第 37 卷,第 1 期 自 然 资 源 遥 感 Vol.37,No.1
2025 年 02 月 REMOTE SENSING FOR NATURAL RESOURCES Feb.,2025

doi:10.6046/zrzyyg.2023237
引用格式:曲海成,王莹,刘腊梅,等.融合 CNN 与 Transformer 的遥感影像道路信息提取[J].自然资源遥感,2025,37(1):38-45.(Qu H C,Wang Y,Liu L M, et al.Information extraction of roads from remote sensing images using CNN combined with Transformer [J].Remote Sensing for Natural Resources, 2025,37(1):38-45.)

融合 CNN 与 Transformer 的遥感影像道路信息提取

曲海成, 王 莹, 刘腊梅, 郝 明
(辽宁工程技术大学软件学院,葫芦岛 125105)

摘要:利用高分辨率遥感影像进行道路信息提取时,深度神经网络很难同时学习影像全局上下文信息和边缘细节信息,为此,该文提出了一种同时学习全局语义信息和局部空间细节的级联神经网络。首先将输入的特征图分别送入到双分支编码器卷积神经网络(convolutional neural networks,CNN)和 Transformer 中,然后,采用了双分支融合模块(shuffle attention dual branch fusion block,SA-DBF)来有效地结合这 2 个分支学习到的特征,从而实现全局信息与局部信息的融合。其中,双分支融合模块通过细粒度交互对这 2 个分支的特征进行建模,同时利用多重注意力机制充分提取特征图的通道和空间信息,并抑制掉无效的噪声信息。在公共数据集 Massachusetts 道路数据集上对模型进行测试,准确率(overall accuracy,OA)、交并比(intersection over union,IoU)和 F_1 等评价指标分别达到 98.04%,88.03%和 65.13%;与主流方法 U-Net 和 TransRoadNet 等进行比较,IoU 分别提升了 2.01 个百分点和 1.42 个百分点,实验结果表明所提出的方法优于其他的比较方法,能够有效提高道路分割的精确度。

关键词:级联神经网络;Transformer;特征融合;注意力机制
中图法分类号:TP 79 **文献标志码**:A **文章编号**:2097-034X(2025)01-0038-08

0 引言

高分辨率遥感影像包含丰富的地物信息[1]。从遥感影像中提取道路信息可以应用于许多领域,如城市规划[2]、自动驾驶[3]、道路信息更新等。深度学习中的语义分割[4]技术会对图像中的每个像素进行分类,将图像分为目标和背景。通过语义分割技术提取道路信息已成为遥感影像道路提取的主流方法[4-5]。当前,卷积神经网络(convolutional neural networks,CNN)[6]作为一种强大的深度学习模型,被广泛应用于图像处理领域。通过对大量标记的遥感影像数据进行训练,CNN 能够学习到特征表示和语义信息,从而实现比较准确的道路信息提取。此外,还有一些基于图像分割的经典算法,如全卷积网络[7](fully convolutional networks,FCN),但 FCN 利用反卷积进行上采样操作时,分割结果受限于局部的感受野中,无法有效地捕获上下文语义信息。因此,出现了许多经典语义分割方法,包括 U-Net[8],SegNet[9],DeepLabV3+[10]等,也在道路提取任务中取得了一定的成果。其中语义分割网络大多是编码器-解码器[11]网络结构,其利用下采样和上采样来捕捉上下文信息并进行精确定位,以恢复空间信息,但是在下采样期间会丢失空间信息。针对上述的问题,提出了许多对语义分割网络进行改进或者变体的网络,应用于遥感影像的道路分割。Gao 等[12]提出改进的编码器-解码器网络,在编码器部分使用连续的卷积进行道路特征提取,使其具有较强的局部信息提取能力。虽然上述方法对遥感影像的道路提取效果较好,但传统的卷积运算忽略了各个维度之间的依赖性,并且在遥感影像中还存在建筑遮挡道路区域、地形复杂等问题,对道路进行提取仍然是一项具有挑战性的任务。王勇等[13]提出结合注意力机制对重要的位置信息和空间结构进行有效捕捉,来提高道路提取的准确性;吴强强等[14]提出空间信息感知语义分割模型用于遥感影像道路的提取,引用坐标卷积和全局信息模块,结果显示该方法对复杂区域的道路提取效果不佳。随着神经网络的发展,基于自注意力机制(self-attention)的 Transformer[15]网络出现在大众面前,其核心

收稿日期:2023-08-02;**修订日期**:2024-05-09
基金项目:国家自然科学基金面上项目"面向数据特性保持的高光谱影像高效压缩方法研究"(编号:42271409)和辽宁省高等学校基本科研项目"基于全脉冲混合神经网络的高效能目标检测"(编号:LIKMZ20220699)共同资助。
第一作者:曲海成(1981-),男,博士,副教授,主要研究方向为遥感图像高性能计算、智能大数据处理等。Email:quhaicheng@lntu.edu.cn。
通信作者:王 莹(1998-),女,硕士研究生,主要研究方向为数字图像处理与模式识别。Email:lntuwangying@163.com。

思想是通过自注意力机制[16]来建立输入序列中各个位置之间的依赖关系。Transformer 使用注意力机制来对序列中的每个位置进行建模,从而实现了并行化计算和长程依赖的建模能力。之后,许多受到 Transformer 启发的网络模型出现在视觉任务中。Dosovitskiy 等[17]提出 vision Transformer(ViT),ViT 将图像作为保留位置信息的非重叠序列块,并使用自注意力机制来构建上下文信息;Yang 等[18]基于 CNN 结合高层语义特征和前景上下文信息对遥感影像中的道路进行提取,提高了对遮挡区域的推断能力和对中心感受野不足的问题,但缺少对道路的空间和位置信息的关注;Dai 等[19]提出的 CoAtNet 网络结合了 CNN 和自注意力机制的优点,也相对提高了道路分割精度,但其不能专注于重要特征信息的提取。

针对 CNN 和 Transformer 在遥感道路分割任务中存在的问题,本文提出了一个双分支级联神经网络,将学习局部高级特征的 CNN 编码器与捕获全局多尺度信息的 Transformer 编码器利用双分支融合模块(shuffle attention dual branch fusion block,SA-DBF)将双分支结合在一起,使全局信息与局部信息得到有效融合并关注到重要信息,进而提高道路的分割精度。

1　本文方法

本文网络整体结构由高效捕获局部特征和深层特征的 CNN 编码器、捕获全局特征的 Transformer 编码器和用于分割的解码器组成,通过 SA-DBF 模块将 CNN 和 Transformer 融合在一起。该方法能够更好地学习遥感影像中所需的道路信息,网络结构如图 1 所示,其中融合模块的主要工作原理为,将通道维度并行处理之前,首先将通道分成多个子特征组,对于每一组子特征,SA-DBF 模块使用置换单元(shuffle unit)来刻画在空间和通道上的特征依赖关系,所有的子特征聚合之后,再使用通道打乱操作加强不同子特征之间的信息交流。这种网络结构能够有效地捕捉全局上下文信息和多尺度特征,并实现对道路的精准提取。

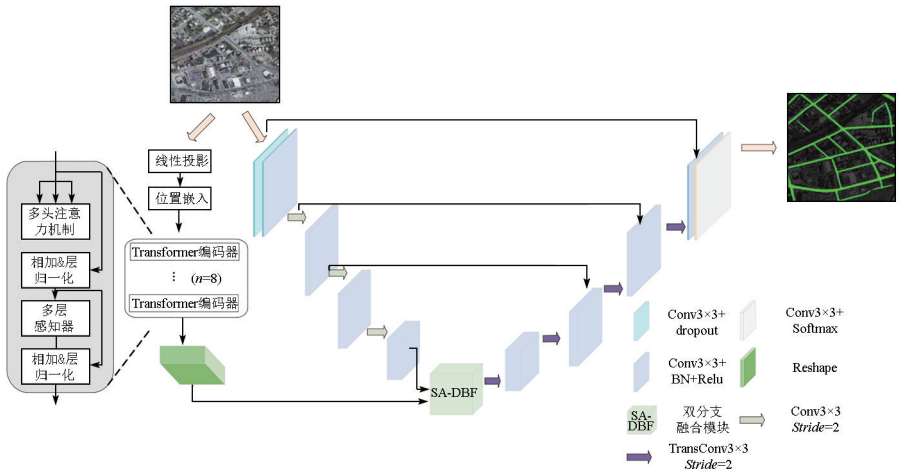


图 1　本文模型整体结构

Fig.1　Overall structure of the model in this paper

1.1　CNN 编码器

CNN 编码器是在图像处理和计算机视觉任务中一种用于提取局部特征的常用工具,可通过多层堆叠的方式逐渐提取更高级别的特征。具体来说,给定一个输入为 $X \in \mathbf{R}^{C\times H\times W}$ 的图像,其中 H 和 W 分别为图像的高度和宽度,C 为通道数,通过一个卷积层和 Dropout 层对输入特征信息进行初始化的操作;然后用 4 个阶段的残差块对图像进行特征提取,4 个阶段的残差块数量分别为 1,2,2,4;下采样过程中采用步长为 2 的卷积代替池化操作,最终得到 128 × 16 × 16 大小的特征图。

1.2　Transformer 编码器

由于卷积运算固有的局部性,CNN 编码器不能有效地捕获输入体素的远距离相关性。为此,使用 Transformer 编码器进行全局上下文建模。首先,将输入特征信息 $X \in \mathbf{R}^{C\times H\times W}$ 重塑为均匀不重叠的块 $X \in \mathbf{R}^{N(P^2\cdot C)}$,其中$(P,P)$为每个块的分辨率,$N=\dfrac{H\times W}{P^2}$为输入序列的长度。使用线性层将块投影到在整个 Transformer 层中保持不变的 K 维嵌入空间中,然后为了对空间信息进行编码,增加了一维可学习的位置嵌入并将其添加到块嵌入中以保留位置信

息。该公式可以表示为：

$$Z_0 = [x^1E, x^2E, \cdots, x^NE] + E_{\text{pos}} \quad (1)$$

式中：$E \in \mathbf{R}^{(P^2\cdot C)\times K}$ 为块嵌入投影；$E_{\text{pos}} \in \mathbf{R}^{N\times K}$ 为位置嵌入。

最终，Transformer 编码器由 L 层多头自注意力（multi-head attention，MHA）和多层感知器（multi-layer perceptron，MLP）块组成，每个子层中采用加性跳跃连接策略来避免梯度消失。因此，第 i 层的输出可以表示成：

$$Z'_i = MHA(LN(Z_i - 1)) + Z_{i-1},\ i = 1,2,\cdots,L \quad (2)$$

$$Z_i = MLP(LN(Z'_i)) + Z'_i,\ i = 1,2,\cdots,L \quad (3)$$

式中：Z_i 和 Z'_i 为第 i 层的输出；$LN()$ 为层归一化；MHA 为多头自注意力；MLP 由具有 GELU 激活函数的 2 个线性层组成；i 为中间块标识符；L 为 Transformer 的层数。

Transformer 编码得到输出特征图后应用一个反卷积层使其与 CNN 编码得到的输出特征图大小相同，通过 SA-DBF 模块将 Transformer 和 CNN 编码阶段学习到的全局信息与局部信息融合。

1.3 SA-DBF 模块设计

为了有效地结合 CNN 和 Transformer 分支的编码特征，提出了一种新的融合模块——SA-DBF 模块，其结构图如图 2 所示。SA-DBF 模块首先将输入沿着通道维度拆分为 G 个组，然后对每一组特征词分为 F_X 和 S_X，拆分后的特征分别利用嵌入平均池化 F_{gp} 和组归一化的操作来生成新的特征，再通过 $F_C()$ 来增强特征的表示，$F_C = W_i + b_i$，其中 W 和 b 为参数，$i=1,2$。用置换单元刻画特征在空域与通道维度上的依赖性，最后将新得到的特征 $\hat{F}_X$ 和 $\hat{S}_X$ 进行集成并通过通道置换（channel shuffle）操作进行组件特征通信。此模块实现了全局信息与局部信息的融合，弥补了 CNN 与 Transformer 只能分别关注单一特征的不足，并通过注意力机制实现了目标特征增强和噪声抑制的目的。

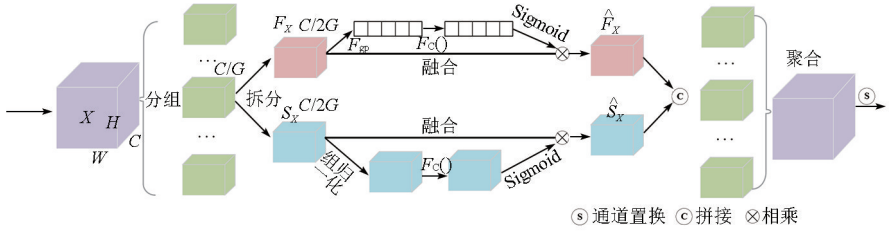


图 2 SA- DBF 模块结构图

Fig.2 SA- DBF module structure diagram

1）特征分组。将输入特征图分为多组，每组为一个子特征（sub-feature）。对于给定的特征图 $X \in \mathbf{R}^{C\times H\times W}$，其中 C，H，W 分别是通道数、空间高度和宽度。首先沿着通道维度将 X 划分为 G 个组，即 $X = [X_1, \cdots, X_G]$，$X_q \in \mathbf{R}^{(CG)\times H\times W}$ 在训练过程中每个子特征 X_q 逐渐地获取一个语义响应，然后通过注意力模块，为每个亚特征都生成一个相应的重要系数。这部分对应上图最左边的“分组”部分。

2）混合注意力。在每个注意力单元开始时刻，X_q 的输入会沿着通道维度，被分为 2 个分支，即 $X_{q1}, X_{q2} \in \mathbf{R}^{\frac{C}{2G}\times H\times W}$，如上图中间“拆分”后的部分，2 个分支分别用不同颜色表示，F_X 分支利用通道间依赖生成通道注意力图，S_X 分支则捕获特征之间的空间依赖生成空间注意力图，这样，模型同时完成了语义和位置信息的注意。对于通道注意力分支，首先通过全局平均池化（global average pooling，GAP）来生成通道统计数据 $S \in \mathbf{R}^{C/2G\times 1\times 1}$ 嵌入全局信息，可以沿着空间维度 $H \times W$ 收缩 X_{q1} 计算得到，公式为：

$$S = F_{\text{gp}}(X_{q1}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{q1}(i,j) \quad (4)$$

此外，用 Sigmoid 激活函数来创建一个紧致特征，从而准确地、自适应地选择。通道注意力的最终输出为：

$$X'_{q1} = \sigma(F_C(S)) \cdot X_{q1} = \sigma(W_1 S + b_1) \cdot X_{q1} \quad (5)$$

式中：δ 为 Sigmoid 函数；$W_1 \in \mathbf{R}^{C/2G\times 1\times 1}$ 和 $b_1 \in \mathbf{R}^{C/2G\times 1\times 1}$ 用于缩放和平移 S。空间注意力与通道注意力不同，是通道注意力的补充。首先，对 X_{q2} 使用组归一化获取空间统计数据。然后使用 $F_C()$ 来增强 X_{q2} 的特征表示。最终的空间注意力输出为：

$$X'_{q2} = \sigma(W_2 \cdot GN(X_{q2}) + b_2) \cdot X_{q2} \quad (6)$$

式中：$GN()$ 为组归一化；W_2，b_2 为形状为 $\mathbf{R}^{C/2G\times 1\times 1}$ 的参数。然后 2 个注意力的结果会被拼接起来，即 $X'_q = [X'_{q1}, X'_{q2}] \in \mathbf{R}^{C/2G\times H\times W}$ 此时与该组的输入尺寸保持一致。

3）特征聚合。所有的特征会被聚合起来，最终与 ShuffleNet v2 相似，本文采用了通道置换的操作，沿着通道维度实现跨组信息交流。空间注意力模块的最终输出与 X 的形状相同，使得 空间注意力模块可以很容易地集成到其他网络中。需要说明的是，

W_1 , b_1 , W_2 , b_2 和 $GN(\,)$的超参数只是空间注意力模块中的参数。在单个空间注意力模块中,每个分支通道的个数是 $C/2G$ 。因此,所有的参数个数就是$\frac{3C}{G}$,通常 G 是 32 或 64,这和网络数以百万计的参数相比微乎其微,因此 SA-DBF 非常轻量。空间注意力模块和通道注意力模块结构如图 3 所示。

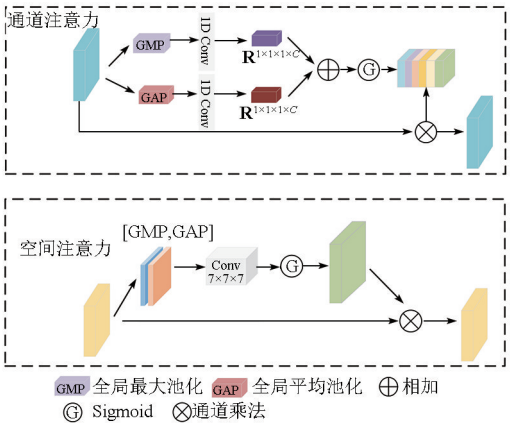


图 3　空间注意力模块和通道注意力模块结构图

Fig.3　SA module and CA module structure diagram

1.4　解码器

解码器是 CNN 架构负责将特征转化目标,使用转置卷积将特征图逐步上采样到输入分辨率 $H \times W$,并且在上采样过程中使用残差块细化特征图。最后,使用 1×1 卷积和 Softmax 激活函数将特征映射为概率分割结果。此外,编码器和解码器之间使用跳跃连接实现浅层信息与深层信息的融合,为解码过程提供更多语义信息,进而提高道路的分割精度。

2　实验及结果分析

2.1　实验环境及参数

1)硬件环境:显卡为 NVIDIA GeForce RTX 3090, Intel(R) Xeon(R) Gold 6330 CPU @ 2.00 GHz处理器,内存为 360 G 内存。

2)软件环境:使用的计算机操作系统为 Ubuntu18.04;使用深度学习框架 PyTorch 进行训练;使用 Python 编程语言进行编写。

3)参数设置:总迭代次数为 100;迭代批量为 8;优化器使用 Adam 随机优化算法更新网络参数,初始学习率为 4E-4 并随每次迭代以 0.9 的幂次进行衰减,同时使用权重衰减率为 1E-5 的 L_2范数进行正则化。

2.2　数据集

本文选用 Massachusetts 遥感道路数据集进行训练和评估所提出的方法。Massachusetts 遥感道路数据集由 1 171 张遥感影像组成,地面分辨率为 1.2 m,涵盖了城市、郊区和农村地区。在 Massachusetts 遥感道路数据集中,将 1 108 张、49 张和 14 张遥感影像分别作为训练集、测试集和验证集。

2.3　数据增强

由于 Massachusetts 遥感道路数据集数据量较少,对其中的图片通过旋转 90°,180°,270°,水平、垂直镜像翻转等方式进行数据增强。最后以 512 为步长制作生成可训练的遥感道路数据集,在测试阶段,使用测试时间增强技术进一步提高模型的性能,该技术已得到验证。最终得到 9 963 张训练集遥感影像,567 张测试集和验证集遥感影像。

2.4　评价指标

2.4.1　总体评价指标

为了验证本文所提方法的有效性,采用了 3 种用于评估遥感影像分割效果的评价指标来衡量模型的性能,包括总体精度(overall accuracy, OA)、F_1 得分和交并比(intersection over union, IoU)。公式分别为:

$$OA = (TP + TN)/(TP + TN + FN + FP)\ , \tag{7}$$

$$F_1 = 2P \cdot R/(P + R)\ , \tag{8}$$

$$P = TP/(TP + FP)\ , \tag{9}$$

$$R = TP/(TP + FN)\ , \tag{10}$$

$$IoU = TP/(TP + FP + FN)\ , \tag{11}$$

式中:TP 为正确预测为道路的像素数;FP 为将背景错误预测为道路的像素数;FN 为将道路错误预测为背景的像素数;TN 为正确预测非道路像素数目;P 为精确率;R 为召回率。

2.4.2　像素损失评价指标

实验标签为背景和道路,Dice 损失函数把一个类别的所有像素作为一个整体,并计算 2 个类别的交集在整体中的比例,所以不受大量背景像素的影响,在样本不平衡的情况下可以达到更好的效果;Focal 损失函数为一个动态缩放的交叉熵损失,通过一个动态缩放因子,可以动态降低训练过程中易区分样本的权重,从而将重心快速聚焦在那些难区分的样本。因此,结合 Dice 损失函数和 Focal 损失函数的优点,本文设计一种复合型的损失函数 L,定义如下:

$$L = L_{\text{Focal}} + L_{\text{Dice}}\ , \tag{12}$$

$$L_{\text{Focal}} = -\,a\,(1-p)^{\lambda}\ln(p)\ , \tag{13}$$

$$L_{\text{Dice}} = 1 - \frac{\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{14}$$

式中：N 为像素总数；g_i 表示像素 i 的真实标签值；p_i 表示像素 i 的预测值；a 为类别权重，用来衡量正负样本不均衡问题；λ 为难以判断的样本权重，用来衡量难分样本和易分样本。当 $a=1$ 时，表示类别的权重相等；当 $a<1$ 时，表示正样本赋予更高的权重，以增加正样本的影响力，从而应对不平衡类别分布；当 $a>1$ 时，表示对负样本赋予更高的权重，以减少正样本的影响力，用于降低对背景类别的关注。通常 $\lambda=0$ 时，即所有样本权重相等，不区分难易样本；λ 接近 0 时，对难以判断的样本引入轻微的权重，但相对均匀的处理样本；λ 较大时，更强烈地关注难以判断的样本，以便训练模型更好地处理难以不均衡问题。

实验损失率变化趋势如图 4 所示，本实验训练和验证的 Epoch 为 150，由图可以看出训练和验证损失在 Epoch 小于 20 时下降较快，验证损失小于 80 时震荡稍不稳定，训练损失保持平滑下降，最后收敛在 0.1 左右，收敛较好。

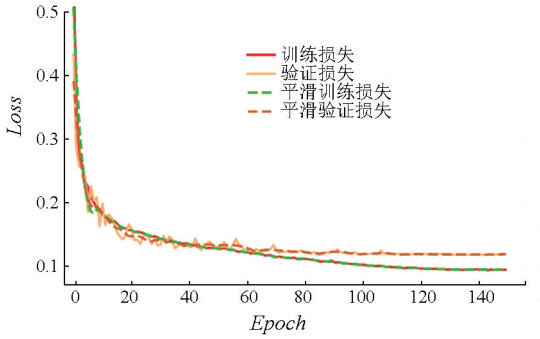


图 4 实验损失率变化趋势

Fig.4 Change Trend of Experimental Loss

2.5 实验结果和分析

本节讨论消融实验搭建的模型和本文改进的模型，在 Massachusetts 道路数据集上的实验结果及与主流方法的对比分析。

2.5.1 消融实验

通过引入 CNN 和 Transformer 融合的级联神经网络、SA-DBF 融合模块对 U-Net 模型进行改进，并利用改进后的模型对遥感影像中的道路进行分割，共做了 4 组消融实验，验证各个改进点的有效性，实验结果见表 1。表中，Transformer+ U-Net 模型是在 U-Net 的基础上引入了 Transformer 结构；U-Net+SA-DBF 模块模型是在 U-Net 模型中引入了 SA-DBF 模块；Transformer +U-Net+SA-DBF 模型是以上 3 种网络结构的结合，加粗字体为最优结果。

表 1 不同模块消融实验的对比结果

Tab.1 Comparison results of different modules (%)

方法	*OA*	F_1	*IoU*
U-Net	96.39	84.12	63.12
Transformer+ U-Net	97.08	85.97	64.01
U-Net+SA-DBF	96.27	86.36	64.37
Transformer +U-Net+SA-DBF	**98.04**	**88.03**	**65.13**

从表 1 可以看出，利用 U-Net 模型进行实验时，F_1 和 *IoU* 值较其他方法都比较低，而在 U-Net 模型中分别引入 Transformer 和 SA-DBF 模块后，增强了模型对全局和局部上下文信息的理解能力，能够更准确地划分像素，从而提高语义分割任务的性能，使得模型在 F_1 和 *IoU* 值都有一定的提升。其中，U-Net+SA-DBF 模块的 *OA* 比 U-Net 模型的 OA 低了 0.12 百分点，但 F_1 比 U-Net 模型的 F_1 高了 2.24 百分点，说明在 U-Net 模型上引入 SA-DBF 模块对模型性能的提升有很大的作用。

2.5.2 注意力模块对比

在该实验中，对比了本文所提的 SA-DBF 模块与其他 4 种常见的注意力模块（SENet，CBAM，SGE-Net，ECA-Net）在道路信息提取任务上的性能表现，实验结果见表 2，表中加粗字体为最优结果。

表 2 不同注意力模块性能对比

Tab.2 Performance comparison of different attention modules

注意力模块	*OA*/%	F_1/%	*IoU*/%	参数量/10^6MB
SENet[20]	96.39	84.12	63.12	28.08
CBAM[21]	97.08	85.97	64.01	28.09
SGE-Net[22]	96.27	86.36	64.37	25.55
ECA-Net[23]	97.04	87.13	64.13	25.65
SA-DBF	**98.04**	**88.03**	**65.13**	**24.20**

从表 2 可以看出，SA-DBF 模块在所有性能指标上表现最优。在整体精度 *OA*，F_1 和 *IoU* 上分别达到了 98.04%，88.03%和 65.13%，这意味着在语义分割任务中，SA-DBF 模块能够更准确地进行像素分类和边界划分。同时，SA-DBF 模块的参数量为 24.20×10^6 MB，较其他注意力模块更小，说明在保持高性能的同时内存消耗也相对较低。

2.5.3 Transformer 规模分析

隐藏层尺寸和 Transformer 层数决定 Transformer 的规模。因此，本文通过消融实验来验证 Transformer 规模对分割性能的影响。“基础”（Base）模型的隐藏层尺寸和注意力头数分别设置为 512 和 8，而

“大型”(Large)模型的超参数设置为 768 和 12,实验结果见表 3。从表 3 中可以看到较大的模型使道路分割性能只得到了略微提升,但这带来额外的计算成本,使得模型训练时间增长。为了提高效率,减小计算成本,本文采用 Base 模型进行所有实验。

表 3 Transformer 规模对模型的影响结果

Tab.3　Influence of Transformer scale on the model

(%)

Transformer 规模	*OA*	F_1	*IoU*
Large	97.08	85.97	64.01
Base	96.82	84.87	63.85

2.5.4　与主流方法对比实验

与近几年提出的 D-LinkNet[24], TransRoadNet[18], CoAtNet 模型和经典的遥感道路的分割模型 SegNet, DeeplabV3+, U-Net 在 Massachusetts 遥感数集上的数值对比结果见表 4,表中加粗字体为最优结果。所有模型采用同一个数据集和实验环境,为了比较上述模型在不同尺度道路上的提取性能,分别选取不同分辨率和地域环境复杂程度的道路图片,进行对比分析。

表 4　不同模型的实验对比结果

Tab.4　Experimental comparison results of different models

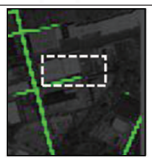
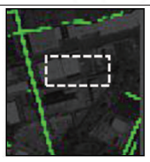
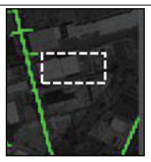
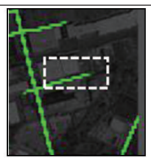
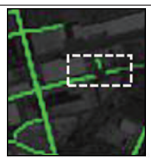
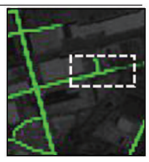
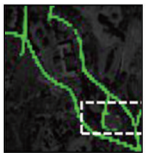
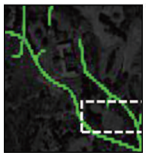
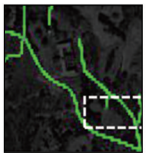
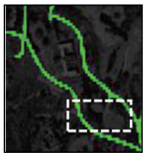
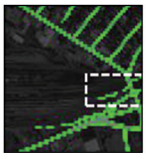
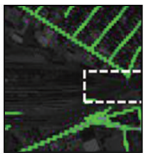
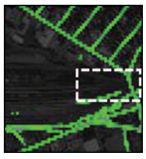
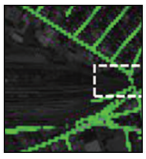
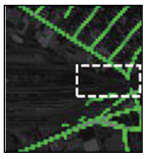
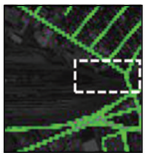
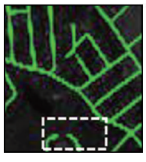
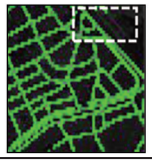
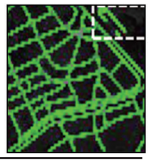
方法	*OA*/%	F_1/%	*IoU*/%	时间/s	参数量/10^6MB
SegNet	95.27	81.34	60.63	43.2	30.6
DeeplabV3+	96.21	83.42	63.08	43.5	30.2
U-Net	96.39	84.12	63.12	42.6	25.3
D-LinkNet	97.32	85.98	63.29	41.5	30.9
TransRoadNet	97.49	85.26	63.71	40.3	31.4
CoAtNet	97.51	86.24	63.92	40.6	27.6
本文方法	**98.04**	**88.03**	**65.13**	**39.1**	**24.2**

由表 4 可以看出,本文模型与近几年 TransRoadNet 和 CoAtNet 2 种遥感图像最优分割方法相比,*OA*, F_1, *IoU* 分别提高 0.55, 2.77, 1.42 百分点和 0.53, 1.79, 1.21 百分点。与 U-Net 分割方法相比,*OA*, F_1, *IoU* 分别提高 1.65, 3.91, 2.01 百分点。在运行时间上,本文模型推理 49 张图片仅需 39.1 s,相比于经典模型 SegNet 快了 4.1 s。在训练参数上面,改进的网络只需要训练 24.2×10^6MB 的参数,大大节约了计算成本。

为了更直观地对比不同模型的道路提取效果,选取 6 个实验网络的道路提取结果,如表 5 所示。

表 5　不同网络的实验对比结果

Tab.5　Experimental comparison results of different networks

序号	原图	DeepLabV3+	U-Net	SegNet	TransRoadNet	D-LinkNet	CoAt	本文方法
1								
2								
3								
4								
5								

通过分别对比表 5 的每个网络,本文网络像素目标提取效果明显优于其他网络,对于道路的边缘细节信息提取得更精确。其次,针对于偏小的道路信息的图片,其他网络对于道路的整体信息都很难提取出来,本文网络不仅可以整体提取道路像素点而且可以更好地提取偏小的道路边缘像素。此外,由图中白色框区域可以看出,DeepLabV3+,U-Net,SegNet,TransRoadNet 和 D-LinkNet 的预测结果较为粗糙,预测图中出现较多的孤立点,道路的断裂现象明显,在提取树木遮挡的道路上效果不佳,其中 SegNet 在提取此类道路方面表现最差,丢失程度最高。本文提出的模型道路预测结果要更加平滑,没有出现孤立点; 在道路被树木、建筑等障碍物部分或完全遮挡的情况下,提取的结果要更加准确和完整。通过对 DeepLabV3+,U-Net,SegNet,TransRoadNet,D-LinkNet 和 CoAtNet 等 6 种网络提取结果的分析,可以发现本文提出的网络方法可以更有效、更全面地提取道路,能够准确地分割道路边缘,并且可以有效地解决树木、建筑物等背景特征带来的干扰,最终提取的道路目标的完整度更高,与标签有更高的相似度。

3 结论

本文提出了一种双分支级联网络,将全局信息与局部信息相结合用于高分遥感影像道路信息提取。其中,CNN 分支通过卷积运算提取输入特征图的局部信息; Transformer 分支通过 MHA 和 MLP 学习输入影像全局上下文信息。在编码阶段结束后,2 个分支使用 SA-DBF 模块结合在一起,通过上采样操作最终生成道路的分割图。

经过实验,加入融合模块后可以更好地提取到关键信息,与没加入之前相比,F_1 和 IoU 分别提高了 2.06 和 2.01 个百分点。与近几年 2 种主流方法 TransRoadNet 和 CoAtNet 相比,OA,F_1,IoU 分别提高 0.55,2.77,1.42 百分点和 0.53,1.79,1.21百分点。

所提方法实现了全局信息与局部信息的有效融合,提高了道路的分割精度,为有关需要道路信息更新的领域带来了有用价值。在后续工作中,将使用更加轻量化的 Transformer 结构,使得模型在保持分割精度的基础上减少运算量。

参考文献(References):

[1] He D,Zhong Y,Wang X,et al.Deep convolutional neural network framework for subpixel mapping[J].IEEE Transactions on Geoscience and Remote Sensing,2021,59(11):9518-9539.

[2] Huang B,Zhao B,Song Y.Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery[J].Remote Sensing of Environment,2018,214:73-86.

[3] Xu Y,Chen H,Du C,et al.MSACon Mining spatial attention-based contextual information for road extraction[J].IEEE Transactions on Geoscience and Remote Sensing,1809,60:5604317.

[4] Yuan Q,Shen H,Li T,et al.Deep learning in environmental remote sensing achievements and challenges[J].Remote Sensing of Environment an Interdisciplinary Journal,2020,241:111716.

[5] Zhu Q,Zhang Y,Wang L,et al.A global context-aware and batch-independent network for road extraction from VHR satellite imagery[J].ISPRS Journal of Photogrammetry and Remote Sensing,2021,175:353-365.

[6] Yang K,Yi J,Chen A,et al.ConDinet++:Full-scale fusion network based on conditional dilated convolution to extract roads from remote sensing images[J].IEEE Geoscience and Remote Sensing Letters,2021,19:8015105.

[7] He D,Shi Q,Liu X,et al.Generating 2m fine-scale urban tree cover product over 34 metropolises in China based on deep context-aware sub-pixel mapping network[J].International Journal of Applied Earth Observation and Geoinformation,2022,106:102667.

[8] Shelhamer E,Long J,Darrell T.Fully convolutional networks for semantic segmentation[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence.IEEE,2017:640-651.

[9] Ronneberger O,Fischer P,Brox T.U-net convolutional networks for biomedical image segmentation[C]// IEEE Springer International 2015:234-241.

[10] 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.

[11] Chen L C,Zhu Y,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.

[12] Gao L,Song W,Dai J,et al.Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network[J].Remote Sensing,2019,11(5):552.

[13] 王 勇,曾祥强.集成注意力机制和扩张卷积的道路提取模型[J].中国图象图形学报,2022,27(10):3102-3115.
Wang Y,Zeng X Q.Road extraction model derived from integrated attention mechanism and dilated convolution[J].Journal of Image and Graphics,2022,27(10):3102-3115.

[14] 吴强强,王 帅,王 彪,等.空间信息感知语义分割模型的高分辨率遥感影像道路提取[J].遥感学报,2022,26(9):1872-1885.
Wu Q Q,Wang S,Wang B,et al.Road extraction method of high-resolution remote sensing image on the basis of the spatial information perception semantic segmentation model[J].National Remote-Sensing Bulletin,2022,26(9):1872-1885.

[15] Vaswani A,Shazeer N,Parmar N,et al.Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems.2017,Long Beach.ACM,2017:6000-6010.

[16] Sanchis-Agudo M, Wang Y, Duraisamy K, et al. Easy attention: A simple self-attention mechanism for Transformers[J/OL]. 2023: arXiv: 2308.12874. http://arxiv.org/abs/2308.12874.

[17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J/OL]. 2020: arXiv: 2010.11929. http://arxiv.org/abs/2010.11929.

[18] Yang Z, Zhou D, Yang Y, et al. TransRoadNet: A novel road extraction method for remote sensing images via combining high-level semantic feature and context[J]. IEEE Geoscience and Remote Sensing Letters, 1973, 19: 6509505.

[19] Dai Z, Liu H, Le Q V, et al. CoAtNet: Marrying convolution and attention for all data sizes[J/OL]. 2021: arXiv: 2106.04803. http://arxiv.org/abs/2106.04803.

[20] Cao Y, Xu J, Lin S, et al. GCNet: Non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South). IEEE, 2019: 1971-1980.

[21] 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.

[22] Su R, Huang W, Ma H, et al. SGE NET: Video object detection with squeezed GRU and information entropy map[C]//2021 IEEE International Conference on Image Processing (ICIP). Anchorage, AK, USA. IEEE, 2021: 689-693.

[23] Wang Q, Wu B, Zhu P, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle. IEEE, 2020: 11531-11539.

[24] Zhou L, Zhang C, Wu M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City. IEEE, 2018: 192-1924.

Information extraction of roads from remote sensing images using CNN combined with Transformer

QU Haicheng, WANG Ying, LIU Lamei, HAO Ming

(*School of Software*, *Liaoning Technical University*, *Huludao* 125105, *China*)

Abstract: Deep learning-based methods for information extraction of roads from high-resolution remote sensing images face challenges in extracting information about both global context and edge details. This study proposed a cascaded neural network for road segmentation in remote sensing images, allowing both types of information to be simultaneously learned. First, the input feature images were sent to encoders CNN and Transformer. Then, the characteristics learned by both branch encoders were effectively combined using the shuffle attention dual branch fusion (SA-DBF) module, thus achieving the fusion of global and local information. Using the SA-DBF module, the model of the features learned from both branches was established through fine-grained interaction, during which channel and spatial information in the feature images were efficiently extracted and invalid noise was suppressed using multiple attention mechanisms. The proposed network was evaluated using the Massachusetts Road dataset, yielding an overall accuracy rate (OA) of 98.04%, an intersection over union (IoU) of 88.03%, and an F1 score of 65.13%. Compared to that of mainstream methods U-Net and TransRoadNet, the IoU of the proposed network increased by 2.01 and 1.42 percentage points, respectively. Experimental results indicate that the proposed method outperforms all the methods compared and can effectively improve the accuracy of road segmentation.

Keywords: cascaded neural network; Transformer; feature fusion; attention mechanism

(责任编辑: 张 仙)