

火星表面石块的卷积自注意力网络识别方法

曹学欢^{1,2}, 彭嫚², 万文辉², 王彪^{2,3}, 王晔昕², 邸凯昌², 李禄¹

1. 北京信息科技大学 自动化学院, 北京 100192;

2. 中国科学院空天信息创新研究院 遥感与数字地球全国重点实验室, 北京 100101;

3. 中国科学院大学, 北京 100049

摘要: 火星表面广泛分布的石块对于火星探测车的安全行驶构成潜在威胁, 同时石块的分布特性也为研究火星着陆区域的地质演化提供了重要线索。然而, 火星探测车图像中的石块识别面临多重挑战: 石块和背景边缘模糊导致轮廓提取困难, 表面纹理特征相似性引发误判, 并且现有的火星石块真实数据集稀缺制约模型训练。为了实现火星探测车图像中石块的精确识别, 本文提出一种基于卷积自注意力网络的石块自动识别模型, 实现图像的像素级分割。该模型采取编码器—解码器架构, 其中编码器基于卷积神经网络提取图像特征, 并嵌入改进的自注意力模块以增强模型对上下文信息的感知能力; 解码器则负责将编码器提取的特征映射回图像空间, 实现精准分割。为了验证模型性能, 本文对“祝融号”火星车影像进行标注, 构建了天问数据集, 结合模拟石块数据集Synmars、Simmars6k、“好奇号”影像数据集MarsData-v2等多个数据集对模型的性能进行测试和验证。此外, 本文还将该模型与DeepLabv3+、Unet++、Segformer、Marsnet等多种方法进行精度比较。采用平均像素精度、召回率、交并比等指标进行评价, 结果表明该模型能够准确识别石块, 其在模拟数据集上精度和召回率均达到90%以上, 在真实数据集上的精度和召回率均最优。

关键词: 火星, 石块提取, 卷积神经网络, Transformer, 特征提取

中图分类号: P2

引用格式: 曹学欢, 彭嫚, 万文辉, 王彪, 王晔昕, 邸凯昌, 李禄. 2025. 火星表面石块的卷积自注意力网络识别方法. 遥感学报, 29(10): 3020-3033

Cao X H, Peng M, Wan W H, Wang B, Wang Y X, Di K C and Li L. 2025. Convolutional self-attention network for automated rock detection in complex martian terrain. National Remote Sensing Bulletin, 29(10): 3020-3033 [DOI:10.11834/jrs.20255015]

1 引言

火星表面分布着丰富的地形地貌, 如撞击坑、沙地、丘陵、干涸河床等(苟盛等, 2021; 谢涓等, 2021), 同时还存在大量不同类型的石块, 部分石块分布密集, 因此火星表面石块自动识别方法对火星巡视器的安全行驶和探测任务的完成具有重要作用(刘召芹等, 2014)。此外, 火星石块的大小、形状、风化程度和空间分布等, 提供了行星表面地质历史和演化的重要信息。

基于探测车影像的火星石块自动识别方法可以分为3类: 基于非监督方法、基于机器学习的方法

法和基于深度学习的方法。基于非监督的识别方法包括两种: 一种是利用像素颜色与空间位置信息的相似性进行归类合并, 如通过设置阈值进行识别的阈值法(韩思奇和王蕾, 2002)、根据图像颜色、纹理结构的突变进行提取的边缘检测法(段瑞玲等, 2005)、根据像素的特征进行分类的区域法(陈方昕, 2008)等; 另一种是统计每个像素的位置关系再进行几何计算分割图像(Rozenberszki等, 2024)。这些方法能够实现对石块的简单分割, 但此类方法对影像的噪声和光照条件较敏感。

基于机器学习的方法主要利用图像的颜色、

收稿日期: 2025-01-16; 预印本: 2025-03-05

基金项目: 中国科学院战略性先导科技专项(B类)(编号:XDB41000000); 国家自然科学基金(编号:62471049); 遥感科学国家重点实验室开放基金(编号:OFSLRSS202317)

第一作者简介: 曹学欢, 研究方向为行星制图与遥感。E-mail: 2023020486@bistu.edu.cn

通信作者简介: 彭嫚, 研究方向为行星制图与遥感。E-mail: pengman@aircas.ac.cn

灰度、纹理等特征进行特征提取，并通过迭代优化和训练提高模型的准确性。代表性方法包括支持向量机（丁世飞等，2011）、决策树（方匡南等，2011）、K-Means（Jain，2008）和条件随机场（钱生等，2015；郑玮等，2008）等，但机器学习的方法提取的特征信息有限，难以实现复杂图像的准确分割。

深度学习技术通过增加网络的深度来提取更丰富的语义信息，并结合数据预处理和模型优化等提高图像分割的精度。基于卷积神经网络CNN（Convolutional Neural Network）的方法在行星遥感图像分割方面取得了显著进展（Ogohara和Gichu，2022；Jia等，2023；Yang等，2020；闵令通等，2023；宋宝贵等，2025）。然而，与地球上的图像相比，火星表面的石块具有不同的形状和相似的颜色及纹理特征。因此，为了实现精确的石块分割，获取多尺度特征并实现图像的语义融合，研究者们引入了特征金字塔方法。这种方法扩大了感受野，并整合了多尺度的上下文信息。U-Net（Ronneberger等，2015）是一种在医学图像分割和分析中广泛应用的多尺度语义分割网络，它在解码器中通过上采样将特征图扩展到与原始图像相同的尺寸，实现多尺度语义融合。一些研究工作者（Lv等，2023b；龙丽红等，2023）通过改进U-Net网络，提出了一种多模态图像变化检测方案。NU-Net（Chen等，2023）采取了一种嵌套式U-Net，对边界模糊且特征相似的目标分割效果较好。其他多尺度语义分割网络，如FCN（Shelhamer等，2017）、PSPNet（Zhao等，2017）和DeepLabV3+（Chen等，2018）也被应用于行星石块探测。然而，这些方法难以从CNN获得的局部特征中提取全局信息，较难从火星场景中获取全局上下文信息。

Transformer（Vaswani等，2017）最初是为自然语言处理（NLP）中的序列到序列建模而开发的，在图像处理中也展现了优越的性能，特别是在场景分割任务中（Xie等，2021；Liu等，2021）。MarsFormer（Xiong等，2023）基于Transformer的架构，通过集成特征增强模块和波束变换模块，构建了编码器—解码器结构，并在虚拟和真实火星数据集上评估了其性能。然而，基于Transformer的方法在处理低层次局部特征时存在局限性，对小石块的分割效果不佳。

目前，学者们通过结合卷积神经网络（CNN）和Transformer设计网络结构，充分利用两者的优点提高石块分割的性能。TransUNet（Chen等，2021）利用基于Transformer的编码器进行鲁棒的图像特征提取，并结合传统的CNN和跳跃连接来实现精确的特征图上采样。然而，TransUNet在图像特定的特征提取方面的能力有限，比如位置和通道。Swin-Unet（Cao等，2021）将Swin Transformer块与U-Net结构相结合，取得了良好的效果。此外，相关研究（Li等，2021；Lv等，2023a）通过结合基于VGG（Simonyan和Zisserman，2015）网络和Transformer的结构，分别提取细节信息和上下文信息，并利用跳跃连接实现特征图上采样和多种信息的融合，完成了精确的分割。EDR-TransUNet（Jia等，2024）整合了增强的双重关系模块、异构核卷积模块和TransUNet混合框架。通过集成全局和局部信息、空间信息和通道信息，并采用Transformer和EDR-Block提高了火星石块图像分割的性能。然而增加大量的Transformer块会导致参数数量增加，但并未显著提高精度。

尽管现有深度学习方法能有效识别火星表面石块，但石块自动识别中仍然存在一些挑战。首先，由于风化原因和地质环境变化，探测车影像中的石块与周围背景的边界不清晰，影响石块识别的准确性；其次，部分区域石块密集分布，彼此间距较小，且这些区域的纹理特征复杂，增加了特征提取的困难。针对以上问题，本研究提出了一种基于卷积自注意力网络的石块自动识别模型。该模型采用VGG作为主干网络，并在编码器网络中加入改进后的Transformer层，然后使用U-Net网络作为解码器，实现石块的准确识别。

2 卷积自注意力网络模型与评价方法

2.1 网络结构

本模型设计了一种混合结构，如图1所示。该结构主要采用编码器解码器架构（Shelhamer等，2017），并结合VGG、卷积自注意力模块及Unet设计而成。VGG网络首先通过一系列卷积层、池化层和激活函数逐步压缩特征图的空间维度，同时增加网络的深度，依次提取出5个不同层次的特征图。随后，网络引入卷积自注意力模块，将最深层的特征图输入到卷积自注意力模块并进行计算。

该模块充分利用区域之间的相关性,通过窗口滑动的方式将模型的注意力逐步扩散至全图,从而提高模型的全局感知能力。在这个模块中,特征图首先经过卷积操作,然后将结果转换为一维数据以执行多头自注意力计算。计算完成后,所得结果再次经过卷积操作,最终输出结果。此后,

卷积自注意力模块的输出经过上采样以恢复特征图的原始尺寸,并与前一层的特征图进行跳跃连接和特征融合。这一上采样、跳跃连接和特征融合的过程不断重复,直至所有层级的特征图被融合,从而构建起U型结构的网络(Ronneberger等, 2015),实现图像的解码和分割任务。

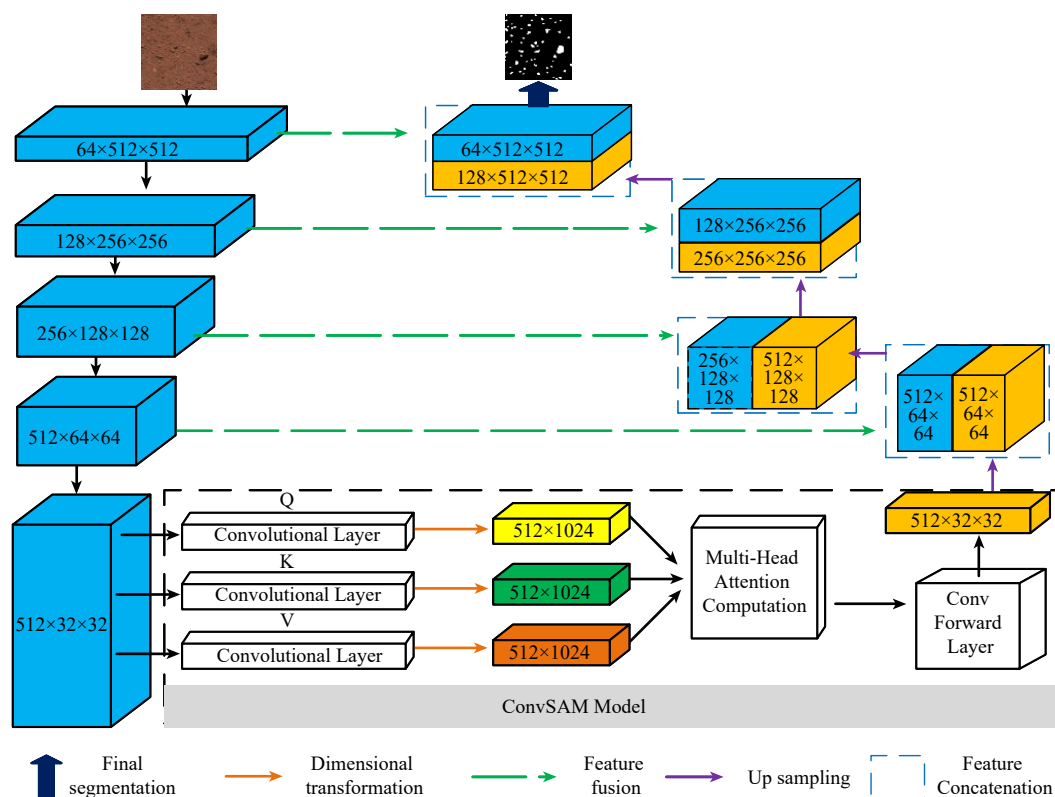


图1 卷积自注意力网络模型结构(维度变换(Dimensional transformation)将多维矩阵 $[512, 32]$ 扁平化为 $[512, 32 \times 32]$ 。特征融合(Feature fusion)对不同深度的特征层进行融合。特征拼接(Feature Concatenation)将两个矩阵按照通道维度进行拼接。上采样(Up sampling)包含卷积,主要恢复特征图维度,以及调整通道数。卷积自注意力模块于图2有详细介绍)

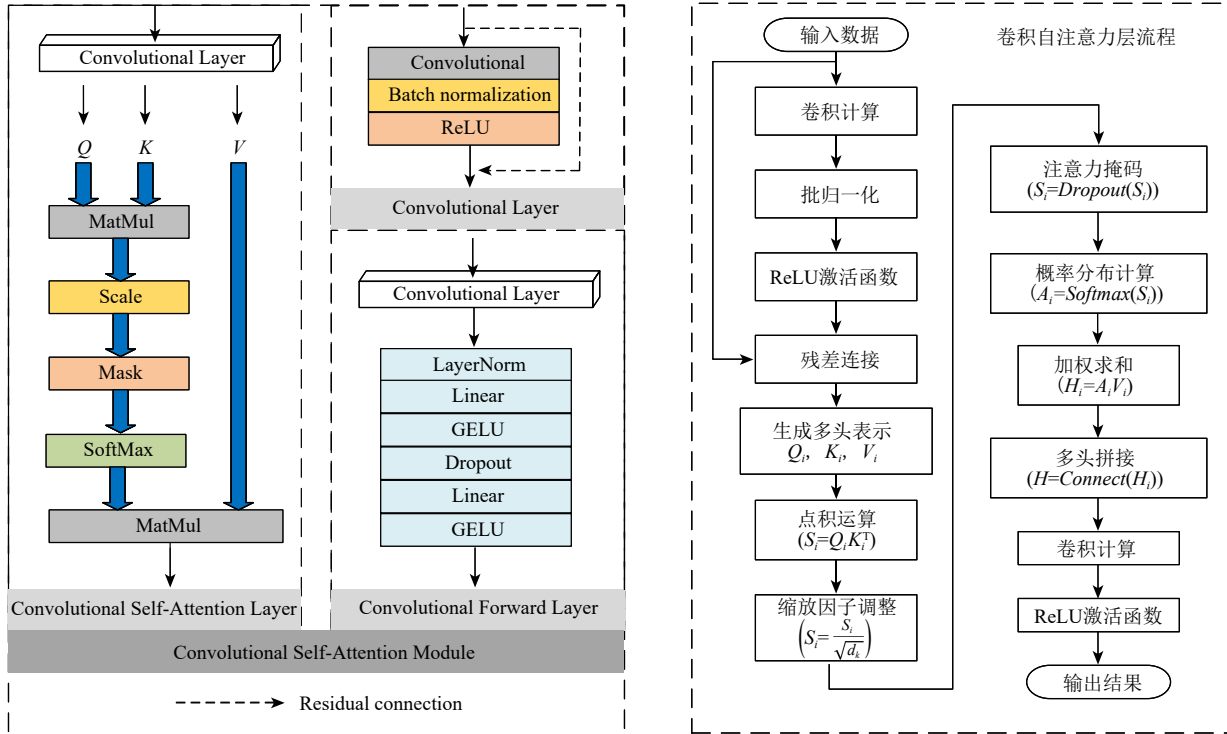
Fig. 1 Structure of convolutional self attention network model (Dimension transformation flattens the multidimensional matrix $[512, 32]$ into $[512, 32 \times 32]$. Feature fusion combines feature layers of different depths. Feature concatenation concatenates two matrices according to the channel dimension. Upsampling includes convolution, which mainly restores the dimensionality of feature maps and adjusts the number of channels. The convolutional self attention module is described in detail in Fig. 2)

2.2 卷积自注意力模块

本文中的卷积自注意力模块由两部分组成:卷积自注意力层和卷积前向传播层。整个卷积自注意力模块流程如图2(a)所示。在卷积自注意力层中,特征图首先通过查询、键和值的卷积操作进行处理,这一过程旨在捕捉特征图中的复杂关系。随后,通过批归一化(He等, 2016)和ReLU激活函数(Nair和Hinton, 2010)的处理,进一步增强了模块处理非线性特征的能力。最终,通过卷积前向传播层的计算,得到模块的输出结果。

2.2.1 卷积自注意力层

卷积自注意力层是卷积自注意力模块的核心部分,主要步骤包括卷积层计算与多头自注意力计算。在卷积层计算阶段,主要包括卷积操作、批量归一化和ReLU激活函数,并通过残差连接增强网络的信息流通。多头自注意力计算(Vaswani等, 2017)则通过点积运算、缩放因子调整、注意力掩码的应用以及概率分布计算等步骤,精确获取特征间的相互关系,多头自注意力计算如式(1)~(3)所示,其中单头注意力计算如式(1)所示。



(a) 卷积自注意力模块(Convolutional Self-Attention Module)
(包含:卷积自注意力层(Convolutional Self-Attention Layer)、
卷积前向传播层(Convolutional Forward Layer)以及卷积层
(Convolutional Layer))

(a) The convolutional self attention module (Includes convolutional
self attention layer, convolutional forward propagation layer, and
convolutional layer)

(b) Provide a flowchart and some calculation formulas for
the convolutional self attention layer

图2 卷积自注意力模块

Fig. 2 Convolutional self-attention module

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

式中, Q , K , V 在卷积自注意力计算中通过卷积层计算后获得, 查询(Query)、键(Key)、值(Value)矩阵是自注意力计算的核心。 QK^T 为点积计算, 即为矩阵乘法, K^T 为 K 的转置矩阵。 $\sqrt{d_k}$ 为缩放因子, 防止点积值过大导致梯度消失, 其中 $\sqrt{d_k} = \frac{d_{\text{model}}}{h}$, d_{model} 为 K 的维度, h 为多头自注意力计算的头数。 softmax 为概率分布计算。

按照单头自注意力计算后可以获得多头自注意力计算中每个头的注意力权重, 如式(2)所示:

$$H_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

最后将所有头的注意力权重进行拼接, 以及维度变换, 并输出最终结果, 如式(3)所示:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(H_1, \dots, H_h)W^O \quad (3)$$

式中, Concat 为多头拼接, 并将维度恢复。 W^O 为线性变换输出, 在本文中, W^O 为卷积计算。

为了解决图像分割中前景与背景特征界限的不清晰, 一些学者(Yuan等, 2020; Zheng等, 2020; Yu等, 2020)通过增强上下文依赖关系实现更精确的识别。Transformer模型在提取上下文特征方面具有显著优势, 其核心机制是自注意力机制。自注意力计算是将特征图展平为序列, 并计算序列中所有元素之间的相互关系, 但这种机制在处理像素间联系时过于简单。此外, 自注意力计算需要对整个序列进行计算, 当特征图的尺寸较大时, 序列的长度会呈平方级增长, 导致计算资源的大量消耗。而在石块密集分布区域, 图像纹理复杂, 石块间的色彩与纹理特征相似性较强, 需要对这些区域进行更深层次的特征提取。采用卷积神经网络可对石块密集区域进行特征提取, 通过卷积核的滑动窗口分割石块密集区域,

提取出细节特征。小卷积核在提取细节特征方面更有效,但其感受野较小,因而在全局信息处理方面存在局限(易琰等,2023;Morid等,2021)。如果仅依赖卷积操作,可能导致提取的特征过于多样化和复杂,同时缺乏全局上下文的联系,难以准确识别分割区域中的目标。

为了解决上述问题,本文融合卷积与自注意力机制,设计了卷积自注意力层,如图2(b)卷积自注意力层流程图所示。该层首先使用卷积操作对细节特征进行预处理,本研究将自注意力模块中的查询、键和值的线性变换层改为卷积层。卷积层能够通过滑动窗口分割特征图,获取区域的独立信息,从而减少纹理特征的复杂性。此外,将特征图中的各个元素与相邻元素进行卷积计算,增强了局部信息感知能力。本研究在卷积层中加入了批归一化和残差连接,有助于减少梯度爆炸或梯度消失的风险,确保特征信息的有效传递。多头自注意力计算首先通过点积计算其中两个特征层,获得不同区域特征之间的相关性;为了控制计算规模,防止梯度消失或爆炸,随后对点积后的相关性结果进行缩放,以此帮助稳定训练过程;接着,采用注意力掩码的方式,忽略不重要的信息,提升模型的泛化能力;然后计算概率分布,将注意力分数转换为概率,以便模型精准的聚焦于重要信息上;最后,将概率分布和最后一层特征层进行点积运算,获得最终的区域相关性结果。由于采用多头自注意力计算,因此需要将计算后的结果进行拼接,并采用卷积与激活函数对结果进行通道转换,以便于后续对特征图进行解码。此方法结合了卷积与自注意力机制的优势,首先通过小卷积核的卷积计算,实现细节特征的提取,然后利用自注意力机制将相邻区域的细节特征关联起来,以捕获局部区域的上下文联系,最后通过卷积进一步整合区域特征并建模全局依赖关系,最终实现从局部细节到全局语义信息的渐进式特征提取,有效增强了模型的上下文感知能力。由于卷积的参数主要取决于卷积核的大小,因此卷积核大小固定的情况下,即使特征图尺寸增加,卷积自注意力模块的参数量仍保持不变。这种机制解决了Transformer中上下文联系性单一、特征图过大而导致参数数量激增的问题,以及卷积在处理区域间联系时,上下文特征提取不足的问题。

2.2.2 卷积前向传播层

卷积前向传播层对卷积自注意力层的输出数据进行深入特征提取和非线性学习,其核心步骤包括卷积层计算、线性变换、GELU激活函数(Hendrycks和Gimpel,2023)、随机失活等,以增强对图像特征的深层次提取和表示,从而提高网络的识别效率。当卷积自注意力层完成对区域特征的相关性计算后,模型获取了特征图中各个区域特征之间的相关性,随后将相关性结果输入到前向传播层中。在前向传播层中,卷积操作逐层提取边缘、纹理、语义信息等局部特征,这一模式实现了特征的渐进式提取,进一步强化了各个区域特征间的联系,形成特征联系的结果。随后,线性变换操作整合所有特征结果,以获得最终的特征信息。同时GELU激活函数引入了非线性变换,使模型能够拟合复杂特征关系,增强了特征的表达能力。随机失活(Dropout)机制通过随机屏蔽部分神经元,实现了有效的正则化,防止模型过度依赖某些特点特征,从而提高模型的泛化能力。这一过程协同了卷积、线性变换、激活函数和随机失活操作,使模型能够通过局部特征的渐进式提取、非线性映射、正则化处理,逐步挖掘输入数据的深层语义信息,不仅加强了特征之间的联系,而且提升了模型对输入数据深层特征的捕捉和理解能力。

2.3 损失函数

交叉熵损失函数(Mao等,2023)是语义分割中广泛使用的损失函数,用以衡量模型预测输出与真实标签之间的差异。通过计算寻找最小化损失,模型的输出概率分布更接近于真实标签,从而提高模型的准确性和鲁棒性。然而,当数据集中存在类别不平衡现象,即不同类别的像素的分布不均匀时,交叉熵损失函数可能会偏向像素数量较多的类别,从而影响模型的整体表现。在火星表面石块识别中,探测车影像的石块像素总数通常远低于背景像素总数,本研究对所有标签中石块像素进行了统计,经计算平均每张图像中,石块像素仅占5%。

为了解决类别不平衡问题,本研究引入了焦点损失函数(Lin等,2020)。焦点损失函数是在交叉熵损失函数的基础上引入了一个调节因子。该因子能降低那些模型预测概率接近0或者1的样

本的损失权重，同时增加模型预测概率接近0.5的样本的损失权重，通过这种方式有效解决样本不平衡的问题。

Dice 系数损失函数 (Sudre 等, 2017) 通过计算模型输出与真实标签之间的交集与并集的比例来评估它们的相似度。这种计算方法不仅促进了模型的训练过程，而且直接优化了分割效果的评估指标。因此，Dice 系数损失函数对处理前景与背景像素不平衡的问题具有独特的优势。本文采取将焦点损失函数与 Dice 系数损失函数相结合的方式，如式 (4)–(6) 所示。

$$\begin{cases} L_{\text{focal_loss}} = (1 - p_t)^\gamma \cdot \log(p_t) \\ p_t = p, \gamma = 1 \\ p_t = 1 - p, \text{其他} \end{cases} \quad (4)$$

$$L_{\text{dic_loss}} = 1 - \frac{2 \times |X \cap Y|}{|X| + |Y|} \quad (5)$$

$$L_{\text{loss}} = L_{\text{focal_loss}} + L_{\text{dic_loss}} \quad (6)$$

式中， P_t 表示模型预测为正类别（石块）的概率。 γ 为平衡系数，用于调整正类别和负类别（背景）的平衡，其范围一般在 $[0, 5]$ ，为了确定 γ 的值，本研究对 γ 设置不同值并进行实验，其结果主要对比 F1 Score (F1 Score 将在 3.2 节评价指标和实现细节中提到) 的大小来确定，实验结果如图 3 所示，当 γ 为 2.3 时效果最佳，因此将 γ 设置为 2.3。

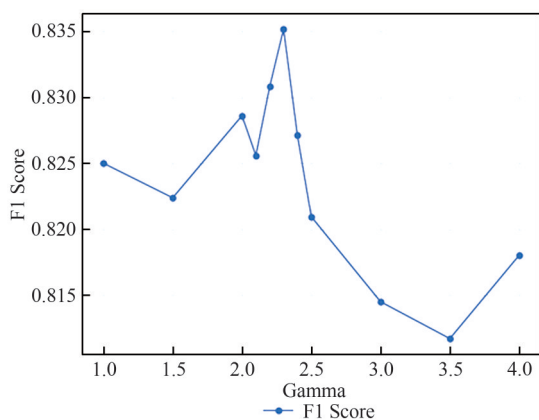


图3 不同Gamma值对比实验

Fig. 3 Comparison experiment of different Gamma values

式 (5) 中， X 表示真实分割图像的像素标签， Y 表示模型预测分割图像的像素类别， $X \cap Y$ 是 X 和 Y 之间的交集， $|X|$ 和 $|Y|$ 分别表示 X 和 Y 的元素的个数。为了确保分母在重复计算后取值范围保持 $[0, 1]$ 之间，将分子乘以 2。若如果分子不乘 2，分母中的 $|X| + |Y|$ 会比分子多出一个交集部

分，导致计算不准确。

3 实验结果与分析

3.1 数据集

本研究使用两个模拟数据集 Simmars6k (Ma 等, 2024) 和 Synmars (Xiong 等, 2023)，以及两个火星车影像真实数据集评价模型的性能。模拟数据集不仅模拟了火星地表还包括天空等背景，提供了更精确的标签能够充分测试模型的性能。并且两个模拟数据集生成了随机分布的不同形状的大石块与小石块，可以测试模型对不同大小的目标的识别能力。各个数据集具体信息如下：

(1) Simmars6k 数据集包含 6300 对图像与标签，图像中模拟了不同太阳高度角的阴影分布和光照强度，能够充分模拟火星环境，从而评估模型的真实性能。

(2) Synmars 数据集包含 150 张 1920×1080 像素图像与标签，经裁剪后扩展到 300 张图像。与 Simmars6k 相比，Synmars 图像环境更为单一，纹理特征更明显，主要以小石块为主。对模型的细节特征提取能力要求更高。

(3) MarsDataV2 数据集 (Liu 等, 2023)。这个数据集包括了“好奇号”火星车使用 Mastcam 相机从 2012 年 8 月至 2018 年 11 月获取的图像，共 8390 张 512×512 像素的带注释图像。在实验中，将 RGB 标签通过线性像素映射转换为灰度标签，并将其分为 5870 张训练图像，1680 张验证图像和 840 张测试图像。

(4) 天问数据集。本研究对“祝融号”火星车导航相机拍摄的 664 张图像进行筛选与裁剪，图像大小为 2048×2048 像素，最终得到了 753 张影像，并通过数据增广扩充到 8000 张。数据增广主要采用翻转、随机裁剪、缩放、平移、旋转、随机亮度、添加噪声等方法，随后采取随机选择两项进行组合的方式来进一步补充数据。将其中的 5600 张用于模型训练，1600 张用于验证，800 张用于测试。所有图像大小均为 512×512 像素，并标注为石块和背景两个类别。每张标签影像中所有标注单独计算，将标注为石块的像素值设置为 1，并统计得到单个石块的像素总数。为了与模拟数据集进行对比，本研究以 100 像素为界限，将所有大于这一数值的石块标注出来。为了便于后续实

验分析, 本文将标注的石块分为大小岩石两种类别。为了更加科学定义石块大小, 相关学者 Shen 等 (2025) 以岩石的最大直径 (50 像素、150 像素、250 像素) 作为划分标准, 而 Lv 等 (2023a) 虽未明确区分大小石块, 但其仅对可能会造成探测车行驶障碍的石块 (大于或等于 30×30 像素) 进行了标注。鉴于以上研究, 结合保证探测的行驶安全, 因此本研究将石块面积大于或等于 900 像素的定义为大石块, 而小于 900 像素的定义为小石块。通过这种方法, 本文构建了包含大石块与小石块的天问数据集。

3.2 评价指标和实现细节

本研究采用多项指标全面评估模型的性能, 包括像素精度 PA (Pixel Accuracy)、平均交并比 MIoU (Mean Intersection over Union)、召回率 (Recall) 和 F1 score。其中, PA 表示分割结果中正确分类的像素占整个图像像素总数的整体比例。平均交并比衡量模型的输出和真实标签之间的重合程度, 是评估分割效果的另一个关键指标。Recall 表示石块分割结果中, 被正确预测的像素所占石块总像素的比例, 反映了模型对石块的识别能力。F1 分数是精确率和召回率的调和平均值, 用以综合评估模型的精确性和召回率。通过计算这些像素级别的指标, 可以得到 PA、Recall、MIoU 和 F1 Score, 它们的定义如式 (7)–(10) 所示:

$$PA = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$MIoU = \frac{1}{2} \left(\frac{TP}{TP + FP + FN} + \frac{TN}{TN + FN + FP} \right) \quad (9)$$

$$F1_Score = \frac{2 \times PA \times Recall}{PA + Recall} \quad (10)$$

式中, TP (真正例) 指的是模型正确预测为正例的样本。FP (假正例) 是指模型错误预测为正例的样本, FN (假反例) 是指模型错误预测为反例的正例样本, 而 TN (真反例) 是指模型正确预测为反例的样本。

本实验的操作系统是 Ubuntu 18.04, 使用 NVIDIA RTX 4090 GPU 进行运算。网络模型采用端到端训练, 以确保模型从输入到输出的整个流程能够协同优化。在所有实验中, 训练周期设定

为 100 轮, 模型训练的批量大小设定为 8, 这些参数值主要是基于 GPU 内存容量确定的。训练过程中, 系统会自动保存在验证集上表现最优 (即损失最低) 的两个模型快照, 以及训练结束后的最终模型输出。选用 Adam 优化器进行参数优化, 其初始学习率设定为 1E-5。训练在达到预设的最大迭代次数终止, 或当模型在连续 40 个训练周期内性能未见提升时自动终止。对于实验中使用的数据集, 将所有输入图像统一调整至 512×512 像素的尺寸, 并将图像标签通过线性像素映射转换为灰度标签, 以便网络能够输出具有不同类别值的灰度图像, 从而实现精确的图像分割。

3.3 对比实验分析

为了验证本文所提出模型的性能, 本研究采用多种语义分割模型进行对比分析。这些模型包括经典的卷积网络 DeepLabv3+、U-Net、U-Net++, 以及基于 Transformer 架构的 SETR (Zheng 等, 2021)、Segformer (Xie 等, 2021), 还有融合了卷积与 Transformer 的 SwinUNet 与 Marsnet。为了全面评估和比较这些模型的性能, 本研究采用了多个数据集进行测试。所有参与对比的模型都使用了 ImageNet 数据集进行预训练, 以此获得初始化的权重, 并且所有模型均应用了相同的损失函数。

(1) 真实火星数据集结果与分析。本研究首先在 MarsDataV2 和天问数据集上进行实验, 定量分析结果如表 1 所示, 其中最高分值用粗体标记, 本模型较其他模型更优的结果用下划线表示。结果显示, 对比方法在天问数据集的 PA、Recall、MIoU 和 F1 score 最高值分别为 90.00%、80.14%、70.85% 和 82.98%; 在 Marsdatav2 数据集上, 这些指标的最高值分别为 96.49%、96.41%、91.25% 和 96.45%。本文通过引入卷积自注意力模块的实验结果, 在多数指标都优于其他模型。在天问数据集上, Recall、MIoU 和 F1 score 最高值分别提升至 81.10%、72.18% 和 85.53%。由于天问数据集中属于石块类别的像素数量远少于背景类别, 导致前景和背景分布的显著不平衡。本研究采取的损失函数旨在衡量两者的相似性进而促进模型训练, 使得精度、召回率与交并比较为平衡。本研究的卷积自注意力模块通过层次融合策略有效联系了全局和局部特征, 从而有利于提高分割预测的准

确性。Marsdatav2数据集的图像特征更加明显，各个模型间的差异主要体现在石块与背景之间的边缘分割上，因此本模型在Marsdatav2改进的效果相对较小。在天问数据集的对比中，U-Net++与DeeplabV3+模型在精度上仅比本模型高2.05%和

0.36%，但在召回率上比本模型低16.3%和24.37%。在Marsdatav2数据集的对比中，Marsnet模型在精度上仅比本模型高0.06%，但在召回率上比本模型低0.16%，因此本模型在综合性能上优于其他模型。

表1 各模型在真实数据集实验对比结果
Table 1 Experimental comparison results of various models on real datasets

模型	天问				Marsdatav2			
	PA	Recall	MIoU	F1	PA	Recall	MIoU	F1
DeeplabV3+	88.31	56.73	51.50	69.08	93.98	93.18	84.06	93.58
U-Net	87.59	70.85	63.72	78.34	94.75	91.45	83.41	93.07
U-Net++	90.00	64.80	59.25	75.35	95.70	87.29	80.06	91.30
SETR	67.31	32.27	27.28	43.63	89.61	79.32	65.01	84.15
Segformer	87.94	38.91	37.68	53.95	94.76	84.13	74.33	89.13
SwinUnet	71.44	53.25	42.65	61.02	91.53	84.03	71.44	87.62
Marsnet	86.04	80.14	70.85	82.98	96.49	96.41	91.25	96.45
Ours	87.95	<u>81.10</u>	<u>72.18</u>	<u>85.53</u>	<u>96.65</u>	96.35	<u>91.38</u>	<u>96.50</u>

注:粗体为最高分值,下划线为本文模型较其他模型更优结果。

图4显示了天问数据集和Marsdatav2数据集的实验结果。通过对比结果可知，各模型对于特征明显的石块能够实现较为完整的识别。然而在石块与背景的边缘区域，由于图像特征从石块到背景的过渡缺乏明显界限，导致特征模糊，模型在这些区域的识别能力还有待提高。在真实数据集中，基于卷积神经网络的模型，如Deeplabv3+、U-Net、U-Net++，能够完整识别石块，对小石块也能有效提取，但在处理大石块时仍存在一些空洞。而基于Transformer架构的模型，如SETR、Segformer、Swinunet，在大石块上则没有空洞，图像分割的较为连续，但在识别小石块时存在较多遗漏。Marsnet与本文模型结合了卷积与Transformer的优势，能够全面识别大石块与小石块。对比天问与Marsdatav2数据集的结果，天问数据集中的石块纹理较为模糊，导致识别结果不如Marsdatav2，但本文模型在精度和召回率上仍能保持在80%以上。此外，对比分析结果表明，卷积网络在细节目标提取方面拥有更强的能力，而Transformer在提取上下文特征方面表现更为出色。本研究的模型结合了这两种方法的优势，因此石块识别结果更为均衡和有效。

(2) 模拟数据集结果与分析。本模型在模拟数据集的结果优于其他方法，具体结果如表2所

示。在Synmars数据集上，本模型的MIoU与F1 score上达到了最高值，分别是90.62%和96.17%。在Simmars6k数据集上，本模型在PA、MioU和F1 score上达到了最高值，分别为94.92%、88.98%、94.67%。由于Synmars与Simmars6k数据纹理特征存在差异，Synmars纹理特征更加明显，因此U-Net、U-Net++等基于卷积的模型在细节提取上表现更佳，并且对纹理明显的图像的上下文信息需求更少，故各类型模型对Synmars数据的分割精度普遍较高。本模型通过引入卷积自注意力模块，增强了上下文联系，在Synmars数据集上虽然精度上略有下降，但召回率得到了提升。在Simmars6k数据集中，尽管U-Net++在召回率上表现最佳，但其精度相对较低。本文模型的PA、MioU和F1 score均最高。与真实数据集相比，各个模型在模拟数据集上的主要差异体现在对小石块的处理。由于小石块在图像上的总像素数较低，因此评价指标的差异较少，在图5的结果对比中更明显。

如图5所示，实验结果显示了各模型在模拟数据集提取小石块的差异。由于小石块在图像中占比相对较低，导致表2中各指标之间的差异不大。尽管各个模型均能较完整识别出较大石块，但在识别小石块时表现各异。基于卷积的网络模型，如Deeplabv3+、Unet、Unet++对小石块识别的较为

准确，但未能识别出小石块的情况较多。而基于Transformer架构的模型如Segformer和Swinunet在识别小石块时遗漏较少，但模型的误识别较多，尤其是在石块密集、纹理比较复杂的区域。卷积神经网络虽然能够提取细节信息，但在石块密集区域缺乏上下文联系，导致许多小石块未能识别。

相反，Transformer能够提取整个区域的上下文信息，但在密集区域由于纹理复杂，缺少细节特征，导致图像中小石块区域出现了大面积的误识别。本研究引入的卷积自注意力模块充分结合了卷积与Transformer的优势，在小石块的识别性能上取得了显著提升。

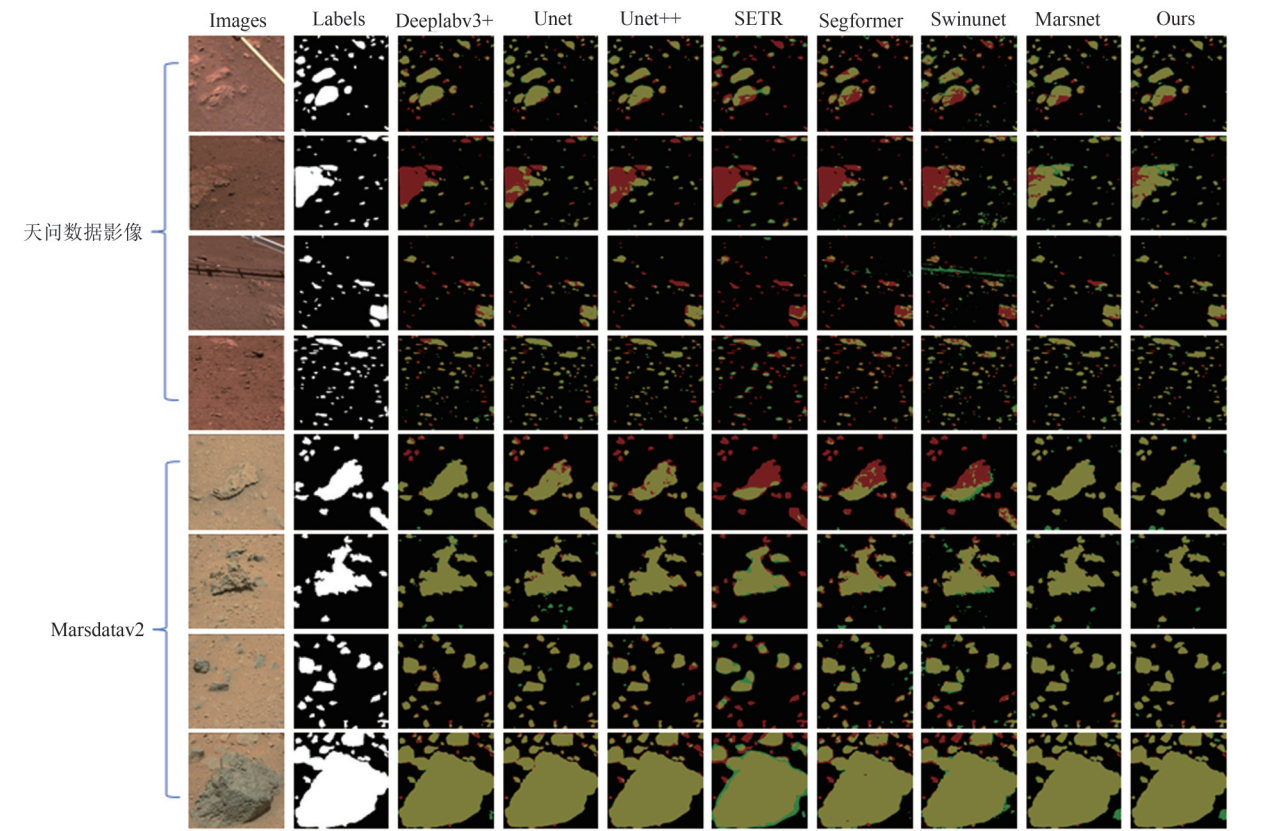


图4 各模型在真实数据集实验结果(其中红色部分为FN,黄色部分为TP,绿色部分为FP)

Fig. 4 Experimental results of each model on a real dataset (The red part represents FN, the yellow part represents TP, and the green part represents FP)

表2 各模型在模拟数据集实验对比结果

Table 2 Comparison results of various models in simulated datasets

模型	Synmars				Simmars6k				/%
	PA	Recall	MIoU	F1	PA	Recall	MIoU	F1	
DeeplabV3+	92.44	92.35	81.76	92.39	91.02	89.42	82.72	90.21	
U-Net	96.91	93.22	88.00	95.03	91.24	92.90	87.71	92.06	
U-Net++	96.86	94.76	89.27	95.80	85.90	94.61	87.94	90.05	
SETR	79.12	67.04	52.35	72.58	60.76	62.29	43.56	61.51	
Segformer	89.68	87.02	74.10	88.33	76.36	87.50	70.68	81.55	
SwinUnet	87.30	87.43	76.19	87.36	78.13	86.26	71.44	82.00	
Marsnet	95.86	96.24	90.45	96.05	94.79	93.35	88.92	94.07	
Ours	96.18	96.16	<u>90.62</u>	<u>96.17</u>	<u>94.92</u>	94.48	<u>88.98</u>	<u>94.67</u>	

注:粗体为最高分值,下划线为本文模型较其他模型更优结果。

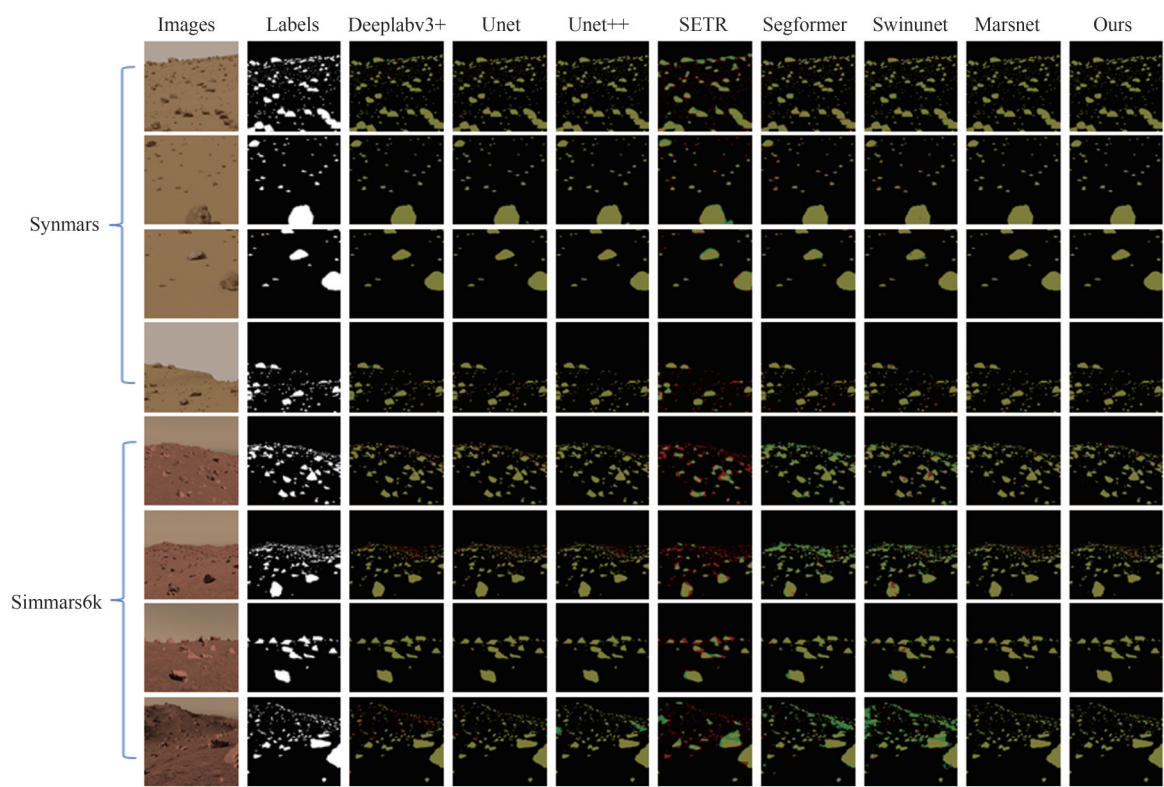


图5 各模型在模拟数据集实验结果(其中红色部分为FN,黄色部分为TP,绿色部分为FP)

Fig. 5 Experimental results of each model on simulated dataset (The red part represents FN, the yellow part represents TP, and the green part represents FP)

3.4 消融实验分析

本研究利用天问数据集进行消融实验，深入分析卷积自注意力模块对模型性能的影响。消融实验主要讨论两个问题：一是评估不同数量的卷积自注意力模块对模型的影响，以及这些模块在不同特征层级上的集成对模型性能的影响；二是比较卷积自注意力层、卷积前向传播层、传统Transformer层以及传统前向传播层对模型性能的具体影响。

如表3所示，该表展示了在最后一层特征层应用不同数量的卷积自注意力模块对模型性能的影响，以及这些模块能否增加网络的深度。同时也分析了模块在不同特征层的应用对模型性能的优化作用，其中最优结果以粗体显示。通过对比1到5层卷积自注意力模块的结果，本研究发现增加模块数量并未显著提升模型性能，反而导致精度下降，召回率也表现出不稳定的波动。说明仅集成一层卷积自注意力模块就能达到最佳效果。随着模块数量的增加，模型因为计算复杂度上升遭遇梯度消失问题，导致性能下降。此外，添加更大尺度的特征层也未能提升模型性能，可能是因为

在这些尺度上的特征图缺乏足够的细节信息，使得自注意力计算后仍然无法捕捉到足够的特征间相关性。这种计算在反向传播过程中对更深层特征层的影响较大，导致性能进一步下降。实验结果表明，单层卷积自注意力模块已经能够提供最优的特征提取能力，实现最佳结果。

如表4所示，实验对5种不同的网络结构组合进行了性能比较：卷积自注意力模块、单独的卷积自注意力层、卷积自注意力层与传统前向传播层的组合、传统自注意力模块，以及传统自注意力层与卷积前向传播层的组合。与传统的Transformer模块相比，卷积自注意力模块在各项性能指标上都有所提升；当模型仅集成单独的卷积自注意力层时，召回率显著下降；当模型结合卷积自注意力层与传统前向传播层时，虽然各项性能指标有所下降，但降幅相对较小；模型采用传统自注意力层与卷积前向传播层的组合时，各项性能指标同样有所下降。前向传播层极大地增强了模型学习非线性特征的能力，对模型性能有着重要的影响。与传统前向传播层相比，卷积前向传播层与卷积自注意力层的结合更为适宜。经

过卷积自注意力层处理的特征，尽管局部特征得到了强化，仍然缺少足够的上下文信息和非线性特征，因此需要卷积前向传播层对特征进一步处理，获取全局特征，并提高模型的泛化能力。这

种结构的协同作用，使得本文模型能够更有效地捕捉和利用图像中的复杂特征，从而提升分割任务的性能。

表 3 模块不同数量与作用位置消融实验

Table 3 Experiments on ablation of modules with different quantities and positions of action

Conv_SAM32	Conv_SAM64	Conv_SAM128	Conv_SAM256	Conv_SAM512	PA/%	Recall/%	MioU/%	F1/%
1					88.0	81.1	72.2	85.5
2					85.9	78.9	69.5	82.2
3					85.4	81.2	70.8	83.3
4					85.0	79.4	69.6	82.1
5					78.8	76.0	59.9	77.1
1	1				84.4	66.8	62.4	74.6
1	1	1			88.8	73.6	64.6	80.5
1	1	1	1		79.7	78.2	64.9	78.9
1	1	1	1	1	83.0	70.3	61.3	76.1

表 4 不同层组合消融实验

Table 4 Experiments on ablation of different layer combinations

CovSAM_layer	Conv_Foward_layer	SelfAteetion_layer	Foward_layer	PA/%	Recall/%	MioU/%	F1/%
√	√			88.0	81.1	72.2	85.5
√				84.9	45.6	31.8	59.4
√			√	86.0	79.3	70.2	82.5
		√	√	87.1	80.7	70.8	83.8
	√	√		85.6	80.1	70.6	82.8

4 结 论

火星表面石块因陨石撞击、风力侵蚀及尘土覆盖呈现形态多样、边界模糊的特性。石块与背景区分困难，探测影像中石块密集区域的误识别率高，影响了探测车环境感知精度。

针对此问题，本研究提出了 CNN-Transformer 融合模型，并设计卷积自注意力模块（ConvSAM）。该模块通过卷积层提取局部细节特征，同时利用自注意力机制捕获全局上下文关联，再结合 U-net 解码器实现多尺度特征的融合。该不仅模型提升了石块密集区域的分割精度，还增强了边界模糊区域的适应能力，有效解决了传统方法因纹理交错导致的误分割问题。

实验结果表明，模型在 Synmars、SimMars6k 模拟数据集及 MarsdataV2、天问真实影像上均优于现有方法。此外，消融实验进一步验证了卷积自注意力模块在提升模型性能方面的重要作用。

本研究的相关成果可为嵌入探测车实时感知系统提供理论参考与方法支撑，从而提升其自主导航能力和科学探测效能。未来研究将在轻量化上进一步优化模型，满足火星探测任务的实际需求。

参考文献(References)

Cao H, Wang Y Y, Chen J, Jiang D S, Zhang X P, Tian Q and Wang M N. 2021. Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv: 2105.05537 [DOI: 10.48550/arXiv.2105.05537]

Chen F X. 2008. Image segmentation technology based on region growing method. Science and Technology Information, (15): 58-59 (陈方昕. 2008. 基于区域生长法的图像分割技术. 科技信息, (15): 58-59)

Chen G P, Li L, Zhang J X and Dai Y. 2023. Rethinking the unpretentious U-Net for medical ultrasound image segmentation. Pattern Recognition, 142: 109728 [DOI: 10.1016/j.patcog.2023.109728]

Chen J N, Lu Y Y, Yu Q H, Luo X D, Adeli E, Wang Y, Lu L, Yuille A L and Zhou Y Y. 2021. TransUNet: transformers make strong en-

- coders for medical image segmentation. arXiv: 2102.04306 [DOI: 10.48550/arXiv.2102.04306]
- Chen L C, Zhu Y K, Papandreou G, Schroff F and Adam H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv: 1802.02611 [DOI: 10.48550/arXiv.1802.02611]
- Ding S F, Qi B J and Tan H Y. 2011. An overview on theory and algorithm of support vector machines. *Journal of University of Electronic Science and Technology of China*, 40(1): 2-10 (丁世飞, 齐丙娟, 谭红艳. 2011. 支持向量机理论与算法研究综述. *电子科技大学学报*, 40(1): 2-10) [DOI: 10.3969/j.issn.1001-0548.2011.01.001]
- Duan R L, Li Q X and Li Y H. 2005. Summary of image edge detection. *Optical Technique*, 31(3): 415-419 (段瑞玲, 李庆祥, 李玉和. 2005. 图像边缘检测方法研究综述. *光学技术*, 31(3): 415-419) [DOI: 10.3321/j.issn:1002-1582.2005.03.028]
- Fang K N, Wu J B, Zhu J P and Xie B C. 2011. A review of technologies on random forests. *Statistics and Information Forum*, 26(3): 32-38 (方匡南, 吴见彬, 朱建平, 谢邦昌. 2011. 随机森林方法研究综述. *统计与信息论坛*, 26(3): 32-38) [DOI: 10.3969/j.issn.1007-3116.2011.03.006]
- Gou S, Yue Z Y, Di K C and Niu S L. 2021. Rampart craters in the Isidis Planitia, Mars: remote sensing analysis and environment implications. *National Remote Sensing Bulletin*, 25(7): 1374-1384 (苟盛, 岳宗玉, 邸凯昌, 牛胜利. 2021. 火星伊西底斯平原的壁垒撞击坑: 遥感分析及环境启示意义. *遥感学报*, 25(7): 1374-1384) [DOI: 10.11834/jrs.20211052]
- Han S Q and Wang L. 2002. A survey of thresholding methods for image segmentation. *Systems Engineering and Electronics*, 24(6): 91-94, 102 (韩思奇, 王蕾. 2002. 图像分割的阈值法综述. *系统工程与电子技术*, 24(6): 91-94, 102) [DOI: 10.3321/j.issn:1001-506X.2002.06.027]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Hendrycks D and Gimpel K. 2023. Gaussian error linear units (GELUs). arXiv:1606.08415 [DOI: 10.48550/arXiv.1606.08415]
- Jain A K. 2008. Data clustering: 50 years beyond K-means//Daelemans W, Goethals B and Morik K, eds. *Machine Learning and Knowledge Discovery in Databases*. Antwerp: Springer: 3-4 [DOI: 10.1007/978-3-540-87479-9_3]
- Jia Y T, Su Z J, Wan G, Liu L and Liu J. 2023. AE-TransUNet+: an enhanced hybrid transformer network for detection of lunar south small craters in LRO NAC images. *IEEE Geoscience and Remote Sensing Letters*, 20: 6007405 [DOI: 10.1109/LGRS.2023.3294500]
- Jia Y T, Wan G, Li W H, Li C, Liu J, Cong D W and Liu L. 2024. EDR-TransUnet: integrating enhanced dual relation-attention with transformer U-Net for multiscale rock segmentation on mars. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 4601416 [DOI: 10.1109/TGRS.2024.3432533]
- Li H C, Qiu L W, Li Z, Meng B, Huang J B and Zhang Z M. 2021. Automatic rocks segmentation based on deep learning for planetary rover images. *Journal of Aerospace Information Systems*, 18(11): 755-761 [DOI: 10.2514/1.1010925]
- Lin T Y, Goyal P, Girshick R, He K M and Dollár P. 2020. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318-327 [DOI: 10.1109/TPAMI.2018.2858826]
- Liu H Q, Yao M B, Xiao X M and Xiong Y G. 2023. RockFormer: a U-shaped transformer network for Martian rock segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 4600116 [DOI: 10.1109/TGRS.2023.3235525]
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: hierarchical vision transformer using shifted windows. arXiv: 2103.14030 [DOI: 10.48550/arXiv.2103.14030]
- Liu Z Q, Wan W H, Peng M, Zhao Q, Xu B, Liu B, Liu Y L, Di K C, Li L C, Yu T Y, Wang B F, Zhou J L and Chen H M. 2014. Remote sensing mapping and localization techniques for teleoperation of Chang'e-3 rover. *Journal of Remote Sensing (in Chinese)*, 18(5): 971-980 (刘召芹, 王文辉, 彭曼, 赵强, 徐斌, 刘斌, 刘一良, 邸凯昌, 李立春, 于天一, 王保丰, 周建亮, 陈宏敏. 2014. 遥感制图与导航定位技术在嫦娥三号遥操作中的应用. *遥感学报*, 18(5): 971-980) [DOI: 10.11834/jrs.20144072]
- Long L H, Zhu Y T, Yan J W, Liu J J and Wang Z Y. 2023. New building extraction method based on semantic segmentation. *National Remote Sensing Bulletin*, 27(11): 2593-2602 (龙丽红, 朱宇霆, 闫敬文, 刘敬瑾, 王宗跃. 2023. 新型语义分割D-UNet的建筑物提取. *遥感学报*, 27(11): 2593-2602) [DOI: 10.11834/jrs.20211029]
- Lv W K, Wei L H, Zheng D, Liu Y and Wang Y M. 2023a. MarsNet: automated rock segmentation with transformers for Tianwen-1 mission. *IEEE Geoscience and Remote Sensing Letters*, 20: 3506605 [DOI: 10.1109/LGRS.2022.3227338]
- Lv Z Y, Huang H T, Sun W W, Lei T, Benediktsson J A and Li J H. 2023b. Novel enhanced UNet for change detection using multi-modal remote sensing image. *IEEE Geoscience and Remote Sensing Letters*, 20: 2505405 [DOI: 10.1109/LGRS.2023.3325439]
- Ma C H, Li Y, Lv J Y, Xiao Z X, Zhang W M and Mo L S. 2024. Automated rock detection from mars rover image via Y-shaped dual-task network with depth-aware spatial attention mechanism. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 4600418 [DOI: 10.1109/TGRS.2024.3371684]
- Mao A Q, Mohri M and Zhong Y T. 2023. Cross-entropy loss functions: theoretical analysis and applications. arXiv: 2304.07288 [DOI: 10.48550/arXiv.2304.07288]
- Min L T, Fan Z M, Xie X X and Lü Q Y. 2023. Feature reassembly and self-attention for oriented object detection in remote sensing images. *National Remote Sensing Bulletin*, 27(12): 2716-2725 (闵令通, 范子满, 谢星星, 吕勤毅. 2023. 特征重组和自注意力的遥感图像有向目标检测. *遥感学报*, 27(12): 2716-2725) [DOI: 10.11834/jrs.20233456]
- Morid M A, Borjali A and Del Fiol G. 2021. A scoping review of trans-

- fer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine*, 128: 104115 [DOI: 10.1016/j.combiomed.2020.104115]
- Nair V and Hinton G E. 2010. Rectified linear units improve restricted Boltzmann machines//*Proceedings of the 27th International Conference on Machine Learning*. Haifa: Omnipress: 807-814
- Ogohara K and Gichu R. 2022. Automated segmentation of textured dust storms on mars remote sensing images using an encoder-decoder type convolutional neural network. *Computers and Geosciences*, 160: 105043 [DOI: 10.1016/j.cageo.2022.105043]
- Qian S, Chen Z H, Lin M Q and Zhang C B. 2015. Saliency detection based on conditional random field and image segmentation. *Acta Automatica Sinica*, 41(4): 711-724 (钱生, 陈宗海, 林名强, 张陈斌. 2015. 基于条件随机场和图像分割的显著性检测. *自动化学报*, 41(4): 711-724) [DOI: 10.16383/j.aas.2015.c140328]
- Ronneberger O, Fischer P and Brox T. 2015. U-Net: convolutional networks for biomedical image segmentation. *arXiv*: 1505.04597 [DOI: 10.48550/arXiv.1505.04597]
- Rozenberszki D, Litany O and Dai A. 2024. UnScene3D: unsupervised 3D instance segmentation for indoor scenes//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 19957-19967 [DOI: 10.1109/CVPR52733.2024.01886]
- Shelhamer E, Long J and Darrell T. 2017. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4): 640-651 [DOI: 10.1109/TPAMI.2016.2572683]
- Shen Y, Pan D, Cao H T, Yuan B F, Jia Y, He L B and Zou M. 2025. Terrain classification and rock abundance analysis at Utopia Planitia using Zhurong image data based on deep learning algorithms. *Journal of Terramechanics*, 117: 101022 [DOI: 10.1016/j.jterra.2024.101022]
- Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition. *arXiv*: 1409.1556 [DOI: 10.48550/arXiv.1409.1556]
- Song B G, Shao P, Shao W, Zhang X D and Dong T. 2025. A two-branch building extraction network integrating main body and edge separation and multi-scale information extraction. *National Remote Sensing Bulletin*, 29(8): 1-13 (宋宝贵, 邵攀, 邵文, 张晓东, 董婷. 2025. 集成主体边缘分离和多尺度信息提取的双分支建筑物提取网络. *遥感学报*, 29(8): 1-13) [DOI: 10.11834/jrs.20253549]
- Sudre C H, Li W Q, Vercauteren T, Ourselin S and Cardoso M J. 2017. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations//*Proceedings of the 3rd International Workshop, DLMIA 2017, and 7th International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Québec City: Springer: 240-248 [DOI: 10.1007/978-3-319-67558-9_28]
- Vaswani A, Shazeer A, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc.: 6000-6010
- Xie E Z, Wang W H, Yu Z D, Anandkumar A, Alvarez J M and Luo P. 2021. SegFormer: simple and efficient design for semantic segmentation with transformers//*Proceedings of the 35th International Conference on Neural Information Processing Systems*. Virtual Event: Curran Associates Inc.: 12077-12090
- Xie J, Yan K, Kang Z Z, Xu X J, Xue B, Yang J F and Tao J Y. 2021. Verification study for the mineral and rock identification using multispectral camera of the “Zhurong” Mars Rover on the earth. *National Remote Sensing Bulletin*, 25(7): 1385-1399 (谢涓, 闫凯, 康志忠, 徐箫剑, 薛彬, 杨建峰, 陶金有. 2021. “祝融号”火星车多光谱相机岩矿类型识别的地面验证研究. *遥感学报*, 25(7): 1385-1399) [DOI: 10.11834/jrs.20211065]
- Xiong Y G, Xiao X M, Yao M B, Liu H Q, Yang H and Fu Y G. 2023. MarsFormer: Martian rock semantic segmentation with transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 4600612 [DOI: 10.1109/TGRS.2023.3302649]
- Yang C, Zhao H S, Bruzzone L, Benediktsson J A, Liang Y C, Liu B, Zeng X G, Guan R C, Li C L and Ouyang Z Y. 2020. Lunar impact crater identification and age estimation with Chang’E data by deep and transfer learning. *Nature Communications*, 11(1): 6358 [DOI: 10.1038/s41467-020-20215-y]
- Yi Q, Zhang Y H, Zong Y T and Dai Y B. 2023. A survey of hyperspectral image classification algorithms based on convolutional neural networks. *Electronics Optics and Control*, 30(3): 70-77 (易琰, 张宇航, 宗艳桃, 戴颜斌. 2023. 基于卷积神经网络的高光谱图像分类算法综述. *电光与控制*, 30(3): 70-77) [DOI: 10.3969/j.issn.1671-637X.2023.03.013]
- Yu C Q, Wang J B, Gao C X, Yu G, Shen C H and Sang N. 2020. Context prior for scene segmentation//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle: IEEE: 12413-12422 [DOI: 10.1109/CVPR42600.2020.01243]
- Yuan Y H, Chen X L and Wang J D. 2020. Object-contextual representations for semantic segmentation//*Vedaldi A, Bischof H, Brox T and Frahm J M, eds. 16th European Conference Computer Vision-ECCV 2020*. Glasgow: Springer: 173-190 [DOI: 10.1007/978-3-030-58539-6_11]
- Zhao H S, Shi J P, Qi X J, Wang X G and Jia J Y. 2017. Pyramid scene parsing network. *arXiv*: 1612.01105 [DOI: 10.48550/arXiv.1612.01105]
- Zheng S X, Lu J C, Zhao H S, Zhu X T, Luo Z K, Wang Y B, Fu Y W, Feng J F, Xiang T, Torr P H S and Zhang L. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville: IEEE: 6877-6886 [DOI: 10.1109/CVPR46437.2021.00681]
- Zheng W, Kang G W, Chen W F and Li X W. 2008. Unsupervised segmentation of remote sensing images based on fuzzy Markov random field model. *Journal of Remote Sensing (in Chinese)*, 12(2): 246-252 (郑玮, 康戈文, 陈武凡, 李小文. 2008. 基于模糊马尔可夫随机场模型的遥感图像无监督分割. *遥感学报*, 12(2): 246-252)

夫随机场的无监督遥感图像分割算法. 遥感学报, 12(2): 246-252) [DOI: 10.11834/jrs.20080232]
 Zheng Z, Zhong Y F, Wang J J and Ma A L. 2020. Foreground-aware relation network for geospatial object segmentation in high spatial

resolution remote sensing imagery//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE: 4095-4104 [DOI: 10.1109/CVPR42600.2020.00415]

Convolutional self-attention network for automated rock detection in complex martian terrain

CAO Xuehuan^{1,2}, PENG Man², WAN Wenhui², WANG Biao^{2,3}, WANG Yexin², DI Kaichang², LI Lu¹

1.College of Automation, Beijing Information Science and Technology University, Beijing 100192, China;

2.State Key Laboratory of Remote Sensing and Digital Earth, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China;

3.University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The Martian surface is characterized by a significant presence of rocks, whose size distribution, spatial density, and morphological characteristics are critical factors in determining the safety of landing site selection for Mars missions. These factors also directly influence the path planning and motion control of rovers during exploration activities. Moreover, the spatial distribution of rocks holds substantial value for studying the geological evolution of landing sites. However, images captured by Mars rovers often exhibit blurred boundaries between rocks and the background, as well as ambiguous textural features of the rocks themselves. These issues complicate the task of distinguishing rocks from the surrounding terrain. The scarcity of real Martian rock datasets further exacerbates these challenges, making it increasingly difficult to achieve automatic and accurate rock identification on the Martian surface. To address these challenges and achieve precise rock identification in Mars rover images, this study proposes a novel model for automatic segmentation of Martian surface rocks based on a convolutional self-attention network. The model implements pixel-level segmentation of images, adopting an encoder-decoder network architecture. The encoder utilizes a convolutional neural network to extract image features. Meanwhile, a convolutional self-attention module is incorporated to enhance the model's ability to understand contextual information within the images and improve its rock detection performance. This module enables the model to focus effectively on important spatial features across the images, thereby capturing dependencies and contextual relationships between different regions. Finally, the model employs skip connections to facilitate feature fusion, forming a U-shaped decoder network that produces pixel-wise classification outputs, enabling precise segmentation of the images. In this paper, we present the Tianwen Mars Surface Image Dataset, a manually annotated dataset of Martian surface imagery captured by the Zhurong rover. We integrate multiple datasets, including the artificially simulated rock datasets SynMars and SimMars6K, as well as the MarsData-v2 dataset from Curiosity rover images, to conduct rigorous testing and validation of the model's performance. Its performance is compared with that of several other advanced methods, including DeepLabv3+, UNet++, SegFormer, and MarsNet, using evaluation metrics such as average pixel accuracy, recall, and intersection over union. Results demonstrate that our proposed model excels in rock extraction, achieving accuracy and recall rates exceeding 90% on the simulated datasets. On real datasets, the model, with the highest accuracy and recall, outperforms other methods, highlighting its superior performance in rock identification. The experimental comparisons indicate that the model has the potential for accurate and reliable detection of rocks in Mars rover images, which is crucial for autonomous exploration and scientific research on Mars. The proposed convolutional self-attention module effectively combines the strengths of convolutional operations and Transformer architectures. It enhances the model's ability to extract contextual information while retaining the capacity to capture detailed features. This dual capability not only improves the model's segmentation accuracy in areas with dense rock clusters but also enhances its adaptability to regions with blurred boundaries. As a result, the model effectively addresses the challenges associated with accurately identifying surface rocks in the Martian environment, paving the way for reliable and efficient Mars exploration missions.

Key words: Mars, rock extraction, convolutional neural networks, Transformer, feature extraction

Supported by B-type Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB41000000); National Natural Science Foundation of China (No. 62471049); Open Fund of the State Key Laboratory of Remote Sensing Science (No. OFSLRSS202317)