多任务协同的多模态遥感目标分割算法

毛秀华* 张 强 阮 航 杨雨昂
(北京跟踪与通信技术研究所 北京 100094)
(天基综合信息系统全国重点实验室 北京 100094)

摘 要：利用语义分割技术提取的高分辨率遥感影像目标分割具有重要的应用前景。随着多传感器技术的飞速发展，多模态遥感影像间良好的优势互补性受到广泛关注，对其联合分析成为研究热点。该文同时分析光学遥感影像和高程数据，并针对现实场景中完全配准的高程数据不足导致两类数据融合分类精度不足的问题，提出一种基于多模态遥感数据的多任务协同模型(UR-PSPNet)，该模型提取光学图像的深层特征，预测语义标签和高程值，并将高程数据作为监督信息嵌入，以提升目标分割的准确性。该文设计了基于ISPRS的对比实验，证明了该算法可以更好地融合多模态数据特征，提升了光学遥感影像目标分割的精度。

关键词：语义分割；遥感影像；多模态；深度学习；高程估计
中图分类号：TP751 **文献标识码**：A **文章编号**：1009-5896(2025)03-0001-09
DOI: 10.11999/JEIT231267

Multitask Collaborative Multi-modal Remote Sensing Target Segmentation Algorithm

MAO Xiuhua ZHANG Qiang RUAN Hang YANG Yuang
(*Beijing Institute of Tracking and Telecommunications Technology*, *Beijing* 100094, *China*)
(*National Key Laboratory of Space Integrated Information System*, *Beijing* 100094, *China*)

Abstract: The use of semantic segmentation technology to extract high-resolution remote sensing image object segmentation has important application prospects. With the rapid development of multi-sensor technology, the good complementary advantages between multimodal remote sensing images have received widespread attention, and joint analysis of them has become a research hotspot. This article analyzes both optical remote sensing images and elevation data, and proposes a multi-task collaborative model based on multimodal remote sensing data (United Refined PSPNet, UR-PSPNet) to address the issue of insufficient fusion classification accuracy of the two types of data due to insufficient fully registered elevation data in real scenarios. This model extracts deep features of optical images, predicts semantic labels and elevation values, and embeds elevation data as supervised information, to improve the accuracy of target segmentation. This article designs a comparative experiment based on ISPRS, which proves that this algorithm can better fuse multimodal data features and improve the accuracy of object segmentation in optical remote sensing images.

Key words: Semantic segmentation; Remote sensing images; Multi-modal data; Deep learning; Elevation estimation

1 引言

随着通信手段和遥感技术的快速发展，各类新型传感器不断涌现，为对地观测相关研究提供更加便捷的方式和手段，多源传感器能够从不同的时间、空间维度采集图像，进而为资源勘测、灾难预警、城市规划等现实应用场景提供了海量数据。遥感领域传统研究多是基于单源数据开展，但受到传感器等设备成像原理的限制，对于不同设备所采集的图像，其蕴含的特征信息在空间分辨率、时间分辨率和光谱分辨率等方面都存在较大差异，因此对于单源遥感数据的研究在实际应用中有一定的局限性。考虑到各类遥感图像间的信息可以进行互补利用，对不同模态的数据源进行综合分析成为目前的研究热点之一[1]。

目标语义分割和高程估计两个任务作为获取图像内容特征信息的重要手段，是计算机视觉在遥感领域的研究热点内容。目标语义分割任务对图像进

收稿日期：2023-11-05；改回日期：2024-03-27；网络出版：2024-05-11
*通信作者： 毛秀华， 13718227932@139.com

行了精细的像素级密集预测和边界划分，从而使每一个像素点都被标记为对应的类别标签。高程估计则是推断物体的高程信息，为场景理解提供了像素级的指导。近些年，基于深度卷积神经网络的研究在计算机学科发展迅速，相关研究内容被广泛应用于图像分类、目标检测和分割等领域，取得了较为突出的成果。

单模态遥感目标语义分割任务通常只从单一角度观测场景的特点，缺乏实例目标信息的完整性和丰富性。利用多模态丰富的特征可以弥补单一模态特征的不足，提高复杂遥感场景中实例目标解译的性能。例如，遥感图像通常是由传感器设备以向下的视角拍摄成像的，这种单一视角的图像具有地物的空间和纹理信息，却缺乏地物目标的高程信息。近年来，随着遥感成像技术的不断进步，有些传感器也实现了用数字化的形式对地面地形信息进行模拟的功能。如生成数字地表模型(Digital Surface Model, DSM)[2]，该模型能够以数字阵列的模式，对包含了树木、建筑物等地表要素的整体地形进行数字描述。根据较新的研究结果，DSM所蕴含的高程信息可以显著地提高目标语义分割任务的分类结果。

在目前单一的目标语义分割任务中，主要存在两类问题：具有相似外观的物体很容易被错误分类，因为不同的物体在遥感图像中可能具备相似的特征，这称为类间相似性；具有不同外观的同类物体，也容易被错误分类到不同的标签，这称为类内相异性。

在深度学习的目标语义分割方法研究方面，研究人员进行了大量的研究。Long等人[3]提出全连接卷积神经网络(Fully Convolutional Networks, FCN)，首次将卷积神经网络应用于目标语义分割，提出了编码器、解码器网络，提升了图像分类结果的准确性。Zhao等人[4]提出了金字塔场景分析网络(Pyramid Scene Parsing Network, PSPNet)，使用金字塔池化模块(Pyramid Pooling Module, PPM)，用于聚合上下文的特征信息，有效编码丰富的语义信息。Mou等人[5]提出了一个端到端的包含残差学习结构的卷积-反卷积网络架构，模拟从单目的遥感图像到高程数据之间的映射关系。Ghamisi等人[6]使用生成对抗网络模型，提出了从单个光学图像模拟数字表面模型的架构，能够对不存在相应高程信息的目标场景进行高程信息的预测。Yuan等人[7]提出一种基于变换器和卷积神经网络的多尺度通道融合模型，通过学习全局到局部的上下文信息，增强了语义特征，并在处理低分辨率图像细节和提取全局图像特征方面效果较好。Weng等人[8]提出一种双路并行网络结构模型，通过特征耦合模块融合局部信息和全局语义信息，可处理高分辨图像中的细节信息，并提升了模型的泛化能力。Hao等人[9]提出了一种基于U-net改进网络的多目标语义分割算法，该方法可以提升目标区域内各类型特征的识别效率。Lü等人[10]提出了一种混合注意力的语义分割网络，该网络通过对多尺度目标的大视野来提取目标及其周围环境，提升了语义分割的准确性。Zhang等人[11]提出了一种多模态分割基础模型，利用多模态融合提升了语义分割的鲁棒性。

为提高复杂场景下遥感图像的目标分割的性能，本文在PSPNet的基础上，改进并设计了一种基于多模态遥感图像的多任务模型，进行语义分割和高程估计，以完成对目标分割，具体工作如下所示。(1)在ISPRS[12]提供的竞赛数据集上，获取光学图像和高程图像两种不同模态的数据，同时获得了可用于语义分割的对应标签数据，并对原始数据进行划分和增强处理。(2)在PSPNet基础上，改进并设计了一种基于多模态遥感图像的多任务模型，同时进行语义分割和高程估计的任务，以完成对实例目标分割。通常情况下高程信息仅作为输入信息，但本文创新性的提出将高程信息作为模型计算过程中的监督信息输入模型，以提升模型分类精度。同时，在实际应用场景中，难以获取到完全配准的多模态图像，因此，本文设计的模型可以仅通过输入光学图像，即可获得高程估计和语义分割的预测结果。在共享编码器的基础上，在解码器部分设计两个分支，对多模态数据的真实值进行多任务损失计算并不断优化模型参数。与不加入高程信息的单模态语义分割模型相比，本文提出的模型模型能够同时完成多任务的预测，且指标正常收敛，表明本文设计的融入高程信息的方式有效。(3)考虑到多任务的直接相关性，对加入了高程信息的多模态模型进行调整，设计了一个多任务协同的多模态网络模型，使完成语义分割和高程估计的两部分数据共享编码器，同时，在解码器部分设计协同模块，使得语义特征与高程特征为彼此提供额外有益信息，进而提升任务的精度，达到相互促进、互相指导的作用。

最终在ISPRS的Vaihingen和Potsdam语义标签竞赛数据集上进行实验，验证了本文提出的方法可以在一定程度上提升目标分割的准确性。

2 基于多模态遥感图像的目标分割模型分析

2.1 模型整体结构

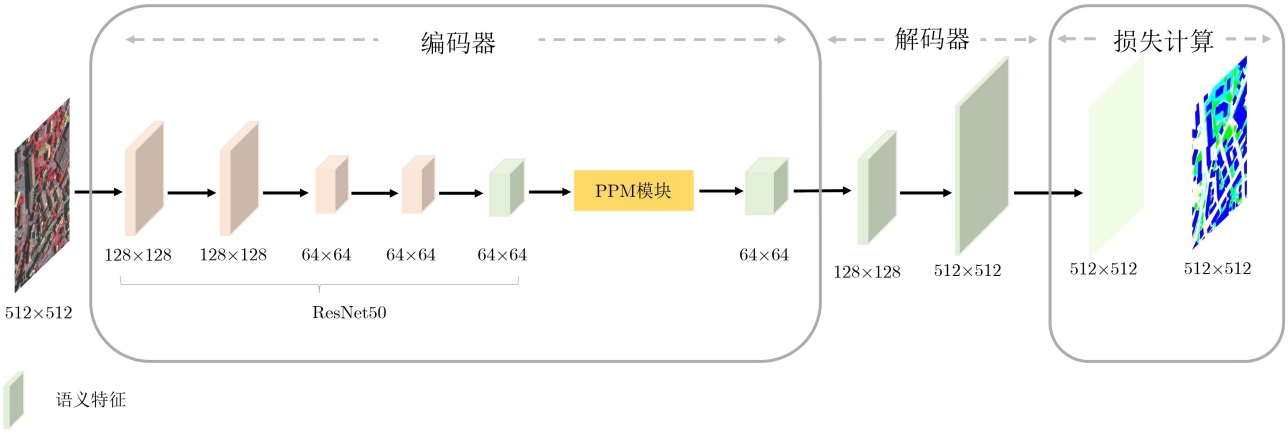
语义分割和高程估计在各自的计算过程中本身就具有很强的相关性，且本次实验的主要目的是妥善处理多任务学习中存在的共性、特性和依赖性，实现不同任务间的语义交互，提高准确度。因此，

为了考虑充分利用语义特征和高程信息的相关性，并联合完成多任务的目标。本文基于PSPNet模型，提出了多模态遥感图像的多任务协同模型UR-PSPNet(United Refined-PSPNet)，如图1(b)所示，在编码器阶段，首先原始图像经过改进后的ResNet50网络，并分解为语义及高程两路分支，然后两路分支逐像素相加后经过PPM模块，最后输出编码器结果。在解码器阶段，在语义分支中利用协同模块拼接高程特征，对经过下采样的图像补充缺失的纹理和高程信息，进而最终提升目标分割精度。该模型改善了单模态遥感图像下易出现模型分类混淆的问题，并且能同时完成语义分割、高程估计和多任务目标。

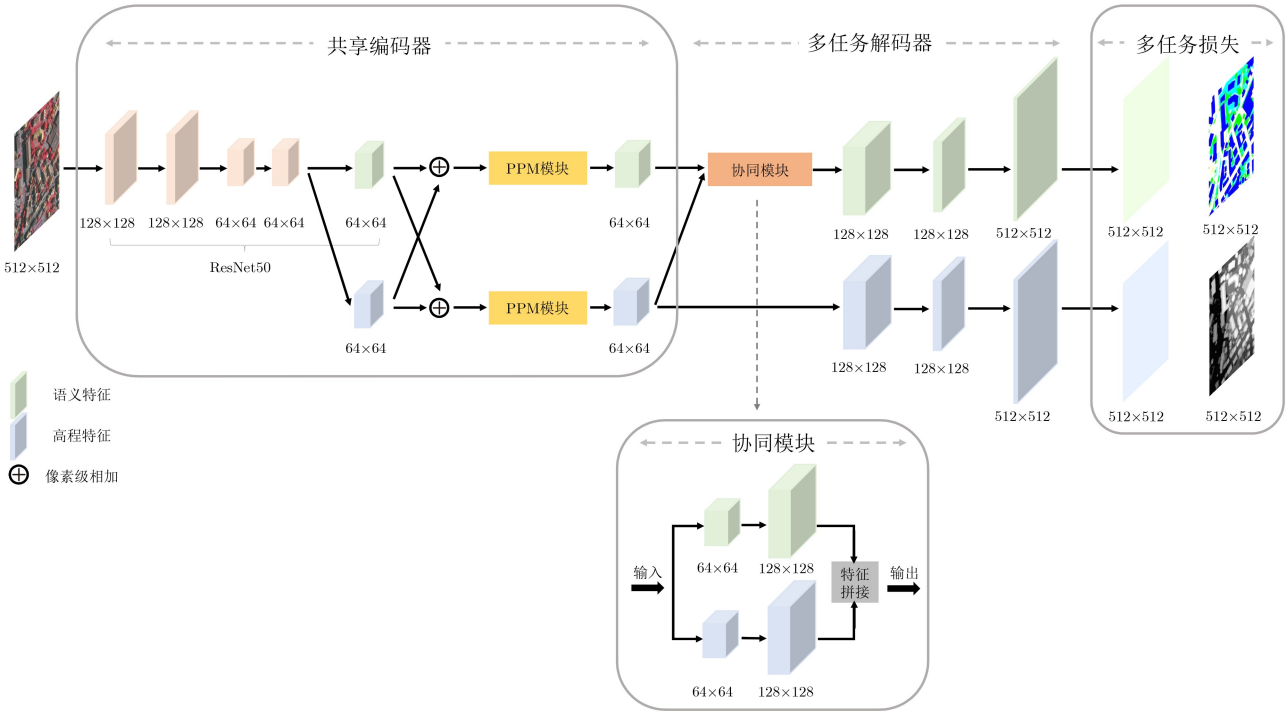
UR-PSPNet网络模型的整体计算流程如下：(1)以单独的光学遥感图像作为输入，通过一个部分共享的编码器提取其深层特征；(2)深层特征通过两个特定任务的解码器，分别对应各自的任务并进行特征恢复；(3)在解码器中，加入协同模块，对语义分割及高程估计任务的特征进行有效的融合，进而提升任务精度；(4)分别预测语义分割的标签图和高程估计任务中每一个像素点的高程值；(5)利用多任务损失函数进行联合的损失计算并反向传播，迭代并优化模型。

2.2 编码器

本文设计的UR-PSPNet网络模型由主干网ResNet50和金字塔池化PPM模块连接组成的部分



(a) PSPNet网络结构



(b) UR-PSPNet网络结构

图 1 PSPNet以及UR-PSPNet网络结构

共享的编码器结构，图像首先经过修改过的ResNet50主干结构后，提取深层特征，生成尺寸为原图大小1/8的特征图，并形成两个分支，进一步经过PPM模块完成多尺度特征的拼接操作，主要目的为更好的聚合上下文的特征信息。

在本实验中，编码器选择ResNet50结构作为主干网络，并对其进行如下修改：

首先，原始的ResNet网络的最后一层为全连接层，而本次实验的主要任务是逐像素输出类别的语义分割，因此取消原始的全连接层，而是将提取到的深层特征直接与后面的网络结构相连。

其次，为了后期能够妥善处理多任务学习中存在的共性，从两个角度同时对光学图像提取特征用于实验，强化语义交互过程。在本实验中，对ResNet50网络中Block4的部分进行改变，扩展了一个高程分支。即特征经过Block3后，同时进行了两个分支的特征提取任务，通过这种方式，将整体模型划分流向。

最后，在进入Block4模块之前，对于原图提取到的语义和高程的深层特征，分别再通过一个额外分支计算损失，并直接预测语义分割和高程估计任务的结果，作为辅助损失，按照设置好的权重，用于参与本次实验的多任务损失计算过程，因此，最终的损失被分解为四项。辅助损失仅在训练过程中承担一部分损失的传递，在验证和测试的过程中不参与预测。之后特征将分别进入各自的金字塔池化结构，即PPM模块。将对两个分支中的特征图各自进行多尺度池化处理，由于以不同尺度进行池化后的特征图，以不同的感受野捕获了图像的信息，因此在将多个尺度进行拼接后获得的复合特征图，对图像的全局和局部信息充分聚合以联合预测。本次实验分别采用了1*1(*表示卷积运算),2*2,3*3和6*6的尺度进行池化操作后，均连接了1*1的卷积层，从而获得尺寸不同，但通道数相同的子特征图，再将子特征图以插值上采样的方式，还原至与池化前的图像相同的大小，并进行整体拼接，语义和高程分支均得到了通道数为4096的特征图。

2.3 解码器

多任务解码器结构在通常情况下，解码器部分用于将抽象的深层特征恢复至原图的分辨率大小，还原重建其本身的空间结构信息。在解码器部分使用反卷积和双线性插值的方法更为普遍，但由于在编码器的特征提取过程中，已经对图像进行了极易丢失原本纹理和空间信息的大幅度下采样，图像在解码恢复后很难完全恢复原始的特征。因此，本文提出了高程估计和语义分割协同模块，使两个任务在上采样阶段充分考虑双方特征，分支之间互补缺失的纹理或高程信息，提高两个任务的精度结果。

本文提出在解码器部分使用多任务协同模块，多任务协同模块具体方法为首先将编码器输出的语义特征和高程特征两个分支分别进行卷积操作提取图像特征，然后利用双线性插值的上采样操作，使语义分支和高程分支图像由65×65恢复至130×130的分辨率大小，通道数目由4096压缩至2048，最后将高程分支的特征拼接至语义分支，使语义分支融合高程分支特征，通道数由2048扩充为4096。经由协同模块输出的语义特征分支再和高程特征分支分别进行卷积核为3*3和1*1的两次卷积操作，进行最后的特征提取，此时语义分支的通道数为6，即语义分割的类别数，高程分支的通道数为1。最后，经过上采样操作，得到恢复至原图大小的语义特征图和高程真值。

本文中上采样方式均为双线性插值，由于该过程不需要学习，能通过不增加额外参数的方式恢复图像的纹理特征，运行速度上与反卷积相比具有明显优势。

2.4 损失函数

损失函数在深度学习中十分重要，根据具体任务妥善选择损失函数的计算方法，对整体模型的更新提升至关重要。对于一个传统的语义分割任务，假设给定一个输入图像满足$X \in R^{3\times H\times W}$，训练的模型为$M_\theta$，真实的类别标签值为$S$，那么传统训练过程则可以看作为一个完全优化问题，如式(1)所示。其中，E表示统计期望，ℓ为语义分割的损失函数，如交叉熵损失。

$$\min_{\ell} \mathrm{E}\left[\ell\left(M_\theta(X), S\right)\right] \tag{1}$$

本文为实现多任务而提出的多模态模型M的优化问题可以归纳为式(2)。

$$\min_{\ell_1,\ell_2} \mathrm{E}\left[\ell_1\left(M_\zeta(X), H\right) + \ell_2\left(M_\eta(X), S\right)\right] \tag{2}$$

其中，$\mathrm{E}\left[\ell_1\left(M_\zeta(X), H\right)\right]$为高程损失的估算项，这里$H$代表了包含几何信息的高程数据，$\ell_1\left(M_\zeta(X), H\right)$是训练高程估计网络模型$M_\zeta$的回归损失值。第2项$\mathrm{E}\left[\ell_2\left(M_\eta(X), S\right)\right]$，类似于式(1)，用于语义分割模型提取语义信息，其中M_ζ与M_η对编码器中共享的部分共享权重。

即在本次实验提出的UR-PSPNet模型中，同时基于语义分割和高程估计这两个预测任务进行了联合的损失计算，在追求损失函数最小化的过程中，使模型根据约束，不断提高学习性能和预测能力，减少与真值之间的差距。

2.4.1 语义分割损失函数

语义分割的目标是分析图像中的每一个像素点的所属类别，并将其标记为对应的标签。本次实验对该任务部分采用交叉熵损失函数(Cross entropy loss function)，如式(3)所示。其中，N表示图像中物体的类别数，P_i表示第i个类别预测结果的概率分布，$\tilde{p}_i$表示第i个类别真实值的概率分布。交叉熵损失能够度量真实标签和模型的预测标签之间的误差，再将误差值反向传播，为模型的优化方向提供指导。

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\tilde{P}_i \cdot \ln\left(\frac{\mathrm{e}^{p_i}}{\sum_{j=1}^{N}\mathrm{e}^{P_i}}\right) \tag{3}$$

2.4.2 高程损失函数

在高程估计任务中使用平滑的L_1损失，如式(4)所示。

$$L_h = \sum_i^n \begin{cases} 0.5(h_i - H_i)^2, |h_i - H_i| \le 1 \\ |h_i - H_i| - 0.5, \text{其他} \end{cases} \tag{4}$$

其中，i为位置索引，n表示图像中的像素总数。这里h_i和H_i分别表示位置i对应的预测的输出高程和真实高程，通过二者的差值表示目前的损失。

2.4.3 多任务损失函数

为了实现两个任务的协同处理，并实现损失的反向传递，本次实验同时计算两个任务的损失函数，并对两个数值线性组合目标值作为模型整体的损失值，进行计算和反向传递，计算为

$$L = \lambda_1 L_{\mathrm{S}} + \lambda_2 L_{\mathrm{a}} + \lambda_3 L_{\mathrm{h}} + \lambda_4 L_{\mathrm{a'}}, \tag{5}$$

L_{S}和L_{h}分别为上文中定义的语义分割和高程估计任务中的损失计算结果。在ResNet50的第3阶段之后，对低分辨率的图像直接进行了一步卷积和预测，该预测结果与真实值的损失作为任务的辅助损失，在多任务损失函数中分别表达为语义分割任务的辅助损失L_{a}，高程估计的辅助损失$L_{\mathrm{a'}}$，以帮助优化学习过程。辅助损失仅参与训练过程，不参与验证和测试过程。参数λ_1, λ_2, λ_3和λ_4为4项损失的权重。λ_1和λ_3对应着语义分割和高程估计任务的重要性，在本次实验中，目标是对两项任务进行协同优化，因此重要性相同，参考PSPNet模型的参数设置，将λ_1和λ_3参数值设为1，λ_2和λ_4为辅助损失的权重，参数值设为0.4。

3 实验结果与分析

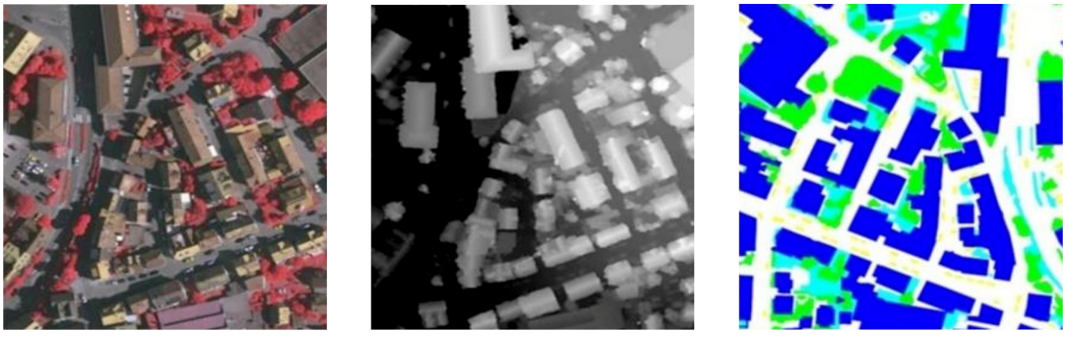
3.1 实验数据集详情

本次实验采用的多模态数据具体指由不同成像手段采集的观测数据，数据采用ISPRS提供的Vaihingen和Potsdam语义标签竞赛数据集。其中Vaihingen数据集同时包含光学数据和高程数据两种模态各33张。其图像的空间分辨率为9cm，其中有16张图像提供了可用于训练的对应的标签图像，包含6个目标类型，包括不透水的表面、建筑、低植被、树木、汽车和背景干扰，Vaihingen数据集示例如图2所示。Potsdam数据集是来自于同一张大的遥感正射图像中提取图像，每张遥感图像的分辨率为5cm, Potsdam数据集示例如图3所示。

本次实验按照该数据集通用的划分方法，在训练、验证和测试集的选取上与其他基于此数据集的研究方法保持一致。训练集具有11对影像(对应编号为1, 3, 5, 7, 13, 17, 21, 23, 26, 32和37)，验证集具有5对影像(对应编号为11, 15, 28, 30和34)。由于GPU资源有限，本次实验使用一个重叠的(50%)滑动窗口将原始的大图像裁剪成大小为512×512的切片。最后，得到721个切片用于训练，325个切片用于验证。在测试阶段，将剩余的17张图像裁剪为398个切片来进行测试。在本次研究中，训练集和验证集中的图像是相互独立的，这保证了实验结果的有效性，划分信息如表1所示。

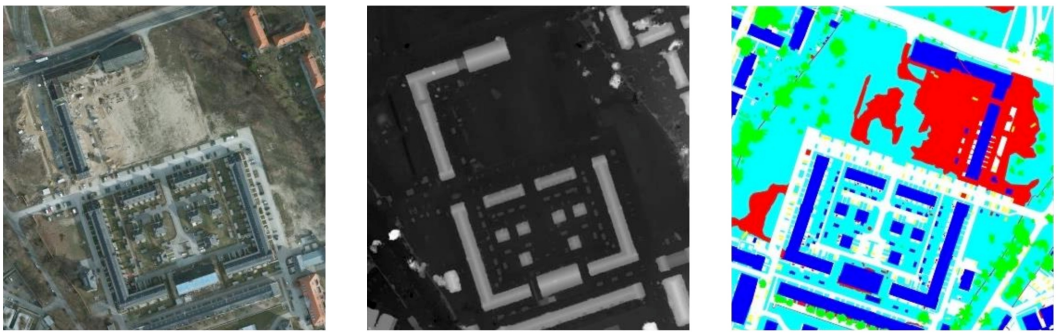
3.2 软硬件环境及实验设置

本次实验使用64位的Linux-3.10.0系统，硬件运行环境为NVIDIA Telsa P100显卡，深度学习框



(a) 示例子图　(b) 高度标签　(c) 语义标签

图 2 Vaihingen数据集提供图像



(a) 示例子图 (b) 高度标签 (c) 语义标签

图 3 Potsdam数据集提供图像

表 1 数据集划分信息

数据集	训练集样本数	验证集样本数	测试集样本数
Vaihingen	721	325	398
Potsdam	721	325	398

架基于Python编程语言，使用Pytorch1.1.0框架进行实验。

为了使网络更加轻量化，优化计算效率，本次实验采用了ResNet-50和金字塔池化模块PPM作为部分共享的编码主干。采用ImageNet图像集初始化预训练模型，经过编码器后的输出图像分辨率大小设置为原图的1/8。该网络采用随机梯度下降(Stochastic Gradient Descent, SGD)进行端到端训练，权值衰减为0.001，动量为0.9。使用幂为0.9的poly学习率策略，初始学习率设置为0.005。引入了迭代为100和比值为0.1的线性热身，以避免初始训练阶段的不稳定。此外，实验将输入的图像裁剪为512×512大小，并将批处理大小设置为4，迭代次数设置为323次，引入了数据增强策略，如随机翻转、调整大小、旋转和高斯模糊等。

本文设计的编码器整体结构和参数设置如表2所示，解码器的整体结构和参数设置如表3所示。

3.3 评价指标

3.3.1 语义分割指标

语义分割任务选择了通用的评估指标，包括总体精度(Overall Accuracy, OA)，交并比(Intersection over Union, IoU)，F_1分数的均值等。首先，假设TP代表真正类(True Positive)的数量，TN代

表 2 编码器具体网络结构及参数

模块	网络层	类型	核尺寸	输出图像尺寸
修改的ResNet50主干网模块	Conv1	卷积层×3	-，-	128×128
	Block1	残差块×3	-，-	128×128
	Block2	残差块×4	-，-	64×64
	Block3	残差块×6	-，-	64×64
	语义分支1	卷积层×2	3*3，256 3*3，6	64×64
	语义上采样1	双线性插值	-，6	512×512
	高程分支1	卷积层×2	3*3，256 3*3，1	64×64
	高程上采样1	双线性插值	-，1	512×512
	语义Block4	残差块×3	-，-	64×64
	高程Block4	残差块×3	-，-	64×64
	相加1	相加层	-，2048	64×64
PPM模块	语义分支2_1	全局平均池化	-，512	64×64
	语义分支2_2	卷积层	1*1，512	64×64
	语义拼接1	通道拼接层	-，2048	64×64
	高程分支2_1	全局平均池化	-，512	64×64
	高程分支2_2	卷积层	1*1，512	64×64
	高程拼接1	通道拼接层	-，2048	64×64

表真负类(True Negative)的数量，FP代表假正类(False Positive)的数量，FN代表假负类(False Negative)的数量。精确率(Precision)和召回率(Recall)的计算方法如式(6)和式(7)所示。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

F_1, OA和IoU等评价指标在语义分割任务的评估中较为通用，且数值越高，代表预测效果越好，定义为

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{9}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{10}$$

在这些评价标准中，计算时考虑了数据集中的每一个类，但考虑整体均值时，将最后一类，即背景和干扰类排除，以进行更加合理的比较，该计算方法也与类似的相关研究保持一致。

3.3.2 高程估计指标

在高程估计任务中，本次实验选择相对误差(Relative Error, Rel)、均方根误差(Root Mean Squared Error,Rmse)两项指标来评估高程估计分支的性能，定义为

$$\text{Rel} = \frac{1}{T}\sum\nolimits_{y^{\text{h}} \in T} \frac{\left|\hat{y}^{\text{h}} - y^{\text{h}}\right|}{y^{\text{h}}} \tag{11}$$

$$\text{Rmse} = \sqrt{\frac{1}{|T|}\sum\nolimits_{y^{\text{h}} \in T} \left\|\hat{y}^{\text{h}} - y^{\text{h}}\right\|^2} \tag{12}$$

其中，y^{h}代表了真实的高度数值，$\hat{y}^{\text{h}}$代表预测的高度值，T代表一幅图像中所有像素点的数目，上述公式度量的是预测值与真实值的差异，数值越小时，说明预测效果越好。

3.4 实验结果分析

3.4.1 定量分析

为了评估本文提出的模型性能，在Vaihingen和Potsdam数据集上从定量角度进行比较分析，如表4、表5所示。

通过表4的对比实验可以看出，在Vaihingen数据集上，FCN模型的mIoU为72.69，mF_1为83.74，OA为86.51。单模态的PSPNet模型mIoU为79.62，mF_1为88.48，OA为89.65。其他多模态模型包括MLHS, BAML以及I^2HN的实验结果如上表所示。

若仅考虑多任务学习模型，通过表5的对比实

表 3　解码器具体网络结构及参数

模块	网络层	类型	核尺寸	输出图像尺寸
协同模块	高程分支3	卷积层	1*1，2048	64×64
	高程上采样2	双线性插值	-，2048	128×128
	语义分支3	卷积层	1*1，2048	64×64
	语义上采样2	双线性插值	-，2048	128×128
	拼接1	通道拼接层	-，4096	128×128
解码器其他部分	语义分支4_1	卷积层	3*3，512	128×128
	语义分支4_2	卷积层	1×1，6	128×128
	高程分支4_1	卷积层	3*3，512	128×128
	高程分支4_2	卷积层	1*1，1	128×128
	高程上采样3	双线性插值	-，6	512×512
	语义上采样3	双线性插值	-，1	512×512

表 4　在Vaihingen数据集的实验结果

模型	mIoU	mF_1	OA	Rel	Rmse
FCN[3]	72.69	83.74	86.51	-	-
PSPNet[4]	79.62	88.48	89.65	-	-
MLHS[13]	77.03	85.68	86.70	0.2818	1.3416
BAML[14]	78.88	86.84	87.58	0.2580	1.3692
I^2HN[15]	79.62	87.41	89.13	0.2342	1.0143
UR-PSPNet(本文)	80.02	88.73	89.88	0.2182	0.9218

验可以看出，在Potsdam数据集上，UR-PSPNet模型的mIoU, mF_1及OA略高于MLHS, BAML以及I^2HN模型，Rel和Rmse略低于MLHS, BAML以及I^2HN模型。

利用多模态数据的多任务协同模型相较于不协同的模型而言，在语义分割和高程估计的任务精度上都有提升。以在Vaihingen数据集上的实验结果举例，本文提出的UR-PSPNet算法模型，mIoU为80.02，mF_1为88.73，OA为89.88，Rel为0.218 2，Rmse为0.921 8，运算速度为11.63帧/秒，相较于MLHS、BAML以及I^2HN模型而言，在检测结果中精准率、召回率、检测精度较高，进而虚警、漏警较少，该模型算法精度提升的主要原因为该算法是多任务协同的多模态网络模型，该模型在语义分割和高程估计层面共享部分编码器、解码器，一定程度上提高了复杂遥感场景下目标解译的性能，同时降低了受类间相似性和类内相异性的影响，对多模态的遥感数据源进行综合分析，进而提升了模型性能。经过分析，多任务协同的多模态模型UR-PSPNet拥有更好的性能。

3.4.2 定性分析

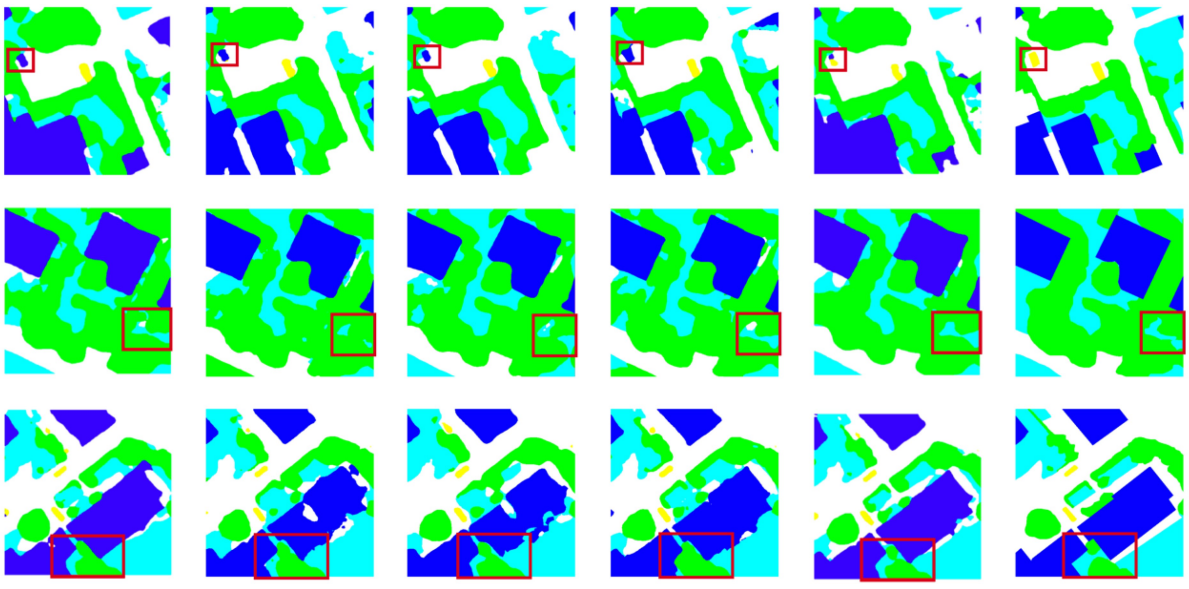
采用不同的网络模型PSPNet, MLHS, BAML, I^2HN和UR-PSPNet进行了多组实验，从定性的角度分析所提出的多任务协同模块对语义和高程联合任务的影响。以在Vaihingen数据集上的实验结果举例，实验结果表明，相较于单模态和多模态不协同的模型，多任务协同的多模态模型在提高分割性能上具有一定作用，如图4所示，不透水表面为白色，建筑物为蓝色，低植被为青色，树木为绿色，汽车为黄色，在要素颜色相似或红色框选的区域，因为高程数据的有效监督和跨任务交互，能够避免语义分割任务在面临相似纹理特征目标时的错误分类，用高程值对相似外观的类别进行修正，因此，多任务协同的多模态模型在分割性能上优于对比网络。

4 结束语

本文在PSPNet的基础上，改进并设计了一种基于多模态遥感图像的多任务模型UR-PSPNet，进行复杂场景下的语义分割和高程估计的任务，提升了目标分割的性能。在该模型中，考虑到多任务的直接相关性，设计了一个多任务协同的多模态网络模型，在语义分割和高程估计层面共享部分编码器、解码器，一定程度上提高了复杂遥感场景下目标解译的性能，同时降低了受类间相似性和类内相

表 5 在Potsdam数据集的实验结果

模型	mIoU	mF_1	OA	Rel	Rmse
MLHS [13]	81.65	88.52	86.97	0.080	1.095 4
BAML[14]	83.69	89.71	88.23	0.076 8	1.072 1
I^2HN[15]	83.72	89.63	87.45	0.061 7	0.618 6
UR-PSPNet(本文)	84.27	90.03	88.61	0.059 2	0.599 1



(a) PSPNet (b) MLHS (c) BAML (d) I^2HN (e) UR-PSPNet (f) 标签真值

图 4 实验可视化结果

异性的影响，对多模态的遥感数据源进行综合分析，以提升模型性能。在实测数据集ISPRS的Vaihingen和Potsdam语义标签竞赛数据集上进行实验，实验结果证明其精确率、召回率、检测精度较高，印证了本文提出的方法可以显著提升目标分割的准确性。

从本文的实验结果上可见，基于多模态遥感图像的多任务模型UR-PSPNet，在目标分割的性能上得到了提升，但同时我们也是认识到本文所适用的范围有限，还需要在实际业务中进行更深入及广泛的研究。例如在ISPRS数据集中，存在着较为明显的类别不均衡问题，如在图像中有大区域面积的目标是植被，小区域面积的是汽车，在模型的训练过程中很容易使小数目的类别分类效果不佳，需要在未来的实验中需要针对该问题改善模型，并在更多相关数据集上进行验证。

参 考 文 献

[1] 李树涛, 李聪妤, 康旭东. 多源遥感图像融合发展现状与未来展望[J]. 遥感学报, 2021, 25(1): 148–166. doi: 10.11834/jrs.20210259.
LI Shutao, LI Congyu, and KANG Xudong. Development status and future prospects of multi-source remote sensing image fusion[J]. *National Remote Sensing Bulletin*, 2021, 25(1): 148–166. doi: 10.11834/jrs.20210259.

[2] QIN Rongjun and FANG Wei. A hierarchical building detection method for very high resolution remotely sensed images combined with DSM using graph cut optimization[J]. *Photogrammetric Engineering & Remote Sensing*, 2014, 80(9): 873–883. doi: 10.14358/PERS.80.9.873.

[3] LONG J, SHELHAMER E, and DARRELL T. Fully convolutional networks for semantic segmentation[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3431–3440. doi: 10.1109/CVPR.2015.7298965.

[4] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, *et al*. Pyramid scene parsing network[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 6230–6239. doi: 10.1109/CVPR.2017.660.

[5] MOU Lichao and ZHU Xiaoxiang. IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network[J]. 2018. doi: 10.48550/arXiv.1802.10249. **(查阅网上资料,未能确认文献类型,请确认文献类型及格式是否正确)**.

[6] GHAMISI P and YOKOYA N. IMG2DSM: Height simulation from single imagery using conditional generative adversarial net[J]. *IEEE Geoscience and Remote Sensing Letters*, 2018, 15(5): 794–798. doi: 10.1109/LGRS.2018.2806945.

[7] YUAN Min, REN Dingbang, FENG Qisheng, *et al*. MCAFNet: A multiscale channel attention fusion network for semantic segmentation of remote sensing images[J]. *Remote Sensing*, 2023, 15(2): 361. doi: 10.3390/rs15020361.

[8] WENG Liguo, PANG Kai, XIA Min, *et al*. Sgformer: A local and global features coupling network for semantic segmentation of land cover[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023, 16: 6812–6824. doi: 10.1109/JSTARS.2023.3295729.

[9] HAO Xuejie, YIN Lizeyan, LI Xiuhong, *et al*. A multi-objective semantic segmentation algorithm based on improved U-Net networks[J]. *Remote Sensing*, 2023, 15(7): 1838. doi: 10.3390/rs15071838.

[10] LV Ning, ZHANG Zenghui, LI Cong, *et al*. A hybrid-attention semantic segmentation network for remote sensing interpretation in land-use surveillance[J]. *International Journal of Machine Learning and Cybernetics*, 2023, 14(2): 395–406. doi: 10.1007/s13042-022-01517-7.

[11] ZHANG Jiaming, LIU Ruiping, SHI Hao, *et al*. Delivering arbitrary-modal semantic segmentation[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 1136–1147. doi: 10.1109/CVPR52729.2023.00116.

[12] ROTTENSTEINER F, SOHN G, JUNG J, *et al*. The ISPRS benchmark on urban object classification and 3D building reconstruction[J]. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2012, 1(1): 293–298. doi: 10.5194/isprsannals-I-3-293-2012. **(查阅网上资料,未能确认文献类型,请确认文献类型及格式是否正确)**.

[13] CARVALHO M, LE SAUX B, TROUVÉ-PELOUX P, *et al*. Multitask learning of height and semantics from aerial images[J]. *IEEE Geoscience and Remote Sensing Letters*, 2020, 17(8): 1391–1395. doi: 10.1109/LGRS.2019.2947783.

[14] WANG Yufeng, DING Wenrui, ZHANG Ruiqian, *et al*. Boundary-aware multitask learning for remote sensing imagery[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, 14: 951–963. doi: 10.1109/JSTARS.2020.3043442.

[15] HE Qibin, SUN Xian, DIAO Wenhui, *et al*. Multimodal remote sensing image segmentation with intuition-inspired hypergraph modeling[J]. *IEEE Transactions on Image Processing*, 2023, 32: 1474–1487. doi: 10.1109/TIP.2023.3245324.

毛秀华：女，助理研究员，硕士，研究方向为人工智能、目标识别.
张　强：男，副研究员，博士，研究方向为人工智能、目标识别.
阮　航：男，副研究员，博士，研究方向为人工智能、目标识别.
杨雨昂：男，研究实习员，学士，研究方向为人工智能、目标识别.

责任编辑：马秀强