引文格式：张彦芬. CNN-VIT双分支遥感影像建筑提取方法研究[J]. 地理空间信息, 2026, 24(1): 1-6. doi:10. 3969/j.issn. 1672-4623.2026.01.001

CNN-VIT 双分支遥感影像建筑提取方法研究

张彦芬
（山西省测绘地理信息院，山西 太原 030001）

摘　要：针对现有模型提取遥感影像建筑目标时存在的提取结果数量缺失、形态不完整等问题，构建了一种结合卷积神经网络（CNN）与视觉转换器（VIT）的双分支编码器建筑提取模型。通过参数化重混排卷积单元搭建CNN分支，可变形自注意力转换器搭建VIT分支，在两组分支独立进行特征编码。在编码器与解码器间引入三元注意力重构的语义细节融合层，对多尺度特征图进行融合和负样本过滤。在解码器中，利用动态上采样单元对大尺寸特征图进行精密重建，并以广义Dice函数与焦点函数的加权组合计算训练损失。在MDD与WHU-Buliding（East Asia）数据集上的实验结果表明，该模型的F_1-score达到91.57%与92.34%，比Swin-UNet模型提高了7.14%与6.51%，提取的建筑目标形态更真实完整，提取精度和泛化性优于当前主流语义分割模型。

关键词：遥感建筑提取；CNN-VIT双分支编码器；三元注意力；语义细节融合；动态上采样
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2026）01-0001-06

Research on Building Extraction Method of Remote Sensing Image Based on CNN-VIT Dual Branch

ZHANG Yanfen
(Shanxi Institute of Surveying, Mapping and Geographic Information, Taiyuan 030001, China)

Abstract: Aiming at the problems of missing number and incomplete shape of extraction results when the existing models extracting building object from remote sensing images, we constructed a dual branch encoder building extraction model combining convolution neural network (CNN) and vision transformer (VIT). We used the parameterized convolution shuffle unit to build CNN branch, used the deformable self-attention transformer to build VIT branch, and performed the independent feature coding in the two branches. Between the encoder and decoder, we introduced the semantic and detail infusion layer of triplet attention reconstruction to fuse and filter the multi-scale feature map. In the decoder, we used the dynamic upsampling unit to precisely reconstruct the large-scale feature map, and used the weighted combination of the generalized Dice loss function and the focal loss function to calculate the training loss. The experimental results on MDD and WHU-Building dataset show that the F_1-score of this model is 91.57% and 92.34%, which is 7.14% and 6.51% higher than that of Swin-UNet model. The extracted building target shape is more authentic and complete, and is superior to the current mainstream semantic segmentation model in terms of extraction accuracy and generalization.

Key words: remote sensing building extraction, CNN-VIT dual branch encoder, triplet attention, semantic and detail infusion, dynamic upsampling

光学遥感卫星作为主要对地观测手段[1-2]，能采集大范围高分辨率遥感影像，其中包含各类地物目标。基于遥感影像的建筑检测能提取区域内的建筑个体[3-4]，为城市建设规划、土地资源调查、军事侦察等应用提供技术支持[5]。部分学者利用像素法、面向对象法[6]提取遥感影像内的建筑目标。这些方法通常涉及复杂的预处理过程，同时提取模型对建筑目标的深层特征学习能力有限，易受影像中相似目标、阴影等干扰，出现错检、漏检情况，导致提取精度和鲁棒性不理想[7]。然而，基于编码器—解码器架构的卷积神经网络（CNN）与视觉转换器（VIT）模型具备强大的深层特征学习能力，现已在多个语义分割领域得到广泛应用[8-11]。在各种提取模型中，CNN架构的编码器更关注感受野内目标的局部特征，但难以建立大尺度下的目标上下文依赖关系；而VIT架构的编码器虽能捕获全局尺度的目标上下文注意力，但更易丢失目标的边缘和纹理信息。遥感影像内部存在各类树木、阴影遮挡以及相似块状地物，对于细节特征稀疏的建筑目标，单一CNN或VIT架构易出现提取结果缺失、非建筑误提等情况，在准确性与完整性方面不够理想。基于现有模型实现高精度、高完整度的建筑提取，仍是一项具有挑战性的任务。鉴于此，本文提出

作者简介：张彦芬（1987—），硕士，高级工程师，从事大地测量、地形测量等方面的工作，E-mail：yy19899_YY@163.com。

了一种结合CNN与VIT架构的双分支编码器建筑提取模型。具体工作步骤为：①基于CNN-VIT架构构建双分支特征编码器，在不同分支上独立计算局部注意力和全局上下文注意力，并在编码器末端进行特征合并；②在编码器与解码器间设置了带有三元注意力（TA）单元的语义细节融合（SDI）模块，对编码器输出的多尺度特征图进行二次筛选与融合，为解码器提供多个尺度的融合特征图；③在特征解码阶段采用动态上采样机制对特征图进行精密重建，并以广义Dice（GDL）函数与焦点（FL）函数的加权组合构建新的损失函数。

1 双分支建筑提取模型

1.1 模型结构简介

模型采用编码器—解码器架构，编码器以单幅影像为输入，通过双分支结构进行特征编码，在编码器末端进行双分支特征融合；在编码器与解码器层之间，采用语义与细节融合结构对来自骨干网络的多尺度特征图进行融合与筛选。

1.2 双分支编码器

为充分捕获建筑目标的细节特征与全局上下文目标特征，以单幅遥感影像为输入，利用双分支编码器进行特征捕获。为限制特征编码过程中模型整体计算量，通过干细胞层无损下采样将输入特征图沿通道维度切分，每个分支仅使用一半通道的特征图进行特征编码。VIT分支采用视觉转换器机制计算全局尺度目标上下文注意力。目前主流的可移动窗口转换器[9]、金字塔视觉变压器[10]等VIT单元均在不带任何偏置的展平特征图上直接计算全局上下文多头注意力（MHSA）；而遥感影像内的建筑特征远少于其他背景地物，以无偏输入直接计算全局尺度的建筑上下文注意力，必然会产生海量背景噪声，导致所求的上下文注意力带有偏置。因此，为准确提取建筑目标的上下文注意力，利用可变形自注意力转换器（DSAT）搭建VIT分支。DSAT单元在尺寸为 $H\times W\times C$ 的特征 x 上绘制一个 $H_g\times W_g\times 2$ 的矩形均匀格网，其中 $H_g=H/r$，$W_g=W/r$，$r=2$ 为下采样步长，格网内均匀分布的参考点是后续偏移值累加的基础。通过偏移单元在查询（Query，$\boldsymbol{q}$）的基础上，计算得到变形偏移值；再将偏移值累加至各参考点，进行双线性插值采样，得到偏移后特征，即

$$\tilde{x}=\phi(x;p+\Delta p) \tag{1}$$

式中，x、$\tilde{x}$ 为偏移前、后的特征；$\phi(;)$ 为双线性插值采样函数；p 为各参考点上初始为0的偏移值；$\Delta p=\text{offset}(q)$ 为累加偏移值。

在 $\tilde{x}$ 上通过投影计算得到带有变形偏移的键与值向量，即

$$\boldsymbol{q}=\boldsymbol{x}\boldsymbol{W}_q,\tilde{\boldsymbol{k}}=\tilde{\boldsymbol{x}}\boldsymbol{W}_k,\tilde{\boldsymbol{v}}=\tilde{\boldsymbol{x}}\boldsymbol{W}_v \tag{2}$$

式中，$\boldsymbol{q}$、$\tilde{\boldsymbol{k}}$、$\tilde{\boldsymbol{v}}$ 分别为查询、键、值3组向量；$\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$ 为对应的投影矩阵。

在各组 $\tilde{\boldsymbol{k}}$ 与 $\tilde{\boldsymbol{v}}$ 共享一组 $\boldsymbol{q}$ 的前提下，根据 $\boldsymbol{q}$、$\tilde{\boldsymbol{k}}$、$\tilde{\boldsymbol{v}}$ 计算得到带有偏移值的MHSA。偏移单元由采样步长为 r 的3×3深度卷积（DWConv）、1×1的逐点卷积（PWConv）和高斯误差线性单元（GeLU）激活函数组成；DSAT由共享一组 q 的多个计算头构成（图1a）。对于第 m 个计算头，其上的自注意力计算过程为：

$$z^{(m)}=\sigma\left(\frac{q\tilde{\boldsymbol{k}}^{(m)\top}}{\sqrt{d}}+\phi\left(\hat{\boldsymbol{B}};\boldsymbol{R}\right)\right)\tilde{\boldsymbol{v}}^{(m)} \tag{3}$$

式中，d 为各计算头维度；$\hat{\boldsymbol{B}}$ 为各组 $\boldsymbol{q}$ 与 $\tilde{\boldsymbol{k}}^{(m)}$ 之间的相对位置偏移向量；$\boldsymbol{R}$ 为相对偏移矩阵；σ 为Softmax非线性激活函数。

对于第 m 个计算头的计算结果，采用按通道拼接的方式进行组合，即

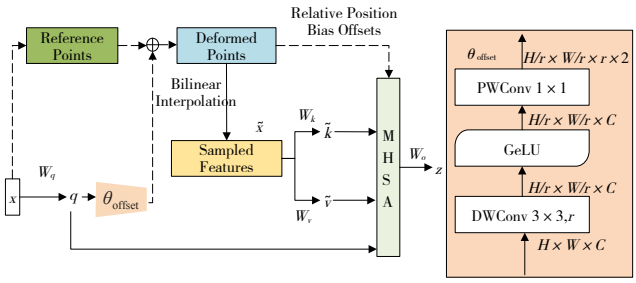
$$z=\text{Concat}(z^{(1)},z^{(2)},\cdots,z^{(m)})\times\boldsymbol{W}_0 \tag{4}$$

式中，z 为多头注意力；$\boldsymbol{W}_0$ 为投影矩阵。

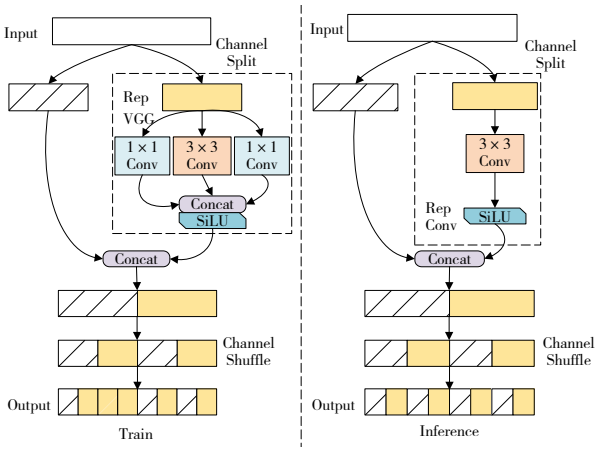
CNN分支采用标准卷积层计算多层感受野下的目标细节特征。为在充分捕获目标细节特征的同时，控制模型整体计算量，利用参数化重混排卷积（RCS）单元搭建CNN分支的编码层。RCS先切分输入特征通道，再输出一半特征图参与提取计算。在训练阶段，利用1×1和3×3尺度的两个并列卷积核同时提取不同尺度下的目标局部特征；再将所得特征图进行拼接与非线性激活，并与未参与提取的特征图进一步拼接；最后采用通道混排的方式打乱各通道间排列顺序，以保证各通道间特征信息流通。在推理阶段，为保证模型的计算效率，利用卷积核可加性，将1×1 Conv与3×3 Conv进行参数级合并，合并后的等效卷积核用于特征提取计算（图1b）。双分支编码器结构见图1c。

1.3 多层次特征融合—筛选模块

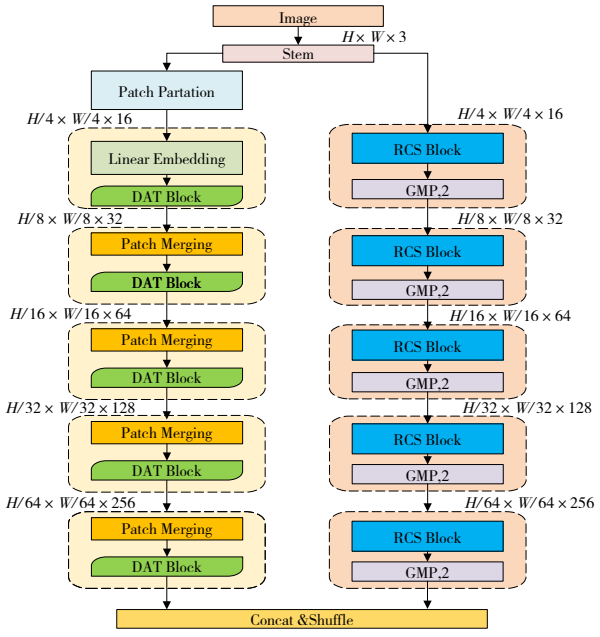
为了更有效地融合不同层级特征图中的语义与细节信息，并筛选无关负样本，本文在双分支编码器与解码器之间引入SDI单元，过滤融合来自编码器网络



a DSAT单元结构图



b CNN分支结构图



c 完整双分支编码器

图1　双分支编码器结构图

的多尺度特征图，为解码器输出多个尺寸的特征图。标准SDI单元利用通道—空间双层注意力单元对输入特征进行过滤，但上述两组注意力权重在计算过程中相互独立，因此不存在相互依赖关系，对负样本特征的过滤能力有限。因此，本文引入TA对SDI层进行改进，通过计算长度（H）、宽度（W）、深度（D）间的交互注意力权重，进一步过滤与目标无关的负样本特征。TA单元通过3组平行计算分支，单独计算 H、W、D 三个维度间的交互注意力权重。假设输入SDI层TA单元的特征图尺寸为 $C\times H\times W$，该特征图同时输入3组平行分支，上层分支首先将特征图绕 H 方向旋转以建立水平与通道 D 方向的联系，并将旋转后的特征通过Z–Pool池化为 $2\times H\times W$ 大小，计算公式为：

$$\mathrm{ZPool}(X)=[\mathrm{MP}(X),\mathrm{AP}(X)] \tag{5}$$

式中，X 为输入特征图；MP、AP分别为最大池化与平均池化。

然后通过 7×7 卷积—批量归一化—乙型激活函数（CBS）单元对特征进行处理以获取权重图。最后通过特征赋权与反向旋转，获得带有水平—通道方向间的交互注意力。中间分支则将特征图绕 W 方向旋转，其余计算过程与上层分支一致，建立 W 方向与通道方向间的交互注意力。下层分支将特征图直接以送入Z–Pool与CBS处理层，以捕获 H 与 W 之间的空间维度依赖关系。最终通过平均求和的方式聚合3组注意力图。改进后的SDI层结构见图2。

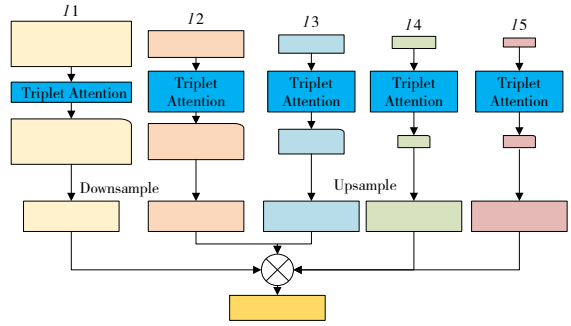


图2　改进的SDI层结构图

1.4　优化解码器

解码器结构的主要功能是融合来自编码器和TSDI层的多尺寸特征图，并逐级上采样，最终得到恢复尺寸且细节和上下文特征丰富的特征图。现有的解码器结构多采用双线性插值的方式进行尺寸放大，不能精细还原特征图内的目标特征。因此，本文引入动态采样（DyS）单元，通过点采样计算核精确重建大尺度特征图。DyS层首先通过 $H\times W\times 2gs^2$ 大小的线性变换层对 $H\times W\times C$ 大小的特征图进行压缩处理；然后通过像素混洗层对压缩后特征图进行偏移混排，得到 $sH\times sW\times 2g$ 大小的偏移矩阵 $\boldsymbol{O}$；再通过采样网格 G 对 $\boldsymbol{O}$ 进行像素级逐点相加，得到采样集 S；最后，以 S 为模板，按照网格采样的方式将输入特征图 X 上采样为 $2H\times 2W\times C$ 大小的输出特征图 X'，具体计算过程为：

$$S=PS(\mathrm{Linear}(X))+G \tag{6}$$

$$X' = \text{Grid Sample}(X, S) \tag{7}$$

标准U-Net、Swin-UNet等语义分割模型以Dice系数损失（DL）与交叉熵损失（CL）加权组合为损失函数[11]，其中CL函数用于评估模型对各像素点的分类情况，DL函数用于评估模型对建筑的整体分割效果。遥感影像中建筑样本与背景样本间的像素比例差异较大，在分割结果出现小幅偏差的情况下，基于交并比计算的DL函数变化不够显著，导致模型在训练阶段的学习收敛效率较低；而CL函数基于样本的平均损失来评估分类结果，同样会导致模型在训练阶段更关注背景像素，从而降低对建筑特征的学习效果。因此，本文引入GDL与FL函数的加权组合构建新损失函数。GDL函数的表达式为：

$$\begin{cases} \text{GDL} = 1 - 2 \times \dfrac{\sum_{l=1}^{L} W_l \sum_{n=1}^{N} r_n p_n}{\sum_{l=1}^{L} W_l (\sum_{n=1}^{N} r_n + \sum_{n=1}^{N} p_n)} \\ W_l = \dfrac{1}{\sum_{n=1}^{N} r_n^2} \end{cases} \tag{8}$$

式中，p_n、r_n为各像素的真实、预测标签；W_l为类别平衡权重；L为包括负样本在内的样本数量；N为总像素数量。

GDL通过引入W_l对不同类别样本进行加权，使低占比的建筑特征在损失计算中占据更高的比例，以提高损失函数对于检测结果变化的敏感性。FL函数的表达式为：

$$FL_{(p_n)} = -a_t (1 - p_n)^{\gamma} \log(p_n) \tag{9}$$

式中，a_t为类别权重，用于平衡正、负样本；γ为侧重难分类样本的调节参数。

相较于静态CL函数，FL函数通过γ动态调整困难或简单分类样本对于损失的贡献，以提升模型对建筑样本的关注度。在模型训练阶段，组合损失函数Loss的加权求和过程为：

$$\text{Loss} = 0.4 \times \text{GDL} + 0.6 \times FL \tag{10}$$

2　实验与结果分析

2.1　实验数据集

在集马萨诸塞建筑（MDD）数据集与武汉大学建筑数据集（东亚）（WHU-Buliding（East Asia））上开展模型训练与性能验证实验。MDD数据包括美国波士顿地区151张尺寸为1 500 px × 1 500 px，空间分辨率为1 m的卫星遥感影像；WHU-Buliding（East Asia）数据集包括17 388张尺寸为512 px × 512 px，空间分辨率为0.45 m的影像。按照8∶1.5∶0.5的比例将两个数据集划分为独立的训练、验证、测试子集；再采用迁移学习策略，让模型在两个数据集上完成独立训练。

2.2　实验参数

模型训练与测试实验均在台式工作站完成。工作站CPU为Intel Xeon(R) Platinum 8352V CPU，显卡为Nvidia RTX4090，运行内存大小为64 GB，硬盘容量为1 TB，操作系统为Ubuntu 18.04。模型和对比模型均在PyTorch 1.11.1框架下编写。训练阶段采用随机梯度下降参数优化器，动量参数为0.925 0，初始学习率为0.002 0，权重衰减系数为0.000 1，训练迭代次数为100次。每批次选取18张影像参与训练，训练过程中对图像进行随机裁剪反转。

为全面评估本文模型对遥感影像建筑目标的提取性能，选取精确率（P）、召回率（R）、F_1-score、平均交并比（mIoU）作为评价指标，对模型测试结果进行评价。

2.3　模型消融实验

为验证CNN-VIT双分支编码器在建筑目标特征捕获上的性能优势，以U-Net模型为基准，通过嵌入不同编码器构造多组消融模型，在MDD数据集上测试各消融模型的实际表现。为保证实验的客观性，模型的CNN分支均采用相同结构。结果表明，采用单一CNN分支的U-Net模型，F_1-score达到71.46%（mIoU为59.49%）；嵌入双层CNN分支后，F_1-score增至75.53%（mIoU为64.34%）；当编码器切换为VIT架构，Swin-UNet模型的F_1-score增至84.43%（mIoU为69.57%）；当采用CNN-VIT双分支编码器后，CNN-VIT-UNet的F_1-score达到86.39%（mIoU为78.48%），说明相较于单一分支编码器，双分支编码器模型的特征捕获能力更强，CNN-VIT双分支编码器更因其局部—全局上下文特征的联合捕获能力，在提取精度方面具有更明显的优势。部分消融实验结果见图3，可以看出，引入CNN-VIT双分支的消融模型，所提取建筑目标在准确性与完整性方面均优

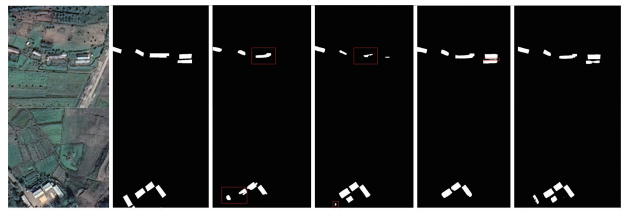


图3　编码器消融实验结果对比

于其他模型，证明了混合双分支编码器的性能优势。

针对在编码器与解码器之间插入SDI层并优化单元注意力的改进方案，以U-Net模型为基准，先后在U-Net模型中插入SDI和TSDI层构建消融模型，在MDD数据集上进行测试。结果表明，插入SDI层与TSDI层后，UNet-SDI、UNet-TSDI、U-Net模型的F_1-score分别为74.84%、78.53%、71.46%；mIoU分别为62.37%、69.58%、59.49%，证明了在编码器与解码器间引入语义特征融合层以及利用TA单元进一步筛选特征对于提高建筑目标分割精度的有效性。部分消融实验结果见图4，可以看出，引入中间层后对应消融模型的提取结果比U-Net模型更准确，而引入TSDI层的提取结果优于使用原生SDI层的消融模型。

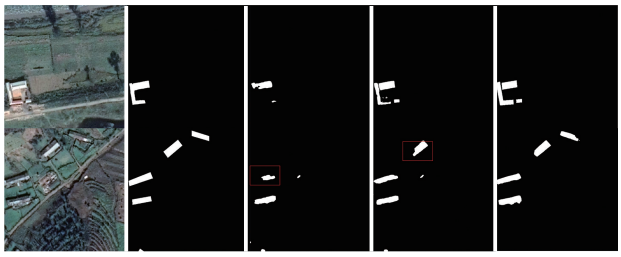


图4　SDI层改进消融实验结果对比

为验证在解码器中引入动态上采样结构与改进损失函数的有效性，以U-Net为基准模型，分别设置了针对上述两项优化方案的消融模型。结果表明，引入动态上采样层与改进损失函数后，UNet-DyS、UNet-GDL&FL、UNet-DyS-GDL&FL的F_1-score分别为73.53%、72.89%、76.42%，较U-Net模型上涨2.07%、1.43%、4.96%；mIoU分别为61.24%、60.83%、64.75%，说明采用上述措施改进解码器能提高建筑目标的提取精度。部分消融实验结果见图5，可以看出，相较于U-Net模型，各消融模型的提取结果更完整，且与真实标签在形态和目标数量上更接近，进一步证明了上述改进措施的有效性。

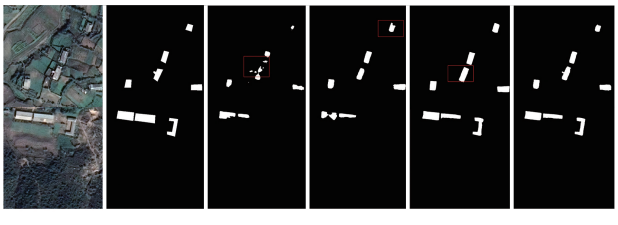


图5　解码器改进消融实验结果对比

2.4　模型对比实验

以当前主流语义分割模型PSPNet[12]、U-Net、SGCN[13]、UNet3+[14]、Swin-UNet[15]、MSA-Net[16]、SCA-Net[17]为对照组，分别在MDD与WHU-Buliding（East Asia）数据集上与本文模型进行对比实验。实验结果见表1、2，可以看出，在MDD数据集上，本文模型的P、R、F_1-score、mIoU优于对照组内的其他主流模型；在WHU-Buliding（East Asia）数据集上，本文模型的F_1-score和mIoU分别达到92.34%与83.26%，较PSPNet、U-Net、SGCN、UNet3+、Swin-UNet、MSA-Net、SCA-Net模型，分别提高了21.91%、19.49%、10.97%、9.05%、6.51%、3.14%、0.37%和25.40%、23.03%、17.02%、14.47%、12.91%、3.61%、0.43%，再次证明了本文模型在遥感建筑提取方面的性能优势。

表1　不同模型在MDD数据集上的对比实验结果/%

模型	P	R	F_1-score	mIoU
PSPNet	67.45	69.93	68.67	55.73
U-Net	70.56	72.42	71.48	59.49
SGCN	78.63	79.86	79.24	64.86
UNet3+	81.32	82.20	81.76	66.34
Swin-UNet	84.23	84.63	84.43	69.57
MSA-Net	88.57	87.34	87.95	78.62
SCA-Net	92.86	89.45	91.12	80.73
本文模型	92.54	90.62	91.57	81.62

表2　不同模型在WHU-Buliding（East Asia）数据集上的对比实验结果/%

模型	P	R	F_1-score	mIoU
PSPNet	69.34	71.55	70.43	57.86
U-Net	71.62	74.12	72.85	60.23
SGCN	80.54	82.22	81.37	66.24
UNet3+	82.65	83.94	83.29	68.79
Swin-UNet	84.79	86.90	85.83	70.35
MSA-Net	89.74	88.67	89.20	79.65
SCA-Net	93.36	90.63	91.97	82.83
本文模型	92.58	92.10	92.34	83.26

本文模型与各对照组模型在MDD与WHU-Buliding（East Asia）数据集上的提取结果对比见图6、7，可以看出，对照组模型的提取结果普遍存在漏检或提取

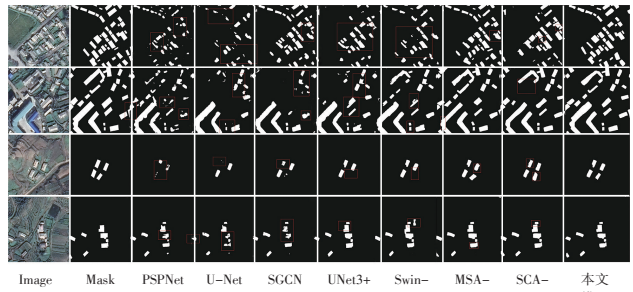


图6　不同模型在MDD数据集上的建筑提取结果

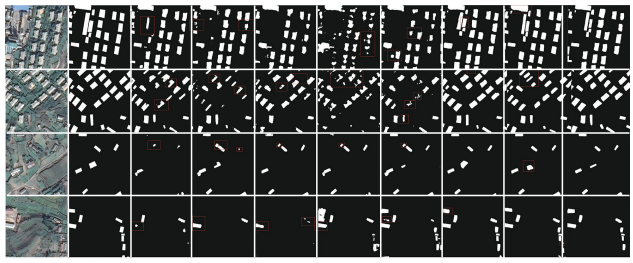


图7 不同模型在WHU-Buliding(East Asia)数据集上的建筑提取结果

结果空洞、断裂等问题，本文模型能提取出数量更准确、形态更完整且真实的建筑目标，说明CNN-VIT双分支编码器模型在遥感地面建筑提取方面具有更显著的优势。

3 结 语

针对现有模型受特征比例稀疏、背景要素复杂等因素影响，难以准确提取遥感影像内建筑目标的问题，本文提出了一种结合CNN-VIT单元的双分支遥感影像建筑提取模型。在双分支编码器中，CNN分支以重参数化重混排卷积提取多重感受野下的目标局部特征，VIT分支以DSAT捕获全局上下文注意力。为实现不同尺度特征的充分融合，以TA单元构建SDI层，对不同尺度的特征图进行融合和负样本过滤。在解码器中，使用动态上采样层实施精密重建，并以GDL与FL函数的加权组合作为损失函数。实验结果表明，本文模型在MDD和WHU-Buliding（East Asia）数据集上的F_1-score分别达到91.57%、92.34%，提取的建筑目标数量上更接近真实掩膜，且形态更真实完整，能为城市建设规划、土地资源调查等场景提供技术支持。在下阶段的研究中，将同时以光学高分辨率影像和SAR影像为输入源协同编码，实现多类地物提取。

参考文献

[1] 戴诗涛,张晶. 基于注意力机制与可变形卷积的遥感图像变化检测方法[J]. 地理空间信息,2024,22(10):22-25

[2] 覃现,郝容,邢鹏,等. 基于深度学习的国土资源遥感影像分类模型[J]. 地理空间信息,2024,22(10):34-37

[3] 景红霞. 尕海则岔自然保护区地表覆盖遥感监测[J]. 地理空间信息,2022,20(10):86-89

[4] 陆妍如,毛辉辉,贺琰,等. 基于Deeplab v3+的高分辨率遥感影像地物分类研究[J]. 地理空间信息,2022,20(6):1-6

[5] 杨彬,毛银,陈晋,等. 深度学习的遥感变化检测综述:文献计量与分析[J]. 遥感学报,2023,27(9):1 988-2 005

[6] Li J, Huang X, Tu L, et al. A Review of Building Detection from very High Resolution Optical Remote Sensing Images[J]. GIScience and Remote Sensing , 2022 , 59(1): 1 199-1 225

[7] 邢云飞,刘萍,谢育斑,等. 基于高分辨率遥感影像的面向对象建筑物分级提取方法[J]. 航天返回与遥感,2023,44(4):88-102

[8] Fu W, Xie K, Fang L. Complementarity-aware Local-global Feature Fusion Network for Building Extraction in Remote Sensing Images[J]. IEEE Transactions on Geoscience and Remote Sensing,2024,62:1-13

[9] Zhou Z, Zhang J, Gong C. Hybrid Semantic Segmentation for Tunnel Lining Cracks Based on Swin Transformer and Convolutional Neural Network[J]. Computer-aided Civil and Infrastructure Engineering,2023,38(17):2 491-2 510

[10] Wang W, Xie E, Li X, et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions[C]//Proceedings of the IEEE/CVF International Conference on Computer Cision,2021:568-578

[11] Gao Z J, He Y, Li Y. A Novel Lightweight Swin-UNet Network for Semantic Segmentation of COVID-19 Lesion in CT Images[J]. IEEE Access,2022,11:950-962

[12] Yin Y, Guo Y, Deng L, et al. Improved PSPNet-based Water Shoreline Detection in Complex Inland River Scenarios[J]. Complex and Intelligent Systems,2023,9(1):233-245

[13] Shi L, Wang L, Long C, et al. SGCN: Sparse Graph Convolution Network for Pedestrian Trajectory Prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021:8 994-9 003

[14] 周家厚,普运伟,陈如俊,等. 改进的UNet3+网络高分辨率遥感影像道路提取[J]. 激光杂志,2024,45(2):161-168

[15] Cao H, Wang Y, Chen J, et al. Swin-UNet: UNet-like Pure Transformer for Medical Image Segmentation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland,2022:205-218

[16] Chang J, He X, Li P, et al. Multi-scale Attention Network for Building Extraction from High-resolution Remote Sensing Images[J]. Sensors,2024,24(3):1 010

[17] Wang Y, Zhao Q, Wu Y, et al. SCA-Net: Multi-scale Contextual Information Network for Building Extraction Based on High-resolution Remote Sensing Images[J]. Remote Sensing, 2023,15(18):4 466