Computer Science and Application 计算机科学与应用, 2024, 14(4), 13-23
Published Online April 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.144072

Hans汉斯

面向交通场景的图像分割网络

高程阳，郁　湧，秦江龙

云南大学软件学院，云南　昆明

收稿日期：2024年3月3日；录用日期：2024年4月2日；发布日期：2024年4月9日

摘　要

语义分割技术是自动驾驶的重要基础技术之一。在交通场景的图像语义分割中，图像语义分割希望对图像的每一个像素进行分类，并进行颜色标注，使车辆能够准确地检测道路上的交通参与者和可以行驶的路面区域。目前，典型的图像语义分割算法通常融合骨干网络生成的不同阶段的特征图，以提高分割性能。简单的融合方法不能充分利用这些特征信息。这将导致相似物体之间的分割错误和小物体边界分割粗糙，使车辆无法准确感知周围环境，从而影响上层的决策，甚至对其他交通参与者造成严重的安全隐患。针对这一问题，本文设计了一种密集特征融合与边界细化网络(DFBNet)，它包括两部分：特征融合网络(FFN)利用多个不同感受野大小的分支和卷积核提取特征图中的信息，并利用注意力机制为提取的特征分配权重；边界细化网络(BRN)利用空间注意力对每个像素位置赋予给予权值，使得目标对象的边界区域分割得更精细。我们在两个数据集上进行实验：Cityscapes和Camvid数据集。我们取得了良好的分割结果，在Cityscapes验证集上的平均交并比(mIoU)为79.47%。在Camvid测试集上的mIoU为75.13%。

关键词

图像语义分割，图像分类，卷积神经网络

Image Segmentation Network for Traffic Scenes

Chengyang Gao, Yong Yu, Jianglong Qin

School of Software, Yunnan University, Kunming Yunnan

Received: Mar. 3rd, 2024; accepted: Apr. 2nd, 2024; published: Apr. 9th, 2024

Abstract

This Semantic segmentation technology is one of the important basic technologies of automatic driving. In automatic driving, image semantic segmentation hopes to classify every pixel of the

文章引用: 高程阳, 郁湧, 秦江龙. 面向交通场景的图像分割网络[J]. 计算机科学与应用, 2024, 14(4): 13-23.
DOI: 10.12677/csa.2024.144072

image and make color labeling, so that the vehicle can accurately detect the traffic participants on the road and the pavement area that can be used. At present, typical image semantic segmentation algorithms usually fuse the feature maps of different stages generated by the backbone network to improve segmentation performance. The simple fusion method cannot fully utilize this feature information. This will result in segmentation errors between similar objects and rough boundary segmentation of small objects, making the vehicle unable to accurately perceive the surrounding environment, thus affecting the decision-making of the upper level, and even causing serious safety hazards to other traffic participants. To solve this problem, this paper designs a dense feature fusion and boundary refinement network (DFBNet), which includes two parts: Feature Fusion Network (FFN) using multiple branches and convolutional kernels of different receptive field sizes to extract the information in the feature map and using the attention mechanism to assign weights to the extracted features; Boundary Refinement Network (BRN) using spatial attention to give weights to each pixel position, making the boundary area of the target object more finely segmented. We experimented on two datasets: Cityscapes and Camvid. We achieved good segmentation results with an average intersection (mIoU) of 79.47% on the Cityscapes validation set and 75.13% on the Camvid test set.

Keywords

Image Semantic Segmentation, Image Classification, Convolutional Neural Networks

1. 引言

语义分割是许多计算机视觉应用领域(例如场景理解，人类解析和自动驾驶等)的基本技术，语义分割是指根据图像的不同特征，把图像划分为互不相交的部分，使得同一区域内的特征尽可能表现出一致性。因此语义分割是一个非常大的挑战。目前流行的分割方法通常是将主干网络所产生的不同阶段的特征图进行融合来丰富特征信息，进而生成分割结果。然而，这些方法忽略了不同阶段特征图所包含的语义信息和分辨率的差异，只是把它们进行一些简单的融合操作，这种方法使得模型对目标边界分割模糊，相似物体之间分割精度较低。在交通场景的图像分割中会使得车辆无法准确识别周围的环境信息和交通信号。这很有可能给自动驾驶的上层决策系统提供错误的信息，从而影响交通安全。

如图 1 为全景特征金字塔网络(FPN)的结构示意图。它只是逐步上采样不同阶段的特征图，然后再进行相加操作。这种简单的融合方式，使得目标边界分割模糊或类别间分割错误。模型将地形误分类为植被，人行道边缘分割存在明显错误。在交通场景的图像分割中这个问题是十分有必要解决的。

综上所述，我们的主要贡献如下：

提出了一种新的特征融合网络(FFN)。采用密集连接的方法，最大限度地弥补特征图下采样过程中丢失的信息，并在特征图的拼接中对每个通道的特征赋权，使模型能够识别出需要学习的重要通道。与其他一些简单的网络特征融合方法相比，我们的方法可以获得更高质量的特征图。它能更好地整合特征图中的语义信息和空间信息。在交通场景图像语义分割中，可以更准确地识别出目标特征，并能准确地识别和分割出相似的目标。为上层自动驾驶系统提供更准确的周边环境信息。

提出了一种边界优化网络(BRN)。它使用注意力机制为每个像素位置分配权重。在交通场景中，它帮助车辆感知小物体，如交通灯、可行驶的道路边界和物体的边界。在 FFN 网络获得高质量特征图的基

础上，使得分割结果更加准确，它为自动驾驶汽车的安全驾驶提供了重要的支撑。在上述两种网络的基础上，我们提出了一种新的分割网络 DFBNet，该网络采用密集连接的方法，可以弥补下采样过程中空间信息的损失，从而使特征更好地融合，目标边界处的分割也更加精细。

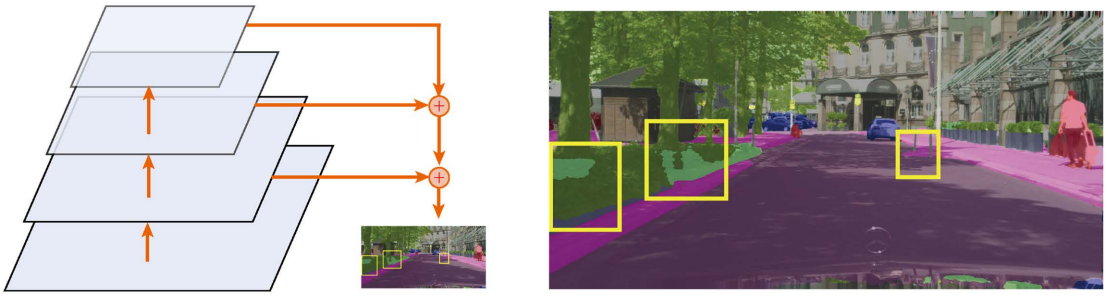

Figure 1. FPN structure diagram and segmentation result
图 1. FPN 网络结构图和分割结果

2. 相关工作

图像语义分割可以看作是图像分类从图像级到像素级的延伸，目前的研究方法大致可分为两类，一类是基于卷积神经网络为代表的分割算法，另一类是基于 Transformer 为代表的分割算法。

FCN (Fully Convolutional Networks) [1]是第一个将卷积神经网络引入图像分割的算法。FCN 采用端到端的训练方法，整个网络由卷积层组成，在输出阶段对特征图进行上采样操作来恢复尺寸。此外，FCN 还将浅层特征与深度特征深度融合，逐像素进行分类，实现了图像语义分割，在一定程度上开创了图像语义分割的新时代。此后，研究人员对不同方面进行了改进和研究。Ronneberger 等人提出了 U-Net [2]，该网络在医学图像分割中表现优异。U-Net 采用编码器–解码器结构，编码网络从图像中提取特征，5 个池化层获取不同尺度的特征图，并采用跳跃连接方法实现网络对图像多尺度特征的识别。但是，U-Net 的缺点也很明显。模型的效率非常低。网络会计算图像的每个邻域，所以特征图尺寸越大，消耗的计算资源就越多。卷积网络的卷积核在提取特征时只计算一个像素和它周围某一区域的像素，这使得卷积网络能够很好地提取局部信息，具有很好的泛化能力但同时，这也导致了卷积神经网络对全局信息的把握能力较弱。

随着研究者们不断地探索，在语义分割方面，Zheng 等人提出的 SETR [3]来证明了在图像语义分割任务中使用 Transformer 的可行性。他们对分割模型的设计进行重新思考，并提出一种新的方法用纯 Transfomer 来代替基于叠加层的逐步降低分辨率的编码器，从而产生了一种新的分割模型，称为分割转换器(SETR)。该编码器将输入图像看作序列，并用全局自注意机制对序列进行变换，以实现识别性的有限元表示学习。具体来说，该方法首先将图像分解成一个由固定大小的贴片组成的网格，形成一个切片序列。将线性嵌入层应用于每个切片的平坦像素向量，然后获得特征嵌入向量序列作为 Transformer 学习的特征，然后使用解码器恢复原始图像分辨率。关键的是，在空间分辨率上没有下采样，而是在编码转换器的每一层进行全局上下文建模，从而为语义分割问题提供了一个全新的视角。

无论是以卷积网络还是以 Transformer 为基础的语义分割方法都采用了编码器解码器的设计结构，编码器是预训练的主干模型，用于生成不同阶段的特征图，解码器将这些特征图进行相加、相乘或者拼接等简单的方式进行逐步融合，恢复分辨率，并负责生成分割结果。但是它们都忽略了特征图所包含的语义信息的差异和大小的差异。这些方法使用的特征融合方法与我们的方法相比过于简单。它没有为解码器提供高质量的特征图，容易造成分割精度低的问题。

3. 方法

在本节中，我们将详细介绍所提出的交通场景图像分割网络 DFBNet。DFBNet 遵循编码器解码器的结构设计，编码器主要是生成特征的骨干网络。解码器由特征融合网络与边界细化网络组成：

(1) 编码器中的骨干网络对输入图像进行卷积操作提取特征，并且对图像进行下采样操作降低特征图的尺寸。(2) 解码器中的特征融合网络(FFN)用来融合不同阶段的特征图，将高级特征图与低级特征图中所包含的特征信息充分融合。(3) 边界细化网络(BRN)对融合好的特征图进行混合注意力机制，给每个像素位置分配权重，使模型自动学习到重点关注区域的权重，通过调整权重，增强特征区域的表达能力，使图中小物体以及物体边界区域分割更加精细。

3.1. 网络整体架构

DFBNet 模型结构如图 2 所示，其中编码器中我用预训练的 Resnet [4]网络作为主干网络进行特征提取。FFN (特征融合网络)用来对特征融合；BRN (边界细化网络)负责产生最后的分割结果。给定一个大小为 H × W，通道数为 3 的 RGB 图像，首先经过主干网络 res-1 至 res-4 进行特征提取，生成 4 个阶段的特征图。它们分别是原始图片分辨率的 1/4，1/8，1/16，1/32。通道数记为 C_1，C_2，C_3，C_4 接着使用 1 × 1 的卷积进行降维处理，把它们的通道数都降低为 C_1，这样做是为了使得各个阶段的特征图在经过特征融合网络时，不会因为某一阶段的特征图通道样本数量所占比重较大而影响模型的学习，接着我们将这些特征图进行融合，融合的方式为：

$$\boldsymbol{H}_l = F\left(\boldsymbol{X}_l, \boldsymbol{X}_{l+1} \cdots \boldsymbol{X}_4\right) \tag{1}$$

其中 $\boldsymbol{H}$ 为融合之后的特征，$\boldsymbol{X}$ 为主干网 res-1 到 res-4 经过卷积降维后产生的特征矩阵。$\boldsymbol{X}_4$ 是主干网络所生成的最后一个特征，所包含的语义信水平最高。l 越大，表示语义信息的层次越高，但尺寸越小。$l \in [1,2,3]$，F 表示特征融合网络。

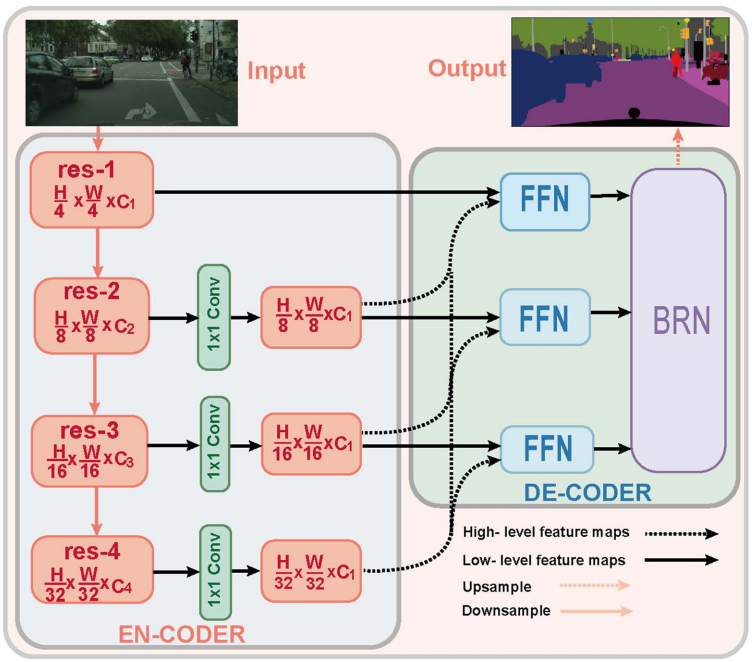

Figure 2. Network model diagram DFBNet
图 2. DFBNet 模型架构图

3.2. 特征融合网络

我们的特征融合的方法与其他网络使用的简单的方式不同，我们考虑到特征图所包含的语义信息和分辨率的不同，因此将特征图在通道这一维度上都拼接在一起，来丰富特征信息。考虑到数据集中目标物体的形状大小都不一样，所以我们决定利用 4 个不同空洞率的组卷积分支来对特征图中不同尺寸的目标进行特征提取，之后利用注意力机制进行权重分配，再将 4 个分支互相通信，让特征更有效的融合，如图 3 所示。

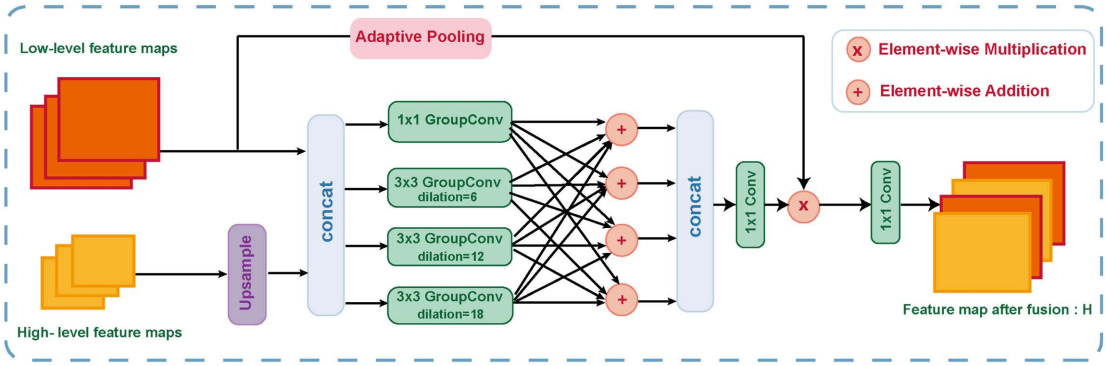

Figure 3. FFN: feature fusion network
图 3. 特征融合网络

首先将高级特征图上采样至与低级特征图尺寸大小一致，并进行通道拼接，得到多级特征，记为 $E=[e_1,e_2,\cdots,e_l]$。将拼接好的特征 E 经过 4 个分组空洞卷积并进行融合，得到特征 $E'=[e_1',e_2',\cdots,e_l']$。接着将低级特征图 $\boldsymbol{X}$ 经过平均池化操作，得到 1 × 1 大小，通道数为 C_1 的一组权重 $\boldsymbol{W}$

$$\boldsymbol{W}=AveragePooling(\boldsymbol{X}) \tag{2}$$

接着将得到的权重 $\boldsymbol{W}$ 按照通道的维度与融合好的特征 $\boldsymbol{E}'$进行相乘得到特征 $\boldsymbol{X}'$。

$$\boldsymbol{X}'=\boldsymbol{W}\boldsymbol{E}' \tag{3}$$

将 $\boldsymbol{X}'$经过 1 × 1 的卷积进行降维处理，与可学习的特征变换矩阵 $\boldsymbol{W}_K$ 相乘将通道数降为与数据集类别数 N 一致，得到融合之后的特征，σ 是激活函数(本研究采用 ReLU 作为激活函数)。

$$\boldsymbol{H}=\sigma(\boldsymbol{W}_K\boldsymbol{E}'+\boldsymbol{b}_1) \tag{4}$$

3.3. 边界细化网络

由于图中物体种类繁多，在这种情况下，模型容易对边界处的像素进行错误分类，于是我们将三个特征融合网络融合好的特征图逐步使用上采样操作并进行相加，然后拼接在一起，采用通道注意力来为不同的通道分匹配权重，学习出重要的通道，接着再利用空间注意力为每个像素分配权重，使网络注意到重要的位置进行学习。这样做可以有效的提升目标边界分割效果。如图 4 所示。

将三个已经融合好的特征图逐步上采样并进行相加，之后进行通道拼接得到特征 $\boldsymbol{Z}$；接着用 1 × 1 卷积将特征 $\boldsymbol{Z}$ 与可学习的特征变换矩阵 $\boldsymbol{W}_q$ 相乘将通道数降为与数据集类别数一致得到特征 $\boldsymbol{Z}'$。

$$\boldsymbol{Z}'=\sigma(\boldsymbol{W}_q\boldsymbol{Z}+\boldsymbol{b}_2) \tag{5}$$

接着经过残差结构进行全局平均池化操作，得到一组权重 $\boldsymbol{W}_v$，将特征 $\boldsymbol{Z}'$与权重 $\boldsymbol{W}_v$ 在通道维度上进行相乘得到特征 $\boldsymbol{K}$。

$$\boldsymbol{K} = \boldsymbol{W}_v \boldsymbol{Z}' \tag{6}$$

将特征 H 经过 sigmoid 激活函数，生成一组可学习权重矩阵 W_b，再与特征 $\boldsymbol{K}$ 对应位置相乘，最后将得到的结果经过 softmax 函数得到最终的分割结果 $\boldsymbol{Y}$，完成端到端的训练过程。

$$\boldsymbol{Y} = softmax\left(\boldsymbol{W}_b \boldsymbol{K}\right) \tag{1}$$

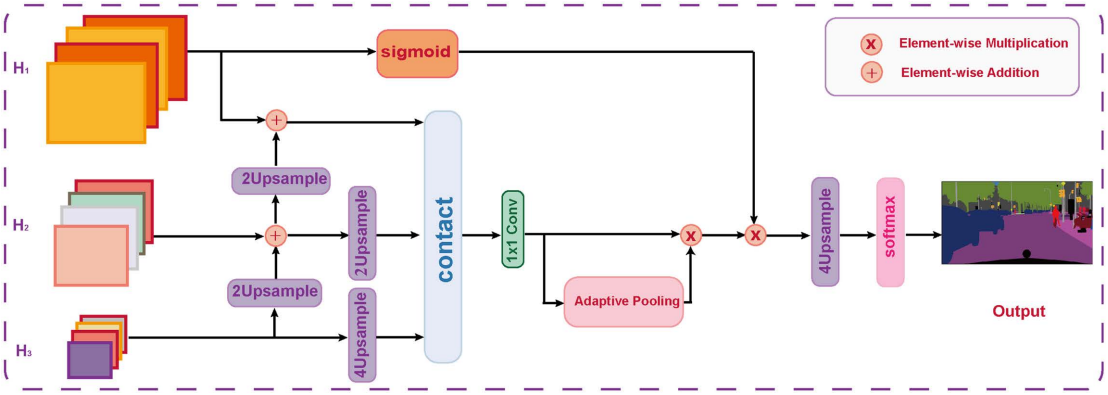

Figure 4. BRN: boundary refinement network
图 4. 边界细化网络

4. 实验结果与分析

为了评估本文提出的分割方法的有效性，我们在两个数据集 Cityscape [5]和 CamVid [6]上验证了我们的方法，并与其他算法进行了比较。

4.1. 数据集简介

Cityscape：Cityscape 是一个基于汽车视角的语义场景分析数据集。它包含 5000 幅精细标注的图像，分为训练集、验证集和测试集，分别包含 2975 幅、500 幅和 1525 幅图像。标注包括 30 类其中 19 类用于语义分割任务，图像分辨率高，为 2048 × 1024。

Camvid：剑桥驾驶标记视频数据库(Camvid)是一个从驾驶角度拍摄的道路场景数据集。该数据集包含从视频序列中提取的 701 幅标注图像，其中 367 幅用于训练，101 幅用于验证，233 幅用于测试。图像分辨率为 960 × 720，32 个语义类别，其中 11 个类别子集用于分割实验。

4.2. 实验过程

我们使用 MMSegmentation [7]代码库，并在服务器上用两张 NVIDIA RTX 3090 进行训练，在 ImageNet-1K 数据集上预训练编码器，并随机初始化解码器。在训练过程中，Cityscape 采用随机调整大小、随机水平翻转、随机裁剪 0.5~2.0 倍的方法进行训练，使用全尺寸进行训练。我们使用 SGD 优化器训练模型，在 Camvid 上进行 80 K 迭代。我们使用的批量大小是 8。将学习速率设置为初始值 0.01，然后使用 Poly 学习速率下降测量值，默认系数为 1.0。

4.3. 与相关方法的比较分析

对 Cityscapes 数据集的分割结果如表 1 所示，我们的模型分割效果最好在 MIoU 评价指标上高于 PointRend、FPN、SETR 等经典分割网络。在 MIoU 指标上，我们的方法比 PointRend 高 0.8%，比 FPN 高 1.7%，比 SETR 高 0.08%。PointRend 将计算机图形学中的分割问题视为“渲染”问题，对自适应选择

的位置进行基于点的分割预测，可以有效地渲染高分辨率图像，该方法可以使模型对目标对象的边缘更加精细。然而，模型提取图像整体特征的能力较弱。因为 PointRend 只在像素值与相邻值相差较大的位置进行计算，而对于其他所有位置该值都是通过插值已经计算出的输出值得到的。这种方法没有过多考虑图像整体特征的提取。我们提出的方法考虑了图像的整体特征，也考虑了目标边缘的像素预测问题，所以我们的方法略优于 PointRend。FPN 被称作特征金字塔网络，它将特征图逐级进行上采样并且逐像素相加，忽略特征图所包含语义信息的差异。因此我们的方法明显优于 FPN。SETR 是一种基于 Transformer 设计的网络，Transformer 具有很强的全局信息控制能力，但由于我们提出的方法包含了边界细化网络，SETR 网络的分割性能略逊于我们的方法，这也是 Transformer 缺陷。我们的模型的参数仅为 46.9 M，是 SETR 的 1/6，这也说明模型实现的分割性能不依赖于参数的叠加，我们的方法不仅分割更准确，而且占用的计算量更少，这意味着我们的模型能够应用于现实世界。

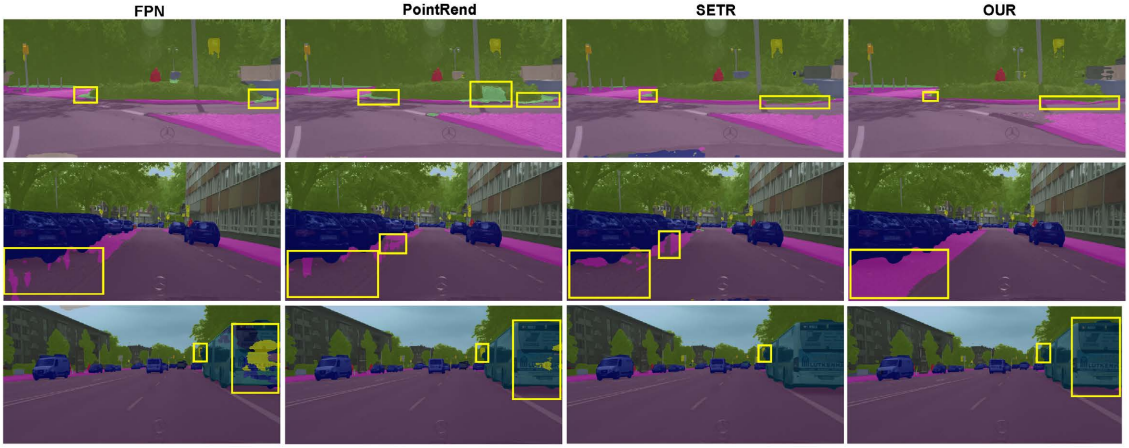

Figure 5. Segmentation results of the Cityscapes dataset
图 5. Cityscapes 数据集分割结果展示

如图 5 所示，FPN 的特征融合方法比较简单，所以从 FPN 的分割效果图中可以看出，FPN 很容易对目标的主体部分进行误分类。公共汽车的尾部错分为交通灯，物体边缘的像素分割也比较粗糙，地形和植被边缘的像素分割也比较模糊。在图 5 中，PointRend 的分割结果中，可以看到图片中有很大一部分道路没有被正确识别出来，在公交车末端也有部分分割错误。SETR 基于 Transformer，具有良好的全局信息控制能力，所以从分割图中可以看出，对象的主体划分基本没有错误，对主体进行了正确的分割，只有边界处的像素分割有点粗糙，把公交车的后视镜误当成了交通灯。对局部信息的控制能力差也是 Transformer 的弱点。我们的模型分割效果最好。这是因为我们提出的方法将每个阶段生成的特征图与前一阶段生成的特征图进行融合，从而弥补了图像下采样所造成的特征图空间信息的丢失。在 FFN 网络中，使用不同感受野大小的卷积分支来感知图中不同大小的物体。然后利用注意机制使网络学习到重要通道的特征，使最终的特征图能够包含更多的空间信息和语义信息。图像的分割效果得到了很大的提高。在最后阶段，我们的 BRN 网络为每个像素位置赋权，使模型注意到物体的重要位置，使得分割结果更加准确。这就是我们的特征融合网络和边界优化网络的用武之地。

Camvid 数据集的结果如表 2 所示，我们的模型达到了 75.13% MIoU。超越一些高性能网络，如 STDC2seg、BiseNetV1 和 V2。我们的方法比 STDC2seg 高 1.2%，比 BiseNetV1 高 9%，比 BiseNetV2-L 高 1.9%。BiseNetV1 也是一个经典的语义分割网络，虽然它意识到语义信息和空间信息的区别，并专门提出了语义分支和空间分支，但其融合方法相对简单，仅进行乘法处理，这就是为什么 BiseNetV1 的分

割性能比我们的差。STDC2seg 设计了一个短期密集级联模块，用于提取接收域可变、多尺度信息的深度特征。并提出细节聚合模块，以更准确地保存底层的空间细节。但其特征融合方法只是简单的加法运算，分割结果不如我们的好。

Table 1. Performance results on the Cityscapes validation set
表 1. Cityscapes 验证集分割结果对比

方法	主干网络	Parameters 参数量	MIoU (%)
PSPNet [8]	Res101	68.10 M	78.50
PSANet [9]	Res101-d8	--	77.90
FPN [10]	Res101	47.51 M	77.70
GCNet [11]	Res101	46.90 M	78.10
PointRend [12]	Res101	--	78.60
SETR [3]	SETR-PUP	318.31 M	79.39
Trans4PASS [13]	PVT-T	22.20 M	79.10
Buffer lader [14]	DeepLabV3+	--	76.88
OUR	Res101-d8	46.90 M	79.47

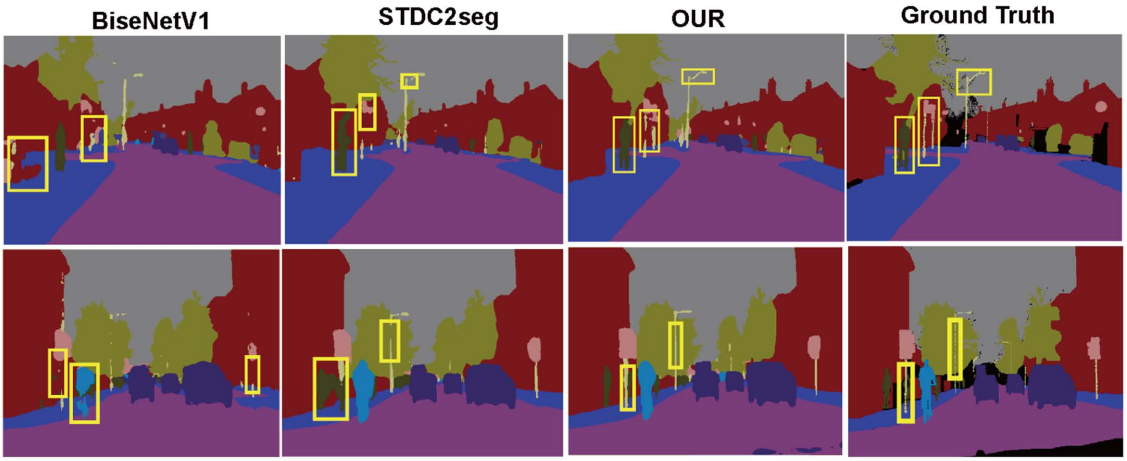

Figure 6. Segmentation results of the Camvid dataset
图 6. Camvid 数据集分割结果展示

从图 6 中可以看出，BiseNetV1 的分割比较粗糙，没有将行人划分为一个整体，轮廓非常粗糙。STDC2seg 没有完全划分道路两侧的边缘，我们的方法是比较准确的。这再次证明了我们提出的特征融合方法的有效性。

Table 2. Performance results on the Camvid test dataset
表 2. Camvid 测试集分割结果对比

方法	主干网络	MIoU (%)
PSPNet [8]	Res50	69.10
BiseNetV1 [15]	Xception39	65.60
ICNet [16]	PSPNet50	67.10

续表

DenseDecoder [17]	Res101	70.90
BiseNetV2-L [18]	Res101	73.20
STDC2seg [19]	STDC2	73.90
PP-LiteSeg [20]	STDC2	75.00
Light-Deeplabv3+ [21]	DeeplabV3+	74.54
OUR	Res101-d8	75.13

4.4. 消融实验

为了证明我们所提出方法的有效性，我们建立了以下消融实验，首先，我们分别去除 FFN 网络和 BRN 网络，并在 cityscape 数据集上进行实验。实验结果如表 3 所示。并选取部分分割结果进行展示，如图 7 所示。从图 7 中也可以清楚地看到，去除特征融合网络后，分割效果相对较差，即使不是相似的对象，也容易错分，而去除边界细化网络的分割结果，整体上没有明显的错分，但在对象的边缘处只存在一个轻微的分割缺陷。

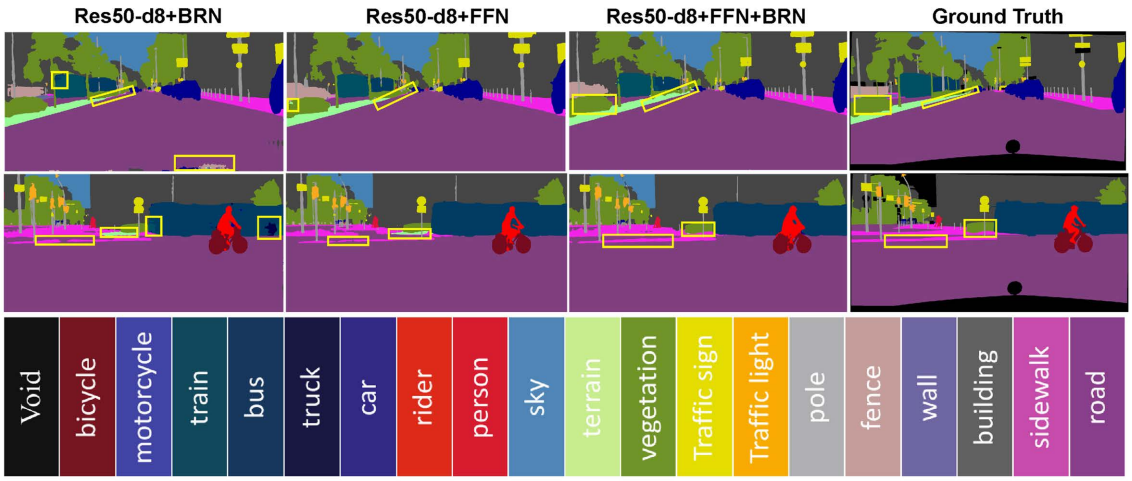

Figure 7. The segmentation results of the Cityscapes ablation experiment are displayed
图 7. Cityscapes 消融实验分割结果展示

Table 3. Results of ablation experiments on the Cityscapes validation set
表 3. Cityscapes 数据集消融实验结果

方法	MIoU (%)
Res50-d8 + FFN	78.07
Res50-d8 + BRN	72.88
Res50-d8 + FFN + BRN	79.02
Res101-d8 + FFN + BRN	79.47

从表 3 中可知，去掉特征融合网络后网络对图像中较大的目标的分割精度都大幅度降低，例如人行道、墙壁、植被等类别分割效果变差。这些物体在图中占比较大，去掉了特征融合网络后，网络的对于特征的融合能力下降，对于此类占比较大的物体无法准确提取其特征，才造成分割精度急剧下降。相较

于去掉边界细化网络后，分割精度仅有微小下降，这证明我们从特征融合和边界细化的方向思考改进图像分割是正确的，实验表明特征融合的质量对分割结果起着决定性的作用，而对象边缘像素的分割也对整体分割结果起到了积极的作用。

5. 结论

本文首先指出，以往分割网络中存在的特征融合方法过于简单，不能充分结合特征图中包含的语义信息和空间信息，导致目标对象划分错误，目标边界划分粗糙等问题。然后我们提出了我们的网络，DFBNet，来解决这个问题。它包括特征融合网络和边界优化网络两个子网络。特征融合网络从特征图大小和语义信息差异的角度出发，利用不同膨胀率的多组卷积从特征图中提取不同大小的对象特征，然后利用注意机制分配权重，弥补了特征图之间的差距，使得不同大小的特征图能够充分的融合，并通过实验证明了该方法的有效性。边界细化网络融合了不同阶段的分割结果。它利用空间注意力为每个像素位置分配权重，从而使模型能够对目标对象的边界进行精细分割。实验表明，该方法在 Cityscape、Camvid 数据集上取得了良好的分割效果。我们的工作证明了从特征融合的角度提高语义分割准确率的想法是完全可行的。

参考文献

[1] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[2] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, 5-9 October 2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[3] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S. and Zhang, L. (2021) Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), 20-25 June 2021, 6877-6886. https://doi.org/10.1109/CVPR46437.2021.00681

[4] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[5] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S. and Schiele, B. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 3213-3223. https://doi.org/10.1109/CVPR.2016.350

[6] Brostow, G.J., Fauqueur, J. and Cipolla, R. (2009) Semantic Object Classes in Video: A High Definition Ground Truth Database. *Pattern Recognition Letters*, **30**, 88-97. https://doi.org/10.1016/j.patrec.2008.04.005

[7] Contributors, M. (2020) MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation

[8] Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017) Pyramid Scene Parsing Network. *Proceedings of the IEEE Conference* on *Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2881-2890. https://doi.org/10.1109/CVPR.2017.660

[9] Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D. and Jia, J. (2018) Psanet: Point-Wise Spatial Attention Network for Scene Parsing. *Proceedings of the European Conference on Computer Vision* (*ECCV*), 8-14 September 2018, 267-283. https://doi.org/10.1007/978-3-030-01240-3_17

[10] Kirillov, A., Girshick, R., He, K. and Dollar, P. (2019) Panoptic Feature Pyramid Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 6399-6408. https://doi.org/10.1109/CVPR.2019.00656

[11] Cao, Y., Xu, J., Lin, S., Wei, F. and Hu, H. (2019) Gcnet: Non-Local Networks Meet Squeeze Excitation Networks and Beyond. 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*), Long Beach, 16-17 June 2019, 1971-1980. https://doi.org/10.1109/ICCVW.2019.00246

[12] Kirillov, A., Wu, Y., He, K. and Girshick, R. (2020) Pointrend: Image Segmentation as Rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 9799-9808. https://doi.org/10.1109/CVPR42600.2020.00982

[13] Zhang, J., Yang, K., Ma, C., Reiß, S., Peng, K. and Stiefelhagen, R. (2022) Bending Reality: Distortion-Aware Transformers for Adapting to Panoramic Semantic Segmentation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 16896-16906. https://doi.org/10.1109/CVPR52688.2022.01641

[14] Liu, Z. and Lei, Z. (2023) Buffer Ladder Feature Fusion Architecture for Semantic Segmentation Improvement. *Signal, Image and Video Processing*, **18**, 475-483. https://doi.org/10.1007/s11760-023-02754-1

[15] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G. and Sang, N. (2018) Bisenet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. *Proceedings of the European Conference on Computer Vision* (*ECCV*), 8-14 September 2018, 325-341. https://doi.org/10.1007/978-3-030-01261-8_20

[16] Zhao, H., Qi, X., Shen, X., Shi, J. and Jia, J. (2018) ICNET for Real-Time Semantic Segmentation on High-Resolution Images. *Proceedings of the European Conference on Computer Vision* (*ECCV*), 8-14 September 2018, 405-420. https://doi.org/10.1007/978-3-030-01219-9_25

[17] Bilinski, P. and Prisacariu, V. (2018) Dense Decoder Shortcut Connections for Single-Pass Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 6596-6605. https://doi.org/10.1109/CVPR.2018.00690

[18] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C. and Sang, N. (2021) Bisenet V2: Bilateral Network with Guided Aggregation for Realtime Semantic Segmentation. *International Journal of Computer Vision*, **129**, 3051-3068. https://doi.org/10.1007/s11263-021-01515-2

[19] Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J. and Wei, X. (2021) Rethinking Bisenet for Realtime Semantic Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 19-25 June 2021, 9716-9725. https://doi.org/10.1109/CVPR46437.2021.00959

[20] Peng, J., Liu, Y., Tang, S., Hao, Y., Chu, L., Chen, G., Wu, Z., Chen, Z., Yu, Z., Du, Y., Dang, Q., Lai, B., Liu, Q., Hu, X., Yu, D. and Ma, Y. (2022) Pp-Liteseg: A Superior Realtime Semantic Segmentation Model.

[21] Ding, P. and Qian, H. (2023) Light-Deeplabv3+: A Lightweight Real-Time Semantic Segmentation Method for Complex Environment Perception. *Journal of Real-Time Image Processing*, **21**, Article No. 1. https://doi.org/10.1007/s11554-023-01380-x