计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(8):68−77 [doi: 10.15888/j.cnki.csa.009596]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 ViT-D-UNet 的双分支遥感云影检测网络①

李远禄[1,2], 王键翔[1], 范小婷[1], 周　昕[1], 吴明轩[1]

[1](南京信息工程大学 自动化学院, 南京 210044)
[2](江苏省大气环境与装备技术协同创新中心, 南京 210044)
通信作者: 王键翔, E-mail: 202212490003@nuist.edu.cn

摘　要: 云及其阴影的有效分割是遥感图像处理领域中重要的问题, 它对于地表特征提取、气候检测、大气校正等有很大帮助. 然而云和云影遥感图像特征复杂, 云分布多样不规则, 且边界信息模糊易受背景干扰等特点, 导致其特征难以准确提取, 也少有专门为其设计的网络. 针对以上问题, 本文提出一种 ViT (vision Transformer) 和 D-UNet 双路网络. 本文网络分为两个分支: 一路是基于卷积的局部特征提取模块, 在 D-UNet 的膨胀卷积模块基础上, 引入深度可分离卷积, 提取多尺度特征的同时, 减少参数; 另一路通过 ViT 在全局上理解上下文语义, 加深对整体特征提取. 两支路间存在信息交互, 完善提取的特征信息. 最后通过独特设计的融合特征解码器, 进行上采样, 减少信息丢失. 模型在自建的云和云影数据集以及 HRC_WHU 公开数据集上取得优越的性能, 在 *MIoU* 指标上分别领先次优模型 0.52% 和 0.44%, 达到了 92.05% 和 85.37%.

关键词: 遥感; 云检测; 语义分割; 特征融合

引用格式: 李远禄,王键翔,范小婷,周昕,吴明轩.基于 ViT-D-UNet 的双分支遥感云影检测网络.计算机系统应用,2024,33(8):68–77. http://www.c-s-a.org.cn/1003-3254/9596.html

Bi-branch Remote Sensing Cloud and Shadow Detection Network Based on ViT-D-UNet

LI Yuan-Lu[1,2], WANG Jian-Xiang[1], FAN Xiao-Ting[1], ZHOU Xin[1], WU Ming-Xuan[1]

[1](School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China)
[2](Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing 210044, China)

Abstract: Effective segmentation of clouds and their shadows is a critical issue in the field of remote sensing image processing. It plays a significant role in surface feature extraction, climate detection, atmospheric correction, and more. However, the complex features of clouds and cloud shadows in remote sensing images, characterized by their diverse, irregular distributions and fuzzy boundary information that is easily disturbed by the background, make accurate feature extraction challenging. Moreover, there are few networks specifically designed for this task. To address these issues, this study proposes a dual-path network combining vision Transformer (ViT) and D-UNet. The network is divided into two branches: one is a convolutional local feature extraction module based on the dilated convolution module of D-UNet, which introduces a multi-scale atrous spatial pyramid pooling (ASPP) to extract multi-dimensional features; the other branch comprehends the context semantics globally through the vision Transformer, enhancing feature extraction. Finally, the study performs an upsampling through a feature fusion decoder. The model achieves superior performance on both a self-built dataset of clouds and cloud shadows and the publicly available HRC_WHU dataset, leading the second-best model by 0.52% and 0.44% in the *MIoU* metric, achieving 92.05% and 85.37%, respectively.

Key words: remote sensing; cloud detection; semantic segmentation; feature fusion

① 基金项目: 国家自然科学基金 (61671010)
收稿时间: 2024-02-28; 修改时间: 2024-03-28; 采用时间: 2024-04-03; csa 在线出版时间: 2024-06-28
CNKI 网络首发时间: 2024-07-01

遥感技术飞速发展, 帮助人们更好地了解地表信息. 而云及云影的检测对于遥感图像处理有重大意义. 云影检测可以更好评估土地覆盖情况, 掌握太阳能资源空间分布和时间分布, 为农业、新能源等生产活动的选址和运转提供依据. 同时, 观测云影的变化和分布也对气候变化, 预测, 防范灾害性天气有重要意义. 因此, 精准的分割云和云影是遥感领域一个关键问题.

传统的检测方式主要有阈值法[1], 以及手工提取特征的方式[2]. 通过分析遥感影像的各个波段光谱的特征, 设置阈值来分离云影和其他地貌, 而它们的特征一般是通过大量的人工样本分析得到, 较为繁琐. 有代表性的方法是 Zhu 等人提出的 Fmask 算法[3], 利用卫星影像的大气顶层反射和亮度温度数据计算地貌和云影特征并通过设置阈值来分割遥感影像. 而后, 随着遥感分辨率的提高, 针对更多样特征的改进方法随之提出, 有 MFmask[4] , Tmask[5]等. 但是这些方法仍然依赖光谱信息分析, 不适用于多传感器图像, 而且由于不同卫星所捕获的光谱波段不同, 这些方法往往泛化性较弱.

近年来, 深度学习得到了迅猛发展, 在各个领域的出色表现引人注目, 在遥感图像领域也得到了广泛应用. 深度学习技术可以自动捕捉人工[6,7]难以注意的微小特征信息, 具有较高的准确率. 以 CNN[8–10]为基础的网络模型在图像分类任务中表现出色, 同时为完成像素级分类任务即语义分割的发展做了铺垫. Long 等人[11]于 2015 年, 首次引入了全卷积神经网络的概念 FCN, 通过对 CNN 网络的改进实现了图像像素的分类, 达到了图像分割的效果. 在医学影像领域, Ronneberger 等人[12]设计了一种基于编码器和解码器实现的 U 形网络, 数据不足的情况下达到了比较好的效果. 同样是基于编码器解码器的方式, Chen 等人[13]提出了 DeepLab, 创新地使用了空洞卷积 (atrous convolution) 扩大每层网络的感受野以及全连接随机场 CRF 用于细化特征信息, 优化了边缘细节, 同时也加快了模型推理速度. 然而这些网络往往忽视了局部和全局特征之间的交互, 使得存在复杂特征或干扰噪声时, 无法准确理解信息.

目前, 云影遥感图像主要有以下几个特点: 云影在遥感图像中的大小和形状变化多样. 由于云影的覆盖范围没有规律, 云层结构不固定, 亮度和对比度不均匀; 背景地貌特征十分丰富, 不同季节和地理条件下, 云影的形状和分布变化差异会十分大; 云影遥感图像中经常存在噪声、阴影、伪影等问题. 同时, 一些地貌, 比如雪覆盖区域, 建筑群密集区域, 水体, 沙漠裸露地表, 丘陵和山地的部分特征都与云相似, 使得网络难以区分.

针对以上问题, 本文在 D-UNet 的 U 形编码器和解码器基础上改进了线性特征提取过程, 提出了关注局部和全局的双分支遥感云影检测方法, 与上述方法比较有以下几点改进和贡献.

(1) 特征提取卷积支路基于 D-UNet 的膨胀卷积模块, 改善最终提取特征的感受野. 在前 4 层, 引入了设计的深度可分离卷积, 提取不同尺度特征信息, 在局部尽可能提取完整特征信息, 同时减少参数量.

(2) 在全局特征提取部分, 通过 ViT 模块提取全局特征, 增强整体上下文语意理解的同时, 增加对抽象特征的提取. 并且两支路间有特征交互. 完善各路特征信息.

(3) 在解码阶段, 通过特征融合对局部, 全局以及深层特征信息进行整合. 优化最终提取到的特征, 防止信息丢失, 提升分割准确率.

1 基础网络及相关模块

1.1 D-UNet

本文基本结构采用 D-UNet, 是龙丽红等人[14]基于 UNet 提出的变体模型, 网络依然采用了基于编码和解码的结构, 主要针对遥感影像设计改动. D-UNet 的编码阶段主要引入了带空洞卷积模块, 由多层卷积与最大池化层组成的, 在压缩提取特征的同时逐渐减少特征张量的分辨率, 增加通道数. 在解码过程中直接使用上采样或反卷积对特征的分辨率进行恢复, 并且使用跳跃残差机制来拼接浅层特征, 以加强对各层次的语义信息理解. 而损失函数采用联合策略, 可以提高模型的准确度、鲁棒性和泛化能力, 防止过拟合. 具体的实现细节和改动会在本文后续详细介绍.

1.2 深度可分离卷积

深度可分离卷积最早由 MobileNet[15]中提出. 其中的逐通道卷积用于对每个输入通道独立执行卷积操作. 逐点卷积作用为将逐通道卷积后得到的拼接特征的通道数改变, 以用于组合输出.

如图 1 所示, 若特征张量输入的大小为 $D_f \times D_f \times M$, 其中 D_f 是输入图像的高和宽大小, M 为的通道数. 在深度卷积过程中, 假设输出特征图大小为 $D_g \times D_g \times M$, 分别是输出图像的宽高和通道数, 通道数与输入图像

一致, 被用作下一个卷积的输入. 对于逐点卷积, 卷积核的大小为 1×1, 每个卷积核上的通道数必须与输入特征映射的通道数相同. 若卷积核数为 N, 经过卷积后输出的特征图大小为 $D_g \times D_g \times N$.

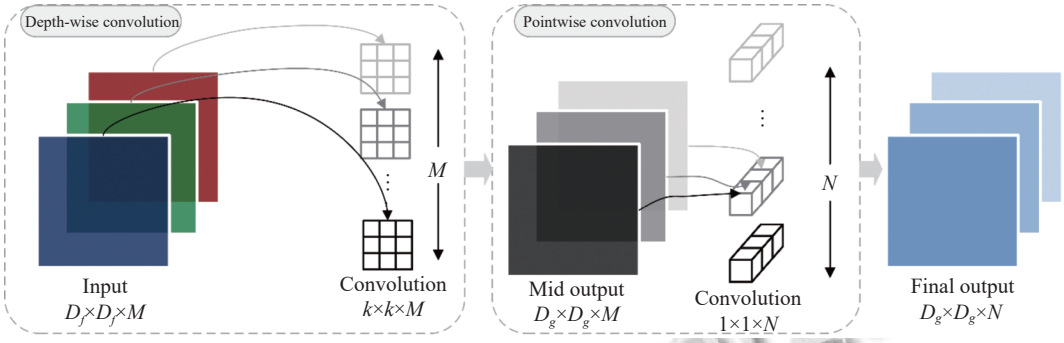


图 1 深度可分离卷积原理

对于输入特征张量 H, 其尺寸是 $D_f \times D_f$, 卷积核 K 的尺寸是 $k \times k$. 输入通道的数量核输出通道的数量分别是 M 和 N. 输出特征张量 G 的尺寸是 $D_g \times D_g$, 标准卷积操作可以如下定义:

$$G_j = \sum_{i=1}^{M} H_i \cdot K_i^j, \ j = 1, \cdots, N \tag{1}$$

其中, H_i 是在 H 中第 i 个特征张量, G_i 是在 G 中的第 i 个特征张量, 而 K_i^j 是在第 j 个卷积核中第 i 个切片. 此外, 符号“·”代表卷积操作. 训练参数量设为 P (包含卷积核参数和偏执参数), 浮点运算数为 F, 则在标准卷积过程中, 它们可以按照如下公式计算:

$$P_1 = k \times k \times M \times N \tag{2}$$

$$F_1 = k \times k \times M \times N \times D_g \times D_g \tag{3}$$

从式 (2) 看出, 参数量取决于核大小、输入通道数 M 和输出通道数 N. 式 (3) 显示浮点运算数取决于参数 P_1 核输出特征大小 $D_g \times D_g$.

在逐点卷积中, 如图 1 中所示, 每个核只占 1 个像素大小, 它是为了卷积改变每个输入通道, 这个过程可以定义为如下公式:

$$G_j = H_i \cdot K_j, \ i = 1, 2, \cdots, M \tag{4}$$

其中, K_j 第 j 个深度卷积核及逐点卷积核, 只改变通道数, 并不是结合张量创建新的特征. 因此, 不需要额外的操作和层, 即用 1×1 的标准卷积来实现.

对于深度可分离卷积过程, 参数 P_2 和浮点计算 F_2 是深度卷积和逐点卷积的总和. 因此 P_2 和 F_2 可以如式 (5)、式 (6) 计算得出.

$$P_2 = k \times k \times M + M \times N \tag{5}$$

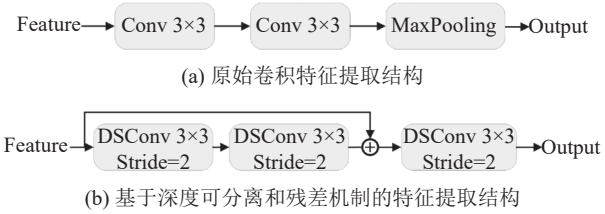
$$F_2 = k \times k \times D_g \times D_g \times M + D_g \times D_g \times M \times N \tag{6}$$

式 (5) 和式 (2) 的比例以及式 (6) 和式 (3) 的比例显示如下:

$$\frac{P_2}{P_1} = \frac{1}{N} + \frac{1}{k^2} \tag{7}$$

$$\frac{F_2}{F_1} = \frac{1}{N} + \frac{1}{k^2} \tag{8}$$

可以清楚地看到, 深度可分离卷积的参数和计算量只有标准卷积的 $\frac{1}{N} + \frac{1}{k^2}$ 倍, 这大大减少了模型中参数量和计算成本. 且深度可分离卷积同样具有标准卷积一样或更高的性能. 由于连续的池化操作会大量丢失信息, 降低图像分辨率, 深度可分离卷积一定程度上能缓解该现象. 因此本文将局部卷积部分的特征提取任务交给深度可分离卷积, 并引入残差机制. 具体结构如图 2 所示, 图 2(a) 为原始卷积结构, 图 2(b) 为本文中改进卷积模块, 除了将标准卷积替换为深度可分离卷积, 并且引入了残差机制减少信息丢失.



(a) 原始卷积特征提取结构

(b) 基于深度可分离和残差机制的特征提取结构

图 2 本文中的深度可分离卷积

1.3 融合多尺度特征的空间金字塔模块 ASPP

如图 3 所示, 多尺度特征的空间金字塔 ASPP[16]是

一种高效的融合策略. ASPP 的核心是空洞卷积 (atrous convolution), 用以扩大感受野, 捕获更广泛的上下文语信息. 具体的最佳空洞率配置由实验得出. ASPP 还包括一个全局平均池化分支, 这个分支旨在捕捉图像的全局上下文信息.

通过对特征图进行全局平均池化, 然后通过 1×1 卷积调整通道数, 最后将结果上采样到与原始特征图相同的尺寸. 这一步骤有助于模型理解整个图像的全局语义信息. 另外 ASPP 还包含一个逐点卷积分支, 这个分支用于处理空间信息, 调整通道数, 减少参数量或增加额外的非线性. 多尺度特征的空间金字塔模块 ASPP 可以增加对大小不一, 形状不规则对象特征的捕捉能力, 增强模型的鲁棒性, 比如在不同尺寸、角度、光照以及环境条件下的对象识别能力得到增强. 并且可以提高边界定位和类别判断能力, 尤其是边界区域和小目标的识别上. 其结构如图 3 所示, 过程如式 (9) 所示:

$$ASPP_{\text{out}} = Concat(AConv_{1,1}(X), AConv_{3,7}(X), AConv_{3,11}(X), AConv_{3,11}(X), GAP(X)) \tag{9}$$

其中, $ASPP_{\text{out}}$ 是融合后输出的特征张量, X 是处理的输入特征张量. $DConv_{i,j}$ 是空洞卷积, 其中 i, j 分别代表卷积核的大小和空洞率. 每个支路主要由空洞卷积进行特征提取, 空洞卷积最早是 Yu 等人[17]于 2015 年提出的. 主要是为了解决语义分割在下采样时空间信息丢失的问题.

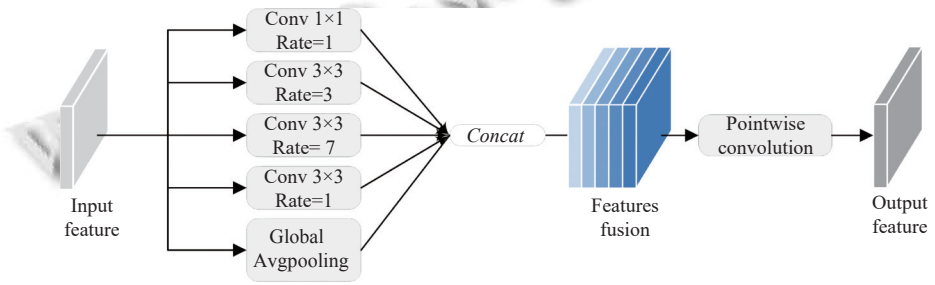
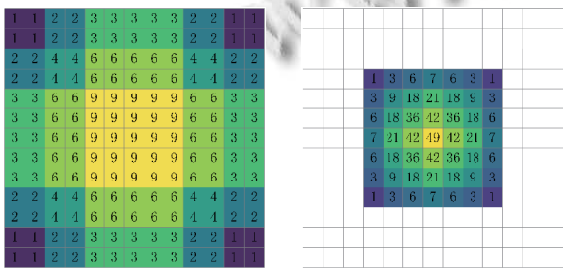


图 3 ASPP 尺度特征的空间金字塔模块

如图 4 所示, 显示了不同的空洞率对感受野的变化, 以及对比传统卷积的优势. 可以观察到, 在空洞卷积中, 由于卷积核元素之间插入了固定数量的空洞空间, 这种方法不会改变卷积核的实际尺寸, 但会扩大它的有效感受野, 使其能够在更大的区域上进行操作. 空洞卷积可以在不增加参数的情况下扩大感受野, 保持特征张量分辨率不变的同时, 获得了更多的空间信息.



(a) 3次空洞卷积后感受野
空洞率分别为 1, 2, 3

(b) 3次普通卷积
后感受野

图 4 膨胀卷积感受野变化

1.4 基于 ViT 的全局特征提取模块

与 CNN 卷积模块不同, Transformer[18]使用了自注意力方式, 可以有效地处理和分析序列. 最初 Transformer 主要用于处理自然语言 (NLP), 该模型用一系列分块作为输入, 并利用多头注意力机制, 在这些分块之间学习提取全局关系.

自注意力机制是 Transformer 的基础, 自注意力是每个特征张量而言, 分别乘上 3 个权重矩阵, W_Q, W_K, W_V 得到查询矩阵 (Q), 键矩阵 (K), 值矩阵 (V). 这些权重矩阵是通过训练可以学习的. 通过查询和键的相互作用计算注意力分数. 然后根据分数缩放, 并通过 *Softmax* 归一化得到最终的权重. 其公式如式 (2) 所示:

$$Atten(Q,K,V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{10}$$

其中, d_k 是键的维度.

Transformer 中实际使用了多头自注意力. 它是在自注意力机制的基础上演变而来的, 通过拆分成多个查询 (Q_i)、键 (K_i)、值 (V_i) 矩阵来增强对信息的利用. 在本文中, 将 Q_i 拆分为 $Q_{i,1}$–$Q_{i,12}$, 而 K_i 同样拆分为 $K_{i,1}$–$K_{i,12}$, V_i 拆分为 $V_{i,1}$–$V_{i,12}$. 每组 $Q_{i,j}$ 和 $K_{i,j}$ 送入注意力得到后经过 *Softmax* 计算后得到 $B_{i,j}$, 最终 $B_{i,j}$

合并得到自注意力中的 B_i.

Vision Transformer (ViT) 将 Transformer 迁移到图像任务中, 具体结构如图 5 所示.

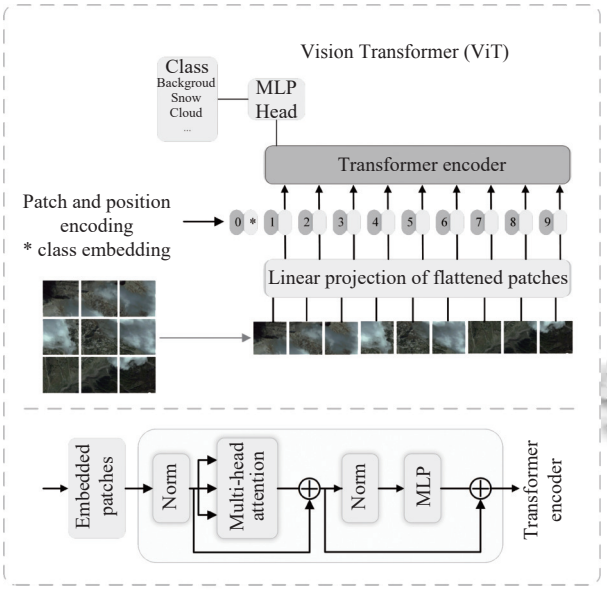


图 5 ViT 模块结构

基于位置编码的思想, ViT 首先需要将图片分割成多个小块, 然后将每个小块通过线性变换成固定长度. 对于图像中的每个块, 都会生成位置编码, 这个编码有固定长度, 并且与图像块的线性投影具有相同维度. 然后将位置编码添加到图像块的线性投影上, 这样每个图像块就包含了位置信息. 进而通过多头注意力机制计算. 与传统卷积相比, 可以更好地捕捉图像中远距离依赖关系, 更好理解全局中的上下文语意信息.

2 ViT-D-UNet 网络模型

本文是以带膨胀卷积模块的 D-UNet 作为基础网络, 在特征提取和解码阶段都进行了改进, 损失函数采用了联合损失函数. 具体的步骤如图 6, 实现如下描述.

(1) 编码阶段

编码阶段由两个支路构成, 分别提取局部特征的卷积支路和提取全局特征的 ViT 支路.

对于卷积支路, 将前 4 层传统卷积改为上述改进设计的深度可分离卷积, 降低参数的同时, 减少提取信息的丢失, 每次提取特征后, 特征张量分辨率减少一半. 对于 ViT 支路, 是一个并行分支, 处理同样的特征, 提取全局信息, ViT 支路首次提取特征降低分辨率为 1/4, 后续每次分辨率减少一半. 两支路之间存在特征交互, 通过拼接的方式叠加局部和全局的特征, 用卷积修改通道数以便后续处理. 对于卷积特征加至 ViT 支路, 使用全局平均池化以达到特征张量的分辨率匹配, 对于 ViT 支路叠加至卷积支路的, 使用双线性插值以达到特征张量的分辨率匹配.

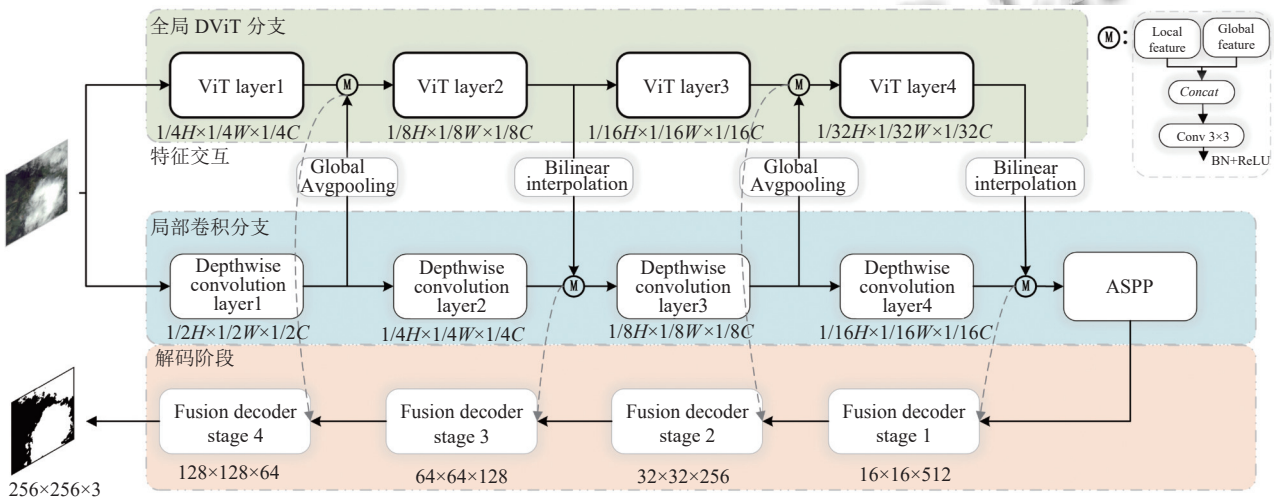


图 6 ViT-D-UNet 网络结构

(2) 解码阶段

在提取编码阶段的类别信息后, 需要通过解码器恢复原始图片大小. 为在解码阶段保持分类效果并提高像素精度, 避免由于上采样导致的模糊和信息丢失. 简单的双线性插值上采样会导致边缘粗糙和低准确度, 为此我们提出了融合特征的解码模块, 对应图 6 中的

Fusion Decoder. 通过改进的自注意力传递浅层特征, 由 Wang 等人[19]提出的 Non-local 模块利用图像中两点的相似性来权重特征, 以强化像素间的相关性. 本文改进自注意力模块使用 3×3 卷积来学习上下文信息以获取键, 然后将查询和上下文信息合并后进行学习, 接着使用两个连续的 1×1 卷积来提取局部信息. 这个操作可以只用于两个连续的像素间关系, 从而扩展到像素周围的语义, 以便它能够在全局级别掌握像素间的长距离依赖性, 使浅层特征具有更准确的位置信息. 图 7 展示了融合特征的解码模块, 使浅层特征 F_1 通过改进的自注意力, 得到具有增强位置信息的浅层特征 F_{11}.

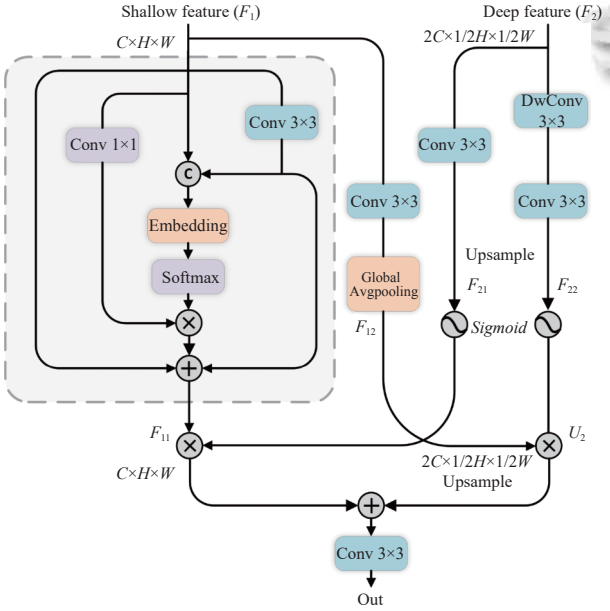


图 7 融合深浅特征的上采样模块

此外, 让深层特征 F_{22} 通过 3×3 卷积和双线性插值上采样改变其大小为 F_{12}, 与浅层特征一样, 然后通过 *Sigmoid* 函数进行权重, 相应元素与浅层特征相乘, 以获得具有增强分类能力的浅层特征 U_1. 浅层特征 F_{11} 通过 3×3 卷积和双线性插值上采样得到 U_1, 然后 F_{12} 乘以经过 3×3 卷积的深层特征 F_{22}, 经 *Sigmoid* 函数得到增强空间信息的深层特征 U_2. 最后, U_2 通过双线性插值和 3×3 卷积上采样, 乘以 U_1 得到的结果是最终输出 U_3, 可以通过对比特征融合和深层特征下采样得到. 具体过程表达式如下:

$$F_{11} = atten(F_1) \tag{11}$$

$$F_{12} = Avgpooling\{BN[Conv_{3\times3}(F_2)]\} \tag{12}$$

$$F_{21} = BN[Conv_{3\times3}(F_2)] \tag{13}$$

$$F_{22} = BN[Conv_{3\times3}(DWConv_{3\times3}(F_2)] \tag{14}$$

$$U_1 = F_{11}\times Sigmoid(F_{21}) \tag{15}$$

$$U_2 = F_{11}\times Sigmoid(F_{22}) \tag{16}$$

$$Out = ReLU(Conv_{3\times3}(U_1 + U_2)) \tag{17}$$

其中, $Conv_{3\times3}$ 是大小为 3 的卷积, *DWConv* 代表深度可分离卷积. *BN* 代表批量归一化操作, *ReLU* 代表激活函数.

2.1 联合损失函数

本文采用联合损失函数策略作为损失函数, 用 *Dice* 系数损失和交叉熵损失进行联合. 联合使用这两种损失函数可以在多个层面优化模型性能, 既考虑到了像素级别的分类准确性, 也重视了分割质量, 即预测和真实分割之间的重叠度. 这种联合损失函数有助于在细节上获得更精确的分割结果, 同时保持对大范围结构的高度敏, 具体描述公式如下:

$$L_B = \lambda L_B + (1-\lambda)L_D \tag{18}$$

其中, λ 为权重参数, L_B 为交叉熵损失函数 (binary cross entropy loss), L_D 是 *Dice* 系数损失函数. 二值交叉熵的基于信息论中的交叉熵概念, 用于衡量两个概率分布之间的差异. 在二分类问题中, 对于每个样本, 我们有一个预测概率和一个真实标签. 其表达式如下:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot \log(1-\hat{y}_i)] \tag{19}$$

其中, N 是样本的数量. $\hat{y}_i$ 是第 i 个样本的真实标签, $y_i = 1$ 表示正类, $y_i = 0$ 表示负类. log 是自然对数.

Dice 系数 (*Dice* coefficient), 也被称作 Sørensen-*Dice* 系数或者 *Dice* 相似性系数, 是一个用于衡量两个样本集合相似性的指标. 特别在图像分割处理领域, *Dice* 系数被用来作为损失函数, 帮助评估预测结果与真实结果之间的相似度. 二值交叉熵函数在样本不平衡情况下, 表现十分不好, 会导致训练效果下滑, 但这种情况下 *Dice* 系数损失函数表现优越. *Dice* 系数是基于两个集合的大小以及它们的交集大小. 给定两个集合 A 和 B, *Dice* 系数定义为:

$$Dice = \frac{2|A\cap B|}{|A|+|B|} \tag{20}$$

在实际应用中, A 和 B 可以表示为模型预测值和

真实标签的二值图像, 像素值 1 代表前景, 像素值 0 代表背景. 因此, *Dice Loss* 可以通过以下方式计算:

$$Dice\ Loss = 1 - \frac{2\times\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{21}$$

其中, N 是图像中像素的综述, p_i 是预测图像中第 i 个像素的值, g_i 是真实图像中第 i 个像素的值. 最后, 对于联合损失, 实验最终表明, 当 $\lambda = 0.68$, 即 $L = 0.68L_B + 0.32L_D$ 时, 分割效果最佳.

3 实验分析

3.1 HRC_WHU 数据集

HRC_WHU[20]高分辨率云覆盖数据集. 数据来源于武汉大学实验室, 由 150 张高分辨率遥感图像组成, 分辨率主要在 0.5–15 m, 原始尺寸为 1280×720. 地貌包括了植被, 雪, 沙漠, 城市以及水面. 将图像裁剪为 224×224 的子图进行训练. 最终得到了 3 600 张图片. 并以 8:2 的比例分组, 分别作为训练集和验证集.

3.2 自建云影数据集

本数据集遥感图像主要由美国的 Landsat 卫星拍摄, 以及收集自谷歌地球 (GE) 中选出的高分辨率遥感图像. Landsat8 卫星携带了 9 个波段的陆地成像仪, 该数据集主要使用了 2 号蓝波段 (0.450–0.515 μm), 3 号绿波段 (0.525–0.600 μm), 4 号红波段 (0.630–0.680 μm). 将原始图片统一裁剪为 224×224 分辨率以方便训练. 最终得到 10 843 张图片, 并以 8:2 的比例对图片分组, 分别作为训练集和验证集

3.3 实验及评估参数

所有的实验工作均是 PyTorch 框架实现, 版本为 1.10.1. 显卡使用英伟达 RTX2080Ti, 显存为 11 GB. 训练两个数据集的 batchsizes 均设为 16, 训练周期为 300 轮, 优化器为 Adam. 学习率采用等间隔学习率 (StepLR), 初始学习率为 0.001, 衰减系数为 0.9, 每 3 轮更新学习率, 学习率计算公式如下:

$$lr_N = lr_0 \cdot \beta^{N/s} \tag{22}$$

其中, lr_N 为第 N 次训练的学习率, lr_0 为初始学习率, β 为衰减系数, s 为更新间隔次. 损失函数使用交叉熵损失函数作为训练损失函数, 公式如下:

$$\begin{aligned} Loss(x,\ class) &= -\log\left(\frac{e^{x[clas]}}{\sum_i e^{x[i]}}\right) \\ &= -x[class] + \log\left(\sum_i e^{x[i]}\right) \end{aligned} \tag{23}$$

在评估模型的性能时, 我们使用像素精度 (*PA*), 平均像素进度 (*MPA*) 以及平均交并集 (*MIoU*) 等指标评价模型表现, 它们的计算公式如下:

$$PA = \frac{\sum_{i=0}^{k} p_{ij}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \tag{24}$$

$$MPA = \frac{1}{k}\sum_{i=0}^{k} \frac{p_{ij}}{\sum_{j=0}^{k} p_{ij}} \tag{25}$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{26}$$

其中, k 为类别数; 而 p_{ij} 为第 i 类像素被预测为第 j 类像素的数量; p_{ji} 为第 j 类像素被预测为第 i 类的像素数量; p_{ii} 为第 i 类像素, 并预测为 i 类像素的数量.

3.4 实验评价

为了评估所提出的网络性能, 我们在自建数据集上进行了对比实验, 将我们的网络与目前主流优秀的语义分割模型进行了对比实验. 另外, 我们也比较了一些最新为遥感设计的网络, 如 Dual-branch network (DBNet). 我们的网络具有最好的性能, 如表 1 中结果所示, 在 *MIoU* 上领先次优网络 0.52%.

表 1 在自建数据集上性能表现 (%)

Model	Class pixel accuracy			Overall results		
	Cloud	Shadow	Background	*PA*	*MPA*	*MIoU*
FCN	94.32	92.11	95.18	94.24	94.12	88.31
U-Net	94.21	90.53	94.43	93.25	92.32	88.12
PSPNet[21]	94.73	92.52	95.69	**95.67**	94.42	91.53
CloudNet[22]	95.03	91.26	95.43	94.82	94.26	89.18
DABNet[23]	95.12	92.85	95.35	95.45	94.43	91.34
D-UNet	94.98	92.52	95.78	95.12	94.24	91.11
DBNet[24]	94.47	92.23	95.48	94.74	94.32	90.52
本文方法	**95.45**	**93.37**	**96.25**	95.51	**95.15**	**92.05**

我们的模型在针对不同层次的特征采取不同策略以适应性的提取多尺度特征. 在提取特征阶段, 卷积分

支可以很好地提取局部多尺度特征. 而 ViT 分支可以很好地提取到全局特征和抽象语义信息, 通过设计的融合特征的解码模块可以很好地结合局部特征和全局特征, 并进行上采样, 减少信息丢失, 增加模型的鲁棒性和泛化性能.

从分割的对比图 8 中可以看出, 我们的网络相比较新的 DBNet 网络在特征模糊的区域准确度更高, 对边界的界定更加准确. 而对于不同尺寸, 不规则的小目标依然有良好的性能. 相对于其他网络, 无论是误检和漏检更少, 对细小目标和不规则目标检测精度更高.

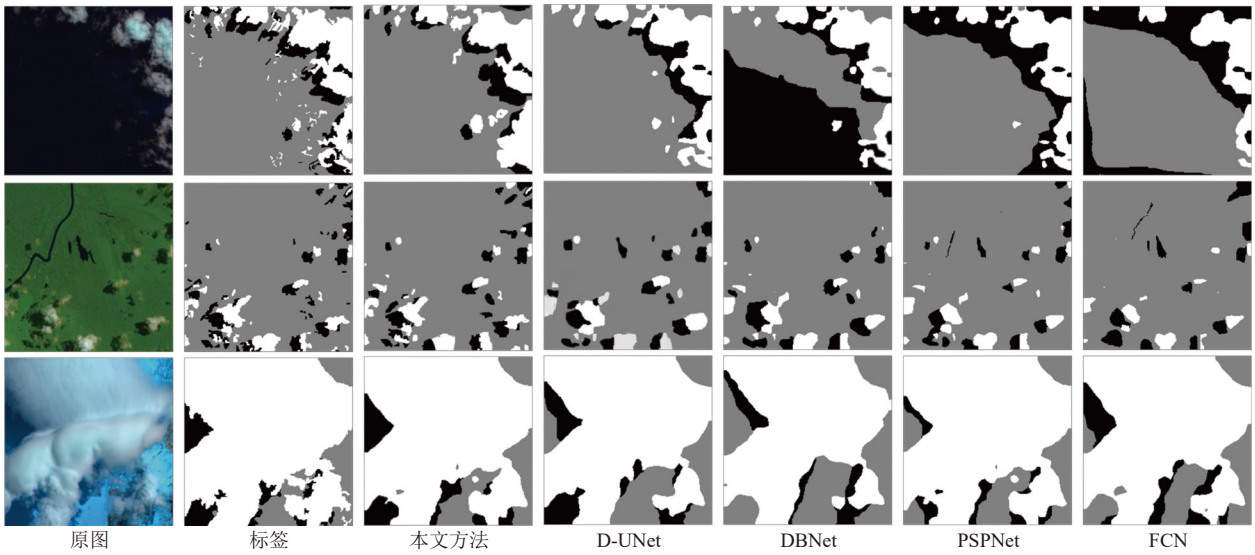


图 8 在自建数据集上分割效果

为了进一步探究模型的性能, 验证我们模型的泛化性, 我们还在 HRC_WHU 数据集上做了对比实验. HRC_WHU 数据集拥有和自己建数据集风格不一致的特征数据. 比如强光下的沙地拥有和稀薄云层相似的特征, 比较容易混淆, 而模型应然能准确地分割背景和目标的边缘, 提取丰富特征. 我们依然与一些流行的模型进行了对比. 实验结果如表 2 所示.

表 2 在 HRC_WHU 数据集上性能表现 (%)

Model	*PA*	*MPA*	*MIoU*
DenseASPP[16]	89.43	87.39	80.56
ENet[25]	89.62	88.16	81.15
CloudSegNet[26]	89.93	89.17	82.57
PVT[27]	89.61	89.27	82.61
DeepMask[28]	89.87	88.41	82.89
ACFNet[29]	90.23	89.44	83.21
DeepLabv3+[30]	91.43	90.27	84.93
本文方法	92.11	91.46	85.37

从结果上看, 我们的模型在 HRC_WHU 数据集上仍然能保持最优性能. 比次优模型在 *MIoU* 指标上仍然领先了 0.44%. 针对该数据, 我们主要比较了在云特征相似的图片上的效果, 以及模糊较难提取特征的图片.

如图 9 所示, 对于第 1 行, 我们的云和雪特征有相似特征, 它们的边缘特征略有差异. 在识别时, 得益于丰富的局部和全局特征的全面理解, 我们的模型在大致轮廓上识别准确, 没有散雪被识别成云, 而对比网络对相似特征的界定进度大大下降, 对于右边模糊的雪甚至出现了很多误检. 对于第 2 行, 我们选取的遥感图像中云边界难以界定, 是一张厚云过度消散为薄云的图片. 从对比图 9 中可以观察到, 我们的网络对边界的界定仍然保持一定的精准度, 甚至没有错过左下角小目标. 然而对比网络出现了比较大的误检以漏检.

3.5 消融实验

为了验证各个模块的有效性, 本节使用了两个数据集的数据作为训练集训练. 探究各个模块的有效性. 将网络 ViT-D-UNet 的 ViT 分支减去, 记为 ViT-D-UNet/ViT$^-$, 将本文网络的 ASPP 模块替换为普通卷积记为 ViT-D-UNet/ASPP$^-$, 将网络的深度可分离卷积替换为普通卷积模块记为 ViT-D-UNet/DWConv$^-$. 而将融合多尺度的上采样替换为双线性插值上采样, 记为 ViT-D-UNet/Decoder$^-$, 进行性能评估. 根据表 3 显示, 在合并的大数据上, 各个模块也能显著提示模型的性能效果.

图 9 在 HRC_WHU 数据集上分割效果

表 3 在 HRC_WHU 数据集上性能表现 (%)

Model	*ACC*	*MPA*	*MIoU*
ViT-D-UNet/ViT$^-$	89.24	89.78	86.21
ViT-D-UNet/ASPP$^-$	90.31	90.19	87.03
ViT-D-UNet/DWConv$^-$	90.30	91.11	87.47
ViT-D-UNet /Decoder$^-$	89.42	90.73	86.41
ViT-D-UNet	**91.15**	**92.35**	**88.49**

4 总结

本文提出了一种高效的双支路云影检测网络. 在卷积支路内部署深度可分离卷积与空间金字塔模块, 使模型能够在局部特征层面上捕获丰富的多尺度信息, 同时相较于普通卷积操作, 有效减少了参数数量. 在全局支路中, 采用 ViT 模块对全局特征及抽象特征进行深度提取. 两个支路的交互促进了信息的完善和检测目标的细化. 最后使用融合深层与浅层特征的上采样模块, 实现对图像分辨率的恢复, 减少信息的丢失. 在 HRC_WHU 公开数据集和自建云影数据集上的对比实验验证了本网络的良好性能. 相较于其他模型, 本网络不仅聚焦于多尺度特征的捕获, 还注重局部与全局特征的提取与整合, 因此对边缘以及模糊和抽象特征都达到了精确的分割效果. 未来的研究将探索自监督学习方法的引入, 以降低对数据集标注的依赖, 进一步提升检测方法的自动化水平.

参考文献

1 Moses WJ, Philpot WD. Evaluation of atmospheric correction using bi-temporal hyperspectral images. Israel Journal of Plant Sciences, 2012, 60(1–2): 253–263.

2 Tapakis R, Charalambides AG. Equipment and methodologies for cloud detection and classification: A review. Solar Energy, 2013, 95: 392–430. [doi: 10.1016/j.solener.2012.11.015]

3 Zhu Z, Woodcock CE. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sensing of Environment, 2012, 118: 83–94. [doi: 10.1016/j.rse.2011.10.028]

4 Qiu S, He BB, Zhu Z, *et al*. Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4–8 images. Remote Sensing of Environment, 2017, 199: 107–119. [doi: 10.1016/j.rse.2017.07.002]

5 Zhu Z, Woodcock CE. Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. Remote Sensing of Environment, 2014, 152: 217–234. [doi: 10.1016/j.rse.2014.06.012]

6 Li S, Wang M, Wu J, *et al*. CloudDeepLabV3+: A lightweight ground-based cloud segmentation method based on multi-scale feature aggregation and multi-level attention feature enhancement. International Journal of Remote Sensing, 2023, 44(15): 4836–4856. [doi: 10.1080/01431161.2023.2240034]

7 Wang ZW, Xia M, Lu M, *et al*. Parameter identification in power transmission systems based on graph convolution network. IEEE Transactions on Power Delivery, 2022, 37(4): 3155–3163. [doi: 10.1109/TPWRD.2021.3124528]

8 Ayala C, Sesma R, Aranda C, *et al*. A deep learning approach to an enhanced building footprint and road detection in high-resolution satellite imagery. Remote Sensing, 2021, 13(16): 3135. [doi: 10.3390/rs13163135]

9 Prathap G, Afanasyev I. Deep learning approach for building detection in satellite multispectral imagery. Proceedings of the 2018 International Conference on Intelligent Systems (IS). Funchal: IEEE, 2018. 461–465.

10 Xie WY, Fan XY, Zhang X, *et al*. Co-compression via superior gene for remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 5604112.

11 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440.
12 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Proceedings of the 18th International Conference on Medical Image Computing and Computer-assisted Intervention. Munich: Springer, 2015. 234–241.
13 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]
14 龙丽红, 朱宇霆, 闫敬文, 等. 新型语义分割 D-UNet 的建筑物提取. 遥感学报, 2023, 27(11): 2593–2602.
15 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.
16 Yang MK, Yu K, Zhang C, *et al*. DenseASPP for semantic segmentation in street scenes. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3684–3692.
17 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. Proceedings of the 4th International Conference on Learning Representations. San Juan, 2016.
18 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.
19 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803.
20 Li ZW, Shen HF, Cheng Q, *et al*. Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. ISPRS Journal of Photogrammetry and Remote Sensing, 2019, 150: 197–212. [doi: 10.1016/j.isprsjprs.2019.02.017]
21 Zhao HS, Shi JP, Qi XJ, *et al*. Pyramid scene parsing network. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6230–6239.
22 Mohajerani S, Saeedi P. Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery. Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama: IEEE, 2019. 1029–1032.
23 Li G, Kim J. DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. Proceedings of the 30th British Machine Vision Conference 2019. Cardiff: BMVA Press, 2019.
24 Lu C, Xia M, Qian M, *et al*. Dual-branch network for cloud and cloud shadow segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5410012.
25 Paszke A, Chaurasia A, Kim S, *et al*. ENet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147, 2016.
26 Dev S, Nautiyal A, Lee YH, *et al*. CloudSegNet: A deep network for nychthemeron cloud image segmentation. IEEE Geoscience and Remote Sensing Letters, 2019, 16(12): 1814–1818. [doi: 10.1109/LGRS.2019.2912140]
27 Wang WH, Xie EZ, Li X, *et al*. Pyramid vision Transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the 2021 IEEE/CVF international Conference on Computer Vision. Montreal: IEEE, 2021. 548–558.
28 Xu K, Guan KY, Peng J, *et al*. DeepMask: An algorithm for cloud and cloud shadow detection in optical satellite remote sensing images using deep residual network. arXiv:1911.03607, 2019.
29 Zhang F, Chen YQ, Li ZH, *et al*. ACFNet: Attentional class feature network for semantic segmentation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6797–6806.
30 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 833–851.

(校对责编: 张重毅)