引用格式: 汤泊川, 帕力旦·吐尔逊, 柏洁馨, 等. 结合 CNN 和 Transformer 的遥感图像土地覆盖分类方法[J]. 微电子学与计算机, 2024, 41(4): 64-73.
TANG B C, PALIDAN Tuerxun, BAI J X, et al. Land cover classification method for remote sensing images using CNN and Transformer[J]. Microelectronics & Computer, 2024, 41(4): 64-73.
DOI: 10.19304/J.ISSN1000-7180.2023.0240

结合 CNN 和 Transformer 的遥感图像土地覆盖分类方法

汤泊川[1,2], 帕力旦·吐尔逊[1,2,3], 柏洁馨[1], 齐然然[1,2]
(1 新疆大学 软件学院, 新疆 乌鲁木齐 830046;
2 新疆维吾尔自治区信号检测与处理重点实验室, 新疆 乌鲁木齐 830046;
3 新疆师范大学 计算机科学与技术学院, 新疆 乌鲁木齐 830046)

摘 要: 利用遥感图像进行语义分割是一种有效的土地覆盖分类方法。然而由于主流框架存在边缘分割不准确、缺乏全局信息导致错误分类等问题, 阻碍了其在土地覆盖分类中的应用。针对以上问题, 提出了一种用于遥感图像土地覆盖分类的卷积神经网络(Convolutional Neural Networks, CNN)和 Transformer 混合网络 CTHNet, 结合了 CNN 的局部细节提取能力和 Transformer 的全局信息提取能力。同时设计了自适应融合模块, 融合来自对应级别的 CNN 和 Transformer 特征, 自适应融合模块的输出进入分割头得到最终的预测结果。最后, 结合边界检测分支为语义分割提供边缘约束。在两个公开的土地覆盖分类数据集上的实验结果表明, 该方法优于当前主流的方法, 分别实现了 90.53% 和 64.33% 的平均交并比(mIoU), 对遥感图像中的大目标和边界也有更好的识别效果。

关键词: 土地覆盖分类; 遥感图像; 特征融合; 卷积神经网络; Transformer
中图分类号: TP391 **文献标识码:** A **文章编号:** 1000-7180(2024)04-0064-10

Land cover classification method for remote sensing images using CNN and Transformer

TANG Bochuan[1,2], PALIDAN Tuerxun[1,2,3], BAI Jiexin[1], QI Ranran[1,2]
(1 College of Software, Xinjiang University, Urumqi 830046, China;
2 Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Urumqi 830046, China;
3 College of Computer Science and Technology, Xinjiang Normal University, Urumqi 830046, China)

Abstract: Semantic segmentation using remote sensing images is an effective land cover classification method. However, its application in land cover classification is hindered by the problems of inaccurate edge segmentation and lack of global information leading to misclassification in mainstream frameworks. To address these problems, a Convolutional Neural Networks (CNN) and Transformer hybrid network CTHNet for land cover classification of remote sensing images is proposed, which combines the local detail extraction capability of CNN and the global information extraction capability of Transformer. The adaptive fusion module is also designed to fuse the CNN and Transformer features from the corresponding levels, and the output of the adaptive fusion module enters the segmentation head to get the final prediction results. Finally, the boundary detection branch is combined to provide edge constraints for semantic segmentation. Experimental results on two publicly available land cover classification datasets show that the method outperforms current mainstream methods, achieving 90.53% and 64.33% of the mean Intersection over Union (mIoU), respectively, and also has

收稿日期: 2023-03-19; **修回日期:** 2023-04-16
基金项目: 国家自然科学基金(62266043); 新疆维吾尔自治区自然科学基金(2022D01A99)

better recognition of large targets and boundaries in remote sensing images.

Key words：land cover classification；remote sensing images；feature fusion； CNN；Transformer

1　引言

随着无人机技术的快速发展，航空成像已成为一种越来越流行的图像捕获技术。基于像素的土地覆盖分类，也被定义为语义分割，使用具有极高空间分辨率的图像，在土地资源管理[1]、城市规划[2]、变化检测[3] 和其他领域[4-5] 发挥着重要作用。近几十年来，土地覆盖分类已成为一个高度活跃的分类领域。传统的视觉判断方法在满足土地覆盖分类产品及时更新、大面积土地覆盖制图等应用时费时、低效、不合理。因此，实现稳健而高效的自动土地覆盖分类方法至关重要。

早期的分类技术[6-8] 仅依赖于像素的光谱波长，导致分类精度低，这是由于土地覆盖类别存在类内差异大和类间差异小的特点。高分辨率的遥感图像包含形状各异的物体，同时其纵横比和颜色纹理也存在复杂变化，例如道路、屋顶、建筑阴影、低矮的植物和树枝。对于大型目标，需要足够的感受野来捕捉丰富的上下文语义信息，以便进行准确的分类。对于小规模物体，需要足够的空间分辨率来确保结构信息的完整性。因此，从遥感图像中实现精确的土地覆盖分类是极其困难的。

近年来，深度学习作为一种图像处理方法，被广泛应用于像素级分类(如语义分割)任务，并取得了良好的效果。特别是基于深度学习方法的卷积神经网络(Convolutional Neural Networks, CNN)和 Transformer 模型，由于其优异的性能，在土地覆盖分类任务中越来越受到关注。基于 CNN，Long 等[9] 提出了完全卷积网络，用于以端到端的方式预测语义分割任务的像素级标签。然而，这种结构通过非线性地叠加卷积层并使用下采样来捕获语义信息，从而减少了原始图像的空间信息。为了解决这一问题，U-Net[10] 采用跳跃连接进行特征融合，重用底层特征，并在一定程度上保留空间细节。为了获取全局上下文信息，DeepLab[11] 和 PSPNet[12] 引入了基于金字塔结构的多尺度特征融合模块，以聚合来自不同感受野的全局上下文信息。Guo 等[13] 设计了一个特征增强模块和一个双门融合模块，增强全局信息的提取并促进与局部信息的融合。Wei 等[14] 提出了一种双编码器注意力网络，使用双分支结构更彻底地融合编码器的全局特征。DANet[15] 引入了一个双重注意力模块来丰富特征表示。

CNN 具有的结构特性包括局部性和平移不变性。这是由卷积运算的局部连接和权重共享决定的。这两种策略有效地减少了基于 CNN 模型的参数数量，提高了特征学习的效率。由此可见，CNN 在局部特征提取方面具有优势，但其捕获全局信息的能力仍然不足，这对遥感图像土地覆盖分类非常重要。尽管 DeepLab 和 PSPNet 扩展了感受野以获得多尺度的全局上下文，但全局信息仍然局限于局部区域。同时，注意力机制为全局信息建模提供了一种良好的模式，但受到模块数量少和计算负担大的限制。因此，需要一种纯粹基于注意力的架构来实现足够的全局信息提取。

Transformer 是 2017 年提出的一种通过自我注意机制学习特征的结构[16]，该结构首次应用于自然语言领域，并取得了优异的性能。目前，Transformer 已被认为是自然语言处理任务的首选模型。受此启发，研究人员将 Transformer 应用于计算机视觉任务。ViT 首次使用纯 Transformer 结构作为特征提取器来处理图像识别任务[17]。在大规模数据集上进行预训练的条件下，ViT 的精度略高于基于 CNN 的模型，这证明了 Transformer 模型强大的特征提取能力及其在图像处理领域的潜力。Zhang 等[18] 提出了第一个使用纯 Transformer 结构作为主干用于语义分割任务的网络 SETR。Swin Transformer[19] 提出了一种分层 Transformer 结构,该方法实现了在各个小窗口内部进行自注意力计算，并通过移动窗口建立跨窗口连接，大大提高了自注意力计算的效率。此外，Swin Transformer 结构可以输出分层特征图，能够完美地集成到现有的语义分割模型中。Transformer 结构将原始图像转换为序列到序列的图像块作为输入，能够从全局的角度提取图像特征，但对于局部的细节信息掌握较差，同时普遍存在需要大量内存和计算资源的问题。

综上，CNN 结构缺乏全局信息，但有足够的局部信息，Transformer 结构缺乏局部信息，但具有足够的全局信息。因此，本文考虑结合 CNN 和 Transformer 各自的优势，提出 CNN 和 Transformer 混合网络(CNN and Transformer Hybrid Network, CTHNet)。CTHNet 同时包含 CNN 和 Transformer 两种结构，以保留各自结构的优势。

2 方法与实现

2.1 模型框架

针对遥感场景中大目标识别较差与边缘不清晰的问题，本文提出用于遥感图像土地覆盖分类的 CNN 和 Transformer 混合网络(CTHNet)，如图 1 所示。图中虚线箭头表示下采样操作，实线箭头表示不改变特征图尺度的卷积操作，点划线箭头表示上采样操作。总体来说，CTHNet 由以下模块组成：

(1)基于 CNN 的编码器模块。选择使用带有预训练权重的 Res2Net50[20] 作为主干，该网络可以有效地提取多尺度语义特征，并且可以输出多个尺度的特征图用于后续的特征融合以及上采样。

(2)基于 Transformer 的语义提取模块。与基于 CNN 的主干相比，Transformer 结构可以更好地模拟图像中的远距离空间相关性。同时，与其他基于 Transformer 的结构相比，Swin Transformer 具有更低的计算复杂度、更快的推理速度，同时由于移动窗口的设计对于位置信息也更加敏感。

(3)自适应融合模块。有效融合分别来自 CNN 和 Transformer 的特征，实现优势互补。

(4)模型的尾部设计使用了分割头。在将特征传递到自适应融合模块之后，输出的分层特征具有强大的语义和丰富的空间细节。所提出的分割头将它们上采样至相同尺度后叠加，然后使用两个 1×1 卷积层来预测分割图，具体如图 1 所示。

(5)模型还使用双任务设计，添加边界检测分支以辅助土地覆盖分类任务。

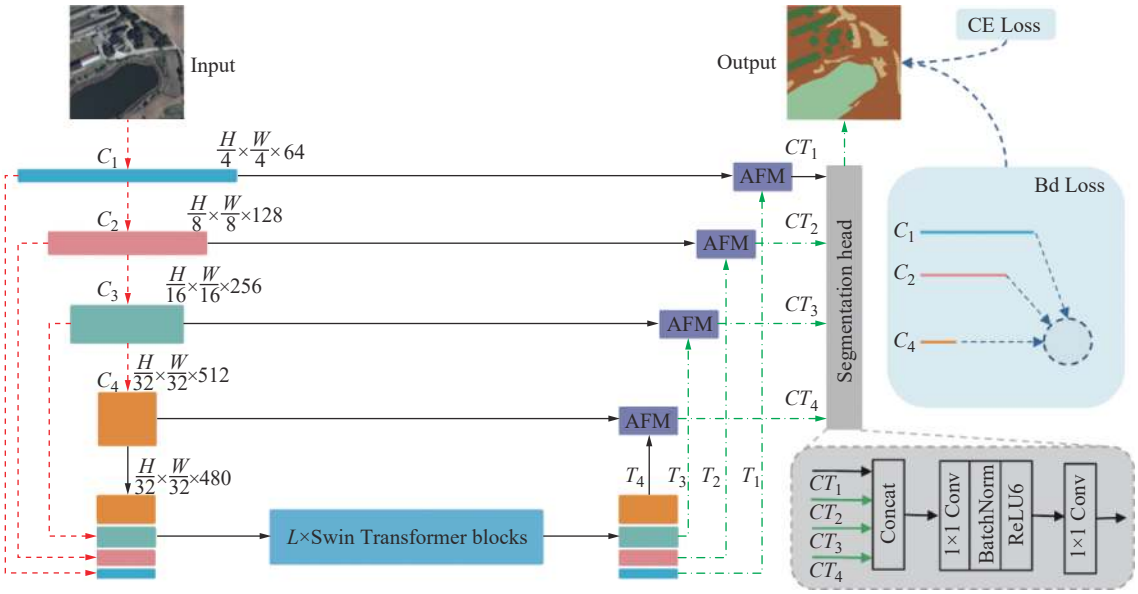

图 1 CTHNet 整体框架

Fig. 1 CTHNet Overall Framework

2.2 基于 CNN 的编码器模块

残差网络(ResNet)[21] 被提出用于解决非常深的 CNN 模型的梯度衰退问题。它使用残差连接来连接不同的卷积层，使得来自浅层的特征信息可以传播到深层。ResNet 由 4 个阶段的 Resblock 组成，每个阶段以比例因子 2 对特征图进行下采样。原始图像经过 ResNet 主干的具体尺度变化如图 1 所示。给定 $H\times W\times 3$ 的原始输入，在经过第一阶段的 Resblock 之后，分辨率降低到原图的 1/4，其中通道维度变为 64。然后，特征经过 4 个阶段后分别生成 C_1、C_2、C_3 和 C_4 这 4 个特征图，其分辨率依次降低一半，长宽分别变为原始图像的 1/4、1/8、1/16 和 1/32，通道依次增加到 C、$2C$、$4C$ 和 $8C$。本文的实验选择 Res2Net50 模型，4 个阶段分别包含 3、4、6 和 3 个 Resblock；输入特征通过残差方式被添加到主分支输出特征，随后是 ReLU 激活函数以增强模型的非线性表达能力。瓶颈架构和残差连接的设计使训练过程更容易。Res2Net50 因其容易根据不同的任务需求扩展到不同的模型大小而被选为 CNN 分支。此外，C_1、C_2、C_3 和 C_4 的多尺度输出特征适合于处理多尺度对象。

2.3 基于 Transformer 的语义提取模块

语义提取模块由几个堆叠的 Swin Transformer 块组成，堆叠的数量为 8。Swin Transformer 块是

Swin Transformer 结构的核心，其包含两种形式的块：基于窗口的多头自注意块(W-MSA)和基于移动窗口的多头自注意块(SW-MSA)。W-MSA 和 SW-MSA 被顺序连接以更有效地获得全局空间相关性，其结构如图 2 所示。每个 Swin Transformer 块包含两个子层。第一个子层是 W-MSA，它以非重叠的方式将特征图划分为单独的窗口，然后在这些局部窗口中计算自注意。对于局部窗口大小为 $m\times m$ 的特征映射 $X\in R^{H\times W\times C}$，其计算复杂度$\Omega$为

$$\Omega(\mathrm{MSA})=4HWC^2+2(HW)^2C \tag{1}$$

$$\Omega(\mathrm{W\text{-}MSA})=4HWC^2+2(HW)^2C \tag{2}$$

由于窗口大小比图像小得多，因此计算复杂度显著降低。第二子层(MLP)是完全连接的层。在 W-MSA 和 MLP 之间添加残差连接，以抵消特征权重的梯度消失和退化。SW-MSA 块的结构与 W-MSA 块的结构几乎相同，只是在 SW-MSA 层的计算中存在一半窗口大小的偏移。

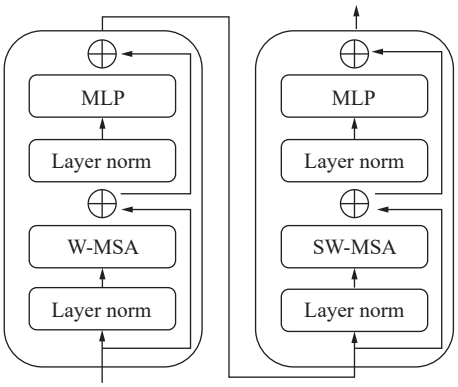

图 2　Swin Transformer 块

Fig. 2 Swin Transformer block

语义提取模块的前后连接情况如图 1 所示，该模块将来自不同尺度的特征图作为输入。为了进一步减少计算量，使用 1×1 卷积操作将堆叠的包含不同尺度的特征图通道数缩小到 480。来自不同规模的特征图集合具有相同的分辨率，它们被连接在一起作为语义提取模块的输入。Swin Transformer 可以获得完整的图像感受野和丰富的语义。具体地说，全局自注意力沿着空间维度在特征图之间交换信息，1×1 卷积层使来自不同尺度的卷积神经网络特征图之间交换信息。在每个 Swin Transformer 块中，在交换来自所有尺度的特征信息之后学习残差映射，然后将残差映射添加到特征图中以增强表示和语义。最后，在经过几个 Swin Transformer 块之后，模型获得了全局尺度的语义信息。特征在经过语义提取模块后生成与基于 CNN 的编码器模块相对应的 T_1、T_2、T_3 和 T_4 这 4 个特征图。

2.4　自适应融合模块

为了解决具有挑战性的 CNN 和 Transformer 特征融合问题，本文提出了自适应融合模块(AFM)。AFM 学习来自 CNN 和 Transformer 结构中的特征权重，并为 CNN 提取的局部特征和 Transformer 提取的全局特征分配更大的权重，学习有利的特征信息，同时抑制其他次要的特征信息。使得 CNN 可以从 Transformer 获得全局信息补充，Transformer 可以从 CNN 获得局部信息补充。

如图 3 所示，首先，来自 CNN($C_1\sim C_4$)和 Transformer($T_1\sim T_4$)的输入特征分别经过 1×1 卷积。接下来，使用 Concat 操作合并两个特征，并将它们发送到下一个 1×1 卷积，以在 CNN 和 Transformer 特征之间进行交互。然后，通过切割操作分离特征，再分别进行 1×1 卷积和 Sigmoid 函数，以将像素值从 0 归一化到 1，以避免出现最大值和最小值。最后，使用堆栈操作来并行连接通道维度中的两个特征，并应用 Softmax 函数来获得像素级权重。然后将生成的 CNN 和 Transformer 的特征权重图与先前的特征逐像素相乘，以执行像素级重加权操作。此外，还使用残差连接(虚线)添加先前的特征，以加速模型优化并降低特征权重图的学习难度。自适应融合模块最终输出 $CT_1\sim CT_4$ 这 4 个特征图进入分割头。

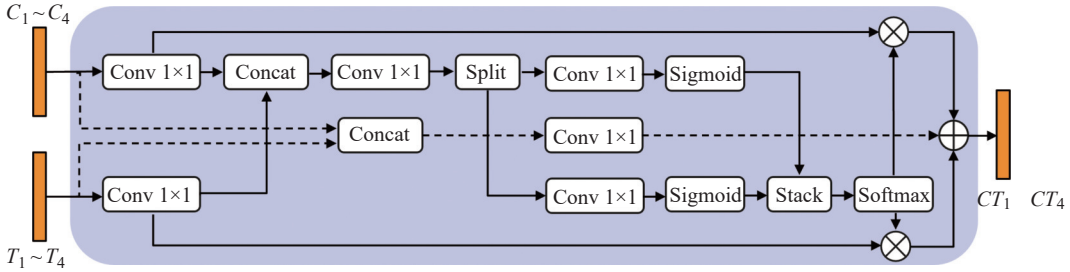

图 3　自适应融合模块

Fig. 3 Adaptive fusion module

基于 Transformer 和 CNN 结构学习到的特征权重分别如图 4 所示，两种结构源自完全不同的特征提取原理。为了更好了解两者提取的特征图的特征，本文将其各自的特征图可视化以进行比较分析，以便于设计更合理的网络结构。图中显示，CNN 主干更善于提取图像的基本信息，如点、线和局部纹理，而对于深层次的高级语义信息，视觉效果并不明显；而对于 Transformer 主干，图像的基本特征图则显得“抽象”，而深层次的语义特征则显示出明显的“聚类”效果。这主要是因为 CNN 通过卷积核提取特征，而 Transformer 通过自注意力提取特征。这也证明了 CNN 和 Transformer 在处理局部和全局信息方面各有优势。

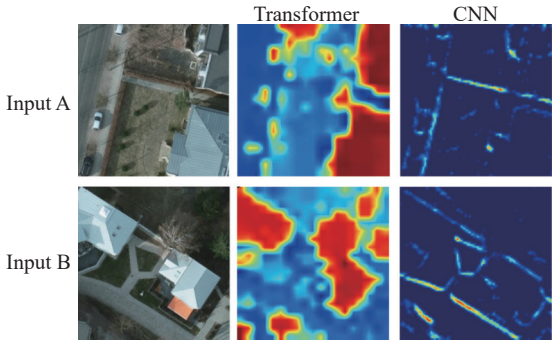

图 4 Transformer 和 CNN 特征权重图

Fig. 4 Transformer and CNN feature weight map

2.5 边界检测分支

基于 Zhang 等[22]研究的探索，在基于 CNN 的遥感图像分割网络中添加边缘约束可以有效提高模型的边界检测性能。因此，本文采用了双任务设计，使用边界检测分支提高边界像素定位的准确性，并减少用于语义分割的椒盐噪声。由于边界检测任务和语义分割任务共享相同的主干，因此，可以避免引入大量的新参数，即基于 CNN 的编码器模块输出的特征图 C_1、C_2、C_3 和 C_4 直接用于边界像素约束。应当注意，特征图的有效使用与视觉任务的准确性密切相关。如图 1 所示，本文在这项工作中使用主干网络的 1、2 和 4 阶段输出的特征图(即 C_1、C_2 和 C_4)来约束边界像素，完成边界损失计算。这种设计思路借鉴了基于 CNN 的边缘提取方法的经验[23]，使用浅层特征来保留局部细节，并使用深层特征来抑制复杂纹理。边界检测的损失函数 BD 如式(3)所示。

$$\mathrm{BD}=1-2\times\frac{P\times R}{P+R} \tag{3}$$

式中：P 为精度；R 为召回率。

该损失可以全面评估边界检测结果的召回和精度。本文使用最广泛使用的交叉熵损失来评估语义分割的准确性。因此，总损失是边界损失和分割损失的加权和，表达式为

$$L_{\mathrm{total}}=\lambda_1 CE+\lambda_2 DL \tag{4}$$

式中：λ_1 和 λ_2 分别是交叉熵损失和边界检测损失的权重。一般情况下，λ_1 和 λ_2 的设置是动态的。但是根据本文使用预训练模型对多组 λ_1 和 λ_2 组合进行的实验，其值对模型的收敛速度和精度几乎没有影响。具体来说，当 $\lambda_1=\lambda_2=1.0$ 时，模型的精度为 90.48%；当 $\lambda_1=1.0$ 和 $\lambda_2=0.1$ 时，模型的精度为 90.51%；当 $\lambda_1=1.0$ 和 $\lambda_2=0.5$ 时，模型的精度为 90.53%。因此，在本文的双任务实验中，将 λ_1 的权重统一设置为 1.0，将 λ_2 的权重统一为 0.5。

3 实验部分

3.1 实验环境与数据集

本文基本实验平台使用 NVIDIA GeForce RTX 2080Ti GPU 显卡，包含 11 GB 内存，并配备 CUDA 10.2 和 cuDNN7.6.5。训练和测试均在该平台上进行。为了训练网络，使用 Pytorch 和 AMSGrad 实现了 Adam 优化器，权重衰减为 2×10^{-5}。学习率(lr)衰减设置为 $[1-(\text{cur epoch}/\text{max epoch})]^{0.9}$。

实验使用的第一个数据集 LandCover.ai[24] 是在波兰拍摄的 41 幅 RGB 图像的数据集，可用于从航空图像自动绘制土地覆盖图。具体而言，数据集包含 33 张空间分辨率为 25 cm(9 000×9 500 像素)的正射影像和 8 张空间分辨率为 50 cm(4 200×4 700 像素)的正射影像，分别覆盖 $1.767\,6\times10^8\,\mathrm{m}^2$ 和 $3.951\times10^7\,\mathrm{m}^2$；所有图像的总覆盖面积达 $2.162\,7\times10^8\,\mathrm{m}^2$。就土地覆盖类型而言，包含 3 种常见的土地覆盖类型：“建筑物”、“林地”和“水”。同时，在本文的实验中，将未分类的区域设置为背景类，并根据划分数据集的官方方法将数据集划分为训练、测试和验证集。本文将每幅原图像分别划分为大小为 512×512 像素的非重叠图像集，并丢弃边缘无法整除的像素，最终获得了 10 674 张分辨率为 512×512 的小图，随机划定其中 7 470 张小图用于训练，1 602 张小图用于验证，1 602 张小图用于测试。

为了验证模型的可迁移性，本文还在 GID-15[25] 数据集上开展了实验。GID-15 是基于 15 个地物类别的高分卫星图像数据集，是一个大规模的土地覆盖分类数据集。GID-15 由于其覆盖范围大、分布广、注释精细和空间分辨率高，与其他土地覆盖

分类数据集相比具有很大优势。数据集包含 10 幅空间分辨率为 1 m 的 RGB 图像及其相应的标签。每幅图像覆盖 5.06×10^8 m^2 的地理区域，分辨率为 4 200×4 700 像素。采用同样的方式将每幅图像划分为大小为 256×256 像素的图像集，最终获得 7 280 张小图，其中 4 368 张小图用于训练，1 456 张小图用于验证，1 456 张小图用于测试。数据集部分示例如图 5 所示。

(a) 原图　(b) 标签
(a) Original　(b) Label

图 5　部分数据集

Fig. 5 Partial dataset

3.2　评价指标

在本文中，使用了两个主流指标来验证模型的性能：平均像素精度(mPA)和平均交并比(mIoU)，同时展示了每个网络的参数量。mPA 是像素位置感知度量，而 mIoU 是更符合人类视觉评估的基于区域的度量，其表达式分别为

$$\mathrm{mPA}=\frac{1}{N}\sum_{c=1}^{N}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c} \tag{5}$$

$$\mathrm{mIoU}=\frac{1}{N}\sum_{c=1}^{N}\frac{\mathrm{TP}_c}{\mathrm{TP}_c+\mathrm{FP}_c+\mathrm{FN}_c} \tag{6}$$

式中：TP_c 为真正像素的数目；FP_c 为真负像素的数目；TN_c 为假正像素的数量；FN_c 为假负像素的数目；N 为数据集中的类别数；c 为某个特定类别。

3.3　模型结构消融实验

在本文中，首先在 LandCover.ai 数据集上进行了消融实验，以讨论不同组件的影响，包括基于 CNN 的编码器模块、基于 Transformer 语义提取模块、自适应融合模块和边界检测分支。然后在两个公共数据集上进行实验，详细描述了实现细节，并将结果与语义分割任务的其他先进模型进行了对比，所有结果都是通过对训练集的训练和对测试集的评估获得的。

对于基于 CNN 的编码器的有效性评定，本文从两个方面设计了消融实验：

(1)基于 CNN 的编码器选择不同尺度特征作为语义提取模块输入的影响。

(2)基于 CNN 的编码器选择不同尺度特征作为自适应融合模块输入的影响。

如表 1 所示，本文开展了消融实验，首先进行试验的是将来自不同尺度的堆叠特征图作为基于 Transformer 的语义提取模块的输入，与之进行对比的是只将最后一层特征图作为语义提取模块的输入。为了公平比较，附加了一个 1×1 卷积层来扩展使之与堆叠的特征图拥有相同的通道数。实验结果证明了堆叠使用全尺度的特征图作为输入的有效性。

表 1　语义提取模块输入消融实验

Tab. 1 Semantic extraction module input ablation experiment

语义提取模块的输入	mIoU/%	参数量/个
全尺度特征图	90.53	58.48×10^6
仅最后一层	88.76	57.06×10^6

在通过基于 Transformer 的语义提取模块之后，自适应融合模块将把语义特征注入到基于 CNN 编码的多尺度特征图中。为了在精度和计算成本之间寻求更好的权衡，本文尝试从不同的尺度中选择特征图进行自适应融合。如表 2 所示，使用来自{C_1, C_2, C_3, C_4}的特征图可以以最大的参数量实现最佳性能。仅使用来自{C_3, C_4}的特征图则以最小的参数量实现最差的性能。实验结果表明，选择使用{C_2, C_3, C_4}中的特征图，更能实现精度和计算成本之间的较好权衡。在本文所有的其他实验中，为了追求更高的精度，选择使用{C_1, C_2, C_3, C_4}中的特征图。

表 2　自适应融合模块输入消融实验

Tab. 2 Adaptive fusion module input ablation experiment

自适应融合模块输入	mIoU/%	参数量/个
C_1, C_2, C_3,C_4	90.53	58.48×10^6
C_2, C_3, C_4	90.45	57.04×10^6
C_3, C_4	88.86	56.23×10^6

全面的模型结构消融实验如表 3 所示，结果表明所有对比试验都得到了改进。具体而言，在基于

CNN 的编码器模块之后特征进去基于 Transformer 的语义提取模块，比较表 3 中前两行的结果可得，该模块将模型的 mIoU 从 86.17% 提高到 89.04%，增加了 2.87%。这是一个巨大的提升，表明了基于 Transformer 的语义提取模块的有效性，以及在土地覆盖分类任务中全局信息的重要性，通过 Transformer 结构提取全局信息对于预测结果的正确分类很重要。在增加了自适应融合模块后，对比直接将两种结构的特征相加的方式，模型的 mIoU 提升了 0.79%，这证明了来自两种结构的特征需要设计合适的融合模块才能实现充分的优势互补。最后，通过表 3 也可以得出，边界检测分支在几乎不添加计算量的情况下，实现了 0.7% 的 mIoU 提升。

此外，本小节也在 GID-15 数据集上进行了模型整体结构的可视化消融实验，结果如图 6 所示。从右下角到左上角依次为基线模型到逐步添加了所有模块的 CTHNet 的预测结果，以及标签与原图。从可视化的结果可以看出，在引入了基于 Transformer 的语义提取模块之后，模型对大型地物目标有了更好的识别效果，在引入边界检测分支后，对地物的边界有了更好分割效果，充分证明了 CTHNet 模块设计的有效性。

表 3 全面消融实验

Tab. 3 Comprehensive ablation experiment

方法	语义提取模块	自适应融合模块	边界检测分支	mIoU/%	mPA/%	参数量/个
CTH Net				86.17	89.79	35.64×10^6
	√			89.04	93.15	56.15×10^6
	√	√		89.83	94.54	58.09×10^6
	√	√	√	90.53	95.03	58.48×10^6

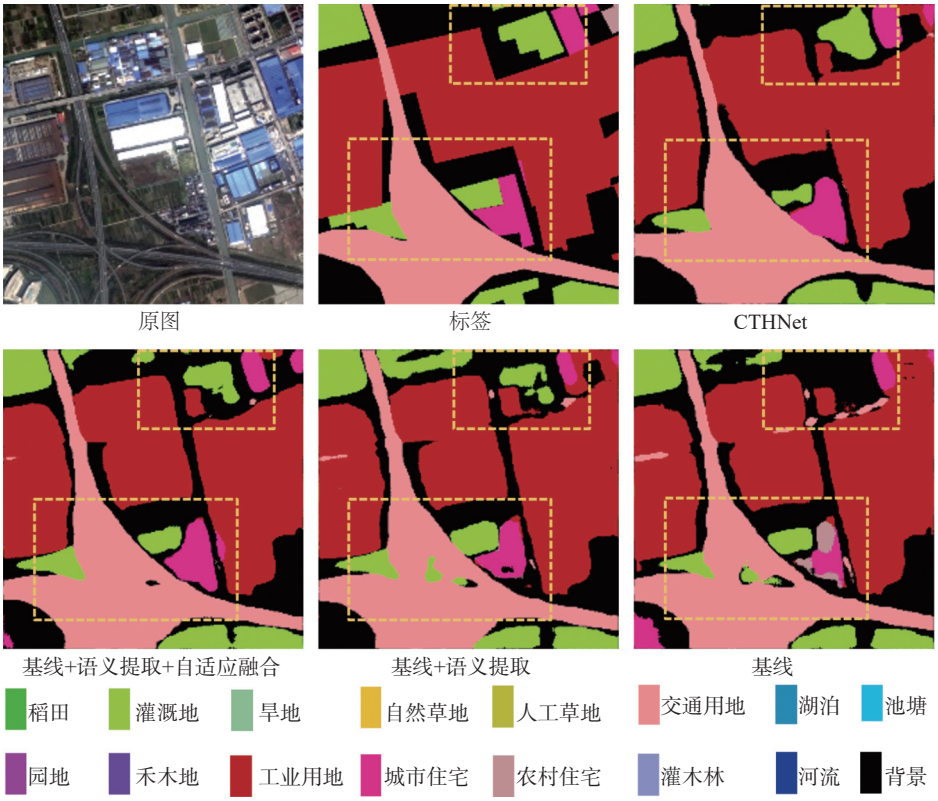

图 6 结构消融实验可视化结果

Fig. 6 Visualization results of structural ablation experiments

3.4 与主流方法比较

在本文中，为了验证 CTHNet 的有效性，使用 LandCover.ai 数据集对比 CTHNet 与目前最先进的方法。对比模型包括一些经典模型，如 UNet[6]、PSPNet[7]、SETR[10] 和 Swin Transformer[11]，以及最新使用该数据集发表论文的模型 DGFNet[16] 和 DEANet[17]。其中，本文方法采用带有预训练权重的 Res2Net50[18] 作为主干网络，UNet、PSPNet 均采用带有预训练权重的 ResNet50 作为主干网络，SETR、Swin Transformer 也使用了官网提供的预训

练权重。

在 LandCover.ai 数据集上的实验定量结果如表 4 所示，并且还包括模型参数量大小。图 7 和图 8 显示了在该数据集上的可视化结果，其中每一行表示测试集中不同样本的预测结果。从定量结果可以看出，本文的模型在两个数据集上的所有指标上都取得了最好的性能。相比于主流的语义分割方法，本文提出的 CTHNet 在引入了 Transformer 结构之后，更好地提取了全局特征，同时设计模块有效地融合了局部和全局特征，在参数量降低的情况下，各项指标均进一步提升，可见本文模型结构设计的合理性。

表 4　与主流方法对比

Tab. 4　Comparison with mainstream methods

方法名称	LandCover.ai		GID-15		参数量/个
	mIoU/%	mPA/%	mIoU/%	mPA/%	
UNet	86.24	91.02	57.69	69.58	72.44×10^6
SETR	88.44	92.89	59.06	70.89	85.93×10^6
PSPNet	89.52	94.15	61.23	72.56	50.34×10^6
Swin T	89.93	94.08	62.14	75.03	87.59×10^6
DGFNet	88.87	93.28	–	–	44.53×10^6
DEANet	90.28	94.54	–	–	60.29×10^6
CTHNet	**90.53**	**95.03**	**64.33**	**76.46**	58.48×10^6

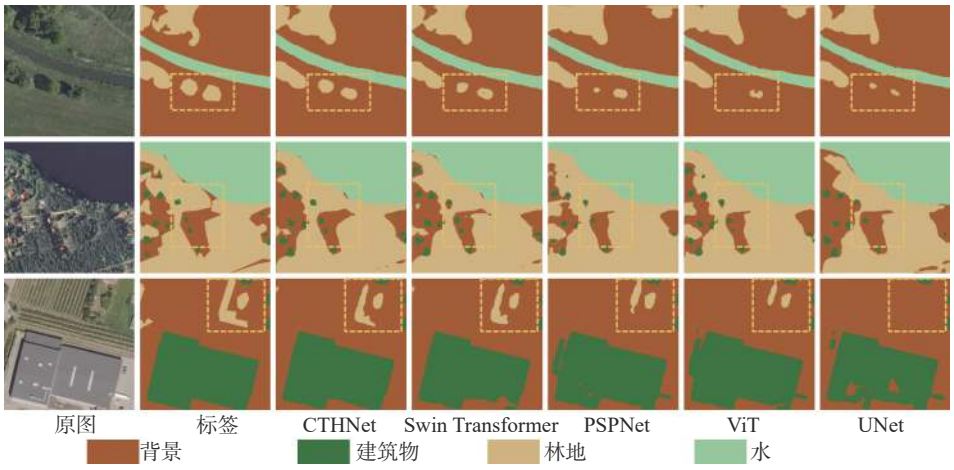

图 7　LandCover.ai 可视化结果对比

Fig. 7 Comparison of LandCover.ai visualization results

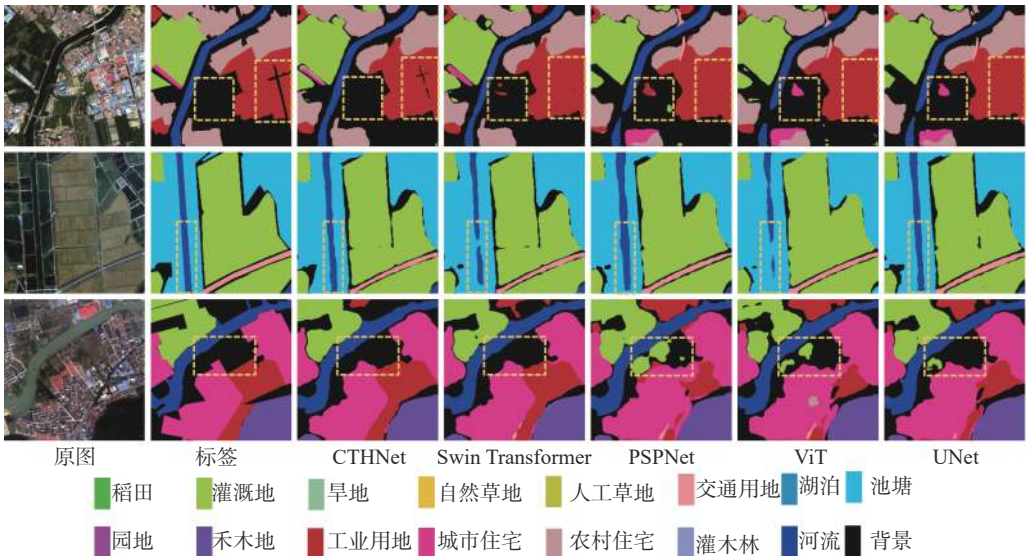

图 8　GID-15 可视化结果对比

Fig. 8 Comparison of GID-15 visualization results

与其他拥有 CNN 或 Transformer 结构的经典模型相比，就最重要的评价指标 mIoU 而言，在

LandCover.ai 数据集上，CTHNet 以 0.6% 的优势超过了第二好的模型，与最新的模型 DEANet 相比，CTHNet 提升了 0.25%；在 GID-15 数据集上，以 2.42% 的优势超过了第二好模型。可以看出，本文的模型显著提高了遥感图像土地覆盖分类的分割效果。另一个可以通过实验结果得出结论是，CTHNet 带来的分割性能的提高不是由于参数量的改变，而是由于模型结构设计的合理性。可以通过比较模型参数更大的模型的分割效果得出这个结论，包括 UNet 和 SETR。与这些模型相比，CTHNet 具有更小的参数量，但本文的模型可以获得更高的性能指标。与更大的模型 Swin Transformer 相比，这一结论也可以得到验证。

在可视化结果方面，图 7 和图 8 中用虚线框标明了一些重要区域，观察可得：一方面，本文的模型明显能够更好地识别地物的类别，错误分类明显减少，这是由于 Transformer 结构引入了更多全局信息，模型可以从全局的角度判断地物的类别；另一方面，CTHNet 对于地物的边界分割效果明显更平滑，这得益于边界检测损失的设计，使得预测图拥有更好的可视化效果。

4 结束语

本文提出了一种新的 CNN 和 Transformer 混合网络 CTHNet，结合了 CNN 结构在局部建模和 Transformer 结构在全局建模中的优势，用于高分辨率的遥感图像土地覆盖分类。首先采用基于 CNN 的编码器结构用于特征提取，因为 CNN 的架构可以更好地提取局部信息，随后使用多个连续堆叠的 Swin Transformer 块构建语义提取模块，因为 Transformer 结构可以更好地建模图像中的长距离空间相关性，即全局特征。然后提出了自适应融合模块有效地融合两种结构提取的特征，在最终生成分割预测图之前，使用双任务设计引入边界检测分支，进一步提升地物边界分割效果。通过实验结果，证明了本文提出模型中各个模块和结构的有效性，并对比了最先进的基于 CNN 和 Transformer 方法，实验结果证明了 CTHNet 的优越性。

参考文献：

[1] SHIRMOHAMMADI B, MALEKIAN A, SALAJEGHEH A, et al. Scenario analysis for integrated water resources management under future land use change in the Urmia Lake region, Iran[J]. Land Use Policy, 2020, 90: 104299. DOI: 10.1016/j.landusepol.2019.104299.

[2] ZHANG C, SARGENT I, PAN X, et al. Joint deep learning for land cover and land use classification[J]. Remote Sensing of Environment, 2019, 221: 173-187. DOI: 10.1016/j.rse.2018.11.014.

[3] SULLA-MENASHE D, GRAY J M, ABERCROMBIE S P, et al. Hierarchical mapping of annual global land cover 2001 to present: the MODIS Collection 6 Land Cover product[J]. Remote Sensing of Environment, 2019, 222: 183-194. DOI: 10.1016/j.rse.2018.12.013.

[4] 徐进勇, 汪潇, 张增祥, 等. 土地资源多尺度遥感智能解译分类体系研究[J]. 地理信息世界, 2022, 29(5): 112-117. DOI: 10.3969/j.issn.1672-1586.2022.05.020.
XU J Y, WANG X, ZHANG Z X, et al. Research on the land resources classification system for multi-scale remote sensing intelligent interpretation[J]. Geomatics World, 2022, 29(5): 112-117. DOI: 10.3969/j.issn.1672-1586.2022.05.020.

[5] AJMAL H, REHMAN S, FAROOQ U, et al. Convolutional neural network based image segmentation: a review[C]//Proceedings of SPIE: 10649, Pattern Recognition and Tracking XXIX. Bellingham: SPIE, 2018: 191-203. DOI: 10.1117/12.2304711.

[6] TALUKDAR S, SINGHA P, MAHATO S, et al. Land-use land-cover classification by machine learning classifiers for satellite observations—A review[J]. Remote Sensing, 2020, 12(7): 1135. DOI: 10.3390/rs12071135.

[7] QIN R J, TIAN J J, REINARTZ P. 3D change detection–Approaches and applications[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 122: 41-56. DOI: 10.1016/j.isprsjprs.2016.09.013.

[8] ZHAO W Z, DU S H. Spectral–spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(8): 4544-4554. DOI: 10.1109/TGRS.2016.2543748.

[9] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3431-3440. DOI: 10.1109/CVPR.2015.7298965.

[10] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Heidelberg: Springer, 2015: 234-241. DOI: 10.1007/978-3-319-24574-4_28.

[11] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 801-818. DOI: 10.1007/978-3-030-01234-2_49.

[12] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 2881-2890. DOI: 10.1109/CVPR.2017.660.

[13] GUO Y J, WANG F, XIANG Y M, et al. DGFNet: dual gate fusion network for land cover classification in very high-resolution images[J]. Remote Sensing, 2021, 13(18): 3755. DOI: 10.3390/rs13183755.

[14] WEI H R, XU X Y, OU N, et al. DEANet: dual encoder with attention network for semantic segmentation of remote sensing imagery[J]. Remote Sensing, 2021, 13(19): 3900. DOI: 10.3390/rs13193900.

[15] FU J, LIU J, TIAN H J, et al. Dual attention network for scene segmentation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3146-3154. DOI: 10.1109/CVPR.2019.00326.

[16] KHAN S, NASEER M, HAYAT M, et al. Transformers in vision: a survey[J]. ACM Computing Surveys, 2022, 54(10s): 200. DOI: 10.1145/3505244.

[17] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//Proceedings of the 9th International Conference on Learning Representations. San Diego: ICLR, 2021.

[18] ZHENG S X, LU J C, ZHAO H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 6881-6890. DOI: 10.1109/CVPR46437.2021.00681.

[19] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 10012-10022. DOI: 10.1109/ICCV48922.2021.00986.

[20] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662. DOI: 10.1109/TPAMI.2019.2938758.

[21] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

[22] ZHANG C, JIANG W S, ZHAO Q. Semantic segmentation of aerial imagery via split-attention networks with disentangled nonlocal and edge supervision[J]. Remote Sensing, 2021, 13(6): 1176. DOI: 10.3390/rs13061176.

[23] BOKHOVKIN A, BURNAEV E. Boundary loss for remote sensing imagery semantic segmentation[C]//Proceedings of the 16th International Symposium on Neural Networks. Heidelberg: Springer, 2019: 388-401. DOI: 10.1007/978-3-030-22808-8_38.

[24] BOGUSZEWSKI A, BATORSKI D, ZIEMBA-JANKOWSKA N, et al. LandCover. ai: dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2021: 1102-1110. DOI: 10.1109/CVPRW53098.2021.00121.

[25] TONG X Y, XIA G S, LU Q K, et al. Land-cover classification with high-resolution remote sensing images using transferable deep models[J]. Remote Sensing of Environment, 2020, 237: 111322. DOI: 10.1016/j.rse.2019.111322.

作者简介:

汤泊川　硕士研究生, tbczzz@163.com

帕力旦·吐尔逊(通信作者)　博士, 副教授, pldtrs@xjnu.edu.cn