

基于特征融合注意力的 RGB-D 语义分割算法

江 晟¹, 路 启¹, 夏森磊², 王稼禹¹

(1. 长春理工大学 物理学院, 长春 130022; 2. 温州理工学院 建筑与能源工程学院, 浙江 温州 325035)

摘 要: 针对室内场景感知任务中, 现有的多模态融合方法主要侧重于单一模态 (如 RGB 或深度图像) 的注意力, 而对模态融合后联合特征空间中的注意力缺乏有效利用, 且许多网络的参数量较大, 在参数量减少时模型精度通常会显著下降等问题, 提出了一种多模态语义分割模型, 该模型通过 ConvNeXt 和 MiT 分别提取 RGB 特征和深度特征, 并引入跨模态特征校正模块 (CMFCM), 通过空间和通道特征增强促进模态间的交互与校正。此外, 模型还设计了跨模态注意力融合模块 (CMAFM), 引入了联合特征空间中的注意力, 实现多尺度特征的有效融合, 充分利用两种模态的分割信息, 同时通过深度可分离卷积显著降低计算成本。最终, 在 NYU-Depth V2 数据集上的实验表明, 与 CMX、Dformer 和 SA-Gate 模型相比, 该方法在分割精度提升 1% 的同时参数量减少了 15.1%, 显著降低了计算复杂度。

关键词: 语义分割; 多模态融合; 室内感知; 注意力机制

中图分类号: TP301.6 **文献标志码:** A **文章编号:** 1671-5497(2026)09-2435-09

DOI: 10.13229/j.cnki.jdxbgxb.20250159

RGB-D semantic segmentation algorithm based on feature fusion attention

JIANG Sheng¹, LU Qi¹, XIA Miao-lei², WANG Jia-yu¹

(1. College of Physics, Changchun University of Science and Technology, Changchun 130022, China; 2. College of Architecture and Energy Engineering, Wenzhou University of Technology, Wenzhou 325035, China)

Abstract: A multimodal semantic segmentation model was proposed for indoor scene perception tasks. Existing multimodal fusion methods mainly focus on a single modality, such as RGB or depth image attention, while lacking effective utilization of attention in the joint feature space after modality fusion. Additionally, many networks suffer from a significant drop in accuracy when the number of parameters is reduced, due to their large parameter sizes. The RGB and depth features are extracted using ConvNeXt and MiT, respectively. A Cross-modal Feature Correction Module (CMFCM) is introduced to enhance spatial and channel features, facilitating interaction and correction between modalities. Additionally, a Cross-modal Attention Fusion Module (CMAFM) is designed to apply attention in the joint feature space,

收稿日期: 2025-02-28.

基金项目: 国家自然科学基金青年基金项目 (62405032).

作者简介: 江晟 (1985-), 男, 副教授, 博士. 研究方向: 多维智能感知与协同控制. E-mail: js1985_cust@163.com

通信作者: 夏森磊 (1986-), 男, 讲师, 硕士. 研究方向: 智能交通系统. E-mail: 240654931@qq.com@163.com

enabling effective multi-scale feature fusion and maximizing the utilization of segmentation information from both modalities. Depthwise separable convolution is also employed to significantly reduce computational costs. Finally, experiments on the NYU-Depth V2 dataset show that, compared with CMX, Dformer, and SA-Gate models, the proposed method improves segmentation accuracy by 1% while reducing the number of parameters by 15.1%, significantly lowering computational complexity.

Key words: semantic segmentation; multimodal fusion; indoor perception; attention mechanism

0 引言

语义分割作为理解图像场景的核心技术,已成为图像处理和计算机视觉领域的研究热点。由于室内场景存在语义类别多、互相遮挡和光照不均匀等问题,室内语义分割仍是最具挑战性的分割任务之一^[1]。传统的目标分割技术通常基于RGB图像,但仅通过RGB图像的分割结果容易受到光照条件的影响,具有一定的局限性。深度相机通过结构光或飞行时间(Time of flight, ToF)技术发射红外光或激光以获取深度信息,进而获取目标距离和三维结构等几何信息。即使在低光照条件下,深度信息仍然可以获取距离和形状信息,对RGB数据进行补充,改善语义分割效果。近年来,RGB与深度图像结合已成为语义分割研究的一个重要方向。通过利用RGB数据的纹理、颜色特征以及深度数据的几何和空间信息,可以显著提升分割的准确性。

一些早期的研究,如RedNet^[2],在主干中采取逐元素相加的方式融合了RGB和Depth,但这种方法并未充分利用两种特征之间的互补关系,影响了最终的特征表示。ESANet网络结构中,通过加入通道注意力,在不同通道中进行特征选择,并加权使模型更关注重要的特征,同时抑制无关或冗余的信息,提升了模型的特征表达能力和分割效果^[3]。PSCNet网络进一步将通道关注扩展到空间和通道方向,空间关注的主要优点在于精细化特征选择和区域性特征增强,结合通道注意力后能够显著提升模型对细小区域的分割能力^[4]。近年来,不少学者提出了更复杂的模型,利用跨模态信息RGBD的融合特征。例如,Chen等^[5]提出了一种门控注意力机制,可以利用两种模态的通道相关性和空间相关性有效抑制深度图像噪声。Zhang等^[6]提出了一种新颖的具有全局感受野的交叉注意力机制,融合不同尺度下的模态特征。通过在不同尺度上提取RGB和深度特

征,捕获不同尺度下的物体信息,使模型能够兼顾大范围背景和细小物体细节。DFormer网络^[7]通过大量且多样化的RGBD图像进行预训练,使模型学习到更丰富、更具泛化性的特征表示,这种策略显著提升了模型对复杂场景的理解能力,但是高质量的RGBD图像获取难度高,限制了其在实际应用中的普及性。Bui等^[8]提出了一种基于扩散模型和可变形注意力的RGB-D语义分割模块,能够有效处理深度图中常见的无效区域,实现高精度的像素级分割。虽然扩散模型训练效率高,但在推理阶段通常很慢。Dong等^[9]针对现有多模态语义分割方法计算开销大、训练效率低的问题,提出DPLNet语义分割模型,利用提示学习思想对冻结的预训练RGB模型进行适配,引入多模态提示生成器和特征适配器两个轻量模块,生成多层级的多模态提示并注入到冻结的主干网络中实现高效的特征融合和引导。该方法在提升分割精度的同时显著减少了参数更新,兼顾了分割性能和计算资源的双重需求。缪君等^[10]构建了一种3个并行支流的双向融合体系结构,确保在每个编码和解码层对并行的RGB图像特征和深度图像特征进行融合。同时,为避免重要特征丢失且充分融合两种特征,构造了两个互补的注意力机制,使两种特征能得到局部和全局的互补。王雪等^[11]提出一种基于多尺度自动感知和多层语义适配的深度卷积编解码网络模型,增强了不同尺度特征学习的表征能力,避免了非目标区域的干扰,但目标较多时分割性能下降。周大可等^[12]针对各网络预测分量不同,所关注的图像特征信息也不同的特点,引入双重注意力机制模块,从而有效提高了网络提取特征的速度和能力。以上多模态融合方法主要侧重于模态的注意力融合,而缺乏对模态融合后联合特征空间中注意力的有效利用。因此,如何在融合过程中充分整合跨模态信息,仍是多模态语义分割面临的关键挑战。此外,大部分现有模型的参数量较大,对计算

资源需求高,难以在保持较高准确率的同时有效控制参数量,在资源受限的场景下无法使用。

针对上述问题,本文提出了一种双模态融合的语义分割模型。使用ConvNeXt和MiT分别提取深度图像局部特征和彩色图像全局特征,设计了跨模态特征校正模块(Cross-modal feature correction module, CMFCM),用于纠正模态间不一致性并提升特征对齐效果,校正二者的图像特征。针对传统注意力机制的计算瓶颈和融合不充分问题,设计了跨模态注意力融合模块(Cross-modal attention fusion module, CMAFM),融合多尺度特征,并加入融合后的注意力,充分利用两种模态分割图像,并使用深度可分离卷积降低了运算参数量。最后,在NYU-Depth V2数据集上进行验证,

结果表明,与其他现有的语义分割网络相比,本文模型所需的参数量减少,但检测精度得到提升。

1 模型构建

本文采取基于编码器-解码器类型的网络架构,提出一种RGB-D信息互补的语义分割网络,搭建了一个双流网络,结合彩色和深度信息进行图像语义分割,如图1所示。针对两种图像的特点,分别使用不同的特征提取策略提取图像特征,CMFCM增强模态间的交互,从空间和通道两个维度校正图像特征;CMAFM计算跨模态相似度,融合跨模态特征,充分利用两种模态分割图像,并利用深度可分离卷积降低运算参数量。

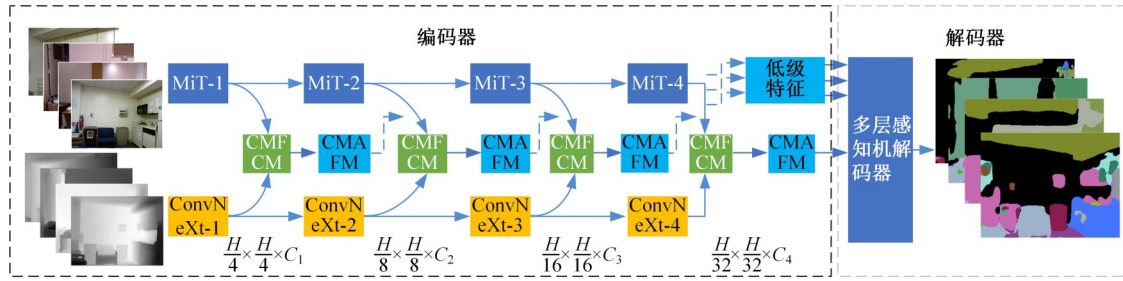


图1 模型框架图

Fig. 1 Diagram of model framework

1.1 特征提取模块

1.1.1 RGB特征提取

本文RGB图像特征提取骨干采用了MiT内部的Transformer模块^[13],该模块运用了重叠块嵌入(Overlap patch embeddings, OPE)技术对输入的图片进行特征提取和下采样处理。随后,这些特征会被进一步送入高效多头自注意力(Efficient multihead self-attention, EMSA)层和混合前馈(Mix feed forward network, MixFFN)层进行深化处理。OPE通过将图像切割成许多重叠的图像块,并赋予每个图像块向量化的表征,从而实现对图像信息的深度挖掘。随后,每个图像块都会经过一个预训练的卷积神经网络(Convolutional neural networks, CNN)模型,进而转化为一个高维度的向量表示。MiT能够输出多尺度的特征,并且避免了位置编码的插值问题。

1.1.2 深度特征提取

Transformer具有更好的全局特征提取能力,而卷积神经网络则更关注局部特征。ConvNeXt^[14]网络结构完全基于卷积神经网络,其骨干网络由深度卷积、归一化层、逐点卷积、GELU

激活函数、缩放层和路径丢弃(Drop path)组成,如图2所示。深度卷积用于提取局部特征并处理空间相关性;归一化层对通道数据进行归一化,确保特征分布的稳定性;逐点卷积和GELU激活函数共同增强特征表达能力;路径丢弃则通过随机深度增加正则化效果,有效防止模型过拟合。此外,模块通过残差连接将输入与输出直接相加,从而高效利用深度信息,同时减少激活函数的使用以简化模型。下采样层通过增加Layer Norm以稳定训练过程。经过多个ConvNeXt模块和下采样层后,最终通过全局平均池化提取全局特征,并结合归一化和线性变换完成深度特征的输出。这种设计充分发挥了局部和全局特征的提取能力,使其适用于深度信息分析任务。

1.2 特征校正模块

不同模态的特征中可能存在冗余信息,RGB图像和深度图像都可以反映物体边缘,但深度图中可能含有噪声,利用来自RGB图像模态的特性,可以过滤深度图像的噪音信息^[15]。为此,本文提出了一种新的跨模态特征校正模块用于在特征提取的每个阶段进行特征校正(见图3)。通过

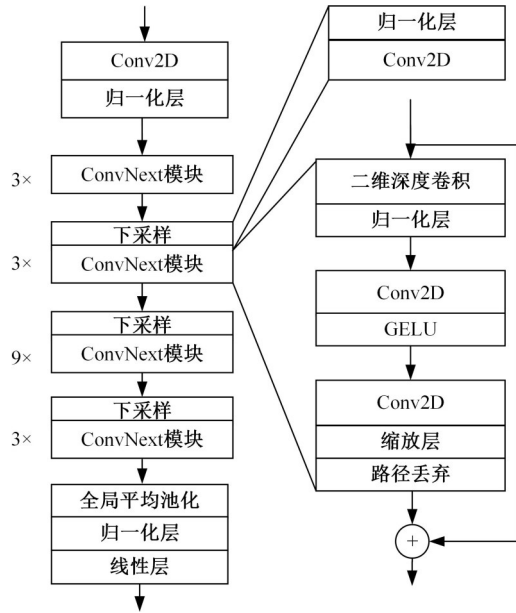


图2 深度特征提取模块

Fig. 2 Depth feature extraction module

跨模态连接和跳跃连接融合多尺度特征,CMF-CM对不同模态下的噪声和不确定性进行处理,将包括通道和空间特征校正在内的两个维度特征进行加权处理,为之后的特征融合提供更准确的特征。

对RGB和深度图像进行通道校正可以确保不同来源的特征被适当加权,提高特征表示的效率。通道特征校正,首先沿通道维度对双模态特征 RGB_{in} 和 $Depth_{in}$ 应用全局最大池化和全局平均池化生成4个特征,通过最大池化减少通道数和空间维度,通过全局平均池化生成每个通道的全

局特征。然后,将生成的4个结果向量拼接连接起来,得到大小 $1 \times 1 \times 4C$ 的特征 X ,再应用多层感知机(Multilayer perceptron, MLP)通过两层全连接层逐步降低特征维度并在每一层后,应用ReLU激活函数进行非线性变换,提高模型的表达能力。最后,将Sigmoid函数归一化的特征拆分为 W_{RGB}^C 和 W_{Depth}^C :

$$W_{RGB}^C, W_{Depth}^C = F_{split}(\text{Sigmoid}(\text{MLP}(X))) \quad (1)$$

式中: F_{split} 为通道拆分操作; W_{RGB}^C 和 W_{Depth}^C 分别为RGB特征和深度特征所对应的通道注意力权重向量。

通道校正可表述为:

$$RGB_{rec}^C = W_{Depth}^C * Depth_{in} \quad (2)$$

$$Depth_{rec}^C = W_{RGB}^C * RGB_{in} \quad (3)$$

式中:“*”代表逐元素相乘。

由于前文提到的通道特征校正模块主要专注于学习全局权重进行全局校准,在此基础上引入了空间特征校正模块,以校准局部信息。双模态输入 RGB_{in} 和 $Depth_{in}$ 将连接并嵌入到两个空间权重矩阵中: $W_{RGB}^S \in R^{H \times W}$ 和 $W_{Depth}^S \in R^{H \times W}$ 。嵌入操作由两个 1×1 卷积与一个ReLU函数构成,通过两层卷积最终输出两个形状为 $2 \times H \times W \times 1$ 的权重张量,再应用Sigmoid函数将输出值限制在 $[0, 1]$ 内作为权重,并将其进一步分割为两个权重映射。获取空间权重矩阵的过程为:

$$W_{RGB}^S, W_{Depth}^S = F_{split}(\text{Sigmoid}(\text{MLP}(Y))) \quad (4)$$

与通道校正类似,空间校正特征可表述为:

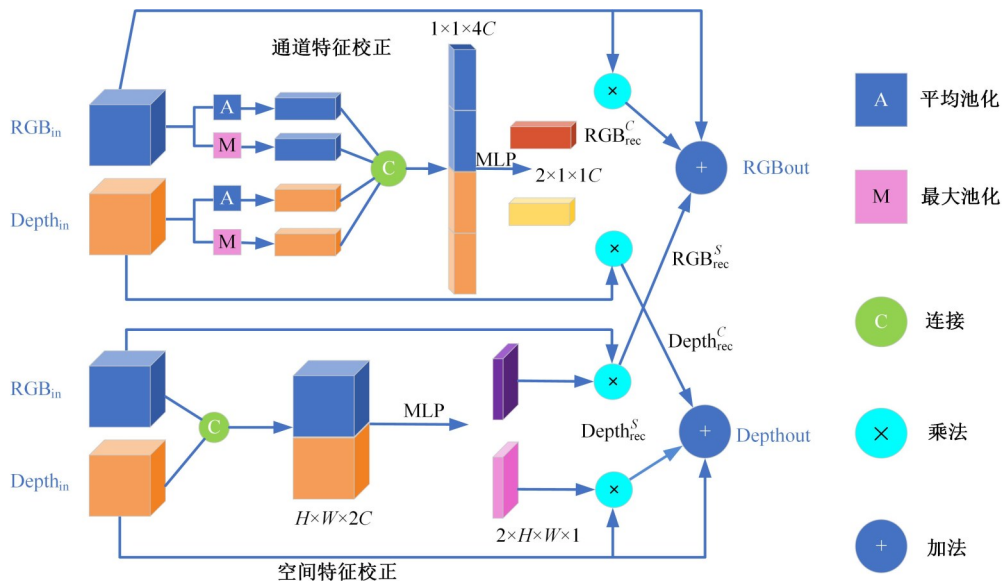


图3 跨模态特征校正模块

Fig. 3 Cross-modal Feature Correction Module

$$\text{RGB}_{\text{rec}}^S = W_{\text{Depth}}^S * \text{Depth}_{\text{in}} \quad (5)$$

$$\text{Depth}_{\text{rec}}^S = W_{\text{RGB}}^S * \text{RGB}_{\text{in}} \quad (6)$$

利用通道和空间上的相互关系,动态调整彩色图像和深度图像的输入特征以提高网络模型的表现。校正过程在原有彩色图像和深度图像输入的基础上,使用通道权重和空间权重分别对校正后的RGB特征和Depth特征进行加权求和,得到最终的校正特征 RGB_{out} 和 $\text{Depth}_{\text{out}}$ 。

$$\text{RGB}_{\text{out}} = \text{RGB}_{\text{in}} + k_C \text{RGB}_{\text{rec}}^C + k_S \text{RGB}_{\text{rec}}^S \quad (7)$$

$$\text{Depth}_{\text{out}} = \text{Depth}_{\text{in}} + k_C \text{Depth}_{\text{rec}}^C + k_S \text{Depth}_{\text{rec}}^S \quad (8)$$

式中: k_C 和 k_S 分别为通道权重和空间权重,均设置为0.5。

1.3 特征融合模块

Transformer中的原始多头自注意力仅限于在单一模态内学习自相似性,当网络输入扩展到多模态时,如何联合利用多模态信息挖掘特征成为表征学习的新目标。因此,本文提出一种多模态自相似度的计算方式,假设在Key和Query中

嵌入RGB和Depth特征,对于像素 (i_0, j_0) 与其他像素 (i, j) 间的跨模态自相似度 $W(i, j)$ 可以定义为:

$$W(i, j) = \sum_{n=1}^N (K r_{n, i, j} \cdot Q r_{n, i_0, j_0}) + \sum_{n=1}^N (K d_{n, i, j} \cdot Q d_{n, i_0, j_0}) \quad (9)$$

式中: $K r_{n, i, j}$ 为RGB像素 (i, j) 第 n 个Key特征向量; $Q r_{n, i, j}$ 为RGB像素 (i, j) 第 n 个Query特征向量; $K d_{n, i, j}$ 、 $Q d_{n, i, j}$ 分别为Depth像素 (i, j) 第 n 个Key和Query特征向量。

现有的跨模态注意力方法主要集中在处理RGB和深度图像单一模态的注意力,而忽视了模态融合后联合特征空间中的注意力,导致模态信息利用不充分;另外,尽管注意力机制可以显著提升模型在多模态学习中的表现,但它的计算复杂度较高^[16],因此必须考虑如何减少计算复杂度。针对以上问题,本文提出了CMAFM模块,计算空间注意力权重,进行特征选择和融合,如图4所示。

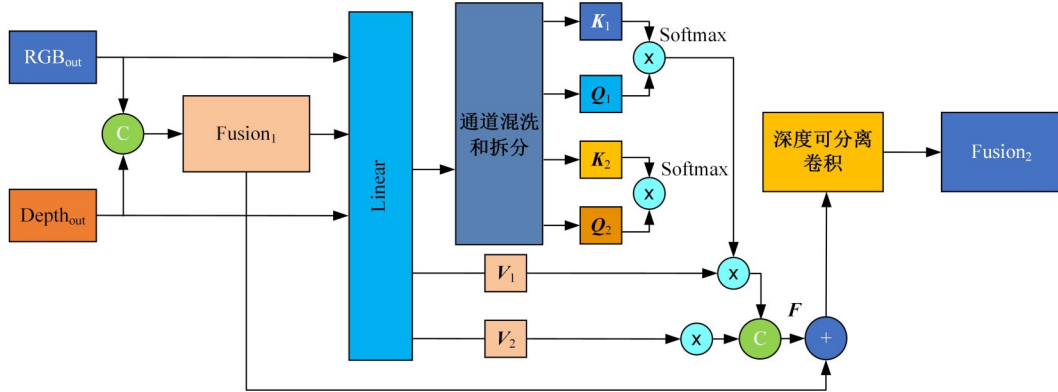


图4 跨模态注意力融合模块

Fig. 4 Cross-modal attention fusion module

在CMAFM的计算过程中,首先将使用校正后的输入特征 RGB_{out} 、 $\text{Depth}_{\text{out}}$ 和拼接特征 Fusion_1 嵌入到向量空间中,生成对应于不同模态特征的Key和Query以及表示拼接特征的Value向量,并使用独立的线性层将不同特征嵌入到向量空间中。计算后,得到嵌入特征向量分别为 K_{RGB} 、 K_{Depth} 、 Q_{RGB} 和 Q_{Depth} ,此外,为了学习两个子空间的特征,将 Fusion_1 的嵌入向量Value在通道方向上划分为两个独立的向量 V_1 和 V_2 。

所提出的跨模态融合模块通过以下步骤挖掘RGB图像与深度图像之间的信息,在通道方向上连接 K_{RGB} 、 K_{Depth} 、 Q_{RGB} 和 Q_{Depth} 得到Key和Query。为了计算两个不同子空间特征的自相似度,需要

将Key和Query分开,目的是为了明确目标特征和参考特征的角色,建立不同子空间之间的联系。但将Key和Query直接分开后不能同时包含RGB和Depth特征。因此,引入了Shuffle机制,执行通道重排操作,从而在网络中增加特征的多样性和表示能力。它将输入张量按给定的组数进行重新排列,以增强特征之间的信息交互。以确保Key和Query的拆分结果 K_1 、 K_2 、 Q_1 和 Q_2 都包含来自两种模态的信息。以Key值为例,处理过程如图5所示。

为了得到融合后的特征,计算跨模态注意力矩阵。首先,计算 Q_1 与 K_1 、 Q_2 与 K_2 的点积,即跨模态自相似度,为了防止数值过大导致梯度爆炸,

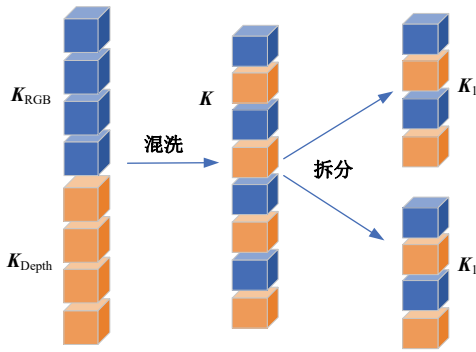


图 5 通道重排

Fig. 5 Channel rearrangement

再除以 $\sqrt{C_1/4}$ 进行缩放, C_1 为输入特征的通道数^[17]。最后,对注意力矩阵进行归一化得到两个子空间的注意力矩阵 W_1 和 W_2 :

$$\begin{cases} W_1 = \text{Softmax}\left(\frac{Q_1 \cdot K_1^T}{\sqrt{C_1/4}}\right), \\ W_2 = \text{Softmax}\left(\frac{Q_2 \cdot K_2^T}{\sqrt{C_1/4}}\right) \end{cases} \quad (10)$$

随后, W_1 和 W_2 通过点积将信息嵌入到 V_1 和 V_2 中,所使用的融合特征 F 是将上述计算结果在通道维度上进行拼接:

$$F = \text{Cat}[W_1 \cdot V_1, W_2 \cdot V_2] \quad (11)$$

式中: Cat 为拼接操作。

最后, CMAFM 通过线性层将计算结果转换为与 Fusion₁ 使用的通道数相同,并逐像素地将其与使用的残差连接相加,并将结果输入到深度可分离卷积 (Depthwise separable convolution, DSC), 如图 6 所示。深度可分离卷积^[18]是一种高效的卷积操作,与传统卷积不同,它将卷积过程分为两步:深度卷积和点卷积。在深度方向对每一通道进行空间卷积,然后使用跨通道的 1×1 卷积进行信息融合。在传统卷积中,卷积核是三维的,每个卷积核会同时作用于输入特征图的所有通道,提取空间和通道的联合特征。而在深度可分离卷积中,深度卷积中的每个卷积核只作用于单个通道,独立提取空间特征,不与其他通道信息融合;随后,通过点卷积将所有通道的信息进行融合,得到融合结果 Fusion₂,并传递给解码器进行解码操作。

2 实验结果及分析

2.1 数据集

本文选用 RGBD 语义分割常见的数据集

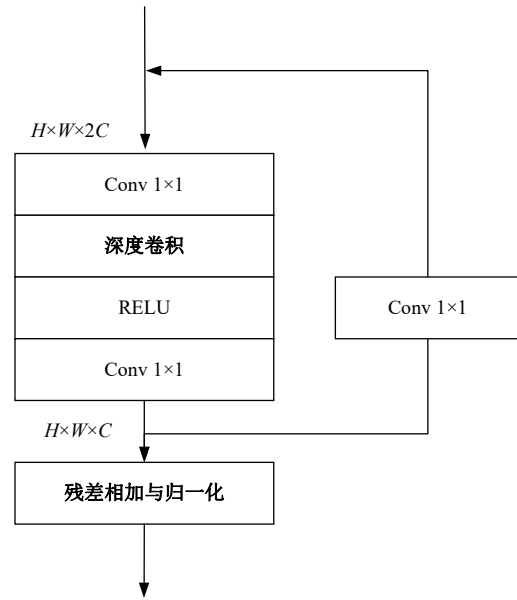


图 6 深度可分离卷积

Fig. 6 Depth-separable convolution

NYU-Depth V2 进行训练。NYU-Depth V2 数据集源自纽约大学,涵盖了各类室内场景的视频序列。这些数据均由 Microsoft Kinect 深度相机采集而成。该数据集提供了总计 1 449 对经过密集标记且精准对齐的 RGB 与深度图像,图像格式统一为 640×480 的分辨率,并且对原始深度图进行补全,在多达 40 个语义类别上进行了详尽的标注,提供了丰富的标签信息。

2.2 实验配置和步骤

本文采用的深度学习框架为 Pytorch,实验主要硬件和环境配置为:Pytorch 版本为 2.0.1; CUDA 版本为 12.0;显卡配置为 NVIDIA GeForce GTX 4090GPU;CPU 为 Intel Xeon Gold 6430 处理器;运行内存 32 GB。采用在 ImageNet 上预训练的 Mit-B2 作为 Backbone 对 RGB 特征进行提取,充分利用预训练模型的特征表达能力。在配置好运行环境后,将预训练模型的参数加载到当前网络中,并使用任务数据集进行微调,通过更新模型权重,使其更好地适应特定任务需求。在所有数据集的训练过程中,通过随机翻转和随机缩放对原始图像数据进行数据增强。训练过程中的部分参数设置如下:输入图像批次设为 8;网络优化方法采用 Adam 优化器;学习率设置为 5×10^{-5} ,最大迭代数为 200 次。

2.3 对比实验

在 NYU-Depth V2 数据集上,将本文模型和 SA-Gate^[5]、CMX-B2^[6]和 Dformer^[7]三个模型的

训练结果进行对比,如表 1 所示。在本文实验中分别对比了像素准确率(Pixel accuracy,PA)和平均交并比(Mean intersection over Union,MIoU)两项评价指标。其中,像素准确率是指分割结果中正确的像素点数占总像素点的比例,其定义为:

$$PA = \frac{\sum_i p_{ii}}{\sum_i \sum_j p_{ij}} \tag{12}$$

式中: $\sum_i p_{ii}$ 为真实类别和预测类别都为 i 的像素总数; $\sum_i \sum_j p_{ij}$ 为真实类别为 i 、预测类别为 j 的像素数量。

表 1 对比实验结果

Table 1 Comparative experimental results			
方法	PA/%	MIoU/%	参数量/M
CMX-B2	82.8	49.7	67
Dformer	82.7	52.0	53
SA-Gate	82.5	49.4	65
本文	83.7	52.8	45

交并比是真实值和预测值两个集合的交集与并集之比,其值越高,表明两者重合部分越大,网络预测性能越好,而平均交并比为所有类别交并比的均值,其表达式如下:

$$MIoU = \frac{1}{N} \sum_{i=1}^N \frac{p_{ii}}{\sum_j p_{ij} + \sum_j p_{ji} - p_{ii}} \tag{13}$$

式中: N 为标签中的类型总数; p_{ii} 为类别 i 预测正确的像素数; $\sum_j p_{ij}$ 为真实类别为 i 的所有像素数; $\sum_j p_{ji}$ 为预测为类别 i 的所有像素数。

分析表 1 结果可知,本文模型的像素准确率相较于 CMX-B2、Dformer 和 SA-Gate 分别提升了 0.9%、1% 和 1.2%,在平均交并比(MIoU)上分别提升了 3.1%、0.8% 和 3.4%。同时,参数量仅为 45 M,显著低于 CMX-B2(67 M)、Dformer(53 M)和 SA-Gate(65 M),在保证分割精度提升的同时大幅降低了模型复杂度,体现出优越的精度和效率平衡能力。

选取数据集集中的 6 组图像进行可视化分析,结果如图 7 所示。图 7 中,第 1 列为 RGB 图像,第 2 列为深度图像,第 3 列为深度标签图像,第 4~6 列为 CMX-B2、Dformer、SA-Gate 网络训练结果,第 7 列为本文模型的训练结果。由图 7(a)可见,本文模型对较小的灯开关成功分割并正确分类;由图 7(b)可见,本文模型对桌子边缘分割效果更好,且其他网络模型对壁橱的分割效果并不理想;图 7(c)中,只有本文和 SA-Gate 的训练结果成功分割了椅子的缝隙;通过结合彩色图像,本文模型能够检测到图 7(d)桌面上较小的台球;图 7(e)中,针对远处的人和花瓶其他模型未能完成较为明显的分割,对椅子的分割效果并不理想;

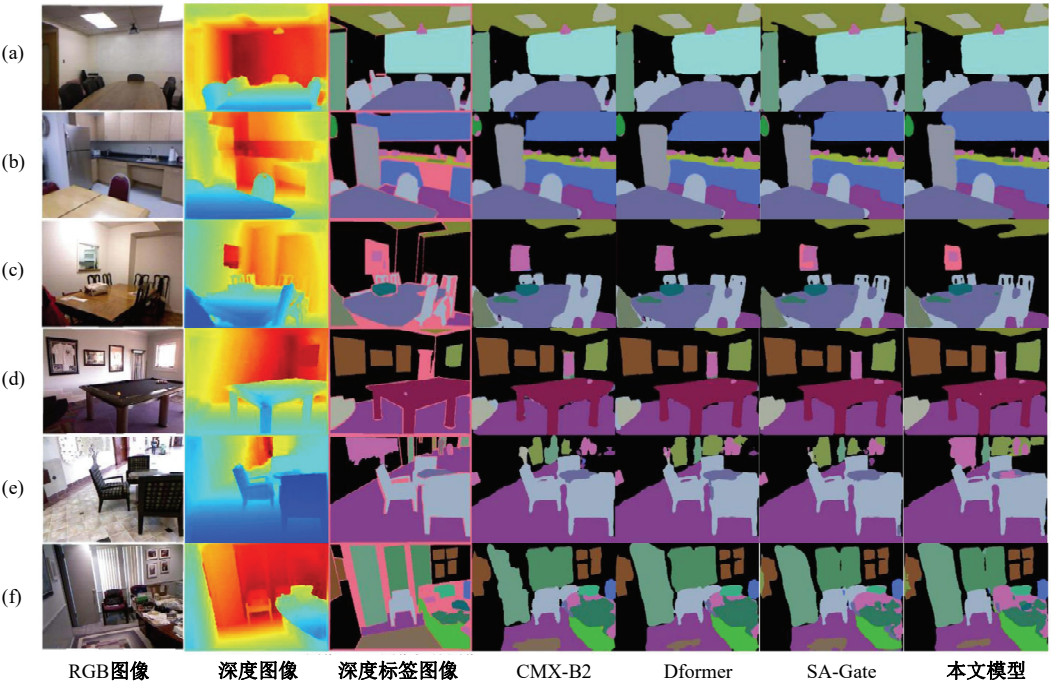


图 7 可视化分割结果对比

Fig. 7 Comparison of visual segmentation results

图 7(f)中,只有本文的分割结果更接近标签图像。

综上所述,本文模型通过有效融合 RGB 和深度图像信息,提升了对小目标、边缘区域、复杂结构和远距离目标的分割性能,与现有的 CMX-B2、Dformer 和 SA-Gate 模型相比,在细节恢复和全局一致性方面具有明显优势,适用于更复杂的场景分割任务。

2.4 消融实验

为分析本文 RGB 图像和深度图像融合方法以及模型中的主要模块 CMFCM 和 CMAFM 对模型总体性能的具体影响,进一步验证模型中模块设计的合理性、有效性以及多模态融合相较于单模态的优势,在保留特征提取模块的基础上,对各模块通过控制变量的方法进行消融实验,实验结果表 2 所示。

表 2 消融实验结果
Table 2 Ablation results

CMFCM	CMAFM	输入	PA/%	MIoU/%
×	×	RGB	72.1	47.6
×	×	RGB-D	74.7	48.5
✓	×	RGB-D	76.6	49.8
×	✓	RGB-D	77.2	50.5
✓	✓	RGB-D	83.7	52.8

从实验结果可以得到以下几个结论:

(1)在未引入任何模块的条件下,双通道输入(RGB-D)相比单通道 RGB 输入,PA 提升了 2.6%,MIoU 提升了 0.9%。虽然提升幅度有限,但仍显示出深度信息作为辅助模态在提升分割精度方面的潜力,验证了引入 Depth 通道分支的合理性和必要性。

(2)在单模块加入的对比实验中,分别分析了 CMFCM 和 CMAFM 对模型性能的影响。当仅引入 CMFCM 时,PA 提升了 1.9%,MIoU 提升了 1.3%。这表明 CMFCM 在一定程度上改善了多模态间的低层特征对齐,但由于其作用主要集中在特征校正阶段,对整体语义理解能力的提升有限。而在仅引入 CMAFM 的情况下,PA 和 MIoU 分别提升了 2.5% 和 2.0%,说明该模块通过注意力机制有效加强了模态间语义信息的融合,从而显著提升了模型的全局感知能力。

(3)当同时引入 CMFCM 和 CMAFM 时,模型在 PA 和 MIoU 上分别达到了 83.7% 和 52.8%,这一结果表明,尽管 CMAFM 在融合高层语义信息方面表现突出,但在其基础上进一步

引入 CMFCM 能够在低层特征校正的基础上实现多层次、多维度的信息融合,进一步提升了模型整体性能。相较于单独引入单一模块,两模块协同工作表现出更强的互补性,验证了本文模块设计的合理性和有效性。

综上所述,多模态输入相较于单模态的输入更具优势,并且通过 CMFCM 和 CMAFM 的配合使用,评价指标相较于两个模块单独作用时的提升更为显著,模型总体性能大幅改善。这进一步证明了本文模型不仅整体上较其他模型更具优势,并且各个模块的设计也都具备合理性。

3 结束语

针对语义分割任务中现有模型融合不充分、参数量大等问题,提出了一种结合 RGB 和深度图像的双模态特征融合网络。特征提取部分采用 ConvNeXt 和 Mix-Transforme 对 RGB 和深度模态的特征进行提取,保证了局部和全局的特征提取能力。通过引入 CMFCM,实现了 RGB 与深度图像特征之间的校正补偿,有效缓解了模态间信息对齐和噪声干扰的问题。同时,CMAFM 融合了多尺度特征的信息,加强了不同模态信息的表征能力。在公开数据集 NYU-Depth V2 上的实验结果表明,本文模型在语义分割任务中表现优异,较现有方法在多个评估指标上取得了提升,分割精度提升 1% 的同时参数量减少了 15.1%。同时,模型在不同场景下均展示出了较强的鲁棒性,尤其在光照变化、遮挡和深度噪声等复杂条件下,依然能够保持较高的分割性能。最后,通过多组消融实验验证了各模块的有效性。总体而言,本文方法有效解决了传统语义分割方法中对单一模态依赖较强、模态间信息利用不足、模型参数量大的问题,为双模态语义分割任务提供了一种新思路。

参考文献:

- [1] 王文俊. 基于深度卷积神经网络的点云数据语义分割方法研究[D]. 成都:电子科技大学信息与通信工程学院, 2024.
Wang Wen-jun. Research on semantic segmentation of point cloud data based on deep convolutional neural networks[D]. Chengdu: School of Information and Communication Engineering, University of Electronic Science and Technology of China, 2024.

- [2] Jiang Jin-dong, Zheng Lu-nan, Luo Fei, et al. Red-net: residual encoder-decoder network for indoor RGB-D semantic segmentation[J/OL]. [2025-02-10]. <https://doi.org/10.48550/arXiv.1806.01054>.
- [3] Seichter D, Köhler M, Lewandowski B, et al. Efficient RGB-D semantic segmentation for indoor scene analysis[C]//2021 IEEE International Conference on Robotics and Automation, Xi'an, China, 2021: 13525-13531.
- [4] Du S Q, Tang S J, Wang W X, et al. PSCNET: efficient RGB-D semantic segmentation parallel network based on spatial and channel attention[J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2022, 1: 129-136.
- [5] Chen X, Lin K Y, Wang J, et al. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation[C]//European Conference on Computer Vision, Glasgow, UK, 2020: 561-577.
- [6] Zhang Jia-ming, Liu Hua-yao, Yang Kai-lun, 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.
- [7] Yin Bo-wen, Zhang Xu-ying, Li Zhong-yu, et al. Dformer: rethinking RGBD representation learning for semantic segmentation[J/OL]. [2025-02-10]. <https://arxiv.org/abs/2309.09668v2>.
- [8] Bui M, Alexis K. Diffusion-based RGB-D semantic segmentation with deformable attention transformer[J/OL]. [2025-02-10]. <https://arxiv.org/html/2409.15117v3>.
- [9] Dong S, Feng Y, Yang Q, et al. Efficient multimodal semantic segmentation via dual-prompt learning [C]//2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, Abu Dhabi, UAE, 2024: 14196-14203.
- [10] 缪君, 严杰, 杜荣华, 等. 基于双向特征融合的物体位姿估计方法[J]. 吉林大学学报: 工学版, 2026, 56(2): 523-532.
Miao Jun, Yan Jie, Du Rong-hua, et al. Object pose estimation method based on bidirectional feature fusion[J]. Journal of Jilin University (Engineering and Technology Edition), 2026, 56(2): 523-532.
- [11] 王雪, 李占山, 吕颖达. 基于多尺度感知和语义适配的医学图像分割算法[J]. 吉林大学学报: 工学版, 2022, 52(3): 640-647.
- [12] 周大可, 张超, 杨欣. 基于多尺度特征融合及双重注意力机制的自监督三维人脸重建[J]. 吉林大学学报: 工学版, 2022, 52(10): 2428-2437.
Zhou Da-ke, Zhang Chao, Yang Xin. Self-supervised 3D face reconstruction based on multi-scale feature fusion and dual attention mechanism[J]. Journal of Jilin University (Engineering and Technology Edition), 2022, 52(10): 2428-2437.
- [13] Xie E, Wang W, Yu Z, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.
- [14] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 11976-11986.
- [15] Song Y, Wen J, Liu D, et al. Deep robotic grasping prediction with hierarchical RGB-D fusion[J]. International Journal of Control, Automation and Systems, 2022, 20(1): 243-254.
- [16] Liu Z, Tan Y, He Q, et al. SwinNet: swin transformer drives edge-aware RGB-D and RGB-T salient object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(7): 4486-4497.
- [17] 钱白云, 吕朝阳, 张维宁, 等. 基于多传感器信息融合与混合感受野残差卷积神经网络的调相机转子故障诊断[J]. 计算机测量与控制, 2023, 31(9): 29-35.
Qian Bai-yun, Lv Chao-yang, Zhang Wei-ning, et al. Phase condenser rotor fault diagnosis based on multi-sensor information fusion and mixed receptive field residual convolutional neural network[J]. Computer Measurement & Control, 2023, 31(9): 29-35.
- [18] 孙启超, 恩擎, 段立娟, 等. 基于多模态自适应卷积的RGB-D图像语义分割[J]. 计算机辅助设计与图形学学报, 2022, 34(8): 1272-1282.
Sun Qi-chao, En Qing, Duan Li-juan, et al. RGB-D image semantic segmentation based on multimodal adaptive convolution[J]. Journal of Computer-Aided Design & Graphics, 2022, 34(8): 1272-1282.