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

面向复杂场景的多层次目标检测与分割

Multi-level Object Detection and Segmentation in Complex Scenes

论文作者 吴宇寰 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 周国栋教授 评 阅 人 匿名评审

南开大学

二〇二二年十月

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	面向复杂场景的多层次目标检测与分割				
姓　　名	吴宇寰	学号	1120180141	答辩日期	2022 年 11 月 24 日
论 文 类 别	博士 √　学历硕士 □　硕士专业学位 □　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	17694804367		电子邮箱	wuyuhuan@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、本科的学位论文。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

目标检测与分割是计算机视觉领域最重要的研究领域之一，它是诸多下游应用的基础，它的任务是检测与分割感兴趣的目标。然而，目标的尺寸、形状、颜色、位置以及所处的环境都是复杂多变的，这些难题就使得目标检测与分割成为计算机视觉最具有挑战的问题之一。

设计面向复杂场景的目标检测与分割算法主要有以下挑战：（1）现有的算法对目标的定位能力不够高效，更多考虑了如何完美地恢复目标具体的细节，而忽略了对目标的定位能力；（2）现有的算法多通道特征融合的效率低，使得算法的实时性受到了限制，难以应用到体积小、功耗低的手机等移动设备上；（3）许多场景下的数据精细标注获取难，使得算法的泛化能力受到了限制。（4）现有的算法多尺度建模的效率低，难以胜任复杂场景下的目标检测与分割任务。

为了解决以上不足，本文提供了不同的改进方案来解决以上四大主要挑战难题。具体的研究内容和主要贡献如下：

1. 本文提出了极致下采样技术，聚焦于复杂场景中高效目标定位难的挑战。它不断地对特征下采样直至其变为一维特征向量，来学习图像的全局视图，从而使算法获得强大的全局先验，在消除模型对目标定位所需高分辨率要求的同时提升了目标的定位精度。该技术应用于显著性目标检测中，在五大主流数据集上进行了评测，并与目前主流的方法进行了性能对比，新算法取得了最佳性能。
2. 本文提出了用于多通道高效特征融合的隐式信息恢复技术，聚焦于现有算法多通道特征融合效率低的挑战。隐式信息恢复通过在输出端对感兴趣的信息进行重建恢复，在最粗糙的层次上进行融合就能达到较好的效果，大幅提升了特征融合的效率。该技术应用于 RGB-D 显著性目标检测中，在六大主流数据集上与其他方法进行了比较，新的算法在相比其他方法提速 15 ～ 150 倍的情况下取得了相当的性能。
3. 本文聚焦于复杂图像精细标注难以获取的挑战，提出了基于注意力融合的二元感知技术。注意力融合帮助算法使用分类、分割的二元感知来充分利用了更多的数据，特别是在低对比度的情况下，能够更精准地检测

目标区域。

4. 本文聚焦于现有算法多尺度建模效率低的挑战，提出了基于金字塔池化的骨干特征提取技术。它非常高效，在降低模型计算复杂度的同时提升了多尺度特征表达能力。通过该技术，本文构建了一整套全新的骨干网络，并应用于语义分割、物体检测、实例分割等多个经典的目标检测与分割任务，在多个主流数据集上与其他骨干网络进行了对比，都取得了最佳性能，同时保持了较小的网络参数量和计算量。此外，该技术与本文提出的极致下采样、隐式信息恢复、二元感知等技术相结合，还能进一步大幅提升算法性能。

关键词：目标检测与分割；目标定位；高效融合；注意力融合；多尺度学习

Abstract

Object detection and segmentation is one of the most significant research area in computer vision. They are the basis of many downstream applications. The goal of them is to detect and segment interesting objects. However, the size, shape, color, and location of objects are diverse in complex scenes, making object detection and segmentation one of the most challenging problem in computer vision. It is suboptimal to solve this task with a single level, since it can not capture all the features of objects. Therefore, designing multi-level algorithms can meet the demands of object detection and segmentation in complex scenes.

However, designing multi-level object detection and segmentation algorithms in complex scenes is not an easy task. There are some major challenges: (1) current algorithms have difficulties in efficiently locating objects; they considered too much on how to recover the complete object details, missing the considerations of object localization; (2) current algorithms have high computational cost on multi-channel feature fusion, limiting the real-time speed of them, so they are not applicable to small and low-power devices like mobile phones; (3) data annotations are hard to collect in many scenarios, limiting the generalization ability of algorithms; (4) current algorithms have low efficiency on multi-scale modeling, whose limited performance can lead to the fact of difficultly detect and segment objects in complex scenes.

To solve the above challenges, this thesis proposes several different solutions. The main research contents and contributions are as below:

1. This thesis proposes an extreme downsampling method, focusing on the problem of difficulty in efficiently locating objects in complex scenes. It continuously downsamples the feature until it becomes a global vector feature, which helps the algorithm to learn a global view of the image and enables the algorithm to obtain a strong global prior. It eliminates the requirement of object localization with high resolution, and meanwhile improves the accuracy of object localization. This method is applied to salient object detection, and has

been evaluated on five popular datasets. Compared with recent state-of-the-art methods, the proposed method achieves the best performance.

2. This thesis proposes the implicit information restoration for efficient fusion, focusing on the problem of low efficiency on multi-channel feature. Implicit information restoration can achieve better results by reconstructing and recovering the information of interest at the output. It enables fusing at the coarsest feature level with good performance, greatly improving the efficiency for the feature fusion. Compared with other state-of-the-art methods on six popular datasets, the proposed method achieves competitive performance with $15 \sim 150\times$ speed.
3. This thesis proposes a joint perception method based on attentive fusion, focusing on the challenge of difficultly collecting data annotations for complex images. Attentive fusion helps the algorithm to leverage more data of classification and segmentation tasks, and more effectively detect object regions especially with low contrast.
4. At last, this thesis proposes a new backbone network with pyramid pooling, focusing on the problem of current algorithms that have low efficiency on multi-scale modeling. It is very efficient, reducing the model complexity and improving the multi-scale representation simultaneously. Based on this method, this thesis proposes a series of backbone networks with different network complexities, and apply them to representative object detection and segmentation tasks such as semantic segmentation, object detection, and instance segmentation. Compared with other state-of-the-art backbones, the proposed network achieves the best performance, keeping small network parameters and low computational cost. Moreover, this method is combined with extremely downsampling, implicit information restoration, and joint perception, further largely improving the performance.

Key Words: Object detection and segmentation; object localization; efficient fusion; attention fusion; multi-scale learning

目录

主要缩略词列表

缩略词	全称
ASPP	Atrous Spatial Pyramid Pooling
AFF	Attentive Feature Fusion
BCE	Binary Cross Entropy
BN	Batch Normalization
CNN	Convolutional Neural Network
COVID-19	Corona Virus Disease 2019
CPR	Compact Pyramid Refinement
CT	Computed Tomography
EDB	Extremely-Downsampled Block
EDN	Extremely-Downsampled Network
EFM	Enhanced Feature Module
E_ξ	E-mesaure
FC	Fully-Connected
FCN	Fully Convolutional Network
FFN	Feed-Forward Network
FLOPS	Floating-Point Operations Per Second
FPN	Feature Pyramid Network
FPS	Frames Per Second
F_β^w	加权 F 度量
F_β	F 度量
GAM	Grouped Atrous Module
GAP	Global Average Pooling
IoU	Intersection over Union
IDR	Implicit Depth Restoration
IRB	Inverted Residual Block
JCS	Joint Classification and Segmentation
MAE	Mean Absolute Error
MHSA	Multi-Head Self-Attention
NLP	Natural Language Processing
P2T	Pyramid Pooling Transformer
PSNR	Peak Signal to Noise Ratio
P-MHSA	Pooling-based Multi-Head Self-Attention
RGB-D	Red, Green, Blue, Depth
SCPC	Scale-Correlated Pyramid Convolution
SE(Net)	Squeeze-Excitation (Network)
SGD	Stochastic Gradient Descant
SSIM	Structural Similarity
S_α	S-measure

第 1 章 绪论

1.1 研究背景与意义

机器在智能算法的帮助下，也能够具备人眼功能，这便是计算机视觉的核心内容。计算机视觉利用摄像机或其他传感器甚至多种传感器融合的信息，代替人眼自动地识别、跟踪、定位图像载体中的内容，从而实现自动驾驶 [1]、智能安防 [2]、工业控制 [3]、医学影像分析 [4] 等应用。近年来，随着人工智能、深度学习以及高性能计算机算力的发展，计算机视觉成为其中最热门的研究话题之一。利用计算机视觉技术，可以自动地分析图像数据，从而实现自动化的图像处理、图像识别等功能。而要实现这些功能，最核心的问题之一是检测与分割感兴趣的目标，从而为后续的任务打好重要基础。在复杂场景下，算法所感兴趣目标的尺寸、形状、颜色以及位置都是多变的，可以小到米粒大到汽车、形状规则或不规则、颜色单一或复杂、位置在中心或边缘。这些难题使得目标检测与分割成为计算机视觉中最具挑战性的问题之一。

而想要解决这些难题，设计单一层次的目标检测与分割算法是次优的，因为它们难以覆盖到感兴趣目标的所有特征范围，无法适应各类复杂场景。而人类在分析图像时，也是分层次、逐步了解图像内容。出于收集信息的本能，人类首先会将视线固定在图像最显著的部分，从中提取关键信息，再转向到另一个聚焦点上，直至其分析完所有的部分。所以，解决复杂场景下的目标检测与分割问题的关键在于如何构建多层次算法。

计算机的思维与人脑不同，人脑的理解方式是抽象的，而计算机所看到的图像是一个数值矩阵，它对细节的捕捉如物体边缘，更为敏感，但缺少抽象的理解。这种差异，使得计算机理解图像的方式与人类截然不同。早期，学者们所设计的多层次目标检测与分割算法主要基于手工设计的特征，它们挖掘图像中的纹理、颜色、梯度等低级特征，再从这些低级特征中向上聚合，从而实现目标检测与分割。这种方式是次优的，因为它们难以覆盖到感兴趣目标的所有特征范围，也无法适应复杂多变的目标尺寸、形状、颜色、位置等等。卷积神经网络的出现，使得计算机也能够抽象地理解图像。简要来说，可以先通过卷积操作提取

图像中的局部特征，再通过池化操作将特征图变小，这样之后的卷积操作就可以获取更大的感受野，提取更高层的特征。通过堆叠卷积操作和池化操作，可以在不同层级上获取低层次和高层次的特征，达成多层次的特征提取，进而实现多层次目标检测与分割。回顾近几年的相关研究，它们的主要思想均源于多层次的特征提取，并研究如何有效地利用多层次特征解决目标检测与分割问题。

在目标检测与分割中，如何精准定位感兴趣的目标是其中的基础。即使算法对物体或区域的边缘十分敏感，但如果不能高效准确地定位目标的位置，那么就会导致最终目标检测与分割的结果不准确，预测不到目标，甚至预测错误的目标。在保证精确定位目标的前提下，算法再通过特征融合，就可以实现高精度的目标检测与分割。

不仅目标定位十分重要，目标检测与分割还依赖于多通道特征融合。目标检测与分割是在图像分类之上更复杂的任务，不仅需要确定目标的类别，还需要确定目标的具体位置和区域。因此，设计目标检测与分割算法依赖于多通道如 RGB-D 特征的帮助。然而，不同层次、不同模态之间的特征差异较大。高层次特征往往存储了图像目标的定位等全局信息，而低层次特征则大部分为边缘纹理等局部信息。来自深度图像的特征大幅削弱了自然图像中含有的纹理信息，有许多噪声，但却保留大量的空间线索。因此，算法需要花费大量的算力来精准地融合来自于不同层次与模态的特征，利用它们来提高目标检测与分割的准确率。

除了聚焦于算法对目标的敏感度和特征融合速度，还应该关注标注数据带来的挑战。因为，标注数据在目标检测与分割中也十分重要，它决定了算法所能够学习的内容。目标检测与分割通常运用于自然场景，自然场景随处可见，有非常多的对照数据可以学习。人们自小便无时无刻不在学习如何理解自然场景，积累了大量的先验。算法在学习如何检测与分割自然图像中的目标前，一般先在一些大规模的自然图像数据集上，如 ImageNet [5]，进行预训练，从而先积累大量的先验知识，以降低目标检测与分割的难度。然而在许多场景下，如医学图像场景，标注样本是难以获得的，算法也难以依赖大规模的自然图像预训练知识。其标注数据依赖于有经验的专家，不同数据之间的差异性很小。这些场景下的算法应用价值却很高，可以大幅提高工作和生产效率。设计这类算法的难题在于，算法可学习的内容很少，但又需要获得可靠的知识，以获得高精度的结果。

在复杂场景下，目标检测与分割算法还依赖于强大的多尺度建模能力。多尺度建模能力主要基于强大的特征先验，即先采用一个在大规模数据集上的预训练的骨干网络提取骨干特征，再在这个骨干特征的基础上设计多层次的检测或分割头。骨干特征的好坏对于目标检测与分割的效果有着至关重要的影响。然而，如何设计骨干网络来保证强大的多尺度建模能力，也是一个巨大的挑战。首先，骨干网络应具有强大的特征表达能力，能尽最大限度提取多层次的一般化特征，也能够适用于各类场景下的目标检测与分割。骨干网络应保持鲁棒性，即能够适应各类图像的变化，如图像尺度、光照、噪声、模糊、遮挡等。骨干网络还应具有高效性，否则会成为下游目标检测与分割算法的速度瓶颈。

1.2 研究现状

上节谈到了设计面向复杂场景的多层次目标检测与分割算法所需考虑的因素，主要有四点：目标定位、特征融合、数据以及特征表达能力。就各个因素存在不同的难题与挑战，本节对各个难题的关键技术进行讲述与探讨。

目标定位。已有的算法更多地将研究注意力集中于如何融合骨干网络的多层次特征，以及如何设计多层次检测与分割头。也有部分算法采用大感受野的模块，如 Liu 等人 [6] 采用金字塔场景理解模块（PSP）[7] 提取更强大的多层次池化语义特征。Zeng 等人 [8] 和 Zhao 等人 [9] 则利用膨胀金字塔模块 [10] 来提取大尺度感受野的特征。它们确实能在许多场景下取得良好的结果，但它们仍然在各类复杂场景下有一定的缺陷，特别是难以定位到算法所应感兴趣的目标。本文为此设计了极致下采样技术，并将该技术应用于图像显著性目标检测算法，大幅度提高了算法的定位感兴趣目标的能力。

多通道特征融合。为了实现更精细化的检测与分割结果，需要将多通道特征进行融合，获取清晰的边缘，或者引入多模态特征融合，在更多的信息的帮助下，如边缘信息、深度信息等，来进一步增强检测与分割精度。比如，Hou 等人 [11] 利用侧方向的多个短连接有效融合骨干网络的多层次特征。Zhao 等人 [12] 通过引入额外的边缘信息，提高了高层特征对低层边缘特征的敏感度。Ji 等人 [13] 设计了一个灵活的深度校准模块，为检测模型提供了可靠的深度补充信息。Zhao 等人 [14] 设计了一个基于对比度先验的网络，在深度图像强大的对比度信息的帮助下，对骨干网络特征提供更强大的深度特征。已有的方法虽然有很高的精度，但是它们仍然存在一定的局限性，因为它们花费了太多的计算

量在各类特征融合上以保证精度，难以满足许多移动设备的要求。而许多移动设备的计算能力有限，因此需要设计高效的特征融合方法。为此，本文设计了基于隐式信息恢复技术的特征融合技术，将其应用至 RGB-D 显著性目标检测中，在算法速度相对目前算法提升 15 ∼ 150 倍的基础上，还取得了相当的性能。

数据标注。在许多场景下，样本是难以获取的，标注数据的成本也十分昂贵，数据差异也较小。医学图像场景是一个典型案例。该场景下主要存在两大问题，一个是数据少，难以支撑大规模的训练，另一个是病灶区域的高度可变性、不均匀性，以及病灶与周围组织的低对比度。已有的方法 [15, 16] 主要通过 U-Net [17]、U-Net++ [18] 提取多尺度特征，从而捕获多种尺度的病灶，但它们难以捕获低对比度的病灶区域，且受限于可训练的已详细标注的样本数据大小，这些问题导致它们的泛化能力不足。为了解决以上两大问题，本文提出了基于注意力融合的图像二元感知技术，利用注意力融合技术多层次地捕获低对比度的病灶区域，并充分利用了仅经过简单类别标注的样本数据。该技术应用到了新冠肺炎 CT 病灶分割中，取得了最佳的性能并具有较好的鲁棒性。

多尺度建模能力。目前的算法主要依赖于经大规模预训练的骨干网络来获取强大的多尺度建模能力。前文在考虑目标定位、特征融合、数据多样性时，也采用了主流的骨干网络提取特征。目前的骨干网络主要基于卷积神经网络。然而，卷积神经网络只能提取局部特征，需要经过层层堆叠来获取更大的感受野，如常见的 VGG [19]、ResNet [20]、DenseNet [21] 等。因此，它们的特征表达能力相对不足，容易成为算法的瓶颈。为了解决这一问题，本文提出了基于金字塔池化的 Transformer 技术，它能够在特征提取阶段，在全局范围内以多尺度的方式自适应地选择最需要关注的特征，从而提升特征表达能力。利用该技术，本文提出了一整套不同复杂度的骨干网络，并在多种目标检测与分割任务下进行了实验验证，相对目前主流的网络取得了更好的性能。本文所提出的骨干网络与其他方法，特别是本文所提出的其他方法形成互补关系，能够进一步大幅提升它们的性能。

1.3 研究目标与主要贡献

本文的研究目标是解决面向复杂场景的多层次目标检测与分割的主要挑战，如目标定位难、多通道特征融合慢、标注样本获取难、多尺度建模能力差等，所研究的内容主要关系如图 1.1所示。本文的不同章节主要聚焦于解决不同的问题，

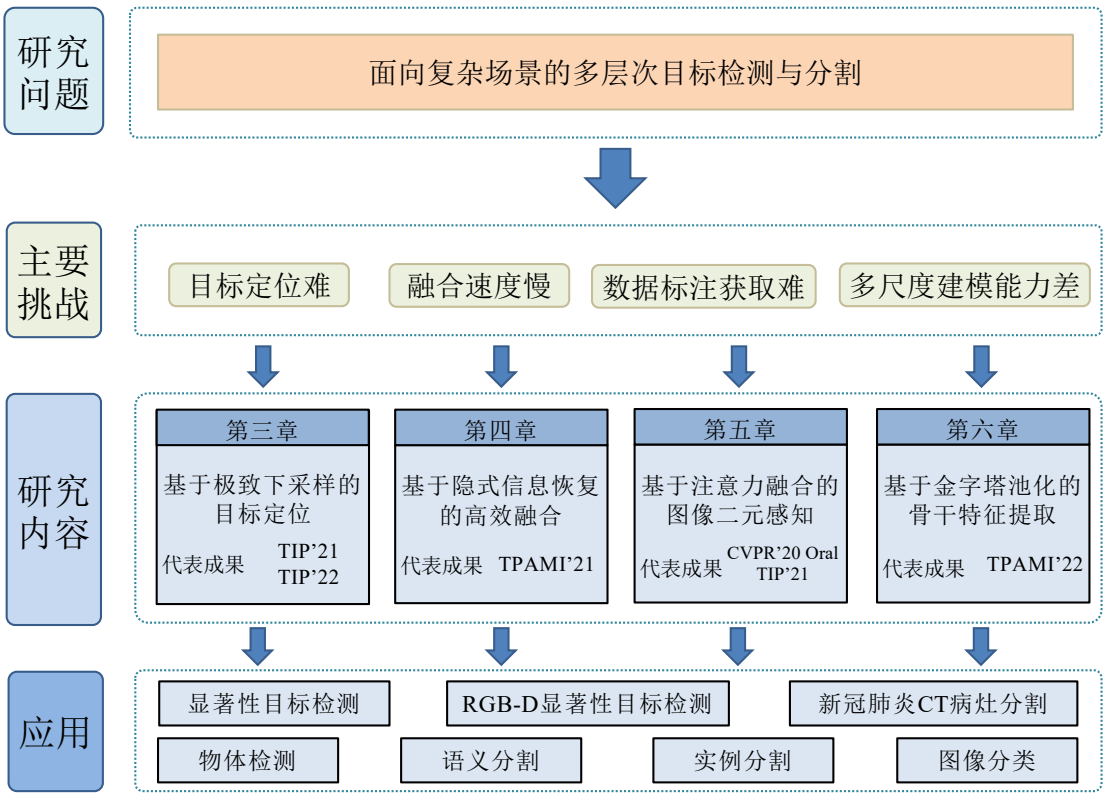


图 1.1 **本文的主要框架。**

具体而言，本文的主要贡献如下：

1. 本文聚焦于高效目标定位问题，并提出了基于极致下采样的目标定位技术。技术验证方面，选择了图像显著性目标检测作为具体任务，它主要用于检测和分割出图像中人们最感兴趣的目标位置及区域，更需要精确的目标定位能力。目前图像显著性目标检测的算法主要是为了处理低级特征学习，而忽略了高级特征学习的问题，导致了较低的目标定位精度。但显著性目标检测更需要定位到目标的位置，它需要学习到整个图像的全局视图。本文将基于极致下采样的目标定位技术应用于图像显著性目标检测算法，使用极致下采样策略逐渐对特征图进行下采样，直到其成为一个特征向量。而在下采样过程中，该策略不断学习深层特征，而随着特征图的变小，学习到的特征会变得更加全面，这样就可以获取到整个图像的全局视图，从而更准确的定位显著性目标。在五大数据集上，本文提出的方法都取得了最佳性能，且拥有较快的速度。

2. 本文聚焦于多通道特征融合速度问题，提出了基于隐式信息恢复的特征融合技术。选择了 RGB-D 显著性目标检测作为具体任务。目前的 RGB-D 显著性目标检测方法特征融合的方式和策略过于复杂和笨重，难以应用到对算力敏感性高的移动设备。本文基于隐式信息恢复的特征融合技术，提出了极致高效的 RGB-D 显著性目标检测方法。它使用隐式深度恢复的策略来保证特征融合前的深度信息不被丢失，强制模型从高层次的深度网络特征中恢复深度图，且仅需在最粗糙的层级进行 RGB 特征和深度特征的融合，对降低计算成本有着至关重要的作用。在六大主流数据集上的实验表明，本文提出的方法能够在速度提升 15 ～ 150 倍的情况下取得与其他主流方法相当的性能。
3. 本文提出了基于注意力融合的图像二元感知技术，聚焦于标注样本获取难的情况下的算法设计问题。以新冠肺炎 CT 病灶分割任务为例，它所涉及的样本较难获得，且数据间差异小，较难发掘对比度低的病灶区域，非常具有代表性。目前的主流分割方法在该任务上的分割准确率也较低。本文提出的算法利用注意力融合保持对病灶位置的高度敏感性，同时在注意力融合的帮助下使用分类、分割的二元图像感知来充分利用更多的样本数据。该方法在增强精度的同时保证了网络的鲁棒性，相对其他方法取得了最佳性能，且仅在 0.8% 的测试 CT 图像帧上失效。
4. 本文提出了基于金字塔池化的骨干特征提取技术，聚焦于目前算法多尺度建模效率低的挑战，与其他技术形成互补关系。目前主流的算法都基于骨干网络所提取的骨干特征。它们的问题在于其网络基础结构的多尺度特征提取能力有限，难以适应复杂场景下的算法要求。该章提出的骨干网络金字塔池化 Transformer 在降低模型复杂度的同时，还提升了算法的多尺度特征表达能力。在各种典型的目标检测与分割任务上的比较实验显示，本文提出的骨干网络取得了与其他主流骨干网络相比更好的效果，同时也保持了较低的模型复杂度和计算量。
5. 第七章对以上研究内容进行了总结，并对未来相关的研究方向进行了展望。

1.4 本文的组织结构

第二章回顾了多种目标检测与分割具体任务、骨干网络设计的相关工作。第三章提出了基于极致下采样的目标定位技术，应用于显著性目标检测任务。第四章提出了基于隐式信息恢复的特征融合技术，应用于 RGB-D 显著性目标检测任务。第五章提出了基于注意力融合的图像二元感知技术，聚焦于标注数据难获取情形下的算法设计问题，应用于新冠肺炎 CT 病灶分割任务。第五章提出了基于金字塔池化的 Transformer 技术，主要聚焦于目前算法多尺度建模效率低的问题，应用于骨干网络设计任务。第七章对以上研究内容进行了总结与展望。

第 2 章 相关工作

目标检测、分割的目标是给定二维或多维图像，找出感兴趣的目标位置及目标的完整区域。它们是计算机视觉领域近几十年最受关注的领域之一，也是最具有挑战性的领域之一。本文的研究内容主要是解决几大主要挑战，如目标定位难、多通道特征融合慢、标注数据难获取、多尺度建模效率低等问题。为了验证本文提出的技术方案，本文选择图像显著性目标检测、RGB-D 显著性目标检测、新冠肺炎 CT 病灶分割、骨干网络架构等具体任务作为研究对象，并分别在 §2.1、§2.2、§2.3及 §2.4介绍了相关工作。

2.1 图像显著性目标检测

图像显著性目标检测的目的是自动检测和分割出自然场景图像中最突出的物体或区域 [22]。早期学者利用手工设计的特征，开发了许多浅度学习方法 [23–26]。此外，启发式的显著性先验也在图像显著性目标检测领域得到了大量的应用，包括色彩对比 [27]，中心先验 [28]，背景先验 [29]，等等。然而，这些方法性能均相对较低，因为手工设计所表达的特征能力有限，在一些较为复杂的场景下这些方法会出现边缘模糊、部分检测区域缺失等问题。

近年来，深度学习在计算机视觉领域中的应用已取得显著成效，基于深度学习的方法也已经成为显著性目标检测的主流方法。早期基于深度学习的方法对图像区域进行处理和分类以进行显著性预测 [30–32]，它抛弃了输入图像的空间布局。在全卷积神经网络（Fully Convolutional Network，FCN）优势的激励下 [33]，学者们将注意力转移到了端到端（end-to-end，即图像到图像间）的显著性目标检测上 [11, 14, 34–37]。大量研究表明，顶部网络层的高层次语义特征可以有效地定位显著性目标位置，而底部网络层的低层次细粒度特征在发现物体细节方面效果更好。最近许多研究都致力于如何完成各个阶段网络特征的多尺度特征融合，以获取位置精确、轮廓完整的显著性目标检测结果。

多尺度特征融合。大多数基于 CNN 的显著性目标检测方法通过设计先进的网络架构进行多级特征融合以实现多尺度学习。最终融合的特征既包含高层次的语义，也包含低层次的精细细节。这些方法的架构通常是基于 HED [11, 35]，

超列（Hypercolumns）[36,38–40]，或者经典 U-Net 结构 [34,41–54,54,55]。它们的目标是在不削弱高层次特征的表达能力的情况下，将低层次的细粒度特征加入到融合的特征中，在保留清晰的边缘的同时分割出显著性目标。

边缘感知。除了多尺度特征融合，目前关于显著性目标检测的研究主要是直接使用边缘信息来提高物体边缘的显著性目标检测精度 [34,35,37,38,41–43,45–47,56]。例如，EGNet [12] 将边缘监督应用于低层次特征学习，从而增强显著性目标检测结果的边缘精度。它首先在网络的最初阶段对低层次特征引入了隐式的边缘信息监督信号，随后将用边缘信息监督增强后的低层次特征传导至更高的特征层级，以提高不同层级下对边缘信息的感知强度，从而提高网络性能。PoolNet [6] 则对每个侧面输出的显著性目标和目标边缘进行了联合监督。它首先利用 PSP 生成了高层次的金字塔池化特征，随后将该特征引入到从上至下每一个阶段的特征融合中，它还对每个阶段融合后的特征引入边缘监督，以提升网络的性能。ITSD [57] 设计了一个双分支网络，两个分支交互地学习边缘细节和显著性目标的位置。它先将显著性目标区域划分为边缘区域和中心区域，然后在从上至下的特征融合中引入两个相互关联的分支，分别学习边缘区域的边缘细节和中心区域的显著性目标的位置的特征。

高层次特征学习。虽然显著性目标检测的相关研究已经取得了巨大的进展，但现有的显著性目标检测方法主要是探索低层次特征的融合或增强，从而对目标边缘区域进行细化，而对高层次特征学习的研究较少。为了加强高层次特征学习，许多研究 [6,8,37,38,40,58] 往往采用一些为语义分割而开发的主流模块，如 PSP [7]，ASPP [10]，DenseASPP [59] 等。例如，Liu 等人 [6] 采用 PSP [7] 提取更强大的全局语义特征。Zeng 等人 [8] 和 Zhao 等人 [9] 则利用 ASPP [10] 来提取大尺度感受野的特征。Liu 等人 [58] 受 DenseASPP [59] 启发，设计了一个轻量级的 DenseASPP 模块，并以该模块为基础进行堆叠，以提取更强大的高层次特征。由于显著性目标检测和语义分割之间的自然差异，目前的显著性目标检测方法在定位显著性目标时只能达到次优的精度。本文对此进行了优化提升，提出了一种极致下采样技术，以更好地学习显著性目标检测中的高层次特征表达。

2.2 RGB-D 显著性目标检测

RGB-D 显著性目标检测是指在 RGB-D 图像中检测显著性目标，它与 RGB 显著性目标检测的主要区别在于，RGB-D 显著性目标检测需要同时利用 RGB

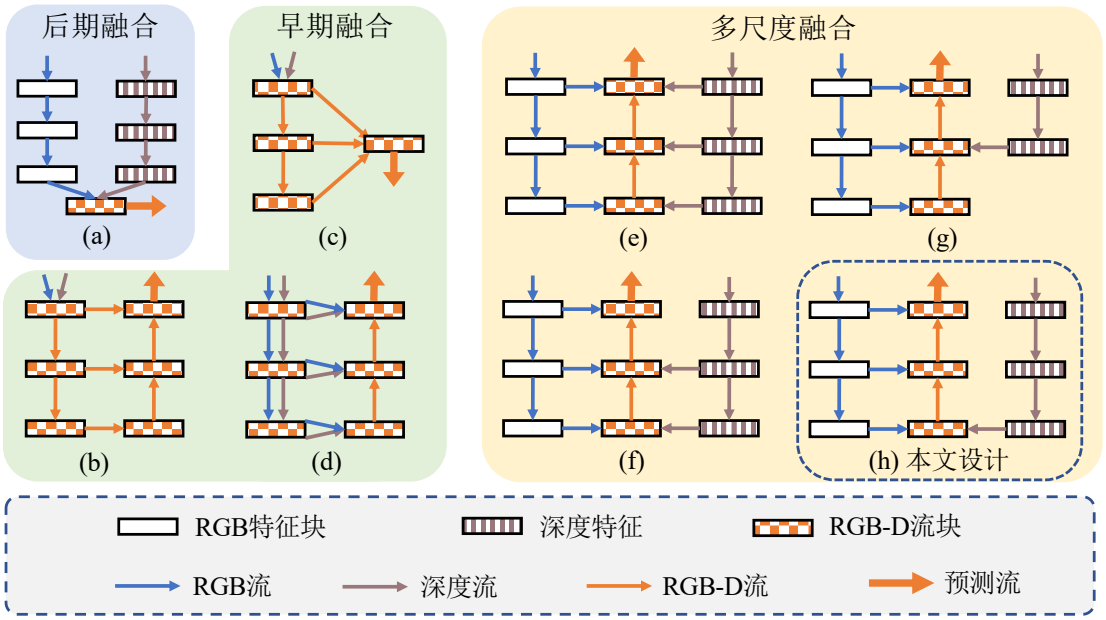


图 2.1 **不同的 RGB-D 特征融合方式。** (a): 后期融合。(b)-(d): 早期融合。(e)-(h): 多尺度融合。

和深度信息。与早期的显著性目标检测方法一样，传统的 RGB-D 显著性目标检测工作从 RGB 图像和深度图中提取手工设计的特征，并将它们融合在一起 [60–65]。最近，由于基于普通二维自然场景图像下的显著性目标检测难以在复杂的图像纹理特征分辨出前背景，基于 RGB-D 的显著性目标检测获得了更多的关注，特别是基于深度学习的 RGB-D 显著性目标检测得到了快速发展 [9,66–75]。例如，Zhao 等人 [9] 提出了一个基于对比度先验的网络，能在深度强大的对比度信息的帮助下，为主干网络特征提供强大的深度特征增强。Piao 等人 [69] 提出通过深度信息诱导的多尺度递归注意力机制来细化深度信息。Huang 等人 [74] 提出通过联合跨模态和单模态特征进行 RGB-D 融合，提供更全面的 RGB-D 分析。Zhang 等人 [66] 提出另一种通过条件变化的自动编码器对不确定的 RGB-D 显著性进行透视。Chen 等人 [76] 首次将 3D CNNs 引入 RGB-D 显著性目标检测中，提供更丰富的空间语义信息。Ji 等人 [13] 提出了一个灵活的深度校准模块，为显著性目标检测模型提供可靠的补充信息。Zhao 等人 [72] 提出了一个自监督学习框架，其只利用了图像级别的标注，节省了进行大规模数据标注的成本。

RGB-D 特征融合方式。如图 2.1所示，可以将目前的方法分为后期融

合 [62,77,78]、前期融合 [66–68,76] 以及多尺度融合 [9,69,71–74,79,80]。后期融合策略出现在特征提取的最后阶段，只根据融合后的特征预测结果 [62,77,78]。早期融合策略直接将输入的 RGB 图像和深度图连接起来，然后通过使用编码器-解码器（encoder-decoder）网络 [68]、超列（hypercolumn）网络 [66] 或者 3D CNN [76]，从这样的 RGB-D 输入中提取显著性图。多尺度融合策略首先分别提取 RGB 特征和深度特征，然后将 RGB-D 特征在所有层次 [69,74]、中高层次 [81] 或中间层次 [71] 进行融合。尽管早期融合策略更有效，但通过多尺度融合得到的结果更加准确。为了确保高效率，本文提出的 MobileSal 方法只在最粗糙的层次上以小的分辨率融合了 RGB 特征和深度特征（如图 2.1 (h) 所示）。记着，隐式深度恢复以一种不需要消耗计算资源的方式被应用于高效的 RGB-D 特征融合，并在从上至下的特征融合中使用紧凑金字塔细化来得到完整而精细的显著性目标边缘。

2.3 新冠肺炎 CT 病灶分割

新冠肺炎（COVID-19）是由新型冠状病毒（SARS-CoV-2）引起的一种急性呼吸道感染，自从 2020 年来已经成为全球性的流行病。世界卫生组织（WHO）也将 COVID-19 列为全球性的公共卫生紧急事件 [82]。COVID-19 的病毒感染会导致肺炎，严重时甚至会导致死亡。新冠肺炎的 CT 肺部感染病变表现也多种多样，如浑浊、实变等。其次，不同患者在不同阶段的 CT 图像也会有不同的表现，它们的病变形状不规则，病灶区域与周围组织的对比度也很低。此外，在 CT 切片上获取高质量的像素级病灶分割标注需要耗费巨大的人力资源，而影像学专家的人力资源极其有限，在短时间内难以收集到足够的标记数据进行大规模的训练。这些困难使得新冠肺炎 CT 病灶分割任务变得非常艰巨。

Rajinikanth 等人 [83] 通过分水岭变换技术 [84] 进行分割，它先利用灰度直方图将肺部区域切割出来，再在切割出的区域使用分水岭算法获取病灶区域。因其使用的是手工设计的思路，得到的特征表达能力有限，所以其分割结果较为粗糙且精度有限。U-Net [17] 是医学图像中常用的分割网络，它的结构简单，训练速度快。Li 等人 [15] 尝试使用 U-Net 来解决新冠肺炎 CT 图像病灶目标分割任务；Jin 等人 [16] 使用 U-Net 的改进版本 U-Net++ [18] 来解决新冠肺炎 CT 图像病灶目标分割任务。但因为新冠肺炎 CT 病灶对比度低的特性且利用的数据量过小，它们容易因过拟合到少样本而丧失了鲁棒性。Zhou 等人 [85] 开发了一

个具有注意力机制的 U-Net，在 CTSeg [86] 数据集上 Dice 得分为 69.1%，但它的训练和测试分割只有 88 和 22 张已感染图片，也不能保证较好的鲁棒性。Fan 等人 [87] 提出了基于半监督策略的 Inf-Net，通过随机选择传播策略，将少量的标注数据传播到大量的无标注数据上，从而使模型能够利用更多的数据。

以上方案在对比度较低的情况下定位和分割病灶的精度较低，或在数据量较少的情况下进行训练，这些都会导致模型的鲁棒性较差。因此，本文提出了基于注意力融合的二元感知模型 JCS，并构建了一个大型的新冠肺炎分类分割数据集，允许 JCS 在分类、分割两个任务上的二元感知。JCS 利用了注意力融合技术保持对病灶位置的高度敏感性，并充分利用了分类特征中丰富的新冠肺炎病灶的对比特征，从而在新冠肺炎 CT 图像病灶分割任务中保持了鲁棒性。

2.4 骨干网络架构

骨干网络是目标检测与分割模型的基础，它用来提取图像的骨干特征。最初，骨干网络只运用于图像分类任务，例如最早期用于手写数字识别的 LENet [88] 与用于 ImageNet 分类的 AlexNet [89]。后来，骨干网络也逐渐成为目标检测与分割等任务的基础，例如用于目标检测的 Faster R-CNN [90]、Mask R-CNN [91]，以及用于分割的 FCN [92]、PSPNet [7]、DeepLab [10]。以上工作的各类实验结果表明，骨干网络的选择对于检测与分割网络的最终性能有着非常重要的影响。因为，目标检测、分割一般使用骨干网络先对输入数据进行一般性的特征提取，然后再使用特定的检测、分割头或进行感兴趣的目标检测或分割。而检测、分割头的预测性能与由骨干网络提取到的特征质量有着密切的关系。

2.4.1 卷积神经网络

自 AlexNet [89] 在 ILSVRC-2012 竞赛 [5] 中获得冠军以来，研究者已经发明了许多先进技术来改进卷积神经网络，在计算机视觉领域中已经变成了各类主流的模型架构。比如，VGG [19] 和 GoogleNet [93] 首先尝试加深卷积神经网络从而获得更好的图像识别效果。然后，ResNets [20] 在残差连接的帮助下成功构建了非常深的卷积神经网络，它在残差连接的基础还引进了瓶颈模块，以达成更鲁棒的特征表达。ResNeXts [94] 在 ResNet [20] 的基础，通过探索其基数操作（Cardinal Operation）来改进 ResNets [20]。具体而言，ResNeXts [94] 将 ResNet 的基本模块替换为多组分组卷积，每组分组卷积可以学习到不同侧重的特征表达，

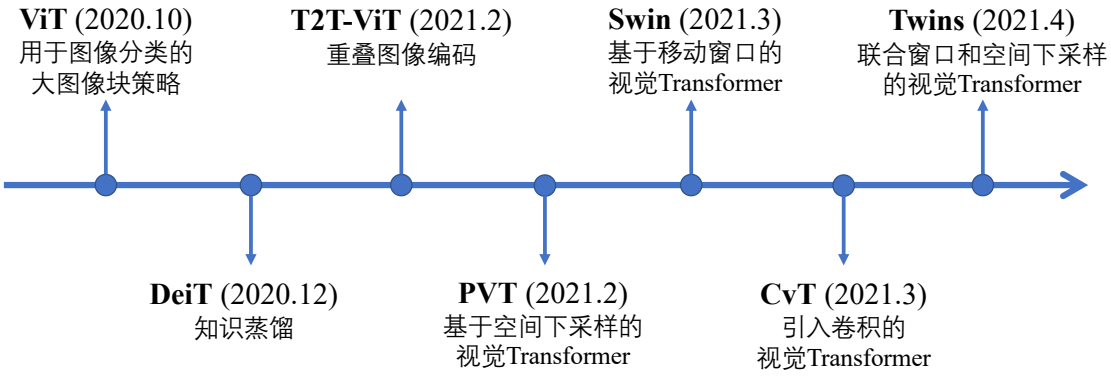


图 2.2 **视觉 Transformer 的主要进展。**

最后通过特征融合的方式将这些分组卷积的特征融合在一起。DenseNets [21] 通过引入了密集连接将每一层连接到其后所有的网络层来实现更好的梯度优化。它的每个阶段中的每层特征都与该阶段之前的每层特征所连接，这种连接方式使得每个阶段的每个层级的特征都能够获得更多的上下文信息。受用于图像描述的 SCA-CNN [95] 启发，SENet [96] 引入了基于挤压激励的通道注意力机制，来改进卷积神经网络的特征提取能力。它将每个模块的输出先通过全局平均池化得到通道向量，再通过两个输出维度不等的全连接层，最后通过 Sigmoid 激活函数得到一个原始输出特征的通道注意力权重，再将其与输出特征相乘，从而得到一个新的特征图。SKNet [97] 引入了基于选择性卷积的通道注意力机制，它通过在每个模块的输出特征上引入一个通道注意力权重，来改进卷积神经网络的特征提取能力。Res2Net [98] 在 ResNeXt [94] 的基础上，引入了更多的多尺度连接以及通道注意力机制，来改进卷积神经网络的特征提取能力。

2.4.2 视觉 Transformer 网络

Transformer 的提出最初运用于自然语言处理领域的机器翻译任务 [99]。通过多头自注意力模块，Transformer 完全依赖于自注意力机制对词符和词符之间的依赖关系进行全局建模。考虑到计算机视觉任务对全局关系的要求也较高，因此容易联想到采用 Transformer 用于改善视觉任务。但是，Transformer 的提出原本是用于处理序列数据，因此无法直接用来处理图像信息。为此，一些研究人员使用卷积神经网络来提取图像二维特征然后扁平化送入 Transformer [100–104]。其中，DETR [100] 是这个方向的里程碑式工作。DETR 首先将 Transformer 引入

到目标检测中，它将每个待检测结果视为一个查询词符，然后利用骨干网络的特征来对查询词符进行编码-解码，得到每个待检测结果的位置和类别。接着，它使用二分图匹配算法来对每个检测框与标注框匹配，得到前景框和背景框，从而在测试中无需传统方法常用的非极大值抑制算法即可得到最终的检测结果，大幅简化了目标检测的流程。

与前者依赖于卷积神经网络骨干进行特征提取不同的是，Dosovitskiy 等人 [105] 提出了第一个视觉 Transformer（ViT）。它们将图像分割成一个个小图像块，并把每个图像块视为自然语言处理研究中的单词或词符。因此，只要配合标签词符（Class Token），纯 Transformer 网络可以直接用来处理图像分类任务。通过大批量的数据以及极强的图像增强策略，它们的方法也在 ImageNet 数据集上取得了非常具有竞争力的结果 [5]。然后，DeiT [106] 通过知识蒸馏技术来减轻训练 ViT [105] 所需的资源，如数据资源要求。它先使用复杂的 RegNet [107] 对图像进行预训练，然后将其作为教师模型，将 ViT [105] 作为学生模型，通过知识蒸馏技术来训练 ViT [105]。T2T-ViT [108] 建议图像分割时保留一些重叠部分从而更好地保留图像局部结构，即引入了重叠图像编码，可以让编码的词符之间有信息交互。CvT [109] 引入了深度卷积，用于在多头自注意力模块中的查询（Query）、索引（Key）和值（Value）的计算。CPVT [110] 建议通过深度卷积将绝对位置编码替换为条件位置编码。它的思想是通过深度卷积可以学习到图像中的局部特征，从而可以更好地学习到图像的位置信息。一些研究人员提出像卷积神经网络一样为视觉 Transformer 构造金字塔结构 [111–114]。其中，PVT [113] 和 MViT [114] 率先采用单层池化操作来减少计算多头自我自注意力模块计算时的词符数量。它们在计算自我注意力前，查询值不变，将索引和值进行大尺度的下采样，从而大幅缩小索引和值的特征长度，达到大幅减少多头自我注意力模块的计算时间和占用内存。但是通过这样方式，它们实际上是进行了“词符-区域”的关系建模，而不是预期的“词符-词符”之间的建模。Swin Transformer [112] 通过牺牲多头自我注意力模块的感受野，将输入的特征划分为了一系列相同大小的窗口特征，从而在小窗口内计算多头自注意力模块来减少了多头注意力所需的计算量。Swin Transformer 通过窗口移位建模来逐渐实现全局相互关系，类似卷积神经网络的机制即通过堆叠更多网络层 [115] 来扩大网络的感受野。Twins [116] 则将 PVT [113] 与 Swin Transformer [112] 的思想进行结合，从而获取更佳的性能，但它仍然没有解决 PVT [113] 和 Swin Transformer [112] 的

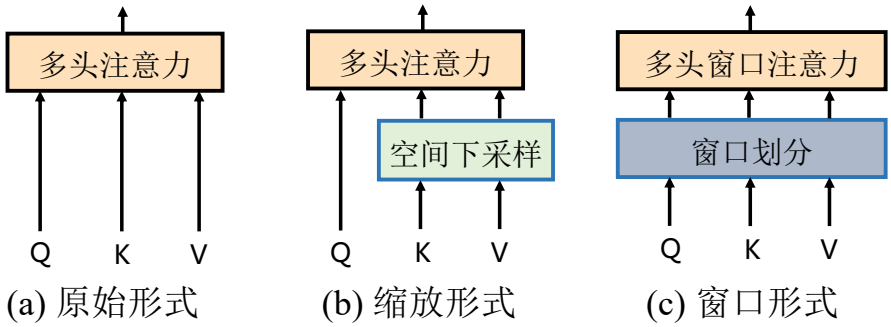


图 2.3 **主流的视觉 Transformer 计算方式。**

内在问题。以上主要工作的时间线如图 2.2所示。

视觉 Transformer 基础模块的计算方式如图 2.3所示，主要有 3 种：原始形式、缩放形式、窗口形式。原始形式的代表工作有 ViT [105] 和 DeiT [106] 等。缩放形式的代表工作有 PVT [113] 与 PVTv2 [117] 等。窗口形式的代表工作有 Swin [112]、Twins [116] 等。以 ViT [105] 为代表的原始形式的问题在于其特征分辨率一般较低，因此其在目标检测、图像分割等下游任务应用的性能较差。以 PVT [113] 为代表的缩放形式通过引入缩放操作使其能支持各类高分辨率的任务，但它们一般使用单层空间采样操作，层次性较弱，所提取的池化特征并不强大。以 Swin Transformer [112] 为代表的窗口形式通过引入窗口划分操作使其无需缩放操作也能支持高分辨率任务，然而它们牺牲了视觉 Transformer 的一个基本特征，即*直接全局关系建模*。

上文提到，PVT [113] 和 MViT [114] 利用单层池化操作提取的池化特征并不强大。与以往研究不同的是，本文将金字塔池化的想法应用于视觉 Transformer 在减少了序列长度的同时，也学习到了多尺度的更加有效的上下文特征。有了更加有效的上下文特征，金字塔池化可能比单层池化更好地计算多头自注意力模块中的自注意力关系。此外，金字塔池化计算效率非常高，因此引入带来的计算开支可以忽略不计。实验表明，本文所提出的 P2T 相比以前的基于卷积神经网络和基于 Transformer 的网络的性能均有明显提升。此外，P2T 的设计还兼容其他 Transformer 技术，比如图像块字典学习（Patch Embedding）[118]、位置编码（Positional Encoding）[110] 和前馈网络（Feed-Forward Network）[119–121]。

2.4.3 轻量化网络

最近，人们越来越关注在移动端的目标检测与分割，对高效的骨干网络产生了很高的需求。自动驾驶载具、机器人和智能手机等移动设备只拥有有限的计算资源，因此传统的繁琐网络，如 VGG [19] 和 ResNets [20] 等骨干网络，并不适合应用在这些平台。为此，研究人员提出了一些轻量化的骨干网络用以解决这一问题。例如，MobileNet [122] 通过引入深度可分离卷积（Depth-wise Seperable Convolution）来大幅减少网络的计算量，深度可分离卷积将普通的 3×3 卷积解耦为逐通道的 3×3 卷积和逐点的 1×1 卷积。MobileNetV2 [123] 受残差连接 [20] 的启发，通过引入倒置残差块来增强 MobileNet 网络主干的特征表达。ShuffleNets [124, 125] 通过将 1×1 卷积替换为分组的 1×1 卷积和交换通道操作（Channel Shuffle Operation）来进一步减少 MobileNets [123, 126] 的延迟。EfficientNet [127] 和 MnasNet [128] 通过采用神经架构搜索（Neural Architecture Search，NAS）来获取到最优的轻量化网络架构。同时，研究人员也提出了一些高效的网络，如 ESPNet 用于语义分割 [129]，EfficientDet 用于目标检测 [130]，以及 HVPNet、SAMNet 用于普通的针对 RGB 图像的显著性目标检测 [58, 131]。这些高效的网络具有较低的计算成本，因此可以灵活地部署在移动平台上。本文的 MobileSal 和 EDN-Lite 方法采用了 MobileNetV2 [123] 骨干网络，从而在高效的显著性目标检测上具有良好的基础。

第 3 章 基于极致下采样的目标定位

在复杂场景下，目标定位在目标检测与分割中是一个巨大的挑战。根据第一章的分析，虽然目前的算法能够利用多层次信息来完成检测与分割，但这些算法存在目标定位难的问题。本章主要聚焦于该问题，提出了基于极致下采样的目标定位技术。极致下采样通过不断地下采样，最终将特征转化为一维特征向量，高效地对目标位置进行建模，可以很好地解决目标定位的问题。它作为目标定位的核心技术应用于图像目标检测与分割中的代表性任务显著性目标检测中，并相对已有的算法取得了最佳性能。第一节介绍了相关研究背景、动机。第二节讲述了基于极致下采样技术的显著性目标检测方法 EDN。第三节对 EDN 方法进行了实验验证。第四节对全章进行了总结。

3.1 引言

显著性目标检测试图模拟人类视觉系统来检测自然图像中最突出的物体或区域 [27,28,132]。它广泛用于各类计算机视觉下游任务，如视觉跟踪 [133]，场景分类 [134]，图像检索 [135]，以及弱监督学习 [136,137]。它最近的研究中已经取得了很大进展 [35,42,43,45,138–140]。然而，要在复杂的场景中准确地检测出完整的显著性目标，仍然是巨大的挑战。

在过去几年里，卷积神经网络（CNN）在显著性目标检测中取得了巨大的成功 [32,37,41,44,141,142]。这些网络通常采用多尺度学习来利用高层次的语义特征和细粒度的低层次表征，其中前者能有效地准确定位显著性目标，后者在发现物体细节和边缘方面效果更好。此外，这种多尺度学习是一种自然的解决方案，可以解决实践中的大规模变化。因此，最近在显著性检测方面的许多研究都致力于设计先进的网络架构，以加强网络多尺度学习的能力 [11,34–38,41,42,143]。

现有的显著性目标检测中的多尺度学习方法主要是为了处理低级特征学习，以更好地显式、隐式地捕捉细粒度的物体细节、边缘特征。关于显式地捕捉细粒度的细节，最近许多研究 [6,12,39,57,138,144–149] 试图通过加强边缘表征和直接对预测结果施加边缘监督来提高显著性目标边缘的准确性。为了探索细粒度的细节，许多研究 [34,35,37,38,41–43,45–47] 设计各种多层次的特征融合策

略，以促进高层次语义与低层次细节的融合，例如，被人熟知的 U-Net [17] 或基于编码器-解码器的架构 [34, 41–47]。许多现有的方法可以很好地处理物体的边缘。然而，在进一步提高性能方面的努力已经达到了一个瓶颈期。

为了突破显著性目标检测的这一瓶颈，一种直观的想法是研究多尺度学习的另一个方面，即，*高层次特征学习*，它在场景理解和进一步定位显著性目标方面起着至关重要的作用。遗憾的是，这个方向的研究比较少。为了更好地进行高层特征学习，现有的显著性目标检测方法 [6, 8, 37, 38, 40] 通常直接应用一些常见的模块来开发以用于语义分割，如空洞空间卷积池化金字塔（Atrous Spatial Pyramid Pooling，ASPP）[10] 和金字塔场景理解（Pyramid Scene Parsing，PSP）[7] 模块。然而与语义分割不同，显著性目标检测需要其他类型的高层次特征学习方式。具体而言，语义分割需要学习每个像素和所有其他像素之间的关系，这样就可以根据这种关系做出准确的预测。因此，语义分割方法通常旨在扩大感受野，以提取每个像素的大规模特征 [7, 10, 150, 151]。另一方面，显著性目标检测需要定位显著性目标，这需要对图像的整体理解。有了显著性目标的位置，就可以用解码器轻松恢复物体的细节。与以往的显著性目标检测方法一样，解码器侧重于低层次的特征学习。如图 3.6中所示，由于高层次特征学习的限制，定位显著性目标的准确性最近已经饱和了。总的来说，语义分割需要学习每个像素的全局关系，而显著性目标检测需要学习*整个图像的全局视图*。因此，直接将语义分割方法应用于显著性目标检测，只能达到次优的性能。

为此，本文旨在加强高层次特征学习，这有望为显著性目标检测的未来发展开辟一条新的道路。本文提出了一种“极致下采样模块”（Extremely-Dowsampled Block，EDB）来学习整个图像的全局视图。EDB 逐渐对特征图进行下采样，直到它成为一个大小为 1×1 的特征向量。在这样的下采样过程中，它不断学习更高层次的特征。随着特征图的变小，学习到的特征会变得更加全面。通过逐渐下采样到一个特征向量，它获得了整个图像的全局视图，这样就可以准确定位显著性目标。由于 EDB 在一个非常低的特征分辨率上运行，它的计算开销很小。为了从全局视图中恢复完整的显著性目标，本文建立了一个简单的解码器，从上至下逐步聚合多级特征。为了实现这一目标，本文构建了一个尺度相关金字塔卷积（Scale-Correlated Pyramid Convolution，SCPC），以便在解码器中进行有效的特征融合。与*传统方法*（如 ASPP [10] 和 PSP [7]）*只分别采用多个平行分支提取多尺度特征不同*，SCPC 在各个分支、尺度之间加入了*相关性*。利用 EDB 和 SCPC，

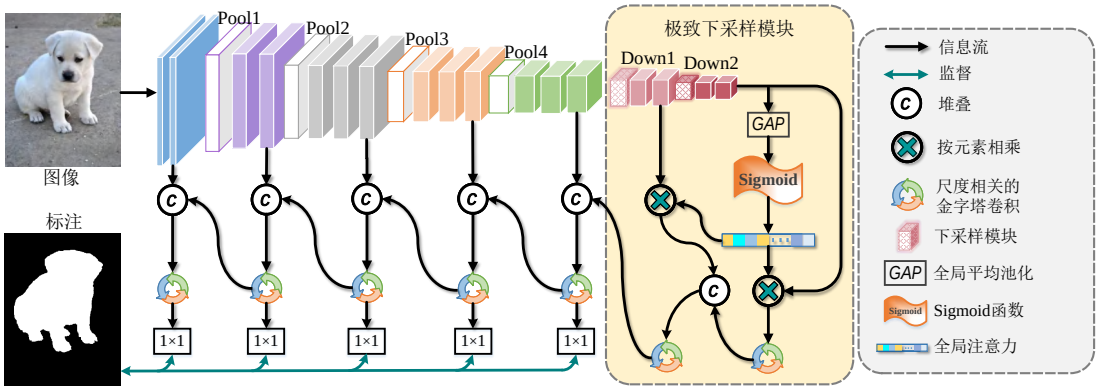


图 3.1 **EDN 的整体网络结构图。**

本文所提出的**极致下采样网络 (Extremely-Downsampled Network，EDN)** 在五个具有挑战性的数据集上以较快的速度和少量的参数达到了最佳性能。为了加快 EDN 的速度，本文用 MobileNetV2 替换了 EDN 的骨干网络 [123] 并构建了一个轻量级网络 EDN-Lite。它在 316fps 的速度下与最近的重型骨干网方法相比，实现了有竞争力的性能。

3.2 方法

在本节中，本文首先在 §3.2.1中对 EDN 方法进行概述。然后，本文在 §3.2.2中介绍了极致下采样技术。最后，本文在 §3.2.3和 §3.2.4中分别介绍 SCPC 和训练中采用的混合损失函数。

3.2.1 方法概述

图 3.1介绍了本文提出的 EDN 的整体结构。由于 VGGs [19]、ResNets [20] 和 MobileNets [123] 的架构都具有类似的 5 个阶段，在不失一般性的前提下，本文以 VGG16 [19] 作为骨干网络的例子来介绍 EDN。本文按照之前的研究 [4, 41–43, 46, 47] 去除最后的池化层和所有全连接层，得到一个全卷积神经网络（Fully Conovolutional Network，FCN）[33] 用于图像-图像的显著性目标检测。到目前为止，VGG16 有 13 个卷积层，由 4 个池化层分开。因此，本文的编码器有五个卷积阶段，其输出分别表示为 E_1, E_2, E_3, E_4 和 E_5，其尺度分别为 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$ 和 $\frac{1}{16}$。

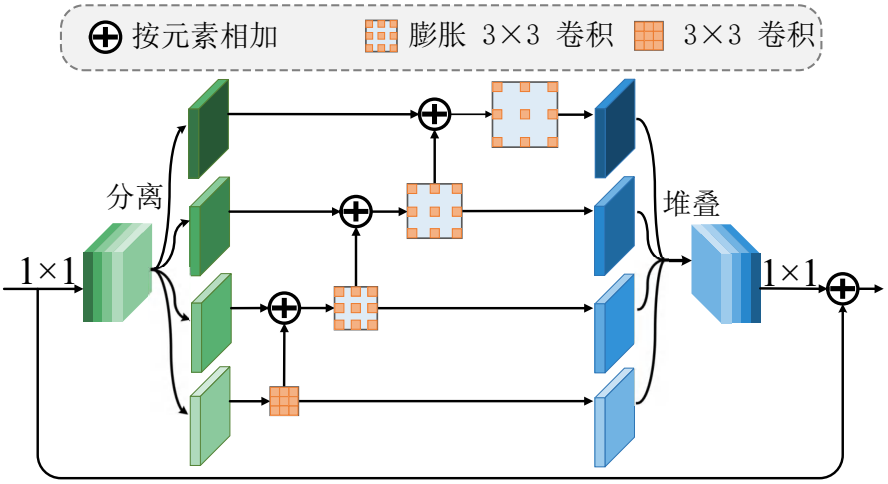


图 3.2 **用于有效特征融合的 SCPC 图解。**

高层特征学习

如上所述，本文提出一个极致下采样块（Extremely Downsampled Block，EDB）来学习整个图像的全局视图。通过应用 EDB，本文可以准确定位显著的物体。假设 $\mathcal{F}$ 表示 EDB 的函数。本文将 EDB 堆叠在 VGG16 之上，输出结果可用下式表示：

$$D_6 = \mathcal{F}(E_5), \tag{3.1}$$

其中 D_6 的尺度为 $\frac{1}{32}$。在这里，本文认为极致下采样通过学习整个图像的全局视图，对显著性目标检测任务有很大好处。EDB 的结构将在 §3.2.2介绍。

多层次特征融合

在 EDB 之后，本文进行自上而下的多级特征整合，以预测具有精细细节的突出性地图。为了完成多级特征的融合。本文构建了尺度相关的金字塔卷积（Scale-Correlated Pyramid Convolution，SCPC）。SCPC 的细节将在 §3.2.3中介绍。本文的解码器由 5 个融合阶段组成。对于每个阶段，本文堆叠 2 个 SCPC，并使用 $\mathcal{H}$ 来作为表示它们。因此，本文的解码器可以简洁地表述为：

$$\begin{aligned} D'_{i+1} &= \text{Upsample}(\text{Conv}_{1\times1}(D_{i+1})), \\ D_i &= \mathcal{H}(\text{Concat}(\text{Conv}_{1\times1}(E_i), D'_{i+1})), \end{aligned} \tag{3.2}$$

其中 $i \in \{1,2,\cdots,5\}$。$\text{Conv}_{1\times1}(\cdot)$ 代表一个 1×1 卷积操作，其后有批量归一化和 ReLU 层。$\text{Upsample}(\cdot)$ 对其输入的特征图按 2 的比例进行上采样。$\text{Concat}(\cdots)$ 将

输入的特征图沿通道维度进行拼接。通过这种方式，本文以一种简洁的方式有效地融合多层次的特征，并获得解码器的输出 D_1, D_2, D_3, D_4, D_5 和 D_6.

3.2.2 极致下采样块

在本章引言中，本文已经讨论了现有的显著性目标检测方法只关注于学习或利用低层次的特征，而忽略了高层次的特征学习。因此，本文提出 EDB 通过学习整个图像的全局视图来加强高层次特征，从而实现更准确的显著物体定位。在这一部分，本文将阐明 EDB 的设计细节。

假设一个 EDB 的输入是 X。本文首先设计一个简单的下采样块，对输入的特征图进行 2 倍的下采样（图 3.1中的 “Down1”）。这可以被表述为：

$$X_1 = \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{3\times3}(\mathrm{Downsample}(X))), \tag{3.3}$$

其中 Downsample$(\cdot)$ 将输入做 2 倍下采样。$\mathrm{Conv}_{3\times3}(\cdot)$ 是一个卷积核大小为 3×3、通道数为 256 的卷积，其后有批量归一化和 ReLU 层。本文重复这个模块而得到 X_2 (图 3.1中的 “Down2”)。X_2 的特征大小很小，因此 X_2 的每个像素有非常大的感受野。为了获得整个图像的全局视图，本文使用全局平均池（Global Average Pooling，GAP）将 X_2 进一步下采样为一个特征向量，可以写为：

$$X_3 = \sigma(\mathrm{GAP}(X_2)). \tag{3.4}$$

X_3 的值范围用一个 sigmoid 函数挤压成 $[0,1]$。尽管 X_3 是输入图像的全局表示，但其单一像素的大小使其不适合从它开始解码。相反，本文把它作为一种自我注意来重新校准 X_2：

$$X_2' = X_2 \odot X_3, \tag{3.5}$$

其中 $\odot$ 代表矩阵元素相乘，X_3 在相乘前被复制成与 X_2 相同大小。本文还采用 X_3 作为非自注意力地来重新校准 X_1，就像式 (3.5) 一样。这样一来，X_1' 和 X_2' 就被全局表示所增强。然后，本文将 X_1' 和 X_2' 融合，可以表述为

$$\begin{aligned} X_2'' &= \mathrm{Upsample}(\mathrm{Conv}_{1\times1}(\mathcal{H}(X_2'))), \\ Y &= \mathcal{H}(\mathrm{Concat}(\mathrm{Conv}_{1\times1}(X_1'), X_2'')), \end{aligned} \tag{3.6}$$

其中 Y 是输出，即 $Y = \mathcal{F}(X)$。Y 将包含整个图像的全局视图，以更好地定位显著性目标。

3.2.3 尺度相关的金字塔卷积

本文构建 SCPC 是为了更好地融合多层次的特征，这也是多尺度学习的一个重要方面。本文的动机来自于现有的模块通常分别进行多尺度特征提取。例如，ASPP [10]，PSP [7]。以及它们的众多变种都分别使用分支来提取多尺度特征。不同的分支负责不同的特征尺度。一个直观的想法是，不同尺度的特征提取应该是相互关联的，相互受益的。假设 M 代表 SCPC 的输入。本文首先应用 1×1 卷积进行过渡：

$$M_1 = \mathrm{Conv}_{1\times1}(M). \tag{3.7}$$

然后，M_1 被分成四个特征图，沿通道维度均匀分布，即：

$$M_2^1, M_2^2, M_2^3, M_2^4 = \mathrm{Split}(M_1). \tag{3.8}$$

接下来，本文以尺度相关的方式进行多尺度学习，可以用下式表示：

$$\begin{aligned} M_3^1 &= \mathrm{Conv}_{3\times3}^{a_1}(M_2^1), \\ M_3^i &= \mathrm{Conv}_{3\times3}^{a_i}(M_2^i + M_3^{i-1}), \quad i \in \{2,3,4\}, \end{aligned} \tag{3.9}$$

其中 $\mathrm{Conv}_{3\times3}^{a_i}(\cdot)$ 是一个 3×3 的膨胀卷积，膨胀率为 a_i。最后，本文将多尺度特征串联起来，并添加一个残差连接：

$$O = \mathrm{Conv}_{1\times1}(\mathrm{Concat}(M_3^1, M_3^2, M_3^3, M_3^4)) + M, \tag{3.10}$$

其中 O 为输出，即 $O = \mathcal{H}(M)$。除了式 (3.10) 将 1×1 的卷积的 ReLU 放在与 M 的残差和之后，SCPC 中的所有卷积都是在批量归一化和 ReLU 激活之后进行的，就像常用的那样 [20]。SCPC 与 Res2Net [98] 的基础模块有类似的联系，但不同点在于，SCPC 通过利用膨胀转卷积来加强多尺度表示的学习。具体而言，SCPC 通过使用小尺度的特征（具有小的膨胀率）来填补大尺度特征（具有大的膨胀率）的漏洞，通过式 (3.9) 来有效学习尺度相关的特征。

3.2.4 损失函数

本文继续介绍 EDN 所使用的的混合损失函数，以更好的完成对 EDN 的训练。$\mathcal{L}$ 代表常用的二元交叉熵（Binary Cross-Entropy，BCE）损失 $\mathcal{L}_{bce}$ 和 Dice

图 3.3 **本文的方法在有或没有 EDB 的情况下的可视化例子。**红, 绿, 和 蓝 颜色的标注显著性图中的像素分别表示显著性目标的中心、边缘和其他像素。

损失 $\mathcal{L}_{dice}$ [152] 的组合，可以用下式进行定义：

$$\begin{aligned}\mathcal{L}_{bce}(P,G) &= G\log P+(1-G)\log(1-P),\\ \mathcal{L}_{dice}(P,G) &= 1-\frac{2\cdot G\cdot P}{||G||+||P||},\\ \mathcal{L}(P,G) &= \mathcal{L}_{bce}+\mathcal{L}_{dice},\end{aligned} \tag{3.11}$$

其中 P 和 G 分别表示算法预测的和人类标注的显著性图。“$\cdot$” 操作表示点积。$||\cdot||$ 表示 ℓ_1 范数。Dice 损失是缓解前背景不平衡的一种有效方法。训练 EDN 的

总损失可以计算为

$$\begin{aligned} P_i &= \sigma(\text{Upsample}(\text{Conv}_{1\times1}(D_i))), \\ L &= \sum_{i=1}^{5} \mathcal{L}(P_i, G), \end{aligned} \tag{3.12}$$

其中 $\text{Conv}_{1\times1}(\cdot)$ 没有批量规范化和 ReLU 激活。Upsample$(\cdot)$ 将预测结果上采样到输入图像的大小。$\sigma(\cdot)$ 是标准的 sigmoid 函数。由于 D_6 的尺寸较小，本文没有在式 (3.12) 中使用它。在测试期间，P_1 被视为 EDN 的最终输出预测。

3.3 实验

3.3.1 实验设置

实施细节。本文所提出的方法是用 PyTorch [162] 和 Jittor [163] 实现的。所有实验的训练都是使用 Adam [164] 优化器进行的，相关参数为 $\beta_1 = 0.9, \beta_2 = 0.99$，权重衰减 10^{-4}，批量大小为 24。本文采用 poly 学习率调度器，即第 n 次训练迭代的学习率是 $init\_lr \times \left(1 - \frac{n}{max\_epoch}\right)^{power}$，其中 $init\_lr = 5 \times 10^{-5}$，$power = 0.9$。一共训练 30 轮。在基于 ResNet [20] 的 EDN 和基于 MobileNetV2 [123] 的 **EDN-Lite** 中，本文分别用瓶颈 [20] 和倒置的残差块 [123] 取代极致下采样中的 $\text{Conv}_{3\times3}$ 块。在 EDN-Lite 中，本文将所有 SCPC 的 $\text{Conv}_{3\times3}$ 操作替换为深度可分离的 3×3 的卷积运算。在训练中，EDN 和 EDN-Lite 的骨干网络都在 ImageNet 上进行了预训练，并且本文按照通常的做法冻结了骨干网络的批量规范化层。在测试中，输入图像都被调整为 384×384 的大小。

数据集。本文在五个数据集上广泛地评估了 EDN。包括 DUTS [153]，ECSSD [154]，HKU-IS [31]，PASCAL-S [155]，和 DUT-OMRON [29] 数据集。这五个数据集分别包括了 15572, 1000, 4447, 850 和 5168 张自然图像与相应的像素级标签图像。依照最近的研究 [36,38,42,143]，本文在 DUTS 训练集（DUTS-TR）上训练 EDN，并在 DUTS 测试集（DUTS-TE）和其他四个数据集上评估。

评价标准。本文用三个广泛使用的指标对 EDN 与以前最先进的方法进行评估。即 F 度量 (F-measure，F_β)，平均绝对误差（Mean Absolute Error，MAE），和加权 F 度量（Weighted F-measure，F_β^w）。F 度量是精度（Precision）和召回率（Recall）的加权谐波平均值：

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{3.13}$$

表 3.1 **EDN 与其他主流显著性目标检测方法的比较结果。**粗体字代表的结果实现了每一列的最佳性能。

方法	速度	# 参数量	DUTS-TE [153]			DUT-OMRON [29]			HKU-IS [31]			ECSSD [154]			PASCAL-S [155]		
	(FPS)	(M)	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE
VGG 骨干网络 [19]																	
DHSNet [44]	10	0.059	0.807	0.705	0.066	-	-	-	0.889	0.816	0.053	94.04	0.906	0.841	0.820	0.731	0.092
ELD [32]	1	43.09	0.727	0.607	0.092	0.700	0.592	0.092	0.837	0.743	0.074	0.868	0.731	0.079	0.770	0.665	0.121
NLDF [141]	18.5	35.49	0.806	0.710	0.065	0.753	0.634	0.080	0.902	0.838	0.048	0.905	0.839	0.063	0.822	0.732	0.098
DSS [11]	7	62.23	0.813	0.700	0.065	0.760	0.643	0.074	0.900	0.821	0.050	0.908	0.835	0.062	0.829	0.742	0.095
Amulet [34]	9.7	33.15	0.778	0.657	0.085	0.743	0.626	0.098	0.897	0.817	0.051	0.915	0.840	0.059	0.807	0.707	0.109
UCF [156]	12	23.98	0.772	0.595	0.112	0.730	0.573	0.120	0.888	0.779	0.062	0.903	0.806	0.069	0.819	0.670	0.127
PiCANet [42]	5.6	32.85	0.745	0.054	0.766	0.691	0.068	0.916	0.847	0.042	0.926	0.865	0.047	0.837	0.852	0.767	0.078
C2S [138]	16.7	137.03	0.811	0.717	0.062	0.759	0.663	0.072	0.898	0.835	0.046	0.911	0.854	0.053	0.843	0.765	0.081
RAS [35]	20.4	20.13	0.831	0.739	0.059	0.785	0.695	0.063	0.914	0.849	0.045	0.920	0.860	0.055	0.828	0.735	0.100
PoolNet [6]	43.1	52.51	0.866	0.783	0.043	0.791	0.710	0.057	0.925	0.864	0.037	0.939	0.735	0.045	0.863	0.782	0.073
AFNet [148]	28.4	35.98	0.857	0.784	0.046	0.784	0.717	0.057	0.921	0.869	0.036	0.935	0.782	0.042	0.861	0.797	0.070
CPD [157]	68	29.23	0.864	0.799	0.043	0.794	0.715	0.057	0.924	0.879	0.033	0.936	0.895	0.040	0.861	0.796	0.072
EGNet [12]	10.7	108.07	0.871	0.796	0.044	0.794	0.728	0.056	0.928	0.875	0.034	0.942	0.892	0.041	0.856	0.788	0.077
GateNet [139]	-	-	0.866	0.785	0.045	0.784	0.703	0.061	0.927	0.872	0.036	0.938	0.788	0.042	0.868	0.797	0.068
ITSD [57]	53	17.08	0.875	0.813	0.042	0.802	0.734	0.063	0.926	0.881	0.035	0.939	0.797	0.040	0.869	0.811	0.068
MINet [140]	22.3	47.56	0.870	0.812	**0.040**	0.780	0.719	**0.057**	0.929	0.889	0.032	0.942	0.811	0.037	0.864	0.808	**0.065**
EDN (Ours)	43.7	21.83	**0.881**	**0.822**	0.041	**0.805**	**0.746**	**0.057**	**0.938**	**0.900**	**0.029**	**0.948**	**0.915**	**0.034**	**0.875**	**0.815**	0.066
ResNet 骨干网络 [20]																	
SRM [38]	12.3	43.74	0.826	0.721	0.059	0.769	0.658	0.069	0.906	0.835	0.046	0.917	0.853	0.054	0.838	0.752	0.084
BRN [143]	3.6	126.35	0.827	0.774	0.050	0.774	0.709	0.062	0.910	0.875	0.036	0.922	0.891	0.041	0.849	0.795	0.072
CPD [157]	32.4	47.85	0.865	0.794	0.043	0.797	0.719	0.056	0.925	0.875	0.034	0.939	0.898	0.037	0.859	0.794	0.071
BASNet [144]	36.2	87.06	0.859	0.802	0.048	0.805	0.751	0.056	0.928	0.889	0.032	0.942	0.904	0.037	0.854	0.793	0.076
PoolNet [6]	40.5	68.26	0.874	0.806	0.040	0.792	0.729	0.055	0.930	0.881	0.033	0.943	0.896	0.039	0.862	0.793	0.075
EGNet [12]	9.9	111.69	0.878	0.814	0.039	0.792	0.738	0.053	0.932	0.886	0.031	0.946	0.903	0.037	0.862	0.795	0.074
GCPANet [158]	51.7	67.06	0.881	0.820	0.038	0.796	0.734	0.057	0.935	0.889	0.032	0.946	0.903	0.036	0.865	0.808	0.063
GateNet [139]	-	-	0.883	0.808	0.040	0.806	0.729	0.055	0.931	0.880	0.034	0.945	0.894	0.041	0.869	0.797	0.068
ITSD [57]	47.3	26.47	0.882	0.822	0.041	0.818	0.750	0.061	0.934	0.894	0.031	0.947	0.910	0.035	0.870	0.812	0.066
MINet [140]	31.1	162.38	0.880	0.824	0.038	0.795	0.738	0.056	0.934	0.897	0.029	0.946	0.911	0.034	0.865	0.809	0.064
EDN (Ours)	51.7	42.85	**0.893**	**0.844**	**0.035**	**0.821**	**0.770**	**0.050**	**0.940**	**0.908**	**0.027**	**0.950**	**0.918**	**0.033**	**0.879**	**0.827**	**0.062**
轻量化方法																	
CSNet [159]	186	0.78	0.804	0.643	0.075	0.761	0.620	0.080	0.896	0.777	0.060	0.912	0.806	0.066	0.826	0.691	0.104
EDN-LiteEX (Ours)	915	1.80	0.836	0.759	0.051	**0.786**	0.716	0.059	0.911	0.857	0.040	0.922	0.869	0.050	0.836	0.755	0.084
EDN-Lite (Ours)	316	1.80	**0.856**	**0.789**	**0.045**	0.783	**0.721**	**0.058**	**0.924**	**0.879**	**0.034**	**0.934**	**0.890**	**0.043**	**0.852**	**0.788**	**0.073**

其中，依照相关的工作 [6, 11, 34, 42]，为了强调检测结果精确性的重要性，本文将 β^2 设为 0.3。本此外，本文使用了在不同的二值化阈值下的最大 F_β，即，$F_\beta^{\max}$。越高的 F-measure 代表越好的算法精度。第二个指标，MAE，衡量预测的

表 3.2 **EDN 和其他主流方法在 S 度量 S_α [160]、最大 E 度量 $E_\xi^{\max}$ [161]、和平均 E 度量 E_ξ^{mean} [161] 上的比较结果。**

方法	速度	# 参数量	DUTS-TE [153]			DUT-OMRON [29]			HKU-IS [31]			ECSSD [154]			PASCAL-S [155]		
	(FPS)	(M)	S_α	$E_\xi^{\max}$	E_ξ^{mean}	S_α	$E_\xi^{\max}$	E_ξ^{mean}	S_α	$E_\xi^{\max}$	E_ξ^{mean}	S_α	$E_\xi^{\max}$	E_ξ^{mean}	S_α	$E_\xi^{\max}$	E_ξ^{mean}
VGG 骨干网络 [19]																	
DHSNet [44]	10	94.04	0.820	0.880	0.855	-	-	-	0.870	0.929	0.905	0.884	0.928	0.909	0.810	0.865	0.845
ELD [32]	1	43.09	0.753	0.835	0.804	0.750	0.826	0.790	0.820	0.897	0.877	0.841	0.900	0.883	0.761	0.821	0.804
NLDF [141]	18.5	35.49	0.816	0.871	0.852	0.770	0.820	0.798	0.879	0.935	0.914	0.875	0.922	0.900	0.805	0.859	0.844
DSS [11]	7	62.23	0.826	0.884	0.851	0.789	0.842	0.811	0.881	0.938	0.907	0.883	0.927	0.903	0.809	0.858	0.847
Amulet [34]	9.7	33.15	0.804	0.852	0.816	0.781	0.834	0.793	0.886	0.933	0.909	0.894	0.932	0.909	0.801	0.847	0.825
UCF [156]	12	23.98	0.782	0.844	0.774	0.760	0.821	0.760	0.875	0.926	0.886	0.884	0.922	0.890	0.802	0.855	0.796
PiCANet [42]	5.6	32.85	0.860	0.907	0.872	0.826	0.866	0.833	0.905	0.949	0.922	0.914	0.947	0.923	0.848	0.896	0.869
C2S [138]	16.7	137.03	0.831	0.886	0.863	0.799	0.845	0.824	0.889	0.940	0.921	0.896	0.937	0.919	0.839	0.889	0.872
RAS [35]	20.4	20.13	0.838	0.889	0.871	0.812	0.858	0.844	0.889	0.941	0.923	0.894	0.932	0.917	0.801	0.854	0.841
PoolNet [6]	43.1	52.51	0.875	0.917	0.888	0.829	0.869	0.841	0.908	0.952	0.927	0.915	0.947	0.927	0.854	0.897	0.879
AFNet [148]	28.4	35.98	0.867	0.910	0.893	0.826	0.861	0.846	0.905	0.949	0.934	0.913	0.947	0.935	0.849	0.895	0.883
CPD [157]	68.0	29.23	0.866	0.911	0.902	0.818	0.856	0.845	0.904	0.948	0.940	0.910	0.944	0.938	0.845	0.888	0.882
EGNet [12]	10.7	108.07	0.878	0.918	0.898	0.836	0.870	0.853	0.912	0.953	0.938	0.919	0.950	0.936	0.848	0.889	0.878
GateNet [139]	-	-	0.870	0.915	0.893	0.821	0.858	0.840	0.910	0.951	0.934	0.917	0.948	0.932	0.857	0.901	0.886
ITSD [57]	53	17.08	0.877	0.919	0.906	0.829	0.866	0.853	0.906	0.950	0.938	0.914	0.949	0.937	0.856	0.902	0.891
MINet [140]	22.3	47.56	0.875	0.917	0.907	0.822	0.856	0.846	0.912	0.952	0.944	0.919	0.950	0.943	0.854	0.900	0.894
EDN (Ours)	43.7	21.83	**0.883**	**0.922**	**0.912**	**0.838**	**0.871**	**0.863**	**0.921**	**0.959**	**0.950**	**0.928**	**0.959**	**0.951**	**0.860**	**0.903**	**0.896**
ResNet 骨干网络 [20]																	
SRM [38]	12.3	43.74	0.836	0.891	0.854	0.798	0.844	0.808	0.887	0.943	0.913	0.895	0.937	0.912	0.834	0.880	0.857
BRN [143]	3.6	126.35	0.842	0.898	0.894	0.806	0.853	0.849	0.894	0.949	0.944	0.903	0.946	0.942	0.836	0.890	0.885
CPD [157]	32.4	47.85	0.869	0.914	0.898	0.825	0.868	0.847	0.905	0.950	0.938	0.918	0.951	0.942	0.848	0.891	0.882
BASNet [144]	36.2	87.06	0.865	0.903	0.896	0.836	0.871	0.865	0.909	0.951	0.943	0.916	0.951	0.943	0.838	0.886	0.879
PoolNet [6]	40.5	68.26	0.883	0.923	0.904	0.836	0.871	0.854	0.915	0.954	0.939	0.921	0.952	0.940	0.849	0.891	0.880
EGNet [12]	9.9	111.69	0.886	0.926	0.907	0.841	0.878	0.857	0.917	0.956	0.942	0.925	0.955	0.943	0.852	0.892	0.881
GCPANet [158]	51.7	67.06	0.890	0.929	0.912	0.839	0.868	0.853	0.920	0.958	0.945	0.927	0.955	0.944	0.864	0.907	0.895
GateNet [139]	-	-	0.885	0.928	0.906	0.838	0.876	0.856	0.915	0.955	0.938	0.920	0.952	0.936	0.858	0.904	0.887
ITSD [57]	47.3	26.47	0.884	0.930	0.914	0.840	0.880	0.865	0.917	0.960	0.947	0.925	**0.959**	0.947	0.859	**0.908**	0.895
MINet [140]	31.1	162.38	0.883	0.927	0.917	0.833	0.869	0.860	0.919	0.960	0.952	0.925	0.957	0.950	0.856	0.903	0.896
EDN (Ours)	51.7	42.85	**0.892**	**0.934**	**0.925**	**0.849**	**0.885**	**0.878**	**0.924**	**0.962**	**0.955**	**0.927**	0.958	**0.951**	**0.865**	**0.908**	**0.902**
轻量化方法																	
CSNet [159]	186	0.78	0.822	0.875	0.820	0.805	0.853	0.801	0.881	0.933	0.883	0.893	0.931	0.886	0.814	0.860	0.815
EDN-LiteEX (Ours)	915	1.80	0.848	0.903	0.882	0.823	**0.867**	0.851	0.894	0.945	0.928	0.899	0.938	0.925	0.820	0.869	0.853
EDN-Lite (Ours)	316	1.80	**0.862**	**0.910**	**0.895**	**0.824**	0.861	**0.848**	**0.907**	**0.950**	**0.938**	**0.911**	**0.944**	**0.933**	**0.842**	**0.890**	**0.878**

显著性图 P 和人类标注的显著性图 G 之间的相似度，可以用下式计算：

$$\text{MAE}(P,G) = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|P_{i,j} - G_{i,j}\right|, \tag{3.14}$$

其中 H 和 W 分别表示显著性图的高度和宽度。MAE 越低代表所测试的显著性

表 3.3 **EDB 各种设计的消融实验**。“GA” 和 “ED” 分别表示全局注意力和极致下采样。

No.	Method	DUTS-TE [153]			DUT-OMRON [29]			HKU-IS [31]			ECSSD [154]			PASCAL-S [155]		
		$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE
1	单独的骨干网络	0.779	0.691	0.065	0.682	0.573	0.094	0.883	0.819	0.049	0.886	0.816	0.068	0.814	0.733	0.091
2	No. 1+ 解码器	0.871	0.816	**0.039**	0.780	0.725	**0.054**	0.932	0.896	0.030	0.938	0.904	0.037	0.864	0.806	0.068
3	No. 2+EDB (只有一个)	0.874	0.820	0.041	0.794	0.741	0.056	0.934	0.899	**0.029**	0.943	0.911	0.035	0.871	0.818	**0.064**
4	No. 2+EDB (没有 GA)	0.876	**0.822**	0.041	0.803	**0.747**	0.056	0.936	**0.901**	**0.029**	0.944	0.912	0.035	0.873	**0.819**	0.066
5	No. 2+EDB (没有 ED)	0.861	0.797	0.047	0.798	0.728	0.062	0.931	0.893	0.031	0.941	0.905	0.036	0.865	0.805	0.068
6	No. 2+EDB (默认)	**0.881**	**0.822**	0.041	**0.805**	0.746	0.057	**0.938**	0.900	**0.029**	**0.948**	**0.915**	**0.034**	**0.875**	0.815	0.066

表 3.4 **在 MAE 指标方面，对有无 EDB 的基线方法进行评估。** “相对提升” 表示应用极致下采样后的相对改善。

设置	评估区域类型	DUTS-TE	DUT-OMRON	HKU-IS	ECSSD	PASCAL-S
基线方法	中心区域	0.084	0.178	0.053	0.053	0.110
+EDB		0.062	0.124	0.043	0.039	0.082
相对提升		**26.6%**	**30.1%**	**19.3%**	**26.4%**	**25.7%**
基线方法	边缘区域	0.243	0.335	0.202	0.195	0.262
+EDB		0.226	0.291	0.196	0.180	0.236
相对提升		7.0%	13.0%	3.1%	7.7%	10.2%
基线方法	其他区域	0.093	0.181	0.073	0.071	0.141
+EDB		0.076	0.133	0.065	0.055	0.112
相对提升		18.1%	26.2%	11.5%	22.5%	20.7%

表 3.5 **在五个数据集上，EDB 与其他替代方案的比较结果。**基线方法指的是去除掉 EDB 后的 EDN 方法。

方案	DUTS-TE [153]			DUT-OMRON [29]			HKU-IS [31]			ECSSD [154]			PASCAL-S [155]		
	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE
基线	0.871	0.816	**0.039**	0.780	0.725	0.054	0.932	0.896	0.030	0.938	0.904	0.037	0.864	0.806	0.068
+ASPP [10]	0.873	0.816	**0.039**	0.790	0.735	**0.053**	0.933	0.893	0.031	0.940	0.902	0.039	0.856	0.800	0.070
+PSP [7]	0.870	0.812	0.042	0.789	0.732	0.056	0.934	0.898	0.030	0.939	0.901	0.038	0.869	0.810	0.068
+NL [165]	0.869	0.815	0.040	0.784	0.725	0.055	0.931	0.896	0.030	0.936	0.902	0.037	0.870	0.809	0.068
+DenseASPP [59]	0.866	0.813	0.040	0.775	0.721	0.056	0.930	0.895	0.029	0.936	0.899	0.038	0.864	0.808	0.065
+EDB	**0.881**	**0.822**	0.041	**0.805**	**0.746**	0.057	**0.938**	**0.900**	**0.029**	**0.948**	**0.915**	**0.034**	**0.875**	**0.815**	**0.066**

目标检测方法就越好。第三个度量，加权的 F-measure F_β^w，解决了 F 度量可能引起的插值缺陷、依赖缺陷和等值缺陷的问题 [166]。本文使用原作者默认设置的官方代码来进行评估。加权的 F-measure 越高，性能就越好。

最近，S 度量 [160] 和 E 度量 [161] 在许多工作中被广泛用于显著性目标检

表 3.6 默认的通道式全局注意力与其他替代方案的比较结果。

方案	DUTS-TE [153]			DUT-OMRON [29]			HKU-IS [31]			ECSSD [154]			PASCAL-S [155]		
	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE
矩阵乘	0.874	0.819	0.042	0.798	0.740	0.059	0.936	**0.902**	**0.029**	0.944	0.913	0.035	0.872	0.815	0.066
空间注意力	0.876	0.821	**0.041**	0.802	0.742	**0.056**	0.937	**0.902**	**0.029**	0.942	0.909	0.036	0.873	**0.817**	**0.065**
通道注意力	**0.881**	**0.822**	**0.041**	**0.805**	**0.746**	0.057	**0.938**	0.900	**0.029**	**0.948**	**0.915**	**0.034**	**0.875**	0.815	0.066

测评价 [140, 167]。本文也采用这两个指标将 EDN 与其他方法进行比较。S 度量计算预测的显著性图和 ground-truth 之间的结构相似度。E 度量计算二值化预测图和二值化 ground-truth 图的相似性。在这里，本文计算所有将预测图二进制化的阈值中的最大和平均 E 度量。本文使用官方代码来计算 S 度量和 E 度量的分值。关于这两个衡量标准的更多细节可以参考相应的指标工作内容 [160, 161]。

3.3.2 实验结果

在这一部分，本文将提出的 EDN 与现有的 20 种今年来的主流方法进行比较，包括 DHSNet [44]、ELD [32]、NLDF [141], DSS [11]、Amulet [34], UCF [156]、PiCANet [42]、C2S [138]、RAS [35]、PoolNet [6]、AFNet [148]、CPD [157]、EGNet [12]、GateNet [139]、ITSD [57]、MINet [57]、BRN [143]、SRM [38]、BASNet [144] 以及 GCPANet [158]。本文使用 VGG16 [19] 和 ResNet-50 [20] 骨干网络对其进行评估。本文还将基于 MobileNetV2 的 EDN-Lite 算法与最近的轻量级显著性目标检测方法 CSNet [159] 进行了比较。为了进一步提高 EDN-Lite 的速度，本文还构建了 EDN-LiteEX，它是以较小的输入大小 (224×224) 测试的 EDN-Lite。由于 DHSNet [44] 使用了 DUT-OMRON [29] 数据集进行训练，本文不报告它在 DUT-OMRON [29] 数据集上的结果。为了公平比较，本文使用原作者提供的显著性图，如果没有提供，本文直接使用它们的官方代码和模型来计算缺少的显著性图。本文还报告了每种方法的速度和参数数量以供参考。速度是使用每种方法的官方代码和单个 NVIDIA TITAN Xp GPU 测试的。

量化比较结果。本文将结果展示在表 3.1 (F 度量, 加权 F 度量及 MAE) 和表 3.2 (S 度量, 最大 E 度量及 E 度量)。此外，本文在图 3.4还对三个最大的数据集，即 DUTS-TE、DUT-OMRON 和 HKU-IS 的速度和准确性进行了可视化比较。在大多数情况下，EDN 达到了最佳性能。而在剩下的少数情况下，EDN 也非常接近最佳性能。EDN 还具有实时速度和相对较少的参数数量。EDN 的轻量级版本 EDN-Lite 与最近的主流方法相比，取得了具有竞争力的性能，且有 10 倍的平

表 3.7 **对 SCPC 的各种膨胀率设置的评估。**最后一行是本文采用的默认设置。

No.	膨胀率设置			DUTS-TE [153]			DUT-OMRON [29]			HKU-IS [31]			ECSSD [154]			PASCAL-S [155]		
	L	H	EH	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE
1	(b)	-	-	0.877	0.824	0.041	0.802	0.745	0.057	0.937	**0.901**	**0.029**	0.945	0.911	0.035	0.870	0.813	0.067
2	(c)	-	-	0.880	**0.825**	0.040	0.804	**0.750**	**0.054**	0.935	0.899	0.030	0.945	0.913	**0.034**	0.874	**0.822**	**0.064**
3	-	(a)	-	0.875	0.820	0.042	0.798	0.740	0.059	0.935	0.900	**0.029**	0.945	0.913	0.035	0.869	0.814	0.067
4	-	(c)	-	0.878	0.824	**0.039**	0.800	0.747	**0.054**	0.935	0.900	**0.029**	0.946	0.913	0.036	0.873	0.818	0.066
5	-	-	(a)	0.873	0.809	0.044	0.803	0.741	0.059	0.933	0.893	0.032	0.943	0.906	0.038	0.872	0.813	0.068
6	-	-	(b)	0.873	0.811	0.045	0.801	0.737	0.061	0.935	0.896	0.030	0.947	0.910	0.036	0.870	0.811	0.070
7	-	-	-	**0.881**	0.822	0.041	**0.805**	0.746	0.057	**0.938**	0.900	**0.029**	**0.948**	**0.915**	**0.034**	**0.875**	0.815	0.066

表 3.8 **在五个数据集上，SCPC 与其他替代方案的比较。**

方案	DUTS-TE [153]			DUT-OMRON [29]			HKU-IS [31]			ECSSD [154]			PASCAL-S [155]		
	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE
卷积	0.837	0.776	0.048	0.740	0.662	0.070	0.919	0.880	0.034	0.924	0.876	0.049	0.855	0.790	0.074
ASPP [10]	0.864	0.805	0.042	0.774	0.712	0.057	0.929	0.890	0.032	0.935	0.899	0.040	**0.868**	**0.806**	**0.068**
SCPC	**0.871**	**0.816**	**0.039**	**0.780**	**0.725**	**0.054**	**0.932**	**0.896**	**0.030**	**0.938**	**0.904**	**0.037**	0.864	**0.806**	**0.068**

表 3.9 **在不同损失函数上的性能表现。**可以发现混合损失比单独的 BCE 或者 Dice 损失要更加适合 EDN 的训练。

No.	损失选择	DUTS-TE [153]			DUT-OMRON [44]			HKU-IS [31]			ECSSD [154]			PASCAL-S [155]		
		$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE	$F_\beta^{\max}$	F_β^w	MAE
1	BCE	0.880	0.817	0.041	0.803	0.741	0.059	0.936	0.895	0.030	0.946	0.908	0.037	**0.875**	0.817	0.066
2	Dice	0.876	**0.835**	**0.038**	0.802	**0.757**	**0.054**	0.934	**0.907**	**0.028**	0.946	**0.921**	**0.033**	0.868	**0.819**	**0.065**
3	BCE + Dice	**0.881**	0.822	0.041	**0.805**	0.746	0.057	**0.938**	0.900	0.029	**0.948**	0.915	0.034	**0.875**	0.815	0.066

均速度。而另一种轻量级方法 CSNet [159] 与最近的主流方法相比，仍然有很大的性能差距。

定性比较结果。定性比较显示在图 3.5。虽然其他竞争者可能无法检测到全部显著的物体或者在困难的情况下甚至找不到一些显著的物体。EDN 能够以清晰的边缘分割出显著性目标。

3.3.3 消融实验

在这一部分中，本文对包含了 EDB 和 SCPC 的 EDN 进行消融研究。这一部分的所有实验都是基于 VGG-16 骨干网 [19] 进行的。其他设置与 §3.3.1相同。

极致下采样的作用。在实验之前，本文首先讨论所提出的极致下采样技术的作用。上文已经阐明，现有的显著性目标检测方法主要集中在学习或更好地利用低层次的细粒度特征来促进多尺度学习。然而，本文通过加强高层特征的学习，探索了多尺度学习的另一个方向，即学习整个图像的全局视图。在此，本

文用数据说明了极致下采样的优点。为此，本文将标注好的显著性图的前景区域划分为边缘、中心区域和其他区域。边缘区域是指与最近的背景像素的欧氏距离小于 5 个像素的前景区域，而中心区域涵盖了与最近的背景像素的欧氏距离在前 20% 的前景像素。其他区域指的是除边缘和中心区域以外的前景区域。这种划分的一些可视化例子显示在图 3.3的第 3 列。

通过上述定义，本文分别计算出中心、边缘和其他区域的平均绝对误差（MAE）。关于指标和数据集的更多细节，请见 §3.3.1。需要指出的是，当本文为一种类型的区域计算 MAE 时，其他两种类型的区域被忽略了。统计结果显示在 表 3.4。本文将 EDN 中的 EDB 删除，作为基线。表 3.4中的相对提升指的是 ΔMAE 占基线的 MAE 的分数的比例，其中 ΔMAE 是在基线中加入 EDB 后 MAE 的减少量。通过应用 EDB，本文观察到中心区域的相对提升要比边缘和其他区域的提升大得多，这表明 EDB 带来的改善主要来自对显著物体的准确定位。图 3.6显示，自 2019 年以来，显著物体的定位精度已经趋于饱和。而 EDB 大大提升了这种准确性。因此，EDB 通过更好的显著性目标定位实现了提高显著性目标检测的目标。此外，有趣的是，本文发现 EDB 在边缘方面也有一些改进，尽管它是为高层特征学习设计的。一个潜在的原因是，强大的高层特征使解码过程更容易，导致低层特征的更好利用。图 3.3中提供了一些可视化的例子。EDB 可以帮助系统检测所有显著的物体。如果没有 EDB，一些显著的物体会完全丢失（第 1，3，4 行）或部分丢失（第 2，5 行）。

此外，本文还参考以往的文献 [5]，定义了一个定位指标来衡量显著物体定位的准确性。本文计算 ground-truth 和预测结果之间的交并比（IoU）。如果 IoU 不优于一个特定的阈值（如 0.7 作为一个严格的阈值 [168]），本文就定义预测结果不能很好地定位显著的物体。因此，如果它们在准确定位显著物体方面存在缺陷，仅考虑如何提取清晰边缘的方法在这个指标上不会得到好的结果。本文在图 3.6中介绍了本文的 EDN 与其他主流方法的比较结果。可以看出，自 2019 年以来，显著性目标定位的准确性已经达到了饱和。

对 EDB 各种设计选择的影响。除了在 §3.3.3展示整个 EDB 的效果之外，本文还对 EDB 的内部设计选择进行了分析。具体而言，本文控制了 EDB 中下采样操作的数量和对输出特征的全局注意力的宽容度。结果总结在表 3.3。“骨干”是指直接从 VGG16 骨干的最后阶段预测的显著性图。“EDB (w/ 1 block)” 表示 EDB 只有一个下采样块（图 3.1中的 Down1）。“EDB (w/o GA)” 表示 EDB 没有全

局注意力（式 (3.4) - 式 (3.5)）。“EDB (w/o ED)” 只删除了下采样操作，但保留了所有的卷积和全局注意力。可以看出，“EDB（默认）” 优于 “EDB w/o GA”，表明全局注意力在 EDB 中很重要。此外，“EDB（默认）” 大大超过了 “EDB（w/o ED）” 和无 EDB 的基线。这体现了下采样和全局注意力在 EDB 中的重要性，去除任意设计都会对 EDB 的性能产生重大影响。

EDB 与其他替代方案的比较。在这里，本文使用其他模块来代替 EDB，进行高层次特征学习, 如 ASPP [10], PSP [7], Non-local (NL) [165], 和 DenseASPP [59] 模块。ASPP, PSP, 和 DenseASPP modules 使用多个独立的分支进行多尺度特征学习。其结果展示在了表 3.5。本文发现，在基线上增加 ASPP、PSP、NL 或 DenseASPP 模块，只能取得稍好甚至更差的性能。相比之下，EDB 比 ASPP、PSP、NL、DenseASPP 和基线都要好很多，体现了本文的极致下采样技术的优越性。

全局注意力的选择。正如在 §3.2.2中所述。本文在计算全局注意力时，默认使用通道注意力，即通道级的按元素乘法作为默认注意力策略。为了验证这一选择的有效性，本文将使用空间注意或矩阵乘法来进行消融研究。结果显示在表 3.6。本文可以观察到，空间注意和矩阵乘法的性能都比默认策略差。因此，默认的通道注意力是最佳选择。

SCPC 的膨胀率设置。 EDN 有六个下采样操作，每次对特征图下采样一半。相应地，有七个 SCPC 模块，其膨胀率是根据特征图的大小来设置的，如图 3.1所示。本文在 表 3.7中展示了 SCPC 的不同膨胀率设置的结果。本文将七次多级特征融合分为三组。“L”（低）包括前两个阶段，输出具有最高分辨率的特征图。“H” (高）包括第 3，第 4 和第 5 阶段。“EH” (极高）包括 EDB 中最后两个额外大小的特征图。对于不同的组，本文应用不同的膨胀率设置。默认情况下，SCPC 中 “L” 组、“H” 组和 “EH” 组的四个分支的膨胀率被分别设置为 {1, 2, 4, 8} (a), {1, 2, 3, 4} (b), 和 {1, 1, 1, 1} (c)。在 表 3.7中，本文为每组尝试了两种其他类型的膨胀率设置。本文可以观察到，结果只随不同的膨胀率而略有波动，表明所提出的 SCPC 对于不同的膨胀率设置是稳健的。由于第 7 的设置在表 3.7中实现了整体的最佳性能，本文采用它作为 SCPC 的默认设置。

比较 SCPC 和其他方案。在这一部分，本文将 SCPC 与原始卷积（“Conv”）、ASPP 进行比较。具体而言，本文首先用等量的 3×3 卷积来代替 SCPC 模块，其输出通道的数量与 SCPC 相同，从而得到一个类似于 U-Net 的解码器 [17]。然

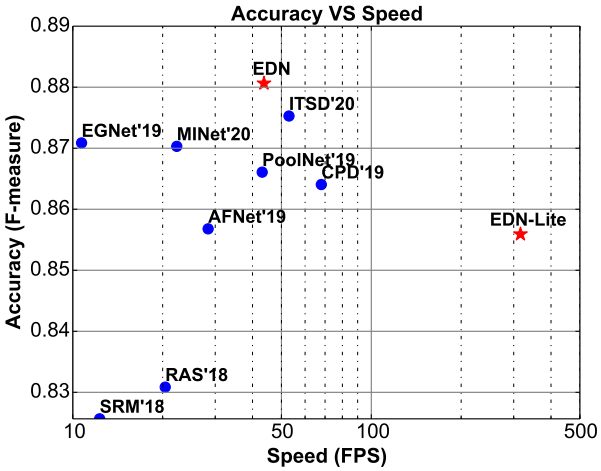
后，本文通过去除 SCPC 中的尺度相关性，用 ASPP 替代 SCPC，即去除 M_3^{i-1} 在式 (3.9) 中的和项。结果显示在 表 3.8。结果显示 ASPP 明显优于 “Conv”，而 SCPC 进一步大幅提高了使用 ASPP 下的算法性能，表明 SCPC 在特征融合中的有效性。

损失函数的讨论。默认情况下，本文使用混合损失，其中包括交叉熵（BCE）损失和 Dice 损失。为了验证这一设计选择，本文还测试了使用单一损失函数（仅 BCE 损失或 Dice 损失）训练的性能。结果显示在 表 3.9。可以看出，Dice 损失可以帮助提高 F_β^w 和 MAE 但降低了 F_β 的得分。由于 F_β 被视为显著性目标检测的主要指标，本文采用 BCE 损失和 Dice 损失的混合方法作为默认设置。

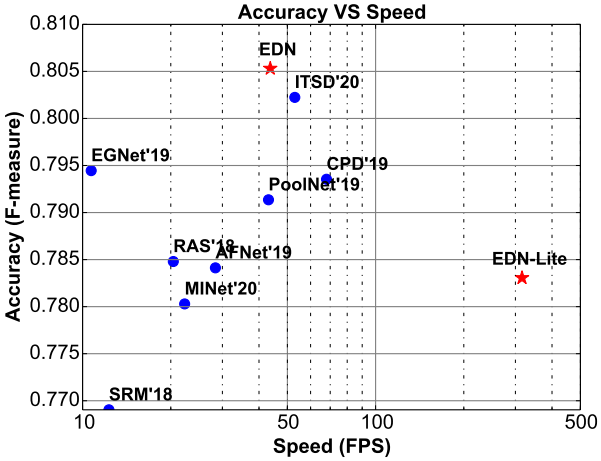
关于失败案例的讨论。尽管本文提出的 EDN 算法在显著性目标检测的全局视图学习方面取得了巨大的成功，但仍有很大的改进空间。本文在图 3.7中提出了一些有代表性的失败案例。可以看出，EDN 在一些令人困惑的情况下失败了。例如，EDN 可能会预测错误的突出区域（在图 3.7中的第 1 号）。EDN 可能预测了最大的显著对象，但没有预测最具鉴别力的显著对象（图 3.7中的第 2、3 号）。EDN 可能将鉴别性的车道视为非显著区域（图 3.7中的第 4 号）。即便如此，在 §3.3.2中定量和定性比较的改进表明，EDN 可以很好地处理大多数情况，取得了最佳性能。

3.4 总结

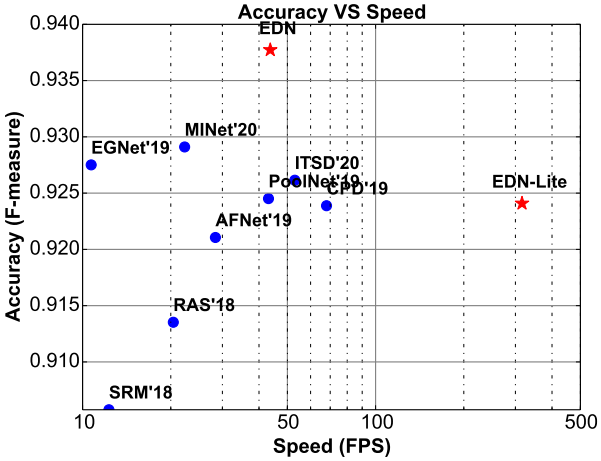
在显著性目标检测中，高层次的语义特征对显著性目标定位是有效的，而低层次的精细细节能很好地捕捉到物体的边缘 [11, 34–38, 41, 42, 48, 143, 169]。这一观察结果引发了对增强低层次特征的广泛研究 [6, 12, 34, 35, 37–39, 41–47, 57, 138, 144–149]，但有趣的是，用于目标定位的高层次特征学习方法几乎没有被深入研究。本文提出了 EDN 网络，通过极致下采样来学习整个图像的更好的全局视图，从而准确地定位显著性目标，扩展了高层次特征学习的研究，提出了新的研究视角。为了配合极致下采样技术，本文还提出了一个尺度相关金字塔卷积策略，以建立一个简洁有效的解码器，从极致下采样的目标定位特征中恢复物体的细节。EDN 网络在五大数据上都相对其他主流方法取得了更好的的性能，其轻量化版本 EDN-Lite 在速度快得多的情况下也取得了非常接近的性能。



(a) DUTS-TE [153]



(b) DUT-OMRON [44]



(c) HKU-IS [31]

图 3.4 与常规显著性目标检测方法的速度和准确性在不同数据集上的比较结果。本文的 EDN 在很大程度上优于其他主流方法，EDN-Lite 与最近主流方法相比也极具竞争力。

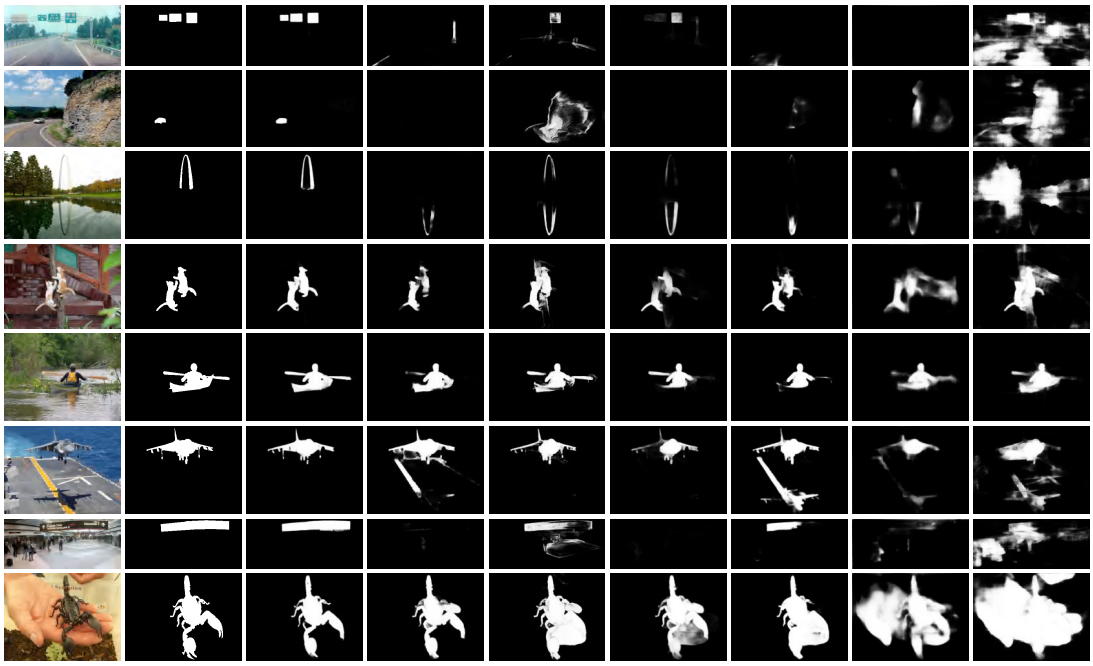


图 3.5 **定性比较。**本文提出的 EDN 方法的预测结果与人类标注上的预测更为相似。

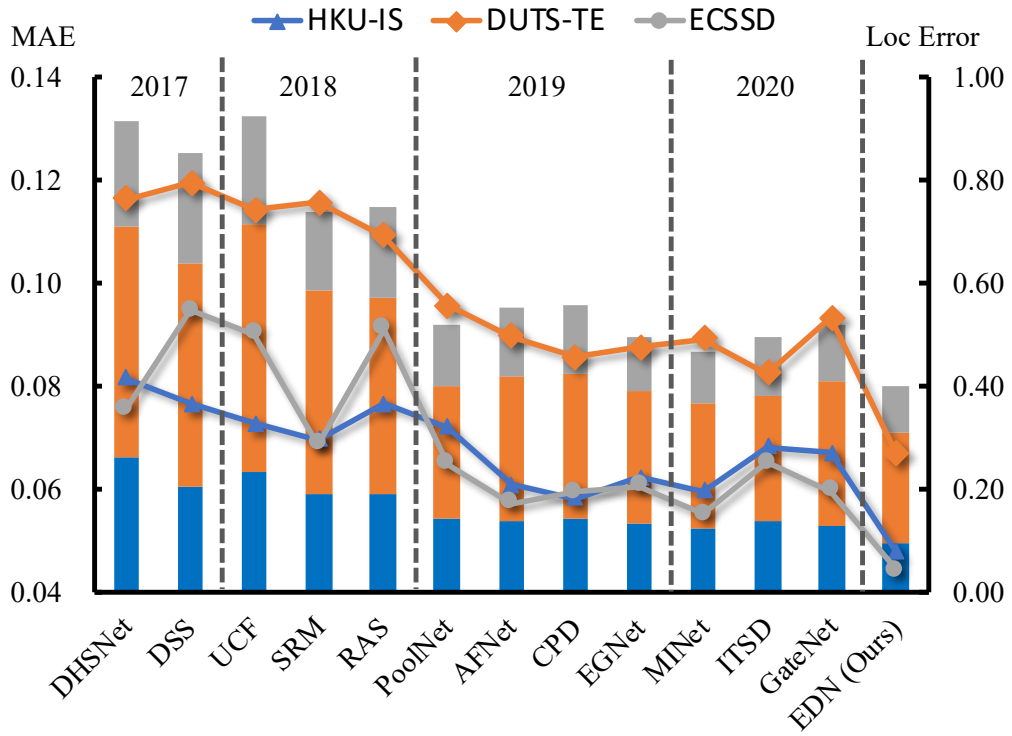


图 3.6 **近期显著性目标检测方法对显著性目标中心的 MAE 和累积定位误差（Loc Error）。**MAE 和 Loc Error 的结果分别以线图和累积直方图的形式呈现。

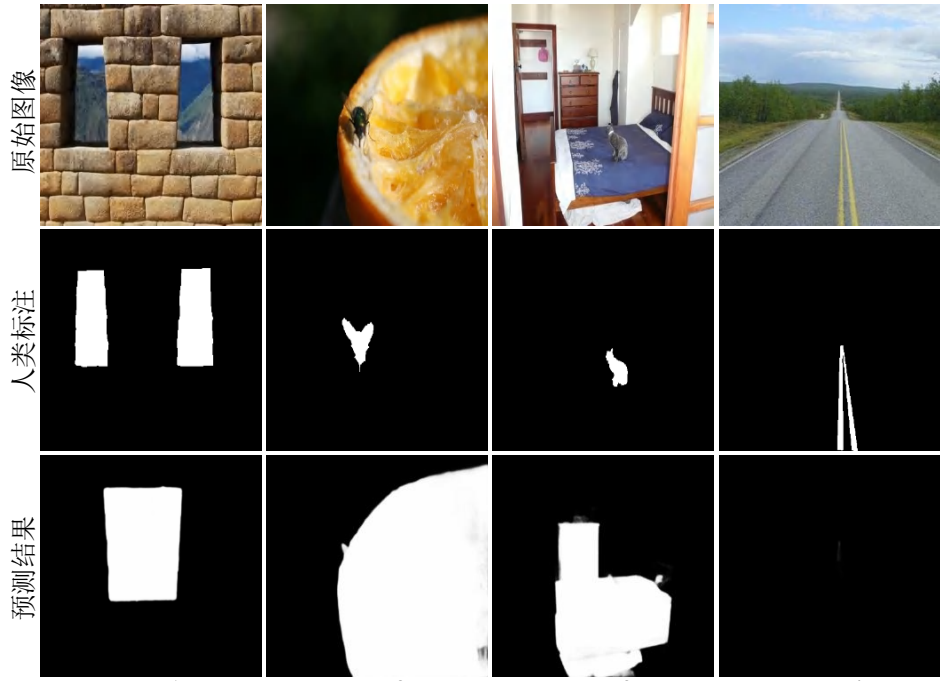


图 3.7 **EDN 的代表性失败案例。**

第 4 章 基于隐式信息恢复的高效融合

目标检测与分割仅使用单种特征很难得到很高的精度，所以需要多通道特征融合的帮助。根据第一章的分析，虽然目前的算法能依靠多通道如 RGB-D 特征融合实现较高的精度，但它们在特征融合方面花费了太多的计算量以保证精度，使得它们在移动端的部署变得困难。本章聚焦该问题提出了基于隐式信息恢复的高效融合技术，可以帮助算法完成高效的多通道特征融合，同时还提高了算法的精度。该技术应用于 RGB-D 显著性目标检测中，提出了一种极致高效的网络 MobileSal。MobileSal 先利用隐式深度恢复使算法仅需在最粗糙的特征层次上完成 RGB-D 特征融合，还提高了算法精度，再利用紧致金字塔细化来融合多尺度的特征，从而恢复边缘清晰的目标。在六大知名数据集上的结果显示，MobileSal 在相比已有方法快 15 ～ 150 倍的前提下取得了与目前算法相当的性能。第一节介绍了相关研究背景、动机。第二节讲述了 MobileSal 网络的结构。第三节对所提出的算法进行了实验验证。第四节对全章内容进行了简要总结。

4.1 引言

显著性目标检测的目的是定位和分割自然图像中最突出的物体或区域。它是目标检测与分割中的基础问题，也是很多计算机视觉任务的前置步骤，如视觉追踪 [170]，图像编辑 [171]，和弱监督学习 [136] 等。目前显著性目标检测方法主要是针对 RGB 图像进行设计 [34, 139, 140]，但它们通常因无法区分前景和背景纹理而无法达到理想结果。为此，研究人员将易采集的深度传感器信息作为 RGB 图像相对应的重要补充，并在 RGB-D 显著性目标检测任务中取得了一定的进展 [9, 66–70, 79, 172]。

虽然卷积神经网络（CNNs）在 RGB-D 显著性目标检测任务中发挥了重要作用，但其高精确度的结果也伴随着模型本身高昂的计算成本和巨大的模型尺寸。这使得大多数方法无法应用于现实世界的场景中，尤其是应用在对计算功耗敏感和计算能力较弱的拥有深度传感器的移动设备上。因此，设计一个能够精确地进行 RGB-D 显著性目标检测的高效网络是十分重要的。为实现这一目标，一个简单的解决方案是采用轻量级骨干网络，如 MobileNets [123, 126] 和

ShuffleNets [124, 125]，进行深度特征提取，而不是采用常用的大型骨干网络，如 VGG [19] 和 ResNets [20]。但这种做法的问题在于，轻量级网络的特征表征学习方面通常不如繁琐的网络强大，这一点已被研究人员广泛验证。这个问题将阻碍轻量级网络在 RGB-D 显著性目标检测任务中取得精确结果，降低了其性能。

为了克服这一问题，本文注意到，如果适当利用图像所对应的深度信息，可以进一步加强 RGB-D 显著性目标检测的特征表达能力 [9, 69]。与现有的一些显式利用深度信息的研究 [66, 68] 不同，本文提出了一种隐式深度恢复（Implicit Depth Restoration, IDR）技术，以促进更高效的特征融合，并加强轻量级骨干网络在特征表示方面的学习能力，从而确保 RGB-D 显著性目标检测结果的准确性。更重要的是，IDR 仅在训练阶段使用，在测试阶段其将被忽略，因此该方法在推理阶段是不需要消耗计算资源的。具体而言，本文强制模型从高层次的骨干网络特征中恢复深度图，通过这种方式，可以对深度流进行监督，使轻量级骨干的特征学习能力变得更加强大。除了 IDR 模块，本文还提出了两个组件以确保模型的高效性：（1）本文只在最粗糙的层次进行 RGB 信息和深度信息的融合，因为最粗糙的层次的特征分辨率较小（输入大小的 1/32），对降低计算成本有着至关重要的作用；（2）本文提出了一个紧凑金字塔细化（Compact Pyramid Refinement, CPR）模块，以有效地融合多尺度的深度特征，实现边缘清晰的显著性目标检测。因为本文所提出的模型是轻量化的，因此该模型被命名为 MobileSal。

以 MobileNetV2 为骨干网络，当输入尺寸为 320×320，本文的 MobileSal 在单个 NVIDIA RTX 2080Ti GPU 上达到了 450FPS，比现有的 RGB-D 显著性目标检测方法快数十倍。在六个具有挑战性的数据集上进行的大量实验表明，与著名的方法相比，MobileSal 在性能上更加具有竞争力（在 NJU2K [173] 和 DUTLF [69] 数据集上的最大 F 度量分别为 91.4% 和 91.2%），而且参数更少 (6.5M)。如此高的效率、精度和小的模型尺寸将使许多现实世界的应用受益。

4.2 方法

在这一节中，本文首先在 §4.2.1对 MobileSal 方法进行了概述 §4.2.1。然后，本文在 §4.2.2介绍了跨模态特征的融合方案，在 §4.2.3介绍了隐式深度恢复技术，在 §4.2.4介绍了紧凑金字塔细化技术。最后，本文在 §4.2.5中提出了混合损失函数。

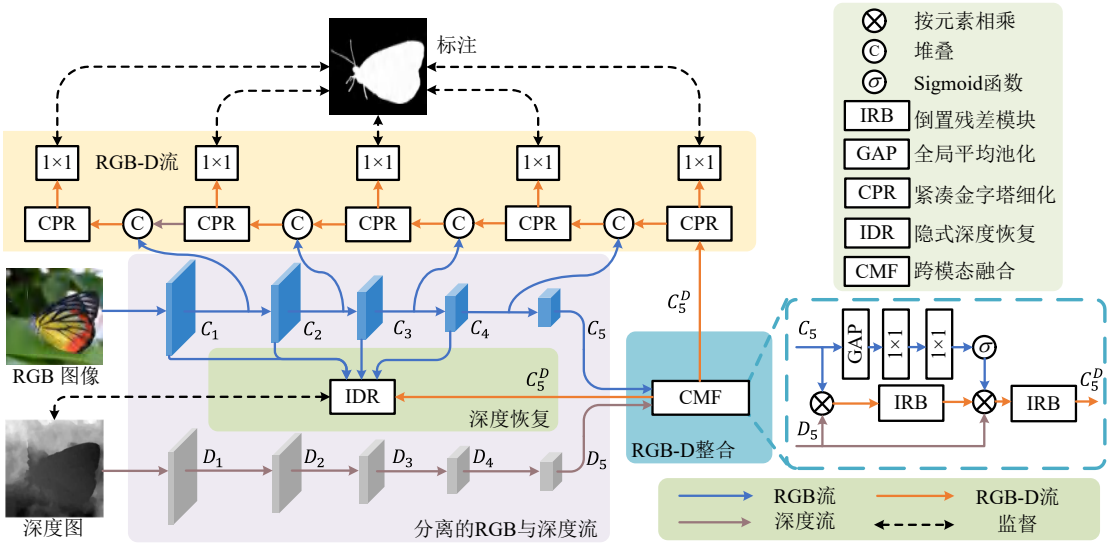


图 4.1 **MobileSal 的总体流程。**

4.2.1 概述

图 4.1描述了本文方法的整体架构。本文分别使用 RGB 信息流和深度信息流进行单独的特征提取。

RGB 信息流。本文采用 MobileNetV2 [123] 作为 MobileSal 的骨干网络。为了使其适应显著性目标检测任务，本文将骨干网络中的全局平均池化层和最后一个全连接层删除。对于 RGB 信息流，每个阶段后都有一个步长为 2 的卷积层，因此在每个阶段后，特征图的分辨率将被降采样为原来的一半。为方便起见，本文把五个阶段的输出映射表示为：$\mathcal{C}_1,\mathcal{C}_2,\mathcal{C}_3,\mathcal{C}_4,\mathcal{C}_5$, 步长分别为 $2,2^2,2^3,2^4,2^5$。

深度信息流。与 RGB 信息流类似，深度信息流也有五个阶段，其步长相同。由于深度图包含的语义信息比相应的 RGB 图像少，本文建立了一个更加轻量的深度网络，其卷积块比 RGB 信息流使用的少。每个阶段只堆叠两个**倒置残差块（Inverted Residual Block, IRB）**[123]。这样的设计降低了计算的复杂度，符合高效 RGB-D 显著性目标检测的目标。在每个 IRB 中，本文首先通过 1×1 的卷积将特征图沿通道维度扩展 M 倍，然后进入输入和输出通道数相同的深度可分离的 3×3 卷积 [126]。接着，通过另一个 1×1 卷积，将特征通道压缩至原来的 $1/M$。在这里，每一个卷积之后都有一个批量标准化（Batch Normalization, BN）[174] 层和 ReLU [175] 层，除了最后一个 1×1 卷积只有一个 BN 层。倒置

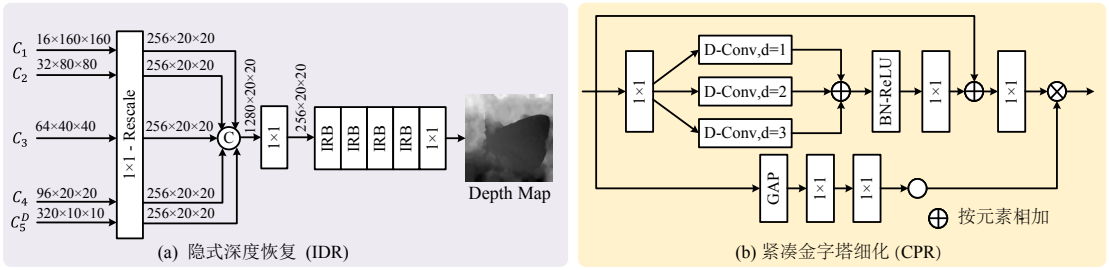


图 4.2 **IDR 和 CPR 的具体结构。**“D-Conv” 代表着深度可分离卷积。

残差块的最终输出是初始输入和初始输入按顺序通过上述三个卷积产生的输出之和。对于每个阶段的第一层，深度可分离卷积的步长被设为 2。深度信息流五个阶段的输出特征图被表示为 $\mathcal{D}_1,\mathcal{D}_2,\mathcal{D}_3,\mathcal{D}_4,\mathcal{D}_5$，前四个阶段的输出分别有 $16,32,64,96$ 个通道。$\mathcal{D}_5$ 和 $\mathcal{C}_5$ 的通道数和步长都相同。

如图 4.1所示，有了 RGB 信息流和深度信息流的输出，本文首先将提取的 RGB 特征 $\mathcal{C}_5$ 和深度特征 $\mathcal{D}_5$ 进行融合，以生成 RGB-D 特征 $\mathcal{C}_5^D$。本文提出的 IDR 技术可以从 $\mathcal{C}_1,\mathcal{C}_2,\mathcal{C}_3,\mathcal{C}_4,\mathcal{C}_5^D$ 恢复深度图，这部分将由输入的深度图进行监督，以加强模型对特征表示的学习能力。对于显著性预测，本文设计了一个以 CPR 模块作为基本单元的轻型解码器。最终预测的显著性图为解码器在最后一个阶段的输出。更多的细节将在之后的章节中进行说明。

4.2.2 RGB-D 跨模态融合

深度图解释了彩色图像的空间线索，这有助于区分前景物体和背景，特别是对于具有复杂纹理的场景。正如之前的研究 [9,69,71,73,79,80] 所关注的，如何将 RGB 特征和深度特征进行融合对于准确的 RGB-D 显著性目标检测是至关重要的。在这里本文主要考虑的是确保本文方法的高效性，所以本文没有像流行的方法那样进行多尺度的 RGB-D 融合 [4,9,69,71,73,79,80]，而只是在最粗糙的层次上融合了 RGB 特征和深度特征，因为小的特征分辨率会降低计算成本。

根据上述分析，本文只融合 RGB 特征图 $\mathcal{C}_5$ 和深度特征图 $\mathcal{D}_5$。如图 4.1所示，本文设计了一个轻量级的**跨模态融合（Cross-Modal Fusion, CMF）**模块来完成高效的 RGB-D 融合。直观地说，语义信息主要存在于 RGB 图像中。深度图则传达了平滑的深度区域先验，这些区域能够大致表示完整目标或物体的形状和结构。因此，本文视深度特征为一个门，将深度特征与 RGB 信息的语义特

征相乘来增强 RGB 信息的语义特征，这个操作可以看作一个强正则化操作。需要指出的是，按元素相加或按特征通道拼接操作只能以平等地对待各特征为基础来融合两个特征图，这些操作与本文的目标是刚好正交而不相符的。§4.3.3中的实验也验证了本文的假设。

具体而言，本文首先将 RGB 特征和深度特征与上述的 IRB 方法结合起来，得到过渡阶段的 RGB-D 特征图，其可以表述为

$$\mathcal{T} = \mathrm{IRB}(\mathcal{C}_5 \otimes \mathcal{D}_5), \tag{4.1}$$

其中，$\otimes$ 是按元素相乘的运算符。同时，本文将 $\mathcal{C}_5$ 输入到一个全局平均池化（Global Average Pooling，GAP）层以获得特征向量，然后用两个全连接层来计算 RGB 信息的注意力向量 $\mathbf{v}$：

$$\mathbf{v} = \sigma(\mathrm{FC}_2(\mathrm{ReLU}(\mathrm{FC}_1(\mathrm{GAP}(\mathcal{C}_5))))), \tag{4.2}$$

其中，FC 和 ReLU 分别表示全连接层和 ReLu 层。FC_1 和 FC_2 的输出通道数与输入通道数相同。σ 表示标准 sigmoid 函数。在 $\mathcal{T}$ 和 $\mathbf{v}$ 被计算出来后，$\mathbf{v}$、$\mathcal{T}$ 和 $\mathcal{D}_5$ 将依次相乘，其结果将被输入 IRB 中：

$$\mathcal{C}_5^D = \mathrm{IRB}(\mathbf{v} \otimes \mathcal{T} \otimes \mathcal{D}_5), \tag{4.3}$$

其中，$\mathcal{C}_5^D$ 表示 CMF 模块输出的特征图。需要指出的是，在乘法之前，$\mathbf{v}$ 将通过复制的方法来达到和 $\mathcal{T}$ 一样的大小。式 (4.3) 再一次通过乘以 $\mathcal{D}_5$ 过滤了 RGB 语义特征，同时通道注意力（channel attention）$\mathbf{v}$ 被用来重新调整融合后的特征。式 (4.3) 在融合了 RGB 特征和深度特征后，可以得出骨干网络的特征，包括 RGB 特征 $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4$ 以及融合之后的 RGB-D 特征 $\mathcal{C}_5^D$。

4.2.3 隐式深度恢复

正如研究人员普遍认为，轻量级的骨干网络的特征表示学习能力比复杂网络更弱 [123, 125]。为了确保 RGB-D 显著性目标检测的准确性，本文希望加强移动网络学习特征表示的能力。本文观察到，深度图传达的深度光滑区域通常代表物体、物体的某部分或光滑的背景。因为从直觉上讲，一个完整的物体或一个相连区域通常具有相似的深度。这一观察促使本文将深度图作为一个额外的监督源来指导模型学习特征表示，这将有助于移动网络抑制物体或相连区域内的纹理变化，并突出它们之间的差异。通过这种方式，显著性物体和背景之间

的差异也将得到加强。根据这个想法，本文设计了一个**隐式深度恢复（Implicit Depth Restoration, IDR**）技术。在这里，本文使用 “隐式” 这个词是因为 IDR 只在训练阶段采用，在测试阶段不运行，这使得它在实际部署中不需要消耗计算资源。

本文继续介绍如何使用 $\mathcal{C}_1,\mathcal{C}_2,\mathcal{C}_3,\mathcal{C}_4,\mathcal{C}_5^D$ 进行上述的辅助监督。如图 4.2（a）所示，IDR 的流程很简单，即只是将多层次的特征图连接起来，然后将其融合。具体地来说，本文首先使用一个 1×1 卷积来将 $\mathcal{C}_1,\mathcal{C}_2,\mathcal{C}_3,\mathcal{C}_4,\mathcal{C}_5^D$ 的通道数压缩到相同的数量，即 256。然后，将得到的特征图调整到与 $\mathcal{C}_4$ 相同的大小，接着将它们连接起来。为节省计算成本，一个 1×1 卷积用来将连接后的特征图从 1280 个通道变为 256 个通道。接下来，按照四个连续的 IRB 来融合多级特征，这样本文就可以得到强大的多尺度特征。最后，本文使用一个简单的 1×1 卷积将融合后的特征图转换为一个单通道映射。通过一个标准的 sigmoid 函数和双线性上采样，本文可以得到与输入图像大小相同的恢复后的深度图。IDR 的训练损失采用了著名的 SSIM 指标 [176] 来衡量恢复的深度图 $\mathcal{D}_r$ 和输入的深度图 $\mathcal{D}_g$ 之间的结构相似度，可以写为：

$$\mathcal{L}_{\text{IDR}} = 1 - \text{SSIM}(\mathcal{D}_r, \mathcal{D}_g), \tag{4.4}$$

其中 SSIM 使用默认设置，即结构化窗口大小为 11×11。需要指出的是，本文在测试过程中省略了上述操作，因此 IDR 在推理过程中将不消耗计算资源。

4.2.4 紧凑金字塔细化

研究人员普遍认为，骨干网络中的高层特征包含语义上的抽象特征，而低层特征则传达了细粒度的细节。为了精确地进行显著性目标检测，充分利用高层次和低层次的特征是至关重要的。目前该方面的文献很多 [9,67,69,79,80]，但现有的方法通常是在不优先考虑效率的情况下设计较为复杂的解码器以达到这个目的。在这里，本文的解码器不仅要有效地融合多层次的特征，而且要尽可能的高效。

本文提出的解码器使用**紧凑金字塔细化 (CPR)** 模块作为基本单元。为了提高效率，CPR 使用 1×1 和深度可分离卷积 [126] 替代之前方法 [67,68,79,80] 中使用的默认的普通卷积。由于多层次特征的表现为多尺度的特征表征，高层次对应较粗的尺度，反之亦然，多尺度学习对于多层次特征融合是必要的。因此，CPR 模块采用了一种轻量级的多尺度学习策略来加强这种融合。本文假设 CPR

模块的输入是 $\mathcal{X}$。如图 4.2 (b) 所示，CPR 首先使用 1×1 卷积将输入的通道数拓展 M 倍。接着，三个膨胀率分别为 $1,2,3$ 的 3×3 深度可分离卷积进行平行连接，以实现多尺度融合。这可以被表述为

$$\begin{aligned}
\mathcal{X}_1 &= \mathrm{Conv}_{1\times1}(\mathcal{X}),\\
\mathcal{X}_2^{d_1} &= \mathrm{Conv}_{3\times3}^{d_1}(\mathcal{X}_1),\\
\mathcal{X}_2^{d_2} &= \mathrm{Conv}_{3\times3}^{d_2}(\mathcal{X}_1),\\
\mathcal{X}_2^{d_3} &= \mathrm{Conv}_{3\times3}^{d_3}(\mathcal{X}_1),\\
\mathcal{X}_2 &= \mathrm{ReLU}(\mathrm{BN}(\mathcal{X}_2^{d_1}+\mathcal{X}_2^{d_2}+\mathcal{X}_2^{d_3})),
\end{aligned} \tag{4.5}$$

其中，d_1、d_2 和 d_3 为膨胀率，分别为 $1,2,3$。BN 是批量标准化的简称 [174]。一个 1×1 卷积将被用来压缩通道，使其通道数与输入相同，即：

$$\mathcal{X}_3 = \mathrm{Conv}_{1\times1}(\mathcal{X}_2)+\mathcal{X}, \tag{4.6}$$

这里使用残差连接以达到更好的优化。式 (4.2) 中的注意力机制被应用于 $\mathcal{X}$ 以计算注意力向量 $\mathbf{v}'$：

$$\mathcal{Y} = \mathbf{v}'\otimes\mathrm{Conv}_{1\times1}(\mathcal{X}_3). \tag{4.7}$$

式 (4.7) 使用了全局上下文信息来重新调整融合后的特征。

如图 4.1所示，在每一个解码器阶段，本文首先使用 1×1 卷积分别把来自顶部解码器和相应编码器阶段的两个特征图的通道数减少到一半，然后将这些结果按通道维度拼接起来，接着用 CPR 模块进行特征融合。通过这种方式，本文的轻量级解码器将融合从上到下的多层次特征。

4.2.5 混合损失函数

如图 4.1所示，在每个解码器阶段，本文将 CPR 模块的输出依次通过一个单通道的 1×1 卷积、一个 sigmoid 函数和双线性上采样层，以预测显著性图。因此，本文可以分别得到五个阶段预测的显著性图 $\mathcal{P}_i(i=1,2,\cdots,5)$。本文设真实标注为 $\mathcal{G}$。每一侧输出的损失可以计算为

$$\mathcal{L}_{sal}^{i} = \mathrm{BCE}(\mathcal{P}_i,\mathcal{G})+\mathrm{Dice}(\mathcal{P}_i,\mathcal{G}), \tag{4.8}$$

其中，BCE 表示二元交叉熵损失函数:

$$\mathrm{BCE}(\mathcal{P}_i,\mathcal{G}) = \mathcal{G}\cdot\log\mathcal{P}_i+(1-\mathcal{G})\cdot\log(1-\mathcal{P}_i), \tag{4.9}$$

表 4.1 **在六个具有挑战性的数据集测试得到的量化结果**。最好、次好和第三好的结果分别用红色、蓝色和**加粗**强调。本文的方法实现了速度-准确度的最佳平衡。

方法名称		DESM	LHM	ACSD	DCMC	CTMF	PCF	TANet	CPFP	DMRA	D3Net	JLDCF	S2MA	UCNet	DANet	BiANet	MobileSal
# 发表年份	[Ref]	2014 [65]	2014 [64]	2014 [173]	2016 [177]	2017 [77]	2018 [178]	2019 [172]	2019 [9]	2019 [69]	2020 [70]	2020 [79]	2020 [80]	2020 [66]	2020 [68]	2021 [179]	2021
参数量 (M)		-	-	-	-	-	133.4	232.4	69.5	59.7	43.2	137.0	86.7	**33.3**	26.7	49.6	6.5
速度 (FPS)		-	-	1	-	8	17	14	6	16	65	9	9	17	32	**50**	450
NJU2K	$F_\beta^{\max}\uparrow$	0.767	0.703	0.749	0.759	0.857	0.887	0.888	0.890	0.896	**0.910**	0.912	0.898	0.908	0.904	0.908	0.914
	MAE ↓	0.286	0.204	0.200	0.171	0.085	0.059	0.060	0.053	0.051	0.047	0.041	0.054	**0.043**	0.047	0.044	0.041
	$S_\alpha\uparrow$	0.671	0.515	0.708	0.686	0.849	0.877	0.878	0.878	0.886	0.900	**0.902**	0.894	0.897	0.897	0.904	0.905
	$E_\xi^{\max}\uparrow$	0.807	0.738	0.814	0.805	0.913	0.924	0.925	0.923	0.927	0.939	0.944	0.930	0.936	0.936	**0.941**	0.942
	排名 ↓	15	15	13	13	12	11	10	9	8	4	2	7	5	6	**3**	1
DUTLF-D	$F_\beta^{\max}\uparrow$	0.728	0.652	0.212	0.419	0.811	0.782	0.804	0.740	0.887	0.748	0.884	0.882	0.836	0.869	**0.885**	0.912
	MAE ↓	0.293	0.162	0.320	0.232	0.095	0.100	0.092	0.100	0.053	0.099	**0.053**	0.054	0.064	0.054	0.048	0.041
	$S_\alpha\uparrow$	0.659	0.568	0.361	0.499	0.831	0.801	0.808	0.749	0.888	0.775	0.906	**0.903**	0.863	0.889	0.906	0.896
	$E_\xi^{\max}\uparrow$	0.800	0.734	0.590	0.654	0.899	0.856	0.861	0.811	0.933	0.834	**0.943**	0.937	0.904	0.931	0.946	0.950
	排名 ↓	13	14	16	15	8	10	9	12	4	11	**3**	5	7	6	2	1
NLPR	$F_\beta^{\max}\uparrow$	0.680	0.693	0.664	0.706	0.841	0.863	0.877	0.888	0.888	0.907	0.925	0.910	0.915	0.907	0.921	**0.916**
	MAE ↓	0.316	0.104	0.163	0.112	0.056	0.044	0.041	0.036	0.031	0.030	0.022	0.030	**0.025**	0.031	0.024	**0.025**
	$S_\alpha\uparrow$	0.573	0.631	0.684	0.729	0.860	0.874	0.886	0.888	0.899	0.912	0.925	0.915	**0.920**	0.909	0.927	**0.920**
	$E_\xi^{\max}\uparrow$	0.808	0.763	0.800	0.795	0.929	0.925	0.941	0.932	0.947	0.953	0.963	0.953	0.956	0.949	0.962	**0.961**
	排名 ↓	16	14	15	13	12	11	10	9	8	6	1	5	4	7	2	**3**
STERE	$F_\beta^{\max}\uparrow$	0.738	0.752	0.682	0.789	0.848	0.875	0.878	0.889	0.895	0.904	0.913	0.895	0.908	0.895	0.908	0.906
	MAE ↓	0.301	0.172	0.197	0.148	0.086	0.064	0.060	0.051	0.047	0.046	0.040	0.051	0.039	0.048	0.042	**0.041**
	$S_\alpha\uparrow$	0.642	0.562	0.692	0.731	0.848	0.875	0.871	0.879	0.886	0.899	0.903	0.890	0.903	0.892	0.904	0.903
	$E_\xi^{\max}\uparrow$	0.811	0.771	0.806	0.819	0.912	0.925	0.923	0.925	0.938	0.938	0.947	0.932	0.944	0.930	0.944	0.940
	排名 ↓	14	14	14	13	12	10	10	9	6	5	2	7	1	7	**3**	4
SSD	$F_\beta^{\max}\uparrow$	0.720	0.633	0.709	0.755	0.744	0.833	0.835	0.801	0.858	0.856	0.860	0.878	0.881	0.878	0.870	0.863
	MAE ↓	0.313	0.195	0.204	0.169	0.098	0.062	0.063	0.082	0.059	0.059	0.053	0.053	0.049	0.050	**0.052**	**0.052**
	$S_\alpha\uparrow$	0.602	0.566	0.675	0.704	0.776	0.841	0.839	0.807	0.857	0.857	0.860	**0.868**	0.866	0.869	0.870	0.862
	$E_\xi^{\max}\uparrow$	0.769	0.717	0.785	0.786	0.865	0.894	0.897	0.852	0.906	0.910	0.902	**0.909**	0.907	0.907	0.907	0.914
	排名 ↓	15	16	14	13	12	9	9	11	8	6	7	**3**	1	**3**	2	5
SIP	$F_\beta^{\max}\uparrow$	0.720	0.634	0.788	0.680	0.717	0.860	0.849	0.869	0.852	0.880	0.903	0.891	0.896	0.900	0.895	**0.898**
	MAE ↓	0.303	0.184	0.175	0.186	0.140	0.071	0.075	0.064	0.086	0.063	0.049	0.057	0.051	0.054	0.051	0.053
	$S_\alpha\uparrow$	0.616	0.511	0.732	0.683	0.716	0.842	0.835	0.850	0.806	0.860	0.880	0.872	0.875	**0.878**	0.884	0.873
	$E_\xi^{\max}\uparrow$	0.770	0.716	0.838	0.743	0.829	0.901	0.895	0.903	0.875	0.909	0.925	0.919	0.919	**0.921**	0.928	0.916
	排名 ↓	14	16	12	15	13	9	10	8	11	7	1	6	4	3	2	5

其中 "·" 表示点乘运算。Dice 表示 Dice 损失 [152]：

$$\mathrm{Dice}(\mathcal{P}_i,\mathcal{G}) = 1 - \frac{2\cdot\mathcal{G}\cdot\mathcal{P}_i}{||\mathcal{G}|| + ||\mathcal{P}_i||}, \tag{4.10}$$

其中 $||\cdot||$ 表示 ℓ_1 范数。本文将每个尺度预测得到的显著性图损失相加，进行深监督。有了深监督和 IDR 分支，训练阶段的损失函数可以被表述为

$$\mathcal{L} = \sum_{i=1}^{5}\mathcal{L}_{sal}^{i} + \lambda\cdot\mathcal{L}_{\mathrm{IDR}}, \tag{4.11}$$

其中 λ 是一个平衡权重。在测试阶段，$\mathcal{P}_1$ 是模型最终预测的显著性图结果。

4.3 实验

本文首先在 §4.3.1中列出了实验环境和设置。接着，本文在 §4.3.2中与著名的 RGB-D 显著性目标检测方法进行了比较，并且在 §4.3.3中进行了全面的消融研究。本文也在 §4.3.3中讨论了 IDR 在其他任务上的应用。

4.3.1 实验环境及设置

实现细节。本文通过 PyTorch [162] 和国内自主研发的计图深度学习框架 [163] 实现了本文的网络。本文使用 MobileNetV2 [123] 作为骨干网络。深度信

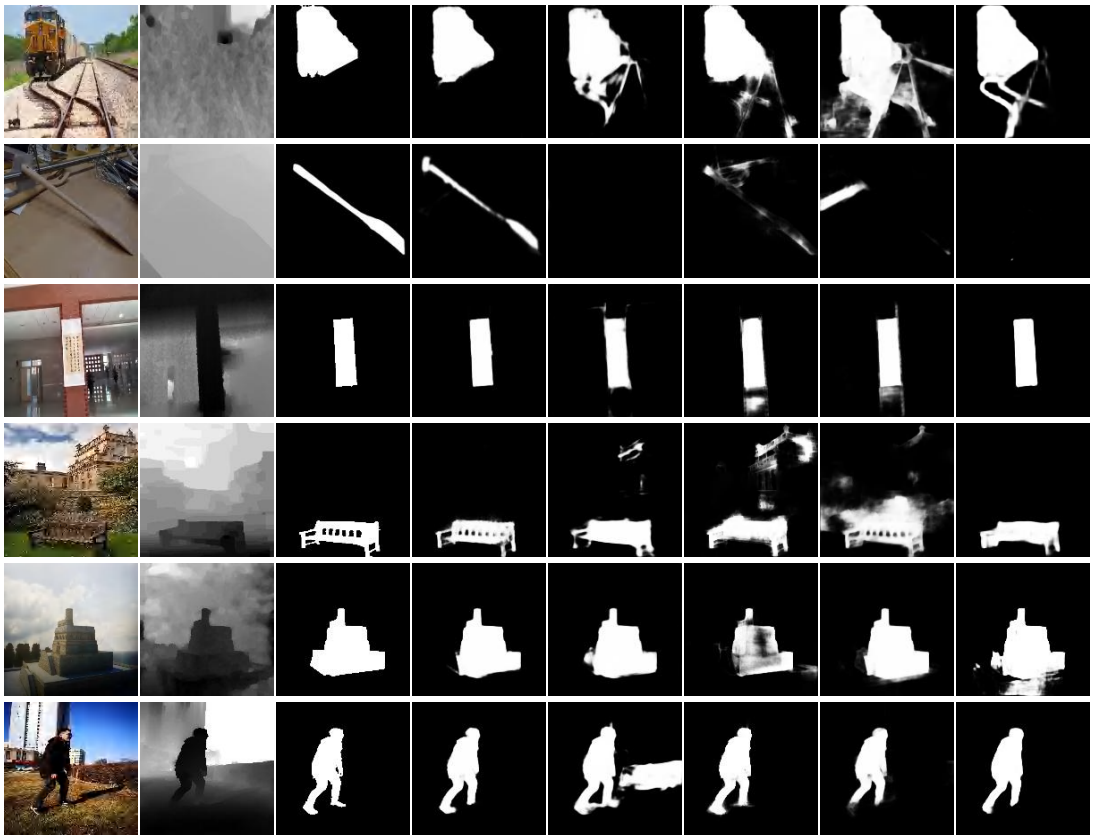


图 4.3 **六个具有挑战性的数据集的可视化比较结果。**结果从上到下分别来自 NJU2K、DUTLF-D、NLPR、STERE、SSD 和 SIP 数据集。

息流、CPR 模块和 IDR 分支的 M 值分别被设置为 4、4 和 6。本文将 RGB 图像和深度图像的大小调整为 320×320。本文使用水平翻转和随机裁剪作为消融实验中默认的数据增广方式。在保持模型设计和参数不变后，本文采用多尺度训练，即 在训练中每个图像的大小被依次调整为 $[256, 288, 320]$，但测试图像的大小不变。本文使用单个 RTX 2080Ti GPU 进行训练和测试。初始学习率 lr 为 0.0001，批图像数量为 10。本文对本文的网络进行了 60 个 epoch 的训练。本文使用 *poly* 学习率策略，因此每个循环 cur_epoch 的真实学习率为 $(1 - \frac{cur\_epoch}{60})^{power} \times lr$，其中 $power$ 为 0.9。本文使用 Adam 优化器 [180] 优化本文的网络。其中，动量（momentum）系数、权重衰减系数、β_1 和 β_2 分别设置为 0.9、0.0001、0.9 和 0.99。

数据集。本文在六个广泛使用的数据集上进行了实验，其中包括 NJU2K [173]、DUTLF-D [69]、NLPR [64]、STERE [181]、SSD [182] 和 SIP [70]。它们分别包含 1985、1200、1000、1000、80、927 张图片。按照 [9,66,68,80]，本文使用

表 4.2 不同方法的 CPU 推理时间。

方法名称	MobileSal	JLDCF [79]	UCNet [66]
输入大小	320 × 320	320 × 320	352 × 352
推理时间 (ms)	**43** (1×)	7246 (150×)	784 (18×)
方法名称	D3Net [70]	S2MA [80]	DMRA [69]
输入大小	224 × 224	256 × 256	256 × 256
推理时间 (ms)	677 (15×)	3049 (70×)	2381 (55×)

表 4.3 恢复到的深度图、不同尺度真实深度图相对真实深度图的相似度。

设置	(a)	(b)	(c)	(d)
尺度	1/16	1/8	1/32	1/8
数据来源	IDR	GT	GT	GT
上采样方式	双线性插值	双线性插值	双线性插值	最近邻插值
PSNR	22.86	30.17	22.55	24.27
SSIM	.8687	.9194	.8445	.8170

NJU2K [173] 的 1500 张图片和 NLPR [64] 的 700 张图片进行训练，NJU2K [173] 的另外 485 张图片和 NLPR [64] 的另外 300 张图片用做测试。除了 DUTLF-D [69]，其他数据集直接用于测试。在 DUTLF-D [69] 数据集上，本文按照 [68, 69]，使用其中的 800 张图片用作训练，400 张图片用做测试。

评估指标。依照最近的研究 [9, 66, 172, 178]，本文采用两个广泛使用的指标对模型进行评估。第一个指标是 F 度量 F_β，根据之前研究的 [9, 68, 69, 183] 建议，β 被设为 0.3 以强调精确性（precision）的重要性。本文计算不同阈值下的最大 F_β 作为 $F_\beta^{\max}$。更高的 $F_\beta^{\max}$ 表明模型性能更优异。第二个指标是平均绝对误差（MAE），该指标越低越好。本文也列出了最近提出的 S 度量 S_α [160] 和在不同阈值下的最大 E 度量 $E_\xi^{\max}$ [161]，以供参考。本文根据官方文献计算了 S_α 和 $E_\xi^{\max}$。本文根据上述四个指标列出每个方法在每个数据集上的总排名。此外，本文还列出了每种方法的参数数量和运行时间，以进行效率分析。

4.3.2 与主流方法的比较结果

本文首先将本文的方法与最近在六个广泛使用的数据集上使用的 15 种著名的方法进行比较。除了基于 ResNet-101 的 JLDCF 和基于 ResNet-50 的 UCNet，

表 4.4 **RGB-D 融合和 IDR 分支的消融研究**。表格第 6 号和第 12 号的策略（用**粗**字体标注）只在最粗糙的特征层次融合了 RGB 和深度特征。

序号	用于融合的特征					IDR	$F_\beta^{\max}$	MAE
	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$			
1	✔	✔	✔	✔	✔	✔	0.899	0.052
2	✔	✔				✔	0.894	0.050
3	✔	✔	✔			✔	0.897	0.047
4		✔	✔	✔		✔	0.902	0.048
5			✔	✔	✔	✔	0.902	0.046
6					✔	✔	**0.906**	**0.045**
7	✔	✔	✔	✔	✔		0.895	0.047
8	✔	✔					0.892	0.049
9	✔	✔	✔				0.896	0.048
10		✔	✔	✔			0.895	0.048
11			✔	✔	✔		0.898	0.048
12					✔		**0.896**	**0.047**
13							0.887	0.052

表 4.5 **RGB-D 融合策略的比较结果**。默认融合策略的结果使用了**粗**字体进行标注。

度量指标	单分支		双分支	
	IDR ✔	IDR ✖	IDR ✔	IDR ✖
$F_\beta^{\max}$	0.900	0.894	**0.906**	**0.896**
MAE	0.048	0.051	**0.045**	**0.047**

大多数方法都是基于 VGG-16 [19] 的。其他方法的显著性图是来自它们公布的结果（如果提供的话），否则是由它们公布的模型计算得到的。

量化比较。表 4.1展示了表格化的比较结果。图 4.4展示了在模型速度上的图形化的比较结果。可以发现，本文的方法的运行速度为 450fps，并且只有 6.5M 的参数。因为其他方法的计算更加复杂，具有更多参数，它们相对本文的方法也慢得多。例如，JLDCF [79] 模型的速度要慢 50 倍，参数要多 20 倍。UCNet 则慢了 26 倍，参数多了 5.5 倍。此外，本文的方法在 NJU2K [173] 和 DUTLF-D [69]

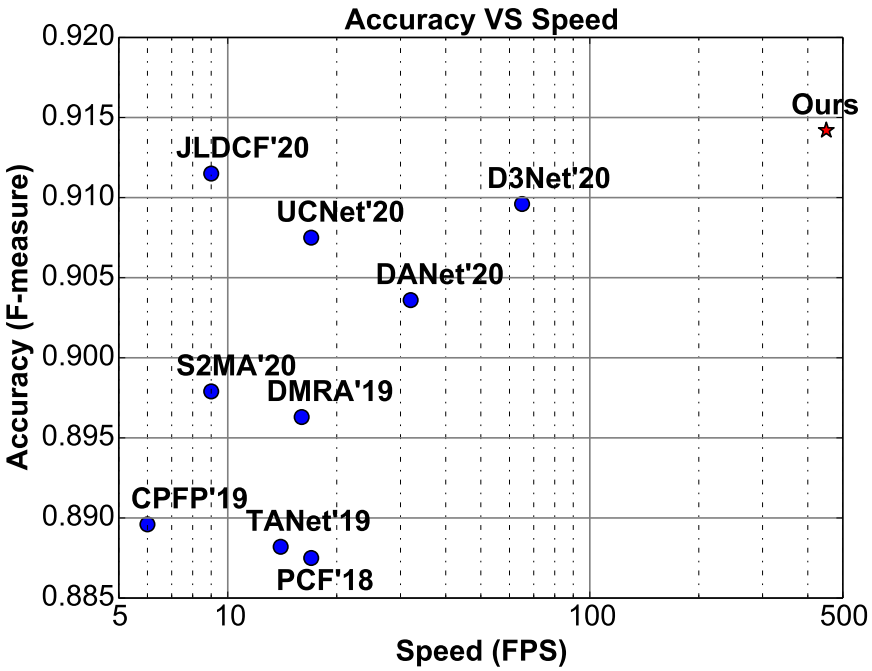


图 4.4 **在 NJU2K [173] 数据集与其他的主流方法在速度上的比较结果。**本文的方法 MobileSal 显示出极具竞争力的算法精度和更快的速度。

数据集上也优于其他方法，在其他 4 个数据集上位列 3 到 5 名。上述结果体现了 MobileSal 方法的高效性和高准确性。

定性比较。图 4.3展示了比较结果。由于篇幅有限，这里本文只将本文的方法与 CPFP [9]、JLDCF [79]、S2MA [80] 和 UCNet [66] 在所有涉及到的数据集上进行比较。本文的方法可以在多种包含噪声深度信息的复杂场景中很好地工作，而其他方法在这些场景中可能会失败。

CPU 推理时间（inference time）。本文在英特尔 i7 8700K@3.7GHz CPU 单核上测试了不同方法的 CPU 推理时间。结果在表 4.2中展示。虽然其他著名的方法的 CPU 推理时间（677 ∼ 7246 ms）远远没有达到实时速度（∼50 ms）的标准，但本文的方法对于每个 RGB-D 输入的 CPU 推理时间仅为 43 毫秒。

4.3.3 消融研究

本文使用 NJU2K [173] 数据集的测试集部分对每个提出的组件进行评估。本文主要使用 $F_\beta^{\max}$ 和 MAE 作为指标。

不同的 RGB-D 融合策略。表 4.4显示了在不同阶段进行 RGB-D 融合的结

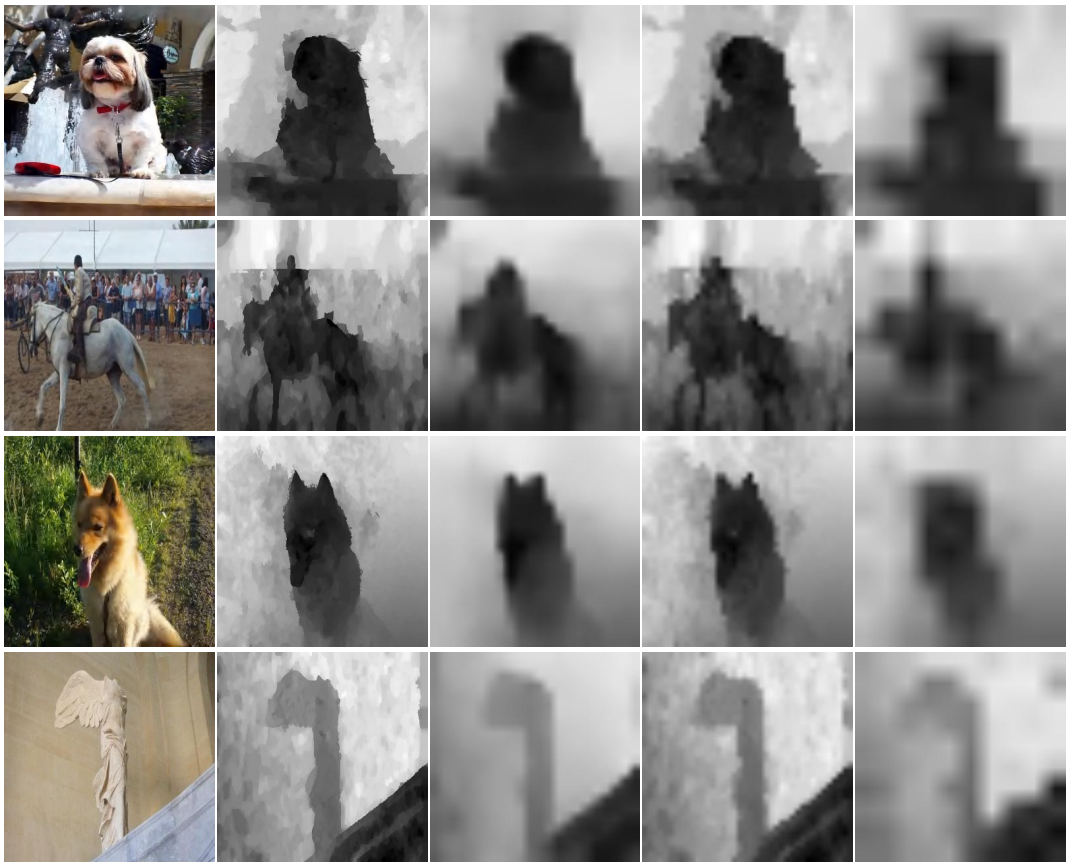


图 4.5 **通过 IDR 对不同尺寸的深度输入映射恢复的深度图在视觉上的比较**。最后 3 列的结果已经通过双线性插值的方法进行了上采样，以保持与输出大小相同。

果。当使用 IDR 分支训练时，模型在最粗糙的层次上融合 RGB 特征和深度特征，其性能最好（No.6）。在没有 IDR 分支的情况下，模型融合后三层的 RGB 和深度特征，其训练结果是最好的（No.11）。在大多数情况下，IDR 分支大幅提高了性能（No. 3 - 6）。且 No.1 和 2（融合了低层次特征）的 MAE 偏高。这验证将融合策略和 IDR 分支的组合是有效的。本文还将所使用的融合策略与早期融合策略进行了比较，后者在输入阶段便将输入的 RGB 图像和深度图连接起来。虽然后者的策略效率更高，但本文的初始融合策略明显优于它（表 4.5）。因此，为了确保准确性和效率，本文在最粗糙的层次上融合了 RGB 和深度特征。

为 RGB-D 融合节省的时间。在 MobileSal 中，本文只在最粗糙的层次上进行 RGB 和深度信息的融合。如果在所有层次都进行 RGB-D 特征融合，本文的方法将达到 260FPS。因此，本文将节省 42% $(1 - \frac{260}{450})$ 的时间。换句话说，只在

表 4.6 λ 系数的消融研究。

No.	1	2	3	4	5	6
λ	0	0.1	0.3	0.5	1.0	2.0
$F_\beta^{\max}$	0.896	0.902	**0.906**	0.902	0.903	0.902
MAE	0.047	0.046	**0.045**	0.046	0.044	0.046

最粗糙的层次融合特征，将使本文的方法加快 73% ($\frac{450}{260}-1$)。

深度信息恢复质量。本文采用主流的图像质量评价指标 PSNR（Peak Signal to Noise Ratio）和 SSIM（Structural Similarity）[176] 来探讨 IDR 中的深度信息恢复质量。IDR 中恢复的深度图的尺寸是输入的 1/16。为了比较，本文评估了通过最邻近插值和双线性插值方法得到尺寸为输入大小 1/8 的深度图的质量。由于 IDR 分支接收尺寸为输入大小 1/32 的深度特征，本文也列出了通过双线性插值得到的尺寸为输入大小 1/32 的真实深度图的质量。表 4.3中展示了结果。本文观察到，恢复后的深度图比通过上采样恢复的尺寸为输入大小 1/32 的真实深度图更接近输入的深度图。本文还在图 4.5中对上述进行了视觉比较。本文可以看到，在 IDR 中恢复的深度图质量良好，并且比通过上采样恢复的尺寸为输入大小 1/8 的真实深度图的噪声更小。

IDR 损失函数的选择。如 §4.2.3所述，本文摒弃琐碎的 L1/L2 损失函数，选择 SSIM 指标 [176] 作为 IDR 分支的损失。本文通过在 L1/L2/SSIM 损失函数下训练本文的 MobileSal 模型来验证这一点。本文发现，就 $F_\beta^{\max}$ 而言，采用 L1/L2 损失函数的 IDR 分支可以将模型的性能提高 0.5%/0.7%，比采用 SSIM 损失函数的 IDR 分支低 0.5%/0.3%。这是因为 SSIM 损失函数提供的是结构相似性，而不是 L1/L2 损失函数计算的简单的点对点误差。基于上述讨论，本文选择 SSIM 损失函数以监督 IDR 分支的训练。

损失函数中的 λ 常数。如式 (4.11) 所述，λ 决定了深度恢复损失函数所占的权重。本文使用不同的 λ 常数对本文的方法进行了实验。结果在表 4.6中。在不同的 λ 下，IDR 分支为本文的方法的稳健性带来了实质性的改善。由于第三列的性能最好，所以本文采用 $\lambda = 0.3$ 作为模型训练时的默认设置。

深度信息和 CMF 模块。在表 4.7中的结果展示了深度信息和 CMF 模块的效果。有深度图输入的结果是使用 IDR 分支训练的。没有 CMF 模块的深度图只使用了 RGB-D 融合中的元素级乘法。结果表明，即使是非常简单的操作，深度

表 4.7 **跨模态融合的有效性**。

No.	RGB 特征	CPR	深度特征	跨模态融合	$F_\beta^{\max}$	MAE
1	✔				0.852	0.068
2	✔	✔			0.887	0.052
3	✔	✔	✔		0.894	0.048
4	✔	✔	✔	✔	**0.906**	**0.045**

表 4.8 **对式 (4.1) 中初始 RGB-D 融合进行不同操作的比较**。“乘” 和 “加” 的操作是按元素计算的。拼接操作是沿着通道进行的。默认融合策略的结果是使用 **粗**字体标注。

操作	乘		加		拼接	
指标	$F_\beta^{\max}$	MAE	$F_\beta^{\max}$	MAE	$F_\beta^{\max}$	MAE
结果	**0.906**	**0.045**	0.897	0.048	0.900	0.046

信息也对 RGB-D 显著性目标检测非常有帮助。更具体地说，本文也观察到使用 CMF 模块后模型的实质性改进。

CMF 模块中的初始 RGB-D 融合操作。正如在式 (4.1) 设计的那样，按元素相乘被用在初始的 RGB-D 融合操作中。为了验证所选操作的有效性，本文将这种设计与广泛使用的按元素级相加进行比较。结果在表 4.8中。就 $F_\beta^{\max}$ 而言，按元素相乘操作在很大程度上优于按元素级相加和拼接操作，分别高出 0.9% 和 0.6%。这表明按元素相乘在 CMF 模块的初始 RGB-D 融合中具有较大优势。

CMF 模块的混合策略。如式 (4.3) 所述，RGB 信息的注意力向量 $\mathbf{v}$ 和 RGB-D 特征 $\mathcal{T}$ 都将和深度信息特征 D_5 进行融合。本文通过去除 $\mathbf{v}$ 或 $\mathcal{T}$ 来验证上述设计的有效性。实验结果在表 4.9中。本文观察到在去除 $\mathcal{T}$ 后，模型的性能下降了很多。这是因为在式 (4.3) 的特征融合中，只有 $\mathcal{T}$ 包含空间级的 RGB 特征。此外，$\mathbf{v}$ 是一个由 RGB 特征产生的注意力向量，而 D_5 是一个纯粹的深度特征。去除 $\mathbf{v}$ 也对模型的性能产生了一些影响，因为它在 RGB-D 融合中提供了通道级的 RGB 注意力。

紧凑金字塔细化。表 4.10显示了 CPR 的结果，其中使用了不同的膨胀策略。本文分别测试了默认设置（No. 1）、不同膨胀率的单一卷积（No. 2 - 4）以及膨胀率稀疏组合的多种卷积（No. 5、6）。默认设置的紧凑膨胀率（1，2，3）明显优于其他设置，说明在 MobileSal 中紧凑金字塔细化是必要的组件。

混合损失函数。为了验证 Dice 损失函数的有效性，本文测试了只用二元交

表 4.9 **式 (4.3) 中的 CMF 模块的融合分析**。No. 1：通过元素级乘法实现的 RGB-D 融合操作；No. 2-3：在式 (4.3) 中去除 $\mathbf{v}$ 或 $\mathcal{T}$ 的变体；No. 4：默认设置。

No.	深度输入	CMF	$\mathbf{v}$	$\mathcal{T}$	$F_\beta^{\max}$	MAE
1	✔				0.894	0.048
2	✔	✔	✔		0.895	0.049
3	✔	✔		✔	0.902	0.046
4	✔	✔	✔	✔	**0.906**	**0.045**

表 4.10 **CPR 膨胀率的消融研究**。可以发现紧凑的膨胀率（No.1）取得了最佳的性能。

No.	1	2	3	4	5	6
膨胀率	1, 2, 3	1	2	3	1, 3, 6	1, 4, 8
$F_\beta^{\max}$	**0.906**	0.900	0.892	0.897	0.903	0.901
MAE	**0.045**	0.047	0.048	0.047	0.046	0.048

叉熵损失函数训练的模型的性能。本文发现，使用 Dice 损失函数将提供高对比度，使 MAE 降低 0.1% ∼ 0.2%，但其不会影响 $F_\beta^{\max}$。

IDR 在其他任务中的应用

对于推理阶段的 RGB-D 输入，IDR 分支加强了骨干网络的特征表示，且不需要消耗计算资源，它不限于仅用于 RGB-D 显著性目标检测。为了显示 IDR 分支在其他任务上的潜力，本文评估了 IDR 分支应用在 RGB-D 语义分割中带来的性能增益。这项任务的目标是给每个像素分配语义标签，与预测每个像素的显著性概率的 RGB-D 显著性目标检测相似。

实验环境及设置。本文选取两个最近的具有的代表性的方法 [184, 185] 作为基线方法，并使用作者提供的官方代码来进行实验。根据 [184, 185]，本文在 NYUDv2 [186] 数据集上进行了实验，该数据集由 1449 张 RGB 图像组成，每张图片有对应的深度图和像素级语义标签，共包含 40 个语义类别。在这个数据集中，795 张图片用于训练，654 张图片用于测试。训练和测试的设置均按照官方文件进行 [184, 185]。和 MobileSal 类似，对于 Chen 等人 [185] 的工作，由每个阶段的深度特征重新校准的 RGB 特征将被送入 IDR 分支。但是，对于 SGNet [184]，和 MobileSal 不同的是，本文只将前四个阶段的输出特征输入到 IDR 分支。这是因为在 SGNet [184] 中，最后一个阶段的特征将直接用于预测语义分割的结果。

表 4.11 **IDR 分支对 RGB-D 语义分割方法的影响。**

方法	SGNet [184]			Chen 等人 [185]		
评价指标 (%)	mIoU	Acc	mAcc	mIoU	Acc	mAcc
基线结果	49.6	75.6	61.9	51.4	77.1	62.9
+ IDR	**50.5**	**76.3**	**62.2**	**52.2**	**77.3**	**64.0**
提升值	+0.9	+0.7	+0.3	+0.8	+0.2	+1.1

评估指标。根据之前的工作 [184, 185]，本文使用平均 IoU（mIoU）作为主要评价指标。本文也列出了像素精度（Acc）和平均精度（mAcc）供参考。关于上述三个指标的计算方法，可以参考 SGNet [184] 的详细介绍。

实验结果。表 4.11显示了 IDR 在 RGB-D 语义分割任务上的提升效果。在加入不需要消耗计算资源的 IDR 分支之后，本文观察到两种方法的性能 [184, 185] 在 mIoU 指标方面都有很大的改善。这表明，IDR 的理念也适用于 RGB-D 语义分割，而且其功能强大，不需要消耗额外的计算资源。

4.4 总结

为了实现高效的 RGB-D 显著性目标检测，本文提出了一个新的方法 MobileSal，其速度达到了 450FPS。本文提出了隐式深度恢复技术来保证 RGB-D 特征融合的深度信息不被丢失，且仅需在最粗糙的层级进行 RGB-D 特征融合，极大地降低了计算成本。本文还在六个主流的数据集上与其他著名的方法进行实验比较，结果显示 MobileSal 在参数量更小、速度提升 15 ～ 150 倍的前提下，仍然保持了与其他著名方法相当的精度性能。本文也提供了详细的消融实验，分析了 MobileSal 中每个组件的有效性，以及隐式深度恢复技术在 RGB-D 语义分割中的有效性。作者计划将来把 MobileSal 应用到其他的目标检测与分割任务中。

第 5 章 基于注意力融合的图像二元感知

标注数据是训练目标检测与分割方法的基础。根据第一章的分析，许多实际场景下，标注数据是难以获取的。其中，医学图像是极具代表性的场景类型，它的特征与常见的自然图像还存在较大差异，比如自然图像通常有丰富的背景，而医学图像只显示了人体相关的组织。首先，在图像规模上，因为数据隐私原因，医学图像的规模也远远小于自然图像，难以获得大规模的数据集。其次，在标注成本上，因为经验丰富的医学专家十分稀缺，所以医学图像的标注成本也十分高昂。为了解决标注数据获取难的问题，本章提出了基于注意力融合的图像二元感知方法，并应用于新冠肺炎 CT 病灶分割。具体而言，该方法利用组化膨胀金字塔卷积，获取更丰富的不同感受野的特征图，从而提高病灶的定位准确率。它再利用注意力融合技术，将不同感受野的特征图进行融合，在融合过程更注重病灶的定位信息，防止病灶特征丢失，从而提高病灶的分割准确率。接着，它利用分类和分割的图像二元感知，额外利用了大量类别标签数据训练的分类特征，使用注意力融合技术进一步提高了病灶的分割准确率。它应用于新冠肺炎 CT 病灶分割任务中，与已有的先进方法相比，它具有最佳的性能。本章中，第一节介绍了新冠肺炎 CT 病灶分割的相关研究背景以及研究动机。第二节讲述了解决方案 JCS 方法。第三节讲述了本文如何构建新冠肺炎的数据集。第四节对 JCS 方法在新冠肺炎 CT 病灶分割任务上进行了实验验证。第五节对本章内容进行了总结。

5.1 引言

为了阻止新冠肺炎的迅速传播，医学图像分析被用作 RT-PCR 检测的重要补充工具 [187]，这是因为大多数新冠肺炎患者的胸部 X 射线影像临床表现出肺部感染 [188]。相关工作 [189, 190] 也显示了使用 CT 图像扫描检测新冠肺炎具有较高敏感度。此外，CT 图像扫描测试是监测疾病严重性的必要工具 [191]。然而，CT 图像扫描测试所需的诊断时间是其主要的应用限制：经调查 [192]，即使是有充分经验的放射科医生也需要 21.5 分钟来分析每个病例的 CT 检测结果。而在传染病大流行爆发期间，经验丰富的放射科医生严重缺乏，难以及时识别

潜在感染患者。因此，自动诊断系统的需求十分迫切，利用它可以过滤健康人员或自动分割病灶。大大缩短医生的诊断时间。

由于深度卷积神经网络强大的识别能力，人工智能技术正在改变医学图像处理界。深度卷积神经网络通常需要在大规模数据集上进行训练才能显示其能力，然而，由于大多数现有的新冠肺炎的 CT 影像扫描数据集 [193, 194] 仅包含几十个病例的百余张 CT 图像，故这些数据集不能满足深度卷积神经网络的训练需求。而且，大多数数据集仅提供患者级别的标签（即类标签）表明此人是否被感染，但这些数据集缺少像素级别的精细标注。因此，用这些数据集训练的 CNN 模型通常忽略了解释最终预测的有价值的信息。尽管已经建立了多个 CT 影像扫描诊断系统 [190, 195–199] 来检测疑似新冠肺炎病例，但大多数都有两个缺点：（1）它们是在小规模的数据上训练的，因此对于新冠肺炎感染检测来说不够鲁棒；（2）由于病灶特征的高度可变性以及病灶与周围组织的低对比度，它们往往难以分割出完整的病灶区域。

为了缓解上述缺点，本文（1）提出了一种基于注意力融合的图像二元感知模型（Joint Classification and Segmentation，JCS）来为与新冠肺炎战斗的医护人员提供精细化的 CT 病灶分割结果；（2）建立了一个包含标签级别和像素级别的大规模数据集 COVID-CS 来对 JCS 进行更充分的训练；具体而言，JCS 通过注意力融合技术，将不同感受野的特征图进行融合，且在融合过程中更侧重保留病灶的高层特征，防止病灶丢失，从而提高了病灶的分割精度。通过额外的分类模型，JCS 利用了大量的类别标签标注的健康和患病数据，进一步提高了分割模型的性能。实验证明，JCS 在 COVID-CS 数据集上相对其他方法实现了显著（8.8%）的提升，且有极高的鲁棒性，仅在 0.8% 的测试样本失效。利用分类模型结果和相应的细粒度分割结果，基于 JCS 的诊断系统大大简化和加速了放射科医生或其他医学专家的诊断过程。本文还在临床上对基于 JCS 的诊断系统进行了速度和精度测试。实验证明，基于 JCS 的诊断系统仅需 22 秒来判断一个感染病例，仅需 1 秒来排除一个未感染病例，比 RT-PCR 测试和经验丰富的放射科医生使用 CT 图像扫描分析快得多。在本文 JCS 诊断系统的协助下，有经验的放射科医生仅需要 55.4 秒（32.4 秒用于医生诊断，22.0 秒用于 JCS 推理）判断一个感染病例，需 1 秒来排除一个未感染病例，并保持与纯人工相同的特异性和敏感性。因此，基于 JCS 辅助的诊断速度提升和诊断精度的相同表明了本文 JCS 系统的优越性。

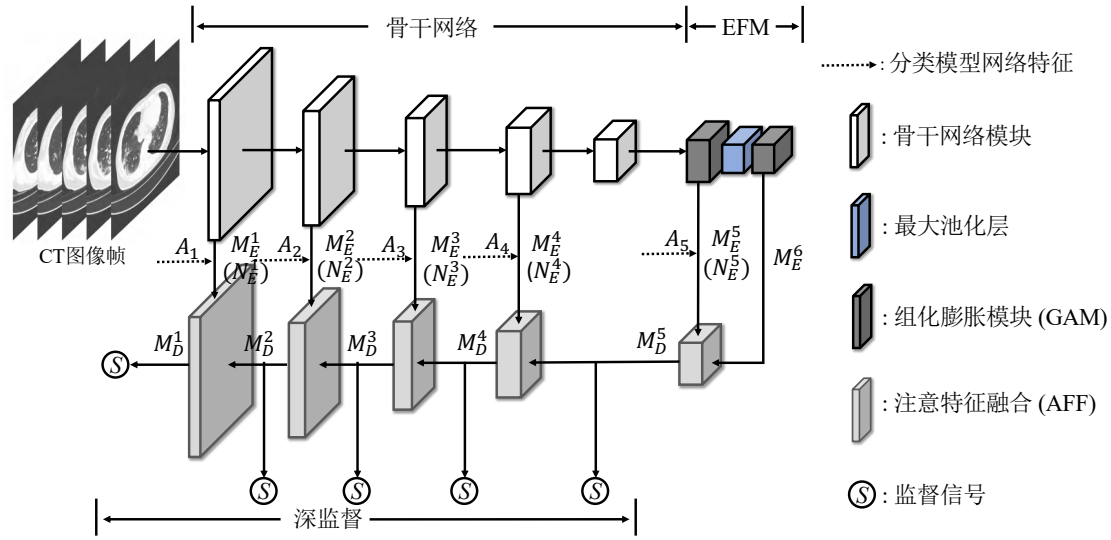


图 5.1 **JCS 网络结构图。** EFM 代表增强特征模块（§5.2.1）。AFF 代表注意特征融合策略（§5.2.1）。如果不与分类模型相结合，$M_E^1 \sim M_E^5$ 将被输入解码器；否则，结合后的 $N_E^1 \sim N_E^5$ 将被输入解码器（图 5.3，§5.2.2）。本文运用深层次的监督来训练本文的分支（§5.2.1）。

5.2 方法

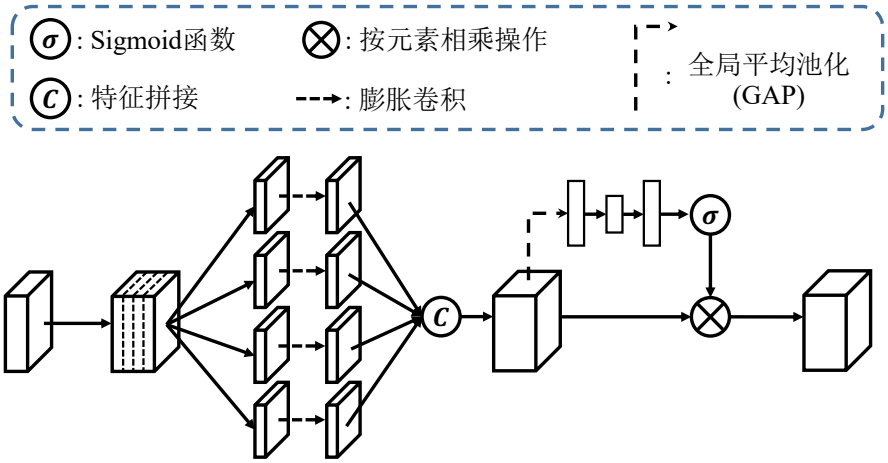
混浊（opacification）是新冠肺炎患者 CT 病灶的基本表现 [200]，它体现的是肺实质的密度增高现象 [201]。本节的 CT 新冠肺炎病灶目标分割方法检测的也是新冠肺炎患者 CT 扫描图像的浑浊区域，利用精确标注的 CT 图像训练分割模型，对病变进行细粒度分割。同时，利用类别标签训练的分类模型通过注意力融合的方式进一步提升分割模型的性能。本节将介绍 CT 新冠肺炎病灶分割方法的整体框架，包括网络结构、损失函数和与分类模型融合的策略。

5.2.1 分割模型构建

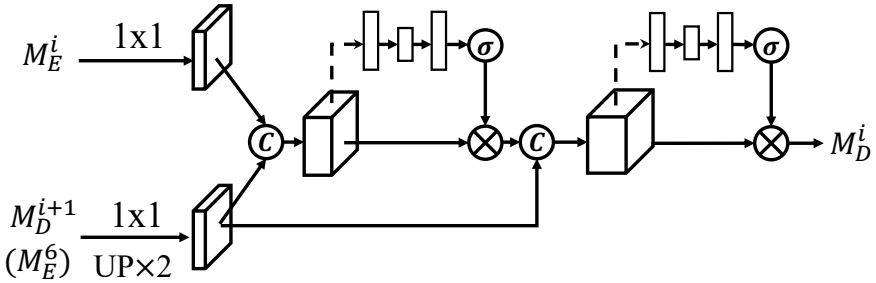
本文的分割模型旨在从新冠肺炎患者的 CT 图像中发现确切的病变区域。图 5.1显示了本文的结合了/未结合分割和分类模型分割分支的结构。这种组合的细节如图 5.3所示。

编码器-解码结构器

本文的分割模型由一个编码器和一个解码器组成。以下部分将详细介绍编码器和解码器的结构。



(a) 组化膨胀模块 (GAM)



(b) 注意特征融合 (AFF)

图 5.2 **本文为分割网络提出的（a）GAM 和（b）AFF**。在 AFF 中，若 $i=5$，M_D^{i+1} 将被 M_E^6 替代。立方体表示三维特征图，矩形表示特征向量。

编码器。编码器基于 VGG-16 [19] 骨干网络，去掉了最后 2 层全连接层。5 个 VGG 块分别定义为 $\{E_1, E_2, E_3, E_4, E_5\}$。VGG-16 主干首先输入 CT 图像，并从五个 VGG 块的最后一层生成多尺度特征图。为了将输入特征图缩小一半，每个块的前面（第一个块除外）是一个步长为 2 的 *max pooling* 函数。块 E_1 生成的特征图包含分辨率最高的最精确特征，而块 E_5 的特征图则最粗糙，分辨率最低。为了获取更好的性能，本文提出了增强特征模块（EFM）以提高本文编码器的表征能力。EFM 模块在块 E_5 的最后一层 *conv*5_3 后加入。其包含 2 个组膨胀卷积（GAM）以在更大的感受野上提取更强力的特征图。GAM 模块生成一个更小的特征图，与 VGG-16 主干的最粗糙特征图相比，大小为一半。它还增强了块 E_5 生成的特征图的表征能力。因此本文的编码器输出 6 个等级的特征图 $\{M_E^1, M_E^2, M_E^3, M_E^4, M_E^5, M_E^6\}$，其步长分别为 $\{1, 2, 4, 8, 16, 32\}$。因为本文采用了 U

形编解码器架构 [202]，所有这六个特征映射将在解码器中使用，稍后将介绍。

解码器。本文的解码器有 5 个不同大小的输出端。这里，本文不预测最小特征图的子输出，步长为 32，因此没有子输出与最小特征图 M_E^6 的大小匹配在本文的解码器中，本文提出了一种注意特征融合（AFF）算法，将不同阶段的特征图进行聚合，预测各阶段的子输出。本文的 AFF 强调了顶层特征图的重要性，并利用注意机制从底层特征图中过滤出有用的特征。最后一次输出与输入的 CT 图像具有相同的分辨率，作为最终的预测。

增强特征模块 EFM。本文提出的 EFM 模块在 VGG-16 编码器的最后一层 E_5 之后添加。它由两个连续的组化膨胀模块（Grouped Atrous Module，GAM）和它们之间的 *max pooling* 函数组成。如图 5.2 (a) 所示，GAM 模块的第一层是一个 11 卷积层，用于扩展特征图的通道。然后将特征图平均分为 4 组。与普通组卷积不同的是，本文将具有不同膨胀率的膨胀卷积 [203] 使用在 4 个组中，以获得具有不同感受野的更丰富的特征图。膨胀卷积可以极大地扩大卷积滤波器的感受野并与正常卷积保持相同的计算成本。在 2D 情况下，卷积核大小为 3×3 的膨胀卷积可以简单地表示如下：

$$q[i,j] = bias + \sum_{k=-1}^{+1}\sum_{l=-1}^{+1}(x[i+k\cdot n, j+l\cdot n]\cdot w[k+1,l+1]), \tag{5.1}$$

其中 n 表示膨胀率，w 是卷积权重，其大小为 3×3，q 和 x 分别是输出和输入特征图，i 和 j 是特征图位置。注意，$n=1$ 的特例即是普通卷积。为了充分利用有用的特性，本文在网络中采用了 Squeeze-Excitation（SE）块 [204]，也就是说，使用注意机制重新校准信道卷积特征响应。更具体地说，输入特征图的每个通道将乘以由 SE 块计算的通道权重。SE 块由两个线性层组成，后跟一个 sigmoid 函数。全局平均池后的输入特征映射将被输入到这个块中，本文可以为每个输入特征通道导出一个范围为 0,1 的通道权重。本文将 SE 块中的缩减率设置为 4，这意味着本文将第一个线性层的输出数设置为输入通道的 1/4。为了将输出通道减少一半，本文在 SE 块之后添加了一个 1×1 的卷积层。

最后，本文使用一个 3×3 卷积层，其中通道数等于输入特征图的通道数，作为到下一个模块的过渡层。

注意特征融合

自顶向下解码器的传统融合策略 [202, 205] 平等对待输入特征映射。为了更好地聚集特征图，本文提出了一种注意特征融合（Attentive Feature Fusion，AFF）

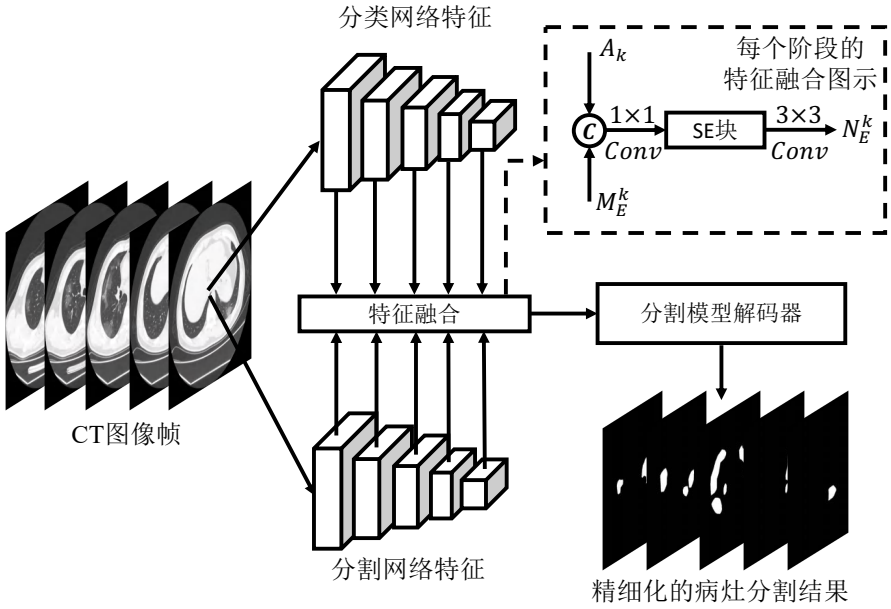


图 5.3 **分割和分类模型的结合。**本文将分割模型的编码器特征与分类模型的主干特征相结合。

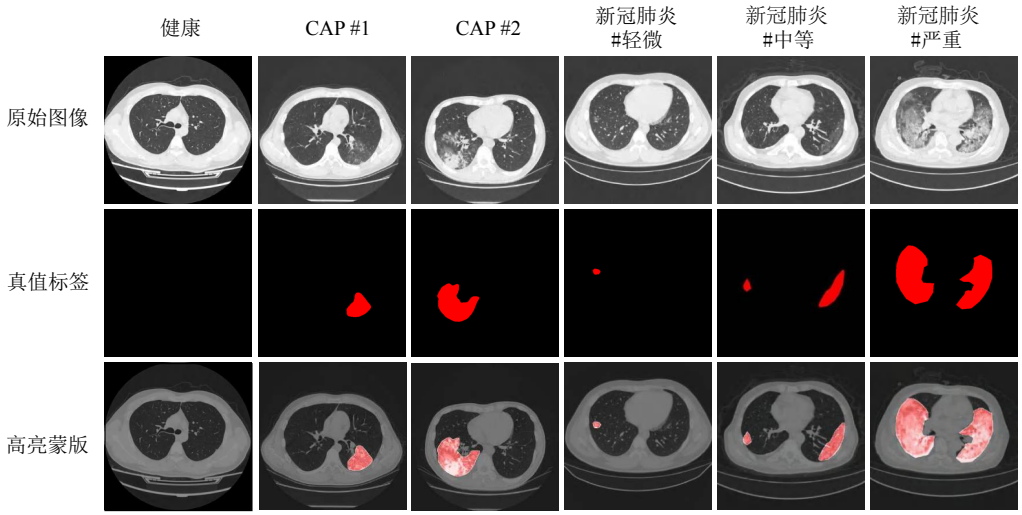


图 5.4 **本文的 COVID-CS 数据集示例，**包括 1 例正常人（第 1 列）、2 例社区获得性肺炎（CAP）病例（第 2 和 3 列）和 3 例新冠肺炎患者的 CT 影像扫描图像和标签（第 4 ∼ 6 列）。

策略。在本文的 AFF 融合策略中，较小尺寸的特征图更有价值。如图 5.2 (b) 所示，当前阶段中的输入特征图 M_E^i 和 M_D^{i+1} 通过 1×1 的卷积层缩小到一半大小。然后通过双线性插值对缩减后的 M_D^{i+1} 进行上采样，输出一个双倍大小的特征

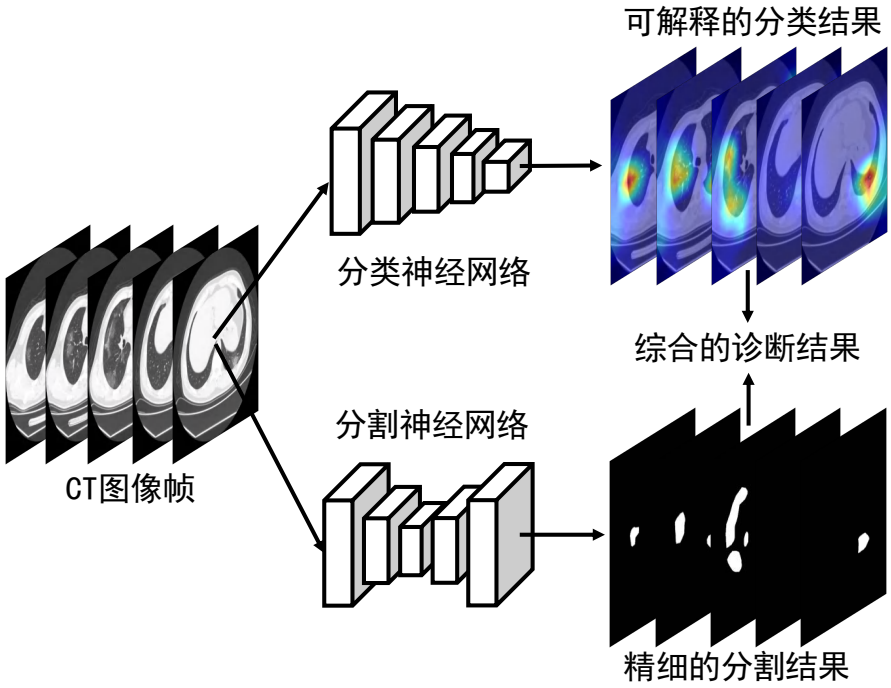


图 5.5 针对新冠肺炎 CT 辅助诊断而设计的联合分类分割（JCS）诊断系统的图示。JCS 系统将只在分类分支报告阳性新冠肺炎预测时执行分割诊断。

图。本文将两个输出连接起来，并应用 SE 块（也用于 GAM）来生成增强的特征图。后将该特征映射与前一阶段的双上采样输出的特征映射拼接起来。在拼接之后，本文使用另一个 SE 块来再次增强特征图。在每个 SE 块之后，本文使用一个 3×3 卷积层作为过渡层，该层具有与输入相同的通道数。一个神经元的一个 1×1 卷积层将被用来预测一个特征映射作为当前阶段的子输出。

深度监督损失

虽然最终的预测仅来自最后一个子输出，但本文将深度监督策略 [206] 应用于不同大小的所有子输出。对于每侧输出，本文将其采样到客观事实图像的大小，并计算标准二进制交叉熵损失和 Dice 损失 [207] 之和，如下所示：

$$L = BCE(\mathbf{P},\mathbf{G}) + 1 - \frac{\mathbf{P}\cdot\mathbf{G}}{\|\mathbf{P}\|_1 + \|\mathbf{G}\|_1}, \tag{5.2}$$

其中二分类交叉熵（BCE）取所有 $H\times W$ 个像素的平均值，$p_{i,j}$ 为通过 *sigmoid* 函数计算的在 (i,j) 位置像素的置信得分，“·” 代表点乘。$\mathbf{P}$ 和 $\mathbf{G}$ 为分别预测图和客观事实图，$\|\mathbf{P}\|_1$ 和 $\|\mathbf{G}\|_1$ 代表对应的 ℓ_1 范数。

5.2.2 二元感知

分类模型构建

通过分类模型，利用更多仅经类别标签标注的数据，可以进一步帮助分割模型学习更多的特征。所以，本文还构建了一个分类模型辅助分割模型。根据患者的 CT 图像，预测其是否为新冠肺炎阳性是一项二分类任务。由于设计新的分类模型不是本文的重点，因此本文构建了基于 Res2Net 网络 [208] 的分类器。Res2Net [208] 是 ResNet [20] 的增强版网络。它的最后一层被修改为具有两个通道（阳性与阴性）的完全连接层，从而输出新冠肺炎感染的概率。如果阳性通道的概率大于阴性通道的概率，则患者被诊断为新冠肺炎阳性，反之亦然。每一个病人的 CT 图像被逐个送入分类模型，如果感染的 CT 图像数量高于一个阈值，则诊断为新冠肺炎阳性。在训练分类模型时，本文使用图像混合 [209] 技术来缓解 CT 图像数据带来的偏差（bias），从而减轻分类模型对数据的过拟合。除此之外，本文使用类别激活映射技术 [210] 来确保分类模型的可解释性。

联合分类分割

如上文所述，本文设计了两个模型，一个用于新冠肺炎 CT 病灶分割，另一个用于新冠肺炎分类。分割网络依赖于精细化的专家标注数据，而分类网络仅需类别标签标注的数据即可训练。通过将它们结合在一起，可以利用更多的标注数据，从而获得更好的性能。受此启发，本文利用分类模型的特征来增强分割模型的特征。如题图 5.3所示，将分割模型的编码器和分类模型的主干点的各个阶段的特征映射合并在一起。分割模型编码器的特征图如 §5.2.1中，定义为 $M_E^1, M_E^2, M_E^3, M_E^4, M_E^5$。Res2Net [208] 分类模型主干有 5 个阶段，本文使用 $k \in [1,5]$ 阶段的最后一个特征图 A_k 进行特征结合。在合并阶段 k 的特征时，本文有两个用于合并的特征图 A_k, M_E^k。本文首先调整较小的 A_k 的大小，使其与较大的 M_E^k 大小相同，然后将它们拼接在一起。然后，本文应用一个简单的 11 卷积层来减少特征通道，使输出特征图的通道数与 M_E^k 相同。这样的 11 卷积层后面是一个还原率为 4 的 SE 块。最后，本文使用了一个输入和输出通道数相同的 33 卷积层作为过渡层。输出 N_E^k 将被视为增强的编码器特征，并被输入到分割模型的解码器中（图 5.1）。然后预测结果如 §5.2.1中所述。

表 5.1 **不同数据集的主要信息。**本文所搜集的 COVID-CS 数据包含了 750 个样本的超 10 万张 CT 图像。

数据集	日期	数据类型	# 图像数	# 样本人数
PLXR [211]	2020/03/23	X-rays	98	70
8023Dataset [193]	2020/03/25	X-rays	229*	-
CTSeg [86]	2020/03/28	CT	110	60
COVID-CT [194]	2020/03/30	CT	746*	-
COVID-CS (Ours)	2020/04/12	CT	**>144K**†	**750**

辅助诊断系统构建

由于医学专家资源紧缺，可以用来对分割模型训练的数据也非常有限。然而，与分割模型相比，本文的分类器是用新冠肺炎感染和未感染病例的 CT 图像进行训练的，以较低的标注代价获得更多的训练数据。尽管本文的分类器可以通过激活映射技术提供新冠肺炎的可解释的病变位置，但它不能进行准确和完整的病变分割。为此，本文的分割模型进一步提供了补充分析，可以发现肺部完整的病变并估计新冠肺炎患者的严重程度。但是由经验丰富的放射科医生注释大量的分割标签成本很高。为了更好地发挥它们的优势，本文通过联合解释分类和分割模型开发了新冠肺炎诊断系统，如图 5.5所示。在实践中，本文的分类模型将首先预测可疑病例的 CT 图像是否为新冠肺炎阳性。如果预测为阳性，疑似病例极有可能被新冠肺炎感染。本文的分割模型将在 CT 图像上进行深入分析，并在每个 CT 图像中发现整个混浊区域。

5.3 数据集构建

数据在基于深度学习的人工智能诊断系统中起着至关重要的作用。具有大规模样本或细粒度像素级标记的公开新冠肺炎 CT 数据集很少。为了填补这个空白，本文构建了一个新的新冠肺炎 CT 病灶分类和分割 (**COVID-CS**) 数据集。如表 5.1所示，本文的数据集与其他公开数据集进行了数据规模上的对比。在本节中，本文将详细介绍这个新数据集的数据收集、专业标注流程，以及相关的统计信息。图 5.4展示了本文 COVID-CS 数据集（普通样本和新冠肺炎样本）的一些样例，同时也展示了一些社区获得性肺炎患者的样例。图 5.6展示了从本文 COVID-CS 数据集的分割集中提供的多样信息。

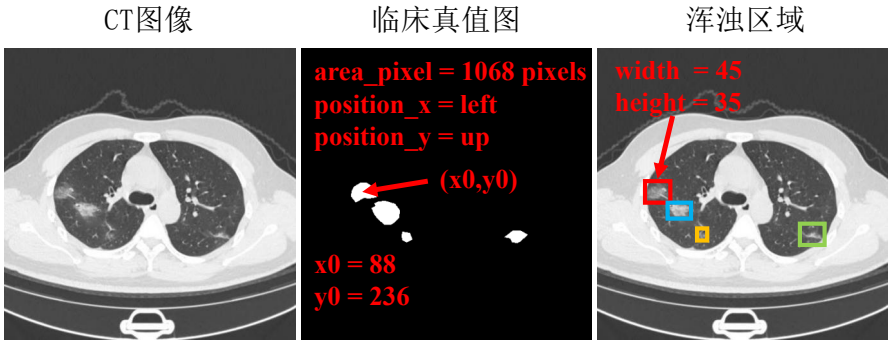


图 5.6 **COVID-CS 数据集提供的多样化标注信息。**图中为 COVID-CS 数据集中关于浑浊的区域（像素级）、位置（x0, y0）、方向（左，上）与宽度/高度的各种信息的说明。

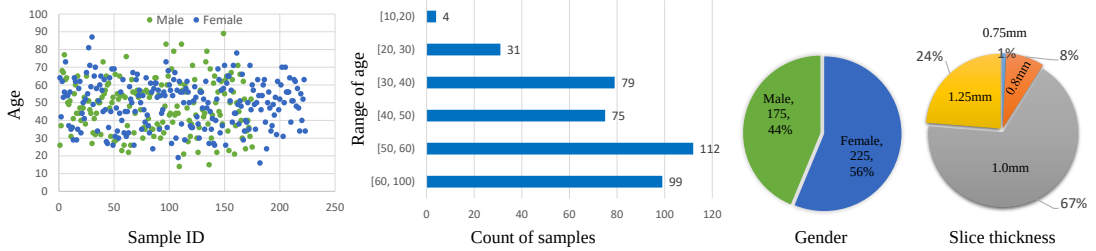
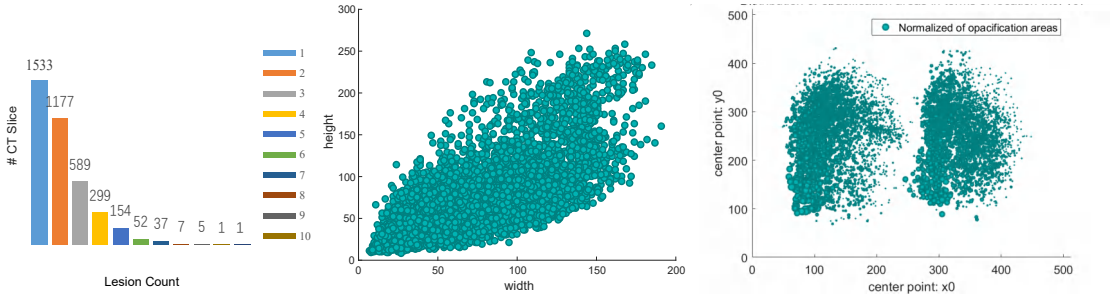


图 5.7 **年龄（左、中左）、性别（中右）、及 CT 切片厚度（右）的分布统计图。**可以通过放大来获取更多的细节。



(a) (b) (c)

图 5.8 **在本文的 COVID-CS 数据集中的分割集（200 个新冠肺炎病例）的统计数据。**(a) 病变计数分布；(b) 浑浊区域宽度和高度的分布；(c) 浑浊区域与位置的关系。

5.3.1 数据收集

为了保护患者的隐私，本文在构建数据集时省略了患者的个人信息。本文收集了 750 例患者的 144167 张 CT 影像扫描图像，其中 400 例为新冠肺炎阳性，

表 5.2 CT 扫描仪的制造商以及收集到的阳性病例人数。

CT 扫描仪制造商	产品型号	# 收集到的阳性样例人数
GE Medical Systems	Revolution CT	1
GE Medical Systems	LightSpeed VCT	6
GE Medical Systems	Discovery CT750 HD	12
GE Medical Systems	BrightSpeed	12
Toshiba	Aquilion ONE	33
GE Medical Systems	LightSpeed16	64
United Imaging Healthcare	uCT 780	272

表 5.3 分割模型中 EFM 和 AFF 的消融实验。基线是基于 VGG16 的分割模型，且没有 EFM 和 AFF（No. 1）。本文分别添加了 EFM 和 AFF，并展示了它们的有效性（No.2 和 No.3）。第四个结果是分割模型的完整版本。

No.	EFM	AFF	Dice	IoU	E_ϕ
1			71.0%	57.7%	88.0%
2	✔		74.3%	61.4%	88.9%
3		✔	75.9%	63.4%	90.9%
4	✔	✔	**77.5%**	**65.4%**	**92.0%**

其余 350 例为阴性，这些样例均经 RT-PCR 检测证实。参照以往研究 [212]，本文不考虑社区获得性肺炎（Community Acquired Pneumonia，CAP）患者（详情见图 5.4）。尽管 CAP 患者的 CT 图像也有相似的混浊，本文提出的诊断系统可能会诊断为新冠肺炎阳性，但 CAP 的威胁远小于新冠肺炎。本文的目的是快速开发一个自动诊断系统，并尽快诊断出疑似病例。此外，利用 CAP/新冠肺炎分类器 [212]、RT-PCR 测试与医师的经验，CAP 患者可简单诊断为新冠肺炎阴性。

所有患者均行标准胸部 CT 影像扫描。每个病例有 250∼400 个 CT 图像，每个病例的 CT 图像数量仅由 CT 影像扫描仪的类型及其扫描设置决定。CT 影像扫描仪包括 BrightSpeed、Discovery CT750 HD、LightSpeed VCT、LightSpeed16、GE Medical Systems 的 Revolution CT、东芝的 Aquilion ONE 和 United Imaging Healthcare 的 uCT 780。不同扫描器的病例数总结在表 5.2中。CT 重建片的厚度为 0.75∼1.25mm（详情见图 5.7）。

5.3.2 专家标注

COVID-CS 数据集为收集的 CT 影像扫描图像提供两个方面的标签，以便实现联合分类和分割任务。如上所述，本文的数据集分为 400 个新冠肺炎病例和 350 个未感染病例。对于细分任务，本文通过以下策略进行专家标注：

- 因为专家标注人力资源紧缺，放射科医生为每个病人选择最多 30 个离散的 CT 影像扫描图像，在这些图像中可以观察到感染并作进一步的标注。在这一步中，本文的目标是用像素级别的注释标记每个浑浊区域。
- 为了生成高质量的标注，本文首先邀请放射科医生，让其根据其临床经验标记尽可能多的混浊区域。然后，本文邀请另一位资深放射科医生对标记进行多次细化，以进行交叉验证。在此步骤之后，将修复一些不准确的标签。

通过实现上述标记过程，本文最终得到 200 位新冠肺炎患者的 3,855 张像素级标注的 CT 影像扫描图像，图像分辨率为 512×512。由于放射科医生数量有限，其他 200 例新冠肺炎患者的 64,771 张 CT 图像没有像素级标注，但这些数据将用于分类试验。从图 5.4的最后三栏可以看出，本文的数据集覆盖了轻微、普通、严重的新冠肺炎病例。

5.3.3 数据统计

数据数量。在本文数据集中，200 个阳性样本的 3,855 张图像被在像素级别进行了标注。64,771 张另外 200 个阳性样本的图像被在病人级别进行了标注，另有 75,541 张从 350 个阴性样本得到的图像。

年龄。 400 例新冠肺炎患者（男 175 例，女 225 例）年龄 14-89 岁，平均 48.9 岁。图 5.7显示了年龄分布、年龄范围内的样本计数和性别百分比。

病变计数。本文在图 5.8 (a) 中展示了病变计数的分布。本文观察到，病变计数分布在 1 至 10 在每个 CT 影像扫描图像。

混浊区域。本文在图 5.8 (b) 中绘制了混浊区域的宽度和高度。宽度和高度的范围分别为 7 ∼ 191 和 8 ∼ 271 个像素，具有多样化的分布。

位置。本文还在图 5.8 (c) 中显示了每个混浊区域和相应的中心位置（x0, y0）之间的关系。可以看出，归一化的混浊区域的范围从最小尺寸（35/28452 像素）到最大尺寸（28452/28452 像素）。它还显示，在本文的 COVID-CS 数据集中，混浊区域均匀分布在不同的位置，这些区域也均匀分布在肺部。

表 5.4 分割模型与分类模型相结合的消融研究。

序号	分割模型	+ 分类特征	Dice	IoU	E_ϕ
1	✔		77.5%	65.4%	92.0%
2	✔	✔	**78.5%**	**66.4%**	**92.7%**

5.4 实验

5.4.1 实验设置

训练/测试设置。对于分割任务，本文的训练集包含来自 150 名新冠肺炎患者的 2794 张图像，测试集包含来自其他 50 名新冠肺炎患者的 1061 张图像。对于分类任务，训练集包含分割集中 150 个新冠肺炎感染病例的 2794 张图像。另外，本文随机选取 150 例未感染病例，7500 张 CT 图像作为阴性病例进行训练。本文也对分类模型进行测试，其分类精度可以作为分类模型特征表达能力的验证。分类测试集包含随机选取的 200 例感染病例的 64711 张图像和 200 例未感染病例的 68041 张图像。

评估指标。本文使用两个标准度量，即，Dice 分数 [213] 与 IoU。为了提供更全面的评估，本文进一步使用了用于评价前景图的广泛使用的度量——增强对齐度量 E-measure(E_ϕ) [161]。

方法对比。为了深入评估本文的 JCS 模型，本文将其与多种主流模型进行比较，即，如用于医学图像的 U-Net [202]，以及用于显著性目标检测的 DSS [11]，PoolNet [6]，和 EGNet [12]。

实现细节。在本文的 JCS 中，分类模型和分割模型分别训练。对于分类模型，本文在 4 个 GPU 上以 256 的批量大小对其进行训练。为了提高计算效率，将 CT 图像的大小调整为 224×224。本文采用 SGD 优化器，初始学习率为 0.1，每 30 个周期除以 10。分类器训练 100 个周期。本文采用了随机水平翻转和随机裁剪技术作为分类模型的数据增强方式，并采用了图像混合技术 [209] 来消除数据的偏差。图像混合的 Beta 分布中的 α 设置为 0.5。每个小批量中的 CT 图像数始终为 4，并且输入 CT 图像的大小不变，为 512×512。本文的分割模型的主干是在 ImageNet 上预先训练的 [214]。两个序列 GAM 中四个空洞卷积的膨胀率分别为 $\{1,3,6,9\}$ 和 $\{1,2,3,4\}$。初始学习率为 2.5×10^{-5}。本文采用 *poly* 学习率策略，即实际的网络学习率将乘以系数 $(1-\frac{cur\_iter}{max\_iter})^{power}$，其中 *power* 为 0.9。分割模型进行了 21000 次迭代训练。本文使用 Adam [180] 优化器，将 β_1, β_2 分别设置为 0.9 和 0.999。本文使用随机水平翻转和随机裁剪作为分割模型的数据增强

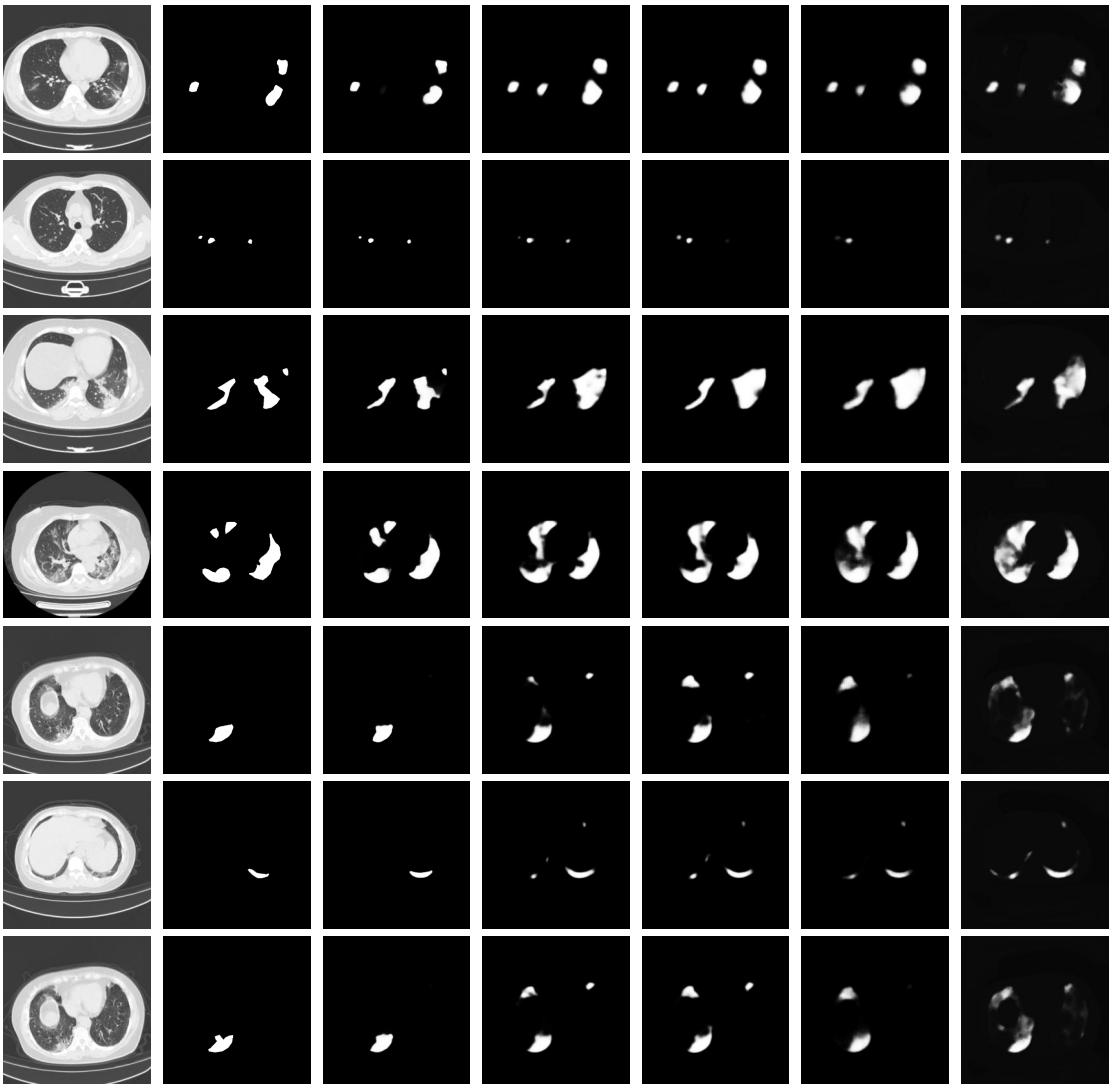


图 5.9 **不同方法在分割测试集上的定性比较。**本文对比了多种程度感染下的不同方法的分割结果。

方法。

5.4.2 结果

分割分支中 EFM 和 AFF 消融实验。在 JCS 模型内，本文引入了两个新的分割模块 EFM 和 AFF。EFM 的目的是提高编码器在分割分支中的表示能力。在特征融合阶段，在 AFF 算法中，较小尺寸的特征图在计算中占据更大的考虑比重，而传统的融合策略对输入的特征图“平等对待”。EFM 和 AFF 的消融研究如

表 5.5 分割测试集的定量结果。

方法	出版物	Dice	IoU	E_ϕ
U-Net [202]	MICCAI'15	65.1%	54.1%	79.7%
DSS [11]	TPAMI'19	65.7%	51.7%	79.9%
EGNet [12]	ICCV'19	69.3%	55.4%	83.6%
PoolNet [6]	CVPR'19	69.7%	55.9%	83.9%
JCS (本文)	TIP'21	**78.5%**	**66.4%**	**92.7%**

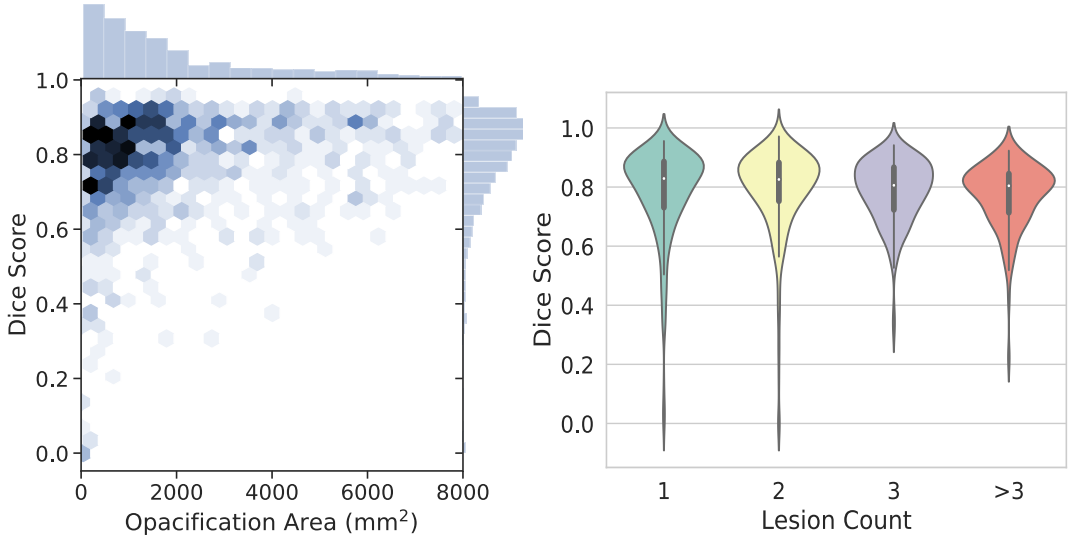


图 5.10 在分割测试集上对本文的分割模型进行统计分析。（a）每个 CT 图像的混浊区域大小与相应 Dice 评分之间的关系。（b）病灶个数与 Dice 得分概率分布的关系。

表 5.3所示。第一个结果是没有 EFM 和 AFM 的基准性能。在将所提出的 EFM 和 AFF 分别应用于基准之后，性能在 Dice 度量方面分别提高了 3.3% 和 4.9%。因此，EFM 和 AFF 对分割分支都有很大的帮助。当结合 EFM 和 AFF 时，本文在 Dice 度量方面获得了 6.5% 的提升。在 IoU 和 E-measure [161] 度量方面的改进与 Dice 度量类似。因此，本文提出的 EFM 和 AFF 对分割模型具有重要促进作用。

分割模型与分类模型相结合的消融实验。如 §5.2.2中所述，将分类模型与分割模型相结合，得到更丰富的特征。为了验证这样的选择，本文运行了如

表 5.6 **不同方法诊断新冠肺炎的平均时间。**“†” 代表诊断未感染病例的耗时。

方法	RT-PCR	CT R.	CT R. + JCS	JCS
用时	∼4h [215]	21.5min [192]	1s†/54.4s	1s†/22.0s

表 5.4中所示的实验。基准是单一分割模型（No.1, 表 5.4）。但本文也观察到分类模型和分割模型的组合选择（No.2, 表 5.4）在 Dice 度量方面有 1.0% 的改进，并且表明分类模型的特征可以帮助分割模型预测更好的结果。

分割性能比较。表 5.5 列出了 4 种最新的分割方法和本文的分割模型的定量比较。可以看出，该模型在所有三个指标上都取得了最优的结果。与第二名的 PoolNet [6] 相比，它在 Dice 得分、IoU 和 $E\phi$ 上分别提高了 8.8%、10.5% 和 8.8%。另外, PoolNet [6] 和 EGNet [12] 在 3 个指标上获得了相似的结果。U-Net 在 IoU 上比 DSS [11] 的表现好，尽管它们在 Dice 得分上相当。图 5.9显示了可视化的定性比较结果。可以看出，其他主流方法在轻度、中度和重度新冠肺炎感染的 CT 图像中对病变区域的预测不准确甚至出现错误。但是本文的分割模型准确地发现了所有新冠肺炎感染水平上的整个病变区域。

统计分析。为了进一步研究它的稳定性，本文在分割测试集上对分割模型进行了统计分析。图 5.10 (a) 显示了本模型的 Dice 评分与 CT 图像混浊区域的相关性。需要指出的是，由于 $\geq 8000mm^2$ 不透明区域的 CT 图像只占所有 CT 图像数量的 1.0%, 它们未绘制在图 5.10 (a) 中。本文观察到 95.9% 的 CT 图像的 Dice 评分为 $[0.6, 1]$，而其他 3.3% 的 CT 图像的 Dice 评分在 $[0.1, 0.6)$ 之间，被认为是不良病例。CT 图像中只有 0.8% 的图像 Dice 得分低于 0.1，以失败病例为例。本文还从不同的角度探讨了每个切片的病变计数与 Dice 评分之间的关系。如图 5.10 (b) 所示，CT 图像中的病灶计数对 Dice 评分的概率分布影响不大。对于 4 种不同的病变计数，中位 Dice 得分高于 0.8，95.0% 置信区间为 $[0.5, 1]$。本文还观察到失败病例的病变计数 ≤ 2。在分割病变区域方面的持续良好性能和低失败概率（0.8%）证实了本文的分割模型的稳定性。

诊断用时。本文对基于 JCS 的诊断系统在单张 RTX 2080Ti 上进行速度测试，表 5.6显示了不同情况下的诊断速度。该表中，“CT R.” 代表 CT 专家单独诊断，“CT R. + JCS ” 表示本文的 JCS 系统与 CT 专家共同诊断。假设每个疑似病例都有 300 张 CT 图像，本文 JCS 系统的分类模型只需 1.0 秒左右的时间就可以确定是否被感染。如果被感染，分割模型将花费 21.0 秒进行细粒度病变分割。因此，

在基于 JCS 的诊断系统中，每个感染病例的系统成本为 22.0 秒，每个未感染病例的系统成本为 1.0 秒。值得注意的是，无论病例是否感染，完整的 RT-PCR 检测和放射科医生的 CT 诊断分别需要 4 小时和 21.5 分钟。

5.5 本章总结

为了训练用于新冠肺炎诊断的卷积神经网络模型，本文系统地构建了一个大规模的新冠肺炎分类与分割 (COVID-CS) 数据集。本文还开发了一个精细化的新冠肺炎病灶分割方法。为了提供精细化的像素级预测，本文实现了一个基于注意力融合的图像二元感知模型 JCS 来发现新冠肺炎患者 CT 图像中的细粒度病变区域，同时还引入了分类模型的特征从而充分利用了多样的特征表达。与其他主流方法相比，如 PoolNet [6]，本文的分割模型获得了 8.8% 的 Dice 性能提升。不仅如此，本文的 JCS 模型结果还十分稳定。从本文的统计分析上看，在 COVID-CS 分割测试集上，JCS 仅在 0.8% 的 CT 图像上失效，在 95.9% 的 CT 图像中获得了 $[0.6, 1]$ 的 Dice 得分。在 JCS 的帮助下，医生的诊断用时缩短了 20 倍，仅需花费不到 1 分钟就可以完成对一个感染病例的诊断，还保持了与纯人工诊断下相同的准确率。

第 6 章 基于金字塔池化的骨干特征提取

目标检测与分割通常基于已有的骨干网络，它们的性能也依赖于已有的骨干网络的多尺度建模能力。按照第一章的分析，目前的骨干网络主要存在多尺度建模能力不足的问题，主要原因来自于它们的多尺度特征提取能力有限，且只能在局部提取特征。本文聚焦该问题，提出了基于金字塔池化的 Transformer（Pyramid Pooling Transformer，P2T）。它首先将 Transformer 引入到了骨干网络中，用于提取全局特征。然而，Transformer 对输入特征长度有着平方关系的计算复杂度，这一特性使得 Transformer 难以直接运用到特征尺度很大的目标检测与分割中。为了解决这一问题，本文提出将金字塔池化引入到了 Transformer 中的多头自我注意力中，降低 Transformer 的计算复杂度的同时还能提取多尺度的特征，增强了上下文表征能力。基于这一改进，本文将 P2T 在几大目标检测与分割的主流任务，如语义分割、物体检测、实例分割等任务上进行了实验验证，发现 P2T 在各类任务上都取得了最佳的性能。此外，本文还将 P2T 引入到第三、四、五章所涉及的算法中，实验发现 P2T 显著提升了以上三章所提出的算法性能。在本章内，本文先在第一节介绍相关的研究背景与动机，再在第二节介绍 P2T 的设计与实现，在第三节介绍实验结果与分析，最后在第四节进行本章总结。

6.1 引言

近十多年来，卷积神经网络（Convolutional Neural Networks，CNNs）已经统治了机器视觉的各个领域并且取得了非常优秀的进展 [19–21, 89, 93, 96, 127, 169]。卷积神经网络在许多大型数据集上的视觉任务上都取得了最先进的结果。在另一个平行的领域里面，如自然语言处理（Natural Language Processing, NLP）上，一种流行的技术是 Transformer，它全部依赖于其自注意力机制来获取大范围的全局相关性，取得了很好的成功。因为全局信息对于视觉任务来说非常重要，Transformer 也可以被用来解决卷积神经网络的局限性，即卷积神经网络通常只能利用不断堆叠的方式来增大它的感受野。许多研究者在如何将 Transformer 用于视觉任务的问题上做了很多的努力 [99]。早期的研究者尝试用 Transformer 来

进一步处理由卷积神经网络提取到的深层次特征 [19, 20]，从而能够拟合目标问题 [100, 101, 216]。Dosovitskiy 等人 [105] 在单纯利用 Transformer 来解决图像分类这个问题上取得了巨大的突破。它们将一张图片给分割成多个小块，并且将每个小块图片当作自然语言处理中的单词（Word）或词符（Token），因此 Transformer 能够直接地用于对图像的处理。这种简单的方法在 ImageNet [5] 数据集上得到了非常具有竞争力的结果。因此，计算机视觉领域出现了一个新的概念，即视觉 Transformer。在视觉 Transformer 概念出现的短时间内，涌现了大量改进视觉 Transformer 的研究工作 [105]，同时也带来了比卷积神经网络更加优秀的性能 [111–114, 121, 217]。

尽管如此，视觉 Transformer 仍然存在一个具有挑战性的问题，即数据序列的长度。当将图像块视为自然语言处理中的词符时，编码得到的序列长度仍然比自然语言处理中要长得多。例如，在自然语言处理中，著名的 WMT2014 英德数据集 [218] 包含了 5000 万个英语单词和 200 万个句子，其平均序列长度为 25。相比之下，在计算机视觉领域，本文通常使用 224×224 图像分辨率的 ImageNet 数据集 [5] 来进行图像分类任务。如果本文使用普遍采用的 4×4 图像块，那么编码的得到的序列长度将为 3136。由于 Transformer 中的多头自注意力模块（Multi-Head Self-Attention, MHSA）的计算复杂度是图像大小的二次方（而不是卷积神经网络中的线性），直接将 Transformer 应用于视觉任务对计算资源的要求很高。为了实现一个可用于图像分类的纯 Transformer 网络，ViT [105] 使用较大的图像块来减少序列长度，例如 16×16 或者 32×32，并在图像分类上取得了巨大的成功。后来，许多 Transformer 工作 [109, 111–114, 116, 121, 219] 通过引入金字塔结构来提高 ViT [105] 的性能。其中，输入层首先使用 4×4 小图像块和然后通过合并相邻图像块来逐渐减小序列的长度。

为了进一步降低多头自注意力模块的计算开支，PVT [113] 和 MViT [114] 在多头自注意力模块的计算中使用了单个池化操作对特征图进行下采样。通过使用池化后的特征，它们模拟的是词符到区域（Token-to-Region）的关系，而不是经常使用的词符对词符（Token-to-Token）的关系。Swin Transformer [112] 提出在建模相互关系的时候，选择只计算小窗口内的多头自注意力模块而不是在整个输入窗口上进行计算。跟卷积神经网络类似，它通过使用移动窗口（Window Shift）的策略以及堆叠更多的层数来逐渐扩大网络的感受野 [115]。尽管如此，视觉 Transformer 的一个基本特征是它的直接全局关系建模，这也是众多研究者

从卷积神经网络转移到 Transformer 的原因。

在这篇文章里面，本文考虑如何改进 PVT [113] 和 MViT [114]，因为它们使用单层池化操作提取到的池化特征似乎都不太强大。如果本文能够在压缩特征序列长度的同时获取更强的特征表征，那么本文可能会在各类任务下取得更好的性能。为此，本文注意到金字塔池化 [7, 220–222] 是一种具有较长历史的计算机视觉技术，它通过提取上下文信息并利用具有不同感受域的多层池化操作在输入特征上进行多尺度运算。这种简单的技术已在各种下游目标检测与分割任务中，例如语义分割 [7] 和物体检测 [222]，被证明是有效的。然而，最近的金字塔池化方法高度依赖于预训练的卷积神经网络骨干，因此它们仅限于一些特定的视觉任务。换言之，金字塔池化技术在具有广泛应用的骨干网络设计中尚未被探索。为了弥补这一差距，本文将金字塔池化应用于视觉 Transformer 模块中，从而减少序列长度并且同时学习到强大的上下文表征。金字塔池化操作的计算效率也非常高，它给视觉 Transformer 带来的计算开支几乎可以忽略不计。

本文通过提出一个新的 Transformer 骨干网络来实现这一目标，即**金字塔池化 Transformer**（Pyramid Pooling Transformer，P2T）。本文将金字塔池化的想法应用于视觉 Transformer 的多头自注意力模块中，不但减少了该模块的计算开支，同时也获取到了丰富的上下文信息。通过将基于新池化的多头自注意力模块用于 Transformer 中，P2T 在特征学习和视觉识别方面的表现都比其他基于单层池化操作的 PVT [113] 和 MViT [114] 更加强大。本文利用各种典型的视觉任务，例如图像分类、语义分割、物体检测和实例分割等，来评估 P2T 的性能。本文还在将 P2T 骨干网络应用到第三、四、五章所涉及到的算法中，并评估了 P2T 所带来的性能提升幅度。大量实验表明，对于这些基本视觉任务，P2T 的性能均优于所有以前基于卷积神经网络和 Transformer 的骨干网络。

6.2 方法

在该小节中，本文首先在 §6.2.1介绍整个 P2T 网络结构的大体思路，在 §6.2.2 展示基于池化多头自注意力模块的 P2T 的网络架构，最后在 §6.2.3 介绍一些网络的实现细节。

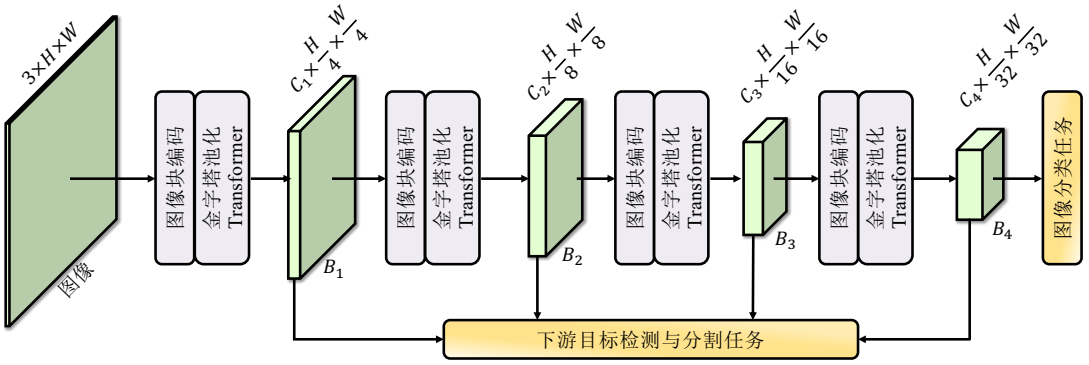


图 6.1 **本文提出的 P2T 网络架构。**特征 $\{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \mathbf{B}_4\}$ 可以用于下游的目标检测与分割任务。

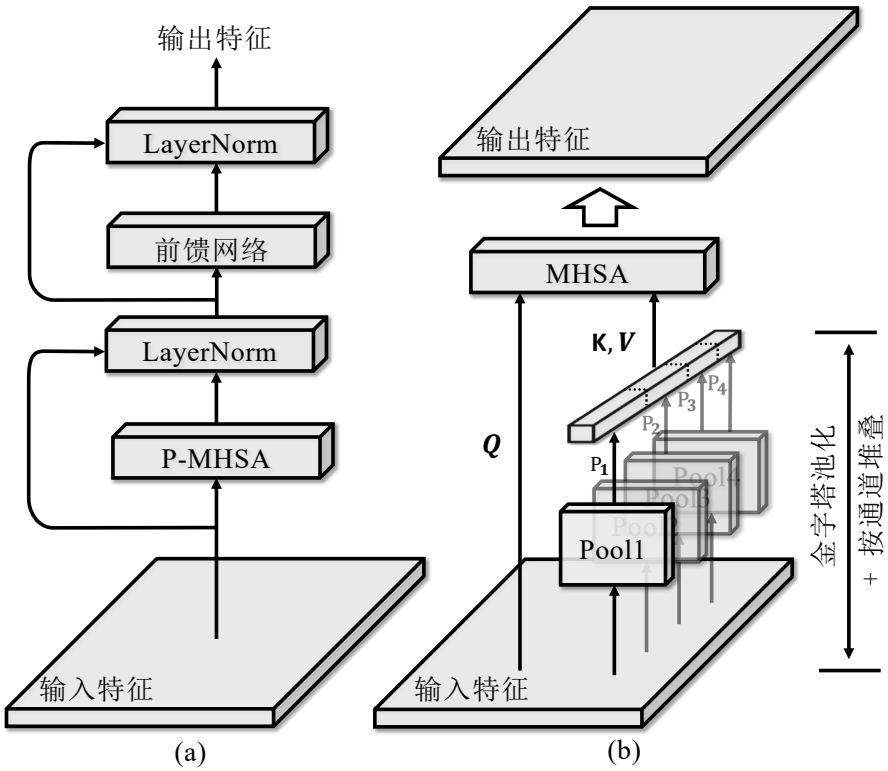


图 6.2 **金字塔池化 Transformer 的基本结构。** (a) 金字塔池化 Transformer 的简单结构示意图。(b) 基于池化的多头自注意力模块的详细结构。

6.2.1 概述

P2T 的整体架构如图 6.1所示。以自然彩色图像作为输入，P2T 首先将其拆分为 $\frac{H}{4}\times\frac{W}{4}$ 个块，每个块都被展平为 48（$4\times4\times3$）个元素。与先前的工作 [113]

一致，P2T 将这些扁平化后的图像块输入到一个图像块编码模块中；它由一个线性投影层组成；然后加上可学习的位置编码。图像块编码模块将 48 个元素的特征维度扩展到 C_1。然后，本文堆叠了多个金字塔池化 Transformer 模块（§6.2.2）。整个网络可以分为四个阶段，它们的特征维度分别为 C_i $(i=\{1,2,3,4\})$。在每两个阶段之间，每个 2×2 图像块组被拼接起来，并从 $4\times C_i$ 线性投影到 C_{i+1} 维度 $(i=\{1,2,3\})$。这样，四个阶段的尺度就变成了 $\frac{H}{4}\times\frac{W}{4}$，$\frac{H}{8}\times\frac{W}{8}$，$\frac{H}{16}\times\frac{W}{16}$，和 $\frac{H}{32}\times\frac{W}{32}$。从四个阶段，本文可以分别推导出四个特征表示即 $\{\mathbf{B}_1,\mathbf{B}_2,\mathbf{B}_3,\mathbf{B}_4\}$。其中只有 $\mathbf{B}_4$ 将用于图像分类的最终预测，但是所有金字塔特征都可以用于下游目标检测与分割任务。

6.2.2 基础结构

金字塔池化已被广泛用于许多与卷积神经网络协作的目标检测与分割任务中 [7, 34, 38, 48, 222–227]。然而，现有文献通常基于已设计好的骨干网络，并在它们基础之上利用金字塔池化额外设计一些模块，用于提取特定任务的全局上下文信息。相比之下，本文首次探索了 Transformer 和骨干网络之中的金字塔池化，目的是为了普遍改善各种目标检测与分割任务。为此，本文将金字塔池化操作与 Transformer 结合，在减少多头自注意力模块的计算负荷的同时捕获丰富的上下文信息。

P2T 的基本单元结构如图 6.2（a）所示。输入特征首先通过基于金字塔池化的多头自注意力模块，其输出与短接的自身相加，然后通过 LayerNorm [228]。如同传统的 Transformer 模块 [105, 106, 113]，特征在接下来被输入到前馈网络（FFN）中，并应用残差连接和 LayerNorm。上述过程可以用下式表述：

$$\begin{aligned}\mathbf{X}_{att} &= \text{LayerNorm}(\mathbf{X}+\text{P-MHSA}(\mathbf{X})),\\ \mathbf{X}_{out} &= \text{LayerNorm}(\mathbf{X}_{att}+\text{FFN}(\mathbf{X}_{att})),\end{aligned} \tag{6.1}$$

其中 $\mathbf{X}$，$\mathbf{X}_{att}$ 和 $\mathbf{X}_{out}$ 分别表示多头自注意力模块的输入、输出和 Transformer 块的输出。**P-型多头自注意力模块**（P-MHSA）是基于池化的多头自注意力模块的代称。

基于池化的多头自注意力模块（P-MHSA）

在这里，本文将介绍基于池化的多头自注意力模块的设计。其结构如图 6.2（b）所示。首先，$\mathbf{X}$ 被重塑为二维特征。然后，本文在重塑的 $\mathbf{X}$ 上应用不同比

例的多个平均池化层，用以生成金字塔特征图，具体如下式所示：

$$\begin{aligned}\mathbf{P}_1 &= \text{AvgPool}_1(\mathbf{X}),\\ \mathbf{P}_2 &= \text{AvgPool}_2(\mathbf{X}),\\ &\cdots,\\ \mathbf{P}_n &= \text{AvgPool}_n(\mathbf{X}),\end{aligned} \tag{6.2}$$

其中 $\mathbf{P}_1, \mathbf{P}_2, ..., \mathbf{P}_n$ 表示生成的金字塔特征图，n 是池化层的数量。接下来，本文将金字塔特征图送入深度卷积进行相对位置编码：

$$\mathbf{P}_i^{enc} = \text{DWConv}(\mathbf{P}_i) + \mathbf{P}_i, \quad i = 1, 2, \cdots, n, \tag{6.3}$$

其中 DWConv$(\cdot)$ 表示深度卷积，核大小为 3×3，$\mathbf{P}_i^{enc}$ 为相对位置编码的 $\mathbf{P}_i$。由于 $\mathbf{P}_i$ 是池化特征，所以在 式 (6.3) 中的操作只需要非常少的计算开支。之后，本文对这些金字塔特征图进行扁平化和拼接操作:

$$\mathbf{P} = \text{LayerNorm}(\text{Concat}(\mathbf{P}_1^{enc}, \mathbf{P}_2^{enc}, ..., \mathbf{P}_n^{enc})). \tag{6.4}$$

公式中为了简单起见，省略了扁平化操作。这样，如果池化率足够大，$\mathbf{P}$ 可以是一个比输入 $\mathbf{X}$ 短得多的特征序列。此外，$\mathbf{P}$ 包含了输入 $\mathbf{X}$ 的上下文抽象特征，因此在计算多头自注意力模块时可以作为输入 $\mathbf{X}$ 的有力替代。

假设多头自注意力模块 [105] 中的查询、键和值张量分别为 $\mathbf{Q}$、$\mathbf{K}$ 和 $\mathbf{V}$。与其采用传统的如下方法：

$$(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\mathbf{X}\mathbf{W}^q, \mathbf{X}\mathbf{W}^k, \mathbf{X}\mathbf{W}^v), \tag{6.5}$$

本文提出采用

$$(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = (\mathbf{X}\mathbf{W}^q, \mathbf{P}\mathbf{W}^k, \mathbf{P}\mathbf{W}^v), \tag{6.6}$$

其中 $\mathbf{W}^q$, $\mathbf{W}^k$ 和 $\mathbf{W}^v$ 表示生成查询、键和值的线性变换的权重矩阵。

然后，$\mathbf{Q}, \mathbf{K}, \mathbf{V}$ 被送入注意力模块，以计算注意力特征 $\mathbf{A}$。它可以被表述为如下形式：

$$\mathbf{A} = \text{Softmax}(\frac{\mathbf{Q} \times \mathbf{K}^\mathbf{T}}{\sqrt{d_K}}) \times \mathbf{V}, \tag{6.7}$$

其中 d_K 是 $\mathbf{K}$ 的通道尺寸。和 $\sqrt{d_K}$ 可以作为一个近似的标准化。Softmax 函数是沿着矩阵的行来应用的。为了简介，式 (6.7) 省略了多头的概念 [99, 105]。

由于 **K** 和 **V** 的长度比 **X** 小，本文所提出的 P-型多头自注意力模块比传统的多头自注意力模块更有效率。此外，由于 **K** 和 **V** 包含高度抽象的多尺度信息，所提出的 P-型多头自注意力模块在全局上下文依赖性建模方面有更强的能力，这对目标检测与分割很有帮助 [7,38,222,223,225–227]。从不同的角度来看，金字塔池化通常被用作连接在已有骨干网络基础之上的有效技术；相比之下，本文通过 Transformer 在骨干网络内利用金字塔池化，从而为目标检测与分割提供强大的特征表示学习。通过上述分析，P-型多头自注意力模块有望比传统的多头自注意力模块更高效、更有效 [99,105]。

计算复杂度分析。正如在式 (6.2) 中所描述的，所提出的基于金字塔池化的注意力利用多个池化操作来生成金字塔特征图。金字塔池化操作只有可忽略的 $O(NC)$ 计算复杂度，其中 N 和 C 分别代表序列长度和特征尺寸。因此，计算多头自我注意力的计算复杂度可以表述为：

$$O(\text{P-MHSA}) = (N+2M)C^2 + 2NMC \tag{6.8}$$

其中 M 是所有池化特征的总和序列长度。对于默认的金字塔池化比例 {12, 16, 20, 24}，本文有 $M \approx \frac{N}{66.3} \approx \frac{N}{8^2}$，这与 PVT 中多头自注意力模块的计算开支相当 [113]。

前馈网络。它是 Transformer 中用于特征增强的一个重要组成部分 [99,229]。以前的 Transformer 通常应用全连接网络作为前馈网络 [99,105,113]，完全依靠注意力来捕捉像素间的依赖关系。虽然有效，但这种架构并不擅长学习二维近邻关系，而二维近邻关系在目标检测与分割中起着关键作用。为此，本文按照 [119,120] 将深度卷积插入到前馈网络中，这样得到的 Transformer 可以同时继承 Transformer（长距离依赖性建模）和卷积神经网络（二维近邻关系）的优点。具体而言，本文采用倒置残差模块（Inverted Bottleneck Block，IRB）作为前馈网络。

为了使倒置残差模块适用于视觉 Transformer，本文首先将输入序列 $\mathbf{X}_{att}$ 变换为二维特征图 $\mathbf{X}^I_{att}$：

$$\mathbf{X}^I_{att} = \text{Seq2Image}(\mathbf{X}_{att}), \tag{6.9}$$

其中 Seq2Image$(\cdot)$ 是将一维序列重塑为二维特征图的操作。考虑到输入的 $\mathbf{X}^I_{att}$，

表 6.1 **本文 P2T 的具体配置。**

阶段序号	输入大小	运算操作	P2T-Tiny	P2T-Small	P2T-Base	P2T-Large
1	224×224	7×7 conv.	$C=48, S=4$	$C=64, S=4$		
2	56×56	P-MHSA IRB	$\begin{bmatrix}C=48\\E=8\end{bmatrix}\times2$	$\begin{bmatrix}C=64\\E=8\end{bmatrix}\times2$	$\begin{bmatrix}C=64\\E=8\end{bmatrix}\times3$	$\begin{bmatrix}C=64\\E=8\end{bmatrix}\times3$
3	28×28	P-MHSA IRB	$\begin{bmatrix}C=96\\E=8\end{bmatrix}\times2$	$\begin{bmatrix}C=128\\E=8\end{bmatrix}\times2$	$\begin{bmatrix}C=128\\E=8\end{bmatrix}\times4$	$\begin{bmatrix}C=128\\E=8\end{bmatrix}\times8$
4	14×14	P-MHSA IRB	$\begin{bmatrix}C=240\\E=4\end{bmatrix}\times6$	$\begin{bmatrix}C=320\\E=4\end{bmatrix}\times9$	$\begin{bmatrix}C=320\\E=4\end{bmatrix}\times18$	$\begin{bmatrix}C=320\\E=4\end{bmatrix}\times27$
5	7×7	P-MHSA IRB	$\begin{bmatrix}C=384\\E=4\end{bmatrix}\times3$	$\begin{bmatrix}C=512\\E=4\end{bmatrix}\times3$	$\begin{bmatrix}C=512\\E=4\end{bmatrix}\times3$	$\begin{bmatrix}C=640\\E=4\end{bmatrix}\times3$
	1×1	-	全局平均池化，维度为 1000 的全连接层，Softmax 层			
参数量			11.6M	24.1M	36.1M	54.5M
计算开支（FLOPS）			1.8G	3.7G	6.5G	9.8G

IRB 可以直接应用，比如说：

$$\begin{aligned}\mathbf{X}^{1}_{\text{IRB}} &= \text{Act}(\mathbf{X}^{I}_{att}\mathbf{W}^{1}_{\text{IRB}})\\ \mathbf{X}^{\text{out}}_{\text{IRB}} &= \text{Act}(\text{DWConv}(\mathbf{X}^{1}_{\text{IRB}}))\mathbf{W}^{2}_{\text{IRB}},\end{aligned} \tag{6.10}$$

其中 $\mathbf{W}^{1}_{\text{IRB}}, \mathbf{W}^{2}_{\text{IRB}}$ 表示 1×1 卷积的权重矩阵，Act 表示非线性激活函数，$\mathbf{X}^{\text{out}}_{\text{IRB}}$ 是 IRB 的输出。由于 $\mathbf{X}^{\text{out}}_{\text{IRB}}$ 是一个二维特征图，本文最后将其转化为一维序列：

$$\mathbf{X}^{S}_{\text{IRB}} = \text{Image2Seq}(\mathbf{X}^{\text{out}}_{\text{IRB}}), \tag{6.11}$$

其中 Image2Seq$(\cdot)$ 是将二维特征图重塑为一维序列的操作，$\mathbf{X}^{S}_{\text{IRB}}$ 是前馈网络的输出，其形状与 $\mathbf{X}_{att}$ 相同。

6.2.3 实现细节

不同深度的 P2T 设置。遵循之前的骨干架构 [20, 94, 112, 113, 121]，本文通过在每个阶段堆叠不同数量的金字塔池 Transformer 来构建不同深度的 P2T。通过这种方式，本文提出了四个版本的 P2T，即 P2T-Tiny、P2T-Small、P2T-Base 和 P2T-Large，其参数量分别与 ResNet-18 [20]、ResNet-50 [20]、ResNet-101 [20] 和 PVT-Large [113] 相当。除了 P2T-Tiny 中每个注意力头有 48 个特征通道，P-型多头自注意力模块的每个头有 64 个特征通道。不同版本的 P2T 的其他配置显示在表 6.1。对于第一阶段，本文应用一个 7×7 卷积与 C 输出通道和 S 的步幅用于

图像块编码。每个 IRB 使用大小为 E 的扩展比率。为简单起见，本文省略了图像块嵌入操作，即在第 t（$t=\{2,3,4\}$）阶段之后的 3×3 卷积，步长为 $S=2$ 。计算开支的结果是在输入大小为 224×224 的情况下得到的。

金字塔池化设置。本文根据经验将 P-型多头自注意力模块中并行池化操作的数量设置为 4。在不同的阶段，金字塔池化的池化比率 Transformer 是不同的。第一阶段的池化比率根据经验设定为 $\{12,16,20,24\}$。除了在最后阶段，接下来每个阶段的池化比率都除以 2。在最后阶段，它们被设定为 $1,2,3,4$。在每个 Transformer 块中，P-型多头自注意力模块的所有深度卷积（在式 (6.3)）都有相同的参数。

其他设置。虽然深度卷积（在式 (6.3)）的内核大小较大（如 5×5）可以带来更好的性能，但为了提高效率，所有深度卷积的内核大小被设置为 3×3。本文选择 Hardswish [230] 作为非线性激活函数，因为它比 GELU [231] 节省了很多内存。除此之外，Hardswish [230] 在实际使用上也很有效。和 PVTv2 [117] 一样，本文采用了重叠图像块编码。也就是说，本文使用 3×3 的卷积，步长为 2，用于从第二阶段到最后阶段的图像块编码，而本文应用 7×7 的卷积，步长为 4，用于第一阶段的图像块编码。

6.3 实验

本文在 §6.3.1首先介绍了在图像分类、语义分割、物体检测和实例分割任务上的实验结果，验证了 P2T 在各类主流目标检测与分割任务的有效性。然后，本文在 §6.3.2 中进行了消融实验，以更好地理解本文的方法内部模块的有效性。

6.3.1 对比结果

图像分类

图像分类是评估骨干网络性能最常见的任务。它的目的是为每个自然图像输入分配一个类别标签。许多任务通过应用建立在图像分类之上的分类网络作为特征提取的骨干。

实验设计：如 §6.2.1 中所述，这里只利用了最后阶段的输出特征 $\mathbf{B}_4$。按照常规的卷积神经网络网络 [20,21,98]，本文在 $\mathbf{B}_4$ 上面附加了一个全局平均池层和一个全连接层，以获得最终的分类分数。本文在 ImageNet-1K 数据集 [5] 上训练本文的网络，它有 128 万张训练图像和 5 万张验证图像。为了公平比较，本文

表 6.2 **数据集 ImageNet-1K [5] 上的图像分类结果。**“Top-1”表示 top-1 的准确率。“*”表示用知识蒸馏的结果 [106]。“P#(M)”表示参数量（单位为百万）。

方法	#P (M) ↓	GFLOPs ↓	Top-1 (%) ↑	FPS ↑
ResNet-18 [20]	11.7	1.8	68.5	1410
DeiT-Tiny/16* [106]	5.7	1.3	72.2	1212
ViL-Tiny [232]	6.7	1.3	76.7	441
PVT-Tiny [113]	13.2	1.9	75.1	608
PVTv2-B1 [117]	13.1	2.1	78.7	502
P2T-Tiny (本文)	**11.6**	**1.8**	**79.8**	**473**
ResNet-50 [20]	25.6	4.1	78.5	483
ResNeXt-50-32x4d [94]	25.0	4.3	79.5	407
Res2Net-50 [98]	25.7	4.5	80.3	430
DeiT-Small/16* [106]	22.1	4.6	79.9	489
PVT-Small [113]	24.5	3.8	79.8	336
T2T-ViT$_t$-14 [108]	21.5	5.2	80.7	305
Swin-T [112]	29.0	4.5	81.3	349
Twins-SVT-S [116]	24.0	2.9	81.7	439
ViL-Small [232]	25.0	4.9	82.4	187
PVTv2-B2 [117]	25.4	4.0	82.0	284
P2T-Small (本文)	**24.1**	**3.7**	**82.4**	**284**
ResNet-101 [20]	44.7	7.9	79.8	288
ResNeXt-101-32x4d [94]	44.2	8.0	80.6	228
Res2Net-101 [98]	45.2	8.3	81.2	265
PVT-Medium [113]	44.2	6.7	81.2	216
T2T-ViT$_t$-19 [108]	39.2	8.4	81.4	202
Swin-S [112]	50.0	8.7	83.0	207
ViL-Medium [232]	40.4	8.7	83.5	114
MViT-B-16 [114]	37.0	7.8	83.1	222
PVTv2-B3 [117]	45.2	6.9	83.2	189
P2T-Base (本文)	**36.1**	**6.5**	**83.5**	**182**
ResNeXt-101-64x4d [94]	83.5	15.6	81.5	147
MViT-B-24 [114]	53.5	10.9	83.0	151
ViL-Base [232]	57.0	13.4	83.7	67
PVT-Large [113]	61.4	9.8	81.7	152
DeiT-Base/16* [106]	86.6	17.6	81.8	161
ViT-Base/16 [105]	86.6	17.6	77.9	49
Swin-B [112]	88.0	15.4	83.3	140
Twins-SVT-L [116]	99.2	14.8	83.3	143
PVTv2-B4 [117]	62.6	10.1	83.6	133
PVTv2-B5 [117]	82.0	11.8	83.8	120
P2T-Large (本文)	**54.5**	**9.8**	**83.9**	**128**

按照 PVT [113] 采用与 DeiT [106] 相同的训练方式（没有知识蒸馏），这也是当前训练视觉 Transformer 的标准训练方式。具体而言，本文使用 AdamW [233] 作为优化器，初始学习率为 10^{-3}，权重衰减为 0.05，每个小批次为 1024 张图像。本文用余弦学习率衰减策略训练 P2T 300 个迭代单位。用于训练和测试的图像

大小被调整为 224 × 224。模型在前五个迭代单位中进行预热。在对 P2T 的训练中，其数据增强方式也与 [106, 113] 相同。

实验结果：定量的比较总结在表 6.2。除了按照 ViT [105] 官方以 384 × 384 的输入大小进行训练和评估，其他所有的模型都是以 224 × 224 的输入大小进行训练和评估的。FPS 是在单个 RTX 2070 GPU 上测试的。分析结果发现，P2T 的性能上大幅超越了 ResNets [20] 和 ResNeXts [94] 等常规卷积神经网络模型。例如，尽管 P2T-Tiny/Small/Base/Large 的运行时间是 ResNet-18/50/101 [20] 和 ResNeXt-101-64x4d [94] 的 2.98/1.70/1.58/1.15 倍，但 P2T-Tiny/Small/Base/Large 的最高准确率分别比 ResNet-18/50/101 [20] 和 ResNeXt-101-64x4d [94] 高 11.3%/3.9%/3.7%/2.4%。另一方面可以看出，与最近最先进的 Transformer 模型相比，本文的 P2T 也取得了优异的性能。例如，P2T-Small/Base/Large 比 Swin Transformer [112] 好 1.1%/0.5%/0.6%，且网络参数少，同时保持较低的计算开支。尽管 PVTv2 [117] 比 PVT [113] 有很大的改进，本文的 P2T-Tiny/Small/Base/Large 仍然比 PVTv2-B1/B2/B3/B4 [117] 有 1.1%/0.4%/0.3%/0.3% 的优势，参数更少，同时计算开支更低。P2T 在计算自注意力时应用了四个并行池化操作，仍然取得了与 PVTv2 [117] 相当的速度。虽然 ViL [232] 实现了与 P2T 相当的性能，但是 ViL [232] 的速度比本文的 P2T 慢得多，ViL [232] 的计算开支也比 P2T 大很多。在参数量少的情况下，P2T 在很大程度上也优于 ViT [105] 和 DeiT [106]，这意味着 P2T 在没有大量训练数据和知识蒸馏的情况下就能达到更好的性能。因此，P2T 是非常适合用于图像分类任务的。

语义分割

给定一个自然图像输入，语义分割的目的是为每个像素分配一个语义标签。它是计算机视觉中最基本的稠密预测任务之一。

实验设置：本文在 ADE20K [235] 数据集上评估了 P2T 及其他最近的著名方法。ADE20K 数据集是一个具有挑战性的目标检测与分割数据集，有 150 个细粒度的语义类别。这个数据集有 20000 张训练图片、2000 张验证图片和 3302 张测试图片。和 [113, 116] 一样，Semantic FPN [234] 被选为基本方法，用来进行公平比较。本文将 Semantic FPN [234] 的骨干网络替换成各种网络架构。所有 Sematic FPN 的方法的骨干网络都在 ImageNet-1K [5] 数据集上进行了预训练，其他层则使用 Xavier 方法初始化 [238]。所有的网络都训练了 80000 迭代次数。本文应用 AdamW [233] 作为网络优化器，初始学习率为 10^{-4}，权重衰减为 10^{-4}。

表 6.3 **ADE20K 验证集的语义分割实验结果。**

骨干网络	Semantic FPN [234]			
	参数量（M）↓	计算量（G）↓	mIoU (%) ↑	每秒帧数 ↑
ResNet-18 [20]	15.5	31.9	32.9	68
PVT-Tiny [113]	17.0	32.1	35.7	36
PVTv2-B1 [117]	17.8	33.1	41.5	30
P2T-Tiny (本文)	**15.4**	**31.6**	**43.4**	**31**
ResNet-50 [20]	28.5	45.4	36.7	35
PVT-Small [113]	28.2	42.9	39.8	26
Swin-T [112]	31.9	46	41.5	26
Twins-SVT-S [116]	28.3	37	43.2	27
PVTv2-B2 [117]	29.1	44.1	46.1	21
P2T-Small (本文)	**27.8**	**42.7**	**46.7**	**24**
ResNet-101 [20]	47.5	64.8	38.8	26
ResNeXt-101-32x4d [94]	47.1	64.6	39.7	20
PVT-Medium [113]	48.0	59.4	41.6	19
Swin-S [112]	53.2	70	45.2	18
Twins-SVT-B [116]	60.4	67	45.3	17
PVTv2-B3 [117]	49.0	60.7	47.3	15
P2T-Base (本文)	**39.8**	**58.5**	**48.7**	**16**
ResNeXt-101-64x4d [94]	86.4	104.2	40.2	15
PVT-Large [113]	65.1	78.0	42.1	15
Swin-B [112]	91.2	107	46.0	13
Twins-SVT-L [116]	102	103.7	46.7	13
PVTv2-B4 [117]	66.3	79.6	48.6	11
PVTv2-B5 [117]	85.7	89.4	48.9	10
P2T-Large (本文)	**58.1**	**77.7**	**49.4**	**12**

采用 $\gamma = 0.9$ 的 *poly* 学习率计划。每个小批次有 16 张图像；用于训练的图像被调整大小并随机裁剪为 512×512。还启用了跨 GPU 的同步批次规范化。在测试过程中，图像短边被调整到 512，长边按比例进行调整。多尺度测试和翻转功能被禁用。与 [113] 一致，本文同样使用 MMSegmentation 工具箱 [239] 来实现上述实验。

实验结果：定量比较结果显示在表 6.3，可视化比较结果显示在图 6.3。表 6.3中 GFlops 的数量是以 512×512 的输入大小计算的，FPS 是在单张 RTX 2070 GPU 上测试的，P2T 骨干网络的结果以**粗体**标记。本文将 P2T 与 ResNets [20]、ResNeXts [94]、PVT [113]、Swin Transformers [112]、Twins [116] 和 PVTv2 [117] 进行比较。每个网络的结果都来自于官方论文或使用官方训练配置重新训练的结果。受益于金字塔池化技术，采用 P2T 骨干的 Semantic FPN [234] 的结果要比其他卷积神经网络和 Transformer 的竞争对手好得多。在参数量和计算量更小的

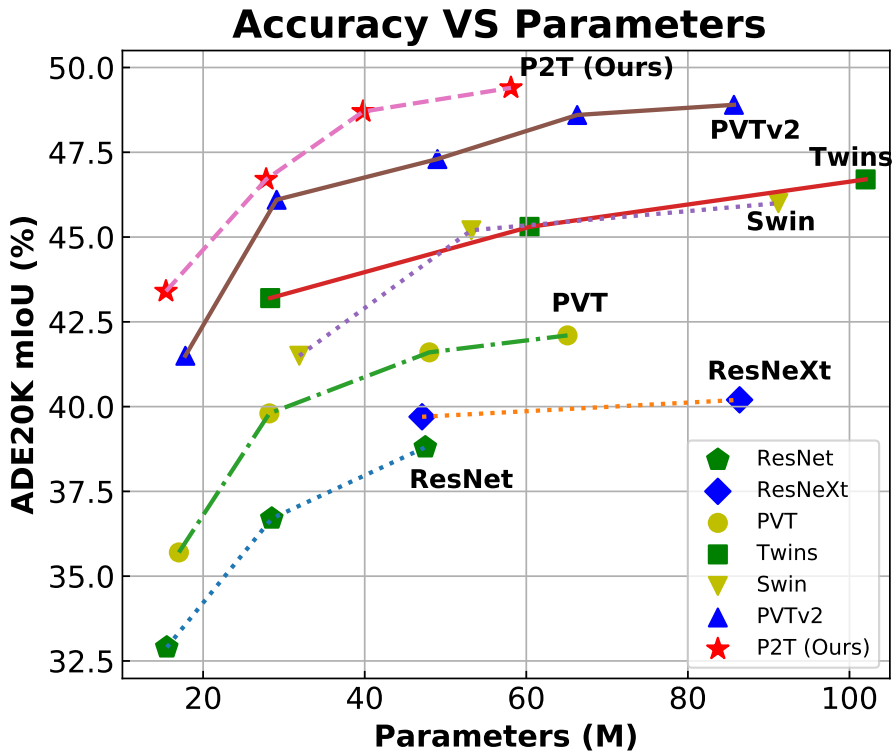


图 6.3 **在 ADE20K 数据集上语义分割的实验结果 [235]**。本文采用 Semantic FPN [234] 作为基本方法，并使用了包括 ResNet [20]、ResNeXt [94]、PVT [113]、Twins [116]、Swin Transformer [112]、PVTv2 [117] 以及本文提出的 P2T 等不同的骨干网络。

前提下，P2T-Tiny/Small/Base/Large 的 mIoU 性能分别比 ResNet-18/50/101 [20] 和 ResNeXt-10-64x4d [94] 提高了 10.5%/10.0%/9.9%/9.2%。与 Swin Transformer [112] 相比，P2T-Small/Base/Large 分别比 Swin-T/S/B [112] 实现了 5.2%/3.5%/3.4% 的提升，表明全局关系建模对于视觉识别具有重要意义。Twins [116] 结合了 Swin Transformers [112] 的局部自我注意力和 PVT [113] 的整体自我注意力。可以看出，Twins [116] 比 Swin Transformers [112] 表现更好，这说明全局自我注意力很重要。与 Twins [116] 不同的是，本文通过金字塔池化来应用纯粹的全局自我注意力，学习更丰富的场景信息。P2T-Small/Base/Large 的 mIoU 性能相对 Twins-SVT-S/B/L [116] 提升了 3.5%/3.4%/2.7%。PVTv2 [117] 是 PVT [113] 的改进版，作为本文 P2T 的最强竞争对手。P2T-Tiny/Small/Base/Large 分别比 PVTv2-B1/B2/B3/B4 [117] 要强 1.9%/0.6%/1.4%/0.8%。此外，P2T-Tiny/Small/Base/Large 总是比相应的 PVTv2-B1/B2/B3/B4 [117] 有更少的参数，更少的计算开支，以及更快的速度。最后，本文发现 P2T-Tiny 的性能相对 ResNeXt-101-64x4d [94] 提

表 6.4 在 **MS-COCO** `val2017` **数据集 [236] 上，用 RetinaNet [237] 进行物体检测的结果。**

骨干网络	参数量 (M) ↓	计算量 (G)↓	每秒帧数 ↑	RetinaNet [237] AP ↑	AP_{50}	AP_{75}	AP_S ↑	AP_M	AP_L
ResNet-18 [20]	21.3	190	19.3	31.8	49.6	33.6	16.3	34.3	43.2
ViL-Tiny [232]	16.6	204	4.2	40.8	61.3	43.6	26.7	44.9	53.6
PVT-Tiny [113]	23.0	205	10.7	36.7	56.9	38.9	22.6	38.8	50.0
PVTv2-B1 [117]	23.8	209	8.5	40.2	60.7	42.4	22.8	43.3	54.0
P2T-Tiny (本文)	**21.1**	**206**	**9.3**	**41.3**	**62.0**	**44.1**	**24.6**	**44.8**	**56.0**
ResNet-50 [20]	37.7	239	13.0	36.3	55.3	38.6	19.3	40.0	48.8
PVT-Small [113]	34.2	261	7.7	40.4	61.3	43.0	25.0	42.9	55.7
Swin-T [112]	38.5	248	9.7	41.5	62.1	44.2	25.1	44.9	55.5
ViL-Small [232]	35.7	292	3.4	44.2	65.2	47.6	28.8	48.0	57.8
Twins-SVT-S [116]	34.3	236	8.5	43.0	64.2	46.3	28.0	46.4	57.5
PVTv2-B2 [117]	35.1	266	5.8	43.8	64.8	46.8	26.0	47.6	59.2
P2T-Small (本文)	**33.8**	**260**	**7.4**	**44.4**	**65.3**	**47.6**	**27.0**	**48.3**	**59.4**
ResNet-101 [20]	56.7	315	9.8	38.5	57.8	41.2	21.4	42.6	51.1
ResNeXt-101-32x4d [94]	56.4	319	8.5	39.9	59.6	42.7	22.3	44.2	52.5
PVT-Medium [113]	53.9	349	5.7	41.9	63.1	44.3	25.0	44.9	57.6
Swin-S [112]	59.8	336	7.1	44.5	65.7	47.5	27.4	48.0	59.9
PVTv2-B3 [117]	55.0	354	4.5	45.9	66.8	49.3	28.6	49.8	61.4
P2T-Base (本文)	**45.8**	**344**	**5.0**	**46.1**	**67.5**	**49.6**	**30.2**	**50.6**	**60.9**
X-101-64x4d [94]	95.5	473	6.2	41.0	60.9	44.0	23.9	45.2	54.0
PVT-Large [113]	71.1	450	4.4	42.6	63.7	45.4	25.8	46.0	58.4
Twins-SVT-B [116]	67.0	376	5.1	45.3	66.7	48.1	28.5	48.9	60.6
PVTv2-B4 [117]	72.3	457	3.4	46.1	66.9	49.2	28.4	50.0	62.2
PVTv2-B5 [117]	91.7	514	3.2	46.2	67.1	49.5	28.5	50.0	62.5
P2T-Large (本文)	**64.4**	**449**	**3.8**	**47.2**	**68.4**	**50.9**	**32.4**	**51.6**	**62.2**

升了 3.2%，速度快一倍。基于上述观察，本文可以得出结论，P2T 非常能胜任语义分割任务。

物体检测

物体检测是计算机视觉领域几十年来最基本和最具挑战性的任务之一。它的目的是检测和识别自然图像中一般化的语义对象的实例。在这里，本文在 MS-COCO [236] 数据集上评估了 P2T 及其他主流方法。

实验设置： MS-COCO [236] 是一个大规模的比赛数据集，用于物体检测、实例分割和关键点检测。MS-COCO `train2017`（118k 张图像）和 `val2017`（5k 张图像）集在本文的实验中分别用于训练和验证。本文采用 RetinaNet [237] 网络作为基本框架，因为它已经被社区广泛认可 [112, 113]。在训练中，每个小批

次有 16 张图像，初始学习率为 10^{-4}。按照 MMDetection 工具箱 [240]，本文对每个网络进行 12 个迭代单位的训练，在 8 和 11 个迭代单位之后，学习率被除以 10。网络优化器是 AdamW [233]，权重衰减被设置为 10^{-4}。在训练和测试过程中，输入图像的短边被调整为 800 像素，长边将保持图像的比例在 1333 像素以内。在训练阶段，只有随机水平翻转被用于数据增强。本文使用标准的 `COCO API` 进行评估，使用 AP、AP_{50}、AP_{75}、AP_S、AP_M 和 AP_L 等指标报告结果。AP_S、AP_M 和 AP_L 分别是指小型、中型和大型对象的 AP 得分，具体定义在 [236]。AP 通常被看作是主要指标。对于每个指标来说，更高的分数代表更好的性能。除此之外，本文还报告了每个方法的参数量、计算开支、速度以供参考。

实验结果：对 MS-COCO 数据集的评估结果在表 6.4进行了总结。其他网络的结果来自官方论文或使用官方配置重新实现。计算量是以 800×1280 的输入大小计算的，每秒帧数是在单个 RTX 2070 GPU 上测试的。P2T 骨干网的结果以**粗体**标记。下面的讨论如果没有说明，则是指 AP 的度量。本文可以观察到，本文的 P2T 在所有微小/小/大的复杂性设置下都取得了最好的性能。例如，P2T-Small 比 Swin-T [112]、Twins-SVT-S [116] 和 PVTv2-B2 [117] 分别实现了 2.9%、1.4% 和 0.6% 的 AP 提升。P2T-Tiny 比 PVTv2 [117] 好 1.1%。与 ViL [232] 相比，P2T-Tiny/Small 分别比 ViL-Tiny/Small [232] 好 0.5% 和 0.2%。需要指出的是，ViL [232] 的运行速度比 P2T 慢得多，如表 6.4所示。在基本复杂度设置下，P2T-Base 比 Swin-S [117] 高出 1.0%，比最佳竞争对手 PVTv2-B3 高出 0.2%。在大复杂度设置下，P2T-Large 比 PVTv2-B4 [117] 和 Twins-SVT-B [116] 分别取得了 1.1% 和 1.9% 的更好 AP。在所有的复杂程度上，P2T 总是优于 PVTv2 [117]，网络参数更少，计算开支更低，速度更快。P2T-Tiny/Small/Base/Large 分别比 ResNet-18/50/101 [20] 和 ResNeXt-101-64x4d [94] 好 9.5%/8.1%/7.0%/6.2%。综上所述，P2T 在物体检测方面有很强的能力。

实例分割

实例分割是另一项基本的目标检测与分割任务，它可以被看作是在物体检测之上更高级的任务。相对物体检测来说，它输出的是细粒度的物体掩码区域而不是物体检测中的边缘框。

实验设置：本文在著名的 MS-COCO 数据集 [236] 上评估实例分割的性能。MS-COCO `train2017` 训练集和 `val2017` 验证集在本文的实验中被用于训练和验证。Mask R-CNN [241] 被用作基本框架，使用不同的骨干网络。训练设置与本

表 6.5 在 **MS-COCO** `val2017` **数据集 [236] 上，用 Mask R-CNN [241] 进行实例分割的结果。**

骨干网络	参数量 (M) ↓	计算量 (G) ↓	每秒帧数 ↑	Mask R-CNN [241]					
				AP^b ↑	AP^b_{50}	AP^b_{75}	AP^m ↑	AP^m_{50}	AP^m_{75}
ResNet-18 [20]	31.2	209	17.3	34.0	54.0	36.7	31.2	51.0	32.7
ViL-Tiny [232]	26.9	223	3.9	41.4	63.5	45.0	38.1	60.3	40.8
PVT-Tiny [113]	32.9	223	10.0	36.7	59.2	39.3	35.1	56.7	37.3
PVTv2-B1 [117]	33.7	227	8.0	41.8	64.3	45.9	38.8	61.2	41.6
P2T-Tiny (本文)	**31.3**	**225**	**8.8**	**43.3**	**65.7**	**47.3**	**39.6**	**62.5**	**42.3**
ResNet-50 [20]	44.2	260	11.5	38.0	58.6	41.4	34.4	55.1	36.7
PVT-Small [113]	44.1	280	7.0	40.4	62.9	43.8	37.8	60.1	40.3
Swin-T [112]	47.8	264	8.8	42.2	64.6	46.2	39.1	61.6	42.0
ViL-Small [232]	45.0	310	3.2	44.9	67.1	49.3	41.0	64.2	44.1
Twins-SVT-S [116]	44.0	254	7.7	43.4	66.0	47.3	40.3	63.2	43.4
PVTv2-B2 [117]	45.0	285	5.4	45.3	67.1	49.6	41.2	64.2	44.4
P2T-Small (本文)	**43.7**	**279**	**6.7**	**45.5**	**67.7**	**49.8**	**41.4**	**64.6**	**44.5**
ResNet-101 [20]	63.2	336	9.1	40.4	61.1	44.2	36.4	57.7	38.8
X-101-32x4d [94]	62.8	340	7.9	41.9	62.5	45.9	37.5	59.4	40.2
PVT-Medium [113]	63.9	367	5.3	42.0	64.4	45.6	39.0	61.6	42.1
Swin-S [112]	69.1	354	6.6	44.8	66.6	48.9	40.9	63.4	44.2
PVTv2-B3 [117]	64.9	372	4.2	47.0	68.1	51.7	42.5	65.7	45.7
P2T-Base (本文)	**55.7**	**363**	**4.7**	**47.2**	**69.3**	**51.6**	**42.7**	**66.1**	**45.9**
ResNeXt-101-64x4d [94]	101.9	493	5.7	42.8	63.8	47.3	38.4	60.6	41.3
PVT-Large [113]	81.0	469	4.1	42.9	65.0	46.6	39.5	61.9	42.5
Twins-SVT-B [116]	76.3	395	4.6	45.2	67.6	49.3	41.5	64.5	44.8
PVTv2-B4 [117]	82.2	475	3.2	47.5	68.7	52.0	42.7	66.1	46.1
PVTv2-B5 [117]	101.6	532	3.0	47.4	68.6	51.9	42.5	65.7	46.0
P2T-Large (本文)	**74.0**	**467**	**3.5**	**48.3**	**70.2**	**53.3**	**43.5**	**67.3**	**46.9**

文在 §6.3.1 中用于物体检测的设置相同。本文以 AP^b、AP^b_{50}、AP^b_{75}、AP^m、AP^m_{50} 和 AP^m_{75} 指标报告对象检测和实例分割的评估结果，其中"b" 和"m" 分别表示边缘框（Bounding Box）和掩码指标（Mask Metric）。AP^b 和 AP^m 被设定为主要评价指标。

实验结果： P2T 与其对比方法的比较显示在表 6.5。Flops 的数量是以 800×1280 的输入大小计算的。FPS 是在单个 RTX 2070 GPU 上测试的。与现有的卷积神经网络和 Transformer 骨干网络相比，P2T 在所有的网络复杂层级上实现了最佳性能。在边缘框指标 AP^b 上，P2T-Small/Base 比 Swin-T/S [112] 好 3.3%/2.4%，P2T-Small/Large 比 Twins-SVT-S/B [116] 好 2.1%/3.1%。P2T-Tiny/Small/Base/Large 分别比 PVTv2-B1/B2/B3/B4 [117] 好 1.5%/0.2%/0.2%/0.8%，且参数少、计算开

表 6.6 基于 **P2T-Small** 的 **EDN** 网络在显著性目标检测上的性能。

方法	$F_\beta^{\max}$	MAE	S_α	$E_\xi^{\max}$
EDN	0.893	0.035	0.892	0.934
+ P2T-Small	0.910	0.030	0.909	0.948

表 6.7 基于 **P2T-Tiny** 的 **MobileSal** 网络在 **RGB-D** 显著性目标检测上的性能。

方法	$F_\beta^{\max}$	MAE	S_α	$E_\xi^{\max}$
MobileSal	0.906	0.045	0.896	0.934
+ P2T-Tiny	0.920	0.038	0.909	0.948

表 6.8 基于 **P2T-Small** 的 **JCS** 网络在新冠肺炎 **CT** 病灶检测上的性能。

方法	Dice	IoU	E_ϕ
JCS	0.775	0.654	0.920
+ P2T-Small	0.796	0.674	0.935

支低、速度快。在掩码指标 AP^m 方面，本文也观察到与使用边缘框指标 AP^b 类似的提升。与基于 ResNet 的骨干网相比，P2T 在所有复杂程度上都明显优于 ResNets [20] 和 ResNeXts [94]。同样令人惊讶的是，本文最轻的 P2T-Tiny 比 ResNeXt-101-64x4d [94] 在边缘框和掩码指标方面分别好 0.5% 和 1.2%。综上所述，P2T 对于实例分割来说是非常有能力的。

显著性目标检测

在第三章中，本文提出了基于极致采样技术的显著性目标检测网络 EDN，它基于已有的 VGG [19] 或 ResNet-50 网络 [20]。本文将 EDN 的骨干网络更换为 P2T-Small，其计算复杂度与 ResNet-50 相当。在替换骨干网络后，本文以相同的实验设置进行训练，在 DUTS-TE [153] 数据集上进行测试。对比实验结果如表 6.6所示。尽管原始 EDN 算法已经大幅优于目前主流的算法，基于 P2T-Small 的 EDN 算法性能还能显著优于原始的 EDN 算法，即在所有评价指标上都有较大的提升幅度。该结果显示，P2T 能够大幅提升原始 EDN 网络在骨干网络方面的特征表达能力，从而大幅提升算法性能。P2T 与基于极致下采样的目标定位技术形成互补关系，两者结合可以更好地完成目标的检测与分割。

表 6.9 **对多个金字塔池化比率的消融研究。**“Top-1” 表示 ImageNet-1K 验证集 [5] 上的 top-1 分类准确率，“mIoU” 表示 ADE20K 数据集 [235] 上的语义分割结果。

序号	池化系数	下采样比率↑	Top-1 (%)↑	mIoU (%)↑
1	24	576	70.6	27.5
2	16	256	72.5	33.0
3	12	144	73.9	34.3
4	8	64	73.9	34.4
5	12, 24	115	74.4	34.8
6	12, 16, 20, 24	66	74.7	35.7

RGB-D 显著性目标检测

在第四章中，本文提出了极致高效的 RGB-D 显著性目标检测网络 MobileSal，它基于 MobileNetV2 网络 [123]。MobileNetV2 的计算复杂度与 P2T-Tiny 最接近，因此本文将 MobileSal 的骨干网络更换为 P2T-Tiny，并以相同的实验设置进行训练，在最常用的 NJU2K 数据集 [173] 上进行测试。评价指标采用最大 F 度量、MAE、S 度量及最大 E 度量。对比实验结果如表 6.7所示。结果显示，经 P2T-Tiny 优化后的 MobileSal 算法相比原始的 MobileSal 算法在四种评价指标上均有大幅提升。P2T 能很好地与基于隐式信息恢复的高效融合技术结合，使后者拥有更高的上限。

新冠肺炎 CT 病灶检测

第五章介绍了基于图像二元感知的新冠肺炎 CT 病灶检测方法 JCS，它基于 VGG16 网络 [19]。本文将其用 P2T-Small 替代，作为 JCS 的全新骨干网络，并评估替代后的性能提升。具体来说，本文仍采用与原始 JCS 算法相同的训练策略，在 COVID-CS 数据集上进行测试。评价指标采用 Dice、IoU 及 E 度量。实验结果如表 6.8所示。结果显示，P2T 使得 JCS 的性能在三大评价指标上均有较大幅度的提升，这体现了 P2T 与 JCS 方法的互补性。

6.3.2 消融实验

实验设置：在本节中，本文进行了消融实验，以分析每个设计选择在 P2T 中的作用。本文评估了各种实验设置在语义分割和图像分类上的表现。由于计

表 6.10 **在不同阶段用多个池化操作取代单一池化操作的消融实验。**由于本文的网络的第 2 阶段和第 3 阶段都只有两个基本块，本文把它们合并为一个选择。

序号	网络的阶段序号			Top-1 (%) ↑	mIoU ↑
	[2, 3]	4	5		
1				73.9	34.4
2	✔			74.1	34.9
3	✔	✔		74.5	35.5
4	✔	✔	✔	74.7	35.7

表 6.11 **关于池化操作选择的消融研究。**可以发现，其他选择的效果比平均池化差。

池化类型	Top-1 (%) ↑	mIoU (%) ↑
平均值池化	74.7	35.7
最大值池化	73.0	33.2
深度卷积	73.8	35.5

算资源有限，本文只在 ImageNet 数据集 [5] 上对每组设置训练 100 个迭代循环，而其他训练设置保持与 §6.3.1 中相同。然后，本文在 ADE20K 数据集 [235] 上微调 ImageNet 预训练模型，训练设置与 §6.3.1 相同。

探究多种金字塔池化比率：为了验证使用多种池化比率的意义，本文进行了实验，以评估 P2T 与一个/两个/四个平行池化操作的性能。基线是没有相对位置编码、IRB 和重叠图像块编码的 P2T-Small。结果显示在表 6.9。可以看出，具有大池子比率的单一池子操作（例如 16, 24）对序列长度有很大的下采样比率。尽管如此，它在图像分类和语义分割方面的性能都非常差。然而，当单一池化操作的池化率为 12 时，如果本文进一步降低池化率，性能将达到饱和。当本文采用两个平行的池化操作时，即使有很高的下采样比率，对图像分类和语义分割来说，性能仍然变得更好。当本文有四个并行的池化操作时，本文的下采样率与池化率为 8（PVT 中的设置）的下采样率相同，且达到了最佳性能。

探究金字塔池化在不同网络阶段的效果：本文对 P2T 的金字塔池化设计进行了不同阶段的消融研究。由于第 1 阶段只包含下采样的卷积，本文不在第 1 阶段进行这种消融研究。基线与上次消融研究相同。单个池化操作的池化率被设置为 8，以确保相同的下采样率。结果显示在表 6.10。本文可以看到金字塔池

表 6.12 对固定池化大小的消融实验。

池化操作	计算开支（G）↓	显存占用 ↓	Top-1 (%) ↑	mIoU (%) ↑
固定池化比率	41.6	3.3	74.7	35.7
固定池化大小	38.9	2.9	74.4	33.3

表 6.13 相对位置编码、IRB 和重叠图像块编码的消融实验。

相对位置编码	IRB	重叠图像块编码	Top-1 (%) ↑	mIoU (%) ↑
			74.7	35.7
✔			76.4	37.4
✔	✔		79.5	42.7
✔	✔	✔	79.7	44.1

化可以提高所有阶段的性能。当更多的阶段应用多个池化操作时，性能变得更高。从结果来看，在第 4 阶段（表 6.10 的第 3 号）应用多个池化操作的改进比其他阶段（表 6.10 的第 2、4 号）更大，因为第 4 阶段比第 [2, 3] 阶段和第 5 阶段有更多的基本模块数。

探究池化操作的选择：本文对不同的池化操作进行了实验，如表 6.11 中所示。有三种典型的选择，即最大池化、深度卷积和默认的平均池化。深度卷积的核大小与最大/平均池化相同，以保持相同的下采样率。很明显，不同的池化类型并不影响计算的复杂性，它们只影响下采样核的参数数量。关于 ImageNet 分类准确率的结果 [5] 和 ADE20K 分割 mIoU 的结果 [235]，平均池化要比其他两种选择好得多。因此，本文应用平均池化作为默认的池化选择。

探究固定的池化大小：当使用固定的池化比率时，池化特征图的尺寸会随着输入特征图的变化而变化。在这里，本文试图将所有阶段的池化大小固定为：1, 2, 3, 6。同时在所有阶段，都使用自适应平均池化。结果显示在表 6.12。与本文的默认设置相比，固定的池化大小大约节省了 10% 的内存用量和 12% 的计算开支。然而，top-1 的分类精度下降了 0.3%。而语义分割的性能则降低了 2.4%。因此，本文选择使用固定的池化比率，而不是固定的池化大小。

探究激活函数的选择：本文使用 Hardswish 函数 [230] 进行非线性激活函数，以减少训练阶段的 GPU 内存使用。通常情况下，当本文在 ImageNet [5] 上训练 P2T-Small、批次大小为 64 的情况下时，GELU [231] 的 GPU 内存使用量

为 10.5GB，比 Hardswish [230] 多 3.6GB（+52%）。本文还发现，如果 P2T 采用 Hardswish [230]，准确率没有明显下降。

探究其他设计：为了验证其他设计选择的有效性，如相对位置编码、IRB 和重叠图像块编码，本文在基线上逐一添加这些组件。实验结果显示在表 6.13。可以看出，相对位置编码对图像分类和语义分割都有明显的改善。在大的池化比率下，池化后的特征会有较小的尺度，所以相对位置编码只需要可忽略的计算开支（对于 224×224 的输入大小，仅需 5M Flops）。前馈网络中额外的深度卷积，即 IRB，也显示出明显的性能提升，说明捕捉二维近邻关系在 P2T 中也是必要的。

6.4 本章总结

本节将金字塔池化引入多头自注意力模块，以减轻多头自注意力模块在视觉 Transformer 中的高计算开支。与多头自注意力模块中应用单一池化操作的策略相比，本节基于金字塔池化的多头自注意力模块不仅减少了序列长度，而且通过金字塔池化同时学习了强大的上下文表征学习。通过基于金字塔池化的多头自注意力模块，本节构建了一个新的骨干网络，称为金字塔池化 Transformer，简称 P2T。本节在几个基础视觉任务上进行了广泛的实验，包括图像分类及具有代表性的目标检测与分割任务，如语义分割、物体检测和实例分割。实验结果表明，P2T 明显优于以前基于卷积神经网络和 Transformer 的骨干网络。通过 P2T 网络，本文还进一步大幅度地提升了第三、四、五章所提出算法的性能。

第 7 章 总结与展望

目标检测与分割作为计算机视觉最重要的研究领域之一，是诸多下游应用的基础。仅在单一层次考虑检测与分割很容易丢失目标。构建多层次目标检测与分割算法，才能够满足复杂多变环境下的目标检测与分割要求。而在复杂场景下，设计多层次算法存在诸多挑战。本文聚焦于目标定位难、多通道特征融合慢、标注数据难获取、多尺度建模能力不足等主要挑战，提出了对应的解决方案。本章将对各章内容进行总结，并对相关的未来研究方向进行展望。

7.1 本文工作总结

首先，本文在第一章介绍了计算机视觉中目标检测与分割的重要意义，以及复杂场景中设计多层次算法存在的主要挑战，主要为四点：目标定位难、多通道特征融合慢、标注数据难获取以及多尺度建模能力不足。通过对它们的详细分析，本文提出了基于极致下采样的目标定位技术，聚焦于目标定位难的问题。然后，通过隐式信息恢复技术，保证算法特征融合的高效性，从而使算法能够应用到体积小、功耗低的移动设备上。接着，本文提出的基于注意力融合的图像二元感知技术可以在标注数据难获取的情况下保证算法的精度和鲁棒性。最后，本文提出了基于金字塔池化的骨干特征提取技术，降低模型复杂度的同时还提升了多尺度建模能力。

在第二章中，本文回顾了多种具体的目标检测与分割任务的相关工作，即显著性目标检测、RGB-D 显著性目标检测、新冠肺炎 CT 病灶分割，并分析了目前骨干网络设计相关的研究工作以及它们所存在的不足。

在第三章中，本文提出了基于极致下采样的目标定位技术，聚焦于更好的目标定位。该技术通过不断地下采样至一维向量，消除了以往算法定位目标需要大分辨率的要求，降低了模型复杂度，还可以更好地学习整个图像的全局视图，从而更准确地定位显著性目标。基于极致下采样的目标定位技术，本文构建了 EDN 网络，应用与显著性目标检测任务。为了验证 EDN 网络的有效性，本文在五大知名数据集上与过去数年的方法使用六个常用的评价指标进行了对比。实验结果显示 EDN 不仅模型小、速度快，而且在以上数据集上综合所有指标取

得了最好的精度性能。EDN 的轻量版本 EDN-Lite 速度是目前轻量方法的 1.7 倍，且其在最具挑战性的 DUTS-TE [153] 数据集上取得了 4.0% 的 $F_\beta^{\max}$ 性能提升。EDN-LiteEX 进一步提速了约 3 倍，仍相对已有的轻量方法在 DUTS-TE [153] 数据集上取得了 2.6% 的 $F_\beta^{\max}$ 分数提升。本文还研究了目前主流方法对目标定位精确度的提升趋势，发现近年来的方法的目标定位能力趋于饱和，而本文设计的 EDN 的目标定位能力大幅超越了目前所有的主流方法。

第四章主要聚焦于多通道特征融合慢的问题，并提出了基于隐式信息恢复的特征融合技术，应用于 RGB-D 显著性目标检测任务中。目前的算法因花费大量算力在多通道特征融合上，导致算法速度过慢，难以在体积小功耗低的移动设备上使用，限制了其在真实世界中的应用。本文提出的隐式信息恢复技术，确保了算法在特征融合上的高效率，还提高了算法的精度。基于隐式信息恢复技术，本文提出了极致高效的 RGB-D 显著性目标检测算法 MobileSal。MobileSal 先提出隐式深度恢复技术来保证 RGB-D 特征融合时的深度信息不被丢失，再在最粗糙的特征层级进行 RGB-D 特征融合，大幅降低了特征融合的计算成本。为了验证 MobileSal 模型的有效性，本文在六大知名数据集上进行了定量和定性对比实验，结果显示算法相对已有主流算法在速度上提升了 15 ～ 150 倍，同时保持了相当的精度性能，甚至在 NJU2K [173] 数据集取得了 91.4% 的 $F_\beta^{\max}$ 分数，比已有的算法还要高 0.2%。

本文在第五章提出了基于注意力融合的图像二元感知技术，其聚焦于解决标注数据获取难的情况下的算法设计问题，应用于新冠肺炎 CT 病灶分割。它通过注意力融合充分发掘病灶区域的对比信息，保持病灶位置的高度敏感性，同时使用分类、分割的二元图像感知来更充分地利用更多的样本数据，提升了精度且保持了鲁棒性。本文还搜集了一个大型的新冠肺炎 CT 数据集 COVID-CS，包含了 750 例已知类别的样本的超过 144K 张 CT 图像，并同时标注了类别标签和像素级的病灶分割位置，便于分类、分割的二元感知。在 COVID-CS 数据集上，本文所提出的算法相对其他主流方法取得了 8.8% 的 Dice 分数提升，且仅在 0.8% 的测试图像上失效。基于该算法的诊断系统也帮助医生更快地给出诊断结果，速度相对仅靠医生自行判断提升了 20 余倍，且保持了相同的诊断精度。

第六章主要讨论如何提升目标检测与分割算法的多尺度建模能力，并提出了基于金字塔池化的 Transformer 即 P2T。它不仅能够降低模型复杂度，还能够提升算法的多尺度特征表达能力。通过堆叠不同数量的 P2T 基础模块可以构建

出不同复杂的骨干网络 P2T-Tiny/Small/Base/Large，并根据其复杂度而应用于不同环境和算力要求下的任务。P2T 骨干网络在各种典型的目标检测与分割任务上，如语义分割、物体检测、实例分割等，P2T 取得了与其他主流骨干网络更好的性能，同时还保持了较低的模型复杂度和计算量。在 MS-COCO [236] 数据集上的物体检测和实例分割任务以及在 ADE20K [235] 数据集上的语义分割任务上，实验还发现最轻量的 P2T-Tiny 骨干网络性能比卷积神经网络中最复杂的 ResNeXt-101-64x4d [94] 骨干网络还要强，所以本文所提出的 P2T 网络相对主流基于卷积神经网络的骨干网络具有强得多的特征表达能力。此外，P2T 可以作为新的骨干网络直接应用到第三、四、五章所提出的算法中，且经实验验证，P2T 显著提升了它们的性能，与它们构成互补关系。

7.2 未来研究展望

在复杂场景下，如何设计多层次目标检测与分割算法是巨大的挑战。本文主要分析了关于目标定位、特征融合、数据多样性、特征表达等四方面的主要挑战，并提出了相应的解决方案。但是这些挑战仍值得更深层次的研究，具体来说：

1. 第三章介绍了基于极致下采样的目标定位技术，它利用更深的下采样来达到高效、精确定位目标的目的。虽然该技术帮助算法取得了最佳性能，但是在下采样的过程中，不可避免地发生了特征信息丢失。也导致了在一些场景下会定位到错误的目标或遗失目标。如何在下采样的过程减少特征信息丢失，以提高定位的准确性，是未来值得进一步研究的话题。
2. 第四章介绍了基于隐式信息恢复的特征融合技术。本文基于该技术提出了极致高效的 RGB-D 显著性目标检测算法，它利用隐式深度恢复策略来保证算法在特征融合不会遗忘输入特征，并保持了特征融合的高效性，同时还大幅提升了算法精度。然而，目前最强大的标准方法的精度仍有一定距离，如何进一步保持其高效架构的前提下进一步提升其精度，如引入边缘信息、自监督预训练、额外的训练监督、更有效的特征融合策略等等，是值得进一步研究的话题。
3. 第五章提出了基于注意力融合的图像二元感知技术，可以在获取标注数据难的情况下精准地检测目标，利用更多的仅类别标签的分类特征，提高了算法精度的同时还保证了鲁棒性。然而，其仍然基于多个数据来源

的数据集，不具有数据隐私性。在实际应用中，数据隐私是一个非常重要的问题，如何在保证数据隐私的前提下，充分利用数据，如引入联邦学习等，训练出精度较高的模型，是值得进一步研究的话题。

4. 第六章介绍了基于金字塔池化的的 Transformer 技术，在多个主流的目标检测与分割任务上取得了最好的结果。然而，其架构还可以进一步改进，比如可以在计算注意力的过程中引入更先进的相对值编码，还可以将其与自监督预训练的方式结合起来，对监督类别标签进行解绑，获取更加鲁棒的骨干特征。

参考文献

[1] J. Janai, F. Güney, A. Behl, A. Geiger, et al., Computer vision for autonomous vehicles: Problems, datasets and state of the art, Foundations and Trends® in Computer Graphics and Vision 12(1–3) 2020, 1–308.

[2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Int. Conf. Comput. Vis., 2015.

[3] S. Xu, J. Wang, W. Shou, T. Ngo, A.-M. Sadick, X. Wang, Computer vision techniques in construction: a critical review, Archives of Computational Methods in Engineering 28(5) 2021, 3383–3397.

[4] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, R.-G. Zhang, M.-M. Cheng, JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation, IEEE Trans. Image Process. 30 2021, 3113–3126.

[5] 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, Int. J. Comput. Vis. 115(3) 2015, 211–252.

[6] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[7] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[8] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, H. Lu, Towards high-resolution salient object detection, in: Int. Conf. Comput. Vis., 2019.

[9] 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: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[10] 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 Trans. Pattern Anal. Mach. Intell. 40(4) 2017,

834–848.

[11] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P. H. Torr, Deeply supervised salient object detection with short connections., IEEE Trans. Pattern Anal. Mach. Intell. 41(4) 2019, 815.

[12] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, M.-M. Cheng, EGNet: Edge guidance network for salient object detection, in: Int. Conf. Comput. Vis., 2019.

[13] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, et al., Calibrated rgb-d salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[14] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, L. Zhang, Contrast prior and fluid pyramid integration for RGBD salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[15] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, et al., Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct, Radiology (2020).

[16] S. Jin, B. Wang, H. Xu, C. Luo, L. Wei, W. Zhao, X. Hou, W. Ma, Z. Xu, Z. Zheng, et al., Ai-assisted ct imaging analysis for covid-19 screening: Building and deploying a medical ai system in four weeks, MedRxiv (2020).

[17] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Med. Image. Comput. Comput. Assist. Interv., Springer, 2015.

[18] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imaging 39(6) 2019, 1856–1867.

[19] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Int. Conf. Learn. Represent., 2015.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[21] G. Huang, Z. Liu, G. Pleiss, L. Van Der Maaten, K. Weinberger, Convolutional networks with dense connectivity, IEEE Trans. Pattern Anal. Mach. Intell. (2019).

[22] 张平平, 图像内容的显著性与相似性研究, Ph.D. thesis, 大连理工大学 (2020).

[23] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, J. Yang, Saliency propagation from simple to difficult, in: IEEE Conf. Comput. Vis. Pattern Recog., 2015.

[24] W.-C. Tu, S. He, Q. Yang, S.-Y. Chien, Real-time salient object detection with a minimum spanning tree, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[25] C. Xia, J. Li, X. Chen, A. Zheng, Y. Zhang, What is and what is not a salient object? Learning salient object detector by ensembling linear exemplar regressors, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[26] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, C. Hou, An iterative co-saliency framework for RGBD images, IEEE Trans. Cybernetics 49(1) 2017, 233–246.

[27] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37(3) 2015, 569–582.

[28] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, N. Zheng, Salient object detection: A discriminative regional feature integration approach, Int. J. Comput. Vis. 123(2) 2017, 251–268.

[29] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: IEEE Conf. Comput. Vis. Pattern Recog., 2013.

[30] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: IEEE Conf. Comput. Vis. Pattern Recog., 2015.

[31] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: IEEE Conf. Comput. Vis. Pattern Recog., 2015.

[32] G. Lee, Y.-W. Tai, J. Kim, Deep saliency with encoded low level distance map and high level features, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[33] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation., IEEE Trans. Pattern Anal. Mach. Intell. 39(4) 2017, 640–651.

[34] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, in: Int. Conf. Comput. Vis.,

2017.

[35] S. Chen, X. Tan, B. Wang, X. Hu, Reverse attention for salient object detection, in: Eur. Conf. Comput. Vis., 2018.

[36] Y. Zeng, H. Lu, L. Zhang, M. Feng, A. Borji, Learning to promote saliency detectors, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[37] L. Zhang, J. Dai, H. Lu, Y. He, G. Wang, A bi-directional message passing model for salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[38] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Int. Conf. Comput. Vis., 2017.

[39] J. Su, J. Li, Y. Zhang, C. Xia, Y. Tian, Selectivity or invariance: Boundary-aware salient object detection, in: Int. Conf. Comput. Vis., 2019.

[40] T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[41] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Salient object detection with recurrent fully convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 41(7) 2018, 1734–1746.

[42] N. Liu, J. Han, M.-H. Yang, PiCANet: Pixel-wise contextual attention learning for accurate saliency detection, IEEE Trans. Image Process. 29 2020, 6438–6451.

[43] M. A. Islam, M. Kalash, N. D. Bruce, Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[44] N. Liu, J. Han, DHSNet: Deep hierarchical saliency network for salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2016.

[45] S. He, J. Jiao, X. Zhang, G. Han, R. W. Lau, Delving into salient object subitizing and detection, in: Int. Conf. Comput. Vis., 2017.

[46] Z. Li, C. Lang, Y. Chen, J. Liew, J. Feng, Deep reasoning with multi-scale context for salient object detection, arXiv preprint arXiv:1901.08362 (2019).

[47] S. Jia, N. D. Bruce, Richer and deeper supervision network for salient object detection, arXiv preprint arXiv:1901.02425 (2019).

[48] Y.-H. Wu, Y. Liu, L. Zhang, W. Gao, M.-M. Cheng, Regularized densely-connected pyramid network for salient instance segmentation, IEEE Trans. Im-

age Process. 30 2021, 3897–3907.

[49] Y. Liu, M.-M. Cheng, X.-Y. Zhang, G.-Y. Nie, M. Wang, DNA: Deeply supervised nonlinear aggregation for salient object detection, IEEE Trans. Cybernetics (2021).

[50] Z. Chen, R. Cong, Q. Xu, Q. Huang, DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection, IEEE Trans. Image Process. 30 2021, 7012–7024.

[51] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, S. Kwong, Dense attention fluid network for salient object detection in optical remote sensing images, IEEE Trans. Image Process. 30 2021, 1305–1317.

[52] R. Cong, Y. Zhang, L. Fang, J. Li, Y. Zhao, S. Kwong, Rrnet: Relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 60 2021, 1–11.

[53] C. Fang, H. Tian, D. Zhang, Q. Zhang, J. Han, J. Han, Densely nested top-down flows for salient object detection, arXiv preprint arXiv:2102.09133 (2021).

[54] D. Zhang, H. Tian, J. Han, Few-cost salient object detection with adversarial-paced learning, Adv. Neural Inform. Process. Syst. 33 2020, 12236–12247.

[55] D. Zhang, J. Han, Y. Zhang, D. Xu, Synthesizing supervision for learning deep saliency network without human annotation, IEEE Trans. Pattern Anal. Mach. Intell. 42(7) 2019, 1755–1769.

[56] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, Q. Tian, Label decoupling framework for salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[57] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, L. Yang, Interactive two-stream decoder for accurate and fast saliency detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[58] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, M.-M. Cheng, Lightweight salient object detection via hierarchical visual perception learning, IEEE Trans. Cybernetics 51(9) 2021, 4439–4449.

[59] M. Yang, K. Yu, C. Zhang, Z. Li, K. Yang, DenseASPP for semantic segmentation in street scenes, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[60] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, S. Yan, Depth matters: Influence of depth cues on visual saliency, in: Eur. Conf. Comput. Vis., 2012.

[61] J. R. Arridhana Ciptadi, Tucker Hermans, An in depth view of saliency, in: Brit. Mach. Vis. Conf., BMVA Press, 2013.

[62] K. Desingh, K. M. Krishna, D. Rajan, C. Jawahar, Depth really matters: Improving visual salient region detection with depth, in: Brit. Mach. Vis. Conf., 2013.

[63] X. Fan, Z. Liu, G. Sun, Salient region detection for stereoscopic images, in: International Conference on Digital Signal Processing, IEEE, 2014.

[64] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: a benchmark and algorithms, in: Eur. Conf. Comput. Vis., 2014.

[65] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in: International Conference on Internet Multimedia Computing and Service, 2014.

[66] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, N. Barnes, UC-Net: uncertainty inspired RGB-D saliency detection via conditional variational autoencoders, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[67] C. Li, R. Cong, Y. Piao, Q. Xu, C. C. Loy, RGB-D salient object detection with cross-modality modulation and selection, in: Eur. Conf. Comput. Vis., 2020.

[68] X. Zhao, L. Zhang, Y. Pang, H. Lu, L. Zhang, A single stream network for robust and real-time RGB-D salient object detection, in: Eur. Conf. Comput. Vis., 2020.

[69] Y. Piao, W. Ji, J. Li, M. Zhang, H. Lu, Depth-induced multi-scale recurrent attention network for saliency detection, in: Int. Conf. Comput. Vis., 2019.

[70] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, M.-M. Cheng, Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks, IEEE Trans. Neural Net. Learn. Syst. 32(5) 2020, 2075–2089.

[71] Y. Pang, L. Zhang, X. Zhao, H. Lu, Hierarchical dynamic filtering network for RGB-D salient object detection, in: Eur. Conf. Comput. Vis., 2020.

[72] X. Zhao, Y. Pang, L. Zhang, H. Lu, X. Ruan, Self-supervised representation learning for RGB-D salient object detection, arXiv preprint arXiv:2101.12482 (2021).

[73] S. Chen, Y. Fu, Progressively guided alternate refinement network for RGB-D salient object detection, in: Eur. Conf. Comput. Vis., 2020.

[74] N. Huang, Y. Liu, Q. Zhang, J. Han, Joint cross-modal and unimodal features for RGB-D salient object detection, IEEE Trans. Multimedia 23 2021, 2428–2441.

[75] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, RGBD salient object detection via deep fusion, IEEE Trans. Image Process. 26(5) 2017, 2274–2285.

[76] Q. Chen, Z. Liu, Y. Zhang, K. Fu, Q. Zhao, H. Du, Rgb-d salient object detection via 3d convolutional neural networks, in: AAAI, 2021.

[77] J. Han, H. Chen, N. Liu, C. Yan, X. Li, CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion, IEEE Trans. Cybernetics 48(11) 2017, 3171–3183.

[78] N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 2019, 55277–55284.

[79] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[80] N. Liu, N. Zhang, J. Han, Learning selective self-mutual attention for RGB-D saliency detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[81] N. Huang, Y. Yang, D. Zhang, Q. Zhang, J. Han, Employing bilinear fusion and saliency prior information for rgb-d salient object detection, IEEE Trans. Multimedia 24 2021, 1651–1664.

[82] WHO, Coronavirus disease (covid-19) outbreak situation (2020).
URL https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[83] V. Rajinikanth, N. Dey, A. N. J. Raj, A. E. Hassanien, K. C. Santosh, N. S. M. Raja, Harmony-search and otsu based system for coronavirus disease (covid-19) detection using lung ct scan images, arXiv (2020). arXiv:2004.03431.

[84] J. B. Roerdink, A. Meijster, The watershed transform: Definitions, algorithms and parallelization strategies, Fundamenta Informaticae 41(1,2) 2000, 187–228.

[85] T. Zhou, S. Canu, S. Ruan, An automatic covid-19 ct segmentation based on u-net with attention mechanism, arXiv preprint arXiv:2004.06673 (2020).

[86] H. B. Jenssen, Covid-19 ct segmentation dataset, http://medicalsegmentation.com/covid19/ (2020).

[87] 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.

[88] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Computation 1(4) 1989, 541–551.

[89] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Adv. Neural Inform. Process. Syst., 2012.

[90] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: Adv. Neural Inform. Process. Syst., 2015.

[91] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Int. Conf. Comput. Vis., 2017.

[92] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conf. Comput. Vis. Pattern Recog., 2015.

[93] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: IEEE Conf. Comput. Vis. Pattern Recog., 2015.

[94] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[95] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[96] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, IEEE Trans. Pattern Anal. Mach. Intell. 42(8) 2020, 2011–2023.

[97] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[98] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. H. Torr, Res2Net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal.

Mach. Intell. 43(2) 2021, 652–662.

[99] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Adv. Neural Inform. Process. Syst., 2017.

[100] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Eur. Conf. Comput. Vis., 2020.

[101] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, arXiv preprint arXiv:2010.04159 (2020).

[102] H. Wang, Y. Zhu, H. Adam, A. Yuille, L.-C. Chen, MaX-DeepLab: End-to-end panoptic segmentation with mask transformers, in: IEEE Conf. Comput. Vis. Pattern Recog., 2021.

[103] R. Liu, Z. Yuan, T. Liu, Z. Xiong, End-to-end lane shape prediction with transformers, in: IEEE Winter Conf. Appl. Comput. Vis., 2021.

[104] J. Hu, L. Cao, L. Yao, S. Zhang, Y. Wang, K. Li, F. Huang, R. Ji, L. Shao, ISTR: End-to-end instance segmentation with transformers, arXiv preprint arXiv:2105.00637 (2021).

[105] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Int. Conf. Learn. Represent., 2021.

[106] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Int. Conf. Mach. Learn., 2021.

[107] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[108] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, S. Yan, Tokens-to-token ViT: Training vision transformers from scratch on ImageNet, in: Int. Conf. Comput. Vis., 2021.

[109] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, CvT: Introducing convolutions to vision transformers, in: Int. Conf. Comput. Vis., 2021.

[110] X. Chu, Z. Tian, B. Zhang, X. Wang, X. Wei, H. Xia, C. Shen, Conditional positional encodings for vision transformers, arXiv preprint arXiv:2102.10882 (2021).

[111] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, S. J. Oh, Rethinking spatial dimensions of vision transformers, in: Int. Conf. Comput. Vis., 2021.

[112] 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: Int. Conf. Comput. Vis., 2021.

[113] 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: Int. Conf. Comput. Vis., 2021.

[114] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, C. Feichtenhofer, Multiscale vision transformers, in: Int. Conf. Comput. Vis., 2021.

[115] Q. Han, Z. Fan, Q. Dai, L. Sun, M.-M. Cheng, J. Liu, J. Wang, Demystifying local vision transformer: Sparse connectivity, weight sharing, and dynamic weight, arXiv preprint arXiv:2106.04263 (2021).

[116] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: Revisiting the design of spatial attention in vision transformers, in: Adv. Neural Inform. Process. Syst., 2021.

[117] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, Computational Visual Media 8(3) 2022, 415–424.

[118] Z.-H. Jiang, Q. Hou, L. Yuan, D. Zhou, Y. Shi, X. Jin, A. Wang, J. Feng, All tokens matter: Token labeling for training better vision transformers, Adv. Neural Inform. Process. Syst. 34 2021, 18590–18602.

[119] Y. Li, K. Zhang, J. Cao, R. Timofte, L. Van Gool, LocalViT: Bringing locality to vision transformers, arXiv preprint arXiv:2104.05707 (2021).

[120] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, W. Wu, Incorporating convolution designs into visual transformers, in: Int. Conf. Comput. Vis., 2021.

[121] Y. Liu, Y.-H. Wu, G. Sun, L. Zhang, A. Chhatkuli, L. Van Gool, Vision transformers with hierarchical attention, arXiv preprint arXiv:2106.03180 (2021).

[122] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (04 2017).

[123] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[124] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, ShuffleNet v2: Practical guidelines for efficient CNN architecture design, in: Eur. Conf. Comput. Vis., 2018.

[125] X. Zhang, X. Zhou, M. Lin, J. Sun, ShuffleNet: An extremely efficient convolutional neural network for mobile devices, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[126] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861 (2017).

[127] M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Int. Conf. Mach. Learn., 2019.

[128] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, Q. V. Le, MnasNet: Platform-aware neural architecture search for mobile, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[129] S. Mehta, M. Rastegari, L. Shapiro, H. Hajishirzi, ESPNetv2: A light-weight, power efficient, and general purpose convolutional neural network, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[130] M. Tan, R. Pang, Q. V. Le, EfficientDet: Scalable and efficient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[131] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, M.-M. Cheng, SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection, IEEE Trans. Image Process. 30 2021, 3804–3814.

[132] Y.-H. Wu, Y. Liu, X. Zhan, M.-M. Cheng, P2t: Pyramid pooling transformer for scene understanding, arXiv preprint arXiv:2106.12011 (2021).

[133] V. Mahadevan, N. Vasconcelos, Saliency-based discriminant tracking, in: IEEE Conf. Comput. Vis. Pattern Recog., 2009.

[134] Z. Ren, S. Gao, L.-T. Chia, I. W.-H. Tsang, Region-based saliency detection and its application in object recognition., IEEE Trans. Circ. Syst. Video Technol. 24(5) 2014, 769–779.

[135] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, X. Wu, Visual-textual joint relevance learning for tag-based social image search, IEEE Trans. Image Process. 22(1) 2013, 363–376.

[136] Y. Liu, Y.-H. Wu, P.-S. Wen, Y.-J. Shi, Y. Qiu, M.-M. Cheng, Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 44(3) 2022, 1415–1428.

[137] P.-T. Jiang, L.-H. Han, Q. Hou, M.-M. Cheng, Y. Wei, Online attention accumulation for weakly supervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[138] X. Li, F. Yang, H. Cheng, W. Liu, D. Shen, Contour knowledge transfer for salient object detection, in: Eur. Conf. Comput. Vis., 2018.

[139] X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, Suppress and balance: A simple gated network for salient object detection, in: Eur. Conf. Comput. Vis., 2020.

[140] Y. Pang, X. Zhao, L. Zhang, H. Lu, Multi-scale interactive network for salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2020.

[141] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, P.-M. Jodoin, Non-local deep features for salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[142] G. Li, Y. Xie, L. Lin, Y. Yu, Instance-level salient object segmentation, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[143] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, A. Borji, Detect globally, refine locally: A novel approach to saliency detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[144] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, M. Jagersand, BASNet: Boundary-aware salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[145] W. Wang, S. Zhao, J. Shen, S. C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: IEEE Conf. Comput. Vis. Pattern Recog.,

2019.

[146] Z. Wu, L. Su, Q. Huang, Stacked cross refinement network for edge-aware salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[147] X. Wang, H. Ma, X. Chen, S. You, Edge preserving and multi-scale contextual neural network for salient object detection, IEEE Trans. Image Process. 27(1) 2017, 121–134.

[148] M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[149] Y. Wang, X. Zhao, X. Hu, Y. Li, K. Huang, Focal boundary guided salient object detection, IEEE Trans. Image Process. 28(6) 2019, 2813–2824.

[150] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, CCNet: Criss-cross attention for semantic segmentation, in: Int. Conf. Comput. Vis., 2019.

[151] Z. Zhu, M. Xu, S. Bai, T. Huang, X. Bai, Asymmetric non-local neural networks for semantic segmentation, in: Int. Conf. Comput. Vis., 2019.

[152] F. Milletari, N. Navab, S.-A. Ahmadi, V-Net: Fully convolutional neural networks for volumetric medical image segmentation, in: Int. Conf. 3D Vis., 2016.

[153] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan, Learning to detect salient objects with image-level supervision, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[154] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2013.

[155] Y. Li, X. Hou, C. Koch, J. M. Rehg, A. L. Yuille, The secrets of salient object segmentation, in: IEEE Conf. Comput. Vis. Pattern Recog., 2014.

[156] P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, in: Int. Conf. Comput. Vis., 2017.

[157] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[158] Z. Chen, Q. Xu, R. Cong, Q. Huang, Global context-aware progressive aggregation network for salient object detection, in: AAAI, 2020.

[159] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, S. Yan, Highly efficient salient object detection with 100k parameters, in: Eur. Conf. Comput. Vis., 2020.

[160] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: Int. Conf. Comput. Vis., 2017.

[161] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in: Int. Joint Conf. Artif. Intell., 2018.

[162] 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: Adv. Neural Inform. Process. Syst., 2019.

[163] 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.

[164] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Int. Conf. Learn. Represent., 2015.

[165] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[166] R. Margolin, L. Zelnik-Manor, A. Tal, How to evaluate foreground maps?, in: IEEE Conf. Comput. Vis. Pattern Recog., 2014.

[167] J. Zhao, Y. Zhao, J. Li, X. Chen, Is depth really necessary for salient object detection?, in: ACM Int. Conf. Multimedia, 2020.

[168] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88(2) 2010, 303–338.

[169] Y.-H. Wu, Y. Liu, J. Xu, J.-W. Bian, Y. Gu, M.-M. Cheng, MobileSal: Extremely efficient RGB-D salient object detection, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[170] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in: Int. Conf. Mach. Learn., 2015.

[171] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Trans. Pattern Anal. Mach. Intell. 41(7) 2018, 1531–1544.

[172] H. Chen, Y. Li, Three-stream attention-aware network for RGB-D salient object

detection, IEEE Trans. Image Process. 28(6) 2019, 2825–2835.

[173] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: Int. Conf. Image Process., IEEE, 2014.

[174] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).

[175] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Int. Conf. Mach. Learn., 2010.

[176] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13(4) 2004, 600–612.

[177] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, C. Hou, Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion, IEEE Signal Process. Let. 23(6) 2016, 819–823.

[178] H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[179] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, D.-P. Fan, Bilateral attention network for rgb-d salient object detection, IEEE Trans. Image Process. 30 2021, 1949–1961.

[180] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Int. Conf. Learn. Represent., 2015.

[181] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: IEEE Conf. Comput. Vis. Pattern Recog., 2012.

[182] C. Zhu, G. Li, A three-pathway psychobiological framework of salient object detection using stereoscopic technology, in: Int. Conf. Comput. Vis. Worksh., 2017.

[183] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2009.

[184] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, M.-M. Cheng, Spatial information guided convolution for real-time RGBD semantic segmentation, IEEE Trans. Image Process. 30 2021, 2313–2324.

[185] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, G. Zeng, Bi-directional

cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation, in: Eur. Conf. Comput. Vis., 2020.

[186] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: Eur. Conf. Comput. Vis., 2012.

[187] Y. Wang, H. Hou, W. Wang, W. Wang, Combination of ct and rt-pcr in the screening or diagnosis of covid-19, Journal of Global Health 10(1) (2020).

[188] J. Zhang, Y. Xie, Y. Li, C. Shen, Y. Xia, Covid-19 screening on chest x-ray images using deep learning based anomaly detection (2020). arXiv:2003.12338.

[189] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases, Radiology (2020).

[190] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, W. Ji, Sensitivity of chest ct for covid-19: comparison to rt-pcr, Radiology (2020).

[191] S. Inui, A. Fujikawa, M. Jitsu, N. Kunishima, S. Watanabe, Y. Suzuki, S. Umeda, Y. Uwabe, Chest ct findings in cases from the cruise ship “diamond princess” with coronavirus disease 2019 (covid-19), Radiology: Cardiothoracic Imaging 2(2) 2020, e200110.

[192] Z. Huang, S. Zhao, Z. Li, W. Chen, L. Zhao, L. Deng, B. Song, The battle against coronavirus disease 2019 (covid-19): Emergency management and infection control in a radiology department, Journal of the American College of Radiology (2020).

[193] J. P. Cohen, P. Morrison, L. Dao, Covid-19 image data collection, arXiv 2003.11597 (2020).
URL https://github.com/ieee8023/covid-chestxray-dataset

[194] J. Zhao, Y. Zhang, X. He, P. Xie, Covid-ct-dataset: a ct scan dataset about covid-19, arXiv preprint arXiv:2003.13865 (2020).

[195] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, J. Xia, Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct, Radiology 2020, 200905.

[196] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu,

Y. Yang, Z. A. Fayad, A. Jacobi, K. Li, S. Li, H. Shan, Ct imaging features of 2019 novel coronavirus (2019-ncov), Radiology 295(1) 2020, 202–207.

[197] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, S. Li, H. Shan, A. Jacobi, M. Chung, Chest ct findings in coronavirus disease-19 (covid-19): Relationship to duration of infection, Radiology 2020, 200463.

[198] Y. Wang, C. Dong, Y. Hu, C. Li, Q. Ren, X. Zhang, H. Shi, M. Zhou, Temporal changes of ct findings in 90 patients with covid-19 pneumonia: A longitudinal study, Radiology 2020, 200843.

[199] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, C. Zheng, Radiological findings from 81 patients with covid-19 pneumonia in wuhan, china: a descriptive study, The Lancet Infect Disease 20(4) 2020, 425–434.

[200] Y. Xiong, D. Sun, Y. Liu, Y. Fan, L. Zhao, X. Li, W. Zhu, Clinical and high-resolution ct features of the covid-19 infection: comparison of the initial and follow-up changes, Investigative radiology (2020).

[201] A. Leung, R. Miller, N. Müller, Parenchymal opacification in chronic infiltrative lung diseases: Ct-pathologic correlation., Radiology 188(1) 1993, 209–214.

[202] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Med. Image. Comput. Comput. Assist. Interv., Springer, 2015.

[203] L. 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 Trans. Pattern Anal. Mach. Intell. (2017).

[204] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[205] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[206] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in: Artificial intelligence and statistics, 2015.

[207] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural net-

works for volumetric medical image segmentation, in: Int. Conf. 3D Vis., IEEE, 2016.

[208] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43(2) 2021, 652–662.

[209] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopezpaz, mixup: Beyond empirical risk minimization, in: Int. Conf. Learn. Represent., 2018.

[210] 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: Int. Conf. Comput. Vis., 2017.

[211] N. Sajid, Covid-19 patients lungs x ray images 10000, https://www.kaggle.com/nabeelsajid917/covid-19-x-ray-10000-images, 记录日期 4 月 10 日, 2020 (2020).

[212] F. Shi, L. Xia, F. Shan, D. Wu, Y. Wei, H. Yuan, H. Jiang, Y. Gao, H. Sui, D. Shen, Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification, arXiv preprint arXiv:2003.09860 (2020).

[213] F. Shan+, Y. Gao+, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, D. Shen, Y. Shi, Lung infection quantification of covid-19 in ct images with deep learning, arXiv preprint arXiv:2003.04655 (2020).

[214] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F. Li, Imagenet: A large-scale hierarchical image database, in: IEEE Conf. Comput. Vis. Pattern Recog., 2009.

[215] J. Won, S. Lee, M. Park, T. Kim, M. Park, B. Choi, D. Kim, H. Chang, V. Kim, C. Lee, Development of a laboratory-safe and low-cost detection protocol for sars-cov-2 of the coronavirus disease 2019 (covid-19)., Experimental neurobiology (2020).

[216] R. Hu, A. Singh, Transformer is all you need: Multimodal multitask learning with a unified transformer, arXiv preprint arXiv:2102.10772 (2021).

[217] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, SwinIR: Image restoration using Swin Transformer, in: Int. Conf. Comput. Vis., 2021.

[218] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, et al., Findings of the 2014 workshop on

statistical machine translation, in: The Workshop on Statistical Machine Translation, 2014.

[219] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, M. Douze, LeViT: a vision transformer in ConvNet's clothing for faster inference, in: Int. Conf. Comput. Vis., 2021.

[220] K. Grauman, T. Darrell, The pyramid match kernel: Discriminative classification with sets of image features, in: Int. Conf. Comput. Vis., Vol. 2, IEEE, 2005.

[221] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: IEEE Conf. Comput. Vis. Pattern Recog., Vol. 2, IEEE, 2006.

[222] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37(9) 2015, 1904–1916.

[223] J.-R. Chang, Y.-S. Chen, Pyramid stereo matching network, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[224] J.-J. Liu, Q. Hou, Z.-A. Liu, M.-M. Cheng, Poolnet+: Exploring the potential of pooling for salient object detection, IEEE Trans. Pattern Anal. Mach. Intell. (2022).

[225] D. Park, K. Kim, S. Young Chun, Efficient module based single image super resolution for multiple problems, in: IEEE Conf. Comput. Vis. Pattern Recog. Worksh., 2018.

[226] H. Zhang, V. Sindagi, V. M. Patel, Image de-raining using a conditional generative adversarial network, IEEE Trans. Circ. Syst. Video Technol. 30(11) 2019, 3943–3956.

[227] H. Zhang, V. M. Patel, Densely connected pyramid dehazing network, in: IEEE Conf. Comput. Vis. Pattern Recog., 2018.

[228] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[229] Y. Dong, J.-B. Cordonnier, A. Loukas, Attention is not all you need: Pure attention loses rank doubly exponentially with depth, arXiv preprint arXiv:2103.03404 (2021).

[230] 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: Int. Conf. Comput. Vis., 2019.

[231] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), arXiv preprint arXiv:1606.08415 (2016).

[232] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, J. Gao, Multi-scale vision Longformer: A new vision transformer for high-resolution image encoding, in: Int. Conf. Comput. Vis., 2021.

[233] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: Int. Conf. Learn. Represent., 2019.

[234] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic feature pyramid networks, in: IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[235] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ADE20K dataset, in: IEEE Conf. Comput. Vis. Pattern Recog., 2017.

[236] 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: Eur. Conf. Comput. Vis., 2014.

[237] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Int. Conf. Comput. Vis., 2017.

[238] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Int. Conf. Artif. Intell. Stat., 2010.

[239] M. Contributors, MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark, https://github.com/open-mmlab/mmsegmentation (2020).

[240] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al., MMDetection: Open MMLab detection toolbox and benchmark, arXiv preprint arXiv:1906.07155 (2019).

[241] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, IEEE Trans. Pattern Anal. Mach. Intell. 42(2) 2020, 386–397.

个人简历

吴宇寰，出生于 1996 年 10 月 13 日，在 2018 年毕业于西安电子科技大学大学电子信息工程专业并获得学士学位。于 2018 年至今在南开大学直接攻读博士研究生。

研究生期间主要的发表论文：

注：“*” 表示共同贡献。

1. **Yu-Huan Wu***, Yun Liu*, Xin Zhan, and Ming-Ming Cheng. P2T: Pyramid Pooling Transformer for Scene Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2022. (SCI 一区，CCF-A 类期刊，影响因子 24.314)
2. Yun Liu*, **Yu-Huan Wu***, Peisong Wen, Yujun Shi, Yu Qiu, and Ming-Ming Cheng. Leveraging Instance-, Image-and Dataset-Level Information for Weakly Supervised Instance Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2022. (SCI 一区，CCF-A 类期刊，影响因子 24.314)
3. **Yu-Huan Wu**, Yun Liu, Jun Xu, Jia-Wang Bian, Yu-Chao Gu, and Ming-Ming Cheng. MobileSal: Extremely Efficient RGB-D Salient Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021. (SCI 一区，CCF-A 类期刊，影响因子 24.314)
4. **Yu-Huan Wu**, Shang-Hua 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 (TIP), 2021. (SCI 一区，CCF-A 类期刊，影响因子 11.041)
5. **Yu-Huan Wu***, Yun Liu*, Le Zhang, Ming-Ming Cheng, and Bo Ren. EDN: Salient object detection via extremely-downsampled network. IEEE Transactions on Image Processing (TIP), 2022. (SCI 一区，CCF-A 类期刊，影响因子 11.041)

6. **Yu-Huan Wu**, Yun Liu, Le Zhang, Wang Gao, and Ming-Ming Cheng. Regularized Densely-connected Pyramid Network for Salient Instance Segmentation. IEEE Transactions on Image Processing (TIP), 2021. (SCI 一区，CCF-A 类期刊，影响因子 11.041)
7. Yun Liu*, **Yu-Huan Wu***, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. (CCF-A 类会议，口头报告)
8. 刘云, 陆承泽, 李仕杰, 张乐, **吴宇寰**, 程明明. 基于高效的多尺度特征提取的轻量级语义分割. 计算机学报, 2022. (中国卓越期刊，影响因子 5.402)

研究生期间主要的发表专利：

1. 吴宇寰，刘云，程明明，基于正则化稠密连接特征金字塔的实例显著性检测方法，专利号 ZL202010328784.6，授权日 2022 年 4 月 15 日；
2. 程明明，吴宇寰，刘云，基于隐含深度信息恢复的图像显著性物体检测方法及系统，专利号 ZL202011500709.X，授权日 2022 年 4 月 12 日；
3. 程明明，吴宇寰，刘云，基于极致下采样的图像显著性物体检测方法及系统，专利号 ZL202011500724.4，授权日 2022 年 9 月 16 日。

研究生期间主要获得的荣誉奖励：

1. 南开十杰，2022 年
2. 国家奖学金，2022 年