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

基于注意力机制的物体定位与归因研究

Object Localization and Attribution Research Based on Attention Mechanism

论文作者 姜鹏涛 指导教师 程明明教授

申请学位 博士 培养单位 南开大学

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 评 阅 人

南开大学研究生院

二〇二二年四月

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	基于注意力机制的物体定位与归因研究				
姓　　名	姜鹏涛	学号	1120190166	答辩日期	
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	17320260072		电子邮箱	pt.jiang@mail.nankai.edu.cn	
通讯地址 (邮编)：300000					
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

在图像识别领域，卷积神经网络中的注意力机制受到了大量的关注。注意力机制从卷积神经网络中为预测类别生成注意力图，随后通过注意力图定位到物体上的判别性区域。由于注意力图具有物体定位的能力，它可以被用于基于类别标签的弱监督任务中去，减小数据标注成本。此外，判别性区域代表了对于分类网络决策起到重要作用的输入特征，这种特性使得注意力图可以被用于卷积神经网络的归因研究。

因为注意力图被广泛应用于弱监督任务和网络归因任务中，所以本文对它存在的以下三种问题进行深入研究。1）定位的物体区域粗糙：注意力图从网络深层特征生成，分辨率较低。当上采样到输入图像的尺寸时，注意力图定位的物体区域非常粗糙，很难满足弱监督任务对于准确物体区域的需求。2）定位的物体区域不完整：注意力图只能定位到物体上的判别性区域，区域较小，很难满足弱监督任务对于完整物体区域的需求。3）归因能力有限：注意力图只能归因输入和输出之间的关系，无法归因网络中的特征通道在决策过程中起到的作用。

针对于以上问题，本文改进注意力机制来提升它在不同任务中的表现。对于弱监督任务，本文从定位物体区域的完整性和准确性出发，提出了两种物体定位方案来提升弱监督任务的性能。对于网络归因任务，本文深入卷积神经网络内部，提出一种改进的注意力方案来剖析网络内部特征通道的重要性。本文的具体贡献如下：

1. 针对于注意力图定位的物体区域粗糙的问题，提出了一种基于层次化注意力的物体定位方法，这种方法可以从网络不同层生成注意力图。考虑到浅层特征空间上像素差异较大，本文采用局部权重来代替全局权重代表每个像素对于决策的重要性。本文将从浅层注意力图定位的物体细节信息和深层注意力图定位的一般位置信息结合起来，更好的定位物体边界。在 ILSVRC 数据集的验证集上，该方法可以取得 47.24% 的 Top-1 定位分数。
2. 针对于注意力图定位的物体区域不完整的问题，提出了一种基于在线注意力累积的物体定位方法。利用分类网络在不同训练时刻的注意力图定

位的物体区域互补的特性，将这些注意力图定位的物体区域记录到累积注意力图中。为了让注意力移动的范围尽可能大，本文引入了一个注意力遮挡层来进一步提升定位的完整性。此外，通过完整注意力学习进一步增强累积注意力图中注意力值小的区域。在 PASCAL VOC 2012 基准的测试集上，该方法可以取得 67.2% 的 mIoU 分数。

3. 针对于注意力图归因能力有限，提出了一种基于层次化注意力分解的归因方法。具体地，本文设计了一个基于梯度的激活回传模块，将网络决策从深层依次向浅层分解，得到一组强相关的层次化证据，用于解释特征通道对于决策的影响以及特征通道之间的联系。

关键词： 卷积神经网络；注意力机制；物体定位；归因；

Abstract

Attention mechanisms in convolutional neural networks have received extensive attention in the field of image recognition. Such methods generate attention maps for the predicted classes from the classification network, which are then used to localize discriminative object regions. Due to the localization ability of the attention maps, they can be used in weakly supervised tasks based on class labels, reducing the cost of data annotation. In addition, the discriminative regions represent the input features that play an important role in the decision-making process of the classification network, which allows attention maps to be used in attribution studies of convolutional neural networks.

Because attention maps are widely used in weakly supervised tasks and network attribution tasks, this paper makes an in-depth study on the following three problems. 1) Rough object regions: the attention map with a low resolution is generated from the deep features of the network. When the attention map is up-sampled to the size of the input image, the object regions localized by the attention map are very rough. It is difficult to meet the needs of weakly supervised tasks for accurate object regions. 2) Incomplete object regions: the attention map can only locate the discriminative object regions, where the regions are small. It is difficult to meet the needs of the weakly supervised task for the complete object regions. 3) Limited attribution ability: the attention map can only attribute the relationship between input and output but cannot attribute the role of feature channels in the decision-making process.

To deal with the above problems, this paper improves the attention mechanism to improve its performance in different tasks. For weakly supervised tasks, this paper proposes two schemes to improve their performance, starting from the integrity and accuracy of the localized object regions. For the network attribution task, this paper goes deep inside the network and proposes an improved attention scheme to dissect the importance of feature channels inside the network.

The specific contributions of this paper are as follows:

1. To deal with the problem of the rough object regions localized by the attention

map, this paper proposes an object localization method based on hierarchical attention, which can generate attention maps from different layers of the network. Considering the large pixel difference in the shallow features, this paper uses local weights instead of global weights to represent the importance of each pixel for decision-making. Besides, this paper combines the object detail information localized by the shallow attention map with the general location information localized by the deep attention map to better locate the object regions. On the validation set of ILSVRC dataset, this method can obtain 47.24% top-1 localization score.

2. To deal with the problem of incomplete object regions localized by the attention map, this paper proposes an object localization method based on an online attention accumulation strategy. Taking advantage of the complementary characteristics of the object regions localized by the attention maps at different training times, the object regions localized by these attention maps are accumulated into the cumulative attention map. To make the range of attention movement as large as possible, this paper introduces an attention drop layer to further improve the localization integrity. Furthermore, attention regions with small values in the cumulative attention map are further enhanced by integral attention learning. On the test set of Pascal VOC 2012 benchmark, this method can obtain 67.2% mIoU score.
3. To deal with the limited attribution ability, this paper proposes an attribution method based on hierarchical attention decomposition. Specifically, this paper designs an efficient gradient-based activation propagation module, which decomposes network decisions from deep layers to shallow layers iteratively, and obtains a set of strongly correlated hierarchical evidence, explaining the influence of feature channels on decision-making and the relationship between feature channels.

Key Words: Convolutional neural network; attention mechanism; object localization; attribution;

目录

第一章 绪论

第一节 研究背景与意义

随着智能手机、高性能计算机和嵌入式设备的普及，图像和视频已经成为人类获取信息的重要手段。如何让计算机像生物视觉系统一样，快速分析并理解图像，辅助人类的生产生活，已成为智慧生活、智慧医疗和智慧城市等领域的研究热点。在计算机中，图像的表达形式是二进制数值矩阵，计算机很难理解这一堆数字组合的含义。早期，研究者们利用手工设计的特征提取算子来从图像矩阵中获取视觉特征，根据提取的特征对物体进行定位识别与理解。然而，由于所提取的特征的表达能力较差，在实际场景中应用效果并不理想。近年来，高性能图像处理器（GPU）出现使得训练一个参数量很大的卷积神经网络变成现实。相比于手工设计的特征提取算子，卷积神经网络拥有更强的特征提取能力。在计算机视觉领域，卷积神经网络在图像识别[1–3]、物体检测与跟踪[4–6]和语义理解[7–9]等任务上取得了突破性的提升，被广泛应用到自动驾驶、安防系统和医学诊断等重要领域。

卷积神经网络在计算机视觉领域虽然获得了巨大的成功，但它有两个明显的缺陷限制了它的实际应用。第一是卷积神经网络的训练**需要大量的数据标签**，而获取数据标签通常需要耗费巨大的人力物力。在图 1.1中，本文给出了几种标签的示例。由 Zhang 等人[10] 指出，标注一个类别标签需要耗时 1 秒, 标注一个物体边界框需要耗时 10 秒, 标注一个物体分割标签需要耗时 78 秒。随着标注难度的增大，标注时间也急剧增加。当在实际场景中应用一个语义分割算法时，我们首先需要获取实际场景的训练数据，随后对图像数据进行精细的像素级标注，最后使用像素级标签学习语义分割模型，对于图像的精细标注会大大增加卷积神经网络的应用成本。因此，研究者们尝试使用标注成本较低的弱标签，例如类别标签，来进行模型学习，从而达到减少标注成本的目的。然而，基于类别标签学习的模型性能通常很差，因为类别标签没有提供任何物体位置信息。如何在弱标签的基础上获取物体位置先验对于弱监督任务至关重要。卷积神经网络的第二个缺陷是它的**决策过程不透明**。随着卷积神经网络层数的不断加深，模

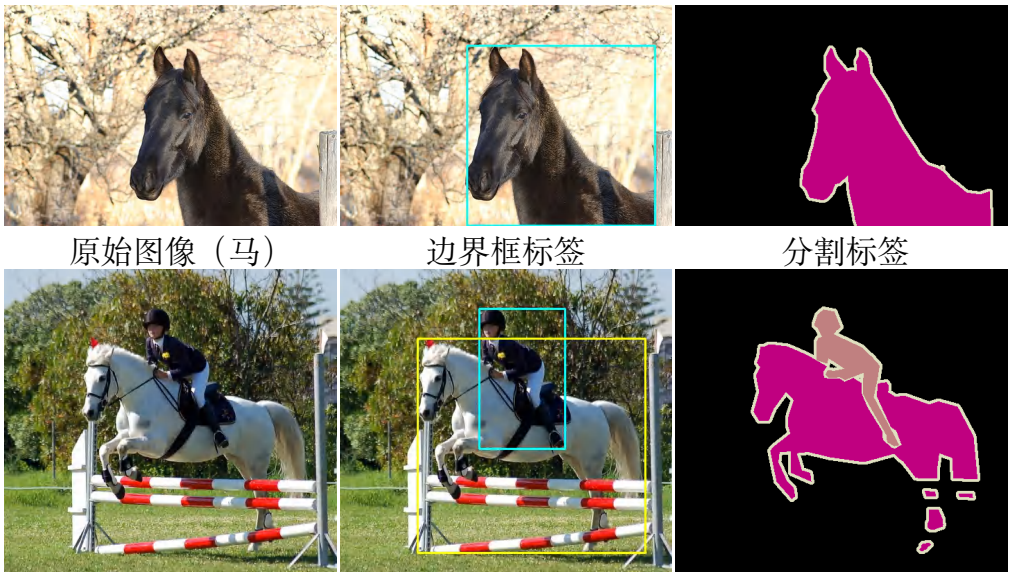


图 1.1 图像的各种标签示例。从类别标签到边界框标签，再到分割标签，标注难度不断增大。

型的非线性度也会随之增加，其决策过程也更加难以分析。当将基于卷积神经网络的算法应用在重要的领域时，例如自动驾驶、医疗诊断等领域，网络做出的决策很难获得人类信任。因此，卷积神经网络的归因研究对于其在重要领域的应用也是至关重要的。

在认知心理学领域，Treisman 等人[11] 提出了特征整合理论来解释生物视觉系统感知机制。该理论将视觉感知过程分为两个阶段，分别是前注意阶段和集中注意力阶段。在前注意阶段，大脑提取视觉感受野中的基础特征，例如颜色，形状以及运动信息等。在集中注意力阶段，大脑通过注意力关注到物体上并结合特征进行感知。注意力机制可以帮助大脑从大量的视觉特征中快速定位物体特征进行感知。卷积神经网络模拟生物视觉感知系统，同样也基于一个相似的规律。Zhou 等人[12] 在 2016 年提出了一种注意力模型，该模型可以从基于卷积神经网络的分类模型中生成注意力图，也叫类别激活图。如图 1.2中红框所示，当分类模型识别图中的物体的类别时，注意力图中的高响应区域通常定位在物体上的某一部分。

分类网络训练时使用的类别标签只提供了图像中的物体类别信息，没有使用任何物体位置信息的监督。通过注意力模型，我们可以从分类网络中生成的注意力图来定位物体区域。由于注意力图的物体定位能力，它被广泛应用于弱

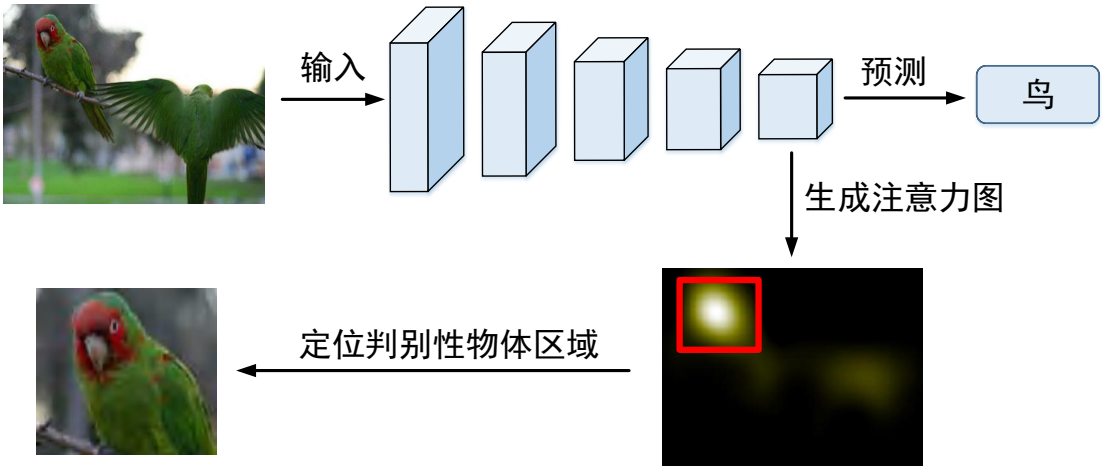
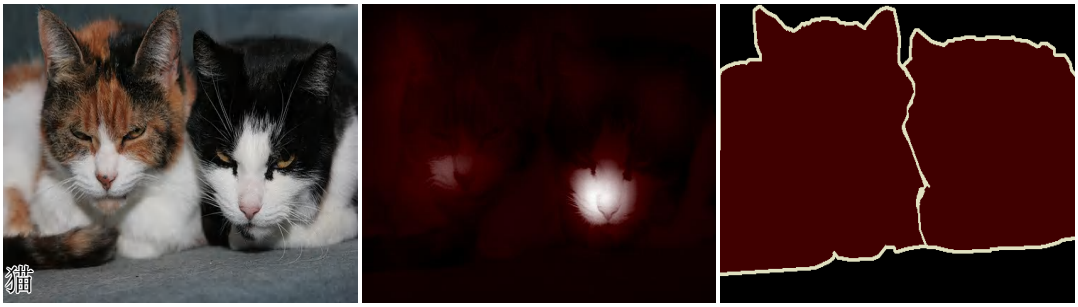


图 1.2 分类网络中注意力图的概念。注意力图中的红色框表示判别性物体区域。

监督任务中[13, 14]，例如基于类别标签来定位物体边界或者语义分割，在一定程度上节省了数据的标注成本。此外，注意力图中的高响应区域通常代表着对分类结果有着重要影响的图像区域。在这个例子中，分类模型预测图像中存在鸟这个类别主要是根据输入图像中左侧鸟的头部特征来判定的。注意力图定位判别性物体区域的能力在一定程度上解释了卷积神经网络的输入与预测之间的关系，可以用于卷积神经网络的归因。注意力图的**物体定位能力和归因能力**可以被用于缓解前文提到的卷积神经网络的两个缺陷。具体来说，注意力图主要有以下几种具体用途：

1. 基于类别标签的弱监督物体定位任务。在这类任务中，算法需要通过类别标签来检测到目标物体的边界框。注意力图定位物体边界框的一般流程是先二值化注意力图，随后找最大连通体的外接矩形框作为物体的定位结果。
2. 基于类别标签的弱监督语义分割任务。在这类任务中，算法需要通过图像的类别标签来对图像中每一个像素做分类。研究者们通常会用注意力图定位的判别性物体区域为基础，进一步挖掘更多的物体区域。
3. 基于注意力图的卷积神经网络的归因。在这类研究中，研究者们使用注意力图定位的物体区域解释分类网络的决策，增加人类对深度神经网络模型的信任。



原始图像　　注意力图　　物体区域

图 1.3　注意力图的示例。注意力图中的高响应区域定位在物体上。与完整物体区域相比，注意力图定位的物体区域粗糙且不完整。

第二节　研究现状

注意力机制虽然在弱监督任务和卷积神经网络归因任务中被广泛应用，但是其在两种任务中的应用仍然有许多难点，主要表现在于以下几个方面。

一、注意力图定位的**物体区域粗糙**。已有注意力模型，例如工作[12, 15, 16] 都是从卷积神经网络深层特征来生成注意力图。由于网络深层特征分辨率较低，在上采样放大到原始图像尺寸时，注意力图只能定位非常粗糙的物体区域，如图 1.3所示。浅层特征分辨率较高，但已有注意力模型[17] 生成的注意力图噪声很大。弱监督任务需要准确且完整的物体区域信息。例如弱监督物体定位需要对物体边界进行精确定位，弱监督语义分割需要对物体上的每个像素进行正确分类。因此，如何从浅层特征生成可靠的具有细节信息的注意力图来帮助深层注意力图更准确的定位物体，是本文的一个研究难点。

二、注意力图定位的**物体区域不完整**。已有的注意力模型，例如工作[12, 15, 16] 生成的注意力图定位只能定位到物体的部分区域，如图 1.3所示。注意力图的定位能力有限，很难定位到完整的物体区域，限制了弱监督任务的性能提升。Wei 等人[14] 和张等人[13] 尝试使用对抗擦除策略，他们将注意力图定位的区域从图像或特征中擦除，让网络生成的注意力图关注到物体上的其它区域，从而定位到更多的物体区域。然而，随着训练的不断进行，注意力图经常会关注到背景区域上，使得定位区域变得不准确。此外，Wei 等人[18] 也尝试使用空洞卷积来定位更多的物体区域，然而空洞率较大的卷积会定位到大量的背景区域，同样会引入很多噪声。因此，如何让注意力图更好的定位到完整的物体区域，是本文的另一个研究难点。

三、注意力图**归因能力有限**。注意力图定位的判别性物体区域代表着对于

分类网络决策重要的图像区域，在一定程度上可以帮助人类理解网络决策。但是仅通过注意力图，人类无法进一步分析网络内部的运行机制，即解释网络内部特征通道对于决策的影响以及特征通道之间的影响。此外，已有的归因方法[19–21] 都是专注于学习输入与输出之间的关系，同样无法深入理解卷积神经网络内部的机制。一些注意力方法[22, 23] 虽然可以归因特征通道与网络决策关系，但是特征通道之间的联系仍然无法得到。因此，如何利用注意力图来归因网络特征通道以及特征通道之间的联系是本文的另一个研究难点。

第三节　研究目标和主要贡献

在前文中，本文已经提到了从分类网络生成的注意力图被广泛应用于弱监督任务和卷积神经网络的归因。针对注意力机制在不同领域应用的难点，本文分别对注意力机制进行改进，以此来适应不同任务的需求。本文的主要研究工作之间的关系如图 1.4所示。

在第二章中，本文首先回顾了注意力机制的相关工作的研究现状，随后又给出了注意力机制的不同应用领域的研究现状，即弱监督物体定位、弱监督语义分割和卷积神经网络归因研究。

在第三章中，针对注意力图**定位的物体区域粗糙**的问题，本文提出了基于层次化注意力的物体定位方法。由于浅层特征分辨率较高，包含有大量的细节信息，本文提出利用神经网络浅层特征来生成物体位置的细节信息，将深层定位的大致位置信息和浅层定位的细节信息结合起来得到更准确的物体区域。然而，直接将已有注意力模型应用在浅层特征上生成的注意力图噪声很大，无法准确的提供物体的细节信息。因此，本文重新考虑了注意力图的生成方法，利用局部权重替代原有的全局权重，这种方式可以充分的考虑特征图中每个点的重要性。此外，本文考虑了负梯度对于注意力图定位的缺陷，并将它的影响去掉。

在第四章中，针对注意力图**定位的物体区域不完整**的问题，本文提出了基于在线注意力累积的物体定位方法来定位完整的物体区域。由于在分类网络训练过程中，不同训练时刻分类模型生成的注意力图定位的区域有所不同，并且互相补充，因此在线注意力累积策略将这些注意力图融合在一起生成累积注意力图。此外，为了让不同时刻的注意力图定位的物体区域尽可能的不同，本文提出了一个注意力遮挡层，它通过对已经定位到的物体区域进行遮挡，增大注意力在物体上的移动范围，从而使分类网络生成的注意力图定位到更多的物体区

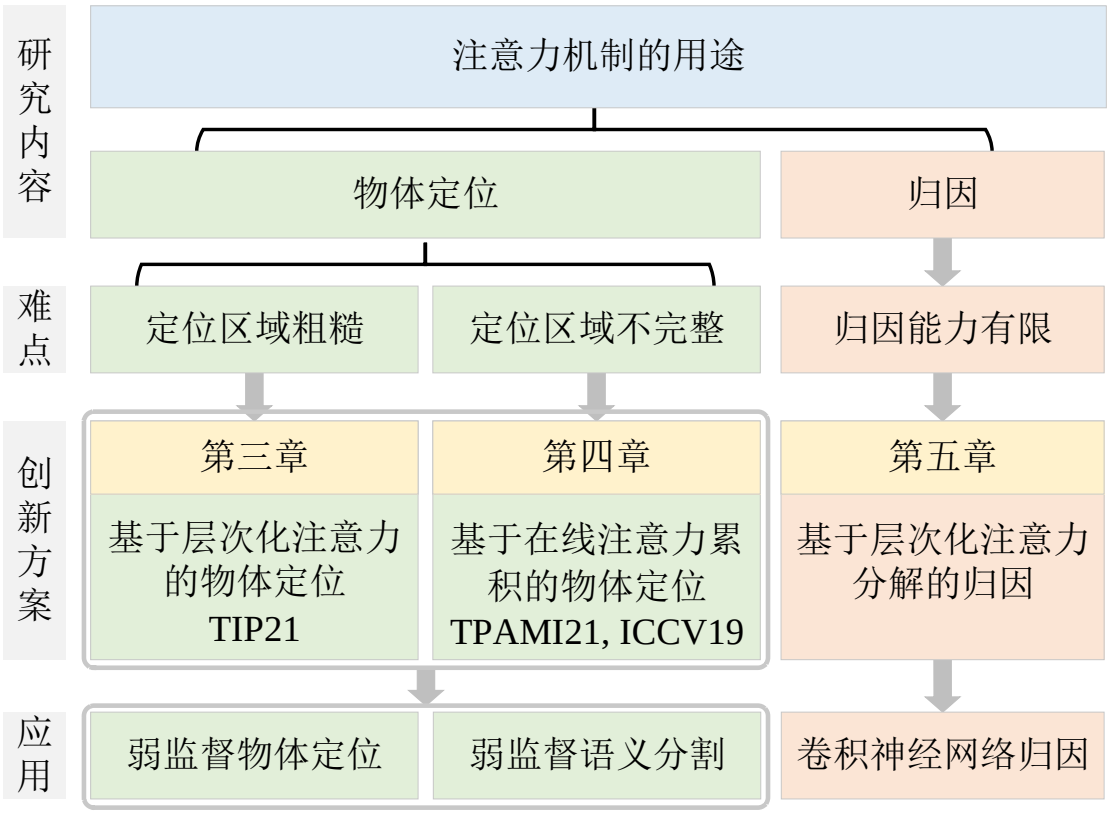


图 1.4 本文的组织结构和主要研究工作之间的关系。本文工作围绕注意力机制，研究注意力机制在不同应用领域的遇到的难点并提出具体改进方案。

域，使最终的累积注意力图挖掘到完整的物体区域。本文还提出了一种完整注意力学习策略来增强累积注意力图中激活较弱的区域。具体地，通过改进传统的 sigmoid 交叉熵损失函数，本文提出了一种新型混合损失函数，其中包含两项损失，增强损失和约束损失，来进一步增强累积注意力图的定位区域的完整性。

在第五章中，针对注意力图**归因能力有限**的问题，（即不能归因网络内部特征通道之间以及特征通道和预测之间的关系），本文提出了基于层次化注意力分解的归因方法。具体地，本文首先提出了一个基于梯度的激活回传模块，它可以将任何神经网络层的决策向浅层分解，并从浅层生成支撑证据（即重要特征通道的注意力图）。基于激活回传模块，本文将决策迭代的向底层分解，直到网络最底层。由于向下分解的过程中，会产生超过人类认知负荷的支撑证据，因此需要对整个分解过程进行简化。首先，本文通过选择卷积神经网络每个阶段的最后的一个卷积层来代替这个阶段，这样选择的原因是同一阶段特征具有相同的分辨率，获取的信息在一个尺度上。其次，对于每个卷积层，只选择最重要的

几个特征通道向下分解来减小冗余。通过注意力的层次化分解，可以得到一个支撑决策的层次化的证据，在这个层次化证据中，特征通道之间是互相关联的。

在第六章中，本文分别对第三、四和五章提出的方法进行总结，并对于每个方法提出可以继续改进的方向。

第四节　本文的组织结构

第二章将介绍相关工作的研究现状。第三章详细分析已有注意力方法在网络浅层定位失败的原因，并介绍基于层次化注意力的物体定位方法。最后通过弱监督物体定位和语义分割实验验证本文方法的有效性。第四章介绍基于在线注意力累积的物体定位方法，包括在线注意力累积机制，注意力遮挡层以及完整注意力学习策略。第五章引入基于层次化注意力分解的归因方法。首先介绍基于梯度的激活传播模块，并详细回顾了和已有的注意力方法的区别，最后将基于梯度的激活传播模块用于层次化注意力分解来生成层次化解释。第六章对本文的工作进行了总结和展望。

第二章 相关工作介绍

在本章中，本文分别对注意力机制以及注意力机制的不同应用领域进行相关工作介绍，方便读者了解国内外研究现状。在第一节，本文首先介绍注意力机制的相关工作。在第二节中，本文将介绍弱监督物体定位领域的相关工作。第三节介绍弱监督语义分割领域的相关工作。第四节介绍图像识别领域中卷积神经网络归因的相关工作。第五节介绍本文所使用的数据集。

第一节 注意力机制研究现状

近年来，学者们提出了许多注意力机制的方法，它们利用基于卷积神经网络的分类模型[1, 24–26] 来生成注意力图，进而通过注意力图中的高响应区域定位判别性物体区域。弱监督任务[27–32] 以及网络归因等任务，都受益于注意力图的定位能力。注意力机制的方法可以大体分为以下几类：基于反向传播的方法、基于类激活映射的方法以及基于特征扰动的方法等。值得注意力的是有一些注意力方法可能同时属于几个类，本文将这些方法只分配到一个类里。下面将分别对每一类方法的发展现状进行介绍。

2.1.1 基于反向传播的方法

这一类方法利用反向传播机制将梯度传到输入层，来生成注意力图，也叫做归因图。在早期，Sung 等人[33] 通过敏感性分析、模糊曲线和均方误差的变化三种工具来对反向传播网络的不同输入变量的重要性进行排序。Baehrens 等人[34] 通过计算决策函数的梯度来识别特定实例的特征重要性。Simonyan 等人[35] 反向传播预测相对于输入图像的梯度，利用梯度生成一个注意力图，指示图像中每个像素的重要性。以上方法的本质都是基于非线性分类器的输出相对于输入的偏导数。Springenberg 等人[36] 和 Zeiler 等人[37] 在反向传播通过 ReLU 层时，使用不同的反向传播逻辑，它们的共同点是都将负梯度置为零。Sundararajan 等人[20] 考虑反向传播的方法存在饱和度和阈值问题，他们提出沿从基础图像到输入图像的路径累积梯度来计算注意力图。

另一组方法[19, 38–40] 提出了不同的自上而下相关性的反向传播规则。Bach

等人[19] 提出了一种通用的理解分类决策的方案，即像素级分解，计算输入图像的每个像素对于决策的贡献。Zhang 等人[41] 提出了一种自上而下的反向传播方案，c-MWP，它基于赢家通吃策略将信号从网络顶层传递到网络底层。赢家通吃策略只保留和决策最相关的特征通道。Montavon 等人[42] 提出了 DeepTayor 算法，将深度泰勒展开算法应用在反向传播中。Shrikumar 等人[21] 通过比较神经元的激活和参考激活的差异来计算神经元对于决策的贡献并将其反向传播。

不同于以上手工设计回传规则，Yang 等人[43] 尝试自动学习反向传播规则来生成归因图，这种方法的好处是自动学习的规则对于模型和输入都具有很强的鲁棒性。Smilkov 等人[44] 锐化基于梯度的归因图以减少视觉噪声，增强用户对于归因图的信任。此外，一些方法[22, 23] 也尝试用基于反向传播的方式来测量网络内部特征通道对预测的重要性。这些方法可以从卷积神经网络的不同层中找出最重要的特征。Kim 等人[45] 通过研究高级概念来解释卷积神经网络的内部状态。他们利用方向导数来量化高级概念对分类结果的重要性。

2.1.2 基于类别激活映射的方法

这类注意力方法[12, 16, 17, 46] 基于卷积神经网络最终卷积层的输出特征来生成注意力图。在图 2.1中，本文展示了这类方法的一般流程。这类方法对所有的特征图进行加权求和生成注意力图。作为早期的尝试，Zhou 等人[12] 首先提出了类别激活映射（CAM）的方法。该方法应用在特殊设计的分类网络，即将 VGG-16 网络[2] 的第一个全连接层替换成全局平均池化层，特征在经过池化层后，连接到一个全连接层来输出每个类别的预测得分。由于全连接层的权重和最后的卷积层的特征是一一对应的，所以 CAM 结合权重和卷积层的特征，来生成特定类别的注意力图。

从注意力图中，可以定位到语义物体上的判别性区域，它预示着输入图像上对于分类网络决策起到重要作用的区域，可以用来作为可解释性的工具。此外，从注意力图中定位的语义物体区域还可以被用于基于类别标签的弱监督任务中去，作为初始的物体位置先验。随后，基于 CAM，Selvaraju 等人[15] 提出了一项新技术，叫 Grad-CAM，其通过将梯度反向传播到最后的卷积层来生成注意力图。Grad-CAM 通过对特征图的梯度取平均值作为该特征图的权重。不同于 CAM 只能用于图像分类，Grad-CAM 可以为任意一种目标概念生成视觉解释，例如图像分类，视觉问答和图片描述等任务。Chattopadhay 等人[16] 提出了 Grad-CAM++，他们通过考虑激活图中每一个像素的重要程度，进一步提高了

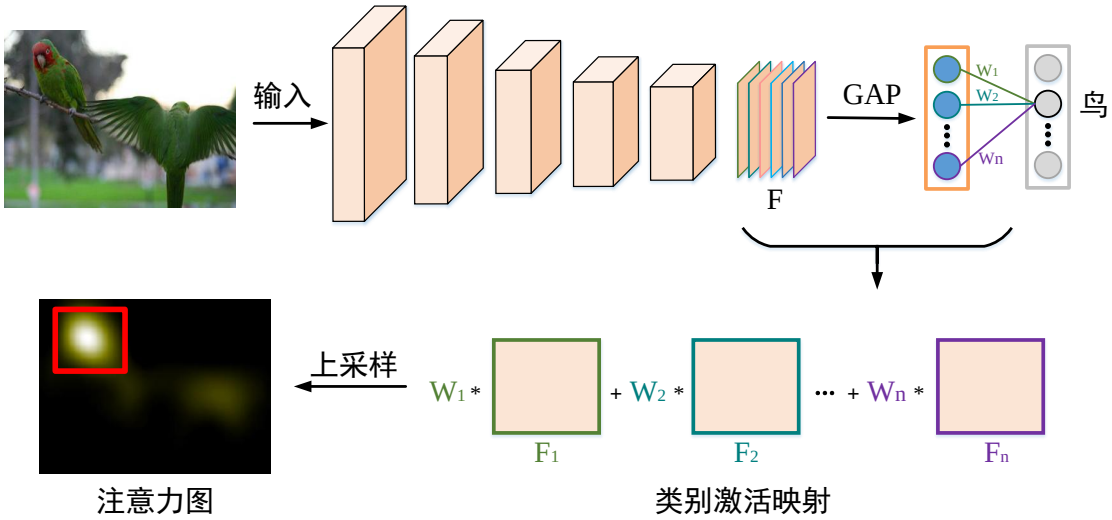


图 2.1 基于类别激活映射的注意力方法的一般流程。F 表示分类网络最后一个卷积层的输出特征。GAP 表示全局平均池化层。$W_1, W_2, ..., W_n$ 是连接类别“鸟”和特征图之间的权重。对于类别“鸟”，特征图加权后经上采样得到注意力图。概念图的绘制参考以前的一些工作[12, 47]。

Grad-CAM 的定位能力。Grad-CAM 和 Grad-CAM++ 相比于 CAM 虽然不再局限于特定网络和特定任务，但是它们在生成注意力图的时候依赖梯度，而卷积神经网络需要执行一次前传和反传才能得到梯度，这样严重限制了 Grad-CAM 和 Grad-CAM++ 在其它任务上的应用。

Zhang 等人[13] 提出了一种和 CAM 等效的方法，他们将最后的卷积层的通道数设置和数据集中的类别数相同，这样每个通道对应一个语义类别，进而将这个卷积层的特征输入到全局平均池化层得到分类概率。每个类的注意力图可以从对应的特征通道生成，这种方法使得注意力图在训练过程中很容易得到，有利于注意力图的应用。此外，Score-CAM[46] 通过网络前向推理分数而不是梯度来计算每个激活图的权重。最近，不同于以上用于定位最具有判别性区域的注意力方法，一些工作[13, 18, 48–50] 尝试生成可以定位完整语义物体区域的注意力图，这对弱监督任务是有益的。本文将在弱监督物体定位和语义分割小节介绍这一类方法。

2.1.3 基于特征扰动的方法

这些方法通过扰动输入来观察输出的变化。Zeiler 等人[37] 滑动一个灰度正方形来遮挡输入图像的不同区域，观察深层特征图和输出的变化。Zintgraf 等人[51] 不仅识别支持网络决策的重要区域，而且识别对决策起相反作用的区域。

不同于以上均匀采样的方法，Petsiuk 等人[52] 随机采样图像区域并遮挡它观察输出的变化。Ribeiro 等人[53] 利用超像素块来选择被遮挡的图像区域，他们通过学习一个局部线性模型来计算每个超像素块的贡献。此外，最近的方法[54–56] 尝试学习了一个扰动图，当将扰动图应用于输入图像时可以最大程度地影响预测。此外，Fong 等人[55] 也采用输入归因方法研究深层网络的重要通道。

第二节 弱监督物体定位研究现状

弱监督物体定位任务在仅使用类别标签的条件下定位目标物体的边界框。早期一些研究人员[57–60] 试图将弱监督物体定位任务建模为多实例学习框架。另一类方法[61–64] 首先生成大量的物体候选框，并从物体推荐框候中选择最合适的矩形框作为物体的边界框。其中，Teh 等人[63] 利用注意力网络为每个候选框计算注意力分数。Zhu 等人[65] 首先提出了软候选框的概念，软候选框实际上是一个概率图的概念，他们根据像素之间的关系不断更新软候选框。

Zhou 等人[12] 首先将注意力图用于弱监督物体定位任务，他们将注意力图二值化后，随后用最大连通体的外接矩形作为目标物体的边界框。由于注意力图的定位能力，它被广泛应用于弱监督物体定位中。然而，由于从分类网络生成的注意力图通常会定位到小的判别性区域，其定位性能是有限的。近年来，研究者们提出了很多基于注意力机制的弱监督物体定位的方法[13, 17, 50, 66, 67]。他们通过改进注意力图的定位能力来提升弱监督物体定位任务的性能。Singh 等人[67] 提出了“遮挡-寻找”策略。他们在训练过程中随机遮挡一部分图像块，迫使分类网络从物体的其它相关区域进行识别，从而让注意力图发现物体上的其它区域。不同于这种随机遮挡策略，Zhang 等人[13] 对判别性区域进行遮挡，这种方式显式的让网络关注到其它物体区域。具体来说，他们在卷积神经网络上连接了两个分类分支，从一个分支生成注意力图，将强响应对应的区域从另一个分支中的输入特征中擦除，以此来从另一个分支中定位到新的物体区域。最后他们将两个分支的注意力图结合起来定位物体边界框。Xue 等人[66] 利用卷积神经网络的层次化特征来学习不同的层次激活区域，并将它们结合起来生成最终的注意力图。

第三节 弱监督语义分割研究现状

近年来，由于很多优秀的方法的提出，弱监督语义分割任务取得了长足的进步。弱监督语义分割任务常用标签主要有以下几种：边界框[68]、草图[69]、点[70]、和类别标签[71]。本文主要介绍基于类别标签的弱监督语义分割方法的研究现状，帮助读者快速了解以往的经典方法。早期的研究者利用多示例学习和期望最大化算法端到端的训练分割网络。后来，随着注意力机制的发展，注意力图的定位能力被广泛应用在弱监督语义分割领域。主流的大部分方法[18, 72–74] 都是基于注意力图设计的弱监督语义分割方案，他们通常利用注意力图定位的语义区域作为初始的物体位置先验。除了注意力机制，一些研究者[18, 31, 74] 还发现自底向上的显著性物体线索[75–78] 对于提取背景区域和物体形状信息非常有用。下面将对几种经典的弱监督语义分割方法进行介绍，分别为基于端到端的方法、基于种子点扩张的方法、基于对抗擦除策略的方法以及基于语义关系学习的方法。

2.3.1 基于端到端的方法

在早期，Pinheiro 等人[71] 将弱监督语义分割算法建模为多示例学习，他们将最后的卷积层输出的特征图聚合成分类概率。Papandreou 等人[79] 提出用期望最大化算法端到端的训练分割网络，他们首先用预测的得分图生成分割标签，进而用分割标签监督得分图。Zhang 等人[80] 提出联合训练分类和分割任务，在训练过程中，利用从分类网络生成的注意力图来制作伪分割标签，然后利用伪分割标签训练分割网络。Araslanov 等人[81] 构建了一个自监督的分割网络，他们利用图像的外观先验来优化从得分图生成的分割标签并用分割标签训练网络。

2.3.2 基于种子点扩张的方法

Kolesnikov 等人[82] 首先提出了种子点扩张的思想。他们用注意力图提取语义物体区域并用显著性图提取背景区域，制作定位先验图。由于这两种图定位的物体和背景区域较小，定位先验图也可以称为种子点图。随后他们引入了 3 种损失函数，分别名为种子损失、扩张损失和边界限制损失。扩张损失让初始种子区域扩张，边界限制损失约束分割图的边界和图像中物体的边界吻合。Huang 等人[73] 改进了这一思路，他们用经过种子区域增长算法扩张后的定位先验图来作为语义分割的像素级监督。

2.3.3 基于对抗擦除策略的方法

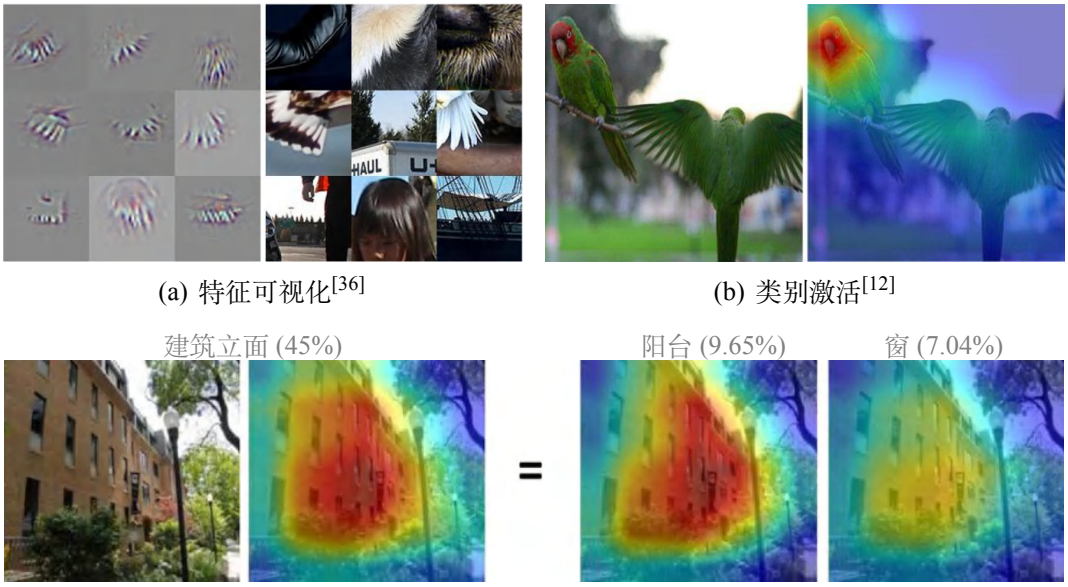
基于对抗擦除策略的方法的基本思想是将注意力图定位的物体区域从图像或者特征中遮住，让分类网络从图像中物体的其它区域进行识别，进而使得新生成的注意力图定位到新的物体区域。Wei 等人[14] 提出了一种对抗擦除策略（AE-PSL）来逐步挖掘物体的不同区域。从一个小的初始区域开始，他们通过不断的将遮挡的图像送入分类网络生成注意力图，来定位新的物体区域。最终他们将所有的注意力图融合到一起定位更加完整的物体区域。AE-PSL 的过程相当复杂，需要重复训练多个分类模型以获得不同的物体区域。不同于以上方法使用固定阈值来从注意力图中选定被擦除的图像区域，Li 等人[48] 使用一个软擦除策略。他们改进了对抗擦除策略，通过损失函数让分类网络从遮挡后的图像学习不到任何类别信息，从而迫使网络将注意力集中于目标整体上。Hou 等人[49] 观察到对抗擦除策略会让注意力扩张到背景区域上，因此他们提出了一种自擦除策略来抑制注意力向背景区域扩张。

2.3.4 基于语义关系学习的方法

注意力图可以定位到部分物体区域，并将对应的语义类别分配给这个区域里所有像素。然而，图像中仍然有大部分像素的语义类别是未知的。Ahn 等人[72] 提出从具有语义的像素向周围区域的未知像素传递语义。为了能够学习语义像素和未知像素之间的关系，Ahn 等人首先利用注意力图去构建了一种语义关系标签。随后通过神经网络学习这种语义关系，最后根据学习到的语义矩阵将已知像素的语义通过随机游走算法传递给未知像素。后来，Ahn 等人[83] 改进了这一策略，他们学习每个物体的大致位置和边界，并在物体的边界内传递语义关系。

第四节 卷积神经网络归因研究现状

归因研究一直是卷积神经网络研究的一大热点。近年来，研究者们在归因研究的几个不同方面取得了重大进展，包括特征归因、特征可视化、可解释模型的知识蒸馏和内在可解释模型。下面将会依次介绍它们的相关工作。



(a) 特征可视化[36] (b) 类别激活[12]

(c) 高层分解[84]

图 2.2 不同可解释性方法的说明。

2.4.1 特征归因

特征归因方法通常会生成一个注意力图（归因图）来解释输入图像是如何影响网络输出的。本文将特征归因方法大体分为三类: 基于反向传播的方法、基于扰动的方法和基于类别激活的方法。对于这几类方法，本文已经在前文注意力机制小节进行了详细的介绍。此外，Zhou 等人[84] 尝试将网络决策分解为几个语义组件并研究每个组件的贡献。如图 2.2 (c) 所示，“建筑立面”主要被分解为“阳台”和“窗”。

上述归因方法主要集中在生成注意力图（归因图）来研究输入如何影响输出预测。尽管一些归因方法可以衡量中间特征对输出预测的重要性，但它们通常忽略了不同中间特征之间的关系。正如 Olah 等人[85] 指出的那样，不同中间特征之间的关系对于解释网络预测也很重要。本文不仅分解了网络决策，还分解了中间特征来从它们的前几层中找到支持证据，解释这些相关的中间特征如何相互影响。虽然 LRP[19] 方法将特征重要性传播到中间特征，但不同通道的特征重要性在反向传播过程中是耦合的。该方法为整个网络行为生成简单的解释，而不是层次化解释。

2.4.2 特征可视化

可视化卷积神经网络中间层的特征可以深入了解这些层的学习内容。对于卷积神经网络的第一层，可以直接将其三通道权重投影到图像空间中。为了可视化更高层的特征，研究人员提出了许多替代方法。其中，Erhan 等人[86] 和 Simonyan 等人[35] 利用梯度上升算法在图像空间中找到最大化神经元激活的最优激励。其它方法[36, 37, 87] 从数据集中识别图像块，以最大化神经元激活，如图 2.2 (a) 所示。Springenberg 等人[36] 和 Zeiler 等人[37] 也利用自上而下的梯度来发现中间层学习的模式。基于自然图像的先验知识，特征反转方法[88–92] 学习一副新图像以重建神经元激活。此外，最近的一些方法[93–95] 试图检测卷积神经网络中间层学习的概念。上述特征可视化方法探索了中间特征检测到了什么，但它们没有回答网络如何组装单独的特征来进行预测。

2.4.3 将知识蒸馏到可解释模型

最近，另一条研究路线试图将卷积神经网络的强大能力转移到可解释模型，例如决策树或线性模型，以近似原始模型的行为。陈等人[96] 将将卷积神经网络的知识提炼成一个可解释的加法模型。Ribeiro 等人[53] 利用局部线性模型来逼近原始模型，研究输入如何影响分类器的决策。Frosst 等人[97] 和刘等人[98] 将卷积神经网络的学习知识提炼到决策树中。这些方法仅在网络决策和输入之间架起桥梁，然而它们无法帮助用户了解卷积神经网络的内部特征如何影响网络决策以及相互影响。本文在第五章提出的层次化注意力分解也是对原始模型的近似。与上述方法不同，层次化注意力分解不仅突出了对于网络决策的重要特征，而且建立了来自不同层的特征通道之间的关系。从本文的方法中，可以获得内部特征的状态以及内部特征之间如何相互影响。

2.4.4 内在可解释模型

除了对训练后的卷积神经网络进行可解释性分析外，一些研究人员还试图探索固有的可解释模型。陈等人[99] 提出了一种称为原型部分网络的深度网络架构。该网络有一个透明的推理过程，首先计算图像块和原型之间的相似度分数，然后网络根据相似度得分的加权和进行预测。概念瓶颈模型[100–102] 本质上也是可解释的。与那些利用人类特定概念生成解释的后处理方法[93, 94] 不同，它们在训练时直接预测一组人类特定概念，然后使用这些概念进行预测，其中推理过程是可解释的。最近的一些内在可解释模型[99, 100] 利用 VGGNet[2] 或 ResNet[1]

图 2.3 ILSVRC 数据集上的一些图像示例。

首先提取高级特征，并对高级特征进行推理过程。

第五节 常用数据集介绍

本文对前文提到的任务使用的数据集进行简单介绍，包括 ILSVRC 数据集、CUB-200-2011 数据集、DAGM-2007 数据集以及 PASCAL VOC 2012 数据集。

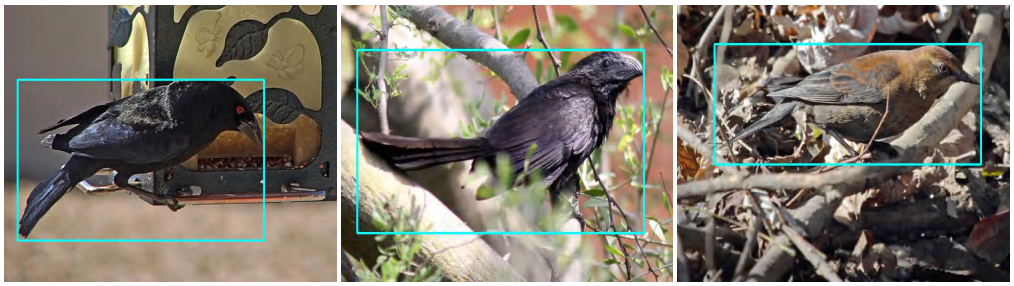
2.5.1 ILSVRC 数据集

ILSVRC 数据集是一个超大规模的图像识别数据集。这个数据集具有约 120 万张图像，包含 1000 个类别，其中验证集包含 50000 张图像。数据集中的每个图像都含有类别标签和边界框标签，可以用来进行分类和弱监督物体定位任务。如图 2.3所示, 本文给出了一些 ILSVRC 数据集上的图像示例。

2.5.2 CUB-200-2011 数据集

CUB-200-2011 数据集[103] 是一个细粒度识别数据集，包含了 200 个鸟的种类，其中图片是从 flicker 网站[1] 收集并经过人工筛选而来。数据集中含有 5,994 张训练图像和 5,794 张测试图像。该数据集提供了几种不同类型的标签，包括类别标签、属性标签、边界框标签以及粗分割标签。本文使用 CUB-200-2011 数据集上的类别标签和边界框标签来测试弱监督物体定位性能。在图 2.4中，本文给出了一些图像示例。

[1] https://www.flickr.com/photos/



铜色牛鹂　　沟嘴犀鹃　　锈色黑鹂

图 2.4　CUB-200-2011 数据集图像示例。本文给出了图像的类别和边界框标签。

2.5.3　DAGM-2007 数据集

DAGM-2007 数据集是关于工业图像表面缺陷的数据集，用于研究者研究缺陷检测算法来辅助人类检测产品质量。数据集中的包含多个子集，每个子集中是一种类型的缺陷。每个子集包含 1000 张无缺陷的正常工业图像和 150 张标记缺陷位置的异常工业图像。

2.5.4　PASCAL VOC 2012 数据集

PASCAL VOC 2012 数据集是由 Everingham 等人[104] 提出的，其中图像都来源于 flicker 网站[2]。这个数据集包含多种标签，例如类别标签，边界框，分割标签和实例标签，可以用来验证多种任务的模型。PASCAL VOC 2012 数据集包含 20 个常见的语义类别，例如人、鸟、自行车和公共汽车等。数据集被分成 3 个集合，训练集（1464 张）、验证集（1449 张）和测试集（1456 张）。Hariharan 等人[105] 将训练数据扩充到 10582 张图像。本文的多分类模型和分割模型都是在 PASCAL VOC 2012 数据集上训练生成的。当评估在测试集上的分割结果时，需要提交到官方的评测网站上 [3]。在图 2.5中，本文给出了一些较难多类别示例图像。

第六节　本章小结

在本章中，本文对研究内容相关的工作进行了详细介绍。具体来说，本章首先介绍了注意力机制的相关工作，包括基于反向传播的方法、基于类别激活映射的方法以及基于基于特征特征扰动的方法。随后本章对注意力机制的不同应

[2]https://www.flickr.com/photos/
[3]http://host.robots.ox.ac.uk:8080/

原始图像　　分割标签

图 2.5　PASCAL VOC 2012 数据集较难图像示例。第一张图像中包含人和自行车两个类别。第二张图像中包含人和马两个类别。分割标签为图像中所有像素分配一个类别，黑色代表背景。

用领域依次进行了详细介绍。本章介绍了弱监督物体定位任务的相关文献，包括早期基于候选框的方法和后来的基于注意力图的方法。本章介绍了弱监督语义分割任务的相关研究工作，包括基于端到端的方法、基于种子点扩张的方法、基于对抗擦除策略的方法以及基于语义关系学习的方法。本章介绍了卷积神经网络归因领域的相关工作，包括特征归因、特征可视化、可解释模型的知识蒸馏和内在可解释模型。在最后一节，本文对不同任务使用的数据集进行了简单介绍。

第三章 基于层次化注意力的物体定位算法研究

第一节 引言

3.1.1 背景知识

近年来，学者们提出了大量的基于注意力的物体定位方法[12, 16, 17]。这些注意力方法从基于卷积神经网络的图像分类器生成注意力图，图中的高响应区域对应目标物体的判别性区域。由于类别标签只提供了目标物体的类别信息，不提供任何目标物体在图像中位置信息，注意力图的定位能力可以很好地为类别标签提供物体位置信息。因此，注意力图的出现进一步促进了基于类别标签的弱监督任务的发展，例如，弱监督语义分割[48, 49, 72, 106] 和弱监督物体定位[13, 107]。

Zhou 等人[12] 在 2016 年提出了一个基于类别激活映射的注意力模型（CAM），该模型对网络深层的特征图进行加权求和来生成注意力图。他们利用特定的网络结构生成注意力图，这种结构将图像分类器的全连接层替换为全局平均池化层。后来，在 Zhou 等人的工作的基础上，Selvaraju 等人[17] 提出了 Grad-CAM 方法，它能够为任何现成的基于卷积神经网络的图像分类器的生成注意力图，增强了这种方法的泛化能力。在 Grad-CAM 的设计中，他们利用特征图的平均梯度来表示该特征图对目标类别的重要性。虽然以上这两种方法可以有效定位目标物体的位置，它们之间存在的一个共同问题是它们都依赖于分类网络最终的卷积层来生成注意力图。由于最终卷积层的特征图空间分辨率非常低，基于这些特征图生成的注意力图的分辨率也会很低，在对注意力图进行上采样后，注意力图只能定位到粗略的物体区域。如图 3.1所示，本文展示了使用 Grad-CAM 从 VGG-16[2] 的 `conv5_3` 层生成的注意力图，可以发现注意力图只能定位马的粗略位置，缺乏定位马的精细细节的能力，例如马腿。

3.1.2 解决方案的动机

从最终卷积层生成的的注意力图定位的物体区域粗糙，限制了基于类别标签的弱监督任务的性能上限。因此，本文希望获得更细粒度的物体位置信息来更好地定位目标物体。由于神经网络浅层的特征有更高的空间分辨率，它们对

图 3.1 Grad-CAM[17]（顶行）和 LayerCAM（底行）生成的注意力图。注意力图是从 VGG-16[2] 的 conv3_3 层和 conv5_3 层生成的。

细粒度的细节信息更加敏感。因此，本文尝试用现有的注意力方法从浅层生成注意力图。在图 3.1顶行，本文展示了用 Grad-CAM 从 VGG-16 的 conv3_3 层生成的注意力图。可以看出，图中强响应的位置散布在整个图像，注意力图不能准确的定位物体区域。本文分析这是由于 Grad-CAM 只考虑捕捉每个特征图全局信息，忽略了像素之间的局部差异。因此，本文重新考虑特征图和权重之间的关系来从浅层生成可靠的注意力图来获得更准确的细粒度物体位置信息。

第二节 相关工作

在下面的章节中，本文首先回顾两种和本文紧密相关的注意力方法，详细介绍 Grad-CAM 和 Grad-CAM++ 的机制并分析它们在浅层失败的原因，随后引出本章的基于层次化注意力的物体定位算法，LayerCAM。

3.2.1 Grad-CAM 和 Grad-CAM++

形式上，让 f 表示图像分类器，θ 表示其参数。对于给定图像 I，将其输入分类器，本文可以通过下式获得目标类别 c 的预测分数 y^c：

$$y^c = f^c(I, \theta). \tag{3.1}$$

让 A 代表 CNN 中最终卷积层的输出特征，A^k 为 A 中的第 k 个特征图。预测得分 y^c 相对于特征图中的一个空间位置 ij 的梯度可以通过 $g_{ij}^{kc} = \frac{\partial y^c}{\partial A_{ij}^k}$ 计算得到。要生成目标类别 c 的注意力图，Grad-CAM 和 Grad-CAM++ 会为每个特征图 A^k

分配一个通道级权重 w_k^c，然后对特征 A 中的所有特征图进行线性加权求和。最后，应用 ReLU 操作从注意力图中去除负响应，公式为：

$$M^c = \mathrm{ReLU}(\sum_k w_k^c \cdot A^k). \tag{3.2}$$

对于 Grad-CAM，特征图 A^k 的通道级权重 w_k^c 是通过对特征图中所有位置的梯度取平均得到的，公式为：

$$w_k^c = \frac{1}{N}\sum_i\sum_j g_{ij}^{kc}, \tag{3.3}$$

其中 N 代表特征图 A^k 中空间像素的数量。对于 Grad-CAM++[16], 通道级别权重 w_k^c 可以通过以下公式计算获得：

$$w_k^c = \sum_i\sum_j \alpha_{ij}^{kc} \cdot \mathrm{ReLU}(g_{ij}^{kc}), \tag{3.4}$$

其中 α_{ij}^{kc} 由下式得到：

$$\alpha_{ij}^{kc} = \frac{\left(g_{ij}^{kc}\right)^2}{2\left(g_{ij}^{kc}\right)^2 + \sum_a\sum_b A_{ab}^k (g_{ij}^{kc})^3}, \tag{3.5}$$

其中 ab 代表 A^k 中的空间位置。Grad-CAM 和 Grad-CAM++ 的区别在于前者只使用梯度生成通道级权重，而后者利用特征图和梯度的结合生成通道级权重。在图像中出现多个物体实例时，Grad-CAM++ 会表现出比 Grad-CAM 更好的定位能力。

虽然 Grad-CAM 和 Grad-CAM++ 可以从最终的卷积层生成可靠的注意力图，但定位的物体区域很粗糙。本文希望找到更多细粒度的定位信息来补充来自最终卷积层的注意力图，从而更好地定位目标物体。众所周知，卷积神经网络的浅层特征具有更大的空间分辨率，这使得它们能够捕获目标物体的细粒度的细节。因此，获得细粒度物体细节信息的一个自然想法是将 Grad-CAM 或 Grad-CAM++ 应用于卷积神经网络浅层特征。然而，根据实验，由 Grad-CAM 或 Grad-CAM++ 从浅层生成的注意力图通常包含许多噪声，如图 3.2所示。在下文中，本文首先分析为什么 Grad-CAM 和 Grad-CAM++ 无法为浅层生成可靠的注意力图，然后介绍本文的方法 LayerCAM。

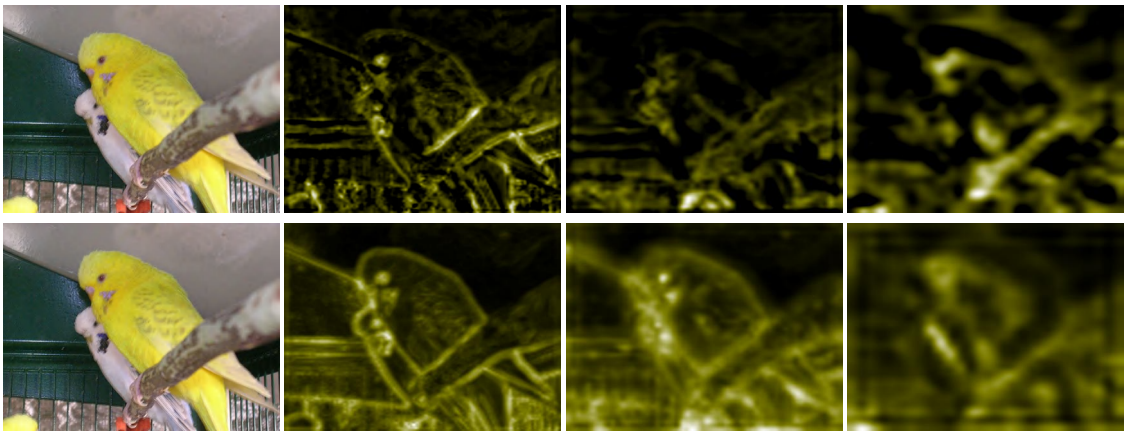


图 3.2 Grad-CAM（顶行）和 Grad-CAM++（底行）在 VGG-16 不同层生成的注意力图。如图所示，Grad-CAM 和 Grad-CAM++ 在卷积神经网络浅层生成的注意力图不能准确的定位物体区域。

3.2.2 分析

Grad-CAM 和 Grad-CAM++ 都为每个特征图 A^k 分配了一个全局权重 w_k^c，其中 A^k 中的每个位置具有相同的权重 w_k^c。然而，如图 3.3(c) 所示，浅层中的特征图倾向于捕获细粒度的细节，不管它们属于目标物体还是背景。因此，全局权重不能消除特征图中的背景区域，这使得生成的注意力图无法准确定位目标物体。

此外，本文还对全局权重是否可以代表一个特征图中每个位置的重要性进行了数值分析。对于 Grad-CAM，本文计算梯度 g^{kc} 的方差，其中方差表示每个位置的梯度与平均梯度 w_k^c 的差异。对于 Grad-CAM++，本文计算 $\alpha^{kc} \cdot \mathrm{ReLU}(g_{ij}^{kc})$ 的方差。本文从 VGG-16 中选择每个阶段的最后一个卷积层。如图 3.3(a-b) 所示，对于 VGG-16 的最后一个阶段，可以看到大多数特征图对应的方差趋于零。这表明特征图中每个空间位置的权重近似等于全局权重。因此，在最后阶段，Grad-CAM 和 Grad-CAM++ 使用的全局权重都可以表示特征图中每个空间位置的重要性。然而，在网络浅层，大多数特征图对应的方差非常大。全局权重不能代表特征图中不同位置对于目标类别的重要性。因此，Grad-CAM 和 Grad-CAM++ 无法从网络浅层生成可靠的注意力图。

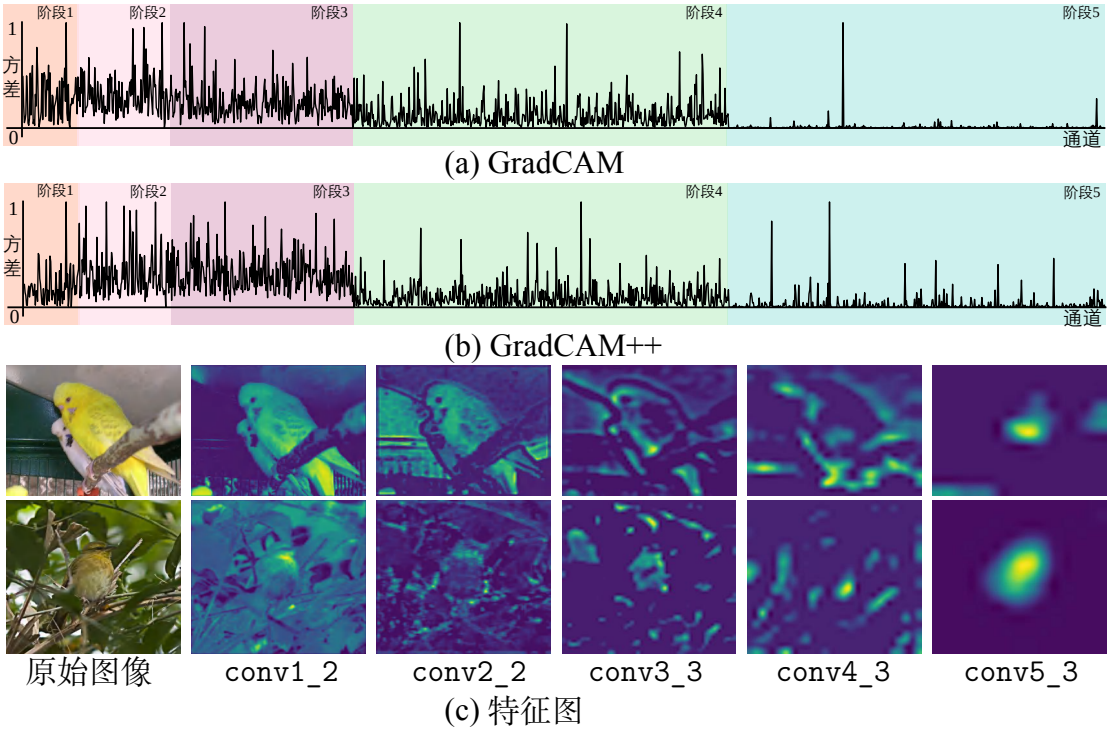


图 3.3 (a-b) 展示了 Grad-CAM 和 Grad-CAM++ 在 VGG-16 的不同阶段每个特征图对应的梯度的方差。(c) 展示了从 VGG-16 不同阶段随机选择的特征图。

第三节 层次化注意力

3.3.1 解决方案概述

在图 3.4中，本文给出了 LayerCAM 的框架图。本文重新思考了特征图以及它们对应的权重之间的关系。与之前的注意力方法只考虑每个特征图的全局信息不同，本文利用局部权重来突出显示特征图中每个位置对目标类别的重要性。通过这样的操作，可以有效保留目标物体的细粒度细节，同时去除背景中的噪声。**(1)** 通过以上的方式，LayerCAM 不仅可以从最终的卷积层生成可靠的注意力图，还可以从卷积神经网络的浅层生成可靠的注意力图，这样通过 LayerCAM，既可以获得粗略的空间位置又可以获得细粒度的物体细节。**(2)** 来自不同层的注意力图通常是互补的。这一优势促使本文将它们结合起来以生成更精确和完整的物体区域，这将显著有益于弱监督任务。LayerCAM 非常简单，容易应用于现有在基于卷积神经网络的图像分类器上，并无需修改网络架构和反向传播方式。

为了证明本文方法生成的注意力图的质量，本文将它们应用于弱监督物体定位任务和弱监督语义分割任务。在这两个任务上的实验表明，LayerCAM 比以前的注意力方法实现了更好的物体定位能力，证明了 LayerCAM 的有效性。此

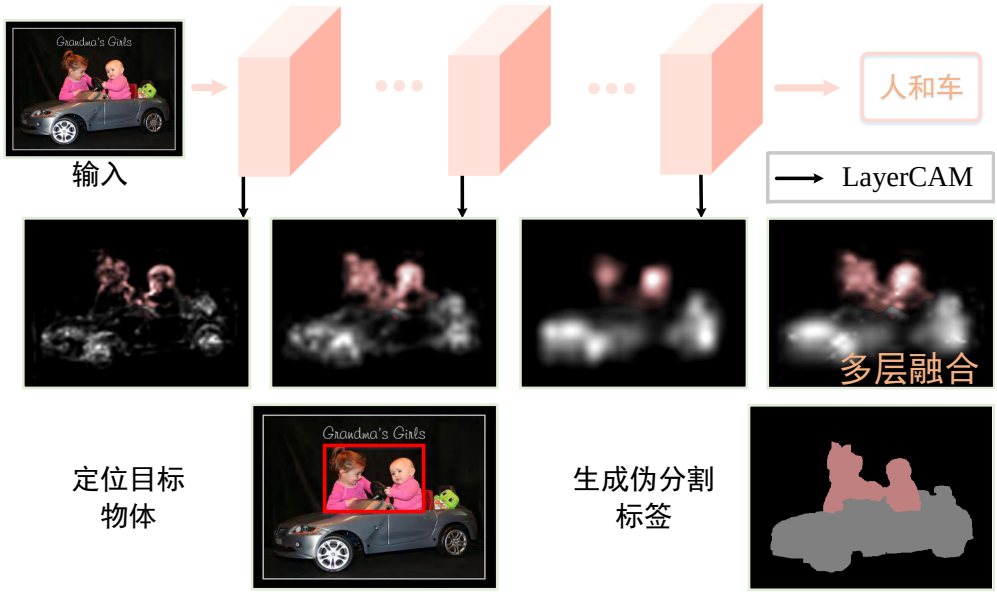


图 3.4 LayerCAM 的示意图. LayerCAM 可以应用于任何现成的基于卷积神经网络的模型并从不同层生成注意力图。不同阶段的注意力图的融合对于物体定位是有益的。

外，从卷积神经网络模型的浅层生成的注意力图也可以被用于定位工业图像中的微小缺陷。

3.3.2 层次化注意力 LayerCAM

本文基于从目标类别回传的梯度给特征图中的每个空间位置生成单独的权重。正如以前工作[16] 中的经验验证一样，与特征图中某个位置对应的正梯度表明增加该位置激活的强度将对目标类别的预测分数产生积极影响。对于具有正梯度的位置，本文使用它们的梯度作为权重。对那些具有负梯度的位置，本文设置权重为零。

正式地，第 k 个特征图中空间位置 ij 的权重可以写为:

$$w_{ij}^{kc} = \mathrm{ReLU}(g_{ij}^{kc}). \tag{3.6}$$

为了获得某一层的注意力图，LayerCAM 首先将特征图中每个位置的激活值乘以权重:

$$\hat{A}_{ij}^{k} = w_{ij}^{kc} \cdot A_{ij}^{k}. \tag{3.7}$$

最后将结果 $\hat{A}^k$ 沿通道维度求和，得到注意力图，公式如下:

$$M^c = \mathrm{ReLU}(\sum_k \hat{A}^k). \tag{3.8}$$

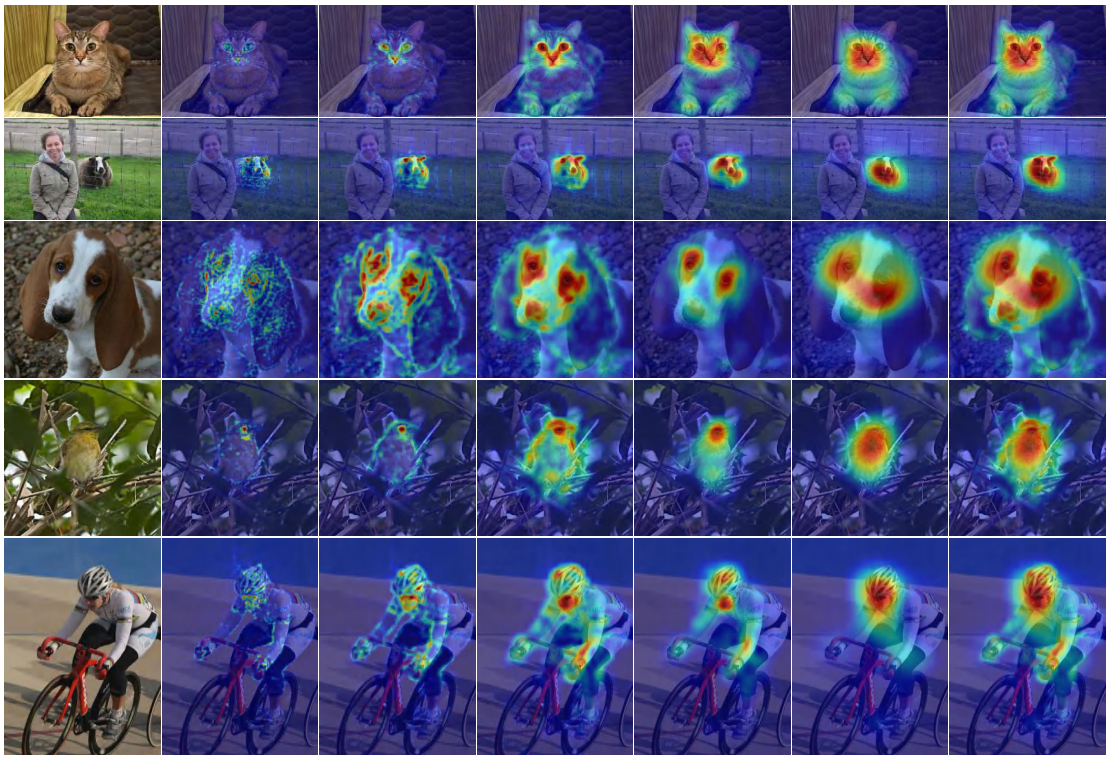


图 3.5 LayerCAM 从不同阶段生成的注意力图之间的比较。图像是从 PASCAL VOC 数据集中随机选择的[104]。**阶段 5** 代表注意力图是从 VGG-16 中第 5 阶段的最后一个卷积层生成的。**融合**表示将阶段 3、阶段 4 和阶段 5 的注意力图融合起来的结果。

基于上述操作，从浅层生成的注意力图可以捕获可靠的细粒度物体定位信息，如图 3.5所示。本文认为 LayerCAM 的成功得益于既考虑了不同通道的重要性，又考虑了特征图中不同空间位置的重要性。每个位置的单独权重可以反映特征图中不同位置对于目标类别的重要性。本文将在实验部分进行更多的定性和定量分析。

第四节 实验

本节首先进行弱监督物体定位实验，验证 LayerCAM 的定位能力。然后本文进行图像遮挡实验来测试来自最终卷积层的注意力图的定位能力的可靠性。此外，本文还将进行工业图像表面缺陷检测实验，以表明 LayerCAM 从浅层生成的注意力图可以定位到可靠的细粒度物体定位信息。最后，本文证明来自不同层的注意力图的组合有利于弱监督语义分割任务。

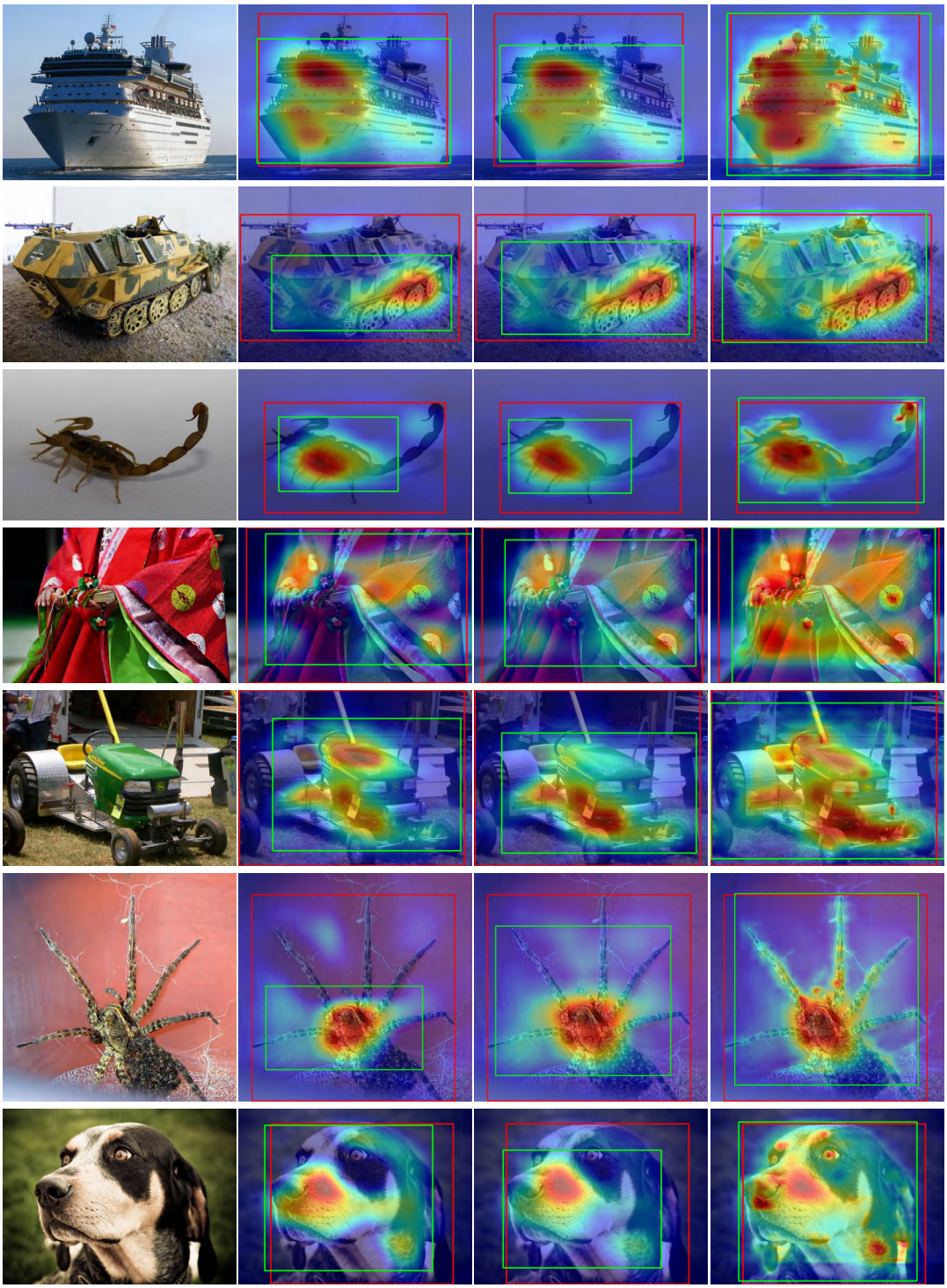


图 3.6 不同方法之间定位结果的比较。图像是从 ILSVRC 验证集中随机选择的[108]。红框表示真值框，绿框表示预测框。LayerCAM 的多层融合的注意力图比 Grad-CAM 和 Grad-CAM++ 更精确地定位物体边界框。

3.4.1 弱监督物体定位实验

ILSVRC 基准[108] 中提出了弱监督物体定位实验，旨在为目标类别定位物体边界框。本文在 ILSVRC 基准的验证集上评估 LayerCAM 的定位能力，其中验证集包含 50000 张图像。定位精度由 *loc1* 和 *loc5* 指标衡量。*loc1* 指标表示如果预测的边界框和真实边界框之间的交并比（IoU）大于或等于 0.5，同时最高概率的预测类别正确，则预测的边界框正确。*loc5* 指标表示如果预测的边界框和真实边界框之间的交并比（IoU）大于或等于 0.5，同时概率最高的前五个预测类别中有目标类别，则预测的边界框正确。

表 3.1 不同阶段注意力图定位精度的比较。第一行中的“S”表示 VGG-16 中的“阶段”。**S5-S1** 表示 VGG-16 中每个阶段的最后一个卷积层。

方法	指标 (%)	S5	S4	S3	S2	S1
Grad-CAM[15]	*loc1*	43.62	18.32	8.87	19.59	13.95
	loc5	53.99	22.70	11.05	23.85	17.27
Grad-CAM++[16]	*loc1*	45.44	41.11	35.33	31.70	31.32
	loc5	56.42	50.97	43.86	39.40	38.90
ScoreCAM[46]	*loc1*	39.51	33.08	31.15	29.90	29.63
	loc5	49.63	41.63	39.30	37.80	37.46
NormGrad[109]	*loc1*	38.94	40.85	38.67	32.05	29.94
	loc5	49.19	51.98	49.56	41.37	38.69
LayerCAM	*loc1*	**46.62**	**44.05**	**41.83**	**43.18**	**43.71**
	loc5	**57.83**	**55.02**	**52.28**	**53.60**	**54.34**

3.4.1.1 实现细节

为了从注意力图中生成物体边界框，如一些工作[12, 15] 中所做的那样，本文直接用注意力图中最大值的 15% 的作为阈值对注意力图进行二值化，然后找到最大连通体的外接矩形框。本文选择 VGG-16 中不同阶段的最后一个卷积层来生成注意力图。对于 LayerCAM 从 `conv1_2` 和 `conv2_2` 层生成的注意力图，其中具有强值的物体位置倾向于分散在物体上。因此，按照[35] 中的做法，本文应用 GraphCut 算法[110] 来生成连通的区域。

表 3.2 公式 (3.9) 中缩放因子 γ 的消融实验。第一行中的 “S” 表示 VGG-16 中的 “阶段”。**S5-S1** 表示 VGG-16 中每个阶段的最后一个卷积层。

设置	指标 (%)	1	2	3	4
S5+S4+S3	*loc1*	47.18	**47.22**	47.17	47.04
	loc5	58.63	**58.72**	58.62	58.46
S5+S4+S3+S2	*loc1*	47.19	**47.24**	47.17	47.02
	loc5	58.63	**58.74**	58.63	58.46
S5+S4+S3+S2+S1	*loc1*	47.20	**47.23**	47.19	46.97
	loc5	58.65	**58.74**	58.64	58.39

表 3.3 不同缩放函数的消融实验。no scale: 无缩放函数。第一行中的 “S” 表示 VGG-16 中的 “阶段”。**S5-S1** 表示 VGG-16 中每个阶段的最后一个卷积层。

设置	指标	no scale	$tanh(x)$	$\sqrt[2]{x}$	$tan(x)$
S5+S4+S3	*loc1*	47.01	**47.18**	44.52	42.91
	loc5	58.40	**58.63**	55.62	53.39
S5+S4+S3+S2	*loc1*	47.00	**47.19**	44.53	40.07
	loc5	58.39	**58.63**	55.64	49.96
S5+S4+S3+S2+S1	*loc1*	46.91	**47.20**	44.51	38.67
	loc5	58.27	**58.65**	55.62	48.27

3.4.1.2 实验结果

在表 3.1中，本文首先展示了来自 VGG-16 不同阶段的注意力图的定位能力。本文发现 LayerCAM 的定位性能相较于 Grad-CAM[15]、Grad-CAM++[16]、ScoreCAM[46] 和 NormGrad[109] 有大幅度提高，尤其是在浅层。这一事实表明，LayerCAM 可以从浅层获得比 Grad-CAM、Grad-CAM++、ScoreCAM 和 NormGrad 更可靠的细粒度物体定位信息。LayerCAM 为特征图中的每个位置分配单独权重，这样的操作可以充分考虑不同位置对目标类别的重要性。

此外，本文还展示了融合不同阶段的注意力图的定位性能。对于来自浅层的注意力图，激活值远低于来自深层的激活值。如表 3.3所示, 当不使用缩放函数时，融合注意力图的定位性能不会得到提升。因此，当结合来自不同层的注意力图时，本文首先通过缩放函数缩放浅层生成的注意力图，其中缩放图中空

表 3.4 不同阶段融合注意力图定位精度的比较。第一行中的“S”表示 VGG-16 中的“阶段”。**S5-S1** 表示 VGG-16 中每个阶段的最后一个卷积层。

方法	指标 (%)	S5	+S4	+S3	+S2	+S1
Grad-CAM[15]	*loc1*	43.62	40.56	40.03	35.87	33.96
	loc5	53.99	50.11	49.47	44.48	42.17
Grad-CAM++[16]	*loc1*	45.44	42.72	37.25	32.51	31.60
	loc5	56.42	52.95	46.19	40.40	39.23
ScoreCAM[46]	*loc1*	39.51	37.26	31.88	29.94	29.52
	loc5	49.63	46.78	40.19	37.84	37.33
NormGrad[109]	*loc1*	38.94	36.59	36.45	36.41	36.26
	loc5	49.19	46.02	45.86	45.79	45.59
LayerCAM	*loc1*	46.62	47.17	47.22	**47.24**	47.23
	loc5	57.83	58.67	58.72	**58.74**	58.74

间位置 ij 的值由下式计算

$$\hat{M_{ij}}^{c} = \tanh\left(\frac{\gamma * M_{ij}^{c}}{\max(M_{ij}^{c})}\right), \tag{3.9}$$

其中 γ 是缩放因子。然后本文利用一个简单的元素最大值操作来融合来自不同层的注意力图。从表 3.2可以看出，当 γ 设置为 2 时，LayerCAM 取得了最好的定位结果。本文还探索了不同种类的缩放函数，如表 3.3所示。当使用 $\sqrt[2]{x}$ 缩放函数时，定位性能变得更差。这是因为 $\sqrt[2]{x}$ 缩放函数将接近 0 的值放大太多，从而增强了噪声强度。例如，0.01 缩放到 0.1。当使用 $tan(x)$ 缩放函数时，性能也会变得更差。$tan(x)$ 缩放函数将 1 附近的大值缩放很多，这将抑制归一化后较低值的放大。可以看出，当使用 $tanh(x)$ 缩放函数时，可以获得更好的融合结果。如表 3.4所示，不同层注意力图的融合可以逐渐提高定位性能。

在表 3.5中，本文展示了不同方法之间定位性能的比较。最右边两列的注意力方法都是基于原始的 VGG-16 模型。最左边的三种定位方法都采用 VGG-16 架构，用全局平均池化层代替全连接层。与 c-MWP、CAM、Grad-CAM 和 GradCAM++ 相比，可以看出 LayerCAM 将 $loc1$ 性能分别提高了 18.16%、4.44%、3.62% 和 1.80%。本文的方法也取得了比一些最先进的定位方法 ACoL[13] 和 ADL[111] 更好的结果，其中它们是专门为解决物体定位任务而设计的。比较表

表 3.5 不同方法之间定位精度的比较。其他方法的注意力图都是从最后的卷积层生成的。星号 * 表示实验结果来自这篇文章[17]。

方法	ACoL[13]	ADL[111]	CAM[12]*	c-MWP[41]*	LayerCAM
loc1 (%)	45.83	44.92	42.80	29.08	**47.24**
loc5 (%)	59.43	-	54.86	36.96	**58.74**

明，LayerCAM 的注意力图可以提供更可靠的物体定位信息。本文在图 3.6中展示了不同方法定位效果可视化的示例图。

此外，本文还在 CUB-200-2011 细粒度数据集上验证本文 LayerCAM 的定位效果。定位精度是由 *loc1* 和 *loc5* 指标衡量。注意力图二值化的阈值为最大值的 5%。从表 3.6中可以看出，LayerCAM 多层融合的注意力图取得了最好的定位结果。相比于 CAM[12] 从网络最深层生成注意力图，本文的注意力图是结合网络不同阶段生成注意力图。浅层的注意力图可以获取到更多细节信息，帮助深层注意力图定位更准确的物体边界。和其它专门用于弱监督物体定位的方法相比，本文的方法取得了更好的定位结果，证明了本章提出的基于层次化注意力的物体定位方法的有效性。

表 3.6 在 CUB-200-2011 测试集上不同方法的物体定位准确率的比较。这些方法都是基于 VGG-16 分类网络[2] 生成的注意力图。相比于其它方法，可以看出 LayerCAM 多层融合的注意力图取得了最好的定位结果。

方法	CAM[12]	ACoL[13]	SPG[50]	ADL[111]	LayerCAM
loc1 (%)	41.46	45.92	46.64	52.36	**56.08**
loc5 (%)	51.36	56.51	57.72	-	**69.18**

3.4.1.3 负梯度的影响

在公式 (3.6) 中，LayerCAM 利用 ReLU 过滤掉负梯度。本文首先做实验来研究负梯度对定位能力的影响。如表 3.7所示，LayerCAM-normal 表示 公式 (3.6) 去掉 ReLU, 可以看到 LayerCAM-normal 比 LayerCAM 的定位性能低很多，尤其是在浅层。这一事实证明，LayerCAM 中的负梯度会降低注意力图的定位能力。

表 3.7 不同阶段注意力图定位精度的比较。第一行中的“S”表示 VGG-16 中的“阶段”。**S5-S1** 表示 VGG-16 中每个阶段的最后一个卷积层。

方法	指标	S5	S4	S3	S2	S1
LayerCAM-normal	*loc1*	42.09	37.63	34.74	34.12	30.86
	loc5	52.10	46.37	43.09	42.52	39.01
LayerCAM	*loc1*	46.62	44.05	41.83	43.18	43.71
	loc5	57.83	55.02	52.28	53.60	54.34

3.4.2 图像遮挡实验

对于 LayerCAM 从最终卷积层生成的注意力图，本文进行了 Zeiler 等人[37]的工作中提出的图像遮挡实验，以验证注意力图定位的置信区域的可靠性。图像遮挡实验是用来测试被遮挡的图像区域对最终预测的重要性。如果置信区域很重要，当输入被遮挡的图像时，目标类别的预测分数将大大降低。具体来说，在 ILSVRC 验证集上，本文首先选择出被 VGG-16 分类网络正确预测的图像。对于这些正确预测的图像，本文用 0.7 的阈值从注意力图中选择遮挡的区域对输入图像进行遮挡，然后将它们输入到网络中。如表 3.8所示，本文分别展示了真实类别的平均预测分数，top-1 分类准确率和 top-5 分类准确率。

表 3.8 图像遮挡实验的分类精度比较。**Confidence**: 表示真实类别的平均预测分数。分数越低，效果越好。

方法	原始分类性能	Grad-CAM[15]	Grad-CAM++[16]	LayerCAM
Top-1 Acc (%)	68.74	50.36	50.07	**48.26**
Top-5 Acc (%)	88.57	75.62	75.26	**73.43**
Confidence (%)	68.64	50.24	49.99	**48.12**

本文在表 3.8中展示了图像遮挡实验的实验结果。LayerCAM 实现了比 Grad-CAM 和 Grad-CAM++ 更低的分类精度，这表明 LayerCAM 从最终卷积层生成的注意力图可以为目标类别发现更重要的物体区域。该实验验证了 LayerCAM 定位的置信区域的可靠性。在对图像进行遮挡时，本文可以发现 LayerCAM 去除置信区域后比 Grad-CAM 和 Grad-CAM++ 更显著的降低了预测分数。

3.4.3 工业表面缺陷定位实验

利用计算机辅助工具检查工业产品的质量是提高工业生产质量和效率的重要途径。为了定位到工业图像中的表面缺陷位置，许多研究人员[112–115] 使用强监督训练的网络。具体来说，他们需要在工业图像中标注缺陷的位置，然后训练一个分割或检测网络。虽然这类方法取得了优异的性能，但标注缺陷位置却相当困难。这是因为工业图像表面上的缺陷及其周围的区域往往具有非常低的对比度，如图 3.7(a) 所示。

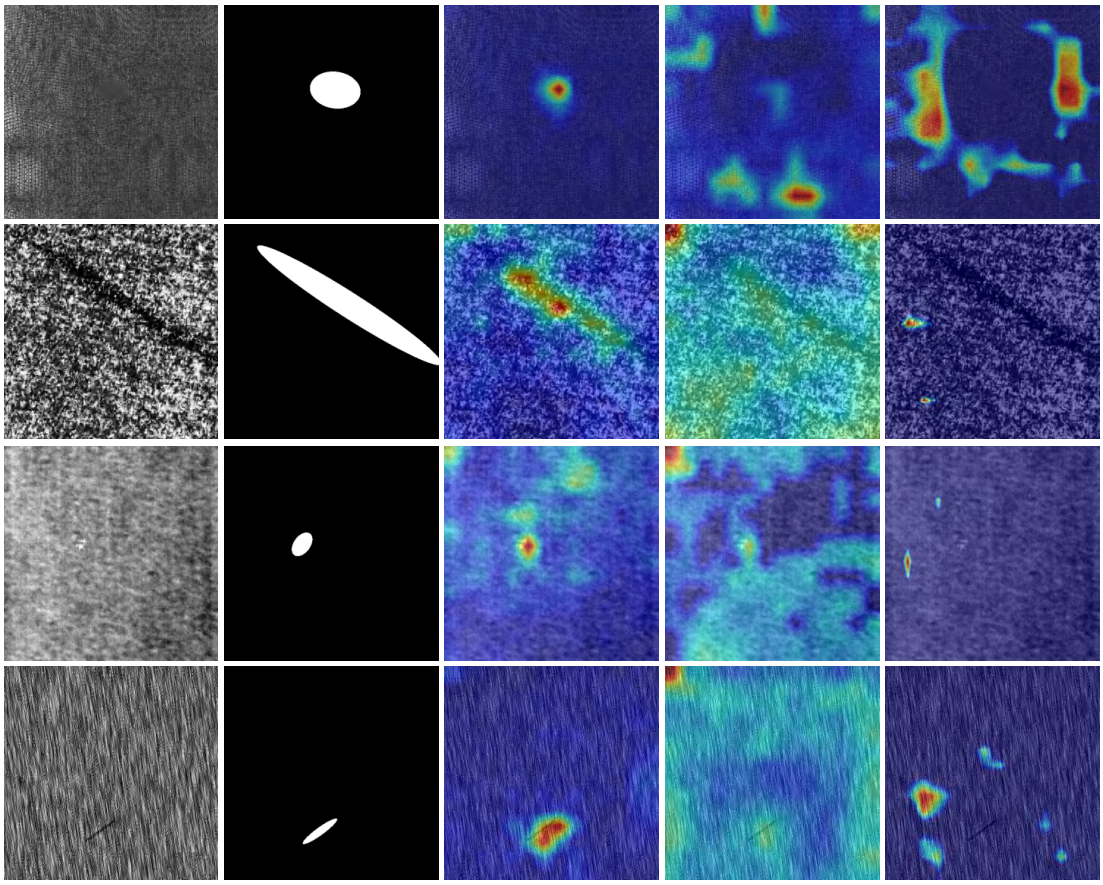
此外，在特定场景下检查工业图像中的缺陷往往需要专业知识，同时也需要大量的人力和时间。因此，弱监督方法值得研究，因为它们可以显著降低标注成本。本文利用从浅层生成的注意力图来定位工业图像中各种形状的微小缺陷，因为来自浅层的注意力图对细粒度的物体细节很敏感。本文将此问题视为图像中存在或不存在缺陷的二元分类问题，然后使用类别标签来训练基于 ResNet50[1] 的分类器。最后，本文应用 LayerCAM 来定位工业图像中的缺陷。

3.4.3.1 实现细节

本文在 DAGM-2007 缺陷数据集[116] 上进行了实验，其中包含 3550 张训练图像和 400 张测试图像。该数据集包含不同纹理表面上的多种类型的缺陷，如图 3.7(a) 所示。本文在这个数据集上训练一个缺陷图像分类器。本文使用 SGD 来优化分类网络并将网络训练 15 轮，其中批大小为 32。初始学习率设置为 0.001，并在第 5 轮和第 10 轮处依次衰减 10 倍。在推理时，本文分别将 LayerCAM、Grad-CAM 和 Grad-CAM++ 应用于 ResNet-50 的 `layer3` 以生成注意力图。随后，本文将生成的注意力图首先阈值化为二值图，并计算二值图和真实标签之间的 IoU 分数，最后在所有图像上计算平均 IoU 分数（mIoU），即可由以下公式计算得到：

$$mIoU = \frac{1}{N} \times \sum_{i=1}^{N} \frac{|A_i \cap GT_i|}{|A_i \cup GT_i|} \tag{3.10}$$

其中 A_i 表示第 i 张二值图像中像素值为 1 的像素集合，GT_i 表示第 i 张真实标签中像素值为 1 的像素集合，N 表示图片数量。本文为每种注意力方法寻找最佳阈值并报告它们的最佳性能。此外，本文还测试了不同方法的帧率 (FPS)。



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

图 3.7 工业缺陷的注意力图。(a) 原始图像 (b) 真值。(c-e) 显示了由 LayerCAM、Gard-CAM++ 和 Grad-CAM 从 ResNet50 的 `layer3` 生成的注意力图。这些图像是从 DAGM-2007 缺陷数据集[116] 中随机选择的。LayerCAM 生成的注意力图比 Grad-CAM 和 Grad-CAM++ 更精确地定位微小缺陷的位置。**放大以获得最佳视图**.

表 3.9 缺陷检测任务中不同方法性能的比较。SegNet 和 RefineNet 是全监督方法，其他方法是基于类别标签的弱监督方法。结果中星号 * 表示结果来自论文[112]。

方法	mIoU (%)	帧率
SegNet[117]	21.95*	17.92*
RefineNet[9]	**32.90***	31.05*
Grad-CAM[15]	0.35	60.97
Grad-CAM++[16]	6.46	60.24
LayerCAM	**27.26**	60.61

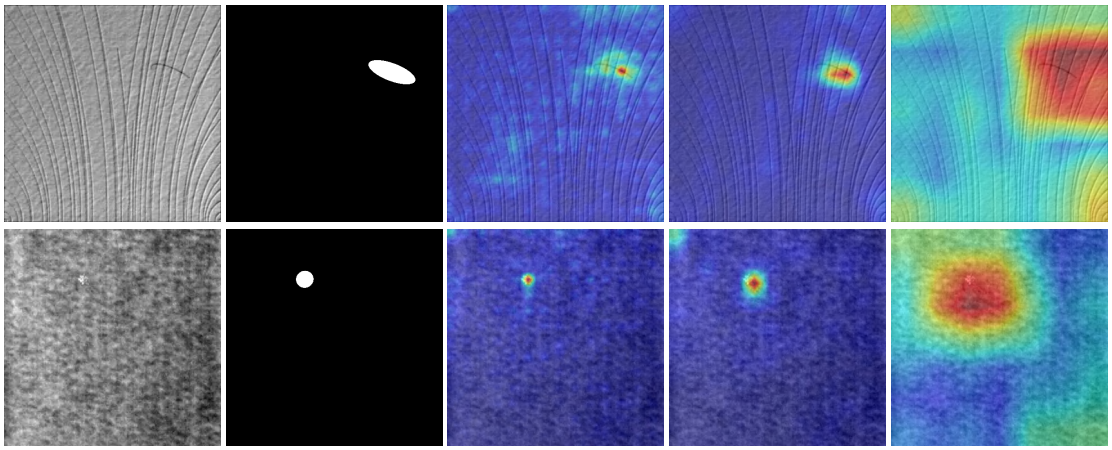


图 3.8 缺陷检测任务中从 ResNet50 不同层生成的注意力图的示例。

3.4.3.2 实验结果

本文在表 3.9中展示了不同方法之间的定量比较。实验结果表明，本文的方法可以比 Grad-CAM 和 Grad-CAM++ 更准确地定位缺陷的位置。本文还展示了两种使用像素级标签训练的全监督方法 SegNet[117] 和 RefineNet[9] 的结果。与它们相比，LayerCAM 实现了可比的性能，但速度大约是它们的两倍。此外，本文还在图 3.7中展示了不同方法之间的定性比较。与 Grad-CAM 和 Grad-CAM++ 相比，LayerCAM 可以定位各种形状的缺陷，而 Grad-CAM 和 Grad-CAM++ 无法准确地定位缺陷位置。

表 3.10 缺陷检测任务中不同层的注意力图定位精度的比较。

设置	layer4	layer3	layer2	layer1	layer4+layer3	layer3+layer2
mIoU (%)	11.59	**27.26**	19.37	13.10	12.28	24.51

在表 3.10中，本文还展示了从 ResNet50 不同层生成的注意力图的定位精度。从表中可以看出，对于工业图像表面缺陷定位任务，layer3 的性能优于多层融合的性能。这是因为缺陷通常具有小尺寸和各种形状。如图 3.8所示，由于从 layer4 生成的注意力图空间分辨率较低，它只能粗略地定位缺陷位置，这不利于注意力图的融合。从 layer2 和 layer1 生成的注意力图定位的缺陷区域小且有噪声。因此，对于工业图像表面缺陷定位任务，本文仅利用来自 layer3 的注

意力图而不是多层融合进行缺陷定位。

3.4.4 弱监督语义分割实验

为了进一步测试 LayerCAM 生成的注意力图的质量，本文将它们应用于需要更多像素位置信息的弱监督语义分割任务。本文利用注意力图和超像素[118]来生成伪分割标签。受 Zhou 等人[119] 的启发，本文使用注意力图作为查询从超像素中收集物体掩码，通过平均每个超像素中的注意力值来计算类别 c 在超像素块中存在的概率，

$$S_c = \left(\frac{1}{|O|} \sum_{j \in O} M_j^c \right), \tag{3.11}$$

其中 O 表示超像素。M 是注意力图，其值归一化到 [0, 1] 的范围内。然后在所有目标类别中选择最大概率，将其对应的类别分配给超像素块中的所有像素。如果最大概率小于固定的低阈值（在本文的实验中，阈值设置为 0.3），则将超像素中的像素分配给背景类别。在为每个超像素分配语义类别后，本文利用它们构成伪分割标签来训练分割网络。

3.4.4.1 实现细节

本文在 PASCAL VOC 2012 数据集[104] 上进行分割实验。该数据集包含 20 个语义类别和背景。数据集被分为三个集合，1464 张训练图像，1449 张验证图像和 1456 张测试图像。按照[105] 中的设置，本文利用具有 10582 张图像的增强训练集来训练分割模型，然后在验证和测试集上将本文的方法与 Grad-CAM 和 Grad-CAM++ 进行比较。为了便于比较，本文使用[12] 中提出的卷积神经网络分类器。具体地，分类网络中具有 1000 个通道的全连接 (FC) 层被修改为 20 个通道。本文采用在 ImageNet[108] 上预训练的 VGG-16 模型来初始化分类网络并使用交叉熵损失对其进行优化。在推理期间，本文选择从 VGG-16 中每个阶段的最后一个卷积层生成注意力图。对于 Grad-CAM 和 Grad-CAM++，本文只利用来自最终卷积层的注意力图，因为他们在浅层生成的注意力图效果很差。

本文采用基于 VGG-16[2] 的 Deeplab-LargeFOV[8] 架构作为分割网络。根据[120] 中的设置，本文还训练基于 ResNet-101[1] 的 Deeplab-LargeFOV[8]。用于训练分割网络的超参数如下：学习率：1e-3；学习率策略：poly，批大小：10。本文使用 SGD 作为优化器，并训练模型 16000 次。学习率在 12000 次迭代时衰减 10 倍。在推理时，本文使用平均交并比（mIoU）指标来评估分割结果。值得

表 3.11 PASCAL VOC 2012 数据集上的弱监督分割结果。本文的 LayerCAM 取得了最好的 mIoU 分数。

方法	Val (%)	Test (%)
Grad-CAM[15]	55.6	56.3
Grad-CAM++[16]	55.5	56.1
LayerCAM (VGG-16)	**60.8**	**61.4**
LayerCAM (ResNet101)	**63.0**	**64.5**

表 3.12 PASCAL VOC 2012 验证集上，不同阶段组合的注意力图的 mIoU 分数的比较。

阶段 5	阶段 4	阶段 3	阶段 2	阶段 1	mIoU (%)
✓					55.6
	✓				55.0
		✓			50.8
			✓		50.5
				✓	46.0
✓	✓				57.1
✓	✓	✓			60.4
✓	✓	✓	✓		**60.8**
✓	✓	✓	✓	✓	60.2

注意的是，这里用的平均交并比和前文工业缺陷检测实验中用到的 mIoU 有所不同。语义分割使用的 mIoU 指标是取所有类别上的 IoU 分数的平均，即可由以下公式计算得到:

$$mIoU = \frac{1}{N_{cl}} \times \sum_{i} \frac{N_{ii}}{\sum_j N_{ij} + \sum_j N_{ji} - N_{ii}} \tag{3.12}$$

其中 N_{cl} 表示类别数量，N_{ij} 表示目标类别为 i、预测类别为 j 的像素数量，N_{ji} 表示目标类别为 j、预测类别为 i 的像素数量，N_{ii} 表示目标类别为 i、预测类别也为 i 的像素数量。

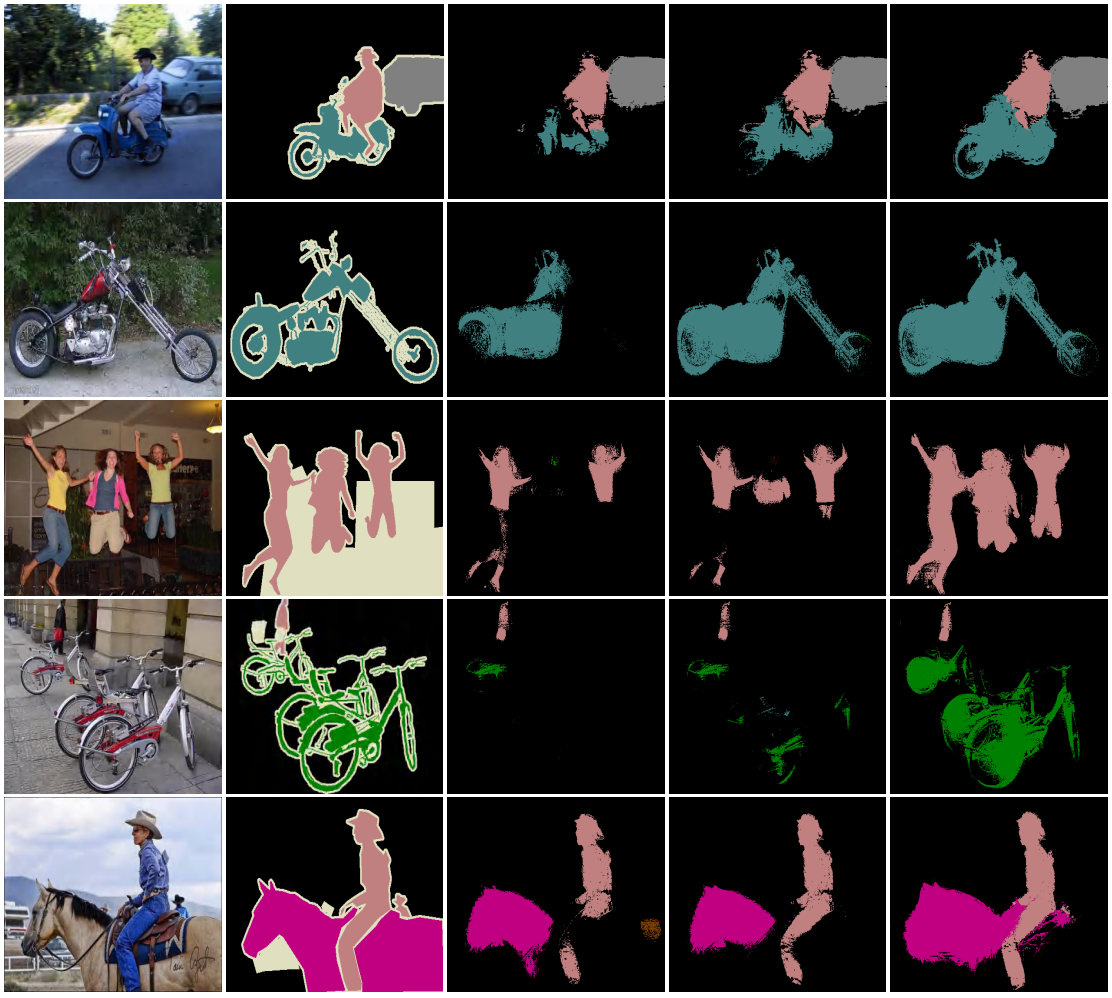


图 3.9 本文的方法产生的分割结果示例。当结合更多来自于浅层的注意力图时，分割效果逐渐提高。

3.4.4.2 实验结果

在 表 3.11 中，本文报告了 LayerCAM 的 mIoU 分数。本文的 LayerCAM 的性能优于 Grad-CAM 和 Grad-CAM++，超过约 5% 的 mIoU 分数。本文还报告了使用不同阶段的注意力图融合的性能，如表 3.12所示。使用 VGG-16 第 5 阶段的注意力图做分割的 mIoU 分数为 55.6%。本文观察到，当连续将第 4、第 3 和第 2 阶段的注意力图与到第 5 阶段通过元素最大值操作融合时，mIoU 分数逐渐增加（从 55.6% 到 60.8%）。这一事实验证了融合注意力图可以获得更多的物体定位信息，这有利于语义分割任务。此外，当将阶段 1 的注意力图融合到最终

表 3.13 DGCN[121] 使用不同注意力图的性能比较。在 PASCAL VOC 2012 验证集和测试集上，基于 LayerCAM 的 DGCN 比基于 CAM 好大约 3% 的 mIoU 分数。

设置	Val (%)	Test (%)
DGCN-CAM	64.0	64.6
DGCN-LayerCAM	67.1	67.6

注意力图中时，性能从 60.8% 下降到 60.2%。本文分析，与其他阶段的注意力图相比，第 1 阶段的注意力图缺乏类别区分度。

本文还将融合的注意力图应用于更高级的弱监督语义方法 DGCN[121]。如表 3.13所示，当用 LayerCAM 种子替换 CAM 种子时，分割结果可以进一步提高约 3% 的 mIoU 分数。实验结果验证了 LayerCAM 生成的种子比 CAM 生成的种子具有更好的定位能力，这有利于弱监督语义分割方法。如图 3.9所示，本文展示了一些定性的分割结果。可以看出，融合 VGG-16 不同阶段的注意力图可以逐渐提高分割结果的质量。

3.4.5 讨论

本章提出的基于层次化注意力的定位方法是基于梯度实现的，因此该方法不受网络结构限制，可以像 Grad-CAM 一样应用于任何目标概念的卷积神经网络。此外，该定位方法不受图像类型限制，既可以在自然图像上进行物体定位，例如 ILSVRC 数据集和 PASCAL VOC 数据集，也可以在合成的工业缺陷图像上应用，例如 DAGM-2007 数据集。最后，该方法既可以用于弱监督物体定位任务，也可以用于弱监督语义分割任务。值得注意的是，本文的方法应用于弱监督语义分割任务时，并没有专门设计更加复杂的算法将任务性能提到最好，本文只是借助于弱监督语义分割任务证明浅层注意力图和深层注意力图结合的好处。

第五节 本章小结

在本章中，提出了一种层次化注意力方法 LayerCAM，它可以有效地从卷积神经网络的不同层生成可靠的注意力图。来自深层的注意力图可以定位物体的粗略物体位置，来自浅层的注意力图可以定位细粒度物体位置。来自不同层的注意力图的组合可以找到更多的物体位置，这有利于提高弱监督任务的性能。实验表明，LayerCAM 比当前的注意力方法具有更好的物体定位能力。此

外，LayerCAM 易用于任何现成的基于卷积神经网络的图像分类器，无需修改网络架构和改变反向传播方式。

第四章　基于在线注意力累积的物体定位算法研究

第一节　引言

4.1.1　背景知识

得益于先进的卷积神经网络（CNN）架构[1, 2, 122]，全监督语义分割方法[7, 9, 123–127] 取得了优异的性能。然而，这些基于卷积神经网络的分割方法通常需要大量的具有分割标注的数据训练分割模型。构造一个像素精确的分割数据集是十分昂贵的，需要耗费大量的人力和时间。为了节约成本，研究者提出使用弱监督的方式学习语义分割任务。常见的用于弱监督语义分割的标签有边界框[68]、草图[69]、点[70] 和类别标签[71]。在这些弱监督标签中，类别的标注比其余标签更加容易获得[128]，也因此被广泛应用于语义分割任务中。

通过类别标签学习语义分割的难点在于类别标签只能说明哪些类别出现在图像中，但是不能提供这些类别的实例物体位置和形状信息。最近，很多工作[18, 49, 72, 73, 82, 129] 尝试基于注意力图[12] 学习语义分割。因为注意力图具有物体定位能力，所以它被广泛应用于弱监督语义分割任务，用于生成目标类别的初始区域。然而，原始的注意力图通常只关注语义物体的一小部分，**注意力图定位的物体区域不完整**，限制了分割网络学习丰富的像素级语义知识的能力。后来的一些方法使用对抗擦除策略[13, 14, 130] 来扩大注意力图定位的区域。不幸的是，随着训练过程的继续，这种方法会扩大定位区域的范围，使得一些不需要的背景区域也被预测为语义区域。Wei 等人[18] 使用空洞卷积来生成完整的物体区域。然而，当卷积的空洞率增大的时候也发生了和对抗擦除策略类似的问题。虽然以上的方法各不相同，但是它们具有一个共同点，即都使用了最终的分类模型来生成注意力图。本文从另一个视角，即分类模型的训练过程，来考虑注意力生成的方式。下面将介绍解决方案的动机。

4.1.2　解决方案的动机

本文观察到，在分类网络达到收敛之前，不同训练阶段产生的注意图定位的判别性区域不断在语义物体上的不同部分移动。主要原因可以总结如下：

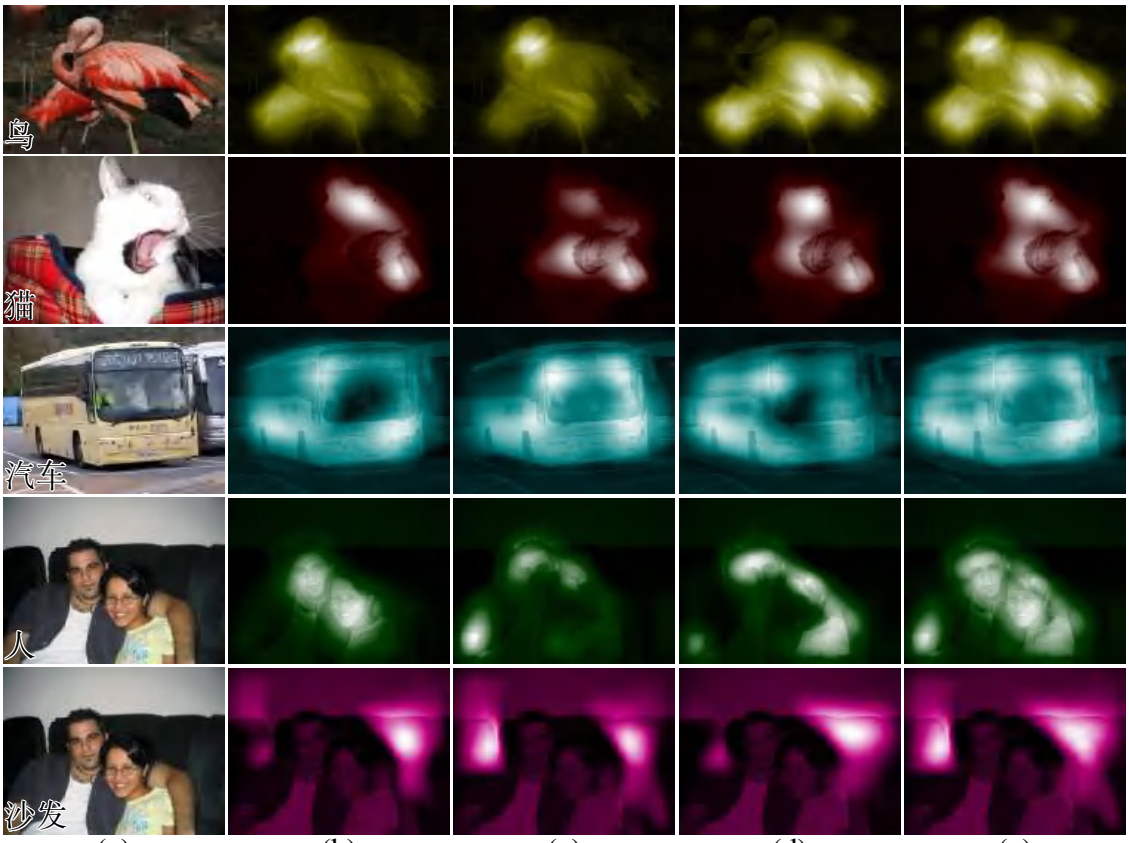


图 4.1 本文的观察。(a) 原始图像；(b-d) 分类网络在不同训练阶段生成的中间注意图；(e) 是将 (b)、(c) 和 (d) 中的注意图通过简单的元素级最大值操作组合起来生成的累积注意图。将 (b)、(c) 和 (d) 的注意力图进行比较，可以很容易地观察到注意力在语义物体上的不断移动。与（b)、(c）和（d）相比，(e）中的累积注意力图可以记录大部分的语义区域。

- 首先，一个强大的分类网络通常寻找对特定类别鲁棒的通用模式，从而使分类模型可以识别该类中所有的图像。如 Arpit 等人[131] 指出的那样，网络会首先优先考虑学习简单的模式。那些难以被正确分类的训练样本将促使网络在选择通用模式方面做出改变，导致注意力图关注的判别性区域的不断变化，直到网络达到收敛。Zeiler 等人[37] 也表明了网络的深层通常需要经过相当多轮迭代学习才能达到收敛。因此，随着网络深层参数不断改变，由深层所生成的注意力将会在训练阶段不断移动。
- 第二，在训练时，现有的分类模型的注意力图主要被之前输入的图像所影响，因此，具有不同内容的图像和训练图像的不同顺序均会导致在训练过程中注意力图中关注的区域发生变化。

图 4.2 加入注意力遮挡层的动机。(a) 原始图像；(b-d) 在不同训练阶段的中间过程注意力图；(e) 是将 (b-d) 中通过简单的元素级最大值操作将注意力图组合起来产生的累积注意力图。可以看到，(b-d) 的注意力图定位在相近的物体区域，导致累积注意力图难以找到新的物体区域。

如图 4.1(b-d) 所示，不同时刻注意力图定位的判别性区域有所不同。除此之外，本文还观察到它们定位的判别性区域往往是互补的，如图 4.1(e) 所示。这个现象促使本文去记录中间阶段的注意力图定位的物体区域，从而基于类别标签来定位完整的语义物体区域。此外，通过观察，本文发现对于一些训练图像，注意力的移动范围并不大。如图 4.2所示，中间过程的注意力图，其注意力总是在鸟的头部或左侧羊所在的位置轻微移动。通过结合图 4.2(b-d)，累积注意力图依旧不能找到鸟的全身和右侧的羊。因此，对于这些图像，增大注意力的移动范围可以更好的定位完整物体区域。

第二节 在线注意力累积

4.2.1 解决方案概述

基于前文的观察，本文引入了一个简单且有效的方法来生成注意力，其能够将分类模型在不同训练阶段中生成的注意力图纳入考虑范围之中。方法的一般流程如图 4.3所示。如图 4.3左边部分所示，本文提出了一个基于在线注意力累积（OAA）的物体定位方法。本文为每张图像中每个目标类别维护一个累积注意力图，并积累由不同训练阶段中的注意力图产生的判别性区域。训练过程中注意力图的互补性，使最终的累积注意力图发掘完整物体成为可能。此外，本文进一步提出一个注意力遮挡层，并将其结合到在线注意力累积过程中。其目标是在训练过程中，扩大注意力的移动范围，从而发掘新的目标区域。具体来

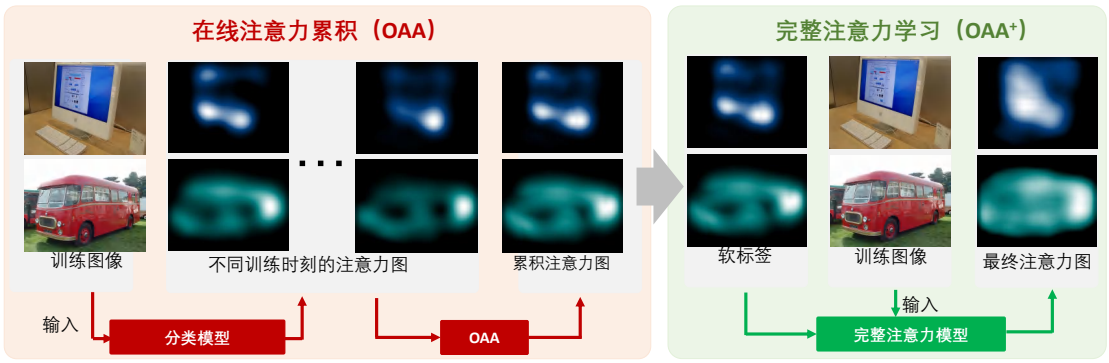


图 4.3 完整注意力学习（OAA$^+$）的一般流程。分类网络在不同训练时间产生的注意图被融合到累积注意力图中，以尽可能完整地挖掘物体区域。所获得的累积注意力图被用作训练完整注意力模型的像素级监督，完整注意力模型进一步提高了注意力图的质量。

说，注意力遮挡层会以一定的概率将注意力图中的强激活区域输入图像进行遮挡，并将其送入分类网络。之后 OAA 将把新的物体区域积累到累积注意力图中。通过利用简单的注意力遮挡层，OAA 可以在不需要额外的训练过程的情况下，定位更加完整的物体区域。

尽管与 CAM[12] 相比，累积注意力图的注意力区域相对完整，但是一些物体区域的注意力值还不够强。为了提高这种情况，本文还设计了一个混合损失函数（增强损失和约束损失的结合），通过将累积注意力图作为软标签来训练一个完整注意力模型，称为 OAA$^+$，如图 4.3右边部分所示。新的注意力模型改进了累积注意力图并且可以生成更完整的物体区域。实验证明，在 PASCAL VOC 2012 基准[104] 的测试集上，本文的方法可以取得 67.2% 的 mIoU 分数。为了更好的理解在线注意力累积方法，本文将首先介绍注意力的生成方法，随后分别介绍在线注意力累积方法中的各种策略，包括在线注意力累积以及完整注意力学习。

4.2.2 在线注意力累积

为了从类别标签定位物体区域，本文采用 CAM[12] 作为默认的注意力图生成器。具体来说，根据以前的工作[12, 13]，本文也采用最后一个卷积层输出的特征图生成注意力图，这在训练阶段非常容易生成。和大多数之前的工作一样[13, 18]，本文也把 VGG-16[2] 作为主干模型。在主干模型的最后一个卷积层后连接三个卷积层来做非线性变换。然后，输出被连接到一个特定于类的卷积层，核大小为 1×1，通道数为 C，用于捕捉注意力，其中 C 表示类别的数量。如

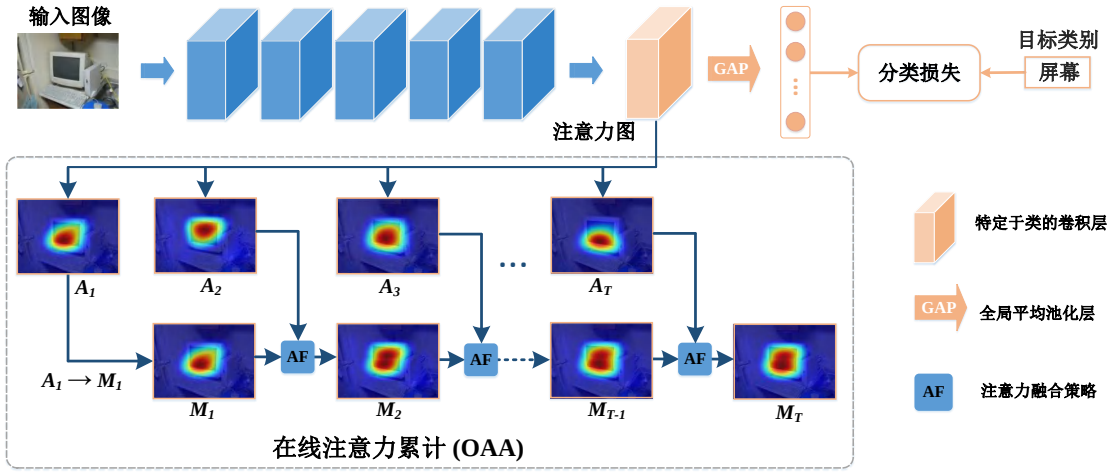


图 4.4 本文的在线注意力累积（OAA）过程的图示。图的顶部展示了分类网络的训练过程，即输入图像到卷积神经网络，然后根据预测和目标计算损失进行优化。图的下半部分展示了累积过程的示意图，注意力图是由特定于类的卷积层生成的，本文的 OAA 利用不同的训练阶段注意力图定位的判别性区域，通过一个简单的注意力融合策略逐步将它们整合到累积注意力图中。

图 4.4顶部所示，注意力图在全局平均池化层（GAP）前由最后一个卷积层所生成。考虑到一些图像可能有不止一个类别的事实，本文把整个训练过程视为 C 个二元分类问题。因此，目标类别 c 的概率可以通过以下方式预测：

$$p^c = \sigma\big(\mathrm{GAP}(F^c)\big) \tag{4.1}$$

其中 $\sigma(\cdot)$ 是 sigmoid 函数，F^c 表示最后一个卷积层的第 c 个特征图。本文采用交叉熵损失函数优化网络。

给定一张图像 I，为了生成目标类别 c 的注意力图，在训练时，本文首先对特征图 F^c 进行 ReLU 操作，然后通过以下方式将其归一化：

$$A^c = \frac{\mathrm{ReLU}(F^c)}{\max(F^c)}. \tag{4.2}$$

基于上述公式，本文可以从分类模型生成注意力图。下面将介绍如何在训练过程中累积注意力图。

根据观察，在不同的训练阶段，注意力图通常关注在目标物体的不同部分。因此，本文提出了一个在线注意力累积（OAA）策略，以在训练中充分使用注意力图。对于任何一张训练图像，OAA 将不同训练阶段生成的注意力图累积到一个累积注意力图中。如图 4.4所示，给定一个训练图像 I，本文为目标类别 c 创建一个累积注意力图 M^c，以记录在每个训练轮次注意力图定位到的判别性区域。

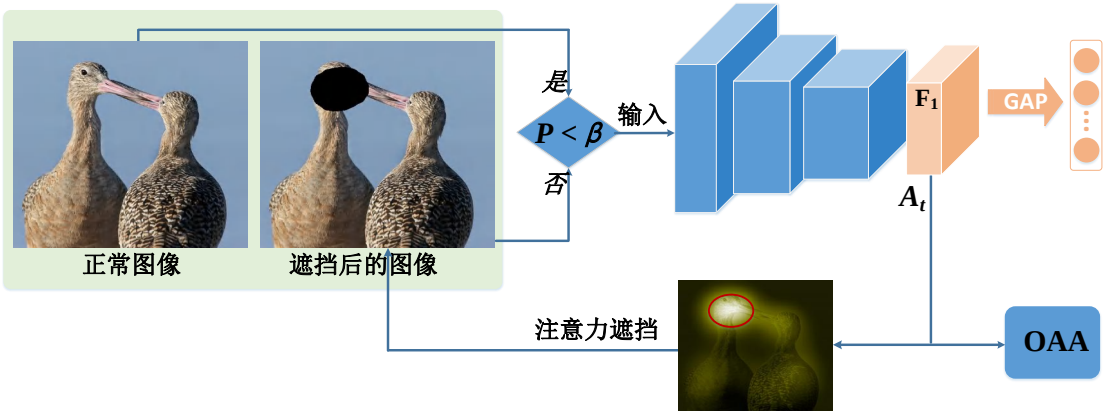


图 4.5 插入一个注意力遮挡层到在线注意力累积（OAA）过程。红色圆圈表示注意力值大于 δ 的区域。

具体来说，当训练图像 I 在第一轮迭代中被送入分类网络时，目标类别 c 生成的注意力图 A_1^c 被用来初始化累积注意力图 M_1^c。为简单起见，本文将在下面的符号中省略类别 c。然后，当该图像第二次输入网络时，OAA 根据以下融合策略将 M_1 和新生成的注意图 A_2 结合起来，更新累积注意力图：

$$M_2 = \mathrm{AF}\big(M_1, A_2\big), \tag{4.3}$$

其中 $\mathrm{AF}(\cdot)$ 表示注意力融合策略。用同样的方法，第 t 轮迭代中生成的注意力图 A_t，被用来更新累积注意力图 M_{t-1}，结果为：

$$M_t = \mathrm{AF}\big(M_{t-1}, A_t\big). \tag{4.4}$$

OAA 不断重复上述更新过程，直到分类模型收敛，得到最终的累积注意力图。在上述更新过程中，注意力融合策略负责保留这些中间注意力图的判别性区域，以定位更加完整的物体区域。

如第 4.1.2 节所述，本文观察到在训练过程中，一些图像的注意力在物体区域上的移动范围很小。通过在线注意力累积方法，很难为这些图像挖掘未出现的物体区域。最近，一些研究者[13, 14, 111, 130] 试图遮挡输入图像或特征中的注意区域，以促使注意力定位到物体上的非判别性区域。受这些工作的启发，本文试图将擦除策略整合到在线注意力累积过程中，以扩大训练过程中注意力转移的范围，累积更多未发现的物体区域，从而获得完整的物体区域。基于这种考虑，本文提出将一个注意力遮挡层插入到在线注意力累积过程中, 如图 4.5所示。具体来说，注意力遮挡层是根据遮挡率 β 决定它的使用频率。对于第 t 轮迭代

Algorithm 1 结合注意力遮挡层的在线注意力累积方法

输入: 输入 I, 分类模型 $f(\cdot)$ 遮挡层 $S(\cdot)$, 遮挡率 β, 阈值 δ, 第 $t-1$ 轮迭代生成的注意力图 A_{t-1},
生成一个随机概率：$\mathrm{r} = \mathrm{Rand}(0,1)$
if $\mathrm{r} < \beta$ **then**
    遮挡输入图像: $I^* = S(I, A_{t-1}, \delta)$
    输入遮挡图像, 得到 $f(I^*)$
else
    输入普通图像 I，得到 $f(I)$
end if
从 $f(\cdot)$ 生成当下的注意力图 A_t
由公式 (4.4) 更新累积注意力图 M_t

中的输入图像 I，已有它上一轮迭代中的注意力图 A_{t-1}。在使用注意力遮挡层的时候，它将遮挡 I 中对应于 A_{t-1} 中注意力值大于阈值 δ 的区域。随后被遮挡的图像被送入到分类网络中，生成注意力图 A_t，它将通过公式 (4.4) 更新累积注意力图，得到 M_t。算法 1详细地描述了该算法。

由于注意力遮挡层在训练中有效地扩大了注意力移动的范围，OAA 可以累积到更多的物体区域。在下面的章节中，我们把结合注意力遮挡层的在线注意力累积方法称为 OAA++。在图 4.6中，可以看到来自 OAA 的最终累积注意力图在加上注意力遮挡层后能定位到更多的物体区域。在消融实验中，本文对遮挡率 β 和注意力阈值 δ 进行敏感性分析，分析它们对于分割结果的影响。

4.2.3 完整注意力学习

OAA 结合了不同训练时刻的注意力图，来挖掘更加完整的物体区域。然而，OAA 的弱点是，分类模型本身不能增强一些注意力值较低的物体区域。考虑到这种情况，本文引入了一个新的损失函数，将累积注意力图作为监督来训练一个完整注意力模型，以进一步改进本文的 OAA，该模型被命名为 OAA^+。

具体来说，如 Wei 等人[132] 所做的那样，本文也将累积注意力图作为软标签。累积注意力图被归一化至 [0,1]，其中每个值都被看作是该位置属于相应目标类别的概率。本文采用图 4.4中所示的没有全局平均池化层和分类损失的分类网络作为本文的完整注意力模型。给定一个由特定于类的卷积层产生的分数图 $\hat{F}$，位置 j 属于某个类别 c 的概率可以用 $q_j^c = \sigma(\hat{F}_j^c)$ 表示，其中 σ 表示 sigmoid

图 4.6 OAA++ 表示 OAA 结合了注意力遮挡层。由 OAA 和 OAA++ 生成的注意力图示例。相比于 OAA, OAA++ 新发现了红框中的区域。

函数。交叉熵损失可以写为如下形式：

$$-\frac{1}{N}\sum_{j\in N}\left(p_j^c\log(q_j^c)+(1-p_j^c)\log(1-q_j^c)\right), \tag{4.5}$$

其中 p_j^c 表示归一化后的累积注意力图中位置 j 的值，N 表示特征图中的像素数量。通过交叉熵损失优化后，增强的注意力图可以由特定于类的卷积层直接生成。然而，在上述多标签交叉熵损失函数的作用下，注意力图只能定位很小的语义物体区域。原因是公式 (4.5) 中的损失函数更倾向于将注意力值低 $(p_j^c < 1-p_j^c)$ 的像素分类为背景。

根据上述讨论，本文提出了一种改进的混合损失函数。鉴于累积注意力图

范围从 0 到 1，本文首先将其分为软增强区域 N_+^c 和软约束区域 N_-^c，其中 N_-^c 包括 $p_j^c = 0$ 的像素，N_+^c 包含其它像素。对于像素集 N_+^c，本文移除了公式 (4.5) 的最后一项，以进一步对注意力区域增强，但不抑制注意力值低的区域。

$$\mathcal{L}_+^c = -\frac{1}{|N_+^c|}\sum_{j\in N_+^c} p_j^c \log(q_j^c). \tag{4.6}$$

由于只有类别标签，累积注意力图中的注意力区域往往包含非目标像素。因此，在公式 (4.6) 中，本文使用 p_j^c 作为真实标签，而不是 1。在累积注意力图中，非语义区域较低的注意力值对网络几乎没有负面影响。对于 N_-^c，其中 $p_j^c = 0$，公式 (4.5) 中的损失函数变成以下形式：

$$\mathcal{L}_-^c = -\frac{1}{|N_-^c|}\sum_{j\in N_-^c} \log(1 - q_j^c). \tag{4.7}$$

因此，本文的完整注意力模型的总混合损失 $\mathcal{L}$ 可以通过以下方式计算出来

$$\mathcal{L} = \sum_{c\in C}(\mathcal{L}_+^c + \mathcal{L}_-^c). \tag{4.8}$$

这样，根据公式 (4.6) 中的损失函数，软增强区域的低值也有助于优化。公式 (4.7) 约束了注意力区域向背景的扩张。基于混合损失函数，本文可以进一步训练一个完整注意力模型，以加强目标物体区域的低注意力值。在测试阶段，注意力图可以直接从完整注意力模型的特定于类的卷积层获得。在图 4.7中，本文给出了一些注意力图的视觉结果。更多的定量分析是在第 4.3.4 节。

第三节 实验

为了验证在线注意力累积方法的有效性，本文将其生成的累积注意力图作为启发式线索应用于弱监督语义分割任务。注意力图被用来生成伪分割标签。本文采取与文献[18] 中类似的方式来生成伪分割标签，即用注意力图来提取物体线索，用显著性图[75] 来提取背景线索。具体来说，本文首先将注意力图和显著性图归一化到 [0,1] 范围内，随后比较不同类别的注意力图，将对应于最大值的类别标签分配给伪分割标签中的像素。在多个图中都超过一定阈值的像素在训练中会被忽略。最后，从上述方法产生的伪分割标签被用来训练分割模型。在分割模型的推理阶段，本文利用多尺度测试，并应用 DenseCRF[139] 来平滑分割图。在下面的小节中，本文提供了一系列的消融研究，并将本文的方法与之前

表 4.1 与以往最先进的方法的定量比较。分割模型是基于 VGGNet[2] 骨干网络。**S:** 显著性图。**IS:** 实例显著性图。**WI** 和 **WV**: 网络抓取的图像和视频。**P:** 分割标签。OAA$^+$ 表示注意图是由完整注意力模型生成。OAA++ 表示结合了注意力遮挡层的 OAA。OAA++$^+$ 表示 OAA$^+$ 利用 OAA++ 的注意力图作为监督。

方法	监督	Val (%)	Test (%)
分割骨干网络: VGGNet[2]			
CCNN[129]	10K	35.3	-
EM-Adapt[79]	10K	38.2	39.6
DCSM[133]	10K	44.1	45.1
SEC[82]	10K	50.7	51.7
AugFeed[68]	10K+S	54.3	55.5
STC[132]	10K+S+WI	49.8	51.2
Roy et al.[134]	10K	52.8	53.7
Oh et al.[135]	10K+S	55.7	56.7
AE-PSL[14]	10K+S	55.0	55.7
Hong et al.[106]	10K+WV	58.1	58.7
WebS-i2[136]	10K+WI	53.4	55.3
DCSP[120]	10K+S	58.6	59.2
TPL[137]	10K	53.1	53.8
GAIN-SEC[48]	10K	55.3	56.8
GAIN[130]	10K	59.4	59.6
DSRG[73]	10K+S	59.0	60.4
MCOF[74]	10K+S	56.2	57.6
AffinityNet[72]	10K	58.4	60.5
MDC[18]	10K+S	60.4	60.8
AISI[138]	10K+IS	61.3	62.1
SeeNet[49]	10K+S	61.1	60.7
OAA	10K+S	61.6	62.0
OAA++	10K+S	**63.0**	**62.7**
OAA$^+$	10K+S	63.1	62.8
OAA++$^+$	10K+S	**63.7**	**63.2**
性能上限	10K+P	70.8	71.2

表 4.2 与以往最先进的方法的定量比较。分割模型是基于 ResNet[1] 骨干网络。**S**：显著性图。**IS**：实例显著性图。**P**：分割监督。OAA^{+} 表示注意图是由完整注意力模型生成。OAA++ 表示结合了注意力遮挡层的 OAA。$OAA++^{+}$ 表示 OAA^{+} 利用 OAA++ 的注意力图作为监督。

方法	监督	Val (%)	Test (%)
分割骨干网络: ResNet[1]			
DCSP[120]	10K+S	60.8	61.9
DSRG[73]	10K+S	61.4	63.2
MCOF[74]	10K+S	60.3	61.2
AffinityNet[72]	10K	61.7	63.7
AISI[138]	10K+IS	63.6	64.5
SeeNet[49]	10K+S	63.1	62.8
SSDD[140]	10K	64.9	65.5
SEAM[141]	10K	64.5	65.7
ScE[142]	10K	66.1	65.9
MCIS[143]	10K+S	66.2	66.9
OAA	10K+S	63.9	65.6
OAA++	10K+S	**64.9**	**66.3**
OAA^{+}	10K+S	65.2	66.4
$OAA++^{+}$	10K+S	**66.1**	**67.2**
性能上限	10K+P	75.4	75.7

最先进的方法进行了比较。本文的在线注意力累积方法是专门为弱监督语义分割任务设计的，为了证明该方法的泛化性，本文将它用于弱监督物体定位任务。

4.3.1 数据集和评价指标

本文在 PASCAL VOC 2012[104] 分割基准上评估分割结果，该基准包含 20 个语义类别和背景。该数据集中的图像被分成三个集合：训练集、验证集和测试集，分别包括 1464、1449 和 1456 幅图像。如同以前的大多数工作一样，本文也使用增强的训练集[105] 进行模型训练。因此，训练集中总共有 10,582 张图像。在推理过程中，本文分别在验证集和测试集上将本文提出的方法与之前最先进的方法在平均交并比（mIoU）[7] 的评价指标方面进行了比较。由于测试集的分割

标签并不公开，本文将预测的分割结果提交到官方的评测服务器[1] 以获得在测试集上的 mIoU 分数。

4.3.2 网络设置

本文的方法基于 Caffe 库[144] 实现。对于分类网络，超参数设置如下：批大小（5），权重衰减（0.0002），以及动量（0.9）。初始的学习率为 1e-3，在 20000 次迭代后衰减 10 倍。本文将分类网络总共运行 30000 次迭代。本文使用没有全局平均池化层和分类损失的分类网络作为完整注意力网络。完整注意力网络的超参数与分类网络的参数相同。对于添加注意力遮挡层的分类网络，本文也运行 30000 次迭代。本文使用 DeepLab-LargeFOV[8] 作为分割网络，就像以前大多数工作中做的那样。分割网络在批大小为 10 的情况下训练，直至 15000 次迭代后停止。分割网络的其它超参数都与[8] 相同。本文分别报告了基于 VGG-16[2] 和 ResNet-101[1] 两个主干网络的分割结果。

4.3.3 实验结果

本文首先将分割结果与之前最好的弱监督语义分割方法进行比较。在表 4.1中，本文列出了这些方法和本文的方法在验证集与测试集上的所有分割结果。可以很容易地看到，无论使用哪种主干网络，本文提出的方法与以前最先进的方法相比，都取得了有竞争力的结果。一些工作，例如 STC[132] 和 WebS-i2[136] 使用了更多的额外图像。此外，Hong 等人[106] 利用额外视频数据提供的丰富的时序动态信息，很容易从视频数据中找到整体语义对象。虽然只使用了 10k 张训练图像，本文的 OAA++$^+$ 在验证集上的结果比上述三种方法分别提高了 13.9%、10.3% 和 5.6%。这一事实表明，由本文的完整注意力模型产生的注意图可以有效地检测出更加完整的物体区域，这有利于伪分割标签的质量。

AE-PSL[14] 需要训练多个分类模型，并进行多个擦除和挖掘步骤，以构成最终的注意力图。与 AE-PSL 相比，本文的 OAA++ 实现了更好的 mIoU 得分（63.0% *v.s.* 55.0%），而且不需要训练多个分类模型。此外，GAIN[130] 以端到端的方式采用了自我引导擦除策略，但本文方法的分割结果在 mIoU 分数上超过了 GAIN 大约 4% 的 mIoU 分数（63.7% *v.s.* 59.4%）。在 MDC[18] 中，Wei 等人利用空洞卷积来发掘完整的物体。然而，它通常会引入一些不相关的像素，因为大空洞率的卷积往往关注到目标区域的外部。不同的是，本文的方法不使用大

[1] http://host.robots.ox.ac.uk:8080/

表 4.3 弱监督语义分割方法在 PASCAL VOC 2012 验证集上的每个类别的 IoU 分数。得分最高的三个的方法依次被标记为红色，绿色和蓝色。自行.: 自行车；公汽：公共汽车；摩托.: 摩托车。

方法	背景	飞机	自行.	鸟	船	瓶子	公汽	轿车	猫	椅子	牛
CCNN[129]	68.5	25.5	18.0	25.4	20.2	36.3	46.8	47.1	48.0	15.8	37.9
DCSM[133]	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9
SEC[82]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5
STC[132]	84.5	68.0	19.5	60.5	42.5	44.8	68.4	64.0	64.8	14.5	52.0
TPL[137]	82.8	62.2	23.1	65.8	21.1	43.1	71.1	66.2	76.1	21.3	59.6
MCOF[74]	85.8	74.1	23.6	66.4	36.6	62.0	75.5	68.5	78.2	18.8	64.6
GAIN[130]	87.6	76.7	33.9	74.5	58.5	61.7	75.9	72.9	78.6	18.8	70.8
DSRG[73]	87.5	73.1	28.4	75.4	39.5	54.5	78.2	71.3	80.6	25.0	63.3
AffinityNet[72]	87.2	57.4	25.6	69.8	45.7	53.3	76.6	70.4	74.1	28.3	63.2
MDC[18]	89.5	85.6	34.6	75.8	61.9	65.8	67.1	73.3	80.2	15.1	69.9
OAA	89.7	84.3	35.6	77.4	60.4	65.6	75.4	72.2	80.4	17.2	69.3
OAA$^+$	90.0	84.1	34.7	77.6	62.7	66.4	80.5	74.6	82.4	18.8	73.0
OAA++	90.1	83.8	35.3	77.3	60.6	66.9	79.0	75.1	82.4	17.7	72.5
OAA++$^+$	90.2	85.6	36.0	75.6	62.0	66.6	82.5	73.6	83.9	18.7	75.3
性能上限	92.4	82.8	35.6	82.1	64.5	72.8	88.0	81.0	85.3	32.5	79.0

方法	桌子	狗	马	摩托.	人	盆栽	羊	沙发	火车	屏幕	平均
CCNN[129]	21.0	44.5	34.5	46.2	40.7	30.4	36.3	22.2	38.8	36.9	35.3
DCSM[133]	32.7	54.9	48.6	57.4	51.8	38.2	55.4	32.2	42.6	39.6	44.1
SEC[82]	28.3	65.8	57.8	62 3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
STC[132]	22.8	58.0	55.3	57.8	60.5	40.6	56.7	23.0	57.1	31.2	49.8
TPL[137]	35.1	70.2	58.8	62.3	66.1	35.8	69.9	33.4	45.9	45.6	53.1
MCOF[74]	29.6	72.5	61.6	63.1	55.5	37.7	65.8	32.4	68.4	39.9	56.2
GAIN[130]	14.1	68.7	69.6	69.5	71.3	41.5	66.5	16.4	70.2	48.7	59.4
DSRG[73]	25.4	77.8	65.4	65.2	72.8	41.2	74.3	34.1	52.1	53.0	59.0
AffinityNet[72]	44.8	75.6	66.1	65.1	71.1	40.5	66.7	37.2	58.4	49.1	58.4
MDC[18]	8.1	75.0	68.4	70.9	71.5	32.6	74.9	24.8	73.2	50.8	60.4
OAA	21.3	73.4	67.4	72.5	72.9	37.4	73.7	28.4	64.4	55.2	61.6
OAA$^+$	22.7	76.8	70.8	72.7	74.8	37.7	73.4	28.4	68.6	55.3	63.1
OAA++	22.7	77.1	71.5	72.1	73.8	39.9	72.2	29.8	68.8	54.7	63.0
OAA++$^+$	18.7	77.9	73.3	73.0	75.4	40.0	76.4	29.7	68.9	54.6	63.7
性能上限	55.8	80.4	77.9	74.3	80.0	52.2	79.9	44.7	79.9	65.2	70.8

表 4.4 弱监督语义分割方法在 PASCAL VOC 2012 测试集上的每个类别的 IoU 分数。得分最高的三个的方法依次被标记为红色，绿色和蓝色。自行.: 自行车；公汽：公共汽车；摩托.: 摩托车。

方法	背景	飞机	自行.	鸟	船	瓶子	公汽	轿车	猫	椅子	牛
CCNN[129]	-	42.3	24.5	56.0	30.6	39.0	58.8	52.7	54.8	14.6	48.4
DCSM[133]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3
SEC[82]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0
STC[132]	85.2	62.7	21.1	58.0	31.4	55.0	68.8	63.9	63.7	14.2	57.6
TPL[137]	83.4	62.2	26.4	71.8	18.2	49.5	66.5	63.8	73.4	19.0	56.6
MCOF[74]	86.8	73.4	26.6	60.6	31.8	56.3	76.0	68.9	79.4	18.8	62.0
GAIN[130]	88.2	79.3	33.7	67.9	50.5	62.5	76.0	72.2	77.6	20.3	65.8
AffinityNet[72]	88.0	61.1	29.2	73.0	40.5	54.1	75.2	70.4	75.1	27.8	62.5
MDC[18]	89.8	78.4	36.2	82.1	52.4	61.7	64.2	73.5	78.4	14.7	70.3
OAA	90.1	77.8	36.0	80.6	49.9	61.4	73.6	73.5	78.5	21.4	68.6
OAA$^+$	90.3	77.0	35.4	80.5	50.0	61.3	77.2	75.4	79.6	21.7	71.8
OAA++	90.3	80.6	35.2	78.8	49.9	59.9	76.4	76.6	80.3	21.1	69.2
OAA++$^+$	90.3	81.5	36.8	76.7	48.9	61.1	78.7	75.1	80.2	20.6	70.7
性能上限	92.7	86.3	37.4	79.8	61.8	68.5	87.7	81.3	84.7	30.3	76.9

方法	桌子	狗	马	摩托.	人	盆栽	羊	沙发	火车	屏幕	平均
CCNN[129]	34.2	52.7	46.9	61.1	44.8	37.4	48.8	30.6	47.7	41.7	-
DCSM[133]	34.8	50.3	48.9	69.0	49.7	38.4	57.1	34.0	38.0	40.0	45.1
SEC[82]	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
STC[132]	28.3	63.0	59.8	67.6	61.7	42.9	61.0	23.2	52.4	33.1	51.2
TPL[137]	35.7	69.3	61.3	71.7	69.2	39.1	66.3	44.8	35.9	45.5	53.8
MCOF[74]	36.9	74.5	66.9	74.9	58.1	44.6	68.3	36.2	64.2	44.0	57.6
GAIN[130]	19.5	72.6	73.0	75.2	71.4	42.4	72.8	21.4	61.5	48.6	59.6
AffinityNet[72]	51.4	78.4	68.3	76.2	71.8	40.7	74.9	49.2	55.0	48.3	60.5
MDC[18]	11.9	75.3	74.2	81.0	72.6	38.8	76.7	24.6	70.7	50.3	60.8
OAA	28.2	73.3	72.3	79.0	73.4	44.3	74.5	27.7	63.5	53.9	62.0
OAA$^+$	29.5	75.4	73.4	78.6	74.3	44.8	76.5	27.9	65.4	52.6	62.8
OAA++	27.6	75.5	72.6	78.8	74.5	46.6	75.8	28.1	67.2	52.6	62.7
OAA++$^+$	27.6	76.4	75.6	79.7	75.1	45.6	76.7	28.5	68.7	52.8	63.2
性能上限	61.5	80.0	75.2	81.9	80.6	55.4	81.6	53.6	76.1	62.4	71.2

空洞率的卷积，可以削弱不相关的像素的影响。

如表 4.1所示, 本文的方法在验证集和测试集上相对于 MDC[18] 都提升了大约 3% 的性能。与使用显著性图作为背景线索的方法（AugFeed[68], AE-PSL[14], DCSP[120], DSRG[73], MCOF[74], MDC[18], AISI[138] and SeeNet[49]）相比，本文 OAA 的性能超出了它们很多。与那些基于擦除的方法[14, 48, 49, 130] 的比较，揭示了使用中间注意力图是更有效的。本文提出的注意力遮挡层也是基于擦除策略的。它可以提高注意力的移动范围来积累更多的目标物体区域，从而进一步提高在线注意力累积策略。插入注意力遮挡层到本文的框架中进一步在验证集和测试集上大幅度提升了 OAA 和 OAA^+。除此之外，在表 4.2中，本文还展示了基于 ResNet-101 主干网络的分割结果。本文所提出的方法在 PASCAL VOC 2012 取得了最好的结果。本文还在表 4.3和表 4.4 中提供了每个类别的 IoU 分数的详细信息。

4.3.4 消融实验

本文进行了一系列的消融实验，并给出了详细的分析。所有的消融实验都是基于 VGGNet 的 DeepLab-LargeFOV 分割模型，且所有的分割结果都在单尺度下进行评估的。

4.3.4.1 累积策略

本文首先研究了不同注意力累积策略对于 OAA 的影响。注意力累积策略在 OAA 中被用于积累不同训练阶段中的注意力图所发掘的判别性区域。除了最大值融合策略外，本文还研究了平均融合策略，可以表述为：

$$M_t = \frac{1}{t}\big((t-1)M_{t-1} + A_t\big). \tag{4.9}$$

如表 4.5所示，在没有 OAA 的情况下，使用 CAM 的注意力图[12] 在验证集上得到了 53.9% 的 mIoU 分数。当使用基于平均融合策略的 OAA 时，结果提升到 57.0%。当用最大值融合策略取代平均融合策略的时候，得到了 58.6% 的 mIoU 分数，这相对于基于 CAM[12] 的结果有了很大的提高。此外，最大值融合策略比平均融合策略更加高效，这个结果是因为平均融合策略平均了所有的中间注意图的注意值，从而使得最终累积注意图中的注意值减小。因此，在下文中，本文将最大值融合策略作为 OAA 的默认融合策略。

表 4.5 在 PASCAL VOC 2012 验证集上，本文方法在不同设置下 mIoU 分数的比较。**AVE:** OAA 使用平均融合策略. **MAX:** OAA 使用最大值融合策略. **MCE:** OAA^+ 使用多标签交叉熵损失，即公式 (4.5). **EP:** OAA^+ 只使用增强损失，即公式 (4.6). **HL:** OAA^+ 使用本文的混合损失，即公式 (4.8).

No.	AVE	MAX	MCE	EP	HL	mIoU (%)
1						53.9
2	✓					57.0
3		✓				58.6
4		✓	✓			51.2
5		✓		✓		53.4
6		✓			✓	59.6

表 4.6 完整注意力学习中软增强区域 N_+^c 和软约束区域 N_-^c 的不同划分阈值的定性比较。本文计算 OAA^+ 的注意力图的噪声率和召回率。使用最大注意力值的 50% 将注意力图阈值化为二值图。

No.	阈值	噪声率 (%)	召回率 (%)	mIoU (%)
1	0.0	43.9	59.4	59.6
2	0.1	40.1	54.9	59.5
3	0.2	37.4	51.0	58.9
4	0.3	35.4	47.4	58.4

4.3.4.2 OAA^+ 的损失函数

如第 4.2.3 节中所述，累积注意力图被用作软标签来训练完整注意力模型，以产生具有更完整和准确的物体区域的注意力图。在表 4.5中, 本文展示了基于不同损失函数的定量结果。可以观察到，当将标准多标签交叉熵损失函数（Multi-label Cross-Entropy loss，MCE）[132] 替换为本文提出的混合损失（hybrid loss，HL）时，性能提高了 8.4% 。当应用多标签交叉熵损失时，输出的注意力图总是覆盖很小的物体区域。与之相反，本文所提出的混合损失可以进一步提升由 OAA 生成的累积注意力图的质量。本文还进行了一个关于混合损耗的消融实验，测试了本文方法仅使用增强型损失下的效果，即公式 (4.6)。使用混合损失的结果比只使用增强损失的结果要好 6.2% 的 mIoU 分数。这一事实表明约束性损失的重要性，即公式 (4.7)。

此外，本文还研究注意力图中的背景噪声对混合损失造成的影响。本文使用

表 4.7 在 PASCAL VOC 2012 验证集上，使用不同数量的训练图像下的分割结果。注意图像是从数据集中随机选择的。weak: 类别标签；pixel: 分割标签。

No.	训练图像	训练图像占总图像比例	mIoU (%)
1	2,116 weak	20	54.6
2	5,291 weak	50	57.3
3	8,466 weak	80	58.9
4	10,582 weak	100	59.6
5	8,466 weak + 2,116 pixel	-	61.6
6	5,291 weak + 5,291 pixel	-	63.7
7	2,116 weak + 8,466 pixel	-	65.1
8	10,582 pixel	-	66.1

不同的阈值来划分 N_+^c 和 N_-^c，其中阈值越低，N_+^C 的背景噪声越高。在表 4.6中，可以看到当阈值从 0.3 降低到 0.0 的时候，混合损失会将更多的背景噪声引入到注意力图中（35.4% 到 43.9%）。因此，混合损失对于背景噪声是很敏感的。然而，阈值的下降可以帮助注意力图挖掘更多具有潜力的物体像素（47.4% 到 59.4%）。在生成伪分割标签时，本文使用显著性图来生成背景区域，这可以帮助过滤掉一些背景噪声。伪分割标签不容易受到背景噪声的影响，但更容易受到目标对象召回的影响。因此，当把阈值从 0.3 降到 0.0 时，分割结果增加了很多。

4.3.4.3 不同策略的结果

如表 4.5所示，本文列出了使用以不同策略来训练的分割网络的注意力图的 mIoU 分数。在表 4.5的第三行和最后一行，可以看出，使用 OAA^+ 可以将验证集上的 OAA 结果进一步提高 1.0%。这一结果表明，完整注意力模型与所提出的混合损失函数结合可以提高累积注意力图的质量。

4.3.4.4 训练图像的数量

为了进一步研究基于本文的注意力图生成的伪分割标签的质量，本文尝试使用不同数量的分割标签来训练分割网络。本文使用 OAA^+ 产生的注意力图来生成伪分割标签。正如表 4.7中所示，随着更多的图像被用于训练，mIoU 的分数会逐渐提高。更有趣的是，当只使用 2116 张训练图像时，本文的分割网络仍然可以达到 54.6% 的 mIoU 分数，这比基于 CAM[12] 并使用整个训练集训练的分割结果更好。这一结果间接表明，本文的注意力图质量很高，有利于语义分割

表 4.8 关于注意力遮挡层的超参数的消融实验。

No.	δ	β	mIoU (%)
1	0.1	0.5	59.2
2	0.4	0.5	59.4
3	0.5	0.5	59.7
4	0.6	0.5	**59.9**
5	0.7	0.5	59.4
6	0.8	0.5	58.9
7	0.6	0.3	59.4
8	0.6	0.1	59.3
9	0.6	0.7	59.3
10	0.6	0.9	59.1

任务。本文还使用伪分割标签和真实标签一起来训练分割网络。随着更多的真实标签被用于训练，mIoU 得分增加，这表明伪分割标签仍有很大的改进空间。

4.3.4.5 视觉比较

如图 4.7所示，本文展示了一些定性的结果。图中分别给出了 CAM[12]、OAA、OAA++ 和 OAA^{+} 的注意力图。图中所展示的图像包括多样的场景，例如图像中含有不同大小的物体、多个物体或者多个类别。从所有展示的例子来看，本文的累积注意里图可以在不同的尺度上发现几乎完整的目标对象，而由 CAM[12] 产生的注意图不能。第四行展示了包含多个物体的图像。可以看到在这种条件下，本文的累积注意力图仍然可以覆盖最多的语义区域。最后两行，本文还展示了一些包含多个类别的例子。本文的积累注意力图可以成功地区别不同的类别并完整地定位目标物体。由 OAA++ 和 OAA^{+} 产生的注意图比 OAA 的累积注意力图能发现更多的完整物体区域。此外，在图 4.8中，本文展示了这些方法的分割结果之间的定性比较。

4.3.4.6 注意力遮挡层的超参数

注意力遮挡层有两个超参数：注意力阈值（δ）和遮挡率（β）。本文进行消融实验来分析它们对分割结果的影响。在研究其中一个参数时，本文会将另一个参数值固定。研究 δ 时，β 为 0.5。研究 β, 将 δ 设置为 0.6，因为这能达到最好的性能。

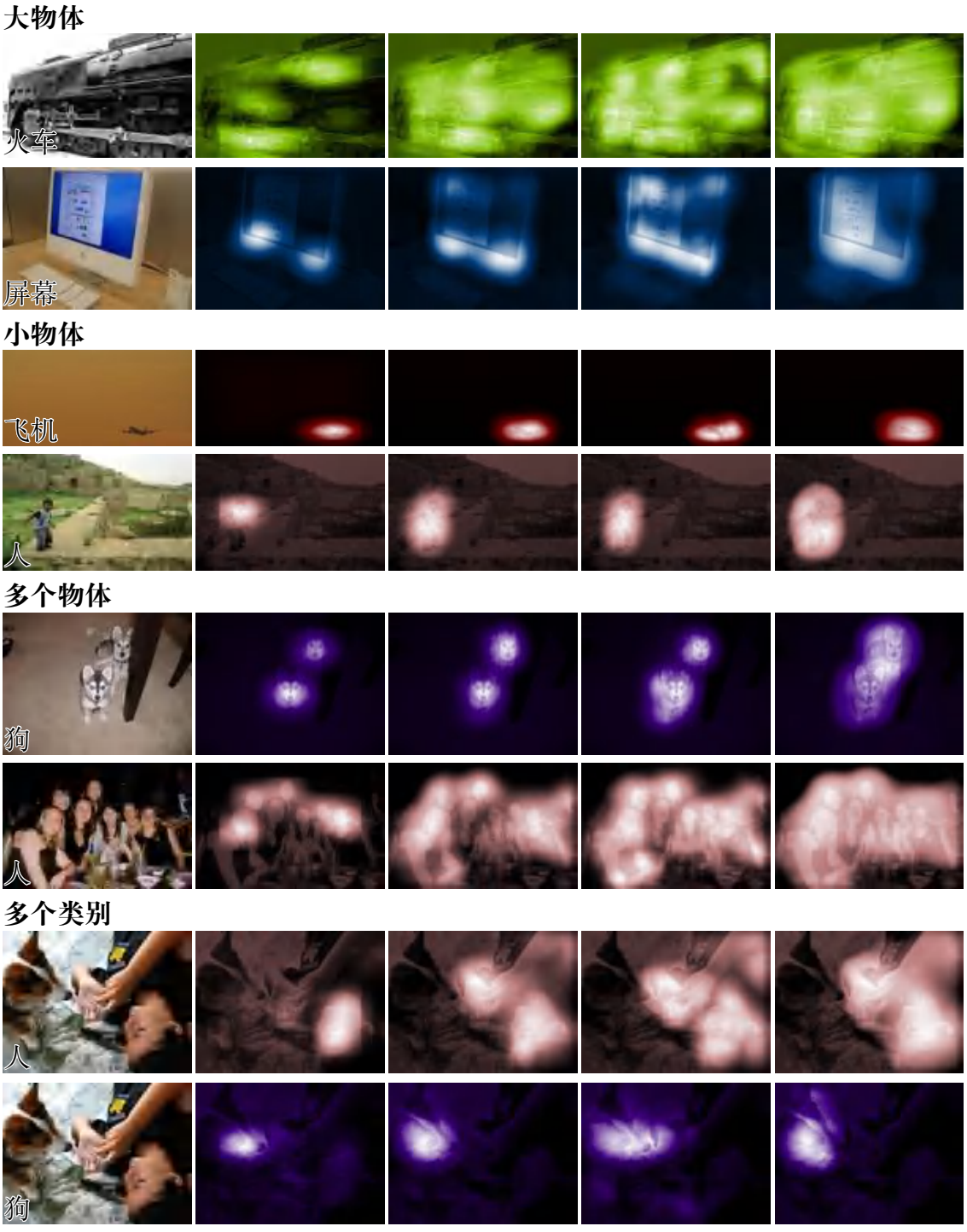


图 4.7 由 CAM[12]，OAA，OAA++，和 OAA^+ 生成的注意力图的视觉比较。OAA++ 代表 OAA 结合注意力遮挡层。OAA^+ 代表完整注意力模型。

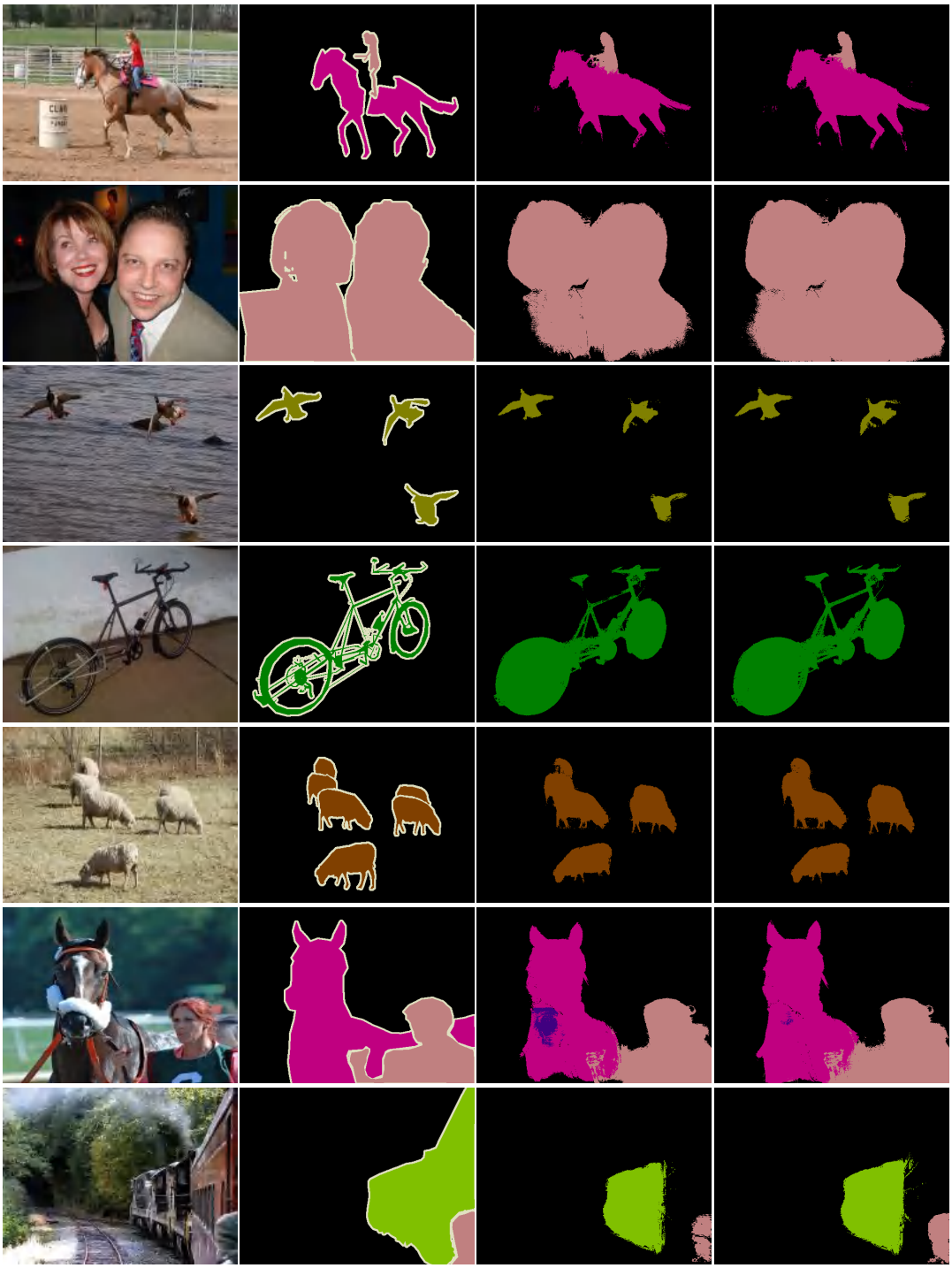


图 4.8 在 PASCAL VOC 2012 上，OAA 和 OAA^+ 的分割结果示例。

表 4.9 在 PASCAL VOC 2012 验证集上注意力遮挡层不同变种的比较。**OAA-drop**：具有注意力遮挡层的 OAA；**OAA-drop-feat**：具有特征层次上的注意力遮挡层的 OAA；**OAA-drop-model**：由伴随注意力遮挡层进行训练的模型产生的注意力图。* 表示结果来自[49]。

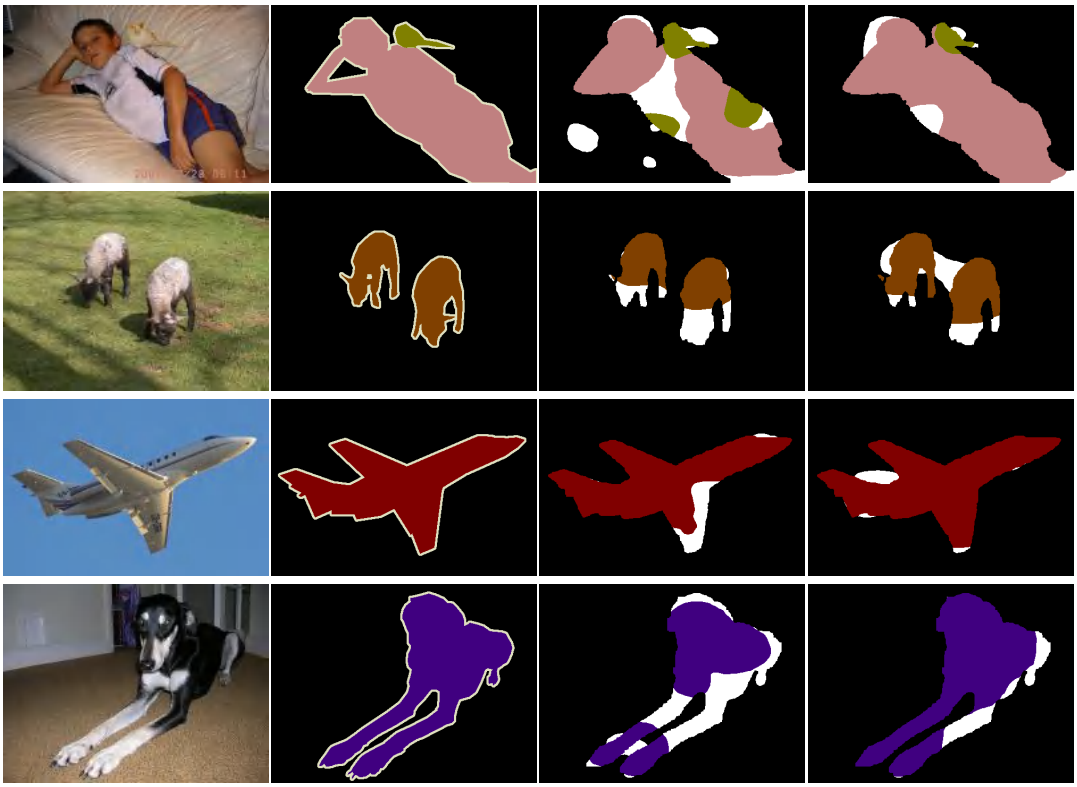
策略	mIoU (%)
OAA	58.6
OAA-drop	59.9
OAA-drop-feat	59.5
OAA-drop-model	58.3
SeeNet*[49]	57.3
ACoL*[13]	56.1

如表 4.8中所示，当 δ 接近 1 时，注意力遮挡层带来的提升会大大减少（59.9% *v.s.* 58.9%）。当 δ 被设置为一个较大值时，遮挡区域变得非常小，导致 mIoU 的分数接近于没有注意力遮挡层的 OAA。当 δ 是一个较小值的时候们可以看到分割结果也下降了。造成这个结果的原因是，当 δ 被设置为一个较小值的时候。注意力区域包含非常大的物体区域。遮挡层可能会迫使注意力移动到无目标的区域，造成累积注意力图中具有过多的噪声。对于遮挡率 β，当 β 被设置为 0.5 的时候，分割结果可以达到最好的性能。

4.3.4.7 注意力遮挡层的变体

在第 4.2.2 节中，本文提出将注意力遮挡层整合到 OAA 中 (OAA++，这里表示为 OAA-drop)，这样可以进一步提高累积注意力图的质量。注意力遮挡层遮挡了输入图像中具有强注意力值的区域。本文还测试了在特征中进行区域遮挡时的性能（OAA-drop-feat）。本文从分类网络的最后一个卷积层中选择特征。如表 4.9所示，遮挡特征的性能比遮挡图像的性能要略低（59.9% *v.s.* 59.5%）。

此外，本文的注意力遮挡层是基于擦除的策略。因此，本文将 OAA-drop 与基于擦除的方法[13, 49] 进行了比较，这些方法都是由最终的分类模型生成注意力图。与基于擦除的方法相比，例如 SeeNet[49]，OAA-drop 远远超出了它们。本文还用具有注意力遮挡层训练的最终分类模型（OAA-drop-model）生成注意力图。使用 OAA-drop 的累积注意力图的分割结果比使用 OAA-drop-model 生成的注意力图的结果要好 1.6%。在图 4.9中，本文还展示了几个分别使用 OAA-drop 和 OAA-drop-model 的注意力图产生的伪分割标签。来自本文 OAA-drop 的伪分



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

图 4.9 伪分割标签的比较。(a) 原始图像；(b) 分割标签；(c) 基于由注意力遮挡层训练的最终分类模型的注意力图生成的伪分割标签；(d) 基于带有注意力遮挡层的 OAA 的累积注意力图生成的伪分割标签。

割标签更加准确和完整。这些事实都验证了在线注意力累积方法和注意力遮挡层的有效性。

4.3.4.8 注意力遮挡层的训练

本文继续探讨将注意力遮挡层和 OAA 结合时，分类网络的不同训练迭代次数对注意力图的影响。在表 4.10中，可以看到，当训练迭代次数从 20000 次增加到 30000 次时，性能从 59.4% 提高到 59.9%。当训练迭代次数设置为 20000 次时，分类模型并没有收敛。因此，当把训练迭代次数从 20000 次增加到 30000 次时，注意力仍然会发现新的物体区域。当进一步增加训练迭代次数到 45000/60000 时，性能没有提高。额外的训练迭代次数不能定位新的物体区域，这是因为当分类模型收敛时，注意力不再发生大的变化。

表 4.10 在不同训练迭代次数下基于注意力遮挡层的 OAA 的分割结果。

迭代次数	20000	30000	45000	60000
mIoU (%)	59.4	59.9	59.7	59.8

表 4.11 不同策略下注意力图质量的对比。

策略	CAM	OAA	OAA-drop
mIoU (%)	53.9	58.6	59.9
召回率 (%)	36.7	52.3	61.9
噪声率 (%)	32.7	37.4	41.1

4.3.4.9 累积注意力图的质量

本文使用召回率和噪声率标准来评测累积注意力图的质量。具体地说，注意力图首先被阈值化为一个二值图，其中阈值设置为最大注意力值的 50%。与分割标签相比，本文计算召回率来表示注意力图定位的物体区域的占整个物体区域比率，计算噪声率来表示注意力图定位的物体区域含有噪声的比率。可以看到和 CAM 相比，OAA 在不引入过多噪声的情况下取得了更高的召回率。当插入注意力遮挡层到 OAA 中的时候，召回率被进一步提升。这一比较表明了累积注意力图的质量。

表 4.12 基于不同分割网络的分割结果。分割结果经过了多尺度测试和 DenseCRF 后处理。

方法	DeepLab-LargeFOV[8]	DeepLabv2[123]
OAA	63.9	67.6
OAA++	64.9	68.4
OAA^{+}	65.2	68.2
$OAA++^{+}$	66.1	68.9
性能上限	75.4	77.9

4.3.4.10 更先进的分割网络

DeepLabv2[8] 是 DeepLab-LargeFOV[123] 的升级版本。DeepLabv2 使用空洞金字塔池化模块来分割不同尺度的物体。本文用 DeepLabv2 进行了分割实验。如表 4.12所示，当用 DeepLabv2 代替 DeepLab-LargeFOV 时，性能得到了很大的

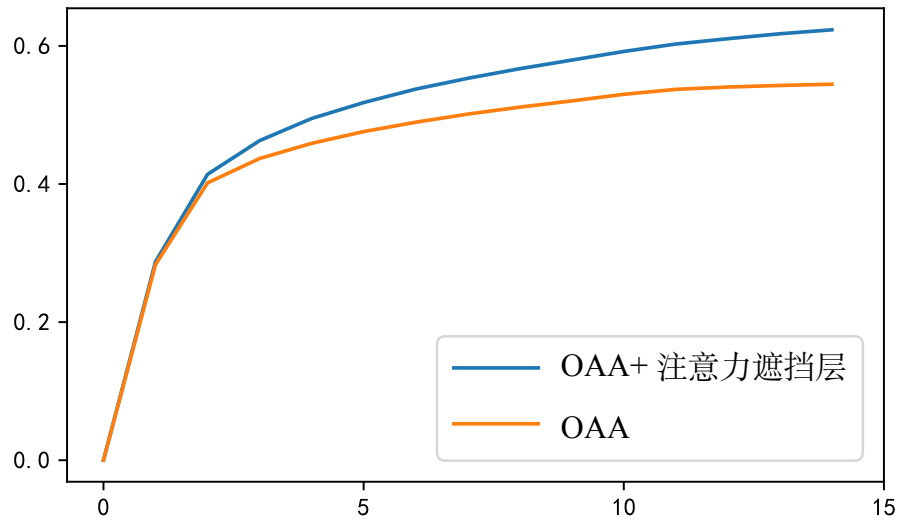


图 4.10 不同训练阶段的注意力图的演进。OAA 表示在线注意力累积。曲线表示被发现的物体区域和整个物体区域的比例。

表 4.13 由注意力图发掘的目标物体区域在每个类别上召回率比较。自行.: 自行车；公汽：公共汽车；摩托.: 摩托车。

方法	飞机	自行.	鸟	船	瓶子	公汽	轿车	猫	椅子	牛
CAM[12]	55.2	43.0	45.4	46.3	43.4	32.5	38.1	22.7	31.9	39.5
OAA	68.2	60.4	58.0	59.5	57.9	49.2	59.2	34.2	51.4	51.4
OAA++	78.5	69.2	73.1	67.3	68.0	59.6	69.0	49.9	58.4	60.1
方法	桌子	狗	马	摩托.	人	盆栽	羊	沙发	火车	屏幕
CAM[12]	31.7	30.0	34.9	37.7	35.9	42.4	45.2	26.7	31.2	46.9
OAA	47.6	39.2	52.7	54.0	54.3	55.7	61.7	42.5	38.4	60.4
OAA++	57.1	51.4	62.7	61.6	62.9	64.8	70.1	50.0	47.7	65.0

提升。

4.3.4.11 注意力的演进过程

在不同的训练阶段中，注意力不断地在目标物体区域上移动。本文进行了一个消融实验，来研究注意力图随着训练的进行产生的演化。本文计算不同训练阶段的累积注意力图定位的物体区域占整个目标物体区域的比例，即召回率。目标物体区域的真值是从分割标签中提取的。

如图 4.10，本文给出了在训练过程中召回率的变化。在训练过程中，无论是否有注意力遮挡层的 OAA，累积注意力图所发现的目标物体区域都会逐渐变大。

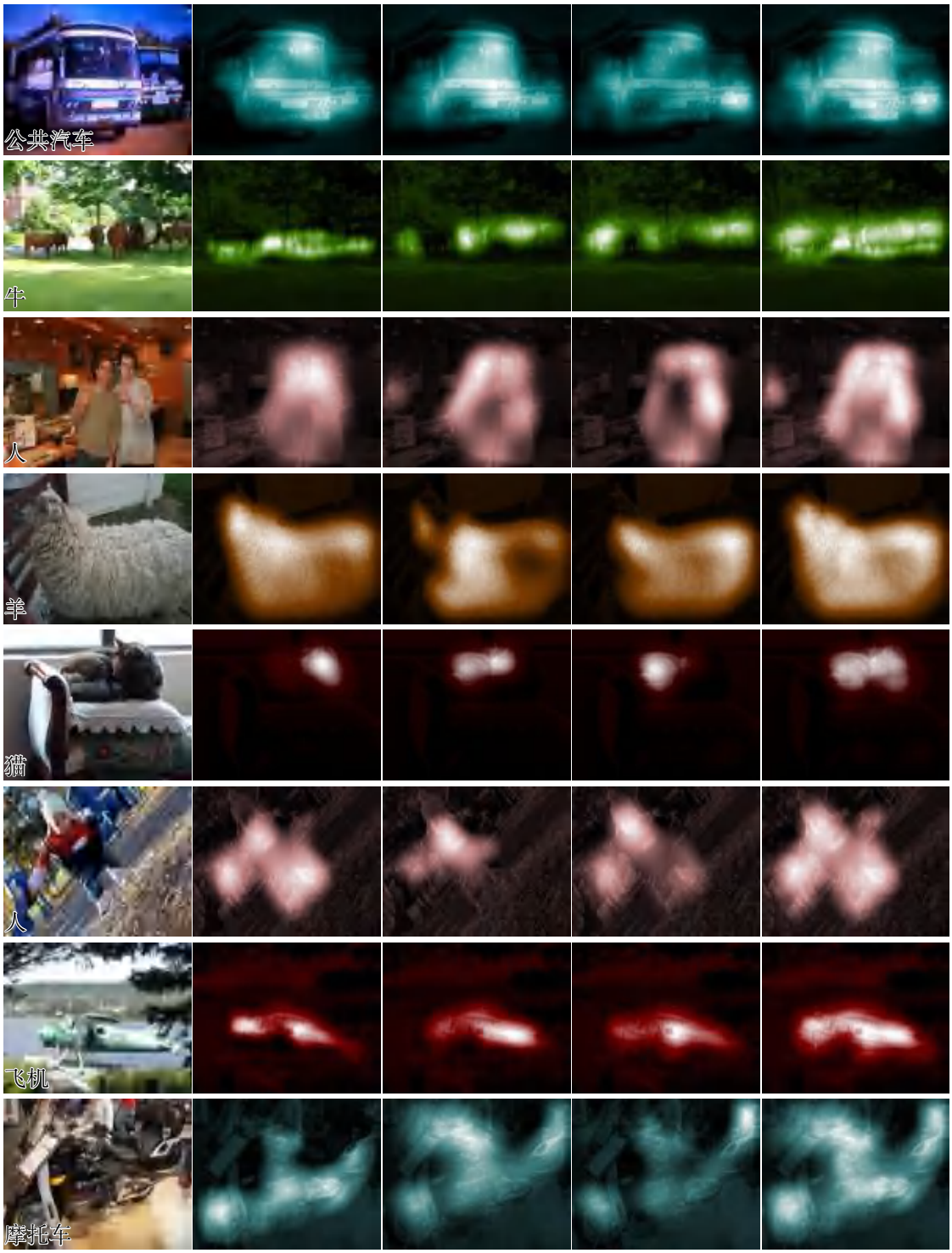


图 4.11　训练时注意力图的变化过程。(a) 原始图像；(b-d) 中间阶段注意力图；(e) 累积注意力图。

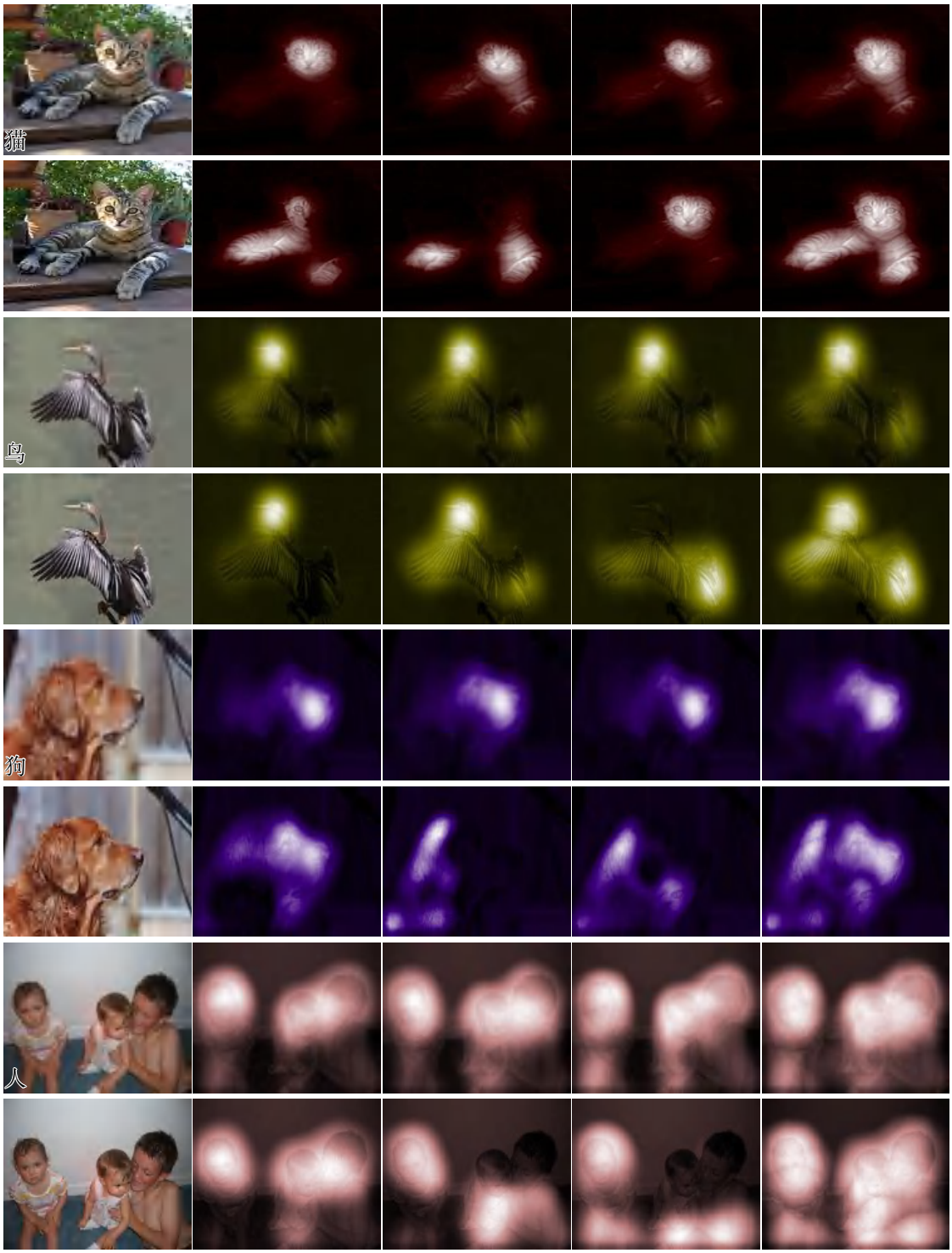


图 4.12　训练过程中注意力图的演化过程。(a) 原始图像；(b-d) 中间阶段注意力图；(e) 累积注意力图。注意力图分别由普通训练过程和具有注意力遮挡层的训练过程产生。

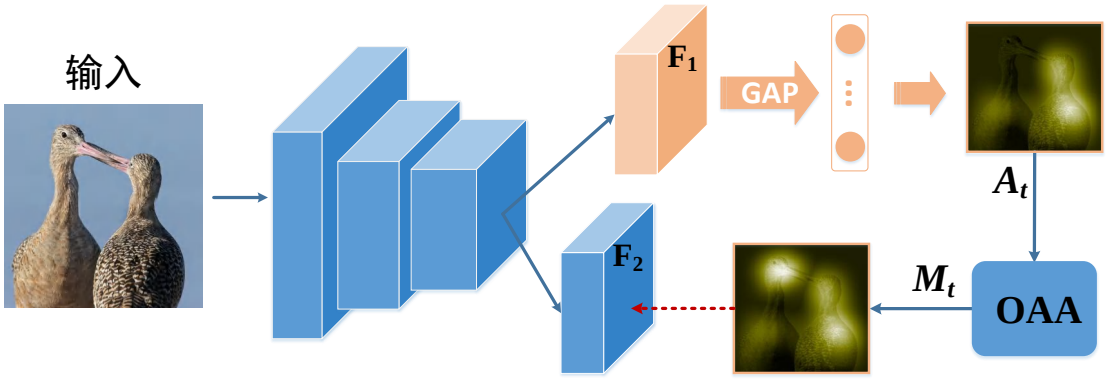


图 4.13 OAA 用于弱监督物体定位。GAP 表示全局平均池化层。红色虚线表示用 OAA 的累积注意力图作监督。

在训练阶段的后期，新发现的物体区域逐渐减少，召回率增加缓慢。当图像分类器趋于收敛时，其参数在一个小范围内波动，导致注意力在物体区域上没有很大变化。当把注意力遮挡层整合到 OAA 框架中时，随着训练过程的进行，累积注意力图可以发现更多的物体区域。遮挡部分判别性区域可以有效地迫使注意力转移到新的物体区域，扩大注意力的移动范围。表 4.13 展示了 CAM、OAA 和 OAA++ 的每个类别注意力图的召回率。

本文还展示了更多关于训练期间注意力演变的直观例子。在图 4.11中，可以看到训练时的注意力图，即图 4.11(b-d)，关注不同的物体部分以及最终的累积注意力图，即图 4.11(e)，可以发现更完整的物体区域。在图 4.12中，在训练过程中，一些图像的注意图定位在相似的物体区域，导致最后的累积注意图只获得小的物体区域。在注意力遮挡层的帮助下，注意力移动的范围变得更大，有利于累积注意力图挖掘完整的物体区域。

表 4.14 在 CUB-200-2011 测试集上不同方法的物体定位准确率的比较。这些方法都是基于 VGG-16 分类网络[2] 生成的注意力图。

方法	*loc1* (%)	*loc5* (%)
CAM[12]	41.46	51.36
ACoL[13]	45.92	56.51
SPG[50]	46.64	57.72
ADL[111]	52.36	-
OAA	**56.23**	**69.68**

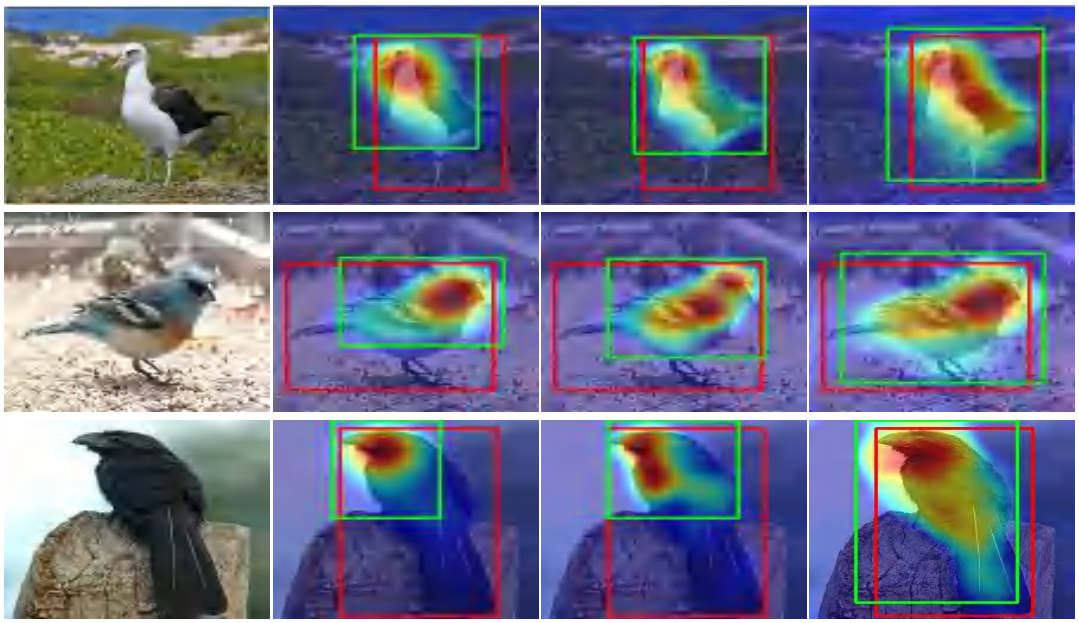


图 4.14 在 CUB-200-2011 数据集上不同方法定位结果的视觉比较。真实的边界框为红色，预测的边界框为绿色。

4.3.5 弱监督物体定位

本章的物体定位算法是专门为弱监督语义分割任务设计的。在本节，本文将在线注意力累积算法应用于弱监督物体定位任务，验证该方法的泛化能力。弱监督物体定位的目标是为目标物体生成一个紧密的边界框。大多数弱监督物体定位方法[12, 13, 50, 111]都是利用注意力图来定位目标物体。由于 OAA 被用于训练过程，无法为测试图像生成注意力图，本文稍微调整了在线注意力累积方法。本文将 OAA 和混合损失结合到一个网络中，来为测试图像生成注意力图。具体来说，本文利用如图 4.13所示的网络。这个网络含有两个分支，一个分支 ($F1$) 用来在训练期间积累注意力图。另一个分支 ($F2$) 接受累积注意力图的监督。在推理时，本文使用 $F2$ 分支来为测试图像生成注意力图。

本文使用和弱监督语义分割任务中相同的网络结构，即 VGG-16 分类网络。定位实验在 CUB-200-2011 数据集[103] 上进行，其有 200 个种类，包括 5,994 张训练图像和 5,794 张测试图像。本文依照[12] 的方式来生成目标物体的边界框，利用 *loc1* 和 *loc5* 来评估定位性能。如表 4.14所示，本文的 OAA 表现优于基线 CAM[12] 很多。与最先进的定位方法 ADL[111] 相比，OAA 取得了更高的定位精度，证明了 OAA 对于弱监督物体定位的有效性。如图 4.14所示，本文给出了不

同方法的定位效果图。

4.3.6 讨论

本章提出的基于在线注意力累积的物体定位方法旨在定位完整的物体区域，是专门为弱监督语义分割任务设计的方案。在弱监督物体定位任务上的实验是用来证明该方法具有一定的泛化性。此外，该物体定位方法主要解决的场景是包含具体主体的自然图像场景。对于不包含主体物体的图像，例如风景图像，本文的方法并不适用。

第四节 本章小结

在本章中，本文探索了一个基于在线注意力累积的物体定位方法，以发现更加完整的目标物体区域。在线注意力累积策略为每张图像中的每个类别维护一个累积注意力图，用于在训练阶段将分类网络生成的注意力图中不同的判别性区域保存下来。为了增大注意力在物体区域上的移动范围，本文还将一个注意力遮挡层嵌入到在线注意力累积策略中去。另外，本文还利用累积注意力图作为软标签以训练完整注意力模型，进一步增强累积注意力图。本文的方法非常简单且可以很容易的嵌入到任何分类网络的训练过程中，以完整地发现目标物体区域。大量的实验表明当弱监督语义分割等任务应用本文的注意力图时，本文方法的性能比以往最先进方法更好。

第五章 基于层次化注意力分解的归因算法研究

第一节 引言

5.1.1 背景知识

卷积神经网络在各种计算机视觉任务上取得了显著的提升，例如图像识别[1–3]、物体检测[4, 5, 145]、语义分割[7–9, 124]、交通环境分析[146, 147] 以及医学图像理解[148, 149]。虽然卷积神经网络具有很高的性能，但其内部决策过程不透明。此外，最近大量的研究[150–152] 指出，以前成功的卷积神经网络模型仍然可能被对抗性示例所愚弄，对抗性示例相对于原始图像的变化甚至无法被人眼辨别。在此前提下，人们很难相信表现良好但不透明的卷积神经网络模型。因此，卷积神经网络的可解释性和性能一样重要，特别是在一些关键的应用中。

一个完全可解释的卷积神经网络是深度学习研究人员长期以来梦寐以求的目标。为此，研究人员提出了各种各样的技术。特征归因方法[20, 21, 35] 为网络的可解释性提供了强大的工具。他们将卷积神经网络的预测归因到输入图像，其中生成的注意力图可以告诉用户哪些像素对预测很重要。这种能力有助于人类理解输入如何影响预测。这类特征归因方法的网络归因能力有限，因为它们无法归因网络内部特征对于决策的作用。即使一些特征归因方法[22, 23] 可以测量中间特征对预测的重要性。但是它们忽略了中间特征[85] 之间的关系，这种关系对于理解预测同样很重要，但很少受到关注。

5.1.2 解决方案的动机

卷积神经网络可以逐渐抽象图像内容并在不同语义级别生成特征，例如边缘、纹理和对象部分[37]。虽然以上归因算法发现的重要特征可以为预测提供丰富的证据，但孤立的证据远不如证据链[153] 或证据金字塔[154] 具有说服力。根据 Treisman 等人[11] 提出的特征集成理论，人脑首先提取基本特征，然后利用注意力组合各个特征来感知对象。理想情况下，本文希望得到如图 5.1所示的层次化证据，它将卷积神经网络决策归因于多个关键特征，并且每个关键特征都可以递归地归因于更基本的特征。在这个例子中，通过联系“头”、“脸”、“眼

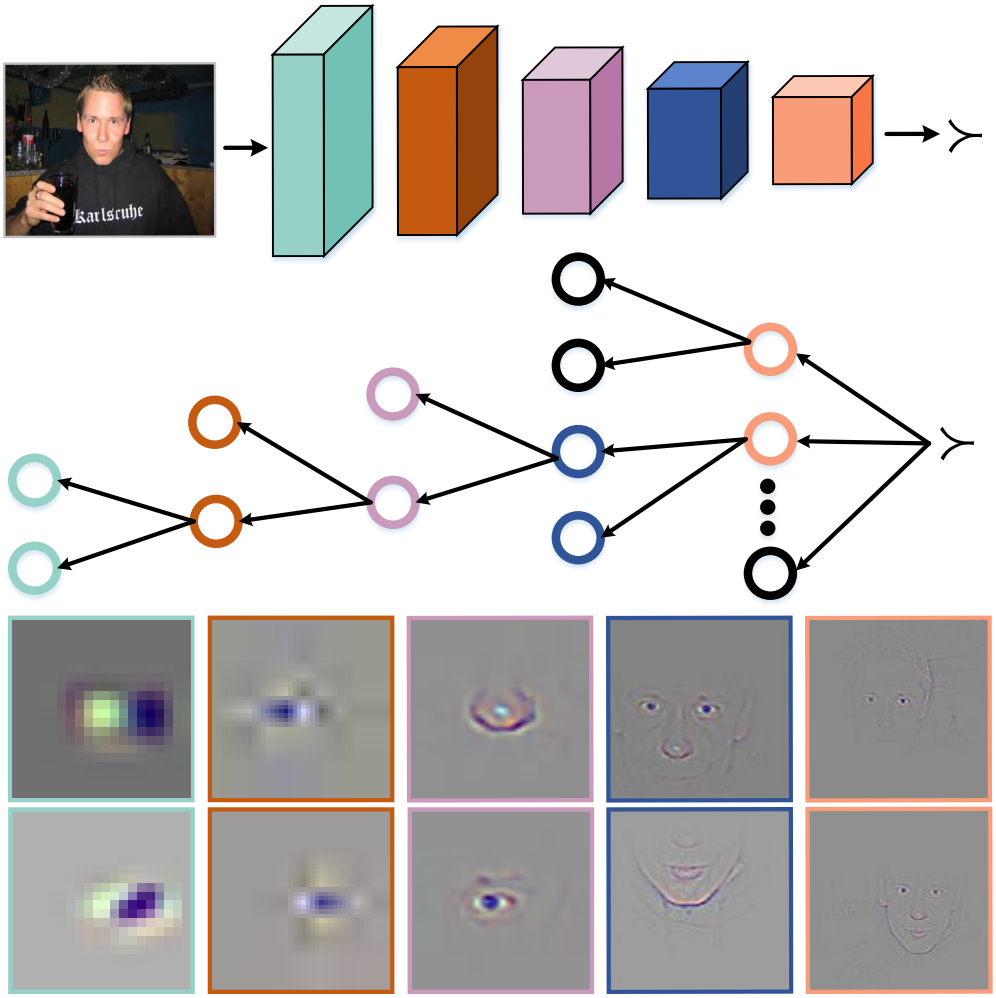


图 5.1 本文的方法示意图。中间是用于解释网络预测的证据金字塔，底部显示了来自 VGG-16 不同阶段的重要特征。中间圆圈的颜色和底部特征边框的颜色是对应的。本文检测特征之间的相互作用，并展示它们如何在决策过程的不同层次结构中组合。

睛”、“鼻子”和“边缘”等中间特征，一组与网络内部状态相对应的强有力的层次化证据正在出现。

第二节 层次化注意力分解

5.2.1 解决方案概述

现有的特征归因方法要对卷积神经网络的决策实现层次化分解主要面临以下几个挑战。首先，直接分解所有特征通道以及每个特征通道中的数百万个特征响应在计算上是不可行的，并且对人类来说是认知过载。同时，一些特征归因方法[22, 23] 非常耗时，因为它们需要多次重复反向传播过程。此外，另一些归

因方法[12, 15] 为整个层生成一个注意力图，而不是为每个特征通道都生成一个注意力图。通道注意力图对于迭代分解过程至关重要，因为它们可以指示要分解的特征通道中最重要的神经元。为了缓解这些问题，本文提出了一种高效的基于梯度的激活传播（gAP）模块，该模块可以将网络不同层的特征响应分解到其较低层。由于激活传播模块为每个特征通道生成一个注意力图，本文可以在每层的分解中选择几个最重要的特征通道作为关键证据，获得人类认知规模下的解释。对于每个选定的特征通道，可以迭代分解最活跃的空间位置对应的网络特征。通过避免在过多空间位置分解特征，可以进一步将潜在可视化的数量减少到人类认知尺度上。本文所提出的分解框架可以有效地生成层次化解释（见图 5.1），从而在关键的特征通道之间建立关系。本文在几个方面进行了广泛的实验，包括对激活回传模块的合理性检验以及用层次化分解诊断网络决策。实验证明了本文提出的框架在解释网络决策方面的有效性。

本文方法的大致流程如下:

- 首先，本文提出了一个高效的基于梯度的激活传播（gAP）模块，它可以分解了网络决策和任何中间特征，以便从前一层中找到支撑它们的关键支持证据。
- 本文利用 gAP 模块构建了一个层次化注意力分解框架，它在重要的特征通道之间建立联系，使层次化解释可以达到人类认知规模。

5.2.2 基于梯度的激活传播模块

本文首先定义卷积神经网络的符号。如图 5.2所示，在卷积神经网络的第 l 层中，**特征** $\mathbf{F}^l$、**梯度** $\mathbf{G}^l$ 和相应的**激活** $\mathbf{A}^l$ 是具有相同大小的 3D 张量，即 $\mathbf{G}^l, \mathbf{A}^l, \mathbf{F}^l \in \mathbb{R}^{K^l \times H^l \times W^l}$，其中 K^l 是通道数，$H^l \times W^l$ 是第 l 层输出的空间大小。为了找到网络决策或任何中间特征响应的支持证据，本文提出了一种基于梯度的激活传播 (gAP) 方法。

如图 5.2所示，在卷积层 $l+1$、通道 k' 和空间位置 (x, y) 处分解一个 CNN 特征 $\mathbf{F}^{l+1}_{k',x,y}$（感兴趣的决策），以在其先前的卷积层 l 中找到支持证据, 其中 $\mathbf{F}^{l+1}_{k',x,y}$>0。在典型的卷积神经网络架构中，例如 VGG-16[2]，第 $l+1$ 层的具有正激活的特征是由第 l 层经过 ReLU 后的特征线性组合得到的。对于 $\mathbf{F}^{l+1}_{k',x,y}$，有以下公式

$$\mathbf{F}^{l+1}_{k',x,y} = \mathrm{ReLU}(\mathbf{w^1} \cdot \mathbf{F^1}) = \mathbf{w^1} \cdot \mathbf{F^1}, \tag{5.1}$$

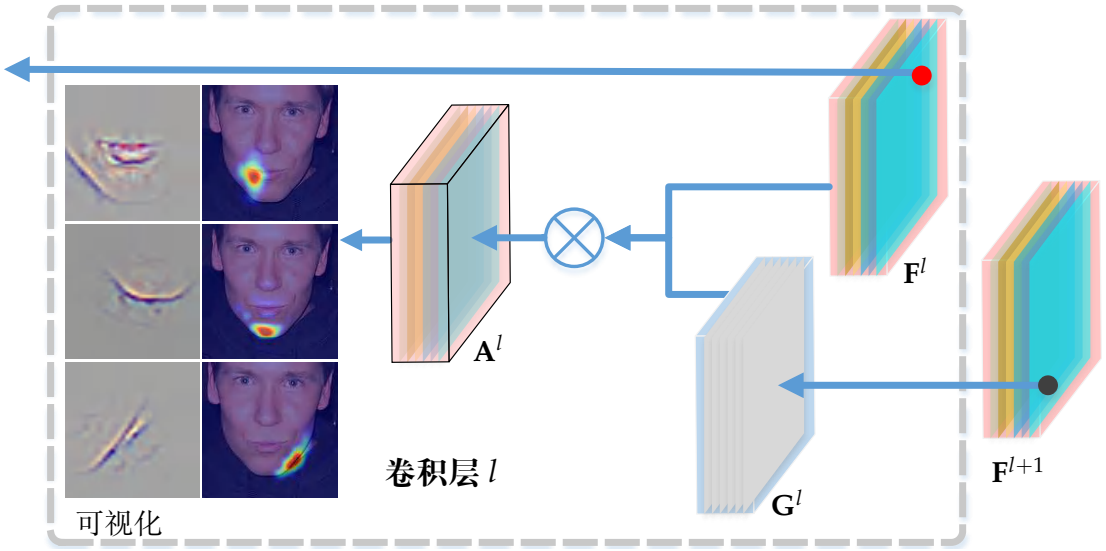


图 5.2 本文的基于梯度的激活传播 (gAP) 方法解释了一个感兴趣的决策 $\mathbf{F}^{l+1}_{k'}(x^{l+1}_{k'}, y^{l+1}_{k'}) \in \mathbb{R}$，即黑点所示的网络特征，通过在其浅层中定位最相关的特征通道。

其中 $\mathbf{w}^{\mathbf{l}}_{\mathbf{k}}$ 是 $\mathbf{F}^l$ 是第 k 个特征通道的线性权重。由于每个特征通道的线性权重和其梯度是等价的，因此为了获取权重 $\mathbf{w}^{\mathbf{l}}_{\mathbf{k}}$，本文使用反向传播来计算特征 $\mathbf{F}^{l+1}_{k',x,y}$ 关于特征通道 $\mathbf{F}^l_k$ 的梯度 $\mathbf{G}^l_k$，公式如下：

$$\mathbf{w}^{\mathbf{l}}_{\mathbf{k}} = \mathbf{G}^{\mathbf{l}}_{\mathbf{k}} = \underbrace{\frac{\partial \mathbf{F}^{l+1}_{k',x,y}}{\partial \mathbf{F}^l_k}}_{\text{通过回传得到的梯度}}. \tag{5.2}$$

梯度 $\mathbf{G}^l_k$ 捕获了特征通道 $\mathbf{F}^l_k$ 对决策 $\mathbf{F}^{l+1}_{k',x,y}$ 的“重要性”。

本文使用梯度 $\mathbf{G}^l_k$ 来为特征通道 $\mathbf{F}^l_k$ 生成注意力图 $\mathbf{A}^l_k$，即

$$\mathbf{A}^l_k = \mathbf{G}^l_k \cdot \mathbf{F}^l_k. \tag{5.3}$$

注意力图表示 $\mathbf{F}^l_k$ 中的每个特征对决策 $\mathbf{F}^{l+1}_{k',x,y}$ 的贡献。基于其对应的注意力图，每个特征通道对决策的贡献可以通过下式计算

$$\alpha^l_k = \frac{1}{Z^l} \sum_x^{H^l} \sum_y^{W^l} \mathbf{A}^l_{k,x,y'} \tag{5.4}$$

其中 $Z^l = H^l \times W^l$ 表示注意力图 $\mathbf{A}^l_k$ 中的空间大小。本文还可以在第 k 个特征通道中通过以下公式识别出对决策贡献最大的特征 $\mathbf{F}^l_{k,\hat{x},\hat{y}}$，

$$(\hat{x}, \hat{y}) = \underset{(x,y)}{\arg\max}\, \mathbf{A}^l_{k,x,y}. \tag{5.5}$$

因此，对于每个决策，可以根据公式 (5.4) 计算的贡献 α_k^l 找到最重要的特征通道 $\mathbf{F}_k^l$。在最重要的特征通道中，还可以根据公式 (5.5) 识别出对决策贡献最大的特征 $\mathbf{F}_{k,\hat{x},\hat{y}}^l$。在图 5.3的第一行中，本文展示了在 `conv4_3` 层中三个最重要的注意力图 $\mathbf{A}_{131}^4, \mathbf{A}_{255}^4, \mathbf{A}_{452}^4$，用于支撑 $\mathbf{F}_{277}^5$ 的决策。这些注意力图提供了空间通道响应，有利于人类理解。此外，本文通过 Guided Backpropagation[36] 生成清晰的可视化图来展示最有贡献的特征。图 5.3的底行显示了一个示例。

与 CAM 和 Grad-CAM 的区别： 本文提出的基于梯度的激活回传（gAP）模块受到 CAM[12] 和 Grad-CAM[15] 的启发，它们通过注意力图来解释卷积神经网络的决策。为了解释与 gAP 的关系和差异，本文首先重新回顾 CAM 和 Grad-CAM。Selvaraju 等人[15] 已经证明 Grad-CAM 是 CAM 的严格推广。不失一般性，本文使用和 CAM[12] 中相同的分类网络。对于一个图像分类网络，它使用全局平均池化层对最后一个卷积层的特征 $\mathbf{F}^L(\in \mathbb{R}^{K^L \times H^L \times W^L})$ 进行空间池化以获得特征向量。随后网络将特征向量输入到全连接层来执行特征向量的**线性组合**。让 C 表示类别的数量。在 softmax 函数之前，每个类 $c \in \{1,2,\ldots,C\}$ 的分类分数 S^c 可以通过以下公式计算得到

$$\begin{aligned} S^c &= \sum_k^{K^l} w_k^c \overbrace{\frac{1}{Z^L}\sum_x^{H^L}\sum_y^{W^L}}^{\text{全局平均池化}} \mathbf{F}_{k,x,y}^L \\ &= \frac{1}{Z^L}\sum_x^{H^L}\sum_y^{W^L}\sum_k^{K^L} w_k^c \mathbf{F}_{k,x,y}^L, \end{aligned} \tag{5.6}$$

其中 $Z^l = H^L \times W^L$，(x,y) 表示特征图中的空间位置，w_k^c 是连接第 k 个特征图和第 c 个类的权重。特征 $\mathbf{F}_{k,x,y}^L$ 对 S^c 的贡献为 $w_k^c \mathbf{F}_{k,x,y}^L$。CAM 通过对所有特征图加权求和来生成注意力图 M^c，

$$\mathbf{M}^c = \sum_k^{K^L} w_k^c \mathbf{F}_k^L, \tag{5.7}$$

其中 M^c 中的每个值表示每个空间位置对 $\mathbf{M}^c$ 的贡献。

对于线性函数，权重也等于梯度。因此，本文也可以通过计算反向传播梯度来获得权重，

$$w_k^c = \sum_x^{H^L}\sum_y^{W^L} \frac{\partial S^c}{\partial \mathbf{F}_{k,x,y}^L}. \tag{5.8}$$

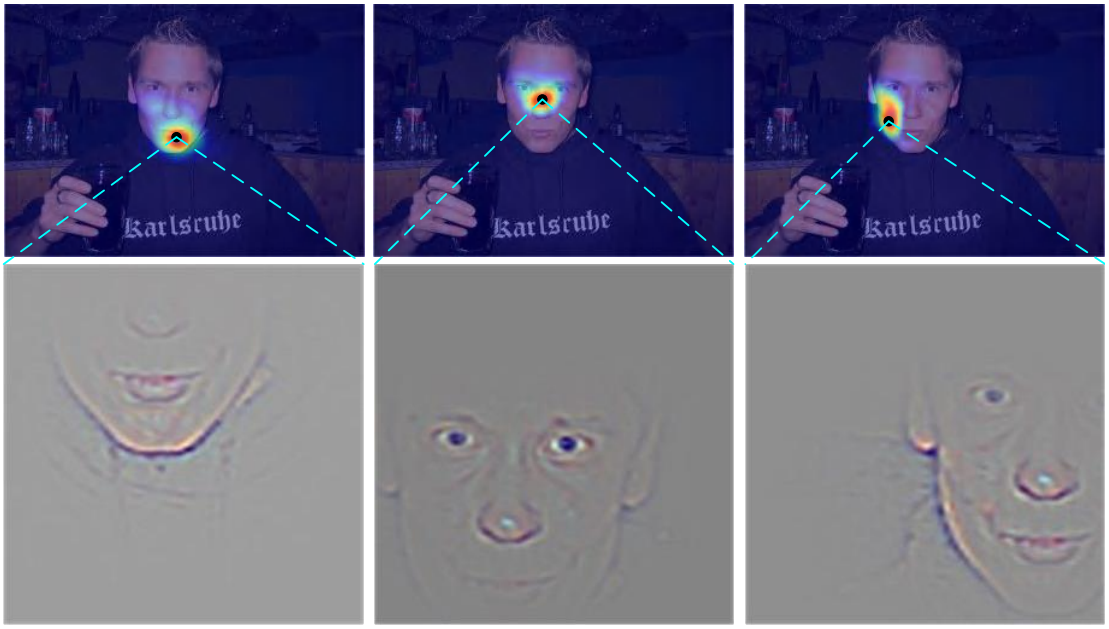


图 5.3 在 conv4_3 层中最重要的注意力图（上一行）及其相应的可视化（下一行）示例。黑点表示注意力图中的峰值位置。

w_k^c 的详细推导过程可以从 Grad-CAM[15] 中找到。公式 (5.8) 也是 Grad-CAM 计算权重 w_k^c 的方式。稍有不同的是 Grad-CAM 将 w_k^c 乘以比例常数，即

$$w_k^c = \frac{1}{Z^L} \sum_x^{H^L} \sum_y^{W^L} \frac{\partial S^c}{\partial \mathbf{F}_{k,x,y}^L}, \tag{5.9}$$

将比例常数 $1/Z^L$ 归一化。

让我们将分数 $\{S^c\}$ 看作具有 C 个通道和空间大小 1×1 的特征，即 $\mathbf{F}_c^{L+1}(1,1) = S^c$，可以带入公式 (5.2) 到 公式 (5.9) 得到以下公式

$$w_k^c = \frac{1}{Z^L} \sum_x^{H^L} \sum_y^{W^L} \mathbf{G}_{k,x,y}^L. \tag{5.10}$$

由于全局平均池化层，$\mathbf{F}_k^L$ 中每个特征的梯度相同，即 $\forall_{x,y}, \mathbf{G}_{k,x,y}^L = w_k^c$。注意力图 $\mathbf{M}^c$ 可以写成

$$\mathbf{M}^c = \sum_k^{K^L} \mathbf{G}_k^L \cdot \mathbf{F}_k^L = \sum_k^{K^L} \mathbf{A}_k^L. \tag{5.11}$$

公式 (5.11) 表明 Grad-CAM 生成的注意力图 $\mathbf{M}^c$ 可以通过简单地将 gAP 为每个通道生成的注意力图 $\mathbf{A}_k^L$ 相加来生成。

gAP 和 Grad-CAM/CAM 之间的区别总结是：

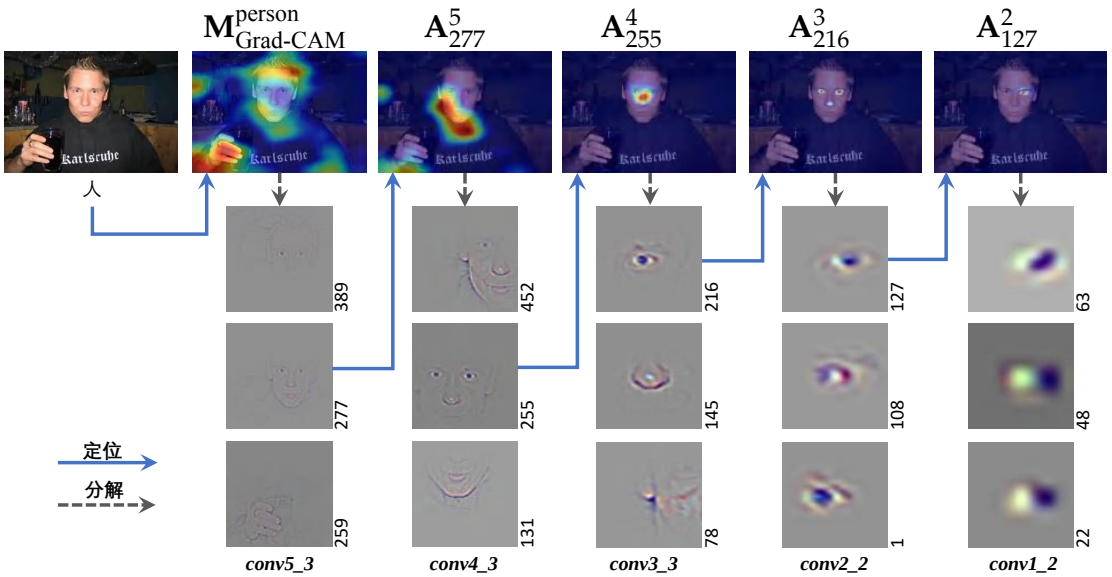


图 5.4 本文的层次化分解过程的图解。每个可视化的数字表示 VGG-16[2] 的通道 id。在每个阶段，本文将前 3 个最重要的特征通道之一分解到较低层。沿着蓝线，放大决策的注意力图。灰色虚线表示对应于最大激活的特征响应的分解。此外，本文还使分解过程具有交互性。在每个阶段，用户可以选择任何决策并将其分解。

- Grad-CAM/CAM 结合所有通道注意力图以生成单个类注意力图 $\mathbf{M}^c$，它突出显示支持预测的重要区域。本文的 gAP 模块通过生成一组注意力图 $\{\mathbf{A}_k\}$ 来解释感兴趣的决策。每个注意力图对应一个特征通道，这对于迭代分解过程至关重要。
- Grad-CAM/CAM 从最后一个卷积层生成类注意力图来解释预测。本文的 gAP 推广了这个想法，并迭代地将任何层的决策分解到其较低层。

虽然上述推导适用于相邻层，但根据经验发现，在卷积神经网络不同阶段的两层之间应用 gAP 模块时，也可以获得令人满意的分解结果（参见第 5.3.1 节）。在下文中，本文将描述如何为网络决策构建层次化解释。

5.2.3 层次化注意力分解

在图 5.4中，本文展示了层次化注意力分解过程的一个示例。首先，本文将网络决策分解到最后一个卷积层，并找到前几个最重要的支持特征。随后，本文继续将每个支持特征分解到它们的前一层，并迭代地重复分解过程直到底层。如第一节所述，构建层次化注意力分解的关键挑战在于分解过程中会产生大量的可视化结果，这将成为人类的认知负担。即使只分解每个通道中的最大贡献特征（另见 公式 (5.5)），直接分解 VGG-16 中的所有通道也会生成

$512^3 \times 512^3 \times 256^3 \times 128^2 \times 64^2 \approx 2.5 \times 10^{22}$ 个证据。

为了获得人类认知规模下的层次化证据，本文提出了两种减少可视化数量的策略。首先，本文只分解每一层最重要的几个特征。实验（参见第 5.3.1 节）已经证实，一个层中的一小部分特征通道占了对决策的大部分贡献。因此，本文选择前几个最重要的通道。此外，本文通过利用每个阶段的最后一个卷积层来简化自上而下的决策分解过程。当前流行的卷积神经网络[1, 2] 通常在每个阶段之后减小特征图的空间大小，其中一个阶段是由一组具有相同输出分辨率的卷积层组成。每个阶段学习不同的模式，例如边缘、纹理和对象部分[36, 37]。实验验证，当在两个连续阶段的两层之间使用 gAP 模块时，本文仍然可以获得视觉上有意义的分解结果（见图 5.4）。通过这两种策略，可以大大减少可视化的数量以获得人类认知规模的层次化解释。

如图 5.4所示, 本文展示了 VGG-16 分类网络的一个例子。本文选择 `conv1_2`、`conv2_2`、`conv3_3`、`conv4_3` 和 `conv5_3` 并将这些层索引为 $\{1,2,\ldots,L\}$, 其中 $L=5$。Softmax 之前的网络输出可视为第 6 层的特征，表示为 $\mathbf{F}^6 \in \mathbb{R}^{C\times 1\times 1}$。分解过程从网络决策 $\mathbf{F}^6_c(1,1)$ 开始，其中 c 对应于“人”这个类别。使用 gAP 模块，本文首先将决策 $\mathbf{F}^6_c(1,1)$ 分解到第 5 层。该分解在第 5 层为 $\mathbf{F}^6_c(1,1)$ 生成一组注意力图 $\{\mathbf{A}^L\}$。本文使用公式 (5.4) 选择前 N 个（例如，N=3）重要的注意力图，即 $\mathbf{A}^5_{389}$、$\mathbf{A}^5_{277}$ 和 $\mathbf{A}^5_{259}$。本文继续分解来自 $\mathbf{F}^5_{389}$、$\mathbf{F}^5_{277}$ 和 $\mathbf{F}^5_{259}$ 的决策，并分别找到它们在第 4 层的前 N 个最重要的注意力图。但是，直接分解特征通道并不容易。因为并非特征通道中的所有特征都有助于决策（参见图 5.3顶行中的注意力图）。本文选择对决策贡献最大的最具代表性的特征并分解该特征。本文利用公式 (5.5) 找到最大激活对应的特征 $\mathbf{F}^l_{k,\hat{x},\hat{y}}$。然后本文使用 gAP 模块将其分解到第 $l-1$ 层。这种层次化分解过程递归运行，知道网络决策被分解到最底层。

可视化的数量 N 是一个灵活的参数，它控制在每次分解期间选择重要特征通道的数量。为了便于人类认知，图 5.4中 N 设置为 3。此外，本文的使层次化注意力分解过程具有交互性，以便用户可以选择要分解的特征，轻松访问所需的信息。在图 5.4中，可以看到在高层检测到的特征可以分解为在低层检测到的不同部分。层次化分解过程跟踪重要特征并使用来自较低层的证据递归地解释证据。例如，“人”的分类结果被分解为“面部”和“手部”证据。然后“面部”证据进一步分解为“眼睛”、“鼻子”和“下颌”。这个过程一直持续到最底层，这些层通常检测边缘特征。

一些归因方法，例如 LRP[19]，以层次化方式将重要性分层传播到输入。它们生成单个注意力图，指示输入中每个像素的重要性。与它们不同，本文的方法将重要性传播链解耦，并产生丰富的层次化注意力图和相应的可视化。为了解释一个人的图像，本文的方法首先找到了一组证据，例如“面部”、“手”等的注意力图。每个证据都与它们自己的支持证据相关联，例如，“面部”有“眼睛”、“鼻子”等的支持注意力图。本文提出的方法提供特征通道以及特征通道之间关系。

第三节 实验

在本节中，本文首先进行实验来验证决策分解的正确性和效率。然后，利用层次化注意力分解分析网络特征，解释网络决策。本文在两个流行的数据集 ILSVRC[108] 和 PASCAL VOC[104] 上进行了实验。在 PASCAL VOC 数据集上，使用包含 10582 张训练图像的增强训练集对不同的分类网络进行微调。本文的层次化注意力分解方法基于 Pytorch 库[1] 实现。所有实验都是在单个 RTX 2080Ti GPU 上进行测试的。

5.3.1 gAP 实验

5.3.1.1 gAP 的有效性

在层次化注意力分解过程中，最重要的是 gAP 模块计算特征通道贡献的准确性。因此，本文首先检验特征通道对决策贡献的准确性。按照以前工作[22, 95, 155] 中的设置，本文将每次移除一个特征通道时的决策分数减少的值作为该通道贡献的真值。具体来说，给定一个输入图像 I，设 f^{l+1} 为第 $l+1$ 层的决策得分。$\hat{f}^{l+1}$ 表示将第 l 层的第 k 个特征通道设置为平均激活时的决策得分。$\hat{\alpha}_k^l = f^{l+1} - \hat{f}^{l+1}$ 表示第 k 特征通道对该决策的真值贡献。

皮尔逊相关系数 (Pearson Correlation Coefficient, PCC)[156] 用于测量真值贡献 $\hat{\alpha}^l \in \mathbb{R}^{K^l}$ 与公式 (5.4) 估计的贡献 $\alpha^l \in \mathbb{R}^{K^l}$ 之间的线性相关程度。当 PCC 值为 1 时，两个变量之间存在线性相关（0 表示不线性相关，-1 表示总线性负相关）。PCC 度量由下式计算

$$\rho = \frac{\mathbb{E}\left[(\alpha^l - \mu_{\alpha^l})(\hat{\alpha}^l - \mu_{\hat{\alpha}^l})\right]}{\sigma_{\alpha^l} \cdot \sigma_{\hat{\alpha}^l}}, \tag{5.12}$$

[1] https://pytorch.org/

表 5.1 不同设置下的皮尔逊相关系数（PCC）。→ 表示分解。**S5-S1** 表示 VGG-16[2] 中 5 个不同阶段的最后一个卷积层。**AA**：平均激活。**MA**：最大激活。**AG**：平均梯度。**MG**：最大梯度。**T**：目标类别。使用平均激活的设置，gAP 可以达到最佳分解效果。

ILSVRC	T → S5	S5 → S4	S4 → S3	S3 → S2	S2 → S1
AA	0.985	0.959	0.933	0.898	0.895
MA	0.897	0.912	0.894	0.864	0.890
AG	0.623	0.421	0.497	0.545	0.472
MG	0.454	0.456	0.567	0.594	0.606
VOC	**T → S5**	**S5 → S4**	**S4 → S3**	**S3 → S2**	**S2 → S1**
AA	0.987	0.961	0.932	0.899	0.893
MA	0.917	0.913	0.892	0.856	0.897
AG	0.702	0.492	0.525	0.564	0.480
MG	0.575	0.525	0.536	0.583	0.669

其中，μ 和 σ 分别表示均值和方差，$\mathbb{E}$ 表示期望值。

如表 5.1所示，本文研究了几种不同策略用于计算特征通道对决策的贡献。从表中可以看出，通过平均激活计算的贡献，即公式 (5.4)，获得了与真值之间最高的 PCC 值。对于 VGG-16 中的所有阶段，公式 (5.4) 计算的贡献与真值之间存在很强的线性相关性。这种高度相关性验证了 gAP 模块的有效性。一些工作[22, 95, 155] 依次计算每个特征通道的贡献，通过计算得分下降来衡量通道贡献是一种非常耗时的方式。相比之下，当 gAP 计算特征通道对决策的贡献时，只需要一次反向传播过程。在 VGG-16 骨干网络上，计算一幅图像的真值贡献需要大约 10 秒，而 gAP 模块只需要大约 50 毫秒，gAP 快了近 200 倍。利用 gAP 模块的效率优势，本文的层次化注意力分解可以高效为网络决策的生成层次化解释。

5.3.1.2 特征通道的贡献分布

如图 5.5的前五条曲线所示，可以观察到卷积层中通道贡献呈现长尾分布。少数特征通道对感兴趣的决定起着最重要的作用。在网络的更深层次，重要特征通道的比例降低。在高层中，特征通道通常更具判别力，这一事实符合公认的概念[37]。此外，本文还检验了卷积层中有多少特征通道对高层的决策有贡献。

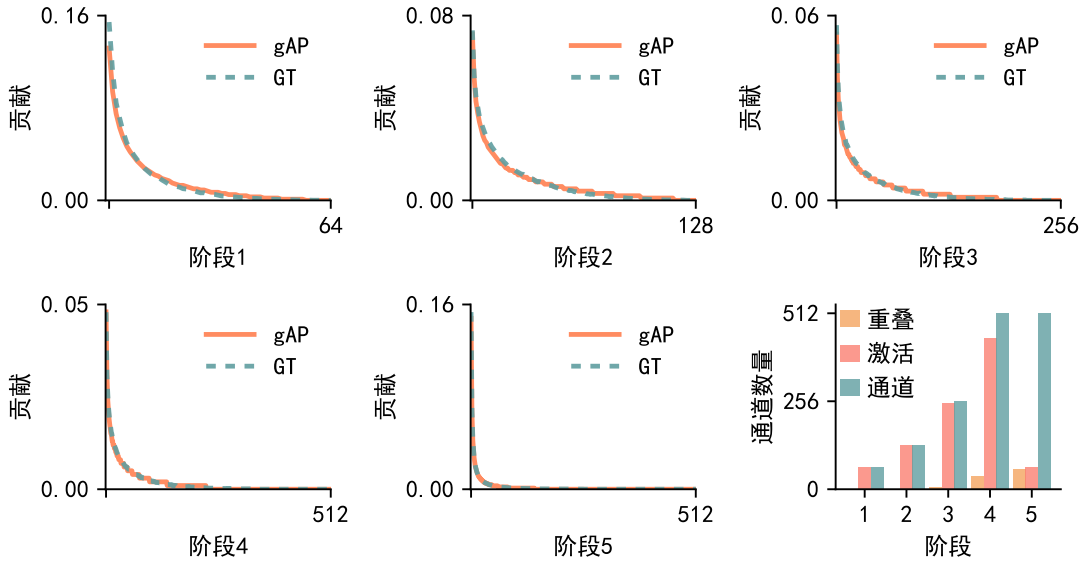


图 5.5 (a) 前五张图描绘了卷积层中每个通道对决策的贡献。通道对决策的贡献表示它对决策的影响程度。gAP 表示提出的基于梯度的激活传播方法。GT 表示去除特征通道的方法。“Stage1-5” 表示卷积神经网络每个阶段的最后一个卷积层。通道贡献按降序排列。gAP 计算的贡献分布与真值几乎保持一致。此外，卷积层中特征通道的贡献呈现长尾分布规律。(b) 最后一张图表绘制了卷积层中激活通道的数量，具有重叠作用的通道数量以及总通道的数量。在高层中，有许多激活的通道具有对决策相似的效果。

本文将 $\alpha_k^l > 0$ 的通道称为激活通道，并在决策分解过程中计算激活通道的数量。如图 5.5的最后一张图所示，当将决策从 `conv2_2` 层分解到 `conv1_2` 层时，发现 `conv1_2` 层中几乎所有的通道都被激活。但是，对于从最终决策到 `conv5_3` 层的分解，可以看到激活的通道数远小于 `conv5_3` 层中总通道的数量。

5.3.1.3 通道效应重叠

本文观察到由相同决策分解而来的一些通道的注意力图之间通常在相似的空间位置具有强激活。这样的空间位置通常表示一个潜在的概念[93, 94]，有助于决策。在展示层次化注意力分解的可视化时，本文将合并这些具有相似效果的重复通道，以便人类更好地理解。具体来说，当将感兴趣的决策分解到较低层时，本文将获得对应于该层中每个通道的注意力图。本文首先将所有通道的注意力图阈值化为二值图，然后计算它们之间的交并比 (IoU)。然后应用非极大值抑制算法[157] 来抑制 IoU 分数大于 0.9 的注意力图，其中注意力图使用公式 (5.4) 计算的贡献分数进行排序。如图 5.5 所示，本文展示了有多少激活的通道彼此之间具有较大的重叠。在底层中，具有较大重叠的激活通道的数量非常少。但是

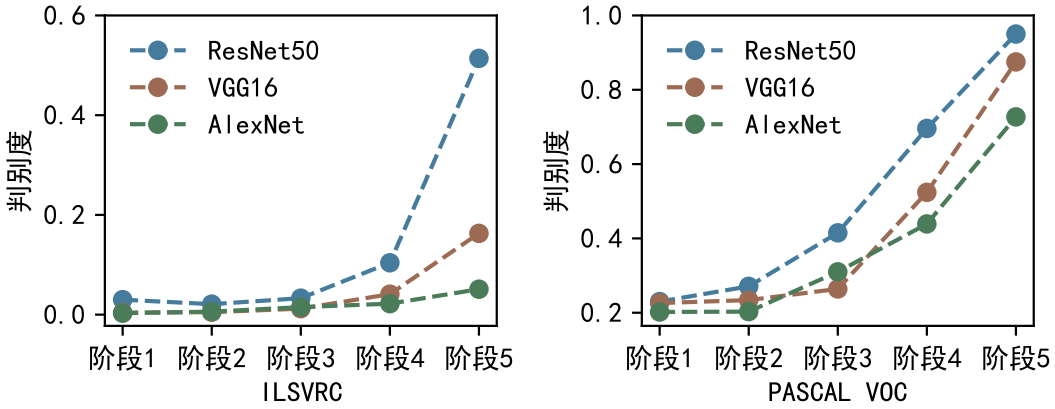


图 5.6 在 PASCAL VOC[104] 和 ILSVRC[108] 的验证集上，来自不同卷积神经网络的通道的判别度。

在高层中，有许多激活的通道具有对决策类似的效果。

5.3.1.4 通道判别性分析

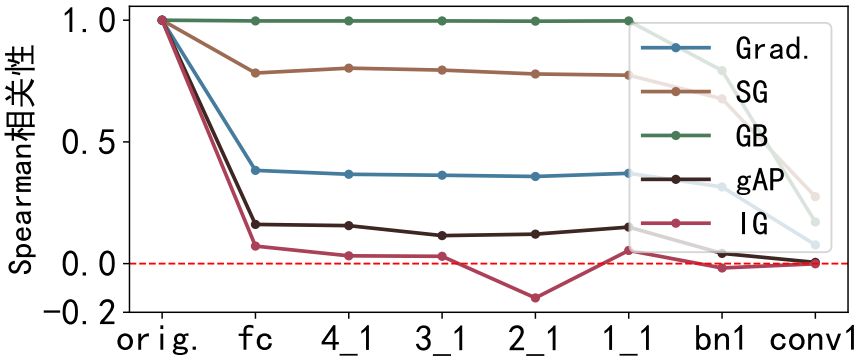
本文利用层次化注意力分解来探索不同层中特征通道的判别信息。具体来说，本文定义了一个判别度 D 来衡量一个特征通道的判别信息。在对带有标签 c 的图像执行层次化注意力分解过程时，本文计算在每个分解过程通道 k 对决策的贡献排名在前 3 时的次数，即 N_c。N_c 是对来自验证集的所有图像求和。然后判别度 D 由下式计算

$$D = \frac{\max_c N_c}{\sum_{c=1}^{C} N_c}, \tag{5.13}$$

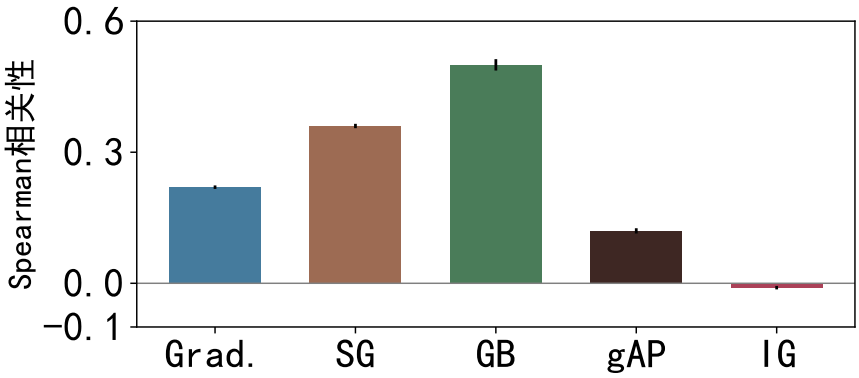
其中 C 表示数据集中的类别数。当通道 k 都是从一个单一类别分解来时，判别度 $D=1$。此外，当通道从每个类别等倍分解来时，本文可以获得 D 的最小值：$D=1/C$ 。

如图 5.6所示，浅层的通道判别度非常小。它们通常对多个类别有很强的激活作用。这一事实表明，浅层通道检测到的基本特征在不同类别之间共享，缺乏用于分类的判别信息。然而，在网络的高层中，通道判别度比低层大很多。因为卷积神经网络的高层逐渐结合低层的基本特征，形成更具判别力的特征。在高层中，不同的类别倾向于突出自己的判别性通道。这些结果为 Zeiler 等人[37]的结论提供了额外的证据。

本文将层次化注意力分解应用于不同的卷积神经网络。对于不同卷积神经网络的高层，通道的判别度随着网络深度的增长而逐渐增加（ResNet-50[1]



(a) 模型随机化试验



(b) 数据随机化试验

图 5.7 使用级联模型参数和数据随机化测试对不同归因方法进行健全性检验。SG：SmoothGrad[44]，GB：Guided Backprop[36]，IG：Integrated Gradient[20]。Spearman 相关性指标[159] 用于测量原始模型的归因图和随机化模型的归因图之间的相关性。低相关性意味着归因方法对模型参数和数据标签敏感，适合解释模型决策。本文的 gAP 在这两个测试中获得了低相关值。

>VGG-16[2] > AlexNet[158]）。这种差异表明 ResNet-50 的高层具有更强的类判别能力。通道强大的判别能力可以有效减少不同类别之间的混淆，这有助于 ResNet-50 实现比 VGG-16 和 AlexNet 更高的分类准确率。

5.3.1.5 gAP 的健全性检验

Adebayo 等人[160] 提出了用于视觉归因方法健全性检验的模型参数和数据随机化测试。这两个测试用于检验归因方法是否对模型参数和数据标签敏感。对模型参数和数据标签不敏感的归因方法不足以调试模型和解释依赖于实例和数据标签之间关系的机制。为了从 gAP 生成归因图，本文的层次化注意力分解策略将决策分解直到输入层，并对每个分解的梯度求和。本文对预训练的

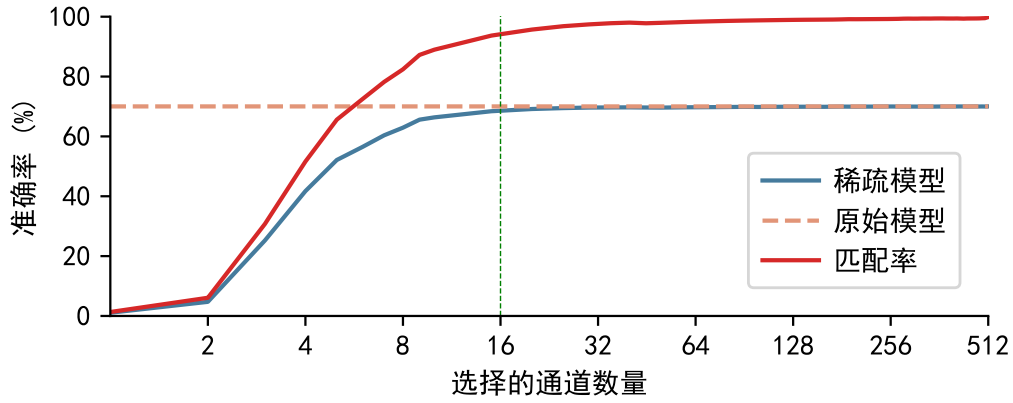


图 5.8 gAP 生成的稀疏模型与原始模型的分类精度的比较。匹配率表示稀疏模型与原始模型之间的预测一致性。

ResNet-18 模型[1] 进行模型参数随机化测试，并以级联方式从顶层到底层随机初始化模型参数。本文利用 spearman 秩相关度量来计算来自原始模型和随机初始化模型的归因图之间的差异。此外，本文通过比较使用真实标签和扰乱标签训练的模型的归因图来进行数据随机化测试。

在图 5.7(a) 中，低 spearman 相关性表明原始模型和随机初始化模型的归因图有很大不同，这表明 gAP 对模型参数很敏感。在图 5.7(b) 中，低 spearman 相关性也表明 gAP 对数据的标签很敏感。实验结果验证了本文的方法可以用于调试模型。

5.3.1.6 top-k 分解是对原始模型的良好近似吗？

当应用层次化注意力分解时，被分解的卷积层会保留最重要的几个特征通道，这样就可以得到代替原始模型的稀疏代理模型。本文测试了稀疏模型的分类精度。此外，本文通过比较稀疏模型和原始模型之间的预测来衡量匹配率。具体来说，本文从预测类别的决策分解到底层，并在每个分解中选择前 k 个重要特征进行预测。如图 5.8所示，当使用 top-16 分解时，稀疏代理模型具有与原始模型相似的分类精度。根据匹配率，当使用 top-16 分解时，稀疏代理模型和原始模型的预测几乎在所有样本上都是一致的。选择少量特征通道的稀疏代理模型可以很好地逼近原始模型。

5.3.1.7 与基于个体重要性的方法进行比较

基于个体重要性的方法[22, 23] 计算来自不同层的每个特征通道对最终网络决策的重要性。与基于个体重要性的方法相比，gAP 可以帮助读者探索不同特征

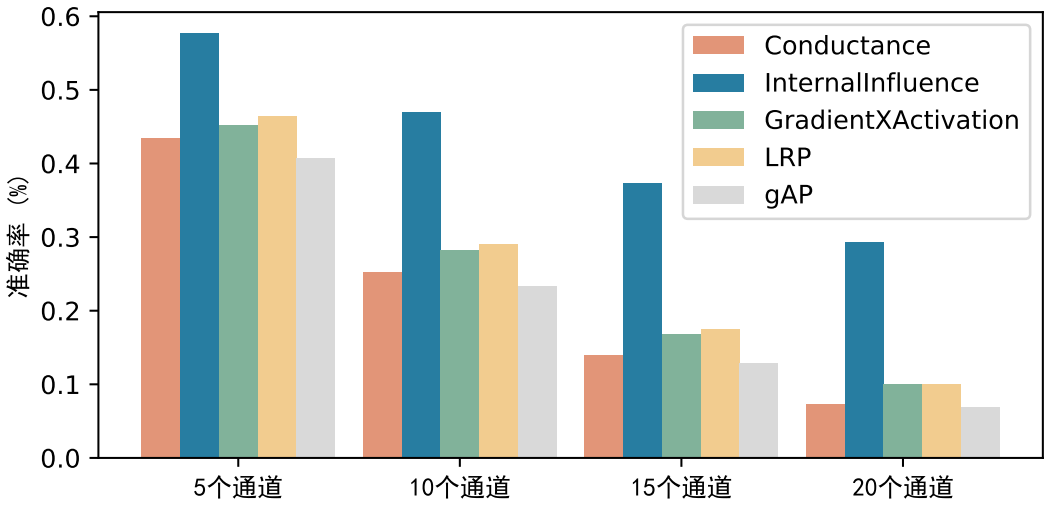


图 5.9 去除前几个最重要的特征通道后的分类精度比较（越低越好）。y 轴是 ILSVRC 验证集[108] 上的分类准确度。x 轴表示要去除的重要特征通道的数量。Conductance[22]，InternalInfluence[23]，LRP[19]。

通道之间的关系。为了直接与它们进行比较，本文将顶层的每个选定通道的重要性传播到浅层。本文从 VGG-16 的不同层中选择前 N 个最重要的通道并消融它们以观察分类精度的变化。本文在 ILSVRC 验证集[108] 进行了实验。如图 5.9 所示，当去除前几个最重要的特征通道时，gAP 获得的分类精度低于其他基于个体重要性的方法。本文分析 gAP 只将那些重要特征通道的贡献传播到较低层，从而减少了其他特征通道的干扰。与基于个体重要性的方法相比，gAP 不仅可以有效地检测重要特征，还可以检测这些特征如何相互影响。

5.3.2 卷积神经网络归因实验

在本节，本文首先对失败案例和对抗样本进行归因研究，随后分析卷积神经网络识别不同类别时使用的上下文信息。

5.3.2.1 失败案例归因

以前的工作[15] 可以为网络预测生成类注意力图，突出支持网络决策的最重要的图像区域。然而，这样的解释并不足以提供足够的信息。层次化注意力分解可以进一步为网络决策提供层次化解释。本文将网络的决策迭代分解到底层，并在不同层找到最重要的特征通道。可以计算出每个通道对网络决策的贡献。此外，还可以研究重要通道及其相应的注意力图。

如图 5.10所示，本文使用层次化分解来检验网络的错误决策。图 5.10展示

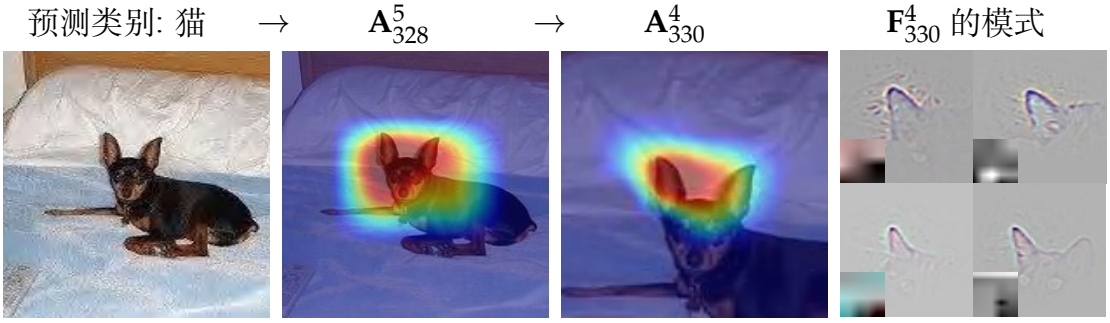


图 5.10 失败案例归因。最左边是被错误分类为 猫类别的狗图像。本文将网络决策分解到 conv4_3 层。最右边是最大激活第 330 个通道的模式。在这个例子中，对猫类别属性敏感的通道具有很强的激活，导致 VGG-16 做出错误的决定。

了一个失败案例。一张狗图像以 99% 的概率被错误分为猫类别。本文首先将网络决策分解到 conv5_3 层，并找到最重要的通道，即第 328 个通道，贡献率为 32.3%。本文进一步展示了从第 328 个通道到 conv4_3 层的分解，并找到了最重要的通道，即第 330 个通道。注意力图 $\mathbf{A}^4_{330}$ 在耳朵区域具有强激活。此外，最大激活第 330 个通道的模式是猫耳朵图像块。本文发现在这个例子中狗的耳朵与猫的耳朵图像块具有相似的形状。本文进一步遮挡了狗耳朵的图像区域，观察到网络能够以 65% 概率正确预测狗类别。在这个例子中，通过层次化分解，本文发现网络做出错误预测是因为它把狗的耳朵识别成猫的耳朵。

此外，在图 5.11中，本文提供了失败的案例，并用层次化注意力分解方法对卷积神经网络预测进行归因。一般来说，本文总结失败案例可分为两类。第一类是目标类别和错误预测的类别之间在图像中具有相似的特征，这使得卷积神经网络对其进行了错误分类。如图 5.11中第一行所示，磁带播放器和火炉都有圆状物体，卷积神经网络将火炉的圆状物体识别成了此外播放器。第二类是图像同时包含目标标签和预测标签的特征，其中 CNN 更关注预测标签的特征。如图 5.11中第四行所示，公牛和森林狼在存在在图里，卷积神经网络更加关注公牛的特征，使得最终公牛类别的预测分数更高。这种案例可以被多分类卷积神经网络解决。

5.3.2.2 对抗性攻击归因

当前的卷积神经网络的模型容易受到对抗性攻击。当对抗性攻击算法对原始图像添加一个小的扰动时，这些模型很容易对扰动后的图像进行错误分类。为了了解对抗性图像如何成功地欺骗卷积神经网络的模型，按照 Bau 等人[95] 的

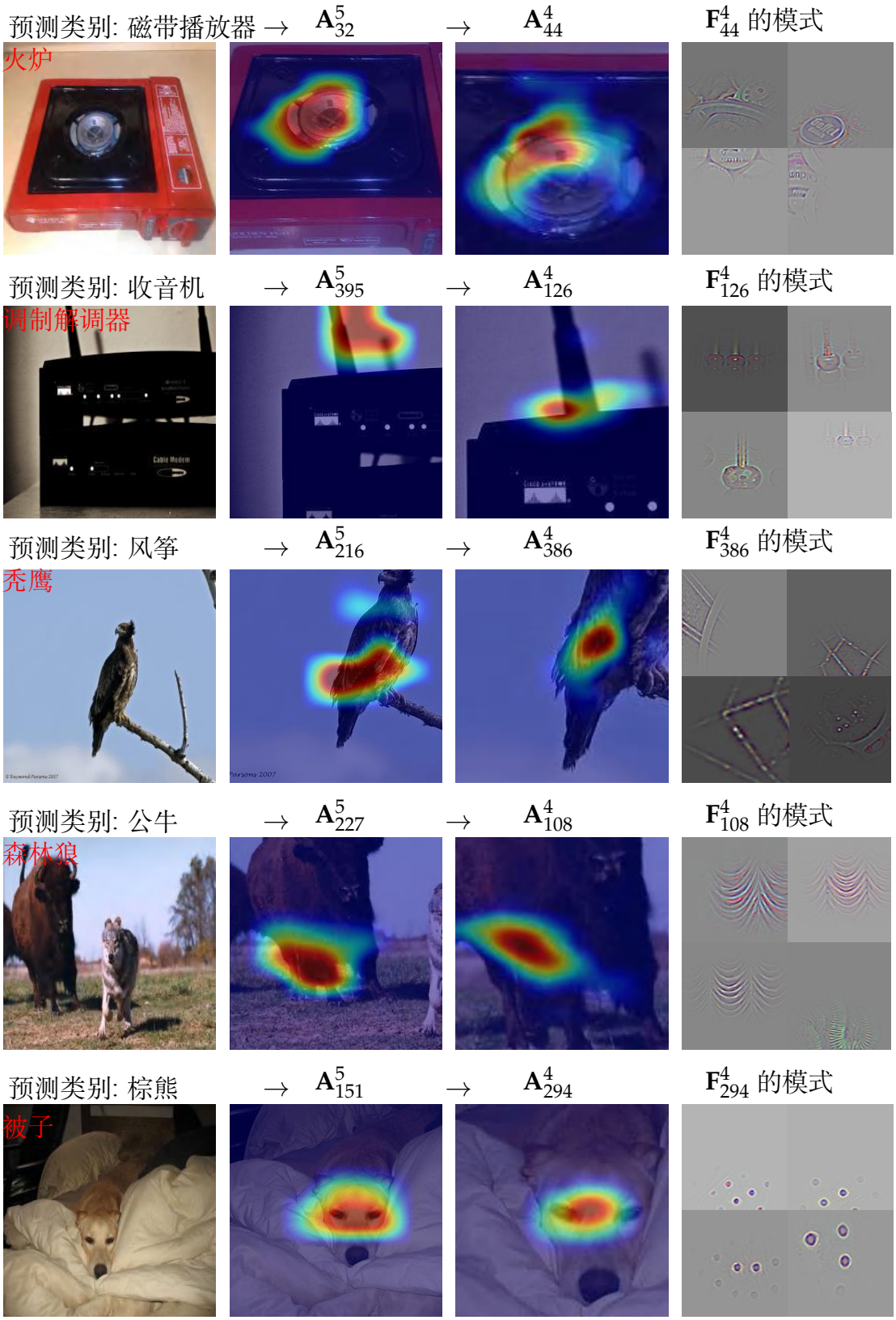
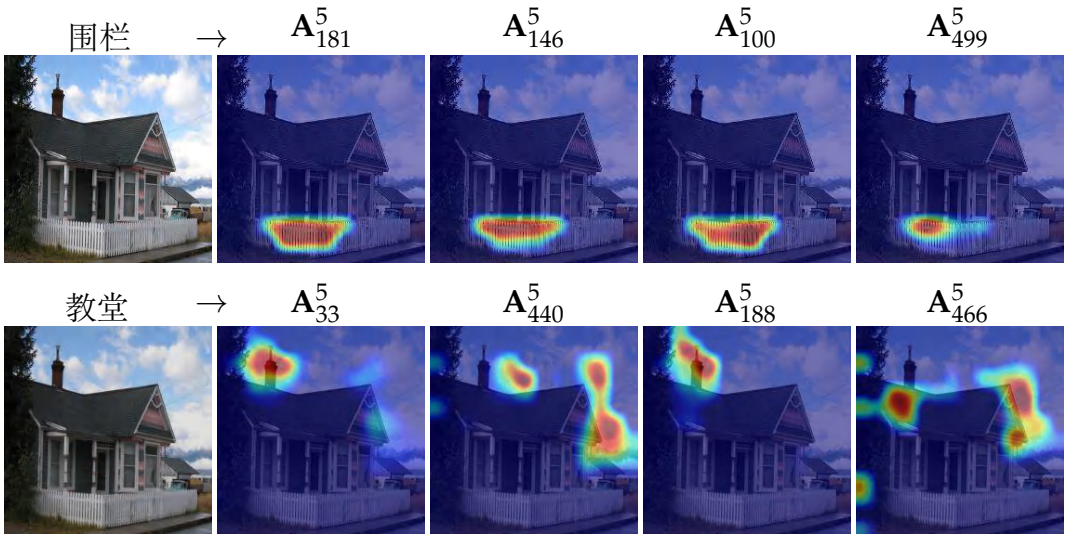
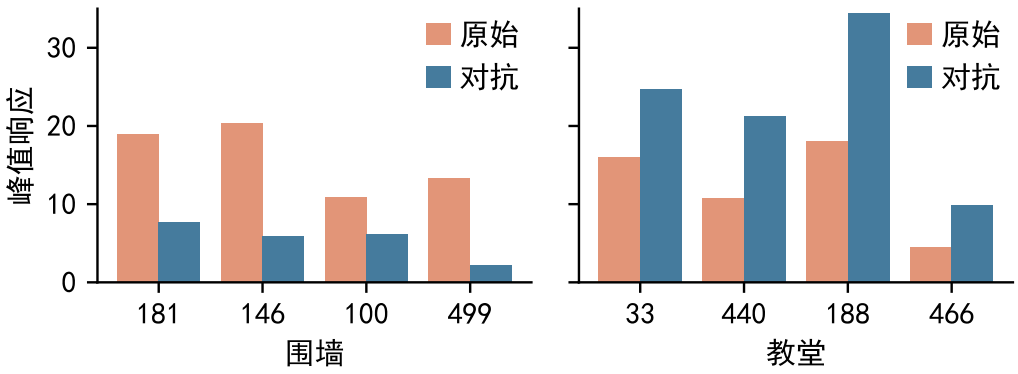


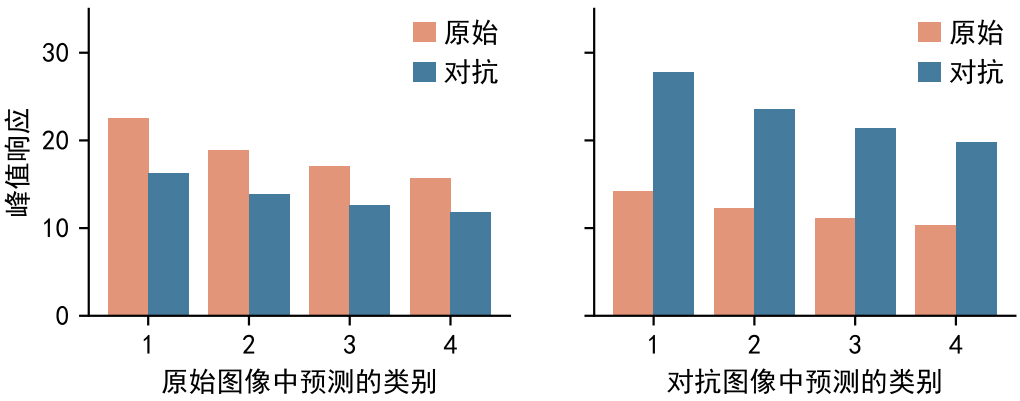
图 5.11 失败案例归因的更多例子。→ 表示注意力分解。目标标签被标记为红色。前三行的错误预测图像归因于目标类别和预测类别之间存在相似特征。最后两行的图像归因于图像中同时存在目标类别和预测类别的特征。



(a) 分解



(b) 峰值响应



(c) 平均峰值响应

图 5.12 对抗性攻击的例子。(a) 顶部是原始图像，底部是对抗图像。→ 表示分解。(b) 绘制原始图像和对抗图像中网络决策最重要通道的峰值响应。(c) 绘制整个 ILSVRC 验证数据集[108] 上原始图像和对抗图像中网络决策的最重要通道的平均峰值响应。正确类别的重要通道峰值大幅下降，错误类别的重要通道峰值大幅增加。

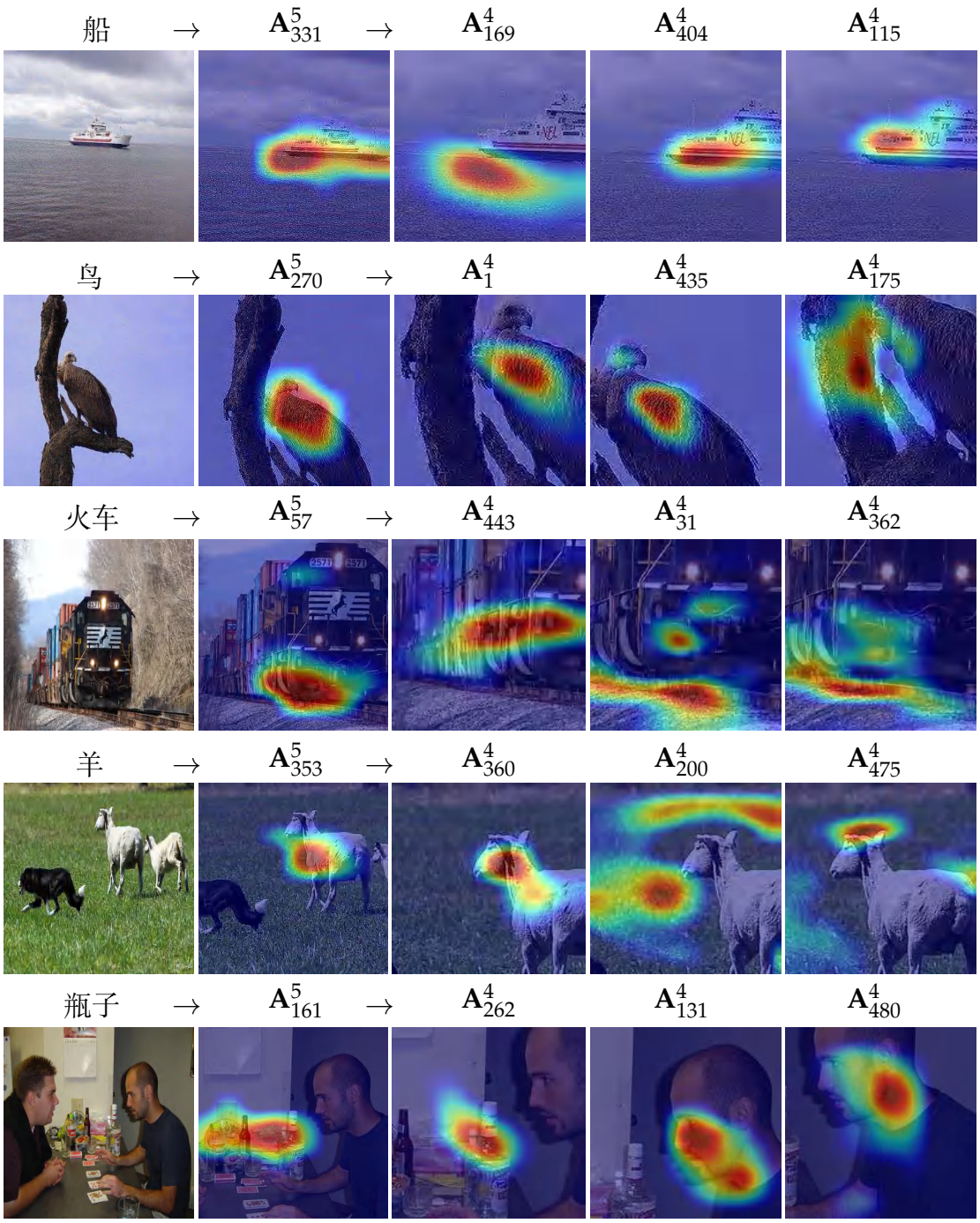


图 5.13 注意力图中的上下文信息。→ 表示分解。目标类别通常出现在特定的语义环境中。第一行中对于“船”这个类别，conv4_3 中第 169 个通道定位大海区域。第二行中对于“鸟”这个类别，conv4_3 中第 175 个通道定位树枝区域。第三行中对于“火车”这个类别，conv4_3 中第 31 个通道定位铁轨区域。第四行中对于“羊”这个类别，conv4_3 中第 200 个通道定位草地区域。第五行中对于“瓶子”这个类别，conv4_3 中第 131 和 480 个通道定位人区域。

方式，本文研究了重要通道的特征响应的变化。如图 5.12(a) 所示，本文展示了原始图像（顶行）和对抗性图像（底行）。对抗性图像由一种流行的对抗攻击算法[161] 生成。VGG-16 将原始图像分类为栅栏类别（概率 92%），将对抗图像分类为教堂类别（概率 100%）。通过从网络决策到 `conv5_3` 层的分解，本文分别找到了对于栅栏和教堂类别预测的重要的前几个特征通道。

如图 5.12(b) 所示，当将对抗图像与原始图像进行比较时，可以观察到栅栏类别的重要通道，即第 181、146、100 和 499 个通道的峰值特征响应，大幅下降 11.3、14.5、4.7 和 11.1。然而，教堂类别的重要通道，即第 33、440、188 和 466 个通道的峰值特征响应大幅增加了 8.7、10.4、16.4 和 5.5。如图 5.12(c) 所示，本文还计算了在整个 ILSVRC[108] 验证集上前 4 个重要特征通道的平均峰值响应。对抗性攻击算法改变重要通道的特征响应以影响最终的网络决策。对于重要通道，对抗性算法减少了正确类别的特征响应，增加了错误类别的特征响应。

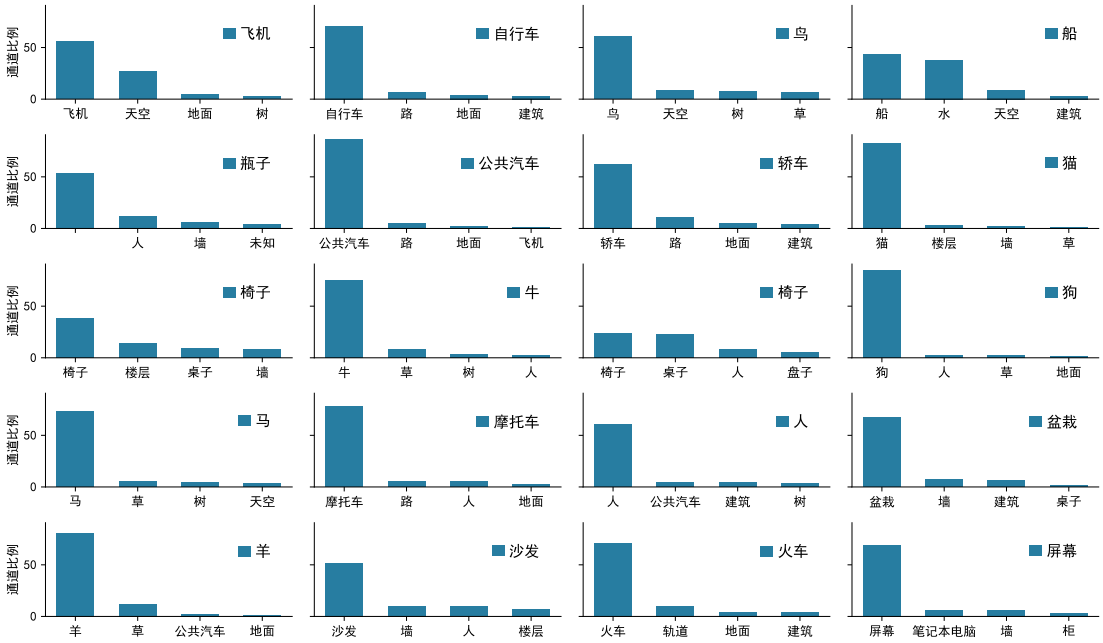


图 5.14　PASCAL VOC 2012 数据集中不同类别的上下文信息。可以看出每个类别都有特定的上下文信息。

5.3.2.3　注意力图中的上下文

上下文信息[162, 163] 对于卷积神经网络识别图像至关重要。一个已知的先验是目标类别通常出现在特定的语义环境中。例如，船通常出现在大海或湖泊中，鸟类经常站在树枝上。通过决策分解，可以在注意力图中找到了一些上下

文来支持卷积神经网络的预测。图 5.13的第一行显示了 conv5_3 层中的第 331 个通道对图像的“船”区域有强烈的响应。本文将注意力图指示的峰值点分解到 conv4_3 层。conv4_3 层的第 169、第 404 和第 115 个通道是最重要的前三个特征通道。最重要的通道是第 169 个通道，其对应的注意力图定位在大海区域。

为了定量分析注意力图中包含的上下文信息，本文利用 PASCAL-Context 数据集[164] 进行评估。本文从 PASCAL VOC 验证集[104] 中选择带有上下文标签的图像（标签从 PASCAL-Context 数据集获取），并为每个类别计算最频繁的上下文标签。具体来说，本文首先将注意力图阈值化为二值图，然后计算二值图和每个上下文区域之间的 IoU。注意力图被分配了对应于最大 IoU 的上下文区域的标签。在图 5.14中，本文展示了不同类别（例如鸟、船和火车）的上下文信息。这些类别通常出现在特定环境中。这一事实表明，对象的上下文信息对于卷积神经网络的决策至关重要。卷积神经网络可以学习到每个类别特定的语义信息偏置。

5.3.3 讨论

目前，本文将该方法应用于经典的卷积神经网络模型，例如 AlexNet、VGGNet 和 ResNet。网络中的组件有以下几种，卷积层、批归一化层、激活层和池化层。其它的带有新组件的卷积神经网络模型是一个可以继续研究的方向。

第四节 本章小结

本章提出了一种新颖的基于梯度的激活传播 (gAP) 模块，该模块可以将任何卷积神经网络的决策以及中间特征分解到其较低层。基于 gAP，网络决策可以被层次化分解为与模型所有层相关的丰富证据金字塔集。层次化注意力分解允许用户以自上而下的方式深入研究卷积神经网络的决策过程。本文已经通过实验验证了所提出方法的有效性，并展示了它理解和诊断卷积神经网络预测的能力。基于梯度的激活传播模块的准确性决定了层次化证据的质量。为了验证这个模块的准确性，本章进行了一系列健全性实验。实验结果表明，基于梯度的激活传播模块可以准确的分解决策，同时相比其它方法具有速度优势。此外，本章还基于层次化注意力分解机制来对卷积神经网络进行诊断，包括错误决策分析，对抗样本分析以及上下文语义分析等。

第六章 总结与展望

注意力机制在计算机视觉领域是至关重要的，本文主要研究了注意力机制在几个不用视觉任务中的应用。根据不同任务的需求，本文分别提出了基于注意力机制的改进策略。在本章，本文将对论文每章介绍的内容进行总结，并对每个方法的改进方向进行展望。

第一节 本文工作总结

本文在第一章首先介绍了注意力机制的研究背景和意义，并详细分析了目前的研究难点，随后对本文的研究目标与主要贡献进行简单概述。在第二章，本文介绍了注意力机制的研究现状，并介绍了大量的相关工作。接下来，本文分别对注意力机制的几个应用领域进行了相关工作介绍。

在第三章，本文提出了基于层次化注意力的物体定位方法，用于从网络不同阶段的卷积层生成注意力图。由于已有方法从卷积神经网络深层特征生成注意力图，而深层特征分辨率较低，导致定位的物体区域粗糙。本文分析了已有方法在浅层生成注意力图质量不好的原因，即特征图的全局权重无法代替每个像素对于决策的贡献。因此本文提出使用局部权重代替全局权重，成功的在浅层生成可靠的注意力图。实验结果证明，本文的注意力方法生成的注意力图不仅在浅层定位效果远远好于其它注意力模型，在深层同样好于它们。此外，从浅层生成的注意力图可以定位更加精细的物体位置，本文将浅层和深层的注意力图结合起来以定位更准确的物体区域。

第四章介绍了基于在线注意力累积的物体定位方法。由于分类网络需要寻找共同的模式来让数据集中同一类的所有图像都被识别成功，因此注意力图通常定位在很小的判别性物体区域处。然而弱监督任务，例如语义分割，需要完整的物体位置信息。本文发现了分类网络在训练过程中的不同时刻生成的注意力图定位的区域通常是不同的并且互补，因此本文提出在线注意力累积策略将注意力图累积起来，从而从最终的累积注意力图来定位物体区域。虽然在线注意力累积策略很有效，但是对于一些图像，注意力图定位的区域在训练阶段的不同时刻发生的变化较小，本文提出了注意力遮挡层来解决注意力在物体区域

上移动范围小的问题。此外，虽然累积注意力图比从最终模型生成的注意力图定位的物体区域更加完整，但是本文发现累积注意力图中存的一些注意力值较小的物体区域，本文提出了完整注意力学习策略来进一步提升累积注意力图定位物体的完整性。实验结果表明，当本文方法生成的注意力图应用于弱监督语义分割等任务中时，超越了已有最好方法的性能。

在第五章，注意力图虽然可以归因输入和决策之间的关系，但是它无法被用来深入理解网络内部特征通道。为了能够分析网络内部特征通道对决策的影响以及理解特征通道之间的关系，本文提出了基于层次化注意力分解的归因方法。为了实现层次化注意力分解，本文首先提出了一个基于梯度的激活回传模块。与已有的注意力方法生成一个整体的注意力图不同的是，该模块可以为每个特征通道生成注意力图。这个模块的优点是可以将决策分解到每个通道，然后迭代的将每个通道向浅层继续分解，直到网络底层。通过层次化注意力分解方法，本文可以得到支持决策的一系列相互关联的层次化证据。这些证据说明了特征通道对于决策的重要性以及特征通道之间的关系。

第二节 未来工作展望

本文提出了基于注意力机制的物体定位和归因方法，但是这些方法仍然存在缺陷。接下来，本文分别对每一种方法提出一些解决办法。

第三章提出的基于层次化注意力的物体定位方法，该方法可以从卷积神经网络的深层和浅层生成可靠的类别注意力图。当本文将不同层次的注意力图融合起来时，注意力图的定位能力可以进一步获得提升。本文使用了最简单的融合方法来结合不同层次的注意力图，研究更加有效的结合方式是一个可以提升的点。此外，从浅层生成的注意力图目前只用于物体定位，如何用浅层生成的注意力图解释网络特性是另一个可以研究的点。目前物体定位算法只能用于处理单个物体的定位，当图像出现同一类别的多个物体时，定位算法通常无法将它们区分开。不同实例物体的定位信息可以帮助弱监督物体检测以及实例分割任务。一个可行的方案是将该物体定位算法和峰值响应图[119] 结合起来定位单个物体区域。由于过分割图有很好的物体边界，为了提升物体定位算法在边界处的精度，一个可行的方案是将注意力图和过分割图[165] 结合起来提取精确的物体边界。

第四章提出了基于在线注意力累积的物体定位方法。虽然可以通过这种方

法挖掘到完整的物体区域，但是注意力图在物体边界上定位不准确，因此提升注意力图在物体边缘处的定位精度是一个可以继续提升的方向。一种可行的方案是结合过分割图[165] 定位精确的物体边界。此外，本文利用了显著性图生成背景的位置信息，如何移除对于显著性图的依赖是另一个可以提升的方向。第三点是虽然该方法在 PASCAL VOC 2012 数据集上取得了非常好的性能，但是由于这个数据集包含的图片数量有限，且图像中类别也有限，在复杂数据集上的表现效果未知。探索在大规模数据集上，例如 COCO 数据集[166]，本文方法的性能是一个可以研究的方向。

第五章提出了基于层次化注意力分解的归因方法。它可以生成丰富的层次化证据对卷积神经网络决策进行归因，但是它还存在着两大缺陷。首先，本文所提出的层次化注意力分解方法通过从卷积神经网络的不同层中选择一组强相关通道来解释个体决策。这些特征通道提供了丰富的证据层次。然而，让人类基于特征通道去理解网络推理过程仍然很难，因为并非所有示例都像含有人类别的图像一样容易理解。因此，本文将尝试在选定的特征通道和人类特定概念[93] 之间建立联系，帮助人类更好的理解特征通道的作用。此外，本文依次消融每个通道以研究它们的贡献。然而，正如一些工作[94, 167] 中所验证的那样，特征表示通常分布在多个通道中。本文同样观察到这一现象，即从同一决策分解的一些特征通道的注意力图通常在相似的空间位置具有强激活。这种现象表明多个特征通道通常一起产生作用。单独研究每个通道的作用会忽略这种分布的存在。一种可能解决方案是通过测量注意力图之间的重叠来找到具有相似效果的特征通道。然后将这些特征通道结合到一起来分析它们对网络决策的影响。

参考文献

[1] 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.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. [C] // International Conference on Learning Representation: 2015.

[3] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2017: 4700–4708.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2014: 580–587.

[5] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. [C] // Advances in Neural Information Processing Systems: 2015: 91–99.

[6] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2016: 779–788.

[7] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2015: 3431–3440.

[8] CHEN L.-C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40 (4): 834–848.

[9] LIN G, MILAN A, SHEN C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2017: 1925–1934.

[10] ZHANG D, HAN J, CHENG G, et al. Weakly supervised object localization and detection: a survey. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[11] TREISMAN A M, GELADE G. A feature-integration theory of attention. [J]. Cognitive psychology, 1980, 12 (1): 97–136.

[12] 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.

[13] 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.

[14] 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.

[15] 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 Workshops: 2017: 618–626.

[16] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. [C] // IEEE Winter Conference on Applications of Computer Vision: 2018: 839–847.

[17] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. [J]. International Journal of Computer Vision, 2020, 128 (2): 336–359.

[18] 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.

[19] BACH S, BINDER A, MONTAVON G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. [J]. PloS one, 2015, 10 (7): e0130140.

[20] SUNDARARAJAN M, TALY A, YAN Q. Axiomatic attribution for deep networks. [C] // International Conference on Machine Learning: 2017.

[21] SHRIKUMAR A, GREENSIDE P, KUNDAJE A. Learning important features through propagating activation differences. [C] // International Conference on Machine Learning: 2017.

[22] DHAMDHERE K, SUNDARARAJAN M, YAN Q. How important is a neuron? [J]. International Conference on Learning Representation, 2019.

[23] LEINO K, SEN S, DATTA A, et al. Influence-directed explanations for deep convolutional networks. [C] // 2018 IEEE International Test Conference: 2018: 1–8.

[24] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2017: 1492–1500.

[25] ZAGORUYKO S, KOMODAKIS N. Wide Residual Networks. [C] // British Machine Vision Conference: 2016.

[26] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2016: 2818–2826.

[27] YANG S, KIM Y, KIM Y, et al. Combinational Class Activation Maps for Weakly Supervised Object Localization. [C] // IEEE Winter Conference on Applications of Computer Vision: 2020: 2941–2949.

[28] SHI Z, HOSPEDALES T M, XIANG T. Bayesian Joint Modelling for Object Localisation in Weakly Labelled Images. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37 (10): 1959–1972.

[29] CHEN Y, LIN Y, YANG M, et al. Show, Match and Segment: Joint Weakly Supervised Learning of Semantic Matching and Object Co-segmentation. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[30] MENG Q, WANG W, ZHOU T, et al. Weakly Supervised 3D Object Detection from Lidar Point Cloud. [C] // European Conference on Computer Vision: 2020: 515–531.

[31] HOU Q, MASSICETI D, DOKANIA P K, et al. Bottom-up top-down cues for weakly-supervised semantic segmentation. [C] // International Workshops on Energy Minimization Methods in Computer Vision Pattern Recognition: 2017: 263–277.

[32] CHOLAKKAL H, SUN G, KHAN F S, et al. Object counting and instance segmentation with image-level supervision. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2019: 12397–12405.

[33] SUNG A. Ranking importance of input parameters of neural networks. [J]. Expert systems with Applications, 1998, 15 (3-4): 405–411.

[34] BAEHRENS D, SCHROETER T, HARMELING S, et al. How to explain individual classification decisions. [J]. The Journal of Machine Learning Research, 2010, 11: 1803–1831.

[35] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: Visualising image classification models and saliency maps. [C] // International Conference on Learning Representation Workshops: 2014.

[36] SPRINGENBERG J T, DOSOVITSKIY A, BROX T, et al. Striving for simplicity: The all convolutional net. [C] // International Conference on Learning Representation Workshops: 2015.

[37] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks. [C] // European Conference on Computer Vision: 2014: 818–833.

[38] KINDERMANS P.-J, SCHÜTT K T, ALBER M, et al. Learning how to explain neural networks: PatternNet and PatternAttribution. [C] // International Conference on Learning Representation: 2018.

[39] KIM B, SEO J, JEON S, et al. Why are saliency maps noisy? cause of and solution to noisy saliency maps. [C] // IEEE International Conference on Computer Vision. IEEE: 2019: 4149–4157.

[40] SRINIVAS S, FLEURET F. Full-gradient representation for neural network visualization. [C] // Advances in Neural Information Processing Systems: 2019: 4124–4133.

[41] ZHANG J, LIN Z, BRANDT J, et al. Top-Down Neural Attention by Excitation Backprop. [C] // European Conference on Computer Vision: 2016: 543–559.

[42] MONTAVON G, LAPUSCHKIN S, BINDER A, et al. Explaining nonlinear classification decisions with deep taylor decomposition. [J]. Pattern Recognition, 2017, 65: 211–222.

[43] YANG Y, QIU J, SONG M, et al. Learning propagation rules for attribution map generation. [C] // European Conference on Computer Vision: 2020: 672–688.

[44] SMILKOV D, THORAT N, KIM B, et al. Smoothgrad: removing noise by adding noise. [C] // International Conference Mach. Learning Workshops: 2017.

[45] KIM B, WATTENBERG M, GILMER J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). [C] // International Conference on Machine Learning: 2018: 2668–2677.

[46] WANG H, WANG Z, DU M, et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. [C] // IEEE Conference on Computer Vision and Pattern Recognition Workshops: 2020: 24–25.

[47] 廖南星, 周世斌, 张国鹏, 等. 基于类激活映射-注意力机制的图像描述方法. [J]. 山东大学学报: 工学版, 2020, 50 (4): 28–34.

[48] 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.

[49] HOU Q, JIANG P, WEI Y, et al. Self-Erasing Network for Integral Object Attention. [C] // Advances in Neural Information Processing Systems. Vol. 31: 2018: 549–559.

[50] ZHANG X, WEI Y, KANG G, et al. Self-produced guidance for weakly-supervised object localization. [C] // European Conference on Computer Vision: 2018: 597–613.

[51] ZINTGRAF L M, COHEN T S, ADEL T, et al. Visualizing deep neural network decisions: Prediction difference analysis. [C] // International Conference on Learning Representation: 2017.

[52] PETSIUK V, DAS A, SAENKO K. Rise: Randomized input sampling for explanation of black-box models. [C] // British Machine Vision Conference: 2018.

[53] RIBEIRO M T, SINGH S, GUESTRIN C. "Why should i trust you?" Explaining the predictions of any classifier. [C] // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining: 2016: 1135–1144.

[54] FONG R C, VEDALDI A. Interpretable explanations of black boxes by meaningful perturbation. [C] // IEEE International Conference on Computer Vision Workshops: 2017: 3429–3437.

[55] FONG R, PATRICK M, VEDALDI A. Understanding deep networks via extremal perturbations and smooth masks. [C] // IEEE International Conference on Computer Vision Workshops: 2019: 2950–2958.

[56] DABKOWSKI P, GAL Y. Real time image saliency for black box classifiers. [C] // Advances in Neural Information Processing Systems: 2017.

[57] CINBIS R G, VERBEEK J, SCHMID C. Weakly supervised object localization with multi-fold multiple instance learning. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39 (1): 189–203.

[58] GOKBERK CINBIS R, VERBEEK J, SCHMID C. Multi-fold mil training for weakly supervised object localization. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2014: 2409–2416.

[59] GALLEGUILLOS C, BABENKO B, RABINOVICH A, et al. Weakly supervised object localization with stable segmentations. [C] // European Conference on Computer Vision: 2008: 193–207.

[60] KANTOROV V, OQUAB M, CHO M, et al. Contextlocnet: Context-aware deep network models for weakly supervised localization. [C] // European Conference on Computer Vision: 2016: 350–365.

[61] LI D, HUANG J.-B, LI Y, et al. Weakly supervised object localization with progressive domain adaptation. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2016: 3512–3520.

[62] JIE Z, WEI Y, JIN X, et al. Deep self-taught learning for weakly supervised object localization. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2017: 1377–1385.

[63] TEH E W, ROCHAN M, WANG Y. Attention Networks for Weakly Supervised Object Localization. [C] // British Machine Vision Conference: 2016: 1–11.

[64] XU W, WU Y, MA W, et al. Adaptively denoising proposal collection for weakly supervised object localization. [J]. Neural Processing Letters, 2020, 51 (1): 993–1006.

[65] ZHU Y, ZHOU Y, YE Q, et al. Soft proposal networks for weakly supervised object localization. [C] // IEEE International Conference on Computer Vision Workshops: 2017: 1841–1850.

[66] XUE H, LIU C, WAN F, et al. Danet: Divergent activation for weakly supervised object localization. [C] // IEEE International Conference on Computer Vision Workshops: 2019: 6589–6598.

[67] SINGH K K, LEE Y J. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. [C] // IEEE International Conference on Computer Vision Workshops. IEEE: 2017: 3544–3553.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] HOU Q, CHENG M.-M, HU X, et al. Deeply supervised salient object detection with short connections. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41 (4): 815–828. DOI: 10.1109/TPAMI.2018.2815688.

[76] 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. DOI: 10.1007/s11263-016-0977-3. ISSN: 1573-1405.

[77] 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. DOI: 10.1109/TPAMI.2014.2345401.

[78] LIU J.-J, HOU Q, CHENG M.-M, et al. A simple pooling-based design for real-time salient object detection. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2019: 3917–3926.

[79] PAPANDREOU G, CHEN L.-C, MURPHY K P, et al. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. [C] // IEEE International Conference on Computer Vision Workshops: 2015: 1742–1750.

[80] ZHANG B, XIAO J, WEI Y, et al. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. [C] // AAAI Conference on Artificial Intelligence. Vol. 34. 07: 2020: 12765–12772.

[81] ARASLANOV N, ROTH S. Single-stage semantic segmentation from image labels. [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2020: 4253–4262.

[82] 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.

[83] AHN J, CHO S, KWAK S. Weakly supervised learning of instance segmentation with inter-pixel relations. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2019: 2209–2218.

[84] ZHOU B, SUN Y, BAU D, et al. Interpretable basis decomposition for visual explanation. [C] // European Conference on Computer Vision: 2018: 119–134.

[85] OLAH C, CAMMARATA N, SCHUBERT L, et al. Zoom in: An introduction to circuits. [J]. Distill, 2020, 5 (3): e00024–001.

[86] ERHAN D, BENGIO Y, COURVILLE A, et al. Visualizing higher-layer features of a deep network. [J]. University of Montreal, 2009, 1341 (3): 1.

[87] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Object detectors emerge in deep scene cnns. [C] // International Conference on Learning Representation: 2015.

[88] MORDVINTSEV A, OLAH C, TYKA M. Inceptionism: Going Deeper into Neural Networks. 2015.

[89] YOSINSKI J, CLUNE J, NGUYEN A, et al. Understanding neural networks through deep visualization. [C] // International Conference on Machine Learning Workshops: 2015.

[90] MAHENDRAN A, VEDALDI A. Understanding deep image representations by inverting them. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2015: 5188–5196.

[91] DOSOVITSKIY A, BROX T. Inverting visual representations with convolutional networks. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2016: 4829–4837.

[92] OLAH C, MORDVINTSEV A, SCHUBERT L. Feature visualization. [J]. Distill, 2017, 2 (11): e7.

[93] BAU D, ZHOU B, KHOSLA A, et al. Network dissection: Quantifying interpretability of deep visual representations. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2017: 6541–6549.

[94] FONG R, VEDALDI A. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2018: 8730–8738.

[95] BAU D, ZHU J.-Y, STROBELT H, et al. Understanding the role of individual units in a deep neural network. [J]. Proceedings of the National Academy of Sciences, 2020.

[96] CHEN R, CHEN H, REN J, et al. Explaining neural networks semantically and quantitatively. [C] // IEEE International Conference on Computer Vision Workshops: 2019: 9187–9196.

[97] FROSST N, HINTON G. Distilling a neural network into a soft decision tree. [C] // CEX workshop at AIIA: 2017.

[98] LIU X, WANG X, MATWIN S. Improving the interpretability of deep neural networks with knowledge distillation. [C] // IEEE International Conference Data Mining Workshops. IEEE: 2018: 905–912.

[99] CHEN C, LI O, TAO D, et al. This Looks Like That: Deep Learning for Interpretable Image Recognition. [C] // Advances in Neural Information Processing Systems. Vol. 32: 2019: 8930–8941.

[100] KOH P W, NGUYEN T, TANG Y S, et al. Concept bottleneck models. [C] // International Conference on Machine Learning: 2020: 5338–5348.

[101] KUMAR N, BERG A C, BELHUMEUR P N, et al. Attribute and simile classifiers for face verification. [C] // IEEE International Conference on Computer Vision Workshops: 2009: 365–372.

[102] LAMPERT C H, NICKISCH H, HARMELING S. Learning to detect unseen object classes by between-class attribute transfer. [C] // IEEE Conference on Computer Vision and Pattern Recognition. IEEE: 2009: 951–958.

[103] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 Dataset. [R]. CNS-TR-2011-001. California Institute of Technology, 2011.

[104] 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.

[105] HARIHARAN B, ARBELÁEZ P, BOURDEV L, et al. Semantic contours from inverse detectors. [C] // IEEE International Conference on Computer Vision Workshops: 2011: 991–998.

[106] 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: 7322–7330.

[107] LI D, HUANG J, LI Y, et al. Progressive Representation Adaptation for Weakly Supervised Object Localization. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42 (6): 1424–1438.

[108] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge. [J]. International Journal of Computer Vision, 2015.

[109] REBUFFI S.-A, FONG R, JI X, et al. There and back again: Revisiting backpropagation saliency methods. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2020: 8839–8848.

[110] BOYKOV Y Y, JOLLY M.-P. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. [C] // IEEE International Conference on Computer Vision Workshops. Vol. 1. IEEE: 2001: 105–112.

[111] CHOE J, SHIM H. Attention-based dropout layer for weakly supervised object localization. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2019: 2219–2228.

[112] DONG H, SONG K, HE Y, et al. PGA-Net: Pyramid Feature Fusion and Global Context Attention Network for Automated Surface Defect Detection. [J]. IEEE Transactions on Industrial Informatics, 2019.

[113] SU B, y. CHEN H, CHEN P, et al. Deep Learning-based Solar-Cell Manufacturing Defect Detection with Complementary Attention Network. [J]. IEEE Transactions on Industrial Informatics, 2020.

[114] TANG Z, TIAN E, WANG Y, et al. Non-Destructive Defect Detection in Castings by Using Spatial Attention Bilinear Convolutional Neural Network. [J]. IEEE Transactions on Industrial Informatics, 2020.

[115] LU S, FENG J, ZHANG H, et al. An Estimation Method of Defect Size From MFL Image Using Visual Transformation Convolutional Neural Network. [J]. IEEE Transactions on Industrial Informatics, 2019, 15 (1): 213–224.

[116] WIELER M, HAHN T. Weakly supervised learning for industrial optical inspection. 2007.

[117] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (12): 2481–2495.

[118] ARBELÁEZ P, PONT-TUSET J, BARRON J T, et al. Multiscale combinatorial grouping. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2014.

[119] ZHOU Y, ZHU Y, YE Q, et al. Weakly supervised instance segmentation using class peak response. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2018: 3791–3800.

[120] CHAUDHRY A, DOKANIA P K, TORR P H. Discovering class-specific pixels for weakly-supervised semantic segmentation. [C] // British Machine Vision Conference: 2017: 20.1–20.13.

[121] FENG J, WANG X, LIU W. Deep graph cut network for weakly-supervised semantic segmentation. [J]. Science China Information Sciences, 2021, 64 (3): 130105.

[122] GAO S.-H, CHENG M.-M, ZHAO K, et al. Res2Net: A New Multi-scale Backbone Architecture. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43 (2): 652–662. DOI: 10.1109/TPAMI.2019.2938758.

[123] 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 Representation: 2015.

[124] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2017: 2881–2890.

[125] ZHANG H, DANA K, SHI J, et al. Context encoding for semantic segmentation. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2018: 7151–7160.

[126] HOU Q, ZHANG L, CHENG M.-M, et al. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2020: 4003–4012.

[127] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation. [C] // IEEE International Conference on Computer Vision Workshops: 2019: 603–612.

[128] PAPADOPOULOS D P, CLARKE A D, KELLER F, et al. Training object class detectors from eye tracking data. [C] // European Conference on Computer Vision: 2014: 361–376.

[129] PATHAK D, KRAHENBUHL P, DARRELL T. Constrained convolutional neural networks for weakly supervised segmentation. [C] // IEEE International Conference on Computer Vision Workshops: 2015: 1796–1804.

[130] LI K, WU Z, PENG K.-C, et al. Guided attention inference network. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42 (12): 2996–3010.

[131] ARPIT D, JASTRZĘBSKI S, BALLAS N, et al. A closer look at memorization in deep networks. [C] // International Conference on Machine Learning: 2017: 233–242.

[132] 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. DOI: 10.1109/TPAMI.2016.2636150.

[133] SHIMODA W, YANAI K. Distinct class-specific saliency maps for weakly supervised semantic segmentation. [C] // European Conference on Computer Vision: 2016: 218–234.

[134] 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.

[135] 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: 5038–5047.

[136] JIN B, ORTIZ SEGOVIA M V, SUSSTRUNK S. Webly supervised semantic segmentation. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2017: 3626–3635.

[137] KIM D, YOO D, KWEON I S, et al. Two-phase learning for weakly supervised object localization. [C] // IEEE International Conference on Computer Vision Workshops: 2017: 3534–3543.

[138] 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.

[139] KRÄHENBÜHL P, KOLTUN V. Efficient inference in fully connected crfs with gaussian edge potentials. [C] // Advances in Neural Information Processing Systems. Vol. 24: 2011: 109–117.

[140] SHIMODA W, YANAI K. Self-supervised difference detection for weakly-supervised semantic segmentation. [C] // IEEE International Conference on Computer Vision Workshops: 2019: 5208–5217.

[141] WANG Y, ZHANG J, KAN M, et al. Self-supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2020: 12275–12284.

[142] CHANG Y.-T, WANG Q, HUNG W.-C, et al. Weakly-Supervised Semantic Segmentation via Sub-Category Exploration. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2020: 8991–9000.

[143] SUN G, WANG W, DAI J, et al. Mining cross-image semantics for weakly supervised semantic segmentation. [C] // European Conference on Computer Vision: 2020: 347–365.

[144] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional architecture for fast feature embedding. [C] // ACM International Conference on Multimedia: 2014: 675–678.

[145] GIRSHICK R. Fast r-cnn. [C] // IEEE International Conference on Computer Vision Workshops: 2015: 1440–1448.

[146] ZHU Z, LIANG D, ZHANG S, et al. Traffic-sign detection and classification in the wild. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2016: 2110–2118.

[147] HOU Y, MA Z, LIU C, et al. Learning lightweight lane detection cnns by self attention distillation. [C] // IEEE International Conference on Computer Vision Workshops: 2019: 1013–1021.

[148] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation. [C] // International Conference Medical image computing and computer-assisted intervention: 2015: 234–241.

[149] LITJENS G, KOOI T, BEJNORDI B E, et al. A survey on deep learning in medical image analysis. [J]. Medical image analysis, 2017, 42: 60–88.

[150] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples. [C] // International Conference on Learning Representation: 2014.

[151] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world. [C] // International Conference on Learning Representation Workshops: 2017.

[152] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples. [C] // International Conference on Machine Learning: 2018.

[153] GIANNELLI P C. Chain of custody and the handling of real evidence. [J]. Am. Crim. L. Rev., 1982, 20: 527.

[154] MURAD M H, ASI N, ALSAWAS M, et al. New evidence pyramid. [J]. BMJ Evidence-Based Medicine, 2016, 21 (4): 125–127.

[155] ZHANG Q, YANG Y, MA H, et al. Interpreting cnns via decision trees. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2019: 6261–6270.

[156] BENESTY J, CHEN J, HUANG Y, et al. Pearson correlation coefficient. [G] // Noise reduction in speech processing: Springer, 2009: 1–4.

[157] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression. [C] // 18th International Conference on Pattern Recognition (ICPR'06). Vol. 3. IEEE: 2006: 850–855.

[158] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks. [C] // Advances in Neural Information Processing Systems: 2012: 1097–1105.

[159] SEDGWICK P. Spearman' s rank correlation coefficient. [J]. Bmj, 2014, 349.

[160] ADEBAYO J, GILMER J, MUELLY M, et al. Sanity Checks for Saliency Maps. [C] // Advances in Neural Information Processing Systems. Vol. 31: 2018.

[161] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks. [C] // International Conference on Learning Representation: 2018.

[162] OQUAB M, BOTTOU L, LAPTEV I, et al. Is object localization for free?-weakly-supervised learning with convolutional neural networks. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2015: 685–694.

[163] KUMAR S, HEBERT M. A hierarchical field framework for unified context-based classification. [C] // IEEE International Conference on Computer Vision Workshops. Vol. 2. IEEE: 2005: 1284–1291.

[164] MOTTAGHI R, CHEN X, LIU X, et al. The role of context for object detection and semantic segmentation in the wild. [C] // IEEE Conference on Computer Vision and Pattern Recognition: 2014: 891–898.

[165] 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, 2016, 39 (1): 128–140.

[166] LIN T.-Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context. [C] // European Conference on Computer Vision: 2014.

[167] LEAVITT M L, MORCOS A S. Selectivity considered harmful: evaluating the causal impact of class selectivity in DNNs. [C] // International Conference on Learning Representation: 2020.

致谢

首先，本人由衷感谢我的导师程明明教授对我的悉心指导和栽培。从程老师身上学习到很多终身受益的技能。感谢北京交通大学的魏云超老师、侯淇彬学长对我的指导与帮助，和他们的合作让我的科研工作更加顺利。

其次，感谢我的同学曹洋、刘姜江、赵凯、赵嘉星、许刚和张鹏对我的支持和鼓励，他们陪伴我度过了博士生涯中的艰难坎坷，能和他们一起同窗，是我的荣幸也是我的幸运。特别感谢和我合作过的师弟韩凌昊、杨雨奇、张长彬，他们在我的科研工作中帮助了我很多。

最后，感谢我的家人，在选择读博的时候，他们一直支持我的决定，并陪伴我度过艰难的日子。在人生的道路上，家人的支持就是我的动力。

个人简历

个人介绍:

姜鹏涛，出生于 1995 年 2 月 13 日。在 2013 年进入西安电子科技大学就读，在 2017 年毕业并获得学士学位。在 2017 年在南开大学就读博士研究生至今。

公开发表的学术论文:

[1] **Peng-Tao Jiang**, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, Hongkai Xiong. "Integral Object Mining via Online Attention Accumulation." IEEE/CVF International Conference on Computer Vision (ICCV), 2019. PDF

[2] **Peng-Tao Jiang***, Ling-Hao Han*, Qibin Hou, Ming-Ming Cheng, Yunchao Wei. "Online Attention Accumulation for Weakly Supervised Semantic Segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021. PDF

[3] **Peng-Tao Jiang***, Chang-bin Zhang*, Qibin Hou, Ming-Ming Cheng, Yunchao Wei. "LayerCAM: Exploring Hierarchical Class Activation Maps for Localization." IEEE Transactions on Image Processing (TIP) 2021. PDF

[4] **Peng-Tao Jiang**, Yuqi Yang, Qibin Hou, Yunchao Wei. "L2G: A Simple Local-to-Global Knowledge Transfer Framework for Weakly Supervised Semantic Segmentation." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022. PDF

[5] Chang-bin Zhang*, **Peng-Tao Jiang***, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, Ming-Ming Cheng. "Delving Deep into Label Smoothing." IEEE Transactions on Image Processing (TIP) 2021. PDF

[6] Qibin Hou, **Peng-Tao Jiang**, Yunchao Wei, Ming-Ming Cheng. "Self-Erasing Network for Integral Object Attention." Neural Information Processing Systems (NeurIPS), 2018. PDF

[7] Yun Liu, **Peng-Tao Jiang**, Vahan Petrosyan, Shi-Jie Li, Jiawang Bian, Le Zhang, and Ming-Ming Cheng. "DEL: Deep Embedding Learning for Efficient Image Seg-

mentation.” International Joint Conferences on Artificial Intelligence (IJCAI), 2018. PDF

[8] Yu Zhang, Chang-bin Zhang, **Peng-Tao Jiang**, Feng Mao, Ming-Ming Cheng. “Personalized Image Semantic Segmentation.” IEEE/CVF International Conference on Computer Vision (ICCV), 2021. PDF

公开发表的专利:

程明明，姜鹏涛，张长彬，侯淇彬，曹洋，基于在线注意力累积的挖掘目标物体区域的方法，申请号：201910715341.X，申请日：2019-08-05。