



中国科学院大学

University of Chinese Academy of Sciences

## 硕士学位论文

基于分歧与协同的弱监督视觉目标定位

作者姓名: 薛昊岚

指导教师: 叶齐祥 教授

中国科学院大学

学位类别: 工学硕士

学科专业: 信号与信息处理

培养单位: 中国科学院大学电子电气与通信工程学院

2020年6月



**Divergent and collaborative Learning for Weakly**  
**Supervised Object Localization**

**A thesis submitted to**  
**University of Chinese Academy of Sciences**  
**in partial fulfillment of the requirement**  
**for the degree of**  
**Master of Science in Engineering**  
**in Signal and Information Processing**

**By**

**Xue Haolan**

**Supervisor Professor Ye Qixiang**

**School of Electronic, Electrical and Communication Engineering**

**University of Chinese Academy of Sciences**

**June 2020**





**中国科学院大学直属院系**  
**研究生学位论文原创性声明**

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。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日 期：2020年5月20日

**中国科学院大学直属院系**  
**学位论文授权使用声明**

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

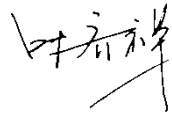
涉密的学位论文在解密后适用本声明。

作者签名：

日 期：2020年5月20日

导师签名：

日 期：2020年5月20日





## 摘要

目标识别及定位是视觉领域的基础技术，是实现很多视觉应用的先决条件。随着技术的飞速发展和数据的爆发式增长，海量数据的标注给目标识别定位带来了新的问题。在此背景下，如何在弱标注下提升目标检测与定位的性能变得尤为重要。本课题针对弱监督目标定位任务开展研究，标注量的减少使其可以扩展到大规模数据集上进行训练，得到泛化性能更好的模型，目前已经应用到医学图像等标注困难的实用场景中。除此之外，其基础研究也与神经网络可视化、可解释神经网络密切相关。在当前的弱监督定位框架下，分类网络提供了物体的特异性特征，而准确定位的关键问题仍在于如何捕获物体的本质特征，本文也以此为切入点展开研究，所完成的主要工作包括：

(1) 探究了解决弱监督目标定位问题的有效思路。提出了基于协同学习的定位器改良方案，构建了基于类别协同和训练过程协同的两种弱监督定位器，通过挖掘更多辅助信息丰富了弱监督定位的目标表达；提出了基于分歧思想的定位器改良方案，构建了基于分歧的弱监督定位器，通过扩张网络特征空间丰富了弱监督定位的目标表达。

(2) 提出了基于分歧与协同思想的弱监督定位器 (divergence and collaboration, DAC)，并将其部署到 VGGNet、GoogLeNet、ResNet 三种主流网络结构上，在两种主流数据集上进行实验，达到了目前领先的弱监督目标定位性能，验证了分歧与协同思想的互补性与有效性。

(3) 将基于分歧与协同的弱监督定位器应用到实际场景中，在医学数据集 ChestX-ray8 中验证了分歧与协同思想在病理学定位上的有效性。

本文探究了弱监督目标定位问题并提出了基于分歧与协同思想的解决方案，达到了目前领先的弱监督目标定位性能，对于本领域的研究具有积极推动作用。同时，该方法在医学图像中的应用验证了其应用价值，对于相关问题的研究具有借鉴意义。

**关键词：**弱监督目标定位，分歧与协同，视觉模式可视化



## Abstract

As a basic problem in the field of computer vision, object recognition and localization are the foundation of many applications. With the rapid development of technology and the explosive growth of data, massive data has brought new problems to data annotation in object recognition and localization. The thesis is based on the weakly supervised object localization(WSOL) task, which only uses image-level labels to learn the object localization model. The reduction of annotation and convenience of acquisition make it scalable to large-scale datasets for training, resulting in better performance. The model has been applied to medical images, security inspection images and other practical scenes with difficulty in annotations. In addition, for that weakly supervised object localization obtains the corresponding position of the object from the response graph, its basic research is also closely related to the neural network visualization and interpretable neural network. Under the current WSOL framework, the classification network provides the specific characteristics of the object, but the key problem of accurate localization is still how to capture the essential characteristics of the object, we will use this insight as an entry point to conduct research. The primary works are as follows:

1. We explore the bottleneck and solution of WSOL task. We propose to improve WSOL task with collaborative thought, and construct two types of weakly supervised locators based on category pyramid collaboration and training process collaboration. This enriches the vision expression of WSOL by mining more auxiliary information. Thus, we propose to improve WSOL task with divergent thought and construct a weakly supervised locator based on divergence. This can help the locator to mine more vision expressions from the inherent WSOL structure, and to reduce the information loss in the training process by expanding the feature space.

2. We propose a new WSOL method based on divergence and collaboration (DAC) thought, and deploye it on three mainstream network structures, VGGnet, GoogLeNet, and ResNet. Experiments are conducted on two mainstream datasets and achieve the

current optimal WSOL performance, verifying the complementarity and effectiveness of divergent and collaborative ideas.

3. We apply the proposed DAC method to actual scenarios. the effectiveness of the divergence and collaboration thought in pathological location was verified in the medical data set ChestX-ray8.

This thesis explores the bottleneck of the problem of WSOL task and proposes solutions based on divergence and collaborative thought, which achieves the current best performance of WSOL and plays an active role in promoting research in this field. In addition, the application in medical images verifies the practical application value of the method, which has reference significance for the study of related issues.

**Key Words:** Weakly Supervised Object Localization, Divergence and Collaboration , Vision Pattern Visualization

## 目 录

第 1 章 绪论 .....	1
1.1 研究背景与意义 .....	1
1.2 国内外研究现状 .....	2
1.3 本文的研究内容 .....	6
1.4 本文的组织结构 .....	7
第 2 章 相关工作概述 .....	9
2.1 弱监督目标发现 .....	9
2.1.1 弱监督目标定位 .....	9
2.1.2 弱监督目标检测 .....	12
2.1.3 其他相关工作 .....	13
2.2 神经网络可视化与可解释性 .....	14
2.3 分歧与协同弱监督目标发现 .....	15
2.4 本章小结 .....	16
第 3 章 基于协同的弱监督目标定位 .....	17
3.1 基于类别协同的弱监督目标定位 .....	17
3.1.1 构建类别金字塔 .....	17
3.1.2 网络结构 .....	19
3.1.3 实验结果及分析 .....	21
3.2 基于训练过程协同的弱监督目标定位 .....	23
3.2.1 各阶段协同弱监督定位的有效性 .....	24
3.2.2 基于训练过程协同的弱监督目标定位器 .....	26
3.2.3 实验结果及分析 .....	28
3.3 本章小结 .....	32

第 4 章 基于分歧的弱监督目标定位 .....	33
4.1 研究动机及建模过程 .....	33
4.2 研究结构框架 .....	34
4.3 实验结果及分析 .....	36
4.3.1 消融实验 .....	37
4.3.2 可视化结果 .....	40
4.4 本章小结 .....	42
第 5 章 基于分歧与协同的弱监督目标定位 .....	43
5.1 基于分歧与协同的弱监督目标定位 .....	43
5.1.1 基于分歧与类别协同的弱监督目标定位 .....	44
5.1.2 基于分歧与训练过程协同的弱监督目标定位 .....	45
5.1.3 比较分析 .....	46
5.2 实验结果与分析 .....	47
5.2.1 性能对比 .....	47
5.2.2 统计及可视化分析 .....	49
5.3 病理学定位应用 .....	50
5.4 本章小结 .....	52
第 6 章 总结与展望 .....	53
参考文献 .....	55
致 谢 .....	61
作者简历及攻读学位期间发表的学术论文与研究成果 .....	63



## 图目录

图 1.1 弱监督目标定位的意义 .....	1
图 1.2 弱监督目标定位框架 .....	4
图 1.3 新冠肺炎病毒 CT 诊断影像 .....	6
图 2.1 神经网络可视化用于图片问答 .....	14
图 3.1 鸟类的类别金字塔 .....	18
图 3.2 ILSVRC 的类别金字塔 .....	19
图 3.3 基于类别协同的弱监督定位器网络结构 .....	20
图 3.4 类别协同对 CUB-200-2011 数据集上定位结果的影响 .....	22
图 3.5 训练过程中的激活区域变化 .....	23
图 3.6 训练过程定位结果聚合网络 .....	24
图 3.7 基于训练过程协同的弱监督目标定位 .....	26
图 4.1 基于分歧的弱监督定位模块 .....	35
图 4.2 SPG 网络结构框图 .....	36
图 4.3 K 对于基于分歧的弱监督定位器的影响 .....	37
图 4.4 分歧方法对训练过程的影响 .....	41
图 4.5 分歧定位结果可视化 .....	41
图 5.1 基于分歧与类别协同的弱监督目标定位 .....	44
图 5.2 基于分歧与训练过程协同的弱监督定位 .....	45
图 5.3 基于分歧与协同的弱监督定位方法 IoU 统计结果 .....	49
图 5.4 基于分歧与协同的弱监督定位方法在 CUB 和 ILSVRC 上的定位结果可 视化 .....	50
图 5.5 Chest X-ray8[10]中包含的八种常见胸部疾病影像 .....	50
图 5.6 病理学定位结果图 .....	52



## 表目录

表 3.1 关于类别协同层级数的消融实验 .....	22
表 3.2 聚合方式对基于训练过程协同弱监督定位的影响 .....	25
表 3.3 网络结构对基于训练过程协同弱监督定位的影响 .....	27
表 3.4 正负样本平衡对基于训练过程协同弱监督定位的影响 .....	29
表 3.5 松弛阈值对基于训练过程协同弱监督定位的影响 .....	31
表 4.1 聚合方式对基于分歧的弱监督定位方法的影响 .....	38
表 4.2 分歧学习与集成学习的差别 .....	40
表 5.1 两种分歧与协同定位器的比较分析 .....	46
表 5.2 基于分歧与协同的弱监督定位方法与主流方法在 CUB 上的对比 ..	47
表 5.3 基于分歧与协同的弱监督定位方法与现有方法在 ILSVRC 上的对比 .....	48
表 5.4 分类结果 ROC 曲线的 AUC 值在 ChestX-ray8 数据集上的对比 ...	51
表 5.5 病理学定位准确性在 ChestX-ray8 数据集上的对比 .....	51



## 第1章 绪论

本章节共包含四部分内容，首先描述课题研究背景与意义，然后介绍国内外研究现状，最后概括了本文的研究内容和论文的组织结构。

### 1.1 研究背景与意义

近年来，计算机视觉技术正加速服务于我们的社会，在各个方面给我们的生活带来了很大程度上的便利，然而随着技术的发展与数据的爆发式增长，海量数据给数据标注带来了新的问题。在此背景下，本课题针对弱监督目标定位任务开展研究。弱监督目标定位（Weakly Supervised Object Localization, WSOL）仅使用图像中出现的物体类别标号来学习目标定位模型，标注量的减少及标注信息的易获取使其可以扩展到大规模数据集上进行训练，甚至利用网上的海量数据来辅助训练，得到泛化性能更好的模型。图 1.1 显示了弱监督目标定位的适用场景，图 1.1 左显示了需要密集标注的场景，该场景下标注费时且容易引入标注噪音，图 1.1 右显示了医学图像的标注场景，这类图像需要有专业知识的医生来标注，成本较高。

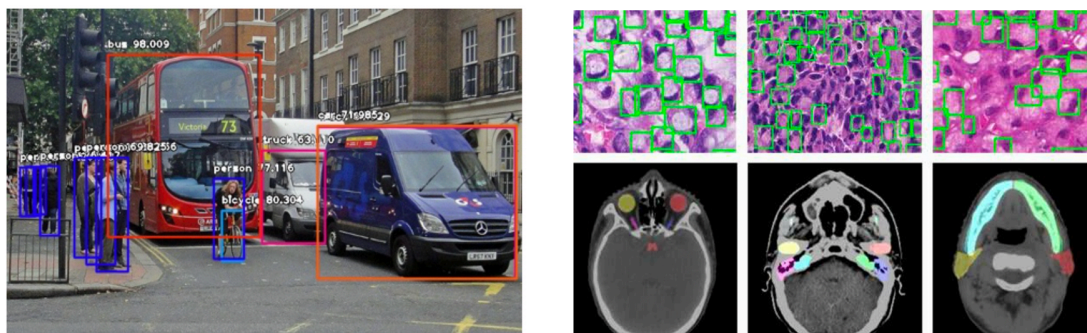


图 1.1 弱监督目标定位的意义

Figure 1.1 The significance of weakly supervised object localization

目标识别及定位作为视觉领域的基础技术，是实现很多视觉应用的先决条件。传统的目标定位技术需要给定图片中出现的物体类别及其位置坐标进行模型训练，这就给数据集的搜集和标注带来了一定的困难；同时，人工标注的数据可能

会存在一些噪音，这些噪音会对训练带来一定的干扰导致模型的泛化性能变差。弱监督学习作为无监督学习和全监督学习的折中，利用不完整的标注信息来学习物体的位置信息。其中，弱监督目标定位仅使用图像中出现的物体类别标号来学习目标定位模型，降低了识别定位网络对数据标注的要求，是该领域的一个重要分支。除此之外，由于弱监督目标定位从响应图中获得目标的相应位置，其基础研究也与神经网络可视化、可解释神经网络密切相关。

弱监督目标定位可应用在医学图像、安检图像等标注困难的实用场景中。在医学图像的病理学诊断中，仅需要标注该图像的患者是否患病，通过弱监督的定位方法就可以确定其病灶位置，从而更好地辅助治疗。在安检图像的查验中，安检人员常常因为重复的单调工作导致漏检、误检，弱监督定位可以辅助确定违禁物品的位置，再由工作人员确认，不仅减轻了人力资源的消耗，还提高了检测的准确性。在以上应用场景中，一方面弱标注下存在海量图片可以被用来训练模型，数据量的扩展可以给模型带来更好的泛化性能；另一方面，弱标注解放了大量的时间和人力，推动了目标识别与定位的发展与应用。本研究将以深度网络模型为基础，探究解决弱监督定位问题的有效思路。

## 1.2 国内外研究现状

目标识别及定位是视觉领域的基础技术，近年来受到了广泛关注，深度神经网络的发展更是进一步提高了目标检测识别的性能，目前的研究热点也从算法研究转到了算法的应用转化。如前文所述，海量数据的标注问题成为了限制该领域继续发展了关键问题，本课题针对弱监督目标定位任务开展研究，从分歧与协同的角度分析解决该问题的有效思路。本章节将首先介绍目标检测识别与弱监督检测定位的国内外研究现状，继而引出弱监督目标定位的应用场景与发展前景。

作为诸多计算机视觉应用的关键一环，目标检测经历了从传统图像识别方法到基于深度神经网络的检测方法的发展历程。目标检测识别主要分为两个步骤：图像特征提取与分类器的搭建，而传统方法与深度学习方法的主要区别在于是否依赖于手工设计特征。

传统目标检测识别方法利用 SIFT、HOG 等手工设计特征与特定分类器来实现图像分类。其中 SIFT 特征是一种局部特征检测的算法，该算法将图像中的特征点进行特征点匹配并计算出其位置信息，结合尺度、方向信息进行特征描述，

适用于在海量特征数据库中进行快速检测；而 HOG 用于检测物体的特征描述，通过计算和统计图像局部区域的梯度方向直方图来构建特征，其他手工设计特征还包括 Harr 特征、Surf 特征等；支持向量机（SVM）作为分类器的典型代表，通过在特征空间构建分类面并使支持向量距离更远来进行分类。手工设计特征鲁棒性较低，并无法与分类器联合训练，阻碍了目标识别定位性能的进一步提升。

随着神经网络的发展，研究人员利用神经网络得到了更多更具判别性的深度特征来表示目标，并构造了神经网络分类器来完成端到端的检测器训练。深度特征与端对端的训练为目标检测识别带来了丰富的特征与鲁棒的检测性能，自此目标检测识别得到了飞速发展。目标检测器主要分为单阶段检测器和双阶段检测器，其中单阶段检测器直接从深度特征图中获得候选框在各类别的置信度与位置偏移量，双阶段检测器首先区分前景与背景，再对前景候选框进行分类与位置的回归预测。单阶段检测器在目标检测速度上有较多优势，而双阶段检测器可以得到更精准检测结果。

目标检测识别性能很大程度上依赖于数据量，更大的数据量往往意味着模型更高的鲁棒性。以此为依据，各研究机构纷纷涉足海量数据下目标检测识别模型的开发。然而对于大量的数据，其标注往往比较费时，且不同的标注人员对于标注标准的理解不同也会给模型的训练带来一定的难度。基于该问题，弱标注下的目标检测与定位逐步发展为一个新兴的热门方向，即弱监督目标检测与弱监督目标定位。

弱监督学习指的是利用不完整的标注信息来完成学习任务，弱监督目标检测与弱监督目标定位均利用物体类别标签来学习目标的检测识别。两者共同来源于 2016 年周博磊在[1]中的发现：在分类网络的训练过程中，网络倾向于关注目标最具判别力的区域来识别物体，网络中的卷积滤波器可以作为检测器来激活该最具判别力的区域来达到分类的目的。其中，弱监督目标检测通过预设候选框，通过各候选框的激活结果来判断其是否包含目标；而弱监督目标定位仅根据全图激活结果来框选目标位置。两者的差别在于：在应用场景上，弱监督目标检测可以检测同类别的多个物体并分别给出检测结果，而弱监督目标定位仅根据全图激活结果框选出某类别的一个检测框，只能针对单目标图片进行检测定位；在效率上，弱监督目标检测由于使用了大量的候选框，具有更低的检测效率和更大的计算资

源消耗；在科学意义上，弱监督目标检测关注于如何在弱标签的情况下检测回归出更好的检测结果，实质上是对于全监督目标检测在弱标签限制下的泛化与探究，弱监督目标定位由于流程的简化，更多的与神经网络可视化、可解释神经网络密切相关。

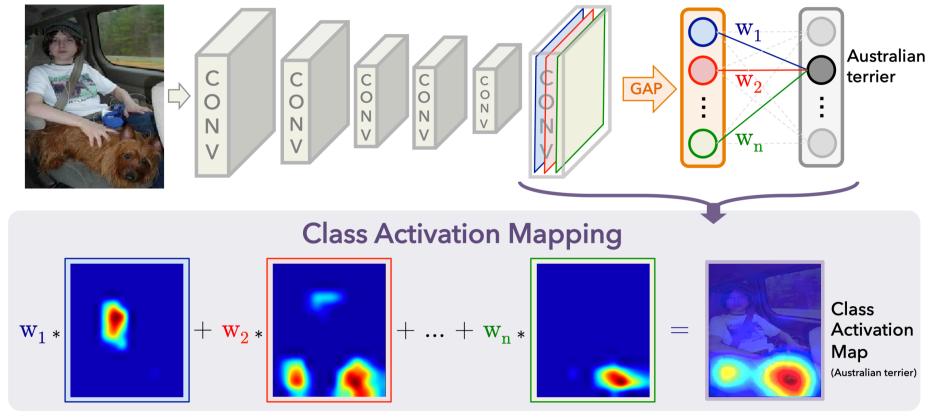


图 1.2 弱监督目标定位框架

Figure 1.2 Framework of weakly supervised object localization

本研究旨在探究限制弱监督目标定位效果的关键因素并给出有效思路。弱监督目标定位任务仅根据图像类别标签判断物体的位置，其依据在于分类网络训练的过程中，目标类别区别于其他类别的关键特征会被激活，体现在特征图上就是该关键位置被激活，我们称之为最具判别力的区域，然而由于分类网络仅关注物体最具判别力的区域，导致类激活图只能激活物体该区域，从而影响了弱监督定位性能。这种现象是由于神经网络训练过程中倾向于寻找最紧致的特征，利用类别监督来优化图像分类的过程中，网络学习的唯一目标是获得输入图像与类别标签最相关的视觉模式[2]，利用类别标签训练网络，隐含地判断了输入图片所有视觉模式与其的相关性，导致在训练过程中无关的视觉模式被抑制，最具判别力的模式作为图像分类的最优特征表示被保留优化分类。后续的一些弱监督目标定位工作为改进这一现象而展开。其中包括图传播[3]、数据增广[4]、空洞卷积[5]和对抗擦除方法[6][7]等。[4]随机遮挡训练集图片上的区域来学习不同的判别力区域，迫使网络不再仅仅关注物体的小部分判别区域，[6]、[7]通过训练多分支的网络来逐步激活物体上的不同区域，他们分别在原图和特征图上遮挡第一个分支的激活区域并用第二个分支训练分类网络，从而得到更多的物体定位区域,[3]用



随机游走的过程模拟目标区域之间的语义关联,迭代地改进弱监督目标定位激活区域,使网络逐步激活更多具有判别力的区域。[8]、[9]通过改进弱监督激活图与分类置信度的映射关系来增益弱监督目标定位,在有效抑制背景的前提下得到更鲁棒的目标定位结果。

弱监督目标定位问题发源于图像数据爆炸增长背景下的标注困难问题。以此为依据,其应用前景主要集中在数据量大、标注成本高的场景中。例如图片标注成本高的医学图像诊断领域,首先在各医院的数据库中已存有海量医学图像数据,电子病例中也带有关于病灶、病情的描述,由医生给出与该疾病相关的关键字就可以生成大量疾病类别标签。对比全监督目标检测算法,其不需要由医生对每张医学图像标注具体病灶位置,然而在医学影像中病灶位置往往不具有清晰的界限,且不同的医生可能有不同的理解,这就会导致标注效率的下降与标注结果的不一致,会给后续模型训练带来一定的困难。在医学图像领域,现已发布了数据集 ChestX-ray8 数据集[10],其中包含数万名患者的胸部 X 光图像与 8 种常见的胸部病例分类标注。实验表明,弱监督定位可以较好的帮助确定病灶的位置,目前的方法已经可以做到 32%的定位准确性[11]。在 2020 年的新型冠状病毒肺炎的疫情中,病人量大,医生人手不足导致了疫情早期的患者积压,并一定程度导致了疫情的进一步扩散。对于 CT 胸片,一位病人的 CT 影像大概在 300 张左右,平均一个病例医生靠肉眼分析需要 5 至 15 分钟,如果可以用计算机视觉技术进行预判将大大提高诊断效率,就可以减少误诊漏诊的情况发生。同时新型冠状病毒肺炎的诊断主要依据病灶区的占比判断患病的轻重程度,该需求与弱监督目标定位的适应场景具有一致性,弱监督目标定位可以在激活图上给出病灶的位置进而计算出其占比。新型冠状病毒肺炎的诊疗方案中指出肺部影像学显示 24 至 48 小时内病灶明显进展>50%者按重型管理,利用计算机视觉技术计算病灶区在肺实质中的体积占比,可以辅助医生快速识别轻型中的进展较快的患者,提前进行干预,改善患者的预后。2020 年 2 月,达摩院基于 5000 多个病例的 CT 影像样本研发出的医疗 AI,分析只需要 20 秒就能快速完成,直接地帮助一线医生提升诊断效率。

另一种是以安检图像、交通图像为代表的图像量大、标注费时的典型场景。以安检图像举例说明,随着运输行业客流和物流量的增加,安检人员的工作强度

也在增大，针对这种情况，弱监督目标定位技术可以辅助实现自动化/半自动化的安检流程并提高安检效率。安检图像数据集 SIXray[12]中包含了百万幅安检图像和六种违禁物品的类别标签，仅利用类别标号训练网络就可以定位出违禁物品在图片中的位置。实验表明，使用现有方法可以达到 77.20% 的分类性能和 52.23% 的定位性能。

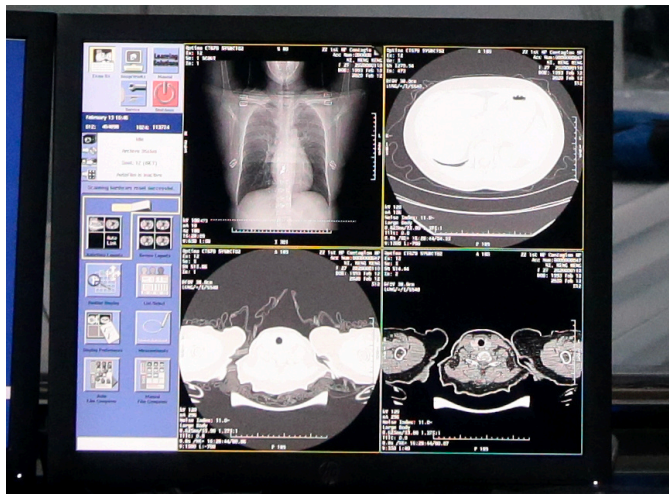


图 1.3 新冠肺炎病毒 CT 诊断影像

Figure 1.3 CT diagnostic images of 2019-nCov

### 1.3 本文的研究内容

本文主要探究了目前弱监督目标定位任务的瓶颈，并提出了基于分歧与协同的弱监督目标定位方法。弱监督目标定位作为目标识别及定位领域的一个分支，旨在利用不完整的目标标注信息——目标类别标号来学习物体的定位信息。当前弱监督定位框架[1]主要从分类网络中学习分类相关的视觉模式，由于分类网络仅仅关注物体最具判别力的区域，导致类激活图只能激活物体的该部分区域，从而影响了定位性能。在当前的弱监督定位框架下，分类网络提供了物体的特异性特征，而准确定位的关键问题仍在于如何捕获物体的本质特征，本文也将以此为切入点展开研究。我们的方法希望基于分歧和协同思想构建弱监督定位器，寻找更多视觉模式，达到最佳的弱监督定位效果。本文的主要贡献主要分为以下几个方面：

1、探究了协同思想对于弱监督定位的有效性。协同思想旨在挖掘出更多的可利用信息来辅助弱监督目标定位学习器的学习，我们探究了基于类别协同和训

练过程协同两种弱监督目标定位器。其中基于类别协同通过挖掘类别标号中的隐藏信息，寻找并保留更多与物体本身相关而不仅是类别相关的视觉模式，通过类别合并得到更丰富的物体表征特征，在定位激活图中激活该部分特征来补全之前被定位器忽略的目标区域；观察到网络训练过程中神经网络不停地捕获相关特征并抑制无关视觉模式来压缩深度特征，基于训练过程协同挖掘并保留了训练过程中被抑制的早期目标表示，通过将训练过程中的有效特征积累起来，更好地完成了弱监督目标定位。

2、探究了分歧思想对于弱监督定位的有效性。与协同思想通过多源信息的融合（类别层级信息、训练过程信息）来增益视觉模式不同，分歧思想构建多个分歧互补的定位器，改善了分类损失函数导致的神经网络特征紧凑性问题，同时减轻了视觉模式的压缩现象，通过挖掘充分的视觉模式，更好地完成了弱监督目标定位。

3、我们将改善学习器的分歧思想与融合多源信息的协同思想结合起来，构建了基于分歧与协同的弱监督定位器，结果表明两者具有互补性，可以共同作用优化弱监督目标定位。我们将该方法部署在三种网络结构上，验证了该方法对不同网络结构的兼容性；在主流数据集上进行实验，与前沿方法对比取得了领先的定位性能，验证了该方法的有效性；进一步的，我们在医学图像数据集 ChestX-ray8[10]中进行病理学定位，取得了高于基线方法的结果，验证了该方法在医学诊断上的应用价值。

#### 1.4 本文的组织结构

本论文的组织结构如下：

第一章，绪论。论述了弱监督目标定位的研究背景和研究意义，介绍了国内外该领域的研究现状并概述了本文的主要研究内容。

第二章，相关工作与技术。介绍了弱监督目标发现的现状和相关工作，其中包含弱监督目标定位、弱监督目标检测和其他相关技术；介绍了与弱监督目标定位基础研究相关的神经网络可视化与可解释性相关工作；最后介绍了基于分歧与协同思想的相关工作，该思想是本文研究工作的基础。

第三章，基于协同的弱监督目标定位器。介绍了基于类别协同和训练过程协同的两种弱监督目标定位器，通过挖掘出更多的可利用信息来辅助弱监督定位器

的学习。我们通过实验证明了两种定位器的有效性，其中类别协同旨在挖掘类别标号中的隐藏信息，训练过程协同旨在挖掘并保留训练过程中被抑制的早期目标表示。

第四章，基于分歧的弱监督目标定位器。介绍了分歧思想对于弱监督定位器的有效性，改善了分类损失函数导致的神经网络特征紧凑性问题，同时减轻了视觉模式的压缩现象。实验证明，分歧思想能较好地提升弱监督定位器的性能。

第五章，基于分歧与协同的弱监督定位器。该定位器既综合了协同思想中多源信息所带来的丰富视觉模式，又包含了分歧思想对神经网络特征紧凑性和视觉模式压缩现象的改善，从两个方面加强了弱监督网络的定位能力，使其仅从分类标签中就能得到良好的目标定位结果。我们在主流数据集上进行了实验，与当前主流弱监督定位算法相比，本方法取得了领先性能。进一步的，我们在医学数据集 ChestX-ray8[10]中进行病理学定位，取得了高于基线方法的结果，验证了该方法在医学诊断上的应用价值。

第六章，总结与展望。本章中我们总结了本文所提出的基于分歧与协同的弱监督定位器的理论和实验成果，并进一步分析了弱监督目标定位方向未来仍需面对的问题。

## 第2章 相关工作概述

本章主要分为三部分内容。首先介绍弱监督目标发现技术，包括弱监督目标定位、弱监督目标检测和其他相关技术；接下来介绍与弱监督目标定位基础研究相关的神经网络可视化与可解释性相关工作；最后介绍基于分歧与协同思想的相关工作，该思想是本文研究工作的基础。

### 2.1 弱监督目标发现

近年来，目标检测识别技术得到了飞速发展，其准确性与检测效率均都有了大幅度的提高。然而在算法应用的过程中，往往伴随着数据量大、标注困难的问题，这就将研究者的目光带到了不完全标注的目标检测算法研究中。全监督的检测识别技术需要标定出训练图片上每个目标的类别与位置信息，不但加大了标注的工作量也会不可避免地带来一定的噪声。本节将介绍弱监督目标发现的相关技术：弱监督目标定位、弱监督目标检测与其他相关的弱标注目标发现算法。

#### 2.1.1 弱监督目标定位

弱监督目标定位指的是：在训练时仅给出每张训练图片中包含的物体类别，算法根据图片类别学习该类物体的位置，在测试时给出预测的目标定位结果。本节中将介绍弱监督目标定位的相关工作并简单介绍该任务的主要数据集与评测方法。

2016年周博磊在[1]中发现：在分类网络的训练过程中，网络倾向于关注目标最具判别力的区域来识别物体，网络中的卷积滤波器可以作为检测器来激活该区域从而指导分类。最具判别力的区域指的是分类网络训练的过程中，该类别区别于其他类别的关键特征位置，然而由于分类网络仅关注物体最具判别力的区域，导致类激活图只能激活物体该区域，从而影响了弱监督定位性能。这种现象是由于神经网络训练过程中倾向于寻找最紧致的特征，利用分类监督来优化图像分类的过程中，网络学习的唯一目标是获得输入图像与类别标签最相关的视觉模式[2]，利用类别标签训练网络，隐含地判别了输入图片所有视觉模式与其的相关性，

网络抑制了无关的视觉模式来获得优化图像分类的最优特征表示。后续的一些弱监督目标定位工作为改进这一现象而展开。

其中一类代表性工作是 2017 年 Dahun Kim 等人提出的两阶段学习方法[6]和 2018 年张晓琳等人发表的对抗互补方法[7]。两者均针对弱监督目标定位部分激活现象进行网络结构改进,通过多阶段的激活来完整激活定位目标。其中两阶段学习方法构造了两阶段的定位器,通过“第一阶段定位-定位结果反馈-第二阶段定位-定位结果融合”的流程得到最终的定位结果;对抗互补方法构造了使用相同特征提取器的两个定位器,通过在网络学习过程中不断擦除定位器二输入特征中定位器一的已激活区域,使两定位器关注不同特征,从而更完整地定位。[4]与上述工作从相同的思路出发,不同的是其改进点在数据预处理而不是网络结构,该方法随机地遮挡训练集图片的部分区域来学习不同的判别力区域,迫使网络不再仅仅关注物体的一部分,本质上是一种数据增广方法。ADL [13]在以上方法的基础上减少了网络训练的存储、时间与算力开销,在单个模型中进行一次前向后向传播就完成了寻找并删除最具区别性区域的操作。这类工作从弱监督目标定位“部分激活”的现象出发,通过增加网络训练过程中对不同区域的关注来增加最具判别力区域的数量,达到更大范围激活目标的目的。

另一类代表性工作旨在发现最具判别力的视觉模式与其他视觉模式的相关关系,利用该关系改进网络特征提取模块,将部分激活区域传导至目标其他位置,并提高定位准确性。2017 年朱艺等人的 Soft Proposal[3]方法采用图传播思想,用随机游走的过程模拟目标区域之间的语义关联,该方法构造了两种距离函数来将已激活区域的置信度传导至目标的其他区域,迭代地改进了弱监督目标定位激活结果。2018 年魏云超等人利用空洞卷积[14]改进了弱监督定位网络特征提取模块。空洞卷积在不产生额外开销的前提下可以有效地扩大卷积核感受野的范围,该特性与作者将激活传播到其他相关部分的目的一致,通过扩大感受野,低响应区域可以通过感知周围的高响应区域来获得更好的辨别力。该团队同年发表了基于自监督思想的 SPG 工作[15],同样是旨在发现已激活区域与未激活区域的联系,将已激活区域作为“种子”,在网络训练过程中逐步发现其与其他区域的关联性,并学习发现更可靠的目标区域。值得注意的是,这一类工作跳出了之前工作中扩

大激活区域的单一目标，转向寻找目标区域与背景区域特征层面的差别，并用图传播/空洞卷积/自学习的思想逐步优化该特征，在提高特征判别力的同时有效抑制了背景，得到了更鲁棒的目标定位结果。

近两年来也有一些其他的工作从不同的角度改善弱监督定位方法。例如，Grad-CAM[16]使用了特征图的梯度信息来判断最具判别力的区域；Wildcat[8]改进了池化层并同时利用了前景和背景的信息，分类和定位效果均有了一定的提高；CCAM[17]发现多种类别定位激活图的组合可以突出前景目标并抑制背景区域，并基于此观察改进了弱监督目标定位结果的生成方式，有效改善了弱监督目标定位结果。

2020 年的工作 PSOL[18]受 selective search[19]和 Faster-RCNN[20]的启发，认为 WSOL 中的定位部分应该为类不可知的，与分类无关。基于这个观察，将 WSOL 问题分为类不可知目标定位以及目标分类两部分，直接通过伪框标注进行模型更新。该方法取得了目前弱监督目标定位问题的最优性能，实验表明该方法在不同数据集之间具有较好的迁移能力。值得注意的是，该方法与其他方法采用了不同的定位框架，希望通过框回归来增益定位，不完全属于弱监督定位的范畴。

在当前的弱监督定位框架下，分类网络提供了物体的特异性特征，之前的方法通过将该部分特异性特征扩大激活来增益定位，而准确定位的关键问题仍在于如何捕获物体的本质特征，本文也将以此为切入点展开研究。

弱监督目标定位分为点定位和框定位两种，目前学界主要专注于框定位算法的改进。点定位正确指的是：图片分类正确，且定位激活图上最高值的位置在目标上；框定位正确指的是：图片分类正确，且框定位结果与真实值的交并比 (Intersection over Union, IoU) 大于 0.5。该任务主要使用 CUB-200-2011[21]和 ImageNet-1000[22]两个数据集评测，评测指标有 Loc error 与 Corloc[23]两种。其中 Loc error 通常计算 Top-1 错误率和 Top-5 错误率，将网络输出结果与真实值比较，将分类正确且定位结果与物品的真实位置框  $\text{IoU}>0.5$  的结果视作正例，而错误率计算的是反例的占比；Corloc 计算正确率，将已知目标类别，定位结果与位置真值的  $\text{IoU}>0.5$  的样本视作正例。

## 2.1.2 弱监督目标检测

如前文所述,弱监督目标检测与弱监督目标定位的区别在于是否预设候选框,弱监督目标检测由于预设了候选框可以检测同一图像上的多个物体,也带来了更多效率和计算资源的开销。典型的弱监督目标检测框架由三个部分组成:候选框生成、弱监督目标定位和弱监督检测器学习。其中,候选框生成是保证目标检测查全率的前提,弱监督目标定位是弱监督学习的核心问题,而检测器学习最终保证了目标检测的效果。弱监督目标定位方法在前文中已经详述,下面将介绍常用的候选框生成方法和弱监督检测器学习方法。

弱监督目标检测中候选框生成模块继承自全监督目标检测[20][24][25],多为无监督方法。包括滑动窗口方法( Sliding window )和基于区域的候选框生成方法。滑动窗口方法以不同大小、长宽比、步长对整个图片扫窗,是一种穷举的思想;基于区域的方法包括 Selective Search[19]、Edge boxes[26]、MSER[27]等,在保证查全率的前提下大大提高了候选框生成效率。Selective Search 以 P. F. Felzenszwalb 在 2004 年发表的基于图的图像分割方法[28]为基础,同时考虑了图片的色彩、纹理和尺寸的相似度; Edge boxes 先使用边界检测算法得到一些初始边界,再用非极大值抑制(Non-Maximum Suppression, NMS)[29]方法进行边界合并,最后得到稳定的目标边界; MSER 方法基于分水岭的概念,递增地设定阈值并将图像进行二值化,在得到的所有二值图像中,某些连通区域变化很小甚至没有变化,则该区域就被称为最大稳定极值区域。弱监督目标检测中用的最多的是 Selective Search 方法。

弱监督检测器学习方法主要分为三种:基于多示例学习的检测器学习、基于聚类方法的检测器学习和基于隐变量支持向量机的检测器学习方法。其中近年来受到关注最多的是基于多示例学习的弱监督目标检测器。

多示例学习[30]由 Dietterich 等人在 1997 年提出,在学习过程中以多示例包为训练单元,每个多示例包含有若干个没有分类标签的示例。如果至少含有一个正示例,则该包被标记为正类多示例包;如果所有示例都是负示例,则该包被标记为负类多示例包。多示例学习的目的是:通过对具有分类标签的多示例包的学习,建立多示例分类器,并将该分类器应用于未知多示例包的预测。对于弱监督



目标检测任务,通过候选框生成和弱监督目标定位两个步骤得到一些可能含有目标的候选框,并将他们划分为多示例包,根据图片的类别标签可以进一步将其分为正类多示例包和负类多示例包,进而将弱监督目标检测任务建模为多示例学习过程。近年来,研究人员将卷积神经网络与多示例学习结合起来,构建了端到端训练的弱监督目标检测框架[31][32][33]。Bilen 和 Vedaldi 提出了一种弱监督深度检测网络[32],该网络使用了一种新颖的加权 MIL 池化策略,并结合了空间正则化以实现更好的性能。在[32]的基础上,Kantorov 等人[33]进一步提出了考虑上下文信息的弱监督目标检测框架,得到了更优的检测效果。

基于聚类的弱监督检测学习器使用无监督的聚类方法初始化候选框和相关类别等的潜在变量,将样本中聚类性突出的子集作为正例样本,迭代地优化该正例样本集合,达到弱监督检测器学习的目的。Wang C 等人[34]提出了基于潜在类别学习的弱监督检测器,使用典型的隐语义空间分析(pLSA)来聚类学习潜在类别,遗憾的是该方法仍需要调整聚类编号来获得各类别的有效聚类。与此相比,Hakan Bilen[35]等人提出的方法仅利用前景的类内方差聚类,不需要显式建模相关背景,此外,该方法可以自动确定最佳聚类数目,进一步简化了弱监督检测器的学习。

隐变量支持向量机(LSVM)作为另一种隐变量学习方法,将图像中目标位置作为隐变量,通过求解一个非凸目标函数,实现最大间距思想下的图像级分类与弱监督检测建模。Song H O 等人使用平滑的 LSVM 作为弱监督检测问题的初始化策略,并开发了负样本挖掘技术来提高算法对负例框的鲁棒性[36]。

### 2.1.3 其他相关工作

本节将简单介绍其他非完整监督的目标发现技术,包括少样本目标检测方法、显著性检测方法等。

少样本目标检测(Few Example Detection)使用少数有标注的样本和大量未标记样本学习目标的定位,其中每种类别只有极少数(2-4个)有类别与位置标注的样本,其关键在于尽可能多地生成可靠的样本标注。[37]迭代地产生高置信度样本标注并训练网络,在训练中首先生成简单的样本并对初始化不佳的模型进行改进,随着模型变得越来越具有判别力,将选择更有挑战性的样本并进行

另一轮模型改进。该方法用共同训练了多个检测模型，以挑选高置信度的伪标签样本，达到了与弱监督目标定位可比的性能。

显著性检测 (saliency detection) 任务旨在从无标注或弱标注的训练样本中学习如何区分前景与背景区域，即检测到图像中的“显著性区域”。其与弱监督目标发现的区别在于不需要判断目标的类别标号，只需要区分前景和背景。无监督的显著性检测方法[38][39][40]多是基于颜色、对比度、显著性先验等低级特征。但是由于缺乏空间相关性推断和图像语义对照检测，这些低级特征并不一定对所有图像都适用。2018年，受到弱监督目标定位任务的启发，中山大学提出了利用图像标号的显著性检测方法[41]。该方法迭代地检测并纠正传统无监督方法产生的噪声标签，将分类网络中的目标激活图与无监督方法生成的显著图结合起来作为像素级注释。该模型通过空间一致性来纠正目标内部标签的噪声，通过全卷积网络纠正跨图像的语义歧义，同时更新该粗略的激活图以进行下一次迭代，是一种简单有效的显著性检测算法。

## 2.2 神经网络可视化与可解释性

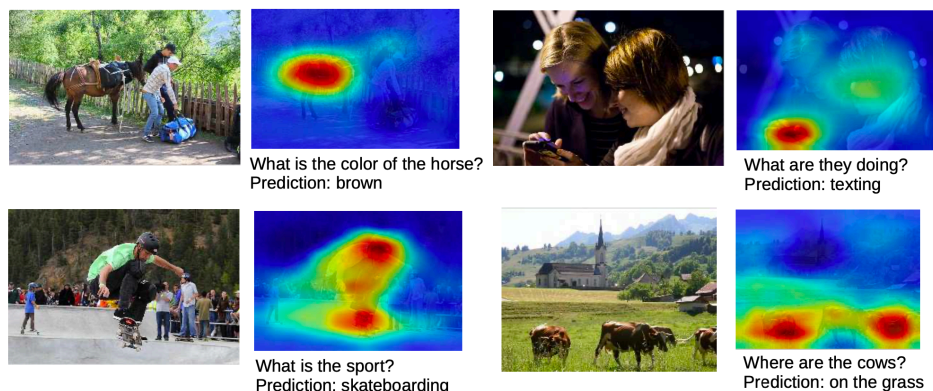


图 2.1 神经网络可视化用于图片问答

Figure 2.1 Interpreting visual question answering

随着深层神经网络在各种任务上超越人类，人们开始探索这些深度模型的学习模式，近年来与深度神经网络可解释性相关的工作越来越受到关注。弱监督目标定位通过学习分类网络中有判别力的区域来定位目标，而这部分有判别力的区域可以帮助生成显著图像特征，与神经网络的可解释性和特征可视化息息相关。

[1]中的实验表明弱监督目标定位有助于场景检测、概念学习、文字检测、图片问答等任务中的关键模式识别,观察识别到的重要区域可以帮助理解各视觉任务中视觉模式的发现规律。

人们研究神经网络可视化与神经网络可解释性,旨在可视化神经网络的深层视觉表示,并为模型的调优提供定性指导。一些工作[42][43]通过采样隐藏单元激活的图像块来寻找神经网络行为的线索;也有其他工作[44]利用神经网络的反向传播来生成显著图像特征。这些工作虽给出了神经网络学习的可视化结果,但是并没有进一步给出这些可视化结果的有效解释。

[45]构造了基础分解框架来解释神经网络学习过程中所关注的视觉模式,该框架将用于图像分类预测的证据分解为可解释的语义组件,并量化每个组件对最终预测的贡献。具体来说,该方法根据注意力热图的生成结果,学习了特征空间中的一组可解释向量,并用这些向量分解表示最终预测。实验证明该方法的视觉模式结果比其他方法解释性更强。

2020年,Lei Chen等人[46]基于对Grad-CAM[16]的观察,即反向传播中的梯度可以作为权重来解释网络决策,提出了一种用于嵌入网络的Grad-CAM方法。该方法从多个训练示例中汇总梯度,并采用了一种有效且无需反向传播的权重传递方法,解决了反向传播相关方法应用在大型网络上的痛点问题。与原始Grad-CAM方法相比,该方法可以获得更准确的视觉注意结果。

### 2.3 分歧与协同弱监督目标发现

分歧与协同思想的相关研究始于Blum和Mitchell的工作[47],该思想较少受到模型假设、损失函数非凸性和数据规模问题的影响,方法简单有效、理论基础相对坚实、适用范围较为广泛。其假设对于同一数据可以从不同的角度进行学习,并训练出不同的模型,由于这些模型是从不同角度训练出来的,其互补性可以提高模型精度,就如同从不同角度可以更好地理解事物一样。该思想的适用性很广,它既不需要大量冗余的视图,也不需要对所采用的学习算法施加任何约束,已成功应用于域自适应[48]、图像分类[49][50]、数据分割[51]、基于图像的搜索[52]等众多任务。后续研究指出,多学习器间的分歧对此类学习的成效至关

重要。本文提出的弱监督定位方法就是基于该思想，接下来将介绍一些基于分歧与协同思想的相关工作。

近年来，分歧与协同思想多应用于半监督学习[53][54][55]中，先使用多个分歧的分类器对无标签样本进行分类，再将置信度高的样本加入训练集中。分歧的分类器主要有两种实现方式：1) 使用不同的网络架构；2) 使用多样化的训练方法。2005年，周志华团队提出了一种使用三个分类器进行协同训练的方法[53]，训练过程中学习器选择要标记的示例后，由不同分类器产生多个伪标签，该方法在保证效率的同时可以得到泛化能力较好的模型。[54]扩展了该协同训练框架，利用神经网络来学习分歧模型。观察到多个模型在学习过程中很容易相互折叠，作者引入了对抗性示例生成方法来形成模型之间的分歧，实验表明，与原模型相比，基于分歧的方法可以加速模型收敛并显著提高模型准确性。2020年澳洲国立大学进一步优化了协同训练框架，对各个特征提取器进行差异约束，使得模型能够学习独特且具有判别力的特征[56]。文章指出协同训练中的多个学习器应具有相同的模型框架以防止学习能力差异导致的网络快速收敛，而仅使用不同初始化或不同训练样本并不能保证学习器的绝对差异，基于这些观察该团队提出了最大差异协同训练框架，通过鼓励网络在统计上的差异来实现学习器的多样性。

## 2.4 本章小结

本章介绍了弱监督目标定位的相关技术。首先介绍了弱监督目标发现的研究现状和相关工作，其中包含弱监督目标定位、弱监督目标检测和其他相关技术；其次介绍了与弱监督目标定位基础研究相关的神经网络可视化与可解释性相关工作；最后介绍了基于分歧与协同思想的相关工作，该思想是本文研究工作的基础。

## 第3章 基于协同的弱监督目标定位

弱监督目标定位以图片类别为唯一的标注信息，基于协同的弱监督目标定位器旨在挖掘出更多的可利用信息来辅助其学习。本章节中将介绍基于类别协同和训练过程协同的两种弱监督目标定位器。其中类别协同旨在挖掘类别标号中的隐藏信息，通过类别合并得到更丰富的物体表征特征；训练过程协同旨在挖掘并保留训练过程中被抑制的早期目标特征表示。

### 3.1 基于类别协同的弱监督目标定位

基于类别协同的弱监督目标定位器希望通过图片标号引入更多的监督，并通过合并相似类别来获得不同语义层级的特征信息。从弱监督定位器部分激活目标区域的现象出发，我们认为很可能是过于精细的目标类别限制了视觉模式的发现。在神经网络根据类别标号训练分类器和定位器的过程中，由于以降低分类交叉熵损失为唯一目标，网络倾向于保留对降低该分类损失、提高分类正确率最有用的视觉模式，而忽略了其他目标相关视觉模式的表达。该观察也符合 Tishby N 等人关于神经网络信息理论的研究[2]：神经网络训练过程中倾向于寻找最紧致的特征，利用分类监督来优化图像分类的过程中，网络学习的唯一目标是获得输入图像与类别标签最相关的视觉模式，利用类别标签训练网络，隐含地判別了输入图片所有视觉模式与其的相关性，最终网络抑制了无关的视觉模式并获得了图像分类的最优特征表示。基于此观察，基于类别协同的弱监督定位旨在寻找并保留更多与物体本身相关而不仅仅是类别相关的视觉模式，通过在定位激活图中激活该部分特征来补全之前被定位器忽略的目标区域，达到准确定位的目的。

#### 3.1.1 构建类别金字塔

为了寻找并保留更多与物体本身相关的视觉模式，我们需要构建新的标签来替代或补充原来的精细标签。具体来说，我们根据类别的相似性来构建类别金字塔，并认为在同一“大类”中的类别具有相似的视觉特征，通过额外给出新的“大类”标签来挖掘对于“大类”分类有益而对“子类”分类无益的特征。这部分特

征更多地保留了物体本身的视觉模式,在现有弱监督目标定位框架中却常常被抑制,在定位激活图中该部分不会被激活,进而影响了弱监督目标定位算法的效果。

我们在实验中的观察也印证了以上想法:在弱监督目标定位的主流数据集 CUB-200-2011 和 ILSVRC 中,精细的分类类别标签导致分类网络在训练的过程中专注于发掘类别间的微小差异,我们观察到此时的定位结果只能激活这些微小差异而不能定位激活物体的全部。

对于生物数据集,我们采用了生物学上的层级划分,依照生物分类学中的“界门纲目科属种”来构建“类别金字塔”,由于生物分类学中的类别层级划分很大程度参考了生物的形态结构,所以使用这种方法来构造金字塔对于寻找共同的视觉模式具有一定的合理性。数据集 CUB-200-2011[21]包含了以北美鸟类为主的 200 种鸟类图像,我们首先根据生物学中的分类(如图 3.1)将其划分为了 11 目、37 科、85 属、200 类,并为各层级中的每一个“大类”设立单独的标签,这样我们构建了具有四个层级的类别金字塔。

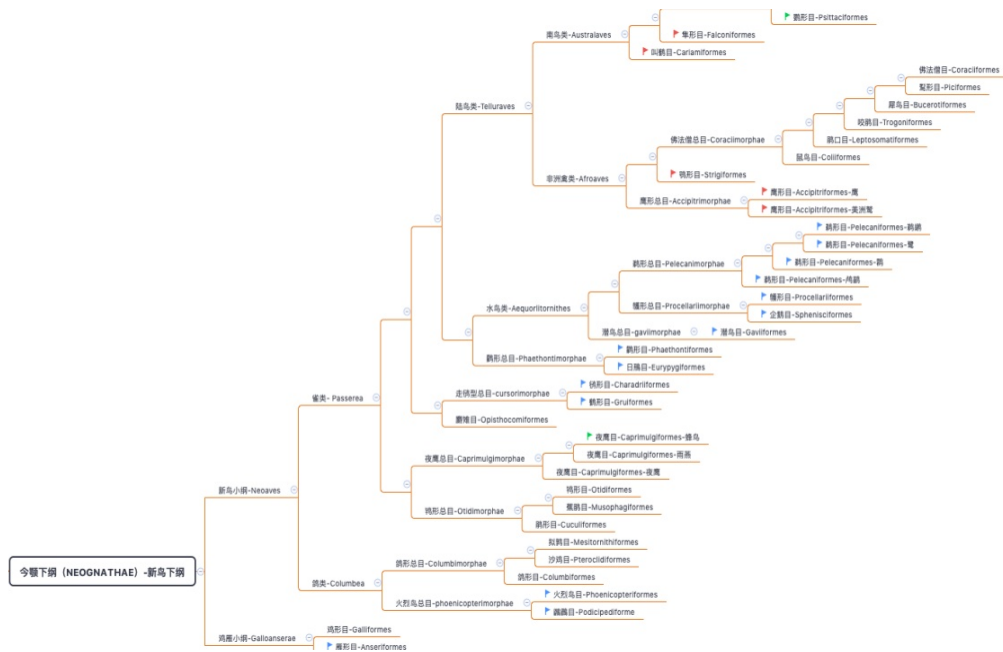


图 3.1 鸟类的类别金字塔

Figure 3.1 Pyramid of birds species

对于非生物数据集,我们结合知识图谱中的类别树获取各类别间的结构关系。知识图谱是由 Google 公司在 2012 年提出来的概念,在知识图谱的帮助下,搜索

引擎能够洞察用户查询背后的语义信息，返回更为精准、结构化的信息，更大可能地满足用户的查询需求。从学术的角度，我们可以将其理解为结构化的语义知识库，描述了物理世界中的概念及其相互关系；而从实际应用的角度，可以简单地把知识图谱理解成多关系图（Multi-relational Graph）。构建知识图谱时需要错综复杂的文档、数据进行有效的加工、处理、整合，并转化为简单、清晰的“实体-关系-实体”的三元组，以实现知识的快速响应和推理。图 3.2 是 ILSVRC[22] 数据集提供的类别层级结构，该结构来源于知识图谱 WordNet[57]。WordNet 中名词的连接关系使用了蕴涵关系（上位 / 下位关系），基于该层次关系我们可以构造相应的类别金字塔并用于类别协同的弱监督目标定位器，值得注意的是，在其他数据集中也可以采用类似方法构建类别层级结构。

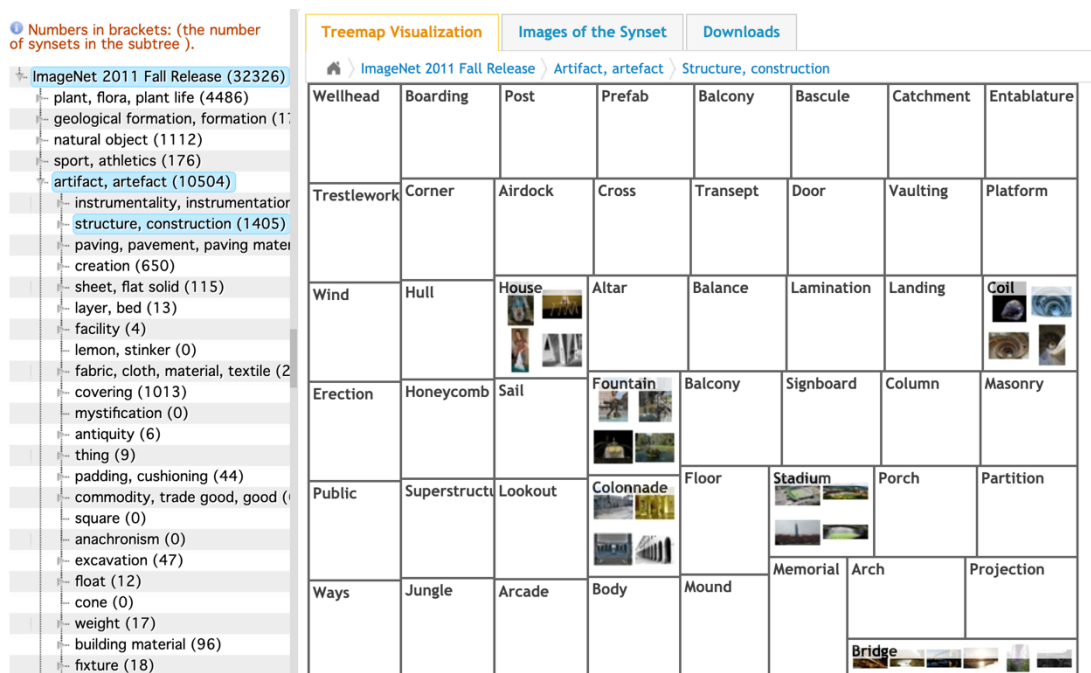


图 3.2 ILSVRC 的类别金字塔

Figure 3.2 Category pyramid of ILSVRC datasets

### 3.1.2 网络结构

如上文所述，我们将类别金字塔引入基于类别协同的弱监督目标定位器，旨在寻找并保留多层次标签带来的多种语义层级的视觉模式，这就要求我们在网络中同时构建多个定位器来分别对应各层级标签。目标检测中常用特征金字塔



[58][59]来解决多尺度的问题，金字塔的各层对应了不同分辨率的检测结果。特征金字塔的有效性说明网络在提取了基本特征之后，不同深度的层可以学习不同层次的语义特征：浅层分辨率高，学习目标的细节特征；深层分辨率低，学习目标的深层语义特征。受到特征金字塔的启发，我们将类别金字塔的各层对应到深度神经网络的各层中；粗粒度的分类器对应神经网络的浅层，学习简单的语义特征；细粒度的分类器对应神经网络的深层，学习复杂的语义特征。网络结构如图 3.3 所示。

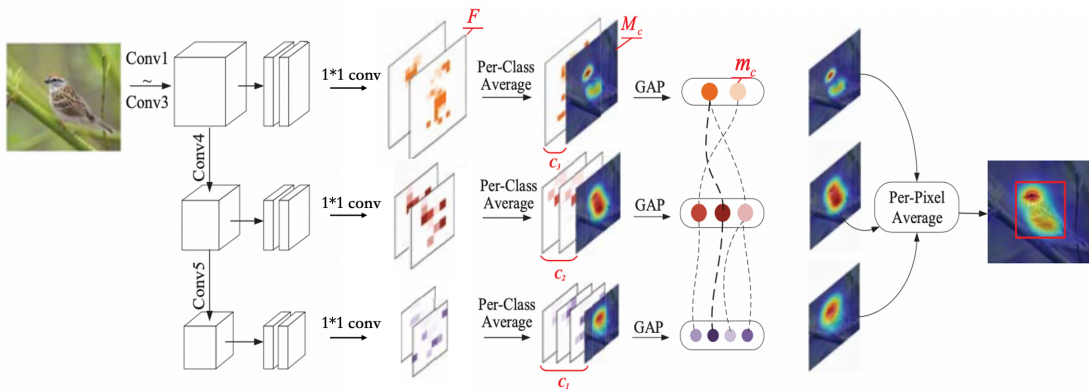


图 3.3 基于类别协同的弱监督定位器网络结构

Figure 3.3 Network architecture of weakly supervised locator based on category collaboration

以 VGGNet-16 网络为例，图 3.3 描述了基于类别协同的弱监督目标定位器的网络结构。我们在 VGGNet-16 网络的多个卷积组后面添加分类定位模块，图中给出的是使用三层类别金字塔的网络结构。我们使用 Conv1 和 Conv2 作为基本特征提取模块，Conv3、Conv4、Conv5 的输出分别连接粗粒度监督到细粒度监督的分类定位模块。对于每一级分类定位模块，我们添加两层带 ReLU 激活层的  $3 \times 3$  卷积层，并采用 CAM[1] 中的全局平均池化层 (Global Average Pooling, GAP) 来替代全连接层，以得到对应的分类和定位结果： $F$  代表每一级分类定位模块的最终特征， $M_c$  表示经过  $1 \times 1$  卷积得到的  $C_i$  张与类别标签一一对应的定位激活图， $i$  表示类别金字塔的层级， $C_i$  表示该层级的类别数目，经过 GAP 层之后，每张定位激活图被转化为对应类别的分类置信度  $m_c$ ， $m_c$  经过  $softmax$  计算得到  $p_c$ ，网



络通过计算 $p_c$ 与类别标签的交叉熵损失来训练分类定位网络。

该交叉熵损失表示如下：

$$\mathcal{L} = \sum_{i=0}^n \mathcal{L}^i = \sum_{i=0}^n \frac{1}{C_i} \sum_c y_c^i \log(p_c^i) \quad (3.1)$$

$$p_c^i = \frac{\exp(m_c^i)}{\sum_c \exp(m_c^i)} \quad (3.2)$$

其中， $i$ 表示类别层级的标号，总体交叉熵损失由各类别层级的交叉熵损失加和得到。 $y_c^i \in \{0,1\}$ 表示该训练样本是否含有 $c$ 类目标， $p_c^i$ 是分类置信度 $m_c^i$ 的 $softmax$ 结果，表示该训练样本中含有 $c$ 类目标的置信度。

### 3.1.3 实验结果及分析

**实验设置：**我们在 CUB-200-2011 数据集上进行探究实验，在 VGGNet-16 分类网络中添加各层级的类别监督，具体来说，在网络的浅层到深层分别添加了 11 目、37 科、85 属、200 类的分类监督，训练时各层级分类损失共同作用于网络参数优化，各层级分类定位器学习对应层级下的鸟类差异，测试时各层级分类定位器分别产生对应的定位激活图，我们通过融合这些激活图得到最终定位结果。实验中我们用分类错误率 Top-1、Top-5 和定位错误率 Top-1、Top-5 这四个指标来指示分类定位网络的分类定位能力，通过观察指标变化来探究类别协同对于弱监督目标定位器性能的影响。其中，Top-1 错误率表示每张测试样例置信度最高的结果预测错误的样例比例，Top-5 错误率表示每张测试样例置信度最高的 5 个结果预测均错误的样例比例。

**可视化结果：**为了验证类别协同的有效性，我们将类别协同的定位结果与基线方法做了可视化对比，可视化结果如图 3.4 所示。当分类标签粒度比较粗时，网络倾向于学习各大类之间的明显差异，在激活图上激活了更大范围的物体区域。从图中可以观察到：利用细粒度标签分类时（如图 3.4 左），弱监督目标定位激活的位置集中在鸟类的头部，这是由于仅在精细类别的监督下，网络致力于提高分类性能，从而仅保留了每类最具判别力的特征，抑制了类别间的公共特征，体现在定位激活图上就是仅有头部（最具判别力的区域）被定位器激活；而当使用粗粒度标签进行分类时（如图 3.4 右），“大类”间的差异被作为有判别力的特征

保留了下来。当我们使用多层次类别标签时，由于各语义层级具有差异，各层级激活的最具判别力视觉特征是不同的，就可以在定位激活图上激活目标各部分，定位到整个物体。

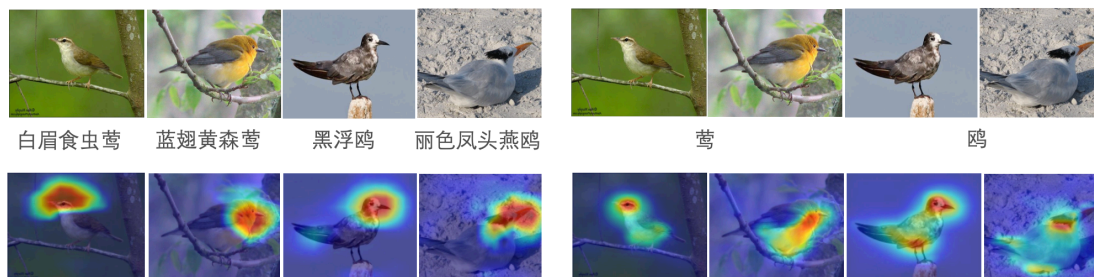


图 3.4 类别协同对 CUB-200-2011 数据集上定位结果的影响

Figure 3.4 Effects of category collaboration on WSOL results on CUB-200-2011 dataset

**消融实验：**为了探究基于类别协同的弱监督定位中类别层数的影响，我们对该部分做了探究实验，实验结果如表 3.1。从表格中可以观察到：随着类别层级的增加，定位错误率总体在逐渐降低，在采用三级层级结构时达到最低。当继续增加层级数时，错误率出现了回升，这是由于在特征金字塔结构的浅层上添加分类器时，浅层特征不足以支持分类器的学习，进而影响了深层特征和分类器的学习。

表 3.1 关于类别协同层级数的消融实验

Table 3.1 Ablation study on levels of category collaboration

方法	分类错误率 (%)		定位错误率 (%)	
	Top-1	Top-5	Top-1	Top-5
基线方法	23.42	7.47	55.85	47.84
类别协同 (两级)	24.27	7.28	52.80	44.56
<b>类别协同 (三级)</b>	<b>24.13</b>	<b>6.96</b>	<b>50.71</b>	<b>43.48</b>
类别协同 (四级)	24.20	7.13	51.75	43.90

**实验结论：**对比表 3.1 中第 1 行基线方法与第 2 至 4 行类别协同方法的实验结果，我们可以观察到：基于类别协同的弱监督定位方法可以有效地提高弱监督

定位的准确性,但是多级分类定位器的学习对于神经网络的特征提取有一定的影响,导致添加了类别协同监督的网络在分类性能上均有一定的下降。同时我们可以观察到虽然分类性能的 Top-1 错误率有一定程度的上升,但是 Top-5 错误率却有了略微的降低,这种现象有一定的合理性:当在定位器的训练过程中引入粗粒度标签时,神经网络的浅层特征中混入了粗粒度类别相关的特征,这导致网络对于细粒度的区分能力受到了抑制,Top-1 分类错误率提升;而此时虽然对相同粗粒度标签下的子类别分辨能力下降,但这些子类别的分类置信度均会上升,所以当测试范围扩展到置信度最高的前 5 个时,Top-5 分类错误率会相应下降。值得注意的是,虽然不同级数的类别金字塔对于弱监督目标定位增益的幅度不同,但是即使分类结果略微变差,定位结果均会更优。当采用合适的类别金字塔级数时,本方法可以达到 5.14%/4.36%的 Top-1/Top-5 定位错误率降低;结合可视化结果图 3.4,我们可以得出结论:基于类别协同的弱监督定位方法通过引入粗粒度类别相关的特征更好地完成了弱监督目标定位任务。

### 3.2 基于训练过程协同的弱监督目标定位

如图 3.5,我们发现弱监督分类定位器在训练的过程中,同其他计算机视觉任务一样,网络关注的位置呈现跳跃状态,在损失函数的驱动下不断寻找最具不变性的特征、最具判别力的区域来完成网络收敛。前文提到,在该过程中,神经网络不停地捕获相关特征并通过抑制与类别标签无关的视觉模式来压缩深度特征,而对于弱监督目标定位来说,保留该部分被压缩的特征可以增益定位性能。上一节中,我们使用类别协同补充了粗粒度标签相关的特征,通过类别层级关系的辅助信息减少了这种压缩带来的信息损失,本节中我们利用训练过程中神经网络关注的不同位置,通过将训练过程中的有效特征积累起来来丰富目标的视觉表达,增益目标定位性能。

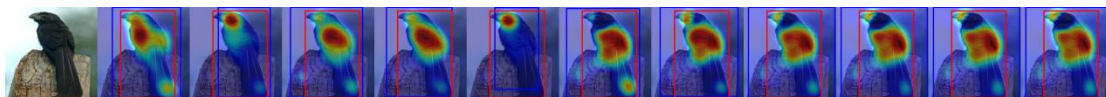


图 3.5 训练过程中的激活区域变化

Figure 3.5 Changes of activated regions during training

### 3.2.1 各阶段协同弱监督定位的有效性

首先，我们验证了聚合弱监督定位中各阶段定位结果的有效性。对于 VGGNet-16 框架，首先为了保证定位激活图的分辨率并提高定位的准确性，我们将 Conv-5 的池化层删除，同时将网络后两层的全连接层替换为带 ReLU 激活层的  $3 \times 3$  卷积层。为了得到与类别对应的定位激活图，我们添加了  $1 \times 1$  卷积层来得到对应的  $C$  张定位激活图，其中  $C$  表示物体类别。我们用  $F$  来表示  $1 \times 1$  卷积层的定位激活输出，与前述一样采用全局池化层和 *softmax* 来获得分类置信度。

$$p_c = \text{softmax}(\text{GAP}(F_c)) \quad (3.3)$$

之后采用简单的归一化操作来得到相应的定位激活图。

$$A_c = \frac{\text{ReLU}(F_c)}{\max(F_c)} \quad (3.4)$$

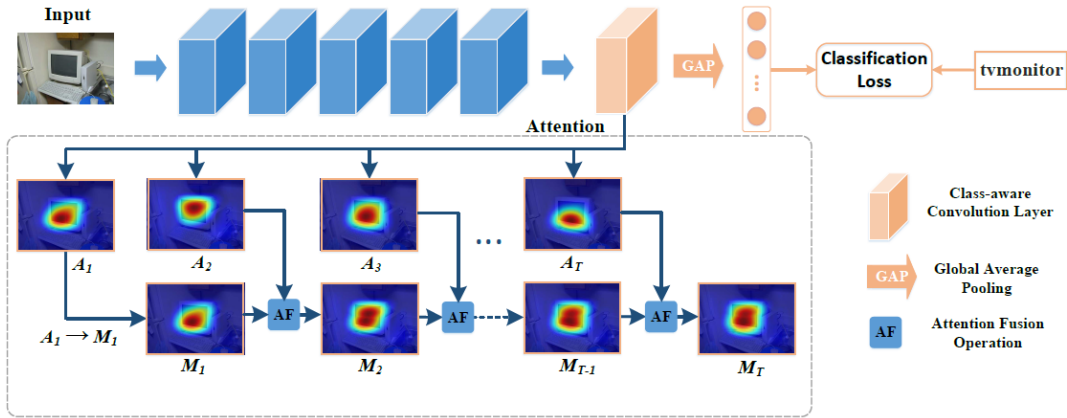


图 3.6 训练过程定位结果聚合网络

Figure 3.6 Architecture of results aggregation network during training process

为了聚合各训练阶段的定位结果，我们探究了不同聚合方法对弱监督定位结果的影响。如图 3.6 所示，对于输入图片  $I$ ，在不同的训练阶段我们可以获得其定位激活图  $A_1, A_2, \dots, A_T$ ，使用  $A_1$  初始化聚合定位结果  $M_1$ ，之后使用以下公式逐步更新  $M_t$ ：

$$M_t = \text{AF}(M_{t-1}, A_t) \quad (3.5)$$

具体的我们采用以下两种方式更新  $M_t$ ：

$$M_t = \text{AF}(M_{t-1}, A_t) = \max(M_{t-1}, A_t) \quad (3.6)$$

$$M_t = AF(M_{t-1}, A_t) = \frac{1}{t}((t-1)M_{t-1} + A_t) \quad (3.7)$$

表 3.2 聚合方式对基于训练过程协同弱监督定位的影响

**Table 3.2 Impact of aggregation methods on WSOL method based on training process collaboration**

方法	分类错误率 (%)		定位错误率 (%)	
	Top-1	Top-5	Top-1	Top-5
基线方法	23.68	6.90	57.11	48.48
最大值聚合-最终类别	-	-	55.54	46.05
最大值聚合-当前类别	-	-	54.77	45.23
平均值聚合-最终类别	-	-	58.11	48.54
平均值聚合-当前类别	-	-	57.63	49.03

表 3.2 中，“最大值聚合”表示按照公式 3.6 进行聚合结果更新，聚合结果中各点的取值采用各训练阶段定位激活结果在该位置的最大值，而“平均值聚合”表示按照公式 3.7 进行聚合结果更新，各点取值为各训练阶段结果在该位置激活值的平均；“最终类别”表示根据训练结束时的分类结果来选取各阶段激活定位结果，而“当前类别”表示根据当前训练阶段的分类结果来选取激活定位结果。从表中的实验结果可以看出来：采用“最大值聚合”要比“平均值聚合”得到更好的定位结果，这与我们的实验设计初衷比较一致：可以保留各阶段对于定位激活有益的视觉特征。由于训练过程中网络的收敛逐渐放缓，而我们的聚合在训练的各阶段等间隔采样，若采用“平均值聚合”方法，训练前期有用的特征就会在平均的操作中被后续结果抑制，同时各阶段的噪声也被保留下来，最终的定位结果要逊于基线方法；而采用“最大值聚合”方法时，各阶段仅保留有意义的视觉特征，我们得到了优于基线方法的定位结果。当我们横向比较“最终类别”和“当前类别”的定位性能时，可以发现虽然两者都有优于基线方法的结果，但“当前类别”在两种设定下都优于“最终类别”的定位性能（约 1%），这可能是由于在网络训练的过程中，即使分类还没有收敛，某些阶段选用的是错误类别的定位结果，但是由于定位激活的都是前景信息，这部分信息对最终的定位结果也是有益

的。为了验证该猜想，我们进一步的进行了下面的实验：探究弱监督目标定位中其他类别的协同辅助作用。

我们选取最终分类结果的 Top-5 类别来进行定位激活聚合，与上文一样采用“最大值聚合”和“平均值聚合”两种方式，当采用“最大值聚合”时得到了 53.68% 的 Top-1 定位错误率，而采用“平均值聚合”时得到了 59.89% 的 Top-1 定位错误率。由此可见，分类置信度靠前的类别都激活了有意义的前景信息，这部分信息对于弱监督定位是有益的。

### 3.2.2 基于训练过程协同的弱监督目标定位器

上文中验证了测试过程中聚合训练各阶段的定位结果对于弱监督定位的有效性，但是由于该方法需要保留弱监督定位器各阶段的模型，在测试时对同一张样例图片进行多次测试，这在时间效率和空间存储上都造成了一定的冗余。除此之外，我们希望探究用历史信息监督的定位将如何影响神经网络浅层特征，以及是否会对分类定位结果产生增益，在下面的章节中我们将构建训练过程协同的弱监督定位器，给出基于训练过程协同的弱监督定位器的调优方案。

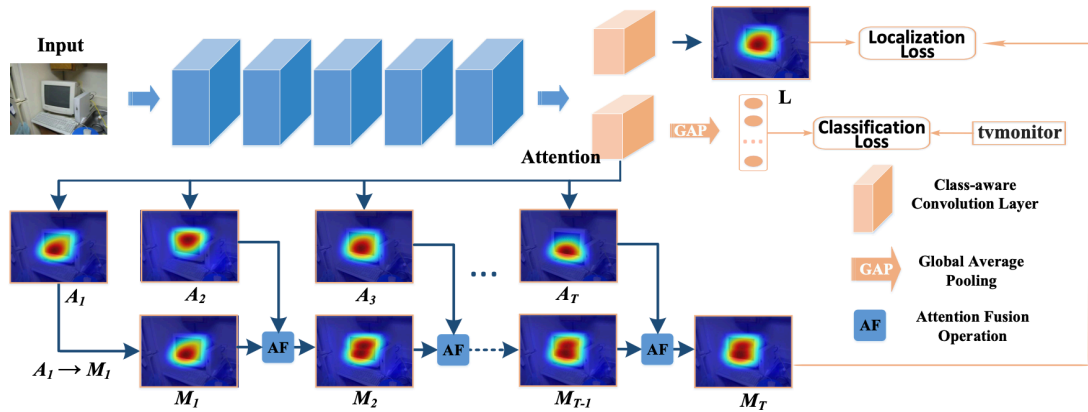


图 3.7 基于训练过程协同的弱监督目标定位

Figure 3.7 Architecture of weakly supervised locator based on training process collaboration

首先我们构建了单分支和双分支两种基于训练过程协同的弱监督定位器。根据上节中对于聚合方式的探究实验结果，两种网络结构中我们均采用了“最大值聚合”的方式。在单分支的定位器中，我们采用基础的弱监督定位器结构[1]，保留每次网络前传得到的定位结果 $A_t$ ，通过“最大值聚合”方式不断更新 $M_t$ ，并将

$M_t$ 作为 $t + 1$ 轮迭代的定位真值，通过计算分类损失和点对点（pixel-to-pixel）的定位损失更新网络参数。在双分支的定位器中，如图 3.7 所示，我们在特征提取器（例如 VGGNet-16 的 Conv5）的后面添加分类和定位两个分支，分类分支与原始的弱监督定位器结构相同，定位分支也同样采用了两层带 ReLU 激活层的  $3 * 3$ 卷积层和一层  $1 * 1$ 卷积层，我们用训练样例类别标签和本轮聚合方式结果  $M_t$ 分别作为双分支定位器中分类分支和定位分支的监督。值得注意的是，双分支定位器中的聚合方式结果  $M_t$ 来自于分类分支而不是定位分支，这是由于我们希望保留分类监督下神经网络捕获的相关特征，而不是训练不完全的定位分支结果。

表 3.3 给出了上述两种结构弱监督定位器的实验结果。从表格中可以看到，若使用单分支同时分类和定位，会导致分类错误率的大幅度提高，从而影响了定位性能；若使用双分支定位器，虽然仍然会小幅度影响分类性能，但是也大幅度降低了定位错误率（4.02%）。当我们将该训练过程监督下的定位结果与上一节中训练过程聚合的定位结果比较时，也可以看到：在分类性能下降 1.14%（23.68%/24.82%）的情况下，仍然取得了 1.64%（54.77%/53.09%）的定位性能提升。该结果说明在弱监督定位任务下，用定位监督影响神经网络浅层特征会对定位结果产生积极的影响。

表 3.3 网络结构对基于训练过程协同弱监督定位的影响

Table 3.3 Effect of network architecture on WSOL method based on training process collaboration

方法	分类错误率 (%)		定位错误率 (%)	
	Top-1	Top-5	Top-1	Top-5
基线方法	23.68	6.90	57.11	48.48
单分支	28.48	9.60	59.32	48.69
双分支	24.82	8.58	<b>53.09</b>	<b>44.10</b>

在双分支定位网络的训练过程中，我们使用中间过程产生的定位激活结果  $M_{t-1}$ 来监督本轮定位。对于分类分支，与前述一样，用 *softmax*交叉熵函数来监督分类；对于定位分支，我们采用上一轮聚合结果  $M_{t-1}$ 作为真值来监督定位，同



时采用点对点的  $MSE$  loss 作为监督函数来计算定位损失。

$$\underset{\alpha}{\operatorname{argmin}}\{\mathcal{L}_C(\alpha) + \mathcal{L}_L(\alpha)\} = \underset{\alpha}{\operatorname{argmin}}\left\{\mathcal{L}_C(\alpha) + \sum_{c \in \mathcal{C}} (\mathcal{L}_+^c + \mathcal{L}_-^c)\right\} \quad (3.8)$$

$$\mathcal{L}_+^c = \frac{1}{N_+^c} \sum_{j \in N_+^c} \operatorname{MSE}(L_j^c, M_{t-1,j}^c) = \frac{1}{N_+^c} \sum_{j \in N_+^c} (M_{t-1,j}^c - L_j^c)^2 \quad (3.9)$$

$$\mathcal{L}_-^c = \frac{1}{N_-^c} \sum_{j \in N_-^c} \operatorname{MSE}(L_j^c, 0) = \frac{1}{N_-^c} \sum_{j \in N_-^c} (L_j^c)^2 \quad (3.10)$$

其中,  $M_{t-1}$  表示采用上述聚合方法  $AF(M_{t-1}, A_t)$  得到的训练过程定位结果,  $L$  表示本轮定位分支的定位结果,  $N_+^c$  表示  $M_{t-1}$  上非零像素点的个数,  $N_-^c$  为零像素点的个数,  $j$  表示像素位置。

### 3.2.3 实验结果及分析

**实验设置:** 我们在 CUB-200-2011 数据集上做了进一步的探究实验, 在 VGGNet-16 分类网络上构建了基于训练过程协同的双分支定位网络, 在测试时从分类分支得到分类结果, 定位分支得到定位结果, 同样通过分类错误率 Top-1、Top-5 和定位错误率 Top-1、Top-5 来指示模型的分类和定位性能, 从而探究训练过程协同对于弱监督目标定位器的影响。

实验中我们探究了定位分支中定位损失函数的影响, 我们使用了图像分割任务中常用的三种损失函数: L1 loss、MSE loss 和 BCEloss。其中 L1 loss 和 MSE loss 将该定位问题建模为回归问题, 让网络输出结果回归到聚合结果  $M_t$ ; 而 BCEloss 将定位问题建模为分类问题, 将每个像素点的定位结果建模为该点为正样本的置信度。在本实验中使用两种回归损失都可以保证定位分支的收敛, 但是使用 BCEloss 损失函数时定位模块输出的结果出现“斑驳”现象, 这是由于需要用阈值界定正负样本, 而随着训练的进行, 合适的阈值无法被明确界定, 反而对训练造成了困难。进一步的, 我们探究了两种回归损失对定位模块的影响: MSE loss 函数曲线光滑、连续、处处可导, 便于使用梯度下降算法, 随着误差的减小, 梯度也在减小, 有利于收敛; 而 L1 loss 大部分情况下梯度都是相等的, 即使对于小的损失值, 其梯度也比较大, 不利于损失函数的收敛和模型的学习。最终我们选择了 MSE loss 作为定位模块的损失函数。



**消融实验：**对比直接聚合定位结果和在用聚合结果监督定位模块的实验结果，我们发现网络监督学习可以得到更好的定位性能（53.09%/54.77%）。这是由于当我们使用随机梯度下降训练网络时，网络参数会朝着梯度下降最快的方向变化，对应到根据聚合结果 $M_t$ 学习定位激活的学习目标，网络优先学习了统计上出现频率高的视觉模式，这部分收敛之后再去学习统计频率低的视觉模式，导致后者可能无法完全收敛。由于我们的聚合结果 $M_t$ 并不是完全正确的真值，而是含有噪声的伪标签，这种学习方式给定位模块带来了更好的“容错性”，从而增益了弱监督定位性能。接下来的实验中，我们将从这一观察出发，进一步改进基于训练过程协同的弱监督定位器。

**正负样本平衡：**在分割任务中，背景像素数量往往多于前景像素数量，这就造成了正负样本不平衡的问题。样本不均衡问题分为内在不平衡和外在不平衡两种，其中内在不平衡指的是数据本身特性所造成的不平衡性，外在不平衡指的是由于对训练数据的采样不足，造成了某种不平衡，往往可以通过增加训练样本解决。本模型的定位模块与分割任务一样，属于数据本身分布造成的内在不平衡问题。对于内在不平衡数据的处理，基本的方法是采用合理的性能评价指标，避免大样本类淹没了小样本类，比如对不同类别赋予不同重要程度的权值，小样本赋予更大的权值，这样对小样本类别的网络前传结果也能进行有效的损失计算与回传。本节中采用这种方法来处理前景/背景数量差异大造成的正负样本不平衡问题。

表 3.4 正负样本平衡对基于训练过程协同弱监督定位的影响

Table 3.4 The effect of class imbalance on WSOL method based on training process collaboration

正负样本 loss 权重比例	分类错误率 (%)		定位错误率 (%)	
	Top-1	Top-5	Top-1	Top-5
1:1	24.82	8.58	53.09	44.10
3:1	23.18	7.92	<b>52.66</b>	<b>43.46</b>
5:1	23.09	7.70	52.99	43.79

表 3.4 中探究了正负样本平衡对基于训练过程协同弱监督定位的影响。从表格中可以看出，当设置正负样本 loss 权重比为 3:1 时，可以达到最低的定位错误率。对比第一行中不针对正负样本平衡做优化的实验结果，可以将定位错误率降低 0.43% (52.66%/53.09%)。

**上采样模块：**在测试过程中，弱监督定位器首先将定位激活结果上采样到图片大小，再通过设定的阈值将该结果分为前景和背景并得到定位结果。由于定位结果与图片分辨率往往差别较大，导致上采样时没办法捕捉到细节信息从而影响了准确的定位。受上文中“神经网络对伪标签噪声容错率较好”这一实验观察的启发，我们将聚合结果 $M_t$ 进行 2 倍上采样并将该结果作为定位模块新的伪标签，这样测试时定位模块输出结果的分辨率也可以相应提高，可以部分减少测试过程中分辨率低带来的定位不准问题。实验表明，这种策略可以有效增益弱监督定位（错误率 52.14%/52.66%）。

**松弛阈值：**受前文中“神经网络对伪标签噪声容错率较好”这一实验观察的启发，本节中我们希望加强定位分支对伪标签中噪声的包容性。弱监督目标定位中通过全局平均池化得到样例在各类的分类置信度，并将全局平均池化之前的特征图作为该类的定位激活图。这种方式实际上将特征图中各像素点激活值的大小都看作了该点在该类别的置信度。对于目标类的前景像素点，我们希望其激活值接近于“1”；对于背景像素点，我们希望其激活值接近于“0”。在网络训练的前期，网络参数没有完全向着目标函数收敛，前景/背景类没有完全分开，这时会出现一定的混杂现象，而该现象表现为伪标签聚合结果 $M_t$ 中的噪声。经过上述分析我们知道：我们希望保留更多训练过程中跟目标相关的视觉模式，而不仅仅是最具判别力的视觉模式，但同时我们也保留了更多的噪声，接下来的实验中我们希望通过改进正负样本的判定方式来尽可能在保留更多视觉模式的同时剔除噪声的影响，这样我们就可以为定位分支提供更准确的伪标签。

我们采用更改前景/背景样本的判定方式来为定位分支提供容错率更高的伪标签。具体实现方式为：在伪标签聚合结果 $M_t$ 上设定阈值，用该阈值圈出伪标签中的可忽略区域，该区域对于分类分支的预测结果贡献较小，但是将其设定为前景会干扰定位分支的定位结果。表 3.5 中我们探究了该区域的合适阈值 $th$ ，伪标

签中激活值为 $(0, th)$ 的位置为忽略区域，该区域的定位损失不计入神经网络训练的反传损失中。当激活值阈值 $th$ 设定为为 0.05 时，分类错误率降低了 0.43% (22.75%/23.18%)，定位错误率降低了 2.76% (49.91%/52.66%)，而当阈值进一步增大时，定位错误率大幅度反弹。这组实验说明通过设置伪标签中合适的忽略区域，可以进一步增加定位分支的容错能力从而提高弱监督定位的性能。

表 3.5 松弛阈值对基于训练过程协同弱监督定位的影响

Table 3.5 Ablation study of threshold relaxation on WSOL method based on training process collaboration

th	分类错误率 (%)		定位错误率 (%)	
	Top-1	Top-5	Top-1	Top-5
0 (基线方法)	23.18	7.92	52.66	43.46
0.05	22.75	7.40	<b>49.91</b>	<b>40.65</b>
0.10	23.28	7.58	55.49	46.70

在测试时我们也需要通过设定阈值来判定定位激活图中的前景和背景，从而框出目标位置，而该阈值要大于忽略区域的最佳阈值 0.05。由此可见该方法的有效原因不仅仅在于纠正了伪标签，更多的在于：剔除了该部分噪声之后，定位分支的网络学习目标更一致，网络可以收敛得更好。随着训练的进行，分类分支判断前景/背景的能力提高，一些在训练初期被判定为忽略区域的像素点在伪标签中更准确地被标记为前景/背景，而无需作为伪标签噪声伴随在定位分支训练的始终，对定位分支造成影响。

**实验结论：**我们探究了基于训练过程协同的弱监督目标定位器的有效性，并给出了模型细节相关的消融实验结果。由于弱监督定位器仅由类别标签驱动，网络收敛时定位激活图中仅保留了对分类有益的最具判别力的视觉模式，而训练过程中网络所发现的目标相关但是判别力较差的视觉模式会被忽略，我们的方法通过保留这部分视觉模式来增益弱监督定位。实验表明，基于训练过程协同的弱监督目标定位器需要将分类分支与定位分支分开来保证分类分支的有效收敛，通过正负样本平衡可以提高定位分支对于前景/背景区域的判别力。进一步的，我们

通过添加上采样模块和设置伪标签中的忽略区域来增加模型对于伪标签噪声的容错能力，使得弱监督定位器的性能得到了大幅提高。

### 3.3 本章小结

本章中我们描述了两种基于协同思想的弱监督定位器，协同思想对于弱监督定位任务有效的原因在于：挖掘出了更多目标相关的视觉模式，并在定位时激活了该区域。基于类别协同的弱监督定位方法通过引入类别层级结构先验信息，从类别间的层级关系中挖掘共性视觉特征来辅助弱监督目标定位，将特征的获取来源从精细类别标签扩展到粗粒度标签，通过放松分类任务中类别间的相互抑制来获取更多的前景视觉特征，更好地完成了弱监督定位。基于训练过程协同的弱监督定位方法挖掘并保留训练过程中被抑制的早期目标特征表示，实验中我们探究了历史信息的保留方式和弱监督定位器合理的网络结构，并通过增加模型对伪标签噪声的容错能力进一步提高了定位器的性能。

## 第4章 基于分歧的弱监督目标定位

上一章中我们探究了基于协同思想的弱监督定位器,通过多源信息的融合来增益弱监督定位器的视觉模式。本章中我们旨在改善神经网络训练过程中的特征紧凑性问题并减少其对视觉模式的压缩现象,构造基于分歧思想的弱监督定位器,从原始类别标签中挖掘更多的隐含信息。

### 4.1 研究动机及建模过程

在弱监督分类定位器训练的过程中,倾向于保留最具有不变性的特征来支撑分类,这种现象归因于卷积特征的固有紧凑性。利用分类目标函数来训练网络,网络学习的唯一目标是捕获与类别标签最相关的视觉模式[2]。类别标签隐含地确定了深度特征中的相关特征和不相关特征,训练过程中网络将通过抑制无关的视觉模式来压缩深度特征。考虑到深度特征和定位激活图之间的对应关系,深度特征的压缩会导致定位激活图的稀疏,从而影响弱监督定位结果。为了减轻上述视觉特征的抑制现象,需要改善单一目标函数导致的特征紧凑性的问题。

本章中我们通过分歧的思想来解决上文中提到的特征紧凑性的问题。如第二章中分歧思想相关技术的描述,其假设对于同一数据可以从不同的角度进行学习,并训练出不同的模型,由于这些模型是从不同角度训练出来的,其互补性可以提高模型精度,就如同从不同角度可以更好地理解事物一样。该思想的适用性很广,因为它既不需要大量冗余的视图,也不需要对所采用的监督学习算法施加任何约束。基于分歧的学习器有两种构建形式,一种是使用对抗性示例,另一种是约束多个分类器之间的差异,我们采用第二种构建方式:对各个特征提取器进行差异约束,以得到分歧但具有判别力的特征。该框架有两点需要注意的地方:一是各学习器中特征的表达方式,一是多个学习器的差异约束方式。关于神经网络的特征表达有很多相关研究,近年来主要集中在跨域自适应学习(Domain Adaptation)方向,该方向旨在将在已知样本域训练的学习器推广到未知的样本域,其关键点在于已知域和未知域的特征表达与对齐。跨域自适应学习中特征表达的研究重点

在于寻找更多可传递的表示，包括类别准则（利用类标记或度量学习作为特征表示的引导）、统计准则（利用 KL 散度、 $\mathcal{H}$ 散度等统计方法表示特征）、结构准则（使用自适应批量归一化、域引导 dropout 等技术调整网络结构改善特征）、几何准则（通过将源域和目标域特征投影到构建的中间子空间来对齐分布）。对于特征表示的差异衡量方式也有很多种，包括直观的距离衡量方式（L1 距离，L2 距离等）、统计距离衡量方式（海林格距离等）、基于类别的概率条件分布等。

在本方法中我们采用定位激活图作为特征表示方式。一方面，定位激活图与类别相关，在定位激活图上进行差异约束实际上是对该类的相关特征进行约束，每张激活图均与网络浅层特征有权值连接，该连接可以将差异约束的影响传导到相关的浅层特征上，有效地抑制了差异约束对网络分类能力的不良影响；另一方面，定位激活图与定位结果直接相关，差异约束下更多的区域被激活，节约存储和计算开销的同时实现了差异约束的目标。具体来说，我们计算了定位激活图之间的余弦相似度，并将该值作为网络的损失函数来约束特征之间的差异。随着网络训练的进行，在余弦相似度损失函数的引导下，同类定位激活图之间的相似度降低，不同的定位激活图激活了不同的区域，达到了学习分歧的弱监督定位器的目的。

## 4.2 研究结构框架

具体来说，我们引入了分歧目标函数来与分类目标函数联合优化，以在定位激活图上保留尽可能多的视觉模式。

对于分类损失我们采用常用的交叉熵公式。对于输入图片 $I$ ，弱监督定位器产生类别 $c$ 的定位激活图 $F_c$ ，经过全局池化层， $f_c = \frac{\sum_{i,j} F_c(i,j)}{H*W}$ ，和softmax层我们得到了该类的分类置信度 $p_c = \frac{\exp(f_c)}{\sum_c \exp(f_c)}$ 。联合优化目标函数具体形式如下：

$$\operatorname{argmin}_{\alpha} \{ \mathcal{L}_C(\alpha) + \lambda \mathcal{L}_D(\alpha) \} \quad (4.1)$$

$$\mathcal{L}_C(\alpha) = -\frac{1}{C} \sum_c y_c \log(p_c) \quad (4.2)$$

其中 $\mathcal{L}_C$ 表示分类目标函数，我们采用的是交叉熵函数（公式 4.2）， $y_c$ 表示图片类

别标签。 $\mathcal{L}_D$ 表示类内分歧损失， $\lambda$ 表示权重因子。

为了得到多个分歧的定位激活图，我们首先将类别 $c$ 的定位激活图由 1 张扩展为 $K$ 张，标记为 $A_c^k (k = 1 \dots K)$ 。我们希望每张 $A_c^k$ 可以激活定位到目标的不同位置，其激活结果相互分歧，采用了如下的目标函数：计算类定位激活图 $A_c^k$ 两两之间的余弦相似度，训练网络使其相似度降低。如图 4.1 所示，当仅仅将激活图扩展为 $K$ 张时，多张激活图的激活区域比较相近，而引入了分歧激活目标函数之后，多张激活图倾向于激活不同位置，达到了分歧定位的目的。

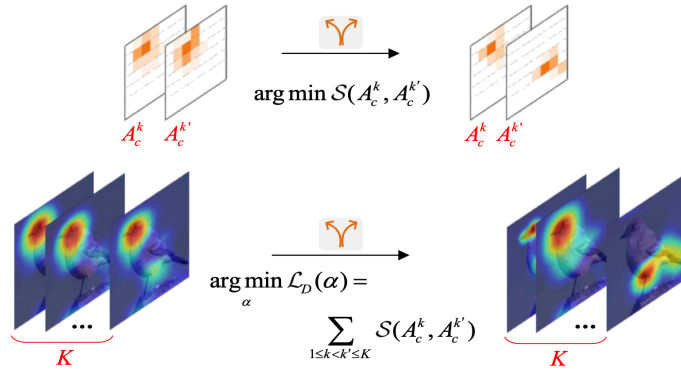


图 4.1 基于分歧的弱监督定位模块

Figure 4.1 WSOL module based on divergence

在训练中我们将 $K$ 张定位激活图两两组合，得到 $C_R^2$ 组定位结果。之后计算每组定位激活图之间的相似度，并将其相加作为分歧激活的相似度损失。为了让网络更快地收敛，实验中我们只随机选取 $C_R^2/2$ 组结果的相似度纳入本轮迭代的分歧损失中反传；为了节约模型的存储并提高训练效率，我们只针对训练样例的目标类作相似度约束。差异约束损失函数定义如下：

$$\mathcal{L}_D(\alpha) = \sum_{1 \leq k < k' \leq K} \mathcal{S}(A_c^k, A_c^{k'}) \quad (4.3)$$

$$\mathcal{S}(A_c^k, A_c^{k'}) = \frac{A_c^k \cdot A_c^{k'}}{\|A_c^k\| \|A_c^{k'}\|} \quad (4.4)$$

其中， $A_c^k$ 表示训练样例 $I$ 在目标类 $c$ 的第 $k$ 张定位激活图， $\mathcal{S}(A_c^k, A_c^{k'})$ 表示第 $k$ 张和第 $k'$ 张的余弦相似度。将 $C_R^2$ 组余弦相似度相加就得到了分歧激活损失。

值得注意的是将定位激活图扩展为 $K$ 张之后，训练过程中，我们将 $K$ 张激活

图聚合得到一张定位激活图，再进行全局池化和计算交叉熵损失等操作；而测试过程中，我们先对每张定位激活结果进行取正值（Relu）操作，再聚合得到最终的定位结果，相关的实验与阐述将在下一节中详细介绍。

### 4.3 实验结果及分析

本节中我们将在 2018 年张晓琳的弱监督定位工作 SPG[15]的基础上进行实验，该模型在发表当年达到了弱监督目标定位领域领先的性能。选择其作为基线方法是为了进一步增益弱监督定位性能，同时验证我们基于分歧定位方法的有效性与可推广性。

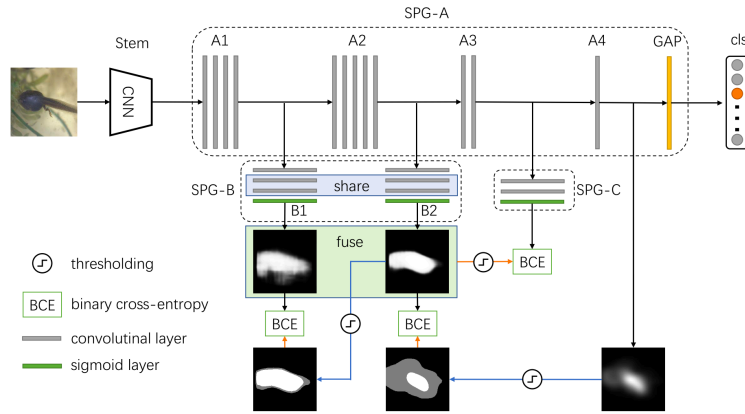


图 4.2 SPG 网络结构框图

Figure 4.2 Network architecture of SPG method

如第二章中对于弱监督定位相关工作的介绍，SPG 方法旨在发现已激活区域与未激活区域的联系，将已激活区域作为“种子”，在网络训练过程中逐步发现其与其他区域的关联性，并学习发现更可靠的目标区域。其网络结构如图 4.2 所示，网络不同深度的特征分别产生定位结果，通过逐级监督来更准确地定位目标。我们将定位分歧模块搭建在 A4 模块后面，该模块对于每张训练样本的每类输出 K 张定位激活图，计算相似度损失之后，取其平均得到该类聚合结果，再进行全局池化操作并计算分类损失。

我们在基于 VGGNet-16 的 SPG 定位网络上搭建我们的定位分歧模块，并在 CUB-200-2011 数据集上进行弱监督定位实验，同样用分类错误率 Top-1、Top-5



和定位错误率 Top-1、Top-5 来反映模型的分类定位能力，探究分歧模块对于弱监督目标定位的影响。

### 4.3.1 消融实验

首先我们进行了关于分歧模块中分歧定位结果张数 $K$ 的消融实验，结果如图4.3所示。需要说明的是，此时还没有添加分歧损失约束。从图表中可以看出，随着定位激活结果个数 $K$ 的增加，分类错误率和定位错误率都呈先下降再升高的趋势，这说明 $K$ 在合适范围内取值时，增加定位激活结果个数会对弱监督定位结果产生正面影响。观察图表中各指标随 $K$ 的变化趋势，我们发现：当 $K = 4$ 时可以得到最佳的分类结果，而随着 $K$ 的继续增加，分类错误率持续上升；Top-1 定位错误率在 $K \in [4,16]$ 时变化并不明显，而受分类错误率升高影响较小的 Top-5 定位错误率却在 $K = 16$ 时达到了最低，这说明在 $K \in [4,16]$ 时我们的定位网络的定位能力越来越好，但是由于受到分类错误率升高的牵制，Top-1 定位错误率结果在该区间变化并不明显；当 $K$ 取值继续上升时，定位能力的提高并不能平衡分类能力下降带来的损失，定位错误率和分类错误率均超过了我们的基线方法 SPG。

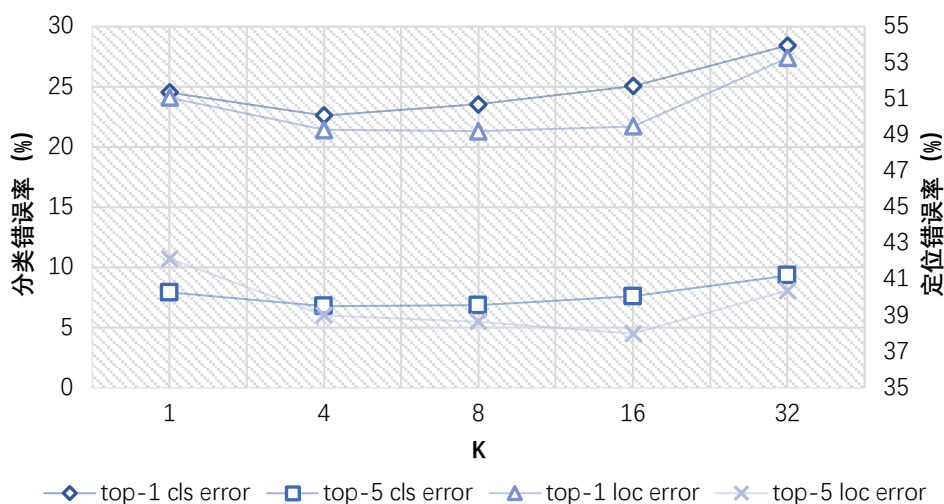


图 4.3  $K$  对于基于分歧的弱监督定位器的影响

Figure 4.3 Ablation study of  $K$  on WSOL method based on divergence

从上述消融实验中我们得出结论：随着 $K$ 的增加，弱监督定位网络的定位能力在增强，而分类能力在减弱。这是由于将定位激活结果扩展为 $K$ 个这一操作会

对分类网络特征的学习造成影响。在小范围内扩展 $K$ 值时，由于分类有益的视觉模式也在增加，会提高模型的分类能力；但继续增加 $K$ 值时，网络引入了更多分类无关的视觉模式，其分类能力受到了影响，所以 $K$ 最佳的取值应为网络分类能力与定位能力的平衡点。当 $K = 8$ 时，网络的定位错误率相较于基线方法 SPG 降低了 1.86% (49.21%/51.07%)，说明多视角的思想对弱监督定位任务有效，且能在高基线方法的基础上进一步提高性能。

表 4.1 聚合方式对基于分歧的弱监督定位方法的影响

Table 4.1 Impact of aggregation methods on WSOL based on divergence

实验编号	分类聚合方式	Cls error (%)		定位聚合方式	Loc error (%)		分歧约束
		Top-1	Top-5		Top-1	Top-5	
1	$\sum_{k=1}^K A_c^k$	23.54	6.85	$\sum_{k=1}^K A_c^k$	49.21	38.66	无
				$\sum_{k=1}^K Relu(A_c^k)$	49.17	38.57	
2	$\sum_{k=1}^K A_c^k$	23.78	7.56	$\sum_{k=1}^K A_c^k$	49.91	40.90	有
				$\sum_{k=1}^K Relu(A_c^k)$	45.54	35.24	
3	$\sum_{k=1}^K Relu(A_c^k)$	27.44	8.96	$\sum_{k=1}^K A_c^k$	46.96	34.73	有
				$\sum_{k=1}^K Relu(A_c^k)$	44.98	32.00	

进一步的，我们探索了基于分歧的弱监督定位器中分歧约束的影响以及分类定位结果的合理聚合方式。我们进行了三组实验，均采用  $K = 8$  的实验设定，这时网络分类能力与定位能力较为均衡。三组实验中，测试阶段均采用了两种聚合方式，即直接聚合  $\sum_{k=1}^K A_c^k$  和先对每张结果取正值 (Relu) 再聚合  $\sum_{k=1}^K Relu(A_c^k)$ 。

实验一中，我们未采用分歧性约束，训练阶段先直接聚合各激活结果图，再全局池化来得到各类别置信度 ( $m_c = GAP(\sum_{k=1}^K A_c^k)$ )，实验表明两种聚合方式的定位性能相差不大。实验二中，我们仍采用实验一的分类聚合方式，同时添加了有合适权重 ( $\lambda$ ) 的分歧约束，可以看到分类未受到分歧约束的影响，而定位聚合方式对定位性能影响较大，先取正值再聚合的方式具有较大优势 (45.54%/49.91%)。实验三中我们先将取正值的定位结果聚合，再进行全局池化得到各类别置信度 ( $m_c = GAP(\sum_{k=1}^K Relu(A_c^k))$ )，从表 4.1 中可以看到，相较于实验一分类错误率大幅提升 (27.44%/23.54%)，而定位性能仍然是先取正值再聚合的方式更具优势。

根据上述实验结果我们得出结论：对于弱监督定位任务，分歧性约束在多视角定位器的基础上能进一步增益网络的定位能力，相对于基线方法取得了 6.09% (44.98%/51.07%) 的定位性能提升。值得注意的是，分歧模块需要与  $\sum_{k=1}^K Relu(A_c^k)$  的定位聚合方式共同作用，才能得到更好的定位效果，而不同的聚合方式对不使用分歧约束的网络几乎没有影响。增加分歧之后，定位激活图中出现了更多视觉模式，这些定位激活结果中有正有负，在分类中负值意味着与本类相悖的激活区域，而在定位中正值负值都是前景信息。没有扩展特征空间之前，忽略负值带来的增益十分有限；在扩展特征空间之后，特征空间中增加了其他类的特征表征方式，甚至包括与本类相悖的区域，这部分抑制被忽略，这种前景激活与分类有悖但是对定位有益。

为了进一步分析分歧思想对于弱监督定位器的影响，我们通过实验对比探究了分歧思想与多模型集成学习的差别。我们进行了两组实验，第一组采用分歧学习思想，采取表 4.1 中实验二的实验设定，先聚合分歧定位结果再进行全局池化来得到分类置信度  $p_c = softmax(GAP(\sum_{k=1}^K A_c^k))$ ；第二组采用集成学习思想，先对每张定位结果进行全局池化来得到每个子分类器的置信度，再聚合各子分类器的结果  $p_c = \sum_{k=1}^K softmax(GAP(A_c^k))$ ，相当于在相同的特征空间中构建多个子定位器再将结果集成起来。两组实验中我们都采用了  $K = 8$  的设定，并添加了合适比例的分歧性约束。

对比两组实验结果我们发现：集成学习思想的弱监督定位器增益了分类性能 (23.11%/23.78%)，却与分歧思想的定位性能相差较大 (48.98%/45.54%)。

由此可见，我们的方法能够增益弱监督定位任务的原因在于分歧思想，而不是多个子模型的集成。此外，我们可视化了两种模型的定位结果，观察到集成思想的定位器激活区域更小，这可能是由于多分类器的限制较为强硬，模型为使分类目标函数收敛，将原本的激活区域分解，而没有进一步扩大激活空间。集成思想的多分类器在分类上优于分歧思想，得到了多组最优最小稳定区域，这些区域可以提高模型对数据的泛化能力；分歧思想的定位器是将有辨别能力的稳定区域扩大，虽然保留了有辨别能力的区域，却无法进一步提高分类性能。

表 4.2 分歧学习与集成学习的差别

Table 4.2 The difference between contrastive learning and integrated learning

算法思想	分类方式	Cls error (%)		定位聚合方式	Loc error (%)	
		Top-1	Top-5		Top-1	Top-5
分歧学习	$\text{softmax}(\text{GAP}(\sum_{k=1}^K A_c^k))$	23.78	7.56	$\sum_{k=1}^K A_c^k$	49.91	40.90
				$\sum_{k=1}^K \text{Relu}(A_c^k)$	45.54	35.24
集成学习	$\sum_{k=1}^K \text{softmax}(\text{GAP}(A_c^k))$	23.11	6.99	$\sum_{k=1}^K A_c^k$	55.51	48.27
				$\sum_{k=1}^K \text{Relu}(A_c^k)$	48.98	39.18

### 4.3.2 可视化结果

我们观察了基线方法与分歧定位方法的训练过程，如图 4.4 所示。其中奇数行为基线方法 SPG 的结果，偶数行为我们分歧定位方法的结果，从图中可以看出基线方法一旦“聚焦”在一个“最小稳定区域”，就不会再扩大激活区域来寻找对于分类有判别力的其他视觉模式，而增加了分歧性约束之后，网络会震荡地寻找其他激活的可能性，所以可以从较小的稳定区域扩大成大范围的激活，从而更好地定位目标。

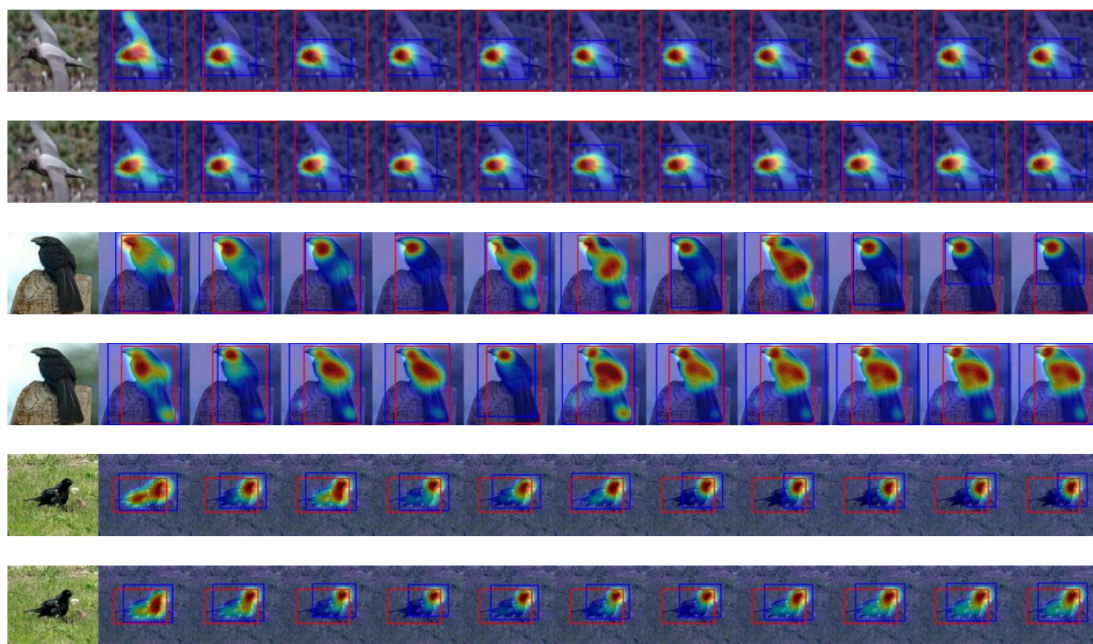


图 4.4 分歧方法对训练过程的影响

Figure 4.4 The effect of divergent methods on training process

图 4.5 中我们可视化了模型中各分歧定位激活图的激活结果与定位聚合结果。前几列中红色圈出的区域为该激活图的前景区域，我们可以观察到不同的分歧性激活图分别定位到了目标的不同区域，而当多个结果聚合之后可以很好地定位到目标的全身，该结果与我们的研究动机一致，验证了分歧思想对弱监督定位的有效性。

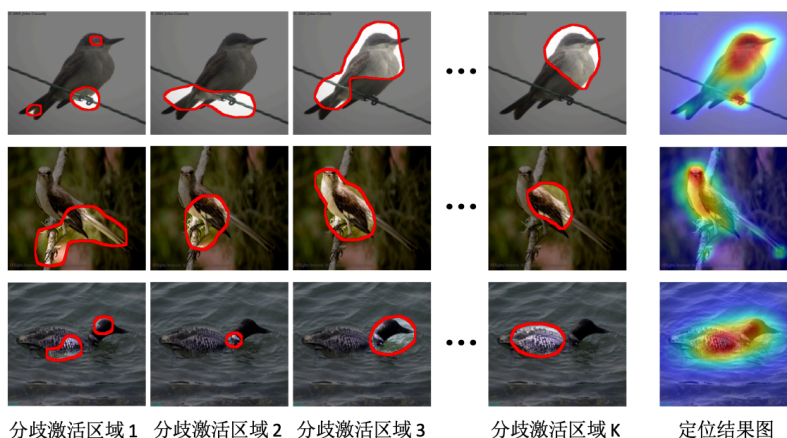


图 4.5 分歧定位结果可视化

Figure 4.5 Visualization of divergent localization results

#### 4.4 本章小结

本章中我们以分歧思想为指导,改善了神经网络训练过程中的特征紧凑性问题和并减少了视觉模式的压缩现象,增强了弱监督定位器的定位能力。我们将定位激活图扩展为多张并对其添加分歧性约束,在网络训练过程中同时收敛分类和分歧约束两种目标函数,不断寻找并扩展更多的视觉模式。我们可视化了训练过程中的定位激活结果和分歧激活定位图,验证了模型与研究思想的一致性。最后,我们通过实验验证了分歧思想对于弱监督定位的有效性,在基线方法的基础上进一步提高了弱监督定位性能。

## 第5章 基于分歧与协同的弱监督目标定位

### 5.1 基于分歧与协同的弱监督目标定位

本研究旨在构建基于分歧与协同思想的弱监督定位器。前文中的论述与实验分别验证了协同思想与分歧思想对于弱监督定位任务的有效性：协同思想认为弱监督目标定位以类别标签为唯一标注信息，需要挖掘出更多的可利用信息来辅助定位，用多源信息的融合来挖掘更多视觉模式。我们探索了类别协同和训练过程协同两种弱监督定位器，他们分别为目标定位提供了额外的类别层级信息和训练过程信息。我们发现构建合理的类别金字塔可以引入被精细分类忽略的视觉模式，通过类别合并可以得到更丰富的物体表征特征，在定位激活图中激活该部分特征可以补全被定位器忽略的目标区域；在弱监督定位器的训练过程中，神经网络不停地捕获相关特征并抑制无关视觉模式来压缩深度特征，最终在损失函数的驱动下保留了最具判别力的区域，即最具不变性的特征。为了增益弱监督定位，我们通过保留训练过程信息来保留该部分被压缩的特征，减少了这种压缩带来的信息损失。实验证明两种基于协同思想的弱监督定位器都能得到良好的定位性能，证明了协同思想对于弱监督定位的有效性。

与协同思想不同，分歧思想改善了神经网络训练过程中的特征紧凑性问题和对视觉模式的压缩现象，从原始类别标签中挖掘到了更多的隐含信息。在弱监督定位器训练的过程中，网络倾向于保留最具有不变性的特征来支撑分类，学习的唯一目标是捕获并表示与对象类别标签最相关的视觉模式，分歧思想通过学习分歧的定位器保留了更多视觉模式，各定位器分别关注目标的不同区域，将这些定位器的定位结果聚合可以得到更准确的弱监督定位结果。

本章中我们将两种思想融合，构建了基于分歧与协同思想的弱监督定位器。该定位器既综合了协同思想中多源信息所带来的丰富视觉模式，又包含了分歧思想对神经网络特征紧凑型 and 视觉模式压缩现象的改善。本章中我们提供了两种弱监督定位解决方案：基于分歧与类别协同的定位器和基于分歧与训练过程协同的定位器，本节中将首先介绍相关的网络结构与实验设置。



## 5.1.1 基于分歧与类别协同的弱监督目标定位

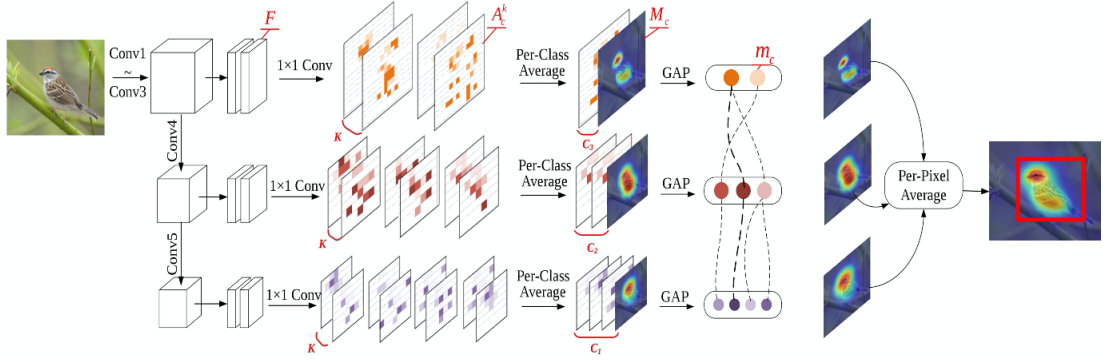


图 5.1 基于分歧与类别协同的弱监督目标定位

Figure 5.1 WSOL based on divergence and category pyramid collaboration

图 5.1 描述了基于分歧与类别协同的弱监督定位器的网络结构，我们在类别协同定位器的各层中分别添加了分歧模块。如前所述，对于输入样例  $I$ ，网络中 Conv3、Conv4、Conv5 分别连接对应层级的分类定位模块，对于层级  $i$  中的类别  $c_i$ ，我们将其定位激活图由 1 张扩展为  $K$  张，标记为  $A_{c_i}^k (k = 1 \dots K)$ ，这里我们依据第四章中的实验结果，设置  $K = 8$ 。将  $K$  张定位激活图聚合，即得到了层级  $i$  的定位激活结果  $M_c$ ，再经过全局池化层和 *softmax* 层即可得到样例  $I$  在该层级的分类置信度  $p_{c_i}$ ，网络通过计算  $p_{c_i}$  与类别标签的交叉熵损失来训练分类定位网络。

基于分歧与类别协同的弱监督定位方法的优化函数如下：

$$\operatorname{argmin}_{\alpha} \{ \mathcal{L}_C(\alpha) + \lambda \mathcal{L}_D(\alpha) \} \quad (5.1)$$

$$\mathcal{L}_C(\alpha) = \sum_{i=0}^n \mathcal{L}^i = \sum_{i=0}^n \frac{1}{C_i} \sum_c y_{c_i} \log(p_{c_i}) \quad (5.2)$$

$$\mathcal{L}_D(\alpha) = \sum_{1 \leq k < k' \leq n * K} \mathcal{S}(A_c^k, A_c^{k'}) \quad (5.3)$$

其中， $\mathcal{L}_C$  表示分类损失， $i$  表示类别层级的标号， $n$  表示类别层级个数，每个层级均使用交叉熵函数作为损失函数； $\mathcal{L}_D$  表示分歧损失， $A_c^k$  表示训练样例  $I$  在目标类  $c$  的第  $k$  张定位激活图， $\mathcal{S}(A_c^k, A_c^{k'})$  表示第  $k$  张和第  $k'$  张的余弦相似度。



## 5.1.2 基于分歧与训练过程协同的弱监督目标定位

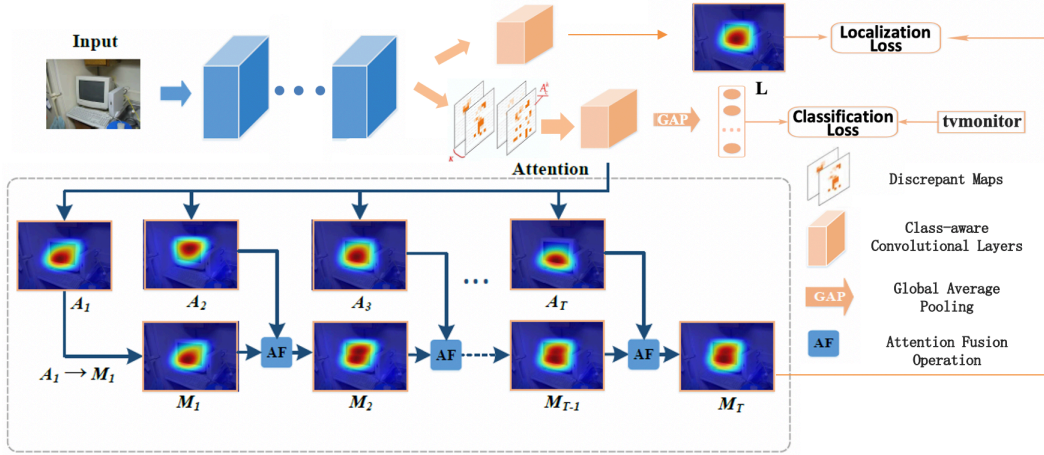


图 5.2 基于分歧与训练过程协同的弱监督定位

Figure 5.2 WSOL based on divergence and training process collaboration

图 5.2 描述了基于分歧与训练过程协同的弱监督定位器的网络结构，我们在训练过程协同定位器分类分支中添加了分歧模块。我们将输入样例  $I$  送入网络中的特征提取模块，得到深度特征  $F$ ，再将其分别送入分类分支和定位分支。我们在分类分支中添加了分歧模块，对于类别  $c$ ，首先将其定位激活图由 1 张扩展为  $K$  张，标记为  $A_c^k (k = 1 \dots K)$ ，这里同样采用  $K = 8$  的设定，在训练过程中同样用分歧损失函数  $\mathcal{S}(A_c^k, A_c^{k'})$  约束定位激活图之间的相似度。我们采用跟 3.2 节中相同的设定，用定位激活结果  $M_{t-1}$  来监督定位分支的定位。基于分歧与训练过程协同的弱监督定位方法的优化函数如下：

$$\operatorname{argmin}_{\alpha} \{ \mathcal{L}_C(\alpha) + \mathcal{L}_L(\alpha) + \lambda \mathcal{L}_D(\alpha) \} \quad (5.4)$$

$$\mathcal{L}_C(\alpha) = \frac{1}{C} \sum_c y_c \log(p_c) \quad (5.5)$$

$$\mathcal{L}_L(\alpha) = \sum_{j \in N^c} \text{MSE}^*(L_j^c, M_{t-1,j}^c) \quad (5.6)$$

$$\mathcal{L}_D(\alpha) = \sum_{1 \leq k < k' \leq K} \mathcal{S}(A_c^k, A_c^{k'}) \quad (5.7)$$

其中， $\mathcal{L}_C$  表示分类损失，这里同样采用交叉熵损失； $\mathcal{L}_L$  表示定位损失，这里采用带正负样本平衡的均方误差损失， $M_{t-1,j}^c$  表示上一轮迭代的定位聚合结果， $L_j^c$  表

示定位分支的定位结果； $\mathcal{L}_D$ 表示分歧损失， $A_c^k$ 表示训练样例在目标类 $c$ 的第 $k$ 张定位激活图， $\mathcal{S}(A_c^k, A_c^{k'})$ 表示类别 $c$ 第 $k$ 张和第 $k'$ 张激活图的余弦相似度。

### 5.1.3 比较分析

表 5.1 两种分歧与协同定位器的比较分析

**Table 5.1 Comparison of two weakly supervised locators based on divergence and collaboration**

方法	Cls error (%)		Loc error (%)	
	Top-1	Top-5	Top-1	Top-5
分歧	24.07	7.03	49.79	40.23
类别协同	24.13	<b>6.96</b>	50.71	43.48
训练过程协同	<b>22.75</b>	7.40	49.91	40.65
分歧+类别协同	24.60	7.70	<b>47.48</b>	<b>38.04</b>
分歧+训练过程协同	24.94	7.82	49.26	40.31

我们在 CUB-200-2011 上对上述两种结构进行了实验，实验结果如表 5.1。观察表中的实验结果，我们发现：基于分歧与类别协同的定位器定位结果明显优于两者独立，而基于分歧与训练过程协同的定位器定位结果仅略优于两者独立，这说明分歧与类别协同之间的互补性要更高。这可能是由于训练过程协同挖掘到的辅助信息仍是从分类框架中挖掘训练初期被抑制的特征，该部分特征与分歧损失约束所能挖掘到的信息具有重叠的部分，而类别协同的定位器中挖掘到的是隐藏在知识图谱中的类包含关系，这部分信息从未出现在神经网络的训练过程中。观察表格中的分类结果，我们发现：训练过程协同的 Top-1 分类错误率最低，这是由于我们采用了双分支分类定位网络，将分类分支和定位分支剥离开来，排除了定位分支对分类分支的干扰；同时，类别协同的 Top-5 分类错误率最低，这是由于多层级分类帮助网络捕获了更多相似类的特征，当判别条件放松到置信度最高的前 5 种时，这种优势被体现出来；而当分歧约束加入到两种协同定位网络中时，其分类性能均出现下降，这也与我们之前的观察：分歧约束会影响分类特征的学习相一致。在接下来的实验中，我们采用该方法的结果与前沿方法进行比较。

## 5.2 实验结果与分析

### 5.2.1 性能对比

为了验证分歧与协同的思想对于弱监督目标定位的有效性，我们在数据集 CUB-200-2011[21]和 ILSVRC[22]上与现有主流弱监督目标定位方法进行了对比。在两个数据集上，我们均评测了 Top-1 和 Top-5 分类性能错误率 (%) 与 Top-1 和 Top-5 定位性能错误率 (%)。值得注意的是，定位性能以分类准确性为基础，在图片分类正确且定位结果与真值交并比大于 0.5 时才被视为定位正确；而 Top-5 定位错误率相对于 Top-1 来说受分类性能影响较小，当 Top-5 定位性能提升较 Top-1 定位性能提升明显时，我们认为弱监督网络的分类影响了定位。

表 5.2 基于分歧与协同的弱监督定位方法与主流方法在 CUB 上的对比

Table 5.2 Comparison of WSOL method based on divergence and collaboration with existing methods on CUB dataset

方法	Cls error (%)		Loc error (%)	
	Top-1	Top-5	Top-1	Top-5
GoogLeNet-CAM[1]	<b>26.2</b>	<b>8.5</b>	58.94	49.34
GoogLeNet-SPG[15]	-	-	53.36	42.28
GoogLeNet-ours	28.8	9.4	<b>50.55</b>	<b>39.54</b>
VGGNet-CAM[1]	<b>23.4</b>	<b>7.5</b>	55.85	47.84
VGGNet-ACoL[7]	28.1	-	54.08	43.49
VGGNet-SPG[15]	24.5	7.9	51.07	42.15
VGGNet-ours	24.6	7.7	<b>47.48</b>	<b>38.04</b>

我们在 GoogLeNet 和 VGGNet 两种基网上部署了基于分歧与协同的弱监督定位方法，并在 CUB-200-2011 数据集上与主流方法进行了对比。CUB-200-2011 是鸟类精细分类的数据集，包含 5994 张训练图片和 5794 张测试图片，实验性能如表 5.2 所示。在分类性能上，弱监督定位方法中错误率最低的是基线方法 CAM[1]，其他在定位性能上虽然都有提升但是一定程度上损害了分类性能；在定位性能上，我们的分歧与协同方法在两种基网上均达到了最优的定位性能，对

比之前的前沿方法 SPG 分别取得了错误率 2.81% (50.55%/53.36%) 和 3.59% (47.48%/51.07%) 的降低。为了验证本方法在先进基网上的优越性，我们在 ResNet-50 上部署了本方法，取得了 18.4% 的 Top-1 分类错误率和 38.9% 的 Top-1 定位错误率，相对于 VGGNet 的弱监督定位性能进一步提升了 8.6% (38.9%/47.5%)，可见我们的方法在先进基网上具有更强的弱监督定位能力。

表 5.3 基于分歧与协同的弱监督定位方法与现有方法在 ILSVRC 上的对比

Table 5.3 Comparison of WSOL method based on divergence and collaboration with existing methods on ILSVRC dataset

方法	Cls error (%)		Loc error (%)	
	Top-1	Top-5	Top-1	Top-5
VGGNet-Backprop[60]	-	-	61.12	51.46
VGGNet-CAM[1]	33.4	12.2	57.20	45.14
VGGNet-ACoL[7]	32.5	12.0	54.17	40.57
GoogLeNet-Backprop[60]	-	-	61.31	50.55
GoogLeNet-CAM[1]	35.0	13.2	56.40	43.00
GoogLeNet-HaS-32[4]	-	-	54.53	-
GoogLeNet-ACoL[7]	29.0	11.8	53.28	42.58
GoogLeNet-SPG[15]	-	-	<b>51.40</b>	<b>40.00</b>
GoogLeNet- ours	<b>27.5</b>	<b>8.6</b>	52.47	41.72

为验证方法在不同数据集的可扩展性，我们同样在大规模数据集 ILSVRC 上进行了实验，结果如表 5.3 所示。ILSVRC 包含 1000 个类，120 万训练图片和 5 万张测试图片。在 ILSVRC 上，相对于基线方法弱监督定位错误率明显下降 (3.93%)，取得了与前沿方法 SPG 可比的性能 (52.47%/51.40%)。值得注意的是，定位性能上我们的方法在 ILSVRC 数据集上没有达到 CUB 上的明显优越性；而分类性能上与 CUB 数据集略降的表现相反，取得了一些提升。对于这种差异，我们认为在大规模数据集上，类间差异小导致的多类共用视觉模式抑制较少，同时对比前文中的消融实验我们发现类别协同思想会影响分类性能，由此我们得

出两种数据集上产生这些差异的原因: 在大规模数据集上本方法中类别协同的部分可能并没有发挥出其优势。这也印证了类别协同思想的适用范围——精细分类数据集。

需要说明的是, 我们与主流方法的对比表格中并没有包含 2020 CVPR 发表的 PSOL[18]方法, 该方法将弱监督定位问题分为类不可知目标定位以及目标分类两部分, 通过伪框标注进行模型更新。由于该方法采用了新的弱监督定位框架, 需要先产生伪标注, 再通过目标检测的相关方法进行定位框的回归, 与主流弱监督框架中从定位激活图得到定位框不同, 我们并没有放在表格中进行比较。该方法采用了候选框预提取和框回归技术, 流程上与弱监督检测方法更加相似, 但是针对弱监督定位数据集的特点——单目标定位进行了改良。

### 5.2.2 统计及可视化分析

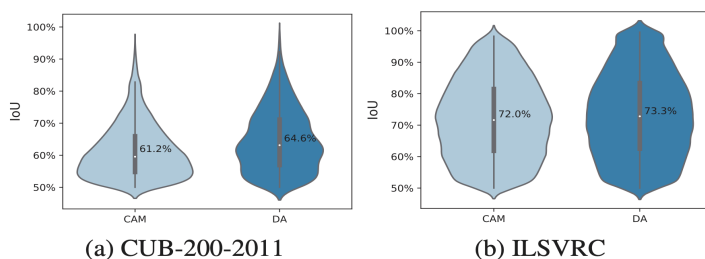


图 5.3 基于分歧与协同的弱监督定位方法 IoU 统计结果

Figure 5.3 IoU statistical results of WSOL method based on divergence and collaboration

进一步的, 我们对比分析了基于分歧与协同的弱监督定位方法对于定位 IoU 统计结果的影响。图 5.3 统计了 CUB-200-2011 和 ILSVRC 两测试数据集中定位正确样例的 IoU 统计分布。从该图可以看出, CUB 数据集中, 基线方法 CAM[1] 的正样例 IoU 主要集中在 50%~60%, 而我们的方法 IoU 整体上移; ILSVRC 数据集中, 本方法正样例 IoU 明显集中在 95%及以上的区域。两数据集上的结果都说明该方法不仅使定位错误率降低、正样例比例上升, 还使正样例定位结果与真值之间差异更小, 也就是提高了定位框的“质量”。

图 5.4 中给出了本方法与基线方法在两数据集上的定位结果。对比该结果我们发现, 基于分歧与协同方法的定位激活图会激活更多的目标区域, 而不是集中在最具判别力的区域, 这与我们通过寻找更多视觉模式来准确定位的出发点一致。

进一步对比激活结果，我们发现：本方法结果中的显著区域仍集中在基线方法的激活位置，这说明在利用分类器来学习目标定位的框架下，分类器仍以分类中最具判别力的区域作为判断的依据，我们的方法只是以弱监督目标定位为目的、在定位激活图上激活了更多区域，并部分影响分类器的学习。由此可见，分类器与定位器的学习存在天然的矛盾，现有的弱监督定位方法更多的是在分类器与定位器的学习间寻找一定的平衡，若希望得到好的分类结果则定位激活图会集中在少数视觉模式上；若希望得到均匀激活的定位结果则分类器的学习会受到影响。

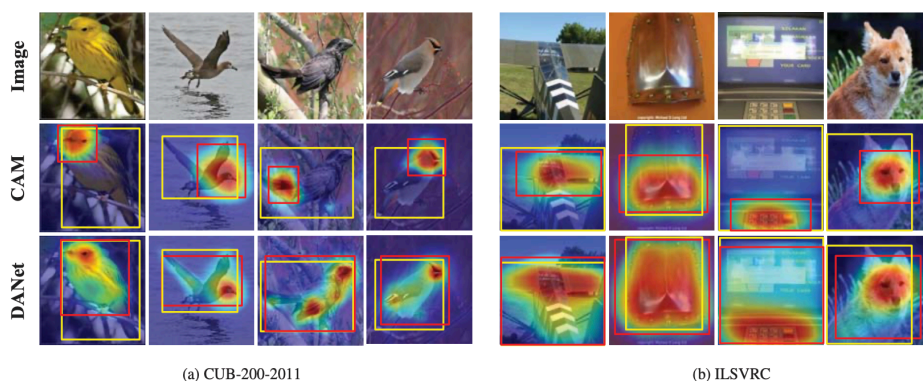


图 5.4 基于分歧与协同的弱监督定位方法在 CUB 和 ILSVRC 上的定位结果可视化

Figure 5.4 Location results of WSOL method based on divergence and collaboration on CUB and ILSVRC dataset

### 5.3 病理学定位应用

本节中，我们将基于分歧与协同的弱监督目标定位方法用于胸片 X 光数据集 ChestX-ray8[10]中，以验证该方法在医学诊断上的应用价值。

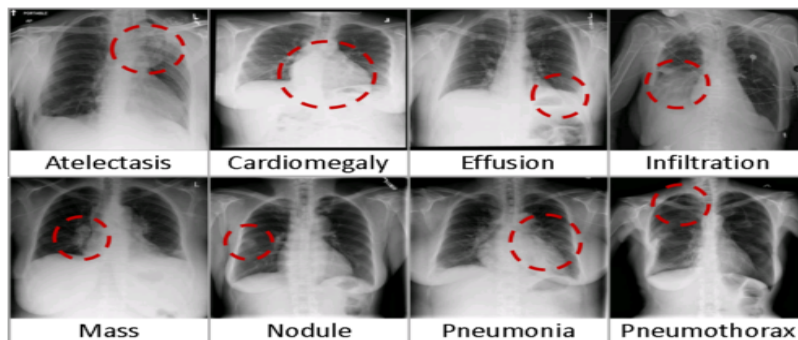


图 5.5 Chest X-ray8[10]中包含的八种常见胸部疾病影像

Figure 5.5 Eight common thoracic diseases observed in ChestX-rays

ChestX-ray8 是 2017 年发布的胸片数据集,其中包含数万名患者的胸部 X 光图像与 8 种常见的胸部病例分类标注,所有图像均带有类别标注信息。该数据集包括肺不张、积液、肺炎、气胸等八种胸部疾病,与前面实验所用数据集不同的是,同一张 X 光图像中可能出现多种肺部疾病。如图 5.5 所示为八种常见胸部疾病的 X 光图像,红色圆圈圈出来的是疾病的位置。医学设备生成的 X 光图像的大小为 3000\*2000 像素,该数据集中图像调整为 1024\*1024 的大小,并且请放射科医生标注了少量的疾病的位置信息作为评价疾病定位性能的依据。

表 5.4 分类结果 ROC 曲线的 AUC 值在 ChestX-ray8 数据集上的对比

Table 5.4 AUCs of ROC curves for multi-label classification on ChestX-ray8

	肺不张	心脏肥大	积液	浸润	弥撒	结核	肺炎	气胸
ChestX-ray8 [10]	0.69	<b>0.84</b>	<b>0.76</b>	<b>0.64</b>	0.64	<b>0.69</b>	0.69	0.82
本方法	<b>0.70</b>	0.83	0.75	0.63	<b>0.68</b>	<b>0.69</b>	<b>0.70</b>	<b>0.83</b>

表 5.5 病理学定位准确性在 ChestX-ray8 数据集上的对比

Table 5.5 Pathology localization accuracy on ChestX-ray8

	肺不张	心脏肥大	积液	浸润	弥撒	结核	肺炎	气胸	综合
ChestX-ray8[10]	0.22	<b>0.98</b>	0.31	0.30	0.14	0.01	<b>0.42</b>	0.16	0.32
本方法	<b>0.25</b>	0.93	<b>0.33</b>	<b>0.41</b>	<b>0.19</b>	<b>0.05</b>	0.40	<b>0.17</b>	<b>0.34</b>

我们将基于分歧与协同的弱监督目标定位方法部署到 ChestX-ray8 数据集上,以验证其在医学图像诊断中的有效性。我们在 ResNet-50 上部署了该算法,将胸片图像裁剪为 512\*512 大小输入网络,并采用带权重的二分类交叉熵函数作为分类的损失函数,以适应数据集类不平衡特性。中给出了 ChestX-ray8[10]中算法与本算法分类性能的对比,表 5.5 中给出了病理学定位性能的对比。其中分类性能采用多分类任务中常用的 AUC 值作为评价指标,病理学定位性能采用数据集论文中提出的 IoBB>0.5 的 X 光图像占比作为指标,指的是病灶区域在定位结果中的占比。从分类和定位结果中我们可以看到,基于分歧与协同的定位方法在 8 种胸腔疾病的分类上与原方法结果相差不大,而定位结果中多数超过了原方法的性能,整体提高了 2%,这说明本方法在病理学定位中具有一定的应用价值。



图 5.6 中给出了部分病理学定位结果图，图中蓝框表示病理学定位真值，红框表示本方法的病理学定位结果，图中可以观察到本方法较好地定位到了胸片 X 光图像中的病灶区域。与弱监督目标定位不同的是，病灶区域没有固定的位置，从图中可以观察到相同疾病的病灶位置可能是完全不同的，这说明病灶区域的定位更依赖于 X 光图片中的纹理信息。本方法在病理学定位中的有效也说明了“捕捉更多的视觉模式”这一思想对更加依赖于纹理信息的病理学定位的同样适用。

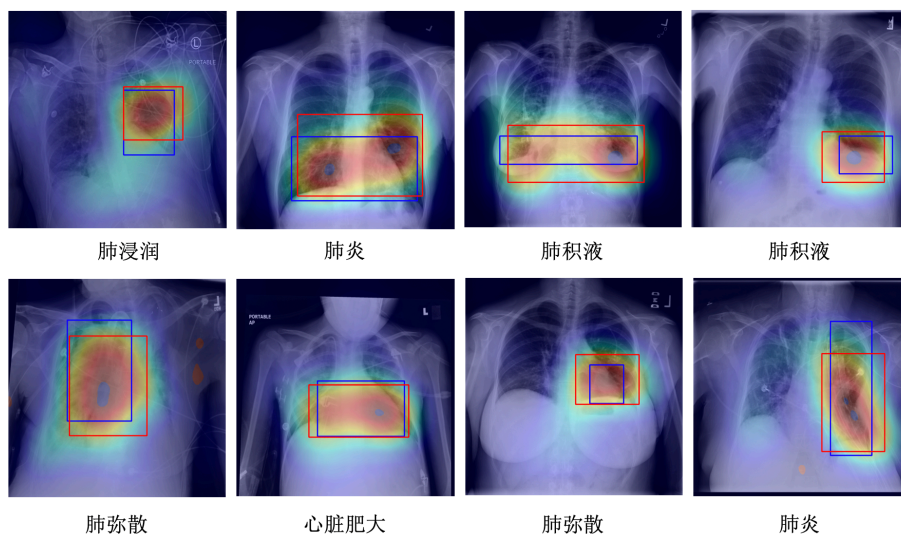


图 5.6 病理学定位结果图

Figure 5.6 Pathology localization results

#### 5.4 本章小结

本章中我们将第三章中的协同思想与第四章中的分歧思想结合起来，构成了基于分歧与协同的弱监督定位器，并分析了两种思想对于弱监督目标定位的互补性，实验表明两者可以共同作用来优化弱监督目标定位。我们将该方法部署在三种网络结构上，验证了该方法对于不同网络结构的兼容性；同时，我们将实验结果与前沿方法进行比较，达到了前沿性能。通过对该方法定位结果的统计分析可视化分析，我们进一步验证了该方法的优越性。进一步的，我们在医学数据集 ChestX-ray8 中进行实验，对比基线方法，提高了 2% 的定位性能，验证了该方法在病理学定位上的应用价值。



## 第6章 总结与展望

弱监督目标定位作为目标识别及定位领域的一个分支,旨在利用不完整的目标标注信息——目标类别标号来学习物体的识别。标注量的减少及标注信息的易获取使其可以扩展到大规模数据集上进行训练,甚至利用网上的海量数据来辅助训练,得到泛化性能更好的模型。除此之外,由于弱监督目标定位从响应图中获得目标位置,其基础研究也与神经网络可视化、可解释神经网络密切相关。在应用方面,弱监督目标定位可以应用在医学图像、安检图像等典型场景中,医学图像的病理学诊断中标注成本较高,安检场景中数据量较大造成标注困难。除了以上场景,也可以利用其标注成本低的特点使用网络数据来学习目标定位。

本研究以深度网络模型为基础,探究弱监督定位问题的解决方案。周博磊的工作[1]表明,分类网络通过关注目标的视觉模式来识别物体,这使得我们可以在分类网络上构建弱监督定位器。由于分类网络仅关注物体最具判别力的区域,导致类激活图只能激活物体的部分区域,从而影响了弱监督定位性能。我们的方法构建了基于分歧和协同思想的弱监督定位器,通过寻找更多的视觉模式,达到更佳的弱监督定位效果。

弱监督目标定位以图片类别为唯一的标注信息,协同思想希望挖掘出更多的可利用信息来辅助弱监督定位器的学习。我们探究了基于类别协同和训练过程协同两种弱监督目标定位器。其中类别协同旨在挖掘类别标号中的隐藏信息,寻找并保留更多与物体本身相关的视觉模式,通过类别合并得到更丰富的物体表征特征,在定位激活图中激活该部分特征来补全被定位器忽略的目标区域;训练过程协同观察到神经网络通过抑制类别无关的视觉模式来压缩深度特征,旨在挖掘并保留训练过程中被抑制的早期目标特征表示,减少这种压缩。与协同思想通过多源信息融合来增益视觉模式不同,分歧思想改善了神经网络的特征紧凑性问题和对于视觉模式的压缩现象,正是这种紧凑性和压缩现象带来了稀疏的定位激活图。分歧思想构建了多个定位器,他们之间是分歧且互补的,而这种互补性为弱监督定位器带来了充分的视觉模式辅助定位。

我们将改善学习器的分歧思想与融合多源信息的协同思想结合起来,构建了基于分歧与协同的弱监督定位器,并分析了两种思想对于弱监督目标定位的互补性,实验表明两者可以共同作用优化弱监督目标定位任务。我们将该方法部署在三种网络结构上,验证了该方法对于不同网络结构的兼容性;同时,我们将实验结果与前沿方法进行比较,达到了前沿性能。通过对该方法定位结果的统计分析与可视化分析,我们进一步验证了该方法的优越性。进一步的,我们在医学数据集 ChestX-ray8 中进行实验,对比基线方法,提高了 2%的定位性能,验证了该方法在病理学定位上的应用价值。

本研究在弱监督目标定位方向上取得了一定的成果,但是弱监督定位方向未来仍有一些需要探索的问题。首先是在方法上,我们通过比较分类与定位性能并观察定位结果发现:分类目标函数与定位任务存在一些分歧,现有方法更多的是在寻找两者之间的平衡,在不过多损害分类性能的基础上提高定位效果,并没有完全解决特征紧凑性和定位任务的天然矛盾。其次是在应用上,弱监督目标定位目前在医疗和安检领域有了一些应用,发挥了其标注简单的优势,未来仍可以探索更多的应用场景,包括利用大量互联网数据进行定位模块的预训练、将轻量的定位模块部署到手机端等;除此之外,该任务还可以发挥其数据量大、训练快的优势,为其他计算机视觉任务提供有效的方法与理论支持。

## 参考文献

- [1] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.
- [2] Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle[C]//2015 IEEE Information Theory Workshop (ITW). IEEE, 2015: 1-5.
- [3] Zhu Y, Zhou Y, Ye Q, et al. Soft proposal networks for weakly supervised object localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1841-1850.
- [4] Singh K K, Lee Y J. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization[C]//2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 3544-3553.
- [5] Wei Y, Xiao H, Shi H, et al. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7268-7277.
- [6] Kim D, Cho D, Yoo D, et al. Two-phase learning for weakly supervised object localization[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3534-3543.
- [7] Zhang X, Wei Y, Feng J, et al. Adversarial complementary learning for weakly supervised object localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1325-1334.
- [8] Durand T, Mordan T, Thome N, et al. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 642-651.
- [9] Wang L, Lu H, Wang Y, et al. Learning to detect salient objects with image-level supervision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 136-145.

- [10] Wang X, Peng Y, Lu L, et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2097-2106.
- [11] Li Z, Wang C, Han M, et al. Thoracic disease identification and localization with limited supervision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8290-8299.
- [12] Miao C, Xie L, Wan F, et al. SIXray: A Large-scale Security Inspection X-ray Benchmark for Prohibited Item Discovery in Overlapping Images[J]. arXiv preprint arXiv:1901.00303, 2019.
- [13] Choe J, Shim H. Attention-based dropout layer for weakly supervised object localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2219-2228.
- [14] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[C]// International Conference on Learning Representations. 2016.
- [15] Zhang X, Wei Y, Kang G, et al. Self-produced guidance for weakly-supervised object localization[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 597-613.
- [16] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.
- [17] Yang S, Kim Y, Kim Y, et al. Combinational Class Activation Maps for Weakly Supervised Object Localization[C]//The IEEE Winter Conference on Applications of Computer Vision. 2020: 2941-2949.
- [18] Zhang C L, Cao Y H, Wu J. Rethinking the Route Towards Weakly Supervised Object Localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020.
- [19] Uijlings J., Sande K., Gevers T., Smeulders A. Selective Search for Object Recognition [J]. International Journal of Computer Vision. 2013, 104(2): 154-171.
- [20] S. Ren, K. He, R. B. Girshick, Jian Sun: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. CoRR abs/1506.01497, 2015.

- 
- [21] Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[J]. 2011.
- [22] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. *International journal of computer vision*, 2015, 115(3): 211-252.
- [23] Deselaers T, Alexe B, Ferrari V. Weakly supervised localization and learning with generic knowledge[J]. *International journal of computer vision*, 2012, 100(3): 275-293.
- [24] Girshick R. Fast r-cnn[C]//*Proceedings of the IEEE international conference on computer vision*. 2015: 1440-1448.
- [25] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: 779-788.
- [26] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges[C]//*European conference on computer vision*. Springer, Cham, 2014: 391-405.
- [27] Matas J, Chum O, Urban M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. *Image and vision computing*, 2004, 22(10): 761-767.
- [28] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation[J]. *International journal of computer vision*, 2004, 59(2): 167-181.
- [29] Neubeck A, Van Gool L. Efficient non-maximum suppression[C]//*18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, 2006, 3: 850-855.
- [30] Dietterich T G, Lathrop R H, Lozano-Pérez T. Solving the multiple instance problem with axis-parallel rectangles[J]. *Artificial intelligence*, 1997, 89(1-2): 31-71.
- [31] Oquab M, Bottou L, Laptev I, et al. Is object localization for free-weakly-supervised learning with convolutional neural networks[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015: 685-694.
- [32] Bilen H, Vedaldi A. Weakly supervised deep detection networks[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016: 2846-2854.
- [33] Kantorov V, Oquab M, Cho M, et al. Contextlocnet: Context-aware deep network models for weakly supervised localization[C]//*European Conference on Computer Vision*. Springer, Cham, 2016: 350-365.
- [34] Wang C, Ren W, Huang K, et al. Weakly supervised object localization with latent category learning[C]//*European Conference on Computer Vision*. Springer, Cham, 2014: 431-445.

- [35] Bilen H, Pedersoli M, Tuytelaars T. Weakly supervised object detection with convex clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1081-1089.
- [36] Song H O, Girshick R, Jegelka S, et al. On learning to localize objects with minimal supervision[J]. arXiv preprint arXiv:1403.1024, 2014.
- [37] Dong X, Zheng L, Ma F, et al. Few-example object detection with model communication[J]. IEEE transactions on pattern analysis and machine intelligence, 2018.
- [38] Li G, Yu Y. Deep contrast learning for salient object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 478-487.
- [39] Liu N, Han J. Dhsnet: Deep hierarchical saliency network for salient object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 678-686.
- [40] Wang L, Wang L, Lu H, et al. Saliency detection with recurrent fully convolutional networks[C]//European conference on computer vision. Springer, Cham, 2016: 825-841.
- [41] Li G, Xie Y, Lin L. Weakly supervised salient object detection using image labels[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- [42] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(1): 142-158.
- [43] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. Springer, Cham, 2014: 818-833.
- [44] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 5188-5196.
- [45] Zhou B, Sun Y, Bau D, et al. Interpretable basis decomposition for visual explanation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 119-134.
- [46] Chen L, Chen J, Hajimirsadeghi H, et al. Adapting Grad-CAM for Embedding Networks[C]//The IEEE Winter Conference on Applications of Computer Vision. 2020: 2794-2803.

- 
- [47] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the eleventh annual conference on Computational learning theory. ACM, 1998: 92-100.
- [48] Chen S, Bortsova G, Juárez A G U, et al. Multi-Task Attention-Based Semi-Supervised Learning for Medical Image Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019: 457-465.
- [49] Qiao S, Shen W, Zhang Z, et al. Deep co-training for semi-supervised image recognition[C]//Proceedings of the european conference on computer vision (eccv). 2018: 135-152.
- [50] Han B, Yao Q, Yu X, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels[C]//Advances in neural information processing systems. 2018: 8527-8537.
- [51] Chai Y, Lempitsky V, Zisserman A. Bicos: A bi-level co-segmentation method for image classification[C]//2011 International Conference on Computer Vision. IEEE, 2011: 2579-2586.
- [52] Gong Y, Ke Q, Isard M, et al. A multi-view embedding space for modeling internet images, tags, and their semantics[J]. International journal of computer vision, 2014, 106(2): 210-233.
- [53] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge & Data Engineering, 2005 (11): 1529-1541.
- [54] Qiao S, Shen W, Zhang Z, et al. Deep co-training for semi-supervised image recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 135-152.
- [55] Xia R, Wang C, Dai X Y, et al. Co-training for semi-supervised sentiment classification based on dual-view bags-of-words representation[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015, 1: 1054-1063.
- [56] Han Y, ROY S, Petersson L, et al. Learning from Noisy Labels via Discrepant Collaborative Training[C]//The IEEE Winter Conference on Applications of Computer Vision. 2020: 3169-3178.
- [57] Miller G A, Beckwith R, Fellbaum C, et al. Introduction to WordNet: An on-line lexical database[J]. International journal of lexicography, 1990, 3(4): 235-244.
- [58] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European

- conference on computer vision. Springer, Cham, 2016: 21-37.
- [59] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
- [60] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. arXiv preprint arXiv:1312.6034, 2013.
- [61] Xu Z, Tao D, Huang S, et al. Friend or foe: Fine-grained categorization with weak supervision[J]. IEEE Transactions on Image Processing, 2016, 26(1): 135-146.
- [62] He X, Peng Y. Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification[C]//Thirty-first AAAI conference on artificial intelligence. 2017.
- [63] Wei Y, Xiao H, Shi H, et al. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7268-7277.
- [64] Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[J]. arXiv preprint arXiv:1611.03530, 2016.



## 致 谢

在这个学生时代最漫长的寒假，我还未来的及好好告别就走到了毕业的节点。三年的研究生生活转瞬即逝，回首三年来在国科大科研与生活的点滴，我在学习上和思想上都受益匪浅，在此我要感谢一路以来帮助与支持我的老师、同学与家人们。

2017年秋天，我进入了模式识别与智能系统开发实验室，三年来，受到了导师叶齐祥教授的无私指导与帮助，叶老师宽仁的性格，实验室优越的条件，自由轻松的科研氛围，实验室同学科研热情与耐心帮助，为我研究生阶段的学习注入了活力与热忱。在此，我要感谢叶老师对于本课题工作的指导与付出，三年来叶老师对于科研选题与实验思路时常与我讨论，给予了我专业且耐心的指导，对于我遇到的瓶颈与困难总是不吝鼓励与帮助，他对科研的严谨敬业以及对学生的耐心负责也时时感染着我，这将使我在未来的工作学习中受益终生。同时也要感谢焦建彬教授在科研上对我的指导与鼓励，在生活上对我的帮助和关心。

感谢万方师兄、刘畅师兄以及张天亮师兄对于我科研学习上的细心指导，感谢学霸与小松在程序调试上的耐心帮助，感谢丁瑶师姐与雨婷的陪伴鼓励，还有实验室其他师兄师姐师弟师妹们，你们构成了暖心的 SDL 大家庭，让温馨与欢乐陪伴了我的研究生生涯。

感谢我的家人及男朋友，他们的每一次开导与鼓励都让我倍感温暖与宽慰，他们的支持是我永远的后盾，让我不论何时、不论遇到什么困难都能勇敢前行。

最后，感谢参与论文开题和中期的各位老师，他们用丰富的经验和渊博的知识把握论文方向以及指点整个研究工作。

虽然我没有继续科研这条道路，但是对于这条路上继续前行的老师们及师兄师姐们我深感羡慕与崇敬，愿你们在未来的科研中披荆斩棘，斩获丰硕的成果。

薛昊岚

2020年5月



## 作者简历及攻读学位期间发表的学术论文与研究成果

### 作者简历:

2013年09月——2017年07月,在浙江大学,信息与电子工程学院获得学士学位。

2017年09月——2020年07月,在中国科学院大学,电子电气与通信工程学院,攻读硕士学位。

### 获奖情况:

中国科学院大学三好学生(2020年)

浙江大学优秀毕业论文(2017年)

浙江大学三好学生(2014年、2015年)

浙江大学学业奖学金(2014年、2015年)

### 已发表的学术论文:

1. H. Xue, C. Liu, F. Wan, J. Jiao, X. Ji and Q. Ye, "DANet: Divergent Activation for Weakly Supervised Object Localization," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 6588-6597.

### 已经申请的专利:

1. 万方, 薛昊岚, 刘畅, 付梦莹, 叶齐祥, 韩振军, 焦建彬, 一种基于分歧学习的弱监督定位方法(申请号: 2019109425654)

