中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

博 士 学 位 论 文

基于视觉注意机制的语义分割自主学习

Autonomic Learning of Semantic Segmentation Based on Visual Attention Mechanisms

论文作者 侯淇彬 指导教师 程明明 教授

申请学位 工学博士 培养单位 计算机学院

学科专业 计算机科学与技术 研究方向 计算机视觉

答辩委员会主席 胡清华 教授 评 阅 人 匿名评阅人

南开大学研究生院

二〇一九年五月

南开大学学位论文使用授权书

本人完全了解《南开大学关于研究生学位论文收藏和利用管理办法》关于南开大学 (简称“学校”) 研究生学位论文收藏和利用的管理规定，同意向南开大学提交本人的学位论文电子版及相应的纸质本。

本人了解南开大学拥有在《中华人民共和国著作权法》规定范围内的学位论文使用权，同意在以下几方面向学校授权。即：

1. 学校将学位论文编入《南开大学博硕士学位论文全文数据库》，并作为资料在学校图书馆等场所提供阅览，在校园网上提供论文目录检索、文摘及前 16 页的浏览等信息服务；

2. 学校可以采用影印、缩印或其他复制手段保存学位论文；学校根据规定向教育部指定的收藏和存档单位提交学位论文；

3. 非公开学位论文在解密后的使用权同公开论文。

本人承诺：本人的学位论文是在南开大学学习期间创作完成的作品，并已通过论文答辩；提交的学位论文电子版与纸质本论文的内容一致，如因不同造成不良后果由本人自负。

本人签署本授权书一份（此授权书为论文中一页），交图书馆留存。

学位论文作者暨授权人 (亲笔) 签字：_____

20　　年　　月　　日

南开大学研究生学位论文作者信息

论 文 题 目	基于视觉注意机制的语义分割自主学习				
姓　　名	侯淇彬	学号	1120160128	答辩日期	2019 年 5 月 27 日
论 文 类 别	博士 ☑　学历硕士 ☐　专业学位硕士 ☐　同等学力硕士 ☐　划☑选择				
学院（单位）	计算机学院		学科/专业 (专业学位) 名称		计算机科学与技术
联 系 电 话	13052259892		电子邮箱	houqibin@mail.nankai.edu.cn	
通讯地址 (邮编)：300350					
非公开论文编号			备注		

注：本授权书适用我校授予的所有博士、硕士的学位论文。如已批准为非公开学位论文，须向图书馆提供批准通过的《南开大学研究生申请非公开学位论文审批表》复印件和“非公开学位论文标注说明”页原件。

南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行研究工作所取得的研究成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本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。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：_____ 年 月 日

非公开学位论文标注说明

(本页表中填写内容须打印)

根据南开大学有关规定，非公开学位论文须经指导教师同意、作者本人申请和相关部门批准方能标注。未经批准的均为公开学位论文，公开学位论文本说明为空白。

论文题目			
申请密级	☐ 限制 (⩽2 年) ☐ 秘密 (⩽10 年) ☐ 机密 (⩽20 年)		
保密期限	20 年 月 日至 20 年 月 日		
审批表编号		**批准日期**	20 年 月 日

南开大学学位评定委员会办公室盖章 (有效)

注：限制 ★2 年 (可少于 2 年); 秘密 ★10 年 (可少于 10 年); 机密 ★20 年 (可少于 20 年)

摘要

近年来，作为计算机视觉领域中的热门研究方向之一，语义分割已经取得了长足进展，尤其在深度卷积神经网络的出现之后。另外，随着含有成千上万张图像的大规模数据集的出现，基于深度学习的语义分割方法的分割精度也在不断提升。然而，基于深度卷积神经网络的全监督语义分割模型严重依赖于大量人工标注数据，因而在学习新的类别信息时仍然需要庞大的人力成本来标注新数据。弱监督语义分割技术，由于仅依赖图像类别标签等轻量级标注数据，也因此正在成为一大学术研究热点。

早期的弱监督语义分割模型主要利用注意力模型等工具来生成带有语义信息的种子区域进而训练语义分割网络。此外，显著性物体分割模型，由于其检测出图像前景区域的能力，也逐渐被大多数算法所采用。将检测到的显著性区域与注意力模型生成的类别激活图像相结合可以给类别无关的显著性区域加上类别标签进而用来训练语义分割模型。以上讨论都基于一个假设方案，即所有图片的类别标签都是已知的且准确的。

考虑到人工标注的成本，当给定一个类别标签集合，如何利用免费的网络大数据来索引相关图像进而学习语义分割模型具有重大研究意义。针对上述难题，本文将从两种不同的视觉注意机制（显著性物体检测与注意力区域检测）出发，提出一种有效的方案来解决这一更为通用的弱监督语义分割问题。与此同时，针对现有显著性物体检测与注意力区域检测模型的不足之处，本文也将提出改进方案。本文的具体贡献如下：

1. 提出一种自顶向下的基于短连接结构的显著性物体检测模型。通过在已有的分类网络的不同阶段后接入侧向路径并在不同侧向路径之间引入一系列自顶向下的短连接，可以使得卷积神经网络高层含有的高级语义特征被传递到低层侧向路径，同时低层特征含有的边缘信息可以进一步丰富高层语义特征模糊的边缘信息。5 个被广泛使用的数据集上的实验结果表明该方法已明显优于现有方法。
2. 提出了一种基于自擦除策略的注意力模型。该方法在现有的对抗擦除策略的基础上引入了背景先验知识，并设计两种不同的自擦除策略，可以有效

地解决基于对抗擦除策略的模型在训练过程中难以控制可辨别区域不断扩散的弊端。该方法生成的类别激活图像不但具有较高的质量，并且在弱监督语义分割任务中可以取得较高的分割精度。

3. 提出了如何智能地从互联网资源中挖掘有用知识来自主地学习语义分割模型。为了解决互联网数据中大量的类别噪声以及复杂的背景，该方法提出了噪声擦除网络的概念。通过从可辨别区域中学习语义知识，可以对显著性物体检测模型提取的前景物体的类别进行推断并擦除其中与检索关键词不相关的区域。

关键词： 语义分割；自主学习；显著性物体检测；注意力模型；卷积神经网络

Abstract

In recent years, as one of the hot research topics in computer vision, semantic segmentation has made great progress, especially after the emerging of convolutional neural networks (CNNs). Because of the large-scale datasets with tens of thousands of training images, the precision of CNN-based segmentation models rises. However, fully-supervised semantic segmentation networks heavily rely on tremendous human-labeled annotations and hence needs significant human labors when dealing with new categories. Weakly-supervised semantic segmentation, due to its dependence on less annotation data, gradually becomes another hot research topic.

Early weakly-supervised semantic segmentation methods mostly leverage tools like attention models to generate seed areas to train segmentation models. In addition, salient object detection models, due to their ability to capture foreground regions, have been being adopted by many approaches. By appropriately combining attention maps with saliency maps, these approaches can allocate each pixel in the class-agnostic saliency maps a class-specific tag, which makes training segmentation models possible. The aforementioned description is based on an assumption that the image-level labels associated with each image are precise.

In fact, given a collection of category keywords, how to extract useful information from the free web images to train segmentation models is of great interests. Taking the above challenge into account, this dissertation aims at presenting an effective way to tackle this general weakly-supervised semantic segmentation problem by leveraging two different visual attention mechanisms (salient object detection and attentive region detection). Beyond that, regarding the drawbacks of existing saliency and attention models, this dissertation also introduces useful thoughts to advance them. The contributions of this dissertation can be summarized as follows:

1. Presenting a salient object detection model based on top-down short connections. By connecting side paths to the stages of classification networks and building short connections between each pair of side paths, high-level semantic features

can be delivered to lower side paths and low-level features with rich edge information can help refine coarse high-level features. Experiments on 5 widely-used benchmarks show that the proposed approach improves all existing methods.

2. Presenting a self-erasing strategy for attention models. Based on the adversarial erasing strategy, the proposed approach takes the location of background as priors and designs two different self-erasing strategies, which are able to well solve the problem of the unstoppable expansion of the discriminative regions to the background for models based on the adversarial erasing strategy. Experiments show that the resulting attention maps are with high quality and lead to high segmentation precision when applied to the weakly-supervised semantic segmentation task.
3. Presenting how to intelligently learn semantic segmentation by extracting useful knowledge from the Internet. To deal with the large amount of label noise presenting in web images and their complex background, the concept of noise erasing network (NENet) is proposed. By learning semantic knowledge from discriminative regions by attention models, NENet is able to assign each foreground regions extracted by saliency models a prediction label and erase regions that are unrelated to the query keyword.

Key Words: Semantic segmentation; automatic learning; salient object detection; attention model; convolutional neural network

目录

第一章 绪论

第一节 研究背景和意义

计算机视觉是一门研究如何智能地利用工具来模仿人的视觉系统的科学。我们知道人类的视觉系统可以快速并且准确地理解复杂的场景信息并将提取的视觉信息传递给大脑进行分析。如何模拟人类的视觉系统在现实生活中起着重要作用，并且具有很多应用场景，比如：自动驾驶、智能监控、家用机器人、智慧交通等。

与人类的视觉系统类似，在计算机视觉领域中，语义分割是必不可少的一部分。语义分割将传感器采集的视觉数据进行特征提取并进行分析，从而得到感兴趣类别的分割结果。图1.1给出了语义分割期望得到的分割结果。近年来，作为计算机视觉领域中的基础研究方向，语义分割一直具有着较高的关注度以及较多的应用场景。

随着计算机存储速度以及计算能力的提升，大数据驱动的语义分割模型逐渐成为了目前研究的主流方向。虽然这类模型的分割精度在不断地提升，在理解新的类别信息时仍需要大量的标注数据来训练模型。为了解决现有语义分割模型对于大量高精度标注数据的依赖问题，越来越多的国内外学者开始考虑使用较少量的标注信息来训练语义分割模型。具体来说，现有的语义分割模型可以大致分为三类：

- 基于全监督的方法。在这类方法中，每个类别的物体对应着数百乃至上千张训练图像并且每张图像都拥有像素级别精确的标注数据。
- 基于弱监督的方法。在这类方法中，每个类别的物体也对应着数百乃至上千张训练图像，但与基于全监督的方法不同的是，每张图像对应的标注数据信息量较少，通常为物体的位置信息（物体的外矩形框）或者仅为图像中所包含的类别信息。
- 基于半监督的方法。在这类方法中，除了提供基于弱监督的方法中的少量标注外还包括少量高精度的像素级别的标注数据。其目的是借助少量的高精度的标注数据来改善基于弱监督模型的分割精度。

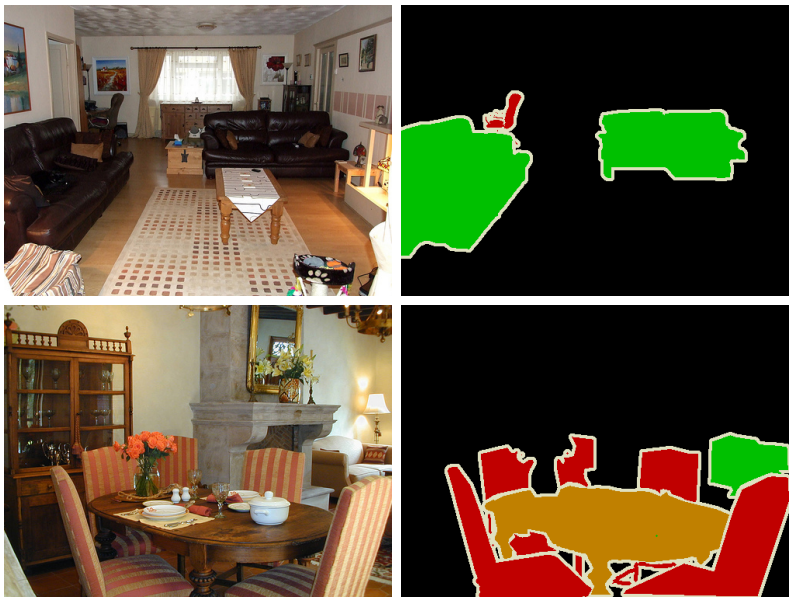


图 1.1 选自 PASCAL VOC 2012 数据集[1] 中的样例图像以及对应的标注结果。右侧图中，彩色部分为感兴趣的语义区域。语义分割的目的在于将给定图像中感兴趣的区域分割出来并分配给其中每一个像素一个标签用来标记该像素的类别信息。

与基于全监督的方法相比，基于弱监督的方法大大减少了对于大量标注数据的依赖，但其仍需要少量的人工标注。

人类在认知的过程中能够快速地学习到场景中的知识并且能够快速地将物体进行分割并完成分类任务。然而，现已有的相关分割方法不仅依赖于大量的图像数据同时需要相应人工标注的数据，使得语义分割模型在学习新的类别时变得十分困难。因而，在给定待学习的类别集合的同时，如何使模型自主地从免费数据源学习到语义信息正逐步成为新的研究热点。庞大的互联网数据提供了免费的样本用于训练，这也使得利用网络图像及其检索关键词生成用于训练语义分割模型的伪标注数据具有重大意义。该思路也将促成基于自主学习的语义分割模型的诞生。

目前为止，基于弱监督的语义分割算法的主要研究内容为如何生成高质量的伪标注数据，也即如何利用已有的弱监督信息生成用于训练语义分割网络的伪标注数据。因而，现有的基于弱监督的语义分割算法的主要流程如图1.2所示。首先，采用类别无关的特征提取器（例如显著性物体检测模型、边缘检测模型等）以及注意力模型对输入图像进行特征提取；然后将得到的特征图进行处理得到伪标注数据；最后将伪标注数据与输入图像送入语义分割网络中得到最终

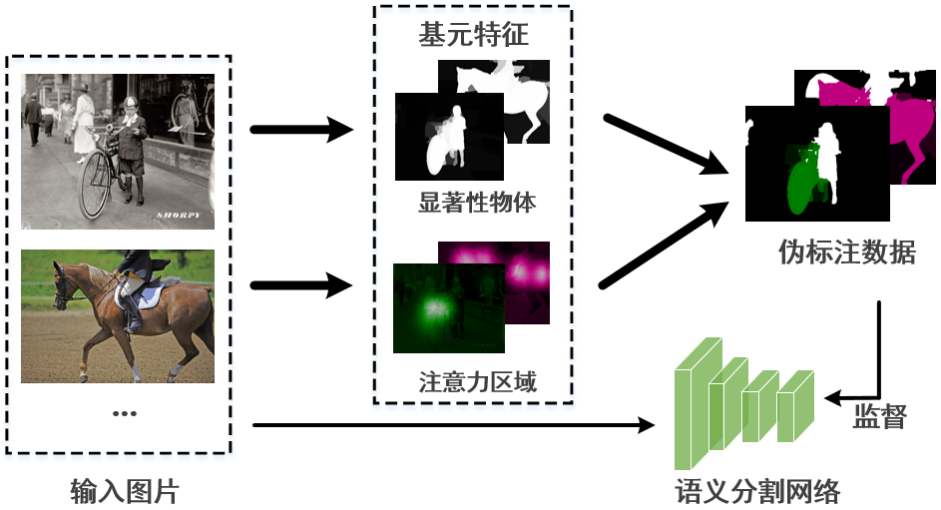


图 1.2 现有主流弱监督语义分割模型框架示意图。首先采用类别无关的特征提取器（例如显著性物体检测模型、边缘检测模型等）或类别相关的注意力模型对输入图像进行特征提取。然后将得到的特征转换为伪标注数据。最后将伪标注数据与输入图像送入语义分割网络中进行训练并得到最终的分割结果。

的分割结果。

本文将从两种不同的视觉注意机制（包括显著性物体检测以及图像注意力机制）入手，主要研究内容为如何利用两种不同的视觉注意机制实现语义分割的自主学习（也即在仅给定类别标签的情况下自主地从免费的数据源学习语义分割）。在以上方案中，显著性物体检测模型以及注意力模型的性能将对最终生成的语义分割模型有重大影响。因此，如何从获取的免费图像训练数据中提取高质量的显著性图以及注意力区域图从而生成高精度的伪标注数据是本文的研究重点。

第二节 研究难点

本文的主要研究内容为如何保证显著性物体检测模型以及注意力模型的高性能（显著性图与类别激活图像的生成）以及如何利用以上两种视觉注意机制实现弱监督语义分割任务乃至语义分割自主学习任务。本文中具体的研究难点如下：

- 显著性物体检测模型可以将复杂度相对较低的图像中的前景物体分割出来。基于卷积神经网络的显著性物体检测算法与经典的方法相比具有一定优势。然而由于卷积神经网络的金字塔结构[2]，深层特征通常含有丰富的

图 1.3 卷积神经网络不同层级特征生成的显著性图。上半部分分别为原始图像与其对应的标注图，下半部分分别为浅层特征以及深层特征生成的显著性图。

语义信息但其尺寸通常较小使得生成的显著性图较为模糊且缺少边缘信息[3]。与此相反，浅层特征尺寸较大但仅含有低层级特征。图1.3 给出了不同层级特征生成显著性图的示意图。因此，如何有效地将深层特征与浅层特征相结合是一大研究难点。

- 注意力模型[4–6] 可以在给定图像及其类别标签的情况下生成每个类别对应的可辨别区域（即类别标签对应的语义物体）。现有的注意力模型虽然可以较为准确地定位到给定图中某个类别标签对应的物体，但其得到的可辨别区域通常具有不规则的形状且仅能覆盖到部分语义物体。因而，如何生成含有较强边缘信息且能较为完整地覆盖语义物体的可辨别区域为一大研究难点。
- 对于语义分割的自主学习而言，与传统基于弱监督[7, 8] 的语义分割相比具有更高难度。如图1.4所示，弱监督语义分割需要在仅给定图像的类别标签时对语义相关的物体完成分割且其分割结果的精度需要是像素级的（如图1.1右侧所示）。然而语义分割的自主学习需要在仅给定关键词的情况下完成像素级精度的分割。因而，如何获取高质量的训练图像以及如何在仅有类别关键词的情况下生成高质量的伪标注数据来训练语义分割网络皆为该方向的研究难点。

火车　　人、狗

图 1.4　选自 PASCAL VOC 2012 数据集中的样例图像以及对应的类别标签数据。

第三节　常用数据集介绍

本节将针对显著性物体检测与弱监督语义分割任务中常用的数据集进行简要介绍。

1.3.1　显著性物体检测常用数据集

自 2000 年以来，显著性物体检测研究方向中出现了十数个数据集，其中较为常用的数据集包含 MSRA-B[9]、ECSSD[10]、HKU-IS[11]、PASCALS[12] 以及 SOD[13, 14]。图1.5 给出了从以上数据集中选取的图像样例以及对应的标注图像。下面将对以上数据集进行简要介绍：

1. MSRA-B 数据集包含 5,000 张图像。其显著性物体来自于上百个不同的类别。由于数据的多样性，该数据集也被认为是显著性物体检测领域最被广泛使用的数据集之一。该数据集中多数图像含有一个显著性物体，因此其也逐渐被看作评测显著性物体检测模型处理简单场景能力的数据集。
2. ECSSD 包含了 1,000 张语义上有意义但场景相对复杂的自然图像。该数据集中图像全部为测试图像。
3. HKU-IS 为另一包含 4000 多张训练以及测试图像的大规模数据集。该数据集中多数图像的显著性物体数量多于一个且图像中前景与背景之间有较低的对比度。
4. PASCALS 包含了 850 张包含多个显著性物体的测试图像。该数据集中所有图像均取自于标准的 PASCAL VOC 2010 分割数据集，因而含有较为复杂的背景。
5. 最后一个测试数据集为 SOD。该数据集总共包含 300 张图像且均选自 BSDS 数据集。

图 1.5　选自不同显著性物体检测数据集中较难的样例以及标注数据。

1.3.2　弱监督语义分割常用数据集

现有的基于弱监督的语义分割算法大多采用 PASCAL VOC 2012 分割数据集进行模型训练与测试。该数据总共含有 4369 张生活场景图像，其中包含 1464 张训练集图像、1449 张验证集图像以及 1456 张测试集图像。训练集以及验证集图像皆有精准的像素级标注而测试集仅有训练图像被公开，因而实验人员需要提交生成的分割结果图到服务器端进行评测。为了提升训练的效果，Hariharan 等[15] 将该数据集的训练集图像数量提升到 10582 张。VOC 2012 数据集总共包含了 20 个前景类以及 1 个背景类。前景类的名称具体如下：“喷气式飞机”、“自行车”、“鸟”、“船”、“杯子”、“公共汽车”、“汽车”、“猫”、“椅子”、“奶牛”、“餐桌”、“狗”、“马”、“摩托车”、“人”、“盆装植物”、“绵羊”、 “沙发”、“火车”以及 “显示器”。由于该数据集中图像皆来自生活中的场景图，因而大多数图像含有 2-5 个不等的语义类别且同种物体的尺寸大小差异较大。图1.6给出了一些较难的样例图像及其对应的标注数据。

第四节　研究目标与主要贡献

近年来，弱监督语义分割算法的精度已经逐步逼近于全监督模型的精度，但其主要缺陷在于仍需要利用少量的人工标注信息。为了解决语义分割算法对于人工标注的依赖，本文提出语义分割自主学习的概念。在给定待分割的关键词集合，自主学习算法可以自动从互联网中检索关键词相关图像并利用多种视觉注意机制（如显著性物体检测模型以及注意力模型等）自动提取图像中与关键词相关的语义区域并将其作为伪标注数据来训练语义分割网络。本文主要工作之

图 1.6 VOC 2012 数据集中较难的样例以及标注数据。

间的关系如图1.7所示。下面将针对本文的主要研究内容以及贡献进行简要介绍。

- 第二章对与本文工作相关的文献进行介绍，包括显著性物体检测算法、基于神经网络的注意力模型以及基于弱监督的语义分割算法。
- 第三章提出了基于短连接的显著性物体检测模型。该模型将卷积神经网络的低层级特征与高层级特征有效地结合在一起。现有的模型在利用卷积神经网络的多层级特征时仅将不同层级的特征简单地串接或加在一起。本章提出的基于短连接的模型将所有高层级特征以短连接的方式稠密地连接到低层级特征中，其目的在于用含有丰富的语义信息但较小尺寸的高层级特征来指导低层级特征寻找显著性物体的位置。同时，低层级特征丰富的细节特征可以帮助细化粗糙的高层级特征从而使得检测出的显著性物体具有较好的边缘信息。为了使每个层级融合后的特征皆能检测出质量较高的显著性物体，本模型采用了深度监督的策略并最终将所有层级中显著图像线性叠加在一起得到最终的显著图像。由于短连接的使用，本模型在各个显著性物体检测数据集上皆优于现有的基于深度学习以及非深度学习的方法。
- 第四章提出了基于自擦除策略的注意力模型。该模型在对抗擦除策略的基础上引入了自擦除的概念。基于对抗擦除模型在训练的过程中很难把控何时停止，因而容易使激活区域扩散到背景区域，大大降低了生成的类别激

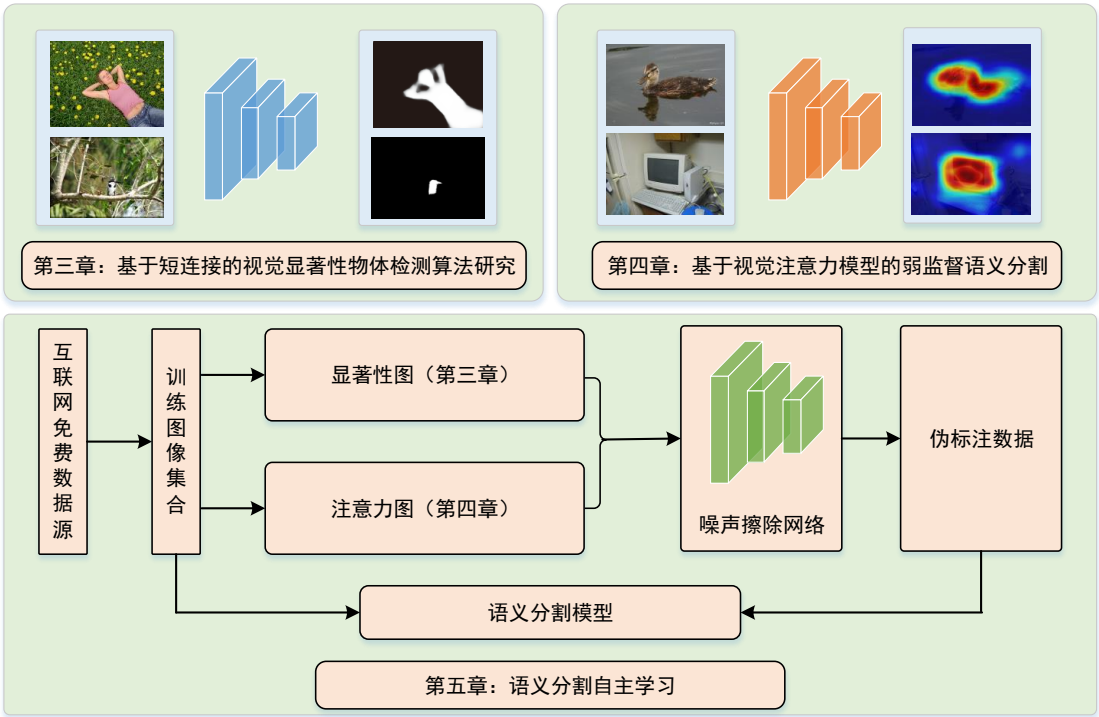


图 1.7 本文主要工作之间相互关系的示意图。

活图像的质量。自擦除概念的引进可以有效利用图像背景这一先验知识。通过将背景区域内的激活值的符号进行反转，可以达到抑制激活区域扩散的目的。为了测试本模型生成类别激活图像的质量，本章将生成的类别激活图像应用到弱监督语义分割任务中。通过将类别激活图像与显著图像以调和均值的方式进行融合可以生成用于训练语义分割模型的伪标注数据。在 VOC 2012 数据集上的实验结果表明基于自擦除策略的注意力模型明显优于现有方法。

- 第五章提出了语义分割自主学习的概念。由于从互联网中检索的图像除了含有检索关键词相关的语义物体外也大概率包含其它类别相关的语义物体，因而如何防止生成的伪标注受到类别噪声的干扰至关重要。为了解决这一问题，本章提出噪声擦除网络。噪声擦除网络首先以显著图像以及注意力模型生成的类别激活图像为输入并从类别激活图像中的种子区域学习语义知识，然后对网络图像中的前景物体中的分割块[16, 17] 进行预判并从训练样本中擦除掉预判标签与关键词不一致的分割块。基于 VOC 2012 分割数据集的结果表明本方法生成的分割图明显优于大部分弱监督语义分割方法。当将 VOC 2012 训练图像及其类别标签用于训练时，本方法可以在

目前所有弱监督语义分割方法中取得最佳效果。

- 第六章对本文所有工作进行总结，提出本文工作中可改进的部分，并对未来工作进行展望。

第二章 相关工作综述

本章主要针对本文所研究内容的相关工作进行介绍。第一节主要介绍显著性物体检测相关模型；第二节对图像注意力机制相关文献进行介绍；第三节主要针对现有的弱监督语义分割算法进行概述；第四节对本章内容进行总结。

第一节 显著性物体检测

显著性物体检测旨在将给定图像中最为显著（或用户感兴趣的区域）分割出来[18]。显著性物体检测方法可以大致分为经典方法（基于非深度学习的方法）以及基于深度学习的方法。而经典方法又可以大致分为局部方法以及全局方法两类。下面将针对经典的显著性物体检测算法以及基于深度学习的显著性物体检测算法分别进行介绍。

2.1.1 经典显著性物体检测方法

2.1.1.1 基于局部对比的方法

基于局部对比的方法通常利用图像区域相对于其周边局部邻域的对比度计算得到最终的显著性区域。其中，比较有代表性的工作为 Itti 等[19] 在 2001 年提出的利用图像在不同多尺度下计算不同位置与中心的差异来得到显著性图。与 Itti 等不同的是，Liu 等[9] 提出在不同尺度下通过计算高斯图像金字塔不同尺度图像之间的对比度并将其线性叠加，得到最终的显著图像。Harel 等[20] 提出将 Itti 等工作提取的显著图像进行归一化操作来进一步突显出显著区域的物体。该方法的另一个优点是可以简单地与其它方法的显著图像相结合从而得到质量更高的显著图像。Goferman 等[21] 提出了一种上下文已知的显著图像生成算法。该算法同时考虑到了局部信息、全局信息、视觉组织规则以及表层特征并将这些所有上下文信息进行建模来找到显著物体。虽然上述方法可以找到显著性物体的大概位置，但由于它们多数仅利用了局部区域内的不同特征之间的对比信息使得最终输出的显著图像通常仅在显著物体边缘处具有较大的显著性值而非整个显著物体。这一弱点使得它们很难被直接应用到其它依赖显著性物体检测的方法[22–24] 中。

2.1.1.2 基于全局对比的方法

在基于全局对比的方法中，较为典型的是 Cheng 等[25] 在 2015 年提出的 GC 算法。该算法首先采用过分割算法[26] 对图像进行分割，然后对每个分割块提取其颜色直方图。通过计算每个分割块的直方图与其它所有分割块直方图的对比度来判断当前分割块是否显著。虽然该方法与上述基于局部对比的方法相比可以检测到整个显著性物体，但在计算不同分割块直方图的对比时所有的权重都是常数值，这一弱点使得该方法在一些复杂的场景中并不适用。Jiang 等[27] 针对 GC 算法的缺陷提出了采用机器学习的方式学习每种特征在计算不同分割块对比度时的系数。另外，该方法对每个分割块提取了更多的特征来解决场景较为复杂的情况。基于全局对比的方法与基于局部对比的方法相比有了一定程度上的提升，但由于多数特征都依赖于研究人员手工设计使得这些算法仅在一些简单的场景中较为适用。

2.1.2 基于卷积神经网络的显著性物体检测模型

由于大规模数据集的出现，基于卷积神经网络以及机器学习的模型已逐渐成为显著性物体检测的主流方向。与传统的基于人工设计的特征的显著性物体检测方法相比，基于卷积神经网络的显著性物体检测模型由于其大量的可学习参数以及可以提取局部与全局上下文信息的能力已逐步成为该研究领域的主流方向。下面将主要介绍近些年来基于卷积神经网络以及机器学习的方法。与此同时，本节也将回顾关于短（跳跃）连接的神经网络架构。

2.1.2.1 早期方法

早期的基于深度神经网络的方法大多借助过分割算法[16, 26, 28, 29] 以及多层感知机来估计每个过分割区域的显著性值。He 等[30] 提出了一种基于多层级对比特征的超像素级别的卷积神经网络架构。该方法每次从所有尺度的超像素中选取两个作为对比区域并送入一维卷积神经网络中来提取更高级特征。最后该方法用学到的权重对不同尺度的超像素的显著性值做加权处理得到最终的显著性图。Li 等[11] 也利用从深度卷积神经网络中提取的多尺度特征来预测图像的显著性图。不同尺度的过分割图被分别送到不同的卷积神经网络中，然后所有的输出被送到多层感知机中来判断每个超像素是否显著。Wang 等[31] 利用了两个卷积神经网络并通过引入局部估计以及全局搜索的概念来解决显著性物体检测问题。第一个卷积神经网络被用来学习局部特征并为每一个像素估计一个显著

性值，然后得到的局部显著性图、全局对比信息以及几何特征等信息被送到第二个卷积神经网络中作为输入来判断每一个区域的显著性值。Zhao 等[32] 提出了一种基于多个上下文信息的深度学习架构来检测显著性物体。两个不同卷积神经网络分别被用来学习每个过分割区域的全局以及局部的上下文信息。得到的全局与局部特征被送到一个多层感知机中来估计每个过分割区域的显著性值。Lee 等[33] 结合了从深度神经网络中提取的高级语义特征以及人工设计的低级特征。为了更好地将二者相融合，其采用了多层感知机来对提取的低级与高级特征进行预判并输出每个过分割区域的显著性值。

2.1.2.2 全卷积神经网络

全卷积神经网络首次被 Long 等人提出[34]，其主要思想是借助于二维卷积来实现端到端的视觉任务学习，如语义分割。Liu 等[35] 提出了一种两阶段的全卷积神经网络。在第一阶段，原始图像被送入卷积神经网络中来预测一个形状较为粗糙的显著性图。在第二阶段中，上述得到的粗糙的显著性图以及从第一阶段提取的一些边缘特征被送入另一个全卷积神经网络中从而得到一个清晰度较高的显著性图。Li 等[36] 引入了两个分支来分别预测像素级的显著性分割图以及超像素级的显著性分割图。第一个分支将 VGGNet[37] 不同阶段的特征进行融合。第二个分支则以超像素为单位预测每一个超像素的显著性值。两个阶段的显著性图加权后的结果被送入一个全连接的条件随机场（Conditional Random Fields）[38] 中来进一步提高显著性图的空间相关性。Wang 等[39] 首次提出了利用循环神经网络的概念来解决显著性物体检测。循环神经网络的多次内部迭代可以使得每次得到的显著性图被持续改善并修正错误的预判，最终输出一个质量较高的显著性图。

2.1.3 跳跃连接结构

近年来，由于卷积神经网络比较灵活的架构，其在显著性分割以及语义分割领域已经取得了突破性进展。在现有的多样的架构中，基于跳跃连接的结构已被广泛采用。其原因在于这种结构能够有效地融合多尺度以及多层级的特征。早期的跳跃连接结构，例如 Hypercolumn[40] 以及 DCL[36]，在各自研究领域已经取得了一定进展。然而，这些方法仅仅将从卷积神经网络不同阶段提取的多尺度特征进行了简单的融合。与上述结构不同的是，基于 FCN[34] 的方法提出了一种自上而下的方法来利用从主干网络中提取的多层级特征。各阶段的特征并非

被简单地融合，而是采用了一种自顶向下的融合策略，使得来自不同层级的特征被逐步融合并得到较高分辨率的显著图像。Xie 和 Tu[41] 提出了一种基于深度监督的网络架构。在网络各阶段提取出不同层级特征之后，该架构在每个阶段之后分别添加一个侧向输出并加以监督。其主要目的在于让每个阶段提取的特征尽可能保留更多细节信息并起到解决 VGGNet 等神经网络因为引入较深的结构而产生的梯度消失问题。

尽管多层级以及多尺度特征已被广泛应用于各种视觉任务中，由于卷积神经网络的灵活性，基于跳跃连接的结构仍有很大改进空间。本文第三章将提出更为高效的利用多尺度与多层级特征的架构并将其应用到显著性物体检测任务中。

第二节 注意力模型

计算机视觉中的注意力模型，在仅给定图像类别标签的情况下，可以较为精准地找到语义物体的大体位置或者说可以找到帮助模型做预判的激活区域。作为理解卷积神经网络的一种工具，注意力模型已被广泛应用到多个计算机视觉领域，包括物体位置定位、弱监督语义分割等。本小节将对现有的注意力模型进行简要介绍。主要内容包括：基于卷积神经网络的注意力模型基础、基于对抗擦除的注意力模型以及基于空洞卷积的注意力模型。

2.2.1 基于卷积神经网络的注意力模型基础

对于早期的注意力模型[42, 43] 而言，其采用的方法主要是通过误差回传的方式来可视化卷积神经网络所学到的知识。后来，MIT 的 Zhou 等提出了 Class Activation Mapping（CAM）模型[4]。该模型以大规模分类网络 VGGNet[37] 为主干网络。与原始 VGGNet 不同的是，CAM 将 VGGNet 最后一个池化层以及所有全连接层换成一个 1024 通道的卷积层加上一个全连接层。该全连接层可根据分类时的类别数而定。通过建立最后一个卷积层每个通道与全连接层权重之间的关系，可以得到每个类别对应的类别激活图像。图2.1列出了几张图像对应的来自 CAM 模型[4] 的类别激活图像。可以看出，类别激活图像不仅可以较为精准地定位到语义物体的具体位置并且可以提供语义物体大致的轮廓信息。当应用于物体定位时，可以根据激活值的大小采用一固定阈值来得到一个包含语义物体的矩形框。当应用于弱监督语义分割任务时，可辨别区域可以作为语义物体的种子区域进而用来训练语义分割模型。

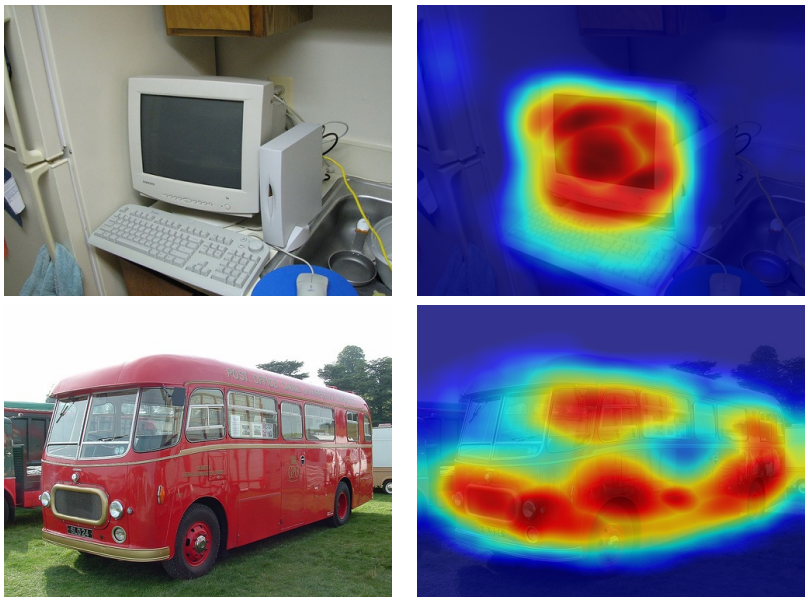


图 2.1 选自 PASCAL VOC 2012 数据集中的样例图像以及对应的类别激活图。其中，第一张图像对应的标签为“显示器”，第二张图像对应的标签为“公共汽车”。类别激活图像由蓝到红代表着激活值强度的增加。可以看出，类别激活图像可以较为精准地定位到语义物体的位置并给出物体大致的轮廓信息。

基于 CAM 的思想，Selvaraju 等提出了 Grad-CAM[44] 模型。该模型的可视化功能使其可以成功地应用到图像分类、场景问答[45–49] 等任务中。与上述方法不同的是，Zhang 等[6] 提出了一种新的基于梯度回传的方法来可视化可辨别区域。受人的视觉系统启发，该方法采用了一种“赢者通吃”（winner-take-all）的概念来将深层中激活的神经元逐步回传到低层中。以上方法的共性在于类别激活图像仅由模型迭代一次生成，因而得到的类别激活图像并不能够将语义区域完整地显示出来。

2.2.2 基于对抗擦除的注意力模型

虽然 CAM 可以提供语义物体较为精准的定位，但其生成的类别激活图像并不能够突显出语义物体边缘的位置，也即生成的可辨别区域通常为语义物体的大致轮廓。另外，当语义物体的形状较为庞大时，仅部分语义物体能够被挖掘出来而非语义物体整体。其主要原因在于注意力模型在训练过程中仅通过语义物体的部分区域即可判断出该物体是否存在于输入图像中。例如，在检测图像中是否含有“狗”时，狗的头部通常作为敏感区域被挖掘出来而其身体部分通常被忽略掉。

为了解决上述难题，Wei 等[8] 提出了一种基于对抗擦除策略的注意力模型。其基本思想在于将已挖掘出的可辨别区域从原始输入图像中擦除掉，然后将剩余的图像继续送入注意力模型中并用相同的类别标签进行训练直到模型无法辨别出图像的类别。这样做的目的在于告诉注意力模型输入图像中未被擦除的区域仍含有类别标签对应的语义物体。在这种激励下，注意力模型将继续从未被擦除的区域中寻找属于对应类别标签的区域。以这种方式，通过不停地迭代模型可将较大物体的整体区域挖掘出来。

由于该种方案在训练过程中需要多次训练模型，使得其训练过程较为繁琐。为了简化该方案，Li 等[50] 提出了一种可以实现端到端的对抗擦除策略，如图2.2（a）所示。在此基础上，Zhang 等[51] 进一步改进了上述擦除过程。其示意图可参见图2.2（b）。该方案并非对原始图像进行擦除操作而是对来自主干网络的特征映射图进行擦除。因而，在训练过程中，网络的每一部分仅需执行一遍即可完成操作。具体而言，首先第一阶段生成器采用了与 CAM 类似的结构并生成类别标签对应的类别激活图像并将得到的类别激活图像进行二值化处理得到可辨别度最高的语义区域。然后，将来自主干网络特征映射中对应的区域擦除掉并将其送入第二阶段生成器中。该阶段采用与第一阶段相同的训练损失继续挖掘第一阶段并未被挖掘到的语义区域。在测试阶段，将两个分支的类别激活图像进行合并得到最终的类别激活图像。

虽然基于对抗擦除的注意力模型可以检测到较为完整的物体区域，但其一大缺陷在于：随着模型的不断迭代，越来越多的语义区域将被挖掘出来，使得许多背景区域亦被判断为语义区域。本文将在第四章中提出一种基于自擦除策略的注意力模型来缓解上述缺陷。

2.2.3 基于空洞卷积的注意力模型

除了对抗擦除模型外，另一种改进 CAM 的策略为基于空洞卷积（Dilated Convolution）的方案[5]。与普通卷积不同之处在于空洞卷积可以通过设置空洞的大小来检测远处点与当前待处理点的关系。通过在 CAM 模型的基础上加入多个包含不同空洞大小的卷积层分支，可以使模型智能地判断图像中与语义物体相关的部分。

虽然引入空洞卷积的概念对检测语义物体的整体有所帮助，但由于语义物体的大小是不可控因素，使得空洞大小的选择较为困难。另外，空洞卷积的卷积核为非可形变的，使其难以完全区分语义物体与背景。因而，基于空洞卷积

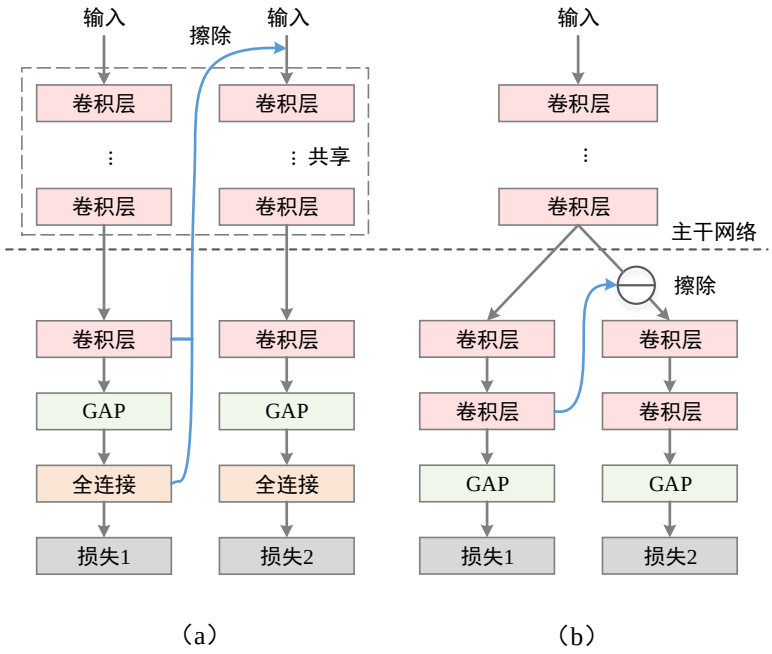


图 2.2 两种典型的基于对抗擦除策略的注意力模型。（a）在输入图片上进行擦除操作。（b）在神经网络中间特征图上进行擦除操作。其中“GAP”表示全局均值池化层。

的模型得到的可辨别区域通常含有背景，从而引入较多的杂质区域。

第三节 弱监督语义分割

基于全监督的语义分割技术[34, 52–59]已经取得了重大进展，但基于网络数据的语义分割自主学习技术尚在起步阶段。由于语义分割自主学习亦可作为基于弱监督的语义分割领域的研究方向之一，因而本小节将主要针对现有的基于弱监督的语义分割方法进行简要介绍。

2.3.1 早期工作

早期的工作[60, 61]主要是利用期望最大化算法[62]或是多实例学习等方法来解决弱监督语义分割任务。Papandreou 等[60]设计了一种在线的期望最大化（EM）算法来训练语义分割模型。根据多次 EM 迭代，该方法对每一个像素值的类别进行预测。Pinheiro 等[61]采用多实例学习方法来实现分割。多实例学习（Multiple Instance Learning，MIL[63]）属于机器学习的一个分支。在多实例学习中，多个实例组成了一个“包”。当一个“包”内有一个正样本时，则该“包”的标签为正。相反，当一个“包”内所有样本标签皆为负时，该“包”的样本为

负。Pinheiro 等[61] 将多实例学习与神经网络相结合。在该方法中，神经网络被用来提取高层级特征以及进行初步像素级分类，然后 MIL 将“包”内（超像素）的样本聚合为图像级标签。该方法将输入图像送入到卷积神经网络中并将其看作多个二进制分类任务，然后根据最后一层特征映射的激活值大小来选取一些种子区域。最后，通过超像素分割算法以及物体外边框等先验对得到的种子区域进行进一步平滑得到最终的分割结果。Pathak 等[64] 提出了一种约束神经网络来解决弱监督语义分割任务。该方法在优化卷积神经网络基础上加上各种不同的约束（如前景约束、背景约束、物体大小约束等）来达到学习语义分割的目的。与上述方法类似，Qi 等[65] 采用了物体矩形边框信息来得到物体的伪标注。该伪标注可直接用来训练卷积神经网络。

2.3.2 基于注意力模型的方法

早期方法大多利用各种先验约束来得到语义物体的大致区域。与上述方法不同的是，基于注意力模型的方法[5, 7, 66–70] 能够通过神经网络的定位这一特性得到更为精准的语义物体区域。通过将这些可辨别的种子区域作为伪标注数据来训练卷积神经网络可以得到较好的效果。目前为止，最受欢迎的注意力模型当属 Zhou 等[4] 提出的 CAM 模型。该模型在分类网络的基础上通过对不同特征映射加权平均得到每个类别对应的可辨别区域。得到的类别激活图像具有丰富的语义信息。

Kolesnikov 等[7] 作为最早采用注意力模型来解决弱监督语义分割任务的学者之一，采用了 3 中不用的损失函数来得到种子区域。具体来说，在训练分类模型的基础上，该方法通过引入扩散损失函数使得可辨别的种子区域不断扩大。同时，该方法在训练过程中引入了条件随机场来消除错误的扩散，从而保证种子区域的正确性。Roy 等[66] 在训练分类网络的过程中将得到的初始类别激活图像与过分割图相结合。通过条件随机场的限制，可以用类别无关的过分割图来平滑形状不规则的类别激活图像，进而得到形状较为规则的用来训练卷积神经网络的伪标注。

以上方法虽然可以得到较为精准且可以用于训练语义分割模型的伪标注数据，但由于注意力模型生成的可辨别区域较小，因而限制了这些方法的分割精度。与其不同的是，Wei 等[8] 提出了一种基于对抗擦除的注意力模型来挖掘更多的可辨别区域。该方法在训练完一个注意力模型后，将其中已检测到的可辨别区域从原始输入图像中擦除掉，然后将得到的图像进一步送入分类网络中进

行训练。通过不停地迭代，则可以不断挖掘出更多的可辨别区域。为了防止挖掘的区域并非语义物体，该方法加入了一限制条件：当分类模型无法对辨认出输入图像是否含有其对应的语义类别后则完成训练。将每次迭代得到的注意力区域结合起来则可以得到最终的可辨别区域。通过把得到的可辨别区域与显著性物体检测得到的背景区域相结合，则可以输出较为精准的伪标注数据。尽管该方法能够逐步地将语义物体挖掘出来，其一大弱点在于迭代的过程较为繁琐、终止条件较难设定且挖掘出的区域很容易扩散到背景上。为了解决这一问题，Zhang 等[51] 和 Li 等[50] 分别提出了端到端的对抗擦除模型。通过在原注意力模型的基础上二次挖掘语义区域可以避免 Wei 等方法不停迭代的过程，大大简化了训练的难度。

2.3.3 基于显著性物体检测的方法

图像的显著性图提供了较为丰富的前景信息。由于显著性物体检测模型通常是类别无关的，因而可以被用来检测含有任意类别图像的前景物体。当输入的图像的标签已知且较为简单时（图像中仅含有一类语义物体且前景与背景的对比度较大），显著性物体检测模型可以直接将前景物体分割出来并分配给其类别标签。得到的含有标注的数据则可以直接用来训练语义分割网络。最早将显著性图用于弱监督语义分割任务的方法当属 Wei 等[71] 提出的 STC（Simple-to-Complex）算法。该算法的思想是从简单的图像学起然后慢慢过度到较为困难的图像。该算法首先从网络中爬取大量的简单图像并采用显著性物体检测算法[27] 提取每张图像的前景区域。网络图像通常仅含有一类物体（即用于检索的关键词），因而得到的显著性图以及关键词可以直接被当作伪标注数据来训练语义分割模型。该算法得到的模型仅在简单的图像上进行训练，当被应用到较为复杂的场景时（如 PASCAL VOC 分割数据集[1]），分割结果的质量并不能得到保证。为此，STC 将经网络图像训练后的模型对 PASCAL VOC 2012 训练集中的图像进行推断并用图像的标签数据进行过滤，然后用得到的结果与网络图像训练另一语义分割模型。虽然 STC 采用的方法可以实现图像的语义分割，但其较为依赖于简单图像以及简单图像缺乏多样性的缺陷限制了分割结果的质量。

基于 STC 中的思想，Hou 等[72] 将图像显著性检测的结果与注意力模型相结合，提供了更为精准的伪标注数据。Chaudhry 等[73] 采用了对抗擦除策略对显著性检测算法进行了扩展使其可以被应用于更为复杂的场景。该方法计算由注意力模型[4] 得到的可辨别区域与显著性图之间的调和均值并将得到的结果作为伪

标注数据用于训练语义分割网络。后来的方法[5, 8, 50] 多以图像显著性图的背景作为伪标注的背景以注意力模型得到的可辨别区域作为伪标注的前景从而得到用于训练语义分割网络的伪标注数据。

2.3.4 基于显著性实例的方法

与显著性物体检测不同的是，显著性实例检测[74] 可以将图像前景中的每个实例分割出来。与普通的显著图像相比，显著性实例更容易被应用到弱监督语义分割中。每个实例仅代表着其个体，因而仅需对每个实例分配一个类别标签即可生成伪标注数据。由于一些显著性实例并非语义物体，此时需要一些过滤方法筛选出属于语义类别的实例。Fan 等[75] 基于以上思想提出了一种聚类策略来过滤掉非语义实例。首先该方法将显著性实例检测算法[74] 作用到每张图像上并生成其显著性实例，其次采用注意力模型为每个实例根据其与可辨别区域的交集大小分配一个初始的标签，然后采用分类网络对每个实例区域提取特征并进行聚类操作，最后将距离聚类中心较远的实例过滤掉并将剩余的显著性实例与其类别标签作为伪标注数据。

2.3.4.1 基于网络数据的方法

现有的用于语义分割的数据集[1, 76, 77] 中每个类别至多含有数千张不同的图像。但由于现实生活中场景千变万化，在特定数据集上训练的模型很难处理好内容比较复杂的场景。网络时代的到来提供给研究者们大量免费且丰富的资源。近年来，越来越多的研究人员将重点转移到从网络数据中学习语义分割。Pinheiro 等[61] 利用了 70 万张网络图像以及多实例学习来训练神经网络。Wei 等[71] 使用了 4 万多张简单的网络图像以及显著性物体检测算法来生成伪标注数据。Hong 等[78] 从网上检索了数百个带有语义关键词的视频，然后利用从视频中提取的光流信息与条件随机场[38] 结合得到伪标注数据。Shen 等[79] 也利用大量的网络图像并借助 GrabCut[80] 算法来提取前景信息作为伪标注数据。虽然以上方法皆利用了不同数量的网络图像或视频来生成伪标注数据，但它们仍然需要借助注意力模型的辨别能力来为每张图像提取的前景信息分配类别标签。本文第五章将提出语义分割自主学习的概念，并提出有效方案来实现真正意义上的语义分割自主学习。

第四节 本章小结

本章主要针对本文研究内容相关的文献进行了简要介绍。首先，本章介绍了显著性物体检测相关的算法，包括基于深度学习以及非深度学习的方法。其次，本章介绍了基于卷积神经网络的图像注意力模型，包括早期方法（如 CAM 等）、基于对抗擦除的方法以及基于空洞卷积的方法。最后本章介绍了弱监督语义分割相关的算法，包括一些早期的基于多实例学习的方法、基于注意力模型的方法、基于显著性物体检测的方法、基于显著性实例的方法以及基于网络数据的方法。

第三章　基于短连接的视觉显著性物体检测算法研究

本章主要研究基于短连接的图像显著性物体分割。第一节主要介绍相关背景知识、研究目的以及解决方案概要；第二节给出基于深度监督的网络架构并提出改进方案；第三节介绍如何在改进的深度监督的网络架构上实现基于短连接的显著性物体分割方法；第四节给出本文提出的显著性物体分割算法的实验结果并对结果进行分析；最后，第五节对本章内容进行总结。

第一节　引言

3.1.1　背景知识

显著性物体检测与分割已成为计算机视觉领域一大重要研究话题。显著性物体检测算法旨在检测并分割给定场景中最为显著的物体或区域。在人类的视觉系统中，显著性物体（人眼主要关注的区域）以及物体的边缘等底层视觉特征都是帮助理解场景信息不可或缺的一部分。在计算机视觉以及图像处理领域中，受人类的视觉系统启发，显著性物体检测算法也逐渐被应用到多种任务中，包括基于弱监督的语义分割[71]、图像与视频压缩[81, 82]、图像分割[83]、内容已知的图像编辑[22, 84]、物体检测[85]、视觉跟踪[86]、非图像现实成像[87, 88]、照片合成[23, 89]、信息挖掘[90]、图像检索[24, 91] 以及动作识别[92] 等。

早期的显著性物体检测模型主要依赖于手工设计的特征[21, 93–96] 或通过分类器对所提取的多种特征进行分类[27, 97]，从而判断出图像中每个区域的显著性值。近年来，随着深度卷积神经网络（Convolutional Neural Networks，CNNs）的快速发展，越来越多的学者将研究重点放到基于深度卷积神经网络的方法中。在许多视觉任务中，例如大规模图像分类[37, 98]、语义分割[34]、边缘检测[41, 99]、物体检测[100, 101] 以及行人检测[102] 等，卷积神经网络已经成功地突破了传统方法的结果。

卷积神经网络主要由卷积层以及池化层组成。池化层的下采样功能使得从卷积神经网络中提取丰富的多尺度特征成为可能。现有的基于卷积神经网络的显著性物体检测模型也主要以将不同尺度的特征相融合为主来提取更为高级的特征。显著性物体检测较其它低层级视觉任务（如边缘检测）而言更依赖于卷

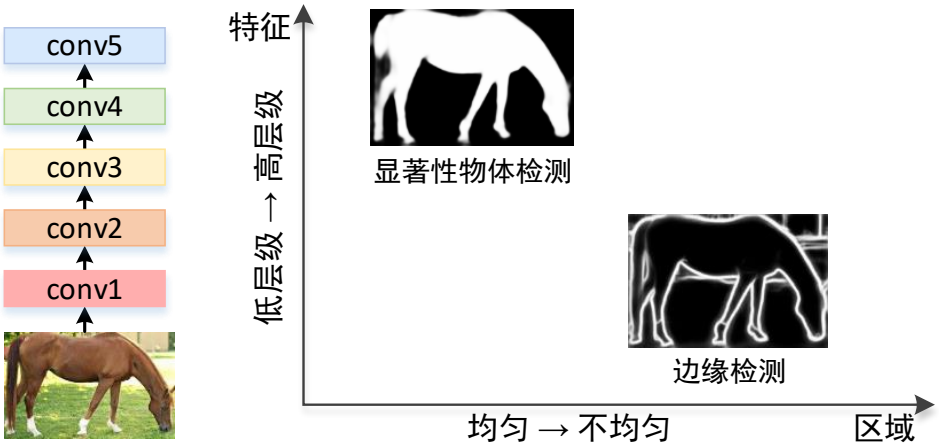


图 3.1 显著性物体检测以及边缘检测对不同特征的依赖。图左为一典型神经网络架构—VGGNet[37]。本图将该网络根据输出尺度的不同把其分为 5 个阶段（conv1-conv5）。自低层到高层分别对应着低层级特征到高层级特征。本章的研究重点在于如何将卷积神经网络高层提取的同质区域与低层提取的丰富的边缘信息相结合使得生成的显著性图像更为完整。

积神经网络的高层语义特征，但其也需要一定量的低层的丰富边缘信息来改善来自高层粗糙的形状不规则的高级语义特征。图3.1给出了显著性物体检测以及边缘检测所依赖的特征类型。因而，如何将不同层级的特征进行有效的融合已然成为构建高质量的显著性物体检测模型的核心问题。

虽然显著性物体检测以及边缘检测对不同层级特征的需求程度是不同的，但其都依赖于多层级特征。因而，现有的显著性物体检测以及边缘检测模型在架构上也较为相似。下面将针对基于深度卷积神经网络的显著性物体检测模型进行介绍。

3.1.1.1 显著性物体检测

随着深度卷积神经网络的发展，越来越多的学者已将其成功地应用到显著性物体检测算法中，尤其是在全卷积网络（Fully Convolutional Networks, FCNs）的出现之后。由于 FCNs 具有大量可学习的参数并且可以进行端到端（end-to-end）的学习，基于 FCNs 的显著性物体检测模型的精度也在不断提升且其运行速度也逐步趋近于实时。本文根据架构的不同简单地将这些方法分为大类，如图3.2所示。

1. 第一类为经典的顺序结构（如图3.2a 所示）。该类方法由于仅采用卷积神经网络的高层级特征而忽略了低层级特征，因而其得到的显著性物体分割

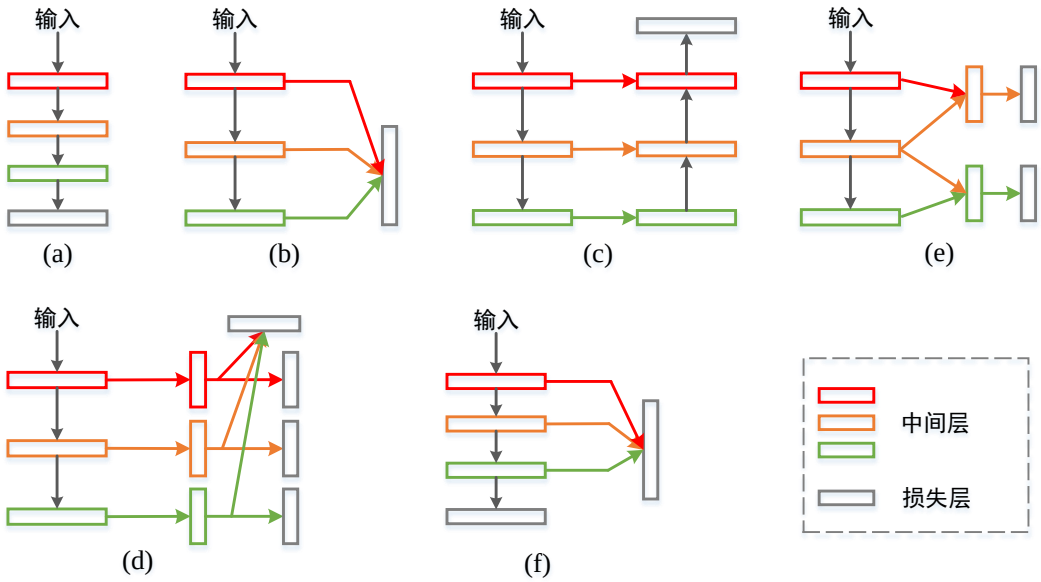


图 3.2 典型的基于深度卷积神经网络的显著性物体检测架构示意图。根据不同方法采用的不同网络架构，本图将现有方法分为六大类。

图通常具有不规则的形状并丢失掉了许多物体的细节信息使得生成的显著图像较为模糊。

2. 第二类（如图3.2b 所示）将不同阶段生成的特征映射先通过双线性插值使其具有相同尺寸，然后将来自各个阶段的特征映射串接起来从而达到应用多尺度信息的目的。由于不同尺度的特征被强行串接在一起，基于该模型的方法生成的结果也因此与第一类相比提升有限。
3. 第三类（如图3.2c 所示）采用了一种编码器—解码器的架构。该类模型首先以典型的分类网络作为编码器并从其不同阶段提取多尺度特征，然后将提取的多尺度特征进行自顶向下的逐步融合。该类方法将高层的高级语义特征逐步传递到低层，取得了较好的效果。
4. 第四类（如图3.2d 所示）与第二类相似都采用了将不同阶段提取的特征进行融合的策略。与第二类方法不同的是，该类方法将相近阶段提取的特征进行融合并采用多个损失函数，避免了第二类方法中将尺度差异较大的特征进行强制融合的弊端。
5. 第五类（如图3.2e 所示）与前四类相比采用了深度监督的概念。除了对所有阶段融合的特征进行监督的同时并对每个阶段的特征进行单独监督。其目的是为了解决卷积神经网络由于层数较多带来的梯度消失问题，同时对

低层特征也进行监督使得低层特征更为丰富。

6. 第六类（如图3.2f 所示）进一步提升了第二类方法。该类方法首先将主干网络不同阶段的特征串接起来并加以深度监督，然后在主干网络的深层加另一监督。在测试阶段，通过将两个预测的显著性图线性叠加即可得到较为完整的显著性物体分割图。

从以上的介绍中可以看出，现有的基于卷积神经网络的显著性物体检测模型大多依赖于将不同阶段提取的多尺度信息进行不同类型的融合从而得到输入图像的显著性分割图，但融合的方式较为简单。

3.1.2 研究意义

显著性物体检测的目标在于识别并分割出给定图像中最为显著的区域。与语义分割不同的是显著性物体检测模型能够快速地从图像中分割出显著的前景区域，这使得显著性物体检测在许多其它视觉任务中成为首要的一环，例如图像和视频的压缩、图像分割、图像编辑、弱监督语义分割、物体追踪、图像合成、图像检索、行为识别等。具体的研究意义可以分为以下几个部分。

- 高效性。显著性区域通常是人们较为关注的区域，因而可以极大地提升物体跟踪、图像编辑等重要的计算机视觉以及图像处理任务。现有主流的显著性区域检测算法大都基于卷积神经网络，因而在 GPU 的帮助下大多可以达到准实时的速度。
- 准确性。由于近年来显著性区域检测的快速发展，出现了越来越多含有数千张训练图像的相关数据集，如 MSRA-B[9] 以及 HKU-IS[11] 等。这些大规模的数据集使得训练高质量的卷积神经网络模型成为可能。丰富的训练数据也使得训练的模型准确度有所保证。
- 通用性。显著性物体检测模型生成的的显著性图通常是连续的灰度图，也即每个像素的值表示该点成为显著性区域的概率是多少。因而，这些模型得到的显著性图通常为类别无关的且能够在一定程度上保持显著性物体的形状信息。这一特性可以有效地应用到基于弱监督的语义分割任务中。通过将得到的前景区域与形状不规则的由注意力模型生成的类别激活图像相结合可以得到质量较高的用于训练语义分割模型的伪标注数据。本文将在第四章与第五章对前述内容进行详细介绍。

根据以上分析，与人的视觉系统相似，在没有任何待检测的场景内容的先验条件下，显著性物体检测模型可以准确地对视觉显著的区域进行自动检测并

分割出相应的物体，因而有着重大研究意义。

3.1.3 解决方案概括

本文提出的显著性物体检测模型基于深度卷积神经网络。由于神经网络的特性，其低层特征通常包含着丰富的细节信息而高层特征通常包含丰富的高级语义信息。图3.3给出了在 VGGNet 不同阶段添加侧向损失后得到的侧向路径效果图。由图可以看出，

1. 高层特征能够较好地帮助找到显著性物体的位置所在，但由于池化层的下采样的效果，得到的显著性图经过双线性插值到图像的原尺寸后，其中显著性物体的形状并不能够得到保证。
2. 与上述内容相反，低层特征能够更有利于找到物体的边缘信息。但这些边缘信息通常含有少量的语义信息，因而许多背景物体的边缘信息也容易被检测出来。

对于卷积神经网络不同阶层特征的特性，本章采用了一种自上而下的深度监督方案。通过在分类网络不同阶段之间引入自上而下的短连接的概念，可以将高层特征的高级语义信息传递到网络低层进而指导低层的特征使其具有判别显著性物体的能力。与此同时，低层特征丰富的细节信息可以与高层特征有效地结合来修复来自高层特征粗糙的边缘，使得生成的显著性物体或区域的形状更为完整。

第二节 基于深度监督的网络架构

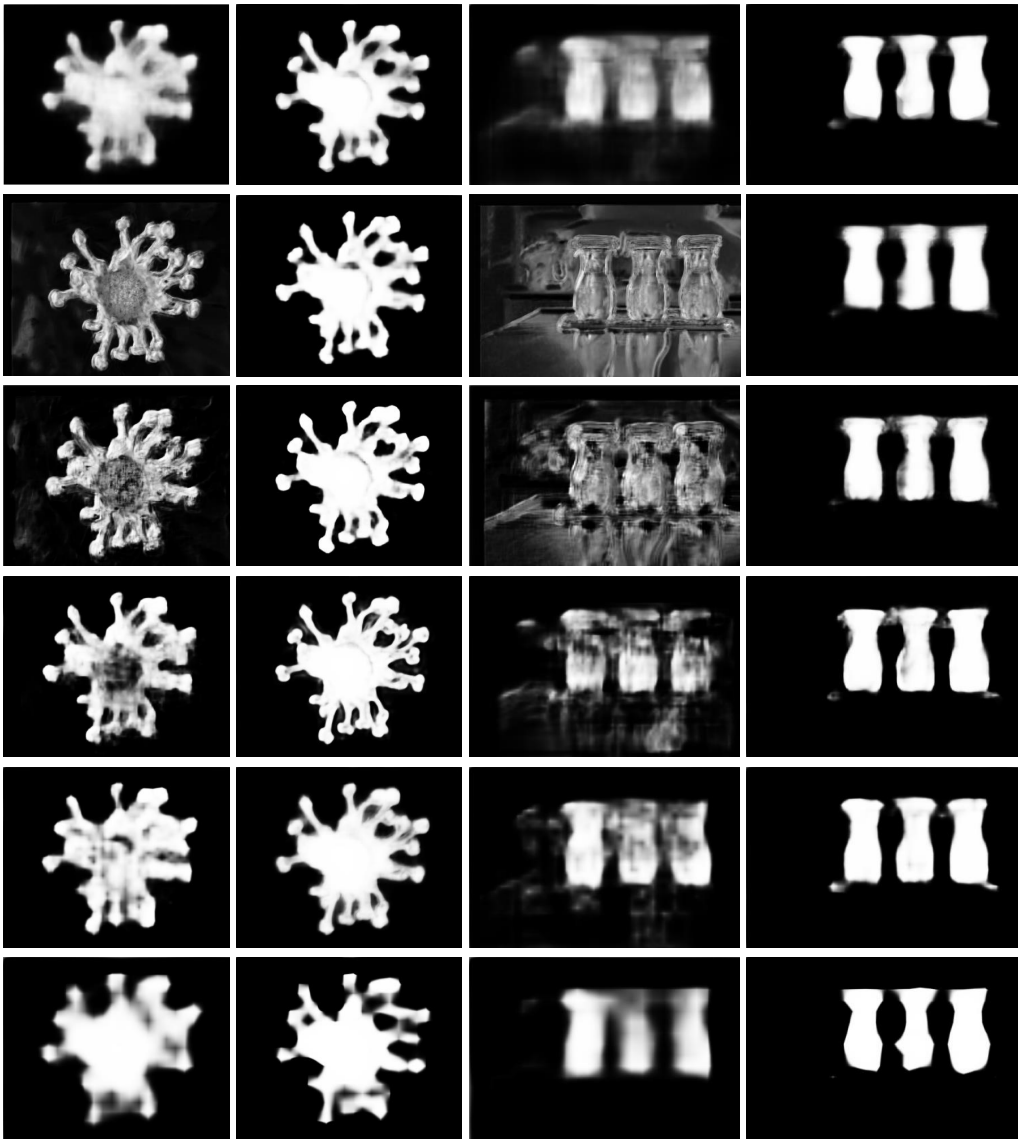
为了更好地理解下文内容，本小节将先简单介绍标准的 HED 架构并提出一种扩展版本来进一步提升其在显著性物体检测任务中的性能。

3.2.1 HED 架构

HED[41] 是一种基于 VGGNet 以及深度监督的新型架构。首先，5 个跳跃连接分别被接到 VGGNet 全卷积部分的五个不同阶段的最后一个卷积层后。每个跳跃连接（侧向路径）分别跟随一个侧向监督（side supervision）。其具体示意图可参见图3.4。

令 $T=\{(X_n,Z_n),n=1,\ldots,N\}$ 表示图像训练集，其中 $X_n=\{x_j^{(n)},j=1,\ldots,|X_n|\}$ 为输入图像且 $Z_n=\{z_j^{(n)},j=1,\ldots,|X_n|\}$，其中 $z_j^{(n)}\in[0,1]$ 为图

(a)



(b)

图 3.3 （a）原始图像以及标注图像；（b）在卷积神经网络不同阶段添加损失函数得到的效果图。其中第一行为基于 HED 方法与本方法的结果图，剩余五行分别为在 VGGNet 不同阶段（conv1-conv5）添加深度监督生成的结果图。奇数列为 HED 对应的结果，偶数列为本方法对应的结果。

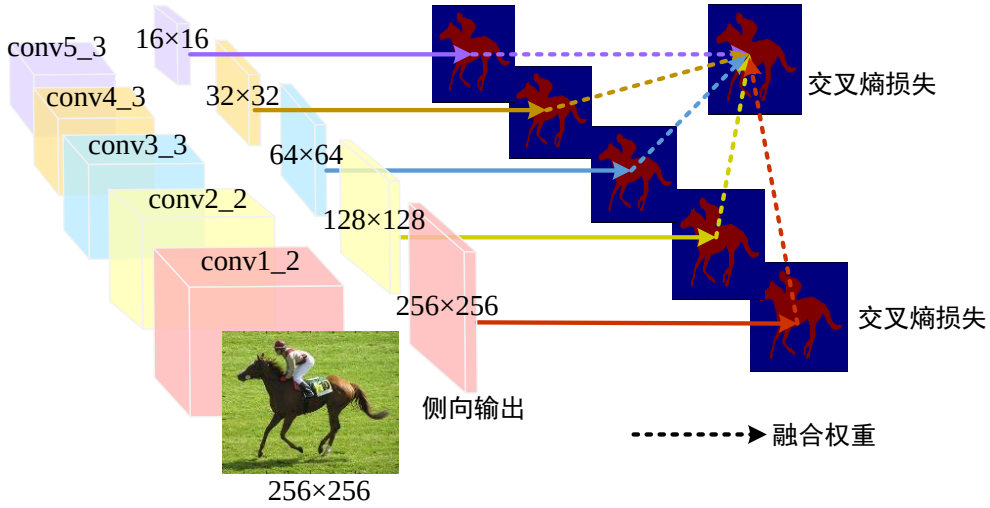


图 3.4 HED 结构[41] 示意图。该方法在 VGGNet 的不同阶段分别引入侧向路径并在每个侧向路径中加上侧向监督。为了更好地结合不同侧向路径的结果，该方法额外引入了一个总损失。

像 X_n 对应的标注图像。在下文中，为了表示方便，下标 n 将被忽略因为所有的输入图像之间都是相互独立的。

令 $\mathbf{W}$ 为卷积神经网络中所有权重的集合。为了不失一般性，这里假设网络中共有 M 个侧向路径（也即跳跃连接）。每一个侧向路径都配有一个分类器，其对应的权重集合可表示为 $\mathbf{w}=(\mathbf{w}^{(1)},\mathbf{w}^{(2)},\ldots,\mathbf{w}^{(M)})$。因此，HED 中的每个侧向目标函数可以表示为

$$L_{\text{side}}(\mathbf{W},\mathbf{w})=\sum_{m=1}^{M}\alpha_m l_{\text{side}}^{(m)}\left(\mathbf{W},\mathbf{w}^{(m)}\right) \tag{3.1}$$

其中 α_m 为第 m 个侧向损失的权重且 $l_{\text{side}}^{(m)}$ 为第 m 个侧向路径中像素级类别平衡交叉熵损失函数[41]。除此之外，一个 1×1 的卷积层以及融合损失层被用来学习每个侧向路径的重要性。该融合损失可以表示为

$$L_{\text{fuse}}(\mathbf{W},\mathbf{w},\mathbf{f})=\sigma\left(Z,h(\sum_{m=1}^{M}f_m A_{\text{side}}^{(m)})\right), \tag{3.2}$$

其中 $\mathbf{f}=(f_1,\ldots,f_M)$ 为每个侧向路径的权重，$A_{\text{side}}^{(m)}$ 为第 m 个侧向路径的激活单元，$h(\cdot)$ 为 sigmoid 函数[1]，$\sigma(\cdot,\cdot)$ 为标注数据与预测显著性图的距离[2]。因此，

[1] $h(x)=\frac{1}{1+\exp^{-x}}$。

[2] 这里采用类别均衡的交叉熵损失[41]。

表 3.1 本模型中每个侧向路径的具体配置。其中，$(n,k\times k)$ 为每个侧向路径中卷积层的通道数以及卷积核大小。“连接点”为每个侧向路径的起始位置。“1”、“2”、“3”分别为每个侧向路径中卷积层的顺序编号。注意，每个侧向路径中的前两个卷积层后都连接一个 ReLU 层[103] 用来达到非线性变换的目的。

编号	连接点	1	2	3
1	conv1_2	$128,3\times 3$	$128,3\times 3$	$1,1\times 1$
2	conv2_2	$128,3\times 3$	$128,3\times 3$	$1,1\times 1$
3	conv3_3	$256,5\times 5$	$256,5\times 5$	$1,1\times 1$
4	conv4_3	$256,5\times 5$	$256,5\times 5$	$1,1\times 1$
5	conv5_3	$512,5\times 5$	$512,5\times 5$	$1,1\times 1$
6	pool5	$512,7\times 7$	$512,7\times 7$	$1,1\times 1$

HED 模型总的损失函数可以表示为

$$L_{\text{final}}(\mathbf{W},\mathbf{w},\mathbf{f}) = L_{\text{fuse}}(\mathbf{W},\mathbf{w},\mathbf{f}) + L_{\text{side}}(\mathbf{W},\mathbf{w}) \tag{3.3}$$

在 HED 架构中，每个侧向路径被分别连接到 VGGNet[37] 每个阶段的最后一个卷积层后，也即 conv1_2，conv2_2，conv3_3，conv4_3，conv5_3。每个侧向路径由一个卷积核大小为 1×1 的卷积层以及一个上采样层组成。通过将所有的侧向路径相结合得到最终的输出结果。

3.2.2 HED 架构的扩展模型

显著性物体检测与边缘检测不同的是其需要分割出同质的区域因而需要更复杂的结构。为此，本章提出了一种 HED 架构的扩展模型。如图3.3所示，卷积神经网络不同阶层提取的特征的特性各不相同。高层可以更好地定位到最为显著的物体的位置。为此，在 HED 结构的基础上，本模型在原始 VGGNet 最后一个池化层（pool5）后额外增加了一个侧向路径。另外，由于显著性物体检测是一种比边缘检测更为复杂的任务，本模型在每一个侧向路径的最前端添加了两个卷积层。由于每一个侧向路径的输入通道数以及尺寸皆不相同，因而每个侧向路径中新加的卷积层的配置也各不相同。低层侧向路径应具有少量的通道数而高层侧向路径应具有较多的通道数。具体信息可以参见表3.1。

与 HED 模型相同，为了使每个侧向路径的预测结果与原始图像具有相同的尺寸，本模型采用双线性插值操作来将小尺寸图像映射到较大尺寸。但与 HED 模型不同的是，本模型采用了标准的交叉熵函数来优化整个模型而非平衡后的

交叉熵函数。其主要原因在于显著性物体检测的正负样本数量较为接近。该损失可以表示为：

$$\hat{l}_{\text{side}}^{(m)}(\mathbf{W},\hat{\mathbf{w}}^{(m)}) = -\sum_{z_j\in Z}\Big[z_j \log \Pr\big(z_j=1|X;\mathbf{W},\hat{\mathbf{w}}^{(m)}\big) + (1-z_j)\log \Pr\big(z_j=0|X;\mathbf{W},\hat{\mathbf{w}}^{(m)}\big)\Big] \tag{3.4}$$

其中 $\Pr\big(z_j=1|X;\mathbf{W},\hat{\mathbf{w}}^{(m)}\big)$ 为第 m 个侧向路径中位置 j 处激活单元显著的概率值。该值可由计算 $h(a_j^{(m)})$ 得到，其中

$$\hat{A}_{\text{side}}^{(m)} = \{a_j^{(m)}, j=1,\ldots,|X|\} \tag{3.5}$$

为第 m 个侧向路径的激活单元的激活值。与 HED 模型[41] 相同，除了以上损失外，本模型也增加了一个全局损失函数来权衡每个侧向路径的重要性。该全局损失函数可以表示为：

$$\hat{L}_{\text{fuse}}(\mathbf{W},\hat{\mathbf{w}},\mathbf{f}) = \hat{\sigma}\big(Z, \sum_{m=1}^{\hat{M}} f_m \hat{A}_{\text{side}}^{(m)}\big) \tag{3.6}$$

其中 $\hat{A}_{\text{side}}^{(m)}$ 为第 m 个侧向路径的新的激活值[3]，$\hat{M}=M+1$，且 $\hat{\sigma}(\cdot,\cdot)$ 为标注图像与最终预测的显著性图像之间的距离。该距离可根据公式3.4计算得到。在本章的第四节将给出扩展版 HED 模型的效果。

第三节 基于短连接的显著性物体分割

由图3.3可以看出，高层的侧向路径可以比较精准地找到显著性物体的位置，但由于高层特征的尺度较小其代价为显著性物体许多细节的丢失。与其相反，低层侧向路径可以捕捉许多显著性物体的细节信息，但其缺点为缺少了获取全局语义信息的能力。该现象说明，如何有效地将高层与低层的侧向路径提取的不同层级的特征进行有效的融合使得显著性物体的位置以及细节都可以被检测到是必要的。本小节在扩展的 HED 模型的基础上引入短连接的概念来实现该目的。

[3]增强的 HED 模型在 HED 模型的基础上添加了一个新的侧向路径。该侧向路径具有更小的尺寸因而可以更好地获取显著性物体的位置信息。

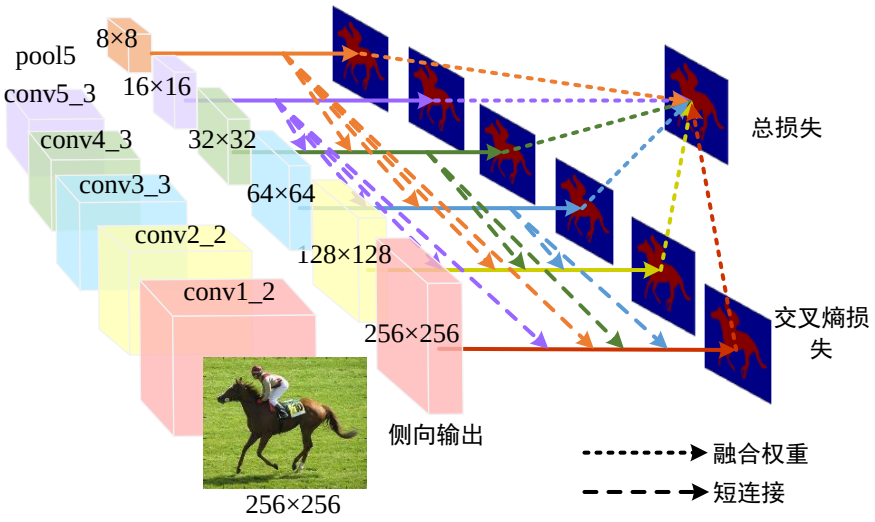


图 3.5 本方法采用的网络架构示意图。由图可知，本方法在已有的 HED 模型的基础上引入了自顶向下的短连接结构。高层的特征信息被传递到低层中。为了更好地将高层特征与低层特征有效地融合，本方法也采用了深度监督的思想。

3.3.1 定义

令 $\tilde{R}_{\text{side}}^{(m)}$ 为第 m 个侧向路径新的激活值，其可表示为

$$\tilde{R}_{\text{side}}^{(m)} = \begin{cases} \sum_{i=m+1}^{\hat{M}} r_i^m \tilde{R}_{\text{side}}^{(i)} + \hat{A}_{\text{side}}^{(m)}, & m = 1,\ldots,5 \\ \hat{A}_{\text{side}}^{(m)}, & m = 6 \end{cases} \tag{3.7}$$

其中，r_i^m 为第 i 个侧向路径到第 m 个侧向路径的短连接（$i > m$）。为了简化模型，r_i^m 可以被设置为 0 来丢弃第 i 个侧向路径到第 m 个侧向路径的短连接。此时，新的侧向损失函数可以被定义为

$$\tilde{L}_{\text{side}}(\mathbf{W}, \tilde{\mathbf{w}}, \mathbf{r}) = \sum_{m=1}^{\hat{M}} \alpha_m \tilde{l}_{\text{side}}^{(m)}(\mathbf{W}, \tilde{\mathbf{w}}^{(m)}, \mathbf{r}) \tag{3.8}$$

以及

$$\tilde{L}_{\text{fuse}}(\mathbf{W}, \tilde{\mathbf{w}}, \mathbf{f}, \mathbf{r}) = \hat{\sigma}\left(Z, \sum_{m=1}^{M} f_m \tilde{R}_{\text{side}}^{(m)}\right) \tag{3.9}$$

其中，$\mathbf{r} = \{r_i^m\}$，$i > m$。注意，此时 $\tilde{l}_{\text{side}}^{(m)}$ 为标准的交叉熵函数，即公式3.4。全局损失函数可以定义为

$$\tilde{L}_{\text{final}}(\mathbf{W}, \tilde{\mathbf{w}}, \mathbf{f}, \mathbf{r}) = \tilde{L}_{\text{fuse}}(\mathbf{W}, \tilde{\mathbf{w}}, \mathbf{f}, \mathbf{r}) + \tilde{L}_{\text{side}}(\mathbf{W}, \tilde{\mathbf{w}}, \mathbf{r}) \tag{3.10}$$

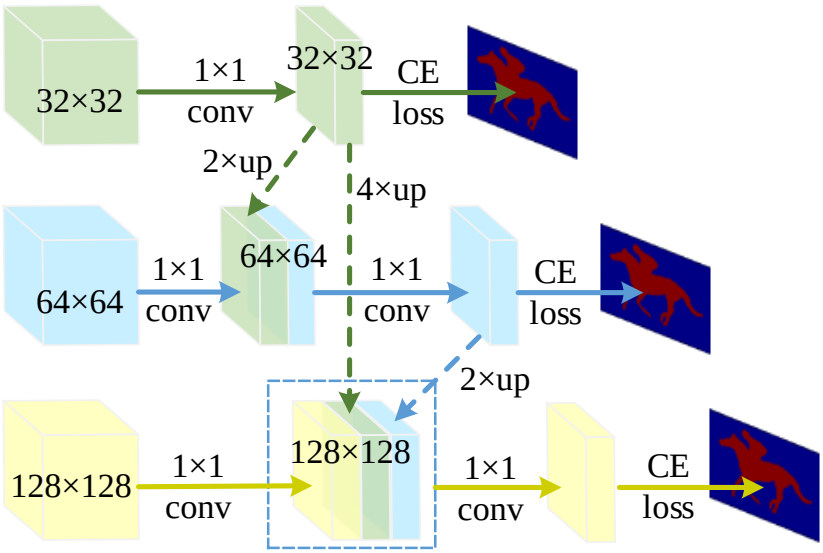


图 3.6 不同侧向路径融合策略。图中“up”表示上采样操作、“conv”为卷积层且“CE Loss”为交叉熵损失。高层的侧向路径被连接到底层的侧向路径中并通过 1×1 的卷积层来完成融合。

3.3.2 网络架构

本模型的主干网络为扩展版的 HED 模型（参见3.2.2）。图3.6给出了如何构建从侧向路径 4 到侧向路径 2 的短连接结构。侧向路径 3 以及 4 中得到的预测映射首先经由双线性插值然后被串接到侧向路径 2 中初始的预测映射上。双线性插值的参数可由上下文直接推导得出。由于显著性物体检测是一个类别无关的视觉任务，本模型采用一个卷积核大小为 1×1 的卷积层来权衡来自三个侧向路径预测映射的权重。具体示意图可参见图3.6中虚线方框内的连接方式。当有更多短连接被连接到某一侧向路径时，相同的策略可以将所有的输入短连接融合起来。

事实上，本架构可以被理解为两个紧密相连的阶段—显著性物体定位阶段以及细节特征精炼阶段。显著性物体定位阶段的主要目的在于寻找给定图像中最为显著的物体。对于细节特征精炼阶段，本模型采用了一种自顶向下的方法，即通过引入一系列的从高层侧向路径到低层侧向路径的短连接结构来达到获取细节特征的目的。这样做的原因在于在高层语义信息的帮助下，低层侧向路径可以既准确地预测到显著的物体又可以利用低层侧向路径的细节特征来优化来自高层侧向路径的粗糙的预测图。本章将在实验小节给出更多实验细节以及本模型的效果。

3.3.3 实现细节

本模型采用可公开获取的 Caffe[104] 工具包来实现。如前几小节所述，为了公平与现有模型相比，本模型采用了 16 层的 VGGNet[37] 作为预训练模型。

3.3.3.1 测试环节

尽管本模型引入了一系列的短连接结构来生成显著性图，但由高层侧向路径或低层侧向路径生成的预测图仍然提升有限。为了生成更高质量的显著性图，在测试阶段本模型将不同的侧向路径得到的预测图进行了一种复杂的组合。令 $\tilde{Z}_1,\cdots,\tilde{Z}_6$ 分别为每个侧向路径生成的预测图。其计算方式可表示为

$$\tilde{Z}_m = h(\tilde{R}_{\text{side}}^{(m)}) \tag{3.11}$$

其中，$h(\cdot)$ 为 sigmoid 函数。因此，最终的输入结果图可以表示为

$$\tilde{Z}_{\text{fuse}} = h\Big(\sum_{m=2}^{4} f_m \tilde{R}_{\text{side}}^{(m)}\Big) \tag{3.12}$$

为了避免来自高层以及低层侧向路径生成的低质量预测图的干扰，本模型将进一步采用 $\tilde{Z}_2$、$\hat{Z}_3$ 以及 $\hat{Z}_4$ 来提升最终的结果图。因此，最终的输出显著性图可以表示为

$$\tilde{Z}_{\text{final}} = \frac{1}{4}(\tilde{Z}_{\text{fuse}} + \tilde{Z}_2 + \tilde{Z}_3 + \tilde{Z}_4) \tag{3.13}$$

3.3.3.2 平滑策略

尽管本模型引入了短连接的结构来提升模型分割显著性物体的能力，但检测到的显著性物体的边缘信息仍旧有所缺失。当输入图像为较为复杂的场景时，这一现象尤其明显。为了进一步提升本模型生成的显著性图的空间相关性，本章采用基于全连接的条件随机场（CRF）[38] 来作为一种可选择的后处理工具进行平滑操作。

CRF 的能量函数可以表示为

$$E(\mathbf{x}) = \sum_i \theta_i(x_i) + \sum_{i,j} \theta_{ij}(x_i, x_j) \tag{3.14}$$

其中，$\mathbf{x}$ 为像素的预测显著性值。为了进一步提升最终结果的质量，本模型并非直接采用来自卷积神经网络的预测结果作为条件随机场的一元势函数而是采用

如下形式

$$\theta_i(x_i) = -\frac{\log \hat{S}_i}{\tau h(\hat{S}_i)} \tag{3.15}$$

其中，$\hat{S}_i$ 为归一化后像素 x_i 的显著性值，$h(\cdot)$ 为 sigmoid 函数，τ 为一系数。条件随机场的点对势函数被定义为

$$\begin{aligned}\theta_{ij}(x_i, x_j) = \mu(x_i, x_j)\Bigg[w_1 \exp\Bigg(-\frac{\|p_i - p_j\|^2}{2\sigma_\alpha^2} - \\ \frac{\|I_i - I_j\|^2}{2\sigma_\beta^2}\Bigg) + w_2 \exp\Bigg(-\frac{\|p_i - p_j\|^2}{2\sigma_\gamma^2}\Bigg)\Bigg]\end{aligned} \tag{3.16}$$

其中，如果 $x_i \neq x_j$ 成立，则 $\mu(x_i, x_j) = 1$ 否则为 0，I_i 和 p_i 分别为像素值以及像素 x_i 的位置。参数 w_1、w_2、σ_α、σ_β 以及 σ_γ 分别为高斯核的权重系数。

本章采用了一种开源的 PyDenseCRF[4] 来实现上述模型。条件随机场的输出即为本方法的最终输出显著性图。

3.3.3.3 超参数设置

本模型采用的超参数包括：学习速率（1e-8），权值衰减（0.0005），动量（0.9），每个侧向路径的权重（1）。为了更好地学习训练图像中的显著性物体，输入图像以全分辨率的形式作为输入且训练时的批量大小为 10。新添加的卷积层中的参数皆采用高斯随机数初始化。用于全连接条件随机场的超参数则由验证集交叉验证后得到。在实验中，这些参数分别被设置为：τ（1.05），w_1（3.0），w_2（3.0），σ_α（60.0），σ_β（8.0），σ_γ（5.0）。

第四节 实验验证

本小节将介绍本章所用的测试数据集以及评测方法，同时也将介绍本模型在这些数据集上的结果。除此之外，为了更好地理解本模型中每个部分的重要性，本小节也将详细地对本模型中每个部分进行敏感性分析。

3.4.1 数据集

本小节采用 5 个标准的、被广泛使用的数据集来进行评测，其中包括 MSRA-B[9]，ECSSD[10]，HKU-IS[11]，PASCALS[12] 以及 SOD[13, 14]。所有的数据集都可公开获取且包含数百乃至数千张训练或测试图像以及高质量的标注数据。

[4]https://github.com/lucasb-eyer/pydensecrf

为了保证与现有方法的公平对比以及评测的完整性，本模型在训练阶段采用了与 DRFI[105] 方法相同的训练数据集用来训练，然后将得到的模型在其它测试数据集上做测试。

3.4.2 评测标准

与多数现有方法类似，本小节采用了三种被广泛使用的标准的评测方法来测评本模型的效果，其中包括精度-召回曲线（precision-recall curves，PR-Curve），F-measure，以及平均绝对误差（mean absolute error, MAE）。关于以上评测方法的具体介绍可参见 Borji 等[106]。

- 首先，对于一个给定的显著性图 S 可设置一个阈值将其转换为一个二值映射图 B。该显著图的精度值以及召回率可由下式计算得到，

$$precision = \frac{|B \cap Z|}{|B|} \quad recall = \frac{|B \cap Z|}{|Z|} \tag{3.17}$$

  其中 $|\cdot|$ 为一映射中所有非零项的总和。给定一个测试数据集，在所有得到的显著性图的基础上计算出其精度值以及召回率则可得到其精度-召回曲线（PR-Curve）。
- 为了定量评测显著性图的质量如何，F-measure 通常被用来作为评测标准。其计算公式可表示为

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall} \tag{3.18}$$

  其中，$Precision$ 为精度值而 $Recall$ 为召回率。上式中 β^2，依照现有工作[25, 94, 106] 建议，通常被设置为 0.3。其目的为在测试阶段更侧重式中精度的重要性。
- 令 $\hat{S}$ 及 $\hat{Z}$ 来表示归一化后的显著性图以及标注数据，则显著性图的平均绝对误差的计算公式为

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |\hat{S}(i,j) - \hat{Z}(i,j)| \tag{3.19}$$

  该式反映了显著性图以及标注数据之间的距离。

3.4.3 模型参数敏感性分析

本部分将对本模型中的参数的敏感性进行分析，主要包括：短连接配置比较、侧向路径内部结构对比、上采样操作、数据增强以及主干网络的影响。

表 3.2 侧向路径中卷积层的不同配置以及基于以上配置模型在 PASCALS[12] 数据集上的量化结果。其中，$(c,k\times k)\times n$ 表示某侧向路径中共有 n 个卷积层且通道数以及卷积核大小分别为 c 和 $k\times k$。注意，每个侧向路径中最后一个卷积层与表3.1中相同。在每一种配置中，仅仅相应的参数被改变而其它所有参数均保持不变。

No.	1	2	3	4
侧向路径 1	$(128,3\times 3)\times 2$	$(128,3\times 3)\times 1$	$(128,3\times 3)\times 2$	$(128,3\times 3)\times 2$
侧向路径 2	$(128,3\times 3)\times 2$	$(128,3\times 3)\times 1$	$(128,3\times 3)\times 2$	$(128,3\times 3)\times 2$
侧向路径 3	$(256,5\times 5)\times 2$	$(256,5\times 5)\times 1$	$(256,3\times 3)\times 2$	$(256,5\times 5)\times 2$
侧向路径 4	$(512,5\times 5)\times 2$	$(256,5\times 5)\times 1$	$(256,3\times 3)\times 2$	$(256,5\times 5)\times 2$
侧向路径 5	$(1024,5\times 5)\times 2$	$(512,5\times 5)\times 1$	$(512,5\times 5)\times 2$	$(512,5\times 5)\times 2$
侧向路径 6	$(1024,7\times 7)\times 2$	$(512,7\times 7)\times 1$	$(512,5\times 5)\times 2$	$(512,7\times 7)\times 2$
F_β	**0.830**	0.815	0.820	**0.830**

3.4.3.1 短连接配置比较

本模型架构（如图3.5所示）的一大特点为其灵活性。因而，本模型可以看做多数现有模型架构（如图3.2所示）的通用版本。为了更好地展示本模型在不同短连接配置情况下的结果，除了现有模型结构（如 Hypercolumns[40] 以及增强版的 HED 模型），本小节主要列出 3 种典型的连接方式用来比较。

第一种方式较为简单，其表达式为

$$\tilde{R}_{\text{side}}^{(m)}=\begin{cases} r_{m+1}^{m}\tilde{R}_{\text{side}}^{(m+1)}+\hat{A}_{\text{side}}^{(m)} & \text{对于 } m=1,\ldots,5 \\ \hat{A}_{\text{side}}^{(m)} & \text{对于 } m=6 \end{cases} \tag{3.20}$$

第二种配置的表达式为

$$\tilde{R}_{\text{side}}^{(m)}=\begin{cases} \sum_{i=m+1}^{m+2} r_{i}^{m}\tilde{R}_{\text{side}}^{(i)}+\hat{A}_{\text{side}}^{(m)} & \text{对于 } m=1,2,3,4 \\ \hat{A}_{\text{side}}^{(m)} & \text{对于 } m=5,6 \end{cases} \tag{3.21}$$

第三种配置的表达式为

$$\tilde{R}_{\text{side}}^{(m)}=\begin{cases} \sum_{i=3}^{6} r_{i}^{m}\tilde{R}_{\text{side}}^{(i)}+\hat{A}_{\text{side}}^{(m)}, & \text{对于 } m=1,2 \\ r_{5}^{m}\tilde{R}_{\text{side}}^{(5)}+r_{6}^{m}\tilde{R}_{\text{side}}^{(6)}+\hat{A}_{\text{side}}^{(m)} & \text{对于 } m=3,4 \\ \hat{A}_{\text{side}}^{(m)} & \text{对于 } m=5,6 \end{cases} \tag{3.22}$$

上述不同配置的量化结果可见表3.3。从该表中可以看出，当在 HED 模型的基础上添加一个额外的侧向路径以及两个额外的卷积层后，模型的 F-measure 值

表 3.3 不同架构配置在 PASCALS 数据集[12] 上的效果对比。本图中‘*’表示本文所采用的模型架构。最好的配置对应的 F-measure 值已被加粗。可以看出，当短连接的数目增加时，得到的效果也有所提升。

方案	架构	F-measure
1	Hypercolumns[40]	0.818
2	原始 HED[41]	0.791
3	增强版 HED	0.816
4	配置 1 (公式3.20)	0.816
5	配置 2 (公式3.21)	0.824
6	配置 3* (公式3.22)	**0.830**

提升了 2.5 个点。另外，随着短连接个数的增加，本模型也逐渐实现了更好的效果。当将配置 1 与配置 2 相比时，得到的 F-measure 值提升了 0.8 个点。当配置改为 3 时，可以得到另外 0.6 个点的提升。

3.4.3.2 侧向路径内部结构对比

本部分的目的在于探索侧向路径的最佳参数配置。在每次实验中，每个侧向路径的详细配置可参见表3.2。为了公平比较，表3.3中的配置 3 被用来作为基准模型。为了对比不同参数的重要性，本部分采用控制变量法使得不同模型中每次只有一个参数发生变化。除此之外，本实验所有的测试结果皆基于 PASCALS 数据集。

通过比较实验 1 与实验 4，可以发现引入更多卷积层通道数之后，F-measure 值无任何提升。在实验 2 中，每个侧向路径的通道数由两个减为 1 个。该操作使得 F-measure 值下降 1.5 个百分点。这一现象表明为每个侧向路径引入额外两个卷积层有一定效果。

除此之外，为了探究上下文信息对实验结果的影响，实验 3 将侧向路径中卷积层的卷积核的大小由大变小。由表可看出，这一操作使得 F-measure 值有了微小的下降。这一现象说明为高层侧向路径引入较大的卷积核可以捕获更多特征图中的上下文信息，进而带来整体效果上的提升。

3.4.3.3 上采样操作

在本实验小节，每个侧向路径中默认的上采样函数都为双线性插值。为了证明双线性插值的有效性，本部分采用了可学习的反卷积操作来实现上采样。

实验结果表明，引入可学习的反卷积操作无任何提升。

为了进一步增大本模型的感受野，每个侧向路径中的卷积操作被空洞卷积所替代。具体相关操作可参见 DCL 模型[36]。实验结果表明，引入空洞卷积降低了本模型的效果。随着生成的显著性图更为稠密，即使 CRF 被用来作为后处理工具后，许多非显著性区域亦被误判成显著区域。该操作使得 F-measure 值下降近 1 个百分点。

3.4.3.4 数据增强

数据增强已被证明为基于学习的视觉任务的一个重要数据处理环节。同现有方法相同，本模型中所有训练图像在训练时被随机横向翻转。该操作使得训练的数据量增加了一倍。实验结果表明，训练图像的随机横向翻转可以带来 0.5 个百分点的提升。这一现象表明，数据增强对于显著性物体检测也有一定效果。

除了简单的横向翻转图像，剪切原始图像到已固定大小（321×321）也被采用。实验结果表明这一操作使得最终结果下降了 0.5 个百分点。这说明用原始图像的尺寸进行训练对最终结果有更大帮助，其原因在于模型可以更容易地从整张图像中学习到显著性物体。

3.4.3.5 主干网络的影响

除了采用 VGGNet 作为本模型的主干网络，ResNet-101 模型[107] 也被用来作为主干网络。考虑到 ResNet-101 网络与 VGGNet 的不同之处，本实验仅采用表3.1中的后 5 个侧向路径。这 5 个侧向路径分别被连接到 ResNet-101 网络的 conv1、res2c、res3b3、res4b22 以及 res5c 层后。其它所有的配置未变动。实验结果可参见表3.4的底部。可以发现，在相同的配置下（除了主干网络），引入 ResNet-101 作为主干网络可以带来额外 1 个点左右的效果提升。

3.4.4 与现有模型的比较

本小节将本模型与 7 种现有的基于卷积神经网络的方法进行比较，其中包括 MDF[11]、DS[108]、DCL[36]、ELD[33]、MC[32]、RFCN[39] 以及 DHS[35]。除此之外，4 中经典的方法也被用于作为对比，包括 GC[25]、CHM[109]、DSR[110] 以及 DRFI[105]。这些经典的方法皆为一总结性论文[106] 中最好的模型。值得一提的是尽管更多的训练数据可能会进一步提升本模型的实验结果，为了公平与现有方法进行比较，本小节仅列出基于 2500 张图像训练的模型，也即本方法的训

表 3.4 本模型与 11 个现有方法在 5 个流行数据集上的量化对比结果。本模型 ResNet-101[107] 版本用符号 † 表示。可以看出，ResNet-101 版本明显优于 VGGNet 版本。为了公平地与其它方法相比，这里将除 ResNet-101 版本以外最好的结果用粗体加以强调。

	MSRA-B[9]		ECSSD[10]		HKU-IS[11]		PASCALS[12]		SOD[14]	
方法	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
GC[25]	0.817	0.138	0.741	0.187	0.726	0.165	0.640	0.225	0.657	0.242
CHM[109]	0.809	0.138	0.722	0.195	0.728	0.158	0.631	0.222	0.655	0.249
DSR[110]	0.812	0.119	0.737	0.173	0.735	0.140	0.646	0.204	0.655	0.234
DRFI[105]	0.855	0.119	0.787	0.166	0.783	0.143	0.679	0.221	0.712	0.215
MC[32]	0.872	0.062	0.822	0.107	0.781	0.098	0.721	0.147	0.708	0.184
ELD[33]	0.914	0.042	0.865	0.981	0.844	0.071	0.767	0.121	0.760	0.154
MDF[11]	0.885	0.104	0.833	0.108	0.860	0.129	0.764	0.145	0.785	0.155
DS[108]	-	-	0.810	0.160	-	-	0.818	0.170	0.781	0.150
RFCN[39]	0.926	0.062	0.898	0.097	0.895	0.079	0.827	0.118	0.805	0.161
DHS[35]	-	-	0.905	0.061	0.892	0.052	0.820	0.091	0.823	0.127
DCL$^+$[36]	0.916	0.047	0.898	0.071	0.907	0.048	0.822	0.108	0.832	0.126
DSS	**0.927**	**0.028**	**0.915**	**0.052**	**0.913**	**0.039**	**0.830**	**0.080**	**0.842**	**0.118**
DSS†	0.936	0.030	0.928	0.048	0.920	0.035	0.838	0.092	0.850	0.119

练数据为 MSRA-B 训练集中的 2500 张图像。

3.4.4.1 视觉对比分析

为了对比本模型生成的显著性图与现有显著性检测方法的优势，本小节从多个数据集中选取多个具有代表性的图像进行展示。为了突显本模型处理不同场景的能力，图3.7和图3.8 对每个图像的属性进行了标注，包括场景的难易程度、是否含有中心点偏移、显著性物体的大小、前景背景的对比度、是否为透明物体等。

从图3.7和图3.8可以看出，在不同的输入场景下，本模型可以很好地将显著性物体从背景中提取出来。不仅如此，本模型在处理显著性物体侧向上与其它方法相比更具有一定优势。这是由于本模型可以合理地将高层语义特征与低层的细节信息更好地融合。这一特性使得本模型生成的显著性图更接近原始标注图像。

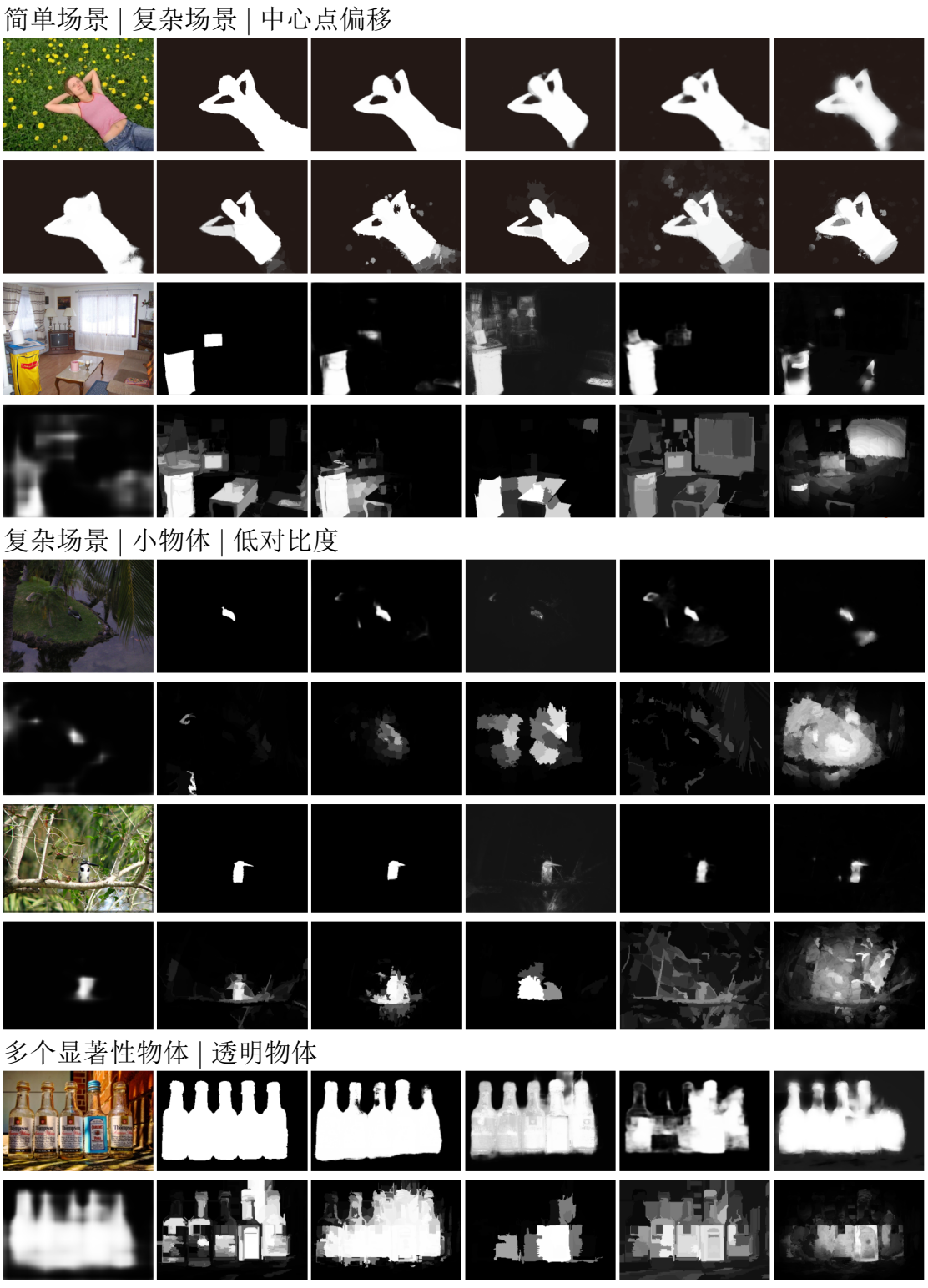


图 3.7 来自不同数据集的显著性图（一）。根据图像的不同属性，这里将其分为多个部分。每组图像对应的标签分别为：输入图像、标注图像、本方法、DCL、DHS、RFCN、DS、MDF、ELD、MC、DRFI 以及 DSR。

图 3.8 来自不同数据集的显著性图（二）。根据图像的不同属性，这里将其分为多个部分。每组图像对应的标签分别为：输入图像、标注图像、本方法、DCL、DHS、RFCN、DS、MDF、ELD、MC、DRFI 以及 DSR。

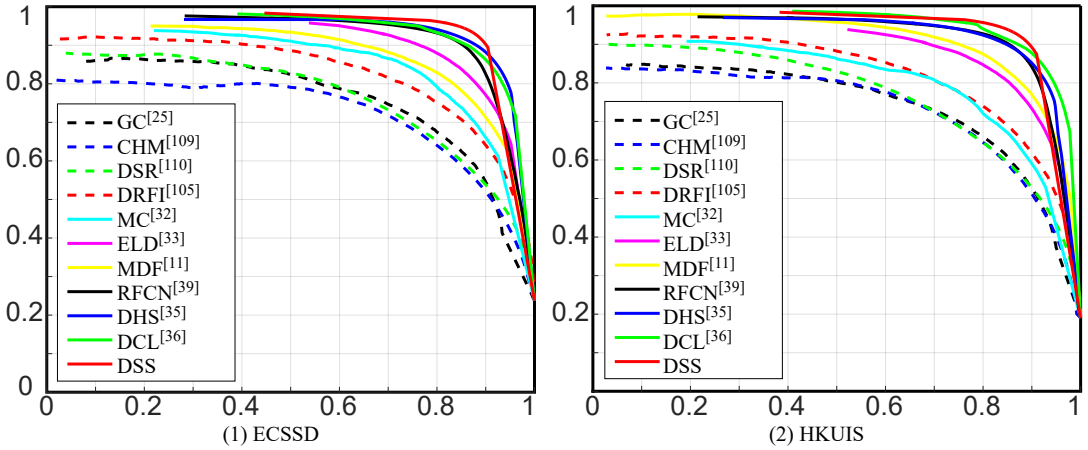


图 3.9 本模型（DSS）在数据集 ECSSD 以及 HKU-IS 上的精度（纵轴）-召回（横轴）曲线（PR-Curve）。从两张图中可以看出，本方法的精度-召回曲线（由红色实线表示）明显优于其它所有方法的曲线，尤其在坐标系的右上角区域。

3.4.4.2 PR 曲线

本小节将以精度-召回曲线（PR 曲线）的形式与其它方法相比。图3.9给出了本模型在 ECSSD 以及 HKU-IS 数据集上的 PR 曲线以及与其它模型的对比。可以看出，基于全卷积网络的模型在 PR 曲线评测上有较好的效果。其中，本模型的结果更优于其它基于全卷积网络的模型。当召回率接近 1 时，本方法的结果得到的曲线明显高于其它方法。

3.4.4.3 F-measure 以及 MAE

除了 PR 曲线，本部分采用 F-meature 以及 MAE 来评测本方法以及与现有方法的区别。定量的结果可参见表3.4。可以看出，本方法的结果在 F-measure 以及 MAE 上的评测结果在所有 5 个数据集上明显优于其它所有方法。例如，在 ECSSD 以及 SOD 数据集上，本方法与现有最好的方法相比有近一个点的提升，且已非常接近理想值。对于 MAE 而言，本方法在 MSRA-B 以及 PASCAL-S 数据集上也有近一个点的优势。这些现象表明，本方法生成的结果与其它方法相比拥有较少的误判区域。

除此之外，从表3.4中仍可观察到本方法在较难的数据集上（例如 HKU-IS[11]、PASCAL-S[12] 以及 SOD[13, 14]）与简单数据集上相比具有更好的效果。这表明本方法能够更好地检测以及分割出显著性物体。

3.4.5 显著与否

目前为止，绝大多数显著性物体检测方法主要专注于至少含有一个显著性物体的数据集上。事实上，多数真实生活场景中并不一定含有显著性物体。因此，直接假设每一场景中都含有显著性物体可能会导致错误判断，且例如 F-measure 等评测指标并不能够直接被用来评测不含有显著性物体的图像。为了解决这个问题，本方法在原有模型的基础上引入了另一个分支用来预测输入的场景中是否含有显著性物体。如果本模型预判出某场景中含有显著性物体，则同时输出相应的显著性图像。

新引进的分支由一个全局均值池化层以及一个多层感知机组成。与多数分类网络[37, 107] 相似，其主要目的为识别输入图像中是否含有显著性物体。在该结构中，全局均值池化层首先被用来将具有不同尺度的特征图映射到相同的大小从而使得得到的特征向量可以被送到多层感知机中做出预测。与现有识别网络[37, 101] 类似，本方法采用的多层感知机包含 2 个含有 1024 个神经元的全连接层以及一个含有两个神经元的全连接层。该分支所用回归函数与现有识别网络相同[37]。

在实验中，本方法采用了与 SSVM[111] 相同的数据集用来训练。该数据集包含了 5000 张背景图像（也即没有显著性物体的图像）以及 5000 张选自 MSRA10K[25] 的图像。对于所有的背景图像，来自显著性物体检测部分的梯度信息并不会被回传。其目的是为了减少检测显著性分支对显著性物体检测部分的影响。

训练时，模型的超参数与显著性物体检测模型的超参数相同。不同的是，由于训练图像的增多，新的模型采用了 24000 次迭代并且在 20000 次迭代时学习速率除以 10。为了测试新模型的效果，三个数据集被采用，包括 JSOD[111]、MSRA-B[9] 以及 ECSSD[10]。表3.5给出了本方法在这些数据集上与现有方法 SSVM[111] 以及王等[112] 的对比结果。由于现有数据集的显著性物体都比较容易被发现，因而得到的结果也都接近理想值（1）。尽管如此，本方法在相对较难的数据集（MSRA-B 以及 ECSSD）中得到更好的结果。

3.4.6 运行效率

由于本模型采用的是全卷积神经网络，因而与现有方法相比在运行速度上也具有一定优势。当采用 MSRA-B 中的 2500 张图像作为训练集时，模型迭代

表 3.5 本模型（DSS）在检测图像中是否具有显著性物体上的实验结果。主要的比较方法为 SSVM[111] 以及王等[112]。每一列中的最好方法的结果已用黑体加粗。

方法	JSOD[111]	MSRA-B[9]	ECSSD[10]
王等[112]	90.64%	89.26%	70.50%
SSVM[111]	**99.22%**	98.66%	94.40%
DSS	98.84%	**99.05%**	**96.8%**

12000 次共需要 8 个小时左右。更有趣的是，尽管 10000 次迭代已经足以使本模型达到收敛状态，但继续迭代模型 2000 次可以使相应结果的 MAE 值得到进一步提升。

在测试阶段，当处理一张尺寸大小为 300 × 400 的图像时仅需要 0.08 秒的时间。这个速度与现有方法（例如 DCL[36] 等）相比具有明显优势。用 CRF 作为后处理工具时每张图像将耗费额外 0.4s 的时间。因此，处理一张尺寸大小为 300 × 400 的图像总的运行时间为 0.5 秒左右。

3.4.7 错误结果分析

本小节将对本方法所产生的错误结果进行简要分析。其主要目的是提供给研究人员更多有用的信息来开发更为先进的显著性物体检测架构以及更为有用的损失函数。

图3.10给出了本方法在处理不同图像时的失败案例。可以看出，这些错误结果可以大致被分为以下三种情况。

- 第一种为大多数基于卷积神经网络模型的共同缺点，即显著性物体并不能被完整的检测出来。换句话说，少部分显著性物体始终被模型判成背景。典型的例子为图3.10 中的前两行。解决该类问题的办法可从增大卷积网络的感受野以及设计更为复杂的网络结构来考虑。
- 第二种情况为图像中显著性物体的主体部分没有被检测到，或者说部分（背景中）非显著性区域被误判成显著性物体。如图3.10中间部分所示，该类情况的主要原因在于输入图像的背景通常非常复杂或含有对比度非常低的前背景。该种情况可以通过增加用于训练的数据量或者采用更多有效的数据增强的方法来解决。
- 第三种情况通常是由于显著性物体的透明性引起的。如图3.10底部所示，尽管本方法的结果可以部分检测出透明显著性物体，但将其完整的分割出

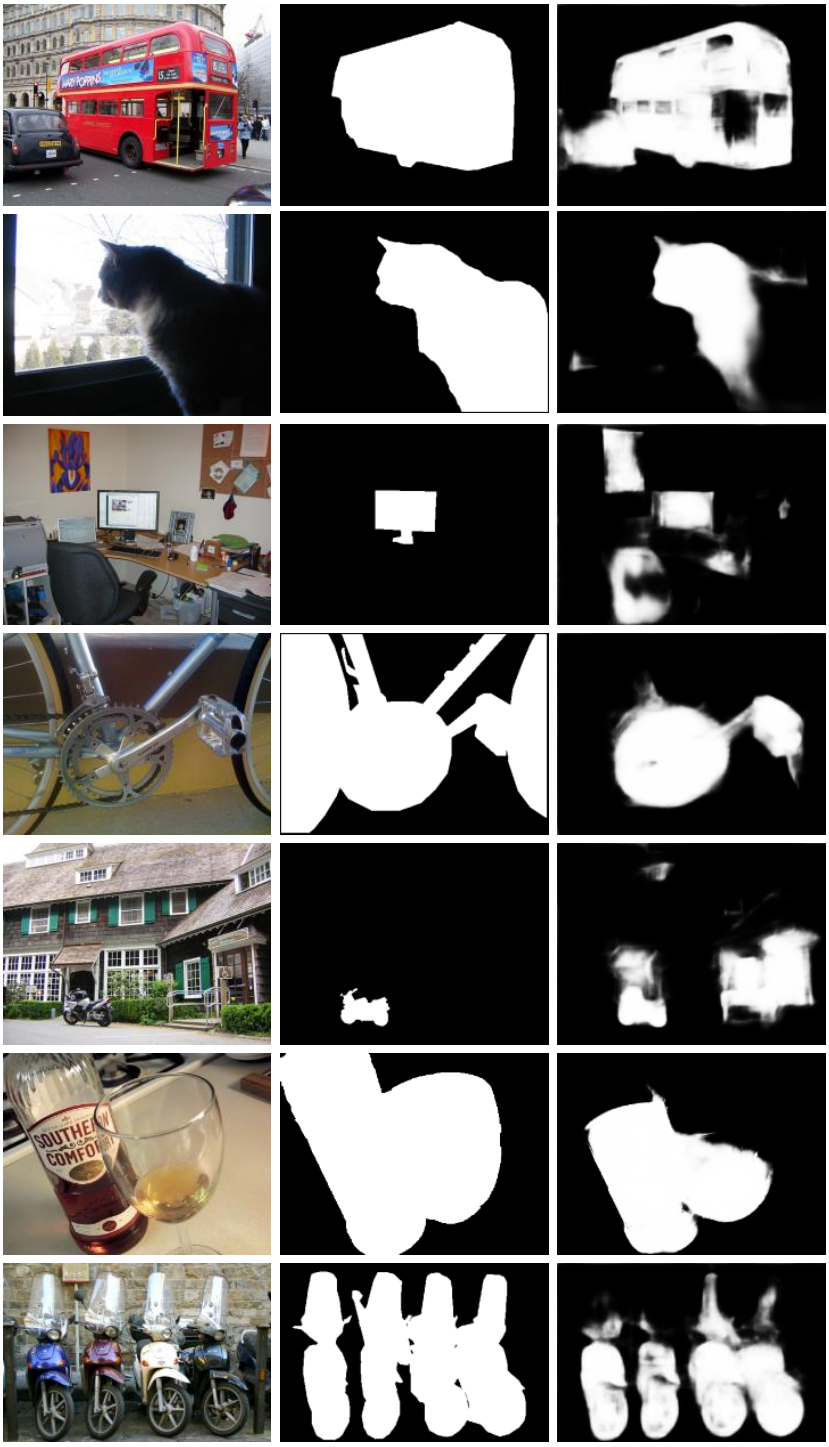


图 3.10 错误结果分析。从左到右分别为：输出图像、标注图像以及预测结果。该图中的图像来自于多个测试数据集。由该图可以看出，主要分割错误的图像均是包含较为复杂场景或是具有较低的前背景对比度的图像。另外，对于部分含有透明物体的图像（如下面两张图像），本方法也容易失败。其主要原因为训练集中类似的图像数量较少。

来仍为一大难点。解决该问题的主要方法与上一点类似，也即通过增加数据量或者采用更多的数据增强方法。

除了以上讨论的方法，在卷积神经网络中引入超像素级别的信息也可作为解决以上问题的一种方法。超像素可以有效地将同质区域分割出来，从而可以帮助解决显著性物体不完整的情形。

第五节 本章小结

本章主要介绍了基于短连接以及深度监督的显著性物体检测算法。本方法在著名的 HED 边缘检测架构的基础上，并非将损失层直接接到每个阶段引出的侧向路径中，而是通过引入一系列的短连接来将不同的侧向路径紧密地结合起来。有了这些短连接之后，高层侧向路径提取的高层级语义特征可以被传递到低层侧向路径中，因而低层侧向路径提取的边缘信息可以帮助高层级语义特征从而生成质量较高的显著性图。除此之外，本章改进了现有的基于全连接的条件随机场使其更适用于显著性检测领域。实验结果表明本章提出的方法可以较为精准地定位到显著性物体的位置并且分割出的区域的边缘信息也得以保留。本方法的实验结果与现有的显著性物体检测算法相比具有更多优势。除此以外，本章也对模型不同的组成部分的重要性进行了评估并对模型的不足之处进行了简要的分析。

第四章　基于视觉注意力模型的弱监督语义分割

近年来，由于基于对抗擦除策略的注意力模型在定位语义物体可辨别区域的能力，其已被深入研究。然而，该策略的主要问题之一在于随着模型不停地迭代，检测出的可辨别区域也将持续扩散到非语义区域（背景区域）中。这一缺陷使得基于该策略的得到的类别激活图像的质量严重下降。

为了解决上述问题以及更好地挖掘更高质量的可辨别区域，本章提出了一种高效地基于自擦除策略的注意力模型（self-erasing network，SeeNet）来阻止可辨别区域无端地扩散到背景区域中。SeeNet 采用了两种不同的自擦除策略来促使注意力模型能够利用已挖掘的可靠的可辨别区域以及背景先验来获取更多高质量的区域。这种方式可以使得语义物体的完整区域被尽可能地挖掘出来。为了测试 SeeNet 生成的可辨别区域的质量，本章将其应用到弱监督语义分割任务中。实验结果表明 SeeNet 生成的伪标注在被用于语义分割模型时可在 Pascal VOC 2012 分割数据集上取得良好的效果。

本章的主要结构如下：第一节主要介绍注意力模型的背景知识；第二节主要对本章提出的自擦除策略进行详细介绍；第三节介绍如何将 SeeNet 生成的类别激活图像应用到弱监督语义分割任务中；第四节 对 SeeNet 生成的类别激活图像以及语义分割结果进行分析并将其与现有方法进行比较；第五节对本章工作进行总结。

第一节　引言

4.1.1　背景知识

语义分割的主要目的在于给定一个任意输入图像对其中的每一个像素进行分类。对于全监督的语义分割模型而言[34, 52–55]，对大规模训练数据的需求大大限制了其通用性[73]。对于已标注的数据而言，其多样性对于训练模型有着重大作用。对于新的类别而言，由于模型的训练需要大量像素级表的标注因而需要大量的人力参与。为了解决这一难题，研究人员开始逐步探索使用较弱的监督数据用来训练网络，包括线条[113]、物体的边界框[65]、以及精准的像素点[114]（随机从每个语义物体上采一个点）等。

图 4.1 选自 PASCAL VOC 2012 数据集中的样例。每张图下对应着图内语义物体的类别标签。

虽然这些相对较弱的监督数据仍可以使得分割模型取得较好的效果，但其仍需要一定量的人工参与。除此之外，研究人员也开始采用图像类别标签作为监督数据[7, 8, 64, 71, 72]。类别标签数据仅需要对每张图像内语义物体的类别做出标注，大大减少了人力的参与。图4.1列出了来自 PASCAL VOC 2012 数据集中的样例。可以看出，每张图像的标注数据仅限于图像内存在的语义物体的类别而非任何像素级别的标注。因而，对于分割新的类别需求时，标注者仅需给出每张图像里对应的语义物体的类别标签。这使得每张图像的标准工作可在数秒内完成。

4.1.2 研究动机

本章的研究内容为如何在检测到更多语义物体区域的同时尽可能将可辨别区域集中到语义物体上而非背景上。为了实现这一目的，本章将以对抗擦除的概念为基础进而引入自擦除的概念。

上面提到，基于对抗擦除的方案在模型持续迭代的过程中会检测到越来越多语义不相关的物体。如图4.2（b-e）所示，随着对抗擦除模型不停地迭代，越来越多非语义区域被误判为语义区域。引起该现象的主要原因在于随着越来越

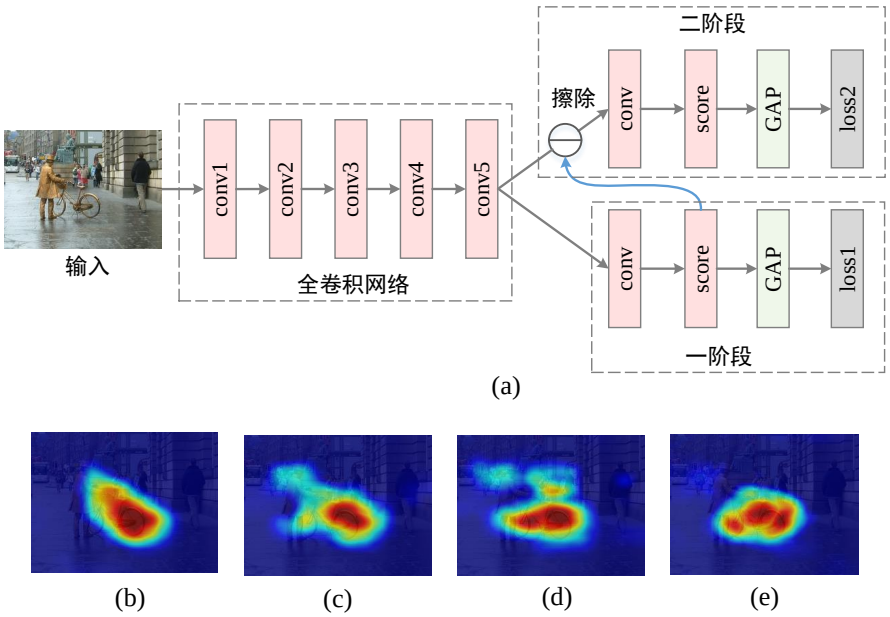


图 4.2 （a）一种典型的基于对抗擦除策略的注意力模型[51]。该模型主要由三部分组成：第一部分为主干网络，也即图中所标注的全卷积网络；第二部分为初始可辨别区域生成器（第一阶段）；第三部分为擦除来自主干网络的特征映射后的可辨别区域生成器。（b-d）在训练该模型不同时期得到的类别激活图像（自行车）。（e）本方法生成的类别激活图像。其中“conv”表示卷积层、“score”表示类别相关的卷积层、“GAP”表示全局均值池化层、“loss”为损失层（交叉熵损失）。

多的语义区域被擦除，损失层仍传递给模型擦除后的图像中仍有语义物体的信号。这使得模型仍需在背景中寻找带有语义的区域。图中，当自行车对应的区域被擦除干净后，分类器将无法从图像的剩余区域内找到自行车对应的区域，因而会把自行车区域周边的背景区域判断成自行车，从而引入噪声。再例如，火车通常在铁轨上行驶。然而，当火车对应的区域被擦除掉后，铁轨部分则很容易被判断成火车从而继续被检测出来。

根据以上讨论，如何在训练的过程中限制语义物体的恶意扩散可以有效改进基于对抗擦除的模型。为此，本章采用了自擦除的策略来改进现有的基于对抗擦除的模型。由于在检测过程中大部分非语义物体区域仍被判成背景区域，因而可以利用不同图像背景之间的相似性使已检测到的具有高置信度的背景区域作为先验作用到注意力模型中。

4.1.3 研究内容概要

本章将提出一种能够有效地解决基于对抗擦除的注意力模型在迭代过程中可辨别区域不停地扩散的方案。基于对抗擦除模型的主要弊端在于其迭代过程

中并未对图像中的背景区域加以标注。其后果即为可辨别区域不停地向背景区域扩散。因而，如果能够提供给分类模型有关背景区域的先验，则可以有效地抑制可辨别区域的随意扩散。本章将从该角度出发，提出自擦除网络的概念来检测尽可能完整的语义物体。

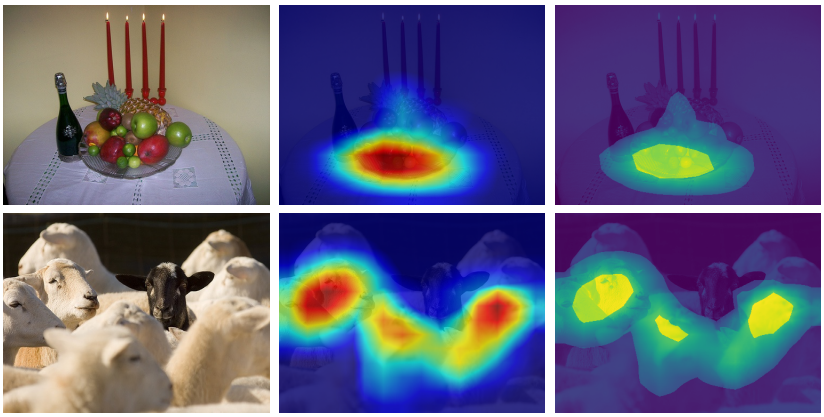
在注意力模型的训练过程中，除了可以检测到语义区域（可辨别区域）外，背景区域也可通过将可辨别区域擦除掉而得到。由于真实场景中的背景大多具有较强的相似性，因而可以借助不同图像中背景之间的相似性对每张图像提取出其大致的背景区域。如图4.2所示，当忽略掉已检测出来的可辨别区域后（红色区域），图像中剩余部分已能够较为准确地表示背景区域。通过将提取的背景区域显式地传递到注意力模型中则可以在一定程度上限制可辨别区域的无端扩散。

根据以上发现，本研究将图像根据已得到的初始类别激活图像（第一阶段）分为 3 个部分：前景部分，待检测部分以及背景部分。前景部分对应着激活值比较高的部分。与对抗擦除模型类似，本方法在模型第二阶段将前景区域擦除掉。对于背景部分而言，为了防止可辨别区域扩散到该部分，本方法将采用一种自擦除的策略显式地将背景部分特征值的符号反转从而抑制可辨别区域的扩散。待检测部分即为激活值较小且未被模型判成前景的部分。注意力模型仅被允许在该部分挖掘更多的可辨别区域而非其它区域，因而可辨别区域的随意扩散可以得到有效的控制。

为了进一步测试本模型生成类别激活图像的质量，本章将得到的类别激活图像应用到近年来比较流行的基于弱监督的语义分割任务中。通过将类别激活图像与上一章中的显著性图进行有效的结合可以生成高质量的伪标注数据。同时，本章将以弱监督语义分割的结果为准来对比自擦除策略的模型较对抗擦除策略的优势。

第二节　自擦除网络

本小节将详细介绍基于自擦除策略的注意力网络（**Se**lf-**E**rasing **N**etwork，SeeNet）。由于本方法的提出是受人类视觉系统的启发，因而在对自擦除策略的整体流程进行详细说明之前，本小节首先将对人类视觉系统中的注意力机制进行简要介绍。



（a）原始图像　　（b）类别激活图像　　（c）三值模板

图 4.3　本方法生成三值模板的过程。（a）原始图像；（b）由注意力模型第一阶段产生的类别激活图像；（c）阈值切割后的三值模板。给定初始类别激活图像后，首先通过设置两个固定的阈值将其分成 3 个不相交的部分。图（c）中的黄色部分对应着激活值比较大的部分。颜色较暗区域对应着背景区域。在训练过程中，该部分将会被用作背景先验来抑制可辨别区域的扩散。中间区域即为大概率含有语义物体的部分。需要注意的是，图（c）实际上为一三值模板。

4.2.1　视觉系统中注意力机制的工作原理

如第一节中所述，随着网络训练次数的增加，基于对抗擦除策略的模型会检测到越来越多非语义区域。因此，训练该类模型的一大难点在于何时终结模型的训练。当处于欠拟合状态时，模型的定位能力将有所减弱。相反，当模型处于过拟合状态时，得到的可辨别区域通常会扩散到背景部分。图4.2（b-e）给出了一些样例。“自行车”类别对应的可辨别区域通常会扩散到周围的背景甚至覆盖了类别“人”对应的区域。事实上，我们人类在观察某一地方的同时总会有意识地忽略掉并不感兴趣的区域[115]。当关注一个较大的物体时，人类通常会先观测该物体中比较显著的部分然后慢慢将注意力移动到其它区域。在这个过程中，人类总是能够无意识且成功地忽略掉背景的影响。因而，如何使注意力模型具有这一特性十分必要。

然而，对于注意力网络来说，它们自己本身并不具备这样的能力。因此，如何显式地把背景引入到注意力模型中将会大大抑制可辨别区域的随意扩散。受人类认知过程的启发，如现有工作[8, 50, 51] 中所述，除了简单地擦除掉已检测到的初始可辨别区域外，本方法将非可辨别区域当作背景来达到抑制的功能。下面将对本模型进行详细介绍。

4.2.2 自擦除的概念

为了突显出语义区域且防止被检测到的可辨别区域扩散到背景部分，本小节在训练阶段引入了自擦除的概念。给定一初始的类别激活图像（由图4.4中 S_A 分支生成），根据功能首先将输入图像划分成三个不相交的部分：内部注意力部分、外部背景部分以及中间的待检测部分。具体划分结果可参见图4.3c 中的三值模板。通过引入背景先验，本章提出的自擦除模型的目的在于驱使注意力模型达到一种自擦除的状态。在这种状态下，可观测的部分能够被限制在非背景部分，从而避免可辨别区域随意地扩散到背景区域同时使已达到完美状态的可辨别区域保持不变。为了实现这一目标，两个关键问题需要首先被解决：

- 给定图像的类别标签，如何定义以及获取背景部分。
- 如何将自擦除的思想引入到注意力模型中使其得到的类别激活图像能够尽量保持完整。

4.2.2.1 背景先验

考虑到注意力模型受弱监督的限制，获取精准的背景部分十分困难。因此，本章退而求其次采用一种阈值分割的方法来获取较为精准的背景先验。给定一由 S_A 生成的初始类别激活图像 M_A，除了设置一个固定的阈值 δ 来得到一个二值模板外[51]，本章同时考虑使用另一小于 δ 的阈值来得到一个三值模板 T_A。为了表达方便，这里使用 δ_h 以及 δ_l（$\delta_h > \delta_l$）来表示这两个阈值。激活值小于 δ_l 的区域将被看作背景部分。因此，三值模板 T_A 可以被定义为

$$T_{A,(i,j)} = 0, \quad \text{当} M_{A,(i,j)} \geq \delta_h \tag{4.1}$$

$$T_{A,(i,j)} = -1, \quad \text{当} M_{A,(i,j)} < \delta_l \tag{4.2}$$

$$T_{A,(i,j)} = 1, \quad \text{当} \delta_l \leq M_{A,(i,j)} < \delta_h \tag{4.3}$$

以上公式表明，背景部分仅对应着 T_A 中值为-1 的区域。由实验可以发现，由以上方法得到的背影区域对于绝大多数输入场景来说包含了其中大部分的真实背景。其主要原因在于由 S_A 分支生成的可辨别区域已经可以较为精准地定位到大部分语义区域。

4.2.2.2 条件反转线性单元（Conditionally Reversed Linear Units，C-ReLUs）

根据以上得到的背景先验，可以很容易地通过反转特征映射中背景区域对应的激活值的符号来实现自擦除策略。这一策略可以使得待检测区域的激活值

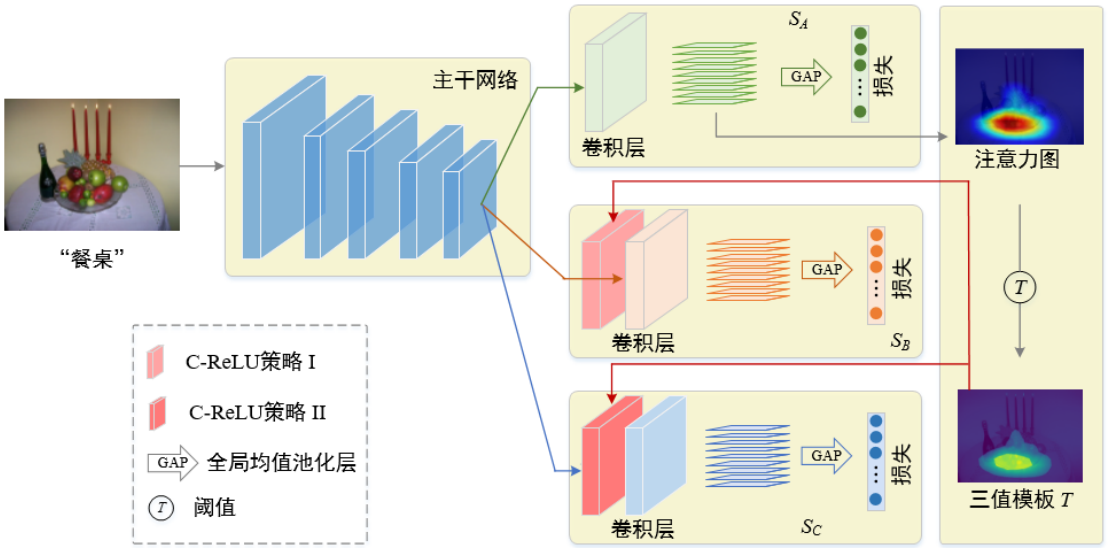


图 4.4 本模型的流程图。可以看出，除了最上层的初始类别激活图像生成器，本模型还包含两个额外的分支，分别实现不同的自擦除策略来抑制可辨别区域的随意扩散。

皆为正值，因而可以更容易地从中挖掘可辨别区域。为了实现这一目标，本部分扩展了传统的线性整流函数（ReLU）[103]，使其具有更为一般的形式。线性整流函数可以被简单定义为

$$\mathrm{ReLU}(x) = \max(0, x) \tag{4.4}$$

其主要作用是消除未激活的（负值）单元对下一层的影响且达到非线性变换的目的。更一般地，条件反转线性单元函数将一个二值模板参与到计算中，其公式为

$$\text{C-ReLU}(x) = \max(x, 0) \times B(x) \tag{4.5}$$

其中 B 为一取值为 $\{-1, 1\}$ 的二值模板。与线性整流函数不同的是，线性整流函数的输出均为非负值，而条件反转线性单元函数在其基础上将根据输入的二值模板有条件地反转激活值的符号。反转背景区域对应激活值的符号的目的在于使注意力模型更为关注含有正激活值的区域而忽略掉含有负激活值的区域。这一功能将促使注意力模型从待检测区域中发掘更多语义物体而非背景区域，从而达到抑制可辨别区域随意扩散的目的。

4.2.3 自擦除网络

图4.4给出了本模型的大体流程图。由图4.4可以看出，本模型主要包含 3 个分支。为了便于阐述，本小节将用 S_A、S_B 以及 S_C 来分别表示 3 个不同的分支。

这三个分支共享同一个主干网络。主干网络的选择可根据不同的需求而定。与Zhang 等所用方法[51] 相同，本模型中的 S_A 分支采用了与其相似的结构来生成初始的类别激活图像。S_B 和 S_C 分支与 S_A 分支类似，但其输入首先被送入到C-ReLU 函数中来实现自擦除的目的。

4.2.3.1 自擦除策略 I

在 S_B 分支中，本模型首先引入第一个自擦除策略。给定由 S_A 分支生成的类别激活图像 M_A，可以根据4.2.2中所述方法得到一三值模板 T_A。当把 T_A 作为输入送到 S_B 分支的 C-ReLU 层中时，可以简单地调整 T_A 使其变为一二值模板。例如，可将所有的非负值皆设为 1 来实现。

当考虑引入擦除策略时，可以将 C-ReLU 函数中的二值模板简单地扩展到三值情况。因此，公式4.5可被重新表示为

$$\text{C-ReLU}(x) = \max(x,0) \times T_A(x) \tag{4.6}$$

公式4.6的示意图可参见图4.3c。图中的黄色区域对应着由 S_A 分支产生的注意力部分。在主干网络的输出特征映射中，该部分对应的区域在训练过程中也将被擦除掉。背景部分对应的正激活值的符号将被取反从而可以更为突出待检测区域的重要性。在训练过程中，S_B 分支将会进入到一种自擦除状态。该状态将会阻止背景部分的非语义物体被检测到。与此同时，可以确保待检测区域的语义物体更容易被检测到。

4.2.3.2 自擦除策略 II

除了 S_B 分支，本模型加入了 S_C 分支来实现第二种自擦除策略。第二种策略的目的在于进一步避免背景区域中的非语义物体被检测出来。具体来说，三值模板 T_A 首先被转换为一二值模板。具体方法为设置背景区域对应的值为 1，而其他区域对应的值为 0。以这种方式，C-ReLU 的输出中背景区域对应的部分将仅有非 0 的激活值。在训练过程中，通过设置背景区域属于任何语义类别的标注数据的概率为 0 来达到自擦除的目的。由于不同场景中的背景都共有一定的相似性，S_C 分支可以进一步帮助本模型矫正一些背景区域中的物体被判为语义区域的情况，从而避免了错误的可辨别区域的生成。

4.2.3.3 损失函数

根据以上定义，本模型总的损失函数可以定义为

$$\mathcal{L} = \mathcal{L}_{S_A} + \mathcal{L}_{S_B} + \mathcal{L}_{S_C} \tag{4.7}$$

所有分支中的多类别分类任务可以被看作 M 个独立的二值分类问题。其中 M 为语义类别的个数。因此，每一个分支都可以采用交叉熵损失来达到分类的目的。给定一图像 I 以及其中的语义标签 $\mathbf{y}$，S_A 与 S_B 分支对应的标签向量可以表示为

$$\mathbf{l}_n = 1, \quad \text{当} n \in \mathbf{y} \tag{4.8}$$

$$\mathbf{l}_n = 0, \quad \text{当} n \notin \mathbf{y} \tag{4.9}$$

其中 $|\mathbf{l}| = M$。S_C 分支中的标签向量为一零向量，即向量中的所有值皆为 0。其作用为暗示注意力模型背景区域中没有任何语义物体。

4.2.3.4 类别激活图像的生成

为了得到最终的类别激活图像，在测试阶段，S_C 分支将被忽略。其原因在于该分支的作用仅为抑制可辨别区域的随意扩散。令 M_B 表示 S_B 分支生成的类别激活图像。M_A 与 M_B 首先将被归一化到 $[0,1]$ 区间。令 $\hat{M}_A$ 与 $\hat{M}_B$ 分别表示其归一化后的结果。融合后的类别激活图像 M_F 可以被定义为

$$M_{F,i} = \max(\hat{M}_{M,i}, \hat{M}_{B,i}) \tag{4.10}$$

令 M_H 表示将输入图像水平翻转后送入注意力模型中得到的类别激活图像，则最终的类别激活图像 M_{final} 可以表示为

$$M_{final,i} = \max(M_{F,i}, M_{H,i}) \tag{4.11}$$

第三节 弱监督语义分割

为了测试本模型生成的类别激活图像的质量，本章将其应用到近年来比较流行的基于弱监督的语义分割任务。为了公平地与现有最好的相关工作进行比较，本章采用了一种将显著性图与类别激活图像相结合的方法[73]。与该方法不同的是，并非将擦除策略简单地应用到显著性物体的挖掘中，本章采用一个流

Algorithm 1: 训练语义分割模型 “伪标注” 的生成

Input : 含有 N 个像素的图像 I；类别标签 $\mathbf{y}$
Output: 伪标注 G

$Q = \text{zeros}(M+1, N)$, M 为语义类别的个数;
$D = \text{Saliency}(I)$; ⇐ 得到显著性图
for $i \in I$ **do**
  $A_{\mathbf{y}} = \text{SeeNet}(I, \mathbf{y})$; ⇐ 生成类别激活图像
  $Q(0, i) \leftarrow 1 - D(i)$; ⇐ 位置 i 为背景的概率
  for $c \in \mathbf{y}$ **do**
    $Q(c, i) \leftarrow \text{harm}(D(i), A_c(i))$; ⇐ 调和均值
  end
end
$G \leftarrow \text{argmax}_{l \in \{0, \mathbf{y}\}} Q$;

行的显著性物体检测模型[3] 来提取图像的背景先验[50]。通过设置一个固定的阈值，可以简单地将显著性图的前景与背景分离开。

具体来说，给定一张输入图像 I，首先将其显著性图归一化到 $[0,1]$，可得 D。令 $\mathbf{y}$ 表示图像 I 的类别标签，其取值空间为 $\{1,2,\ldots,M\}$。其中，M 为语义物体的类别数。令 A_c 为标签 $c \in \mathbf{y}$ 对应的类别激活图像。本方法采用的 “伪标注” 可根据算法1来得到。为了计算每个类别对应的种子区域，调和均值函数被用来计算图像中每个点 I_i 的类别为 c 的概率。具体公式如下：

$$\text{harm}(i) = \frac{w+1}{\left(w/(A_c(i)) + 1/D(i)\right)} \tag{4.12}$$

其中，参数 w 被用来控制类别激活图像的重要性。在本章实验中，w 的值被默认设为 1。

第四节 实验结果

为了验证本章提出的自擦除策略的有效性，本小节将本模型生成的类别激活图像应用到基于弱监督的语义分割任务中。通过根据算法1将得到的类别激活图像嵌入一简单的方法中，本方法得到的语义分割图可以取得较好效果。

4.4.1 实现细节

本小节主要介绍训练本模型需要的数据集以及在弱监督语义分割任务中所用的评测指标。

4.4.1.1 数据集与评测指标

本章采用流行的 PASCAL VOC 2012 图像分割数据集[1] 作为训练集以及测试集。该数据集共包含 20 个语义类别以及一个背景类别。与现有方法类似，本模型采用训练集中的 1400 多张图像以及后增加的 8000 多张图像[15] 来训练注意力模型以及语义分割模型。因而，训练集的总量为 10582 张。本小节将汇报本模型在验证集以及测试集上的结果。这两个数据集分别包含 1449 以及 1456 张图像。与之前工作类似，在评测语义分割结果阶段，本章采用平均交并比（mean intersection-over-union，mIoU）对生成的分割图进行评测。其计算公式为

$$\mathrm{IoU} = \frac{Res \cap Anno}{Res \cup Anno} \tag{4.13}$$

其中，Res 为预测的分割结果、$Anno$ 为图像对应的标注。

4.4.1.2 网络配置

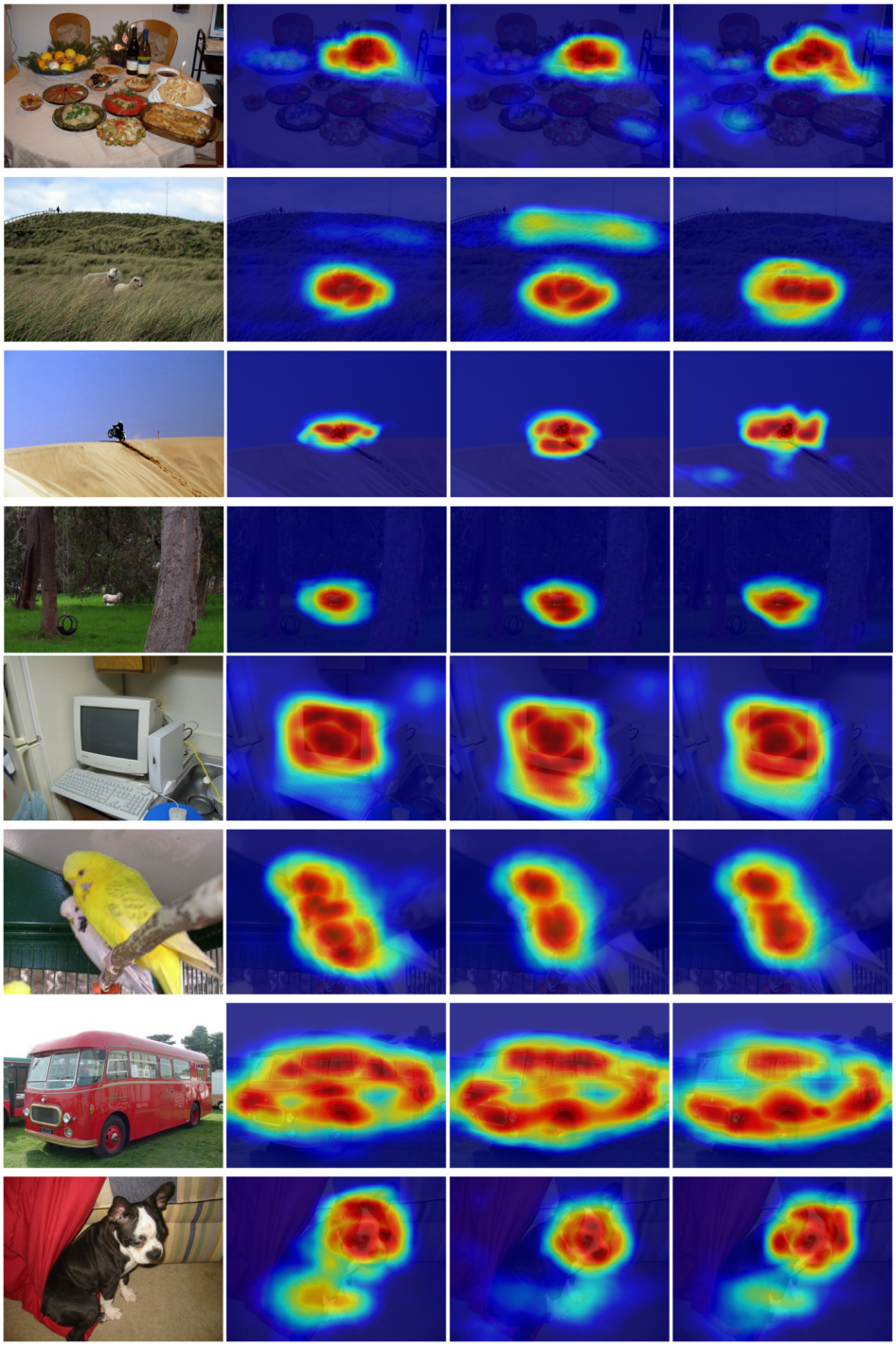
对于注意力模型而言，与现有工作[8, 51] 类似，本模型采用 16 层的 VGGNet[37] 作为基础网络。本模型首先去掉 VGGNet 中最后 3 个全连接层以及最后一个池化层，然后在剩余网络后面连接 3 个通道数为 512、卷积核大小为 3×3 的卷积层。由于 VOC 数据集中共有 20 个语义类别，因而网络的最后部分为一个 20 通道的卷积层以及一个全局均值池化层用来预测每个类别对应的概率。

在训练过程中，批量大小为 16、权重衰减因子为 0.0002、初始学习速率为 0.001。每迭代 15000 次后学习速率缩小 10 倍。网络总的迭代次数为 25000 次。在数据增强方面，本模型采用了与 ResNet[107] 相同的策略。在 S_B 分支中的阈值 δ_h 与 δ_l 分别被设置为类别激活图像中最大值的 0.7 倍与 0.05 倍。S_C 分支中的阈值被设为 $(\delta_h + \delta_l)/2$。

对于语义分割任务，为了公平地与现有方法进行比较，本章采用了标准的 DeepLab-LargeFOV 模型[52] 作为分割网络。该网络的预训练模型为在 ImageNet 数据集[116] 上预训练过的 VGGNet[37]。与部分现有方法相同[73]，本章同样汇报本方法基于 ResNet 版本的 DeepLab-LargeFOV 的结果。网络的参数以及条件随机场的参数可参见 Chen 等[52]。

4.4.1.3 测试阶段

在注意力模型的测试阶段，输入图像首先被线性插值到固定大小 224×224，在得到类别激活图像之后，再将其线性插值回原图大小。在分割任务中，与 Lin



(a) (b) (c) (d)

图 4.5 不同网络配置下生成的视觉效果图。（a）输入图像；（b）本方法生成的类别激活图像；（c）由 ACoL 方法[51] 生成的类别激活图像；（d）由配置 2 对应网络生成的类别激活图像。前四行图像中均包含尺寸较小的语义物体，而后四行图像则包含尺寸较大的语义物体。由该图可以看出，本方法可以更好地抑制可辨别区域的随意扩散并且可以生成完整性更好的语义物体。

表 4.1 与其它两个配置的模型在 PASCAL VOC 2012 验证集上的量化对比。该表中的分割结果由分割模型直接生成且无多尺度测试、条件随机场等后处理操作。

配置	训练集	监督类型	mIoU（验证集）
1（ACoL[51]）	10,582 VOC	弱监督	56.1%
2（C-ReLU 中无符号反转）	10,582 VOC	弱监督	55.8%
3（SeeNet）	10,582 VOC	弱监督	57.3%

等[55] 相似，本章采用了多个尺度的输入图像进行测试并将其相加的结果作为条件随机场的输入。条件随机场的介绍可参见 Chen 等[52]。

4.4.2 自擦除的优势

为了进一步体现本章提出的自擦除策略的重要性，本小节列出了一系列本模型在不同配置下的实验结果。除了列出本模型生成的结果图外，本小节重新复现了两个其它网络架构的结果作为对比。首先，本小节复现了 Zhang 等[51] 提出的基于对抗擦除的注意力模型（配置 1）—ACoL。在复现过程中，所有的超参数以及网络架构都与原文[51] 中的默认配置保持一致。该模型没有使用本章提出的 C-ReLU 层以及第三个分支 S_C。另外，为了体现出本章提出的条件符号翻转操作的有效性，第二种模型将背景区域对应的特征值皆设为 0 且保持其余部分不变（配置 2）。

4.4.2.1 类别激活图像的质量

图4.5给出了从 PASCAL VOC 2012 测试集中选出的样例以及当采用不同配置时得到的类别激活图像。如图4.5中前 4 行所示，当语义物体较小时，SeeNet 与其它两个模型相比能够更精准地覆盖语义物体对应区域。这主要是因为 SeeNet 中的 S_C 分支可以很好地限制可辨别区域的随意扩散，因而得到的类别激活图像并不会覆盖较多的背景区域。对于大物体而言（图4.5中的后 4 行），SeeNet 相对而言可以检测到整个语义物体，而其它两个模型生成的类别激活图像则不具有整体性。这一现象说明，条件符号翻转操作除了能够抑制类别激活图像的随意扩散外还能够使模型更关注语义区域，从而保证了 SeeNet 检测整个语义物体的能力。

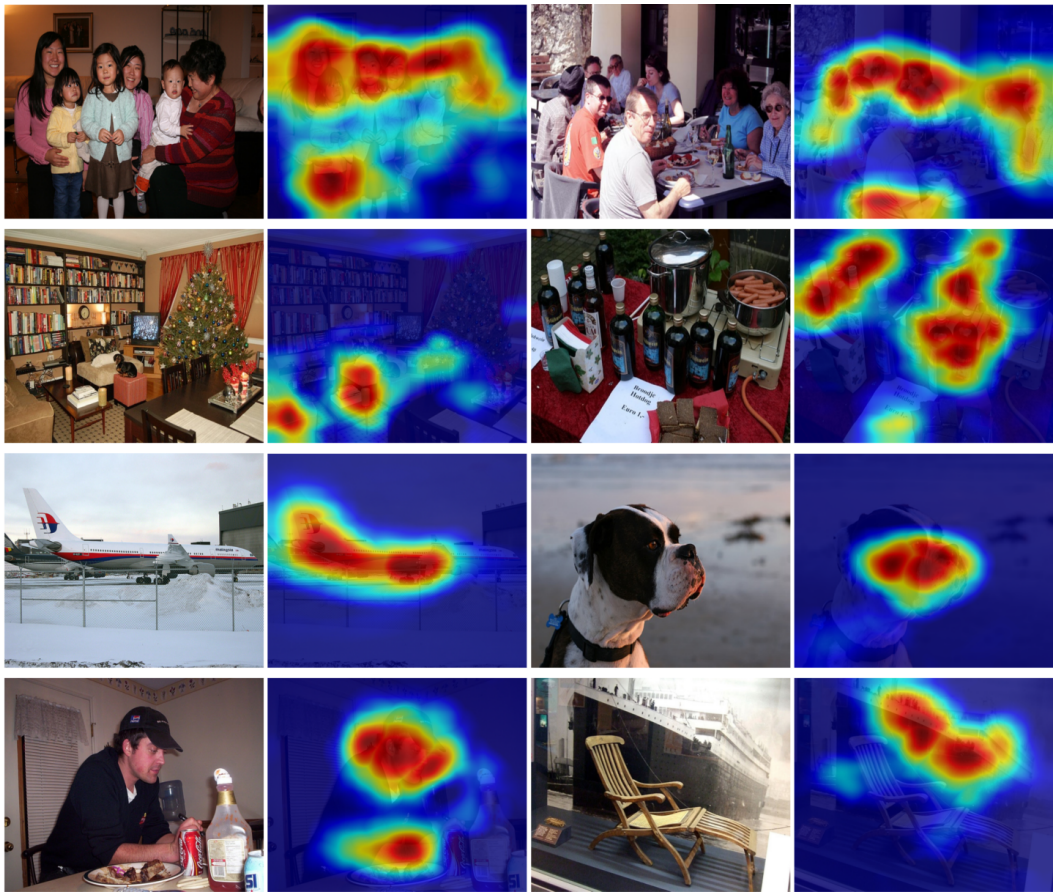


图 4.6 本模型生成的质量较差类别激活图像的案例。

4.4.2.2 量化结果对比

除了以上视觉效果对比，本部分将汇报将 SeeNet 生成的类别激活图像应用到弱监督语义分割任务时的量化结果以及与采用其它两种配置模型生成的分割结果的量化对比。在得到 SeeNet 生成的类别激活图像后，首先根据第三节所述方法生成训练集图像对应的伪标注，然后使用得到的伪标注数据来训练语义分割网络。

本模型在 VOC 验证集上的量化结果可参见表4.1。值得一提的是，本部分所有的分割结果都基于单尺度的测试并且无任何后处理工具参与。由表4.1可以观察到，在采用相同的显著性图的情况下，SeeNet 生成的类别激活图像有助于分割模型生成质量更高的伪标注。与张等[51] 生成的类别激活图像相比，SeeNet 生成的类别激活图像导出的伪标注使得最终语义分割结果有 1.2 个百分点的提升。这也反映出本模型生成的类别激活图像具有较高的质量。

表 4.2 与现有方法在 VOC 验证集以及测试集上的结果对比。其中“weak”表示该模型仅使用了图像的类别标签来训练网络；“sal”以及“bbox”分别表示该模型使用了显著性图以及物体外矩形框用来训练网络。当无特殊声明时，该表中所列出的方法均使用 VGGNet[37] 作为预训练网络。CRF 表示条件随机场。

方法	出版物	监督类型	mIoU (验证集)		mIoU (测试集)
			无 CRF	有 CRF	有 CRF
CCNN[64]	ICCV'15	10K	33.3%	35.3%	-
EM-Adapt[60]	ICCV'15	10K weak	-	38.2%	39.6%
MIL[61]	CVPR'15	700K weak	42.0%	-	-
DCSM[117]	ECCV'16	10K weak	-	44.1%	45.1%
SEC[7]	ECCV'16	10K weak	44.3%	50.7%	51.7%
AugFeed[65]	ECCV'16	10K weak + bbox	50.4%	54.3%	55.5%
STC[71]	PAMI'16	10K weak + sal	-	49.8%	51.2%
Roy et al.[66]	CVPR'17	10K weak	-	52.8%	53.7%
Oh et al.[67]	CVPR'17	10K weak + sal	51.2%	55.7%	56.7%
AE-PSL[8]	CVPR'17	10K weak + sal	-	55.0%	55.7%
Hong et al.[78]	CVPR'17	10K + video weak	-	58.1%	58.7%
WebS-i2[118]	CVPR'17	19K weak	-	53.4%	55.3%
DCSP-VGG16[73]	BMVC'17	10K weak + sal	56.5%	58.6%	59.2%
DCSP-ResNet101[73]	BMVC'17	10K weak + sal	59.5%	60.8%	61.9%
TPL[119]	ICCV'17	10K weak		53.1%	53.8%
GAIN[51]	CVPR'18	10K weak + sal	-	55.3%	56.8%
SeeNet (VGG16)	-	10K weak + sal	59.9%	61.1%	60.7%
SeeNet (ResNet101)	-	10K weak + sal	62.6%	63.1%	62.8%

4.4.3 与现有方法对比

本小节将本方法得到的分割结果与现有的弱监督语义分割的模型生成的结果进行对比。此处对比的所有方法皆基于类别标签级别监督。详细的数据对比可参见表4.2。该表除了汇报了每个方法在验证集上的结果外，也汇报了其在测试集上的结果。

从表4.2中可以看出，当所有模型均采用 VGGNet[37] 作为预训练模型时，本方法与现有的其它方法相比具有一定优势。特别地，当与 DCSP[73] 相比时（该方法利用了与本方法相似的方式生成伪标注数据用于训练语义分割网络），本方法得到的语义分割结果在验证集上比其高出超过两个百分点。该方法使用了原始的 CAM[4] 模型作为其类别激活图像生成器。这一现象表明，本方法生成的类别激活图像与原始 CAM 模型相比具有一定优势。当使用 ResNet 作为预训练模

型时，本方法得到的分割结果与 ResNet 版本的 DCSP 相比在验证集上有了 2.3 个百分点的提升。

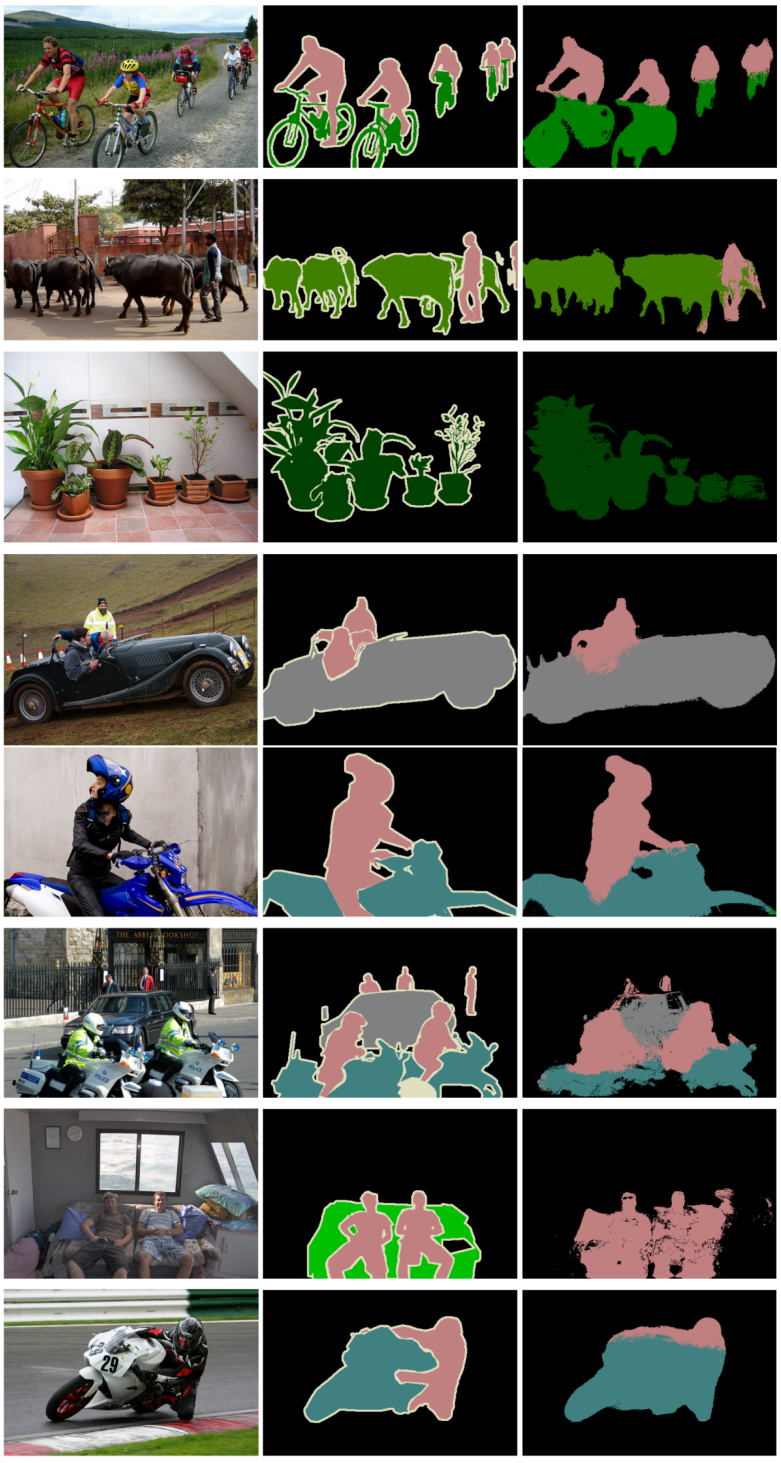
为了进一步比较本方法生成的类别激活图像与基于对抗擦除策略生成的类别激活图像，这里将本方法生成的分割结果与 AE-PSL[8] 以及 GAIN[50] 生成的分割结果进行对比。从表4.2中可以观察到，本方法得到的结果明显优于上述两个基于对抗擦除策略的方法。这也间接地说明了本方法生成的类别激活图像与另外两种方法相比具有更高的质量。

4.4.4 讨论

为了更好地理解本章提出的方法，图4.7给出了一些本模型生成的语义分割效果图。由图中可以看出，本方法能够较为精准地分割出语义物体的主要原因在于 SeeNet 生成的高质量类别激活图像。另外，尽管大部分分割结果质量都比较高，但仍有一少部分图像的分割结果不尽人意。在图4.7 底部列出了一些质量较差的分割结果。这些质量较差的分割结果主要是由于带有不同类别的语义物体时常一同出现在某些图像中，从而使得注意力模型很难分开它们。例如，桌子和椅子两个类别通常会一同出现，因而注意力模型很难区分开哪些物体是桌子，哪些物体是椅子。

图4.6中给出了更多结果用来分析本方法的利弊。由该图中结果可以看出，一些含有较为复杂或前景背景对比度较低的场景是导致分割结果差的主要原因。尽管本方法可以将一些背景信息作为先验送入注意力模型中，但当处理这些场景较为复杂的图像时，本模型生成的结果也会出现定位错误。除此之外，当处理含有较多语义物体的场景时，本模型也容易定位失败。具体结果可参见图4.6。

对于注意力模型而言，找到图像中存在的简单类别通常较为容易，而检测出所有语义物体的位置及其完整的区域是极为困难的。一个可行的方案为在训练的过程中引入少量带有像素级别标注的图像。这些像素级精度的标注数据将帮助注意力网络矫正错误的定位并且改善检测到的语义物体的边缘信息，使得输出的类别激活图像的形状更为完整。在未来的工作中，作者将针对这一方案进行进一步探究。



(a) (b) (c)

图 4.7 本方法生成的语义分割结果图。（a）原始图像；（b）像素级标注数据；（c）本方法生成的语义分割结果图。除了累出质量较好的结果外（前 6 行），本图中也列出了一些质量较低的分割结果（后两行）用来帮助读者更好地理解本模型的优缺点。

第五节 本章小结

本章引入了自擦除的概念到注意力网络中。通过将初始类别激活图像生成器得到的结果分成 3 个区域可以得到较为精准的背景信息。给定这些背景信息，本章设计了两种不同的自擦除策略。其主要思想为抑制类别激活图像的随意扩散从而保证检测到的语义区域的完整性。基于以上两种不同的自擦除策略，本章设计了一种新颖的自擦除网络。该自擦除网络可以有效地限制可辨别区域的扩散，并且可以实现模型端到端的模型训练。

为了验证上述自擦除网络生成类别激活图像的质量，本章将得到的类别激活图像应用到弱监督语义分割任务中。通过将其简单地与已有的显著性图进行融合可以生成用于训练语义分割的伪标注数据。实验结果表明，本方法得到的分割结果明显优于现有的弱监督语义分割模型生成的结果。这一现象反映出本章提出的自擦除网络生成的类别激活图像具有较高的质量。

第五章 语义分割自主学习

收集大量手工标注的训练数据是非常耗时的。对于语义分割任务而言，由于其需要像素级精确的标注数据，收集的任务则更为艰巨。本章将着重解决基于网络大数据的语义分割任务。给定一个待分割的关键词集合，本章的目标在于利用仅有的关键词信息，从网络中爬取大量相关图像并完成语义分割的学习工作。该任务的一大挑战在于从网络中爬取的数据可能含有标签噪声（图像中实际存在的语义类别与关键词不匹配）。为了解决这一难题，本章采用一种特征拼接策略从网络图像中提取精准的种子区域来挖掘更多的语义区域。同时，本章设计一种噪声擦除网络。该网络从上述种子区域中学习语义知识并擦除图像中含有标签噪声的区域。给定以上工具，本章提出一种新颖的框架来实现语义分割自主学习。实验结果表明，本章所提方法在 PASCAL VOC 2012 验证数据集上可以实现 62.0% 的 mIoU 值。当将 PASCAL VOC 2012 中的图像以及类别信息参与语义分割网络的训练时，本章所提方法可以在弱监督语义分割中实现 66.1% 的 mIoU 值。

本章的主要结构如下：第一节主要介绍基于弱监督以及网络监督的背景知识；第二节提出语义分割自主学习的问题定义；第三节引入特征拼接技术以及噪声擦除网络并详细地介绍其工作原理；第四节验证本章提出的特征拼接技术的有效性，同时将噪声擦除网络生成的伪标注数据的效果进行分析。第五节对本章工作进行总结。

第一节 引言

近年来，基于全监督的语义分割算法在大规模的数据集上（例如 PASCAL VOC[1]，Cityscapes[76]，以及 MS COCO[77] 等）已经取得了重大进展[120]。然而，这些方法的一大缺陷在于所有方法皆需要像素级精度的标注数据来训练模型。由于真实的场景较为复杂且多变，收集大量数据且进行人工标注耗时又费力。为了解放大量的人工参与，研究人员逐步将语义分割的研究重点从全监督模型放到半监督或弱监督模型中。其中，较为典型的方法大多利用了物体外矩形框[60, 65]、一条曲线[113]、关键点[114] 以及类别标签[7, 8, 71–73, 118] 作为监督来训练

图 5.1 典型的选自网络的图像。蓝色的类别名称为索引图像所使用的关键词，而红色的类别名称为待解决的类别噪声。在训练过程中，仅蓝色关键词已知。根据统计数据，在所有网络图像中，大约百分之十五的图像含有类别噪声。

语义分割模型。

5.1.1 研究动机

当处理真实世界中的场景时，新的类别将会不断加入到训练中。为了解决标注数据严重缺少这一关键问题，利用图像类别标签作为监督可以加快工程的进行速度。在利用图像类别标签作为监督的过程中，虽然需要标注者仅对每一场景中相关的语义类别进行标注，但标注一张图像仍需要 20 秒左右的时间[79, 114]。当为每一个类别收集数千张以上的训练图像时，此种方法也需要一定量的人工参与。当处理数百个新的类别时，此种方法则需要标准数百万张图像的类别信息[121]。

针对以上难题，一个待解决的关键问题为：怎样使一个机器模型能够从互联网的大数据中自动学习语义分割相关知识且不依赖于任何人工标注。该问题与目前基于网络监督的弱监督语义分割任务[78, 79, 118] 较为相似但也存在以下两点不同之处：

- 首先，现有的基于网络数据的弱监督语义分割任务[78, 79, 118] 使用了大量已被标注的数据（例如 PASCAL VOC 数据集[1] 中的类别标签）来训练一个初始的弱监督网络，然后用已训练好的弱监督网络来过滤掉一些含有错误标签的网络数据，以防其在训练语义分割模型中起反作用。然而，当处理新的数据类别时，这些方法通常会失效因为并没有任何已标注好类别标签

的数据集可供其使用。

- 其次，这些方法[78, 79, 118] 的主要做法为从网络中选取较为简单的图像进行学习。这些简单的图像通常包含简单的背景以及单个类别的语义物体。当处理较为复杂的场景时，由于缺少较为困难的图像（包含复杂背景以及多个类别的语义物体）参与训练，这些方法通常会得到较差的结果。

类别噪声的存在使得从网络数据中学习语义分割成为一大难题。总的来说，以 PASCAL VOC 数据集为例，网络图像中的类别噪声可以大致分为两种：

- 第一种为网络图像的关键词与图像中的语义物体不一致。例如，检索关键词为“马”但检索到的图像中仅含有类别“牛”。
- 第二种为网络图像包含了更多类别的语义物体。例如，检索关键词为“马”但检索到的图像中含有“马”以及“人”。

现有的弱监督语义分割算法[8, 72, 73, 78, 79, 118] 通常检索较为简单的图像（仅包含一类语义物体）以避免以上情况。然而，当处理图5.1最后一行所示的场景时，这些方法的有效性并不能够得到保证。因而，如何解决以上两个问题来实现真正的语义分割自主学习具有重大研究意义。

5.1.2 语义分割自主学习框架

近年来，基于全监督的语义分割技术[34, 52–59] 已经取得了重大进展。与此同时，基于不同视觉注意力机制的弱监督语义分割算法也取得了长足的发展。但基于网络数据的语义分割自主学习技术尚在起步阶段。

本章提出了一种解决语义分割自主学习的学习框架并将从一个全新角度来解决网络图像中附带的类别噪声问题。其主要思想为：给定一些网络图像后，将其中带有噪声的区域（也即与图像关键词不相关的语义区域）擦除掉而非直接忽略掉所有带有类别噪声的图像。为了实现这一方案，本章提出噪声擦除网络（Noise Erasing Network）模型来过滤掉所有不相关的噪声区域。其基本思想为从已挖掘的可辨别区域中学习语义知识来推断其它前景区域的类别。如果推断的类别与图像的关键词不同，则将相应的区域标记为噪声区域并将其擦除掉。如图5.2所示，噪声擦除网络将从深绿色对应的可辨别区域（可由注意力模型得到）学习语义知识，并对浅绿色对应的前景区域进行推断。其最终目的为擦除掉有红色十字标识的噪声区域。

由于网络图像的质量不尽相同以及其多样性，直接使用现有的注意力模型[4] 并不能够直接提取出精度较高的可辨别区域。其主要原因为受类别噪声的

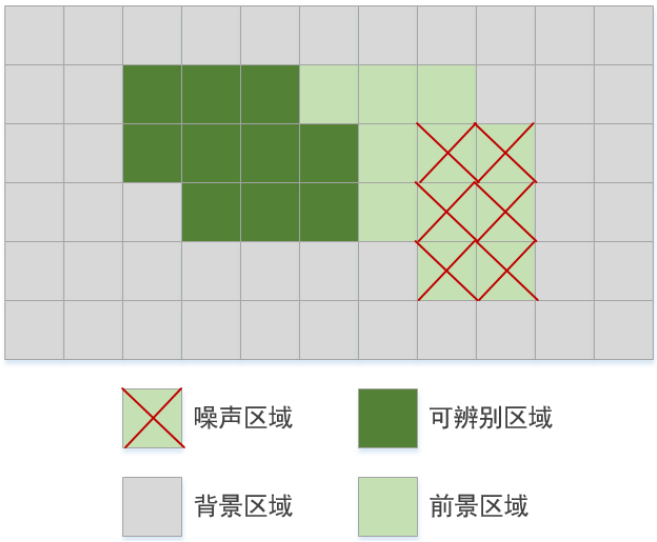


图 5.2 本章提出的噪声擦除网络的主要思想。噪声擦除网络将从深绿色对应的可辨别区域（可由注意力模型得到）学习语义知识，并对浅绿色对应的前景区域进行推断。其最终目的为擦除掉有红色十字标识的噪声区域。

干扰，一些图像中并不含有关键词应对的物体或含有多个不相关的语义物体。为了解决这一问题，本章提出一种特征拼接策略。该策略可以弥补图像中语义物体与关键词不匹配这一问题，使得提取的可辨别区域更为精准。

为了验证本章提出的语义分割自主学习框架的有效性，PASCAL VOC 2012 分割数据集[1] 将被用来作为测试集。给定 20 个来自 PASCAL VOC 数据集中的关键字，本系统首先为每个语义类别从互联网中检索 2000 张图像并采用本章提出的语义分割自主学习框架进行学习且无任何人工参与。实验结果表明本系统在仅有网络监督的条件下可以在测试集中实现 62.0% 的 mIoU 值。为了进一步与现有的弱监督语义分割方法进行对比，本系统也将 PASCAL VOC 数据集中的类别标签作为监督数据。将精准的类别标签参与训练进一步提升了本系统的分割效果。

第二节 问题定义

目前为止，获取训练数据最简单且有效的方法即为从互联网中检索。给定一组关键词，通过设定一系列的过滤信息（如图像的亮度、饱和度等）可以将较为复杂的图像过滤掉。该过程并不需要任何人工干预。给定检索到的网络图像，通过将每张图像的关键词与低层级任务的结果图（如显著性图与过分割图

等）相结合，可以得到用于训练语义分割模型的伪标注数据。与现有的弱监督语义分割方法不同的是，现有的方法大多依赖于具有精准类别标签的数据，而本章即将解决的问题为如何在仅有关键词信息的基础上解决语义分割任务。与弱监督方法相比，其难点在于图像中不相关的语义物体的干扰[122]将直接影响伪标注数据的质量（如图5.1所示）。由于每张图像仅有一个索引关键词，因而不相关的语义物体区域很容易被判成关键词所属类别（如图5.9c 所示）。

令 $\mathcal{I}=\{I_i\}_{i=1}^{N}$ 为一从互联网中索引的图像集合，其中每张图像对应一关键词。令 $\mathcal{L}=\{l_1,l_2,\ldots,l_L\}$ 为所有关键词的集合，其中 $L=|\mathcal{L}|$ 为所有关键字类别的数量。为了区别已有关键词，本章用 l_0 来表示背景类别。每张来自 $\mathcal{I}$ 集合的图像 I_i 将对应一个伪标签 y_i（也即其关键字）以及其真实的标签集合 $\hat{\mathbf{y}}_i$（包含该图像中所有的语义类别标签）。由于类别噪声的干扰，标签 y_i 通常并不等价于真实的标签集合 $\hat{\mathbf{y}}_i$。此时，y_i 与 $\hat{\mathbf{y}}_i$ 中不同的标签被定义为类别噪声。本章的主要目的在于使用 $\{I_i\}$ 以及 $\{y_i\}$ 来达到语义分割自主学习的目的。

第三节 噪声擦除网络

由于网络图像的多样性，少部分图像通常会带有类别噪声，使得从其中提取有用信息变得格外困难。如何消除类别噪声的影响并产生高质量的伪标注数据对于语义分割来说尤为重要。本节将提出一种有效的去除噪声干扰的方法。

5.3.1 工作流程概述

本章所用方法的流程可参见图5.3。其大概可以分为三个模块：一个可辨别区域检测模块、一个噪声去除模块以及一个语义分割模块。

注意力模型近年来已被广泛应用到弱监督语义分割领域。给定训练图像及其对应的关键词，注意力模型可以检测到每个关键词对应的可辨别区域。但注意力模型的缺陷之一在于其得到的可辨别区域的形状并不能够得到保证。因而，这些可辨别区域很难被直接用作伪标注来训练语义分割网络。与多数现有工作[71, 73, 123]类似，本章将可辨别区域与显著性图相结合来生成形状较为规则的伪标注数据。

虽然显著性物体检测模型可以将网络图像的前景区域分割出来，但由于其中类别噪声的存在，得到的前景区域通常包含一定大小的背景物体以及与关键词不相关的语义物体。如图5.7所示，第一行图像对应的关键词为“餐桌”，经过

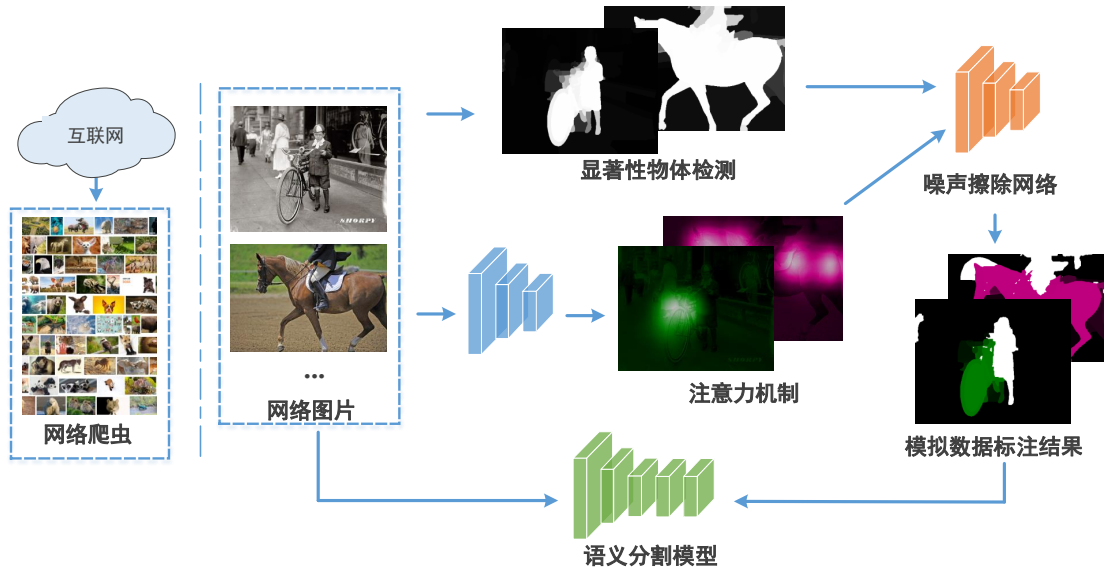


图 5.3 本方法的流程图。给定一些从网络检索的图像，首先注意力网络以及显著性物体检测模型被用来产生关键词类别对应的可辨别区域以及前景区域。得到的可辨别区域与前景区域被送入到噪声擦除网络中。该网络的作用是将与关键词相关的语义区域留住并去除与关键词不相关的区域。噪声擦除网络进而被用来生成伪标注数据，从而完成语义分割任务。

显著性物体检测后得到的显著性图中，“椅子”对应的区域也被分割出来。如果直接将前景区域的类别标注为“餐桌”则将引入大量的噪声区域。针对这一问题，本章提出如下方案：首先由注意力模型生成精度较高的可辨别区域（图5.7d中白色区域）；然后利用神经网络的泛化能力从这些精度较高的可辨别区域中学习语义知识，并对剩余的前景区域进行分类。如果某一前景区域的分类结果与图像对应的关键词不同，则将该区域标记为噪声区域（在学习语义分割模型的过程中，这部分噪声区域将被忽略掉）。剩余的前景区域则被保留。下面将针对流程图中的各个模块进行详细介绍。

5.3.2 可辨别区域挖掘

与多数现有弱监督语义分割方法[7, 8]类似，本章采用 CAM（Class Activation Mapping）[4] 作为注意力模型来挖掘高质量的可辨别区域。考虑到一些图像中会含有类别噪声，作者首先用所有的网络图像以及其对应的关键词训练了一个分类网络[107]，并用得到的分类模型对所有网络图像进行推断并丢弃每个类别中损失最大的 10% 的图像。

虽然经过上述过滤策略可以简单过滤掉一些带有类别噪声的图像，但剩余

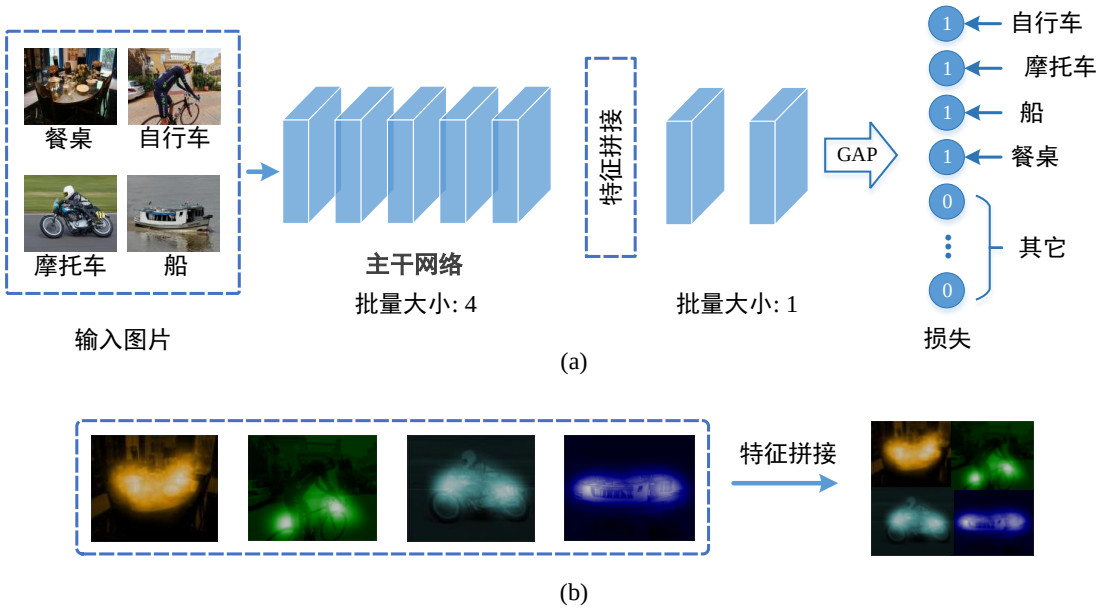


图 5.4 （a）本章提出的带有特征拼接技术的注意力模型。其中“GAP”为全局池化函数（Global Average Pooling）。在此图中，批量大小为 4 的输入图像被送入到网络中，经过特征拼接后，其批量大小为 1。特征拼接的目的在于将所有的输入图像得到的中间特征进行拼接来尽可能弥补图像中的类别的缺失。（b）特征经拼接后的效果图。经过特征拼接后，每一个新的特征映射将对应至多 4 个不同类别。新拼接后的特征映射将被送入下一层中。特征拼接将使得注意力模型学习到来自 4 张图像中的交叉信息，从而可以使最终模型更准确地定位到语义物体的位置。

含有类别噪声的图像仍然会影响注意力模型挖掘可辨别区域的能力。例如，用“摩托车”检索出的图像通常含有关键词“人”对应的物体，这将使得摩托车和人对应的区域皆被判成摩托车从而引入噪声区域。因而，直接将现有注意力模型[4] 直接应用到本任务中对挖掘出的可辨别区域的质量有一定影响。为了解决这一问题，本小节提出在 CAM 模型的基础上引入一种特征拼接技术。

本特征拼接技术的流程图可参见图5.4。传统的注意力模型[4] 在训练的过程中通常将批量中的每一张图像看成一个独立的样本。与其相反的是，特征拼接技术考虑将每一个批量样本拆分成更小的组从而使得注意力模型可以从每一个组的样本中学习到交叉信息。具体而言，令 F 为卷积神经网络中某一中间层输出的特征映射，其维度为（N, C_F, H_F, W_F）。其中，N、C_F、H_F 以及 W_F 分别为批量大小、卷积层的通道数、特征映射的高以及宽。特征拼接技术首先将每一批量中的样本分成多个小组，每个组中包含 N_T（$N_T = n^2, n \in \mathbb{N}^+$）个样本组成的组。在每一组中，所有的特征映射被拼接在一起。具体拼接方法可参见图5.4b。

新得到的特征映射的维度则为（$1, C_F, nH_F, nW_F$）。因而，在特征拼接后，新的批量大小为 $N' = N/N_T$[1]。

一个组内的所有特征映射经拼接后组成了一个新的特征映射。令新组成的特征映射 $\hat{F}$ 的维度为（N', C_F, nH_F, nW_F）。$\hat{F}$ 中的每个特征映射将对应着至多 N_T 个不同的类别。在引入特征拼接后，类别噪声将会在一定程度上被消除。其原因在于一张图像的类别噪声可能存在于另外 $N_T - 1$ 个图像中。在实验中，VGGNet 的 conv5 生成的特征映射被用来做特征拼接。图5.5中给出一些具体的例子来说明特征拼接的有效性。由图5.5可以看出，CAM 在引入特征拼接技术后能够更准确地定位到关键词对应的语义物体并在一定程度上去除了噪声区域的干扰。

5.3.3 噪声区域擦除

本小节主要介绍用于去除网络图像中类别噪声的噪声擦除网络。在正式介绍噪声擦除网络前，本小节首先介绍噪声擦除网络的研究意义。

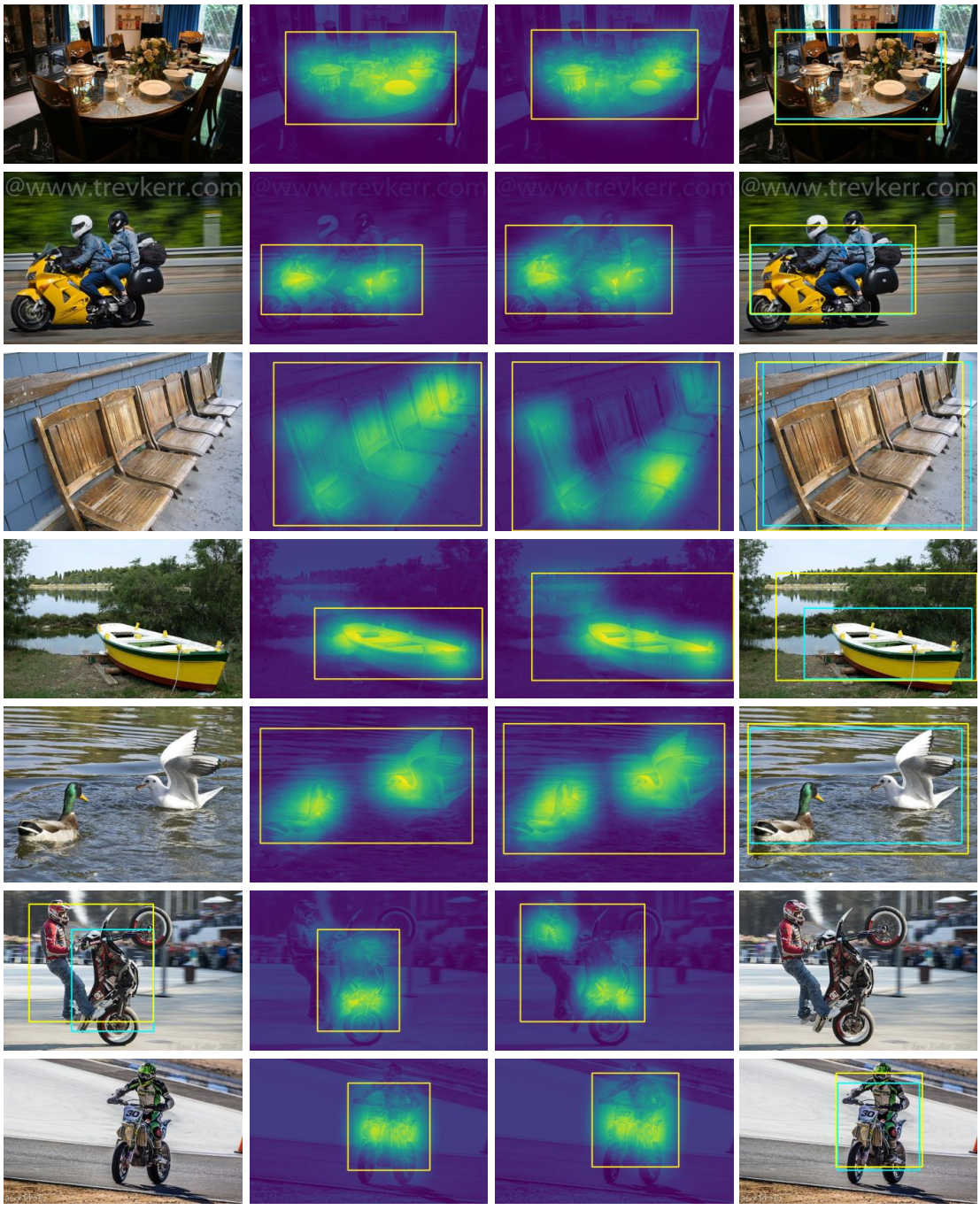
5.3.3.1 研究意义

如上一小节所述，显著性图尽管包含了丰富的前景与背景的信息，但其分割出的前景区域很大概率上会包含类别噪声对应的语义物体（如图5.9c 所示）。为了去除前景中的杂质区域，本小节提出一种噪声擦除网络来擦除掉与关键词不相关的前景区域（属于类别噪声或背景的物体）。

图5.6a 展示了一个简化的样例来说明噪声擦除网络的工作原理。给定一张图像 I 与来自 $\mathcal{L}$ 中的关键词 y 及其对应的显著性图 S，通过设定一阈值来切割显著性图 S 可以简单地将图像 I 划分为两个不相交的区域：背景区域（灰色部分）与前景区域（非灰色部分）[2]。令 R 为某一过分割算法[17] 生成的分割图（在图5.6中，过分割的区域用不同颜色的方块来表示）。本小节的目的在于擦除前景区域中属于类别噪声或背景物体的区域（图5.6中带有红色十字叉的前景区域）。

为了实现这一目的，本节提出一种噪声擦除网络。该网络将从图5.6中的置信区域（由注意力模型生成的可辨别区域）学习语义知识并用训练后的模型对其余的前景区域进行推断并分配给每一个区域一个类别标签。如果图像 I 中的某一前景块对应的类别标签与该图像所属的关键词不一致，则将该前景块标记

[1]在本方法中，N 需要满足以下条件：$N = n^2 \cdot N'$（$n, N' \in \mathbb{N}^+$）。
[2]在本章中，该阈值的大小被设置为 0.2。



(a) (b) (c) (d)

图 5.5 在 CAM 模型[4] 中引入特征拼接前后的对比图。（a）来自互联网中的网络图像；（b）CAM 在引入特征拼接后生成的类别激活图像；（c）由 CAM 生成的类别激活图像；（d）由（b）以及（c）中类别激活图像导出的物体外边框（分别被标记为青色以及黄色）。本章采用阈值 0.2 来判断类别激活图像中每一点是否为前景。由图中可以看出，CAM 在引入特征拼接技术后能够更准确地定位到关键词对应的语义物体并在一定程度上去除了噪声区域的干扰。

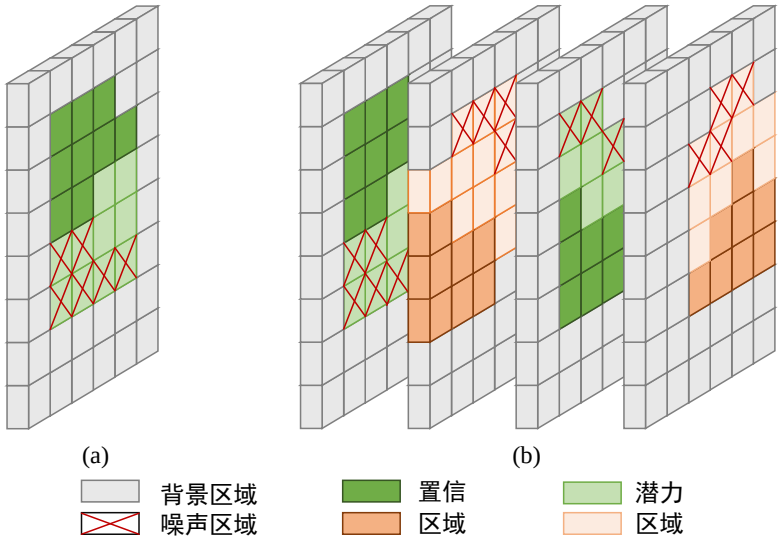


图 5.6 噪声擦除网络的工作原理。在该图中，不同颜色表示不同类别对应的区域。通过用一固定阈值将显著性图的前背景分开，输入图像可以被分为两个区域：背景区域（灰色块表示）以及前景区域（非灰色块表示）。噪声擦除网络在训练过程中从置信区域（由注意力模型生成的小的深色的可辨别区域）学习其中的语义知识。在推断阶段，训练好的模型被用来给每一个前景区域分配一个标签。其目的在于保留（属于关键词）潜力区域对应的块并消除（属于其它关键词或背景物体）噪声区域。（a）对应着批量大小为 1 的情况。每个批量样本中仅含有一个关键词。（b）对应着批量大小大于 1 时的情况。在这种情况下，噪声擦除网络将从同一关键词对应的置信区域中学习它们之间的相似性。

为噪声区域并将其擦除掉。因此，如何保证置信区域中的所有块所对应的类别标签皆正确对于训练噪声擦除网络而言有着至关重要的作用。

5.3.3.2 训练样本选择

如上一小节所述，注意力模型可以挖掘出图像中给定类别相关的可辨别区域。因此，给定图像的显著性图以及类别激活图像，训练的样本应该从二者的交集中选取。

具体来说，令 A_y 为图像 I 的类别激活图像。此时，其前景区域可以被进一步拆分为两个区域。具体的计算方式为计算显著性图 S 与类别激活图像 A_y 的调和均值。其公式如下：

$$H = \frac{1+w}{\frac{1}{S}+\frac{w}{A_y}} \tag{5.1}$$

其中 w 为类别激活图像的权重[3]。公式5.1的作用在于：在保证可辨别区域不变的

[3]在本章中，w 的值被设为 1。

情况下尽可能去除掉可能属于噪声的前景区域。

为了便于说明，H 中被保留的区域被称作置信区域（也即图5.6中深颜色对应的块），剩余的前景区域被称作潜力区域（也即图5.6中淡颜色对应的块）。潜力区域，顾名思义，意味着该区域大概率属于关键词对应的语义物体。被红色十字叉标记的区域则被定义为噪声区域（该区域属于与关键词不相关的语义类别或者背景）。图5.7d 给出一些样例。由图可以看出，置信区域大概率对应着与关键词相关的语义物体。

根据上述定义，分割图 R 可以被简单划分为三个不相交的子集：R_B、R_C 以及 R_P，其分别表示背景中的分割块、置信区域内的分割块以及潜力区域内的分割块。为了方便表示，本章用

$$R_F = R_C \cup R_P \tag{5.2}$$

来表示前景区域内的分割块。在训练过程中，集合 R_C 内部的分割块将分别被用来提取特征进行训练噪声擦除网络。在测试过程中，训练完的模型会为集合 R_F 中的每一个分割块预测一个标签来判断其是否属于噪声区域。

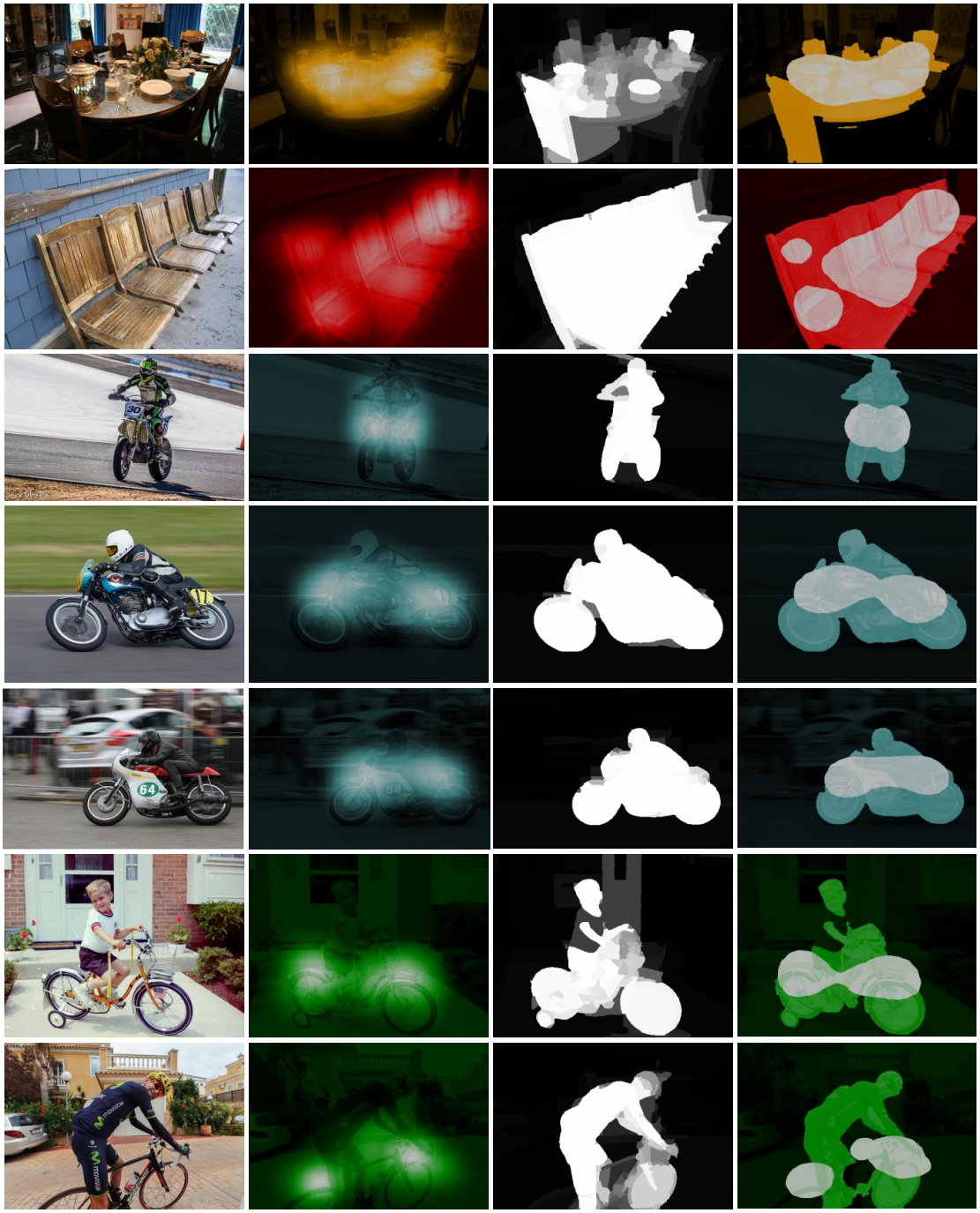
5.3.3.3 噪声擦除网络

本章所用噪声擦除网络由一个全卷积神经网络以及一个区域池化层组成。全卷积神经网络主要来自用于大规模图像分类的网络的全卷积部分。区域池化层则为每一个给定区域生成一个特定长度的特征向量，这里用 $\mathbf{v}$ 来表示。令 f 为全卷积神经网络生成的含有 K 个通道的映射图。对于每一个由过分割算法生成的分割块 R_i，其在第 k 个维度对应的特征向量值 v_i^k 的计算方式为：

$$v_i^k = \frac{1}{|R_i|} \sum_{j \in R_i} f_j^k \tag{5.3}$$

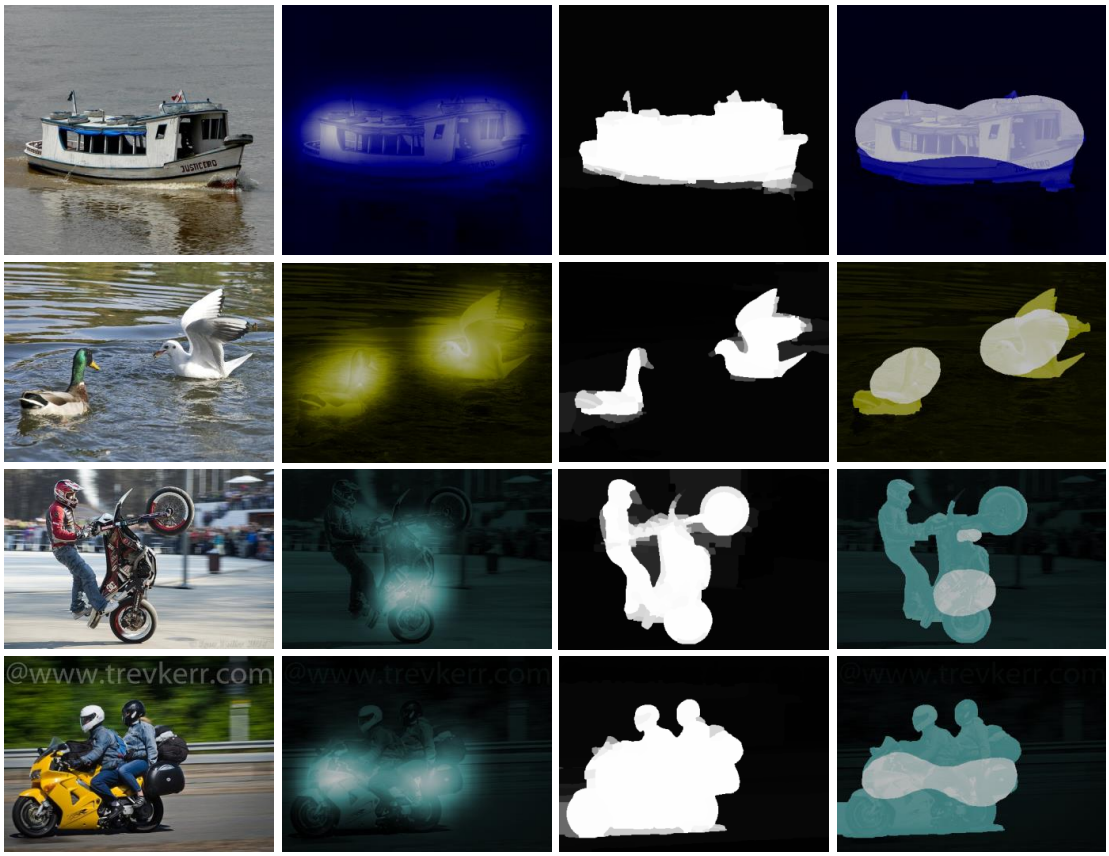
这里 f^k 即为 f 在第 k 个维度的通道映射。

在得到每一个分割块对应的特征向量后，噪声擦除网络可以被简单地看作一个 $|\mathcal{L}|$ 类的分类器。因而，采用简单的在分类网络中常用的交叉熵损失即可实现分类的目的。然而，由于注意力模型通常很难去捕捉到语义物体的边缘信息，置信区域内的分割块也因此会含有错误的标签。当把噪声擦除网络简单看成一个分类网络时，使用错误的标签进行训练将会大大影响其性能，从而导致在推断过程中为潜力区域的分割块分配了错误的标签。



(a) (b) (c) (d)

图 5.7 不同区域示意图（一）。（a）网络图像；（b）类别激活图像；（c）显著性图；（d）用来表示背景区域、置信区域以及潜力区域的示意图。首先，显著性图被二值化为两个区域：前景区域（d 中非黑色的部分）以及背景区域（d 中黑色的部分）。前景区域进而被继续划分成置信区域（白色部分）以及潜力区域（彩色部分）。具体计算方式为二值化类别激活图像与显著性图的调和均值。



(a) (b) (c) (d)

图 5.8 不同区域示意图（二）。（a）网络图像；（b）类别激活图像；（c）显著性图；（d）用来表示背景区域、置信区域以及潜力区域的示意图。首先，显著性图被二值化为两个区域：前景区域（d 中非黑色的部分）以及背景区域（d 中黑色的部分）。前景区域进而被继续划分成置信区域（白色部分）以及潜力区域（彩色部分）。具体计算方式为二值化类别激活图像与显著性图的调和均值。

解决以上问题的一个办法为在分类网络中引入聚类的概念使得属于同一个语义类别的分割块被划分到一起。为了实现这一目的，除了简单地加入交叉熵函数，本章也同时引入一个新的损失项来使得属于同一个语义类别的分割块对应的特征向量之间的距离尽可能小而不属于同一个语义类别的分割块对应的特征向量之间的距离尽可能大，从而达到聚类的目的。

独立地把每张训练图像看成一个样本（如多数分类网络文献[37, 107, 124] 所述）并不包含任何图像之间的交叉信息，因而并不能使噪声擦除网络学到不同类别语义物体之间的差异。在训练噪声擦除网络的过程中，可将同一个批量图像中所有来自置信区域的分割块对应的特征向量看成独立的个体。由于一个批

量中提取的分割块可能属于多个不同的语义物体，从而实现了学习不同类别语义物体之间差异的目的。例如，在图5.6b 中，来自不同置信区域（不同颜色的块）的一对特征向量在训练过程中将会被惩罚。相反，来自同一种颜色的置信区域内的特征向量对之间的距离应该接近 0。因此，本章所用噪声擦除网络的损失函数的定义具体如下。

损失函数令 M 为训练过程中的批量大小。噪声擦除网络损失函数的第一项 L_{cls} 为一 $|\mathcal{L}|$ 类交叉熵函数。则第 m 张图像中区域 $R_{m,i}$ 属于类别 y（该图像对应的关键词）的概率可表示为：

$$p_{m,i}^{y} = \frac{\exp(\sum_{k=1}^{K} w_{y}^{k} \times v_{m,i}^{k})}{\sum_{l \in \mathcal{L}} \exp(\sum_{k=1}^{K} w_{l}^{k} \times v_{m,i}^{k})} \tag{5.4}$$

其中 w 为可学习的参数。此时，该批量对应的损失 L_{cls} 可以被定义为：

$$L_{cls} = -\sum_{m=1}^{M} \sum_{R_{m,i} \in R_{m,C}} \log(p_{m,i}^{y}) \tag{5.5}$$

其中，$R_{m,C}$ 为图像 I_m 对应的置信区域。

为了使噪声擦除网络实现聚类的功能，下面将介绍其损失函数的第二项 L_{simi}。L_{simi} 的主要作用为增大同一个批量样本中不属于同一个语义类别的特征向量对的距离并缩小属于同一个语义类别的特征向量对的距离。具体来说，当前批量第 n 张图像内特征向量的中心点可被定义为：

$$\bar{\mathbf{v}}_n = \frac{1}{|R_{n,C}|} \sum_{R_{n,i} \in R_{n,C}} \mathbf{v}_{n,i} \tag{5.6}$$

考虑到注意力模型的缺陷（不能够完美地定位到所有的语义物体区域），提取的置信区域时而会包含不属于关键词对应的区域。为了解决这一问题，本模型并非直接计算两个不同的特征向量之间的距离而是计算每一个特征向量与不同特征向量中心点之间的距离。因此，一个批量对应的 L_{simi} 的表达式为：

$$\begin{aligned} L_{simi} = -\sum_{m=1}^{M} \sum_{R_{m,i} \in R_{m,C}} \sum_{n=1}^{M} \Big[& \mathbb{I}_{\{y_m = y_n\}} \log(d(\mathbf{v}_{m,i}, \bar{\mathbf{v}}_n)) \\ & + \mathbb{I}_{\{y_m \neq y_n\}} \log(1 - d(\mathbf{v}_{m,i}, \bar{\mathbf{v}}_n)) \Big] \end{aligned} \tag{5.7}$$

其中 $\mathbb{I}$ 为一示性函数，$d(\cdot,\cdot)$ 为两个特征向量之间的相似性距离。该距离可定义为

$$d(\mathbf{v}_{m,i}, \bar{\mathbf{v}}_n) = \exp(-||\mathbf{v}_{m,i} - \bar{\mathbf{v}}_n||_2^2) \tag{5.8}$$

Algorithm 2: 伪标注的生成

输入: 输入图像 I 及其标签 y；分割图 R；概率值 $\mathbf{p}$；显著性图 S
输出: 伪标注 G

for $R_i \in R_F$ **do**
  $C_m \leftarrow \text{argmax}_{l \in \mathcal{L}} p_i^l$
  if $C_m \notin \mathbf{y}$ **then**
    $G_j \leftarrow l_s, \ \forall j \in R_i$ ⇐ 擦除
    continue
  end
  for $j \in R_i$ **do**
    $G_j \leftarrow S_j$ ⇐ 保留
  end
end
输出 G

上述定义可以使得置信区域内带有不同标签的特征向量对被惩罚，因而使得噪声擦除网络能够更为容易地分辨出给定区域的类别。此时，噪声擦除网络的最终损失函数为

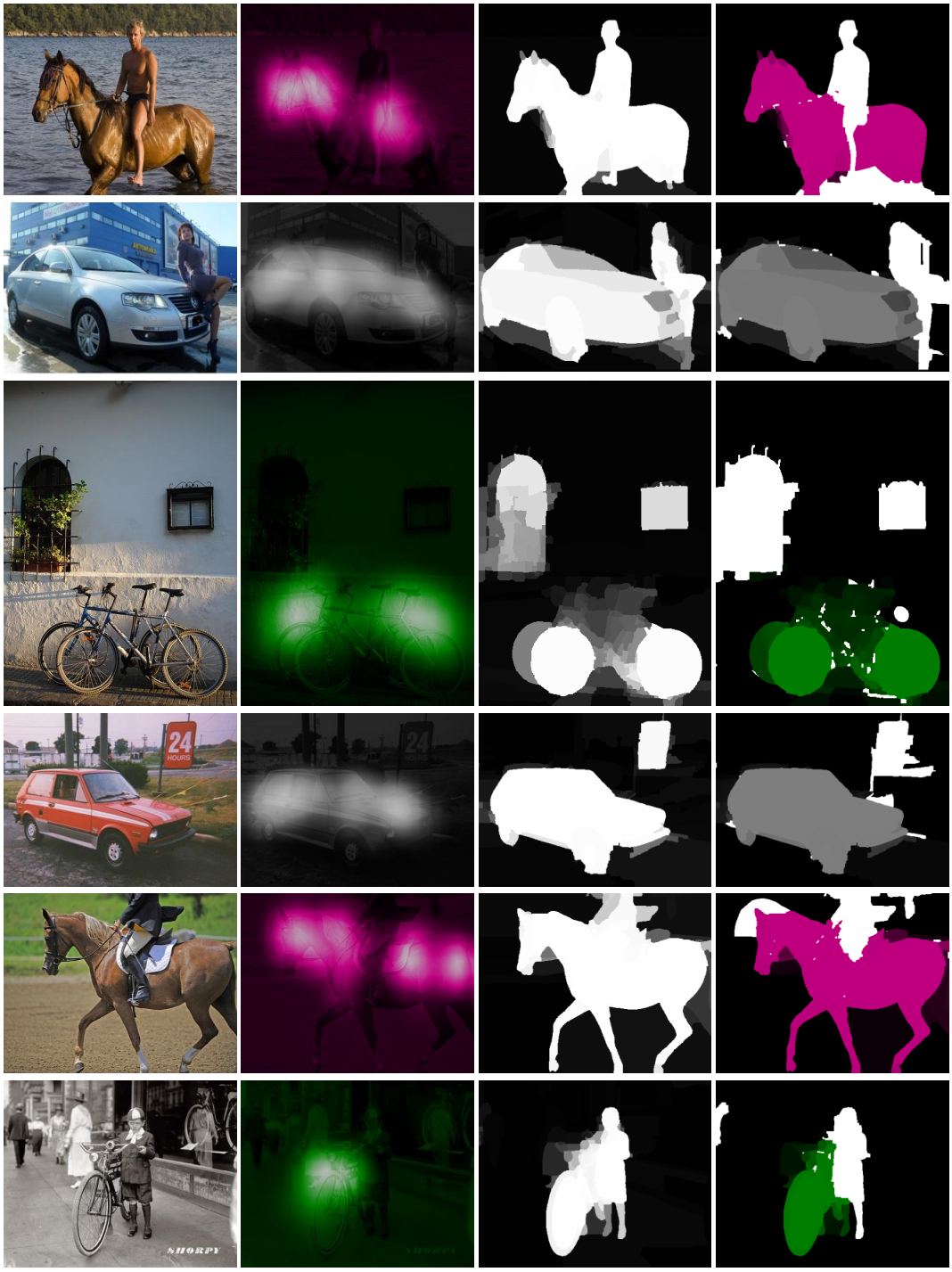
$$L = L_{cls} + \lambda L_{simi} \tag{5.9}$$

其中 λ 为一标量用来控制 L_{simi} 的重要性[4]。训练过程中，SGD 被用来优化整个模型。更多具体参数可参见实验部分。

测试阶段在测试阶段，噪声擦除网络被用来给每一个前景区域内的分割块预测一个标签并根据预测的标签判断该分割块是否被擦除。给定一张图像 I 及其关键词 y 和分割图 R，首先根据公式5.4算出每个前景分割块属于每一个语义类别的概率。然后根据算法2来生成该图像对应的伪标注。具体而言，如果某个前景分割块 R_i 被预测的标签与 y 不一致，则在伪标注中给该分割块分配一特殊类别 l_s（意味着在学习语义分割网络时，对应的区域将不返回梯度信息）。

图5.9给出了一些噪声擦除网络生成的结果图。可以看出，显著性区域很容易包含很多背景物体或与关键词不相关的语义物体。当经过噪声擦除网络擦除后，部分不相关的物体已经明显地被擦除掉。实验部分将给出更多数值上的结果对比。

[4]在本文中，λ 被设置为 0.2。



(a) (b) (c) (d)

图 5.9 （a）输入图像；（b）类别激活图像；（c）显著图像；（d）噪声擦除网络生成的结果。在（d）中，彩色部分为保留的语义区域而白色区域为被擦除掉的区域。在训练语义分割模型的过程中，被擦除的区域将被忽略掉。由图可以看出，通过简单的聚类策略，噪声擦除网络可以较为精准地找到不属于关键词的区域。

5.3.4 语义分割模块

与显著性图类似，由噪声擦除网络生成的伪标注是灰度图，其中每个像素值代表其对应位置成为前景的概率（被标记为特殊标签的区域除外）。因此，参照 Wei 等[71]，本章采用以下交叉熵函数来优化整个语义分割网络：

$$L_{SS}(\theta) = \sum_{n=1}^{N} \left[\hat{q}_n^0 \log q_n(l_0|I;\theta) + \hat{q}_n^c \log q_n(y|I;\theta) \right] \tag{5.10}$$

其中 N 为图像 I 中像素的个数，θ 为网络参数，$\hat{q}_n^c = G_n$ 表示第 n 个像素为前景的概率值，

$$\hat{q}_n^0 = 1 - \hat{q}_n^c \tag{5.11}$$

以及 $q_n(\mathbf{y}|I;\theta)$ 表示第 n 个像素属于类别 y 的概率。该概率值可以由语义分割网络[52] 直接得到。

第四节 实验验证

本小节将对本章提出方法与现有基于网络监督以及弱监督语义分割算法进行对比。除此之外，本小节也将对模型内不同模块的作用以及方法的不足之处进行分析。

5.4.1 实现细节

数据集给定 PASCAL VOC 2012 数据集[1] 内的 20 个类别，首先为每个类别在 Flickr 网站上根据相关性检索 2000 张图像。考虑到检索到的图像的多样性（例如，一些图像可能含有非常复杂的背景或非常低的对比度），这里采用了一系列过滤策略来自动筛掉质量比较低的图像。为了评测图像的复杂性，作者首先采用拉普拉斯方差[125, 126] 来判断每张网络图像是否模糊。通过用拉普拉斯算子对图像进行卷积并计算得到结果的方差，可以得到每张图像对应的得分。在本章实验中，如果网络图像的得分在 50 以上，则判断该图像为复杂图像并将其滤掉。其次，带有较大饱和度以及亮度的图像也将被滤掉。通过将图像从 RGB 色彩空间变换到 HSV 空间，可以计算出 H 以及 V 通道的均值。如果任何一个通道的均值大于 20，则该图像也将被滤掉。最终，经过过滤后的图像数量为 33000 张（每个类别大约剩余 1650 张图像）。为了表示方便，这里用 $\mathcal{D}(W)$ 来表示该网络图像数据集。

表 5.1 噪声擦除网络的抗噪声能力。为了体现本章提出噪声擦除网络的抗噪声能力，本表选取 9 个类别在是否使用噪声擦除网络时对应的分割 IoU 值，其中包含 6 个含有噪声较多的类别以及 3 个含有噪声最少类别。

类别	NENet（VGG）		NENet（ResNet）	
	✗	✓	✗	✓
自行车	26.1%	30.0%	29.7%	33.3%
椅子	9.2%	14.6%	10.8%	13.5%
马	54.1%	60.1%	64.5%	74.2%
摩托车	55.8%	57.0%	60.3%	61.8%
餐桌	6.2%	8.1%	7.0%	10.2%
狗	65.3%	68.5%	76.9%	79.2%
瓶子	56.5%	58.9%	69.0%	68.0%
公共汽车	81.0%	80.4%	82.2%	83.3%
猫	69.2%	66.4%	79.3%	81.9%
均值	54.0%	56.9%	57.4%	61.6%

评测方法与多数语义分割算法类似，这里采用预测结果与标注的交并比（intersection-over-union，IoU）来评测分割结果的质量。IoU 的计算方式可参见公式4.13。

模型参数设置除了特别声明，本章所用注意力模型的参数与 CAM 模型[4]中参数基本一致。参数 n 以及批量大小分别被设置为 2 与 16，因此批量内组的大小为 4 且特征拼接后的批量大小亦为 4。

对于噪声擦除网络，本章采用 ResNet-50[107] 作为主干网络。为了提高生成的类别激活图像的分辨率，ResNet-50 中最后一个阶段的滑动步长被设为 1。具体的超参数设置如下：学习速率（0.001）、权重衰减因子（0.0005）、动量（0.9）以及批量大小（16）。网络训练周期为 5。

对于分割网络，本章采用与多数现有算法相同的 DeepLab-LargeFOV[52] 网络。所有的超参数也与 DeepLab-LargeFOV 中所用超参数相同。除此之外，本章采用了一个图模型[113] 作为后处理工具来进一步平滑生成的分割图像。为了公平地与现有算法进行对比，这里将汇报 DeepLab-LargeFOV 基于 VGGNet[37] 以及 ResNets[107] 的结果。

5.4.2 敏感性分析

为了分析本方法中每个模块的重要性，本小节采用了控制变量法对模型进行了敏感性分析。

特征拼接的作用： 图5.5给出了使用原始 CAM 模型[4] 以及带有特征拼接的 CAM 模型得到的类别激活图像。从图5.5可以看出，本章提出的特征拼接策略可以更为准确地定位到关键词对应的语义物体。为了更准确地对比两个方法的结果，这里采用一固定阈值对得到的类别激活图像进行二值化并用矩形框对可辨别区域进行标注。由图5.5d 可以看出，由本方法得到的可辨别区域可以更好地覆盖在语义物体上。以第 4 张图像为例，本方法生成的可辨别区域几乎完全覆盖到“船”对应的语义物体上，而 CAM 生成的模型则将部分背景区域也看作可辨别区域。这些现象表明，本章提出的特征拼接可以较好地学习到图像之间的交互信息从而可以更为精准地定位到语义物体的位置。

抵抗噪声的能力： 为了验证本方法处理类别噪声的能力，本段主要阐述在是否使用类别噪声网络情况下的对比分析。表5.1给出了本方法在不同语义类别上的 mIoU 值。为了更清楚地表明噪声擦除网络在有类别噪声的情况下有效且不影响类别噪声较少的类别，表5.1将 9 个类别分成两组，分别包含 6 个含有最多类别噪声的类别以及 3 个含有较少类别噪声的类别。可以看出，在使用 VGGNet 以及 ResNet 作为主干网络时，本方法在 9 个语义类别上的结果都有一定提升。特别地，对于含有较多类别噪声的类别，使用噪声擦除网络后得到的 mIoU 提升更为明显。对于整个数据集的 mIoU 而言，使用 VGG 版本的 DeepLab-LargeFOV 模型可以得到近 3 个百分点的提升而当使用 ResNet 版本的 DeepLab-LargeFOV 模型时，提升的幅度达到 4.2 个百分点。这些实验结果表明，噪声擦除网络在去除网络图像中的类别噪声方面有明显效果。

训练图像的数量： 为了进一步衡量由噪声擦除网络过滤后得到的伪标注数据的质量，作者随机从每一个类别中抽取不同数量的训练图像并用其训练语义分割网络。选取训练图像的具体百分比值可参见表5.3。从表中可以看出，当减少训练图像的数量到原来的一半时，语义分割模型生成分割结果的 mIoU 值仅下降了 1.1 个百分点。当进一步减少训练图像的数量到原来的百分之三十时，语义分割模型生成分割结果的 mIoU 值下降了 2.6 个百分点。值得注意的是，当仅用少于 10000 张经噪声擦除网络过滤后的标注图像来训练语义分割模型时，得到的分割结果仍比使用 30000 多张无噪声过滤的训练图像训练出的模型效果明

表 5.2 本方法的敏感性分析。值得一提的是，本表中所有结果皆直接来自于卷积神经网络且没有任何后处理工具被使用。其中，低层级特征指的是显著性图像。当噪声擦除网络（NENet）被使用时，最终的分割结果有了 4.2 个百分点的提升。当将 VOC 2012 中的训练图像与类别标签参与训练时，可以得到额外的 4.2 个百分点的提升。本表中所有结果皆基于 ResNet 版本 DeepLab-LargeFOV 模型。用于评测的数据集为 VOC 2012 验证集。最好的结果已被加粗强调。

序号	低层级特征	NENet	VOC 数据	mIoU (%)
1	✓			57.4
2	✓	✓		$61.6_{+4.2}$
3	✓	✓	✓	$\mathbf{65.8}_{+4.2}$

表 5.3 使用不同数量网络图像时的 mIoU 值对比。其中“百分比”项为随机从每个类别图像中抽取的百分比数。本表中所有结果皆基于 ResNet 版本 DeepLab-LargeFOV 模型且用于评测的数据集为 VOC 2012 验证集。最好的结果已被加粗强调。从表中可以看出，当仅有百分之三十的图像被用于训练时，mIoU 值仅有少于 3 个百分点的下降。这一现象说明噪声擦除网络能够有效地过滤掉多数类别噪声。

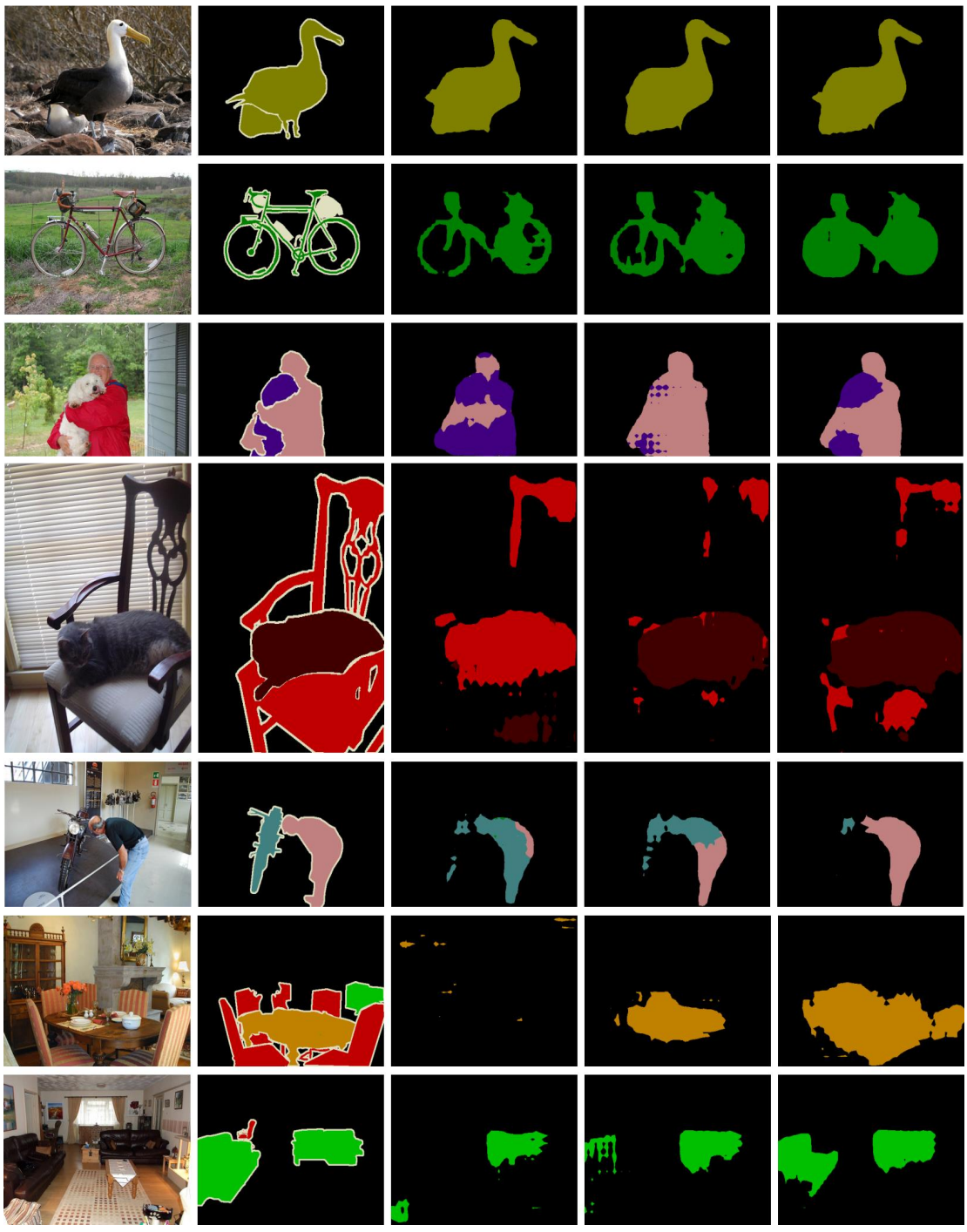
序号	网络图像数量	百分比	NENet	mIoU (%)
1	33,000	100%	✗	57.4
2	33,000	100%	✓	**61.6**
3	16,500	50%	✓	60.5
4	9,900	30%	✓	59.0

显。由表5.3可以看出，仅使用 9900 张经噪声擦除网络过滤后的标注图像得到的分割结果比无噪声擦除网络过滤的图像参与训练得到的分割结果高 1.6 个百分点。这一现象表明，噪声擦除对于去除网络图像中的类别噪声有明显作用。

5.4.3 引入 VOC 训练图像

以上所有实验结果皆基于纯网络图像数据集 $\mathcal{D}(W)$。每张图像都可能含有类别噪声。为了更为公平地与现有弱监督语义分割方法进行比较，本小节将 PASCAL VOC 2012 训练集中的图像以及类别标签参与训练。该数据集共包含 10582 张图像[1, 15]。为了方便表示，本小节用符号 $\mathcal{D}(V)$ 来表示该数据集。

对于某张来自数据集 $\mathcal{D}(W)$ 或 $\mathcal{D}(V)$ 中的图像 I 以及其对应的类别标签集合 $\mathbf{y}$，令 $Z \in \mathbb{R}^{T\times(L+1)}$ 为在 $\mathcal{D}(W)$ 数据集上训练得到的分割模型以 I 为输入生成的得分图像，其中 T 为图像中所有像素点的个数。根据已得到的 Z，可以将图像 I 划分成一二值图 B，其中语义区域表示前景所在的位置。特别地，对于图



(a) (b) (c) (d) (e)

图 5.10 使用不同配置时的分割效果图。（a）原始图像；（b）标注图像；（c）无擦声擦除网络时的分割结果；（d）有擦声擦除网络时的分割结果；（e）将 VOC 训练图像参与训练时的分割结果。由图可以看出，在使用擦声擦除网络后，本模型生成的分割结果能够更准确地找到语义物体。

表 5.4 本方法与现有方法在 VOC 2012 验证集（val）以及测试集（test）上的量化结果对比。其中，33K 网络图像指的是 $\mathcal{D}(W)$ 图像集且 $\mathcal{D}(V)$ 指的是 10582 张 PASCAL VOC 训练集。仅用类别标签作为监督数据的方法被标记为“弱监督”。含有少量像素级精度标签数据的方法被标记为“半监督”。仅有关键词作为监督数据的方法被标记为“纯网络监督”。基于 VGGNet 以及 ResNets 基础模型的最好结果被分别用绿色以及红色字体表示。由表中数据可以看出，本方法在仅有网络图像以及带有类别噪声的关键词作为监督时得到的结果已经优于大部分现有方法。当将 VOC 训练集的图像以及类别标签参与训练时，本方法的结果则明显优于现有所有方法。

方法	训练数据	监督方式	主干网络	mIoU val	mIoU test
SEC[7]	$\mathcal{D}(V)$	弱监督	VGGNet	50.7	51.7
AugFeed[65]	$\mathcal{D}(V)$	弱监督	VGGNet	54.3	55.5
Oh 等[67]	$\mathcal{D}(V)$	弱监督	VGGNet	55.7	56.7
AE-PSL[8]	$\mathcal{D}(V)$	弱监督	VGGNet	55.0	55.7
DCSP[73]	$\mathcal{D}(V)$	弱监督	VGGNet	58.6	59.2
DCSP[73]	$\mathcal{D}(V)$	弱监督	ResNets	60.8	61.9
DSRG[68]	$\mathcal{D}(V)$	弱监督	VGGNet	59.0	60.4
DSRG[68]	$\mathcal{D}(V)$	弱监督	ResNet	61.4	63.2
MCOF[70]	$\mathcal{D}(V)$	弱监督	VGGNet	56.2	57.6
Ahn 等[69]	$\mathcal{D}(V)$	弱监督	VGGNet	58.4	60.5
Wei 等[5]	$\mathcal{D}(V)$	弱监督	VGGNet	60.4	60.8
GAIN[50]	1464 像素 + $\mathcal{D}(V)$	半监督	VGGNet	60.5	62.1
Fan 等[75]	$\mathcal{D}(V)$	弱监督	ResNet	63.6	64.5
基于网络数据方法语义分割结果					
STC[71]	40K 网络 + $\mathcal{D}(V)$	弱监督	VGGNet	49.8	51.2
WebS-i2[118]	19K 网络 + $\mathcal{D}(V)$	弱监督	VGGNet	53.4	55.3
Hong 等[78]	网络视频 + $\mathcal{D}(V)$	弱监督	VGGNet	58.1	58.7
Shen 等[79]	80K 网络 + $\mathcal{D}(V)$	弱监督	VGGNet	58.8	60.2
Shen 等[79]	80K 网络 +$\mathcal{D}(V)$	弱监督	ResNet	63.0	63.9
WebSearch	33K 网络 + $\mathcal{D}(V)$	弱监督	VGGNet	62.5	62.2
WebSearch	33K 网络 + $\mathcal{D}(V)$	弱监督	ResNets	65.8	66.1
Shen 等[79]	80K 网络	纯网络监督	VGGNet	56.6	-
WebSearch	33K 网络	纯网络监督	VGGNet	59.5	59.3
WebSearch	33K 网络	纯网络监督	ResNet	61.6	62.0

I 中的位置 t，如满足：

$$\underset{\{l_0,\mathcal{L}\}}{\operatorname{argmax}}(Z_t) \in \mathbf{y} \tag{5.12}$$

则令 $B_t = 1$，否则为 0。

给定二值图 B，通过计算 B 与类别激活图像[5] 之间的调和均值（公式5.1）可以得到新的得分图 $Q \in \mathbb{R}^{T\times(L+1)}$。不难发现，$Q_t^c$ 即为位置 t 属于类别 c 的概率值。本小节默认设置 $Q_t^0 = 0.1$，其意义为每一个位置属于背景类别的概率为 0.1。此时，新的用于训练语义分割模型的伪标注数据在位置 t 处的值可以由下式计算得到

$$\hat{G}_t = \underset{\{l_0,\mathcal{L}\}}{\operatorname{argmax}}(Q_t) \tag{5.13}$$

给定 $\hat{G}$ 后，可以通过优化一个新的语义分割模型[52] 来生成新的分割图。

引入 VOC 图像的效果： 由表5.2以及表5.4可以看出，当将 PASCAL VOC 2012 数据集中的训练图像与类别标签参与训练后，得到的分割结果在 VOC 验证集上有较大提升。当用 VGGNet[37] 作为主干网络时，提升的幅度为接近 3 个百分点。当用 ResNet-101[107] 作为主干网络时，提升的幅度为 4.2 个百分点。这一现象在 VOC 2012 测试集中也较为明显。

除了量化结果，图5.10给出了不同模型配置下的分割图。对于较为简单的场景图像，基于 $\mathcal{D}(W)$ 训练集的模型已经可以生成质量较高的分割结果。对于含有多个类别的场景图像，引入噪声擦除网络可以有效地抵抗类别噪声带来的干扰。由于类别噪声的去除，更多语义区域可以被精准地分割出来。

5.4.4 与现有方法的对比

本小节将本方法得到的分割结果与现有方法得到的结果进行对比。由于多数相关的工作皆以类别标签作为监督数据，因而本小节将以引入 VOC 2012 训练图像的分割模型与其进行对比。为了公平比较，所有的方法皆基于 DeepLab-LargeFOV[127] 分割模型。

表5.4列出本方法与现有方法在 VOC 2012 验证集以及测试集上的结果对比。可以看出，在仅使用纯网络图像 $\mathcal{D}(W)$ 参与训练时，本方法生成的分割结果可以实现 61.6% 的 mIoU 值。这个结果已经明显优于大部分已有基于弱监督的语义分割算法。当进一步将 VOC 2012 中训练图像引入训练后，本方法得到的 mIoU

[5]本章采用上一章中在 $\mathcal{D}(V)$ 数据集上训练得到的注意力模型。

值进一步提升到 65.8%。这一结果已明显优于所有现有基于弱监督的语义分割算法。

与现有基于网络数据的方法相比，本方法也明显优于现有方法。STC[71]、WebS-i2[118]、Hong 等[78]、Shen 等[79] 方法在 VOC 2012 验证集上的结果分别是 49.8%、53.4%、58.1%、56.6%，而本方法在仅使用网络数据的情况下得到的分割结果为 59.5%。这一结果明显优于上述提到的所有方法，且比 mIoU 值最高的方法（Hong 等[78]）多 1.4 个百分点[6]。另外，与以上方法不同的是，本方法除了使用含有简单场景的网络图像外，仍包含了含有较难场景的网络图像。这一现象说明提高输入数据的多样性对提升弱监督语义分割模型的性能有一定效果。上述现象表明，去除复杂图像中的类别噪声是一个实现语义分割自主学习的有效方法。

第五节 本章小结

本章提出了一个新颖的计算机视觉任务，即在从含有类别噪声的纯网络数据中学习语义分割模型。现有的基于弱监督的语义分割方法皆利用大量已有数据集的训练图像以及精准的类别标签来生成分割结果。基于网络数据的方法虽然采用了互联网中图像来训练语义分割模型，但其仍依赖于精准的类别标签训练出的注意力模型来生成初始的种子区域。与以上方法不同的是，本章提出的方法不依赖任何含有精准类别标签的数据集且将网络图像中含有较为复杂场景的图像参与训练以提升训练数据的多样性。

在算法方面，考虑到网络图像中的类别噪声，本章提出了一种特征拼接策略。在训练注意力模型时，可以将多张图像的特征进行拼接，从而组成一张分辨率更高的特征图像。这一操作将使得注意力模型能够学习到多个不同图像之间的交叉信息。给定以上方法得到的注意力模型，本章将其生成的类别激活图像与图像的显著性图相结合并提出了噪声擦除网络来擦除掉网络图像中的噪声区域。该噪声擦除网络主要从类别激活图像中的可辨别区域中学习语义知识并将学习到的语义知识作用于所有前景区域从而生成用于训练语义分割模型的伪标注数据。

为了测试本章提出方法的效果，本章将生成的语义分割模型在 PASCAL

[6]值得注意的是，Hong 等[78] 方法虽然利用了互联网中的免费视频图像作为训练数据，其也应用了含有精准类别标签的 Pascal VOC 2012 数据集训练的注意力模型对训练数据进行处理。

VOC 2012 的验证集以及测试集进行了验证。实验结果表明，本章方法在仅依赖于互联网图像时已优于多数现有基于弱监督的语义分割方法。当将 PASCAL VOC 2012 中训练图像集与类别标签参与训练时，本章方法可以进一步提升分割结果并在所有基于弱监督的语义分割方法中取得最佳效果。

第六章 总结与展望

语义分割已然成为计算机视觉领域中最为重要的分支之一。基于全监督的语义分割任务需要大量人工标注的像素级别精度的标注数据。弱监督语义分割由于仅依赖图像级别的标注，因而大大简化了人工标注的难度并减少对于人工标注的依赖。本文受人类视觉系统启发提出了语义分割自主学习的概念，即在仅给定待分割物体的类别标签且不依赖于任何人工干预的情况下完成语义分割模型的自主训练。

第一节 本文工作总结

为了实现语义分割自主学习，本文采用了两种不同的视觉注意机制（显著图像与类别激活图像）来生成用于训练语义分割模型的伪标注数据。由于显著图像与类别激活图像的质量将直接影响到生成的伪标注数据的质量，考虑到现有显著性物体检测模型与注意力模型的缺陷，本文分别提出了更为先进的神经网络模型。下面将对本文主要工作按章节顺序进行总结。

本文首先介绍了语义分割的发展背景以及语义分割几种不同的研究方向，包括全监督语义分割、半监督语义分割以及弱监督语义分割，并提出了语义分割自主学习的概念。

第二章结合文本的研究内容对显著性物体检测模型、注意力模型以及现有弱监督语义分割算法进行了回顾。

第三章提出了一种基于短连接与卷积神经网络的显著性物体检测网络架构。该方法利用了卷积神经网络的多层级特征并将深度监督的概念引进到了显著性物体检测方向中。通过将卷积神经网络中含有高级语义信息的深层特征经过短连接的方式与含有丰富细节信息的浅层特征相结合可以将网络中多个层级特征有效结合在一起。高层级特征可以较为精准地找到场景中显著性物体的具体位置且低层级特征可以准确地勾勒出显著性物体的边缘。二者相结合可以使生成的显著性物体具有较高的质量。为了定量验证本章方法的有效性，本章对本算法在 5 个该领域中被广泛使用的数据集上进行了测试。实验结果表明，在不同评测指标下，本算法生成的显著性图像在 5 个数据集上皆有较大提升。

第四章提出了基于自擦除策略的注意力模型。考虑到现有基于对抗擦除模型的缺陷（即在训练过程中很难选择合适的停止时机从而导致可辨别区域的随意扩散），本章提了自擦除的概念并设计了两种自擦除策略来达到生成高质量类别激活图像的目的。第一种策略将输入图像分为三个区域：背景区域、潜力区域、前景区域。在训练过程中，通过将背景区域对应特征值的符号进行反转可使模型仅在潜力区域内寻找可辨别区域从而达到抑制可辨别区域随意扩散的目的。第二种策略仅考虑背景区域内的特征并用交叉熵损失通知模型该区域内没有语义物体。为了测试以上两种自擦除策略的有效性，本章将得到的类别激活图像应用到弱监督语义分割任务。通过计算类别激活图像与显著性图像的调和均值可以生成伪标注数据进而训练语义分割模型。实验结果表明，本章提出的注意力模型可以在弱监督语义分割领域取得较好效果。

第五章提出了如何从免费的互联网资源中自主地学习语义分割模型。考虑到互联网图像的多样性，其中大多数图像会含有类别噪声。为了解决类别噪声并从图像中提取出高质量的伪标注数据，本章提出了噪声擦除网络。该网络以显著图像与类别激活图像为输入，从类别激活图像中的可辨别区域学习每个类别对应的语义知识并对前景中的区域进行推断，从而得到每个前景区域对应的类别标签。为了测试本章所提方法的有效性，本章将网络图片以及伪标注数据用于训练语义分割网络并与现有的弱监督方法进行对比。当仅采用网络图片进行训练时，本章所提方法明显优于大部分现有的弱监督语义分割网络。当将 VOC 2012 训练图像及其类别标签参与训练时，本章所提方法实现了最好的效果。

第二节 未来工作展望

语义分割自主学习完全仿生了人类的视觉系统，可以在无人工标注的情况下实现语义分割任务。本文提出了将显著图像与类别激活图像相结合来训练噪声擦除网络的方法进而生成每张图像的伪标注数据。因而，如何生成高质量的显著图像以及类别激活图像具有重要研究意义。针对语义分割自主学习这一新研究方向，本文仅提出了一种可行方案，但仍有许多重点项目可以继续深入研究。下面将针对本文每章中可研究的方向以及未来的工作进行简要介绍。

第三章提出了基于短连接以及深度监督的显著性物体检测网络。虽然本方法与现有方法比已具有明显优势，但究其结构仍有很大改进空间。首先，本方法仅考虑将不同层级的特征进行简单的融合，如何将融合后的特征进一步有效

地融合从而构建出更丰富的特征表达是一个值得研究的地方。另外，本文在侧向路径中仅采用了简单的卷积层来实现特征变换，如何设计更有效的模块来取代简单的卷积层也是一个值得研究的方向。

第四章提出了基于自擦除的注意力模型。本方法将图像的背景知识作为先验并用其辅助注意力模型生成类别激活图像。由于背景区域仅通过二值化第一阶段的类别激活图像而得，因而仍然缺乏精准的边缘信息。针对这一缺陷，一大值得研究的方向为通过将少量带有像素级别标注的数据参与训练来改善类别激活图像的质量。

第五章提出了使用噪声擦除网络来过滤掉网络图像中的类别噪声。通过学习类别激活图像中的语义知识可以对图像中的前景物体进行推断从而过滤掉与类别标签无关的前景区域。由于该问题较为新颖，因而有多种探索途径来实现语义分割自主学习，如使用扩张损失函数、基于超像素的条件随机场等。

参考文献

[1] EVERINGHAM M, ESLAMI S A, VAN GOOL L, et al. The Pascal Visual Object Classes Challenge: A Retrospective[J]. International Journal of Computer Vision, 2015, 111 (1): 98–136.

[2] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE, 1998, 86 (11): 2278–2324.

[3] HOU Q, CHENG M.-M, HU X, et al. Deeply Supervised Salient Object Detection with Short Connections[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3203–3212.

[4] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning Deep Features for Discriminative Localization[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2921–2929.

[5] WEI Y, XIAO H, SHI H, et al. Revisiting Dilated Convolution: A Simple Approach for Weakly-and Semi-Supervised Semantic Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7268–7277.

[6] ZHANG J, BARGAL S A, LIN Z, et al. Top-down neural attention by excitation backprop[J]. International Journal of Computer Vision, 2018, 126 (10): 1084–1102.

[7] KOLESNIKOV A, LAMPERT C H. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation[C] // European Conference on Computer Vision. 2016: 695–711.

[8] WEI Y, FENG J, LIANG X, et al. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1568–1576.

[9] LIU T, YUAN Z, SUN J, et al. Learning to Detect a Salient Object[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33 (2): 353–367.

[10] YAN Q, XU L, SHI J, et al. Hierarchical Saliency Detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2013: 1155–1162.

[11] LI G, YU Y. Visual Saliency Based on Multiscale Deep Features[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2015: 5455–5463.

[12] LI Y, HOU X, KOCH C, et al. The Secrets of Salient Object Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2014: 280–287.

[13] MARTIN D, FOWLKES C, TAL D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C] // IEEE International Conference on Computer Vision. 2001: 416–423.

[14] MOVAHEDI V, ELDER J H. Design and Perceptual Validation of Performance Measures for Salient Object Segmentation[C] // IEEE CVPRW. 2010: 49–56.

[15] HARIHARAN B, ARBELÁEZ P, BOURDEV L, et al. Semantic Contours from Inverse Detectors[C] // IEEE International Conference on Computer Vision. 2011: 991–998.

[16] PONT-TUSET J, ARBELAEZ P, BARRON J T, et al. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (1): 128–140.

[17] MANINIS K.-K, PONT-TUSET J, ARBELAEZ P, et al. Convolutional Oriented Boundaries: From Image Segmentation to High-Level Tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40 (4): 819–833.

[18] CHENG M.-M, MITRA N J, HUANG X, et al. Global Contrast based Salient Region Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37 (3): 569–582.

[19] ITTI L, KOCH C. Computational Modeling of Visual Attention[J]. Nature reviews neuroscience, 2001, 2 (3): 194–203.

[20] HAREL J, KOCH C, PERONA P. Graph-Based Visual Saliency[C] // Advances in Neural Information Processing Systems. 2006: 545–552.

[21] GOFERMAN S, ZELNIK-MANOR L, TAL A. Context-Aware Saliency Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34 (10): 1915–1926.

[22] CHENG M.-M, ZHANG F.-L, MITRA N J, et al. RepFinder: Finding Approximately Repeated Scene Elements for Image Editing[C] // ACM Transactions on Graphics. Vol. 29. 4. 2010: 83.

[23] HU S.-M, CHEN T, XU K, et al. Internet visual media processing: a survey with graphics and vision applications[J]. 2013, 29 (5): 393–405.

[24] CHENG M.-M, HOU Q.-B, ZHANG S.-H, et al. Intelligent Visual Media Processing: When Graphics Meets Vision[J]. Journal of Computer Science and Technology, 2017, 32 (1): 110–121.

[25] CHENG M, MITRA N J, HUANG X, et al. Global Contrast Based Salient Region Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37 (3): 569–582.

[26] FELZENSZWALB P F, HUTTENLOCHER D P. Efficient Graph-Based Image Segmentation[J]. International Journal of Computer Vision, 2004, 59 (2): 167–181.

[27] JIANG H, WANG J, YUAN Z, et al. Salient Object Detection: A Discriminative Regional Feature Integration Approach[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2083–2090.

[28] ZHAO J, REN B, HOU Q, et al. FLIC: Fast Linear Iterative Clustering With Active Search[C] // Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[29] ARBELAEZ P, MAIRE M, FOWLKES C, et al. Contour Detection and Hierarchical Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011.

[30] HE S, LAU R, LIU W, et al. SuperCNN: A Superpixelwise Convolutional Neural Network for Salient Object Detection[J]. International Journal of Computer Vision, 2015, 115 (3): 330–344.

[31] WANG L, LU H, RUAN X, et al. Deep Networks for Saliency Detection via Local Estimation and Global Search[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3183–3192.

[32] ZHAO R, OUYANG W, LI H, et al. Saliency Detection by Multi-Context Deep Learning[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1265–1274.

[33] GAYOUNG L, YU-WING T, JUNMO K. Deep Saliency with Encoded Low level Distance Map and High Level Features[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[34] LONG J, SHELHAMER E, DARRELL T. Fully Convolutional Networks for Semantic Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431–3440.

[35] LIU N, HAN J. DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016: 678–686.

[36] LI G, YU Y. Deep Contrast Learning for Salient Object Detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[37] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C] // International Conference on Learning Representations. 2015.

[38] KRÄHENBÜHL P, KOLTUN V. Efficient inference in fully connected crfs with gaussian edge potentials[C] // Advances in Neural Information Processing Systems. 2011: 109–117.

[39] WANG L, WANG L, LU H, et al. Saliency Detection with Recurrent Fully Convolutional Networks[C] // European Conference on Computer Vision. 2016.

[40] HARIHARAN B, ARBELÁEZ P, GIRSHICK R, et al. Hypercolumns for Object Segmentation and Fine-Grained Localization[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2015: 447–456.

[41] XIE S, TU Z. Holistically-Nested Edge Detection[J]. International Journal of Computer Vision, 2017, 125 (1): 3–18.

[42] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps[C] // Workshop on International Conference on Learning Representations. 2014.

[43] ZEILER M D, FERGUS R. Visualizing and Understanding Convolutional Networks[C] // European Conference on Computer Vision. 2014: 818–833.

[44] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization[C] // IEEE International Conference on Computer Vision. 2017: 618–626.

[45] ANTOL S, AGRAWAL A, LU J, et al. Vqa: Visual question answering[C] // IEEE International Conference on Computer Vision. 2015: 2425–2433.

[46] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C] // Advances In Neural Information Processing Systems. 2016: 289–297.

[47] SHIH K J, SINGH S, HOIEM D. Where to look: Focus regions for visual question answering[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4613–4621.

[48] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6077–6086.

[49] WU Q, WANG P, SHEN C, et al. Ask me anything: Free-form visual question answering based on knowledge from external sources[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4622–4630.

[50] LI K, WU Z, PENG K.-C, et al. Tell Me Where to Look: Guided Attention Inference Network[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9215–9223.

[51] ZHANG X, WEI Y, FENG J, et al. Adversarial Complementary Learning for Weakly Supervised Object Localization[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1325–1334.

[52] CHEN L.-C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. 2018, 40 (4): 834–848.

[53] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, et al. Conditional Random Fields as Recurrent Neural Networks[C] // IEEE International Conference on Computer Vision. 2015: 1529–1537.

[54] ZHAO H, SHI J, QI X, et al. Pyramid Scene Parsing Network[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2881–2890.

[55] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-Path Refinement Networks with Identity Mappings for High-Resolution Semantic Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[56] LIU Z, LI X, LUO P, et al. Semantic Image Segmentation via Deep Parsing Network[C] // IEEE International Conference on Computer Vision. 2015.

[57] NOH H, HONG S, HAN B. Learning Deconvolution Network for Semantic Segmentation[C] // IEEE International Conference on Computer Vision. 2015: 1520–1528.

[58] LIN G, SHEN C, van den HENGEL A, et al. Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3194–3203.

[59] MOSTAJABI M, YADOLLAHPOUR P, SHAKHNAROVICH G. Feedforward semantic segmentation with zoom-out features[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3376–3385.

[60] PAPANDREOU G, CHEN L.-C, MURPHY K, et al. Weakly-and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation[C] // IEEE International Conference on Computer Vision. 2015: 1742–1750.

[61] PINHEIRO P O, COLLOBERT R. From Image-Level to Pixel-Level Labeling with Convolutional Networks[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1713–1721.

[62] DEMPSTER A P, LAIRD N M, RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the royal statistical society. Series B (methodological), 1977: 1–38.

[63] DIETTERICH T G, LATHROP R H, LOZANO-PÉREZ T. Solving the Multiple Instance Problem with Axis-Parallel Rectangles[J]. Artificial Intelligence, 1997, 89 (1-2): 31–71.

[64] PATHAK D, KRAHENBUHL P, DARRELL T. Constrained Convolutional Neural Networks for Weakly Supervised Segmentation[C] // IEEE International Conference on Computer Vision. 2015: 1796–1804.

[65] QI X, LIU Z, SHI J, et al. Augmented Feedback in Semantic Segmentation under Image Level Supervision[C] // European Conference on Computer Vision. 2016: 90–105.

[66] ROY A, TODOROVIC S. Combining Bottom-Up, Top-Down, and Smoothness Cues for Weakly Supervised Image Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3529–3538.

[67] OH S J, BENENSON R, KHOREVA A, et al. Exploiting Saliency for Object Segmentation from Image Level Labels[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4410–4419.

[68] HUANG Z, WANG X, WANG J, et al. Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7014–7023.

[69] AHN J, KWAK S. Learning Pixel-level Semantic Affinity with Image-level Supervision for Weakly Supervised Semantic Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4981–4990.

[70] WANG X, YOU S, LI X, et al. Weakly-Supervised Semantic Segmentation by Iteratively Mining Common Object Features[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1354–1362.

[71] 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.

[72] HOU Q, DOKANIA P K, MASSICETI D, et al. Bottom-Up Top-Down Cues for Weakly-Supervised Semantic Segmentation[C] // Energy Minimization Methods in Computer Vision and Pattern Recognition. 2017: 263–277.

[73] CHAUDHRY A, DOKANIA P K, TORR P H. Discovering Class-Specific Pixels for Weakly-Supervised Semantic Segmentation[J]. British Machine Vision Conference, 2017.

[74] FAN R, HOU Q, CHENG M.-M, et al. S4Net: Single Stage Salient-Instance Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[75] FAN R, HOU Q, CHENG M.-M, et al. Associating Inter-Image Salient Instances for Weakly Supervised Semantic Segmentation[C] // European Conference on Computer Vision. 2018: 367–383.

[76] CORDTS M, OMRAN M, RAMOS S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3213–3223.

[77] LIN T.-Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common Objects in Context[C] // ECCV. 2014: 740–755.

[78] HONG S, YEO D, KWAK S, et al. Weakly Supervised Semantic Segmentation Using Web-crawled Videos[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3626–3635.

[79] SHEN T, LIN G, SHEN C, et al. Bootstrapping the Performance of Webly Supervised Semantic Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1363–1371.

[80] ROTHER C, KOLMOGOROV V, BLAKE A. Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts[J]. 2004, 23 (3): 309–314.

[81] GUO J, REN T, HUANG L, et al. Video Salient Object Detection via Cross-Frame Cellular Automata[C] // IEEE ICME. 2017: 325–330.

[82] GUO C, ZHANG L. A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression[J]. IEEE Transactions on Image Processing, 2010, 19 (1): 185–198.

[83] DONOSER M, URSCHLER M, HIRZER M, et al. Saliency Driven Total Variation Segmentation[C] // IEEE International Conference on Computer Vision. 2009: 817–824.

[84] ZHANG G.-X, CHENG M.-M, HU S.-M, et al. A Shape-Preserving Approach to Image Resizing[J]. Comput. Graph. Forum, 2009, 28 (7): 1897–1906.

[85] RUTISHAUSER U, WALTHER D, KOCH C, et al. Is Bottom-up Attention Useful for Object Recognition?[C] // IEEE Conference on Computer Vision and Pattern Recognition. Vol. 2. 2004.

[86] BORJI A, FRINTROP S, SIHITE D N, et al. Adaptive Object Tracking by Learning Background Context[C] // IEEE CVPRW. 2012.

[87] ROSIN P L, LAI Y.-K. Artistic Minimal Rendering with Lines and Blocks[J]. Graphical Models, 2013, 75 (4): 208–229.

[88] HAN J, PAUWELS E J, DE ZEEUW P. Fast Saliency-Aware Multi-Modality Image Fusion[J]. Neurocomputing, 2013: 70–80.

[89] CHEN T, CHENG M.-M, TAN P, et al. Sketch2Photo: Internet Image Montage[J]. ACM Transactions on Graphics, 2009, 28 (5): 124:1–10.

[90] ZHU J.-Y, WU J, WEI Y, et al. Unsupervised Object Class Discovery via Saliency-Guided Multiple Class Learning[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2012: 3218–3225.

[91] GAO Y, WANG M, TAO D, et al. 3-D object retrieval and recognition with hypergraph analysis[J]. IEEE Transactions on Image Processing, 2012, 21 (9): 4290–4303.

[92] ABDULMUNEM A, LAI Y.-K, SUN X. Saliency Guided Local and Global Descriptors for Effective Action Recognition[J]. Computational Visual Media, 2016, 2 (1): 97–106.

[93] HOU X, ZHANG L. Saliency Detection: A Spectral Residual Approach[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2007: 1–8.

[94] PERAZZI F, KRÄHENBÜHL P, PRITCH Y, et al. Saliency Filters: Contrast Based Filtering for Salient Region Detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2012: 733–740.

[95] QI W, CHENG M.-M, BORJI A, et al. SaliencyRank: Two-Stage Manifold Ranking for Salient Object Detection[J]. Computational Visual Media, 2015, 1 (4): 309–320.

[96] YANG C, ZHANG L, LU H, et al. Saliency Detection via Graph-Based Manifold Ranking[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3166–3173.

[97] JIANG H, CHENG M.-M, LI S.-J, et al. Joint Salient Object Detection and Existence Prediction[J]. Front. Comput. Sci., 2018.

[98] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet Classification with Deep Convolutional Neural Networks[C] // Advances in Neural Information Processing Systems. 2012: 1097–1105.

[99] LIU Y, CHENG M.-M, HU X, et al. Richer Convolutional Features for Edge Detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3000–3009.

[100] LI J, LIANG X, WEI Y, et al. Perceptual Generative Adversarial Networks for Small Object Detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[101] GIRSHICK R. Fast R-CNN[C] // IEEE International Conference on Computer Vision. 2015: 1440–1448.

[102] ZHANG L, LIN L, LIANG X, et al. Is Faster R-CNN Doing well for Pedestrian Detection?[C] // European Conference on Computer Vision. 2016: 443–457.

[103] NAIR V, HINTON G E. Rectified Linear Units Improve Restricted Boltzmann Machines[C] // International Conference on Machine Learning. 2010: 807–814.

[104] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional Architecture for Fast Feature Embedding[C] // ACM International Conference on Multimedia. 2014: 675–678.

[105] WANG J, JIANG H, YUAN Z, et al. Salient Object Detection: A Discriminative Regional Feature Integration Approach[J]. International Journal of Computer Vision, 2017, 123 (2): 251–268.

[106] BORJI A, CHENG M.-M, JIANG H, et al. Salient Object Detection: A Benchmark[J]. IEEE Transactions on Image Processing, 2015, 24 (12): 5706–5722.

[107] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770–778.

[108] LI X, ZHAO L, WEI L, et al. DeepSaliency: Multi-Task Deep Neural Network Model for Salient Object Detection[J]. IEEE Transactions on Image Processing, 2016, 25 (8): 3919–3930.

[109] LI X, LI Y, SHEN C, et al. Contextual Hypergraph Modeling for Salient Object Detection[C] // IEEE International Conference on Computer Vision. 2013: 3328–3335.

[110] LI X, LU H, ZHANG L, et al. Saliency Detection via Dense and Sparse Reconstruction[C] // IEEE International Conference on Computer Vision. 2013: 2976–2983.

[111] JIANG H, CHENG M.-M, LI S.-J, et al. Joint Salient Object Detection and Existence Prediction[J]. Front. Comput. Sci., 2017.

[112] WANG P, WANG J, ZENG G, et al. Salient Object Detection for Searched Web Images via Global Saliency[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2012.

[113] LIN D, DAI J, JIA J, et al. Scribblesup: Scribble-Supervised Convolutional Networks for Semantic Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3159–3167.

[114] BEARMAN A, RUSSAKOVSKY O, FERRARI V, et al. What’s the point: Semantic segmentation with point supervision[C] // European Conference on Computer Vision. 2016: 549–565.

[115] LI F F, VANRULLEN R, KOCH C, et al. Rapid Natural Scene Categorization in the near Absence of Attention[J]. Proceedings of the National Academy of Sciences, 2002, 99 (14): 9596–9601.

[116] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115 (3): 211–252.

[117] SHIMODA W, YANAI K. Distinct Class-Specific Saliency Maps for Weakly Supervised Semantic Segmentation[C] // European Conference on Computer Vision. 2016: 218–234.

[118] JIN B, ORTIZ SEGOVIA M V, SUSSTRUNK S. Webly Supervised Semantic Segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3626–3635.

[119] KIM D, YOO D, KWEON I S, et al. Two-Phase Learning for Weakly Supervised Object Localization[C] // IEEE International Conference on Computer Vision. 2017: 3534–3543.

[120] 魏云超, 赵耀. 基于 DCNN 的图像语义分割综述[J]. 北京交通大学学报, 40 (4): 82.

[121] MITCHELL T M, COHEN W W, HRUSCHKA JR E R, et al. Never Ending Learning[C] // AAAI. 2015: 2302–2310.

[122] FRÉNAY B, KABÁN A, et al. A Comprehensive Introduction to Label Noise.[C] // ESANN. 2014.

[123] HOU Q, JIANG P.-T, WEI Y, et al. Self-Erasing Network for Integral Object Attention[C] // Advances in Neural Information Processing Systems. 2018: 547–557.

[124] LEE C.-Y, XIE S, GALLAGHER P, et al. Deeply-Supervised Nets[C] // AISTATS. 2015.

[125] PECH-PACHECO J L, CRISTÓBAL G, CHAMORRO-MARTINEZ J, et al. Diatom Autofocusing in Brightfield Microscopy: A Comparative Study[C] // International Conference on Pattern Recognition. 2000.

[126] PERTUZ S, PUIG D, GARCIA M A. Analysis of Focus Measure Operators for Shape-from-Focus[J]. Pattern Recognition, 2013.

[127] CHEN L.-C, PAPANDREOU G, KOKKINOS I, et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[C] // International Conference on Learning Representations. 2015.

致谢

本人由衷地感谢指导教师程明明教授在本人 3 年博士期间在科研上的精心指导以及在职业选择上的悉心帮助与细心规划。是程明明教授的引导，让本人从初入研究的迷茫状态中走向了科研的正轨。在程明明教授的介绍下，本人得以有机会与计算机视觉以及图形学多位资深学者进行合作。在此，本人也感谢牛津大学 Philip Torr 教授、卡迪夫大学 Paul Rosin 教授、加州大学圣地亚哥分校屠卓文教授、新加坡国立大学冯佳时教授等在本人博士攻读期间的指导。

本人感谢三年来一直陪伴我成长的实验室同学，感谢博士期间给予本人帮助的学校工作人员，感谢华为公司研究人员（陈心等）与本人一道将 CVPR 2017 的工作嵌入华为多部智能手机中作为实时人像分割的工具。

最后，本人感谢我的父母以及其他家人在我最为迷茫时对我做出攻读博士学位这一决定后的大力支持，感谢在业余时间陪伴我一起体育锻炼的小伙伴们，感谢那些让我懂得回馈与奉献的好心人们。春去春又来，我依然是那个对科学充满无限热忱的追梦人。

个人简历

个人介绍:

侯淇彬，出生于 1991 年 9 月 23 日。在 2015 年 7 月毕业于东北大学电子与通信工程专业并获得硕士学位。于 2016 年 9 月至今在南开大学就读博士研究生。

研究生期间已发表论文:

1. **Qibin Hou**, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply Supervised Salient Object Detection with Short Connections[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019, 41(4): 815-828. (SCI 源刊, 中科院一区, CCF A 类期刊, 影响因子 9.455.)
2. **Qibin Hou**, Peng-Tao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-Erasing Network for Integral Object Attention[C]. Advances in Neural Information Processing Systems (NeurIPS), 2018, 547-557. (EI 源刊, CCF A 类会议.)
3. **Qibin Hou**, Daniela Massiceti, Puneet Kumar Dokania, Yunchao Wei, Ming-Ming Cheng, Philip HS Torr. Bottom-up Top-down Cues for Weakly-Supervised Semantic Segmentation[C]. Energy Minimization Methods in Computer Vision and Pattern Recognition, 2017, 263-277. (EI 源刊.)
4. **Qibin Hou**, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply Supervised Salient Object Detection with Short Connections[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 3203-3212. (EI 源刊, CCF A 类会议.)
5. Ming-Ming Cheng, **Qibin Hou**, Song-Hai Zhang, and Paul L. Rosin. Intelligent Visual Media Processing: When Graphics Meets Vision[J]. Journal of Computer Science and Technology, 2017, 32(1): 110-121. (SCI 源刊, CCF B 类期刊.)
6. Jiang-Jiang Liu*, Qibin Hou*, Ming-Ming Cheng, Jiashi Feng, Jianmin Jiang. A Simple Pooling-Based Design for Real-Time Salient Object Detection[C]. IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), 2019. (EI 源刊, CCF A 类会议. * 表示并列第一作者)

7. Ruochen Fan, **Qibin Hou**, Ming-Ming Cheng, Gang Yu, Ralph R. Martin, and Shi-Min Hu. Associating Inter-Image Salient Instances for Weakly Supervised Semantic Segmentation[C]. European Conference on Computer Vision (ECCV), 2018, 367-383. (EI 源刊, CCF B 类会议.)
8. Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, **Qibin Hou**, and Ali Borji. Salient Objects in Clutter: Bringing Salient Object Detection to the Foreground[C]. European Conference on Computer Vision (ECCV), 2018, 186-202. (EI 源刊, CCF B 类会议.)
9. Jiaxing Zhao, Bo Ren, **Qibin Hou**, Ming-Ming Cheng, and Paul Rosin. FLIC: Fast Linear Iterative Clustering With Active Search[C]. AAAI Conference on Artificial Intelligence, 2018. (EI 源刊, CCF A 类会议.)
10. Ruochen Fan, Ming-Ming Cheng, **Qibin Hou**, Tai-Jiang Mu, Jingdong Wang, Shi-Min Hu. S4Net: Single Stage Salient-Instance Segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. (EI 源刊, CCF A 类会议.)

研究生期间申请专利:

1. 侯淇彬; 程明明; 白蔚; 周迅溢. 图像显著性物体检测方法和装置 [P]. 中国专利: CN109118459A, 2019-01-01.
2. 程明明; 刘云; 侯淇彬; 白蔚. 图像分割方法及装置 [P]. 中国专利: CN107871321A, 2018-04-03.
3. 刘姜江; 程明明; 侯淇彬; 范登平; 谭永强. 一种基于深度网络的多类型任务通用的检测方法 [P]. 中国专利: CN108428238A, 2018-08-21.

研究生期间参与课题:

1. 场景语义智能识别与理解技术. 天津市新一代人工智能科技重大专项. 项目号: 18ZXZNGX00110.

2. 认知规律启发的弱监督图像场景理解. 天津市杰出青年科学基金. 项目号: 17JCJQJC43700.
3. 3D 多视点全景视频的室内场景重构理论及算法的国际合作研究. 国自科国际合作重点. 项目号: 61620106008.
4. 移动设备上的图像交互式分析与编辑. 国自科面上项目. 项目号: 61572264.