中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

弱监督视觉目标检测

作者姓名：万方

指导教师：叶齐祥　教授

中国科学院大学

学位类别：工学博士

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

2019 年 6 月

Weakly Supervised Visual Object Detection

A dissertation submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Signal and Information Processing

By

Fang Wan

Supervisor: Professor Qixiang Ye

School of Electronic, Electrical and Communication Engineering

June 2019

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：

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：　　导师签名：

日　　期：　　日　　期：

摘 要

目标检测旨在从图像中定位出待检测目标并辨识其类别，是计算机视觉领域中最基本和最具挑战性的问题之一。在深度学习时代，考虑到模型包含大量待学习参数，现有目标检测算法在训练过程中需要大量的精确标注样本。当面对形态复杂、种类繁多的视觉场景时，标注训练样本需要耗费大量的人工劳动。

弱监督的目标检测方法通过用图像级标注代替样本框的标注，从而显著降低人工标注的工作量。在弱监督设定下，互联网上存在的大量带有图像级标号的样本可以用来学习目标检测模型，实现“网络监督”的自主学习建模，显著降低训练数据标注的成本。

典型的弱监督目标检测框架有三个部分组成：候选框生成、弱监督目标定位和检测器学习。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，而检测器学习最终保证目标检测的效果。

本文首先通过实验观察到传统弱监督目标定位学习中两个问题：定位过程具有很强随机性和定位容易受到目标部件（Part）干扰；进而通过理论分析确定了这些实验现象背后的原理在于模型的非凸性。基于实验观察与理论分析，提出了最小熵隐变量模型、渐进示例学习方法和弱监督 X 光图像违禁品定位算法，从模型构建、模型优化、方法应用三方面系统研究了弱监督视觉目标建模的科学问题。

本文主要的贡献总结如下：

1）提出了最小熵隐变量模型。该模型利用最小熵对训练过程中目标定位的随机性进行建模，通过降低目标定位熵，降低了训练过程中目标定位随机性问题，显著提升了模型的定位精度。

2）提出了渐进多示例学习模型。隐变量模型的非凸性使得传统的弱监督目标检测算法容易陷入局部最优，从而错误的定位到背景或者目标局部。本文将渐进优化的方法引入到多示例学习的框架，创造性地提出了渐进多示例学习方法解决非凸优化这一科学难题。

3）提出了弱监督 X 光图像违禁品定位算法框架。针对实际的 X 光安检场合中违禁品正反例样本比例失衡问题，提出了弱监督 X 光图像违禁品类平衡分类算法，结合建模和优化将弱监督问题研究推进至实际应用场景。

本文的研究成果表明，通过降低定位随机性、缓解隐变量学习的非凸性能够解决弱监督学习的本质问题，显著提升弱监督目标检测的性能。本论文涉及的算法框架、渐进优化方法、弱监督目标定位算法开拓了视觉目标检测的新方向，为深度学习框架中的非凸优化问题、不完全标注下的模型估计与样本标注问题提供了新思路。展望未来，本论文的研究成果为目标检测模型的自主进化提供了坚实的理论基础。

关键词：弱监督学习，目标检测，隐变量学习，多示例学习，连续优化

Abstract

Object detection is aiming at localizing the objects in images and classifying them. It is one of the most fundamental and challenging tasks in computer vision. In the deep learning era, the models always need to learn millions of parameters, which requires a huge amount of accurate annotations, especially for the object detection task. The situation will be worse when the scene in images is complicated.

The weakly supervised object detection requires only the image-level annotation indicating the existence of a class of objects in images. This remarkably eases the annotation work. Under the weakly supervised settings, the large-scale data in the internet could be used to train object detectors, which can be referred to as webly supervised. This can further decrease the annotation cost.

The weakly supervised object detection task consists of object proposal generation, weakly supervised object localization and detection estimation. In this thesis, we first point out the problems of weakly supervised object detection, i.e., localization randomness and falsely localized object part. To solve these problems, we then propose min-entropy latent model, continuation multiple instance learning and weakly supervised X-ray prohibited item discovery framework, from the aspects of modeling, optimization and application.

The main contributions of this thesis are as follows:

1) A min-entropy latent model has been proposed. This model focuses on the localization randomness problem by modeling it with entropy models, which is able to reduce the randomness and thus significantly improve the performance.

2) A continuation multiple instance learning method. The latent models for weakly supervised object detection are always non-convex, which is prone to fall into local minima and then falsely localize the background/object part. We then novelly introduce the continuation optimization to solve the non-convex problem.

3) A weakly supervised X-ray prohibited item discovery framework. To solve

the imbalance problem in the X-ray security inspection, we then propose the balance classification framework to model and optimize the problems for practical application.

The experiments show that the proposed min-entropy latent model and continuation multiple instance learning have reduced the localization randomness, alleviated the non-convex problem and significantly improved the detection results. The methods including modeling, optimization, and application framework have extended the horizon of weakly supervised learning and weakly supervised object detection with incomplete annotations. It also provided fresh insights for relevant computer vision areas.

Key Words：Weakly supervised learning, Object detection, Latent model, Multiple instance learing, Continuation optimization

目 录

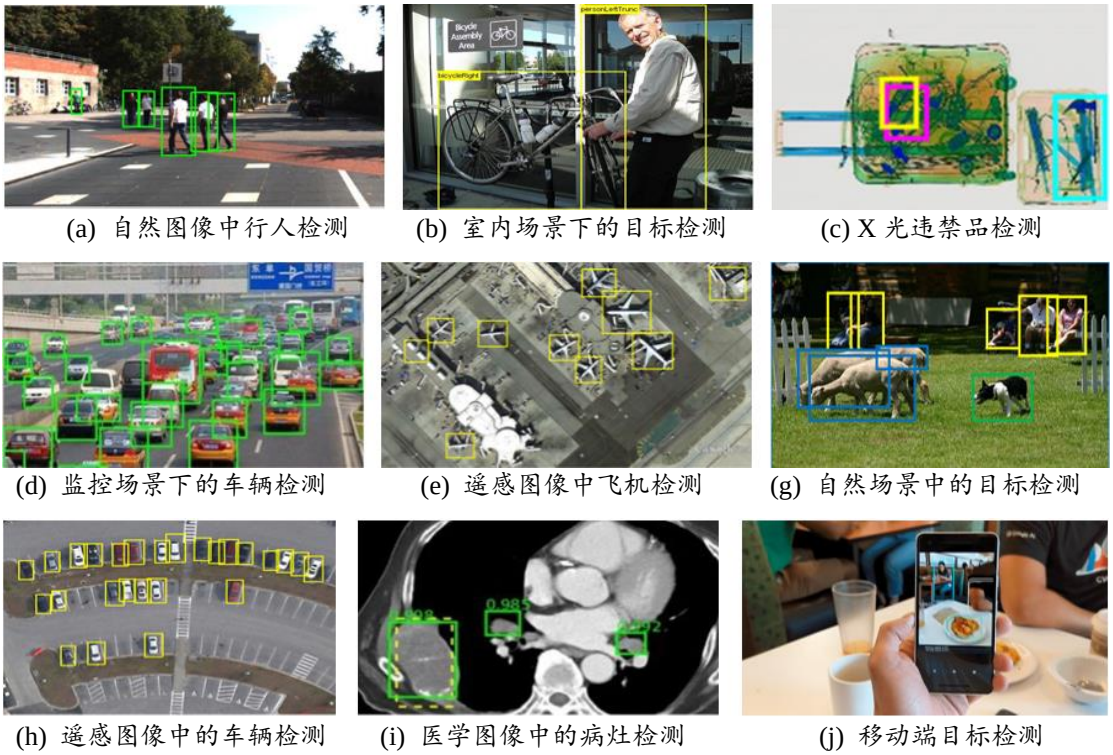
图目录

表目录

第1章 绪论

1.1 研究背景与意义

视觉目标检测是计算机视觉里最重要和最基础的任务之一，是众多高层视觉任务，如活动或事件检测、行为理解和分析、场景识别和解析等的重要前提，对推动计算机视觉和人工智能的发展有着非常重要的意义[1][2][3][4] [5]。视觉目标检测任务不仅要判断给定的图像中是否存在感兴趣的目标、识别该目标所属的类别，还需要以框的形式确定每个感兴趣的目标的具体位置。



(a) 自然图像中行人检测 (b) 室内场景下的目标检测 (c) X 光违禁品检测

(d) 监控场景下的车辆检测 (e) 遥感图像中飞机检测 (g) 自然场景中的目标检测

(h) 遥感图像中的车辆检测 (i) 医学图像中的病灶检测 (j) 移动端目标检测

图 1.1 视觉目标检测应用示例

Figure 1.1 Illustration of applications of visual object detection

视觉目标检测的应用场景非常广泛，包括智能辅助驾驶、智能视频监控、机器人导航、工业检测、航空航天等，如图 1.1 所示。目标检测在信息、控制与智能系统中运用非常广泛，是构成智能监控视频、视觉目标检索和机器人导航系统等的核心技术。传统的目标检测任务分为两类：目标示例检测（Object Instance Detection）和目标类别检测（Object Class Detection）[6]。目标示例检测

任务要求识别并定位输入图像中已知的特定物体，例如检测图像中的某一只特定的猫。在该任务中，测试集中的目标和训练集中的目标是同一个目标的不同形态和环境下的成像，本质上是将测试图像中的目标与训练集合中的目标进行匹配，检测模型需要成像条件如光照、角度等的变化等鲁邦；目标类别检测任务则是更关注于检测目标的类别。该任务要求识别并定位感兴趣的目标，与目标示例检测的主要区别在于训练样本集中的目标和待检测的目标样本不是同一个特定的目标，而是属于同一类别。相比之下，后者更具有挑战性。其原因在于同一类别的目标在语义上虽然很接近，但实际的物理特性如颜色、纹理、形状可能会有非常大的差异。本文的研究内容属于目标类别检测这一任务，后文中均将该任务简称为目标检测。

在过去的几十年中，大批的研究人员投入到了目标检测的研究中，尝试并提出了多种有效的目标检测方法。这些方法主要基于机器学习的方法，学习并建立目标检测框架。其中，应用最为广泛的机器学习算法是监督学习方法。监督学习下的目标检测是指利用已知类别的训练样本集合，学习目标检测器，使得检测器能够准确分类和定位测试集中未知标号的目标样本。在监督学习过程中，算法往往依赖于人工给定的样本类别标注信息，监测任务中还需要给出目标的具体位置。在训练分类和检测模型之前，人们需要对图像数据集中所有的目标样本进行标注。为了能够涵盖多视角、多姿态、多形态的目标，以增强模型的学习效率和鲁棒性，监督学习目标建模过程中往往需要大量精确的人工标注信息，包括目标的类别和位置。这个过程往往十分复杂、耗时耗力[13]。相比之下，对于其他视觉任务如图像分类、场景分类，标注者只需要对图像中目标的类别进行标注。

近年来，人工智能中的深度学习技术（例如，深度卷积神经网络、长短时记忆神经网络等）在经典计算机视觉任务（例如，图像分类、物体检测等）上面取得了巨大的成功，极大地促进了计算机视觉的发展。但是，现有的深度模型对训练数据有很高的要求，为保证模型的性能，需要大量人工标注好的样本来训练网络参数。同时，随着社会的持续发展，海量图像视频数据、城市安全监控数据（如 X-光安检图像），一直在爆发性地增长。而这里面绝大部分的数

据都是没有标注的，对其进行人工标注的代价非常高昂。而且，人工标注往往会因为标注者的疲劳和理解不同等原因，产生标注错误。如何从理论和应用的角度出发，设计新的学习模型，实现以下目标：（1）从海量未标注的数据中挖掘有价值的视觉目标信息，（2）实现（极）弱监督下的特征与模型学习，（3）大幅度地提升模型的通用性和适应性，成为了学术界和工业界的研究热点。

另一方面，人类能够在非常少量的样本或者先验信息（弱监督）驱动下获得物体认知能力（图 1.2）。人类视觉认知机制能够通过“自主学习”不断增强认知能力。机器学习与计算机视觉能否仿照人类认知机制在少量监督或者先验信息的辅助下进行学习？本文涉及的弱监督目标建模是指：只需要给出图像中是否包含待检测目标（图像级标注）。

图 1.2 认知科学领域的成果表明：人类视觉认知呈现弱监督与自学习特性

Figure 1.2 The results of the recongnition research show that human recognition process are weakly- and self-supervised

弱监督目标检测相比于监督目标检测而言，不需要对所有图像中的目标所在的位置进行精确标注[14][15][16][17][18][19]。为了减少标注的工作量，弱监督学习中往往需要让标注的量减少或是让标注变得更容易。比较典型的减少标注的标注方法有以下两种：第一种方式是采用半监督[13]方法，即只对一少部分样本进行标注，而不对其余的样本进行标注；第二种方法则是对所有的样本进行弱监督标注。所谓弱监督标注也就是标注信息不完全，在标注的过程中只给出了标注样本的部分信息，如类别信息，另一部分信息如位置信息则需要通过其他方式（先验或学习）等途径获取。对于视觉目标检测任务而言，其标注的工作量主要集中在于对图像中的多个目标的位置进行标注，而该过程工作量往往非常大并且十分繁琐，导致标注的过程比较长、记录的数据比较复杂等，这些问题

都给实际的标注工作带来很大负担。

本文采用上述的第二种标注方法降低标注的工作量，去掉最繁琐的位置标注过程，只对每张图像只进行弱监督标注。在标注的过程中，该标注方式实际只需给出图像中的类别标号，该标号表示在图像中是否存在待指定类别的目标，通过这种方式显著减少样本标注工作量。

弱监督目标检测算法可以以更少的人工标注代价扩展数据集，并可以利用大多数主流的图像分类数据集。由于对标注的要求较低，弱监督目标检测算法可以从网络中扩展数据，例如使用关键词在网络上搜集图像数据。这些特性增强了弱监督对大规模数据集的利用。但是，由于标注信息相对于目标检测任务而言不完整，如何学到好的检测模型这一问题非常具有挑战性。虽然有很多研究人员致力于解决该问题并提出了很多相关方法，但是弱监督的目标检测算法离监督目标检测之间的差距仍然非常巨大。大数据时代的到来导致数据井喷式的增长，这使得各类算法对弱监督标注的需求量越来越大，能否将弱监督的算法投入实际使用这一问题变得越来越重要。

综上可知，弱监督视觉目标检测任务有着非常重要科学研究意义与实际应用价值。

1.2 研究现状和存在的问题

1.2.1 研究现状

弱监督的图像目标检测框架主要包括三个步骤：候选框的提取[20-25]、特征提取[7][26-31]和弱监督学习[13-19]。前两个步骤中的方法一般都是沿用监督目标检测中的方法。弱监督学习算法中，传统方法包括聚类法、隐变量学习算法和多示例学习算法。随着深度学习的兴起，传统的弱监督学习算法逐渐和深度学习结合，主要包括非端到端和端到端两种。

下文首先介绍弱监督目标检测的基本框架，再介绍传统弱监督目标检测算法和基于深度学习的弱监督目标检测算法。

1.2.1.1 弱监督检测的框架

候选框的提取：最早的候选框提取的方法就是扫窗（Sliding Window）[7]。

该方法非常简单有效，具体的做法是：首先对图像做多尺度变换，形成图像金字塔；然后利用多个尺度、多个比例的窗口对图像金字塔中的每层采用自左向右、自上向下的逐像素扫描方式提取候选框，作为目标的定位候选框。该方法的好处在于，多尺度和多比例的方式，配合逐像素扫描，能够保证无论目标出现在图像中的位置和尺度如何变化都能至少保证一个框定位到目标。也就是说，扫窗方法能够保证数据集中目标预定位的查全率（查全率指的是定位到的目标总数和数据集中目标总数的比值）。查全率是目标检测前提，也是目标检测的上限。如果查询率低，例如查全率只有 60%，那么无论后续的检测器检测能力有多强，目标检测的性能不会超过 60%，另外 40%在检测器识别之前就已经丢失。扫窗的方法极大的保证了查全率，但是，该方法却有一个非常大的弊端，就是候选窗口过多。通常情况下，一张普通图像（例如像素为 500×375）能产生百万级数量的候选窗。检测器需要对每一个候选框进行分类，这极大地增加了计算复杂度，降低了检测的实时性。

为了在保证高目标查全率，同时尽可能减少候选框的个数以增加检测效率，人们提出了很多候选框提取算法，例如 Selective Search[20]、Bing[21]、MCG[22]、Rantalankila[23]、Edge Boxes[24]和 MSER[25]。这些方法生成候选框的过程不需要任何标注信息，只利用图像的纹理、边缘等底层信息生成候选框，最终产生只有 2000 左右个候选框，而查全率（Recall）也能达到 80%~90%。由于这一类方法能够以三到四个数量级的规模减少候选框个数，极大的提升了检测效率，近年来受到了越来越多的重视。本文所使用的候选框是 Selective Search 和 Edge Boxes，在所有候选框算法中这两种方法兼顾了查全率高和候选框个数少的特性。

特征提取：特征提取算法要求对不同尺寸的候选框输出相同维度的特征，以保证分类器的输入要求。此外，对候选框提取的特征应当尽量简洁，保留对分类有利的信息，去除冗余信息；对同一个类别的候选框而言，特征的距离应当尽可能接近，而不同类别的候选框的特征应当尽可能远离。特征提取的结果直接决定了其对目标的表示，是目标检测中非常关键的一环。

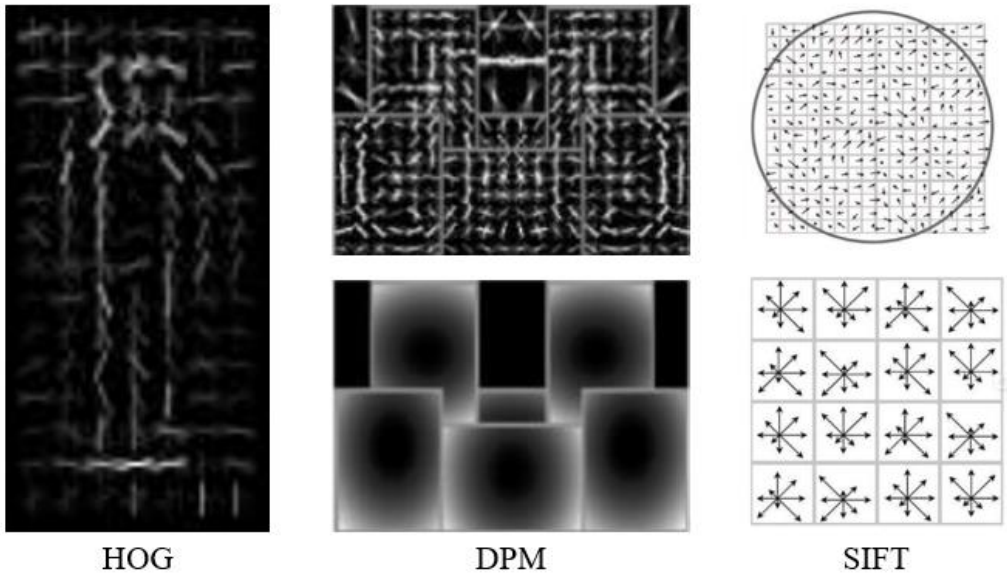


图 1.3 手工设计的特征示例[7,31,32]

Figure 1.3 Illustration of hand-craft features

特征提取的算法可以分为两类：手工设计特征和深度学习特征。在早期的机器学习中，特征提取方法主要是手工设计的方式，而早期的手工设计特征往往都是低层次的特征，比如 Haar-like 特征[26]、HOG（Histogram of Oriented Gradient）特征[7]、LBP（Local Binary Pattern）[27]特征及其改进版 HOG-LBP[28]、SIFT （Scale-Invariant Feature Transform）[29]以及对上述方法的改进和组合。这些特征均为像素级特征，针对图像中的梯度、边缘和纹理特性等进行特征提取。但是这些特征的表示能力往往比较低，对分类十分不利。为此，研究人员进一步提出了更高层次的手工设计的特征，这些特征往往是基于上述底层特征设计的，例如 DPM（Deformable Parts Model）是基于 HOG 特征的、Fisher Vector 是基于 SIFT 特征。随着机器学习的兴起，深度特征被越来越多的研究人员所采用，如 RBM（Restricted Boltzmann Machine）[30]，CNN（Convolutional Neural Network）[31][32]等特征。图 1.3 是几种特征的举例。

弱监督学习：弱监督学习是算法一般以多示例学习为主。在多示例学习中，图像被看做是示例包，而图像中提取到的候选框被看做是示例包中的示例。经过上述两部分获得示例（候选框）及其对应的特征之后，弱监督学习的目的是在只给定示例包标号，而不给定示例的标号的情况下学习到关于示例的分类器，同时定位目标。为了能够简洁的区分弱监督目标检测中的标注信息和其他监督

的标注信息的区别，我们给出了各个监督框架下的目标检测任务中的标注示意。由图中可以看出，在全监督目标检测中，图像中所有的目标均有准确的位置标注；而弱监督中只有图像信息的标注，没有目标位置和个数的标注；半监督问题则是图像数据集中只有部分图像有全监督的标注，而其余图像中没有任何标注。

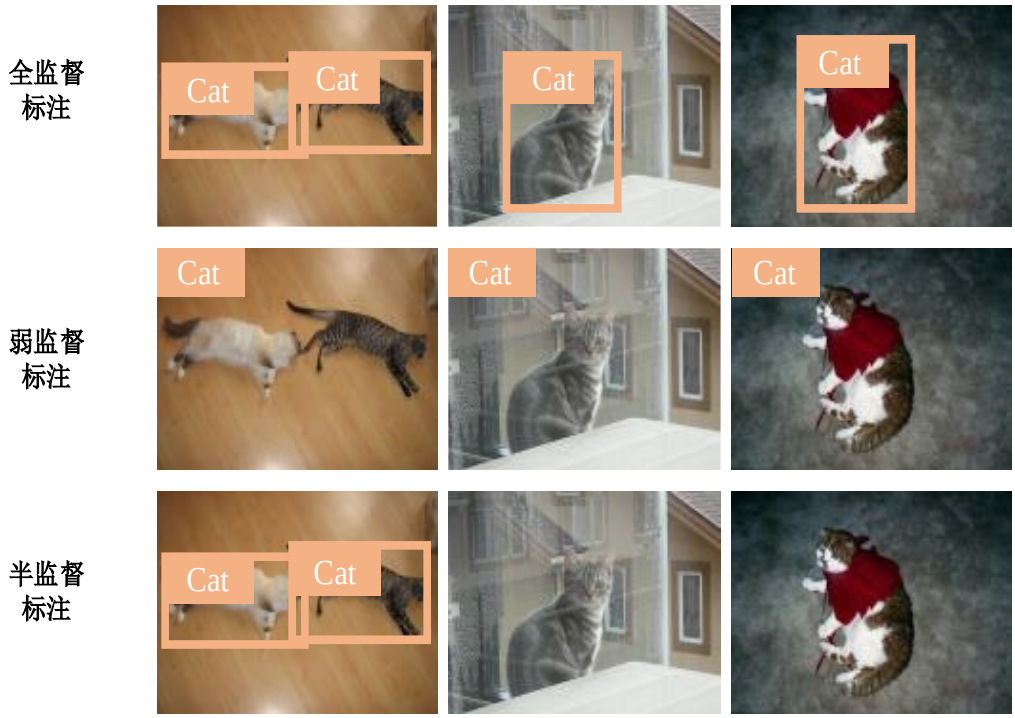


图 1.4 正反例图像和样本举例

Figure 1.4 Samples of positive and negative images/sampes

对于不同的监督信息，在训练阶段，各个任务之间的训练过程也会有所区别，其关系和区别如图 1.5 所示。图 1.5 可以看出，相对全监督问题和半监督问题而言，弱监督目标检测的过程需要同时确定目标的标号和目标的位置，其核心问题是如何在只有图像的标号，而没有样本标号的情况下学习得到关于样本的分类器。

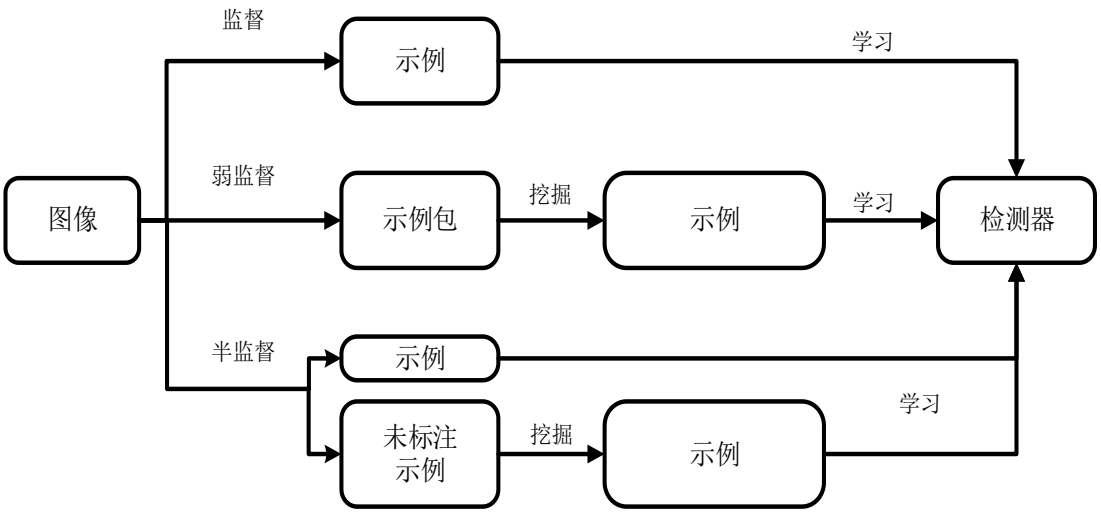


图 1.5 弱监督和全监督学习框架对比

Figure 1.5 Comparison of weakly and fully supervised learning frameworks

针对以上问题，加州伯克利大学计算机系的学者提出了一种通过使用图覆盖理论进行弱监督的目标检测的方法[37]；法国 INRIA 实验室的学者使用多示例学习方法（MIL）的方法[38]；新加坡国立大学的学者使用了自动样本标注的方法[39]；中科院自动化所的学者使用了隐语义聚类和多示例学习[40][41]相结合的方法；哈尔滨工业大学的学者[42]研究了基于随机游走图模型的弱监督机器学习方法；西安交大的学者使用协调分割的方法对特定的视频场景中的目标进行无监督建模[43]；解放军信息工程大学[44]、山东大学[45]、西北工业大学[46]等高校的学者对弱监督分类的方法进行了研究。南京大学计算机系的学者提出了确保半监督的学习方法在学习过程中不降低性能的理论[47]；武汉理工大学的学者提出了一种去噪的受限玻尔兹曼机的弱监督特征学习方法[48]。

1.2.1.2 传统弱监督检测算法

传统弱监督视觉目标检测的建模方法大致可以分成三大类：聚类方法[17][41][49]、多示例学习方法(Multi-Instance Learning, MIL)[38][44]与隐变量支持向量机(Latent SVM, LSVM)[16][17][49]方法。聚类方法主要通过无监督的学习方式在所有样本的集合中找到一个聚类性突出的子集，该集合可以较好的涵盖正例样本的集合。该方法的研究者使用了隐语义空间分析的方法[49]来实现高层次的语义

级别的聚类。Bilen 等人将隐变量的目标求解过程与聚类方法相结合，依靠聚类算法的中目标方程是凸函数这一性质提高隐变量求解中的目标函数的凸性。

多示例学习（MIL）的方法则将每个图像视为一个包含多个目标候选区域的“包”，在学习过程中迭代选择得分高的候选区域作为正例样本并以此更新模型。为了避免高得分样本是反例（非目标区域），Multi-fold MIL[38]将整个样本的集合划分为多个子集，在训练过程中进行交叉验证与协同训练。中科院自动化所针对大规模数据的学习提出了 MILlinear 方法[40]，该方法通过将求解 MIL 的传统的梯度方法替代为信赖域的方法，最终显著提高了学习速度。

隐变量支持向量机（LSVM）将图像中目标样本位置作为隐变量，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建模。但是，与其他弱监督学习方法类似，LSVM 的目标函数是非凸的，这就使得 LSVM 在学习过程中无法像 SVM 那样求得全局最优解。一个通常的方法是将其非凸目标函数写成两个凸函数的差的形式，在学习过程每次迭代优化时找到一个线性函数作为第二个目标函数的线性上界，进而保证每一步迭代中目标函数是凸函数。为了阻止非凸目标函数迅速陷入局部最优，Song 等人[37]采用聚类方法对隐变量进行初始化，并采用 Nesterov 平滑方法，通过将目标函数转化为欧拉二次型而增加凸性。Bilen 等人[16]提出了根据先验知识在目标函数中加入两个凸函数进行正则化，以提高 LSVM 解的质量。后续研究中，Bilen 等人[16]将非凸最优化方法用来求解初始解，而采用凸聚类为主要方法进行求解。

1.2.1.3 基于深度学习的弱监督检测算法

随着深度学习的兴起传统的深度学习方法，越来越多的方法采用深度学习的框架[57-67]。最早的深度学习框架下的弱监督目标检测算法是 WSDDN 方法，该方法利用两个全连接层分别取计算图像中的各个候选框的定位信息和类别信息，然后将所有的候选框的类别信息相加之后用于预测图像标号。通过这种方式，深度网络很好的构建了目标检测和图像分类之间的关系。该方法因此也被很多弱监督目标检测框架采用。其中包括引入上下文信息、增加检测重新精调、分割信息等。

1.2.2 存在的问题

尽管国内外的研究机构开展了弱监督目标检测工作并取得成果，弱监督方法的性能相对于全监督方法仍然相差甚远，如下图所示。

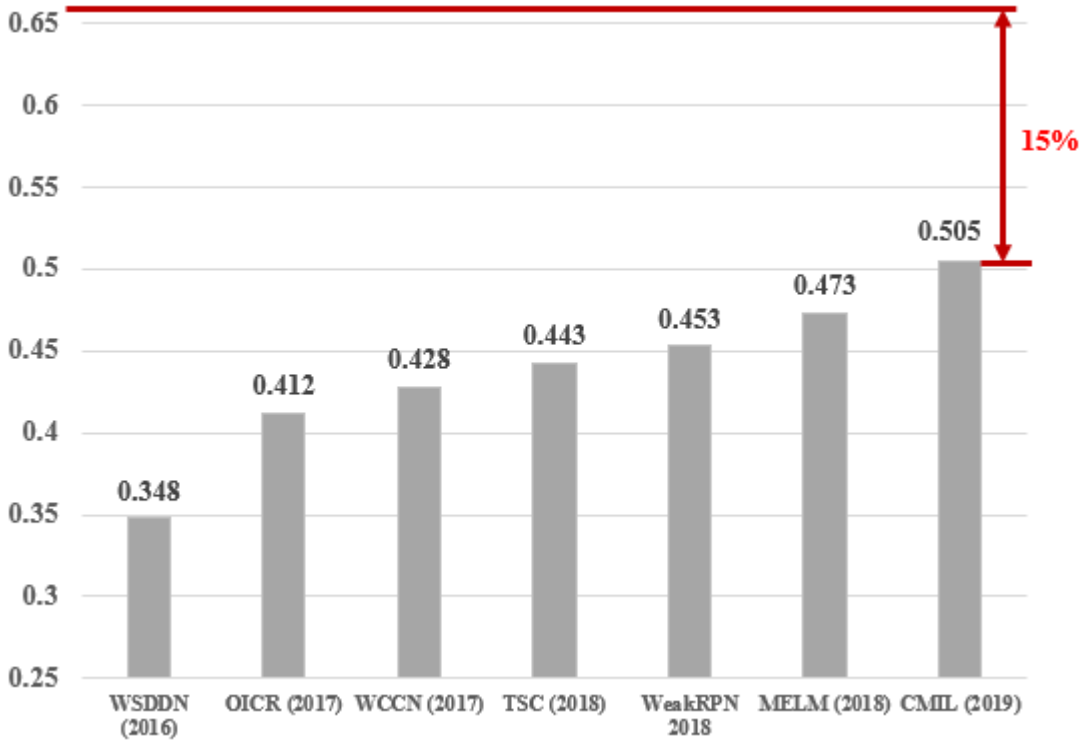


图 1.6 弱监督目标检测算法的性能发展以及和全监督性能的对比

Figure 1.6 Detection mAP of weakly supervised object detection methods and comparison to supervised method.

在建模方面，弱监督学习方法的目标函数一般是非凸函数，容易陷入局部最优。传统的多示例学习、LSVM 方法的求解思路应用到图像目标建模时也会遇到此问题。传统的多示例学习、LSVM 的目标函数通常是以图像级分类准确率为导向，而不是样本分类准确度为导向。在实验中经常观测到的现象是：弱监督算法将目标的部件（Part）当成目标本身，实现了图像级分类精度最高，模型很快收敛到局部最优解，结果造成很多目标部件或者具有相关性的其他目标被错误当成目标本身。在目标位置不能确定的情况下进行模型学习与求解，其解空间是巨大的。在特征表示方面，现有深度特征一般都依赖于大量准确标注的样本，如何利用弱标注样本提高特征表示能力的问题也需要更好地解决。

1.3 本文的研究内容与主要贡献

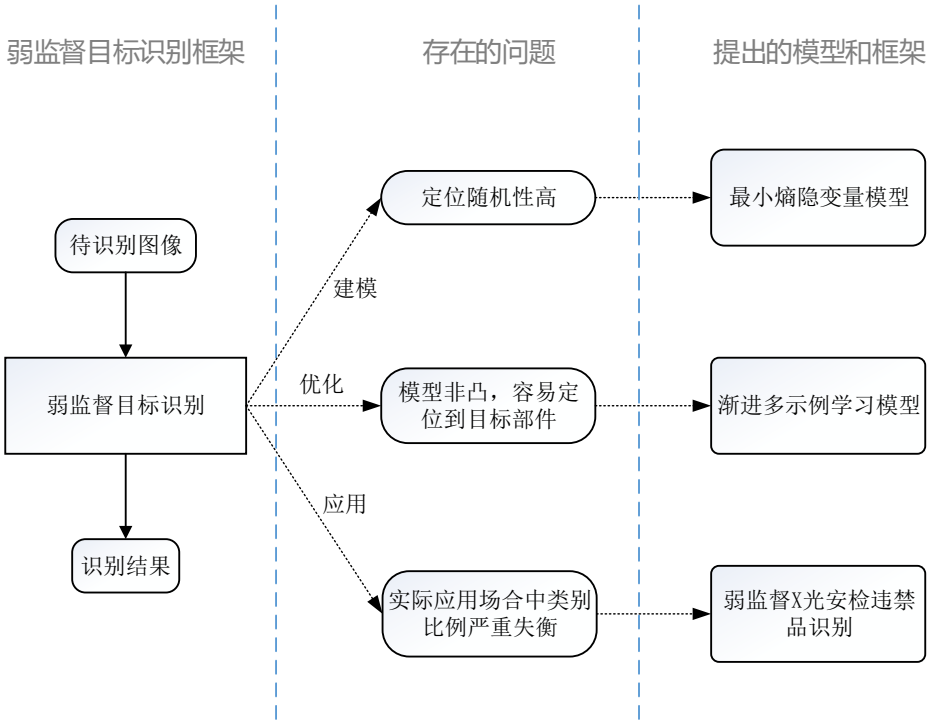


图 1.7 建模、优化和应用三方面对弱监督目标识别任务的研究

Figure 1.7 Weakly supervised object recognition task from the aspects of modeling, optimization and application.

本文的研究内容如图 1.7 所示，对于弱监督目标识别问题，本文从建模、优化和应用三个方面，分别解决弱监督目标识别中存在的定位随机性高、模型非凸导致容易定位到目标部件和实际应用场合中类别比例严重失衡等问题，分别提出了最小熵隐变量模型、渐进多示例学习方法，并实现了弱监督 X 光安检违禁品识别这一实际应用框架。具体研究内容包括以下几个方面：

（1）提出了一种有效的深度学习模型，称为最小熵隐变量模型，用于弱监督目标检测任务。最小化熵能减少系统的随机性，通过引入最小熵隐变量模型，算法在训练阶段的定位随机性得到了降低，因此能够更稳定的学习目标特征，提升目标注位的准确性。最小熵隐变量模型的贡献总结如下：一是采用深度神经网络结合最小熵隐变量模型以便更有效地挖掘到目标候选框，并且最小化学习过程中的定位随机性；二是采用一个候选框团更好地搜集目标的信息并激活完整的目标区域，从而能够更准确的检测到目标。三是用一个循环学习算法分别将图像分类和目标检测看做一个预测器和一个校正器，并且利用连续优化

（Continuation Optimization）的方法解决非凸优化问题。四是在 PASCAL VOC 数据集上取得了 state-of-the-art 的分类、定位和检测性能。

（2）提出了一种有效的弱监督目标检测方法，称为渐进多示例学习。渐进多示例学习的方法致力于解决传统多示例学习方法的非凸优化问题。通过引入一个序列的对原函数的平滑损失函数，在训练过程中以一个容易求解的凸损失函数为起点，逐渐优化该序列中的平滑损失函数，直至损失函数退化成原损失函数。该平滑过程是通过引入示例子集的方式完成的。渐进多示例学习显著提升了弱监督目标检测和弱监督目标注位的性能，并超过了目前最新已发表的工作。

这些现象背后的原理在于：当使用渐进优化模型和深度网路结合时，模型在训练过程中通过搜集目标或者目标部件的方式激活了目标的完整区域，从而最终学习到语义稳定极值区域。这为弱监督视觉目标检测任务带来了新思路。

（3）提出了类平衡分层激活网络，该模型通过增加深层特征对中间层特征的监督，使得中间层特征能够得到更精细的视觉线索并且能够过滤掉一些不相关的信息。除此之外，设计了类别均衡的损失函数，通过减少反例样本的数量，尽量的使正例和反例的数量达到一个平衡，并且该损失函数依赖于分层的网络结构，使得深层侧输出的损失函数对浅层侧输出的损失函数有指导作用。在 SIXray 数据集中三个子集上，与多种方法进行对比实验后发现，本章提出的类平衡分层激活网络性能有显著提升，除此之外，在自然场景数据集上，我们的方法也表现出了比较好的性能，提现出方法的泛化能力与可扩展性。

1.4 本文的组织结构

第一章，绪论。论述了弱监督目标识别的研究背景和研究意义，分析了当前弱监督目标识别中存在的难点和常见问题，并陈述了本文的研究内容和研究贡献。

第二章，相关工作与技术。介绍了弱监督目标识别的研究现状和相关工作，从传统的基于手工设计的特征的弱监督目标识别算法，到基于深度卷积神经网络的弱监督目标识别算法，再到最近的端到端的网络架构。另外介绍了和弱监

督目标识别的相关任务和方向。

第三章，最小熵隐变量模型。首先分析了弱监督目标检测任务中的定位随机性的问题，然后对定位随机性的问题进行建模，通过训练最小熵隐变量模型解决定位随机性的问题，随后从模型优化的角度论述了最小熵隐变量模型的作用，并在最后对该模型进行了充分的实验分析。

第四章，渐进多示例学习。首选论述了传统多示例学习方法的模型非凸问题，以及模型非凸所造成的影响，随后对原非凸模型做出凸性分析，并指出具体需要解决的问题；然后根据弱监督目标检测问题的特性，提出了基于示例团划分的优化方法，并结合渐进优化，提出了渐进多示例学习的优化模型。最终对该模型进行实验分析和验证。

第五章，弱监督 X 光图像违禁品定位。首先论文在实际使用场景，X 光安检场合下，违禁品和非违禁品类别差异巨大，导致模型学习失效的问题；然后利用卷积神经网络自身的特性，提出了逐层优化的方式，逐渐降低类别不平衡的问题造成的影响，并给出实验验证。

第六章，总结与展望。总结了本文的主要内容，并对未来的工作进行了展望，包括无候选框弱监督目标检测算法、元学习驱动的弱监督目标检测等。

第2章 相关工作与技术

上一章介绍了国内外相关的弱监督目标检测算法的研究意义和背景，本章作为上一章的扩展，详细综述了弱监督目标检测任务中常见的相关工作和技术。由于到弱监督目标检测中大多数方法从全监督目标检测框架中继承而来，下文首先介绍全监督目标检测相关的方法，然后介绍弱监督目标检测的相关工作。

2.1　全监督目标检测

2.1.1 候选框提取算法

候选框提取算法是弱监督目标检测的前提。为了能够检测到目标，算法往往首先需要得到目标潜在的位置，然后再对所有潜在的位置进行区分，排除背景位置，最终检测到目标。检测的结果是目标所在的位置和类别。本文中的弱标注信息是图像级的，也就是每张图像只有包含目标的类别标号，没有目标的具体位置。因此，如何有效的定位到目标对于检测的问题而言十分重要。最简单可行的方法就是穷举法，也就是对目标可能出现的位置进行穷举式的搜索；由于穷举式的方法计算代价太高，研究人员提出了利用目标的先验信息来预测目标候选框的算法。

穷举式的候选框提取算法对所有可能的区域进行遍历搜索，也就是扫窗。该过程中窗口会以多尺度多比例的方式对图像逐像素扫描。穷举式的候选框提取算法的特点是具有非常高的查全率。但是其缺点也非常突出，即候选框个数太多。穷举式的方法往往会生成百万级数量的窗口，这会为后面的检测分类带来巨大的计算量。因此，该方法逐渐被其他基于先验的方法替代。

Selective Search[20]是目前候选框算法中查全率和窗口数量两个方面都很出色的算法。该方法综合了穷举法和图像分割的方法。与传统的单一策略相比，其结合了多种策略，同时大幅度降低了候选框的数量，让更复杂和精细的识别算法可以被使用。该算法使用了分割的方法将图像分割成小的图像区域，然后再分层合并。

近年来涌现出一批基于机器学习的候选框提取算法。和以往的候选框提取算法不同，基于机器学习的候选框提取算法使用的不是先验知识，而是人工标

注的候选框训练样本。通过学习地方法进行建模比传统的候选框提取算法中所依据的一些规律性的先验知识更为准确，指导性更强。目前，比较新的基于机器学习的候选框提取的算法有 Philipp Krähenbühl 等人提出的方法[52]和何凯明等人提出的 Faster-RCNN 框架中的 RPN（Region Proposal Networks）网络[36]。然而对于弱监督目标检测任务而言，由于缺少目标的标注信息，基于学习的候选框提取算法难以应用。

2.1.2 特征提取

上一节中讲到了如何在图像中提取样本（候选框）。样本提取完成之后，为了能让分类器正常处理所得的样本，需要先对样本提取特征，使得其满足后续分类器的输入。特征提取算法对于检测而言非常重要，好的特征往往会让后续的分类任务事半功倍。因此，近年来大量的学者对特征的提取进行了研究并提出了很多高效的特征提取算法。这些特征主要可以分为手工设计的特征和基于学习的特征。下文将分别两类特征提取算法进行介绍。

最初的特征提取方法主要是手工设计的。早期的手工设计特征往往都是低层次的特征，这些特征都是像素级的，主要是提取图像中的梯度、边缘等。本文主要介绍两个最为经典的手工设计的特征：HOG 和 SIFT。

方向梯度直方图[7]（Histogram of Oriented Gradient，HOG）特征提取方法就是对一个图像提取不同方向的梯度，然后统计图像中各个区域的关于梯度的直方图。具体步骤如下：HOG 特征通过直方图的方式来表示图像中的梯度分布情况。对于图像的局部，HOG 特征只描述梯度的分布特性而舍弃了梯度的空间分布特性，最后又使用 block 的方式建立各个图像局部之间的空间位置关系描述。该特性使得 HOG 特征对于图像的光照、阴影和噪声好的鲁棒性。然而，HOG 特征只关注梯度信息，在提取过程中舍弃掉了其他信息，使得其目标表示能力效果受到了限制。

尺度不变特征变换[29]（Scale-invariant feature transform，SIFT）是一种用来描述和检测图像特征点的算法。该算法最早由 Lowe 于 1999 年所提出，现在已经广泛应用于图像匹配、图像拼接、3D 建模、姿态对比等应用。SIFT 特征不仅具有尺度不变性，而且对图像的旋转、图像亮度的变化或拍摄角度的改变具有很好的鲁棒性。SIFT 在图像的尺度不变特征提取方面的优势非常明显，他对旋

转、尺度缩放、光照变化等都能保持不变性。但是，SIFT 的计算实时性不够高，对边缘光滑的图像无法准确提取特征点。

2.1.3 特征学习

随着机器学习算法的快速发展，越来越多的学者开始使用机器学习的方法来提取特征。特征的学习过程使用机器学习的算法，目的是要让学习的目标方程或是损失函数在设定的意义上达到最优。最终在实际中使用的特征一般都是机器学习算法达到最优时的靠近输出的中间值。

卷积神经网络（Convolutional Neural Networks，CNN）[31][32][53]的由于其出色的特征表示能力而被广泛采用。普通的神经网络输入都是向量，因此，在一张图像用作输入的时候，首先应该对图像提取特征使其变成一个固定维度的向量，然后再进行神经网络的运算。这样使得提取特征和训练网络分离开来，二者之间相互独立的进行，不能更好的适应和匹配对方。卷积神经网络则不同于此，它的输入就是二维（灰度）甚至三维（彩色）图像，输出是分类数。整个卷积神经网络将提特征和分类融为一体，不需要采取任何其他的人为操作，所有参数都是自动学习的。但是，直观上来看，增加输入的维度会使得整个网络的参数以指数次方增长。事实上，卷积神经网络的这一举动确实会使网络参数大幅增加，如果没有任何应对措施，卷积神经网络就无法做到更大更深，也就发挥不出其大数据大网络的优势。为了解决这个问题，卷积神经网络使用了局部接受野和权值共享的策略，这极大的降低了参数的个数，使卷积神经网络的计算得到广泛应用。下文将详细介绍其原理。

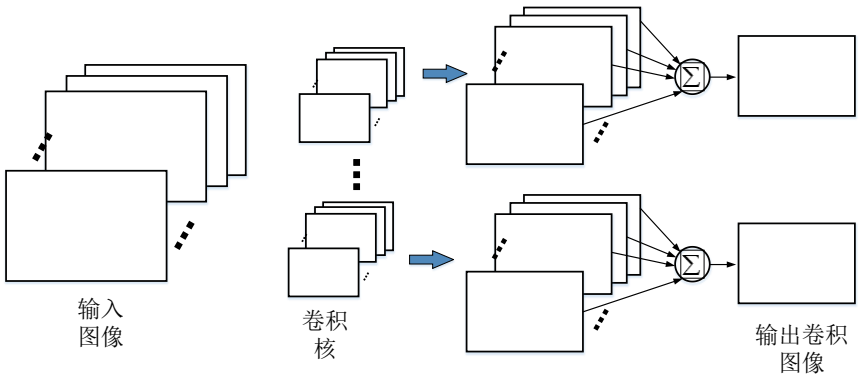


图 2.1 卷积运算示意图

Figure 2.1 Illustation of convolutional operation

一个基本的卷积神经网络包含：卷积层、池化层和全连接层。首先介绍卷

积层，如图 2.1 所示，该图为卷积层的运算过程。其中，卷积核的定义的由来就是局部接受野的理论概念，也就是输出图像的每一个点只与输入图像的一个局部有关。这样避免了输出层的每一个点都需要与输入层的每一个点进行运算。另外，对于每一个输出卷积图像而言，它的每一个点都是由原图和相同的卷积核计算得来的，这就是权值共享。权值共享使得最终卷积层的参数就是卷积核的参数的个数，这使得训练所需的参数大幅减少。

另一个重要的结构就是池化层。池化层的存在是为了减小卷积图的大小，本质上是降低计算复杂度。另外，由于池化的特性，池化之后的输出对于图像的一些平移、缩放等变换都具有一定的鲁棒性。池化层的方式有很多种，比较常用的有最大池化，平均池化等。

最后，卷积神经网络会接连一个全连接的人工神经网络。由于类别信息为向量，而输入信息是矩阵，卷积神经网络最终将输入的矩阵展开，构成向量，然后用一个全连接层将该向量投影到类别空间。

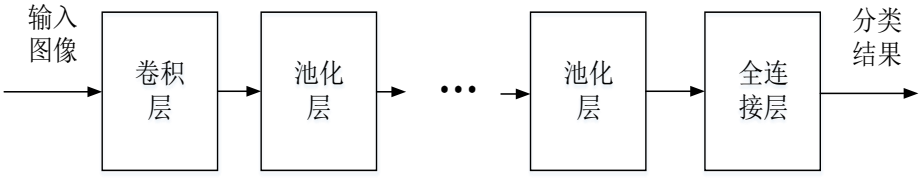


图 2.2 卷积神经网络示意图

Figure 2.2 Illustration of convolutional neural networks

卷积神经网络的结构如图 2.2 所示。输入图像之后，重复进行卷积和池化，重复若干次之后，再接全连接层作为输出。

2.2 弱监督目标检测

学习算法是弱监督目标检测的核心，主要解决在标注信息不全或是不准确的情况下如何学习好的分类模型。目前，国内外有很多学者致力于弱监督学习算法的研究，本文综述了应用最为广泛的几类弱监督学习算法，包括：多示例学习（MIL）、聚类法以与隐变量支持向量机（LSVM）。弱监督的学习算法大多数都依赖于监督学习算法，是监督学习算法的改进。本节首先介绍监督分类学

习算法，为后文介绍弱监督学习算法做铺垫；后续小节将分别介绍上述的几类弱监督算法。

弱监督分类算法主要针对分类样本标注不全或不确定的情况，通过比较弱的标注信息和先验知识学习分类模型。弱监督分类算法主要是以监督分类器为基础，通过增加惩罚项或是先预测样本再监督分类。弱监督标注的方式分为以下三种：线标注，点标注和图像及标注，如图 2.3 所示。线标注只需要给给出图像中目标的对称轴的大致方位；点标注相比线标注更加简洁，只需要给出目标的大概的中心位置；图像级的标注在这三者中标注量最少，只需要给出图像中存在的目标的类别。

本文采用图像级的标注作为弱监督的标注方式，其原因有以下几点：首先图像级的标注最能减少标注工作量；其次图像级的标注使用范围更广，除了标注工作量最少之外，还可以利用现有的图像分类数据集和网络中的关键词搜索信息，具有非常好的扩展性。

图 2.3 几种弱监督标注的示意图

Figure 2.3 Illustration of weakly supervised annotations

弱监督目标检测相关方法可以分为传统方法和基于深度学习的方法，其中，传统方法包括聚类方法、多示例学习和隐变量学习；基于深度学习的方法包括非端到端的方法和端到端的方法两种，如图 2.4 所示。

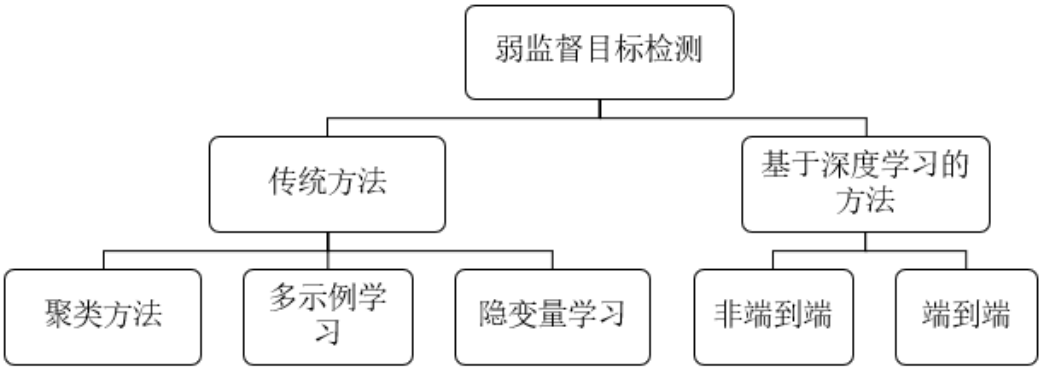


图 2.4 弱监督方法相关工作总结

Figure 2.4 Summarization of weakly supervised object detection methods

2.2.1 传统方法

多示例学习：多示例学习（Multiple Instance Learning，MIL）[38][44]最早是由药物分子学科的学者提出的。该算法的提出是为了解决同分异构体所造成的混淆，最终找到同分异构体中真正起作用的那一种。多示例学习分为两个层级，一个是示例（样本）级，一个是示例包级，每个示例包中包含多个示例样本。传统的监督学习可以看作是每个示例包只包含一个示例，也就是示例和示例包是一一对应的。因此，多示例学习是传统监督学习的一个延伸。多示例学习由于其广泛的应用场景和独有的学习性质而受到了计算机视觉领域的关注和重视。

多示例学习中，样本的标注是以示例包为单位的。示例包中包含若干个示例样本，样本分为正例样本和反例样本。当一个示例包中出现至少一个正例样本时，示例包被标记为正（标号为 1）；当示例包中没有出现正例样本，也就是示例包中所有样本都为反例样本的时候，示例包被标记为负（标号为-1）。多示例学习的输入为示例包的集合，示例包的标号取值为{1，-1}。在反例包中，所有样本的标号的确定的，都是反例；但是正例包中则不同，正例包中的样本标号未知，但可以确定的是至少有一个样本标号为正。

多示例学习的框架分为两种，第一种是基于示例分类的框架（mi-SVM），第二种是基于示例包分类的框架（MI-SVM）。可以看出，聚类的核心目的是寻找正例样本，在找到正例样本之后，再结合反例图像中的反例样本训练最终的

分类模型。多折多示例学习（Multi-fold MIL 该方法可以视为对多示例学习的一种改进方法。多示例学习的一个很大的问题在于对其对模型初始化过于敏感，初始化不当容易导致模型往错误的方向学习，从而使得模型更容易陷入局部最优。聚类法很好的优化了这个问题。但是，聚类法过度依赖正例样本之间的聚类性，在样本的类内间距比较大的情况下容易选错正例样本，从而导致模型学习失败。

隐变量支持向量机：隐变量支持向量机（LSVM）[16][17][49]是本文研究工作的基础。LSVM 所解决的问题和 MIL 一样，样本的标注是以示例包为单位的。正例包中至少包含一个正例样本，反例包中一定不包含正例样本。与 MIL 不同，LSVM 给出了分类器的损失函数，分类器的学习过程就是对损失函数求最小。损失函数的形式如下：

$$L(w)=\frac{1}{2}\|w\|^2+\lambda\sum_{i=1}^{n}\max\left(0,1-y_i f\left(x_i,w\right)\right) \tag{2.1}$$

其中 $f\left(x_i,w\right)$ 是图像 x_i 在分类器 w 下的得分，在传统的方法中，该分类器一般为 SVM 分类器。

由上述公式可以看出：当图像的得分与图像的标号不一致的时候，该图像会造成比较大的损失；当图像的得分与图像的标号一致的时候，损失比较小，甚至为 0。可见图像的得分影响着整个优化过程，一般图像的得分的表达式形式如下：

$$f\left(x,w\right)=\max_{z} w\cdot\phi\left(x,z\right) \tag{2.2}$$

其中 z 为隐变量，表示目标可能的位置。图像的得分一般是取图像中样本的最大得分，这个定义是根据弱标注的形式来定的。

2.2.2 基于深度学习的方法

基于深度学习的方法大致分为两类：非端到端的方法和端到端的方法。其中，非端到端的方法通常采用多示例学习的方式，将传统的分类器替换成深度卷积神经网络，如图 2.5 所示。

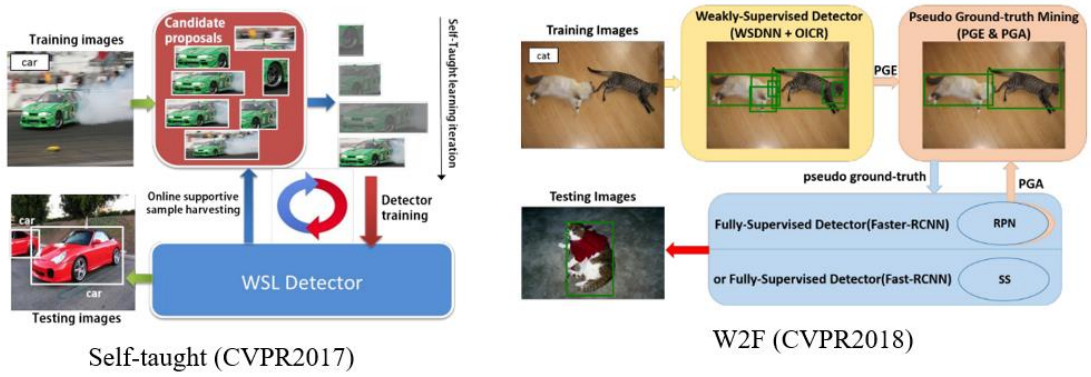
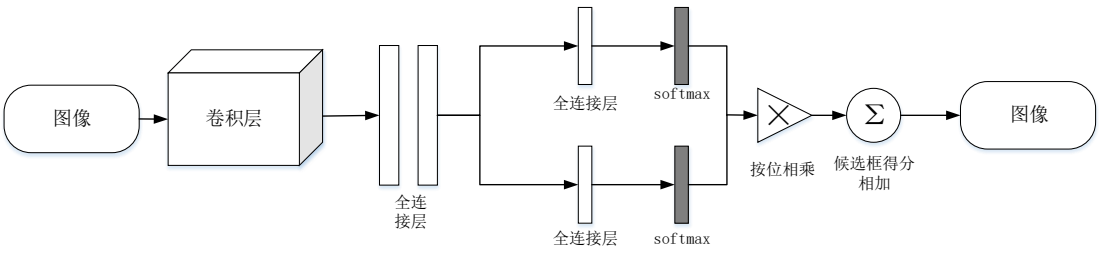


图 2.5 非端到端的弱监督深度检测方法

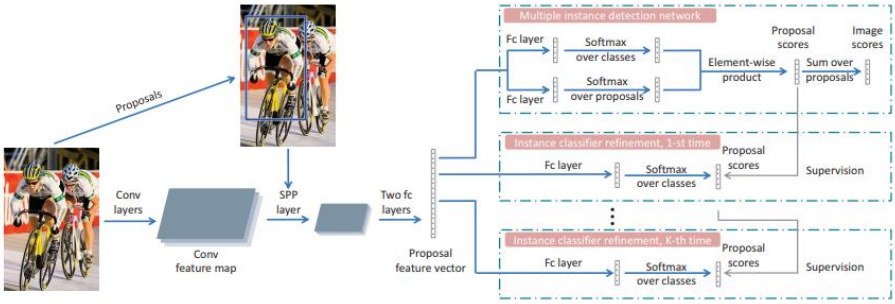
Figure 2.5 None-end-o-end methods for weakly supervised object detection network.

最新已发表的非端到端的方法有 Self-taught[61]和 W2F[96]，这两个方法的核心是挖掘样本，然后利用挖掘的样本重新按监督学习的方式训练检测器。这一类的方法通过设计策略挑选更为准确的窗口，从而提升检测性能。在 Self-taught 中，作者使用了一种图传的方式去挖掘样本，通过目标之间的相似性度量，挖掘潜在的目标，并用该目标训练 Fast RCNN 检测器；训练好 Fast RCNN 检测器之后，又重新更新所有候选框的得分，对目标候选框重新挖掘。该方式迭代多次直到模型收敛。而 W2F 的框架则是基于其他弱监督方法的结果，以设计融合策略的方式更新目标的位置。该方法中的融合的方式是基于对基准方法的检测结果的经验观察，并利用先验信息设计候选框修正和融合策略，使得最终得到的候选框更加准确。在的大候选框之后，该方法训练了 Faster RCNN 检测器，取得了很好的性能。

然而其弊端则是训练过程十分复杂和耗时，需要调整的参数也非常多，对学习十分不利。由于这些问题，越来越多的研究人员开始研究端到端的弱监督检测算法。



(a) WSDDN[59]



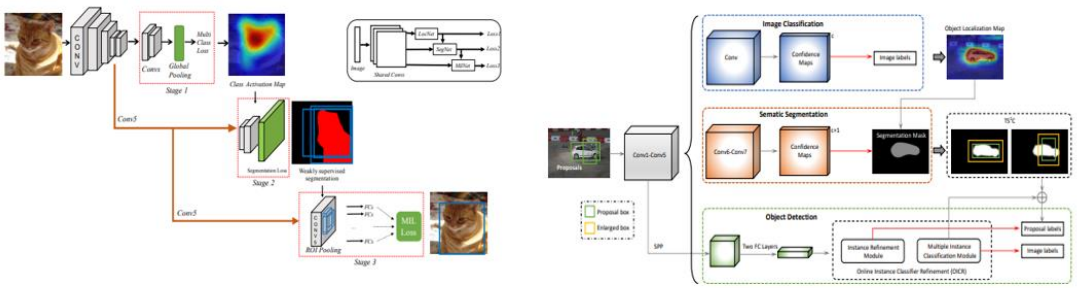
(b) OICR[63]

图 2.6 弱监督深度检测网络（WSDDN）及其改进算法 OICR

Figure 2.6 Weakly supervised deep detection network (WSDDN) and its improved method

弱监督深度卷积神经网络在 2016 年有 Bilen[59]提出，该方法的结构如图 2.6 所示，网络结构以 Fast RCNN 为基础，在最后一个全连接层后加上两个全连接层的分支，分别用于定位和分类，最终综合定位和分类的结果预测图像的类别标号。该方法网络结构简单，容易训练，被越来越多的研究人员使用。

另外一种端到端的弱监督检测算法则是基于分类的框架，如下图所示：



WCCN (CVPR2017)　　TS²C (ECCV2018)

图 2.7 基于分类网络和语义分割信息的弱监督检测算法 WCCN[58]和 TS²C[66]

Figure 2.7 Classification and semantic based weakly supervised object detection methods WCCN 和 TS²C

这两中方法分别成为弱监督级联卷积网络（WCCN）[58]和 TS²C[66]，其通过图像分类的框架激活卷积特征，从而定位出目标的大致位置，并结合语义分割的信息预测目标的位置，再用改位置信息训练检测器。该方法的优点是计算效率高，但是性能往往受到限制。

2.3 特征学习和建模

根据本文涉及的研究内容，对相关的特征学习和弱监督建模分别进行了综述。

2.3.1 无监督特征预学习

视觉特征表示的无监督学习可以粗略地分为两种：（1）生成式方法，（2）聚类与关联方法。早期的生成模型方法包括自编码器[69][70]和受限玻尔兹曼机（RBMs）[71][72]。例如，Le 等[73]在 YouTube 视频的大规模数据集上训练多层自编码器：在无图像类别标注的情况下学习的高层神经元可以识别猫和人脸。最近的生成模型，如生成对抗网络[74]和变分自编码器[75]，能够在给定部分样本的情况下生成风格不同的图像并扩张样本空间，从而为模型“自我更新”提供了基础。

基于不变性特征的区域关联是自学习特征的重要方法。视觉不变性可以通过在视频帧序列中拍摄的相同实例/场景来捕获，基于不变性的特征进行匹配、跟踪或聚类进而自主生成图像类别标注，用于特征自学习。自学习特征表示的关键是挖掘目标/区域的共同特性，主要由图像块的相对位置[76]反映出目标空间布局，或者表观一致性[77, 78]。在此过程中的目标与区域匹配都非常依赖于特征的（旋转、尺度等）不变性与适应性。如果特征不变性得不到保证，在初始化匹配与跟踪过程中的精度就会变的非常低，从而显著降低卷积网络与特征的性能。这是本研究要解决的问题。

2.3.2 不变性特征

为了实现目标特征自学习与模型自学习，需要进行大量的区域关联与匹配，具有不变性的特征表示是研究的前提。在深度学习框架中，为了解决目标的视角与尺度不变性问题，常用的办法是以大量数据驱动深度网络“记忆”各种尺度与视角变换。谷歌公司提出 TI-Pooling[79]算法要求对训练样本进行尺度、旋转、仿射变换等一系列变化，然后采用并行的卷积网络分别学习每一种变化，最后通过权值共享与多示例学习策略来选择输入样本最优的变换进行训练。TI-Pooling 中大量的输入样本使得网络训练变得十分困难，耗费更多计算资源。散射不变卷积网络[80]利用小波散射变换计算具有旋转不变性的图像表示的特点，它具有双层结构，一层输出类似于 SIFT 的结构。第二层得到用于分类的不

变性特征编码，并且采用了在小波转换卷积中加入非线性系数和 Average Pooling 运算。Spatial Transform Network (STN)[81]在 CNN 中引入了一个用于计算样本空间变换参数的子网络，子网络通过学习仿射变换参数实现对样本尺度、平移、旋转变化进行校正。但是，STN 对于复杂背景下、角度较大的变换适应性仍然不强，子网络对于复杂图像空间变换的估计精度不高。

CNN 的卷积特征图呈现出金字塔式的多分辨率结构，但是当目标物体尺度变化较大时，CNN 仍然无法自适应地提取对尺度鲁棒的特征。目前，在深度学习框架内解决尺度适应性的方法主要有以下几种：1）尺度金字塔，如传统特征 SIFT 等。这类方法通过构建多尺度的滤波器金字塔以应对目标的尺度变化。2）图像仿射变换。这类方法不需要提供额外的监督信息，通过一个子网学习出输入图片相对于此目标类别“标准”的仿射变换，然后用其逆变换对输入图片进行矫正之后提取深度特征，以此获得对旋转、尺度、平移的适应性。3）对图片预处理之后提取各种尺度的目标候选框，再针对每个目标候选框提取深度特征。第一类方法的不足在于难以与现有深度学习框架统一，且提取的是区域局部特征，无法构建样本的整体不变性；第二类、三类方法由于对输入进行仿射变化或者尺度枚举，导致运算开销大；如何设计更加合理的卷积神经网络结构以提高深度学习特征的旋转不变性与尺度适应性需要进一步研究。

2.3.3 弱监督目标建模

在过去的二十年间，以 Adaboost 方法为基础的人脸目标检[82]、以 DPM 模型为代表的人体目标检测[83]、以深度学习为代表的多类目标检测方法极大的促进了相关领域的发展[84-87]。这些方法中大都工作依赖于精确的样本类别与位置标注，没有涉及减少监督信息的研究。

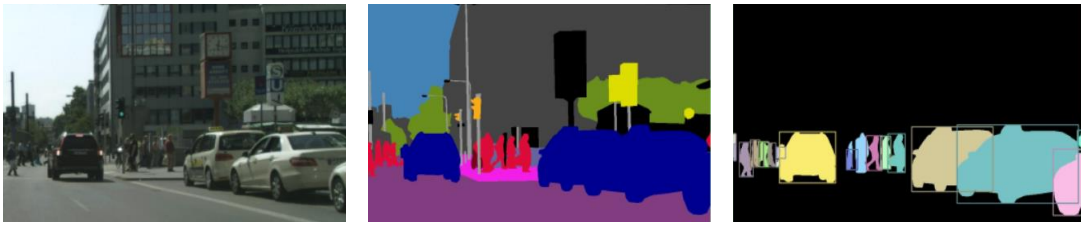
弱监督学习(Weakly Supervised Learning，WSL)通过将目标位置标注简化为图像类别标注，从而大大减少人工标注工作量。在弱监督图像目标建模中，通常假设是不同类别的目标在特征空间内相距比较远，而同类目标能够在特定的特征空间形成一个聚类。在这样的假设下，可以使用图匹配、跟踪、聚类与多示例学习等方法去发现感兴趣的目标的同时学习检测器。

中科院自动化所提出了能处理大规模数据的 MILineal[19]方法，通过信赖域的方法替代传统的梯度方法求解 MIL。MILinear 通过模型得分将 MIL 的正例

“包”中的样本划分为正反例，进而将容易混入正例的反例样本排除，显著改善了目标定位性能。Bilen 等人[57]使用凸聚类来防止陷入错误的标记中。以上两种方法通过引入关于目标先验的正则化项来缓解局部最优问题。为了实现高层语义级的聚类，中科院自动化所采用隐语义空间分析(pLSA)[59]方法并取得更好效果。隐变量支持向量机（LSVM）是另外一种常用的隐变量学习方法。LSVM 将图像中目标样本位置作为隐变量，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建模。与多示例学习一样，LSVM 的目标函数是非凸的，在学习优化过程同样容易出现局部最小值。LSVM 的目标函数是以图像级的分类精度为导向，而不是目标分类精度为导向。这可能导致算法将目标的一部分(Part)当成目标本身，模型很快收敛到局部最优解，造成随机性定位错误。Bilen[60]采用凸聚类正则化提高 LSVM 解的质量，Song[61]采用聚类方法为隐变量学习设定初始解，并通过将目标函数转化为欧拉二次型来增加目标函数的凸性。后续研究中，特征学习与弱监督学习进行了充分结合，其主要思路是将卷积层像素或目标候选区域当成“示例”，将目标定位与特征学习联合求解。但是，局部最小值与随机性定位错误并未从理论上进行解决。

2.4 弱监督语义分割与实例分割

对于目标识别问题而言，与本文内容有较大关联的任务还有弱监督语义分割和示例分割。对于分割任务而言，在监督学习框架中，算法的训练需要更高强度的标注信息，即所有的目标像素均需要被标注，尤其是在示例分割中，同一张图像中的不同示例还需要有不同的标注。



原图　　语义分割　　示例分割

图 2.8 语义分割和实例分割

Figure 2.8 Semantic segmentation and instance segmentation

弱监督语义分割和实例分割[88-92]的任务是要利用弱监督信息来解决像素级的语义分割问题。和本文的问题类似，在训练过程中，弱监督语义分割没有像

素级别的标号，只有图像的类别标注信息，弱监督模型除了需要能找到前景像素外，还需要学会区分背景像素和不同类别的前景像素。全监督的分割算法通常利用图像和其对应的全监督的像素级标注信息训练分割模型，该模型在测试的过程中可以对图像中的每个像素进行分类，从而得到语义分割的结果。

与语义分割不同的是，目标检测只需要用方框定位目标，因此该方法相对而言更加简便，在实际应用场合中对目标的位置要求不是很高的情况下更为实用，并且也能为进一步精细的语义分割或示例分割问题提供良好的初始值。

2.5 本章小结

本章介绍了弱监督目标识别问题中的各种相关技术。首先，针对弱监督方法对全监督方法的继承性，并介绍了全监督下的目标检测方法，然后对弱监督目标检测框架中的各种方法做了详细的综述。针对弱监督目标识别中的特征和建模等问题，单独展开论述。最后，介绍了与弱监督目标检测相关的两个任务：弱监督的语义分割和实例分割。

第3章 最小熵隐变量模型

3.1 问题简介

目标检测任务是要识别并定位出给定未知图像中所有感兴趣的目标。当前的目标检测框架以“预定位+精分类”的方式为主。其中，预定位的过程是根据一定的先验规则如纹理、颜色、前景分类器等，提取图像中所有可能是目标的位置，作为目标的候选区域（Region Proposal），该区域通常以方框的形式表示，因此，目标的候选区域也成为目标候选框（Object Proposal）。预定位的作用是通过先验信息，以尽可能少的目标候选框定位到全部的目标。预定位的引入在极大的减少了目标位置的搜索范围、降低计算代价的同时，也能保证较高的查全率（Recall）。精分类的过程则是通过训练目标候选框的分类器，对预定位生成的目标候选框进行精确分类，从众多的目标候选框中识别出目标，并给定目标的类别。

在全监督的目标检测的训练过程中，目标的位置信息是已知的。目标检测算法在精分类的分类器训练过程中，可以根据目标的位置信息确定每一个目标候选框的类别标号。通过这种监督方式，全监督的目标检测算法取得了巨大的进步。然而，在弱监督目标检测的训练过程中，目标的位置信息是未知的，只有目标的存在与否是已知的。因此，弱监督算法在训练过程中除了需要学习目标候选框的分类器之外，还需要先定位目标的位置。如图 3.1 所示。

隐变量模型是弱监督视觉目标检测的最常用的模型之一。在隐变量模型中，目标候选框为隐变量，在训练过程中标号未知。通过隐变量的引入，隐变量模型能够简历图像标号和目标候选框的关系，并在训练过程中迭代预测隐变量的标号和优化图像分类损失，直至模型收敛。

然而，弱监督目标检测模型（如隐变量模型）的损失函数通常是非凸的，导致该模型容易陷入局部最优；另一方面，由于优化目标往往是最优化图像分类失，这与目标检测的任务并不一致。这种不一致会导致很多局部最优解（如目标的部件等）也能使得图像分类损失达到最小。模型的非凸问题、损失函数

和优化目标不匹配的问题导致模型在训练过程中具有很强的随机性，模型对目标的定位结果可能会在多个局部最优解之间切换，导致目标特征学习不稳定，使得模型最终收敛到错误的位置，如图 3.2 前两行所示。近年来，虽然很多研究人员提出了图像分割、上下文信息、分类器精调等正则项的方式，但是对于如何从原理上降低定位随机性这一问题仍然有待解决。

本章提出了基于候选框团的最小熵隐变量模型（Min-Entropy Latent Model，MELM），通过该模型最小化目标定位的随机性。最小熵隐变量模型受热力学原理的启发，即最小化熵能减少系统的随机性。通过引入最小熵隐变量模型，算法在训练阶段的定位随机性得到了降低，因此能够更稳定的学习目标特征，提升目标定位的准确性，如图 3.2 下面两行所示。

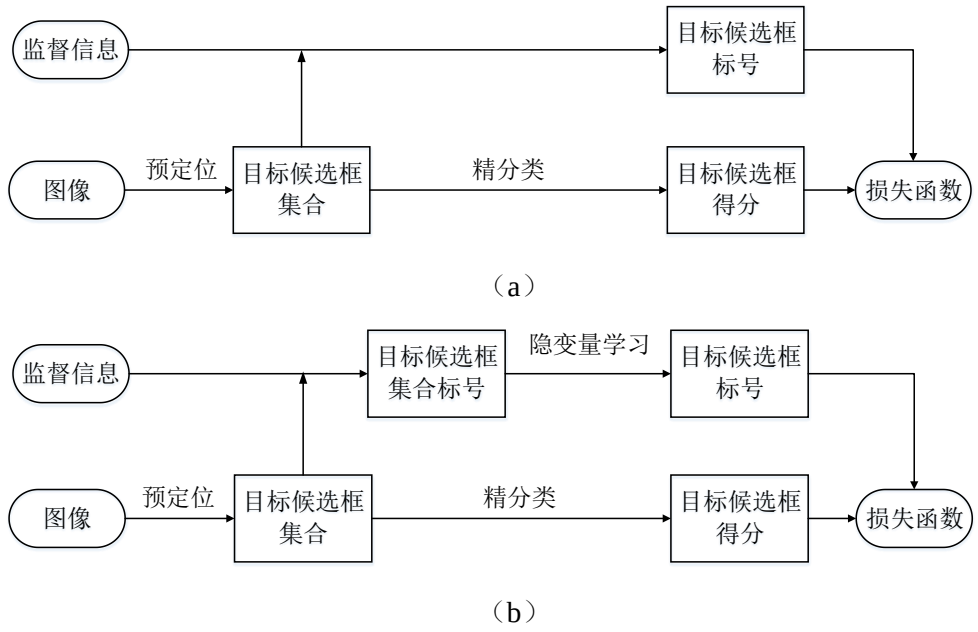


图 3.1 全监督和弱监督目标检测的训练过程对比

Figure 3.1 Comparison of the training procedures of fully and weakly supervised object detection

本章后续两节将详细介绍最小熵隐变量模型的建模和具体分析结果。其中，3.2 节介绍模型构建，3.3 节介绍模型的属性和特征，3.4 节给出模型在深度网络中的实现，3.5 节中呈现具体的实验结果和分析，最后在 3.6 节总结本章内容。

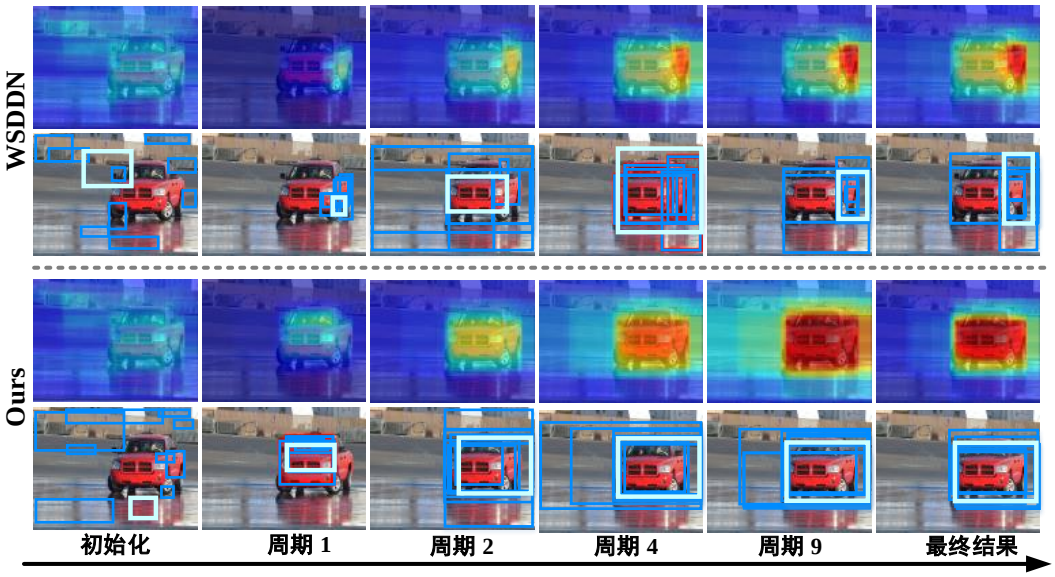


图 3.2 传统弱监督目标检测框架（WSDDN 为例）和本文提出的最小熵隐变量模型训练过程中定位结果的对比。图中第一行和第三行是候选框得分叠加之后的置信度图。第二行和第四行中，蓝（深）色框为得分前十的候选框，白（浅）色框为最高得分框，即定位结果。

Figure 3.2 Comparison of the localization results of WSDDN and our proposed Min-Entropy Latent Model (MELM). The fisrt and third rows are localization confidence map. The second and fourth rows are localization results where the blue boxes are top 10 proposals and white ones are localization results.

3.2 最小熵隐变量模型

最小熵隐变量模型由三部分构成：

（1）候选框团划分模块，用于目标候选框（目标或目标局部）的搜集；

（2）全局最小熵隐模型，用于发现包含目标的候选框团；

（3）局部最小熵隐模型，用于对目标的精确定位，如图 3.3 所示。

其中，候选框团的定义是一个候选框集合，该集合中的候选框相互之间具有空间关联性（空间位置相互重叠）和类别关联性（属于同一个目标类别）。候选框团的引入有助于减少候选框之间的冗余，减小弱监督学习的解空间，从而优化模型的求解过程。结合最小熵隐模型，本文所提出的方法能够搜集目标候选框，并最小化目标定位的随机性、激活更完整的目标区域和压制背景，如图 3.2 所示。

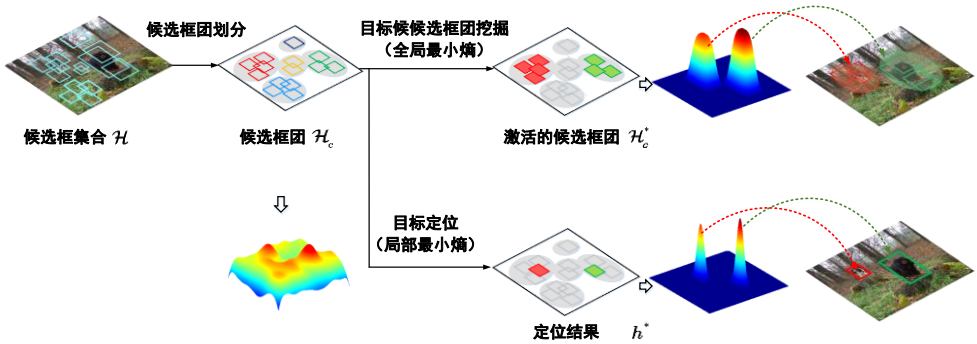


图 3.3 最小熵隐变量模型示意图。首先，候选框团划分模块在大量嘈杂的候选框集合中搜集与目标相关的候选框团；基于这些候选框团，我们定义了全局最小熵模型，其作用是发现目标候选框团；最后，局部最小熵模型对背景进行压制，目标进行精确定位。三个模块在训练过程中迭代优化。

Figure 3.3 Illustration of the min-entropy latent model (MELM). A clique partition module is proposed to collect objects/parts from redundant proposals; Based on the cliques, a global min-entropy model is defined for object clique discovery; Within discovered cliques, a local min-entropy model is proposed to suppress object parts and select true objects. The three components are iteratively performed.

在介绍最小熵隐变量模型之间，先定义相关符号如下：

$x \in X$：图像 x 属于图像数据集合 X。

$y \in Y$：图像标号 y，其取值范围为标号集合 $Y = \{1, 0\}$，其中 $y = 1$ 表示图像中包含感兴趣的目标，即为正例图像；$y = 0$ 表示图像中不包含感兴趣的目标，即为反例图像。

$h \in \mathcal{H}$：候选框 h，其取值范围为候选框集合 $\mathcal{H}$。

$H_c \subseteq \mathcal{H}$：候选框团，是候选框集合的子集。

θ：最小熵隐变量模型的参数。

$E(\cdot)$：最小熵模型。

$L(\cdot)$：损失函数。

根据上述符号，最小熵隐变量模型的方程定义如下：

$$\begin{aligned} \{h^*, \theta^*\} &= \arg\min_{h,\theta} E_{(X,Y)}(h, \theta) \\ &= \arg\min_{h,\theta} E_{(X,Y)}(H_c, \theta) + \lambda E_{(X,Y,H_c)}(h, \theta) \\ &\Leftrightarrow \arg\min_{h,\theta} L_{(X,Y)}(H_c, \theta) + \lambda L_{(X,Y,H_c)}(h, \theta) \end{aligned} \tag{3.1}$$

其中，$E_{(X,Y)}(H_c,\theta)$ 和 $E_{(X,Y,H_c)}(h,\theta)$ 分别是全局最小熵和局部最小熵，优化此方程用于发现包含目标的候选框团并进一步精确定位目标。λ 是正则项因子，其决定了局部最小熵在优化过程中的权重。$L_{(X,Y)}(H_c,\theta)$ 和 $L_{(X,Y,H_c)}(h,\theta)$ 是模型的损失函数，分别基于全局最小熵模型 $E_{(X,Y)}(H_c,\theta)$ 和局部最小熵模型 $E_{(X,Y,H_c)}(h,\theta)$ 定义。

3.2.1 候选框团划分

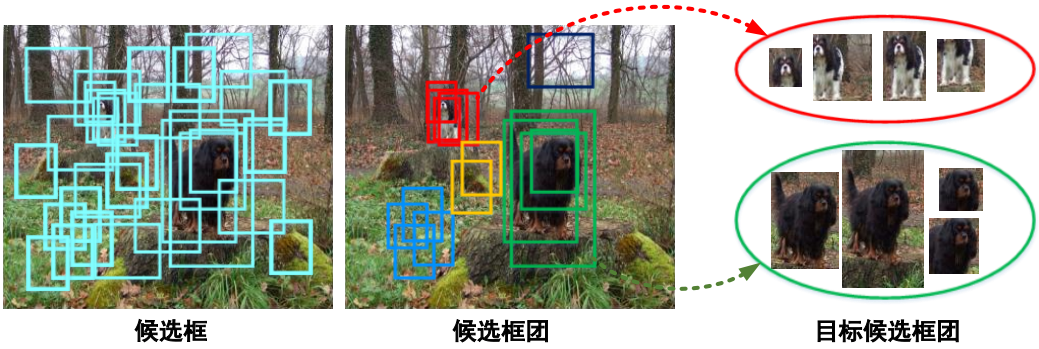


图 3.4 候选框团示意图。首先选取候选框集合中得分高的候选框，并根据其空间位置关系和类别的关系，动态的划分成多个候选框集合。候选框团的作用是搜集目标或目标部件的信息，以激活完整的目标区域。

Figure 3.4 Illustration of the proposal cliques. The proposals of high scores are selected and dynamically partitioned into same cliques if they are spatially related (overlapping with each other) and class related (having similar objectclass scores).Clique partition targets at collecting object/object parts and activating true object extent.

候选框团是一个候选框的集合，是整个候选框集合的子集。候选框团中的候选框相互之间具有空间关联性（空间位置相互重叠）和类别关联性（属于同一个目标类别），如图 3.4 所示。候选框团的作用是搜集目标或目标部件的信息，以激活完整的目标区域。

由于定位随机性通常发生在得分高的候选框之间，在候选框团生成之前，我们首先通过候选框的置信度将低得分的候选框视为背景，只保留得分高的候选框集合 $\tilde{H}$（经验设定为得分最高的 200 个候选框），其中 $\tilde{H} \subseteq \mathcal{H}$。候选框团是高得分候选框集合 $\tilde{H}$ 的最小充分覆盖，其满足如下关系：

$$\begin{cases} \bigcup_{c=1}^{C} H_c = \tilde{H} \\ \forall c \neq c', H_c \cup H_{c'} = \varnothing \end{cases} \tag{3.2}$$

其中$c, c' \in \{1, 2, ..., C\}$，$C$ 是候选框团的个数，其数值随着候选框生成过程动态变化。对于一张输入图像，其候选框团的生成过程分为以下几个步骤：

（1）对所有候选框根据其预测置信度进行排序；

（2）在候选框集合中选取最大置信度的候选框，并以该样本为候选框团的基准候选框；

（3）计算所有候选框和基准候选框的交比并集（IoU），选取 IoU 大于 τ 的候选框，这些候选框和基准候选框形成一个候选框团H_c；

（4）从候选框集合中去除步骤（3）得到的候选框团中的所有候选框，形成新的候选框集合$\tilde{H}$；

（5）回到步骤（2）直至候选框集合$\tilde{H}$为空集。

3.2.2 全局最小熵隐模型

候选框团的引入起到了搜集目标或目标部件的信息的作用，为激活完整的目标区域的提供了良好的保证。然而，在弱监督模型训练过程中，目标候选框团的选取仍然具有随机性，如果模型没有成功的找到包含目标或者目标部件的候选框，那么完整目标区域的激活也会受到限制。因此，我们引入了全局最小熵隐模型，用于以最小的随机性定位到目标候选框团。全局最小熵隐模型的定义如下：

$$\begin{aligned} H_c^* &= \arg\min_{H_c} E_{(X,Y)}(H_c, \theta) \\ &= \arg\min_{H_c} -\log \sum_c p(y, H_c; \theta) \end{aligned} \tag{3.3}$$

其中$p(y, H_c; \theta)$是候选框团H_c属于类别 y 的概率，该概率是基于候选框得分$s(y, h; \theta)$计算得到的，具体公式如下：

$$p\left(y,H_c;\theta\right)=\frac{\exp\left(1/\left|H_c\right|\sum_{h\in H_c}s\left(y,h;\theta\right)\right)}{\sum_c\sum_y\exp\left(1/\left|H_c\right|\sum_{h\in H_c}s\left(y,h;\theta\right)\right)} \tag{3.4}$$

其中，$\left|H_c\right|$表示候选框团H_c中候选框的个数。候选框得分$s\left(y,h;\theta\right)$是通过深度卷积神经网络和目标候选框分支中的全连接层计算得到。

为了保证被发掘的候选框团除了包含目标和目标部件之外，还能用于区分正反例图像，我们进一步提出了类别关联的权重w_{H_c}。候选框的置信度和图像分类的置信度是具有很大的关联的。基于这个先验信息，全局最小熵定义如下：

$$E_{(X,Y)}\left(H_c,\theta\right)=-\log\sum_c w_{H_c}p\left(y,H_c;\theta\right) \tag{3.5}$$

其中，候选框团的权重w_{H_c}定义为

$$w_{H_c}=\frac{p\left(y,H_c;\theta\right)}{\sum_y p\left(y,H_c;\theta\right)} \tag{3.6}$$

公式（3.5）中定义的全局最小熵属于 AD 熵（Acz´el and Dar´oczy (AD) Entropy）家族，并且是可导的。从公式（3.6）可以看出，当$y=1$时，$w_{H_c}\in\left[0,1\right]$。此时$w_{H_c}$和候选框团属于正例的得分是正相关的，并且与其他类别（反例）的得分是负相关的。

根据上述定义，将全局最小熵应用于深度卷积神经网络中的目标候选框团发现分支，其对应的损失函数定义为：

$$\begin{aligned}L_{(X,Y)}\left(H_c,\theta\right)=&yE_{(X,Y)}\left(H_c,\theta\right)\\&-\left(1-y\right)\sum_h\log\left(1-p\left(y,h;\theta\right)\right)\end{aligned} \tag{3.7}$$

对于正例图像而言，即当$y=1$时，公式（3.7）中的第二项等于 0，此时只有全局最小熵模型被优化；对于反例图像，即当$y=0$时，公式（3.7）中的第一项为 0，此时公式中的第二项（图像分类损失）被优化。在训练阶段，图像分类和目标定位熵均得到了优化，这使得模型不仅能够正确分类图像，同时还降低了训练过程中候选框团定位的随机性。

3.2.3 局部最小熵隐模型

由全局最小熵隐模型挖掘的目标候选框团为最终目标的定位提供了良好的初始解，但是，该初始解仍然可能包含随机噪声如目标部件或包含了背景的目标部件等。原因在于全局最小熵隐模型的目标方程，也就是公式（3.7）中，模型的最终目标是选择最具有判别性的候选框团去区分图像的类别，而没有关注目标的定位是否准确。

我们进一步提出了局部最小熵隐模型用于对目标的位置精确定位。局部最小熵隐模型首先利用全局最小熵隐模型挖掘的目标候选框团，选取得分最高的候选框，如下述公式所示：

$$h^{*}=\arg\min_{h\in H_c^{*}} E_{\left(X,Y,H_c^{*}\right)}\left(h,\theta\right) \tag{3.8}$$

其中，局部最小熵定义如下：

$$E_{(X,Y,H_c)}\left(h,\theta\right)=-\sum_{h\in\Omega_{h^{*}}} w_h\, p\left(y,h;\theta\right)\log p\left(y,h;\theta\right) \tag{3.9}$$

局部最小熵同样属于 AD 熵家族且可导。在全局最小熵，即公式（3.5）中，模型通过对全部候选框团的得分概率加和的结果来预测图像类别的概率。与公式（3.5）不同的是，局部最小熵——公式（3.9）的作用是在局部范围区分每个候选框是否是目标或背景。w_h 是候选框 h 的权重，其定义如下

$$w_h=\frac{\sum_{h\in\Omega_{h^{*}}} g\left(h,h^{*}\right)p\left(y,h;\theta\right)}{p\left(y,h;\theta\right)\sum_{h\in\Omega_{h^{*}}} g\left(h,h^{*}\right)} \tag{3.10}$$

其中 Ω_h^{*} 表示最高得分候选框 h^{*} 的领域。$g\left(h,h^{*}\right)=e^{-a\left(1-O\left(h,h^{*}\right)\right)^2}$ 是一个高斯核函数，其参数为 a。$O\left(h,h^{*}\right)$ 表示候选框 h 和 h^{*} 的交集比并集（Intersection over Union， IoU）。$O\left(h,h^{*}\right)$ 的值越大，也就是候选框 h 和 h^{*} 的空间位置接近时，高斯核函数 $g\left(h,h^{*}\right)$ 的值也越大；反之，当 $O\left(h,h^{*}\right)$ 的值越小，也就是候选框 h 和 h^{*} 的空间位置远时，高斯核函数 $g\left(h,h^{*}\right)$ 的值也越小。可以看出，公式（3.10）本质上是定义了一种“软”候选框标号赋值策略。从实验验证可以观察到，该形式的标号赋值策略相比“硬”标号赋值策略，也就是阈值策略，其对噪声更

加鲁邦。

根据上述公式，目标的定位损失函数定义为

$$L_{(X,Y,H_c)}(h,\theta)=E_{(X,Y,H_c^*)}(h,\theta) \tag{3.11}$$

根据公式（3.10）中定义的权重 w_h，和 h^* 在空间位置上比较接近的候选框将趋向于和 h^* 同一类别；而离 h^* 较远的候选框则更倾向于是成为反例或者难反例。优化损失函数，也就是公式（3.11）的结果是，随着局部最小熵在训练过程中逐渐变小，H_c^* 中目标候选框的定位结果越来越稀疏且背景逐渐被压制。

3.3 网络结构和实现

最小熵隐变量模型的实现结合了深度卷积神经网络。如图 3.4 所示。首先通过卷积神经网络对全图提取特征，并在最后一个卷积层利用 ROI-Pooling 和两个全连接层对每一个候选框提取特征，其中候选框提取算法为 Selective Search 算法。以候选框特征为输入，后续加入一个候选框团划分模块和两个网络分支。第一个分支为候选框团挖掘分支，该分支利用全局最小熵隐模型，降低候选框团定位的随机性，并优化图像分类的损失函数，使得在图像分类达到最优的同时，模型能够挖掘到目标候选框团。第二个分支为目标定位分支，其利用局部最小熵隐模型对目标的位置进行定位，通过将候选框划分为目标和难反例的形式，在训练过程中逐渐学会判别正反例候选框。

在优化过程中，最小熵隐变量模型引入了循环学习的策略，如图 3.5 所示。该策略将在 3.4 小节中详细介绍。在网络的前向传播过程中，网络能够学习到稀疏并稳定的目标候选框；在网络的反向传播过程中，网络根据所选候选框的梯度更新参数。通过循环学习的方式，在目标定位分支中的目标得分以循环的方式返回到候选框团挖掘分支中。在测试阶段，候选框团挖掘分支被删除，只保留候选框定位分支，因此测试阶段网络的速度稍快于监督框架中的 Fast RCNN。

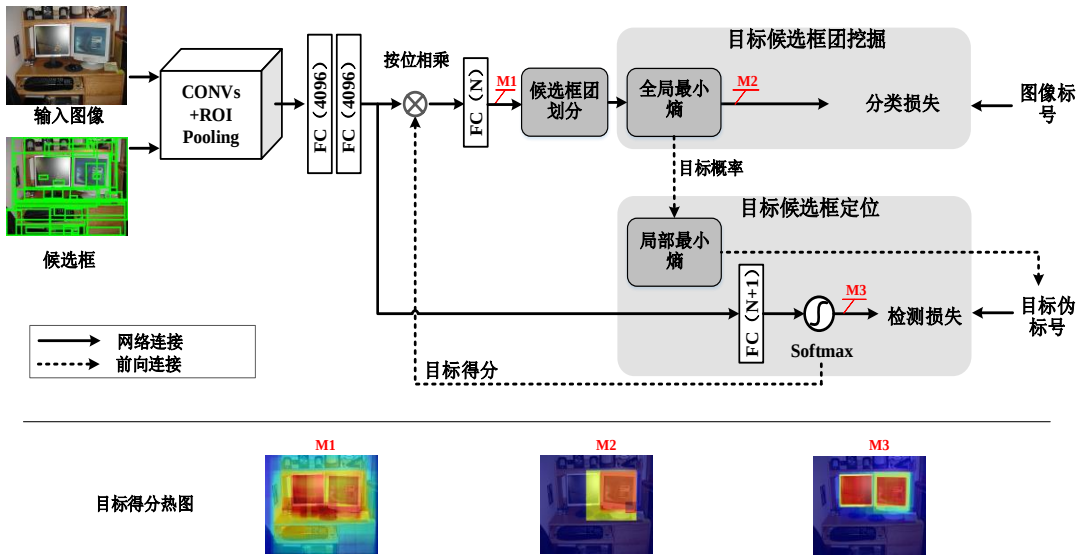


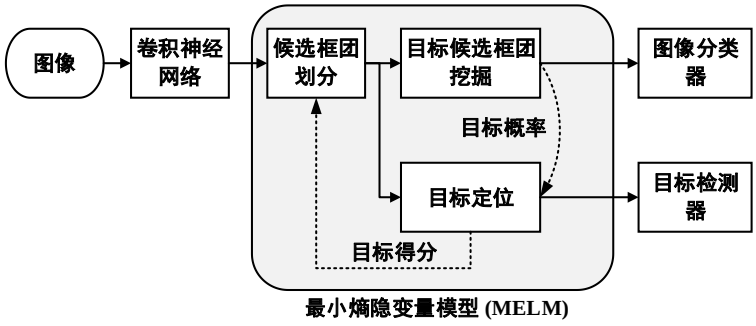
图 3.5 最小熵隐变量模型在深度学习框架中的实现。该框架由基网、候选框团划分模块以及两个用于候选框团挖掘和候选框定位的分支构成。该两分支的结构融合了特征学习，并加入了循环优化的策略。“M1”、“M2”、“M3”分别表示没有熵时的候选框得分图、全局和局部最小熵的候选框得分图。N 表示目标类别数。

Figure 3.5 The min-entropy latent model (MELM) is deployed as a clique partition module and two network branches for object clique discovery andobject localization. These two network branches are unified with feature learning and optimized with a recurrent learning algorithm. “M1”, “M2” and“M3” are heatmaps about proposal scores without min-entropy, with global min-entropy, and with local min-entropy, respectively. N is the number of object categories.

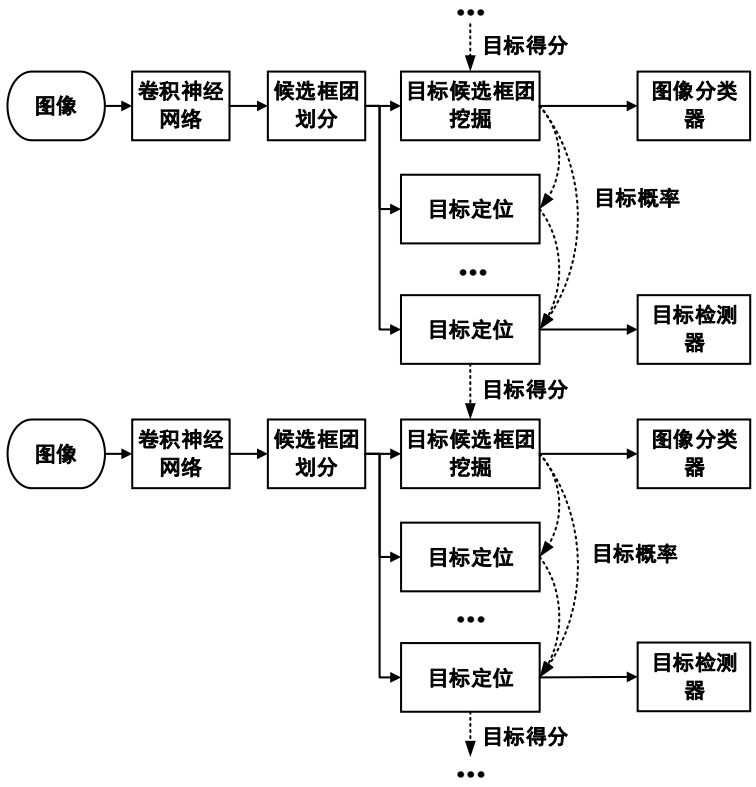
最小熵隐变量模型的目标是在训练的过程中将图像类别的监督信息迁移到目标位置，并且在最小熵约束的前提下降低定位随机性。训练过程中，我们引入了循环学习的优化方式，是的图像类别监督信息和目标位置之间的信息迁移更加鲁棒。

循环学习：循环学习的流程图如图 3.6（a）所示。循环学习通过前向传播和反向传播的的过程将图像类别监督信息逐渐迁移到目标位置。在前向传播过程中，通过全局最小熵隐模型的学习，网络挖掘到目标候选框团，并将候选框团中最大得分的候选框视为伪标号。

3.4 模型优化



（a）



(b)

图 3.6 循环学习。（a）循环学习流程图；（b）展开后的累加循环学习示意图。其中实现箭头表示网络连接，虚线箭头表示该连接只前向传播，不反向传播。

Figure 3.6 The flowchart of (a) the proposed recurrent learning algorithm and (b) unfolded accumulated recurrent learning algorithm. The solid lines denote network connections and dotted lines denote forward-only connections.

通过这种方式，图像类别标号的信息被迁移到了目标位置；通过局部最小熵的学习，模型学习到更具有判别性的候选框分类模型，从而进一步加强了前一个分支迁移得到的目标位置信息，得到了新的目标得分。在循环学习中，我

们将该新的目标得分重新传递回前一个分支，具体的方式是将该得分作为权重以按位相乘的方式对相应的候选框进行加权。通过这种方式，目标的判别性被引入到全局最小熵模型中，从而使得全局最小熵模型能够在考虑到目标判别性的同时学习图像分类，进而挖掘到新的目标候选框团。在反向传播过程中，目标候选框团挖掘分支和目标候选框挖掘分支在随机梯度下降的框架中被联合优化。该过程如算法 1 所示::

算法 1：循环学习

输入：图像 $x \in X$，图像标号 $y \in Y$ 以及候选框 $h \in \mathcal{H}$。

输出：网络参数 θ。

1. 初始化所有候选框的目标得分 $s(h)=s(y,h;\theta)=1$；
2. 通过网络前向传播计算候选框特征 ϕ_h；
3. $\phi_h \leftarrow \phi_h \cdot s(h)$
4. 候选框团划分：
5. $H_c \leftarrow$ 公式（3.2） //划分候选框
6. 候选框团挖掘：
7. $H_c^* \leftarrow$ 公式（3.5） //通过全局最小熵计算目标候选框团
8. $L_{(X,Y)}(H_c,\theta) \leftarrow$ 公式（3.7） //计算损失函数
9. 目标候选框定位：
10. $h^* \leftarrow$ 公式（3.8） //通过局部最小熵计算目标候选框
11. $L_{(X,Y,H_c)}(h,\theta) \leftarrow$ 公式（3.11） //计算损失函数
12. 网络参数更新：
13. $\theta \leftarrow$ 公式（3.7）和公式（3.11）
14. $s(h) \leftarrow$ 更新之后的网络参数 θ
15. 如果没有达到迭代终止条件，则返回第 4 行。

累加循环学习：累加循环学习的流程如图 3.6（b）所示。和循环学习的区别是，累加循环学习通过引入多个目标定位分支，其中，每个分支都能够挖掘到一个目标候选框的结果。在前向传播的过程中，每个分支挖掘到的结果都累加到下一个分支，作为下一个分支的伪标号。由于各个分支之间挖掘到的候选框可能会不同，因此，累加循环学习不仅能够定位到图像中的多个目标，并且还能通过增加候选框样本的多样性来增强弱监督模型的鲁棒性。

3.5 模型分析

通过引入候选框团的划分和循环学习策略，最小熵隐模型可以看做是一种基于连续优化的方法，该方法有助于求解弱监督目标检测中的非凸优化问题。

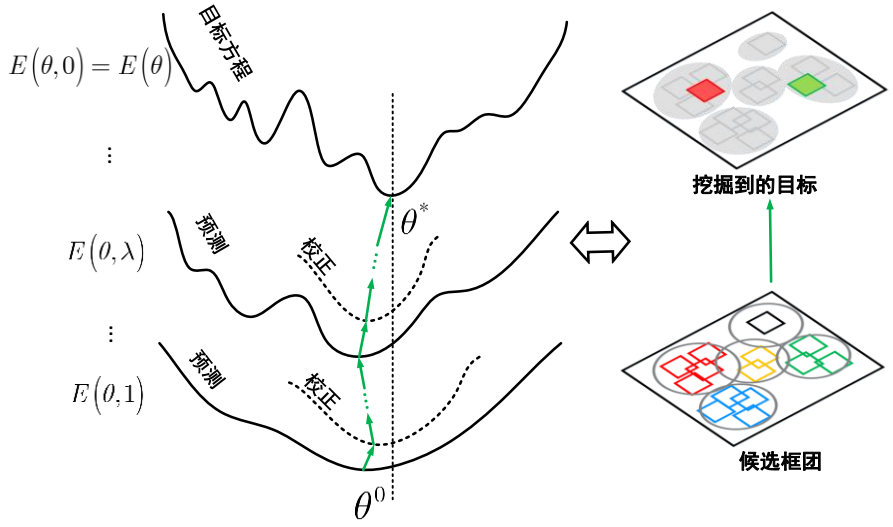


图 3.7 渐进优化示意图。通过目标候选框团挖掘（预测过程）和目标候选框定位（校正过程），原非凸函数的优化问题变成了一个逐渐近似求解的过程。该过程以基于候选框团的模型为起点逐渐准换成原来的基于候选框的目标方程。其中，近似的目标方程由候选框团划分构建，其近似的方式是对候选框团中的候选框的概率取平均。

Figure 3.7 Continuation optimization. With object clique discovery (prediction) and object localization (correction), the non-convex optimization problem is turned into a proximate problem, which has a smoothed function and is easier to be optimized. The smoothed function is achieved by reducing the solution space from thousands of proposals (denoted by square boxes) to tens of cliques (denoted by circles) in each image and averaging the class probability of all proposals in each clique.

对于一个复杂的非凸优化问题 $E(\theta)$，其中 θ 表示模型的参数。优化 $E(\theta)$ 的目的是为了求解全局最优解 θ^*，如下式所示

$$\theta^* = \arg\min_{\theta} E(\theta) \tag{3.12}$$

然而，直接优化公式（3.12）容易是模型陷入局部最优。为了能更好的的优化该模型，防止模型过早的陷入局部最优，我们可以引入公式（3.12）的近似方程 $E(\theta;\lambda)$，通过渐进优化的方式来求解该方程，定义如下

$$E(\theta,\lambda)=E(\theta)-\lambda\mathcal{E}(\theta) \tag{3.13}$$

其中参数$\lambda \in [0,1]$控制着近似方程的平滑程度，$\mathcal{E}(\theta)$是校正函数。传统的基于“预测-校正”的渐进优化模型通常需要定义一组函数序列，该序列以$(\theta^0,1)$为起点逐渐向最优解$(\theta^*,0)$还原。其中θ^0为$\lambda=1$时的模型参数。在训练过程中，如果$E(\theta;\lambda)$是$E(\theta)$的平滑，并且和$E(\theta)$的解接近。根据“预测-校正”渐进优化策略，我们将$E(\theta;\lambda)$看做是对$E(\theta)$的一次预测。那么每次预测之后，只需要对预测结果进行一次校正，取弥补预测函数因为平滑而造成的误差即可。这个过程可以通过定义一组预测和校正函数的形式来迭代近似求解原复杂的非凸函数，并最终得到模型的解θ^*。

在最小熵隐变量模型中，公式（3.1）中定义的目标方程是用于求解$\{h^*,\theta^*\}$，如下述公式所示

$$E_{(X,Y)}(H_c,\theta)=E_{(X,Y)}(h,\theta)-\lambda E_{(X,Y,H_c)}(h,\theta) \tag{3.14}$$

这个公式（3.13）中定义的目标方程一致。$E_{(X,Y)}(H_c,\theta)$是基于候选框团定义的，该方程可以看做是原目标方程$E_{(X,Y)}(h,\theta)$的近似，如图 3.7 所示。该近似是通过候选框团划分构建，并对候选框团中的候选框的概率取平均的方式完成的。这样通过将候选框的发掘问题转换成候选框团的挖掘问题，减小了模型的求解空间，并通过平均目标概率的方式平滑了目标方程。

在定义了近似的目标方程之后，我们可以求解到原来的目标方程的近似解，通过循环的调用预测-校正渐进优化方式，我们对近似解不断校正。由于近似方程$E_{(X,Y)}(H_c,\theta)$和原方程$E_{(X,Y)}(h,\theta)$的差异主要体现在近似方程的目标是挖掘候选框团而后者则是挖掘候选框，并且后者的定位结果被包含于前者的定位结果中，该误差可以通过定义校正方程$E_{(X,Y,H_c)}(h,\theta)$得到解决。校正方程的作用是辨别候选框团中的目标候选框和背景候选框。通过两个步骤的结合，原来的目标方程在循环优化的过程中逐渐被求解。

通过“预测-校正”渐进优化策略，原来的弱监督学习问题被分解为目标候

选框团挖掘（预测）和目标定位（校正）两个子问题。原来的非凸优化问题被转换成一个近似目标函数序列，使得复杂的非凸优化问题更容易被求解。

3.6 实验结果与分析

为了验证本章所提出的最小熵隐模型的有效性，本文使用 VGGF 和 VGG16 作为实验的基础网络，在目前较为常用的目标检测的数据集 PASCAL VOC 2007、2010、2012 以及 ILSVRC 2013 和 MSCOCO 2014 五个数据集中实验了本章提出的方法。后面几个小节将分别介绍相关的实验设定、实验分析以及和 stateof-the-art 方法的对比。

3. 6. 1 实验设定

数据集：PASCAL VOC 数据集一共包括 20 个目标类别。VOC 2007 数据集一共包含 9963 张图像，其中 5011 张图像是训练集和验证集，4952 张图像用于测试集。VOC 2010 数据集包含 19740 张图像，其中 10103 张图像用于训练和验证集，9637 张图像用于测试集。VOC 2012 包含 22531 张图像，其中 11540 张用于训练和验证集，10991 张图像用于测试集。ILSVRC 2013 数据集一共包含 200 个类别的目标，该数据集包含 464278 张图像用于训练集，424126 张图像用于验证集，40152 张图像用于测试集。为了能和之前的工作公平的对比，我们采用了 RCNN 中的划分方式，将 ILSVRC 2013 中的验证集划分成两部分（*val*1 和 *val*2），分别用于训练和测试。相比 PASCAL VOC 数据集而言，ILSVRC 2013 数据集虽然图像总数要更多，单个类别的图像数量要远少于 PASCAL VOC 数据集，因此非常具有挑战。MSCOCO 2014 数据集一共包含 80 个类别的目标，其中包括了多目标，多类别和小目标等挑战。在 PASCAL VOC 和 ILSVRC 2013 数据集中，我们采用了平均准确率（mean Average Precision, mAP）的评测方式。在 MSCOCO 2014 数据集中，我们采用了基于多个 IoU 的 mAP 评测方式。

评测标准：本章用到的评测标准有三种：mAP，CorLoc 和 mAP@IoU。其中 PASCAL VOC 数据集目标检测中常用的 mAP 是 mAP@IoU 的特殊情况，即 mAP@0.5。下面将介绍 mAP 和 CorLoc 两个评测标准。

mAP：在介绍 mAP 之前，需要先介绍查全率和准确率。候选框定位结果

的查全率（Recall）就是所有定位结果真正定位到图像数据集中目标站整个数据集中所有目标的比例。查全率越低，说明检测算法漏检的目标越多，反之则漏检的目标越少。而判断候选框是否成功定位到目标的标准如图 3.8 所示。其中，候选框 O 表示目标物体，候选框 P 表示定位结果。图 3.8 左侧中阴影部分的区域 I 即为 O 和 P 两个候选框的交集，图 3.8 右侧中的阴影部分的区域 U 即为 O 和 P 两个候选框的并集。候选框 P 和目标 O 的交比并集（Intersection over Union， IoU）的计算公式如下所示：

$$IoU（O,P）=I/U \tag{3.15}$$

在 PASCAL VOC 数据集中，当 IoU>0.5 时，则认为 P 定位到了目标，此时 P 为正例样本，在 MSCOCO 数据集中，P 定位到目标的标准会根据 IoU 阈值的的大小变化。Recall 和 IoU 的关系可以表示如下：当 Recall = 90%时，数据集中的有 90%的目标被检测算法定位到，即该 90%的目标的定位 IoU 均大于 0.5，而剩下的 10%的目标则为定位 IoU 小于 0.5。

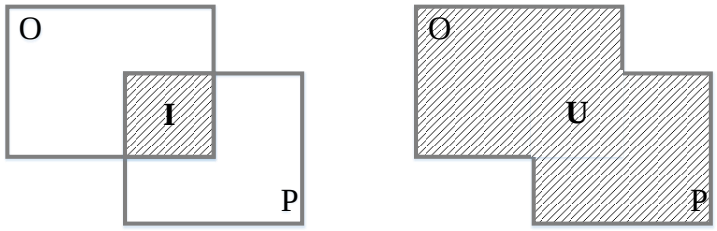


图 3.8 候选框的交集和并集示意图

Figure 3.8 Illustration of intersection and union between two proposals

而准确率（Precision）的计算和 IoU 则有着直接的关联。准确率表示检测算法定位到的所有结果中，与目标候选框的 IoU 大于 0.5 的候选框占所有检测结果的比例。例如整个数据集只有一个检测结果且该结果与目标的 IoU 大于 0.5，则检测算法的准确率为 100%。由此可以看出，准确率的评测标准并不关注整个数据集的目标是否全部被检测到,而是只关注检测器输出结果的正确性，因此，这两个指标经常被一起用于检测性能的评测，即平均性能（Average Precision, AP），检测算法的 AP 的计算过程如下（以 PASCAL VOC 为例）：

（1）计算整个数据集的测试集中图像的所有候选框的检测得分；

（2）对于根据检测得分，对所有候选框按得分从大到小的顺序进行排序；

（3）根据 IoU 的阈值判断候选框是否定位到某类目标，判断标准为 IoU>0.5 为定位正确，反之则为定位失败；

（4）选取所有大于检测得分阈值 δ 的候选框，计算这些候选框的准确率。

（5）将检测得分阈值 δ 按从大到小的顺序不断调整，根据步骤 4 中方法计算出不同阈值下的准确率。

（6）对上述准确率取平均，得到平均性能（AP）。

（7）重复上述所有步骤，计算对数据集中的所有类别的平均性能，对个各类别的平均性能取平均，得到整个数据集的 mAP。

CorLoc：CorLoc 是用于评测训练过程弱监督算法对目标的定位准确性。其计算过程同样基于 IoU，具体计算过程如下：

（1）对于训练集中的某一个类别的图像，计算所有图像的候选框的得分；

（2）对每张图像，取得分最大的候选框作为该图像的定位结果；

（3）如果图像的定位结果和图像中的任何一个目标的 IoU > 0.5，则算法对该图像定位准确。

（4）计算该类别所有图像中，定位准确的图像所占的比例，作为该类别的正确定位（Correct Localization, CorLoc）的性能；

（5）对所有类别的正确定位性能取平均，得到整个数据集的正确定位性能。

预训练模型：预训练模型采用的是目前比较主流的 VGG 网络，分别为 VGGF 和 VGG16。这两个模型均在 ILSCVR 2012 的图像分类任务中预训练。VGGF （VGG-CNN-F）和 AlexNet 的网络结构类似，拥有 5 个卷积层和 3 个全连接层。VGG16 拥有 13 个卷积层和 3 个全连接层。对于这两个基网，我们去掉了最后一个空间最大池化层，用一个 ROI-Pooling 层替代。同时去掉了最后一个全连接层，并使用一个随机初始化的全连接层替代，该全连接层的节点个数和数据集类别个数对应。

候选框生成算法：候选框生成算法采用了 Selective Search 算法和 Edge Boxes 算法。对于每张图像，算法大概生成了 2000 个左右的候选框。对于 Selective Search 算法，我们使用了其 fast 模式生成候选框。在训练过程中，我们去掉了宽或高小于 20 个像素的候选框。

训练参数：和众多已发表的弱监督目标检测算法一样，我们采用了多尺度训练策略，在训练过程中将输入图像的长或者宽随机缩放至下述五个尺度中的一个：{480,576,688,864,1200}。同时，训练图像还会被随机左右翻转。在测试的时候，我们将所有的五个尺度，包括翻转之后的图像共计 10 张图像的检测结果取平均，得到最终的检测结果。在循环学习过程中，我们使用了随机梯度下降（SGD），其中动量参数为 0.9，权重衰减系数为 5e-4，单次输入图像的数量（Batch Size）为 1。模型在整个数据集上一共迭代 20 个周期，其中前 15 个周期的学习率是 5e-3，后 5 个周期的学习率是 5e-4。

3.6.2 候选框团的影响和分析

由图 3.9（a）可以看出，候选框团挖掘过程中，属于背景的候选框团得到了抑制，而包含目标或目标局部的候选框团保留下来。通过该方式，目标候选框团，包括目标区域和目标局部均在训练过程中参与网络的正反传，从而保证了完整的目标区域被激活，如图 3.9（b）所示。

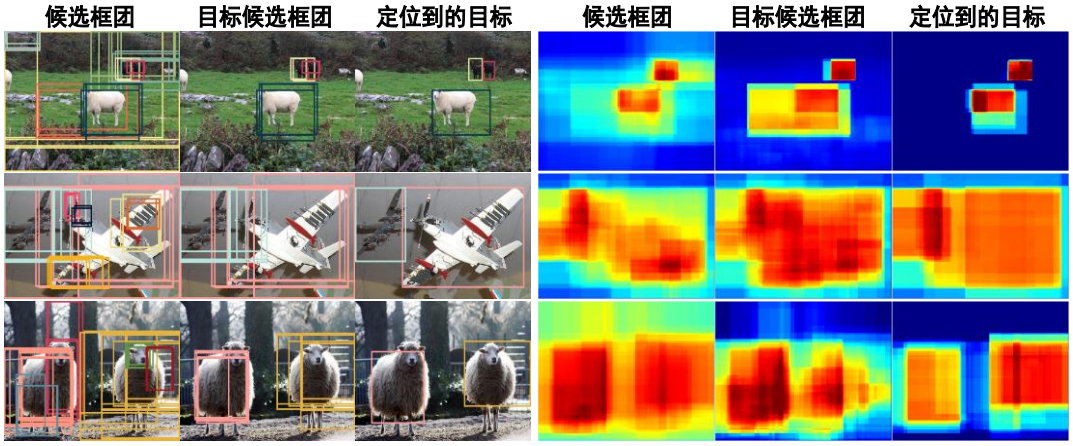


图 3.9 候选框团挖掘和定位可视化。（a）不同颜色的框属于不同的候选框团，（b）候选框团和候选框的得分图。

Figure 3.9 Visualization of the clique partition, object clique discovery, and object localization results. (a) Bounding boxes of different colors denote proposals from different cliques. (b) Score maps of cliques and objects.

以包含目标或目标局部的候选框团为基础，目标定位分支进一步压制背景和难反例，最终定位到完整的目标图 3.9 呈现了训练过程中目标候选框团的演变过程。可以看出，在训练初期（Epoch 2），目标候选框包含了包括目标和目

标部件在内的候选框。通过该方式，完整目标区域的激活得到了保证。在训练过程中，随着最小熵优化的迭代，目标候选框团中的背景和目标部件逐渐被抑制（Epoch 4）。最后，最小熵隐模型成功了抑制了背景和目标部件，并定位到了完整的目标（Epoch 20）。

3.6.3 定位随机性分析

本小节对定位随机性的分析包括两个方面。第一个方面是从量化的角度去评价训练过程中的定位随机性，以及训练过程中最小熵优化的各项与定位随机性相关的数据和指标。第二方面则是通过可视化训练过程中定位结果的变化，对定位随机性做定性分析。

定位随机性的定量分析：对随机性的分析中，我们用到了以下几个量，分别为熵、梯度、定位准确性和定位方差。图 3.10 呈现了上述四个量在 PASCAL VOC 2007 数据集中训练和验证集的训练最小熵隐变量模型的演变过程。从图 3.10（a）可以看出，在网络优化的过程中，全局和局部最小熵均在不断降低，这也意味之定位随机性在不断降低。

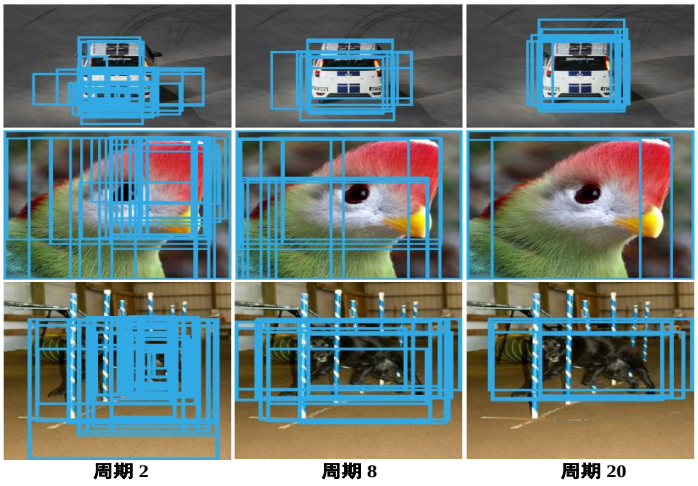


图 3.10 候选框团在训练过程中的演变

Figure 3.10 Evolution of cliques during training

图 3.10（b）中呈现的是全局和局部最小熵分支中的全连接层的梯度，从梯度变化的角度可以看出两个分支对网络特征在不同阶段的影响是不用的。在训练初期，全局最小熵模型的梯度要大于局部最小熵的梯度，此时网络主要聚焦于优化全局最小熵，迅速压制背景并找到目标候选框团，以优化图像分类的损

失；在训练后期，全局最小熵的梯度逐渐降低，此时局部最小熵越来越占据主导地位，这意味着在训练后期，网络的主要聚焦于优化局部最小熵，以训练目标定位模块（检测器）。

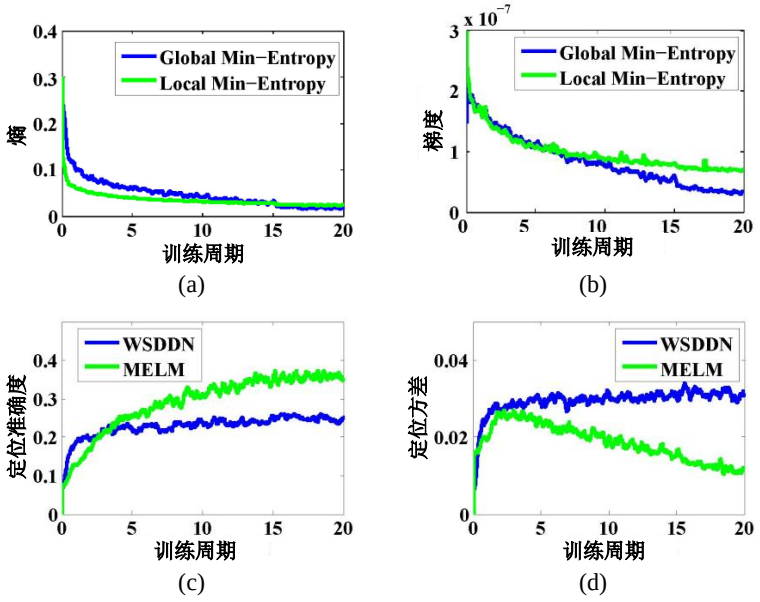


图 3.11 PASCAL VOC 2007 数据集的训练和验证集中，熵、梯度、定位结果在训练过程中的演变。（a）熵的演变过程；（b）梯度的演变过程；（c）定位准确性的演变过程；（d）定位方差的演变过程。

Figure 3.11 Entropy, gradient, and localization on the PASCAL VOC 2007 trainval set. (a) The evolution of entropy. (b) The evolution of gradient. (c) Localization accuracy. (d) Localization variance.

为了更全面的分析定位随机性在训练过程中的变化和影响，本文进一步评测了模型在训练过程中的定位准确度和定位随机性。其中定位准确度的计算过程是对高得分的候选框是否准确，准确性的度量是该候选框与目标的标注候选框之间的 IoU，IoU 越大表示该候选框定位越准确。对于每张图像，我们将所有高得分的候选框的定位准确性取加权平均，其中权重是候选框的概率 $p(y,h;\theta)$，由此得到该图像的定位准确性。定位方差则是定位准确性的加权方差，该权重为 $p(y,h;\theta)$。图 3.10（c）和（d）中呈现了训练过程中定位准确性和定位方差对的演变过程，对比的方式是 WSDDN。从图中可以看出，最小熵隐变量模型（MELM）比 WSDDN 的定位准确性更高，并且定位方差要比 WSDDN 更低，这个优势在训练后期更加明显。这个现象非常直观的说明了本章所提出的方法

在引入最小熵隐模型之后，定位随机性的问题得到了很大的改善，模型在定位的过程中定位结果更加鲁邦，准确性也得到了提升。

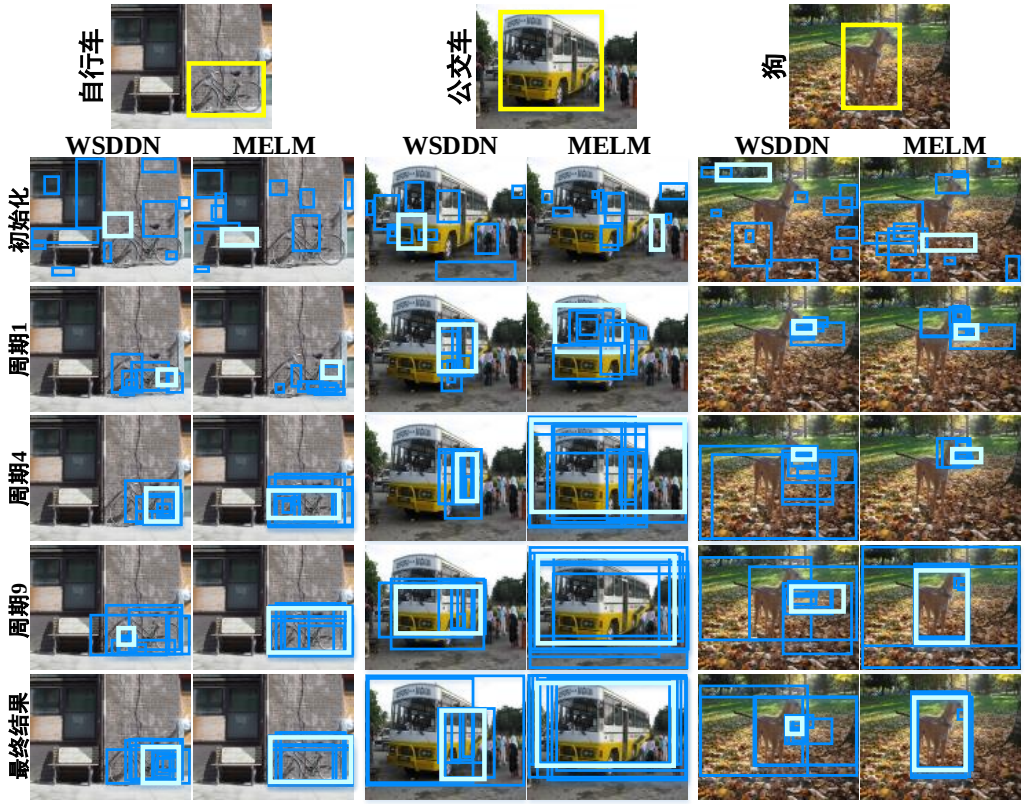


图 3.12 MELM 和 WSDDN 的定位结果的对比。其中，黄色方框表示标注位置，蓝色方框为高得分候选框，白色方框为定位结果。从图中可以看出，WSDDN 由于定位随机性较大，导致其定位结果不稳定，最终定位失败。而本章所提出的方法通过最小熵优化，降低了定位随机性，定位结果更加一致，最终成功定位到目标。

Figure 3.12 Comparison of the learned object locations by WSDDN and the proposed MELM. The yellow boxes in the first column denote groundtruth objects. The white boxes denote the learned object locations and the blue boxes denote the high-scored proposals. It can be seen that for WSDDN the learned object locations evolved with larger randomness, i.e., switch among the proposals around the objects. In contrast the object locations learned by MELM are more consistent with each other with less randomness.

定位随机性的定性分析：通过可视化定位结果，我们进一步对比了本章所提出的方法和 WSDDN 的定位随机性结果。图 3.11 中给出了三张图像的定位结果随着训练过程的演变过程。从图像可以看出本章所提出的最小熵隐变量模型极大的降低了定位结果的随机性，并取得相比 WSDDN 更为准确的定位结果。以图 3.11 中的“Bicycle”为例，在训练初期，WSDDN 和最小熵隐变量模型均

定位失败了。但是随着训练的进行，最小熵隐变量模型通过降低了定位随机性，逐渐定位到目标；而 WSDDN 收到定位随机性的影响，其定位结果不断的在目标和目标局部之间切换，导致其最终定位失败。

3.6.4 模型拆解分析

表 3.1 PASCAL VOC 2007 测试集上的目标检测性能。最小熵隐变量模型的拆解实验。

Table 3.1 Detection mean average precision (%) on the PASCAL VOC 2007 test set. Ablation experimental results of MELM

CNN	Method	mAP
VGGF	MELM-base	31.5
	MELM-base+Clique	33.9
	MELM-D	33.6
	MELM-L	36.0
	MELM-D+RL	34.1
	MELM-L+RL	**38.4**
VGG16	MELM-base+Clique	29.5
	MELM-D	32.6
	MELM-L	40.1
	MELM-D+RL	34.5
	MELM-L+RL	42.6
	MELM-D+ARL	37.4
	MELM-L1+ARL	46.4
	MELM-L2+ARL	**47.3**

模型的拆解分析如表 3.1 所示，下面将从候选框团的影响、最小熵模型拆解分析、循环优化、累加循环优化这四个方面介绍各个模块对模型的影响。

Baseline：基准（baseline）方法是对公式（3.7）中定义的全局最小熵简化版，其可以看做是没有空间正则项的 WSDDN。该方法只有一个损失函数，就是最小化图像分类的损失函数。基准方法在表格中用“MELM-base”表示，在使用 VGGF 作为基网的时候，该方法的性能是 31.5%。

候选框团的影响：通过将候选框根据其空间位置和类别关联划分成候选框团之后，我们将基准方法“MELM-base”改进成“MELM-base+Clique”。由表 3.1 可以看出，当引入了候选框团策略之后，弱监督模型的性能从 31.5%提升到了 33.9%（提升 2.4%）。这是因为候选框团的引入降低了隐变量求解的解空间，减少了候选框之间的冗余，因此有助于定位到更好的目标位置。对于 3.2.1 节中

定义的候选框团划分阈值 τ，我们对其做了分析实验，实验结果如表 3.2 所示，由此可以看出，当 τ 取值范围在[0.5~0.7]之间时，检测性能最好。在后续的实验中，τ 的取值都设定为 0.7。

表 3.2 PASCAL VOC 2007 验证集上的目标检测性能。候选框团划分阈值 τ 对最小熵隐变量模型影响的实验结果。

Table 3.2 Detection mean average precision (%) on the PASCAL VOC 2007 *val* set. Performance with different clique sizes (controlled by τ) of MELM.

τ	0.1	0.3	0.5	0.7	0.9	1
mAP	32.6	34.3	34.4	**35.3**	33.5	34.4

最小熵模型拆解分析：我们将最小熵隐变量模型的目标候选框团挖掘和目标候选框定位两个分支分别表示成“MELM-D”和“MELM-L”。在训练过程中，只讲两个分支级联，而不是用循环优化策略。从表 3.1 可以看出，当使用 VGGF 时，目标候选框团挖掘和目标候选框定位两个分支分别取得了 33.6%和 36.0%的性能这比基准模型的性能分别提升了 2.1%和 5.5%；当使用 VGG16 时，“MELM-L”对“MELM-base-Clique”提升非常显著，检测性能从 29.5%提升到了 40.1%，一共提升了 10.6%。这充分验证最小熵模型的引入对弱监督检测模型起到了非常关键的作用。

循环优化：循环优化的结果在表 3.1 中被表示成“MELM-D-RL”和“MELM-L-RL”，分别对应了目标候选框团挖掘和目标候选框定位两个分支在循环优化框架中的检测性能。从表 3.1 可以看出，这两个分支分别取得了 34.1%和 38.4%的性能。相比不使用循环优化的“MELM-D”和“MELM-L”，循环优化的引入分别提升了两个分支 0.5%和 2.4%的性能。当使用 VGG16 作为基网时，“MELM-D-RL”和 “MELM-L-RL”两个分支分别取得了 34.2%和 42.6%的性能。相比不使用循环优化的“MELM-D”和“MELM-L”，循环优化的引入分别提升了两个分支 1.9%和 2.5%的性能。这些性能提升也验证了循环学习的有效性，如图 3.6 所示。循环学习通过循环交换两个分支的定位置信度，使得两个分支的性能同时得到了提升。

累加循环优化：累加循环学习的结果在表 3.1 中被表示成“MELM-D-ARL”、“MELM-L1-ARL”和“MELM-L2-ARL”，其分别对应了目标候选框团挖掘和多个目标候选框定位分支。在训练过程中，前一个分支的最高得分以伪标注的形式传累加递到下一个分支。当使用两个目标候选框定位分支时，“MELM-L1-ARL”的性能达到了 46.4%，相比“MELM-L-RL”提升了 3.8%。而“MELM-L2-ARL”的性能进一步提升到了 47.3%。

3.6.5 实验结果和对比

PASCAL VOC：对于 PASCAL VOC 数据集，我们采用了目标检测性能，目标定位性能和图像分类性能的评测对比。

弱监督目标检测：目标检测的性能对比见表 3.3。我们将提出的最小熵隐变量模型和最新已发表的方法做了对比。在 PASCAL VOC 2007 数据集上，本章所提出的最小熵隐变量模型使用 VGGF 和 VGG16 分别取得了 38.4%和 47.3%的性能。其中，使用 VGG16 模型所得到的的性能比最新的方法 OICR，Self-Taught，WCCN，WeakRPN 和 TS2C 等方法分别提升了 by 6.1% (47.3% vs. 41.2%), 5.6% (47.3% vs. 41.7%), 4.5% (47.3% vs. 42.8%), 3.0% (47.3% vs. 44.3%) 和 2.0% (47.3% vs. 45.3%)。对于非常具有挑战的弱监督目标检测任务而言，这些性能上的提升十分显著。同时，我们也评测了多模型的装配（MELM-Ens.），该方法使用了 VGGF 和 VGG16 的结果的平均，取得了 47.8%的性能。相比于 OICR 的多模型装配 OICR-Ens，MELM-Ens 的性能要高出 5.8%。为了更进一步的验证我们的方法的性能，我们将 VGG16 模型的检测结果当成标注信息，用 ResNet101 网络为基网，训练了 Fast-RCNN 检测器并取得了 49.0%的性能。

在表 3.4 中，我们在 VOC 2010 和 VOC 2012 两个数据集上对比了实验性能。可以看出，最小熵隐变量模型基本上超过了所有的对比方法。在 VOC 2010 数据集中，当使用 VGGF 时，大幅超过了 WCCN 7.5% (36.3% vs. 28.8%)的性能；当使用 VGG16 时，检测结果是可比的。在 VOC 2012 数据集上，当使用 VGGF 时，分别超过 WCCN 和 OICR 8.0%和 1.8%；当使用 VGG16 时，超过 WCCN，Self-Taught，OICR，TS^2C 4.5% (42.4% vs. 37.9%), 4.1% (42.4% vs. 38.3%), 4.5% (42.4% vs. 37.9%) 和 2.4% (42.4% vs. 40.0%)。在 VOC 数据集的

20 个类别中，“bicycle”、“cow”、“dining-table”和“dog”这几个类分别被大幅提升 4.5%、8.5%、14.7%和 9.6%。实验充分说明了模型的有效性。

表 3.3 PASCAL VOC 2007 测试集上的检测性能。最小熵隐变量模型和最新方法的对比。

Table 3.3 Detection mean average precision (%) on the PASCAL VOC 2007 test set. Comparison of MELM to the state-of-the-arts.

CNN	*Method*	*aero*	*bike*	*bird*	*boat*	*bttle*	*bus*	*car*	*cat*	*char*	*cow*	
VGGF	MILlinear[19]	41.3	39.7	22.1	9.5	3.9	41.0	45.0	19.1	1.0	34.0	
	Multi-fold[15]	39.3	43.0	28.8	20.4	8.0	45.5	47.9	22.1	8.4	33.5	
	PDA[59]	49.7	33.6	30.8	19.9	13.0	40.5	54.3	37.4	**14.8**	39.8	
	LCL+Cont[41]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	
	WSDDN[57]	42.9	56.0	32.0	17.6	10.2	61.8	50.2	29.0	3.8	36.2	
	ContextNet[62]	**57.1**	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	
	WCCN[58]	43.9	**57.6**	**34.9**	**21.3**	14.7	**64.7**	52.8	34.2	6.5	41.2	
	OICR[63]	53.1	57.1	32.4	12.3	15.8	58.2	56.7	**39.6**	0.9	44.8	
	MELM	56.4	54.7	30.9	21.1	**17.3**	52.8	**60.0**	36.1	3.9	**47.8**	
VGG16	WSDDN[57]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	
	PDA[59]	54.5	47.4	**41.3**	20.8	**17.7**	51.9	63.5	46.1	21.8	57.1	
	OICR[63]	**58.0**	62.4	31.1	19.4	13.0	65.1	62.2	28.4	24.8	44.7	
	Self-Taught[61]	52.2	47.1	35.0	26.7	15.4	61.3	66.0	**54.3**	3.0	53.6	
	WCCN[58]	49.5	60.6	38.6	**29.2**	16.2	**70.8**	56.9	42.5	10.9	44.1	
	TS2C[66]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	
	WeakRPN[64]	57.9	70.5	37.8	5.7	21.0	66.1	69.2	59.4	3.4	57.1	
	MELM	55.6	**66.9**	34.2	29.1	16.4	68.8	**68.1**	43.0	**25.0**	**65.6**	
Ens.	OICR-Ens.[63]	58.5	63.0	35.1	16.9	17.4	63.2	60.8	34.4	8.2	49.7	
	MELM-Ens.	**60.3**	**65.0**	**39.5**	**29.0**	**17.5**	**66.1**	**66.4**	**44.8**	**18.6**	**59.0**	

	Method	tble	dog	hrse	mbke	prsn	plnt	shep	sofa	train	tv	mAP
VGGF	MILlinear[19]	16.0	21.3	32.5	43.4	**21.9**	19.7	21.5	22.3	36.0	18.0	25.4
	Multi-fold[15]	23.6	29.2	38.5	47.9	20.3	20.0	35.8	30.8	41.0	20.1	30.2
	PDA[59]	9.4	28.8	38.1	49.8	14.5	**24.0**	27.1	12.1	42.3	39.7	31.0
	LCL+Cont[41]	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	31.6
	WSDDN[57]	18.5	31.1	45.8	54.5	10.2	15.4	36.3	45.2	50.1	43.8	34.5
	ContextNet[62]	**49.2**	**42.0**	47.3	56.6	15.3	12.8	24.8	**48.9**	44.4	47.8	36.3
	WCCN[58]	20.5	33.8	47.6	56.8	12.7	18.8	**39.6**	46.9	**52.9**	45.1	37.3
	OICR[63]	39.9	31.0	**54.0**	**62.4**	4.5	20.6	39.2	38.1	48.9	48.6	37.9
	MELM	35.5	28.9	30.9	61.0	5.8	22.8	38.8	39.6	42.1	**54.8**	**38.4**
VGG16	WSDDN[57]	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
	PDA[59]	22.1	34.4	50.5	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
	OICR[63]	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	**64.3**	**62.6**	41.2
	Self-Taught[61]	24.7	43.6	48.4	65.8	6.6	18.8	51.9	43.6	53.6	62.4	41.7
	WCCN[58]	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	42.8
	TS2C[66]	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
	WeakRPN[64]	57.3	35.2	**64.2**	**68.6**	**32.8**	28.6	50.8	49.5	41.1	30.0	45.3
	MELM	**45.3**	**53.2**	49.6	**68.6**	2.0	25.4	**52.5**	**56.8**	62.1	57.1	**47.3**
Ens.	OICR-Ens.[63]	41.0	31.3	51.9	64.8	**13.6**	23.1	41.6	48.4	58.9	58.7	42.0
	MELM-Ens.	**48.4**	**53.2**	**53.0**	**67.2**	11.0	**26.5**	**50.0**	**55.7**	**63.1**	**62.4**	**47.8**

表 3.4 最小熵隐变量模型和最新方法在 VOC 2010,2012 和 ILSVRC2013 的检测性能对比。

Table 3.4 Detection mean average precision (%) on the PASCAL VOC 2010, 2012, and the ILSVRC 2013 datasets. Comparison of MELM to the state-of-the-arts.

Dataset	CNN	Method	Dataset Splitting	mAP
PASCAL VOC 2010	VGGF/ AlexNet	PDA[59]	train/val	21.4
		WCCN[58]	trainval/test	28.8
		MELM	train/val	**35.6**
		MELM	trainval/test	**36.3**
	VGG16	PDA[59]	train/val	30.7
		WCCN[58]	trainval/test	39.5
		MELM	train/val	**37.1**
		MELM	trainval/test	**39.9**
PASCAL VOC 2010	VGGF/ AlexNet	PDA[59]	train/val	22.4
		MILinear[19]	train/val	23.8
		WCCN[58]	trainval/test	28.4
		ContextNet[62]	trainval/test	35.3
		OICR-VGGM[63]	trainval/test	34.6
		MELM	train/val	**36.2**
		MELM	trainval/test	**36.4**
	VGG16	PDA[59]	train/val	29.1
		Self-Taught[61]	train/val	39.0
		WCCN[58]	trainval/test	37.9
		OICR[63]	trainval/test	37.9
		Self-Taught[61]	trainval/test	38.3
		TS2C[59]	trainval/test	40.0
		MELM	train/val	**40.2**
		MELM	trainval/test	**42.4**
ILSVRC 2013	VGGF/ AlexNet	MILinear[19]	-	9.6
		PDA[59]	val1/val2	7.7
		WCCN[58]	-	9.8
		MELM	val1/val2	**13.4**

然而，尽管最小熵隐变量模型对各个 VOC 数据集检测性能提升很大，但是在“person”这个类别却出现了较大的性能下降。这是因为“person”这个类别是弱监督目标检测任务中最具有挑战的类别之一，该类别的不同目标之间往往有着非常大的差异，例如服装、姿态和遮挡等情况。同时，“person”目标的定义也导致很多同类目标非常不一致，例如“person”的目标在图像中可能是行人、上半身肖像或者甚至就是人脸。其类别定义的模糊性也导致了模型在学习该类别特征的时候的不准确性。为了能够稳定而准确的定位到“person”目标，算法只能选取其最稳定的判别区域，该区域通常就是人脸。尽管“person”

这个类性能降低了，但是其他的类别的性能均有大幅提升。

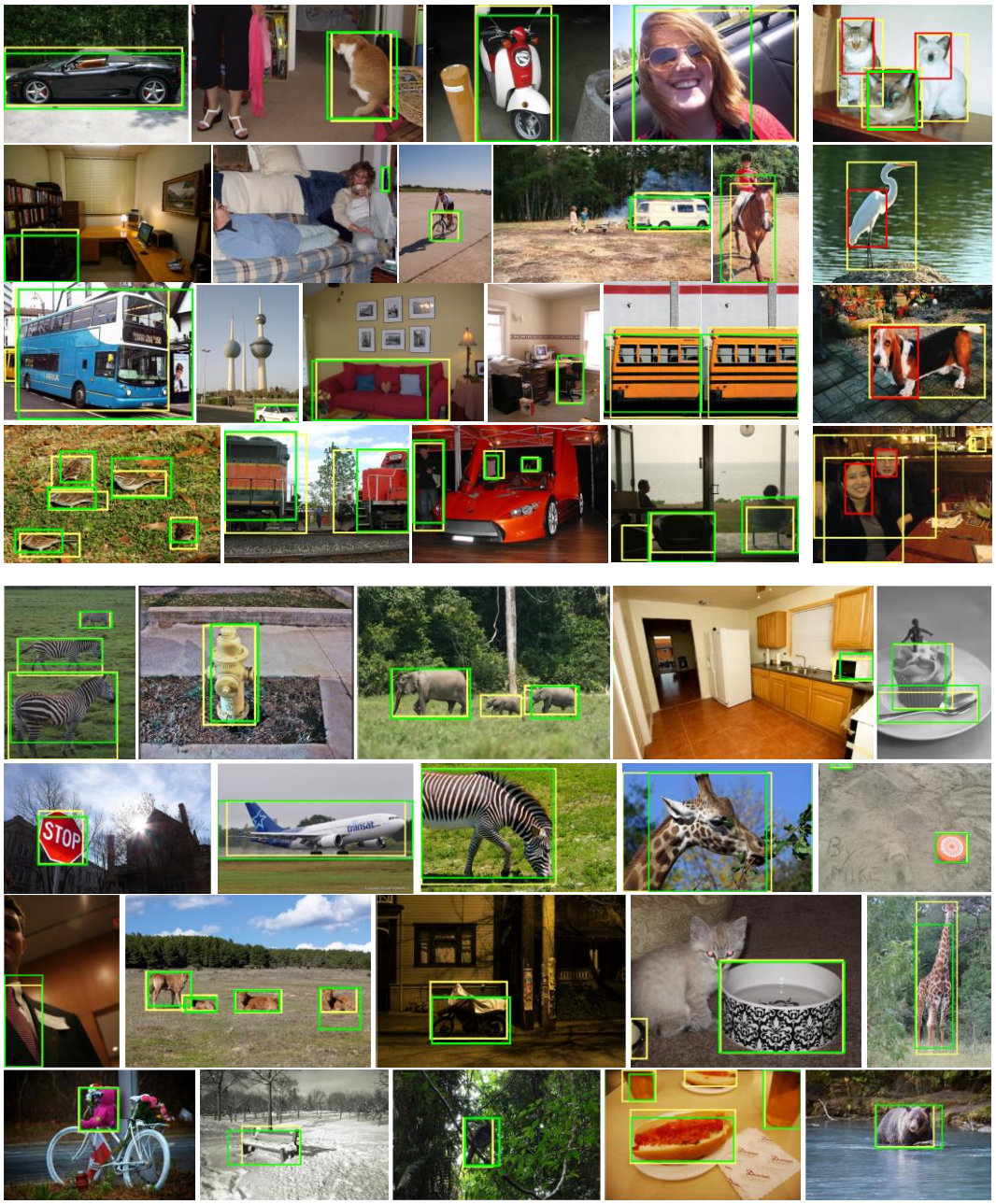


图 3.12 PASCAL VOC 2012 和 MSCOCO 2014 数据集的检测结果示例。其中黄框表示目标真实位置，绿框表示检测结果正确，红框表示检测失败。

Figure 3.12 Object detection examples on the PASCAL VOC 2012 and MS COCO 2014 datasets. Yellow bounding boxes denote ground-truth annotations, green boxes correct detection results and red boxes false detection results.

对于很多外观差异非常大的目标类别，我们发现算法通常能够正确分类目标，但是却不能完全准确的定位目标，通常定位结果和目标真实位置的 IoU 小于 0.5。为了能够进一步验证算法的定位能力，我们使用了点定位的评测。点定

位的评测过程和 CorLoc 类似，唯一的区别是，CorLoc 是判断图像中最高得分的候选框是否定位到目标，而点定位则是评断最高得分的像素点是否落在目标候选框的范围内。通过评测点定位，我们发现对于“person”这个类别，点定位的性能达到 97.1%，这也表明了算法在对定位要求较低的场合的应用潜力。

图 3.12 中给出了一些最小熵隐变量模型的检测结果可以看出，最小熵隐变量模型能够准确的定位到复杂背景中的目标，并且能够检测到单张图像中的多个目标，说明了其具有很好的判别性。

弱监督目标定位：弱监督目标定位的评测在 PASCAL VOC 2007 训练验证集（*trainval*）上完成，其目的是评测模型在训练过程中对目标定位的准确度。从表 3.5 可以看出，当使用 VGGF 作为基网时，最小熵隐变量模型分别超过最

表 3.5 PASCAL VOC 2007 上的目标定位性能。最小熵隐变量模型和最新方法对比。

Table 3.5 Correct localization rate (%) on the PASCAL VOC 2007 trainval set. Comparison of MELM to the state-of-the-arts.

CNN	Method	mAP
VGGF/AlexNet	MILinear[19]	43.9
	LCL+Context[41]	48.5
	PDA[59]	49.8
	WCCN[58]	52.6
	Multi-fold MIL[15]	54.2
	WSDDN[57]	54.2
	ContextNet[62]	55.1
	MELM	**58.4**
VGG16	PDA[59]	52.4
	WSDDN[57]	53.5
	WCCN[58]	56.7
	MELM	**61.4**

WSDDN、WCCN 4.2% (58.4% vs. 54.2%) 和 5.8% (58.4% vs. 52.6%)；当使用 VGG16 作为基网时最小熵隐变量模型分别超过最新已发表的算法 WSDDN、WCCN 7.9% (61.4% vs. 53.5%) 和 4.7% (61.4% vs. 56.7%)。值得注意的是，对于“bus”、“car”、“chair”和“table”这些类，最小熵隐变量模型超过了这些最新已发表的工作 7~15%。这充分验证了使用了候选框团的最小熵隐变量模型比对比方法 WCCN 中引入分割策略的方法更为有效。

图像分类：在本章提出的最小熵隐变量模型中，目标候选框团挖掘和目标样本定位模块能够抑制背景，激活更完整的目标区域，这些特点同样有助于正确分类图像。在表 3.6 中，我们评测了最小熵隐变量模型在 PASCAL VOC 2007 上的图像分类性能。当使用 VGGF 时，图像分类的性能达到了 87.8%；当使用 VGG16 时，图像分类的性能高达 93.1%，这些性能比最新发表的方法 WSDDN 和 WCCN 分别高出 3.4% (93.1% vs. 89.7%) 和 2.2% (93.1% vs. 90.9%)。值得注意的是，使用 VGG16 的最小熵隐变量模型的图像分类性能比 VGG16 单独训练图像分类时的性能提升高达 3.8% (93.1% vs. 89.3%)。

表 3.6 PASCAL VOC 2007 测试集上的图像分类性能。最小熵隐变量模型和最新方法对比。

Table 3.6 Image classification mAP (%) on the PASCAL VOC 2007 test set. Comparison of MELM to the state-of-the-arts.

CNN	Method	mAP
VGGF/ AlexNet	MILinear[19]	72.0
	AlexNet[31]	82.4
	WSDDN[57]	85.3
	WCCN[58]	**87.8**
	MELM	**87.8**
VGG16	VGG16[32]	89.3
	WSDDN[57]	89.7
	WCCN[58]	90.9
	MELM	**93.1**

ILSVRC 2013 和 MSCOCO 2014 数据集的实验结果和对比：除了 PASCAL VOC 数据集以为，我们还在大规模数据集 ILSVRC 2013 和 MSCOCO 2014 上进行了实验对比。其中，在拥有 200 个目标类别的 ILSVRC 2013 数据集上，使用 VGGF 作为基网的最小熵隐变量模型取得了 13.4%的检测性能，超过了 WCCN 3.6%。在 MSCOCO 2014 数据集中，我们评测了图像分类，点定位和目标检测的性能。其中，图像分类的评测标注包括宏/微准确率（P-C 和 P-O）、宏/微召回率（R-C 和 R-O）和宏/微 F1-测度（F1-C 和 F1-O）。从表 3.7 可以看出，最小熵隐变量模型的图像分类的性能比最好的算法 SPN 的性能高 23.1%（79.1% vs. 56%），点定位性能比 SPN 高 9.8% （65.1% vs. 55.3%），同时检测性能也超

过了 WSDDN。

表 3.7 MSCOCO 2014 数据集的图像分类、目标检测依据点定位性能。最小熵隐变量模型和最新方法对比。

Table 3.7 Image classification, detection and localization performance (%) on MSCOCO 2014. Comparison of MELM to the state-of-the-arts.

Image Classification							
Method	mAP	F1-C	P-C	R-C	F1-O	P-O	R-O
CAM[33]	54.4	-	-	-	-	-	-
SPN[38]	56	-	-	-	-	-	-
ResNet-101[38]	75.2	69.5	**80.8**	63.4	74.4	82.2	68
MELM-VGG16	**79.1**	**72**	79.3	**68.6**	**76.8**	**82.5**	**71.9**

Pointing Localization (with class prediction)					
Method	WeakSup[35]	Pronet[63]	DFM[43]	SPN[38]	MELM
mAP	41.2	43.5	49.2	55.3	**65.1**

Object Detection			
Method	CNN	mAP@.5	mAP@[.5,.95]
WSDDN[57]	VGGF	10.1	3.1
MELM	VGGF	11.9	4.1
	VGG16	**18.8**	**7.8**

3.7 本章小结

本章提出了一种有效的深度学习模型，称为最小熵隐变量模型，用于弱监督目标检测任务。最小化熵能减少系统的随机性，通过引入最小熵隐变量模型，算法在训练阶段的定位随机性得到了降低，因此能够更稳定的学习目标特征，提升目标定位的准确性。最小熵隐变量模型的贡献总结如下：

一是采用深度神经网络结合最小熵隐变量模型以便更有效地挖掘目标候选框，并最小化学习过程中的定位随机性；二是采用候选框团更好地搜集目标的信息并激活完整的目标区域，从而能够更准确的检测目标。三是用循环学习算法分别将图像分类和目标检测看做“predictor”和“corrector”，并且利用连续优化（Continuation Optimization）的方法解决非凸优化问题。四是在 PASCAL VOC 数据集上取得了 state-of-the-art 的分类、定位和检测性能。

第4章 渐进多示例学习

传统的弱监督目标检测模型通常采用多示例学习框架及其改进框架，而该框架的目标方程往往是非凸的。目标方程非凸的问题会造成模型在优化的过程中可能会陷入局部最优。另外，对于弱监督目标检测任务而言，多示例学习的优化目标是图像分类，并非目标检测。图像分类的框架会造成网络聚焦到最具有判别性的区域，该区域通常只是目标区域的局部。模型非凸和容易聚焦到目标局部这两个原因，使得传统的弱监督目标检测算法容易陷入局部最优，从而错误的定位到背景或者目标局部。

在本文第三章中，我们提出的最小熵隐变量模型以凸正则的形式，一定程度上缓解了弱监督框架的非凸优化问题。此外，还有一些研究人员通过引入空间先验、上下文信息和分类器精调等正则项方式来缓解模型非凸所造成的影响。但是对于如何系统性的处理模型的非凸问题，从优化的角度解决弱监督目标检测容易陷入局部最优的问题等仍然缺乏研究。

本章将渐进优化的方法引入到多示例学习的框架，提出了渐进多示例学习模型来解决弱监督目标检测的非凸优化问题，如图 4.1 所示。其中，图 4.1（b）是传统的多示例学习示意图。从图中可以看出，由于目标/损失函数是非凸的，模型在训练的过程中容易陷入到局部最优，从而导致最终定位到目标的最具有判别行的局部（绿色方框为定位结果）。根据分析，我们发现多示例学习框架之所以容易陷入局部最优，是因为模型在训练过程中需要选择最具有判别性的一个示例（候选框），并使用该候选框去区图像的标号。对于目标整体而言，目标具有判别性的局部的模式更加稳定，更有利于图像分类，因此更容易在学习的过程中被选中。该过程往往在训练初期就已经发生，而弱监督目标检测算法因为缺乏精确的位置标注信息而无法去除这些定位错误的目标局部，导致这些定位不准确的目标候选框在后续的训练过程中被保留下来，并对模型的训练造成了持续的负面影响。为了避免这个问题，我们引入了渐进优化的学习方式，如图 4.1（a）所示。与传统的多示例学习框架而言，渐进多示例学习方法没有直

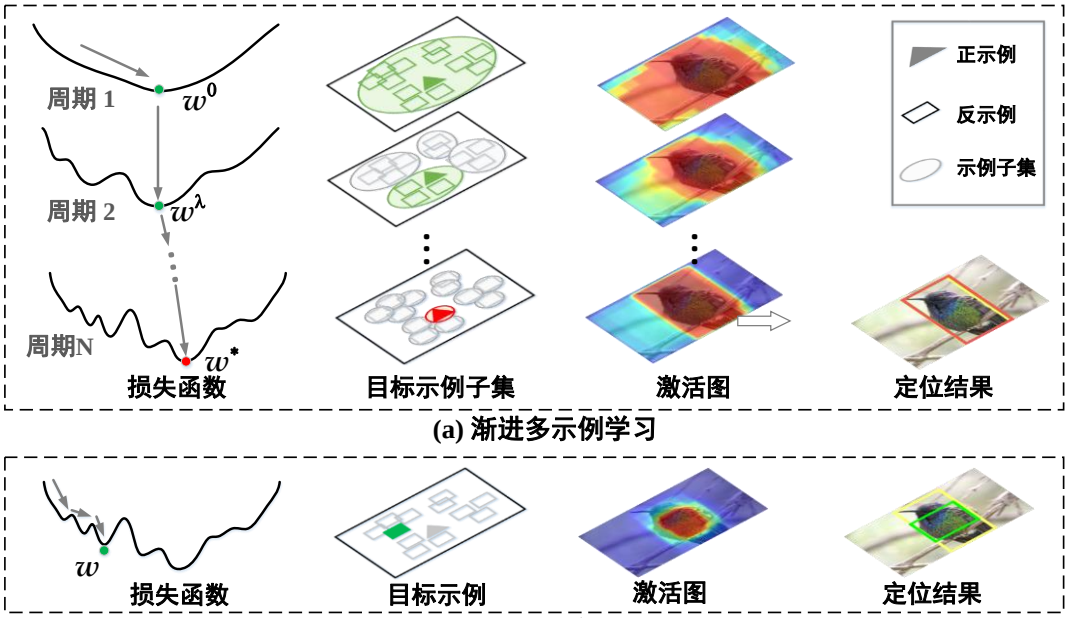


图 4.1 多示例学习和渐进多示例学习的对比。由于多示例学习损失函数非凸，该方法容易陷入到局部最优。渐进多示例学习通过引入一个序列的原函数的近似且更容易求解的损失函数，来处理原损失函数非凸所导致的优化问题，最终定位到目标整体。

Figure 4.1 Comparison of the optimization procedures of MIL and C-MIL approaches. Due to the non-convex loss function MIL often falls into local minima and falsely localizes an object part. By constructing a series of functions which are easier to optimize to approximate the original loss function, C-MIL alleviates the non-convexity problem and localizes full object extent.

接选择候选框去优化图像分类，而是首先对候选框进行划分，然后使用划分之后的候选框子集去优化图像分类，通过这种方式避免模型的定位结果陷入到某一个目标局部。连续优化的思路是，将一个复杂的非凸优化问题用一个凸函数近似，然后通过一个连续变量控制该凸函数与原函数的近似程度，在训练的过程中逐渐从该与原函数近似的凸函数，逐渐还原成原来的复杂的非凸函数，如图 4.1（a）自上而下所示。对原函数的近似则通过对候选框的动态划分来完成。当一张图像中所有的候选框均被划分到一个子集时，此时原函数蜕变为单纯分类函数，此时目标方程问凸函数。通过候选框子集的划分系数的逐渐调整，候选框逐渐被划分成多个自己，在训练过程中，这个划分不断的调整，使得最终候选框子集的个数和图像中候选框的个数相同。也就是每个候选框子集只包含一个候选框，此时的目标方程和原多示例学习方程一致。通过这种方式，渐进多示例学习能够以一个和原目标方程近似的凸函数为起点，通过渐进优化的方

式逐渐逼近原函数，从而使得优化问题更容易求解，模型对非凸问题更鲁棒。

本章后续章节首先在 4.1 节回顾多示例学习的建模和在弱监督目标检测问题中的应用；然后在 4.2 节对多示例学习的非凸优化问题进行具体的分析，并结合弱监督目标检测任务提出具体的解决思路；在 4.3 节中，我们将介绍渐进多示例学习算法的建模；在 4.4 节和 4.5 节中，我们将分别介绍渐进多示例学习的网络结构和实现细节、以及具体的实验结果和分析。

4.1 多示例学习回顾

在多示例学习框架中，图像被看作是一个示例包，而图像中的候选框被看作是一个示例。多示例学习的目标是在只给定示例包的标号的情况下，学习关于示例的分类器。其中示例包分为正示例包和反示例包，示例分为正示例和反示例。其关系是，如果示例包中至少有一个正示例，则该示例包为正示例包；只有当示例包中的全部示例均为反示例时，该示例包为反示例包。为了能更加清晰的对多示例学习的问题进行建模和分析，我们首先介绍下面的符号表示：

$B_i \in \mathcal{B}$：示例包（图像）B_i 属于示例包集合（图像数据集合）$\mathcal{B}$。

$B_{ij} \in B_i$：示例（候选框）B_{ij} 是示例包 B_i 中的一个示例。

$y_i \in Y$：示例包标号 y_i，其取值范围为标号集合 $Y = \{1, -1\}$，其中 $y_i = 1$ 表示示例包中包含正示例（目标），即为正例图像；$y_i = -1$ 表示图像中不包含正示例，即为反例图像。

$y_{ij} \in Y$：示例标号 y_{ij}，其中 $j \in \{1, 2, ..., N\}$，N 是示例包 B_i 中示例的个数。示例标号 y_{ij} 的取值范围为标号集合 $Y = \{1, -1\}$，其中 $y_{ij} = 1$ 表示图像中包含感兴趣的目标，即为正例图像；$y_{ij} = -1$ 表示图像中不包含感兴趣的目标，即为反例图像。

w：模型的参数。

$L(\cdot)$：损失函数。

与第三章中的符号体系不同的是，本章的符号表示侧重于表示示例（候选

框）和示例包（图像）之间的关系，因此采用了双索引的方式，使得该关系一目了然，而第三章中的符号体系则侧重表表示隐变量，因此使用了单独的符号来突出隐变量在学习过程中的作用。

根据上述定义，多示例学习在弱监督目标检测中的学习过程可以被总结为示例挖掘和检测器（示例分类器）学习两个步骤：

（1）示例挖掘。正示例挖掘的步骤是学习一个示例的挖掘器 $f\left(B_{ij}, w_f\right)$，其参数为 w_f，然后用该挖掘器从示例包中挖掘到正示例（目标）B_{ij^*}，具体过程如下述公式所示

$$B_{ij^*} = \arg\max_j f\left(B_{ij}, w_f\right) \tag{4.1}$$

其中，j^* 表示示例包 B_i 中的最高得分示例的索引，如图 4.2 所示。利用最高得分的示例样本，我们可以通过不用示例之间的空间位置关系，将其余示例划分为正示例和反示例。其中，两个示例之间的空间位置关系通过其对应候选框之间的 IoU 来衡量。

（2）检测器学习。检测器的学习过程是利用上一个步骤中挖掘到的示例样本训练检测器 $g_z\left(B_{ij}, B_{ij}^*, w_g\right)$，其中 $z \in Y$。w_f 和 w_g 分别是示例挖掘器和检测器的参数。

在传统多示例学习框架中，示例挖掘其 $f(\cdot)$ 和检测器 $g(\cdot)$ 是同一模型并享有相同的参数。在训练过程过程中，这两个部分交替迭代，直到所有的示例标号不再更新。每次示例挖掘其挖掘到示例样本之后，模型的参数都会重置成随机初始化。

然而，在深度学习框架中，这种学习方式十分耗时耗力。为了能够更容易的训练多示例学习模型，我们将示例挖掘其 $f(\cdot)$ 和检测器 $g(\cdot)$ 定位为两个互相独立的模型，并将其加入到同一个网络的不同分支中，使得两个模型既能独立学习，也能相互配合。深度框架中的多示例学习损失函数定义如下：

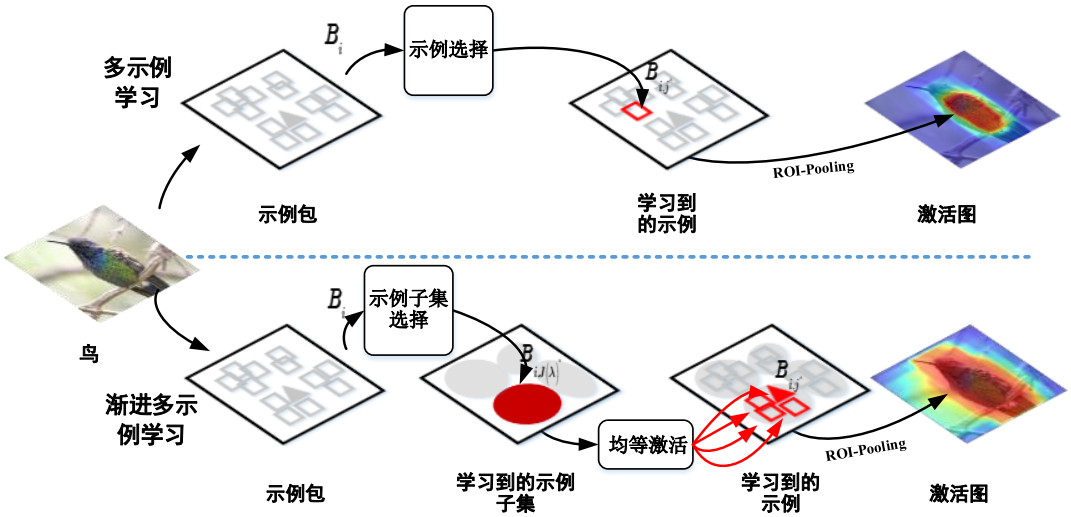


图 4.2 多示例学习与渐进多示例学习的激活对比。多示例学习由于只选择一个示例用于图像分类，当该示例为目标局部的时候，容易只激活目标局部。渐进示例学习通过引入示例子集，每次均等选择一个示例集合，因此更容易激活目标整体。

Figure 4.2 For a "bag" of instances (region proposals), CMIL and MIL use different strategies to selection instances. MIL tends to select the most discriminative instance and activate the object part. In contrast, C-MIL selects the most discriminative instance subset. The instances in the subset are activated equally during back-propagation and thus the object extent is activated.

$$L(\mathcal{B},w)=\sum_i L_f\left(B_i,w_f\right)+L_g\left(B_i,B_{ij}^*,w_g\right) \tag{4.2}$$

其中，第一项损失函数为目标示例挖掘的损失函数，具体定义形式如下

$$L_f\left(B_i,w_f\right)=\max\left(0,1-y_i\max_j f\left(B_{ij},w_f\right)\right) \tag{4.3}$$

该损失函数为标准的 hingle 损失函数。第二项为检测器的损失函数，具体定义形式如下

$$L_g\left(B_i,B_{ij*},w_g\right)=-\sum_z\sum_j \delta_{z,y_{ij}}\log g_z\left(B_{ij},w_g\right) \tag{4.4}$$

其中，示例的标号 y_{ij} 根据 PASCAL 中的正反例区分标准定义如下：

$$y_{ij}=\begin{cases}1, if\ \ IoU\left(B_{ij},B_{ij*}\right)\geq 0.5\\ -1, otherwise\end{cases} \tag{4.5}$$

其中，公式（4.4）中 $\delta_{a,b}$ 为克罗内克函数，当 $a=b$ 时其值为 $\delta_{a,b}=1$，否则为 $\delta_{a,b}=0$。

在训练过程中，网络通过公式（4.2）中定义的损失函数联合优化，最终得

到关于示例的分类器，即目标检测器。

4.2 非凸分析

在对损失函数公式（4.2）进行凸性分析之前，我们先给定两个结论：

（1）线性函数是凸函数；

（2）对若干个凸函数取最大，得到的函数仍为凸函数。

根据上述两个结论，我们可以看出，当$y_i=-1$的时候，公式（4.3）是凸函数；但$y_i=1$时，公式（4.3）是非凸的。由于公式（4.3）是公式（4.2）其中的一项，由此可以推断出公式（4.2）是非凸的。非凸的目标方程会导致模型在训练的时候容易陷入局部最优，影响目标示例挖掘的准确性，进而影响检测器的学习。这个问题是目前弱监督视觉目标检测中的一个关键问题，主要影响算法在正示例包中对正示例的挖掘。

根据上述分析，传统多示例学习的方法中存在两个待解决的问题：

（1）如何优化多示例学习的非凸目标方程，并竟可能找到全局最优解；

（2）在示例挖掘器还没有训练充分的时候，也就是训练初期，如何避免挖掘错误的示例。

4.3 渐进多示例学习

为了解决上一节提到的两个问题，我们提出了渐进示例学习，系统解决这两个问题。在渐进多示例学习中，我们没有直接使用正则项去优化原来的损失函数，而是直接从优化的角度解决这一问题。在弱监督目标检测中，我们通过引入候选框子集，对原始候选框集合进行划分，将原来多示例学习挖掘示例的过程转变成挖掘示例子集的过程，通过该方式对公式（4.3）进行平滑，通过逐渐调整示例子集的大小，我们定义了一个函数序列去逐渐逼近原始的非凸目标方程。

渐进多示例学习采用了传统的渐进优化策略。渐进优化的思想是追溯一组明确定义的平滑函数序列的轨迹，该轨迹从一个初始点$\left(w^0,0\right)$到结束点$\left(w^*,1\right)$，

其中， w^0 是损失函数 $L(\mathcal{B},w,\lambda)$ 在 $\lambda=0$ 时的解； w^* 是损失函数 $L(\mathcal{B},w,\lambda)$ 在 $\lambda=1$ 时的解。$L(\mathcal{B},w,\lambda)$ 是平滑之后的目标方程，其平滑程度受到渐进参数 λ 的控制。当 $\lambda=0$ 时， $L(\mathcal{B},w,\lambda)$ 为凸函数；当 $\lambda=1$ 时， $L(\mathcal{B},w,\lambda)$=$L(\mathcal{B},w)$，此时函数退化成多示例学习的损失函数。

根据上述渐进优化的思路，我们定义了一组关于渐进参数 λ 的序列为 $0=\lambda_0<\lambda_1<...<\lambda_T=1$，其中 T 为迭代次数。由此，公式（4.2）更新为

$$\begin{aligned} w^* &= \arg\min_{w} L(\mathcal{B},w,\lambda) \\ &= \arg\min_{w_f,w_g} \sum_i L_f\left(B_i,B_{i,J(\lambda)},w_f\right)+L_g\left(B_i,B_{i,J(\lambda)},w_g\right) \end{aligned} \tag{4.6}$$

其中， $B_{i,J(\lambda)}$ 表示表示示例子集， $J(\lambda)$ 是示例子集中示例的索引集合，该集合由渐进参数 λ 控制。$L_f\left(B_i,B_{i,J(\lambda)},w_f\right)$ 和 $L_g\left(B_i,B_{i,J(\lambda)},w_g\right)$ 分别是渐进示例挖掘器和渐进检测器的损失函数。

4.3.1 渐进示例挖掘

在学习示例挖掘器时，示例包首先被划分成示例子集。示例子集中的所有示例相互之间具有空间关联性（空间位置相互重叠）和类别关联性（属于同一个目标类别）。示例子集是示例包的最小充分覆盖，其满足如下关系 $\bigcup_J B_{i,J}=B_i$ 和 $B_{i,J}\cap B_{i,J'}=\varnothing,\forall J\neq J'$。首先所有的示例根据其示例挖掘器的得分 $f\left(B_{ij},w_f\right)$ 排序，接下来循环执行下面两个步骤：

（1）用最高得分且不属于任何一个示例子集的示例构建新的示例子集。

（2）在其他未被划分的示例集合中，搜索与该最高得分示例 IoU 大于等于阈值 λ 的示例，并将其加入该示例子集。

当 $\lambda=0$ 时，示例包 B_i 中的所有示例被划分到一个示例子集中；当 $\lambda=1$ 时，示例包 B_i 中的示例子集的个数等于示例数，也就是每个示例子集中只包含一个示例，这种情况和多示例学习模型是等价的。当给定参数 $\lambda\in[0,1]$ 时，示例子

集挖掘的损失函数定义如下：

$$L_f\left(B_i, B_{i,J(\lambda)}, w_f\right) = \max(0, 1 - y_i \max_{J(\lambda)} f(B_{i,J(\lambda)}, w_f)) \quad (4.7)$$

其中，$f(B_{i,J(\lambda)}, w_f)$是示例子集$B_{i,J(\lambda)}$的得分，其定义如下

$$f\left(B_{i,J(\lambda)}, w_f\right) = \frac{1}{\left|B_{i,J(\lambda)}\right|} \sum_j f\left(B_{i,j}, w_f\right) \quad (4.8)$$

其中$\left|B_{i,J(\lambda)}\right|$表示示例子集$B_{i,J(\lambda)}$中示例的个数，$B_{i,j} \in B_{i,J(\lambda)}$。

根据公式（4.8）示例子集的得分是示例子集中所有示例的平均得分，该过程可以看作是平滑滤波，从而使得公式（4.7）相比公式（4.3）而言更为平滑。因此，公式（4.6）相比公式（4.2）更为平滑，当不断调整λ的数值时，我们便能得到相应的一个序列的平滑函数去缓解原目标方程的非凸问题，如图 4.1（a）所示。

在训练过程中，渐进多示例学习平等的利用示例子集中的示例去更新网络参数，并且在反传的过程中，示例子集所覆盖的区域能被均匀的激活。由于一个示例子集中的示例互相之间是有空间位置关联的，因此，示例子集中容易包含包括目标、目标部件等区域，因而能够激活目标的完整区域。当$\lambda = 0$时，由于每个示例包只包含一个示例子集，因此，公式（4.7）中第二项 max 函数变为线性函数。根据“两个线性函数取最大得到的函数是凸函数” 这一结论可以推断出公式（4.7）此时变成凸函数。当$\lambda = 1$时，每个示例子集只包含一个示例，此时模型退化为多示例学习模型；当$\lambda \in (0,1)$时，公式（4.7）是公式（4.3）的平滑后的结果，如图 4.1 所示。

4.3.2 渐进检测器学习

在训练过程中，得分最高的示例子集中$B_{i,J(\lambda)^*}$被用于检测器的学习。考虑到检测器的训练过程中没有框标注，并且被选中的示例子集$B_{i,J(\lambda)^*}$中可能会包含目标部件或者背景，我们进一步提出渐进的检测器学习思路。

首先，我们将示例包中的示例根据渐进参数λ划分成正示例和反示例。我们将渐进示例挖掘输出的示例子集$B_{i,J(\lambda)^*}$中最高得分的示例表示成B_{i,j^*}，那么

示例划分的过程则可表示成

$$y_{i,j}=\begin{cases}+1, if\ \ IoU\left(B_{i,j},B_{i,j^*}\right)\geq 1-\lambda/2\\ -1, if\ \ IoU\left(B_{i,j},B_{i,j^*}\right)<\lambda/2\end{cases} \tag{4.9}$$

由公式（4.9）可以得出，在训练过程中，与 B_{i,j^*} IoU 大于 $1-\lambda/2$ 的示例为正示例；与 B_{i,j^*} IoU 小于 $\lambda/2$ 的示例为反示例；与 B_{i,j^*} IoU 介于 $[\lambda/2,1-\lambda/2]$ 的示例在训练过程中将被忽略。

在训练过程中，随着渐进参数 λ 从 0 到 1 逐渐变化，正示例的阈值 $1-\lambda/2$ 从 1 逐渐减小到 0.5，而反示例的阈值 $\lambda/2$ 从 0 逐渐增大到 0.5。这个过程中，正反示例的样本数量由于阈值的变化而逐渐增多，当渐进参数 $\lambda=1$ 时，公式（4.9）退化成公式（4.5），即和原多示例学习一致。根据上述定义的正反例样本，渐进检测器的学习可以通过优化以下损失函数实现

$$F_g\left(B_i,B_{i,J(\lambda)},w_g\right)=-\sum_z\sum_j\delta_{z,y_{ij}}\log g_z\left(B_{i,j},w_g\right) \tag{4.10}$$

4.4 网络结构和实现

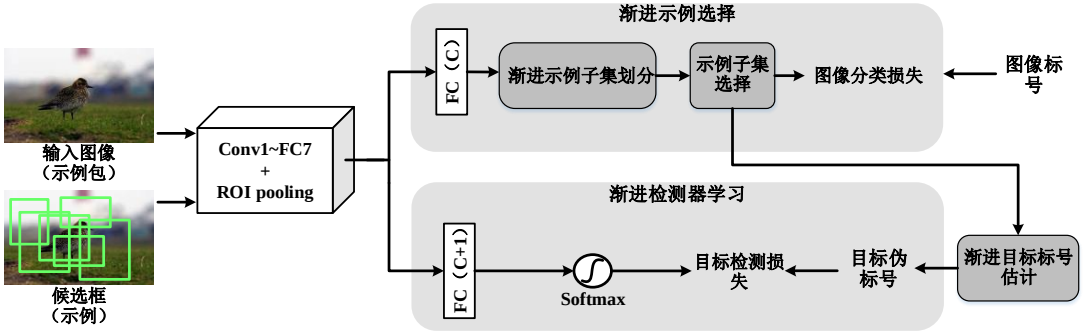


图 4.3 渐进多示例学习与深度学习框架结合用于弱监督目标检测。C 表示目标类别数。

Figure 4.3 The modules of continuation instance selection and continuation detector estimation are implemented atop a deep network for weakly supervised object detection. C is the number of object categories.

渐进多示例学习的在深度网络框架中的模块图如图 4.3 所示，输入图像首先根据候选框算法生成候选框，然后利用卷积神经网络、ROI-Pooling 和两个全连接层对所有候选框（示例）提取特征。在两个全连接层后面，我们将上了渐进示例挖掘模块和渐进检测器学习模块。在前向传播中，渐进多示例学习算法选

择示例子集中的正示例，并将其视为伪标注信息，利用该信息去训练检测器。在反向传播过程中，示例挖掘器和检测器在随机梯度下降的框架下联合优化。通过网络的正向和反向传播，网络参数不断更新，最终学习到相应的目标检测器。

4.5 实验结果与分析

为了验证本章所提出的渐进多示例学习的有效性，本文使用 VGGF 和 VGG16 作为实验的基础网络，在目前较为常用的目标检测的数据集 PASCAL VOC 2007 和 VOC 2012 数据集中实验了本章提出的方法。后面几个小节将分别介绍相关的实验设定、实验分析以及和 stateof-the-art 方法的对比。

4.5.1 实验设定

数据集：PASCAL VOC 数据集一共包括 20 个目标类别。VOC 2007 数据集一共包含 9963 张图像，其中 5011 张图像是训练集和验证集，4952 张图像用于测试集。VOC 2012 包含 22531 张图像，其中 11540 张用于训练和验证集，10991 张图像用于测试集。

评测标准：本章用到的评测标准有两种：mAP，CorLoc。其中，各个评测标准的定义见 3.6.1 节。

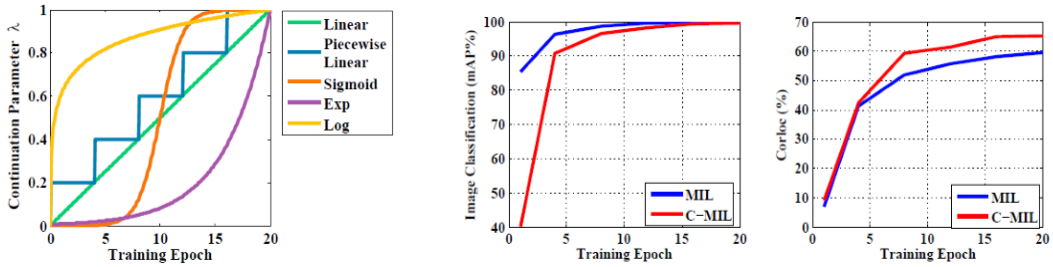
预训练模型：预训练模型采用的是目前比较主流的 VGG 网络，分别为 VGGF 和 VGG16。这两个模型均在 ILSCVR 2012 的图像分类任务中预训练。VGGF（VGG-CNN-F）和 AlexNet 的网络结构类似，拥有 5 个卷积层和 3 个全连接层。VGG16 拥有 13 个卷积层和 3 个全连接层。对于这两个基网，我们去掉了最后一个空间最大池化层，用一个 ROI-Pooling 层替代。同时去掉了最后一个全连接层，并使用一个随机初始化的全连接层替代，该全连接层的节点个数和数据集类别个数对应。

候选框生成算法：候选框生成算法采用了 Selective Search 算法和 Edge Boxes 算法。对于每张图像，算法大概生成了 2000 个左右的候选框。对于 Selective Search 算法，我们使用了其 fast 模式生成候选框。在训练过程中，我们去掉了宽或高小于 20 个像素的候选框。

训练参数：和众多已发表的弱监督目标检测算法一样，我们采用了多尺度训练策略，在训练过程中将输入图像的长或者宽随机缩放至下述五个尺度中的一个：{480,576,688,864,1200}。同时，训练图像还会被随机左右翻转。在测试的时候，我们将所有的五个尺度，包括翻转之后的图像共计 10 张图像的检测结果取平均，得到最终的检测结果。在循环学习过程中，我们使用了随机梯度下降（SGD），其中动量参数为 0.9，权重衰减系数为 5e-4，单次输入图像的数量（Batch Size）为 1。模型在整个数据集上一共迭代 20 个周期，其中前 10 个周期的学习率是 5e-3，后 10 个周期的学习率是 5e-4。

4.5.2 连续优化方法评测

渐进参数 λ：为了验证渐进参数 λ 对渐进优化过程的影响，我们验证了五种渐进参数 λ 的变化函数，如图 4.4（a）所示。评测结果如表 4.1 所示。从表中可以看出，引入了渐进优化策略之后，模型的检测性能提升了 1.1%~4.7%，模型的定位性能提升了 1.4%~4.5%。



(a) (b) (c)

图 4.4 (a)渐进参数的变化函数； (b) 和 (c) 图像分类性能和目标定位性能在训练过程中的演变。

Figure 4.4 (a) Functions of continuation parameter. (b) and (c) Evolution of image classification and object localization performance during training.

从表 4.1 中可以看出，“Log”函数取得了最好的性能。该函数的变化趋势是，在训练初期的时候，渐进参数 λ 迅速变大；而在训练后期的时候，渐进参数 λ 的变化速度逐渐放缓，最终达到最大值 1。这和训练过程是吻合的。也就是说，模型在训练初期学习大的示例子集，这样有助于模型激活目标完整区域，避免陷入目标部件等局部最优；而在训练后期，示例子集趋于稳定，此时模型更聚焦于训练检测器的学习。

表 4.1 渐进参数 λ 按五种函数曲线变化的检测和定位结果，实验数据集为 PASCAL VOC 2007，基网为 VGGF。

Table 4.1 Comparison of five functions controlling the change of continuation parameter λ. Detection and localization performance (%) on the VOC 2007 dataset with VGGF.

Method	Approaches/ Continuation Functions	mAP	CorLoc
MIL	ContextNet[22]	36.0	55.0
C-MIL(Ours)	Linear	37.9	58.9
	Piecewise Linear	37.6	57.4
	Sigmoid	38.3	58.4
	Exp	37.1	56.4
	Log	**40.7**	**59.5**

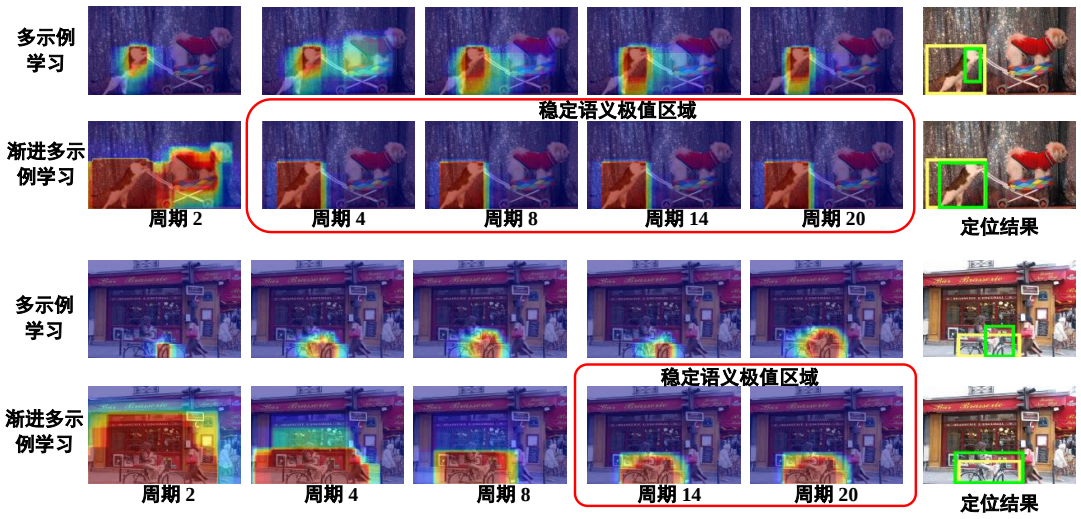


图 4.5 稳定语义极值区域。

Figure 4.5 Stable Semantic Extremal Regions (SSERs).

渐进优化策略：表 4.2 中呈现了渐进示例挖掘器和渐进检测器学习的拆解实验。从表中可以看出，与基准模型相比，当只使用渐进示例挖掘器时，检测性能提升了 3.0% (39.0% vs. 36.0%)；当只使用渐进检测器学习的时候，检测性能提升了 1.4% (37.4% vs. 36.0%)；而当两个渐进优化策略均被使用时，检测性能超过基准模型 4.7% (40.7% vs. 36.0%)。这些结果清晰的验证了渐进优化策略的引入对多示例学习的作用。

表 4.2 渐进多示例学习的拆解实验。实验数据集为 PASCAL VOC 2007，基网为 VGGF。

Table 4.2 Ablation experimental results of C-MIL. Detection performance (%) on the VOC 2007 dataset with VGGF.

Method	Instance Selector	Object Detector	mAP
MIL[22]	-	-	36.0
C-MIL(Ours)	√		39.0
		√	37.4
	√	√	40.7

图 4.4（b）和（c）呈现了图像分类和目标定位在训练过程中演变结果。多示例学习算法在训练初期的图像分类性能和目标定位性能均高于渐进多示例学习；在后续的训练中，渐进多示例学习的定位性能逐渐追上并超过多示例学习。其中的原因是，多示例学习以优化图像分类损失为主，并未考虑定位能力，其图像分类能力在训练初期能得到更好的优化。然而，多示例学习因为聚焦于寻找最具有判别性的图像区域去区分图像类别，因此容易定位到目标局部。与之不同的是，渐进多示例学习通过学习示例子集同时优化图像分类和目标定位，因此最终能够成功定位到完整的目标。

4.5.3 语义稳定极值区域

为了进一步的理解渐进优化，我们将训练过程中学习到的示例子集做了可视化，其结果如图 45 所示。从图中可以看出当渐进参数 λ 由小变大时，激活的区域逐渐变小。在训练初期，示例子集的作用是尽可能的搜集目标或者目标部件的信息。随着训练的进行，激活的区域减小的速度逐渐稳定下来，该区域在目标边缘处逐渐稳定。我们将这些区域称为稳定的语义极值区域，而这些区域的出现通常能观察到目标的完整区域被成功定位到。

稳定的语义极值区域的出现表明渐进多示例学习在训练过程中逐渐压制背景并激活目标区域。这个过程和最大稳定极值区域（Maximally Stable Extremal Regions，MSERs）有相似之处。不同的地方是，最大稳定极值区域是在原图像的像素空间中定义并以无监督的方式提取的，而稳定的语义极值区域则是在语义的基础上提取并通过弱监督的方式学习的。

4.5.4 实验性能和对比

表 4.3 中给出了渐进多示例学习和最新已发表方法在 PASCAL VOC 2007 上的目标检测性能的对比。渐进多示例学习在使用 VGGF 和 VGG16 作为基网时分别取得了 40.7%和 50.5%的性能。当使用 VGGF 时，渐进多示例学习分别超过了 WCCN[14], OICR[34] 和 MELM[37] 3.4% (40.7% vs. 37.3%), 2.8% (40.7% vs. 37.9%) 和 2.3% (40.7% vs. 38.4%)；当使用 VGG16 时，渐进多示例学习分别超过了 WeakRPN[35], TS2C[38], 和 MELM[37] 6.2% (50.5% vs. 44.3%), 5.2% (50.5% vs. 45.3%) 和 3.2% (50.5% vs. 47.3%)，这些提升在弱监督目标检测这一非常具有挑战的任务中是十分显著的。

我们进一步使用上述基于 VGG16 的渐进多示例学习模型的检测结果，将其用作伪标号，训练了 Fast-RCNN 检测器。由表 4.3 中的结果可以看出，使用 Fast-RCNN 重训之后的检测模型的性能进一步提升到了 53.1%，该结果超过了目前最新已发表方法 2.7%~6.1%。其中，“aeroplane”、“bird”、“cat”和“train”等类别的性能都得到了大幅提升。

表 4.4 中呈现了渐进多示例学习和最新已发表方法在 PASCAL VOC 2012 上的目标检测性能的对比，基网为 VGG16。可以看出，渐进多示例学习分别超过了 WeakRPN[35], TS2C[38] 和 MELM[37] 5.9% (46.7% vs. 40.8%), 6.7% (46.7% vs. 40.0%) 和 4.3% (46.7% vs. 42.4%)。图 4.6 中给出了部分检测结果示例。

表 4.3 VOC 2007 数据集上的实验性能对比

Table 4.3 Detection performance (%) on the VOC 2007 test set. Comparison of C-MIL to the state-of-the-arts.

CNN	*Method*	*aero*	*bike*	*bird*	*boat*	*bttl*	*bus*	*car*	*cat*	*char*	*cow*	
VGGF/ AlexNet	PDA[59]	49.7	33.6	30.8	19.9	13.0	40.5	54.3	37.4	14.8	39.8	
	LCL+Context[41]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	
	WSDDN[57]	42.9	56.0	32.0	17.6	10.2	61.8	50.2	29.0	3.8	36.2	
	ContextNet[62]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	
	WCCN[58]	43.9	57.6	34.9	21.3	14.7	64.7	52.8	34.2	6.5	41.2	
	OICR[63]	53.1	57.1	32.4	12.3	15.8	58.2	56.7	39.6	0.9	44.8	
	MELM[65]	56.4	54.7	30.9	21.1	17.3	52.8	60.0	36.1	3.9	47.8	
	C-MIL(Ours)	54.5	55.5	34.4	20.3	16.7	53.4	59.2	44.6	8.4	46.0	
VGG16	WSDDN[8]	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	
	PDA[59]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	
	OICR[63]	58.0	62.4	31.1	19.4	130.0	65.1	62.2	28.4	24.8	44.7	
	WCCN[58]	49.5	60.6	38.6	29.2	16.2	70.8	56.9	42.5	10.9	44.1	

	TS2C[66]	59.3	57.5	43.7	27.3	13.5	63.9	61.7	59.9	24.1	46.9	
	WeakRPN[64]	57.9	70.5	37.8	5.7	21.0	66.1	69.2	59.4	3.4	57.1	
	MELM[65]	55.6	66.9	34.2	29.1	16.4	68.8	68.1	43.0	25.0	65.6	
	C-MIL(Ours)	62.5	58.4	49.5	32.1	19.8	70.5	66.1	63.4	20.0	60.5	
FRCNN Re-train	OICR-Ens.[63]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	
	TS^{2}C[66]	-	-	-	-	-	-	-	-	-	-	
	WeakRPN-Ens.[64]	63.0	69.7	40.8	11.6	27.7	70.5	74.1	58.5	10.0	66.7	
	C-MIL(Ours)	61.8	60.9	56.2	28.9	18.9	68.2	69.6	71.4	18.5	64.3	

	Method	*tble*	*dog*	*hrse*	*mbke*	*prsn*	*plnt*	*shep*	*sofa*	*tran*	*tv*	*mAP*
VGGF/ AlexNet	PDA[59]	9.4	28.8	38.1	49.8	14.5	24.0	27.1	12.1	42.3	39.7	31.0
	LCL+Context[41]	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	31.6
	WSDDN[57]	18.5	31.1	45.8	54.5	10.2	15.4	36.3	45.2	50.1	43.8	34.5
	ContextNet[62]	49.2	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
	WCCN[58]	20.5	33.8	47.6	56.8	12.7	18.8	39.6	46.9	52.9	45.1	37.3
	OICR[63]	39.9	31.0	54.0	62.4	4.5	20.6	39.2	38.1	48.9	48.6	37.9
	MELM[65]	35.5	28.9	30.9	61.0	5.8	22.8	38.8	39.6	42.1	54.8	38.4
	C-MIL(Ours)	40.2	40.8	47.7	63.2	22.8	23.2	39.4	44.3	53.8	52.3	40.7
VGG16	WSDDN[57]	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
	PDA[59]	22.1	34.4	50.5	61.8	16.2	29.9	40.7	15.9	55.3	40.2	39.5
	OICR[63]	30.6	25.3	37.8	65.5	15.7	24.1	41.7	46.9	64.3	62.6	41.2
	WCCN[58]	29.9	42.2	47.9	64.1	13.8	23.5	45.9	54.1	60.8	54.5	42.8
	TS^{2}C[66]	36.7	45.6	39.9	62.6	10.3	23.6	41.7	52.4	58.7	56.6	44.3
	WeakRPN[64]	57.3	35.2	64.2	68.6	32.8	28.6	50.8	49.5	41.1	30.0	45.3
	MELM[65]	45.3	53.2	49.6	68.6	2.0	25.4	52.5	56.8	62.1	57.1	47.3
	C-MIL(Ours)	52.9	53.5	57.4	68.9	8.4	24.6	51.8	58.7	66.7	63.5	50.5
FRCNN Re-train	OICR-Ens.[63]	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
	TS^{2}C[66]	-	-	-	-	-	-	-	-	-	-	48.0
	WeakRPN-Ens.[64]	60.6	34.7	75.7	70.3	25.7	26.5	55.4	56.4	55.5	54.9	50.4
	C-MIL(Ours)	57.2	66.9	65.9	65.7	13.8	22.9	54.1	61.9	68.2	66.1	**53.1**

表 4.4 VOC 2012 数据集上的检测和定位性能对比

Table 4.4 Detection and localization performance (%) on the VOC 2012 dataset using VGG16. Comparison of C-MIL to the state-of-the-arts.

Method	mAP	CorLoc
WCCN[14]	37.9	-
Self-Taught[21]	38.3	58.8
OICR[34]	37.9	62.1
TS2C[38]	40.0	64.4
WeakRPN[35]	40.8	64.9
MELM[37]	42.4	-
C-MIL(Ours)	46.7	**67.4**

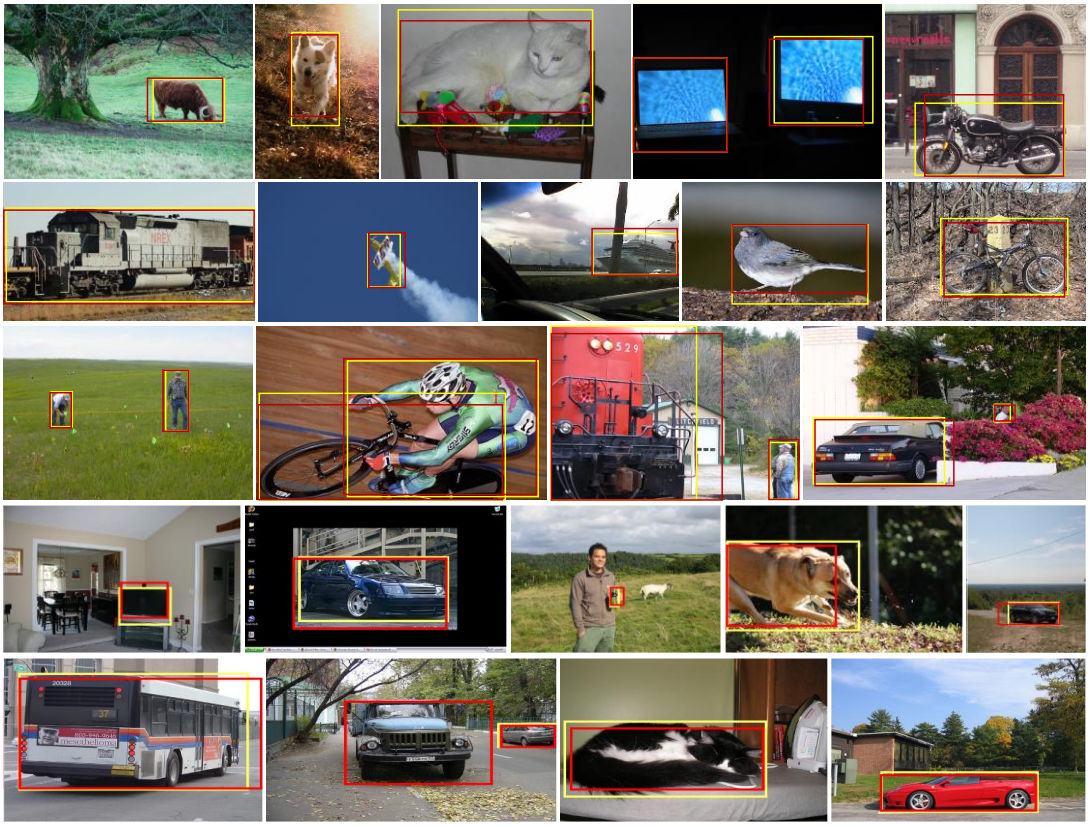


图 4.6 PASCAL VOC 2012 数据集的检测结果示例。

Figure 4.6 Object detection examples on the PASCAL VOC 2012 dataset.

在表 4.4 和 4.5 中，我们评测了渐进多示例学习的目标定位性能。在 PSACAL VOC 2007 数据集上，渐进多示例学习分别超 WeakRPN[35] 和 TS2C[38] 1.2% (65.0% vs. 63.8%) 和 4.0% (65.0% vs. 61.0%)；VOC 2012 数据集上，渐进多示例学习分别超 WeakRPN[35] 和 TS2C[38] 3.0% (67.4% vs. 64.4%) and 2.5% (67.4% vs. 64.9%)。

表 4.5 VOC 2007 目标定位性能对比

Table 4.5 Localization performance (%) on the VOC 2007 trainval set. Comparison of C-MIL to the state-of-the-arts.

CNN	Method	mAP
VGG16	WSDDN[57]	53.5
	WCCN[58]	56.7
	OICR[63]	60.6
	TS2C[66]	61.0
	WeakRPN[64]	63.8
	C-MIL(Ours)	**65.0**

4.6 本章小结

本章更深入研究了涉及非凸目标函数的弱监督问题的优化，提出了渐进多示例优化方法。该方法致力于解决传统多示例学习方法的非凸优化问题。通过引入一个序列的对原函数的平滑损失函数，在训练过程中以一个容易求解的凸损失函数为起点，逐渐优化该序列中的平滑损失函数，直至损失函数退化成原损失函数。该平滑过程通过引入示例子集完实现。

渐进多示例学习显著提升了弱监督目标检测和定位的性能，并超过了目前最新已发表的工作。这些现象背后原理在于：当使用渐进优化模型和深度网路结合时，模型在训练过程中通过搜集目标或者目标部件的方式激活了目标的完整区域，从而最终学习到语义稳定极值区域。本章的研究拓展了弱监督视觉目标检测与弱监督学习问题的求解思路。

第5章 弱监督 X 光图像违禁品检测

前面两章介绍了从建模和优化两个方面对弱监督视觉目标检测任务的理论研究。本章从应用的角度以 X 光图像违禁品检测为背景，解决弱监督目标检测的实际应用问题。接下来首先对 X 光图像违禁品的应用做简单介绍并指出相应的问题，然后给出对应的模型方法，最后通过实验验证提出的方法的有效性。

5.1 问题简介

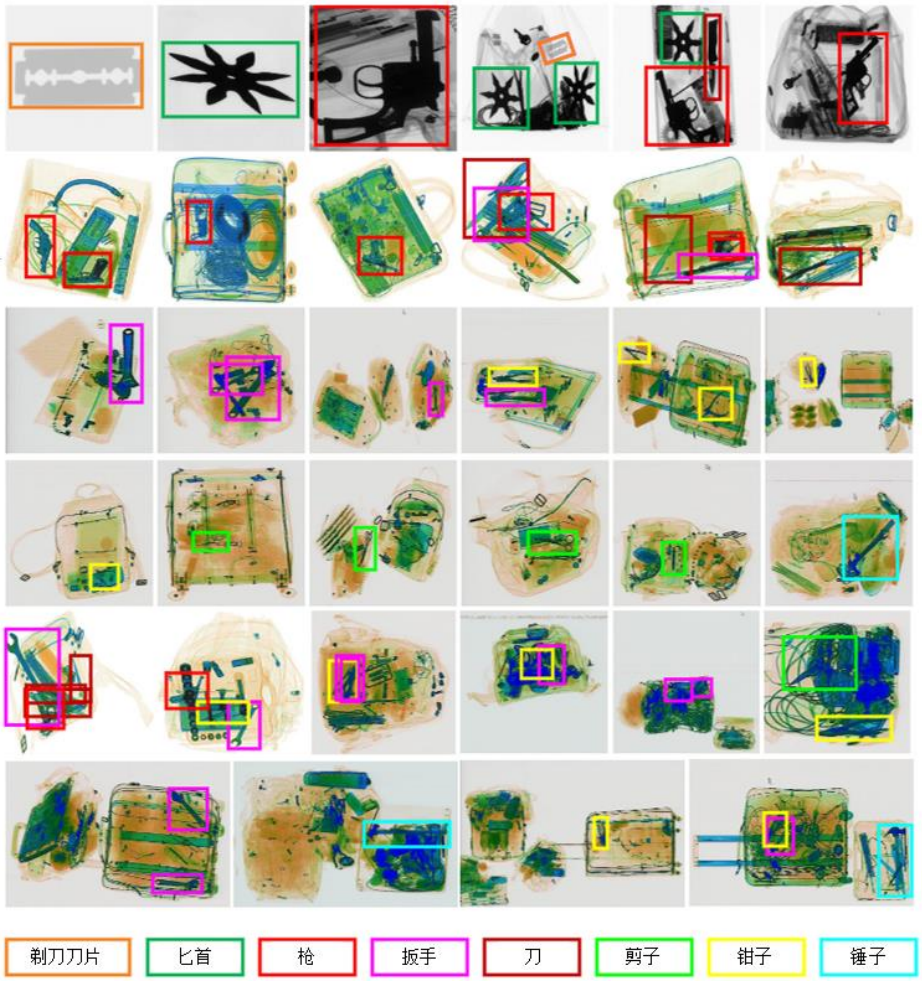


图 5.1 X 光图像中的各类违禁品示例

Figure 5.1 Examples of prohibited items in X-ray images.

安检领域违禁品目标定位问题是计算机视觉在实际应用场合中的一个典型应用，违禁品目标自动发现对于辅助安检人员进行违禁品定位、提高安检与通

关效率具有重大意义。安检领域违禁品检测的 X 光图像如图 5.1 所示。

X 光安检图像是由 X 光射线对目标穿透成像，通过计算 X 光的穿透率等反向计算生成 X 光图像，其反应的是各类物理材质对 X 光的吸收情况，因此成像结果只和目标实际的材质有关。在实际工作中，物体的摆放经常会相互重叠，这一特点使得 X 光图像具有非常显著的重叠特性。同时由于物体在成像时存在着多角度、多视角以及多尺度等问题，违禁品和很多背景噪声难以区分。X 光安检数据的标注往往需要专业人员的参与，标注过程十分耗时耗力。同时 X 光安检数据非常庞大，但是实际违禁品出现的概率非常低，这使得违禁品数据集的搜集十分困难。在大规模的数据集中，标注的难度非常大，同时包含违禁品的图像（正例）和不包含违禁品的图像（反例）的比例非常大，这导致传统的机器学习算法在该问题上会失效。

为了解决这一问题，本章提出了弱监督的 X 光图像违禁品定位的框架。在该框架中标注者只需要给出图像中有无违禁品以及违禁品的类别，不需要对违禁品的位置精确标注，从而极大的减少了标注工作量，使得深度学习在大规模 X 光数据集上的使用成为可能。

5.2 弱监督 X 光违禁品定位网络

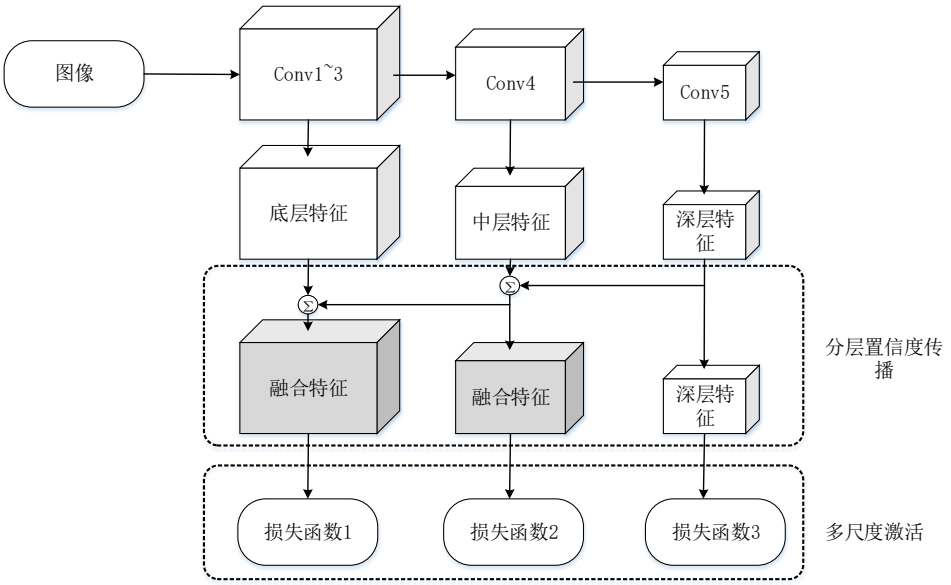


图 5.2 类平衡分层激活网络结构示意图

Figure 5.2 The overall architecture of the class-balanced hierarchical network

5.2.1 分层置信度传播

分层置信度传播的作用是利用卷积神经网络的分层特性，对不同层之间的特征进行传播，通过融合不同之间的特性来达到增强特征表示的作用。分层置信度传播包括层间传播和层内传播两个部分。

层间传播：层间传播采用特征金字塔的结构，通过融合不同层来实现多层信息融合，实现特征对尺度的鲁邦性。卷积神经网络中第 l+1 层的置信度图首先通过上采样的方式，使得其和第 l 层的置信度图有相同的空间分辨率，然后再和第 l 层特征图进行级联，接着通过一个 1×1 的卷积层将两层的特征进行了融合。由此 l+1 层的置信度传播至 l 层，传播公式如下

$$M^l \leftarrow W^l * \wedge\left(M^{l+1}, F^l\right) \tag{5.1}$$

其中，$F^l \in R^{K \times N \times N}$ 表示第 l 层卷积特征图有 K 个维度并且每个维度的大小为 $N \times N$，$\wedge(\cdot)$ 表示将 l+1 层的特征图 M^{l+1} 上采样后和第 l 层特征图 F^l 进行级联的操作。W^l 是 1×1 卷积层中的参数，“*”表示卷积操作。当 L 是卷积层数时，也就是最后一个卷积层时，$M^L = F^L$。

层内传播：由公式 5.1 可以看出，置信度传播图 M^l 是在自深而浅传播的过程中由第 l 层和第 $l+1$ 层的特征图融合得到。在得到置信度传播图之后，我们对该特征图进行层内传播，通过特征之间的位置关系和相似度来增强特征表达。层内传播通过使用传播图和周围像素交互作用更新置信度传播图，主要作用在于通过抑制噪声、聚焦相关区域得到更准确的违禁品定位信息。将每一个卷积层的像素看作一个马尔科夫链，第 l 层的传播图 $A^l \in R^{N \times N}$ 通过随机游走的方法计算得到，该传播图通过转换概率矩阵 $G^l \in R^{N^2 \times N^2}$ 不断循环的更新每个像素的状态。当马尔科夫链的平衡分布通过不断积累使得像素与其周围像素具有很高不相似度时，A^l 将会达到一个稳定的状态。

$v_{ij}^l \in R^K$ 表示置信度传播图 M^l 中第 K 个维度上位置 (i, j) 对应的向量，转

换概率矩阵通过M^l中像素之间的连接得到。定义两个像素(i,j)和$(i^{'},j^{'})$间的转换概率矩阵：

$$G^l_{(i,j),(i^{'},j^{'})}=\left\|v^l_{ij}-v^l_{i^{'}j^{'}}\right\|\cdot D\left((i,j),(i^{'},j^{'})\right) \tag{5.2}$$

$\|\cdot\|$表示$L2$正则化，正则化的空间距离公式：

$$D\left((i,j),(i^{'},j^{'})\right)=\exp\left(\left(i-i^{'}\right)^2+\left(j-j^{'}\right)^2\right)\Big/\sigma^2 \tag{5.3}$$

σ表示距离参数根据经验值在实验中设置为$0.2\times N$。A^l中的每个元素都被初始化为$1\big/_{N^2}$，根据A^l当前的状态和转换概率矩阵G^l就可以得到A^l的下一个状态，重复此操作直到A^l到达一个稳定的状态，我们就可以得到传播图A^l，置信度特征图M^l根据如下公式进行更新：

$$M^l_k\leftarrow A^l\otimes M^l_k,\forall \mathrm{k}\in\{1,2,...,K\} \tag{5.4}$$

$\otimes$表示对应像素相乘，M^l_k表示M^l的第k个维度。层内传播基础是基于深层特征中的邻近像素呈现出语义相关性以及同一个类别的像素有相似的特征向量。相当于采用软分割的过程来聚合之前的激活。

5.2.2 多尺度激活

激活的过程是由弱监督驱动的，图像的标号是对整个网络的监督。在弱监督定位的任务中，激活特征图通过激活深层卷积特征图上的显著性区域得到，该过程是为了发现图像中目标的位置。然而直接将图像分类的网络用来做定位的任务还是有一些缺点：（1）深层的神经元对应着原图很大的面积但是空间精度较低；（2）浅层的神经元有更精确的定位但是其感受野比较小，只能看到原图中局部的信息。受监督学习定位方法的启发，本文采用分层激活的网络结构相当于在多层激活。

对于类别c，其l层的激活图定义为：

$$T^l_c=\sum\nolimits_k w^c_k M^l_k \tag{5.5}$$

w_k^c 定义为最后一个全连接层中类别 c 与神经元 k 之间的权重。在本文中根据公式（5.3）和（5.4）可以得到第 l 层的激活图：

$$T_c^l = \sum_k w_k^c \left(A^l \otimes \left(W^l * \wedge \left[M^{l+1}, F^l \right] \right)_k \right) \tag{5.6}$$

传统方法 CAM 和 SPN 都是只利用最后一层卷积层的特征图计算得到的单一的激活图。而 HPA 可以从分层的卷积特征图中得到更丰富的可以用于图像分类和违禁品定位的特征信息。

5.3 实验结果及分析

类平衡分层激活网络的基网是当前流行的卷积神经网络，为了验证本章提出的方法对正反例不均衡问题的有效性，从 SIXray 数据集中选择了正反样本比例不同的三个子集，然后在多个子集上进行算法性能的评测，本节首先介绍实验中基本参数设置以及评测方法，其次介绍 CHR 与其他方法在 SIXray 数据子集上的对比实验，然后验证模型各个模块的效果以及对类别不均衡问题进行深入分析，最后评测 CHR 模型在自然场景数据集上的实验性能。

5.3.1 实验设置及评测

为了探究图像正反例不均衡问题，我们根据正例和反例的不同比例从 SIXray 数据集中选取三个不同的子集，命名为 SIXray10，SIXray100，SIXray1000 分别对应的是反例与正例的比例为 10,100,1000。在 SIXray10 和 SIXray100 的数据集中，所有的 8929 个正例都被使用，然后从反例图像中随机取 10 倍、100 倍的反例图像，SIXray100 的数据分布与真实场景中的分布是最接近的。为了最大程度探索我们的算法处理图像不均衡问题的能力，我们构造了 SIXray1000 数据集，该数据集是由随机从每个类别中选择的 1000 个正例图像以及所有的 1,050,302 反例图像构成。每一个子数据集都会被随机分为 train 和 test，其中 train 占总数据集的 80%，test 占总数据集的 20%，训练集和测试集图像比例为 4:1。划分后对于每一个方法来说训练和测试数据都是固定的，对于评测，分类采用评测类别的平均精度即 mAP，定位采用评测点定位的正确率。

在实验中，我们采用 5 个常用的网络结构包括 ResNet34 层、50 层和 101

层网络结构，Inception-v3[93]和 DenseNet121[94]。我们在这些网络基础上添加设计的模块，并且在 CHR 中令 L=3 即我们只利用三层的特征进行融合，当然如果增大 L 采用更多的特征也是可以的，但是在实验中我们发现 L=3 已经提供了足够多的的特征信息。

5.3.2 数据集简介

SIXray 数据集中的图像均来源于地铁站，该数据集共包含 1,059,231 幅 X 光安检图像，其中含有违禁品的图像有 8,929 幅，违禁品类别主要包含六种：枪、刀、扳手、钳子、剪刀以及锤子。如图 5.3 所示为 SIXray 数据集中的 X 光安检图像，在图中依次给出了包含六个类别违禁品的图像，图中红色的圆圈是标出来的违禁品的位置，最右边一列给出的是不含有违禁品的图像。从图中可以看出该数据集中的图像与实际安检场景中的图像是一致的。都是由 X 光机生成的伪彩色图像。不同材质的物体会被投影为不同的颜色。相同材质的物体会被投影为相同颜色。至少包含一种违禁品的图像称为正例图像，不包含任何违禁品的图像称为反例图像。

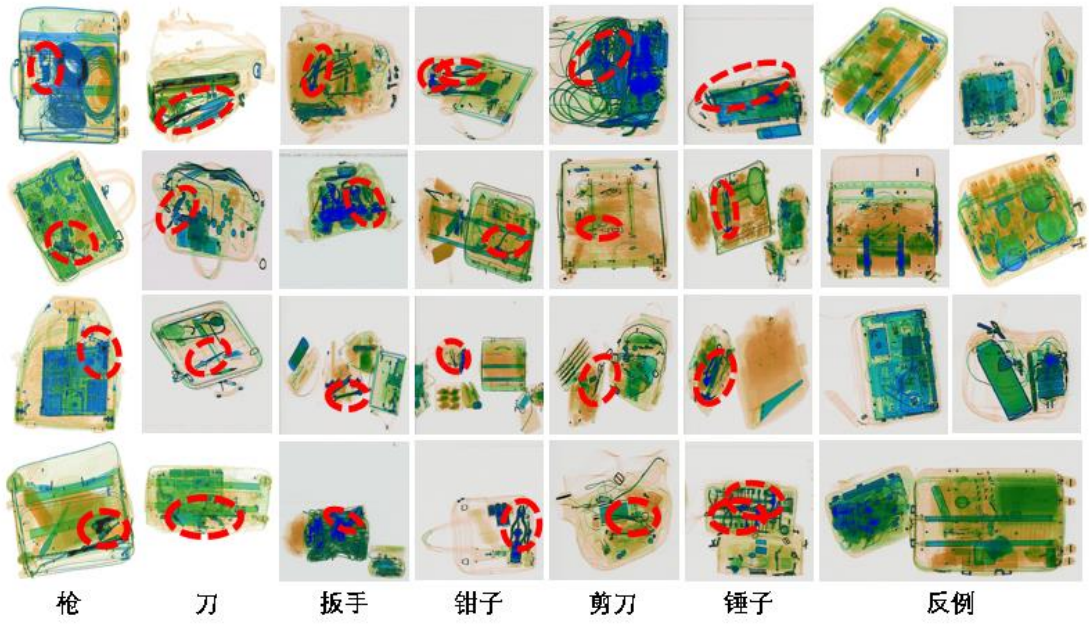


图 5.3 SIXray 数据集图像展示

Figure 5.3 Image examples in SIXray dataset

如表 5.1 所示为 SIXray 数据集中各个类别图像数目的统计信息，从表中

表 5.1 SIXray 数据集统计表

Table 5.1 SIXray dataset statistics

SIXray 数据集 (1,059,231)						
正例 (8,929)						反例
枪	刀	扳手	钳子	剪刀	锤子	
3,131	1,943	2,199	3,961	983	60	1,050,302

可以看出：正例图像数目很少只有将近 9,000 幅，而反例图像的数目是非常巨大的，有 105 万幅左右，可以看出正反样本的比例存在着严重的不平衡性。但是这样的数据分布跟现实场景中的数据分布是一致，在现实场景中可能安检一天也不会查到一两个违禁品，所以数据本身就是不含违禁品的图像数量多，包含违禁品的图像数量少。从表中可以看出，不同类别的违禁品的数目存在着较大的差异，锤子这一类只有 60 幅图像，相对于整个数据集的 100 多万幅太少了，所以在实验中并没有用锤子这一个类别的数据，即实验中正例图像只有 5 类。从表中发现：所有类别的图像相加总和是要大于正例图像的总数的，说明在正例图像中存在一幅图像中包含多个违禁品的情况。所有图像都人工标注了图像级别的标注信息，即图像中是否含有违禁品，并且为了评估模型的定位性能，在测试集上给出了违禁品的 bounding box 信息，将其存入 xml 文件中。数据集中图像的平均大小为 100K 像素，所有的图像都被存为 JPEG 格式。

我们对测试集中目标的信息做了更详细的统计如图 5.4 中分别展示了目标的角度、长宽比和面积的分布情况，从图中可以看出在 0 到 180 度范围内角度的分布比较均匀，图像中的违禁品存在着角度多样性；长宽比和面积的分布相对比较集中，长宽比大多数都小于 2，目标的面积大多数都小于 5 万像素点。

SIXray 数据集主要有以下特点：一、SIXray 数据集中的图像都是过安检时通过用 X 射线扫描旅客的包裹得到的，包裹内的物品都是随机摆放的，当 X 射线对包裹进行扫描时，X 射线的穿透属性甚至可以看到包裹中被遮挡的部分，也就是之前介绍的 SIXray 数据集跟自然场景数据集不同：重叠特性；二、违禁品存在着多角度、多视角的问题，甚至同一个类中的违禁品也存在着一些不同的子类，这些都会导致类内差异性增大，增大安检图像识别的难度；三、图像内容非常杂乱，因为包裹中可能什么东西都有，很难预料在背景区域会出现什

么，使得背景信息非常复杂；四、如之前所述，正例图像数量非常少，使得网络在训练过程中很容易将更多的参数用来学习反例样本的信息，因为将所有的数据都预测为反例，也会有一个非常高的正确率，所以如何使保持训练的过程更稳定是一个大的挑战。

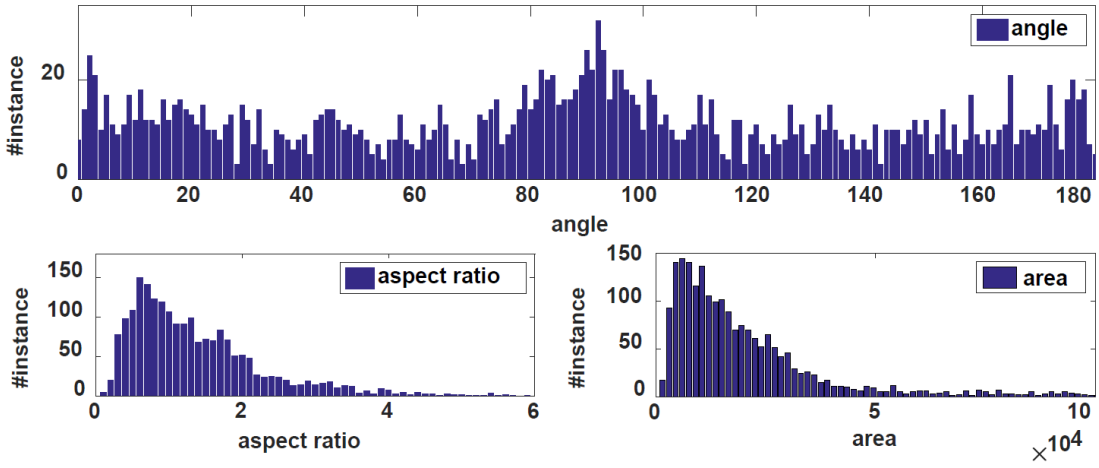


图 5.4 SIXray 测试集中目标角度、长宽比和面积分布

Figure 5.4 Distribution of object angle, aspect ratio and area in SIXray test set

5.3.3 分类和定位实验

SIXray10、SIXray100、SIXray1000 的分类实验结果分别在表 5.2 表 5.3 表 5.4 中，从表中可以看到 CHR 在每个数据集上均取得了比基本实验网络好的性能，从表格中观察发现 CHR 在更深的网络中会有更好的性能，对于 Inception-v3 和 DenseNet，CHR 在 SIXray1000 数据集上分别有 8.22%和 9.08%性能的提升。

接下来分别对五个类别的性能进行分析，CHR 在每一类上性能的提升是不相同的，以 DenseNet 为例，对于枪这一类，分类的性能在每个数据集上都没有提升，这是因为枪这一类训练样本较多，增加了训练样本较少的类的权重。但是我们观察到：对于除枪以外的其他类别都是有性能提升的，尤其对于剪刀这一类，性能最高可以提升 30%。从表 5.1 中可以看出，剪刀这一类别的图像数

表 5.2 SIXray10 **数据集上分类性能**

Table 5.2 Classification performance on SIXray10 dataset

方法	枪	刀	扳手	钳子	剪刀	平均
ResNet34[93]	89.71	85.46	62.48	83.50	52.99	74.83
ResNet34+CHR	87.16	87.17	64.31	85.79	61.58	77.20
ResNet50[93]	90.64	87.82	63.62	84.80	57.35	76.85
ResNet50+CHR	87.55	86.38	69.12	85.72	60.91	77.94
ResNet101[93]	87.65	84.26	69.33	85.29	60.39	77.38
ResNet101+CHR	85.45	87.21	71.23	88.28	64.68	79.37
Inception-v3[94]	90.05	83.80	68.11	84.45	58.66	77.01
Inception-v3+CHR	88.90	87.23	69.47	86.37	65.50	79.49
DenseNet[95]	87.36	87.71	64.15	87.63	59.95	77.36
DenseNet+CHR	87.05	85.89	70.47	88.34	66.07	79.56

表 5.3 SIXray100 数据集上的分类性能

Table 5.3 Classification performance on SIXray100 dataset

方法	枪	刀	扳手	钳子	剪刀	平均
ResNet34[93]	83.06	78.75	30.49	55.24	16.14	52.74
ResNet34+CHR	81.96	77.70	36.85	64.56	14.49	55.11
ResNet50[93]	84.75	77.92	28.49	50.53	19.39	52.22
ResNet50+CHR	82.64	79.60	41.19	58.02	27.89	57.87
ResNet101[93]	82.83	76.16	35.59	54.82	20.63	54.01
ResNet101+CHR	83.25	77.53	42.02	68.01	32.33	60.63
Inception-v3[94]	81.18	77.28	32.47	66.89	22.63	56.09
Inception-v3+CHR	79.22	73.48	37.20	69.01	31.81	58.15
DenseNet[95]	83.23	77.24	37.72	62.69	24.89	57.15
DenseNet+CHR	82.06	78.75	43.22	66.75	28.80	59.92

表 5.4 SIXray1000 数据集分类性能

Table 5.4 Classification performance on SIXray1000 dataset

方法	枪	刀	扳手	钳子	剪刀	平均
ResNet34[93]	72.05	56.42	16.47	14.24	7.12	33.26
ResNet34+CHR	73.35	60.46	23.72	17.98	18.19	38.74
ResNet50[93]	74.19	59.82	16.03	16.59	2.87	33.90
ResNet50+CHR	73.43	61.32	18.88	12.32	19.03	37.00
ResNet101[93]	76.04	63.53	13.65	15.57	11.28	36.01
ResNet101+CHR	75.38	64.80	15.27	19.02	16.21	38.14
Inception-v3[94]	75.52	56.33	24.01	16.75	20.72	38.67
Inception-v3+CHR	76.91	61.29	29.60	19.11	47.56	46.89
DenseNet[95]	75.00	65.55	23.57	18.09	14.18	39.28
DenseNet+CHR	74.87	71.23	29.79	21.57	44.27	48.36

量是五个类别中最少的，所以在训练的过程中会有更多的权重，同时 CHR 通过添加分层的监督信息极大的抑制了干扰信息。

最后，我们对不同数据子集上存在的正反例样本不均衡问题做了具体的分析，SIXray10、SIXray100 和 SIXray1000 中分别对应着反例与正例的比例为 10、100 和 1000，从图 5.5 中可以看出类平衡分层激活网络在分类和定位的性能上都比基网络方法有明显提高，除此之外还发现，随着反例和正例样本比例的增大，我们提出的方法比基网络方法性能提升幅度也在对应增大，充分体现了 CHR 算法对于处理类别不均衡问题的有效性。

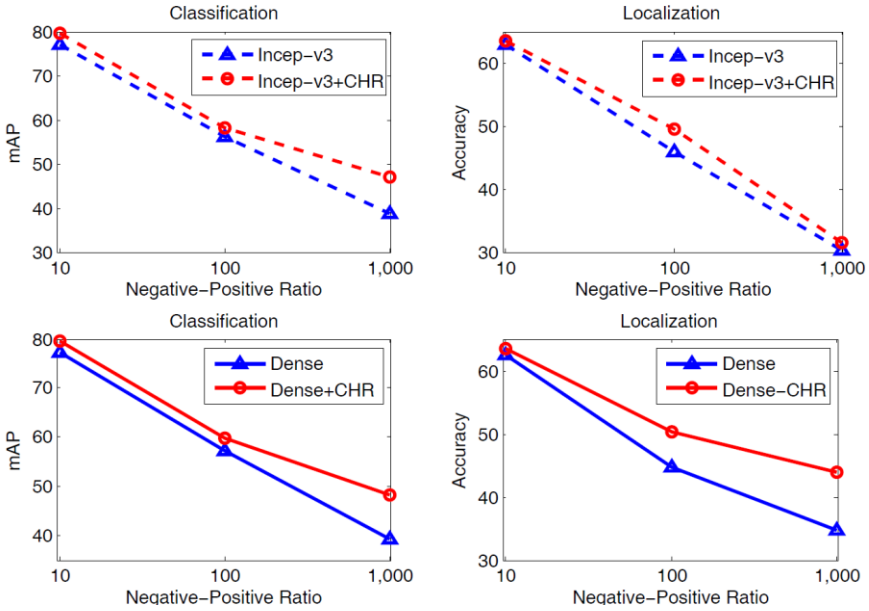


图 5.5 不同正反类别比例下的分类性能对比

Figure 5.5 The accuracy gain of CHR becomes better with larger negative-positive ratio

为了验证类平衡分层激活网络不仅在分类实验中性能有较大的提升，我们借鉴 CAM 生成类别响应图。 表 5.5、表 5.6、表 5.7 中展示了算法在不同数据子集上的定位性能。在 SIXray100 子集中，CHR 比 DenseNet 性能高 5.61%（50.31% vs 44.70%），在 SIXray1000 子集中，DenseNet+CHR 比 DenseNet 性能高 9.26%（43.87% vs 34.61%）。尤其对于 SIXray1000 数据集中扳手这个类别，Inception-v3+CHR 比 Inception-v3 性能提高 16.04%（23.53%vs7.49%），除此之外，可以发现越深的网络结构会有更好的性能。

表 5.5 SIXray10 数据集定位性能

Table 5.5 Localization performance on SIXray10

方法	枪	刀	扳手	钳子	剪刀	平均
ResNet34[93]	71.60	51.28	43.32	68.88	22.16	51.45
ResNet34+CHR	75.62	55.38	52.41	58.44	19.32	52.23
ResNet50[93]	63.89	57.44	49.73	68.88	17.05	51.40
ResNet50+CHR	68.83	58.46	54.01	77.04	15.91	54.85
ResNet101[93]	73.77	65.13	28.34	62.24	21.02	50.10
ResNet101+CHR	80.86	73.85	52.41	9.30	40.34	51.35
Inception-v3[94]	79.94	75.38	59.36	59.58	40.34	62.92
Inception-v3+CHR	78.70	74.36	52.41	59.96	52.27	63.54
DenseNet[95]	74.38	71.28	59.89	71.54	35.23	62.46
DenseNet+CHR	79.01	76.92	59.36	72.49	40.34	65.62

表 5.6 SIXray100 数据集定位性能

Table 5.6 Localization performance on SIXray100

方法	枪	刀	扳手	钳子	剪刀	平均
ResNet34[93]	50.62	55.38	26.74	34.54	7.95	35.05
ResNet34+CHR	60.19	63.08	35.83	53.70	0.00	42.56
ResNet50[93]	47.53	52.82	28.34	39.85	1.70	34.05
ResNet50+CHR	57.72	49.23	41.18	49.91	15.34	42.67
ResNet101[93]	73.15	64.10	25.13	31.50	11.36	41.05
ResNet101+CHR	79.32	69.23	27.81	48.39	6.25	46.20
Inception-v3[94]	64.81	65.64	40.11	32.83	26.14	45.91
Inception-v3+CHR	67.59	63.08	23.53	54.27	39.20	49.53
DenseNet[95]	71.60	62.05	24.60	55.60	9.66	44.70
DenseNet+CHR	78.40	62.56	41.71	63.76	5.11	50.31

表 5.7 SIXray1000 数据集定位性能

Table 5.7 Localization performance on SIXray1000

方法	枪	刀	扳手	钳子	剪刀	平均
ResNet34[93]	53.93	38.97	22.46	13.69	6.82	27.17
ResNet34+CHR	70.41	26.15	37.97	25.10	2.27	32.38
ResNet50[93]	42.32	48.72	19.79	19.77	2.84	26.69
ResNet50+CHR	60.67	37.44	22.46	20.91	13.64	31.02
ResNet101[93]	70.41	60.00	15.51	14.07	5.68	33.13
ResNet101+CHR	79.03	61.54	21.93	17.11	19.32	39.78
Inception-v3[94]	71.16	52.31	7.49	18.63	1.70	30.26
Inception-v3+CHR	73.41	41.54	23.53	7.60	11.36	31.49
DenseNet[95]	58.05	56.92	26.20	20.53	11.36	34.61
DenseNet+CHR	76.78	57.95	39.04	39.92	5.68	43.87

5.3.4 模型验证实验

在 SIXray 数据集上对各个模块进行分析，首先对于分层细化的网络结构，自顶向下的连接（ResNet34+HR）在 SIXray100 数据集上比（ResNet34+H）分类和定位精度分别提高 1%和 6.52%，在 SIXray1000 数据集上分别提升 3.15%和 2.13%，主要原因是后者提供的更多是低维特征信息。我们分析了不同损失函数影响，如表 5.8 所示 ResNet34+CH 表示添加了类别均衡损失函数，分类和定位的结果分别提升 1.00%和 3.77%在 SIXray100 数据集、3.10%和 3.44%在 SIXray1000 数据集上。通过将分层细化的网络结果与类平衡损失函数结合（ResNet34+CHR），分类和定位性能在 SIXray100 数据集上分别提升 2.37%和 7.51%，在 SIXray1000 数据集上分别提升 5.48%和 5.11%比基网络模型 ResNet34。性能的提升只需要比较少的额外的计算量。ResNet34 需要 7.68ms 处理一幅测试图像，而 ResNet34-CHR 需要 8.28ms 测试环境均在 Tesla V100 GPU 仅需要 7.81%额外的时间开销。

表 5.8 CHR 在 SIXray 数据集上的分类和定位性能

Table 5.8 Classification and localization performance using different options of CHR

方法	SIXray10		SIXray100		SIXray1000	
ResNet34	74.83	51.45	52.74	35.05	33.26	27.17
ResNet34+H	74.43	49.91	53.59	38.70	34.78	28.68
ResNet34+CH	76.28	48.01	54.59	42.47	37.87	32.12
ResNet34+HR	75.87	50.19	53.72	41.57	36.41	29.30
ResNet34+CHR	77.20	52.23	55.11	42.56	38.74	32.28

5.4 本章小结

在本章中，我们提出了类平衡分层激活网络，该模型通过增加深层特征对中间层特征的监督，使得中间层特征能够得到更精细的视觉线索并且能够过滤掉一些不相关的信息。除此之外，设计了类别均衡的损失函数，通过减少反例样本的数量，尽量的使正例和反例的数量达到一个平衡，并且该损失函数依赖于分层的网络结构，使得深层侧输出的损失函数对浅层侧输出的损失函数有指导作用。在 SIXray 数据集中三个子集上，与多种方法进行对比实验后发现，本章提出的类平衡分层激活网络性能有显著提升，除此之外，在自然场景数据集

上，我们的方法也表现出了比较好的性能。

本章节的研究面向实际应用场景，通过集成前面章节实现的弱监督目标检测方法，并且进一步加入处理类别不平衡的策略，实现了弱监督目标检测在实际场景中的应用。

第6章 总结与展望

6.1 本文工作总结

本文围绕弱监督目标检测问题，从建模、优化和实际应用多个方面进行了研究，并提出了系统的方法和算法框架。

（1）提出了一种有效的最小熵隐变量模型，用于弱监督目标检测任务。算法在训练阶段的定位随机性得到了降低，因此能够更稳定的学习目标特征，提升目标定位的准确性。最小熵隐变量模型的贡献总结如下：一是采用深度神经网络结合最小熵隐变量模型以便更有效地挖掘到目标候选框，并且最小化学习过程中的定位随机性；二是采用一个候选框团更好地搜集目标的信息并激活完整的目标区域，从而能够更准确的检测到目标。三是用一个循环学习算法分别将图像分类和目标检测看做一个预测和校正，并且利用连续优化的方法解决非凸优化问题。四是在几个常用的公开数据集上取得了 state-of-the-art 的分类、定位和检测性能。

（2）提出了一种有效的弱监督学习优化方法，称为渐进多示例学习（CMIL）。渐进多示例学习的方法致力于解决传统多示例学习方法的非凸优化问题。通过引入一个序列的平滑损失函数，在训练过程中以一个容易求解的凸损失函数为起点，逐渐优化该序列中的平滑损失函数，直至损失函数退化成原损失函数。该平滑过程是通过引入示例子集的方式完成的。渐进多示例学习显著提升了弱监督目标检测和弱监督目标定位的性能，并超过了最新已发表的工作。当使用渐进优化模型和深度网络结合时，模型在训练过程中通过搜集目标或者目标部件的方式激活了目标的完整区域，从而最终学习到语义稳定极值区域。所提出的面向弱监督学习的优化方法拓展了相关计算视觉问题的研究思路。

（3）提出了类平衡分层激活网络，该模型通过增加深层特征对中间层特征的监督，使得中间层特征能够得到更精细的视觉线索并且能够过滤掉一些不相关的信息。设计了类别均衡的损失函数，通过减少反例样本的数量，尽量的使正例和反例的数量达到一个平衡，并且该损失函数依赖于分层的网络结构，使

得深层侧输出的损失函数对浅层侧输出的损失函数有指导作用。在实验室发布来源于实际安全检测场景的 SIXray 数据集上，所提出的渐进优化发放与类别均衡化的分层细化模型体现出实际应用价值。

6.2 未来工作展望

弱监督视觉目标检测不仅具有显著的科学意义，而且具有明确的社会经济价值。但是，现有弱监督视觉目标检测算法和全监督的检测算法在性能上还有较大的差距，阻碍了其在计算机视觉实际任务中的应用。未来可以从本质上解决传统弱监督目标检测的固有问题，在性能上减少和全监督模型的距离。具有潜力的研究问题与方向包括：

（1）弱监督语义稳定极值区域学习：语义稳定极值区域学习和目标检测算法利用了目标区域和背景区域的语义分布特性，是一种新的目标定位的判断依据。该方法对将弱监督视觉目标检测的问题转换为语义稳定极值区域的搜索问题，为弱监督视觉目标检测提供了全新的解决思路和方法论。

（2）无候选框弱监督视觉目标检测算法：无候选框弱监督视觉目标检测算法是首个一阶段（one-stage）弱监督目标检测方法。该方法不仅能解决传统弱监督算法对候选框算法的依赖而降低了检测效率的问题，还能保证候选框的查全率，减少模型学习过程中噪声样本的影响。该模型为弱监督目标检测算法提供了全新的框架，对弱监督算法的实际应用有着非常重要的意义。

（3）弱监督主动学习视觉目标检测。结合主动学习视觉目标检测是将弱监督算法推向实际应用的重要步骤。主动学习的研究关注点在于以最小的人工标注量，来达到使用了所有标注量的相似性能。该问题通过少量的人机交互，结合模型的训练策略。然而，传统的主动学习算法中，用户反馈的标注信息均为全监督的格式，这对用户有着较高的要求。相比于全监督框架下的主动学习算法，弱监督的主动学习框架极大的降低了用户的标注门槛，减少了用户标注的工作量，同时极大的降低了标注之间的歧义和误解。在面对海量数据搜集数据时，该方向具有显著的优势。

参考文献

[1] 袁国武. 智能视频监控中的运动目标检测和跟踪算法研究[D]，云南大学, 2012.

[2] Aggarwal J. K., Ryoo M. S. Human Activity Analysis: A Review [J]. ACM Computing Surveys. 2011, 43(3):194-218.

[3] Datta, R., Joshi, D., Li, J., Wang, J. Z. Image Retrieval: Ideas, Influences, and Trends of The New Age [J]. ACM Computing Surveys. 2008, 40(2), Article 5.

[4] Volker K., Danica K., Aleš U., Christopher G. The meaning of action: a review on action recognition and mapping [J]. Advanced Robotics. 2007, 21(21): 1473-1501.

[5] Palmese M., Trucco A. From 3-D Sonar Images to Augmented Reality Models for Objects Buried on The Seafloor [J]. IEEE Transactions on Instrumentation and Measurement. 2008, 57(4): 820-828.

[6] Bin W., 目标检测简要综述[Z]. 2015-6, http://imbinwang.github.io/blog/object-detection-review.

[7] Dalal N., Triggs B. Histograms of Oriented Gradients for Human Detection [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2005, 1: 886-893.

[8] Felzenszwalb, P., Girshick, R., Mcallester, D., Ramanan, D. Object Detection with Discriminatively Trained Part-based Models [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010, 32(9), 1627-1645.

[9] Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.

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

[11] 蔡强, 刘亚奇, 曹健,等. 图像目标类别检测综述[J]，计算机科学与探索，2015, 9(3): 257-265.

[12] Szegedy C., Toshev A., Erhan D. Deep Neural Networks for Object Detection [C]. Advances in Neural Information Processing Systems. 2013: 2553-2561.

[13] Zhu X., Goldberg A. B. Introduction to Semi-supervised Learning [J]. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2009, 3(1):1–130.

[14] Andrews S., Tsochantaridis I., Hofmann T. Support Vector Machines for Multiple-instance Learning [C]. Advances in Neural Information Processing Systems. 2002: 561-568.

[15] Cinbis R. G., Verbeek J., Schmid C. Multi-fold MIL Training for Weakly Supervised Object Localization [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2409-2416.

[16] Bilen H., Pedersoli M., Tuytelaars T. Weakly Supervised Object Detection with Posterior

Regularization [C]. British Machine Vision Conference. 2014, 147(8):1997-2005.

[17] Bilen H., Pedersoli M., Tuytelaars T. Weakly Supervised Object Detection with Convex Clustering [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1081-1089.

[18] Liang, X., Liu, S., Wei, Y., Liu, L., Lin, L., Yan, S. Towards Computational Baby Learning: A Weakly-Supervised Approach for Object Detection [C]. IEEE International Conference on Computer Vision. 2015:999-1007.

[19] Ren W., Huang K., Tao D., Tan T. Weakly Supervised Large Scale Object Localization with Multiple Instance Learning and Bag Splitting [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016, 38(2):405-416.

[20] Uijlings J., Sande K., Gevers T., Smeulders A. Selective Search for Object Recognition [J]. International Journal of Computer Vision. 2013, 104(2): 154-171.

[21] Cheng M. M., Zhang Z., Lin W. Y., Torr. BING: Binarized Normed Gradients for Objectness Estimation at 300fps [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2014:3286-3293.

[22] Arbelaez, P., Pont-Tuset, J., Barron, J., Marques, F., Malik, J. Multiscale Combinatorial Grouping. IEEE Conference on Computer Vision and Pattern Recognition. 2014:328-335.

[23] Rantalankila P., Kannala J., Rahtu E. Generating Object Segmentation Proposals using Global and Local Search [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2417-2424.

[24] Zitnick C. L., Dollár P. Edge Boxes: Locating Object Proposals from Edges [C]. Springer, European Conference on Computer Vision. 2014: 391-405.

[25] Matas, J., Chum, O., Urban, M., Pajdla, T. Robust Wide-baseline Stereo from Maximally Stable Extremal Regions. Image and Vision Computing. 2004, 22(10): 761-767.

[26] Viola P., Jones M. Rapid Object Detection using A Boosted Cascade of Simple Features [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2001, 1(I): 511-518.

[27] Ojala T., Pietikäinen M., Harwood D. A Comparative Study of Texture Measures with Classification Based on Featured Distributions [J]. Pattern Recognition. 1996, 29(1): 51-59.

[28] Wang X., Han T. X., Yan S. An HOG-LBP Human Detector with Partial Occlusion Handling [C]. Computer Vision, IEEE International Conference on Computer Vision. 2009: 32-39.

[29] Lowe D. G. Object Recognition from Local Scale-invariant Features [C]. IEEE International Conference on Computer Vision. 1991, 2: 1150-1157.

[30] Hinton G. A Practical Guide to Training Restricted Boltzmann Machines [J]. Momentum. 2010, 9(1): 926.

[31] Krizhevsky A., Sutskever I., Hinton G. E. Imagenet Classification with Deep Convolutional Neural Networks [C]. Advances in Neural Information Processing Systems. 2012: 1097-1105.

[32] Donahue J., Jia Y., Vinyals O., Hoffman J., Zhang N., Tzeng E. Decaf: A Deep Convolutional

Activation Feature for Generic Visual Recognition [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2013: 647-655.

[33] Schapire R. E., Singer Y. Improved Boosting Algorithms using Confidence-rated Predictions [J]. Machine Learning. 1999, 37(3): 297-336.

[34] Cortes C., Vapnik V. Support Vector Machine [J]. Machine Learning. 1995, 20(3): 273-297.

[35] Girshick R. Fast R-CNN [C]. IEEE International Conference on Computer Vision. 2015: 1440-1448.

[36] Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks [C]. Advances in Neural Information Processing Systems. 2015: 91-99.

[37] Hyun O., Ross G., Stefanie J., Julien M., Zad H., Trevor D. On Learning to Localize Objects with Minimal Supervision. International Conference on Machine Learning. 2014: 1611-1619.

[38] Ramazan G., Jakob V., Cordelia S. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning. IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2409-2416.

[39] Xiaodan L., Si L., Yunchao W., Luoqi L., Liang L., Shuicheng Yan. Towards Computational Baby Learning: A Weakly-Supervised Approach for Object Detection, IEEE International Conference on Computer Vision. 2015: 999-1007.

[40] Weiqiang R., Kaiqi H., Dacheng T., Tieniu T. Weakly Supervised Large Scale Object Localization with Multiple Instance Learning and Bag Splitting. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016, 38(2): 405-416.

[41] Chong W., Weiqiang R., Kaiqi H., Tieniu T. Weakly Supervised Object Localization with Latent Category Learning. Springer, European Conference on Computer Vision. 2014: 431-445.

[42] 程圣军，基于带约束随机游走图模型的弱监督学习算法研究[D]，博士论文，哈尔滨工业大学，2014 年 6 月.

[43] Le W., Gang H., Rahul S., Jianru X., Nanning Z. Video Object Discovery and Co-segmentation with Extremely Weak Supervision. Springer, European Conference on Computer Vision. 2014: 640-655.

[44] 赵永威，李弼程，柯圣财，基于弱监督 E2LSH 和显著图加权的目标分类方法[J]，电子信息学报，2016, 38(1): 38-46.

[45] 岳亚伟，基于弱监督空间金字塔模型的图像分类研究[D]，山东大学博士论文，2013 年 4 月.

[46] 陈燕，耿国华，贾晖，基于密度中心图的弱监督分类方法[J]，计算机工程与应用，2015, 51(6): 6-10.

[47] Yu-Feng L., Zhi-Hua Z. Towards Making Unlabeled Data Never Hurt. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015, 37(1): 175-188.

[48] 杨杰，孙亚东，张良俊，刘海波，基于弱监督学习的去噪受限玻尔兹曼机特征提取算法[J]，电子学报，2014, 12(2): 22-34.

[49] Chong W., Kaiqi H., Weiqiang R., Junge Z., Stephen J. Maybank: Large-Scale Weakly Supervised Object Localization via Latent Category Learning. IEEE Transactions on Image Processing. 2015, 24(4): 1371-1385.

[50] Song, O., Yong, J.., Jegelka, S., Darrell, T. Weakly-supervised Discovery of Visual Pattern Configurations. Advances in Neural Information Processing Systems. 2014, 2: 1637-1645.

[51] Everingham M., Van Gool L., Williams C. The Pascal Visual Object Classes (VOC) Challenge [J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[52] Krahenbuhl P., Koltun V. Learning to propose objects [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1574-1582.

[53] Hubel D. H., Wiesel T. N. Receptive Fields, Binocular Interaction and Functional Architecture in The Cat's Visual Cortex [J]. The Journal of Physiology. 1962, 160(1): 106-154.

[54] Sivic J., Zisserman A. Efficient Visual Search of Videos Cast as Text Retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2009, 31(4): 591-606.

[55] Perronnin F., Dance C. Fisher Kernels on Visual Vocabularies for Image Categorization [C]. IEEE Conference on Computer Vision and Pattern Recognition. 2007: 1-8.

[56] Perronnin F., Sánchez J., Mensink T. Improving the Fisher Kernel for Large-scale Image Classification [C]. Springer, European Conference on Computer Vision. 2010: 143-156.

[57] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 2846–2854, 2016.

[58] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 5131–5139, 2017.

[59] Li Dong, Huang Jia Bin, Li Yali, Wang Shengjin, and Yang Ming Hsuan. Weakly supervised object localization with progressive domain adaptation. In Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 3512–3520, 2016.

[60] Mingfei Gao, Ang Li, Ruichi Yu, Vlad I Morariu, and Larry S Davis. C-wsl: Count-guided weakly supervised localization. 2018.

[61] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 4294–4302, 2017.

[62] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In Proc. Europ. Conf. Comput. Vis. (ECCV), pages 350–365, 2016.

[63] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In Proc. IEEE Int. Conf. Comput. Vis. Pattern

Recognit. (CVPR), pages 3059–3067, 2017.

[64] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In Proc. Europ. Conf. Comput. Vis. (ECCV), pages 352–368, 2018.

[65] FangWan, PengxuWei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 1297–1306, 2018.

[66] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: tight box mining with surrounding segmentation context for weakly supervised object detection. In Proc. Europ. Conf. Comput. Vis. (ECCV), pages 434–450, 2018.

[67] Qixiang Ye, Tianliang Zhang, Qiang Qiu, Baochang Zhang, Jie Chen, and Guillermo Sapiro. Self-learning scenespecific pedestrian detectors using a progressive latent model. In Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR), pages 2057–2066, 2017.

[68] B. A. Olshausen and D. J. Field, “Sparse Coding with an Over-complete Basis Set: A Strategy Employed by V1?” Vision Research, 37(23):3311–3325, 1997.

[69] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, “Extracting and Composing Robust Features with Denoising Autoencoders,” In Proc. Int’l Conf. Machine Learning, 2008.

[70] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” Science, 2006.

[71] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy Layer-wise Training of Deep Networks,” in Proc. of Neural Information Processing System, 2007.

[72] Q. V. Le, M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng., “Building High-level Features using Large-Scale Unsupervised Learning,” In Proc. Int’l Conf. Machine Learning, 2012.

[73] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in Proc. of Neural Information Processing System, 2014.

[74] D. Kingma and M. Welling, “Auto-encoding Variational Bayes,” In Proc. Int’l Conf. Learning Representation, 2014.

[75] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised Visual Representation Learning by Context Prediction,” in Proc. of IEEE Int’1 Conf. Computer Vision 2015.

[76] R. Zhang, P. Isola, and A. A. Efros, “Colorful Image Colorization,” In Proc. ECCV, 2016.

[77] R. Zhang, P. Isola, and A. A. Efros, “Split-brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction,” In Proc. IEEE Computer Vision and Pattern Recognition , 2017.

[78] D. Laptev, N. Savinov, J. M. Buhmann, and M. Pollefeys, “TI-POOLING: Transformation-Invariant Pooling for Feature Learning in Convolutional Neural Networks,” IEEE CVPR 2016.

[79] J. Bruna and S. Mallat, “Invariant Scattering Convolution Networks,” IEEE Trans. Pattern Anal. Mach. Intell. 35(8):1872–1886, 2013.

[80] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial Transformer Networks,” in Proc. of Neural Information Processing System, pp.2017–2025, 2015.

[81] P. A. Viola, M. J. Jones, D. Snow, “Detecting Pedestrians Using Patterns of Motion and Appearance,” International Journal of Computer Vision，63(2): 153-161, 2015.

[82] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, D. Ramanan, “Object Detection with Discriminatively Trained Part-Based Models,” IEEE Trans. Pattern Anal. Mach. Intell. 32(9): 1627-1645, 2010.

[83] R. B. Girshick, J. Donahue, T. Darrell, J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in Proc. IEEE Computer Vision and Pattern Recognition, 2014.

[84] 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.

[85] X. Wang, T. X. Han, S. Yan, “An HOG-LBP Human Detector with Partial Occlusion Handling,” in Proc. of IEEE Int’1 Conf. Computer Vision, 2009.

[86] Y. Tian, P. Luo, X. Wang, X. Tang, “Deep Learning Strong Parts for Pedestrian Detection,” in Proc. of IEEE Int’1 Conf. Computer Vision, 2015.

[87] Vezhnevets A, Ferrari V, Buhmann J M. Weakly supervised semantic segmentation with a multi-image model[C] ICCV. 2011, 1(2): 3.

[88] Wei Y, Liang X, Chen Y, et al. Stc: A simple to complex framework for weakly-supervised semantic segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(11): 2314-2320.

[89] Papandreou G, Chen L C, Murphy K P, et al. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation[C] Proceedings of the IEEE international conference on computer vision. 2015: 1742-1750.

[90] Khoreva A, Benenson R, Hosang J, et al. Simple does it: Weakly supervised instance and semantic segmentation[C] Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 876-885.

[91] Huang Z, Wang X, Wang J, et al. Weakly-supervised semantic segmentation network with deep seeded region growing[C] Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7014-7023.

[92] Zhou Y, Zhu Y, Ye Q, et al. Weakly supervised instance segmentation using class peak response[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3791-3800.

[93] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C] Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[94] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings

of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[95] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[96] Zhang Y, Bai Y, Ding M, et al. W2f: A weakly-supervised to fully-supervised framework for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 928-936.

致 谢

博士三年时间很快就过去了，在此期间，我在学习上和思想上都受益匪浅。这除了自己的努力之外，和各位老师、同学、朋友和家人的关心、鼓励与支持是分不开的。

首先，我要感谢我的导师叶齐祥教授。叶齐祥教授在我攻读博士期间给了我很大的帮助，不管是在学术上的悉心指导还是在生活中的热心照顾，在我遇到困难的时候，总是给与我帮助和鼓励，为我提供科研的思路，并教我如何思考。叶齐祥教授对科研的严谨和对工作的敬业都深深的感染着我，让我明白作为一个优秀者科研工作者应该保持什么样的心态。

其次，我要感谢焦建彬教授在学习和生活中给予我的关心和指导，焦建彬教授为人和蔼，对实验室的每一位同学都很关心，是一位令人敬重的教授。感谢韩振军副教授和秦飞副教授对我的耐心指导，在我的研究生期间给予了很多指导和建议，让我加深了对科研的理解和热情。

感谢实验室的师兄和师姐们，感谢你们的帮助和鼓励。感谢邹佳凌师兄和崔妍婷师姐在我刚进实验室时的时候的指导和帮助，让我迈出了研究生科研的第一步；感谢魏鹏旭师姐一直以来的帮助和指导，在我科研中遇到问题时的热心帮助和指导；感谢高山师兄和柯炜师兄，给了我很多科研的建议。感谢实验室的所有同学，我们在这个大家庭下一起学习和成长，互帮互助。

感谢我的家人，无论我在哪里，无论我处于什么环境，总是给予我无私的关怀和鼓励。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。感谢他们为我所做的一切。

感谢参加开题及中期评阅的各位老师和专家们，他们丰富的经验和无私的工作对论文方向和研究进度的把握和指点给整个研究工作带来了巨大的帮助。

最后，感谢参加论文评审和答辩的各位老师。

万　方

2019 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：

2009/09~2013/06，武汉大学，测控技术与仪器，本科

2013/09~2016/06，中国科学院大学，电子与通信工程，硕士研究生

2016/09~2017/06，中国科学院大学，计算机应用技术，博士研究生

2017/09~2019/07，中国科学院大学，信号与信息处理，博士研究生

获得表彰与奖励：

[1] 博新计划，国家博士后管理委员会资助 60 万元，2019 年

[2] 中国科学院院长奖，候选人，2019 年

[3] 骨架目标检测竞赛 IEEE CVPR 2019，一等奖，2019 年

[4] 高分辨率对地观测感软件解译竞赛，一等奖、二等奖，2017 年

[5] 中国科学院大学，三好学生， 2016 年

已发表（含接受）的学术论文

[1] **Fang Wan**, Pengxu Wei, Zhenjun Han, Jianbin Jiao, Qixiang Ye, “Min-Entropy Latent Model for Weakly Supervised Object Detection,” ***IEEE transactions on pattern analysis and machine intelligence (TPAMI)***, 2019. **(中科院一区国际期刊，影响因子 9.45)**

[2] **Fang Wan**, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, Qixiang Ye, “C-MIL: Continuation Multiple Instance Learning for Weakly Supervised Object Detection,” ***in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR, Oral)***, 2019.（**CCF A 类会议**）

[3] **Fang Wan**, Pengxu Wei, Zhenjun Han, Jianbin Jiao, Qixiang Ye, “Min-Entropy Latent Model for Weakly Supervised Object Detection,” ***in Proc.***

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.（**CCF A 类会议**）

[4] Chang Liu, **Fang Wan**, Yuan Yao, Xiaosong Zhang, Wei Ke, Qixiang Ye, “Orthogonal Decomposition Network for Pixel-wise Binary Classification”, in ***Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***, 2019.（**CCF A 类会议**）

[5] Caijing Miao, Lingxi Xie, **Fang Wan**, Chi Su, Hongye Liu, Jianbin Jiao, Qixiang Ye, “SIXray: A Large-scale Security Inspection X-ray Benchmark for Prohibited Item Discovery in Overlapping Images”, in ***Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***, 2019.（**CCF A 类会议**）

[6] Penxu Wei, Fei Qin, **Fang Wan**, Yi Zhu, Jianbin Jiao and Qixiang Ye,"Correlated Topic Vector for Scene Classification", ***IEEE Transactions on Image Processing (TIP),*** 26 (7): 3221‐3234, 2017. **(中科院二区国际期刊, 影响因子 5.40)**

[7] Wei Ke, Tianliang Zhang, Jie Chen, **Fang Wan**, Qixiang Ye, Zhenjun Han, “Texture Complexity based Redundant Regions Ranking for Object Proposal”, in ***Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop***, 2016.（**CCF A 类会议 Workshop**）

授权国家发明专利：

[1] 焦建彬，崔妍婷，邹佳凌，**万方**，叶齐祥，韩振军，基于集群轨迹分类的集群场景智能监控方法及系统, 201510100197.0, 中国发明型专利，2018 年 5 月接收。