分类号：TP391.4　　学校代码：10363
密　级：公开　　学　　号：2210320103

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于 SwiftNet 面向室内 RGB-D 场景的高效语义分割研究

论 文 作 者　王博
指 导 教 师　许钢
学科（专业）　控制科学与工程
研 究 方 向　智能信息处理及应用

论文提交日期：2024 年 5 月 31 日

分类号：TP391.4　　单位代码：10363

密　级：公开　　学　　号：2210320103

题　　目 基于 SwiftNet 面向室内 RGB-D 场景的高效语义分割研究

英文并列

题　　目 Research on Efficient Semantic Segmentation for Indoor RGB-D Scenes Based on SwiftNet

学 生 姓 名：王 博

指 导 教 师：许 钢（教 授）

专　　业：控制科学与工程

研 究 方 向：智能信息处理及应用

论文答辩日期：2024 年 05 月 23 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：王博

日期：2024 年 5 月 31 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：王博　　指导教师签名：

日期： 2024 年 5 月 31 日　　日期： 2024 年 5 月 31 日

基于 SwiftNet 面向室内 RGB-D 场景的高效语义分割研究

摘 要

室内场景语义分割是实现智能家居移动机器人环境智能感知理解的关键技术。由于室内场景环境的复杂性和光线照明的多变性，传统的基于单 RGB 的方法难以准确进行分割。随着深度学习技术的突破和高性能传感器的普及，RGB-D 数据（结合彩色图像与深度信息）已经成为获取室内环境丰富信息的关键手段。这种数据不仅包含了传统的彩色图像信息，还提供了每个像素的深度信息，引入深度数据能够显著提高分割精度。此外，实时性和效率也是室内 RGB-D 语义分割研究中的重要考虑因素，尤其是对于需要快速响应和处理的实际应用场景。同时，为了在后续移动设备上应用，分割效率也是必须考虑的因素。本研究针对上述问题提出了一种基于神经网络的高效语义分割方法，本文的主要研究内容如下。

（1）首先针对复杂室内场景使用单 RGB 图像分割精度低的问题，研究在多尺度道路场景轻量化网络 SwiftNet 中引入深度数据，以提高分割精度。通过利用深度图像的颜色稳定性和距离信息，减轻光线、颜色和距离等因素对结果的影响，实现高效的多模态语义分割。采用了一维非瓶颈块简化网络参数，同时改进了深度聚合金字塔池化模块以提高分割效率，称之为快速聚合金字塔池化模块（FAPPM）。在 NYUDv2 和 SUN RGB-D 数据集上分割精度指标 mIoU 分别达到 50.9%和 48.5%，领先于大多数先进语义分割算法模型。

（2）其次针对深度数据的空间几何和形状信息提取问题，引入了深度特征形状卷积块，以充分利用深度数据的形状信息。将深度局部区域分解为基础分量和形状分量，并通过基础内核和形状内核进行处理。最终通过元素相加的方式将处理后的分量组合起来，得到形状

感知图像块，有利于后续的语义分割任务。在 NYUD v2 和 SUN RGB-D 数据集上，改进后的算法将 mIoU 提升了 1.34%和 2.33%，同时提升了对形状的感知能力。

（3）最后为了解决特征融合学习中部分特征通道被忽略的问题，采用了加权通道交换方法。通过批量正则化的缩放因子来衡量每个通道的重要性，并用另一模态的平均值替换与接近零的因子相关的通道。通过共享卷积滤波器，不同模态之间的通道被嵌入相同的映射，以更好地对模态共同的统计进行建模。引入稀疏性约束，促使模型自动学习哪些通道是重要的，哪些可以被忽略。未充分利用的通道重新进入训练更新部分，促进了不同模态间的信息交互和共享，改善了分割性能。在 NYUD v2 和 SUN RGB-D 数据集上，相较于基线模型 mIoU 分别提高了 1.94%和 0.60%，进一步验证了该模型的有效性。

关键词：室内 RGB-D 场景，多尺度特征融合，形状特征提取，加权通道交换，语义分割

Research on Efficient Semantic Segmentation for Indoor RGB-D Scenes Based on SwiftNet

ABSTRACT

Indoor scene semantic segmentation is a crucial technology for enabling intelligent perception and understanding of the environment in smart home mobile robots. Due to the complexity of indoor environments and the variability of lighting conditions, traditional RGB-based methods struggle to achieve accurate segmentation. With the breakthroughs in deep learning and the widespread use of high-performance sensors, RGB-D data (combining color images with depth information) has become a key means of acquiring rich information about indoor environments. This data not only includes traditional color image information but also provides depth information for each pixel, significantly improving segmentation accuracy. Additionally, real-time performance and efficiency are important considerations in indoor RGB-D semantic segmentation research, especially for practical applications that require quick response and processing. Furthermore, segmentation efficiency is a crucial factor for future applications on mobile devices. This study addresses these issues by proposing an efficient semantic segmentation method based on neural networks.

Firstly, to address the issue of low segmentation accuracy using single RGB images in complex indoor scenes, we investigate the integration of depth data into the lightweight SwiftNet network tailored for multiscale road scenes to enhance segmentation accuracy. By leveraging the color stability and distance information provided by depth images, we mitigate the impact of factors such as lighting, color, and distance on the

segmentation results, achieving efficient multimodal semantic segmentation. We employ one-dimensional non-bottleneck blocks to simplify network parameters and enhance segmentation efficiency. Additionally, we improve the depth aggregation pyramid pooling module to parallelize context extraction, thereby enhancing segmentation efficiency, referred to as the Fast Aggregation Pyramid Pooling Module (FAPPM). On the NYUDv2 and SUN RGB-D datasets, the segmentation accuracy, as measured by the mIoU metric, reaches 50.9% and 48.5%, respectively, outperforming most state-of-the-art semantic segmentation algorithms.

Secondly, to address the issue of extracting spatial geometry and shape information from depth data, we introduce the Depth Feature Shape Convolutional Block to fully utilize the shape information embedded in the depth data. We decompose the depth local regions into base components and shape components, which are then processed separately using base and shape kernels. Finally, we combine the processed components through element-wise addition to obtain shape-aware image blocks, which are beneficial for subsequent semantic segmentation tasks. On the NYUD v2 and SUN RGB-D datasets, the improved algorithm enhances the mIoU by 1.34% and 2.33%, respectively, while also enhancing the perception of shapes.

Lastly, to address the problem of some feature channels being ignored during feature fusion learning, we adopted a weighted channel exchange approach. This method utilizes batch normalization scaling factors to measure the importance of each channel and replaces channels with factors close to zero with the average value of another modality. By sharing convolutional filters, channels from different modalities are embedded into the same mapping, enabling better modeling of shared statistics across modalities. Introducing sparsity constraints prompts the model to

automatically learn which channels are important and which can be ignored. Unused channels are reintroduced into the training update process, facilitating information exchange and sharing among different modalities, thereby improving segmentation performance. On the NYUD v2 and SUN RGB-D datasets, compared to the baseline model, the mIoU was increased by 1.94% and 0.60%, respectively, further confirming the effectiveness of this model.

Wang Bo(Control Science and Engineering)

Supervised by Professor Xu Gang

KEY WORDS: Indoor RGB-D scene, Multi-scale feature fusion, Shape feature extraction, Weighted channel exchange, Semantic segmentation

目录

第1章 绪论

1.1 研究背景与意义

随着技术的进步和社会的发展，图像理解变得越来越关键。随着图像数据的激增，传统的人工分析已经难以满足日益增长的需求。在这种情况下，计算机视觉技术应运而生，并在图像的智能化处理中扮演着至关重要的角色。其目的是利用计算机模拟人类视觉，实现对图像的高效处理。同时，图像语义分割是计算机视觉领域的一项重要任务，其目的是对输入图像的每个像素点进行语义标签的预测，进而解析图像所呈现的整体场景。语义分割技术对于在很多领域展现出巨大价值，包括车辆智能驾驶[1]、医疗影像分析[2]、道路交通规划[3]、智能家居[4]等。

在自动驾驶方面，图像语义分割可以用于实现道路场景中车辆和行人的检测和跟踪。主要通过将图像中的不同对象进行分类和标注，为自动驾驶系统提供更加准确和可靠的数据支持，帮助智能车机系统更好地理解和避免潜在的危险。此外，语义分割技术还可辅助自动驾驶系统理解和解析交通场景，为车辆的自主导航提供关键信息。在医疗诊断领域，图像语义分割作为一项有益辅助，可帮助医生进行疾病的诊断和治疗规划。通过对医学影像数据进行语义分割，能够自动检测病变区域，有助于医生更精确地评估病情并制定治疗策略。这一技术不仅有望提升医学诊断的精准度和效率，同时为医生提供客观、全面的医学影像分析工具。在交通规划方面，图像语义分割可用于对道路交通场景进行分类和分析。通过对交通监控视频数据进行语义分割，可以自动识别出车辆、行人、车道线等不同对象，为交通管理和安全预警提供数据支持。此外，语义分割技术还可用于交通场景的态势感知和解析，为智能交通系统的实现提供必要技术支持。在智能家居领域，通过对室内图像进行语义分割可以实现室内不同区域和空间的自动识别，如客厅、卧室、厨房等，为智能家居系统提供更精准的空间感知和环境理解。这一技术的发展有望推动智能家居导航、环境监测以及自动化控制等功能的智能化发展。

传统的图像语义分割方法通常需要手动提取低级特征，这一过程耗时且容易造成信息丢失，对模型性能产生不利影响。近年来，卷积神经网络[5]（CNN）在

图像分割中得到了广泛的应用，高效提取图像特征的结构取得了显著的进展。尽管如此，特殊场景条件如光照不均、复杂遮挡、物体多样性和相似颜色纹理等仍然是挑战。特别是在室内场景中，受到家居风格、光照条件等因素的影响，物体颜色相近、种类繁多、易于遮挡，这导致了语义分割面临着分割边缘模糊和识别错误等问题。因此，本文旨在深入分析深度学习在图像语义分割中的突破，并探讨室内场景条件下的挑战及应对策略。

多项研究已经指出，仅使用 RGB 图像进行室内场景的分割难以实现高准确度[6-8]。随着 Kinect、Xtion 等 3D 传感器的大量成熟的应用，获取物体的三维几何信息变得更加便捷，推动了 RGB-D 语义分割技术的发展。通过对真实世界几何信息进行编码，RGB-D 图像能够克服 2D 仅在投影图像空间中显示光度外观属性的限制，并且丰富了 RGB 图像的表征。RGB 和深度图像呈现信息的方式完全不同。RGB 图像捕获了投影图像空间中的光度外观属性，而深度图则提供了丰富的补充信息，可用于获得局部几何的外观信息。因此，在语义分割任务中，增强和融合 RGB 和深度数据的优势至关重要。RGB-D 图像的融合也为更全面的语义理解提供了可能。通过将 RGB 和深度数据进行有效融合，可以实现对物体的更精确的边界和细节的捕捉，从而改善分割结果的准确性。这种融合策略通常涉及到特征的提取、融合和分类等多个步骤，需要综合考虑不同数据源的特点和互补性。

如图 1-1 所示，分别展示 RGB、Depth、HHA 以及标签图像，RGB 图像中，每个像素由红、绿、蓝三个颜色通道的值组成，用于捕捉场景的颜色信息。Depth 图像则表示场景中每个像素到摄像头的距离或深度信息。HHA 是基于深度图像的特征表示，包括水平距离、垂直高度和表面法线三个组成部分，能够有效地捕捉场景的几何结构和空间布局信息。而 Label 图像表示场景中每个像素或区域的语义类别或标签，在语义分割任务中，每个像素被标记为特定的类别，如“门”、“窗”、“桌子”等。通过对标签图像进行学习和推理，模型可以识别并分割出场景中不同类别的物体和区域，从而实现对场景的语义理解。

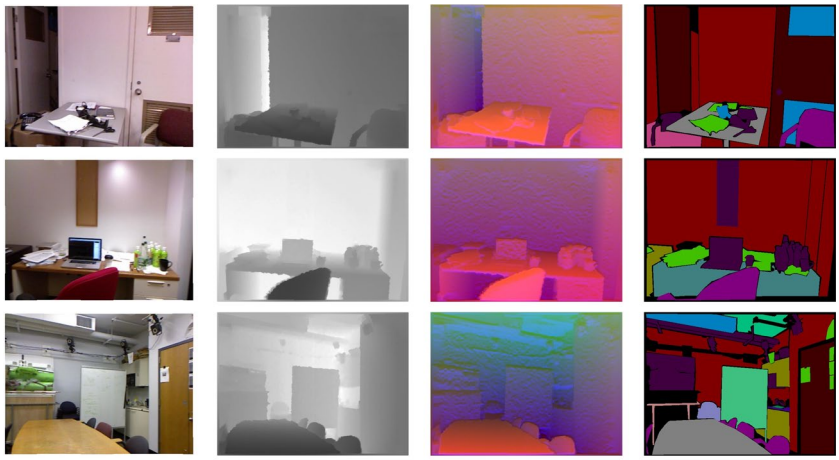


图 1-1 基于 RGB-D 语义分割图像

室内场景涵盖了家居、商场、展馆、医院和办公楼等多样化环境，人们在这些场所生活和工作的时间较长。因此，对室内场景图像进行语义分割具有重要的研究意义。首先，它有助于理解和解析室内环境，为智能家居、机器人导航等领域提供基础性支持。其次，在实际应用中具有广泛的前景，如智能监控、安防、增强现实等领域。此外，随着人工智能技术在社会生活中的应用，室内 RGB-D 场景语义分割的研究能够更客观地解析室内场景的复杂结构，实现对场景中不同物体的精确识别与分类，还有助于推动相关领域如机器人导航、智能家居等的技术进步，深入探索室内 RGB-D 场景的语义分割技术，对于促进视觉技术的发展与应用具有重要的科研价值。

1.2 国内外研究现状

在图像语义分割领域，国内外科研人员已经进行了大量深入的研究，主要的研究算法可分为两大类。首先是传统图像语义分割算法方面，这些算法主要依赖于人工设计的特征提取和基于规则的分割技术。虽然这些方法在简单场景中表现出色，但在面对复杂场景时存在一定局限性。其次是基于深度学习的算法方面，特别是卷积神经网络（CNN）和 U-Net 结构等代表性模型。这些算法实现了更精准的像素级别分割，并且随着研究者逐渐引入 RGB-D 图像，并结合颜色和深度信息，分割的准确性和鲁棒性得到了进一步提升。本节将重点介绍这两个方面的研究进展。

1.2.1 传统图像语义分割

传统图像语义分割是计算机视觉领域的经典研究方向之一，它致力于将图像分割成具有语义信息的区域[9]，为计算机对图像内容的深入理解和高级认知提供了基础。在这一领域中，研究者们主要利用传统的计算机视觉技术，如图像处理、特征提取和分类等方法，尝试解决图像中不同物体和场景的精确分割问题[10, 11]。

首先，基于阈值的分割[12]是一种简单而直观的方法。通过设定像素值或颜色阈值将图像分割成具有不同特征的区域。然而，这种方法在处理复杂图像，特别是包含多样化颜色和纹理的场景时效果有限。其次，区域生长方法是一种基于相似性的分割技术[13]，从种子点开始逐渐将相邻像素合并成相似区域。这种方法在纹理较为明显、区域较为规则的情况下表现较好，但对于复杂纹理和不规则物体边界的处理能力较弱。边缘检测与分割是另一类常见的传统方法，它通过边缘检测算法（如 Canny 算法[14]）找到图像中的边缘信息，并利用边缘将图像分割成不同的区域。尽管边缘提供了一些物体边界的信息，但这种方法容易受到噪声和干扰的影响。

图割算法是一种基于图论的分割技术，它将图像表示为图结构，并通过最小割或最大流等方法实现图像的分割[15]。这种方法在处理一些复杂场景的语义分割具有一定优势，但计算复杂度相对较高，在实际应用中有很大的挑战。另一种常见的方法是颜色空间变换，它将图像从 RGB 颜色空间转换到其他颜色空间，如 HSV 以利用不同颜色通道的信息进行分割[16]。这种方法在一定程度上解决了颜色信息对分割的影响，但对于光照变化等问题仍然较为敏感。

形态学分割利用形态学运算（如腐蚀和膨胀）识别图像中的物体边界和区域，适用于处理形状规则的物体，然而，它在处理复杂场景时存在一定的局限性。基于纹理的方法则通过分析像素的纹理特征进行分割，对于包含大量纹理信息的图像更具适应性。传统的图像语义分割方法在简单场景中表现良好，但在面对复杂、多样化的真实场景时，其性能相对有限。

1.2.2 基于深度学习的 RGB-D 图像的语义分割

传统图像语义分割方法在一定程度上能够解决图像分割问题，但在处理复杂

场景和具有多样化特征的图像时存在局限性，因此，基于深度学习的方法逐渐成为了图像语义分割的主流研究方向。在国际上，RGB 场景高效语义分割引起了广泛关注，随着卷积神经网络的发展，许多研究者借助深度学习技术取得了显著的成果。首先，在数据集构建方面做出了重要贡献，推动了语义分割领域的发展。大规模标注的 RGB-D 数据集，如 NYU v2、SUN RGB-D 等，为算法的训练提供了充足的数据支持。此外，许多国际研究团队在模型设计上进行了创新，提出了一系列高效的网络架构。例如，DeepLab 系列[17-20]通过结合深度卷积网络、全连接条件随机场（CRF）和空洞卷积来处理多尺度上下文信息，提高分割准确性或引入可分离空洞卷积以降低计算复杂度，从而提高模型效率和实用性。U-Net 系列[21-24]等网络在语义分割任务上表现出色，通过独特的 U 形结构，结合图像的全局和局部信息或采用三维卷积处理三维医学图像，增强了上下文信息的捕捉能力，或通过更深的网络结构和残差连接提高分割性能。引入注意力机制模块，有效提高增强模型对关键区域的识别能力。在实时性方面，一些研究者还专注于设计轻量级网络结构，如 ICNet[25]、ENet[26]、BiSeNet[27]，以在嵌入式设备上实现实时语义分割。还有基于深度学习技术的目标检测扩展版本，YOLOv5-seg[72]保持了快速的实时目标检测特性，通过深度卷积神经网络和端到端训练，它实现了优秀的语义分割效果，在处理实时图像时表现出色。RTMDet-seg[73]是基于实时目标检测模型的语义分割扩展。RTMDet-seg 保留了快速推理速度和高准确性的优点，同时引入了语义分割任务，为实时应用提供更精细的图像理解和处理。

此外，对于 RGB-D 场景的独特性质，一些研究者还尝试将深度信息与 RGB 信息充分融合，以提升语义分割的准确性[28-31]。深度信息不仅用于增强物体边缘的分割效果，还可以在遮挡等复杂场景中提供额外的信息。然而，国际上的研究也面临一些挑战，例如对于实时性和模型轻量化的要求，以及在复杂场景中的鲁棒性等问题。这些问题仍然是当前国际研究面临的热点和难点。

研究人员正致力于通过多种手段提升语义分割的准确性、鲁棒性及实时性能。包括数据增强、多尺度特征融合以及模型轻量化等方法。数据增强技术[32]被广泛采用，通过对有限的数据进行处理来扩充训练数据集以增强模型的泛化能力，特别是在训练样本有限的情况下效果最佳。此外，多尺度特征融合技术[33]将来自不同尺度的特征整合到一个统一的特征表示，网络会提取输入图像的不同尺度特征，这些特征分别反映了图像在不同层次上的信息，从而在不同尺度上捕捉语义信息。

例如，一种简化的多尺度特征融合网络 SwiftNet[34]提出了一种新颖的方法，在多个分辨率上融合共享特征，以提升语义分割的准确性和效率。虽然这些方法在某些情况下取得了一定的成功，但大多数都只关注 RGB 图像输入，根据文献[35]所述，RGB 数据蕴含了丰富颜色和纹理信息，但缺乏对空间几何信息的直接描述，仅凭借 RGB 数据往往难以区分具有相似颜色和纹理的物体实例，也难以理解在环境中的上下文。多模态数据融合[36]已被证实为一种有效的策略。具体来说，深度图像能够提供每个像素点的距离，能够准确反映出位置、形状等几何特征。

在国内，RGB-D 场景高效语义分割的研究也取得了显著的进展。研究[37-39]表明，将深度信息直接应用于现有的 RGB 框架，或者简单地将两种模态的数据融合，往往难以获得理想的分割效果。近年来，研究人员提出了多种方法来提升 RGB-D 语义分割的性能，其中有，针对 RGB-D 数据设计专用的网络架构[40-43]，例如跨模态注意力融合网络[41]采用双分支编码器结构来提取多尺度的 RGB 和深度特征。ACNet[42]提出了一个注意互补模块(ACM)选择性地从 RGB 和深度分支中收集特征，使网络能够有效地处理 RGB 和深度图像在特征分布上存在的显著差异。CENNet[43]通过交换前向传播过程中动态的交换通道以充分利用所有通道信息。还有一项研究[44]通过将深度数据进行 HHA 编码转换，获取包含物体边缘、立体特征、场景结构和相机视角下的空间布局等信息的特征表示，提高对深度信息的利用。尽管之前的研究在 RGB-D 语义分割中取得显著效果，但多数使用相同卷积算子平等地提取 RGB 和深度特征，未考虑它们的本质差异。一些研究提出了特定定制的卷积方法[45-49]，如通过 2.5D 卷积核[45]或改进 2.5D 卷积[46]在特征图上同时捕获空间和深度信息，以及空间信息引导卷积[48]结合空间和深度信息来指导卷积操作，以优化语义分割的性能。此外，还有使用 3D 邻域卷积[49]构建 3D 邻域以捕获像素之间的空间关系。然而，这些技术可能会带来计算开销的增加，在实际的室内场景 RGB-D 语义分割应用中，不仅需要高精度，还需要尽可能地轻量化，以满足在嵌入式系统和移动设备上实现实时性能的需求。

为了满足实时或移动需求，并在推断速度和准确性之间找到最佳平衡，研究者提出了许多高效和有效的语义分割模型。这些模型大致可分为两类：一是轻量级编码器和解码器，二是双分支网络架构。具体而言，SwiftNet 采用低分辨率输入获取高级语义，并使用高分辨率输入为其轻量级解码器提供足够的细节。BiSeNet[50]提出了一种新的双边分割网络，通过设计空间路径和语义路径，结合

特征融合模块，实现实时语义分割的同时保持高精度。与此同时，现有的许多 RGB-D 分割方法[51-54]通常采用双分支结构，分别用于处理 RGB 图像和深度图像，这种设计使得每个分支能够专注于提取与其对应模态相关的特征。然而，“先提取，后融合”的策略虽然在一定程度上实现了模态间的互补，但也可能忽略了在特征提取过程中跨模态信息的交互。FuseNet[55]的研究结果表明，多阶段特征融合可以显著提高模型性能，这意味着在不同层次上融合特征能够更好地捕获图像语义信息，从而提高了分割的准确性。

以上这些方法极大地促进了 RGB-D 图像语义分割技术的发展，RGB-D 图像语义分割方法的精度和泛化性能都得到了有效提高。然而，这些方法仍然存在一定的局限性，对于实时性和模型轻量化的要求，以及在复杂场景中的鲁棒性等问题仍然是需要解决的热点和难点。因此，本文将以这些方法特别是编解码网络为基础，针对室内场景的具体问题展开研究，提出创新性的解决方案，来进一步提升室内 RGB-D 语义分割任务的性能。

1.3 主要研究内容及章节安排

1.3.1 主要研究内容

针对室内场景，本文首先提出了一种基于 SwiftNet 面向室内 RGB-D 场景语义分割算法，该算法模型设计理念借鉴了其在道路 RGB 场景上的成功应用，但经过对室内场景特性的深入理解和调整，使其在室内 RGB-D 场景中表现更为出色。首先，进行 RGB 图像和深度图像的特征提取，然后将它们融合，在经过解码器和上采样的处理后进行输出。结果显示，模型取得了领先的分割效果。

后来针对现有方法中忽略 RGB 和深度特征之间内在差异的问题，通过将深度特征分解为形状分量和基础分量，并独立地处理局部特征，以更有效地利用深度特征中的形状信息，使用形状卷积算子将深度特征分解为形状分量和基础分量，引入两个可学习的权重，以独立地与形状分量和基础分量组合，对重新加权的组合进行卷积，通过这种方式能够同时捕获形状和基础位置信息，并利用它们进行语义分割。实验结果显示，相较于基线分割算法，改进后的算法在 NYU v2 和 SUN RGB-D 两个数据集上的最终语义分割精度指标均有所提升，同时并未增加

额外的计算或内存需求。

最后，针对现有方法在平衡跨模态融合和单模态处理之间的权衡方面存在不足，引入了一种新的多模态融合方法。该方法通过动态交换不同模态子网络之间的通道，以实现更好的模态间和模态内信息的融合。具体而言，通道交换过程是由每个通道的重要性所引导的，重要性由训练期间批归一化缩放因子的幅度来衡量。为了进一步提高模型的效率和解释性，引入了稀疏性约束。通过在模型的设计中引入稀疏性约束，鼓励网络学习具有稀疏性质的特征表示，从而降低模型的复杂度并提高泛化能力。具体来说，通过在损失函数中添加对通道交换后的通道表示的 L1 正则化项。同时，通过共享卷积滤波器并保持单独的批归一化层，使得多模态网络几乎与单模态网络一样紧凑，从而进一步提高模型的效率。

1.3.2 章节安排

第 1 章为绪论部分，简要介绍室内 RGB-D 场景语义分割的研究背景与意义，概述当前国内外的研究现状与发展趋势，明确本研究的目的和主要研究内容，强调了本课题对于实现更精准的室内场景理解与应用的重要性，以及在实际场景中对智能系统的推动作用。并对整篇论文的结构安排进行说明。

第 2 章详细介绍了与室内 RGB-D 场景语义分割相关的基础理论和相关技术，首先简述了深度学习的原理，并对卷积神经网络的组成进行了介绍，其次重点阐述深度学习在语义分割中的应用，包括经典的语义分割算法、主流数据集以及评价指标等，为后续的算法实验提供了数据支撑。

第 3 章介绍了基于 SwiftNet 的室内 RGB-D 场景语义分割网络，首先介绍了一种高效双分支多尺度语义分割算法 SwiftNet。接着，简要介绍了快速聚合上下文模块、一维非瓶颈模块以及注意力融合模块等。然后，将详细阐述基于 SwiftNet 的语义分割网络的结构和训练方法。最后，将通过实验设计和结果分析来验证该网络在室内 RGB-D 场景语义分割任务中的有效性。

第 4 章介绍了基于形状感知卷积的室内场景 RGB-D 图像高效语义分割算法，是由第三章的算法模型改进得到，首先，通过对先前提到的 SwiftNet 算法的不足之处进行分析，引出了形状感知卷积的改进思路，然后介绍了形状感知卷积和可学习上采样的基本原理，最后进行实验分析并讨论了算法模型的实验结果。

第 5 章介绍了基于自适应阈值通道交换的室内场景 RGB-D 图像高效语义分割算法，通过对前文算法的不足进行分析，引入自适应阈值通道交换的改进思路。详细介绍了自适应阈值通道交换的原理和网络整体结构，最后进行实验分析并讨论了算法模型的实验结果。

第 6 章是总结和展望，首先对全文的主要内容和成果进行简明扼要的概括，分点概括了本研究的采用的方法和主要结果。同时，还将对研究中存在的不足和局限性进行分析，讨论本研究的潜在应用前景和社会影响，并提出未来的研究方向和展望。

第2章 基础理论与相关技术

随着深度学习的蓬勃发展，其在计算机视觉领域的应用日益广泛，尤其在RGB-D 图像的语义分割任务中展现出巨大的潜力。图像语义分割是计算机视觉领域的重要任务之一，深度学习技术在语义分割中取得了显著的突破，并成为目前最为有效的方法之一。本章首先介绍深度学习的基础理论，包括本文的主干网络 ResNet，其次对将深入探讨语义分割的相关技术，如语义分割流程、数据集、评价指标、损失函数进行了详细阐述。

2.1 深度学习基础

深度学习是机器学习的一个分支，其核心在于构建并训练深度神经网络模型，使其能够从大量数据中自动学习并提取出有用的特征表示。这种方法对于处理海量数据、复杂的非线性关系以及自我学习具有显著的优势。深度学习的应用领域广泛，包括图像识别、语音识别、自然语言处理等。深度学习起源于 20 世纪 80 年代，当时神经网络模型开始被用于解决一些复杂的问题，如图像和语音识别。然而由于当时计算机硬件性能的限制以及可用于训练的数据量较小，深度学习的研究在很长一段时间内并未取得显著的进展。直到近年来，随着计算机硬件性能的显著提升和大数据的快速增长，深度学习才得以实现突破性的发展。

人工神经网络是深度学习的基本模型，特别是深度神经网络。一个典型的深度神经网络包括输入层、隐藏层和输出层。输入层负责接收外部输入的数据，隐藏层通过一系列非线性变换将输入数据转换为更高级的特征表示，最后输出层将隐藏层的结果输出为最终的预测结果。深度神经网络的一个重要特点是其非线性变换能力。通过使用非线性激活函数，如 ReLU[56]、sigmoid[57]等，深度神经网络能够学习和识别复杂的非线性模式。此外，深度神经网络还具有自动特征学习的能力，能够从原始数据中自动提取有用的特征，而无需人工进行特征工程。

主要训练方法主要包括前向传播和反向传播算法。前向传播算法是将输入数据通过神经网络得到输出结果的过程。反向传播算法则是根据输出结果与真实结果的损失反向调整神经网络的权重和偏置值的过程。反向传播算法的核心思想是梯度下降法。具体来说，算法通过计算输出层和真实数据之间的损失，然后反向

传播这个损失，更新网络的权重和偏置值以最小化损失。在训练过程中，网络通过不断地调整权重和偏置值，逐渐学习到对训练数据的有用特征和模式，从而提高其对新数据的分类和识别能力。深度学习的优势在于其对复杂数据的处理能力。通过构建多层神经网络，深度学习能够捕捉到数据中的复杂特征和模式，从而实现更高的分类和识别准确率。此外，深度学习还具有自动特征学习的能力，能够从原始数据中自动提取有用的特征，避免了手工进行特征选取的繁琐过程。

2.2 卷积神经网络基础

在计算机视觉领域，深度学习已经取得了显著的进展，尤其是卷积神经网络的应用。CNN 是一种模拟人脑神经元连接方式的深度学习算法，适用于处理图像和视觉任务。在室内 RGB-D 场景的语义分割任务中，CNN 被广泛应用于特征提取和分类。本章将详细介绍 CNN 的基本概念、原理及组成以及损失函数等相关技术。

2.2.1 卷积神经网络概念

卷积神经网络（CNN）是深度学习领域中最具影响力的网络结构之一。其概念的起源可以追溯至上世纪 60 年代，当时 Hubel 和 Wiesel 通过对猫视觉皮层细胞的研究[58]，提出了感受野这一概念，感受野是指视觉皮层中的神经元对视觉输入空间的子区域的敏感程度。这一概念为后来的卷积神经网络提供了理论基础。到 80 年代，Fukushima 在感受野概念的基础之上提出了神经认知机的概念[59]，通过模拟动物视觉系统结构，被视为卷积神经网络的初步实现，其位置特征的平移不变性和对形状畸变的不敏感性使其在图像处理和计算机视觉任务中具有广泛的应用前景，从而推动了视觉系统模型化的发展。

1998 年，纽约大学的 Yann Lecun 等人正式提出了卷积神经网络的概念，并成功将其应用于手写数字识别等问题[60]。随着时间的推移，CNN 不断被改进和扩展，以适应更广泛的应用场景和问题。研究人员提出了各种变体和改进，如深度 CNN、残差网络、注意力机制等，进一步提升了模型的性能和效率。CNN 的成功也激发了对神经网络其他方面的研究，如循环神经网络（RNN）和生成对抗网络（GAN）等。

除了图像识别领域，CNN 还被成功应用于语音识别、自然语言处理等领域，拓展了其在人工智能领域的应用范围。随着硬件技术的不断进步和深度学习算法的不断演进，CNN 在各个领域都取得了巨大的成功，并且成为了当今人工智能研究和应用中不可或缺的重要组成部分。

综上所述，卷积神经网络是一种基于生物视觉机制的深度学习网络结构，它通过局部连接和权值共享的方式降低了模型的复杂度，提高了运算效率，特别适用于处理图像等二维数据。随着研究的深入和技术的不断发展，卷积神经网络在人工智能领域的应用将会越来越广泛。

2.2.2 卷积神经网络原理及组成

卷积神经网络模拟了人脑神经元之间的连接方式，即通过卷积运算对输入数据进行特征提取。在 CNN 中，每个神经元都与输入数据的一个局部区域相连，并通过卷积运算提取该局部区域中的特征。这种局部连接方式有效地减少了模型的参数数量，提高了模型的泛化能力。

CNN 的设计受到生物视觉系统的启发，特别是对视觉皮层局部感受野的理解。局部感受野意味着视觉神经元仅对其视野中的小部分刺激做出响应。此思想在 CNN 中通过卷积层得以运用，每个卷积层的神经元也只响应前一层中的一小区域，即感受野。卷积神经网络原理主要基于卷积运算和神经网络的结构。卷积运算是一种特殊的矩阵运算，主要用于图像处理和模式识别等领域。在 CNN 中，卷积运算被用来提取输入数据中的局部特征，并通过共享权重的方式实现对全局特征的捕捉。其基本结构有输入层、卷积层、激活函数层、池化层和全连接层等组成部分，这些层次协同工作，共同完成特征提取和分类任务。

输入层：负责接收原始数据。在图像处理任务中，输入层通常表示输入图像。每个输入神经元对应输入数据的一个特征或一个像素值。输入层的主要作用是将原始数据转换成网络能够处理的格式。

卷积层：CNN 的核心部分。卷积操作通过在输入数据的局部区域上滑动一个小的窗口（卷积核）来提取特征。这个操作能够保留输入数据的空间结构，使网络能够更好地理解图像的局部特征。卷积操作共享权重，这意味着同一个卷积核在整个输入图像上共同使用，减少了需要学习的参数数量，从而提高了网络的

训练效率。卷积根据作用在不同维度数据进行卷积操作分为一维卷积和二维卷积。

一维卷积通常用于处理序列数据，比如文本、时间序列等。这些数据通常具有单一维度。一维卷积的具体定义如式(2-1)所示。

$$(f*g)(t)=\sum_{a=-\infty}^{\infty} f(a)g(t-a) \tag{2-1}$$

其中，f 是输入序列，g 是卷积核，*号表示卷积操作 t 是卷积操作的输出位置。卷积核是一个包含一维权重的滤波器。在每个位置 t，卷积核与输入序列的对应位置上的元素逐一相乘，随后对乘积累求和，生成输出的一个元素。

二维卷积通常用于处理图像数据，如图 2-1 所示，图像数据是二维的，具有高度和宽度两个维度。每个维度上的值代表像素的亮度或颜色信息。二维卷积的数学表达如式(2-2)所示。

$$(I*K)(x,y)=\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} I(x-m,y-n)K(m,n) \tag{2-2}$$

其中，I 是输入图像，K 是二维卷积核，(x,y) 是卷积操作的输出位置。卷积核是一个包含二维权重的滤波器。在每个位置 (x,y)，卷积核与输入图像的局部区域进行对应相乘，然后将结果相加，得到输出图像的一个像素值。

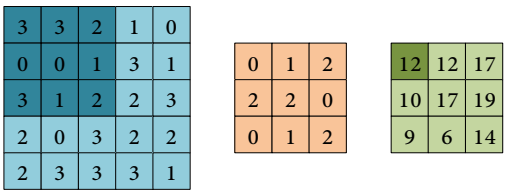


图 2-1 二维卷积过程

池化层通过下采样操作降低特征图的尺寸，从而不仅减少了计算复杂度，还进一步提取和精炼了关键特征。池化层通常位于卷积层之后，通过选择性地保留局部区域中的最大值或平均值来降低特征图的尺寸。最为广泛应用的两种池化方式为最大池化（Max Pooling）和平均池化（Average Pooling），图 2-2 展示了两种池化方式。池化层的作用是进一步简化网络结构、减少参数数量和提高模型的泛化能力。同时还能提高网络的平移不变性，使模型对图像的小位移更加鲁棒。

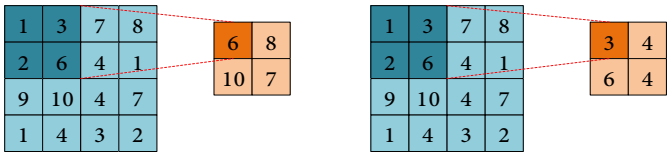


图 2-2 最大池化和平均池化

激活函数层：卷积操作之后，通常会加入激活函数引入非线性。这是因为实际世界的数据往往是非线性的，激活函数能够帮助网络学习到更复杂的模式。常用的激活函数包括 ReLU、Sigmoid 和 Tanh 等，图 2-3 展示了常用的激活函数曲线，这些激活函数可以将负值置为 0，将正值保持不变，从而增加网络的稀疏性。通过合理地选择和应用激活函数，卷积神经网络不仅能够更好地拟合训练数据，还能够提高模型的泛化能力，从而在实际应用中表现出更好的性能。

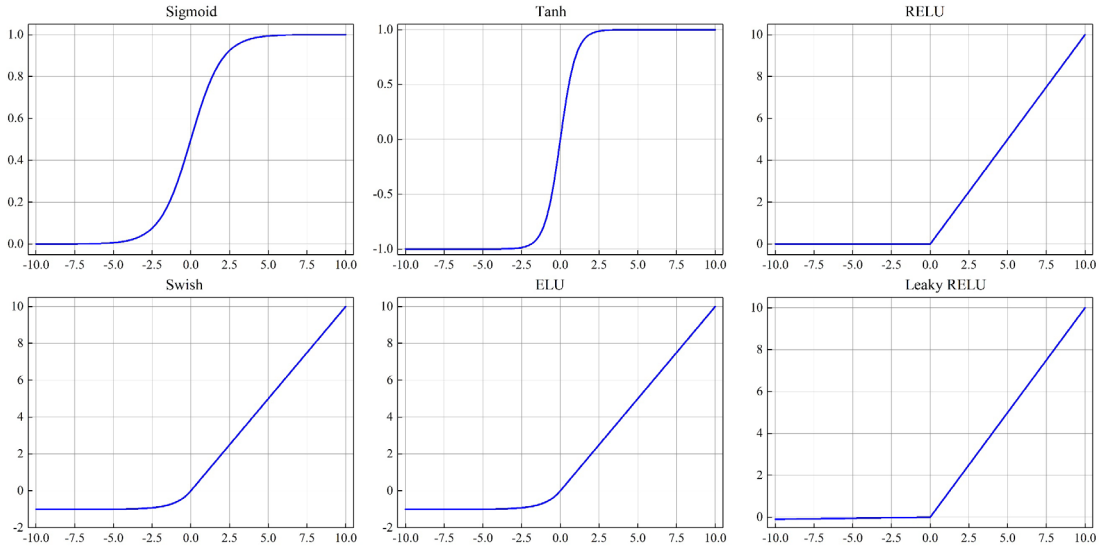


图 2-3 常用的激活函数

全连接层：CNN 的最后一层，负责将前面层的输出聚合起来，进行最终的分类或回归任务。在语义分割任务中，全连接层将提取到的特征表示映射到目标类别空间，实现像素级别的分类。在处理分类任务时，全连接层通常采用 Softmax 函数进行分类任务的输出处理，将每个类别的概率值归一化到 0-1 之间，提供了清晰可靠的分类结果。

2.2.3 深度学习语义分割流程

深度学习在语义分割领域的应用日益增多，这种复杂任务的完成依赖于精心设计的流程。图 2-4 展示了从数据采集到模型部署的完整步骤。

首先，采集数据是整个流程的基础。这一步骤涉及收集与项目目标紧密相关的图片数据集，数据的质量直接决定了模型性能的上限。接着，数据增强通过对原始数据进行变换，如旋转、翻转、缩放等，以此提高模型的泛化能力和鲁棒性。数据预处理则确保数据格式统一，包括图像大小调整、归一化等，让数据适配后

续模型输入。在数据准备阶段，数据标注对于监督学习至关重要，要求为每个像素分配相应的类别标签，数据划分阶段则是将这些已标注数据合理地分为训练集、验证集和测试集三个部分，以支撑模型的训练和评估。

选择合适的模型架构是关键一步，当前流行的如 U-Net、FCN 等，都是语义分割领域的主流选择。模型设计涉及网络层的具体设置，如卷积层、激活函数、池化层的配置。模型初始化为训练过程提供起点，合理的初始化方法可以加速模型收敛。前向传播计算模型的输出，而损失函数则评估模型输出与真实标签之间的差异。反向传播是学习过程的核心，它负责计算损失相对于模型参数的梯度。紧随其后的参数更新步骤，根据梯度调整模型权重，通常采用 SGD 或 Adam 等优化算法。模型训练过程中，超参数调优和验证集监控是不可或缺的。调优涉及学习率、批量大小等参数的调整，而验证集监控则帮助追踪模型性能，防止过拟合。通过是否最后一轮的判断，确定是否终止训练循环。

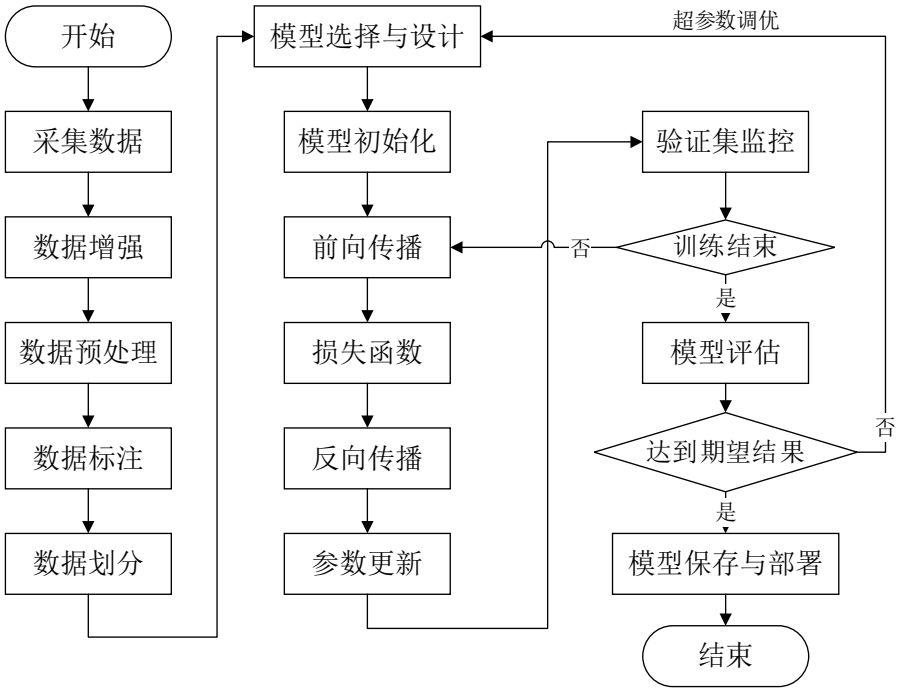


图 2-4 深度学习语义分割流程图

当模型经过若干轮次迭代和优化后，需要在独立的测试集上进行模型评估，目的是验证模型在未见过的数据上泛化能力。若测试结果不满足预期，可能需要返回进行模型架构的修改或超参数的调整。一旦模型达到满意的性能标准，便进行模型保存。在部署前，还需要对模型进行优化和压缩，以减小其占用空间和提高运行效率，确保模型高效地应用到实际场景。

2.2.4 损失函数

神经网络的损失函数（Loss Function）是用来度量模型预测输出与真实标签之间的偏差或不一致性。损失函数在训练神经网络时起到了至关重要的作用，它是优化算法的目标函数，通过最小化损失函数，模型可以更准确地拟合训练数据，提高在未见过数据上的泛化能力。不同类型的任务（如分类、回归、生成等）通常需要选择不同的损失函数。以下是常见分类任务神经网络的损失函数：

交叉熵损失函数（Cross-Entropy Loss）如式(2-3)所示。

$$Loss = -\sum_{i=1}^{C} y_i \log(p_i) \tag{2-3}$$

其中，y_i 是真实标签的概率分布，p_i 是模型预测的概率分布，C 是类别数。适用于多类别分类问题。

二元交叉熵损失函数（Binary Cross-Entropy Loss）如式(2-4)所示。

$$Loss = -\sum_{i=1}^{2} y_i \log(p_i) \tag{2-4}$$

其中，y_i 是真实标签的概率分布，p_i 是模型预测的概率分布。适用于二分类问题。

多标签分类损失函数（Multi-Label Classification Loss）：常用的损失函数是二元交叉熵损失函数的扩展版本如式(2-5)所示，在多标签分类中，对每个类别使用二元交叉熵损失函数，然后将所有类别的损失值进行求平均损失。

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left(y_{ij}\log\left(p_{ij}\right)+\left(1-y_{ij}\right)\log\left(1-p_{ij}\right)\right) \tag{2-5}$$

其中，N 是样本数量。y_{ij} 是第 i 个样本属于第 j 个类别的真实标签。p_{ij} 是模型对第 i 个样本属于第 j 个类别的预测概率。通过合理选择损失函数，神经网络能够在训练过程中不断优化其预测性能，从而提高在各类任务上的准确性和泛化能力。

2.3 语义分割相关技术与评价指标

2.3.1 经典的语义分割算法

语义分割技术的发展历程中，涌现了许多经典的算法，这些算法为理解和处理视觉信息提供了强大的工具。卷积神经网络的核心思想最早由 Yann LeCun 等人在 20 世纪 80 年代末和 90 年代初提出。将神经网络与卷积运算结合起来，并引入了局部感受野和权值共享的概念，用于处理图像和视觉任务。这种结构的卷积神经网络被称为 LeNet，成为了当时手写数字识别任务的先驱。如图 2-5 所示，LeNet 网络输入层接收灰度图像作为输入，图像尺寸为 32×32 像素。第一层是一个卷积层，包含 6 个卷积核，每个卷积核的尺寸为 5×5，步长为 1，使用 Sigmoid 激活函数。第二层是一个平均池化层，采用 2×2 的窗口进行下采样，步长为 2。第三层是另一个卷积层，包含 16 个卷积核，每个卷积核的尺寸为 5×5，步长为 1，同样使用 Sigmoid 型激活函数。第四层是另一个平均池化层，采用 2×2 的窗口进行下采样，步长为 2。第五层是一个全连接层，包含 120 个神经元，每个神经元与前一层的所有神经元相连。第六层是另一个全连接层，包含 84 个神经元。最后一层是输出层，根据任务需要确定输出神经元的数量，对于手写数字识别，输出层通常有 10 个神经元，每个神经元对应一个数字类别，LeNet-5 网络如图 2-5 所示。

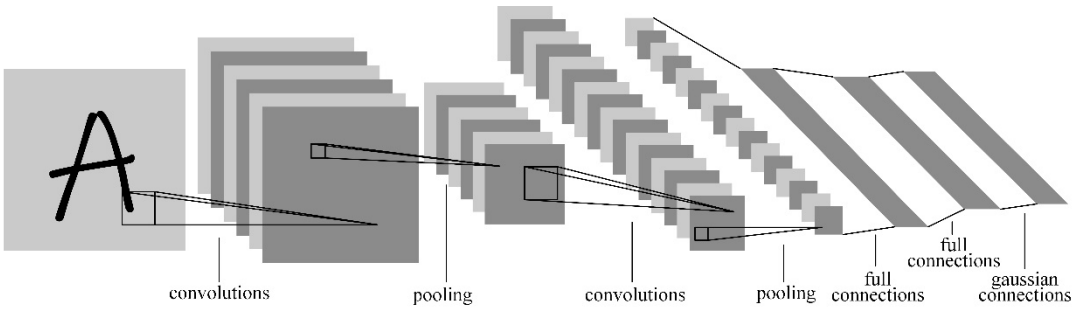


图 2-5 LeNet-5 识别手写字母

近年来，得益于计算能力的显著增强以及大规模数据集的日益丰富，卷积神经网络在计算机视觉和其他领域取得了重大突破。AlexNet 模型在 ImageNet 图像识别挑战赛中取得了巨大的成功，AlexNet 比 LeNet 更加深层，通过小卷积叠加来替换单个的大卷积，将深度学习推上了新的高度，自此以后，卷积神经网络成为了计算机视觉领域的主流模型，并在图像分类、目标检测、语义分割等任务上取得了许多重要的进展。经典的网络还有 2013 年在 ImageNet 挑战赛上取得了优异的成绩的 ZFNet、2014 年的冠军网络 GoogleNet、2015 年的冠军网络 ResNet。

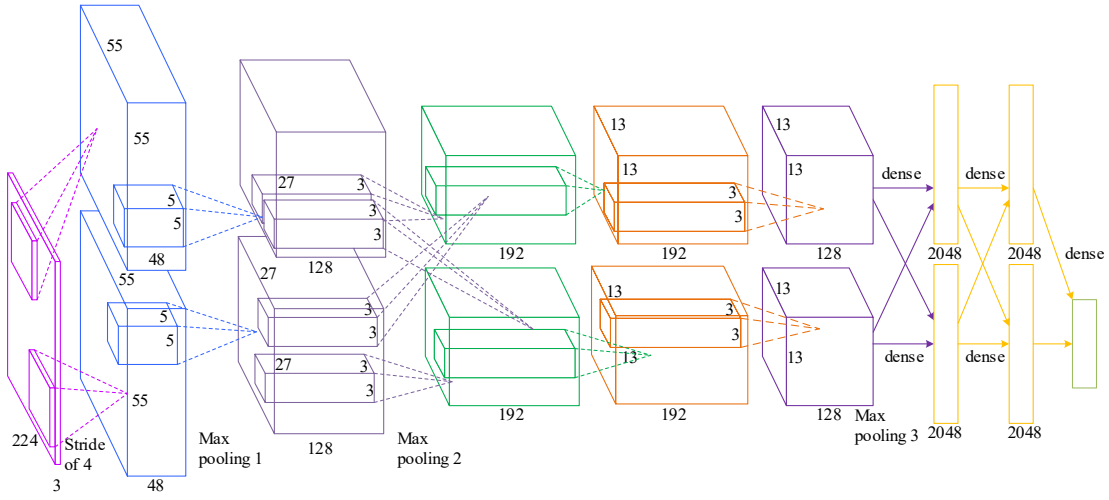


图 2-6 AlexNet 可视化流程图

在语义分割任务中，通过引入残差学习的概念，成功地解决神经网络中常见的梯度消失或梯度爆炸等难题，ResNet 可以作为编码器提取图像特征，将提取的特征传递给解码器进行像素级别的语义分割。具体来说，ResNet 首先将输入图像经过一系列的卷积层和池化层，然后通过多个残差模块进行特征提取。每个残差模块包含一个或多个卷积层，以及一个跳跃式连接，用于将输入直接加到输出上，从而保留原始特征并避免特征逐层消失。如表 2-1 所示，ResNet 有多种不同层数的网络配置，包括 18 层、34 层、50 层、101 层和 152 层，根据网络层数的不同，残差模块的类型也会有所不同。在较浅层的网络如 ResNet18 和 ResNet34 中，主要使用基本的残差模块（Basic Block），而在较深的网络如 ResNet50、ResNet101 和 ResNet152 中，则使用瓶颈残差模块（Bottleneck Block）。

表 2-1 不同深度的 ResNet 结构配置

层名称	输出尺寸	18 层	34 层	50 层	101 层	152 层
conv1_x	112×112	7×7, 64, 步幅 2				
conv2_x	56×56	3×3, 最大池化, 步幅 2				
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
conv4_x	14×14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$

续表 2-1 不同深度的 ResNet 结构配置

层名称	输出尺寸	18 层	34 层	50 层	101 层	152 层
conv5_x	7×7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
1×1		平均池化, 1000-d fc, softmax				
浮点数		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

2.3.2 语义分割主流数据集

高质量的数据集对于语义分割算法的训练和评估至关重要。主流的语义分割数据集包括 PASCAL VOC[61]、MS COCO[62]、Cityscapes[63]、NYU v2[64]、SUN RGB-D[65]等。

（1）RGB 数据集：PASCALVOC 是一个经典的语义分割数据集，提供了详细的分割标注，用于识别和分割图像中的对象。MS COCO 由微软研究院发布，数据集拥有大量的图像，并提供了丰富的类别标注，适用于多种计算机视觉任务，包括语义分割。Cityscapes 是一个专注于城市场景的数据集，提供高质量的像素级分割标注，主要用于自动驾驶和城市场景理解。ADE20K 数据集由 MIT 计算机科学和人工智能实验室提供，包含多种场景和对象的详细分割标注。

（2）RGB-D 数据集：NYU v2 数据集由纽约大学提供，包含室内场景的 RGB 和深度信息，适用于场景理解和深度预测。SUN RGB-D 数据集包含从不同室内场景捕获的 RGB-D 图像，用于对象检测和语义分割。KITTI 是实际交通场景数据集，尽管 KITTI 数据集主要用于自动驾驶场景，但它也包含了 RGB-D 语义分割任务所需的 RGB-D 图像。

这些数据集通常用于训练和验证语义分割模型的性能。RGB 数据集在视觉识别方面非常有用，因为它们模拟了人眼对环境的感知。而 RGB-D 数据集则提供了更多维度的信息，深度通道能帮助模型更好地理解场景的三维结构，这对于精确分割任务尤为重要。

2.3.3 语义分割评价指标

在语义分割中，评估语义分割模型的性能至关重要。以下是几种常用的评价

指标，它们能够全面地反映模型的性能。

像素准确率（Acc）：这是最直观的评价指标，它计算的是分类正确的像素数占总像素数的比例。定义如式(2-6)所示。

$$Acc = \frac{\sum_i n_{ii}}{\sum_i t_i} \tag{2-6}$$

其中，n_{ii}是正确分类到类别的像素数，t_i是类别的总像素数。尽管像素准确率简单易懂，但它偏向于面积较大的类别，可能忽视了小类别的分割效果。

平均准确率（mAcc）：为了弥补 PA 的不足，平均准确率计算的是每个类别的像素准确率的平均值。这使得所有类别都同等重要，无论其大小。

交并比（IoU）：也称为 Jaccard 指数，如式(2-7)所示，这是衡量预测区域与真实区域重叠程度的指标。对于每一个类别，它的定义是预测区域与真实区域的交集大小与它们的并集大小之比。

$$\mathrm{IoU} = \frac{\sum_i n_{ii}}{\sum_i \left(t_i + \sum_j n_{ji} - n_{ii} \right)} \tag{2-7}$$

其中，n_{ji}是被错误分类为类别 j 的类别 i 的像素数。IoU 是一个非常强大的指标，因为它同时考虑了真实的正样本和错误分类的样本。

平均交并比（mIoU）：为了全面评价，在所有类别上计算 IoU 的平均值。mIoU 是语义分割中最常用的评价指标之一，因为它能够平衡各种类别大小，提供更加公正的性能评估。

频权交并比（fwIoU）：此指标对每个类别的 IoU 进行加权，权重是该类别的频率。这考虑了类别出现频率对性能的影响，对常见类别给予更高的权重。

Dice 系数：通常用于医学图像分割中，Dice 系数计算的是两倍的预测区域与真实区域交集大小与它们各自大小之和的比例。这个指标与 IoU 类似，但在某些情况下，可能更适用于数据不平衡的问题。

除了模型准确性外，在实际应用中，也需要考虑模型的效率。模型的效率可以从训练时间、推理时间、每秒处理帧数（FPS）等方面进行评估。

训练时长是影响模型迭代次数和最终性能的关键因素。为了有效地缩短训练时间并提高模型的收敛速度，可以采取多种策略。其中包括调整批量大小以平衡

内存使用与计算效率，利用 GPU 加速进行并行计算，以及应用数据增强技术来增加训练样本的多样性。下面是一个训练代码示例。

```
model = SegmentationModel(num_classes)
criterion = nn.CrossEntropyLoss()
optimizer = optim.Adam(model.parameters(), lr=0.001)
for epoch in range(num_epochs):
    for i, (inputs, labels) in enumerate(train_loader):
        optimizer.zero_grad()
        outputs = model(inputs)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()
```

推理时间的长短会影响模型在实际场景中的应用效果。可以通过调整网络结构、使用轻量化的模型、使用加速器等方式来加速推理过程。下面是一个推理代码示例。

```
model.eval()
with torch.no_grad():
    for i, (inputs, _) in enumerate(test_loader):
        outputs = model(inputs)
        outputs = F.softmax(outputs, dim=1)
        _, predicted = torch.max(outputs.data, 1)
```

每秒处理帧数是衡量模型在推理阶段的速度的重要指标，FPS 越高，表示模型能够在单位时间内处理更多的图像，速度越快。在推理阶段，通过输入图像并记录模型推理所花费的时间，FPS 可以通过推理时间的倒数计算得到。为了得到更准确的结果，可以对多个图像的推理时间进行平均，然后再计算 FPS 平均值。

2.4 本章小结

本章首先介绍深度学习的基础和相关组成原理，其次介绍了经典语义分割算法包括本文的主干网络 ResNet，其次详细阐述了语义分割的相关技术，包括完整的语义分割流程、数据集、损失函数、评价指标进行了详细阐述，这些内容的详尽介绍为后续研究工作提供了全面的理论支撑和实践指导。

第3章 基于 SwiftNet 的室内 RGB-D 场景语义分割网络研究

3.1 引言

针对当前室内场景语义分割的挑战，其中包括精度不高和计算效率低下的问题，本章的主要目标是解决这些问题。为了应对室内环境的复杂性和多样性，本章主要内容为在一个轻量级多尺度道路 RGB 场景语义分割算法 SwiftNet 中引入深度图像。该方法旨在有效地结合多模态信息提升分割精度。通过将深度信息与 RGB 图像相结合，网络能够更准确地理解场景中的物体边界和结构信息。为了提高计算速度，网络中采用了 Non-Bottleneck-1D 替代传统的瓶颈结构，从而降低了网络的复杂度并提高了推理速度。另外，为了进一步提升特征的表征能力，还引入了注意力融合模块，通过对不同尺度特征的加权融合，使网络能够更好地关注对分割结果有贡献的信息，从而提高整体的分割性能。同时，为了扩大网络的感受野并捕获更广泛的语境信息，通过快速聚合上下文模块有效地结合局部和全局信息提高感受野，进一步提高了网络对场景语义的理解能力。最后在 NYU v2 和 SUN RGB-D 数据集上进行了试验，并与主流分割方法进行了对比，同时结合消融实验验证了模块的有效性。

3.2 相关原理技术介绍

3.2.1 SwiftNet 网络介绍

SwiftNet 是一种针对城市道路场景理解和语义分割任务的深度神经网络方法。其网络结构以多尺度交错金字塔形式设计，利用预训练的 ImageNet 架构实现实时语义分割任务。图 3-1 展示了 SwiftNet 的原始网络结构。该网络是基于 RGB 输入的轻量级语义分割网络，采用双流结构处理多尺度 RGB 信息。两个编码器同时处理两种不同分辨率的输入图像，以实现对不同尺度信息的充分利用。此外，通过横向连接，编码器与解码器之间建立了连接，将特征张量拼接在一起，以增强特征的表征能力和信息传递。参数共享是该网络的另一个重要特征，两个编码器共享参数，使用相同的权重处理不同分辨率的图像，从而提高了模型的效

率，减轻了对模型容量的要求。此外，网络还采用了多个上采样模块，将低分辨率特征图上采样到原始输入图像的分辨率，以恢复分割结果的空间信息。SwiftNet 在道路驾驶数据集 Cityscapes 上的表现优于同类型网络，在实际道路场景理解任务中具有有效性和性能优势。

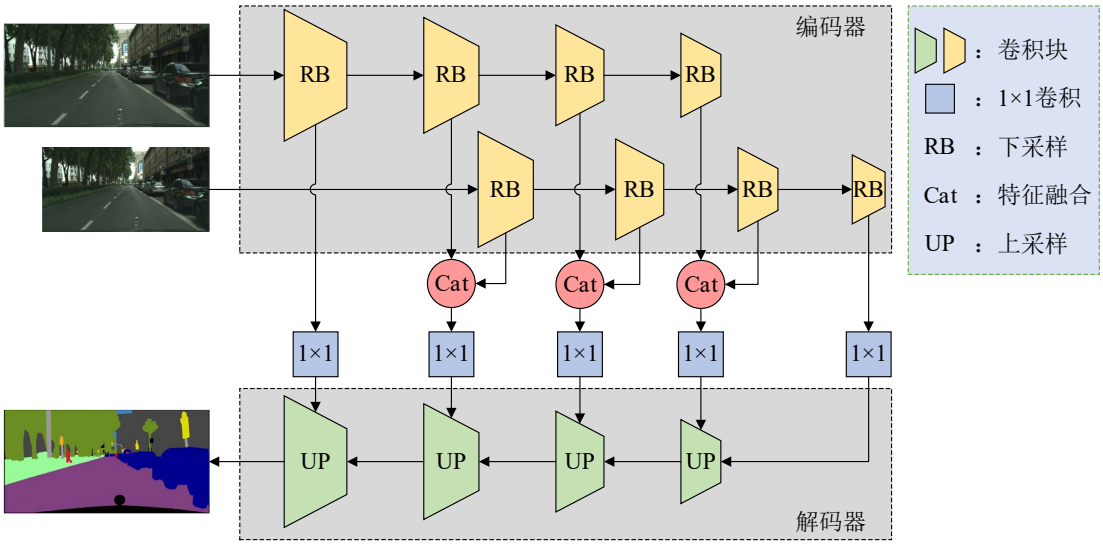


图 3-1 面向道路场景的多尺度语义分割网络 SwiftNet 网络框架

3.2.2 快速聚合上下文模块

由于编码器中采用残差网络的感受野受限，受到了金字塔场景解析网络（PSPNet）的启发，通过在模型中加入了金字塔池化模块（PPM）在不同尺度上聚合特征来附加上下文信息，模块的作用是在不同尺度上聚合特征，以增加网络的感受野并促进局部与全局特征的交互。深度聚合金字塔池化模块（DAPPM）从低分辨率特征图中提取上下文信息。然而，DAPPM 的串行计算特性以及各尺度通道数较多的问题，导致其在处理时间上的开销较大，并不适用于轻量级模型。为克服上述不足，如图 3-2 所示，提出了一种快速聚合金字塔池化模块（FAPPM）。该模块通过在不同尺度上实施池化操作，能够捕捉更广泛范围内的信息。其中，较大的池化核大小（K）和步长（S）能够覆盖更大范围的图像区域，提供全局信息；而较小的 K 和 S 则能够更细致地分析局部特征，从而充分理解图像的细节结构，并进一步扩大网络的感受野。同时，对其进行了改进，使其支持并行操作，从而提高了计算速度。具体而言，在通过金字塔池化提取到不同层级的特征信息后，不再要求原特征信息逐一与四个空洞卷积后的特征进行融合，而是采用并行

方式与四个分支同时进行融合。这种并行计算方式能够更好地利用硬件并行计算能力，特别是在处理大规模图像和计算需求较高的任务时，能够显著加快推理速度。

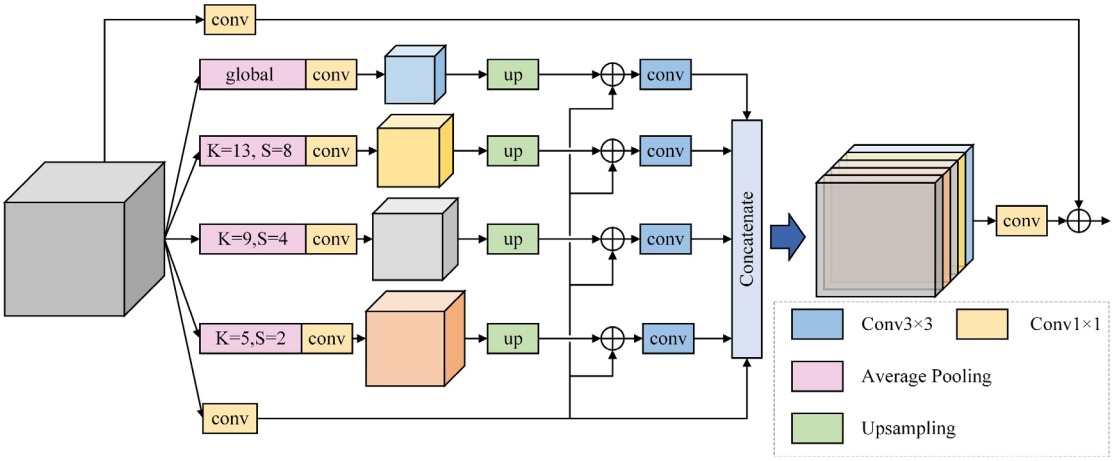
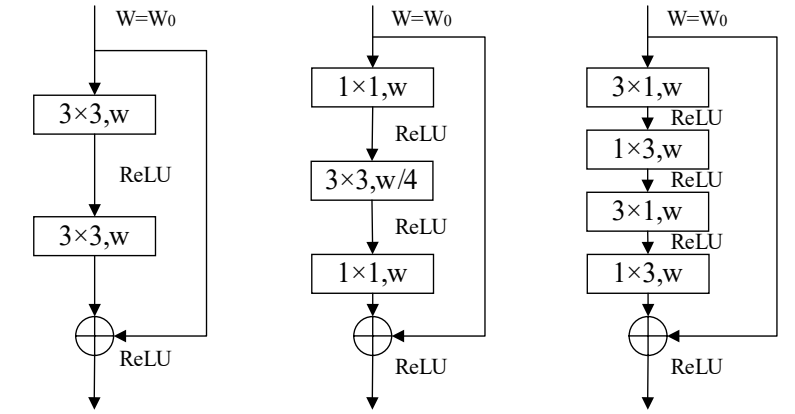


图 3-2 快速聚合金字塔池模块

3.3.3 一维非瓶颈模块

一维非瓶颈模块（Non-Bottleneck-1D）如图 3-3 中（c）所示，这种结构设计灵感来自残差网络的 Bottleneck 的结构，如图 3-3 中（b）所示，利用跳跃连接来绕过一个或多个层。一维非瓶颈模块特别针对一维特征映射，通过分解卷积操作来减少参数数量和计算复杂度。这种块结构通常包含两个平行的一维卷积路径，一个用于提取特征，另一个用于形成跳跃连接，确保网络可以学习到残差函数。这样的设计有助于网络在不增加过多计算负担的情况下学习更深层次的特征表示，从而在视觉处理任务中实现更好的性能。



(a) Non-Bottleneck　　(b) Bottleneck　　(c) Non-Bottleneck-1D

图 3-3 Non-Bottleneck，Bottleneck，Non-Bottleneck-1D 模块

在参数量上，Non-Bottleneck 由于不缩减特征维度，在使用3×3标准卷积核时，参数量相对较高为$3\times3\times w^2$。相比之下，Bottleneck 结构通过减小通道数量至$w/2$来降低参数量，参数数量为$3\times3\times(w/2)^2$，尽管有所减少，但在恢复原始维度时，其参数量仍然高于 Non-Bottleneck-1D 结构。Non-Bottleneck-1D 参数量最少为$1\times3\times w^2$。可以看出，Non-Bottleneck 的参数量最大，因为它不能减小特征的维度。Bottleneck 通过减小和恢复特征维度来减少参数量，但仍然比一维非瓶颈结构多，因为在瓶颈层后恢复了原始特征维度，但中间的参数数量减少了一半。Non-Bottleneck-1D 结构的参数量最少，因为处理的是一维数据，并且没有减小特征的维度，能够更加高效地利用输入特征。一维卷积在计算上更加高效，因为它只在一个方向上进行滑动，减少了计算量，使得 Non-Bottleneck-1D 结构在模型推理过程中表现更加迅速。

3.3.4 注意力融合模块

图 3-4 展示了注意力融合模块，注意力融合模块分别作用于每个残差编码层之后，通过引入一个压缩（Squeeze）操作和一个激励（Excitation）操作来对通道之间的关系进行特征融合建模。在 Squeeze 阶段通过全局平均池化操作将 RGB 和深度特征图分别压缩成一个特征向量，以获取每个通道的全局信息。然后，在 Excitation 阶段通过两个全连接层，将全局信息进行处理，产生一个用于缩放每个通道权重的激励值，得到的激励值应用到融合后的特征上调整每个通道的权重，网络能更加灵活地关注 RGB 和深度数据中的重要性不同的特征，增强网络对 RGB 和深度信息的综合利用。

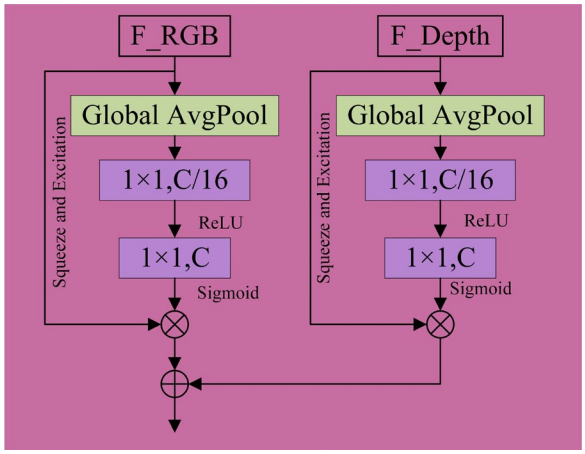


图 3-4 RGB-D 注意力融合模块

3.3 网络结构

SwiftNet 分割算法只在道路场景下进行了验证，其在更复杂的室内 RGB-D 场景下的表现还不明确。此外，室内场景与道路场景存在很大差异，如物体类别、光照条件等。直接应用 SwiftNet 到室内场景可能难以适应，本节将研究基于 SwiftNet 适合室内 RGB-D 场景的分割网络。网络采用双分支编码器-解码器结构，在编码器部分，采用了在 ImageNet 数据集上预训练的 ResNet34 网络作为主干特征提取网络。ResNet34 原本是用于 RGB 图像特征提取，使用一维非瓶颈模块替换每层的残差块，每个 3×3 卷积被一个 3×1 和一个 1×3 卷积所取代，其中包含一个 ReLU。同时考虑到需要同时处理相同尺度的 RGB 和深度两个模态，为充分利用两个模态之间的互补信息，在第一个特征提取块（RB）就开始对两个模态的特征进行了注意力融合（Cat），深度特征被融合到 RGB 编码器中。两种模式的特征首先用压缩和激励（SE）模块重新加权，然后自主地对元素进行求和，这种早期特征融合设计可以充分利用 RGB 和深度两个模态在低级特征上的互补关系。

如图 3-5 所示，解码器模块扩展了 SwiftNet 的解码模块，由一个固定数量为 128 通道的 3×3 卷积和后续的双线性上采样组成。实验表明，对于室内 RGB-D 分割需要一个更复杂的解码器。因此，在第一个解码器模块（UP）中使用 512 个通道，并随着分辨率的增加减少每个 3×3 卷积中的通道数量。此外，加入了三个额外的 Non-Bottleneck-1D（NBt1D）块，以进一步提高细分性能。尽管尺寸被放大了，结果的特征图仍然缺乏在编码器下采样过程中丢失的细粒度细节。因此，设计了从相同分辨率的编码器到解码器阶段的跳跃式连接采用融合的 RGB-D 编码器特征映射，用 1×1 卷积将编码器层级输出投影到解码器中使用的相同数量的通道，并将其添加到解码器特征映射中，通过跳跃连接到解码器减少信息丢失。最后使用双线性插值对图像尺寸进行放大输出。

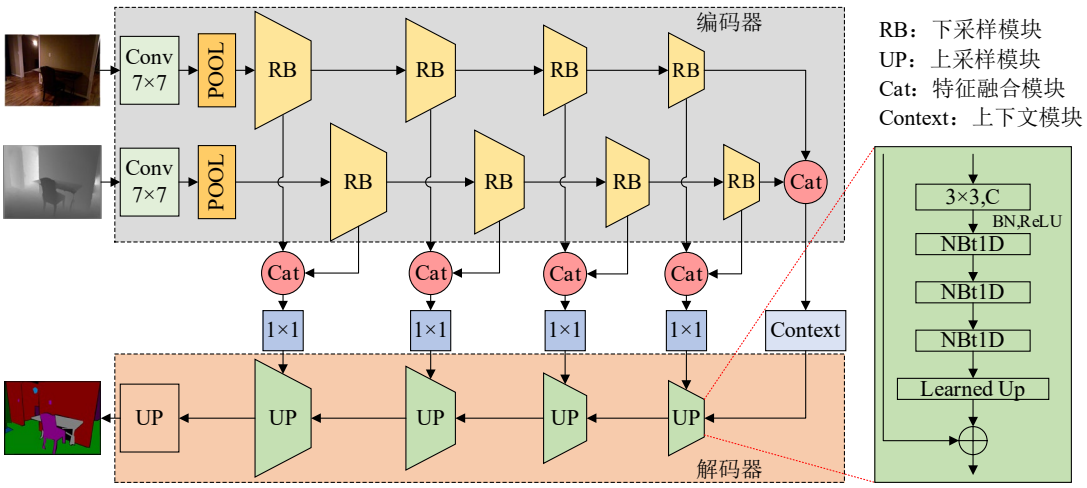


图 3-5 整体网络结构

3.4 实验设计与结果分析

3.4.1 实验数据集及评价指标

实验在两个主流 RGB-D 室内场景数据集上对提出的方法进行了评估，这两个数据集分别是 SUN RGB-D 和 NYU v2，同时，也在 Cityscapes 数据集上展示了该方法的适用性。为了深入理解各网络组件的作用，还进行了一系列消融实验。以下是对实验数据集和评价标准的详细说明。

NYU v2 数据集由 1449 张室内场景的 RGB-D 图像组成，其中包括 795 张训练图像和 654 张测试图像。该数据集中图像被分为 40 个语义类别。实验采用了这一标准类别设置。消融实验主要在该数据集上进行，以加速训练过程。尽管数据集规模较小，但先前研究显示，在此子集上训练得到的模型能够进行有效的模型选择。网络的输入分辨率设定为 640×480。为了处理类别不平衡问题，本研究在数据预处理阶段应用了类别平衡策略。由于网络结构导致的下采样，上下文模块的输入分辨率减小至 20×15，为了适应该分辨率，上下文模块进行全局平均池化的同时进行卷积调整通道。

SUN RGB-D 数据集包括 10335 张室内场景的 RGB-D 图像，涉及 37 个类别，包括 5285 张训练图像和 5050 张测试图像。该数据集中同样采用了 640×480 的输入分辨率，并实施了类似的类别平衡策略。上下文模块的分支设计与 NYU v2 数据集相同。

Cityscapes 数据集含有 5000 张精细标注的城市街景 RGB 图像，分辨率为 2048×1024，标注了 19 个类别。其中包括 2975 张训练图像、500 张验证图像和 1525 张测试图像。此外，还提供了 20000 张粗略标注的图像，但未将其用于训练。本研究从视差图中计算得到深度图像，并将网络输入分辨率调整为 1024×512。这导致上下文模块的输入分辨率为 32×16。在上下文模块中，采用了四个分支：一个用于全局平均池化，其他三个分支的池化窗口分别为 16×8、8×4 和 4×2。

3.4.2 网络参数设置及试验方法

模型的主干分别选择了 ResNet18、ResNet34 和 ResNet50 等不同深度的经典残差网络作为模型的主干。在进行实验前，先对输入数据进行了预处理，包括图像归一化和深度信息处理，为了避免数值不稳定性，将 RGB 和深度值映射到[0,1]的范围内。对于 RGB 图像还在 HSV 空间中应用了轻微的颜色抖动，以增加数据的多样性。针对深度数据中可能存在的缺失值，采用了均值填充法进行缺失值填充。在训练过程中，为了避免过拟合，使用了随机缩放、裁剪和翻转等数据增强技术。

针对本章网络框架，训练时使用二元交叉熵损失函数，如式(3-1)所示。

$$L(y,\hat{y})=-\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log(\hat{y}_i)+(1-y_i)\log(\hat{y}_i)\right] \tag{3-1}$$

其中，N 是样本数量，y_i 是真实标签（通常为 0 或 1），$\hat{y}_i$ 是模型的预测结果。

本章模型的构建、训练、测试都是在 PyTorch 深度学习平台上进行，具体运行环境见表 3-1。具体实验参数设置见表 3-2。

表 3-1 实验环境

配置	版本型号
处理器	Intel(R) Core^{TM}i5-6500 CPU 3.20GHZ
显卡	NVIDIA GeForce GTX 1080 Ti
显存	11G
内存	16GB
操作系统	Windows10 Professional
CUDA/CUDNN	CUDA10.1/ CUDNN7.6.3
编程语言	Python 3.7
深度学习平台	PyTorch 1.3

表 3-2 参数设置

参数	参数选择		
GPU 数量	1		
迭代次数	500		
批大小	8		
学习率	策略	初始值	幂
	Poly	0.16	0.9
优化器	策略	动量	权重衰减
	SGD	0.9	0.0001

3.4.3 实验结果分析

与基线模型对比：表 3-3 为在 Cityscapes 数据集上与基线模型 SwiftNet 对比结果，在分辨率为 1024×512 下进行评估 mIoU 和 FPS，首先关注 1024×512 的小分辨率，因为此分辨率通常用于高效语义分割。此外，由于大多数方法仅依赖 RGB 作为输入，实验也比较了模型的 RGB 单模态输入的结果。SwiftNet 网络的 mIoU 更高处理速度更快。尽管如此，在 1024×512 的输入分辨率下，本章使用 ResNet34 主干的模型结果仍然超过 30FPS，比所有其他有效方法至少高出 2.2% mIoU。结合深度信息进一步提高细分性能。但是，性能增益没有室内数据集 NYU v2 那么高。所以认为这可以归结为城市景观的视差图像不如 NYU v2 和 SUN RGB-D 的室内深度图像精确。为了完整性还在 2048×1024 的全分辨率上评估了本章的网络。与其他方法相比，本方法在 mIoU 和推理时间上处于移动（SwiftNet、BiSeNet）和非移动方法之间。然而，与 SwiftNet（RGB, 2048×1024）相比，使用 ResNet34 主干的模型在处理 RGB-D 输入时具有 1024×512 较小的输入分辨率，实现了相似的分割性能和更快的推理。

表 3-3 与基线模型对比

图像	模型	1024×512		2048×1024	
		mIoU(%)	FPS	mIoU(%)	FPS
RGB	SwiftNet	70.20	64.5	75.40	20.8
	ResNet18	71.48	37.2	77.95	9.8
	ResNet34	72.70	32.2	78.47	8.3
	ResNet50	73.88	24.9	79.23	6.5
RGB-D	ResNet18	74.65	28.9	79.25	7.6
	ResNet34	75.22	23.4	80.09	6.2
	ResNet50	75.66	16.9	79.97	4.0

与当前领先的技术（SOTA）进行对比：在评估 RGB-D 分割方法时，针对两个室内数据集的结果进行了统计。特别是对于较大的 SUN RGB-D 数据集，可观察到与其他数据集类似的趋势。与目前先进算法相比，使用较小网络结构能够实现与更深网络结构相媲美的分割精度。考虑到机器人嵌入式硬件的推理时间约束，利用 NVIDIA Jetson AGX Xavier 平台和 NVIDIA TensorRT 对所有可用方法的推理时间进行了评测。结果证实，使用 TensorRT 优化后，网络性能达到相对于原生 PyTorch 实现高达 5 倍的推理速度提升。如表 3-4 所示，优化后的方法能够在保证性能的同时实现更快的推理速度。使用 ResNet34 主干在 NYU v2 和 SUN RGB-D 上 mIoU 高达 50.5%和 48.3%，比同主干模型的 RedNet 高出 5.49%和 1.07%，且 FPS 可以达到 33.5，显示出研究的有效性和计算效率，在深度任务中表现良好，同时保持了较高的计算效率。使用 ResNet50 主干在 NYU v2 上取得了最高的 mIoU 达 50.9%，同时在 SUN RGB-D 上也能达到 48.5%的 mIoU，在性能和计算效率上取得了平衡，适用于多样的任务场景。在不同数据集和任务下，研究的模型都表现出色，兼顾了性能和计算效率。这种全面性和平衡性是研究的显著优势，为实际应用提供了可靠的解决方案。

表 3-4 与 SOTA 模型对比

模型	BackBone	Depth	mIoU(%)		Param(M)	FPS
			NYU v2	SUN RGB-D		
FuseNet[55]	VGG16	Depth	36.8	37.3	-	-
FCN[66]	VGG16	HHA	34.9	-	272.2	8.0
3DGNN[67]	VGG16	HHA	42.1	-		5.0
RedNet[68]	ResNet34	Depth	45.2	46.8	47.2	26.0
SSMA[69]	ResNet50	Depth	44.9	44.4	-	12.4
MMAF-Net[70]	ResNet50	Depth	46.9	45.5	-	N/A
RDFNet[29]	ResNet50	HHA	47.7	46.1	-	7.2
ACNet[42]	ResNet50	Depth	48.3	48.1	116.6	18.0
ESANet[71]	ResNet50	Depth	50.5	48.3	-	22.6
SegNet[11]	ResNet101	Depth	49.0	-	58.3	28.0
Ours-ResNet34	ResNet34	Depth	47.7	47.3	48.8	33.5
Ours-ResNet50	ResNet50	Depth	50.9	48.5	56.4	24.2

在特定应用中，选择了配备 ResNet34 主干和非瓶颈（NBt1D）模块的实现，因为它在推理速度与分割性能之间提供了最佳权衡。最后，对比在 NYU v2 和 SUN RGB-D 数据集的结果，能够得出对于数据量有限的目标数据集，相较于使

用更深层次的主干结构，优先采用额外的合成数据进行预训练更加高效。

消融实验：本研究通过大量的消融实验，主要在 NYU v2 和 SUN RGB-D 数据集上验证了所提出模块的有效性。特别关注 Non-bottleneck-1D 和 Bottleneck 两个关键模块的性能，并深入探讨了它们在不同组合下对语义分割任务的影响。针对 Bottleneck 模块的有效性验证，实验结果见表 3-5，在不含 Non-bottleneck-1D 的情况下，包含 Bottleneck 的配置在 NYU v2 数据集上表现出色，达到了 mIoU 46.4%。此外，当同时去除 Bottleneck 和 Non-bottleneck-1D 时，在 SUN RGB-D 数据集上观察到显著的性能提升，mIoU 为 46.4%。这一结果清晰地证实了 Bottleneck 在提高语义分割性能方面的关键作用。对于 Non-bottleneck-1D 模块，实验数据表明，在不使用 FAPPM 的情况下，引入 Non-bottleneck-1D 的配置在两个数据集上均实现了显著的 mIoU 提升（47.7% NYU v2, 47.3% SUN RGB-D），同时保持了相对较高的 FPS（33.5）且相较于 Bottleneck 提升 7.7%。另一方面，当使用 FAPPM 而不包含 Non-bottleneck-1D 时，在两个数据集上也表现出良好的性能，mIoU 分别为 47.9%和 47.4%，FPS 为 31.1。这一发现进一步证实了 Non-bottleneck-1D 在提高性能和维持计算效率平衡方面的重要作用。此外，还深入探讨了 FAPPM 在不同配置下对性能的影响。实验结果显示，在使用 PPM 的配置在 NYU v2 上 mIoU 为 46.3%，FPS 为 33.1。而在使用 DAPPM 的情况下，配置在 SUN RGB-D 上取得了较高的 mIoU（47.9%）。使用 FAPPM 的配置在 NYU v2 数据集上 mIoU 和 FPS 分别比 PPM 和 DAPPM 高出 4.0%、0.8%、1.2%和 2.4%，这一发现揭示了 FAPPM 在不同配置下对性能表现的影响，特别是在计算效率方面的重要性。

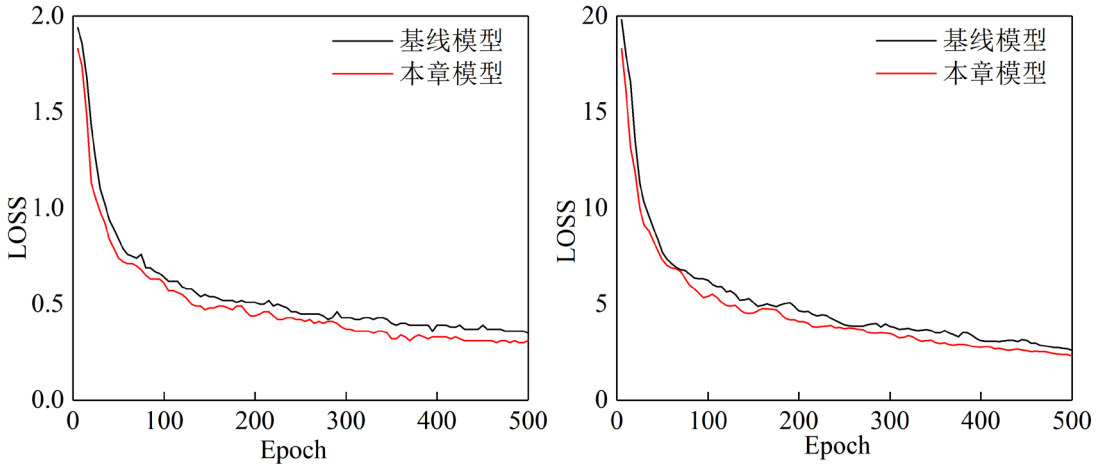
表 3-5 不同模块对结果的影响

上下文			主干		mIoU(%)		FPS
PPM	DAPPM	FAPPM	Bottleneck	Non-bottleneck-1D	NYU v2	SUN RGB-D	
√	×	×	√	×	46.4	46.7	31.4
×	√	×	√	×	47.5	47.9	30.8
×	×	√	√	×	47.9	47.4	31.1
√	×	×	×	√	46.3	46.4	33.1
×	√	×	×	√	47.3	47.9	32.7
×	×	√	×	√	**47.7**	**47.3**	**33.5**

通过对 FAPPM 和 Non-bottleneck-1D 组合的合理性分析，研究展现了在语义

分割任务中模型设计的灵活性和创新性。采用 FAPPM 的配置在性能上具备一定优势，而结合 Non-bottleneck-1D，不仅在性能上取得了提升，还在计算效率上实现了平衡。模型在不同数据集上表现出色，突显了通用性，为实际应用提供了可行的解决方案。

通过在 SUN RGB-D 数据集上分别训练了基线模型 SwiftNet 和本章语义分割模型，图 3-6 分别展示了模型训练和验证过程中的损失函数曲线，本章引入深度的模型在训练和验证集上的损失曲线都明显低于基线模型，本章模型收敛速度更快，说明了本章改进模型相对于基线模型具有更好的性能，主要原因是深度信息的加持。



(a) 模型训练损失函数曲线　　(b) 模型验证损失函数曲线

图 3-6 模型训练和验证过程中的损失函数曲线

可视化：图 3-7 展示了在 NYU v2 数据集上的分割效果，图中列出的图像顺序依次为：原始的 RGB 图像、HHA 可视化图像、Depth 深度图像、真实的分割标签 GT 以及本章模型的预测输出 Pred。并在下方提供了各类别的颜色示例。从图中可以清晰地观察到，模型成功地区分了椅子、桌子、家具等目标，这体现了本方法在整合深度信息方面的优秀能力。即便是在仅凭 RGB 图像难以准确识别的目标上，本章模型也展现出了其强大的识别与分割能力。特别是在图 3-7 中红框标注的案例中，尽管 RGB 图像中沙发的边缘并不明显，但模型依然能够精确地识别出沙发的轮廓。这主要归功于深度信息为模型提供了丰富的几何数据和模型对两种模态数据的整合能力。

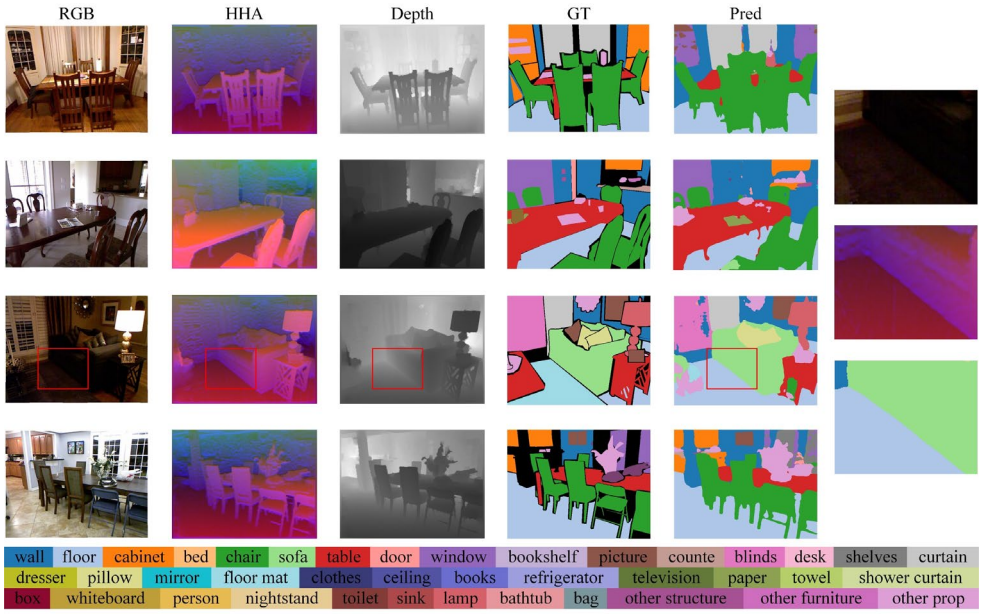


图 3-7 结果可视化

在图 3-8 中，呈现了室内图像语义分割的特征图可视化。通过将本章中提出的方法与领先的 RGB-D 语义分割技术（ESANet[71]）进行了对比，可以更直观的展示网络性能。主要提取了输出特征中椅子的通道，通过特征图可以观察到，本章方法特征图中的椅子的轮廓信息相对清晰，对椅子的形状、边缘的细节特征信息捕捉的更好。

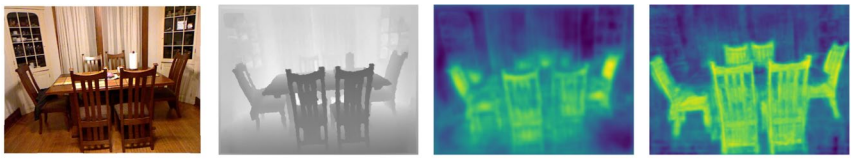


图 3-8 椅子通道的特征图

3.5 本章小结

本章主要针对 RGB 图像分割存在的问题，提出了一种基于 SwiftNet 的室内 RGB-D 场景语义分割网络。首先，介绍了 SwiftNet 网络的整体架构及其在整合深度数据后的模型改进，结合快速聚合上下文模块、一维非瓶颈模块以及注意力融合模块等核心组件共同协作，使网络能够更高效地处理 RGB-D 图像，并提取出有用的特征信息。在实验部分，采用了主流的实验数据集和评价指标，对基于 SwiftNet 的语义分割网络进行了全面的评估。实验结果表明，该网络在 RGB-D 图像语义分割任务上表现出色，验证了其可行性和优越性。此外，还通过消融实

验进一步验证了所提模块的有效性，通过展示与基线模型和先进模型对比分割结果的可视化，呈现了网络的分割效果。

第4章 基于形状感知卷积 RGB-D 图像高效语义分割

4.1 引言

基于现有的 RGB-D 图像语义分割方法大多采用同质卷积算子来处理 RGB 和深度特征，忽略了它们之间的内在差异。本章提出了一种基于形状感知卷积的方法，旨在进一步改进第 3 章中提出的 SwiftNet 网络框架，用于 RGB-D 图像的语义分割。由于 RGB 和深度信息在本质上存在差异，RGB 值主要反映投影图像空间中的光度外观属性，而深度特征则包含有关局部几何体的形状以及周围环境的信息，因此深度信息不能简单地与 RGB 数据等同对待。为了更有效地处理深度特征，本章方法在深度图像编码器分支中引入了形状感知卷积，以实现对深度数据形状信息的提取。随后，采用可学习的上采样模块进行轻量化上采样，以减少网络计算的开销。最后，阐述了网络的整体结构和特点，并在 NYU v2 和 SUN RGB-D 数据集上进行了消融和对比实验，以验证本章方法的可行性和先进。

4.2 相关原理介绍

4.2.1 形状感知卷积

为了更好地进行语义分割，编码器需要充分考虑 RGB 和深度信息的特点。对于相同物体，希望其对应的特征能够一致，因为它们具有相同的形状。形状是物体最基本的属性之一，与语义标签密切相关。为了实现这一目标，需要在学习过程中注重形状的不变性，为了改进 RGB-D 语义分割的效果，使用形状感知卷积处理深度数据，目的是自适应地平衡深度数据层中形状和基础信息的重要性，使网络在处理时能够更加关注形状信息。首先，将每个局部区域分解为两个独立的部分：基础分量和形状分量。基础分量描述了面片在更大背景下的位置，而形状分量则反映了面片中的相对变化，从而提供了关于底层几何体形状的信息。然后，使用两个可学习的权重来处理这两个分量，分别是基础核和形状核。基础乘积和形状乘积这两种操作分别处理基础分量和形状分量，生成相应的特征图。最后，将这两者的输出相加，形成形状感知补丁，该补丁再与正常的卷积核进行卷

积操作。与传统的卷积操作相比，形状感知卷积层能够利用形状核自适应地学习形状特征，而基础核则用于平衡形状和基础对最终预测的贡献。在推理阶段，由于基础核和形状核都变成了常数，可以将其融合到下面的卷积核中，从而形成一个与普通卷积层相似的网络结构。这种设计的优点在于，它可以很容易地插入到大多数 CNN 中，作为语义分割中的传统卷积操作的一种替代。这种简单的替换将原本为 RGB 数据设计的 CNN 转变为更适合处理 RGB-D 数据的 CNN，而不会增加推理阶段的计算和内存负担。

图 4-1 为形状感知卷积的方法流程，该方法的核心是将深度特征分解为位置分量（L）与形状分量（S），通过引入可学习的权重（W_L 和 W_S）进行调整，随后重新组合加权生成最终的特征表示（F）。

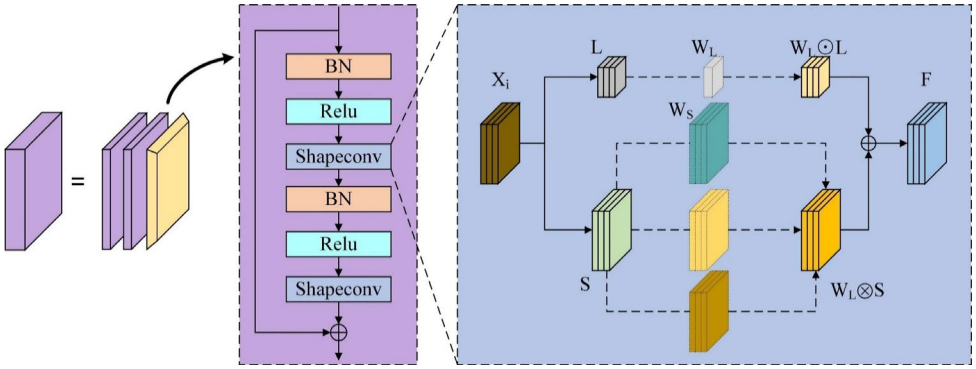


图 4-1 形状感知卷积流程

在标准卷积操作中，输入 $X \in R^{H_{in} \times W_{in} \times C_{in}}$，输出特征图 $F \in R^{H_{out} \times W_{out} \times C_{out}}$ 如式(4-1)所示。

$$F = Conv(K, X) = \sum_{i=1}^{K_h \times K_w \times C_{in}} K_{i,C_{out}} \times X_i \tag{4-1}$$

其中，$K \in R^{K_h \times K_w \times C_{in} \times C_{\text{out}}}$ 是卷积核，K_h、K_w 分别是卷积核的高度和宽度，C_{in}、C_{out} 分别是输入输出特征图的通道数，X_i 是输入特征图 X 中的第 i 个局部区域，$K_{i,C_{out}}$ 是卷积核 K 中的第 i 个权重。为了解决视角和距离因素带来的问题，将输入的局部区域 X_i 分解成一个描述位置的基础分量 $L \in R^{1 \times 1 \times C_{in}}$ 和一个描述形状的分量 $S \in R^{K_h \times K_w \times C_{in}}$。$L$ 和 S 的定义如式(4-2)所示。

$$\begin{cases} L = mean(X_i) \\ S = X_i - mean(X_i) \end{cases} \tag{4-2}$$

其中，$mean(\cdot)$ 表示求所有元素的平均值操作，L 对局部区域的所有像素值进行

求平均操作，实际上就是计算了该局部区域的中心位置，代表了该区域的整体位置和空间分布。S 是局部区域每个像素值与该区域的平均值之间的差异，反映了局部区域相对于其平均位置的偏移情况，即局部区域的形状信息。使用卷积核 K 和两个可学习的权重 W_L 和 W_S，同时定义两个操作符⊙和⊗用于加权 L 和 S，如式(4-3)所示。

$$\begin{aligned} L = W_L \odot L \Rightarrow L_{1,1,C_{in}} &= W_L \times L_{1,1,C_{in}} \\ S = W_S \otimes S \Rightarrow S_{K_h,K_w,C_{in}} &= \sum_i^{K_h \times K_w} W_{S_{i,K_h,K_w,C_{in}}} \times S_{i,C_{in}} \end{aligned} \tag{4-3}$$

将 L 和 S 的加权结果结合到一起，形成一个新的局部区域 X_i'，如式(4-4)所示。

$$X_i' = L + S \tag{4-4}$$

普通卷积核 K 对新的输入 X_i' 进行卷积，以获得最终的输出 F，同时引入两个可学习的权重 W_L 和 W_S，用于加权 L 和 S。输出 F 如式(4-5)所示。

$$\begin{aligned} F &= SConv(K, W_L, W_S, X_i) \\ &= Conv(K, W_L \odot L + W_S \otimes S) \\ &= Conv(K, L + S) \\ &= Conv(K, X_i') \end{aligned} \tag{4-5}$$

这样，F 就包含了局部区域 X 的基础和形状信息，并且是通过学习的权重 W_L 和 W_S 进行加权的，在推理过程中，附加的权重 W_L 和 W_S 会保持不变，不会更新这些权重，而是直接使用预训练的权重。这种方式能够更好地利用输入数据的位置和形状信息，因为与普通卷积中的卷积核共享相同的张量大小，所以在推理过程中不会增加额外的计算成本。

4.2.2 可学习上采样

在语义分割中，上采样通常用于增加特征图的尺寸，以便进行后续处理。常见的上采样方法有双线性插值和最近邻插值。然而，传统的转置卷积计算成本高且可能引入不希望的伪像，因此引入了一种轻量级学习上采样方法，获得了更好的分割结果。最近邻上采样：如果有一个输入特征图 X，其尺寸为 $H \times W$，需要将其上采样 r 倍，即将其尺寸变为 $rH \times rW$。最邻近上采样操作如式(4-6)所示。

$$Y(i,\ j)=X\left(\lfloor i/r\rfloor,\lfloor j/r\rfloor\right) \tag{4-6}$$

其中，Y 是上采样后的特征图，i 和 j 分别表示上采样后特征图中的行和列索引，$\lfloor\cdot\rfloor$ 表示向下取整操作，r 是上采样倍数。

对上采样后的特征图 Y 应用3×3深度卷积，产生特征图 Z 。假设卷积核权重为 W_{conv} ，则卷积后的特征如式(4-7)所示。

$$Z(i,\ j)=\sum_{m=-1}^{1}\sum_{n=-1}^{1}W_{conv}(m,\ n)\cdot Y(i-m,\ j-n) \tag{4-7}$$

其中，Z 是卷积后的特征图，i 和 j 分别表示卷积后特征图中的行和列索引，W_{conv} 是卷积核权重。

传统的转置卷积方法通常需要大量的可训练参数来学习上采样过程中的权重，这会增加模型的参数量。相比之下，轻量级学习上采样方法采用了最近邻上采样和3×3深度卷积的组合，这两种操作都具有较少的参数量，使得模型整体的参数量较少。

在初始化卷积核时，使其模仿双线性插值的效果。具体来说，初始化卷积核权重为双线性插值的权重值。在训练过程中，通过反向传播算法，调整卷积核的权重 W_{conv} 的值，能够适应训练中的权重，因此，通过学习上采样方法，模型能够适应训练数据中的特征分布，并学习到如何以更有用的方式组合相邻的特征。这种方法能够提高分割模型的性能，使其在处理复杂场景时表现更好。

4.3 网络结构

如图 4-2 所示，本章节提出的网络架构是基于 SwiftNet 的双分支编码器-解码器结构。其中，图 4-2 下面淡蓝色部分为一个浅层的双分支编码器与上下文模块紧密结合，并通过跳跃链接将浅层融合特征送入解码器模块，有助于还原图像的局部细节。使用在 ImageNet 预训练的 ResNet34 作为主干网络对 RGB 和深度图像特征进行提取。网络策略是在初始阶段就融合尺寸一致的多模态数据，减少有价值信息的丢失，同时还能充分挖掘输入图像在各个阶段的特征。在深度提取分支中，使用整合形状感知卷积后的残差块进行深度特征提取。

图 4-2 上方淡蓝色部分表示解码器分支，在首个解码器模块，开始使用 512

个通道，随后随着分辨率提升逐步减少每层的通道数。如图 4-2 淡红色部分所示，每个解码模块都嵌入了三个参数量较小的 NBt1D。接着通过一个 3×3 卷积调整通道数以适应每个像素的预测类别数。最后，特征图通过两次上采样操作扩大为原始尺寸的四倍，如图 4-2 右侧浅蓝色部分所示，为提高结果质量，首先应用最近邻上采样放大图像，接着通过 3×3 深度卷积整合邻近特征。初始阶段，上采样模仿双线性插值。然而，训练过程中网络权重的适应性使其能够学习如何更有效地整合邻近特征，从而进一步增强分割效果。

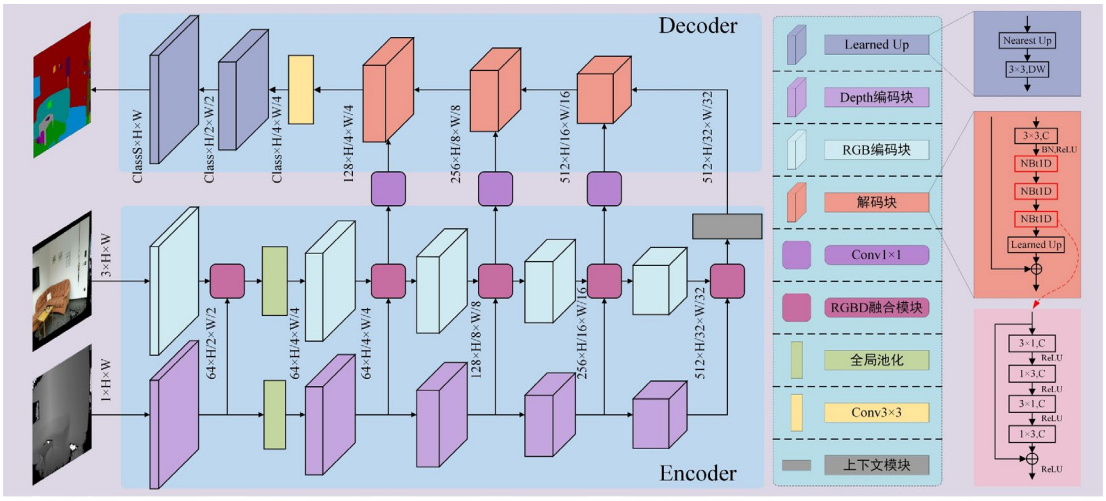


图 4-2 网络整体结构概述

4.4 实验设计及结果分析

4.4.1 实验设置

在 RGB-D 语义分割实验中，数据集选用了广泛使用的 NYU v2，该数据集包含 RGB 图像和深度信息以确保实验结果的可比性和通用性。实验平台采用了 NVIDIA GPU（GeForce RTX 1080 Ti）、i5-6500CPU，并运行在 Windows 10-1909 操作系统上，以确保实验过程的顺利进行。选择了 ResNet34 和 ResNet50 等不同深度的经典卷积神经网络作为实验模型，在 NYU v2 上进行了 500 个 epoch 的训练，在 SUN RGB-D 上进行了 150 个 epoch，批量大小、动量和权重衰减分别被设置为 8、0.9 和 1×10^{-4}。在进行实验前，先对输入数据进行了预处理，包括图像归一化和深度信息处理。主要评估指标包括 Acc、mAcc、mIoU 和 FPS，分别用于衡量模型的语义分割精度和计算效率。为全面评估模型性能，设计了多组实验，

包括用第 3 章模型作为基线模型进行对比，每组实验均经过多次运行以获得稳定的结果。通过这样的实验设置，可以为 RGB-D 语义分割模型的性能评估提供可靠的科学基础。

4.4.2 结果分析

与基线模型对比：对于第 3 章基线方法，使用形状卷积替换了普通卷积层，没有对其他设置进行任何更改。这保证了实验的公平性和一致性，仅通过调整卷积层的形状，可以独立地评估新的卷积结构对模型性能的影响，而不受其他超参数调整的影响。

表 4-1 展示了通过对 Baseline 和本章节的模型在 NYU v2 和 SUN RGB-D 数据集上的实验结果进行综合分析，本章节的模型在两个主干网络（ResNet34 和 ResNet50）配置下均实现了性能的显著提升。相较于 ResNet34，ResNet50 在两个数据集上均实现了显著提升的 mIoU，分别为 50.88%和 51.56%。特别是在 ResNet50 配置下，本章节模型相对于 Baseline 在 NYU v2 和 SUN RGB-D 数据集上的 mIoU 分别提高了 1.34%和 2.33%，展示了在更复杂场景中学习语义信息的卓越能力，而在计算效率上仅下降了约 0.21 FPS。这数字化差异进一步强调了本章模型在性能提升的同时，对计算效率的有效管理，尤其是在 ResNet34 配置下，模型的计算效率高于对应 Baseline，凸显了在性能与效率平衡上的优越性。强调了在实际应用中的可行性。总体而言，本章研究在性能优越性、灵活性和性能与效率平衡方面取得了显著成果，为 RGB-D 语义分割任务提供理论基础。

表 4-1 与基线模型对比

模型	主干	mIoU(%)		Param(M)	FPS
		NYU v2	SUN RGB-D		
Baseline	ResNet34	47.65	47.25	48.83	33.56
	ResNet50	50.88	48.50	56.44	24.23
Ours	ResNet34	47.96	47.68	48.80	33.44
	ResNet50	51.56	49.63	56.45	24.24

与当前领先的技术（SOTA）进行对比：表 4-2 展示了在 NYU v2 数据集上相较于先进的模型的对比实验，如 FCN、LSD-GF、D-CNN 等，本章模型在关键

指标上取得了显著的优势。具体而言，本章节的方法在像素准确率、平均准确率和平均交并比方面分别达到了 75.8%、62.8%和 50.2%的高水平，显示出其在图像语义分割任务中的卓越性能。

表 4-2 NYU v2 数据集上与其他方法的性能比较

模型	Acc(%)	mAcc(%)	mIoU(%)
FCN[66]	65.4	46.1	34.9
LSD-GF[40]	71.9	60.7	45.9
D-CNN[38]	-	61.1	48.4
MMAF-Net[70]	72.2	59.2	46.9
ACNet[42]	-	-	48.3
CFN[9]	-	-	47.7
3DGNN[67]	-	55.7	42.1
RDFNet[29]	76.0	62.8	47.7
M2.5D[52]	76.9	-	50.9
SegNet[11]	76.8	63.3	49.0
Ours	75.8	62.8	51.6

表 4-3 展示了在 SUN RGB-D 数据集的对比实验，本章的方法在像素准确率上达到了 82.0%，在平均准确率和平均交并比方面分别为 58.5%和 49.6%。这一优异的性能与 MMAF-Net 和 SegNet 相媲美，在 Acc 和 mIoU 方面超越了它们。特别值得注意的是，本章的模型在 mIoU 上实现了 49.6%，在 CFN、3DGNN 等其他方法之上，突显了在像素级别的分割任务中的卓越性

表 4-3 SUN RGB-D 数据集的性能比较。

模型	Acc(%)	mAcc(%)	mIoU(%)
3DGNN[67]	-	55.7	44.1
D-CNN[38]	-	53.5	42.0
MMAF-Net[70]	81.0	58.2	47.0
SegNet[11]	81.0	59.8	47.5
CFN[9]	-	-	48.1
3DGNN[67]	-	57.0	45.9
Ours	82.0	58.5	49.6

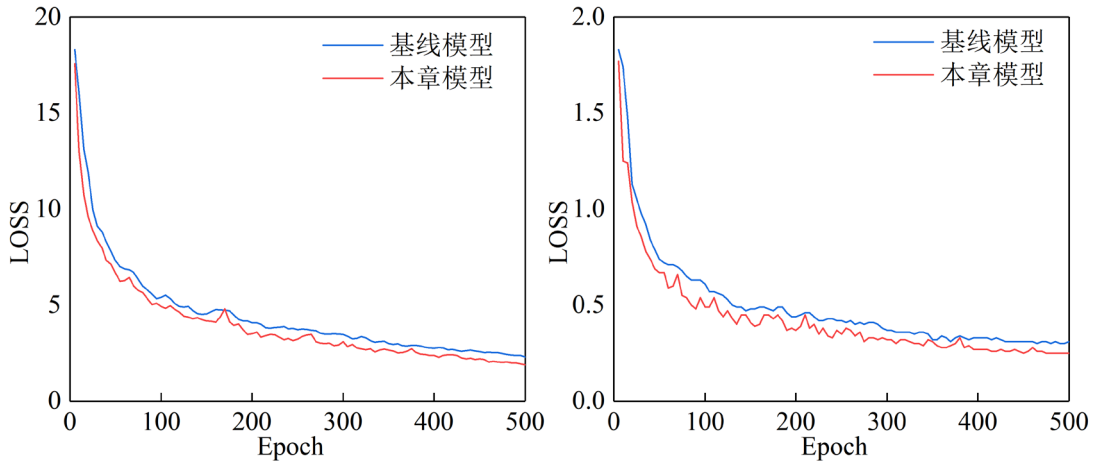
表 4-4 展示了在 NYU v2 数据集上基线模型和改进模型的 mIoU，分别表示基线模型和本章模型的结果，可以看出在绝大多数类别上，改进模型的 mIoU 都

高于基线模型，特别是对于具有明显形状特征的物体，例如桌子、门、窗户等，这些物体通常具有明显的形状特征，形状卷积能够更好地捕获这些特征，从而提高了物体的分割准确性，表明本章模型在语义分割任务中的性能优于基线模型。

表 4-4 与基线模型每一类 mIoU 对比

	墙	地板	橱柜	床	椅子	沙发	桌子	门	窗户	书架
基线	**0.773**	**0.836**	0.634	0.621	0.662	0.685	0.485	0.466	**0.523**	0.464
Ours	0.760	0.801	**0.651**	**0.729**	**0.686**	**0.695**	**0.508**	**0.483**	0.505	**0.499**
	图片	柜台	百叶窗	书桌	架子	窗帘	梳妆台	枕头	镜子	地垫
基线	**0.550**	0.688	**0.605**	0.266	**0.222**	0.501	0.521	0.511	**0.451**	**0.425**
Ours	0.501	**0.724**	0.592	**0.305**	0.211	**0.596**	**0.554**	**0.548**	0.403	0.411
	衣物	天花板	书籍	冰箱	电视	纸巾	毛巾	淋浴	盒子	板子
基线	0.241	**0.779**	**0.305**	0.611	0.653	0.329	0.333	0.412	**0.181**	**0.783**
Ours	**0.264**	0.765	0.289	**0.652**	**0.696**	**0.339**	**0.358**	**0.459**	0.152	0.720
	人	床头柜	马桶	水槽	灯	浴缸	包	其他物品	其他家具	其他物产
基线	0.812	0.451	**0.871**	**0.626**	**0.496**	**0.584**	0.062	**0.302**	0.318	0.315
Ours	**0.828**	**0.469**	0.847	0.597	0.474	0.55	**0.122**	0.226	**0.328**	**0.328**

通过在 SUN RGB-D 数据集上分别训练了基线模型和本章语义分割模型，图 4-3 分别展示了模型训练和验证过程中的损失函数曲线，本章模型在训练和验证集上的损失曲线都明显低于基线模型，本章模型收敛速度更快，说明了本章改进模型相对于基线模型具有更好的性能，主要原因是深度信息对深度信息的充分利用。



(a) 模型训练损失函数曲线　　(b) 模型验证损失函数曲线

图 4-3 模型训练和验证过程中的损失函数曲线

4.4.3 可视化

图 4-4 展示了两组在 NYU v2 数据集的可视化结果，列从左到右依次表示 RGB、Depth 图像、标签、基线和本章（Ours）的图像，其中的标签中黑色区域表示被忽略（void）的类别。如图 4-4 所示，相较于基线网络，本章网络可以很好地利用深度信息，尤其是细节信息来提取对象特征。例如，示例中的椅子和桌子区域的颜色逐渐变化，因此很难预测基线方法的准确分割边界。与传统的卷积层相比，形状卷积学习的形状特征使得根据几何提示进行精确切割。与基线相比，形状卷积还可以显著改善边缘区域的分割结果。

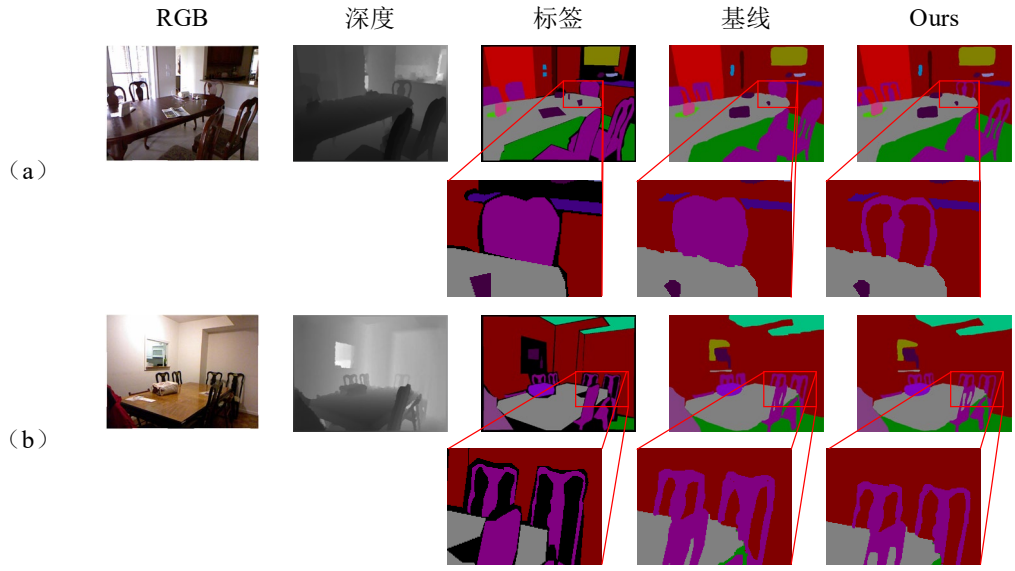


图 4-4 NYU v2 数据集的可视化结果

4.5 本章小结

本章介绍了在 RGB-D 语义分割网络中引入深度形状感知卷积和可学习上采样方法的相关内容，并进行了大量实验。首先介绍了传统卷积神经网络在处理形状不规则对象时的局限性，以及在 RGB-D 语义分割任务中面临的挑战。接着，详细讨论了深度形状感知卷积的原理和操作方法，从而更好地捕捉对象的形状信息。同时，还介绍了可学习上采样方法，避免了传统方法中的计算成本高和可能引入不希望的失真的问题，并提高了分割性能。本章在第 3 章基于 SwiftNet 的 RGB-D 图像语义分割网络的基础上进行改进，在实验部分，设计了一系列实验来验证深度形状感知卷积和可学习上采样方法的有效性，并与主流先进的方法进

行了比较。实验结果表明，本章提出的方法在 RGB-D 语义分割任务中取得了显著的性能提升，能够更准确地捕捉对象的形状信息，并有效地提高了分割精度。综上所述，本章所介绍的深度形状感知卷积和可学习上采样方法为 RGB-D 语义分割任务带来了新的思路和解决方案，具有重要的研究和应用价值。

第5章 基于自适应阈值通道交换 RGB-D 图像高效语义分割

5.1 引言

多模态数据的融合在图像处理领域中日益受到关注。RGB-D 图像多模态数据融合技术，通过结合 RGB 图像和深度图像的信息提高图像理解和分析的性能。在实际应用中，传统的融合方法往往依赖于简单的特征叠加或串行处理方式，无法有效地融合 RGB 和深度信息以及平衡融合过程中的模态间和模态内处理之间的权衡。本章提出了一种基于自适应阈值通道交换面向室内场景 RGB-D 图像高效语义分割方法，网络整体框架是基于本文第 4 章网络模型进一步改进实现。通过动态地交换通道来实现 RGB 和深度信息之间的有效融合。通道的重要性是由训练期间的批量标准化缩放因子的大小来自动衡量的，从而实现了对不同通道的自适应调整。此外，通过共享卷积滤波器并保持独立的批量归一化层，保证了模态间信息的有效交互和模态内处理的平衡。

在本章将详细介绍自适应阈值通道交换方法的原理和实现细节，并通过实验验证其在室内场景 RGB-D 图像语义分割任务中的有效性和性能优势。该方法的提出旨在为实现高效的多模态图像分析提供新的思路和方法，从而推动图像理解和场景理解领域的发展。

5.2 自适应阈值通道交换网络原理

多模态数据融合技术长期以来一直备受关注，并且已经做了很多工作。就融合结构而言，现有的方法可以分为早期融合、中期融合、晚期融合，早期融合通常在数据预处理阶段进行，中期融合发生在特征提取阶段，而晚期融合则是在决策阶段进行。就融合方式而言，现有的方法又可以分为基于聚合的融合、基于对齐的融合以及它们的混合。基于聚合的融合方法采用的操作如加权法、平均法、拼接法和自注意力机制，将不同模态的信息直接组合起来。而基于对齐的融合将来自不同模态的图像在某种特定的表示空间或特征空间上对齐，然后再进行融合。

尽管在多模态融合方面取得了显著成果，但如何整合各模态的共同信息，同时保留各模态的独特特性仍然是一个巨大的挑战。传统的基于聚合的融合倾向于

忽视模态内部的细节传播，而基于对齐的融合则由于通过训练对齐损失来实现模内信息交换，但这种交换往往较为微弱导致模态间融合的有效性不足。为了平衡模间融合和模内处理，可以采用聚合和对齐融合的分层组合方式提高性能，但这也伴随着额外的计算和工程开销。

由于模式内处理的弱点，基于聚合的融合方法在融合特征时倾向于平衡所有模态的子网络，以弥补模式内处理的不足。然而，这种方法的性能在很大程度上受到所选融合层的影响。相比之下，基于对齐的方法则通过调整相似度来对齐多模态特征，其中最大平均差异是衡量这种相似度的一个常用指标。尽管如此，单纯追求整体的对齐可能会忽视每个模态或领域的独特信息。为此，可以基于调制方法去关联模态的共性特征，同时保留模态的特定信息入手。

针对上述问题，本章使用了一种基于自适应阈值通道交换的方法，用于室内场景 RGB-D 图像的高效语义分割。该方法根据批量标准化的缩放因子来评估不同通道特征在训练过程中的重要性，使模型自主判断哪些通道的特征在训练过程中不会对参数更新产生显著影响，利用通道交换技术使这些原本被忽略的通道特征再次参与到模型的参数更新中，从而有效地提高模型的性能。

为了对齐多模态特征并处理模态间的差异，采用了一种自适应阈值通道交换网络。该方法旨在确保模态间的充分互动并有效地学习模态内的特征，首先利用批量标准化（Batch Normalization，BN）层对每个批次的输入进行标准化处理来消除协变量偏移。假设 $x_{m,l}$ 表示第 m 个子网络的第 l 层特征图，$x_{m,l,c}$ 表示第 c 个通道。BN 层对 $x_{m,l,c}$ 进行归一化处理，然后进行仿射变换，如式(5-1)所示。

$$x'_{m,l,c}=\gamma_{m,l,c}\frac{x_{m,l,c}-\mu_{m,l,c}}{\sqrt{\sigma^2_{m,l,c}+\varepsilon}}+\beta_{m,l,c} \tag{5-1}$$

其中，$\mu_{m,l,c}$ 和 $\sigma_{m,l,c}$ 分别计算当前小批数据的所有像素位置上所有激活的平均值和标准偏差；$\gamma_{m,l,c}$ 和 $\beta_{m,l,c}$ 分别是可训练的缩放系数和偏移量；ε 是一个小常数，以避免被零除。第 $(l+1)$ 层以 $\{x'_{m,l,c}\}_c$ 作为输入，再通过一个非线性函数进行变换后输出。

式(5-1)中的因子 $\gamma_{m,l,c}$ 评估了训练期间输入 $x_{m,l,c}$ 和输出 $x'_{m,l,c}$ 之间的相关性。如果 $\gamma_{m,l,c}\to 0$，损失的梯度将接近 0，意味着 $x_{m,l,c}$ 将失去对最终预测的影响，

即当前信道 $x_{m,l,c}$ 在某一训练步骤中由于 $\gamma_{m,l,c} \to 0$ 而变得多余，那么此后它几乎就不会参与梯度更新。

假设有 M 个模态的第 i 个输入数据，$x^{(i)} = \{x_m^{(i)} \in R^{C\times(H\times W)}\}_{m=1}^{M}$，其中 C 表示通道的数量，H 和 W 表示特征图的高度和宽度。将 N 定义为批次大小。深度多模态融合的目标是确定一个多层网络 $f(x^{(i)})$，其输出 $\hat{y}^{(i)}$ 有望尽可能地适合目标 $y^{(i)}$。这可以通过最小化损失函数来实现，如式(5-2)所示。

$$\min_{f} \frac{1}{N}\sum_{i=1}^{N} L\left(\hat{y}^{(i)} = f\left(x^{(i)}, y^{(i)}\right)\right) \tag{5-2}$$

整体方法的优化目标如式(5-3)所示，定义为最小化损失函数和约束条件来确定不同模态输出的加权组合，同时通过正则化项对模型进行正则化。

$$\min_{f_{1:M}} \frac{1}{N}\sum_{i=1}^{N} L\left(\sum_{m=1}^{M}\alpha_m f_m\left(x^{(i)}, y^{(i)}\right)\right) + \lambda\sum_{m=1}^{M}\sum_{l=1}^{L}\left|\hat{\gamma}_{m,l}\right|, \quad s.t.\sum_{m=1}^{M}\alpha_m = 1 \tag{5-3}$$

其中，子网络 $f_m(x^{(i)})$ 与融合方程相对，通过信道交换融合多模态信息，每个子网络配备包含第 l 层的缩放因子 $\gamma_{m,l}$ 的 BN 层，将惩罚其某部分缩放因子 $\hat{\gamma}_{m,l}$ 的 L1 范数。子网络 f_m 除了 BN 层外，共享相同的参数，以方便通道交换以及进一步压缩结构。集合输出的决策分数 α_m，由类似于基于对齐的方法的 softmax 输出来训练。

通过式(5-3)的设计，进行了跨模态的无参数信息融合，同时保持每个子网络的自我传播，以便描述每个模态的特定统计。此外，信道交换融合是自适应的，很容易嵌入到各个子网络中，可推导出式(5-4)。

$$x'_{m,l,c} = \begin{cases} \gamma_{m,l,c}\dfrac{x_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma^2_{m,l,c} + \varepsilon}} + \beta_{m,l,c}, \ if \ \ \gamma_{m,l,c} > \theta \\ \dfrac{1}{M-1}\displaystyle\sum_{m'\neq m}^{M}\gamma'_{m,l,c}\dfrac{x_{m',l,c} - \mu_{m',l,c}}{\sqrt{\sigma^2_{m',l,c} + \varepsilon}} + \beta_{m,l,c}, \ \ \text{else;} \end{cases} \tag{5-4}$$

其中，如果当前通道的缩放系数小于某个阈值 $\theta \approx 0+$，则用其他通道的平均值替换。简而言之，如果一个模态的一个通道对最终预测的影响很小，那么就用其他模态的平均值来代替它。在将每个模态送入非线性激活之前，对每个模态应用式(5-4)，然后在下一层中进行卷积。梯度从被替换的通道中分离出来，并通过新的

通道反向传播。

图 5-1 形象的展示了特征通道交换过程。将整个通道分为 M 个相等的子部分，并且只在每个不同的子部分对不同的模式进行通道交换。用 $\gamma_{m,l}$ 来表示允许被替换的缩放因子。为了优化模型性能，在目标函数中引入了稀疏性约束，以便识别出不必要的通道，从而提高整体计算效率。式(5-1)中的交换是一个只在通道的一个子部分内的定向过程，它不仅有望在其他 $M-1$ 个子部分中保留模态特定的传播，而且还能避免无用的交换，因为 $\gamma_{m_0,l,c}$ 与 $\gamma_{m,l,c}$ 不同，所以不在稀疏性约束范围内。

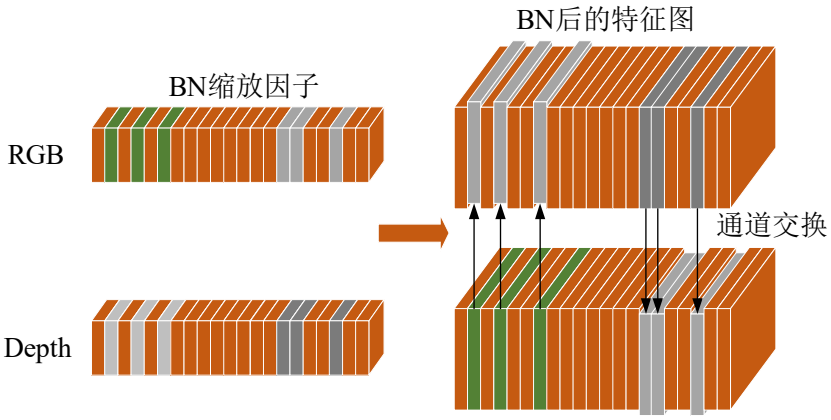


图 5-1 通道交换过程

5.3 网络结构

模型采用了一种基于 SwiftNet 双分支编码器-解码器结构，其中 ResNet50 被用作骨干网络。该结构旨在有效处理多模态数据。在此结构中，两种模态首先经过7×7卷积进行特征提取和降维，然后数据经过图 5-2 中四个下采样模块（RB）进行特征提取，每个具有相同分辨率的残差块的输出到图 5-2 中特征融合模块（Fuse）进行特征融合，两种模态的特征进行了注意力融合，以学习哪种模态的哪些特征对于任务最为重要，并抑制不相关的特征。自适应阈值通道交换的方法应用于双分支之间，用于进一步优化模型的性能。自适应阈值通道交换的主要目标是根据模型的需要自动调整通道之间的信息交换，以实现更好的特征学习和表示。解码器与编码器之间，编码部分中的多级特征经过 1×1 卷积来灵活地调整通道数，然后通过跳跃连接分别到图 5-2 中上采样模块（UP）进行分辨率恢复，实现信息的无缝传递，保持了高层次特征的空间信息，有助于提高模型对局部细

节的捕捉能力。这种结构的目的是确保网络能够充分利用输入数据的信息，同时保持梯度的稳定性的同时使其更易于训练。

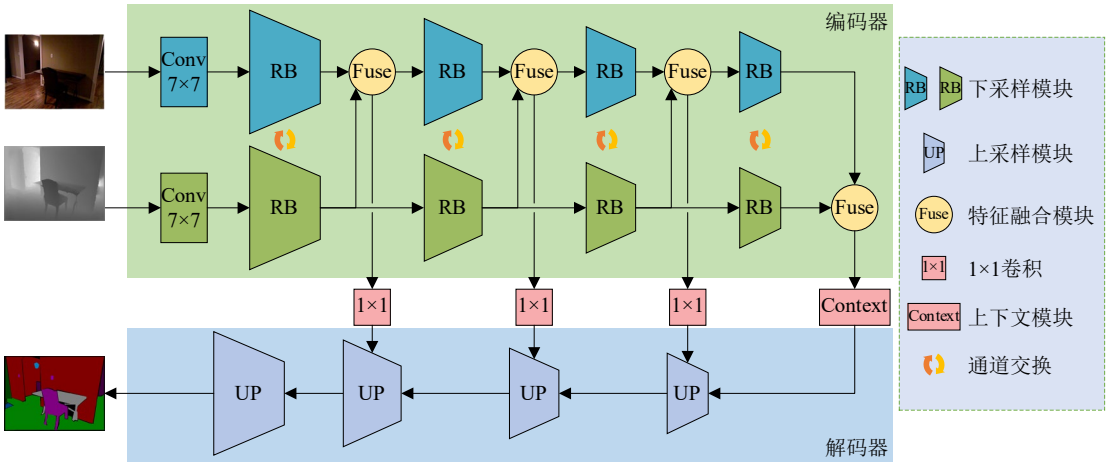


图 5-2 网络整体结构

5.4 实验设计及结果分析

5.4.1 实验设置

模型的主干分别选择了 ResNet50 经典残差网络作为模型的主干。在进行实验前，先对输入数据进行了预处理，包括图像归一化和深度信息处理。对本章节的模型进行了在两个广泛被使用的公共数据集（NYU v2 和 SUN RGB-D）的评估。这两个数据集都将 RGB 和深度信息作为输入。在 NYU v2 数据集上，遵循标准设置，使用 795 张图像进行训练，654 张图像进行测试，并采用标准的 40 个语义类别进行预测。至于 SUN RGB-D，它是室内语义分割领域中最具挑战性的大规模基准之一，包含 10335 张 RGB-D 图像，涵盖 37 个语义类别。在训练过程中，使用了第 4 章提出的 RGB-D 图像语义分割框架，编码器和解码器的初始学习率分别设置为 5×10^{-4} 和 3×10^{-3}。在 NYU v2 上进行了 500 个 epoch 的训练，在 SUN RGB-D 上进行了 150 个 epoch。批量大小、动量和权重衰减分别被设置为 6、0.9 和 1×10^{-4}。训练过程中设置了参数 λ 为 5×10^{-3}，并将阈值 θ 设定为 0.02。采用平均交比、像素准确度以及平均准确度作为评价指标，全面评估模型的性能。

5.4.2 结果分析

与基线模型对比：通过在 NYU v2 和 SUN RGB-D 数据集上与基线模型进行比较，突显了本研究的先进性。基线模型采用 ResNet50 作为主干，在 NYU v2 上获得了 75.8%的像素准确率和 62.8%的平均像素准确率，以及 51.5%的平均交并比。相较之下，本章的模型（ResNet50）在两个数据集上均取得了更高的性能，分别达到了 77.7%和 81.8%的像素准确率，相比基线模型提高了 2.51%和 2.76%，65.3%和 62.1%的平均像素准确率相比基线模型提高了 3.98%和 0.32%，以及 52.5%和 49.9%的平均交并比相比基线模型提高了 1.94%和 0.60%。这表明本章的方法在语义分割任务中的卓越表现，超越了基线模型的性能水平。

表 5-1 与基线模型对比结果

模型	主干	NYU v2			SUN RGB-D		
		Acc(%)	mAcc(%)	mIoU(%)	Acc(%)	mAcc(%)	mIoU(%)
基线	ResNet50	75.8	62.8	51.5	79.6	61.9	49.6
Ours	ResNet50	77.7	65.3	52.5	81.8	62.1	49.9

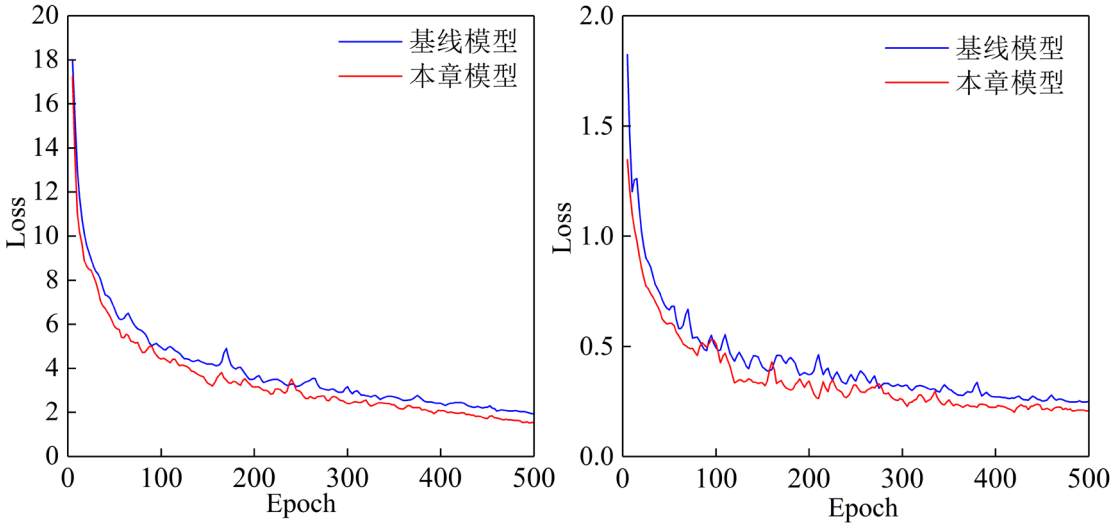
与当前领先的技术（SOTA）进行对比：在 NYU v2 数据集上使用 ResNet50 和单尺度推理策略进行公平比较，展示了 40 个类别的分割结果。遵循前面提到的训练方法，设置参数 λ 为 0.005，并将阈值 θ 设定为 0.02，本章的方法仍然可以实现 77.5%的像素准确率，65.3%的平均像素准确率，和 52.5%的平均交并比，优于当前最先进的方法。在 SUN RGB-D 数据集上，还将本章的方法与当前最先进的方法进行了比较。本章的模型实现了 81.8%的像素准确率，62.1%的平均像素准确率和 49.9%的平均交并比，在 mAcc 方面比基线模型结果更佳。

表 5-2 与主流算法对比实验结果

模型	主干	NYU v2			SUN RGB-D		
		Acc(%)	mAcc(%)	mIoU(%)	Acc(%)	mAcc(%)	mIoU(%)
FuseNet[55]	VGG16	68.1	50.4	37.9	68.4	41.1	29.0
FCN[66]	-	65.4	46.1	34	68.2	38.4	27.4
ACNet[42]	ResNet50	-	-	48.3	-	-	48.1
SSMA[69]	ResNet50	75.2	60.5	48.7	81.0	58.1	45.7
CFN[9]	ResNet152	-	-	47.7	-	-	48.1
RDFNet[29]	ResNet101	75.6	62.2	49.1	80.9	59.6	47.2
REDNet[68]	ResNet50	-	62.6	47.2	81.3	60.3	47.8
ACNet[42]	ResNet50	-	63.1	48.3	-	60.	48.1
Ours	ResNet50	77.7	65.3	52.5	81.8	62.1	49.9

通过在 SUN RGB-D 数据集上分别训练了基线模型和本章基于通道交换的语

义分割模型，图 5-3 并分别展示了模型训练和验证过程中的损失函数曲线，改进模型在训练和验证集上的损失曲线都明显低于基线模型，本章模型收敛速度更快，说明了改进模型相对于基线模型具有更好的性能。



(a) 模型训练损失函数　　(b) 模型验证损失函数

图 5-3 模型训练和验证过程中的损失函数曲线

5.4.3 可视化

如图 5-4 所示，通过引入通道交换网络，成功地提高了椅子轮廓的分割清晰度。具体地，可以观察到椅子轮廓得到了显著的增强。网络能够更好地区分椅子与周围环境的边界，确保分割结果更加精细且贴合实际场景。这种通道交换机制使得网络能够聚焦于椅子类别的关键特征，通过巧妙地交换和调整通道信息，本章节的网络能够捕获到椅子的细节和边缘信息，从而产生更加准确和清晰的分割结果。提高了语义分割的特征的充分利用。

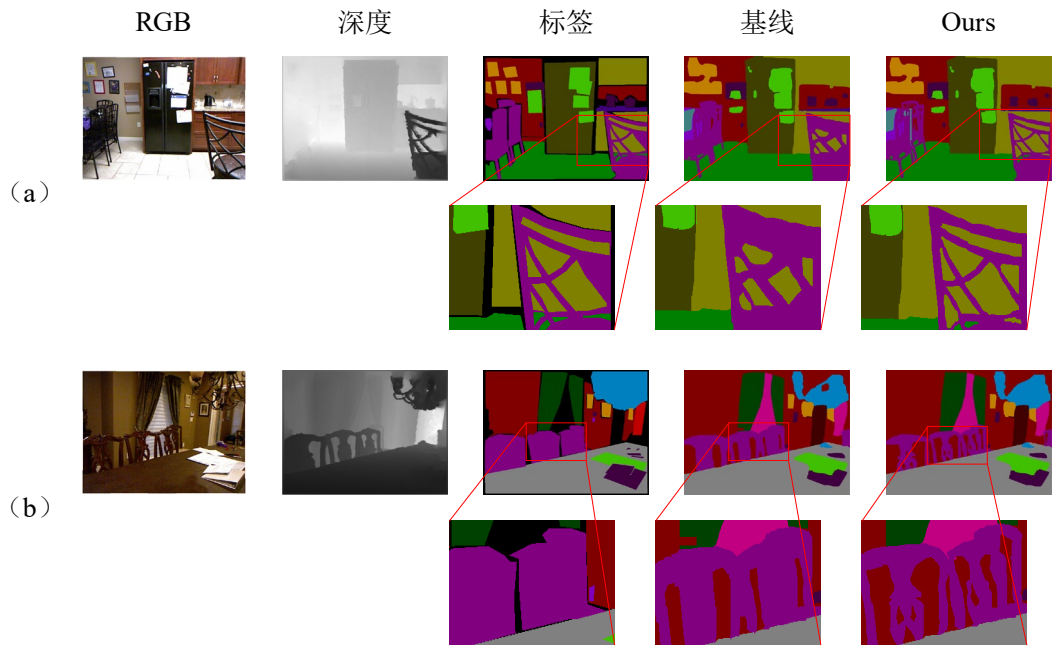


图 5-4 结果可视化

5.5 本章小结

本章主要工作是在 RGB-D 语义分割网络中引入了自适应阈值通道交换优化方法，并进行了大量实验验证。首先，回顾了批量归一化层的原理和作用，以及深度多模态融合的目标和优化目标。然后，介绍了自适应阈值通道交换的原理，该方法通过对缩放因子进行稀疏惩罚和自适应的通道交换融合，实现了跨模态的无参数信息融合，以及减少模态间的冗余信息，有助于进一步提高模型的性能和鲁棒性。在实验部分，设计了一系列实验来验证自适应阈值通道交换方法的有效性，并与传统方法进行了比较。实验结果表明，本章提出的方法在 RGB-D 语义分割任务中取得了显著的性能提升，能够更准确地捕捉对象的形状信息，并有效地提高了分割精度。综上所述，本章所介绍了一种基于 SwiftNet 的自适应阈值通道交换优化方法为 RGB-D 语义分割网络的性能提升提供了新的思路和解决方案，具有重要的研究和应用价值。

第6章 总结与展望

6.1 总结

本文主要研究了面向室内 RGB-D 场景的语义分割问题，通过引入新的方法和技术，取得了一定的研究成果。具体来说，本文的创新点主要包括基于 SwiftNet 的室内 RGB-D 场景语义分割网络、基于形状感知卷积的高效语义分割方法以及基于自适应阈值通道交换的语义分割网络，本文的主要研究工作和研究成果包含以下几个方面：

（1）针对 RGB 图像分割的局限性，提出了基于 SwiftNet 的室内 RGB-D 场景语义分割网络，通过引入注意力机制和优化特征融合方式，提高了网络的分割精度。此外，利用 Non-Bottleneck-1D 和改进的快速聚合上下文模块，增强了网络的速度和全局特征提取能力，设计了一种高效的网络结构。进行试验验证模型的有效性。

（2）针对目前 RGB-D 分割方法大都以同样的方式去处理深度数据，没有充分利用深度模态对 RGB 模态的本质差别，由此采用了深度形状感知卷积针对性提取深度特征中的形状特征，从而更好地捕捉对象的形状信息，提高深度模态特征的充分利用。同时针对传统上采样导致信息损失或混叠效应问题，利用可学习上采样方法通过将卷积核进行转置来实现上采样，卷积核的权重是可学习的，可以通过训练来优化上采样的效果。实验结果验证了该方法对物体形状的轮廓能够更好地进行分割。

（3）针对在语义分割研究中，现有的基于聚合和基于对齐的 RGB-D 融合方法在平衡模态间融合和模态内处理之间权衡不足的问题，提出了一种基于自适应阈值通道交换的语义分割网络。通过在不同模态的子网络之间动态地交换通道，确保了信息流通的高效性。同时交换过程的有效性是通过共享卷积滤波器来保证的，在不同的模态之间保持独立的批处理标准化层，有效减少了模态间信息的浪费，并在训练过程中充分利用了数据特征，提升了语义分割的性能。

6.2 未来展望

未来的研究方向将集中在进一步优化和改进室内 RGB-D 场景语义分割技术。首先将致力于扩展和完善数据集，探索更复杂和真实的场景数据，以提高模型的泛化能力和适用性。同时将继续改进网络结构和算法，探索更有效的语义分割方法，并将其应用于智能监控、机器人导航等实际场景，为智能系统提供更精准和可靠的场景理解能力。接下来可以从以下方面进行深入研究：

（1）当前实验数据集主要依赖于公共数据集，对于模型的鲁棒性和泛化能力等方面的研究尚不够深入。未来的研究方向可以是构建一套新的数据集，涵盖多种室内环境，包括家居、购物中心、超市、办公室等场景。这样的数据集将有助于更全面地评估模型在不同环境下的性能，并提供更具挑战性和实用性的测试平台。计划通过采集多角度、多尺度的图像来细分研究，同时根据需求进行数据调整和增强，以提高模型的泛化能力。

（2）改进方法在处理分类不均衡问题时遇到了挑战，导致不同物体类别的分割准确率存在显著差异。虽然部分类别的分割效果较为理想，但其他类别的分割结果相对较差，从而影响了整体的分割质量。为了实现更均衡的分割结果，并提高各类别的分割精度，需要进一步优化改进方法。尽管本文提出的两种方法在一定程度上提升了分割精度，但仍有进一步改进的空间。因此，如何更有效地解决分类不均衡问题，提高各类别物体的分割准确率，将成为后续研究的重要方向。

参考文献

[1] Sagar A, Soundrapandiyan R K. Semantic segmentation with multi scale spatial attention for self driving cars[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 2650-2656.

[2] Shen Y, Zhang H, Fan Y, et al. Smart health of ultrasound telemedicine based on deeply represented semantic segmentation[J]. IEEE Internet of Things Journal, 2020, 8(23): 16770-16778.

[3] Kherraki A, Maqbool M, El Ouazzani R. Traffic scene semantic segmentation by using several deep convolutional neural networks[C]//2021 3rd IEEE Middle East and North Africa COMMunications Conference (MENACOMM). IEEE, 2021: 1-6.

[4] Liu Y, Miura J. RDS-SLAM: Real-time dynamic SLAM using semantic segmentation methods[J]. IEEE Access, 2021, 9(2): 23772-23785.

[5] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] Silberman N, Fergus R. Indoor scene segmentation using a structured light sensor[C]//2011 IEEE international conference on computer vision workshops (ICCV workshops). IEEE, 2011: 601-608.

[7] Zhang Z. Microsoft kinect sensor and its effect[J]. IEEE Multimedia, 2012, 19(2): 4-10.

[8] Wang J, Wang Z, Tao D, et al. Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer International Publishing, 2016: 664-679.

[9] Ji J, Shi R, Li S, et al. Encoder-decoder with cascaded CRFs for semantic segmentation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(5): 1926-1938..

[10] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[11] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.

[12] Otsu N. A threshold selection method from gray-level histograms[J]. IEEE transactions on systems, man, and cybernetics, 1979, 9(1): 62-66.

[13] Yanowitz S D, Bruckstein A M. A new method for image segmentation[J]. Computer Vision, Graphics, and Image Processing, 1989, 46(1): 82-95.

[14] Canny J. A computational approach to edge detection[J]. IEEE Transactions on pattern analysis and machine intelligence, 1986,1(6): 679-698.

[15] 刘松涛, 殷福亮. 基于图割的图像分割方法及其新进展[J]. 自动化学报, 2012, 38(6): 911-922.

[16] Cheng H D, Jiang X H, Sun Y, et al. Color image segmentation: advances and prospects[J]. Pattern recognition, 2001, 34(12): 2259-2281.

[17] Tani Y, Ono K, Yamakawa S, et al. Semantic Segmentation of Spine and Femur Bone Using Atrous Spatial Pyramid Pooling-based U-Net with Fully Connected CRF[C]//2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2023: 4967-4972.

[18] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[19] Gao R. Rethinking dilated convolution for real-time semantic segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 4675-4684.

[20] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[21] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[22] Çiçek Ö, Abdulkadir A, Lienkamp S S, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation[C]//Medical Image Computing and

Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. Springer International Publishing, 2016: 424-432.

[23] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). IEEE, 2016: 565-571.

[24] Guo C, Szemenyei M, Yi Y, et al. Sa-unet: Spatial attention u-net for retinal vessel segmentation[C]//2020 25th international conference on pattern recognition (ICPR). IEEE, 2021: 1236-1242.

[25] Zhao H, Qi X, Shen X, et al. Icnet for real-time semantic segmentation on high-resolution images[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 405-420.

[26] Miao S, Du Y, Lu X, et al. Semantic segmentation of vehicle vision based on two-branch Enet network[C]//2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA). IEEE, 2023: 477-481.

[27] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 325-341.

[28] Wang C, Wang C, Li W, et al. A brief survey on RGB-D semantic segmentation using deep learning[J]. Displays, 2021, 70(1): 102080-102088.

[29] Park S J, Hong K S, Lee S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4980-4989.

[30] Sun L, Yang K, Hu X, et al. Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images[J]. IEEE robotics and automation letters, 2020, 5(4): 5558-5565.

[31] Fooladgar F, Kasaei S. A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks[J]. Multimedia Tools and Applications, 2020, 79(7-8): 4499-4524.

[32] Shorten C, Khoshgoftaar T M. A survey on image data augmentation for deep learning[J]. Journal of big data, 2019, 6(1): 1-48.

[33] Zhang H, Zhang H, Wang C, et al. Co-occurrent features in semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 548-557.

[34] Orsic M, Kreso I, Bevandic P, et al. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12607-12616.

[35] 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.

[36] Atrey P K, Hossain M A, El Saddik A, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia systems, 2010, 16: 345-379.

[37] Su Y, Yuan Y, Jiang Z. Deep feature selection-and-fusion for RGB-D semantic segmentation[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021: 1-6.

[38] Wang W, Neumann U. Depth-aware cnn for rgb-d segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 135-150.

[39] Zhang G, Xue J H, Xie P, et al. Non-local aggregation for RGB-D semantic segmentation[J]. IEEE Signal Processing Letters, 2021, 28: 658-662.

[40] Cheng Y, Cai R, Li Z, et al. Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3029-3037.

[41] Fang A, Zhao X, Zhang Y. Cross-modal image fusion guided by subjective visual attention[J]. Neurocomputing, 2020, 414: 333-345.

[42] Hu X, Yang K, Fei L, et al. Acnet: Attention based network to exploit complementary features for RGB-D semantic segmentation[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 1440-1444.

[43] Wang Y, Huang W, Sun F, et al. Deep multimodal fusion by channel exchanging[J]. Advances in neural information processing systems, 2020, 33: 4835-4845.

[44] Xing Y, Wang J, Chen X, et al. Coupling two-stream RGB-D semantic segmentation network by idempotent mappings[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 1850-1854.

[45] Xing Y, Wang J, Chen X, et al. 2.5 D convolution for RGB-D semantic segmentation[C]//2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 1410-1414.

[46] Xing Y, Wang J, Zeng G. Malleable 2.5 d convolution: Learning receptive fields along the depth-axis for rgb-d scene parsing[C]//European Conference on

Computer Vision. Cham: Springer International Publishing, 2020: 555-571.

[47] Wang W, Neumann U. Depth-aware cnn for rgb-d segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 135-150.

[48] Chen L Z, Lin Z, Wang Z, et al. Spatial information guided convolution for real-time RGB-D semantic segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 2313-2324.

[49] Chen Y, Mensink T, Gavves E. 3D neighborhood convolution: Learning depth-aware features for RGB-D and RGB semantic segmentation[C]//2019 International Conference on 3D Vision (3DV). IEEE, 2019: 173-182.

[50] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 325-341.

[51] Wang L, Li D, Zhu Y, et al. Dual super-resolution learning for semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3774-3783.

[52] Jiao J, Wei Y, Jie Z, et al. Geometry-aware distillation for indoor semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2869-2878.

[53] Khan S H, Bennamoun M, Sohel F, et al. Integrating geometrical context for semantic labeling of indoor scenes using RGB-D images[J]. International Journal of Computer Vision, 2016, 117: 1-20.

[54] Fan M, Lai S, Huang J, et al. Rethinking bisenet for real-time semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9716-9725.

[55] Hazirbas C, Ma L, Domokos C, et al. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture[C]//Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13. Springer International Publishing, 2017: 213-228.

[56] 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.

[57] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. nature, 1986, 323(6088): 533-536.

[58] Hubel D H, Wiesel T N. Receptive fields and functional architecture of monkey striate cortex[J]. The Journal of physiology, 1968, 195(1): 215-243.

[59] Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological cybernetics, 1980, 36(4): 193-202.

[60] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[61] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88: 303-338.

[62] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.

[63] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 3213-3223.

[64] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGB-D images[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. Springer Berlin Heidelberg, 2012: 746-760.

[65] Song S, Lichtenberg S P, Xiao J. Sun rgb-d: A rgb-d scene understanding benchmark suite[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 567-576.

[66] 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.

[67] Qi X, Liao R, Jia J, et al. 3d graph neural networks for rgbd semantic segmentation[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5199-5208.

[68] Peng X, Feris R S, Wang X, et al. Red-net: A recurrent encoder–decoder network for video-based face alignment[J]. International Journal of Computer Vision, 2018, 12(6): 1103-1119.

[69] Qian Y, Deng L, Li T, et al. Gated-residual block for semantic segmentation using RGB-D data[J]. IEEE Transactions on Intelligent Transportation Systems, 2021,

23(8): 11836-11844.

[70] Nazir D, Pagani A, Liwicki M, et al. Semattnet: Toward attention-based semantic aware guided depth completion[J]. IEEE Access, 2022, 10(1): 120781-120791.

[71] Seichter D, Köhler M, Lewandowski B, et al. Efficient rgb-d semantic segmentation for indoor scene analysis[C]//2021 IEEE international conference on robotics and automation (ICRA). IEEE, 2021: 13525-13531.

[72] Chen X, Ma N, Wang M. YOLOv5-pothole: An improved pothole perception method based on YOLOv5-seg[C]//2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR). IEEE, 2023: 75-79.

[73] Achanccaray P, Gerke M, Wesche L, et al. On the Assessment of Instance Segmentation for the Automatic Detection of Specific Constructions from Very High Resolution Airborne Imagery[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2023, 48(1): 1303-1309.

攻读学位期间所发表的学术论文

[1] 王博,许钢,苏世林.一种基于 SwiftNet 面向室内 RGBD 场景的高效语义分割算法 [J/OL]. 重庆工商大学学报(自然科学版),1-10[2024-05-27].http://kns.cnki.net/kcms/detail/50.1155.N.20231218.1126.002.html.

[2] Su S, Xu G, Wang B. Research on semantic segmentation of night infrared image for road scene[C]//Third International Conference on Signal Image Processing and Communication (ICSIPC 2023). SPIE, 2023, 12916: 314-320.

致谢

站在毕业的门槛上，回首过去的三年，心中涌动着深深的喜悦和感激之情。在此刻，这三年是个人和学业发展的旅程，充满了挑战、胜利和丰富的学习机会。我对这个过程产生了真挚的喜爱，全情投入其中，我不禁要表达对我所珍视的经历和在我成长中扮演重要角色的人们的感激之情。

在即将完成硕士学位论文的这一时刻，我首先要衷心感谢我的导师许钢教授。您温文尔雅的待人风格让人感到温馨而亲切，您的卓越学识和严谨科研态度一直是我学术生涯中的楷模。在整个研究过程中，您的悉心指导和无私帮助使我能够克服难题，深入研究每一个问题。每次与您的学术讨论都让我受益匪浅，您的建议和批评使我的研究更加严谨。在此，我向您表达最衷心的感谢和最崇高的敬意。

同时，我要感谢实验室的所有同学们，苏世林、俞泽、李想、朱志敏，在这个研究旅程中，有了你们的支持与合作，我才得以顺利进行实验室工作。为我的研究提供了宝贵的帮助。我们共同协作的时光充满了团队精神和合作意识，成为我学术道路上难忘的一部分。

特别感谢我的家人，感谢你们无条件的爱和支持，谢谢你们为我付出的一切。有你们的支持和陪伴，我才能毫无顾虑地追求自己的梦想，你们是我最坚实的后盾。

最后，我要感谢所有给予我支持和帮助的人，无论是直接还是间接地。每一个人的贡献都是我能够完成这篇论文的重要因素。在这段学习旅程中，我汲取了丰富的知识和人生经验，我将珍惜这段经历，并在未来的学术道路上继续努力。

谨以此致谢，表达我深深的感激之情。