中图分类号: 学校代码: 10055
UDC: 密级: 公开

南開大學

硕 士 学 位 论 文

基于对比度先验和流动金字塔的 RGBD 显著性检测方法

Contrast Prior and Fluid Pyramid Integration for RGBD Salient Object Detection

论文作者 曹洋 指导教师 程明明教授
申请学位 工学硕士 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 杨巨峰 评 阅 人 杨巨峰、卢少平

南开大学研究生院

二〇二〇年四月

南开大学学位论文使用授权书

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

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

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

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

3. 非公开学位论文在解密后的使用权同公开论文。

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

本人签署本授权书一份（此授权书为论文中一页），交图书馆留存。

学位论文作者暨授权人 (亲笔) 签字：曹洋

2020 年 5 月 30 日

南开大学研究生学位论文作者信息

论 文 题 目	基于对比度先验和流动金字塔的 RGBD 显著性检测方法				
姓 名	曹洋	学号	2120170448	答辩日期	2020 年 5 月 21 日
论 文 类 别	博士 ☐ 学历硕士 ☑ 专业学位硕士 ☐ 同等学力硕士 ☐ 划☑选择				
学院（单位）	计算机学院		学科/专业 (专业学位) 名称		计算机科学与技术
联 系 电 话	17602216660		电子邮箱	2120170448@mail.nankai.edu.cn	
通讯地址 (邮编)：天津市津南区同砚路 38 号南开大学信息栋楼 425(300350)					
非公开论文编号			备注		

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

南开大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下进行研究工作所取得的研究成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：  2020 年 5 月 30 日

非公开学位论文标注说明

(本页表中填写内容须打印)

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

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

南开大学学位评定委员会办公室盖章 (有效)

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

摘要

显著性检测致力于检测出场景中最具备显著性的区域，在计算机视觉的相关研究方向和工业界中有着广泛的应用。近些年，受益于高性能显卡设备和大体量数据集的出现，基于深度网络模型的显著性检测方法的效果卓越，但在面临例如低对比度场景和背景复杂场景等困难场景时，针对于 RGB 信息搭建的模型难以取得好的检测效果。随着深度传感器在更多终端上的部署，获取深度信息越来越便利，RGB 图片的困难场景在深度层面却可能易于模型处理，因此在显著性检测任务中引入深度信息的思路得到学者们的重视。现有的方法主要在模型前段、中段和末段对深度信息和 RGB 信息进行综合处理，对显著性检测效果进行了优化，但仍存在以下亟待解决的问题:

1. 缺乏高质量深度图。从深度传感器捕获的深度图通常具备噪声，不易于模型处理。同时没有像 ImageNet 这样的大规模深度图数据集，所以缺乏能被广泛应用的预训练模型 (例如在 ImageNet 上的 VggNet 或 ResNet 预训练模型)。

2. 有待优化的多尺度跨模态融合方法。RGB 图和深度图中具备两种不同模态的信息，二者具备较大的差异。比如，和其他颜色相比，绿色和“植物”类别相关性更大。然而，各种类别在深度层面的分布却不存在类似的相关性，因此在进行如通道拼接之类的简单融合时，两种模态的差异很可能导致信息的不兼容问题。

基于对以上问题的思考，本文提出了从深度图中提取深度对比度先验的方法，并通过残差形式的跨模态融合方法将其用于优化 RGB 分支特征，在基础网络的不同模块后得到了多个尺度的跨模态特征，而后本文设计了流动金字塔融合方法对多尺度跨模态特征进行了更加充分的融合。在实验部分，本文在五个公开数据集基于四种评价指标和先前的九种 RGBD 显著性检测方法进行了对比，发现本方法领先于先前方法。最后，本文进行了消融实验，以进一步讨论本方法中各个模块的重要性。

关键词： 显著性检测；深度学习；多尺度跨模态特征融合；深度信息

Abstract

Saliency detection, which has a wide range of applications, is one research direction of computer vision. The goal is finding the most salient areas in the scene. Recently, thanks to the emergence of high-performance graphics card and large datasets, deep learning based methods have good performances. But they still meet troubles in difficult scenes which have low contrast or complex background. It is difficult to find the salient regions completely with models built for RGB scene. Because depth sensor is applied on more terminals. It is more convenient to obtain depth maps. And difficult scenes of RGB images may be easy to deal at the depth level. Therefore, scholars pay attention to the idea of introducing depth information into the saliency detection task. Existing methods mainly fuse the depth information and RGB information in the early, middle or late stage of the methods. The performances are improved. However, there are still the following troubles:

1) Shortage of high-quality depth maps. Depth maps captured from depth sensors are much noisier and textureless than RGB images, posting a challenge for the depth feature extraction. We lack well pretrained backbone networks for extracting powerful features from depth maps, as an ImageNet like large scale depth maps dataset is unavailable.

2) Suboptimal multi-scale cross-modal fusion. The two modalities, i.e., depth and RGB, have very different properties, making an effective multi-scale fusion of both modalities difficult. For instance, compared with the rest colors, ‘green’ color has a much stronger correlation with the ‘plants’ category. However, none depth value has such a correlation. The inherent difference between the two modalities may cause incompatibility problems when simple fusion strategies such as linear combination or concatenation are employed.

Based on the consideration of the above problems, the paper proposes a method to extract contrast prior from depth maps. And the method uses cross-modal fusion strategy in residual connection way to enhance RGB features by contrast prior. The

multi-scale cross-modal features are obtained in the end of five blocks of basic networks. Then the method use a novel fluid pyramid integration, which can make better use of multi-scale cross-modal features. Comprehensive experiments on 5 challenging benchmark datasets demonstrate the superiority of the architecture CPFP over 9 state-of-the-art alternative methods. Finally, ablation experiments are carried out to further discuss the importance of modules in this method.

Key Words: saliency detection; deep learning; multi-scale cross-modal features fusion; depth information

目录

第一章 引言

人工智能是一个极具潜力的研究领域，主要研究如何使计算机像人类智能一样思考和活动，在未来极有可能帮助便利化人类生活，甚至成为社会生产的重要推动力和人们健康生活的助力，图 1.1展示了通过结合热成像摄像机和人工智能算法快速检测高温人员的示例。

图 1.1 结合热成像摄像机和 AI 技术进行体温检测 (图片来源web)

对于人类智能而言，视觉是重要的感知窗口，关系到信息获取和理解，和人类的个人行为以及社会行为息息相关。与之相应的，如何让人工智能高效地捕获视觉信息和处理视觉信息也尤为重要。计算机视觉领域[1] 的研究致力于研究探索相关的问题，例如物体检测[2]、显著性检测[3]、实例分割[4]、边缘检测[5]、行人姿态估计[6]、图像去噪[7] 和图像超分辨[8] 等等。

处理视觉任务的传统方法，通常是针对不同任务的不同特点，基于简明准确的数学理论进行建模，例如对于显著性检测而言，早期的方法[9] 和[10]，前者通过在频域对数域中对图片数据进行建模，然后计算光谱残差来得到图片中的显著性区域，后者将像素点的显著性定义为像素级别的对比度，通过引入全局对比度来帮助计算显著性值。

近些年，随着大体量数据集 ImageNet[11] 的出现, 该数据集为 1400 多万张图片提供标注，极大的扩张了在解决问题的过程中可以凭借的数据量，为深度学

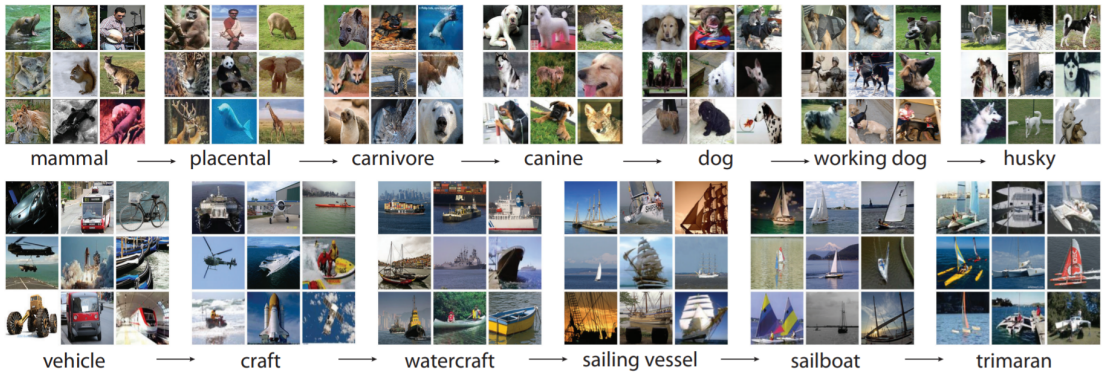

图 1.2 ImageNet 的类别划分示例 (图片来源[11])

习在计算机视觉任务中的应用提供了数据基础。ImageNet 中的图片分类模拟真实世界中物种的树状分类，由树根节点到叶子节点层层细分，图 1.2是两个由树根到树叶的分类示例，第一行是哺乳动物细化到哈士奇的类别划分，第二行是交通工具细化到三体帆船的类别划分。这种相对客观的分类方法为模型在客观世界中的应用提供相对合理的类别分布。

随着相关硬件设备 (尤其是高性能显卡) 的发展，高效的矩阵运算得以实现，因此基于矩阵级别计算的深度学习的硬件条件得以满足，再加上上文所讲的大体量数据集的出现，越来越多的计算机视觉任务都开始采用深度学习的方法进行处理。经典的网络模型相继出现，例如多伦多大学的 Hinton 研究组于 2012 年提出的 AlexNet[12]，该模型提出了 Relu 作为非线性激活层。为了一定程度地避免模型过拟合，提出了 Dropout。AlexNet 在多 GPU 上进行训练，相较于 CPU 训练，训练速度得到极大提升，同时受益于深度模型在大体量数据集上的适用性，以及相较于传统方法具有更大参数量，来帮助模型在更大的函数空间内依据训练数据进行拟合，也就是说具备了更强的学习能力。AlexNet 在 2012 年的 ImageNet 竞赛上夺得第一名。自此，深度学习受到计算机视觉领域内众多研究人员的广泛关注。

2015 年，牛津大学的 Visual Geometry Group 联合 Google Deepmind 提出了另一个经典的网络模型 VggNet[13]，相较于 AlexNet，VggNet 进一步增加了网络深度，并且论证了相较于大卷积核 (7×7,11×11 和 5×5 等)，连续使用 3×3 卷积核可以在保持同等感受野的同时具备更少的参数量，因而修改了模型通用卷积核大小。该模型具有更深的结构和更多的参数。受益于合理的卷积设置和更深的网络结构，VggNet 比 AlexNet 取得了更好的效果。

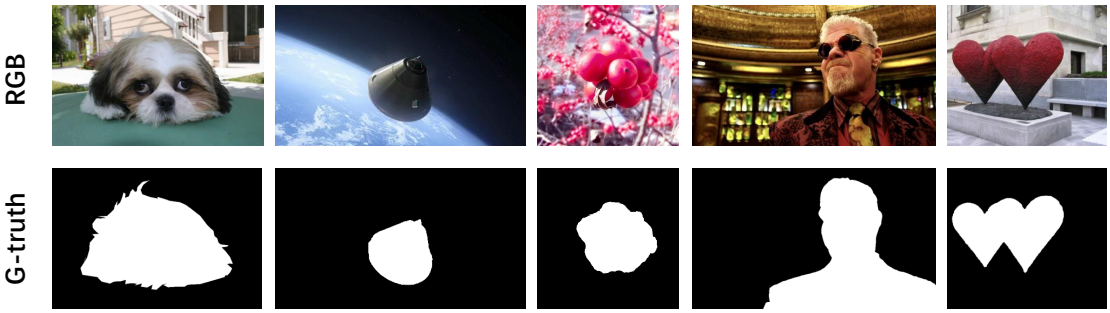

图 1.3 原图与显著性检测人工标注。首行是原图，次行是显著性检测人工标注。

VggNet 在文章中[13] 证明，随着网络的加深，模型的学习效果越来越好。但在实践中，伴随着网络进一步加深，其在训练过程中很可能会出现梯度消失，即梯度趋近于零，导致训练效果下降。微软亚洲研究院的研究组在 2016 年提出了 ResNet[14]，该模型在基本的卷积串联模式中引入残差连接，来减弱梯度消失带来的负面影响。残差连接为梯度提供了更多可能的传播路径，有效地避免了梯度消失，因此进一步提供了加深网络的可行性。ResNet 相较于其他结构，有效地避免了梯度消失，以致于顺利加深网络结构，优化网络的学习能力，ResNet 最终取得了 2015 的 ImageNet detection、ImageNet localization、ImageNet classification、COCO segmentation 和 COCO detection 的第一名, 证明了残差连接的优越性。

在通过上文了解了深度学习经典模型的发展过程后，本文继续将介绍计算机视觉中的任务：显著性物体检测。在这个任务中，显著性区域的定义是：在人关注该场景时，首先会注意到的区域。进一步地，显著性检测旨在区分场景中最具视觉显著性的区域。如图 1.3所示：在第一行的原图场景中，当人们看向该场景时，首先最容易注意到，即最显著的区域，是第二行的人工标注区域。显著目标检测具有广泛的应用，包括视频/图像分割[15, 16]、视觉跟踪[17]、前景图评估[18]、图像检索[19]、内容感知[20] 和弱监督语义分割[21] 等。

显著性检测任务应用广泛，可以用在分割任务和检测任务上。传统的显著性检测方法通常是对于图片用数学方法进行建模，比如[10] 和[9]，前者引入了全局对比度来辅助计算显著性值，后者在频域对数域中对图片信息进行建模。此类方法的优势是可解释性强，理论框架严密，计算量小，实现算法后运行速度快。但受限于有限的模型复杂度和参数量，效果有进一步的提升空间。

在深度学习发展后，学者们将其引入显著性检测领域，并针对于特定的任务特点进一步设计和改善，例如[3] 在深度网络中引入了深浅层间连接。在深度网络中，深层信息通常是维度较高的抽象信息，如位置和语义。浅层信息通常

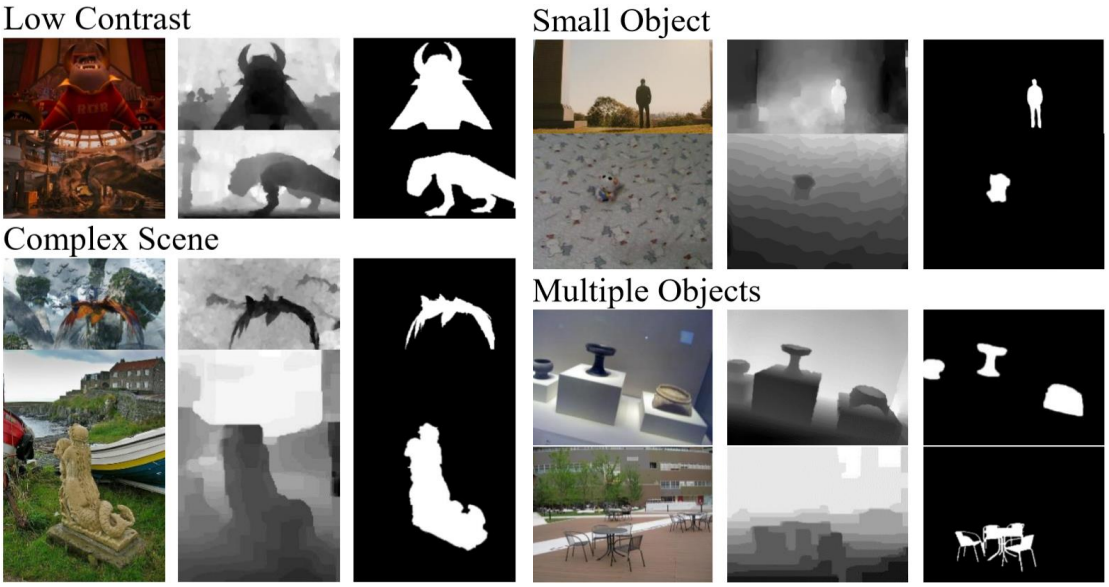

图 1.4 较难检测的困难场景。

是维度较低的具体信息，如边缘和纹理。[3] 通过引入的深浅层连接，进一步地将二者结合，同时通过旁侧监督去指导结合抽象信息和具体信息，改善了检测效果。[22] 发现显著性检测通常容易忽视边缘的地方，而边缘的细节会较大程度地影响检测的视觉效果，于是该工作在网络中加入对于物体边缘的监督，以此来指导网络重视边缘的学习，最终提升了检测结果的边缘效果。

随着学者们的努力，显著性检测领域的工作在不断发展，效果也在进步。但存在一些困难的场景，如图 1.4，图中所示小物体场景，含有多个物体的场景，低对比度场景和复杂场景，很难通过在 RGB 场景建模，来将显著性区域完整地检测出来。

近些年，随着深度图采集技术的普及，很多终端如 iphone X 和 Kinect 都可以较为方便地采集深度图。图 1.4所示，对于 RGB 场景下的一些困难场景，如前文所述背景繁杂场景和前背景对比度低场景，在深度场景下却存在较为明显的差别，所以研究如何引入深度信息，如何结合深度信息和 RGB 信息，是本文的主要研究内容。从信息融合的角度来看，现有的 RGBD 场景下的检测方法主要有如图 1.5所示的三类。

第一类[23] 如图中 (a) 所示，该类方法在较早阶段就融合了深度信息和 RGB 信息。具体地说，该方法将 RGB 图和深度图在通道维度进行拼接。而后一起送入模型中进行学习，模型对混合的信息进行映射后输出最终的预测结果。

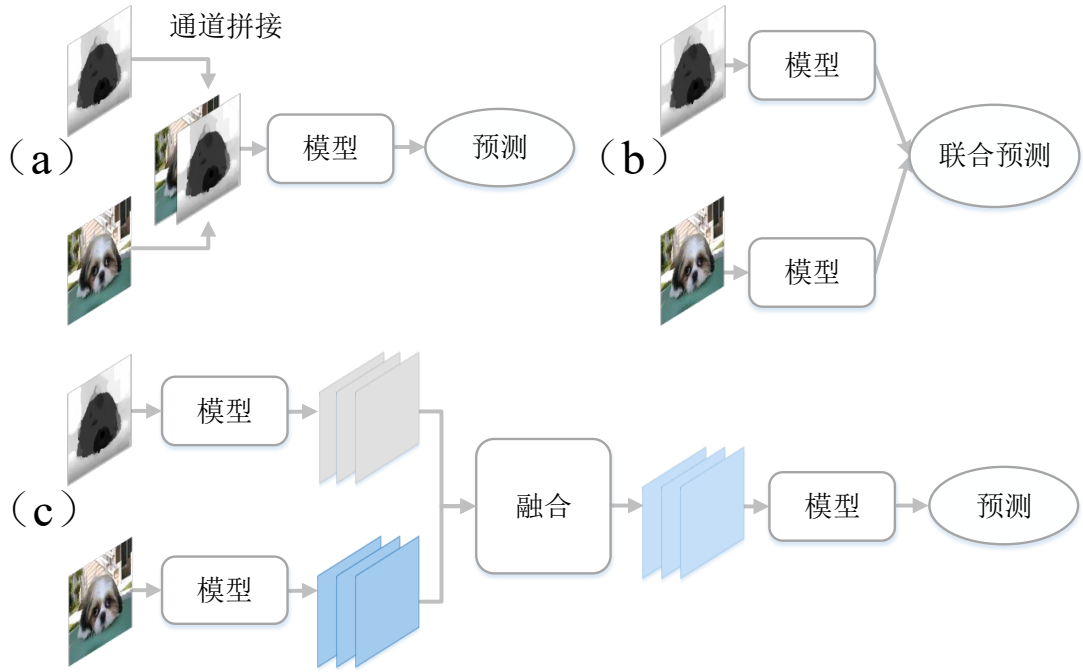

图 1.5 三类 RGBD 显著性检测模型。

第二类如图中 (b) 所示，该类模型[24] 分别用两个模型对 RGB 图和深度图进行预测，再根据预测结果进行联合预测，得到最终的输出。更具体地说，在通过模型得到深度图和 RGB 图像的预测结果后，对两个结果进行后处理以得到最终的结果。例如，[25] 结合深度对比以及深度加权的彩色特征对比来衡量区域的显著性值。[26] 通过计算彩色空间和深度空间视觉显著性得到最终的检测结果。此外，[27] 采用了非线性支持向量机来融合关于深度图和 RGB 图的预测结果。

第三类如图中 (c) 所示，通过模型抽取出深度图和 RGB 图上的特征，并对两者特征进行融合。例如，[28] 设计了显著性特征来捕捉角度方向上的扩散。[29] 提出的方法去捕捉背景区域和低维度的显著性信息。

现有的 RGBD 显著性检测方法存在以下亟待解决的问题:

1) 缺少高质量的深度图。从深度传感器捕获的深度图比 RGB 图像噪声大且无纹理，这对深度特征的提取提出了挑战。同时由于缺乏像 ImageNet 这样的大规模深度图数据集，所以没有强有力的基本网络的预训练模型 (如在 ImageNet 上预训练的 VggNet 或 ResNet 等)。

2) 有待优化的多尺度跨模态融合方法。深度图和 RGB 图中两种不同模态的信息具备较大的差异，使得高效地融合多尺度跨模态信息面临困难。例如，和其他颜色相比，绿色和“植物”类别相关性更大，因为在分布层面，大部分植物

都是绿色的。然而，各种类别在深度层面的分布却不存在此类的相关性，在进行如拼接之类的简单融合时，两种模态的差异很可能导致信息的不兼容问题。

与在 ImageNet 上预训练后提取深度信息并将其与 RGB 信息融合的思路不同，本文从深度图上提取出深度对比度先验，然后用深度对比度先验去增强 RGB 分支的特征，从而得到较好的结果。详细地讲，通过设计深度对比度增强分支，本文综合使用深度图分支特征的先验和 RGB 分支特征，同时针对深度图的特点，本文设计了深度分支的适用网络结构。

如何合理地融合多尺度跨模态特征是处理 RGBD 场景下的显著性检测任务的重点内容。本文考虑到在融合多尺度跨模态特征时出现的信息兼容性问题，在传统金字塔结构上引入了更加丰富的层间连接，为金字塔的各个融合节点提供更加丰富的来自不同尺度的跨模态信息来源，继而通过网络训练来在多尺度层面更好地融合跨模态信息。

在接下来的章节安排中，为帮助读者了解相关的背景，本文在章节 第二章中介绍了深度学习的背景知识和相关的显著性检测工作。在章节 第三章详细介绍了本方法的细节，包括基础网络、特征增强模块和流动金字塔融合方法。在章节 第四章中，本文在五个数据集上，基于四种评价指标和九种先前的 RGBD 显著性检测方法进行了对比实验，此外，本文对各个模块进行了消融实验以进一步探讨其对整体方法的重要性。在章节 第五章中，本文对全文进行了全面总结，同时基于完成本文过程中的思考，对未来 RGBD 显著性检测领域中有潜力的探索方向进行了讨论。

第二章 相关背景

本文基于深度学习，设计了结合深度对比度先验和流动金字塔的深度模型来完成 RGBD 场景下的显著性检测。为给后续介绍本方法做铺垫，本章节将介绍与本方法相关的深度学习和显著性检测背景知识。首先是介绍了深度学习的相关背景知识，包括基础网络结构、参数初始化算法、损失函数和优化算法，接着本文介绍了最近的 RGB 场景和 RGBD 场景中显著性检测领域的相关工作。

第一节 深度网络

在本节，本文首先对深度学习的相关基础知识进行介绍，其中具体包括网络结构、损失函数和学习策略。

2.1.1 基础网络结构

首先本小节介绍卷积神经网络。在传统神经网络中，每个神经元接受来自上层所有神经元的输出，因此参数量会随着神经网络层数的增加而以指数速度上升。受限于硬件条件 (如显存限制)，这种过于稠密的连接会对网络的深度造成严格的限制，进而影响模型的学习能力。对于图片数据，图片的局部信息大都具有旋转不变性、平移不变性等性质，具备参数共享处理的理论前提。因此卷积神经网络基于卷积的思想，放弃了传统神经网络中不必要的稠密连接，通过滑动的卷积核对图片中的局部信息进行特征提取。

卷积神经网络具有两个特点: 1. 局部连接。卷积神经网络通过滑动卷积核对输入数据进行特征提取，卷积核每次运算都仅仅计算感受野内的像素点，因此更利于提取局部信息，并符合信息的平移不变性等性质。2. 参数共享。在显存消耗方面，对应模型需要为一次映射存储的可学习参数只有该卷积核，相较于传统神经网络的稠密运算，这极大地减小了参数量，为进一步加深网络提供了可能性。

卷积神经网络中有代表性的是 Hinton 研究组提出的 AlexNet[12]。如图 2.1所示，该网络模型的主体部分由卷积层构成，受益于卷积层的参数共享性质，每层所需存储的参数量相较于传统神经网络大大减小，因此 AlexNet 有着较深的

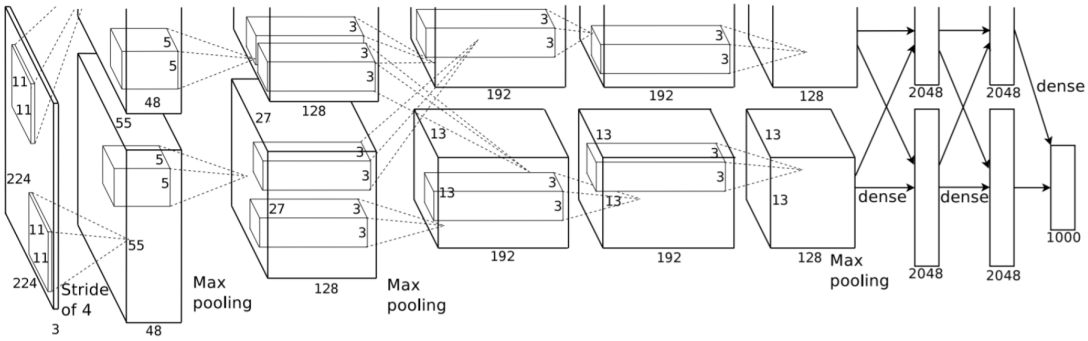

图 2.1 Alexnet 的结构示意图 (图片来源[12])

网络结构，具备较强的学习能力。

卷积神经网络通过采用卷积层为增大网络深度提供了可行性，随着进一步加深网络，网络模型在训练过程中很可能会存在梯度消失的现象，导致训练减慢甚至效果下降。其背后的理论原因是由梯度回传的特点导致，梯度的反向传播遵循链式法则，在计算时要逐层累乘，若训练过程中出现某层梯度小于 1，经过逐层累乘后会越来越小，造成梯度消失。网络深度越大，在反向传播过程中因累乘而造成梯度消失的可能性越大。因此梯度消失现象的存在限制了所设计网络的深度，进而限制了模型的表现效果。

为应对梯度消失现象，基于残差网络结构的 ResNet[14] 在 2016 年被提出。该工作的核心思想是通过残差连接来降低梯度消失对网络拟合的影响。如图 2.2所示，x 是此模块的输入，$F(x)$ 表示 WeightLayer-Relu-WeightLayer 的串联结构所表示的模型映射，残差连接即是在得到 $F(x)$ 再加上其输入 x，最终输出是 $F(x)+x$。这样做的优势在于，反向传播时，假如本模块内的函数 ($F(x)$) 计算所得梯度过小，又其他层传递过来的梯度依然可以由恒等映射分支继续传递，有效地避免了梯度消失。从梯度反传的角度理解，残差连接为梯度提供了更多的有效传播路径。

ResNet 的残差连接有效地减弱了梯度消失带来的负面影响，因此该网络的深度得到了进一步增大，和上文介绍的 AlexNet(图 2.1，8 层) 相比，ResNet 的深度得到了成倍增加，效果因此进一步提升。

2.1.2 模型初始化方法

在模型开始训练前，首先要对其进行初始化，赋予其初始参数。在很多任务场景中，都会对模型进行预训练，譬如在显著性检测、物体检测和分割等任

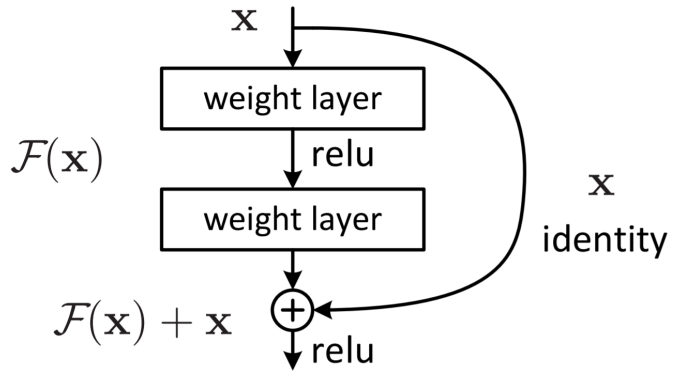

图 2.2 残差连接示意图 (图片来源[14])。)

务中，大部分方法会使用对应基础网络结构（如 ResNet、VggNet 等）在分类任务上训得的预训练模型做初始化参数。但也有些任务如超分辨、图像生成等任务较少采用预训练模型，而采用重新初始化开始训练。本小节将介绍较有代表性的 Xavier 算法[30] 和 He 算法[31]。

在网络训练的过程中，通常不希望中间层的输出分布差异过大（差异过大很可能是某层特征分布异常，可能导致损失值过大，模型收敛失败），因此初始化算法的核心需求是尽量使网络中间特征的分布相近。首先介绍 Xavier 算法，其核心思路是通过模型初始化使模型的中间层特征分布的方差尽可能接近。欲对模型中间层特征的分布分析，首先表示第 i 层的输出由公式 2.1表示。

$$P_i = W_i X_i + b_i. \tag{2.1}$$

可继续得到第 i 层输出的方差为公式 2.2。

$$Var(P_i) = Var(W_i X_i). \tag{2.2}$$

又因对参数进行初始化时，W_i 中的元素为独立同分布的，此处假设 X_i 中元素同样独立同分布，可得公式 2.3。

$$\begin{aligned} Var(p_i) &= n_i Var(w_i x_i) \\ &= n_i \left\{ E[(w_i x_i)^2] - [E(w_i x_i)]^2 \right\} \\ &= n_i \left\{ E(w_i^2) E(x_i^2) - [E(w_i)]^2 [E(x_i)]^2 \right\}, \end{aligned} \tag{2.3}$$

其中，n_i 表示 W_i 中的元素个数，对卷积层而言即为卷积核中元素个数，w_i、x_i 和 p_i 分别表示 W_i、X_i 和 P_i 中的单个元素。在参数初始化时，令 w_i 为 0 均值，

又 x_i 为 0 均值，所以可得公式 2.4。

$$\begin{aligned} Var(p_i) &= n_i E(w_i^2) E(x_i^2) \\ &= n_i Var(w_i) Var(x_i). \end{aligned} \tag{2.4}$$

因此，若使 $Var[p_i]$ 和 $Var[x_i]$ 相等，需保证 $Var[w_i] = 1/n_i$。综上所述，Xavier 算法为保证层间分布相近，进行方差为 $Var[w_i] = 1/n_i$ 且均值为 0 的初始化。

然而，对某些模型中采用 Xavier 算法进行初始化会遇见麻烦，详细地说，若网络中采用 Relu 作为层间激活层，那么 x_i 就无法满足 0 均值的假设，则公式 2.4中的推导不成立。He 算法[31] 指出，此时因 $x_i = max(p_{i-1}, 0)$，可得 $E(x_i^2) = Var(p_{i-1})/2$, 从而有公式 2.5。

$$\begin{aligned} Var(p_i) &= n_i E(w_i^2) E(x_i^2) \\ &= n_i Var(w_i) \times Var(p_{i-1})/2. \end{aligned} \tag{2.5}$$

因此，若使 $Var(p_i)$ 和 $Var(p_{i-1})$ 相等，需保证 $Var(w_i) = 2/n_i$。He 算法为保证层间分布相近，进行方差为 $Var[w_i] = 2/n_i$ 且均值为 0 的初始化。

2.1.3 损失函数

损失函数是梯度反向传播的起点，定义了整个网络模型的学习目标，在模型训练的环节非常重要。本小节将介绍和本方法相关的交叉熵损失函数 (Cross Entropy Loss) 以及其在面对类别不均衡情况下的变体 Focal Loss。

交叉熵损失函数 (后简写为 CE-loss) 用于分类任务，比如在显著性检测中，检测显著性区域等价于对于显著与非显著的分类，显著性程度以在显著类别上的得分衡量。CE-loss 的计算公式如公式 2.6所示。

$$l = -[Y \log P + (1 - Y) \log(1 - P)]. \tag{2.6}$$

分析采用 CE-loss 的理论原因，首先要从分类任务对损失函数的需求开始。损失函数的意义是指导模型在可能的模型空间中拟合出相对更好的函数，以完成从输入数据的分布 X 到输出结果的分布 P 的映射。其中，P 越接近真实标注的分布 Y，模型的效果越好。因此，损失函数用于指导函数拟合，使得 P 接近于 Y。在数学上，常选用 KL 散度[32](信息领域称相对熵) 来衡量两种分布的差异。

以本任务举例，P 和 Y 是关于 X 的两种分布，则 Y 相对于 P 的 KL 散度如公式 2.7:

$$D(Y||P) = \sum_{x \in X} Y(x) \log \frac{Y(x)}{P(x)}. \tag{2.7}$$

散度值越大说明 Y 与 P 两种分布差异越大，反之越为接近，此性质符合在分类任务中对损失函数的需求。对公式 2.7进行展开得到公式 2.8

$$D(Y||P) = F_1 + loss_{ce}, \tag{2.8}$$

其中，F_1 和 CE-loss 含义如公式 2.9所示。

$$F_1 = \sum_{x \in X} Y(x) \log Y(x), loss_{ce} = -\sum_{x \in X} Y(x) \log P(x). \tag{2.9}$$

观察上述公式可以发现，CE-loss 是 Y 与 P 的交叉熵，在网络拟合的过程中，人工标注的分布 Y 不变，F_1 不变，只有 CE-loss 随网络的拟合状态变化而改变，因此选用 CE-loss 去指导网络拟合。

综上所述，指导网络拟合的过程在数学上可通过 KL 散度公式 2.7进行建模，Y 与 P 的分布越接近，其 KL 散度越小，又根据 F_1 的客观不变性，可等价于 CE-loss 越小，以上即为 CE-loss 的理论背景。

通过可视化可进一步理解 CE-loss，如图 2.3所示，当标签 Y 为 1 时，P 越接近 1，loss 值越小，当标签 Y 为 0 时，P 越接近 0，loss 值越小，这符合二分类任务对 CE-loss 的需求。

CE-loss 图 2.3可以很好地在分类任务中监督网络的训练，Lin 等研究人员指出[33]，当分类场景中存在严重的类别不平衡时，分类的学习效果会因不平衡而受到影响。例如，对于单阶段的物体检测模型而言，需要做分类的建议物体数量巨大，甚至每个像素点都有多个建议物体需要分类，但其中真正是物体的可能仅有几个，这中间就存在着严重的类别不平衡问题，会为模型的训练带来麻烦。具体而言，虽然非物体的简单样本随着网络的拟合会很容易被分类，但它们数量相对尚未分类成功的困难样本巨大，因此即便每个样本点回传的梯度不大，但由于简单样本的数量过多，梯度累积作用不容忽视，仍然会对模型产生很大的影响。

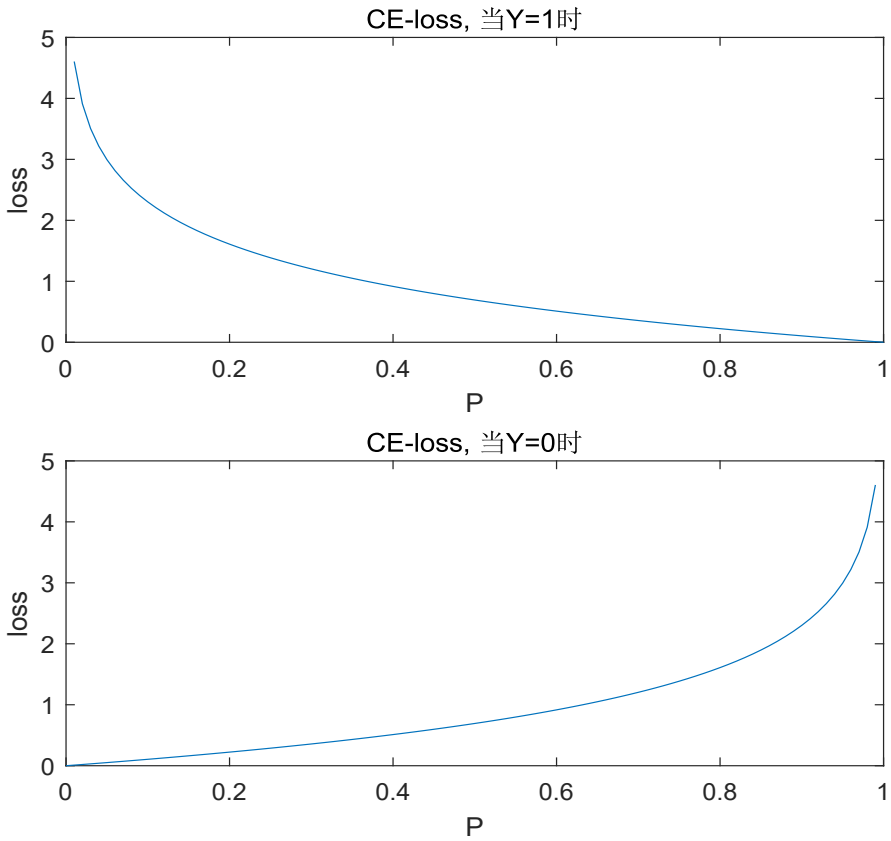

图 2.3 CE-loss 的函数曲线，上边是当标注 Y=1 时的曲线，下边是当标注 Y=0 时的曲线。

面对样本分布不平衡问题，Focal loss 提出的思路是: 降低简单样本对应梯度的权重，减弱它们对模型的影响。此处需对何为简单样本进行定义，具体地说，在网络的学习过程中，对于 $Y=1$ 的样本，预测结果 P 越接近 1，说明模型预测正确该样本的把握越大，该样本为相对简单样本。P 越接近于 0，说明模型预测正确该样本的把握越小，该样本为相对困难样本。与之相对地，对于 $Y=0$ 的样本，若预测结果 P 越接近 0，模型预测正确该样本的把握越大，该样本为相对简单样本。反之，P 越接近于 1，说明模型预测正确该样本的把握越小，该样本为相对困难样本。

基于此逻辑基础，Focal loss 在计算 loss 时把预测得分映射到损失权重上，数学表达如公式 2.10所示。

$$loss_{foc} = -(1-P_t)^{\gamma} log(P_t), \tag{2.10}$$

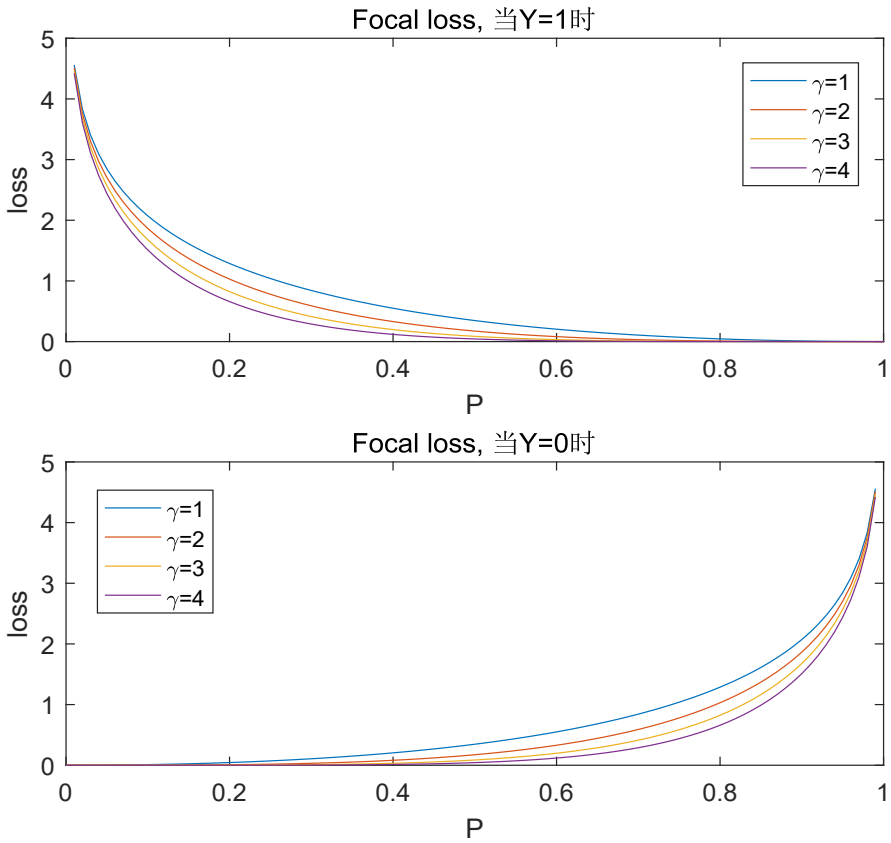

图 2.4 Focal loss 的函数曲线，上边是当标注 Y=1 时的曲线，下边是当标注 Y=0 时的曲线。

其中，P_t 的定义如公式 2.11所示。

$$P_t = \begin{cases} P, & Y = 1 \\ 1 - P, & Y = 0. \end{cases} \tag{2.11}$$

由数学表达可以观察到，当对于 $Y = 1$ 的样本，当 P 越接近于 1，样本的定义越倾向于简单，相应的在公式 2.10中给予的权重越小，弱化它回传的梯度。对于 $Y = 0$ 的样本，当 P 越接近于 0，样本的定义越倾向于简单，相应的在公式 2.10中给予的权重越小，弱化它回传的梯度。从全局来看，简单样本的梯度受到弱化，那么困难样本的梯度就会得到重视，减弱了样本不均衡带给训练的负面影响。

通过可视化可进一步理解 Focal loss，如图 2.4所示，γ 越大，曲线越向内“凹”。对于 $Y = 1$ 的样本，较大的 P 处的曲线越“平”，受到的抑制越大，梯度越小，较小的 P 处的曲线越“陡”，受到的抑制越小，梯度越大。对于 $Y = 0$ 的

样本，较小的 P 处的曲线越“平”，受到的抑制越大，梯度越小，较大的 P 处的曲线越“陡”，受到的抑制越小，梯度越大。视觉上的直观趋势符合上文所述对简单样本的抑制，弱化了简单样本回传梯度对困难样本的影响，从而一定程度地解决了样本不均衡的问题。

2.1.4 深度学习优化算法

在上一小节，本文介绍了深度学习领域中基础的损失函数，在网络的训练过程中，通过损失函数可以计算得到梯度，再对梯度进行反向传播，继而根据链式法则计算得到模型中所有的可学习参数的梯度，而得到梯度后如何更新可学习参数对网络的拟合效果也尤为重要，优化算法[34] 就是在研究如何根据梯度去更新网络中的可学习参数。

在优化算法中研究的主要对象有待优化参数 w、初始学习率 lr、和目标函数 $f(w)$，首先介绍最为基础的 Batch Gradient Descent[35](下文简写为 BGD)，BGD 的主要思路是，对整个数据集的数据计算梯度，然后再根据公式 2.12进行更新。

$$w_{t+1} = w_t - lr \times \nabla_{w_t} f(w_t, X, Y), \tag{2.12}$$

其中，w_t 表示经过 t 次更新的参数，X 表示整个数据集的输入，Y 表示整个数据集的标注。BGD 存在的问题是：1. 遇到数据集非常大的场景时，进行一次梯度更新会非常慢。2. 如数据集过大，受限于存储空间的限制将无法采用 BGD 算法更新参数。

另一种基础的优化方法 Stochastic Gradient Descent[36](下文简写为 SGD)，SGD 的主要思路是，对每个样本计算梯度，而后再根据公式 2.13进行更新。

$$w_{t+1} = w_t - lr \times \nabla_{w_t} f(w_t, x_i, y_i), \tag{2.13}$$

其中，w_t 表示经过 t 次更新的参数，x_i 表示第 i 个样本数据，y_i 表示第 i 个样本的人工标注。SGD 存在的主要问题是由于对每个样本都单独计算梯度并更新参数，可能会造成训练过程不稳定的情况。

一种结合了二者的优势的优化算法是 Mini-batch gradient descent[35](下文简写为 Mini-BGD)，其主要思路是，每次对单个训练批次，即 n 个样本 (n 小于样本总数，通常根据任务场景和显存大小确定) 进行梯度计算，继而依照公式

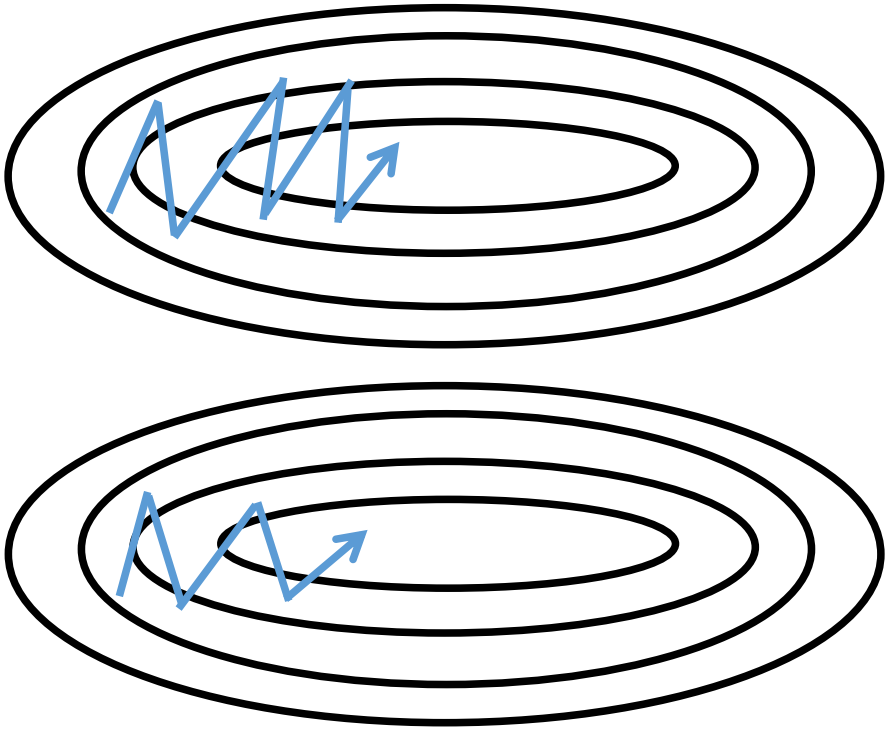

图 2.5 模型在函数空间中的拟合示意图。上边是未采用惯性动量的拟合示意图，下边是采用惯性动量的拟合示意图。

2.14更新参数。

$$w_{t+1} = w_t - lr \times \nabla_{w_t} f(w_t, x_{i:i+n}, y_{i:i+n}), \tag{2.14}$$

其中，w_t 表示经过 t 次更新的参数，$x_{i:i+n}$ 表示第 i 个训练批次的样本数据，$y_{i:i+n}$ 表示第 i 个训练批次的人工标注。Mini-BGD 相较于以上两种优化算法存在优势，首先比 SGD 每次更新计算的样本多，训练过程相对更稳定，其次比 BGD 每次更新计算的样本少，提升了更新的速度，放松了对硬件空间的限制。

以上几种算法的研究路线主要是沿根据多少样本计算梯度更新一次参数进行，另一条研究思路是，在得到梯度后如何计算参数的步进值，上述方法是直接乘上学习率 lr, 这样可能在存在一个问题，在网络拟合的过程中由于函数空间内对应位置的梯度较大，而导致模型在理想的拟合路线附近左右震荡，可视化效果如图 2.5的上半部分。面对这个问题，学者们引入了物理学中的惯性定义，希望模型在函数空间中移动时，一定程度地保留历史移动状态的影响，这样会相对减弱新算得梯度的作用，让网络对其不过分“敏感”，从而减弱震荡的程度。

将第 t 次更新时的惯性动量用 m_t 表示，则得 m_{t+1} 的计算如公式 2.15：

$$m_{t+1} = \gamma m_t + lr \times \nabla_{w_t} f(w_t), \tag{2.15}$$

其中，γ 是预设定的用于调和历史惯性动量对后续拟合影响大小的权重，根据经验[34]，γ 通常设置在 0.9 附近。得到 m_{t+1} 后，根据公式 2.16更新梯度。

$$w_{t+1} = w_t - m_{t+1}. \tag{2.16}$$

在引入惯性动量后的模型拟合示意图如图 2.5的下半部分所示，因考虑了历史动量，当下梯度会一定程度地考虑历史梯度，模型拟合的震荡程度有所减弱。

在模型训练的过程中，除震荡问题外还存在着一个值得考虑的问题：深度网络模型通常包含有较多的可学习参数，在模型拟合的过程中，不同参数的更新程度是不同的，例如一个数据集中某些大量的样本会导致特定的参数频繁更新，而其他参数更新程度相对小，那么会一定程度地降低模型在其他样本上的学习效果。面对这种参数更新程度不均衡问题，Adagrad 提出了一种可行的方案，在更新的过程中累积历史的梯度平方和，梯度平方和可以反应该梯度的历史更新程度，梯度平方和越大则更新程度越大，为 lr 增加惩罚，降低更新程度较大参数对应的学习率，梯度平方和越小则更新程度越小，相应地提升对应参数的学习率。整体的优化思路如算法 1所示。

Algorithm 1 Adagrad 算法

Input: 全局学习率 lr
Input: 目标函数 $f(w)$
Input: 初始化可学习参数 w_0
Input: 稳定小值 δ，一般设为 $1e^{-8}$(保持除法运算的数值稳定)
Output: 模型完成收敛
初始化时间戳 $t \leftarrow 0$
初始化累计变量 $s_0 \leftarrow 0$
repeat
计算 $f(w)$ 在 w_t 处的梯度：$g_t \leftarrow \nabla f(w_t)$
累计梯度平方：$s_{t+1} \leftarrow s_t + g_t \odot g_t$
计算参数调整的步进值：$d_{t+1} \leftarrow \frac{lr}{\sqrt{s_{t+1}+\delta}} \odot g_t$
更新参数：$w_{t+1} \leftarrow w_t - d_{t+1}$
until 训练结束

对比引入惯性动量 (Momentum) 算法和 Adagrad 算法可以发现，前者在数学

含义上累积了历史梯度的一阶矩，以防止模型在更新参数时会对局部梯度过于敏感而发生严重的震荡，后者在数学含义上累积了历史梯度的二阶矩，以自适应地为不同参数进行不同程度的更新。二者优势并不耦合，Adam 算法将二者核心思想结合起来，同时累积一阶矩和二阶矩，在参数更新环节防止过度震荡的同时自适应地更新不同的参数，整体的优化思路如算法 2所示。

Algorithm 2 Adam 算法

Input: 全局学习率 lr
Input: 目标函数 $f(w)$
Input: 初始化可学习可学习参数 w_0
Input: 稳定小值 δ，一般设为 $1e^{-8}$(保持除法运算的数值稳定)
Input: 矩估计的稳定衰减系数 β_1, β_2, 一般分别设为 0.9,0.999
初始化时间戳 $t \leftarrow 0$
初始化一阶矩累计变量 $m_0^1 \leftarrow 0$
初始化一阶矩累计变量 $m_0^2 \leftarrow 0$
Output: 模型完成收敛
repeat
计算 $f(w)$ 在 w_t 处的梯度：$g_t \leftarrow \nabla f(w_t)$
累积一阶矩: $m_{t+1}^1 \leftarrow \beta_1 m_t^1 + (1-\beta_1) g_t$
累计二阶矩: $m_{t+1}^2 \leftarrow \beta_2 m_t^2 + (1-\beta_2) g_t \odot g_t$
修正一阶矩偏差：$\hat{m}_{t+1}^1 \leftarrow \frac{m_{t+1}^1}{1-\beta_1^t}$
修正二阶矩偏差：$\hat{m}_{t+1}^2 \leftarrow \frac{m_{t+1}^2}{1-\beta_2^t}$
计算参数调整的步进值：$d_{t+1} \leftarrow \frac{lr}{\sqrt{\hat{m}_{t+1}^2}+\delta} \odot \hat{m}_{t+1}^1$
更新参数：$w_{t+1} \leftarrow w_t - d_{t+1}$
until 训练结束

可以看到，算法 2结合了 Momentum 和 Adagrad 的优点。在更新参数的过程中考虑了拟合方向的稳定性和参数更新的自适应性。

第二节 显著性检测

在深度学习发展后，基于深度网络的显著性检测工作涌现出来。主要以全卷积神经网络、基于多层感知器的神经网络和混合结构的深度网络为主。全卷积神经网络[15] 最早应用在语义分割 (Semantic Segmentation) 领域，网络结构如图 2.6。在此之前，深度网络主要用于做图像分类 (Image classification) 任务，分类问题的通用结构是在若干卷积层后接上若干个全连接层, 将卷积层抽取的特

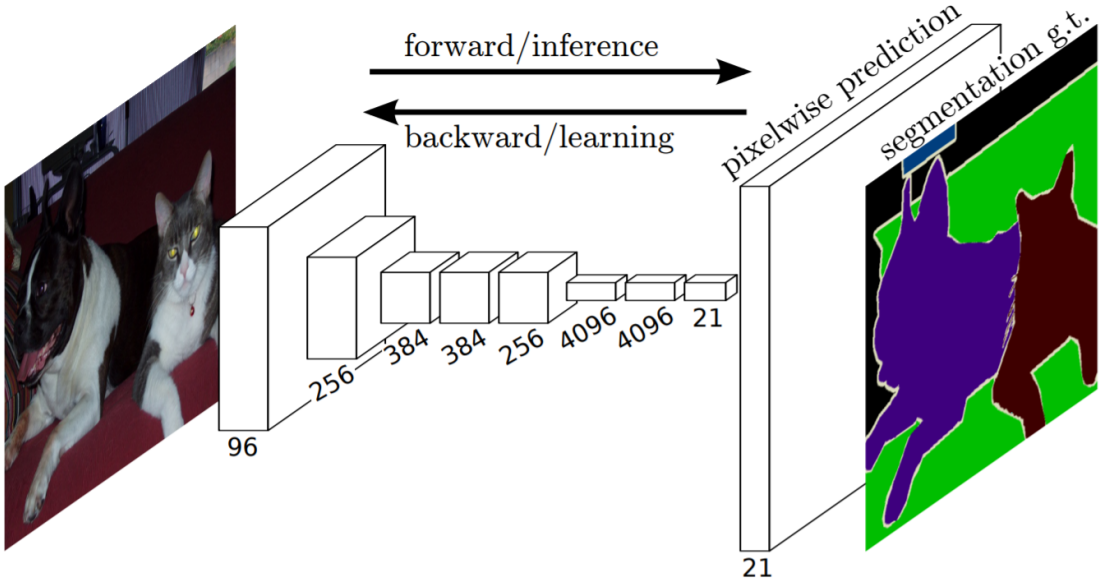

图 2.6 全卷积神经网络结构图 (图片来源[15])。可以看到，该结构实现了像素级别的预测。

征图映射到特定向量，即分类任务中的类别概率向量。全卷积深度网络，将此模型搭建逻辑泛化至图到图的映射，将全连接层去掉并且换成卷积层，在整个映射过程中始终保持空间分布，据此实现了像素级的分类，适配了语义分割的任务定义。此外，由于全卷积神经网络内部只有卷积层，对各个层的输入没有空间维度的约束，所以该网络能够处理任意尺寸的输入图像，这极大地提高了网络的适用性。显著性检测在维度角度同样是由图到图的映射，所以全卷积神经网络的基本网络结构同样可以应用在显著性检测领域。基于全卷积神经网络，涌现了一批值得关注的工作。例如单主干网络系列中的[37], 该模型的基本架构由卷积层和膨胀卷积层构成。膨胀卷积有助于在不减小特征尺度的情况下提升网络的感受野，相较于步进值 (stride) 大于 1 的卷积而言更优。在通过基本架构提取初始显著图后，对初始显著图在超像素 (super-pixel) 层面进行优化。多主干网络系列中,[38] 包括了三个自底向上/自顶向下的网络结构来处理三种尺度的输入图片。得到三种输出后，网络通过一个可学习的注意力模块 (attention module) 对其进行融合。对于深度网络而言，浅层 (即靠近输入的层) 处理的信息通常是低维度的具体信息，比如边缘和纹理等。而深层 (即靠近输出的层) 处理的信息通常是高维度的抽象信息，比如大小和位置等。考虑到网络在不同深度抽取的信息类型不同，有一批科研工作开始结合不同层面的不同信息，例如[3], 该模型在深层和浅层之间增加了层间连接，通过深层来帮助浅层定位显著性区域的位

置，通过浅层来帮助丰富显著性区域的细节。基于多层感知器的工作通常是基于超像素和目标建议区 (object proposal) 进行的。例如，[39] 是基于超像素进行处理的，它为每个超像素预测了两种手工设计的特征，进一步通过卷积层去得到显著性得分。还有一些工作采用混合结构去解决这个问题，例如[40] 从像素层面和超像素层面两个层面生成显著图，该模型采用[41] 算法自适应地生成超像素层面的显著图，同时通过融合全连接深度网络的后两层特征来得到像素层面的显著图，继而通过融合层将其融合得到最终的结果。

第三节 RGBD 显著性检测

对于 RGBD 场景下的显著性检测而言，很重要的一个研究内容是如何使用深度信息来帮助显著性检测。在先前的研究工作中，一部分工作是在模型的前阶段对深度信息和 RGB 信息进行融合。例如，[42] 的方法主要分为三个模块来从 RGBD 信息中检测出显著性检测区域。详细地说，该方法首先基于[43] 算法生成超像素区域，据此从 RGBD 信息中抽取不同的基于色彩和深度的显著性特征。然后将其输入到卷积神经网络模型中进行映射得到显著性预测图。最后通过拉普拉斯转移方法对之前的预测结果进行后处理，以得到最终的显著性检测结果。[44] 是一个两阶段检测方法，该模型根据局部的对比和全局的先验生成最终的显著图。详细地说，该方法首先将 RGB 图片转换到 CIELAB 色彩空间，然后通过[43] 等超像素分割方法将 RGB 图分割成超像素图。接着，该方法结合全局先验和局部对比度生成显著图，其中，全局先验包括深度先验、背景先验和表面方向先验。在第二个阶段，该方法采用了基于 PageRank 的采样算法和基于 MRF 的显著性恢复方法来进一步优化检测结果。[23] 结合低维特征对比度、高维定位先验、中间维特征加权因子，基于多尺度区域性分割得到了多尺度特征，接着基于随机森林在每个尺度上生成显著图，最后融合跨尺度的显著图以生成检测结果。

另一类工作是在模型的中间阶段融合 RGB 信息和深度信息，例如[45]，该工作首先将深度图编码成 HHA 格式，然后采用两个并行的同样结构的分支分别对 RGB 和深度信息进行处理，然后在不同尺度对二者进行融合以促使二者相互帮助，并且从小尺度特征依次引入信息到大尺度特征中对其进行再调整，最终得到显著性检测结果。[28] 首先对 RGB 图进行超像素分割，然后量化了背景之外的物体边界比例，得到了 LBE 特征，然后综合应用了深度先验、空间先验和背景

先验，最后采用 Grabcut 算法对边缘进行优化以得到更好的检测结果。

在模型的靠后阶段融合深度信息和 RGB 信息的方式也被相关工作采用，例如[46] 采用了一种双分流后阶段融合的方式融合两种不同模态的信息，整个网络分阶段进行训练，最终取得了不错的效果。[24] 在方法中分为三步生成显著性检测结果，首先将 RGBD 数据分成单独的 RGB 和深度图，然后分别对 RGB 和深度图进行显著性检测，其中深度图基于对局部中心旁侧相关性、全局独特性和背景信息进行显著性检测，RGB 图基于现有的 RGB 显著性检测模型进行检测，最后将二者逐像素相乘以得到综合考虑了深度信息和 RGB 信息的初步显著性检测结果，接着该方法提出了一种后处理方法。详细地讲，首先通过阈值化来得到初始的显著性种子，显著性种子中包括了显著性值较大的区域，然后在加权图中采用最小生成树生成算法 (Prim 算法[47]) 来选取显著性值较大的区域作为显著性区域。该过程不断重复，直到遍历了所有种子，得到最终的显著性检测结果。[48] 分别对超像素分割后的 RGB 图和深度图进行单独检测，然后该方法对二者进行融合以得到一个高精度的初始显著图，最后采用元胞自动机在初始显著图上迭代扩散显著值，以得到区域更完整的检测结果。[49] 首先对 RGB 图和深度图分别进行显著性检测，再将二者检测结果融合得到初始的显著图，同时该方法由输入生成中心显著性先验，将其与暗元色先验融合得到中心暗元色先验。最终将初始显著图和中心暗元色先验融合得到最终的显著图。

本方法和上述相关方法的对比在章节 第四章中进行了详细介绍。

第三章 基于对比度先验和流动金字塔的 RGBD 显著性检测方法

RGBD 显著性检测的人物场景中，需要探讨的重要问题是如何引入深度信息以及怎样将深度信息和 RGB 信息合理地结合使用。在对于 RGB 信息的处理上，本方法对 VGG-16[13] 进行修改得到基础网络结构。在对于深度信息的处理上，本方法致力于提取深度对比度先验，设计了特征增强模块提取深度对比度先验并用深度对比度先验去增强 RGB 分支的特征。在对 RGB 信息和深度信息的综合使用上，本方法在传统金字塔的基础上引入了更加丰富的多尺度连接，用以更加充分地融合多尺度跨模态信息。综上所述，本模型的整体架构由基础网络、特征增强模块 (Feature Enhanced Module, 简写为 FEM) 和流动金字塔融合方法 (Fluid Pyramid Integration，简写为 FPI) 组成。其中基础网络结构基于 VGG-16 调整产生，FEM 用于从深度图上提取深度对比度先验，FPI 为多尺度的跨模态特征提供更加充分的融合方式，详细内容将在后续小节介绍。

第一节 基础网络

在 RGB 场景的诸多任务中，很多方法会在基础网络的基础上进行修改，而基础网络通常是在分类任务上定义并且预训练的，是整个方法的根基，常用的有 VGG-16[13]、ResNet[14]、DenseNet[50]、Res2Net[51] 等，本方法基于 VGG-16 结构做了调整以适用于显著性检测任务。

VGG-16 是由牛津大学的 Visual Geometry Group 提出的一个经典的分类任务基础模型，其主要思想是，相较于先前的网络结构进一步加深了网络深度，验证了网络越深网络模型的学习能力越强的潜在规律，最终取得了更好的结果。VGG-16 模型共由五个模块串联组成，不同的模块内部分别都采用了最大池化层以对特征尺度进行缩减，因此输出不同尺度的特征。每个模块内部详细构成由若干卷积层和最大池化层串联，其中卷积层对其输入进行特征提取，在每个模块最后接的最大池化层对特征进行降采样，减小其空间尺寸，每个模块后得到的特征相较于其输入特征长和宽分别减半。信息依次经过五个模块后，特征的长宽减小比例分别是 1/2,1/4,1/8,1/16,1/32。

对于分类任务，首先对特征进行提取，之后要将特征映射到类别向量，以

表 3.1 VGG-16 结构。不同的 block 串联连接，每个 block 内部的层串联连接。对于每个层而言，conva-b 中，a 表示卷积核大小，b 表示输出通道数

block1	block2	block3	block4	block5	fc-block
conv3-64	conv3-128	conv3-256	conv3-512	conv3-512	FC-4096
conv3-64	conv3-128	conv3-256	conv3-512	conv3-512	FC-4096
max pooling	max pooling	conv3-256	conv3-512	conv3-512	FC-1000
		max pooling	max pooling	max pooling	

类别概率对输入进行分类。因此 VGG-16 在特征经过第五个模块后，后接一个全连接模块，模块内部串联三个全连接层，将空间特征映射为向量特征，最后接 softmax 得到类别概率。其内部的设置如表 3.1所示。

由表 3.1可以看到，VGG-16 网络结构最开始有五个模块，后接全连接模块。其中，第一个模块中有两个卷积层和一个最大池化层串联，两个卷积层中的卷积核大小为 3×3，其输入为 RGB 图片，因此第一个模块的输入通道数为 3，其输出通道数设置为 64，该模块经过最大池化层得到的特征相较于输入的空间尺度下降了 $1/2$。同样地，第二个模块中有两个卷积层和一个最大池化层串联，两个卷积层中的卷积核大小为 3×3，输入通道数为 64，输出通道数为 128，该模块经过最大池化层得到的特征相较于输入的空间尺度下降了 $1/4$。第三个模块中有三个卷积层和一个最大池化层串联，三个卷积层中的卷积核大小为 3×3，输入通道数为 128，输出通道数为 256，该模块经过最大池化层得到的特征相较于输入的空间尺度下降了 $1/8$。对于第四个模块而言，其内部有三个卷积层和一个最大池化层串联，三个卷积层中的卷积核大小为 3×3，输入通道数为 256，输出通道数为 512，该模块经过最大池化层得到的特征相较于输入的空间尺度下降了 $1/16$。第五个模块内部同样有三个卷积层和一个最大池化层串联，三个卷积层中的卷积核大小为 3×3，输入通道数为 512，输出通道数为 512，该模块经过最大池化层得到的特征相较于输入的空间尺度下降了 $1/32$。第五个模块后接上全连接模块，用于将空间特征映射到类别向量，三个全连接层分别有 4096、4096 和 1000 个节点，最后一个全连接层的 1000 个节点用于将特征映射到 ImageNet 的 1000 个类别的概率上。

在迁移到本任务上时，首先考虑每个模块内部卷积核大小的设置。相较于更早的工作 (如 AlexNet[12])，VGG-16 在卷积核选择上保持了简单的设置。用连续的 3×3 卷积核代替了更大的卷积核 (7×7，11×11 和 5×5 等)。其背后原理是通过串联的 3×3 卷积核可以保持和更大卷积核相同的感受野，但同时有较

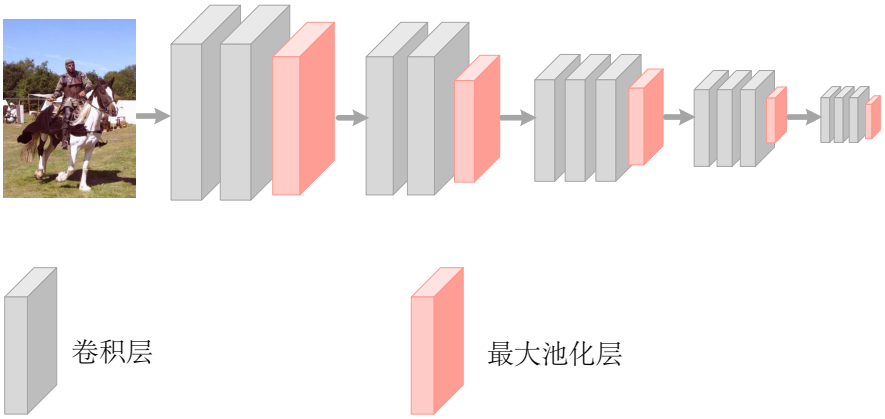

图 3.1 基础网络示意图。

小的参数量和计算量。此处以 5×5 的卷积核举例，和两个串联的 3×3 做对比。假设在一个网络模型中，第 n 层的感受野为 RF_n, 叠加步进值 st_n，第 $n+1$ 层的卷积核大小为 ks_{n+1}。那么在第 $n+1$ 层经过 5×5 的卷积核 ($ks_{n+1}=5$，步进值为 1) 做的卷积操作后的感受野如公式 3.1。

$$\begin{aligned} RF_{n+1}^{5x5} &= RF_n + st_n * (ks_{n+1} - 1) \\ &= RF_n + st_n * 4. \end{aligned} \tag{3.1}$$

若用两个串联的 3×3 卷积 ($ks_{n+1}=3$，步进值设为 1) 取代 5×5 卷积核，感受野如公式 3.2。

$$\begin{aligned} RF_{n+2}^{3x3} &= RF_{n+1} + st_{n+1} * (ks_{n+2} - 1) \\ &= RF_n + st_n * (ks_{n+1} - 1) + st_{n+1} * (ks_{n+2} - 1) \\ &= RF_n + st_n * 4. \end{aligned} \tag{3.2}$$

可以发现，5×5 的卷积核和两个串联的 3×3 感受野相同。再来对比二者的参数量，设输入通道数和输出通道数都为 c。则一个 5×5 的卷积核的参数量如公式 3.3。

$$\begin{aligned} Para^{5x5} &= c \times (c \times h \times w + 1) \\ &= 25c^2 + c. \end{aligned} \tag{3.3}$$

用两个串联的 3×3 卷积核取代 5×5 卷积核，参数量如公式 3.4。

$$\begin{aligned} Para^{3x3-3x3} &= c \times (c \times h \times w + 1) \times 2 \\ &= 18c^2 + 2c. \end{aligned} \tag{3.4}$$

又有 $c>1$，可得 $Para^{5x5}>Para^{3x3-3x3}$，因此串联的 3×3 卷积核参数量更少，同样地，用三个串联的 3×3 卷积核代替 7×7 卷积核后，感受野和参数量的论证同上。推广到更大的卷积核也可以证明，用串联的 3×3 卷积核代替大卷积核更有优势，可以在保持同样感受野的同时减少参数量。综上所述，基础网络的卷积核设置保留 VGG-16[13] 的方案，在不同的模块内用 3×3 串联对特征进行提取。

再考虑模块间的串联，因为 VGG-16 的每个模块中都通过最大池化层进行降采样，所以经过不同模块得到不同尺度的中间特征。详细地说，在五个模块后分别得到尺度缩减了 $1/2,1/4,1/8,1/16,1/32$ 的中间特征。而在深度网络的不同深度，得到的不同尺度的中间特征具备不同的特点，例如浅层的特征通常描述的是偏具体的细节特征，如边缘和纹理等。而深层的特征通常描述的是偏抽象的特征，如区域的位置和形状。可以发现，不同的特征具备不同的特点，而不同特点的特征对最终的检测可以提供不同的帮助。在显著性检测任务中，浅层特征可以帮助网络预测好细节信息，而深层特征可以帮助网络把区域定位准确，因此可以据此结合不同尺度的中间特征来生成最终的检测结果。但显著性检测与分类任务不同的地方在于，显著性检测需要保持特征的空间分布，最终完成的是从图片到图片的预测。因此本方法不对空间特征进行降采样，而是将 VGG-16 的全连接模块和 softmax 去掉，不把特征降维到向量层面，仅保留前五个模块，以抽取 RGB 图片中五种不同尺度的特征，修改后的基础网络如图 3.1所示。可以看到，该基础网络共有五个模块，每个模块内由卷积层和最大池化层组成，其中前两个模块由两个卷积层和一个最大池化层串联组成，后三个模块由三个卷积层和一个最大池化层串联组成，每个模块输出的特征维度依次递减。对于这些不同尺度的 RGB 特征，如何引入深度信息，以及如何在多尺度层面对其进行融合将在接下来的小节进行介绍。

第二节　特征增强模块

上一小节介绍了用于抽取 RGB 特征的基础网络模型，本节将介绍如何引入深度信息去帮助检测。在 RGBD 场景可以使用从深度图中提取的信息去增强 RGB 分支的特征，但简单直接地使用深度图可能会带来噪声，所以本文设计了一种对比度增强模块去增强深度信息。在特征增强模块中，本方法通过设计一种对比度增强网络来从深度图中提取深度对比度先验，并且在特征增强模块中

设计了一种简洁的跨模态融合方法去将深度信息作用于 RGB 信息。

3.2.1 对比度增强网络

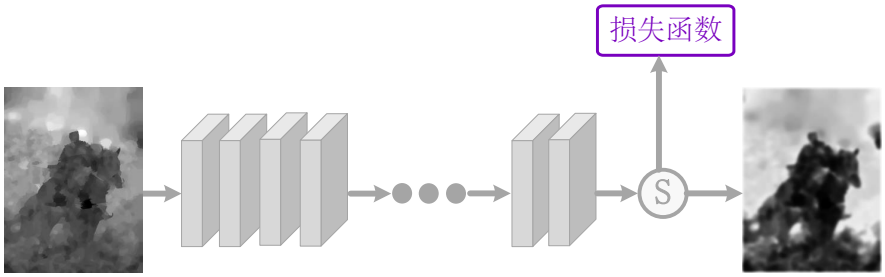

图 3.2 对比度增强网络。

在考虑如何引入深度信息时，首先要对原有深度图进行探讨。受限于深度采集设备，原始深度图通常有分布不均匀和对比度不明显等特点。具体而言，原始的深度图可能存在前景区域和背景区域对比度不够明显、前景区域分布不够均匀和背景区域分布不够均匀等问题。而受[18] 的启发，在显著性检测任务中，前景和背景的对比度以及前背景内部的均匀分布对检测结果而言较为重要，比如均匀的前景有助于更完整地检测出显著性区域，均匀的背景有助于更完整地去除背景区域，较大的前背景对比度有助于将二者更好地区分开。因此，如果能够从原始深度图中提取出具备均匀前景、均匀背景和较大对比度的深度先验，可以为辅助 RGB 分支完整地检测区域提供助力。为了合理地使用这种先验，需要由网络模型去合理地提取这种先验，构建怎样的网络结构是本文关心的下一个问题。

首先，本方法考虑该部分网络的构成，考虑网络的结构需要先分析网络处理的输入，此处也就是深度信息。通常情况下，RGB 场景包括的信息更为丰富，有丰富的语义信息和纹理信息等，而深度图中的信息只是描述对应像素点到采集设备的深度远近信息，信息类型和信息量较于 RGB 场景的信息更少。所以，本方法在处理深度信息时，没有采用用于处理 RGB 信息的复杂网络结构，而是简洁地采用若干串联的卷积层去提取深度对比度先验 (细节设置见章节 4.2.2)。此外，由于深度图片包含的信息和 RGB 图片包含的信息差异较大，在对该部分网络进行初始化时本文没有采用在分类任务上预训练的模型，而是通过高斯初始化方式对其中的可学习参数进行初始化。

其次，本方法对能提供较大帮助的深度对比度先验进行探讨，在显著性检

测任务场景中，深度对比度先验主要用于帮助扩大前景和背景区域在 RGB 分支特征上的差异，这能使得后续网络结构更容易地检测出前景区域。同时希望前景区域和背景区域对应的深度图层面的信息更为均匀，这样不会为 RGB 分支的信息引入更多的额外噪声。在考虑区域的均匀性时，本文引入方差来辅助促进前景区域和背景区域对应的深度信息更加均匀。在考虑前景区域和背景区域的对比度时，本文引入前景区域均值和背景区域均值的距离来辅助促进二者的均值的距离变得更大，即增强二者区域的对比度。结合以上考虑，本文设计了对比度损失函数指导对比度增强网络提取对比度先验。

综上所述，本方法中对比度增强网络的结构如图 3.2所示。考虑到深度图片相较于 RGB 图片包含有相对少量的信息，该网络简洁地选用若干卷积层串联而成，由一个对比度损失函数辅助最终的显著性检测损失函数指导训练。考虑到，显著性检测场景对于深度对比度先验的需求，希望深度对比度先验的前景对应区域更加均匀，深度对比度先验的背景对应区域更加均匀，前景和背景对应区域的对比度更大，该损失函数由三项组成：前景分布损失函数 l_f、背景分布损失函数 l_b 和整体分布损失函数 l_w, l_f 与 l_b 由公式3.5计算得到。

$$\begin{cases} l_f = -\log(1 - 4 * \sum_{p \in F} \frac{(p - \hat{p_f})^2}{N_f}) \\ l_b = -\log(1 - 4 * \sum_{p \in B} \frac{(p - \hat{p_b})^2}{N_b}), \end{cases} \tag{3.5}$$

其中，F 和 B 是显著性检测人工标注中对应的前景点集合和背景点集合。N_f 表示人工标注中前景区域点总个数，N_b 表示背景区域中像素点的总个数。p 表示深度信息增强图中相应点对应的显著性值，$\hat{p_f}$ 和 $\hat{p_b}$ 表示深度信息增强图中前景对应区域的均值和背景对应区域的均值 (此处前背景区域是人工标注中定义的前背景区域)，可由公式3.6计算得到。我们将特征经过逻辑回归函数 (Sigmoid) 映射到 [0, 1] 区间内，以生成深度信息增强图，可知该深度信息增强图方差的分布范围为 [0, 0.25]，乘上 4 的值域变换到 [0, 1]。可以看出，l_f 可以辅助指导网络去将深度图的前景区域点的分布的方差变小，物理意义上，会使得该前景区域点的分布更加均匀，l_b 可以辅助指导网络去将深度图的背景区域点的分布的

图 3.3 深度图和深度信息增强图。

方差变小，物理意义上，会使得该背景区域点的分布更加均匀。

$$\hat{p_f} = \sum_{p \in F} \frac{p}{N_f}, \hat{p_b} = \sum_{p \in B} \frac{p}{N_b}. \tag{3.6}$$

整体分布损失函数 l_w 的数学描述如公式3.7所示，其指导网络学习去扩大背景均值与前景均值的距离，即背景区域和前景区域的对比度被增强，以便用它去增强 RGB 分支对应区域特征的差异，为后续完整地区分两者提供帮助。

$$l_w = -\log(\hat{p_f} - \hat{p_b})^2. \tag{3.7}$$

对比度损失函数的数学表达如公式3.8所示，其中 α_1、α_2 和 α_3 分别为 5,5 和 1。如图 3.3所示，深度信息增强图和原始深度图相比，前景区域和背景区域的值的分布更为均匀，并且增强了背景区域和前景区域的对比度。

$$l_c = \alpha_1 l_f + \alpha_2 l_b + \alpha_3 l_w. \tag{3.8}$$

3.2.2 跨模态特征融合

上一小节介绍了如何从原始深度图中提取能对检测显著性区域提供帮助的深度对比度先验，在得到深度信息增强图后，如何用它去增强 RGB 分支特征，如何用它去辅助检测显著性区域是本文探讨的另一个问题。参考注意力图[52, 53]

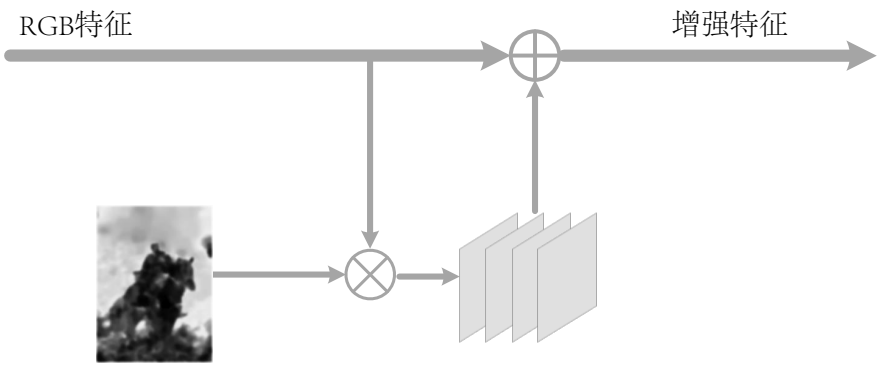

图 3.4 跨模态融合示意图。用图 3.2中得到的深度增强图与原始的 RGB 特征逐像素相乘，得到增强后的 RGB 特征后，再将其与原始的 RGB 特征逐像素相加，以得到增强后的跨模态融合特征。

的常见使用方法，深度信息增强图中的对比度先验可以用通过加权的方式给予 RGB 分支特征不同的权重，通过前景区域和背景区域的对比度差异来加权 RGB 分支特征对应区域，以扩大 RGB 分支对应区域特征的差异，以此来进一步帮助区分前景和背景区域，但同时部分深度图的特征分布可能存在会对 RGB 分支特征产生负面影响的噪声，此时需要合理地保留原始 RGB 分支特征，使得后续网络能接收到原始的 RGB 特征，因此，本文在特征增强模块中，设计了跨模态特征融合方法，用于融合得到深度对比度先验增强的 RGB 特征和原始的 RGB 特征。其主要思路是，依上文所述得到的单通道的深度信息增强图，在这里采用类似于注意力图的用法。详细地说，本文将深度信息增强图与 RGB 分支对应位置的特征逐像素相乘，来帮助增强显著性区域与非显著性区域的对比度。同时，为保留原始 RGB 分支特征，这里采用残差连接[14] 的形式融合原始 RGB 特征和使用深度信息增强图加权后的特征，以此在保存原始 RGB 特征的基础上引入深度信息，融合后得到的特征为跨模态增强特征。融合示意图如图 3.4，可以看到，将由深度对比度网络提取的深度增强图与 RGB 特征逐像素相乘，得到增强后的 RGB 特征后，再将其与原始的 RGB 特征逐像素相加，以得到增强后的跨模态融合特征。

跨模态融合的数学表达如 公式 3.9所示。可以看到，跨模态融合方法参考了注意力图[52, 53] 的用法对原始 RGB 特征进行了逐像素加权。同时，结合了残差连接[14] 的方法，对原始 RGB 特征和由深度对比度增强图增强后的 RGB 特征

进行逐像素加和，得到由深度对比度增强图增强后的跨模态融合特征。

$$\tilde{F} = F + F \odot D_E, \tag{3.9}$$

其中，F 是 RGB 分支的特征，D_E 是由上文所介绍的对比度增强网络生成的深度信息增强图。$\odot$ 表示像素级别的乘法。

综上所述，在特征增强模块中，首先通过对比度增强网络得到对比度增强图后，采用跨模态融合方法，在保留原始 RGB 特征的基础上，用深度对比度先验增强 RGB 分支特征。具体做法是，先用对比度增强图和原始 RGB 特征进行逐像素相乘，再将相乘后的结果与原始 RGB 特征逐像素相加。

本方法将特征增强模块加在基础网络图 3.1的五个模块后，以在五种尺度层面结合深度信息和 RGB 信息，可以得到多个尺度的跨模态增强特征，分别用 $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4$ 和 $\tilde{F}_5$ 表示。在五个模块后得到多个尺度的跨模态增强特征后，如何将其在多尺度层面融合将在下一个小节介绍。

第三节 流动金字塔

上一小节介绍过，在基础网络的五个模块后加上特征增强模块得到五种不同尺度的跨模态增强特征。而不同深度得到的不同尺度的特征，通常具备不同的特点。比如在网络的浅层得到的大尺度特征，通常是偏具体的特征，如边缘和纹理等。在网络的深层得到的小尺度特征，通常是偏抽象的特征，如位置和形状等。因此，在多尺度层面将不同深度的特征融合起来，有助于充分利用不同类型的信息帮助得到最终的显著性检测结果。

为设计合理的多尺度融合方法，本文首先回顾了两种常用的多尺度融合方案。第一种是[54] 中采用的方法，如图 3.5所示, 其主要思路是将不同尺度的特征直接在一个节点进行融合，可以看到该方法全面地融合了各个尺度的特征。

第二种是[55] 中采用的金字塔式的融合方法，如图 3.6所示，其主要思路是用金字塔类的结构，逐层地融合相邻尺度的特征，最终在金字塔顶得到融合结果。通过金字塔式的融合方法去将不同尺度的特征融合较于第一种方法，包含了更多尺度间连接，使得不同尺度的特征融合的更为充分。

本文在章节 4.5.2中对以上方法进行讨论和对比，发现如何将多尺度跨模态特征进行进一步更加充分的融合是在多尺度层面对跨模态特征融合的改进思路。因此，本文设计了一种流动金字塔融合方式 (Fluid pyramid integration)，用于在

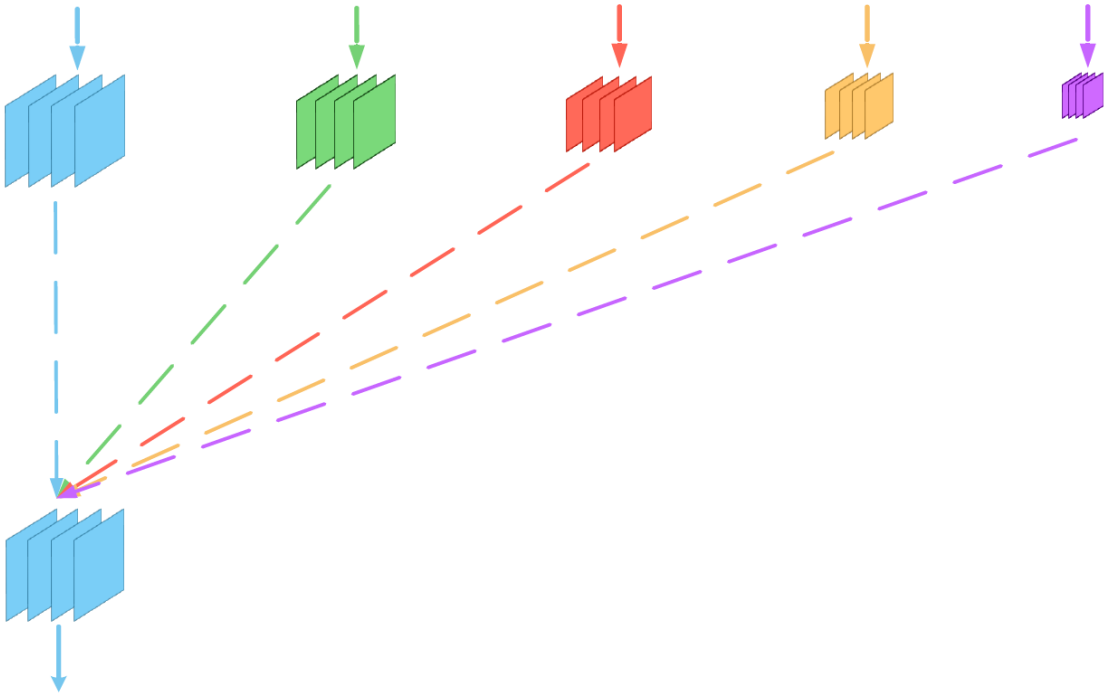

图 3.5 直接的多尺度融合方法。

多尺度层面更加充分地融合跨模态特征，从而加强两种模态信息的兼容性。

如图 3.7所示，在章节 第一节中介绍过的基础网络由 5 个模块串联，5 个模块后分别输出 5 种不同尺度的特征。相应地，流动金字塔共有 5 层。金子塔顶记为第一层，金字塔底记为第五层。第五层有 5 个节点，每个节点分别是 5 种尺度的跨模态增强信息。对于第四层而言，本方法通过上采样将 $\tilde{F}_2^5, \tilde{F}_3^5, \tilde{F}_4^5, \tilde{F}_5^5$ 上采样到和 $\tilde{F}_1^5$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第四层的第一个节点 $\tilde{F}_1^4$。本方法通过上采样将 $\tilde{F}_3^5, \tilde{F}_4^5, \tilde{F}_5^5$ 上采样到和 $\tilde{F}_2^5$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第四层的第二个节点 $\tilde{F}_2^4$。本方法通过上采样将 $\tilde{F}_4^5, \tilde{F}_5^5$ 上采样到和 $\tilde{F}_3^5$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第四层的第三个节点 $\tilde{F}_3^4$。本方法通过上采样将 $\tilde{F}_5^5$ 上采样到和 $\tilde{F}_4^5$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第四层的第四个节点 $\tilde{F}_4^4$。

对于第三层而言，本方法通过上采样将 $\tilde{F}_2^4, \tilde{F}_3^4, \tilde{F}_4^4$ 上采样到和 $\tilde{F}_1^4$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第三层的第一个节点 $\tilde{F}_1^3$。本方法通过上采样将 $\tilde{F}_3^4, \tilde{F}_4^4$ 上采样到和 $\tilde{F}_2^4$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第三层的第二个节点 $\tilde{F}_2^3$。本方法通过上采样将 $\tilde{F}_4^4$ 上采样到和 $\tilde{F}_3^4$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第三层的第三个节点 $\tilde{F}_3^3$。

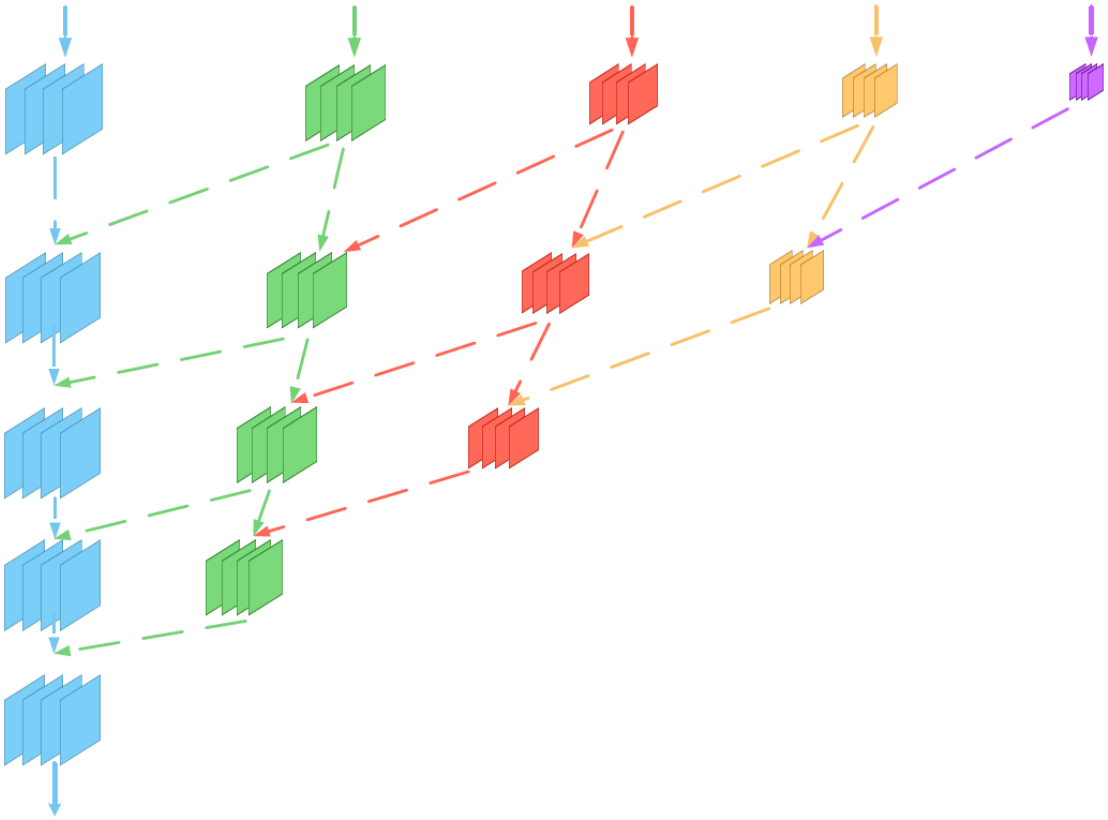

图 3.6 金字塔式的多尺度融合方法。

对于第二层而言，本方法通过上采样将 $\tilde{F}_2^3, \tilde{F}_3^3$ 上采样到和 $\tilde{F}_1^3$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第二层的第一个节点 $\tilde{F}_1^2$。本方法通过上采样将 $\tilde{F}_3^3$ 上采样到和 $\tilde{F}_2^3$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第二层的第二个节点 $\tilde{F}_2^2$。

对于第一层而言，本方法通过上采样将 $\tilde{F}_1^2$, 上采样到和 $\tilde{F}_2^2$ 相同的尺度，并将它们逐像素相加，得到流动金字塔第一层的节点 $\tilde{F}^1$。该融合方法的数学表达为公式 3.10。

$$\tilde{F}_i^j = \tilde{F}_i^{j-1} + UpS(\sum_{k=i+1}^{j-1} \tilde{F}_k^{j-1}), \tag{3.10}$$

其中，UpS 代表的是上采样操作。$\tilde{F}_i^j$ 表示第 j 层的第 i 个节点，有 $j \in [1,5]$。对于第 j 层，有 $i \in [1,j]$。

依照上述思路，金字塔底 (金字塔的第 5 层) 接受来自基础网络结构 (图 3.1) 的 5 种尺度的跨模态信息，然后经过上述融合过程，逐层对多尺度跨模态特征

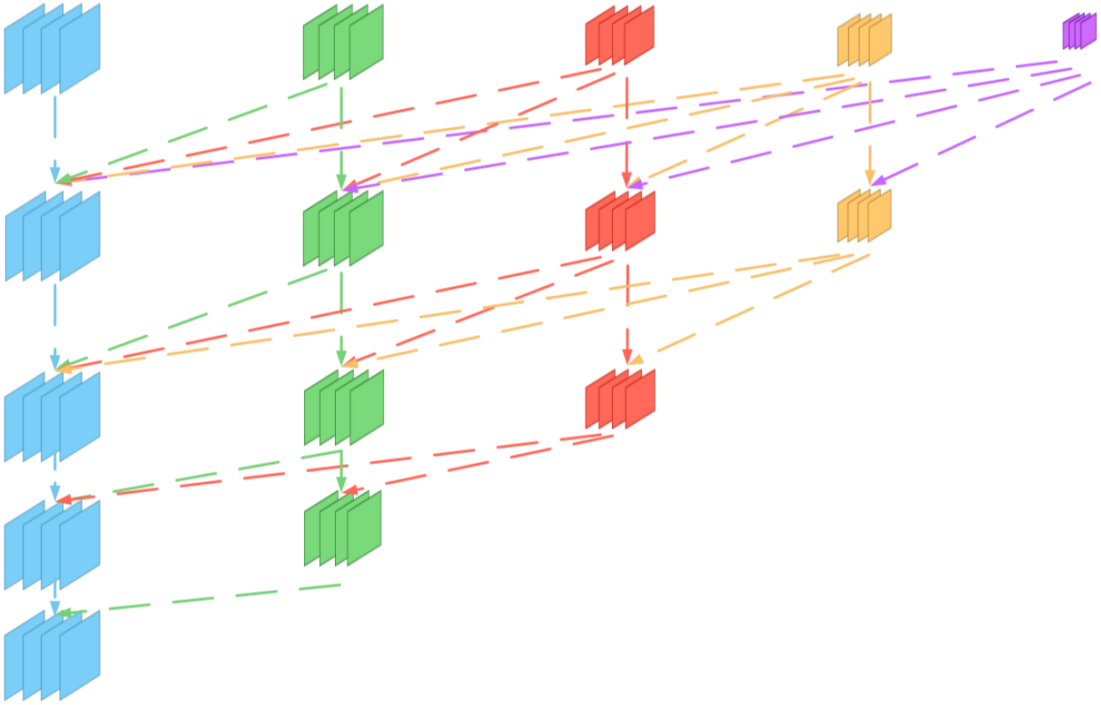

图 3.7 流动金字塔融合方法示意图。

进行融合，直到在金字塔顶合为一个节点，然后通过卷积层和 Sigmoid 层得到最终的输出 P。可以发现，相对于传统的金字塔结构，流动金字塔融合方法在金字塔的每一层，通过更丰富的连接，为金字塔的每个节点的低维度特征引入更多的高维度特征，进而为后续的融合提供更多的信息选择。

受[56] 启发，本方法为深度信息增强图在不同的尺度加上旁侧监督，因此，损失函数形式如公式3.11:

$$L = l_s + \sum_{i=1}^{5} l_{c_i}, \tag{3.11}$$

其中，l_s 表示预测的显著图和人工标注计算得到的交叉熵损失，l_{c_i} 表示第 i 个特征增强模块的对比度损失函数。交叉熵损失的计算方法如公式3.12:

$$l_f = -[Y \log P + (1 - Y) \log(1 - P)], \tag{3.12}$$

其中，P 和 Y 分别表示预测结果图和人工标注图。

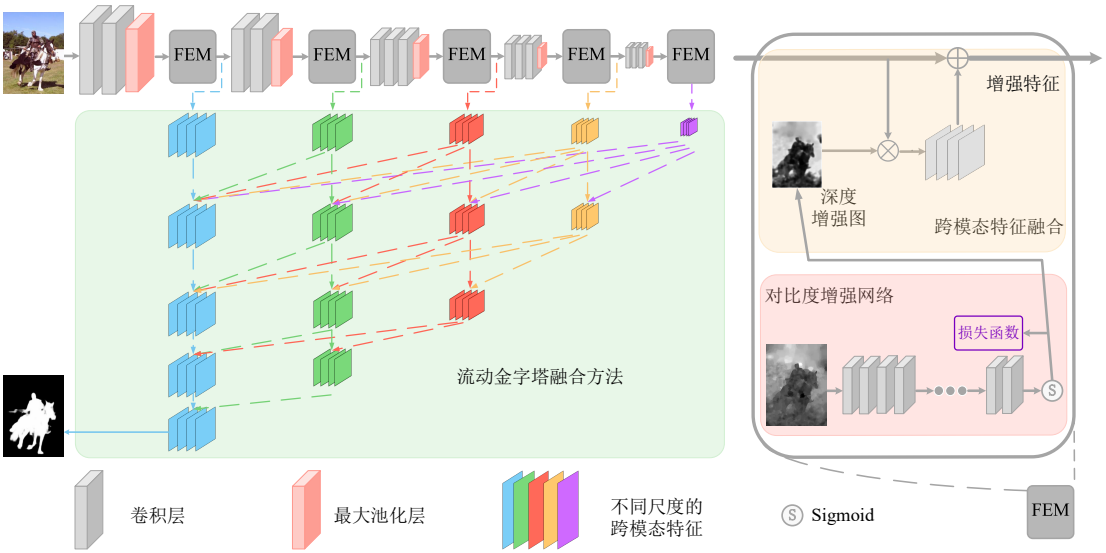

图 3.8 方法整体架构图。

第四节 整体结构

在本章的前述章节分别介绍了基础网络、特征增强模块和流动金字塔融合方法。综上所述，本方法的整体架构如图 3.8所示，基于 VGG-16 修改的基础网络用于提取 RGB 图片上的特征，在基础网络的五个模块后得到五个不同尺度的特征。其中，在基础网络的五个模块后加入特征增强模块 (章节 第二节)，以从深度图中提取深度对比度先验，而后通过跨模态融合方法增强 RGB 分支特征。在得到了五个模块后五种尺度的跨模态特征之后，本方法采用流动金字塔融合方法，通过向传统金字塔中引入更加丰富的层间连接，对五种不同尺度的跨模态特征进行充分地融合，在金字塔顶得到最终的输出结果。

第四章 实验与讨论

本章节对章节 第三章中讲述的方法进行实验，实验主要包括两大部分，一是在五种学术界公开数据集上和九种 RGBD 显著性检测方法进行对比，五种公开数据集包括：SSB[57]、NJU2K[58]、LFSD[59]、RGBD135[26] 和 NLPR[24]。其中，SSB 包含有 1000 张图片，NJU2K 包含有 2003 张图片，LFSD 包含有 100 张图片，RGBD135 包含有 135 张图片，NLPR 包含有 1000 张图片，九种 RGBD 显著性检测方法分别是 LHM[24]、GP[44]、LBE[28]、SE[48]、CTMF[49]、DF[42]、MDSF[23]、CDCP[46] 和 PCF[45]。其中，LHM、GP、LBE、SE 等是传统方法，DF、PCF 等是基于深度学习的方法。数据集内容和对比实验的详细内容将在章节 第三节中进行介绍。二是对方法进行消融实验，通过对比基础网络方案、原始深度图代替深度对比度先验、深度对比度先验、直接多尺度融合方法、传统金字塔多尺度融合方法和流动金字塔融合方法进行组合比较，以探讨各个模块对于模型的意义，详细内容将在章节 第五节中进行介绍。

第一节 评价指标选取

首先对评测 RGBD 显著性常用的 4 种评测指标进行介绍，4 种评测指标分别是：S-measure、mean F-measure、max F-measure 和 mean absolute error(MAE)。其中，mean F-measure、max F-measure 和 mean absolute error(MAE) 主要考察预测结果和人工标注中点与点的对齐情况，S-measure 主要从区域结构相似度考察预测结果和人工标注中的相似程度。本方法采用以上四个评测指标来综合评测不同方法的具体表现。

首先介绍 F-measure，F-measure 是准确率 (precision) 和召回率 (recall) 的调和平均值，综合考虑了召回率 (recall) 和准确率 (precision) 两个方面，计算公式参照公式 4.1。参照[60] 的建议, 在对预测结果二值化的环节，本文通过设置不同的阈值（0-255) 来计算 mean F-measure 和 max F-measure。

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{4.1}$$

表 4.1 TP、FP 和 FN 的含义表。

Y \ P	1	0
1	TP (True Posive)	FN (False Negative)
0	FP (False Posive)	TN (True Negative)

其中，Precision 和 Recall 的计算参照公式 4.2。

$$\begin{cases} Precision = \dfrac{TP}{TP+FP} \\ Recall = \dfrac{TP}{TP+FN}. \end{cases} \tag{4.2}$$

TP,FP,FN 的含义参照表 4.1，表中 P 表示显著性预测图，Y 表示人工标注结果，可以看到 *Precision* 主要统计预测结果图中为 1 的像素点中，有多少比例是预测准确的。*Recall* 主要统计人工标注图中为 1 的像素点中，有多少比例是被预测结果预测中的。

可以看到，F-measure 主要考察点与点的对齐情况。依照[61] 建议，本文将 β 设置为 $\beta^2 = 0.3$，以此来给予准确率 (Precision) 更多的重视。

此处选用 P 表示显著性预测图，Y 表示人工标注结果，二者的值域都是 $[0,1]$。本文参照[60]，通过公式4.3计算 MAE。

$$\varepsilon = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - Y(x,y)|, \tag{4.3}$$

其中，W 和 H 分别表示显著图的宽和高。可以看到，MAE 是在统计预测结果和人工标注的平均 L1 距离。

通过计算方式可以看出来，MAE、mean F-measure 和 max F-measure 在计算过程中，主要考虑预测的点是否与人工标注中的点匹配，比较重视显著图与人工标注中点与点的关系，而一定程度地忽视了结构相似性。行为视觉研究表明[18]，人类的视觉对场景中的结构信息很敏感。因此本文另外选用了 S-measure[18] 来衡量区域信息的相似度。S-measure 结合了区域结构相似度因子（S_r）和物体结构相似度因子（S_o）来计算结构层面的相似度，具体计算公式为4.4

$$S - measure = \alpha * S_o + (1-\alpha) * S_r, \tag{4.4}$$

其中，$\alpha \in [0,1]$ 是平衡参数，这里参数设置参照[18] 将其设置为 0.5。

第二节 细节设置

4.2.1 训练集和验证集设置

实验部分的实现平台选用伯克利研究组提出的 Caffe 框架[62]。参照[45]，训练集设置上，本方法从 NJU2K[58] 中随机取出 1400 张图片，从 NLPR[24] 随机取出 650 张用于训练。同时，验证集设置上，本方法从 NJU2K[58] 中随机取出 100 张图片，从 NLPR[24] 中取出 50 张图片用于验证，其余的图片用于测试。此外，在训练时候，通过随机翻转来对数据进行扩充。

4.2.2 参数设置

为保证深度增强图和待增强的 RGB 分支特征有着相同的尺度大小，本方法通过迭代增加两个卷积层和非线性激活层 (Relu) 来改变深度增强图的大小。其中，第一个卷积层的输出通道数、卷积核大小和步长分别设置为 (32, 4, 2)。第二个卷积层的输出通道数、卷积核大小和步长分别设置为 (32, 3, 1)。迭代通过 (卷积层-卷积层-非线性激活层) 这样的映射组合得到和 RGB 分支特征图尺度相同的深度特征图，然后该深度特征图再通过两个卷积层，这两个卷积层的输出通道数、卷积核大小和步长分别设置为 (32, 3, 1) 和 (1, 3, 1)，最后该特征再通过 Sigmoid 层得到最终的深度增强图。Sigmoid 层用于保证输出值域为 [0, 1]。

4.2.3 训练设置

在训练阶段，本方法的具体设置见表 4.2。在整个训练过程中，本方法共迭代 10000 次 (10000 iterations)。起始的学习率被设置为 $1e-7$, 在经历 7000 次 iterations 后衰减为 $1e-8$，以对整个模型进行更细致的调整。动量参数 (momentum) 设置为 0.9。此外，权重衰减参数 (weight decay) 设置为 0.0005。本方法实验基于一块 NVIDIA TITAN X GPU 显卡进行。Batch size 和 iter size 分别设置为 1 和 10。基础网络之外的新加入的卷积层通过高斯分布初始化方法进行初始化。对于长或宽大于 400 的图片数据，有可能会存在显存不够的客观限制，因此本方法在保证长宽比的基础上通过双线性插值将图片数据大小调整至新的尺寸，新的长和宽中较大的值为 400。

4.2.4 预测设置

在4.2.3中提到，为保证模型运行时满足显存要求，本方法会在数据预处理阶段会改变输入尺寸大小，所以在测试阶段，本方法会通过双线性插值将图片

表 4.2 设置参数表。

基准学习速率	1e-7
权重衰减常数	0.0005
总迭代次数	10000
学习率更新间隔	7000
学习率衰减倍数	0.1
动量参数	0.9
batch size	1
iter size	10
基础网络初始化方法	VGG-16 在 ImageNet 上的预训练模型
特征增强模块初始化方法	高斯分布初始化

数据大小调整至新的尺寸，使其与原始输入一致，以和人工标注结合进行评测。

第三节 对比实验

本节中，在 5 种现有的公开数据集上本文进行了对比实验，数据集分别是 SSB[57]、NJU2K[58]、LFSD[59]、RGBD135[26] 和 NLPR[24]。共计对比 9 种 RGBD 场景下的显著性检测模型，包括 LHM[24]、GP[44]、LBE[28]、SE[48]、CTMF[49]、DF[42]、MDSF[23]、CDCP[46] 和 PCF[45]。各种方法的显著性检测结果通过运行源代码或由原作者生成，在这里向他们的工作致以谢意。

在接下来的小节中，本文将逐小节介绍在不同数据集上和先前方法的对比实验。

4.3.1 SSB

SSB 是由波特兰州立大学的研究组在[57] 提出的 RGBD 显著性检测数据集，其图像数据来源于 Stereoscopic Image Gallery3、Flickr2 和 NVIDIA 3D Vision Live4，首先从以上数据来源下载了 1250 张图片数据，然后由三位志愿者对这 1250 张数据集用矩形框框出最显著性的物体，根据框的结果选取前 1000 张一致性最强的图片。最后由一位志愿者对这 1000 张图片的显著性物体区域进行标注，得到最终带有原图和人工标注，数据量为 1000 张图片的数据集。图 4.1是 SSB 数据集中的示例，由于空间限制，此处示例中只展示了本方法和最近的基于深度学习的其他方法的预测结果，和更多方法的图像比较示例在章节 第四节。如图 4.1，可以看到，受益于深度对比度先验，在大部分示例中本方法都取得了较好的效果。其中深度图分布不理想的场景 (如第一行场景)，本方法的特征增强

模块中采用了残差连接形式的跨模态融合，保留了原始 RGB 特征，因此依然取得了不错的预测效果。

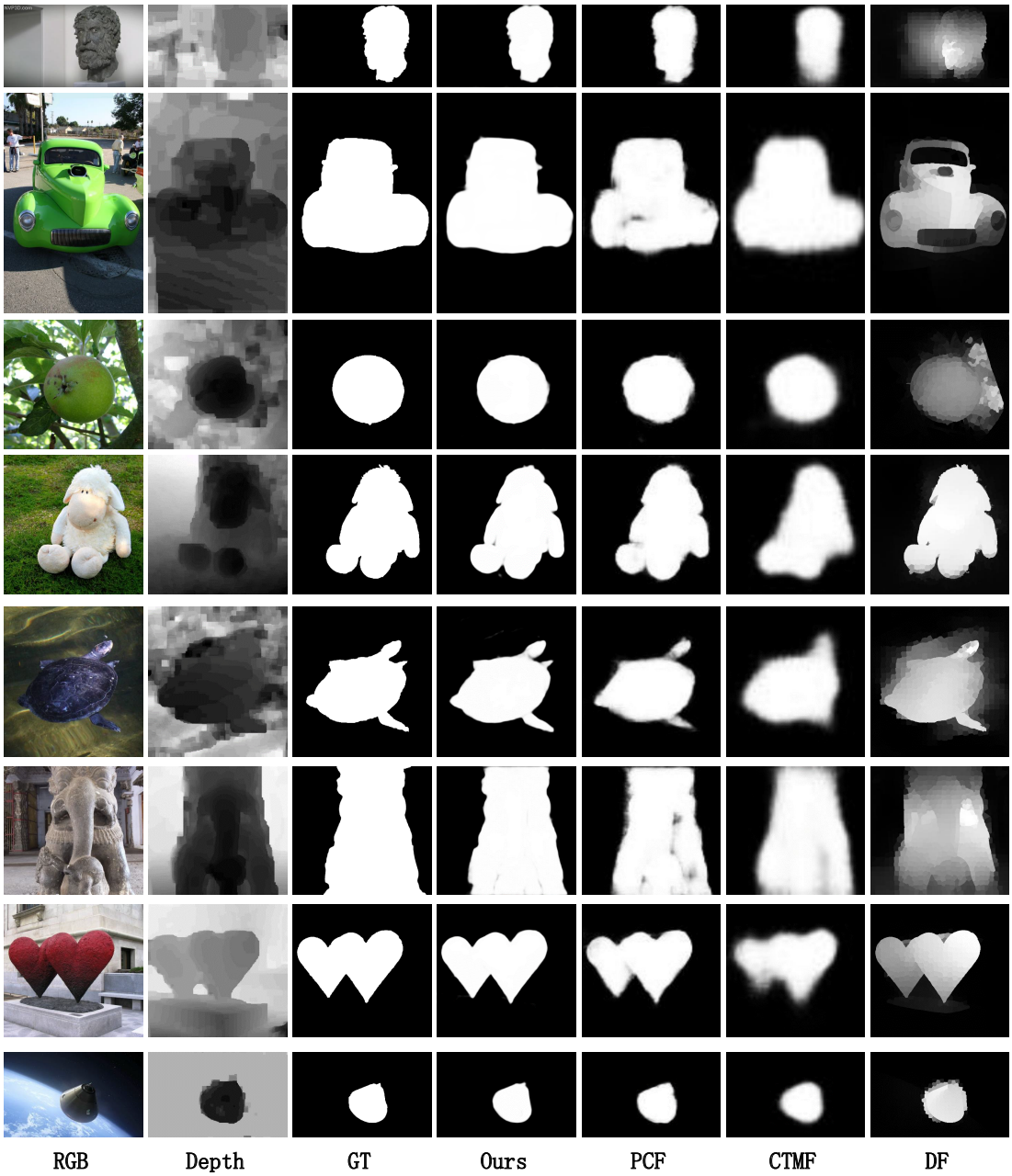

图 4.1 SSB 数据集示例。

在 SSB 数据集上进行实验，对比结果如表 4.3所示，可以看到基于深度学习的方法[45, 46] 排名较高，其中，本文所设计方法取得了领先的效果。

表 4.3 在 SSB 上的对比实验表格。包括 4 种评价指标：S-measure, mean F-measure, maximum F-measure 和 MAE。↑ & ↓ 分别表示数值越大越好或者越小越好。每行得分位列前三的分别用红色, 蓝色, and 绿色表示。

Dataset	Metric	LHM [24]	GP [44]	LBE [28]	SE [48]	CDCP [49]	DF [42]	MDSF [23]	CTMF [46]	PCF [45]	Our CPFP
SSB	S-m ↑	0.562	0.588	0.660	0.708	0.713	0.757	0.728	**0.848**	**0.875**	**0.879**
	meanF ↑	0.378	0.405	0.501	0.610	0.643	0.616	0.527	**0.758**	**0.818**	**0.842**
	maxF ↑	0.683	0.671	0.633	0.755	0.668	0.756	0.719	**0.831**	**0.860**	**0.873**
[57]	MAE ↓	0.172	0.182	0.250	0.143	0.149	0.141	0.176	**0.086**	**0.064**	**0.051**

4.3.2 NJU2K

NJU2K 是由南京大学研究组[58] 提出的 RGBD 显著性数据集，其中包括 2003 张带有多个物体和复杂场景的富有挑战的 RGB 场景图、深度场景图和人工标注，是目前数据量较大的 RGBD 显著性检测数据集。图片的数据来源是 3D 影视作品、互联网数据和 Fuji W3 stereo camera 拍摄的场景图，由四名志愿者对其进行标注。图 4.2是 NJU2K 数据集中的示例，由于空间限制，此处示例中只展示了本方法和最近的基于深度学习的其他方法的预测结果，和更多方法的图像比较示例在章节 第四节。如图 4.2，可以看到，本方法在各个示例中都取得了较好的结果，但在如第五行的汽车场景中，本方法检测结果中车的引擎盖附近有些许噪声，这表明本方法在边缘优化上还存在着提升空间。

在 NJU2K 数据集上进行实验，对比结果如表 4.4所示，可以看到相较于传统方法，基于深度学习的方法[45, 46] 排名较高。其中，本文所设计方法取得了领先的效果。

表 4.4 在 NJU2K 上的对比实验表格。包括 4 种评价指标：S-measure, mean F-measure, maximum F-measure 和 MAE。↑ & ↓ 分别表示数值越大越好或者越小越好。每行得分位列前三的分别用红色, 蓝色, and 绿色表示。

Dataset	Metric	LHM [24]	GP [44]	LBE [28]	SE [48]	CDCP [49]	DF [42]	MDSF [23]	CTMF [46]	PCF [45]	Our CPFP
NJU2K	S-m ↑	0.514	0.527	0.695	0.664	0.669	0.763	0.748	**0.849**	**0.877**	**0.878**
	meanF ↑	0.328	0.357	0.606	0.583	0.594	0.663	0.628	**0.779**	**0.840**	**0.850**
	maxF ↑	0.632	0.647	0.748	0.747	0.621	0.815	0.775	**0.845**	**0.872**	**0.877**
[58]	MAE ↓	0.205	0.211	0.153	0.169	0.180	0.136	0.157	**0.085**	**0.059**	**0.053**

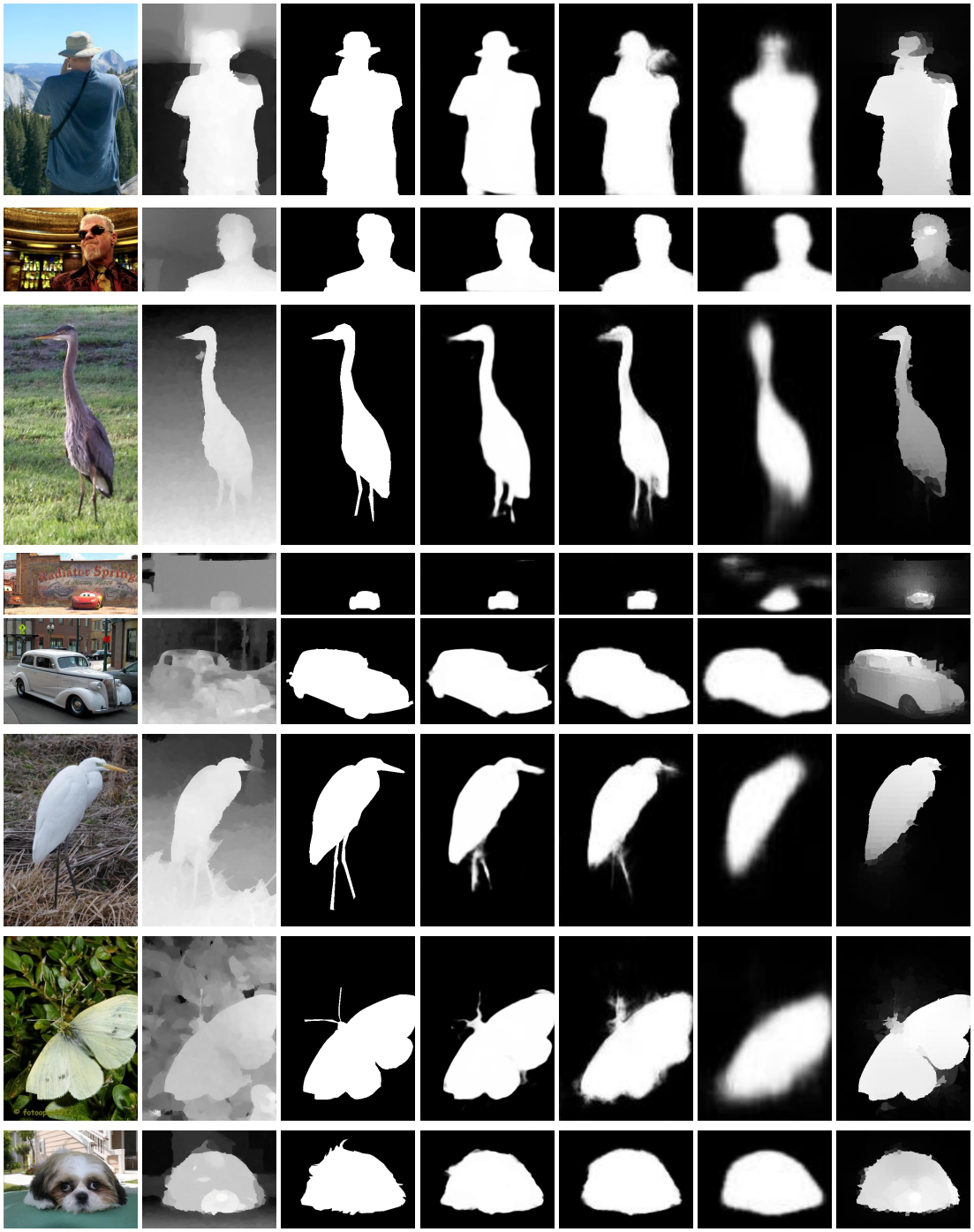

图 4.2 NJU2K 数据集示例。

4.3.3 LFSD

LFSD 是由特拉华州立大学研究组[59] 提出的 RGBD 显著性检测数据集，是一个包含 100 张人工标注、深度图和原图的小体量数据集。图片数据通过 Lytro light field camera 采集，有 60 张室内场景和 40 张室外场景。每张图片由三位志愿者手工进行人工标注。图 4.3是 LFSD 数据集中的示例，由于空间限制，此处示例中只展示了本方法和最近的基于深度学习的其他方法的预测结果，和更多方法的图像比较示例在章节 第四节。如图 4.3，可以看到，该数据集场景中的色彩和类型较为丰富，本方法受益于深度对比度先验带来的深度层面的先验信息和流动金字塔进行的更加充分的跨模态多尺度融合，在各个示例中都取得了较为理想的预测结果。

在 LFSD 数据集上开展实验，如表 4.5所示，基于深度学习的方法[45, 46] 依然保持着优势，但由于在小体量数据集上的测试可能因为数据分布存在波动，相较于前两种数据集的排名有所改变，但本文所设计方法依然取得了领先的效果。

表 4.5 在 LFSD 上的对比实验表格。包括 4 种评价指标：S-measure, mean F-measure, maximum F-measure 和 MAE。↑ & ↓ 分别表示数值越大越好或者越小越好。每行得分位列前三的分别用红色, 蓝色, and 绿色表示。

Dataset	Metric	LHM [24]	GP [44]	LBE [28]	SE [48]	CDCP [49]	DF [42]	MDSF [23]	CTMF [46]	PCF [45]	Our CPFP
LFSD	S-m ↑	0.557	0.640	0.736	0.698	0.717	0.791	0.700	**0.796**	**0.794**	**0.828**
	meanF ↑	0.396	0.519	0.611	0.640	0.680	0.679	0.521	**0.756**	**0.761**	**0.811**
	maxF ↑	0.712	0.787	0.726	0.791	0.703	**0.817**	0.783	**0.791**	0.779	**0.826**
[59]	MAE ↓	0.211	0.183	0.208	0.167	0.167	0.138	0.190	**0.119**	**0.112**	**0.088**

4.3.4 RGBD135

RGBD135 是由中国科学院信息安全实验室[26] 提出的 RGBD 显著性检测数据集，该数据集包括了 7 种室内场景，共计 135 张室内图片。这些数据的分辨率为 640 × 480，通过 Microsoft Kinect 采集。图 4.4是 RGBD135 数据集中的示例，由于空间限制，此处示例中只展示了本方法和最近的基于深度学习的其他方法的预测结果，和更多方法的图像比较示例在章节 第四节。如图 4.4，可以看到，该数据集主要是室内场景，在此类场景中本方法表现良好，在第四行的植物场景中，由于植物叶片边缘的复杂性，本方法预测结果中边缘部分未能完整地呈现，表明本方法在边缘部分的细化还有提升空间。

在 RGBD135 数据集上进行实验，对比结果如表 4.6所示，可以发现和章节

图 4.3 LFSD 数据集示例。

4.3.3中介绍的同为小体量数据集的 LFSD 结果相似，相较于大体量数据集 (在章节 4.3.1章节 4.3.2介绍) 的实验结果，小体量数据集上的测试可能因为数据分布存在波动，导致排名变动，本文所设计方法大体依旧领先。

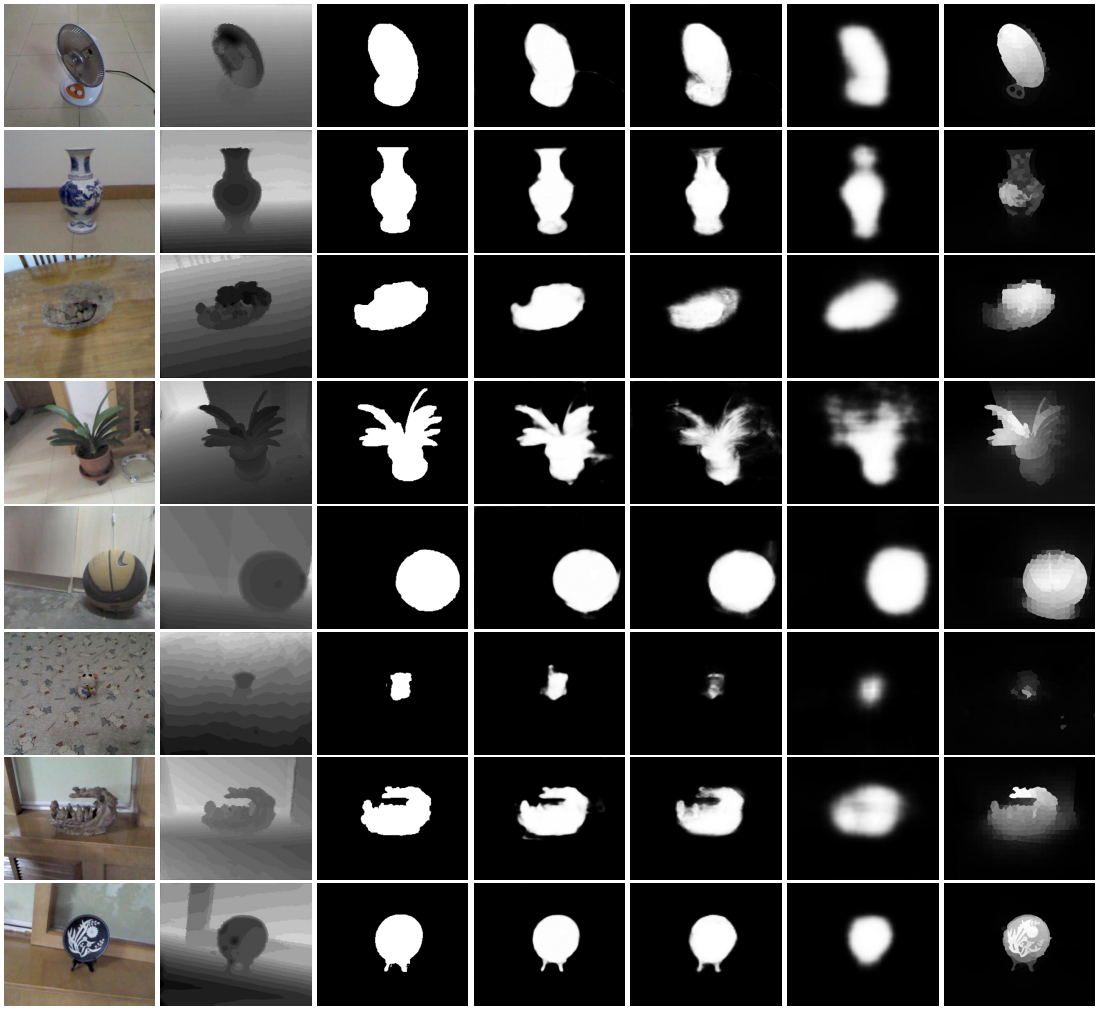

图 4.4 RGBD135 数据集示例。

4.3.5 NLPR

NLPR 是由中国科学院自动化研究院[24] 提出的 RGBD 显著性检测数据集，该数据集包括 1000 张图片，部分图片中有多个显著性物体，通过 Microsoft Kinect 采集。组建数据集的过程基于 2000 张图片数据展开，由五位志愿者通过矩形框框出最显著的区域，然后根据其结果挑选一致性排前 1000 的图片，最后由两名志愿者通过 Adobe Photoshop 去手工标注出显著性的区域。图 4.5是 NLPR 数据集中的示例，由于空间限制，此处示例中只展示了本方法和最近的基于深度学习的其他方法的预测结果，和更多方法的图像比较示例在章节 第四节。如图 4.5，可以看到，该数据集有多种角度拍摄的场景，本方法在该数据集

表 4.6 在 RGBD135 上的对比实验表格。包括 4 种评价指标：S-measure, mean F-measure, maximum F-measure 和 MAE。↑ & ↓ 分别表示数值越大越好或者越小越好。每行得分位列前三的分别用红色, 蓝色, and 绿色表示。

Dataset	Metric	LHM [24]	GP [44]	LBE [28]	SE [48]	CDCP [49]	DF [42]	MDSF [23]	CTMF [46]	PCF [45]	Our CPFP
RGBD135	S-m ↑	0.578	0.636	0.703	0.741	0.709	0.752	0.741	**0.863**	**0.842**	**0.872**
	meanF ↑	0.345	0.411	0.576	0.619	0.585	0.604	0.523	**0.756**	**0.765**	**0.815**
	maxF ↑	0.511	0.600	0.788	0.745	0.631	0.766	0.746	**0.844**	**0.804**	**0.838**
[26]	MAE ↓	0.114	0.168	0.208	0.089	0.115	0.093	0.122	**0.055**	**0.049**	**0.037**

的表现较好，受益于深度对比度先验，检测结果的区域更为均匀。

在 NLPR 数据集上开展实验，对比结果如表 4.7所示，可以发现和大体量数据集 (在章节 4.3.1章节 4.3.2介绍) 的实验结果相似，相较于小体量数据集 (在章节 4.3.4 章节 4.3.3介绍) 实验结果的波动性更小，基于深度学习的方法[45, 46] 效果排名靠前。其中，受益于特征增强模块和流动金字塔融合方法，本文所设计方法领先。

表 4.7 在 NLPR 上的对比实验表格。包括 4 种评价指标：S-measure, mean F-measure, maximum F-measure 和 MAE。↑ & ↓ 分别表示数值越大越好或者越小越好。每行得分位列前三的分别用红色, 蓝色, and 绿色表示。

Dataset	Metric	LHM [24]	GP [44]	LBE [28]	SE [48]	CDCP [49]	DF [42]	MDSF [23]	CTMF [46]	PCF [45]	Our CPFP
NLPR	S-m↑	0.630	0.654	0.762	0.756	0.727	0.802	0.805	**0.860**	**0.874**	**0.888**
	meanF ↑	0.427	0.443	0.626	0.624	0.621	0.684	0.649	**0.753**	**0.809**	**0.840**
	maxF ↑	0.622	0.603	0.745	0.720	0.655	0.792	0.793	**0.834**	**0.847**	**0.869**
[24]	MAE ↓	0.108	0.155	0.081	0.099	0.117	0.078	0.095	**0.063**	**0.052**	**0.036**

第四节 图像效果对比

上一节，本方法在五种数据集上，通过四种评价指标和先前工作进行了对比实验。在图 4.6中，本文展示了更多方法的图像结果，总结了一些在显著性检测任务上较为复杂的场景，分别是低对比度（low contrast）、复杂场景（complex scene）、小物体（small object）和多物体场景（multiple objects）。其中，对于低对比度场景，由于前景区域和背景区域在该场景中的对比度较低，所以模块在对该类场景进行检测时，模型难以区分前景区域和背景区域，出现漏检和边缘模糊的情况。在复杂场景，由于背景的复杂性，在模型检测中容易引入额外的噪声，亦或是前景背景区分不明显。对于小物体场景，由于物体在图片中尺寸较

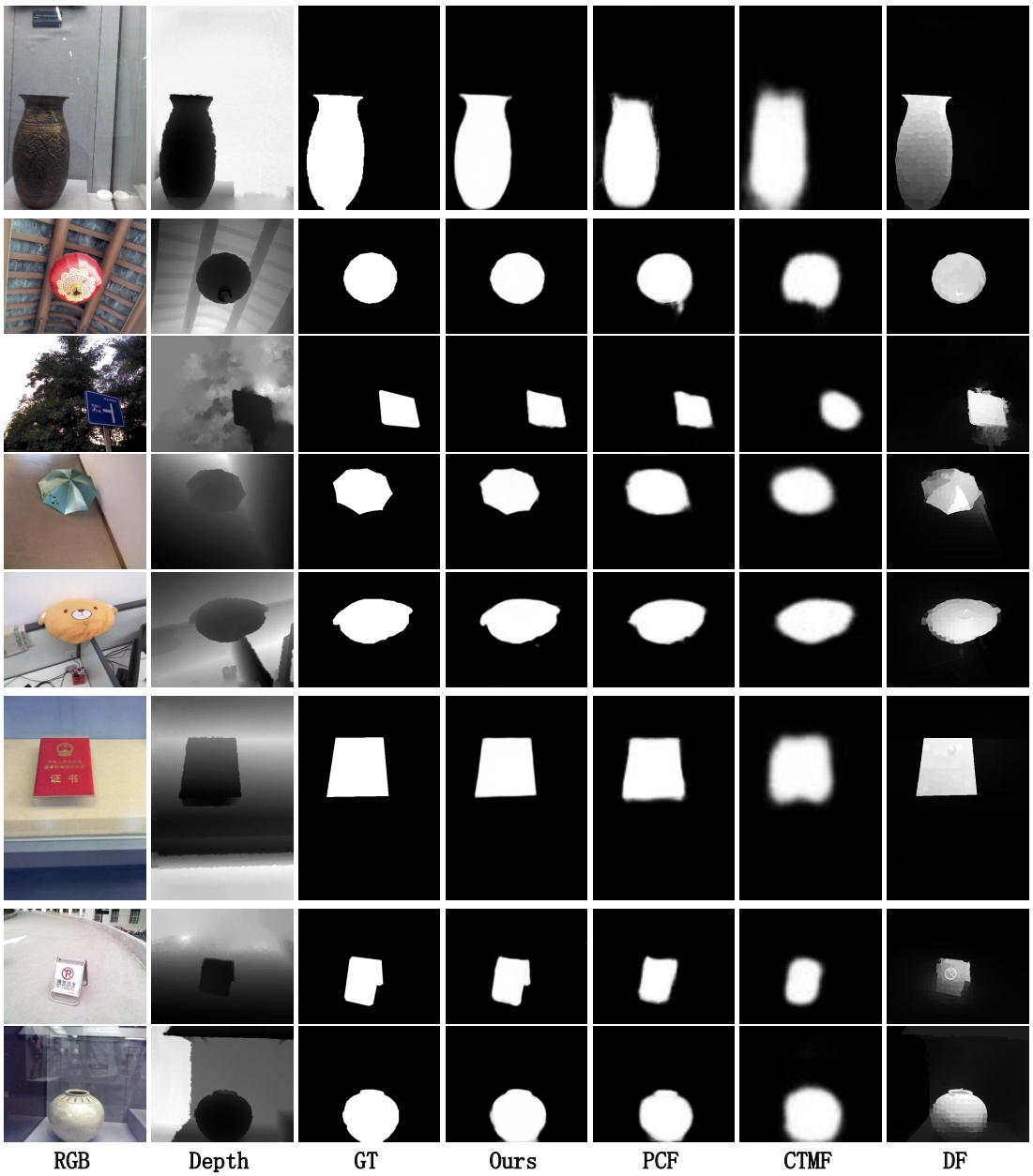

图 4.5 NLPR 数据集示例。

小，很可能出现形状不完整和没检测到的情况。在含有多个物体的场景中，容易出现模型检测遗漏的情况。

如图 4.6所示，第一栏是较为简单的场景，其中，前景区域和背景区域的区分较为明显，背景中没有很多复杂的噪声分布，可以看到大部分方法的检测结果视觉效果都挺不错。受益于参数量，基于深度的方法效果普遍较好。

在第二栏是低对比度的场景，在该场景中显著性区域和背景之间的对比差

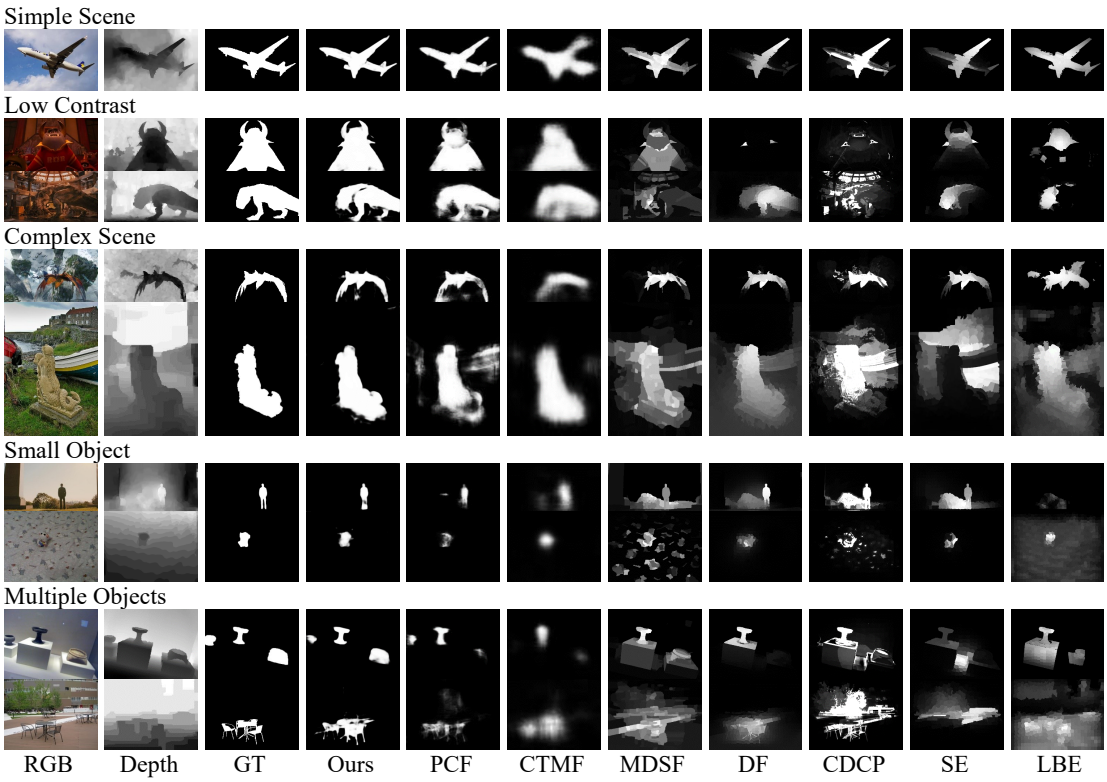

图 4.6 和更多方法效果的视觉对比。

异不明显，此时深度场景上前背景对比可以提供帮助，本方法可以基于深度信息去提取深度对比度先验，同时结合流动金字塔在多尺度层面对跨模态信息进行更加充分的融合，能够较好地检测出区域信息和细节信息。和较早的方法比（右侧），本方法检测到的区域较为完整。和基于深度学习的方法（如 PCF[45] 和 CTMF[49]）对比，本方法的细节信息相对更好。

第三栏是复杂场景，在该场景中，背景的构成较为复杂，可以看到大部分模型的检测结果会存在一些噪声，区域边界也因为背景较为复杂而变得模糊。本方法受益于深度对比度先验，在深度信息的帮助下取得了较好的结果。

第四栏是小物体场景，由于物体在整个场景中尺寸较小，大部分方法的检测结果中小物体区域的检测结果不够完整，效果模糊。本方法受益于流动金字塔中更加充分的跨尺度连接，充分地融合了多个尺度的特征，把小物体的区域检测的更为完整。

第五栏是多个物体场景，在该类场景中有多个物体，大部分方法的检测结果中存在物体漏掉的情况。本方法的流动金字塔融合可以更加充分地融合多个尺度的特征，在网络深层的小尺度特征包含较为抽象的信息 (例如位置)，可以

帮助定位不同位置的物体，网络浅层的大尺度特征包含较为具体的信息 (例如边缘)，结合不同尺度的特征，本方法可以较好地处理含有多个物体的场景。

第五节 消融实验和分析

在本章节，本文对本方法在 NJU2K 数据集[58] 上进行了消融实验和分析，主要包括对特征增强模块和流动金字塔两部分进行分析。

4.5.1 特征增强模块消融实验

表 4.8 不同模块的消融实验。B 表示基础网络结构，D 表示深度图，B + D 将原始深度图作为增强图使用，C 表示特征增强模块，M 表示简单的多尺度融合方法，如图 3.5所示，P 表示传统的金字塔融合方法，如图 3.6所示，FP 表示流动金字塔融合方法，以上细节在章节 4.5.1进行了介绍。

Model	meanF↑	maxF↑	MAE↓
B	0.714	0.791	0.115
B + D	0.708	0.788	0.121
B + C	0.756	0.806	0.094
B + FP	0.758	0.814	0.092
B + C + M	0.748	0.824	0.105
B + C + P	0.789	0.844	0.078
B + D + FP	0.783	0.842	0.081
B + C + FP	0.851	0.877	0.053

为了进一步分析特征增强模块，本文对模型的不同变种进行对比：仅使用基础网络模型（用 B 表示），在基础网络模型上增加特征增强模块（用 B+C 表示）。如表 4.8所示，通过对比第一行和第三行，可以看到特征增强模块带来了显著的提升。除此之外，本文在图 3.3中展示了深度图和深度信息增强图的视觉效果。可以看到，和原始深度图相比，深度信息增强图中显著性区域和非显著性区域之间的对比变得相对明显，同时显著性区域点的分布也更加均匀，背景点的分布同样更加均匀。

在进行了有无特征增强模块的方案对比后，为进一步研究深度信息增强图和原始深度图的效果对比，本文将深度信息增强图替换为原始深度图进行实验，该实验结果在表 4.8中的第二行，用 B+D 表示，可以看到这种替换方案使得效果有一定程度的降低，这本身是合理的，因为从图 3.3可以看到，原始深度图中显著性区域和背景的对比不够明显，同时显著性区域和背景内部分布不够均匀。与之相对的，在基础网络模型上增加特征增强模块（B+C）取得了较好的效果。

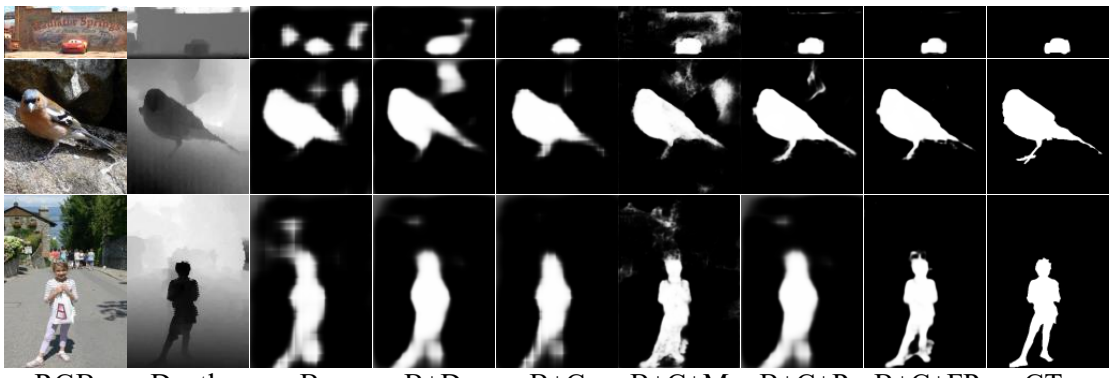

图 4.7 不同替换方案的视觉对比。其中，各个索引的含义可以在表 4.8的标注中看到。

同样的，视觉效果对比可以在图 4.7 中看到，仅使用基础网络模型（B）和在基础网络模型上使用原始深度图（B+D）效果不够好，，但当在基础网络模型上采用特征增强模块（B+C）后，受益于增强后的深度信息，多出来的噪声被一定程度地减弱了。这表明当在 RGB 场景中检测显著性物体有困难时可以借助增强后的深度信息来帮助检测，其背后的理论支持是在 RGB 场景中存在噪声或过于复杂的区域，在深度层面可能会存在较为均匀的分布。

4.5.2 流动金字塔融合方法消融实验

和传统的多尺度融合方法[54, 55] 相比，流动金字塔融合方法可以更加充分地融合多尺度跨模态信息，从而进一步地提高了跨模态信息的互补性和兼容性。在表 4.8 中可以看到采用流动金字塔融合方法 (用 B+C+FP 表示，在表 4.8中最后一行) 和不采用流动金字塔融合方法 (用 B+C 表示，在表 4.8中第三行) 的对比，可以看到流动金字塔融合方法为模型带来了很大的提升。

为进一步探讨流动金字塔融合方法的有效性，本文首先对比了一种传统的多尺度融合方案[14]，其融合方案如图 3.5所示，其主要思路是通过将不同尺度的特征变换到同一尺度后一起进行融合，该方案用 B+C+M 表示，其效果在表 4.8的第四行，结果显示该融合方法带来的帮助是有限的。

接着，本文对比了一种传统的金字塔融合方案[55]，其融合方案如图 3.6，主要思路是通过金字塔结构去逐层融合不同尺度的特征，该方案用 B+C+P 表示，其效果在表 4.8的第五行，可以看到相较于直接融合方案（B+C+M），该方案的效果更佳。进一步地，本方案在传统金字塔结构中增添了更多的跨尺度连接，检测效果在表 4.8的第六行，模型表现得到了进一步改善。视觉效果如图 4.7所示，

通过对比第五列 (B+C) 和第六列 (B+C+M)，可以看到融合多尺度信息后检测结果的边缘部分得到了改善，但有一部分在深度增强信息帮助下 (第五列，B+C) 得到减弱的噪声浮现出来，其原因是在引入传统的多尺度融合方法后，跨模态信息在多尺度层面的兼容性存在问题，导致噪声出现。在采用传统的金字塔融合方法后，非显著性噪声得到减弱，说明在增强跨尺度连接后跨模态信息的兼容性得到增强。本方法继续在传统金字塔融合方法中加入更多的跨尺度连接 (B+C+FP)，在金字塔的每个节点引入更多的不同尺度信息，可以看到噪声得到进一步减弱，显著性区域的定位更加准确，跨模态信息的兼容性得到进一步增强。

第五章 总结与展望

显著性检测在科研领域具有广泛的应用，可以用在视频/图像分割[15, 16]、内容感知[20]、弱监督语义分割[21, 63] 和视觉跟踪[17] 等方向上。例如，在弱监督语义分割领域,[63] 通过结合带有区域信息的实例显著性分割结果和带有类别信息和部分空间信息的注意力图，生成带有区域信息和类别信息的语义分割伪标注，再用伪标注和原图去训练现有的语义分割模型[4]，在这个过程中避免了稠密的语义分割标注，实现了弱监督的语义分割流程。在工业界，显著性检测同样正发挥着重要的作用，例如在 Huawei Mate 10 等智能手机中，进行拍摄时可以通过显著性检测去增强显著区域，以在整个场景中强调并优化显著性区域 (如自拍场景中的人脸)，达到更好地拍摄效果。

深度学习发展以来，受益于较大的数据集和更多的参数量，显著性检测模型的效果得到了进一步提升。而后，有学者基于全卷积神经网络提出了进一步改进的方法，例如[3] 引入短连接来结合深浅层信息,[64] 引入了新的池化层设计,[22] 引入了边缘指导网络训练，显著性检测领域得到了发展，然而在一些困难的场景，如低对比度场景和复杂场景，由于 RGB 场景数据分布较为复杂，使得 RGB 显著性检测模型效果受限。但在这些困难的 RGB 场景，其对应的深度图中，前景和背景的分布可能较为均匀，因此如何更好地结合深度信息，完成 RGBD 场景下的显著性检测是学者们关注的另一个主要问题。

现有的 RGBD 场景的显著性检测工作，通常是对深度信息和 RGB 信息在网络模型的早段[23, 24]、中段[25] 和末段[28] 进行融合。其中存在着以下两个方面的问题:

1) 数据层面: 现有的 RGBD 显著性检测数据集中缺少大体量的高质量深度图，深度传感器采集得到的深度图通常含有噪声，这为模型带来了一定麻烦。同时深度图的数据量受限，例如其中较大的数据集 NJU2000[58] 也只有两千个场景，和含有 1400 多万张图片的 ImageNet[11] 比相差甚远，所以没有在大体量数据集上的进行预训练的基础网络用于处理深度信息。

2) 跨模态信息融合层面:RGB 图片和深度图包含有两种不同模态的信息，本身存在着较大差异。例如在 RGB 场景中，“植物”和绿色的类别相关性更大，因

为在分布层面，植物的大部分颜色是绿色。但深度图包含的信息是对应像素点到深度采集设备的距离信息，所以不会存着这种相关性。那么对这两种模态的信息进行简单的融合会导致跨模态信息不兼容问题。

针对于以上问题，本文使用深度对比度先验增强 RGB 分支特征，并用流动金字塔融合方法在多尺度层面对跨模态信息进一步融合。详细地讲，与采用在 ImageNet 上预训练过的基础网络去处理深度信息的常用方法不同，本文设计了特征增强模块去提取深度对比度先验，深度对比度先验中对应的背景区域和前景区域的分布相对均匀，同时背景区域和前景区域的对比度较大。采用深度对比度先验增强 RGB 分支的特征，同时考虑到深度图上可能存在噪声，本文采用残差连接的方式去融合跨模态信息。在得到五种尺度的跨模态特征后，本文为了从多尺度层面增强跨模态信息的兼容性，在传统的金字塔融合方法的基础上引入更加充分的层间连接，以流动金字塔的结构对跨模态信息进行更加充分的融合，以在多尺度层面增强跨模态信息的兼容性。为了对比本文方法和先前方法的效果，本文在五种 RGBD 显著性检测数据集上，与九种 RGBD 显著性检测方法，通过四种评价指标进行了对比。同时本文对本方法的各个模块进行了消融实验，以进一步研究各个模块对于本方法的重要性。

在未来的研究工作中，有以下几个方向可以继续尝试:

1) 设计兼顾速度和效果的模型: 如想要增强 RGBD 显著性检测的应用价值，模型的预测效果之外，模型的预测速度同样也很重要。例如实时预测的模型可以方便地处理视频场景，并且可以应用于手机、相机等重视计算速度的终端。这个方面可以尝试研究小网络模型 (如 MobileNet[65] 和 ShuffleNet[66] 等) 的设计思想，看能否在特定的 RGBD 场景中，结合深度信息的特点做出适配的改进。

2) 更为合理的网络结构: 在如何更好地处理深度信息和 RGB 信息这个问题上，可以通过网络结构搜索[67] 的方式去自动求解更为优越的网络结构。

3) 大体量的数据集: 现有的数据集中，较大的 NJU2000[58] 也只有 2003 张图片，少量的数据集中，由于数据分布的问题，网络模型的表现可能存在波动 (参照章节 第三节)。在大体量数据集中训练得到的模型偏置会相对更小，在各种场景中的泛化能力会进一步提升。

4) 更有针对性应用场景的数据集: 在工业界，很多智能终端 (Iphone, 华为手机) 等都加上了深度传感器，用户定位更多的是考虑到通过自拍或者他拍的人像拍摄需求，因此在各种场景下的人像深度数据集具备很强的需求，针对人像主

导的显著性场景中检测任务也有很强的研究意义。

参考文献

[1] FORSYTH D A, PONCE J. Computer vision: a modern approach. [M]. Prentice Hall Professional Technical Reference, 2002.

[2] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. [C] // Advances in neural information processing systems. 2015: 91–99.

[3] HOU Q, CHENG M.-M, HU X, et al. Deeply supervised salient object detection with short connections. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3203–3212.

[4] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn. [C] // Proceedings of the IEEE international conference on computer vision. 2017: 2961–2969.

[5] LIU Y, CHENG M.-M, HU X, et al. Richer convolutional features for edge detection. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3000–3009.

[6] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation. [C] // European conference on computer vision. Springer. 2016: 483–499.

[7] BUADES A, COLL B, MOREL J.-M. A non-local algorithm for image denoising. [C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 2. IEEE. 2005: 60–65.

[8] DONG C, LOY C C, HE K, et al. Image super-resolution using deep convolutional networks. [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38 (2): 295–307.

[9] HOU X, ZHANG L. Saliency detection: A spectral residual approach. [C] // 2007 IEEE Conference on computer vision and pattern recognition. Ieee. 2007: 1–8.

[10] CHENG M.-M, MITRA N J, HUANG X, et al. Global contrast based salient region detection. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 37 (3): 569–582.

[11] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database. [C] // 2009 IEEE conference on computer vision and pattern recognition. Ieee. 2009: 248–255.

[12] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks. [C] // Advances in neural information processing systems. 2012: 1097–1105.

[13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition. [J]. ArXiv preprint arXiv:1409.1556, 2014.

[14] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778.

[15] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431–3440.

[16] FAN D.-P, WANG W, CHENG M.-M, et al. Shifting more attention to video salient object detection. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 8554–8564.

[17] BORJI A, FRINTROP S, SIHITE D N, et al. Adaptive object tracking by learning background context. [C] // 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE. 2012: 23–30.

[18] FAN D.-P, CHENG M.-M, LIU Y, et al. Structure-measure: A new way to evaluate foreground maps. [C] // Proceedings of the IEEE international conference on computer vision. 2017: 4548–4557.

[19] CHENG M.-M, HOU Q.-B, ZHANG S.-H, et al. Intelligent visual media processing: When graphics meets vision. [J]. Journal of Computer Science and Technology, 2017, 32 (1): 110–121.

[20] ZHU J.-Y, WU J, XU Y, et al. Unsupervised object class discovery via saliency-guided multiple class learning. [J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37 (4): 862–875.

[21] WEI Y, LIANG X, CHEN Y, et al. Stc: A simple to complex framework for weakly-supervised semantic segmentation. [J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39 (11): 2314–2320.

[22] ZHAO J.-X, LIU J.-J, FAN D.-P, et al. EGNet: Edge guidance network for salient object detection. [C] // Proceedings of the IEEE International Conference on Computer Vision. 2019: 8779–8788.

[23] SONG H, LIU Z, DU H, et al. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. [J]. IEEE Transactions on Image Processing, 2017, 26 (9): 4204–4216.

[24] PENG H, LI B, XIONG W, et al. Rgbd salient object detection: a benchmark and algorithms. [C] // European conference on computer vision. Springer. 2014: 92–109.

[25] FAN X, LIU Z, SUN G. Salient region detection for stereoscopic images. [C] // 2014 19th International Conference on Digital Signal Processing. IEEE. 2014: 454–458.

[26] CHENG Y, FU H, WEI X, et al. Depth enhanced saliency detection method. [C] // Proceedings of international conference on internet multimedia computing and service. 2014: 23–27.

[27] DESINGH K, KRISHNA K M, RAJAN D, et al. Depth really Matters: Improving Visual Salient Region Detection with Depth. [C] // BMVC. 2013.

[28] FENG D, BARNES N, YOU S, et al. Local background enclosure for RGB-D salient object detection. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2343–2350.

[29] SHIGEMATSU R, FENG D, YOU S, et al. Learning RGB-D Salient Object Detection using background enclosure, depth contrast, and top-down features. [C] // Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 2749–2757.

[30] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks. [C] // Proceedings of the thirteenth international conference on artificial intelligence and statistics. 2010: 249–256.

[31] HE K, ZHANG X, REN S, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. [C] // Proceedings of the IEEE international conference on computer vision. 2015: 1026–1034.

[32] KULLBACK S, LEIBLER R A. On information and sufficiency. [J]. The annals of mathematical statistics, 1951, 22 (1): 79–86.

[33] LIN T.-Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection. [C] // Proceedings of the IEEE international conference on computer vision. 2017: 2980–2988.

[34] RUDER S. An overview of gradient descent optimization algorithms. [J]. ArXiv preprint arXiv:1609.04747, 2016.

[35] HINTON G, SRIVASTAVA N, SWERSKY K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. [J]. Cited on, 2012, 14 (8).

[36] TSURUOKA Y, TSUJII J, ANANIADOU S. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. [C] // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. Association for Computational Linguistics. 2009: 477–485.

[37] HU P, SHUAI B, LIU J, et al. Deep level sets for salient object detection. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2300–2309.

[38] LI G, XIE Y, LIN L, et al. Instance-level salient object segmentation. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2386–2395.

[39] HE S, LAU R W, LIU W, et al. Supercnn: A superpixelwise convolutional neural network for salient object detection. [J]. International journal of computer vision, 2015, 115 (3): 330–344.

[40] TANG Y, WU X. Saliency detection via combining region-level and pixel-level predictions with CNNs. [C] // European Conference on Computer Vision. Springer. 2016: 809–825.

[41] ZHAO R, OUYANG W, LI H, et al. Saliency detection by multi-context deep learning. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1265–1274.

[42] QU L, HE S, ZHANG J, et al. RGBD salient object detection via deep fusion. [J]. IEEE Transactions on Image Processing, 2017, 26 (5): 2274–2285.

[43] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state-of-the-art superpixel methods. [J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34 (11): 2274–2282.

[44] REN J, GONG X, YU L, et al. Exploiting global priors for RGB-D saliency detection. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015: 25–32.

[45] CHEN H, LI Y. Progressively complementarity-aware fusion network for RGB-D salient object detection. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3051–3060.

[46] HAN J, CHEN H, LIU N, et al. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. [J]. IEEE transactions on cybernetics, 2017, 48 (11): 3171–3183.

[47] PRIM R C. Shortest connection networks and some generalizations. [J]. The Bell System Technical Journal, 1957, 36 (6): 1389–1401.

[48] GUO J, REN T, BEI J. Salient object detection for RGB-D image via saliency evolution. [C] // 2016 IEEE International Conference on Multimedia and Expo (ICME). IEEE. 2016: 1–6.

[49] ZHU C, LI G, WANG W, et al. An innovative salient object detection using center-dark channel prior. [C] // Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 1509–1515.

[50] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700–4708.

[51] GAO S, CHENG M.-M, ZHAO K, et al. Res2net: A new multi-scale backbone architecture. [J]. IEEE transactions on pattern analysis and machine intelligence, 2019.

[52] WANG F, JIANG M, QIAN C, et al. Residual attention network for image classification. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3156–3164.

[53] JIANG P.-T, HOU Q, CAO Y, et al. Integral Object Mining via Online Attention Accumulation. [C] // Proceedings of the IEEE International Conference on Computer Vision. 2019: 2070–2079.

[54] LI G, YU Y. Deep contrast learning for salient object detection. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 478–487.

[55] ZHAO K, SHEN W, GAO S, et al. Hi-fi: Hierarchical feature integration for skeleton detection. [J]. ArXiv preprint arXiv:1801.01849, 2018.

[56] XIE S, TU Z. Holistically-nested edge detection. [C] // Proceedings of the IEEE international conference on computer vision. 2015: 1395–1403.

[57] NIU Y, GENG Y, LI X, et al. Leveraging stereopsis for saliency analysis. [C] // 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2012: 454–461.

[58] JU R, GE L, GENG W, et al. Depth saliency based on anisotropic center-surround difference. [C] // 2014 IEEE international conference on image processing (ICIP). IEEE. 2014: 1115–1119.

[59] LI N, YE J, JI Y, et al. Saliency detection on light field. [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 2806–2813.

[60] BORJI A, CHENG M.-M, JIANG H, et al. Salient object detection: A benchmark. [J]. IEEE transactions on image processing, 2015, 24 (12): 5706–5722.

[61] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection. [C] // 2009 IEEE conference on computer vision and pattern recognition. IEEE. 2009: 1597–1604.

[62] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional architecture for fast feature embedding. [C] // Proceedings of the 22nd ACM international conference on Multimedia. 2014: 675–678.

[63] FAN R, HOU Q, CHENG M.-M, et al. Associating inter-image salient instances for weakly supervised semantic segmentation. [C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 367–383.

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

[65] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. [J]. ArXiv preprint arXiv:1704.04861, 2017.

[66] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices. [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848–6856.

[67] ELSKEN T, METZEN J H, HUTTER F. Neural architecture search: A survey. [J]. ArXiv preprint arXiv:1808.05377, 2018.

致谢

宝贵的研究生阶段就要结束了，过去三年的生活历历在目，回顾其中的点点滴滴，没有遗憾，唯有珍惜与感恩。

首先感谢我的导师程明明教授。他为实验室营造了浓厚的科研氛围，同时大力支持着我们在科研上的各种需求，让每位同学都能专心于科研工作。同时在科研的过程中老师始终以身作则，勤勉又专注的工作作风影响着我们每一位同学，他又像一位学长，和我们没有隔阂，经常和我们分享他的求学经历和科研经历，例如最开始我们需要口头用英语作报告时不太习惯，导致说出来的内容断断续续，老师和我们分享他学生时代的实验室同学们相似的经历，大力鼓励我们主动积极地开口说，后来大家的报告都流畅了很多。此外，程老师大力支持我们出国交流，鼓励我们去更广阔的平台上吸收交流更深刻的观点，激发我们的科研兴趣。程老师分享给我们的，不仅仅是做科研的态度和方法，还有很多定义问题和思考问题的方式。我现在做得不够好，希望将来能坚持进步。

回顾过去的学习和生活，还要感谢实验室的同学们，大家在科研中相互交流，生活中一起欢笑，一起运动，为我的生活增添了厚度和色彩。尤其是范登平师兄、侯淇滨师兄、赵凯师兄、刘云师兄、李仕杰师兄、刘笑畅师兄、胡晓伟师兄、王亚慧师兄、姜鹏涛、刘姜江和赵嘉星，希望大家事业顺利，天高路远，未来再见。感谢我的舍友刘姜江和刘亚飞，出门在外，宿舍就是家，舍友就是家人。还要感谢宿管阿姨，有时候回宿舍比较晚，阿姨耐心地起床为我开门。

从高中时期就在外求学，给予家人的陪伴少之又少，但他们总会以最大的耐心和关怀来理解我和支持我，这是我面对困难能坚持下去的动力，是我面对挑战最大的底气，希望他们身体健康，也希望自己更有担当，有能力照顾好家庭。

真诚地感谢各位评审组老师，在忙碌中抽出时间对我的论文进行审阅。同时感谢研究生期间的各位任课老师，从你们的课堂中收获了很多宝贵的知识。

最后，感谢南开大学对我的培养。

个人简历

2013 年考入西安电子科技大学，在 2017 年本科毕业并获得通信工程专业工学学士学位。于 2017 年至今在南开大学就读计算机科学与技术专业研究生。

研究生期间发表论文:

- Enhanced-alignment Measure for Binary Foreground Map Evaluation. Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, Ali Borji. International Joint Conferences on Artificial Intelligence (IJCAI), 2018.(CCF A 类，第三作者)
- Contrast Prior and Fluid Pyramid Integration for RGBD Salient Object Detection. Jiaxing Zhao, Yang Cao, Deng-Ping Fan, Xuan-Yi Li, Le Zhang, Ming-Ming Cheng. Computer Vision and Pattern Recognition (CVPR), 2019. (CCF A 类，并列第一作者)
- Integral Object Mining via Online Attention Accumulation. Peng-Tao Jiang, Qibin Hou, Yang Cao, Ming-Ming Cheng, Yunchao Wei, Hong-Kai Xiong IEEE International Conference on Computer Vision (ICCV), 2019 (CCF A 类，第三作者)
- EGNet: Edge guidance network for salient object detection. Jiaxing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, Ming-Ming Cheng. International Conference on Computer Vision (ICCV), 2019. (CCF A 类，第四作者)

所获荣誉:

- 新生入学奖学金 2017
- 公能一等奖学金 2019