文章编号:1007-5321(2025)05-0159-08 **DOI**:10.13190/j.jbupt.2024-151

基于非对称特征校正与融合的 RGB-D 语义分割

游新冬[1], 沈文涛[1], 韩 晶[1], 吕学强[1,2], 才藏太[2]
(1. 北京信息科技大学 网络文化与数字传播北京市重点实验室, 北京 100101;
2. 青海师范大学 省部共建藏语智能信息处理及应用国家重点实验室, 西宁 810008)

摘要: 针对红绿蓝 3 通道颜色模型(RGB)单模态包含语义信息单一、易受噪声干扰且分割性能不佳的问题,提出了一种基于非对称特征交互方法的红绿蓝 3 通道颜色模型-深度信息(RGB-D)语义分割算法。首先,使用双流网络分别提取 RGB 模态和深度模态特征,并通过非对称特征校正模块,以利用一种模态校正另一种模态的特征,达到抑制模态内噪声的效果。然后,通过非对称融合模块进一步增强模态间的信息交互。此外,在解码器中引入了多尺度特征融合,并在训练过程中采用对抗性训练作为辅助,从而有效利用上下文信息并整体提升准确性。实验结果表明,所提算法有效抑制了模态内的噪声,增强了模态间有效语义信息的交互,在纽约大学深度数据集第 2 版(NYUDepthv2)和斯坦福大学 RGB-D 数据集(SUN-RGBD)上的平均交并比(mIoU)分别达到 57.4% 和 52.1%。

关 键 词: 红绿蓝 3 通道颜色模型-深度信息语义分割; 编码器-解码器; 红绿蓝 3 通道颜色模型-深度信息互补; 深度学习
中图分类号: TP391.41 **文献标志码**: A

RGB-D Semantic Segmentation Based on Asymmetric Feature Rectification and Fusion

YOU Xindong[1], SHEN Wentao[1], HAN Jing[1], LYU Xueqiang[1,2], CAI Zangtai[2]
(1. Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing 100101, China;
2. The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Qinghai Normal University, Xining 810008, China)

Abstract: To address the issue that the single red-green-blue (RGB) modality contains limited semantic information, is susceptible to noise interference, and exhibits suboptimal segmentation performance, this paper proposes an RGB-depth (RGB-D) semantic segmentation algorithm based on an asymmetric feature interaction method. First, a two-stream network is employed to extract features from the RGB and depth modalities separately. By incorporating an asymmetric feature correction module, features from one modality are used to correct those of the other, thereby suppressing intra-modal noise. Then, an asymmetric fusion module is applied to further enhance information interaction between the modalities. Additionally, multi-scale feature fusion is introduced in the decoder, and adversarial training is adopted as an auxiliary strategy during the training process to effectively leverage contextual information and improve overall accuracy. Experimental results demonstrate that the proposed algorithm effectively

收稿日期: 2024-07-17
基金项目: 国家自然科学基金项目(62171043);北京市自然科学基金项目(4232025);北京市教委科研计划科技一般项目(KM202311232003);青海省创新平台建设专项(2022-ZJ-T02)
作者简介: 游新冬(1979—), 女, 教授, 硕士生导师。
通信作者: 韩 晶(1990—), 女, 讲师, 硕士生导师, 邮箱:hanjing@bistu.edu.cn。

suppresses intra-modal noise and enhances the interaction of valid semantic information across modalities, achieving mean intersection over union (mIoU) scores of 57.4% and 52.1% on the New York University depth dataset v2 (NYUDepthv2) and the Stanford University RGB-D dataset (SUN-RGBD), respectively.

Key words: red-green-blue-depth semantic segmentation; encoder-decoder; red-green-blue-depth information complementary; deep learning

近年来,基于红绿蓝3通道颜色模型-深度信息(RGB-D, red-green-blue-depth)数据的语义分割任务的主流方法之一是采用单网络架构实现模态交互。例如,Cao等[1]在单个主干网络中使用形状感知卷积层,动态调整RGB卷积权重。Chen等[2]则设计空间引导卷积,利用深度生成几何亲和矩阵优化特征聚合。这类方法通过在卷积操作中嵌入几何建模,显著超越传统输入级融合策略。然而,它们仍受限于深度噪声敏感性与严格模态对齐依赖,在复杂场景中泛化能力不足。

除了前述基于单个主干网络的算法之外,另一种主流算法是采用双分支编码解码器架构,RGB图像和深度图像的特征分别由2个单独的主干网络提取,并添加额外的特征融合模块实现2种模态间的交互,提高了性能。例如,Hu等[3]先使用通道注意力机制分别处理2种模态的特征,之后通过简单的相加操作实现尺度间和模态间的信息交互。Chen等[4]引入一种跨模态引导编码器,以期利用2种模态的通道相关性和空间相关性,通过模态间的交互抑制特征噪声,并校正RGB和深度特征。此外,Wang等[5]则通过模态间的通道交换实现信息融合,在没有引入额外参数的情况下动态地控制融合过程。不过,这些算法均采用基于卷积神经网络(CNN, convolutional neural network)的主干网络,没有考虑远程依赖。

随着变换器架构在自然语言处理领域的巨大成功,将其引入视觉领域的语义分割任务也成为了一个流行趋势。Xie等[6]利用变换器架构创建层次结构来提取多分辨率特征,并用于RGB图像的语义分割。基于这一工作,Zhang等[7]提出的跨模态融合语义分割方法(CMX, cross-modal fusion for RGB-X)采用双分支架构,提出了一个特征校正模块,使用通道和空间注意力聚合跨模态特征,并通过交叉注意力实现了跨模态特征融合。Gao等[8]则通过交叉注意力校正特征,通过跨场融合模块融合特征,并引入边界监督优化边界。然而,上述算法虽然利用了远程依赖,但是特征校正模块和融合模块均采用对称架构,没有考虑到RGB特征和深度特征包含语义信息的差异性。

针对以上问题,笔者提出了一种基于非对称特征校正与融合的RGB-D语义分割算法(AFRF-Seg, asymmetric feature rectification and fusion segmentation)。该算法在充分利用不同模态信息的基础上,采用了非对称架构,在特征校正与融合阶段进行模态间的信息交互。具体是在编码器的每个阶段设计了一种非对称的特征校正模块(AFR, asymmetric feature rectification)和非对称的特征融合模块(AFF, asymmetric feature fusion)。其中,AFR模块使用与交叉注意力相似的架构获得基于RGB特征的全局特征评估,使用普通卷积获得局部特征评估,并共同用于2种模态的特征校正。AFF模块则使用交叉注意力实现深度特征对RGB特征的互补融合。考虑到校正模块和融合模块目的不同,2个模块进行交叉注意力的操作也进行了差异化处理。此外,为了实现尺度间的特征融合,AFRF-Seg基于空洞空间金字塔池化提出了一种语义分割解码器,以利用来自不同编码器层级的特征之间的上下文信息。在训练过程中,还采用了对抗性训练策略,以进一步提升模型的整体准确性。实验结果显示,AFRF-Seg在2个公开的数据集上表现出色,相较于现有的RGB-D语义分割算法,实现了更加优秀的场景分割效果。

1 算法原理

1.1 整体框架

AFRF-Seg基于双分支的编码解码器的网络架构,使用2个混合变换器网络(MiT, mix transformer)分别从RGB图像和深度图像中提取特征。如图1所示,这2个并行分支使用阶段性交互的方式,通过AFR模块用一种模态的特征校正另一种模态。被校正后的RGB模态和深度模态特征,分别被送入解码器的下一阶段,同时在经过

AFF 模块后被送入含有多尺度特征融合的解码器中。此外，受 Taghavi 等[9]工作的启发，采用了对抗性训练的方法，通过衡量预测和真实分布之间的分布差异，提高算法的整体准确性。整体的损失函数如下：

$$L = L_{\text{seg}} + \lambda L_{\text{gan}} \tag{1}$$

其中：L_{seg} 表示分割损失，L_{gan} 表示判别器损失，λ 为超参数，默认设置为 0.1。

网络的整体流程如图 1 所示。其中，特征提取阶段采用的 MiT 编码器能够在给定输入图像的情况下，生成类似卷积神经网络的多级特征，这些特征包含高分辨率的细粒度特征和低分辨率的粗粒度特征，通常可以提高语义分割的性能。在解码器阶段中，由 1 × 1 卷积实现的多层感知机（MLP，multilayer perceptron）用于将每个阶段输出的特征嵌入到相同数量的通道中。

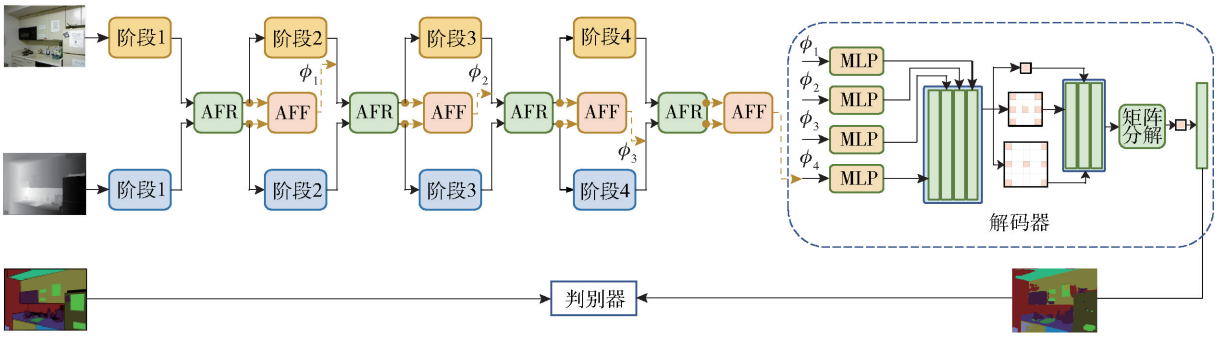

图 1　整体网络结构图

1.2　非对称特征校正模块

编码器阶段的 RGB 特征和深度特征通常是互补的，且 2 种模态的数据在采集和特征提取的过程中，不可避免地会引入噪声。从原始图像的角度进行解释，当 RGB 图像中的颜色或纹理过于接近难以区分时，模型应该更加关注深度图像。同样地，当深度图像中的深度值难以用来区分不同目标类别时，模型应该更加关注 RGB 图像。此外，特征校正还可以用其中一种模态抑制另一种模态中噪声的影响。如图 2 所示，AFR 模块对 2 种模态的交互处理包括基于交叉注意力的全局特征评估和基于卷积的局部特征评估，2 者共同提供特征校正。

1）基于交叉注意力的全局特征评估。考虑到 RGB 模态包含的语义信息更加丰富，AFR 模块在进行全局特征评估时，采用非对称架构，仅计算 2 种模态对 RGB 模态的注意力。同时，为了减少因像素增加而 2 次增长的计算量，先将 2 种模态的特征在通道方向上串联，然后使用自适应平均池化对串联后的特征进行下采样，再通过 1 个 1 × 1 卷积得到 2 种模态共同的查询 $\boldsymbol{Q}_{\text{AFR}}$，以将 2 次全局注意力的计算合并为 1 次。为了获得计算注意力时所需要的键值（$\boldsymbol{K}_{\text{AFR}}$，$\boldsymbol{V}_{\text{AFR}}$），仅对 RGB 特征进行变换。上述过程可以表述为

$$\boldsymbol{Q}_{\text{AFR}} = B_{1\times1}(P_{k_1\times k_2}(F_{\text{cat}}(\boldsymbol{X}_{\text{in}}^{\text{rgb}}, \boldsymbol{X}_{\text{in}}^{\text{d}}))) \tag{2}$$

$$\boldsymbol{K}_{\text{AFR}}, \boldsymbol{V}_{\text{AFR}} = F_{\text{split}}(B_{1\times1}(\boldsymbol{X}_{\text{in}}^{\text{rgb}})) \tag{3}$$

其中：$F_{\text{cat}}(\cdot)$ 表示沿通道方向的串联操作，$P_{k_1\times k_2}(\cdot)$ 表示在空间维度的自适应平均池化，将特征图下采样到 $k_1 \times k_2$ 大小，$B_{1\times1}(\cdot)$ 是 1 个 1 × 1 普通卷积，用来执行通道间的线性变换，$F_{\text{split}}(\cdot)$ 表示按通道方向进行划分。基于生成的 $\boldsymbol{Q}_{\text{AFR}} \in R^{k_1\times k_2\times C}$，$\boldsymbol{K}_{\text{AFR}} \in R^{h\times w\times C}$ 和 $\boldsymbol{V}_{\text{AFR}} \in R^{h\times w\times C}$（其中，$h$，$w$ 和 C 是当前阶段特征图的高度、宽度和通道数），全局特征评估如下：

$$\boldsymbol{X}_{\text{attn}} = F_{\text{UP}}\left(\boldsymbol{V}_{\text{AFR}} @ \text{softmax}\left(\frac{\boldsymbol{Q}_{\text{AFR}} @ \boldsymbol{K}_{\text{AFR}}^{\text{T}}}{\sqrt{C/N_{\text{head}}}}\right)\right) \tag{4}$$

其中：$F_{\text{UP}}(\cdot)$ 表示双线性插值以进行上采样，将特征图大小从 $k_1 \times k_2$ 转换为 $h \times w$，@ 表示矩阵乘法。

2）基于卷积的局部特征评估。局部特征评估需要综合对比 2 种模态特征包含的语义信息，利用普通卷积实现通道间的交互。同时，由于特征包含语义信息的质量需要考虑一定大小区域内的信息，因此卷积模块使用了 3 × 3 的卷积核，以实现空间方向的对比评估。具体操作如下：

$$\boldsymbol{X}_{\text{local}} = B_{3\times3}(\text{ReLU}(B_{3\times3}(F_{\text{cat}}(\boldsymbol{X}_{\text{in}}^{\text{rgb}}, \boldsymbol{X}_{\text{in}}^{\text{d}})))) \tag{5}$$

$$\boldsymbol{X}_{\text{local}}^{\text{rgb}}, \boldsymbol{X}_{\text{local}}^{\text{d}} = F_{\text{split}}(\text{sigmoid}(\boldsymbol{X}_{\text{local}})) \tag{6}$$

其中 $B_{3\times3}$ 表示 3 × 3 的普通卷积。

与 CMX 类似，获得 2 种模态共同的全局特征评估和各自的局部特征评估后，采用如下的方法对特征图进行校正：

$$\boldsymbol{X}_{\text{out}}^{\text{rgb}} = \lambda_1 \boldsymbol{X}_{\text{attn}} + \lambda_2 \boldsymbol{X}_{\text{local}}^{\text{d}} \boldsymbol{X}_{\text{in}}^{\text{d}} + \boldsymbol{X}_{\text{in}}^{\text{rgb}} \tag{7}$$

$$\boldsymbol{X}_{\text{out}}^{\text{d}} = \lambda_1 \boldsymbol{X}_{\text{attn}} + \lambda_2 \boldsymbol{X}_{\text{local}}^{\text{rgb}} \boldsymbol{X}_{\text{in}}^{\text{rgb}} + \boldsymbol{X}_{\text{in}}^{\text{d}} \tag{8}$$

其中：λ_1 和 λ_2 是2个超参数，遵从CMX的设置，均设为默认值0.5，$\boldsymbol{X}_{\text{out}}^{\text{rgb}}$ 和 $\boldsymbol{X}_{\text{out}}^{\text{d}}$ 是综合校正后的特征，被送入主干网络的下一个阶段，同时送入融合模块进行各个阶段的特征融合。

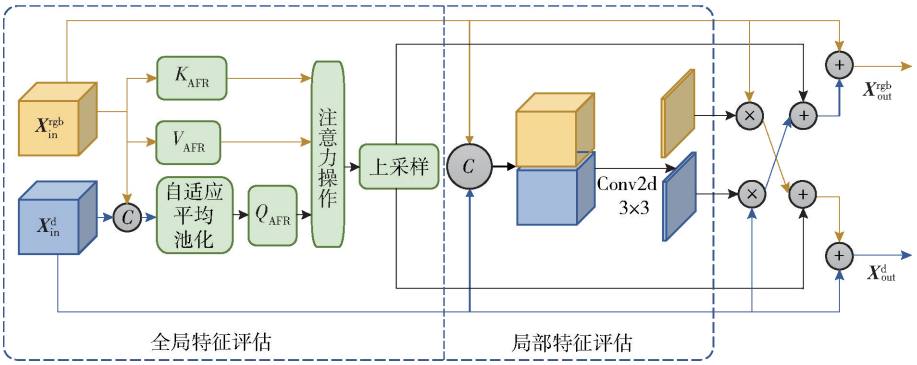

图2 非对称特征校正模块AFR

1.3 非对称特征融合模块

特征校正模块利用2种模态之间的交互，对各自包含的噪声进行抑制，并将处理后的特征送到主干网络的下1个阶段进行特征提取。为了在各个阶段的特征送入到解码器前增强信息的交互，构建了1个特征融合模块。考虑到2种模态语义信息的差异性，该特征融合模块仍使用非对称架构，整体结构如图3所示。

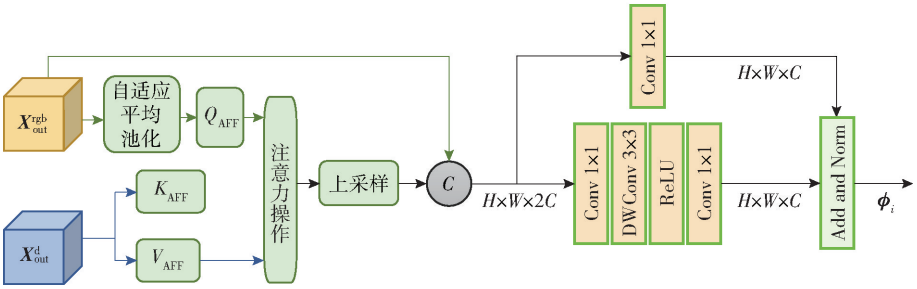

图3 非对称特征融合模块AFF

在进行特征融合时，以RGB特征为主，通过交叉注意力获得RGB特征对深度特征的关注，从而提取出深度特征中对RGB特征有效的互补信息。具体做法与特征校正模块相似，为了减少提取互补信息时的计算量，使用自适应平均池化将RGB特征下采样并通过1个1×1卷积得到查询 $\boldsymbol{Q}_{\text{AFF}}$，与通过深度特征得到的键值（$\boldsymbol{K}_{\text{AFF}}$，$\boldsymbol{V}_{\text{AFF}}$）进行交叉注意力操作，然后使用双线性插值将注意力结果上采样至输入大小。具体操作过程如下：

$$\boldsymbol{Q}_{\text{AFF}} = B_{1\times1}(P_{k_1\times k_2}(\boldsymbol{X}_{\text{out}}^{\text{rgb}})) \tag{9}$$

$$\boldsymbol{K}_{\text{AFF}}, \boldsymbol{V}_{\text{AFF}} = F_{\text{split}}(B_{1\times1}(\boldsymbol{X}_{\text{out}}^{\text{rgb}})) \tag{10}$$

$$\boldsymbol{X}_{\text{cross-attn}} = F_{\text{UP}}\left(\boldsymbol{V}_{\text{AFF}} @ \text{sigmoid}\left(\frac{\boldsymbol{Q}_{\text{AFF}} @ \boldsymbol{K}_{\text{AFF}}^{\text{T}}}{\sqrt{C/N_{\text{head}}}}\right)\right) \tag{11}$$

经过上述操作，得到了深度特征中对RGB特征具有互补作用的语义信息。将该互补信息和RGB特征在通道方向上进行串联，然后使用如图3所示的跳跃连接，将大小为 $R^{H\times W\times 2C}$ 的串联特征融合为大小为 $R^{H\times W\times C}$ 的特征，以进行特征解码。

1.4 多尺度融合的解码器

先前使用变换器主干网络的语义分割工作通常仅在解码器中利用局部信息。由于早期特征为图像的语义分割提供了有价值的低级语义信息，AFRF-Seg的解码器为了增加算法性能和稳健性，不仅考虑了瓶颈特征的上下文，还利用来自不同编码器层级的上下文信息。解码器的架构如图1所示。在进行多尺度特征融合之前，通过1×1卷积将每个 $\boldsymbol{\phi}_i$ 嵌入到相同数量的通道中，然后将各个层级的特征上采样到 $\boldsymbol{\phi}_1$ 大小，并将它们拼接起来。之后使用多个并行的3×3深度可分离卷积和1个1×1卷积进行多尺度特征融合。受Geng等[10]工作的启发，在后续阶段加入矩阵分解，以保证解码器处理后的

信息不会因信息冗余或缺失而被损坏。最终，通过 1 个 1×1 卷积输出分割结果。

2　实验

为了验证 AFRF-Seg 对于 RGB-D 语义分割任务的有效性，分别在 2 个 RGB-D 数据集上进行实验，其评价指标包括平均交并比（mIoU，mean intersection over union）和像素精度（PA，pixel accuracy），并与最先进的算法进行了比较。同时，进行全面的各个模块的消融研究，以验证不同模块的效果，最后进行了可视化实验。

2.1　数据集

在纽约大学深度数据集第 2 版（NYUDepthv2，New York University depth dataset v2）和斯坦福大学 RGB-D 数据集（SUN-RGBD，Stanford University RGB-D dataset）上对 AFRF-Seg 进行微调和评估。NYUDepthv2 数据集包含 1 449 张带有 40 类标签的 RGB-D 图像，其中 795 张图像用于训练，其余 654 张图像用于测试。所有 RGB 图像和深度图像的分辨率统一为 480×640。SUN-RGBD 包含 10 335 张分辨率为 530×730 的 RGB-D 图像，其中对象分为 37 个类别，5 285 张图像用于训练，其余 5 050 张图像用于测试。

2.2　实验方法

AFRF-Seg 采用预训练的 MiT 编码器作为主干网络，并将多尺度融合阶段的嵌入维度设置为 256，矩阵分解时的嵌入维度设置为 512。特征提取网络选择权重衰减为 0.01 的 AdamW 优化器，其中 NYUDepthv2 数据集的初始学习率设置为 0.000 06，SUN-RGBD 数据集的初始学习率设置为 0.000 08，并采用交叉熵作为损失函数。对抗性训练中的鉴别器则统一采用衰减为 0.000 1、初始学习率为 0.000 1 的设置。在微调过程中，2 个数据集均被裁剪为分辨率，并只采用 2 种常见的数据增强策略，即随机水平翻转和随机缩放（从 0.50 到 1.75）。对 MiT-B2 和 B3 模型的批量大小均设置为 8，其中 NYUDepthv2 数据集训练的轮次数设置为 500，SUN-RGBD 数据集训练的轮次数设置为 300。

与训练过程类似，在评估 NYUDepthv2 和 SUN-RGBD 的测试结果时，同样使用带有水平翻转的多个尺度（从 0.50 到 1.75）进行推理。所有实验均在 4 个 V100 型号的图形处理器上进行。

2.3　实验结果分析与评价

2.3.1　对比实验

将 AFRF-Seg 与 4 种最新的 RGB-D 语义分割算法的实验结果进行比较，结果如表 1 所示。从表 1 中可以看出，AFRF-Seg 取得了良好的性能水平。在同样采用 MiT-B2 作为主干网络的情况下，AFRF-Seg 在 NYUDepthv2 数据集中的 mIoU 相比 TokenFusion[11] 和 CMX 分别提升了 2.5% 和 1.4%，在 SUN-RGBD 上的 mIoU 和 PA 相比 CMX 则分别提高了 1.2% 和 0.5%。值得注意的是，AFRF-Seg 在采用 MiT-B3 作为主干网络时，在 NYUDepthv2 数据集上的性能已超过了基于 MiT-B5 的 CMX 和基于

表 1　NYUDepthv2 和 SUN-RGBD 的实验结果

模型	主干网络	参数量/M	NYUDepthv2				SUN-RGBD			
			输入尺寸	运算量/G	mIoU/%	PA/%	输入尺寸	运算量/G	mIoU/%	PA/%
TokenFusion	MiT-B2	26.0	480×640	55.2	53.3	–	530×730	71.1	–	–
TokenFusion	MiT-B3	45.9	480×640	94.4	54.2	–	530×730	122.1	–	–
CMX	MiT-B2	66.6	480×640	67.6	54.4	79.9	530×730	86.3	49.7	82.8
CMX	MiT-B4	139.9	480×640	134.3	56.3	79.9	530×730	173.8	52.1	83.5
CMX	MiT-B5	181.1	480×640	167.8	56.9	80.1	530×730	217.6	52.4	83.8
CMNext	MiT-B4	119.6	480×640	131.9	56.9	–	530×730	170.3	51.9	–
DFormer	DFormer-T	6.0	480×640	11.8	51.8	–	530×730	15.1	48.8	–
DFormer	DFormer-S	18.7	480×640	25.6	53.6	–	530×730	33.0	50.0	–
DFormer	DFormer-B	29.5	480×640	41.9	55.6	–	530×730	54.1	51.2	–
DFormer	DFormer-L	39.0	480×640	65.7	57.2	–	530×730	83.3	52.5	–
AFRF-Seg	MiT-B2	73.0	480×640	56.3	55.8	79.7	530×730	72.3	50.9	83.3
AFRF-Seg	MiT-B3	112.6	480×640	91.1	57.4	80.5	530×730	117.6	52.1	83.6

MiT-B4 的跨模态下一代模型(CMNext, cross-modal next generation model)[12],在参数量和运算量较小的情况下,均获得了 0.5% 的 mIoU 提升,在 SUN-RGBD 数据集上的 PA 仅比基于 MiT-B5 的 CMX 低了 0.2%。同时,与基于强大的 RGB-D 预训练模型的深度变换器(DFormer, depth transformer)[13]相比,AFRF-Seg 作为基于双流网络的算法,虽然参数量和运算量较大,但是在性能水平上亦有优势。在不同数据集上的实验结果表明,AFRF-Seg 可以有效构建 RGB 特征和深度特征之间的互补融合,从而更加准确地进行 RGB-D 数据的语义分割。

2.3.2 消融实验

为了探索算法的不同模块如何影响分割性能,还进行了一系列的消融实验。除非另有说明,在消融实验中均使用 MiT-B2 作为主干网络,以模型在 NYUDepthv2 测试集上的语义分割性能为标准进行评估。

1)AFR 和 AFF 模块的有效性。为了验证所提模块的有效性,把去除校正模块且以特征图相加作为融合方法的模型作为基准模型,之后分别添加 AFR 和 AFF 模块进行实验。此外,为了证明 AFR 和 AFF 模块的有效性,对 CMX 模型中引入的跨模态特征校正模块(CM-FRM, cross-modal feature rectification module)和特征融合模块(FFM, feature fusion module)也进行了实验。

表 2 AFR 和 AFF 模块的消融结果

校正模块	融合模块	参数量/M	mIoU/%	PA/%
–	相加	61.0	53.4	78.5
CM-FRM	相加	71.1	54.4	78.8
AFR	相加	69.9	54.9	78.8
–	FFM	67.6	54.2	78.7
–	AFF	64.1	54.6	78.9
CM-FRM	FFM	77.6	55.0	79.2
AFR	AFF	73.0	55.8	79.7

如表2结果所示,与基准模型相比,仅使用 AFR 模块时,mIoU 和 PA 分别提高了 1.5% 和 0.3%。仅使用 AFF 模块时,mIoU 和 PA 分别提高了 1.2% 和 0.4%。同时,引入 AFR 和 AFF 模块,其最终的 mIoU 和 PA 分别提高了 2.4% 和 1.2%。值得注意的是,对比仅使用 CM-FRM 与仅使用 FFM 的实验结果,单独使用 AFR 或 AFF 时,不仅参数量更低,而且性能表现均更优,mIoU 分别获得了 0.5% 和 0.4% 的提升。同时使用 AFR 和 AFF 模块时,相比同时使用 CM-FRM 和 FFM 模块,模型的 mIoU 获得了 0.8% 的提升。

此外,实验还发现,AFRF-Seg 的性能会受到式(2)和式(9)中自适应平均池化输出特征图的大小影响,具体结果如表3所示。以全阶段使用 7×7 的输出大小作为对比基准,动态地调整输出尺寸时,模型的 mIoU 和 PA 分别提高了 1.3% 和 1.0%。考虑到 NYUDepthv2 和 SUN-RGBD 数据集的输入尺寸均为 480×640,在自适应平均池化阶段,将输出尺寸的长宽比设置为 3∶4,可以使池化后的每个像素点对应的区域保持为正方形,从而保证语义信息不会被破坏。当自适应平均池化输出的长宽比设置为 3∶4且在不同阶段动态调整大小时,模型获得了 mIoU 为 55.8%、PA 为 79.7% 的最优性能。同时,由于平均池化输出尺寸的变化不会引入需要学习的参数,因此模型性能提高的同时,其参数量并不会发生变化。

表 3 AFR 和 AFF 模块上的实验结果 %

$k_1\times k_2$				评价指标	
阶段1	阶段2	阶段3	阶段4	mIoU	PA
7×7	7×7	7×7	7×7	54.2	78.2
56×56	28×28	14×14	7×7	55.5	79.2
24×32	12×16	6×8	3×4	55.8	79.7

2)解码器的消融实验。实验综合对比了使用多层感知机解码器(MLPDecoder, multilayer perceptron decoder)、Hamburger 解码器与所提解码器的分割效果,同时为了探索不同层级的特征对性能的影响,进行了针对解码器输入不同层级特征的实验。实验结果如表4所示。其中,"*"表示解码器输入特征的层级索引为"1,2,3",其余为"0,1,2,3"。

表 4 不同解码器的消融实验结果

解码器	参数量/M	mIoU/%	PA/%
MLPDecoder	62.0	54.7	78.8
Hamburger*	66.7	55.5	79.3
Hamburger	67.9	55.3	79.1
所提解码器*	72.5	55.5	79.4
所提解码器	73.0	55.8	79.7

与 MLPDecoder 相比,Hamburger 解码器和 AFRF-Seg 所采用的解码器中含有的矩阵分解模块,

能够将编码器学习到的特征分解为子矩阵，以恢复干净的低秩特征子空间，确保解码器处理的信息不会存在冗余或缺失的问题，故而在性能表现上均优于 MLPDecoder。在仅采用 1，2，3 级别特征时，由于特征多样性的缺乏，AFRF-Seg 解码器中引入的多尺度融合模块优势并不明显。在使用全部尺度的特征时，多尺度融合模块丰富了特征的多样性及其之间的交互，相比使用同样尺度的 Hamburger 解码器参数量仅提高了 5.1 M，但获得了 0.5% 的 mIoU 提升和 0.6% 的 PA 提升，同时获得了 mIoU 为 55.8%、PA 为 79.7% 的最优性能。

3）对抗性训练的消融实验。考虑到模型推理结果和真实标签的巨大差异，在训练过程中，将鉴别器引入到网络架构中可以获得性能提升。结果如表 5 所示，相比在训练过程未引入对抗性训练，AFRF-Seg 获得了 mIoU 和 PA 分别为 0.3% 和 0.2% 的性能提升。值得注意的是，在训练过程中引入预测标签和真实标签的鉴别器，并不会影响模型在推理时的参数量和计算量，在不增加计算开销的情况下，提升了模型语义分割的性能。

表 5　对抗性训练的消融实验

对抗性训练	参数量/M	mIoU/%	PA/%
×	73.0	55.5	79.5
√	73.0	55.8	79.7

2.3.3　定性分析

为了进一步验证 AFRF-Seg 提出的特征校正模块、融合模块和算法整体的功能与优势，将第 1 个特征校正模块前后的 RGB 特征图、深度特征图和融合之后的特征图进行了可视化。同时，进行了 AFRF-Seg、CMX（MiT-B5）和 DFormer-L 的语义分割结果之间的定性比较。

具体做法：将 1 个 $\boldsymbol{R}^{c\times h\times w}$ 的特征图视为 $h\times w$ 个维度为 c 的向量，利用主成分分析（PCA，principal component analysis）算法进行主成分分析后，取前 3 个重要的特征向量，并将 $h\times w$ 个向量对应的 3 个权重值作为可视化结果的 RGB 值。具体结果如图 4 所示，其中图 4（b）、图 4（c）、图 4（e）和图 4（f）分别代表 RGB 特征、深度特征在 AFR 模块前后的可视化结果，图 4（g）为校正后的特征图经过 AFF 模块融合后的可视化结果。从图 4 中可以看出，特征校正模块 AFR 能够很好地抑制 2 种模态的特征中的噪声，进而突出有用的语义特征，而特征融合模块 AFF 能够进一步增强 2 种模态间的交互，以 RGB 模态为主体，利用深度特征中的有效语义信息进行补充，得到包含噪声较少、语义信息丰富的特征图。

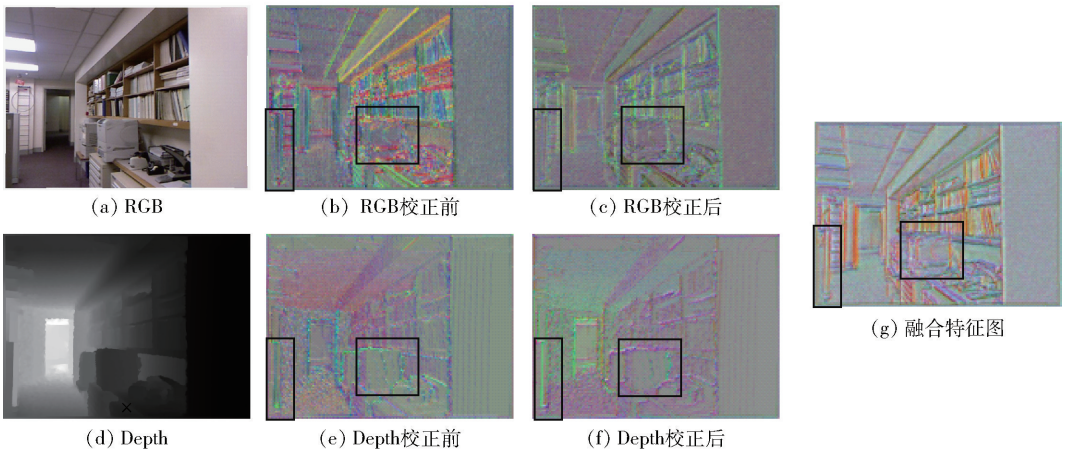

图 4　特征校正模块和特征融合模块在 NYUDepthv2 数据集上的效果可视化

AFRF-Seg、CMX（MiT-B5）和 DFormer-L 的定性比较结果如图 5 所示。可以看出，AFRF-Seg 不仅能够很好地识别出物体的语义类别，而且能够高效应对 RGB 图像中由于反射产生的干扰纹理，而这种反射对于 DFormer 这种基于单个主干网络的算法来说，更加难以处理。

3　结束语

对于 RGB-D 语义分割，考虑到 2 种模态包含语义信息的丰富性不同，笔者基于双流网络，引入了非对称特征校正模块和特征融合模块，有效抑制了 2 种模态中的噪声并增强了模态间的互补性。同

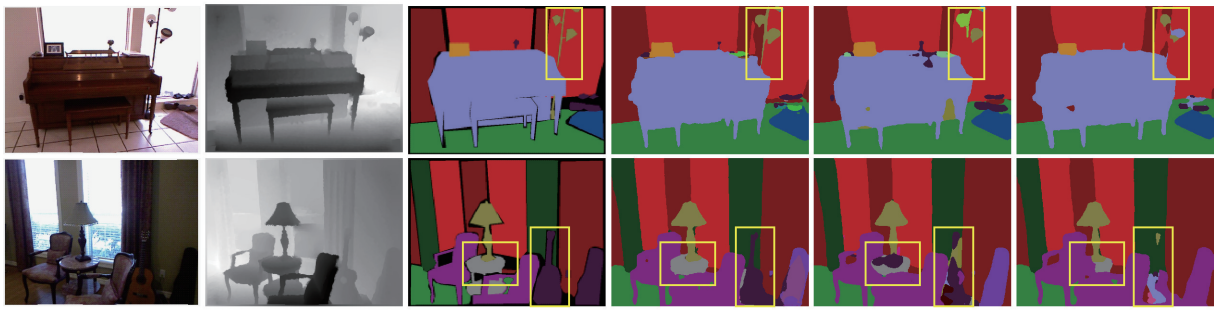

图5 AFRF-Seg(MiT-B3)、DFormer-L和CMX(MiT-B5)在NYUDepthv2数据集上的定性比较

时,采用了多尺度特征融合和对抗性训练的方法,充分利用上下文信息,提高了算法整体准确性。实验结果表明,AFRF-Seg在室内RGB-D数据集上均获得了具有竞争力的分割性能。

然而,AFRF-Seg基于双流网络,模型的计算量和参数量相比其他算法优势并不明显。如何将这种非对称特征交互的方法融入到单个网络,同时保持单个模态语义信息的独特性,是未来研究的一个可行方向。同时,提出的非对称特征交互方法仅针对主干网络的特征提取,因此可以尝试将该特征校正和融合方法应用于其他密集预测任务,如实例分割、全景分割、显著性目标检测等。

参考文献:

[1] CAO J, LENG H, LISCHINSKI D, et al. Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 7088-7097.

[2] CHEN L Z, LIN Z, WANG Z, et al. Spatial information guided convolution for real-time RGB-D semantic segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 2313-2324.

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

[4] 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. Cham: Springer International Publishing, 2020: 561-577.

[5] WANG Y, HUANG W, SUN F, et al. Deep multimodal fusion by channel exchanging[J]. Advances in Neural Information Processing Systems, 2020, 33: 4835-4845.

[6] 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.

[7] ZHANG J, LIU H, YANG K, 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.

[8] GAO S, YANG X, JIANG L, et al. Global feature-based multimodal semantic segmentation[J]. Pattern Recognition, 2024, 151: 110340.

[9] TAGHAVI P, LANGARI R, PANDEY G. SwinMTL: A shared architecture for simultaneous depth estimation and semantic segmentation from monocular camera images[J]. arXiv preprint arXiv: 2403. 10662, 2024.

[10] GENG Z, GUO M H, CHEN H, et al. Is attention better than matrix decomposition? [J]. arXiv preprint arXiv: 2109. 04553, 2021.

[11] WANG Y, CHEN X, CAO L, et al. Multimodal token fusion for vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 12186-12195.

[12] ZHANG J, LIU R, SHI H, et al. Delivering arbitrary-modal semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 1136-1147.

[13] YIN B, ZHANG X, LI Z, et al. DFormer: Rethinking RGB-D representation learning for semantic segmentation[J]. arXiv preprint arXiv: 2309. 09668, 2023.