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

基于空间信息引导卷积的实时 RGBD 语义分割方法

Real-Time RGBD Semantic Segmentation based on Spatial Information Guided Convolution

论文作者 陈林卓 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 匿名评审

南开大学研究生院

二〇二一年五月

南开大学学位论文使用授权书

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

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

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

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

3. 非公开学位论文在解密后的使用权同公开论文。

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

本人签署本授权书一份（此授权书为论文中一页），交图书馆留存。

学位论文作者暨授权人 (亲笔) 签字：_____

20　　年　　月　　日

南开大学研究生学位论文作者信息

论 文 题 目	基于空间信息引导卷积的实时 RGBD 语义分割方法				
姓　　名	陈林卓	学号	2120180449	答辩日期	2021 年 5 月 12 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	17691192703		电子邮箱	linzhuochen@foxmail.com	
通讯地址 (邮编)：天津市津南区海河教育园区同砚路 38 号 (300350)					
非公开论文编号			备注		

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

南开大学学位论文原创性声明

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

学位论文作者签名：_____ 年 月 日

非公开学位论文标注说明

(本页表中填写内容须打印)

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

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

南开大学学位评定委员会办公室盖章 (有效)

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

摘要

对输入图片进行像素级别分类的语义分割任务是计算机视觉领域的重要课题。深度传感器的普及使得空间信息易于获取。空间信息可以补充场景投影到2D平面过程中丢失的尺度与深度信息，对图片的语义分析有着重要作用。同时随着无人驾驶，机器人等领域的兴起，探索高效利用空间信息的方法来有效快速地提升网络在语义分割任务上的结果，对这些领域有着重要影响。

近年来，很多工作都在探索如何利用空间信息来提升语义分割的性能。深度图也可被视作2D图片，可以用卷积操作提取其特征，因此大部分工作将空间信息视为卷积网络的额外输入，即使用额外分支的网络提取空间信息的特征，并探索与主干网络RGB特征高效融合的方法。这类方法提升了语义分割的精度，但也极大地增加了网络的计算量与推理速度，限制了网络的实时应用。同时针对复杂的室内环境，2D卷积的结构不变性很难适应动态场景变化。探索高效充分地利用空间信息来提升语义分割性能的方法是当前RGBD语义分割的研究方向。

针对当前方法分别处理RGB图片与空间信息的效率瓶颈，以及二维卷积的固定结构与场景动态的空间变换之间存在的矛盾，本文提出了空间信息引导卷积(S-Conv)。S-Conv能够在输入图片空间信息的指导下推断卷积核的权重与采样偏移量，帮助卷积层自适应调整感受野并适应物体的几何变换。由于空间信息的直接输入，S-Conv可以直接分析出物体的尺度和空间变换，并生成对应的权值与卷积核分布，从而更好地感知场景中物体的空间关系与几何形状。S-Conv可以在增加少量计算量和参数量的情况下，充分地利用空间信息并显著地提升语义分割网络的性能。本文基于S-Conv进一步设计了一个实时RGBD语义分割网络，名为空间信息引导卷积网络(SGNet)。SGNet在通用数据集上，例如NYUDv2数据集与SUNRGBD数据集，达到了实时推理速度，并与其他方法相比有着最优的性能。

关键词： RGBD语义分割；卷积神经网络；动态卷积

Abstract

Semantic segmentation is an important task in computer vision. With the popularity of depth sensors, spatial information becomes easy to be obtained. Spatial information can supplement the loss of scale and depth information in image, playing an important role in image semantic analysis. Meanwhile, with the popularity of autonomous driving, robotics, and other fields, how to effectively and quickly improve the results of semantic segmentation by utilizing spatial information has important impact on these fields.

In recent years, a lot of work explored the methods to improve the results of semantic segmentation by utilizing spatial information. Depth maps can also be seen as 2D images, and their features can be extracted by convolution. Therefore, most work regards spatial information as additional input, using additional branches to extract its features, and exploring efficient fusion methods with RGB features. This kind of method improves the results of semantic segmentation, but also greatly increases the computational cost and inference time, limiting its scope for real-time applications. Meanwhile, in complex indoor scene, the structural invariance of 2D convolution can not adapt to dynamic scene changes. How to effectively and fully utilize spatial information to improve the performance of semantic segmentation becomes the main research direction in RGBD semantic segmentation.

Aiming at the efficiency bottleneck of processing RGB images and spatial information respectively, alone with the contradiction between the fixed structure of 2D convolution and the varying spatial transformation, this paper proposes **S**patial information guided **Conv**olution (S-Conv), which can infer the weight and sampling offset of the convolution kernel guided by spatial information , helping the convolution layer adjust the receptive field adaptively and adapt to the geometric transformation of the object. Due to the direct input of spatial information, S-Conv can directly analyze the scale and spatial transformation of object to generate corresponding weight and kernel distribution, perceiving the spatial relationship and geometric shape of the object. S-conv can

make full use of spatial information and significantly improve the performance of semantic segmentation with a little computational cost and parameters. Based on S-Conv, this paper further designs a real-time network, named **S**patial Information **G**uided Convolutional **Net**work (SGNet), which can achieve real-time inference speed on general datasets, such as NYUDv2 and SUNRGBD. SGNet achieves the state-of-the-art performance compared with other methods.

Key Words: RGBD Semantic Segmentation; Convolutional Neural Network; Dynamic Convolution

目录

第一章 绪论

语义分割是计算机视觉领域中的重要基础任务，促进了很多应用的发展，如自动驾驶，机器人，推荐系统等。随着 3D 传感器的应用普及，图片的空间信息变得易于获取。相较于传统语义分割任务中的 2D 图片输入，3D 空间信息 (深度图) 能够补充物体投影到相机平面过程中丢失的空间信息，进而提升分割网络的语义理解能力。如何高效充分地使用这些空间信息来指导语义分割任务是一个具有实用价值和挑战的课题。本文提出了一种利用空间信息指导卷积的操作来高效充分地利用空间信息提升语义分割精度。该方法可以显著提升网络在语义分割任务上的表现，并且仅增加少量参数量和计算量。

第一节 研究背景及意义

语义分割是计算机视觉的关键任务之一，旨在从像素级别上，识别图像的语义信息。随着互联网时代与 5G 时代的到来，语义分割技术赋能于各种实际应用中，例如短视频特效，推荐系统，无人驾驶，机器人等应用场景。相比于图片分类，语义分割任务能从像素级别进行分类，进而获得更精细的语义信息。同时，语义分割也经常作为其他高级语义分析任务的基本模块，其分割性能也对上游任务的表现至关重要。语义分割任务通常以 RGB 图片作为输入，输出为像素级的分类结果。然而在 3D 场景投影到 2D 相机平面的过程中丢失了物体的空间信息。因此网络输入的 RGB 图片缺少对应的空间信息。但人的视觉认知不仅仅停留在 2D 层面。当人在观察自然景物时，会本能的通过双眼估计出景物的空间结构辅助语义信息的判断，所以空间信息对于语义感知十分重要。本文以图 1.1为例进行说明：一些物体可能会反射其他物体的纹理，例如镜子或者图 1.1(a) 中的桌子。如果辅助空间信息 (深度图) 进行判断，即可发现桌子中反射纹理的部分是一个平面，而并非其反射物体的空间结构。这样就可以过滤掉对应的干扰信息。另一类例子是图 1.1(b)(c) 中的椅子，由于在图片中的椅子对比度较低，网络很难通过 RGB 信息恢复其详细的结构。图片中对应的空间信息却一目了然。如果充分利用图片对应的空间信息，会帮助网络分割出更加精细的几何形状。因此，从直观上来看，深度图可以辅助语义分割网络进行语义识别与分割，并提升

图 1.1 来自于 NYUDv2 数据集的 RGB 图，空间信息 (深度图)，与分割标签示例。图中圈出来的部分代表在 RGB 图中神经网络较难识别的区域。

语义分割的性能。尤其是对于一些较难判断的种类，例如镜子这类容易反射其他物体纹理的类别，椅子与桌子这类具有丰富空间变换的类别，冰箱，墙面和浴缸这类不具有代表性纹理的种类。近些年来，也有很多的相关工作围绕如何利用深度信息提升语义分割的精度来展开。这些工作通过不同的方法引入空间信息，进一步提升了神经网络语义分割的精度。这也证明了空间信息对于语义感知的重要性。同时随着自动驾驶的火热，3D 传感器的快速发展，使得深度信息变得易于获得。因此该领域吸引了很多企业和高校的科研人员。RGBD 语义分割相关技术已经开始应用于自动驾驶，机器人，AR/VR 等相关应用场景中。所以，高效充分地利用空间信息提升网络的语义感知能力，对这些场景至关重要。

由于深度图也可被视作 2D 图片，可以用卷积操作提取特征。因此有大量的相关工作围绕如何将深度图的特征与对应的 RGB 图片特征进行融合。这是很自然也很有效的一种方法。目前的研究重点在于如何将空间信息特征与 RGB 图片

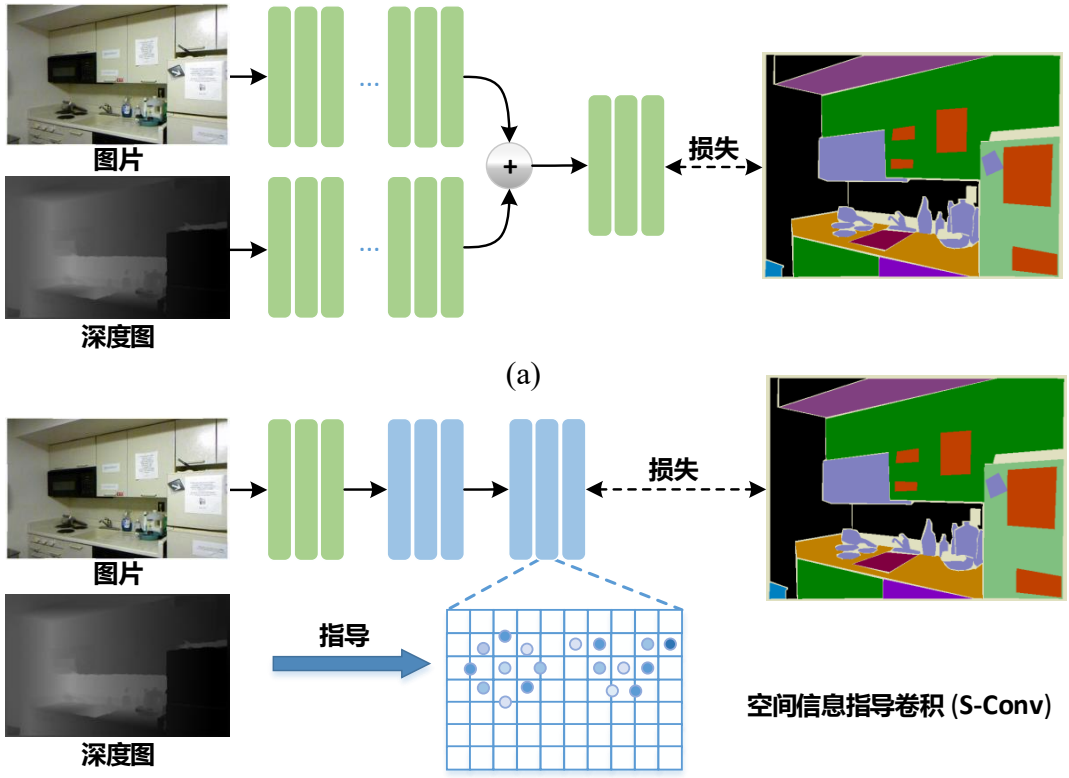

图 1.2 不同模态融合方法的示例：(a) 为双流网络的网络结构，该结构使用额外的主干网络提取深度信息的特征，并与主干网络提取的 RGB 特征进行融合。这类方法取得了较好的结果，但极大的增加了参数量与计算量，很难部署到实时应用中。(b) 为本文提出的空间信息引导卷积方法，本文使用额外的深度信息，来指导卷积核的空间权重与分布，使得卷积操作能够根据物体的尺度与空间变换，自适应的调整卷积核的分布与感受野。这种方法能够在增加少量参数量和计算量的情况下，显著地提升语义分割任务的精度。

特征充分融合来进一步提升网络的精度。例如探索双流网络不同阶段的融合以及不同的特征融合方式。然而，目前的方法存在很多尚未解决的问题，限制了其在实时场景下的应用。这些问题主要体现在以下几个方面：

第一，同时利用卷积神经网络提取深度图与 RGB 图的特征，通常的方法是使用两个网络分别提取对应的特征最后合并，如图 1.2 (a) 所示。这种方法极大的增加了网络的参数量和计算量，相对于其提升的性能来说，性价比不高。例如[1–5] 等工作，使用一个额外的主干网络来提取深度图的特征，并探索与 RGB 图特征模态融合的方法。然而为了利用空间信息，额外增加了一个主干网络的计算量与参数量。同时提升性能也相对有限，很难达到实时推理速度。一些方法[2, 5] 在两个网络的多个层进行充分特征融合，进一步提升了网络的表现，同时

也进一步加大了网络的计算量。另外，这些工作通常会使用深度图编码的 HHA 信息，来作为空间信息进行特征提取。然而这种编码过程相对耗时[2]，限制了其在实时场景的应用。第二，室内场景相对于室外环境，有着更复杂的空间关系与几何变换，这对网络的特征提取性能有着较高的要求。然而大规模应用的卷积操作，并不能很好的适应物体的空间变换，例如卷积操作缺乏旋转不变性以及尺度不变性。因此，为了得到更好的分割结果，神经网络通常需要大量的训练数据进行穷举拟合，或者添加一些池化操作。然而这些大量数据所需的大量训练成本，以及实际场景中物体的空间变换复杂多样，使得网络在复杂的室内环境下性能受限。如何将空间信息融入到卷积操作的结构中，即让卷积操作自适应于空间变换，也是目前值得探索的方向。

本文的研究重点围绕解决上述两个问题展开，旨在探索高效利用空间信息提升语义分割精度的方法。为了同时解决上述两类问题，本文提出了一种利用空间信息指导卷积的方法，名为空间信息引导卷积 (**S**patial information Guided **Conv**olution)，简称为 S-Conv。如图 1.2 (b) 所示：相比于双流网络的方法，该方法能够在增加少量参数量和计算量的情况下，显著地提升语义分割的精度。具体来说，本文学习利用空间信息指导卷积核的权重和分布，使其自适应于复杂场景的动态空间关系，增强网络的空间变换适应能力和感受野调节能力，并将空间信息融入特征提取中，进一步达到高效利用空间信息的目的。同时由于其动态自适应的结构，增强了网络在复杂场景下的泛化能力。本文同时以该方法为基础，提出一种实时的 RGBD 语义分割网络架构，空间信息引导卷积网络 (**S**patial information **G**uided Convolution **Net**work)，简称为 SGNet。SGNet 可以达到实时推理速度，并在现有主流数据集上性能超过了其他方法。本文接下来，将介绍对应领域的国内外研究现状。

第二节 国内外研究现状

1.2.1 语义分割

近年来，卷积神经网络（CNN）的发展为语义分割的研究提供了新的思路[6, 7]。FCN[8] 是将 CNN 应用在语义分割上的先驱工作，在各个语义分割数据集上取得了令人信服的结果。FCN 如今成为各大像素级分类任务的基本框架，例如边缘检测[9]，显著性检测[10, 11]，骨架检测[12] 等。随着语义分割任务的发展，科研人员为了进一步提升 FCN 语义分割的精度，提出了很多改进的网络结

构。目前的方法可以依据网络结构分为两类，包括基于空洞卷积的方法[13–16] 和编码器-解码器的架构[17–22]。这两类架构，都旨在通过解决 FCN 输出特征分辨率不足的问题来提升语义分割的性能。还有一些工作，致力于研究实时语义分割架构，寻找网络速度和性能之间的权衡点。本文接下来将分别介绍这些工作。

基于空洞卷积的方法

为了保证语义分割的精度，足够大的感受野至关重要[18]。FCN 网络架构依赖于步长大于 2 的卷积或者池化来减少网络特征图的输出尺寸，进而保证一个足够大的感受野。然而，增大感受野的代价是减少了特征图的分辨率，限制了语义分割的精度[13]。为了解决这一问题，目前的方法通过空洞卷积，在不损失特征图分辨率的情况下，增大网络的感受野。DeepLabv3[23] 将空洞卷积用于主干网络，保证了主干网络有足够大的感受野，并提出 ASPP 模块，通过不同尺度的空洞卷积带来多尺度的感受野组合。ESPNet[24] 提出了基于空洞卷积的空间金字塔来高效地多尺度提取网络特征，并减少了网络的参数量和计算复杂度。DenseASPP[25] 将 ASPP 模块与 DenseNet[26] 的思想结合，提出了 DenseASPP 模块，进一步扩大了网络多尺度的感受野组合。同时，也有一些工作提出基于池化的方法来提升网络的感受野。PSPNet[27] 基于池化操作提出 PPM (Pyramid Pooling Module) 来聚合不同尺度的信息，并更好的捕捉场景下的上下文信息。同时其主干网络也采用了空洞卷积来进一步提升感受野。工作[28] 将自注意力机制与 PPM(Pyramid Pooling Module) 结合，进一步提升网络的性能。Hou 等人提出 Strip Pooling[29]，通过一个长且窄的池化层来更好的捕获图片的的上下文语义信息。

另外，其他一些方法也基于空洞卷积的主干网络，提出了一些上下文语义的聚合方法[30–33]。Non-local[30] 网络考虑特征图上所有信息，生成每个位置上的注意力权重来关联上下文信息，在视频分类领域取得了很好的效果，但需要耗费相对较多的计算资源与时间成本。为了缓解这一问题，CCNet[33] 提出了一种近似方法 (criss-cross attention module) 来减少 Non-local 网络的计算量，利用相对较少的计算量充分关联了上下文信息。为了保证相对快速的推理速度与精度，本文在网络设计上，主要考虑以空洞卷积为基础的网络结构设计。

基于编码器-解码器的架构

另一种方法使用编码器-解码器的架构，即由于中间特征图的分辨率较低，而底层特征图的分辨率较高，因此可以通过解码器的底层特征来逐渐恢复预测图的分辨率并修复细节[17–22]。DeconvNet[20] 通过一系列的反卷积操作来通过低分辨率的中间特征图生成一个高分辨率的预测图。SegNet[19] 通过编码器中最大池化的索引来指导解码器恢复特征图的细节。UNet[34] 通过添加编码器与解码器的跳层连接来提升分割性能与细节。RefineNet[17] 通过使用编码器中的底层特征来帮助解码器精修特征图的细节，进而提升网络的分割结果。DeepLabv3+[18] 将空洞卷积与编码器-解码器方法结合，即通过空洞卷积增大感受野。DeepLabv3+ 在保持中间分辨率的同时，使用编码器中的底层特征进行修复，进一步提升了分割的结果。基于编码器-解码器的架构相对来说能取得更为精细的结果，然而相比于其他方法，需要更多的推理时间与计算成本。

实时语义分割

实时的语义分割架构对于实际应用场景十分重要。本质上，实时语义分割的架构是在寻找网络速度和性能表现的权衡。ENet[35] 为实时语义分割架构的先驱。ICNet[36] 提出了实时语义分割的多尺度架构，同时引入了标签引导来进一步提升性能。ERFNet[37] 通过残差连接和因式卷积来高效地提升网络的性能，进而达到快速的推理速度。ESPNet[24] 通过提出基于空洞卷积的空间金字塔来高效的提取图片特征。BiseNet[38] 通过提出两路架构，一路架构提取空间特征，另一路架构提取上下文特征来高效地提取图片的语义特征。

1.2.2 RGBD 语义分割

在上个小节中，本文详细介绍了语义分割的方法。RGBD 语义分割任务本质上是在语义分割的基础上，研究充分高效地引入空间信息提升语义分割精度的方法。工作[39] 首先将深度图编码成 HHA 信息，然后用网络提取 HHA 特征，来进行对语义分割和目标检测任务的性能提升。很多工作[1–3, 40, 41] 将空间信息作为网络的另一个输入，着重于如何从空间信息中提取特征，并与 RGB 特征更充分高效的融合。一些工作[1–5] 使用双流网络来分别提取 RGB 与空间信息的特征，并在中间层以及最后一层合并结果。这类方法取得了很好的结果，但为了利用空间信息引入了新的主干网络，因此极大的增加了参数量和推理时间。一些方法[42–44] 使用 3D CNN 或者 3D KNN 图网络来利用空间信息，然而这种方法

需要占用大量的计算成本与推理时间。另一类方法是将空间信息融入到操作中。Cheng 等人[45] 使用几何信息来构建一个特征亲和矩阵来指导在平均池化和上采样池化操作。Lin 等人[46] 依据深度信息，将图片分成不同的分支。Wang 等人[47] 提出深度感知 CNN，它将深度信息作为一种权重的先验，提升了网络的结果, 然而这种先验是手动设定的而并非从数据中学习到的，因此在复杂的其他场景性能受限。其他方法，使用多任务学习[40, 48–52] 或者引入时间与空间的分析[53] 来进一步提升网络的结果。Wang 等人[54] 提出了一个多任务框架网络：同时预测深度图与分割图来使得两种任务相互促进，同时使用一种层级结构的 CRF 来修正最后的结果，但最后的层级 CRF 耗时较多。Kokkinos 等人[55] 提出了 UberNet, 可以同时处理多种不同的视觉任务，包括表面张量估计，语义分割等。Zhang 等人[52] 提出了模式仿射 (Pattern-Affinitive) 的传播方式来同时预测深度，表面张量与语义分割结果。

1.2.3 3D 点云的识别与分割

3D 点云的识别任务，也在探索如何利用其自身的空间信息来进行点云的识别与分割。RGBD 语义分割任务中的空间信息 (深度图) 通过相机内参可恢复成点云图。因此这两种领域的方法的思路可以相互借鉴参考。目前 3D 点云识别领域主要分为 3D 卷积与多视图方法和 3D 点直接输入的方法。

基于 3D 卷积与多视图的方法

基于 3D 卷积的方法将点集转换为常规三维栅格并利用 3D 卷积来进行处理[56, 57]。然而，3D 卷积通常会引入大量计算，同时由于点云的稀疏会引入很多不必要的计算。一些工作[31, 58–61] 致力于通过设计新的数据结构来减少 3D 卷积的计算量。例如, 稀疏的 3D 点云可以被表示为 Octree[58], Kd-树[61]。在工作[59] 中, 作者使用一种以特征为中心的投票方案来实现快速的三维卷积。而在[60] 中，引入了一种新的稀疏卷积运算对稀疏数据进行有效的 3D 卷积。多视图方法将三维点集投影为为二维视图的集合，以便对转换后的数据使用 2D 卷积运算。例如，多视图 CNN[62] 为每个视图构造 CNN，并使用池化操作来聚合每个视图的提取特征。

基于点云输入的方法

PointNet[63] 是第一个将点云坐标直接输入到神经网络并提取特征的方法。它通过共享全连接层提取每个点的特征，最后通过次序无关的聚合操作，例如

最大池化，来聚合所有点的特征得到最后的预测结果。然而，PointNet[63] 不能提取局部点云的空间特征。为了解决这个问题，PointNet++[64] 利用最远距离采样算法来选取点云的中心点，并采样中心点附近的临近点，使用共享全连接层与最大池化来提取中心点的特征。通过逐步减少中心点的方式 (类似于图像中步长大于 2 的卷积) 来得到最后的结果。

还有其他一些方法来使用深度学习处理点云, 比如[65–69]。具体来说，SO-Net[65] 使用自组织网络来处理点云。RSNet[66] 使用循环神经网络 (RNN) 来提取点集的特征。KCNet[67] 提出核相关算法来提取邻域的信息。PointCNN[68] 从点集中学到一种自适应变换矩阵来置换点云的顺序，其也可被视为一种空间自适应权重。在[69] 中, 将点云的特征投影到常规区域中，这样可以利用 CNN 来提取特征。3D 点云同样可以表示为网格[70] 或者图[71, 72]。一些工作着重于提取这些表示的点云特征。

1.2.4 CNN 中的动态结构

很多研究工作致力于利用 CNN 的动态结构来处理网络的动态输入。空洞卷积[13, 14] 可以在不损失特征图分辨率的情况下增大网络的感受野，同时也可以用来构造多尺度模块。空间变换网络[73] 通过对特征图进行变形来适应空间变换。动态卷积[74] 可以根据输入自适应的调整权重。此外，基于自注意力机制的方法[30, 75–77] 通过从中间特征图生成注意图，调整每个位置的响应来自适应地捕获远程上下文信息。SENet[76] 通过生成逐通道的注意力权重来增强通道内特征的联系。OCNet[78] 通过自注意力机制来关联语义分割上下文的信息。CCNet[33] 通过十字交叉的注意力机制来聚合上下文的信息，同时通过一种循环的方式来获取更丰富的上下文信息。同时，也有一些将卷积从 2D 图像扩展到 3D 点云的工作，通过将卷积修改为动态结构来适应点云的动态输入。PointCNN[68] 将 CNN 从 2D 图像应用到 3D 点云中。也有一些方法[79–81] 致力于使用自定义的动态结构来处理 3D 点云信息。可变形卷积[82, 83] 可以生成自适应分布与自适应权重的卷积核，然而这种自适应分布是从中间特征图中推断得到。SV Conv[77] 通过自注意力机制，联系上下文信息生成语义自适应权重来提取关联上下文的信息。这些动态的卷积结构自适应于 RGB 图输入，本文在空间自适应卷积结构上做了进一步探索。上述任务的总结如表 1.1所示。

表 1.1 RGBD 语义分割任务与其他相关任务的介绍。

任务类型	任务目标	主要技术路线
语义分割	对输入 RGB 图片进行像素级别分类。	主要分为编码器-解码器架构与基于空洞卷积的方法。
实时语义分割	实时地对输入 RGB 图片的进行像素级别分类。	主要通过多尺度空洞卷积，高效的分组卷积，以及网络架构来寻求速度和性能的最佳权衡点。
RGBD 语义分割	对输入 RGB 图片以及深度进行像素级别分类。	主要探索充分利用空间信息提升语义分割结果的方法。目前的主流方法是双流网络。
3D 点云识别与分割	对输入点云进行整体分类与像素级别分类。	目前主要通过探索将卷积操作扩展到点云的数据结构上来提取特征。

第三节 本文研究内容

本文的主要研究内容为，探索充分高效利用空间信息的方法，提升神经网络的空间感知能力进而提升语义分割的结果。本文首先针对室内 RGBD 语义分割问题，介绍相关的基础知识，包括卷积神经网络架构，损失函数，评测指标，与常用的空间信息格式等。接着，本文针对室内 RGBD 语义分割任务，设计一个基础网络，为了方便实时应用场合，该网络需要良好的性能和快速的推理速度。经过实验验证，该基础网络在常用公共数据集上表现良好的同时，可以达到实时推理速度。接着，本文提出了空间自适应权重，该方法通过建立起邻域空间结构与卷积核权重之间的关系，使得卷积操作可以更好地感知邻域结构。本文通过实验证明了基础网络设计的合理性，并同时将空间自适应权重应用于基础网络的卷积操作中以证明空间自适应权重的有效性。最后，本文提出了空间信息引导卷积 (S-Conv)。S-Conv 在空间自适应权重的基础上，其卷积核分布随空间信息自适应变化，进而增强网络的空间变换适应能力和感受野调节能力，进一步达到高效利用空间信息的目的。该方法仅仅需要增加少量的参数量和计算量，即可显著提升语义分割的精度。本文通过实验将 S-Conv 与其他动态卷积方法进行对比，包括可变形卷积[83], 深度感知卷积[47] 等。本文将 S-Conv 引入到基础网络中，名为空间信息引导卷积网络 (SGNet)，使得 SGNet 在保持快速推理速度的基础上，通过引入 S-Conv 达到精确的语义分割结果。本文与当前的主流方法在 NYUDv2[84]，SUNRGBD[85, 86] 数据集上进行对比，并取得了最快速先进

的结果。本文同时通过可视化 SGNet 与基础网络在 NYUDv2 测试集上的比较结果，充分的说明 S-Conv 的有效性和高效性。最后，本文对 S-Conv 在网络不同层的感受野进行可视化来直观的说明 S-Conv 的工作原理。

总结来说，本文的主要贡献如下：

- 对于多模态任务来说，例如 RGBD 语义分割，本文提出了一种创新的视角和方法。具体来说，与其他双流网络方法相比，本文通过利用空间信息引导卷积过程的方式来进行模态融合，极大的减少了网络的参数量和计算量，并达到了更优的效果与实时的推理速度。这为模态融合的方法提供了一种新的思路。
- 本文提出了一种创新的 S-Conv 卷积操作。S-Conv 可以通过输入的 3D 空间信息分析图片中物体的尺度和空间变换，并以此调整网络卷积核的权重与采样分布来感知物体的几何形状。它仅仅使用了少量的参数量和计算量即可显著的提升语义分割的精度。基于 S-Conv，本文提出的 SGNet 可以在 NYUDv2 和 SUNRGBD 数据集上达到实时推理速度，并取得最优的结果。

第四节　论文结构安排

本文的工作为探索充分高效地空间信息利用方法来改善提升语义分割的精度。首先本文设计了一个高效的语义分割基础网络，接着引入空间自适应权重，提升基础网络的语义分割精度。最后以空间自适应权重为基础，设计了一种利用空间信息指导卷积过程的方法，名为空间信息引导卷积 (S-Conv)。该方法可以利用空间信息显著提升语义分割的精度。相比于双流网络方法，该方法仅仅需要增加少量的参数量和计算量。本文将 S-Conv 应用于基础网络，提出了一种实时语义分割网络，名为 SGNet。SGNet 可以达到实时推理速度，并且在公开数据集上，例如 NYUDv2 与 SUNRGBD，性能超过了当前所有的双流网络方法和其他方法。本文总计分为四章，其中第二章第三章为本文方法部分的介绍，关系图如图 1.3所示。整体的论文结构安排如下所示：

第一章为绪论，本文首先介绍 RGBD 语义分割任务的研究背景及意义，接着介绍了 RGBD 语义分割相关任务的国内外研究现状：包括语义分割，RGBD 语义分割，3D 点云的识别，CNN 的动态结构等。最后说明了本文的研究内容以及结构安排。

第二章，本文介绍 RGBD 语义分割的基础知识，包括常用的网络结构，损

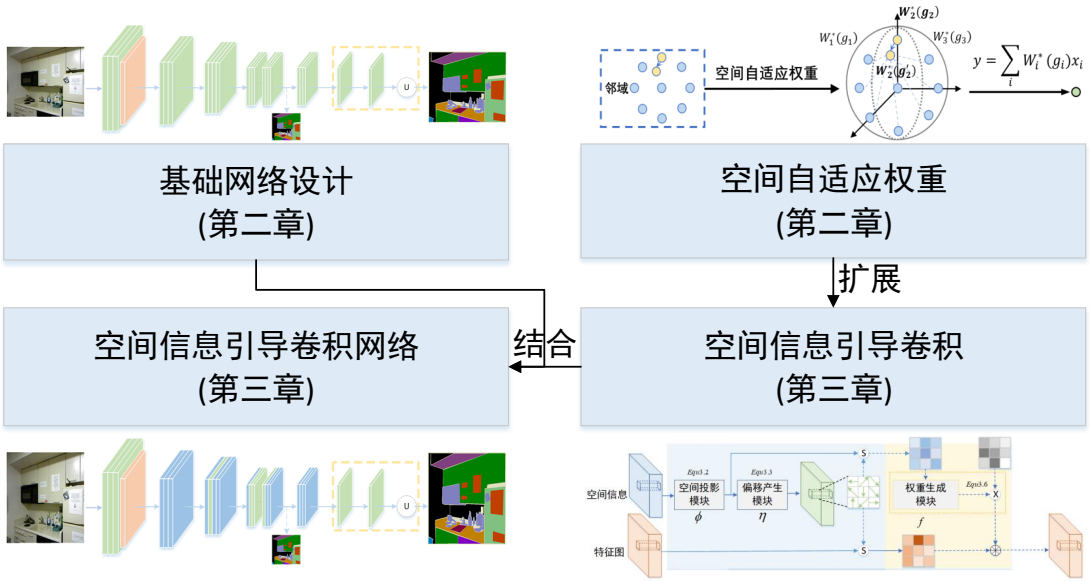

图 1.3 本文的章节安排，以及各个章节中方法之间相互关系的示意图。

失函数，评测指标等。本文设计了一个基础网络，该网络在满足实时性能的同时，可以达到较好的分割结果。本文通过实验来说明基础网络设计的合理性，包括推理速度与性能测试，消融实验等，并替换基础网络的解码器来进一步提升网络的性能。接着，本文通过在不同层的卷积引入空间自适应权重改善基础网络的结果，并通过实验分析利用空间信息的效率。同时将空间自适应权重与深度感知卷积[47] 等方法进行对比。最后，本文将展示分割结果，来进一步直观感受自适应权重对空间信息的特性。

第三章，以自适应权重为基础，本文提出一种高效利用空间信息的操作，名为空间信息引导卷积 (S-Conv)。该操作仅需要少量的计算量和参数量即可提升语义分割的精度。本文同时将 S-Conv 引入基础网络中，名为空间信息引导卷积网络 (SGNet)。本文通过实验分析 S-Conv 利用空间信息的效率，并将 S-Conv 与可变性卷积[82, 83], 深度感知卷积[47] 等方法进行对比。本文将 SGNet 与当前最先进的方法在 NYUDv2, SUNRGBD 数据集上进行比较，证明 S-Conv 的有效性和高效性。本文展示了 SGNet 与基础网络在 NYUDv2 测试集上的比较结果，说明了 S-Conv 利用空间信息改善语义分割性能的特点。最后本文对 S-Conv 在网络不同层的感受野进行可视化，直观地展示 S-Conv 的工作原理。

第四章中，本文总结了前三章的工作，并对未来的工作进行展望。

第二章 语义分割基础网络与空间自适应权重

第一节 卷积神经网络与语义分割

2.1.1 背景知识介绍

语义分割是计算机视觉领域研究的热点问题，其本质是对输入图片进行像素级别的分类。随着卷积神经网络在 ImageNet[6] 分类任务上的成功应用，FCN[8] 被提出，成功地将纯卷积架构应用在语义分割的任务上，并成为语义分割架构的基础模板。FCN 由一个编码器 (encoder) 和一个解码器 (decoder) 组成，为了增大网络的感受野，解码器将输入图片编码成特征图，输出特征图的分辨率为输入图片的 1/32 (FCN 包含 5 个步长为 2 的池化层)。因此，解码器 (decoder) 需要将特征图上采样 32 倍，导致分割结果不够精细。为了解决这一弊端，目前科研人员主要提出了两类方法，如图 2.1所示: 一类方法使用空洞卷积，在增大感受野的同时，可以保持特征图的分辨率，使得图片的细节部分被保留。代表工作有[13–16]。这类方法相较于 FCN 可以获得更精细的分割结果。另一类方法，由于中间特征图分辨率较低，而底层特征图的分辨率较高，因此可以使用编码器 (encoder) 的底层特征来指导解码器 (decoder) 的预测图细节恢复，主要代表工作有[17–22]。这类方法取得了较好的结果，但需要耗费相对较多的计算资源与成本。其中 Deeplabv3+[18] 网络将两种方法结合并取得了更好的结果。但网络计算量较大，不适合实时应用场合。

目前的语义分割网络大体上都在这两类方法上进行改进。针对实时场合的应用，本文采用第一类，即采用空洞卷积为基础的方法来进行网络的设计。

2.1.2 基础网络设计

本小节中，本文将详细介绍基础网络的设计思想。本文设计的基础网络结构如图 2.2所示，具体细节如下：1) 主干网络 (网络的编码器 (encoder) 部分) 采用 ResNet101[87]，同时在第三层和第四层中使用空洞卷来提升网络感受野并保证特征图的空间分辨率。2) 为了提升网络推理速度，整个网络的输出步长为 16。为了优化网络速度，解码器部分由两个 3×3 卷积和一个双线性插值上采样操作

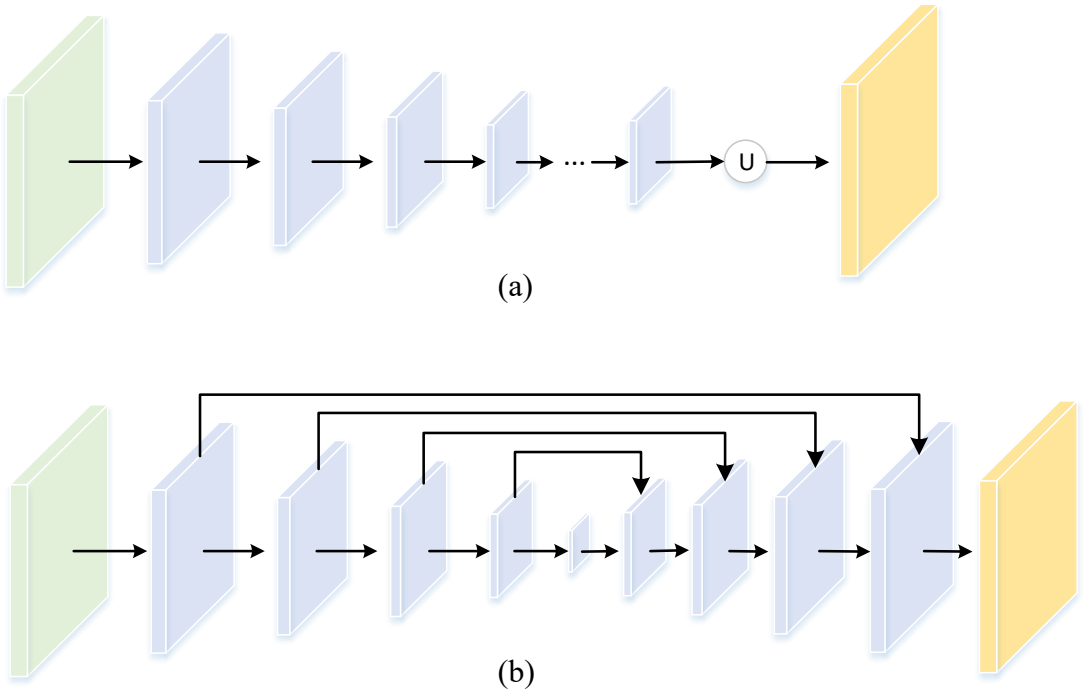

图 2.1 语义分割任务的主流架构：(a) 基于空洞卷积方法的网络架构。该架构保证了特征图分辨率的同时，利用空洞卷积增大网络感受野。(b) 基于编码器-解码器的架构。该架构的中间特征分辨率较小，保证了足够大的感受野，但由于低分辨率的特征图，很难得到精细的结果。因此使用了编码器的浅层特征来指导解码器对特征图分辨率进行修复。该架构有着较高的计算量和参数量。

组成。3) 为了提升网络优化性能，作者在第三层和第四层中间添加一个深层监督信号。该基础网络能够保证快速推理速度的情况下，达到较好的语义分割结果。本文将在接下来的实验中对该网络的性能进行分析。

同时，为了尽可能的提升基础网络的速度，编码器的设计较为简单，仍有进一步的提升空间。为了进一步提升网络的多尺度感知能力，相对于图 2.2 中较为简单的解码器 (decoder)，本文可以通过修改解码器的网络结构，牺牲一些推理速度来换取进一步的性能提升。常用的解决方案为 Deeplabv3[23] 中的 ASPP (Atrous Spatial Pyramid Pooling) 模块与 PSPNet[27] 中的 PPM (Pyramid Pooling Module) 模块。ASPP 模块与 PPM 模块的网络结构如图 2.3所示：

ASPP 由不同膨胀率 (dilated rate) 的空洞卷积并行组成，每一个分支都有着不同大小的感受野。同时，通过独立的分支引入全局池化 (global pooling) 来作为图片全局的表示。最后 ASPP 将不同感受野的分支进行合并，并通过 1×1 卷

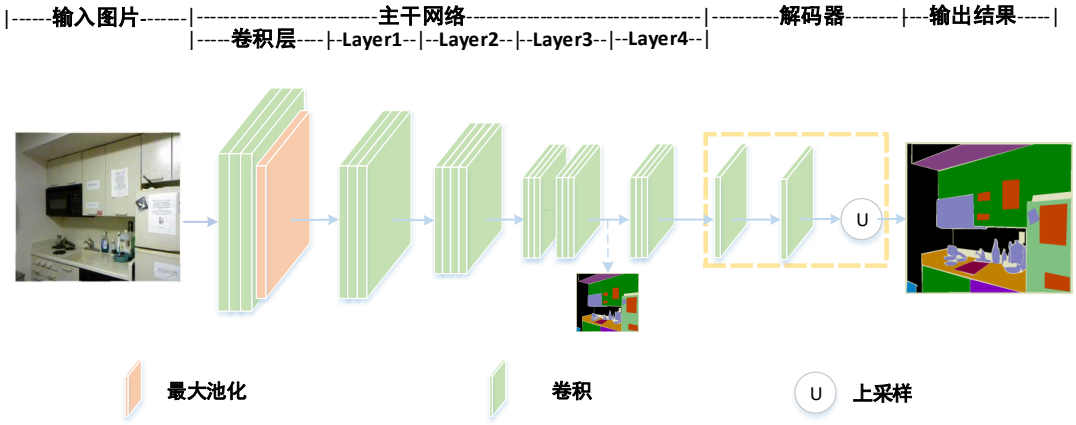

图 2.2 室内语义分割基础网络结构。虚线部分代表额外的监督信号，主干网络采用 ResNet101。

积进行融合，达到多尺度感知的目的。PPM 模块通过引入每一分支不同尺度的池化操作来对特征图进行由局部到全局的表示。这两种模块设计从思想上都考虑了由局部到整体的过程。本文接下来将通过对比实验详细验证两个模块的解码效果。

2.1.3 评测指标与损失函数

本文首先对 RGBD 语义分割任务的评测指标进行介绍，之后介绍在语义分割任务中常用的损失函数。

在语义分割领域中，通常以准确率 (Acc), 平均准确率 (mAcc), 平均交并比 (mIoU) 作为评测指标，其定义如公式 (2.1)：

$$\begin{aligned} Acc &= \sum_i \frac{p_{ii}}{g}, \\ mAcc &= \frac{1}{p_c}\sum_i \frac{p_{ii}}{g_i}, \\ mIoU &= \frac{1}{p_c}\sum_i \frac{p_{ii}}{g_i + \sum_j p_{ji} - p_{ii}}, \end{aligned} \tag{2.1}$$

其中 p_{ij} 为预测为类别 j，真实类别为 i 的像素数量, p_c 类别数目, g_i 是真实类别为 i 的像素数量。$g = \sum_i g_i$. 一般语义分割任务中，通常使用交叉熵作为损失函

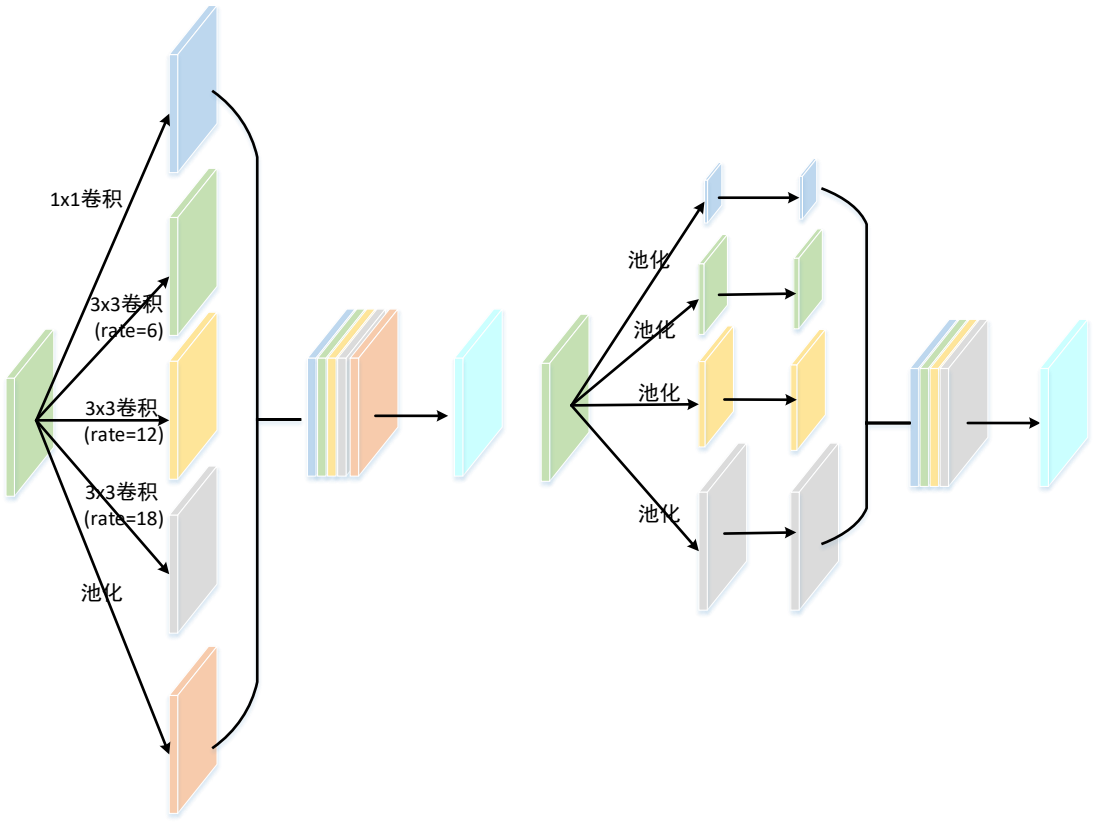

图 2.3 ASPP 模块与 PPM 模块的对比。(a) 为 Deeplabv3 中的 ASPP 模块，(b) 为 PSPNet 中的 PPM 模块。

数，其定义如下：

$$L = -\sum_{i}^{c} y_i log(p_i), \tag{2.2}$$

其中 c 为类别数，y 为真实标签，p_i 为网络预测类别为 i 的概率。然而对于类别数目严重不平衡的数据集，例如 SUNRGBD[85, 86], 会由于数据集中类别不平衡严重影响数目较少类别的分类结果。因此，本文通常会在不同的类别上施加权重来提升小类别的分割精度：

$$L = -\sum_{i}^{c} w_i y_i log(p_i), \tag{2.3}$$

其中 w_i 由不同类别的数量比值计算得到。通过对不同类别的损失施加权重可以缓解由于数据集类别不平衡带来的负面影响。

第二节 空间信息介绍

2.2.1 常用的空间信息结构

目前 RGBD 语义分割领域使用的空间信息主要分为深度图 (depth map)，HHA[39]，与 3D 坐标[68]。本文接下来介绍这常见的三种格式，以及各自的特点。

深度图也被称为距离影像，通常描述的是像素在相机坐标系下 z 轴的坐标，也通常被称为 2.5D 图像。其像素的 3D 坐标可以通过相机内参和深度图恢复。目前深度图主要通过双目摄像机或者深度相机获得。双目摄像机主要通过立体匹配与三角测量来获得像素的深度值。深度相机主要通过飞行时间法 (TOF) 或者结构光的方法来获取深度值，但深度相机获取的距离影响受光照影响明显，对于室外场景的深度，估计偏差较大。

HHA[39] 为深度图 (depth map) 的编码结果，由三通道组成 (水平视差，对地高度，法向角度)。该编码图像首先被[39] 采用，之后被大量应用在 RGBD 语义分割任务上[2, 4, 5, 8, 39]，并相对于深度图，取得了更好的效果。然而从深度图编码到 HHA 图片十分消耗计算资源与时间[2]。因此以 HHA 作为空间信息输入的网络很难部署到实时应用场景。

3D 坐标信息通常被点云识别与分割任务采用[63, 64, 68]，同时在 RGBD 语义分割任务上也被用来作图网络的构建。3D 坐标信息可以通过相机内参和深度图转换得到。转换公式如下所示：

$$\begin{aligned} x &= \frac{(u - c_x) * z}{f_x}, \\ y &= \frac{(v - c_x) * z}{f_y}, \\ z &= z, \end{aligned} \tag{2.4}$$

其中 x,y,z 为空间坐标值，f_x,f_y 为相机的焦距，c_x,c_y 为相机光心的坐标，u,v 为像素坐标值。本文通过上述式子即可利用深度图来恢复像素的 3D 空间坐标信息。不同空间信息的示例图如图 2.4所示。

人眼可以根据左右眼看到的图像差异来估计场景的空间结构。在室外场景数据集中 (例如 Cityscapes[88]) 提供经过矫正的双目相机的左右视图。通过左右视图的匹配点，本文可以计算左右视图的视差图，并依次得到图片中每个像素

RGB 图　　深度图　　HHA

图 2.4　常用的空间信息结构。从左到右分别为：RGB 图，深度图，HHA。

的深度信息，计算公式如下：

$$Z = f\frac{T}{x_r - x_l}, \tag{2.5}$$

其中 Z 为深度信息，f 为相机的焦距，T 为左右相机的距离。x_r, x_l 分别为左右相机匹配的像素点。本文通过公式 (2.5) 可以利用视差图计算得到深度图。由于视差最小为一个像素，因此理论上存在一个深度的最大值。

在之前的工作中，主要使用的空间信息结构为由深度图 (depth map) 编码得到的 HHA。相较于深度图 (depth map) 来说，HHA 取得了相对更好的结果。然而这种空间信息结构很难应用到实时场景中。这同时也说明在之前的工作中，网络不能充分从深度图中提取对应的 HHA 特征。

本文在接下来的章节中，将通过实验探索使用不同形式的空间信息结构对网络预测结果的影响。

2.2.2 常用的空间信息利用方法

随着 3D 传感器技术的发展，空间信息变得易于获取。额外的空间信息输入对于提升神经网络的语义感知能力十分重要。归功于高效的卷积神经网络架构与和易于获取的空间信息，最近的方法在室内场景语义分割的任务上取得了很好的结果。然而由于室内环境的复杂性，以及额外输入的空间信息所需的算力，特别对于需要实时推理的应用，仍然是一个巨大的挑战。

为了得到深度图的特征，这些方法会采用额外的主干网络提取深度图的空间特征，最后采用不同的方法与 RGB 特征进行融合，取得了较好的效果。但双流网络架构极大的增加了参数量和计算量。对于一些实时性要求较高的应用场合，例如机器人，自动驾驶等，这类方法的落地性受限。同时，一些方法[2, 4, 5, 8, 39]将原始的空间信息 (深度图) 编码成由水平视差，对地高度, 和法向角度组成的 HHA 信息，然而这种编码过程十分消耗时间[2]。一些方法将深度图投影到 3D 空间，并采用 3D CNN 或者图网络来提取空间特征，然而这种方法需要相对较多的计算成本与推理时间。也有一些工作利用多任务学习 (利用空间信息作为监督) 来利用空间信息。

另外，室内场景物体的颜色和纹理通常不具有代表性。而物体的几何形状通常在语义感知中起到关键作用。举例来说，冰箱和墙壁具有相似的纹理，为了识别冰箱和墙壁，区分点在于两者的几何形状，而不是两者相似的纹理。然而 2D 卷积并不能充分考虑到这种几何形状区别，很容易受到两者相似纹理的干扰。因此，通过合理的利用空间信息即可解决这类问题。深度感知卷积[47] 被提出来解决这个问题，它迫使与卷积核中心深度相似的像素具有比其他像素更高的权重。由于在权重中融入了空间信息，深度感知卷积具有一定的空间感知能力。然而这种先验是人工手动设定的，并非从训练样本中学习得到。因此，当室内环境的空间关系变得更为复杂时，该方法的性能相对受限。该领域需要一种更充分高效利用空间信息的方法。

为了探索更高效的空间信息利用方式，克服分别提取空间信息与 RGB 图片特征的效率瓶颈，以及动态复杂的室内场景变化与固定结构的卷积操作之间的矛盾。本文从卷积操作的本身入手，旨在将空间信息充分融入到卷积操作中去，来使得卷积操作本身更好的感知几何结构，从而达到高效利用空间信息的目的。因此本文提出了空间自适应权重，建立起权重与空间信息之间的关系，使得卷积核的权重随着其本身空间信息自适应变化，进而使得卷积操作更好地感知邻

域的几何形状和空间结构。

为了减少参数量与计算量，2D 卷积采用了权重共享策略，即图片中每个位置的卷积核权重共享参数。这种做法极大的减少了参数量，但不能充分地感知动态变化的几何结构。比如，2D 卷积并没有尺度不变性，因此网络需要大量的训练数据来进行拟合。空间自适应的核心思想是，建立起邻域空间结构与其卷积核权重的联系，使得卷积核权重随着邻域空间结构自适应变化，从而更好地感知邻域内的空间结构，同时达到高效利用空间信息的目的。本文在图 2.5举例对比了 2D 卷积与空间自适应权重的区别。2D 卷积可以被视为空间中的像素点放在固定的权重空间中，忽略了像素点之间的相互关系与空间信息。黄色的点即将改变空间位置的像素点，其对应的权重不会发生变化。所以 2D 卷积不能捕捉到这个几何形状的变化。空间自适应权重可以视为将邻域内的像素点放置在了一个 3D 的空间权重中，因此每个像素点的权重都与其自身和其他点的空间关系有关。当图中黄色的像素点空间位置发生变化之时，权重随即发生变化，因此卷积的输出也会变化。网络可以捕捉到这种差异，从而更好地感知邻域内的几何形状。本文接下来将通过公式说明，空间自适应权重的技术细节。

第三节 空间自适应权重

2.3.1 2D 卷积

本文首先回顾 2D 卷积的流程。本文使用 $\mathbf{A}_i(\mathbf{j}), \mathbf{A} \in \mathcal{R}^{c\times h\times w}$ 来指示一个张量。为了方便说明，本文将非标量的符号加黑。其中 i 为第一维度的索引，$\mathbf{j} \in \mathcal{R}^2$ 为第二维度与第三维度的索引。

本文使用 $\mathbf{X} \in \mathcal{R}^{c\times h\times w}$ 来指代卷积的输入特征。为了方便说明，本文将在 2D 的情况下进行讨论，所以 $c = 1$。将其推广到 3D 情况是很直接的。空间自适应权重在整个通道维度上的操作保持一致。传统卷积输入 $\mathbf{X}$ 得到 $\mathbf{Y}$ 的过程如公式 (2.6) 所示：

$$\mathbf{Y}(\mathbf{p}) = \sum_{i=1}^{K} \mathbf{W}_i \cdot \mathbf{X}(\mathbf{p} + \mathbf{d}_i), \tag{2.6}$$

其中 $\mathbf{W} \in \mathcal{R}^K$ 为卷积核的权重，大小为 K, 其中 $K = k_h \times k_w$，k_h,k_w 为卷积核的尺寸。$\mathbf{p} \in \mathcal{R}^2$ 为卷积核的中心，$\mathbf{d} \in \mathcal{R}^{K\times 2}$ 为卷积核的空间分布。举例来说，对

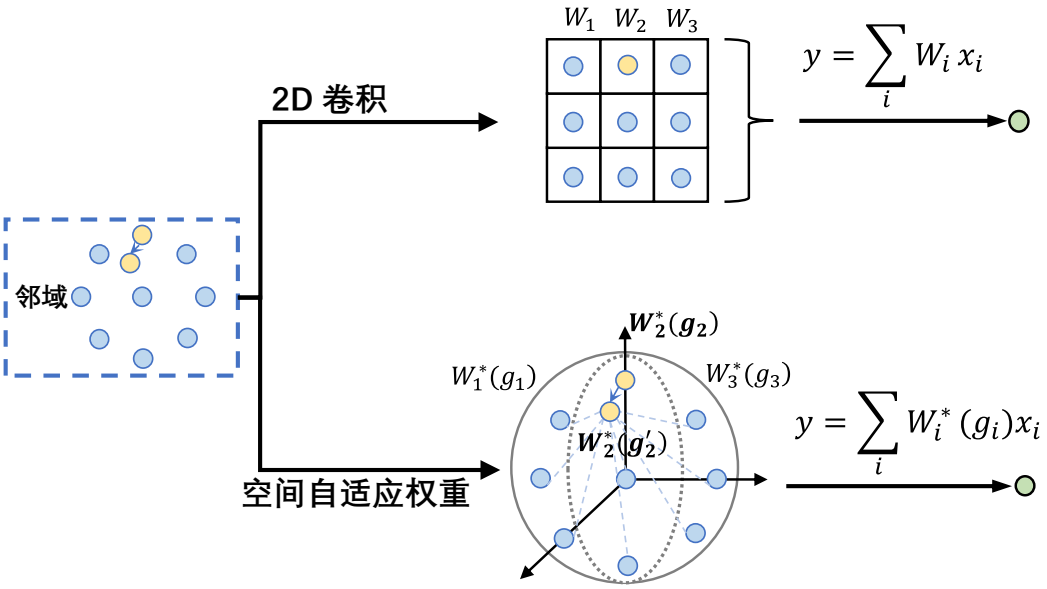

图 2.5 2D 卷积与空间自适应权重 W^* 之间的关系。黄色的像素点代表空间位置将随着箭头改变的点。2D 卷积的说明在上方，空间自适应权重在下方。可以看到，当黄色像素点的空间位置发生变化时，2D 卷积的输出保持不变，因此不能捕捉到像素空间位置的变化。而引入空间自适应权重后，其对应的卷积核权重随像素空间位置自适应变化，输出也会因此自适应变化，进而更好的感知邻域的空间结构。

于 3×3 卷积来说，d 为：

$$\mathbf{d} = \{[-1,-1],[-1,0],...,[0,1],[1,1]\}. \tag{2.7}$$

由式公式 (2.6) 可得，对于不同的卷积中心 p，d 的分布和 W 的权重保持不变，并与图像中物体的尺度和空间变换无关。本文希望在 RGBD 场景下，能够将物体的空间信息引入到卷积过程中，使得卷积对空间信息自适应。即，权重 W 能够随着输入图片中的物体尺度与空间变换自适应变换，进而提升网络的空间结构理解能力与泛化能力。

2.3.2 空间自适应权重应用于 2D 卷积

本文基于上述想法提出了空间自适应权重。通过将空间自适应权重引入到 2D 卷积操作中提升卷积操作的空间感知能力。卷积需要两个输入：一个是同卷积一样的特征图 $\mathbf{X}$，另一个是空间信息 $\mathbf{S} \in \mathcal{R}^{c' \times h \times w}$。$S$ 可以为深度图 ($c' = 1$)，

空间坐标图 ($c' = 3$) 或者 HHA 图 ($c' = 3$)。输入特征图 $\mathbf{X}$ 中并不包含空间信息。

首先本文对输入的空间信息 $\mathbf{S}$ 进行升维，方便后续处理：

$$\mathbf{S}' = \phi(\mathbf{S}), \tag{2.8}$$

其中 ϕ 为空间转换函数，本文可以用卷积网络实现。$\mathbf{S}' \in \mathcal{R}^{64\times h\times w}$，其中 S' 的维度高于 S。S' 为 S 升维后的结果。S' 中包含着输入图片中物体的高阶空间信息，例如物体的尺度，空间变换或者形变等等。接下来本文依靠这些信息生成卷积核的自适应权重。本文首先采集中心点邻域内像素的空间信息：

$$\mathbf{S}^*(\mathbf{p}) = \{\mathbf{S}'(\mathbf{p} + \mathbf{d}_i)|_{i=1,2,...,K}\}, \tag{2.9}$$

其中 $\mathbf{S}^*(\mathbf{p}) \in \mathcal{R}^{64K}$ 对应着卷积核内所有像素的空间特征。本文根据这些空间特征生成卷积核的空间自适应权重：

$$\mathbf{W}^*(\mathbf{p}) = \sigma(f(\mathbf{S}^*(\mathbf{p}))) \cdot \mathbf{W}, \tag{2.10}$$

其中 $\mathbf{W} \in \mathcal{R}^K$ 为卷积核权重，可以通过梯度下降法更新。$\mathbf{W}^*(\mathbf{p}) \in \mathcal{R}^K$ 对应着卷积的空间自适应权重，其权重由位置 p 对应的空间信息决定。f 为非线性函数，可以由全连接网络实现。σ 为 Sigmoid 激活函数。

本文最后引入空间自适应权重的卷积操作如下：

$$\mathbf{Y}(\mathbf{p}) = \sum_{i=1}^{K} \mathbf{W}_i^*(\mathbf{p}) \cdot \mathbf{X}(\mathbf{p} + \mathbf{d}_i)), \tag{2.11}$$

其中 $\mathbf{W}_i^*(p)$ 建立起了空间关系与权重之间的关系，其值取决于 p 位置处的空间特征 (物体的尺度，形状等特征)。相比于 2D 卷积，本文引入了空间自适应权重 $\mathbf{W}_i^*$，该权重可以随着输入空间信息自适应变化，从而有着更强的几何感知能力。

第四节 实验结果对比与分析

在这一小节中，本文进行 RGBD 语义分割任务的实验。首先，本文对章节 2.1.2中提出的基本模型进行速度和性能测试，并对其结构设计消融实验，验证本文设计的合理性。接着，本文将空间自适应权重引入基础模型中的每一层的卷积中，通过实验测试其结果提升。随即，本文测试不同种类的空间信息对

引入自适应权重网络结果的影响。最后本文展示引入空间自适应权重的网络的分割结果来说明空间自适应权重利用空间信息改善语义分割结果的特点。

2.4.1 实验介绍

数据集和评测指标：本文在下列数据集上验证基础网络和其引入空间自适应权重后在 RGBD 语义分割任务中的性能。

- NYUDv2[84]：这个数据集包括 1,449 张 RGB 图片和对应的深度图与分割标注图。遵循之前的方法[89]，本文使用 795 张图片用于训练，654 张图片用于测试。本文使用 40 类别的设置来进行实验。

本文使用公式 (2.1) 来作为评测指标，即精度 (Acc), 平均精度 (mAcc), 和平均交并比 (mIoU)。

实现细节：对于 RGBD 语义分割任务，本文使用 ResNet101[87] 作为特征提取的主干网络，并在 ImageNet[6] 上进行预训练。本文默认设置整个网络的输出步长为 16，并使用 Pytorch 深度学习框架来实现整个语义分割系统。优化器设置为随机梯度下降 (SGD), 并使用"poly" 的学习率衰减策略, 这个衰减策略在 NYUD 数据集上每迭代 40 轮应用一次。网络的初始学习率设为 5e-3，网络的权重衰减设置为 5e-4。网络的激活函数默认设置为 ReLU，训练批大小默认设置为 8，网络使用语义分割任务中常见的数据增广策略，包括随机尺度变换，随机剪裁，随机反转。剪裁大小为 480×640。其中随机尺度变换的尺度范围为 [0.5, 2.25]。本文在测试阶段将图片下采样到训练时剪裁的大小，即 480×640。之后将预测的结果上采样到输入大小，上采样的方法为双线性插值。空间自适应权重的实现从可变形卷积[83] 修改而来。

2.4.2 基础网络测试

首先，本文测试基础网络在 NYUDv2[84] 数据集上的基础性能，接着本文将针对基础网络设置消融实验来验证各个设置的有效性，包括空洞卷积，额外监督。最后，本文将替换基础网络的编码器模型来验证 ASPP, PPM 模块的有效性。

基础网络的测试：本文的基础网络结构如图 2.2所示，其中主干网络采用 ResNet101[87], 网络的输出步长默认为 16，并在第三层和第四层中间添加一个深层监督信号。本文在 NYUDv2 训练数据集上进行训练，在测试数据集上进行测试，本文的速度测试在 NVIDIA 1080TI 硬件环境下进行，每一类别的结果和速度如表 2.1所示：

表 2.1 在 NYUDv2 测试集基础网络的平均交并比结果。mIoU: 平均交并比，FPS：帧数。表中结果为百分数。

墙壁	地板	储柜	床	椅子	沙发	桌子	门	窗户	书架	图片
77.4	82.0	57.3	65.1	54.5	56.6	41.3	42.7	45.4	41.6	58.5
柜台	窗帘	桌子	架子	帘子	橱柜	枕头	镜子	垫子	衣服	顶板
59.4	58.8	15.7	17.3	48.2	48.2	34.6	39.2	24.9	19.4	68.6
书籍	冰箱	电视	纸张	毛巾	喷头	箱子	板子	人	货摊	马桶
29.5	52.8	61.6	25.0	30.5	17.0	9.4	65.5	66.0	40.9	69.3
碗槽	台灯	浴缸	背包	其他 1	其他 2	其他 3			mIoU	FPS
51.2	38.1	26.9	5.6	28.5	12.3	34.1			43.0	34

表 2.2 基础网络在 NYUDv2 测试集上的消融实验。DC: 空洞卷积，AS：额外的监督信号。

AS	DC	准确率 (%)	平均准确率 (%)	平均交并比 (%)
		69.3	50.9	39.3
✓		69.9	51.0	39.7
✓	✓	**72.1**	**54.6**	**43.0**

作者发现，本文的基础网络推理时间较为快速，同时取得了较为不错的实验结果。但是在镜子等可以反射其他物体纹理的类别，和一些不具有代表性纹理的类别，例如冰箱和浴缸，还有一定的提升空间。这两种类别由于其相似的纹理，需要从几何形状入手来进行判断，但基础网络缺乏一定的空间感知能力。本文期望通过高效的利用空间信息来提升网络针对这些类别的表现。

消融实验：本文将验证空洞卷积，额外的监督信号对基础网络的结果影响。其中空洞卷积用来扩大网络的感受野，额外的监督信号可以提升网络的优化能力。其网络配置和训练细节与基础网络保持一致，实验结果如表 2.2所示：

本文通过实验结果发现，1) 空洞卷积对提升网络的感受野，改善网络的结果有着正向作用。2) 额外的监督信号对网络的优化和性能有正向作用。因此，本文的基础网络采用空洞卷积和额外的监督信号配置。上述实验进一步验证了本文基础网络配置的合理性。但为了优化网络的速度，基础网络的编码器设置较为简单。接下来，本文将优化基础网络的编码器设置，并测试对于网络的结果提升。

ASPP 与 PPM：为了进一步提升网络的多尺度感知能力，本文将基础网络的编码器替换为 ASPP[23] 模块与 PPM[27] 模块，这两类模块分别利用不同尺度的空洞卷积与池化操作来提升网络的多尺度感知能力。本文分析这两类模块的

表 2.3 基础网络的解码器消融实验。origin: 简易解码器，PPM：PSPNet 中的 PPM 模块，ASPP：Deeplabv3 中的 ASPP 模块。

origin	PPM	ASPP	准确率 (%)	平均准确率 (%)	平均交并比 (%)
✓			72.1	54.6	43.0
	✓		72.1	55.0	43.3
		✓	**72.4**	**55.9**	**43.7**

参数量，性能与时耗，实验结果如表 2.3所示：

作者发现，ASPP 与 PPM 都对结果有一定的正向提升。ASPP 模块的性能相对于 PPM 更好，在增加了少量和参数量的同时，提升了网络的性能。为了保持基础网络的速度，在接下来的实验中，本文默认不使用额外的多尺度解码器(ASPP)。在后面的章节中，本文会使用 ASPP 模块进一步提升网络的性能。

2.4.3 空间自适应权重测试

空间自适应权重替代实验：从上述实验结果中，作者发现基础网络可以在保持实时推理速度的同时，达到较为理想的分割结果，但是在一些类别的表现上，例如不具有代表性纹理的冰箱和浴缸，或者可以反射其他物体纹理的镜子，还有一定的提升空间。为了区分这两种类别，网络需要一定的空间感知能力。因此，本文将空间自适应权重引入到基础网络的卷积操作中来验证空间自适应权重的效果。在这一小节中，本文通过在网络的不同阶段将空间自适应权重引入卷积层探索其对网络结果的影响。本文在网络的每一层的后 3 个卷积操作中引入空间自适应权重。实验结果如表 2.4所示。

作者发现，基础网络的推理速度最快，但分割性能相对较弱，若将空间自适应权重引入卷积操作，会显著的提升语义分割的精度。同时，本文通过实验发现，在前三层引入空间自适应权重，效果最好。由于第四层对应的空间信息分辨率较低，因此生成的自适应权重不能有效地提升网络的性能。所以本文在接下来的默认在 Layer1, Layer2, Layer3 中的后三个卷积操作中引入空间自适应权重。网络中没有显式的空间信息输入，空间信息仅仅影响网络的权重。

空间自适应权重测试：接下来，本文将验证空间自适应权重在不同类别上的性能提升。网络默认使用深度图作为空间信息输入。本文将基础网络前三层中的后三个卷积操作引入空间自适应权重，并分析在 NYUDv2 测试数据集上每一类别上性能的提升。实验结果如表 2.5所示，网络引入空间自适应权重后在每

表 2.4 网络在不同层引入空间自适应权重在 NYUDv2 测试集上的表现。

Layer1	Layer2	Layer3	Layer4	准确率 (%)	平均准确率 (%)	平均交并比 (%)
				72.1	54.6	43.0
✓				74.1	57.7	46.2
✓	✓			74.1	58.2	46.3
✓	✓	✓		**74.4**	**58.8**	**46.9**
✓	✓	✓	✓	74.0	57.8	46.0

表 2.5 基础网络引入空间自适应权重后在 NYUDv2 测试集基础网络的结果。mIoU: 平均交并比，FPS：帧数。表中结果为百分数。

墙壁	地板	储柜	床	椅子	沙发	桌子	门	窗户	书架	图片
79.0	85.7	60.0	69.6	60.7	61.7	43.0	41.3	51.2	44.5	61.2
柜台	窗帘	桌子	架子	帘子	橱柜	枕头	镜子	垫子	衣服	顶板
64.8	61.9	18.0	19.3	51.6	40.4	39.8	50.9	30.0	19.6	75.2
书籍	冰箱	电视	纸张	毛巾	喷头	箱子	板子	人	货摊	马桶
31.3	54.5	62.2	26.8	35.7	16.1	10.2	78.8	72.4	46.5	72.8
碗槽	台灯	浴缸	背包	其他 1	其他 2	其他 3			mIoU	FPS
57.1	44.5	44.7	10.2	30.8	13.7	37.5			46.9	28

一类别上的提升如图 2.6所示。作者通过实验结果发现，在引入了空间自适应权重后，网络在大部分类别上的性能表现均有提升，特别是镜子等具有反射其他物体纹理性质类别，和不具有代表性纹理的类别 (浴缸，板子)，交并比有显著的提升。同时网络能够保持较快的推理速度。因此可以看出，本文的空间自适应权重可以通过高效充分的利用空间信息提升网络在语义分割任务上的精度。

空间信息的消融实验：在之前的章节中，本文介绍了常用的空间信息的格式以及对应的优缺点。在这一小节中，本文探索在应用空间自适应权重中使用不同种类的空间信息对结果的影响。本文所使用的空间信息包括：深度图 (depth map)，RGB 特征，HHA，和像素的 3D 坐标，实验结果如表 2.6 所示：

作者发现，使用 HHA 的性能与 3D 空间坐标与深度图相近效果均优于 RGB 特征。与之前的工作中使用 HHA 信息的结果明显优于深度图的结论不同。这说明引入空间自适应权重可以从深度图中充分提取 HHA 特征。然而，本章节提出的空间自适应权重仍有进一步的改进空间。本文将在第三章中对空间自适应权重的方法进行进一步扩展。

与其他备选方案的对比：在上一小节中，本文通过实验验证了不同种类的空间信息对空间自适应权重的影响。本文接下来将空间自适应权重方案与其他

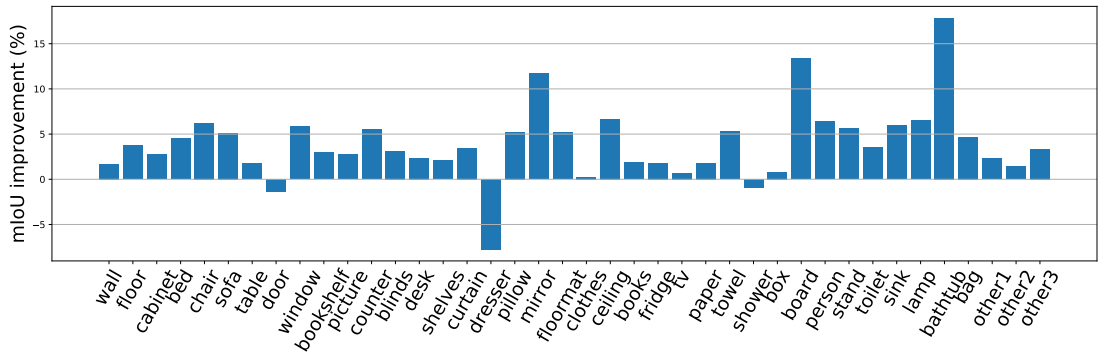

图 2.6 基础网络引入空间自适应权重后，在 NYUDv2 数据集上每个类别的提升。

表 2.6 使用不同的空间信息在 NYUDv2 测试集上的比较结果。

输入空间信息	准确率 (%)	平均准确率 (%)	平均交并比 (%)
深度图	74.6	59.0	47.4
RGB 特征	72.1	55.5	43.7
HHA	**74.8**	**59.3**	**47.5**
3D 空间坐标	74.6	59.1	47.5

方法进行对比。这些方法包括双流网络方法，深度感知卷积方法[47] 等。其中深度感知卷积方法将深度信息作为一种权重的先验，即通过深度信息来影响权重，这与本文空间自适应权重的思路一致，但深度感知卷积通过空间信息影响权重的方式是手工设置的，而并非从数据中学习到。本文的空间自适应权重与空间信息的关系是从数据中学习到的。本文在表 2.7中与其他方法进行了比较。可以看到，本文通过将空间自适应权重引入到卷积层中，其性能超过了双流网络方法与深度感知卷积方法，这证明了本文空间自适应权重的高效性，以及从数据中习得权重与空间信息关系的必要性。

分割结果展示：为了直观展示空间自适应权重的有效性，本文展示了基础网络引入空间自适应权重后，在 NYUDv2 测试集上的分割结果图。如图 2.7所示：可以看到，网络在 NYUDv2 测试数据集上取得了较为精细的分割结果，例如图 2.7 (a) 中具有丰富空间变换的椅子，以及图 2.7 (b,c) 中反射其他物体纹理的镜子等。

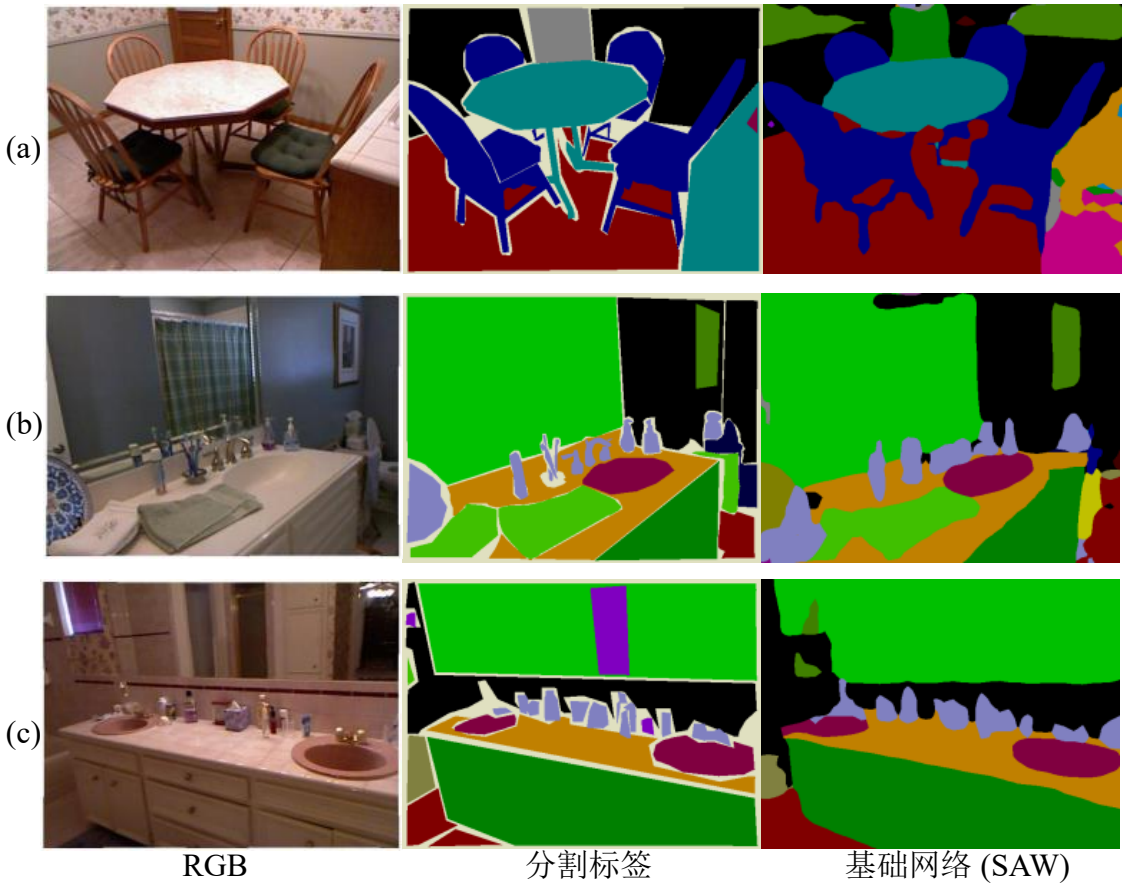

图 2.7 NYUDv2 数据集上的分割结果展示。从左到右，RGB，分割标签，基础网络引入空间自适应权重的结果。

第五节 本章小结

在本章中，本文首先介绍了语义分割任务以及常用的方法，并参考之前的工作，设计了 RGBD 语义分割基础模型，介绍了语义分割的评测指标，并介绍了常用的空间信息结构 (深度图，HHA，3D 空间坐标)。之后，本文介绍了空间自适应权重，并将其应用在 RGBD 语义分割任务。接着，本文测试了基础网络的性能，设计了消融实验，替换不同的编码器来验证本文配置的有效性。最后，本文测试了空间自适应权重应用于基础网络的性能，并将空间自适应权重与其他备选方法进行对比。作者认为，本文提出的利用空间信息指导卷积过程的方法取得了一定的效果。但目前的卷积过程，仍然无法较好的适应物体的空间变换。在接下来的章节中，本文将进一步探索利用空间信息指导卷积操作的方法

表 2.7 NYUDv2 测试集上不同备选方案的比较结果。DAC: 深度感知卷积[47]，SAW: 空间自适应权重，HHANet：使用额外分支 (ResNet101) 提取空间特征并与主干的 RGB 特征在网络的最后阶段合并。

模型	准确率 (%)	平均准确率 (%)	平均交并比 (%)
基础网络	72.1	54.6	43.0
基础网络 +HHANet	73.5	56.8	45.4
基础网络 +DAC	73.8	57.1	45.4
基础网络 +SAW	**74.5**	**58.4**	**46.8**

提升语义分割的精度。本文将利用空间信息增强卷积对物体空间几何变换的适应性。该方法以空间自适应权重为基础，进一步的提升了语义分割任务的精度。本文将通过原理阐述与实验证明方法的有效性。同时本文将该方法引入到基础网络中，并与当前最先进的 RGBD 语义分割方法进行比较。

第三章　基于空间信息引导卷积的实时语义分割网络

深度信息可以补充图像投影到 2D 相机平面过程中损失的深度和尺度信息，对神经网络分析图片的语义信息有着重要的作用。目前语义分割任务主要采用深度学习网络。其中大多数主流方法利用深度信息的方式为双流网络架构，即一路网络提取 RGB 图片的特征，另一路提取深度图的特征，最后使用模态融合的思路融合两路的特征，输出最后的分割结果。这些方法可以通过额外的空间信息输入提升语义分割的结果，但由于采用了双流网络的架构，极大的增大了网络的参数量和计算量，限制了其实时场景下的应用。

本文认为，高效充分的利用空间信息对 RGBD 语义分割的实际应用至关重要。在上一章节中，本文提出了空间自适应权重的思想来高效利用空间信息。在空间自适应权重的基础上，本文首先提出了一种高效利用空间信息的操作：空间信息引导卷积 (**S**patial information Guided **Conv**olution)，简称为 S-Conv。S-Conv 可以将卷积核空间自适应权重与空间自适应分布相结合，并根据图片的空间信息，在卷积的不同位置生成空间自适应变化的卷积核的分布与权重，进而增强网络的空间适应能力与感受野自我调节能力。该操作在增加少量计算量和参数量的情况下，充分高效地利用空间信息提升语义分割的性能。接下来，基于 S-Conv，本文设计了一个实时语义分割网络：空间信息引导卷积网络 (**S**patial information **G**uided Convolution **Net**work)，简称 SGNet。得益于 S-Conv 高效充分利用空间信息的能力，该网络在 NYUDv2 数据集实现了实时推理速度，并达到最优效果。最后，本文将 SGNet 在公开数据集 (NYUDv2[84], SUNRGBD[85, 86]) 上进行验证，均取得最优的实验结果。

第一节　研究动机以及贡献

在上一章中，本文首先参考国内外工作设计了语义分割基础网络，在保持一定推理速度的情况下，在各个数据集上取得了较为良好的结果。同时本文引入了空间自适应权重，高效地利用空间信息并改善语义分割的结果。值得注意的是，室内场景相对于室外场景有着相对更复杂的空间关系。因此，室内场景对网络的空间变换适应性有着更高的要求。然而在大部分深度分割网络中所采

图 3.1 NYUDv2 数据集中的示例图片。从左到右分别为，RGB 图，深度图，分割标签。

取的卷积操作，卷积核采用固定分布，例如 3×3 或者 1×1 卷积。这类方法主要通过加深网络和卷积核堆叠的方法增大感受野，在一些相对简单的场景下取得了较为理想的结果。但其卷积核无法根据输入图片物体的尺度大小和形变自适应的调整感受野和空间变换，只能通过大量的样本学习进行穷举。而在很多场景下，例如室内，大多物体具有复杂的空间变换与远近关系，这对神经网络的空间适应能力提出了很高的要求。如图 3.1中所示的室内场景中的椅子，桌子上的摆件和可以反射纹理的镜子。因此这种方法使得分割网络在室内场景下的泛化能力受限。

可以看到，结构固定的 2D 卷积与物体间动态变化的空间关系存在矛盾，同时也存在着 RGB 与空间信息分开处理的效率瓶颈。因此针对上述问题，结合上一章空间自适应权重的思想，本文提出一个创新的方法，名为 S-Conv (**S**patial information guided **Conv**olution), 其卷积操作随输入空间信息动态变化，从而达到利用空间信息提升语义分割精度的目的。具体来说，S-Conv 卷积操作可以根据

输入的空间信息，在卷积的不同位置生成空间自适应变化的卷积核分布，从而增强网络的空间变换适应能力和感受野调节能力。另外，S-Conv 建立起卷积核权重与其对应像素的空间信息的联系，将空间信息融入卷积核的权重当中，从而更好的感知场景的空间结构。这种方法相比于采用额外的主干网络提取深度图的空间特征，大大减少了计算量和参数量。本文的方法如 图 1.2 (b) 所示：可以看到，相比于双流网络的方法，本文没有使用额外的主干网络来提取空间特征，而是利用深度图中包含的物体尺度和空间变换信息来指导卷积核的空间分布与感受野大小，从而达到高效利用空间信息的目的。由于 S-Conv 中空间信息 (深度图) 的直接输入，图片中物体的尺度和空间变换可以直接被分析出来，并产生对应的空间自适应卷积核分布和权重，从而更好地感知物体的几何形状。

本文提出的 S-Conv 结构轻巧而灵活，只需要少量的额外参数与计算开销即可显著的提升网络的性能，因此十分适用于实时应用。同时，S-Conv 也可被视为一种创新高效的模态融合方法。具体来说，与双流网络方法相比，本文通过使用空间信息指导卷积的过程达到多模态融合的目的。这种方法的性能优于其他基于双流网络的方法，同时极大的减少了参数量和计算量。本文设计了大量的实验说明 S-Conv 的有效性与高效性。本文首先设计了消融实验，将 S-Conv 与双流网络方法，可变形卷积[82, 83], 和深度感知卷积[47] 进行比较，说明 S-Conv 的优势。本文同时测试深度图，HHA 和 3D 坐标等不同类型空间信息的输入对 S-Conv 的影响，验证了 S-Conv 对空间变换的适用性。受益于 S-Conv 对空间变换的自适应能力与空间结构感知能力，本文基于 S-Conv 提出了 SGNet (**S**patial information **G**uided convolutional **Net**work)，在 NYUDv2[84] 数据集和 SUNRGBD[85, 86] 数据集上以实时的推理速度取得了高质量的结果。最后本文可视化了 SGNet 在不同数据集上的分割结果，并将 SGNet 与基线网络进行对比，进一步从直观上阐述了 S-Conv 的有效性。总结来说，本章节的贡献如下：

- 本文提出了一种创新的 S-Conv 操作。该操作可以自适应的调节感受野和和适应空间变换，同时可以以较低的计算资源感知图片内部的空间结构。
- 基于 S-Conv，本文提出了 SGNet。SGNet 可以在 NYUDv2 和 SUNRGBD 数据集上以实时的速度达到最优的分割结果。

接下来，本文将详细介绍 S-Conv 的理论和细节实现部分。

第二节 S-Conv: 空间信息引导卷积

在本节中，本文会介绍 S-Conv (**S**patial information Guided **Conv**olution) 的实现细节，作者发现，S-Conv 可以视为 2D 卷积引入空间信息的推广。本文同时会讨论 S-Conv 与其他的自适应卷积的区别。

3.2.1 S-Conv 原理介绍

与第二章相同，本文使用 $\mathbf{A}_i(\mathbf{j}), \mathbf{A} \in \mathcal{R}^{c\times h\times w}$ 来指示一个张量。为了方便，本文将非标量的符号加黑。其中 i 为第一维度的索引，$\mathbf{j} \in \mathcal{R}^2$ 为第二维度与第三维度的索引。

首先，2D 卷积操作的公式如公式 (3.1) 所示：

$$\mathbf{Y}(\mathbf{p}) = \sum_{i=1}^{K} \mathbf{W}_i \cdot \mathbf{X}(\mathbf{p} + \mathbf{d}_i), \tag{3.1}$$

由式公式 (3.1) 可得，对于不同的卷积中心 p，d 的分布和 W 的权重保持不变，并与图像中物体的尺度和空间变换无关。作者希望在 RGBD 场景下可以将物体的空间信息引入到卷积过程中，使得卷积对空间信息自适应。即：卷积核的分布 d 与权重 W 可以随着输入图片中的物体尺度与空间变换自适应变换，从而提升网络的语义理解能力与泛化能力。基于上述想法，本文提出了 S-Conv。该卷积需要两个输入：一个是与卷积相同的特征图 $\mathbf{X}$，另一个是空间信息 $\mathbf{S} \in \mathcal{R}^{c'\times h\times w}$。$S$ 可以为深度图 ($c' = 1$)，空间坐标图 ($c' = 3$) 或者 HHA 图 ($c' = 3$)。输入特征图 $\mathbf{X}$ 中并不包含空间信息。

首先，在 S-Conv 中，与第二章相同，本文对输入的空间信息 $\mathbf{S}$ 进行升维，方便后续处理：

$$\mathbf{S}' = \phi(\mathbf{S}), \tag{3.2}$$

其中 ϕ 为空间转换函数，本文用卷积网络实现。$\mathbf{S}' \in \mathcal{R}^{64\times h\times w}$，其中 S' 的维度高于 S。S' 为 S 升维后的结果。与第二章的空间自适应权重类似，S' 中同样包含着输入图片中物体的高阶空间信息，例如物体的尺度，空间变换或者形变等等。接下来本文依据这些信息，生成卷积核的空间分布：

$$\mathbf{\Delta d} = \eta(\mathbf{S}'), \tag{3.3}$$

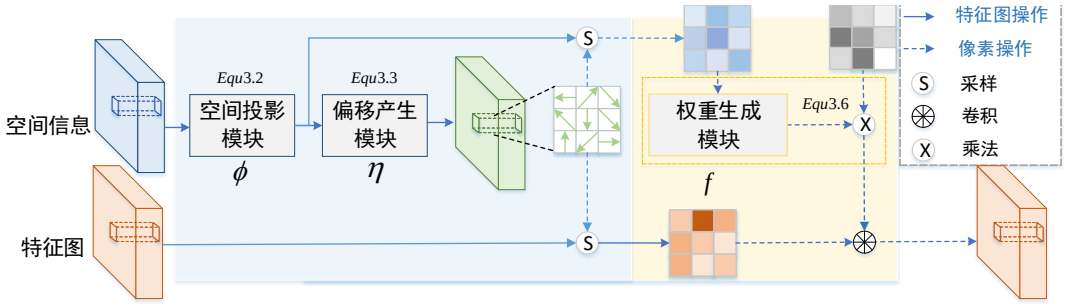

图 3.2 S-Conv 的说明：首先，3D 空间信息会通过空间投影模块将空间信息升维匹配特征图。接着，通过偏移产生模块，本文将升维后的空间信息转化为卷积核的空间分布。最后，通过偏移后的卷积核，本文采集到偏移后的空间信息，并生成对应的权重。

其中 $\mathbf{\Delta d} \in \mathcal{R}^{K\times h'\times w'\times 2}$，$h', w'$ 为卷积后特征图的尺寸。$K = k_h \times k_w, k_h, k_w$ 为卷积核的大小。对于 3×3 卷积来说，$\mathbf{\Delta d} \in \mathcal{R}^{9\times h'\times w'\times 2}$。$\eta$ 是个非线性函数，与 ϕ 类似，η 在本文中用几层卷积来实现。

$\mathbf{\Delta d}$ 包含了卷积核的分布信息，在引入空间自适应分布后，本文的卷积流程更新如下：

$$\mathbf{Y}(\mathbf{p}) = \sum_{i=1}^{K} \mathbf{W}\cdot\mathbf{X}(\mathbf{p}+\mathbf{d}_i+\Delta\mathbf{d}_i(\mathbf{p})), \tag{3.4}$$

其中对于不同的 p, 分布 $\Delta\mathbf{d}_i(\mathbf{p})$ 会动态变化，其值由 p 所在的空间信息决定。而 $\mathbf{d}_i$ 保持不变。在引入 $\Delta\mathbf{d}_i(\mathbf{p})$ 后，卷积核的分布引入了空间自适应性，其感受野和几何变换都自适应于输入图片中物体的几何信息。

接下来，如第二章的空间自适应权重，本文尝试将空间信息融入卷积核的权重中，使得卷积过程有着更强的空间自适应性。与第二章不同的是，由于卷积核已经发生了偏移，其对应的空间信息发生了改变。本文需要首先采样卷积核发生偏移后所对应的空间特征：

$$\mathbf{S}^*(\mathbf{p}) = \{\mathbf{S}'(\mathbf{p}+\mathbf{d}_i+\Delta\mathbf{d}_i(\mathbf{p}))|_{i=1,2,\dots,K}\}, \tag{3.5}$$

其中 $\Delta\mathbf{d}_i(\mathbf{p})$ 为卷积核的偏移量，$\mathbf{S}^*(\mathbf{p}) \in \mathcal{R}^{64K}$ 对应着卷积核发生偏移后的空间特征。最后本文根据偏移后的空间特征，生成卷积核的空间自适应权重：

$$\mathbf{W}^*(\mathbf{p}) = \sigma(f(\mathbf{S}^*(\mathbf{p})))\cdot\mathbf{W}, \tag{3.6}$$

其中 $\mathbf{W} \in \mathcal{R}^K$ 为卷积核权重，可以通过梯度下降法更新。$\mathbf{W}^*(\mathbf{p}) \in \mathcal{R}^K$ 对应着

S-Conv 的空间自适应权重，其权重由位置 p 对应的空间信息决定。f 为非线性函数，可以由全连接网络实现。σ 为 Sigmoid 激活函数。

最后，在引入空间自适应分布和权重之后，本文的卷积过程更新如下：

$$\mathbf{Y}(\mathbf{p}) = \sum_{i=1}^{K} \mathbf{W}_i^*(\mathbf{p}) \cdot \mathbf{X}(\mathbf{p} + \mathbf{d}_i + \Delta\mathbf{d}_i(\mathbf{p})), \tag{3.7}$$

其中 $\mathbf{W}_i^*$ 建立起了空间关系与权重之间的关系，$\Delta\mathbf{d}_i(\mathbf{p})$ 建立起了空间信息与卷积核分布之间的关系。相比于 2D 卷积，本文的 S-Conv 通过额外引入的空间信息，有着更强的几何感知能力。同时 $\Delta\mathbf{d}$ 为网络预测的结果，其值为浮点数，因此同[82]，本文使用双线性插值的方法来计算 $\mathbf{X}(\mathbf{p} + \mathbf{d}_i + \Delta\mathbf{d}_i(\mathbf{p}))$：

$$X(\mathbf{p}) = \sum_{q} H(\mathbf{q}, \mathbf{p}) \cdot X(\mathbf{q}), \tag{3.8}$$

其中 p 为像素的浮点位置，例如公式 (3.7) 中的 $\mathbf{p} + \mathbf{d}_i + \Delta\mathbf{d}_i$，$\mathbf{q}$ 为所有特征图的空间位置。H 的表达式如下所示：

$$H(\mathbf{q}, \mathbf{p}) = h(q_x, p_x) \cdot h(q_y, p_y), \tag{3.9}$$

其中 $h(q,p) = max(0, 1 - |q - p|)$。本文通过公式 (3.8), 公式 (3.9) 即可算出特征图浮点位置的特征值。ϕ, η, f 都可以通过反向梯度传播与梯度下降算法更新参数，通过训练数据学习自适应地产生卷积核的权重与分布。

上述讨论的过程如图 3.2所示。首先，空间信息通过空间投影模块提取空间信息特征，接着，通过偏移量产生模块产生自适应卷积核分布。最后，依据偏移后的卷积核，通过权重生成模块生成对应的自适应权重。每个公式对应的过程都在图 3.2中标出。

3.2.2 S-Conv 与其他方法的关系

在上一小节中，本文从公式角度来详细介绍了 S-Conv 的工作原理。本文接下来从原理上，介绍 S-Conv 与其他方法的通性与区别。本文可以将 2D 卷积视为 S-Conv 没有空间信息的特殊情况。在缺乏空间信息输入时，本文若将公式 (3.7) 中生成的 $W_i^*(p)$ 和 $\Delta d_i(p)$ 移除，公式 (3.7) 就会转换成传统 2D 卷积。同时在 RGBD 的情况下，本文可以通过 S-Conv 中的空间自适应权重在 3D 点云层面而非 2D 层面提取特征，如图 2.5所示。SV Conv[77] 是一个应用在语义分割

任务上的卷积操作，它和 S-Conv 类似，都是随位置变化的卷积操作。本文的 S-Conv 和 SV Conv 有以下几点不同：1）在研究动机上，SV Conv 通过基于语义相关区域的位置卷积来限定其上下文区域，侧重于理解上下文语义而非空间结构。本文的 S-Conv 利用深度图而非特征图来生成空间自适应偏移量和权重。本质上，S-Conv 的目的是提取深度图的空间信息以辅助语义信息的判断。在实现方面，SV Conv 实现了一个卷积运算符，它的权值是由上下文的语义信息决定的。与 SV Conv 不同，S-Conv 的卷积核的分布和权重是空间自适应的。这两类方法的区别如图 3.3所示。可变形卷积也可以生成不同分布和权重的卷积核，但

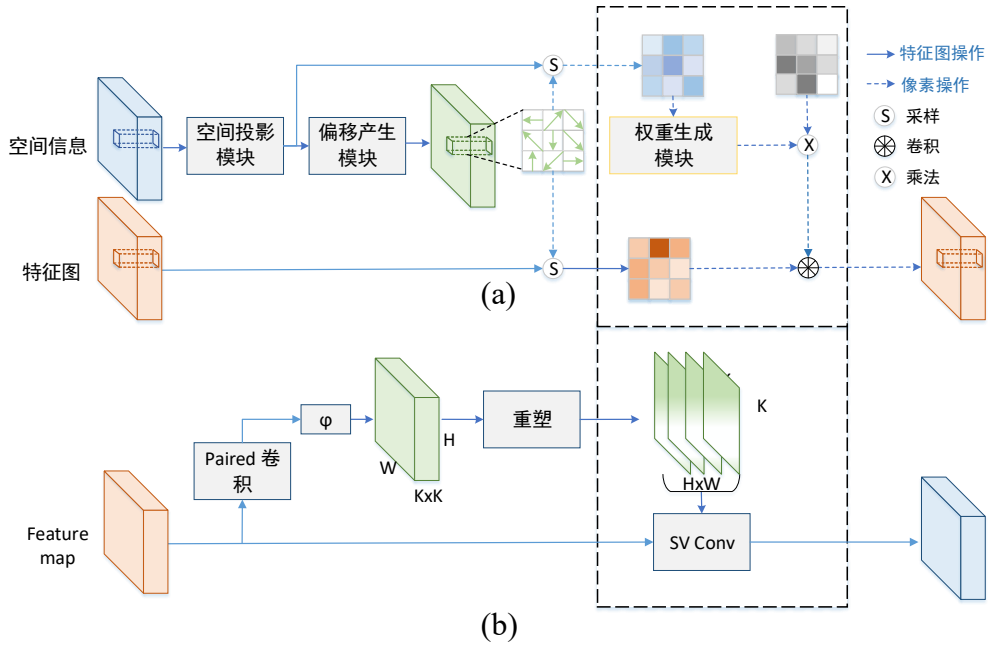

图 3.3 S-Conv (a) 与 SV Conv (b) 的比较。

它们的分布由中间特征图决定而非 3D 空间信息决定。本文的 S-Conv 直接通过空间信息推理出物体的尺度和空间变换，生成卷积核的空间分布。从而帮助卷积核自适应空间变换与调整感受野。本文后面通过实验证明，通过 3D 空间信息生成的分布与权重性能优于使用中间特征图。相比于 3D 点云分割方法中的 KNN 寻找最近邻的方法，本文的 S-Conv 通过空间自适应的方法寻找最近邻而不是固定且更消耗计算资源的 KNN 方法。

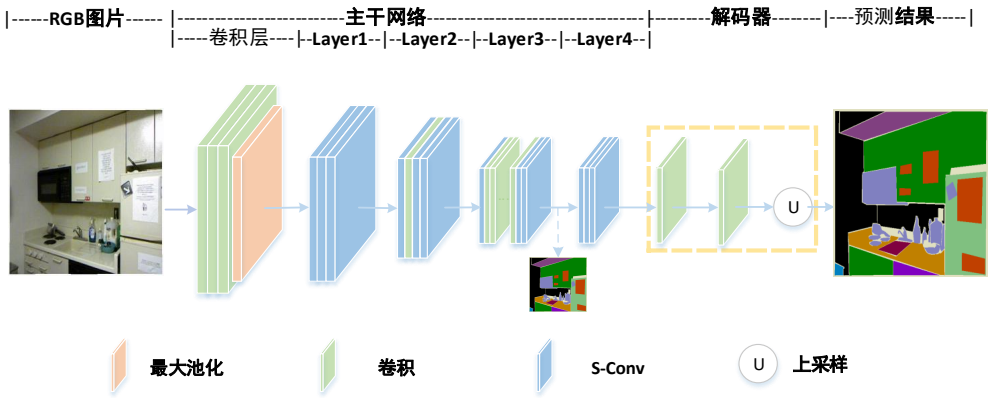

图 3.4 SGNet 的网络结构：SGNet 包括一个主干网络和解码器。本文在第三层和第四层添加了一个深层监督来提升性能。

3.2.3 SGNet 的结构

本文提出的实时 RGBD 语义分割网络，名为 SGNet, 其为第二章中的基础网络引入 S-Conv 得到。本文 SGNet 的网络结构如图 3.4所示：本文使用 ResNet101 作为主干网络，并将每层的第一个和后两个卷积（3×3 卷积）替换为 S-Conv。作者在主干网络的后面加了几层卷积网络，最后使用双线性插值得到最后的分割结果。公式 (3.2) 中的 ϕ 的实现为 3 个 3×3 卷积。即带有非线性激活层的 Conv(3, 64) - Conv(64, 64) - Conv(64, 64) 结构。公式 (3.3) 中的 η 和公式 (3.6) 中的 f 分别使用一个和两个卷积层实现，通道数均为 64。S-Conv 的实现参考于可变形卷积。类似于 PSPNet[27], 本文在第三层和第四层添加了一个深层监督来提升性能。在上一章中的实验得出，深层监督对网络的优化有正向作用。在接下来的章节中，本文将设计实验，在 NYUDv2[84] 与 SUNRGBD[85, 86] 数据集上验证 S-Conv 与 SGNet 的有效性，并进行详细分析。

第三节 实验结果对比与分析

在这一节中，本文首先探索 S-Conv 在主干网络不同层上的表现；设计消融实验证明 S-Conv 的有效性；使用不同的信息来生成卷积核的偏移量与权重并对比结果。接下来本文对比 SGNet 与其他先进方法在 NYUDv2[84] 与 SUNRGBD[85, 86] 上的表现。最后，本文可视化 S-Conv 的感受野和 SGNet 在 NYUDV2[84] 与 SUNRGBD[85, 86] 上的语义分割结果。上述实验说明了本文提出

的 S-Conv 高效充分地利用了空间信息提升语义分割的结果。

3.3.1 实验介绍

数据集和评测指标：本文在下列数据集上，验证 S-Conv 与 SGNet 分割网络的性能。

- NYUDv2[84]：此数据集包括 1,449 张 RGB 图片和对应的深度图与分割标注图。与之前的方法[89] 保持一致，本文使用 795 张图片用于训练，654 张图片用于测试。本文使用 40 类别的设置来进行实验。
- SUNRGBD[85, 86]：此数据集包括 10,335 张 RGB 图片和对应的深度图与分割标注图，有 37 个类别，其中 5,285 张图片用于训练，5,050 张图片用于测试。
- Cityscapes[88]: 此数据集为室外场景数据集。本文将这个数据集分为训练，测试和验证集，分别有 2,975, 500 和 1,525 张图片。

本文使用 3 个常见的评测指标来验证，包括精度 (Acc)，平均精度 (mAcc)，和平均交并比 (mIoU)。本文使用深度图作为默认的空间信息作为 S-Conv 的输入。

实现细节：与[13] 一致，本文使用在 ImageNet[6] 上预训练后的 ResNet101[87] 作为本文特征提取的主干网络。默认的输出步长为 16。本文使用 Pytorch 实现整个系统。网络使用 SGD 作为优化器，并采用以下的衰减策略：$lr = initial\_lr \times (1 - \frac{iter}{total\_iter})^{power}$。这个衰减策略在 NYUDv2 数据集每迭代 40 轮应用一次，在 SUNRGBD 上每 10 轮应用一次。其中网络的初始学习率在消融实验上为 5e-3，在 NYUDv2[84] 数据集上为 8e-3，在 SUNRGBD[86] 数据集上为 1e-3。权重衰减设置为 5e-4。网络默认使用 ReLU 激活函数，批大小默认为 8，同[5]，网络使用常用的数据增广策略，包括随机尺度变换，随机剪裁，随机翻转。剪裁的大小为 480×640。在测试阶段，网络将图片下采样到训练裁剪的大小，然后预测的结果上采样到输入大小。网络使用交叉熵损失，由于 SUNRGBD 严重类别不平衡，本文依据类别的分布，对不同的类别添加不同的权重。本文使用两张 NVIDIA 1080Ti 显卡进行训练，在 NYUDv2 数据集上训练 500 轮，在 SUNRGBD 上训练 200 轮。

表 3.1 在 NYUDv2 测试集上将卷积替换为 S-Conv 的结果。“layerx_y” 代表替换第 x 层的第 y 个残差网络的 3 × 3 卷积。

layer3_0	layer3_1	layer3_2	layer3_20	layer3_21	layer3_22	其他	平均交并比 (%)	参数量 (M)	FPS
							43.0	56.8	34
✓							47.0	56.9	34
✓	✓	✓					46.6	57.2	33
			✓	✓	✓		46.5	57.2	33
✓				✓	✓		47.8	57.2	33
✓				✓	✓	✓	**49.0**	58.3	26

3.3.2 S-Conv 的分析

作者首先在 NYUDv2[84] 数据集上进行消融实验。本文使用第二章介绍的基础网络作为基线模型。基线模型如图 2.2所示。在这一小节中，本文着重于分析 S-Conv 的结构设计，性能，以及与其他替代方案的比较。

S-Conv 替代卷积实验：在上一章中，本文简单的在网络中每一层的最后三个卷积引入空间自适应权重。为了更加精细充分的利用 S-Conv 提升基础网络的性能，本文在这一小节中通过将 S-Conv 在基础网络的不同位置卷积的替换来探索在不同位置的 S-Conv 对网络结果的影响。本文通过在不同层将 3 × 3 卷积替换为 S-Conv 来研究其对结果的影响。为了方便说明，本文首先在第三层做实验，接着将这个探索到的规则推广到其他层。本文的 FPS (Frames Per Second) 在 NVIDIA 1080Ti 上测试，输入图像分辨率为 480 × 640。实验结果如表 3.1所示：

作者可以从表 3.1得到以下结论：1) 基线网络的推理速度最快，但表现相对一般。将卷积替换为本文的 S-Conv 可以大大提升网络的精度，并只需要少许额外的参数量和计算量。2) 作者发现，除了第一个步长为 2 的卷积，替换后部分的卷积效果比较明显。可能因为空间信息可以更好的指导下采样。所以本文选择替换每一层的第一个和最后两个 3 × 3 卷积。本文将第三层得到的结论推广到其他层，达到了更好的效果。上述实验说明本文的 S-Conv 可以在增加少量参数的情况下。显著的提升网络的结果。值得说明的是，本文的网络没有显式空间信息输入。空间信息仅仅影响卷积核的空间分布和权重。本文同时探索 S-Conv 应用在不同层的表现，结果如表 3.2所示。作者发现，网络的效果会随着 S-Conv 层数的加深提升。

为了直观的感受到 S-Conv 带来的性能提升，本文同时在图 3.5展示了相对基线模型每个类别的提升。作者发现 S-Conv 在绝大多数类别都有提升。尤其是在缺少纹理信息的类别：例如镜子，板子和浴缸。这些类别在基线网络中表

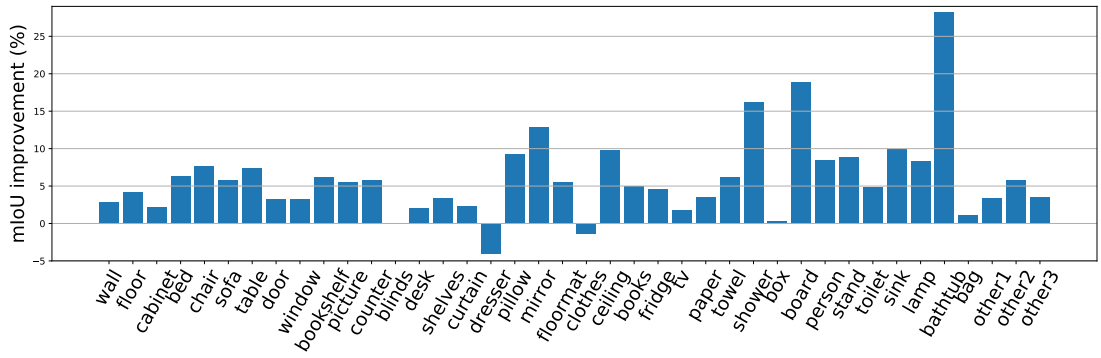

图 3.5 引入 S-Conv 后，基础网络在 NYUDv2 测试集上每个类别的提升。

表 3.2 S-Conv 替换网络不同层在 NYUDv2 测试集上的表现。结果为百分数。

Layer1	Layer2	Layer3	Layer4	准确率	平均准确率	平均交并比
				72.1	54.6	43.0
✓				74.3	58.1	46.3
✓	✓			74.4	58.4	46.7
✓	✓	✓		75.2	60.3	48.5
✓	✓	✓	✓	**75.5**	**60.9**	**49.0**

现较差，但在引入了 S-Conv 后，这些类别的性能得到了显著提升，尤其是浴缸 (+27%)，板子 (+18%)。S-Conv 在一些具有丰富空间变换类别上依然有显著提升，例如椅子和桌子。这说明 S-Conv 在推理阶段充分的利用了空间信息。

S-Conv 结构消融实验：在上述实验中，本文探索了 S-Conv 在不同位置对网络结果的影响，并应用此规律来将网络的性能最大化。在这一小节中，本文将验证 S-Conv 结构设计的合理性。为了验证本文提出的 S-Conv 中每个模块的效果，本文设计了消融实验，实验结果如表 3.3 所示。本文根据表 3.1得到的结论，每一层仅替代第一个和最后两个 3×3 卷积。作者发现，S-Conv 中的偏移产生模块，权重产生模块，空间投影模块，均对结果产生了正向作用。其中偏移产生模块的作用最为显著。权重生成模块对应着本文的空间自适应权重，这说明本文 S-Conv 的设计均对结果产生了正向作用。

与其他备选方案的比较：在上一小节中，本文探索了 S-Conv 的设计合理性，设置了消融实验来验证不同模块对结果的影响。在这一小节中，本文将 S-Conv 与其他的空间信息利用方法进行比较来进一步验证 S-Conv 的优势。这些方法包括双流网络方法，可变形卷积方法，深度感知卷积方法等。之前的大部分方法[2, 5, 41, 90] 使用双流网络来提取不同模态的特征然后进行混合。本文的 S-Conv

表 3.3 SGNet 在 NYUDv2 测试集上的消融实验。OG: S-Conv 的偏移产生模块；WG: S-Conv 的权重产生模块；SP: S-Conv 的空间投影变换模块。

SP	OG	WG	准确率 (%)	平均准确率 (%)	平均交并比 (%)
			72.1	54.6	43.0
	✓		73.9	58.2	46.3
✓	✓		75.2	60.0	48.4
✓		✓	74.5	58.4	46.8
✓	✓	✓	**75.5**	**60.9**	**49.0**

表 3.4 NYUDv2 测试集上的比较结果。DCV2: 可变形卷积[83]；DAC: 深度感知卷积[47]；SP: S-Conv 中的空间投影模块；WG: S-Conv 中的权重产生模块。表中的结果为百分数。

模型	准确率	平均准确率	平均交并比
基础网络	72.1	54.6	43.0
基础网络 +DCV2	73.0	56.1	44.5
基础网络 +HHANet	73.5	56.8	45.4
基础网络 +DAC	73.8	57.1	45.4
基础网络 +HHANet+DCV2	74.3	58.4	47.0
基础网络 +DAC+DCV2	74.5	58.3	46.5
基础网络 +SP+WG	74.5	58.4	46.8
基础网络 +S-Conv(SGNet)	**75.5**	**60.9**	**49.0**

专注于利用空间信息来改善特征提取的过程。这里本文比较了双流网络，可变形卷积[82, 83]，和深度感知卷积[47]。作者使用和上述实验一致的基线网络，其由一个 ResNet101 网络，一个深层额外监督和简单的解码器组成。在与双流网络方法对比的过程中，作者添加了一个额外的 ResNet101 网络，名为 HHANet，来提取空间信息的特征并在网络的最后一层与基线网络合并。为了与深度感知卷积与可变形卷积比较，和 SGNet 相同，本文替代基线网络每一层的第一个和最后两个卷积。对于 “基础网络 +DAC+DCV2” 配置，由于可变形卷积 (DCV2) 在底层表现一般[83]，因此本文在前两层 (Layer1, Layer2) 使用深度感知卷积，后两层 (Layer3, Layer4) 使用可变形卷积。结果如表 3.4所示。作者发现，S-Conv 的效果要优于双流网络，可变形卷积[82, 83]，深度感知卷积[47]，和它们的组合。这说明了本文的 S-Conv 可以高效充分的利用空间信息。同时作者发现，仅仅使用 S-Conv 中的权重生成模块 (基础网络 +SP+WG) 效果优于深度感知卷积[47]。这说明本文的自适应权重策略要优于深度感知卷积[47] 中的手动设置的权重策略。

空间信息的消融实验：上述实验中，本文通过将 S-Conv 与其他方法进行比

表 3.5 使用不同的空间信息在 NYUDv2 测试集上的比较结果。

空间信息	准确率 (%)	平均准确率 (%)	平均交并比 (%)
深度图	75.5	60.9	**49.0**
RGB 特征	73.9	58.5	46.4
HHA	**75.7**	60.8	48.9
3D 空间坐标	75.3	**61.2**	48.5

较说明了 S-Conv 的优势。在这一小节中，和自适应权重类似，本文同时探索在 S-Conv 中，使用不同类型的空间信息的效果。可以使用的空间信息形式包括深度图 (depth map)，RGB 特征，HHA 和像素的 3D 坐标。其中 HHA 为 Depth 的编码结果，在之前的工作中效果最好，但带来了大量的计算量。这同时也说明之前的网络无法充分的利用空间信息。本文的实验结果如表 3.5所示：可以看到，使用深度图的效果和空间坐标，HHA 的效果基本一致。即使没有手工设计的编码特征，S-Conv 使用深度图与 HHA 的效果接近，这说明本文的 S-Conv 提取到了手工设计的编码特征，相比于之前的工作更加充分的利用了空间信息。同时，使用深度图 (Depth)，HHA 和像素的 3D 坐标的结果显著优于 RGB 特征，这说明相比于使用 RGB 图片网络中间特征，空间信息更适合用来生成权重与卷积核分布。这也证明了本文利用空间信息生成卷积核偏移量与权重思路的正确性。

推理速度测试: 之前的实验中，本文探索了不同空间信息对 S-Conv 结果的影响，在这一小节中，本文对网络的推理速度进行了测试来说明 S-Conv 的轻便性。本文在这一小节分析推理的计算量和速度。本文同时比较了 S-Conv 与双流网络的推理速度的对比，输入图像的大小为 480×640。本文的结果如表 3.6所示。作者发现，S-Conv 相比于双流网络的方法，仅仅在基础网络上增加了少量计算量。同时本文的 SGNet 在 480×640 的输入下，使用 ResNet101 与 ResNet50 主干网络可以达到实时推理速度。

3.3.3 与其他主流方法的对比

在上述实验中，本文验证了 S-Conv 充分利用空间信息的能力和其较小的参数量和计算量。本文将 SGNet 与其他的先进方法在 NYUDv2 数据集和 SUNRGBD 数据集上进行比较。SGNet 的网络结构如图 3.4所示。

NYUDv2 数据集：在这一小节中，本文测试 SGNet 在 NYUDv2 数据集上的表现。作者的参数配置与设置和之前消融的实验保持一致。SGNet 的比较结果

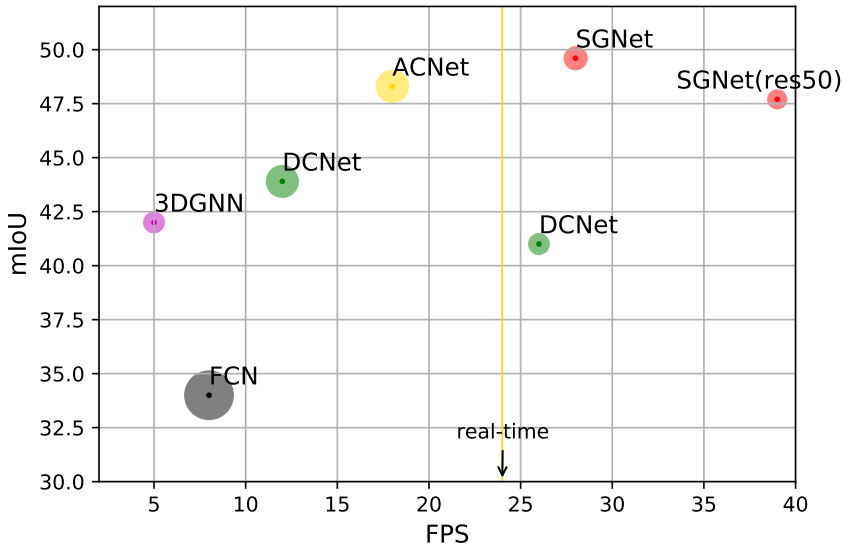

图 3.6 不同方法使用 NVIDIA 1080Ti 在 NYUDv2 测试集上的参数量，mIoU 和 FPS：圆的半径代表参数量。DCNet[47] 和 3DGNN[44] 的结果来自[47]。所有的速度比较均为单尺度，输入图片大小为 425 × 560。本文的 SGNet 达到了最快的推理时间和最先进的性能。

如表 3.7和图 3.6所示: 作者将初始学习率从 5e-3 调整为 8e-3, 作者将输入图片下采样到 480 × 640 并上采样其预测图来得到最终结果。为了和其他方法比较速度。本文和[47] 保持一致，测试图片大小为 425 × 560，在 NVIDIA 1080Ti 下测试。DCNet[47] 和 3DGNN[44] 的速度来自于[47]。同时本文测试了其他方法的单尺度速度，另外，SGNet 在图片分辨率为 480 × 640 的推理速度如表 3.6所示。值得注意的是，表 3.7里有些方法并没有开源或者报告参数量，所以本文只列举了它的 mIoU，没有使用额外的网络提取空间特征。本文的 SGNet (ResNet50) 可以在 39FPS 的速度下达到富有竞争力的效果，并有着最少的参数量。本文的 SGNet (ResNet101) 可以达到最好表现和实时的推理速度。这些都得益于 S-Conv 可以在增加少量的计算量充分高效的利用空间信息。另外，本文的 S-Conv 可以不使用 HHA 信息，依然达到很好的效果。这使得本文的网络在实时场景中得到应用。这同时说明了本文的 S-Conv 可以更好的提取空间特征。由第二章的结论，通过增加 ASPP[23] 多尺度模块，本文的”SGNet*” 可以进一步的提升结果，效果优于使用多尺度测试的，两个 ResNet152 主干网络，效果最好的 RDFNet[5] 和 CFNet[46]。同时本文的”SGNet*” 也可以保持实时推理速度。在使用多尺度测试的方式后，本文的”SGNet” 效果进一步得到提升。mIoU 达到 51.1%。本文同时在图 3.6中，直观比较了每种方法的速度，参数量和交并比。可以看到，在这三

表 3.6 480 × 640 输入图片分辨率下模型的推理速度测试。OG: S-Conv 的偏移量产生模块, †: SGNet 中不应用产生的偏移量与权重, HHANet: 额外的主干网络 (ResNet101) 来处理空间信息。

模型	时间 (秒)	帧率 (FPS)	参数量 (M)
基础网络	0.029	34	56.8
基础网络 +OG	0.033	30	57.7
基础网络 +HHANet	0.053	18	99.4
SGNet(ResNet50)	0.028	36	39.3
SGNet†	0.032	31	58.3
SGNet	0.037	26	58.3

种参考标准的权衡下，本文的方法 SGNet 表现最优。

SUNRGBD 数据集：本文的 SGNet 在 NYUDv2 数据集上取得了最优的结果并达到了实时推理速度。在这一小节中，本文将验证 SGNet 在 SUNRGBD 数据集上的表现。由于 SUNRGBD 的样本比例极度不平衡，本文采用公式 (2.3) 作为损失函数。SGNet 的 SUNRGBD 数据集上的比较结果如表 3.8所示。在表 3.7中的某些方法并没有报告 SUNRGBD 数据集的结果。表 3.8中的推理时间与参数量与表 3.7 相同。SGNet 相比于其他方法达到了富有竞争力的结果，例如使用了两个 ResNet152 网络，并采用多尺度验证方法的 RDFNet[5] 和 CFNet[46]。本文的 SGNet 使用了最少的参数量，并有着实时推理速度。在使用了多尺度测试的方式后，SGNet 在所有的方法中可以达到最好的结果。

Cityscapes 数据集：上述实验中，SGNet 在两个室内场景下取得了较为理想的结果。为了验证 SGNet 的通用性，本文将验证 SGNet 在室外场景的效果。本文首先在 SGNet 后面添加 ASPP[23] 模块，增加其多尺度表达能力，并设置输出步长为 8，命名为 *SGNet-8s*。网络的学习率设置为 1e-2，批大小设置为 8。本文使用的数据增广包括随机尺度变换，随机剪裁，随机翻转。裁剪的大小为 769 × 769，随机尺度变换的范围为 [0.7, 2.1]。由于 Cityscapes 数据集只有视差图，本文通过公式 (2.5) 将视差图转换成深度图。作者使用 SGD 作为优化器，并采用 “Poly” 的学习率下降策略。作者在训练集的 2975 张图片上进行训练，并在验证集上进行验证。本文同时提供了在 Cityscapes 服务器上的测试结果。比较结果如表 3.9所示：由于室外数据集，深度信息噪声较大，充分利用其深度信息指导语义分割较为困难。可以看到，SGNet 在 Cityscapes 数据集上达到了富有竞争力的结果，这说明了 SGNet 的有效性和通用性。同时本文的 SGNet 网络的效果

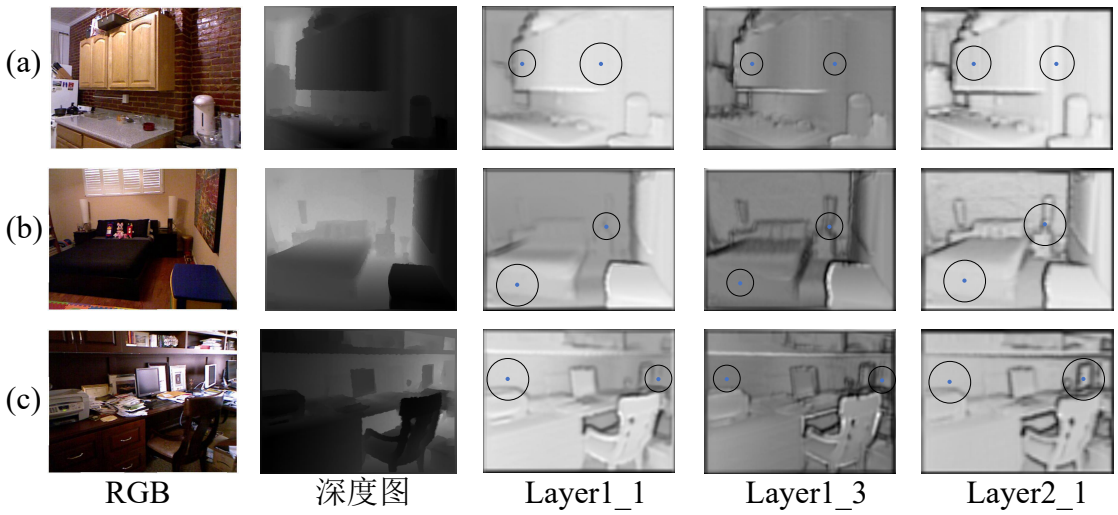

图 3.7 S-Conv 中的感受野可视化。其中从左到右分别为：输入图片，深度图，Layer1_1 的可视化结果，Layer1_3 的可视化结果，和 Layer2_1 的可视化结果。圆圈半径代表感受野大小。

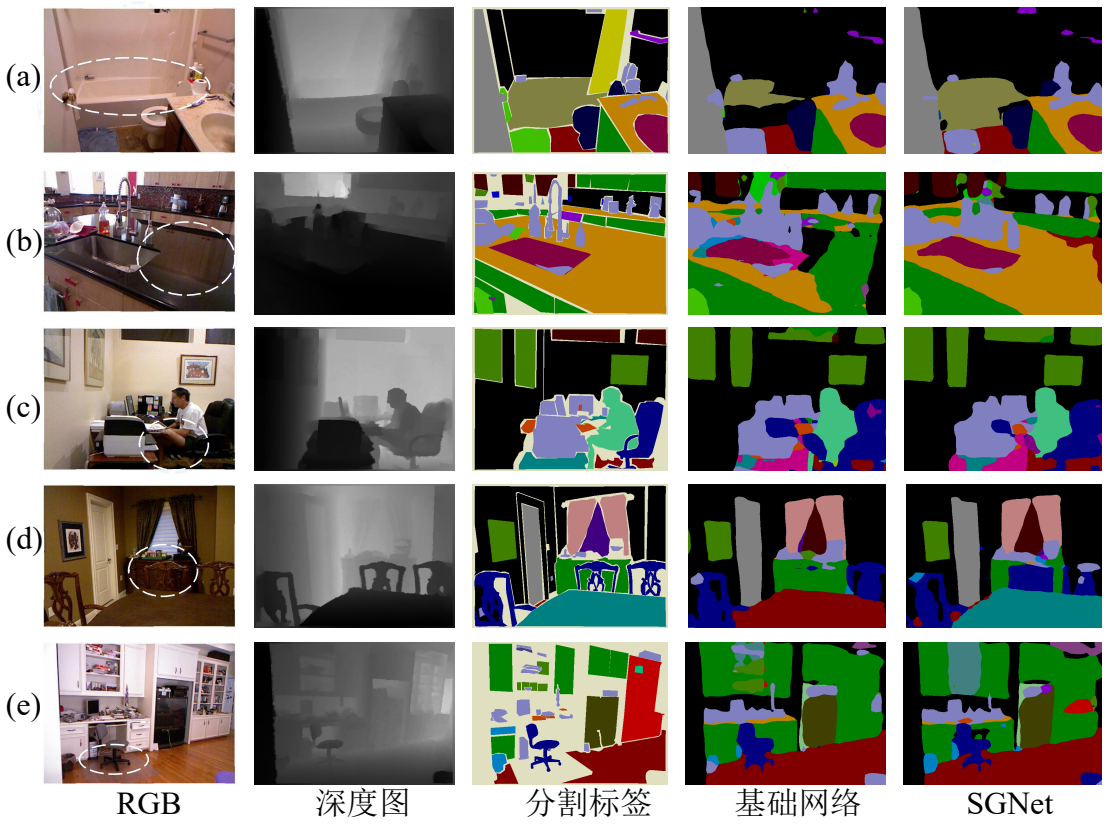

图 3.8 NYUDv2 测试集上的分割结果可视化。

表 3.7 NYUDv2 测试集上的比较结果。MS: 多尺度测试; SI: 空间信息，Acc: 准确率 (%)，mIoU：平均交并比 (%)，param：参数量 (M)。输入测试图片的大小为 425 × 560，在 NVIDIA 1080Ti 环境下，速度均为单尺度速度。本文在 SGNet 的最后一层添加了 ASPP 模块[23] , 命名为 “SGNet*”。

网络	主干模型	MS	SI	Acc	mIoU	FPS	param
FCN[8]	2×VGG16		HHA	65.4	34.0	8	272.2
LSD-GF[45]	2×VGG16		HHA	71.9	45.9	-	-
3DGNN[44]	VGG16		HHA	-	42.0	5	47.2
D-CNN[47]	2×ResNet152		Depth	-	48.4	-	-
ACNet[41]	3×ResNet50		Depth	-	48.3	18	116.6
RefineNet[17]	ResNet152	✓	-	73.6	46.5	16	129.5
RDFNet[5]	2×ResNet152	✓	HHA	76.0	50.1	9	200.1
RDFNet[5]	2×ResNet101	✓	HHA	75.6	49.1	11	169.1
CFNet[46]	2×ResNet152	✓	HHA	-	47.7	-	-
SGNet	ResNet50		Depth	75.0	47.7	**39**	**39.3**
SGNet	ResNet101		Depth	75.6	49.6	28	58.3
SGNet*	ResNet101		Depth	76.1	50.2	26	64.7
SGNet*	ResNet101	✓	Depth	**76.8**	**51.1**	26	58.3

优于基础网络，这验证了 S-Conv 可以充分有效的利用空间信息来改善语义分割的结果。

3.3.4 可视化分析

SGNet 在室内与室外数据集上均达到了较为理想的结果，并超过了目前的所有方法。为了直观体验 S-Conv 充分利用空间信息的性能，在接下来的实验中，本文将对 SGNet 中 S-Conv 的感受野进行可视化。

S-Conv 的感受野可视化：合适的感受野对于场景识别十分重要。本文可视化了在 SGNet 不同层中 S-Conv 产生的自适应感受野。具体来说，本文通过累计 S-Conv 操作期间其卷积核偏移量的模长来获得每个像素的感受野。接着本文将感受野的大小归一化到 [0,255] 来可视化感受野的大小。感受野可视化的结果如图 3.7所示：像素越亮代表相对感受野越大。本文同时使用了圆的半径来代表感受野的大小。作者发现感受野随着深度变化有一定的规律，比如在 layer1_1, 感受野随着深度呈反比。通过不同层不同的感受野的组合可以帮助网络更好的感知室内场景的空间关系与结构。

分割结果比较：在可视化了 S-Conv 的感受野之后，针对第一章中的分割示例图，为了定性展示网络在难区分类别上的性能，本文展示了 SGNet 在 NYUDv2

表 3.8 SUNRGBD 测试集上的比较结果。MS: 多尺度测试，SI: 空间信息，Acc: 准确率 (%)，mIoU：平均交并比 (%)，param：参数量。本文在 SGNet 的最后一层添加了 ASPP 模块[23], 命名为 “SGNet*”。

网络	主干模型	MS	SI	Acc	mIoU	param (M)
LSD-GF[45]	2×VGG16		HHA	-	-	-
RefineNet[17]	ResNet152	✓	-	80.6	45.9	129.5
CGBNet[22]	ResNet101		-	82.3	48.2	-
3DGNN[44]	VGG16	✓	HHA	-	45.9	47.2
D-CNN[47]	2×VGG16		HHA	-	42.0	92.0
ACNet[41]	3×ResNet50		HHA	-	48.1	272.2
RDFNet[5]	2×ResNet152	✓	HHA	81.5	47.7	200.1
CFNet[46]	2×ResNet152	✓	HHA	-	48.1	-
SGNet	ResNet101		Depth	81.0	47.1	58.3
SGNet*	ResNet101		Depth	81.0	47.5	64.7
SGNet*	ResNet101	✓	Depth	82.0	**48.6**	64.7

表 3.9 在 Cityscapes 验证集上的比较结果。‡: 在测试集上的测试结果。

网络	主干模型	迭代次数	多尺度	平均交并比 (%)
基础网络	ResNet101	40k		78.2
SGNet-8s*	ResNet101	40k		79.2
SGNet-8s*	ResNet101	65k	✓	80.6
SGNet-8s*	ResNet101	65k	✓	81.2‡

数据集上的定性比较结果，如图 3.8所示。对于图 3.8 (a)，浴缸和墙壁没有具有代表性的纹理信息，对于基线网络来说很难区分。一些物体比如图 3.8 (b) 中的桌子会有倒影反射其他物体的纹理来干扰网络进行语义感知。本文的 SGNet 通过 S-Conv 可以更好的解决上述情况。图 3.8 (c,d) 中的椅子由于其对比度很低，很难被基线网络区分，但可以被受益于 S-Conv 的 SGNet 更好的区分。另外，S-Conv 也能更好的恢复物体的几何形状，如图 3.8(e) 中的椅子所示。通过对上述难区分类别上的分割结果对比，证实了本文 S-Conv 的有效性。本文同时展示了一些在 SUNRGBD 数据集上和 Cityscapes 数据集上的结果，如图 3.9与图 3.10所示。可以看到本文的 SGNet 在 SUNRGBD 测试集上得到了精细的分割结果，同时在 Cityscapes 数据集上，“SGNet-8s*” 可以得到高质量的分割结果。这说明本文的 SGNet 在室内外数据集均有着良好的表现和泛化性能。

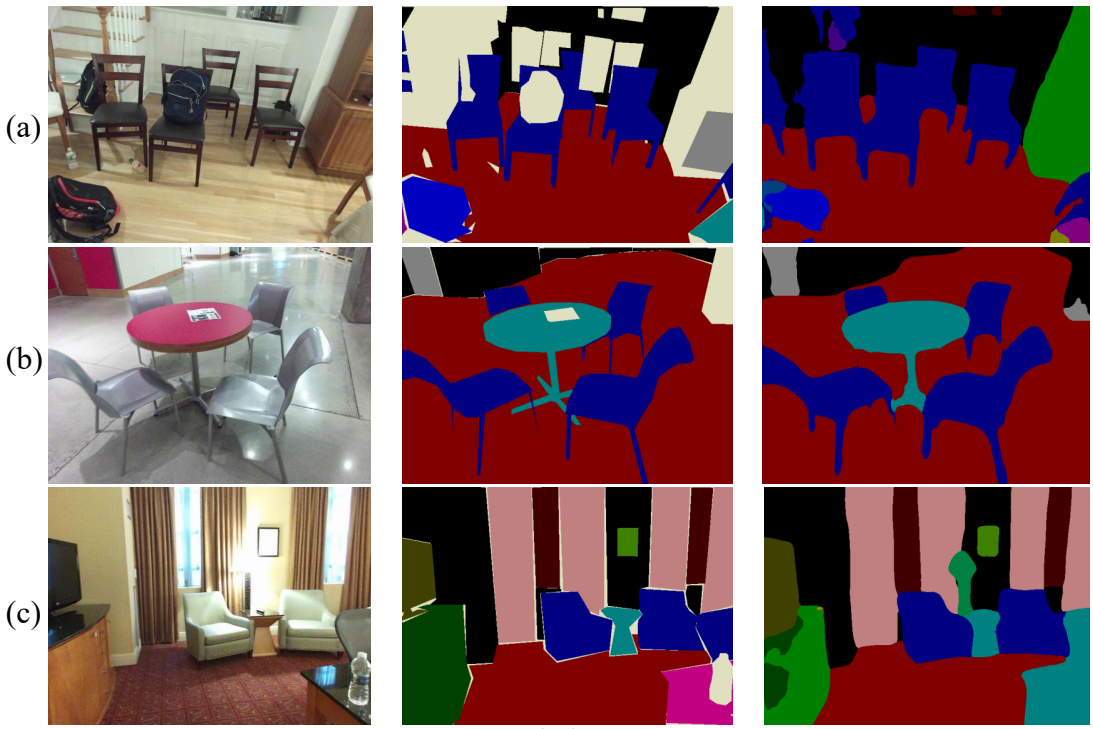

图 3.9 SUNRGBD 测试集上的分割结果可视化。从左到右，RGB，分割标签，SGNet 的结果。

第四节 本章小结

通过在卷积中引入空间自适应权重可以高效利用空间信息提升语义分割精度，并仅仅使用少量的计算量和参数量。为了更充分的利用空间信息指导 RGB 网络卷积过程，以空间自适应权重思路为基础，在本节中，本文提出了 S-Conv 操作，相比于传统的 2D 卷积，S-Conv 可以自适应的根据输入的空间信息来自适应的调整卷积核的分布与权重，进而在增加少量的参数和计算量的条件下，更好的感知几何结构，提升语义分割网络的精度和表现。本文也提出了基于 S-Conv 的 SGNet，在实时推理速度下在 NYUDv2 数据集和 SUNRGBD 数据集上达到富有竞争力的结果。本文将 S-Conv 与当前主流的备选方案进行比较，例如可变形卷积，深度感知卷积，双流网络等，充分证明了 S-Conv 的高效性和有效性。本文也比较了用深度图和中间特征图来生成卷积核分布的性能，说明了使用深度图能产生相对更好的偏移量与权重。最后，本文可视化了每层 S-Conv 的感受野来直观的说明 S-Conv 的工作原理和有效性。同时，S-Conv 也为其他多模态任务提供了一种新的模态融合思路。与通过额外分支提取模态特征，并与主干 RGB

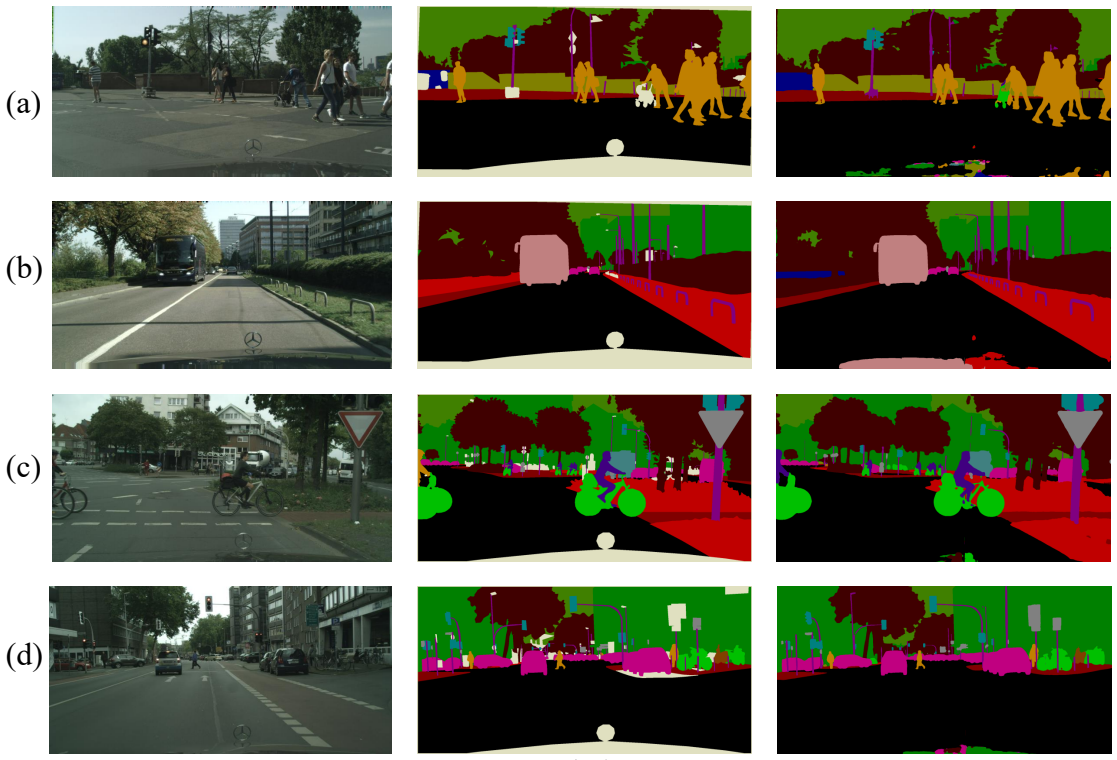

图 3.10 Cityscapes 验证集上的分割结果可视化。从左到右，RGB，分割标签，SGNet 的结果。

网络特征融合的思路相比，本文通过利用额外模态信息 (空间信息) 指导 RGB 网络卷积操作的思路可以极大的减少网络的参数量，并充分利用该模态信息提升网络的性能表现。同时，本文也可以结合这两种思路，使得双流网络和指导卷积的方法相互增益。

第四章 总结与展望

第一节 全文总结

随着深度传感器的大规模应用，深度信息变得越来越普及，利用深度信息提升神经网络的感知能力将变得十分有意义。对无人车，机器人等领域也影响甚远。本文首先总结了目前语义分割，RGBD 语义分割的研究现状，分析了当前 RGBD 语义分割方法中存在的问题，即目前的双流网络方法需要大量的参数量和计算量来利用空间信息。为了高效充分的利用空间信息并使其应用到实时场景。本文首先设计了一种基础网络，并通过设计实验验证了结构设计的合理性。接着，针对室内场景动态的空间场景变化与结构固定 2D 卷积的矛盾，本文提出了空间自适应权重，使得卷积核的权重随着其本身的空间位置自适应变化，来使其更好的感知邻域的几何形状与空间结构。将空间信息充分的引入到卷积操作中。本文通过实验初步验证了空间自适应权重的有效性，并将其应用于 RGBD 语义分割任务中来提升分割任务的精度。最后以空间自适应权重为基础，本文提出了空间信息指导卷积 (S-Conv)，该操作可以根据动态的空间信息输入，在卷积的不同位置生成空间自适应变化的卷积核分布，从而增强网络的空间变换适应能力与感受野调节能力。同时，S-Conv 建立起卷积核权重与对应像素空间信息的关联，将空间信息融入权重之中，从而更好的感知物体的空间结构。S-Conv 可以在增加少量计算量和参数量的情况下充分提升语义分割的性能。本文通过 S-Conv 替代卷积，结构消融，和其他备选方案进行对比，不同空间信息类型的比较，与时耗分析实验来验证 S-Conv 操作的有效性。最后，基于 S-Conv, 本文提出了空间信息引导卷积网络 (SGNet)，该 SGNet 能够达到实时推理速度，并在 NYUDv2,SUNRGBD 数据集上达到最先进的性能。本文同时展示了 SGNet 在 NYUDv2 数据集，SUNRGBD 数据集上的分割结果，展示了其在不同种类环境下基础网络与 SGNet 的分割效果对比，充分说明了 S-Conv 的有效性。本文也可视化了 S-Conv 在不同层上的感受野，详尽展示了 S-Conv 在不同层感受野随不同空间信息输入的变化。这说明 S-Conv 可以自适应的根据空间信息，在不同层自适应的调整感受野，从而更好的感知场景的空间结构。

第二节 未来展望

目前，RGBD 语义分割研究领域处于起步阶段，双流网络思路已经有了足够深入的研究，但通过自适应操作来利用空间信息的方法仍需要进一步的探索。这两种思路各有特点，也可以相互补足。在未来，作者会探索 S-Conv 的思想在其他多模态任务下的应用，包括人脸识别，姿态识别，场景识别，显著性检测等等。作者会将双流网络方法与 S-Conv 的方法结合，使得这两种方法彼此受益。作者同时也会探索 S-Conv 在不同场景下的应用，例如姿态检测与 3D 物体检测。同时将 S-Conv 推广到 3D 场景下，例如点云识别，检测与分割。将其与 3D 卷积的优势结合，也是一个值得探索的方向。

参考文献

[1] MA L, STÜCKLER J, KERL C, et al. Multi-view deep learning for consistent semantic mapping with rgb-d cameras. [C] // Intelligent Robots and Systems (IROS). IEEE. 2017: 598–605.

[2] HAZIRBAS C, MA L, DOMOKOS C, et al. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. [C] // Asian Conference on Computer Vision (ACCV). Springer. 2016: 213–228.

[3] WANG J, WANG Z, TAO D, et al. Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. [C] // European Conference on Computer Vision (ECCV). Springer. 2016: 664–679.

[4] LI Z, GAN Y, LIANG X, et al. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. [C] // European Conference on Computer Vision (ECCV). Springer. 2016: 541–557.

[5] PARK S.-J, HONG K.-S, LEE S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. [C] // The IEEE International Conference on Computer Vision (ICCV). 2017: 4980–4989.

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

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

[8] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2015: 3431–3440.

[9] LIU Y, CHENG M.-M, HU X, et al. Richer convolutional features for edge detection. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 3000–3009.

[10] HOU Q, CHENG M.-M, HU X, et al. Deeply supervised salient object detection with short connections. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 3203–3212.

[11] 范登平, 季葛鹏, 秦雪彬, 等. 认知视觉启发的物体分割评价标准及损失函数. [J]. 中国科学：信息科学, -. DOI: https://doi.org/10.1360/SSI-2020-0370.

[12] SHEN W, ZHAO K, JIANG Y, et al. Object skeleton extraction in natural images by fusing scale-associated deep side outputs. [C] // Computer Vision and Pattern Recognition (CVPR). 2016: 222–230.

[13] CHEN L.-C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017, 40 (4): 834–848.

[14] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions. [J]. ArXiv preprint arXiv:1511.07122, 2015.

[15] DING H, JIANG X, LIU A Q, et al. Boundary-aware feature propagation for scene segmentation. [C] // The IEEE International Conference on Computer Vision (ICCV). 2019: 6819–6829.

[16] SHUAI B, DING H, LIU T, et al. Toward achieving robust low-level and high-level scene parsing. [J]. IEEE Transactions on Image Processing, 2018, 28 (3): 1378–1390.

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

[18] CHEN L.-C, ZHU Y, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. [C] // European Conference on Computer Vision (ECCV). 2018: 801–818.

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

[20] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation. [C] // The IEEE International Conference on Computer Vision (ICCV). 2015: 1520–1528.

[21] DING H, JIANG X, SHUAI B, et al. Context contrasted feature and gated multi-scale aggregation for scene segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2018: 2393–2402.

[22] DING H, JIANG X, SHUAI B, et al. Semantic segmentation with context encoding and multi-path decoding. [J]. IEEE Transactions on Image Processing, 2020, 29: 3520–3533.

[23] CHEN L.-C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation. [J]. ArXiv preprint arXiv:1706.05587, 2017.

[24] MEHTA S, RASTEGARI M, CASPI A, et al. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. [C] // European Conference on Computer Vision (ECCV). 2018: 552–568.

[25] YANG M, YU K, ZHANG C, et al. Denseaspp for semantic segmentation in street scenes. [C] // Computer Vision and Pattern Recognition (CVPR). 2018: 3684–3692.

[26] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 4700–4708.

[27] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 2881–2890.

[28] HE J, DENG Z, ZHOU L, et al. Adaptive pyramid context network for semantic segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2019: 7519–7528.

[29] HOU Q, ZHANG L, CHENG M.-M, et al. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. [C] // Computer Vision and Pattern Recognition (CVPR). 2020.

[30] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks. [C] // Computer Vision and Pattern Recognition (CVPR). 2018: 7794–7803.

[31] RIEGLER G, OSMAN ULUSOY A, GEIGER A. Octnet: Learning deep 3d representations at high resolutions. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 3577–3586.

[32] LI X, ZHONG Z, WU J, et al. Expectation-maximization attention networks for semantic segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2019: 9167–9176.

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

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

[35] PASZKE A, CHAURASIA A, KIM S, et al. Enet: A deep neural network architecture for real-time semantic segmentation. [J]. ArXiv preprint arXiv:1606.02147, 2016.

[36] ZHAO H, QI X, SHEN X, et al. Icnet for real-time semantic segmentation on high-resolution images. [C] // European Conference on Computer Vision (ECCV). 2018: 405–420.

[37] ROMERA E, ALVAREZ J M, BERGASA L M, et al. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. [J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 19 (1): 263–272.

[38] YU C, WANG J, PENG C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation. [C] // European Conference on Computer Vision (ECCV). 2018: 325–341.

[39] GUPTA S, GIRSHICK R, ARBELÁEZ P, et al. Learning rich features from RGB-D images for object detection and segmentation. [C] // European Conference on Computer Vision (ECCV). Springer. 2014: 345–360.

[40] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. [C] // The IEEE International Conference on Computer Vision (ICCV). 2015: 2650–2658.

[41] HU X, YANG K, FEI L, et al. ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation. [J]. ArXiv preprint arXiv:1905.10089, 2019.

[42] SONG S, YU F, ZENG A, et al. Semantic scene completion from a single depth image. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 1746–1754.

[43] SONG S, XIAO J. Deep sliding shapes for amodal 3d object detection in rgb-d images. [C] // Computer Vision and Pattern Recognition (CVPR). 2016: 808–816.

[44] QI X, LIAO R, JIA J, et al. 3d graph neural networks for rgbd semantic segmentation. [C] // The IEEE International Conference on Computer Vision (ICCV). 2017: 5199–5208.

[45] CHENG Y, CAI R, LI Z, et al. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 3029–3037.

[46] LIN D, CHEN G, COHEN-OR D, et al. Cascaded Feature Network for Semantic Segmentation of RGB-D Images. [C] // The IEEE International Conference on Computer Vision (ICCV). 2017.

[47] WANG W, NEUMANN U. Depth-aware cnn for rgb-d segmentation. [C] // European Conference on Computer Vision (ECCV). 2018: 135–150.

[48] JIAO J, WEI Y, JIE Z, et al. Geometry-Aware Distillation for Indoor Semantic Segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2019: 2869–2878.

[49] WANG P, SHEN X, LIN Z, et al. Towards unified depth and semantic prediction from a single image. [C] // Computer Vision and Pattern Recognition (CVPR). 2015: 2800–2809.

[50] HOFFMAN J, GUPTA S, DARRELL T. Learning with side information through modality hallucination. [C] // Computer Vision and Pattern Recognition (CVPR). 2016: 826–834.

[51] KOKKINOS I. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 6129–6138.

[52] ZHANG Z, CUI Z, XU C, et al. Pattern-Affinitive Propagation Across Depth, Surface Normal and Semantic Segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2019.

[53] HE Y, CHIU W.-C, KEUPER M, et al. Std2p: Rgbd semantic segmentation using spatio-temporal data-driven pooling. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 4837–4846.

[54] WANG P, SHEN X, LIN Z, et al. Towards Unified Depth and Semantic Prediction From a Single Image. [C] // Computer Vision and Pattern Recognition (CVPR). 2015.

[55] KOKKINOS I. Ubernet: Training a Universal Convolutional Neural Network for Low-, Mid-, and High-Level Vision Using Diverse Datasets and Limited Memory. [C] // Computer Vision and Pattern Recognition (CVPR). 2017.

[56] MATURANA D, SCHERER S. Voxnet: A 3d convolutional neural network for real-time object recognition. [C] // Intelligent Robots and Systems (IROS). 2015: 922–928.

[57] QI C R, SU H, NIEßNER M, et al. Volumetric and multi-view cnns for object classification on 3d data. [C] // Computer Vision and Pattern Recognition (CVPR). 2016: 5648–5656.

[58] WANG P.-S, LIU Y, GUO Y.-X, et al. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. [J]. ACM Transactions on Graphics (TOG), 2017, 36 (4): 72.

[59] ENGELCKE M, RAO D, WANG D Z, et al. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. [C] // Robotics and Automation (ICRA). 2017: 1355–1361.

[60] GRAHAM B, ENGELCKE M, van der MAATEN L. 3d semantic segmentation with submanifold sparse convolutional networks. [C] // Computer Vision and Pattern Recognition (CVPR). 2018: 9224–9232.

[61] KLOKOV R, LEMPITSKY V. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. [C] // The IEEE International Conference on Computer Vision (ICCV). 2017: 863–872.

[62] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3d shape recognition. [C] // The IEEE International Conference on Computer Vision (ICCV). 2015: 945–953.

[63] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 652–660.

[64] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. [C] // Advances in Neural Information Processing Systems (NIPS). 2017: 5099–5108.

[65] LI J, CHEN B M, LEE G H. SO-Net: Self-Organizing Network for Point Cloud Analysis. [C] // Computer Vision and Pattern Recognition (CVPR). 2018: 9397–9406.

[66] HUANG Q, WANG W, NEUMANN U. Recurrent Slice Networks for 3D Segmentation of Point Clouds. [C] // Computer Vision and Pattern Recognition (CVPR). 2018: 2626–2635.

[67] SHEN Y, FENG C, YANG Y, et al. Mining point cloud local structures by kernel correlation and graph pooling. [C] // Computer Vision and Pattern Recognition (CVPR). 2018: 4548–4557.

[68] LI Y, BU R, SUN M, et al. PointCNN: Convolution On X-Transformed Points. [C] // Advances in Neural Information Processing Systems (NIPS). 2018: 828–838.

[69] SU H, JAMPANI V, SUN D, et al. Splatnet: Sparse lattice networks for point cloud processing. [C] // Computer Vision and Pattern Recognition (CVPR). 2018: 2530–2539.

[70] MONTI F, BOSCAINI D, MASCI J, et al. Geometric deep learning on graphs and manifolds using mixture model CNNs. [C] // Computer Vision and Pattern Recognition (CVPR). Vol. 1. 2017: 3.

[71] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering. [C] // Advances in Neural Information Processing Systems (NIPS). 2016: 3844–3852.

[72] YI L, SU H, GUO X, et al. SyncSpecCNN: Synchronized Spectral CNN for 3D Shape Segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2017: 6584–6592.

[73] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks. [C] // Advances in Neural Information Processing Systems (NIPS). 2015: 2017–2025.

[74] JIA X, DE BRABANDERE B, TUYTELAARS T, et al. Dynamic filter networks. [C] // Advances in Neural Information Processing Systems (NIPS). 2016: 667–675.

[75] LI X, WANG W, HU X, et al. Selective Kernel Networks. [C] // Computer Vision and Pattern Recognition (CVPR). 2019: 510–519.

[76] HU J, SHEN L, SUN G. Squeeze-and-Excitation Networks. [C] // Computer Vision and Pattern Recognition (CVPR). 2018.

[77] DING H, JIANG X, SHUAI B, et al. Semantic correlation promoted shape-variant context for segmentation. [C] // Computer Vision and Pattern Recognition (CVPR). 2019: 8885–8894.

[78] YUAN Y, WANG J. Ocnet: Object context network for scene parsing. [J]. ArXiv preprint arXiv:1809.00916, 2018.

[79] CHEN L.-Z, LI X.-Y, FAN D.-P, et al. LSANet: Feature Learning on Point Sets by Local Spatial Aware Layer. [J]. ArXiv preprint arXiv:1905.05442, 2019.

[80] XU Y, FAN T, XU M, et al. Spidercnn: Deep learning on point sets with parameterized convolutional filters. [C] // European Conference on Computer Vision (ECCV). 2018: 87–102.

[81] WANG C, SAMARI B, SIDDIQI K. Local spectral graph convolution for point set feature learning. [C] // European Conference on Computer Vision (ECCV). 2018: 52–66.

[82] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks. [C] // The IEEE International Conference on Computer Vision (ICCV). 2017: 764–773.

[83] ZHU X, HU H, LIN S, et al. Deformable convnets v2: More deformable, better results. [C] // Computer Vision and Pattern Recognition (CVPR). 2019: 9308–9316.

[84] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from rgbd images. [C] // European Conference on Computer Vision (ECCV). Springer. 2012: 746–760.

[85] SONG S, LICHTENBERG S P, XIAO J. Sun rgb-d: A rgb-d scene understanding benchmark suite. [C] // Computer Vision and Pattern Recognition (CVPR). 2015: 567–576.

[86] JANOCH A, KARAYEV S, JIA Y, et al. A category-level 3d object dataset: Putting the kinect to work. [G] // Consumer depth cameras for computer vision. Springer, 2013: 141–165.

[87] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition. [C] // Computer Vision and Pattern Recognition (CVPR). 2016: 770–778.

[88] CORDTS M, OMRAN M, RAMOS S, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding. [C] // Computer Vision and Pattern Recognition (CVPR). 2016.

[89] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from rgbd images. [C] // European Conference on Computer Vision (ECCV). Springer. 2012: 746–760.

[90] JIANG J, ZHENG L, LUO F, et al. RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation. [J]. ArXiv preprint arXiv:1806.01054, 2018.

致谢

三年的研究生时光转瞬即逝。同时也标志着我十九年求学生涯的结束。这一路走来，欢笑与泪水并存。没有痛苦相伴的成长是毫无意义的，倘若不牺牲些什么，就什么也得不到。借此机会，我想感谢在我的人生道路上，帮助过我的所有人。

首先，我要感谢我的导师程明明教授。程老师是我硕士期间学术方向的指导者，是他带领我从一个学术小白，到可以进行学术研究的硕士毕业生。本文也是在程老师的指导下完成。程老师严谨认真的科研作风和习惯深深的影响了我。借此机会，我也要感谢王恺老师，卢少平老师和巴斯大学的杨永亮老师对我在学术上的悉心指导和帮助。杨永亮老师在论文上每一条中肯实用的建议都让我受益匪浅。同时，我也要感谢所有给予我帮助的任课老师，每一堂课都是我珍贵的回忆。感谢老师们的无私付出和教诲。

接着，感谢我硕士期间论文合作者们：林铮，李炫毅，汪子钦，张钊和范登平师兄。和你们的合作让我收获良多，我们一起的坚持和努力最终都有了回报。感谢媒体计算实验室所有一起奋斗过的小伙伴们，大家一起闲聊，聚餐的场景还历历在目。祝大家都有着光明的前程和美好的未来。感谢李炫毅，林铮，吴宇寰，谭永强，张宇，许刚，高尚华，张长彬，李振，朱子悦，以及姜鹏涛师兄，曹洋师兄，胡晓伟师兄对我的帮助和支持。在此对你们表达衷心的感谢和祝福，和你们做同学是我的荣幸。

最后，我要感谢一直以来支持我的朋友和家人们。感谢我的好朋友秦李浩对我一直以来的关怀，帮助和支持，他对事物的建议和看法总是会让我豁然开朗，希望他能一切顺利。感谢我的女朋友对我的陪伴，关心和鼓励。感谢我的爸爸妈妈，他们的对我的养育之恩，无条件的帮助和扶持是我求学路上的精神支柱。

个人简历

- 南开大学，计算机科学与技术专业，硕士，2018.09-2021.06
- 西安电子科技大学，电子信息工程专业，学士，2014.08-2018.06

研究生期间发表论文:

- **Lin-Zhuo Chen**, Zheng Lin, Ziqin Wang, Yong-Liang Yang, Ming-Ming Cheng. Spatial Information Guided Convolution for Real-Time RGBD Semantic Segmentation, IEEE Transactions on Image Processing (TIP), 2021 (第一作者，CCF A 类，SCI 一区)
- Zheng Lin, Zhao Zhang, **Lin-Zhuo Chen**, Ming-Ming Cheng, Shao-Ping Lu. Interactive Image Segmentation with First Click Attention, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020 (第三作者，CCF A 类)

在申专利:

- 程明明，**陈林卓**，李炫毅，基于空间注意力机制的点云特征提取方法，申请号：201910235177.2，申请日：2019-03-27
- 程明明，**陈林卓**，林铮，一种利用空间信息提升语义分割精度的高效方法，申请号：202010031390.4，申请日：2020-01-13

研究生期间其它成果:

- 南开大学新生奖学金
- 南开大学研究生公能奖学金二等