智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

结合全局注意力机制的实时语义分割网络

李涛, 高志刚, 管晟媛, 徐久成, 马媛媛

引用本文:
李涛,高志刚,管晟媛,徐久成,马媛媛. 结合全局注意力机制的实时语义分割网络[J]. 智能系统学报, 2023, 18(2): 282–292.
LI Tao,GAO Zhigang,GUAN Shengyuan,XU Jiucheng,MA Yuanyuan. Global attention mechanism with real–time semantic segmentation network[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(2): 282–292.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202208027

DOI: 10.11992/tis.202208027
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20220909.1714.002.html

结合全局注意力机制的实时语义分割网络

李涛[1,2], 高志刚[3], 管晟媛[4], 徐久成[1,2], 马媛媛[1]

（1. 河南师范大学 计算机与信息工程学院, 河南 新乡 453007; 2. 智慧商务与物联网技术河南省工程实验室, 河南 新乡 453007; 3. 河南师范大学 软件学院, 河南 新乡 453007; 4. 中国人民公安大学 国家安全学院, 北京 100038）

摘　要: 针对轻量化网络结构从特征图提取有效语义信息不足, 以及语义信息与空间细节信息融合模块设计不合理而导致分割精度降低的问题, 本文提出一种结合全局注意力机制的实时语义分割网络 (global attention mechanism with real time semantic segmentation network ,GaSeNet)。首先在双分支结构的语义分支中引入全局注意力机制, 在通道与空间两个维度引导卷积神经网来关注与分割任务相关的语义类别, 以提取更多有效语义信息; 其次在空间细节分支设计混合空洞卷积块, 在卷积核大小不变的情况下扩大感受野, 以获取更多全局空间细节信息, 弥补关键特征信息损失。然后重新设计特征融合模块, 引入深度聚合金塔池化, 将不同尺度的特征图深度融合, 从而提高网络的语义分割性能。最后将所提出的方法在 CamVid 数据集和 Vaihingen 数据集上进行实验, 通过与最新的语义分割方法对比分析可知, GaSeNet 在分割精度上分别提高了 4.29%、16.06%, 实验结果验证了本文方法处理实时语义分割问题的有效性。

关键词: 实时语义分割; 全局注意力机制; 多尺度特征融合; 混合空洞卷积; 卷积神经网络; 金字塔池化; 感受野; 特征提取

中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)02−0282−11

中文引用格式: 李涛, 高志刚, 管晟媛, 等. 结合全局注意力机制的实时语义分割网络 [J]. 智能系统学报, 2023, 18(2): 282–292.

英文引用格式: LI Tao, GAO Zhigang, GUAN Shengyuan, et al. Global attention mechanism with real-time semantic segmentation network[J]. CAAI transactions on intelligent systems, 2023, 18(2): 282–292.

Global attention mechanism with real-time semantic segmentation network

LI Tao[1,2], GAO Zhigang[3], GUAN Shengyuan[4], XU Jiucheng[1,2], MA Yuanyuan[1]

(1. College of Computer and Information Engineering, He'nan Normal University, Xinxiang 453007, China; 2. Engineering Lab of He'nan Province for Intelligence Business & Internet of Things, Xinxiang 453007, China; 3. College of Software, He'nan Normal University, Xinxiang 453007, China; 4. National Security Academy, People's Public Security University of China, Beijing 100038, China)

Abstract: The lightweight network structure cannot sufficiently extract effective semantic information from feature maps, and the unreasonable design of the semantic information and spatial detail information fusion block leads to a decrease in segmentation accuracy. To address these problems, a global attention mechanism with a real-time semantic segmentation network (GaSeNet) is proposed in the paper. First, a global attention mechanism is introduced into the semantic branch of the dual-branch structure. The convolutional neural network is then guided in the two dimensions of channel and space to focus on the semantic categories related to the segmentation task to extract remarkably effective semantic information. Second, a mixed hole convolution block is designed in the spatial detail branch, and the receptive field is enlarged while maintaining the size of the convolution kernel to obtain additional global spatial detail information and compensate for the loss of key feature information. The feature fusion module is then redesigned, and the deep aggregation pyramid pooling module is introduced to fuse feature maps of different scales comprehensively, thereby improving the semantic segmentation performance of the network. Finally, the proposed method is tested on CamVid and Vaihingen datasets. Compared with the latest semantic segmentation algorithm, GaSeNet improves the segmentation accuracy by 4.29% and 16.06%. Experimental results verify the effectiveness of this method in dealing with real-time semantic segmentation problems.

Keywords: real-time semantic segmentation; global attention mechanism; multiscale feature fusion; hybrid dilated convolution; convolutional neural network; pyramid pooling; receptive field; feature extraction

语义分割是计算机视觉中一个重要的任务, 其主要思想是对图像中的每一个像素点做语义类别预测, 进而实现图像中不同的语义类别分割。语义分割在医疗影像分割、无人驾驶、遥感图像

收稿日期: 2022−08−19.　**网络出版日期:** 2022−09−13.
基金项目: 国家自然科学基金项目(61976082, 62002103); 河南省高等学校重点科研项目(22B520013); 河南省科技攻关计划项目(222102210169).
通信作者: 李涛. E-mail: litao0116@163.com.

分割、场景理解分析等方面被广泛应用。传统机器学习分割方法主要包括基于聚类的分割、基于边缘检测的分割、基于图的分割、基于区域的分割和基于阈值的分割 5 大类。上述方法大多是根据图像的颜色、纹理特征、灰度值等底层特征信息, 把像素值相近的区域划分为一个类别, 并标记为相同的语义标签, 从而实现图像分割。虽然传统机器学习分割方法可以完成图像分割任务, 但是存在分割精度低, 无法划分多语义类别, 且需要人工调参等缺陷, 无法满足实际情况中高精度多语义类别的需求。随着深度学习的快速发展以及高性能 GPU 的应用, 基于神经网络的方法为求解语义分割的问题提供了解决方案。深度学习中的卷积神经网络从海量图像数据集中学习到多层次特征图像, 然后将特征图像映射为语义信息, 并把不同的语义信息标记为不同的语义标签, 进一步改善了语义分割效果。

近年来, 国内外学者针对语义分割问题进行研究, 并取得较好的成果。例如 Long 等[1](2015) 提出全卷积神经网络 (fully convolutional network, FCN), 使用反卷积层替换 VGG16[2] 卷积网络中的全连接层, 以全卷积的方式实现了端到端的图像语义分割。在全卷积网络的基础上, Ronneberger 等[3](2015) 提出 U-Net 网络结构模型, 该模型采用 U 型对称结构, 并将编码器提取的特征图与解码器提取的特征图进行融合。虽然 U-Net 提高了分割精度, 但其对称结构计算开销较大, 且未充分提取关键空间信息。为提取更多空间信息, Yu 等[4](2016) 提出空洞卷积网络结构, 其主要思想是在在卷积核元素之间填充空值, 以扩大感受野, 来捕获更多的上下文信息。虽然空洞卷积在语义分割任务上取得较好结果, 但在小尺度目标分割任务上表现较差。为适应多尺度目标分割任务, Chen 等[5](2017) 提出 DeepLabV3 网络结构, 该结构引入级联空洞卷积来提取多尺度特征图, 并采用金字塔池化将多尺度特征图进行特征融合。DeepLabV3 虽然能提高图像分割的精度, 但仍存在丢失空间信息的问题。之后, Wang 等[6](2018) 提出混合空洞卷积 (hybrid dilated convolution) 网络结构, 该网络结构解决了空洞卷积出现的网格现象, 并提升了网络分割多尺度目标能力。然而, 在现实生活中不仅要达到高精度的目的, 而且要满足一定的实时性, 如车辆自动驾驶和移动手机设备等领域对实时性与精度要求较高。针对该问题, Zhao 等[7](2018) 提出了 ICNet 网络结构, 该网络结构以级联多尺度图像作为输入, 以平衡级联特征融合模块中速度与精度。例如对于 1024 像素×2048 像素的高分辨率输入, 其分割精度 MIoU 达到 70.6%, 且分割速度达到 30 f/s。虽然 ICNet 满足了一定的实时性要求, 但是级联结构引入较多的参数使得结构不够轻量化。Li 等[8](2019) 提出 DFANeT 网络结构, 采用相互连接的编码流来提取特征, 通过编码周期来融合特征图, 减少了参数规模, 同时保持了分割速度与分割精度的平衡性。在最新的实时语义分割网络研究中, 研究者提出了双分支的网络结构, 例如 Yu 等[9](2021) 提出的 BiSeNetV2 网络结构, 该网络结构可以提取高级语义特征和空间细节信息, 在满足网络结构轻量化与实时性同时提高了分割精度。Hong 等[10](2021) 提出主干–分支双分辨率网络 DDRNet(deep dual-resolution networks) 网络结构, 该网络结构的主干通过多次下采样提取高层语义信息, 而高分辨率分支通过卷积来维持特征图分辨率, 该结构大幅度降低了网络参数规模, 并进一步提高了网络实时性和分割精度。

基于上述分析可知, 实时性语义分割研究已经取得了一定的成果, 但是仍存一些不足, 具体如下: 1) 为保证网络具有较好实时性, 一般采用精简的轻量化网络结构, 但会出现网络结构提取特征信息能力减弱的问题。2) 采用多分支的实时语义分割网络, 各分支提取到的不同特征图进行特征融合时, 特征融合策略设计不够合理导致分割精度不佳。

为此, 本文提出了结合全局注意力机制的实时语义分割网络 (global attention mechanism with real time semantic segmentation network, GaSeNet)。所提的网络结构创新如下: 1) 在 BiSeNetV2 语义分支引入了全局注意力机制 (global attention mechanism, GAM), 采用全局注意力模块来处理 BiSeNetV2 语义分支输出的特征图, 以增强在通道和空间维度上的特征交互, 进而提升语义分支挖掘特征信息的能力。2) 在 BiSeNetV2 空间细节分支中引入改进的混合空洞卷积, 不仅可以扩大感受野来获得更多细节信息, 还能捕获多尺度的特征信息, 进一步增强空间细节分支的特征信息提取能力。3) 在特征融合过程中引入深度聚合金塔池化模块 (deep aggregation pyramid pooling module, DAPPM), 以解决多尺度特征图融合不充分的问题, 并通过减少原双向交互融合结构中的通道数来降低计算量, 进而改善网络的分割效率。

1　GaSeNet 网络结构

为了提取更加丰富的有效语义信息与提高不同尺度特征图的融合能力，本文提出了 GaSeNet 网络结构，如图 1 所示，主要由语义分支、空间细节分支、特征融合模块 3 部分组成。其中，语义分支的原卷积结构不变，语义分支尾部引入全局注意力机制[11]来提取更多有效的高层语义信息。空间细节分支采用混合空洞卷积来增大感受野，这有助于提取多尺度低层次空间细节信息[12]。特征融合模块将语义分支的高层语义信息和空间细节分支的低层空间细节信息进行融合。因为在特征融合的过程中充分利用了不同尺度的特征图[13]，所以有助于提高图像分割的精度。

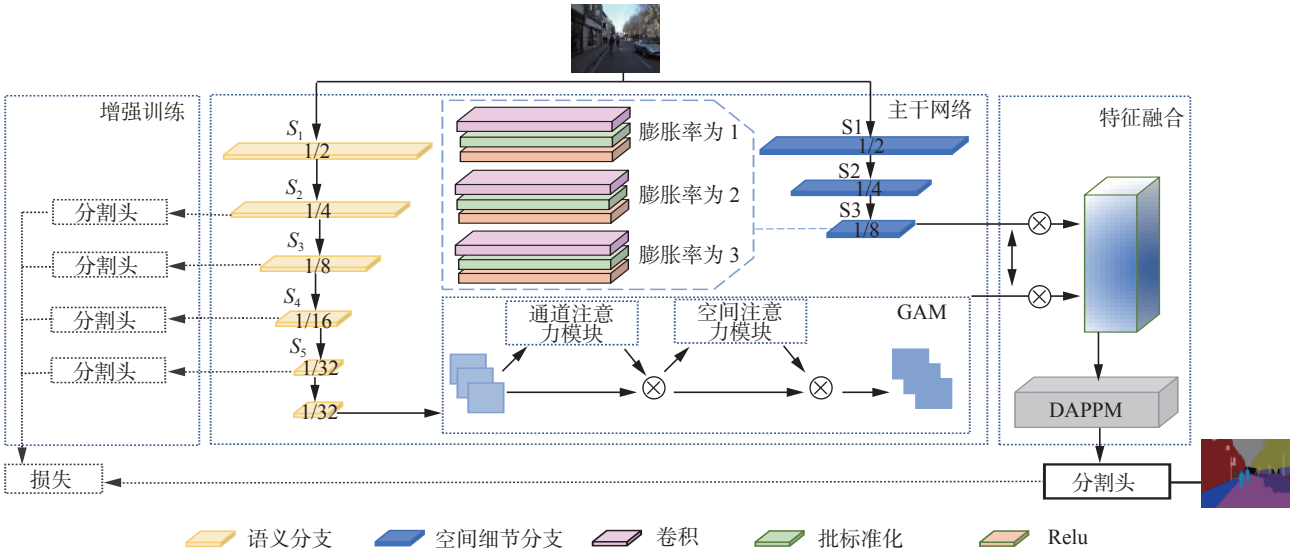


图 1　GaSeNet 网络结构

Fig. 1　GaSeNet network structure

1.1　全局注意力机制

注意力机制本质是对特征图的加权处理，强化特征图与语义信息的映射关系。由于卷积下采样造成的细节信息丢失和数据集噪声影响，预测结果可能出现边界分割模糊与小尺度语义信息预测错误[14]。为提高有效特征信息的利用率和增强网络全局场景理解能力，本文在语义分支引入全局注意力机制，其由通道注意力机制和卷积空间注意力机制组成。全局注意力结构如图 2 所示。

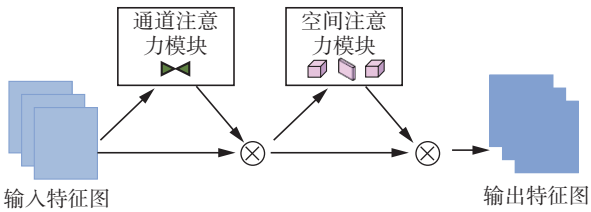


图 2　全局注意力机制

Fig. 2　Global attention mechanism

语义分支输出的特征图在全局注意力机制的处理过程如式 (1) 和 (2) 所示，式中$\boldsymbol{F}_1$为输入的语义分支特征图，$\boldsymbol{F}_2$为通道注意力子模块输出特征图，$\boldsymbol{F}_3$为空间注意力子模块的特征图输出，M_C与M_S分别代表通道注意力子模块和空间注意力子模块。

$$\boldsymbol{F}_2 = M_C(\boldsymbol{F}_1)\otimes\boldsymbol{F}_1 \tag{1}$$

$$\boldsymbol{F}_3 = M_S(\boldsymbol{F}_2)\otimes\boldsymbol{F}_2 \tag{2}$$

1）通道注意力子模块。

通道注意力机制主要思想是关注特征图有意义的通道，并且抑制无关的通道，从而实现卷积输出通道中特征图加权。卷积层输出通道的特征图映射为对应的语义信息，但不同语义信息量权重相同，不利于特征表达，通道注意力机制赋予通道中语义信息不同权重[15]，关注更重要的语义信息。通道注意力子模块结构如图 3 所示。

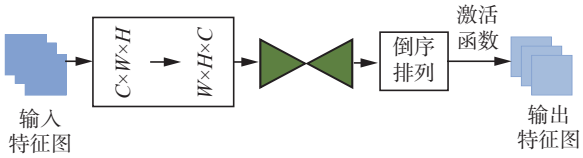


图 3　通道注意力子模块

Fig. 3　Channel attention submodule

在通道注意力中使用两层 MLP 感知机来增强通道之间的特征信息交互，层间使用 ReLU 激活函数预防梯度消失和梯度爆炸。在第 1 层 MLP 感知机中以大小为 r 的压缩比率将通道数压缩，来减少注意力机制中的计算量。然后在第 2 层 MLP 感知机将通道数恢复为原通道数，最后使用 Sigmoid 函数生成通道权重系数，将得到的通道权

重系数与输入特征图相乘做加权处理[16]，具体计算过程如下所示：

$$y = x_1 \boldsymbol{F}_1^{\mathrm{T}} + b_1 \tag{3}$$

$$\boldsymbol{F}_2 = \mathrm{sigmoid}[\mathrm{ReLU}(x_2 y + b_2)]^{\mathrm{T}} \boldsymbol{F}_1 \tag{4}$$

式中：$\boldsymbol{F}_1$为语义分支输入特征图，其大小为 $[C, W, H]$；$\boldsymbol{F}_2$为通道注意力输出特征图；x_1、x_2和b_1、b_2分别为两层 MLP 感知机的随机初始权重值和偏置项。

2）空间注意力子模块。

虽然通道注意力实现了在通道维度关注不同的特征图，但是不能实现在空间维度关注特征图的局部信息，而空间注意力机制利用特征之间的空间关系生成空间注意力映射，来关注更有意义的特征图局部信息[17]。结构如图 4 所示。

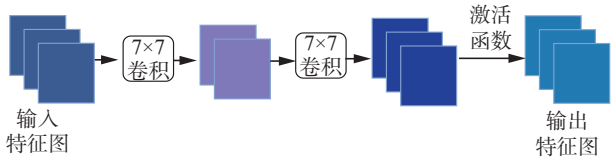


图 4　空间注意力子模块

Fig. 4　Spatial attention aubmodule

通道注意力主要由两层卷积组成，其卷积核大小为 7，步长为 1，补零填充为 3，并采用与通道注意力模块相同的压缩比率 r。在卷积层中特征图输出尺寸计算如下：

$$W_{\mathrm{out}} = \frac{(W_{\mathrm{in}} - k + 2p)}{S} + 1 \tag{5}$$

$$H_{\mathrm{out}} = \frac{(H_{\mathrm{in}} - k + 2p)}{S} + 1 \tag{6}$$

式中：W_{in}、H_{in}是数入特征图的宽和高；W_{out}、H_{out}是输出特征图的宽和高；p 为补零填充，k 为卷积核的大小；S 为步长。

由式 (5) 和式 (6) 计算过程可得出两层卷积均不进行下采样操作，所以图像分辨率保持不变，减少了空间信息的损失。卷积层采用的大感受野提取到的全局特征图对特征图各部分做权重计算，并将特征图数量以 r 的压缩比进行压缩，将不同通道中相同位置的特征图的特征信息进行空间信息融合，用于减少计算开销。卷积层之间使用批标准化（batch normalization）和 ReLU 激活函数，以防止梯度消失和梯度爆炸。第 2 层卷积后恢复输入特征图数量。特征提取过程如图 5 所示。

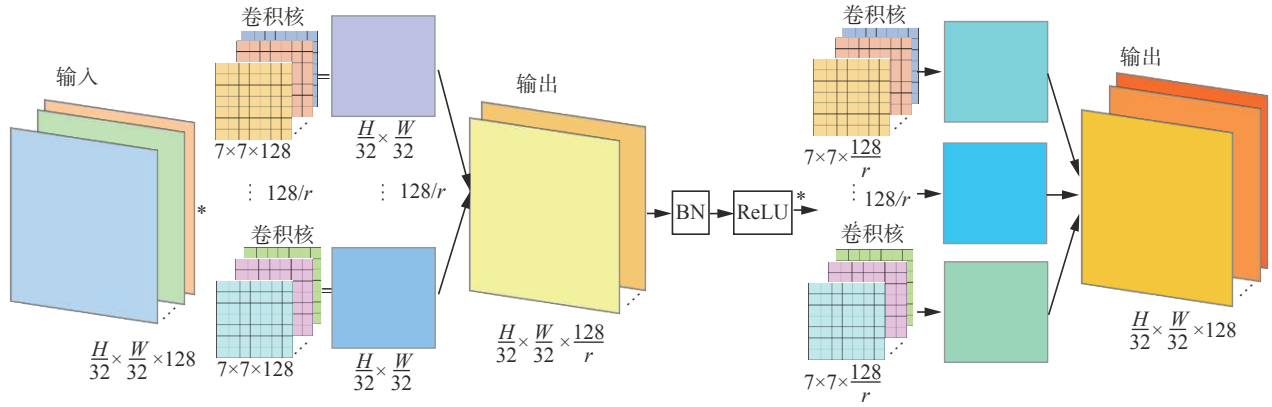


图 5　特征提取过程

Fig. 5　Feature extraction process

最后使用 Sigmoid 函数生成空间权重系数，将空间权重系数与输入特征图相乘做加权处理，实现对特征图区域的不同关注。计算过程如下：

$$F_3 = M_S(F_2) \otimes F_2 = \mathrm{sigmoid}[\mathrm{ConvBN}(\mathrm{ConvBNReLU}(F_2))] \tag{7}$$

1.2　混合空洞卷积

空洞卷积在保证图像分辨率不变的条件下，可以扩大感受野[18]，有利于大尺度目标的分割，而且特征图像分辨率的保持有利于精确定位目标。但由于空间细节分支主要是提取细节，如果在空洞卷积中设定固定的膨胀率会固定感受野的大小，不利于多尺度的细节信息提取，而且相同膨胀率空洞卷积多层叠加，会造成网格效应产生，导致空洞卷积层之间的结果缺乏相关性，造成局部信息的丢失。而混合空洞卷积是由一组有着不同膨胀率的空洞卷积叠加，不但有利于多尺度细节信息提取混合卷积，还解决了空洞卷积叠加的网格效应。但在设计混合空洞卷积时其叠加卷积的膨胀率不能有大于 1 的公约数，且要满足：

$$M_i = \max[M_{i+1} - 2d_i, M_{i+1} - 2(M_{i+1} - d_i), d_i]$$

式中：d_i是 i 层的膨胀率，M_i是在 i 层最大膨胀率，当网络有 n 层时设$M_n = r_n$，当M_2不大于卷积核的大小时，可以避免网格效应的产生。空间细节分支输出分辨率为输入图像的 1/8，卷积层中卷积核大小为 7×7，膨胀率过大时，出现感受野范围超过

特征图大小情况。所以扩大合适的感受野，需要合理设计混合空洞卷积的膨胀率，故本文提出的膨胀率为 [1,2,3]，使感受野增大在合理的范围内。

特征图空洞卷积运算：

$$k_e = k + (k-1)(d-1) \tag{8}$$

$$W_{out} = \frac{(W_{in} - (k-1)d + 2p - 1)}{S} + 1 \tag{9}$$

$$H_{out} = \frac{(H_{in} - (k-1)d + 2p - 1)}{S} + 1 \tag{10}$$

其中：d 为膨胀率的大小；k_e 是经过膨胀卷积中的感受野的大小。

由式 (8) 中可以得出空洞卷积与普通卷积相比较，卷积核的大小相同，而空洞卷积的膨胀率 d>1，空洞卷积在卷积核中会塞入 (d−1) 个空值，比普通卷积相比扩大了感受野的范围，而且还能保持和普通卷积相同的分辨率。

1.3　特征融合模块

语义分支提取的语义信息表征能力强，分辨率低，空间细节的表征能力弱，而空间细节分支提取的细节信息表征能力强，分辨率高，但是语义信息表征能力弱。因此要在语义分割中设计合理策略去融合这些特征图[19]，提高分割精度。本文设计一个新的特征融合模块，该模块能同时兼顾细节分支与语义分支的特征信息。BiSeNetV2 特征融合采用语义分支和空间细节分支特征图双向交互融合的结构，虽然这种结构相较于简单的特征融合方式提高了精度，但是仅通过不同尺度的交互特征融合方法，仍会出现分割精度降低问题。为解决特征图融合不充分问题本文做出改进，在特征融合部分保留了原分支特征图相互指导融合部分，其次将输出特征图通道数由 128 压缩至 64，提高特征图信息量并降低计算量，最后特征图输入深度聚合金塔池化模块[10]，扩大感受野并融合多尺度特征图。深度聚合金塔池化模块如图 6 所示。

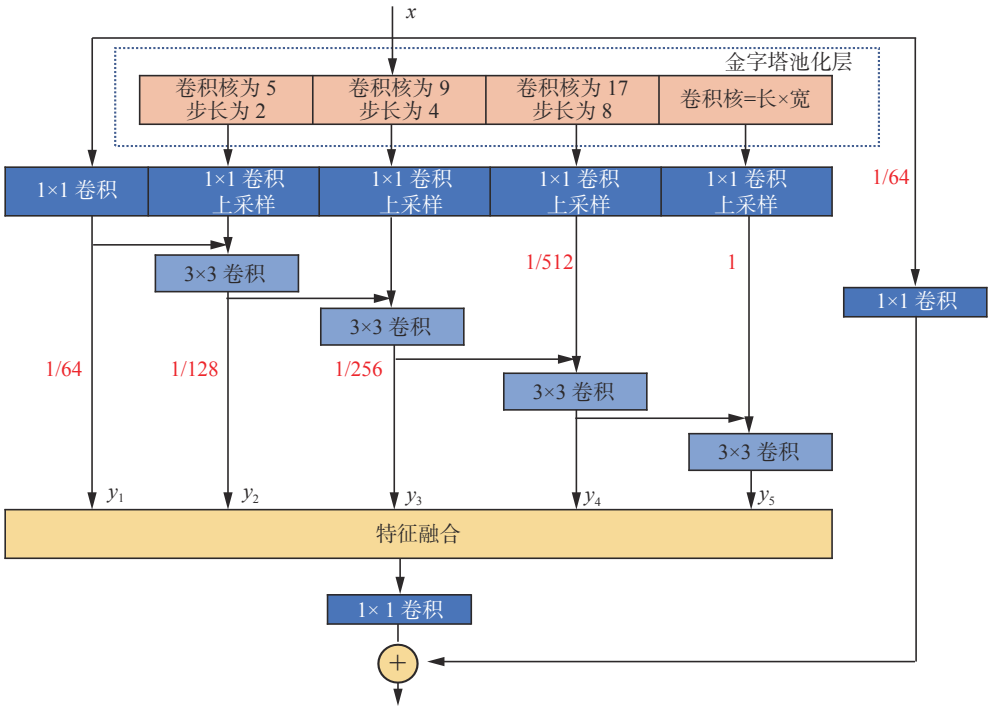


图 6　特征融合模块

Fig. 6　Feature fusion module

金字塔池中通过平均池化层不同的池化卷积核和步长，生成特征图分辨率分别为原图的 1/128、1/256、1/512 的特征图，并且输入 1/64 分辨率的特征图和由全局池化生成的图像级别信息也被利用。

如图 6 所示，从左到右 6 条分支分辨率分别为 1/64、1/128、1/256、1/512、1(全局池化)、1/64，输出特征图分辨率y_i可用公式表达如下：

$$y_i = \begin{cases} \mathrm{Covn}_{1\times1}(x), & i = 1 \\ \mathrm{Covn}_{3\times3}(\mathrm{U}(\mathrm{Covn}_{1\times1}(P_{2^i+1,2^{i-1}}(x))) + y_{i-1}), & 1 < i < n \\ \mathrm{Covn}_{3\times3}(\mathrm{U}(\mathrm{Covn}_{1\times1}(P_{\mathrm{globle}}(x))) + y_{i-1}), & i = n \end{cases}$$

式中：Covn表示卷积；U 表示上采样；p_{ij}表示池化层；i 表示池化层中卷积核大小；j 表示步长。

平均池化表达式为

$$\mathrm{out}(h,w) = \frac{1}{kH \cdot kW} \sum_{m=0}^{kH-1} \sum_{m=0}^{kH} \mathrm{in}(s \times h + m, s \times w + n)$$

式中: H 和 W 为输入特征图的高和宽; h 和 w 为输出特征图的高和宽, 池化核尺寸是(kH,kW), s 为步长大小。

经过金字塔池化输出不同分辨率的特征图后, 先通过 concatenate 函数在通道维度将特征图拼接, 即增加了描述图像本身的特征数(通道数), 但通道中特征图表征的信息量未增加。然后在 $\mathrm{Covn}_{1\times1}$卷积层下跨通道做特征图聚合, 实现深度特征融合, 最后与输入特征图做特征图的叠加, 实现特征图数量保持不变, 描述的信息量增加。新的特征融合模块通过生成多尺度特征, 充分融合了来自语义分支和空间细节信息的特征图, 进而提高了分割精度。

2　实验与分析

2.1　数据集描述

本文实验采用 Cambridge-driving Labeled Video Database(CamVid) 和 ISPRS 航空影像 Vaihingen 两个公开数据集来验证改进模型的有效性, 数据集的具体信息如下:

1)Camvid 数据集。该数据集从驾驶汽车的角度拍摄的街景图像, 标签包含 32 个语义类别, 从中选择 11 个关注的类别。该数据集图像大小 720 像素×960 像素, 共有 701 张数据集, 包含 367 张训练集图像, 101 张验证集图像, 233 张测试集。

2)Vaihingen 数据集。该数据集是在德国上空所捕获共有 33 张图像, 该数据集包含有 6 个语义类, 分别为不透水面、低矮植被、建筑、树、汽车、杂乱背景。数据集图像大小不一致, 图像平均大小为 2494 像素×2064 像素。由于图像过大, 受计算机显卡硬件限制, 将图像裁剪为 512 像素×512 像素大小, 获得 3300 张图像, 以 6∶4 比例随机划分数据, 充分利用数据集。

2.2　实验参数设置

实验基于 Pytorch 1.9.0 深度学习框架和 Cuda 11.1 库,由 Python3.8 编写, 在 Linux 操作系统下使用一块 NVIDIA GeForce GTX 3090 进行模型训练, 一块 NVIDIA GeForce GTX 1080Ti 用于模型验证, CPU 为 E5-2650 v4。迭代次数设置为 600, 采用 Adam 优化器, 起始学习率设置为 0.001。损失函数实验中采用交叉熵损失函数 (CrossEntropyLoss), 计算公式为

$$\mathrm{CrossEntropyLoss}=-\frac{1}{N}\sum_{n=1}^{N}(wr_n\log(p_n)+(1-r_n)\log(1-p_n))$$

其中 $w=\dfrac{N-\sum_n p_n}{\sum_n p_n}$, r_n为像素的原类别, p_n为像素预测的类别。

2.3　数据集处理与数据增强

通过对实验数据集的分析, 数据集存在类别不均衡问题, 在训练计算损失函数的过程中, 类别频率高的损失函数值与类别频率低的损失函数值直接求和不利于网络的收敛, 这不利于求解语义分割问题, 为此, 本文在训练时采用等级权重来平衡类别[20], 计算各类权重需要计算出图像像素点数和各类标签的像素数。因此, 类别权重的计算公式为

$$\mathrm{freq}_c=\frac{\mathrm{sum}_c}{\mathrm{sum}_{pc}}$$

$$\mathrm{weight}_c=\frac{\mathrm{median\_freq}_c}{\mathrm{freq}_c}$$

式中:freq_c代表第 c 类像素在训练集标签中出现频率; sum_c代表包含第 c 类训练标签的总像素数。weight_c代表第 c 类像素的权重; $\mathrm{median_freq}_c$表示所有语义类别的频率中值。

CamVid 数据集和数据集类别频率和权重值如图 7 所示。从图 7 可以看出语义类别杆、信号标志、栅栏、汽车、行人、自行车所占的比例较小分别为 1.00%、1.10%、1.50%、1.90%、0.70%、0.50%, 所以赋值权重较高。

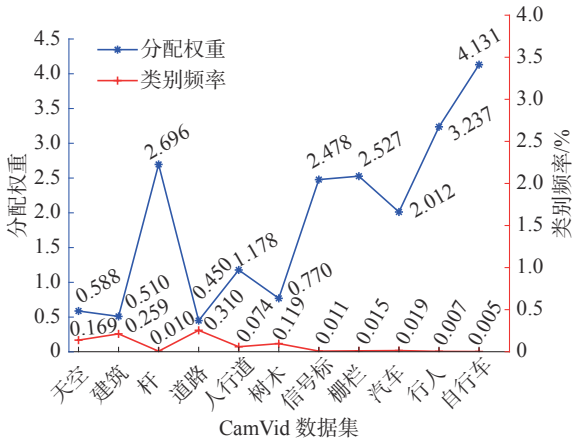


图 7　CamVid 数据集和数据集类别频率和权重值

Fig. 7　CamVid dataset and dataset class frequencies and weight values

Vaihingen 数据集和数据集类别频率和权重值如图 8 所示。从图 8 可以清楚看出类别不透水面、低矮植被、建筑、树所占的比例分别为 28.178%、20.016%、26.127%、23.4740% 所占的比例较大, 而汽车、杂乱背景比例为 1.1313%、0.891% 所占的比例较小, 类别失衡较为严重, 分配类别权重后

可以缓解此问题[21]。另外，在训练的过程中，为防止模型过拟合，采用垂直翻转策略对两个数据集都进行数据增强。

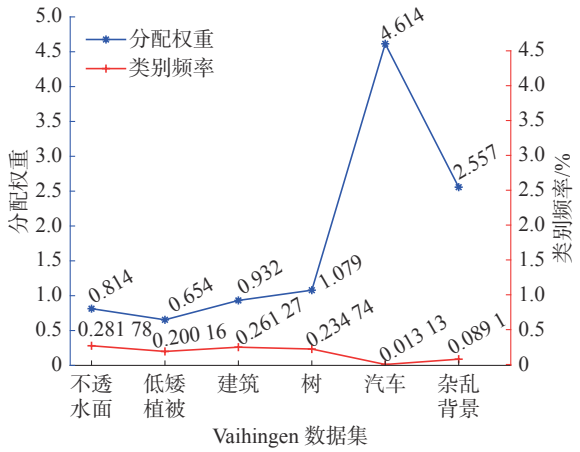


图 8　Vaihingen 数据集和数据集类别频率和权重值

Fig. 8　Vaihingen dataset and dataset class frequencies and weight values

2.4　采用的评价指标

在实验中采用交并比 (intersection over Union, IoU)，准确率 (Acc)，F_1 分数 (F_1-Score)，平均交并比 (mean intersection over union, MIoU)，平均准确率 (mean accuracy, MA) 作为评价指标，可被定义为

$$\mathrm{IoU}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

$$\mathrm{MIoU}=\frac{1}{k}\sum_{i=0}^{k}\mathrm{IoU}$$

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

$$\mathrm{MA}=\frac{1}{k}\sum_{i=0}^{k}\mathrm{Acc}$$

$$\mathrm{F_1\text{-}Score}=2\times\frac{P\times R}{P+R}$$

其中，TP 为预测是正确的正样本；FP 为预测是错误的正样本；TN 为预测是正确的负样本；FP 为预测是错误的负样本。R 为召回率，$R=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$，$P$ 表示预测为正的样本中实际的正样本的数量，$P=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$。

2.5　消融实验

为验证本文研究方法的正确性,在 CamVid 测试集和 Vaihingen 测试集做了如表 1 所示的消融实验。

表 1　CamVid 测试集和 Vaihingen 测试集消融实验

Table 1　Ablation experiments on CamVid test set and Vaihingen test set　　%

	Camvid			Vaihingen		
	平均交并比	平均准确率	平均F_1分数	平均交并比	平均准确率	平均F_1分数
BiSeNetV2	72.40	—	—	87.72	93.13	93.35
混合空洞卷积、DAPPM、BiSeNetV2	76.45	85.19	82.83	88.47	**93.77**	93.01
混合空洞卷积、DAPPM、全局注意力机制、BiSeNetV2	**76.69**	**85.82**	**83.73**	**88.53**	92.70	**93.79**

CamVid 数据集实验结果表明，相比于 BiSeNetV2 的 MIoU，加入全局注意力机制、DAPPM 模块后，网络分割的性能有明显提升，提高了 4.05%，在此基础上引入的混合空洞卷在 MIoU 提高了 0.25%，平均准确度从 82.83% 提高到了 83.73%，平均 F_1 分数从 85.19% 提高到了 85.82%。在 Vaihingen 数据集的实验结果显示，加入全局注意力机制和 DAPPM 模块后 MIoU 提高了 0.75%，但平均 F_1 分数降低了 0.34%，引入混合空洞卷积后 MIoU 提高了 0.81%，平均准确度降低了 0.43%，平均 F_1 分数提高了 0.44%。两个数据集实验结果看出本文提出的方法在 MIoU 评价指标上有一定幅度的提升。

实时性语义分割网络具备实时性和轻量性结构，以输入图像尺寸为 720 像素×960 像素的 Camvid 数据集为例，本文网络模型复杂度对比分析如表 2 所示。

表 2　网络模型复杂度对比分析

Table 2　Network structure calculation comparison

模块	BiSeNetV2		GaSeNet	
	存储空间/MB	GFLOPs	存储空间/MB	GFLOPs
语义分支	1.16	1.763	1.57	2.101
空间细节分支	0.519	15.152	0.519	15.497
特征融合模块	1.672	12.627	**1.302**	**10.997**
总和	3.634	31.8	3.674	**28.595**

本文采用浮点运算次数 (GFLOPs) 分析时间复杂度, 从总参数量和各层输出特征图所占存储空间分析空间复杂度。假设深度为 D 层的神经网络模型结构中, 第 l 层的特征图尺寸为 $H\times W$, 卷积核大小为 K, 输入与输出通道数分别为 $C_{\text{in}}^{l}, C_{\text{out}}^{l}$, 则时间复杂度与空间复杂度可表示为

$$T(n)\sim O\left(\sum_{l=1}^{D} H\times W\times C_{\text{in}}^{l}\times K_l^2\times C_{\text{out}}^{l}\right)$$

$$S(n)\sim O\left(\sum_{l=1}^{D} K_l^2\times C_{\text{in}}^{l}\times C_{\text{out}}^{l}+\sum_{l=1}^{D} H\times W\times C_{\text{out}}^{l}\right)$$

浮点运算次数是用来计算整个网络模型中乘法或加法的运行次数, 可反应出模型的计算量, 是衡量模型的时间复杂度的重要指标。从表 2 中得到本文提出的网络模型结构浮点运算次数为 28.594 GFLOPs 比 BiSeNetV2 的浮点运算次数 31.8 GFLOPs 降低了 3.206 GFLOPs, 与原模型相比降低了计算量。这是由于本文在引入 DAPPM 模块时调整了原来的双向交互结构语义分支, 减少了中间层的输入输出通道数, 从而降低了计算量。虽然 GaSeNet 网络结构引入全局注意力机制和混合空洞卷积, 增加了部分计算量, 但是网络的整体计算量对比 BiSeNetV2 的整体计算量, 降低了计算的复杂度, 从理论上加快了推理速度。实时语义分割网络应具备轻量化特点, 推理过程中应占用较小的存储空间, 为分析空间复杂度, 输入一张尺寸为 720 像素×960 像素的图像, 得到网络模型总参数与各层输出特征图共占用存储空间 3.674 MB, 相比于 BiSeNetV2 的 3.634 MB 仅增加了 0.04 MB, 故本文重新设计的结构具备轻量性。

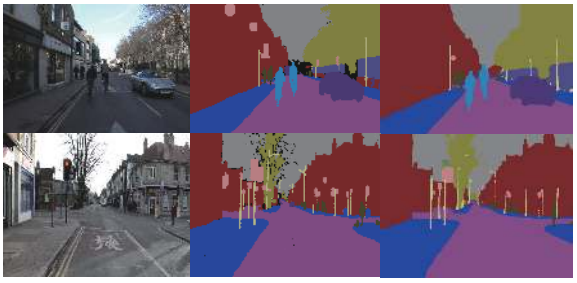
2.6　实验结果

表 3 给出了本文方法在 CamVid 数据集各个类别的交并比 (IoU)、准确率 (Acc) 和 F_1分数 (F_1-Score)。图 9 为 CamVid 数据集 RGB 图像、相应标签和采用本文方法的预测结果图像。由表 3 可以看出 MIoU 为 76.69%, 平均准确率和平均 F_1 分数分别达到了 83.73%、85.82%。类别杆、信号标识、行人、自行车交并比(IoU)分别为 47%、57%、62.35%、64.64%, 属于目标尺寸小, 类别边界模糊, 出现频率低的类别, 分割任务具有一定的难度, 故结果精度较低。从图 9 预测分割结果来看, 整体来说各类别分割出来的轮廓较为清晰, 在类别频率高、尺度大的类别如道路类别, 与标签对比两者无太大的差异。但在小目标如杆类, 虽然出现了边界模糊现象, 但可以辨认出其所属类别, 以及出现了将自行车错误预测为信号标识的错误分类情况。

表 3　CamVid test 数据集上的评估结果

Table 3　Evaluation results on the CamVid test dataset　%

类别	交并比	准确率	F_1分数
天空	93.9	96.88	96.86
建筑	90.47	94.88	95.00
杆	47.00	54.94	63.95
道路	96.93	98.12	98.44
人行道	87.23	94.36	93.18
树木	81.83	91.49	90.01
信号标识	57.27	59.14	72.83
栅栏	72.21	80.87	83.86
汽车	89.78	95.46	94.62
行人	62.35	84.04	76.81
自行车	64.64	70.91	78.52
平均值	76.69	83.73	85.82



(a) 原图　(b) 标签　(b) 预测图

图 9　CamVid 数据集 RGB 图像、标签和预测图

Fig. 9　CamVid dataset RGB images, labels and prediction results

本方法与其他方法在 CamVid Test 数据集对比实验如表 4 所示。在表 4 对比实验结果中, 为提高本文实验的可信度, 将本文网络模型与公开的对比网络模型, 统一部署在相同的 NVIDIA GeForce GTX 1080Ti 进行对比实验。结果表明本文 MIoU 为 76.69%, 分割精度要优于其他对比模型方法, 其中比 BiSeNetV2 提高了 4.45%, 且高于 DDRNet-23 的 76.3% 和 PP-LiteSeg-T[22] 的 73.3%。模型实际推理速度为 65.59 f/s, 对比 BiSeNetV2 速度降低了 3.36 f/s, 虽然本文从浮点运算次数的理论上分析得出时间复杂度要低于 BiSeNetV2, 但由于增加网络模型的并行分支会受实验环境 IO 读取速率影响, 进而出现实际推理速度降低的情况。对比 BiSeNet[Xception39][23] 速度低了 49.25 f/s, 但对比以 BiSeNetV2 基础上增大卷积深度和宽度, 用以提升网络性能的 BiSeNetV2-L, 提高了 34 f/s。本文网络计算量 GFLOPs 为 28.595

要低于 BiSeNet[ResNet18][23]、BiSeNetV2、BiSeNetV2-L 其理论计算复杂度更低，但比更轻量化的 DDRNet-23 的计算量高了 16.495 GFLOPs。因此，本文的网络模型在提高了分割精度的同时保持了实时性与轻量化特点。

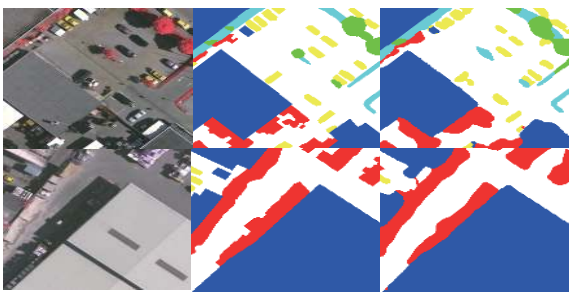
表 4　不同方法在 CamVid test 数据集上的表现结果
Table 4　Performance results of different methods on the CamVid test dataset

网络结构	平均交并比/%	速度/(f·s^{-1})	GFLOPs
BiSeNet [ResNet18]	65.60	**175**	34.0
BiSeNet [Xception39]	68.70	116.25	—
BiSeNetV2	72.40	68.95	31.80
BiSeNetV2-L	73.20	33	118.5
DDRNet-23	76.30	90	**12.10**
PP-LiteSeg-T	73.30	91	15.3
GaSeNet	**76.69**	66.5	28.595

本文方法在 Vaihingen 数据集评价指标如表 5 所示，图 10 为 RGB 图像、相应标签和采用本文方法的预测结果图像。

表 5　Vaihingen 数据集上的评估结果
Table 5　Evaluation results on the Vaihingen dataset　%

类别	交并比	准确率	F_1分数
不透水面	92.65	96.6	96.19
低矮植被	86.45	92.43	92.73
建筑	95.77	98.11	97.84
树木	88.62	93.82	93.97
汽车	76.10	80.98	86.43
杂乱背景	91.58	94.27	95.60
平均值	88.53	92.70	93.79



(a) 原图　(b) 标签　(b) 预测图

图 10　Vaihingen 数据集 RGB 图像、标签和预测图
Fig. 10　Vaihingen dataset RGB images, labels and prediction results

本文提出的 GaSeNet 网络结构在 Vaihingen 数据集 MIoU 为 88.53%，平均准确度达到了 92.7%，平均 F_1-Score 达到了 93.79%，建筑类别的交并比（IoU）、准确度（Acc）和 F_1-Score 分别达到了 95.77%、98.11%、97.84%。其中汽车类别交并比为 76.1%，分割结果对比其他类别来说较差。从预测结果图像中对于小的目标如汽车、低矮灌木丛能够做的准确定位，并且类别与类别之间的边界较为清晰。但由于在航空遥感图像中受光照拍摄角度影像，出现了预测结果中类别轮廓比较模糊的问题[24]。

从表 6 看出本文方法在该数据集上取得了显著性提高，本文方法在 MIoU 的结果分别比 FCN、SegNet、DeepLabV3+[25] 高出了 18.672%、13.086%、15.511 5%，表明本文方法在航空影像语义分割数据集中能获得较高的分割精度。

表 6　不同方法在 Vaihingen 数据集上的表现结果
Table 6　Performance results of different methods on the Vaihingen dataset　%

Classes	FCN	SegNet	DeepLabV3+	GaSeNet
不透水面	81.001	86.294	83.371	**92.65**
低矮植被	59.378	66.889	63.456	**86.45**
建筑	83.339	91.603	84.344	**95.77**
树木	73.789	78.934	72.665	**88.62**
汽车	48.017	56.562	59.977	**76.10**
杂乱背景	73.624	72.383	74.298	**91.58**
平均值	69.858	75.444	73.018 5	**88.53**

3　结束语

本文提出结合全局注意力机制的实时语义分割网络 (GaSeNet)，首先语义分支引入全局注意力机制扩大通道间的全局交互能力，增强了语义特征的表达，其次空间细节分支设计了混合空洞卷积块，保持图像分辨率扩大感受野的范围，提高细节信息的提取准确率，使得不同的膨胀率的空洞卷积适应不同尺度的目标，最后将金字塔池化与深度特征融合策略引入新的特征融合模块，该模块能有效融合来自语义分支和空间细节分支不同级别的特征图，提高特征信息的利用率。实验结果表明本文方法在未增加计算开销的情况下，提高了实时语义分割的正确率，并有效平衡了推理速度与分割性能。在未来的研究中，我们将在大型数据集如 COCO-Stuff、Cityscapes 上做模型泛化性研究，并进一步提高轻量级实时语义分割网络精度。

参考文献:

[1] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015: 3431−3440.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015−04−10)[2022−08−06].https://doi.org/10.48550/arXiv.1409.1556.

[3] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234−241.

[4] YU F, KOLTUN V. MULTI-SCALE context aggregation by dilated convolutions[EB/OL]. (2016−04−30)[2022−08−06].https://doi.org/10.48550/arXiv.1511.07122.

[5] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017−12−05)[2022−08−06].https://doi.org/10.48550/arXiv.1706.05587.

[6] WANG P, CHEN P, YUAN Y, et al. Understanding convolution for semantic segmentation[C]//2018 IEEE winter conference on applications of computer vision. Nevada: IEEE, 2018: 1451−1460.

[7] ZHAO Hengshuang, QI Xiaojuan, SHEN Xiaoyong, et al. ICNet for Real-Time Semantic Segmentation on High-Resolution Images[C]//European Conference on Computer Vision. Cham: Springer, 2018: 418−434.

[8] LI Hanchao, XIONG Pengfei, FAN Haoqiang, et al. DFANet: deep feature aggregation for real-time semantic segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 9514−9523.

[9] YU Changqian, GAO Changxin, WANG Jingbo, 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.

[10] HONG Yuanduo, PAN Huihui, SUN Weichao, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes[EB/OL]. (2021−09−01)[2022−08−06].https://doi.org/10.48550/arXiv.2101.06085.

[11] LIU Yichao, SHAO Zongru, HOFFMANN N. Global attention mechanism: retain information to enhance channel-spatial interactions[EB/OL]. (2021-12-10)[2022-08-06].https://doi.org/10.48550/arXiv.2112.05561.

[12] 丁宗元, 孙权森, 王涛, 等. 基于融合多尺度标记信息的深度交互式图像分割 [J]. 计算机研究与发展, 2021, 58(8): 1705–1717.
DING Zongyuan, SUN Quansen, WANG Tao, et al. Deep interactive image segmentation based on fusion multi-scale annotation information[J]. Journal of computer research and development, 2021, 58(8): 1705–1717.

[13] 张噢琦, 亢宇鑫, 武卓越, 等. 基于多尺度特征和注意力机制的肝脏组织病理图像语义分割网络 [J]. 模式识别与人工智能, 2021, 34(4): 375–384.
ZHANG Aoqi, KANG Yuxin, WU Zhuoyue, et al. Semantic segmentation network of pathological images of liver tissue based on multi-scale feature and attention mechanism[J]. Pattern recognition and artificial intelligence, 2021, 34(4): 375–384.

[14] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 3−19.

[15] 杨昆, 常世龙, 王尉丞, 等. 基于 sECANet 通道注意力机制的肾透明细胞癌病理图像 ISUP 分级预测 [J]. 电子与信息学报, 2022, 44(1): 138–148.
YANG Kun, CHANG Shilong, WANG Yucheng, et al. Predict the ISUP grade of clear cell renal cell carcinoma using pathological images based on sECANet Chanel attention[J]. Journal of electronics & information technology, 2022, 44(1): 138–148.

[16] 张志华, 温亚楠, 慕号伟, 等. 结合双注意力机制的道路裂缝检测 [J]. 中国图象图形学报, 2022, 27(7): 2240–2250.
ZHANG Zhihua, WEN Yanan, MU Haowei, et al. Dual attention mechanism based pavement crack detection[J]. Journal of image and graphics, 2022, 27(7): 2240–2250.

[17] FU Jun, LIU Jing, TIAN Haijie, et al. Dual attention network for scene segmentation[C]//2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3146−3154.

[18] 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.

[19] 徐硕, 郑锋, 唐俊, 等. 双分支特征融合网络的步态识别算法 [J]. 中国图象图形学报, 2022, 27(7): 2263–2273.
XU Shuo, ZHENG Feng, TANG Jun, et al. Dual branch feature fusion network based gait recognition algorithm[J]. Journal of image and graphics, 2022, 27(7): 2263–2273.

[20] 刘万军, 佟畅, 曲海成. 空洞卷积与注意力融合的对抗式图像阴影去除算法 [J]. 智能系统学报, 2021, 16(6): 1081–1089.
LIU Wanjun, TONG Chang, QU Haicheng. An antagonistic image shadow removal algorithm based on dilated convolution and attention mechanism[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1081–1089.

[21] 吴止锾, 高永明, 李磊, 等. 类别非均衡遥感图像语义分割的全卷积网络方法 [J]. 光学学报, 2019(4): 393−404.
WU Zhihuan, GAO Yongming, LI Lei, et al. Fully convolutional network method of semantic segmentation of class imbalance remote sensing images[J]. Acta optica sinica, 2019(4): 393−404.

[22] PENG J, LIU Y, TANG S, et al. PP-LiteSeg: A superior real-time semantic segmentation model[EB/OL]. (2022−04−06)[2022−08−06].https://doi.org/10.48550/arXiv.2204.02681

[23] YU Changqian, WANG Jingbo, PENG Chao, et al. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation[C]//European Conference on Computer Vision. Cham: Springer, 2018: 334−349.

[24] 张小娟, 汪西莉. 完全残差连接与多尺度特征融合遥感图像分割 [J]. 遥感学报, 2020, 24(9): 1120–1133.
ZHANG Xiaojuan, WANG Xili. Image segmentation models of remote sensing using full residual connection and multiscale feature fusion[J]. Journal of remote sensing, 2020, 24(9): 1120–1133.

[25] CHEN L C, ZHU Yukun, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[M]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 833−851.

作者简介：

李涛, 讲师, 博士, 主要研究方向为智能信息处理、数据挖掘。参与或主持国家自然科学基金、省级自然科学基金和省级科技攻关项目 5 项。发表学术论文 10 余篇。

高志刚, 本科生, 主要研究方向为深度学习、图像语义分割、目标检测、计算机视觉。

管晟媛, 硕士研究生, 主要研究方向为深度学习、数字水印、计算机视觉。