计算机科学
COMPUTER SCIENCE

基于边缘优化和全局建模的多路径语义分割

陈乔松, 张羽, 蒲柳, 谭冲冲, 邓欣, 王进, 孙开伟, 欧阳卫华

引用本文

陈乔松, 张羽, 蒲柳, 谭冲冲, 邓欣, 王进, 孙开伟, 欧阳卫华.基于边缘优化和全局建模的多路径语义分割[J]. 计算机科学, 2023, 50(6A): 220700137-7.

CHEN Qiaosong, ZHANG Yu, PU Liu, TAN Chongchong, DENG Xin, WANG Jin, SUN Kaiwei, OUYANG Weihua. Multi-path Semantic Segmentation Based on Edge Optimization and Global Modeling[J]. Computer Science, 2023, 50(6A): 220700137-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700137

基于边缘优化和全局建模的多路径语义分割

陈乔松[1] 张 羽[1] 蒲 柳[1] 谭冲冲[2] 邓 欣[1] 王 进[1] 孙开伟[1] 欧阳卫华[1]
1 重庆邮电大学计算机科学与技术学院数据工程与可视计算重点实验室 重庆 400065
2 重庆邮电大学自动化学院 重庆 400065
(chenqs@cqupt.edu.cn)

摘 要 目前的语义分割卷积网络中,空间信息和细节信息随着卷积层的加深而逐渐丢失,造成物体边界和细小物体的分割效果不准确。同时,卷积的局部特征能力限制了网络获取有效的全局建模能力,造成物体内部分割混淆。针对这些问题,文中设计了基于边缘优化和全局建模的多路径语义分割算法。该算法提出了多路径邻近错位融合的网络,4条不同的分辨率路径邻近之间细节信息融会,高分辨率路径尾部与低分辨率路径首部间的语义信息交融,以此减少空间信息和细节信息的丢失。文中提出了自适应边缘特征模块得到边缘特征,融入网络中间层和深度监督层,增强边缘特征的表达能力和细小物体的分割效果,提出了Transformer全局特征模块,采用不同卷积进行下采样操作,缩短自注意力序列的长度,再融合通道信息与自注意力信息,从而获取有效的高层语义的全局信息。实验结果表明,在CamVid测试集和Cityscapes验证集上mIoU值分别达到76.2%和79.1%。

关键词:语义分割;多路径;边缘优化;深度监督;全局建模
中图法分类号 TP391.4

Multi-path Semantic Segmentation Based on Edge Optimization and Global Modeling

CHEN Qiaosong[1],ZHANG Yu[1],PU Liu[1],TAN Chongchong[2],DENG Xin[1],WANG Jin[1],SUN Kaiwei[1] and OUYANG Weihua[1]
1 Key Laboratory of Data Engineering and Visual Computing,School of Computer Science and Technology,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
2 School of Automation,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract In the current semantic segmentation convolutional network,the spatial and detail information is gradually lost with the deepening of the convolutional layer,resulting in inaccurate segmentation of boundary parts and small objects. Meanwhile,the local feature capability of convolution restricts the network's ability to obtain effective global modeling,resulting in confusion of internal segmentation of objects. Aiming at these problems,a multi-path semantic segmentation algorithm based on edge optimization and global modeling is designed. The algorithm proposes a multi-path adjacent dislocation fusion network. Four branches of different resolutions are interlaced and fused adjacently. In order to reduce the loss of spatial information and detail information,the detail information between the adjacent four different resolution paths is fused,and the semantic information is fused between the tail of the high-resolution path and the header of the low-resolution path. The adaptive edge feature module is proposed to obtain edge features which are integrated into the middle layer and depth supervision layer of the network to enhance the expressive ability of edge features and the segmentation effect of small objects. The Transformer global feature module is proposed,which uses different convolutions for downsampling operations to reduce the length of self-attention sequences and fuse channel information and self-attention information to obtain effective high-level semantic global information. Experimental results show that the mIoU value on the CamVid test set reaches 76.2%,and the mIoU value on the Cityscapes validation set reaches 79.1%.

Keywords Semantic segmentation,Multi-path,Edge optimization,Deep supervision,Global modeling

1 概述

作为像素级别的多分类任务,语义分割在计算机视觉领域有着无比重要的地位。语义分割旨在通过对不同像素点进行类别预测,来将不同的目标从背景中分割出来。在实际生活中,语义分割发挥着不可忽视的作用,特别是在医学影像分析诊断、自动驾驶、卫星图像处理和环境分析等领域应用广泛。

语义分割在对不同目标进行分割提取时,图像的边界细节信息和场景语义信息都不可忽视。在早期的语义分割相关研究工作中,主要采用基于阈值的分割算法[1]、基于边缘的分割算法[2]、基于区域的分割算法以及图割法[3]等。传统的提取边缘轮廓的方法对图像的分割有不错的促进作用。其中,Canny边缘检测法[4]较为突出,能在噪声抑制和检测之间取得较好的平衡。这些传统的图像分割算法,以人工设计的方式提取图像颜色、纹理和边缘等浅层特征,再针对单个目标进行分割操作。在这一过程中,许多参数的设置不能自适应

基金项目:国家重点研发计划(2022YFE0101000)
This work was supported by the National Key Research and Development Program of China(2022YFE0101000).
通信作者:张羽(1262912010@qq.com)

每张图片，不能灵活地处理不同的分割任务。

近年，替代传统分割算法的深度神经网络语义分割方法不断地快速发展。Long 等提出的全卷积网络(Fully Convolutional Networks，FCN)[5]取得了良好的分割效果。在此基础上，基于全卷积神经网络的各种分割技术不断发展[6]，其语义分割网络框架主要分为两种。一种是编码器-解码器的网络框架，其编码器部分不断下采样以获取更高的语义信息，其解码器部分不断上采样来恢复空间维度和细节信息。其中，UNet[7]，DeconvNet[8]，SegNet[9]均采用该框架。但因其解码器部分的卷积层简单，丢失了细节信息，上采样得到的结果较为粗糙。即使将同一层级的编码特征与解码特征进行融合以弥补信息，但缺乏全局的上下文信息提取的能力，容易将复杂场景中的物体内部像素点错误分类。另一种是特征融合的语义分割网络架构，其利用骨干网络提取各层特征图信息，结合上下文信息提取模块进行分割。但这种架构需要有效的全局上下文信息作为指导，才可避免像素点错分的情况。其中，文献[10]提出了金字塔池化模块，获得了多尺度的特征信息。基于此，大量学者多采用金字塔的池化方式聚合上下文信息，但因普通的卷积感受野有限，全局信息的获取受到限制。为了增大感受野，文献[11]采用空洞卷积代替传统的普通卷积。而这种方式丢失了局部信息，存在网格效应问题。此外，为了获取更有效的全局信息，各种注意力机制涌现。文献[12-15]均采用类似的注意力机制，利用分配权重的方式，突出重要信息，抑制不重要的信息，探究通道维度上的全局信息。但卷积网络具有无法充分获取全局依赖的局限性，这些注意力机制均是通过卷积和池化操作来获取全局信息，无法避免这种局限性。而 Transformer[16]具有天然的全局建模能力，是一种基于自注意力机制的全新网络架构。SegFormer[17]网络就是其代表之一，采用若干个 Transformer 模块堆叠获取图像特征信息，由于其自注意力和全连接层的计算量过大，因此不易应用于工程实践。

在目前的卷积神经网络中，随着卷积层的加深，图像的语义信息逐渐丰富，图像的细节信息与空间信息不断地丢失。这种信息的丢失在后续的上采样操作中不能得到有效的恢复，从而影响分割效果。同时，大部分网络将图像的形状、纹理和颜色等信息均放在一个网络中进行提取，而边缘部分和细小物体的像素点所占比例较小，导致边界和细小物体的细节信息丢失更为严重，从而使得边缘分割粗糙和细小物体分割不准确。此外，卷积网络感受野的局限性不利于获取有效的全局信息，物体的内部分割效果有待提高。基于上述内容，本文提出了一种基于边缘优化和全局建模的多路径语义分割网络(Multi-path Interleaved Network，MINet)。本文的主要贡献如下：

(1)本文提出了多路径交织网络(Multi-path Interleaved Network Base，MINet-Base)作为本文的骨干网络，低分辨率分支首阶段信息仅与相邻的高分辨分支尾阶段信息错位融合，以此降低开销并增强高分辨率分支的语义信息和网络的特征提取能力。保持 4 条不同分辨率分支两两交错融合，将高分辨率分支与相邻的低分辨率分支交错融合，以减少空间和细节信息的损失。

(2)本文提出了自适应边缘特征模块(Adaptive Edge Feature Module，AEFM)，利用 Canny 自适应算法提取边缘信息，自动计算不同图像的高低阈值。再将边缘信息与网络中的高分辨分支融合，增加网络的边缘特征表达能力。同时，将边缘信息与骨干网络中间输出的特征信息融合，将其作为深度监督辅助损失函数的输入，使得网络对边缘部分的关注进一步得到增强，从而提升边缘分割的效果。

(3)本文提出了 Transformer 全局特征模块(Transformer Global Feature Module，TGFM)，利用 Transformer 的全局建模能力，获取高层语义分支的全局上下文信息，以有效地提升卷积网络的全局建模能力。针对其自注意力计算部分，采用不同卷积核在不同通道上进行下采样操作，融合通道信息与自注意力信息，使得其既保留空间信息，也降低模块的开销。将全局信息与高分辨率分支信息融合，使得局部特征的全局感知能力和全局表示的局部细节能力相结合，以此提升图像中物体内部分割的效果。

2 基于边缘优化和全局建模的多路径语义分割方法

本文提出多路径交织骨干网络作为基础的特征提取网络，利用 AEFM 模块提取边缘轮廓并将其融入到多路径交织骨干网络中，将边缘信息与网络中间层输出作为深度监督的损失函数的输入，优化边缘分割效果。利用 TGFM 模块提取全局建模信息，捕获像素点之间的依赖关系，提高分割精度。本文提出的网络模型的总体结构如图 1 所示。

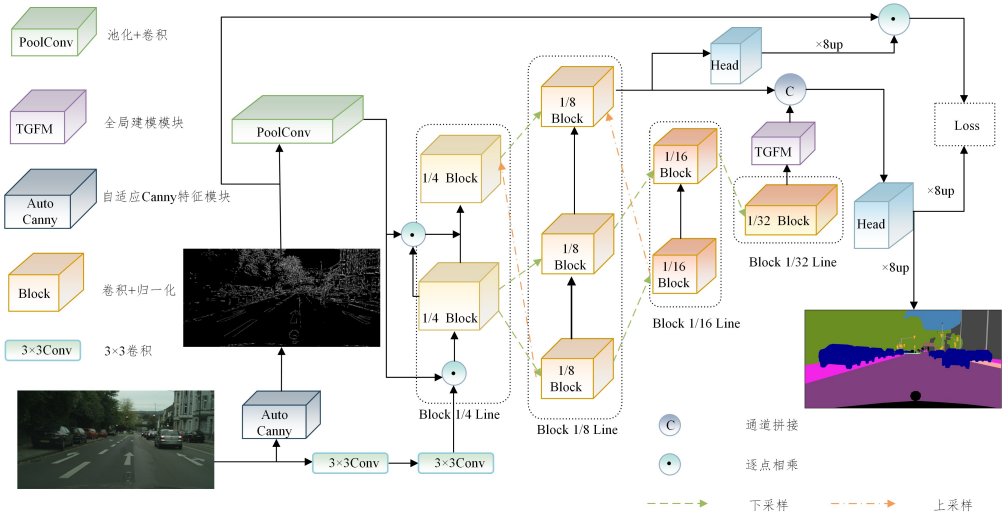


图 1 MINet 网络模型的整体架构

Fig. 1 Overall structure of MINet

2.1 多路径交织骨干网络(MINet-Base)

本文提出多路径交织骨干网络作为基础网络，考虑到高分辨率输入图像的计算量较大，将特征图的输出分辨率降为原始图像分辨率的 1/4。将始终保持为原始图像分辨率的 1/4，1/8，1/16 和 1/32 的 4 个分辨率分支，对应图 1 中的 1/4 Block Line，1/8 Block Line，1/16 Block Line 和 1/32 Block Line，相邻分辨率分支采取两两交织下采样、错位融合上采样的操作，使得不同层级的特征信息得到充分交互，以消除不同层的语义信息的差异性，减少空间信息和细节信息的丢失。每个分支中的 Block 内部包含若干个卷积层，每个分支上保持该分支的分辨率，Block 内部不进行任何的下采样操作。

其中交织融合的上采样和下采样操作如图 1 所示。两两交织下采样采用 3×3 的卷积调整通道和特征图尺寸，将其与原分支信息进行逐点求和，低层的空间和细节信息与相邻高层语义信息之间交织融合，每层的高层语义信息都得到基于低层信息的补充，以便于降低高层分支的空间和细节信息的丢失。补充的空间和细节信息从低到高流动，每层都与累积之前低层特征信息相融合，逐层交织至最顶层。而错位融合上采样操作采用 1×1 的卷积调整通道，同时利用双线性插值的方式将低分辨率特征图尺寸放大到高分辨率特征图的尺寸，将两者进行逐点求和，消除不同阶层的特征信息的差异性。采用这种方式使得高层特征信息与低层特征信息充分融合，以便于空间和细节信息可以更好地在网络中的每层进行传递与累积。同时，考虑模型的计算量，仅采用低分辨率分支的首部阶段 Block 信息与两两相邻的高分辨率分支的尾部阶段 Block 信息进行上采样融合，实现在降低开销的同时，也使得前阶段的高分辨率分支中的空间和细节信息充分融入到后阶段的高语义信息分支中。本文通过这两种不同的新型融合的采样操作，减少了空间信息和细节信息的丢失，有效地提取了图像的语义信息，从而提升了分割效果。

2.2 自适应边缘特征模块(AEFM)

在深度神经网络中，随着卷积和池化的加深会造成边界细节信息的丢失，影响目标边界像素的分类。因此，加强边缘细节信息的提取是有必要的。Canny 边缘检测法采用非极大值抑制将像素宽的边缘缩减成窄边缘，最后通过设置双阈值得到边缘轮廓。而自适应 Canny 算法针对不同的图片自动计算出高低阈值，以提取有效的边缘轮廓。因其根据图像的像素值的中位数计算高低阈值，而非固定值，所以可以很好地与神经网络相结合。将低阈值表示为 lower，高阈值表示为 upper，计算过程如式(1)和式(2)所示：

$$lower=\max(0,median*0.66) \tag{1}$$

$$upper=\min(median*1.33,255) \tag{2}$$

其中，$median$ 为一张图像中的所有像素值的中位数，max 为取区间内的最大值操作，min 为取区间内的最小值操作，系数是根据以往研究者经验所设置的，在该系数的情况下，提取边缘轮廓的效果较好。

本文提出自适应边缘特征模块，采用了自适应 Canny 算法来获取边缘轮廓信息。如图 2 所示，经过自适应 Canny 算法得到边缘轮廓信息 X，经过平均池化操作获取全局的边界信息，再将其经过 1×1 的卷积调整尺寸，与 MINet-Base 中分辨率为原始分辨率的 1/8 的分支融合，进一步增加网络的细节信息。将 X 经过 Sigmoid 函数激活，生成边界注意力特征信息 Y，具体过程如式(3)所示：

$$Y=Sigmoid(Conv(AvgPool(X))) \tag{3}$$

其中，$Conv$ 代表采用卷积核为 1×1 的卷积操作，$AvgPool$ 代表采用卷积核为 3×3 的平均池化操作。

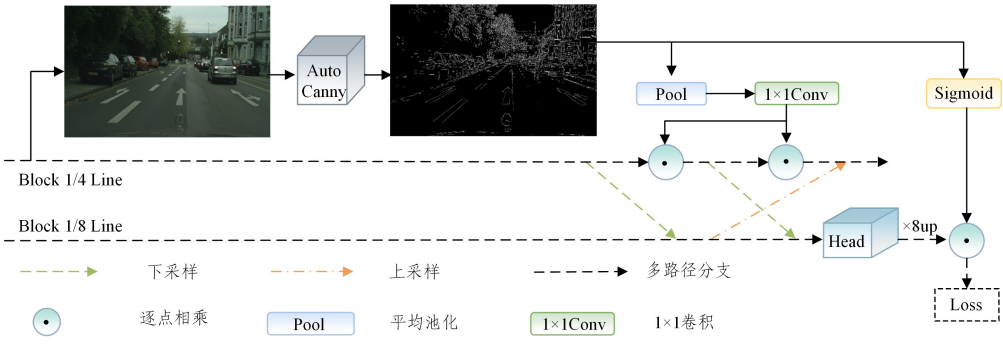


图 2 自适应边缘特征模块

Fig. 2 Adaptive edge feature module

在语义分割任务中，训练阶段的深度监督可以在早期阶段学习特征的稳健性，可有效地避免“梯度消失”现象。同时，引入深度监督学习可以改善大型和深层次网络中损失函数的收敛能力，从而提升性能。本文将边界注意力特征信息与分辨率为原图分辨率的 1/8 的特征信息进行逐点相乘，更新边界的权值信息，提升边界像素点分类的准确性。其结果采用双线性插值方式上采样恢复到原图的尺寸，以此作为辅助损失函数的输入来监督网络的训练，增加网络对边缘分割的关注程度。采用最终的损失函数度量预测与实际的差异程度，根据以往研究者的经验将辅助损失函数的权重设置为 0.4，并且辅助损失函数在测试阶段被弃用，最终损失由式(4)表示：

$$L_1=L+\alpha L_\alpha \tag{4}$$

其中，L_1 代表最终损失，L 代表主要的损失，L_α 代表辅助损失，α 表示辅助损失的权重。

2.3 Transformer 全局特征模块(TGFM)

长距离视觉元素之间的关系对于语义分割任务尤为重要。图像中大量的像素点之间相互联系从而构成不同的物体，在对一个像素点进行预测分类时，不能孤立地去看待，应当充分考虑它周围的像素点的信息。由于卷积网络有限的感受野，因此无法直接获得特征图的全局信息。若直接扩大感受野，则需要更密集且具有破坏性的池化操作，不利于减少空间信息和细节信息的丢失。

本文提出 Transformer 全局特征模块，将通道信息与自注意力信息融合，缩短自注意力机制输入的序列长度，将自注意力机制的全局建模能力和卷积的局部提取能力相结合，得到有效的全局信息。如图 3 所示，多路径交织骨干网络的

高语义分支输出经过 LayerNorm 归一化操作得到高语义信息特征图 A，其维度为 $B\times C\times H\times W$，其中 B,C,H,W 为批量大小、通道维度、特征图的长和宽。将其维度变为 $B\times head\times L\times (C/head)$ 得到 Q。其中 $L=H\times W$，$head$ 为头部数量，本文设置为 8。将特征图 A 的通道一分为二，分别经过 1×1 的卷积和 3×3 的卷积进行两倍下采样操作，减少序列长度并保留空间信息。通过不同的卷积核大小对不同通道上的信息进行下采样操作，减少了后续自注意力机制的输入序列长度，也针对多维度上的特征通道进行重标定，得其维度为 $B\times C\times (H/2)\times (W/2)$ 的 $A3$。将 $A3$ 进行 Reshape 操作得到 V，其维度为 $B\times head\times (L/4)\times (C/head)$。此外，将 $A3$ 依次进行 Reshape 操作和 Transpose 操作得到 $\boldsymbol{K}$，其维度变为 $B\times head\times (C/head)\times (L/4)$。将 $Q,\boldsymbol{K},V$ 进行自注意力计算，得到自注意力信息，过程如式(5)所示：

$$Attention(Q,K,V)=SoftMax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V \tag{5}$$

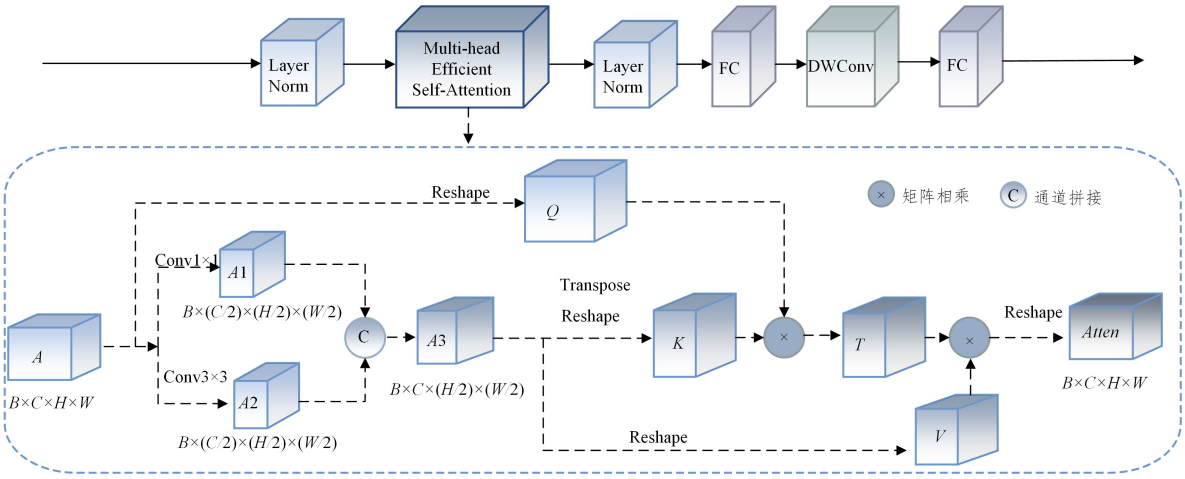


图 3 Transformer 全局特征模块

Fig. 3 Transformer global feature module

其中，$d=C/head$，$\boldsymbol{K}^{\mathrm{T}}$ 是 $\boldsymbol{K}$ 的转置向量。将自注意力信息经过 LayerNorm 归一化操作，再依次经过全连接层、深度可分离卷积层和全连接层，增加网络对复杂特征的学习能力。在两个全连接层中加入深度可分离卷积层，增加其非线性变换能力，以此充分获取 A 中像素点与像素点之间的依赖关系。

此模块的自注意力机制将通道一分为二，经过不同感受野获取空间信息，重新针对通道的特征信息进行筛选，缩短了自注意力机制的输入序列长度，降低了模块的开销。此模块融合了卷积提取的局部细节信息和自注意力机制获取的全局感知信息，得到了有效的全局上下文信息。

3 实验结果和分析

3.1 实验环境与评估指标

本文实验环境为 Ubuntu 系统、Python3.7、基于 PyTorch1.8 框架、GPU 版本为 NVIDIA GeForce GTX 1070 (8G)。针对验证实验，本文使用平均交并比(mean Intersection over Union，mIoU)反映语义分割的精度。针对消融实验，增加像素精度指标(Pixel Accuracy，PA)以充分反映模块的有效性。

本文网络按照 Block 之间下采样的卷积层个数的不同分为 MINet-S 网络和 MINet-M 网络，其中 MINet-S 网络中 1/4 Block Line，1/8 Block Line，1/16 Block Line 和 1/32 Block Line 的 Block 之间下采样的卷积层数分别为 2,2,2,2，而 MINet-M 网络中卷积层数分别为 2,4,6,2。另外，为更好地观测分割结果，针对 MINet-S 网络进行单尺度验证，针对 MINet-M 网络进行多尺度验证，尺度大小比例分别设置为 0.5，0.75，1.0，1.25，1.5，1.75。本文在数据集上的训练配置如表 1 所列。使用 OHEM 交叉熵损失函数衡量预测值与实际值的偏离程度，解决数据中类别不均衡的问题。

mIoU 是计算真实标签值和预测值两个集合的交集和并集之比，计算过程如式(6)所示。PA 指预测正确的像素个数占总像素个数的比例，计算式如式(7)所示：

$$mIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \tag{6}$$

$$PA=\frac{\sum_{i=0}^{k}p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}} \tag{7}$$

其中，$k+1$ 为图像中所有分割类别数目；P_{ij} 代表真实标签值为 i 而被预测为 j 的数量；P_{ji} 代表真实标签值为 j 而被预测为 i 的数量；P_{ii} 代表真实标签值为 i 而被预测为 i 的数量。

在训练过程中，均采用 poly 的学习策略，该策略的学习率衰减公式如式(8)所示：

$$init\_lr=\left(1-\frac{iter}{max\_iter}\right)^{power} \tag{8}$$

其中，$power$ 为衰减系数，$init\_lr$ 为初始学习率，$iter$ 为当前迭代次数，max_iter 为最大迭代次数。

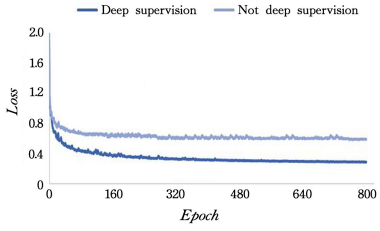


图 4 CamVid 数据集上的损失值曲线图

Fig. 4 Loss value on CamVid dataset

采用深度监督学习的方式，本文将网络中层的特征信息

和边界注意力特征信息融合，并将其作为辅助损失函数的输入来监督网络的训练。引入 OHEM 主损失函数平衡正负样本类别，采用交叉熵辅助损失函数监督网络训练，着重针对边缘分割优化。最终损失的计算过程如式(4)所示，其中 L 为 OHEM 主损失函数，L_a 为交叉熵辅助损失函数。图 4 为本文算法在 CamVid 数据集上的损失值曲线图，位于最下面的曲线为引入深度监督的损失值曲线，其损失值在经过 800 个 epoch 之后已处于平缓状态，从损失值曲线图可知引入深度监督方法使得损失函数收敛得更快更小，以达到提升分割的效果。

3.2 Camvid

在 CamVid 数据集上对本文方法进行对比验证。CamVid 数据集是对视频逐帧提取的道路与驾驶场景图像分割数据集，共有 701 张城市街道图片。以驾驶汽车的角度拍摄，增加了观察目标的数量和异质性。其中包含 421 张训练集图片、112 张验证集图片和 168 张测试集图片。原始图像分辨率大小为 720×960，共包含 11 个道路场景语义类别。本文将输入分辨率大小设置为 360×480，在训练前使用随机尺寸裁剪，以增强数据集的表现效果。数据集训练配置如表 1 所列。

表 1　数据集训练配置

Table 1　Training configuration on dataset

各项配置	CamVid	Cityscapes
数据集尺寸	360×480	1024×1024
Batch Size	4	4
优化器	Adamw	Adamw
学习率	0.0003	0.0003
损失函数	OHEM[18]交叉熵损失	OHEM[18]交叉熵损失
迭代轮数	800	600

本文采用 MINet-S 网络和 MINet-M 网络在 CamVid 数据集上进行验证，与各个网络进行对比实验的结果如表 2 所列。结果表明，本文算法的 MINet-S 网络的 mIoU 值比 SegNet 高 14.1%，即多路径交织的结构比 SegNet 的网络结构更好地保留了空间和细节信息。由表 2 可知，在参数量更少的情况下，本文 MINet-S 网络的 mIoU 值比 BiseNet 提升了 1.0%，MINet-M 网络的 mIoU 值也比 BiSeNet V2[19]网络提升了 3.8%，即本文采用的全局建模模块能更好地促进分割效果。综上所述，本文算法针对语义分割有较好的效果。

表 2　CamVid 测试集上的实验结果

Table 2　Results on CamVid test set

Model	mIoU/%	Resolution	Params
GCN[20](Resnet101)	54.6	512×512	43.0×10^6
RefineNet[21](Resnet101)	55.1	512×512	85.6×10^6
SegNet(VGG16)	55.6	360×480	29.7×10^6
DeepLabV2(Resnet101)	61.6	360×480	37.3×10^6
DFANet A[22]	64.7	720×960	7.8×10^6
BiseNet(Resnet18)	68.7	720×960	49.0×10^6
Ours(MINet-S)	**69.7**	**360×480**	$\mathbf{19.8\times10^6}$
BiSeNet V2	72.4	720×960	65.5×10^6
Ours(MINet-M)	**76.2**	**360×480**	$\mathbf{25.1\times10^6}$

3.3 Cityscapes

在 Cityscapes 数据集上对本文提方法进行对比验证。Cityscapes 作为城市环境中驾驶场景的图像，记录不同城市的不同场景和不同季节的街景图片。它提供了 5000 张精细标注的图像、2000 张粗略标注的图像和 30 类标注物体，其中 19 类用于语义分割任务。本文使用 2975 张训练集图片进行训练，使用 500 张验证集图片进行验证。在训练前使用随机尺寸裁剪和随机亮度变换，以此增强数据集。数据集训练配置如表 1 所列，考虑到实际的设备情况，MINet-M 网络在训练时，Batch Size 设置为 3，其余设置与表 1 一致。

本文采用 mIoU 作为验证实验的对比性能评估指标，MINet-S 网络和 MINet-M 网络在 Cityscapes 数据集的验证集上进行验证，与各个网络进行对比实验的结果如表 3 所列。实验结果表明，本文算法的 mIoU 值达到 79.1%，对比其他网络有更好的分割效果，其参数量也远远小于其他网络。

表 3　Cityscapes 验证集的实验结果

Table 3　Results on Cityscapes validation set

Model	mIoU/%	Resolution	Params
SegNet	56.0	640×360	29.5×10^6
Fast-SCNN[23]	68.6	2048×1024	1.1×10^6
HRNetV2-W18-v1[24]	70.3	512×1024	1.5×10^6
DFANet A	71.3	1024×1024	7.8×10^6
BiseNet(Resnet18)	74.8	2048×1024	49.0×10^6
MDEQ-small[25]	75.1	2048×1024	7.8×10^6
SwiftNetRN18[26]	75.5	2048×1024	11.8×10^6
RepMLPNet-D256[27]	76.2	2048×1024	78.5×10^6
PSPNet(Resnet 50)	76.5	1024×1024	46.6×10^6
Ours(MINet-S)	76.7	1024×1024	19.8×10^6
MDEQ-large	77.8	2048×1024	53.0×10^6
PSPNet(Resnet101)	78.4	2048×1024	65.9×10^6
DeepLabv3[28](Resnet101)	78.5	2048×1024	78.5×10^6
Ours(MINet-M)	79.1	1024×1024	25.1×10^6

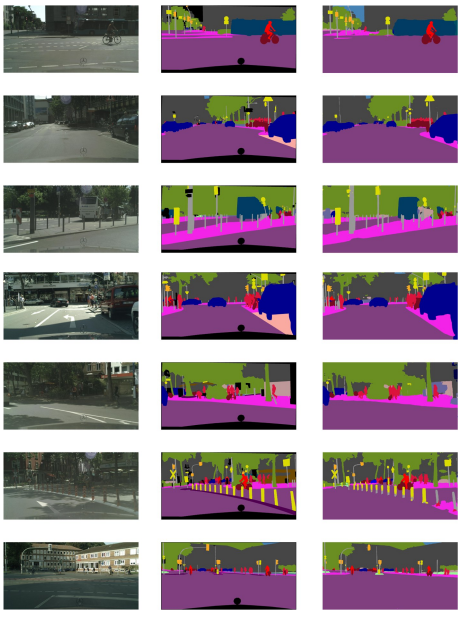
本文在 Cityscapes 数据集的验证集上针对每一类物体的 mIoU 值进行对比，如表 4 所列，表中的所有数值均为百分制。结果显示，本文方法针对人行道、建筑、墙壁、栅栏、杆、指示牌、卡车、公共汽车、火车和自行车类别的 mIoU 值分别达到 84.1%，92.4%，55.9%，58.0%，64.0%，78.1%，78.3%，83.5%，73.7%，76.0%，76.7%。由于 MINet 中多路径交织错位的特殊融合方式保留了大量的细节信息，AEFM 模块向网络增加了边缘信息，本文方法针对细长的电杆、栅栏等小目标分割较为细致，针对大目标边缘部分的分割效果也较为完整。由于 TGFM 模块提取了有效的全局上下文信息，本文的算法针对卡车、火车、公共汽车和建筑等大目标均有着不错的分割效果。

在 Cityscapes 数据集的验证集上的验证结果如图 5 所示，其中图 5(a)为原图，图 5(b)为真实标签图，图 5(c)为本文算法得到的分割图。结果显示，本文提出的 MINet 有效地保留了空间信息，分割效果较为完整，边缘分割粗糙和物体内部混淆问题也得到改善。由于 TGFM 模块得到了有效的全局信息，公交车、汽车内部分割完整准确，建筑物与道路的整体分割效果也较好。而 AEFM 模块促进了边缘的分割效果，针对细长的目标的分割也有促进作用，特别是电杆这种细长的物体的分割较为细致。此外，针对树木上的树叶轮廓边缘，其分割效果也比较清晰。

表 4 Cityscapes 验证集上的各类实验结果

Table 4 Results of each type of experiment on Cityscapes validation set

Model	road	s. walk	build	wall	fence	pole	t-light	t-sign	veg	terrain	sky	person	rider	car	truck	bus	train	motor	bike	mIoU
DABNet[29]	96.8	78.5	90.9	45.3	50.1	59.1	65.2	70.7	92.5	68.1	94.6	80.5	58.5	92.7	52.7	67.2	50.9	50.4	65.7	70.1
RefineNet	98.2	83.3	91.2	47.7	50.4	56.1	66.9	71.3	92.2	70.3	94.7	80.8	63.2	94.5	64.5	76.0	64.2	62.2	69.9	73.6
SwiftNetRN18	**98.3**	83.8	92.2	46.3	52.7	63.2	**70.5**	75.8	**93.1**	**70.3**	**95.4**	**84.0**	**64.5**	**95.2**	63.8	77.9	71.9	**61.5**	73.6	75.5
Ours(MINet-S)	98.0	**84.1**	**92.4**	**55.9**	**58.0**	**64.0**	69.1	**78.1**	92.4	64.7	94.6	81.5	59.4	94.9	**78.3**	**83.5**	**73.7**	59.4	**76.0**	**76.7**



(a) (b) (c)

图 5 Cityscapes 验证集上的分割效果

Fig. 5 Segmentation effect on Cityscapes validation set

3.4 消融对比实验

本文使用 MINet-S 网络在 Camvid 数据集和 Cityscapes 数据集上进行各个模块的验证实验。在 Camvid 数据集的验证集和测试集上验证各个模块的有效性，使用 mIoU 和 PA 指标进行衡量分割效果。如表 5 所列，MINet-Base 为本文提出的多路径交织骨干网络，MINet 为本文算法的网络。在验证集的单尺度测试下，MINet-Base 的 mIoU 值可以达到 67.3%，PA 值可以达到93.9%。在测试集的单尺度测试下，mIoU 值可以达到 68.4%，PA 值可以达到 93.7%。向骨干网络中加入 AEFM 模块，融入边缘信息，以此增加网络对边缘分割部分的关注，从而提升整体的分割效果。其 mIoU 值在验证集上提高了 0.2%，在测试集上提高了 0.6%。其 PA 值在验证集上提高了 0.2%，在测试集上提高了0.3%。在此基础上，加入 TGFM 模块，获取有效的全局上下文信息，以提高分割精度。其 mIoU 值在验证集上提高了0.9%，在测试集上提高了 0.7%。其 PA 值在验证集上提高了 0.1%，在测试集上提高了 0.1%。综上所述，针对分割效果，本文提出的各个模块均有促进的作用。

表 5 CamVid 验证集和测试集上各模块表现对比

Table 5 Performance comparison of each module on CamVid validation set and test set

(单位：%)

Model	mIoU		PA	
	Val	Test	Val	Test
MINet-Base	67.3	68.4	93.9	93.7
＋AEFM	67.5	69.0	94.1	94.0
＋TGFM	68.4	69.7	94.2	94.1

表 6 Cityscapes 验证集上各模块表现对比

Table 6 Performance comparison of each module on Cityscapes validation set

(单位：%)

Model	mIoU	PA
MINet-Base	73.3	95.6
＋AEFM	75.4	95.8
＋TGFM	76.7	95.9

在 Cityscapes 数据集的验证集上验证各个模块的有效性，采用单尺度测试的方式，使用 mIoU 指标和 PA 指标进行衡量分割效果。如表 6 所列，本文算法的骨干网络在单尺度测试下，其 mIoU 值可以达到 73.3%，PA 值可以达到 95.6%。向骨干网络中加入 AEFM 模块，增加了边缘信息的特征表达，提升了整体的分割效果。实验结果表明，加入 AEFM 模块的 MINet 的 mIoU 值提高了 2.1%，PA 值提高了 0.2%。在此基础上，加入 TGFM 模块来获取有效的全局上下文信息，提升目标内部分割准确度，从而使得整体分割效果更好。在单尺度测试下，其 mIoU 值提高了 1.3%，PA 值提高了 0.1%。实验结果证明，针对分割任务，本文提出的模块都有不错的促进效果。

结束语 为了解决目前卷积网络的局限性导致的边缘分割粗糙、细小物体分割不准确和物体内部混淆等问题，本文提出了一种基于边缘优化和全局建模的多路径语义分割网络 MINet。本文探索了一种新的交织错位融合的方式，有效地融合了高层细节信息与低层语义信息，减少了空间信息和细节信息的丢失。同时，本文提出的 AEFM 模块将自适应提取边缘算法与神经网络相结合，以有效地获取边界注意力特征信息。同时，将边缘信息和网络中间输出信息作为深度监督损失函数的输入，加强网络对边缘分割的关注，提升对边缘的分割效果。此外，本文提出的 TGFM 模块使用不同卷积核在不同通道上进行下采样，将重标定的通道信息与自注意力信息融合，既保留了空间信息，又降低了模型的计算量。将卷积与自注意力机制相融合，以获得有效的全局上下文信息。通过在 CamVid 数据集和 Cityscapes 数据集上进行实验，本文算法得到的 mIoU 值分别达到了 76.2%和 79.1%。但是本文算法的 Canny 计算量和引入的 Transformer 模块中的计算量有待进一步地减少。因此，探索更为轻量级的模型是未来的研究方向。

参 考 文 献

[1] ZHAN Z Y, AN Y J, CUI W C. Image Threshold Segmentation Algorithms and Comparative Research [J]. Information and Communication, 2017(4): 86-89.

[2] LIANG Z X, WANG X B, HE T, et al. Research and implementation of instance segmentation and edge optimization algorithms [J]. Journal of Graphics, 2020, 41(6): 939-946.

[3] ROTHER C, KOLMOGOROV V, BLAKE A. ""GrabCut" in-

teractive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics(TOG),2004,23(3):309-314.
[4] CANNY J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1986(6):679-698.
[5] LONG J,SHELHAMER E,DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer vision and Pattern Recognition. 2015:3431-3440.
[6] WANG Y R,CHEN Q L,WU J J. Research on Image Semantic Segmentation for Complex Environments[J]. Computer Science, 2019,46(9):36-46.
[7] RONNEBERGER O,FISCHER P,BROXT. U-net:Convolutional networks for biomedical image segmentation[C]// International Conference on Medical image computing and computer-assisted intervention. Cham:Springer,2015:234-241.
[8] NOH H,HONG S,HAN B. Learning deconvolution network for semantic segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:1520-1528.
[9] BADRINARAYANAN V,KENDALL A,CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(12):2481-2495.
[10] ZHAO H,SHI J,QI X,et al. Pyramid scene parsing network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:2881-2890.
[11] CHEN L C,ZHU Y,PAPANDREOU G,et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018:801-818.
[12] HU J,SHEN L,SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7132-7141.
[13] YU C,WANG J,PENGC,et al. Bisenet:Bilateral segmentation network for real-time semantic segmentation[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 325-341.
[14] 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. 2019:3146-3154.
[15] 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. 2019: 603-612.
[16] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Advances in Neural Information Processing systems. 2017:5998-6008.
[17] 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.
[18] SHRIVASTAVA A,GUPTA A,GIRSHICK R. Training region-based object detectors with online hard example mining [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:761-769.
[19] YU C,GAO C,WANG J,et al. Bisenet v2:Bilateral network with guided aggregation for real-time semantic segmentation [J]. International Journal of Computer Vision,2021,129(11): 3051-3068.
[20] PENG C,ZHANG X,YU G,et al. Large kernel matters--improve semantic segmentation by global convolutional network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:4353-4361.
[21] LIN G,MILAN A,SHEN C,et al. Refinenet:Multi-path refinement networks for high-resolution semantic segmentation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:1925-1934.
[22] LI H,XIONG P,FAN H,et al. Dfanet:Deep feature aggregation for real-time semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:9522-9531.
[23] POUDEL R P K,LIWICKI S,CIPOLLA R. Fast-scnn:Fast semantic segmentation network[J]. arXiv:1902. 04502,2019.
[24] SUN K,ZHAO Y,JIANG B,et al. High-resolution representation for learning pixels and regions[J]. arXiv: 1904. 04514, 2019.
[25] BAI S,KOLTUN V,KOLTER J Z. Multiscale deep equilibrium models[J]. Advances in Neural Information Processing Systems,2020,33:5238-5250.
[26] ORSIC M,KRESO I,BEVANDICP,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. 2019: 12607-12616.
[27] DING X,CHEN H,ZHANG X,et al. Repmlpnet:Hierarchical vision mlp with re-parameterized locality[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:578-587.
[28] YURTKULU S C,ŞAHIN Y H,UNAL G. Semantic segmentation with extended DeepLabv3 architecture[C]// 2019 27th Signal Processing and Communications Applications Conference (SIU). IEEE,2019:1-4.
[29] LI G,YUN I,KIM J,et al. Dabnet:Depth-wise asymmetric bottleneck for real-time semantic segmentation[J]. arXiv: 1907. 11357,2019.

CHEN Qiaosong, born in 1978, Ph. D, associate professor, is a member of China Computer Federation. His main research interests include blockchain, data mining and machine vision.

ZHANG Yu, born in 1998, postgraduate. Her main research interest is machine vision.