汕頭大學
SHANTOU UNIVERSITY

硕士研究生学位论文

题目：基于注意力机制的室内RGBD图像语义分割

姓　　名：	邱本章
学　　号：	111709029
院　　系：	工学院
专　　业：	信息与通信工程
研究方向：	数字图像处理
导师姓名：	范衡教授

二〇二〇年 六月

学位论文原创性声明和学位论文使用授权声明

学位论文原创性声明

本论文是我个人在导师指导下进行的工作研究及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或其它机构已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在论文中以明确方式标明。本人完全意识到本声明的法律责任由本人承担。

作者签名：_____ 日期：_____年____月____日

学位论文使用授权声明

本人授权汕头大学保存本学位论文的电子和纸质文档，允许论文被查阅和借阅；学校可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、缩印或其它复制手段保存和汇编论文；学校可以向国家有关部门或机构送交论文并授权其保存、借阅或上网公布本学位论文的全部或部分内容。对于保密的论文，按照保密的有关规定和程序处理。

作者签名：_____ 导师签名：_____

日期：_____年____月____日 日期：_____年_____月_____日

版权声明

摘要

图像语义分割被广泛应用于自动驾驶、AR/VR 交互、机器人等前沿领域。然而由于图像中拍摄场景多样、拍摄角度广泛、标记类别繁多、环境光线变化大、数据不均衡以及物体之间存在复杂而广泛的上下文关系，图像语义分割是一项具有挑战性地任务。

传统的 RGB 图像语义分割算法由于缺乏三维空间位置信息，在算法层面具有一定的局限性。本文引入深度图像用来帮助 RGB 图像进行语义分割，提出了基于编码器解码器架构的注意力融合网络（AFNet），实现了快速且高精度的室内 RGBD 图像语义分割。此外，为了提高模型的性能，本文提出了基于注意力机制的特征融合模块以及特征解码模块，并在其上构建了 AFNet 网络架构。

本文所提出的方法在主流的数据集上进行了算法评估。实验结果表明，所提出的方法已经获得了与现有先进方法相比更好或相当的 RGBD 图像语义分割结果。

关键词：RGBD 图像，语义分割，注意力机制，特征融合

Indoor RGB-D Semantic Segmentation Based on Attention Mechanism

Benzhang Qiu (Information and Communication Engineering)
Directed by Zhun Fan

ABSTRACT

Image semantic segmentation is widely used in automatic driving, AR / VR interaction, robot and other cutting-edge fields. However, image semantic segmentation is a challenging task because of the variety of scenes, the wide range of shooting angles, the various types of labels, the large changes of environmental light, the imbalance of data and the complex and extensive context between objects.

Traditional RGB image semantic segmentation algorithm has some limitations in algorithm level due to the lack of three-dimensional spatial location information. In this paper, depth image is introduced to help RGB image semantic segmentation, and AFNET based on encoder decoder architecture is proposed to achieve fast and high-precision indoor RGBD image semantic segmentation. In addition, in order to improve the performance of the model, this paper proposes feature fusion module and feature decoding module based on attention mechanism, and constructs AFNET network architecture on them.

The proposed method is evaluated on the mainstream datasets. Experimental results demonstrate that the proposed method has achieved better or equivalent semantic segmentation results compared with the state-of-the-art methods.

KEY WORDS: RGBD Images, Semantic Segmentation, Attention Mechanism, Feature Fusion

目录

第1章 绪论

1.1 研究背景与意义

计算机视觉中的大多数任务都涉及到图像的捕获和分析，以做出决策或执行一些相关任务，如自动监视、机器人应用、自动驾驶、内容驱动检索、目标检测、图像推理、实例分割等[40] 。基于语义分割的场景理解是计算机视觉的主要任务之一，语义分割的目标是为输入图像的每一个像素指定一个标签（也称为密集标签或逐像素标签）。传统的 RGB 图像语义分割在实际应用中具有一定的局限性，如图所示，仅给定 2D 图像，位于桌子上的红点会和微波炉上的蓝点非常接近。然而在 3D 点云图像中没有这样的混淆，桌子上的红点和微波炉上的蓝点在 3D 点云中是非常遥远的。

图 1-1 二维图像与三维点云图像示例[52]

随着深度传感器的普及以及商用化（如微软的 Kinect[51] 、Intel 的 RealSense[53] ），使得 RGBD 语义分割称为可能。与常见的 RGB 图像语义分割相比，RGBD 语义分割可以通过挖掘深度信息来利用真实的几何信息。常见的利用深度信息的方式分两种，一种是将 3D 点云图像作为输入，直接进行三维的语义分割（如 PointNet[54] ）；另一种方式是将深度图像当作一种输入图片，用神经网络分别对 RGB 图像和深度图像进行特征提取，然后将分割结果重新投影回 3D 点云中。相较于后一种方法，前一种方法的 3D 点云非常稀疏，很难有效地学习到想要的特征，而且三维卷积在计算成本上开销过大，难以扩展到大量点云数据的处理，现阶段也几乎满足不了实时性的要求。所以后续论文会主要讨论以 RGB 图像和深度图像作为输入时，神经网络的构建。

语义分割方法主要可分为室内场景和室外场景。由于目前深度传感器在室外精度不佳以及深度捕获距离的限制，目前大部分数据集都是针对室内场景，而且室内场景

可以提供更为重要的独立照明特征，后续论文所讨论的应用场景也只针对室内场景。

1.2 研究难点与挑战

如图所示，该图为室内 RGBD 图像数据集 NYU-V2 的示例图像，其他数据集与之类似，存在如下难点：

（1） 拍摄的场景复杂，拍摄角度广，标记类别多，相同类别之间的差异也很大；

（2） 拍摄光线变化较大，有亮光场景、暗光场景以及如图（c）所示的光线变化场景；

（3） 如图（a）所示，标签物体之间存在相互遮挡，各物体存在复杂而广泛的上下文关系；

（4） 如图所示，数据集的各类别之间存在严重的数据不均衡现象；

（5） 如图（a）（f）所示，数据集中存在很多小物体以及难识别的物体（如镜子：镜子中映射的是其他物体）；

（6） 数据中存在一些漏标记和错标记情况，如图（b）中存在很多“衣服”类漏标，图（c）中书柜中的“书”类没有标记，图（d）中书柜中的“书”类有标记，这些具有歧义的标记都会给分割带来很多难度，需要神经网络拥有很强的鲁棒性。

1.3 主要研究内容

本文以室内 RGBD 图像语义分割为核心任务，以深度学习以及神经网络为主要技术手段，利用深度卷积神经网络的强大图像表征能力，构建基于注意力机制的 RGBD 图像语义分割算法，实现高精度的室内 RGBD 图像语义分割。

具体研究内容如下：

（1） 研究基于编码器-解码器架构的全卷积神经网络，实现端到端的 RGBD 图像语义分割。

（2） 研究 RGB 图像与深度图像的融合方法，实现不同特征向量的有效融合；

（3） 研究编码器-解码器之间的跳连结构，改善解码器对编码器空间信息的利用；

（4） 研究金字塔层级输出，实现多尺度输出，优化分割精度；

（5） 针对类别不均衡现象，研究特定的损失函数，改善分割结果。

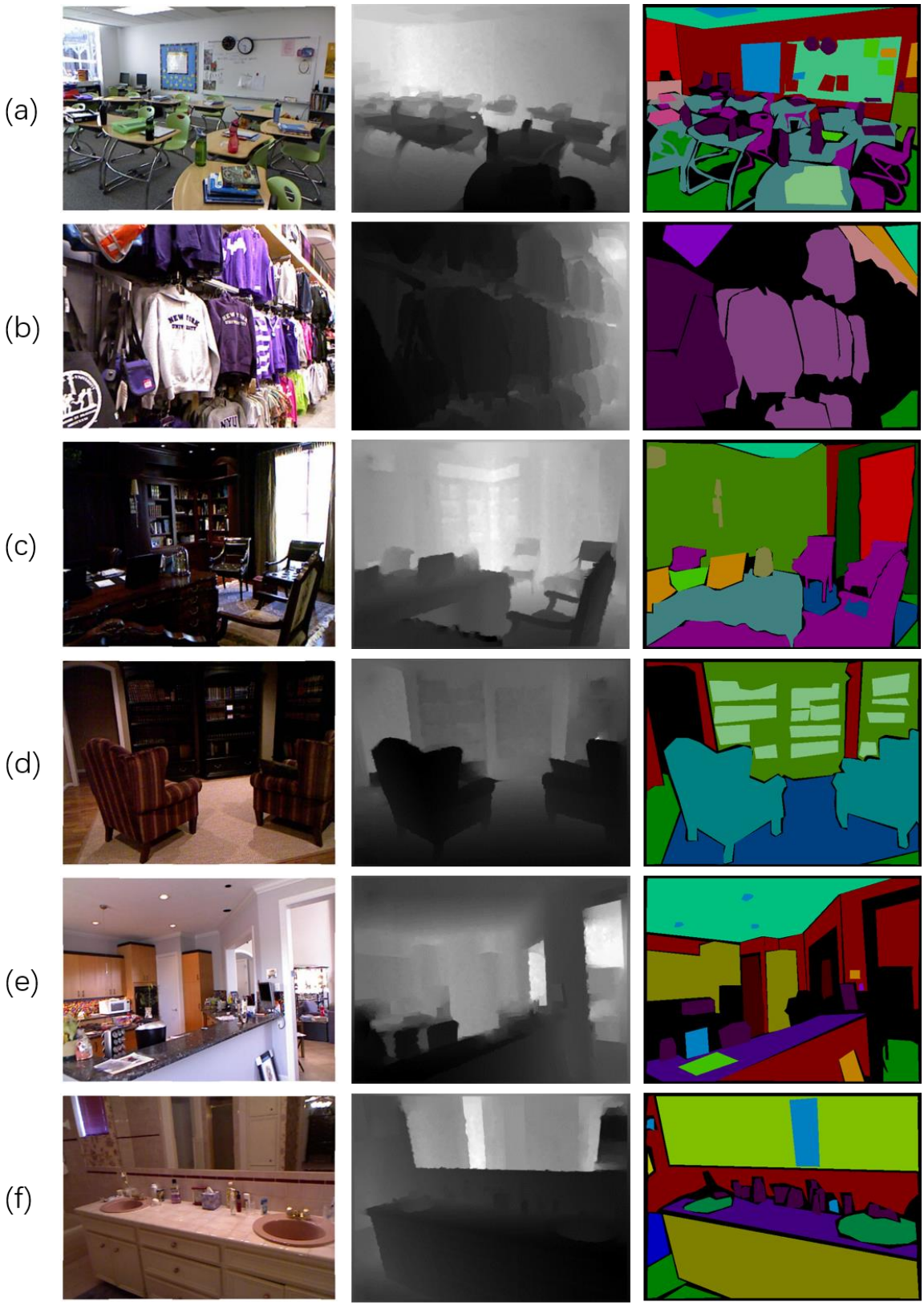


图 1-2 NYU-V2 数据集示例[55]

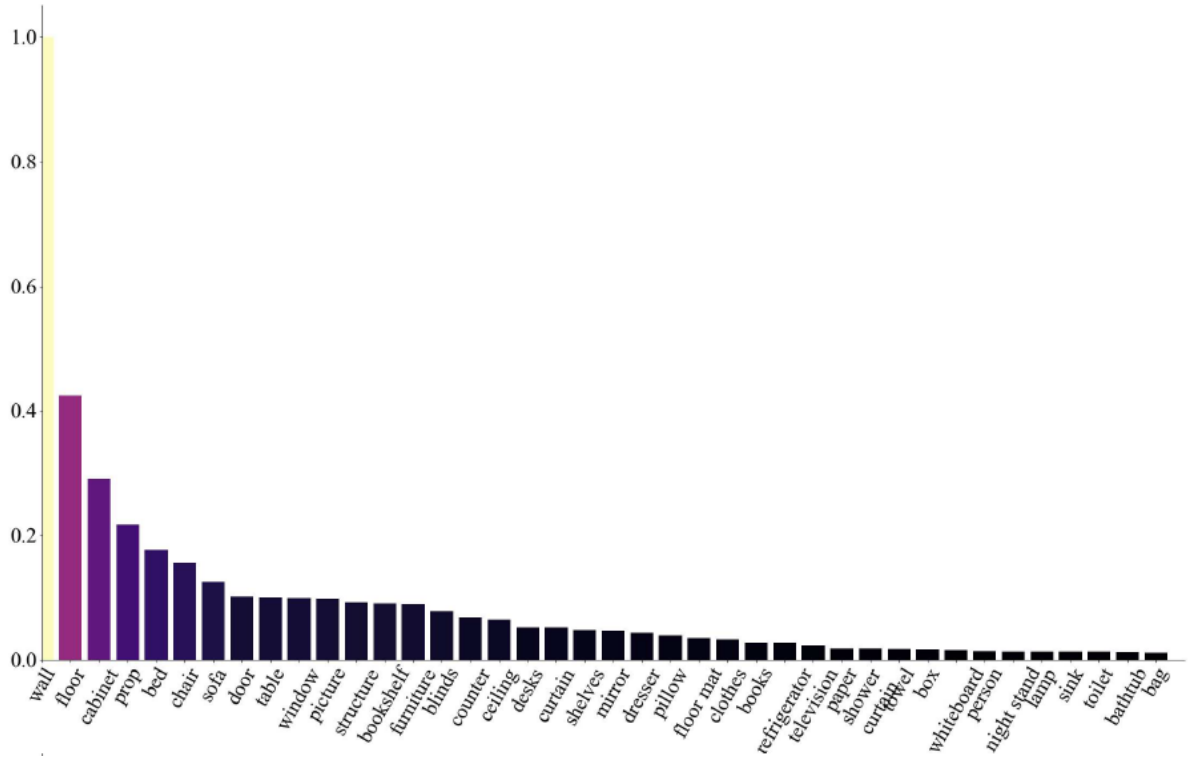


图 1-3 NYU 数据类别数量分布图[56]

1.4 论文章节安排

本文共包含五个章节，各章节具体内容如下：

第 1 章：绪论。本章节主要介绍室内 RGBD 图像语义分割的研究背景及意义，介绍其研究难点和挑战，并确立论文的主要研究内容和论文的章节安排。

第 2 章：相关工作。本章节主要介绍卷积神经网络、RGB 图像语义分割、RGBD 图像语义分割、注意力机制的基础技术以及前沿的相关算法。

第 3 章：基于注意力机制的 RGBD 图像语义分割算法。本章节主要阐述本论文提出的基于编码器-解码器架构的 RGBD 图像分割算法，以及各个子模块的设计。针对类别不均衡现象，对损失函数的设计。

第 4 章：实验结果与分析。本章节主要阐述实验所用的数据库、评价指标，以及证明上述算法设计的有效性的实验设计，记录并分析实验结果。

第 5 章：总结与展望。本章节主要对全文的研究内容进行总结，阐述本论文主要贡献及创新点，并对未来工作的展望。

第2章 相关工作

2.1 卷积神经网络

近些年来，卷积神经网络（CNN）广泛应用于图像处理、计算机视觉、自然语言处理、跨模态学习等领域[1] 。与传统方法相比，卷积神经网络不需要针对特定任务设计复杂的特征提取算子，极大地提高了视觉任务的准确率。

卷积神经网络首先于 20 世纪 60 年代被提出，在对猫的视皮层细胞的实验中，Hubel 和 Wiesel[2] 首次发现视皮层神经元对图像边缘信息敏感，进而提出了“感受野（Receptive Field）”的概念。他们进一步发现了视觉皮层通路中的信息层次处理机制——简单细胞检测位置信息，复杂的细胞整合由简单细胞提供的信息。后来，“感受野”的概念被引入到卷积神经网络的工作中，Fushima 和 Miyake[3] 基于“感受野”提出了“Neocognitron”，这可以看作是卷积神经网络的第一次实现。“Neocognitron”将一个视觉模型分解成若干个子模型，然后在层次化、渐进化的特征平面上对其进行处理，从而在物体发生位移或轻微变形的情况下也能完成识别。“Neocognitron”是第一个基于神经元之间的局部连通性和层次结构而设计出来的人工神经网络。但由于当时缺乏合适的学习算法，网络采用了其他无监督算法，主要应用于手写数字识别。

之后，研究人员尝试用多层感知器来学习特征，而不是人工设计的特征，并用 Paul[4] 首先提出的反向传播（BP）算法来训练模型。通过 Rumelhart[5] 等人的工作，反向传播算法得到了广泛的关注。Lecun[6] 等人提出了 BP 网络在手写数字识别中的应用，这表明大型的 BP 网络可以应用于实际的图像识别问题，而不需要大而复杂的预处理阶段，以及详细的工程。Lecun[7] 等人总结了模块化系统的端到端训练的原理，提出了一种卷积神经网络结构——“LeNet-5”，在当时的标准手写数字识别任务中表现出了优于其他所有技术的性能。

2.1.1 卷积神经网络基本组成单元

基础的卷积神经网络结构如图所示，包含卷积层、池化层和全连接层。针对于其他任务（如目标检测、语义分割等），网络主体结构会有所变化，后续章节会予以介绍，该主体框架主要应用于图像分类任务。

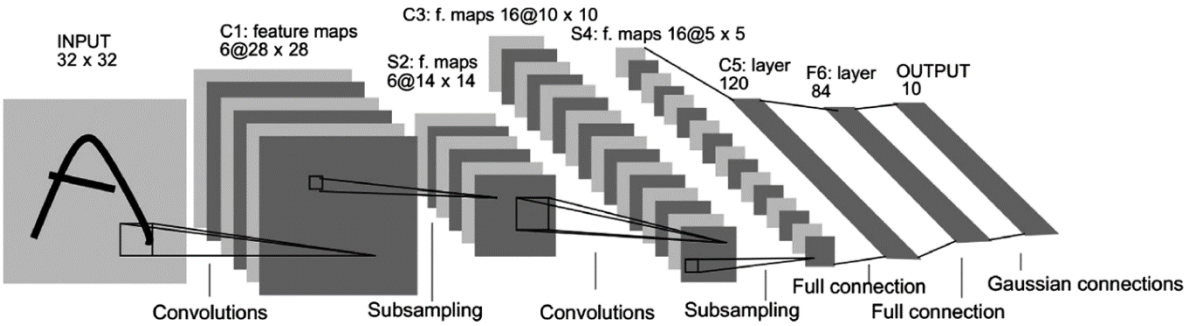


图 2-1 LeNet-5 网络结构图[7]

2.1.1.1 卷积层

卷积层由多个特征映射组成，这些特征映射由卷积核和输入特征图卷积而成。每个卷积核是一个权重矩阵，对于单通道的二维特征图，它可以是3 × 3或5 × 5矩阵。图展示了二维卷积的一个例子。

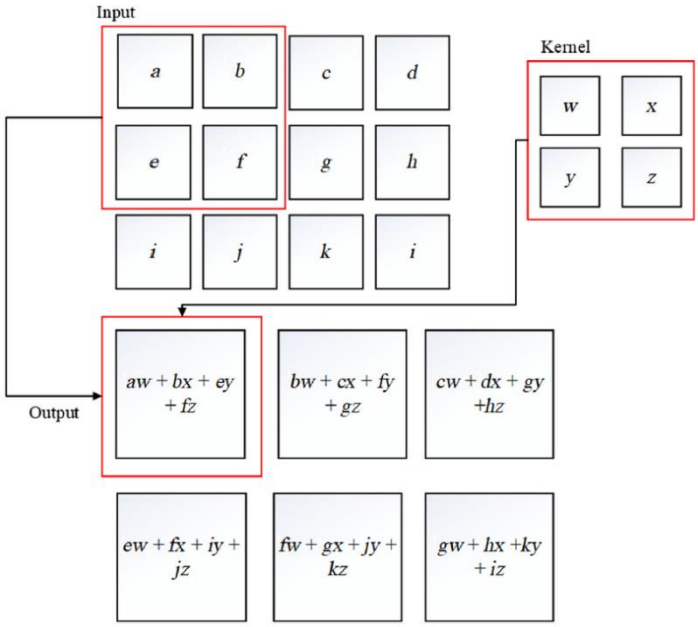


图 2-2 二维卷积例子[1]

卷积运算提供了一种利用卷积核处理可变大小输入的方法，通过卷积运算在卷积层提供不同的输入特征。第一层提取较低级别的特征，如边、端点和角点。然后高层通过对底层特征的处理，提取更复杂、更高层的特征。卷积层主要具有稀疏连接和权重共享的特点。

（1） 稀疏连接：传统的神经网络使用矩阵乘法来建立输入与输出之间的联系，每个输出单元与每个输入单元相互作用。当输入图像包含数千像素时，这种连接将增加模型的存储要求，并增加计算量。与传统的连接方式不同，卷积网络具有稀疏连接的特点，通过控制卷积核的大小远小于输入特征图的大小来实现。稀疏连接的图解说明如图所示，相邻层的神经元之间使用

的是局部连接的方式，m 层中的每个神经元的“感受野”的宽度为 3，接受来自前一层 3 个神经元的输入，但是对于超出自己“感受野”的神经元就不会产生连接。卷积层的稀疏连接不仅降低了模型的存储要求，而且获得输出所需的计算量更少，从而提高了模型的效率。

（2） 权值共享：卷积层还具有权值共享的特点，由卷积核实现。卷积核用于控制参数的数目，并施加空间限制的权值来处理可变大小的输入。权值共享意味着层中的卷积核单元使用相同的权重和偏值（bias）。例如，LeNet-5 的 C1 层是通过计算 6 个卷积核得到的卷积层，每个卷积核在与前一层卷积时具有固定的权重。当输入为单通道信号时，C1 层包含 6 个卷积核，卷积核的大小为$1 \times 5 \times 5$。如果考虑到偏值，C1 层包含共有$63 \times 5 \times 5 + 6 = 156$个参数。相较于全连接网络结构，权值共享在很大程度上降低了网络训练参数，有效地防止了大量参数引起的网络过度拟合，提高了网络运行效率。

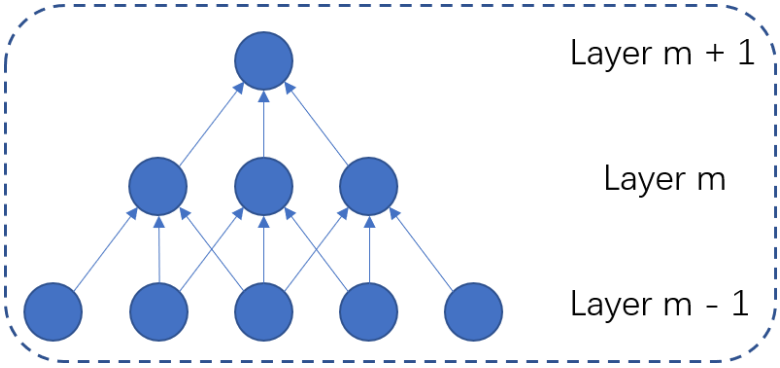


图 2-3 稀疏连接示意图

2.1.1.2 池化层

通常在卷积层之间周期性地插入一个池化层（Pooling Layer），其功能是逐渐减少数据的空间大小，从而减少网络中参数的数量，减少计算资源的消耗。常见的池化方法由最大池化（Max Pooling）和平均池化（Average Pooling）两种，它们分别采用局部感受野中的最大值和平均值作为输出。池化层还可以学习输入的一些不变特征，在采用最大池化时，池化单元只对周围的最大值敏感，而对确切的位置不敏感。因此，通过将获得的特征集合起来，网络可以学习到输入的一些不变特征。在 LeNet-5 中，最大池化层主要使用大小为2×2、步长为 2 的窗口进行卷积，此窗口中的最大值将作为输出结果。

2.1.1.3 全连接层

经过一系列的卷积和池化操作，提取出图像的特征映射，将特征映射中的所有神经元转化为一个完全连通的层，最后，输入可以按 softmax 层进行分类。全连接层的

功能是将卷积层和池化层中的局部信息与类别区分相结合，从而提高整个 CNN 的性能。

LeNet-5 是经典的 CNN 架构。虽然全连接层往往在网络中占较大的内存开销，后续很多网络中也逐渐取消了全连接层，卷积层、池化层和全连接层的结合仍然是现代深度卷积神经网络的基本组成部分。LeNet-5 对深度卷积神经网络的发展具有开创性的意义。

2.1.2 常见卷积神经网络结构

2.1.2.1 AlexNet

由于硬件计算和数据不足，LeNet-5 提出后并没有引起足够的注意。随着计算机硬件的发展和可用于神经网络训练的数据量的增加，2012 年，AlexNet[8] 以远低于第二名的错误率赢得了 ILSVRC-2012 图像分类竞赛。从那时起，深度神经网络开始引起了广泛关注。AlexNet 网络结构如图所示与 LeNet-5 相比，AlexNet 网络框架的改进如下：

（1） ReLU 激活函数：ReLU 可以将非线性和稀疏性引入网络。稀疏性可以选择性地或以分布式的方式激活神经元。它可以学习相对稀疏的特征，实现自动分离。

（2） 数据增广：AlexNet 视同保留标签的图像变换来人工增大数据集。数据增广的方式包括图像平移、图像翻转、改变图像中 RGB 通道的值。

（3） Dropout：神经元可以根据一定的概率从网络中丢弃，以减少网络模型参数，防止过拟合。

（4） 在两台 NVIDIA GTX580 3GB GPU 上进行训练，随着 GPU 并行计算能力的提高，这种方法加快了网络训练的速度。

（5） 局部响应标准化（Local Response Normalization）：在局部神经元的活动中建立竞争机制，使响应较大的值相对变得更大，同时抑制其他反馈相对较小的神经元，增强了模型的泛化能力。

（6） 重叠池化（Overlapping Pooling）：AlexNet 中提出让池化核的步长小于池化核的大小，这样池化层的输出之间就会有覆盖和重叠，提升了特征的丰富性。

AlexNet 是深度卷积神经网络发展的里程碑，引发了新一轮的神经网络研究浪潮，AlexNet 的成功主要取决于计算机硬件的发展和数据集的增强。

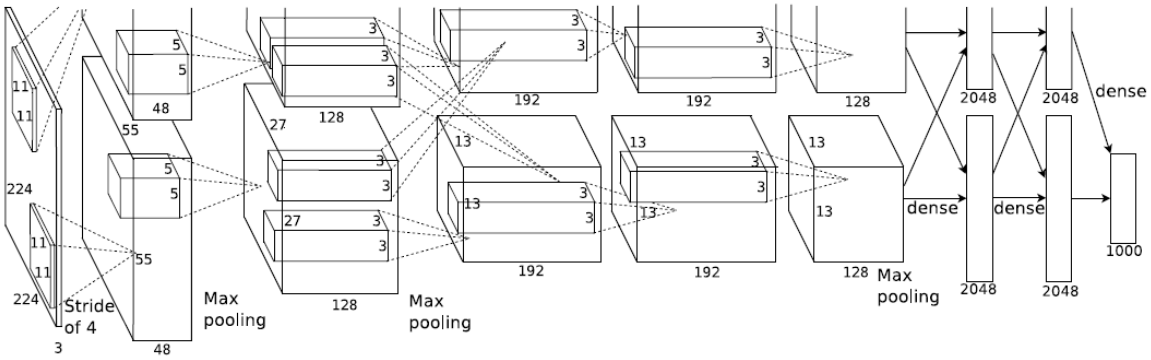


图 2-4 AlexNet 网络结构[8]

ConvNet configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
Input (224×224 RGB image)					
conv3-64	conv3-64	conv3-64	conv3-64	conv3-64	conv3-64
	LRN	**conv3-64**	conv3-64	conv3-64	conv3-64
Maxpool					
conv3-128	conv3-128	conv3-128	conv3-128	conv3-128	conv3-128
		conv3-128	conv3-128	conv3-128	conv3-128
Maxpool					
conv3-256	conv3-256	conv3-256	conv3-256	conv3-256	conv3-256
conv3-256	conv3-256	conv3-256	conv3-256	conv3-256	conv3-256
			conv1-256	**conv3-256**	conv3-256
					conv3-256
Maxpool					
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
			conv1-512	**conv3-512**	conv3-512
					conv3-512
Maxpool					
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
conv3-512	conv3-512	conv3-512	conv3-512	conv3-512	conv3-512
			conv3-512	**conv3-512**	conv3-512
					conv3-512
Maxpool					
FC-4096					
FC-4096					
FC-1000					
Softmax					

图 2-5 VGGNet 网络配置[9]

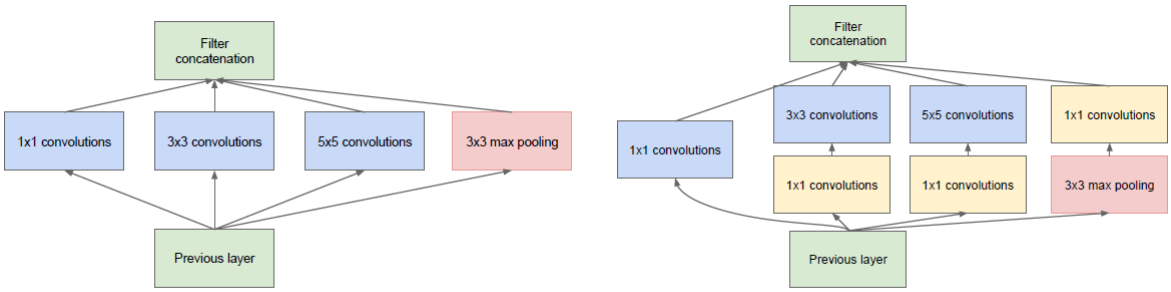
2.1.2.2 VGGNet

浅层卷积神经网络模型在大规模图像识别任务中具有一定的局限性，为了进一步探讨更深层的网络模型的性能，Simonyan 和 Zisserman[9] 提出了 VGGNet。VGGNet 的主要贡献是使用非常小的(3×3)的卷积核的架构来增加网络的深度，通过将深度推到 16 到 19 个权重层，可以实现对现有结果的显著提升。VGGNet 提出了六种不同的网络配置，具体的配置信息如**错误!未找到引用源。**所示。

与 AlexNet 不同的是，VGG 在整个网络的构造过程中使用3×3的小卷积核，并通过叠加3×3的小卷积核来加深网络。在实验中，为了保持各特征层组成构造的计算复杂度大致一致，当特征映射的大小通过最大池化层减少一半时，下一层卷积核的数目增加一倍。**错误!未找到引用源。**中的各种配置几乎具有相同数量的参数，实验结果表明，VGG-19 模型取得了较好的效果，这也证实了增加网络深度有利于提高图像分类的精度。

2.1.2.3 Inception 系列

VGG 通过构造小卷积核建立了更深层次的网络模型，Inception V1[11] 受到 Network in Network（NiN）[10] 的启发，拓宽了网络结构，巧妙地提出了 Inception 模块。Inception V1 带有 Inception 模块的网络使得模型能够更好地描述输入数据的内容，同时进一步增加了网络模型的深度和宽度。初始模块自提出以来一直都在不断更新和改进，初始模块的不同版本如图 2-6 Inception V1 模块结构[11] 所示。



(a) Inception module, naïve version　　(b) Inception module with dimension reductions

图 2-6 Inception V1 模块结构[11]

通过上一小节的介绍，我们可以知道，卷积层的一个功能时通过使用卷积层中的通道（卷积核）数目来减少和增大维度。在 Inception V1 中，主要通过1×1卷积核来降低维度，从而减少网络参数和特征映射的数量。输入特征映射用1×1卷积核卷积，该操作相当于在大小不变的情况下，对原始图像进行尺度变换，可以大大提高图像分类的精度。Inception V1 还使用了3×3和5×5的卷积，拓宽网络的同时也提高了网络对输入图像尺度变换的适应性。

Inception V2[12] 采用 Batch Normalization（BN）[13] 来规范化每一层的参数分布，

增强了网络的鲁棒性，同时加速了网络的收敛。还采用了 VGGNet 的两个3×3卷积核替换掉5×5卷积核，减少网络的参数，增加网络的深度。Inception V3[12] 使用1×3和3×1的卷积核替换掉3×3的卷积核，进一步减少了参数量。Inception V4[14] 则结合了残差模块（Residual Block），进一步加深了网络，提高了模型的精度。关于残差网络，将会在下一小节进行介绍。

2.1.2.4 ResNet

从以上各种卷积神经网络的发展可以看出，增加神经网络深度和宽度可以提高网络性能，但一味的增加容易引起梯度消失、梯度爆炸现象。He[15] 提出，梯度消失的问题在很大程度上已通过标准化初始值和标准化中间层得到解决。但也会出现另外一个问题——退化问题，即当网络层数增加时，训练集的精度反而下降，这不能被解释为过拟合，因为过拟合在训练集中应该表现更好。退化问题表明，深层网络不能很好地进行优化。ResNet[15] 的提出就是为了解决以上问题，ResNet 由残差学习模块可以突破 100 层甚至达到 1000 层。

ResNet 主要由残差模块构成，如图（左）所示，在残差模块中，假设原始网络的能够学习到的表达为$H(x)$，残差学习的表达则为$F(x) = H(x) - x$。这两个表达式具有相同的效果，但优化的难度不同，在极端情况下，如果一个恒等映射是最优的，那么将残差优化到零，比通过一堆非线性层来拟合到一个恒等映射要容易得多。恒等映射的加法连接不向网络中添加额外的参数和计算，但可以大大提高模型的训练速度，并改进训练效果。

ResNet 网络结构中主要使用两个残差模块，一个是将大小为3×3的两个卷积核串联为图（中）所示的一个残差模块，另一个是将1×1、3×3和1×1卷积核连接在一起作为图（右）所示的“Bottleneck”模块。在“Bottleneck”模块中，第一个1×1卷积核主要将特征映射的维度从 256 维降到 64 维，然后用3×3卷积核进行计算，最后用1×1的卷积核将数据维度改到 256。

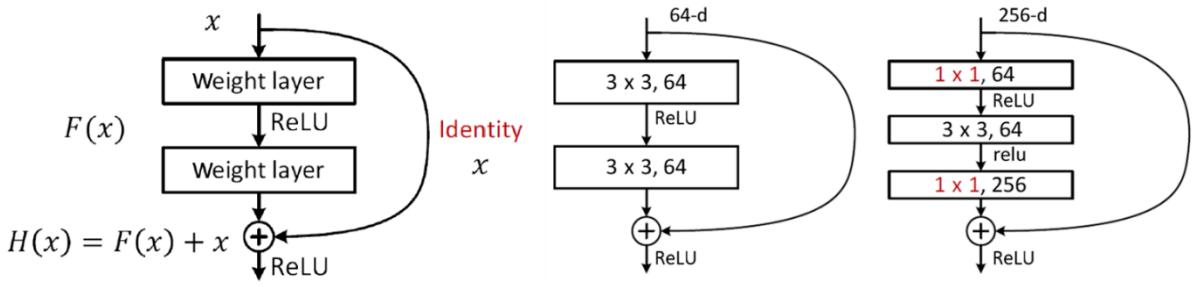


图 2-7 残差模块[15]

2.2 RGB 图像语义分割

RGB 图像分割是许多视觉系统的重要组成部分[16] 。它包括将图像（或视频帧）

分割成多个片段或对象[17]。分割在广泛的应用中起着核心作用[18]，包括医学图像分析（例如：肿瘤边缘提取、组织体积测量、眼底血管检测等）、智能交通（例如：地面导航和行人检测）、视频监控和增强现实等。过去的研究人员已经开发了许多图像分割算法，从最早的阈值化[19]、基于直方图的区域生长[20]、K-means 聚类[21]、分水岭[22]等算法，到更先进的活动轮廓[23]、抠图[24]、条件场和马尔科夫随机场[25]以及基于稀疏性[26]的算法。但在过去的几年里，深度学习网络产生了新一代性能卓越的图像分割模型，其性能得到了显著的提高——通常在流行的基准上达到了最高的准确率，这导致了许多人认为是该领域的范式转变。

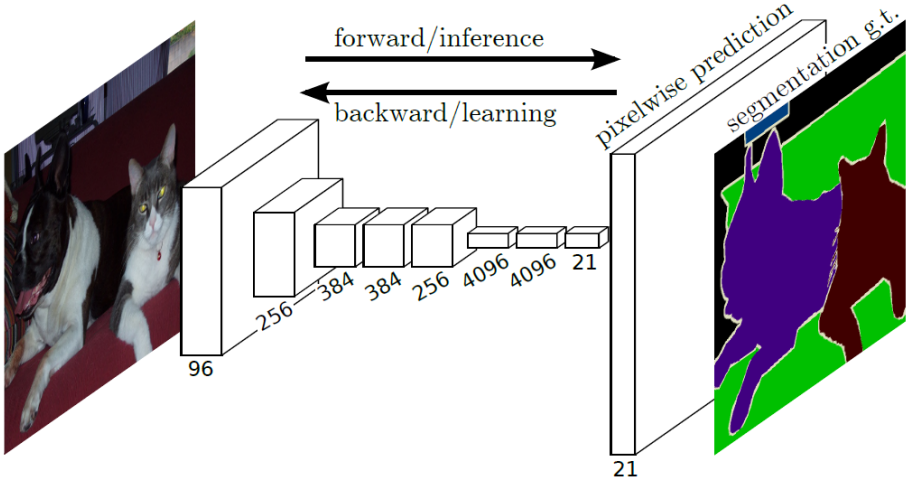


图 2-8 全卷积神经网络（FCN）框架[27]

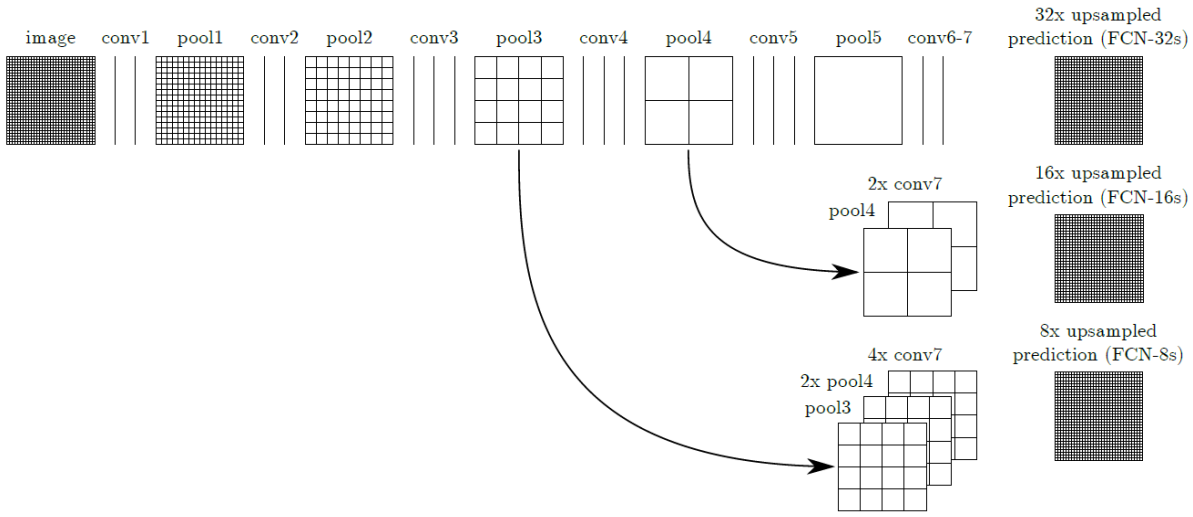


图 2-9 跳连结构将粗糙的高层信息和底层的精细信息相结合[27]

2.2.1 全卷积神经网络

Long[27]等人提出了一种利用全卷积网络（Fully Convolutional Networks）进行图像语义分割的深度学习方法。全卷积神经网络（如

图所示）只包括卷积层，这使得它能够获取任意大小的图像并生成相同大小的分

割图。作者修改了现有的 CNN 框架（如 VGGNet），以管理非固定大小的输入和输出，将所有全连接层替换成全卷积层。结果，该模型输出的是空间分割图而不是分类分数。

如图所示，通过跳连结构，将来自模型最后一层的特征映射被向上采用并与前层的特征映射相融合，该模型将语义信息（来自深层、粗糙层）和表面信息（来自浅层、精细层）相结合，以产生准确和精细的分割。

这项工作被认为是图像分割中的一个里程碑，证明了可以在可变大小的图像上以端到端的方式训练深层网络进行语义分割。然而，传统的 FCN 模型虽然具有普遍性和有效性，但也存在一定的局限性：实时推理速度不够快，不能有效地考虑全局上下文信息，不易移植到三维图像中。

2.2.2 基于编码器解码器的模型

另一个流行的用于图像分割的深度网络模型家族的是基于卷积编码器-解码器体系构成的。大多数基于深度网络的分割工作都是使用某种编码-解码模型。

Noh 等人[28] 发表了一篇关于基于反卷积（也称转置卷积）的论文。该模型（如图所示）由两部分组成，一部分是使用 VGG-16 网络中的卷积层的编码器，另一个部分是以特征向量为输入并生成像素级概率图的反卷积网络。反卷积网络由反卷积层和反池化层组成，用于识别像素级标签和预测分割掩模。

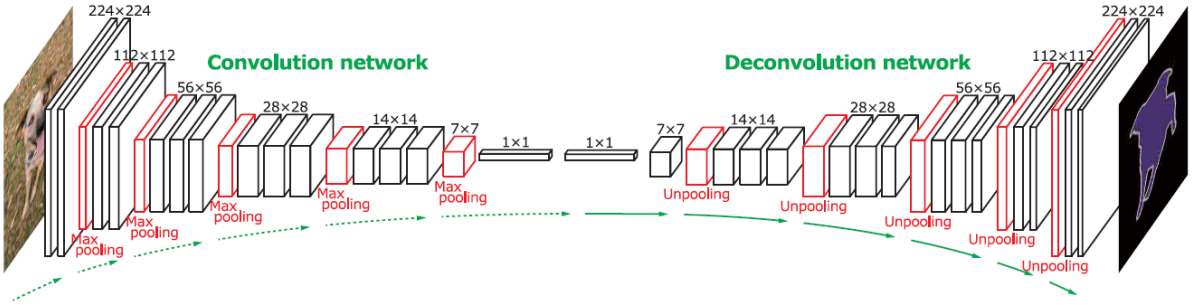


图 2-10 反卷积语义分割网络模型[28]

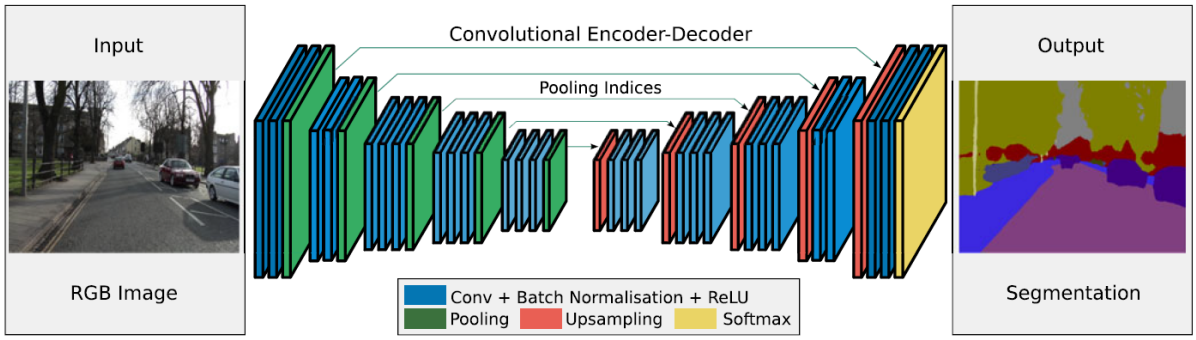


图 2-11 SegNet 网络结构[29]

在另一个著名的网络 SegNet[29] 中，作者提出了一种用于图像分割的卷积编码器

架构（如图所示）。与反卷积网络类似，SegNet 的核心可训练分割引擎有一个与 VGG-16 网络中的 13 个卷积层拓扑结构相同的编码网络和一个对应的解码网络和一个像素级分类层组成。该模型主要创新之处在于解码器对其低分辨率输入特征映射进行上采样，具体地说，它使用在相应编码器的最大池化步骤中计算的池化位置索引来执行非线性来进行上采样。这消除了学习向上采样的必要性。然后用可训练滤波器卷积（稀疏）上采样的映射，以产生密集的特征映射。SegNet 在可训练参数的数量上也明显小于其他竞争架构。同一作者还提出了 SegNet 的 Bayesian 版本，用于建模场景分割的卷积编码器-解码器网络固有的不确定性[30] 。

2.2.3 多尺度以及基于金字塔的模型

多尺度分析（Multi-Scale Analysis）是图像处理中的一个古老的思想，已经被广泛应用于各种神经网络结构中，其中最突出的一种模型是 Lin 等人提出的特征金字塔网络模型（FPN）[31] 。主要用途目标检测，但后来也应用于图像分割。利用深度卷积神经网络固有的多尺度金字塔层次结构构造具有边际额外成本的特征金字塔。为了融合低分辨率和高分辨率特征，FPN 由一个自下而上的结构、一个自上而下的结构和横向连接构成，然后通过3×3卷积处理连接的特征映射，以产生每个阶段的输出。最后，自上而下结构中的每个阶段都生成一个预测结果。对于图像分割，作者使用两个多层感知机（MLPs）来生成掩模。图 2-12 FPN 中特征金字塔构成[31] 显示了横向连接和自上而下路径是如何通过加法合并的。

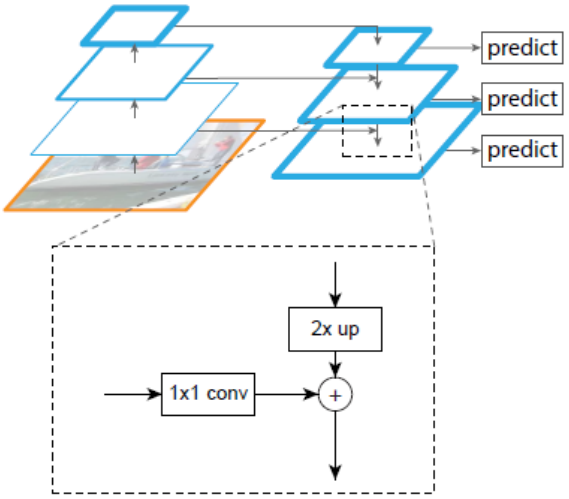


图 2-12 FPN 中特征金字塔构成[31]

Zhao 等人[32] 开发了金字塔场景解析网络（PSPN），这是一个多尺度网络，可以更好地学习场景的全局上下文表示（如图 2-13 所示）。使用残差网络（ResNet）作为特征提取器，通过扩展网络从输入图像中提取不同的模式。然后将这些特征映射输入金字塔模块，以区分不同尺度的模式。他们在四个不同的尺度上集合，每个尺度对应

一个金字塔层，并由1×1卷积层处理以减少它们的维度数。金字塔层的输出被上采样，并与初始特征映射连接，以捕获本地和全局上下文信息。最后，使用卷积层来产生逐像素预测。

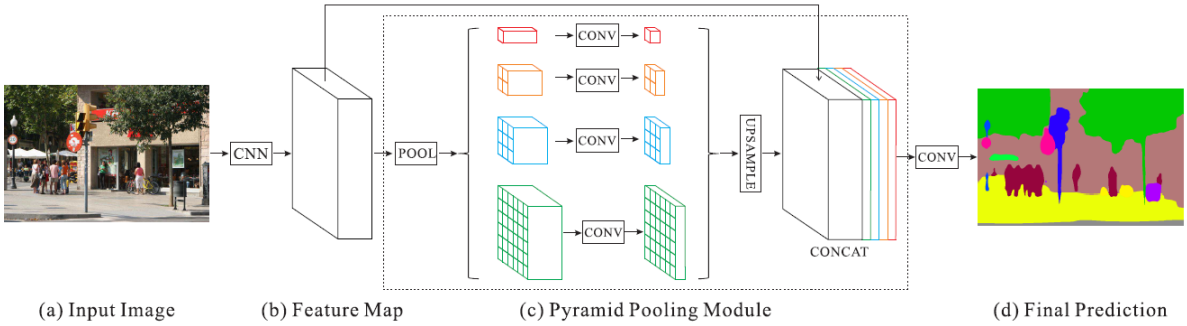


图 2-13 PSPN 网络结构[32]

2.2.4 空洞卷积以及 DeepLab 系列

扩张卷积（Dilated Convolution）为卷积层引入了另一个参数——扩张率。信号$x(i)$的扩张卷积（如图 2-14 所示）被定义为$y_i = \sum_{K=1}^{K} x[i + rk]w[k]$，其中$r$是定义核$w$的权重之间间隔的扩张率。例如，扩张率为 2 的$3 \times 3$的卷积核将具备和$5 \times 5$的卷积核相同大小的接收场。而仅使用 9 个参数，从而在不增加计算成本的情况下扩大接收场。扩张卷积在图像分割领域中得到了广泛的应用，其中一些最重要的包括 DeepLab 系列[33]、多尺度上下文聚合（Multi-scale context aggregation）[34]、密集上采样卷积和混合扩张卷积（Dense upsampling convolution and hybrid dilated convolution）[35]、密集连接的空洞空间金字塔池化（Densely connected atrous spatial pyramid pooling）[36] 和高效神经网络（Efficient neural network）[37]。

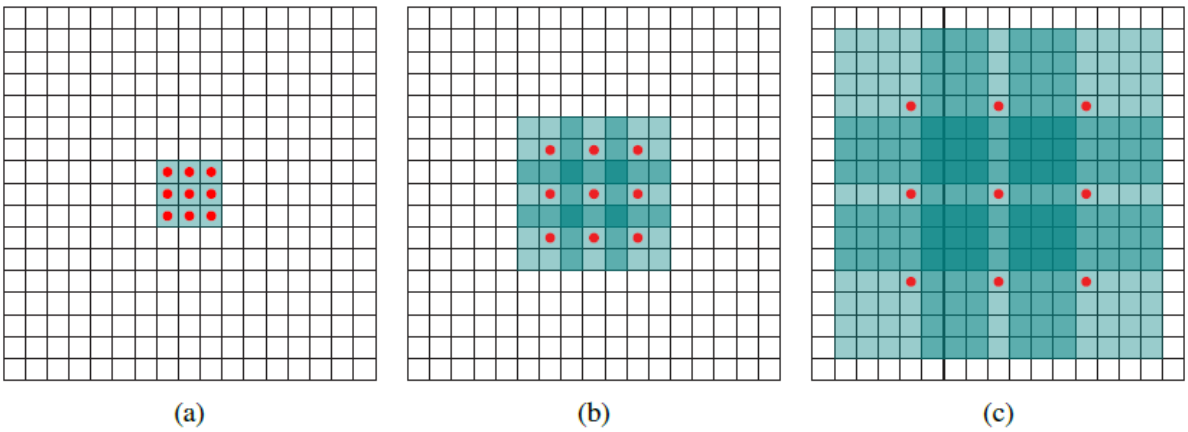


图 2-14 扩张卷积[34]

DeepLab V1 和 DeepLab V2[33] 是最流行的图像分割方法之一，如图所示，该模型主要有三个特点，首先是使用扩张卷积来解决网络中低分辨率降低的问题（由最大池化和跨步引起）；第二是 Atrus 空间金字塔池化（ASPP），它以多采样率使用滤波器

探测传入的卷积特征层，从而在多个尺度上捕获对象和图像上下文，以在多个尺度上可靠地分割对象；第三是结合深度卷积神经网络的概率图模型的方法改进目标边界的定位。

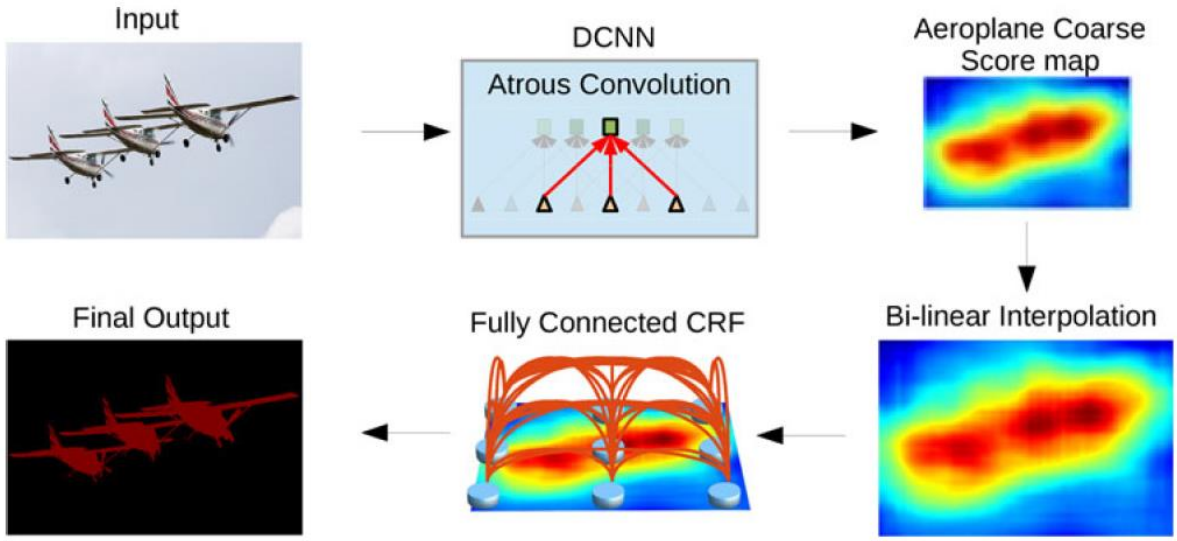


图 2-15 DeepLab 网络结构[33]

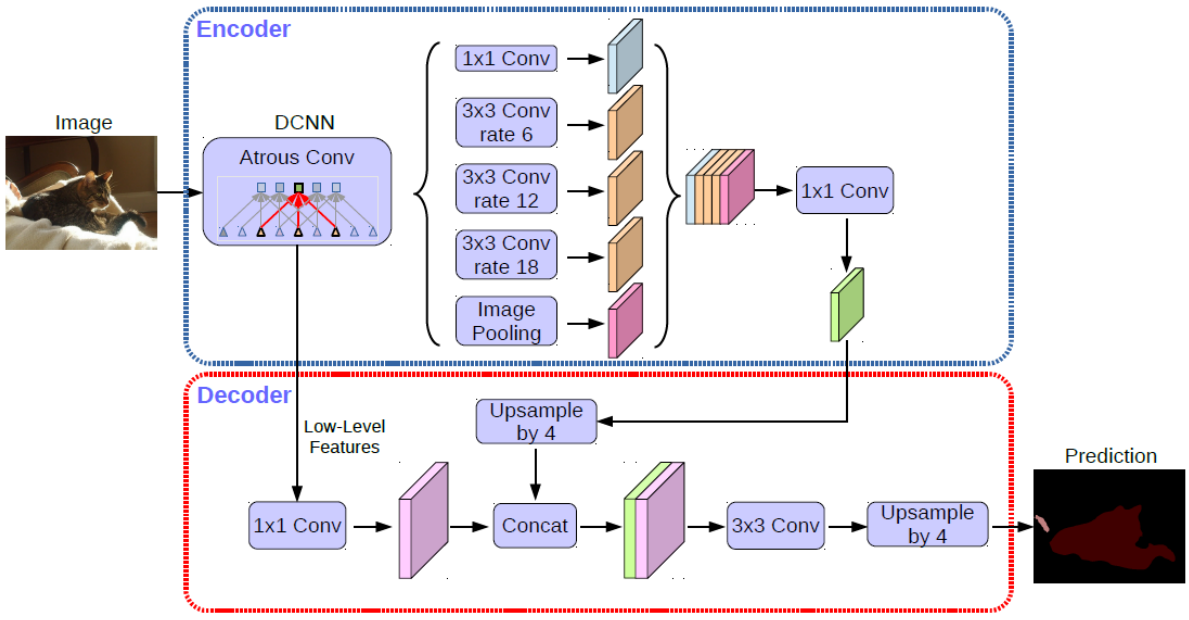


图 2-16 DeepLab V3+网络结构[39]

随后，Chen 等人[38] 提出了 DeepLab V3，它结合了级联和并行的扩张卷积模块，其中并行卷积模块分组在 ASPP 中。在 ASPP 中加入了1×1卷积和 BN，所有的输出被连接起来，并由另一个1×1卷积处理，以创建每个像素的最终输出。

在 2018 年，Chen 等人[39] 提出了 DeepLab V3+，它使用编码器-解码器结构（如图所示），其解码器包括由一个反卷积（输入的每个通道的空间卷积）和点卷积（以反卷积作为输入的1×1卷积）组成的反卷积构成，其编码器使用 DeepLab V3 作为主

框架。

2.3 RGBD 图像语义分割

在上一节我们介绍了 RGB 图像分割技术，随着 Kinect、RealSense 等商用 RGBD 传感器的问世，人们已经一致证明利用从深度信息中提取的特征有助于减少识别对象的不确定性。深度特征可以描述仅在 RGB 特征中可能丢失的三维几何信息。为了从 RGB 和深度数据中提取有用的特征，开发一种有效地融合两种输入数据的方法至关重要。已有很多人尝试用不同的方法利用深度信息进行语义分割。

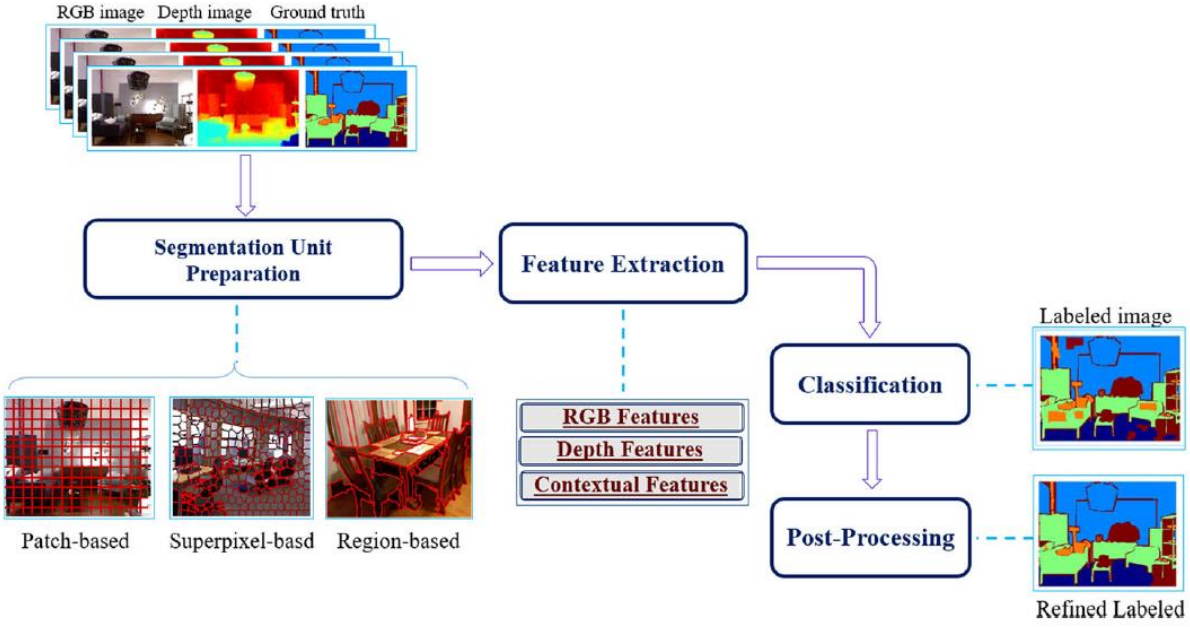


图 2-17 传统的 RGBD 语义分割方法的典型流程图[40]

传统的 RGBD 语义分割方法的典型流程图如图所示，输入包括 RGBD 图像和真值标签。在分割单元准备中，现有的方法利用了分割树（Segmentation Tree）[41] 或其他无监督方法生成的像素、图像块、图像超像素或相干区域。在特征提取考虑到上下文信息的情况下，当分割单元越精细（最精细的是像素级），最后预测的准确率会越高。但是，在这种情况下，时间和计算成本都会增加。而另一方面，从粗粒度区域提取到的特征比从细粒度区域提取到的特征具有更强的识别能力，同时，减少了时间计算成本。因此，如何选择分割单元是一个非常重要的事情。在特征提取过程中，可以提取三种底层特征和高层特征：（1）从 RGB 图像中提取的特征；（2）从深度图像中提取的特征；（3）作为上下文特征的特征。通常使用的 RGB 和深度图像的特征一般有 SIFT、SURF、HOG、Haar 小波、LBP、Lab Color、3D 点云和曲面法线。对于分类器，所有典型的分类器都有被用过，最流行的是随机森林和支持向量机。最后，一些方法应用 CRF 或 MRF 框架来平滑标签或考虑每个分割单元标签之间的上下文关

系。

2.3.1 特征融合方式

当深度卷积神经网络成功在 RGB 图像语义分割取得较大突破后，深度卷积神经网络也被用于 RGBD 图像语义分割，大部分方法还是采用前面提到编码器-解码器架构。RGBD 语义分割面临的主要挑战之一是如何表示和融合 RGB 和深度通道，从而考虑到 RGB 通道和深度通道之间的强相关性，不同的方法在对编码器和解码器使用了不同的策略，这些策略包括早期融合、中期融合和晚期融合（如图 2-18 所示）。早期融合[42]和晚期融合[27]是融合 RGB 通道与深度通道最简单的方式。常见融合手段包括将特征图通道连接（Concatenate）、和通道对应位置叠加（Add）。

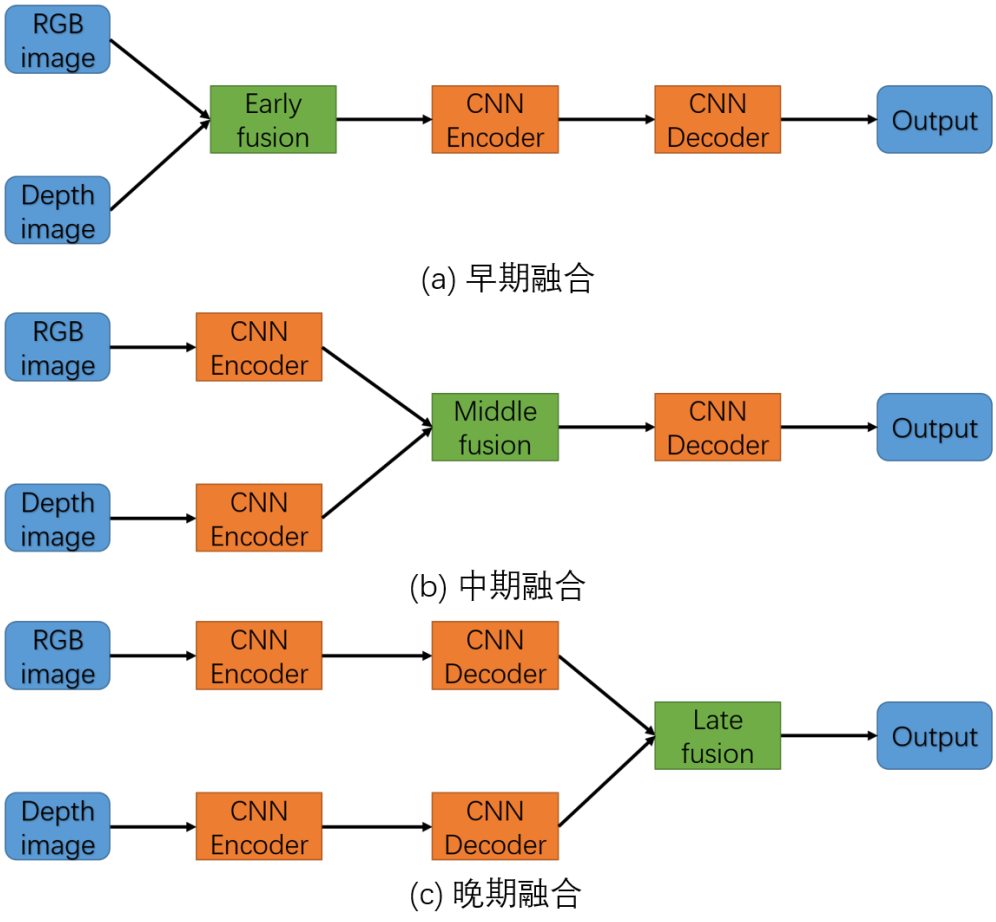


图 2-18 常见 RGBD 特征融合方式

2.3.2 常见 RGBD 图像分割网络

由 Hazirbas 等人提出的 FuseNet[43]（如图所示）实行了深度通道中中间特征映射与 RGB 通道中的中间特征映射元素级求和。他们使用 VGG-16 作为 RGB 和深度通道编码器的两个分支，并将深度编码器分支中的特征映射与 RGB 编码器分支中的特征映射逐层融合，生成一个主编码器流。

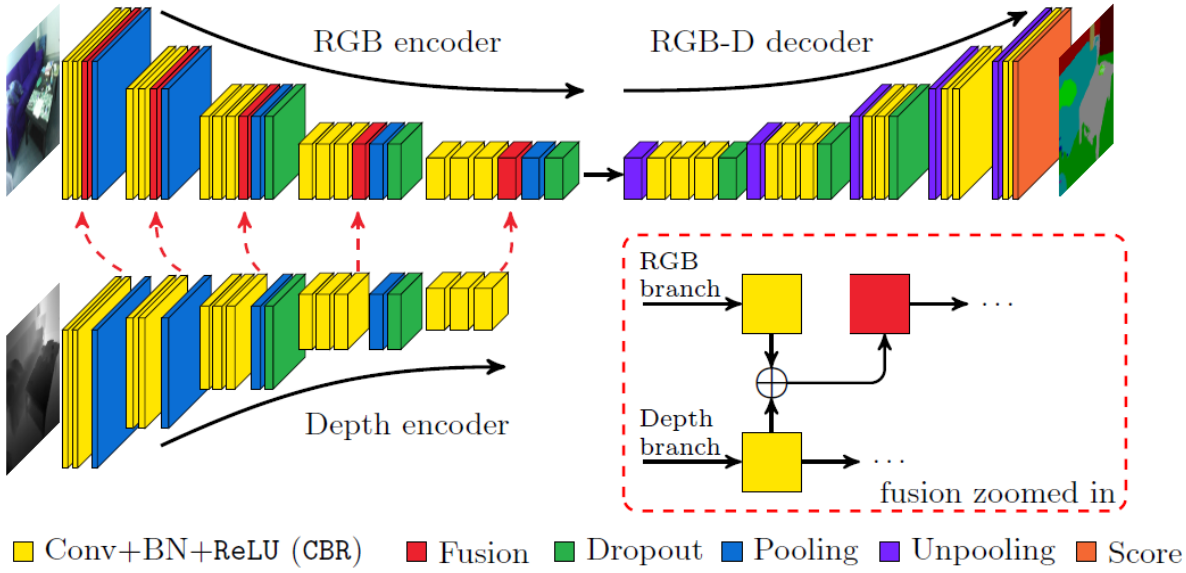


图 2-19 FuseNet 网络结构[43]

Wang 等人[44] 利用反卷积作为解码器，并在中间特征融合部分采用由共同特征和特有特征组成的转换网络来进行学习（如图所示）。通过共同最大化共享信息之间的相似性和特定模态信息之间的差异，网络学会将每个模态的特征分别分解为共同特征和特有特征。并且为了实现鲁棒的预测，网络显式地允许一个模态借鉴其他模式的共同特征，增强共享信息的表现力。

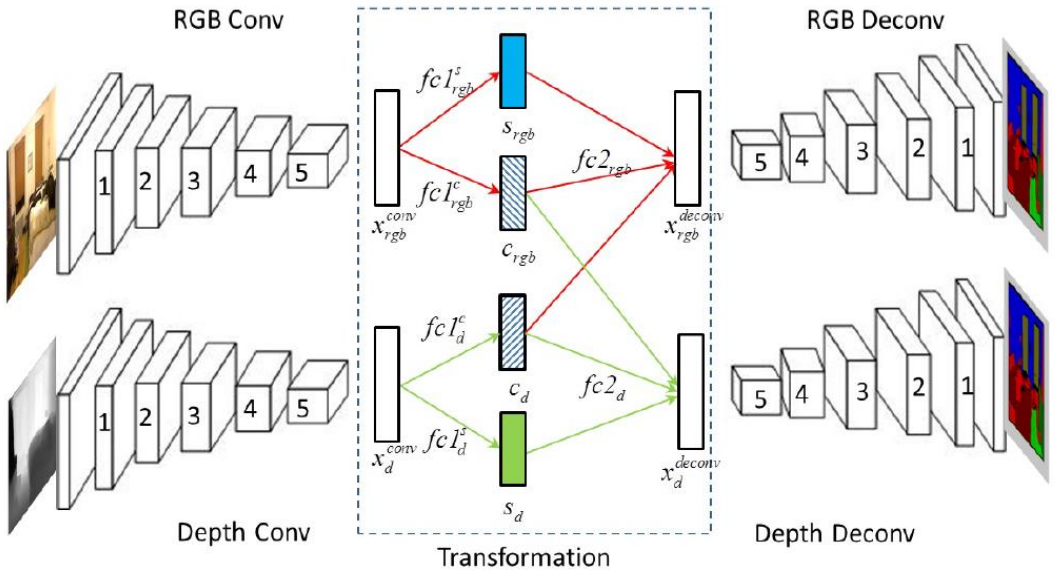


图 2-20 文献[44] 中的网络结构图

在 LSTM-CF[45] 中（如图所示），通过长短时记忆网络（LSTM），从 RGB 和深度通道捕获 2D 全局上下文信息。首先，在对 RGB 和深度通道分别进行卷积后，利用垂直 LSTM 进行信息编码，然后将两个通道的特征进行通道连接，再用水平 LSTM 进行融合，以捕获全局 2D 全局上下文信息。值得一提的是，该模型的特征编码部分

采用的是调整过后的 DeepLab 模型。

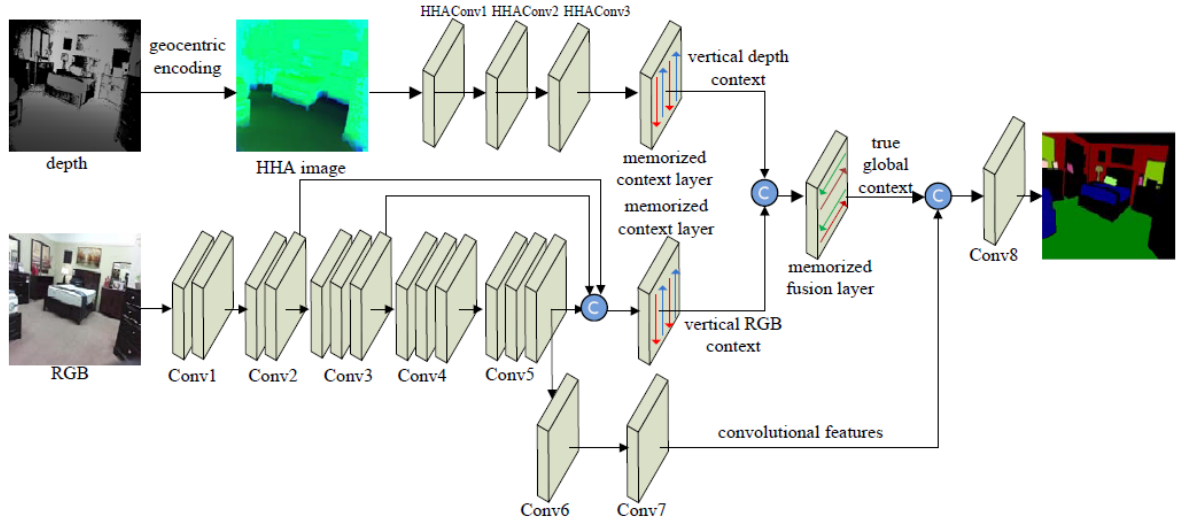
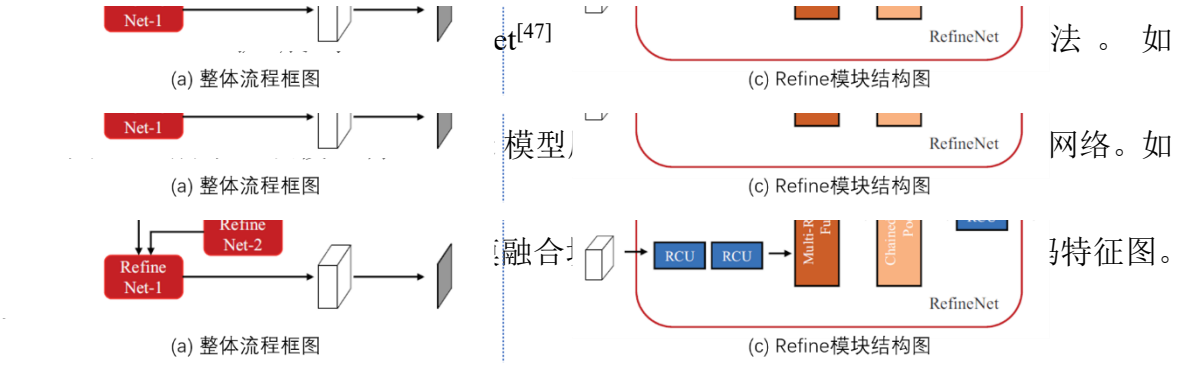


图 2-21 LSTM-CF 网络结构图[45]

et[47] 法。如

模型) 网络。如

融合 特征图。



图（c）所示，RefineNet 中的级联细化块被用来做解码器。

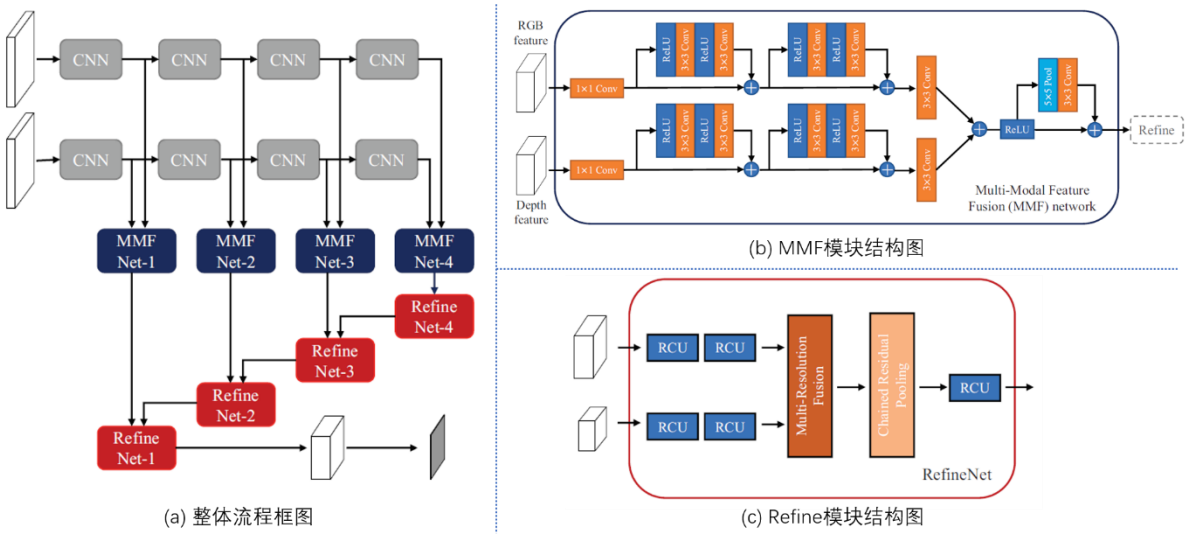


图 2-22 RDFNet 网络结构图[46]

RedNet[48] 编码部分采用和 FuseNet 相同的操作，将深度通道的特征映射逐层叠加在 RGB 通道的特征映射上，并采用跳连结构将编码器与解码器连接，同时采用金

字塔输出提升输出结果。Liu 等人[49] 通过集成 2D 和 3D 信息改进了 HHA 编码。然后，他们扩展了用于 RGBD 语义分割的 VGG 网络，同时提出了 RGB 和深度通道 CNN 模型的加权和，然后用全连通 CRF 来增强预测。LSD-CF[50] 提出了 RGB 和深度通道特征图的后期融合，他们提出了一种基于网络输入的门控制融合方法来学习个模态组合的有效权值。此外，他们研究了相邻 RGBD 像素之间的成对关系，将其作为一个亲和矩阵，以恢复其体系结构中上采样层中更清晰的边界，该亲和矩阵由邻域像素的相似性计算得出。

2.4 注意力机制

2.4.1 注意力机制的起源

在心理学中，收处理瓶颈（Processing bottlenecks）的限制，人类往往会有选择地将注意力集中在一部分信息上，同时忽略其他可感知的信息。上述机制通常称为注意力机制（Attention mechanism）[62] 。例如，在人类的视觉处理中，虽然人的眼睛能够接受到大的视野，但通常只有一小部分被注视，原因是视网膜不同区域有不同程度的处理能力，这种能力通常称为敏锐度。只有视网膜的一小部分（中心凹）有最大的视力。要分配有限的视觉处理资源，首先要选择一个特定的视觉区域，然后再对其进行聚焦。例如，当人类阅读时，通常在特定时刻要阅读的单词会被关注和处理。因此，注意力主要有两个方面：

（1） 决定输入的哪一部分需要重点关注；

（2） 将有限的处理资源分配到重要部分。

心理学中关于注意力机制的定义是非常具象和直观的，因此这一概率被借用并广泛应用于计算机科学领域中[63] 。注意力机制最初被应用于机器翻译[64] 中，现在已经成为神经网络研究中一个重要的概念，现已被广泛应用于自然语言处理（Nature language processing）、统计学习（Statistical learning）、计算机视觉（Computer vision）中。

由于本论文主要讨论 RGBD 语义分割相关工作，后续关于注意力机制的介绍主要针对图像处理领域。

2.4.2 注意力机制的分类

上一小节我们提到注意力机制的核心为聚焦，即关注我们所关心、所需要的部分，在计算机科学领域，我们用加权的方式表征注意力机制，对我们所关注的部分给予较大的权重，对不关注的部分给予较小的权重。就注意力机制的类型来划分，可以划分为硬注意力（Hard attention）和软注意力（Soft attention）。其中，硬注意力就是二值表征，即那些区域是被关注的、那些区域是不被关注的。硬注意力在图像处理中最常

见的应用就是图像裁剪（Image cropping），保留需要的区域。硬注意力在数学上表达是离散的，是一个不可微的注意力，在加入神经网络中通常是通过强化学习（Reinforcement learning）来完成网络的训练过程，如文献[65]很好地将硬注意力机制与强化学习相结合。相较于硬注意力，软注意力（Soft attention）更为常见，软注意力是在 0-1 区间内的连续表征，用连续值表示每个区域被关注程度的大小。由于软注意力的表征是连续的，具有可微的性质，可以在神经网络中算出梯度进行前向和反向传播，进而得到注意力的权重。

就注意力机制关注的域来划分，可以分为空间域注意力（Spatial domain attention）、通道域注意力（Channel domain attention）和混合域注意力（Mixed domain attention）。空间域注意力即对每张特征图的不同区域赋予不同的关注度，用权重表示对特征图特定的关注，同一特征图的不同区域对结果的影响不一样，我们对影响力较大的区域赋予较大的权重。通道域注意力的原理也很简单，我们可以从信号变换的角度来理解，在信号与系统中，任何信号都可以表示成正弦波的线性组合，连续的正弦波在通过时频变换后，可以用一个频率信号数值表示。在卷积神经网络中，图片由 RGB 三通道表示，经过卷积操作后产生新的通道，每个通道的特征图表示该图片在不同卷积核上的分量。类似于时频变换，卷积操作类似于做了傅里叶变换，从而将三个通道的信息分解成多个卷积核上的信号分量，不同分量对结果的影响力肯定不一样，通道域注意力则赋予不同卷积核不一样的权重。简单了解空间域注意力和通道域注意力之后，我们可以发现，空间域注意力忽略了通道间的信息，通道域注意力则忽略了每个通道内的局部信息，结合这两种思想，就可以设计出混合域的注意力机制。关于这几种注意力机制具体在神经网络中的应用，本文会在下一小节进行介绍。

2.4.3 常见注意力机制模型

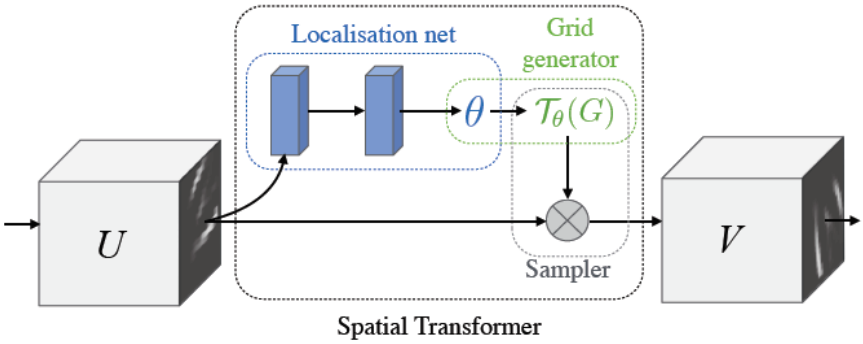


图 2-23 空间转换器[66]

Jaderberg 等人[66]提出了空间转换网络（Spatial transformer networks）模型，其中主要结构空间转换器（Spatial transformer）其本质是一种空间域注意力机制。由于之前的卷积神经网络大多直接采用池化操作（最大池化或平均池化）进行信息整合，作

者认为过于简单粗暴，会丢失很多有用信息，而且作者希望神经网络学习如何对输入图像进行空间变换，以增强模型的几何不变性。例如，网络可以裁剪感兴趣的区域、缩放并更正图像的方向。空间转换器的结构如

图所示，它主要包含三个部分：（1）本地网络（Localisation network），（2）网格生成器（Grid Generator），（3）采样器（Sampler）。其中本地网络由卷积操作或全连接操作组成，得到空间仿射变换参数；网格生成器根据仿射变换参数构建一个采样网格，得到由输入数据经过采样变换后的输出；采样器利用输入的特征图和采样网络的输出特征图进行采样，得到经过空间转换后的结果。

Hu 等人[67] 提出了 SENet（Squeeze and Excitation Networks），其论文中主要创新结构如图所示，其本质是一个通道域注意力机制。Squeeze-and-Excitation 模块主要分为三个部分：（1）挤压（Squeeze）；（2）激励（Excitation）；（3）变换（Scale）。其中挤压就是用全局平均池化（Global Average Pooling）将每个通道内所有特征值相加再平均，得到一个长度与输入特征图通道数一样的特征向量；激励就是利用 Bottleneck 的结构交换特征，先压缩通道数，再重构通道数，最后利用 Sigmod 激活函数得到 0-1 之间的注意力权重；缩放就是将得到的通道注意力权重乘回原始输入特征。

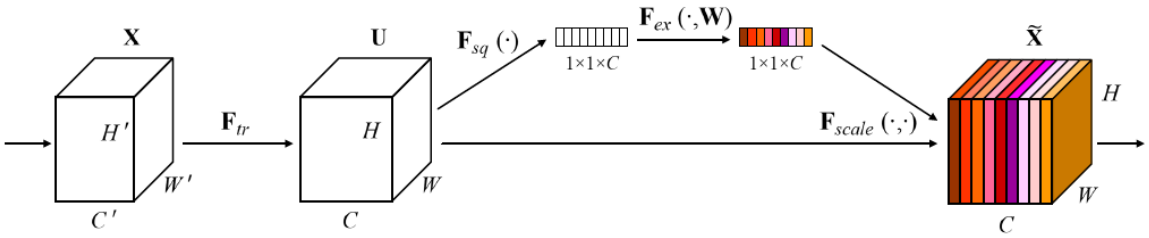


图 2-24 Squeeze-and-Excitation 模块[67]

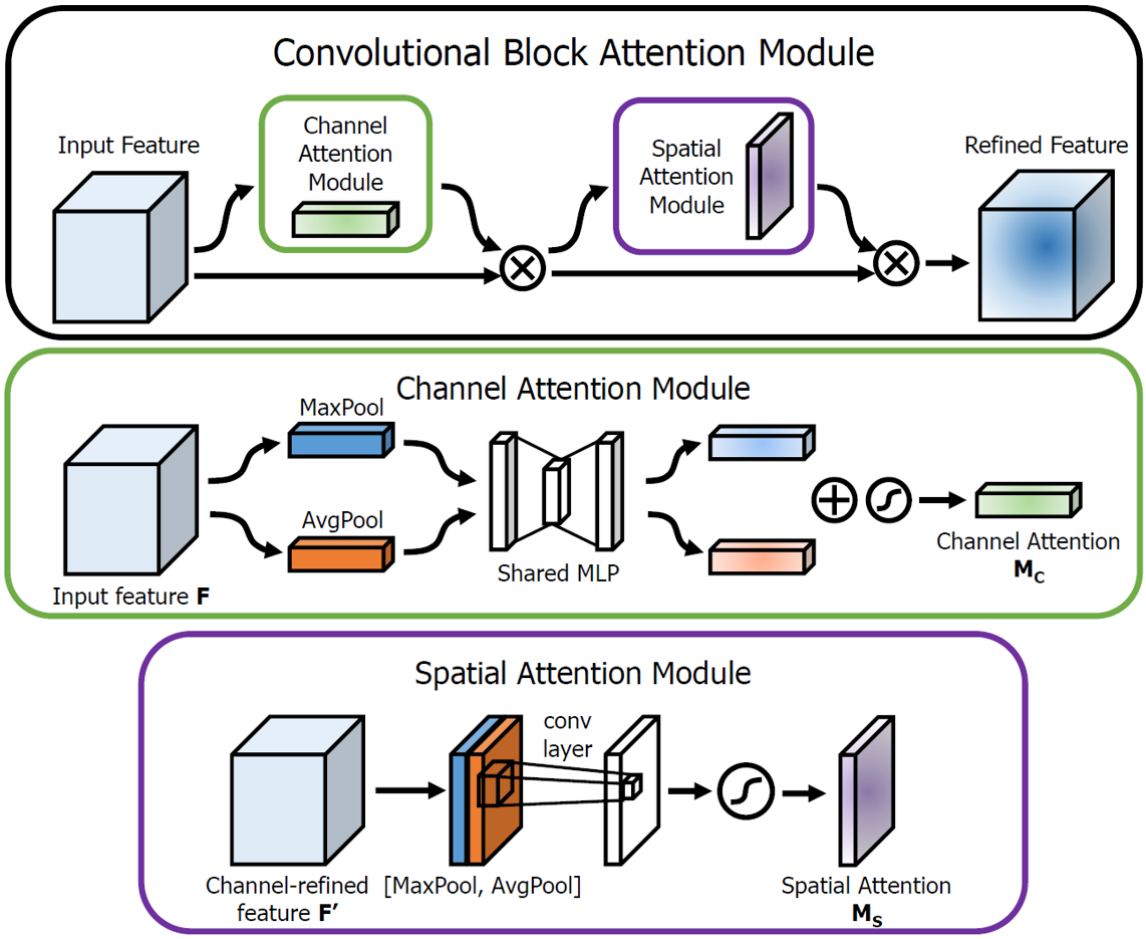


图 2-25 Convolutional Block Attention Module[68]

Woo 等人[68] 提出了 CBAM（Convolutional Block Attention Module），这是一种混合域注意力机制，其网络结构如图所示，CBAM 同时使用了空间域注意力和通道域注意力。其通道域注意力同时采用最大池化和平均池化操作，再经过 Bottleneck 的全连接层得到变换结果，然后分别应用于两个通道，最后使用 Sigmod 激活函数得到通道域的注意力结果。其空间域注意力首先将通道本身进行降维，分别获取最大池化和平均池化的结果，然后将其拼接成一个特征图，进过卷积操作融合特征，最后经过 Sigmod 激活函数得到空间域的注意力结果。

Chen 等人[69] 提出了一种注意力机制，学习在每个像素位置对多尺度特征进行软加权，他们采用了一个强大的语义分割模型，并与多尺度图像和注意力模型联合训练。注意力机制的性能优于平均池化和最大池化，使模型能够评估特征图在不同位置和尺度上的重要性。

与其他训练卷积神经网络来学习表征语义特征的工作不同，Huang 等人[70] 提出了一种基于反向注意力机制的语义分割方法。他们的反向注意力网络（Reverse Attention Network）架构同时训练模型去捕捉相反的特征（即与目标类无关的特征）。反向注意力网络是一个同时执行正向和反向注意力学习过程的三分支网络。

Liu 等人[71] 提出了一种用于语义分割的金字塔注意力网络，该网络模型充分利用了全局上下文信息对语义分割的影响。他们将注意力机制与空间金字塔相结合，提取精确的密集特征用于像素标记，而不是复杂的扩展卷积核人工设计的解码网络。

后来，Fu 等人[72] 提出了一种用于场景分割的双注意力网络，该网络能够基于自注意力机制捕获丰富的上下文依赖关系。他们在扩张卷积 FCN 的基础上附加了两种类型的注意模块，分别在空间域和通道域对语义相关性进行建模，空间注意力模块通过所有位置的特征的加权和选择性地聚集每个位置的特征。

还有其他一些研究探索了语义分割的注意力机制，如 OCNet[73] 提出了一种受自注意力机制启发的目标上下文池化（Object Context Pooling），最大期望注意力网络（Expectation-Maximization Attention）[74]，十字交叉注意力网络（Criss Cross Attention Network）[75]，基于循环注意力的端到端实例分割[76]，用于场景分割的点空间注意力网络（Pointwise Spatial Attention Network）[77] 和判别特征网络（Discriminative Feature Network）[78]。

第3章 基于注意力机制的 RGBD 图像语义分割算法

3.1 整体网络结构设计

本论文针对 RGBD 图像语义分割任务，提出了全新的端到端的 RGBD 语义分割算法框架（AFNet）如图所示，整个网络由特征提取网络、特征融合网络、特征解码网络构成。

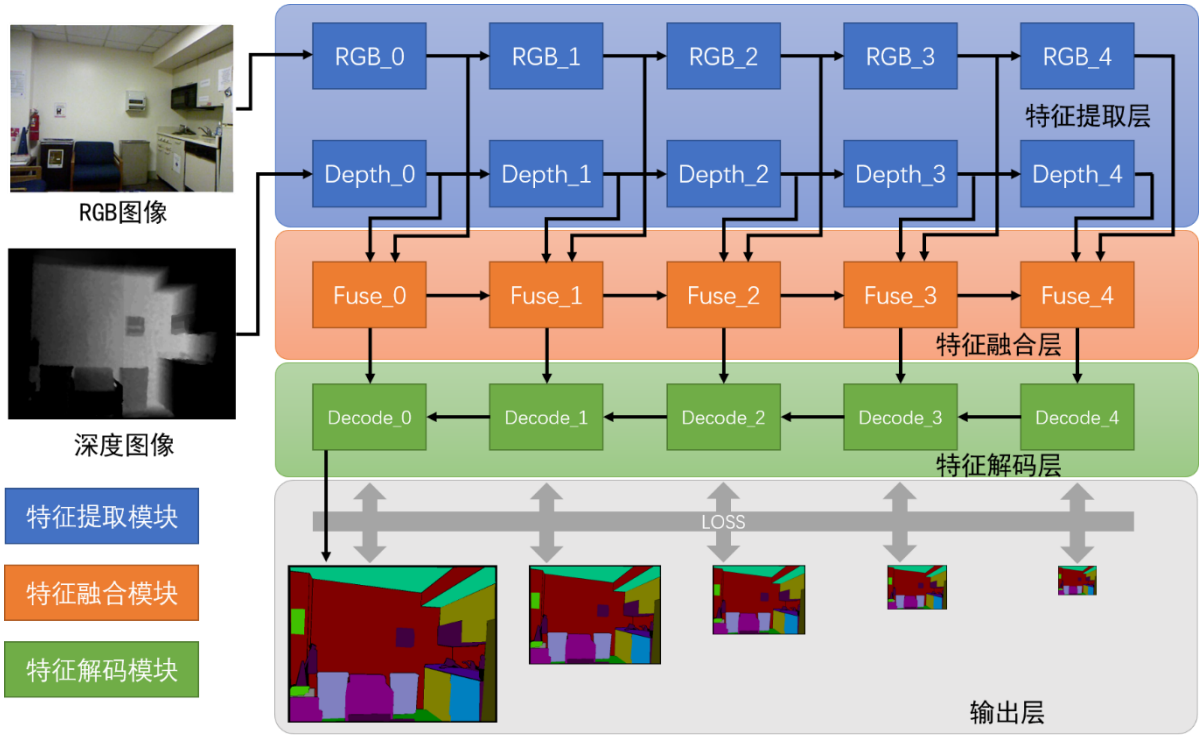


图 3-1 AFNet 网络架构

其中 RGB 图像和深度图像分别经由一路特征提取模块构成神经网络提取特征，然后将每个特征提取模块得到中间特征送入特征融合模块。相较于 Wang 等人[44] 只利用卷积神经网络最后输出层的特征进行融合的方式，本论文采用的多层融合的方式能够更好的融合不同特征尺度的信息以及保留较好的空间信息，更有利于后面的特征解码。相较于 FuseNet[43] 逐层将深度通道的特征图叠加进 RGB 通道特征图的方式，FuseNet 会从一开始就破坏掉了纯净的 RGB 通道的特征信息，本论文采用额外添加一个特征融合网络用于特征融合，使得 RGB 通道和深度通道在特征提取时能够得到纯净的特征，这样也有利于在实验中添加预训练模型时能有更与之匹配的初始参数分布。特征融合网络主要由特征融合模块组成，其输入为特征提取模块得到的 RGB 通道特征、深度通道特征以及上一个特征融合模块得到的特征，其输出为融合后的特征以及送入特征解码模块的跳连特征。特征解码网络由特征解码模块构成，其网络主要由特征融合模块得到的高维卷积特征以及跳连特征进行反卷积得到解码特征图以及语义输出。本论文在输出层引入多尺度输出金字塔，全部计入损失函数，得到更高精度的分割结果。

3.2 注意力模块设计

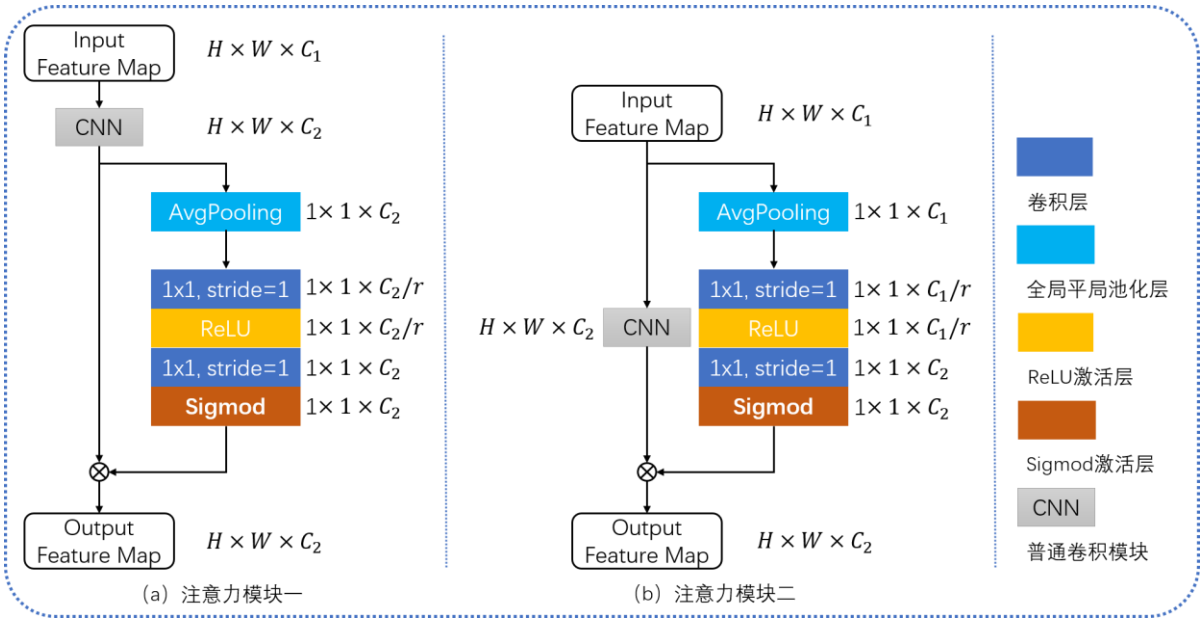


图 3-2 注意力模块

如图所示，本论文主要采用如下两种注意力机制模块，注意力模块一为 SENet[67] 中的的 Squeeze-and-Excitation 模块，其主要步骤如下：

（1） 对原始特征图进行转换操作（对输入的特征图进行卷积），将特征图的尺寸从$H \times W \times C_1$变为 $H \times W \times C_2$。将该操作定义为$conv(*)$。

（2） 接下来对特征图进行挤压（Squeeze）操作，也就是用全局平均池化（Global average pooling）将特征图的尺寸从$H \times W \times C_2$变为 $1 \times 1 \times C_2$。将该操作定义为$Fsq(*)$。

（3） 后面就是对特征图进行激励（Excitation）操作，这里采用的是类似 Bottleneck 的卷积操作，先利用1×1的卷积将通道数降为C_2/r（其中r是一个缩放系数，目的是为了减少通道数量进而降低计算量，本论文中所有的r均设置为 8），经过一个 ReLU 层后，又将通道数变为C_2，起到交换通道特征信息的作用，最后经过 Sigmod 函数，得到每一个通道的权重，输出尺寸为$1 \times 1 \times C_2$。将该操作记为$Fex(*)$。

（4） 最后就是变换（Scale）操作，将得到的通道权重与原特征图相乘，得到经过注意力之后的特征图。该操作记为$F_{scale}(*)$。

综合以上内容，我们定义输入为X，输出为Y，注意力模块一的计算逻辑可表示为：

$$Y=F_{scale}\left(F_{ex}\left(F_{sq}(conv(X))\right),conv(X)\right)=F_{ex}\left(F_{sq}(conv(X))\right)*conv(X) \quad (3\text{-}1)$$

注意力模块一是利用卷积操作的特征通道做注意力加权，我们知道卷积操作后的特征图会丢失掉一部分的空间信息，而本论文所讨论的语义分割任务需要较为完整的空间信息，注意力模块二对此做出了改进，利用卷积操作前的特征通道做注意力加权，这样可以利用到空间信息更为完整的特征图的信息，提高语义分割的精度。注意力模块二的计算逻辑可表示为：

$$Y=F_{scale}\left(F_{ex}\left(F_{sq}(X)\right),conv(X)\right)=F_{ex}\left(F_{sq}(X)\right)*conv(X) \qquad (3\text{-}2)$$

3.3 特征提取模块设计

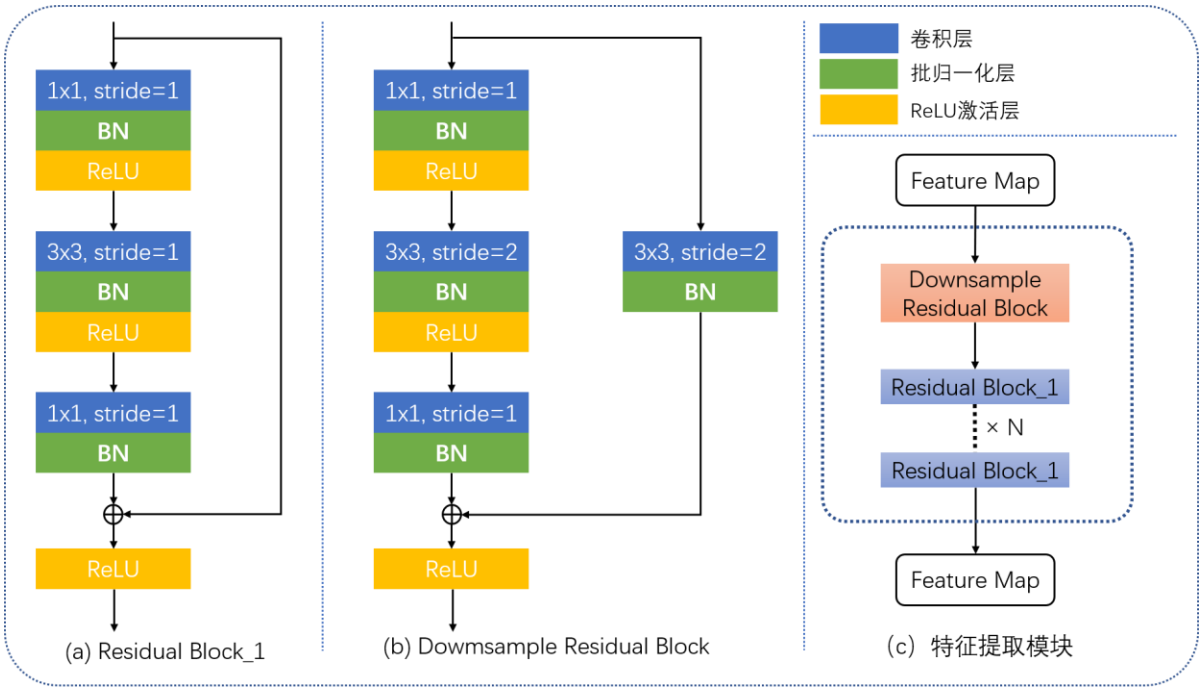


图 3-3 特征提取模块

特征提取网络的主体网络采用的是 ResNet[15]，其基本模块 Residual Block 如图所示，图（a）为基本的残差卷积模块，采用的是 Bottleneck 的结构，先利用1×1的卷积减少通道数，再利用3×3的卷积提取特征，最后利用1×1的卷积还原通道数，利用跳连结构，将输出与输入叠加，得到丰富的特征信息。图（b）为下采样的残差卷积模块，用来减小特征图尺寸。我们所用的特征提取模块由一个下采样残差卷积模块和多个普通残差卷积模块构成，其具体参数会在实验部分予以介绍。

3.4 特征融合模块设计

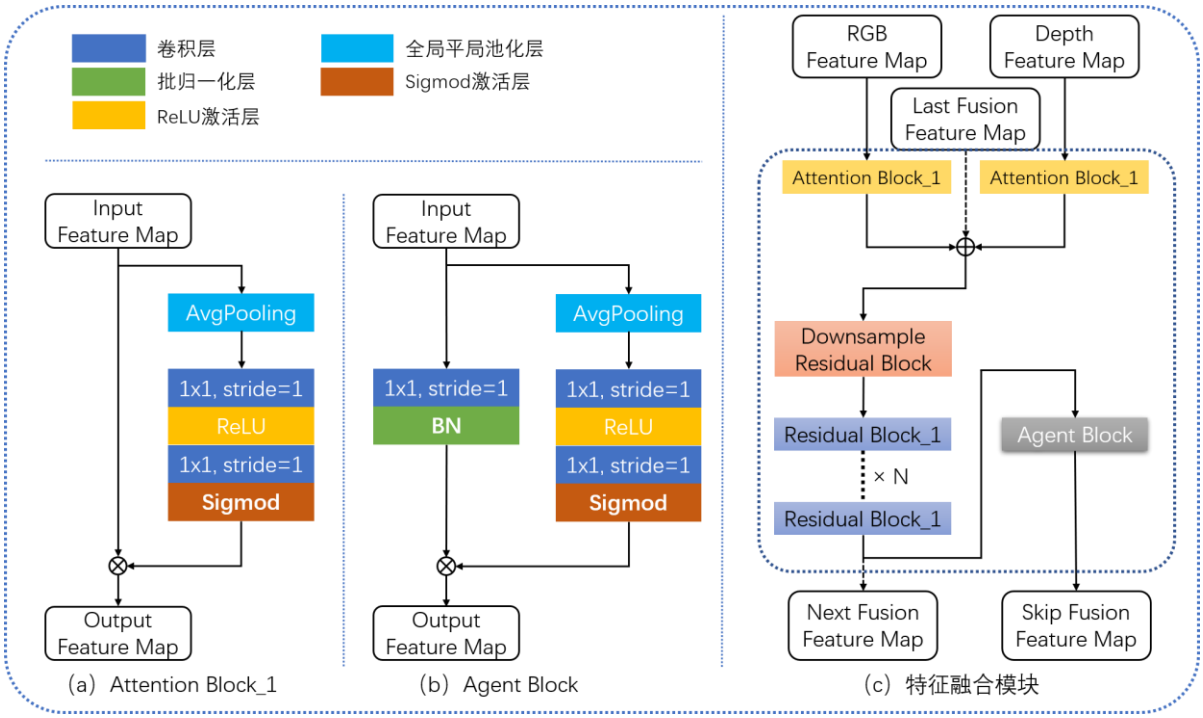


图 3-4 特征融合模块

如图（c）所示，特征融合模块的输入有三个：一个为 RGB 通道特征图，一个为深度通道特征图，另一个为上一个特征融合模块的输出（注：第一个特征融合模块的输入只有 RGB 和深度通道特征图）；特征融合模块的输出有两个：一个输入到下一个特征融合模块，一个输入到特征解码模块（注：最后一个特征融合只有一个输出，直接送入特征解码模块）。

不同于 FuseNet[43] 的直接将 RGB 通道特征图和深度通道特征图相加，本论文所设计的融合模块会先经过如图（a）所示的注意力模块，具体计算逻辑参考章节 3.2。RGB 图像提取到的特征和深度图像提取到的特征是两种不一样的特征，直接相加的方式不能有效地利用其特征值，经过通道域注意力机制后，调整了两通路的特征图映射关系，使其叠加时特征图能够更加匹配。

像素叠加后的特征图会经过与特征提取模块一致的残差卷积网络，一部分输出直接送入下一个特征融合模块，另一部分经过 Agent 模块后送入特征解码模块，Agent 模块的结构如图（b）所示。传统的跳连结构如 UNet[79] 会直接将跳连的特征图与解码器的特征图连接，然后用1×1的卷积降低通道数。这样做的缺点是，连接通道后特征图通道数增多，再进行卷积会增加较多的计算量。本论文所采用的方式是将跳连特征图先进行1×1的卷积操作降低通道数，再与特征解码模块中的特征图直接相加，一次来减少计算量。为减少在1×1卷积过程中通道数降低带来的信息丢失，我们用了章节 3.2 所提到注意力模块二的形式，捕获减少通道数前的特征图全局信息，修正

减少通道数后的特征图的值。

3.5 特征解码模块设计

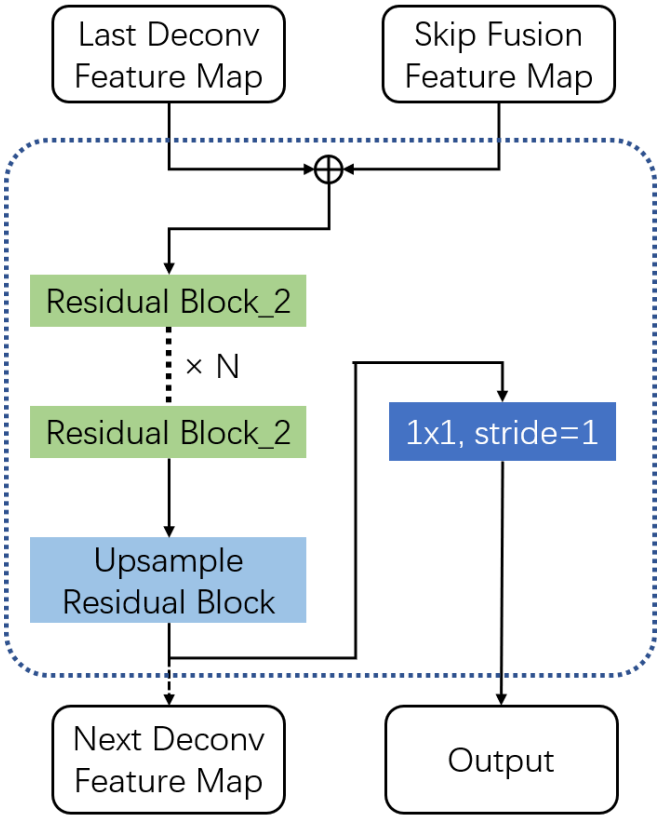


图 3-5 特征解码模块

如图所示，特征解码模块的输入有两个：一个为上一个特征解码模块的输出，另一个为特征融合模块输出的跳连特征（注：第一个特征解码模块的输入只有从特征融合模块输出的特征），两个输入连接的方式为特征图逐像素点叠加。特征解码模块的输出有两个：一个输入到下一个特征解码模块，另一个经过一个1×1的卷积降到通道数直接输出，得到不同尺寸的分割结果。

特征解码模块中的主体结构由残差卷积模块和残差反卷积模块构成，如图所示，本论文在残差卷积模块和残差反卷积模块中加入章节 3.2 中注意力模块二形式的注意力机制。其中图（a）由两个3×3的卷积层构成，通道域注意力关注卷积前的特征图全局信息，加权至卷积后的特征图中，尽可能多的保留空间信息，然后与输入特征图进行叠加。图（b）由一个3×3的卷积层和一个3×3反卷积层构造，采用同样的通道域注意力机制，跳连结构先经过一个3×3的反卷积再与输出叠加。

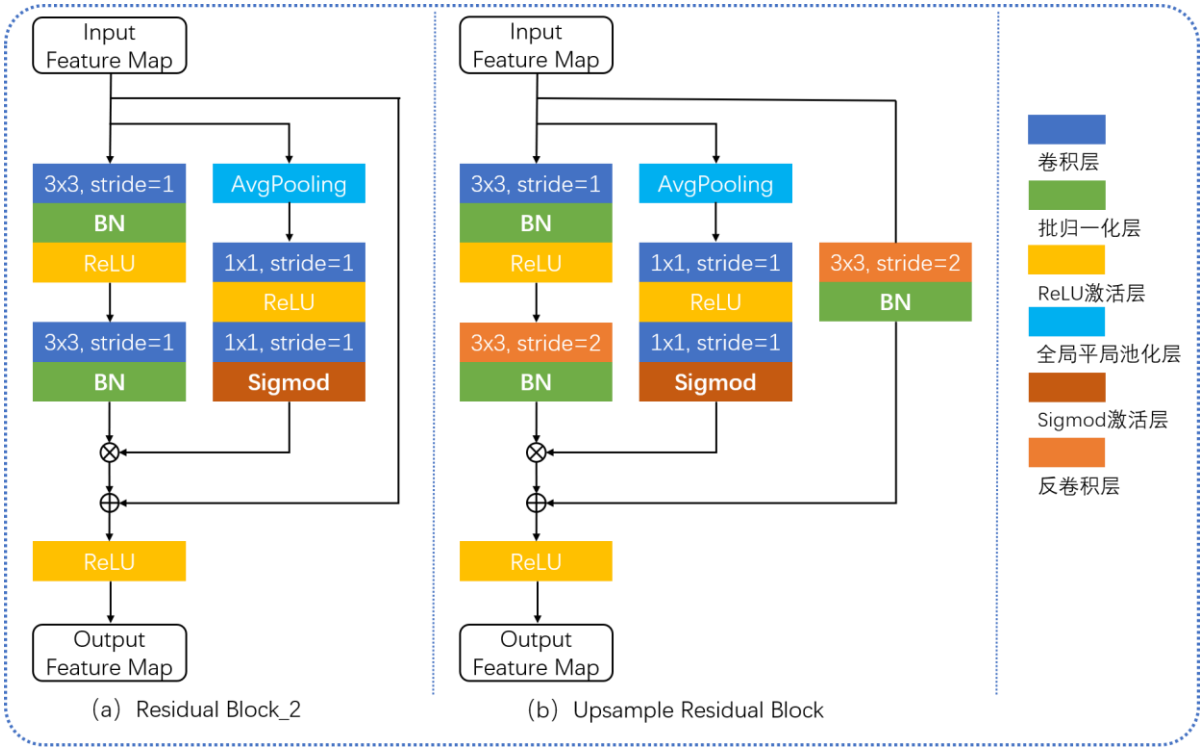


图 3-6 特征解码模块子模块

3.6 损失函数设计

由图可知，数据集存在严重的类别不均衡问题，而且数据集中有些类别的形状小、识别难度大。针对识别难度不同的问题，He 等人[80] 提出 Focal Loss，用于解决类别间识别难度不一的问题，本论文在此基础上扩张了 Focal Loss，采用带权重的 Focal Loss 解决数据集类别不均衡问题，损失函数的计算公式如下：

$$L_k=-\sum_{l}\sum_{c} W_c\times(1-p_{i,c})^2\times l^*\times\log(p_{i,c}) \tag{3-3}$$

其中i表示像素，$c \in 1,2,3,\ldots$表示标签类别，$p_{i,c}$表示预测像素i属于类别c的概率，l^*是标签的真实值，W_c为计算类别c损失时的权重。

权重为统计训练集所有标签值计算得出，计算过程如下：

$$Freq_c=num_c/\sum_{c} num_c \tag{3-1}$$

其中num_c为每一个类别c的总像素数量，$Freq_c$表示每一个类别c出现的频率。由此我们可以计算出每一个类别的权重：

$$W_c=median(Freq_c)/Freq_c \tag{3-2}$$

其中$median(Freq_c)$为$Freq_c$的中位数。

由于我们的网络模型采用金字塔输出结构，有五个不同尺寸的输出结果，总的损失函数为：

$$L_{all}=\sum_{k} L_k \tag{3-3}$$

其中$k \in 1,2,3,4,5$表示不同层级的输出结果。

第4章 实验结果与分析

4.1 数据库

NYU-V2[55] 是最流行的 RGBD 数据库之一，它包含了 646 个不同场景和 26 个不同场景类型的图像，由微软 Kinect 相机以 640*480 的分辨率拍摄而成。数据库还提供了由 Levin 等人[57] 的着色方法计算的修补深度图，本论文用修复后的深度图像进行实验。按照标准的训练集/测试集分割，我们使用 795 个训练图像和 654 个测试图像。整个数据库按照文献[58] 提供的标准划分为 40 个类标签。

4.2 评价指标

RGBD 图像语义分割的评价通常需要将输出的像素级类别图与标记类别图进行比较，判断每个像素是否分类正确。对于一张图像，我们C为标签总类别数；n_{ij}表示标签为i、预测为j的像素个数，统计所有的像素输出，可以得到一个大小为$C \times C$的混淆矩阵（Confusion Matrix），记为N；$t_i = \sum_{j=1}^{C} n_{ij}$表示属于类别$i$的总像素个数。

接下来，我们将介绍四种常见的指标：

（1） 像素准确率（Pixel accuracy）

该指标也被称为全局准确率（Global accuracy），其计算公式如下：

$$Pixel_{acc}=\frac{\sum_{i=1}^{C} n_{ii}}{\sum_{i=1}^{C} t_i} \tag{4-1}$$

像素准确率代表了像素正确预测的比例，此度量偏向于构成数据集中像素较多部分的类标签，在不平衡数据集中，这种度量不够公平，因为存在大量像素属于特定的少数类。如果算法正确预测了像素数量较多的类，及时像素数量较少的类全部预测错误，依然能够得到很高的数值指标。

（2） 类别准确率（Class accuracy）

该指标为类别的平均准确率，换而言之，它是归一化混淆矩阵的平均值，计算公式如下：

$$Class_{acc}=\frac{1}{C}\sum_{i=1}^{C}\frac{n_{ii}}{t_i} \quad (4\text{-}2)$$

该指标考虑到了像素数量较少的类，是一个较好的度量。

（3） 平均交并比（MIoU）

该指标是由每一个类别的预测值与标签值之间的并集比上交集的平均值得到的结果，它是比较两个集合的相似性和多样性的一种相似性度量，是最常用的一种指标，计算方式如下：

$$MIoU=\frac{1}{C}\sum_{i=1}^{C}\frac{n_{ii}}{t_i-n_{ii}+\sum_{j=1}^{C}n_{ji}} \quad (4\text{-}3)$$

此度量优于先前的两种度量，但它没有考虑到每个分割区域的轮廓是否精确。

（4） 加权交并比（WIoU）

为考虑到不同类别出现的频率不一样，故而定义了一个加权版的 IoU 指标，其计算公式如下：

$$WIoU=\frac{1}{\sum_{i=1}^{C}t_i}\sum_{i=1}^{C}\frac{t_i*n_{ii}}{t_i-n_{ii}+\sum_{j=1}^{C}n_{ji}} \quad (4\text{-}4)$$

该指标在其他论文中所用较少。

4.3 实验设计

4.3.1 实验环境

本论文实验以 GPU 集群 Linux 服务器作为硬件平台，其系统版本为 Ubuntu18.04，其主要计算资源为 32 核 Intel Xeon Gold 5115 型号的 CPU，主频为 2.4GHz，以及 8 块 NVIDIA Tesla P40 型号的 GPU，其显存为 24G。本论文方法实现所涉及的神经网络搭建平台为 Pytorch[81] 。

4.3.2 实验参数设置

训练参数方面，模型训练的最大代数为 300，并以此为模型训练终止条件。优化器方面，本实验采用带动量的 SGD[82] 作为模型的优化器，初始学习率设置为 0.002，momentum 设置为 0.9，正则化项 weight decay 设置为 0.0001。本实验还采用了自适应学习率下降的策略，监测指标持续 10 代没有发生变化，则降低当前学习率到 90%。

本实验所做的数据预处理包括统一将图像缩放至 640*480 分辨率，随机左右翻转，随机裁剪，HSV 色域颜色变换，图像像素归一化操作。

本实验的特征提取网络和特征融合网络中的主体残差网络主要采用 ResNet50、ResNet101、ResNet152 三种架构，其网络参数如表所示，其中[*]表示一个残差卷积

模块，其表达式$n \times n, c, s$中n为卷积核大小，c为通道数，s表示步长。当s为标注时，为默认值 1。$s = 1$时的残差模块结构如图（a）所示，$s = 2$时的残差模块结构如图（b）所示。

表 4-1 本实验特征提取模块和特征融合模块所采用残差网络主要参数设置

模块名称	输出尺寸	ResNet50	ResNet101	ResNet152
RGB_0/ Depth_0	320*240	$7 \times 7, 64, s = 2$		
RGB_1/ Depth_1/ Fuse_1	160*120	$3 \times 3\ max\ pool, s = 2$		
		$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
RGB_2/ Depth_2/ Fuse_2	80*60	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128, s = 2 \\ 1 \times 1, 512 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128, s = 2 \\ 1 \times 1, 512 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128, s = 2 \\ 1 \times 1, 512 \end{bmatrix} \times 1$
		$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 7$
RGB_3/ Depth_3/ Fuse_3	40*30	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256, s = 2 \\ 1 \times 1, 1024 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256, s = 2 \\ 1 \times 1, 1024 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256, s = 2 \\ 1 \times 1, 1024 \end{bmatrix} \times 1$
		$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 5$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 22$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 35$
RGB_4/ Depth_4/ Fuse_4	20*15	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512, s = 2 \\ 1 \times 1, 2048 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512, s = 2 \\ 1 \times 1, 2048 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512, s = 2 \\ 1 \times 1, 2048 \end{bmatrix} \times 1$
		$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 2$

本实验所采用的特征解码模块如图和图所示，其主要网络参数设置如表所示，其中T表示该卷积层为反卷积层，当步长$s = 1$时的注意力残差模块结构如图（a）所示，当步长$s = 2$时的注意力残差模块结构如图（b）所示。其中注意力机制的通道数由特征解码模块的输入输出唯一决定。

表 4-2 本实验特征解码模块的网络主要参数设置

模块名称	Decode_0	Decode_1	Decode_2	Decode_3	Decode_4
输出尺寸	640*480	320*240	160*120	80*60	40*30
模块参数	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 5$
	$T, 3 \times 3, n, s$ =2	$\begin{bmatrix} 3 \times 3, 64 \\ T, 3 \times 3, 64, s = 2 \end{bmatrix} \times 1$	$\begin{bmatrix} 3 \times 3, 128 \\ T, 3 \times 3, 64, s = 2 \end{bmatrix} \times 1$	$\begin{bmatrix} 3 \times 3, 256 \\ T, 3 \times 3, 128, s = 2 \end{bmatrix} \times 1$	$\begin{bmatrix} 3 \times 3, 512 \\ T, 3 \times 3, 256, s = 2 \end{bmatrix} \times 1$

4.3.3 实验设计

为证明本论文所采用的各个机制的有效性（包括引入深度图像、采用预训练模型、增加跳连结构、金字塔输出、注意力机制），我们设计了组内对照实验，具体内容如下：

（1） 实验一（RGB）：只采用 RGB 图像作为输入，输出只有最后一个分割图，采用最基本的编码器解码器架构，无跳连结构；

（2） 实验二（RGBD_cat）：在实验一的基础上，将 RGB 图像和深度图像合并成四通道作为网络的输入；

（3） 实验三（RGBD_split）：将 RGB 图像和深度图像分别输入到两路特征提取网络中，并将提取到的特征送入特征融合网络；

（4） 实验四（RGBD_pretrained）：在实验三的基础上，使用 ResNet 在 ImageNet[83] 上的预训练模型参数，改变本实验网络的初始参数部分，使网络收敛的更好；

（5） 实验五（RGBD_skip）：在实验四的基础上，增加跳连结构，将特征融合模块的输出层级引入特征解码模块中；

（6） 实验六（RGBD_pyramid）：在实验五的基础上，引入金字塔输出，输出多个不同尺度的分割图，并均计入损失函数中；

（7） 实验七（RGBD_attention）：在实验六的基础上，在特征融合模块以及特征解码模块中引入注意力机制；

（8） 实验八（RGBD_resnet101）：在实验七的基础上，将特征提取模块中的 ResNet50 替换成 ResNet101；

（9） 实验九（RGBD_resnet152）：在实验七的基础上，将特征提取模块中的 ResNet50 替换成 ResNet152。

4.4 结果分析

如**错误!未找到引用源。**所示，该结果为本文方法的组内实验对比，具体设置参考章节 0。由实验 RGBD_cat 和 RGB 的结果对比可知，引入深度图像可以帮助 RGB 图像进行语义分割，并且可以有较大的提升，大约在$MIoU$这个指标上可以提升 5.35 个百分点；由实验 RGBD_split 和实验 RGBD_cat 的结果对比可知，将深度图像作为一个单独的通路进行输入，比将深度图像合并入 RGB 图像进行输入，分割的效果更好，大约在$MIoU$这个指标上可以提升 2.56 个百分点；由实验 RGBD_pretrained 和实验 RGBD_split 的结果对比可知，引入预训练模型，改变初始模型参数分布，可以很大的提升网络性能，大约在$MIoU$这个指标上可以提升 11.01 个百分点；由实验 RGBD_skip 和实验 RGBD_pretrained 的结果对比可知，引入跳连结构可以在$MIoU$这个指标上提升 1.26 个百分点；由实验 RGBD_pyramid 和实验 RGBD_skip 的结果对比可知，增加金字塔层级输出，可以很好地改善分割精度，大约在$MIoU$这个指标上可以提升 4.71 个百分点；由实验 RGBD_attention 和实验 RGBD_pyramid 的结果对比可知，增加注意力机制，大约在$MIoU$这个指标上可以提升 1.44 个百分点；由实验 RGBD_resnet101 和实验 RGBD_resnet152 可知，增加特征提取网络深度，也可以略微改善网络性能。

表 4-3 本文方法的组内实验在 NYU-V2 数据集上的对比结果

实验名称	编码网络	深度图像	预训练模型	跳连结构	金字塔输出	注意力机制	$Pixel_{acc}$	$Class_{acc}$	$MIoU$	$WIoU$
RGB	Resnet50						0.5277	0.3845	0.2188	0.3928
RGBD_cat	Resnet50	√					0.5821	0.4421	0.2723	0.4439
RGBD_split	Resnet50	√					0.6027	0.4690	0.2979	0.4667
RGBD_pretrained	Resnet50	√	√				0.6976	0.6215	0.4080	0.5714
RGBD_skip	Resnet50	√	√	√			0.7101	0.6036	0.4206	0.5866
RGBD_pyramid	Resnet50	√	√	√	√		0.7272	0.6480	0.4677	0.6042
RGBD_attention	Resnet50	√	√	√	√	√	0.7364	0.6609	0.4821	0.6142
RGBD_resnet101	Resnet101	√	√	√	√	√	0.7461	0.6719	0.4918	0.6256
RGBD_resnet152	Resnet152	√	√	√	√	√	0.7529	0.6723	0.5013	0.6348

表 4-4 本文方法与其他先进方法在 NYU-V2 数据集上的对比结果

方法	$Pixel_{acc}$	$Class_{acc}$	$MIoU$
Ren et al.[84]	0.493	0.211	0.214
Gupta et al.[58]	0.591	0.284	0.291
FCN[27]	0.654	0.461	0.340
Liu et al.[49]	0.703	0.517	0.412
LSTM-CF[45]	-	-	0.494
3D Graph[52]	-	0.557	0.431
D-CNN[85]	-	0.563	0.439
Cheng et al.[50]	0.719	0.600	0.459
Lin et al.[86]	-	-	0.477
RDFNet[46]	**0.760**	0.628	0.501
Propose Networks	0.7529	**0.6723**	**0.5013**

表显示了本文方法与其他新进方法的对比结果，可以看到，本文方法与先前最好的方法在$Pixel_{acc}$和$MIoU$这两个指标上都能有与之相近或更好的结果，在$Class_{acc}$这个指标上，我们的方法要远高于前人的方法。

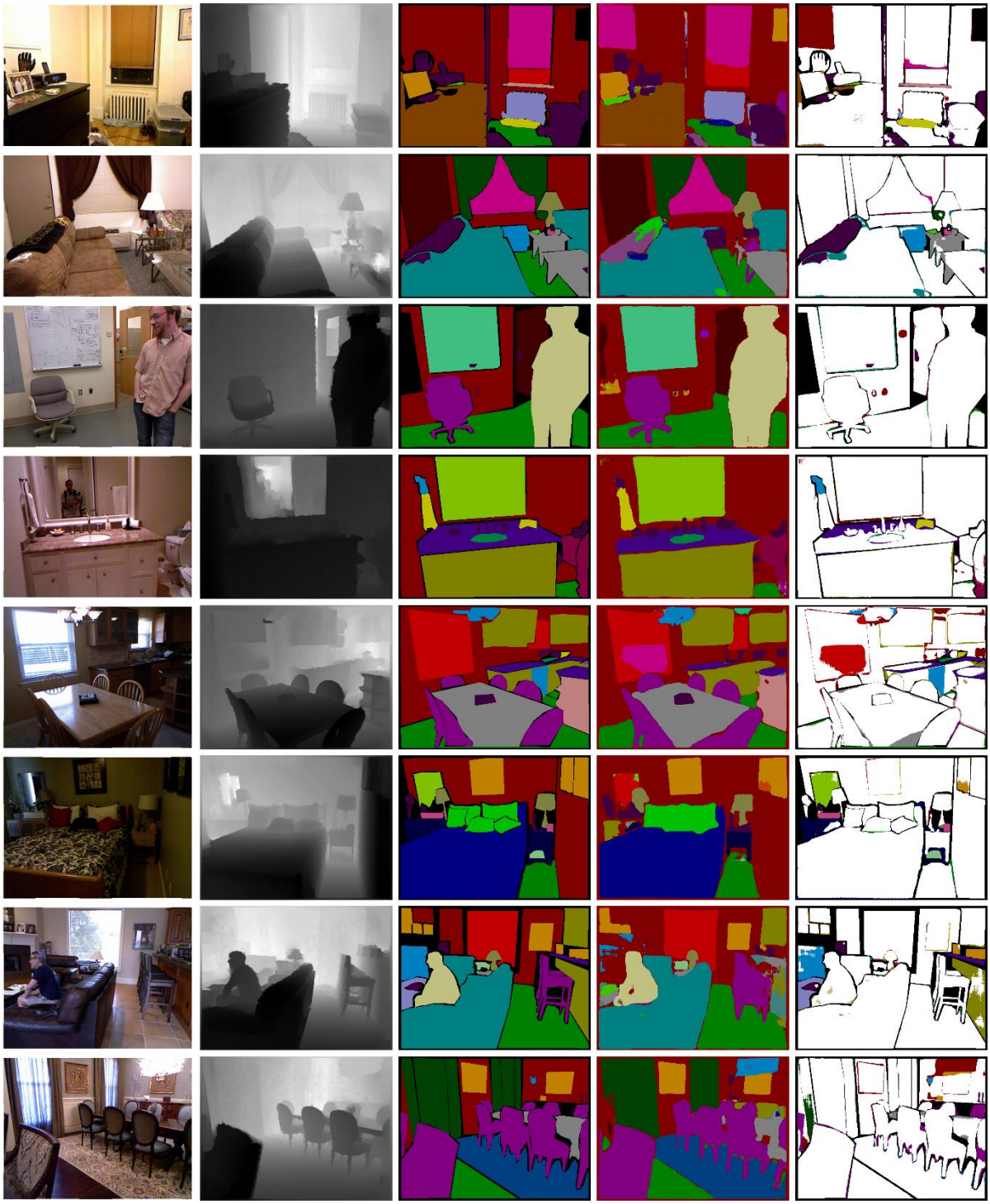
表 4-1 本文方法在 NYU-V2 数据上每一个类别的分类准确率

类别	wall	floor	cabinet	bed	chair	sofa	table	door
IoU	0.744	0.8857	0.615	0.7014	0.6386	0.6331	0.4508	0.4285
类别	window	bookshelf	picture	counter	blinds	desk	shelves	curtain
IoU	0.4649	0.4559	0.626	0.6881	0.6092	0.2146	0.1929	0.6548
类别	dresser	pillow	mirror	floor mat	clothes	ceiling	books	refrigerator
IoU	0.5321	0.4691	0.5001	0.4673	0.2362	0.7614	0.3285	0.5427
类别	television	paper	towel	shower curtain	box	whiteboard	person	night stand
IoU	0.6154	0.3418	0.3853	0.4517	0.132	0.7353	0.8158	0.4289
类别	toilet	sink	lamp	bathtub	bag	other structure	other furniture	other prop
IoU	0.7226	0.5896	0.5063	0.4993	0.0793	0.318	0.2013	0.3891

表 4-1 显示了本文方法在 NYU-V2 数据上每一个类别的分类准确率，可以看到，在一些较难的类（如镜子）上也能得到不错的结果，但在较少的类（如背包）上分割结果还是不令人满意，这也是网络后续需要提升的方向。

图显示了本文方法在 NYU-V2 数据集上的结果示例。第一列为 RGB 原图像，第二列为深度图像，第三列为数据集标定的标签图，第四列网络模型预测的分类图像，第五列为分析图像，白色表示预测对的像素部分，彩色表示预测错误的像素部分，其颜色对应原本应该属于的类标记。由结果可知，本文方法在大部分场景可以有较好的分割结果，分割边缘也较清晰。

图显示了本文方法在NYU-V2 数据集上金字塔层级输出结果示例。第一行为RGB 原图像，第二行为数据集标定的标签图，后面依次为特征解码网络 Decode_0、Decode_1、Decode_2、Decode_3、Decode_4 的层级输出结果，其原本的输出分辨率依次为 640*480、320*240、160*120、80*60、40*30，为方便结果显示，图统一放大到 640*480 的尺寸进行显示。我们可以看到，解码网络在较浅层的时候也能大致输出空间轮廓信息，层级越高时，分割的结果越精确。



（a）原图　（b）深度图　（c）标记图　（d）预测图　（e）分析图

图 4-1 本文方法在 NYU-V2 数据集上的结果示例

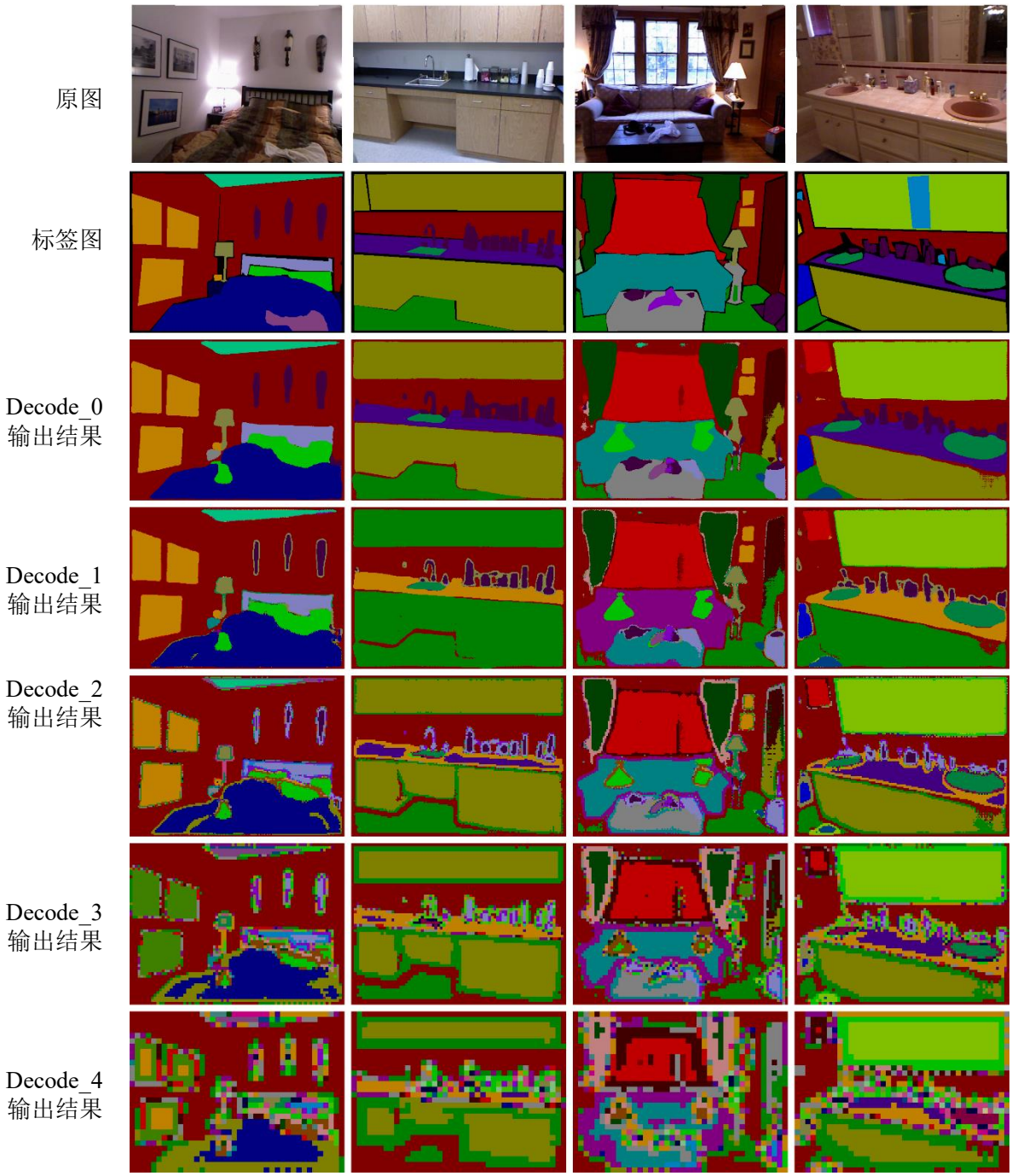


图 4-2 本文方法在 NYU-V2 数据集上金字塔层级输出结果示例

第5章 结论及展望

5.1 结论

图像语义分割被广泛应用于自动驾驶、AR/VR 交互、机器人等前沿领域，在实际研究中存在场景多样、拍摄角度广泛、类别繁多、光线变化大、数据不均衡以及复杂

而广泛的上下文关系。传统的 RGB 图像语义分割算法由于缺乏三维空间位置信息，在算法层面具有一定的局限性。本文引入深度图像用来帮助 RGB 图像进行语义分割，提出了基于编码器解码器架构的注意力融合网络（AFNet），实现了快速且高精度的室内 RGBD 图像语义分割。本文的主要贡献如下：

（1） 提出了基于编码器解码器架构的注意力融合网络（AFNet），实现了端到端的 RGBD 图像语义分割。

（2） 在 RGB 和深度图像融合上，提出了单独的特征融合网络，让 RGB 通路和深度通道分别卷积、互不影响，同时在特征融合模块中引入了注意力机制，实现了两种不同特征图的有效融合；

（3） 在跳连结构中引入注意力机制，降低了通道数，减少了计算量，同时保留了足够的空间信息；

（4） 改进了 Resnet 中的残差卷积模块，利用注意力机制增加其在解码器中的全局信息捕获能力；

（5） 引入金字塔层级输出，实现了多尺度输出，提高了分割精度；

（6） 针对类别不均衡以及难易不均衡现象，引入了带权重的 Focal loss，改善了分割结果。

5.2 研究课题展望

在目前的 RGBD 图像分割领域，目前公开的数据集不多，且存在较大的标注错误问题，同时样本的识别难度很大。在后续工作中，可以研究对标注样本依赖较少的弱监督或无监督学习，这类研究对实际应用具有重要意义。

虽然本文的方法可以快速的实现 RGBD 语义分割，但依旧无法满足机器人应用上实时性的要求，针对语义分割网络的模型轻量化以及压缩，是后续应用上非常重要的一环。

本文利用注意力机制改善了 RGB 图像深度图像的融合方式，但依旧有些局限性，如何更有效地利用深度信息来帮助 RGB 图像进行语义分割，依旧是非常重要的研究课题。

参考文献

[1] Wang, Wei, et al. "Development of convolutional neural network and its application in image classification: a survey." Optical Engineering 58.4 (2019): 040901.

[2] Hubel, David H., and Torsten N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex." The Journal of physiology 160.1 (1962): 106-154.

[3] Fukushima, Kunihiko, and Sei Miyake. "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position." Pattern recognition 15.6 (1982): 455-469.

[4] Werbos, Paul. "Beyond regression:" new tools for prediction and analysis in the behavioral sciences." Ph. D. dissertation, Harvard University (1974).

[5] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." nature 323.6088 (1986): 533-536.

[6] LeCun, Yann, et al. "Handwritten digit recognition with a back-propagation network." Advances in neural information processing systems. 1990.

[7] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.

[8] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[9] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[10] Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400 (2013).

[11] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[12] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[13] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint

arXiv:1502.03167 (2015).

[14] Szegedy, Christian, et al. "Inception-v4, inception-resnet and the impact of residual connections on learning." AAAI. Vol. 4. 2017.

[15] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[16] Minaee, Shervin, et al. "Image Segmentation Using Deep Learning: A Survey." arXiv preprint arXiv:2001.05566 (2020).

[17] R. Szeliski, Computer vision: algorithms and applications. Springer Science & Business Media, 2010.

[18] D. Forsyth and J. Ponce, Computer vision: a modern approach. Prentice Hall Professional Technical Reference, 2002.

[19] N. Otsu, “A threshold selection method from gray-level histograms,” IEEE transactions on systems, man, and cybernetics, vol. 9, no. 1, pp. 62–66, 1979.

[20] R. Nock and F. Nielsen, “Statistical region merging,” IEEE Transactions on pattern analysis and machine intelligence, vol. 26, no. 11, pp. 1452–1458, 2004.

[21] N. Dhanachandra, K. Manglem, and Y. J. Chanu, “Image segmentation using k-means clustering algorithm and subtractive clustering algorithm,” Procedia Computer Science, vol. 54, pp. 764–771, 2015.

[22] L. Najman and M. Schmitt, “Watershed of a continuous function,” Signal Processing, vol. 38, no. 1, pp. 99–112, 1994.

[23] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” International journal of computer vision, vol. 1, no. 4, pp. 321–331, 1988.

[24] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via graph cuts,” IEEE Transactions on pattern analysis and machine intelligence, vol. 23, no. 11, pp. 1222–1239, 2001.

[25] N. Plath, M. Toussaint, and S. Nakajima, “Multi-class image segmentation using conditional random fields and global classification,” in Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009, pp. 817–824.

[26] J.-L. Starck, M. Elad, and D. L. Donoho, “Image decomposition via the combination of sparse representations and a variational approach,” IEEE transactions on image processing, vol. 14, no. 10, pp. 1570–1582, 2005.

[27] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional

networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[28] Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[29] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.

[30] Kendall, Alex, Vijay Badrinarayanan, and Roberto Cipolla. "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding." arXiv preprint arXiv:1511.02680 (2015).

[31] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[32] Zhao, Hengshuang, et al. "Pyramid scene parsing network." IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017.

[33] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2018): 834-848.

[34] Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." arXiv preprint arXiv:1511.07122 (2015).

[35] Wang, Panqu, et al. "Understanding convolution for semantic segmentation." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018.

[36] Yang, Maoke, et al. "Denseaspp for semantic segmentation in street scenes." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[37] Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." arXiv preprint arXiv:1606.02147 (2016).

[38] Chen, Liang-Chieh, et al. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

[39] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." arXiv preprint arXiv:1802.02611 (2018).

[40] Fooladgar, Fahimeh, and Shohreh Kasaei. "A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks." Multimedia Tools and Applications 79.7 (2020): 4499-4524.

[41] Arbelaez P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. IEEE Trans Pattern Anal Mach Intell 33(5):898–916.

[42] Couprie, Camille, et al. "Indoor semantic segmentation using depth information." arXiv preprint arXiv:1301.3572 (2013).

[43] Hazirbas, Caner, et al. "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture." Asian conference on computer vision. Springer, Cham, 2016.

[44] Wang J,Wang Z, Tao D, See S,Wang G (2016) Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In: European conference on computer vision. Springer, pp 664–679.

[45] Li Z, Gan Y, Liang X, Yu Y, Cheng H, Lin L (2016) Lstm-cf: unifying context modeling and fusion with lstms for rgb-d scene labeling. In: European conference on computer vision. Springer, pp 541–557.

[46] Park, Seong-Jin, Ki-Sang Hong, and Seungyong Lee. "Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation." Proceedings of the IEEE International Conference on Computer Vision. 2017, pp 4980-4989.

[47] Lin, Guosheng, et al. "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp 1925-1934.

[48] Jiang, Jindong, et al. "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation." arXiv preprint arXiv:1806.01054 (2018).

[49] Liu, Hong, et al. "RGB-D joint modelling with scene geometric information for indoor semantic segmentation." Multimedia Tools and Applications 77.17 (2018): 22475-22488.

[50] Cheng, Yanhua, et al. "Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp 3029-3037.

[51] Han J, Shao L, Xu D, Shotton J (2013) Enhanced computer vision with microsoft kinect sensor: a review. IEEE Trans Cybern 43(5):1318–1334.

[52] Qi, Xiaojuan, et al. "3d graph neural networks for rgbd semantic segmentation." Proceedings of the IEEE International Conference on Computer Vision. 2017, pp 5199-5208.

[53] Keselman, Leonid, et al. "Intel realsense stereoscopic depth cameras." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017, pp 1-10.

[54] Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp 652-660.

[55] Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. Springer, pp 746–760.

[56] Jiao, Jianbo, et al. "Geometry-aware distillation for indoor semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019, pp 2869-2878.

[57] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. In ACM transactions on graphics (tog), volume 23, pages 689–694. ACM, 2004.

[58] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In Proc. CVPR, pages 564–571, 2013.

[59] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3d object dataset: Putting the kinect to work. In Consumer Depth Cameras for Computer Vision, pages 141–165. Springer, 2013.

[60] J. Xiao, A. Owens, and A. Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In Proc. ICCV, pages 1625–1632, 2013.

[61] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proc. CVPR, pages 567–576, 2015.

[62] John R Anderson. Cognitive psychology and its implications . WH Freeman/Times Books/Henry Holt & Co, 1990.

[63] Wang, F., & Tax, D. M. (2016). Survey on the attention based RNN model and its applications in computer vision. arXiv preprint arXiv:1601.06823.

[64] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[65] Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In Advances in neural information processing systems (pp. 2204-2212).

[66] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In Advances in neural information processing systems (pp. 2017-2025).

[67] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[68] Woo, S., Park, J., Lee, J. Y., & So Kweon, I. (2018). Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19).

[69] Chen, L. C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3640-3649).

[70] Huang, Q., Xia, C., Wu, C., Li, S., Wang, Y., Song, Y., & Kuo, C. C. J. (2017). Semantic segmentation with reverse attention. arXiv preprint arXiv:1707.06426.

[71] Li, H., Xiong, P., An, J., & Wang, L. (2018). Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180.

[72] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3146-3154).

[73] Y. Yuan and J. Wang, “Ocnet: Object context network for scene parsing,” arXiv preprint arXiv:1809.00916, 2018.

[74] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, “Expectation-maximization attention networks for semantic segmentation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9167–9176.

[75] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “Ccnet: Criss-cross attention for semantic segmentation,” in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 603–612.

[76] M. Ren and R. S. Zemel, “End-to-end instance segmentation with recurrent attention,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6656–6664.

[77] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. Change Loy, D. Lin, and J. Jia, “Psanet: Point-

wise spatial attention network for scene parsing," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 267–283.

[78] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Learning a discriminative feature network for semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1857–1866.

[79] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[80] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[81] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., ... & Lerer, A. (2017). Automatic differentiation in pytorch.

[82] Qian, Ning. "On the momentum term in gradient descent learning algorithms." Neural networks 12.1 (1999): 145-151.

[83] Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International journal of computer vision 115.3 (2015): 211-252.

[84] Ren, X., Bo, L., & Fox, D. (2012, June). Rgb-(d) scene labeling: Features and algorithms. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2759-2766). IEEE.

[85] Wang, W., & Neumann, U. (2018). Depth-aware cnn for rgb-d segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 135-150).

[86] Lin, D., Chen, G., Cohen-Or, D., Heng, P. A., & Huang, H. (2017). Cascaded feature network for semantic segmentation of RGB-D images. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1311-1319).

研究生阶段研究成果

一、奖项

2019 年第十五届“挑战杯”广东大学生课外学术科技作品竞赛三等奖

二、论文

- Automated Steel Bar Counting and Center Localization with Convolutional Neural Networks. IEEE Transcations on Industrial Informatics, Under Review. https://arxiv.org/abs/1906.00891
- Retinal Vessel Segmentation via Octave Convolution Neural Network. IEEE Transcations on Medical Imaging, Under Review. http://arxiv.org/abs/1906.12193
- Design and Implementation of Mobile Manipulator System. IEEE-CYBER 2019.（已接收）
- Object Sorting Using a Global Texture-Shape 3D Feature Descriptor. International Journal of Advanced Robotic Systems, Under Review. https://arxiv.org/abs/1802.01116

三、专利

- 一种基于深度卷积神经网络的钢筋端面识别方法，专利申请号：201811618063.8
- 一种基于卷积神经网络的电厂电表字符定位和识别方法，专利申请号：201910316734.3
- 一种基于深度学习的智能抓取系统，专利申请号：201810801897.6
- 一种复合型移动机器人（发明），专利申请号：201810780569.2
- 一种复合型移动机器人（实用新型），专利申请号：201821125302.1
- 一种复合型移动机器人及复合型移动机器人控制系统（发明），专利申请号：201810777333.3
- 一种复合型移动机器人及复合型移动机器人控制系统（实用新型），专利申请号：201821125759.2
- 一种构建三维地图的方法，专利申请号：201810809721.5
- 一种基于二维码的移动机器人导航方法，专利申请号：201810809736.1

致谢

本论文是在范衡老师的悉心指导下完成的。范衡老师作为一名优秀的、经验丰富的教师，具有丰富的数字图像处理知识和经验，在整个论文实验和论文写作过程中，对我进行了耐心的指导和帮助，提出严格要求，引导我不断开阔思路，为我答疑解惑，鼓励我大胆创新，使我在这一段宝贵的时光中，既增长了知识、开阔了视野、锻炼了心态，又培养了良好的实验习惯和科研精神。在此，我向我的指导老师表示最诚挚的谢意！