基于 RGB-D 图像的语义分割方法综述

王晨[1)], 杜晨曦[1)], 刘瑞军[2)*], 齐越[3)]

[1)] (北京工商大学计算机与人工智能学院 北京 100048)
[2)] (北京航空航天大学软件学院 北京 100191)
[3)] (北京航空航天大学虚拟现实技术与系统全国重点实验室 北京 100191)
(liuruijun@buaa.edu.cn)

摘 要: 语义分割技术致力于精确识别并分割图像中的各个物体或场景. 基于 RGB 图像的方法在信息利用上存在局限, 导致性能受限, 随着深度传感器技术的普及, 深度图的引入为语义分割网络注入了丰富的几何信息, 显著地提升了分割精度. 文中介绍了近几年基于 RGB-D 图像的语义分割方面的显著进展和相关方法, 根据对多模态融合特征处理方式的差异, 将基于 RGB-D 图像的语义分割方法归纳为单分支、双分支、三分支网络架构 3 大类. 其中, 单分支网络在同一分支同时处理 RGB 和深度特征, 实现特征的有机结合; 双分支网络利用 RGB 和深度特征之间的互补性, 优化多模态特征的校正与融合; 三分支网络在保留原始的 RGB 和深度特征的同时, 深入挖掘融合特征, 确保信息的全面性. 同时, 总结注意力、模型优化等关键技术, 并归纳常用的数据集和评价指标, 对比分析各种方法在不同数据集上的性能, 最后总结当前 RGB-D 图像语义分割在多模态数据交互与处理方面所面临的挑战, 展望了语义分割技术在跨领域数据融合方向的发展前景.

关键词: RGB-D 图像; 语义分割; 多模态特征融合; 卷积神经网络
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024-00107

A Review of Semantic Segmentation Methods Based on RGB-D Images

Wang Chen[1)], Du Chenxi[1)], Liu Ruijun[2)*], and Qi Yue[3)]

[1)] (*School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing* 100048)
[2)] (*School of Software, Beihang University, Beijing* 100191)
[3)] (*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing* 100191)

Abstract: Semantic segmentation technology is dedicated to accurately identify and segment each object or scene in an image. With the popularity of depth sensor technology, the introduction of depth map has injected rich geometric information into the semantic segmentation network and substantially enhanced the precision of segmentation processes. The paper introduces the significant progress and related methods in the field of semantic segmentation based on RGB-D images in recent years. According to the differences in the processing of multimodal fusion features, the semantic segmentation methods based on RGB-D images are categorized into three categories, namely, single-branch, dual-branch, and three-branch network architectures. Among them, the single-branch network processes both RGB and depth features in the same branch to achieve the organic combination of features; the dual-branch network optimises the correction and fusion of multimodal features by taking advantage of the complementarity between RGB and depth features; and

收稿日期: 2024-03-01; 修回日期: 2024-08-14. **基金项目**: 国家自然科学基金青年基金(62201018); 新一代人工智能国家科技重大专项(2022ZD0119502); 北京市自然科学基金-海淀原始创新联合基金(L222052). **王晨**(1990—), 女, 博士, 讲师, 硕士生导师, CCF 会员, 主要研究方向为虚拟现实、医学图像处理; **杜晨曦**(2000—), 女, 硕士研究生, 主要研究方向为计算机视觉; **刘瑞军**(1982—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 论文通信作者, 主要研究方向为虚拟现实、计算机视觉; **齐越**(1969—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究方向为虚拟现实、计算机图形学.

the three-branch network explores the fusion features in depth while retaining the original RGB and depth features to ensure the comprehensiveness of information. Additionally, the paper summarizes the key techniques such as attention mechanisms and model optimization, and reviews commonly used datasets and evaluation metrics. It compares and analyzes the performance of various methods on different datasets. Finally, the paper summarizes the current challenges of RGB-D image semantic segmentation in the interaction and processing of multimodal data. It also looks forward to the development prospects of semantic segmentation technology in the direction of cross-domain data fusion.

Key words: RGB-D images; semantic segmentation; multimodal feature fusion; convolutional neural network

语义分割作为计算机视觉方向的一项重要任务，旨在将图像中的每个像素分类到不同的语义类别中. 与传统的图像分类任务不同，语义分割需要对图像进行像素级别的分类，使得每个像素都被正确地赋予对应的语义标签. 随着智能化生活的推进，语义分割技术在自动驾驶、虚拟交互、增强现实中发挥着重要的作用.

传统的语义分割方法，如阈值分割法[1]、基于边缘检测的分割方法[2]、基于区域生长的分割方法和基于马尔可夫随机场模型[3]，在一定程度上能够完成像素级别的分类任务，但需要大量的手工设计和复杂的调整，而且对于复杂场景和目标形状不规则的图像效果较差. 随着深度学习技术的发展，基于卷积神经网络(convolutional neural network，CNN)的语义分割已经成为主流的方法，通过端到端的训练，CNN 可以自动地学习图像中的特征表示[4]，并将其应用在语义分割任务中. 其中一些经典的网络架构，如 FCN(fully convolutional networks)[5]，U-Net[6]和 SegNet[7]等方法被广泛应用，并且在性能和泛化能力方面得到显著的提升.

随着深度传感器的普及，采集的图像信息从彩色信息扩展到 RGB-D 信息[8]. 与传统的语义分割大多只考虑 RGB 图像信息相比，RGB-D 图像语义分割引入包含几何信息的深度图像，与 RGB 图像中的纹理信息进行互补. 在光照变化和遮挡等情况下，RGB 图像特征可能会受到影响，导致难以准确地识别物体. 相比之下，深度数据对光照变化和遮挡具有较强的鲁棒性，将深度数据与 2D 视觉信息结合，通过提供 3D 几何形状信息使 RGB 图像信号得到补充，能够更好地区分不同物体.

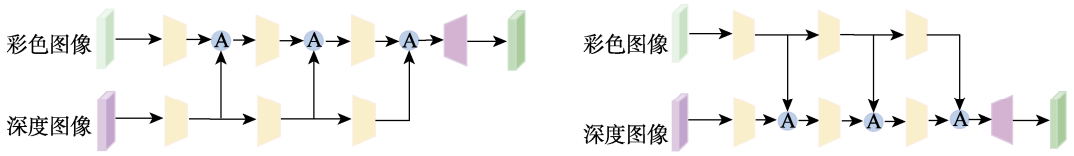
将深度信息整合到 RGB 分割架构中是一项充满挑战的任务，需要解决数据不匹配、模态间差异和复杂场景等难题，以实现深度信息与 RGB 信息的精确且高效的融合. 早期的方法[9-10]通过简单地将 RGB 图像和深度图像的特征拼接起来，增强每个像素的语义信息. 对于深度图像，Gupta 等[11]对其进行水平视差、离地高度、重力与表面法线的夹角(horizontal disparity, height above ground, angle with gravity, HHA)编码，将深度信息编码为 HHA3 个通道，这种编码方式比单纯使用深度通道有了明显的改进. 还有一些方法如 ESANet[12]和 DCANet[13]使用双分支 CNN 作为编码器，分别从 RGB 图像和深度图像中提取特征，将多个阶段提取的特征进行融合. 对于 RGB 图像和深度图像之间存在的模态差异，如何确保网络充分地利用这 2 种模态的信息进行有效的融合，仍然是一个需要解决的问题.

已有的 RGB-D 语义分割的综述文献[14-16]主要从传统机器学习和深度学习角度进行分析，而针对方法中对多模态融合特征的处理方式缺乏详细的对比介绍. 本文将网络架构划分为单分支网络、双分支网络和三分支网络 3 类，对其发展情况进行介绍. 在梳理各分支重要的语义分割方法及最新研究成果的同时，本文还对 RGB-D 语义分割方法中常用的关键技术进行介绍. 此外，本文还对常用的公共数据集以及各种方法在这些数据集上的实验结果进行分析对比. 最后讨论该领域的研究价值、未来发展方向以及亟待解决的问题，对这些问题的深入讨论，有助于为进一步推动 RGB-D 语义分割方法的发展提供有益的思路和指导.

1　基于单分支网络的 RGB-D 语义分割方法

如图 1 所示，在单分支网络中，通常以 RGB 分支为主，专注提取图像的颜色和纹理特征，深度分支捕获的特征作为辅助信息，用于补充 RGB 信息并实现对细节分割的增强，如图 1a 所示；还有一些网络选择将深度分支作为主导，进而将 RGB

特征传播至深度分支中，实现对深度信息进行进一步补充和细化，如图 1b 所示. 单分支使用融合模块将每层网络提取到的单侧特征传播到另一分支中，实现深度特征与 RGB 特征的有机结合，融合后的特征被馈送到下一级网络中进行更高层次的特征提取，实现更精确的语义分割结果.



a. 单分支网络结构1　　b. 单分支网络结构2

图 1　单分支网络结构

在 FuseNet[17]和 RedNet[18]中，RGB 图像和深度图像分别被馈送到 2 个独立的神经网络分支中，在上采样之前进行融合. FuseNet[17]采用编码器-解码器类型的网络结构，包括 Dense Fusion 和 Sparse Fusion 这 2 种融合策略，将深度特征融合到 RGB 分支中，增强了特征的表达能力. RedNet[18]使用残差模块作为编码器-解码器的基本块，先使用元素求和的方法进行 RGB 图像和深度图像融合，再将融合结果通过跳跃连接将编码器的空间信息传递到解码器中，这些结构改善了网络深度增加可能导致的模型退化问题，并且应用多层监督学习改善深层网络训练过程中的梯度消失问题，提高网络性能. 在融合层，FuseNet 和 RedNet 只是使用简单的逐元素相加将深度信息添加到 RGB 分支，并没有使深度信息得到充分的利用.

现有方法通过多阶段融合策略传播深度特征，但在深层网络阶段，使用简单元素相加的传播方式不能充分地利用深度信息，随着网络加深，模型对 RGB 图像中语义信息的理解和表达会变得更加明显和准确. 但是，连续的卷积和下采样操作会导致 2 种模态特征无法完全对齐，深度特征无法为 RGB 特征提供准确的几何信息. 因此，GLPNet[19]在深层网络引入局部上下文融合模块(local context fusion module，L-CFM)和全局上下文融合模块(global context fusion module，G-CFM). 其中，L-CFM 在元素融合之前先动态地对齐 2 种模态的特征，通过预测的偏移量对特征图进行微调，实现更精准地对齐；G-CFM 通过联合建模多模态全局上下文特征将深度特征传播到 RGB 分支中，从全局角度理解并利用深度信息. GLPNet 在解决特征对齐和语义传播方面取得了一定进展，但在网络的计算效率和超参数方面存在一些局限性，需要通过算法优化实现更高效的计算性能.

GLPNet 只在最后一层网络提取全局和局部信息，对于浅层网络仅使用逐元素相加的方式将深度信息传播到 RGB 分支，浅层网络提取到的特征没有得到充分的利用. RFNet[20]在编码器的每个阶段都引入注意特征互补(attention feature complementary，AFC)模块，应用注意力机制将深度特征融合到 RGB 分支. 在 AFC 模块中，对 RGB 特征和深度特征各自使用 SE(squeeze and excitation)模块实现特征增强[21]，SE 模块通过重新加权 2 种模式的特征对元素进行求和，模型根据给定的输入增强或抑制某种特征；然后将 2 种特征加在一起输入到 RGB 分支中，使 RGB 特征与深度特征得到充分、有效的融合，提高分割精度；在上采样之前，使用空间金字塔池化(spatial pyramid pooling，SPP)[22]将 2 个分支融合的 RGB-D 特征聚集在一起，生成具有多尺度信息的特征图，这种高效的网络架构能够实现在保持实时速度的同时增加感知范围，以覆盖大型物体的像素，并且网络使用多源学习策略进行数据泛化，增强了模型对意外小障碍物的检测能力，丰富了模型面对现实世界中不可预见的危险所需的可识别类，但对于更复杂的场景，模型的性能可能会受到影响.

ESANet[12]采用 NBt1D(non-bottleneck-1D-block)模块[23]，将 ResNet 每层中的 3×3 卷积替换为一个 3×1 卷积和一个 1×3 卷积，并在中间添加 ReLU 激活函数层；在编码器的多个阶段，使用 SE 模块将深度信息融合到 RGB 分支中，同时在上下文模块中，采用类似于 PSPNet 中的金字塔池化模块(pyramid pooling module，PPM)[24]的设计，使用多个分支聚合不同尺度的特征信息，获取全局上下文信息. 实验证明，这些高度优化的网络架构减少了网络的推理时间并显著地提升了分割性能，在移动机器人上能够实现高效场景分析，但在计算资源、硬件兼容性和环境泛化能力方面，网络仍需进行深入研究和改进. ESANet 的结构如图 2 所示.

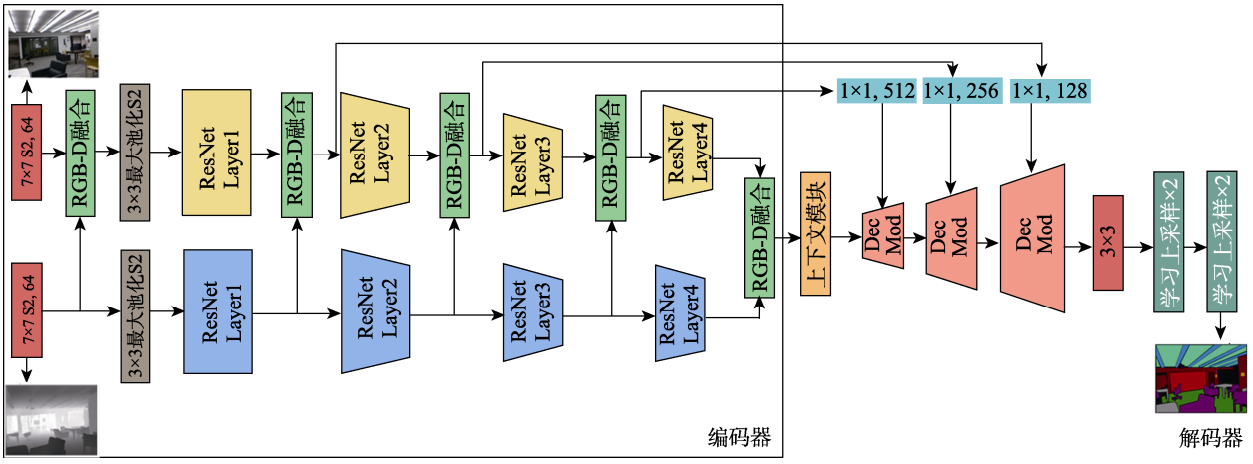


图 2　ESANet 结构[12]

Zhang 等[25]设计一种轻量级编码器-解码器网络 SGACNet，在减少计算参数的同时，保证了算法的鲁棒性. 在编码器部分，使用通道和空间注意融合模块(attention fusion module, AFM)选择性地将深度特征与 RGB 特征融合，实现 RGB 与深度信息的特征增强，利用融合的信息引导下一层的特征提取，有效地捕获了多级 RGB-D 特征. 为了充分地利用编码器的全局信息，引入自适应金字塔上下文(adaptive pyramid context, APC)模块，利用全局引导的局部亲和力扩大感受野，有助于捕捉不同类别之间近邻区域的共享特征. 这种高效的轻量级网络架构利用跨模态特征之间相关性和互补线索，有效地改善了深度数据噪声的不利影响，在保持较快推理速度的同时，具有较高的分割精度，更加适用于移动机器人的场景分析.

如图 3 所示，Du 等[26]提出一种高精度实时语义分割网络 AsymFormer，使用 ConvNeXt[27]进行 RGB 特征提取，同时使用轻量级的 Mix-Transformer 处理深度特征. 这种非对称的网络设计减少冗余参数，实现了网络复杂性和推理速度的优化平衡. 为了实现多模态特征的高效融合，AsymFormer 引入局部注意引导多模态特征选择(local attention-guided feature selection, LAFS)模块，采用可学习的策略提取全局信息，并深入挖掘多模态数据间的依赖性，在空间和通道 2 个维度上进行选择性的特征融合；为了进一步增强特征的表达能力，还设计了跨模态融合(cross modal attention-guided feature correlation embedding, CMA)模块，旨在从融合的特征中提炼更有价值的信息，进一步提取和整合跨模态特征，提升网络对跨模态特征的理解，使网络具有更深层次的语义理解能力.

2　基于双分支网络的 RGB-D 语义分割方法

双分支网络采用与单分支网络相同的网络结构，包含 2 个独立的 CNN 分支，分别用来提取对应的 RGB 特征和深度特征信息，但在特征融合层面更加强调不同模态信息互补的重要性，其架构如图 4 所示. RGB 分支和深度分支都被视为网络的主要分支，并且在相同尺度下进行特征提取，通过

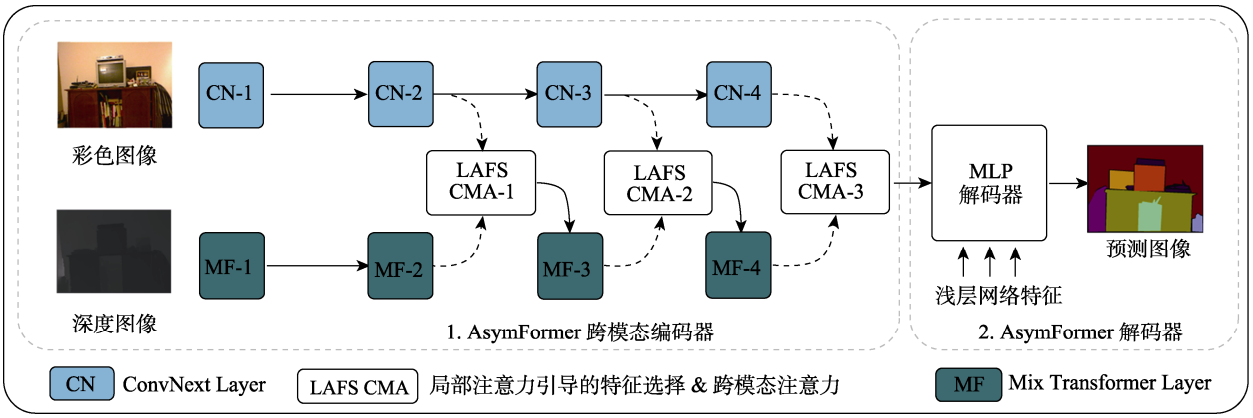


图 3　AsymFormer 网络结构[26]

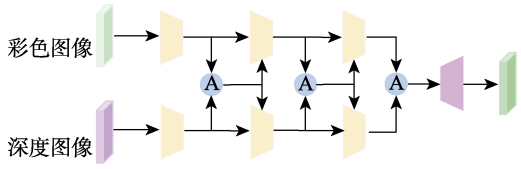


图 4 双分支网络结构

特征融合或特征交互模块将 2 个分支提取的特征进行融合，实现更准确的语义表示；然后将融合后的特征送入各自的分支中进一步提取特征．这种双分支结构能够充分地利用 RGB 和深度信息之间的互补性，实现特征校准和特征融合，提高语义分割的精度和鲁棒性．

由于各种因素的干扰，RGB 和深度数据中的噪声是不可避免的，而且随着网络的深入，数据的准确性逐渐降低．因此基于这一情况，Chen 等[28]提出一种即插即用的 SA-Gate 模块，有效地利用编码器阶段跨模态的互补信息，并且可以将其灵活地集成到其他语义分割网络中．为了保证信息特征在模态之间有效传播，SA-Gate 由特征校准和特征聚合 2 个部分组成，通过对称和双向的方式实现 RGB 特征和深度特征的相互校准，过滤深度数据中的噪声，并在空间维度上对跨模态特征进行互补聚合；然后使用双向多步传播(bi-directional multi-step propagation, BMP)模块传播和融合多模态特征，同时保持它们的特性．SA-Gate 在室内和室外数据集上均表现出良好的性能，但对于复杂环境中深度测量产生的噪声仍需要进一步优化，以应对更严重的噪声条件．

随着 Transformer 的出现[29]，ViT(vision Transformer)[30]将其应用到图像任务中，SETR[31]用 ViT 代替 CNN 的主干网络，将语义分割转化为序列到序列的预测任务，得到了更好的特征表示．CMX[32]和 CMFormer[33]正是基于 Transformer 改进的 RGB-D 语义分割方法，其中，CMX 选择 Swin-Transformer 和 SegFormer 作为骨干网络．为了实现网络对多模态特征更深层的理解和利用，CMX 和 CMFormer 中设计了 2 种模块，分别实现特征校准和特征聚合．在 CMX 中，跨模态特征校准模块(cross-modal feature rectification module, CM-FRM)从特征映射角度，通过在空间和通道 2 个维度上结合其他模态的特征来校准当前模态的特征，有效地减少了噪声的不利影响；并在获取每层的特征映射后，应用一种两阶段特征融合模块(feature fusion module, FFM)增强信息交互，其中，第 1 阶段使用交叉注意机制实现全局信息交换，第 2 阶段实现特征聚合．CMFormer 网络[33]结构如图 5 所示，其中，跨模态多头自注意(cross-modal multi-head self-attention, CM-MHSA)模块对 Transformer 中的多头注意模块进行修改，通过交换查询向量充分地利用 2 种模态的特征，多尺度信道注意校正(multi-scale channel attention correction, MS-CAC)和全局特征聚合(global feature aggregation, GFA)都是基于 CM-MHSA 模块实现特征校准和融合，这些结构有效地改善了多尺度融合和实时性的问题．CMX 和 CMFormer 在特定数据集上都表现良好，但由于 Transformer 这类网络需要更多的参数来进行训练和推理，CMX 和 CMFormer 对计算资源的要求较高，网络仍需探索计算效率与性能的平衡性．

基于 RGB 和深度数据的差异性，为了实现多模态信息的挖掘和整合，NANet[34]和 DCANet[13]中设计了不同角度的融合模块．NANet 中设计了多模

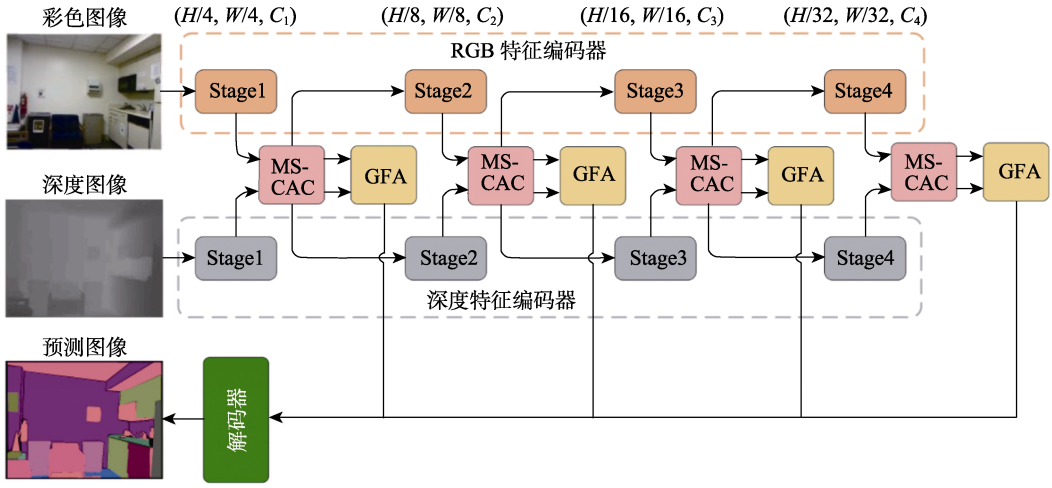


图 5 CMFormer 网络结构[33]

态非局部聚合模块(multi-modality non-local aggregation module, MNAM)，首先通过空间融合模块(spatial fusion module, SFM)在空间维度上建模 RGB-D 远程依赖关系，然后通过通道融合模块(channel fusion module, CFM)沿通道维度建模跨模态上下文依赖关系，以便更充分地利用多个阶段 RGB-D 特征的非局部上下文信息. DCANet 则是对深度特征和 RGB 特征分别使用像素微分卷积注意(differential convolution attention, DCA)和集成微分卷积注意(ensemble differential convolution attention, EDCA)这 2 种即插即用的模块，实现特征增强，然后进行特征聚合. 其中，DCA 融合了局部几何信息，使模型更好地理解数据中的局部特征；EDCA 使用分解的大核卷积，充分地利用 RGB 数据的空间分布特征，有效地传播远程上下文依赖关系，提高模型对全局信息的感知能力. DCA 与 EDCA 对于 RGB 和深度数据的适配性也证明了深度图在局部区域内具有深度连续性，更适合提供物体的几何信息，而 RGB 图像拥有丰富的色彩信息，可以提供全局视图；且 DCA 和 EDCA 的应用减少了计算量，在保证网络分割性能的同时也减少了计算资源的使用.

基于 RGB 和深度数据的互补性, Zhou 等[35]提出 BCINet，包括 3 种模块：(1) 双边跨模态交互模块(bilateral cross-modal interaction module, BCIM)中引入 FEM 实现单侧单模态的特征增强，然后交叉加入到另一侧的对应特征中，以捕捉跨模态的互补信息；(2) 混合金字塔扩展卷积(hybrid pyramid dilated convolution, HPDC)模块在垂直和水平空间方向上扩大感受野，自适应地捕捉不同形状下的环境信息；(3) 上下文引导模块将不同阶段的高级上下文信息与低级特征相结合，实现对分割图的分层逐步细化. 实验结果表明，这些结构能使 BCINet 在室内环境中实现更精准的场景理解，特别是对于一些外观相似的类别，其展现出了卓越的识别性能.

如图 6 所示，HaarNet[36]采用一种结合数学形态学和传统线性网络的方法，在利用深度数据的几何特性的同时保持 RGB 数据的线性特征. HaarNet 使用形态学上采样从低分辨率特征中获取更多信息，并且使用形态学 Haar 小波(morphological Haar wavelet, MHW)模块改善下采样信号，优化 2 种模态特征信号中的高频细节，实现对边界细节的识别. HaarNet 将 ReLU[37]看作形态扩张的一种特殊情况，并且把网络中所有的激活层替换为形态学扩张 FReLU[38]，以提高网络对几何信号的处理能力. 实验结果表明，HaarNet 在室内和室外数据集上均表现优异，并且使用更少的参数实现了卓越的性能，但其对于深度数据质量要求较高，在低质量的深度信号上性能可能受影响.

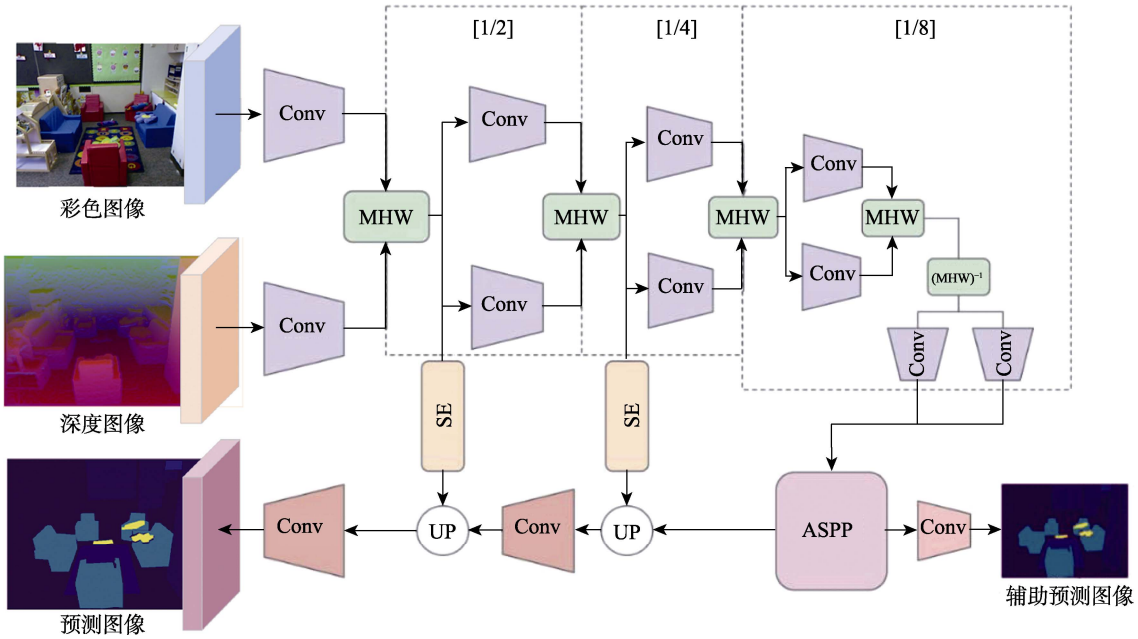
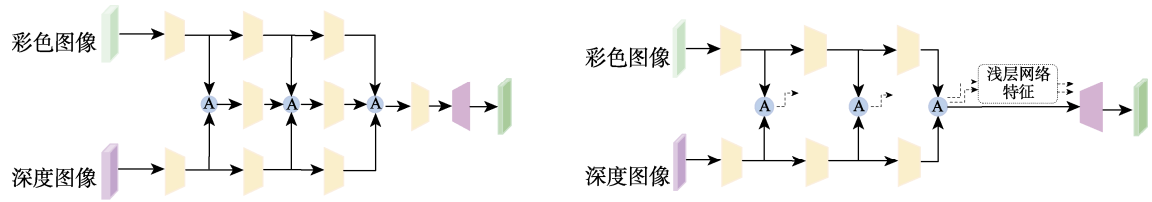


图 6　HaarNet 结构[36]

3 基于三分支网络的 RGB-D 语义分割方法

三分支网络包含 2 类网络结构. 第 1 类网络由 3 个并行的 CNN 分支组成, 分别用于提取 RGB 特征、深度特征和融合特征, 如图 7a 所示, 其中, 融合特征是通过将 RGB 和深度特征进行适当的融合得到的, 而且融合后的特征仅在第 3 分支进行后续的特征提取, 不输入到 RGB 分支和深度分支中. 第 2 类网络由 2 个分支组成, 分别用于提取 RGB 和深度特征, 如图 7b 所示, 与第 1 类网络结构不同的是, 第 3 个分支并非 CNN, 而是对融合后的特征进行额外的特征增强或其他处理操作. 三分支网络旨在保留原始的 RGB 特征和深度特征的同时, 充分地利用融合特征, 从而在整个网络中实现更全面、更准确的语义分割.

如图 8 所示, Hu 等[39]提出一种具有 3 个并行分支的特征提取网络 ACNet, 使用基于 ResNet 的 2 个独立的分支分别提取 RGB 和深度特征, 由于 RGB 图像和深度图像在不同场景中的特征分布具有较大差异, 因此提出一种基于通道注意力的注意互补模块(attention complementary module, ACM), 将每层提取到的特征馈送到 ACM 中实现各模态特征的动态增强, 再将 2 个特征进行逐元素相加后输入到另一个基于 ResNet 的分支中来处理融合的特征, 并在每层上采样之前保留 RGB 和深度特征. 这种结构既能保持原始 RGB-D 特征的流动, 又能有效地整合和传播已融合的特征信息, 提高了特征的利用效率. 与 ACNet 相似, TCANet[40]同样使用 3 个并行 CNN 提取 RGB 特征、深度特征和融合特征, 但 ACNet 中的多模态融合方式比较简单, 没有充分利用 RGB 和深度信息; 而 TCANet 中引入了多模态 FFM(multi-modal FFM, MFF), 设计坐标注意力(coordinate attention, CA)机制来提取 2 种模态的互补信息, 在空间和通道维度上实现 2 种模态信息的聚合, 使网络能够获取丰富的上下文信息, 并将更多的注意力集中在信息丰富的区域, 提高模型对关键信息的感知能力. TCANet 中还嵌入了空洞空间金字塔池化(atrous spatial pyramid pooling, ASPP)模块优化网络中的语义信息, 使网络更好地整合多尺度上下文信息. 该网络虽然有较好的结果, 但较高的参数需求使得其推理时间较长, 在实际部署时可能会有局限性, 尤其是在需要快速推理的场景. CANet[41]也通过 3 个分支引入 RGB 特征、深度特征和融合特征, 采用共同注意融合的方法增强潜在特征, 设计了位置共同注意融合模块(position co-attention fusion module, PCFM)和通道共同注意融合模块(channel co-attention fusion module, CCFM), 分别在空间和



a. 三分支网络结构1　　b. 三分支网络结构2

图 7　三分支网络结构

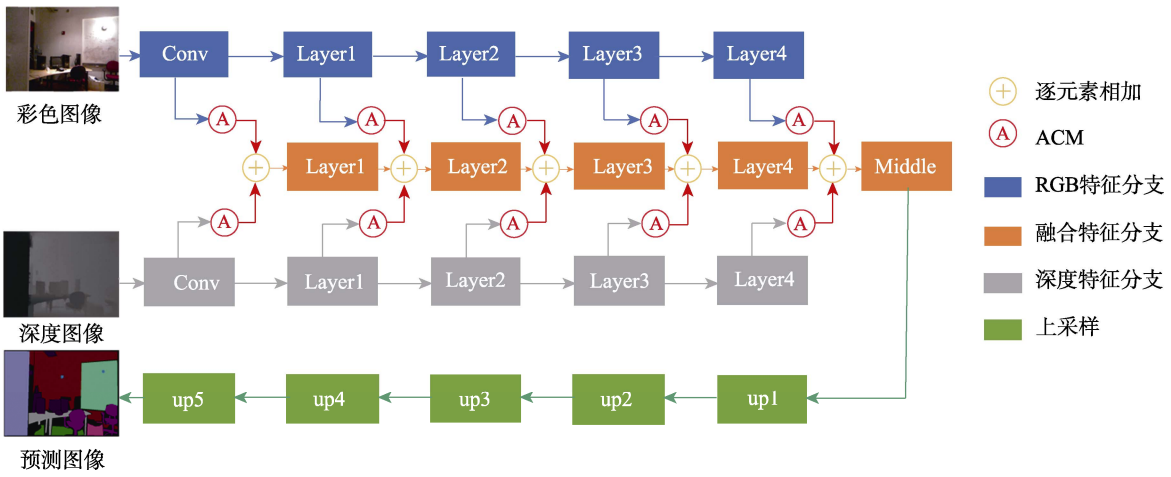


图 8　ACNet 结构[39]

通道维度上利用共同注意机制有效地融合 RGB 和深度特征，克服数据之间的差异，提高特征表达能力；并引入最终融合模块集成 PCFM，CCFM 和融合支路的输出，生成更具有代表性的最终特征．在训练过程中，CANet 应用多尺度损失函数，有效地提升网络对多尺度特征的深入理解和学习能力，使模型更加精准地捕捉和处理细节信息及全局信息，实现更精准的分割．

TSNet[42]使用 2 个非对称编码器结构和一个包含自注意模块的跨模态蒸馏分支：随着网络的加深，因为深度图像中包含的边缘信息会更容易丢失，所以使用浅编码器 VGGNet16 专门提取深度图像的边缘细节，而 RGB 图像则应用深编码器 ResNet34 进行特征的深入挖掘；跨模态蒸馏分支利用自注意力机制整合和提炼来自 2 个非对称输入分支的语义信息；最后融合 3 个分支输出的特征信息，并利用卷积层和 ASPP 进一步优化融合特征．此外，TSNet 还在网络中引入 2 个特征输出分支实现对多尺度特征图的监督，提升网络在多尺度特征学习的精度，但由于其所需参数量相对较高，实际应用时可能会受到限制．

RDFNet[43]中提出 MMF 模块，用于自适应地训练残差特征，有效地融合不同模态的互补特征，同时学习每个模态特征的相对重要性；进一步，将融合的特征输入到 RefineNet 中，实现高分辨率的语义分割．但是，MMF 仅通过卷积和简单的叠加实现模态融合，没有充分地考虑 2 种模态之间的互补性以及不同模态特征之间的重要性差异．如图 9 所示，CMPFFNet[44]中提出一种基于注意机制的多模态自适应对齐融合(multimodal adaptive alignment fusion, MAAF)模块，通过其中的交叉通道注意(cross channel attention, CCA)模块，利用加性注意对 2 种模式的通道维度进行全局调整，并将深度特征的空间信息作为线索调整 RGB 特征的权重，增强 RGB 特征的表达能力，实现 2 种模态的融合；引入反向注意增强(reverse attention augmentation, RAA)模块，将低层特征的具体语义信息传递给抽象的高层特征，提高高层特征的表达能力，并将其送入多层特征渐进融合(multilevel feature progressive fusion, MFPF)模块中进行最终融合，实现对多尺度特征全局信息的整合，提高分割的准确性．Zhou 等[45]在 FRNet 中提出一种新颖的特征构建编码器(feature construction encoder, FCE)，通过反向传播实现多层特征信息的逐层更新，并将特征输入相邻层网络，不断地调整和优化特征表示，保持了信息的连续性和完整性；还引入跨模态感知模块(cross-modality awareness module, CAM)对跨模态信息进行整合，生成具有丰富语义信息的特征节点，获取更全面的上下文信息；同时，引入交叉层次丰富模块(cross-level enriching module, CEM)重建相邻下层的基本特征，获得新的多层特征表示，提高网络的分割能力．

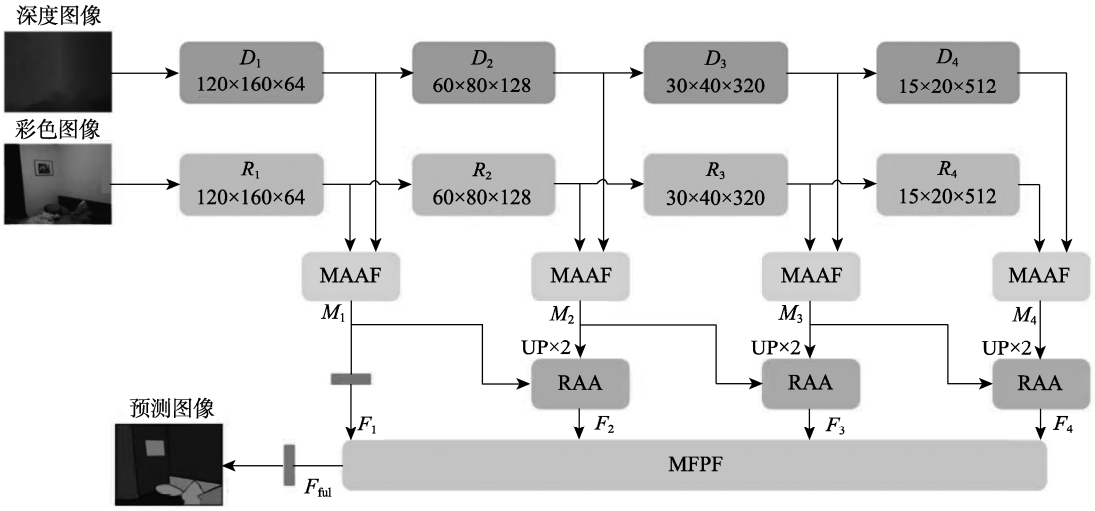


图 9　CMPFFNet 结构[44]

4　关键技术

4.1　注意力机制

通过在 RGB-D 语义分割方法中引入注意力机制，使模型能够更加关注对于语义分割任务更重要的区域，提取更有效的特征信息．在语义分割任务中，全局上下文信息对于正确的像素分类是至关重要的．注意力机制有助于整合全局上下文信息，使

模型能够更好地理解整个场景中的语义关系，提升语义分割的准确性. 单尺度下，注意力机制在融合 RGB 和深度特征的全局特征中发挥着重要作用，能够动态地调整 RGB 和深度特征之间的权重，优化 RGB 与深度图像的融合方式，使得模型能够更准确地理解并利用不同特征之间的关系，实现更有效的特征融合. 这种融合方式能够提高模型对于关键信息的感知能力，从而提升语义分割任务的性能和效果.

一种常见的注意力机制是空间注意力机制，它通过为特征图的不同空间位置分配权重来关注不同的区域，使得网络可以更加关注与目标对象相关的区域，同时抑制无关的背景信息，可以有效地提高语义分割的精度. Google DeepMind 团队提出的 STN(spatial transformer network)[46]具有空间不变性，通过学习输入数据的形变信息，完成适合特定任务的预处理操作.

另一种常见的注意力机制是通道注意力机制，它通过为不同通道的特征图分配权重来调整每个通道的重要性. Hu 等[21]提出即插即用的 SENet 模块，其结构如图 10 所示，从特征通道层面上统计图像的全局信息，通过自适应地学习每个通道的权重，调整每个通道的特征在后续网络层中的重要性，从而提升神经网络的表达能力和性能，使网络能够更有效地学习和利用特征通道之间的相关性.

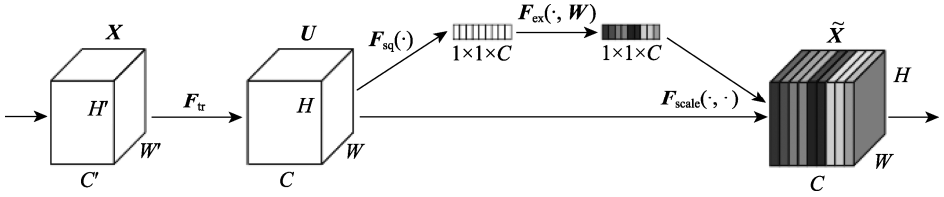


图 10 SENet 结构[21]

ESANet[12]在多个阶段使用 SE 模块，将深度数据编码器中提取到的深度信息融合到 RGB 分支；RFNet[20]则在编码器每个阶段，对 RGB 特征和深度特征各自使用 SE 模块实现特征增强. ACNet[39]应用一种基于通道注意力的 ACM 接收每层提取到的特征，并使用通道注意力机制实现对各种模态特征的增强. 在网络中引入通道注意力可以自适应地选择和增强对于语义分割任务关键的通道特征，提高语义信息的表达能力. 但是，仅依赖通道注意或空间注意中的一种注意机制的处理方法相对单一，导致网络性能提升有限，所以应同时输入特征的通道信息和空间信息. Woo 等[47]提出卷积块注意模块(convolutional block attention module, CBAM)，将通道注意力和空间注意力通过级联及并行的方式进行融合，有效地提升了特征提取能力. SGACNet[25]中的融合模块采用 2 种注意机制过滤掉不重要的特征，并将这 2 种机制融合在一起，能够更全面地挖掘特征信息，提高模型特征提取的准确性和泛化能力.

自注意力与交叉注意力在语义分割任务中也扮演着重要的角色. 随着 Transformer 的发展，自注意力机制[29]作为 Transformer 架构的核心组件，逐渐成为深度学习方向的研究热点，尤其在计算机视觉方向得到广泛应用和深入研究. Segformer[48]和 Swin-Transformer[49]在 Transformer 基础上进行优化，实现了更高效的自注意力机制. CMX[32]正是将 Segformer 与 Swin-Transformer 作为骨干网络，实现更高效的特征提取.

如图 11 所示，自注意力机制关注单个序列内部不同位置之间的相互联系，而交叉注意力机制更关注不同序列之间的联系[50-51]，更适用于多模态的数据处理，使不同模态序列进行深度交互，以实现跨模态信息有效融合，其计算过程如图 12 所示. CMX[32]和 CMFormer[33]中，应用交叉注意力促进 RGB 和深度特征之间的有效融合，实现多模态信息的深度交互.

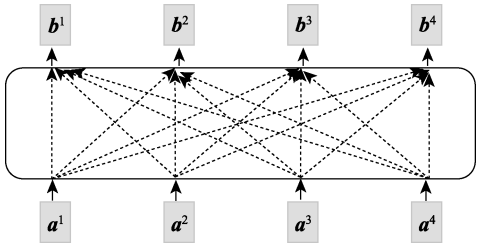


图 11 自注意力机制计算示意图

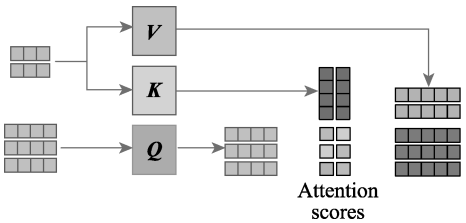


图 12 交叉注意力机制计算示意图

在 RGB-D 语义分割网络中，还有其他多样化的注意力机制不断被提出，这些先进的注意力机制能够显著地提升网络对图像信息的识别和分析能力. 在实际应用中，合理地设计和引入不同类型的注意力机制，可以进一步提升语义分割网络的性能.

4.2　金字塔网络

金字塔网络在 RGB-D 语义分割方法的多尺度特征提取方面发挥着重要作用，通过该网络可以获取图像的细节和全局语义信息，使网络能够更好地理解和表达不同尺度下的特征信息.

PSPNet[24]中设计了 PPM，用于改进全卷积网络中无法有效地处理场景之间的关系及全局信息的情况. PSPNet 通过 PPM 聚合不同区域的上下文信息，从而获取全局上下文信息. ESANet[12]在类似于 PPM 的上下文模块中，使用多个分支聚合不同尺度的特征信息，以获取上下文信息.

Yu 等[52]提出空洞卷积，在不增加网络参数和计算量的情况下增加每个卷积层的感受野，从而提取更多的上下文信息. 在传统的 CNN 中，单一尺度的感受野较难捕获物体的多尺度特征. 基于金字塔池化和空洞卷积，DeeplabV2[53]中提出 ASPP 模块，其结构如图 13 所示，通过引入多尺度的空洞卷积和池化操作，解决了语义分割任务中物体尺度变化和全局上下文信息获取的问题，生成多尺度的、丰富的语义信息表示. TCANet[40]在将特征信息输入到解码器之前，引入 ASPP 模块优化网络中的语义信息，捕获多尺度上下文信息. TSNet[42]在跨模态蒸馏操作中加入 ASPP 模块，对特征信息进行整合提炼，图像分割性能得到显著的提升，尤其是对于大尺度物体和多尺度物体的识别，以及在复杂背景下的适应能力. 类似于 APCNet[54]中的自适应上下文模块，SGACNet[25]中引入 APC 模块，利用全局引导的局部亲和力扩大感受野，有助于理解不同类别近邻域之间的共享特征.

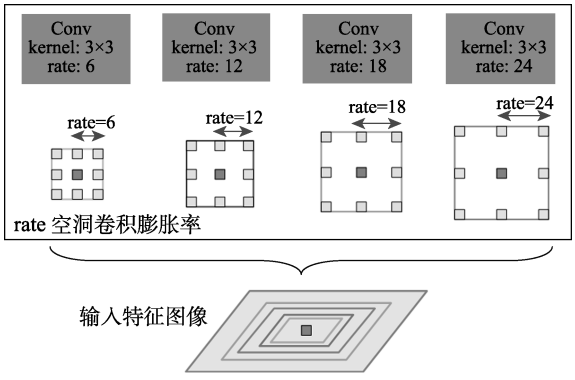


图 13　ASPP 模块结构[53]

金字塔网络通过多尺度分析捕获图像中的空间信息，进一步提高分割的精度；同时，该网络的设计能够充分利用现有的计算资源，提高分割效率.

4.3　跳跃连接

在 RGB-D 图像语义分割任务中，跳跃连接是一种常用的技术[55]，它将输入数据直接添加到网络某一层输出之上，旨在融合来自不同层级的特征信息. 这种设计促进了信息流动，并且保留了原始输入数据中的细节和语义信息，避免信息丢失.

通常，跳跃连接通过求和操作或拼接操作实现. 残差网络[56]通过引入残差块构建网络，使用跳跃连接将输入直接添加到层输出上. 卷积网络使用串联实现跳跃连接，确保特征的可重用性. 图 14 所示为在 U-Net[6]架构中应用跳跃连接结构的情况. 在实际应用中，跳跃连接通常与特征金字塔网络(feature pyramid networks, FPN)[57]等结构结合使用，可以提升网络在不同尺度下的感知和识别能力，并且在网络训练过程中降低计算成本.

值得注意的是，在不同网络中跳跃连接的数量和位置有所不同. SA-Gate[28]和 GLPNet[19]仅将浅层网络的融合信息输入到解码器中，为解码器提供更多细节信息. 相比之下，CMX[32]，CMFormer[33]等模型则将每层的融合信息都传递到解码器中，利用多尺度信息以实现更精确的分割. 因此，在应用跳跃连接技术时，需要根据不同的任务和数据集选择合适的数量和连接位置，确保信息在网络中的流动和高效融合，使模型充分发挥跳跃连接的优势.

4.4　损失函数

在 RGB-D 图像语义分割任务中，常用交叉熵损失函数、Dice 系数损失函数和 Focal Loss 等来计算训练过程中的损失.

将每个像素的预测结果看作一个独立的分类问题，交叉熵损失函数通过计算预测结果与真实标签之间的差异来惩罚错误分类，定义为

$$L=-\sum_{c=1}^{M}\boldsymbol{y}_c\ln\left(p_c\right) \tag{1}$$

其中，M 表示类别数；$\boldsymbol{y}_c$ 表示 one-hot 向量，只有 0, 1 这 2 种取值，如果该类别和样本类别相同则取值为 1，否则取值为 0；p_c 表示预测样本属于 c 的概率.

在 RGB-D 图像语义分割任务中，一些数据中存在不同类别样本数量不均的问题，为此，可以使用 Dice 损失函数[58] L_{Dice}，通过计算预测结果与真

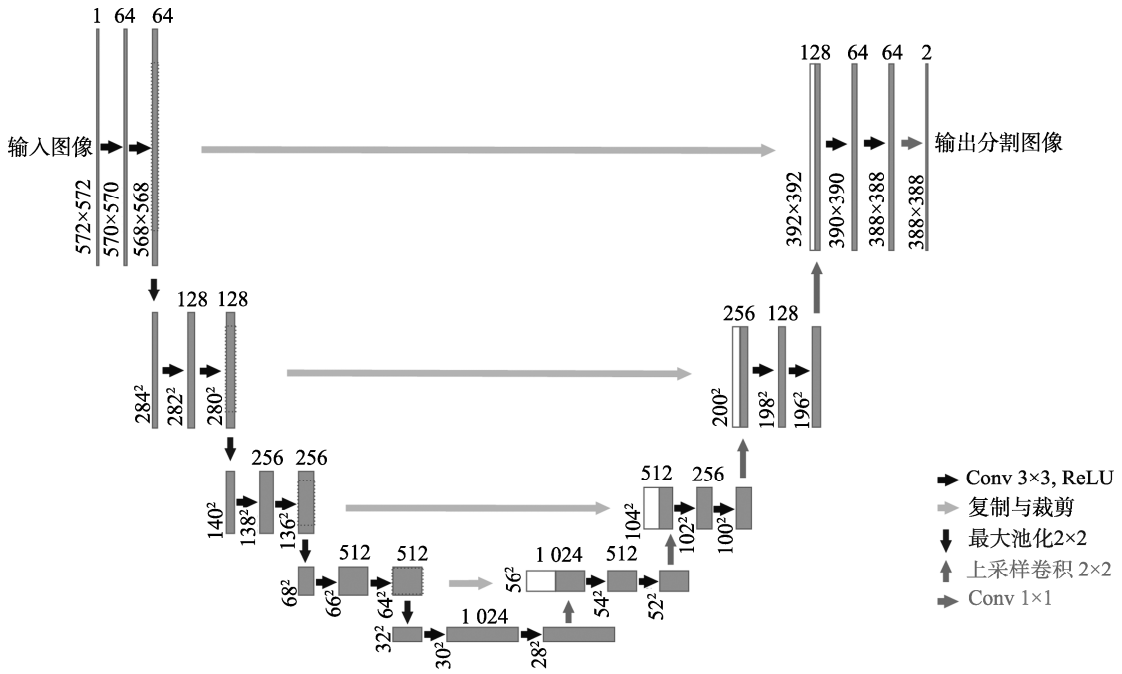


图 14 U-Net 结构[6]

实标签之间的相似度来评估分割质量，该损失函数倾向于提高对少数类别的识别. L_{Dice} 的取值范围为 0~1，当预测结果与真实标签完全一致时，$L_{\text{Dice}}=0$；而当两者完全不一致时，$L_{\text{Dice}}=1$. 因此，L_{Dice} 的值越小，表示模型的预测结果与真实标签的差异越小，说明模型性能越好. L_{Dice} 定义为

$$L_{\text{Dice}}=1-\frac{2|X\cap Y|}{|X|+|Y|} \tag{2}$$

其中，X 表示真实分割图像的像素标签；Y 表示模型预测分割图像的像素类别.

深度数据中的一些像素可能位于物体边界或者光照变化明显的区域，而这些区域往往是难以准确分类的. FL(focal loss)[59]通过减少易分类样本的权重，增加难分类样本的权重，使模型聚焦于难分类样本，提高对边界处的小物体的识别能力. 该损失函数 FL 定义为

$$\text{FL}(p_t)=-\alpha_t(1-p_t)^{\gamma}\ln(p_t) \tag{3}$$

其中，p_t 表示模型预测样本；t 表示属于正样本的概率；α_t 表示样本 t 的权重；γ 表示样本难度的调整因子.

除了单一的损失函数外，还可以结合多种损失函数，如 BCE-Dice Loss[60]为二元交叉熵(binary cross entropy, BCE)损失函数和 Dice 损失函数的组合，它综合考虑分类准确性与分割结果的相似度，提高了模型的泛化能力和对于少数类别识别的准确性.

4.5 模型优化

随着深度学习的发展，一系列技术不断涌现，极大地推动了模型性能的提升. 2023 年，Kirillov 等[61]提出通用分割模型(segment anything module, SAM)，它的推出为图像分割方向带来新的突破. 此外，模型预训练方法的改进以及基于提示的模型调优策略的应用，都在不同维度上对模型进行深入优化，推动了语义分割技术的发展.

4.5.1 通用分割模型

基础模型是人工智能领域的一项重要技术，它通过在大规模数据上进行训练获得强大的表征能力，能够灵活地适用于各种下游任务. 基础模型广泛应用于多个领域中，如自然语言处理(natural language processing, NLP)中的 GPT[62]和计算机视觉中的 CLIP[63]，均在多个任务中表现出色.

SAM 的结构如图 15 所示，由图像编码器、提示编码器和一个轻量级掩码解码器组成，并且开发了一个数据引擎，通过迭代地使用模型辅助数据收集和使用新收集的数据改进模型，同时生成一个大规模多样化的数据集 SA-1B. SAM 通过灵活的提示实现对不同的下游分割任务的实时响应，该模型在零样本迁移方面表现出色，甚至超过之前的全监督结果，其在多个任务中展现的强大的通用性和高效性，使其成为计算机视觉方向的重要研究工具.

但是，SAM 较高的模型复杂度和运算成本限制了它在特定场景中的广泛应用，因此一些研究

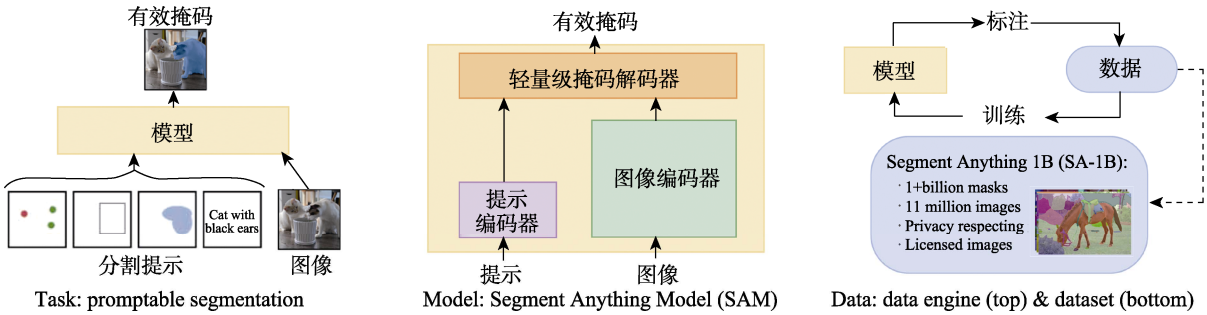


图 15　SAM 结构[61]

对其结构进行优化，以实现更高效的分割. FastSAM[64]将任务分解为所有实例的分割和基于提示的选择 2 个阶段，利用 CNN 高效的计算效率实现高效的运行速度，同时能够保持与 SAM 相当的性能，使其在实时应用中更具优势. EfficientSAM[65]通过学习重构 SAM 图像编码器的特征，实现更有效的视觉表示学习，并采用轻量级的 ViT 架构减少参数量，降低模型复杂度，进一步提升模型效率.

SAM 在 2D 图像处理方面表现出色，但面对 3D 图像分割却存在一定局限性. 针对该问题，SAM-Med3D[66] 精心改进和优化 SAM 的结构，采用一个全面的 3D 架构使其更好地适应 3D 医学图像分割任务. 针对复杂场景的挑战，Ji 等[67]通过一系列实验发现，SAM 在隐蔽场景识别方面存在一定局限性. 因此，针对某些隐蔽较高的复杂场景，研究者还需要借助特定领域的先验知识实现模型优化，增强模型对隐蔽目标的识别能力.

一些研究将 SAM 作为分割工具直接应用在多样化任务中[68]，通过结合 SAM 的优势解决相关问题. Inpaint Anything[69]引入 SAM 解决图像修复相关问题；还有一些研究将 SAM 扩展到视频领域，如 TAM[70]和 SAM-Track[71]应用 SAM 实现动态目标跟踪，SAMText[72]将 SAM 用于边界框注释，实现视频文本定位；SA3D[73]还将 SAM 应用在 3D 数据中，为 3D 重建带来创新方案；Cen 等[74]将 SAM 应用在 RGB-D 图像语义分割任务中，提出 SAD (segment any RGB-D)模型，使用 OVSeg[75]生成包含重要语义信息的粗略分割图像，同时将深度图像渲染到 RGB 空间并输入到 SAM 中提取几何信息，通过语义投票生成最终分割图像，一系列巧妙的设计显著地提升了 SAD 的分割性能. 上述研究成功地将 SAM 应用于各种任务中，拓宽了 SAM 的实际应用范围，为多个领域注入新的活力.

随着技术持续创新和突破，未来 SAM 将在计算机视觉方向发挥更加关键的作用.

4.5.2　模型预训练

预训练是深度学习中一种常见的技术，广泛应用于 NLP 和计算机视觉中，预训练的目的是在进行特定任务之前，通过在大规模数据集上进行训练学习通用特征，大幅提高模型的性能和实用性.

如图 16 所示，目前主流的 RGB-D 图像语义分割方法在预训练期间没有考虑深度信息[76-77]，而

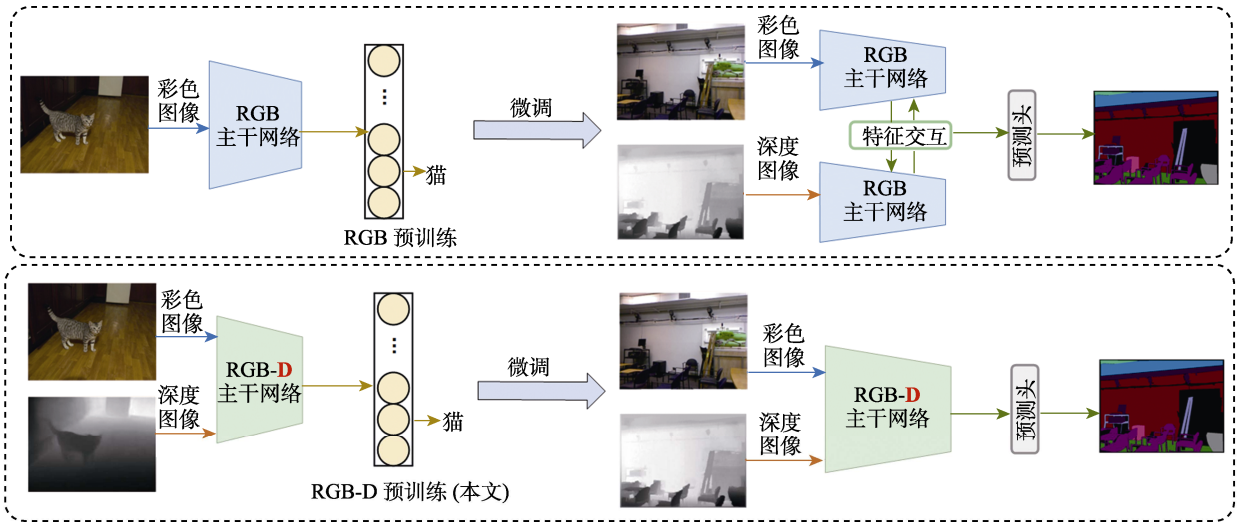


图 16　其他预训练方法与 DFormer 结构对比[78]

是采用 2 个 RGB 预训练主干分别对 RGB 和深度信息进行编码，并在微调阶段进行密集交互，以上设置会导致预训练和微调阶段输入不一致、RGB 与深度编码不匹配以及计算成本较高等问题.

基于上述问题, Yin 等[78]设计一个 RGB-D 预训练框架 DFormer，将 ImageNet-1K 的图像-深度对作为输入，在编码器的构建块中应用全局感知注意力(global awareness attention, GAA)模块和局部增强注意力(local enhancement attention, LEA)模块实现 RGB 与深度特征之间的交互，这种将多模态信息交互提前的方法使 DFormer 具备编码 RGB-D 表示的能力，同时避免了 RGB 预训练主干网络与对应的深度图像中几何关系不匹配的问题；该网络还通过少量通道编码深度信息有效地减小了模型的规模，以较低的计算成本实现显著的性能提升，在 NYUDv2 和 SUN-RGBD 这 2 个数据集上分别取得了平均交并比(mean intersection over union, mIoU)为 57.2%和 52.5%的优异成绩. 实验结果证明了 DFormer 在模型预训练阶段所做改进的有效性，为多模态模型的优化提供了新的途径.

4.5.3 模型提示调优

通过预训练在大规模数据集上学习到通用特征之后，模型会根据不同的下游任务对预训练模型进行微调，其中，大多数模型采用全量微调的方法，但该方法需要为每个下游任务存储和部署单独的主干参数副本，成本较高. 因此，提示调优的方法应运而生，该方法最初源于 NLP 领域，旨在提升预训练模型在特定任务中的通用性和性能.

基于提示调优在 NLP 中的进展[79-80], Jia 等[81]提出一种简单高效的视觉提示调优(visual prompt tuning, VPT)网络[81]，将提示调优应用于计算机视觉中. 如图 17 所示，该网络只在输入空间中引入少量可学习的参数，同时在训练下游任务期间冻结预训练模型的主干网络. E^2VPT[82]在自注意力机制和输入层中引入一组可学习的键值提示和视觉提示，提升了模型微调的有效性，并且设计了高效的提示剪枝策略来剪除一些不重要的提示；实验证明，这些方法有效地减少了计算负担，同时实现模型性能的显著提升.

Dong 等[83]巧妙地将提示调优应用在 RGB-D 图像语义分割任务中，提出一个双提示学习网络 DPLNet，设计了多模态提示生成器(multimodal prompt generation, MPG)和多模态特征适配器(multimodal feature adapter, MFA)这 2 种轻量级模块以适应多模态语义分割，并为网络设置了一个

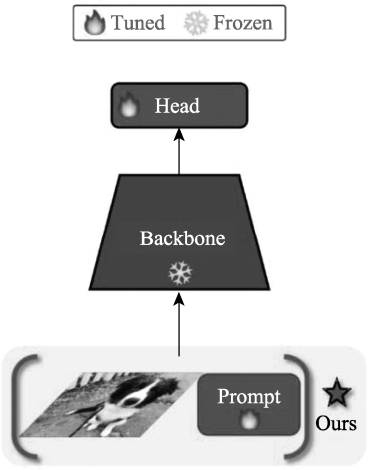


图 17 VPT 网络架构[81]

冻结的预训练 RGB 模型，这些结构有效减少了参数更新，且在多个数据集上表现优异；实验结果表明，在 NYUDV2 数据集上以 59.3%的 mIoU 分数位居语义分割任务排行榜第 2 名，以 52.8%的 mIoU 得分在 SUN-RGBD 数据集上排名第 2 名，彰显了 DPLNet 在 RGB-D 图像语义分割任务中强大的竞争力，同时也证明，提示调优应用在多模态语义分割任务中是卓有成效的.

随着深度学习技术不断发展，提示调优技术应用前景广阔，有望被广泛应用在 RGB-D 图像语义分割任务中，为解决复杂的视觉任务提供创新高效的方案.

5 数据集和性能分析

5.1 数据集

语义分割任务作为计算机视觉中重要的任务之一，其探索过程中不可或缺的是大量数据的支撑. 下面详细介绍 RGB-D 图像语义分割方法中一些常见的公共数据集.

(1) NYUDv2 数据集[9]. 该数据集由美国纽约大学的研究小组于 2012 年开发，包含 1 449 幅经过标注的室内 RGB 图像，这些图像覆盖了 51 个不同的房间，其中，407 幅图像有对应的深度图像，共包括 894 个类别. NYUDv2 数据集是一个具有挑战性的理想数据集，涵盖了多种不同的室内场景、光照条件和物体布局，同时还会存在一定的噪声和误差，用于测试和比较不同的计算机视觉算法的准确性和鲁棒性.

(2) SUN RGB-D 数据集[84]. 该数据集由美国普林斯顿大学的研究小组开发，数据由 4 个不同的

传感器捕获，包含 10 335 幅不同场景下的室内 RGB-D 图像，其中，包括 146 617 个 2D 多边形标注和 58 657 个具有精确对象定位的 3D 物体框.

(3) ScanNet 数据集[85]. 该数据集是一个 RGB-D 视频数据集，包含 1 500 多次不同的室内场景扫描，共包括超过 250 万个视图，并且使用 3D 相机姿势、表面重建和实例级语义标签对其进行注释，有助于室内场景理解和 3D 重建的研究.

(4) Stanford 2D-3D-S 数据集[86]. 该数据集是一个多模态的大规模室内空间数据集，采集的数据来自 6 个室内区域，包含 271 个房间；提供的数据包括 2D, 2.5D 和 3D(点云和网格)形式，由 70 496 幅完整的高清 RGB 图像及对应的深度图像、表面法线、网格和点云组成，共涵盖 13 种类别.

5.2　评价指标

在 RGB-D 图像语义分割任务中，准确性是衡量模型性能的常用的评价指标之一. 目前，语义分割图像的准确性度量评价指标主要包括 mIoU、交并比(intersection over union, IoU)、像素精度(pixel accuracy, PA)和平均 PA(mean PA, mPA)，可以帮助客观地评估语义分割模型的精确度及性能. 根据具体的需求和任务，选择适合的评价指标进行评估和比较.

5.2.1　IoU 与 mIoU

在语义分割任务中，IoU 是常用的评估指标之一，通过计算图像预测分割结果与真实标签之间的交集与并集，得到该像素点的 IoU. 将所有类别的 IoU 进行平均，得到 mIoU，用于评估整体的分割准确性，公式为

$$\mathrm{mIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ij}-p_{ii}} \tag{4}$$

其中，k 表示像素的类别数；p_{ii} 表示正确分类的像素个数；p_{ij} 表示实际类别为 i、预测类别为 j 的像素的数目.

5.2.2　PA 与 mPA

PA 是一个用于衡量模型整体准确性的评价指标，是通过计算预测结果中正确分类的像素数量与总像素数量之比得到，公式为

$$\mathrm{PA}=\sum_{i=0}^{k}p_{ii}\Big/\sum_{i=0}^{k}\sum_{j=0}^{n}p_{ij} \tag{5}$$

mPA 指预测结果中被正确分类的像素数量与总像素数量的比值，定义为

$$\mathrm{mPA}=\frac{1}{k+1}\sum_{i=0}^{k}\left(p_{ii}\Big/\sum_{j=0}^{n}p_{ij}\right) \tag{6}$$

该指标考虑了不同类别之间的像素分类情况，并将其平均化，可以反映模型在不同类别上的表现，提供更全面的性能评估.

5.3　性能比较

在 NYUDv2 和 SUN-RGBD 数据集上，对本文中涉及 3 种 RGB-D 图像语义分割网络及其对应的方法进行整理和对比，同时，对每种方法中使用的关键技术进行总结，结果如表 1 所示.

从表 1 可以看出:

(1) 单分支网络中，在 2 个数据集上, GLPNet[19] 展现出了明显的优势，其中，在 SUN-RGBD 数据集上，mIoU 达到 51.20%，在 NYUDv2 数据集上，mIoU 更是高达 54.60%. 这是因为在最深层网络中，GLPNet 对局部信息进行模态对齐，并且利用多模态全局上下文信息进一步促进深度特征的传播；另外，并行利用局部信息和全局信息，使得该方法能够在不同数据集上取得较高的性能，表现出很好的泛化能力和适应性.

(2) 双分支网络中，在 NYUDv2 数据集上，CMX[32]展现出了最佳性能，其 mIoU 达到 56.90%，而在 SUN-RGBD 数据集上, mIoU 为 52.40%. CMX 通过特征校正与特征融合的方式实现多模态特征增强，在 2 个数据集上都取得了较高的 mIoU，表明该方法在跨模态场景下的特征处理方面具有显著的优势，并且能够有效地提升语义分割的性能.

(3) 在 2 类三分支网络中，第 2 类三分支网络普遍比第 1 类三分支网络展现出更好的分割性能. 在 NYUDv2 数据集上，第 1 类网络的 mIoU 最高为 51.20%，而第 2 类网络的 mIoU 最高为 56.87%，表明在语义分割任务中，三分支网络的设计可以带来较好的性能表现，而第 2 类三分支网络的表现优于第 1 类，可能是因为第 2 类网络能够更好地利用多尺度融合信息，从而提高分割性能.

(4) 3 种分支网络中，双分支网络在 2 个数据集上的表现较为突出，尤其是 CMX 方法展现了最佳性能. 这是因为双分支网络通过将融合特征输入到 2 个分支中进行进一步下采样和特征提取，充分利用了多模态全局特征，有效地提升了分割的准确性. 结果表明，在 RGB-D 语义分割任务中，双分支网络或成为一种更有效的架构设计，有助于提升分割方法的性能.

(5) 关键技术中，注意力机制和跳跃连接技术在语义分割任务中得到了广泛应用，这 2 种技术能

表 1 RGB-D 图像语义分割网络性能对比

类别	方法	出版年	关键技术	骨干网络	mIoU/%	
					NYUDv2	SUN-RGBD
单分支 1	FuseNet[17]	2017		FuseNet-SF5		37.29
				FuseNet-DF1		34.02
	RedNet[18]	2018		ResNet-50		47.80
				ResNet-34		46.80
	RFNet[20]	2020				
	ESANet[12]	2021	跳跃连接、注意力机制、金字塔网络	R34-NBt1D	50.30	48.17
				ResNet-50	50.53	48.31
	GLPNet[19]	2021	跳跃连接、注意力机制	ResNet-101	54.60	51.20
	SGACNet[25]	2023	跳跃连接、注意力机制、金字塔网络	R34-NBt1D	49.40	47.80
				R18-NBt1D	48.20	46.50
单分支 2	AsymFormer[26]	2023	跳跃连接、注意力机制	B0+T	55.30	49.10
双分支	SA-Gate[28]	2020	跳跃连接、注意力机制	ResNet-50	52.40	
	NANet[34]	2021		ResNet-50	51.40	48.00
				ResNet-101	52.30	48.80
	DCANet[13]	2022	跳跃连接、注意力机制	ResNet-101	53.30	49.60
	BCINet[35]	2023	跳跃连接、金字塔网络	ResNet-50	52.95	49.57
	CMFormer[33]	2023	跳跃连接、注意力机制	CMFormer-M	54.12	
				CMFormer-L	55.75	
	CMX[32]	2023	跳跃连接、注意力机制	MiT-B2	54.40	49.70
				MiT-B5	56.90	52.40
	HaarNet[36]	2024	金字塔网络	ResNet-50	50.70	
三分支 1	RDFNet[43]	2017		RDF-101	49.10	
				RDF-152	50.10	47.70
	FRNet[45]	2022		ResNet-34	53.60	51.80
	CMPFFNet[44]	2023	注意力机制	Segformer-B2	53.79	49.33
				Segformer-B5	56.87	51.46
三分支 2	ACNet[39]	2019	注意力机制	ResNet-50	48.30	48.10
	TSNet[42]	2021	跳跃连接、注意力机制、金字塔网络	ResNet-34+ VGGNet16	46.10	
	CANet[41]	2021	注意力机制	ResNet-50	49.60	48.10
				ResNet-101	51.20	49.30
	TCANet[40]	2024	跳跃连接、注意力机制	ResNet-50	50.20	48.90

够增强网络对重要区域和上下文信息的关注，显著地提升了模型分割效果.

对本文涉及的能够实现实时分割的网络的性能进行总结对比，结果如表 2 所示. 其中，单分支网络中实时语义分割网络较多，这可能得益于单分支网络的结构特性，仅以单一模态作为主要分支，这种设计减少了模型参数量，因此更容易满足实时性的要求. 这些对比能够为实时网络应用研究提供有价值的参考信息.

本文综合评述3类分支的代表性网络，深入分析了各自的优势和局限性，并对其适用范围进行总结，结果如表 3 所示. 旨在为研究者提供更多参考信息，以期在未来能够设计出与特定任务更匹配的网络.

表 2 RGB-D 图像实时语义分割网络性能对比

类别	方法	分辨率/像素	GPU	参数量	分割速度/(帧·s^{-1})
单分支	RFNet[20]	2 048×1 024	NVIDIA GTX 2080Ti		22.00
	ESANet[12]	2 048×1 024	NVIDIA Jetson AGX Xavier	29.70M	29.70
	SGACNet[25]	1 024×512	NVIDIA RTX 3090	22.10M	20.70
	AsymFormer[26]	480×640	NVIDIA RTX 3090Ti	33.00M	65.00
双分支	BCINet[35]	480×640	NVIDIA TITAN V	78.38M	27.30
	CMFormer-S[33]	480×640	Intel Core i9 12900K CPU	47.93M	43.00
	CMFormer-M[33]	480×640	64 GB 3 600 MHz RAM	81.75M	27.00
	CMFormer-L[33]	480×640	NVIDIA RTX 3090	131.41M	18.00

表 3　RGB-D 图像语义分割网络总结

类别	方法	优点	缺点	适用范围
单分支 1	FuseNet[17]	提出密集融合和稀疏融合 2 种融合策略，增强特征表达能力；可以端到端学习，无需额外的处理	融合模块通过元素求和实现，可能不是最佳策略	需要精细场景理解的室内环境
	RedNet[18]	应用多层监督学习改善梯度消失，使用残差模块提高网络训练效率	残差网络及金字塔监督增加了计算量	需要精细场景理解的室内环境
	RFNet[20]	高效的网络架构，满足实时操作的需求；结合多数据集训练策略，增强对意外小障碍物的检测能力	计算资源需求较大，在较低端硬件上部署可能受限；对于更复杂场景，模型性能受影响	自动驾驶
	ESANet[12]	高度优化的网络架构，实现快速推理及性能提升	计算资源需求较大，在较低端硬件上部署可能受限	移动机器人
	GLPNet[19]	有效地解决了特征融合过程中空间不一致性和语义信息的传播问题	模型表现受超参数影响，需要细致调整	需要精细场景理解的室内环境
	SGACNet[25]	改善深度信息的噪声问题，且减少计算参数	对深度数据的质量有较高要求，需要进一步优化	自动驾驶 移动机器人
单分支 2	AsymFormer[26]	非对称高效网络减少冗余参数，深化对跨模态特征的理解和利用	依赖特定数据集，在复杂的室内环境中的性能可能较差	移动机器人
双分支	SA-Gate[28]	有效地减小噪声深度测量的影响，灵活性较高，即插即用，易于集成	复杂环境中噪声处理需要优化	室外场景 室内场景
	NANet[34]	捕获 RGB-D 特征的非局部上下文息，整合多尺度信息提高特征表达能力	非局部聚合可能需要较高的计算成本，对于小物体分割识别能力可能较差	需要精细场景理解的室内环境
	DCANet[13]	自适应捕获并融合局部和全局上下文信息，模块即插即用	对数据要求较高，可能不适用于其他数据	室内场景 室外场景
	BCINet[35]	捕获跨模态的补充信息，并通过特征增强提高特征表达能力	主要针对室内场景分割，在其他类型的场景中可能表现不佳	需要精细场景理解的室内环境
	CMFormer[33]	实现跨模态校正和融合，模型精度较高且实时处理能力较强	对小数据量较敏感，处理特定类别时表现不佳	实时场景理解
	CMX[32]	多种模态通用语义分割框架，能够实现全面跨模态交互	需要大量的计算资源，对于特定模态数据需要调整优化	自动驾驶
	HaarNet[36]	结合线性和数学形态学操作，模型参数量较少	对深度信号质量要求较高，在一些具有噪声的深度数据可能表现不佳	需要精细场景理解的室内环境
三分支 1	RDFNet[43]	多级特征细化块有利于高分辨率预测	大规模数据集可能需要较长的训练时间	需要高分辨率预测的室内场景
	FRNet[45]	高效网络能够生成具有丰富模态感知信息的特征节点，提高特征表达的质量	模型较复杂，计算成本较高	需要精细场景理解的室内环境
	CMPFFNet[44]	多尺度特征融合增强对全局信息的整合	由于对多模态融合特征进行额外处理，导致参数量增加	需要精细场景理解的室内环境
三分支 2	ACNet[39]	能够并行处理单一模态特征和多模态融合特征	3 个并行分支使参数量和计算成本增加	需要精细场景理解的室内环境
	TSNet[42]	不对称编码器结构提取有效信息，自注意力机制优化特征融合	不对称网络参数量相对较多，计算成本较高	需要精细场景理解的室内环境
	CANet[41]	通道和空间双注意力提高特征的表达能力	双注意力机制需要更多的计算资源	具有相似颜色和纹理的物体
	TCANet[40]	CA 机制有效地聚合通道和空间不同维度的特征信息	注意力机制和多模态融合可能会增加模型计算负担	需要精细场景理解的室内环境

6 存在问题与未来展望

基于 RGB-D 的语义分割方法在自动驾驶、增强现实、虚拟现实、医学影像分析等多个领域实现对环境的理解和感知. 当前, 该方法面临如下挑战.

(1) 深度数据质量问题. 由于深度传感器本身的特性或环境因素, 获取的深度图像可能存在噪声和信息缺失情况, 这些噪声和缺失会影响整体的分割效果, 特别是对于物体边界和细节区域的识别.

(2) 多模态数据的不一致. 在 RGB-D 数据中, 感知差异或数据噪声等因素可能导致 RGB 图像与深度图像之间存在特征不匹配的情况, 这种不一致会影响网络对深度信息的正确理解和利用, 降低网络提取特征的准确性. 因此, 需要对图像进行校准和特征对齐.

(3) 多模态数据融合和网络架构设计策略. 在 RGB-D 语义分割任务中, 有效地融合 RGB 和深度特征是一个关键问题. 由于 RGB 和深度数据具有不同的特性, 需要设计泛化能力强的网络结构来充分地利用多模态信息. 在单分支网络中, 由于未能充分利用 RGB 对深度特征的特征增强, 可能会导致深度图像的特征信息不完整. 在一些双分支网络的特征融合过程中, 不同模态的特征相互影响, 导致一些原始特征的丢失或混淆. 而在一些三分支网络中, 虽然将融合后的特征作为新的分支, 但其余 2 个分支中 RGB 和深度特征在捕捉目标特征方面具有不同的优势, 而仅使用原始图像可能无法充分利用这种互补性, 导致所提取的特征信息的不完整; 在后续的特征提取阶段, 某些细节信息可能无法得到有效的利用, 从而影响了网络对目标的准确识别能力. 因此, 在选择网络时应综合考虑各种网络模型的优缺点, 并根据具体任务的需求和数据特点选择最合适的网络.

(4) 算法复杂度和计算资源. 相对于仅使用 RGB 图像进行分割的方法, RGB-D 语义分割方法通常更为复杂, 这是因为需要同时处理 RGB 和深度数据, 并进行特征融合与联合优化, 这种复杂性导致算法的计算量增加, 需要更多的计算资源. 在实时应用和嵌入式系统中, 需要寻找更高效的算法和硬件加速方法来应对这一挑战.

未来还可以探索多层次跨领域的数据融合模式, 如语音、激光雷达、生物信息等, 不同类型的数据能够提供更准确、多样化的场景理解和分割结果. RGB-D 图像语义分割为智能化的视觉理解提供了更广阔的可能性, 未来也将不断拓展应用场景、采用更加高效智能的算法和网络架构, 以实现更加精确的语义分割效果, 执行更精细化的语义分割任务.

7 总 结

本文对近年来基于 RGB-D 图像的语义分割方法进行梳理和归纳, 根据对多模态融合特征处理方式的差异, 将这些方法归类为单分支网络、双分支网络和三分支网络, 分别阐述了这些网络结构的优势、特点和局限性. 归纳了图像语义分割常用的公共数据集和性能评价指标, 在此基础上, 分别比较 3 类方法在不同数据集上的性能, 同时对文中提到的 RGB-D 语义分割网络的优缺点及适用范围进行归纳总结, 最后指出了目前该领域存在的问题, 提出未来的研究发展方向. 通过研究和讨论这些问题, 以期为进一步推进 RGB-D 图像语义分割方法的发展提供思路和指引, 为研究者在该领域探索新的解决方案提供有力支持.

参考文献(References):

[1] He Jian, Liu Xinyuan. Ground obstacle detection technology based on fusion of RGB-D and inertial sensors[J]. Journal of Computer-Aided Design & Computer Graphics, 2022, 34(2): 254-263(in Chinese)
(何坚, 刘新远. RGB-D 和惯性传感器融合的地面障碍物检测技术[J]. 计算机辅助设计与图形学学报, 2022, 34(2): 254-263)

[2] Dong Zihao, Shao Xiuli. A multi-category edge perception method for semantic segmentation[J]. Journal of Computer- Aided Design & Computer Graphics, 2019, 31(7): 1075-1085 (in Chinese)
(董子昊, 邵秀丽. 多类别的边缘感知方法在图像分割中的应用[J]. 计算机辅助设计与图形学学报, 2019, 31(7): 1075-1085)

[3] Ma Xiao, Liu Xiabi, Gao Yixuan, *et al*. A review on image co-segmentation[J]. Journal of Computer-Aided Design & Computer Graphics, 2017, 29(10): 1767-1775(in Chinese)
(马骁, 刘峡壁, 高一轩, 等. 图像协同分割方法综述[J]. 计算机辅助设计与图形学学报, 2017, 29(10): 1767-1775)

[4] 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

[5] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440

[6] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of

the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241

[7] 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

[8] Cippitelli E, Fioranelli F, Gambi E, *et al*. Radar and RGB-depth sensors for fall detection: a review[J]. IEEE Sensors Journal, 2017, 17(12): 3585-3604

[9] Silberman N, Hoiem D, Kohli P, *et al*. Indoor segmentation and support inference from RGBD images[C] //Proceedings of the 12th European Conference on Computer Vision. Heidelberg: Springer, 2012: 746-760

[10] Ren X F, Bo L F, Fox D. RGB-(D) scene labeling: features and algorithms[C] //Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 2759-2766.

[11] Gupta S, Girshick R, Arbeláez P, *et al*. Learning rich features from RGB-D images for object detection and segmentation[C] //Proceedings of the 13th European Conference on Computer Vision. Heidelberg: Springer, 2014: 345-360

[12] Seichter D, Köhler M, Lewandowski B, *et al*. Efficient RGB-D semantic segmentation for indoor scene analysis[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2021: 13525-13531

[13] Bai L Z, Yang J, Tian C Q, *et al*. DCANet: differential convolution attention network for RGB-D semantic segmentation[OL]. [2024-03-01]. https://arxiv.org/abs/2210.06747

[14] Zhang Xin, Yao Qing'an, Zhao Jian, *et al*. Image semantic segmentation based on fully convolutional neural network[J]. Computer Engineering and Applications, 2022, 58(8): 45-57(in Chinese)
(张鑫, 姚庆安, 赵健, 等. 全卷积神经网络图像语义分割方法综述[J]. 计算机工程与应用, 2022, 58(8): 45-57)

[15] Gao Changxin, Xu Zhengze, Wu Dongyue, *et al*. Deep learning-based real-time semantic segmentation: a survey[J]. Journal of Image and Graphics, 2024, 29(5): 1119-1145(in Chinese)
(高常鑫, 徐正泽, 吴东岳, 等. 深度学习实时语义分割综述[J]. 中国图象图形学报, 2024, 29(5): 1119-1145)

[16] Barbosa F M, Osório F S. A threefold review on deep semantic segmentation: efficiency-oriented, temporal and depth-aware design[OL]. [2024-03-01]. https://arxiv.org/abs/2303.04315

[17] Hazirbas C, Ma L N, Domokos C, *et al*. FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture[C] //Proceedings of the 13th Asian Conference on Computer Vision. Heidelberg: Springer, 2017: 213-228

[18] Jiang J D, Zheng L N, Luo F, *et al*. RedNet: residual encoder-decoder network for indoor RGB-D semantic segmentation[OL]. [2024-03-01]. https://arxiv.org/abs/1806.01054

[19] Chen S H, Zhu X X, Liu W, *et al*. Global-local propagation network for RGB-D semantic segmentation[OL]. [2024-03-01]. https://arxiv.org/abs/2101.10801

[20] Sun L, Yang K L, Hu X 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

[21] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141

[22] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916

[23] Romera E, Álvarez J M, Bergasa L M, *et al*. ERFNet: efficient residual factorized ConvNet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(1): 263-272

[24] Zhao H S, Shi J P, Qi X J, *et al*. Pyramid scene parsing network[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6230-6239

[25] Zhang Y, Xiong C Y, Liu J J, *et al*. Spatial information-guided adaptive context-aware network for efficient RGB-D semantic segmentation[J]. IEEE Sensors Journal, 2023, 23(19): 23512-23521

[26] Du S Q, Wang W X, Guo R Z, *et al*. AsymFormer: asymmetrical cross-modal representation learning for mobile platform real-time RGB-D semantic segmentation[OL]. [2024-03-01]. https://arxiv.org/abs/2309.14065

[27] Liu Z, Mao H Z, Wu C Y, *et al*. A ConvNet for the 2020s[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 11966-11976

[28] Chen X K, Lin K Y, Wang J B, *et al*. Bi-directional cross-modality feature propagation with separation-and- aggregation gate for RGB-D semantic segmentation[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 561-577

[29] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010

[30] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: transformers for image recognition at scale[OL]. [2024-03-01]. https://arxiv.org/abs/2010.11929

[31] Zheng S X, Lu J C, Zhao H S, *et al*. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6877-6886

[32] Zhang J M, Liu H Y, Yang K L, *et al*. CMX: cross-modal fusion for RGB-X semantic segmentation with transformers[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24 (12): 14679-14694

[33] Ru Q J, Chen G Z, Zuo T Y, *et al*. Cross-modal Transformer for RGB-D semantic segmentation of production workshop objects[J]. Pattern Recognition, 2023, 144: Article No.109862

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

[35] Zhou W J, Yue Y C, Fang M X, *et al*. BCINet: bilateral

cross-modal interaction network for indoor scene understanding in RGB-D images[J]. Information Fusion, 2023, 94: 32-42

[36] Groenendijk R, Dorst L, Gevers T. HaarNet: large-scale linear-morphological hybrid network for RGB-D semantic segmentation[C] //Proceedings of the 3rd International Conference on Discrete Geometry and Mathematical Morphology. Heidelberg: Springer, 2024: 242-254

[37] Fukushima K. Cognitron: a self-organizing multilayered neural network[J]. Biological Cybernetics, 1975, 20(3-4): 121-136

[38] Qiu S, Xu X M, Cai B L. FReLU: flexible rectified linear units for improving convolutional neural networks[C] //Proceedings of the 24th International Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 1223-1228

[39] Hu X X, Yang K L, Fei L, *et al*. ACNET: attention based network to exploit complementary features for RGBD semantic segmentation[C] //Proceedings of the IEEE International Conference on Image Processing. Los Alamitos: IEEE Computer Society Press, 2019: 1440-1444

[40] Jia W K, Yan X C, Liu Q L, *et al*. TCANet: three-stream coordinate attention network for RGB-D indoor semantic segmentation[J]. Complex & Intelligent Systems, 2024, 10(1): 1219-1230

[41] Zhou H, Qi L, Wan Z L, *et al*. RGB-D co-attention network for semantic segmentation[C] //Proceedings of the 15th Asian Conference on Computer Vision. Heidelberg: Springer, 2021: 519-536

[42] Zhou W J, Yuan J Z, Lei J S, *et al*. TSNet: three-stream self-attention network for RGB-D indoor semantic segmentation[J]. IEEE Intelligent Systems, 2021, 36(4): 73-78

[43] Lee S, Park S J, Hong K S. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation[C] // Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 4990-4999

[44] Zhou W J, Xiao Y X, Yan W Q, *et al*. CMPFFNet: cross-modal and progressive feature fusion network for RGB-D indoor scene semantic segmentation[OL]. [2024-03-01]. https://ieeexplore.ieee.org/document/10252155

[45] Zhou W J, Yang E Q, Lei J S, *et al*. FRNet: feature reconstruction network for RGB-D indoor scene parsing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(4): 677-687

[46] Jaderberg M, Simonyan K, Zisserman A, *et al*. Spatial transformer networks[C] //Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 2017-2025

[47] Woo S, Park J, Lee J Y, *et al*. CBAM: convolutional block attention module[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19

[48] Xie E Z, Wang W H, Yu Z D, *et al*. SegFormer: simple and efficient design for semantic segmentation with transformers[C] // Proceedings of the 35th Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc. 2021: 12077-12090

[49] Liu Z, Lin Y T, Cao Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows[C] //Proceedings of the IEEE/CVF international Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 9992-10002

[50] Lin H Z, Cheng X, Wu X Y, *et al*. CAT: cross attention in vision transformer[C] //Proceedings of the IEEE International Conference on Multimedia and Expo. Los Alamitos: IEEE Computer Society Press, 2022: 1-6

[51] Shen J F, Chen Y F, Liu Y, *et al*. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection[J]. Pattern Recognition, 2024, 145: Article No.109913

[52] Yu F, Koltun V, Funkhouser T. Dilated residual networks[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 636-644

[53] 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, 2018, 40(4): 834-848

[54] He J J, Deng Z Y, Zhou L, *et al*. Adaptive pyramid context network for semantic segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 7511-7520

[55] Orhan A E, Pitkow X. Skip connections eliminate singularities[C] //Proceedings of the 6th International Conference on Learning Representations. Washington DC: ICLR, 2018

[56] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[57] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 936-944

[58] Zhao R J, Qian B Y, Zhang X L, *et al*. Rethinking dice loss for medical image segmentation[C] //Proceedings of the IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2020: 851-860

[59] Mukhoti J, Kulharia V, Sanyal A, *et al*. Calibrating deep neural networks using focal loss[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.1282

[60] Xu H Z, He H J, Zhang Y, *et al*. A comparative study of loss functions for road segmentation in remotely sensed road datasets[J]. International Journal of Applied Earth Observation and Geoinformation, 2023, 116: Article No.103159

[61] Kirillov A, Mintun E, Ravi N, *et al*. Segment anything[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 3992-4003

[62] Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training[OL]. [2024-03-01]. https://api.semanticscholar.org/CorpusID:49313245

[63] Radford A, Kim J W, Hallacy C, *et al*. Learning transferable visual models from natural language supervision[C] //Proceedings of the 38th International Conference on Machine Learning. Maastricht: ML Research Press, 2021: 8748-8763

[64] Zhao X, Ding W C, An Y Q, *et al*. Fast segment anything[OL]. [2024-03-01]. https://arxiv.org/abs/2306.12156

[65] Xiong Y Y, Varadarajan B, Wu L M, *et al*. EfficientSAM: leveraged masked image pretraining for efficient segment anything[OL]. [2024-03-01]. https://arxiv.org/abs/2312.00863

[66] Wang H Y, Guo S Z, Ye J, *et al*. SAM-Med3D[OL]. [2024-03-01]. https://arxiv.org/abs/2310.15161

[67] Ji G P, Fan D P, Xu P, *et al*. SAM struggles in concealed scenes-empirical study on “segment anything”[J]. Science China Information Sciences, 2023, 66(12): Article No.226101

[68] Zhang C H, Liu L, Cui Y W, *et al*. A comprehensive survey on segment anything model for vision and beyond[OL]. [2024-03-01]. https://arxiv.org/abs/2305.08196

[69] Yu T, Feng R S, Feng R Y, *et al*. Inpaint anything: segment anything meets image inpainting[OL]. [2024-03-01]. https:// arxiv.org/abs/2304.06790

[70] Yang J Y, Gao M Q, Li Z, *et al*. Track anything: segment anything meets videos[OL]. [2024-03-01]. https://arxiv.org/abs/2304.11968

[71] Cheng Y M, Li L L, Xu Y Y, *et al*. Segment and track anything[OL]. [2024-03-01]. https://arxiv.org/abs/2305.06558

[72] He H B, Zhang J, Xu M Y, *et al*. Scalable mask annotation for video text spotting[OL]. [2024-03-01]. https://arxiv.org/abs/2305.01443

[73] Cen J Z, Zhou Z W, Fang J M, *et al*. Segment anything in 3D with NeRFs[C] //Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2023: Article No.1130

[74] Cen J, Wu Y Z, Wang K W, *et al*. SAD: segment any RGBD[OL]. [2024-03-01]. https://arxiv.org/abs/2305.14207

[75] Liang F, Wu B C, Dai X L, *et al*. Open-vocabulary semantic segmentation with mask-adapted CLIP[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 7061-7070

[76] Wang Y K, Chen X H, Cao L L, *et al*. Multimodal token fusion for vision transformers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 12176-12185

[77] Zhang J M, Liu R P, Shi H, *et al*. Delivering arbitrary-modal semantic segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 1136-1147

[78] Yin B W, Zhang X Y, Li Z Y, *et al*. DFormer: rethinking RGBD representation learning for semantic segmentation[OL]. [2024-03-01]. https://arxiv.org/abs/2309.09668

[79] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 3045-3059

[80] Li X L, Liang P. Prefix-tuning: optimizing continuous prompts for generation[C] //Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 4582-4597

[81] Jia M L, Tang L M, Chen B C, *et al*. Visual prompt tuning[C] // Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 709-727

[82] Han C, Wang Q F, Cui Y M, *et al*. E^2VPT: an effective and efficient approach for visual prompt tuning[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 17445-17456

[83] Dong S H, Feng Y H, Yang Q, *et al*. Efficient multimodal semantic segmentation via dual-prompt learning[OL]. [2024-03-01]. https://arxiv.org/abs/2312.00360

[84] Song S R, Lichtenberg S P, Xiao J X. SUN RGB-D: a RGB-D scene understanding benchmark suite[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 567-576

[85] Dai A, Chang A X, Savva M, *et al*. ScanNet: richly-annotated 3D reconstructions of indoor scenes[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2432-2443

[86] Armeni I, Sax S, Zamir A R, *et al*. Joint 2D-3D-semantic data for indoor scene understanding[OL]. [2024-03-01]. https://arxiv.org/abs/1702.01105