北京大学

博士研究生学位论文

题目：辨识性特征学习及在细粒度分析中的应用

姓　　名：何相腾

学　　号：1401111369

院　　系：信息科学技术学院

专　　业：计算机应用技术

研究方向：图像、视频理解与检索

导　　师：彭宇新教授

二〇二〇年六月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

摘要

细粒度分析旨在对粗粒度的大类进行细粒度的子类划分，如把鸟划分为里海燕鸥、北极燕鸥等子类别。其广泛应用于智能农业、智能医疗等智能产业，具有重要的研究和应用价值。其挑战在于类间差异小、类内差异大。以图像为例，不同子类别在形状、颜色上差异细微，难以区分；相同子类别在姿态、视角上差异显著，容易误分。因此，关键科学问题是：**如何获取细粒度子类别的辨识性信息并有效表达，突破细粒度分析难题**。针对上述问题，本文从**减少标注成本、减少人工先验、提高辨识速度、提高语义关联**四个方面展开**辨识性特征学习**研究，并分别应用于**细粒度图像分类和细粒度跨媒体检索任务**。主要工作总结如下：

1. 在减少标注成本上，提出了基于对象-部件注意力模型的细粒度图像分类方法。在对象级注意力上，提出注意力选择和显著性提取，自动定位对象区域，学习更精细的对象特征。在部件级注意力上，提出空间关联约束和部件语义对齐，实现辨识性部件的有效定位，排除了姿态、视角等差异的干扰。两者结合能够学习到多粒度的辨识性特征，准确率超过了使用对象、部件人工标注的强监督方法。
2. 在减少人工先验上，提出了基于堆叠式深度强化学习的细粒度图像分类方法。首先，层次化地定位图像中的多粒度辨识性区域，并自适应地确定其数目。然后，通过多尺度区域的定位及辨识性特征学习，进一步提升细粒度图像分类准确率。学习过程由语义奖励函数驱动，能够有效捕捉图像中的辨识性、概念性的视觉信息，实现弱监督甚至无监督条件下的辨识性特征学习。
3. 在提高辨识速度上，提出了基于弱监督快速辨识定位的细粒度图像分类方法。首先，提出多级注意力引导的辨识性定位，通过显著图生成伪监督信息，实现了弱监督条件下的辨识性定位。进一步显著图驱动二次定位学习，增强了定位的准确性。然后，提出多路端到端辨识性定位网络，实现多个辨识性区域的同时定位，从而提高了辨识速度。多个辨识性区域之间互补促进，提升细粒度图像分类准确率。
4. 在提高语义关联上，引入文本、视频、音频等跨媒体数据，提出了基于细粒度分类的跨媒体检索方法。建立了首个包含 4 种媒体类型（图像、文本、视频和音频）的细粒度跨媒体检索公开数据集和评测基准 PKU FG-XMedia。提出了能够同时学习 4 种媒体统一表征的深度模型 FGCrossNet，确保统一表征的辨识性、类内紧凑性和类间松散性。实现图像向跨媒体的扩展，分类向检索的扩展。

关键词：细粒度图像分类; 细粒度跨媒体检索; 辨识性特征; 注意力; 强化学习

Discriminative Feature Learning and Its Application to Fine-grained Analysis

Xiangteng He (Computer Application and Techonology)
Directed by Prof. Yuxin Peng

ABSTRACT

Fine-grained analysis is to divide coarse-grained category into fine-grained subcategories, such as dividing the coarse-grained category of "Bird" into the fine-grained subcategories of "Common Tern", "Caspian Tern" and so on. It is widely applied to the intelligent industry, such as intelligent agriculture and healthcare, and has important research and application values. Its challenges lie in small inter-class variance and large intra-class variance. Take image as an example, there only exist subtle distinctions among different fine-grained subcategories, like similar shape and color, so it's hard to classify them. However, there exist large distinctions in the same subcategory, like different poses and views, so it's easy to misclassify them. Therefore, the key scientific problem is how to obtain the discriminative information of the fine-grained subcategories with effective representation, breaking through the fine-grained analysis problem. Therefore, this paper conducts the studies of discriminative feature learning from four aspects: reducing annotation cost and artificial prior, improving discrimination speed and semantic association, and applies them to fine-grained image classification and fine-grained cross-media retrieval respectively. The main work canbe summized as follows:

1. To reduce annotation cost, the object-part attention model is proposed for fine-grained image classification. In object-level attention, attention selection and saliency extraction can automatically localize the object region, learning more refined object-level feature. In part-level attention, spatial correlation constraint and part semantic alignment can effectively localize the discriminative parts, suppressing the interferences of different poses and views. Their combination learn the discriminative feature with multiple granularities to outperform than those supervised methods using object-level or part-level annotations.
2. To reduce artificial prior, the stacked deep reinforcement learning based fine-grained image classification approach is proposed. First, automatically localize the multi-

granularity discriminative regions in a hierarchical manner, and determine the number of discriminative regions in an adaptive manner. Then, the localization and discriminative feature learning of multi-scale regions further boost the fine-grained image classification accuracy. Semantic reward function is proposed to drive the learning process to fully capture the salient and conceptual visual information. It realizes the discriminative feature learning in the weakly supervised or unsupervised manner.

3. To improve discrimination speed, the weakly supervised fast discriminative localization is proposed for fine-grained image classification. First, multi-level attention guided discriminative localization is proposed to realize weakly-supervised discriminative localization via utilizing attention map to generate pseudo supervised information. The secondary localization learning driven by attention map, further boosts the localization accuracy. Then, multi-pathway end-to-end discriminative localization network is proposed to simultaneously localize multiple discriminative regions, boosting the discrimination speed. Besides, multiple discriminative regions provide complementary information to guarantee the fine-grained image classification accuracy.
4. To improve semantic association, this paper imports text, video and audio to propose the fine-grained classification based cross-media retrieval approach. First, construct the first dataset and benchmark for fine-grained cross-media retrieval with 4 media types (i.e. image, text, video and audio), namely PKU FG-Xmedia. Besides, the uniform deep model (i.e. FGCrossNet) for fine-grained cross-media retrieval is proposed, to simultaneously learn the common representation of 4 media types, which guarantees the discrimination, intra-class compactness and inter-class looseness of the common representation. Therefore, implement the extensions from image data to cross-media data, as well as from classification task to retrieval task.

KEYWORDS: Fine-grained Image Classification; Fine-grained Cross-media Retrieval; Discriminative Feature; Attention; Reinforcement Learning

目录

第一章 绪论 **1**
1.1 研究背景 1
1.2 研究难题 3
1.3 研究内容 4
1.4 本文的结构组织 5
第二章 国内外研究现状 **7**
2.1 细粒度图像分类 7
2.1.1 基于定位的方法 7
2.1.2 基于编码的方法 10
2.1.3 基于属性的方法 11
2.2 细粒度图像检索 11
2.3 细粒度视频分类 12
2.4 细粒度跨媒体分析 12
2.4.1 图像-文本之间的跨媒体分析 12
2.4.2 视频-音频之间的跨媒体分析 13
第三章 基于对象-部件注意力模型的细粒度图像分类 **15**
3.1 引言 15
3.2 算法描述 16
3.2.1 对象级注意力模型 16
3.2.2 部件级注意力模型 18
3.2.3 最终预测 22
3.3 实验结果与分析 23
3.3.1 实验数据集和评价指标 23
3.3.2 实验设置 24
3.3.3 与现有方法进行对比 25
3.3.4 基线实验 28
3.4 本章小结 32
第四章 基于堆叠式深度强化学习的细粒度图像分类 **35**
4.1 引言 35

4.2 算法描述 37
4.2.1 问题定义 38
4.2.2 多粒度辨识性定位 38
4.2.3 辨识性定位中的 Q-learning 算法 42
4.2.4 无监督辨识性定位 42
4.2.5 多尺度特征学习 43
4.2.6 最终预测 44
4.3 实验结果与分析 44
4.3.1 实验设置 44
4.3.2 与现有方法进行对比 46
4.3.3 辨识性定位的有效性 47
4.3.4 无监督辨识性定位的有效性 49
4.3.5 本章 StackDRL 方法每个组成部分的有效性 50
4.4 本章小结 52

第五章 基于弱监督快速辨识定位的细粒度图像分类 53

5.1 引言 53
5.2 算法描述 54
5.2.1 多级注意力提取网络 54
5.2.2 辨识性定位网络 55
5.2.3 训练过程 56
5.3 实验结果与分析 57
5.3.1 实验设置 57
5.3.2 与现有方法进行对比 58
5.3.3 本章 WSFDL 方法中每个组成模块的有效性 60
5.3.4 基线实验 61
5.3.5 辨识性定位的有效性 62
5.3.6 不同注意力的不同聚焦点 64
5.3.7 错误细粒度分类分析 65
5.4 本章小结 66

第六章 基于细粒度分类的跨媒体检索 67

6.1 引言 67
6.2 细粒度跨媒体检索数据集和评测基准 69
6.2.1 数据集的构建 70

6.2.2 特点 72
6.2.3 细粒度跨媒体检索任务 73
6.3 算法描述 74
6.3.1 网络结构 74
6.3.2 数据处理 75
6.3.3 损失函数 76
6.3.4 训练和检索 77
6.4 实验结果与分析 78
6.4.1 数据划分和评价指标 78
6.4.2 对比方法 78
6.4.3 与现有方法进行对比 79
6.4.4 基线实验 80
6.5 本章小结 81
第七章 总结与展望 **83**
7.1 工作总结 83
7.2 未来展望 84
参考文献 **85**
个人简历、攻读博士学位期间的研究成果 **95**
致谢 **99**
北京大学学位论文原创性声明和使用授权说明 **101**

插图

1.1 传统图像分类与细粒度图像分类的对比 1
1.2 细粒度分析的两大挑战：类间差异小和类内差异大（以图像为例） . . . 2
1.3 本文主要工作 5

2.1 细粒度图像标注信息示例 7
2.2 Part-based R-CNN 方法示意图 [5] 9
2.3 知识图示意图 [24] 11

3.1 对象-部件注意力模型的总体框架 17
3.2 本章对象-部件注意力模型中显著性提取结果的展示 19
3.3 本章 OPAM 方法部件语义对齐结果展示 21
3.4 谱聚类示意图 22
3.5 本章对象级注意力模型和部件级注意力模型所选择的图像块对比 23
3.6 CUB-200-2011 [3]，Cars-196 [28]，Oxford-IIIT Pet [41] 和 Oxford-Flower-102 [42] 数据集的样例 24
3.7 部件定位的失败例子 29
3.8 本章 OPAM 方法对象定位和部件定位的结果 30
3.9 与现有方法部件自动定位结果对比 32

4.1 人类识别图像时的注意力示意图 35
4.2 本章堆叠式深度强化学习的总体框架 37
4.3 辨识性定位动作示意图 39
4.4 树状执行策略示意图 40
4.5 在 CUB-200-2011 和 Cars-196 两个数据集上的召回率与 IoU 值曲线 . . . 41
4.6 Q-network 结构 43
4.7 CUB-200-2011 数据集上定位召回率与 IoU 曲线 47
4.8 本章对象级强化学习方法在 CUB-200-2011 数据集上定位的召回率与 IoU 曲线 48
4.9 对象级强化学习辨识性定位过程示意图 49
4.10 本章部件级强化学习方法定位到的辨识性区域数目展示 50
4.11 部件级强化学习辨识性定位的结果 50

5.1 本章弱监督快速辨识定位方法示意图 55
5.2 辨识性定位网络定位到的辨识性区域与对象级标注的对比 63
5.3 本章 WSFDL 方法中多级注意力定位到的区域结果 64
5.4 CUB-200-2011 和 Cars-196 两个数据集上最容易误分的细粒度子类别对 . 65
5.5 CUB-200-2011 和 Cars-196 两个数据集上的分类混淆矩阵 66

6.1 跨媒体检索示意图 68
6.2 粗粒度跨媒体检索 VS 细粒度跨媒体检索 69
6.3 本章构造的细粒度跨媒体数据集中样例展示 73
6.4 本章 FGCrossNet 网络结构示意图 75
6.5 文本处理示意图 76

表格

3.1 CUB-200-2011 数据集上实验结果 26
3.2 Cars-196 数据集上实验结果 27
3.3 Oxford-IIIT Pet 数据集上实验结果 27
3.4 Oxford-Flower-102 数据集上实验结果 28
3.5 本章 OPAM 方法每个模块在 CUB-200-2011、Cars-196、Oxford-IIIT Pet 和 Oxford-Flower-102 四个数据集上的结果 28
3.6 对象-部件空间关联约束和部件语义对齐结果 31
3.7 对象级注意力选择的有效性 32

4.1 CUB-200-2011 数据集上实验结果 45
4.2 Cars-196 数据集上实验结果 46
4.3 无监督辨识性定位的有效性 49
4.4 多尺度特征学习的有效性 50
4.5 多粒度辨识性定位的有效性 51
4.6 对象级强化学习有效性实验结果 52
4.7 语义奖励函数的有效性 52

5.1 细粒度图像分类速度对比结果 57
5.2 CUB-200-2011 数据集上的细粒度分类结果 59
5.3 Cars-196 数据集上的细粒度分类结果 60
5.4 多级注意力在细粒度分类准确率上的有效性 61
5.5 辨识性定位网络在细粒度分类速度上的有效性 61
5.6 基线实验对比 62
5.7 定位结果对比 63
5.8 CUB-200-2011 数据集上对于每个部件的 PCL 值 64
5.9 不同 IoU 值的百分比 64

6.1 本章细粒度跨媒体检索数据集与现有常用粗粒度跨媒体检索数据集对比 70
6.2 文本和音频的来源网站 71
6.3 双模态细粒度跨媒体检索结果（MAP） 80
6.4 多模态细粒度跨媒体检索结果（MAP） 80

6.5 三种约束的有效性 81

第一章　绪论

1.1　研究背景

近年来，多媒体内容分析技术[1] 迅猛发展，涉及人工智能、多媒体、计算机视觉等多个领域。分类与检索是其中最基本且重要的研究方向，主要包括图像分类与检索、视频分类与检索、跨媒体检索等，其主要聚焦于对粗粒度的大类（如鸟、车等）进行分析。然而，相比于粗粒度的大类，在日常生活中我们有更加精细化的分析应用需求，通常趋向于获取具体的细粒度子类别信息（如美洲乌鸦、现代伊兰特 2007 等）。显然，传统的分类检索等多媒体内容分析技术无法实现对大类的精细划分，不能满足人类的需求。为了使得计算机趋向于人类智能，细粒度分析便应运而生。与传统的多媒体内容分析不同，细粒度分析旨在对粗粒度的大类进行细粒度的子类划分。以图像分类为例，在传统图像分类[2] 中，计算机只需要判断图 1.1中左边两幅图像属于鸟类，右边两幅图像属于车类即可；而在细粒度图像分类[3] 中，计算机需要精细地识别出左边两幅图像分别属于鸟类中的美洲乌鸦（American Crow）和鱼鸦（Fish Crow），而右边两幅图像分别属于车类中的现代伊兰特 2007（Hyundai Elantra Sedan 2007）和丰田红杉 SUV 2012（Toyota Sequoia SUV 2012）。

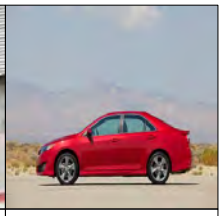
图像				
传统图像分类	**鸟**	**鸟**	**车**	**车**
细粒度图像分类	**美洲乌鸦**	**鱼鸦**	**现代伊兰特 2007**	**丰田红杉 SUV 2012**

图 1.1　传统图像分类与细粒度图像分类的对比

细粒度分析是多媒体内容分析领域中的重要研究方向，能够为智能产业的发展提供关键技术支撑，有着丰富的应用场景：在智能农业上，应用于病虫害识别等；在智能医疗上，应用于皮肤病变、癌细胞检测等；在智能零售上，应用于商品的自动识别以实现自动结账；在智能深海作业上，应用于鱼类的识别分析等。综上，细粒度分析具有重要的研究和应用价值。

在现实生活中，存在着成百上千甚至上万种细粒度子类别，如据统计目前世界上

存在 10,426 种鸟类子类别 ①、34,300 种鱼类子类别 ②、200 多种癌细胞子类别 ③。面对如此浩繁的细粒度子类别，人类很难进行精准辨识。因此，细粒度分析是一项非常具有挑战性的任务。

细粒度分析的挑战主要体现在以下两个方面：1）类间差异小：不同的细粒度子类别在视觉上（即图像、视频）具有相似的形状、颜色等外观；在听觉上（即音频）发出相似的声音；在语言上（即文本描述）具有相似的描述词汇。因此，不同的细粒度子类别由于类间差异小，导致难以区分。2）类内差异大：相同的细粒度子类别在视觉上由于姿态、视角等外界因素影响存在差异；在听觉上由于外界环境的背景声音等而存在差异；在语言上由于描述者的不同而存在差异。因此，相同的细粒度子类别由于类内差异大，导致容易误分。以图像为例，如图 1.2 所示，上面四幅图像分别属于大冠蝇霸鹟 (Great Crested Flycatcher) 、阿卡迪亚霸鹟 (Acadian Flycatcher)、小纹霸鹟（Least Flycatcher）和黄腹纹霸鹟（Yellow-bellied Flycatcher）四个不同的细粒度子类别，但其在形状、颜色等外观上具有非常高的相似度；而下面四幅图像均属于蓝鸦 (Indigo Bunting) 这一细粒度子类别，但由于姿态、光照等外界因素的影响导致其在形状、颜色等外观上具有显著差异。

图 1.2　细粒度分析的两大挑战：类间差异小和类内差异大（以图像为例）

① 国际鸟盟（http://datazone.birdlife.org/species/requestdis）
② 世界鱼类数据库（https://www.fishbase.se/search.php）
③ 英国癌症研究中心（https://www.cancerresearchuk.org/about-cancer/what-is-cancer）

1.2　研究难题

在细粒度分析中，需要解决的科学问题是：如何获取细粒度子类别的辨识性信息并进行有效表达，突破细粒度分析难题。针对上述问题，研究者们主要聚焦于图像数据和分类任务来展开细粒度分析方面的研究，即细粒度图像分类[3]。细粒度子类别在视觉上的辨识性特征主要存在于细小的局部部件中，如鸟喙、胸部、翅膀、尾巴等，因此，如何检测对象及其部件，并对其进行有效表达，便成为了细粒度图像分类的重要研究问题。现有方法虽然取得了一定的研究成果，促进了细粒度图像分类的发展，但其仍存在以下难题：

- **难题 1：标注成本巨大**。现有方法通常依赖于图像级类别信息、对象级位置信息和部件级位置信息等标注来训练对象和部件检测器，从而实现辨识性对象和部件的定位和特征学习[4, 5]。然而，这些标注信息的获取耗时耗力，成本巨大。Gebru 等人指出构造具有 200 万条标注信息的细粒度数据集需要耗费 30 万美元 [6]。由此可见，依赖于成本巨大的标注信息，不利于细粒度图像分类方法向实际应用进行技术转化。
- **难题 2：依赖人工先验**。因为细粒度图像分类的关键在于辨识性区域的定位，即对象及其部件，那么辨识性区域的数目将直接影响细粒度图像分类的准确率。然而，现有方法通常依赖于实验验证等人工先验的方式来设定辨识性区域的数目[7]。因此，需要根据新任务或数据集的数据来重新设定辨识性区域的数目，从而导致现有细粒度图像分类方法在可用性和可扩展性上的局限。
- **难题 3：忽略辨识速度**。现有方法通常聚焦于细粒度图像分类的准确率问题，却忽略了速度这一实际应用中的关键问题。现有方法通常采用两阶段的流程：先进行辨识性定位，再提取辨识性区域对应的特征来训练分类器[8]。多阶段的流程操作导致速度难以满足实际应用的需求。

除了细粒度图像分类方向的研究，研究者们还在细粒度图像检索[9]、细粒度视频分类[10] 上展开细粒度分析研究，但相对较少。尽管任务有所变化，但归其根本依旧与细粒度图像分类研究所关注的问题一致，即视觉信息的辨识性特征学习。然而，图像、文本、视频、音频等跨媒体数据已经成为了当今世界信息传播的主要载体，由此也导致了新的难题:

- **难题 4：忽略语义关联**。现有研究主要聚焦于以图像为主的视觉数据，忽略了与其相关联的文本、视频、音频等跨媒体数据。图像、文本、视频、音频等跨媒体数据之间存在着隐含的语义关联关系，通过分析这种跨媒体隐含语义关联关系，能够进一步促进细粒度分析以及跨媒体数据的有效管理与利用。然而，现有跨媒体的研究通常聚焦于粗粒度的大类，在细粒度分析方面的研究还很匮乏。

1.3 研究内容

针对上述研究难题，本文从减少标注成本、减少人工先验、提高辨识速度、提高语义关联四个方面展开辨识性特征学习研究，并将其分别应用于细粒度图像分类和细粒度跨媒体检索任务。本文的研究内容归纳如下：

- 在减少标注成本上，提出了基于对象-部件注意力模型的细粒度图像分类方法。在对象级注意力上，提出注意力选择和显著性提取，自动定位对象区域，学习更精细的对象特征。在部件级注意力上，提出空间关联约束和部件语义对齐，实现辨识性部件的有效定位，排除了姿态、视角等差异的干扰。两者结合能够学习到多粒度的辨识性特征，准确率超过了使用对象、部件人工标注的强监督方法。
- 在减少人工先验上，提出了基于堆叠式深度强化学习的细粒度图像分类方法。首先，层次化地定位图像中的多粒度辨识性区域，并自适应地确定其数目。然后，通过多尺度区域的定位及辨识性特征学习，进一步提升细粒度图像分类准确率。学习过程由语义奖励函数驱动，能够有效捕捉图像中的辨识性、概念性的视觉信息，实现弱监督甚至无监督条件下的辨识性特征学习。
- 在提高辨识速度上，提出了基于弱监督快速辨识定位的细粒度图像分类方法。首先，提出多级注意力引导的辨识性定位，通过显著图生成伪监督信息，实现了弱监督条件下的辨识性定位。进一步显著图驱动二次定位学习，增强了定位的准确性。然后，提出多路端到端辨识性定位网络，实现多个辨识性区域的同时定位，从而提高了辨识速度。多个辨识性区域之间互不促进，提升细粒度图像分类准确率。
- 在提高语义关联上，引入文本、视频、音频等跨媒体数据，提出了基于细粒度分类的跨媒体检索方法。建立了首个包含 4 种媒体类型（图像、文本、视频和音频）的细粒度跨媒体检索公开数据集和评测基准 PKU FG-XMedia。提出了能够同时学习 4 种媒体统一表征的深度模型 FGCrossNet，确保统一表征的辨识性、类内紧凑性和类间松散性。实现图像向跨媒体的扩展，分类向检索的扩展。

如图1.3所示，本文以辨识性特征学习为基础，针对四个难题展开研究工作，它们之间具有很强的关联：1）对象-部件注意力模型旨在减少标注成本，由现有方法依赖的三种标注信息，即图像级、对象级和部件级标注，减少为仅使用图像级标注信息；2）堆叠式深度强化学习，在此基础上进一步减少人工先验的依赖，从而避免了现有方法在可用性和可扩展性上的局限；3）弱监督快速辨识定位则针对上述方法所忽略的速度问题展开研究，在加快速度的同时保证分类准确率；4）基于细粒度分类的跨媒体检索，将上述细粒度图像分类上的研究进行扩展，实现了图像数据向跨媒体数据的扩展，分类任务向检索任务的扩展。

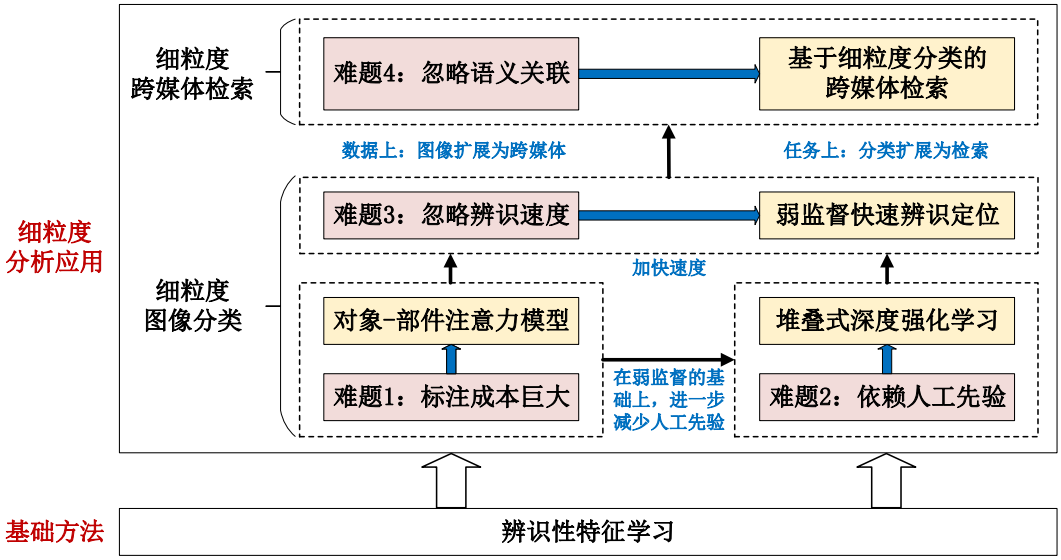


图 1.3　本文主要工作

1.4　本文的结构组织

本文内容组织如下：

第一章，绪论。总览介绍本文选题的研究背景、研究难题、以及研究内容。

第二章，国内外研究现状。主要介绍在细粒度分析上的现有工作，包括细粒度图像分类、细粒度图像检索、细粒度视频分类、细粒度跨媒体分析。

第三、四、五、六章，本文方法部分。分别从减少标注成本、减少人工先验、提高辨识速度、提高语义关联四个方面对本文方法进行介绍。

第七章，总结与展望。对全文进行总结，并介绍未来研究工作的方向。

第二章　国内外研究现状

目前，研究者们主要从细粒度图像分类、细粒度视频分类、细粒度图像检索等方面展开细粒度分析研究。此外，在现有的细粒度图像/视频分类工作中，有研究者不仅利用图像、视频等单一媒体信息，也综合考虑文本、音频等跨媒体数据，以此提升分类准确率。在本章中，将单独介绍此类工作。因此，本章主要从以下几个方面展开介绍：细粒度图像分类、细粒度图像检索、细粒度视频分类、细粒度跨媒体分析。

2.1　细粒度图像分类

现有细粒度图像分类方法一般可以划分为以下三类: 基于定位的方法、基于编码的方法以及基于属性的方法。

2.1.1　基于定位的方法

细粒度图像分类中不同细粒度子类别的差异主要存在于细小的局部部件中，如鸟喙、前额、翅膀等，因此，现有基于定位的细粒度图像分类方法一般分为两个步骤：1）定位辨识性区域，即对象区域（如鸟）及其部件区域（如鸟喙、前额、翅膀等）；2）对定位到的具有辨识性的区域提取特征以完成分类器的训练，并对该辨识性区域进行细粒度分类预测，得到最终的预测结果。

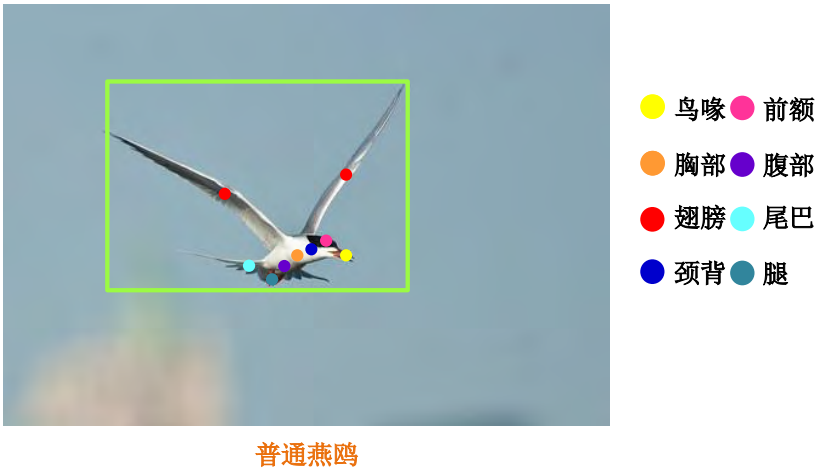


图 2.1　细粒度图像标注信息示例

在细粒度图像分类任务中，对于每幅图像通常有以下三种标注信息：图像级标注、对象级标注和部件级标注。如图2.1所示，图像级标注指的是图像的细粒度子类别信息，

如普通燕鸥（Common Tern）；对象级标注指的是对象在图像中的位置信息，通常用矩形框表示（Bounding Box），如图中绿色矩形框；部件级标注指的是对象中具有辨识性的部件在图像中的位置信息（Part Location），如鸟的喙（Beak）、前额（Forehead）和胸部（Breast）等，具体如图中各种颜色的圆点所示。

现有基于定位的细粒度图像分类方法根据所使用标注信息的不同，一般分为强监督细粒度图像分类方法和弱监督细粒度图像分类方法。其中，强监督细粒度图像分类方法不仅使用了图像级标注信息，同样还使用了对象级、部件级标注信息；而弱监督细粒度图像分类方法仅使用了图像级标注信息。下面分别对两类方法进行介绍。

2.1.1.1 强监督细粒度图像分类方法

由于细粒度图像分类的关键在于准确定位到图像中的辨识性区域，包括对象及其部件，因此，早期研究者们通常利用对象级和部件级的标注信息来训练对象和部件检测器，以此来定位图像中的辨识性区域，进一步提取对应区域的特征来训练分类器。Zhang 等人[4] 将可变部件模型（Deformable Parts Model，简称 DPM）[11] 引入细粒度图像分类，利用 DPM 来训练对象和部件检测器以定位图像中的对象和部件区域，然后提取其特征来训练分类器。随着深度学习的发展，Zhang 等人[5] 进一步提出了 Part-based R-CNN 方法，利用 R-CNN [12] 检测图像中具有辨识性的候选区域，然后通过几何约束筛选候选区域并将其用于训练分类器。如图 2.2 所示，该方法首先采用对象级和部件级标注信息来训练对应的检测器，接着采用选择搜索方法（Selective Search）[13] 为每一张图像生成成百上千的图像候选区域，并对其用已训练的检测器进行评分，从中筛选出具有高分的图像候选区域，即定位得到图像中的对象和部件区域。但是，这样筛选出的不同部件会出现重叠等现象，因此他们进一步提出了一种对定位区域的约束方法。具体约束方法可用公式 (2.1) 表示，其中 x_0 表示对象的位置，x_1 到 x_n 表示 n 个部件的位置。当部件区域 x_i 超出对象区域 x_0 一定像素点时，$c_{x_0}(x_i)$ 为 0，否则为 1。$\delta_i(x_i)$ 为对区域 x_i 计算在训练数据上的混合高斯模型的值。该约束要求所有部件区域不能超出对象区域的某个阈值，同时利用训练数据对区域位置进行约束以保证其可靠性。

$$\Delta_{geometrix}(X) = (\prod_{i=1}^{n} c_{x_0}(x_i))(\prod_{i=0}^{n} \delta_i(x_i)) \tag{2.1}$$

Huang 等人 [14] 提出了一种端到端的部件堆叠的 CNN 网络（Part-Stacked CNN，简称 PS-CNN）方法，在定位对象和部件区域之后，提取对象和部件的卷积特征并融合进行细粒度图像分类。该方法首先利用对象级标注信息将对象从输入图像中裁剪出来，接着利用全卷积神经网络生成特定部件的显著图，然后对显著图进行高斯滤波去噪并生成特征图作为部件定位的结果。完成部件定位后，PS-CNN 使用两个输入流的方法设

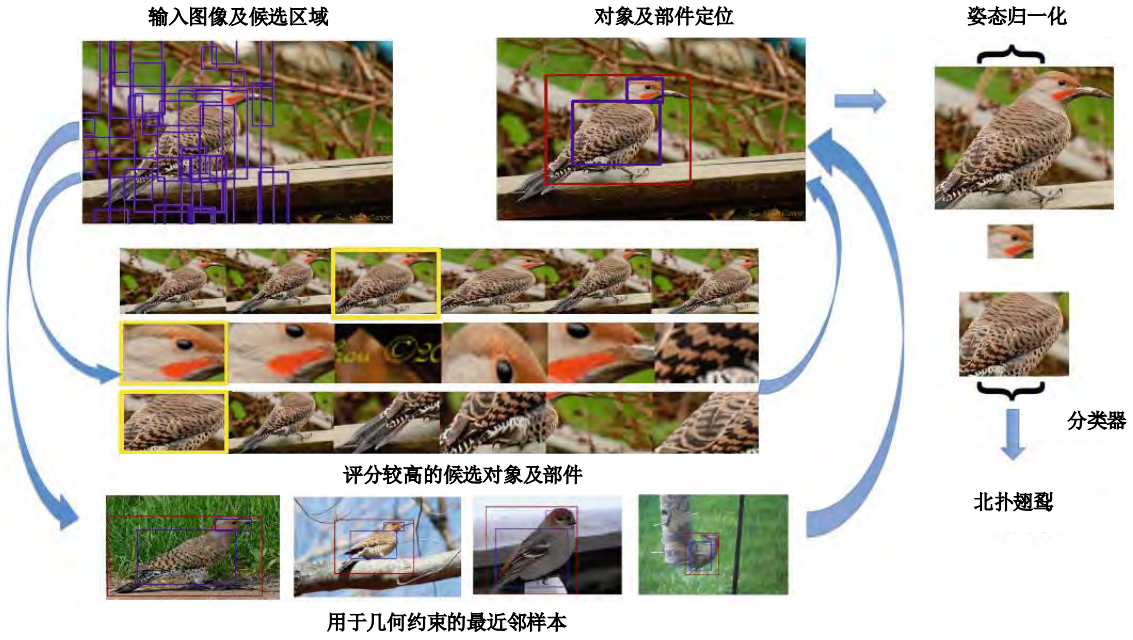


图 2.2　Part-based R-CNN 方法示意图 [5]

计分类网络，分别是对象流和部件流，分别对对象和部件进行特征学习，最后将两者的特征进行拼接输入全连接层并完成细粒度图像分类任务。

上述两种方法不仅使用了对象级标注信息，同时还使用了部件级标注信息。由于部件的标注相比对象要更加繁琐、耗时耗力，Krause 等人[15] 提出了一种基于协同分割和对齐的细粒度图像分类方法，在不使用部件级标注信息的情况下获得部件的自动定位。首先利用协同分割（Co-segmentation）获取图像中的对象区域；然后，提取对象区域的深度特征，构建姿态图（Pose Graph）；进一步，通过对随机点的图传播实现对齐；最后，将这些点扩展为部件区域，从而实现了部件的定位。

从上述方法的分析可以发现，强监督细粒度图像分类方法严重依赖于对象级、部件级的标注信息。然而，这些标注信息耗时耗力，标注成本巨大。Gebru 等人指出构造具有 200 万条标注信息的细粒度图像数据集需要耗费 30 万美元 [6]。由此可见，依赖于标注成本巨大的对象级和部件级标注信息，不利于细粒度图像分类方法走向实际应用。因此，研究者们开始转向弱监督细粒度图像分类方法的研究，**即不使用对象级和部件级的标注信息，仅使用图像级标注信息，大大减少了标注成本。这也是本文工作的第一个目标。**

2.1.1.2　弱监督细粒度图像分类方法

弱监督细粒度图像分类的关键是如何能够不依赖于对象级和部件级的标注信息，自动地定位得到图像中对应的辨识性区域。这已经成为细粒度图像分类任务的主要研

究问题。

Xiao 等人[8] 模拟人类视觉注意力机制，提出了两级注意力模型，在不使用对象级和部件级标注信息的情况下取得了不错的细粒度图像分类效果。而且与一些强监督细粒度图像分类相比，取得了可比的细粒度分类准确率。其主要包含对象级注意力模型和部件级注意力模型。其中，对象级注意力模型通过对由选择搜索方法所生成的图像候选块进行预测，进而根据所对应的类别的预测得分筛选出与类别相关的图像块。这些筛选出的图像块大多包含了对象区域，能够很好地学习对象的细节特征。利用这些图像块训练对象级的分类器；部件级注意力模型则利用了卷积神经网络中间层的卷积核具有特定模式的挖掘属性，通过谱聚类将具有相同模式的卷积核聚合在一起作为部件检测器，以此来获得图像中的部件区域。利用这些部件区域训练部件级分类器。最终，将上述两者的得分融合得到最终的预测结果。受到 Xiao 等人工作的启发，Zhang 等人[7] 同样利用了卷积神经网络中卷积核的特定模式挖掘属性，选取具有强区分性的卷积核作为部件检测器，以此实现部件的定位。但是，这些方法在定位得到辨识性部件后，并未考虑部件间的关联关系。因此 Ge 等人[16] 首先通过互补部件模型定位到一系列的辨识性部件，然后通过 LSTM 将这一系列辨识性部件编码为具有辨识性的特征，从而提高了细粒度图像分类的准确率。

上述方法都是分阶段的，即先定位到辨识性的区域，然后再通过辨识性特征学习实现最终的细粒度分类。为了将定位与特征学习统一到一个框架中，实现端到端的学习，Fu 等人提出递归注意力卷积神经网络 (Recurrent Attention Convolutional Neural Network，简称 RA-CNN)[17]，能够迭代地定位具有辨识性的区域，同时学习多尺度的区域特征表示，从而获得更好的细粒度图像分类准确率。其主要是通过注意力建议子网络（Attention Proposal Sub-Network，简称 APN）由粗到细迭代地生成当前具有辨识性的区域，并作为下一个尺度的输入图像。

弱监督细粒度图像分类方法有效减少了标注的依赖，大大降低了标注成本。但是，这类方法有两个局限性：1）辨识性区域的数目通常依赖于实验验证等人工先验的方式来设置，这大大降低了方法的可用性和可扩展性。**本文工作的第二个目标就是针对这一问题，减少人工先验**；2）忽略了辨识速度，由于定位辨识性区域增加了所用的时间，这也大大减低的方法的实用性。**本文工作的第三个目标就是针对这一问题，提高辨识速度**。

2.1.2 基于编码的方法

一些工作聚焦于特征表示学习, 其主要方法是对卷积神经网络的特征图 (Feature Map) 进行统计编码，以获取更好的特征表示。最具代表性的方法是 Lin 等人[18] 提出的双线性汇合方法 (Bilinear Pooling)。通过计算 CNN 特征图的格拉姆矩阵来捕获特征通

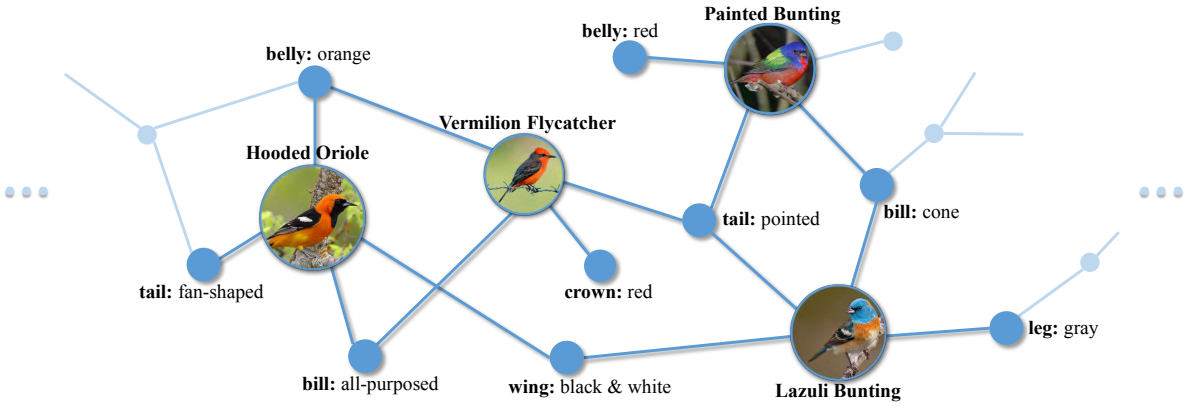


图 2.3　知识图示意图 [24]

道之间成对的相关关系，从而获得更好地特征表示以提升细粒度图像分类准确率。受到双线性汇合方法的启发，Gao 等人[19] 进一步提出了紧凑双线性汇合方法 (Compact Bilinear Pooling)，通过 CNN 特征图低维投影的内积近似二次多项式核来降低双线性汇合方法的高维度。Cui 等人[20] 进一步利用核近似获取更高阶的特征表示。Wang 等人[21] 学习辨识性过滤器，并将其应用到 CNN 中使得其更加关注辨识性特征的学习。

2.1.3　基于属性的方法

上述方法都是依靠图像级类别、对象级或部件级位置等信息来学习图像的辨识性表征，而有一类信息能够直接指出图像的辨识性特征，如红色的鸟喙、灰黑色的翅膀，粉色的脚等，这就是属性标注信息。研究者们希望通过属性标注信息的加入，能够学习到更好的辨识性特征。Zhou 等人 [22] 利用图像级类别标注信息和属性标注信息来构建二分图，并且将二分图的标签建模到卷积神经网络中，以此来学习更好的特征表示。Zhang 等人[23] 通过将层次化、共享属性信息来构建三元组，通过三元组损失函数来学习图像的特征表示。Chen 等人 [24] 基于属性标注信息，提出了一种知识嵌入表示学习方法。首先，他们利用图像的属性标注信息来构建知识图，如图 2.3 所示。其中，顶点分别表示类别和属性，边表示该类别是否具有该属性。然后，通过图神经网络将类别节点的信息在知识图中进行传播，从而学习具有辨识性信息的知识。其能够学习到图像中哪些区域对于最终的细粒度分类预测起决定性作用，从而获得更好的细粒度分类效果。

2.2　细粒度图像检索

近年来，研究者们开始在检索任务上进行细粒度分析的相关研究。Xie 等人[9] 首次提出了细粒度图像检索，将现有的几个细粒度图像数据集以及一般的图像检索数据集

共同搭建了一个具有层次化结构的检索数据集。当输入一个查询样例时，首先判断它属于哪个大类，然后在此大类的数据中再进行细粒度检索。此后，Wei 等人[25] 采用了卷积神经网络，对卷积特征描述子进行选择实现无监督条件下的辨识性区域定位，从而有效提升了细粒度图像检索的准确率。随后，Zheng 等人[26] 提出全局中心排序损失，通过施密特正交化从全局对目标函数进行优化，从而实现细粒度图像检索效果的提升。然而，目前在细粒度图像检索上的工作依旧很少。

2.3 细粒度视频分类

相比于图像，视频通常包含了更丰富的辨识性信息，因此研究者们开始关注细粒度视频分类任务。Saito 等人[27] 构建了一个细粒度视频数据集来探索运动信息在细粒度视频分类中的有效性。Zhu 等人[10] 构建了两个大规模细粒度视频数据集，即 YouTube Birds 和 YouTube Cars 数据集。这两个数据集分别对应了细粒度图像分类中广泛使用的两个数据集：CUB-200-2011[3] 和 Cars-196[28]，他们的类别完全一致。这两个数据集的构建促进了细粒度分析在视频领域的发展，细粒度视频分类工作也相继展开。由于视频帧存在大量冗余信息，Zhu 等人[10] 提出了冗余降低注意力网络来降低 CNN 模型中特征的冗余信息，从而学习得到细粒度的辨识信息。视频分类的准确率与视频帧的采样直接相关，为此 Wu 等人[29] 将视频帧采样过程建模为多并行马尔科夫决策过程，通过提出的多智能体强化学习来解决。Duan 等人[30] 则探讨利用网络数据来训练细粒度视频分类模型的可能性。

2.4 细粒度跨媒体分析

前面介绍的细粒度图像分类/检索、细粒度视频分类，都是对图像或者视频单一媒体进行细粒度分析，学习具有辨识性的特征表示，忽略了与其相关联的其他媒体数据。在现实生活中图像、文本、视频和音频等跨媒体数据经常同时出现，他们之间存在着隐含的语义关联关系，充分发掘这一语义关联关系能够进一步促进细粒度分析的效果。因此，**如何提高跨媒体数据之间的语义关联，从而进一步促进细粒度分析的研究，是一个重要的研究方向。这也是本文工作的第四个目标。**

2.4.1 图像-文本之间的跨媒体分析

文本描述信息能够直接指出图像中最具辨识性的特征，而属性标注信息指出的是图像中对象的所有特征。因此，文本描述能够更好地帮助我们理解图像内容。基于上述想法，He 等人 [31] 将文本描述信息引入到细粒度图像分类，提出了视觉-文本特征表

示学习，通过卷积-循环神经网络来学习图像和文本之间的关联关系，能够对图像和文本进行相似性度量。这样对于一幅图像，既可以得到其图像上的特征表示，也能够得到文本上的特征表示，这二者具有差异性，又具有互补性，能够从多角度对图像进行表示，从而获得更好的细粒度图像分类效果。进一步，Xu 等人 [32] 不仅利用了文本描述信息，也引入了外部知识库（如 WikiPedia），通过两者的嵌入表示学习，实现图像的辨识性特征学习。

2.4.2　视频-音频之间的跨媒体分析

视频是一种天然的跨媒体载体，其中不仅包含了视觉信息还包含了听觉（音频）信息。Zhang 等人[33] 通过挖掘视频中视觉和音频之间的关联关系，联合学习视觉-音频特征，以此获得更好的细粒度视频分类准确率。

上述仅仅是图像-文本、视频-音频两种媒体信息的关联学习，本文在此基础上，进一步探讨图像、文本、视频和音频等 4 种媒体数据之间的关联，以获得更好的细粒度分析效果。

第三章　基于对象-部件注意力模型的细粒度图像分类

3.1　引言

由于细粒度图像分类中类别间的差异主要存在于细小的局部辨识性区域中，如鸟喙、胸部、翅膀、尾巴等。因此，现有方法[5, 7, 34] 一般先定位辨识性区域，即对象及其部件在图像中的位置，然后根据辨识性特征进行细粒度子类别的判别。

为了定位具有辨识性的对象及其部件区域，现有方法通常先通过自底向上的生成方法来获取候选图像块，这些图像块包含了对象及其部件区域。选择性搜索算法（Selective Search）[13] 就是这样一种可以生成上千候选图像块的无监督方法，其广泛应用于现有的细粒度图像分类方法中 [5, 7, 35]。由于自底向上的生成方法具有高召回率、低准确率的特性，因此剔除噪音图像块并保留包含辨识性对象或者部件区域的图像块便成为了一个不可或缺的步骤。而这可以通过自顶向下的注意力模型来实现。在细粒度图像分类中，定位具有辨识性的对象及其部件区域可以被看作两级注意力过程，分别是对象级和部件级。一种自然的想法是在对象级注意力过程中利用对象级标注信息，即对象位置信息（Bounding Box），训练对象检测器来定位图像中的对象区域；在部件级注意力过程中利用部件级标注信息，即部件位置信息（Part Location），训练部件检测器来定位图像中的部件区域。[5, 15, 22, 36] 等方法均采用了如上想法，但是对象级和部件级信息的标注成本是十分巨大的。Gebru 等人指出构造具有 200 万条标注信息的细粒度数据集需要耗费 30 万美元 [6]。由此可见，依赖于成本巨大的标注信息，不利于细粒度图像分类方法向实际应用进行技术转化。

为了解决上述问题，研究者们开始聚焦于如何在弱监督的条件下实现辨识性特征学习，即在训练和测试阶段均不使用对象级和部件级标注信息，自动定位对象及其部件区域并进行特征表达。Zhang 等人 [37] 通过利用部件聚类簇的有用信息来选择具有辨识性的部件。Zhang 等人 [7] 通过合并深度卷积核来进行部件的选择与表达，从而实现弱监督条件下的细粒度图像分类。但是，这些方法在选择辨识性部件的过程中，忽略了对象与其部件之间以及部件相互之间的空间关联关系，从而导致所定位到的辨识性部件具有如下问题：1）辨识性信息少，即所包含的辨识性对象区域面积小，背景区域面积大；2）冗余信息多，即部件相互之间有大面积的重叠，造成了大量的冗余信息，因此可能导致遗漏了真正的辨识性部件。

综上所述，现有方法具有两个局限性：1）依赖对象级和部件级标注信息，这些人工标注成本巨大；2）忽略了对象与部件之间，以及部件相互之间的空间关联关系，而这些空间关联关系对于辨识性部件的定位具有重要作用。因此，本章提出了对象-部件

注意力模型（Object-Part Attention Model，简称 OPAM），通过辨识性对象和部件的特征学习，实现了不使用对象级和部件级标注的弱监督细粒度图像分类。其主要贡献如下：

- **对象-部件注意力模型**：现有方法 [5, 15, 22] 多依赖于成本巨大的对象级或者部件级标注信息，不利于细粒度图像分类方法向实际应用的转化。针对上述问题，本章提出了对象-部件注意力模型以实现弱监督细粒度图像分类，避免了对象级和部件级标注信息的使用，有助于推动细粒度图像分类方法的实际应用。它集合了两级注意力：1）对象级注意力模型利用全局平均池化操作来提取显著图，从而根据显著图的信息定位到图像中的对象区域，实现对象的辨识性特征学习；2）部件级注意力模型首先通过对象-部件空间关联约束选择具有辨识性的部件区域，然后通过对卷积神经网络的卷积核进行谱聚类以实现部件的语义对齐，从而学习更加精细的局部辨识性特征。对象注意力模型聚焦于具有代表性的对象表征，部件级注意力模型聚焦于细粒度子类别之间具有辨识性的部件特征。两者的联合可以进一步促进多视角、多粒度的特征学习，强化对象和部件辨识性特征之间的互补性，从而取得更好的细粒度图像分类准确率。
- **对象-部件空间关联约束**：现有的弱监督细粒度图像分类方法 [7, 37] 忽略了对象与部件之间，以及部件相互之间的空间关联关系。针对上述问题，本章提出了基于对象-部件空间关联约束的部件选择方法，其联合了两种类型的空间关联约束：1）对象空间关联约束确保所选择的部件位于对象区域内，具有较高的代表性；2）部件空间关联约束降低了部件之间重叠区域的面积，根据显著性进行部件的选择。减少了所选择部件的冗余性，增强了其辨识性。两种空间关联约束的结合不仅显著提升了定位的辨识性，而且有效提升了细粒度图像分类的准确率。

3.2 算法描述

本章提出的方法基于人类辨别对象的模式：首先确定辨识性的对象（对象级注意力），然后定位辨识性的部件（部件级注意力）。例如，当识别一张里海燕鸥的图像，首先会找到鸟在这张图像中的位置，然后再仔细观察这个鸟的局部，找到能够与其他细粒度子类别具有区分性的部件，如红色的鸟喙等。如图3.1所示，本章提出了对象-部件注意力模型（OPAM），首先通过对象级注意力模型定位图像中的对象区域以学习对象特征，然后通过部件级注意力模型选择具有辨识性的部件以学习精细的局部特征。

3.2.1 对象级注意力模型

现有弱监督细粒度图像分类方法 [7, 37, 38] 致力于辨识性部件的定位与选择，却忽视了对象的定位。而对象定位可以有效避免背景信息对于细粒度分类的错误影响，能够学

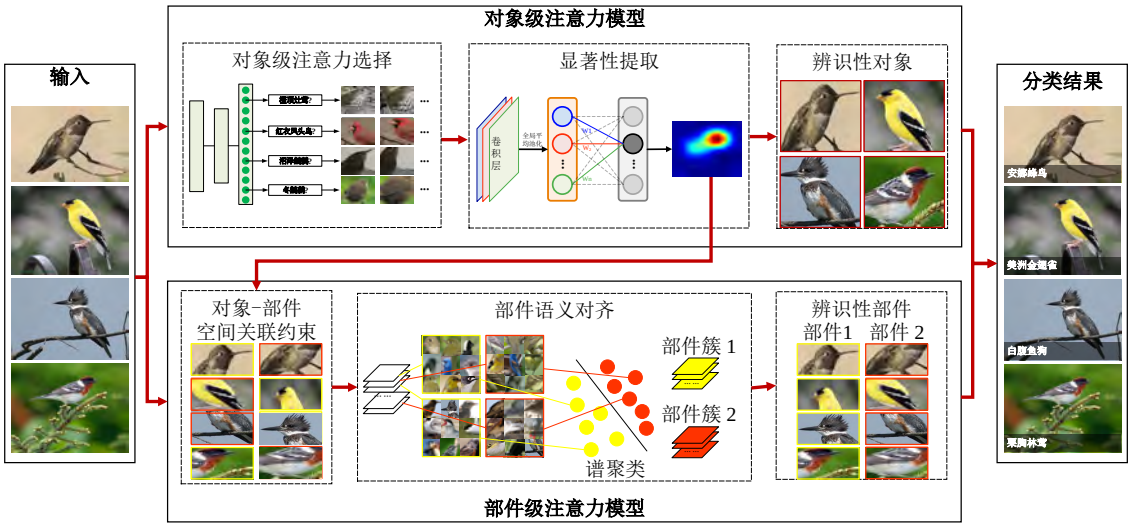


图 3.1　对象-部件注意力模型的总体框架

习更加有用且具有代表性的对象特征。尽管有一些方法同时考虑了对象定位和部件定位，但是他们依赖于成本巨大的对象级和部件级标注信息[4, 5]。针对上述问题，本章提出了对象级注意力模型，自动定位图像中的辨识性对象区域。而且这一模型只在训练阶段使用了图像级的类别标注信息，并不依赖于对象级标注信息。该模型包含两个组成模块：对象级注意力选择和显著性提取。其中，对象级注意力选择的目的是通过选择包含对象区域的图像块，以扩充训练样本，从而训练得到卷积神经网络模型 *ClassNet*。这样可以学习多视角、多尺度的辨识性特征。显著性提取的目的是通过卷积神经网络中的全局平均池化操作来提取图像的显著图，从而根据显著性定位到图像中的对象区域。

3.2.1.1　对象级注意力选择

卷积神经网络能够在计算机视觉领域获得巨大的成功，其中一个重要的原因是有大量的训练数据供其学习。因此，我们首先聚焦于如何扩充训练数据。自底向上的生成方法通过像素点的聚合可以产生上千个可能包含对象区域的候选图像块。这些候选图像块根据其与对象区域的相关性可以被用作扩充的训练数据。因此，选择性搜索（Selective Search）方法[13]被用来生成候选图像块。这些图像块提供了多视角、多尺度的信息，有利于训练得到一个有效地卷积神经网络模型，从而获得更好的细粒度图像分类结果。但是，由于选择性搜索方法具有高召回率、低准确率的特性，这些候选图像块并不能直接用来扩充训练数据。而对象级注意力模型有助于选择与对象区域相关性高的图像块。

我们通过一个预训练的卷积神经网络模型，即 *FilterNet*，来过滤掉包含大面积背景

区域的图像块，保留包含大面积对象区域的图像块。其中，FilterNet 先利用 ImageNet 1K 数据集 [39] 进行预训练，然后在利用特定的细粒度图像数据集进行微调（Fine-tune)，如 CUB-200-2011 数据集 [3]。将 FilterNet 最后一层（即 Softmax 层）中每个神经元的响应值作为选择置信度得分，其表示该候选图像块与对应对象细粒度子类别的相关性。我们设置一个置信度阈值来判断该图像块是否被选择。经过上述过程，可以获得具有多视角、多尺度特性的包含对象区域的图像块。最后，利用这些得到的图像块来训练得到卷积神经网络 *ClassNet*。ClassNet 对于本章提出的对象-部件注意力模型具有两方面的好处：（1）ClassNet 本身能够获得较好的细粒度图像分类准确率；（2）ClassNet 网络的中间层特征对于部件级注意力模型构建部件检测器实现部件语义对齐有非常重要的作用。值得注意的是，这些选择得到的图像块仅在训练阶段使用，而且只使用了图像级的类别标注信息。

3.2.1.2 显著性提取

在这一模块中，CAM [40] 被用来获取图像对应细粒度子类别 c 的显著图 M_c，从而根据显著信息来定位图像中的对象区域。显著图 M_c 表明了图像中具有辨识性的区域，根据这些区域卷积神经网络将图片判定为某个细粒度子类别，如图3.2中每个子图的第二行所示。然后，对显著图进行二值化和最大连通域提取操作以获得图像中的对象区域，如图3.2中每个子图的第三行所示。具体地，对于给定的输入图像 I，用 $f_u(x,y)$ 表示最后一层卷积层上神经元 u 对于空间位置 (x,y) 的激活响应，w_u^c 表示细粒度子类别 c 对于神经元 u 的权值，则显著图中对应空间位置 (x,y) 的值的计算公式为：

$$M_c(x,y) = \sum_u w_u^c f_u(x,y) \tag{3.1}$$

其中，$M_c(x,y)$ 表示了空间位置 (x,y) 的激励响应对于将图像分类到细粒度子类别 c 的重要程度。在此过程中，并没有使用图像级的类别标注信息，而是直接使用了 ClassNet 对于该图像的预测结果。

通过对象级注意力模型，定位得到图像中辨识性对象的位置区域，然后利用对象区域的图像来训练得到卷积神经网络 *ObjectNet*，从而获得最终的对象级的预测结果。

3.2.2 部件级注意力模型

由于具有辨识性的部件，如鸟喙、胸部、翅膀、尾巴等，对于细粒度图像分类是十分重要的，因此现有工作 [5] 通常从候选图像块中选择具有辨识性的图像块作为部件。但是，这些工作严重依赖于成本巨大的部件级标注信息。尽管有一些工作 [7, 8] 开始聚焦于不使用部件级标注信息来获取辨识性部件，但他们忽略了对象与部件之间，以及部件相互之间的空间关联关系。

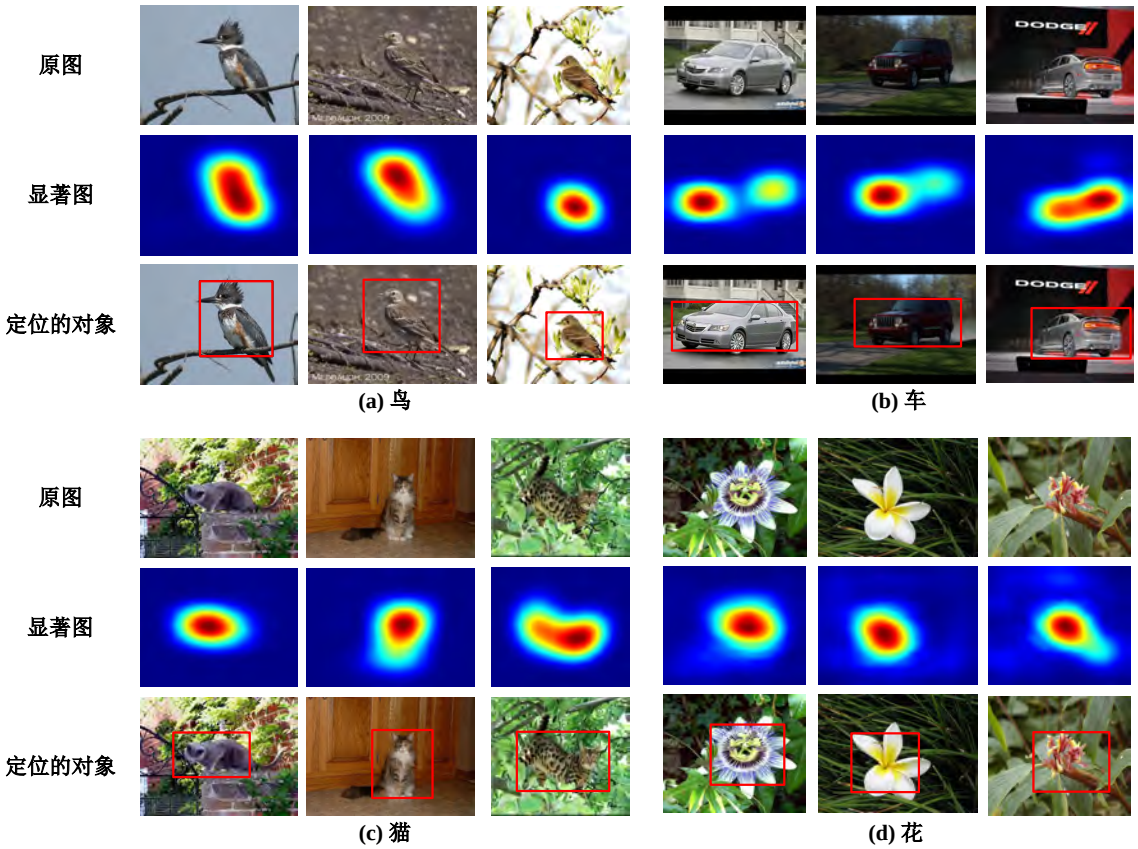


图 3.2　本章对象-部件注意力模型中显著性提取结果的展示

因此，本章提出了一种基于部件级注意力的部件选择方法，能够定位发现图像中精细的局部辨识性区域，从而与其他细粒度子类别区分开来。其并不依赖对象级和部件级标注信息，大大减少了标注成本。其包括对象-部件空间关联约束和部件语义对齐，前者用于选择具有辨识性的部件，后者用于对辨识性部件进行语义对齐。

3.2.2.1　对象-部件空间关联约束

两种空间关联关系被联合考虑：（1）对象空间关联约束定义了对象和部件之间的空间关联关系；（2）部件空间关联约束定义了部件相互之间的空间关联关系。对于给定的输入图像 I，其对应细粒度子类别 c 的显著图 M_c 以及对象区域 b 已经通过对象级注意力模型获得。然后，基于对象-部件空间关联约束的部件选择过程如下：

我们用 $\mathbb{P}$ 表示所有候选图像块的集合，$P = \{p_1, p_2, ..., p_n\}$ 表示从 $\mathbb{P}$ 选择出的 n 个图像块，作为辨识性部件。候选图像块是通过选择性搜索算法来获取的。对象-部件空间关联约束模型通过求解下列优化问题来同时考虑两种约束信息：

$$P^* = arg\max_{\mathbb{P}} \Delta(P) \tag{3.2}$$

其中，$\Delta(P)$ 表示两种空间关联约束的得分函数，其定义如下：

$$\Delta(P) = \Delta_{box}(P)\Delta_{part}(P) \tag{3.3}$$

其确保了所选部件的代表性和辨识性，从而确保了细粒度图像分类的有效性。其包括对象空间关联约束 $\Delta_{box}(P)$ 和部件空间关联约束 $\Delta_{part}(P)$。所选择的部件需同时满足上述两种约束。

对象空间关联约束： 忽略对象与部件之间的空间关联关系可能会导致所选择的部件包含大面积的背景区域，却仅有较小面积的对象区域，这使得所选择的部件的辨识性较差。由于辨识性的部件一定位于对象区域里面，一种直接的空间关联约束定义如下：

$$\Delta_{box}(P) = \prod_{i=1}^{n} f_b(p_i) \tag{3.4}$$

$$f_b(p_i) = \begin{cases} 1, & IoU(p_i) > \tau \\ 0, & otherwise \end{cases} \tag{3.5}$$

其中，$IoU(p_i)$ 定义了对象与部件区域之间交集与并集的比值（Intersection-over-Union，简称 IoU），τ 表示设定的 IoU 阈值。值得注意的是，对象区域是通过对象级注意力模型自动定位得到的，而不是由对象级标注信息提供。

部件空间关联约束： 忽略部件相互之间的空间关联关系可能会导致所选择的部件有大面积重叠区域，一些具有辨识性的部件也因此而被忽略。显著图表示了图像的辨识性信息，有利于选择具有辨识性的部件。因此，我们将显著图信息和空间关联关系联合建模：

$$\begin{aligned} \Delta_{part}(P) = & \, log(A_U - A_I - A_O) \\ & + log(Mean(M_{A_U})) \end{aligned} \tag{3.6}$$

其中，A_U 是 n 个部件区域的并集，A_I 是 n 个部件区域的交集，A_O 是指在对象区域以外的区域，$Mean(M_{A_U})$ 的定义如下:

$$Mean(M_{A_U}) = \frac{1}{|A_U|} \sum_{i,j} M_{ij} \tag{3.7}$$

其中，像素点 (i, j) 位于所选部件并集的区域中，M_{ij} 表示像素点 (i, j) 在显著图中的对应显著值，$|A_U|$ 则表示所选部件并集中的像素点个数。

部件空间关联约束旨在选择最具辨识性的部件，其第一项约束是为了降低所选部件之间的重叠度，通过 $log(A_U - A_I - A_O)$ 来实现。其中，$-A_I$ 确保所选部件之间有最

小的重叠面积，$-A_O$ 确保所选部件能够最大程度的覆盖对象区域；第二项约束是为了最大化所选部件的辨识性，通过 $log(Mean(M_{A_U}))$ 来实现，其表示所选部件并集区域中所有像素点所对应的的显著值的均值。

当一组部件能够同时满足公式（3.4) 和公式3.6并使得公式3.3取得最大值时，则是最终所选的部件。

3.2.2.2　部件语义对齐

通过对象-部件空间关联约束所选择的部件是无序的，并未依据其语义信息进行对齐，如图3.3(a) 所示。所选择的部件因为具有不同的语义信息，因此对于最终的预测有不同的贡献。鉴于此，我们需要将所选部件依据语义信息进行对齐，如图3.3(b) 所示。

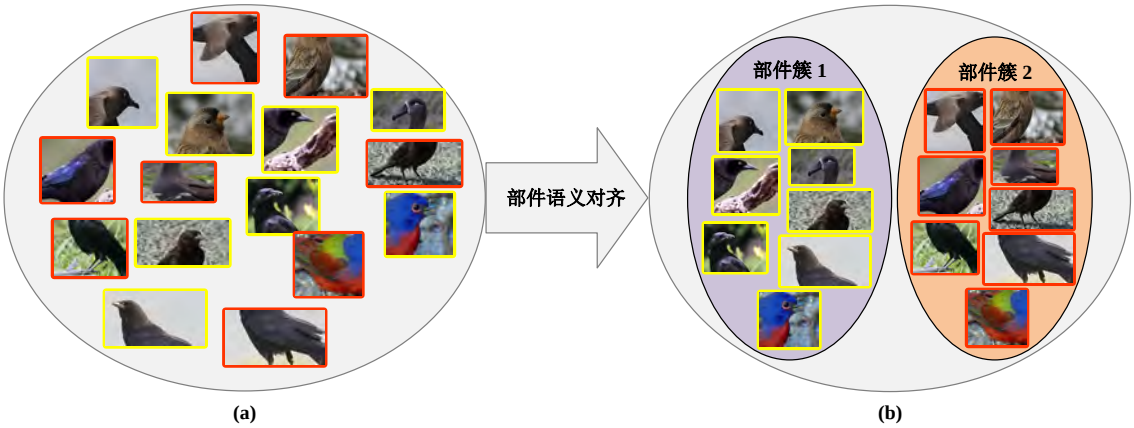


图 3.3　本章 OPAM 方法部件语义对齐结果展示

在对象级注意力模型中得到的 ClassNet 的中间层具有特定的聚类模式，例如，有一些卷积核会对鸟的头部具有较强的激励响应，而另一些卷积核则对鸟的躯干具有较强的激励响应。受到此启发，我们对 ClassNet 中间层的卷积核进行聚类，构建部件检测器以实现部件的语义对齐。我们首先，计算得到相似矩阵 S，其表示为中间层卷积核 u_i、u_j 两者权重的余弦相似度（Cosine Similarity）。然后，通过在相似矩阵上进行谱聚类将卷积核聚类为 m 组。在本章实验中，我们选取最后一层卷积层的卷积核，并且将 m 设置为 2。如3.4所示，其中坐标表示相似矩阵两个最大的特征值。

基于上述得到的卷积核组，我们通过以下步骤对所选部件进行语义对齐：（1）将图像中对应所选部件的区域进行裁剪，并缩放到最后一层卷积核感受野所对应的大小；（2）将步骤（1）中得到的图像作为卷积神经网络的输入，得到最后一层卷积层卷积核的激励响应；（3）将对应卷积核组的响应值相加；（4）将所选部件分配到响应值和最大的卷积核组。经过上述过程，所选部件完成了语义对齐，具有相同语义的部件聚集在一起，从而使得细粒度图像分类的准确率进一步提升。

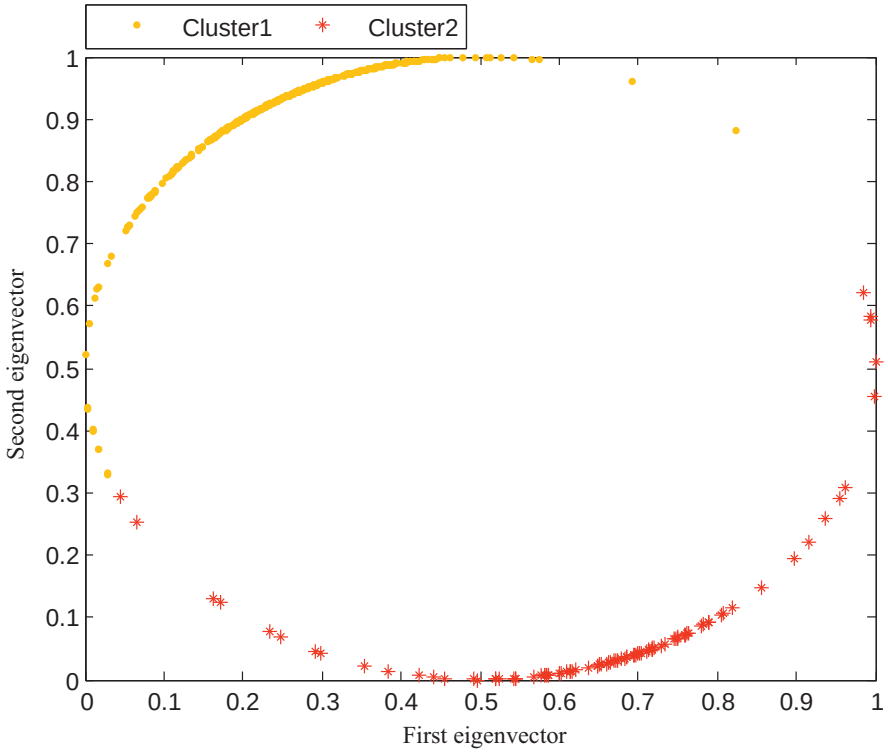


图 3.4　谱聚类示意图

3.2.3　最终预测

为了获得更好的细粒度图像分类效果，我们分别用自动定位得到的对象及其部件区域的图像对 *ClassNet* 进行微调（Fine-tune）获得 *ObjectNet* 和 *PartNet* 两个网络模型，以学习到具有辨识性的特征。*ClassNet*、*ObjectNet* 和 *PartNet* 均为细粒度图像分类器：*ClassNet* 对应原始图像，*ObjectNet* 对应对象，*PartNet* 则对应所选择的辨识性部件。但是，它们对于最终预测结果的影响不同，因为它们聚焦于图像的不同信息。

对象级注意力模型首先驱动 FilterNet 来选择具有多视角、多尺度的与对象相关的图像块，如图3.5 (a) 所示。这些图像块使得 *ClassNet* 学习得到更具代表性的对象特征，并且通过显著性提取定位到图像中的对象区域。部件级注意力模型选择具有辨识性的部件，一般位于图像的局部区域，如图3.5 (b) 所示。不同的注意力（即图像级、对象级、部件级）有不同的聚焦点，能够学习到不同的辨识性特征，但他们相互之间又具有很强的互补性。因此，最终我们将他们的预测结果进行融合，融合方式如下所示：

$$S = \alpha * S_{ori} + \beta * S_{obj} + \gamma * S_{par} \tag{3.8}$$

其中 S_{ori}、S_{obj} 和 S_{par} 分别表示 *ClassNet*、*ObjectNet* 和 *PartNet* 三者的预测结果。α，β 和 γ 通过 k-折交叉验证方法（k-fold cross-validation）来设定。

(a) 对象级注意力模型　　(b) 部件级注意力模型

图 3.5　本章对象级注意力模型和部件级注意力模型所选择的图像块对比

3.3　实验结果与分析

本章在四个广泛使用的细粒度图像分类数据集上进行实验，与 10 多个现有方法进行对比，来验证本章对象-部件注意力模型的有效性。

3.3.1　实验数据集和评价指标

- **CUB-200-2011 数据集** [3]：这是最广泛使用的细粒度图像分类数据集，包含 200 个鸟类的细粒度子类别和 11,788 张图片。其中，训练集包含 5,994 张图片，测试集包含 5,794 张图片。对于每一张图片，有 4 种人工标注信息：1 个图像级的类别标签、1 个对象级位置信息（Bounding Box）、15 个部件级位置信息（Part Location）以及 312 个属性信息。本章方法只使用了图像级的类别标签这一种人工标注信息，其余三种标注信息并未使用。数据集样例如图 3.6 (a) 所示。
- **Cars-196 数据集** [28]：它包含 196 个车类的细粒度子类别，16,185 张图片。其中，训练集 8,144 张图片，测试集 8,041 张图片。对于每一张图片，有 2 种标注信息：1 个图像级的类别标签和 1 个对象级位置信息。数据集样例如图 3.6 (b) 所示。
- **Oxford-IIIT Pet 数据集** [41]：它包含 37 个宠物的细粒度子类别，7,349 张图片。其中，训练集 3,680 张图片，测试集 3,669 张图片。对于每一张图片，有 3 种标注信息：1 个图像级的类别标签、1 个对象头部的位置信息以及身体部位的像素

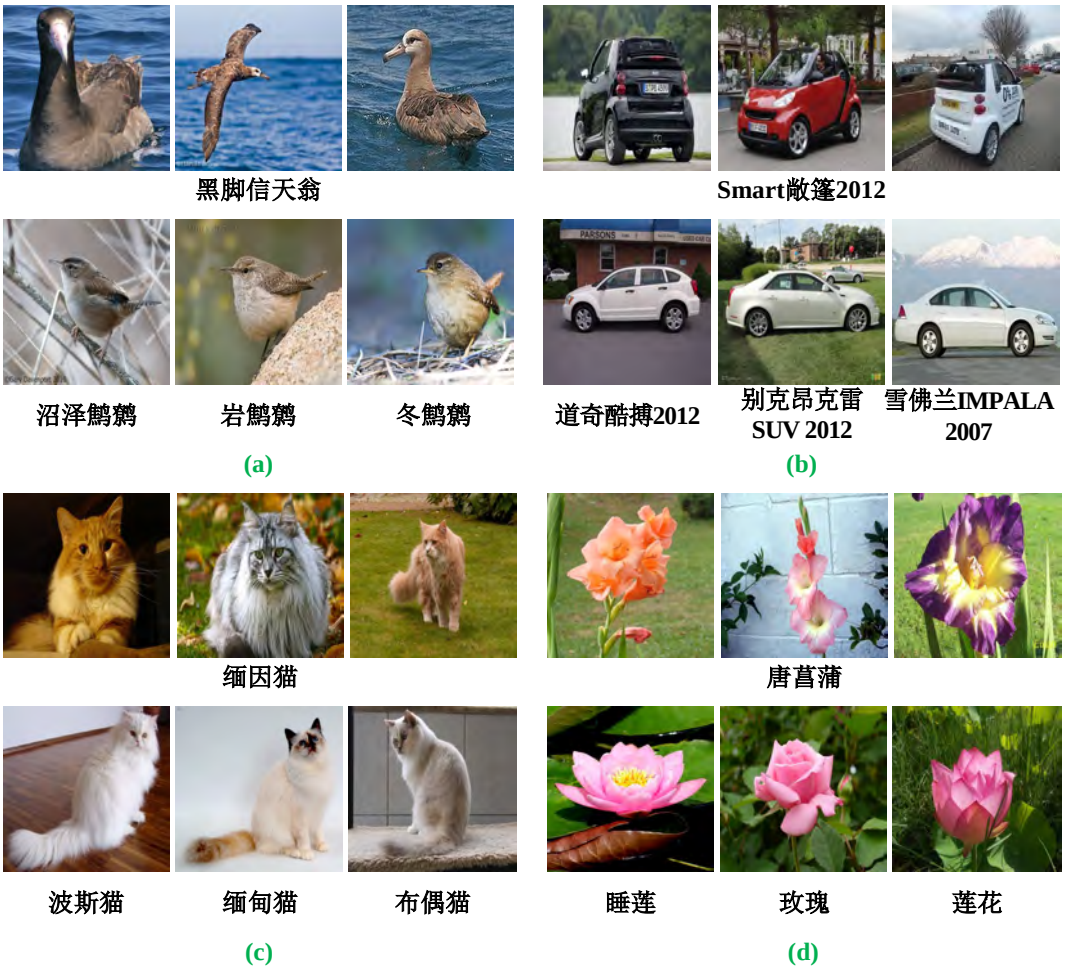


图 3.6　CUB-200-2011 [3]，Cars-196 [28]，Oxford-IIIT Pet [41] 和 Oxford-Flower-102 [42] 数据集的样例

级标签。数据集样例如图 3.6 (c) 所示。

- **Oxford-Flower-102 数据集** [42]：它包含 102 个花类的细粒度子类别，8,189 张图片。其中，训练集 1,020 张图片，验证集 1,020 张图片，测试集 6,149 张图片。每一张图片只有 1 个图像级的类别标签。数据集样例如图 3.6 (d) 所示。

本章采用准确率（Accuracy）来评价对象-部件注意力模型的有效性，其定义如下：

$$Accuracy = \frac{R_a}{R} \tag{3.9}$$

其中，R 表示测试集中所有样例的数目，R_a 表示测试集中正确分类的样例数目。

3.3.2　实验设置

在本章实验中，采用 19 层 VGGNet [43] 作为基础网络模型。需要注意的是，该基础网络模型可以替换为任意卷积神经网络模型。在本章对象-部件注意力模型中，无论是

辨识性区域的定位还是细粒度分类，都与基础网络模型有关。

- 定位：在对象级注意力模型中，*ClassNet* 将用来提取每张图像的显著图以自动定位辨识性对象区域。Zhou 等人[40] 发现，如果最后一层卷积层有一个较高的空间分辨率，则能够取得更加准确的定位准确率。因此，在本章实验中，将 VGGNet 模型 conv5_3 以后的层剔除，从而获得 14×14 的空间分辨率。此外，增加一层卷积层，其包含 1024 个卷积核，核大小为 3×3，步长为 1，填充（Padding）为 1。其后紧跟全局平局池化层和 Softmax 层。修改后的 VGGNet 首先在 ImageNet 1K 数据集 [39] 上预训练，之后再在细粒度图像数据集上进行微调，最后再利用对象级注意力选择的图像块作为扩充数据进一步微调，以得到 *ClassNet*。
- 细粒度分类：前述得到的 *ClassNet* 用来对原始图像进行识别。对于自动定位到的辨识性对象和部件则分别用 *ObjectNet* 和 *PartNet* 来进行识别。他们分别在对象级注意力和部件级注意力所定位到的对象和部件区域的图像数据上进行了微调。最后，通过公式3.8来获得最终的预测结果。与[5] 类似，在本章实验中通过 k-折交叉验证方法（k-fold cross-validation）来设定 (α, β, γ)。最终，在四个数据集上的 (α, β, γ) 分别设定为 $(0.4, 0.4, 0.2)$，$(0.5, 0.3, 0.2)$，$(0.4, 0.4, 0.2)$ 和 $(0.4, 0.3, 0.3)$。

3.3.3　与现有方法进行对比

本节展示了本章对象-部件注意力模型（表示为 OPAM）以及现有方法在上述 4 个细粒度图像分类数据集上的结果与实验分析，以验证本章 OPAM 方法的有效性。表 3.1展示了本文 OPAM 方法在 CUB-200-2011 数据集上的对比结果。使用的对象级标注、部件级标注、CNN 网络模型等信息也列出，以进行公平对比。

从表 3.1 中可以发现：早期工作 [48, 54, 55] 采用 SIFT 特征 [56] 来表示图像，与本章 OPAM 方法相比，它们细粒度分类准确率都非常低。而且[54, 55] 还使用了对象级和部件级的标注信息。在训练和测试阶段均不使用对象级和部件级标注信息的条件下，本章 OPAM 方法取得了最好的细粒度分类准确率，与对比方法中最好的 FOAF 方法 [34] 相比高 1.20%（85.83% vs. 84.63%）。值得注意的是，在预训练阶段，FOAF 方法不仅使用了 ImageNet 1K 数据集 [39]，还使用了 PASCAL VOC 数据集 [57]，而本章 OPAM 方法仅使用 ImageNet 1K 数据集。PD 方法 [7] 的细粒度分类准确率在对比方法中排第二名，但是与本章 OPAM 方法相比要低 1.29%（85.83% vs. 84.54%）。

此外，本文 OPAM 方法与聚焦于 CNN 网络结构的细粒度图像分类方法（例如，STN 方法 [44] 和 Bilinear-CNN 方法 [18]）相比也取得了更好的细粒度图像分类准确率。在 STN 方法中采用带批标准化（Batch Normalization) [71] 的 GoogleNet [72] 作为基础网络，其直接利用 CUB-200-2011 数据集的训练数据进行微调便可以取得 82.30% 的细粒度图像分类准确率。在 Bilinear-CNN 方法中采用了两种网络结构：VGGNet [43] 和 VGG-M

表 3.1　CUB-200-2011 数据集上实验结果

方法	训练集标注		测试集标注		准确率 (%)	CNN
	对象级	部件级	对象级	部件级		
本章 OPAM 方法					**85.83**	VGGNet
FOAF [34]					84.63	VGGNet
PD [7]					84.54	VGGNet
STN [44]					84.10	GoogleNet
Bilinear-CNN [18]					84.10	VGGNet&VGG-M
Multi-grained [45]					81.70	VGGNet
NAC [38]					81.01	VGGNet
PIR [37]					79.34	VGGNet
TL Atten [8]					77.90	VGGNet
MIL [46]					77.40	VGGNet
VGG-BGLm [22]					75.90	VGGNet
InterActive [47]					75.62	VGGNet
Dense Graph Mining [48]					60.19	
Coarse-to-Fine [49]	√				82.50	VGGNet
Coarse-to-Fine [49]	√		√		82.90	VGGNet
PG Alignment [15]	√		√		82.80	VGGNet
VGG-BGLm [22]	√		√		80.40	VGGNet
Triplet-A (64) [50]	√		√		80.70	GoogleNet
Triplet-M (64) [50]	√		√		79.30	GoogleNet
Webly-supervised [51]	√	√			78.60	AlexNet
PN-CNN [36]	√	√			75.70	AlexNet
Part-based R-CNN [5]	√	√			73.50	AlexNet
SPDA-CNN [52]	√	√	√		85.14	VGGNet
Deep LAC [53]	√	√	√		84.10	AlexNet
SPDA-CNN [52]	√	√	√		81.01	AlexNet
PS-CNN [14]	√	√	√		76.20	AlexNet
PN-CNN [36]	√	√	√	√	85.40	AlexNet
Part-based R-CNN [5]	√	√	√	√	76.37	AlexNet
POOF [54]	√	√	√	√	73.30	
HPM [55]	√	√	√	√	66.35	

[73]。这两个方法或采用了更强的基础网络或采用了两个基础网络，但它们的细粒度图像分类准确率均为 84.10%，比本章 OPAM 方法低 1.73%。这说明即使本章 OPAM 采用了较弱的基础网络，但是通过对象-部件级注意力能够挖掘图像中具有辨识性的区域并进行有效表达，很好地提升了细粒度分类的准确率。

进一步，本章 OPAM 方法与 Coarse-to-Fine [49]、PG Alignment [15] 和 VGG-BGLm [22] 等使用对象级标注信息的方法相比也取得了更好的细粒度图像分类准确率。甚至，与同时使用对象级和部件级标注信息的方法 [5, 52] 方法相比，本章 OPAM 方法也能取得更好的细粒度图像分类准确率。避免使用成本巨大的对象级和部件级标注信息有助于细粒度图像分类方法向实际应用转化。

表 3.2　Cars-196 数据集上实验结果

方法	训练集标注		测试集标注		准确率 (%)	CNN
	对象级	部件级	对象级	部件级		
本章 OPAM 方法					**92.19**	VGGNet
Bilinear-CNN [18]					91.30	VGGNet&VGG-M
TL Atten [8]					88.63	VGGNet
DVAN [58]					87.10	VGGNet
FT-HAR-CNN [59]					86.30	AlexNet
HAR-CNN [59]					80.80	AlexNet
PG Alignment [15]	√				92.60	VGGNet
ELLF [60]	√				73.90	CNN
R-CNN [35]	√				57.40	AlexNet
PG Alignment [15]	√		√		92.80	VGGNet
BoT(CNN With Geo) [61]	√		√		92.50	VGGNet
DPL-CNN [62]	√		√		92.30	VGGNet
VGG-BGLm [22]	√		√		90.50	VGGNet
LLC [63]	√		√		69.50	
BB-3D-G [28]	√		√		67.60	

表 3.3　Oxford-IIIT Pet 数据集上实验结果

方法	准确率 (%)	CNN
本章 OPAM 方法	**93.81**	VGGNet
InterActive [47]	93.45	VGGNet
TL Atten [8]	92.51	VGGNet
NAC [38]	91.60	VGGNet
FOAF [34]	91.39	VGGNet
ONE+SVM [64]	90.03	VGGNet
Deep Optimized [65]	88.10	AlexNet
NAC [38]	85.20	AlexNet
MsML+ [66]	81.18	CNN
MsML [66]	80.45	CNN
Deep Standard [65]	78.50	AlexNet
Shape+Appearance [41]	56.68	
Zernike+SCC [67]	59.50	
GMP+p [68]	56.80	
GMP [68]	56.10	
M-HMP [69]	53.40	
Efficient Object Detection [70]	54.30	

在 Cars-196、Oxford-IIIT Pet 和 Oxford-Flower-102 三个数据集上的对比结果如表 3.2 - 3.4 所示。由表可见，与现有方法细粒度图像分类准确率的对比趋势与 CUB-200-2011 数据集一致，本章 OPAM 方法都取得了最好的细粒度图像分类准确率，在三个数据集上分别取得了 92.19%、93.81% 和 97.10% 的准确率，并与对比方法中最高的结果

表 3.4　Oxford-Flower-102 数据集上实验结果

方法	准确率 (%)	CNN
本章 OPAM 方法	**97.10**	VGGNet
InterActive [47]	96.40	VGGNet
PBC [74]	96.10	GoogleNet
TL Atten [8]	95.76	VGGNet
NAC [38]	95.34	VGGNet
RIIR [75]	94.01	VGGNet
Deep Optimized [65]	91.30	AlexNet
SDR [65]	90.50	AlexNet
MML [66]	89.45	CNN
CNN Feature [76]	86.80	CNN
Generalized Max Pooling [68]	84.60	
Efficient Object Detection [70]	80.66	

表 3.5　本章 OPAM 方法每个模块在 CUB-200-2011、Cars-196、Oxford-IIIT Pet 和 Oxford-Flower-102 四个数据集上的结果

方法	准确率 (%)			
	CUB-200-2011	Cars-196	Oxford-IIT Pet	Oxford-Flower-102
本章 OPAM 方法 (Original+Object-level+Part-level)	**85.83**	**92.19**	**93.81**	**97.10**
Original	80.82	86.79	88.14	94.70
Object-level	83.74	88.79	90.98	95.32
Part-level	80.65	84.26	85.75	93.09
Original+Object-level	84.79	91.15	92.20	96.55
Original+Part-level	84.41	91.06	91.82	96.23
Object-level+Part-level	84.73	89.69	91.50	95.66

相比分别提高了 0.89%、0.36% 和 0.70%。

在四个数据集上的对比实验验证了本章 OPAM 方法的有效性，这是由于：1）对象级注意力与部件级注意力的联合，不仅促进了多视角、多粒度的辨识性特征学习，同时还增强了两者之间的互补促进；2）对象-部件空间关联约束的提出，能够更好地挖掘细粒度子类别之间具有辨识性的特征。

3.3.4　基线实验

3.3.4.1　对象级注意力模型和部件级注意力模型的有效性

在本章 OPAM 方法中，最终的预测得分是通过融合 3 种不同图像的预测得分，即原始图像、对象区域的图像以及部件区域的图像，分别表示为 “Original’’、“Object-level’’ 和 “Part-level’’。首先看一下对象级注意力模型和部件级注意力模型的有效性。从表3.5、图3.7和3.8可以发现：

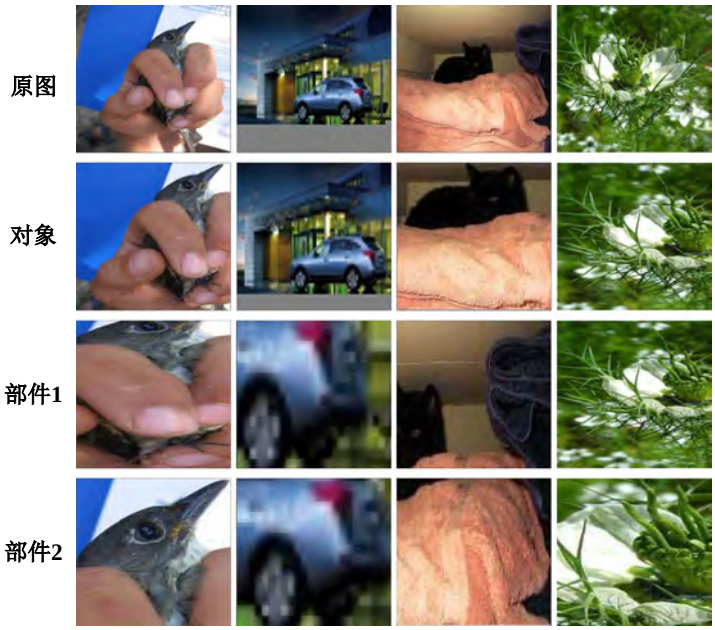


图 3.7　部件定位的失败例子

- 对象级注意力模型通过定位图像中的辨识性对象区域，学习具有代表性的对象特征来提升细粒度图像分类准确率。与使用原图的结果 “Original’’ 相比，分别在四个数据集上提升 2.92%、2.00%、2.84% 和 0.62%。而且联合使用原图和对象区域的图像，准确率能够进一步得到提升，即 3.97%、4.36%、4.06% 和 1.85%。这证明本章的对象级注意力模型能够提取出图像中具有辨识性的对象区域，降低背景区域对于细粒度分类的错误影响。
- 仅使用部件级注意力模型的结果不如使用原图的结果高。图3.7展示了部件选择失败的例子。我们可以总结出本章 OPAM 方法的部件定位可能会在下列两种情况下失效：1）对象很难从图像的背景中区分出来；2）图像中的对象有严重的遮挡。在这两种情况下很难精准地定位到对象区域，因此基于对象定位的部件选择也会容易失效。部件选择失效是只使用部件区域结果低的一个原因。另一个原因是部件注意力聚焦于对象精细的局部特征，相比于原始图像包含了较少的信息。但是，尽管存在上述具有挑战的情况，“Part-level’’ 依然取得了不错的结果，甚至比一些现有方法的结果还高，例如[22, 46]。此外，部件区域与原图之间存在互补关系，联合二者可以取得更好的细粒度图像分类准确率。
- 相比单个注意力，联合对象级和部件级注意力能够取得更好的准确率，例如在 CUB-200-2011 数据集上 84.73% vs. 83.74% 和 80.65%。进一步再联合使用原图，能够相比原图提升更多，即在四个数据集上分别提升 5.01%、5.40%、5.67% 和 2.4%。上述结果表明对象级和部件级注意力之间存在很强的互补关系。两种注

图 3.8　本章 OPAM 方法对象定位和部件定位的结果

意力有不同但却互补的聚焦点：对象级注意力模型聚焦在对象的代表性区别上；部件级注意力模型聚焦在对象精细的具有辨识性的局部部件上，依靠这些辨识性部件能够与其他相似的细粒度子类别区分开。二者的联合使用能够促进多视角、多粒度的特征学习，强化他们之间的互补关系，进一步取得更好的细粒度图像分类准确率。

- 从表3.5中可以发现 “Original+Part-level’’ 的结果要比 “Object-level+Part-level’’ 的结果好，这是因为：1）部件是基于对象区域所选择的，而且依据对象-部件空间关联约束所选择的部件都位于对象区域，并且能够最大程度的覆盖对象区域。

这就导致对象与部件之间相比原图与部件之间的互补性要弱一点。2）对象定位可能会出错，导致定位得到的对象区域并不包含整个对象区域。而被忽略掉的对象区域可能对于最终的细粒度分类很有帮助。但这些区域是包含在原图中的。3）原图包含了一些背景信息，在一定程度下背景也会有助于细粒度分类。所以，综上所述，“Original+Part-level’’ 相比 “Object-level+Part-level’’ 能够提供更加互补的信息。但是，“Original+Object-level+Part-level’’ 能够取得更好的结果，综合了原图、对象、部件三者之间的互补性。

- 图3.8展示了本章 OPAM 方法对象定位和部件定位的结果。第一行是原始图像，第二行是通过对象级注意力模型定位得到的对象区域，最后两行是通过部件级注意力模型定位的辨识性部件区域。对于 CUB-200-2011、Cars-196 和 Oxford-IIIT Pet 三个数据集，所选的部件具有一定的语义，如第三行表示对象的头部（车头）、第四行表示对象的躯干（车身）。对于 Oxford-Flower-102 数据集，其包含两种类型的图像：一种是只包含一朵花，另一种是包含一片花。对于只包含一朵花的图像，对象即花这个个体，部件表示花具有辨识性的局部区域，如花瓣、花蕾或花托。对于包含一片花的图像，对象可能是最显著的一朵花，或者是图像中的整个花丛；部件则表示一朵花的辨识性区域或者花丛中的一朵花。本文 OPAM 方法对于这两种情况都有不错的效果，能够定位到具有辨识性的区域，从而获得不错的细粒度图像分类准确率。在本章 OPAM 方法中，仅使用了图像级标注信息，大大降低了标注成本，促使细粒度图像分类算法向实际应用迈进。

表 3.6　对象-部件空间关联约束和部件语义对齐结果

方法	准确率 (%)			
	CUB-200-2011	Cars-196	Oxford-IIIT Pet	Oxford-Flower-102
OPSC+PA	**80.65**	**84.26**	**85.75**	**93.09**
OPSC	79.74	83.34	83.46	92.33
PA	65.41	68.32	75.42	88.75

3.3.4.2　对象-部件空间关联约束与部件语义对齐的有效性

文献[8] 方法只考虑了部件对齐，本章 OPAM 方法进一步考虑了对象-部件空间关联约束，以选择更好的辨识性部件。对象级空间关联约束能够确保所选部件具有较高的代表性，部件级空间关联约束能够减少所选部件的冗余性，增强其辨识性。两者的结合可以定位发现能够区分相似细粒度子类别的局部辨识性特征。在图3.9和表3.6中，“OPSC’’ 表示对象-部件空间关联约束，“PA’’ 表示部件语义对齐，“OPSC+PA’’ 表示两者的结合，即本章的部件级注意力模型。从图3.9中四个数据集的左边一列可以看出，仅使用 “PA’’ 所选择的部件：1）辨识性信息少，具有大面积的背景区域，小面积的对象区域；2）冗

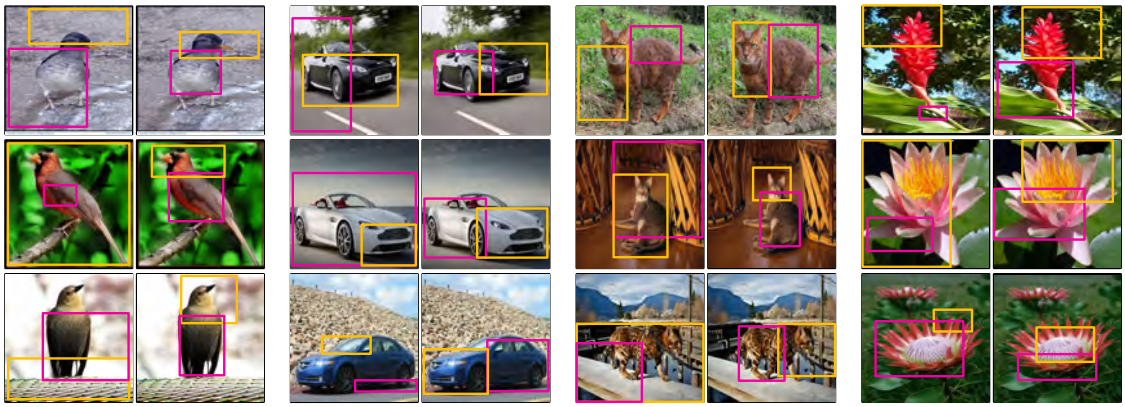


图 3.9　与现有方法部件自动定位结果对比

余信息多，部件之间重叠区域多，造成信息冗余。从表3.6中可以发现，通过对象-部件空间关联约束所选的部件与部件语义对齐所选的部件能够取得更好的细粒度图像分类结果。但是，两者的结合能够取得更好的结果。这充分证明了将具有相同语义的部件聚合在一起能够进一步促进部件注意力模型的细粒度图像分类结果。

3.3.4.3　对象级注意力选择的有效性

在对象级注意力模型中，一些图像块经过过滤选择作为扩充训练图像，其与对象相关，并被用来训练 *ClassNet*，以学习到多视角、多尺度的辨识性特征。在表3.7中，“ft-patches’’ 表示利用对象级注意力选择得到的图像块训练的结果，“ft-original’’ 表示只是用原图训练的结果。我们可以发现，前者能够取得更好的分类准确率，由于其能学习到更加丰富的多视角、多尺度特征。

表 3.7　对象级注意力选择的有效性

方法	准确率 (%)			
	CUB-200-2011	Cars-196	Oxford-IIIT Pet	Oxford-Flower-102
ft-patches	**80.82**	**86.79**	**88.14**	**94.70**
ft-original	80.11	85.76	87.52	93.84

3.4　本章小结

本章提出了对象-部件注意力模型，实现了弱监督的细粒度图像分类，其联合了两种注意力模型：对象级注意力模型定位图像中的辨识性对象区域，部件级注意力模型选择对象中具有辨识性的部件区域。两者的结合能够促进多视角、多粒度的特征学习，

从而强化两者的互补性。此外，提出基于对象-部件空间关联约束的部件选择方法，其联合了两种空间关联约束：对象级空间关联约束能够确保所选部件具有较高的代表性，部件级空间关联约束能够降低所选部件的冗余性，强化其辨识性。两者的结合能够促进局部辨识性区域的定位与特征学习。重要的是，本章对象-部件注意力模型避免使用成本巨大的对象级和部件级标注信息，促进了细粒度图像分类迈向实际应用。在四个细粒度图像分类数据集上的实验结果验证了本章对象-部件注意力模型的有效性，与 10 多个现有方法的对比均取得了最好的结果。

第四章　基于堆叠式深度强化学习的细粒度图像分类

4.1　引言

相似的细粒度子类别的区别一般位于对象的局部细节中，即使是人类也很难捕获，更不必说计算机。研究表明人类在识别时，倾向于先找到对象所在的位置[77]。眼睛通常会寻找定位在具有高特征密度[78]、特殊质地[79]以及颜色对比度高[80]的区域。这些特点都是影响对象辨识度的特征。例如，当人类识别一张图像时，首先会先找到图像中对象所在的位置，然后再从对象本身寻找具有辨识性的部件区域，最后根据这些特征来识别图像的类别，如图4.1所示。

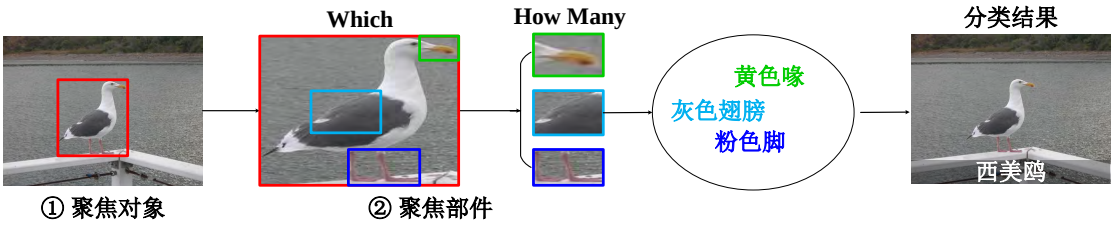


图 4.1　人类识别图像时的注意力示意图

受到人类识别图像过程的启发，现有细粒度图像分类方法通常先聚焦于如何定位图像中具有辨识性的区域，即对象与其部件区域。在辨识性区域定位的过程中，有两个重要的问题：1）“Which’’ 问题：哪些区域是具有辨识性和代表性的，能够将其与其他相似的细粒度子类别区分开？2）“How Many’’ 问题：多少个辨识性的区域对于获得最好的细粒度分类效果来说是必须的？

现有细粒度图像分类方法一般依赖于人工先验来解决上述问题，这严重限制了它们的可用性和可扩展性。Zhang 等人[5]利用数据集中头部和躯干的标注信息来训练 R-CNN[35]，通过几何约束来检测对象及其部件。Huang 等人[14]则利用数据集中的部件级标注信息来训练全卷积神经网络，从而来定位对象的部件区域。这些方法通常依赖于对象级和部件级标注信息来解决 “Which’’ 和 “How Many’’ 问题。但是，并非所有的标注信息都是有利于最终的细粒度图像分类的。例如，在 CUB-200-2011 鸟类数据集[3]中标注了鸟的眼睛在图像中的位置信息，但是由于其在图像中的区域太小，包含了较少的信息，很难利用其进行细粒度分类。这些通过人工先验来设定的标注信息使得辨识性区域的定位具有一定的主观性，并且使得细粒度图像分类算法需要根据不同的任务或数据集进行定制化。

因此，研究者们开始聚焦于如何不依赖这些人工先验的标注信息而自动地定位到图像中的辨识性区域。Zhang 等人通过聚合卷积核来进行部件的选择与表示[7]。在该方法中，辨识性区域的数目是通过另外一种人工先验的方式，即实验验证来确定的，对于不同的数据集会采用不同的数目，从而在当前数据集上取得最好的细粒度图像分类结果。第三章中提出的对象-部件注意力模型同样也是根据实验验证来确定最佳的辨识性区域的数目。这种通过实验验证的人工先验方式使得细粒度图像分类方法不易扩展到其他任务或数据域，大大增加了现有方法的复杂性和不确定性，限制了其可用性和可扩展性。此外，它们对于所有的细粒度类别设定相同的辨识性区域数目以作为区分的依据，忽略了不同子类别甚至不同的图像所具有的辨识性区域数目是不同的，这在一定程度上也影响了细粒度图像分类的准确率。

为了同时解决 “Which’’ 和 “How Many’’ 问题，本章提出了堆叠式深度强化学习方法（Stacked Deep Reinforcement Learning，简称 StackDRL），层次化地定位对象及其辨识性部件区域，并自自适应地选择辨识性区域的数目，避免了现有方法依赖人工先验所造成的可用性和可扩展性上的局限。其主要贡献总结如下：

- **多粒度辨识性定位**：以一种自适应的方式来解决 “Which’’ 和 “How Many’’ 问题，而不像现有方法[8, 81] 一样依赖于人工先验。提出了一种对象-部件两阶段的深度强化学习方法来有序层次化地定位不同粒度的辨识性区域，即对象及其部件，来解决 “Which’’ 问题；同时，利用强化学习的自主学习能力自适应地选择辨识性区域的数目来解决 “How Many’’ 问题。
- **多尺度特征学习**：以避免对象及部件不同尺度对于细粒度图像分类的影响，相比只考虑一种尺度的细粒度图像分类能够取得更好的效果。多尺度体现在两个方面：1）输入图像的多尺度。我们采用了两种尺度的图像作为输入，大尺度图像更加聚焦于细节信息，而小尺度图像更加聚焦于常规信息。2）辨识性区域的多尺度。通过多粒度辨识性定位能够获得多个辨识性的候选部件，它们可能包含相同的语义部件，但是却有不同的尺度，提供了更多的互补信息。因此，多尺度特征学习通过定位不同尺度的更具辨识性的区域，并联合不同尺度的区域特征来提升细粒度图像分类准确率。
- **语义奖励函数**：通过避免在深度强化学习中使用标注成本巨大的对象级和部件级信息，来增强本章堆叠式深度强化学习的可用性和可扩展性。它通过引入语义信息充分利用了图像中的辨识性和概念性信息，其由两种奖励函数构成：1）辨识性奖励函数，聚焦于定位更显著的区域；2）概念性奖励函数，聚焦于定位更具有概念信息、有助于提升细粒度分类准确率的区域。二者的联合能够同时促进细粒度定位与分类。
- **无监督辨识性定位**：探索本章堆叠式深度强化学习在无监督条件下的辨识性区

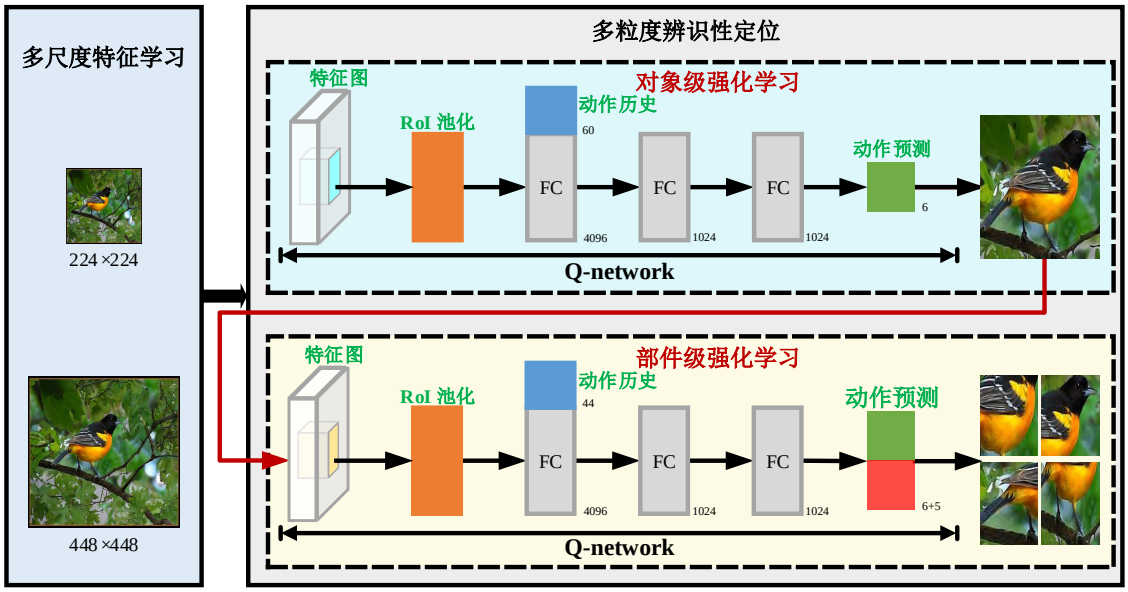


图 4.2　本章堆叠式深度强化学习的总体框架

域定位能力，不使用任何标注信息，包括图像级、对象级和部件级的标注信息。有效避免了巨大的标注成本，极大地增强了本章堆叠式深度强化学习的可用性和可扩展性。

4.2　算法描述

本章提出的堆叠式深度强化学习（StackDRL）的总体框架如图 4.2 所示，其包括多粒度辨识性定位（Multi-granularity Discriminative Localization，简称 MgDL）和多尺度特征学习（Multi-scale Representation Learning，简称 MsRL），能够通过深度强化学习获取多粒度的辨识性的区域注意力以及多尺度的区域特征表示。

为了充分利用图像的信息，本章堆叠式深度强化学习方法采用多尺度的图像作为输入。对于每一种尺度的图像，一种两阶段的深度强化学习被提出以发现图像中辨识性区域的多粒度信息。其中，第一阶段深度强化学习，即对象级强化学习（ObjectDRL），通过一系列缩放操作在图像中定位到对象区域；第二阶段深度强化学习，即部件级强化学习（PartDRL），通过一系列的缩放平移操作进一步在对象区域中挖掘更具辨识性的局部区域。值得注意的是，在此过程中定位到的辨识性区域，对于不同的图像、不同的细粒度子类别均有不一样的数目，这体现了本章堆叠式深度强化学习方法的自适应性，在有效提升细粒度图像分类准确率的同时，也提高了方法的可用性和可扩展性。在学习的过程中，提出语义奖励函数以引导模型获得最优的长期奖励（Long-term Reward）。

4.2.1 问题定义

对于给定的一幅图像 I，我们将辨识性定位定义为在候选图像区域集合 B 最大化置信度得分函数 $f_c : B \rightarrow \mathbb{B}$:

$$b^* = arg\max_{b \in B} f_c(b) \tag{4.1}$$

我们通过马尔科夫决策过程（Markov Decision Process，简称 MDP）来解决这个问题。MDP 能很好地适用于建模离散时间序列决策过程，其包含一个动作集合 A，一个状态集合 S 以及一个奖励函数 R。这三者在 ObjectDRL 和 PartDRL 中的定义有所不同，具体细节见后面小节内容。

4.2.2 多粒度辨识性定位

在深度强化学习的过程中，每执行完一个动作（Action），环境（Environment）的状态（State）就会发生变化，与此同时，一个与动作对应的奖励（Reward）也会随着当前状态一起反馈给智能体（Agent），指导智能体执行下一个动作。这是一个完整的深度强化学习的过程。下面我们从辨识性定位动作（Discriminative Localization Actions）、状态（State）和语义奖励函数（Semantic Reward Function）三个方面来详细介绍。

4.2.2.1 辨识性定位动作

我们将辨识性定位动作 A 定义为两组动作，如图4.3所示。

第一组动作包含 5 个缩放动作以逐渐定位到图像中的辨识性区域，和 1 个特殊动作用来停止当前智能体的动作，即 “Trigger’’。每一个缩放动作都将当前辨识性区域以缩放比率 α 向一个特定的子区域移动，分别对应当前区域的 4 个顶角和中心区域，其中 $\alpha \in [0,1]$。缩放动作可以定位到不同尺度的区域，因此保证了辨识性定位的有效性。

第二组动作包含 4 个平移动作以逐渐定位到图像中的辨识性区域，和 1 个特殊动作用来停止当前智能体的动作，即 “Trigger’’。每一个平移动作都将当前区域以平移比率 β 向上下左右四个方向平移，其中 $\beta \in [0,1]$。平移动作可以使得智能体纠正辨识性定位的错误操作，同时发现更多的不同的辨识性区域。

值得注意的的是对象级强化学习和部件级强化学习采用了不同的动作集合。在对象级强化学习中，只采用了第一组动作。在部件级强化学习中，我们希望智能体能够定位多个具有不同特征的区域，以更好地与其他相似细粒度子类别进行区分。因此，在部件级强化学习中，我们采用了树状执行策略，即一路执行第一组动作，另外一路执行第二组动作，如图 4.4 所示。其目的是：1）自适应地定位对象的辨识性部件；2）两路操作能够起到纠正错误动作的作用。

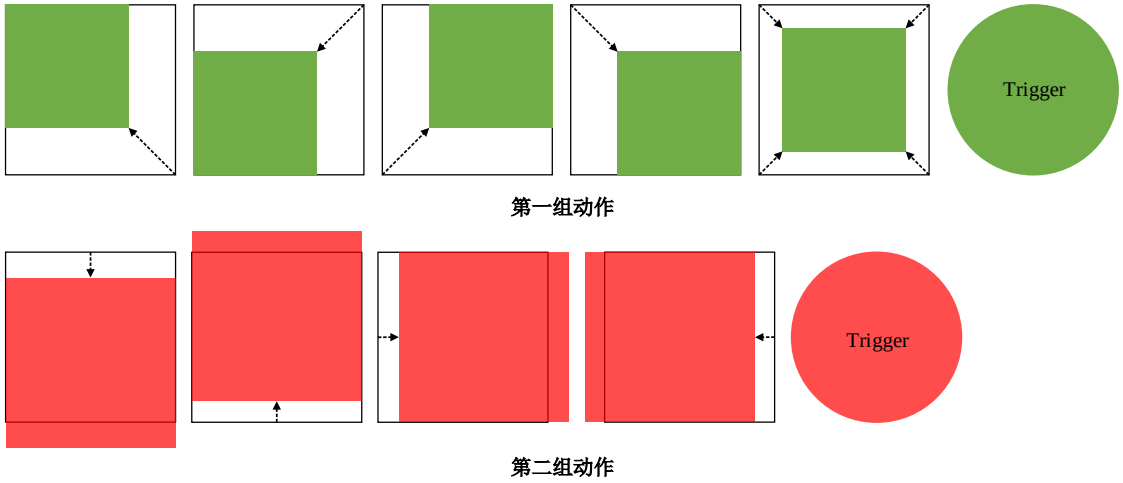


图 4.3　辨识性定位动作示意图

4.2.2.2　状态

当一个动作执行以后，当前定位区域就会发生变化，相应的当前状态也发生变化，我们采用历史向量（History Vector）和特征向量（Feature Vector）来表示。在第 t 步动作，状态表示为 $S_t = (v_t, h_t)$。其中，特征向量 v_t 是当前定位到的辨识性区域的特征表示，历史向量 h_t 记录了智能体之前已经执行过的所有动作。当前状态 S_t 会随着奖励 R 一起反馈给智能体，以指导智能体决定下一步要执行的动作。

特征向量 v_t 是指定位到的辨识性区域从 CNN 模型中提取的特征，该 CNN 模型是在 ImageNet 1K 数据集[39] 上预训练得到的。在本章实验中，将 VGGNet[43] 中 “conv5_3’’ 层的特征图作为初始特征，然后紧跟一个全连接层以生成最终的 4096 维特征向量。受到 Fast R-CNN [82] 的启发，感兴趣区域池化（RoI Pooling）层被用来加速特征提取。

历史向量 $h_t = \{H_1, H_2, ..., H_N\}$ 是一个二值向量，其中 N 表示预先设定的动作执行最大次数。H_i 表示第 i 个执行动作的独热编码（One-hot Encoding），其维度在对象级强化学习中为 6，在部件级强化学习中为 11，这是与其动作执行的次数有关。值得注意的是，在第 t 步动作时，t 以后的元素值均为 0。

4.2.2.3　语义奖励函数

奖励函数 R 用来评价当前执行的动作的优劣，是否在逐步逼近辨识性区域。如果定位到的辨识性区域与之前相比更加准确，则说明当前执行的动作是正确的，会反馈给智能体一个正奖励（Positive Reward）；反之，则反馈一个负奖励（Negative Reward）。本章提出了一种语义奖励函数，充分考虑了图像的辨识性和概念性信息。

(I) 辨识性奖励函数

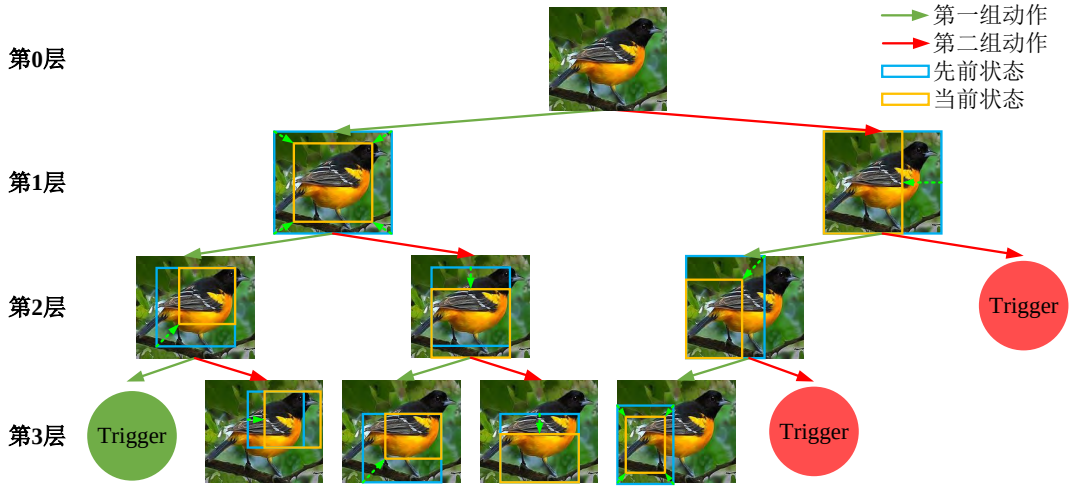


图 4.4　树状执行策略示意图

当前定位区域与标注信息（如 Bounding Box）之间的 IoU 值通常用来评价当前执行动作对于定位效果的影响[83]。我们用 $RA_a(s,s')$ 奖励函数来表示当执行动作 a，状态由 s 变为 s' 时智能体所收到的奖励反馈，其定义如下：

$$RA_a(s,s') = sign(IoU(b',g) - IoU(b,g)) \tag{4.2}$$

其中，b 表示当前定位到的区域，b' 表示在当前区域 b 上执行动作 a 所定位到的区域，g 表示标注信息，$IoU(b,g)$ 表示 b 与 g 两区域的交集与并集面积的比值，具体定义为 $IoU(b,g) = area(b \cap g)/area(b \cup g)$，$IoU(b',g)$ 表示 b' 与 g 两区域的交集与并集面积的比值。然而，上述奖励函数 $RA_a(s,s')$ 严重依赖于成本巨大的标注信息。

因此，本章提出了一种新的基于注意力的辨识性奖励函数，能够有效降低标注成本。Zhou 等人的工作[40] 表明卷积神经网络的卷积核具有在不依赖对象级标注信息的条件下定位对象的能力。受到其工作的启发，本章通过对卷积层的特征图输出进行处理以获得显著图，即

$$M(x,y) = \frac{1}{K}\sum_{u=1}^{K} f_u(x,y) \tag{4.3}$$

其中，$M(x,y)$ 表示了空间位置 (x,y) 的激励响应对于细粒度图像分类的重要程度，$f_u(x,y)$ 表示第 u 个卷积核在空间位置 (x,y) 处的响应值，K 表示卷积核的总数。因为显著图对于最终的细粒度图像分类准确率有重要的影响，我们展示了其辨识性定位的有效性。召回率和 IoU 值曲线如图4.5所示。在 CUB-200-2011 数据集上，AUC（Area Under Curve) 值在训练集和测试集上分别为 0.494 和 0.487。在 Cars-196 数据及上，则分别为 0.478 和 0.471。考虑到并未使用对象级标注信息，取得这样的定位结果已经是很不错的。

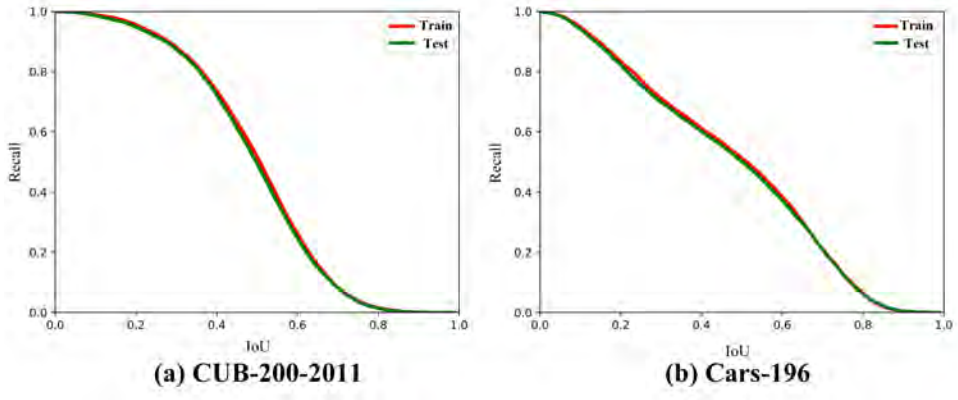


图 4.5　在 CUB-200-2011 和 Cars-196 两个数据集上的召回率与 IoU 值曲线

本章对象级强化学习和部件级强化学习分别设计了不同的辨识性奖励函数。

对象级强化学习中的辨识性奖励函数：首先在显著图上进行大津阈值二值化算法（OTSU）[84]，然后选择最大的连通域作为伪对象级标注信息 g_{atten}。具体的辨识性奖励函数定义如下：

$$RA_a(s,s') = sign(IoU(b', g_{atten}) - IoU(b, g_{atten})) \tag{4.4}$$

基于注意力的辨识性奖励函数充分利用了图像中的注意力信息，而不依赖于对象级标注信息，能够引导智能体学习如何定位到图像中的辨识性对象区域。

部件级强化学习中的辨识性奖励函数：因为需要定位多个辨识性部件区域，无法为每一个部件区域都生成一个伪标注信息，因此我们选取显著图中的显著值作为部件级强化学习中奖励函数的惩罚标准。因此，辨识性奖励函数 RA_a 定义如下：

$$RA_a(s,s') = sign(Mean(b') - Mean(b)) \tag{4.5}$$

其中，$Mean(\cdot)$ 表示当前定位到的区域对应的显著图中像素点的均值。通过树状执行策略以及辨识性奖励函数，我们可以获得对象中不同的具有一定辨识性特征的局部区域，能够有效的提升区域特征的多样性，从而获得更好的细粒度图像分类效果。

(II) 概念性奖励函数

众所周知，图像级的类别标签能够直接提供对应的概念信息，可以用来引导智能体定位得到真正有助于细粒度分类的辨识性区域。因此，本章提出了一种概念性奖励函数，其定义如下：

$$RC_a(s,s') = sign(P_c(b') - P_c(b)) \tag{4.6}$$

其中，$P_c(\cdot)$ 表示当前定位到的区域在对应细粒度子类别 c 上的预测得分，c 是图像级

的类别标注信息。

语义奖励函数 R 综合考虑了上述两种信息，其定义如下：

$$R_a(s,s') = RA_a(s,s') + RC_a(s,s') \tag{4.7}$$

值得注意的是，受到[85] 启发，我们对于 “Trigger’’ 单独定义了一个奖励函数，用来表示当前定位到的区域已经锁定了目标对象或者部件。在对象级强化学习中，其定义如下：

$$RO_{trigger}(s,s') = \begin{cases} +\eta, & if\ IoU(b, g_{atten}) \geq \tau \\ -\eta, & otherwise \end{cases} \tag{4.8}$$

其中，η 是特定的结束反馈奖励。只有当 IoU 值超过设定的阈值 τ，“Trigger’’ 动作才会执行。在部件级强化学习中，其定义如下：

$$RO_{trigger}(s,s') = \begin{cases} +\eta, & if\ Mean(b) \geq \tau \\ -\eta, & otherwise \end{cases} \tag{4.9}$$

表示当当前定位到的区域的显著性均值大于某一阈值时则执行 “Trigger’’ 动作。

通过对象级强化学习首先定位到图像中的对象区域，然后进一步通过部件级强化学习在对象区域上挖掘更多的辨识性局部信息。

4.2.3 辨识性定位中的 Q-learning 算法

本章采用深度强化学习来学习定位策略，以获取最大的奖励反馈。DQN 算法（Deep Q-network）[86] 被用来解决深度强化学习问题。本章的 Q-network 结构如图4.6所示。其包含三路：第一路是为了动作预测，第二路是为了辨识性奖励计算，第三路是为了概念性奖励计算。特别地，每个候选对象区域或者部件区域都通过 RoI 池化来提取特征以减少计算消耗，然后将其作为第三路的输入用于计算概念性奖励，即公式（4.6）中的 $P_c(\cdot)$。在 RoI 池化之前，特征图用来生成对应的显著图。这样辨识性奖励和概念性奖励能够同时用来引导动作的预测。对于动作预测一路，我们将特征向量和动作历史向量拼接在一起，将其作为全连接层的输入。最终，获得下一步动作的预测结果。我们采用了在特定细粒度图像分类数据集上微调的 CNN 模型。这是因为微调后的 CNN 模型能够获得更好的显著图，提取更加有效、更具辨识性的特征。

4.2.4 无监督辨识性定位

在本节，我们探索一种无监督辨识性定位方法，即不使用任何标注信息，包括图像级、对象级和部件级标注信息。我们知道显著图能够指出图像中哪些区域具有辨识

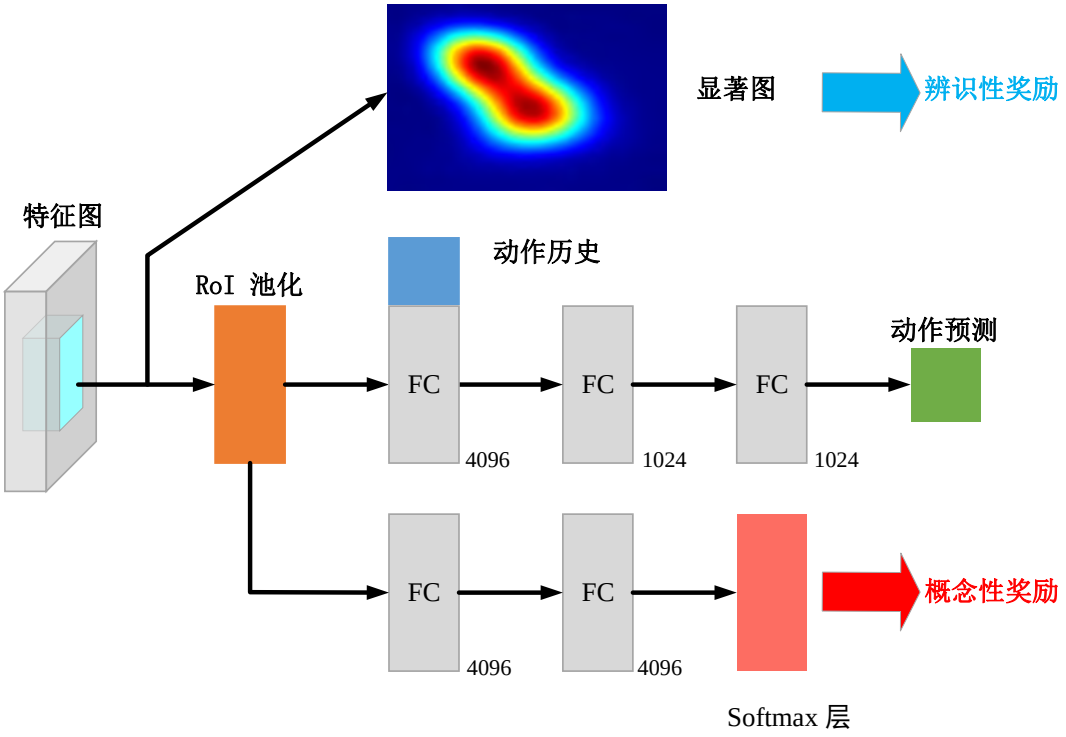


图 4.6　Q-network 结构

性，能够决定最终的细粒度分类结果。此外，先前的研究表明，在 ImageNet 1K 数据集上预训练得到的 CNN 模型具有较好的泛化性。考虑到从预训练 CNN 模型中提取的显著图不能较好地反应对象区域，却能较好地反应一些辨识性的局部区域，我们在无监督条件下仅采用部件级强化学习。

特别地，在无监督辨识性定位中，定位动作和状态与部件级强化学习一样。为了避免使用标注信息，在语义奖励函数的设计上，我们仅采用辨识性奖励函数，其与部件级强化学习中的 RA_a 一致。

$$RU(s,s') = sign(Mean(b') - Mean(b)) \tag{4.10}$$

但是，CNN 模型并未在特定的细粒度图像数据集上进行微调，而且并未使用任何标注信息。此外，对于辨识性定位中的 Q-learning 算法，我们用预训练的 CNN 模型来初始化其卷积层，用零均值正态分布来初始化全连接层的参数。

4.2.5　多尺度特征学习

通过多粒度辨识性定位，对于每幅图像可以得到一系列辨识性区域，其数目不固定。这些辨识性区域对应图像中的对象及部件区域，具有多尺度的特性。目标对象和部件区域存在小尺度的情况，这导致智能体难以定位。因此，我们提出了多尺度特征

学习方法，其包含两个方面：1）输入图像的多尺度。我们采用了两种尺度的图像作为输入，大尺度图像更加聚焦于细节信息，而小尺度图像更加聚焦于常规信息。在实验中，我们采用了 224×224 和 448×448 两种尺度。2）辨识性部件的多尺度。在部件级强化学习中，我们不仅使用树状执行策略的叶节点所定位到的辨识性区域，如图4.4所示，而且采用了除根节点以外节点所定位到的辨识性区域。由于在不同树层的辨识性区域有不一样的尺度，提供了更多的互补信息。因此，多尺度特征学习通过定位不同尺度的更具辨识性的区域并联合不同尺度的区域特征来提升细粒度图像分类准确率。

4.2.6 最终预测

对于给定的图像 I，在对象级强化学习中有不超过 $N_{step}-1$ 个候选区域，在部件级强化学习中有不超过 $2^{N_{level}}-2$ 候选区域，分别对应具有辨识性的对象和部件。每一个候选区域都作为微调的 CNN 模型的输入，并得到其预测得分向量。对于通过对象级强化学习得到的候选区域，我们选取得分最高的区域得分作为最终的对象区域得分，表示为 $\max(SO)$。对于通过部件级强化学习得到的候选区域，我们选取树状结构中每一层得分的最高值，表示为 $\max(SP_l)$。最终的预测结果通过以下方式获得：

$$Score=\lambda\max(SO)+(1-\lambda)\frac{1}{N_{level}}\sum_{l=1}^{N_{level}}\max(SP_l) \tag{4.11}$$

其中，λ 通过 k-折交叉验证方法获得。

4.3 实验结果与分析

我们在广泛使用的细粒度图像分类数据集 CUB-200-2011 [3] 和 Cars-196 [28] 上进行实验。上述两个数据集已经在第一章中进行了详细介绍，在这里不再赘述。在评测指标上，除了采用细粒度图像分类常用的准确率以外，还采用了 IoU 值[87] 来验证辨识性区域定位的准确性，其定义如下：

$$IoU=\frac{area(b\cap g)}{area(b\cup g)} \tag{4.12}$$

其中，b 表示预测得到的辨识性区域的矩形框，g 表示目标区域的标注信息，$b\cap g$ 表示二者的交集，$b\cup g$ 表示二者的并集。

4.3.1 实验设置

本节介绍实验的细节：1）对于辨识性定位动作，将缩放和平移比率分别设置为 0.9 和 0.1。为了在辨识性定位速度和准确率之间取得权衡，在对象级强化学习中将最大

执行次数 N_{step} 设置为 10。对于不同的数据集 N_{step} 保持不变。在部件级强化学习中，$N_{level}=4$。2）对于语义奖励函数，“Trigger’’ 的奖励 η 以及阈值 τ 分别设置为 3 和 0.5。

表 4.1　CUB-200-2011 数据集上实验结果

方法	训练集标注		测试集标注		准确率 (%)
	对象级	部件级	对象级	部件级	
本章 StackDRL 方法					**87.21**
第三章 OPAM 方法					85.83
CVL [31]					85.55
RA-CNN [17]					85.30
HCA [88]					85.30
PNA [81]					84.70
TSC [89]					84.69
FOAF [34]					84.63
PD [7]					84.54
LRBP [90]					84.21
STN [44]					84.10
Bilinear-CNN [18]					84.10
Multi-grained [45]					81.70
NAC [38]					81.01
PIR [37]					79.34
TL Atten [8]					77.90
MIL [46]					77.40
VGG-BGLm [22]					75.90
InterActive [47]					75.62
Dense Graph Mining [48]					60.19
Coarse-to-Fine [49]	√				82.50
Coarse-to-Fine [49]	√		√		82.90
PG Alignment [15]	√		√		82.80
VGG-BGLm [22]	√		√		80.40
Triplet-A (64) [50]	√		√		80.70
Triplet-M (64) [50]	√		√		79.30
Webly-supervised [51]	√	√			78.60
PN-CNN [36]	√	√			75.70
Part-based R-CNN [5]	√	√			73.50
SPDA-CNN [52]	√	√	√		85.14
Deep LAC [53]	√	√	√		84.10
SPDA-CNN [52]	√	√	√		81.01
PS-CNN [14]	√	√	√		76.20
PN-CNN [36]	√	√	√	√	85.40
Part-based R-CNN [5]	√	√	√	√	76.37
POOF [54]	√	√	√	√	73.30

表 4.2　Cars-196 数据集上实验结果

方法	训练集标注		测试集标注		准确率 (%)
	对象级	部件级	对象级	部件级	
本章 StackDRL 方法					**93.25**
RA-CNN [17]					92.50
第三章 OPAM 方法					92.19
Bilinear-CNN [18]					91.30
TL Atten [8]					88.63
DVAN [58]					87.10
FT-HAR-CNN [59]					86.30
HAR-CNN [59]					80.80
PG Alignment [15]	√				92.60
ELLF [60]	√				73.90
R-CNN [35]	√				57.40
PG Alignment [15]	√		√		92.80
BoT(CNN With Geo) [61]	√		√		92.50
DPL-CNN [62]	√		√		92.30
VGG-BGLm [22]	√		√		90.50
LLC [63]	√		√		69.50
BB-3D-G [28]	√		√		67.60

4.3.2　与现有方法进行对比

本节展示了本章堆叠式深度强化学习方法（表示为 StackDRL）以及现有方法在 2 个数据集上的细粒度图像分类准确率与实验分析。实验结果如表4.1和表4.2所示。为了公平对比，表格中列出了方法在训练和测试阶段所使用的的标注信息。值得注意的是，本章 StackDRL 方法既没有使用对象级标注，也没有使用部件级标注，仅使用了图像级的类别标注。

在 CUB-200-2011 数据集上，本章 StackDRL 方法取得了最好的细粒度图像分类准确率，如表4.1所示。对比方法中，取得最高细粒度图像分类准确率的是本文第三章 OPAM 方法，其联合了对象级和部件级的注意力，将辨识性区域的数目设置为 3，包括 1 个对象区域和 2 个部件区域。与其相比，本章 StackDRL 方法取得了 1.38% 的提升。CVL 方法 [31] 联合建模了图像和文本信息，从而获得图像和文本的联合表示。除了图像信息，它在训练阶段额外使用了文本信息。即使如此，本章 StackDRL 方法依然取得了比它高 1.66% 的准确率。RA-CNN 方法 [17] 取得了 85.30% 的准确率，它采用了 3 种不同尺度的区域。当它采用 2 种尺度的区域时，准确率为 84.70%，降低了 0.60%。PNA 方法 [81] 训练了 11 种部件检测器来定位辨识性区域。TSC 方法[89] 通过定位 3 个辨识性区域来获得最好的准确率。从上述分析可以看出，辨识性区域的数目对于最终的细粒

度图像分类结果而言至关重要，但是现有方法通常依靠人工先验的方式来确定。这使得上述方法需要对不同任务或数据域进行特殊定制，严重限制了上述方法的可用性和可扩展性。

而本文 StackDRL 方法则通过对象级和部件级两阶段的强化学习，层次化地自动定位辨识性区域，并自适应地决定辨识性区域的数目，来解决“Which’’ 和 “How Many’’ 问题。辨识性区域数目的设定是在多粒度辨识性定位过程中自适应完成的。其不仅对于不同的细粒度子类别有不同的数目，而且对于不同的图像也有不同的数目，如图4.10所示。我们将在4.3.3.2中具体阐述。由于能够自适应地定位和决定辨识性区域的数目，本章 StackDRL 方法取得了最好的细粒度分类准确率。

即使与那些使用了对象级标注信息的方法相比，本章 StackDRL 方法依旧可以取得最好的准确率。甚至与那些使用了部件级标注信息的方法相比，本章 StackDRL 方法也能取得最好的准确率。

此外，Cars-196 数据集上的实验结果如表4.2所示，其趋势与 CUB-200-2011 数据集一致，本章 StackDRL 方法同样取得了最好的准确率，有 0.75% 的提升，进一步验证了本章 StackDRL 方法的有效性。

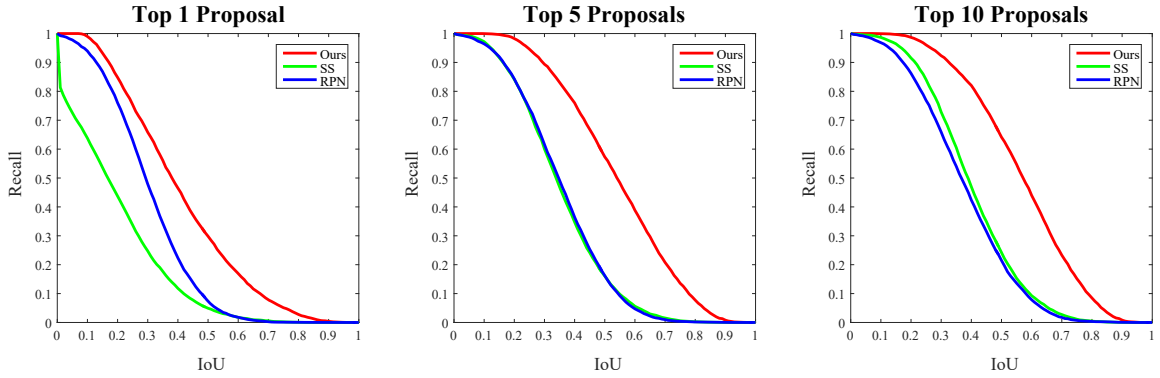


图 4.7　CUB-200-2011 数据集上定位召回率与 IoU 曲线

4.3.3　辨识性定位的有效性

在本章 StackDRL 方法中，对象级强化学习和部件级强化学习顺序执行以定位两种粒度的辨识性区域：对象及其部件。下面我们以 224×224 尺度图像作为输入，在 CUB-200-2011 数据集上分析辨识性定位的有效性。

4.3.3.1　对象级强化学习辨识性定位的有效性

对象级强化学习能够将对象区域从背景区域中辨别出来，以学习对象的整体特征。与[91] 一样，我们计算了不同召回率条件下的候选区域与目标区域标注之间的 IoU 值，如

图4.7展示了生成前（Top）1、5、10 个区域所对应的的曲线。与选择性搜索方法（Selective Search，表示为 SS）[13]、区域生成网络（Region Proposal Network，表示为 RPN）[91] 进行了对比，其 Top N 个候选区域是通过置信度得分来选择的。本章对象级强化学习依次生成 Top N 个候选区域。从图4.7可以看出，SS 和 RPN 的召回率明显低于本章 StackDRL 方法，从而验证了对象级强化学习在辨识性对象定位上的有效性。

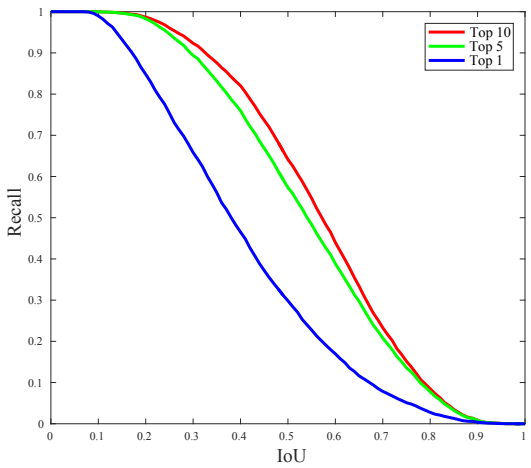


图 4.8　本章对象级强化学习方法在 CUB-200-2011 数据集上定位的召回率与 IoU 曲线

图4.8展示了本章对象级强化学习方法在 CUB 数据集上定位不同 Top N 个辨识性区域的召回率与 IoU 曲线。我们发现，随着定位辨识性区域的数目逐渐增加，其召回率也逐渐增加，这表明了本章对象级强化学习定位的有效性。需要注意的是，这些候选区域都是通过对象级强化学习自动且自适应生成的，对于不同的图像有不同的数目。图4.9展示了本章对象级强化学习方法定位辨识性区域的过程。我们可以发现智能体通过执行一系列最优的动作集合来确保所定位到的区域包含了图像中的对象。其中，红色的矩形框表示最终的定位结果。我们可以发现其定位效果很不错，验证了本章对象级强化学习方法定位的有效性。我们还计算了 AUCs 值，本章方法在 CUB-200-2011 和 Cars-196 数据集上的结果分别为 0.501 和 0.508，而 g_atten 的值分别为 0.494 和 0.487。这也说明了本章对象级强化学习方法的有效性，它能够通过基于语义奖励的深度强化学习方法来促进辨识性定位的准确性。

4.3.3.2　部件级强化学习辨识性定位的有效性

部件级强化学习能够发现对象区域中具有独特属性的区域，其能够与其他相似的细粒度子类别进行区分。图4.10展示了由部件级强化学习方法定位到的辨识性区域的数目。本章 StackDRL 方法尝试以一种自动、自适应的方式来解决 “Which’’ 和 “How Many’’ 问题。因此，定位到的辨识性区域数目对于不同的图像是不同的。在本章实验

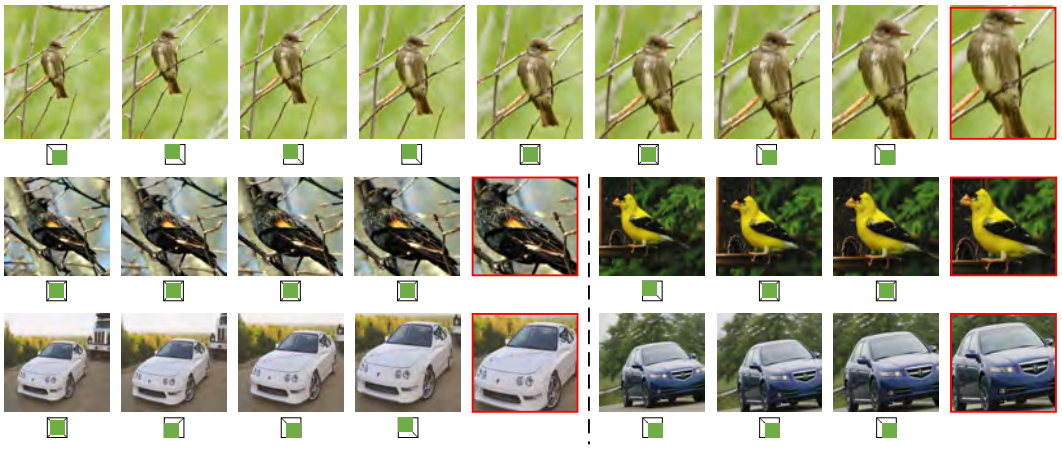


图 4.9　对象级强化学习辨识性定位过程示意图

中，辨识性区域的数目从 1 到 15 不等。由于针对不同图像采取了自适应、非确定数目的定位方式，本章 StackDRL 方法取得了最好的细粒度图像分类准确率，超过了依赖人工先验的方法。图4.11展示了部件级强化学习树状策略的定位结果。其中，‘‘第 0 层’’节点处的图像为对象级强化学习定位到的区域，而不是原始图像。可以看到，部件级强化学习能够发现多个辨识性的区域，而且这些区域是具有不同尺度的，指出了与其他细粒度子类别的区别。此外，由于平移动作的执行，使得不同分支所获得的辨识性区域具有一定的多样性，从而能够提供更多的辨识性信息。黄色和红色矩形框分别是由第一、二组动作定位到的区域，它们是不同的，充分验证了部件级强化学习中辨识性定位动作执行树状策略的有效性。

表 4.3　无监督辨识性定位的有效性

方法	CUB-200-2011	Cars-196
MgDL	**86.61**	**90.98**
UDL	83.29	90.34
PartDRL	83.23	88.98

4.3.4　无监督辨识性定位的有效性

表4.3对无监督辨识性定位的有效性展开探讨，其中，“UDL’’ 表示无监督辨识性定位。我们可以发现，无监督辨识性定位能够取得令人鼓舞的结果。PartDRL 使用了图像级的类别标注，而 UDL 并未使用，但 UDL 却可以取得与 PartDRL 相近的结果，这是非常有趣且重要的现象。这源于在 ImageNet 1K 数据集上预训练的 CNN 模型具有较好的泛化性。无监督辨识性定位甚至比使用了对象级标注信息的方法的效果要好，例如 Coarse-to-Fine (82.50%、82.90%) [49] 和 PG Alignment (82.80%) [15]，如表4.1所示。这启

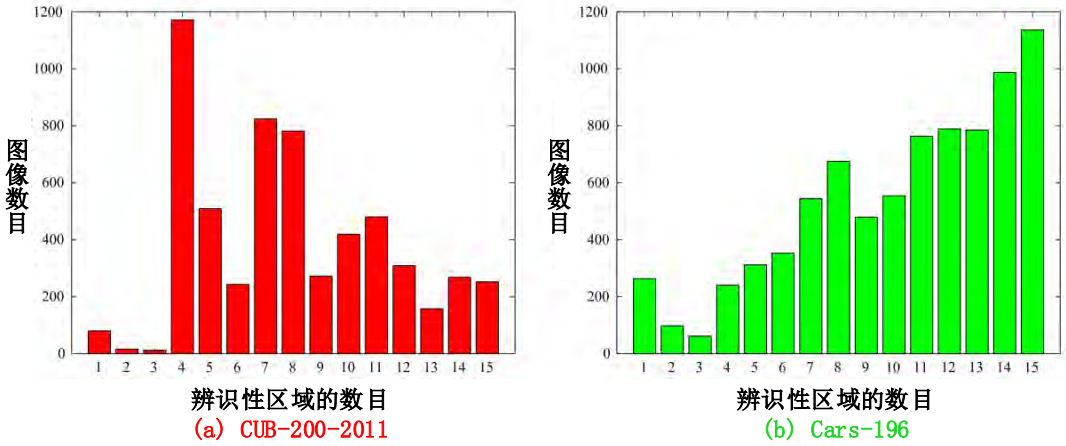


图 4.10　本章部件级强化学习方法定位到的辨识性区域数目展示

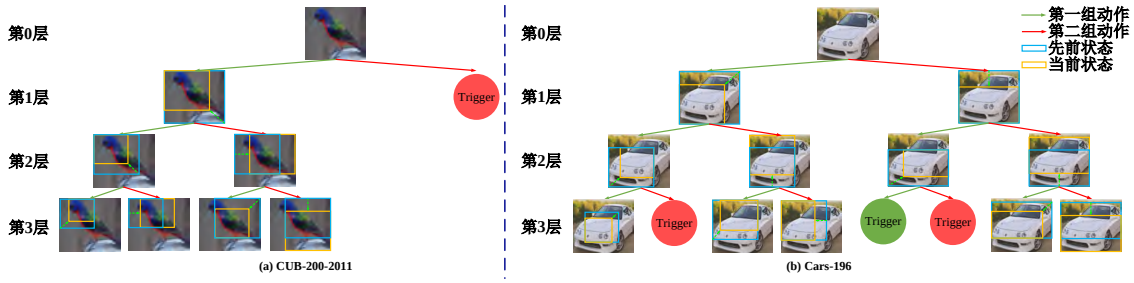


图 4.11　部件级强化学习辨识性定位的结果

示我们进一步探索无监督辨识性定位的应用。

4.3.5　本章 StackDRL 方法每个组成部分的有效性

为了充分验证本章 StackDRL 方法的有效性，我们在 CUB-200-2011 和 Cars-196 两个数据集上，从多尺度特征学习、多粒度辨识性定位和语义奖励函数三个方面来验证本章 StackDRL 方法每个组成部分的有效性。

表 4.4　多尺度特征学习的有效性

方法	CUB-200-2011	Cars-196
本章 StackDRL 方法	**87.21**	**93.25**
MgDL (224 × 224)	86.61	90.98
MgDL (448 × 448)	86.42	92.82

4.3.5.1　多尺度特征学习的有效性

在本章 StackDRL 方法中，不同尺度的图像被采用作为输入，提供了互补信息来促进细粒度图像分类效果。从表4.4可以看出，在两个数据集上多尺度信息能够带来至少

0.6% 的准确率提升。

4.3.5.2　多粒度辨识性定位的有效性

表4.5展示了多粒度辨识性定位在两个数据集上的有效性实验，实验中以 224×224 尺度的图像作为输入。“Baseline’’ 表示直接用在细粒度图像数据集上微调过的 VGGNet 网络识别原图的结果。“ObjectDRL’’ 表示考虑了定位得到的对象区域的结果。“PartDRL’’ 表示考虑了定位得到的辨识性部件的结果。“MgDL’’ 表示同时考虑了对象和部件的结果。我们可以发现：

表 4.5　多粒度辨识性定位的有效性

方法	CUB-200-2011	Cars-196
MgDL	**86.61**	**90.98**
ObjectDRL	85.29	89.93
PartDRL	83.23	88.98
Baseline	80.82	86.79

- 与 “Baseline’’ 相比，“PartDRL’’ 可以在两个数据集上分别能够提高 2.41% 和 2.19% 的准确率。这是由于 PartDRL 能够定位到局域辨识性的区域，能够增强特征的辨识度，同时这些区域还是多尺度的，进一步增强了特征的表示能力。
- 与 “Baseline’’ 相比，“ObjectDRL’’ 能够较大地提升细粒度分类准确率，在两个数据集上分别提高了 4.47% 和 3.14%。与 “PartDRL’’ 相比，也提高了 2.06% 和 0.95%。这是因为通过对象级强化学习定位得到的区域包含了对象的全局信息，同时也包含了反应辨识性信息的局部特征。为了进一步验证对象级强化学习方法的有效性，我们将其与显著性对象检测的方法（如 CAM）以及使用对象级标注信息的方法进行了对比，如表4.6所示。我们采用 CAM 方法生成对象区域，并使用它们去做细粒度分类，取得了比 “Baseline’’ 高的准确率。我们同样展示了使用对象级标注信息的 “Baseline w/ bbox’’ 的结果，其在两个数据集上分别比 CAM 的结果高 1.23% 和 2.57%。但是，考虑到 CAM 并未使用对象级标注信息，这样的结果已经是令人鼓舞的了。本章对象级强化学习方法能够比 CAM 方法高 1.55% 和 1.14%。这是因为两者之间不同的学习策略。在训练过程中，CAM 仅仅从原图中学习，而 ObjectDRL 因为要执行多个动作能够产生多个原图的子图，因此能够从中学习多尺度、多粒度且更具辨识性的特征。需要注意的是，即便没有使用标注信息，与 “Baseline w/ bbox’’ 相比，对象级强化学习依然能够取得甚至更好的结果。
- 对象级强化学习与部件级强化学习的结合能够进一步促进准确率的提升，与 “Baseline’’ 相比提升了 5.79%。这充分体现了两者之间的互补性，以及两阶段

深度强化学习的有效性。二者能够挖掘不同但互补的信息，聚焦在图像的不同辨识性区域，提供更多的辨识性信息。

表 4.6　对象级强化学习有效性实验结果

方法	CUB-200-2011	Cars-196
ObjectDRL	**85.29**	89.93
Baseline w/ bbox	84.97	**91.36**
CAM [40]	83.74	88.79
Baseline	80.82	86.79

4.3.5.3　语义奖励函数的有效性

表4.7展示了语义奖励函数有效性的验证实验结果，其中 “RA’’ 表示辨识性奖励函数，“RC’’ 表示概念性奖励函数。我们可以发现：

- 辨识性奖励函数和概念性奖励函数取得了相近的准确率，这表明注意力信息和类别信息在细粒度分类中起了相近的作用。
- 两者的联合能够进一步提升准确率，这是由于两个奖励函数的侧重点不同，是互补的：辨识性奖励函数能够提供图像中相对局部的显著信息，而概念性奖励函数能够提供图像中相对整体的特征信息。

表 4.7　语义奖励函数的有效性

方法	CUB-200-2011	Cars-196
MgDL	**86.61**	**90.98**
RA	85.79	90.37
RC	85.23	90.00

4.4　本章小结

为了解决 “Which’’ 和 “How Many’’ 问题，本章提出了堆叠式深度强化学习（StackDRL) 方法。首先，通过多粒度辨识性定位层次化地定位得到不同粒度的辨识性区域，同时自适应地决定细粒度分类所需的辨识性区域数目。然后，通过多尺度特征表示学习来帮助定位不同尺度的对象以及获得不同尺度图像的特征表示，通过尺度信息的互补提升细粒度分类准确率。此外，提出语义奖励函数来驱动 StackDRL 的学习，能够充分挖掘图像中的辨识性和概念性信息。进一步，提出无监督辨识性定位，有效避免了标注的巨大成本，极大地增强了方法的可用性和可扩展性。

第五章　基于弱监督快速辨识定位的细粒度图像分类

5.1　引言

细粒度图像分类在智能农业、智能医疗、智能零售等智能产业有着广泛的应用前景。但是，在技术向应用的转化过程中，现有细粒度图像分类方法有两个亟需解决的问题：1）时间消耗：现有方法主要聚焦于如何取得更高的细粒度图像分类准确率，但是忽略了方法的复杂度，导致分类的速度很慢。但是，在实际应用中，需要满足用户对于系统的响应速度要求，因此实时性是实际应用中的一项重要标准。2）标注消耗：正如在本文前面章节中所述，很多现有方法为了定位到图像中具有辨识性的区域，在训练甚至是测试过程中均使用了成本巨大的对象级和部件级标注信息。因此，尽可能少地使用标注信息且能够获得不错的效果是细粒度图像分类方法向实际应用转化的关键。

本文第三章和第四章的工作，主要是致力于解决标注消耗的问题，即在不使用对象级和部件级标注信息的情况下，实现图像中辨识性对象和部件的自动定位，从而使得细粒度图像分类准确率超过使用对象级、部件级标注信息的强监督方法。但是，它们仅仅聚焦于如何在降低标注成本的情况下如何取得较高的细粒度图像分类准确率，却忽略了速度问题。这主要是因为这类方法通常采用两阶段的步骤：首先，通过各种自动的方式来定位图像中的辨识性区域；然后，根据辨识性区域进行辨识性特征的学习，从而进一步实现细粒度图像分类。这种两阶段的方式，使得辨识定位与分类分裂开来，而且通常对象与部件的定位也是分裂的，这就导致辨识速度非常慢。这也就是时间消耗问题。

当然,现有方法中也有部分方法通过设计端到端的网络来解决时间消耗问题。Zhang等人[52] 提出部件堆叠的 CNN 网络（Part-stacked CNN），其包含了一个全卷积神经网络，和两路分类子网络。它首先利用全卷积神经网络来定位辨识性区域，然后采用两路分类子网络分别编码对象级和部件级的特征。它将定位和分类统一到一个网络模型中，有效实现了辨识速度的加速，但是却依赖于成本巨大的图像级、对象级和部件级标注信息。这就是标注消耗的问题。

现有方法通常只针对时间消耗和标注消耗两个问题中的一个，以牺牲另一个的方式来获得提升。因此，同时解决上述两个问题是一项非常重要且具有挑战性的任务。因此，本章提出了弱监督快速辨识定位（Weakly Supervised Fast Discriminative Localization，简称 WSFDL）方法，以同时解决上述两个问题，在加快辨识速度且较少标注依赖的情况下获得细粒度图像分类准确率的提升。其主要贡献归纳如下：

- **多级注意力引导的定位学习**：现有弱监督定位方法通常直接利用显著图生成辨

识性区域，其有两方面的局限性：定位速度慢和细粒度分类准确率低。因此，本章提出了多级注意力引导的定位学习方法，能够同时实现定位与细粒度分类。显著图被应用于二次定位学习，以学习到更加精准的定位，并通过避免现有方法对显著图的复杂处理过程实现定位的加速。此外，多级注意力能够提供多粒度、多尺度的信息，从而使得细粒度图像分类准确率得到提升。整个学习过程都是由显著图的信息驱动的，并未使用对象级和部件级的标注信息，有效避免了标注消耗。

- **多路端到端辨识性定位网络**：现有细粒度图像分类方法通常在一次定位过程中只能定位到一个辨识性区域，忽略了其他辨识性区域的影响。因此，本章提出了一种多路端到端辨识性定位网络，能够一次性定位多个不同的辨识性区域。其包含多个定位网络和一个区域生成网络，前者统一共享使用区域生成网路所提取的特征来减少卷积操作的计算，从而有效避免了由于增加定位多个辨识性区域所带来的的近线性的复杂度增长。

5.2 算法描述

本章提出了弱监督快速辨识定位方法（WSFDL），方法框架如图5.1所示。其包含两个子网络：多级注意力提取网络（Multi-level Attention Extraction Network，简称 MAEN）和辨识性定位网络（Discriminative Localization Network，简称 DLN）。多级注意力提取网络通过提取不同卷积层的特征来获取不同的注意力信息，并能够以此来生成多个初始辨识性区域。然后，这些区域的矩形框被作为辨识性定位网络的伪标注信息，从而避免了使用对象级或部件级标注信息。二者均可以生成辨识性区域，但有不同的特点：1）多级注意力提取网络虽然为辨识性定位网络提供了伪标注信息，但是其定位得到的辨识性区域并不十分准确。需要注意的是多级注意力提取网络仅在训练过程中使用。2）基于多级注意力提取网络提供的辨识性区域伪标注信息，辨识性定位进一步优化学习辨识性定位，从而定位到对于细粒度分类来说更有帮助的辨识性区域。二者联合充分利用了它们的优势，也弥补了其不足，从而获得更好的细粒度图像分类准确率。

5.2.1 多级注意力提取网络

注意力是一种选择性地聚集离散信息的行为和认知过程[92]。Karklin 等人的研究[93]表明，初级视觉皮层（如 V1）的神经元对边缘信息比较敏感，而高级视觉皮层（如 V2、V4）则对图像的特征和形状更感兴趣。同样的发现也出现在卷积神经网络中，不同层的特征图反映了图像的不同信息。不同层的卷积核有着不同的焦点，它们提供的信息互为补充，都能够为最终的细粒度图像分类做出贡献。

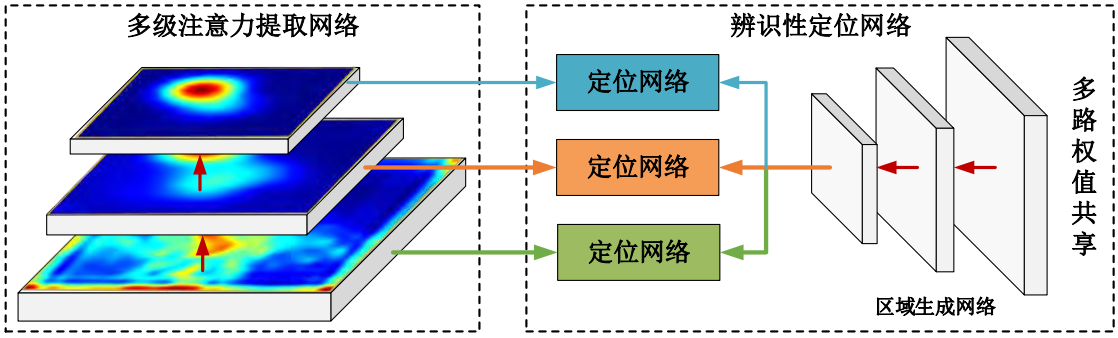


图 5.1　本章弱监督快速辨识定位方法示意图

根据现有的视觉注意力机制的相关研究，本章设计了多级注意力提取网络来生成多个辨识性区域，进一步将其作为伪标注信息指导辨识性定位网络的学习。受到 CAM [40] 启发，我们将卷积神经网络的最后一层前的全连接层去掉，替换为全局平均池化层，然后再跟 Softmax 层。然后，我们将卷积层的输出加权求和得到每一幅图像的显著图。在此阶段中，我们分别提取多个卷积层所对应的的显著图。最后，在这些显著图上进行二值化和最大连通域提取操作得到显著图对应的辨识性区域。这些区域所对应的的矩形框信息作为辨识性区域定位网络的监督信息指导其训练。

对于给定的图像 I，我们生成 n 个显著图，表示为

$$M_i(x,y) = \sum_{u_i} w_{u_i} f_{u_i}(x,y) \tag{5.1}$$

其中，$M_i(x,y)$ 表示空间位置 (x,y) 处对于最终细粒度分类的重要度，$f_{u_i}(x,y)$ 表示第 i 层卷积层中第 u_i 个卷积核在空间位置 (x,y) 处的激励响应，w_{u_i} 表示用于生成显著图的权重。对于不同的卷积层，w_{u_i} 定义如下：

$$w_{u_i} = \begin{cases} w_{u_i}^c, & i = L \\ \frac{1}{|u_i|}, & i \neq L \end{cases} \tag{5.2}$$

其中，$w_{u_i}^c$ 表示最后一层卷积层中卷积核 u_i 对应到细粒度子类别 c 的权重，c 为预测的子类别，$|u_i|$ 表示第 i 卷积层中卷积核的总数，L 表示模型中卷积层的数目。

5.2.2　辨识性定位网络

为了充分利用多级注意力信息，我们基于 Faster R-CNN [91] 设计了多路端到端网络，其包含多个定位网络和一个区域生成网络。Faster R-CNN 是为了加速检测过程，同时提升了检测准确率。我们对 Faster R-CNN 做了以下两个方面的改动：1）在训练过程中，Faster R-CNN 需要图像中辨识性区域的标注作为监督信息，而在本章弱监督快速辨识

定位方法中，我们通过多级注意提取网络来自动获取辨识性区域的伪标注信息，有效避免了成本巨大的对象级和部件级标注信息。2）受到视觉注意力机制相关研究发现的启示，本章弱监督快速辨识定位方法采用了多级注意力。但是，原始 Faster R-CNN 只包含一个区域生成网络和一个定位网络，使得其一次只能定位一个辨识性区域。如果采用训练多个 Faster R-CNN 来获取多级注意力，势必会造成接近线性复杂度的提升，大大提升了时间的消耗。因此，在本章弱监督快速辨识定位方法中，所有定位网络均共享由区域生成网络所得到的的卷积特征。

现有细粒度图像分类方法在定位辨识性区域时，通常会采用选择性搜索算法[13]，而在本章弱监督快速辨识定位方法中，我们采用了区域生成网络，其能够加速候选图像块的生成速度。对于区域生成网络的训练，一个二值类别标签作为监督信息，其表示当前预测区域是否包含对象，由预测区域与目标对象标注区域的 IoU 值决定。但是，在本章弱监督快速辨识定位方法中，我们计算的是预测区域与多级注意力提取网络所获得的伪标注区域的 IoU 值。那么，损失函数定义如下：

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{5.3}$$

其中，p_i 表示区域 i 是辨识性区域的预测结果，p_i^* 表示区域 i 的伪标注信息，由多级注意力提取网络生成的区域的矩形框 t_i^* 决定，t_i 是预测的矩形框，L_{cls} 表示分类损失，L_{reg} 表示回归损失，由鲁棒的 Smooth L_1 损失函数表示。

由于本章弱监督快速辨识定位方法采用了多级注意力，因此我们设计了多个定位网络，每一个都与 Fast R-CNN [82] 相同。所有的定位网络紧跟 RoI 池化层，用来对区域生成网络生成的每个候选区域生成一个固定大小的特征图向量，然后作为定位网络的输入，输出预测类别和预测区域的矩形框。通过辨识性定位网络，我们可以获得具有多级注意力的区域。然后，我们将每个区域的预测得分融合以获得最终的预测结果。

5.2.3 训练过程

多级注意力提取网络学习图像的多级注意力信息，指出图像中对于细粒度图像分类重要的区域，并用来指导辨识性定位网络的训练。辨识性定位网络中的区域生成网络生成候选的辨识性区域。考虑到训练区域生成网络需要多级注意力提取网路生成的伪标注信息，我们采用共享二者卷积层参数的策略以促进定位学习。

第一步，训练多级注意力提取网络，其先在 ImageNet 1K 数据集[39] 上进行预训练，然后再在特定的细粒度图像数据集上进行微调。第二步，训练区域生成网络，其卷积层的参数由多级注意力提取网路的参数进行初始化，并在训练过程中进行微调，而不

表 5.1　细粒度图像分类速度对比结果

方法	平均速度 (FPS)	CNN
本章 WSFDL 方法	**9.09**	VGGNet
Bilinear-CNN [18]	4.52	VGGNet&VGG-M
TSC [89]	0.34	VGGNet
TL Atten [8]	0.25	VGGNet
NAC [38]	0.10	VGGNet
本章 WSFDL 方法	**16.13**	AlexNet
Part-stacked CNN [14]	14.30	AlexNet

是固定不变。第三步，训练定位网络。由于定位网络都共享由区域生成网络提取的特征，在训练定位网络时，我们将区域生成网络的参数固定。

5.3　实验结果与分析

本章在 CUB-200-2011 和 Cars-196 两个广泛使用的细粒度图像数据集上进行实验验证，分别从速度和准确率两个方面来验证本章弱监督快速辨识定位方法的有效性。

5.3.1　实验设置

本章弱监督快速辨识定位方法中的多级注意力提取网络和辨识性定位网络都是基于 VGGNet[43] 的。值得注意的是，该基础网络可以被替换为其他 CNN 网络。对于多级注意力提取网络中，将 VGGNet 模型 conv5_3 以后的层剔除，从而获得 14×14 的空间分辨率。此外，增加一层卷积层，其包含 1024 个卷积核，核大小为 3×3，步长为 1，填充（Padding）为 1。其后紧跟全局平局池化层和 Softmax 层。辨识性定位网络则与多级注意力网络的卷积层共享参数以获得更好的定位和分类效果。

在训练过程中，对于多级注意力提取网络，首先在 ImageNet 1K 数据集 [39] 上进行预训练。然后，在本章所用的数据集上进行微调。具体地，在微调过程中，批量大小（Batch Size）设为 20，权重衰减系数（Weight Decay）设为 0.0005，动量（Momentum）设为 0.9，初始学习率为 0.001，每 5K 次迭代学习率下降 10 倍。在 CUB-200-2011 数据集上训练了 35K 次迭代，在 Cars-196 数据集上训练了 55K 次迭代。因为辨识性定位网络包括 1 个区域生成网络和 n 个定位网络，因此，在训练过程中，依次训练每一个定位网络。首先，利用多级注意力提取网络的卷积层参数初始化区域生成网络，然后再依此训练定位网络。在训练定位网络时，区域生成网络的参数是固定的，只对定位网络的参数进行微调。在微调过程中，批量大小（Batch Size）设为 128，权重衰减系数（Weight Decay）设为 0.0005，动量（Momentum）设为 0.9，初始学习率为 0.001。在 CUB-200-2011 数据集上，每 40K 次迭代学习率下降 10 倍，共训练 90K 次迭代。在

Cars-196 数据集上，每 50K 次迭代学习率下降 10 倍，共训练 120K 次迭代。

5.3.2 与现有方法进行对比

本章从以下两个方面来验证弱监督快速辨识定位方法（表示为 WSFDL) 的有效性：细粒度分类速度和准确率。实验结果表明，与现有方法相比，本章 WSFDL 不仅在细粒度分类速度上，而且在细粒度分类准确率上都取得了最好的结果。

5.3.2.1 细粒度分类速度

表5.1展示了本章 WSFDL 方法与现有方法在细粒度分类平均速度上的结果比对。平均速度是通过计算每秒识别图像数目来评测的，表示为 FPS（Frames Per Second）。由于细粒度分类速度并不受数据集的影响，因此本章以 CUB-200-2011 数据集为例进行验证。实验在一台拥有一个 GPU（NVIDIA TITAN X @1417MHZ）和一个 CPU（Intel Core i7-6900K @3.2GHZ）的台式计算机上运行。与现有方法相比，本章 WSFDL 方法不仅在细粒度分类准确率上，而且在细粒度分类速度上同样取得最好的结果。根据所使用的的 CNN 模型的不同，将现有方法划分为两组：VGGNet [43] 和 AlexNet [94]。细粒度分类速度除了与所使用的硬件环境有关，还与算法的实现方式有关。不同的实现方式会取得不同的细粒度分类速度。为了公平对比，对于现有方法，我们直接运行其作者提供的源码，与本章 WSFDL 方法在相同硬件条件下运行。其中，由于 Part-stacked CNN 方法[14] 并未提供源码，为了公平对比，我们进行了相应换算。其论文中报告的细粒度分类平均速度为 20 FPS，使用的模型为 CaffeNet[95]，GPU 型号为 NVIDIA Tesla K80。我们在相同情况下，模型使用 CaffeNet，GPU 使用 NVIDIA Tesla K80，细粒度分类平均速度为 35.75 FPS。因此我们计算得到其模型在我们环境下的平均速度为 20×35.75÷50=14.30 FPS。与使用 VGGNet 的方法相比，本章 WSFDL 方法与对比房中中最快速度的 Bilinear-CNN 方法相比快了 2 倍，而且在细粒度分类准确率上比它高 1.61%（85.71% vs. 84.10%）。此外，与其他基于时间消耗巨大的区域生成（即 Selective Search 方法）的方法（如 TSC [89]、TL Atten [8] 和 NAC [38]）相比，本章 WSFDL 方法在平均速度上取得了两个数量级的提升。当使用模型由 VGGNet 替换为 AlexNet，本章 WSFDL 方法依旧比对比方法中最快的 Part-stacked CNN 方法要快。而且，需要注意的是，Part-stacked CNN 方法不仅使用了对象级标注，而且还使用了部件级标注。而本章 WSFDL 方法未使用对象级和部件级标注信息，通过辨识性定位网络避免了时间消耗巨大的区域生成，通过定位与细粒度分类的交互学习进一步促进了细粒度分类准确率的提升。这使得本章 WSFDL 方法具有更好的实用性。此外，本章 WSFDL 方法与第三章、第四章的方法相比，在平均速度上都取得了至少两个数量级的提升，而且在分类准确率上仅有小幅下降。其中，与第三章的对象-部件注意力模型相比，在 CUB-200-2011 数据集上细粒度图像分类准

表 5.2　CUB-200-2011 数据集上的细粒度分类结果

方法	训练集标注		测试集标注		准确率 (%)	CNN
	对象级	部件级	对象级	部件级		
本章 WSFDL 方法					**85.71**	VGGNet
TSC [89]					84.69	VGGNet
FOAF [34]					84.63	VGGNet
PD [7]					84.54	VGGNet
STN [44]					84.10	GoogleNet
Bilinear-CNN [18]					84.10	VGGNet&VGG-M
PD (FC-CNN) [7]					82.60	VGGNet
Multi-grained [45]					81.70	VGGNet
NAC [38]					81.01	VGGNet
PIR [37]					79.34	VGGNet
RBF [96]					78.98	ResNet-50
TL Atten [8]					77.90	VGGNet
MIL [46]					77.40	VGGNet
VGG-BGLm [22]					75.90	VGGNet
InterActive [47]					75.62	VGGNet
Coarse-to-Fine [49]	√		√		82.90	VGGNet
PG Alignment [15]	√		√		82.80	VGGNet
Coarse-to-Fine [49]	√				82.50	VGGNet
VGG-BGLm [22]	√		√		80.40	VGGNet
Triplet-A (64) [50]	√		√		80.70	GoogleNet
Triplet-M (64) [50]	√		√		79.30	GoogleNet
AGAL [97]		√ + 属性标注			85.40	ResNet-50
Webly-supervised [51]	√	√			78.60	AlexNet
PN-CNN [36]	√	√			75.70	AlexNet
Part-based R-CNN [5]	√	√			73.50	AlexNet
AGAL [97]	√	√ + 属性标注			85.50	ResNet-50
SPDA-CNN [52]	√	√	√		85.14	VGGNet
Deep LAC [53]	√	√	√		84.10	AlexNet
SPDA-CNN [52]	√	√	√		81.01	AlexNet
Part-stacked CNN [14]	√	√	√		76.20	AlexNet
PN-CNN [36]	√	√	√	√	85.40	AlexNet
Part-based R-CNN [5]	√	√	√	√	76.37	AlexNet
POOF [54]	√	√	√	√	73.30	
HPM [55]	√	√	√	√	66.35	

确率仅下降了 0.12%。

5.3.2.2　细粒度分类准确率

本章 WSFDL 方法不仅能够取得最快的速度，而且在细粒度分类准确率上也有很好的表现。表5.2和表5.3展示了在 CUB-200-2011 和 Cars-196 两个数据集上的细粒度图像分类结果。其中，所使用的对象级、部件级标注以及 CNN 网络模型也列举在表格中，以示公平对比。本文 WSFDL 方法仅使用了图像级的类别标注信息，但依然取得了最

表 5.3　Cars-196 数据集上的细粒度分类结果

方法	训练集标注		测试集标注		准确率 (%)	CNN
	对象级	部件级	对象级	部件级		
本文 WSFDL 方法					**92.30**	VGGNet
Bilinear-CNN [18]					91.30	VGGNet&VGG-M
TL Atten [8]					88.63	VGGNet
DVAN [58]					87.10	VGGNet
FT-HAR-CNN [59]					86.30	AlexNet
HAR-CNN [59]					80.80	AlexNet
PG Alignment [15]	√				92.60	VGGNet
SWP [98]	√				92.30	ResNet-50
ELLF [60]	√				73.90	CNN
R-CNN [35]	√				57.40	AlexNet
PG Alignment [15]	√		√		92.80	VGGNet
BoT(CNN With Geo) [61]	√		√		92.50	VGGNet
DPL-CNN [62]	√		√		92.30	VGGNet
VGG-BGLm [22]	√		√		90.50	VGGNet
BoT(HOG With Geo) [61]	√		√		85.70	VGGNet
LLC [63]	√		√		69.50	
BB-3D-G [28]	√		√		67.60	

好的细粒度图像分类结果。

本章 WSFDL 相比现有方法中最好的方法 TSC [89] 取得了 1.02% 的准确率提升（85.71% vs. 84.69%）。此外，在分类速度上快了 27 倍，如表5.1所示。本章 WSFDL 方法与 Bilinear-CNN [18] 相比取得了 1.61% 准确率上的提升。此外，与使用对象级或者部件级标注的方法相比，本章 WSFDL 方法同样取得了更好的细粒度图像分类准确率。不使用对象级和部件级标注信息，使得本章 WSFDL 方法能更好地转化为实际应用。而且多级注意力的应用使得本章 WSFDL 方法能够进一步促进辨识性区域定位并取得更好的细粒度图像分类准确率。

在 Cars-196 数据集上的细粒度图像分类结果如表5.3所示。其趋势与 CUB-200-2011 一致，本章 WSFDL 方法取得了最好的细粒度图像分类准确率，取得了 1.00% 的准确率提升。

5.3.3　本章 WSFDL 方法中每个组成模块的有效性

5.3.3.1　多级注意力在细粒度分类准确率上的有效性

在本章 WSFDL 方法中，多级注意力被采用，不同的注意力聚焦于图像中具有不同特点和辨识度的区域。它们互不相同却又互补促进，能够提升图像的辨识性特征表示，从而获取更好的细粒度分类准确率。在本章实验中，我们提取了“Conv4_3”、“Conv5_3”和 “Conv_cam” 三个卷积层的特征，并对其有效性进行了评测。从表5.4可以看出，三

种不同注意力的结合能够进一步促进细粒度图像分类的准确率，由此可以证明上述不同卷积层（分别对应不同注意力）辨识性特征具有互补性。其中，“Conv4_3’’ 层在细粒度分类中起到了相对较小的作用，而且采用三级注意力会使得时间消耗较大。因此，在本章实验中，我们仅采用了两级注意力，即 “Conv5_3’’ 和 “Conv_cam’’，以取得细粒度分类准确率和速度上的平衡。

表 5.4　多级注意力在细粒度分类准确率上的有效性

卷积层	准确率 (%)	
	CUB-200-2011	Cars-196
Conv_cam	83.45	89.59
Conv5_3	81.15	84.31
Conv4_3	77.84	78.01
Conv_cam + Conv5_3	84.43	90.29
Conv_cam + Conv4_3	84.36	90.10
Conv4_3 + Conv5_3	81.41	84.68
Conv_cam + Conv5_3 + Conv4_3	84.59	90.30

5.3.3.2　辨识性定位网络在细粒度分类速度上的有效性

由于本章 WSFDL 方法采用了多级注意力，因此有两个选择：1）分别训练多个定位网络，每个定位网络包含 RPN 和 Fast R-CNN，在表5.5中用 “two-level (respectively)’’ 和 “three-level (respectively)’’ 表示。从表中可以看到，这带来了时间消耗的近线增长。2）在本章 WSFDL 方法中，我们设计了端到端的多路辨识性定位网络，包含一个 RPN 和多个定位网络。所有定位网络共享 RPN 生成的候选区域，因此能够有效避免时间消耗的线性增长。在表5.5中表示为 “two-level (DLN)’’ 和 “three-level (DLN)’’。我们可以看到，本章设计的辨识性定位网络能够有效减少时间消耗。

表 5.5　辨识性定位网络在细粒度分类速度上的有效性

方法	平均速度 (FPS)
one-level	10.07
two-level (respectively)	5.04
two-level (DLN)	9.09
three-level (respectively)	3.36
three-level (DLN)	7.69

5.3.4　基线实验

本章 WSFDL 方法是基于 Faster R-CNN [91]、多级注意力提取网络（MAEN）和 VGGNet [43] 的。为了验证其有效性，本节实验展示了与 Faster R-CNN、MAEN 和 VGGNet

在 CUB-200-2011 数据集上的对比结果，如表5.6所示。其中，“VGGNet’’ 表示直接使用在细粒度图像数据集上微调后的 VGGNet 获得的结果，“MAEN’’ 表示采用本章多级注意力提取网络获得的结果，“Faster R-CNN (gt)’’ 表示基于 Faster R-CNN 方法使用了对象级标注信息的结果。从表中可以看到，本章 WSFDL 方法取得了最好的细粒度图像分类结果。在本章 WSFDL 方法中，VGGNet 被采用作为基础网络，但是其结果仅为 70.42%，严重低于本章 WSFDL 方法。这表明了辨识性定位及特征的学习，发现图像中对于细粒度分类有帮助的重要区域，通常包含了与其他细粒度子类别的关键区别，能够有效地促进细粒度图像分类。与 “Faster R-CNN (gt)’’ 相比，本章 WSFDL 方法依然能够取得更好的细粒度分类结果。这是一个令人鼓舞且值得思考的现象。从图5.2从后一行可以看出，并非所有对象级标注（红色矩形框）的区域都是有助于细粒度图像分类的。有些标注区域包含了大面积的背景区域，这些背景区域可能会使得细粒度分类模型受到干扰从而造成误分。因此，从图像中定位到具有辨识性的区域对于提升细粒度图像分类准确率是至关重要的。从表5.7中可以看到，MAEN 与本章 WSFDL 方法有较为接近的定位准确率，但是从表5.6可以发现其有较低的细粒度分类准确率。这主要是因为 MAEN 和 WSFDL 不同的学习能力。在训练阶段，MAEN 仅仅从原始图像中进行学习。而本章 WSFDL 方法首先对于每张图像生成多个候选图像块，然后这些图像块进一步驱动学习多尺度、多粒度且更具辨识性的特征，从而取得更好的细粒度定位和分类。

表 5.6　基线实验对比

方法	细粒度分类准确率 (%)
本章 WSFDL 方法	**85.71**
Faster R-CNN (gt)	82.46
MAEN	77.50
VGGNet	70.42

5.3.5　辨识性定位的有效性

本章 WSFDL 方法聚焦于同时提高细粒度定位和分类的效果。由于辨识性区域通常定位在图像中的对象区域，因此我们采用所定位到的区域与对象级标注信息的 IoU 值来评测定位的准确性。如果 IoU 值超过 0.5，则我们判定定位正确。表5.7展示了基于 “Conv_cam’’ 卷积层的定位结果。可以看到，本章 WSFDL 方法在 CUB-200-2011 和 Cars-196 两个数据集上分别取得了 46.05% 和 56.60% 的定位准确率。考虑到并未使用对象级标注信息，因此这是一个不错的定位结果。此外，由于本章 WSFDL 方法是将由 MAEN 生成的辨识性区域作为伪标注信息进行二次定位学习，因此，其取得了更好的定位效果。

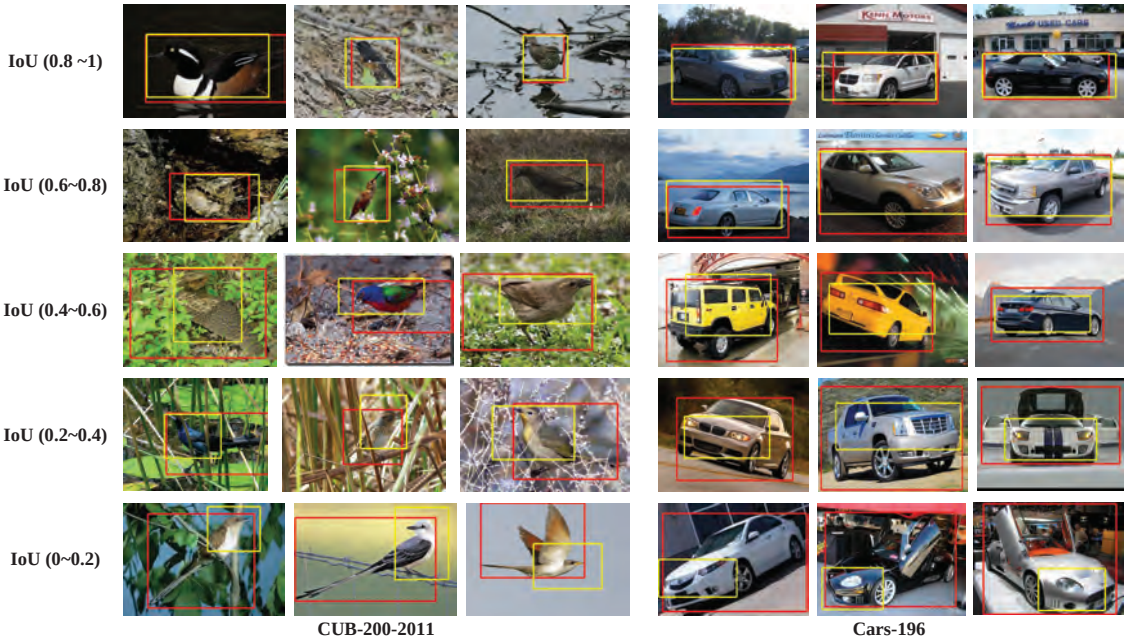


图 5.2　辨识性定位网络定位到的辨识性区域与对象级标注的对比

表 5.7　定位结果对比

方法	定位准确率 (%)	
	CUB-200-2011	Cars-196
本章 WSFDL 方法	**46.05**	**56.60**
MAEN	44.93	55.79

根据定位到的辨识性区域与对象级标注的IoU值的范围，即0~0.2、0.2~0.4、0.4~0.6、0.6~0.8、0.8~1，将 CUB-200-2011 和 Cars-196 两个数据集上的定位结果进行了展示，如图5.2所示。在本章 WSFDL 方法中，存在一些定位到的辨识性区域与对象级标注的 IoU 值低于 0.5，但是从图中可以看到这些区域都包含了对象中的辨识性区域，如头部、躯干等，同样是有助于细粒度图像分类的。这证明了本章 WSFDL 方法能够通过定位对象的辨识性区域来获得细粒度图像分类准确率的提升。

图5.3展示了本章 WSFDL 方法通过 “Conv_cam’’ 和 “Conv5_3’’ 两个卷积层的注意力获取的辨识性区域。不同的注意力聚焦于图像中的不同区域，通过提供互补的信息促进细粒度图像分类准确率。为了进一步验证辨识性定位的有效性，我们做了定量实验，即计算正确定位比率（Percentage of Correctly Localization，PCL）。其表示定位到的辨识性区域是否包含了对象的部件，结果如表5.8所示。CUB-200-2011 数据集提供了 15 个部件级标注信息，表示每个部件在图像中的位置，如背部、胸部、腹部等。从表5.8可以看出，本章 WSFDL 方法定位到的辨识性区域包含了 94.68% 的部件，验证了其能够定位到对象的辨识性区域，从而进一步提升细粒度图像分类的准确率。

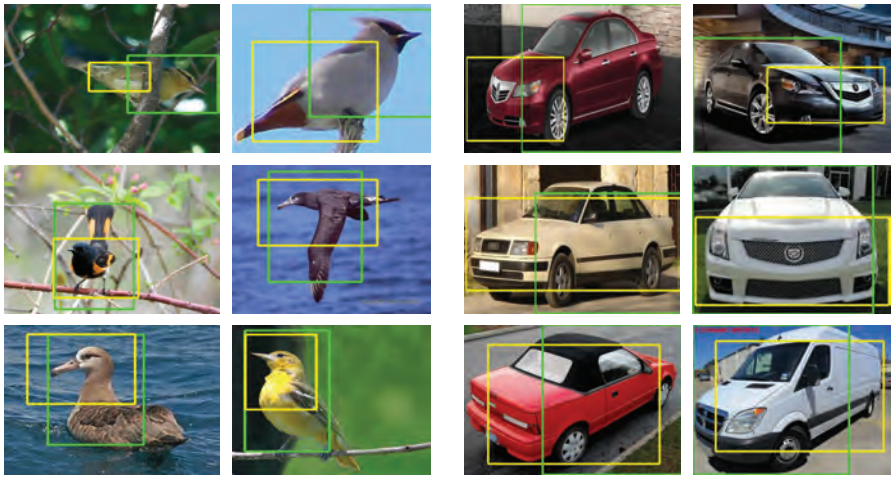


图 5.3　本章 WSFDL 方法中多级注意力定位到的区域结果

表 5.8　CUB-200-2011 数据集上对于每个部件的 PCL 值

部件	背部	鸟喙	腹部	胸部	鸟冠	前额	左眼	左腿
PCL (%)	96.33	96.49	94.00	95.29	97.38	97.07	97.49	89.92
部件	左翅膀	颈背	右眼	右腿	右翅膀	尾巴	喉咙	平均
PCL (%)	92.60	96.60	96.79	91.85	97.00	85.03	96.38	**94.68**

5.3.6　不同注意力的不同聚焦点

正如现有研究 [99, 100] 表明，不同的卷积层能够从简单的视觉元素（如边缘等）中提取出复杂的视觉概念（如部件、对象等）。不同层所表达的视觉概念也不尽相同，他们之间具有一定的互补性，能能够对特定的任务起到促进作用。我们对于不同的卷积层（“Conv_cam’’ 和 “Conv5_3’’）生成了辨识性区域的矩形框，可以看到他们互不相同。进一步，通过计算他们之间的 IoU 值，以验证不同注意力有不同的聚焦点。表5.9展示了 IoU 值超过某一阈值的百分比。可以看出，在 CUB-200-2011 和 Cars-196 两个数据集上，“Conv_cam’’ 和 “Conv5_3’’ 所定位到的辨识性区域之间 $IoU > 0.5$ 的百分比仅为 13.22% 和 4.44%。而当 $IoU > 0.7$ 时，则百分比更小，在 Cars-196 数据集上甚至没有。这都证明了本章 WSFDL 方法中不同注意力所提取的辨识性区域不尽相同，它们互相促进，从而为细粒度图像分类提供多粒度的辨识性特征。

表 5.9　不同 IoU 值的百分比

数据集	IoU>0.5	IoU>0.6	IoU>0.7	IoU>0.8	IoU>0.9
CUB-200-2011	13.22%	6.13%	2.28%	0.50%	0
Cars-196	4.44%	0.85%	0	0	0

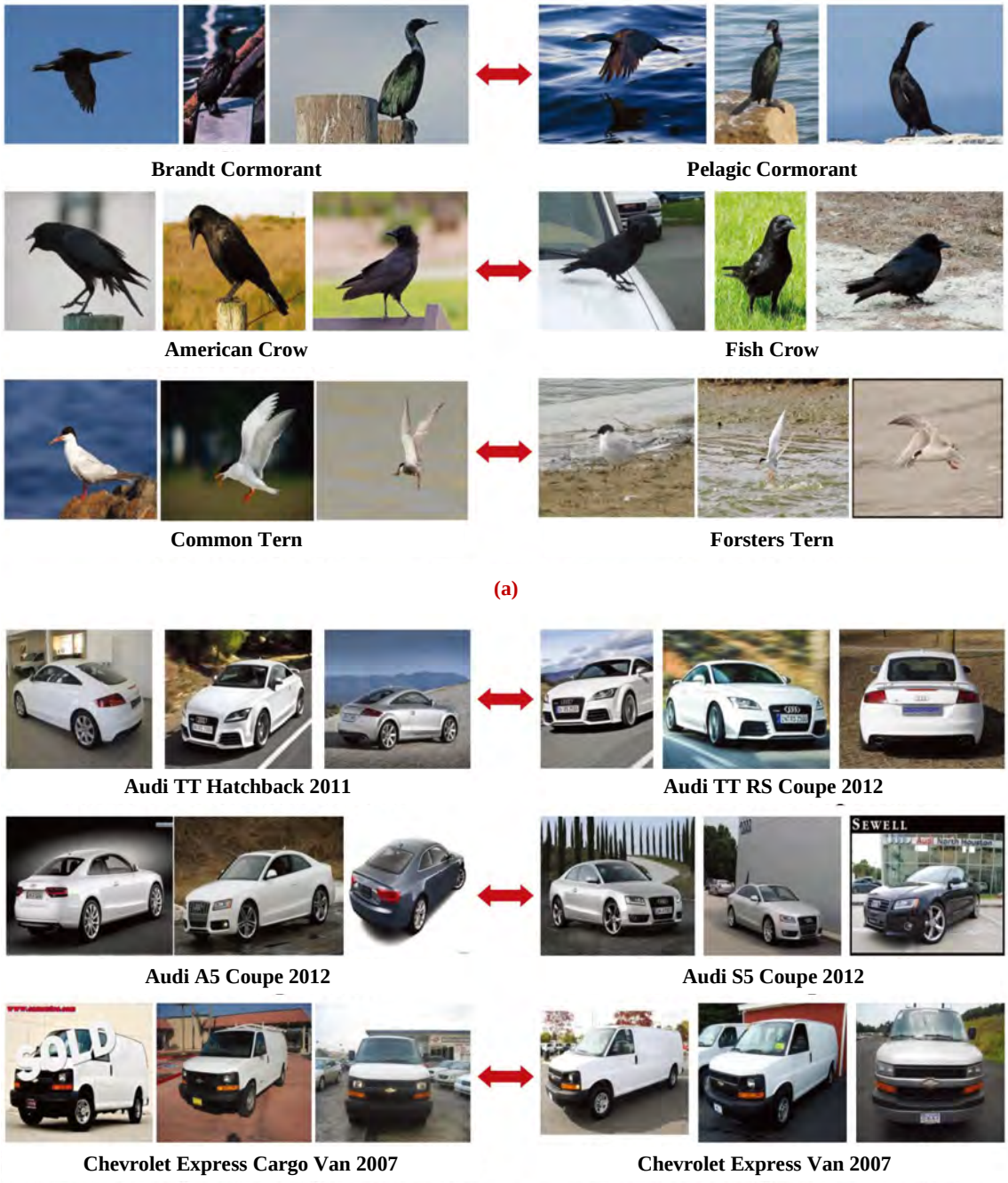


图 5.4　CUB-200-2011 和 Cars-196 两个数据集上最容易误分的细粒度子类别对

5.3.7　错误细粒度分类分析

图5.5展示了 CUB-200-2011 和 Cars-196 两个数据集上的细粒度分类的混淆矩阵。其中，坐标轴表示细粒度子类别，不同的颜色代表不同的误分可能性。由于相似的细粒度子类别在数据集中拥有邻近的类别 ID，因此混淆矩阵中，最容易错分的细粒度子类

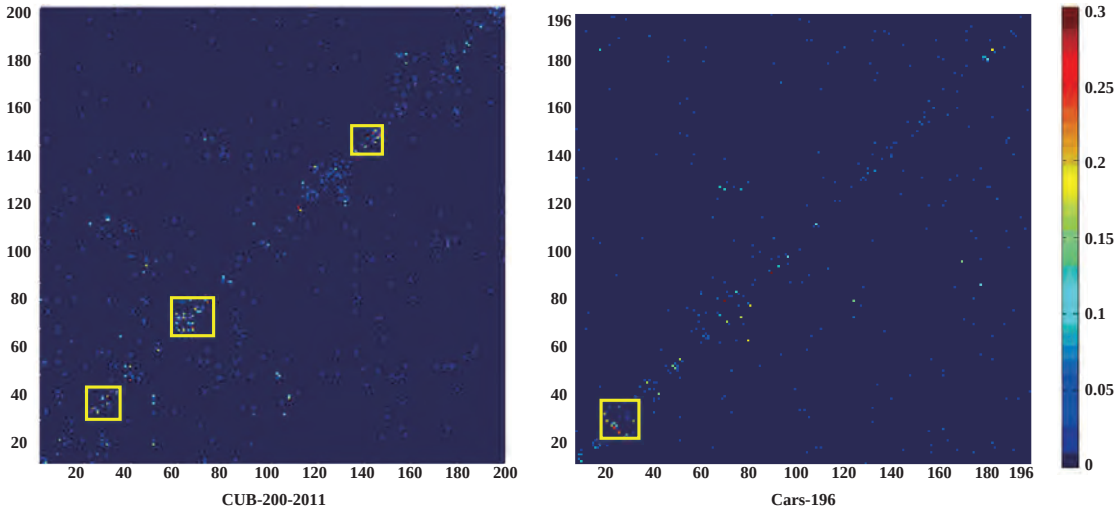


图 5.5　CUB-200-2011 和 Cars-196 两个数据集上的分类混淆矩阵

别对位于对角线上，如黄色矩形框所示。由于相似的细粒度子类别在视觉上具有较小的差异，即使是人类也很难进行区分，这正是细粒度图像分类的挑战性。图5.4展示了两个数据集上最容易误分的细粒度子类别对。同一行的两个子类在外观上很相近，因此容易误分。例如，图5.4第三行左边所示的 “Common Tern’’ 和 “Forsters Tern’’ 在外观上很难进行区分，它们有相近的属性，如白色翅膀和黑色的额头。其误分主要是因为所考虑的辨识性区域还不够局部、细微和全面。在每个卷积层中可能存在多个一个辨识性区域，但在本章 WSFDL 方法中为了取得更快的辨识速度仅考虑具有最大面积的辨识性区域，这在一定程度上忽略了一些辨识性特征，从而导致一些误分情况。

5.4　本章小结

本章提出了弱监督快速辨识定位方法。首先，通过多级注意力引导辨识性定位学习来同时定位多个辨识性区域。再此过程中只使用了图像级的类别标注，避免了成本巨大的对象级和部件级标注的使用。然后，通过多路端到端辨识性定位网络定位辨识区域并进行区域特征学习，不仅提升了辨识速度，同时保证了细粒度图像分类准确率。

第六章　基于细粒度分类的跨媒体检索

6.1　引言

随着互联网和多媒体技术的迅猛发展，图像、文本、视频、音频等跨媒体数据已经成为当前信息传播的主要形式。但是，在前面几章中，本文主要从细粒度图像分类中的辨识性特征学习展开细粒度分析研究，即视觉信息的辨识性特征学习。忽略了与其相关联的文本、视频、音频等跨媒体数据，这些跨媒体数据中存在着隐含的语义关联关系，通过分析这种跨媒体隐含语义关联关系，能够进一步促进细粒度分析。此外，随着数据数量的增长、模态的增多，亟需一种有效的多媒体检索方式来对跨媒体数据进行有效的管理与利用。因此，本章将从数据、任务上对细粒度分析进行扩展，将图像扩展到跨媒体，分类扩展到检索，在细粒度跨媒体检索上展开研究。

跨媒体检索（Cross-media Retrieval）[101] 正是这样一种有效的检索方式，指用户给定任意一种媒体类型数据作为查询样例，系统检索得到与查询样例相关的各种媒体数据。如图6.1所示，当用户给定一张灰背鸥（Slaty-backed Gull）的图像作为查询样例，检索结果包含了图像、文本、视频和音频 4 种媒体数据。

现有跨媒体检索研究一般聚焦在粗粒度跨媒体检索（Coarse-grained Cross-media Retrieval），只是将灰背鸥的图像作为鸟的图像进行分析检索，因此检索结果中会包含各种相似鸟类的媒体数据（如灰翅鸥、银鸥、加州海鸥等），而不是灰背鸥的图像、文本、视频和音频数据，如图6.2(a) 所示。为了克服上述问题，本章提出了细粒度跨媒体检索（Fine-grained Cross-media Retrieval），即用户给定任意一种媒体类型数据作为查询样例，系统检索得到与查询样例细粒度类别相同的各种媒体数据，如图6.2(b) 所示，检索得到灰背鸥的图像、文本、视频和音频数据。

作为一个新兴的研究方向，细粒度跨媒体检索面临三大挑战：

- **缺乏数据集和评测基准**：现有跨媒体数据集一般是针对粗粒度跨媒体检索，而细粒度跨媒体检索还缺乏数据集和评测基准，因此相关研究比较少。
- **异构鸿沟**：这是跨媒体检索面临的经典难题，是指不同媒体类型的数据有着不同的分布和特征表示，导致跨媒体检索十分困难，对细粒度跨媒体检索更是难上加难。
- **类间差异小，类内差异大**：这是细粒度分析面临的挑战。其中，类间差异小是指不同的细粒度类别具有相似的外表（图像、视频）、描述（文本）和声音（音频）；类内差异大是指由于视角、光照、描述、背景等不同，相同的细粒度类别又存在外表、描述和声音差异大的现象。上述问题导致难以准确检索特定细粒

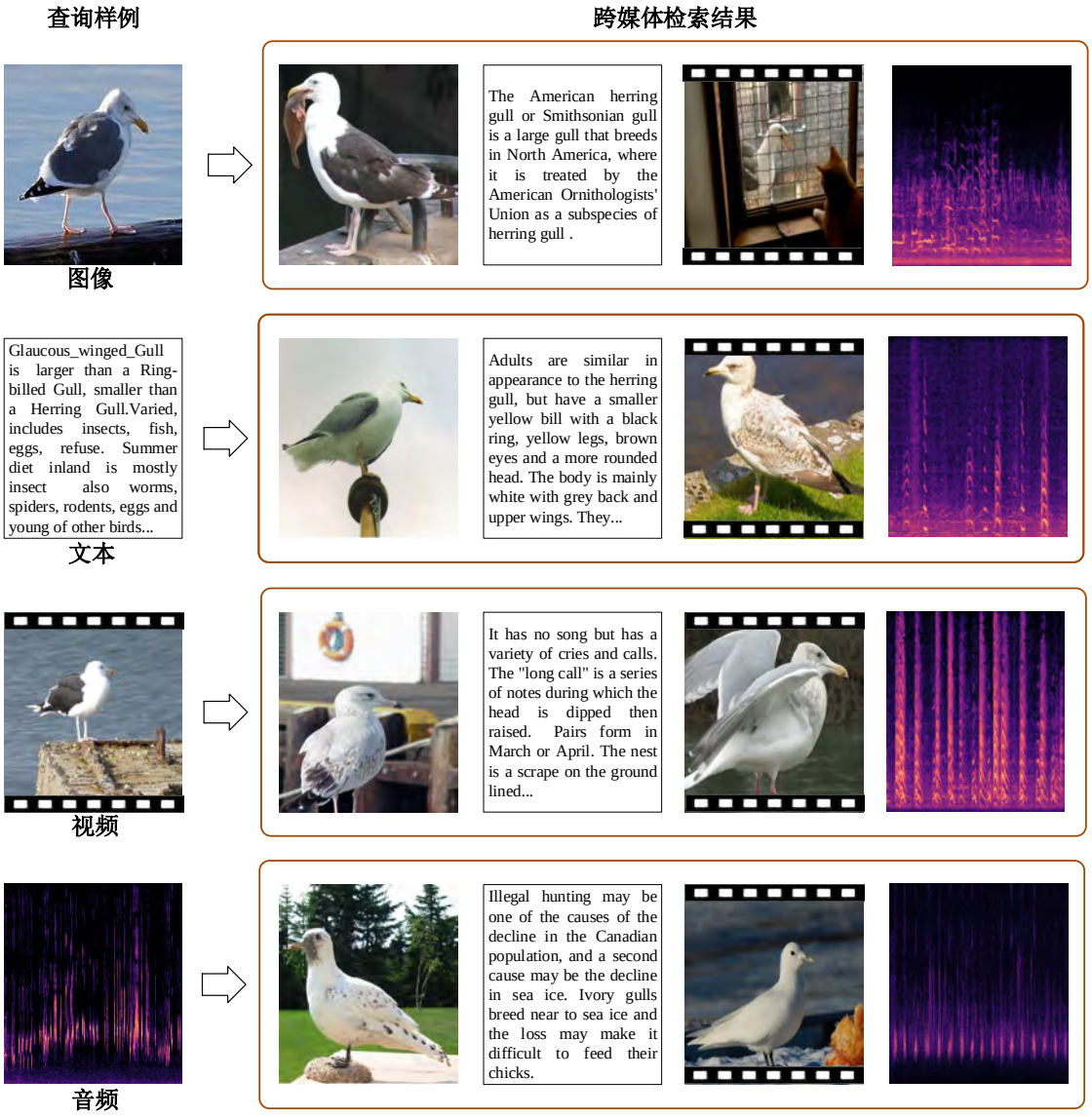


图 6.1　跨媒体检索示意图

度类别的媒体数据，相比粗粒度跨媒体检索更具挑战。

为了解决上述问题，本章首先建立了首个包含 4 种媒体类型（图像、文本、视频和音频）的细粒度跨媒体检索公开数据集和评测基准 PKU FG-XMedia，其次提出了一种能够同时学习 4 种媒体统一表征的深度网络模型 FGCrossNet。其贡献归纳如下：

- **建立了首个包含 4 种媒体类型（图像、文本、视频和音频）的细粒度跨媒体检索公开数据集和评测基准 PKU FG-XMedia**：其有三个优势：1）类别多样性，包含鸟的 200 个细粒度子类别，如灰背鸥、银鸥、加州海鸥和灰翅鸥等；2）媒体多样性，包括图像、文本、视频和音频 4 种媒体类型，据我们所知，这是第一个包含 4 种媒体类型（图像、文本、视频和音频）的细粒度跨媒体检索公开数据

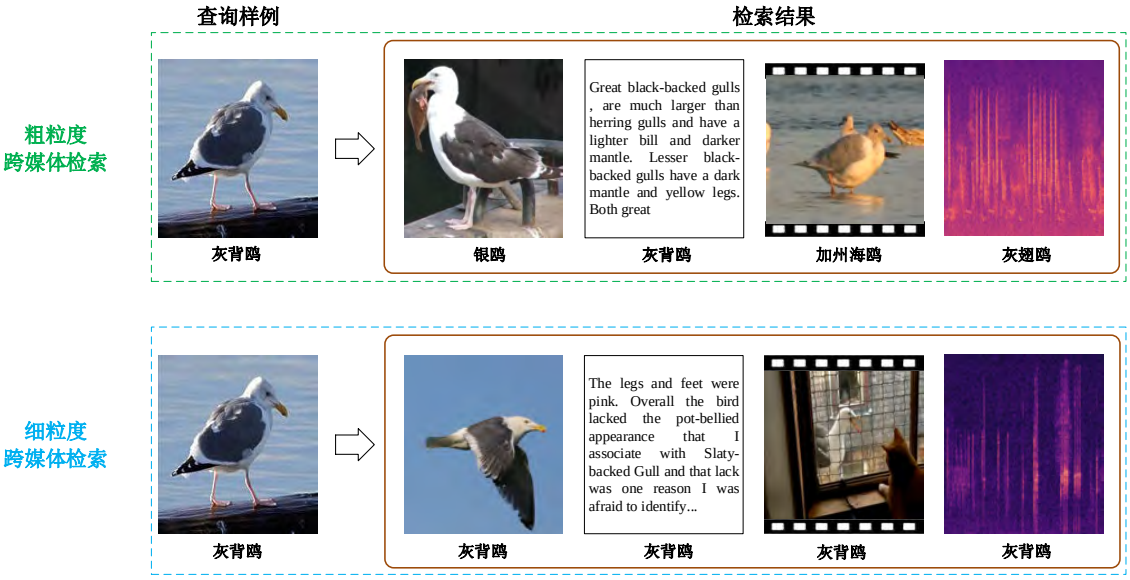


图 6.2　粗粒度跨媒体检索 VS 细粒度跨媒体检索

集。3）源域多样性，数据来源于不同的网站，导致数据质量不同，因此增加了检索的难度。数据集和评测基准已经公布，以方便研究者们在细粒度跨媒体检索上进一步探索。

- **提出了能够同时学习 4 种媒体统一表征的深度网络模型 FGCrossNet**：其联合考虑了三种约束来促进具有辨识性的统一表征学习：1）分类约束，确保细粒度子类别统一表征的辨识性；2）中心约束，确保相同细粒度子类别统一表征的紧凑性；3）排序约束，确保不同细粒度子类别统一表征的松散性。FGCrossNet 在一个统一的分类网络中同时建模这 4 种媒体数据，通过上述三种约束的联合优化，使得模型能够在学习分类的过程中一次性学习 4 种媒体的统一表征。

6.2　细粒度跨媒体检索数据集和评测基准

现有研究者针对粗粒度跨媒体检索任务构建了很多数据集和评测基准，他们的统计信息如表6.1所示。具体地，Rasiwasia 等人构建了最广泛使用的跨媒体数据集 Wikipedia [102]，其包含 10 个粗粒度类别的 2,866 个图像/文本对，如“历史”、“战争”等。Rashtchian 等人从 Pascal VOC 2008[103] 数据集中选取了 1,000 张图像，并且为每一张图像标注了 5 个句子，以此构建了 Pascal Sentences 数据集[104]。随后，一些数据量规模较大的跨媒体检索数据集被建立，以促进粗粒度跨媒体检索技术的发展，如 Flickr-30K 数据集 [105] 和 MS-COCO 数据集 [106]。在上述数据集中，文本信息都是用句子话或段落来表示。而 Chua 等人构建了 NUS-WIDE 数据集 [107]，从互联网中采集了 81 个粗粒度类别的 269,648 张

表 6.1 本章细粒度跨媒体检索数据集与现有常用粗粒度跨媒体检索数据集对比

	Wikipedia	Pascal Sentences	Flickr-30K	MS-COCO	NUS-WIDE	PKU XMediaNet	本章 PKU FG-XMedia
# 图像	2,866	1,000	31,783	123,287	269,648	40,000	11,788
# 文本	2,866	5,000	158,915	616,435	5,018	40,000	8,000
# 视频	N/A	N/A	N/A	N/A	N/A	10,000	18,350
# 音频	N/A	N/A	N/A	N/A	N/A	10,000	12,000
# 类别	10	20	N/A	91	81	200	200
细粒度?	否	否	否	否	否	否	是

图像，每张图像包括多个标签。整个数据集包含 5,018 个不同的标签，以此来表示对应图像的文本信息。上述数据集都只包含两种数据类型，即图像和文本。

为了更全面的评测并促进粗粒度跨媒体技术的发展，Peng 等人提出了 PKU XMediaNet 数据集 [101]。这是目前包含 5 种媒体类型（图像、文本、视频、音频和 3D 图形）的最大跨媒体数据集。它包含 200 个粗粒度类别的 100,000 个样本。它的类别由 WordNet ①选取而来，涵盖了 47 个动物类别（如鸟、狗等）以及 153 个人工产品（如飞机、车等）。但是上述所有数据集所涉及的类别均是粗粒度的，而在我们日常生活中更加趋向于知道或获取具体的细粒度子类别信息，因此它们均不能满足人类细粒度跨媒体检索的需求。

因此，在本章我们构建了一个新的细粒度跨媒体检索公开数据集和评测基准 PKU FG-XMedia，包含 4 种媒体类型，即图像、文本、视频和音频；包含 200 个鸟类细粒度子类别。在接下来的段落，我们从以下三个方面进行介绍：数据集的构建、特点以及细粒度跨媒体检索任务。

6.2.1 数据集的构建

我们从互联网上采集了图像、文本、视频和音频等数据来构造本章的细粒度跨媒体数据集。受到相关细粒度图像/视频分类工作 [3, 10] 的启发，我们构造了一个包含 200 个鸟类细粒度子类别的数据集。依据相同的分类法，研究者们已经构造了包含 200 个相同鸟类细粒度子类别的图像和视频数据集，即 CUB-200-2011 图像数据集 [3] 和 YouTube Birds 视频数据集 [10]。因此，我们基于这两个数据集，直接利用它们作为图像和视频数据，以此来构造细粒度跨媒体数据集。这两个数据集简单介绍如下：

CUB-200-2011 图像数据集 [3]：这是最广泛使用的细粒度图像分类数据集，包含 200 个细粒度子类别和 11,788 张图片，来源于 Flickr 图像网站②。其中，训练集包含 5,994 张图片，测试集包含 5,794 张图片。对于每一张图片，有 4 种人工标注信息：1 个图像级的类别标签、1 个对象级位置信息、15 个部件级位置信息以及 312 个属性信息。需要注意的是，在本章的细粒度跨媒体检索任务中只使用了图像级的类别标签这一种人

① wordnet.princeton.edu/
② https://www.flickr.com/

工标注信息，其余三种标注信息并未使用。

YouTube Birds 视频数据集 [10] ：这是新近构建的大规模细粒度视频数据集，共包含 18,350 个视频。与 CUB-200-2011 数据集相同，涵盖了 200 个鸟类的细粒度子类别，而且二者的细粒度子类别完全相同。视频数据来源于 YouTube 视频网站用户上传的真实视频，每个视频时长不超过 5 分钟。数据集的划分如下：训练集包含 12,666 个视频，测试集包含 5,684 个视频。每个视频仅有一个视频级的类别标签信息。

除了上述图像、视频数据，我们还需要收集文本和音频数据。由于这些数据很容易从互联网中获得，因此我们选取一些专业网站作为数据来源，如表6.2所示。在数据收集过程中主要包括收集和清洗两个步骤。

表 6.2　文本和音频的来源网站

数据	数据来源
文本	(1) www.wikipedia.org (2) www.allaboutbirds.org (3) www.audubon.org (4) birdsna.org (5) birds.fandom.com (6) nhpbs.org (7) ebird.org (8) mnbirdatlas.org (9) sites.psu.edu (10) www.birdwatchersdigest.com (11) folksread.com (12) neotropical.birds.cornell.edu
音频	(1) www.xeno-canto.org (2) www.bird-sounds.net (3) www.findsounds.com (4) freesound.org (5) www.macaulaylibrary.org (6) avibase.bsc-eoc.org (7) soundcloud.com

6.2.1.1　收集

文本数据收集：Wikipedia ①是目前最大的免费百科全书网站，由世界各地的志愿者创建、编辑和维护。从 Wikipedia 上，我们可以输入细粒度子类别作为查询关键词，这样很容易获得对应的文本描述信息。需要注意的是，这里的细粒度子类别与 CUB-200-2011 数据集一致。从 Wikipedia 上，我们获得了 200 个细粒度子类别的文本数据。但是每个类别的文本样例数量并不多。因此，为了得到更多的文本数据，我们采取了以下两种策略：1）选取更多的百科全书网站。除了 Wikipedia，我们又选取了其他 11 个专业网站作为文本数据来源，如 All About Birds、Audubon、Animal Spot 等，如表6.2所示。2）选用更多的查询关键词。目前我们仅使用了 CUB-200-2011 数据集中的图像级类别标注作为查询关键词。但实际上，许多鸟类都有其对应的学名或者别名，将这些作为查询关键词可以获得更多的文本数据。例如，“Black-footed Albatross’’ 对应的学名为 “Phoebastria Nigripes Audubon’’。

① www.wikipedia.org/

音频数据收集：与文本数据收集一样，我们同样选取了多个专业的音频网站作为音频数据的来源，如 xeno-canto ①和 Bird-sounds ②，它们发布了世界各地鸟类的声音。为了获取更多的音频数据，我们同样采取了两种策略：1）选取更多的专业网站。我们一共选取了 7 个网站，如表6.2所示。2）选用更多的查询关键词。

6.2.1.2 数据清洗

文本数据清洗：在我们收集到的数据中，有一些数据并非是我们所需要的。首先，我们将网页中的链接从文本数据中剔除。然后，将一个文本网页的信息依据段落进行划分，每个文本段落作为一个文本样例，即最终的文本数据。因为这些文本数据是从专业的百科全书网站上搜集的，因此它们是已经被标注好的。

音频数据清洗：由于一些所收集到的音频数据的时长过长，如超过一个小时，因此，我们将一个音频片段划分为多个，以获取更多的音频样例。但是，这样的划分会导致一些音频数据并不包含鸟类的声音，因此我们邀请人工标注者将这些音频片段剔除。需要注意的是，一些音频数据并非只包含了鸟类的声音，还包含了其他声音，如人类的说话声，风声等。这从另一个方面增加了细粒度跨媒体检索的挑战性。

经过数据的收集与清洗，我们构建了最终的细粒度跨媒体检索数据集，其样例如图6.3所示。

6.2.2 特点

6.2.2.1 数据规模大

从表6.1中可以看到，本章构建的新细粒度跨媒体检索数据集包含了 4 种媒体类型，即图像、文本、视频和音频。在媒体类型数目上仅比 PKU XMediaNet 数据集[101] 少 3D 图形的数据。而其他的跨媒体检索数据集只包括 2 种媒体类型，即图像、文本。此外，每个媒体类型的数据规模也很大，即 11,788 张图像，8,000 个文本，18,350 个视频片段和 12,000 个音频片段。对于文本数据，每个细粒度子类别包含 40 个文本样例。对于音频数据，每个细粒度子类别包含 60 个音频样例。

6.2.2.2 数据多样性

类别多样性：数据集包含 200 个鸟类的细粒度子类别，是包含媒体类型最多的、规模最大的细粒度跨媒体检索数据集。相似的细粒度子类别带来了类间差异小的挑战：它们有相似的外观（图像、视频），相似的文本描述（文本）以及相似的叫声（音频），因

① www.xeno-canto.org
② www.bird-sounds.net

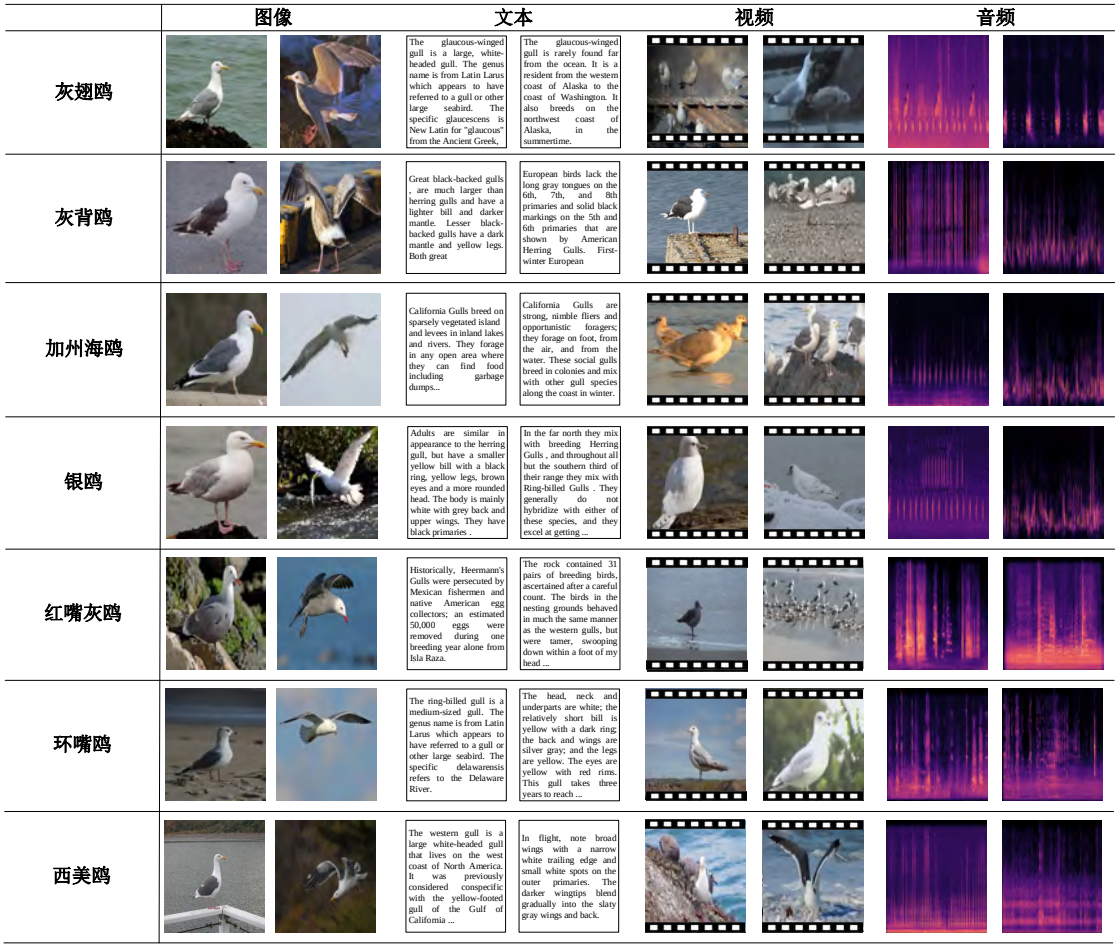


图 6.3　本章构造的细粒度跨媒体数据集中样例展示

此很难对相似的子类别进行区分。例如，如图6.3所示，即使属于不同细粒度子类别的图像，他们在外观上也极其相似。

源域多样性：所有的数据都是从不同的来源收集到的，具有不同的质量，这就导致了数据分布的差异，从而进一步增加了细粒度跨媒体检索的挑战性。对于图像和视频来说，它们具有不同的分辨率、颜色、视角以及光照等。对于文本来说，它们的文本长度不同，有不一样数目的单词。对于音频来说，它们有不一样的时长以及背景声音。音频样例的时长可以从 1 秒到 2,000 秒不等。而且一些音频样例还包含了人声、风声等其他背景声音。

6.2.3　细粒度跨媒体检索任务

为了充分验证本章构建的细粒度跨媒体检索数据集的有效性，我们设定了如下两种细粒度跨媒体检索任务，即双模态细粒度跨媒体检索（Bi-modality Fine-grained Cross-media Retrieval）和多模态细粒度跨媒体检索（Multi-modality Fine-grained Cross-media

Retrieval）。

6.2.3.1 双模态细粒度跨媒体检索

查询样例是任意一种媒体数据，检索结果是另外一种媒体数据，表示为“X → Y’’。例如，查询样例是一张灰背燕鸥（Slaty-backed Gull）的图像，则检索结果可以是灰背燕鸥的文本描述，这一检索过程表述为“I → T’’。在双模态细粒度跨媒体检索中，共有 12 种类似的检索任务，包括“I → T’’、“I → V’’、“I → A’’、“T → I’’、“T → V’’、“T → A’’、“V → I’’、“V → T’’、“V → A’’、“A → I’’、“A → T’’ 和“A → V’’。

6.2.3.2 多模态细粒度跨媒体检索

查询样例是任意一种媒体数据，检索结果是 4 种媒体数据，表示为“X → All’’。例如，查询样例是灰背燕鸥的图像，检索结果是灰背燕鸥的图像、文本、视频和音频样例，这一过程表示为“I → All’’。在多模态细粒度跨媒体检索中，共有 4 种类似的检索任务，包括“I → All’’、“T → All’’、“V → All’’ 和“A → All’’。

6.3 算法描述

为了验证本章新构建的细粒度跨媒体检索数据集的有效性，本章也提出了能够同时学习 4 种媒体统一表征的深度网络模型 FGCrossNet，以分类学习的方式进行统一表征的学习。我们从以下四个方面进行具体介绍：网络结构、数据处理、损失函数、训练和检索。

6.3.1 网络结构

现有的跨媒体检索方法通常对于不同的媒体数据采用不同的网络支路来处理，这就导致了以下问题：1）网络复杂度：不同的媒体数据通过不同的网络来处理。例如，图像一般采用卷积神经网络来处理，如 ResNet [108]；而文本一般采用 LSTM [109] 来处理。因此，最终用于处理 4 种或 5 种媒体的网络模型会具有非常高的复杂度。2）训练复杂度：由于网络结构复杂，因此它的训练也是非常困难的，从而导致复现过程困难。为了简化网络复杂度和训练复杂度，本章提出了一个通用深度网络模型，对于 4 种不同的媒体数据采用相同的网络结构。其网络结构如图6.4所示。我们采用了 ResNet50 [108] 作为基础网络。为了取得更好的结果，我们做了以下几个改动：采用 448×448 的输入大小，在最后一层卷积层之前接一个平均池化层，其核大小为 14，步长为 1。需要注意的是，在本章 FGCrossNet 中可以采用现有先进的网络结构作为基础模型，如 AlexNet [94] 和 VGGNet [43]。

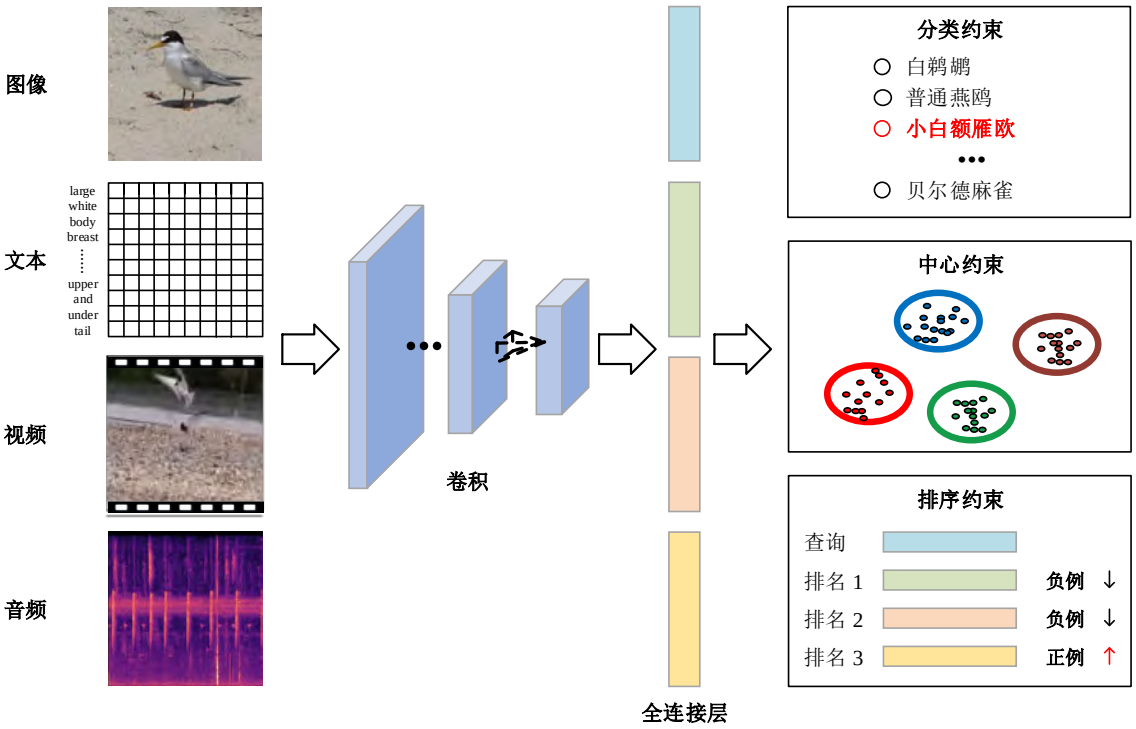


图 6.4　本章 FGCrossNet 网络结构示意图

6.3.2　数据处理

为了能够接受不同媒体类型的数据作为 FGCrossNet 的输入，我们需要首先对数据进行预处理。对于图像，不需要进行额外处理。对于视频，每一个视频样例我们等间隔地抽取 25 帧作为视频数据。对于音频，我们采用短时傅里叶变换（Short-Time Fourier Transformation）[110] 对每一个音频样例生成对应的频谱图。经过上述数据处理过程，本章的 FGCrossNet 可以直接处理音频数据。我们利用 librosa①来为每个音频数据生成频谱图，其大小设置为 448×448，这与音频的长度无关。频谱图的样例如图6.3所示。

对于文本数据，为了满足 FGCrossNet 的输入要求，本章提出了一种文本处理方法，如图6.5所示。给定一个文本样例，我们首先通过独热编码的方式量化每一个字符，将其编码为大小为 $n \times d$ 的向量 [111]，每个字符编码大小为 16。此外，对于每一个文本样例，我们设定最大字符数为 448，所以上述向量大小为 448×16。如果一个文本样例的字符数小于 448，我们把其余行用 0 来填充；如果大于 448，多余的字符直接截取不要。在本章新构建的细粒度跨媒体检索数据集中的文本数据的字符长度均小于 448，因此不会造成信息损失。然后，我们采用两个分别包含 224 和 448 个卷积核、核大小为 3、填充（Padding）为 1、步长为 1 的一维卷积层来处理上述文本向量，以此得到 448×448 的输出向量。最后，再采用一个卷积核数目为 3、核大小为 3、填充为 1、步长为 1 的

① librosa.github.io/librosa/core.html

二维卷积层来处理，得到 $448 \times 448 \times 3$ 的输出向量，以此作为 FGCrossNet 的输入。进一步，我们采用了位移差（Position Shift）[112] 来扩充文本数据已获得更好的特征学习。

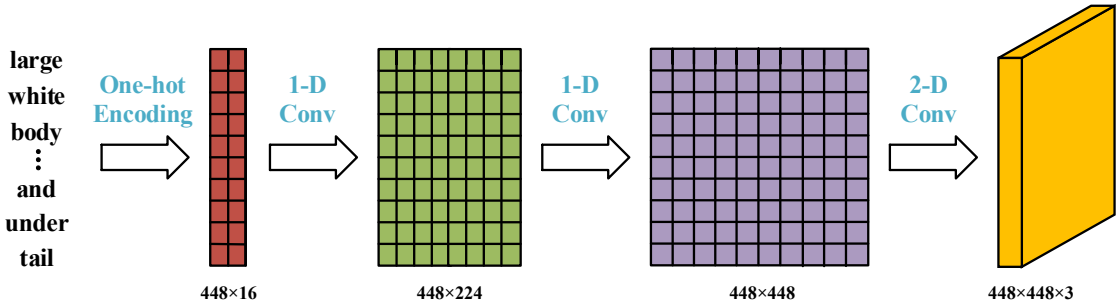


图 6.5　文本处理示意图

6.3.3　损失函数

我们为 FGCrossNet 设计了一个新的损失函数，其联合考虑了三种约束信息以学习具有辨识性的统一表征：分类约束，确保细粒度子类别统一表征的辨识性；中心约束，确保相同细粒度子类别统一表征的紧凑性；排序约束，确保不同细粒度子类别统一表征的松散性。该损失函数定义如下：

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{cen} + \mathcal{L}_{rank} \tag{6.1}$$

下面我们分别介绍上述公式中的三项。

6.3.3.1　分类约束

我们采用交叉熵损失函数来作为分类损失，以使得 FGCrossNet 具有能够区分相似细粒度子类别的能力。例如，将灰背燕鸥（Slaty-backed Gull）与银鸥（Herring Gull）区分开，如图6.3所示。分类约束 $\mathcal{L}_{cls}$ 定义如下：

$$\begin{aligned}\mathcal{L}_{cls} = &\frac{1}{N_I}\sum_{k=1}^{N_I} l(x_k^I, y_k^I) + \frac{1}{N_T}\sum_{k=1}^{N_T} l(x_k^T, y_k^T) + \\ &\frac{1}{N_V}\sum_{k=1}^{N_V} l(x_k^V, y_k^V) + \frac{1}{N_A}\sum_{k=1}^{N_A} l(x_k^A, y_k^A)\end{aligned} \tag{6.2}$$

其中，$l(x_k, y_k)$ 是交叉损失函数，I、T、V 和 A 分别表示图像、文本、视频和音频。以图像为例，N_I 表示训练集中图像的数目，y_k^I 表示第 k 个图像样本的资力度子类别标签，x_k^I 表示第 k 个图像样本的特征向量，在本章实验中为 FGCrossNet 网络模型最后一层全连接层的输出，如图6.4所示。需要注意的是，本章对每个视频样例等间隔抽取 25 帧，因此 N_V 表示的是训练集中视频帧的数目。

6.3.3.2　中心约束

在公式（6.1）中，第二项 $\mathcal{L}_{cen}$ 表示中心约束，其定义如下：

$$\mathcal{L}_{cen} = \frac{1}{2}\sum_{k=1}^{N} ||x_k - c_{y_k}||_2^2 \tag{6.3}$$

为了取得更好的细粒度跨媒体检索效果，相同细粒度子类别的样本（包括图像、文本、视频和音频）在统一空间中应具有相近的特征。因此，本章通过中心约束来减少类内特征的距离以缩短模态之间的差异。受到聚类算法的启发，我们通过最小化特征与其类别中心的距离来优化 FGCrossNet 的学习。在公式（6.3）中，x_k 表示训练集中第 k 个样本的特征，在这里不区分媒体类型，因为目的是使得相同细粒度子类别的所有媒体数据的特征相近。c_{y_k} 表示 y_k 细粒度子类别的质心的特征，其通过计算每个 batch 中所有属于 y_k 的媒体数据的特征进行更新优化，N 表示训练集中样本的数目。

6.3.3.3　排序约束

在公式（6.1）中，第三项 $\mathcal{L}_{rank}$ 表示排序约束，其定义如下：

$$\mathcal{L}_{rank} = \sum_{i,j,k}^{N}[d(x_i,x_j)^2 - d(x_i,x_k)^2 + \alpha_1]_+ + \sum_{i,j,k,l}^{N}[d(x_i,x_j)^2 - d(x_l,x_k)^2 + \alpha_2]_+ \tag{6.4}$$

由于中心约束是为了最小化类内差异，排序约束的目的是最大化类间差异。我们采用四元组损失函数[113] 来驱动 FGCrossNet 使得不同细粒度子类别的样本特征在统一空间中的距离尽可能大。在公式（6.4）中，x 表示训练数据。需要注意的是，x_i、x_j、x_k 和 x_l 分别表示 4 种媒体的输入样本。对于输入样本有两个约束：1）它们分属于 4 种媒体类型，即分属于图像、文本、视频和音频。2）它们属于 3 个细粒度子类别，即其中其中两个样本属于相同细粒度子类别，另外两个样本属于其他两个细粒度子类别。例如，x_i、x_j、x_k 和 x_l 分别表示图像、文本、视频和音频样本，其中，x_i 和 x_j 属于灰背燕鸥（Slaty-backed Gull），而 x_k 属于加州海鸥（California Gull），x_l 属于银鸥（Herring Gull）。这样这 4 个输入样本构成了四元组。它们之间的相似度用 L2 距离来度量，表示为 $d()$，$[z]_+$ 表示 $max(z,0)$。α_1 和 α_2 表示边界阈值用于平衡公式（6.4）中的两项。在本章实验中，采取与[113] 一样的设置，将 α_1 和 α_2 分别设置为 1 和 0.5。

6.3.4　训练和检索

在训练过程中，与以往网络仅输入一个样本不同，我们一次性输入 4 个样本。这 4 个样本分别属于图像、文本、视频和音频，且其中两个样本属于相同细粒度子类别，另外两个样本属于其他两个细粒度子类别。需要注意的是，这里的设置与媒体类型无关，属于相同细粒度子类别的两个样本可以属于任意两种媒体类型，另外两个样本则

对应属于剩余的两种媒体类型，整个过程是随机设定的。由于 FGCrossNet 的输入是图像（图像、视频和音频均以图像的形式输入）或者类图像矩阵（文本），因此我们首先利用图像数据来对 FGCrossNet 进行微调。然后，将 4 种媒体的数据作为输入，根据损失函数 $\mathcal{L}$ 来对 FGCrossNet 进行优化。在训练过程中，三种约束是依此添加的，初始学习率设置为 0.001，每 3 个 eopch 以 0.5 比率下降。

当进行检索时，我们提取 FGCrossNet 网络最后一层全连接层的输出作为 4 种媒体的统一表征。然后，采用余弦距离来进行相似性度量。最终根据相似度返回检索结果。

6.4 实验结果与分析

为了验证本章构建的细粒度跨媒体检索数据集 PKU FG-XMedia 以及提出的 FGCrossNet 的有效性，我们在新数据集上与现有方法进行细粒度跨媒体检索实验对比。

6.4.1 数据划分和评价指标

对于图像和视频，我们与原始数据集的划分一致。具体地，对于图像，训练集包括 5,994 张图像，测试集包括 5,794 图像；对于视频，训练集包括 12,666 个视频，测试集包括 5,684 个视频。对于文本，训练集和测试集分别包括 4,000 个文本。对于音频，训练集和测试集分别包括 6,000 个音频。

与 [101] 一样，我们选取 MAP（Mean Average Precision）作为细粒度跨媒体检索的评价指标。其反应检索结果的准确率和排序情况，值越高越好。首先对每个查询样例，计算 AP（Average Precision）值，然后计算所有查询样例的平均 AP 值作为 MAP 值。

6.4.2 对比方法

我们与现有方法进行了对比，包括 MHTN [114]、ACMR [115]、JRL [116]、GSPH [117]、CMDN [118]SCAN [119]GXN [120]。

- MHTN [114] 通过迁移学习将知识从单媒体数据（图像）迁移到跨媒体数据，能够学习 5 种媒体的统一表征。
- ACMR [115] 通过对抗式学习来学习统一表征。
- JRL [116] 引入半监督规约和系数规约来学习统一表征。
- GSPH [117] 提出了泛化哈希的方法来保持两种媒体数据之间的语义距离。
- CMDN [118] 首先通过多个深度网络来学习每种媒体数据的特征表示，然后再通过一个堆叠的网络来学习统一表征。
- SCAN [119] 考虑了图像区域和文本单词之间的隐含对齐关系，以此学习图像-文本对的相似度。

• GXN [120] 将生成过程引入到特征编码过程中，以此学习统一表征。

由于 SCAN 和 GXN 方法是针对图像和文本之间的跨媒体检索特殊设计的，因此很难将其扩展到 4 种媒体的跨媒体检索。所以，在本章中我们只与其图像和文本之间的跨媒体检索结果进行了对比。而对于其他方法，我们对比了 4 种媒体之间的跨媒体检索。

6.4.3　与现有方法进行对比

在本章实验中，我们通过两种跨媒体检索任务来验证本章构建的细粒度跨媒体检索数据集和提出的 FGCrossNet 的有效性，即双模态细粒度跨媒体检索和多模态细粒度跨媒体检索。结果如表6.3-表6.4所示。

为了公平对比，对于对比方法，我们采用相同的特征作为输入。对于图像和视频，如果输入不是原始的图像或视频（帧），我们提取 ResNet50 网络最后一层的全连接层的 200 维特征作为输入。需要注意的是，这里采用的 ResNet50 网络已经在图像数据上进行了微调。对于文本，我们利用 1,000 维的 BoW 特征作为输入。对于音频，我们利用 128 维的 MFCC 特征作为输入。

6.4.3.1　双模态细粒度跨媒体检索对比

表6.3展示了双模态细粒度跨媒体检索的对比结果。我们可以看到，本章 FGCrossNet 取得了最好的细粒度跨媒体检索结果。在所有对比结果中，MHTN 方法取得了最好的检索准确率。这是由于其从额外的单媒体数据中进行迁移学习的结果。但是，本章 FGCrossNet 与 MHTN 方法相比，在所有 12 种双模态细粒度跨媒体检索任务上均取得了更好的检索准确率。这主要是因为：1）本章 FGCrossNet 同样采用了迁移学习的机制，从图像迁移到文本、视频和音频。与 MHTN 不同的是，这里的图像、文本、视频和音频均来自本章构建的细粒度跨媒体检索数据集。2）本章提出了一个统一的网络模型，能够同时学习 4 种媒体数据的统一表征，有效关联 4 种媒体数据并在一定程度上缩短了异构鸿沟。3）本章 FGCrossNet 联合考虑了分类约束、中心约束和排序约束，有效缩减了类内差异和增大了类间差异。

SCAN 方法采用 Faster R-CNN [91] 来挖掘图像中多个对象所对应的的多个区域，其不适用于本章所构建的数据集。这是因为新构建的数据集中的图像大多只包含一个对象。GXN 利用生成模型来学习统一表征，其同样不适用于本章所构建的数据集。这是因为图像并没有其对应的文本描述。在本章数据集中文本数据更聚焦于描述对应的细粒度子类别，而不是图像。因此，上述两个方法在本章构建的数据集上的表现不好。

表 6.3　双模态细粒度跨媒体检索结果（MAP）

方法	本章 FGCrossNet 方法	MHTN [114]	ACMR [115]	JRL [116]	GSPH [117]	CMDN [118]	SCAN [119]	GXN [120]
I→T	**0.210**	0.116	0.162	0.160	0.140	0.099	0.050	0.023
I→A	**0.526**	0.195	0.119	0.085	0.098	0.009	-	-
I→V	**0.606**	0.281	0.477	0.435	0.413	0.377	-	-
T→I	**0.255**	0.124	0.075	0.190	0.179	0.123	0.050	0.035
T→A	**0.181**	0.138	0.015	0.028	0.024	0.007	-	-
T→V	**0.208**	0.185	0.081	0.095	0.109	0.078	-	-
A→I	**0.553**	0.196	0.128	0.115	0.129	0.017	-	-
A→T	**0.159**	0.127	0.028	0.035	0.024	0.008	-	-
A→V	**0.443**	0.290	0.068	0.065	0.073	0.010	-	-
V→I	**0.629**	0.306	0.536	0.517	0.512	0.446	-	-
V→T	**0.195**	0.186	0.138	0.126	0.126	0.081	-	-
V→A	**0.437**	0.306	0.111	0.068	0.086	0.009	-	-
平均	**0.366**	0.204	0.162	0.160	0.159	0.105	0.050	0.029

6.4.3.2　多模态细粒度跨媒体检索对比

表6.4展示了多模态细粒度跨媒体检索的对比结果。其趋势与双模态细粒度跨媒体检索一致，本章 FGCrossNet 同样取得了最好的细粒度跨媒体检索准确率。需要注意的是，本章 FGCrossNet 在 4 种媒体数据的统一表征学习上有一个特有的优势，即它是一个统一且简单的深度模型，能够同时为图像、文本、视频和音频生成具有辨识性的统一表征。在对比方法中，只有 MHTN 方法可以同时学习 4 种媒体数据的特征，但是，其网络模型是复杂的，对于每一种媒体数据都设计了一种网络。而对于其他对比方法，它们一次只能学习 2 种媒体数据的统一表征，这就增加了训练和检索的复杂度。对于这些方法，我们只能通过两两之间的学习来进行跨媒体检索。以 “I→All’’ 任务为例，我们首先进行双模态细粒度跨媒体检索，即 “I→T’’、“I→V’’ 和 “I→A’’。然后，再将上述结果与 “I→I’’ 的结果融合作为 “I→All’’ 的最终结果。

表 6.4　多模态细粒度跨媒体检索结果（MAP）

方法	I→All	T→All	V→All	A→All	平均
本章 FGCrossNet 方法	**0.549**	**0.196**	**0.416**	**0.485**	**0.412**
MHTN [114]	0.208	0.142	0.237	0.341	0.232
GSPH [117]	0.387	0.103	0.075	0.312	0.219
JRL [116]	0.344	0.080	0.069	0.275	0.192
CMDN [118]	0.321	0.071	0.016	0.229	0.159
ACMR [115]	0.245	0.039	0.041	0.279	0.151

6.4.4　基线实验

为了验证每一种约束在本章 FGCrossNet 中的有效性，我们进行了基线实验，结果如表6.5所示。我们可以发现：1）即使只采用分类约束，本章 FGCrossNet 方法也能比所有的对比方法取得更好的细粒度跨媒体检索准确率。这表明分类约束能够帮助

FGCrossNet 学习细粒度的辨识性特征，以此对相似细粒度子类别进行区分。2）“+ 中心约束’’ 表示在分类约束的基础上加入中心约束。它能够比只采用分类约束取得 0.043 的提升。这是因为中心约束能够强制相同细粒度子类别数据的特征趋近于其细粒度子类别中心。3）“+ 排序约束’’ 表示采用所有的 3 种约束。除了 “I → A’’ 和 “A → I’’ 两个任务，都取得了最好的检索结果。排序约束聚焦于学习不同细粒度子类别数据特征之间的辨识性，以此促进细粒度跨媒体检索的准确率。

表 6.5　三种约束的有效性

方法	分类约束	+ 中心约束	+ 排序约束
I→T	0.132	0.195	**0.210**
I→A	0.485	**0.540**	0.526
I→V	0.579	0.596	**0.606**
T→I	0.181	0.240	**0.355**
T→A	0.126	0.176	**0.181**
T→V	0.146	0.193	**0.208**
A→I	0.514	**0.562**	0.553
A→T	0.100	0.150	**0.159**
A→V	0.410	0.439	**0.443**
V→I	0.597	0.616	**0.629**
V→T	0.126	0.174	**0.195**
V→A	0.396	0.432	**0.437**
平均	0.316	0.359	**0.366**

6.5　本章小结

本章有两个方面的贡献：1）构建了一个新的细粒度跨媒体检索公开数据集和评测基准 PKU FG-XMedia。这是最大的包含媒体类型最多的细粒度跨媒体检索数据集。它能有助于细粒度跨媒体检索的研究。2) 提出了能够同时学习 4 种媒体统一表征的深度网络模型 FGCrossNet，联合考虑了分类约束、中心约束和排序约束，有效地学习跨媒体数据的辨识性统一表征。本章从细粒度跨媒体检索上展开辨识性特征学习，实现了图像向跨媒体的扩展，分类向检索的扩展。

第七章　总结与展望

7.1　工作总结

细粒度分析旨在对粗粒度的大类（如鸟）进行细粒度的子类划分（如里海燕鸥、北极燕鸥等），其广泛应用于智能农业、智能医疗等智能产业，具有重要的研究和应用价值。其关键在于获取细粒度子类别的辨识性信息并进行有效表达。本文针对辨识性特征学习，从减少标注成本、减少人工先验、提高辨识速度、提高语义关联四个方面展开研究，并将其应用于细粒度图像分类和细粒度跨媒体检索等应用。主要工作总结如下：

- 现有方法通常依赖于成本巨大的图像级、对象级、部件级标注信息，为了减少标注成本，提出了基于对象-部件注意力模型的细粒度图像分类方法。在对象级注意力上，提出注意力选择和显著性提取，自动定位对象区域，学习更精细的对象特征。在部件级注意力上，提出空间关联约束和部件语义对齐，实现辨识性部件的有效定位，排除了姿态、视角等差异的干扰。两者结合能够学习到多粒度的辨识性特征，准确率超过了使用对象、部件人工标注的强监督方法。
- 现有方法通常依赖于实验验证等人工先验的方式来进行辨识性特征学习，为了减少人工先验，提出了基于堆叠式深度强化学习的细粒度图像分类方法。首先，层次化地定位图像中的多粒度辨识性区域，并自适应地确定其数目。然后，通过多尺度区域的定位及辨识性特征学习，进一步提升细粒度图像分类准确率。学习过程由语义奖励函数驱动，能够有效捕捉图像中的辨识性、概念性的视觉信息，实现弱监督甚至无监督条件下的辨识性特征学习。
- 现有方法通常聚焦于准确率问题，却忽略了实际应用中的速度问题，为了提高辨识速度，提出了基于弱监督快速辨识定位的细粒度图像分类方法。首先，提出多级注意力引导的辨识性定位，通过显著图生成伪监督信息，实现了弱监督条件下的辨识性定位。进一步显著图驱动二次定位学习，增强了定位的准确性。然后，提出多路端到端辨识性定位网络，实现多个辨识性区域的同时定位，从而提高了辨识速度。多个辨识性区域之间互不促进，提升细粒度图像分类准确率。
- 现有研究主要聚焦于图像数据，忽略了与其相关联的跨媒体数据，为了提高语义关联，引入了文本、视频、音频等跨媒体数据，提出了基于细粒度分类的跨媒体检索方法。建立了首个包含 4 种媒体类型（图像、文本、视频和音频）的细粒度跨媒体检索公开数据集和评测基准 PKU FG-XMedia。提出了能够同时学习 4 种媒体统一表征的深度模型 FGCrossNet，确保统一表征的辨识性、类内紧凑性和类间松散性。实现图像向跨媒体的扩展，分类向检索的扩展。

7.2 未来展望

本文针对细粒度分析展开研究，重点研究了细粒度图像分类和细粒度跨媒体检索中的辨识性特征学习，取得了一些研究成果，但仍有很多问题亟待解决。在未来的研究中，将主要围绕以下三个方向展开工作：

- 大规模细粒度分析：在现有的细粒度图像分类与检索、细粒度视频分类等细粒度分析研究中所广泛使用的数据通常规模较小、细粒度子类别也相对较少。例如，广泛使用的 CUB-200-2011 鸟类数据集，仅有 11,788 张图像，200 个细粒度子类别。而大规模的细粒度分析任务将会使得现有方法的效果明显下降。因此，如何支持大规模细粒度分析，并有效提升分析效果，将是细粒度分析走向实际应用亟待解决的问题。
- 细粒度的视觉推理能力：在图像、视频等视觉领域的细粒度分析方法取得了一定进展，但仍缺乏视觉推理能力，不能很好地解释细粒度类别之间的本质差异。因此，如何借鉴人类的知识，将现有知识构建为知识图谱，将知识嵌入到深度学习中，学习从局部推演出整体，从属性推演出概念的推理能力，这将是细粒度分析在视觉领域的一个重要方向。
- 单媒体向跨媒体的扩展：现有研究主要聚焦于图像这一单媒体的细粒度分析，针对文本、视频、音频的研究还很少。因此，如何基于现有的细粒度图像分析工作，将图像中的知识迁移到文本、视频和音频等跨媒体数据，并对跨媒体数据进行细粒度的关联关系挖掘，充分利用跨媒体数据的大量有用信息，实现跨媒体数据的细粒度分类与检索，是一个极具挑战性的研究任务。

参考文献

[1] 彭宇新, 綦金玮, 黄鑫. 多媒体内容理解的研究现状与展望. 计算机研究与发展, 2019, 56(1): 183–208.

[2] 黄凯奇, 任伟强, 谭铁牛. 图像物体分类与检测算法综述. 计算机学报, 2014, 37(6): 1225–1240.

[3] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona and Serge Belongie. The Caltech-ucsd birds-200-2011 dataset. 2011.

[4] Ning Zhang, Ryan Farrell, Forrest Iandola and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In: International Conference of Computer Vision. 2013: 729–736.

[5] Ning Zhang, Jeff Donahue, Ross Girshick and Trevor Darrell. Part-based r-cnns for fine-grained category detection. European conference on computer vision (ECCV), 2014: 834–849.

[6] Timnit Gebru, Jonathan Krause, Jia Deng and Li Fei-Fei. Scalable annotation of fine-grained categories without experts. In: CHI Conference on Human Factors in Computing Systems. 2017: 1877–1881.

[7] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin and Qi Tian. Picking deep filter responses for fine-grained image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1134–1142.

[8] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015: 842–850.

[9] Lingxi Xie, Jingdong Wang, Bo Zhang and Qi Tian. Fine-grained image search. IEEE Transactions on Multimedia, 2015, 17(5): 636–647.

[10] Chen Zhu, Xiao Tan, Feng Zhou, Xiao Liu, Kaiyu Yue, Errui Ding and Yi Ma. Fine-grained video categorization with redundancy reduction attention. In: European Conference on Computer Vision. 2018: 136–152.

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester and Deva Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(9): 1627–1645.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142–158.

[13] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers and Arnold WM Smeulders. Selective search for object recognition. International Journal of Computer Vision (IJCV), 2013, 104(2): 154–171.

[14] Shaoli Huang, Zhe Xu, Dacheng Tao and Ya Zhang. Part-stacked cnn for fine-grained visual categorization. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1173–1182.

[15] Jonathan Krause, Hailin Jin, Jianchao Yang and Li Fei-Fei. Fine-grained recognition without part annotations. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5546–5555.

[16] Weifeng Ge, Xiangru Lin and Yizhou Yu. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In: IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3034–3043.

[17] Jianlong Fu, Heliang Zheng and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4438–4446.

[18] Tsung-Yu Lin, Aruni RoyChowdhury and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In: International Conference of Computer Vision. 2015: 1449–1457.

[19] Yang Gao, Oscar Beijbom, Ning Zhang and Trevor Darrell. Compact bilinear pooling. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 317–326.

[20] Yin Cui, Feng Zhou, Jiang Wang, Xiao Liu, Yuanqing Lin and Serge Belongie. Kernel pooling for convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2921–2930.

[21] Yaming Wang, Vlad I Morariu and Larry S Davis. Learning a discriminative filter bank within a CNN for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4148–4157.

[22] Feng Zhou and Yuanqing Lin. Fine-grained image classification by exploring bipartite-graph labels. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1124–1133.

[23] Xiaofan Zhang, Feng Zhou, Yuanqing Lin and Shaoting Zhang. Embedding label structures for fine-grained feature representation. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1114–1123.

[24] Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu and Xiaonan Luo. Knowledge-embedded representation learning for fine-grained image recognition. In: International Joint Conference on Artificial Intelligence. 2018: 627–634.

[25] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu and Zhi-Hua Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE Transactions on Image Processing, 2017, 26(6): 2868–2881.

[26] Xiawu Zheng, Rongrong Ji, Xiaoshuai Sun, Baochang Zhang, Yongjian Wu and Feiyue Huang. Towards optimal fine grained retrieval via decorrelated centralized loss with normalize-scale layer. In: AAAI Conference on Artificial Intelligence. 2019: 9291–9298.

[27] Tomoaki Saito, Asako Kanezaki and Tatsuya Harada. IBC127: video dataset for fine-grained bird classification. In: IEEE International Conference on Multimedia and Expo. 2016: 1–6.

[28] Jonathan Krause, Michael Stark, Jia Deng and Li Fei-Fei. 3d object representations for fine-grained categorization. In: International IEEE Workshop on 3D Representation and Recognition. 2013: 554–561.

[29] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen and Shilei Wen. Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In: IEEE International Conference on Computer Vision. 2019: 6222–6231.

[30] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. arXiv preprint arXiv:2003.13042, 2020.

[31] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5994–6002.

[32] Huapeng Xu, Guilin Qi, Jingjing Li, Meng Wang, Kang Xu and Huan Gao. Fine-grained image classification by visual-semantic embedding. In: International Joint Conference on Artificial Intelligence. 2018: 1043–1049.

[33] Hua Zhang, Xiaochun Cao and Rui Wang. Audio visual attribute discovery for fine-grained object recognition. In: AAAI Conference on Artificial Intelligence. 2018: 7542–7549.

[34] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou and Qi Tian. Fused one-vs-all features with semantic alignments for fine-grained visual categorization. IEEE Transactions on Image Processing, 2016, 25(2): 878–892.

[35] Ross Girshick, Jeff Donahue, Trevor Darrell and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580–587.

[36] Steve Branson, Grant Van Horn, Serge Belongie and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. arxiv:1406.2952, 2014.

[37] Yu Zhang, Xiu-Shen Wei, Jianxin Wu, Jianfei Cai, Jiangbo Lu, Viet-Anh Nguyen and Minh N Do. Weakly supervised fine-grained categorization with part-based image representation. IEEE Transactions on Image Processing, 2016, 25(4): 1713–1725.

[38] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In: International Conference of Computer Vision. 2015: 1143–1151.

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248–255.

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva and Antonio Torralba. Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2921–2929.

[41] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman and CV Jawahar. Cats and dogs. In: IEEE Conference on Computer Vision and Pattern Recognition. 2012: 3498–3505.

[42] Maria Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image Processing. 2008: 722–729.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arxiv:1409.1556, 2014.

[44] Max Jaderberg, Karen Simonyan, Andrew Zisserman et al. Spatial transformer networks. In: Advances in Neural Information Processing Systems. 2015: 2017–2025.

[45] Dequan Wang, Zhiqiang Shen, Jie Shao, Wei Zhang, Xiangyang Xue and Zheng Zhang. Multiple granularity descriptors for fine-grained categorization. In: International Conference on Computer Vision. 2015: 2399–2406.

[46] Zhe Xu, Dacheng Tao, Shaoli Huang and Ya Zhang. Friend or foe: fine-grained categorization with weak supervision. IEEE Transactions on Image Processing, 2017, 26(1): 135–146.

[47] Lingxi Xie, Liang Zheng, Jingdong Wang, Alan L Yuille and Qi Tian. Interactive: Inter-layer activeness propagation. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 270–279.

[48] Luming Zhang, Yang Yang, Meng Wang, Richang Hong, Liqiang Nie and Xuelong Li. Detecting densely distributed graph patterns for fine-grained image categorization. IEEE Transactions on Image Processing, 2016, 25(2): 553–565.

[49] Hantao Yao, Shiliang Zhang, Yongdong Zhang, Jintao Li and Qi Tian. Coarse-to-fine description for fine-grained visual categorization. IEEE Transactions on Image Processing, 2016, 25(10): 4858–4872.

[50] Yin Cui, Feng Zhou, Yuanqing Lin and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1153–1162.

[51] Zhe Xu, Shaoli Huang, Ya Zhang and Dacheng Tao. Webly-supervised fine-grained visual categorization via deep domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 40(5): 1100–1113.

[52] Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal and Dimitris Metaxas. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1143–1152.

[53] Di Lin, Xiaoyong Shen, Cewu Lu and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1666–1674.

[54] Thomas Berg and Peter Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. 2013: 955–962.

[55] Lingxi Xie, Qi Tian, Richang Hong, Shuicheng Yan and Bo Zhang. Hierarchical part matching for fine-grained visual categorization. In: International Conference of Computer Vision. 2013: 1641–1648.

[56] David G Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110.

[57] Hossein Azizpour and Ivan Laptev. Object detection using strongly-supervised deformable part models. In: European conference on computer vision. 2012: 836–849.

[58] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. IEEE Transactions on Multimedia, 2017, 19(6): 1245–1256.

[59] Saining Xie, Tianbao Yang, Xiaoyu Wang and Yuanqing Lin. Hyper-class augmented and regularized deep learning for fine-grained image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2645–2654.

[60] Jonathan Krause, Timnit Gebru, Jia Deng, Li-Jia Li and Li Fei-Fei. Learning features and parts for fine-grained recognition. In: International Conference on Pattern Recognition. 2014: 26–33.

[61] Yaming Wang, Jonghyun Choi, Vlad Morariu and Larry S Davis. Mining discriminative triplets of patches for fine-grained classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1163–1172.

[62] Yaming Wang, Vlad I Morariu and Larry S Davis. Weakly-supervised discriminative patch learning via CNN for fine-grained recognition. arxiv:1611.09932, 2016.

[63] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang and Yihong Gong. Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2010: 3360–3367.

[64] Lingxi Xie, Richang Hong, Bo Zhang and Qi Tian. Image classification and retrieval are one. In: ACM on International Conference on Multimedia Retrieval. 2015: 3–10.

[65] Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki and Stefan Carlsson. From generic to specific deep representations for visual recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 36–45.

[66] Qi Qian, Rong Jin, Shenghuo Zhu and Yuanqing Lin. Fine-grained visual categorization via multi-stage metric learning. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3716–3724.

[67] Ahmet Iscen, Giorgos Tolias, Philippe-Henri Gosselin and Hervé Jégou. A comparison of dense region detectors for image search and fine-grained classification. IEEE Transactions on Image Processing, 2015, 24(8): 2369–2381.

[68] Naila Murray and Florent Perronnin. Generalized max pooling. In: IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2473–2480.

[69] Liefeng Bo, Xiaofeng Ren and Dieter Fox. Multipath sparse coding using hierarchical matching pursuit. In: IEEE Conference on Computer Vision and Pattern Recognition. 2013: 660–667.

[70] Anelia Angelova and Shenghuo Zhu. Efficient object detection and segmentation for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2013: 811–818.

[71] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. 2015: 448–456.

[72] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich. Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1–9.

[73] Ken Chatfield, Karen Simonyan, Andrea Vedaldi and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. arxiv:1405.3531, 2014.

[74] Chao Huang, Hongliang Li, Yurui Xie, Qingbo Wu and Bing Luo. PBC: Polygon-based classifier for fine-grained categorization. IEEE Transactions on Multimedia, 2017, 19(4): 673–684.

[75] Lingxi Xie, Jingdong Wang, Weiyao Lin, Bo Zhang and Qi Tian. Towards reversal-invariant image representation. International Journal of Computer Vision, 2017, 123(2): 226–250.

[76] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan and Stefan Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014: 806–813.

[77] Mark B Neider and Gregory J Zelinsky. Searching for camouflaged targets: Effects of target-background similarity on visual search. Vision Research, 2006, 46(14): 2217–2235.

[78] Benjamin W Tatler, Roland J Baddeley and Benjamin T Vincent. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. Vision Research, 2006, 46(12): 1857–1862.

[79] Laurent Itti and Christof Koch. Computational modelling of visual attention. Nature Reviews Neuroscience, 2001, 2(3): 194–203.

[80] Derrick Parkhurst, Klinton Law and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. Vision Research, 2002, 42(1): 107–123.

[81] Xiaopeng Zhang, Hongkai Xiong, Wengang Zhou, Weiyao Lin and Qi Tian. Picking neural activations for fine-grained recognition. IEEE Transactions on Multimedia, 2017, 19(12): 2736–2750.

[82] Ross Girshick. Fast R-CNN. In: International Conference of Computer Vision. 2015: 1440–1448.

[83] Zequn Jie, Xiaodan Liang, Jiashi Feng, Xiaojie Jin, Wen Lu and Shuicheng Yan. Tree-structured reinforcement learning for sequential object localization. In: Neural Information Processing Systems. 2016: 127–135.

[84] Nobuyuki Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62–66.

[85] Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In: International Conference of Computer Vision. 2015: 2488–2496.

[86] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529.

[87] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 2015, 111(1): 98–136.

[88] Sijia Cai, Wangmeng Zuo and Lei Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017: 511–520.

[89] Xiangteng He and Yuxin Peng. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification. In: AAAI Conference on Artificial Intelligence. 2017: 4075–4081.

[90] Shu Kong and Charless Fowlkes. Low-rank bilinear pooling for fine-grained classification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7025–7034.

[91] Shaoqing Ren, Kaiming He, Ross Girshick and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. 2015: 91–99.

[92] John R Anderson. Cognitive psychology and its implications. WH Freeman/Times Books/Henry Holt & Co, 1990.

[93] Yan Karklin and Michael S Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. Nature, 2009, 457(7225): 83–86.

[94] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. 2012: 1097–1105.

[95] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia. 2014: 675–678.

[96] Benjamin J Meyer, Ben Harwood and Tom Drummond. Nearest neighbour radial basis function solvers for deep neural networks. arXiv preprint arXiv:1705.09780, 2017.

[97] Xiao Liu, Jiang Wang, Shilei Wen, Errui Ding and Yuanqing Lin. Localizing by describing: attribute-guided attention localization for fine-grained recognition. In: AAAI Conference on Artificial Intelligence. 2017: 4190–4196.

[98] Qichang Hu, Huibing Wang, Teng Li and Chunhua Shen. Deep CNNs with spatially weighted pooling for fine-grained car recognition. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(11): 3147–3156.

[99] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. 2014: 818–833.

[100] Limin Wang, Yu Qiao and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In: IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4305–4314.

[101] Yuxin Peng, Xin Huang and Yunzhen Zhao. An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges. IEEE Transactions on circuits and systems for video technology, 2018, 28(9): 2372–2385.

[102] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In: ACM International Conference on Multimedia. 2010: 251–260.

[103] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results.

[104] Cyrus Rashtchian, Peter Young, Micah Hodosh and Julia Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In: NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010: 139–147.

[105] Peter Young, Alice Lai, Micah Hodosh and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2014, 2: 67–78.

[106] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C Lawrence Zitnick. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. 2014: 740–755.

[107] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo and Yantao Zheng. NUS-WIDE: a real-world web image database from National University of Singapore. In: ACM International Conference on Image and Video Retrieval. 2009: 48.

[108] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770–778.

[109] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780.

[110] Karlheinz Gröchenig. Foundations of time-frequency analysis. Springer Science & Business Media, 2013.

[111] Alexis Conneau, Holger Schwenk, Loïc Barrault and Yann Lecun. Very deep convolutional networks for text classification. In: European Chapter of the Association for Computational Linguistics. 2016: 1107–1116.

[112] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang and Yi-Dong Shen. Dual-path convolutional image-text embedding with instance loss. arXiv preprint arXiv:1711.05535, 2017.

[113] Weihua Chen, Xiaotang Chen, Jianguo Zhang and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017: 403–412.

[114] Xin Huang, Yuxin Peng and Mingkuan Yuan. Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval. IEEE Transactions on Cybernetics, 2020, 50(3): 1047–1059.

[115] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic and Heng Tao Shen. Adversarial cross-modal retrieval. In: ACM International Conference on Multimedia. 2017: 154–162.

[116] Xiaohua Zhai, Yuxin Peng and Jianguo Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(6): 965–978.

[117] Devraj Mandal, Kunal N Chaudhury and Soma Biswas. Generalized semantic preserving hashing for n-label cross-modal retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4076–4084.

[118] Yuxin Peng, Xin Huang and Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks. In: International Joint Conference on Artificial Intelligence. 2016: 3846–3853.

[119] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu and Xiaodong He. Stacked cross attention for image-text matching. In: European Conference on Computer Vision. 2018: 201–216.

[120] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In: IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7181–7189.

个人简历、攻读博士学位期间的研究成果

个人简历

1991 年 5 月 5 日出生于山东省泰安市。

2010 年 9 月考入南开大学计算机与控制工程学院计算机与信息安全系，2014 年 7 月本科毕业并获得工学学士学位。

2014 年 9 月免试进入北京大学王选计算机研究所，攻读理学博士学位至今。

攻读博士学位期间的研究成果

发表（接收）的学术论文

1. **何相腾**，彭宇新，‘‘基于跨域和跨模态适应学习的无监督细粒度视频分类’’，**软件学报**，2020. (已接收，CCF A 类中文期刊)
2. **Xiangteng He**, Yuxin Peng and Junjie Zhao, “Which and How Many Regions to Gaze: Focus Discriminative Regions for Fine-grained Visual Categorization”, International Journal of Computer Vision (**IJCV**), Vol. 127, No. 9, pp. 1235-1255, Sep. 2019. (CCF A 类国际期刊，影响因子 6.071)
3. **Xiangteng He**, Yuxin Peng and Junjie Zhao, “Fast Fine-grained Image Classification via Weakly Supervised Discriminative Localization”, IEEE Transactions on Circuits and Systems for Video Technology (**TCSVT**),Vol. 29, No. 5, pp. 1394-1407, May. 2019. (CCF B 类国际期刊，影响因子 4.046)
4. **Xiangteng He** and Yuxin Peng, “Fine-grained Visual-textual Representation Learning”, IEEE Transactions on Circuits and Systems for Video Technology (**TCSVT**), Vol. 30, No. 2, pp. 520-531, Feb. 2020. (CCF B 类国际期刊，影响因子 4.046)
5. **Xiangteng He** and Yuxin Peng, “Fine-grained Image Classification via Combining Vision and Language”, 30th IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), pp. 5994-6002, Honolulu, Hawaii, USA, Jul. 21-26, 2017. (CCF A 类国际会议)
6. **Xiangteng He** and Yuxin Peng, “Weakly Supervised Learning of Part Selection Model with Spatial Constraints for Fine-grained Image Classification”, 31th AAAI Conference on Artificial Intelligence (**AAAI**), pp. 4075-4081, San Francisco, California, USA, Feb. 4–9, 2017. (CCF A 类国际会议)

7. **Xiangteng He**, Yuxin Peng and Junjie Zhao, "Fine-grained Discriminative Localization via Saliency-guided Faster R-CNN", 25th ACM International Conference on Multimedia (**ACM MM**), pp. 627-635, Mountain View, CA, USA, Oct. 23-27, 2017. (CCF A 类国际会议)
8. **Xiangteng He**, Yuxin Peng and Junjie Zhao, "StackDRL: Stacked Deep Reinforcement Learning for Fine-grained Visual Categorization", 27th International Joint Conference on Artificial Intelligence (**IJCAI**), pp. 741-747, Stockholm, Sweden, Jul. 13-19, 2018. (CCF A 类国际会议)
9. **Xiangteng He** and Yuxin Peng, "Only Learn One Sample: Fine-Grained Visual Categorization with One Sample Training", 26th ACM International Conference on Multimedia (**ACM MM**), pp. 1372-1380, Seoul, Korea, Oct. 22-26, 2018. (CCF A 类国际会议)
10. **Xiangteng He**, Yuxin Peng and Liu Xie, "A New Benchmark and Approach for Fine-grained Cross-media Retrieval", 27th ACM International Conference on Multimedia (**ACM MM**), pp. 1740-1748, Nice, France, Oct. 21-25, 2019.(CCF A 类国际会议)
11. **Xiangteng He** and Yuxin Peng, "Multi-attention Guided Activation Propagation in CNNs", 1st Chinese Conference on Pattern Recognition and Computer Vision (**PRCV**), pp. 16-27, Guangzhou, China, Nov. 23-26, 2018.
12. Yuxin Peng, **Xiangteng He**, and Junjie Zhao, "Object-Part Attention Model for Fine-grained Image Classification", IEEE Transactions on Image Processing (**TIP**), Vol. 27, No. 3, pp. 1487-1500, Mar. 2018. (CCF A 类国际期刊，影响因子 6.79)
13. Junjie Zhao, Yuxin Peng and **Xiangteng He**, "Attribute Hierarchy based Multi-task Learning for Fine-grained Image Classification", **Neurocomputing**, Vol. 395, pp. 150-159, Jun. 2020.(CCF C 类国际期刊，影响因子 4.072)

申请发明专利

- 彭宇新，**何相腾**，一种基于选择与生成的数据增广方法及图像分类方法，申请号：201811183994.X，申请日：2018 年 10 月 11 日。

国际评测

- 2014 年-2016 年，连续 3 年作为团队成员参加由美国国家标准技术局（NIST）举办的国际评测比赛 TRECVID，在视频语义搜索比赛中获第一名，参赛队伍包括 IBM Watson 研究中心、AT&T 实验室、阿姆斯特丹大学、日本国立情报学研究所等国内外大学和研究机构。

参与项目

- 国家自然科学基金面上项目，视觉注意力驱动的图像视频分类与检索研究，2018年 1 月-2021 年 12 月。
- 国家自然科学基金重点项目，混杂数据的模式识别及敏感内容挖掘理论与方法，2016 年 1 月-2020 年 12 月。

奖励荣誉

- 2019.10　华为奖学金
- 2018.12　百度奖学金（每年全球 8-10 名获奖学生）
- 2018.11　北京大学国家奖学金
- 2018.06　北京大学博士研究生校长奖学金
- 2018.06　北京大学信息科学技术学院 ‘‘学术十杰”
- 2017.10　华为奖学金

致谢

回首攻读博士学位的六年时光，‘‘幸运’’可能是我的关键词。一路走来，虽然也遇到过坎坷，但总体上相对顺利。这都源于学习、工作、生活中导师、父母、同学等对我的大力支持与帮助。

感谢我的导师彭宇新教授，在研究道路上为我授业解惑、排忧解难。在确立研究方向之初，我还对研究一无所知，是彭老师仔细地为我介绍各个方向的优势以及当前的问题，并且发挥我的主观能动性，让我自主选择。正因为如此，在读博期间我一直对我的研究方向保持着浓厚的兴趣，一直保有高昂的热情。在研究过程中，正是彭老师的严格要求和耐心指导，让我逐渐懂得了什么是研究，如何做研究，怎么做好研究。除了学术研究以外，我觉得从彭老师的言传身教中受益最多的是：要认真、细致、负责。这是我今后人生道路上的宝贵财富，不仅对于学术研究，对于以后的工作、生活都会让我受益匪浅。

感谢我的父母和女朋友，他们是我读博道路上的坚实后盾，给予了我巨大的支持，照顾我的生活、疏导我的低落情绪、鼓励我勇敢前行。

感谢研究室的所有同学，特别是在我的研究之初帮助我的肖天骏师兄，在我刚开始接触项目工作时帮助我的张健师兄，以及与我共同合作的赵俊杰、谢柳师弟。也感谢实验室其他帮助我的老师、师兄、同学和师弟们：朱超老师，翟晓华、谢文轩、彭云波、黄雷、唐攀攀、王桂存、赵仕荣、李秋宇师兄，赵韫禛、黄鑫同学，张俊超、袁玉鑫、綦金伟、迟敬泽、袁明宽、叶钊达、孙宏博、赵祥宇、任远志师弟等。

感谢我的室友韩硕，感谢在读博期间对我生活上的帮助，对我的支持与鼓励。

最后，感谢北京大学、王选计算机研究所的各位老师、职工，为我们提供了良好的学习、生活环境，为我们提供了坚实的后勤保障，让我们潜心做好学术研究。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：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。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：何翔腾　日期：2020 年 5 月 27 日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：何翔腾　导师签名：

日期：2020 年 5 月 27 日