一种基于 SwiftNet 面向室内 RGBD 场景的高效语义分割算法

王 博,许 钢,苏世林
安徽工程大学 电气工程学院,安徽 芜湖 241000

摘 要:目的 针对室内场景中的复杂光照、多样化的材质以及空间结构,现有的 RGBD 语义分割算法未能充分利用深度图像提供的形状信息,且计算成本高等问题,提出一种基于 SwiftNet 面向室内 RGBD 场景高效语义分割方法。**方法** 首先,在轻量级多尺度道路 RGB 场景语义分割算法(SwiftNet)中引入深度图像,通过利用深度图像的颜色稳定性和其为每个像素提供的到相机的距离信息,能够降低光线、颜色和距离等因素对分割结果的影响;然后,针对深度图像的几何形状特征进行专门提取,把深度特征分解为位置分量和形状分量,同时引入两个可学习权重以独立地与它们协作,再对这两个分量的重新加权组合应用卷积获取深度数据中固有的几何形状信息,不会在推理阶段引入计算和内存增加;最后,为了更快地捕捉更丰富的上下文信息,改进深度聚合金字塔池化模块使其并行提取上下文信息,称为快速聚合金字塔池化模块(FAPPM)。**结果** 通过在公共室内数据集 NYUv2 和 SUNRGBD 上的评估实验结果表明:相较于当前表现良好的 ESANet 模型,在两数据集上分别获得的 2.21%和 3.2%的 MIoU 提升,同时能够达到 33.36 的 FPS。**结论** 验证了该算法在处理复杂的室内环境语义分割中展现出的高效与准确性,为室内应用的后续智能机器人任务提供了良好的支持。

关键词:RGBD 语义分割;形状感知卷积;室内场景; 特征融合;深度学习
中图分类号:TP391 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0003.011

An Efficient Semantic Segmentation Algorithm for Indoor RGBD Scenes Based on SwiftNet

WANG Bo, XU Gang, SU Shilin
School of Electrical Engineering, Anhui Polytechnic University, Anhui Wuhu 241000, China

Abstract: Objective Existing RGBD semantic segmentation algorithms fail to fully utilize shape information provided by depth images and suffer from high computational costs, particularly for complex lighting, diverse materials, and spatial structures in indoor scenes. This paper proposed an efficient semantic segmentation method for indoor RGBD scenes based on SwiftNet. **Methods** Firstly, in the SwiftNet (a lightweight multi-scale road RGB scene semantic segmentation algorithm), depth images were incorporated. By leveraging the color stability of depth images and the distance information provided for each pixel relative to the camera, this approach reduced the impact of factors such as lighting, color variations, and distances on segmentation results. Next, a specialized extraction of geometric shape features from depth images was conducted. Depth features were decomposed into positional components and shape components, with two learnable weights introduced to independently collaborate with them. Convolution operations were then applied for the reweighting and combination of these two components, securing the intrinsic geometric shape information from the depth

收稿日期:2023-10-07 **修回日期:**2023-12-21 **文章编号:**1672-058X(2025)03-0084-10
基金项目:国家自然科学基金区域创新与发展联合基金项目(U22A2079).
作者简介:王博(1998—),男,安徽淮北人,硕士研究生,从事语义分割研究.
通信作者:许钢(1972—),男,安徽合肥人,教授,从事语音与图像信号处理、机器视觉研究. Email:xugang@ahpu.edu.cn.
引用格式:王博,许钢,苏世林. 一种基于 SwiftNet 面向室内 RGBD 场景的高效语义分割算法[J]. 重庆工商大学学报(自然科学版),2025,42(3):84-93.
WANG Bo, XU Gang, SU Shilin. An efficient semantic segmentation algorithm for indoor RGBD scenes based on SwiftNet[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(3): 84-93.

data without incurring additional computation and memory during the inference phase. Finally, to capture richer contextual information more rapidly, the depth aggregation pyramid pooling module was enhanced to extract context information in parallel, referred to as the Fast Aggregation Pyramid Pooling Module (FAPPM). **Results** Through evaluation experiments on the NYUv2 and SUNRGBD indoor datasets, the results demonstrated that compared with the current well-performing ESANet model, the proposed approach achieved improvements of 2.21% and 3.2% in mean intersection over union (MIoU) on these datasets, respectively. Furthermore, it achieves a processing speed of 33.36 frames per second (FPS). **Conclusion** The validation confirms the algorithm's efficiency and accuracy in handling complex indoor semantic segmentation tasks, providing solid support for subsequent intelligent robot tasks in indoor applications.

Keywords: RGBD semantic segmentation; shape-aware convolution; indoor scene; feature fusion; deep learning

1　引　言

室内场景是日常生活中不可或缺的一部分,随着深度学习技术的迅猛发展,对室内场景的深入理解变得至关重要,因为它为自动化和智能化系统的实现提供了关键支持。在室内场景中进行语义分割能够有效地助力各个领域的应用,包括智能家居[1]、机器人导航[2]、安防监控[3]等。在这一背景下,室内场景语义分割技术的发展显得愈发重要。

近年来,RGB 图像语义分割取得了显著的进展,但由于室内场景中光照变化、各种材质的存在空间结构复杂等因素,使得基于 RGB 图像的语义分割变得具有挑战性。同时为了满足各种实际应用场景的需求,也需强调实时性的重要性。为此,研究人员正在不断努力,从数据增强、多尺度融合、轻量化等多个角度来提高语义分割的精确性、鲁棒性和实时性。利用数据增强技术[4]扩充训练数据集以提高模型泛化能力,尤其是在训练数据有限的情况下。多尺度特征融合[5]利用高层网络的感受野大的低分辨率特征和低层网络语义几何细节信息表征能力强的高分辨率特征进行融合,可以在不同尺度上捕捉语义信息。例如,SwiftNet[6]提出了一种简化的多尺度特征融合网络,在多个分辨率上以一种新颖的方式融合共享特征,以扩大感受野提高语义分割的准确性和效率。虽然这些方法在某些情况下取得了一定的成功,但大多数都只关注 RGB 图像输入,如文献[7]中所述,RGB 信息提供了丰富颜色和纹理模型,但缺乏几何信息,因此很难区分共享相似颜色和纹理的实例和上下文。

多模态数据融合[8]已被证实为一种有效的策略。具体来说,深度图像能够提供物体的距离和形状等几何信息,而 RGB 图像则包含了颜色和纹理等视觉信息,通过将不同模态的数据融合到一个统一的模型中,能够显著提升语义分割的性能。此外,随着深度传感技术的不断发展,深度数据的获取变得更加容易和实时,从而进一步凸显了 RGBD 语义分割的潜力。

研究[8-9]表明,直接将互补的深度信息应用于现有的 RGB 框架中或简单地将两种模态的数据进行融合,通常无法获得理想的分割效果。近些年来,研究人员提出了一系列方法以提高 RGBD 语义分割的性能,其中之一为 RGBD 数据设计专用的网络架构[10-14],例如跨模态注意力融合网络[11]使用双分支编码器结构提取多尺度 RGB 和深度特征。ACNet[12]使用注意力辅助模块(ACM)从 RGB 和深度分支中提取加权特征,使网络能够专注于信息量更大的区域。CENNet[14]通过交换前向传播过程中动态的交换通道以充分利用所有通道信息。还有研究[15]通过将深度数据经过 HHA 编码转换得到水平视差、离地高度和与重力的角度多维度信息,然后在与 RGB 数据融合,从而提高了语义分割的性能。尽管上述研究取得显著的分割效果,大多使用相同卷积算子平等地提取 RGB 和深度特征,没有考虑 RGB 和深度图像本质上彼此不同。有研究设计特别定制的卷积[16-20],使用双流或多级的方式处理深度数据,加强对深度数据的特征提取。但是使用双流或多级方式会导致计算量的大幅增加,在室内场景 RGBD 语义分割实际应用中,不仅需要高准确性,还需要尽可能轻量化,以便在嵌入式系统和移动设备上实现实时性能。

为满足实时或移动需求,同时在推断速度和准确性之间达到最佳平衡,研究人员提出了许多高效和有效的语义分割模型。大致可分为两类:轻量级编码器和解码器、双分支网络架构。具体而言,SwiftNet[6]采用低分辨率输入获取高级语义,并使用高分辨率输入为其轻量级解码器提供足够的细节。BiSeNet[24]提出了一个双分支网络架构,包含两个深度不同的分支,用于上下文嵌入和细节解析,同时还使用特征融合模块来融合上下文和详细信息。受此启发,本文提出了一种基于 SwiftNet 面向室内 RGBD 场景的高效语义分割算法。众多现有 RGBD 分割方法[21-25]通常采用双分支结

构,分别用于处理RGB图像和深度图像,然后在完成所有特征提取操作后进行特征融合,这种“先提取,后融合”的策略使得每个分支专注于提取与其模态相关的特征。FuseNet[26]的研究结果表明,多阶段特征融合可以显著提高模型性能,这意味在不同层次上融合特征能够更好地捕获图像语义信息,从而提高了分割的准确性。本文在多尺度语义分割网络SwiftNet中引入深度数据,专门针对室内RGBD场景。深度数据的引入有助于保持特征尺度的一致性,在多个输出尺度相同的阶段进行多模态特征融合,使得融合操作不仅可以提前发生,而且还多于原先的多尺度融合次数,深度数据能够在所有层次上更好地与RGB数据相融合,从而捕捉图像的更完整语义信息[27]。通过实验证明该语义分割方法表现出色,能够很好地预测每个像素的类别。

此外,RGB和深度信息在本质上存在显著差异,RGB值捕获投影图像空间中的光度外观属性,而深度特征编码了局部几何体的形状信息以及其与周围环境的关系,形状信息通常更加固有且与语义关联更紧密,因此对于分割精度至关重要,不能简单地将深度图像视为和RGB等同的输入[28-29]。本文为了充分发挥深度数据的优势,引入了一个用于处理深度特征的形状感知卷积层,深度特征会首先被分解为形状分量和基础分量,然后引入两个可学习权重以独立地与它们协作,最后对这两个分量的重新加权组合应用卷积。基本卷积核和形状卷积核在推理阶段变为常数,在推理阶段不会引入任何计算和内存增加。

另外,一般的方法包括空洞空间金字塔池化(ASPP)[30]由具有不同采样率的并行空洞卷积层组成,用于关注多尺度的上下文信息。金字塔池化模块(PPM)[31]则在卷积层之前实施金字塔池化,相比ASPP更加轻量和具有更高的计算效率。另一方面,自注意力机制专注于捕捉全局依赖关系,这在某些情况下非常有用。双重注意网络(DANet)[32]结合了位置注意力和通道注意力,以进一步改善特征表示。对象上下文网络(OCNet)[33]则利用自注意力机制来探索对象上下文,即属于相同对象类别的一组像素之间的关系。另外,CCNet[34]引入了交叉注意力机制以提高内存和计算效率。然而,这些上下文提取模块通常设计和优化用于处理高分辨率特征图,这对于轻量级模型来说可能会导致计算开销过大。本文提出了一种名为快速聚合金字塔池化模块的上下文模块,模块输入低分辨率的特征图,通过增加更多尺度的池化操作和深度特征聚合来加强PPM模块,不仅能够提高分割性能,而且在推断速度方面几乎没有显著影响。以下为本文的主要工作:

(1) 在一个轻量级多尺度道路RGB场景语义分割算法SwiftNet中引入深度图像,通过利用深度图像的颜色稳定性和其为每个像素提供的到相机的距离信息,能够降低光线、颜色和距离等因素对分割结果的影响。

(2) 针对深度图像的几何形状特征进行专门提取,把深度特征分解为位置分量和形状分量,同时引入两个可学习权重以独立地与它们协作,再对这两个分量的重新加权组合应用卷积获取深度数据中固有的几何形状信息,不会在推理阶段引入计算和内存增加。

(3) 为了捕捉更丰富的上下文信息,改进上下文信息提取模块降低通道数,并行化连接方式,更快地捕捉丰富的上下文信息。

2 网络系统的设计与构建

本章描述整个网络流系统的总体框架,包括三个主要组成部分:网络结构、形状感知卷积和快速聚合金字塔池化模块。

2.1 网络结构

如图1所示,本文提出的网络架构采用双分支编码器-解码器结构。其中,图1下面淡蓝色部分为一个浅层的双分支编码器与上下文模块紧密结合,并通过跳跃链接将浅层融合特征送入解码器模块,有助于还原图像的局部细节。使用在ImageNet[35]预训练的ResNet34作为主干网络对RGB和深度图像特征进行提取。相较于完全缺失深度信息的SwiftNet,网络策略是在初始阶段就融合尺寸一致的多模态数据,减少有价值信息的丢失,同时还能充分挖掘输入图像在各个阶段的特征。深度提取分支中,使用整合形状感知卷积后的残差块进行深度特征提取,具体实现细节在第2.2节中描述。

图1上方淡蓝色部分表示解码器分支。为了更好地应对室内RGBD分割,这里使用了一种更为精细的解码方式,在首个解码器模块,开始使用512个通道,随后随着分辨率提升逐步减少每层的通道数。此外,如图1淡红色部分所示,每个解码模块都嵌入了三个参数量较小Non-Bottleneck-1D(NBt1D)[28],以确保在保持分割精度的同时占用更少的计算资源。接着通过一个3×3卷积调整通道数以适应每个像素的预测类别数。最后,特征图通过两次上采样操作扩大为原始尺寸的4倍,为提高结果质量,如图1右侧浅蓝色部分所示,首先应用最近邻上采样放大图像,接着通过3×3深度卷积整合邻近特征。初始阶段,上采样模仿双线性插值。然而,训练过程中网络权重的适应性使其能够

学习如何更有效地整合邻近特征,从而进一步增强分割效果。

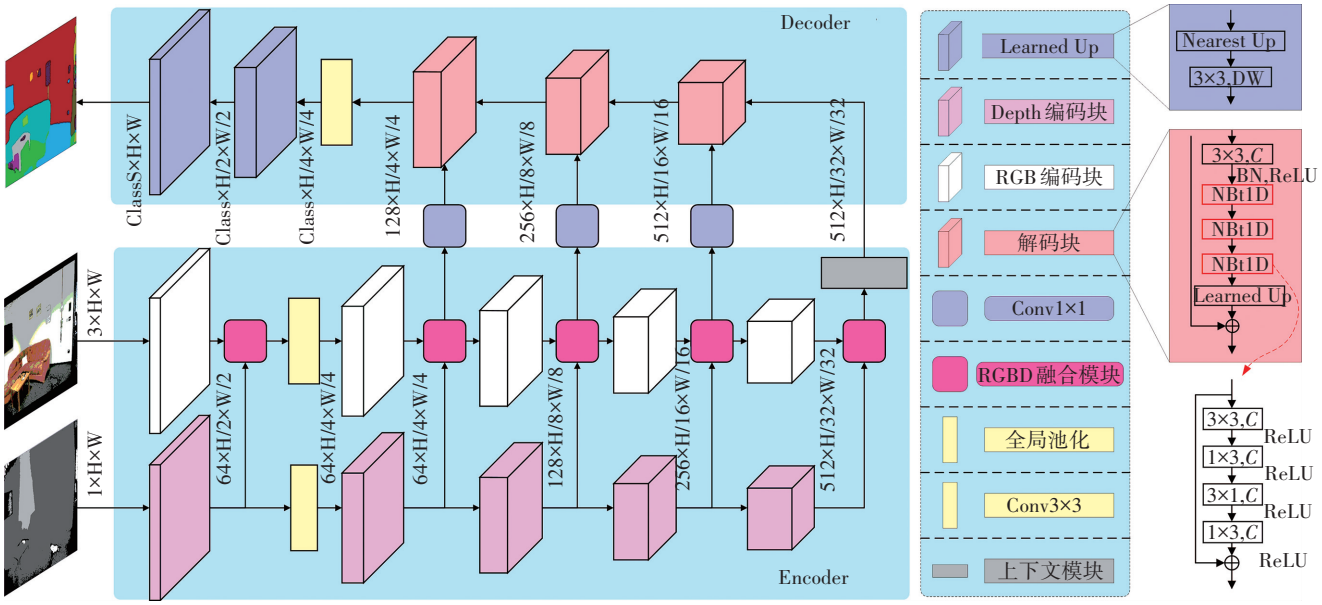

图 1　网络整体结构概述

Fig. 1　Overview of the overall network architecture

如图 2 所示,采用了注意力融合模块[36]分别作用于每个残差编码层之后,注意力模块通过引入一个 Squeeze 操作和一个 Excitation 操作来建模通道之间的关系进行特征融合。在 Squeeze 阶段通过全局平均池化操作将 RGB 和深度特征图分别压缩成一个特征向量,以获取每个通道的全局信息。然后,在 Excitation 阶段通过两个全连接层,将全局信息进行处理,产生一个用于缩放每个通道权重的激励值,得到的激励值应用到融合后的特征上调整每个通道的权重,网络能更加灵活地关注 RGB 和深度数据中的重要性不同的特征,增强网络对 RGB 和深度信息的综合利用。

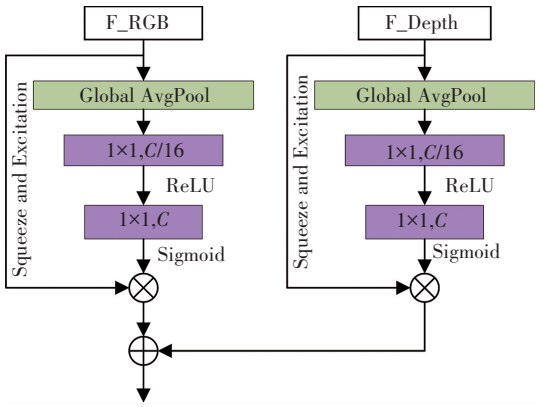

图 2　RGBD 注意力融合模块

Fig. 2　RGBD attention fusion module

2.2　形状感知卷积

图 3 为形状感知卷积的方法流程,方法的核心思想简略概括为将深度特征分解为分别描述位置和形状信息的分量,再通过可学习的权重修饰分量,最后对重新加权组合,以生成最终的特征表示。

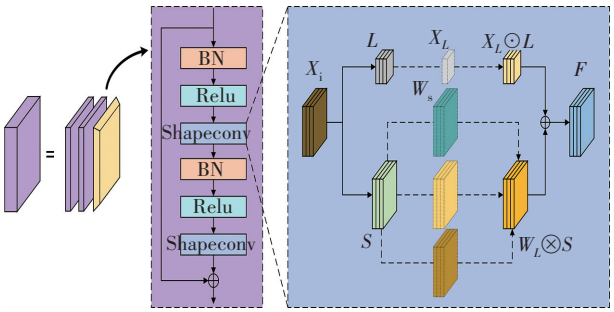

图 3　深度特征提取块

Fig. 3　Deep feature extraction block

在标准卷积操作中,输入 $X \in \mathbf{R}^{H_{\text{in}} \times W_{\text{in}} \times C_{\text{in}}}$,输出特征图 $F \in \mathbf{R}^{H_{\text{out}} \times W_{\text{out}} \times C_{\text{out}}}$ 由以下公式给出:

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

其中,$K \in \mathbf{R}^{K_h \times K_w \times C_{\text{in}} \times C_{\text{out}}}$ 是卷积核,K_h 和 K_w 是核空间维度,C_{in} 和 C_{out} 分别是输入输出特征图的通道谁,X_i 是输入的局部区域。为了解决视角和距离因素带来的问题,将输入的局部区域 X_i 分解成两个组成部分:一个描述位置的基础部分 $L \in \mathbf{R}^{1 \times 1 \times C_{\text{in}}}$ 和一个描述形状的部分 $S \in \mathbf{R}^{K_h \times K_w \times C_{\text{in}}}$。$L$ 和 S 由以下公式定义:

$$\begin{aligned} L &= \text{mean}(X_i) \\ S &= X_i - \text{mean}(X_i) \end{aligned} \tag{2}$$

其中,mean 表示求所有元素的平均值操作,L 包含了局

部区域的平均位置信息,相当于对局部区域进行平滑处理,能反映出局部区域的整体位置和空间分布,S 是与原局部特征与平均值的差值,包含了局部区域相对于平均位置的形状信息,通过差值操作,突出局部区域的形状差异。使用卷积核 K 和两个可学习的权重 W_L 和 W_S,同时定义两个操作符⊙和⊗用于加权 L 和 S:

$$L = W_L \odot L \Rightarrow L_{1,1,C_{\text{in}}} = W_L \times L_{1,1,C_{\text{in}}}$$
$$S = W_S \otimes S \Rightarrow S_{K_h,K_w,C_{\text{in}}} = \sum_i^{K_h \times K_w} W_{S_{i,K_h,K_w,C_{\text{in}}}} \times S_{i,C_{\text{in}}} \tag{3}$$

将 L 和 S 的加权结果结合到一起,形成一个新的局部区域 X_i':

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

普通卷积核 K 对新的输入 X_i' 进行卷积,以获得最终的输出 F,同时引入两个可学习的权重 W_L 和 W_S,用于加权 L 和 S。F 由以下公式给出:

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

这样,F 就包含了局部区域 X 的基础和形状信息,并且是通过学习的权重 W_L 和 W_S 进行加权的,在推理过程不涉及训练中的权重更新,附加的权重 W_L 和 W_S 会保持常数,在推理过程中不会更新这些权重,将其融合为 K_{LS} 与普通卷积中的 K 共享相同的张量大小,因此适用形状卷积处理深度数据时,不引入额外的推理时间。

2.3 快速聚合金字塔池化

由于编码器中采用残差网络的感受野受限,受到了 PSPNet[22] 的启发,通过在模型中加入了金字塔池化模块(PPM)在不同尺度上聚合特征来附加上下文信息,这一模块的作用是在不同尺度上聚合特征,以增加网络的感受野并促进局部与全局特征的交互。深度聚合金字塔池化模块(DAPPM)从低分辨率特征图中提取上下文信息。然而,DAPPM 的深度设计在计算上是串行的,而且 DAPPM 的每个尺度的通道数相对较多导致较高的时间开销,超出了轻量模型的处理范围。为此,如图 4 展示了快速聚合金字塔池化的内部结构,通过在不同尺度上应用池化操作捕捉更广泛范围内的信息,增加网络的感受野。如图 4 所示,对 DAPPM 进行改进使其支持并操作提高计算速度,称之为快速聚合金字塔池化模块(FAPPM),具体地,金字塔塔池化提取到不同层级的特征信息后,不需要让原特征信息逐一地与 4 个空洞卷积后的特征进行融合,而是同时并行地与 4 个分支进行融合,这种并行计算方式可以更好地利用了硬件并行计算的能力,特别是在大规模图像处理和计算需求较高的任务中,可以显著提高推理速度。

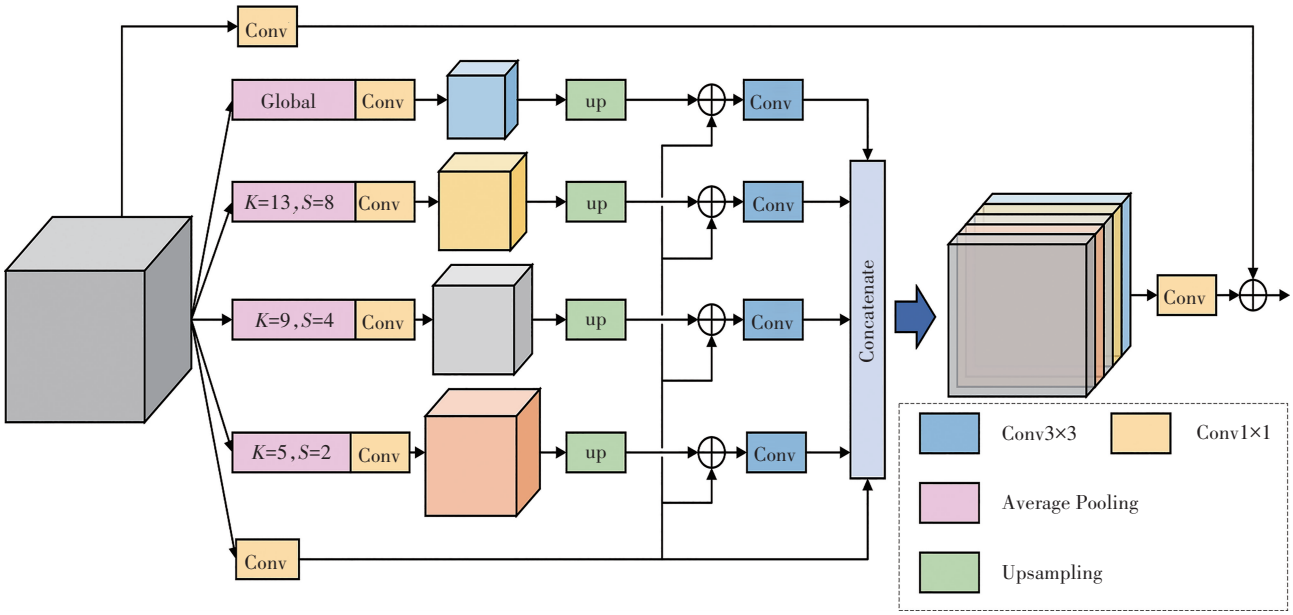

图 4 快速聚合金字塔池模块

Fig. 4 Fast aggregated pyramid pooling module

3 仿真实验与结果分析

为了验证本文方法的有效性,在 NYUv2 和 SUNRGBD 两个主流的且具有代表性的室内数据集上进行试验,这两个数据集都将 RGB 和深度作为输入,使用 Microsoft

Kinect v1、Intel RealSense RGB 和深度摄像机捕获 RGB 和深度图像。具体地,NYUv2 包含 1 449 个 RGBD 场景图像,遵循标准设置,划分为 795 张训练图像和 654 张测试图像,并且每个像素被分类为 40 个不带背景类的类别。SUN RGBD 数据集由 10 355 张 RGBD 图像构成,标注了 37 个类别,数据集划分为 5 285 张的训练部分和 5 050 张的测试部分。同时为了综合评估网络性能,使用每秒 10 亿次的浮点运算数(GFLOPs)、模型参数量(Params)、平均交并比(MIoU)和每秒帧数(FPS)作为实验的评估指标。

3.1　参数设置

使用在 ImageNet 上预训练的 ResNet34 作为主干网络,借助 PyTorch 深度学习框架实现本文的网络结构。在模型训练方面,模型的训练过程使用了 Nvida GTX 1080Ti 12G GPU,并利用 FLUKE TI55FT 拍摄的室内场景红外图像进行额外的测试验证。学习策略采取了“Poly”方法,并设置初始学习率为 0.01。每个批次处理 16 张图片,每张图片尺寸为 640×480。模型优化器选用了 SGD,其中动量参数设定为 0.9,权重衰减为 1×10^{-4}。在训练过程中,对批归一化参数进行微调,并采用了图像的随机缩放、裁剪和翻转进行数据增强以防止模型过拟合。此外,对 RGB 图像在 HSV 颜色空间中引入了轻微的色彩扰动。损失使用二维交叉熵损失函数计算。针对 NYUv2 数据集,模型进行了 500 个 epoch 的训练,而在更大的 SUN RGBD 数据集上,则进行了 300 个 epoch 的训练。而消融实验则选择在较小的 NYUv2 数据集上进行。

3.2　实验结果

在室内场景数据集上的实验结果清晰地展示了本文提出的方法在语义分割领域的显著性能。从表 1 的数据可以明显看出,与先进用 ResNet50 主干的方法相比,本方法在 NYUv2 数据集上的 MIoU 提高了显著的性能,达到了 53.53%。在 SUN RGBD 数据集上,本方法的 MIoU 也达到了 53.33%。在使用 ResNet34 主干时,本文方法在 NYUv2 和 SUN RGBD 数据集上的 MIoU 比 ESANet 提高了 2.21% 和 3.20%。这一提高不仅证明了本文算法的优越性,特别是在目标数据集很小的情况下,更突出了在深度数据特征挖掘上的独到之处。这也意味着模型更能够捕捉到细微的场景差异,并对其进行准确的语义标注。在处理速度方面,尽管本方法的 FPS 与 RedNet 相差不大,但在性能上确实超过了其 0.36%。进一步证明了模型不仅在准确性上表现出色,而且在实际应用中也具有高效的性能。

同时表 1 也展示了 GFLOPs 和 Params,分别从模型计算量和参数量角度分析。在使用 ResNet34 为主干网络时,模型相对于主流轻量化语义分割 RedNet、ESANeT 模型在浮点运算数减少 7.78 和 2.68 GFLOPs,在参数量上分别减少 1.66 M 和 0.53 M。在 ResNet50 为主干网络时,模型相对于 ESANeT 模型在浮点运算数减少 6.42 GFLOPs,在参数量上减少 0.95 M。这表明本研究方法在保持高准确率的同时,具有更低的计算量和参数量。

表 1　NYUDv2 和 SUNRGBD 数据集上对比实验结果

Table 1　Comparative experimental results on the NYUDv2 and SUNRGBD datasets

Model	Backbone	Image	MIoU		GFLOPs	Params/M	FPS
			NYUv2	SUNRGBD			
RedNet[37]	ResNet34	RGBD	—	47.80	104.66	50.49	33.00
SSMA[38]	ResNet50	RGBD	—	44.43	189.81	65.82	12.42
ACNet[12]	ResNet50	RGBD	48.38	48.11	126.47	116.61	16.54
SA-Gate[39]	ResNet50	RGBD	50.43	49.44	204.97	63.46	11.91
RDFNet[13]	ResNet50	RGBD	50.12	47.73	257.45	70.96	5.83
ESANet[2]	ResNet50	RGBD	50.68	48.31	115.50	57.39	22.88
ESANet	ResNet34	RGBD	50.30	48.17	99.56	49.36	33.90
Ours	ResNet50	RGBD	53.53	53.33	109.08	56.44	24.52
Ours	ResNet34	RGBD	52.51	51.37	96.88	48.83	33.36

在 NYUv2 数据上与基线模型进行对比,对比了本文多模态融合与轻量级多尺度网络 SwiftNet 计算开销,同时对比了采用 RGB 和 RGBD 图像的效果。如表 2 所示,与基线模型相比本文的网络比轻量级多尺度网络

在 ResNet50 和 ResNet34 主干 SwiftNet 少 1.39 和 1.17 的 GFLOPs,在引入深度模态时,ResNet50 和 ResNet34 的为主干的网络 MIoU 分别比 RGB 单模态高 12.18% 和 14.94%。结果表明:采用深度图像可以提高分割的准确性,从而更好地支持物体分割任务。

表 2 与基线模型对比

Table 2 Comparison with baseline models

Model	Backbone	Image	GFLOPs	Params/M	MIoU
SwiftNet	ResNet50	RGB	110.47	58.16	42.95
	ResNet34	RGB	98.05	44.32	38.66
Ours	ResNet50	RGB	110.12	56.63	41.35
		RGBD	109.08	56.44	53.53
Ours	ResNet34	RGB	97.23	49.41	37.57
		RGBD	96.88	48.83	52.51

虽然使用 ResNet50 作为骨干网络的方法在两个数据集上都比使用 ResNet34 获得了更高的分割精度,在 NYUv2 和 SUNRGBD 数据集上分别高出 1.02% 和 2.04%,但其推理速度确实受到了影响,特别是在数据集规模较小的情况下。这暗示着对于某些实际应用场景,可能更倾向于选择 ResNet34 作为骨干网络,因为它在维持相对高精度的同时,可以提供更快的推理速度,从而达到更好的综合效果。

3.3 消融实验

针对所设计的网络架构进行了消融实验研究,并在用平均交并比(MIoU)和帧数(FPS)作为评估指标以证明本文方法的可行性。表 3 展示了网络各主要部分对语义分割性能的影响。具体地,当引入形状卷积进行深度特征提取时,分割平均交并比精度 MIoU 实现了约 2%的提升,而 FPS 仅仅增加了 0.02%,基本保持不变,这一结果验证了形状卷积在深度特征提取方面的有效性,其能有效捕获并利用形状信息以增强语义分割的准确性,而无需显著增加计算负担。进一步对比金字塔池化模块、深度聚合金字塔池化模块和快速聚合金字塔池化模块,后者比金字塔池化模块 MIoU 提升 2.71%,表明更长的上下文信息能够提升分割精度。快速聚合金字塔池化比深度聚合金字塔池化在 MIoU 指标上保持一致,在 FPS 上达到了 0.6%提升,表明并行融合方式提高了推理速度。值得注意的是,在不使用跳跃连接(skip)和自注意力融合(se)的网络配置中,分割结果的 MIoU 分别下降了 3.2%和 2.95%,与此同时,FPS 仅分别实现了 0.34%和 0.5%的增加。这两个实验点明确揭示了这两个组件在提升分割精度方面的关键作用,同时也指出在一定程度上,牺牲少量的运算效率能够换取分割性能的显著提升。

表 3 对网络中不同组成部分的有效性

Table 3 Effectiveness of different components in the network

Skip	Fusion		Context			Depth convolution		MIoU	FPS
	add	se	ppm	dappm	fappm	vanilla	shape		
		✓			✓		✓	49.31	33.70
✓	✓				✓		✓	49.56	33.86
✓		✓			✓	✓		50.51	33.34
✓		✓	✓				✓	49.80	33.29
✓		✓		✓			✓	52.44	33.96
✓		✓			✓		✓	**52.51**	**33.36**

如表 4 所示,展示了解码块中 NBt1D 数量变化对实验结果的影响。分别展现了逐渐从 1 到 4 增加了 NBt1D 块的个数下 MIoU 和 FPS 的变化情况,实验结果表明在 NBt1D×1 的情况下,尽管 FPS 略有提升,但 MIoU 实际上略低于使用基本块的情况。而当 NBt1D 卷积块的数量提升至 3 时,MIoU 达到了一个显著的提升,达到 52.51%。然而,进一步增加到 4 块后,虽然 MIoU 略有提升,达到 52.58%,但 FPS 显著下降至 32.78。结果表明:当 NBt1D 块的数量为 3 时网络性能最佳。

表 4 解码块中卷积块对实验结果的影响

Table 4 Impact of convolutional blocks in the decoding block on experimental results

Decoder module	MIoU	FPS
Basic block	50.65	34.03
NBt1D×1	50.07	34.10
NBt1D×2	51.25	33.48
NBt1D×3	**52.51**	**33.36**
NBt1D×4	52.58	32.78

表 5 则展示了不同上采样方法对实验结果的影

响。通过分析最邻近插值、双线性插值以及可学习上采样法在实验中的表现,揭示不同上采样策略在本文语义分割任务中的适用性和效率。在最邻近差值和双线性插值上采样策略中,虽然 FPS 表现出略高的特性(分别为 34.04 和 33.92),这两种策略在 MIoU 指标上的表现却不如可学习的上采样方法。特别是可学习的上采样方法在 MIoU 上以 52.51%的表现,超越了双线性插值的 51.24%,高出 1.27%,而在 FPS 上的损失相对较小。综上所述,在本文语义分割任务中可学习的上采样方法为更优的选择。

表 5　上采样方法对实验结果的影响

Table 5　Influence of upsampling methods on experimental results

Upsampling	MIoU	FPS
Nearest-neighbor	50.88	34.04
Bilinear	51.24	33.92
Leanred upampling	**52.51**	**33.36**

3.4　可视化

图 5 展示了在 NYUv2 数据集上的分割效果,图像从左到右依次为 RGB、HHA、Depth、GT 和 Pred,下方给出每种类别所对应的颜色示例。RGB 代表原始的彩色图像;HHA 和 Depth 表示深度信息的不同可视化方法;GT 是真实的分割标签,Pred 是模型的预测输出。从可视化的结果中,可以明显分辨到椅子、桌子、家具等目标,展现方法在整合深度信息方面的能力,其能够精准地识别并分割仅凭 RGB 图像难以辨别的目标。尤其在红框标出的例子中,可以看到即使在 RGB 图像中沙发的边缘不太明显,模型仍然能够准确地分割出沙发的轮廓,这得益于深度信息提供的额外几何数据。这些结果表明:当融合 RGB 和深度信息时,可以获得更好的物体分割效果,特别是在某些物体在 RGB 图像中难以识别的情况下。深度信息为模型提供了额外的几何和结构信息,有助于模型更好地理解场景,并提高分割的准确性。

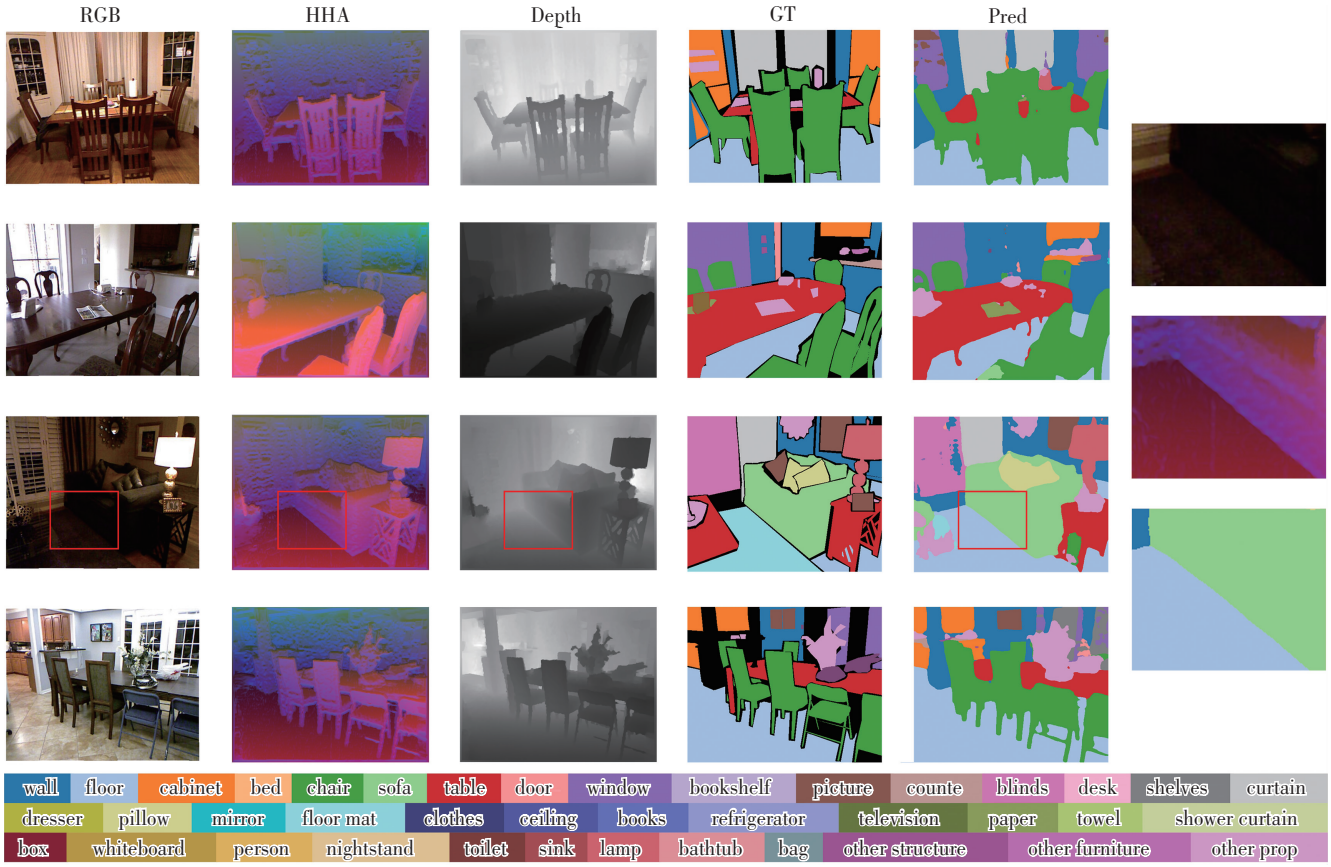

图 5　NYUv2 数据集的可视化结果

Fig. 5　Visualization of the NYUv2 dataset

图 6 展示了本研究方法与 ESANet 在室内图像语义分割上的特征图的可视化效果,针对形状结构复杂的椅子相对应的通道特征。通过观察不难看出椅子的形状轮廓在本文的方法中更加清晰和准确,而 ESANet 对椅子的中间结构形状不够清晰。可以明显看出本研究方法所呈现的特征图在细节和边缘形状捕捉方面表现得更为出色。

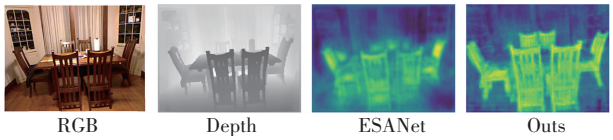

图 6 椅子通道的特征图

Fig. 6 Feature map of chair channel

4 结论与展望

4.1 结 论

在本项研究中,借助 SwiftNet 网络框架,提出了一种针对室内 RGBD 场景的高效语义分割策略。在继承了 SwiftNet 对 RGB 图像实时语义分割的基础之上,将深度图像成功整合至编码器中。为更充分地利用深度图像中的物体形状信息,研究引入了形状感知卷积技术。同时提出快速聚合金字塔池化模块并行处理上下文信息,快速从低分辨率的特征图中抽取出多尺度的上下文信息。经过一系列的对比实验与消融实验,实验证明本方法在室内 RGBD 场景的语义分割中在 MIoU 和 FPS 上表现出色,为室内 RGBD 场景的后续研究和应用提供了有力支持。

4.2 展 望

在未来的工作中,将进一步研究在保持计算效率的同时,如何融合更多的场景理解技术,例如对象检测和实例分割,期望在更广泛的应用场景中实现更准确和更快速地语义理解。同时,还将探索融合多模态传感器数据,包括红外线、热成像等,捕捉更多元的场景信息,提升系统的识别和分析能力,拓宽应用范围以满足更加复杂和多变的实际需求。

参考文献(References):

[1] KIM W, SEOK J. Indoor semantic segmentation for robot navigating on mobile[C]//2018 Tenth International Conference on Ubiquitous and Future Networks. Prague, Czech: IEEE, 2018: 22-25.

[2] SEICHTER D, KÖHLER M, LEWANDOWSKI B, et al. Efficientrgb-d semantic segmentation for indoor scene analysis[C]// International Conference on Robotics and Automation. Xi' an, China: IEEE, 2021: 13525-13531.

[3] ABDULLAH F, JALAL A. Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system[J]. Arabian Journal for Science and Engineering, 2023, 48(2): 2173-2190.

[4] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. Journal of Big Data, 2019, 6(1): 1-48.

[5] ZHANG H, ZHANG H, WANG C, et al. Co-occurrent features in semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 548-557.

[6] ORSIC M, KRESO I, BEVANDIC P, et al. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 12607-12616.

[7] LIN D, CHEN G, COHEN-OR D, et al. Cascaded feature network for semantic segmentation of RGB-D images[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 1311-1319.

[8] NOORI A Y. A survey of RGB-D image semantic segmentation by deep learning[C]//2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). Coimbatore, India: IEEE, 2021, 1: 1953-1957.

[9] FANG T, WEI Z, ZHANG L, et al. Depth dependence removal in RGB-D semantic segmentation model[C]//2023 IEEE International Conference on Mechatronics and Automation (ICMA). Harbin, China: IEEE, 2023: 699-704.

[10] CHENG Y, CAI R, LI Z, et al. Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 3029-3037.

[11] FANG A, ZHAO X, ZHANG Y. Cross-modal image fusion guided by subjective visual attention[J]. Neurocomputing, 2020, 41(4): 333-345.

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

[13] PARK S J, HONG K S, LEE S. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 4980-4989.

[14] WANG Y, HUANG W, SUN F, et al. Deep multimodal fusion by channel exchanging[J]. Advances in Neural Information Processing Systems, 2020, 33(4): 4835-4845.

[15] XING Y, WANG J, CHEN X, et al. Coupling two-Stream RGB-D semantic segmentation network by idempotent mappings[C]//2019 IEEE International Conference on Image Processing (ICIP). Taipei, Taiwan, China: IEEE, 2019: 1850-1854.

[16] XING Y, WANG J, CHEN X, et al. 2.5 D convolution for RGB-D semantic segmentation[C]//2019 IEEE International Conference on Image Processing (ICIP). Taipei, Taiwan, China: IEEE, 2019: 1410-1414.

[17] XING Y, WANG J, ZENG G. Malleable 2.5 d convolution:

Learning receptive fields along the depth-axis for RGB-D scene parsing[C]//European Conference on Computer Vision. Virtual: Springer, 2020: 555-571.

[18] WANG W, NEUMANN U. Depth-aware CNN for RGB-d segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: IEEE, 2018: 135-150.

[19] CHEN L Z, LIN Z, WANG Z, et al. Spatial information guided convolution for real-time RGBD semantic segmentation[J]. IEEE Transactions on Image Processing, 2021, 30(4): 2313-2324.

[20] CHEN Y, MENSINK T, GAVVES E. 3D neighborhood convolution: Learning depth-aware features for RGB-D and RGB semantic segmentation[C]//2019 International Conference on 3D Vision. Québec, Canada: IEEE, 2019: 173-182.

[21] WANG L, LI D, ZHU Y, et al. Dual super-resolution learning for semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 2020: 3774-3783.

[22] JIAO J, WEI Y, JIE Z, et al. Geometry-aware distillation for indoor semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE: 2019: 2869-2878.

[23] KHAN S H, BENNAMOUN M, SOHEL F, et al. Integrating geometrical context for semantic labeling of indoor scenes using RGBD images[J]. International Journal of Computer Vision, 2016, 117(1): 1-20.

[24] FAN M, LAI S, HUANG J, et al. Rethinkingbisenet for real-time semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 2021: 9716-9725.

[25] WU Y, SHI Y, SHEN H, et al. Light-TBFnet: Rgb-d salient detection based on a lightweight two-branch fusion strategy[J]. Multimedia Tools and Applications, 2023, 82(17): 26005-26035.

[26] HAZIRBAS C, MA L, DOMOKOS C, et al. Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture[C]//Computer Vision-ACCV 2016: 13th Asian Conference on Computer Vision. Taipei: IEEE, Taiwan: 2017: 213-228.

[27] JIN W D, XU J, HAN Q, et al. CDNet: Complementary depth network for RGB-D salient object detection[J]. IEEE Transactions on Image Processing, 2021, 30(8): 3376-3390.

[28] ROMERA E, ALVAREZ J M, BERGASA L M, et al. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 19(1): 263-272.

[29] ZHANG X, DU B, WU Z, et al. LAANet: Lightweight attention-guided asymmetric network for real-time semantic segmentation[J]. Neural Computing and Applications, 2022, 34(5): 3573-3587.

[30] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[31] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, US: IEEE, 2017: 2881-2890.

[32] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 3146-3154.

[33] YUAN Y, CHEN X, WANG J. Object-contextual representations for semantic segmentation [C]//Computer Vision-ECCV 2020: 16th European Conference. Glasgow, UK: Springer-Verlag, 2020: 173-190.

[34] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019: 603-612.

[35] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 248-255.

[36] HU J, SHEN L, SUN G. Squeeze and excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.

[37] HUA Z, QI L, DU D, et al. Dual attention based multi-scale feature fusion network for indoor RGBD semantic segmentation [C]//2022 26th International Conference on Pattern Recognition (ICPR). Montreal, QC, Canada: IEEE, 2022: 3639-3644.

[38] VALADA A, MOHAN R, BURGARD W. Self-supervised model adaptation for multimodal semantic segmentation[J]. International Journal of Computer Vision, 2020, 128(5): 1239-1285.

[39] 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, Glasgow, UK: Springer, 2020: 561-57.

责任编辑:陈　芳