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

复杂场景的自适应视觉感知

Adaptive Visual Perception of Complex Scenes

论文作者 高尚华 指导教师 程明明 教授
申请学位 工学博士 培养单位 计算机学院
学科专业 计算机科学与技术 研究方向 计算机视觉
答辩委员会主席 邹北骥 教授 评 阅 人 匿名评审

南开大学研究生院

二〇二三年五月

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)

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	复杂场景的自适应视觉感知				
姓　　名	高尚华	学号	1120200183	答辩日期	2023 年 5 月 23 日
论 文 类 别	博士☑　硕士□　专业硕士□　专业博士□　本科□　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	15009290281		电子邮箱	shgao@mail.nankai.edu.cn	
通讯地址 (邮编)：天津市南开区卫津路 94 号 (300071)					
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。

摘要

视觉感知旨在理解视觉场景中的语义要素。面对复杂场景进行鲁棒的视觉感知，需要感知系统具有强大的自适应处理能力，进而需要在模型架构和表征学习策略两个方面处理如下挑战：（1）面对复杂多变的场景，模型架构需要具备多尺度特征自适应处理能力和相应的感受野。（2）复杂场景产生的大规模和多样化的数据使人工标注过于昂贵，因而要求模型能够在尽可能少的人工干预下完成对数据的表征和理解。（3）复杂场景导致训练数据和模型体积的激增，因此要求新模型能够有效利用现有模型的知识储备自适应地学习更强的表征，从而降低训练学习的计算开销。

针对以上问题，本文着重研究在复杂场景下增强网络架构的多尺度和感受野自适应能力，以及表征学习策略的数据和模型学习的自适应能力。具体而言，本文对以下几个方面的自适应感知能力进行研究。

1）本文提出了基于残差递进多尺度模块的主干网络架构，其多尺度特征的自适应表征能力得到组合爆炸式的提升。所提出的主干网络架构在分类、检测等十多种典型视觉任务中取得了显著性能提升。

2）为了克服传统方法手工指定感受野大小的局限性，本文针对复杂场景实际需求，提出对卷积神经网络感受野进行从全局到局部的自适应搜索的有效算法。该感受野搜索算法可以有效提升多种应用中的模型性能。

3）为了避免昂贵的数据标注，本文提出了大规模无监督语义分割问题，并设计了有效算法。该算法通过自我表征学习从百万量级数据中学习丰富的语义特征，并将自适应总结出的上千个语义类别分配给大规模数据中的每个像素。本文验证了无需人工标注的自适应大规模视觉感知是可行的。

4）本文提出绿色可持续视觉感知模型学习的新概念。通过构建基于掩码重建的目标增强条件化自监督预训练机制，本文方法可以自适应地学习并超越特性各异的已有视觉模型，避免了现有神经网络基础模型从头开始训练导致的计算量和能耗过高的问题。

关键词： 自适应；视觉感知；多尺度；主干网络架构；感受野搜索；大规模无监督语义分割；可持续自监督表征学习;

Abstract

Visual perception aims to understand semantics in visual scenes. Robust visual perception in complex scenes requires the system to have strong adaptive processing capabilities. Enhancing the adaptive visual perception capability of complex scenes requires addressing the following challenges in network architecture and representation learning: (1) In complex scenes, networks have to adaptively process multi-scale features with corresponding receptive fields. (2) The large-scale and diverse data from complex scenes make the human annotation too costly, thus requiring the model to conduct representation learning with limited human intervention. (3) Complex scenes lead to a surge in training data and model size, which requires the new model can effectively utilize the knowledge of the existing model to adaptively learn stronger representation, thereby reducing the training computational cost.

To solve above problems, this paper focuses on enhancing the adaptive visual perception of complex scenes from the perspective of multi-scale ability and receptive field in network architecture, as well as the data representation and model training in representation learning. Specifically, this paper studies the adaptive perception capabilities in the following aspects.

1）This paper proposes a backbone network architecture based on a hierarchical residual-like structure, which significantly enhances the adaptive representation ability of multi-scale features. The proposed backbone network architecture achieves significant performance improvements in dozens of representative visual tasks such as classification and detection.

2) To meet the actual demands of complex senses and overcome the limitations of traditional methods that manually specify receptive fields, this paper proposes an effective global-to-local algorithm for adaptive searching of the receptive field combination of convolutional neural networks. The proposed method for receptive field search can significantly improve the model performance in various

applications.

3) To avoid expensive data annotation, this paper proposes a large-scale unsupervised semantic segmentation problem and designs an effective algorithm for it. The algorithm learns rich semantic features from millions of data by self-supervised learning, and assigns thousands of adaptive summarized semantic categories to each pixel in large-scale data. This paper verifies that adaptive large-scale visual perception without manual annotation is feasible.

4) This paper proposes a new concept for green and sustainable visual perception model learning. By constructing a mask-reconstruction-based target-enhanced conditional self-supervised pre-training mechanism, the proposed method can adaptively learn and surpass existing visual models with diverse characteristics, avoiding the problem of high computational and energy costs caused by training foundation models from scratch.

Key Words: Adaptive; Visual Perception; Multi-scale; Backbone network; Receptive field search; Large-scale unsupervised semantic segmentation; Sustainable self-supervised learning;

目录

摘要 …… I
Abstract …… II
第一章 绪论 …… 1
1.1 研究背景与意义 …… 1
1.2 研究现状 …… 3
1.3 研究目标和主要贡献 …… 5
1.4 本文组织结构 …… 8
第二章 相关工作介绍 …… 9
2.1 多尺度特征提取 …… 9
2.2 感受野 …… 11
2.3 无监督语义分割 …… 14
2.4 自监督表征学习 …… 16
第三章 残差递进尺度自适应表征 …… 19
3.1 多尺度自适应表征简介 …… 19
3.2 残差递进多尺度模块 …… 21
3.3 残差递进多尺度网络 …… 23
3.4 实验与分析 …… 24
3.5 场景自适应能力分析 …… 29
3.6 总结 …… 34
第四章 场景自适应感受野搜索 …… 35
4.1 场景自适应感受野搜索简介 …… 35
4.2 全局到局部的感受野搜索算法 …… 37
4.3 实验与分析 …… 43
4.4 场景自适应能力分析 …… 51
4.5 总结 …… 60
第五章 数据自适应大规模无监督语义分割 …… 61

5.1 大规模无监督语义分割简介 …… 61
5.2 大规模无监督语义分割基准 …… 64
5.3 大规模无监督语义分割方法 …… 71
5.4 实验与分析 …… 76
5.5 总结…… 89
第六章 模型自适应可持续自监督学习…… 91
6.1 可持续自监督学习简介 …… 91
6.2 目标增强的条件掩码重建可持续学习 …… 93
6.3 自适应能力分析 …… 99
6.4 实验与分析 …… 102
6.5 总结…… 108
第七章 总结与展望 …… 109
7.1 本文工作总结 …… 109
7.2 未来工作展望 …… 110
参考文献 …… 113
致谢 …… 145
个人简历 …… 147

第一章 绪论

1.1 研究背景与意义

视觉感知对视觉系统采集到的环境光线进行理解并加以运用，它是自然界中人类等生物与环境进行交互的关键能力。随着数字化技术的发展和移动互联网的普及，用于采集视觉信息的人工摄像设备在工业领域和消费领域被大量应用并产生海量的视觉数据。计算机的视觉感知技术旨在使计算机获得与人类相似的视觉感知理解能力，从而能够通过分析处理视觉数据实现对场景事物的分类、分割、定位、关系建模等能力。随着计算机算力的提升和深度学习技术的飞速发展，视觉感知技术的应用场景愈发广泛，为智能制造、医疗分析、智慧农业、自动驾驶、元宇宙、智能生成等广泛的应用场景提供支撑。随着应用场景的拓展，视觉感知面临的视觉场景逐渐复杂，数据量也进一步激增。同时，高阶应用对视觉感知的精准度和适应性要求也显著提高。因此，实现针对多样化应用需求和复杂环境的自适应视觉感知技术非常重要也充满挑战。

视觉感知的丰富应用，使其面临的视觉场景呈现高度的复杂性和多样性。例如，自动驾驶的公路场景 [1]、工业场景 [2]、遥感识别的卫星图像 [3]、医疗影像 [4] 等差异十分巨大。即使在同一场景下，视觉场景常存在高度的复杂性，例如场景布局和事物类型、尺寸、材质、角度等都具有高度多样性。不同的任务目标也会进一步增加场景理解的复杂度，例如实现语义分割相比场景分类需要处理更多的细节信息。面对复杂和多样的场景进行鲁棒的视觉感知，需要感知系统对场景有强大的自适应处理能力。此外，为了覆盖复杂和多样的视觉场景，需要针对场景采集大量的视觉数据，这进一步提升了视觉感知的学习难度。

自从二十世纪六十年代开始，传统视觉感知技术通过手工设计的规则来完成对某些视觉特征的识别。例如，通过 Canny 滤波器 [5] 提取物体的边缘特征，利用 SIFT 尺度不变特征变换 [6] 描述局部视觉特征。然而，手工设计的特征难以充分描述复杂的视觉场景，因此仅能用于简单场景的视觉感知。随着 AlexNet[7] 在 2012 年的 ImageNet 分类比赛 [8] 中大幅超越传统手工设计的识别算法，利用基于梯度回传 [9] 的深度学习算法进行视觉场景表征开始受到广泛

研究和应用 [10]。基于深度学习的视觉感知系统的核心要素包括深度神经网络架构和表征学习训练。深度神经网络架构确保模型具有高效表征丰富视觉特征的能力，表征学习通过梯度回传优化目标函数从训练数据中学习到所需的视觉场景表征。

虽然基于深度学习的视觉感知技术能处理较为复杂的视觉场景，但随着应用场景的增多和任务要求的提高，复杂场景的视觉感知依然面临很大挑战。视觉场景的复杂性主要体现在应用场景的多样性和特定应用场景内的复杂性。针对复杂场景的视觉感知有如下主要挑战：

- 复杂场景呈现丰富的尺度多样性，即不同物体的尺寸、位置、角度各异。不同场景中数据形式的多样性也进一步丰富了场景的尺度多样性。面对多变复杂的场景，模型架构需要具备自适应的多尺度特征处理能力，以表征各种数据形式下不同尺度的物体。得益于深度神经网络的级联架构，模型能够从局部到全局逐步处理不同尺度的物体，因而具备一定的多尺度处理能力。然而，现有模型架构更多关注粗粒度层级的多尺度处理，缺乏对场景中复杂且细粒度的尺度信息的自适应调整能力。因而，需要更加细粒度的多尺度网络架构对复杂场景进行自适应的多尺度特征表示。
- 复杂场景中存在着复杂多变的物体间关系，需要模型能够根据任务需求建模大量物体间的长距离、短距离关系。网络架构的感受野能够控制模型处理的视觉范围，因此对于建模场景中物体的关系至关重要。现有方法主要根据在特定任务上的经验手工调整感受野。但由于不同任务和场景依赖的感受野范围相差甚大，手工设计需要耗费大量精力且难以调整到最优的感受野。因此，需要针对不同场景和任务进行自适应的感受野调节以确保高效建模复杂场景的物体间关系。
- 基于深度学习的视觉感知依赖大量数据进行表征学习训练。目前视觉感知算法常用的有监督训练需要针对某一任务对数据进行大量的人工标注。随着视觉采集设备的丰富和应用场景的增多，复杂多样的使用场景产生了大规模和多样化的数据，使得人工标注成本过于昂贵。为了降低标注成本的同时充分学习到大量数据蕴含的丰富信息，需要模型能够在尽可能少的人工标注指导的情况下完成对数据的表征和理解。
- 复杂的场景导致训练数据激增，进而需要更大复杂度的模型来学习丰富的信息。训练数据和模型复杂度的激增，进而大幅地增加模型的训练开销。

仅靠硬件的提速难以满足大幅增加的训练开销，因此需要模型能够以尽可能低的训练成本实现高质量的视觉感知性能。目前多数的神经网络基础模型通常从头开始进行训练，这导致大量的计算量和能源的消耗。为大幅降低模型训练成本，需要让新模型绿色高效且自适应地利用现有模型的知识完成对场景的表征学习。

本文希望通过增强对模型架构的尺度、感受野的自适应能力以及对模型表征学习的训练数据和先验模型的自适应能力，高效地完成对复杂场景的视觉感知学习。具备自适应能力的复杂场景感知算法能够高效地以较低的成本快速部署到新的应用场景中，进而推动计算机视觉技术在更多场景的广泛应用。

1.2 研究现状

本节介绍视觉场景感知的关键技术的研究进展，包括网络架构的多尺度和感受野设计，以及针对复杂场景大规模数据下模型训练的无监督语义分割和自监督学习策略。

网络架构多尺度表征 多尺度特征已经被广泛应用于传统的特征设计 [11, 6] 和深度学习 [12, 13] 的网络架构设计中。近几年，拥有更强的多尺度表达能力的主干网络结构 [7, 14, 12, 15, 16, 17, 18, 19] 在众多的视觉任务中取得了世界一流的性能。以 AlexNet [7] 和 VGGNet [14] 为代表的神经网络通过串联堆积卷积层对输入信息遵循由底层细节到高层语义的处理模式，因而具有基本的多尺度特征表达能力。然而，AlexNet 和 VGGNet 通过堆叠卷积层构建网络，使得其每一层都只有相对固定的尺度处理能力。随后，例如 GoogLeNet [12] 和 InceptionNet [20, 21] 等网络利用并行的不同大小或数量的卷积层来增强网络的多尺度表达能力。不过受限于其较低的参数量利用率，该类型网络的多尺度能力往往受计算量限制。ResNet [15] 提出在神经网络中加入短连接从而实现更深的网络结构。其中，短连接允许不同的层之间的组合，使得网络产生大量等效的多尺度特征。DenseNet [16] 和 DPN [13] 在此基础上构建了更加密集的层间短连接，从而获得更好的层间多尺度表达能力。虽然以上网络架构逐渐展现出更强的多尺度表征能力，但这些方法注重增强层间的多尺度表征，难以进行更加细粒度的高效的自适应多尺度表征。

网络架构的感受野 网络架构的感受野决定网络能捕捉信息的范围，在视觉任务的高效感知中起了关键作用 [22, 23, 24]。[23, 24] 研究了有效的感受野需要与

实际任务和网络架构相匹配。然而由于感受野巨大的可选择范围，手工设计难以精确的设定每层的感受野。为此，例如空洞空间金字塔模块 [25]、八度卷积等 [26] 并行感受野模块 [12, 27] 被提出来实现网络更灵活的感受野表达。进一步地，例如多头注意力 [28]、非局部操作 [29]、图网络模块 [30] 等工作通过建模所有位置的连接以形成理论上的任意感受野 [28, 29, 30]。尽管这些方法探索了感受野的巨大潜力，然而它们或难以满足任务所需的实际感受野，或造成巨大的计算冗余。为不同的任务和不同的网络架构选择有效的感受野仍然十分必要但面临挑战。而人工方法难以高效地针对各种视觉任务为网络设定最优的感受野组合。使用搜索算法自适应的针对任务和网络寻找更优的感受野组合是替代人工设计感受野的方案之一。然而，目前网络结构的搜索空间 [31, 32, 33] 大多只支持几个算子的搜索。但是感受野组合的可搜索范围远超常见的网络结构搜索空间，以至于将网络架构搜索算法直接应用于感受野巨大的搜索空间是不切实际的。例如，传统的基于奖励的搜索算法 [34, 35, 36] 由于对每种结构的模型训练和性能评估成本太高，并不适用于具有巨大搜索空间的网络模型。可微的结构搜索方法 [31, 32, 37] 依赖于共享的大网络来节省训练时间，模型大小的限制使其仅支持一层内的几个算子的搜索。此外，该类方法严重依赖于初始的搜索空间，无法搜索到与初始搜索空间相差甚大的新结构组合。

无监督语义分割 作为最具代表性的视觉感知任务，语义分割 [38, 27, 39] 旨在用类别信息来分类图像像素。受限于这项任务的固有挑战和昂贵的人工数据标注，大多数工作集中于在多样性有限 [1, 40, 41] 和数据规模较小 [42, 43] 条件下的语义分割。例如，PASCAL VOC 数据集只包含约两千张图片，而 BDD100K [41] 数据集只关注道路场景。由于复杂场景带来的巨大的数据规模和隐私问题，人为像素级地甚至是图像级地数据标注十分昂贵。基于现有的预训练表征 [44, 45, 46, 47]，一些无监督语义分割模型使用分割排序 [46]，相互信息最大化 [44]，区域对比学习 [47]，以及几何一致性 [48] 等技术完成在小规模数据下无需人工标注的语义分割任务。然而，这些方法关注小数据规模 [44, 45, 46, 47] 下少量（例如 20 个类别）且简单的类别 [44]（例如天空和地面）。从大规模数据中无监督学习到的丰富表征的优势没有被现有方法探索。大规模数据所面临的挑战，例如巨大的计算成本和如何自适应的高效学习场景的信息等问题也被这些方法忽略。

自监督表征学习 通用基础模型的自监督学习通过使用例如实例识别 [49, 50]

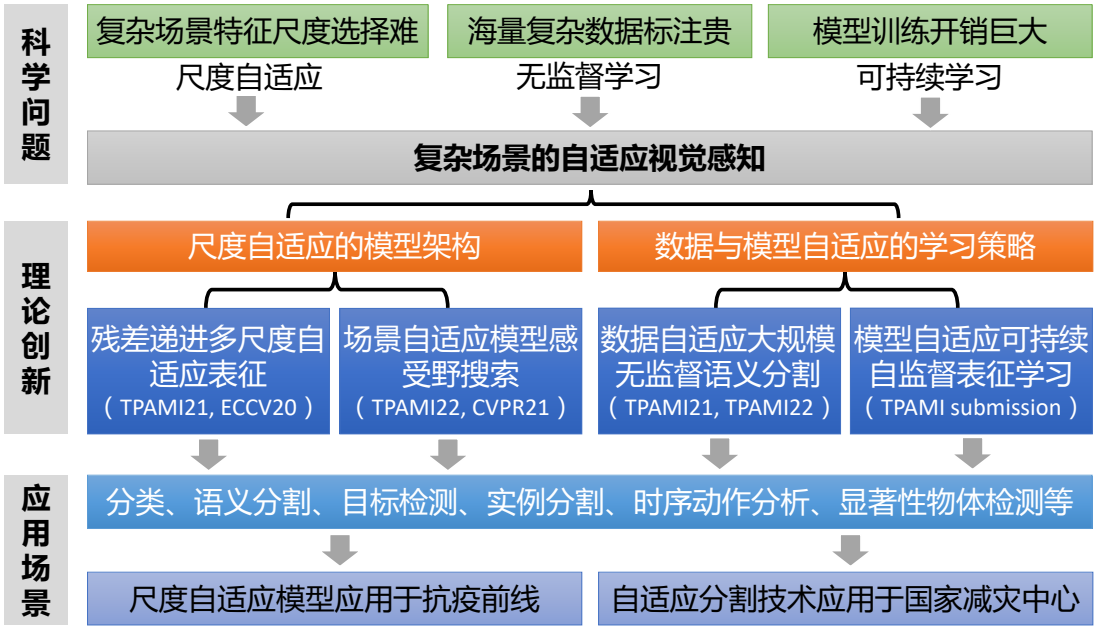


图 1.1 复杂场景自适应视觉感知的研究架构。

和掩码图像建模 [51, 52] 等代理任务进行训练而无需人工标注实现表征预训练。自监督学习在无监督表征学习领域取得了压倒性的成功，自监督预训练的基础模型在分类 [53, 54]、目标检测和分割 [51, 52] 等许多下游任务中具有惊人的性能提升。但由于训练数据的增加和模型复杂度的提升，自监督学习正朝着需要越来越大的训练成本的方向发展。例如，MoCo [49] 需要 200 个迭代轮次，而最近的 MAE [52] 需要 1600 个迭代轮次以释放其潜力。然而大多数研究人员只拥有有限的计算预算，往往无法负担训练大型自监督基础模型的巨额成本。此外，性能稍差的预训练的自监督基础模型在实践中很少使用，导致非最先进的模型很快就会变得无用，浪费了大量的训练资源。因此，构建一个能利用现有基础模型进行自适应学习的可持续的自监督学习框架是十分必要的。

1.3 研究目标和主要贡献

本文旨在研究复杂场景的自适应感知技术，主要研究内容架构如图 1.1所示。复杂场景的自适应感知面临着场景特征尺度选择难、数据海量的复杂人工标注昂贵、模型训练开销巨大等研究挑战。本文将通过增强模型架构和学习策略两方面的自适应能力提升对复杂场景的感知性能。在网络架构层面，本文提

出残差递进尺度自适应表征和场景自适应感受野搜素增强网络架构对复杂多样场景的尺度自适应能力。在学习策略层面，本文提出数据自适应的大规模无监督语义分割和模型自适应的可持续自监督表征学习来实现无需人工标注的情况下对海量数据进行高效低成本的特征学习。以上技术能够在分类、语义分割、目标检测等丰富的视觉任务中得到应用。基于尺度自适应技术的新冠肺炎筛查系统被应用到抗疫前线，产生一定社会效益。自适应图像分割技术也被应用到国家减灾中心的系统中。本文在网络架构多尺度、网络架构感受野、数据表征、模型训练四个方面的自适应感知能力进行研究，具体的研究方案如下：

1）对于众多的计算机视觉任务来说，表达多尺度特征至关重要。为提升对复杂场景的表征能力，先进的主干神经网络不断展现出更强的多尺度特征表达能力，并在大量的任务中取得稳定的性能提升。不过，现存的大多数方法都是通过逐层的方式表达多尺度特征。本文通过设计基于细粒度残差递进模块的多尺度主干网络增强模型，在更细粒度的层面提高对多尺度特征的自适应表征能力。在该模型中，我们为卷积神经网络提出了一种新颖的基础模块，称之为 Res2Net，该模块在一个残差模块中构建类似残差的层级递进的连接结构。Res2Net 可以在更细粒度级别表达多尺度特征，并且可以增加每层网络的特征尺度表征范围和感受野覆盖范围。通过梯度反向传播，模型可以自适应地学习到任务所需的相应尺度的特征。本文提出的 Res2Net 模块可以被加入到例如 ResNet，ResNeXt [18]，DLA [19] 等具有先进性能的模型中。基于 Res2Net 模块的上述模型在 ImageNet 等广泛使用的数据集上展现出相比基准模型更好的性能。进一步的研究表明，在一些如物体检测，类别激活区域映射，显著性物体检测等具有代表性的视觉任务中，Res2Net 具有比其他性能优越的基准模型更好的性能表现。

2）模型的感受野决定其能捕捉场景特征的范围，在众多视觉任务中起着关键作用。大的感受野可以促进长距离关系建模，而小的感受野则可以帮助捕捉局部细节。现有的方法常通过手工设计网络每层感受野的方式构建模型，因而面对新场景需要大量人工对模型进行调整。此外，手工设计的感受野难以达到最优性能。为进一步提升复杂场景对模型感受野的要求，本文考虑通过搜索算法有效地寻找更优的感受野组合，以此来替代手工设计的模式。然而，现有的搜索算法无法处理巨大的感受野搜索空间。为此，本文提出一种高效的全局到局部的感受野搜索策略，可针对任意场景自适应精细化调整模型感受野。本文

的搜索方案利用全局搜索来找到粗略的感受野组合，并利用局部搜索来获得更细化的感受野。全局搜索会发现与人工设计的模式不同的粗略感受野组合。在全局搜索之上，本文提出了一种由期望引导的迭代式局部搜索方案来有效地细化感受野组合。将本文的感受野搜索方案应用到时序动作分割，目标检测，实例分割和语音合成等各种任务的模型中可以显著提升性能。

3）复杂场景中大量的数据为视觉感知学习提供支持的同时也带来了更大的挑战。巨大的数据规模导致人工标注的成本极其高昂，进而使基于人工标注的有监督学习难以为继。借助大型数据集，例如 ImageNet [8] 和 COCO [55]，大规模数据的无监督学习在分类任务方面取得了重大进展。然而，是否能够实现复杂场景下的大规模无监督语义分割仍然是未知的。实现这个任务有两个主要的挑战: （1）我们需要一个评测算法好坏的大规模语义分割基准；（2）我们需要研究新的方法来以无监督的方式同时学习类别和形状表征，并实现语义分割。本文中，我们提出了大规模无监督语义分割这一新问题，并创建了一个基准数据集来跟踪研究进展。基于 ImageNet 数据集，我们提出了 ImageNet-S 数据集，其中包含 120 万张训练图像和约 5 万张用于评测的高质量语义分割标注。我们的基准具有高度的数据多样性和明确的任务目标。此外，本文设计了首个面向大规模场景的无监督语义分割算法，在无需人工标注的情况下实现对数据的自适应表征、聚类和像素级分割。该算法在无任何人工标注数据的情况下，使模型通过自我表征学习从百万量级数据中学习丰富且能够共存的细节和类别语义特征，并将自适应总结出的数百上千个语义类别分配给大规模数据中的每个像素。本文证明完全无需人工标注的自适应大规模视觉感知是完全可行的，为复杂场景的视觉感知提供了一条新的路径。

4）复杂场景导致模型训练成本越来越高，但大多数花费大量成本训练的视觉感知基础模型却没有被充分利用。为此，本文提出一个可持续的自监督表征学习框架，能够自适应地利用现有的预训练基础模型以显著更低的计算成本学习得到更加强大的新模型。要实现一个可持续的自监督表征学习框架，有两个主要挑战：1）以成本友好的方式，基于现有的预训练自监督基础模型，也称为“基”模型，学习一个更强大的“新”自监督基础模型；2）使新模型的训练能与具有各种特性的基模型兼容。为此，本文提出了一个基于掩码重建的目标增强条件化预训练机制。首先，我们提出了区域间关系增强的重建目标来增强基模型提供的特征目标，并鼓励新模型利用不完全输入从基模型中学习语义关系知识。

这种目标增强和预测复杂化有助于新模型超越基模型，因为它们加强了模型额外的区域间关系建模能力来处理不完整的输入。其次，本文提出利用条件适配器自适应地调整新模型的预测以匹配不同基本模型的目标。大量的实验结果表明，本文的可持续学习方案可以加快模型基础模型的学习速度，并进一步提升其视觉感知性能，该方案朝着绿色可持续的视觉感知学习迈出了探索性的一步。

1.4 本文组织结构

本文的组织架构如下：第二章介绍本文的相关工作，包括网络架构的多尺度表征、视觉感知所需的感受野、无监督语义分割以及自监督表征学习技术。第三章介绍了本文提出的残差递进尺度自适应表征主干网络，并在各种任务上证明了该网络架构的自适应能力。第四章介绍了本文提出的场景自适应的感受野搜索算法，并分析了在各种场景和网络架构下不同的感受野需求。第五章介绍了本文提出的数据自适应的大规模无监督语义分割算法，证明了在大规模数据下无需人工标注也可实现高效的视觉感知。第六章介绍了本文提出的模型自适应的可持续自监督学习，展示了该方案能够自适应地利用现有基础模型以明显更低的训练成本实现更好的自监督表征学习。第七章对本文的研究进行总结，并展望了基于现有成果的未来研究方向。

第二章 相关工作介绍

本文将从模型架构的尺度、感受野以及模型训练的数据和表征学习等方面的自适应能力对复杂场景的视觉感知进行研究。本章节将介绍相关工作的研究现状以说明增强视觉感知自适应能力的必要性和挑战。章节2.1介绍网络架构提取视觉任务所需的多尺度特征的相关研究工作。章节2.2进一步介绍网络架构中感受野的表现形式以及不同任务场景中对合适感受野的需求。章节2.3介绍实现数据自适应学习的无监督语义分割相关领域现状。章节2.4介绍在与模型自适应相关的自监督表征学习的研究现状。

2.1 多尺度特征提取

2.1.1 主干网络

主干网络由于其强大的特征表示能力，常被用作各种视觉任务模型的特征提取器 [7, 14, 15, 56]。卷积神经网络由于对输入信息是由细到粗的处理模式，因而具有基本的多尺度特征表达能力。AlexNet [7] 通过串联地堆积卷积层，在物体识别方面相比传统方法有显著的性能突破。不过受限于其网络的深度和卷积核的大小，AlexNet 只有较小的感受野。VGGNet [14] 使用了更小的卷积核并增加了网络的深度。更深的网络可以获得更大感受野，有助于提取大尺度物体的特征。通过堆叠更多的网络层数往往比扩大卷积核能更有效的扩大感受野大小。正因为如此，VGGNet 比 AlexNet 使用了更少的参数却获得了更强的多尺度特征表达能力。不过 AlexNet 和 VGGNet 都是直接堆积卷积层，这会使得其每一层都有一个相对固定的感受野。

NIN [57] 通过将多层感知器作为子网络加入大型网络中，这样做增强了在某一感受野下对局部区域的特征辨别能力。NIN 中 1×1 卷积层也成为了一种优秀的融合特征图信息的方法。GoogLeNet [12] 利用并行的不同卷积核大小的卷积层来增强网络的多尺度表达能力。不过受限于其较低的参数量利用率，该网络的多尺度能力往往受计算量限制。另一方面，ResNet [15] 提出的神经网络中的短连接方式可以减缓梯度消失的现象，从而可以获得一个更深的网络结构。在特征提取过程中，短连接允许组合不同的卷积操作，这使得网络产生大量等

效的多尺度特征图。类似的，DenseNet [16] 中的稠密连接层使得网络可以处理更大尺度范围内的物体。DPN [13] 对于 ResNet 和 DenseNet 的组合使得其拥有 ResNet 的特征复用能力和 DenseNet 的特征提取能力。在最近的 DLA [19] 中，其以类似树的结构来组合网络中的层。树状结构的分层使得网络也能获得更好的逐层的多尺度表达能力。

2.1.2 视觉任务中的多尺度表达能力

卷积神经网络中，多尺度表达能力非常重要。强大的多尺度表达能力在目标检测 [58, 59]、面部分析 [60, 61]、边缘检测 [62]、语义分割 [27]、显著性物体检测 [63, 64] 和骨架检测 [65] 等任务中能有效提升模型的性能。

目标检测 一个高效的卷积神经网络需要能够有效的辨别出一个场景中的各种不同尺度的物体。早期的工作 R-CNN [66] 主要依靠如 VGGNet [14] 等主干网络来提取多尺度特征。He 等人提出的 SPP-Net [67] 方法在主干网络后使用空间金字塔池化来增强多尺度表达能力。后来的 Faster R-CNN [58] 提出的区域候选网络用来生成不同尺度的边界框。基于 Faster R-CNN，FPN [68] 加入特征图金字塔结构来提取单个图像中不同尺度的特征信息。SSD [69] 方法利用不同阶段的特征图来处理不同尺度的视觉信息。

语义分割 语义分割通常需要卷积神经网络处理不同尺度的特征来提取上下文语境信息 [70]。Long [38] 等人提出最早的对于语义分割有效的具有多尺度表达能力的全卷积网络。在 DeepLab 中，Chen 等人 [27, 25] 提出的级联空洞卷积模块可以在保持空间分辨率的同时扩大感受野。最近的 PSPNet [71] 通过金字塔池化的方法汇总来自基于区域特征的全局上下文信息。

显著性物体检测 为了精准定位图像中的显著性物体所在的区域，需要模型提取大尺度上下文信息以确定物体的显著性程度，并且需要小尺度的特征信息来准确定位物体的边界 [72]。更早的方法如 [73] 利用手工设计的全局对比信息 [74] 或者多尺度区域特征信息 [75] 区分显著物体。Li 等人 [76] 提出了最早的能够利用深度多尺度特征来进行显著性物体检测的方法。之后，多语境深度学习 [77] 和多层次卷积特征 [78] 也被提出以提升显著性物体检测的性能。最近，Hou 等人 [79] 提出在不同阶段加入稠密短连接来丰富每层中的多尺度信息以提升显著性物体检测的性能。

2.2 感受野

2.2.1 网络架构中的感受野

感受野的作用已被广泛研究 [22, 23, 24, 25, 26, 80]。虽然理论上的网络感受野可能很大，但 [22] 表明有效感受野仅占据理论感受野的一小部分。[23] 假设一个像素的有用预测信息来自其附近的位置而并非远处的像素，并且，该文建议使用尺度敏感的正则化逐渐抑制远处的像素值。[81] 扩大了感受野以提高图像超分辨率任务的性能。[24] 观察到在图像超分辨率任务上模型深度必须与感受野大小相匹配。[80] 在高斯尺度空间表示的帮助下构建了连续的感受野。并行感受野 [12, 25, 26] 被提出来实现网络层中更灵活的感受野。InceptionNet 系列 [12, 20, 21] 探索使用具有不同感受野的并行非对称卷积来增强模型的表征能力。空洞空间金字塔模块 [27, 25] 证明了并行感受野在语义分割中的有效性。八度卷积 [26] 分解卷积以同时处理两个特征尺度。一些工作对所有位置的连接进行建模以形成理论上的任意感受野 [28, 29, 30]。[28] 提出了多头注意力，以利用注意力机制对每两个像素之间的关系进行建模。类似地，[29] 利用非局部操作来聚合所有位置的特征。[30] 提出了一种基于图的全局推理模块来捕获任意区域之间的关系。尽管这些方法探索了感受野的巨大潜力，但为不同的任务选择有效的感受野仍然是一个悬而未决的问题。

2.2.2 序列性任务的感受野

序列性任务以序列的形式处理数据，例如视频流和音频流。由于序列性任务的序列长度可能有很大的差异，因此需要能覆盖适当范围的有效感受野的模型。其中，时序动作分割是具有代表性的长序列性任务。

时序动作分割 时序动作识别用于分割每个视频帧中的动作，在例如剪辑标记 [82]，视频监控 [83, 84] 和异常检测 [85] 等计算机视觉应用中发挥重要作用。虽然之前工作 [86, 87, 88, 89] 不断刷新包含单个动作的短视频的识别性能，但在未修剪的长视频中密集分割每一帧仍然具有挑战性，因为这些视频包含许多具有不同时长的活动。许多方法已经被提出用于建模时序动作分割的依赖关系。早期的工作 [90, 91, 92] 主要使用滑动窗口 [93, 94, 95] 对外观和动作的变化状态进行建模。因此它们主要关注短距离依赖建模。随后，同时捕获短距离和长距离依赖关系逐渐成为时序动作分割的焦点。

序列性模型 序列性模型以迭代的形式捕获长期依赖关系。Vo 和 Bobick [96]

应用贝叶斯网络来分割由上下文无关语法表示的动作。Tang 等人 [97] 使用隐马尔可夫模型来建模状态和持续动作之间的转换。后来，隐马尔可夫模型被与上下文无关语法 [98]、高斯混合模型 [99] 和循环网络 [100, 101] 相结合，对长序列动作依赖关系进行建模。Cheng 等人 [102] 应用序列记忆器来捕捉从视频中学习到的视觉词汇中的长距离依赖关系。然而，这些序列性模型在并行建模长距离依赖关系时不灵活，并且通常会遭遇信息遗忘 [103, 104] 的影响。

多流结构 一些研究人员 [105, 106, 107, 108] 利用多流模型来建模长距离和短距离的依赖关系。Richard 和 Gall 采用 [105] 动态规划来推理由长度模型、语言模型和动作分类器组成的模型。Singh 等人 [106] 使用双流网络学习短视频块表示，并将这些块传递给双向网络按照顺序预测时序动作分割结果。[107] 中提出了一种包含以自我为中心的线索、空间和时间流的三流结构。Tricornet [108] 利用混合时间卷积和循环网络来捕获局部动作并记住长距离动作依赖关系。CoupledGAN [109] 使用 GAN 模型来利用多模态数据来更好地模拟人类行为的演变。然而，使用多个流捕获长短距离信息会增加计算冗余。

时序卷积网络 最近，时序卷积网络（TCN）通过调整感受野来建模统一结构内不同范围的依赖关系，并且可以并行处理长视频。Lea 等人 [110] 提出了用于时序动作分割的编码器-解码器风格的 TCN 并应用膨胀卷积来扩大感受野，以捕获远距离的时序特征。TDRN [111] 进一步引入了可变形卷积来处理全分辨率残差流和低分辨率池化流。MS-TCN [103, 104] 利用多级膨胀 TCN 和手工设计的膨胀率组合来捕获来自各种时序感受野的信息。然而，感受野的调整仍然依赖于人为设计，导致设计的感受野不符合任务和网络的实际需求。本文提出的高效感受野组合搜索方案可以自动发现更有效的感受野组合，进而改进这些基于 TCN 的方法。

2.2.3 空间性任务的感受野

与主要处理一维序列的序列性任务不同，空间性任务处理具有两个维度的图像，即高度和宽度维度。为了提取场景中各种尺寸物体的特征，模型需要小的感受野来检测小物体，利用大感受野来覆盖大物体或捕获周围的上下文信息。目标检测和实例分割作为主流的基础性视觉任务，依赖合理的网络感受野实现较优性能。

目标检测 目标检测旨在用边界框定位物体并相应地分配类别 [112, 113, 114, 115, 116]。常见的目标检测方法可以分为单阶段 [69, 117, 118, 113, 114] 和

两阶段 [58, 119, 120] 方法。单阶段检测器，例如 SSD [69]、YOLO [117] 和 CornerNet [118]，需要一次推理来端到端定位和分类物体对象，虽然能保证低时延却很难覆盖所有对象。两阶段方法，例如 R-CNN [66]、Faster-RCNN [58] 和 Cascade R-CNN [119]，将目标检测分解为候选区域生成和候选区域中的目标检测，以较慢的推理速度为代价提高了检测质量。尽管存在差异，但目标检测器倾向于增强网络的感受野覆盖范围，从而处理各种尺寸的物体 [69, 58, 119, 120]。SSD [69] 合并来自多个阶段的特征来检测目标。Faster-RCNN [58] 利用特征金字塔网络来聚合具有多个感受野下不同尺度的特征。Cascade R-CNN [119] 和 HTC [120] 执行级联的多阶段特征融合和细化。本文证明，这些方法的正确感受野设置可以进一步提高其目标检测能力。

实例分割 实例分割旨在为实例 [121, 122, 123] 的每个像素分配类别标签。该任务与目标检测十分相关，因为两者都需要定位物体对象。因此，常见的实例分割方法在目标检测器上添加分割分支实现在目标检测器检测到的边界框内分割实例，例如 Mask-RCNN [124] 通过添加目标掩码预测分支来扩展 Faster-RCNN。例如 Cascade R-CNN [119] 和 HTC [120] 等目标检测器中的多感受野形成的多尺度处理能力被自然地被继承到实例分割方法中。一些工作专注于细化分割掩码的边界 [125, 126, 127, 128, 129]，然而它们仍然依赖于具有适当感受野的特征提取器。本文观察到感受野搜索同样有利于实例分割的性能提升。

2.2.4 网络结构搜索

网络结构搜索旨在通过搜索算法找到更优的网络结构。最近许多基于遗传算法 [130] 的搜索方法 [34, 35, 36, 131, 132] 被引入于视觉任务的神经网络架构搜索。一种进化编码方案被应用于遗传卷积神经网络 [36] 来将网络架构编码为二进制字符串。Liu 等人提出了分层表示 [35] 限制搜索空间。Real 等人 [34] 通过一个年龄属性选择操作来正则化遗传进化算法。Sun 等人 [131] 介绍了一种用于高效架构设计的可变长度编码方法。然而，遗传算法需要对每个候选样本进行训练，因此在面对巨大的搜索空间时会消耗过多的计算成本。

可微结构搜索 [31, 133, 134, 135, 136, 37] 通过引入包含具有不同搜索选项的子网络的超网络来节省训练时间。不同搜素选项的重要性由梯度反向传播确定 [9]。然而，这些网络架构搜索方法是为搜索例如卷积、激活函数、批量归一化、短连接等有限数量的操作算子而设计的。因此，由于搜索目标不同，这些方法不能直接用于感受野搜索，例如，它们无法处理巨大的感受野组合搜索空间。

公平 DARTS [133] 解决了由于不同算子之间排他竞争的不公平优势而导致的性能崩溃问题。感受野搜索由于只包含相同的操作算子，因而没有该问题。[134] 通过随机采样一定比例的通道进行操作算子搜索并在短连接中绕过保留的部分，从而降低了超网络的内存成本。这种方法也不适用于感受野搜索，因为短连接不属于感受野搜索空间。[135] 使用掩码机制在进行通道数和特征分辨率搜索时复用特征图。特征图复用的掩码机制也不能应用于由膨胀率表示的感受野搜索，因为不同的膨胀率不属于彼此，不能用不同的掩码进行选择。[136] 缩小了从小数据集搜索到的模型与在大数据集上评估的模型之间的性能差距。从理论上讲，[136] 与感受野搜索空间正交。然而，本文所提出的高效的局部搜索支持直接在大数据集上而不是在小数据集上搜索。本文提出的全局到局部的感受野搜索首先利用全局搜索来处理稀疏采样的巨大搜索空间，然后期望引导的迭代局部搜索将感受野的稀疏搜索空间转移到密集搜索空间以进行精细搜索。

可微搜索的想法 [137] 也被延伸于语义分割 [37, 138] 和其他图像分类之外的任务 [32]。Auto-deeplab [37] 和 DCNAS [138] 专注于搜索语义分割网络中不同阶段特征的分辨率。实验证明本文提出的感受野搜索方案可以基于这些搜索得到的分割网络找到更好的感受野组合。

2.3 无监督语义分割

2.3.1 无监督分割

无监督分割在不使用人工标注的监督情况下实现视觉场景的分割。在深度学习取得最近的进展之前，很多非参数化方法（例如标签转移 [139]，匹配 [140, 141]，距离评测 [142]）和手工设计特征（例如边缘 [143] 和超像素 [144]）已被提出用于分割物体。一些无监督分割方法只关注分割物体而忽略其类别，然而本文研究的大规模无监督语义分割任务同时关注物体的分割和分类。即便如此，无监督分割模型仍然可以为大规模无监督语义分割模型提供先验知识。基于现有的预训练表征 [44, 45, 46, 47]，一些小规模无监督语义分割模型使用分割排序 [46]，相互信息最大化 [44]，区域对比学习 [47]，以及几何一致性 [48] 等技术完成该任务。作为小规模无监督语义分割任务的拓展，大规模无监督语义分割任务不同于前者的是它的大规模数据和类别。然而，以下几个问题限制了小规模无监督语义分割对大规模无监督语义分割任务的适用性：1）现有模型关注于小数据规模 [44, 45, 46, 47] 和少量（例如 20 类左右）且简单的类别 [44]（例

如天空和地面）。因为数据不足，从大规模数据中无监督学习到的丰富表征的优势没有被现有方法探索。大规模数据所面临的挑战（例如巨大的计算成本）也被忽略。2）由于缺少明确的问题定义和标准的评价指标，一些现有方法使用有监督学习的先验知识，例如有监督预训练网络权重 [48]，有监督边缘检测 [46] 和有监督显著性检测 [47, 145, 146]，使得公平评测这些方法变得困难。

2.3.2 语义表征学习

大规模无监督语义分割任务依赖于自监督学习（SSL）提供的语义特征。SSL 方法帮助模型通过代理任务学习语义特征 [147, 148, 149, 150]，例如彩色化 [151, 152, 153]、拼图 [154, 155, 156]、修补 [157]、对抗学习 [158, 159]、上下文预测 [160, 161]、计数 [162]、旋转预测 [163, 156]、跨域预测 [164]、对比学习 [165, 166, 49, 167, 168, 169]、非对比学习 [170, 171, 172] 以及聚类 [173, 174, 175]。本文将介绍与大规模无监督语义分割任务相关的几种 SSL 方法。

基于对比学习的 SSL 作为无监督对比学习方法的核心 [165, 176, 177, 178, 179, 180, 181, 182, 183]，基于对比损失的实例区分方法 [184, 185, 186] 利用图像的不同视角 [168, 169] 或者数据增强 [167, 49] 作为正样本对。进而，该类方法通过使模型产生的负样本对表征互相远离同时拉近正样本对的表征来学习。Wu 等人 [187] 引入了一个记忆库来扩大对比学习可用的负样本数量。MoCo [49] 用一个动量编码器来稳定训练。CMC [168] 提出了多视角的对比学习，而 SimCLR [167] 探讨了不同数据增强的影响。

基于非对比学习的 SSL 一些非对比学习的方法 [171, 188, 189] 最大化图像的不同版本的特征的相似性同时避免使用对比学习常用的负样本对。BYOL [170] 通过预测由动量编码器输出的特征来避免模型输出崩溃为一个常数的平凡解。SimSiam [172] 利用了梯度停止操作以避免模型训练崩溃。然而，因为对比学习和非对比学习的方法都不包含类别信息，所以它们都在类别相关的任务上表现欠佳。例如，这些方法训练的模型对属于同一类别的图像样本不一定输出相似的表征。

基于聚类的 SSL 另一个研究方向是将聚类策略引入无监督学习 [190, 191, 192, 193, 194, 195, 174] 来鼓励一组图像具有接近聚类中心的特征表示。Asano 等人 [173] 提出通过一个优化目标来同时进行聚类和表征学习。Li 等人 [174] 通过期望最大化框架最大化观测数据的对数似然，进而迭代执行聚类和对比学习。SwAV [175] 在对视图进行聚类的同时加强聚类簇之间的平衡性。与其他表征学

习方法相比，聚类策略有助于通过聚类中心实现更强的类别相关表征。

像素级 SSL 一些工作在像素层级而不是图像层级进行自监督学习，以增强向像素级下游任务的转移学习能力 [171, 196, 197]。PixPro [171] 在相邻/其他像素之间应用对比学习，并提出像素传播的一致性机制，以增强表征的空间平滑度。SCRL [196] 随机裁剪局部区域，并使与该位置匹配的其他区域具有一致的空间表征。DenseCL [197] 通过匹配图像的两个视图中最相似的特征向量来选择正样本对。尽管在迁移学习方面表现良好，但这些方法忽略了大规模无监督语义分割任务所需的实例级别类别相关表征能力。

2.4 自监督表征学习

自监督学习通过使用代理任务进行训练而无需人工标注实现表征预训练，例如，实例分区任务和掩码图像建模任务。例如章节2.3.2中介绍的对比、非对比和聚类等实例区分的学习策略通过拉近一个图像的多个视图中提取的表示来学习强类别相关的表示 [167, 170, 172, 189, 175]。从多视图中提取表征在学习到丰富特征的同时，比有监督训练需要更大的训练成本。掩码图像建模任务通过从未掩码部分重建掩码词符的信息来学习语义，比实例区分任务学习更多的空间语义细节。由于训练过程中需要使模型处理不完整的输入，掩码图像建模任务通常需要比实例分区任务更长的训练轮次才能收敛。[198, 199] 探索结合上述两种代理任务的优势，进一步提高自监督学习性能。最近，[200] 揭示了这两种代理任务本质上都是在学习遮挡不变性特征。本文观察到一个趋势，这些自监督表征学习方法需要越来越大的计算成本来实现最优性能，这阻碍了新的自监督表征学习方法的发展。为了解决这个问题，本文通过从预先训练好的自监督表征模型来以更低的计算开销学习新知识，从而探索可持续的自监督表征学习方案。

基于掩码图像建模的自监督表征学习 掩码图像建模任务中不同的重建目标使模型不同语义空间上进行学习。掩码图像建模任务已经探索了各种重建目标，例如，RGB 图像像素和分词器（tokenizers）。为了使图像变成类似于自然语言处理中的离散化语言 [201]，Beit [51] 使用 DALLE 预训练的分词器 [202] 作为预测目标。CAE [203] 进一步将这种代理任务预测与编码器解耦。MAE 和 SimMIM [204] 表明使用 RGB 图像作为重建目标也可以实现具有竞争力的完全微调性能。MaskFeat [205] 揭示了手工设计的 HOG 特征 [206] 是一种有效的目

标形式。Ge2-AE [207] 和 MFM [208] 发现图像频域可以与 RGB 图像目标互补。PeCo [209] 中的感知编码本有助于模型学习语义信息。iBOT 和 data2vec [210] 使用在线动量网络 [49] 来提供持续更新的预测目标。BootMAE [211] 同时利用了 RGB 图像和在线更新的重建目标。[212] 使用掩码方案增强了从大型教师模型到紧凑学生模型的蒸馏。MVP [213] 以 CLIP 预训练模型 [202] 为目标，引入了从视觉语言预训练中学习到的丰富语义。与这些强调特定重建目标的独特属性的工作不同，本文表明所有自监督预训练模型都可以在本文提出的目标增强机制的帮助下作为良好的基模型。本文的可持续表征学习方法中的适配器和目标增强方案使其对各种基模型目标具有良好的适应性。

自监督知识蒸馏 本文提出的模型自适应的可持续的自监督表征学习可以被视为自监督知识蒸馏的特殊情况，因为它们都从自监督预训练的模型中学习。逆知识蒸馏 [214] 表明在有监督训练的范式下，一个弱教师模型也可以使学生模型受益。ClusterFit[194] 对聚类伪标签进行训练，以减少对代理任务的过拟合。SEED [215] 使用对比损失将大型自监督预训练模型中的知识蒸馏到小型模型中。[216] 使用多层感知机进行特征回归，将大型自监督预训练模型蒸馏为紧凑的学生模型。[217] 用教师模型对实例进行分组，并将实例关系知识传递给学生模型。作为特例，[218] 表明特征蒸馏改善了基于对比学习的自监督预训练模型，但对目前性能最优的基于掩码图像建模的 MAE [52] 模型带来的增益有限。本文提出的模型自适应的可持续的自监督表征学习以一种自监督的方式专注于使新模型优于基模型。实验表明，本文的方法比几种先进的自监督蒸馏方法更有优势。

第三章 残差递进尺度自适应表征

多尺度特性在视觉场景中广泛存在。网络模型架构需要具备自适应的多尺度特征处理能力来适应多变复杂的场景。现有模型架构多关注较为粗粒度的尺寸表征能力，缺乏对不同场景尺度的细粒度自适应调整能力。本章提出一种残差递进的多尺度主干网络模型，在细粒度的层面提高对多尺度特征的自适应表征能力。基于该主干网络架构的视觉算法在如分类、检测等数十项具有代表性的视觉任务中取得显著性能提升。章节3.1介绍多尺度表征的背景，并简述本文提出的残差递进的多尺度表征方法。章节3.2和章节3.3分别介绍残差递进的多尺度模块和由其构成的主干网络架构。本文在章节3.4验证该尺度自适应网络架构的有效性，并在章节3.5分析它在各种场景和任务中的自适应能力。章节3.6对本章节进行总结。

3.1 多尺度自适应表征简介

如图 3.1所示，视觉场景中的视觉模式都呈现多尺度特性。首先，同一张图片的不同物体会呈现不同的尺寸，例如图中的沙发和杯子的尺度差异较大。其次，一个物体的关键上下文语境信息需要从比它本身的所占面积更大的区域提取。例如，我们需要图中的桌子作为上下文语境信息来区分桌上的黑色小物体是笔筒还是杯子。此外，感知有着不同尺度的物体的全局或局部信息有助于细粒度分类和语义分割等任务的性能提升。因此，在很多的计算机视觉感知任务中，如图像分类 [7]、目标检测 [58]、注意力预测 [219] 等，设计一个可以自适应地针对任务场景表达多尺度特征的网络结构十分重要。

多尺度特征已经被广泛的应用于传统的特征设计 [11, 6] 和深度学习 [12, 13] 中。在众多的视觉任务中，为了使网络获得多尺度表达能力，需要特征提取器提取不同尺度的物体以及其上下文的信息。卷积神经网络通过堆叠卷积层使得网络由粗到细地学习多尺度特征。卷积神经网络的这种多尺度特征提取能力可以很有效的解决众多视觉任务中的问题。如何设计一个更加有效的网络结构是进一步提升卷积神经网络性能的关键。在过去的几年中，部分主干网络结构 [15, 16] 在众多的视觉任务中取得了重大的进展，并且获得了出色的性能。这些

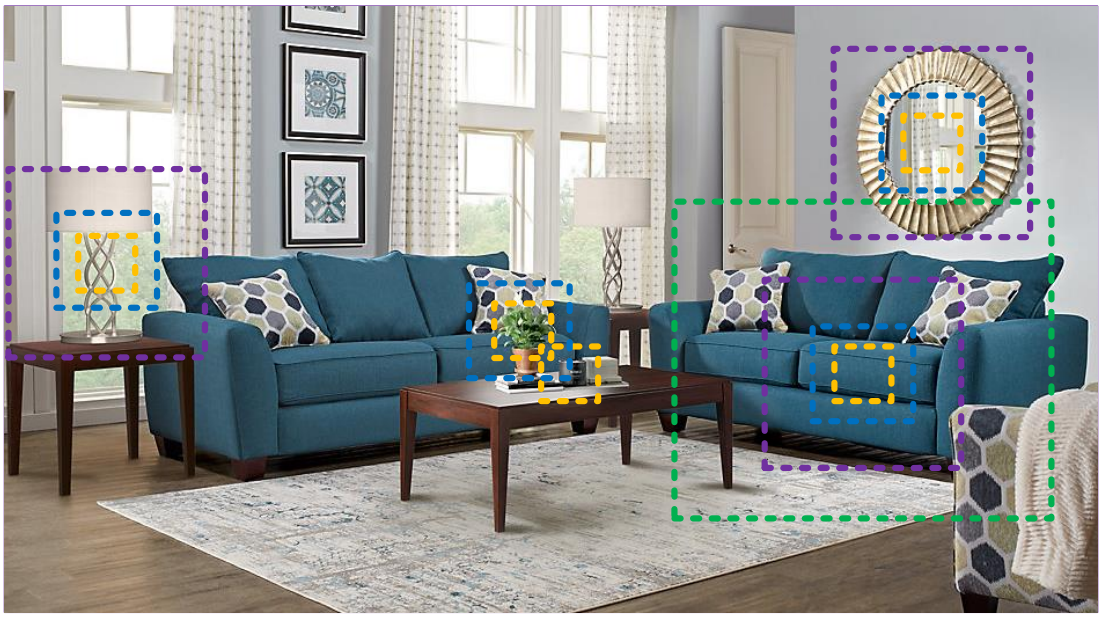


图 3.1 多尺度表征对于如识别物体边界、区域和语义类别等各种视觉任务是十分重要的。即使在最简单的物体识别任务中，对多尺度信息的感知对于理解场景、物体（如图中的沙发、杯子）和它们所处的上下文环境是很重要的（如本文要识别图中黑色小杯子，就需要一个“它在桌子上”的信息）。

卷积神经网络的主干网络结构正在倾向于向更加高效和有效的多尺度表达能力方向发展。

本文提出了一种简单并且有效的自适应多尺度特征处理方法。不同于大多数现有方法注重提升层级的多尺度表达能力，本文提出的方法是在更细粒度的级别提升网络的多尺度表达能力。此外，不同于一些粗粒度的多尺度表征工作[26, 220, 221] 利用具有不同分辨率的特征图来提升多尺度表达能力，本文提出的方法是在更细粒度的尺度上通过多个感受野的组合爆炸效用来提升多尺度表达能力。本文通过将原有的 n 通道 3×3 卷积层替换为一系列有 w 通道的更小的卷积层组（为方便表述，本文令 $n=s\times w$）。如图 3.2，小卷积层组以类似于残差的模式被逐层连接，这样可以增加输出特征能表达各种尺度的数量。具体而言，本文将输入的特征图分为几组，一组卷积层先从一组输入特征图中进行特征提取，得到的该组输出特征图和另一组输入的特征图一起被送到下一组卷积层进行处理。这个过程将一直持续到所有特征图都被处理完毕。最终，所有输出特征图将被拼接在一起并被送到一个 1×1 的卷积进行信息融合。在任意一

条可能的将输入特征图转化为输出特征图的路径上，相应的感受野在经过 3×3 的卷积操作后总会增多，最终会因组合效应表达出丰富的特征尺度。经过训练，网络能够自适应地针对场景和任务所需的尺度进行优化。

本文的 Res2Net 方法扩展出来了一个新的维度，命名为尺度维度（scale, 即 Res2Net 模块中特征图被分成的组数），它将作为网络结构中除深度（depth）[14]、宽度（width）[1]和基数（cardinality）[18] 以外的一个新的关键维度。本文将在章节3.4.4中演示增加尺度维度比增加其他维度更加有效。本文提出的在更细粒度级别扩展网络的多尺度表达潜力的方法和现存的其他层间多尺度方法是没有冲突的。因此本文的 Res2Net 模块可以很容易就被加入到其他现存的网络框架中。大量实验表明，本文的 Res2Net 模块可以很好地提升如 ResNet [15]、ResNeXt [18]、和 DLA [19] 等现有一流网络的表现。

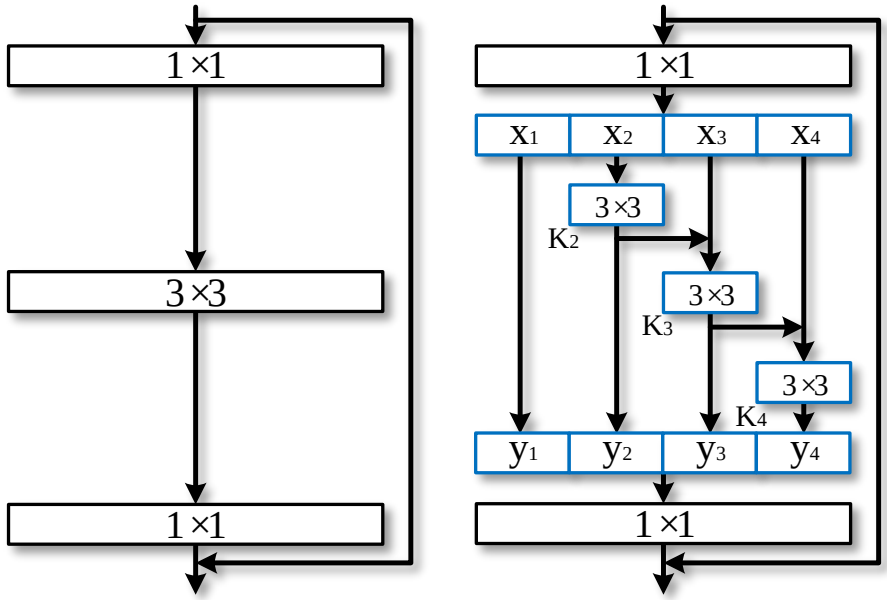
3.2 残差递进多尺度模块

3.2.1 Res2Net 模块设计

如图 3.2(a) 中所示的瓶颈模块是如 ResNet [15]，ResNeXt [18]，和 DLA [19] 等许多现代卷积神经网络的基础构建结构。不同于瓶颈模块利用 3×3 卷积来提取特征，在保持计算量大致不变的情况下，本文旨在设计有更强的多尺度特征提取能力的网络结构。具体而言，本文将一个 3×3 的卷积层替换成一组的卷积层，并且通过类似层次化残差连接的方式将不同的卷积层连接起来。因为本文提出的这个模块在单个残差块中有类似残差的连接，所以本文将其命名为 Res2Net（**Res**idual-like connections in **Res**idual **Net**）。

图 3.2展示了本文提出的 Res2Net 模块和瓶颈模块的差别。在经过 1×1 的卷积层，本文将特征图分为 s 个子集，用 $\mathbf{x}_i$ 表示，其中 $i\in\{1,2,...,s\}$。除了子集的通道数是原来的 $1/s$ 外，每个特征图子集 $\mathbf{x}_i$ 都有着和原始特征图集相同的空间尺寸。除了 $\mathbf{x}_1$，每个特征图子集 $\mathbf{x}_i$ 都有其对应的 3×3 卷积层，用 $\mathbf{K}_i()$ 表示。本文定义 $\mathbf{K}_i()$ 的输出为 $\mathbf{y}_i$。特征图子集 $\mathbf{x}_i$ 和 $\mathbf{K}_{i-1}()$ 相加后被一同送入 $\mathbf{K}_i()$ 进行处理。为了减少参数并增加 s 的数量，本文省略了对 $\mathbf{x}_1$ 所要进行的 3×3 卷积处理，因此 $\mathbf{y}_i$ 可以被表示为：

[1]宽度表示一个层中的通道数 [222]。



(a) 瓶颈模块 (b) Res2Net 模块

图 3.2 瓶颈模块和本文提出的 Res2Net 模块（尺度维度 $s=4$）的对比。

$$\mathbf{y}_i = \begin{cases} \mathbf{x}_i & i = 1; \\ \mathbf{K}_i(\mathbf{x}_i) & i = 2; \\ \mathbf{K}_i(\mathbf{x}_i + \mathbf{y}_{i-1}) & 2 < i \leqslant s. \end{cases} \tag{3.1}$$

每一个 3×3 卷积层 $\mathbf{K}_i()$ 都有可能从它之前所有特征图子集接收信息，即接收 $\{\mathbf{x}_j, j \leq i\}$ 的特征信息。每一次特征图子集 $\mathbf{x}_j$ 通过一个 3×3 的卷积核之后，输出结果就可以有一个比 $\mathbf{x}_j$ 更大的感受野。因为组合爆炸效应，Res2Net 模块的输出包含了不同数量的不同大小以及不同尺度的感受野的组合。

在本文的 Res2Net 模块中，特征子集是以多尺度的方式处理的，这有利于提取局部和全局的信息。为了让不同尺度的信息融合得更好，本文将输出的特征子集在通道维度并联在一起然后通过一个 1×1 的卷积层进行信息融合。这种分组、合并的策略使得卷积层能够更有效地处理特征图。为了减少参数，本文忽略了第一个分组的卷积层，这也是一种特征复用的形式。本文使用参数 s 控制尺度维度。更大的 s 将能更好的学习到更丰富的不同尺寸的感受野，但是这样做将会增加计算量和内存的消耗。

3.2.2 集成先进模块

近几年，大量的神经网络模型被提出，如 Xie 等人 [18] 引入的基数维度还有 Hu 等人 [223] 提出的 SE 模块。Res2Net 模块引入的尺度维度和这些方法是可以共存的。如图 3.3，本文可以很容易地将基数维度 [18] 和 SE 模块 [223] 集成到本文的 Res2Net 模块中。为保证结构简洁，若未额外说明，默认的 Res2Net 不包含基数维度和 SE 模块。

基数维度 基数（cardinality）维度表示一个卷积层内的组的数量 [18]。这个维度将网络的卷积层从单分支变成了多分支，从而提升了模型的表达能力。本文将模块中 3×3 的卷积层替换为一组 3×3 的卷积组，其中用 c 表示组的数量。本文在章节3.4.2和章节3.4.4 中对比了不同的的尺度维度和基数维度对于网络性能的影响。

SE 模块 SE 模块通过显式地建立通道之间的联系来自适应地调整各通道之间的特征响应 [223]。类似于 [223]，本文在 Res2Net 模块的残差连接前面加入了 SE 模块。如章节3.4.2和章节3.4.3所示，SE 模块可以提升 Res2Net 模块的性能。

3.3 残差递进多尺度网络

因为本文的 Res2Net 模块对于网络的基础结构没有严格的要求，并且 Res2Net 模块的这种多尺度特征能力和其他的层级的特征聚合模型可以共存，所以它可以很容易地被集成在如 ResNet [15]、ResNeXt [18]、DLA [19] 和 Big-LittleNet [220] 等一流的网络结构中。上述模型集成本文的模块之后分别对应 Res2Net、Res2NeXt、Res2Net-DLA 和 bLRes2Net-50。

本文提出的尺度维度和之前已经被提出的基数维度 [18] 、宽度维度 [15] 都是可以共存的。因此在固定尺度维度之后，本文将调整基数维度和宽度维度来保证其复杂度和原始模型类似。对于 ImageNet-1K [224] 数据集，本文主要使用了 ResNet-50 [15]、ResNeXt-50 [18]、DLA-60 [19] 和 bLResNet-50 [220] 作为基准模型。本文模型的参数量和其基准模型大致相同。对于一个 50 层的网络来说，其参数量为 $25M$，一张 224×224 图片的每秒浮点运算次数（Floating-point Operations Per Second，FLOPs）在 $4.2G$ 左右。对于 CIFAR [225] 数据集，本文主要使用了 ResNeXt-29 (8c×64w) [18] 作为基准模型。对于模型复杂度的评估和讨论将在章节3.4.4给出。

Res2Net 实现更强的表征 为了进一步探索 Res2Net 的多尺度表征能力，本文

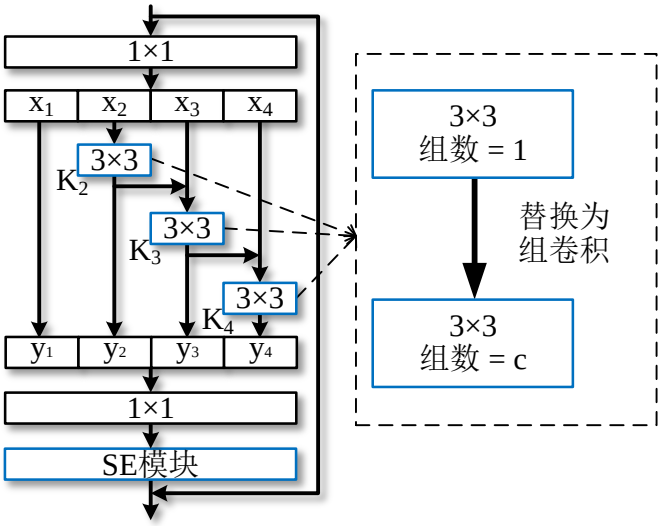


图 3.3 Res2Net 模块可以与基数维度 [18]（用组卷积替换原来的标准卷积）和 SE [223] 模块一起集成在模型中。

遵循 ResNet v1d [226] 的策略来改进 Res2Net，并且使用 CutMix [227] 数据增强技术训练模型。如表 3.1所示，修改后的 Res2Net，即 Res2Net v1b，大幅提升了在 ImageNet-1K 上的分类能力。Res2Net v1b 也进一步提升了在下游任务上的性能。

3.4 实验与分析

3.4.1 实现细节

本文用 Pytorch 框架实现了本文提出的模型。为了对比的公平，本文用 ResNet [15]、ResNeXt [18]、DLA [19] 和 bLResNet-50 [220] 等模型的 Pytorch 实现，并且只将其原始的瓶颈模块替换为 Res2Net 模块做对比。与之前工作类似，在 ImageNet-1K 数据集 [224] 上，本文使用从随机调整过大小的图片中随机剪切出来 224×224 的图片进行训练。本文使用了和 [15, 20] 相同的训练策略。并且和 [15] 类似，本文使用 SGD 优化器，权重衰减为 0.0001，动量为 0.9，使用批大小为 256 在 4 块 Titan Xp 上进行训练。初始的学习率设置为 0.1，并且每 30 个迭代轮次将学习率下降为原来的 0.1 倍。

对于 ImageNet-1K 数据集，所有模型（包括基准模型和 Res2Net 模型）都将在对齐的设置下进行 100 个迭代轮次的训练。对于测试集，本文使用了和 [15]

表 3.1 ImageNet-1K 数据集上 Top-1 和 Top-5 的错误率。

	Top-1 错误率 (%)	Top-5 错误率 (%)
ResNet-50 [15]	23.85	7.13
Res2Net-50	**22.01**	**6.15**
InceptionV3 [20]	22.55	6.44
Res2Net-50-299	**21.41**	**5.88**
ResNeXt-50 [18]	22.61	6.50
Res2NeXt-50	**21.76**	**6.09**
DLA-60 [19]	23.32	6.60
Res2Net-DLA-60	**21.53**	**5.80**
DLA-X-60 [19]	22.19	6.13
Res2NeXt-DLA-60	**21.55**	**5.86**
SENet-50 [223]	23.24	6.69
SE-Res2Net-50	**21.56**	**5.94**
bLResNet-50 [220]	22.41	-
bLRes2Net-50	**21.68**	6.00
Res2Net-v1b-50	19.73	4.96
Res2Net-v1b-101	18.77	4.64
Res2Net-200-SSLD [228]	**14.87**	-

相同的裁剪策略。本文基于 ResNeXt-29 [18] 的实现完成 CIFAR 数据集上的实验。对其他所有的任务，本文使用了基准模型的实现，并且只将瓶颈模块替换为 Res2Net 模块。

3.4.2 ImageNet-1K 分类实验

本文实验使用的数据集 ImageNet-1K [224] 包含了有 1000 种分类标注的 128 万张训练集图片和 5 万张验证集图片。本节主要使用约为 50 层的 Res2Net 系列模型来和一流模型进行性能对比。此外，本文也在 CIFAR 数据集上进行了更多的消融实验。

模型性能 表 3.1展示出了在 ImageNet-1K 数据集上的 Top-1 错误率和 top-5 错误率。简单起见，表 3.1中的所有 Res2Net 模型的尺度维度均为 4（$s = 4$）。在 Top-1 错误率上，本文的 Res2Net-50 相较于 ResNet-50 有 1.84% 的降低。Res2NeXt-50 的 Top-1 错误率也比 ResNeXt-50 降低了 0.85%。并且 Res2Net-

表 3.2 更深的网络在 ImageNet-1K 数据集上的 Top-1 和 Top-5 错误率（%）。

网络架构	Top-1 错误率	Top-5 错误率
DenseNet-161 [16]	22.35	6.20
ResNet-101 [15]	22.63	6.44
Res2Net-101	**20.81**	**5.57**

DLA-60 的 Top-1 错误率比 DLA-60 降低了 1.27%。Res2NeXt-DLA-60 的 Top-1 错误率比 DLA-X-60 降低了 0.64%。SE-Res2Net-50 的 Top-1 错误率比 SENet-50 降低了 1.68%。bLRes2Net-50 的 Top-1 错误率比 bLResNet-50 降低了 0.73%。即使对于 bLResNet 这种利用不同粗粒度尺度特征的网络来说，本文的 Res2Net 模块依然在更细粒度的水平上增强了 bLResNet 的多尺度表达能力。注意，本文使用的 ResNet [15]、ResNeXt [18]、SE-Net [223]、bLResNet [220]、和 DLA [19] 都是现在性能一流的网络。相比较于本身就很优秀的基础网络架构，集成了 Res2Net 模块的网络依然可以获得性能的提升。

本文也把模型和利用不同大小卷积核并行的 InceptionV3 [20] 进行了比较。公平起见，本文使用了 ResNet-50 [15] 作为基准模型，使用和 InceptionV3 模型一样的 299×299 尺寸的图像进行训练。本文的 Res2Net-50-299 在 Top-1 错误率上比 InceptionV3 降低了 1.14%。因此可以得出结论，本文的层次化的残差递进连接比 InceptionV3 的并行卷积能够更有效地处理多尺度信息。相比于 InceptionV3 的卷积组合模式需要精心设计，本文的 Res2Net 模块能简洁而高效的进行模式组合。

更深的 Res2Net 在视觉任务中，更深的网络往往有更好的表达能力 [15, 18]。为了解模型加深后的表现，本文用 101 层的 Res2Net 和 ResNet 进行物体分类性能的比较。如表 3.2所示，本文的 Res2Net-101 在 Top-1 错误率上比 ResNet-101 低 1.82%。需注意本文的 Res2Net-50 在 Top-1 错误率上也比 ResNet-50 低 1.84%。因此，Res2Net 可以和更深的模型结合，从而拥有更好的表现。同样，本文也对比了 DenseNet [16]。Res2Net-101 相较于官方提供的 DenseNet-161 的 Top-1 错误率降低 1.54%。

尺度维度的作用 为了验证本文提出尺度维度的作用，本节实验和分析了有不同的尺度参数的模型。如表 3.3所示，更大的尺度往往有更好的性能。随着尺度的增加，本文的 Res2Net-50（14w×8s）的 Top-1 错误率比 ResNet-50 低了

表 3.3 不同尺度的 Res2Net-50 在 ImageNet-1K 数据集上的 Top-1 错误率和 Top-5 错误率 (%)。其中 w 是卷积层宽度，s 是尺度数量，参见公式 (3.1)。

网络架构	配置	FLOPs	耗时	Top-1 错误率	Top-5 错误率
ResNet-50	64w	4.2G	149ms	23.85	7.13
Res2Net-50	48w×2s	4.2G	148ms	22.68	6.47
（保持	26w×4s	4.2G	153ms	22.01	6.15
复杂度）	14w×8s	4.2G	172ms	21.86	6.14
Res2Net-50	26w×4s	4.2G	-	22.01	6.15
（增加	26w×6s	6.3G	-	21.42	5.87
复杂度）	26w×8s	8.3G	-	20.80	5.63
Res2Net-50-L	18w×4s	2.9G	106ms	22.92	6.67

1.99%。为了保证复杂度不变，在尺度增加的时候，$\mathbf{K}_i()$ 的宽度随之减少。本文也进一步证明了，同时增加尺度和模型复杂度，能进一步提升性能。Res2Net-50（26w×8s）框架比 ResNet-50 的 Top-1 错误率低了 3.05%。Res2Net-50（18w×4s）的 Top-1 错误率也比 ResNet-50 低了 0.93%，并且前者 FLOPs 只有后者 69%。表 3.3也展示了不同尺度模型的运行时间，其为在 ImageNet-1K 验证集上使用 224×224 尺寸图片的平均耗时。尽管本文需要将特征图分割成 $\{\mathbf{y}_i\}$，并且之后需要依次进行层级连接，但是这些 Res2Net 引入额外的运行时间很小可以基本忽略。因为 GPU 能够同时处理的向量数量有限，所以当 Res2Net 的 $s = 4$ 时，可以使得 GPU 单时钟周期内实现高效的并行运算。

更强表征的 Res2Net 配合更先进的表征学习技术的 Res2Net v1b 大幅提升了在 ImageNet-1K 上的分类能力。Res2Net v1b 也进一步提升了在下游任务上的性能。本文在表 3.5和表 3.8 分别展示了 Res2Net v1b 在目标检测和实例分割任务上的效果。Res2Net 更强的多尺度表征能力也被许多下游任务的后续工作证明，例如向量化道路提取 [3]，目标检测 [229]，弱监督语义分割 [230]，显著性物体检测 [145]，交互式图像分割 [231]，视频识别 [232]，伪装物体检测 [233]，和医疗影像分割 [4, 234, 235]。半监督知识蒸馏方法 [228] 也可以被用于 Res2Net 训练，并在 ImageNet-1K 实现了 85.13% Top-1 正确率。

表 3.4 在 CIFAR-100 数据集上的 Top-1 错误率 (%) 和模型大小。参数 c 表示基数的值，w 表示卷积层宽度。

网络架构	参数量	Top-1 错误率
Wide ResNet [222]	36.5M	20.50
ResNeXt-29 (8c×64w) [18] 基准	34.4M	17.90
ResNeXt-29 (16c×64w) [18]	68.1M	17.31
DenseNet-BC (k = 40) [16]	25.6M	17.18
Res2NeXt-29 (6c×24w×4s)	24.3M	16.98
Res2NeXt-29 (8c×25w×4s)	33.8M	16.93
Res2NeXt-29 (6c×24w×6s)	36.7M	**16.79**
ResNeXt-29 (8c×64w-SE) [223]	35.1M	16.77
Res2NeXt-29 (6c×24w×4s-SE)	26.0M	16.68
Res2NeXt-29 (8c×25w×4s-SE)	34.0M	16.64
Res2NeXt-29 (6c×24w×6s-SE)	36.9M	**16.56**

3.4.3 CIFAR

本文也使用 CIFAR-100 [225] 数据集进行分类任务的实验和消融。CIFAR-100 数据集有 100 个类别，有 5 万张图的训练集和 1 万张图的测试集。本文使用的基准模型是 ResNeXt-29 (8c×64w) [18]。本文将原始网络中的基础模块替换为本文的 Res2Net 模块，保持网络其他的配置不变。表 3.4展示了不同大小的模型在 CIFAR-100 数据集上的 Top-1 错误率。实验结果表明，本文的方法比基准模型和其他方法在参数更少的情况下性能更优。本文的 Res2NeXt-29 (6c×24w×6s) 比基准模型的 Top-1 错误率低 1.11%。Res2NeXt-29 (6c×24w×4s) 的参数量甚至只有 ResNeXt-29, 16c×64w 的 35%。对比 DenseNet-BC ($k = 40$)，本文的方法也以更少参数获得更高的性能。相比较于 Res2NeXt-29 (6c×24w×4s)，Res2NeXt-29 (8c×25w×4s) 因为有更大的宽度和基数，所以获得了更好的性能。这也表明了尺度维度和宽度、基数是可以共存的。本文也将先进的 SE 模块集成在了网络结构中。本文的方法可以使用比基准模型 ResNeXt-29 (8c×64w-SE) 以更少的参数获得更高的性能。

3.4.4 改变尺度维度

类似于 Xie 等人 [18]，本文改变网络的不同维度的数值来测试其性能，这些维度包括公式 (3.1) 的尺度、基数 [18] 和深度 [14]。当本文增加模型的一个维度的时候，会保持其他维度不变。这一系列网络将在上述条件的改变下被训练和

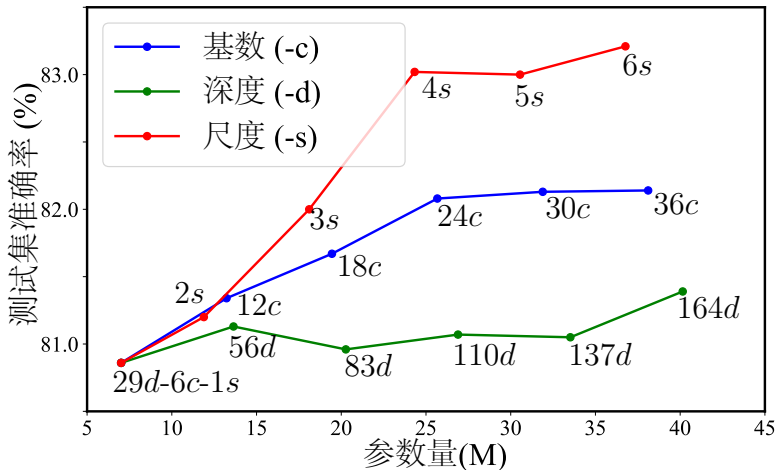


图 3.4 CIFAR-100 数据集上改变了基数（ResNeXt-29）、深度（ResNeXt）和尺度（Res2Net-29）产生的不同参数量的模型性能。

测试。因为 [18] 中已经验证了增加基数比增加宽度更加有效，所以本文只将尺度维度与基数、深度维度进行对比。

图 3.4展示了 CIFAR-100 数据集上的不同大小和参数的模型测试结果。基准模型的深度、基数和尺度分别是 29、6 和 1。实验结果表明了尺度对于模型的性能很重要，这也和章节3.4.2 中在 ImageNet-1K 数据集上的实验结果相吻合。并且增加尺度也比增加其他维度能更快的提升网络性能。如公式 (3.1) 和图 3.2所示，在尺度 $s=2$ 的情况下，本文只是通过增加网络中 1×1 卷积层的参数来增加模型的容量。因此，$s=2$ 时的模型性能会比增加基数略差。对于 $s=3,4$ 来说，本文方法的层次化残差递进连接结构能够产生一系列丰富的等效尺度集合，这有利于获得更好的性能。不过当尺度是 5 和 6 时，模型就只能获得有限的性能提升，这可能是因为 CIFAR 数据集的图象太小（32×32），不需要过于丰富的多尺度信息。

3.5 场景自适应能力分析

本章介绍了 Res2Net 在分类任务以外的各种任务上的性能表现，证明该方法具有针对广泛任务场景的自适应性。

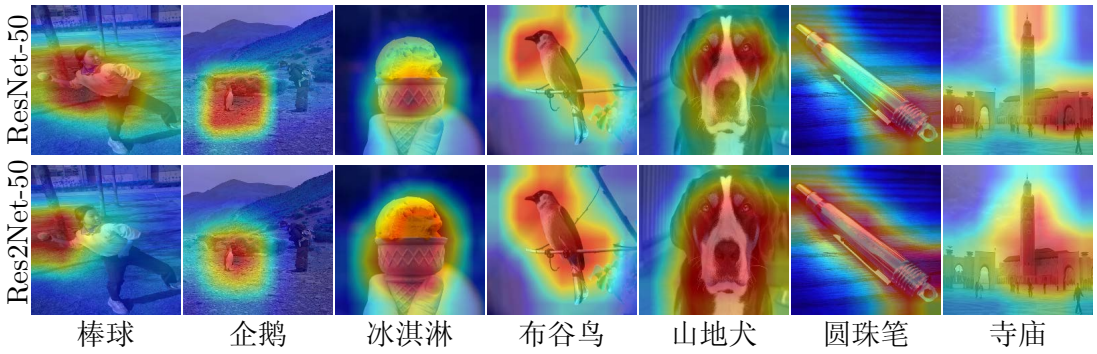


图 3.5 ResNet-50 和 Res2Net-50 的类别激活 [219] 可视化效果对比。

3.5.1 类别激活图

为了更好地理解 Res2Net 的多尺度表达能力，本文使用 Grad-CAM [219] 方法可视化了类别激活图，该方法常用来定位图像分类器的敏感区域。图 3.5的可视化示例中，更强的类别激活响应区域使用了更亮的颜色。相比于 ResNet 的类别激活图，Res2Net 的类别激活图在棒球，企鹅等小物体上的有更集中准确的响应。这两种网络对于冰淇淋等中等大小的物体有着相似的类别激活图。由于有着更好的多尺度表达能力，Res2Net 的类别激活图更倾向于覆盖整个物体，如图中的布谷鸟、山地犬、圆珠笔和寺庙，而 ResNet 的类别激活图则只能覆盖物体的一部分。这种能够在类别激活图中精准定位物体所在区域的能力，对于弱监督语义分割有着可挖掘的潜在价值 [236]。

3.5.2 目标检测

对于目标检测这个任务，本文使用了 Faster R-CNN [58] 作为基准模型，在 PASCAL VOC07 [237] 数据集和 MS COCO [55] 数据集上验证了本文的 Res2Net。本文使用了 ResNet-50 和 Res2Net-50 作为主干网络进行对比，并且公平起见，其他实现细节也都保持一致。表 3.5中展示了目标检测的结果。在 PASCAL VOC07 数据集上，Res2Net-50 模型比其对比的模型平均精度（Average Precision，AP）变优了 2.3%。在 COCO 数据集上，Res2Net-50 模型比其对比的模型 AP 变优了 2.6%，AP@IoU=0.5（AP_{50}）变优了 2.2%。

本文也测试模型在不同尺寸物体上的 AP 和平均召回率（Average Recall，AR），表 3.6为测试结果。Res2Net 模型比其他基准模型取得了很大的性能提升。根据 [55] 的标准，物体按照尺寸不同被分为三类。Res2Net 在小尺寸，中尺寸

表 3.5 在 PASCAL VOC07 和 COCO 数据集上的目标检测结果，使用 AP (%) 和 AP@IoU=0.5 (%) 作为测试标准。Res2Net 和其对比的网络有着相似的模型复杂度。

数据集	主干结构	AP	AP@IoU=0.5
VOC07	ResNet-50	72.1	-
	Res2Net-50	**74.4**	-
COCO	ResNet-50	31.1	51.4
	Res2Net-50	33.7	53.6
	Res2Net-v1b-101	**43.0**	**63.5**

表 3.6 在 COCO 数据集上，模型在不同尺寸物体上的 AP 和 AR 表现。

		物体尺寸			
		小尺寸	中尺寸	大尺寸	所有尺寸
ResNet-50	AP (%)	13.5	35.4	46.2	31.1
Res2Net-50		14.0	38.3	51.1	33.7
性能提升		+0.5	+2.9	+4.9	+2.6
ResNet-50	AR (%)	21.8	48.6	61.6	42.8
Res2Net-50		23.2	51.1	65.3	45.0
性能提升		+1.4	+2.5	+3.7	+2.2

和大尺寸物体的 AP 分别提升了 0.5%、2.9% 和 4.9%，AR 分别提升了 1.4%、2.5% 和 3.7%。因为有着更强的多尺度表达能力，Res2Net 模型可以用更大范围的感受野来覆盖物体，这样提升了其在不同尺寸物体上的表现。

3.5.3 语义分割

语义分割需要卷积神经网络对物体的上下文语境信息有很强的多尺度提取能力。因此本文验证了 Res2Net 在 PASCAL VOC12 数据集上进行语义分割的表现。本文使用的 PASCAL VOC12 数据集 [239] 由包含 10582 张图片的训练集和包含 1449 张图片的测试集组成。本次使用的基准模型是 Deeplab v3+ [238]。除了将主干网络由 ResNet 替换为 Res2Net 之外，其他配置保持和 Deeplab v3+ [238] 相同。在训练和测试时使用的模型步长（strides）都是 16。如表 3.7所示，Res2Net-50 相比基准模型在平均交并比（Mean Intersection over Union，mIoU）上提升了 1.5%。Res2Net-101 相比基准模型在 mIoU 上提升了 1.2%。图 3.6中也将部分语义分割结果进行了可视化。Res2Net 模型倾向于对任何尺寸物体的所有部分都进行覆盖。

表 3.7 在 PASCAL VOC12 数据集上，使用不同尺度的 Res2Net-50 的表现。Res2Net 和其对比模型有着相似的复杂度。

主干结构	配置	mIoU (%)
ResNet-50	64w	77.7
Res2Net-50	48w×2s	78.2
	26w×4s	**79.2**
	18w×6s	79.1
	14w×8s	79.0
ResNet-101	64w	79.0
Res2Net-101	26w×4s	**80.2**

图 3.6 使用 ResNet-101 和 Res2Net-101 作为主干网络的语义分割结果的可视化 [238]。

3.5.4 实例分割

实例分割是目标检测和语义分割的结合。它不仅需要识别出各种尺寸的物体，也要准确的分割出每个物体。正如在章节3.5.2和章节3.5.3中分析的，目标检测和语义分割都需要神经网络有很强的多尺度表达能力。因此，实例分割将能从更优的多尺度表达能力上获益。本文使用了 Mask R-CNN [124] 作为实例分割的算法，本文只将其主干网络由 ResNet-50 替换为 Res2Net-50。在 MS COCO [55] 数据集上的实例分割表现如表 3.8所示。Res2Net-26w×4s 模型比其对比模型 AP 变优了 1.7%，AP_{50} 变优了 2.4%。其也展示了对于不同的尺寸物体的性能提升。对于小尺寸、中尺寸、大尺寸物体，其 AP 提升分别是 0.9%、1.9% 和 2.8%。表 3.8中也展示了 Res2Net 在相同复杂度不同尺度下的性能对比。随着尺度的增加，性能也有上升的趋势。注意 Res2Net-50-26w×4s 相较于 Res2Net-50-48w×2s

表 3.8 不同尺度的 Res2Net-50 在 COCO 数据集上实例分割的性能表现。Res2Net 和其对比模型复杂度近似。

主干网络	设置	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50	64w	33.9	55.2	36.0	14.8	36.0	50.9
Res2Net-50	48w×2s	34.2	55.6	36.3	14.9	36.8	50.9
	26w×4s	35.6	**57.6**	37.6	**15.7**	37.9	**53.7**
	18w×6s	**35.7**	57.5	**38.1**	15.4	**38.1**	53.7
	14w×8s	35.3	57.0	37.5	15.6	37.5	53.4
Res2Net-v1b-101	64w	38.7	61.0	41.4	20.6	42.0	53.2

表 3.9 在不同数据集上的显著性物体检测结果，使用 F-measure 和 MAE 作为评价指标。Res2Net 和其对比模型的复杂度类似。

数据集	主干结构	F-measure↑	MAE ↓
ECSSD	ResNet-50	0.910	0.065
	Res2Net-50	**0.926**	**0.056**
PASCAL-S	ResNet-50	0.823	0.105
	Res2Net-50	**0.841**	**0.099**
HKU-IS	ResNet-50	0.894	0.058
	Res2Net-50	**0.905**	**0.050**
DUT-OMRON	ResNet-50	0.748	0.092
	Res2Net-50	**0.800**	**0.071**

在 AP_L 性能上提升了 2.8 %，而 Res2Net-50-48w×2s 和 ResNet-50 有着相同的 AP_L。本文猜想对于大的物体，模型的性能随着其多尺度表征范围扩大而提升。当尺度表征范围相对较大时，性能提升将不明显。Res2Net 模型能够在训练过程中自适应性地调整尺度范围。当整个图象中的物体已经被合适的感受野覆盖时，模型的性能提升将变得有限。在模型的复杂度不变的情况下，单纯增加模型尺度可能造成每个尺度的通道数减少，这可能会降低模型对于特定尺度特征的处理能力。

3.5.5 显著性物体检测

像是显著性物体检测这种像素层级的视觉任务，也需要卷积神经网络有很强的多尺度表达能力来定位整个物体和其区域边界。本文使用了最新的 DSS [79] 作为本文的基准模型。公平起见，也只将其主干结构替换为 ResNet-50

和 Res2Net-50，同时其他配置参数保持不变。如 [79]，本文使用 MSRA-B 数据集 [240] 进行训练，在 ECSSD [241]、PASCAL-S [242]、HKU-IS [76] 和 DUT-OMRON [243] 数据集上验证结果。本文使用 F 度量（F-measure）和平均绝对误差（MAE）作为检测标准，如表 3.9所示，集成了 Res2Net 的模型相较于其他模型性能均有提升。在 DUT-OMRON 数据集（包含 5168 张图片）上，集成 Res2Net 的模型比集成 ResNet 的模型在 F-measure 上优 5.2%，在 MAE 上优 2.1%。本文的 Res2Net 方法在 DUT-OMRON 数据集上的性能提升最大，因为这个数据集相较于其他数据集，其图像中显著物体的大小变化范围会更大。

3.6 总结

本章节提出了一种可将卷积神经网络的自适应多尺度表达能力提升到更细粒度层次的简洁而高效的模型，命名为 Res2Net。Res2Net 扩展出了一个名叫尺度的维度，这个维度比现存的深度、宽度、基数等维度要更加有效。本文的 Res2Net 模块也可以轻松集成在现有的一流模型上。在 CIFAR-100 和 ImageNet-1K 两个数据集的图像分类任务中，本文的模型也比包括 ResNet、ResNeXt、DLA 等模型在内的其他一流模型有更好的性能。本文在多个场景中的分类、目标检测、显著性物体检测等几个有代表性的视觉任务中证明 Res2Net 的尺度自适应能力。同时，本文将在后续章节证明场景自适应的感受野搜索可以进一步提升 Res2Net 的表征性能。Res2Net 强大的多尺度表达能力也可以应用到第五章节介绍的大规模督语义分割等更具挑战性的任务中。

第四章 场景自适应感受野搜索

复杂场景中存在着复杂多变的物体间关系，需要模型根据任务需求通过调整感受野控制其处理的视觉范围。现有方法主要根据在特定任务上的经验手工调整感受野。但由于不同任务和场景依赖的感受野范围相差甚大，手工设计需要耗费大量精力且难以得到最优的感受野。因此，需要针对不同场景和任务进行自适应的感受野调节以确保模型对复杂场景的高效关系推理。本章节提出一种高效的全局到局部的感受野搜索策略，可针对任意场景自适应精细化调整模型感受野。本章节的搜索方案利用全局搜索来找到粗略的感受野组合，并利用局部搜索来获得更精准的感受野。该感受野搜索算法可插入到各种视觉任务的模型中提升性能。章节4.1对该场景自适用的感受野搜索算法进行简介。章节4.2具体介绍高效的全局到局部的感受野搜索算法。章节4.3验证该算法各部分的有效性。章节4.4在多项任务和多个场景下证明该算法的自适应能力。章节4.5对本章内容进行总结。

4.1 场景自适应感受野搜索简介

由于强大的表征能力，卷积神经网络已经被广泛的应用于视觉识别任务 [124, 27] 以及时序性的感知任务 [104, 244]。卷积网络通过堆叠具有不同感受野的卷积层来处理短距离/长距离特征。用于视觉识别任务的空间卷积网络通过处理局部和全局特征来表示纹理和语义信息。时序卷积网络因其捕捉长期和短期信息的能力而被广泛的用于序列性的任务。网络每层内合适的感受野对于空间卷积网络和时序卷积网络都至关重要，因为大的感受野有助于建模长距离依赖，而小的感受野则有利于捕捉局部细节。最先进的空间卷积网络 [15, 124, 245] 和时序卷积网络 [246, 247] 依赖于人工设计的感受野组合，即人工设定网络每一层的膨胀率或池化大小，以在捕捉长距离依赖和短距离依赖之间达到平衡。这一实现所存在的问题是：是否有其他有效的感受野组合能够与手工设计的模式相媲美，甚至更好？不同场景和任务所需要的感受野组合是否会有所不同？本文提出通过全局到局部搜索通过由粗到细的策略找到更优的感受野组合。

如图 4.1所示，不同于目前网络架构的搜索空间 [31, 32, 33] 仅包括几个不同

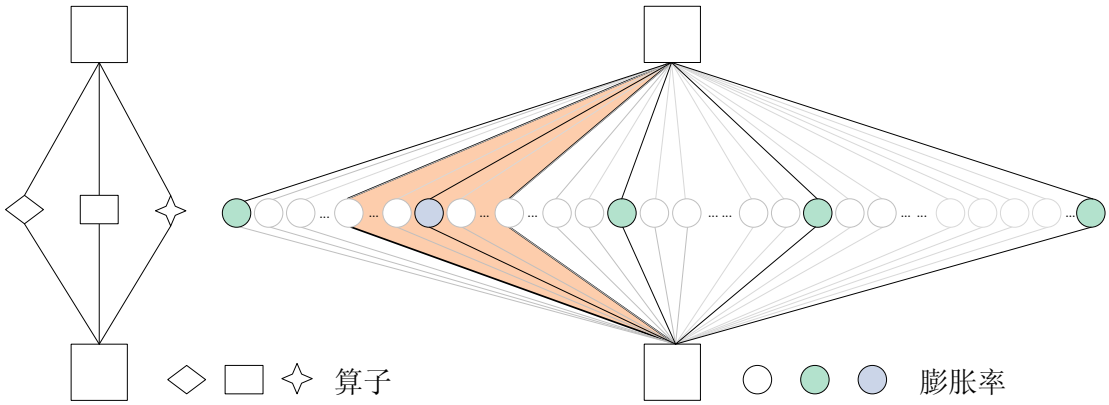


图 4.1 网络架构和感受野组合的搜索空间比较。左：网络架构搜索主要搜索具有不同功能的几个操作。右：感受野组合的搜索空间很大。白色、绿色、蓝色节点和橙色阴影分别表示候选感受野、全局搜索中的稀疏搜索空间、全局搜索结果之一和局部搜索空间。

的算子，感受野组合的可搜索空间可能很大。假设一个网络有 L 个卷积层，每层有 D 个可能的感受野，则共有 D^L 种可行的组合。例如用于长序列动作分割任务的 MS-TCN [103] 有 40 层且每层有 1024 个可能的感受野，共有 1024^{40} 种可能的感受野组合。现有的网络结构搜索算法要么计算成本太高 [36]，要么无法支持大型搜索空间 [37, 31]。因此难以将这些算法直接应用于感受野组合如此巨大的搜索空间。

为了以低成本探索有效的感受野，本文利用基于遗传算法的全局搜索来找到粗略的感受野组合，并利用期望引导的迭代局部搜索（EGI）来获得细化的组合。具体来说，本文遵循许多现有方法 [103, 27, 248, 249] 中的通用设置来使用膨胀率确定每层的感受野。本文提出了一种基于遗传算法的全局搜索方案，以可承受的成本在稀疏采样的搜索空间内找到粗略组合。全局搜索发现了各种新的组合。这些组合实现了比人类设计更好的性能，但具有与后者完全不同的组合模式。基于全局搜索的粗略组合，本文提出了局部搜索来确定细粒度的膨胀率。在局部搜索中，卷积权重共享方案强制使用学习到的不同膨胀率权重来近似概率质量分布。期望引导搜索将离散膨胀率转换为分布，通过计算膨胀率的期望值来实现细粒度的膨胀率搜索。通过迭代搜索过程，局部搜索逐渐以低成本找到更有效的细粒度感受野组合。由本文提出的全局到局部搜索方案增强的模型，即 RF-Next 模型，在许多任务上以令人印象深刻的性能提升超越了人工设计的结构。本节做出了两个主要贡献：

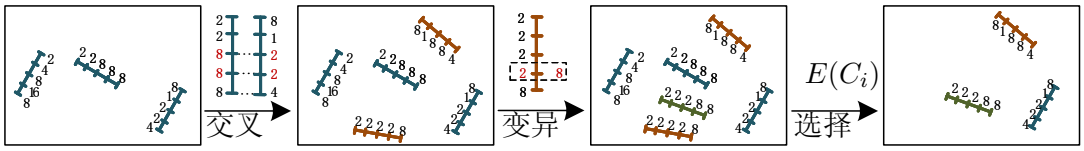


图 4.2 本文提出的基于遗传的全局搜索算法中的一次迭代示意图。步骤 1：随机采样的初始感受野组合；步骤 2：感受野组合片段之间的交叉；步骤 3：随机变异感受野以产生新个体；步骤 4：根据使用早停策略训练的模型的评估性能选择下一次迭代的个体。

- 期望引导的迭代式局部搜索方案能够在密集搜索空间中搜索细粒度的感受野组合。
- 全局到局部搜索针对多样的场景和任务自适应地发现更加有效的感受野组合，其性能优于手工设计的模式。

4.2 全局到局部的感受野搜索算法

本文提出的全局到局部搜索方法的流程有两个组成部分：（i）基于遗传算法的全局搜索算法产生粗略但有效的感受野组合；（ii）期望引导的迭代局部搜索方案进一步局部细化全局搜索的粗略结构。

4.2.1 算法背景及概述

本文的目标是高效地搜索某个场景任务下网络的最佳感受野组合。感受野可以用多种形式表示：膨胀率、卷积核大小、池化大小、步幅和层数。本文的方法最初是为时序动作分割而设计的。因此，本文主要遵循 MS-TCN [103] 使用层中膨胀率的组合来设置感受野，并在搜索过程中优化由膨胀率构成的感受野组合。其他感受野表示只需稍作调整也可以应用于本文提出的全局到局部搜索。虽然本文主要对时序动作分割任务进行了大量实验，但如章节4.2.4中所介绍，本文的感受野搜索方法可以很容易地推广到新任务。

假设一个网络有 L 个卷积层，并且 $D=\{d_1,d_2,...,d_N\}$ 是每一层中可能的膨胀率/感受野。感受野的组合用 $C=\{c_1,...,c_l,...,c_L\}$ 表示，其中 $l\in[1,L]$ 是膨胀卷积层的索引，$c_l\in D$ 是感受野。感受野的可能组合有 $|D|^L$ 种，即当膨胀率从 1 到 1024 时，MS-TCN [103] 中可能的感受野组合数量为 1024^{40}。在如此大的搜索空间中直接搜索有效组合是不切实际的。因此，本文将搜索过程分解为全局搜索和局部搜索，以从粗到细的方式找到感受野组合。

4.2.2 全局搜索

全局搜索旨在巨大的搜索空间中找到可能与手工设计的结构有很大差异的粗略感受野组合，其更多地关注于发现与手工设计相比具有很大多样性的新结构，而不是性能。为了保证新结构的多样性，本文利用随机稀疏采样策略，并应用了具有专门为感受野设计的随机交叉和随机变异操作的遗传搜索算法。为适配感受野搜索，本文提出的基于遗传算法的感受野搜索方案与传统遗传算法有一定差异。

使用逐渐稀疏采样来进行种群初始化 全局搜索的目标是以较低的成本找到粗略的感受野组合。因此，本文通过对层内的膨胀率进行稀疏采样来减少搜索空间。可以采用均匀采样、渐稀疏采样、渐密集采样等多种稀疏离散采样策略来稀疏搜索空间。因为小的感受野有利于提取精确的局部细节，而大的感受野有助于捕捉粗略的长距离依赖。膨胀率逐渐稀疏的采样方案适用于常见的视觉任务。因此，本文将全局搜索中的感受野空间定义为：

$$D_g = \{d_i = k^i, i \in [0, 1, \cdots T]\}, \tag{4.1}$$

其中 k 是搜索空间稀疏度的控制参数，T 决定了最大的感受野。在最大感受野相同的情况下，$|D_g| \ll |D|$，因此搜索空间大大减少。例如，当设置 $k = 2$ 并将最大感受野设置为 MS-TCN 中的 1024 时，搜索空间从 1024^{40} 减少到 11^{40}。感受野组合的种群可以描述为一组候选结构 $P = \{C_i, i \in [1, M]\}$，其中 C_i 是全局搜索空间中的候选结构，M 是总体中的个体数量。

然而，缩小后的感受野组合空间仍然是巨大的，并且使用暴力搜索的计算成本仍然难以负担。本文提出了一种基于遗传算法 [130] 的方法来寻找媲美手工设计或更好的粗略感受野组合。本节将详细介绍本文提出的全局搜索方法中的选择、交叉和变异过程。

根据早停训练的选择 本文需要为每次迭代从感受野组合 P 的总体中选择样本。选择操作根据每个结构 C_i 的估计性能选择要保留在 P 中的个体，用 $E(C_i)$ 表示:

$$E(C_i) = f(V|C_i, \theta_n), \tag{4.2}$$

其中 $f(\cdot)$ 是验证集 V 上特定于任务的评估指标，例如时序动作分割的逐帧准确度，θ_n 是用 n 个迭代轮次训练的模型。全局搜索的主要计算成本在候选结构的性能评估。全局搜索旨在找到有合理的性能的粗略的结构。此外，本文观察到感

算法 1 全局搜索。

Input: 搜索迭代次数 N，训练迭代轮次 n，变异概率 p_m，种群大小 M；
逐渐稀疏的随机采样得到的初始感受野组合种群 P；
for 迭代次数属于 $[1, N]$ **do**
以公式 (4.3) 中计算的概率为交叉操作选择个体；
对两个选中的感受野组合的片段进行交叉操作；
以概率 p_m 随机选择组合，并对其感受野以概率 p_s 在逐渐稀疏的搜索空间内进行变异来生成新个体；
对每个样个体以早停的方式训练 n 个迭代轮次来节省评估开销；
用公式 (4.2) 的评估性能来选择最优的 M 个组合作为新的种群 P；
end for
return P。

受野组合在模型收敛中起着关键作用，即具有良好感受野的模型比配备不良感受野的模型收敛得快得多。为了降低评估成本，本文在训练的模型可以大致显示不同结构的相对性能差距时提前停止候选结构的训练，例如，训练 MS-TCN 5 个迭代轮次可以反映结构性能差异。提前停止训练策略大大降低结构评估成本。

感受野组合段之间的交叉 这一操作生成感受野组合的新样本。种群中的每两个组合在保持局部结构的同时被交换以构成新的组合模式。每个 C_i 将以概率 $p(C_i)$ 被选择用于交叉操作：

$$p(C_i) = \frac{E(C_i)}{\sum_i^M E(C_i)}. \tag{4.3}$$

由于网络表征能力在于感受野组合出的模式，本文希望在交叉期间保留局部感受野组合模式。因此，本文选择交换感受野组合的随机片段，而不是随机交换单个点。具体来说，本文随机选择两个锚点并在两个锚点内交换感受野组合片段以生成新样本。

随机感受野变异 变异操作通过预定义概率 $p_m \in [0,1]$ 选择一个组合，并以预定义概率 $p_s \in [0,1]$ 随机更改所选组合中的每个值来避免陷入局部最优结果。为了降低搜索成本，本文在选择新的感受野值时也应用了逐渐稀疏的采样策略。

全局搜索过程可以概括为算法 1，全局搜索的一次迭代流程如图 4.2所示。通过稀疏搜索空间和全局搜索方法，本文可以找到比人工设计的结构具有相似甚至更好的性能的不同的感受野组合模式。本文进一步提出局部搜索，以在全局搜索结构之上局部地找到更有效的感受野组合。本文在表 4.6b中展示了局部搜索严重依赖初始结构，揭示了全局搜索的重要性。

算法 2 期望引导的迭代式局部搜索。

输入: 搜索迭代次数 N，初始的感受野组合 D；
使用 D 初始化模型的感受野；
for 迭代次数属于 $[1, N]$ **do**
基于 D 对每一层构建 T_l 并以相同的权重初始化 W；
训练模型得到公式 (4.4) 中的 PMF；
通过公式 (4.6) 得到新的感受野；
更新 D；
end for
return 局部搜索得到的结构 D。

4.2.3 期望引导的迭代式局部搜索

局部搜索旨在以低成本在细粒度级别上找到更有效的感受野组合。一种简单的方法是在利用全局搜索得到的初始膨胀率附近对更细粒度的膨胀率进行采样，并应用现有的 DARTS 算法 [32, 31] 来选择合适的结果。然而，即使全局搜索提供了良好的初始结构，细粒度膨胀率的可能范围仍然很大。现有的搜索算法被设计为在每一层中搜索有限的几个算子，因此无法处理具有数百个选择的膨胀率。然而过于稀疏的采样与本节搜索更细粒度的感受野的目标相冲突。此外，DARTS 方法搜索具有不同功能的运算操作 [31]，而对感受野的搜索仅包含一个功能维度。数据集中的不同子集会偏向不同的搜索选项。在感受野这一个功能维度内搜索使我们能够根据所有子集的期望来确定膨胀率，而不是只选择一个多数子集所需的选项。因此，本文提出了一种期望引导的迭代式（Expectation-Guided Iterative，EGI）局部搜索方案来确定全局搜索结构之上的更精细的膨胀率。

假设第 l 层的感受野是 D_l。对于一个数据集，一旦得到 D_l 附近的膨胀率的概率质量分布，就可以通过所有子集所需的膨胀率的加权平均值来获得预期的膨胀率。然而，数据集的膨胀率的概率质量是无法直接得到的。因此，本文利用卷积权重共享方案来强制学习膨胀率的重要性系数来近似概率质量。为了得到膨胀率的近似概率质量函数，本文首先在 $[D_l \pm \Delta D_l]$ 范围内对初始膨胀率 D_l 附近的 S 膨胀率进行均匀采样。该层中可用的膨胀率集是 $T_l = \{d_i | i \in [1, S]\}$，其中 $d_i = D_l - \Delta D_l + (i-1) \cdot 2\Delta D_l/(S-1)$。$\Delta D_l$ 是搜索空间的精细力度的控制参数，来确保得到比全局搜索更密集的采样。

在膨胀率设置为 T_l 的情况下，本文提出了一个由共享卷积权重和具有不同

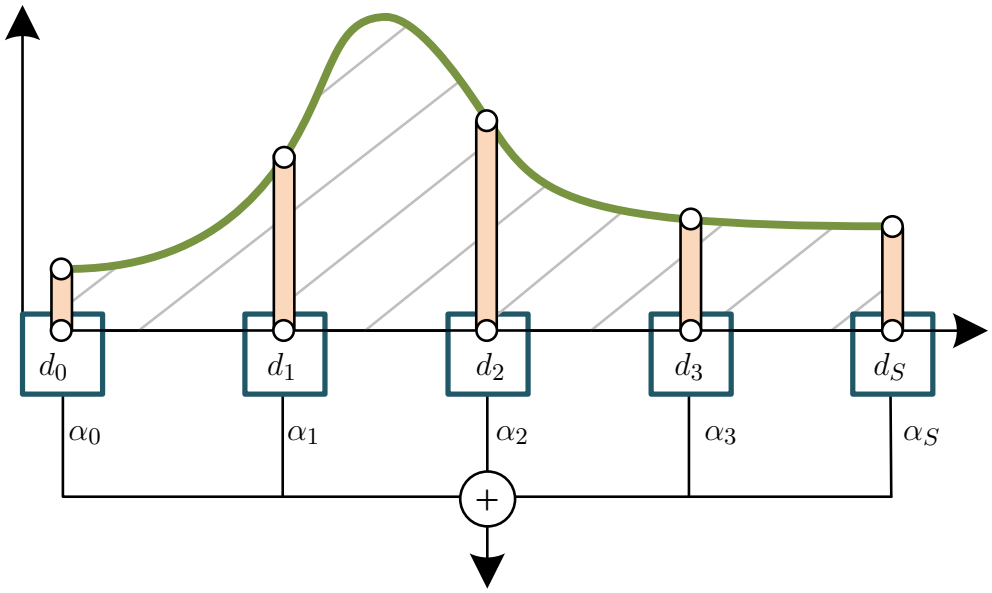


图 4.3 膨胀率的近似概率质量函数由具有共享卷积权重的多膨胀卷积层确定。d_i 是膨胀率，而 α_i 是公式 (4.4) 中的 PMF。

膨胀率的多个分支组成的多膨胀层，如图 4.3 所示。每个分支都有一个独特的系数来确定膨胀率的重要性。在搜索过程中，使用梯度反向传播更新系数以反映数据集的感受野要求。现有的 DARTS 方案 [31, 137] 在每个分支中都有离散的算子权重。相比之下，本文的卷积权重共享策略迫使模型学习感受野的近似概率并加速模型收敛。具体来说，多膨胀卷积层中的膨胀率设置为 T_l。除了共享卷积权重 θ，多膨胀层还包含系数 $W = \{w_1, w_2, ..., w_i, i \in [1, S]\}$ 以确定膨胀率的重要性。θ 和 W 都是可学习的参数，可以通过梯度反向传播进行训练。对于每次迭代，W 中的每个值都使用相同的初始值重新初始化。

由于 W 是无界的，因此不能直接用于确定膨胀率的概率。因此，本文提出了一个归一化函数，通过归一化 w_i 得到膨胀率的近似概率质量函数 $PMF(d_i)$：

$$PMF(d_i) = \alpha_i = \frac{|w_i|}{\sum_i^S |w_i|}. \tag{4.4}$$

给出概率质量函数后，给定输入特征 x，多膨胀卷积层的输出 y 可以写为：

$$y = \sum_i^S \alpha_i \Psi(x, d_i, \theta), \tag{4.5}$$

其中 $\Psi(x, d_i, \theta)$ 为带有共享卷积参数 θ 和膨胀率 d_i 的卷积算子。α_i 通过梯度优化进行更新。一旦本文得到概率质量函数，新搜索到的膨胀率 $D_l^{'}$ 将会由计算期

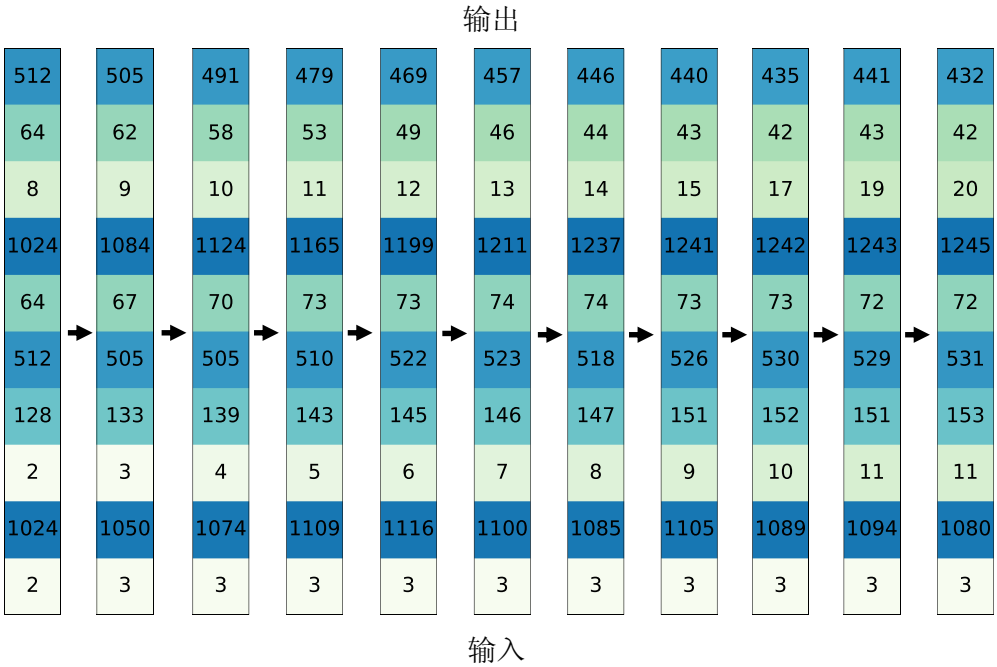


图 4.4 EGI 局部搜索过程中感受野（由膨胀率表示）组合变化的可视化。

望得到：

$$D_l^{'} = \lfloor \sum_{d_i \in T_l} PMF(d_i) \cdot d_i \rfloor. \tag{4.6}$$

为了降低局部搜索过程中的计算成本，本文默认将 T_l 中的膨胀率数量减少到 3，并应用迭代搜索方案根据上次迭代的 $D_l^{'}$ 找到更合适的膨胀率。局部搜索过程可以概括为算法 2。此外，图 4.4 可视化了局部搜索过程中的膨胀率变化。

用于多尺度增强的并行感受野 局部搜索会为每个卷积产生一个膨胀率。然而，本文观察到一些空间任务，例如实例分割和目标检测，需要并行多尺度能力来处理不同大小的对象。本文的期望引导的局部搜索方案可以提供具有不同膨胀率和共享卷积权重的并行多尺度能力。因此，本文通过将膨胀率保持在 T_l 中而不是在最后一次搜索迭代后合并它们来将局部搜索结构扩展到并行版本。并行版本只有与单分支版本相比 $|T_l|$ 个额外参数。并行结构在实例分割和目标检测方面比单个分支结构有显著性能提升。

4.2.4 RF-Next：场景自适应的感受野模型

本文的全局到局部感受野搜索方案适用于使用卷积的各种模型。给定一个初始网络结构，本文将搜索方案应用于卷积核大小大于 1 的卷积。为了便于实现，本文利用膨胀率来表示感受野。全局搜索的目的是找出手工设计之外的感

表 4.1 本文使用的三个时序动作分割数据集的详细信息。

	类别数量	视频数量	帧数	场景
GTEA [92]	11	28	1115	日常活动。
50Salads [250]	17	50	11552	准备沙拉。
BreakFast [251]	48	1712	2097	做早饭。

受野组合，这一步是不是必须的，因为很多模型的感受野已经被手工微调过。局部搜索以较小的额外成本找到合适的细粒度感受野，因此它可以很容易地应用于各种任务的人工设计模型。通过本文的自适应感受野搜索增强，这些配备搜索感受野的模型（即 RF-Next 模型）在例如目标检测、实例分割、语义分割、语音合成和序列建模等许多任务上显示出优势。

4.3 实验与分析

时序动作分割需要较大范围的感受野，适合验证本文提出的全局到局部搜索的有效性。因此，本节主要对时序动作分割任务进行实验。本节介绍了本文提出的全局到局部搜索方案的实现细节，并展示了搜索的感受野组合在时序动作分割任务上优于人工设计的模式。本节还对搜索方案和搜索到的结构的特性进行分析。

4.3.1 实现细节

结构搜索和训练 本文提出的方法是使用 PyTorch [252] 和 Jittor [253] 框架实现的。按照现有的工作 [104, 103]，首先使用 I3D 网络 [88] 从视频中提取特征，然后将其传递给时序动作分割模型以获得时序分割。由于本文提出的全局到局部搜索方案与模型无关，模型评估的训练设置，即训练迭代轮次、优化器、学习率、批大小，与基线方法 [104, 254, 244] 保持相同。在全局搜索阶段，本文设置总迭代次数 $N=100$，公式 (4.1) 中 $k=2$，初始化种群大小 $M=50$，变异概率 p_m=p_s=0.2。公式 (4.1) 中的 T 设置为 10，表示全局搜索空间的最大膨胀率为 1024。本文观察到 5 个迭代轮次的训练可以反映结构性能，因此模型训练 5 个迭代轮次进行评估。在 EGI 局部搜索阶段，ΔD_l 和 S 分别设置为 $0.1D_l$ 和 3。本文在局部搜索期间训练模型 30 个迭代轮次，每 3 个迭代轮次进行一次局部搜索迭代。

表 4.2 本文使用 MS-TCN [103] 作为基线的全局到局部搜索方法的全局和局部搜索阶段的性能。全局搜索找到比基线更好的新感受野组合。局部搜索进一步细化了全局搜索结构以获得更好的性能。

	F@0.1	F@0.25	F@0.5	Edit	Acc
BreakFast					
MS-TCN [103]	52.6	48.1	37.9	61.7	66.3
本文复现	69.1	63.7	50.1	69.9	67.3
全局搜索	72.2	66.0	51.5	71.0	69.2
全局 + 局部搜索	**74.9**	**69.0**	**55.2**	**73.3**	**70.7**
50Salads					
MS-TCN [103]	76.3	74.0	64.5	67.9	80.7
本文复现	78.8	75.3	64.4	71.4	77.8
全局搜索	79.3	76.5	68.1	71.9	81.2
全局 + 局部搜索	**80.3**	**78.0**	**69.8**	**73.4**	**82.2**
GTEA					
MS-TCN [103]	87.5	85.4	74.6	81.4	79.2
本文复现	87.1	83.6	70.4	81.1	75.5
全局搜索	89.1	87.1	74.4	84.2	**78.6**
全局 + 局部搜索	**89.9**	**87.3**	**75.8**	**84.6**	78.5

数据集 依照 [103, 104, 254, 244]，本文在三个流行的时序动作分割数据集上评估本文提出的方法：Breakfast [251]、50Salads [250] 和 GTEA [92]。表 4.1中总结了三个数据集的详细信息。据本文所知，Breakfast 数据集是时序动作分割任务中最大的公共数据集，与其他两个数据集相比，它具有更多的类别和样本。因此，如果没有另外说明，本文主要在 Breakfast 数据集上进行消融实验。按照通用设置 [103, 104, 254, 244]，本文对 Breakfast 和 GTEA 数据集执行 4 折交叉验证，对 50Salads 数据集执行 5 折交叉验证。

评估指标 本文按照以前的工作 [103, 104, 254, 244] 在时间维度上使用逐帧精度（Acc）、分段编辑分数（Edit）[110] 和以阈值 0.1、0.25、0.5（F@0.1、F@0.25、F@0.5）为时序联合交集分段的 F1 分数 [255] 作为评估指标。

4.3.2 性能评估

全局到局部搜索 本文提出的全局到局部搜索旨在找到比人工设计更好的感受野的新组合。本文主要以 MS-TCN [103] 作为基线架构来执行全局到局部搜索。

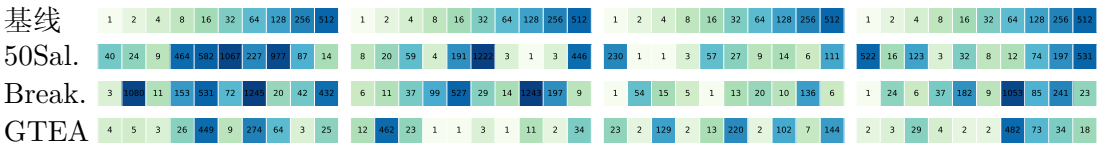


图 4.5 MS-TCN 基线和三个数据集的全局到局部搜索的感受野组合的可视化。每一行代表一个结构的膨胀率组合，MS-TCN 网络包含四个阶段。

表 4.3 与现有的时序动作分割算法结合的效果。本文基于 MS-TCN [103] 执行整个搜索流程。由于计算资源有限，本文只对 MS-TCN++ [104] 和 BCN [254] 进行 EGI 局部搜索，记为 †。SSTDA [244] 使用 MS-TCN [103] 作为主干网络，因此本文直接将搜索到的结构添加到 SSTDA，用 ‡ 表示。

BreakFast	F@0.1	F@0.25	F@0.5	Edit	Acc
ED-TCN [110]	-	-	-	-	43.3
HTK (64) [98]	-	-	-	-	52.0
TCFPN [256]	-	-	-	-	56.3
GRU [100]	-	-	-	-	60.6
GTRM [247]	57.5	54.0	43.3	58.7	65.0
MS-TCN [103]	52.6	48.1	37.9	61.7	66.3
RF-MS-TCN	74.9	69.0	55.2	73.3	70.7
MS-TCN++ [104]	64.1	58.6	45.9	65.6	67.6
RF†-MS-TCN++	72.4	66.8	53.5	70.2	69.6
BCN [254]	68.7	65.5	55.0	66.2	70.4
RF†-BCN	72.5	69.9	60.2	69.0	72.9
SSTDA [244]	75.0	69.1	55.2	73.7	70.2
RF‡-SSTDA	76.3	69.9	54.6	74.5	70.8

在 Breakfast 数据集上测试 MS-TCN 时，本文训练所有模型设置批尺寸为 8 来节省训练时间。表 4.2中展示的复现结果表明更大的批尺寸可以实现更好的性能。表 4.2显示，全局到局部搜索的结构比人工设计的基线实现了相当大的性能改进，即在 F@0.1 指标上，搜索的结构超过了复现的基线 5.8%。全局到局部搜索侧重于感受野组合，因此可以与现有的先进时序动作分割方法结合以提高其性能。如表 4.6c所示，在大规模 BreakFast 数据集上，全局到局部搜索稳定提高 MS-TCN++ [104]、BCN [254] 和 SSTDA [244] 的性能。此外，本文在表 4.4 在两个小规模数据集 50Salads 和 GTEA 上也验证了全局到局部搜索的有效性。

全局搜索 全局搜索通过稀疏搜索空间和本文提出的基于遗传算法的搜索方案降低了计算成本。图 4.6 显示了模型在全局搜索过程中的性能变化。与随机搜索

表 4.4 在 50Salads 和 GTEA 数据集上与现有的时序动作分割方法进行比较。

50Salads	F@0.1	F@0.25	F@0.5	Edit	Acc
Spatial CNN [257]	32.3	27.1	18.9	24.8	54.9
Bi-LSTM [106]	62.6	58.3	47.0	55.6	55.7
Dilated TCN [110]	52.2	47.6	37.4	43.1	59.3
ST-CNN [257]	55.9	49.6	37.1	45.9	59.4
TUnet [258]	59.3	55.6	44.8	50.6	60.6
ED-TCN [110]	68.0	63.9	52.6	59.8	64.7
TResNet [15]	69.2	65.0	54.4	60.5	66.0
TricorNet [108]	70.1	67.2	56.6	62.8	67.5
TRN [111]	70.2	65.4	56.3	63.7	66.9
TDRN [111]	72.9	68.5	57.2	66.0	68.1
MS-TCN++ [104]	80.7	78.5	70.1	74.3	83.7
MS-TCN [103]	76.3	74.0	64.5	67.9	80.7
RF-MS-TCN	**80.3**	**78.0**	**69.8**	**73.4**	**82.2**
BCN [254]	82.3	81.3	74.0	74.3	84.4
RF-BCN	**85.8**	**83.6**	**76.5**	**78.1**	**85.5**
GTEA	F@0.1	F@0.25	F@0.5	Edit	Acc
Spatial CNN [257]	41.8	36.0	25.1	-	54.1
Bi-LSTM [106]	66.5	59.0	43.6	-	55.5
Dilated TCN [110]	58.8	52.2	42.2	-	58.3
ST-CNN [257]	58.7	54.4	41.9	-	60.6
TUnet [258]	67.1	63.7	51.9	60.3	59.9
ED-TCN [110]	72.2	69.3	56.0	-	64.0
TResNet [15]	74.1	69.9	57.6	64.4	65.8
TricorNet [108]	76.0	71.1	59.2	-	64.8
TRN [111]	77.4	71.3	59.1	72.2	67.8
TDRN [111]	79.2	74.4	62.7	74.1	70.1
MS-TCN++ [104]	88.7	87.4	73.5	83.0	78.2
MS-TCN [103]	87.5	85.4	74.6	81.4	79.2
Reproduce	87.1	83.6	70.4	81.1	75.5
RF-MS-TCN	89.9	87.3	75.8	84.6	78.5
BCN [254]	88.5	87.1	77.3	84.4	79.8
RF-BCN	**92.1**	**90.2**	**79.2**	**87.2**	**80.6**

相比，基于遗传算法的全局搜索收敛速度更快。基于遗传搜索算法的模型性能的标准差小于随机搜索，显示了本文提出的搜索方案的稳定性。图 4.5 中性能良好的全局搜索可视化结构证明，全局搜索发现了与人工设计模式完全不同的各种新的感受野组合。表 4.6b 还表明，局部搜索严重依赖全局搜索得到的结构来

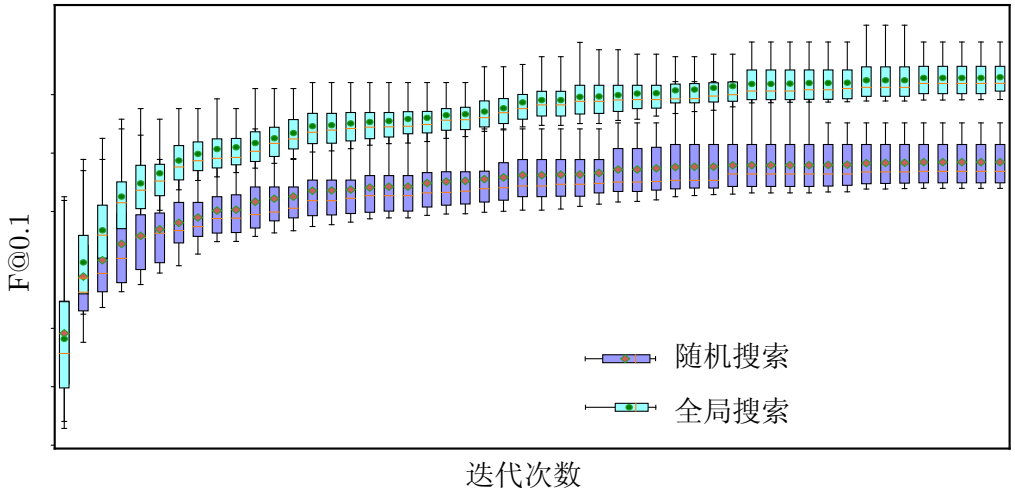


图 4.6 本文提出的基于遗传的搜索和随机搜索在全局搜索阶段的性能比较。

获得更好的性能。

局部搜索 基于全局搜索结构，本文提出的 EGI 局部搜索可以在更精细的搜索空间中对感受野进行微调。正如表 4.6a中所展示的，本文对比了 DARTS [31] 方法和基于全局搜索结构的 EGI 局部搜索法。相较于 DARTS 方法只支持几种搜索选择，本文的 EGI 局部搜索方法可以在稠密空间中迭代的搜索更准确的感受野，从而获得更有利于性能的结构。本文也对比了 EGI 和 DARTS 的几个变种，例如早停策略 [259]、公平 DARTS [133]，结果表明本文的方法相较于它们有明显的优势。如表 4.6c所示，EGI 局部搜索在搜索膨胀率时，对膨胀率的采样数目 S 不敏感。表 4.6b表明，EGI 局部搜索可以提高随机生成、人工设计和全局搜索结构的性能。尽管如此，局部搜索结构的性能还是和初始的结构有关，因为局部搜索的重点是在更精细的局部空间中搜索感受野。在图 4.4中本文可视化展示出了迭代局部搜索的搜索过程。在迭代搜索过程中，各层的膨胀率会逐渐收敛至合适的状态。表 4.6d验证了从系数 w 得到近似概率质量函数 $PMF(d_i)$ 的不同方法。公式 (4.4) 比 sigmoid 函数和 softmax 函数更优，因为它会保持原有的概率分布，而其他两个则会非线性地映射概率分布。

搜索成本 本文展示全局到局部搜索方法的计算开销。当与 MS-TCN 结合时，感受野的组合构成的搜索空间是 1024^{40}。使用现有的搜索方法在如此大的空间上搜索成本是无法担承受的。本文提出的从全局到局部搜索方法将搜索过程分解为全局搜索和局部搜索，由于该搜索方法的主要瓶颈是 GPU 资源，因此本文在表 4.6中报告了本文提出的方法的 GPU 耗时。全局搜索需要更多的计算成本

表 4.5 关于本文提出的 EGI 局部搜索的消融。

BreakFast	F@0.1	F@0.25	F@0.5	Edit	Acc
DARTS [31]	73.4	67.3	53.1	72.7	69.5
+ 早停机制 [259]	73.8	67.6	52.8	72.8	69.3
DARTS+ 早停机制	73.8	67.6	52.8	72.8	69.3
+ Fair DARTS [133]	73.3	67.5	52.9	71.9	69.9
RF-Next	**74.9**	**69.0**	**55.2**	**73.3**	**70.7**

(a) EGI 局部搜索和 DARTS 相关方法的性能比较。

BreakFast	F@0.1	F@0.25	F@0.5	Edit	Acc
随机初始结构	67.7	61.8	48.3	68.4	67.0
随机初始 + 局部搜索结构	73.6	67.8	53.7	72.3	69.9
基线结构 [103]	69.1	63.7	50.1	71.0	69.2
基线结构 + 局部搜索结构	74.1	68.5	55.3	72.3	70.2
全局搜索结构	72.2	66.0	51.8	71.5	69.4
全局搜索 + 局部搜索结构	74.9	69.0	55.2	73.3	70.7

(b) 由不同结构初始化的 EGI 局部搜索的性能。

BreakFast	F@0.1	F@0.25	F@0.5	Edit	Acc
$S=2$	74.8	68.9	55.0	73.4	70.4
$S=3$	74.9	69.0	55.2	73.3	70.7
$S=4$	74.9	68.8	55.1	73.3	70.9

(c) 使用不同分支数目的 EGI 局部搜索的性能。

BreakFast	F@0.1	F@0.25	F@0.5	Edit	Acc
sigmoid	72.7	66.9	52.7	71.8	69.4
softmax	73.2	67.2	52.0	71.6	69.7
公式 (4.4)	**74.9**	**69.0**	**55.2**	**73.3**	**70.7**

(d) EGI 局部搜索中可用的概率质量函数的对比。

来找多个新的相较于手工设计来说性能更好的不同感受野组合。局部搜索需要较小的训练成本在密集的局部空间中微调全局搜索或人工设计的结构的感受野。

4.3.3 搜索得到的感受野组合分析

本节试图探索全局到局部搜索得到的结构中包含的知识。

感受野和数据之间的联系 本文想知道不同数据之间的感受野组合是否不同。

表 4.6 在 RTX 2080Ti GPU 上使用基于 MS-TCN 方法在不同时序动作分割数据集上进行全局和局部搜索的 GPU 小时数。

GPU 小时数	BreakFast	50Salads	GTEA
全局搜索	144h	9h	1h
局部搜索	2.2h	0.15h	0.05h
MS-TCN 训练时间	2.0h	0.14h	0.05h

表 4.7 使用不同数据集的子集 1 搜索到的感受野结构的交叉验证性能（F@0.1）。结构-数据集表示在哪个数据集上搜索得到的结构。

	MS-TCN	结构-50Salads	结构-GTEA	结构-BF
50Salads	67.1	**75.4**	68.8	72.6
GTEA	83.8	82.4	**88.9**	85.6
BF	69.9	75.1	72.5	**76.4**

表 4.8 在 BreakFast 数据集不同子集搜索到的感受野结构的交叉验证性能（F@0.1）。结构-n 表示在子集 n 上搜索得到的结构。

BreakFast	结构-1	结构-2	结构-3	结构-4
子集 1	76.4	76.3	76.2	75.7
子集 2	74.1	75.3	75.1	74.6
子集 3	76.1	76.6	76.1	75.4
子集 4	71.7	72.1	72.0	71.8

因此，本文分别评估了搜索结构在同一数据集的子集和不同数据集上的泛化能力。在 BreakFast 数据集中，本文对一个子集执行全局到局部搜索，然后在其他子集上测试搜索到的结构。表 4.8表明，在不同的子集上几乎没有明显的性能差距，这说明感受野组合在同一个数据集中几乎没有差异。然而，如表 4.7所示，当在不同的数据集间搜索和测试结构时，在不同的数据集上搜索的不同结构有很大的性能差距。本文可以得出结论，不同的数据分布会导致不同的感受野组合。本文在图 4.5中将不同数据集搜索到的结构进行了可视化。搜索的结构同时基于全局搜索和局部搜索。由于全局搜索给每个结构引入了随机性，本文不能公平地比较这些来自不同的数据集的结构。尽管如此，本文还是根据每个结构的搜索的感受野给出了一个粗略的分析。在 Breakfast 数据集和 50Salads 数据集上搜索的结构有往往有更大感受野，而在 GTEA 数据集上搜索的结构则有更小的感受野。表 4.1 表明，视频帧数与感受野大小呈正相关。本文认为数据集的

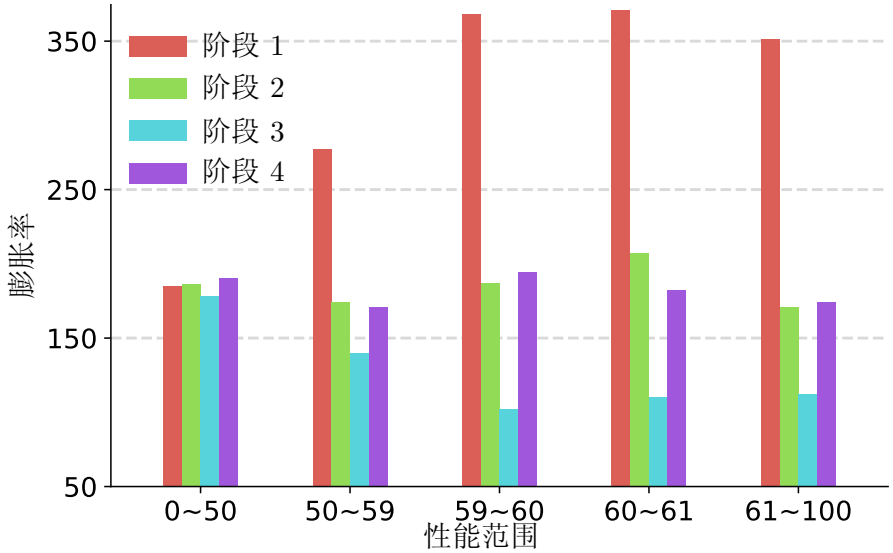


图 4.7 MS-TCN 网络每个阶段的平均膨胀率和全局搜索结构的性能范围的关系。

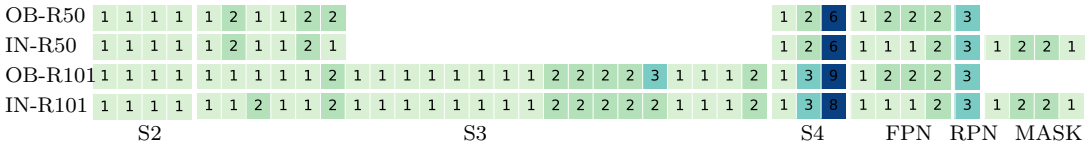


图 4.8 用于目标检测（OB）的 Faster-RCNN [58] 和用于实例分割（IN）的 Mask-RCNN [124] 的局部搜索结构的可视化。S2-S4 表示 ResNet 主干网络的第 2 阶段到第 4 阶段。FPN、RPN 和 MASK 分别表示 Faster-RCNN 和 Mask-RCNN 中的特征金字塔网络、区域提议网络和掩码分割头。

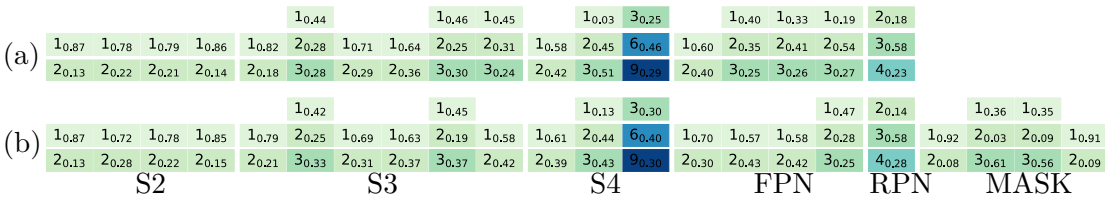


图 4.9 在搜索过程中 S=3 时，基于 ResNet-50 的 Faster-RCNN (a) 和 Mask-RCNN (b) 的并行搜索结构中每个感受野的概率。S2-S4 表示 ResNet 主干的第 2 阶段到第 4 阶段。FPN、RPN 和 MASK 分别表示 Faster-RCNN 和 Mask-RCNN 中的特征金字塔网络、区域提议网络和掩码分割头。

中的视频平均长度会影响结构的感受野。一个拥有更多帧的视频通常需要更大的感受野来捕获长距离的关联。

网络不同阶段的感受野 本文的从全局到局部的搜索是基于 MS-TCN 的。MS-TCN 包含四个阶段，所有阶段在人类手工设计中具有相同的感受野组合。

表 4.9 在 COCO [55] 测试集上使用 Faster-RCNN 作为基线方法在目标检测任务上的局部搜索性能。Local-P 表示如章节4.2.3 中所述的具有并行感受野的局部搜索结构。-R50 和 -R101 分别表示使用 ResNet-50 和 ResNet-101 作为主干网络。S 表示在局部搜索中使用如公式 (4.4) 所示的 S 个分支。

	P	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
Faster-RCNN-R50 [58]		37.8	59.0	41.0	22.1	40.8	46.4
+RF ($S=3$)		39.2	60.9	42.6	22.6	41.8	48.9
+RF ($S=3$)		40.4	62.1	44.0	23.6	43.0	50.5
+RF ($S=2$)		39.1	60.8	42.3	22.7	41.7	48.7
+RF ($S=2$)	✓	40.3	62.1	43.9	23.7	43.0	50.4
Faster-RCNN-R101 [58]		39.7	60.7	43.2	22.5	42.9	49.9
+RF ($S=3$)		41.2	62.8	44.9	23.6	44.1	52.0
+RF ($S=3$)	✓	42.1	63.8	45.8	24.3	45.1	53.3

图 4.5所可视化搜索结构表明，不同的阶段有不同的感受野组合，这与人类手工设计相冲突。本文进一步计算了所有个体中每个阶段平均感受野。各阶段的性能范围和平均膨胀率如图 4.7所示。在高性能结构上，MS-TCN 第一阶段的平均膨胀率往往较大。相比之下，MS-TCN 在第三阶段的平均膨胀率相对较小。本文猜测 MS-TCN 的第一阶段需要大的感受野来获得长程信息进行粗略预测，而接下来的阶段需要小的感受野来局部细化结果。

4.4 场景自适应能力分析

本节展示了 RF-Next 感受野搜索可以自适应地应用于多个场景下的多种网络和多个任务。本文使用前缀 RF 来表示 RF-Next 搜索模型，P 表示并行感受野版本的 RF-Next 搜索模型。具体而言，本节使用提出的搜索方案来为如目标检测、实例分割和语义分割等空间任务寻找适当的感受野。此外，本文还对其他序列任务进行感受野搜索，例如语音合成，P-MNIST 数字分类和复调音乐建模。其中，由于语音合成训练成本高，本文在人工设计的结构上应用局部搜索。本文观察到，适当的感受野显著提高了这些任务的性能。

4.4.1 目标检测

目标检测旨在为每个不同大小的物体分配边界框和类别。本文利用广泛使用的 Faster-RCNN [58] 方法，其中所有卷积的膨胀率都为 1。Faster-RCNN 应

表 4.10 COCO [55] 测试集使用 Mask-RCNN 作为基线方法在实例分割上的局部搜索性能。Local-P 表示章节4.2.3 中所述的具有并行感受野的局部搜索结构。-R50 和 -R101 分别表示使用 ResNet-50 和 ResNet-101 作为主干网络。R50-非局部表示添加非局部模块 [29] 到 ResNet-50 主干的第 4 阶段中的每个残差块。S 表示在局部搜索中使用公式 (4.4) 中的 S 个分支。

	P	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
实例分割掩码							
Mask-RCNN-R50 [124]		34.9	56.4	37.2	18.9	37.5	44.6
+RF ($S=3$)		36.2	58.5	38.5	19.9	38.6	46.8
+RF ($S=3$)	✓	37.1	59.5	39.9	20.5	39.7	48.2
+RF ($S=2$)		36.1	58.2	38.6	19.8	38.4	46.5
+RF ($S=2$)	✓	37.1	59.5	39.7	20.7	39.6	48.3
Mask-RCNN-R101 [124]		36.5	58.3	38.9	19.6	39.2	47.8
+RF ($S=3$)		37.8	60.5	40.4	20.5	40.4	49.7
+RF ($S=3$)	✓	38.5	61.3	41.3	21.0	41.4	50.7
R50-非局部 [29]		36.0	58.2	38.3	19.7	38.6	46.3
+RF ($S=2$)		36.4	58.7	38.9	19.7	38.6	47.2
+RF ($S=2$)	✓	37.3	59.8	39.9	20.3	39.7	48.3
物体检测边界框							
Mask-RCNN-R50 [124]		38.5	59.5	41.8	22.3	41.6	47.4
+RF ($S=3$)		40.0	61.4	43.6	23.2	42.6	49.8
+RF ($S=3$)	✓	41.0	62.5	45.0	24.0	43.8	51.3
+RF ($S=2$)		39.8	61.2	43.4	23.2	42.3	49.4
+RF ($S=2$)	✓	41.0	62.5	44.7	24.1	43.7	51.4
Mask-RCNN-R101 [124]		40.4	61.2	44.1	23.1	43.5	50.8
+RF ($S=3$)		42.0	63.4	45.9	24.2	45.0	53.1
+RF ($S=3$)	✓	42.8	64.1	46.9	24.7	46.1	54.2
R50-非局部 [29]		39.8	61.4	43.3	23.2	42.8	49.1
+RF ($S=2$)		40.4	61.9	44.0	23.2	42.8	50.4
+RF ($S=2$)	✓	41.2	62.9	45.0	23.7	43.7	51.7

用特征金字塔网络来聚合具有多个尺度的特征来处理不同大小的物体。然而，卷积的感受野在该方法被忽略。因此，本文对 Faster-RCNN 中卷积核大于 1 的卷积的膨胀率进行搜索。由于训练成本极大且模型原本初始膨胀率较小，本文只使用高效的局部搜索方案。

如表 4.9所示，RF-Next 模型将使用 ResNet-50 的 Faster-RCNN 的 mAP 提高了 1.4%。RF-ResNet-101 模型的 mAP 也提高了 1.5%。理论上，一个深度更大的网络具有更大范围的感受野。尽管如此，有效的感受野设置在浅模型和深模

型上都有相似的性能提高。如图 4.8所示，本文可视化了基于 RF-ResNet-50/101 的 Faster-RCNN 的搜索膨胀率。浅层需要相对较小的膨胀率，而一些深层的膨胀率较大。有趣的是，与基于 ResNet-50 的模型相比，基于 ResNet-101 的模型在网络的第 4 阶段需要更大的膨胀率。当使用具有并行感受野的 RF-Next 时，基于 ResNet-50 和 ResNet-101 的模型在 mAP 中的性能增益分别为 2.6% 和 2.4%，这说明目标检测任务需要并行的多尺度能力。并行的 RF-Next 中每个感受野的可视化和概率如图 4.9所示。默认情况下，本文使公式 (4.4) 中的膨胀率采样次数为 $S=3$。如表 4.9所示，本文也探索了使用 $S=2$ 进行局部搜索。并行 RF-Next 使用两个/三个分支性能类似，这表明每层都使用两个分支就能提供足够多的多尺度能力。该结果也与表 4.6c中的观察结果相一致。表 4.6c表明，本文所提出的期望引导搜索对采样膨胀率的次数不敏感。在表 4.9中本文分析了其对于不同大小物体的性能增益。基于 ResNet-50 和 ResNet-101 模型，对小、中、大对象的 mAP 增益分别为（1.5%、2.2%、4.1%）和（1.8%、2.2%、3.4%）。随着对象大小的增加，性能增益逐渐增加，这表明 Faster-RCNN 的默认感受野设置不足以处理大型物体。

4.4.2 实例分割

实例分割输出实例分割掩码和类别，其任务目标与目标检测任务类似。为了比较目标检测和实例分割的感受野需求，本文使用了广泛使用的 Faster-RCNN 的扩展算法 Mask-RCNN [124]，与目标检测一样，本文对卷积核大于 1 的卷积应用局部搜索。

本文在表 4.10中给出了搜索到的结构和基线的性能比较。对于 ResNet-50/101 模型来说，使用单分支的 RF-Next 在掩码 mAP 上提升了 1.3%/1.3%，在边界框 mAP 提升了 1.5%/1.6%。并行 RF-Next 进一步提高了性能，其中掩码 mAP 增益为 2.2%/2.0%，边界框 mAP 增益为 2.5%/2.4%。在图 4.8中，本文给出了 Faster-RCNN 和 Mask-RCNN 之间的可视化膨胀率比较。如图 4.9所示，Mask-RCNN 的掩模分割头部的膨胀率概率分布显示，中间两阶段需要不同的感受野，而第一阶段和最后一阶段需要一个较小的感受野。

4.4.3 语义分割

语义分割任务要求为图像的每个像素分配类别标签。感受野对于语义分割这种密集像素级预测至关重要。Deeplab 系列 [27, 25] 利用膨胀率大于 1 的卷积

表 4.11 使用 PASCAL VOC [237] 和 ADE20K 数据集 [262] 进行语义分割的局部搜索性能。Local-P 表示如章节4.2.3 中所述具有并行感受野的局部搜索结构。† 表示将局部搜索应用于网络的所有卷积。-S3 表示网络第三阶段的输出，该位置添加了辅助损失以加速收敛 [25]。

	P	VOC [237]		ADE20K [262]	
		mIoU	mAcc	mIoU	mAcc
DeepLabV3 [25]		76.2	85.7	42.4	53.6
+RF		77.8	87.4	43.2	54.1
+RF	✓	77.9	87.6	43.0	53.7
+RF†		76.3	85.8	-	-
+RF-S3		51.7	64.7	-	-
+RF-S3†		67.7	79.8	-	-

来扩大感受野，这成为语义分割网络 [260, 261] 的默认选择。然而，可能比人类手工设计的更好的感受野尚未被探索。本文使用 DeeplabV3 [25] 网络作为基线网络，使用局部搜索来寻找更高效的感受野。DeeplabV3 在 ResNet 主干网络的第三阶段的输出上应用了一个辅助损失来加速收敛。为了避免辅助损失的影响，本文将局部搜索应用于网络的第四阶段和解码器。

本文利用平均交并比（mIoU）和平均正确率（mAcc）来验证训练好的模型。如表 4.11中所示，RF-Next 在 PASCAL VOC 和 ADE20K 中，mIoU 分别获得了 1.6%、0.8% 的增益。多分支结构实现了与单分支相似的性能。本文认为 DeeplabV3 中膨胀卷积构成的空间金字塔池化结构已经增强了网络的并行多尺度能力。在图 4.10中，搜索到的感受野的可视化表明，在网络的第四阶段需要更大的感受野。如上所述，本文在网络的前三个阶段跳过局部搜索，以避免辅助损失的副作用。在表 4.11中显示，搜索所有卷积可以达到 76.3% 的 mIoU，与人工设计的基线性能相近。与在第三阶段后的搜索相比，前三阶段也进行搜索在网络第三阶段输出的 mIoU 中的性能提高了 15.4%。由于局部搜索依赖于梯度反向传播来寻找感受野，添加辅助损失使局部搜索可以在第三阶段而不是最终输出阶段找到更好的感受野。

4.4.4 语音合成

本节着重于在语音合成中将声学特征转换到语音波形的过程。本文使用 WaveGlow [248] 作为基线方法，这个方法结合了 Glow [263] 的和 WaveNet [249]

图 4.10 DeeplabV3 语义分割任务的第四阶段和解码器的局部搜索感受野可视化。

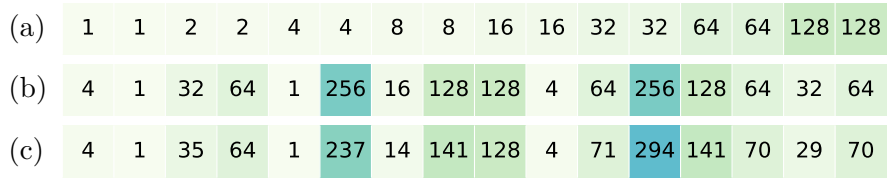


图 4.11 P-MNIST 分类任务中时序卷积网络的基线结构 (a)、全局搜索结构 (b) 和全局到局部搜索结构 (c) 的可视化。

的优点。WaveGlow 网络有 12 层，每层包含 8 层膨胀卷积，采用人工设计的逐渐增大的膨胀率。为了节省计算成本，本文利用局部搜索来基于人工设计的结构寻找更有效的膨胀率。本文在广泛使用的 LJ 语音数据集 [264] 上进行实验。样本首先通过短时的傅里叶变换生成梅尔普图 [265, 266]，之后送入网络进行语音合成。

在表 4.12中，为了验证语音合成的质量，本文使用三种评价指标：梅尔倒谱失真（MCD） [267]，语音质量感知评价（PESQ） [268]，对数似然比（LLR） [269]。MCD 测量了两个语音序列之间的差异，MCD 越小表示合成语音和自然语音越接近。类似地，LLR 测量了两个语音之间的差异。在 MCD 和 LLR 指标上，搜索出来的感受野组合结构比人工设计的感受野表现更好。PESQ 对语音质量进行评估，值越高意味着合成语音质量越好。局部搜索的语音合成结果也有较好的 PESQ 得分，表明更合适的感受野有利于提升语音合成质量。在图 4.12中，本文可视化了感受野和基线方法 WaveGlow [248] 的感受野（膨胀率）。本文观察到，人工设计结构的最大膨胀率比搜索结构大得多，这表明该任务可能不需要太大的感受野。与人类设计的每个阶段具有相同感受野的组合结构不同的是，搜索结构在浅层有较小的感受野在深层有较大的感受野。本文认为语音合成任务需要浅层的局部特征，而较深层负责建模长程依赖关系。

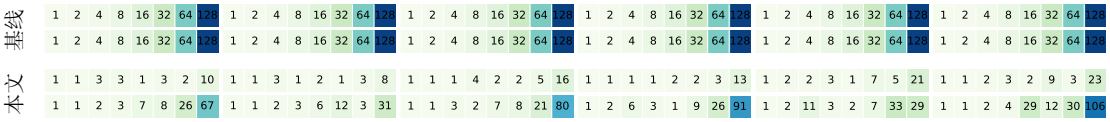


图 4.12 语音合成任务中 WaveGlow [248] 的局部搜索膨胀率的可视化。

表 4.12 WaveGlow [248] 在 LJ 语音数据集 [264] 上基于局部搜索结构的语音合成性能。

	MCD↓ [267]	LLR↓ [269]	PESQ↑ [268]
WaveGlow [248]	5.79	1.29	1.52
RF-WaveGlow	5.59	0.71	1.84

4.4.5 时序卷积网络序列建模

Bai 等人 [270] 验证了时序卷积网络在多个序列建模任务上的性能。本文进一步展示了 RF-Next 模型在复调音乐建模和 P-MNIST 数字分类这两个序列任务上的有效性。

P-MNIST 分类 P-MNIST 分类的目的是对像素顺序被打乱的手写体数字图像进行分类。P-MNIST 数据集 [271, 272] 将 MNIST 数据集 [273] 中的图像随机排列到 784 个长度的序列，用于评测模型长期关系建模能力 [274, 275, 276, 277, 270]。P-MNIST 分类 [270] 采用 8 层的时序卷积网络，其中每一层的卷积核大小为 7，通道数为 25。本文在时序卷积网络上应用全局到局部的搜索来寻找更有效的感受野。在 P-MNIST 中，像素的顺序是随机排列的，本文固定了其在所有实验中的顺序。在全局搜索过程中，本文设置迭代次数 $N = 50$ 和初始种群大小 $M = 25$。每个样本训练轮。在局部搜索过程中，ΔD_l 被设置为 $0.1D_l$，并且该结构被训练 15 轮，结构每 3 轮更新一次。以分类精度作为评价指标。如表 4.13所示，全局搜索的准确率从 97.2% 提高到 97.6%，局部搜索的性能也进一步提高到了 97.8%。图 4.11中的可视化结构显示，搜索到感受野与人工设计的模式非常不同。

复调音乐建模 复调音乐建模的目的是根据已经演奏的音符的来预测随后的音符。复调音乐建模是在广泛使用的 Nottingham 数据集 [278, 279, 280] 上进行的，该数据集包括 1200 首英国和美国的民间曲目。对于复调音乐建模，本文使用了一个包含 4 层的时序卷积网络，其中每一层都有两个卷积，卷积核大小为 5，通道数为 150。在全局搜索中，本文设置迭代次数 $N = 50$，初始种群大小

表 4.13 全局到局部搜索在时序卷积网络模型上的性能 [270]。本文评估了复调音乐建模和 P-MNIST 数字分类任务的性能。

任务	基线	全局搜索	全局 + 局部搜索
P-MNIST 数字分类 (精度 ↑)	97.2	97.6	97.8
复调音乐建模 (NLL↓)	2.97	2.73	2.69

表 4.14 在 COCO 验证数据集上，感受野搜索改进了用于目标检测和实例分割任务的先进注意力/卷积模型。基于 PVT [281, 282] 和 ConvNeXt [283] 的官方实现，PVTv2-B0 和 ConvNeXt-T 分别采用了 Mask-RCNN 检测器和 Cascade Mask-RCNN 检测器。

物体检测	P	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
PVTv2-B0		38.2	60.5	40.7	22.9	40.9	49.6
RF-PVT		38.8	60.9	41.8	23.6	41.2	50.8
RF-PVT	✓	39.1	60.8	42.7	23.3	41.8	51.4
ConvNeXt-T		50.4	69.1	54.8	33.9	54.5	65.1
RF-ConvNeXt		50.6	69.2	54.8	34.1	54.0	65.5
RF-ConvNeXt	✓	50.9	69.5	55.5	34.3	54.6	65.8
实例分割		mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
PVTv2-B0		36.2	57.8	38.6	18.0	38.4	51.9
RF-PVT		36.8	58.4	39.5	18.7	39.0	52.7
RF-PVT	✓	37.1	58.5	40.0	17.8	39.3	53.7
ConvNeXt-T		43.7	66.5	47.3	24.2	47.1	62.1
RF-ConvNeXt		44.0	66.8	47.5	24.8	47.0	62.1
RF-ConvNeXt	✓	44.3	67.3	47.8	24.7	47.4	62.6

$M = 25$，每个样本训练 30 个迭代轮次。对于局部搜索，ΔD_l 被设置为 $0.15D_l$，模型训练 60 个迭代轮次，每 10 轮更新一次结构。在表 4.13中，本文使用负对数似然法 (NLL) 来评测模型。通过全局搜索，NLL 从 2.97 提高到 2.73，局部搜索将性能提高到 2.69。

4.4.6 在先进网络架构上的感受野搜索

本文将感受野搜索方法应用于多种模型结构，如先进的基于注意力/卷积的模型、多尺度模型和搜索得到的模型。

感受野搜索提升先进模型 本文展示了例如 PVT [281, 282]，ConvNeXt [283] 等最近的先进模型均能受益于本文的感受野搜索方法。PVTv2 [281, 282] 是一种

表 4.15 基于 PVTv2-B0 主干网络和 Semantic FPN [282] 方法进行语义分割的感受野搜索性能。

	P	Pascal VOC [237]		ADE20K [262]	
		mIoU	mAcc	mIoU	mAcc
PVTv2-B0		73.7	85.0	37.5	48.3
RF-PVT		74.4	86.0	38.0	48.6
RF-PVT	✓	74.4	85.9	37.8	48.7

基于自注意的金字塔视觉 Transformer，它同时使用全局自注意和深度卷积。本文将感受野搜索应用于 PVTv2 的卷积中，并将其用于目标检测、实例分割和语义分割任务，实现了比 PVTv2-B0 基线模型更好的性能。如表 4.14所示，对于目标检测和语义分割来说，单分支 RF-PVTv2 的边界框 mAP 提升了 0.6%，掩码 mAP 提升了 0.6%。并行 RF-PVTv2 的边界框 mAP 提升了 0.3%，掩码 mAP 提升了 0.3%。表 4.15表明，搜索到的单分支结构在 Pascal VOC 和 ADE20K 数据集上的性能提升为 0.7% 和 0.5%。ConvNeXt [283] 是一个先进的卷积模型，它优于许多基于注意力的先进模型。尽管 ConvNeXt 手动调整卷积核大小以支持更大范围的感受野，但感受野搜索仍然进一步提高了目标检测和实例分割的性能。如表 4.14所示，并行的搜索结构相较于 ConvNeXt-T 模型获得了 0.5% 的边界框 mAP 提升和 0.6% 的掩码 mAP 提升。这两个强大的先进模型的性能提高证明了本文的感受野搜索方法的有效性。

感受野搜索提升多尺度模型 本文将展示感受野搜索相较于几个流行的手工设计的多尺度模型的优势，如上一章节提到的 Res2Net 和 HRNet [284, 285]。Res2Net 在一个块内构建类似于残差的分层连接，以实现细粒度的多个感受野。HRNet 并行处理多分辨率特征形成多尺表征。尽管他们有良好的多尺度能力，在表 4.16中显示，本文的感受野搜索算法仍然提高了它们在目标检测和实例分割任务上的性能。对于 HRNet 来说，单分支和多分支的 RF-HRNet 对目标检测的边界框 mAP 分别提升了 1.3% 和 2.1%，实例分割掩码 mAP 分别提升了 1.2% 和 1.7%。单分支和多分支的的 RF-Res2Net 对目标检测边界框 mAP 分别提升了 0.6% 和 1.6%，实例分割掩码 mAP 分别提升了 0.7% 和 1.5%。因此，本文的感受野搜索方法可以进一步改进多尺度模型，使其具有更好的感受野组合。

与注意力机制比较 注意机制理论上可以形成任意的感受野 [28, 29, 30]。然而，注意机制的实际感受野的表征能力是未知的。因此，本文提出将感受野搜

表 4.16 在 COCO 验证集数据集上测试的目标检测和实例分割任务，感受野搜索改进了手工设计的多尺度模型，例如第三章节提出的 Res2Net 和 HRNet [284, 285]。Cascade Mask-RCNN 方法用作检测器。

物体检测	P	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
Res2Net-101		46.3	64.4	50.5	27.2	50.3	60.5
RF-Res2Net		46.9	65.8	51.2	28.4	50.7	62.1
RF-Res2Net	✓	47.9	66.6	52.2	29.7	51.9	62.8
HRNetV2p-W18		41.6	58.7	45.4	23.5	44.7	54.9
RF-HRNet		42.9	60.8	46.7	25.9	46.2	54.8
RF-HRNet	✓	43.7	61.9	47.7	26.5	47.3	56.7
实例分割		mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
Res2Net-101		40.0	61.7	43.3	22.2	43.8	54.1
RF-Res2Net		40.7	63.2	43.9	20.4	44.0	59.0
RF-Res2Net	✓	41.5	64.0	44.9	21.3	44.6	59.5
HRNetV2p-W18		36.4	56.3	39.3	19.1	39.1	49.5
RF-HRNet		37.6	58.3	40.4	19.0	40.2	53.9
RF-HRNet	✓	38.1	59.3	41.0	19.4	40.7	55.3

索方法与非局部（NonLocal）模块在实例分割任务上进行比较。根据其官方实现 [29]，本文将非局部模块插入到 ResNet50 主干的第 4 阶段的每个残差块中。如表 4.10所示，在 COCO 测试集上，基于非局部的 Mask-RCNN 相较于基线 Mask-RCNN，掩码 mAP 提升了 1.1%，边界框 mAP 提升了 1.3%。具有搜索并行感受野的 Mask-RCNN 比基于非局部的 Mask-RCNN，掩码 mAP 提升 1.1%，边界框 mAP 提升 1.2%，这表明搜索的感受野比非局部模块能提供更好的表征。因此，尽管非局部模块对模型性能有所改进，但它不能提供像本文搜索感受野方法那样强的有效感受野。然后，本文将局部搜索应用于非局部网络，并使搜索的超参数与基准保持一致。本文观察到，搜索的单分支感受野和并行分支感受野进一步提高了网络性能。单分支版本的掩码 mAP 和边界框 mAP 的增益分别为 0.4% 和 0.6%，并行分支版本的掩码 mAP 和边界框 mAP 的增益分别为 1.3% 和 1.4%。感受野搜索方案可以进一步提高非局部模型的性能，说明非局部模块即使在像素之间有密集连接，也不能覆盖所有有效的感受野。

在搜索得到的网络上的感受野搜索 Auto-deeplab [37] 在分割网络的不同阶段中搜索特征的分辨率。为了验证是否有可能进一步调整 Auto-deeplab 的感受

表 4.17 使用 Auto-deeplab [37] 和 Cityscapes 数据集 [1] 对语义分割任务进行局部搜索的性能。Local-P 表示如章节4.2.3 中所述具有并行感受野的局部搜索结构。

	P	mIoU	mAcc
Auto-deeplab [37]		76.0	83.7
RF-Auto-deeplab		76.3	84.2
RF-Auto-deeplab	✓	76.7	84.3

野，本文在 Auto-deeplab 基础上进行了局部搜索。本文遵循 Auto-deeplab 的实现 [37]，在 Cityscapes [1] 数据集上进行实验。如表 4.17所示，单分支和多分支的 RF-Auto-deeplab 均获得了 mIoU 的提升。因此，本文的感受野搜索方案可以进一步提升具有搜索得到的特征分辨率的语义分割模型。

4.5 总结

本章节提出了一种场景自适应的从全局到局部的搜索方案 RF-Next，可以针对各种视觉场景由粗到细地寻找有效的感受野组合。全局搜索可以发现比手工设计具有更好的性能的有效的但是模式不同的感受野组合。期望引导的迭代局部搜索方案能够在密集的搜索空间中搜索细粒度的感受野组合。通过场景自适应地感受野搜索方案增强的 RF-Next 模型，可以插入如动作分割、序列建模、分割、目标检测等多种场景的视觉任务中来进一步提升性能。本章节也证明，RF-Next 感受野搜索算法也可进一步提升第二章提出的尺度自适用主干网络 Res2Net 的性能。

第五章 数据自适应大规模无监督语义分割

在广泛使用的有监督训练范式下，复杂场景产生的大规模和多样化的数据使人工标注成本过于昂贵，因而要求模型在尽可能少的人工干预下完成对数据的表征和理解。为降低数据标注的巨大成本，本章节设计了首个面向大规模场景的无监督语义分割算法，在无需人工标注的情况下实现对数据的自适应表征和像素级语义分割。该算法在无任何人工标注数据的情况下，使模型通过自监督表征学习从百万量级数据中学习丰富的语义特征，并将自适应总结出的上千个语义类别分配给大规模数据中的每个像素。该算法能够实现对视觉感知数据的自适应理解。本章在章节5.1简要介绍了大规模无监督语义分割的背景和思路。章节5.2提出首个针对大规模无监督语义分割的评测基准。章节5.3设计了一个可行的大规模无监督语义分割算法。章节5.4验证本章提出算法的可行性，并分析了大规模无监督语义分割面临的挑战和可能的方向。章节5.5总结了本章的内容。

5.1 大规模无监督语义分割简介

由于语义分割任务的固有挑战，大多数工作集中于在多样性有限 [1, 40, 41] 和数据规模较小 [42, 43] 条件下的语义分割。然而大幅地扩大问题规模往往会导致研究模式的改变，例如从 PASCAL VOC [42] 拓展到 ImageNet [224] 使识别任务难度大幅增加。这促使本文思考一个更具有挑战性的问题：语义分割是否可能用于具有广泛多样性的大规模现实世界环境？然而，巨大的数据规模和隐私问题使基于人工数据标注的有监督大规模语义分割任务的发展受限。当使用数百万张甚至数十亿张图片进行训练，例如 ImageNet、JFT-300M [286] 和 Instagram-1B [287]，分类模型的无监督学习已经展现了和有监督学习相当的能力 [168, 49, 167]。为了实现面向真实世界的语义分割，本文提出了一个新的问题：大规模无监督语义分割（**L**arge-scale **U**nsupervised **S**emantic **S**egmentation，LUSS）。如图 5.1所示，LUSS 任务的目标是在没有人工标注监督的情况下，为大规模图片数据中的每个像素分配类别标签。为了实现这一目标，需要同时解决例如大规模数据下的形状和类别表征学习以及无监督语义聚类等许多挑战。具体来说，模型需要提取具有类别和形状线索的语义表征。类别相关的表征用来区

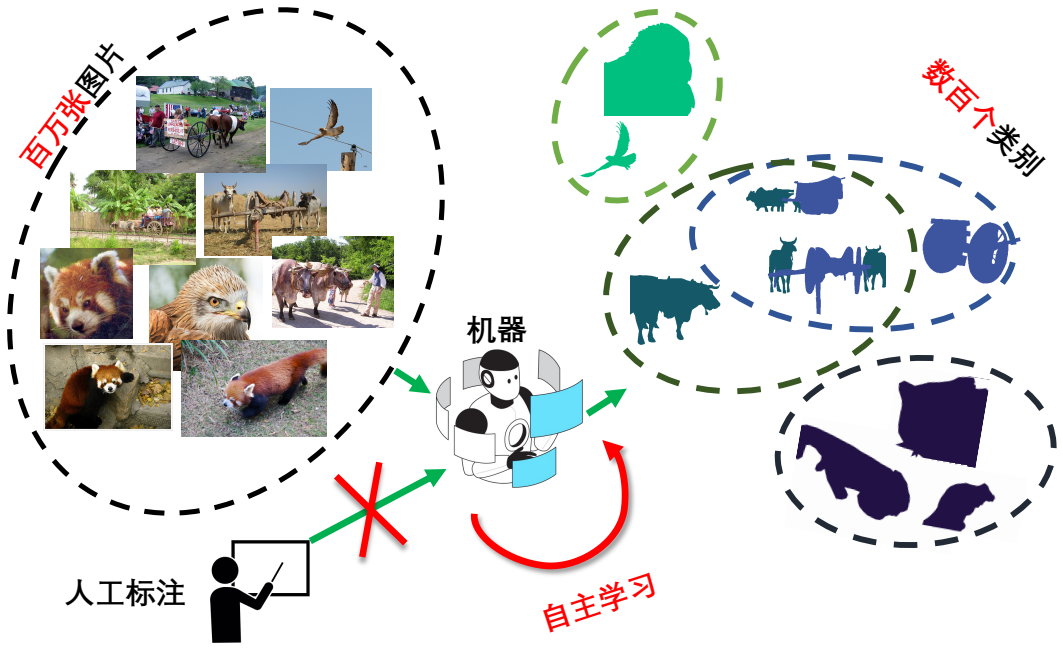


图 5.1 大规模无监督语义分割（LUSS）任务的目标是在没有人类标注的情况下，模型通过自主学习来进行语义分割，将成百上千个自我总结出的类别标签分配给数百万以上的图像中的每个像素。

分不同事物的类别，而例如物体、边缘等形状相关的像素级表征对语义分割至关重要。以上两种表征的有效共存对 LUSS 至关重要，因为冲突的表征可能会导致错误的语义分割结果。从大规模数据中生成类别需要鲁棒和高效的语义聚类算法。为像素指定标签需要区分相关和不相关的语义区域。解决 LUSS 的这些挑战也能帮助许多相关的任务。例如，从 LUSS 中学习到的形状表征可以被用作在数据规模和多样性有限条件下的语义分割 [27, 238] 和实例分割 [124] 等像素级下游任务的预训练表征。此外，利用半监督学习范式来微调 LUSS 模型能够在实际应用中只需人工标注一小部分数据即可实现大规模数据的语义分割。

为助力 LUSS 任务的发展，本节提出了一个评测基准，其中包含高度多样性的大规模数据，无需直接/间接的人工标注的无监督语义分割任务目标，以及从不同角度进行评测的多种指标。高度多样性的大规模数据给 LUSS 带来了挑战的同时也为模型提供获取丰富的表征的来源。因为数据不足，一些无监督分割方法 [44, 45, 46, 47] 主要关注类别和多样性有限的小规模数据的场景，因此不适合 LUSS 任务。基于类别表征学习工作中常用的 ImageNet 数据集 [224]，本文提出了一个用于 LUSS 任务的大规模基准数据集 ImageNet-S。本文移除了

图 5.2 本文提出的 ImageNet-S 数据集的可视化。

ImageNet 中不可分割的类别，例如书店，并且使用剩余的包含919 个类别的大约 120 万张图片用于训练。然后本文在 ImageNet 验证集中标注了 4 万张图片的精准的像素级语义分割掩码用于 LUSS 任务的评测。本文还标注了训练集中大约 9 千张图像，以支撑更全面的评测和对潜在应用的探索。基于 [288] 中更精确的重新标注的图像级标签，本文使 ImageNet 数据集中的单一图片中有多个类别标签。ImageNet-S 数据集提供的大规模和高多样性的数据能够用于 LUSS 方法的公平训练和充分评测对比。

本文提出了一个用于 LUSS 任务的新方法，包括无监督表征学习，标签生成和微调步骤。对于无监督表征学习，本文提出了 1）一种非对比像素级表征对齐策略，以在不损害实例级类别表征的情况下增强像素级形状表征；2）一个提高网络中间层特征表征质量的由深到浅的监督策略。该策略学习到的表征法保证了形状和类别信息的有效共存。本文提出了一种像素注意力方案用来突出有意义的语义区域进而用于像素级标签生成，实现在大规模数据下高效的像素级标签生成和微调。基于提出的方法和 ImageNet-S 数据集，本文分析了本文提出的 LUSS 任务和一些相关工作的关系（例如无监督学习 [49, 167, 175, 170, 171] 和下游任务的迁移学习 [124, 27]）并且确定了 LUSS 任务面临的挑战和可能的研究方向。本章节有两个主要贡献：

- 本文提出了一个新的大规模无监督语义分割问题，以及包含近五万张像素级图像标注、919 个类别和多个评价指标的 ImageNet-S 数据集。
- 本文提出了一种包括增强的表征学习策略和像素注意力机制的新的 LUSS 方法，并且评测了 LUSS 任务的相关工作。

5.2 大规模无监督语义分割基准

LUSS 任务旨在不使用直接/间接人工标注的前提下从大规模图像中学习语义分割。给定大规模图像，LUSS 模型将自学习得到的标签分配给所有图像的每个像素。为了便于理解，本文给出了实现 LUSS 的其中一个方案，见章节5.3。LUSS 模型同时从大规模数据中学习类别和形状表征，而无需人工标注。该模型使用学习的特征表征进行类别标签聚类和分配，以生成图像的像素级标签。然后，根据生成的标签对模型进行微调，以优化分割结果。理想情况下，标签分配和微调步骤可以隐含在无监督的表征学习过程中。

LUSS 面临多重挑战，例如语义表征学习，大规模数据下的类别标签生成，和无监督学习。此外，缺乏评测基准限制了 LUSS 任务的发展。因此，本文制定了具有明确目标、大规模训练数据和全面评价标准的 LUSS 基准。

5.2.1 大规模 LUSS 数据集: ImageNet-S

LUSS 任务非常具有挑战性，因为它不使用人工标注标签进行训练，并且需要大规模数据来学习丰富的表征。原则上，LUSS 所需的训练图像规模随着图像复杂性的增长而增加，例如更多的类别数和复杂的场景需要更多的训练数据。现有的分割数据集由于图像复杂度大而数据规模小，很难支持 LUSS 任务。例如 PASCAL VOC [42] 和 CityScapes [1] 等一些数据集仅包含在少数场景下有限数量的图像。例如 ADE20K [40]、COCO [55] 和 COCO-Stuff [43] 等其他数据集仅有每个类别的样本数量有限的复杂图像，而对于 LUSS 模型来说很难用有限的数据学习复杂场景的丰富表征。

为了弥补这些数据集的缺陷，常见的有监督分割方法 [27, 124, 238] 使用广泛使用的大型 ImageNet 数据集预训练的模型 [15, 18, 245] 进行微调实现分割。然而，最近的研究 [289, 290] 表明，由于数据分布、数据域和任务目标的不稳定性，ImageNet 和下游数据集上的性能并不总是一致的。对于 LUSS 任务来说，微调预训练的模型使得评价复杂化，并且可能导致不公平和有偏见的比较。ImageNet 有更多的类别、更大的数据规模、相对简单的图像，以及针对每个类别的足够图像，这使得模型学习丰富的表征成为可能。因此，ImageNet 被很多无监督学习方法 [167, 175, 171, 168, 49] 广泛使用。然而，ImageNet 仅有图像级的标注，因此不能用于 LUSS 任务中像素级的评测。为了促进 LUSS 任务，本文从 ImageNet 数据集 [224] 中收集数据并提出了一个大规模 ImageNet-S 数据

图 5.3 ImageNet-S 数据集的类别结构树。

集并且为 LUSS 评测标注了像素级标签。本文移除了例如书店等不可分割类别，使用了在 ImageNet 中剩余的919 个类别。ImageNet-S 数据集（见图 5.2）比现存的数据集在图像数量（见表 5.1）和类别多样性（见图 5.3）都更大。

5.2.1.1 图像标注

本文在 ImageNet-S 数据集中标注验证/测试集和部分训练集，以进行 LUSS 评测。因为 ImageNet 数据集有错误的标签并且缺少多类别标签，本文依照 [288] 中重新标注的图像级标签标注像素级语义分割掩码，并进一步更正缺失和错误的标注。图像级标签对应的对象被标注，其他部分被标注为“其他”类别。“其他”类别代表这些类别不经常出现在数据集或只出现在周围环境中。对于验证/测试集，本文标注了919 个所选类别中的所有对象。难以区分的部分标记为“忽略”，不会用于评测。对于训练集，本文对每个类别随机选取十张图像并且标注对应于该类别的对象，而属于919 个类别的其他对象被标注为“忽略”。

纠正缺失/不正确的标签 在标注过程中，本文发现由于 ImageNet 高多样性和大规模特性，在 [288] 的更正标注中仍有一些缺失和错误的图像级标注。因此，本文提出了几种方案来尽可能地纠正标签：1）本文发现有些类别是相互关联的，

表 5.1 ImageNet-S 数据集和现有的语义分割数据集图像类别和数量比较。

数据集	类别数	训练集	验证集	测试集
PASCAL VOC 2012 [42]	20	1,464	1,449	1,456
CityScapes [1]	19	2,975	500	1,525
ADE20K [40]	150	20,210	2,000	3,000
ImageNet-S_{50}	50	64,431	752	1,682
ImageNet-S_{300}	300	384,862	4,097	9,088
ImageNet-S	919	1,183,322	12,419	27,423

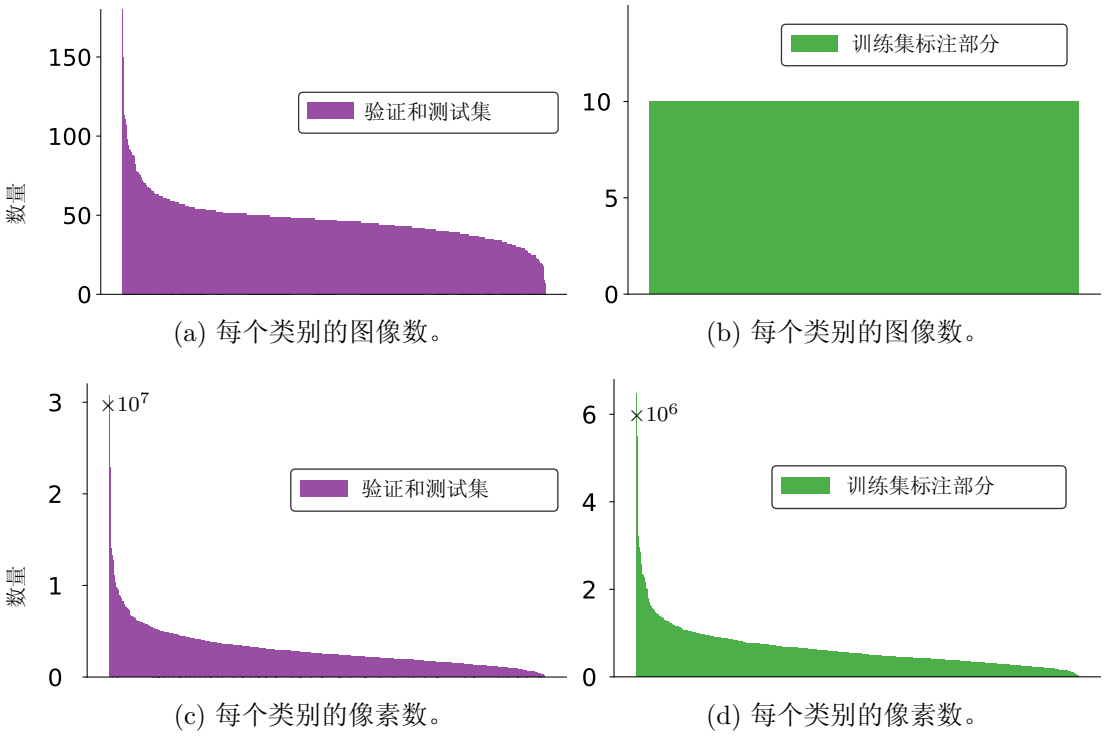
表 5.2 ImageNet-S 数据集中每个图像中的类别数。

	图片数量					
	验证集			测试集		
单张图片内主要类别数量	1	2	>2	1	2	>2
ImageNet-S_{50}	745	7	0	1,676	6	0
ImageNet-S_{300}	3,971	118	8	8,815	264	9
ImageNet-S	11,294	954	171	25,133	1,938	352

例如蜘蛛和蜘蛛网通常出现在同一张图片中。基于最初标注者观察到的缺失类别，本文对该类别与其他类别相关的图像进行了复查。2）本文使用第三章节提出的 Res2Net 和 Swin transformer [291] 等有监督训练图像的分类器，通过检查分类器预测出高置信度但不是真实标签（Ground Truth，GT）标签的类别，来找到缺失的类别标注。通过这些方案，本文纠正了296 个错误标注的图像并且发现了942 个缺少标签的图像。

5.2.1.2 统计和分布

图像数量 如表 5.1所示，在 ImageNet 数据集中移除了例如书店，山谷和图书馆这样不可分割类别之后，ImageNet-S 数据集包含919 个类别，1,183,322 张训练图像, 12,419 张验证图像，和27,423 张测试图像。现有的许多自监督表征学习方法 [49, 171] 使用 ImageNet 数据集训练。为了公平比较，本文使用包含1,281,167 张训练图片的 ImageNet 数据集学习无监督表征，并使用 ImageNet-S 数据集进行 LUSS 的其他步骤。本文用精确的像素级掩码标注了39,842 张验证/测试图片和9,190 张训练图片，并且在图 5.2可视化了一些标注。本文的像素级标注使得 ImageNet-S 数据集在每张图片中有多个类别。表 5.2 给出了在 ImageNet-S 验



(a) 每个类别的图像数。 (b) 每个类别的图像数。

(c) 每个类别的像素数。 (d) 每个类别的像素数。

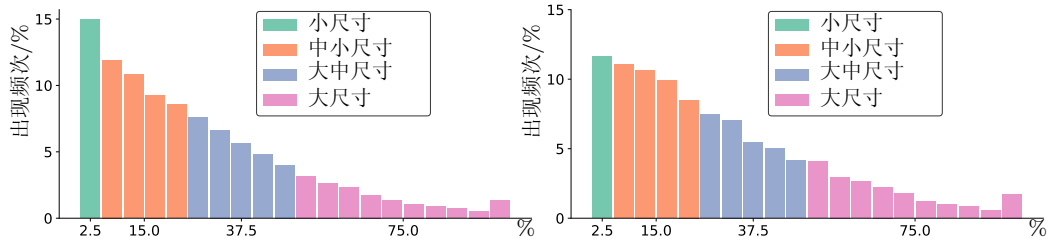
图 5.4 在 ImageNet-S 数据集的类别中实例级/像素级样本的数量分布，即每个类别的图像/像素数。

证/测试集中每张图片的类别数。大量的图像包含一个类别，8.6% 的图像有多于一个类别。ImageNet-S 相比现有的分割数据集有更简单的图像和更多的类别，与 LUSS 任务下没有无标注、大规模的图像和大量的类别相适配。

类别分布 如图 5.3所示，ImageNet-S 数据集中的类别由于是从单词树 [224] 中提取的，因此其展现了一个树形结构。图 5.4 展示了 ImageNet-S 数据集类别与图像、像素的数量分布，即每个类别中包含的图像/像素的数量。训练集和验证/测试集有相似的分布。大多数类别的图像数量是均衡的，而每个类别的像素数量呈现长尾分布。像素级类别分布不平衡可能会带来图像级表征学习中未考虑的新挑战。

物体大小 因为分割更小的物体会更难，本文根据物体与图像的比例将物体分为以下几组，即小尺寸（0%-5%），中小尺寸（5%-25%），大中尺寸（25%-50%），和大尺寸（50%-100%）物体。图 5.5 中展示的物体大小分布展示出大多数的物体相对较小。

位置分布 本节重叠来自验证和测试集的分割掩码，以分析语义物体在数据集



(a) 验证集和测试集的分布。 (b) 训练集已标注图像的分布。

图 5.5 ImageNet-S 数据集中物体大小的分布。物体大小定义为物体与图像大小的比率。

中的位置分布，见图 5.6（上）。ImageNet-S 数据集中的物体具有中心偏向分布，这说明了现有自监督学习方法 [167, 49] 的中心裁切策略的有效性。本文还重叠了物体的边界，见图 5.6 （下）。它表明物体几乎覆盖了所有区域，而不仅仅是图像的中心区域。此外，如图 5.6，本文将 ImageNet-S 数据集与 COCO [55] 和 Open Images [292] 数据集的分布进行比较。ImageNet-S 数据集和其他两个数据集具有相似的分布。所有数据集都观察到了中心偏态分布，本文猜测人类可能倾向于记录更多的中心偏态图像。有趣的是，ImageNet-S 的分布图几乎与 Open Images 数据集相同，而后者以其真实性而闻名。

在有限资源下的 ImageNet-S-50/300 为了在低计算资源预算下促进研究，本文提出了两个包含50 和300 类别的子集，名为 ImageNet-S_{50} 和 ImageNet-S_{300}。考虑到 LUSS 任务的艰巨性，本文为 ImageNet-S_{50} 选择50 个在日常生活中容易区分的类别。ImageNet-S_{300} 由 ImageNet-S_{50} 和250 个随机选取的类别组成。ImageNet-S_{50} 和 ImageNet-S_{300} 中图像的数量见表 5.1。即使是 ImageNet-S_{50} 子集也比大多数语义分割数据集拥有更多的图像。

5.2.2 评测

5.2.2.1 评测方案

因为在训练过程中缺少人为标注类别的监督，LUSS 模型不能像有监督学习得到的模型一样直接测试性能。因此，本文为 LUSS 提出了三个评测方案，包括完全无监督评测、半监督评测以及距离匹配评测。

完全无监督评测方案 完全无监督评测方案在训练期间不需要人为标注标签，只需要验证/测试集进行评测。与有监督任务不同，LUSS 任务中的类别是由模型生成的，在评测期间需要与 GT 类别匹配。本文提出了一个默认的图像级匹

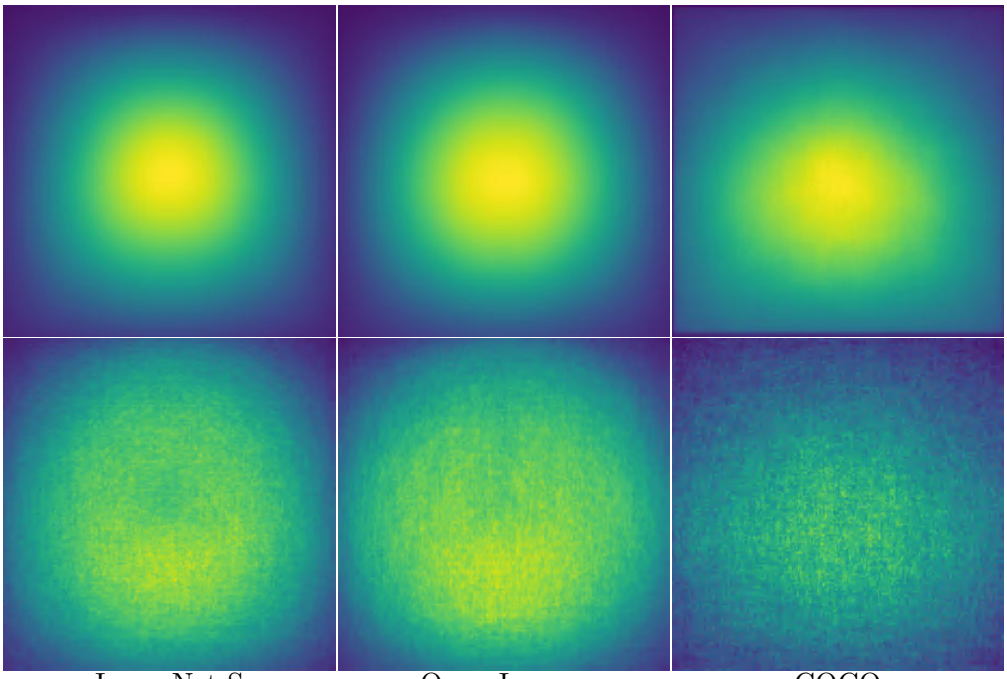


图 5.6 数据集之间的物体位置分布比较：（上）分割掩码的位置分布，（下）掩码边界的位置分布。

配方案，而更有效的匹配方案应该可以进一步提高 LUSS 评测性能。假设匹配集（通常为验证集）具有 N 张图像和 C 个类别。因为数据集有 C 个主要类别，因此类别的数量隐式包含在训练数据集中。本文假设无监督模型在训练过程中应该学会从数据集中生成超过 C 个类别。默认的图像级别匹配方案仅将 C 个生成的类别与 C 个真实类别相匹配。给定图像集 $\mathbf{D}=\{\mathbf{D}_k, k\in[1,N]\}$ 和 GT 标签 $\mathbf{G}=\{\mathbf{G}_k, k\in[1,N]\}$ 和预测的类别 $\mathbf{P}=\{\mathbf{P}_k, k\in[1,N]\}$, $\mathbf{G}_k$ 和 $\mathbf{P}_k$ 分别是图像 $\mathbf{D}_k$ 的 GT 和预测的类别集合。本文计算生成的类别和 GT 类别的匹配矩阵 $\mathbf{S}\in\mathbb{R}^{C\times C}$ 如下，其中 $\mathbf{S}_{ij}$ 表示在第 i 个生成的类别和第 j 个 GT 类别间的匹配度，当两个类别更可能是同一类别时其值更大：

$$\mathbf{S}_{ij}=\sum_{k=1}^{N}\mathbb{I}\{(i,j)\in\mathbf{P}_k\times\mathbf{G}_k\}, \tag{5.1}$$

$\mathbf{P}_k\times\mathbf{G}_k$ 是 $\mathbf{P}_k$ 和 $\mathbf{G}_k$ 的笛卡尔积，并且 (i,j) 属于 $\mathbf{P}_k\times\mathbf{G}_k$ 时 $\mathbb{I}$ 等价于1。利用匹配矩阵 $\mathbf{S}\in\mathbb{R}^{C\times C}$，本文在生成的类别和 GT 类别中使用匈牙利算法最大化 $\sum_{i=1}^{C}\mathbf{S}_{i,\mathbf{f}(i)}$ 找到了双射 $\mathbf{f}:i\mapsto j$。

半监督评测方案 因为本文对大约 1% 的训练图像进行像素级标签标注，所以

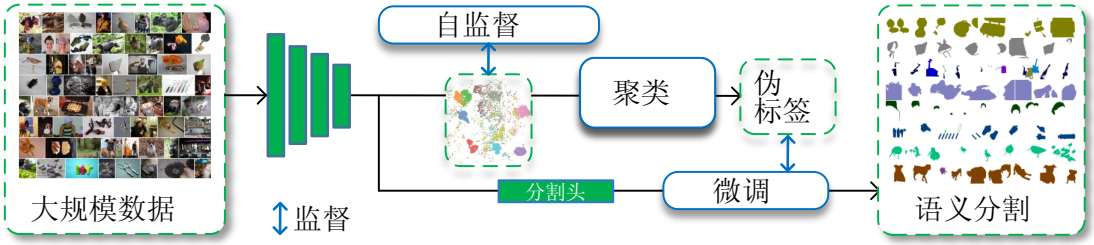


图 5.7 本文提出的实现 LUSS 任务的其中一种方案。

可以进行半监督微调以评测 LUSS 模型。半监督评测方案需要使用人工标注的训练数据对训练好的 LUSS 模型进行微调。因此，该方案不需要匹配生成的类别和 GT 类别。此外，该方案适用于现实世界中大量数据中只有部分图像被人工标注的现实应用场景。

距离匹配评测方案 在距离匹配评测方案中，本文直接利用像素级标注的训练图像获得 GT 类别的特征向量，并将其与验证/测试集中的特征向量进行匹配，以分配类别标签。具体来说，本文得到了训练集中每个类别的所有像素（包括“其他”类别）的平均特征向量和相应的类别标签。然后本文使用 k-NN 分类器 [187] 推理验证/测试集上的分割掩码。对于验证/测试集中的每个像素的特征向量，本文会在训练集中找到前 k 个相似的特征向量和相应的类别标签。每个像素的类别标签由这 k 个特征向量对应的类别投票决定。

5.2.2.2 评价指标

本文使用平均交并比（Mean Intersection over Union，mIoU），边界平均交并比（Boundary mIoU，b-mIoU），图像级准确率（Image-level Accuracy, Img-Acc），和 F-度量（F-measure，F_β）作为 LUSS 任务的评价指标。在评测中，所有图像都使用原始图像分辨率进行评测。mIoU 和 b-mIoU 是综合评测指标，而 Img-Acc 和 F_β 分别从类别和形状方面评测模型性能。

平均交并比 类似于有监督语义分割任务 [42, 40]，本文使用 mIoU 来评测语义分割掩码的质量。除了主要类别外，“其他”类别也被用于计算 mIoU。

边界平均交并比 与上述的评测所有物体区域的掩码的 mIoU 不同，b-mIoU [293] 重点关注边界区域。本文使用 b-mIoU 来评测边界区域的语义分割质量。本文使用 $d = 3\%$ 的 b-mIoU [293]。

图像级准确率 Img-Acc 可以评测模型的类别表征能力。由于许多图像包含多

个标签，本文依照 [288] 将面积最大的预测类别是否属于该图片的 GT 类别集作为分类正确与否的评价标准。

F-度量 除了与类别相关的表征外，本文使用忽略语义类别的 F_β 来评测形状质量 [74]。本文将主要类别视为前景类别，将“其他”类别视为背景类别。

5.3 大规模无监督语义分割方法

5.3.1 概述

本文总结了 LUSS 任务面临的主要挑战：1）模型应该在无需图像级标签监督的情况下学习与类别相关的表征。2）提取语义分割掩码需要模型学习形状表征。3）形状和类别表征应在尽可能减少冲突的情况下共存。4）利用学习到的表征，模型应该高效地为图像中的每个像素分配自学习到的标签。5）大规模的训练数据有助于以无监督学习的方式学习丰富的表征但不可避免地会消耗大量的训练成本，这就要求提高训练效率。

考虑到上述挑战，本文提出了一种新的 LUSS 方法，名为 PASS（见图 5.7），包括四个步骤。1）一个随机初始化的模型通过自监督的代理任务来学习形状和类别表征。经过表征学习，本文得到了所有训练图像的特征集。2）然后，本文应用基于像素注意力的聚类方案来获得伪类别，并将生成的伪类别分配给每个图像像素。3）本文用生成的伪标签微调预训练模型，以提高分割质量。4）在推理时，LUSS 模型与有监督模型相同，即将生成的标签分配给图像的每个像素。注意，本文提出的流程不是 LUSS 任务唯一的选择。下面详细介绍 PASS 方法的每个步骤。为了便于阅读，一些频繁使用的符号见表 5.3。

5.3.2 无监督表征学习

对于本文的 LUSS 方法的第一步，一个随机初始化的模型，例如 ResNet，通过自监督的代理任务来学习语义表征。LUSS 任务需要类别相关表征来区分不同类别的场景，并需要形状相关表征来构建物体的形状。之前的工作已经做了很多努力来学习图像级类别相关表征或像素级表征 [171, 196, 197]。然而，图像级方法通常忽略形状相关的特征。像素级方法更多关注有监督下游任务的迁移学习性能。通过 [294] 的发现，大多数下游任务的性能依赖于网络浅层的低级特征。因此，在下游任务中表现良好的像素级方法可能无法学习到有类别和形状信息的高级语义特征。

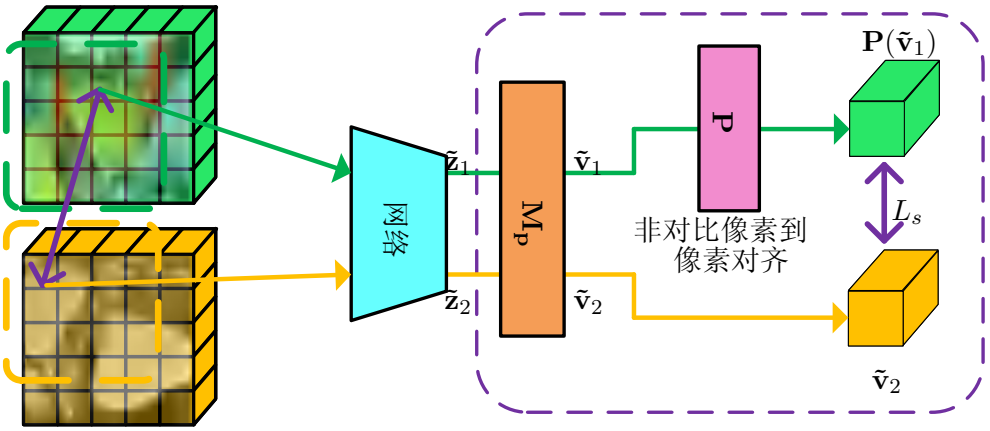


图 5.8 非对比像素到像素表征对齐策略的图示。$\mathbf{M_p}$ 是确保像素级表征减小对类别表征干扰的映射层。$\mathbf{P}$ 是非对称损失的像素级预测器。

表 5.3 本章节常用符号的定义。

符号	维度/类型	含义
$\mathbf{z}$	$L\times H\times W$	一个图像的输出特征
$\mathbf{z}_k$	$L\times H\times W$	第 k 个图像的输出特征
$\mathbf{q}_k$	$(C+1)\times H\times W$	第 k 个图像的像素级伪标签
$\mathbf{y}_k$	$(C+1)\times H\times W$	第 k 个图像的像素级 GT 标签
C	标量	主要类别数
L	标量	输出特征的维度
H	标量	输出特征的高度
W	标量	输出特征的宽度
N	标量	图像数
$\mathbb{P}$	操作	空间维度上的全局平均池化

为了获得强大的表征来支持 LUSS 任务，本文提出了两种自监督学习策略来增强类别和形状表征，包括 1）一种非对比像素到像素表征对齐策略，用于增强像素级形状相关表征，而不会损害实例级类别表征。2）一种由深到浅的监督策略，以提高网络中间层特征的表征质量。

非对比像素到像素表征对齐 像素级形状相关表征旨在增强像素级的特征区分能力，即同一类别或来自同一图像的不同视图的相同位置的像素应具有一致的表征，反之亦然。本文观察到，大多数现有的像素级表征方法在 LUSS 任务上的性能比图像级表征方法差。本文认为现有的像素级方法过于关注像素级的区别，从而导致同一物体实例中像素间的语义差异。为了避免像素级表征对实例

级类别表征的副作用，本文提出了一种非对比像素到像素表征对齐策略，该策略将来自同一图像的不同视图的相同位置的特征对齐，但不刻意增大不同位置的表征差异。

如图 5.8所示，给定从同一个图像的两个视图预测的特征对，本文在重叠区域提取特征图 $(\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2)$ 并且通过映射 $\tilde{\mathbf{v}} = \mathbf{M}_\mathbf{p}(\tilde{\mathbf{z}})$ 获得像素级表征向量对 $(\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2)$，其中 $\mathbf{M}_\mathbf{p}$ 是包含两个 1×1 卷积和激活层的像素级多层感知机（MLP）。本文在章节5.4.3中展示，映射 $\mathbf{M}_\mathbf{p}(\tilde{\mathbf{z}})$ 减少了像素级表征对类别表征的干扰。本文利用像素到像素对齐策略，使用非对称损失将两个视图重叠区域像素的特征向量对齐：

$$L_{I2I} = L_s(\mathbf{P}(\tilde{\mathbf{v}}_1), \mathcal{G}(\tilde{\mathbf{v}}_2)) + L_s(\mathcal{G}(\tilde{\mathbf{v}}_1), \mathbf{P}(\tilde{\mathbf{v}}_2)), \tag{5.2}$$

其中映射 $\mathbf{P}$ 是像素级 MLP 预测器，$\mathcal{G}$ 是为了避免预测器崩溃的停止梯度操作 [172]，L_s 是余弦相似性损失函数。本文提出的非对比像素到像素对齐策略在不同视图之间形成稳健的像素级表征的同时保持了类别表征能力。

由深到浅的监督 网络浅层的低级、中级特征的质量，已被证明对视觉任务至关重要 [295, 290]。Islam 等人 [295] 揭示了网络浅层中具有丰富的低/中级语义的表征，从而能够快速适应新任务。类似地，Kotar 等人 [290] 展示了使用基于对比学习的方法能够有效学习高质量的低级特征。现有的大多数工作都是通过网络高层的间接梯度反向传播来优化中级表征 [49, 167, 174, 175]。本文观察到，由于低/中级特征缺乏语义信息，直接使用它们进行表征学习会导致次优性能。因此，本文提出了一种由深到浅的监督学习策略，以通过高质量高级特征监督的方式来增强低/中级特征的表征质量。

如图 5.9，给定从一幅图像经过数据增强得到的两个视图，本文从网络的 s 阶段得到特征对为 $(\mathbf{z}_1^{(s)}, \mathbf{z}_2^{(s)})$。为了简单起见，本文主要研究图像级由深到浅监督的影响。给定一个具有四个阶段的网络，用于由深到浅监督的图像级特征向量如下：

$$\mathbf{u}_i^{(s)} = \begin{cases} \mathbf{M}_I^s(\mathbb{P}(\mathbf{z}_i^{(s)})) & s = 4; \\ \mathbf{M}_I^s(\mathbb{P}(\mathbf{M}_K^s(\mathbf{z}_i^{(s)}))) & s < 4, \end{cases} \tag{5.3}$$

其中 $\mathbb{P}$ 是空间维度的全局平均池化操作，$\mathbf{M}_I^s$ and $\mathbf{M}_K^s$ 分别是阶段 s 的图像级/像素级 MLP 层。本文观察到直接全局池化中层特征会导致表征崩溃，因此添加 $\mathbf{M}_K^s$ 来避免此问题。在由深到浅的监督策略中，一个视图最后阶段的特征

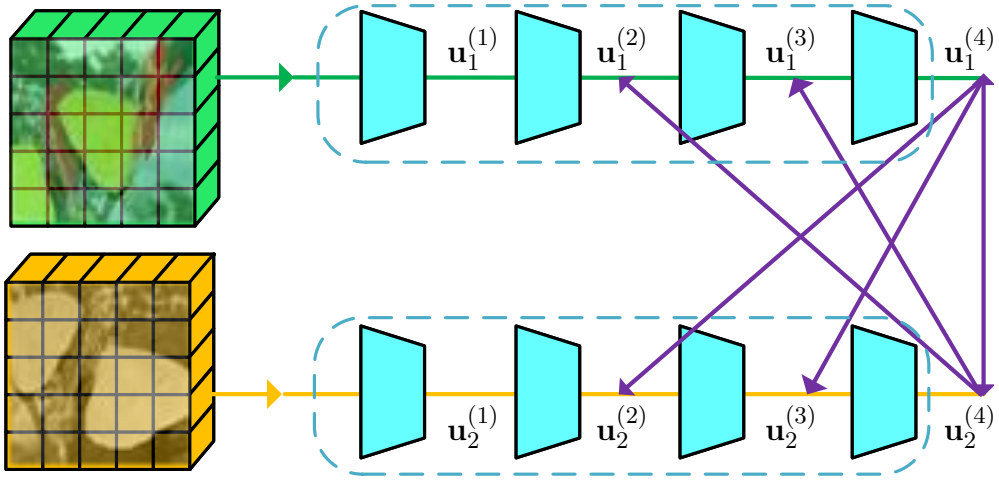


图 5.9 由深到浅监督的图示。紫色线表征使用损失函数 L_I 作为监督。为了简单起见，本图省略了 $\mathbf{M}_I^s$ 和 $\mathbf{M}_K^s$。

向量用于监督另一个视图的所有阶段的特征向量：

$$L_{D2S} = \frac{1}{|S|}\sum_{j}^{j\in S} L_I(\mathbf{u}_1^{(4)}, \mathbf{u}_2^{(j)}) + \frac{1}{|S|}\sum_{j}^{j\in S} L_I(\mathbf{u}_2^{(4)}, \mathbf{u}_1^{(j)}), \tag{5.4}$$

其中 S 是用于由深到浅监督的阶段的集合，L_I 是图像级损失。L_I 可以定义为多种形式，在本文中本文使用聚类损失 [175] 作为 L_I。

表征学习的训练损失 本文提出的像素到像素对齐和由深到浅监督可以与现有方法配合使用，以提高表征质量。无监督表征学习步骤的损失函数如下：

$$L_{sum} = L_{I2I} + L_{D2S} + L_e, \tag{5.5}$$

其中 L_e 是例如 SwAV [175] 和 PixelPro [171] 等现有方法的损失函数。

5.3.3 使用像素注意力生成像素标签

在表征学习之后，本文获得了所有训练图像的特征集合 $\mathbf{Z} = \{\mathbf{z}_k \in \mathbb{R}^{L\times H\times W}, k \in [1, N]\}$，其中 N 是图像的数量，L、H 和 W 是输出特征图的维度、高度和宽度。本文对 $\mathbf{Z}$ 进行聚类，以获得 C 个生成的类别，并将生成的类别分配给每个像素。标签生成的一种简单方法是对训练集中所有像素的特征向量进行聚类，LUSS 中的大规模数据导致聚类成本太高，例如聚类 ImageNet-S 训练集的 7×7 分辨率的像素级特征需要大约114 小时。另一种方法是使用在空间

维度上聚类的图像级特征来节省聚类成本。然而，全局池化后的特征图中包含了许多不相关的特征信息，会影响聚类质量。

本文观察到，模型学习到的特征往往更关注语义更丰富的区域，例如具有更多有用语义信息的像素更有助于无监督表征学习模型的收敛。基于这一观察，本文提出了一种像素注意力方案，以突出有意义的语义区域，便于使用图像级特征生成像素级标签。具体来说，本文在模型的输出端添加一个像素注意力模块，并使用表征学习损失对其进行微调，以过滤掉语义信息较少的区域。具有像素注意力的过滤功能可以减少混合图像级特征向量中的噪声，从而提高聚类质量。此外，像素注意力将语义丰富的区域与语义较少的区域分开，从而在像素级标签生成过程中生成更精确的物体形状。本文在微调和标签生成步骤中给出了像素注意力的实现细节。

微调像素注意力 给定模型预测的一幅图像的特征 $\mathbf{z}$，表征学习方法 [175, 49]，使用池化特征向量 $\mathbf{M}_I(\mathbb{P}(\mathbf{z}))$ 计算损失，其中 $\mathbf{M}_I$ 是图像级 MLP 层。池化操作平均地使用所有像素的特征。因为并非所有像素都表示有意义的语义，该操作不可避免地将噪声引入到图像级特征向量。本文的像素注意力被定义为：

$$\mathbf{c}(\mathbf{z}) = \sigma(\mathbf{M}_A(\|\mathbf{z}\|) + \theta), \tag{5.6}$$

其中 $\mathbf{M}_A$ 是像素级 MLP 层，$\theta \in \mathbb{R}^L$ 是初始化为 0 的可学习参数，σ 是限制输出注意值范围的 sigmoid 函数，$\|\mathbf{z}\|$ 是应用于特征 $\mathbf{z}$ 的通道维度的 L2 正则化操作。$\mathbf{z}$ 的每一个通道都有对应的像素注意力图。本文将像素注意力乘到特征图 $\mathbf{z}$ 上并且获得像素注意力增强的图像级特征向量 $\hat{\mathbf{v}} = \mathbf{M}_I(\mathbb{P}(\mathbf{c}(\mathbf{z}) \cdot \|\mathbf{z}\|))$。在微调过程中，本文将回传到网络的梯度断开，只利用 $\hat{\mathbf{v}}$ 计算的表征损失优化像素注意力模块。本文发现使用聚类损失 [175] 微调的像素注意力模块能获得与形状相关的像素注意力图（见图 5.10）。

基于像素注意力的标签生成 基于像素注意力 $\mathbf{c}(\mathbf{z})$, 本文得到像素注意力增强的图像级别特征 $\hat{\mathbf{Z}} = \{\hat{\mathbf{Z}}_k \in \mathbb{R}^L, k \in [1, N]\}$，其中 $\hat{\mathbf{Z}}_k = \mathbb{P}(\mathbf{c}(\mathbf{z})_k \cdot \|\mathbf{z}_k\|)$。本文在 $\hat{\mathbf{Z}}$ 上构造 k-均值聚类来生成 C 个类别的聚类中心 $K \in \mathbb{R}^{L \times C}$。根据生成的类别，本文需要给图像分配像素级伪标签 $\mathbf{Q} = \{\mathbf{q}_k \in \mathbb{R}^{C+1 \times H \times W}, k \in [1, N]\}$。本文在图 5.10中展示了微调后的像素注意力可以突出图像中的语义区域。因此，本文

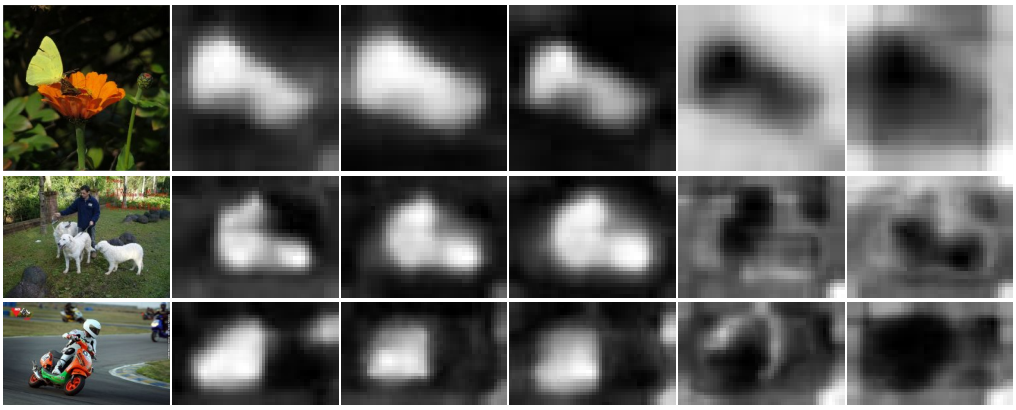


图 5.10 不同通道像素注意力图的可视化。大多数像素注意力图突出显示语义区域，而一些通道突出显示背景区域。

基于像素注意力提取语义信息丰富的区域：

$$\mathbf{d}(\mathbf{z}) = \begin{cases} 0 & \frac{1}{L}\sum_{i}^{0\le i<L}\mathbf{c}(\mathbf{z})_i < \tau; \\ 1 & \frac{1}{L}\sum_{i}^{0\le i<L}\mathbf{c}(\mathbf{z})_i \ge \tau, \end{cases} \tag{5.7}$$

其中 τ 是用于区分主要类别和“其他”类别的预定义的阈值，低于 τ 的区域被视为“其他”类别。对于主要类别区域中的每个像素，本文将聚类中心 K 中与该像素的特征向量距离最小的类别作为其类别。

5.3.4 微调和推理

在微调步骤中，本文加载无监督表征学习得到的预训练权重并添加具有 $L\times(C+1)$ 个通道的 1×1 卷积层作为分割头。使用伪标签 $\mathbf{Q}$ 用交叉熵损失来监督分割头的输出特征 $\mathbf{Y}=\{\mathbf{y}_k\in\mathbb{R}^{(C+1)\times H\times W}, k\in[1,N]\}$ 从而微调模型。LUSS 模型的推理过程与一个完全有监督的语义分割模型的推理过程一致。对于每个在 $\mathbf{y}_k$ 中的像素特征向量 $\mathbf{w}\in\mathbb{R}^{C+1}$，通过如下公式获得分割类别标签：

$$\mathbf{w} = \arg\max_{i\in[1,C+1]}(\mathbf{w}_i). \tag{5.8}$$

5.4 实验与分析

5.4.1 实现细节

表征学习步骤的训练细节 本文在 ImageNet-S_{50} 数据集上使用 ResNet-18 网络，在 ImageNet-S_{300} 和 ImageNet-S 数据集上使用 ResNet-50 网络。为了公

平比较，所有网络使用256 的批大小，在 ImageNet-S_{50} 上训练 200 个迭代轮次，在 ImageNet-S_{300}/ImageNet-S 上训练 100 个迭代轮次。本文分别基于图像级方法 SwAV [175] 和像素级方法 PixelPro [171] 实现了本文提出的表征学习方法。依照 SwAV [175]，LARS 优化器用于更新网络，权重衰减为 1e-6，动量为 0.9。初始学习率为 0.6，并使用余弦学习率调整策略逐渐衰减到 6e-6。对于 ImageNet-S_{300} 和 ImageNet-S 数据集，为与其他方法进行公平比较，本文只使用大小为 224×224 的两个裁剪视图进行训练。与 SwAV 相同，从第15 个迭代轮次开始使用一个长度为3,840 的队列，并且在5,005 次迭代前将聚类中心冻结。当在 ImageNet-S_{50} 上进行训练时，为加速收敛，队列长度设置为 2048，并且聚类中心在1,001 次迭代之前被冻结。在 ImageNet-S_{50} 数据集上训练，本文使用多次裁剪训练策略，其中包含六个大小为 96×96 的裁剪视图，两个裁剪大小为 224×224 的视图。当和 PixelPro [171] 结合时，训练模式与官方设置一致。本文使用 LARS 优化器训练网络，初始学习率为 1.0。经过五个预热迭代轮次后，学习率随着余弦学习率调整策略逐渐下降到 1e-6。

微调步骤的训练细节 为了生成像素级标签，本文首先对像素注意力模块微调20 个迭代轮次，此时固定在无监督表征学习阶段训练好的模型参数。本文默认使用聚类损失 [175] 来微调像素注意力模块。训练策略与表征学习步骤中的策略相同。在微调步骤中，表征学习损失被移除，使用交叉熵损失来监督分割头。本文加载在表征学习步骤上预先训练好的模型权重，并对其进行20 个迭代轮次的微调。本文使用一个权重衰减为 1e-6、批大小为 256、动量为 0.9 的 LARS 优化器来训练网络。初始学习率为 0.6，随着余弦学习率调整策略逐渐衰减到 6e-6。

5.4.2 大规模无监督语义分割性能

在本节中，本文使用完全无监督评测方案在 ImageNet-S 数据集上评测本文提出的 LUSS 方法的性能。表 5.4表明本文的方法在大规模数据上取得了合理的性能。图 5.11中的可视化说明大规模数据的无监督语义分割是可行的。

与无监督语义分割方法的比较 现存的无监督语义分割方法被设计用于相对较小规模的数据，因此由于训练时间限制不能被直接用于 ImageNet-S 数据集。因此，本文在 ImageNet-S_{50} 子集上将本文的 LUSS 方法和现有的小规模无监督语义分割方法进行比较，见表 5.4。为了公平比较，在 ImageNet-S_{50} 数据集上训练的方法都使用了 ResNet-18 网络。严格来说，这种比较并不公平，因为一些现有的小规模无监督语义分割方法没有在完全无监督的条件中进行训练。例如,

表 5.4 在 ImageNet-S 数据集上完全无监督评价方案下本文提出的 PASS 方法和现有的小规模无监督语义分割方法的比较。不同物体尺寸下的测试 mIoU 也在表中提供。† 代表从头开始训练模型 200 个迭代轮次。本文方法的 $_{s/p}$ 分别表示在公式 (5.5) 中使用 SwAV [175] 和 PixelPro [171] 作为 L_e。S 代表该方法使用显著图。I 代表使用有监督 ImageNet-1K 预训练权重作为模型初始化。默认情况下，“其他”类别用于计算 mIoU 和 b-mIoU。

LUSS 任务	先验	mIoU		b-mIoU		Img-Acc		F_β		mIoU				b-mIoU			
		验证	测试	验证	测试	验证	测试	验证	测试	S.	M.S.	M.L.	L.	S.	M.S.	M.L.	L.
ImageNet-S_{50}																	
MDC[191, 48]	-	4.0	3.6	1.4	1.2	14.9	13.4	31.6	31.3	0.4	2.6	3.8	4.9	0.2	1.1	1.4	1.5
MDC[191, 48]	I	14.6	14.3	3.1	3.1	44.8	40.8	33.2	32.6	2.6	10.9	14.6	19.1	0.9	2.2	3.2	4.7
PiCIE[48]	-	5.0	4.5	1.8	1.6	15.8	14.0	14.6	32.2	0.2	3.1	5.0	5.3	0.2	1.2	1.7	1.9
PiCIE[48]	I	17.8	17.6	3.7	4.0	45.0	44.0	32.1	31.6	4.4	13.1	20.1	23.1	1.0	2.7	4.4	5.8
MaskCon[47]	S	24.6	24.2	15.6	15.1	47.9	47.6	65.7	66.2	12.2	25.6	24.7	20.4	10.1	17.0	14.5	10.6
MaskCon[47]†	S	13.9	10.5	8.5	10.5	30.2	22.4	62.6	62.3	2.5	2.1	1.7	1.7	2.4	6.3	6.5	5.7
PASS$_s$	-	29.2	29.3	7.6	7.4	**66.2**	**65.5**	49.0	49.0	6.6	25.0	33.2	32.6	3.3	6.2	8.1	9.5
PASS$_p$	-	32.4	32.0	7.2	7.2	62.9	64.1	48.7	47.9	9.7	26.2	36.5	40.5	5.1	5.8	7.8	10.4
PASS$_p$+ RC [74]	S	42.6	42.1	17.5	17.7	58.8	61.8	62.1	61.3	17.0	38.6	**45.5**	**43.7**	11.2	17.2	19.0	**17.1**
PASS$_p$+ Sal	S	**43.3**	**42.3**	**20.4**	**20.2**	64.6	65.2	**70.0**	**69.9**	**19.0**	**41.7**	45.1	38.3	**14.7**	**22.6**	**20.6**	15.3
ImageNet-S_{300}																	
PASS$_p$	-	16.6	16.0	4.4	4.2	34.7	32.8	34.4	34.3	2.8	12.0	16.4	21.7	1.4	3.2	3.9	6.4
PASS$_s$	-	**18.0**	**18.1**	**5.2**	**5.2**	**43.9**	**42.6**	**47.6**	**47.5**	**4.2**	**13.6**	**19.5**	**23.5**	**2.1**	**4.2**	**5.5**	**7.1**
ImageNet-S																	
PASS$_p$	-	7.3	6.6	2.4	2.1	19.9	18.0	34.8	34.6	1.3	4.6	7.1	8.4	0.6	1.5	2.1	2.8
PASS$_s$	-	**11.5**	**11.0**	**3.8**	**3.5**	**24.0**	**22.3**	**37.1**	**36.9**	**2.4**	**8.3**	**11.9**	**13.4**	**1.3**	**3.0**	**3.8**	**4.3**

MDC [191] 和 PiCIE [48] 使用有监督 ImageNet-1K 预训练权重初始化模型。这两个方法在使用 MoCo [49] 预训练权重时会有大幅度性能下降，这说明有监督预训练对这些方法必不可少。MaskContrast [47] 使用 MoCo 预训练权重初始化模型并使用额外的显著图作为监督进行训练。如果从头开始训练此模型，将导致很大的性能损失。相反，本文的 LUSS 方法是从头开始训练的，没有使用直接或间接的人为监督信息。本文的方法包括新的表征学习策略、标签生成方法和微调方案。为了验证本文方法的通用性，本文基于两种表征学习方法来实现本文的方法，即 SwAV[175] 和 PixelPro[171]。本文的方法在 mIoU 指标上相比现有的无监督语义分割方法有显著提升。得益于额外显著图监督，MaskContrast 比本文的方法有更高的 F_β。当本文的方法也使用相同的显著图，在 F_β 指标明显优于 MaskContrast 并且达到了更高的 mIoU。注意在 [47] 中的显著图不是严格的无监督版本，因为其使用了有监督的 ImageNet 预训练权重。本文还实现了其他无监督语义分割方法，例如 IIC [193]。然而，由于这些方法是为仅有几个类别的简单语义分割而设计的，因此它们无法在 ImageNet-S_{50} 数据集收敛。

不同尺寸物体的性能 如章节5.2.1.2中介绍，ImageNet-S 数据集根据物体大小被分为不同组。本文评测了不同物体尺寸下的测试集 mIoU，见表 5.4。在 mIoU

图 5.11 无监督语义分割结果的可视化。最后三行在标签生成过程中使用显著性物体检测提供的先验信息进行训练，因而有着更好的形状质量。

和 b-mIoU 指标下，小物体的性能比大物体差，这表明小物体需要一个具有更精确像素级表征和分割能力的模型。注意，b-mIoU 中不同物体大小的性能差异比 mIoU 小，因为 b-mIoU 对物体大小变化更为鲁棒。

5.4.3 表征学习的有效性分析

本节在 LUSS 任务上对本文提出的和一些现有的无监督表征学习方法进行了对比。除非另有说明，否则本文使用 ImageNet-S_{300} 数据集进行实验以节省计算成本。为了避免 LUSS 中微调步骤的影响，本文用章节5.2.2.1中介绍的距离匹配评测方案来评测 LUSS 方法。

本文提出的表征学习方法的消融实验 本文在 SwAV[175] 和 PixelPro[171] 上实现了提出的非对比像素到像素（P2P）对齐和由深到浅（D2S）监督。表 5.6a 展示了 PixelPro 由于其缺少 LUSS 任务所需要的类别相关表征能力，效果比 SwAV 差。因此，本文在 PixelPro 中添加了聚类损失 [175] 来针对 LUSS 任务构建一个合理的基准。如表 5.6a，本文的方法在 ImageNet-S_{300} 数据集上分别相较于 SwAV 和 PixelPro 的测试集 mIoU 提升了 2.6% 和 7.6%。具体而言，与图像级方法 SwAV 相比，P2P 对齐在测试集 mIoU 中的增益为 2.2%，在测试集

表 5.5 本文提出的 P2P 对齐和 D2S 监督表征学习策略的消融实验。所有模型训练100 个迭代轮次。D2S3 和 D2S32 分别表示监督网络的第 3 和第 2-3 阶段。

ImageNet-S_{300}	mIoU		Img-Acc		F_β	
	验证集	测试集	验证集	测试集	验证集	测试集
SwAV[175]	22.4	22.6	57.4	57.5	63.5	63.7
+P2P	24.8	24.8	58.4	58.5	64.5	64.8
+P2P-D2S3	25.1	25.2	57.3	57.5	65.0	65.2
+P2P-D2S32	24.8	24.9	56.8	56.6	65.7	66.0
PixelPro[171]	15.5	15.8	44.0	44.3	62.4	62.6
+ 聚类损失函数	20.8	21.3	52.0	52.1	61.5	62.1
+P2P	21.3	22.0	52.2	52.8	61.5	62.1
+P2P-D2S3	22.2	22.8	53.2	53.1	62.2	62.9
+P2P-D2S32	23.0	23.4	53.3	54.3	62.4	63.1

(a) 使用距离匹配评测方案的 LUSS 消融实验。

ImageNet-S_{300}	COCO 实例分割			COCO 物体检测			VOC 语义分割
	AP	AP50	AP75	AP	AP50	AP75	mIoU
SwAV[175]	32.4	52.1	34.6	35.5	54.9	38.6	68.9
+P2P	32.8	52.5	34.9	36.0	55.4	39.1	70.4
+P2P-D2S3	33.5	53.4	35.8	36.7	56.4	39.4	70.8
+P2P-D2S32	33.8	53.7	36.2	37.2	56.6	40.6	70.8
PixelPro[171]	34.7	54.8	37.2	38.2	57.5	41.7	72.8
+ 聚类损失函数	34.9	55.2	37.3	38.4	58.1	41.9	73.3
+P2P	35.3	55.9	37.9	38.9	58.6	42.4	72.3
+P2P-D2S3	35.3	55.9	37.6	38.8	58.6	42.3	73.9
+P2P-D2S32	35.7	56.6	38.3	39.4	59.1	43.1	75.1

(b) 下游任务迁移学习的消融实验。

mIOU 中，它还将使用聚类损失增强后的 PixelPro 提高了 0.5%。与基于 SwAV 和 PixelPro 的基准相比，D2S 监督分别带来 0.4% 和 1.4% 的进一步提升。总之，P2P 对齐有效地增强了图像级方法的像素级表征，D2S 监督丰富了像素级方法的实例级类别表征。P2P 对齐和 D2S 监督分别改进了像素级和图像级表征方法，展示了本文所提策略的鲁棒性。如表 5.7，本文提出的表征学习策略在 ImageNet-S 数据集上也比基准要效果更好。

非对比像素对像素对齐 本文利用非对比 P2P 对齐来增强像素级表征，而不会损害实例级类别表征。本文还比较了不同的像素级对齐策略，包括聚类、对比

和非对比策略。对于聚类和对比损失，本文将两个视图相同位置的像素设置为正样本，其他像素设置为负样本。如表 5.7a，与基准相比这两种像素级对齐策略都具有更高的 F_β，显示出形状表征质量的提升。然而，由于同一物体中像素之间的语义差异，聚类和对比度损失的 mIoU 和 Img-Acc 性能较差。相比之下，由于保持了属于同一语义实例的像素的表征一致性，本文提出的非对比 P2P 对齐在 mIoU 和 Img-Acc 中的性能优于基准方法。本文也在表 5.7b中分析了映射 $\mathbf{M_p}(\tilde{\mathbf{z}})$ 的有效性。有映射 $\mathbf{M_p}(\tilde{\mathbf{z}})$ 的 P2P 对齐可以实现更好的 Img-Acc 性能，因为 $\mathbf{M_p}(\tilde{\mathbf{z}})$ 减小了像素级表征对类别相关表征的干扰。

由深到浅的监督 D2S 监督利用最后阶段的高质量特征来监督早期特征。表 5.7c比较使用相同或最后阶段的特征作为浅层的监督。本文观察到，这两种设置都比基准有所提升，由深到浅的监督在 mIoU 和 Img-Acc 上优于同一阶段监督。默认情况下，本文使用从一个视图的深层特征来监督另一个视图的浅层特征。如表 5.7d，本文研究了使用来自同一视图的特征对 D2S 监督的影响。交叉视图监督略优于相同视图监督。本文观察到，相同视图监督的训练损失低于交叉视图监督。本文认为，相同视图监督会导致训练过拟合，影响测试性能。D2S 监督可以用于网络不同浅层阶段的多个特征。如表 5.6a，本文研究了监督网络不同阶段特征对基于 SwAV 和 PixelPro 的方法的影响。本文发现不同的方法需要监督不同的阶段来获得最优结果，例如在 SwAV 中监督 3 和 2 阶段比监督 3 阶段效果差，但 PixelPro 能从更多的对浅层的监督中获益。本文通过消融实验决定 D2S 监督的阶段。

评测无监督学习方法 为了分析无监督学习方法在 LUSS 任务中的表征能力，本文对包括对比、非对比、聚类和像素级等有代表性的方法进行分类并进行测试。如表 5.7，图像级方法在 mIoU、Img-Acc 和 F_β 上比像素级方法有更明显的优势。像素级方法过于关注像素级特征的差异，导致同一物体实例中像素之间的语义差异较大。相比之下，图像级方法提供了一致的实例级类别相关表征，因为这些方法的损失函数重点优化模型对图像之间的区分能力。然而，像素级表征对 LUSS 任务至关重要，因为本文提出的非对比 P2P 对齐方法相比于图像级方法 SwAV 有相当大的提升。本文发现聚类方法在 Img-Acc 上比对比和非对比方法都要高，但在形状相关的 F_β 上要低。与对比和非对比方法相比，聚类方法鼓励使用聚类中心进行更强的类别相关表征学习。但由于聚类方法中一幅图像的所有像素的特征都接近类别质心，因此主类别与“其他”类别之间的表征差

表 5.6 使用距离匹配评测方案在 ImageNet-S_{300} 测试集对 P2P 对齐和 D2S 监督策略的消融实验。

ImageNet-S_{300}	mIoU	Img-Acc	F_β
SwAV baseline	22.6	57.5	63.7
+ 基于聚类损失的 P2P	21.2	51.8	66.4
+ 基于对比损失的 P2P	18.0	46.4	64.6
+ 基于非对比损失的 P2P	24.8	58.5	64.8

(a) P2P 对齐的不同损失函数形式。

ImageNet-S_{300}	mIoU	Img-Acc	F_β
SwAV 基线	22.6	57.5	63.7
未使用 $\mathbf{M_p}(\tilde{\mathbf{z}})$ 的 P2P	24.6	57.1	64.9
使用 $\mathbf{M_p}(\tilde{\mathbf{z}})$ 的 P2P	24.8	58.5	64.8

(b) $\mathbf{M_p}(\tilde{\mathbf{z}})$ 映射在 P2P 对齐的作用。

ImageNet-S_{300}	mIoU	Img-Acc	F_β
PixelPro+P2P （基线）	22.0	52.8	62.1
+ 同阶段监督	22.6	52.9	63.1
+ 由深到浅监督	23.4	54.3	63.1

(c) D2S 监督中由深到浅层监督与同层监督的对比。

ImageNet-S_{300}	mIoU	Img-Acc	F_β
PixelPro+P2P （基线）	22.0	52.8	62.1
+ 同视图监督	23.1	53.9	63.2
+ 跨视图监督	23.4	54.3	63.1

(d) D2S 监督中使用同一个视图和不同视图的特征进行监督的对比。

异会减弱。图像级有监督方法比聚类方法有更好的类别中心，但在 F_β 上更差。这些结果解释了为什么类别指标更好的聚类方法有更差的 F_β。

类别在 LUSS 任务中作用 为了回答这个问题，本文使用经过图像级有监督训练的模型作为基准。如表 5.7，在 mIoU 指标上，有监督模型的性能优于无监督模型。此外，它在图像级分类精度方面大大优于无监督模型。相反，它在形状相关指标，例如 F_β，比大多数无监督方法要差。这些结果表明，类别特征确实有助于 LUSS 任务。然而，形状特征不能仅通过类别表征学习来得到。

表 5.7 本文的无表征学习增强方法与其他无监督表征学习方法在距离匹配评测方案下的性能比较。本文的 $_{s/p}$ 分别表示使用 SwAV [175] 和 PixelPro [171] 作为公式 (5.5) 中的 L_e。所有模型训练100 个迭代轮次。有监督模型表示使用图像级有监督预训练权重初始化模型。

LUSS 任务	mIoU		Img-Acc		F_β	
	验证集	测试集	验证集	测试集	验证集	测试集
ImageNet-S_{300}						
有监督模型	33.8	33.9	80.4	81.5	60.0	60.0
对比学习						
SimCLR[167]	12.5	12.6	37.7	38.4	63.7	64.0
MoCov2[296, 49]	12.4	12.4	40.3	40.3	64.1	64.4
AdCo[297]	21.1	21.5	55.1	54.8	64.9	65.5
非对比学习						
BYOL[170]	13.4	13.4	38.3	38.0	64.0	64.4
SimSiam[172]	20.1	20.3	56.9	57.5	65.5	66.0
聚类						
PCL[174]	17.4	17.9	48.4	48.0	63.0	63.3
SwAV[175]	22.4	22.6	57.4	57.5	63.5	63.7
PASS$_s$	25.1	25.2	57.3	57.5	65.0	65.2
像素级						
DenseCL[197]	13.9	13.8	36.4	36.8	63.7	63.7
PixelPro[171]	15.5	15.8	44.0	44.3	62.4	62.6
PASS$_p$	23.0	23.4	53.3	54.3	62.4	63.1
ImageNet-S						
有监督模型	30.0	29.8	75.9	76.6	58.7	58.7
PixelPro[171]	7.7	7.5	26.9	26.5	61.8	61.8
PASS$_p$	9.8	9.8	29.4	29.6	61.1	61.3
SwAV[175]	15.1	15.1	43.5	43.3	64.2	64.3
PASS$_s$	15.6	15.6	43.1	42.9	64.3	64.6

5.4.4 标签生成和微调的有效性分析

本文使用章节5.2.2.1节中描述的完全无监督评测方案评测本文提出的基于像素注意力的标签生成和微调方案的有效性。除非另有说明，否则本节使用 ImageNet-S_{50} 子集进行消融实验。

像素标签生成的效果 本文将提出的基于像素注意力的像素标签生成方法与图

表 5.8 使用完全无监督评价方案在 ImageNet-S_{50} 测试集的像素级标签生成和微调步骤的消融实验。

ImageNet-S_{50}	mIoU	Img-Acc	F_β
图像级方法	26.9	57.6	53.0
像素级方法	12.7	37.4	32.9
像素注意力	29.3	65.5	49.0
像素注意力 $^\tau$	29.2	61.7	52.3

(a) 不同伪标签生成方式的对比。$^\tau$ 代表使用图像级方法的推理策略。

	ImageNet-S_{50}	ImageNet-S_{300}	ImageNet-S
图像级方法	2.8×10^0	8.9×10^1	7.5×10^2
像素级方法	3.2×10^2	4.6×10^4	4.1×10^5
像素注意力	2.8×10^0	8.9×10^1	7.5×10^2

(b) 不同伪标签生成方法的聚类时间（秒）。

ImageNet-S_{50}	mIoU	Img-Acc	F_β
共享注意力图	28.4	64.3	48.8
非共享注意力图	29.3	65.5	49.0

(c) 对输出特征是否共享像素注意力图。

ImageNet-S_{50}	mIoU	Img-Acc	F_β
未微调模型	26.0	63.8	44.7
已微调模型	29.3	65.5	49.0

(d) 微调步骤对 PASS 方法的性能影响。

像级和像素级标签生成方法进行了比较。本文首先简要介绍使用图像级和像素级标签生成方法的标签生成和微调过程。图像级别标签生成方法在池化图像级别特征向量上聚类出 C 个类别，并将图像级别标签分配给每个图像。在微调期间，使用图像级标签监督全连接（FC）层。为了获得像素级分割掩码，在推理阶段 FC 层被替换为 1×1 卷积层。由于缺少“其他”类别，本文采用弱监督语义分割方法广泛使用的基于类激活映射的掩码生成方法来生成最终的分割掩码。在大规模 ImageNet-S 数据集上，像素级特征向量聚类的成本太高。因此，本文在 ImageNet-S_{50} 数据集上实现像素级方法来进行比较。本文使用像素级特征向量对 $C+1$ 类别进行聚类，并使用像素级标签对其进行微调。如表 5.9a，本文

提出的基于像素注意力的标签生成方法优于图像级和像素级方法，且具有相当大的优势。

聚类时间比较 本文将基于像素注意力的标签生成的聚类时间与表 5.9b中的其他两种标签生成方法进行了比较。本文的方法与图像级方法具有相同的聚类时间，因为它们都使用图像级特征向量。因为训练集中有大量的像素，即使使用低分辨率 7×7 的输出特征图的像素级方法也比本文的方法慢得多。在完整的 ImageNet-S 数据集上进行聚类时，像素级方法的时间大约为114 小时，这对于实际使用来说是不可接受的。

输出特征共享/非共享像素注意力图 默认情况下，本文为每个输出特征通道生成一个的像素注意力图。本文还研究了对所有通道使用一个共享的像素注意力图的效果。在表 5.9c中的结果说明对每个通道使用非共享像素注意力图可以获得更好的性能。本文在图 5.10中可视化了不同通道的像素注意力图。大多数通道侧重于语义区域，而少数通道突出显示背景区域。此外，每个像素注意力图的关注区域也不相同，这也说明了非共享像素注意力的有效性。

微调的效果 本文基于像素注意力的标签生成方法可以直接生成像素级分割掩码。本文比较了微调步骤前后的性能，以验证本文的 LUSS 方法中微调步骤的效果。如表 5.9d，微调方法将测试集 mIoU 提升了 3.3%，说明生成的像素级标签仍然有噪声，微调进一步提高了语义分割质量。

5.4.5 下游任务迁移学习的自适应性

本节研究 LUSS 任务学习的表征是否能够自适应地迁移到像素级下游任务，例如语义分割、实例分割和目标检测。本文还比较了不同表征学习方法对 LUSS 和下游任务的影响。为了公平比较，除非另有说明，不同的表征学习方法均使用 ResNet-50 [15] 网络在 ImageNet-S_{300} 或 ImageNet-S 数据集上预训练100 个迭代轮次。

实例分割和目标检测 本文使用 MaskRCNN [124] 作为实例分割和目标检测的检测器。模型在 COCO17 [55] 训练集上训练，并且在验证集上评测。依照通用设置 [171, 124, 49]，本文加载在不同表征学习方法上预训练的 ResNet-50 的权重，并应用 1× 训练计划。如表 5.9所示，本文基于 SwAV [175] 和 PixelPro [171] 验证了本文提出的非对比 P2P 对齐和 D2S 监督表征学习策略的有效性。本文首先比较在 ImageNet-S_{300} 数据集上预训练模型的性能。在实例分割中，本文的方法在 mAP 指标上相较 SwAV 和 PixelPro 分别提升 1.4% 和 1.0%。类似地，

表 5.9 在对 ImageNet-S_{300} 和 ImageNet-S 数据集预训练的无监督表征学习方法的迁移学习性能比较。所有模型训练100 个迭代轮次。本文的 $_{s/p}$ 分别代表使用 SwAV [175] 和 PixelPro [171] 作为公式 (5.5) 中的 L_e。有监督模型表示通过图像级有监督预训练权重对模型进行初始化。

迁移学习	COCO 实例分割			COCO 物体检测			VOC 语义分割
	AP	AP50	AP75	AP	AP50	AP75	mIoU
ImageNet-S_{300}							
有监督模型	34.7	55.3	37.0	38.4	58.1	42.0	72.6
对比学习							
SimCLR[167]	31.9	51.1	34.1	35.0	53.7	38.2	66.4
MoCov2[296, 49]	33.7	53.6	36.1	37.1	56.3	40.3	67.8
AdCo[297]	34.3	54.3	36.7	37.9	57.2	41.5	70.0
非对比学习							
BYOL[170]	32.1	51.6	34.2	35.1	54.2	38.2	65.8
SimSiam[172]	33.7	53.3	36.2	36.9	56.0	40.3	61.1
聚类							
PCL[174]	34.3	54.4	36.9	37.8	57.0	41.3	69.6
SwAV[175]	32.4	52.1	34.6	35.5	54.9	38.6	68.9
PASS$_s$	33.8	53.7	36.2	37.2	56.6	40.6	70.8
像素级							
DenseCL[197]	33.7	53.4	36.2	37.0	56.2	40.4	67.7
PixelPro[171]	34.7	54.8	37.2	38.2	57.5	41.7	72.8
PASS$_p$	35.7	56.6	38.3	39.4	59.1	43.1	75.1
ImageNet-S							
有监督模型	36.6	57.5	39.4	40.3	60.5	44.0	76.4
SwAV[175]	34.4	55.0	36.8	37.8	58.0	41.1	73.0
PASS$_s$	35.3	56.0	37.8	38.9	58.8	42.3	75.3
PixelPro[171]	35.9	56.6	38.6	39.5	59.2	43.1	73.9
PASS$_p$	36.5	57.4	39.1	40.2	60.3	44.1	76.1

目标检测的 mAP 分别提升 1.7% 和 1.2%。这些结果证明，本文用于 LUSS 任务的表征学习方法在实例分割和目标检测任务的不同基准上具有稳定的性能增益。像素级方法 PixelPro 优于例如 SwAV、AdCo 和 SimSiam 等其他图像级方法，这证明像素级方法对这两个像素级下游任务具有更强的迁移能力。当使用完整的 ImageNet-S 数据集进行预训练时，本文的方法仍然优于基线，例如基于

PixelPro 的本文方法在实例分割和目标检测任务中分别获得 0.6% 和 0.7% 的 mAP 增益。

语义分割 本文还使用基于 ResNet-50 的 Deeplab V3+ [238] 网络将预训练好的模型权重迁移到 PASCAL VOC 数据集 [237] 上的语义分割任务。模型在 Pascal VOC SBD 训练集上训练并在验证集上评测。在 ImageNet-S_{300} 数据集上进行预训练时，本文的方法在 mIoU 上比 SwAV 和 PixelPro 分别高了 1.9% 和 2.3% 。使用 ImageNet-S 预训练模型时性能在 mIoU 上分别提升了 2.3% 和 2.2%。像素级方法 PixelPro 与其他图像级方法相比具有明显优势，表明像素级表征对于语义分割至关重要。基于对比学习的方法虽然仍是图像级方法，但在语义分割方面优于聚类和非对比学习方法。

LUSS 与迁移学习的关系 本文在表 5.7和表 5.9中分别比较了表征学习方法在 LUSS 和下游任务上的性能。与图像级方法相比，SwAV 聚类方法由于分类精度高，在 LUSS 任务中具有更好的性能。在下游任务上，SwAV 不如许多在 LUSS 任务上性能较差的方法。例如，在下游实例分割任务中，对比学习方法 MoCov2 的 mAP 比 SwAV 提升 1.3%，但 LUSS 任务的 mIoU 有 10% 的差距。这一观察结果与 [290] 的发现一致，即对比方法可以更好地学习低层特征，从而有利于像素级下游任务。在下游任务中，像素级方法 PixelPro 明显优于图像级方法。但在 LUSS 任务中，它的性能比许多图像级方法都差。像素级方法学习下游任务的可区分像素级表征，但缺乏 LUSS 任务需要的足够的类别相关表征。在同一类方法内比较，大多数在 LUSS 任务中表现良好的方法在下游任务中都能取得更好的性能。因此，LUSS 和下游任务需要不同的表征，但都会从高质量的表征中受益。本文证明了本文提出的 P2P 对齐和 D2S 监督在 LUSS 任务（表 5.6a）和下游任务（表 5.6b）的有效性。这两种策略都提高了 LUSS 和下游任务的性能，表明了本文所提出的表征学习方法的通用性。

5.4.6 LUSS 的拓展和应用

本节介绍 LUSS 任务以及其评测方案在大规模半监督语义分割和图像级有监督主干模型评测的应用。

大规模半监督语义分割 半监督语义分割需要使用一小部分标记数据和许多未标记数据进行训练。在 ImageNet-S 数据集的 1% 有像素级标注的训练图像上微调训练的 LUSS 模型，可实现半监督语义分割，这也是章节5.2.2.1中介绍的 LUSS 半监督评测方案。本文遵循在章节5.4.1中介绍的微调步骤的训练策略，唯

表 5.10 使用 ImageNet-S_{50}/ImageNet-S 数据集的半监督语义分割（半监督评测方案）。本文的 $_{s/p}$ 分别代表使用 SwAV [175] 和 PixelPro [171] 作为公式 (5.5) 中的 L_e。有监督模型是指使用图像级监督预训练初始化的模型。

半监督学习	mIoU		Img-Acc		F_β	
	验证集	测试集	验证集	测试集	验证集	测试集
ImageNet-S_{300}						
有监督模型	27.7	27.5	61.1	62.3	64.3	64.9
SimCLR[167]	12.7	12.6	34.4	34.8	59.1	59.6
BYOL[170]	10.5	10.6	30.1	30.5	58.5	59.0
MoCov2[296, 49]	12.6	12.3	33.0	32.5	59.2	59.4
DenseCL[197]	16.2	16.0	34.9	35.7	61.0	60.9
AdCo[297]	19.6	19.6	45.4	45.4	63.8	63.8
PCL[174]	17.3	17.4	41.7	41.8	61.7	61.9
SwAV[175]	23.0	23.3	51.2	51.5	64.0	64.0
PASS$_s$	25.7	25.7	52.3	52.8	65.5	66.0
PixelPro[171]	23.3	23.4	49.0	48.9	66.0	66.6
PASS$_p$	29.7	29.8	56.9	56.9	68.1	68.5
ImageNet-S						
有监督模型	25.7	25.0	57.3	57.4	66.3	66.7
PixelPro[171]	16.0	15.6	36.0	36.2	66.2	66.5
PASS$_p$	18.9	18.6	40.9	41.3	68.0	68.4
SwAV[175]	18.2	17.9	42.8	43.2	66.0	66.2
PASS$_s$	19.4	19.2	43.3	43.4	66.6	66.9

一不同的是此处模型是使用 GT 标签用30 个迭代轮次训练的。半监督语义分割结果见表 5.10。本文提出的 PASS 方法优于 SwAV 和 PixelPro 基准，分别在 ImageNet-S_{300} 和 ImageNet-S 数据集上有可观的提升。本文基于 PixelPro 的方法甚至超过了 ImageNet-S_{300} 数据集上的图像级有监督模型。在半监督范式中，PixelPro 的性能与 SwAV 相似，但 SwAV 在距离匹配评测结果方面比 PixelPro 有很大优势（见表 5.7）。本文推测这是因为具有像素级 GT 标签的微调模型使模型需要较少的自学习得到的类别相关表征能力。

自适应网络架构在大规模语义分割的应用 本节将测试前两章节介绍的自适应多尺度表征主干网络和自适应的感受野搜索对大规模语义分割的作用。本文使用距离匹配评测方案在 ImageNet-S 测试集上对模型的 mIoU 进行基准测试，见表 5.11。作为参考，本文还提供了这些模型在 ImageNet-S 数据集上的

表 5.11 在 ImageNet-S 测试集上使用距离匹配评测方案的有监督主干模型的 mIoU 结果。Top-1 准确率是在 ImageNet-S 测试集上的分类准确率。* 表示模型使用 ImageNet-S 的训练集进行半监督微调训练。

	Top-1 准确率	mIoU
ImageNet-S		
ResNet-50[15]	83.6	29.8
ResNet-101[15]	84.3	31.4
DenseNet-161[16]	84.3	29.8
Inception V3[298]	77.7	29.9
ResNeXt-50[18]	84.4	32.6
ResNeXt-101[18]	85.5	34.8
EfficientNet-B3[299]	85.3	32.3
Res2Net-50	84.8	35.7
Res2Net-101	85.6	37.2
ConvNeXt-T* [283]	-	45.1
RF-ConvNeXt-T* （单分支）	-	46.2
RF-ConvNeXt-T* （多分支）	-	47.0

top-1 分类精度。本文观察到，图像级别的 top-1 精度并不总是与 mIoU 保持一致，这表明具有良好类别表征的模型可能不擅长形状表征。本文第三章介绍的尺度自适用的多尺度网络结构 Res2Net 在类别和形状表征上都表现较优。为了进一步评测 ImageNet-S 数据集的性能在多大程度上受益于更好的感受野组合，本文进而测试了基于第四章提出的自适应感受野搜索 RF-Next 结合 ConvNeXt 模型的表现，命名为 RF-ConvNeXt。该模型通过更合适的感受野组合增强了 ConvNeXt [283] 模型。RF-ConvNeXt 展现了较高的语义分割性能，说明一个好的感受野组合对 ImageNet-S 数据集下的语义分割非常有必要。

5.5 总结

本章节提出了一个新的数据自适应的大规模无监督语义分割问题，以便于在具有丰富多样性和大规模数据的现实环境中进行自适应地语义分割。本章为该任务提供了一个测试基准，其中包含具有高度多样性的大规模数据、明确的任务目标和充分的评测方法。本章提出的大规模无监督语义分割方法可以在没有人工标注监督的情况下，从大规模数据中学习类别和形状表征，最终为图像的像素分配类别标签。该 LUSS 方法包含增强的表征学习和像素注意力辅助的

像素级标签生成策略。本章用多种评测方法评测了本文的方法，并揭示了 LUSS 对如语义分割等像素级下游任务的潜力。本章也验证了第三章提出的尺度自适用表征网络结构和第四章提出的感受野自适用搜索对该任务的帮助。此外，本章对无监督表征学习方法进行了评测和分析，总结了 LUSS 面临的挑战和可能的研究方向。大规模无监督语义分割虽然取得一定效果，但其性能相较于传统有监督语义分割仍有一定差距，因此需要更强的通用表征学习技术和精细高效的聚类技术来进一步提升该任务的性能。

第六章 模型自适应可持续自监督学习

复杂的场景导致训练数据和模型体积的激增，进而大幅地增加模型的训练开销。因此要求新模型能够有效利用现有基础模型的知识储备自适应地学习对新场景的表征，从而降低训练的计算开销。为自适应地利用先验视觉模型的知识，本文提出一个可持续的自监督表征学习框架，能够自适应地从各种现有的预训练模型中学习，从而以显著更低的训练开销学习得到更加强大的新模型。该可持续预训练方案可以加快模型学习速度，并提升模型视觉感知性能，朝着绿色可持续的视觉感知学习迈出了探索性的一步。章节6.1简要介绍了本文提出的可持续自监督学习概念和基本方案。章节6.2介绍了本文提出的用于实现可持续自监督学习的一个基于掩码重建的目标增强条件化预训练机制。章节6.3和章节6.4验证了该方法针对现有模型的自适应学习能力。章节6.5对本章进行总结。

6.1 可持续自监督学习简介

自监督学习（SSL）虽然在视觉场景的表征上取得巨大成功，但由于训练数据的增加和模型复杂度的提升，SSL 正朝着需要越来越大的训练成本的方向发展。此外，大部分的 SSL 预训练基础模型虽然含有视觉感知信息，但没有被充分利用。因此，本节将探索构建一个能够有效利用先验模型知识进行学习的可持续的 SSL 框架。就像人类社会中的知识是在代代相传中逐渐扩充的一样，本文试图让新的 SSL 模型继承先前的预训练 SSL 模型的知识，并进一步获得更强的表示学习能力。以此实现的“可持续”SSL 相比从头开始训练一个新的 SSL 模型提高了学习效率和表征能力。图 6.1给出可持续 SSL 的示意图，其中本文将待训练的新 SSL 模型简称为新模型，将预训练 SSL 模型称为基模型。为了超越基模型，在可持续 SSL 中，新模型不仅要利用隐含的基模型知识，而且要补充基模型中缺乏的知识。不同于有监督学习需要标签实现自我训练 [300, 301]，可持续 SSL 学习过程遵循完全自监督的范式。

本文通过构建一种能高效地学习并超越现有的预训练 SSL 模型的目标增强的条件掩码重建（Target-enhanced Conditional Mask Reconstruction，TEC）训练策略，向可持续 SSL 迈出了探索性的一步。为了实现这一具有挑战性的目标，

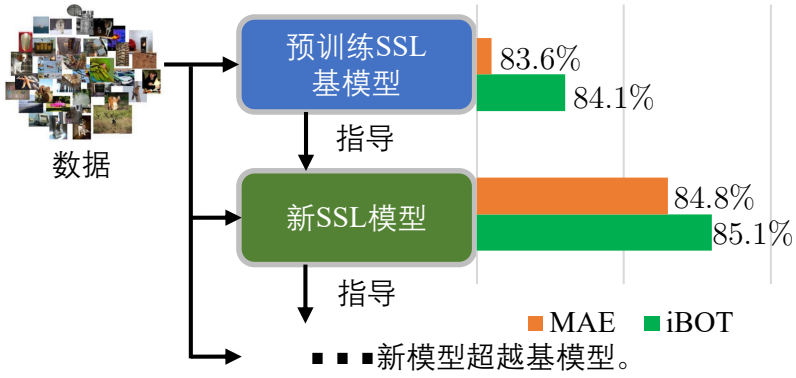


图 6.1 可持续自监督学习的概念。

该策略鼓励新模型不仅学习基模型的知识，还学习更多与语义相关的新知识。因此，本文选择了一种基于掩膜重建 [52] 的 SSL 方案来训练新模型，其中基模型从完整的输入图像中生成重建目标，新模型试图从随机掩盖后的图像输入中预测该重建目标。通过此代理任务，新模型必须学习输入图像的完整语义及区域间关系，以便新模型能够从不完整输入中推理出所需的完整信息。如图 6.2所示，iBOT [53] 预训练策略下的 Vision Transformer（ViT）[302] 模型的注意力图遗漏了一些语义区域，如耳朵，而以 iBOT 预训练模型为基模型经过 TEC 训练得到的 ViT 捕获了所有区域的语义，并有效区分输入图像的不同组成部分。由于 TEC 具有更强大的捕获综合语义的能力，因此它有助于实现具有挑战性的可持续 SSL，并且可以为下游任务提供丰富而灵活的语义表征。

然而，不同的 SSL 基模型由于其不同的训练目标和训练策略而可能具有不同的属性，例如 iBOT 预训练模型具有更多的类别语义特征，而 MAE [52] 预训练模型具有更多的图像细节特征。因此，从基模型中构建高质量且兼容的重建目标非常重要，这样新模型才能自适应地学习得到更全面的信息。本文认为好的重建目标应该揭示特征的空间语义关系，例如展现车轮和车身的关系，这有利于新模型学习能够适用各种下游任务的通用关系特征。为此，本文提出通过使用两个互补的重建目标来提高基模型生成的目标质量: a）空间维度归一化的重建目标通过对基模型输出特征沿空间维度进行归一化，从而增强特征区域间的关系属性; b）利用基模型中语义丰富的词符（Token）对应的注意力图作为重建目标，从而过滤掉噪声并建立整个图像与语义丰富的区域之间的联系。为了自适应地兼容不同基模型的重建目标，本文在新模型中引入了条件适配器，以便新模型的预测结构可以适应具有不同属性的各种基模型。在给定基模型重建

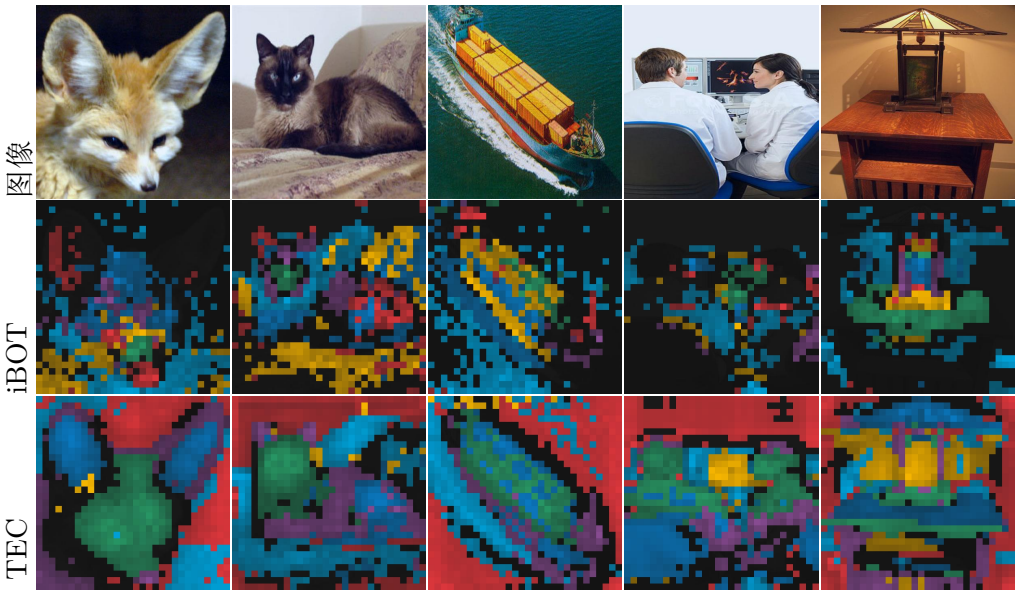


图 6.2 ViT 自注意力模块的可视化图。每种颜色代表一个自注意力头的类别词符对应的注意力图。黑色表示没有自注意力头关注此区域。x

目标的情况下，适配器可以有条件地激活并调整新模型的中间层特征，从而更有效地预测重建目标。这些适配器在预训练完成后被丢弃，但如果保留的话可以用于轻量化微调 [303, 304]。

如图 6.3所示，在 ImageNet 上，本文提出的的 TEC 在无需任何额外训练数据和使用相同网络架构的前提下相比 MAE [52] 和 iBOT [53] 等 SSL 基模型的性能有显著改善。例如，以 1600 迭代轮次的 iBOT 为基模型，只需 800 个迭代轮次的 TEC 提高了 1.0% 分类准确率。此外，本文还发现 TEC 可以显著加快 SSL 模型的学习过程，节省训练成本。例如，基于 300 个迭代轮次的 MAE 基模型，随机初始化训练 TEC 仅 100 个迭代轮次即可优于训练 1600 个迭代轮次的 MAE。该方法迈出了探索可持续 SSL 的第一步，本文希望其在未来激发更多的工作以绿色的方式可持续地改进 SSL。

6.2 目标增强的条件掩码重建可持续学习

6.2.1 总体框架

本文所提出的目标增强的条件掩码重建方法的总体框架如图 6.4所示。TEC 依照 [52, 51] 使用 ViT [302] 网络架构。在掩模重建框架 [52] 下，如图 6.4所示，TEC 由待预训练的新 ViT 编码器、用于条件预训练的条件适配器、用于重建目

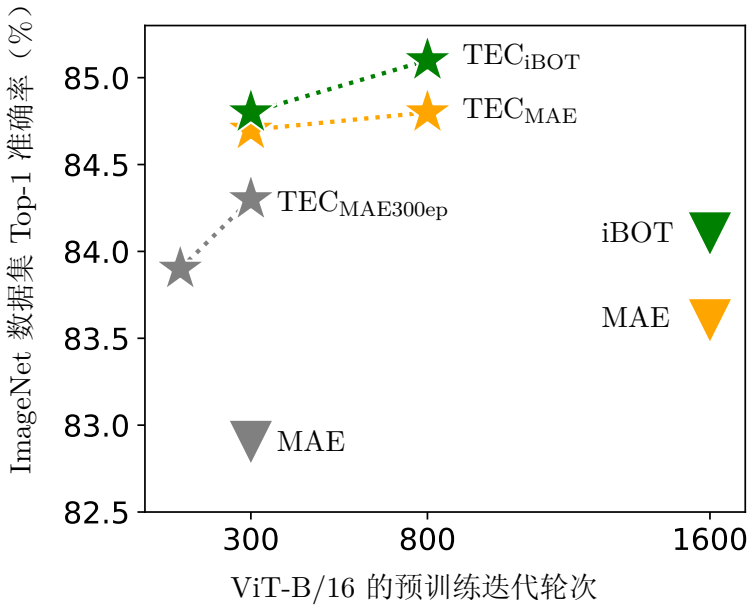


图 6.3 ImageNet-1K 上的 Top-1 准确率。TEC 预训练的新模型与基模型颜色相同。

标预测的多目标解码器、预训练的 SSL 基模型和一个用于从基模型构建特征关系增强重建目标的目标增强模块组成。具体来说，基模型是 SSL 预训练的 ViT 编码器，例如 MAE [52]，并用于生成完整图像的隐式语义特征。然后，目标增强模块对隐式语义进行增强，构建两个互补的重建目标作为新模型的监督。配备适配器的新 ViT 编码器得到掩码图像输入并生成适配过的隐式语义特征，然后将特征输入多目标解码器以预测基模型提供的重建目标。在预训练之后，新 ViT 编码器被保留用于下游任务，而其他部分被移除。下面，本文将在章节6.2.2中介绍通过适配器辅助的条件预训练来帮助新模型有效地预测基模型目标，并在章节6.2.3中详细介绍用于生成高质量基模型重建目标的目标增强模块。

6.2.2 条件预训练

如前所述，基模型通常具有不同的属性，例如，iBOT 中含有更多全局类别语义，而 MAE 中更多的是局部细节。因此，新模型的重建预测应该与任意给定的基模型兼容。为了解决图像像素重建的类似问题，文章 [199, 211, 305] 通过试错的方式从编码器的中间层手动选择某些特征，以更好地与图像像素目标对齐。然而，从某些固定层中手动选择与不同基模型兼容的特征几乎是不可能的，因为它们可能具有不同的属性。因此，为了更好地预测基模型目标，新模型必须

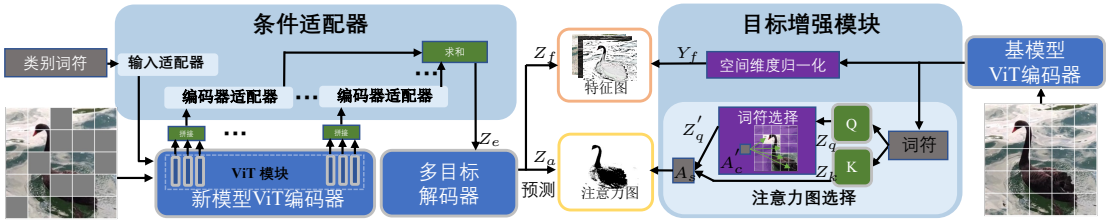


图 6.4 本文提出的 TEC 的总体框架。TEC 中预训练的 SSL 基模型生成区域间关系增强的重建目标，即空间维度归一化特征和语义相关的注意力图。新的 ViT 编码器送入掩码图像和输入适配器增强的类别词符，然后将其生成的特征依次传递给编码器适配器和多目标解码器，以预测基模型给定的重建目标。

对给定的 SSL 基模型具有条件适应能力。

给定一个固定的预训练模型，轻量化微调方案将具有少量参数的可训练额外模块引入预训练模型，使其适应视觉 [303, 304] 和自然语言处理 [306, 307, 308] 域的下游任务。例如，提示（Prompt）方案 [307, 308, 303] 将可学习的输入词符（例如，类别词符）与特征词符连接起来，以激活固定的 ViT 模型的某些适用于特定的下游任务的语义特征。此外，将轻量级适配器模块（例如，MLP [306, 304] 和残差词符 [309]）整合到固定模型中，可以对模型的中间层特征进行调制，从而预测下游任务所需的特征。受这些轻量化微调方案的启发，本文将适配方案引入到预训练阶段，通过为新模型配备条件适配器来自适应地处理基模型的多样性。由于本文的适配器仅用于预训练，并将在微调期间删除，因此它们不会增加额外的推理成本。实际上，表 6.4显示了将这些适配器在推理阶段保留可以增强模型的轻量化微调能力。下面，本文将介绍如何将适配器，即输入和编码器适配器，应用到新模型的 ViT 编码器中。

输入适配器 对于 ViT 网络，通常将类别词符与输入特征词符连接起来，以学习整个输入的全局语义。由于提示方案证明了类别词符的调整能力，本文提出通过增加输入适配器进一步增强类别词符的特征调整能力。如图 6.5所示，由一个小的两层 MLP 层构成的输入适配器增强了类别词符的表示能力，使类别词符能够更好地根据基模型目标激活新模型中的特征。具体来说，使用 MLP 层对 ViT 的类别词符 $T \in \mathbb{R}^C$ 进行处理，得到增强的类别词符 $T^{'} \in \mathbb{R}^C$：

$$T^{'} = \mathbf{MLP}(T),$$

其中 C 为特征向量维度。在预训练期间，$T^{'}$ 被附加到特征词符中。**MLP** 增强

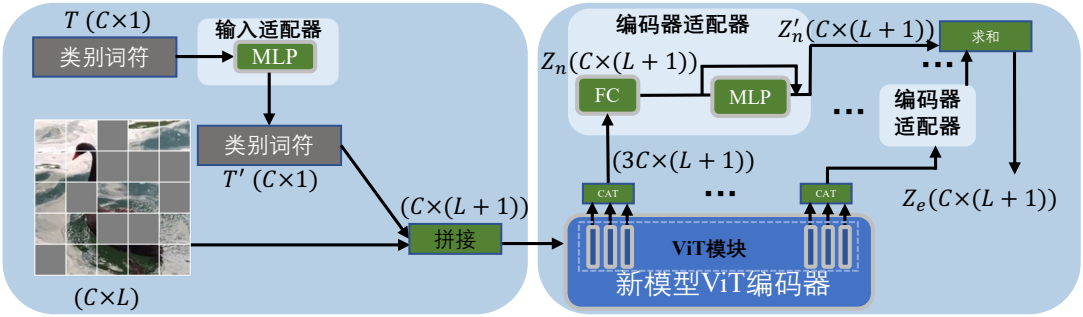


图 6.5 条件预训练输入适配器和编码器适配器的细节图。

了 T 的表示能力，使新模型能够更好地预测基模型目标。在推理时，由于所有输入样本共享 $\mathbf{MLP}(T)$，所以可以提前计算得到新的类别词符 T'，这意味着 $\mathbf{MLP}(T)$ 不会带来额外的推理开销。

编码器适配器 为了调节新模型中的中间层特征，使其能够适应基模型目标，本文在预训练阶段应用了一个带有残差连接的简单 MLP [304] 作为 ViT 的编码器适配器。由于本文希望在预训练后去除适配器以获得更高的推理效率，因此本文需要在去除适配器后保持编码器网络拓扑结构不变。因此，本文将适配器的输入放在编码器的中间，并在编码器的输出位置合并所有的适配器输出。如图 6.5所示，从编码器每个中间层得到特征 $X=\{X_i, i=1,\ldots,D\}$，其中 D 为编码器中间层的编号，本文首先将它们统一分成 N 组，其中每组默认包含 3 个中间层。在第 n 组中合并来自所有该组中间层的特征：

$$Z_n = \mathbf{FC}(\mathbf{Concat}(X_i, ..., X_j)).$$

然后将特征 Z_n 输入到适配器中，得到一个整体特征 Z_e：

$$Z_n^{'} = Z_n + \mathbf{MLP}(Z_n), \qquad Z_e = \sum_{n=1}^{N} Z_n^{'}, \tag{6.1}$$

其中 $\mathbf{MLP}$ 是一个两层全连接的小 MLP。然后将适配器调整后的特征输入多目标解码器来预测基模型目标。

6.2.3 区域间关系增强的重建目标

为了更好地利用基模型的知识实现可持续 SSL，本文的目标增强模块构造了两个增强区域间关系的互补目标：1）空间维度归一化的特征级目标来增强特征词符之间的关系; 2）语义相关的注意力图用于学习高语义特征词符与其他特

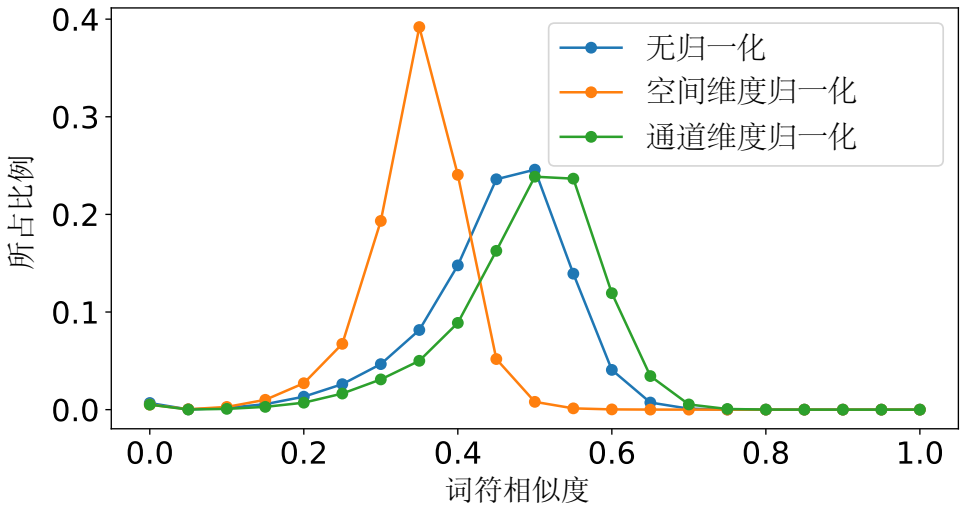


图 6.6 MAE 预训练模型的词符相似度分布。

征词符之间的关系。特征级目标揭示了特定词符的语义，而注意力图更关注特征词符之间的关系。

空间维度归一化的特征级目标 在给定一个基模型的输出特征作为目标，本文提出将该特征沿空间维度进行归一化，以增强空间区域间关系。具体来说，对于一个输入，假设其基模型目标为 $Y \in \mathbb{R}^{L\times C}$，其中 L 和 C 分别表示词符数量和通道维度。然后本文沿着词符维度归一化 Y：

$$Y_f = (Y - \mu_L)/\sigma_L, \tag{6.2}$$

其中，μ_L 和 σ_L 分别为沿词符维度的均值和方差。对于掩膜图像建模（MIM），这种词符维度归一化比广泛使用的在通道维度上的特征归一化 [218, 205, 210] 能更好地增强词符之间的空间关系。这是因为从图 6.6中可以看出，使用 MAE 预训练得到的模型的表征可能表示了图像较多的全局语义，因此不同词符的基模型特征值较为相似。这导致 MAE 预训练模型的特征不能很好地揭示这些词符之间的空间关系。因此，由于掩码词符的特征与可见词符的特征具有较高的相似性，模型可以很容易地重建掩码词符的特征目标。通道维度归一化只考虑了词符内的均值和方差，很难增强词符之间的关系差异。实际上，如图 6.6所示，通道维度归一化甚至扩大了词符之间的相似性。而词符维度归一化则保证了每个通道内的值有明显的差异，如图 6.6所示，通过明显降低词符之间的相似性可以增强词符之间可能存在的空间关系。此外，从表 6.7c可以看出，本文提出的

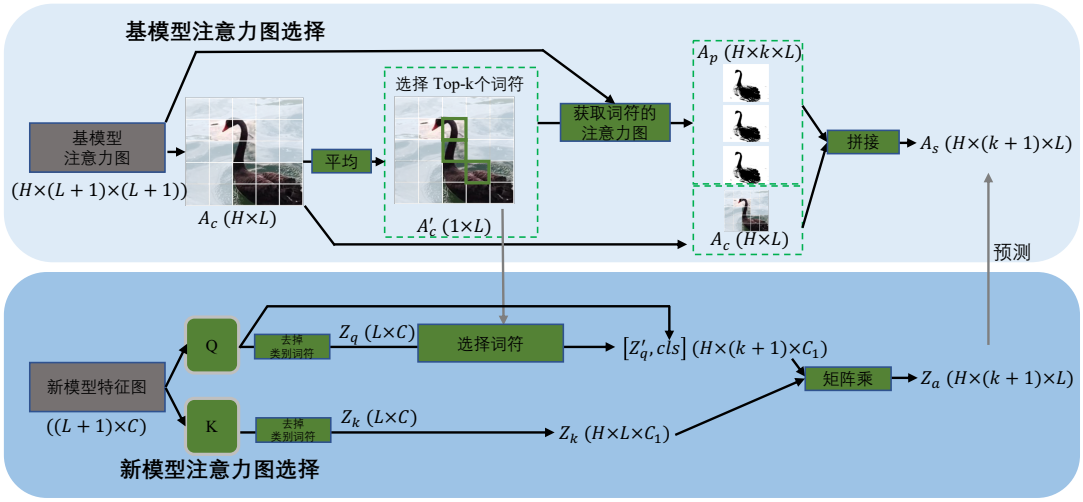


图 6.7 语义相关注意力图选择的细节图。

归一化可以显著提高新模型的性能。在归一化之后，依照 [52]，新模型在解码器后使用一个全连接层生成 Z_f，用于预测掩码区域的基模型目标 Y_f：

$$L_{\text{fea}} = \|M \circ (Y_f - Z_f)\|_2^2, \tag{6.3}$$

其中 M 是掩码矩阵，$\circ$ 表示元素点乘。

语义相关注意力图作为目标 预训练 ViT 模型中的自注意力具有很强的捕获特征词符之间语义关系的能力 [310, 311, 312]。本文随之提出利用自注意力图作为 MIM 的一种重建目标，进一步增强新模型的语义关系建模能力。根据之前关于自注意力图在知识蒸馏中作用的研究 [313, 314]，并不是所有的注意力图都包含有用的语义关系，有严重噪声的注意力图甚至会阻碍模型的学习。因此，有必要选择部分注意力图，以减少可能出现的严重噪声，也有助于降低训练成本。

本文利用包含足够全局语义的基模型的类别词符来选择最相似的特征词符，从而过滤掉可能的噪声。如图 6.7所示，给定基模型中最后一个 ViT 层的类别词符和特征词符之间的注意力图 $A_c \in \mathbb{R}^{H\times L}$，其中 L 和 H 分别表示词符数量和注意力头数量，本文对注意力图 A_c 沿注意力头的维度求平均，得到 $A'_c \in \mathbb{R}^{1\times L}$。然后，本文选取 A'_c 中值最大的 k 个词符，然后计算这 k 个词符与所有特征词符之间的注意力图 $A_p \in \mathbb{R}^{H\times k\times L}$。考虑到类别词符的重要性，本文进一步将其自身与所选 A_p 之间的注意力图合并起来，得到本文最终的重建目标，即 $A_s \in \mathbb{R}^{H\times(k+1)\times L}$。请注意，当本文计算 A_s 时，在 Softmax 操作之前添加了一

个温度 τ 来调整注意力的清晰度。对于新模型，本文分别使用两个全连接层将其解码器输出映射到两个预测 $Z_q \in \mathbb{R}^{L\times C}$ 和 $Z_k \in \mathbb{R}^{L\times C}$ 中。新模型从 Z_q 中选取与 A_s 中相同的词符，形成 $Z_q^{'} \in \mathbb{R}^{k\times C}$。然后将新模型中的类别词符 cls 与 Z_q' 连接，并计算出 KQ 注意力图 $Z_a = \mathrm{Softmax}([Z_q^{'}, \mathsf{cls}]^{\top} Z_k) \in \mathbb{R}^{H\times(k+1)\times L}$。最后，计算预测 Z_a 与目标 A_s 之间的预测熵损失：

$$L_{\mathrm{att}} = -A_s \log Z_a. \tag{6.4}$$

多目标解码器 由于特征目标和注意目标这两个重建目标的性质不同，在新模型中，一个解码器难以同时处理这两个重建目标，且往往会导致预测冲突。但是对每个目标使用单独的解码器会增加可训练参数，从而减慢训练速度。为了解决这个问题，本文使用了一个简单的解码器适应方案，即构造特定于目标的输入特征，然后将它们送到共享解码器中。具体来说，本文将新模型编码器的输出特征 Z_e （见公式. 6.1）输入到一个全连接层中，然后用一个可学习的词符填充掩码区域的词符以获得 $Z_f^{'}$。然后类似地，给定 Z_e，本文也使用一个全连接层和一个可学习的词符来获得 $Z_m^{'}$。接下来，本文分别将 $Z_f^{'}$ 和 $Z_m^{'}$ 输入到一个共享的基于 ViT 模块的解码器中，用于预测基模型的特征和注意力重建目标。与 MAE 中编码器输出和 RGB 图像之间的巨大语义差距不同，基模型目标与新模型预测具有相似的语义。因此，一个浅的 2 层解码器比 MAE 中使用的 8 层解码器的效果更好。这种设计也大大降低了训练成本。

6.2.4 预训练细节

本文在 ImageNet-1K [8] 通过预训练随机初始化的 ViT [302] 模型来评估 TEC 预训练方案。训练使用 16×16 的词符尺寸和 224×224 的图像分辨率，通过 AdamW [315] 以 4,096 的批大小训练 300/800 个迭代轮次。为了确保改进来自 TEC，本文没有使用任何显式或隐式的额外训练数据和或比新模型更强的基模型。实际上，本文分别使用 iBOT [53] 和 MAE [52] 在 ImageNet-1K 上预训练的 ViT 模型作为基模型。基模型是从它们的官方公开发布版本中获得的。本文使用与 MAE 相同的掩码策略，即使用 75% 的随机掩码比。

6.3 自适应能力分析

本节将自适应的可持续自监督表征学习算法在分类、语义分割、实例分割等多个视觉场景中进行测试。同时，本节也验证了该可持续的表征学习算法对

表 6.1 在使用 ViT 进行 ImageNet-1K 分类微调任务中与现有 SSL 方法的比较。† 和灰色表示使用隐式/显式额外数据。TEC 的预训练迭代轮次数是在基模型指导下随机初始化权重的新模型训练的迭代轮次，其中不包括基模型的迭代轮次。相比较的结果来自于其方法官方报告的结果。

模型	方法	迭代轮次	指导	Top-1 准确率
ViT-Base	Deit III [316]	800	Supervised	83.8
	DINO [310]	300	NA	82.8
	MoCov3 [50]	300	NA	83.2
	MixMIM [317]	300	RGB	83.2
	MFM [208]	300	Frequency	83.1
	BEiT [51]	800	DALLE†	83.2
	SplitMask [318]	300	NA	83.6
	ConMIM [319]	800	Momentum	83.7
	SimMIM [204]	800	RGB	83.8
	SIM [320]	1600	Momentum	83.8
	CAE [203]	1600	DALLE†	83.9
	MaskFeat [205]	1600	HOG	84.0
	LoMaR [321]	1600	RGB	84.1
	BootMAE [211]	800	RGB+Momentum	84.2
	data2vec [210]	800	Momentum	84.2
	Mugs [54]	1600	NA	84.3
	MVP [213]	300	CLIP†	84.4
	PeCo [209]	800	Perceptual codebook	84.5
	CMAE [198]	1600	RGB	84.7
	Ge2-AE [207]	800	RGB+Frequency	84.8
	FD-CLIP [218]	300	CLIP†	84.9
	MAE [52]	1600	RGB	83.6
	FD-MAE [218]	300	MAE	$83.8_{+0.2}$
	TEC	300	MAE	$\mathbf{84.7}_{+1.1}$
	TEC	800	MAE	$\mathbf{84.8}_{+1.2}$
	iBOT-ImageNet-22K	-	Momentum	84.4
	iBOT [53]	1600	Momentum	84.1
	SemMAE [322]	800	iBOT	$84.5_{+0.4}$
	TEC	300	iBOT	$\mathbf{84.8}_{+0.7}$
	TEC	800	iBOT	$\mathbf{85.1}_{+1.0}$
ViT-Large	MAE [52]	1600	RGB	85.9
	TEC	300	MAE	$\mathbf{86.5}_{+0.6}$

不同预训练模型的自适应学习能力。

在 ImageNet-1K 上微调 表 6.1总结了在 ImageNet-1K 上的微调性能。可以观察到，以 iBOT 为基模型，从随机初始化训练 300 个迭代轮次，TEC 在 Top-1 准确率指标上超出了基模型 0.7%，经过 800 个训练轮次 TEC 又将准确率提高了 1.0%。同样，在 300/800 训练轮次下，TEC 相对 MAE 基模型分别带来 1.1% 和 1.2% 的提升。这些结果表明，TEC 可以进一步改进基于 MIM 的 MAE 和 iBOT 等先进方法。此外，表 6.1 还显示，在训练成本相似甚至更低的情况下，TEC 优于其他先进的 SSL 方法，包括使用隐式额外数据训练的方法，如 MVP [213] 和 FD-CLIP [218]。更令人惊讶的是，仅使用 ImageNet-1K 数据的

表 6.2 预训练方法使用 Upernet 和 ViT-B 对 ADE20K 进行语义分割的对比。

方法	迭代轮次	mIoU
BEiT	800	47.1
PeCo	800	48.5
GE2-AE	800	48.9
CAE	1600	50.2
CMAE	1600	50.1
MAE	1600	48.1
TEC$_{\text{MAE}}$	800	**49.9**
iBOT	1600	50.0
TEC$_{\text{iBOT}}$	800	**51.0**

表 6.3 预训练方法使用 Cascade MaskRCNN 和 ViT-B 对 COCO 进行实例分割的对比。

方法	AP_{bbox}	AP_{mask}
[53] 的实现版本		
iBOT	51.2	44.2
TEC$_{\text{iBOT}}$	**52.7**	**45.4**
[309] 的实现版本		
MAE	54.0	46.7
TEC$_{\text{MAE}}$	**54.6**	**47.2**

TEC 比使用 ImageNet-22K 训练的 iBOT 提高了 0.7%，这表明 TEC 预训练比更多的训练数据更有效。据本文所知，仅使用 ImageNet-1K 时，TEC 在 ViT-B 模型的 85.1% 性能是新的最优记录，显示了可持续 SSL 学习的潜力。本文还使用 ViT-Large 研究了 TEC 的缩放能力，并观察到 TEC 从随机初始化训练 300 个迭代轮次后比 MAE 预训练的基模型高出 0.6%。

ImageNet-1K 的轻量化微调 例如线性预测（linear probing）等轻量化微调方法旨在微调少量参数以适应下游任务。本文在线性预测设置下测试 TEC，该设置仅对冻结参数的预训练模型的输出线性分类器进行微调。表 6.4给出了线性预测下 VIT-B 在 ImageNet-1K 上的分类准确率。TEC 相比 MAE 基模型提升了 1.8%，显示了学习到的新模型中含有更多与类别相关的语义信息。事实上，本文用于预训练的输入适配器和编码器适配器也可以用于轻量化微调。通过对输入适配器进行微调可显著提高 4.6% 准确率，对输入适配器和编码器适配器进行微调可较 MAE 基模型提高 11.9%。这也显示了本文提出的适配器的优点。

在 ImageNet-S 上的大规模语义分割 为了测试 TEC 预训练模型的像素级表示能力，本文在第五章介绍的大规模语义分割任务的 ImageNet-S 数据集上进行语义分割微调。由于预训练和微调数据没有域偏移，本文使用 ViT-B 作为分割模型。从表 6.5可以看出，TEC_{MAE} 在 mIoU 上将 MAE 基模型提高了 4.6%。当

表 6.4 在轻量化微调下，ImageNet-1K 数据集的 Top-1 准确率。

方法	迭代轮次	设置	Top-1 准确率
MAE	1600	线性预测	68.0
TEC_{MAE}	800	线性预测 + 输入适配器微调 + 编码器适配器微调	69.8 72.6 79.9

表 6.5 TEC 在 ImageNet-S 数据集上的半监督语义分割效果。

预训练机制	方法	迭代轮次	mIoU_{val}
SSL	MAE TEC_{MAE}	1600 800	38.3 **42.9**
SSL+ 有监督微调	MAE TEC_{MAE}	1600+100 800+100	61.0 **62.0**

使用有监督 ImageNet 完全微调的预训练模型时，TEC_{MAE} 比 MAE 获得 1.0% 的性能提升。以上结果证明模型自适用的可持续 SSL 有助于提升数据自适应的大规模语义分割任务的性能。

语义分割 对于 ADE20K [40] 数据集上的语义分割，本文使用带有 ViT-B 的 Upernet [323] 作为分割模型。在 mIoU 指标上，从表 6.2 可以看出，TEC_{iBOT} 比 iBOT 基模型提高了 1.0%，TEC_{MAE} 比其 MAE 基模型提高了 1.8%。因此，与它们的基模型相比，TEC 预训练模型在语义分割方面表现出更强的迁移学习能力。此外，TEC 在预训练轮次较少的情况下，比强大的竞争对手表现出明显的优势。例如，它优于 MAE、CAE [203] 和 CMAE [198] 2.9%、0.8% 和 0.9%，实现了新的最先进结果。

实例分割 对于 COCO 上的实例分割 [55]，为了公平起见，本文将 iBOT [53] 和 ViTDet [309] 实现的 Cascade MaskRCNN [324] 应用于基于 iBOT/MAE 基模型的 TEC 上。从表 6.3可以看出，使用 iBOT 的实现，TEC 在边界框 AP 上超过 iBOT 基模型 1.5%，在掩码 AP 上超出 1.2%。当使用 ViTDet 的实现时，TEC 在边界框 AP 上获得了 0.6% 的提升，在掩码 AP 上也获得了 0.5% 的提升，这表明 TEC 能够稳定的改善性能。

6.4 实验与分析

本节对提出的 TEC 进行了消融实验和分析。默认情况下，模型预训练 300 个迭代轮次，并在 ImageNet-1K 分类任务上进行微调和评估。Top-1 准确率为

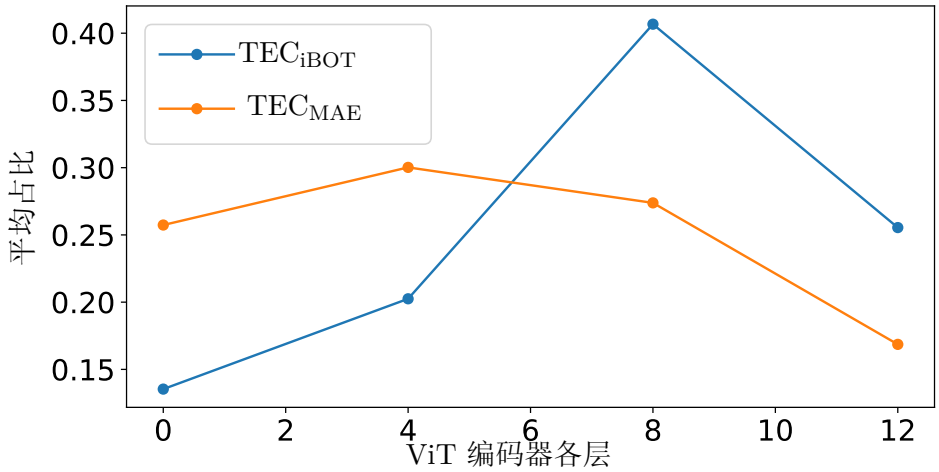


图 6.8 编码器适配器占编码器输出 Z_e 的平均比例。

评测指标。

条件预训练 条件适配器有助于在不同基模型下进行 SSL 预训练。表 6.7a显示，使用 MAE 和 iBOT 作为基模型时，适配器稳定地提高了 0.4% 和 0.2% 的性能。为了观察对不同基模型的适应差异，本文在图 6.8 中显示了编码器适配器占编码器输出的平均比例，即公式. 6.1 中的 Z_n'/Z_e。iBOT 基模型要求适配器从更深的层提供更多的特征，而 MAE 基模型使适配器更关注浅层，该结果与基模型的属性是一致的，即 iBOT 基模型具有更多的高级分类语义，而 MAE 模型具有更多的低级图像细节。

不同维度上的特征归一化 本文在空间的词符维度上对目标特征进行归一化，强调词符之间的相对关系，这不同于现有的在通道维度上对特征进行归一化的方法。在表 6.7c中，对词符维度进行归一化比通道维度归一化获得 0.3% 的提升。相反，通道维度归一化与无归一化相比没有任何效果。通道维度归一化强调通道的特征差异。相反，本文的词符维度归一化强调了词符之间的关系，这与 MIM 方案中的词符预测是匹配的。表 6.7a显示，使用词符维度归一化特征进行训练，相较于 MAE/iBOT 基模型，具有 0.6%/0.4% 的提升，显示了其相对于基模型的适应性。

语义关联的注意力图 KQ 注意力图通常包含词符之间的语义关系，因此被用作增强区域间关系属性的基模型目标。表 6.7a 显示，使用注意力图进一步改善了用词符维度归一化训练的模型。表 6.7f比较了不同类型注意力图的效果。仅使用类别词符的注意力图并没有改善，而使用语义相关词符的注意力图比基线

表 6.6 使用 ViT-B 在 ImageNet-1K 进行分类微调的消融研究。Acc_1 代表 Top-1 准确率。

空间维度归一化特征	语义关联的注意力图	适配器	MAE 基模型	iBOT 基模型
基模型性能			83.6%	84.1%
✓			84.2%	84.5%
✓	✓		84.3%	84.7%
✓		✓	84.6%	84.7%
✓	✓	✓	84.7%	84.8%

(a) 本文提出模块的性能消融。

	Acc_1
MAE 基模型	83.6
无适配器	84.2
+ 输入适配器	84.3
+ 编码器适配器	84.6

(b) 适配器的作用。

归一化	Acc_1
MAE 基模型	83.6
无归一化	83.9
特征维度	83.9
空间维度	**84.2**

(c) 空间维度归一化作用。

	Acc_1
iBOT 基模型	84.1
载入基模型权重	84.4
不载入基模型权重	**84.8**

(d) 新模型初始化权重。

	迭代轮次	Acc_1
MAE	1600	83.6
TEC$_{\text{MAE1600ep}}$	300	**84.7**$_{+1.1}$
MAE	300	82.9
TEC$_{\text{MAE300ep}}$	**100**	**83.9**$_{+1.0}$
TEC$_{\text{MAE300ep}}$	300	**84.3**$_{+1.4}$

(e) TEC 加速 MAE 的训练。

	Acc_1
iBOT 基模型	84.1
无注意力图	84.5
类别词符的注意力图	84.5
所有词符的注意力图	84.6
语义关联词符的注意力图	**84.7**

(f) 语义相关的注意力图的作用。

提高了 0.2%。因此，词符之间的关系有助于训练。与使用所有的注意力图相比，选择语义相关的注意力图可以降低噪声，从而获得更大的性能提升。

加快基模型的训练进程 默认情况下，本文使用完全预训练的 SSL 模型作为基模型。为了验证 TEC 是否可以改善未收敛的 SSL 模型，本文使用 300 个迭代轮次 MAE 预训练的 ViT-B 作为基模型，并从随机初始化训练 TEC 100/300 个轮次。如表 6.7e所示，300 轮次预训练的 MAE 的 Top-1 准确率为 82.9%。相比之下，$\text{TEC}_{\text{MAE300ep}}$ 在 300/100 迭代轮次时达到 84.3%/83.9%，超过 300 个迭代轮次 MAE 基模型 1.4%/1.0%。值得注意的是，$\text{TEC}_{\text{MAE300ep}}$ 在仅进行 100 个轮次训练的情况下，甚至比 1600 个训练轮次预训练的 MAE 提高了 0.3%，表明 TEC 可以显著加速基模型的训练过程。并且，使用 1600 个轮次的 MAE 基模型训练的 $\text{TEC}_{\text{MAE1600ep}}$ 相比 $\text{TEC}_{\text{MAE300ep}}$ 进一步提高了 0.4% 的性能，表明本文的可持续学习策略依靠好的基模型来取得更好的表现。

是否使用基模型的权重初始化新模型 TEC 框架中的新模型是由随机初始化开始训练的。表 6.7d比较了加载或不加载预训练的基模型权重时新模型的性能。

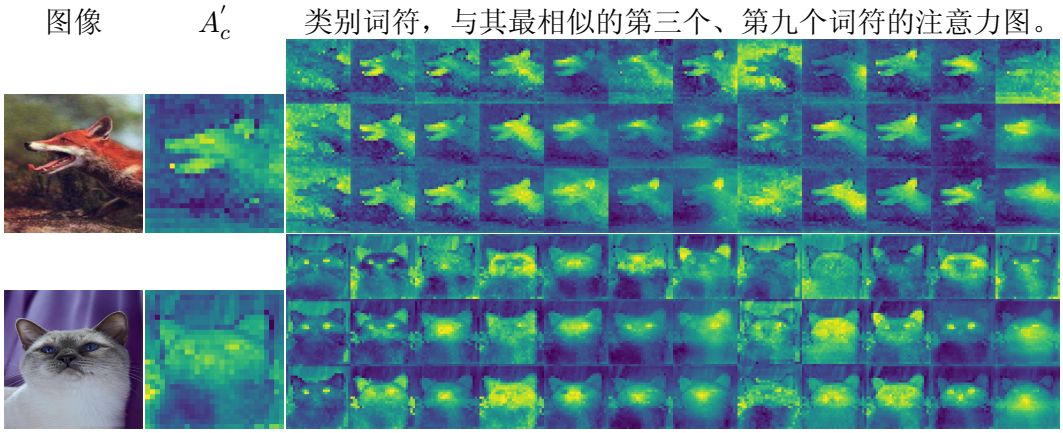


图 6.9 在 iBOT 基模型中选择的语义相关的注意力图目标的可视化。

随机初始化的新模型比使用预训练基模型的权重的新模型性能好 0.4%。本文推测随机初始化使新模型学习到不同的权重分布，从而避免新模型陷入与基模型相同的局部最优。

迈向一般的可持续 SSL 本文旨在基于现有的预训练 SSL 模型向可持续 SSL 迈出第一步。为了验证多次迭代的可持续的 SSL 的可行性，本文使用 TEC 预训练模型作为新一轮 TEC 预训练的基模型。由表 6.7 可知，使用第一轮 TEC 基模型训练的第二轮 TEC 达到了 85.2%。第二轮提高幅度较小的可能原因是网络容量有限或两轮 TEC 预训练学习的知识相似。因此，后续的可持续自监督学习策略需要关注如何保持多轮迭代训练的性能持续提升。

训练成本比较 在一个较小的预训练模型 [325, 326] 的帮助下，加速一个较大语言模型的训练，已经在自然语言处理领域被证明可行。本文依照它们对 FLOPs、训练时间和参数进行比较，结果如表 6.8所示。TEC 需要更短的训练时间来获得比基模型更好的性能。例如，TEC 在训练时间仅为 7%/20% 的情况下，比 iBOT/MAE 的 Top-1 准确率高出 0.7%/1.1%。TEC 具有与 MAE 相似的参数量，因为 TEC 的浅解码器减少参数而适配器增加参数。由于额外的解码器，TEC 和 MAE 与 iBOT 相比具有更多的参数。但是得益于解码器，他们

表 6.7 使用 TEC 作为新的基模型迈向一般的可持续 SSL。

模型	基模型	迭代轮次	Top-1 准确率
iBOT	-	1600	84.1
TEC_{iBOT}	iBOT	800	85.1
TEC	TEC_{iBOT}	800	85.2

表 6.8 预训练方法的训练成本比较。

方法	迭代轮次	时间 (8xA100)	FLOPs (有梯度)	FLOPs (无梯度)	参数量	Top-1 准确率
VIT-B	-	-	17.6G	-	86.6M	-
iBOT	1600	361h	19.2G	19.2G	96.3M	84.1
TEC_{iBOT}	300	25h	8.3G	17.6G	118.6M	84.8
MAE	1600	125h	9.8G	0G	111.9M	83.6
TEC_{MAE}	300	25h	8.3G	17.6G	118.6M	84.7

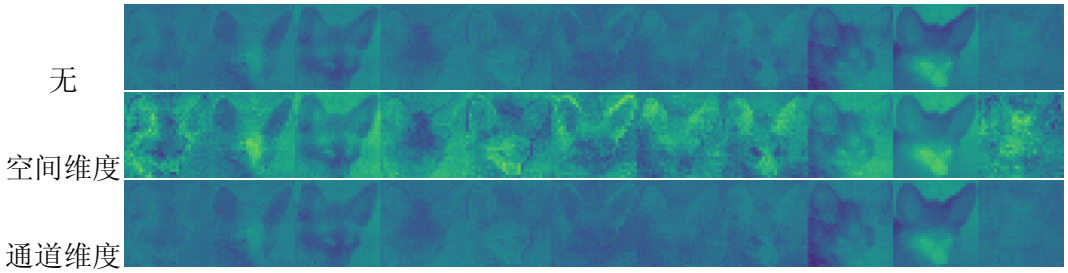


图 6.10 iBOT 基模型的不同维度归一化特征级目标的可视化。

只处理编码器中可见的词符，因此需要比 iBOT 更小的训练成本。由于在某些 SSL 方法中，模型只有部分需要梯度，例如，TEC 中的基模型和 iBOT 中的在线模型不需要反向传播，因此本文比较了有/无梯度的网络部分的 FLOPs。得益于编码器仅处理未掩码词符和浅的两层解码器，TEC 比 iBOT 和 MAE 需要更小的具有梯度的训练 FLOPs。TEC 中基模型的额外 FLOPs（无梯度 FLOPs）小于 iBOT 中的在线网络，因为 TEC 中基模型不需要额外的预测头。与 MAE 相比，TEC 中具有梯度的较小 FLOPs 可以部分平衡基模型的额外 FLOPs。因此，TEC 在每次训练迭代中与 MAE 的训练时间相似。

轻量化微调的比较 表 6.9报告了线性预测、适配器微调和完全微调的准确率。可以观察到，1）TEC 的线性预测性能依赖于基模型，2）适配器微调显著提高了性能。事实上，大多数基于 MIM 的模型，例如 BEiT 和 MAE，它们的线性预测性能要低得多，因为它们没有使用全局语义学习损失，比如聚类损失或实例判别损失。这也解释了 TEC 与 iBOT 等全局语义学习方法相比性能较低的原因。但是通过微调适配器和线性分类器，TEC 相比 iBOT 有 3.9% 的显著优势。这是因为，如图 6.2所示，iBOT 更侧重于区分与全局语义相关的词符，而忽略了其他词符的语义，而 TEC 可以将这些词符分为几个语义组，并进一步识别每个组的语义。这样，微调适配器有助于激活下游任务所需的与全局语义相关的语义组，从而提高模型对全局语义的分辨能力，表现出良好的轻量化微调性能。

与自监督蒸馏方法的比较 本文比较了最近提出的几种自监督蒸馏方法在

表 6.9 线性预测（LP）、适配器微调（Adapter FT）和完全微调（Fully FT）下，ImageNet-1K 数据集上分类任务的 Top-1 准确率。

方法	迭代轮次	设置	Top-1 准确率	Fully FT Top-1 准确率
BEiT	800	LP	56.7	83.2
SimMIM	800	LP	56.7	83.8
BootMAE	800	LP	66.1	84.2
CAE	800	LP	68.6	83.8
SemMAE	800	LP	68.7	84.5
CMAE	800	LP	73.9	84.7
Ge2-AE	800	LP	75.3	84.8
MAE	1600	LP	68.0	83.6
TEC_{MAE}	800	LP	69.8	84.7
TEC_{MAE}	800	Adapter FT	79.9	84.7
iBOT	1600	LP	79.8	84.1
TEC_{iBOT}	800	LP	78.0	84.8
TEC_{iBOT}	800	Adapter FT	**81.9**	**85.1**

表 6.10 TEC 与自监督蒸馏方法的比较。

方法	基模型	网络结构	迭代轮次	Top-1 准确率
MAE	-	ViT-B	1600	83.6
FD_{MAE}	MAE-ViT-B	ViT-B	300	83.8
TEC_{MAE}	MAE-ViT-B	ViT-B	300	84.7
MoCov3	-	ViT-B	300	83.2
$\text{MaskFeat}_{\text{MoCov3}}$	MoCov3-ViT-B	ViT-B	300	83.9
$\text{TEC}_{\text{MoCov3}}$	MoCov3-ViT-B	ViT-B	300	84.5

ImageNet 上的完全微调性能。表 6.10说明，与其他自监督蒸馏方法相比，TEC 方法有明显的提升。在使用 MAE ViT-B 作为基模型时，TEC 比 FD 明显提高了 0.9%。使用 MoCov3 ViT-B 作为基模型时，TEC 与同样采用 MIM 方案的 MaskFeat 相比有 0.6% 的额外性能提升。

词符维度归一化特征级目标和语义注意力级目标的可视化 图 6.9将从 iBOT 基模型中选择的语义注意力图目标进行可视化。类别词符（$A_c^{'}$）的平均注意力图大多集中在高语义对象上，从而使所选词符属于高语义信息。所选词符的注意力图包含了高语义对象与其他区域之间的语义关系。不同的词符有一些不同于其他词符的独特的注意力部分。这些所选词符的注意力图关注相似的语义对象，但在某些部分是互补的，这就解释了为什么使用所选词符的注意图比只使用类别词符注意力图效果更好。图 6.10显示了 iBOT 基模型的词符维度归一化特征级目标的可视化。词符维度归一化特征与原始和通道维度归一化特征相比，具有更好的可区分性。通过词符归一化，特征词符之间的空间关系更加清晰。

6.5 总结

本章节提出通过从预训练的 SSL 模型中自适应学习来探索可持续的自监督学习。本章提出了一种目标增强的条件掩码重建学习方案，以学习并超越现有的 SSL 模型。适配器有助于在预训练期间使新模型自适应地学习各种基模型包含的知识，也可以作为轻量化的微调模块。本章利用掩模重建方案作为超越基模型的基础，构造具有增强空间关系的预测目标，以辅助掩码重建预训练。本章的方法进一步改进了先进的基础模型预训练方法，例如 MAE 和 iBOT，证明了可持续学习的可行性。实验结果也表明可持续自监督学习能够提升第五章提出的大规模语义分割任务的性能。这项工作迈出了向可持续 SSL 的第一步。

第七章 总结与展望

复杂场景的视觉感知是众多计算机视觉任务的基础和核心。视觉场景的复杂性和多样性为模型架构设计和表征学习带来巨大挑战。模型的架构需要满足各种复杂任务和多样场景的需求。复杂场景加剧了模型训练难度，且使人工数据标注成本和模型训练成本大幅增加。为此，本文提出针对复杂视觉场景的自适应感知技术，通过网络架构的多尺度和感受野自适应满足不同场景下任务的尺度和感受野需求。本文提出首个大规模无监督语义分割方法，能够实现无需人工标注的数据自适应表征学习。借助先验模型的知识，本文提出的模型自适应的可持续自监督学习在实现更优视觉感知能力的同时大幅度降低训练开销。本章将对本文贡献进行总结，并依据现有结果展望未来可能的研究方向。

7.1 本文工作总结

本文首先介绍了复杂场景视觉感知的研究背景和意义，并分析了其在网络架构和表征学习上的挑战。通过对相关方向的研究现状进行分析，本文在模型架构方向提出通过增强模型的尺度表达和设置更优的感受野来增强模型对各种任务场景的自适应能力。本文也在学习策略方面提出数据自适应的大规模无监督语义分割和模型自适应的可持续自监督学习来实现复杂场景的低成本高效视觉感知。本文进一步介绍了在以上几个方面的研究现状，并分析了相关工作的不足。

为增强模型尺度表达能力，本文介绍了残差递进的尺度自适应表征方法。本文提出了一种可将卷积神经网络的多尺度表达能力提升到细粒度层次的简洁而高效的模型 Res2Net。Res2Net 扩展出的尺度维度比现存的深度、宽度、组数等维度更加有效。Res2Net 拥有大范围的尺度表征空间，能够有效适应不同任务和场景的尺度需求。实验表明，Res2Net 可以集成在现有的模型上，在复杂场景下的分类、语义分割、实例分割、目标检测等任务上有出色表现。

为了解决网络感受野对不同场景的适应性问题，本文提出了场景自适应的感受野搜索来替代传统手工设置的感受野。该全局到局部的搜索方案，可以由粗到细的寻找有效的感受野组合。其中，全局搜索可以发现相比手工设计具有

更好性能但差异巨大的粗略感受野组合。期望引导的迭代式局部搜索方案能够基于全局搜索结果搜索更加细粒度的感受野组合。该感受野搜索方案，可以插入如动作分割、序列建模、分割、目标检测等多种任务中来进一步提升性能。此外，该方案搜索出的许多结构与传统人工设计的结构差异较大，为理解不同场景和任务下的网络架构的感受野需求提供了新的指导。

为避免针对复杂场景的昂贵人工标注，本文提出了数据自适应的大规模无监督语义分割。本文首先提出了一个新的大规模无监督语义分割问题，用于在具有丰富多样性和大规模数据的现实环境中进行完全自适应的语义分割。本文也为该任务提供了一个基准，包含具有高度多样性的大规模数据、明确的任务目标和充分的评测方法。本文提出的大规模无监督语义分割方法由针对语义分割增强的表征学习策略和像素注意力辅助的像素级标签生成策略构成。它可以在自监督的范式下从大规模数据中自主学习类别和形状表征，并为像素分配自学习到的类别标签。本文验证该方法的有效性，并揭示大规模无监督语义分割对像素级下游任务的作用。此外，本文对相关工作进行评测和分析，总结了该任务面临的挑战和可能的研究方向。

为降低基础模型的预训练成本，本文介绍了模型自适应的绿色可持续自监督视觉感知学习方案。这项工作通过从预训练的自监督表征模型学习来探索可持续的自监督学习。本文提出了一种目标增强的条件掩码重建学习方案，以学习并超越现有的预训练自监督表征模型。其中，该方案中的适配器有助于使新模型适应各种预训练模型，也可以作为轻量化的微调模块。本文构造具有增强空间关系的预测目标来辅助掩码重建预训练。该方法能利用 MAE 和 iBOT 等预训练模型，以更低的训练开销进一步提升基础模型的性能，证明了可持续自监督表征学习的可行性。这项工作迈出了向模型自适应的可持续自监督表征学习的第一步。

7.2 未来工作展望

本文在网络架构的尺度和感受野以及模型训练的数据和模型自适应能力方面依然有可以进一步提升的空间。本节将介绍未来可能的改进方向：

- 残差递进的尺度自适应方法着重提升在细粒度层级的多尺度表征能力。因此，未来工作可以在此基础上增强网络架构的粗粒度的多尺度表征，甚至实现根据需求调整对不同粒度的多尺度特征的使用。此外，本文的实验主

要集中在视觉感知任务，未来工作可以探索利用本文的 Res2Net 模块在非视觉任务上的应用。

- 场景自适应的感受野搜索以卷积的膨胀率作为感受野的表示方式。事实上，感受野的表示方式具有多样性，可通过卷积核大小、网络深度、算子覆盖范围决定。因此，后续工作可以继续探索将本文的全局到局部的感受野搜索算法拓展到更多的感受野表示空间，实现更具通用性的感受野搜索。
- 数据自适应的大规模无监督语义分割证明了该任务的可行性。然而，其性能相较于传统的有监督学习依然有较大差距。后续工作可以考虑在自监督表征学习、类别发现、伪标签生成、模型微调等方面进一步改进大规模无监督语义分割算法。同时，借助于多模态学习和生成模型技术的无监督语义分割也是可以重点探索的方向。
- 虽然本文证明模型自适应的可持续的自监督学习的可行性，但是本文也发现，由于学习信息的影响，同一个学习策略在多次可持续学习迭代后性能会趋向于饱和。因此，后续工作可以探索如何根据现有模型已学到的和欠缺的知识动态调整可持续自监督学习策略。
- 此外，随着 GPT 等通用大模型和生成模型在各项任务上展现出的强大泛化能力，后续的工作可以探索利用生成式通用大模型对于所有的视觉感知任务进行统一。通过将不同视觉感知任务的输出形式和表征学习技术进行统一，将有可能构建出面向复杂场景的具有强大适应能力的统一视觉大模型。

参考文献

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[2] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, D. Lin, MMDetection: Open mmlab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155 (2019).

[3] Y.-Q. Tan, S. Gao, X.-Y. Li, M.-M. Cheng, B. Ren, Vecroad: Point-based iterative graph exploration for road graphs extraction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[4] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, C.-W. Zhao, M.-M. Cheng, Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation, IEEE Transactions on Image Processing 30 2021, 3113–3126.

[5] J. Canny, A computational approach to edge detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 6 1986, 679–698.

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

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2009.

[9] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations

by back-propagating errors, nature 323(6088) 1986, 533–536.

[10] 周飞燕, 金林鹏, 董军, 卷积神经网络研究综述, 计算机学报 40(6) 2017, 23.

[11] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24(4) 2002, 509–522.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.

[13] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, J. Feng, Dual path networks, in: Advances in Neural Information Processing Systems, 2017.

[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2014.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[16] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[17] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[18] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[19] F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, 2016.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning., in: AAAI, Vol. 4, 2017.

[22] W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2016.

[23] M. Loog, F. Lauze, Supervised scale-regularized linear convolutionary filters, in: BMVC, 2017.

[24] R. Wang, M. Gong, D. Tao, Receptive field size versus model depth for single image super-resolution, IEEE Transactions on Image Processing 29 2019, 1669–1682.

[25] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587 (2017).

[26] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, J. Feng, Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Transactions on Pattern Analysis and Machine Intelligence 40(4) 2018, 834–848.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017.

[29] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[30] Y. Chen, M. Rohrbach, Z. Yan, Y. Shuicheng, J. Feng, Y. Kalantidis, Graph-based global reasoning networks, in: Proceedings of the IEEE/CVF Confer-

ence on Computer Vision and Pattern Recognition, 2019.

[31] H. Liu, K. Simonyan, Y. Yang, Darts: Differentiable architecture search, in: International Conference on Learning Representations, 2019.

[32] H. Cai, L. Zhu, S. Han, ProxylessNAS: Direct neural architecture search on target task and hardware, in: International Conference on Learning Representations, 2019.

[33] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[34] E. Real, A. Aggarwal, Y. Huang, Q. V. Le, Regularized evolution for image classifier architecture search, in: Association for the Advancement of Artificial Intelligence, Vol. 33, 2019.

[35] H. Liu, K. Simonyan, O. Vinyals, C. Fernando, K. Kavukcuoglu, Hierarchical representations for efficient architecture search, in: International Conference on Learning Representations, 2018.

[36] L. Xie, A. Yuille, Genetic cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.

[37] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, L. Fei-Fei, Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[38] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.

[39] 侯淇彬, 韩凌昊, 刘姜江, 程明明, 互联网图像驱动的语义分割自主学习, 中国科学: 信息科学 51(7) 2021, 16.

[40] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ade20k dataset, International Journal of Computer Vision 127(3) 2019, 302–321.

[41] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, T. Dar-

rell, Bdd100k: A diverse driving dataset for heterogeneous multitask learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[42] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, International Journal of Computer Vision 111(1) 2015, 98–136.

[43] H. Caesar, J. Uijlings, V. Ferrari, Coco-stuff: Thing and stuff classes in context, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[44] Y. Ouali, C. Hudelot, M. Tami, Autoregressive unsupervised image segmentation, in: European Conference on Computer Vision, 2020.

[45] X. Zhan, Z. Liu, P. Luo, X. Tang, C. Loy, Mix-and-match tuning for self-supervised semantic segmentation, in: Association for the Advancement of Artificial Intelligence, Vol. 32, 2018.

[46] J.-J. Hwang, S. X. Yu, J. Shi, M. D. Collins, T.-J. Yang, X. Zhang, L.-C. Chen, Segsort: Segmentation by discriminative sorting of segments, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[47] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, L. Van Gool, Unsupervised semantic segmentation by contrasting object mask proposals, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[48] J. H. Cho, U. Mall, K. Bala, B. Hariharan, Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[49] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[50] X. Chen*, S. Xie*, K. He, An empirical study of training self-supervised vision transformers, arXiv preprint arXiv:2104.02057 (2021).

[51] H. Bao, L. Dong, F. Wei, Beit: Bert pre-training of image transformers, arXiv preprint arXiv:2106.08254 (2021).

[52] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[53] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, T. Kong, ibot: Image bert pre-training with online tokenizer, in: International Conference on Learning Representations, 2022.

[54] P. Zhou, Y. Zhou, C. Si, W. Yu, T. K. Ng, S. Yan, Mugs: A multi-granular self-supervised learning framework, arXiv preprint arXiv:2203.14415 (2022).

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, 2014.

[56] 张珂, 冯晓晗, 郭玉荣, 苏昱坤, 赵凯, 赵振兵, 马占宇, 丁巧林, 图像分类的深度卷积神经网络模型综述, 中国图象图形学报 26(10) 2021, 21.

[57] M. Lin, Q. Chen, S. Yan, Network in network, in: International Conference on Learning Representations, 2013.

[58] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015.

[59] 陈科圻, 朱志亮, 邓小明, 马翠霞, 王宏安, 多尺度目标检测的深度学习研究综述, 软件学报 32(4) 2021, 27.

[60] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.

[61] M. Najibi, P. Samangouei, R. Chellappa, L. S. Davis, Ssh: Single stage headless face detector, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.

[62] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, J. Tang, Richer convolutional features for edge detection, IEEE Transactions on Pattern

Analysis and Machine Intelligence 41(8) 2019, 1939 – 1946.

[63] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[64] J. Zhao, Y. Cao, D.-P. Fan, X.-Y. Li, L. Zhang, M.-M. Cheng, Contrast prior and fluid pyramid integration for rgbd salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[65] K. Zhao, W. Shen, S. Gao, D. Li, M.-M. Cheng, Hi-Fi: Hierarchical feature integration for skeleton detection, in: International Joint Conference on Artificial Intelligence, 2018.

[66] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014.

[67] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 37(9) 2015, 1904–1916.

[68] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, S. J. Belongie, Feature pyramid networks for object detection., in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vol. 1, 2017.

[69] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, 2016.

[70] 田萱, 王亮, 丁琪, 基于深度学习的图像语义分割方法综述, 软件学报 30(2) 2019, 29.

[71] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[72] K. Zhao, S. Gao, W. Wang, M.-M. Cheng, Optimizing the F-measure for

threshold-free salient object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[73] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Transactions on Image Processing 24(12) 2015, 5706–5722.

[74] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 37(3) 2015, 569–582.

[75] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, N. Zheng, Salient object detection: A discriminative regional feature integration approach, International Journal of Computer Vision 123(2) 2017, 251–268.

[76] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.

[77] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2015.

[78] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[79] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P. Torr, Deeply supervised salient object detection with short connections, IEEE Transactions on Pattern Analysis and Machine Intelligence 41(4) 2019, 815–828.

[80] N. Tomen, S.-L. Pintea, J. Van Gemert, Deep continuous networks, in: International Conference on Machine Learning, PMLR, 2021.

[81] G. Seif, D. Androutsos, Large receptive field networks for high-scale image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018.

[82] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE transactions on affective computing 3(1) 2011, 42–55.

[83] R. T. Collins, A. J. Lipton, T. Kanade, Introduction to the special section on video surveillance, IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8) 2000, 745–746.

[84] R. T. Collins, A. J. Lipton, T. Kanade, H. Fujiyoshi, D. Duggins, Y. Tsin, D. Tolliver, N. Enomoto, O. Hasegawa, P. Burt, et al., A system for video surveillance and monitoring, VSAM final report 2000 2000, 1–68.

[85] V. Saligrama, Z. Chen, Video anomaly detection based on local statistical aggregates, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2012.

[86] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, 2014.

[87] C. Feichtenhofer, A. Pinz, R. P. Wildes, Spatiotemporal multiplier networks for video action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[88] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[89] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[90] A. Fathi, J. M. Rehg, Modeling actions through state changes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2013.

[91] A. Fathi, A. Farhadi, J. M. Rehg, Understanding egocentric activities, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2011.

[92] A. Fathi, X. Ren, J. M. Rehg, Learning to recognize objects in egocentric activities, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2011.

[93] M. Rohrbach, S. Amin, M. Andriluka, B. Schiele, A database for fine grained

activity detection of cooking activities, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2012.

[94] S. Karaman, L. Seidenari, A. Del Bimbo, Fast saliency based pooling of fisher encoded dense trajectories, in: European Conference on Computer Vision, Vol. 1, 2014.

[95] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, M. Shah, Recognition of complex events: Exploiting temporal dynamics between underlying concepts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014.

[96] N. N. Vo, A. F. Bobick, From stochastic grammar to bayes network: Probabilistic parsing of complex activity, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014.

[97] K. Tang, L. Fei-Fei, D. Koller, Learning latent temporal structure for complex event detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2012.

[98] H. Kuehne, J. Gall, T. Serre, An end-to-end generative framework for video segmentation and recognition, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2016.

[99] H. Kuehne, A. Richard, J. Gall, Weakly supervised learning of actions from transcripts, Computer Vision and Image Understanding 163 2017, 78–89.

[100] A. Richard, H. Kuehne, J. Gall, Weakly supervised action learning with rnn based fine-to-coarse modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[101] H. Kuehne, A. Richard, J. Gall, A hybrid rnn-hmm approach for weakly supervised temporal action segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 42(4) 2020, 765–779.

[102] Y. Cheng, Q. Fan, S. Pankanti, A. Choudhary, Temporal sequence modeling for video event detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014.

[103] Y. A. Farha, J. Gall, Ms-tcn: Multi-stage temporal convolutional network for action segmentation, in: Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition, 2019.

[104] S.-J. Li, Y. AbuFarha, Y. Liu, M.-M. Cheng, J. Gall, Ms-tcn++: Multi-stage temporal convolutional network for action segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 0(0) 2020, 1–1.

[105] A. Richard, J. Gall, Temporal action detection using a statistical language model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[106] B. Singh, T. K. Marks, M. Jones, O. Tuzel, M. Shao, A multi-stream bi-directional recurrent neural network for fine-grained action detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[107] S. Singh, C. Arora, C. Jawahar, First person action recognition using deep learned descriptors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[108] L. Ding, C. Xu, Tricornet: A hybrid temporal convolutional and recurrent network for video action segmentation, arXiv preprint arXiv:1705.07818 (2017).

[109] H. Gammulle, T. Fernando, S. Denman, S. Sridharan, C. Fookes, Coupled generative adversarial network for continuous fine-grained action segmentation, in: IEEE/CVF Winter Conference on Applications of Computer Vision, 2019.

[110] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, G. D. Hager, Temporal convolutional networks for action segmentation and detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[111] P. Lei, S. Todorovic, Temporal deformable residual networks for action segmentation in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[112] S. Qiao, L.-C. Chen, A. Yuille, Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution, arXiv preprint arXiv:2006.02334 (2020).

[113] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, P. Luo, Sparse r-cnn: End-to-end object detection with learnable proposals, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[114] Y. Xiong, H. Liu, S. Gupta, B. Akin, G. Bender, Y. Wang, P.-J. Kindermans, M. Tan, V. Singh, B. Chen, Mobiledets: Searching for object detection architectures for mobile accelerators, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[115] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, N. Zheng, End-to-end object detection with fully convolutional network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[116] T. Liang, Y. Wang, Z. Tang, G. Hu, H. Ling, Opanas: One-shot path aggregation network architecture search for object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[117] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[118] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: European Conference on Computer Vision, 2018.

[119] Z. Cai, N. Vasconcelos, Cascade r-cnn: High quality object detection and instance segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 43(5) 2021, 1483–1498.

[120] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, D. Lin, Hybrid task cascade for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[121] M. Hu, Y. Li, L. Fang, S. Wang, A2-fpn: Attention aggregation based feature pyramid network for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[122] X. Shen, J. Yang, C. Wei, B. Deng, J. Huang, X.-S. Hua, X. Cheng, K. Liang,

Dct-mask: Discrete cosine transform mask representation for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[123] H. Ding, S. Qiao, A. Yuille, W. Shen, Deeply shape-guided cascade for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[124] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.

[125] T. Cheng, X. Wang, L. Huang, W. Liu, Boundary-preserving mask r-cnn, in: European Conference on Computer Vision, 2020.

[126] Z. Hayder, X. He, M. Salzmann, Boundary-aware instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[127] J. Liang, N. Homayounfar, W.-C. Ma, Y. Xiong, R. Hu, R. Urtasun, Polytransform: Deep polygon transformer for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[128] Y. Yuan, J. Xie, X. Chen, J. Wang, Segfix: Model-agnostic boundary refinement for segmentation, in: European Conference on Computer Vision, 2020.

[129] C. Tang, H. Chen, X. Li, J. Li, Z. Zhang, X. Hu, Look closer to segment better: Boundary patch refinement for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[130] M. Mitchell, An introduction to genetic algorithms, MIT press, 1998.

[131] Y. Sun, B. Xue, M. Zhang, G. G. Yen, J. Lv, Automatically designing cnn architectures using the genetic algorithm for image classification, IEEE Transactions on Cybernetics 50(9) 2020, 3840–3854.

[132] Z. Lu, I. Whalen, V. Boddeti, Y. Dhebar, K. Deb, E. Goodman, W. Banzhaf, Nsga-net: neural architecture search using multi-objective genetic algorithm, in: The Genetic and Evolutionary Computation Conference, 2019.

[133] X. Chu, T. Zhou, B. Zhang, J. Li, Fair DARTS: Eliminating Unfair Advantages in Differentiable Architecture Search, in: European Conference on Computer Vision, Vol. 12360, 2020.

[134] Y. Xu, L. Xie, X. Zhang, X. Chen, G.-J. Qi, Q. Tian, H. Xiong, Pc-darts: Partial channel connections for memory-efficient architecture search, in: International Conference on Learning Representations, 2020.

[135] A. Wan, X. Dai, P. Zhang, Z. He, Y. Tian, S. Xie, B. Wu, M. Yu, T. Xu, K. Chen, et al., Fbnetv2: Differentiable neural architecture search for spatial and channel dimensions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[136] X. Chen, L. Xie, J. Wu, Q. Tian, Progressive differentiable architecture search: Bridging the depth gap between search and evaluation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[137] A. Zela, T. Elsken, T. Saikia, Y. Marrakchi, T. Brox, F. Hutter, Understanding and robustifying differentiable architecture search, in: International Conference on Learning Representations, 2020.

[138] X. Zhang, H. Xu, H. Mo, J. Tan, C. Yang, L. Wang, W. Ren, Dcnas: Densely connected neural architecture search for semantic image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[139] C. Liu, J. Yuen, A. Torralba, Nonparametric scene parsing via label transfer, IEEE Transactions on Pattern Analysis and Machine Intelligence 33(12) 2011, 2368–2382.

[140] B. C. Russell, A. Efros, J. Sivic, W. T. Freeman, A. Zisserman, Segmenting scenes by matching image composites, in: European Conference on Computer Vision, 2009.

[141] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with superpixels, in: European Conference on Computer Vision, 2010.

[142] T. Malisiewicz, A. A. Efros, Recognition by association via learning per-exemplar distances, in: Proceedings of the IEEE/CVF Conference on Com-

puter Vision and Pattern Recognition, 2008.

[143] D. R. Martin, C. C. Fowlkes, J. Malik, Learning to detect natural image boundaries using local brightness, color, and texture cues, IEEE Transactions on Pattern Analysis and Machine Intelligence 26(5) 2004, 530–549.

[144] P. F. Felzenszwalb, D. P. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision 59(2) 2004, 167–181.

[145] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, S. Yan, Highly efficient salient object detection with 100k parameters, in: European Conference on Computer Vision, 2020.

[146] M.-M. Cheng, S. Gao, A. Borji, Y.-Q. Tan, Z. Lin, M. Wang, A highly efficient model to study the semantics of salient object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 44(11) 2022, 8006–8021.

[147] S. Jenni, P. Favaro, Self-supervised feature learning by learning to spot artifacts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[148] P. Bojanowski, A. Joulin, Unsupervised learning by predicting noise, in: International Conference on Machine Learning, 2017.

[149] L. Zhang, G.-J. Qi, L. Wang, J. Luo, Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[150] Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, L.-P. Morency, Self-supervised learning from a multi-view perspective, in: International Conference on Learning Representations, 2021.

[151] R. Zhang, P. Isola, A. A. Efros, Colorful image colorization, in: European Conference on Computer Vision, 2016.

[152] S. Iizuka, E. Simo-Serra, H. Ishikawa, Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification, ACM Transactions on Graphics 35(4) 2016, 1–11.

[153] G. Larsson, M. Maire, G. Shakhnarovich, Colorization as a proxy task for visual understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[154] M. Noroozi, P. Favaro, Unsupervised learning of visual representations by solving jigsaw puzzles, in: European Conference on Computer Vision, 2016.

[155] M. Noroozi, A. Vinjimoor, P. Favaro, H. Pirsiavash, Boosting self-supervised learning via knowledge transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[156] I. Misra, L. v. d. Maaten, Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[157] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[158] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, in: International Conference on Learning Representations, 2017.

[159] J. Donahue, K. Simonyan, Large scale adversarial representation learning, in: Advances in Neural Information Processing Systems, 2019.

[160] C. Doersch, A. Gupta, A. A. Efros, Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2015.

[161] T. N. Mundhenk, D. Ho, B. Y. Chen, Improvements to context based self-supervised learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[162] M. Noroozi, H. Pirsiavash, P. Favaro, Representation learning by learning to count, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.

[163] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, in: International Conference on Learning Representations, 2018.

[164] Z. Ren, Y. J. Lee, Cross-domain self-supervised multi-task feature learning

using synthetic imagery, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[165] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[166] O. Henaff, Data-efficient image recognition with contrastive predictive coding, in: International Conference on Machine Learning, 2020.

[167] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020.

[168] Y. Tian, D. Krishnan, P. Isola, Contrastive multiview coding, in: European Conference on Computer Vision, 2020.

[169] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, P. Isola, What makes for good views for contrastive learning, in: Advances in Neural Information Processing Systems, 2020.

[170] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Ávila Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap your own latent - a new approach to self-supervised learning, in: Advances in Neural Information Processing Systems, 2020.

[171] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, H. Hu, Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[172] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[173] A. YM., R. C., V. A., Self-labelling via simultaneous clustering and representation learning, in: International Conference on Learning Representations, 2020.

[174] J. Li, P. Zhou, C. Xiong, R. Socher, S. C. Hoi, Prototypical contrastive learning of unsupervised representations, in: International Conference on

Learning Representations, 2021.

[175] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, in: Advances in Neural Information Processing Systems, 2020.

[176] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, in: International Conference on Learning Representations, 2019.

[177] P. Bachman, R. D. Hjelm, W. Buchwalter, Learning representations by maximizing mutual information across views, in: Advances in Neural Information Processing Systems, 2019.

[178] M. Ye, X. Zhang, P. C. Yuen, S.-F. Chang, Unsupervised embedding learning via invariant and spreading instance feature, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[179] Y. Cao, Z. Xie, B. Liu, Y. Lin, Z. Zhang, H. Hu, Parametric instance classification for unsupervised visual feature learning, in: Advances in Neural Information Processing Systems, 2020.

[180] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, S. Jegelka, Debiased contrastive learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, 2020.

[181] F. Wang, H. Liu, D. Guo, F. Sun, Unsupervised representation learning by invariance propagation, in: Advances in Neural Information Processing Systems, 2020.

[182] M. Patacchiola, A. Storkey, Self-supervised relational reasoning for representation learning, in: Advances in Neural Information Processing Systems, 2020.

[183] J. D. Robinson, C.-Y. Chuang, S. Sra, S. Jegelka, Contrastive learning with hard negative samples, in: International Conference on Learning Representations, 2021.

[184] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discrim-

inatively, with application to face verification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2005.

[185] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2006.

[186] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, T. Brox, Discriminative unsupervised feature learning with convolutional neural networks, in: Advances in Neural Information Processing Systems, 2014.

[187] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[188] Y. Tian, X. Chen, S. Ganguli, Understanding self-supervised learning dynamics without contrastive pairs, in: International Conference on Machine Learning, Vol. 139, 2021.

[189] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, arXiv preprint arXiv:2103.03230 (2021).

[190] C. Zhuang, A. L. Zhai, D. Yamins, Local aggregation for unsupervised learning of visual embeddings, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[191] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep clustering for unsupervised learning of visual features, in: European Conference on Computer Vision, 2018.

[192] M. Caron, P. Bojanowski, J. Mairal, A. Joulin, Unsupervised pre-training of image features on non-curated data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[193] X. Ji, J. F. Henriques, A. Vedaldi, Invariant information clustering for unsupervised image classification and segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[194] X. Yan, I. Misra, A. Gupta, D. Ghadiyaram, D. Mahajan, Clusterfit: Improving generalization of visual representations, in: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[195] X. Zhan, J. Xie, Z. Liu, Y.-S. Ong, C. C. Loy, Online deep clustering for unsupervised representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[196] B. Roh, W. Shin, I. Kim, S. Kim, Spatially consistent representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[197] X. Wang, R. Zhang, C. Shen, T. Kong, L. Li, Dense contrastive learning for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[198] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, J. Feng, Contrastive masked autoencoders are stronger vision learners, arXiv preprint arXiv:2207.13532 (2022).

[199] L. Wang, F. Liang, Y. Li, W. Ouyang, H. Zhang, J. Shao, Repre: Improving self-supervised vision transformer with reconstructive pre-training, arXiv preprint arXiv:2201.06857 (2022).

[200] X. Kong, X. Zhang, Understanding masked image modeling via learning occlusion invariant feature, arXiv preprint arXiv:2208.04164 (2022).

[201] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[202] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: International Conference on Machine Learning, PMLR, 2021.

[203] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, J. Wang, Context autoencoder for self-supervised representation learning, arXiv preprint arXiv:2202.03026 (2022).

[204] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, H. Hu, Simmim: A simple framework for masked image modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[205] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, C. Feichtenhofer, Masked

feature prediction for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[206] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2005.

[207] H. Liu, X. Jiang, X. Li, A. Guo, D. Jiang, B. Ren, The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training, arXiv preprint arXiv:2204.08227 (2022).

[208] J. Xie, W. Li, X. Zhan, Z. Liu, Y. S. Ong, C. C. Loy, Masked frequency modeling for self-supervised visual pre-training, arXiv preprint arXiv:2206.07706 (2022).

[209] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, N. Yu, Peco: Perceptual codebook for bert pre-training of vision transformers, arXiv preprint arXiv:2111.12710 (2021).

[210] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, M. Auli, Data2vec: A general framework for self-supervised learning in speech, vision and language, arXiv preprint arXiv:2202.03555 (2022).

[211] X. Dong, J. Bao, T. Zhang, D. Chen, W. Zhang, L. Yuan, D. Chen, F. Wen, N. Yu, Bootstrapped masked autoencoders for vision bert pretraining, in: European Conference on Computer Vision, 2022.

[212] Z. Yang, Z. Li, M. Shao, D. Shi, Z. Yuan, C. Yuan, Masked generative distillation, arXiv preprint arXiv:2205.01529 (2022).

[213] L. Wei, L. Xie, W. Zhou, H. Li, Q. Tian, Mvp: Multimodality-guided visual pre-training, arXiv preprint arXiv:2203.05175 (2022).

[214] L. Yuan, F. E. Tay, G. Li, T. Wang, J. Feng, Revisiting knowledge distillation via label smoothing regularization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[215] Z. Fang, J. Wang, L. Wang, L. Zhang, Y. Yang, Z. Liu, Seed: Self-supervised distillation for visual representation, arXiv preprint arXiv:2101.04731 (2021).

[216] K. Navaneet, S. A. Koohpayegani, A. Tejankar, H. Pirsiavash, Simreg: Regression as a simple yet effective tool for self-supervised knowledge distillation, arXiv preprint arXiv:2201.05131 (2022).

[217] H. Xu, J. Fang, X. Zhang, L. Xie, X. Wang, W. Dai, H. Xiong, Q. Tian, Bag of instances aggregation boosts self-supervised distillation, in: International Conference on Learning Representations, 2021.

[218] Y. Wei, H. Hu, Z. Xie, Z. Zhang, Y. Cao, J. Bao, D. Chen, B. Guo, Contrastive learning rivals masked image modeling in fine-tuning via feature distillation, arXiv preprint arXiv:2205.14141 (2022).

[219] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.

[220] C.-F. R. Chen, Q. Fan, N. Mallinar, T. Sercu, R. Feris, Big-Little Net: An Efficient Multi-Scale Feature Representation for Visual and Speech Recognition, in: International Conference on Learning Representations, 2019.

[221] B. Cheng, R. Xiao, J. Wang, T. Huang, L. Zhang, High frequency residual learning for multi-scale image classification, in: British Machine Vision Conference, 2019.

[222] S. Zagoruyko, N. Komodakis, Wide residual networks, in: British Machine Vision Conference, 2016.

[223] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[224] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International Journal of Computer Vision 115(3) 2015, 211–252.

[225] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, Tech. rep., Citeseer (2009).

[226] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, M. Li, Bag of tricks for

image classification with convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[227] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[228] Y. Ma, D. Yu, T. Wu, H. Wang, Paddlepaddle: An open-source deep learning platform from industrial practice, Frontiers of Data and Domputing 1(1) 2019, 105–115.

[229] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, in: NeurIPS, 2020.

[230] Y. Liu, Y.-H. Wu, P. Wen, Y. Shi, Y. Qiu, M.-M. Cheng, Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 44(3) 2022, 1415–1428.

[231] Z. Lin, Z. Zhang, L.-Z. Chen, M.-M. Cheng, S.-P. Lu, Interactive image segmentation with first click attention, in: IEEE CVPR, 2020.

[232] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, L. Wang, Tea: Temporal excitation and aggregation for action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[233] D.-P. Fan, G.-P. Ji, M.-M. Cheng, L. Shao, Concealed object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 44(10) 2022, 6024–6042.

[234] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2020.

[235] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Inf-net: Automatic covid-19 lung infection segmentation from ct images, IEEE Transactions on Medical Imaging 39(8) 2020, 2626 – 2637.

[236] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, S. Yan, Object region mining with adversarial erasing: A simple classification to semantic segmentation approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[237] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International Journal of Computer Vision 88(2) 2010, 303–338.

[238] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: European Conference on Computer Vision, Vol. 11211, 2018.

[239] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2011.

[240] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Transactions on Pattern Analysis and Machine Intelligence 33(2) 2011, 353–367.

[241] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2013.

[242] Y. Li, X. Hou, C. Koch, J. M. Rehg, A. L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014.

[243] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2013.

[244] M.-H. Chen, B. Li, Y. Bao, G. AlRegib, Z. Kira, Action segmentation with joint self-supervised temporal domain adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[245] S. Gao, Q. Han, D. Li, P. Peng, M.-M. Cheng, P. Peng, Representative batch normalization with feature calibration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[246] J. Li, S. Todorovic, Set-constrained viterbi for set-supervised action segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[247] Y. Huang, Y. Sugano, Y. Sato, Improving action segmentation via graph-based temporal reasoning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[248] R. Prenger, R. Valle, B. Catanzaro, Waveglow: A flow-based generative network for speech synthesis, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2019.

[249] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: A generative model for raw audio, arXiv preprint arXiv:1609.03499 (2016).

[250] S. Stein, S. J. McKenna, Combining embedded accelerometers with computer vision for recognizing food preparation activities, in: ACM international joint conference on Pervasive and ubiquitous computing, 2013.

[251] H. Kuehne, A. Arslan, T. Serre, The language of actions: Recovering the syntax and semantics of goal-directed human activities, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2014.

[252] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019.

[253] S.-M. Hu, D. Liang, G.-Y. Yang, G.-W. Yang, W.-Y. Zhou, Jittor: a novel deep learning framework with meta-operators and unified graph execution, Science China Information Sciences 63(12) 2020, 1–21.

[254] Z. Wang, Z. Gao, L. Wang, Z. Li, G. Wu, Boundary-aware cascade networks for temporal action segmentation, in: European Conference on Computer Vision, 2020.

[255] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition, 2015.

[256] L. Ding, C. Xu, Weakly-supervised action segmentation with iterative soft boundary assignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[257] C. Lea, A. Reiter, R. Vidal, G. D. Hager, Segmental spatiotemporal cnns for fine-grained action segmentation, in: European Conference on Computer Vision, 2016.

[258] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, 2015.

[259] H. Liang, S. Zhang, J. Sun, X. He, W. Huang, K. Zhuang, Z. Li, Darts+: Improved differentiable architecture search with early stopping, arXiv preprint arXiv:1909.06035 (2019).

[260] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[261] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, Denseaspp for semantic segmentation in street scenes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[262] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ade20k dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017.

[263] D. P. Kingma, P. Dhariwal, Glow: Generative flow with invertible 1x1 convolutions, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.

[264] K. Ito, L. Johnson, The lj speech dataset, https://keithito.com/LJ-Speech-Dataset/ (2017).

[265] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., Tacotron: Towards end-to-end speech synthesis, in: Interspeech, 2017.

[266] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen,

Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

[267] R. Kubichek, Mel-cepstral distance measure for objective speech quality assessment, in: Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, 1993.

[268] A. Rix, J. Beerends, M. Hollier, A. Hekstra, Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, 2001.

[269] S. R. Quackenbush, T. P. Barnwell, M. A. Clements, Objective measures of speech quality, Prentice-Hall, 1988.

[270] S. Bai, J. Z. Kolter, V. Koltun, An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271 (2018).

[271] Q. V. Le, N. Jaitly, G. E. Hinton, A simple way to initialize recurrent networks of rectified linear units, arXiv preprint arXiv:1504.00941 (2015).

[272] S. Zhang, Y. Wu, T. Che, Z. Lin, R. Memisevic, R. R. Salakhutdinov, Y. Bengio, Architectural complexity measures of recurrent neural networks, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.

[273] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86(11) 1998, 2278–2324.

[274] S. Wisdom, T. Powers, J. Hershey, J. Le Roux, L. Atlas, Full-capacity unitary recurrent neural networks, in: Advances in Neural Information Processing Systems, Vol. 29, 2016.

[275] T. Cooijmans, N. Ballas, C. Laurent, Ç. Gülçehre, A. Courville, Recurrent batch normalization, in: International Conference on Learning Representations, 2016.

[276] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, N. R. Ke, A. Goyal, Y. Bengio, A. Courville, C. Pal, Zoneout: Regularizing rnns by

randomly preserving hidden activations, in: International Conference on Learning Representations, 2016.

[277] L. Jing, Y. Shen, T. Dubcek, J. Peurifoy, S. Skirlo, Y. LeCun, M. Tegmark, M. Soljačić, Tunable efficient unitary neural networks (eunn) and their application to rnns, in: International Conference on Machine Learning, PMLR, 2017.

[278] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555 (2014).

[279] R. Pascanu, C. Gulcehre, K. Cho, Y. Bengio, How to construct deep recurrent neural networks, arXiv preprint arXiv:1312.6026 (2013).

[280] R. Jozefowicz, W. Zaremba, I. Sutskever, An empirical exploration of recurrent network architectures, in: International Conference on Machine Learning, PMLR, 2015.

[281] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[282] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvtv2: Improved baselines with pyramid vision transformer, Computational Visual Media 8(3) 2022, 1–10.

[283] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[284] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.

[285] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, B. Xiao, Deep high-resolution representation learning for visual recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 43(10) (2021).

[286] C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017.

[287] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, L. Van Der Maaten, Exploring the limits of weakly supervised pretraining, in: European Conference on Computer Vision, 2018.

[288] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, A. v. d. Oord, Are we done with imagenet?, arXiv preprint arXiv:2006.07159 (2020).

[289] E. Cole, X. Yang, K. Wilber, O. Mac Aodha, S. Belongie, When does contrastive visual representation learning work?, arXiv preprint arXiv:2105.05837 (2021).

[290] K. Kotar, G. Ilharco, L. Schmidt, K. Ehsani, R. Mottaghi, Contrasting contrastive self-supervised representation learning pipelines, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[291] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[292] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, V. Ferrari, The open images dataset V4, International Journal of Computer Vision 128(7) 2020, 1956–1981.

[293] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, A. Kirillov, Boundary IoU: Improving object-centric image segmentation evaluation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[294] N. Zhao, Z. Wu, R. W. H. Lau, S. Lin, What makes instance discrimination good for transfer learning?, in: International Conference on Learning Representations, 2021.

[295] A. Islam, C.-F. R. Chen, R. Panda, L. Karlinsky, R. Radke, R. Feris, A broad study on the transferability of visual representations with contrastive

learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[296] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, arXiv preprint arXiv:2003.04297 (2020).

[297] Q. Hu, X. Wang, W. Hu, G.-J. Qi, Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[298] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016.

[299] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, Vol. 97, 2019.

[300] Q. Xie, M.-T. Luong, E. Hovy, Q. V. Le, Self-training with noisy student improves imagenet classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[301] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, D. Mahajan, Billion-scale semi-supervised learning for image classification, arXiv preprint arXiv:1905.00546 (2019).

[302] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[303] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, S.-N. Lim, Visual prompt tuning, arXiv preprint arXiv:2203.12119 (2022).

[304] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, P. Luo, Adaptformer: Adapting vision transformers for scalable visual recognition, arXiv preprint arXiv:2205.13535 (2022).

[305] P. Gao, T. Ma, H. Li, J. Dai, Y. Qiao, Convmae: Masked convolution meets masked autoencoders, arXiv preprint arXiv:2205.03892 (2022).

[306] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for nlp, in: International Conference on Machine Learning, PMLR, 2019.

[307] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, arXiv preprint arXiv:2101.00190 (2021).

[308] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, J. Tang, Gpt understands, too, arXiv preprint arXiv:2103.10385 (2021).

[309] Y. Li, H. Mao, R. Girshick, K. He, Exploring plain vision transformer backbones for object detection, arXiv preprint arXiv:2203.16527 (2022).

[310] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[311] Z.-Y. Li, S. Gao, M.-M. Cheng, Exploring feature self-relation for self-supervised transformer, arXiv preprint arXiv:2206.05184 (2022).

[312] A. Ziegler, Y. M. Asano, Self-supervised learning of object parts for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[313] H. Wu, Y. Gao, Y. Zhang, S. Lin, Y. Xie, X. Sun, K. Li, Self-supervised models are good teaching assistants for vision transformers, in: International Conference on Machine Learning, PMLR, 2022.

[314] S. Wang, J. Gao, Z. Li, J. Sun, W. Hu, A closer look at self-supervised lightweight vision transformers, arXiv preprint arXiv:2205.14443 (2022).

[315] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[316] H. Touvron, M. Cord, H. Jegou, Deit iii: Revenge of the vit, in: European Conference on Computer Vision, Vol. 13684, 2022.

[317] J. Liu, X. Huang, Y. Liu, H. Li, Mixmim: Mixed and masked image modeling for efficient visual representation learning, arXiv preprint arXiv:2205.13137 (2022).

[318] A. El-Nouby, G. Izacard, H. Touvron, I. Laptev, H. Jegou, E. Grave,

Are large-scale datasets necessary for self-supervised pre-training?, arXiv preprint arXiv:2112.10740 (2021).

[319] K. Yi, Y. Ge, X. Li, S. Yang, D. Li, J. Wu, Y. Shan, X. Qie, Masked image modeling with denoising contrast, arXiv preprint arXiv:2205.09616 (2022).

[320] C. Tao, X. Zhu, G. Huang, Y. Qiao, X. Wang, J. Dai, Siamese image modeling for self-supervised vision representation learning, arXiv preprint arXiv:2206.01204 (2022).

[321] J. Chen, M. Hu, B. Li, M. Elhoseiny, Efficient self-supervised vision pre-training with local masked reconstruction, arXiv preprint arXiv:2206.00790 (2022).

[322] G. Li, H. Zheng, D. Liu, B. Su, C. Zheng, Semmae: Semantic-guided masking for learning masked autoencoders, arXiv preprint arXiv:2206.10207 (2022).

[323] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, J. Sun, Unified perceptual parsing for scene understanding, in: European Conference on Computer Vision, 2018.

[324] Z. Cai, N. Vasconcelos, Cascade r-cnn: high quality object detection and instance segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 43(5) 2019, 1483–1498.

[325] Y. Qin, Y. Lin, J. Yi, J. Zhang, X. Han, Z. Zhang, Y. Su, Z. Liu, P. Li, M. Sun, et al., Knowledge inheritance for pre-trained language models, arXiv preprint arXiv:2105.13880 (2021).

[326] C. Chen, Y. Yin, L. Shang, X. Jiang, Y. Qin, F. Wang, Z. Wang, X. Chen, Z. Liu, Q. Liu, bert2bert: Towards reusable pretrained language models, arXiv preprint arXiv:2110.07143 (2021).

致谢

在博士生涯结束之际，回顾过往几年，感慨良多。特此向各位致以最真诚的谢意。

本人衷心感激导师程明明教授在科研中和生活中对我的指导和帮助。程老师对我的信任和支持让我能探索更多科研的可能性，度过充实的博士时光。也感谢帮助我接触并热爱上科研的本科期间导师程文驰教授。感谢颜水成教授、Ming-Hsuan Yang 教授、Philip Torr 教授、王亮教授、周攀和陈云鹏对我在学术上的指导。

感谢实验室和实习期间同学们对我科研上的帮助和生活上的支持和陪伴。其中特别感谢赵凯师兄对我在博士开始阶段的认真指导，也感谢范登平、侯淇滨等师兄给我的宝贵建议和帮助。感谢我的合作者韩琦、李钟毓、谭永强、吴宇寰、陆承泽、林铮、顾宇超等同学与我一起完成一项又一项有挑战又有趣的科研。感谢研究实习期间谢星宇、余玮宸、高晨、陈守法等同学与我一起拼搏一起欢笑。特别感谢实验室的朋友们与我一起打桌游、打球、尝遍津南美食，让我的博士生活丰富多彩，充满欢声笑语。

最后，感谢家人对我的无尽的爱护和无条件的支持，有你们是我莫大的幸运。

生活充满新奇和挑战，我将带着你们对我的支持和帮助，探索未知。

个人简历

高尚华，生于 1996 年 12 月 23 日。于 2014 年就读于西安电子科技大学通信工程学院，并于 2018 年获得通信工程学士学位。在 2018 年进入南开大学师从程明明教授攻读计算机科学与技术博士学位，目前主要研究方向为神经网络结构设计，表征学习，以及场景分割等计算机视觉应用。

博士期间发表的主要学术论文：

1. **Shanghua Gao**, Zhong-Yu Li, Qi Han, Ming-Ming Cheng, and Liang Wang. RF-Next: Efficient Receptive Field Search for Convolutional Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (**IEEE TPAMI**), 2022. doi:10.1109/TPAMI.2022.3183829.
2. **Shanghua Gao**, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (**IEEE TPAMI**) (2022). doi:10.1109/TPAMI.2022.3218275.
3. **Shanghua Gao**, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence (**IEEE TPAMI**), 2021. doi:10.1109/TPAMI.2019.2938758.
4. Ming-Ming Cheng*, **Shanghua Gao***, Ali Borji, Yong-Qiang Tan, Zheng Lin, and Meng Wang. A highly efficient model to study the semantics of salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (**IEEE TPAMI**), 2021. doi:10.1109/TPAMI.2021.3107956.
5. **Shanghua Gao**, Qi Han, Duo Li, Ming-Ming Cheng, and Pai Peng. Representative batch normalization with feature calibration. IEEE / CVF Computer Vision and Pattern Recognition Conference (**CVPR Oral**), 2021. doi:10.1109/CVPR46437.2021.00856.
6. **Shanghua Gao**, Qi Han, Zhong-Yu Li, Pai Peng, Liang Wang, and Ming-Ming Cheng. Global2local: Efficient structure search for video action seg-

mentation. IEEE / CVF Computer Vision and Pattern Recognition Conference (**CVPR**), 2021. doi:10.1109/CVPR46437.2021.01653.

7. **Shanghua Gao**, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. European Conference on Computer Vision (ECCV), 2020. doi:10.1007/978-3-030-58539-6_42.
8. **Shanghua Gao**, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Towards Sustainable Self-supervised Learning. NeurIPS workshop, 2022. doi:10.48550/arXiv.2210.11016.
9. Yu-Huan Wu, **Shanghua Gao**, Jie Mei, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang, and Ming-Ming Cheng. JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation. IEEE Transactions on Image Processing (**IEEE TIP**), 2021. doi:10.1109/TIP.2021.3058783.
10. Yu-Chao Gu, **Shanghua Gao**, Xu-Sheng Cao, Peng Du, Shao-Ping Lu, and Ming-Ming Cheng. iNAS: Integral NAS for Device-Aware Salient Object Detection. IEEE/CVF International Conference on Computer Vision (**ICCV**), 2021. doi:10.1109/ICCV48922.2021.00489.
11. Yong-Qiang Tan, **Shanghua Gao**, Xuan-Yi Li, Ming-Ming Cheng, and Bo Ren. Vecroad: Point-based iterative graph exploration for road graphs extraction. In: IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR), 2020. doi:10.1109/CVPR42600.2020.00893.
12. Zhao, Kai, **Shanghua Gao**, Wenguan Wang, and Ming-Ming Cheng. Optimizing the f-measure for threshold-free salient object detection. IEEE/CVF International Conference on Computer Vision (**ICCV**), 2019. doi:10.1109/ICCV.2019.00894.

博士期间申请的专利成果：

1. 程明明; **高尚华**; 李钟毓. 一种面向大规模数据的无监督语义分割方法及系统. 中国专利: 202110600887.8
2. 程明明; **高尚华**; 韩琦. 基于全局到局部的感受野搜索的动作分割模型获取方法. 中国专利: 202110004845.8 (**已实现专利转化**)
3. 程明明; **高尚华**; 韩琦. 基于增强表现力神经网络批归一化的图像表征方法

及系统. 中国专利: 202011551847.0

4. 程明明; **高尚华**; 谭永强; 陆承泽. 超小参数量的分割模型的实现方法. 中国专利: 202010045961.X
5. 程明明; **高尚华**; 赵凯. 可集成到神经网络架构中的图像多尺度信息提取方法. 中国专利: 201910242489.6
6. 程明明; 李钟毓; **高尚华**. 一种基于多维度关系建模的视觉 Transformer 自监督学习方法及系统. 中国专利: 202210645115.0

研究生期间主要获得的荣誉奖励：

1. 2021 年中国百篇最具影响国际学术论文, 中国科学技术信息研究所，2022 年 12 月发布 [1]

[1] https://www.istic.ac.cn/html/1/284/338/1292211314138981529.html