改进的YOLOv8n轻量化火星表面岩石检测算法

戴 娟[1,2,3]，刘经纬[1,2,3]，苏 中[1,2,3]，朱 翠[4]

（1. 北京信息科技大学 高动态导航技术北京市重点实验室，北京 100192；2. 现代测控技术教育部重点实验室，北京 100192；
3. 北京信息科技大学 自动化学院，北京 100192；4. 北京信息科技大学 信息与通信工程学院，北京 100192）

摘 要：针对火星探测器在复杂地形中自主导航的安全避障需求及星载平台计算资源与能源供应的双重约束，构建YOLOv8-LMD轻量化检测模型，旨在实现火星表面岩石检测算法需兼具的高精度与轻量化特性要求。基于HGNetv2架构重构轻量化主干网络，实现模型参数的初步压缩；设计了一种多尺度特征融合网络结构，通过集成Slim-neck与ASF-YOLO对颈部网络进行重构，有效提升对不同尺度岩石目标的特征表征能力；采用卷积共享策略设计了轻量级检测头，在降低计算复杂度的同时增强分类定位精度；使用剪枝算法针对模型参数冗余进行修剪，使模型进一步压缩，并通过知识蒸馏技术实现精度的补偿优化。通过实验发现，与YOLOv8n相比，YOLOv8-LMD精度提升1.7%，计算量减少68%，参数量减少77%，模型大小减小75%。因此，可认为本文模型更适合应用于火星表面岩石检测任务。

关键词：YOLOv8n；火星表面检测；轻量化；通道剪枝；知识蒸馏

中图分类号：TP391.41 **文献标识码：**A **文章编号：**2096-9287(2025)02-0179-11

DOI:10.15982/j.issn.2096-9287.2025.20250003

引用格式：戴娟，刘经纬，苏中，等. 改进的YOLOv8n轻量化火星表面岩石检测算法[J]. 深空探测学报（中英文），2025，12（2）：179-189.

Reference format: DAI J，LIU J W，SU Z，et al. Lightweight Mars surface rock detection algorithm based on improved YOLOv8n[J]. Journal of Deep Space Exploration，2025，12（2）：179-189.

引 言

目前，随着人类深空探测活动逐渐深入，在月球、火星等深空探测任务取得了诸多珍贵的探测成果[1-5]。而在火星表面岩石探测任务中，火星探测器在复杂地形的自主导航与安全避障至关重要[6]。火星表面复杂的岩石分布对探测器的行驶安全构成了严重威胁，稍有不慎就可能导致探测器损坏。因此，准确检测避开这些岩石障碍物，是保障探测器顺利完成任务的关键环节[7]。同时，星载平台的计算资源与能源供应[8]存在双重限制，这为火星表面岩石检测算法带来了巨大挑战。在这种情况下，要求火星表面岩石检测算法需兼具高精度与轻量化的特性。

近年来，火星表面岩石检测工作在多方面取得进展，文献[9]提出了一个区域火星岩石检测框架，并利用特征和子空间可分性假设设计了两种检测算法，用于火星探测器岩石探测，能以较少延迟生成更精确的岩石检测结果，可以有效地处理包含岩石、阴影和砾石的复杂图像。文献[10]提出了一种新的基于稀疏背景建模的行星探测器自动岩石探测方案，通过区域对比度增强来处理检测问题。通过稀疏表示将标记区域作为字典，在特征空间中重建背景，通过自动阈值直接分割岩石。结果表明，所提的算法产生了高精度的检测结果。文献[11]提出了一种基于火星探测器图像的梯度–区域约束水平集方法，用于火星表面岩石的自动提取，并在各种地形和光照条件下对火星探测器图像进行了实验，结果表明所提出的方法对于自动探测火星表面的小尺度和大型岩石都是鲁棒和高效的。文献[12]提出了一种基于区域对比的自主岩石探测方法，首先根据强度信息和空间布局将图像分割成均匀的区域，然后通过自适应阈值从得到的对比度图中分割岩石，结果表明此算法优于基于边缘的算法。虽然上述检测方法能够明显提升火星表面岩石检测精度，但却忽视了在星载平台严苛的计算资源约束条件下，检测模型的轻量化特性同样重要[13]。鉴于此，本文以YOLOv8n 模型〔YOLO（You Only Look Once）是一种基于深度学习的高效目标检测算法〕为基准模型，提出了一个兼备高精度和轻量化的检测模型应用于火星场景的岩石检测。

收稿日期：2025-01-10 修回日期：2025-04-13

基金项目：国家自然科学基金（61703040，61603047）

1 YOLOv8与改进算法

1.1 YOLOv8算法

YOLOv8作为一种单阶段模型，在YOLOv5的基础上开展了更加实质性的创新性改进。在骨干网络中，YOLOv8n结合了跨阶段局部网络（Cross Stage Partial Network，CSPNet）结构[14]和YOLOv7[15]的高效层聚合网络（Efficient Layer Aggregation Network，ELAN）结构，打造出全新的C2f模块，能够极大地加快模型推理速度并提升目标检测精度。另外在颈部网络层面，YOLOv8n采用先进的路径聚合网络与特征金字塔网络（Path Aggregation Network and Feature Pyramid Network，PAN-FPN）结构[16]构造多尺度金字塔，加强了模型提取特征的能力。在检测头方面，使用功能更为强大的解耦头，根据分类和回归得分加权来分配正负样本，有效地提升模型性能。

与YOLOv5系列模型相比，YOLOv8n运用了更为先进的特征融合技术，采用了更高效的主干网络以及C2f模块，并且在训练阶段使用更先进的数据增强技术，具有更高的检测精度。与YOLOv7和YOLOv9系列模型相比，YOLOv8n拥有更轻量的网络结构和更快的推理速度，其本身更小的模型体积和参数量更适合在火星场景进行硬件部署，所以选择YOLOv8n作为基准模型改进。

1.2 YOLOv8-LMD网络框架

本文在网络结构部分对YOLOv8n的主干网络、颈部网络、检测头做出改进，主要改进包括：主干网络采用HGNetv2的层次化梯度流设计，并使用幽灵卷积构建了Ghost_HGBlock作为主干网络的主要部分；颈部网络主要采用ASF-YOLO架构，主要包括SSFF模块、TFE模块和CPAM模块，同时融入了Slim-neck结构的VOVGSCSP模块和GSConv，构成了一种多尺度特征融合网络结构；在检测头方面，创新性地设计了一种基于共享卷积的轻量型检测头，有效降低了模型复杂度。改进后的YOLOv8-LMD网络如图1所示。

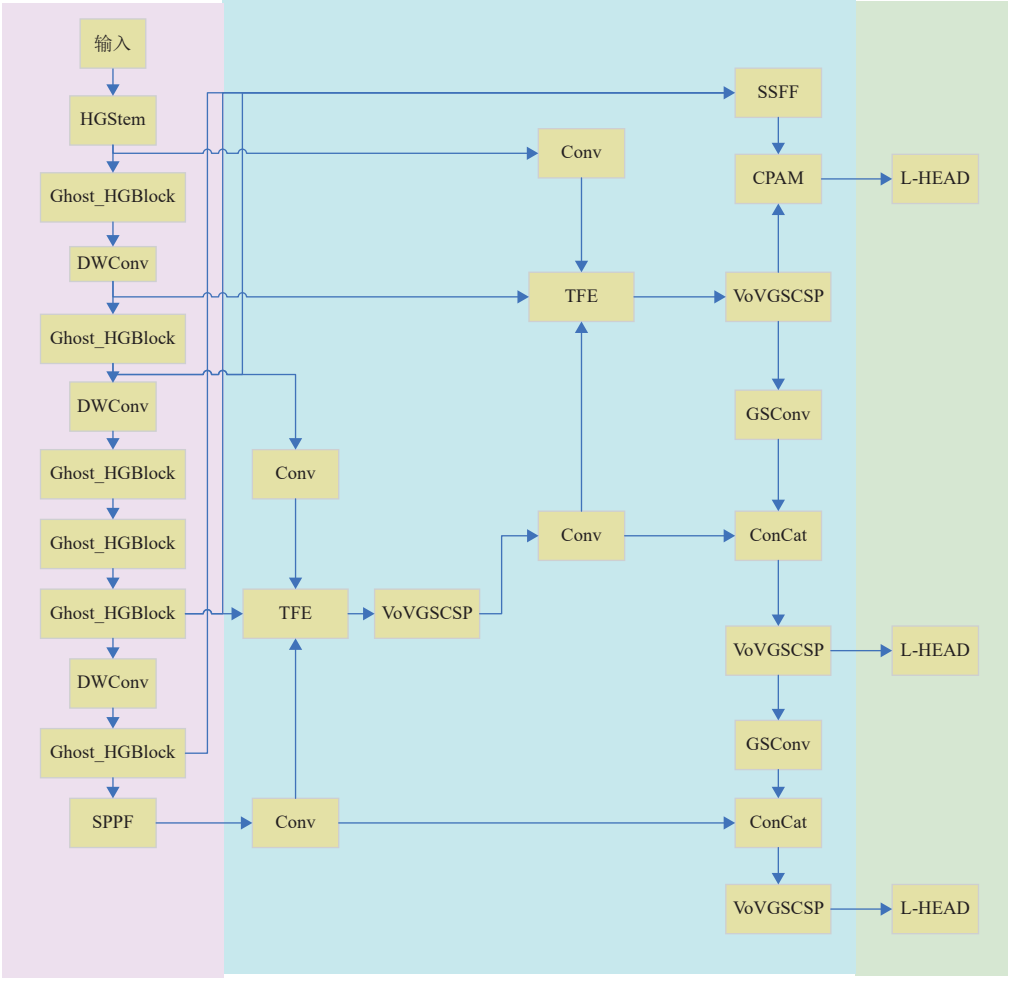


图 1 YOLOv8-LMD网络结构图

Fig. 1 YOLOv8-LMD network structure diagram

算法整体流程：在前处理阶段，使用改进的HGNetv2主干网络对输入图像进行特征提取，再通过多尺度特征融合网络对提取到的特征融合和增强，最后由轻量化检测头输出预测结果。在后优化阶段，采用结构化剪枝对模型冗余参数修剪，并使用知识蒸馏技术对模型精度调整。

1.3　主干网络改进

YOLOv8n的主干结构主要采用了卷积层和C2f模块，以及空间金字塔池化模块（Space Pyramid Pool-Fast，SPPF）来增强模型对多尺度多特征目标的感知能力。由于该主干网络具有较深的网络层和冗余参数，网络训练时间较长，需要较高条件的计算资源才能支持，但是火星探测器搭载的嵌入式平台存在严格的计算功耗限制，难以充分发挥模型性能。基于此，本文使用在目标检测方向拥有更高性能的HGNetv2[17]来作为其新的主干网络，同时使用幽灵卷积[18]对主干网络进一步优化，以达到轻量化效果。此种轻量化设计可减少模型对特定特征的过拟合倾向，使网络更专注于本质特征，提高模型对未知场景的适应能力。

1.3.1　Ghost_HGBlock

幽灵卷积作为一种轻量级卷积，包含两个阶段的卷积操作。第一阶段是使用一半数量的1 × 1卷积核来对目标本身进行特征提取，从而得到整体特征图的一半。第二阶段使用5 × 5大小卷积核通过廉价计算得到另一半特征图。最后通过Concat操作将两个不同的特征图进行拼接。通过引入幽灵卷积可显著地降低计算成本、压缩模型大小，使其更容易在计算资源有限的环境进行硬件部署。

GhostHGNetv2主干网络可以比原主干网络提取更丰富的特征信息，通过引入幽灵卷积所构成的Ghost_HGBlock作为其主要部分，网络结构如图2所示。Ghost_HGBlock的主要特点是可将输入通道分成一个主通道和一个幽灵通道来实现参数共享，这极大地减少了模型参数量并降低了计算复杂度，使模型保持良好的性能。

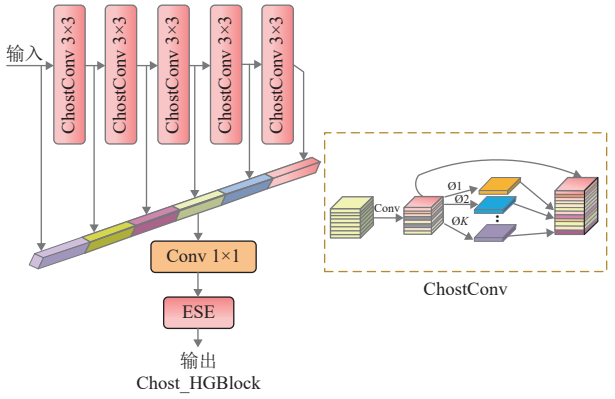


图 2　Ghost_HGBlock结构图

Fig. 2　Ghost_HGBlock structure diagram

1.3.2　深度可分离卷积

YOLOv8n的主干网络采用的是普通卷积，其优点是有强大的表征能力，然而每个卷积核都要与输入的每个通道进行卷积操作，因此其参数数量和计算量较大。与普通卷积相比，深度可分离卷积（Depthwise Convolution，DWConv）是一种更轻量化的卷积操作，主要包含两个过程，分别是逐通道卷积和逐点卷积，其网络结构如图3所示。

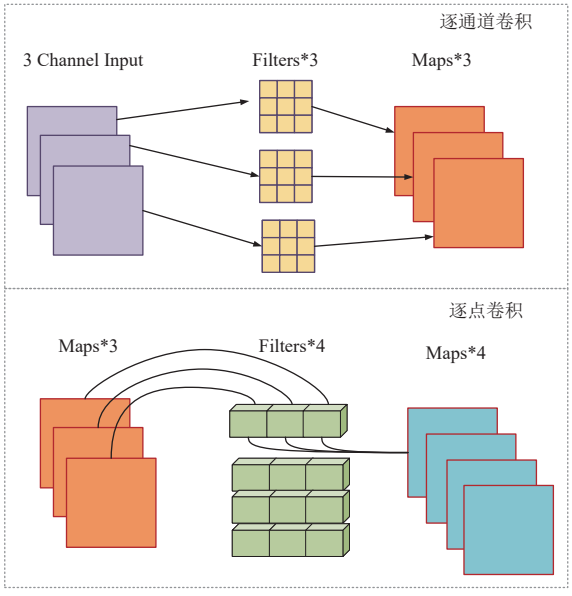


图 3　DWConv结构图

Fig. 3　DWConv structure diagram

逐通道卷积将每一个输入通道都使用单独的卷积核进行卷积操作，避免了像普通卷积那样同时处理所有通道，显著减少计算参数量。逐点卷积使用1 × 1卷积核对逐通道卷积的结果进行卷积。它在深度方向上合并深度卷积产生的特征图，以生成最终的输出特征图。采用深度可分离卷积可以极大地减少模型的参数量和计算量，降低模型复杂度，提高运算效率。

1.4　颈部网络改进

由于火星表面岩石的尺度多样性以及存在复杂背景的干扰，这对目标检测模型的特征融合能力提出了更高的要求，于是本文设计了一种多尺度特征融合网络结构。通过采用Slim-neck结构[19]和ASF-YOLO（Attention-al Scale Sequence Fusion based You Only Look

Once）结构[20]对颈部网络重构。此设计有效弥补了主干网络轻量化造成的精度损失，提升了模型多尺度特征融合能力。

1.4.1 Slim-neck

Slim-neck结构主要由GSConv和VOVGSCSP两部分组成，网络结构图如图4所示。

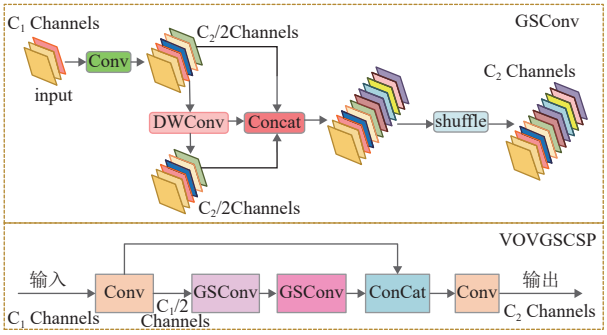


图 4 VOVGSCSP结构图

Fig. 4 VOVGSCSP structure diagram

GSConv先将输入经过普通卷积，再将其进行深度卷积，随后将两个结果拼接起来，最后进行Shuffle操作，将之前两个卷积结果对应的通道拼接在一起。这种结构能够在降低模型复杂度的同时，显著提升模型运行速度，同时结合一般卷积和组卷积对特征的处理方式，通过分组处理特征图，不仅能够捕捉局部特征，保持全局上下文信息，提高模型的感知能力，还能减少卷积核的数量，节省存储空间并提高运行速度。

VOVGSCSP结构首先将特征图经过普通卷积，通道数减半，再经过两次GSConv进行全局稀疏操作，再与原始卷积结果进行拼接，实现了全局信息的有效聚合，此结构有助于模型更好地捕捉图像中的目标，同时提高特征图传递效率。

1.4.2 ASF-YOLO

ASF-YOLO结合空间和尺度特征，以实现准确且快速的目标检测和实例分割。该框架建立在YOLO检测框架，采用尺度序列特征融合（Scale Sequence Feature Fusion，SSFF）模块融合来自多尺度图像的全局语义信息，通过有效地组合多尺度特征图，增强模型对不同尺度图像的信息提取能力。TFE（Triple Feature Encoder）模块通过捕获局部细节和融合多尺度特征来增强小目标检测能力，通过分析不同尺度的形状和外观变化来增强对密集小目标的识别。CPAM（Channel and Position Attention Mechanism ）模块利用空间注意力机制，以整合SSFF和TFE模块，该机制专注于与信息通道和空间位置相关的小物体，以改进检测性能，网络结构如图5所示。

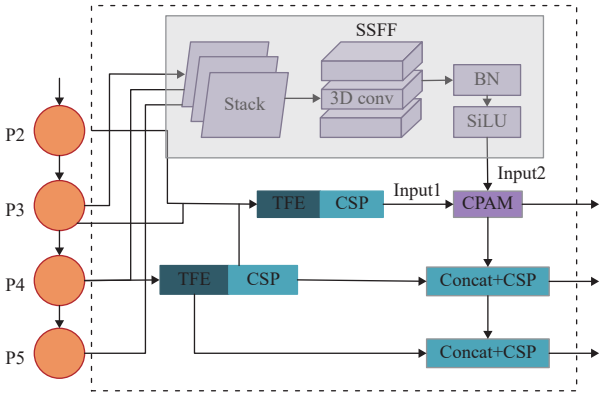


图 5 ASF-YOLO网络结构图

Fig. 5 ASF-YOLO network structure diagram

ASF-YOLO结构不仅在细胞实例分割领域表现出色，其多尺度信息提取、特征融合与信息增强等技术特点使ASF-YOLO在目标检测领域也具有实际的意义。在火星表面岩石检测任务中，ASF-YOLO能够有效地处理不同大小、形状、颜色的目标，包括弥补目标检测中长宽比方面的缺陷，增强对密集目标与小目标特征的敏感度，同时结合Slim-neck可更好地自适应调整对不同通道和空间位置的全局关注，缓解因目标尺度差异导致的漏检问题，进而提升检测模型对不同尺度目标的泛化能力。

1.5 一种轻量化的共享卷积检测头

YOLOv8n模型的检测头采用的是解耦合头，检测头有2个分支，每个分支均包括2个3维卷积以及1个1维卷积，此种检测头计算开销较大。同时，由于火星与地球之间的通信延迟和带宽限制，传输数据的成本非常高昂。为降低数据传输的成本，本文借助共享卷积创新性地设计了一种轻量型检测头LDH（Lightweight Detection Head）来解决上述问题，其结构如图6所示。

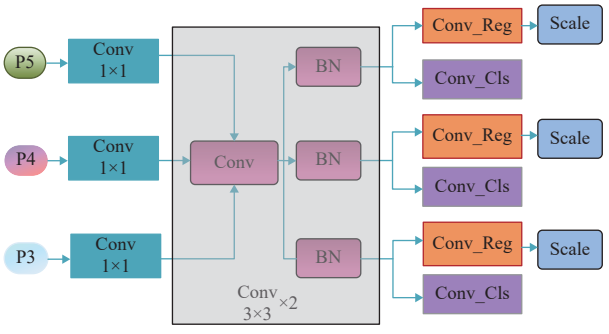


图 6 LDH结构图

Fig. 6 LDH structure diagram

LDH检测头的最大优点在于实现卷积的参数共享，且对BN层进行独立计算。由于不同层级之间的参数量存在差异，导致归一化层不可或缺，但是又由于共享卷积特性，引入归一化会导致存在滑动平均值误差。虽然组归一化[21]能够避免此问题，但其会增加大量的推理开销，所以本文决定在卷积共享的同时，对BN层进行独立计算。

此模块首先从P3、P4、P5检测层提取特征图，然后经过一个1 × 1卷积进行通道调整，再经过共享参数的2个3 × 3卷积提取检测目标的特征，且BN层独立计算。之后经过定位分支和分类分支分离。此设计采用参数共享，同时考虑到检测目标的尺度问题，利用Scale层进行相应大小的缩放，增加对不同尺寸目标的定位能力。LDH检测头通过共享卷积策略有效地减小了模型的参数量和计算量，避免因计算冗余导致的泛化性能下降，其多尺度特征识别能力使模型更容易识别更小的目标，提高模型在不同环境检测性能的稳定性。

1.6　基于LAMP的剪枝方法

虽然YOLOv8n在经过网络改进后已具备轻量化特性，但模型中仍存在过多冗余参数，难以满足火星表面岩石检测任务的实时性要求，因此需采用模型剪枝技术对模型进行进一步压缩。本文使用基于层自适应幅度的剪枝算法[22]（LAYER-ADAPTIVE MAGNITUDE-BASED PRUNING，LAMP）来对改进后的模型进行修剪。LAMP算法是一种深度神经网络模型的剪枝方法，旨在通过剔除模型中的冗余参数来降低模型的计算量和参数量，进一步提升模型的性能。

$$score(u;W):=\frac{(W[u])^2}{\sum_{v\geqslant u}(W[v])^2} \tag{1}$$

$$(W[u]^2)>(W[v]^2)\rightarrow score(u;W)>score(v;W) \tag{2}$$

由于卷积层权重和全连接层权重的维数不同，所以将每个权重张量全部压为一维向量，并计算幅值。在式（1）中，W表示权重，u、v表示权重的索引，$W[u]$表示索引u所对应的权重项，分子代表目标连接的权重幅值的平方，分母代表在同一层其它剩余连接的权重幅值平方之和，由式（1）可计算出第u个索引对应权重的LAMP分数。在式（2）中，权重将按照给定的索引映射进行排序，权重项越大，与之对应的分数就越高，表明参数重要性越高，反之分数越低，则判定该权重项为相对不重要的部分，同时予以剪枝。

$$\begin{pmatrix}U_1\\U_2\\U_3\\\cdots\end{pmatrix}\begin{pmatrix}V_1\\V_2\\V_3\\\cdots\end{pmatrix}\rightarrow\begin{pmatrix}\dfrac{U_1^2}{U_1^2+U_2^2+U_3^2}\\\dfrac{U_2^2}{U_1^2+U_2^2+U_3^2}\\\dfrac{U_3^2}{U_1^2+U_2^2+U_3^2}\\\cdots\end{pmatrix}\begin{pmatrix}\dfrac{V_1^2}{V_1^2+V_2^2+V_3^2}\\\dfrac{V_2^2}{V_1^2+V_2^2+V_3^2}\\\dfrac{V_3^2}{V_1^2+V_2^2+V_3^2}\\\cdots\end{pmatrix}\rightarrow$$
$$\begin{pmatrix}\dfrac{U_1^2}{U_1^2+U_2^2} & \dfrac{V_1^2}{V_1^2}\\\dfrac{U_2^2}{U_1^2+U_2^2} & \varnothing\\\varnothing & \varnothing\\\cdots & \cdots\end{pmatrix} \tag{3}$$

本文采用LAMP分数剪枝操作，修剪流程如式（3）所示，省略号代表向量中其它未展开的元素。主要流程：首先对通过基础模型训练得到的权重文件（待修剪的权重文件）初始化，并计算出连接的权重幅值的平方，通过归一化处理得到同一层所有权重的幅值平方之和，即LAMP分数。依据此分数和预置的剪枝率参数，选择相应的连接修剪。在修剪幅度达到预期后，要对修剪后的模型微调（finetune），目的是使修剪后的模型尽可能达到较高的精度。LAMP剪枝算法的优点在于在计算中，每一层都将有一个评分为1的最优通道，有效避免了层崩溃的发生，将全局剪枝和局部剪枝的优势有效地融合。本文通过使用LAMP剪枝实现模型进一步轻量化的同时维持模型的性能，显著降低了计算资源需求。

1.7　基于CWD的知识蒸馏方法

虽然剪枝可有效地减少参数量和计算量，但模型剪枝造成的精度损失也不可忽视。尽管已经对剪枝后的模型训练来实现精度微调，但根据实验数据来看，微调效果有限，所以本文通过使用知识蒸馏来对剪枝性能劣化进行弥补。

知识蒸馏有两大类：一类为基于标签的知识蒸馏，另一类为基于特征的知识蒸馏。其中标签知识蒸馏方法是将教师模型输出的标签作为学生学习的知识，使学生模型的输出标签与教师模型一致，来达到指导学习的目的。但是这种蒸馏方法有一定的缺点，如果仅仅是依靠教师模型输出的标签来指导学生模型学习，导致学生模型缺乏对特征的理解，所学习到的知识有限，不利于模型性能恢复。学生模型需要学习更多隐藏在教师模型中的检测目标特征知识，来解决通道差异导致特征表达能力不同的问题[23]。因此本文选择采用基于通道的CWD（Channel-Wise Knowledge Distillation）特征知识蒸馏算法[24]，特征知识蒸馏是将

教师模型中某些特征层对检测目标的特征表达作为学生模型的学习内容，学生模型可以直接学习目标特征知识，使学生模型某些层的特征表达可以拟合教师模型对应层的特征表达[25]，其架构如图7所示。

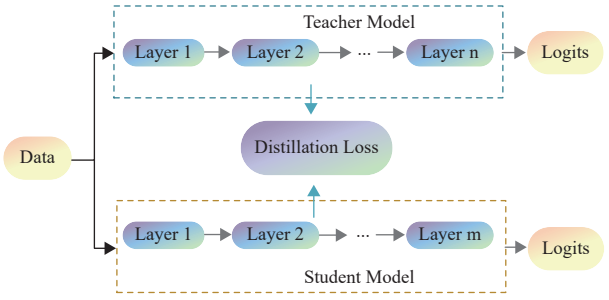


图 7 蒸馏结构图

Fig. 7 Distillation structure diagram

由于剪枝不改变模型的整体架构，所以教师模型和学生模型架构相同，每一层之间一一对应，以便学生可以直接学习教师模型指定层的特征表达。本文选择基于中间层的特征知识蒸馏，所选中间层为颈部网络中对应生成P3、P4、P5特征图的层，具体对应层数为18、21、24，蒸馏损失函数见式（4）。

$$L_{\mathrm{KD}} = \| F^{\mathrm{S}}(x_i) - F^{\mathrm{T}}(x_i) \| \tag{4}$$

其中：L_{KD}为中间特征知识蒸馏损失；F^{S}和F^{T}为维度变换函数。公式特点在于将学生与教师网络输出的特征图转换到相同维度，进而计算蒸馏损失，利用反向传播的方式来优化学生网络。此种计算方式使学生模型的中间层特征表达尽可能与教师模型对应中间层的特征表达相拟合，使学生学习到更多的知识，从而提升特征知识蒸馏的效果。

2 实验对比

针对在火星场景的火星表面岩石检测算法需兼具的高精度与轻量化特性要求，本文以YOLOv8n为基准模型，提出了YOLOv8-LMD模型，具体改进点包括重构主干网络和颈部网络，以及对检测头的创新性设计，同时使用剪枝算法和知识蒸馏技术实现模型进一步压缩和性能提升。为验证各方面改进的有效性，在本节将进行多组对比实验对模型性能进行全面评估。

2.1 实验准备

2.1.1 实验设备

本次实验基于Ubuntu 22.04操作系统，使用python语言实现，应用深度学习框架PyTorch 2.1.0搭建网络模型，Python 3.11编译程序。硬件使用RTX 4090显卡，输入图像大小为640 × 640，batch-size设置为16，训练轮数为250轮，其它参数为默认值。

2.1.2 数据集

为验证本文提出的轻量化模型应用于火星表面岩石检测任务的效果，采用自建火星表面岩石数据集进行验证，图片数据来源于美国国家航空航天局（National Aeronautics and Space Administration，NASA）官网火星图片以及火星“好奇号”（Curiosity）探测器、“毅力号”（Perseverance）探测器等拍摄的真实火星场景。共制作2 000张图片，包括训练、验证、测试3个部分，按照8∶1∶1的比例分为1 600∶200∶200。为进一步增加数据集的丰富性，对原图片翻转、噪声、改变亮度等数据增强处理，用于训练在不同环境和不同角度的检测情况，提高鲁棒性。

2.1.3 评价指标

由于检测模型需同时满足高精度识别与轻量化特性的双重约束条件，所以本文选取指标如下。

以准确率（Precision，P）、召回率（Recall，R）、平均精度均值（mean Average Precision，mAP）来衡量模型检测精度，以计算量（Floating-point Operations Per Second，FLOPs）、参数量（Parameters）、模型体积（Size）衡量轻量化特性。

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$AP = \int_0^1 P \mathrm{d}R \tag{7}$$

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{N} \tag{8}$$

其中：N为标注类别个数；TP为正确识别岩石标注的样本数量；FP为错误识别岩石标注的样本数量；FN为错误识别非岩石标注的样本数量。式（5）和式（6）分别求得模型识别的准确率和召回率，式（7）表示为每个类别的PR曲线与坐标轴围成的面积，与模型优秀程度呈正相关。式（8）代表计算所有类别的平均AP，用于衡量模型在所有类别的性能。

2.2 主干网络实验对比

为验证改进网络主干的有效性，增加与其它轻量型主干网络的对比实验，为避免偏差，在实验中确保除主干网络以外的其它部分和训练参数均保持一致，实验数据见表1。

表 1　主干网络对比

Table 1　Backbone network comparison

主干名称	mAP50/%	Parameters	FLOPs/G	Size/M
Mobilenetv3	80.6	2 841 251	6.2	5.8
Ghostnet	81.7	2 771 463	6.8	5.7
Shufflenetv2	80.1	2 046 522	6.1	4.3
本文主干	81.4	2 314 041	6.9	4.8

本文使用Mobilenetv3[26]、Ghostnet、Shufflenetv2[27]进行主干网络对比，通过实验发现，使用Shufflenetv2为主干时，参数量和计算量最低，模型体积最小，但是检测精度比本文主干低1.3%。使用Ghostnet时，精度虽然比本文主干高0.3%，但是其参数量较多，模型体积比本文主干大0.9 M。使用Mobilenetv3为主干时，计算量比本文主干低0.7 G，但是同时考虑到精度、参数量和模型体积，本文主干仍具优势。综上所述，通过与主流轻量化主干网络对比，本文改进的主干网络具有显著的性能优势。

2.3　剪枝实验对比

通过剪枝实验，发现不同的剪枝率对检测精度、参数量和计算量影响显著，为找到最优的参数，本文进行了剪枝对比实验，通过调整剪枝参数，将剪枝率分别设置为1.5、2.0、2.5、3.0进行对比实验，结果见表2。

表 2　剪枝对比实验

Table 2　Comparative pruning experiments

剪枝率	mAP50/%	Parameters	FLOPs/G	Size/M
1.5	82.9	948 286	3.5	1.9
2.0	82.7	689 374	2.6	1.5
2.5	80.8	528 395	2.1	1.3
3.0	78.1	349 785	1.7	1.0

通过实验发现，当剪枝率大于2.0时，模型检测精度损失严重，当调整为2.5时，虽然参数量和计算量明显降低，但是精度下降3.2%，而调整为3.0时，精度下降5.9%，均超出可接受范围。相比之下，当剪枝率为1.5和2.0时，精度保持良好。其中，剪枝率为2.0的模型相较于剪枝率为1.5的模型虽然精度低了0.2%，但参数量减少27%，计算量降低26%，模型体积缩小0.4M，基于此，选择剪枝率为2.0作为最优的实验参数。

剪枝前后通道结果对比如图8所示，黄色部分代表剪枝移除的通道，红色部分表示剪枝后的剩余通道，图中横坐标表示模型中的特定层，纵坐标表示通道数。根据图8可以很清晰地观察到剪枝前后通道变化情况，更直观了解到剪枝的效果。

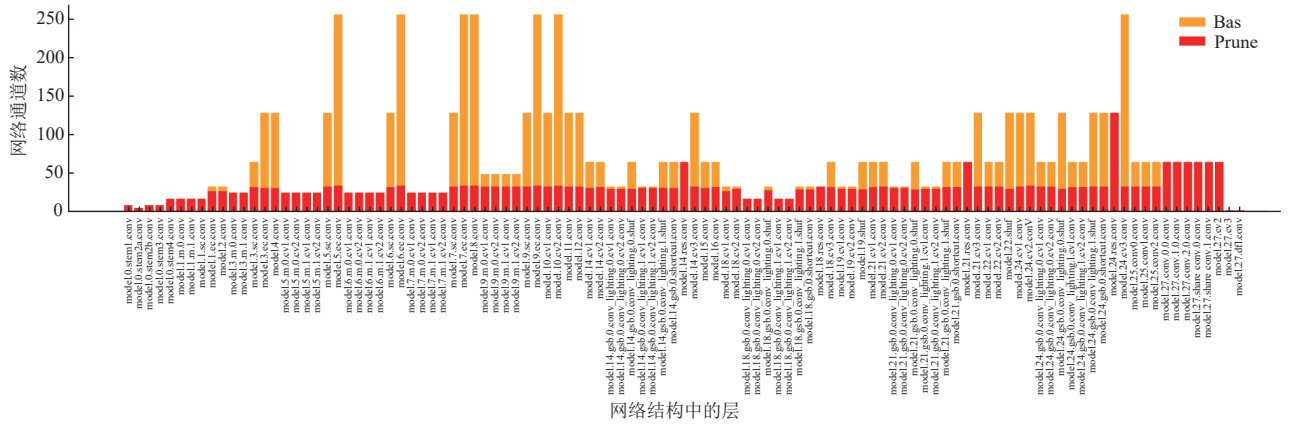


图 8　通道对比图

Fig. 8　Channel contrast diagram

2.4　知识蒸馏实验对比

通过实验发现，不同的蒸馏权重系数对于模型精度有显著影响，本文选择不同的蒸馏权重系数进行对比实验，将系数调整为0.5、1、1.5、2作为4组实验初始数据，实验数据见表3。

另外，选择不同的知识蒸馏系数，准确率和召回率波动幅度较大，说明不同的知识蒸馏系数对于岩石特征信息提取具有一定的影响。当系数调为1或1.5时识别精度较高，其中当系数为1时，检测精度虽然比系数为1.5时的检测精度低了0.1%，但是其识别准确率相对高了1.8%，当模型具有更高的准确率和对应更低的召回率时，表明模型在减少误检率方面表现更优，进而证明检测性能更强，所以本文选择权重系数1为最优的实验参数。

表 3　蒸馏对比实验

Table 3　Comparative distillation experiments

权重系数	*P*/%	*R*/%	mAP50/%
0.5	78.8	74.4	83.1
1.0	79.2	74.1	83.6
1.5	77.4	75.5	83.7
2.0	78.6	74.9	83.3

2.5 消融实验

为验证各项改进措施的作用，开展消融实验，实验数据结果见表4。

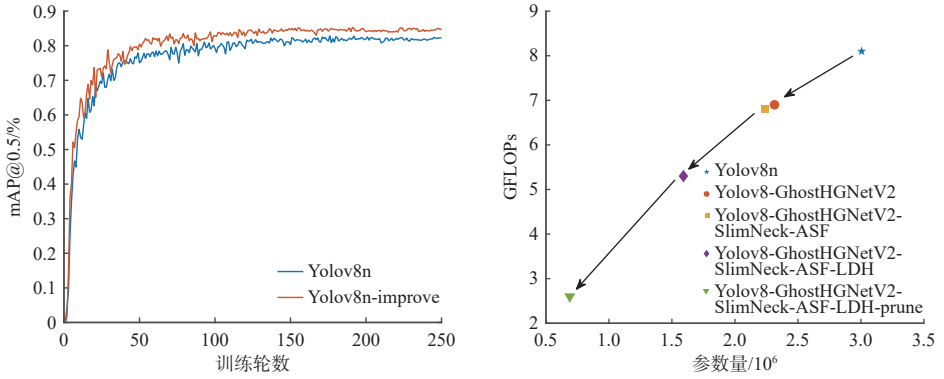
表 4 消融实验

Table 4 Ablation experiments

改进主干	改进颈部	LDH	剪枝	蒸馏	FPS	mAP50/%	Parameters	FLOPs/G	Size/M
—	—	—	—	—	149.2	81.9	3 005 843	8.1	6.1
√	—	—	—	—	125.0	81.4	2 314 041	6.9	4.8
√	√	—	—	—	95.1	82.6	2 235 524	6.8	4.7
√	√	√	—	—	109.3	84.0	1 591 350	5.2	3.4
√	√	√	√	—	99.3	82.7	689 374	2.6	1.5
√	√	√	√	√	106.1	83.6	689 374	2.6	1.5

消融实验结果显示，本文改进主干在精度降低0.5%的同时，参数量减少23%，计算量降低15%，模型体积压缩21%，完成初步轻量化。然后进一步改进颈部网络，精度提升1.2%，与未改进时相比提升了0.7%，同时具有轻量化的效果。之后通过引入轻量化检测头，凭借参数共享的优势，不仅使检测精度提升1.4%，同时计算量降低23%，参数量减少29%。随后进行剪枝操作，虽然精度降低1.3%，但是参数量减少57%，计算量降低50%，模型体积减小56%，使模型得到了进一步压缩。最后通过知识蒸馏，使模型检测精度回调至83.6%。综合来看，改进后的YOLOv8-LMD模型与YOLOv8n相比，精度提升了1.7%，参数量降低了77%，计算量减少了68%，模型体积压缩了75%，并且FPS达到106.1，模型性能得到大幅提升。

改进前后数据对比如图9所示，图9（a）为改进前后的精度对比曲线，在训练达到拟合后，改进后的模型精度高于基准模型。图9（b）为在改进过程中各个改进对模型轻量化效果分析图，横坐标为参数量，纵坐标为计算量。从图9可看出每次改进后，模型的参数量和计算量均向坐标左下方偏移，表明模型轻量程度逐步提升，证明了改进后的有效性。



（a）改进前后精度对比曲线 （b）改进过程中轻量参数变化图

图 9 改进前后数据对比

Fig. 9 Comparison of data before and after improvement

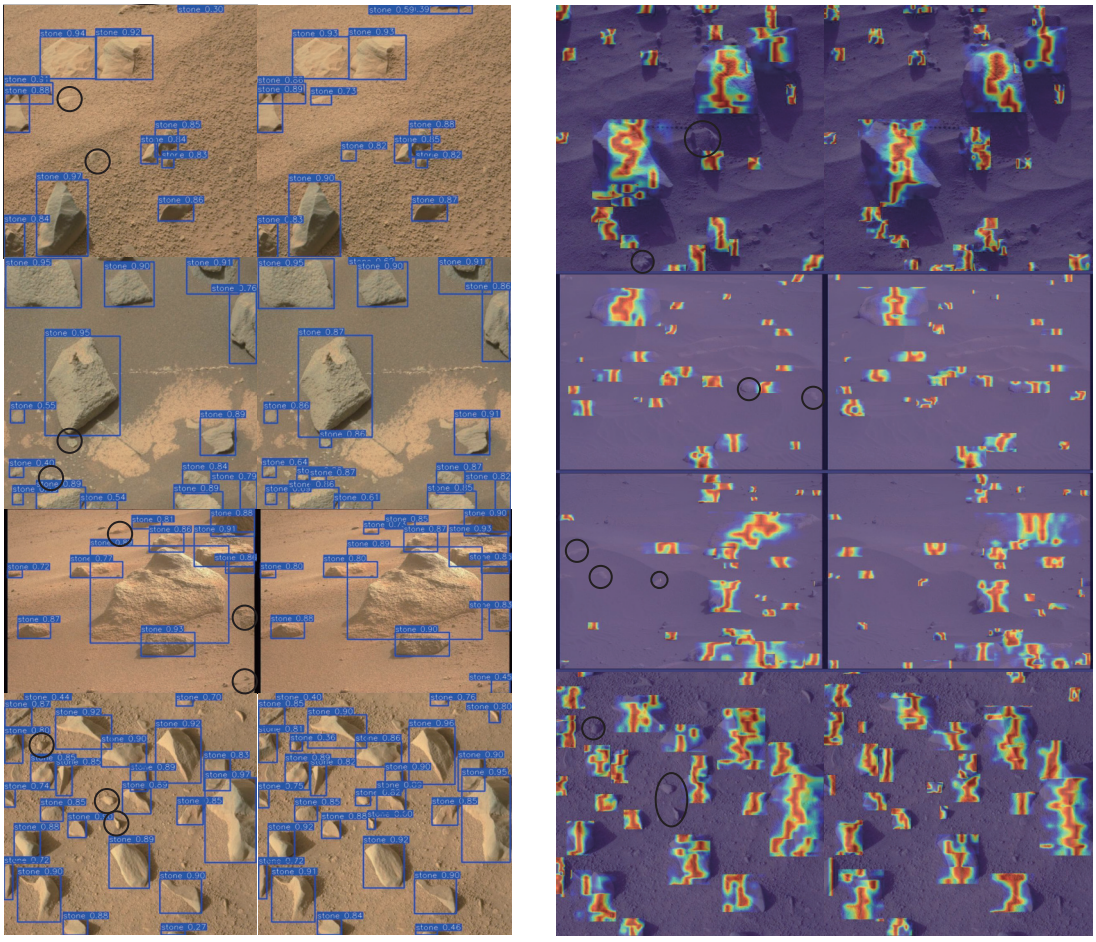
2.6 可视化对比图

可视化对比图结果如图10所示，其中图10（a）为改进前后的效果对比图，图10（b）为热力图的对比可视化结果。左侧为基准模型检测结果，右侧为改进模型检测结果，黑色笔标注区域为基准模型未识别出的目标，可以看出本文改进模型提高了对小目标的检测能力，证明了改进后模型具有更好的性能。

2.7 模型对比实验

为进一步验证本文模型的性能，本文选取15 组主流模型与本文改进模型进行对比，实验结果见表5。

对比发现，本文改进模型相比于Faster-RCNN和SSD，在精度上分别高了9.9%和7.2%，有着显著优势。RT-DETR-L和RT-DETR-R50模型虽然精度较高，但是参数量分别是本文改进模型的41倍和61倍，计算量更是相差接近50倍，轻量程度差距过大。与轻量化模型MobileNetv3、NanoDet和EfficientNet相比，本文模型在精度方面分别高出6.2%、11.4%和9%，展现了性能优势。YOLOv8s和YOLO11n虽然检测精度较高，但是本文改进模型与YOLOv8s和YOLO11n相比在计算量上分别低了91%和59%，在参数量分别低了96%和73%。YOLOv7虽然具备较高的精度，但是YOLOv7的参数量和计算量与本文模型相比分别高了98%和97%。与之相比，YOLOv10n具备良好的性能，但是其检测精度比本文模型低了2.5%，另外参数量和计算量也存在明显差距。同样，YOLOv9的检测精度比本文模型低了2.2%，且参数量冗余的特性不符合本文轻量化的需求。综上所述，通过与其它主流模型相比，本文所改进的YOLOv8-LMD模型具备更好的性能。



（a）效果对比图　　（b）热力图可视化

图 10　可视化对比图

Fig. 10　Visualisation comparison chart

表 5　模型对比实验

Table 5　Model comparison experiments

模型	mAP50/%	Parameters	FLOPs/G	Size/M	FPS
Faster-RCNN	73.7	137 457 896	360.6	260.0	33.0
SSD	76.4	26 074 564	67.2	52.0	60.0
RT-DETR-L	80.9	28 445 315	100.6	56.3	100.0
RT-DETR-R50	81.5	41 936 736	125.6	82.0	116.0
MobileNetv3	77.4	1 658 749	2.2	3.5	98.0
NanoDet	72.2	977 484	1.4	2.0	104.0
EfficientNet	74.6	4 270 434	3.1	6.1	91.0
YOLOv5n	81.1	2 544 652	7.2	5.1	131.2
YOLOv7	82.6	37 196 556	105.1	73.1	56.4
YOLOv7-tiny	80.8	6 015 478	13.3	12.6	127.1
YOLOv8n	81.9	3 005 843	8.1	6.1	149.2
YOLOv8s	83.2	11 125 971	28.4	22.0	129.9
YOLOv9	81.4	7 167 475	26.7	14.9	65.4
YOLOv10n	82.1	2 694 806	8.2	5.6	**237.1**
YOLO11n	83.4	2 582 347	6.3	5.4	108.5
YOLOv8-LMD	**83.6**	**689 374**	**2.6**	**1.5**	106.1

3　结　论

本文针对火星表面岩石检测算法需兼具高精度与轻量化特性的要求，提出了以YOLOv8n为基准改进的YOLOv8-LMD模型。首先通过设计轻量化主干网络使模型初步达到轻量化效果；其次设计了一种多尺度特征融合网络结构，利用Slim-neck和ASF-YOLO重新构造颈部网络，增强模型多尺度特征融合能力，解决了由于主干网络轻量化导致的精度损失问题，同时具有轻量化的效果，然后设计了一种轻量化检测头，其特殊的卷积共享结构使模型轻量的同时能够进一步提高模型对小目标的特征提取能力；另外针对模型存在过多的冗余参数，对其进行剪枝操作，使模型进一步压缩，通过实验调整剪枝参数，实现精度与轻量化的平衡；最后对剪枝后的模型使用知识蒸馏进行最后的精度调整。相比于原始模型，本文改进的YOLOv8-LMD在精度提升1.7%的同时，计算量降低68%，参数量减少77%，模型体积压缩75%，FPS达到106.1，因此，可认为YOLOv8-LMD更适合应用于火星表面岩石检测任务。

参考文献

[1]　于登云，吴学英，吴伟仁. 我国探月工程技术发展综述[J]. 深空探测

学报(中英文)，2017，3(4)：307-314.
YU D Y，WU X Y，WU W R. An overview of the development of China's moon exploration engineering technology[J]. Journal of Deep Space Exploration，2017，3(4)：307-314.

[2] 于登云，孙泽洲，孟林智，等. 火星探测发展历程与未来展望[J]. 深空探测学报(中英文)，2016，3(2)：108-113.
YU D T，SUN Z Z，MENG L Z，et al. Mars exploration development history and future prospects[J]. Journal of Deep Space Exploration，2016，3(2)：108-113.

[3] 崔平远，张成宇，朱圣英，等. 小天体柔性附着技术[J]. 宇航学报，2023，44(6)：805-816.
CUI P Y，ZHANG C Y，ZHU S Y，et al. Flexible attachment technology for small bodies[J]. Journal of Astronautics，2023，44(6)：805-816.

[4] 崔平远，陆晓萱，朱圣英，等. 小天体柔性附着状态协同估计方法[J]. 宇航学报，2022，43(9)：1219-1226.
CUI P Y，LU X X，ZHU S Y，et al. Collaborative estimation of flexible attachment states of small bodies[J]. Journal of Astronautics，2022，43(9)：1219-1226.

[5] 董光亮，李海涛，郝万宏，等. 中国深空测控系统建设与技术发展[J]. 深空探测学报(中英文)，2018，5(2)：99-114.
DONG G L，LI H T，HAO W H，et al. Development and future of China's deep space TT&C system[J]. Journal of Deep Space Exploration，2018，5(2)：99-114.

[6] GERDES，LEVIN，AZKARATE，et al. Efficient autonomous navigation for planetary rovers with limited resources[J]. Journal of Field Robotics，2020，37(7)：1153-1170.

[7] DAFTRY，SHREYANSH，ABCOUWER，et al. MLNav：learning to safely navigate on Martian terrains[J]. IEEE Robotics and Automation Letters，2022，7(2)：5461-5468.

[8] BISWAL M，MALAYA KUMAR，KUMAR V，et al. Power options for human Mars mission[EB/OL]. (2021-8-9)[2025-1-10]. https://arc.aiaa.org/doi/full/10.2514/6.2021-3260

[9] XIAO X M，YAO M B，LIU H Q，et al. A kernel-based multi-featured rock modeling and detection framework for a mars rover[J]. IEEE Transactions on Neural Networks and Learning Systems，2021，34(7)：3335-3344.

[10] XIAO X M，CUI H T，YAO M B，et al. Auto rock detection via sparse-based background modeling for mars rover[C]//Proceedings of 2018 IEEE Congress on Evolutionary Computation (CEC). Rio de Janeiro，Brazil：IEEE，2018.

[11] YANG J T，KANG Z Z. A gradient-region constrained level set method for autonomous rock detection from Mars rover image[J]. The International Archives of the Photogrammetry，Remote Sensing and Spatial Information Sciences，2019，42：1479-1485.

[12] XIAO X M，CUI H T，YAO M B，et al. Autonomous rock detection on mars through region contrast[J]. Advances in Space Research，2017，60(3)：626-635.

[13] KAMATH V，RENUKA A. Deep learning based object detection for resource constrained devices：systematic review，future trends and challenges ahead[J]. Neurocomputing，2023，531：34-60.

[14] WANG C Y，LIAO H Y M，WU Y H，et al. CSPNet：a new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle，WA，USA：IEEE，2020.

[15] WANG C Y，BOCHKOVSKIY A，LIAO H Y M. YOLOv7：trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，BC，Canada：IEEE，2023.

[16] LIU S，QI L，QIN H，et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City，UT，USA：IEEE，2018.

[17] ZHAO Y，LV W，XU S，et al. Detrs beat YOLOs on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，WA，USA：IEEE，2024.

[18] HAN K，WANG Y H，TIAN Q，et al. Ghostnet：more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，WA，USA：IEEE，2020.

[19] LI H L，LI J，WEI H B，et al. Slim-neck by GSConv：a better design paradigm of detector architectures for autonomous vehicles[EB/OL]. (2022-6-6)[2025-1-10]. https://arxiv.org/abs/2206.02424.

[20] KANG M，TING C M，TING F F，et al. ASF-YOLO：A novel YOLO model with attentional scale sequence fusion for cell instance segmentation[J]. Image and Vision Computing，2024，147：105057.

[21] TIAN Z，SHEN C H，CHEN H. Fcos：a simple and strong anchor-free object detector[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022，44(4)：1922-1933.

[22] LEE J，PARK S，MO S，et al. Layer-adaptive sparsity for the magnitude-based pruning[EB/OL]. (2021-5-9)[2025-1-10]. https://arxiv.org/abs/2010.07611

[23] 黄震华，杨顺志，林威，等，知识蒸馏研究综述[J]. 计算机学报，2022，45(3)：624-653.
HUANG Z H，YANG S Z，LIN W，et al. A review of knowledge distillation research[J]. Journal of Computing，2022，45(3)：624-653.

[24] SHU C Y，LIU Y F，GAO J F，et al. Channel-wise knowledge distillation for dense prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal，QC，Canada：IEEE，2021.

[25] 邵仁荣，刘宇昂，张伟，等. 深度学习中知识蒸馏研究综述[J]. 计算机学报，2022，45(8)：1638-1673.
SHAO R R，LIU Y A，ZHANG W，et al. A review of research on knowledge distillation in deep learning[J]. Journal of Computing，2022，45(8)：1638-1673.

[26] HOWARD A，SANDLER M，CHU G，et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Barcelona，Spain：IEEE，2019.

[27] MA N，ZHANG X，ZHENG H T，et al. Shufflenet v2：practical guidelines for efficient cnn architecture design[EB/OL]. (2018-10-9)[2025-1-10]. https://link.springer.com/chapter/10.1007/978-3-030-01264-9_8#chapter-info.

作者简介：
戴娟(1984–)，女，博士，副研究员，主要研究方向：自主导航制导与控制、深空探测器制导与控制、智能控制方法等。**本文通信作者**。
通信地址：北京市海淀区清河小营东路12号北京信息科技大学自动化学院。
E-mail：daijuan@bistu.edu.cn

Lightweight Mars Surface Rock Detection Algorithm Based on Improved YOLOv8n

DAI Juan[1, 2, 3]，LIU Jingwei[1, 2, 3]，SU Zhong[1, 2, 3]，ZHU Cui[4]

（1. University of Beijing Information Science & Technology Beijing Key Laboratory of High Dynamic Navigation Technology，Beijing 100192，China；
2. Key Laboratory of Modern Measurement & Control Technology，Ministry of Education，Beijing 100192，China；
3. School of Automation，Beijing Information Science & Technology University，Beijing 100192，China；
4. School of Information and Communication Engineering，University of Beijing Information Science & Technology，Beijing 100192，China）

Abstract：Due to the demand for safe obstacle avoidance in the autonomous navigation of Mars rover in complex terrain and the double constraints of computational resources and energy supply of the onboard platform，a lightweight detection model, YOLOv8-LMD, was constructed，aiming at realizing the requirements of high precision and lightweight characteristics of the rock detection algorithm on the surface of Mars. First，the lightweight backbone network was reconstructed based on the HGNetv2 architecture to realize the preliminary compression of model parameters. Secondly，a multi-scale feature fusion network structure was designed，and the neck network was reconstructed by integrating slim-neck and ASF-YOLO to effectively improve the feature characterization of rock targets at different scales. In addition，a lightweight detection head was designed by using the convolutional sharing strategy，which reduced the computational complexity and enhanced the classification and localization accuracy at the same time. Finally，a pruning algorithm was used to prune the model parameter redundancy to further compress the model, and the knowledge distillation technique was used to achieve the compensation and optimization of the accuracy. Through experiments， it is found that compared with YOLOv8n，YOLOv8-LMD accuracy was improved by 1.7%，the computational amount was reduced by 68%，the parameter amount was reduced by 77%，and the model size was reduced by 75%. Therefore，it can be concluded that the model proposed in this paper is more suitable for the task of rock detection on the surface of Mars.

Keywords：YOLOv8n；Mars surface detection；light weighting；channel pruning；knowledge distillation

Highlights：

- A lightweight backbone network was designed to achieve initial compression of model parameters.
- A multi-scale feature fusion network structure was designed to effectively improve the feature characterization of rock targets at different scales.
- A lightweight detection head was designed using a convolutional sharing strategy to effectively reduce computational complexity.
- Structured pruning was performed on the improved model to further compress the model.
- The accuracy of the pruned model was adjusted using a knowledge distillation algorithm to bring the model performance to the desired level.

[责任编辑：宋宏，英文审校：宋利辉]