许鸿奎, 郭文涛, 李振业,等. 多特征融合网络在街道场景中的应用[J]. 智能计算机与应用,2025,15(9):124-131. DOI:10.20169/j.issn.2095-2163.250920

多特征融合网络在街道场景中的应用

许鸿奎[1,2], 郭文涛[1], 李振业[1], 赵京政[1], 郭旭斌[1]
(1 山东建筑大学 信息与电气工程学院, 济南 250101; 2 山东省智能建筑技术重点实验室, 济南 250101)

摘 要: 双分支网络结构在实时语义分割任务中显示出其高效和准确性,然而低级细节与高级语义信息融合过程中会导致细节特征被周围的上下文信息所掩盖,导致边缘模糊化。针对此问题提出一种三分支网络结构,该架构具有 3 个分支、分别提取空间信息、上下文信息和边界信息。在语义提取网络中,放弃了传统的 CNN 卷积方式,采用了新型的非跨行卷积方式,并通过深度聚合模块对语义信息进行深度提取,在最后的融合阶段利用边界信息来指导空间信息与高级语义信息的融合,从而提高语义分割网络的性能。最后将所设计的网络结构在城市景观数据集上进行实验,取得了 78.8%的平均交并比,推理速度为 80.2 FPS,在速度与准确性之间达到了平衡。

关键词: 双分支网络; 信息融合; 三分支网络; 非跨行卷积; 深度聚合
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)09-0124-08

Application of multi feature fusion network in street scene

XU Hongkui[1,2], GUO Wentao[1], LI Zhenye[1], ZHAO Jingzheng[1], GUO Xubin[1]
(1 School of Information and Electrical Engineering, Shandong Jianzhu University, Jinan 250101, China;
2 Shandong Key Laboratory of Intelligent Buildings Technology, Jinan 250101, China)

Abstract: The dual-branch network structure has demonstrated its efficiency and accuracy in real-time semantic segmentation tasks. However, the fusion of low-level details and high-level semantic information can lead to the obscuring of detailed features by surrounding contextual information, resulting in edge blurring. To address this issue, a tri-branch network structure is proposed, which consists of three branches for extracting spatial information, contextual information, and boundary information. In the semantic extraction network, the traditional CNN convolution method has been abandoned in favor of a novel non-cross-row convolution approach, and semantic information is deeply extracted through a depth aggregation module. In the final fusion stage, boundary information is utilized to guide the fusion of spatial and high-level semantic information, thereby enhancing the performance of the semantic segmentation network. The designed network structure is tested on a urban landscape dataset, achieving an average intersection over union of 78.8% and an inference speed of 80.2 FPS, striking a balance between speed and accuracy.
Key words: dual-branch network; information fusion; three branch network; SPD-Conv; deep aggregation

0 引 言

图像分割在计算机视觉领域是一个重要的研究方向,旨在实现对图像中的像素进行类别标签的分配。在深度学习方法还未获普及的阶段,人们使用数学和数字图像处理领域的知识去做分割任务。但在传统方法中,提取的特征只能是一些低层次的如颜色、纹理和轮廓等属性。这些特征是通过对图像的底层属性进行深入理解来获得的,然后基于这些特征对图像进行分类和标签。在传统算法中常用的方法包括阈值方法[1]、聚类方法[2]、图划分[3]。尽管传统算法可以从图像中获取大量的特征信息,但这些信息都是一些低层次特征,仍然解决不了分割精度低的问题。近年来,随着深度神经网络的发展,卷积神经网络(CNN)在图像分割领域的应用取得了显著的成效。相比传统的手动特征提取方法,基

基金项目: 山东省重大科技创新工程项目(2019JZZY010120); 山东省重点研发计划项目(2019GSF111054)。
作者简介: 郭文涛(1998—),男,硕士研究生,主要研究方向:计算机视觉; 李振业(1999—),男,硕士研究生,主要研究方向:计算机视觉; 赵京政(2000—),男,硕士研究生,主要研究方向:自然语言处理; 郭旭斌(1999—),女,硕士研究生,主要研究方向:自然语言处理。
通信作者: 许鸿奎(1966—),男,博士,教授,主要研究方向:模式识别,智能信息处理。Email:xhkui2009@163.com。
收稿日期: 2023-12-26

于深度学习的图像分割算法在性能上具有明显的优势。

在最近几年中，深度学习取得的可观进展，也不断推动着图像语义分割领域的技术进步。众多模型结构如 AlexNet[4]、GoogLeNet[5]、ResNet[6]、VGG[7]等被相继提出，这也引发了学术界对分割领域研究的极大兴趣。2015 年，Long 等学者[8]提出了全卷积神经网络（Fully Convolutional Networks，FCN）来进行端到端的图像语义分割。网络结构以 VGG16 为主干网络，在网络的最后一层将全连接层替换为卷积层。虽然该技术优势明显，但是分割后的图像仍然存在边界粗糙和分割精度准确率低的情况，如何更有效地结合和利用多尺度特征，仍是提高语义分割性能的一个重要研究方向。因此，人们开始考虑如何更好地利用网络结构中不同层次的信息，通过将不同层次的信息进行融合，以达到提高分割精度的目的。这种全卷积神经网络不仅展示了如何端到端训练卷积神经网络来进行图像语义分割，而且经实验证实其分割精度相比传统方法有了显著的提高。

1　相关工作

在语义分割任务中，不管是在提取空间信息、或者语义信息的过程中，总会忽略图像中的边缘信息。而能够确保对空间信息、语义信息与边缘信息的充分提取则是提升分割精度的关键步骤。针对图像中存在的小物体如何提取其中的语义信息也是研究的重要问题。

1.1　编码器-解码器结构

早期语义分割方法主要基于编码器-解码器架构。然而，2015 年，Long 等学者[8]提出了全卷积神经网络，该方法在语义分割任务中将卷积神经网络最后一层替换为反卷积层进行上采样，并预测每个像素的类别。这一创新标志着卷积神经网络在语义分割领域的兴起。随后的 2017 年，Badrinarayanan 等学者[9]推出了 SegNet，该网络由编码器和解码器组成。其中，编码器基于 VGG16 模型进行物体信息解析，而解码器则将这些解析后的信息转化为图像形式，实现像素级别的物体信息表达。同年，DeepLab[10]被提出，并改变了 ResNet 网络中的部分下采样操作，这个网络在结构中维持高分辨率，而且在网络结构中存在大膨胀卷积增大感受野。自此，基于扩展卷积和上下文提取模块的主干网络成为各种方法的标准配置，例如 DeepLabV2[11]、DeepLabV3[12]、PSPNet[13]和 DenseASPP[14]。

尽管扩展卷积模型具有强大的性能，编码器-解码器结构在计算和推理时间上的表现则更出色。编码器通常以深度网络的形式存在，主要负责处理和压缩输入数据，从中提取出有用的特征和上下文信息；解码器通过插值或转置卷积技术，将编码器所提取的上下文信息，用来恢复原始图像的分辨率。此外，编码器结构的网络既能在 ImageNet 数据集上做预训练，也可以不做预训练，直接在准备好的数据集上进行训练，就像 ERFNet[15]和 ESPNet[16]。如 FANet[17]则是通过引入快速注意模块和整个网络的额外下采样，成功实现了速度和准确性之间的平衡。最后，SFNet[18]则通过流对齐模块对齐相邻层的特征映射，进一步优化了特征融合的效果。综上可知，在语义分割领域取得的可观成果也相继带动了有关研究的发展。

1.2　双分支结构

虽然编码器-解码器架构在计算量降低方面有不错的表现，但是在图像下采样的过程中可能会存在信息丢失的问题，导致反采样的信息不完整，进而影响了分割精确度。为了解决这一问题，旷视科技团队创新性地提出一种双分支的网络架构。在这种架构中，除了一条用于提取语义信息的路径外，还有一条保持高分辨率的浅路径，能够提供丰富的空间细节作为补充，从而提高了信息的完整性。

2018 年，双分支网络结构 BiseNet[19]被提出。在双路径架构中，一个分支主要负责提取图像的空间特征信息，另一个分支则专注于提取更丰富的语义信息。由于这 2 个分支处理不同类型的特征，其特征维度存在差异。为了更好地融合这 2 个分支的特征，引入了特征融合模块（FFM）和注意力优化模块（ARM）以增强网络性能。接下来在 2019 年，Li 等学者[20]提出了一种名为 DFANet 的网络结构。通过子网和子级联的方式聚合判别性特征，采用深度多尺度特征聚合和轻量级深度可分离卷积，进一步提升了语义分割的精度。最近，在 2021 年，Hong 等学者[21]提出了 DDRNet 网络结构。这个结构将 2 个分支进行深度融合，充分利用了不同阶段的分辨率，从而实现了更高的语义分割性能。这些创新在不断提高语义分割精度的同时，也展示了深度学习在图像处理领域的巨大潜力。

1.3　上下文信息

在语义分割中，捕获更丰富的上下文信息是一个关键挑战。为了解决这一问题，人们提出一种分

层空间金字塔池。这个结构由不同大小卷积核的分层组成,致力于能够将不同尺度的信息进行同时处理。通过在不同尺度的特征图上执行卷积操作,ASPP 能够捕获到更广泛的空间信息,从而更好地理解图像的上下文。PSPNet[13] 中的金字塔池模块(PPM)通过在卷积之前实现金字塔池,比 ASPP 具有更高的计算效率。与卷积核的局部性不同,自注意力机制善于捕获全局依赖性。通过这种方式,DANet[22] 网络结构采用了位置信息和通道信息,得到了更加丰富的特征信息。OCNet[23] 网络结构为了获得更加丰富的上下文信息,利用了注意力机制。总结来说,当前的工作主要在摸索如何通过网络结构获取更加丰富的语义信息以及如何将不同分辨率的空间信息与语义信息更好地结合起来。随着深度学习技术的持续进步和新方法的不断涌现,语义分割领域在未来会取得更多的突破和进展。

本文主要贡献如下:

(1)设计了一种三分之网络结构,多加一条网络分支进行物体边缘信息的提取,并设计了一个特殊模块利用边界信息指导空间信息与语义信息相融合。

(2)在图像下采样过程中,放弃了传统的 CNN 卷积模块,采用了一种非跨行卷积模块,提高了对图像中小物体的检测。

(3)设计了一种深度聚合模块(Deep Aggregation Module,DAM),采用了串联的方式,从而获得了丰富的上下文信息,同时又加快了模型推理速度。

2 多特征融合网络

网络结构如图 1 所示。图 1 中,SPD-Conv 为采用非跨行卷积模块;DAM 表示为深度聚合模块;Bag 表示边界指导融合模块。主干网络结构为三分支,能提取图片空间信息、语义信息和边缘信息,通过双侧融合模块,对空间信息与语义信息,边缘信息与语义信息进行融合,加强了分支之间的关联性。其网络结构详细结构见表 1。

表 1 网络结构图

Table 1 Network structure diagram

名称	操作			输出尺寸
Conv1		SPD-Conv 3 × 3,32		*H*/2 × *W*/2
Conv2		SPD-Conv $\begin{bmatrix}3\times3 & 32\\ 3\times3 & 32\end{bmatrix}\times2$		*H*/4 × *W*/4
Conv3		SPD-Conv $\begin{bmatrix}3\times3 & 64\\ 3\times3 & 64\end{bmatrix}\times2$		*H*/8 × *W*/8
Conv4	$\begin{bmatrix}3\times3 & 64\\ 3\times3 & 64\end{bmatrix}\times2$	SPD-Conv $\begin{bmatrix}3\times3 & 128\\ 3\times3 & 128\end{bmatrix}\times2$	$\begin{bmatrix}3\times3 & 64\\ 3\times3 & 64\end{bmatrix}\times2$	*H*/8 × *W*/8 *H*/16 × *W*/16 *H*/8 × *W*/8
	Bilateral fusion		Bilateral fusion	
Conv5	$\begin{bmatrix}3\times3 & 64\\ 3\times3 & 64\end{bmatrix}\times2$	SPD-Conv $\begin{bmatrix}3\times3 & 256\\ 3\times3 & 256\end{bmatrix}\times2$	$\begin{bmatrix}3\times3 & 64\\ 3\times3 & 64\end{bmatrix}\times2$	*H*/8 × *W*/8 *H*/32 × *W*/32 *H*/8 × *W*/8
	Bilateral fusion		Bilateral fusion	
Conv6	$\begin{bmatrix}1\times1 & 64\\ 1\times1 & 64\\ 1\times1 & 128\end{bmatrix}\times1$	SPD-Conv $\begin{bmatrix}1\times1 & 256\\ 3\times3 & 256\\ 1\times1 & 512\end{bmatrix}\times1$	$\begin{bmatrix}1\times1 & 64\\ 3\times3 & 64\\ 1\times1 & 128\end{bmatrix}\times1$	*H*/8 × *W*/8 *H*/64 × *W*/64 *H*/8 × *W*/8

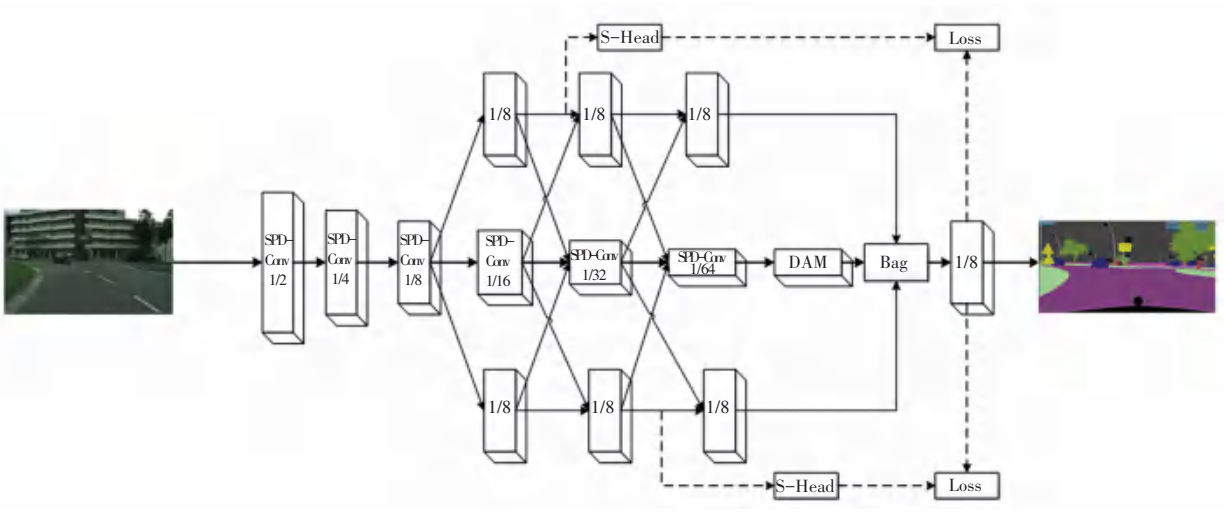


图 1　网络结构图

Fig. 1　Network structure diagram

图像送入网络后，经过新的构建块将图片压缩至 1/8，然后分出 3 个分支。其中，一个分支提取图片中丰富的空间信息；一个分支用于提取图片中的边缘信息；最后一个分支则可提取语义信息。在语义提取的最后，通过本文所设计的深度聚合模块，将图像的 1/64、1/128、1/256、1/512 和 1/1 024 进行深度融合，最大限度地提取语义信息。将 3 个分支提取到的空间信息、边缘信息和语义信息进行融合，融合后的特征采用双线性插值的方式再恢复到原图片分辨率，随后通过 Softmax 函数将图片中的每一个像素进行分类，完成分割任务。

2.1　新型的 CNN 构建块

在计算机视觉任务中、如图像分类和目标检测，卷积神经网络（CNN）已经取得了令人瞩目的成果。然而，当遇到低分辨率图像或较小尺寸的物体时，其性能会受到严重影响。为了解决这个问题，本次研究引入了一种名为 SPD-Conv 的新型 CNN 构建模块，以替代所有的跨步卷积层和池化层，从而优化深度语义信息的提取过程。这样，就可以更有效地处理复杂的视觉任务，尤其是在图像分辨率较低或物体尺寸较小的情况下。SPD-Conv 是由一个空间到深度（SPD）层和一个非跨步卷积层（Conv）组合而成。当输入任意大小为 $H \times W \times C$ 特征图 X 时，将一系列特征映射分割为：

$$
\begin{aligned}
&f_{0,0} = X[0:H:\text{scale},0:W:\text{scale}],\\
&f_{1,0} = X[1:H:\text{scale},0:W:\text{scale}],\cdots,\\
&f_{\text{scale}-1,0} = X[\text{scale}-1:H:\text{scale},0:W:\text{scale}]\\
&f_{0,1} = X[0:H:\text{scale},1:W:\text{scale}],\\
&f_{1,1},\cdots,f_{\text{scale}-1,1} = X[\text{scale}-1:H:\text{scale},1:W:\text{scale}]\\
&\vdots\\
&f_{0,\text{scale}-1} = X[0:H:\text{scale},\text{scale}-1:W:\text{scale}],\\
&f_{1,\text{scale}-1},\cdots,f_{\text{scale}-1,\text{scale}-1} =\\
&X[\text{scale}-1:H:\text{scale}-1,0:W:\text{scale}]
\end{aligned}
\tag{1}
$$

一般地，通过 scale=2 对图片进行下采样，生成 4 个子特征图，即 $f_{0,0}$，$f_{0,1}$，$f_{1,0}$，$f_{1,1}$，每个子特征图的形状为 $\left(\frac{H}{2},\frac{W}{2},C\right)$。此后，通过一个 1×1 的卷积将形成的子特征图进行通道拼接，将通道数压缩到原来大小。图 2 即展示了 scale=2 时的说明。

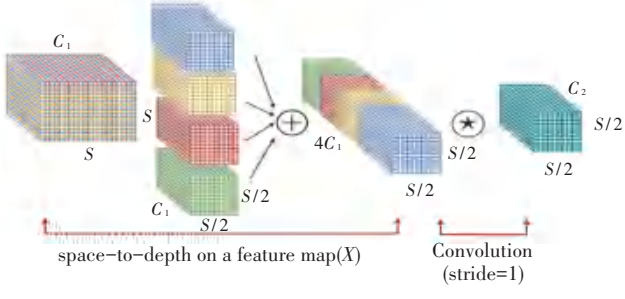


图 2　scale=2 时 SPD-Conv 的下采样方式

Fig. 2　Downsampling method of SPD-Conv at scale=2

2.2　深度聚合模块

SwiftNet[24] 运用空间金字塔池（Spatial Pyramid Pooling，SPP）来解析全局依赖关系，以抽取更深层次的高级语义信息。另外，PSPNet[13] 则采用了金字塔池化模块（PPM），该模块在卷积前将多尺度池化特征图进行连接，从而形成局部和全局的上下文表示。此外，PSPNet[13] 还提出了深度聚合 PPM（Deep Aggregation PPM，DAPPM），使得 PPM 的上下文嵌入能力得到进一步提升，从而展现出卓越的性能。深度聚合图如图 3 所示。本文的串行深度聚合模

块,输入为 1/64,通过串联 4 个池化核为 5 的池化模块,分别获得 1/128、1/256、1/512 和 1/1 024 图像分辨率的特征图,而且将输入特征图与全局池化信息也得到了利用。在多尺度信息融合的过程中,输入特征图被融入各个阶段的上下文信息中,从而更紧密地融合不同尺度的上下文信息。这种融合方式有助于网络更好地理解和利用图像中的上下文信息,提高语义分割的准确性和鲁棒性。其输入设为 x,输出为 y。对此过程可以表示为:

$$y_1 = C_{1\times1}(x) \tag{2}$$

$$y_2 = C_{3\times3}(U(C_{1\times1}(P_{\text{kernel}=5}(x))) + y_1) \tag{3}$$

$$y_3 = C_{3\times3}(U(C_{1\times1}(P_{\text{kernel}=5}(P_{\text{kernel}=5}(x)))) + y_1) \tag{4}$$

$$y_4 = C_{3\times3}(U(C_{1\times1}(P_{\text{kernel}=5}(P_{\text{kernel}=5}(P_{\text{kernel}=5}(x))))) + y_1) \tag{5}$$

$$y_5 = C_{3\times3}(U(C_{1\times1}(P_{\text{kernel}=5}(P_{\text{kernel}=5}(P_{\text{kernel}=5}(P_{\text{kernel}=5}(x)))))) + y_1) \tag{6}$$

其中,$C_{1\times1}$ 表示 1×1 卷积;$C_{3\times3}$ 表示 3×3 卷积;U 表示上采样操作;$P_{\text{kernel}=5}$ 表示核为 5 的池化操作,在最后的阶段,采用 1×1 的卷积将特征映射连接并压缩。此外,为了利于后续优化,还添加了 1×1 投影快捷方式。这种设计能够有效地减少计算量,同时保持较高的分割精度。通过这种结构,网络能够更好地理解和处理图像中的上下文信息,从而提高语义分割的准确性和鲁棒性。

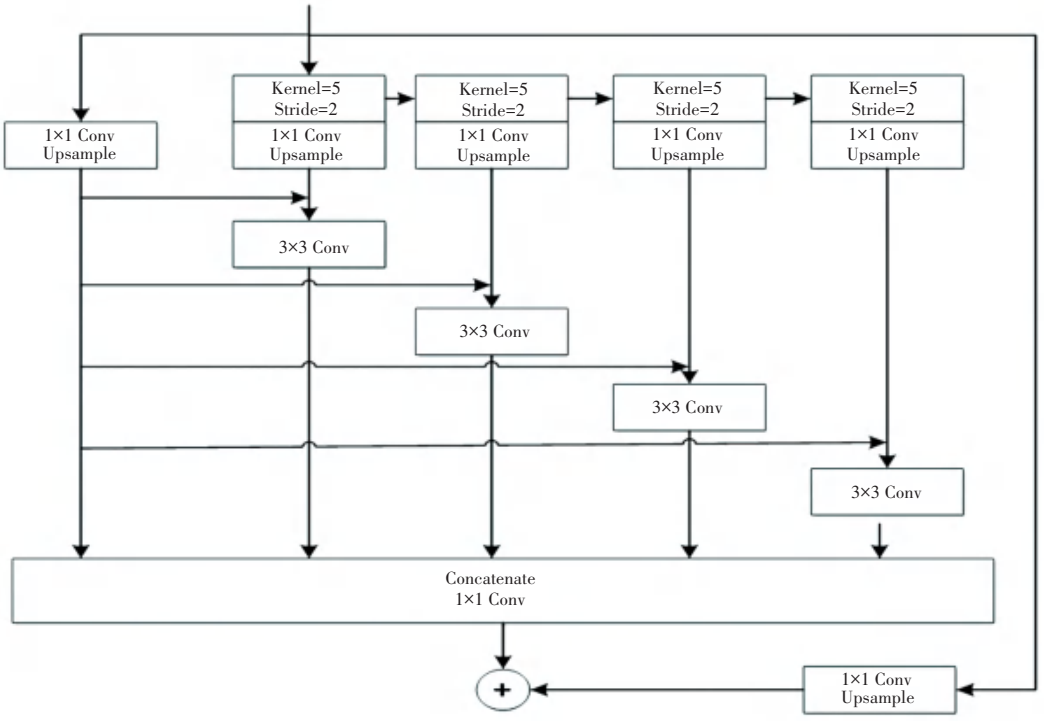


图 3 深度聚合图

Fig. 3 Deep aggregation diagram

2.3 边界指导融合模块

为了将高分辨率的空间信息与丰富的高级语义信息相结合,引入了边界信息作为指导。尽管高级语义信息包含丰富的内容,但却往往丧失了过多的空间和几何细节。细节分支在此方面表现出色,能够更完善地保存空间细节。因此,设计了一种方法,使模型在边界区域更加依赖细节分支,并利用上下文特征来填充对象的内部区域。通过这种方法,本文研发的模块能够更加有效地整合空间信息、高级语义信息和边界信息。

边界指导融合图如图 4 所示。图 4 中,S 表示空间信息,A 表示高级语义信息,E 表示边缘信息。融合公式为:

$$\text{out} = \text{Conv}((\sigma \times \text{Sigmoid}(E) \times S) + (1-\sigma) \times \text{Sigmoid}(E)) \tag{7}$$

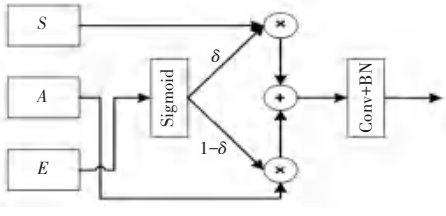


图 4 边界指导融合图

Fig. 4 Border guided fusion diagram

3　实验及分析

3.1　实验环境

本次实验采用的 CPU 为 Intel 处理器，内存为 16 GB，GPU 为 RTX3060，深度学习框架为 Pytorch。详细实验配置见表 2。

表 2　软硬件实验配置环境表

Table 2　Environment table of software and hardware experimental configuration

实验环境	配置
操作系统	Ubuntu 20.04
中央处理器 CPU	11th Gen Intel(R) Core(TM) i5
内存	16 GB
显卡 GPU	GEFORCE RTX 3060
深度学习框架	Pytorch

3.2　实验数据集

本文数据集为大规模城市街道场景语义分割数据集 Cityscapes[25]，主要用于自动驾驶领域，共有 50 个城市街道场景。这个数据集有 2 个版本：一个是精确标注版本，另一个是粗略标注版本。在精确标注版本中，数据集包含 5 000 幅街道场景图像，每幅图像都进行了精确的像素级标注，共涵盖 34 个街景类别。然而，在本次实验中，仅使用了其中的 19 个类别进行实验和评估。具体地，其中 2 975 幅图像用于训练模型，500 幅图像用于验证模型的性能，而剩下的 1 525 幅图像则用于测试模型的最终性能。每幅图像的分辨率均为 1024×2048。

3.3　评估指标

在实验中，采用了平均交并比（Mean Intersection over Union，mIoU）作为算法的评估指标。mIoU 是图像语义分割领域常用的评价标准，衡量了真实标签和预测值之间的重叠程度。研究可知，单个类别的交并比（Intersection over Union，IoU）用于衡量该类别的预测准确度。具体来说，IoU 计算了该类别的真实标签与预测值之间的交集，并除以两者的并集。这个比值越接近 1，表示预测越准确。mIoU 则是所有类别 IoU 的平均值，用于综合评估模型在所有类别上的分割性能。通过计算每个类别的 IoU，并将其取平均值，mIoU 提供了一个整体性能指标，反映了模型在所有类别上的平均分割精度。具体计算公式如下：

$$\mathrm{mIoU}=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{i=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{8}$$

3.4　网络参数设置

预处理将大小统一调整为 1024×1024。在训练的过程中，采用的优化方法为随机梯度下降（Stochastic Gradient Descent，SGD）方法，并将学习率设置为 0.01。为了加速收敛并提高稳定性，给 SGD 添加了动量项，其值为 0.9。此外，还引入了权重衰减来正则化模型，其值为 0.000 5。在训练过程中，使用大小为 32 的批次（Batch_Size）进行迭代。通过这种方式，可以在每个训练步骤中处理一批次的图像数据，并基于这些数据更新模型的权重。将这一过程重复 500 个 epoch，以确保模型充分地学习了训练数据。在测试阶段，采用输入图片尺寸为 1 024×2 048 进行评估。

3.5　对比试验

在 Cityscapes 数据集上将本文方法同 SegNet[9]、ICNet[26]、DFANet[20]、BiSeNet[19] 和 SFNet[18] 等一系列网络结构进行对比。实验结果对比见表 3。从表 3 中可以看出，本文所采用的网络结构在平均交并比上达到了 78.8%，并且其速度达到了 80.2 FPS，在准确率与速度之间达到平衡。

表 3　实验结果对比

Table 3　Comparison of experimental results

Model	FPS	mIoU/%	Params/M
SegNet[9]	16.7	57.0	29.50
ICNet[26]	30.0	69.5	26.50
DFANet[20]	100.0	71.3	7.80
BiSeNet[19]	74.8	65.5	49.00
SFNet[18]	30.4	78.9	12.87
本文	80.2	78.8	9.03

3.6　消融实验对比

为验证三分支网络结构、深度聚合模块与新的 CNN 构建块的有效性，本文进行了消融实验。具体实验结果见表 4。

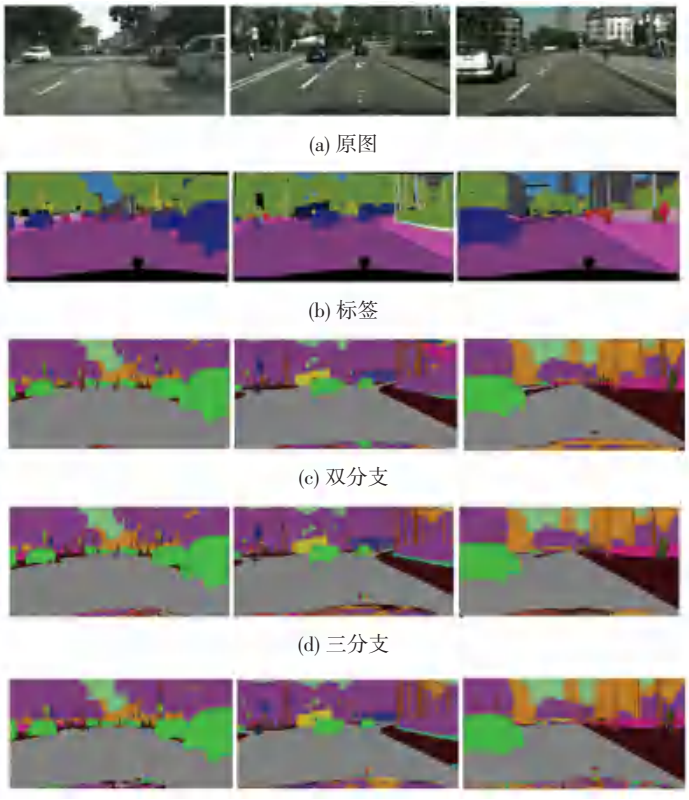
表 4　消融实验对比

Table 4　Comparison of ablation experiments

Group	FPS	mIoU/%	Params/M
双分支	87.4	65.5	5.73
三分支	84.2	70.2	7.70
三分支+SPD-Conv+DAM	80.2	78.8	9.03

3.7　视图可视化

视图可视化结果如图 5 所示。通过对分割图像的可视化，可以看出三分支网络结构、深度聚合模块和新的 CNN 构建块在边缘模糊和小物体检测不明显的问题上得到改善。



(a) 原图

(b) 标签

(c) 双分支

(d) 三分支

(e) +DAM+SPD-Conv

图 5 可视化视图

Fig. 5 Visualization view

4 结束语

本文提出一种新颖的三分支网络结构来进行图像分割。主干网络使用了一种用于小物体的新 CNN 构建块来提取网络特征,并利用深度聚合模块对图片进行深度信息提取;一个分支用于提取图片的空间信息,而另一个分支则专注于提取图片的边缘信息。通过边缘信息来指导空间信息与高级语义信息的融合,实现了更加精准的分割结果。在 Cityscapes 数据集上,本文所提的方法取得了出色的分割精度。未来,将继续本文所提高不同物体的分割精度,并将模型加以简化以加快训练速度,同时确保模型的准确性和实时性。

参考文献

[1] SAHOO P K, SOLTAIN S, WONG A K C. A survey of thresholding techniques[J]. Computer Vision Graphics and Image Processing,1988,41(2):233-260.

[2] KANUNGO T, MOUNT D M, NETANYAHU N S, et al. An efficient k - means clustering algorithm: Analysis and implementation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2002,24(7):881-892.

[3] SHI Jiaobo, MALIK J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-905.

[4] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[5] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions [C]//Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition. Piscataway,NJ:IEEE, 2015: 1-9.

[6] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE, 2016:770-778.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large - scale image recognition[J]. arXiv preprint arXiv,1409. 1556, 2014.

[8] LONG J, SHEIHAMER E, DARRELL T, et al. Fully convolutional networks for semantic segmentation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2015:3431-3440.

[9] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: A deep convolutional encoder - decoder architecture for imge segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,39(12):2481-2495.

[10] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv,1412. 7062, 2014.

[11] 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.

[12]CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation [J]. arXiv preprint arXiv,1706.05587, 2019.

[13]ZHAO Hengshuang, SHI Jianping, QI Xiaojuan. et al. Pyramid scene parsing network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 2881-2890.

[14]YANG Maoke, YU Kun, ZHANG Chi. et al. Denseaspp for semantic segmentation in street scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE, 2018:3684-3692.

[15]ROMERA E, AIVAREZ J M A, 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.

[16]MEHTA S, RASTEGARI M, CASPI A, et al. ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation [C]//Proceedings of the European Conference on Computer (ECCV). Cham:Springer,2018: 552-568.

[17]HU Ping, PERAZZI F, HEILBRON F C, et al. Real-time semantic segmentation with fast attention[J]. IEEE Robobics and Automation Letters,2020,6(1):263-270 .

[18]LI Xiangtai, YOU Ansheng, ZHU Zhen, et al. Semantic flow for fast and accurate scene parsing[C]//Proceedings of the 16th European Conference on Computer Vision. Cham: Springer, 2020:775-793.

[19]YU Changqiang, WANG Jingbo, PENG Chao, et al. BiseNet: Bilateral segmentation network for real-time semantic segmentation[C]//Proceedings of the European Conference on Computer Vision(ECCV). Cham:Springer,2018: 325-341.

[20]LI Hanchao, XIONG Pengfei, FAN Haoqiang, et al. DFANet: Deep feature aggregation for real-time semantic segmentation [C]//Proceedings of 2019 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 9522-9531.

[21]HONG Yuanduo, PAN Huihui, SUN Weichao, et al. Deep Dual-resolution networks for real-time and accurate semantic segmentation of road scenes[J]. arXiv preprint arXiv, 2101.06085, 2021.

[22]FU Jun, LIU Jing, TIAN Haijie, et al. Dual attention network for scene segmentation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2019: 3146-3154.

[23]YUAN Yuhui, HUANG Lang, GUO Jianyuan, et al. OCNet: Object context network for scene parsing [J]. arXiv preprint arXiv,1809.00916, 2018.

[24]WANG Haochen, JIANG Xiaodong, REN Haibin, et al. Swiftnet: Real-time video object segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2021: 1296-1305.

[25]CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE,2016:3213-3223.

[26]ZHAO Hengshuang, QI Xiaojuan, SHEN Xiaoyong, et al. ICNet for real-time semantic segmentation on high-resolution images [C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham:Springer,2018: 405-420.