the contribution of the stage where the discriminative features were located. Finally, the predictions from the leaf nodes are summarized to obtain the final decision. Experimental results show that the proposed method has better performance than other methods.

(3) Aiming at the problem that high-response regions contain coarse-grained information (non-discriminative features), this paper proposes a Transformer fine-grained image classification based on multi-scale feature extraction. Specifically, the proposed method contains a feature aggregation module and a low-temperature freezing module to select effective high-response regions and learn richer discriminative features. On the one hand, in order to select effective high-response regions, the filter aggregation module first filters out

useless low-response regions based on the correlation between feature maps. Then, aggregating features from the true high-response regions further widens the difference. On the other hand, the proposed cryofreezing module is able to learn richer discriminative features. The variable temperature function encourages the model to explore more discriminative regions and freeze weakly correlated features in feature maps at different scales to extract appropriate discriminative features. Experimental results show that the proposed method can remove coarse-grained information and find potential discriminative regions to improve the classification ability of the model.

Key words: fine-grained image classification, multi-scale feature extraction, feature enhancement, attention mechanism

目录

第 1 章 绪论

1.1 研究背景和意义

随着计算机视觉和深度学习技术的快速发展，图像分类任务近年来取得了显著进展。但由于光照、视角、遮挡、背景等干扰因素，导致类间差异小、类内差异大，因此在细粒度图像分类（Fine Grained Image Classification，FGIC）中依然存在挑战。2022 年 8 月科技部等六部门联合印发《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》指出：优先探索精细识别检测技术促人工智能专业的发展具有重要意义。2023 年 3 月 13 日，政策法规研究所召开“人工智能产业政策”研究讨论会，大会指出人工智能是未来国际竞争的焦点和经济发展的新引擎。因此，细粒度图像分类任务具有重要的应用价值。细粒度图像分类作为计算机视觉重要组成部分，在医学、商业、安防和自然保护等领域被广泛应用。

图 1-1 粗粒度与细粒度图像分类示意图

Figure 1-1 Schematic diagram of coarse-grained and fine-grained image classification

细粒度图像分类是对属于同一大类但属于不同子类的对象进行更深入的分类。如图 1-1 所示，传统的图像分类任务主要是粗粒度分类，例如区分鸟类和飞机等大类。这些大类之间的特征差异显著，即使像传统的卷积神经网络（Convolutional Neural Network, CNN）网络也能获得较好的分类效果。然而，细粒度图像识别则需要更高的精度和细化程度。这是因为模型不仅要求识别出图像中的鸟或飞机，还需要进一步准确地识别出具体的子类，例如区分不同种类的鸟或飞机型号。因此，细粒度图像分类对图像的精度和细化程度提出了更高的要求，这成为计算机视觉领域中极具挑战性的问题。

细粒度识别技术能够用于生活的各个方面，如医学领域、安防、智能交通、智能零售、气候变化评估等。具体而言：(1)在医学领域方面，特别是与视网膜相关的疾病，通常在确诊前视网膜会表现出微小的病理变化，并且这些变化在早期症状上不易被察觉，但却能在病理成像中以微小的差异表现出来。若在早期能诊断出视网膜病变，这将极大的增加病人被治愈的概率，因此细粒度图像分类在医学领域发挥重要作用。(2)在安防场景中，细粒度图像分类技术可以结合面部识别技术，用于身份验证和访问控制。这可以确保只有授权人员能够进入敏感区域，提高安全性。(3)在智能交通方面，主要应用与交通监控和车辆识别。例如，通过道路监控摄像头捕捉到的图像，利用细粒度图像分类技术，可以精确地识别和分类经过的车辆，包括车辆的品牌、型号等细粒度信息，这有助于交通管理部门更准确地掌握道路车辆的情况，为交通规划、拥堵预警等提供数据支持。(4)在智能零售场景下，商品识别技术带来的高效性与便利性可帮助企业减少成本的同时并提升收益。例如，超市货架的同一区域通常摆放同一品类但不同品牌和规格的商品，而同品牌商品的不同规格在外观上具有较高的相似度。(5)在气候评估方面的应用主要表现为通过卫星或无人机图像对地表覆盖类型进行分类和识别。这种分类有助于更精确地了解地表的特征，如植被覆盖、土地利用情况等。基于这些信息，气候评估人员可以对地区的气候状况进行更准确的评估，例如判断地区的干旱程度、植被健康状态等。

1.2 研究挑战

尽管卷积神经网络和 Vision Transformer（ViT）在一般图像分类任务中表现出越来越强的实用性和广泛的应用前景，但细粒度图像分类仍然面临诸多挑战。如图 1-2 所示，这一挑战主要源于类间低方差和类内高方差的特性，显著增加了模型区分细粒度类别的难度。

具体而言，类内高方差（红色虚线框）表现为：同一类别的实例在图像中可能存在显著的视觉差异。例如，鸟类的图像因外部因素表现出明显不同的外观特征，这些因素包括光照条件的变化（如第 1 列所示），拍摄角度的多样性（如第 2 列所示），复杂的背景干扰（如第 3 列所示）以及部分图像遮挡（如第 4 列所示）。这些因素提高了模型在提取图像特征时的难度。

图 1-2 FGIC 面临的挑战

FIG 1-2 Challenges faced by FGIC

另一方面，类间低方差（蓝色虚线框）则表现在不同类别的实例之间存在高度相似的外观特征。例如，不同鸟类在形状、颜色和纹理等方面往往高度相近（如同一行所示），这种细微的差别对传统的特征提取方法提出了更高的要求。对于现有的深度学习模型而言，在这种情况下区分不同类别需要模型具备更强的局部特征捕捉能力和对细粒度特征的敏感性。

1.3 研究现状

细粒度图像分类可归为三类：基于局部区域定位的细粒度图像分类、基于注意力机制的细粒度图像分类和基于 Transformer 的细粒度图像分类。其中，基于局部区域定位的细粒度图像方法是通过多尺度特征金字塔（Feature PyramidNetwork，FPN）或候选区域生成网络（Region Proposal Network，RPN）精准锚定目标区域，利用小尺寸卷积核强化边缘、纹理等细微特征的提取能力。随着深度学习对特征动态交互需求的提升，注意力机制因其能够自适应地捕捉图像中与任务密切相关的局部-全局关联性，逐渐成为细粒度分类任务的关键补充。基于注意力机制的细粒度图像分类方法在局部特征的基础上通过卷积注意力（Convolutional Block Attention Module，CBAM）或交叉注意力动态筛选关键信息，抑制背景干扰并提升特征表达的鲁棒性。随着细粒度任务对全局上下文建模需求的深化，基于 Transformer 的细粒度图像分类方法凭借其独特的全局依赖建模能力，为突破局部特征局限、实现更精准的语义关联提供了新的思路。基于 Transformer 的全局建模层突破 CNN 的局部感受野限制，通过 ViT 将图像划分为 Patches 并引入位置编码，获长程依赖关系。本节将对这些方法进一步说明。

1.3.1 基于局部区域定位的细粒度图像分类

随着计算机视觉和深度学习技术的快速发展[1-4]，FGIC 任务近年来取得了显著进展[5-8]。在 FGIC 任务中，准确的定位鉴别性区域是提升分类性能的关键[9]。在近些年的研究中，许多方法利用人工标注的边界框或局部区域来实现目标定位(如图 1-3(a)所示)。例如，Branson 等[10]通过估计待识别图像中的物体的姿态来优化检测到的局部区域；Zhang 等[11]在挖掘局部区域的过程中考虑了不同局部区域之间的位置关系；Lin 等[12]则将不同的局部区域和目标主体进行对齐，并通过一种新颖的 valve linkage 函数来调整检测到的局部区域；Zhang 等[13]则通过关键点定位信息将挖掘到的局部区域与其语义概念相对应，从而调整检测到的局部区域的大小。此外，Huang 等[14]和 Wei 等[15]则通过目标分割技术对局部区域进行的特征提取并关联。由于标注出图像鉴别性区域，所以这类的方法能够提高了模型的分类性能。然而，这类方法局限在于标注难度大、费时费力且非专业人员难以进行标注，这严重限制了该类方法在实际场景中的应用。

最近，研究人员转向了弱监督网络，这种网络在没有人类边界框或部分注释的情况下运行。这些方法通常依赖于滤波响应来识别相关的鉴别区域。这些区域定位方法根据零件标注来识别区分区域，并使用区域建议网络[16-17]生成包含这些区域（如图 1-3(b)所示）。然而，它们忽略了所选区域之间的关系，因此无法确定真正重要的区域。具体而言，图像定位的区域通常需具备明确的语义信息，这限制了模型对细粒度特征信息的挖掘。其原因在于，许多细粒度类别中包含难以明确界定的视觉区域，如鸟喙上的花纹或尾部的纹路等。

1.3.2 基于注意力机制的细粒度图像分类

注意力机制作为一种高效的局部区域定位方法，可引导模型关注不同尺度的视觉区域，从而增强其识别能力。此外，注意力机制在人类感知过程中同样到了关键作用：当人类识别一个物体时，会自然地寻找物体的起判别区域[18]，并能较好地掌握不同视觉区域之间的空间关系[19]以分辨对应的类别。因此，注意力机制也被应用到了 FGIC 任务中（如图 1-4 所示），其中一些方法也取得的较好的效果。Hu 等[20]介绍了一种用于生成裁剪框的注意力数据增强多级裁剪方法。该方法使目标能够检测到有区别的区域，提供了一种更主动的控制方法。

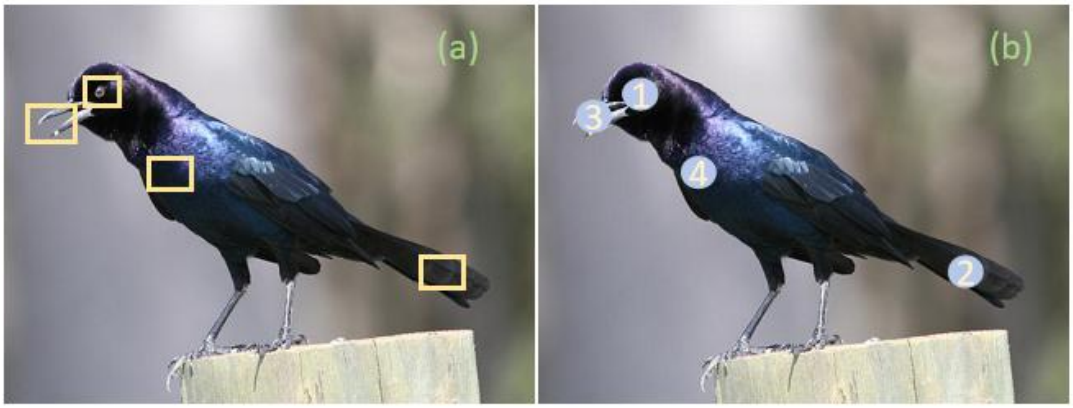

图 1-3 基于人工标注的边界框(a)和局部区域示意图(b)

Figure 1-3 Bounding box (a) and local area (b) based on manual annotation

Zhang 等[21]提出了一个多专家和门控网络的模型。然后，门控网络确定每个专家分配给最终预测的权重。Zheng 等[22]利用注意机制将不同的特征通道分组为不同的视觉模式，然后将其投射到类别预测中。Zhang 等[23]开发了一个共同关注模块来测量同一类内成对样本之间的通道特征相似性。在文献[24]中，作者提出了级联注意力机制来引导 CNN 的不同层和将它们连接起来以获得鉴别表示，作为最终线性分类器的输入。Jetley 等[25]将 CBAM 模块将空间区域关注与特征图关注相结合。Zhu 等[26]探讨了如何拓展自注意力机制以更好地学习微妙的特征嵌入，通过一种双向交叉注意力机制促使模型捕捉图像对之间的关联特征。Shu 等[27]提出了一个自增强注意力机制约束模型，约束模型在学习分类的同时，学习其通过通道注意力机制（Channel Attention Mechanism，CAM） 获得的热力图。尽管这些区域定位方法可以捕捉子类别之间的细微差异，但由于细粒度图像具有类间差异小、类内差异大的特点，过度依赖单一或有限数量的类别特征可能导致上下文信息的丢失。

1.3.3 基于 Transformer 的细粒度图像分类

基于 Transformer 的 ViT 框架已成为许多计算机视觉任务的研究热点。ViT 的核心是多头自注意机制，该机制通过计算输入补丁之间的相互关系，从而捕获全面的对象特征表示。

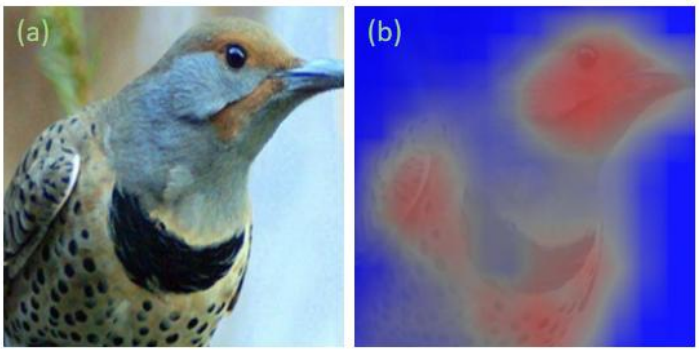

图 1-4 原始图像（a）与注意力图可视化（b）

Figure 1-4 Original image (a) and attention map visualization (b)

Zhang 等[28]提出了一种识别最具区别性的区域并将其重新引入网络的方法，以实现更具区别性的特征表示。Zhu 等[29]介绍了一种双交叉注意学习算法，增强了全局图像特征与局部高响应区域之间的相互作用。DosoVitkiy 等[30]提出使用 ViT 将 2D 图像分割成一组序列，并使用 Transformer 架构来实现与 CNN 在计算机视觉任务中相似的性能。尽管 ViT 可以捕获全局特征，但捕捉局部特征能力有时欠佳，这限制了它在 FGIC 任务中的性能。为了解决这一问题，Wang[31]等人提出了一种用于细粒度识别的 Transformer 架构，该架构将图像分割成重叠的小块，以获得更丰富的特征，并使用有区别的图像小块进行分类。He 等[32]引入了一种专门为细粒度识别设计的新颖的转换器体系结构，称为 transfg。该方法通过将图像划分为重叠的小块来增强特征提取过程，使模型能够捕获更丰富和更细微的特征。Hu 等[33]提出将定位最具鉴别性的区域并再次馈送到网络中进行详细的特征表示。Liu 等[34]提出在预测过程中抑制反应最灵敏的 token，使网络能够找到更丰富的特征，并将其添加到知识库中。Tang 等[35]提出通过挖掘显著斑块的空间上下文关系，通过 ViT 整合结构信息，并采用对比学习策略增强特征的鲁棒性。Zhao 等[36]提出使用多个 Transformer block 来提取全局特征和局部特征。Zhu 等[37]提出了一种双交叉注意学习算法，增强了全局图像与局部高响应区域之间的交互，建立了图像对之间的交互。

尽管如此，上述方法中检测到的高响应区域需要进一步选择，因为一些噪声区域被用作鉴别区域。

1.4 本文的研究工作

为实现一个高准确率的细粒度分类模型，围绕解决深度神经网络的图像分类模型

识别区域容易受到背景干扰的问题、模型难以辨别鉴别性特征所在阶段和检测的高响应区域中存在粗粒度信息的问题，提出了一系列细粒度图像分类方法：（1）通过多阶段鉴别性特征提取学习更多的细节特征，（2）通过树形分支选举探索鉴别性特征所在的阶段，（3）通过多尺度特征提取的 Transformer 方法学习更多的细粒度特征，从而实现对细粒度类别间差异的充分挖掘，提升细粒度分类的精度，并通过大量实验验证了所提方法的有效性。所提出的三种方法从减少背景干扰、评估鉴别特征的重要性和发现潜在鉴别性特征的逻辑依次展开叙述。本文的研究内容与贡献具体如下：

（1）针对细粒度图像存在类间差异小、类内差异大的问题，本文提出了基于多阶段特征提取的细粒度图像分类方法。该模型由两个分支组成，即可调整定位模块和多角度注意模块。具体来说，在可调整定位模块中，首先定位目标感兴趣的区域，并通过背景遮掩对感兴趣区域进行调整，得到调整后的裁剪区域。然后，对调整后的区域进行聚集，定位潜在的鉴别性特征。此外，对于多角度注意模块，逐步最大化原始注意图与随机选择的注意图之间的差异。因此，该方法相较于其它方法具有更好的泛化性。

（2）针对模型难以辨别鉴别性特征所在阶段的问题，本文提出了基于树形分支选举的细粒度图像分类模型。首先，在初始阶段模型通过最大化与粗粒度信息的差异，能够将关注的区域引导到目标上，并为后续阶段提供更有效的指导。随后阶段以前一阶段为基准，对高响应区域进行裁剪，得到新的图像并逐渐发现鉴别性特征。通过决策树集成的多个分支路由模块对决策树的子分支选举，不断增加鉴别性特征所在阶段的贡献。最后，汇总来自叶节点的预测以获得最终决策。实验结果表明，该方法相比于其它方法具有更好的表现。

（3）针对高响应区域中包含粗粒度信息问题，本文提出了基于多尺度特征提取的 Transformer 细粒度图像分类。具体而言，该方法包含过滤聚合模块和低温冻结模块，以选择有效的高响应区域并学习更丰富的鉴别特征。一方面，为了选择有效的高响应区域，过滤聚合模块首先根据特征图之间的相关性过滤掉无用的低响应区域。然后，从真正的高响应区域聚合特征进一步扩大了差异。另一方面，所提出的低温冻结模块能够学习更丰富的鉴别特征。可变化的温度函数鼓励模型探索更多的鉴别性区域并将弱相关特征冻结在不同尺度的特征图中，提取合适的鉴别特征。实验结果表明，该方

法能够去除粗粒度信息的同时，发现潜在的鉴别性区域，以提高模型的分类能力。

1.5 本文的结构安排

本文主要分为六个部分，其中各部分内容和总体路线如下。

第一章，绪论。主要通过 FGIC 的背景、意义和总体研究进展等方面依次进行阐述。随后对论文中提到的研究现状和和存在的一些问题进行阐述，为下文的创新性实验改进进行铺垫。

第二章，相关研究理论。主要介绍论文涉及的重要背景知识，包括概念、公式及推导过程等内容，并为后续创新性研究的论述和展开做好铺垫。

第三章，基于多阶段特征提取的细粒度图像分类方法。本文讨论了可调整定位模块和多角度注意模块的网络结构。主要包括该网络模型及全新模块的原理、内容、细节等。

第四章，基于树形分支选举的细粒度图像分类。这部分主要讨论了模型探索鉴别性特征所属阶段的过程。主要包括该模型的原理、内容、细节等。

第五章，基于多尺度特征提取的 Transformer 细粒度图像分类。这部分主要描述丢弃出粗粒度信息和探索更多细粒度信息的过程，并进行详细解释。

第六章，总结与展望。主要对于本文的整体内容进行总结，并对于未来的研究进行展望。

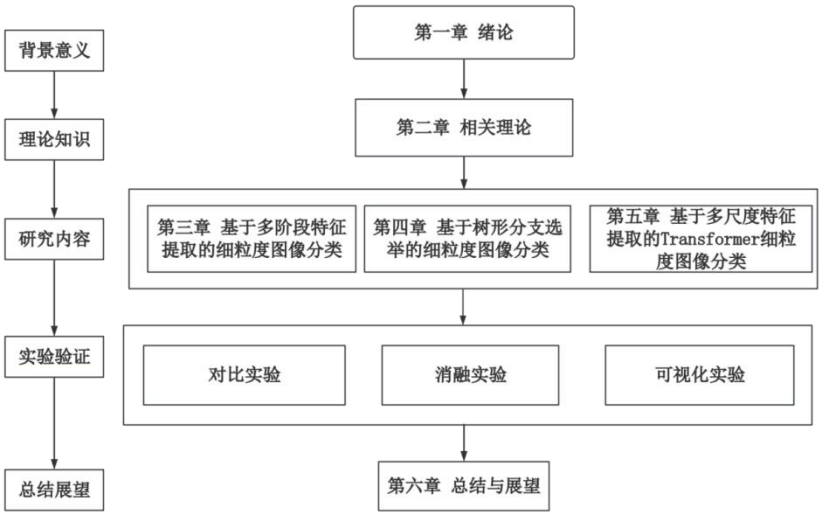

图 1-5　论文的组织结构图

Figure 1-5 Organizational chart of the paper

第 2 章 细粒度图像分类相关理论

本章介绍了 FGIC 方法所需要的相关理论基础。主要包括数据增强方法、特征融合方法、残差神经网络、注意力机制以及 ViT 等内容。相应的技术和内容指标将在第 3、4、5 章节中涉及。

2.1 数据增强方法

在深度学习模型训练过程中，常面临数据集样本不均衡、鲁棒性不足、模型过拟合及泛化能力弱等问题。因此，合理应用数据增强方法显得尤为重要。对于图像数据集，简单而有效的变换方法包括翻转、旋转、裁剪和缩放等。此外，当前 FGIC 任务中还采用了一些更为有效的数据增强技术，如图 2-1 所示：Mixup、CutMix、随机擦除、高斯噪声等。本文将在本节介绍这些方法。

图 2-1 常见的数据增强方法示意图

Figure 2-1 Schematic diagram of common data augmentation methods

2.1.1 随机裁剪

随机裁剪通过在原始图像中随机选取一个子区域进行裁剪，并将其调整为固定尺

寸。它可以去除图像中的某些不必要或多余的背景信息，迫使模型专注于图像中的主要目标。随机裁剪还可以模拟摄像头的不同拍摄角度或视场范围，生成具有轻微偏移或放大的图像。裁剪的过程可描述为：

假设原始图像为 $I \in \mathbb{R}^{H \times W \times C}$，裁剪后的图像为$I'$，其过程可表示为：

$$I' = I[y:y + h, x:x + w,:], \quad (2\text{-}1)$$

式（2-1）中，x,y 是裁剪区域的左上角坐标，h 和 w 是裁剪区域的高度和宽度，并满足以下条件：

$$h = R_h \cdot H，\ w = R_w \cdot H, \quad (2\text{-}2)$$

式（2-2）中 R_h 和 $R_w \in (0,1]$是随机比例。

随机裁剪迫使模型学会捕捉对象的局部和整体特征，而不是依赖固定的像素位置或背景信息，提高模型的鲁棒性。特别是对于目标检测和分类任务，随机裁剪能帮助模型适应实际中变化的场景和视角。

2.1.2 Mixup

Mixup 是一种线性混合策略，通过随机选取两张图像及其标签进行加权叠加，生成新的训练样本。Mixup 在增强样本的多样性的同时，还可以改变类别边界的分布，使模型在分类时更加平滑。当数据量不足或类别分布不平衡时，此方法有效缓解了训练过程中出现的过拟合问题。具体而言，给定两张图像及其标签，（I_1，y_1）和（I_2，y_2），混合样本为：

$$I' = \lambda I_1 + (1 - \lambda) I_2, \quad y' = \lambda\, y_1 + (1 - \lambda) y_2, \quad (2\text{-}3)$$

式（2-3）中，$\lambda \sim \mathrm{Beta}(\alpha,\alpha)$，$\alpha>0$ 为超参数。

Mixup 能够提高模型对噪声的抗干扰能力，同时减小对过拟合的依赖。它生成了许多人工样本，增强了模型在边界附近的区分能力，使得模型对未见数据的泛化性更强。

2.1.3 图像翻转

图像翻转是一种简单但有效的几何数据增强方法。水平翻转可以生成左右镜像的样本，而垂直翻转在某些特殊任务（如遥感影像分析）中同样适用。翻转的原理是对图像像素的位置进行重排，保持原图像内容的对称性，从而增加样本的多样性。具体而言：

对于水平翻转：

$$I'[i, j, :] = I[i, W - j - 1, :] \quad (2\text{-}4)$$

对于垂直翻转：

$$I'[i, j, :] = I[H - i - 1, j, :] \quad (2\text{-}5)$$

翻转操作本质上保留了图像的内容和目标特征，但以不同的空间排列形式呈现。对于自然图像（如动物、人脸）等，模型需要学习不依赖于方向的特征表示，因此翻转有效防止了方向性偏差。此外，这种方法无需要额外计算或复杂操作便可实现。

2.1.4 随机擦除

随机擦除是通过遮盖图像中的某些区域来模拟部分遮挡或损坏的图像场景。这种方法通常随机选择一个矩形区域并填充为常值（如零）或随机噪声值。它迫使模型学会利用全局信息，而不是过于依赖单一区域的特征。假设擦除区域的左上角为（x，y)，高度和宽度为（h，w)，则擦除后的图像为：

$$I'[y : y + h, x: x + w, :] = c, \quad (2\text{-}6)$$

式（2-6）中，c 是填充值，可以是常数或噪声值。随机擦除的效果类似于 Dropout，能够有效抑制模型对局部特征的依赖性，提高模型对遮挡和缺失信息的鲁棒性。在实际应用中，物体可能会部分被遮挡（如挡住的车牌、遮掩的人脸），通过随机擦除的方式能够帮助模型适应这种情况。

2.1.5 CutMix

CutMix 是 Mixup 的变体，它通过用另一张图像的局部区域替换当前图像的一部分来生成新样本，同时对标签进行加权。这种局部替换方法能够有效增强模型对复杂场景和遮挡情况的理解能力。CutMix 适合处理需要局部区域关注的任务，如目标检测和图像分类。CutMix 过程可描述为：给定两张图像及其标签，（I_1,y1）和（I_2,y2），生成的新样本为：

$$I_1[y : y + h, x: x + w, :] = I_2[y : y + h, x: x + w, :], \quad (2\text{-}7)$$

$$y' = \lambda\, y1 + (1 - \lambda)y2, \quad (2\text{-}8)$$

式（2-8）中λ为超参数。

CutMix 通过在空间维度上混合图像，保持了局部区域的真实信息，同时增加了样本的多样性。相比 Mixup，其生成的样本更接近于真实场景（如部分遮挡的物体），

能够提高模型的鲁棒性。

2.1.6 高斯噪声

高斯噪声通过在原始图像中加入随机噪声模拟实际传感器中常见的随机噪声。噪声通常服从零均值的正态分布，其标准差决定了噪声的强度。添加高斯噪声可以使图像在像素级别上发生微小变化，从而提升模型对噪声干扰的抵抗力。给定原始图像 I，加入高斯噪声后的图像为：

$$I' = I + N,\ N \sim n(0,\sigma^2), \tag{2-9}$$

式（2-9）中，N 是服从均值为零、方差为的高斯分布的噪声。

在实际场景中，图像往往会受到噪声的影响，例如低光环境下的拍摄或传感器故障等。通过在训练时引入高斯噪声，模型可以学会在噪声存在时提取有效特征，从而提升在复杂场景中的鲁棒性和稳定性。

2.2 特征融合方法

特征融合方法通过整合多源、多尺度或跨模态特征信息，显著提升了模型的表达能力和泛化性能。其核心优势在于能够突破单一特征的局限性，通过互补性增强对复杂模式的表征能力：浅层特征保留细节信息，深层特征捕获语义抽象，跨模态特征丰富上下文关联。这种融合机制不仅缓解了数据稀疏性问题，还增强了模型对噪声和遮挡的鲁棒性，使其在细粒度分类、目标检测等高精度任务中表现卓越。本节将介绍常用的特征融合方法如图 2-2 所示。

2.2.1 特征拼接

特征拼接是最直接的特征融合方法，通过将不同来源的特征向量在维度上进行连接，形成一个更高维的综合特征向量，其过程如图 2-2(a)所示。假设特征向量 $f_1 \in \mathbb{R}^{d1}$ 和 $f_2 \in \mathbb{R}^{d2}$,拼接后的特征为 $f_{concat} = [f1; f2] \in \mathbb{R}^{d1+d2}$。这种方法不会对原始特征进行任何变换，而是保留其原始信息完整性。拼接操作简单易用，且可以用于不同来源的特征，比如多模态数据(图像、文本、音频)的融合。特征拼接能够完整保留各来源特征的信息，因此适合在特征维度较低且特征重要性均等的场景中使用。通过直接增加维度，拼接可以扩展模型的表达能力，让模型同时利用来自多个来源的细粒度信息。

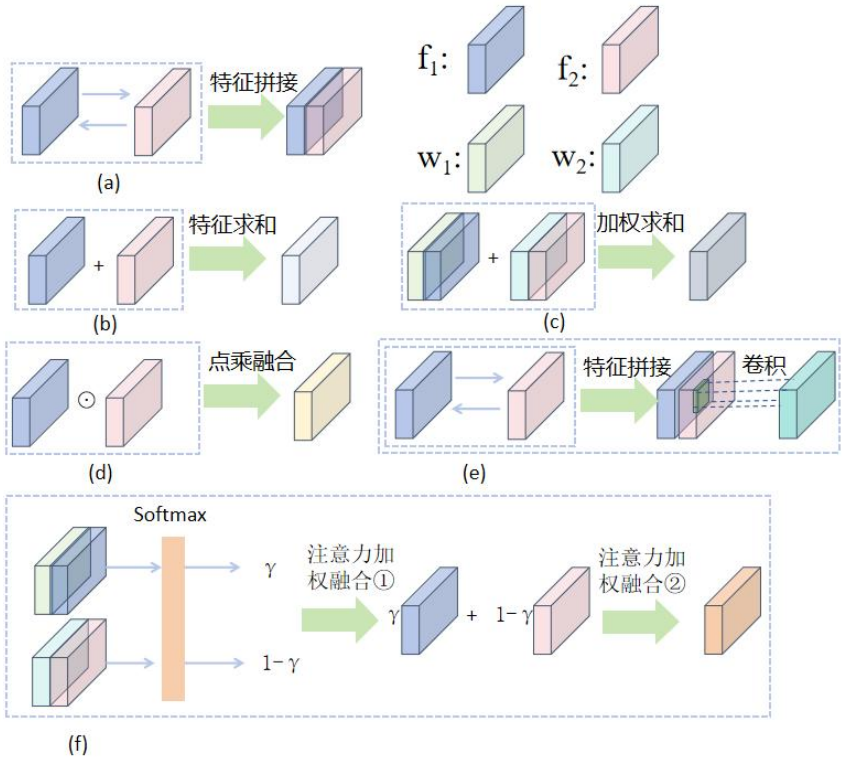

图 2-2 本文涉及的特征融合方法

Figure 2-2 Feature fusion methods involved in this paper

此外，拼接无需复杂运算，对计算资源要求较低，特别适合需要快速实现和简单融合的任务。

2.2.2 特征求和

特征加法是一种简单的融合方法，将两组特征逐元素相加生成新的融合特征，其过程如图 2-2(b)所示。给定 $f_1,f_2 \in \mathbb{R}^d$,融合特征为 $f_{add}=f_1+f_2$。假设两组特征在同一空间中分布，可直接通过相加操作生成有意义的复合特征。

特征加法的计算效率高，不会增加特征维度，同时保留了特征的整体信息。在特征分布较为均匀的情况下，加法能够很好地整合信息而不引入过多计算量，适合轻量级模型或实时应用场景。

2.2.3 特征加权求和

加权求和是一种线性特征融合方法，通过为每组特征分配权重进行线性组合，生成复合特征，其过程如图 2-2(c)所示。给定特征 $f_1,f_2 \in \mathbb{R}^d$,融合特征为 $f_{sum}=w_1f_1+w_2f_2$,其中 $w_1+w_2=1$,w_1,w_2 可人为设定或通过模型学习获得。加权求和能够对特征空间进行加权表示，从而突显重要特征，同时压制冗余或不相关的信息。

加权求和方法高效且易于实现，尤其在资源有限的场景中表现出色。通过调整权重，控制各特征来源对最终结果的影响，增强模型对主要特征的关注。同时，这种方法能够减小特征融合后可能出现的维度爆炸问题，适合特征空间相似的任务，如时间序列预测或多层特征融合。

2.2.4 点乘融合

点乘融合通过对特征向量进行逐元素相乘，实现两组特征的交互表示，其过程如图 2-2(d)所示。假设 f_1，$f_2 \in \mathbb{R}^d$,融合特征为 $f_{mul}=f_1 \odot f_2$,其中⊙表示逐元素乘法。点乘融合可以保留特征间的细粒度交互信息，有助于提取高阶关系。

点乘融合能够捕获特征之间的相关性，帮助模型学习到更复杂的关系模式，适合处理需要理解特征交互的任务，如多模态学习和关系建模。相比简单的加法或拼接，点乘更注重特征的组合效果，使模型关注相似性或互补性信息，同时降低计算复杂度，适合高效训练。

2.2.5 卷积融合

卷积融合将多组特征视为多通道输入，利用卷积操作提取局部交互信息其过程如图 2-2(e)所示。假设输入特征为 $F_1,F_2 \in \mathbb{R}^{H\times W\times C}$,融合特征为 Fconv=Conv([F1·F2]),其中[·]表示通道维度拼接，Conv 为卷积操作。卷积融合擅长捕获空间结构和局部特征，适合处理图像数据或需要保留空间信息的任务。相比简单的加法或拼接，卷积能够提取更加细粒度的特征交互，同时利用权值共享降低计算复杂度，是视觉任务中特征融合的主流方法之一。

2.2.6 注意力机制加权融合

注意力加权融合通过动态分配权重来突出不同特征的重要性，生成加权融合特征，其过程如图 2-2（f）所示。假设特征矩阵为 F=[f_1,f_2 ,⋯,fn],权重向量为 a=Softmax(WF),则融合特征为 $f_{attn}=\sum_{i=1}^{n} a_i f_i$。注意力机制使模型能够根据任务需求动态调整权重，强调对当前任务最有用的特征，同时抑制冗余或噪声特征。

注意力机制的灵活性使其在多模态融合或多层特征融合中表现出色。通过关注关键特征来源，注意力加权融合能够提高模型的有效性和泛化能力。此外，它还能帮助模型在不同任务中自动适应特征的重要性分布，以适应特征重要性不均或信息冗余的复杂场景。

2.3 残差神经网络

神经网络（Neural Network , NN）是一种模拟大脑神经元连接方式的计算模型，由输入层、隐藏层和输出层组成，每一层由多个神经元构成，其结构图如图 2-3 所示。每个神经元执行加权和计算，并通过非线性激活函数（如 ReLU、Sigmoid）产生输出。神经网络的核心公式为：

$$Y=f(Wx+b), \tag{2-10}$$

式（2-10）中，x 是输入向量，W 是权重矩阵，b 是偏置项，f(·)是激活函数。神经网络能够通过训练自动学习数据的复杂非线性关系，适合处理诸如图像、语音、文本等复杂数据。随后通过梯度下降等优化方法，不断调整网络的参数，从而提升模型的预测能力。然而，面对高维数据时（如图像、语音等）的处理效果表现不尽人意。

随着卷积神经网络的不断发展，其以卷积运算为基础的操作方法，逐步取代了传统依赖全连接层构建的神经网络。这一变化在显著的减少了神经网络的参数量，并有效节省计算时间。卷积神经网络主要由卷积层、池化层和全连接层三部分组成。将这三个部分组合构成的网络，能够以较简单的方法拟合复杂数据。

卷积神经网络的学习过程包括两个主要步骤：首先对输入图像进行特征提取，然后将特征向量输入全连接层进行物体分类预测的前向传播。接着，比较预测结果与标签向量的差异，并通过梯度下降进行反馈优化网络层的参数，逐渐向标签结果对齐。一般而言，使用多个卷积核卷积操作，可以提取更多的特征。换句话说，使用卷积层越多，模型学习到的图像特征也越丰富。但随着网络层数的增加，导致模型训练困难，容易出现梯度消失和退化问题。

为了解决神经网络中随着层数增加带来的梯度消失、梯度爆炸等问题，残差神经网络（Residual Network，ResNet）随即被提出，其结构示意图如图 2-4。以梯度消失问题为例，在反向传播中，每次传递梯度信息时都需要将前一层的梯度与小于 1 的误差梯度系数相乘。这一操作会导致层数较深的网络中，梯度信息无法有效更新。而 ResNet 通过跳跃连接强制梯度直接流回浅层，打破了传统网络的梯度传播瓶颈。其数学本质是为反向传播提供了复合路径，确保梯度不会因链式法则的乘积效应而消失。

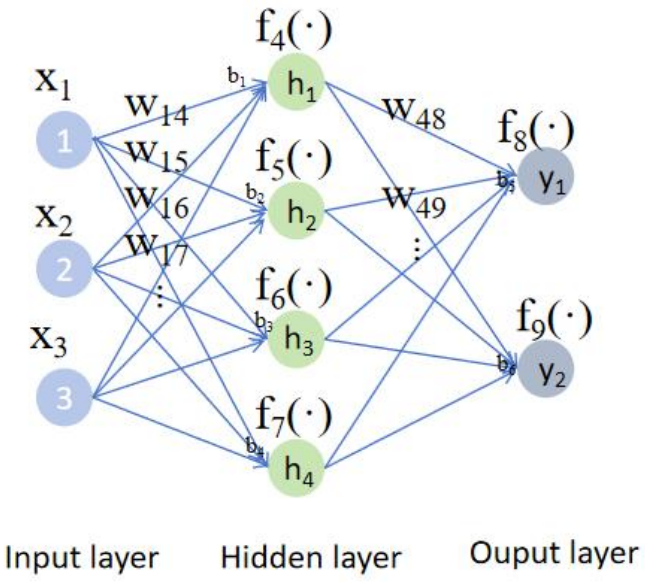

图 2-3 简单神经网络示意图

Figure 2-3 Schematic diagram of a simple neural network

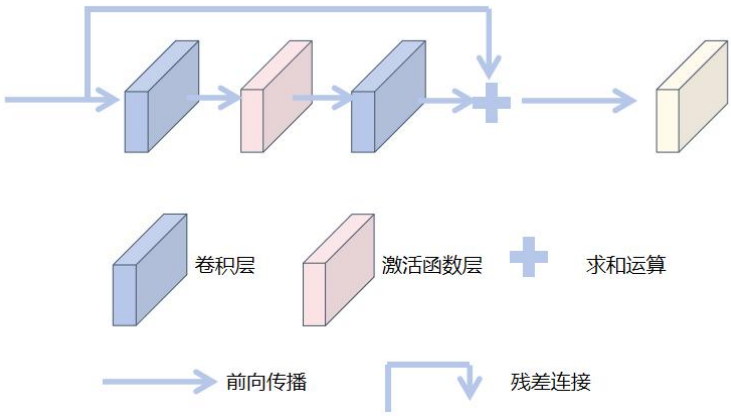

图 2-4 残差网络结构示意图

Figure 2-4 Schematic diagram of the residual network structure

残差网络结构主要由卷积层、归一化层构成的权重层组成。相应的处理过程为：首先将输入向量 x 经过卷积层和激活函数层运算之后输出 $\mathrm{F}(x)$。随后，将原始输入向量 x 与 $F(x)$相加，得到最终的输出结果 $\mathrm{H}(x)$。 $\mathrm{H}(x)$生成过程可由公示表示为：

$$\mathrm{H}(x) = F(x) + x \tag{2-11}$$

2.4 注意力机制

注意力机制的核心在于使网络更加专注于那些需要关注的区域。在注意力机制中，通常可以分为通道注意力机制和空间注意力机制两大类。空间注意力机制的作用

是在多维特征图中，针对包含多种物体的像素点（鸟和汽车等）进行关注。而通道注意力机制的原理是根据各通道的权重差异来判断哪个通道更重要，从而给予该通道更多关注。

通道注意力机制中比较典型的方法是 SENet。SENet 主要关注输入特征图中各通道的权重。其工作原理是首先是对输入特征图进行全局平均池化，接着进行两次全连接计算，其中第一次神经元数量较少，而第二次神经元数量与输入特征图的通道数相同。然后，使用 Sigmoid 函数为每个通道分配一个介于 0 和 1 之间的权重值。最后，将得到的通道权重乘以原输入特征图以评估每个通道中特征的重要性。另一种是以空间注意力机制为主的混合方法：CBAM。传统 CNN 通常依赖卷积核自动学习特征，但这种方法难以显式地关注哪些通道或空间区域的特征对 FGIC 任务更重要。CBAM 通过引入双路径注意力机制，动态地为每个特征图分配权重，使网络能够在特征图中定位哪些空间信息更重要，同时学习到特征通道中重要的信息的位置。

2.5 Vision Transformer

在计算机视觉领域中，卷积神经网络一直处于视觉任务中的主流主干网络结构，直到 Guo 等[53]提出的 ViT 在计算机视觉领域出现。ViT 的核心是多头自注意机制，该机制通过计算输入补丁之间的相互关系，从而捕获全面的对象特征表示。ViT 的内部结构如图 2-5 所示。ViT 首先将数据集中的图像切分为大小均匀的图像块，然后将这些图像块做线性变换并展平为 token 后添加位置编码。接着，将 token 输入到 Transformer 编码器，学习它们之间的自注意力特征。最后将学习到的注意力特征输入到分类头进行分类。

Transformer 编码器的组成结构如图 2-5 虚线右侧所示。其中，ViT 处理图像的步骤可简要描述为：

1.图像预处理：

首先将输入图像被分割成多个大小为 $P\times P$ 的图像块。随后，每个块被展平（Flatten），然后通过一个线性层（Embedding Layer）映射到一个固定的维度 D。这个过程相当于将图像转化为一个序列，类似于自然语言处理（Natural Language Processing，NLP）中将文本转化为单词序列。例如，如果原始图像大小是 H×W×C,

它将被分成 N= $\frac{H}{P} \times \frac{W}{P}$个小块。每个图像块被展平成一个 $P^2 \times C$ 的向量，并通过一个线性映射变换为 D 维的嵌入向量。

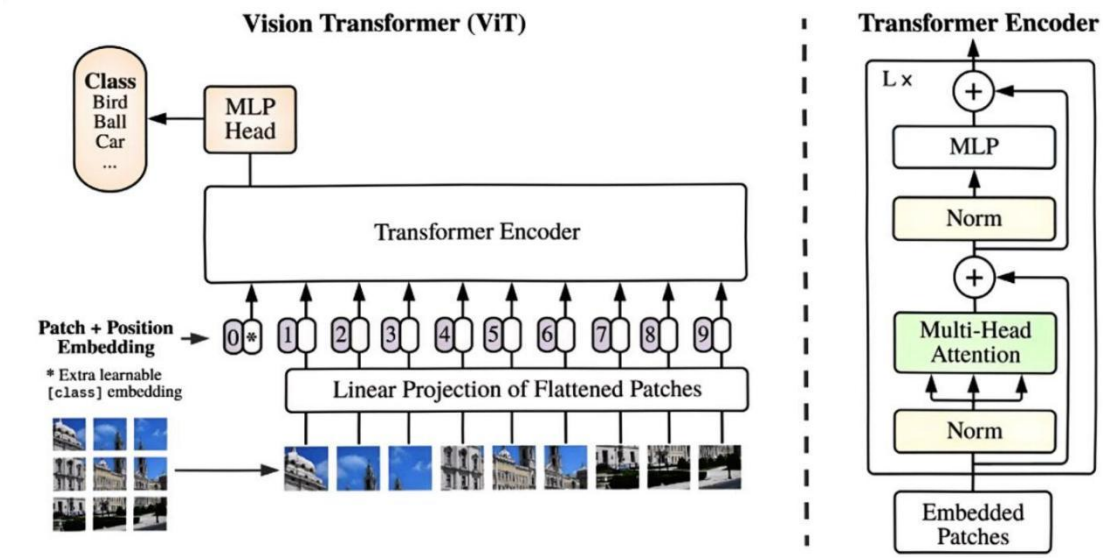

图 2-5　ViT 网络结构图

Figure 2-5 Structural diagram of the ViT network

2. 类别标记：

ViT 在输入序列中加入一个特殊的“类别标记”（Class Token）。这个类别标记在通过 Transformer 编码器的多层处理后，最终会代表图像的整体特征。在分类任务中，最终的类别预测通常来自这个类别标记的输出。类别标记通常是一个 D 维的向量（与其他图像块的嵌入向量维度一致），它被加到每个图像块的序列中，并在 Transformer 编码器的每一层中一起处理。

3. Transformer 编码器：

ViT 中的编码器由多个自注意力层（Self-Attention）和前馈神经网络（Feed-Forward Networks）堆叠组成。每一层的结构包含：

（1）归一化层

每个子层（自注意力机制、多层感知机）后都跟有层归一化（Layer Normalization）操作。层归一化的作用是对每个样本的特征进行标准化，减轻训练过程中内部协变量偏移的问题，从而加速模型的训练和提高稳定性。对于每个输入 x，层归一化的计算过程为：

$$X=\frac{x-\mu}{\sigma+\varepsilon}\cdot\eta+\beta, \quad (2\text{-}12)$$

式（2-12）中，$\boldsymbol{\mu}$和σ是输入 X 的均值和标准差，η和β是学习的缩放和偏移参数，$\boldsymbol{\varepsilon}$是一个数值较小的常数，避免除零错误。这样做的目的是使得每一层的输出具有均值为 0，方差为 1 的分布，从而提高模型的训练速度和稳定性。

（2）多头自注意力机制

自注意力机制是 Transformer 模型的核心，它能够计算输入序列中每个元素（图像块）之间的关系，从而捕捉到全局的信息。为了更好地学习不同的特征表示，ViT 使用了多头注意力机制（Multi-Head Attention）。多头注意力机制的公式： 设输入为一组向量 X，通过线性变换得到查询（Query）、键（Key）和值（Value）：

$$Q=XW_Q\,,K=XW_K\,,V=XWv, \quad (2\text{-}13)$$

式（2-13）中，W_Q，W_K，Wv 是待学习的权重矩阵。接着，计算注意力权重:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2\text{-}14)$$

式（2-14）中，d_k 是键向量的维度。通过自注意力，ViT 能够为每个图像块分配一个全局的权重，捕捉图像中的长程依赖关系。

在多头注意力中，多个不同的查询、键、值组合（即多个注意力头）同时计算，最终将所有头的输出拼接起来，经过线性变换得到最终的输出。这种机制能够捕捉到不同子空间的信息，从而增强模型的表达能力。

（3）多层感知机

在自注意力机制之后，每个编码器层都包含一个前馈神经网络。多层感知机（Multilayer Perceptron，MLP）通常由两个全连接层组成，带有 ReLU 激活函数，用于进一步非线性变换特征。

$$MLP(z)=ReLU(xW_1+b_1)W_2+b_2 \quad (2\text{-}15)$$

式（2-15）中，W_1, W_2 是权重矩阵，b_1,b_2是偏置项，ReLU 是激活函数。MLP 模块通常带有一个 Dropout 层，目的是在训练过程中进行正则化，避免过拟合。

4.位置编码：

由于 Transformer 本身不具备处理序列中元素的位置信息，因此 ViT 对输入图像块的顺序引入了位置编码（Positional Encoding）。位置编码通过加到每个图像块的嵌入向量上，提供了图像块在原始图像中的位置信息。这是为了 Transformer 模型理解图像块之间的相对位置关系。

5.输出：

经过多层 Transformer 编码器处理后，最终的类别标记输出表示了整个图像的高维特征，通常用于分类任务中作为图像的表征。这个输出会经过一个全连接层进行分类。

2.6 本章小节

本章节首先介绍了传统图像分类中数据增强的相关理论；其次，详细介绍了特征融合的种类和原理；然后，介绍了残差神经网络和神经网络的注意力机制；最后介绍了 ViT 的内部建构以及处理图像的流程，为本文后续章节提出的图像分类算法研究奠定了理论基础。

第 3 章 基于多阶段特征提取的细粒度图像分类

3.1 概述

在目前的细粒度图像分类技术中，存在着诸如光线、遮掩、环境等多种因素的干扰导致识别区域容易受到背景干扰的问题，本文提出了基于多阶段特征提取的细粒度图像分类方法（Adjustable Localization and Multi-Angle Attention，ALMA）。该模块由两个分支组成，即可调整定位模块（Adjustable Localization , AL）和多角度注意模块（Multi-Angle Attention , MA）。具体来说，在可调整定位模块中，首先定位目标感兴趣的区域，并通过背景遮掩（Bmask）对感兴趣区域进行调整，得到调整后的裁剪区域。然后，对调整后的区域进行聚集，定位潜在的鉴别性特征。此外，对于多角度注意模块，逐步最大化原始注意图与随机选择的注意图之间的差异。因此，该方法相较于其它方法具有更好的泛化性。

3.2 算法介绍

本文提出的 ALMA 模型整体结构如图 3-1 所示。第一阶段，通过骨干网络提取输入图像的特征，生成特征图。随后，通过 1×1 卷积函数获得注意图。在 AL 模块中，随机选择一个注意图来定位感兴趣的区域，如图 3-1 所示。在背景遮掩模块的引导下，遮掩背景区域，以提示模型专注于对象本身。最终，裁剪区域被放大以捕捉更详细的特征。通过多次类似的操作来定位该对象。在 MA 模块中，引入随机注意图，将其与初始注意图进行比较，逐渐放大两者之间的差异。对两个阶段的初始结果和比较结果进行计算和平均，从而得出结果。所提出的模型算法流程如下图 3-2 所示：

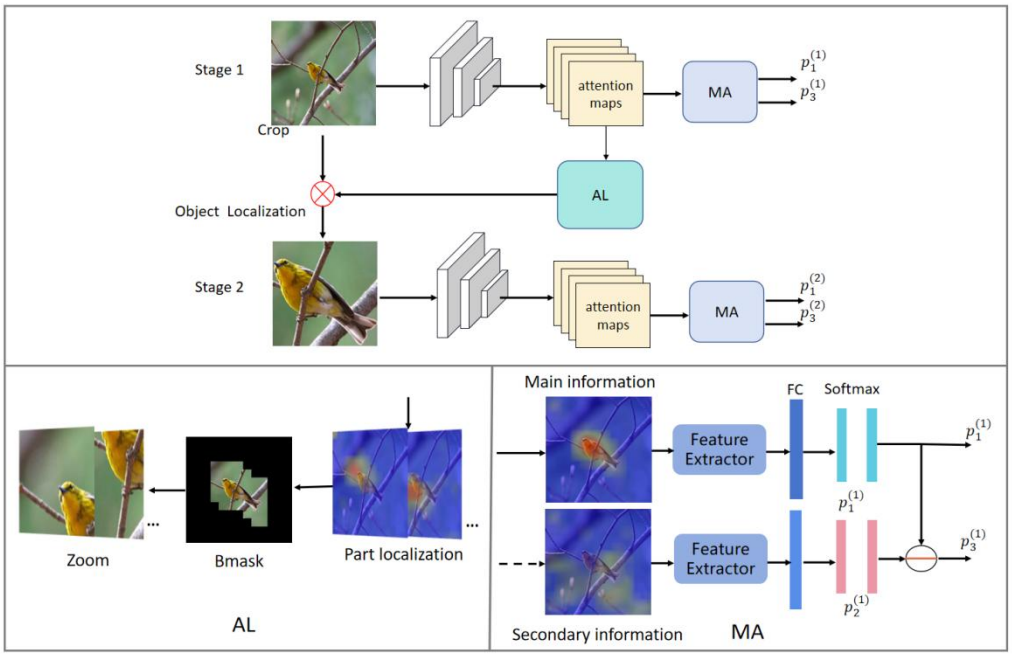

图 3-1 ALMA 算法框架图

Figure 3-1 Framework diagram of ALMA algorithm

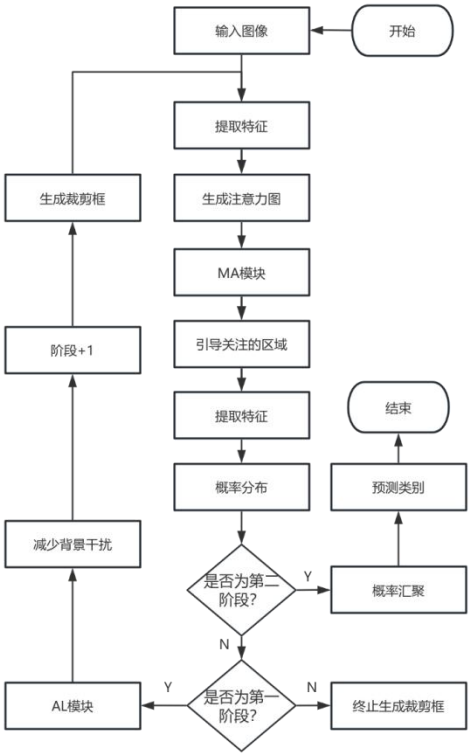

图 3-2　ALMA 算法流程图

Figure 3-2 Flowchart of ALMA algorithm

3.3 模块介绍

3.3.1 可调整定位模块

首先用 Resnet 网络提取给定图像的特征，得到特征图 F，其大小为 H × W × C，其中 H、W 和 C 分别表示特征图的高度、宽度和通道数。注意图 A 由之前的特征图 F 获得：

$$A = f(\mathrm{F}) = \bigcup_{k=1}^{M} A_k, \tag{3-1}$$

式（3-1）中 $f(\cdot)$是卷积函数，其过程可描述为：首先经过 1×1 卷积，随后进行批量归一化，最后输入 ReLu 激活函数。$A_k \in \mathbb{R}^{\mathrm{HxW}}$ 表示感兴趣的区域，M 表示注意图的数量。为了便于后续的逻辑判断，将随机选择的注意图归一化如下：

$$A_k^* = \frac{A_k - min(A_k)}{max(A_k) - min(A_k)}. \tag{3-2}$$

（1）局部定位

考虑到细粒度图像分类的挑战，获得更精确的图像区域信息是必要的。具体来说，是一个阈值θ_c，$\theta_c \in$（0，1）。如果A_k^*超过阈值，将被赋值为 1；否则，它将被赋值为 0。最后，将创建一个最小化的矩形框，以包含赋值为 1 的所有区域。

$$\mathrm{C_k} = \begin{cases} 0, & A^* > \theta_c; \\ 1, & \text{Otherwise.} \end{cases} \tag{3-3}$$

局部定位（Part Localization）区域可能包含大量的背景信息，特别是在小目标的数据集中，这可能会干扰模型定位。为了更好地集中在物体的区分区域，所提出的模型通过差异化区域掩盖背景区域。这使得模型能够更有效地关注对象本身。

（2）背景遮掩

如图 3-1 所示，注意图是随机选择的。随后，背景遮掩区域将基于小于预定阈值θ_m的值生成。裁剪区域和这遮掩部分之间的重叠将被设置为 0，进一步限制感兴趣的区域。

$$B_M = \begin{cases} 0, & A^* < \theta_m; \\ 1, & \text{Otherwise.} \end{cases} \tag{3-4}$$

定位区域将减少对遮挡区域的关注，使关注区域更加准确。随后，放大定位区域，可以更详细地观察目标的特征，以便捕捉更详细的特征。类似的操作重复多次。最后，

生成一个对象位置框，该框最小限度包含的所有局部定位框，用于第二阶段的预测。具体来说，模型中的 Bmask 模块仅用于训练。最后将定位的区域放大，捕捉更详细的特征。

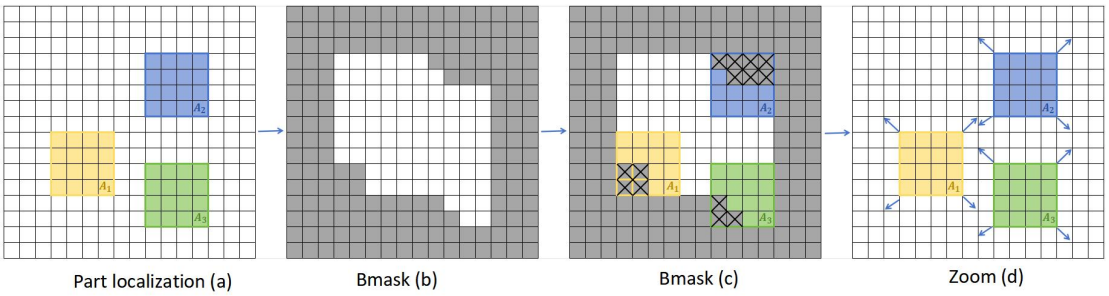

图 3-3 AL 模块处理过程

Figure 3-3 AL module processing process

可调整定位模块的处理过程如图 3-3 所示。(a)选择三个关注区域（A1, A2, A3）。(b)通过归一化 A^*生成掩蔽区域，该掩蔽区域由灰色区域表示。(c) A1，A2，A3 中的 x 表示再次被屏蔽的区域。(d)放大感兴趣的区域以捕捉更详细的特征。

3.3.2 多角度注意模块

由于大多数的研究没有指导模型如何区分主要和次要信息，因此在多角度注意模块中，从多个角度中在感兴趣的区域重新提取特征。

$$P_{\mathrm{j}}^{\mathrm{i}}(F*A)=P_2^*\left(\left((F*A_1),(F*A_2),\ldots,(F*A_M)\right)\right), \tag{3-5}$$

式（3-5）中 $P(\cdot)$表示预测函数，i 表示预测的阶段，j 表示第 i-th 阶段的预测数。M表示注意图的数量。首先，在 MA 模块中输出初始预测的概率，即当 j=1 时。然而，现有的大部分工作主要集中在如何提取检测到的特征，而往往忽略了次要信息的重要性。在输出P_1^i后，为了更好地识别次要信息，引入了随机注意图。概率P_2^i表示第二次预测概率，是通过次要信息得到的。逐渐扩大主要信息和次要信息之间的区别。该模型更好地理解了哪些是重要的信息，从而提高了图像预测的精度，得到了第三次预测概率，记为P_3^i。

$$P_{\mathrm{j}}^{\mathrm{i}}(F*A^*)=P_2^*\left((F*A_1^{\mathrm{r}}),(F*A_2^{\mathrm{r}}),\ldots,(F*A_M^{\mathrm{r}})\right), \tag{3-6}$$

式（3-6）中 A^{r} 表示随机注意图。然后从原始结果$\mathrm{P}_1^{\mathrm{i}}$中去掉随机关注$P_2^{\mathrm{i}}$，得到$P_3^{\mathrm{i}}$：

$$P_3^{\mathrm{i}}=P_1^{\mathrm{i}}-P_2^{\mathrm{i}}. \tag{3-7}$$

3.4 模型训练与推理

模型训练的两个阶段互相补充信息，从而提高了最终预测的能力。为了得到最终的预测概率，将每个 MA 模块产生的两个预测概率进行汇总和平均。此外，利用关注区域质量的角度来监督和指导注意力区域的过程。损失函数计算如下：

$$\zeta = \zeta_{ce}(Y_{final}, Y) + \zeta_{cl} \ , \qquad (3\text{-}8)$$

式（3-8）中 Y 表示分类标签，ζ_{ce}表示交叉熵函数，ζ_{cl}表示中心损失函数。ζ损失函数使模型减少对背景的聚焦，从而更好地集中在物体本身。这种方法有助于识别最具鉴别性的部分，以捕获详细的特征。

3.5 实验结果分析

3.5.1 数据集选取

所提出的模型在三个公开的 FGIC 数据集上进行了广泛的实验：Cub -200-2011[60]，NA-Birds[59],Aircraft[62]和 Stanford Cars[63]。其中这些数据集中的部分图像由图 3-4、图 3-5、图 3-6 和图 3-7 所示。

其中，Cub -200-2011 数据集包含 200 种鸟类和 11,788 张图像，其中包括 5,994 张训练图像和 5,794 张测试图像。Aircraft 数据集由 100 种飞机和 10,000 张图像组成，其中 2/3 用于训练，其余 1/3 用于测试模型。Stanford Cars 数据集包含 96 个汽车分类和 16,185 张图像，其中 8,144 张用于训练，8,041 张用于测试。NA-Birds 中 23929 张图像用于训练，24633 张图像用于测试，这些数据集的图像的划分比例见表 3-1。

3.5.2 实验细节

为了公平比较，本文提出的模型分别以 Inception-v3、Resnet50 和 Resnet101 作为骨干网。注意图的数量设置为 32。在实验中，所有数据集上使用相同的超参数。具体来说，训练轮数设置为 120，批处理大小为 16，图像大小为 448 × 448。权重衰减设置为 1e-5。以 1e-3 的学习率开始训练，随后每 2 epoch 将其降其 0.9 倍。为了更公平的比较，实验中与其他方法对比采用相同环境或者原论文的数值进行对比。

表 3-1　CUB、Aircraft、Cars 数据集的图像数据

Table 3-1 Image data for CUB, Aircraft, and Cars datasets

Dataset	Category	Training	Testing
CUB-200-2011	200	5994	5794
FGIC-Aircraft	100	6667	3333
Stanford Cars	196	8144	8041
NA-Birds	555	23929	24633

图 3-4 Cub -200-2011 数据集图像

Figure 3-4 Image of Cub-200-2011 dataset

图 3-5　FGIC Aircraft 数据集图像

Figure 3-5 Images of the FGIC Aircraft dataset

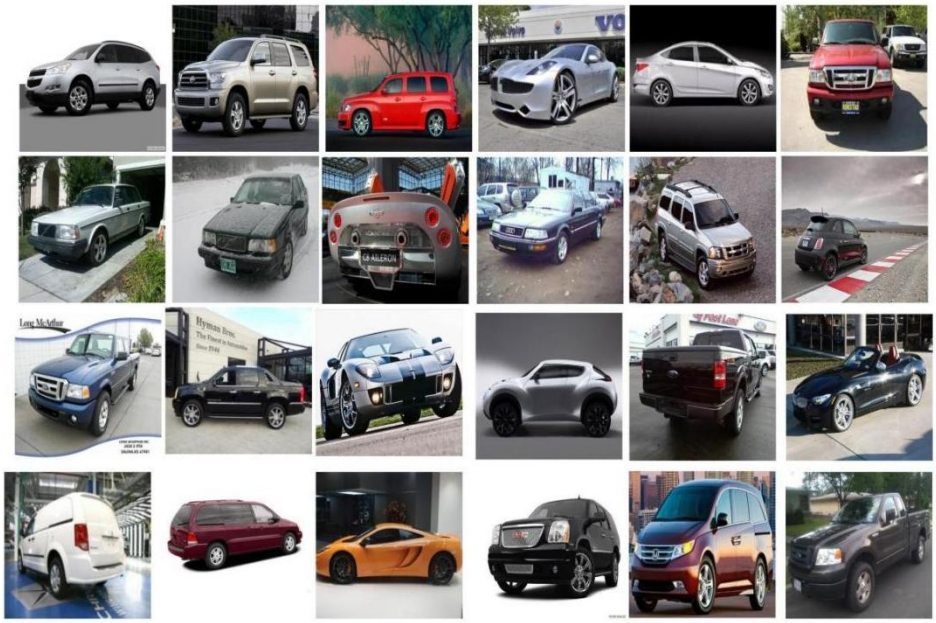

图 3-6 Stanford Cars 数据集图像

Figure 3-6 Images of the Stanford Cars dataset

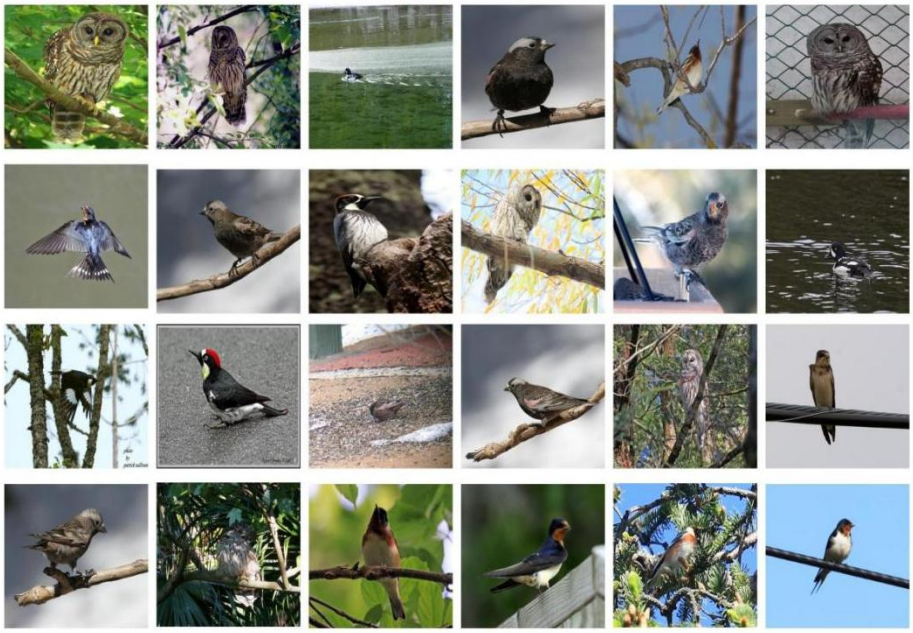

图 3-7 NA-Birds 数据集图像

Figure 3-7 Image of NA-Birds dataset

3.5.3 与现有的方法比较

为了评估所提出的模型的有效性，将其与其他一些最先进的 FGIC 方法进行了比较，对比结果见表 3-2。

与 CIN [41]相比，ALMA 在 Cub -200-2011 和 Aircraft 任务中分别提高了 2.2％和 1.3％。对比结果表明，所提出的 AL 模块和 MA 模块是有效的。CIN 在单个注意图上缺乏额外的约束，导致可能包含大量的背景信息。这进一步限制了模型的性能。在 AL 模块的引导下，ALMA 减少了对背景的聚焦。由于 WS-DAN[20]忽略了次要信息的重要性，因此正确识别这些区域将更好地提高模型的性能。相比之下，提出的 ALMA 较少关注次要信息，更多地关注目标本身。此外，在图 3-8 中说明了提取的注意图的一些可视化。可以看出，与 WS-DAN 相比，ALMA 可以更好地观测整个目标，获得更有效的信息。

另外，根据是否使用人为标注的边界框或额外注释来比较两种类型的基准，如表 3-3 所示。“Train Anno”表示在训练过程中使用边界框进行额外的标注。与基于手工注释的方法 PN-CNN[47]相比，ALMA 显示出 4.4％的改进。两阶段的互补性是取得较好效果的原因。这表明，对于细粒度图像，单一阶段的捕捉的特征有限，物体需通过多阶段的裁剪以捕获更详细的信息。

表 3-2　ALMA 与主流模型性能对比

Table 3-2 Performance comparison between ALMA and mainstream models

Methods	Backbone	Accuracy(%)		
		CUB	Aircraft	Cars
B-CNN [44]	VGG	84.1	84.1	90.6
RA-CNN [39]	VGG	85.4	88.4	92.5
AP-CNN [40]	VGG	86.7	92.9	94.6
CIN [41]	Resnet-50	87.5	92.6	94.1
Cross-x [42]	Resnet-50	87.7	92.7	94.6
PMG [44]	Resnet-50	88.6	93.1	94.3
ALMA	Resnet-50	**89.7**	**93.9**	**95.1**
SPS [45]	Inception-v3	87.0	92.0	93.5
WS-DAN[20]	Inception-v3	89.3	93.0	94.5
ALMA	Inception-v3	**89.8**	**94.2**	**94.8**
CIN [41]	Resnct-101	88.1	92.8	94.5
ELoPE [49]	Resnet-101	88.5	93.5	95.0
FBSD [46]	Resnet-101	88.8	93.1	95.0
ALMA	Resnet-101	**90.3**	**94.2**	**95.2**

表 3-3 基于多阶段的方法对比

Table 3-3 Comparison of methods based on multiple stages

Methods	**TrainAnno**	**Accuracy**
B-CNN [44]	√	85.1
PN-CNN [47]	√	85.4
B-CNN [44]		84.0
MGE-CNN [48]		88.5
WS-DAN[20]		89.3
Stage 1		87.8
Stage 1+2		89.8

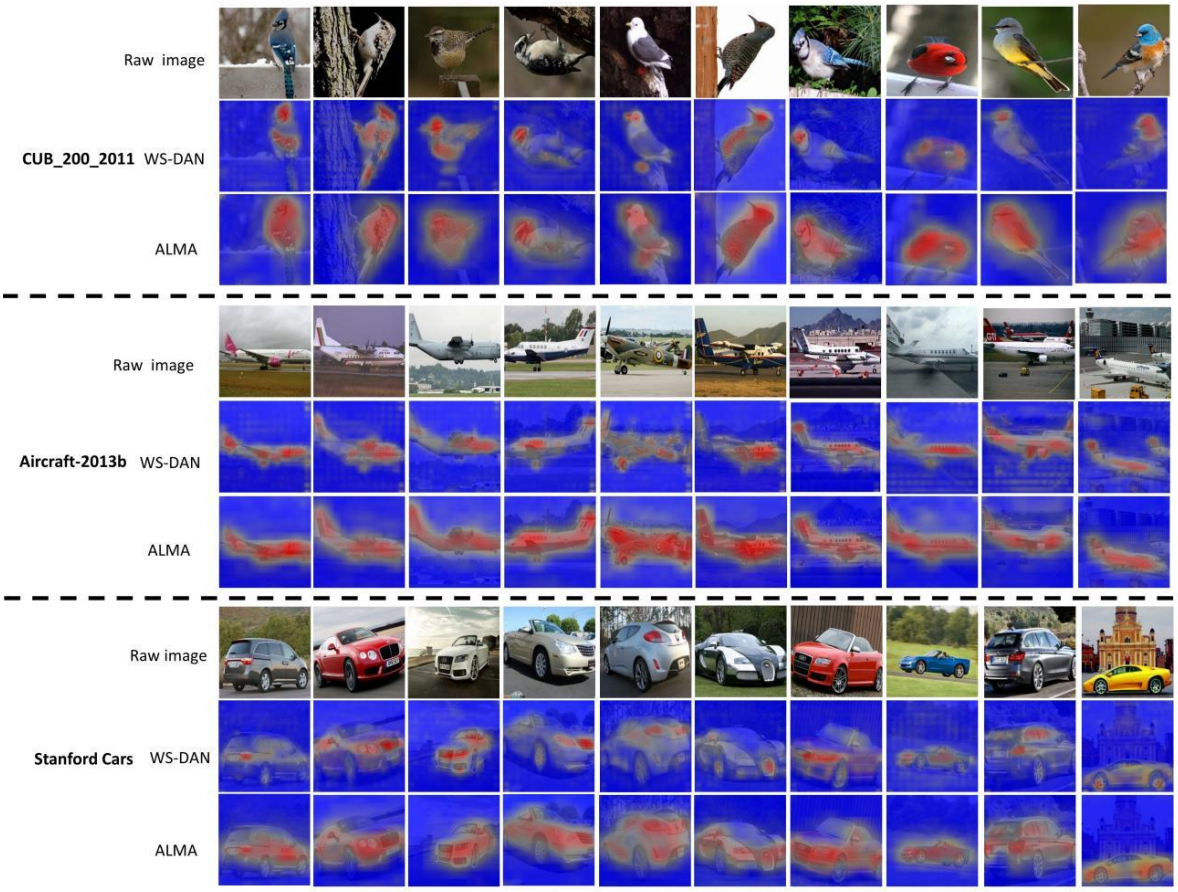

图 3-8 与主流方法的可视化对比

Figure 3-8 Visual comparison with mainstream methods

3.5.4 消融实验

为了评估每个模块的有效性，在不同的数据集上进行了消融实验。过多的关注背景信息会进一步限制细粒度分类的性能。为了减少对背景信息的关注，引入了 Bmask 模块。如表 3-4 所示，增加 Bmask 模块可使性能提高 1.3％。

另外，添加 MA 模块将有 1.35%的性能提升。这表明该模型能够有效区分物体信息和背景信息。随后，将 AL 模块与 MA 模块相结合，所提出的模型准确率达到 89.80%，表明了两个模块之间的有着较好的兼容性。

3.5.5 超参数分析

在该模型中，注意图是至关重要的超参数。这些注意力图专注于图像的特定部分，增加注意力图的数量可以帮助模型捕获更详细的信息。值得注意的是，这也可能提高模型的计算复杂性要求。为了确定最佳性能注意图数量，进行了更详细的实验。如表

3-5 所示，当注意图的数量为 32 时，模型的性能达到最佳，当注意图的数量增加到 64 时，而模型的性能开始下降，这进一步验证了之前的假设。

表 3-4 不同模块的消融实验

Table 3-4 Ablation experiments for different modules

Module				CUB	Aircraft	Cars
AL			MA			
Part Loc	Object Loc	Bmask				
√				87.73	91.15	92.83
√	√			88.26	92.03	93.36
√	√	√		89.56	93.97	94.68
√	√		√	89.61	93.28	94.52
√	√	√	√	**89.80**	**94.24**	**94.75**

表 3-5 不同注意力图数量的效果

Table 3-5 Effects of different numbers of attention maps

Number of Attention Maps	Accuracy
4	87.99
8	88.49
16	89.28
32	89.80
64	89.13

3.6 本章小结

为了解决识别区域容易受到背景干扰的问题的问题，本文提出了一种新的用于细粒度图像分类的 ALMA 模型。本文提出的 ALMA 模型通过引导目标的局部区域，从不同角度区分主次信息，使其能够更有效地观察整个目标，从而提高泛化性能。在三个 FGIC 数据集上进行大量实验验证了所提出的 ALMA 模型在各种 FGIC 任务上的优越性能。

第 4 章 基于树形分支选举的细粒度图像分类

4.1 概述

传统的多阶段的细粒度图像分类的方法虽然能够捕捉到重要的区域，然而却难以对这些区域的重要性做出客观的评估。针对模型难以辨别鉴别性特征所在阶段的问题，本文提出了基于树形分支选举的细粒度图像分类（Attention Binary Navigation Tree , ABNT）模型。首先，在初始阶段模型通过最大化与粗粒度信息的差异，能够将关注的区域引导到目标上，并为后续阶段提供更有效的指导。随后阶段以前一阶段为基准，对高响应区域进行裁剪，得到新的图像并逐渐发现鉴别性特征。通过树形结构集成的多个分支路由模块对决策树的子分支选举，不断增加鉴别性特征所在阶段的贡献。最后，汇总来自叶节点的预测以获得最终决策。实验结果表明，该方法相比于其它方法具有更好的表现。

4.2 算法介绍

本文提出 ABNT 模型有多个模块组成，如图 4-1 所示。该模型由六个模块构成：（a）骨干网（Backbone Network）、（b）分支路由（Branch Routing）、（c）注意模块（Attention Module）、（d）导航模块（Navigation Model）、（e）局部特征注意学习模块（Local Feature Attention Learning）和（f）标签预测模块（Label Prediction）。其中骨干网模块用于提取图像特征并获取特征图。特征图作为分支路由模块的输入，通过卷积池化和其他计算操作产生标量输出，以评估左右分支的重要性。注意模块将特征图作为其输入，根据 1×1 卷积生成注意图。导航模块生成随机注意图，以增强模型区分物体及其周围背景区域的能力。具体来说，在 L_2 中，存在三个输入：特征图、注意图和随机注意图。在池化、连接和其他计算步骤之后，最终生成特征矩阵。最后，每个叶节点根据归属于对应分支路由模块路径的权重累积，将这些值相加以得出最终结果。

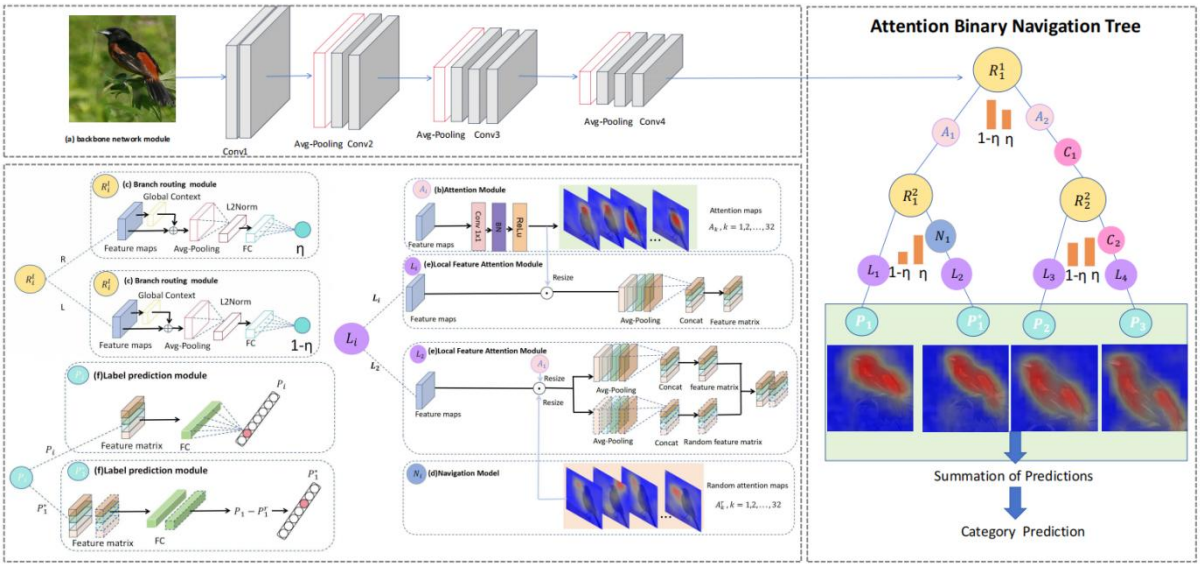

图 4-1 ABNT 框架图

Figure 4-1 ABNT framework diagram

另外，为了更好的说明 ABNT 模型处理流程，其算法处理过程如图 4-2 所示。

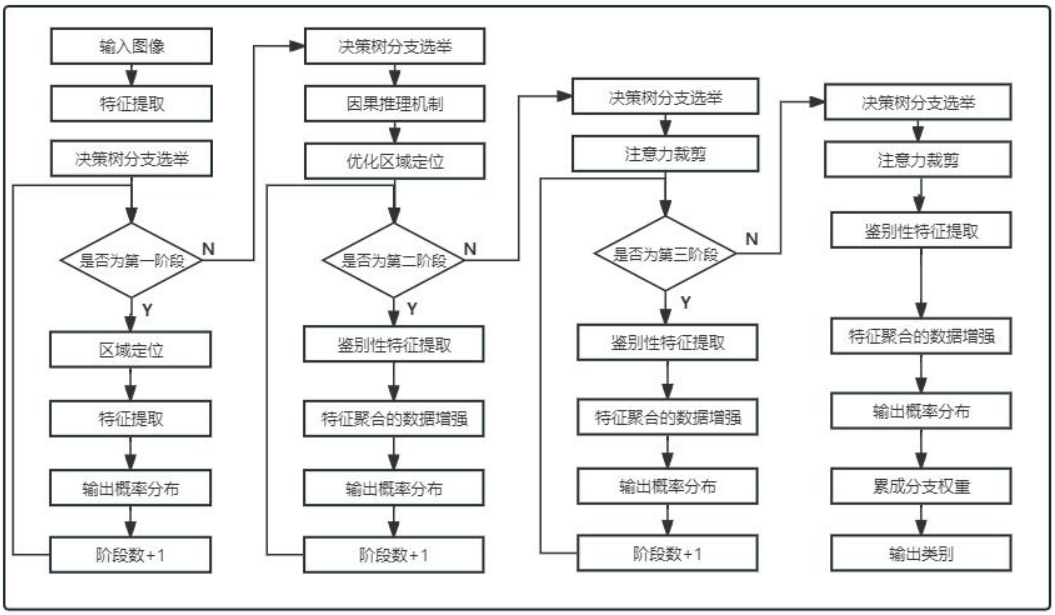

图 4-2　ABNT 的处理过程

Figure 4-2　The processing of ABNT

4.3 模块介绍

4.3.1 骨干网络

细粒度图像中的鉴别区域通常包含较小的鉴别区域，因此需要缩小特征提取的范围。这种减少是通过调整池化层和卷积核的大小和步长来实现的。具体来说，实验设

置中，以 ResNet 网络作为主干网络。与传统方法不同，实验中使用的图像大小为 448 × 448，而不是默认的 224 × 224。为了确保公平的比较，实验种以两种不同的骨干网络（ResNet-50 和 ResNet-101）进行了训练。

4.3.2 分支路由

分支路由模块输出的标量值表现出对左或右分支节点的偏好。如图 4-1 所示，在 l 层 i^{th} 节点 $R^l_i(\cdot)$的路由模块中，全局内容模块（Global Context）用于捕捉图像的全局特征。

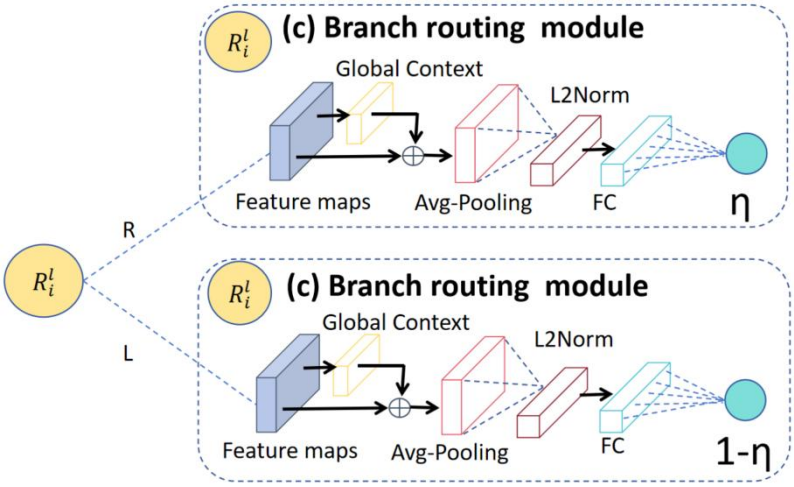

图 4-2 分支路由模块结构图

Figure 4-2 Branch routing module structure

这个轻量级的全局内容模块，从简化的非局部（Non-local，NL）和挤压激励（Squeeze-and-Excitation，SE）方法中提取特征，随后融合。这种融合机制促进了上下文信息的整合，减轻了对局部区域的过度依赖，增强了对象描述能力。分支路由模块的输出概率表示分配给各自子分支的权重。例如，分支路由输出概率记为η，表示对右子分支的重视程度，范围在[0,1]之间。相应地，指向左分支的注意力可以推导为 1-η。如果η超过 0.5，则表示更强调来自正确子分支的分类结果。相反，左侧分支支配最终的分类结果。

4.3.3 导航模块

注意力机制的提出进一步增强了模型感兴趣的区域，并在计算机视觉任务中被广泛应用。本文所提出的注意力模块核心步骤可由图 4-1 表示。然而，在训练样本不平衡时，模型易受到非鉴别区域的干扰。例如，训练样本中含有大量以蓝色的天空为背

景的图像，模型分类时可能也以背景区域作为判别的一部分，这不仅增大模型的计算量，还使得模型的分类性能难以突破。因此进一步提出导航模块，目标是通过深入研究变量之间的因果关系进行分析，从而超越主流方法的限制。如图，FGIC 的主流方法是训练特征图 X 到注意图 A: X—> A 的映射关系。相反，本文目标是构建新的关系，训练特征图 X 到随机注意图$\bar{A}$映射关系：X -->$\bar{A}$，迫使模型生成的注意图不仅仅由特征 X 决定。采用差异化的方法，结合补充的外部信息来建立与初始预测结果的因果关系。

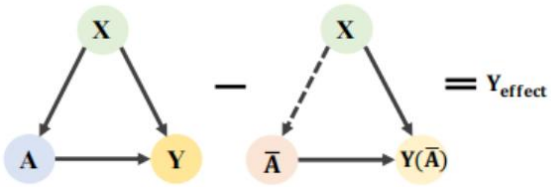

图 4-3 导航模块处理过程

Figure 4-3 Navigation module processing

具体来说，引入了随机注意图$\bar{A}$来对模型施加额外的约束。这些随机注意图是随机生成的，Y_{effect}表示两次预测分布的差异。通过对损失函数的不断优化，随机注意图与初始注意图在焦点区域上的差异会逐渐扩大。目标是随机注意图专门突出背景区域，而注意图专门捕捉对象本身的特征。

4.3.4 局部特征注意力学习

注意图 A 表示对象的各个部分，A_k表示对象的一部分，例如鸟的嘴。局部特征注意学习集成了注意图和特征图学习过程。具体来说，将注意力映射A_k乘以特征图 F，得到部分特征图F_k：

$$F_k = A_k \odot F(k = 1,2, ..., M), \quad (4\text{-}1)$$

式（4-1）中，M表示注意图的个数，$\odot$表示点积算子。采用特征提取函数 e(·)，进行池化、卷积等运算，最终得到局部特征表示F_k:

$$f_k = e(F_k), \quad (4\text{-}2)$$

设$v(\cdot)$表示注意图 A 和特征图 F 之间的整体局部特征注意学习过程。对象的特征是通过将 M 个特征提取函数 e(·)的数量叠加在一起来表示的，每个特征提取函数都包含丰富的局部特征。特征矩阵的表示过程如下：

$$v(A,F)=\begin{pmatrix}e(A_1\odot F)\\ e(A_2\odot F)\\ \cdots\\ e(A_3\odot F)\end{pmatrix}=\begin{pmatrix}f_1\\ f_2\\ \cdots\\ f_M\end{pmatrix},\qquad(4\text{-}3)$$

与之前的方法类似，随机局部特征通过 M 特征提取函数 e(·)叠加表示，每个特征提取函数包含丰富的随机局部特征，如下所示：

$$v(\mathrm{A^r},F)=\begin{pmatrix}e\left(A_1^r\odot F\right)\\ e\left(A_2^r\odot F\right)\\ \cdots\\ e\left(A_3^r\odot F\right)\end{pmatrix}=\begin{pmatrix}f_1^r\\ f_2^r\\ \cdots\\ f_M^r\end{pmatrix},\qquad(4\text{-}4)$$

式（4-4）中，A^r表示随机注意图。

4.3.5 标签预测

对于树结构 T 的每个叶节点，标签预测模块 P_i用于预测其所属的相应子类别。对于第二个叶节点的预测结果，从随机注意图中提取背景区域特征，生成虚拟概率。然后，将虚拟概率预测与第一个叶节点概率预测相减，得到调整后的概率分布：

$$P_1^*=P_1\left(v(A,F)\right)-P_1^{\mathrm{r}}\left(v(A^*,F)\right),\qquad(4\text{-}5)$$

式（4-5）中P_1^*为调整后的概率分布，$P_1(\cdot)$为预测概率函数，$P_1^{\mathrm{r}}(\cdot)$为虚拟预测概率函数。每个叶节点倾向于观察不同的目标识别区域，从不同的角度提取特征。叶节点的预测以输入全连接层的局部特征矩阵作为预测概率。将叶节点预测的概率与路径权值相乘，得到叶节点的最终概率。

4.3.6 模型训练与推理

模型的损失函数由叶节点预测的损失、最终预测的损失和特征中心的损失三部分组成。损失函数计算如下：

$$\zeta=\zeta_{\mathrm{ce}}\left(Y_{final},Y\right)+\sum_{i=1}^{2^{\mathrm{h}-1}}\zeta_{\mathrm{l}}(Y_{\mathrm{i}},Y)+\zeta_{\mathrm{cl}}\qquad(4\text{-}6)$$

式（4-6）Y 表示分类标签，ζ_{ce}表示交叉熵函数，ζ_l表示叶节点交叉熵函数，ζ_{cl}表示中心损失函数。通过这个损失函数，目的是增强第一个叶节点区分背景和对象的能力。这种增强可以更好地关注对象本身，并为后续叶节点提供更可靠的指导。

4.4 实验结果分析

4.4.1 数据集选取

在两个公共 FGIC 数据集上进行了广泛的实验：Cub-200-2011 和 Stanford-Cars 数据集。两个数据集的划分比例详见 3.5.1 节。

4.4.2 实验细节

为了进行公平的比较，提出的 ABNT 模型分别采用 Resnet-50 和 Resnet-101 作为骨干网。在训练中，并使用水平翻转来提高表现。在所有数据集中，实验中使用了统一的超参数。值得注意的是，epoch 数固定为 120，同时使用 16 个批处理大小并保持 448×448 的图像大小。权重衰减配置在 1e-5。训练开始时，学习率为 1e-3，每 2 次学习率递减 0.9 倍。为了更公平的比较，实验中与其他方法对比采用相同环境或者原论文的数值进行对比。

4.4.3 与现有的方法比较

（1）基于主干网络对比

为了评估所提出的 ABNT 模型的有效性，对几种最先进的 FGIC 方法进行了比较分析，详见表 4-1。提出的 ABNT 模型在性能上有显著提高。具体而言，表 4-1 的第三列显示了 CUB-200-2011 的比较结果。MGE-CNN [48]采用 KL 散度来促进多个专家生成多样化的结果，同时使用门控网络为每个专家分配权重。多阶段的图像裁剪方法从不同角度提取图像特征，显著提高了模型的性能。与 MGE-CNN[48]相比，ABNT 在初始阶段对感兴趣区域进行了优化，ABNT 在 Resnet-50 和 Resnet-101 上分别提高了 1.2％和 0.9％。表中的第四列描述了 Stanford Cars 数据集上的结果。与最先进的 FBSD[46]相比，ANBT 在 Cars 数据集上的改进有限。我们推测，这可能是由于汽车具有更均匀和更大的体型。尽管如此，与 FBSD 相比，ABNT 表现出 0.7%的改进。ABNT 区分对象和背景信息的识别能力超过了现有的基于注意力的方法。这是由于 ABNT 中的导航模块，它显著减少了模型对背景区域的关注，从而放大了对目标本身的关注。

（2）基于多阶段对比

根据两种基线类型对人类定义的边界框或部分注释的使用情况并设置两种基准，详见表 4-2。“Train Anno”表示在训练阶段使用边界框或部分注释。值得注意的是，与训练使用手工标注的方法相比，ABNT 显示了 4.3％的改进。与多阶段分类方法相比，ABNT 达到了最先进的性能。

表 4-1 与其他方法对比

Table 4-1 Comparison with other methods

Methods	Backbone	Accuracy(%)	
		CUB	Cars
B-CNN [38]	VGG	84.1	90.6
RA-CNN [45]	VGG	85.4	92.5
AP-CNN [40]	VGG	86.7	94.6
ELoPE [49]	Resnet-50	87.4	94.5
CIN [41]	Resnet-50	87.5	94.1
Cross-x [42]	Resnet-50	87.7	94.6
AC-NET [50]	Resnet-50	88.1	94.6
MGE-CNN[48]	Resnet-50	88.5	94.0
FBSD [43]	Resnet-50	88.5	94.4
Ours	Resnet-50	89.7	95.2
CIN [41]	Resnet-101	88.1	94.5
ELoPE [37]	Resnet-101	88.5	95.0
MGE-CNN[48]	Resnet-101	89.4	93.6
FBSD [43]	Resnet-101	88.8	95.0
Ours	Resnet-101	90.5	95.4

值得注意的是，当 ABNT 使用叶节点（leaf(1)）时，分支路由模块和导航模块的分类性能降低，导致 ABNT 回归到传统 CNN 分类方法的范围内。此外，当存在两个叶节点（leaf(1+1*)）时，导航模块纠正第一个叶节点所关注的区域。这种区分的思想使模型具有区分背景区域和物体本身的能力，增强了对物体的关注。裁剪方法迫使模型优先考虑有区别的区域。然而，如图 4-4 所示，第三个叶节点并没有强调鉴别区域（leaf(1+1*+2)）。为了解决这个问题，加入了第四个叶节点。可以观察到，第四个叶节点（leaf(1+1*+2+3)）高效的捕获了图中的鉴别区域特征。重要的是，当叶节点的数量达到 4 个时，ABNT 将转换为全二叉树。受益于多个分支路由模块的影响，这进一

步提高了 ABNT 的性能。多个分支路由模块的调整确保在四个叶节点之间更公平地分配贡献。

表 4-2 基于多阶段方法对比

Table 4-2 Comparison of the multistage methods

Methods	Train Anno	Accuracy
B-CNN [38]	√	85.1
PN-CNN [47]	√	85.4
B-CNN [38]		84.0
MGE-CNN[48]		88.5
WS-DAN[20]		89.3
Leaf (1)		87.4
Leaf(1+1*)		88.9
Leaf(1+1*+2)		89.5
Leaf (1+1*+2+3)		89.7

4.4.4 超参数分析

注意图构成了所提出模型中的关键超参数，将引导模型到图像中的特定区域。一般来说，注意图的数量越多，最终的表现就越好。增加它们的数量可以帮助模型捕捉更复杂的细节。尽管如此，这种增强可能会增加模型的计算需求。为了确定最佳性能的注意图数量，进行了实验分析。如图 4-4 所示，该模型在使用 32 张注意力图时达到了最高性能。值得注意的是，当数量增加到 64 时，性能开始下降。这一实验证实了本文的假设。

4.4.5 可视化实验

在图 4-5 中展示了叶节点的可视化。从每个数据集中随机抽取三幅图像。在图 4-5 中，第一列和第三列展示了原始图像，而第二列和第四列提供了这些原始图像的注意力图可视化。从图 4-5 中可以明显看出，与 MGE-CNN 相比，ABNT 可以更有效地观察到物体。在主流方法中，只关注对象往往包含大量的背景信息（第一行）。这种过

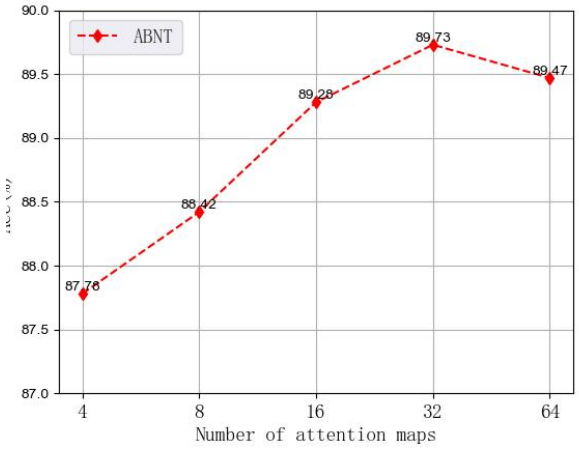

图 4-4 选择不同注意力图数量的效果

Figure 4-4 Effect of selecting different number of attention maps

多的对目标的强调会增加多阶段分类任务的计算需求，并可能误导后续阶段的观察区域。相反，当初始阶段强调的区域太少（第六行）时，后续阶段可能只提取到有限的特征，特别是在鉴别性特征。

另外，在第六列和第八列中，模型倾向于优先识别对象内的鉴别区域。与传统的粗粒度图像分类不同，关注这些鉴别区域可以显著提高模型在细粒度图像分类中的性能。图像中红色强度表示该区域的重要性。在图 4-5 中，ABNT 捕获了物体的最关键区域，例如鸟的头、翅膀和尾巴，以及汽车的门。这些可视化结果验证了所提出方法的可解释性。

4.5 本章小结

传统的多阶段的细粒度图像分类的方法虽然能够捕捉到重要的区域，然而却难以对这些区域的重要性做出客观的评估。针对模型难以辨别鉴别性特征所在阶段的问题，本文提出了一种用于细粒度图像分类的 ABNT 模型。提出的 ABNT 模型通过引导，导航模块从不同角度区分目标和背景信息，使其能够更有效地观察整个目标，从而提高泛化性能。大量的实验证明了所提出的 ABNT 模型在多个 FGIC 数据集上的优越性能。在未来的工作中，将考虑对树结构进行修剪，提高模型的训练速度。

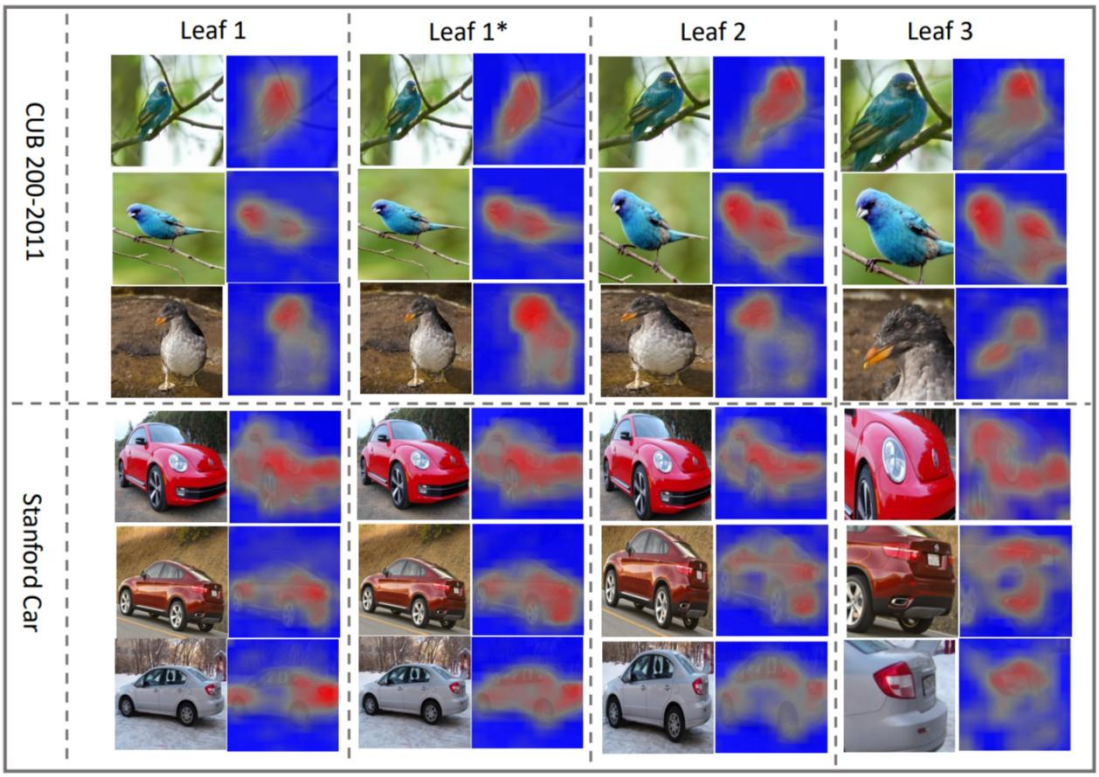

图 4-5ABNT 叶子节点的可视化

Figure 4-5 Visualization of ABNT leaf nodes

第 5 章 基于多尺度特征提取的 Transformer 细粒度图像分类

5.1 概述

主流的 FGIC 模型虽然能够捕捉高响应区域，但这些区域同样也包含着粗粒度信息。针对高响应区域中包含粗粒度信息问题，同时也为了发现更多的细粒度信息，本文提出了基于多尺度特征提取的 Transformer 细粒度图像分类模型（Filter Aggregation and Low-Temperature Freezing , FALT）。具体而言，该方法包含过滤聚合模块（Filter Aggregation, FA）和低温冻结模块(Low-Temperature Freezing, LT)，以选择有效的高响应区域并学习更丰富的鉴别特征。一方面，为了选择有效的高响应区域，过滤聚合模块首先根据特征图之间的相关性过滤掉无用的低响应区域。然后，从真正的高响应区域聚合特征进一步扩大了差异。另一方面，所提出的低温冻结模块能够学习更丰富的鉴别特征。可变化的温度函数鼓励模型探索更多的鉴别性区域并将弱相关特征冻结在不同尺度的特征图中，提取合适的鉴别特征。实验结果表明，该方法能够去除粗粒度信息的同时，发现潜在的鉴别性区域，以提高模型的分类能力。

5.2 算法介绍

在本文中，提出了 FALT 模型来选择真正的高响应区域，并学习更丰富的鉴别特征。FALT 模型的总体结构如图 5-1 所示，过滤聚合模块模块和低温冻结模块模块更详细的结构如图 5-2 所示。模型包括骨干网、路径聚合 （Path Aggregation，PA）模块、过滤聚合模块和低温冻结模块。骨干网络可以 Swin-t 和 ResNet 来构建。多尺度特征分别集成了自顶向下和自底向上两个方向的特征。过滤聚合模块利用特征图的空间关系，选择真实的高响应区域作为目标特征区域，放弃的高响应区域作为噪声区域。随后，对真实高响应区域的特征进行聚合，进一步扩大目标特征区域与噪声区域的差异。低温冻结模块对不同特征尺度的特征图中的弱相关特征进行冻结，检测出合适的鉴别特征。最后，结合层将高响应区特征和鉴别特征融合到分类器中，得到最终结果。

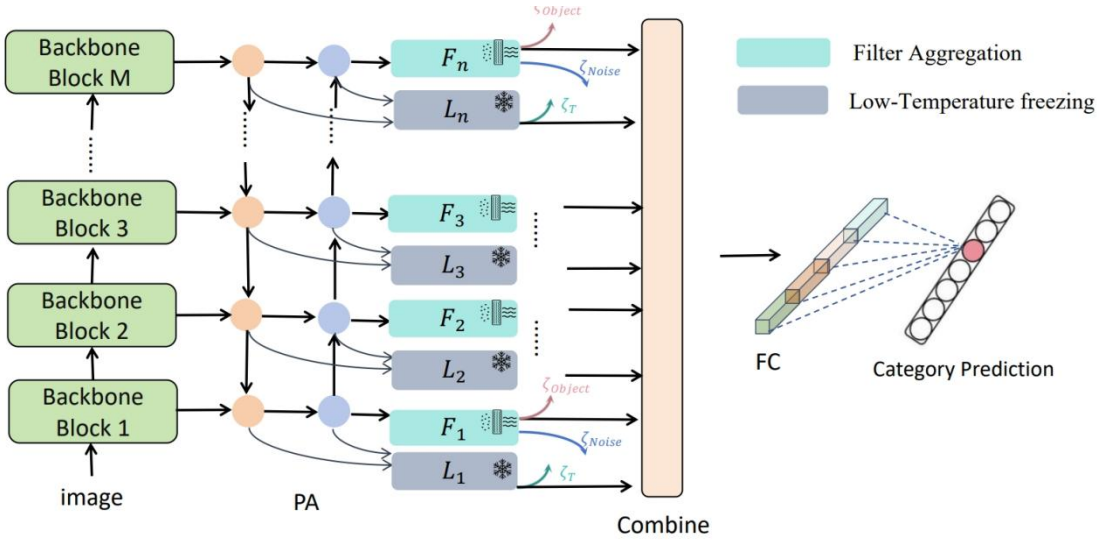

图 5-1 FALT 算法框架图

Figure 5-1 Framework diagram of FALT algorithm

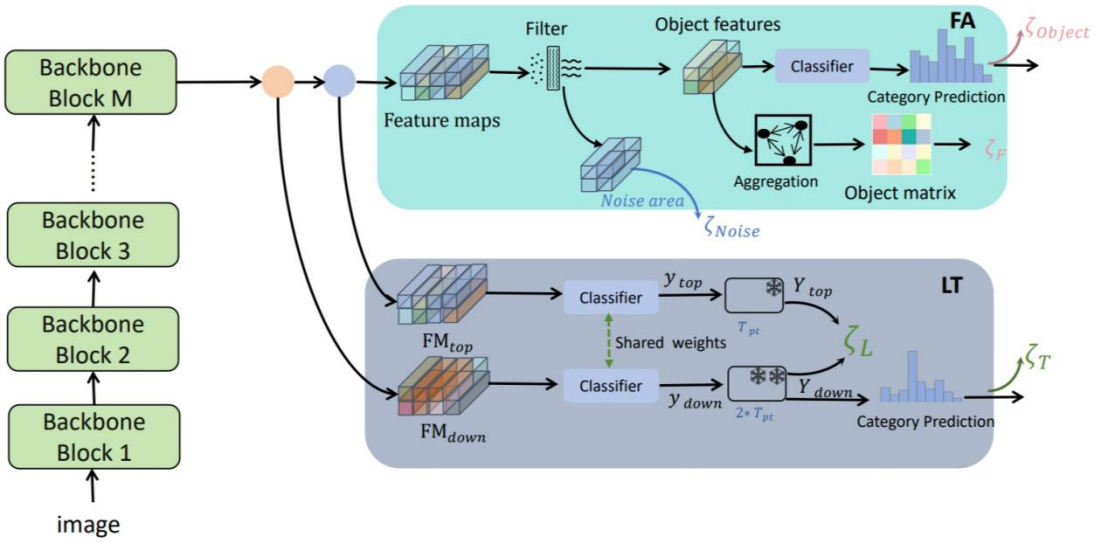

图 5-2 FA 和 LT 模块细节图

Figure 5-2 Detail diagram of FA and LT modules

5.3 模块介绍

5.3.1 过滤聚合模块

令$U^l = [u_1^l, u_2^l, ...u_n^l] \in \mathbb{R}^{W \times H \times C}$，为第 l 层 PA 模块输出的特征图，W，H，C 分别表示特征图的宽，高和通道数，n 表示特征图的数量。接下来，按照预测得分降序对特征图进行排序，可以表示为

$$U_{reset}^1 = F_{de}\left(Soft\max(\delta(U^1))\right) = \left[u_1^1, u_2^1, \ldots, u_n^1\right] \quad (5\text{-}1)$$

式（5-1）中，$F_{de}(\cdot)$表示降序函数，Softmax 表示 Softmax 激活函数，$\delta(\cdot)$表示 1x1 的卷积。随后选择前 $\frac{3n}{4}$ 特征图作为目标特征，其他特征图作为噪声特征。这个过程可以描述如下：

$$U^{l}_{obj}=[u^{l}_{1},u^{l}_{2},...u^{l}_{\frac{3n}{4}}]_{reset} \tag{5-2}$$

$$U^{l}_{nosie}=[u^{l}_{\frac{3n}{4}+1},u^{l}_{\frac{3n}{4}+2}...u^{l}_{n}]_{reset} \tag{5-3}$$

式（5-2）、式（5-3）中的U^{l}_{obj}和U^{l}_{nosie}分别为目标特征区域和噪声区域。值得注意的是，这个过程会持续多次。换句话说，随着主干网层数的增加，U^{l}_{obj}的特征图数量会逐渐减少。

探索特征图之间的相关性以学习特定的语义特征被证明是有效的[27]。为了增加目标特征区和噪声区之间的差异，提出了在同一层中最大化特征相关性的聚合。具体来说，首先对对象特征U^{l}_{obj}全局平均池化（Global Average Pooling，GAP），得到池化特征$F^{l}_{obj}\in\mathbb{R}^{1xn}$并执行 L_2 归一化。然后计算目标特征之间的相关性，形成目标矩阵K^{l}_{obj}，其生成的过程如下：

$$K^{l}_{obj}=F^{l^{T}}_{obj}F^{l}_{obj} \tag{5-4}$$

式（5-4）中，T 表示转置运算。聚集损失函数的目标是最大化K^{l}_{obj}的对角线，从而最大化聚集特征的相关性。聚合的损失函数描述如下：

$$\zeta_{F}=1-\sum_{l=1}^{M}\upsilon_{a}\,\|\,\mathrm{diag}(K^{1}_{obj})\,\|_{2}^{2}, \tag{5-5}$$

式（5-5）中，$||\cdot||_2$表示弗罗贝尼乌斯范数，diag(·)表示提取矩阵的主对角线元素，M 表示最大的网络层数。υ_{a}为超参数，实验设定为 0.1。然后，对目标特征U^{l}_{obj}进行预测，并与真值标签进行比较，其损失函数为

$$\zeta_{Obj}=-\sum_{k=1}^{C_k}Y_{true,k}\log(\beta(U^{l}_{obj})) \tag{5-6}$$

式（5-6）中，$Y_{true,k}$，C_k分别表示第 k 个类别的真实标签和类别的数量。$\beta(\cdot)$表示 Softmax 激活函数。

鉴别性特征聚合的另一个目的是有效地过滤噪声区域。只通过预测分数在区分目

标特征区域和噪声区域方面的作用有限，因为这两种类型的区域有时具有很高的相似性。为了进一步扩大目标特征区与噪声区之间的差异，对噪声区采用均方误差损失函数：

$$\zeta_{\text{Noise}} = \sum_{k=1}^{M} \left(\tau(U_{\text{noise}}^{k}) + 1\right)^{2}, \tag{5-7}$$

式（5-7）中，k 表示噪声区域所在的层数，$\tau(\cdot)$ 和 M 表示 tanh 函数和最大主干网络层数，与 Sigmoid 相比，tanh 激活函数的输出范围更加宽泛。通过优化式（5-5）和式（5-7），鉴别性特征聚合将增加特征的相似度，同时降低U_{obj}^{l}和U_{nosie}^{l}的相关性。这将进一步将目标区域特征从图像中分离出来。综上所述，过滤聚合的损失函数由特征对象损失、聚合损失和噪声损失三部分组成，每一部分都乘以相应的权值，其描述如下：

$$\zeta_{\text{FA}} = u_{\text{f}}\zeta_{\text{F}} + u_{\text{o}}\zeta_{\text{Obj}} + u_{\text{n}}\zeta_{\text{Noise}}, \tag{5-8}$$

式（5-8）中，$u_{\text{f}}, u_{\text{o}}, u_{\text{n}}$是 3 个损失函数的权重系数，在实验中$u_{\text{f}}, u_{\text{o}}, u_{\text{n}}$分别设定为 0.45, 0.2, 0.15。

5.3.2 低温冻结模块

设自底向上和自顶向下的特征图分别为U_{top}^{l}和U_{down}^{l}，其中 l 表示第 l 层骨干网络。其中U_{top}^{l}具有更多的对象特征，而U_{down}^{l}具有更多样化的图像特征。低温冻结的目的是帮助模型从全局的角度逐渐学习到不同的特征。换句话说，低温冷冻更侧重于检测合适的鉴别区域。设U_{down}^{l}和U_{top}^{l}的预测类为y_{down}^{l}和y_{top}^{l}。逐渐学习不同特征的过程可表示为:

$$Y_{\text{down}}^{l} = Sparemax\left(y_{\text{down}}^{l}/(2 * T_{pt})\right), \tag{5-9}$$

$$Y_{top}^{l} = Softmax\left(y_{top}^{l}/T_{pt}\right), \tag{5-10}$$

式（5-9）中，*Sparemax* 为 *Sparemax* 激活函数。$\boldsymbol{T_{pt}}$函数组成如下：

$$T_{pt} = \left(\frac{1}{2}\right)^{\left(\frac{es}{T}\right)}, \tag{5-11}$$

式中 *es* 表示训练轮数，*T* 表示设定温度，在实验中设定为 0.2。值得注意的是，$\boldsymbol{T}_{\text{pt}}$是一个单调递减函数。换句话说，随着训练的进行，$\boldsymbol{T}_{\text{pt}}$减小，鼓励模型探索更多的鉴别

区域，从而提高预测类的准确性。为了发现额外的鉴别性特征，使用 KL 损失函数增大预测类别概率分布的差异：

$$\zeta_L = - Y^{\mathrm{l}}_{top} log\left(\frac{Y^{\mathrm{l}}_{down}}{Y^{\mathrm{l}}_{top}}\right), \tag{5-12}$$

式（5-12）中ζ_L表示 KL 散度损失函数。LT 模块的损耗函数可以描述为：

$$\zeta_{LT} = v_l\zeta_L + v_t\zeta_T\left(Y_{true}, Y^i_{down}\right), \tag{5-13}$$

式（5-13）中v_l和v_t表示损失函数对应的权重系数。FALT 模型的总损失函数是 FA 模块损失函数和 LT 模块损失函数的总和。总损失函数的优化可以帮助 FALT 模型提取更有效的目标特征，并鼓励探索潜在的鉴别特征。

5.4 实验结果与分析

5.4.1 数据集选取

实验部分两个公开的 FGIC 数据集上进行了广泛的实验：CUB-200-2011 和 NA-Birds 数据集。其中 NA-Birds 数据集的部分图像如图 3-7 所示。

5.4.2 实验细节

所提出的 FALT 模型分别采用 Resnet-50 和 Swin-t 作为骨干网。在训练中，使用水平翻转，随机裁剪和随机 GaussianBlur 提高性能。学习率设置为 0.0005，带有余弦衰减和权重。衰减设置为 0.0003。前四层选择的对象特征图数量分别为 256、128、64、32 张。值得注意的是，epoch 数固定为 100，同时 batchsize 设为 16，并保持图像大小为 384 × 384。所有的实验都是在单张 Nvidia GeForce RTX 4090 主机上完成的。在 CUB-200-2011 和 NA-Birds 数据集上的训练分别花费了大约 7 小时和 20 小时。为了更公平的比较，实验中与其他方法对比采用相同环境或者原论文的数值进行对比。

5.4.3 与现有的方法比较

为了评估所提出的 FALT 模型的有效性，将所提出的模型与上述细粒度数据集的最新研究成果进行了比较，详见表 5-1。所提出的 FALT 模型在性能上有显著的提高。从结果来看，所提出的方法在 CUB 和 NA-Birds 两个数据集上取得了具有竞争力的性能。具体而言，表 5-1 的第三列显示了 CUB-200-2011 的比较结果。与目前最好的 IELT[59] 相比，FALT 在 Top-1 精度指标上提高了 1%，与同样以 Swin-t 作为主干网络的方法 ViT-NeT[57]相比提高了 1.2%。NA-Birds 是一个更大的鸟类数据集，不仅在图像数量方

面，而且还有 555 个类别，这大大增加了 FGIC 任务的难度。表 5-1 中的第四列展示了 NA-Birds 数据集上的结果。最好的方法 IELT[59]对比，FALT 有着 1.9%的提高。这可能是由于鸟类的形状更规则。在多个数据集上的实验结果表明，通过过滤特征图进行学习可以改进模型，从而识别出 FGIC 任务的真正高响应区域。此外，识别部位的准确定位可以突出重要区域，促进模型学习更丰富的特征。

5.4.4 消融实验

为了评估每个模块的有效性，在 CUB-200-2011 数据集上进行了消融实验。本实验采用 Resnet-50 和 Swin-t 作为消融实验的骨干网络。消融实验结果如表 5-2 所示。首先，将 PA 模块添加到原来的 Swin-t 上可以提高两个主干网络的精度，例如，在 Resnet-50 上提高 0.8 %，而在 Swin-t 上提高 0.5 %。这些改进表明，PA 可以通过更有效地聚合和传播特征信息来提高特征表示能力。其次，添加 FA 模块两种骨干网的分类准确率分别提高了 1.1%和 0.8%。FA 模块扩大

了目标特征区与噪声区之间的差异，因此选择真高响应区进一步减轻了噪声区的干扰。最后，LT 模块的加入进一步提高了 Cub -200-2011 的分类精度。LT 模块对不同特征尺度的特征图中的弱相关特征进行冻结，检测出合适的鉴别特征。因此，该模型从温度变化中学习特定的细节特征。综上所述，结合这三个模块，所提出的 FALT 可以获得整体上显著的改进。

5.4.5 可视化实验

在图 5-3 中展示了每个模块的可视化结果。第一列是原始图像，这是为了更好地与注意力图可视化进行比较。第二列显示了骨干网络的注意力图可视化。

骨干网虽然识别了一部分目标区域，但也包含了大量的噪声区域。这表明骨

干网很难识别出真正的高响应区域。如第三列所示，在骨干网中加入 PA 模块，该模型减少了噪声区域的干扰。然而，在 FGIC 的任务中，鉴别区域通常位于一个或有限的小区域。然后，将 FA 模块加入骨干网。FA 模块利用特征图和特征之间的关系，选择真实的高响应区域作为目标特征区域，放弃的高响应区域作为噪声区域。为了检测更多的鉴别特征，在骨干网中增加了 LT 模块。温度函数的变化使模型能够学习到更丰富的鉴别特征。最后，在骨干网中加入三个模块，三个模块相互补充，使模型能够减少噪声区域的干扰，学习更丰富的鉴别特征，进一步提高模型的分类性能。

表 5-1 与其他方法进行对比

Table 5-1 Comparison with other methods

Methods	Backbone	Accuracy(%)	
		CUB	NA-Birds
ELoPE[45]	Resnet-50	87.4	-
CIN[41]	Resnet-50	87.5	-
Cross-x[42]	Resnet-50	87.7	86.2
API-NET[51]	Resnet-50	87.7	86.2
AC-NET[50]	Resnet-50	88.1	-
MGE-CNN[48]	Resnet-50	88.5	88.6
FDL[52]	Resnet-50	89.1	-
PAIRS[53]	Resnet-50	89.1	87.9
Ours	Resnet-50	**89.7**	**88.8**
CAMF[54]	Swin-t	91.2	-
AA-Trans[55]	ViT-B	91.4	90.2
SIM-Trans[56]	ViT-B	91.4	90.2
ViT-NeT[57]	Swin-t	91.6	-
TransFG[58]	ViT-B	91.7	90.8
IELT[59]	ViT-B	91.8	90.8
Ours	Swin-t	**92.8**	**92.7**

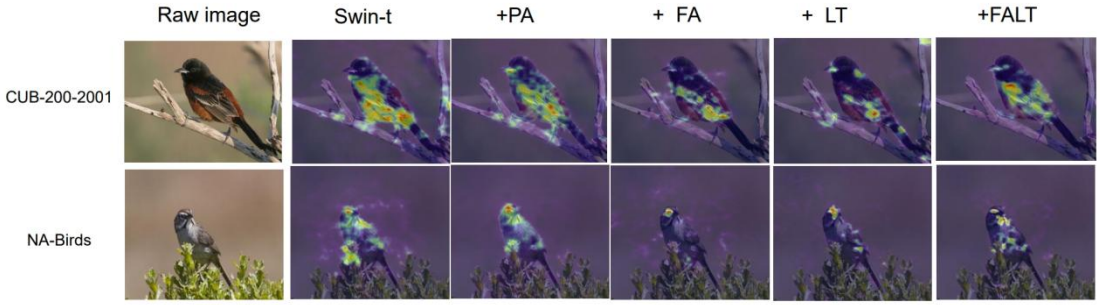

图 5-3 FALT 在不同数据集上的可视化效果

Figure 5-3 Visualization of FALT on different datasets

表 5-2 实验加入不同模块的效果

Table 5-2 The effect of adding different modules to the experiment

FALT			Backbone	
PA	**FA**	**LT**	**Resnet-50**	**Swin-t**
			88.4	91.7
√			89.2	92.2
	√		89.5	92.5
		√	88.9	91.9
√	√	√	**89.7**	**92.8**

5.4.6 超参数分析

为了确定噪声损失权重的影响，分析了超参数υ_n对 CUB-200-2011 数据集的敏感度。不同υ_n值的注意图可视化如图 5-4 所示。很明显，随着υ_n的增大，噪声面积会逐渐减小。然而，当υ_n大于 5 时，一些鉴别特征也逐渐消失，模型的性能逐渐变差。这表明应该将υ_n设置为适当的值。如图 5-5 所示，将υ_n设为 5 时效果更好。υ_n的较大值虽然降低了噪声，但也降低了目标特征，包括整个鉴别区域的特征。这使得模型很难预测对象的类别。反之，υ_n的值越低，可能会使网络集中在噪声较大的区域。这使得模型容易受到噪声区域的影响。因此，在本文中将υ_n设为 5。

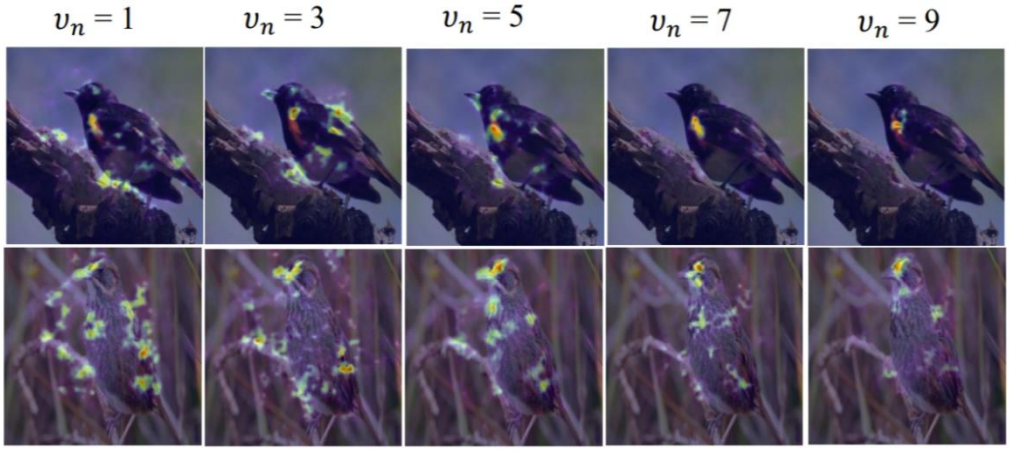

图 5-4 不同υ_n值在数据集上可视化的效果

Figure 5-4 Effect of different υ_n values visualized on the dataset

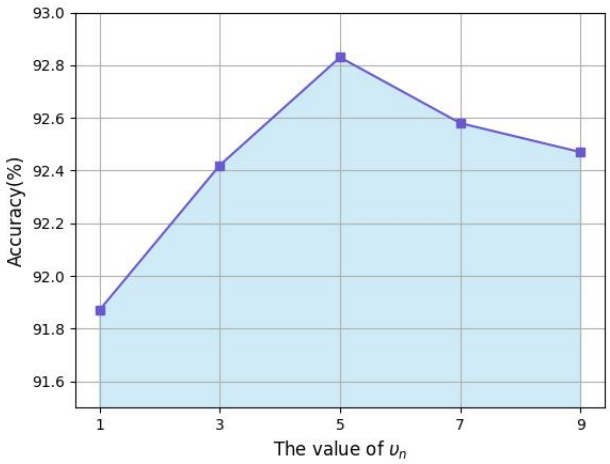

图 5-5　超参数υ_n的敏感度分析

Figure 5-5 Sensitivity analysis of the hyperparameter υ_n

5.5 本章小结

为了更有效的提取出真实高响应区域，同时发现更多的鉴别性特征，提出了一种新的用于细粒度图像分类的 FALT 模型。所提出的 FALT 模型能够识别出真正的高响应区域，学习到更丰富的鉴别特征，从而提高了分类性能。在多个 FGIC 数据集上的大量实验证明了该模型的优越性能。在未来的工作中，我们考虑在保持模型分类性能的同时减少模型的参数，以减少模型的复杂性。

第 6 章 总结与展望

6.1 工作总结

FGIC 是计算机视觉领域的一个重要分支，是图像分割、行为追踪、目标检测等任务的基础。随着深度学习的不断发展，出现三类主流图像分类算法：基于局部定位、基于注意力机制和基于 Transformer 的细粒度图像分类算法。其中，基于局部定位在训练时需要大量手动标注，难以在实际中应用。而基于注意力机制的方法广泛应用于多阶段的图像处理模型中。这类方法中，后续阶段严重依靠前一阶段的定位区域并进一步提取特征。如果优先考虑不感兴趣的区域，不仅会增大计算量，还会误导后续阶段的鉴别性特征提取。另外，这类方法中缺少对每个阶段的客观评估机制，导致模型难以发现鉴别性区域。最后，基于 Transformer 的细粒度图像分类方法中，虽然能够依靠高响应区域捕捉鉴别性特征，但同时也包含粗粒度信息，这对细粒度分类模型分类性能影响较大。同时，只处理已发现的鉴别性特征容易产生局限性，如何发现更多的鉴别性特征又成新的难题。总的来说，本文从三个方面提出一系列方法：（1）从减少背景干扰的角度提出了 ALMA 模型。（2）从评估鉴别性特征重要程度的角度，提出了 ABNT 模型。（3）从增大物体特征和噪声特征差异性以及探测潜在鉴别性特征的角度提出了 FALT 模型。具体工作可总结为：

（1）提出了基于多阶段特征提取的细粒度图像分类方法。ALMA 由两个分支组成，即可调整定位模块和多角度注意模块。具体来说，在可调整定位模块中，首先定位目标感兴趣的区域，并通过背景遮掩对感兴趣区域进行调整，得到调整后的裁剪区域。然后，对调整后的区域进行聚集，定位潜在的鉴别性特征。此外，对于多角度注意模块，逐步最大化原始注意图与随机选择的注意图之间的差异。因此，该方法相较于其它方法具有更好的泛化性。

（2）提出了基于树形分支选举的细粒度图像分类模型。首先，在初始阶段模型通过最大化与粗粒度信息的差异，能够将关注的区域引导到目标上，并为后续阶段提供更有效的指导。随后阶段以前一阶段为基准，对高响应区域进行裁剪，得到新的图像并逐渐发现鉴别性特征。通过决策树集成的多个分支路由模块对决策树的子分支选举，

不断增加鉴别性特征所在阶段的贡献。最后，汇总来自叶节点的预测以获得最终决策。实验结果表明，该方法相比于其它方法具有更好的表现。

（3）提出了基于多尺度特征提取的 Transformer 细粒度图像分类。具体而言，该方法包含过滤聚合模块和低温冻结模块，以选择有效的高响应区域并学习更丰富的鉴别特征。一方面，为了选择有效的高响应区域，过滤聚合模块首先根据特征图之间的相关性过滤掉无用的低响应区域。然后，从真正的高响应区域聚合特征进一步扩大了差异。另一方面，所提出的低温冻结模块能够学习更丰富的鉴别特征。可变化的温度函数鼓励模型探索更多的鉴别性区域并将弱相关特征冻结，提取合适的鉴别特征。实验结果表明，该方法能够去除粗粒度信息的同时，发现潜在的鉴别性区域，以提高模型的分类能力。

6.2 工作展望

本文所提出的一系列方法，虽然解决了图像分类精度的问题，但是由于时间与能力的限制，仍存在一些问题，本文未来需要研究的方向如下：

（1）本文提出的 ALMA 模型通过引导目标的局部区域，从不同角度区分主次信息，使其能够更有效地观察整个目标，从而提高泛化性能。尽管基于多阶段图像处理能够使模型更全面的提取特征，但在其他的计算机视觉任务中没有进行测试。未来可以将算法的通用性提升以适用于不同的计算机视觉领域的任务。

（2）提出的 ABNT 模型通过引导导航模块从不同角度区分目标和背景信息，使其能够更有效地观察整个目标，从而提高泛化性能。尽管如此，在未来的工作中，我们将考虑对树结构进行修剪，提高模型的复杂性。

（3）所提出的 FALT 模型包含过滤聚合模块和低温冻结模块，能够识别出真正的高响应区域，学习到更丰富的鉴别特征，从而提高了分类性能。在多个 FGIC 数据集上的大量实验证明了该模型的优越性能。尽管分类能力提升明显，但在未来的工作中，我们考虑在保持模型分类性能的同时减少模型的参数，以加快模型的训练速度。

参考文献

[1]Barua A, Benharrak K, Chen M, et al. Lotus: Creating Short Videos From Long Videos With Abstractive and Extractive Summarization[C]//Proceedings of the 30th International Conference on Intelligent User Interfaces. 2025: 967-981.

[2]Syahmani S, Rusmansyah R, Leny L, et al. Identifying Students' Undergraduate Concept Understanding and Misconceptions on Amino Acids, Proteins, and Enzymes Using A Four-Tier Diagnostic Test[J]. Journal of Mathematics Science and Computer Education, 2025, 4(2): 145-157.

[3]王波,黄冕,刘利军等.基于多层聚焦 Inception-V3 卷积网络的细粒度图像分类 [J].电子学报,2022,50(01):72-78.

[4]姜昊，凌萍, 陈寸生保. 一种新的基于通道-空间融合注意力及 Swin-t 的细粒度图像分类算法 [J]. 南京师范大学学报(工程技术版), 2023, 23(3): 36–42.

[5]陈权,陈飞,王衍根,等.融合目标定位与异构局部交互学习的细粒度图像分类[J].自动化学报,2024,50(11): 2219-2230.

[6]余悦, 陈楠, 成科扬. 不确定性域感知网络在少样本跨域图像分类中的研究[J].中国图象图形学报,2025,30(02):518-532.

[7]宫智宇, 王士同. 面向重尾噪声图像分类的残差网络学习方法[J].计算机应用,2025,43(01)1-14.

[8]张晓滨, 刘昊. 基于多结构教师蒸馏的服装图像分类方法[J]. 计算机应用与软件,2024,41(12):167-172.

[9]Du R, Xie J, Ma Z, et al. Progressive learning of category-consistent multi-granularity features for fine-grained visual classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(12): 9521-9535.

[10]Branson S, Van Horn G, Belongie S, et al. Bird species categorization using pose normalized deep convolutional nets[J]. arXiv preprint arXiv:1406.2952, 2014.

[11]Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer International Publishing, 2014: 834-849.

[12]Lin D, Shen X, Lu C, et al. Deep lac: Deep localization, alignment and classification for fine-grained recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1666-1674..

[13]Zhang H, Xu T, Elhoseiny M, et al. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1143-1152.

[14]Huang S, Xu Z, Tao D, et al. Part-stacked CNN for fine-grained visual categorization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1173-1182.

[15]Wei X S, Xie C W, Wu J, et al. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization[J]. Pattern Recognition, 2018, 76: 704-714.

[16]Ge W, Lin X, Yu Y. Weakly supervised complementary parts models for fine-grained image classification from the bottom up[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3034-3043.

[17]Liu C, Xie H, Zha Z J, et al. Filtration and distillation: Enhancing region attention for fine-grained visual categorization[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 11555-11562.

[18]Waswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//NIPS. 2017.

[19]Larochelle H, Hinton G E. Learning to combine foveal glimpses with a third-order Boltzmann machine[J]. Advances in neural information processing systems, 2010, 23.

[20]Hu T, Qi H, Huang Q, et al. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification[J]. arXiv preprint arXiv:1901.09891, 2019.

[21]Zhang L, Huang S, Liu W, et al. Learning a mixture of granularity-specific experts for fine-grained categorization[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8331-8340.

[22]Zheng H, Fu J, Mei T, et al. Learning multi-attention convolutional neural network for fine-grained image recognition[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5209-5217.

[23]Zhang T, Chang D, Ma Z, et al. Progressive co-attention network for fine-grained visual classification[C]//2021 International Conference on Visual Communications and Image Processing (VCIP). IEEE, 2021: 1-5.

[24]Jetley S, Lord N A, Lee N, et al. Learn to pay attention[J]. arXiv preprint arXiv:1804.02391, 2018.

[25]Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[26]Zhu H, Ke W, Li D, et al. Dual cross-attention learning for fine-grained visual categorization and object re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 4692-4702..

[27]Shu Y, Yu B, Xu H, et al. Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 449-465..

[28]Zhang Y, Cao J, Zhang L, et al. A free lunch from vit: Adaptive attention multi-scale fusion transformer for fine-grained visual recognition[C]//ICASSP 2022-2022 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, 2022: 3234-3238..

[29]Zhu H, Ke W, Li D, et al. Dual cross-attention learning for fine-grained visual categorization and object re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 4692-4702..

[30]Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020..

[31]Wang M, Zhao P, Lu X, et al. Fine-grained visual categorization: A spatial–frequency feature fusion perspective[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(6): 2798-2812.

[32]He J, Chen J N, Liu S, et al. Transfg: A transformer architecture for fine-grained recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(1): 852-860..

[33]Hu Y, Jin X, Zhang Y, et al. Rams-trans: Recurrent attention multi-scale transformer for fine-grained image recognition[C]//Proceedings of the 29th ACM international conference on multimedia. 2021: 4239-4248..

[34]Liu X, Wang L, Han X. Transformer with peak suppression and knowledge guidance for fine-grained image recognition[J]. Neurocomputing, 2022, 492: 137-149..

[35]Tan M, Yu J, Yu Z, et al. User-click-data-based fine-grained image recognition via weakly supervised metric learning[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2018, 14(3): 1-23..

[36]Zhao Y, Li J, Chen X, et al. Part-guided relational transformers for fine-grained visual recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 9470-9481..

[37]Zhu H, Ke W, Li D, et al. Dual cross-attention learning for fine-grained visual categorization and object re-identification[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 4692-4702..

[38]Lin T Y, RoyChowdhury A, Maji S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1449-1457..

[39]Fu J, Zheng H, Mei T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4438-4446.

[40]Ding Y, Ma Z, Wen S, et al. AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification[J]. IEEE Transactions on Image Processing, 2021, 30: 2826-2836..

[41]Gao Y, Han X, Wang X, et al. Channel interaction networks for fine-grained image categorization[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 10818-10825..

[42]Luo W, Yang X, Mo X, et al. Cross-x learning for fine-grained visual categorization[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8242-8251.

[43]Song J, Yang R. Feature boosting, suppression, and diversification for fine-grained visual classification[C]//2021 International joint conference on neural networks (IJCNN). IEEE, 2021: 1-8.

[44]Du R, Chang D, Bhunia A K, et al. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 153-168..

[45]Huang S, Wang X, Tao D. Stochastic partial swap: Enhanced model generalization and interpretability for fine-grained recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 620-629.

[46]Song J, Yang R. Feature boosting, suppression, and diversification for fine-grained visual classification[C]//2021 International joint conference on neural networks (IJCNN). IEEE, 2021: 1-8.

[47]Branson S, Van Horn G, Belongie S, et al. Bird species categorization using pose normalized deep convolutional nets[J]. arXiv preprint arXiv:1406.2952, 2014..

[48]Zhang L, Huang S, Liu W, et al. Learning a mixture of granularity-specific experts for fine-grained categorization[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8331-8340..

[49]Hanselmann H, Ney H. Elope: Fine-grained visual classification with efficient localization, pooling and embedding[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020: 1247-1256..

[50]Ji R, Wen L, Zhang L, et al. Attention convolutional binary neural tree for fine-grained visual categorization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10468-10477..

[51]Zu Yet O, Rassem T H, Rahman M A, et al. Improved Attentive Pairwise Interaction (API-Net) for Fine-Grained Image Classification[C]//2021 Emerging Technology in Computing, Communication and Electronics (ETCCE). IEEE, 2021: 1-6..

[52]Liu C, Xie H, Zha Z J, et al. Filtration and distillation: Enhancing region attention for fine-grained visual categorization[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 11555-11562..

[53]Guo P, Farrell R. Aligned to the object, not to the image: A unified pose-aligned representation for fine-grained recognition[C]//2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2019: 1876-1885..

[54]Miao Z, Zhao X, Wang J, et al. Complemental attention multi-feature fusion network for fine-grained classification[J]. IEEE Signal Processing Letters, 2021, 28: 1983-1987..

[55]Wang Q, Wang J J, Deng H, et al. Aa-trans: Core attention aggregating transformer with information entropy selector for fine-grained visual classification[J]. Pattern Recognition, 2023, 140: 109547.

[56]Sun H, He X, Peng Y. Sim-trans: Structure information modeling transformer for fine-grained visual categorization[C]//Proceedings of the 30th ACM international conference on multimedia. 2022: 5853-5861..

[57]Kim S, Nam J, Ko B C. Vit-net: Interpretable vision transformers with neural tree decoder[C]//International conference on machine learning. PMLR, 2022: 11162-11172..

[58]He J, Chen J N, Liu S, et al. Transfg: A transformer architecture for fine-grained recognition[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(1): 852-860.

[59]Xu Q, Wang J, Jiang B, et al. Fine-grained visual classification via internal ensemble learning transformer[J]. IEEE Transactions on Multimedia, 2023, 25: 9015-9028..

[60]Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[J]. 2011..

[61]Van Horn G, Branson S, Farrell R, et al. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 595-604..

[62]Maji S, Rahtu E, Kannala J, et al. Fine-grained visual classification of aircraft[J]. arXiv preprint arXiv:1306.5151, 2013..

[63]Krause J, Stark M, Deng J, et al. 3d object representations for fine-grained categorization[C]//Proceedings of the IEEE international conference on computer vision workshops. 2013: 554-561.

致　谢

时光荏苒，岁月如梭。回首这段求知的岁月，我满怀感激之情，想要对那些在我这三年上给予无私帮助的人，致以最深的谢意。

首先，我要衷心感谢我的导师徐晓峰老师。徐老师不仅是我学术上的引路人，更是我人生路上的明灯。在徐老师的悉心指导下，我不仅学到了专业知识，更学会了如何以更加成熟和全面的视角去看待问题、解决问题。徐老师严谨的治学态度、务实的工作作风和宽厚的人格魅力，都深深地影响着我，让我受益终身。能选到这样平易近人的导师，是我人生中莫大的幸运。

同时，我还要感谢 MVIC 实验室的汪军、严楠和戴家树等老师。实验室的老师们在给予了我莫大的鼓励和帮助。在他们的指导下，实验室的硬件设施不断完善，学术氛围日益浓厚，为 MVIC 的研究生们提供了一个良好的学习和研究环境。

此外，我要深深感谢 MVIC 实验室的每一位同学和朋友。MVIC 实验室的同学就像是一群亲密无间的好朋友，共同探讨学术问题，分享彼此的研究心得并共同进步。每当有人遇到困难，总有人伸出援手，并给予宝贵的建议和帮助。这种团结合作的精神，使我在学术的道路上更加有信心。

我深知，家人的付出和支持是无法用言语来表达的。他们为我所做的一切，我都铭记在心，感恩不尽。在此，我再次向所有帮助过我的人表示衷心的感谢。未来的日子里，我将继续努力，不负众望。

尚德教学 唯实惟新