工程科学学报
Chinese Journal of Engineering

基于稀疏注意力卷积ViT模型的锌浮选工况识别

苏越 唐朝晖 谢永芳 高小亮 张虎 马炜烨 汤海琀

Sparse attention convolution-ViT model for working condition recognition in zinc flotation

SU Yue, TANG Zhaohui, XIE Yongfang, GAO Xiaoliang, ZHANG Hu, MA Weiye, TANG Haiyang

引用本文:
苏越, 唐朝晖, 谢永芳, 高小亮, 张虎, 马炜烨, 汤海琀. 基于稀疏注意力卷积ViT模型的锌浮选工况识别[J]. 北科大：工程科学学报, 2025, 47(2): 328–338. doi: 10.13374/j.issn2095–9389.2024.05.13.004
SU Yue, TANG Zhaohui, XIE Yongfang, GAO Xiaoliang, ZHANG Hu, MA Weiye, TANG Haiyang. Sparse attention convolution–ViT model for working condition recognition in zinc flotation[J]. *Chinese Journal of Engineering*, 2025, 47(2): 328–338. doi: 10.13374/j.issn2095–9389.2024.05.13.004

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2024.05.13.004

工程科学学报, 第 47 卷, 第 2 期: 328−338, 2025 年 2 月
Chinese Journal of Engineering, Vol. 47, No. 2: 328−338, February 2025
https://doi.org/10.13374/j.issn2095-9389.2024.05.13.004; http://cje.ustb.edu.cn

基于稀疏注意力卷积 ViT 模型的锌浮选工况识别

苏　越[1)], 唐朝晖[1)], 谢永芳[1)], 高小亮[1)✉], 张　虎[1,2)], 马炜烨[1)], 汤海玥[1)]

1) 中南大学自动化学院, 长沙 410083　2) 中南大学数学与统计学院, 长沙 410083
✉通信作者, E-mail: gxiaoliang@csu.edu.cn

摘　要　准确识别锌浮选工况并用于指导锌浮选操作, 可以提高浮选效率、优化选矿过程. 目前浮选现场主要通过人工肉眼观察泡沫并依据经验判断工况, 这种方法主观性强, 难以客观准确地评价锌浮选工况. 针对该问题, 本文通过研究锌浮选泡沫视觉特征和浮选工况的密切联系, 提出基于稀疏注意力卷积 ViT 模型的锌浮选工况识别方法. 首先, 所提模型融合了卷积神经网络(Convolutional neural networks, CNN)和视觉 Transformer(Vision transformer, ViT)的结构和优点, 同时感知泡沫局部空间信息和全局信息, 完备表征泡沫图像. 其次, 模型引入稀疏的多头注意力机制, 每个注意力头以不同的稀疏程度处理特征, 从不同尺度下感知全局信息, 同时引入注意力门控单元优化特征传递, 最终实现锌浮选工况识别. 实验结果表明, 本文所提工况识别方法在锌浮选泡沫图像数据集上的准确率达到 88.62%, 解决了传统 CNN 和 ViT 模型不能充分利用泡沫图像全局信息, 且无法自适应捕捉泡沫图像重要特征的问题, 为浮选流程优化提供有力支持.

关键词　工况识别; 卷积神经网络; 视觉 Transformer; 稀疏注意力; 泡沫浮选
分类号　TP391.4

Sparse attention convolution-ViT model for working condition recognition in zinc flotation

SU Yue[1)], *TANG Zhaohui*[1)], *XIE Yongfang*[1)], *GAO Xiaoliang*[1)✉], *ZHANG Hu*[1,2)], *MA Weiye*[1)], *TANG Haiyang*[1)]

1) School of Automation, Central South University, Changsha 410083, China
2) School of Mathematics and Statistics, Central South University, Changsha 410083, China
✉Corresponding author, E-mail: gxiaoliang@csu.edu.cn

ABSTRACT　Accurate recognition of working conditions can optimize the zinc flotation process and improve its efficiency. Traditionally, this recognition heavily relies on manual observations of froth appearance, a method prone to human error and subjective judgment. To address this issue and improve recognition accuracy, a sparse attention convolution-ViT model is proposed. This model leverages machine vision techniques to investigate the relationship between froth visual features and the working conditions using real-time froth images from industrial sites. The model aims to recognize zinc flotation working conditions in real time, thereby providing guidance for operations. First, it combines the strengths of convolutional neural networks (CNNs) and vision transformers (ViT) to effectively extract both local and global features from froth images. Specifically, CNNs are adept at capturing local features, such as texture, color, and fine details of the froth, while ViT excels at identifying global features, such as the froth size distribution. By combining these two architectures, the sparse attention convolution-ViT model comprehensively analyzes the froth images. To enhance the global feature processing of froth images, a sparse multi-head attention mechanism is introduced into the ViT component. This mechanism allows the model to process global features with different sparsity levels, reducing computational costs and improving the model's adaptability to different froth appearances. Each attention head in the sparse multi-head attention mechanism targets different

收稿日期: 2024−05−13
基金项目: 国家自然科学基金资助项目(62171476, 62376040, 62233018); 湖南省自然科学基金资助项目(2023JJ30084)

aspects of global features, allowing the model to extract various information from the froth images while maintaining efficiency. Furthermore, an attention gated unit is introduced to refine the feature processing. This unit allows adaptive weighting of extracted features in the image, enhancing model interpretability and optimizing feature transfer. By effectively capturing the relevant features, the attention-gated unit helps the model to focus on critical features of the froth images that can indicate the working conditions. Experimental results demonstrated the effectiveness of the proposed sparse attention convolution-ViT model in recognizing zinc flotation working conditions. The model achieved a recognition accuracy of 88.62% on the zinc flotation froth image dataset, surpassing traditional CNN and ViT models. Ablation experiments highlighted the critical role of the sparse multi-head attention mechanism and the attention-gated unit, contributing to accuracy improvements of 0.92% and 2.63%, respectively. Moreover, gradient-weighted class activation mapping was used to visualize feature weights, confirming the model's capability to effectively characterize froth images by identifying both local and global features. This accurate recognition of zinc flotation conditions underscores the potential of the model in providing reliable real-time recognition, supporting the optimization of the flotation process, thereby improving efficiency and resource utilization in zinc flotation.

KEY WORDS　working condition recognition; convolutional neural networks; vision transformer; sparse attention; froth flotation

浮选是一项从矿石中提取有用矿物的重要技术，通过向矿浆中注入空气和化学药剂并不断搅拌，使矿物颗粒附着在产生的气泡上并随之从矿浆浮出形成泡沫层，用刮板收集泡沫实现矿物富集. 经过多个粗选、精选等流程后，得到矿物含量较高的泡沫产品. 浮选工况直接影响浮选最终浮选效果，对整个浮选流程中有重要意义. 由于浮选槽泡沫的表层视觉特征与浮选工况存在密切联系，能够反应浮选效果，可以通过人工观察泡沫的方式判断浮选工况，这种方法主观性强，难以实现客观的评价与认知，可能造成回收率下降、药剂浪费等问题[1]. 因此，引入机器视觉，将实时采集的锌浮选泡沫图像作为输入，利用一定方法提取泡沫视觉特征，建立不同特征和工况之间的关系，准确、快速地识别工况，为实现锌浮选过程安稳运行及优化奠定基础.

锌浮选流程中，泡沫的颜色、尺寸、形状、稳定性、流速、纹理等表层视觉特征与浮选工况密切相关[2−4]. 在传统方法中，研究者建立这些泡沫表层视觉特征和浮选工况的联系，用于评价浮选工况并指导现场工人操作. 桂卫华等[5] 基于泡沫表面颜色共生矩阵特征，提出一种新的纹理特征提取方法，为锌浮选优化控制提供依据. 刘金平等[6] 利用泡沫大小动态分布评价锌浮选状况. 罗金等[7] 结合图像局部二值模式和灰度共生矩阵，提出一种泡沫动态纹理特征，并在锌浮选工况识别中验证了其有效性和稳定性. 这些泡沫表层视觉特征能在一定程度反映浮选工况，但是存在不能完备、细微地表征泡沫图像的问题.

近年来，随着机器视觉领域快速发展，研究者们尝试利用卷积神经网络（Convolutional neural networks, CNN）模型分析泡沫图像并识别锌浮选工况. 廖一鹏等[8] 设计双通道 CNN 提取并融合泡沫可见光和红外图像特征，用于浮选工况识别. 范影等[9] 针对时间维度上的泡沫尺寸分布特征，设计基于神经网络的浮选工况识别模型. 高小亮等[10] 结合泡沫表层视觉特征和深度特征，提出一种分层感知方法，用于识别锌浮选工况，并提出一种基于双规则库的加权模糊插值推理模块和时间编码器–解码器贝叶斯网络的试剂添加控制策略，实现锌浮选药剂调节[11]. 田灿等[12] 使用双向长短期记忆网络提取时序的泡沫尺寸分布统计特征，用于识别锌浮选工况. 张虎等[13] 提出一种多任务学习策略，集成编码器–解码器模型和 Siamese 时间序列的差分网络，同时监测锌浮选的精矿和尾矿品位. 唐朝晖等[14] 提出一种半监督记忆网络，利用一种混合无监督学习策略，使用未标记的数据进行训练，用于监测锌浮选工况. 因为卷积计算方法的特点，以上基于 CNN 模型的工况识别方法具有归纳偏置特性，能够较好地反映出图像中局部空间信息. 但是，图像中相隔较远的像素受到卷积核大小限制，无法在前几次卷积操作中交互，导致 CNN 不擅长从全局角度处理泡沫尺寸分布等全局信息[15]，在表征泡沫图像时在存在一定局限性.

相较之下，Transformer 模型没有使用卷积操作，而是利用自注意力机制处理信息，能够较好地捕捉全局特征[16]. 后来 Transformer 模型的编码器部分被用于图像识别，成为视觉 Transformer 模型（Vision transformer, ViT），证明了其在机器视觉领域的实用性[17]. 随后，有大量研究使用 ViT 模型完成机器视觉任务，Chen[18] 结合不同大小的图像块（token）及其融合模块，提出一种双分支 ViT 模型

(CrossViT), 用于图像分类任务. 刘泽等[19]提出一种分层的 Transformer 结构, 使用移动窗口对图像进行表征并允许跨窗口连接, 用于包括图像分类、目标检测和语义分割等机器视觉任务. 杨锐等[20]在计算注意力时使用两个因子缩放查询、键和值矩阵维度, 泛化上下文导向, 增强对象敏感性, 完成图像分类任务. 梁礼明等[21]结合 Swin Transformer 和图形推理, 提出了一种用于结直肠图像中息肉的分割方法. 杨本臣等[22]设计了一种基于 Swin Transformer 的双尺度网络, 交互图像局部与全局特征, 准确分割视网膜图像中的血管. ViT 模型能够很好地捕捉图像的全局信息, 但缺乏类似 CNN 的归纳偏置能力, 难以感知图像局部空间信息[23−25], 且注意力计算方式单一, 模型泛化能力不足[26]. 同时, 由于 ViT 模型中的前馈网络(FFN, Feed-forward network)结构简单, 导致模型不能较好地判断不同特征的重要性, 自适应地从图像特征中提取更重要的信息.

在锌浮选中, 泡沫纹理、大小、形状等局部空间信息, 以及泡沫尺寸分布等全局信息与浮选工况密切相关[27]. 然而, CNN 和 ViT 都不能较好地单独处理这两类信息, 因此提出一种融合 CNN 和 ViT 的模型, 完整感知泡沫局部空间信息和全局信息. 但是, 传统 ViT 模型的多头注意力机制中, 每个注意力头都会同时考虑全部信息, 这可能会加重计算负担并限制模型充分利用全局信息. 因此引入稀疏的注意力机制, 在每个注意力头以不同的稀疏程度分析数据, 从不同尺度处理泡沫尺寸分布等全局信息, 提高模型在不同任务上的泛化能力. 同时, 传统 ViT 的前馈网络结构简单, 难以自适应处理不同特征的重要性. 例如, 对于严重过浮选和轻微过浮选两个工况, 两个工况下的泡沫图像都偏暗, 此时仅用颜色特征并不能很好地区分两个工况, 需要用其他比如纹理、泡沫尺寸等特征区分, 需要为这些特征分配不同权重. 因此, 引入注意力门控单元, 自适应优化泡沫图像特征传递, 更好地捕捉重要特征.

综上所述, 本文提出一种基于稀疏注意力卷积 ViT 模型的锌浮选工况识别方法, 主要贡献如下:(1)针对 CNN 和 ViT 无法单独处理泡沫局部空间信息和全局信息的问题, 提出一种融合 ViT 和 CNN 结构的模型, 同时提取泡沫局部空间信息和全局信息, 更准确地识别锌浮选工况. (2)针对传统 ViT 注意力计算方式单一, 难以完整反映泡沫全局信息的问题, 引入稀疏的多头注意力机制, 每个注意力头以不同稀疏程度分析数据, 从不同尺度处理信息, 充分利用泡沫图像全局信息, 获取更完备的泡沫图像特征. (3)针对传统 ViT 的前馈网络不能自适应调节不同特征重要性的问题, 提出注意力门控单元, 优化泡沫图像特征传递, 捕捉泡沫重要特征. (4)对提出的模型进行充分实验验证. 进行消融实验, 证明上述各个方法的必要性, 同时证明所提模型在识别锌浮选工况准确率上高于 CNN 和 ViT 模型, 并使用 Grad-CAM(Gradient-weighted class activation mapping)可视化模型决策权重, 说明模型提取特征的特点.

1　锌浮选回路及锌浮选工况分类

如图 1 所示, 铅浮选回路尾矿作为锌浮选的原始流入矿浆, 进入锌浮选回路进行浮选. 锌浮选回路包括锌快粗选、粗选(Ⅰ和Ⅱ)、扫选(Ⅰ、Ⅱ和Ⅲ)以及精选(Ⅰ、Ⅱ和Ⅲ)四道工序, 各道工序之间通过泡沫和底流相互联系. 矿浆首先经过锌快粗选工序, 经过浮选后得到富含锌矿粒子的泡沫产品并送入锌精选工序, 经三次精选后得到最终的锌精矿产品. 锌快粗选底流则送入锌粗选槽, 经锌粗选、锌扫选工序处理后, 从锌扫选底流处获得锌尾矿产品, 同时将锌粗选工序的泡沫产品送入铅锌混合浮选回路作进一步浮选. 作为锌浮选的第一道工序, 锌快粗选将直接影响浮选质量, 在整个工艺流程中有重要意义. 锌浮选回路入矿矿浆正常时, 若锌快粗选槽底流品位偏高, 则代表泡沫产品的锌含量减少, 将导致锌精矿产品质量和浮选回收率降低, 而底流品位越低, 则代表精矿产品质量和浮选回收率越高. 因此, 可以通过监控锌快粗选状态来评估锌浮选工况.

采用 X-ray 荧光分析仪监测锌快粗选底流品位, 同时通过安置在浮选槽上方的数据采集装置获取锌快粗选槽表层泡沫图像, 监控锌快粗选工序. 如图 2 所示, 数据采集装置由防护罩、工业相机以及固定光源组成. 其中, 工业相机负责实时拍摄泡沫图像, 固定光源提供稳定的光照, 防护罩用于保护工业相机和固定光源, 同时构建一个独立的环境, 减少外部环境对图像的影响, 确保拍摄质量和可靠性. 工业相机位于泡沫层上方垂直距离约 1 m 处, 采集的图像分辨率为 690×516, 监测浮选槽表面泡沫层约 40 cm×30 cm 的矩形区域.

在不同浮选工况下, 锌快粗选槽表层泡沫呈现出多样化的视觉特征, 这些特征直接反映当前浮选工况. 依据荧光分析仪获得的锌快粗选槽底

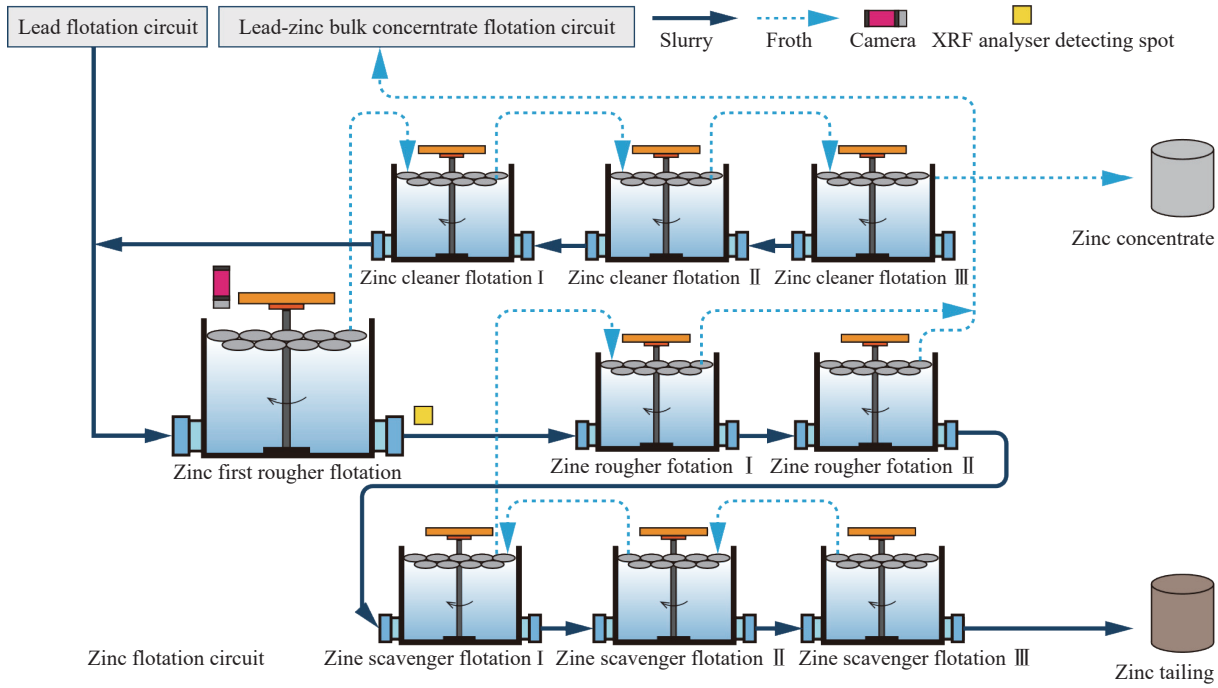


图 1　锌浮选回路

Fig.1　Zinc flotation circuit

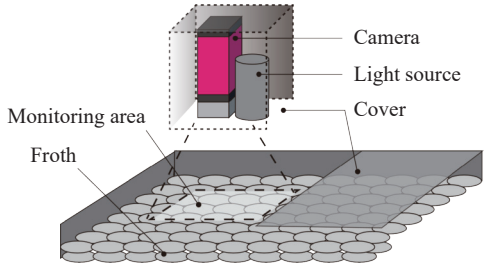


图 2　锌快粗选槽数据采集装置

Fig.2　Data acquisition device of the zinc first rougher tank

流品位，将锌浮选工况划分为严重欠浮选、轻微欠浮选、正常、轻微过浮选和严重过浮选（Ⅰ类、Ⅱ类、Ⅲ类、Ⅳ类和Ⅴ类）五类，作为预测的输出. 这种分类方法有助于更准确地评估锌浮选工况，及时采取相应的措施进行调整和优化，以保证生产过程的稳定性和效率. 五种工况下的泡沫图像如表 1 所示.

2　基于稀疏注意力卷积 ViT 模型的锌浮选工况识别

为准确识别锌浮选工况，指导现场操作，提升矿产资源利用率，提出一种基于稀疏注意力卷积 ViT 模型的锌浮选工况识别方法. 如图 3 所示，稀疏注意力卷积 ViT 模型由多个组件构成，包括 CNN 模块、稀疏注意力 ViT 模块、卷积层（Conv-3×3）和全连接层（Fully connected layer, FC）. 这些组件共同协作，对输入泡沫图像进行处理，准确识别锌浮选工况，稀疏注意力卷积 ViT 模型主要内容包括以下三点.

（1）融合 CNN 和 ViT 的模型结构: 针对 CNN 和 ViT 无法单独完整处理泡沫局部空间信息和全局信息的问题，提出一种融合 CNN 和 ViT 的网络结构. 该网络结合两者优势，使模型能够充分处理泡沫纹理、尺寸、形状等局部空间信息及泡沫尺寸分布等全局信息.

（2）稀疏的多头注意力: 针对 ViT 注意力计算方式单一，不能充分利用图像全局信息的问题，引入一种稀疏的多头注意力机制，每个注意力头以不同的稀疏程度分析输入特征，减少模型计算负担，同时允许模型提取更泛化的泡沫特征.

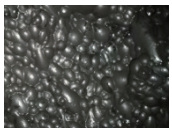
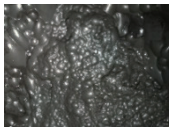
（3）注意力门控单元: 传统 Transformer 的 FFN 没有考虑到不同特征重要性的问题. 为此，引入注意力门控单元，使模型自适应调节注意力输出，更有效地捕捉信息，优化信息传递.

2.1　基于 CNN 模块的图像局部空间信息处理

在稀疏注意力卷积 ViT 模型中，首先利用 CNN 模块处理泡沫图像局部空间信息. 具体来说，CNN 在图像上滑动卷积核时，只关注卷积核的局部像素区域，提取出如泡沫的边缘、纹理和结构等局部空间信息. 这种操作的优势在于权重共享，即使用相同卷积核可以在图像中不同位置感知相似特征. 这种局部感知性使得 CNN 在捕捉细节信息方面表现出色，有助于进一步分析泡沫图像. 在 CNN

表 1　五种锌浮选工况对应的锌快粗选槽泡沫视觉特征及底流品位

Table 1　Visual features of the froth and the slurry grade in the zinc quick rougher flotation cell corresponding to five different zinc flotation working conditions

Froth image	Visual features	Slurry grade	Working condition
	Large size, light color, coarse texture, smooth surface	≤1	Ⅰ
	Relatively large size, lighter color, coarser texture	1–1.25	Ⅱ
	Medium size, uniform distribution, overall stable appearance	1.25–1.5	Ⅲ
	Small size, accumulation present, darker color	1.5–1.75	Ⅳ
	Smaller size, finer texture, with many creased and muddied areas	≥1.75	Ⅴ

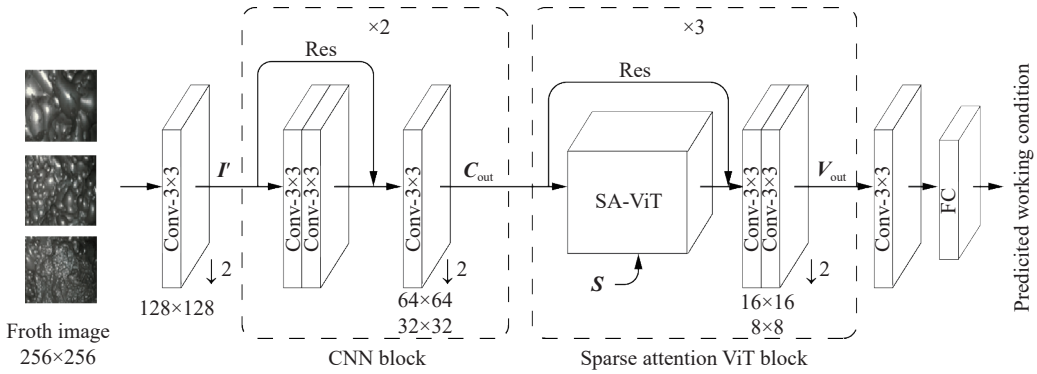


图 3　基于稀疏注意力卷积 ViT 模型的锌浮选工况识别

Fig.3　Sparse attention convolution-ViT model for working condition recognition in zinc flotation

模块的作用下, 输入泡沫图像通过一系列卷积层和其他非线性激活函数处理, 被转换为深度特征图像. 这些深度特征图像中每个像素都包含对应泡沫图像各处的局部空间信息. 如图 3 所示, 输入图像经过一次卷积操作后, 输入 CNN 模块, 具体运算过程如下:

$$\boldsymbol{I}' = \underset{\text{stride}=1}{\text{Conv}}(\boldsymbol{I}) \tag{1}$$

$$\boldsymbol{C}_{\text{out}} = [\underset{\text{stride}=2}{\text{Conv}}([\underset{\text{stride}=1}{\text{Conv}}(\boldsymbol{I}')]^{\times 2} + \boldsymbol{I}')]^{\times 2} \tag{2}$$

其中, Conv(·)代表卷积计算[28], 卷积核的大小为 3×3; stride 代表卷积核移动步长, $\boldsymbol{I} \in \mathbf{R}^{3\times 256\times 256}$代表输入 256×256 像素大小的 3 通道泡沫图像; $\boldsymbol{I}' \in \mathbf{R}^{c\times s\times s}$代表进行一次卷积后输入 CNN 模块的特征图像, c、s分别代表图像通道数和尺寸; $[\cdot]^{\times n}$代表重复进行 n 次括号内的运算; $\boldsymbol{C}_{\text{out}}$ 代表 CNN 模块的最终输出.

2.2　基于稀疏注意力 ViT 模块的全局信息处理

ViT 是一种基于 Transformer 编码器架构的图像处理模型, 将图像视为一系列像素块(patch), 然后将其转换为序列数据输入到 Transformer 中进行处理. 通过自注意力机制, ViT 能够处理图像的全局信息, 从整体上理解、分析图像. 但是, 在处理泡沫图像时, 传统 ViT 仍然存在不能充分利用全局信息的问题. 由于传统 ViT 中每个注意力头都会同时考虑全部信息, 这样会失去例如局部图像中

的泡沫尺寸分布等信息. 因此, 其次, 传统 ViT 模型中的前馈神经网络通常使用结构单一的全连接层对特征进行非线性变换, 并没有自适应地考虑不同特征的重要性, 从而可能导致模型难以优先处理关键特征. 例如, 在正常工况下的泡沫图像中, 有时会出现小范围的泡沫堆积, 模型可能会过度关注这一特征, 忽略其他如尺寸、纹理等特征, 错误地将工况识别为过浮选. 针对这些问题, 采用稀疏的多头注意力, 充分表征泡沫全局信息进行, 提升模型泛化能力; 引入注意力门控单元, 自适应地控制泡沫特征传递, 优化计算过程. 如图 4 所示, 稀疏注意力 ViT 模块的具体计算过程如下:

$$\boldsymbol{V}_{\text{out}} = [\underset{\text{stride}=2}{\text{Conv}}(\underset{\text{stride}=1}{\text{Conv}}(\text{SAViT}(\boldsymbol{C}_{\text{out}}, \boldsymbol{S}) + \boldsymbol{C}_{\text{out}}))]^{\times 3} \quad (3)$$

其中, $\boldsymbol{V}_{\text{out}}$代表稀疏注意力 ViT 块输出, SAViT(·)代表稀疏注意力 ViT(Sparse attention ViT, SA-ViT)计算, $\boldsymbol{S}$ 代表用于稀疏注意力计算的稀疏矩阵.

如图 4 所示, 在稀疏注意力 ViT 中, 首先通过卷积操作对输入特征图像进行相对位置编码[29], 再将结果分割、重组, 得到若干个 token 用于计算注意力, 经 L 个注意力层处理后得到输出 token, 然后通过变换、重组 token 得到新的特征图像, 最后进行一次卷积操作, 得到最终输出特征图像. 每一个注意力层都包含了稀疏的多头注意力和注意力门控单元.

2.2.1　稀疏的多头注意力

为解决传统 ViT 存在处理泡沫尺寸分布等全局信息不充分的问题, 如图 4 所示, 引入稀疏的多头注意力. 每个注意力头依据不同稀疏程度分析数据, 有利于模型充分处理泡沫全局信息, 简化模型计算量并获取更加泛化的泡沫图像特征.

在计算多头注意力时, 模块将输入的多个 token 映射成 h 个不同的查询、键和值($\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$), 用于稀疏注意力 ViT 计算. 稀疏的多头注意力具体计算过程如下:

$$\boldsymbol{S} = (\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_h) \quad (4)$$

$$\begin{aligned} \text{head}_i &= \text{SparseAttention}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) \\ &= \text{Softmax}\left(\boldsymbol{S}_i \odot \frac{\boldsymbol{Q}_i \boldsymbol{K}_i^{\text{T}}}{\sqrt{d_{\text{k}}}}\right)\boldsymbol{V}_i \end{aligned} \quad (5)$$

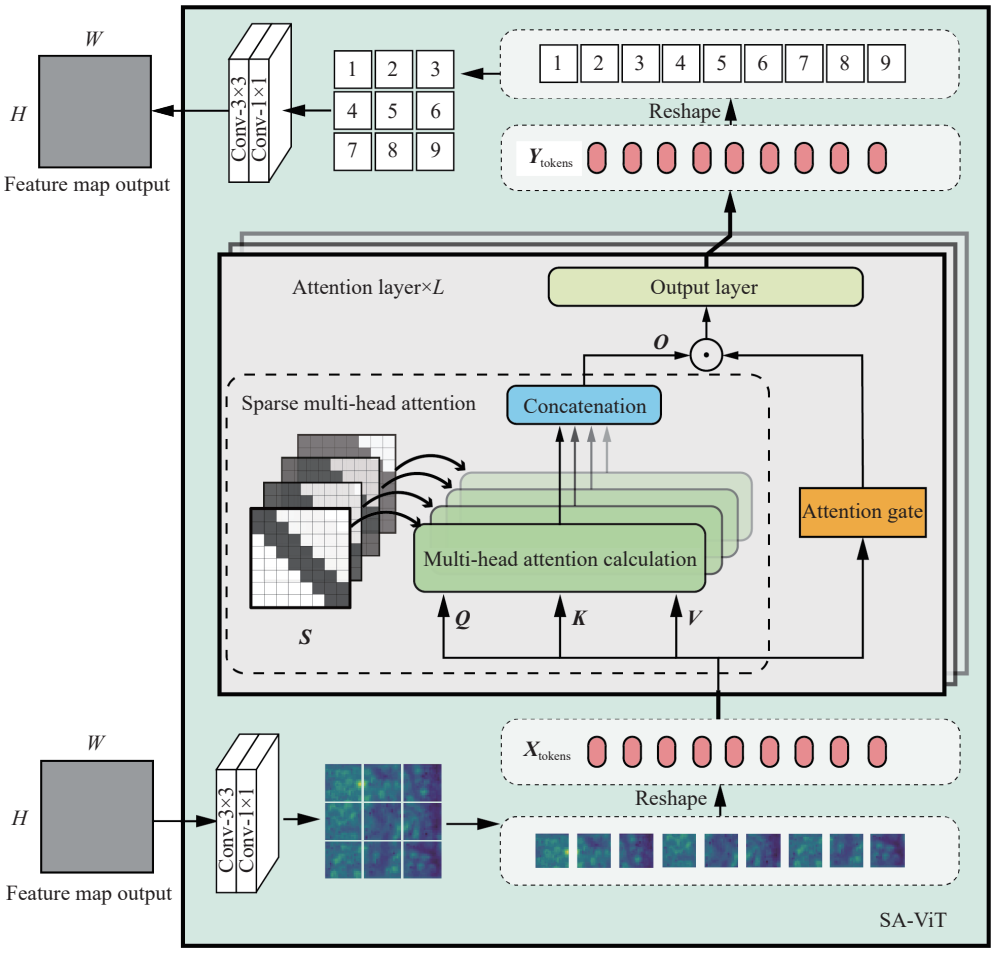


图 4　稀疏注意力 ViT 结构

Fig.4　Structure of Sparse Attention ViT

$$\boldsymbol{O}=\mathrm{Concat}(\mathrm{head}_1,\mathrm{head}_2,\cdots,\mathrm{head}_h) \tag{6}$$

其中, $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbf{R}^{h\times N\times d_k}$, $d_k=d/h$, h 表示多头注意力中头的数量, N、d 分别表示 token 的个数和每个 token 的特征维度. $\boldsymbol{S}\in\mathbf{R}^{h\times N\times N}$ 分别由不同稀疏程度的稀疏矩阵 $\boldsymbol{S}_1,\boldsymbol{S}_2,\cdots,\boldsymbol{S}_h\in\mathbf{R}^{N\times N}$ 组成, Softmax(·)表示归一化操作, ⊙表示两个相同尺寸的矩阵按照对应位置作乘法运算, head_i 代表每个注意力头的输出结果, Concat(·)表示矩阵拼接操作, $\boldsymbol{O}\in\mathbf{R}^{N\times d}$ 表示稀疏的多头注意力输出.

2.2.2　注意力门控单元

在传统 ViT 前馈网络(FFN)中, 全连接层没有自适应考虑不同特征之间重要性, 为此, 提出如图 4 所示的注意力门控单元. 注意力门控单元对输入特征做一次可学习的变换, 所得结果作为门控矩阵, 控制注意力计算结果输出. 这样, 使得模型能够更加灵活地控制特征的传递和调节, 自适应地权衡输入特征的重要性, 有助于模型更好地捕捉到重要信息, 提高整体性能. 具体计算过程如下:

$$W_{\mathrm{u}}(\boldsymbol{X}_{\mathrm{tokens}})=\mathrm{ReLu}(\boldsymbol{X}_{\mathrm{tokens}}\boldsymbol{U}) \tag{7}$$

$$\boldsymbol{Y}_{\mathrm{tokens}}=\boldsymbol{O}\odot W_{\mathrm{u}}(\boldsymbol{X}_{\mathrm{tokens}}) \tag{8}$$

其中, $\boldsymbol{X}_{\mathrm{tokens}},\boldsymbol{Y}_{\mathrm{tokens}}\in\mathbf{R}^{N\times d}$ 分别表示整个注意力计算的输入和输出 token, $\boldsymbol{O}\in\mathbf{R}^{N\times d}$ 表示稀疏的多头注意力输出, $W_{\mathrm{u}}(\cdot)$ 表示对输入进行一次可学习的映射变换, ReLu(·)表示 ReLu 激活函数, $\boldsymbol{U}\in\mathbf{R}^{d\times d}$ 为一个可学习的矩阵.

2.3　基于全连接层的锌浮选工况类别输出

模型采用全连接层作为输出层, 将先前提取和处理得到的高维特征映射到最终五种工况类别: 正常、轻微欠浮选、轻微过浮选、严重欠浮选和严重过浮选. 全连接层将所有输入神经元与输出神经元连接, 每个输出神经元对应一个类别, 通过学习合适的权重和偏置参数, 将输入特征空间映射到对应类别空间. 模型经过训练后, 全连接层将能够根据输入特征向量, 准确预测泡沫图像所属工况类别. 具体计算过程如下:

$$\mathrm{WC}_{\mathrm{pred}}=\mathrm{FC}(\underset{\mathrm{stride}=1}{\mathrm{Conv}}(\boldsymbol{V}_{\mathrm{out}})) \tag{9}$$

其中, $\mathrm{WC}_{\mathrm{pred}}\in\mathbf{R}^{1\times5}$ 代表最终预测的 5 类工况, $\boldsymbol{V}_{\mathrm{out}}$ 代表稀疏注意力 ViT 块输出, FC 代表全连接层运算[28].

3　实验结果和分析

3.1　实验设置

为验证模型有效性, 使用锌浮选泡沫图像数据集进行实验. 图像数据集来源于广东某浮选厂 2020 年 9 月 1 日到 9 月 28 日的锌快粗选槽数据采集装置, 共包含 8186 张大小为 690×516 的 RGB 泡沫图像, 其宽和高分别为 690 和 516 像素. 使用双三次插值的方法将图像的大小缩放为 256×256, 这种方法能够保留原图像中的纹理、轮廓和其他重要特征, 使模型更容易捕捉图像中的重要特征. 训练集、测试集和验证集, 三者数量分别为 4546、1820、1820 张, 其中测试集和验证集的数量稍高, 这种划分比例能够提供足够多的样本用于调节模型超参数、减少过拟合风险, 同时在评估模型时提供更加稳定可靠的结果. 根据 X-ray 荧光分析仪采集的锌快粗槽底流品位, 确定泡沫图像标签, 分为正常、轻微欠浮选、轻微过浮选、严重欠浮选和严重过浮选五类. 其中正常、轻微过浮选、轻微欠浮选、严重过浮选和严重欠浮选图像分别为 1772、1672、1573、1649 和 1511 张, 选择浮选工况预测准确率作为模型评价指标.

使用 Adam 优化器训练模型, 实验环境为 CUDA11.8, python3.9.16 版本, 所有计算在 Intel(R) Core (TM) i5-12400F 和 NVIDIA GeForceRTX3060Ti 上完成. 初始学习率设置为 0.001, 选取 ViT 注意力头数 h=4, 使用如图 5 的滑动窗口的稀疏注意力, 能够有效保留图像中不同尺度下的局部信息, 各个不同稀疏程度的稀疏矩阵 $\boldsymbol{S}_i$ 见图 5.

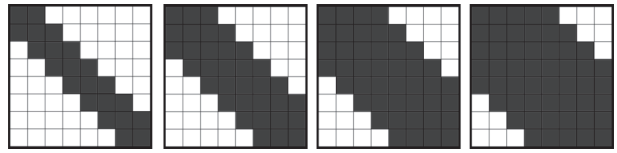


图 5　不同稀疏程度的滑动窗口注意力

Fig.5　Sliding window attention with varying degrees of sparsity

在计算稀疏注意力时, 让序列中各个 token 仅与相邻阴影部分位置的 token 计算注意力. 稀疏注意力卷积 ViT 模型参数如表 2 所示.

3.2　训练过程

为了验证模型的泛化能力、防止模型过拟合, 我们添加了训练集与测试集在训练过程中的损失对比, 训练过程中网络在训练集和测试集在约 200 个训练步长内的损失变化如图 6 所示. 可以看到, 模型对训练集和测试集的损失随着训练逐步下降, 并趋于稳定, 在 150 个训练步长左右均达到稳定, 且二者的损失十分接近, 表明模型具有较好的泛化性. 同时, 使用早停法训练模型, 当损失在一定步长内不再下降时停止训练. 综上所述, 可以

表 2　稀疏注意力卷积 ViT 模型参数

Table 2　Parameters of the sparse attention convolution-ViT model

Layer name	Output size	Output channels	Number
Conv-3×3(stride=2)	128×128	16	1
Conv-3×3(stride=1)	128×128	32	2
Conv-3×3(stride=2)	64×64	48	1
Conv-3×3(stride=1)	64×64	48	2
Conv-3×3(stride=2)	32×32	64	1
SA-ViT	32×32	64	1
Conv-3×3(stride=1)	32×32	64	1
Conv-3×3(stride=2)	16×16	80	1
SA-ViT	16×16	80	1
Conv-3×3(stride=1)	16×16	80	1
Conv-3×3(stride=2)	8×8	96	1
SA-ViT	8×8	96	1
Conv-3×3(stride=1)	8×8	384	1
FC(Fully connected layers)	1	[32, 5]	1

认为模型不存在过拟合现象.

3.3　参数选择

注意力层堆叠数量影响模型的工况识别准确率和复杂程度, 是稀疏注意力卷积 ViT 模型中的重要参数. 采用网格搜索法调整注意力模块中注意力层堆叠数量 L, 以模型工况识别准确率和参数量作为评价标准, 参数 L 的选择范围设定为 {5,6,7,8,9,10}, 网格搜索法通过遍历这一参数集合, 对每一个可能的 L 值训练模型, 并在验证集上评估其准确率和参数量. 不同注意力层数量对应模型准确率和参数量如表 3 和图 7 所示, 模型准确率随 L 增加而提升, 使模型的工况识别更为准确. 然而, 当注意力层堆叠数量 L 超过 9 时, 虽然模型参数量持续增加, 但准确率提升幅度却相对较小. 因此, 选择将注意力层数量设置为 9, 这样既能控制模型参数量增长, 又能保持较高准确率.

3.4　消融实验

为了验证模型中每个部分的有效性, 进行消融实验, 验证稀疏注意力和注意力门控单元的必要性. 建立四个模型, 表示为“Baseline”、“Baseline+稀疏注意力”、“Baseline+注意力门控”和“Baseline+稀疏注意力+注意力门控”. 其中, Baseline 为图 2 所示的卷积 ViT 模型, 但没有使用稀疏注意力和门控注意力, 其他模型表示相对 Baseline 添加了对应机制. 实验结果表明, 逐步引入不同模块可以逐步提升模型准确性. 在处理泡沫图像时, 稀疏的多头注意力机制有利于模型充分利用泡沫图像全局信息, 获取更加泛化的泡沫图像特征, 在工况识别准确率上提升了 0.92%; 而注意力门控能够自适应调节输出, 更有效地捕捉信息并优化信息传递, 提升了 2.63% 的识别准确率. 在同时引入稀疏注意力和注意力门控的情况下, 网络性能提升最为显著, 达到约 4.25% 的准确性提升. 实验结果如表 4 所示.

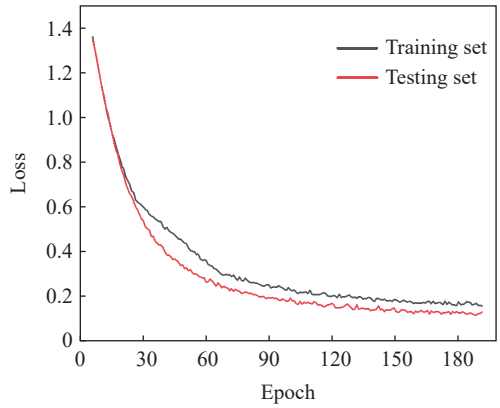


图 6　训练过程中模型对训练集和测试集的损失

Fig.6　Model loss on the training and test datasets during training

表 3　不同注意力层堆叠数量 L 对应的准确率和参数量

Table 3　Accuracy and number of parameters corresponding to different numbers of attention layers L

L	Accuracy/%	Number of parameters/10^6
5	82.28	1.757
6	85.55	1.963
7	87.43	2.169
8	88.36	2.375
9	88.62	2.581
10	88.64	2.788

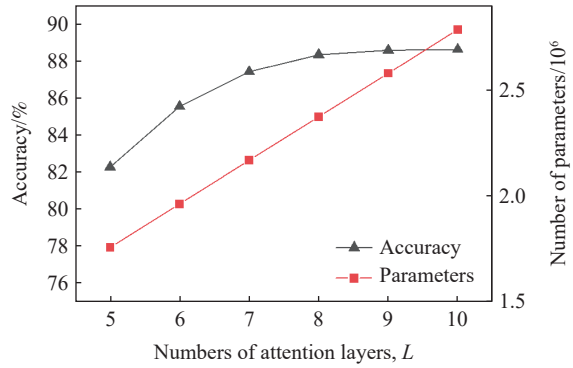


图 7　不同注意力层堆叠数量 L 对应的准确率和参数量

Fig.7　Accuracy and number of parameters corresponding to different numbers of attention layers L

3.5　对比实验

由于本文所提模型结合了 CNN 和 ViT 模型,

表 4　稀疏注意力卷积 ViT 模型的消融实验

Table 4　Ablation study of the sparse attention convolution-ViT model

Model	Accuracy/%	Δ↑/%
Baseline	84.35	—
Baseline + Sparse multi-head attention	85.27	0.92
Baseline + Attention gate	86.98	2.63
Baseline + Sparse multi-head attention + Attention gate	88.62	4.25

因此选择 MobileNetv1、MobileNetv2 MobileNetv3、ESPNetv2、EfficientNet-B0、EfficientNet-B1 和 LightViT 模型作为对比模型, 这些模型是一些和所提模型参数量相似的 CNN 和 ViT 模型. 用这些模型和稀疏注意力卷积 ViT 模型在工况识别准确率和参数量上进行对比. 实验结果表明, 本文所提出的稀疏注意力卷积 ViT 模型的准确率达到 88.62%(表 5), 高于对比模型, 能够更准确地识别锌浮选工况, 并且拥有较小的参数量, 更易于浮选现场部署.

利用混淆矩阵反映不同模型识别工况的具体情况, 如图 8 所示, 选取 MobileNetv3、ESPNetv2、LightViT、EfficientNet-B1 和稀疏注意力卷积 ViT 四个不同模型进行预测实验, 从结果来看, 在 1820 张泡沫图像中, 本文所提模型错误识别 207 张泡沫图像, 正确识别其余 1613 张; ESPNetv2 错误识别 266 张泡沫图像, 正确识别其余 1554 张; LightViT 错误识别 271 张泡沫图像, 正确识别其余 1549 张; MobileNetv3 错误识别 277 张泡沫图像, 正确识别其余 1543 张; EfficientNet-B1 错误识别 251 张泡沫

表 5　稀疏注意力卷积 ViT 模型与其他模型对比

Table 5　Comparison between sparse attention convolution ViT model and other models

Model	Accuracy/%	Number of parameters/10^6
MobileNetv1	82.27	4.2
MobileNetv2	84.85	3.4
MobileNetv3	84.78	2.9
ESPNetv2	85.38	3.5
EfficientNet-B0	85.88	5.3
EfficientNet-B1	86.21	7.8
LightViT	85.11	9.4
Ours	**88.62**	**2.6**

图像, 正确识别其余 1569 张. 从混淆矩阵来看, 五个模型在预测严重过浮选和严重欠浮选这两个工况时准确率极高, 这是由于这两个工况的泡沫与其他工况截然不同, 因此能够十分准确地识别这两个工况. 模型错误识别的工况集中在正常、轻微过浮选和轻微欠浮选这三种工况, 其中正常和轻微过浮选这两种工况的错误率较高, 原因是这两种工况下的泡沫十分相似, 区分难度大. 但是, 从总体上看, 本文所提的稀疏注意力卷积 ViT 模型具有更高的精度, 能够准确识别大部分工况, 准确率达到 88.62%, 较 MobileNetv3、ESPNetv2、LightViT 和 EfficientNet-B1 模型准确率分别提升 3.84%、3.24%、3.51% 和 2.41%, 误判率更低. 综上所述, 本文所提的稀疏注意力卷积 ViT 模型与这些 CNN 或 ViT 模

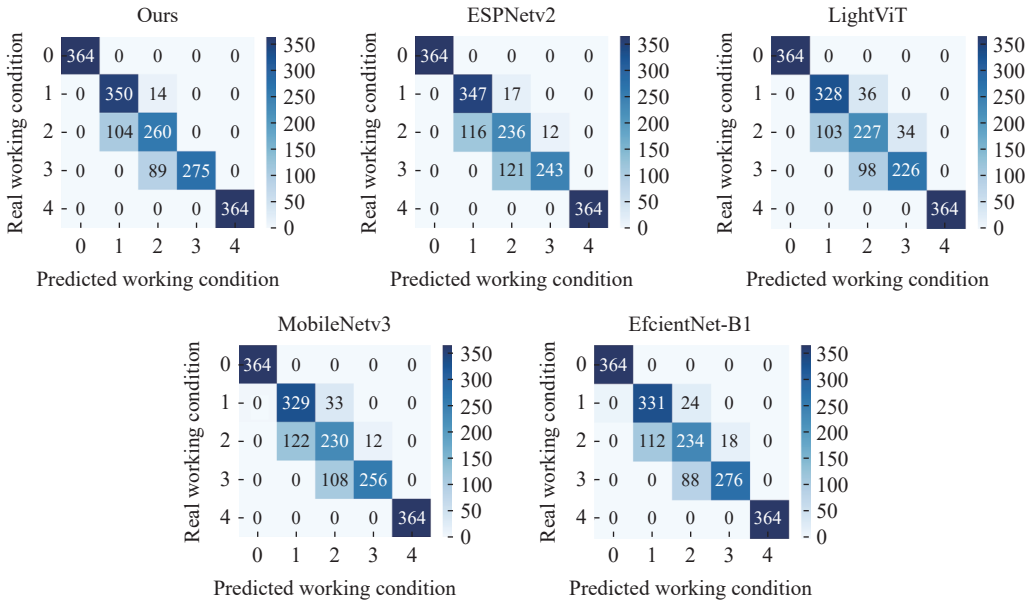


图 8　不同模型预测结果混淆矩阵

Fig.8　Confusion matrix of different model prediction results

型相比性能更好, 可以更好地识别锌浮选工况.

3.6　网络权重可视化

使用 Grad-CAM[30] 对网络关注的泡沫图像重点进行可视化. Grad-CAM 是一种用于解释深度学习模型预测的方法, 可以帮助理解模型的决策过程. Grad-CAM 通过计算特征图梯度来找出模型决策的关键区域, 生成热力图, 直观地展示模型在识别对象时最关注的区域.

对 MobileNetv3、ESPNetv2、LightViT 和本文所提稀疏注意力卷积 ViT 四个不同模型进行可视化计算, 如图 9 所示, 热力图中越红的地方代表模型对输入图像对应区域关注度和分类权重越高, 从红色到蓝色代表权重逐步减小. 结果表明, MobileNetv3 和 ESPNetv2 两个模型主要关注输入泡沫图像局部, 表明 CNN 模型主要关注泡沫图像的局部空间信息, 但一定程度上忽视了全局信息; 而 LightViT 模型具备较广泛的视野, 但是视野内部不细致, 表明 ViT 模型能够关注泡沫图像中的全局信息, 却没有细致地处理局部空间信息. 本文提出的稀疏注意力卷积 ViT 模型权重分布广泛且细致, 验证了模型同时感知泡沫图像局部空间信息和全局信息的能力.

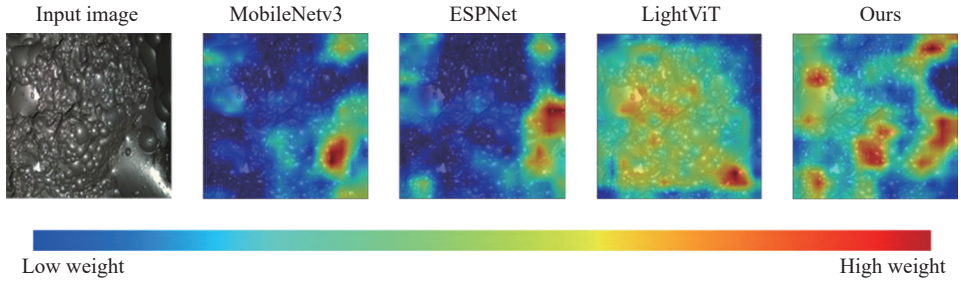


图 9　各个模型的 Grad-CAM 可视化结果

Fig.9　Grad-CAM visualization results for various models

4　结论

本文提出一种基于稀疏注意力卷积 ViT 模型的锌浮选工况识别方法. 经过实验验证, 得出以下结论:

(1) 模型融合了 CNN 和 ViT 模型结构, 同时处理泡沫图像中泡沫尺寸、纹理等局部空间信息和尺寸分布等全局信息, 更完备地表征泡沫图像, 准确识别锌浮选工况.

(2) 模型计算注意力时, 每个注意力头以不同的稀疏程度分析数据, 减少模型计算负担. 使用稀疏注意力后, 模型准确率提升了 0.92%, 说明模型能同时充分利用泡沫全局信息, 获取更泛化的特征.

(3) 模型引入注意力门控机制后, 准确率提升了 2.63%, 表明注意力门控机制看自适应调节注意力输出, 有效地捕捉重要信息.

(4) 最后, 在数据集上验证本文所提模型准确率, 并与其他参数相近的 CNN 和 ViT 分类模型在准确率和参数量上进行比较. 模型的准确率为 88.62%, 较 MobileNetv3、ESPNetv2、LightViT 和 EfficientNet-B1 模型准确率分别提升 3.84%、3.24%、3.51% 和 2.41%.

实验结果表明, 与模型参数量相似的 CNN 或 ViT 模型相比, 本文提出的模型具有更高准确率和更低的参数量, 能够更加准确地识别锌浮选工况, 有助于提升锌浮选最终产品质量, 提高整个锌浮选流程效率.

参　考　文　献

[1] Gui W H, Yang C H, Xu D G, et al. Machine-vision-based online measuring and controlling technologies for mineral flotation—a review. *Acta Autom Sin*, 2013, 39(11): 1879
(桂卫华, 阳春华, 徐德刚, 等. 基于机器视觉的矿物浮选过程监控技术研究进展. 自动化学报, 2013, 39(11): 1879)

[2] Liu J P, He J Z, Xie Y F, et al. Illumination-invariant flotation froth color measuring *via* Wasserstein distance-based CycleGAN with structure-preserving constraint. *IEEE Trans Cybern*, 2021, 51(2): 839

[3] Liu J P, Zhou J M, Tang Z H, et al. Toward flotation process operation-state identification *via* statistical modeling of biologically inspired Gabor filtering responses. *IEEE Trans Cybern*, 2020, 50(10): 4242

[4] Zhang H, Tang Z H, Xie Y F, et al. Long short-term memory-based grade monitoring in froth flotation using a froth video sequence. *Miner Eng*, 2021, 160: 106677

[5] Gui W H, Liao X, Yang C H, et al. A new texture extraction method for flotation froth images. *China Sci*, 2012, 7(4): 277
(桂卫华, 廖茜, 阳春华, 等. 一种新的浮选泡沫图像纹理特征提取方法. 中国科技论文, 2012, 7(4): 277)

[6] Liu J P, Gui W H, Tang Z H, et al. Dynamic bubble-size-

distribution-based health status analysis of reagent-addition in froth flotation process. *Contr Theory Appl*, 2013, 30(4): 492
(刘金平, 桂卫华, 唐朝晖, 等. 基于泡沫大小动态分布的浮选生产过程加药量健康状态分析. 控制理论与应用, 2013, 30(4): 492)

[7] Luo J, Tang Z H, Zhang H, et al. LTGH: A dynamic texture feature for working condition recognition in the froth flotation. *IEEE Trans Instrum Meas*, 2021, 70: 5008110

[8] Liao Y P, Zhang J, Wang Z G, et al. Flotation performance recognition based on dual-modality multiscale CNN features and adaptive deep learning KELM. *Opt Precis Eng*, 2020, 28(8): 1785
(廖一鹏, 张进, 王志刚, 等. 结合双模多尺度 CNN 特征及自适应深度 KELM 的浮选工况识别. 光学精密工程, 2020, 28(8): 1785)

[9] Fan Y, Guo Y Q, Tang Z H, et al. A dynamic size-based time series feature and application in identification of zinc flotation working conditions. *J Cent South Univ*, 2020, 27(9): 2696

[10] Gao X L, Tang Z H, Xie Y F, et al. A layered working condition perception integrating handcrafted with deep features for froth flotation. *Miner Eng*, 2021, 170: 107059

[11] Gao X L, Tang Z H, Xie Y F, et al. Dual-rule-based weighted fuzzy interpolative reasoning module and temporal encoder–decoder Bayesian network for reagent addition control. *IEEE Trans Fuzzy Syst*, 2024, 32(7): 3891

[12] Tian C, Tang Z H, Zhang H, et al. Operating condition recognition in zinc flotation using statistic and temporal correlation features. *IEEE Trans Instrum Meas*, 2022, 71: 2514412

[13] Zhang H, Tang Z H, Xie Y F, et al. ES-net: An integration model based on encoder–decoder and Siamese time series difference network for grade monitoring of zinc tailings and concentrate. *IEEE Trans Ind Electron*, 2023, 70(11): 11819

[14] Tang Z H, Zhang J, Xie Y F, et al. Semisupervised contrastive memory network for industrial process working condition monitoring. *IEEE Trans Instrum Meas*, 1070, 72: 5025110

[15] Peng Z L, Huang W, Gu S Z, et al. Conformer: Local features coupling global representations for visual recognition // 2021 *IEEE/CVF International Conference on Computer Vision*. Montreal, 2021: 357

[16] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inform Process Syst*, 2017, 30

[17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [J/OL]. *arXiv preprint* (2020-10-22) [2024-05-13]. https://arxiv.org/abs/2010.11929

[18] Chen C F R, Fan Q F, Panda R. CrossViT: Cross-attention multi-scale vision transformer for image classification // 2021 *IEEE/CVF International Conference on Computer Vision*. Montreal, 2021: 347

[19] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows // 2021 *IEEE/CVF International Conference on Computer Vision*. Montreal, 2021: 9992

[20] Yang R, Ma H L, Wu J, et al. ScalableViT: Rethinking The context-oriented generalization of Vision Transformer [J/OL]. *arXiv preprint* (2022-03-21) [2024-05-13]. https://arxiv.org/abs/2203.10790

[21] Liang L M, He A J, Yang Y, et al. Colorectal polyp segmentation method based on the Swin Transformer and graph reasoning. *Chin J Eng*, 2024, 46(5): 897
(梁礼明, 何安军, 阳渊, 等. 基于 Swin Transformer 和图形推理的结直肠息肉分割方法. 工程科学学报, 2024, 46(5): 897)

[22] Yang B C, Wang J Y, Jin H B. DS-TransFusion: Automatic retinal vessel segmentation based on an improved Swin Transformer. *Chin J Eng*, 2024, 46(10): 1889
(杨本臣, 王建宇, 金海波. DS-TransFusion: 基于改进 Swin Transformer 的视网膜血管自动分割. 工程科学学报, 2024, 46(10): 1889)

[23] Wang W X, Chen W, Qiu Q B, et al. CrossFormer: A versatile vision transformer hinging on cross-scale attention. *IEEE Trans Pattern Anal Mach Intell*, 2024, 46(5): 3123

[24] Chen Y P, Dai X Y, Chen D D, et al. Mobile-former: Bridging MobileNet and transformer // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans 2022: 5260

[25] Lou M, Zhou H Y, Yang S B, et al. TransXNet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition [J/OL]. *arXiv preprint* (2023-10-30) [2024-05-12]. https://arxiv.org/abs/2310.19380

[26] Yuan K, Guo S P, Liu Z W, et al. Incorporating convolution designs into visual transformers // *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021: 579

[27] Tian C, Tang Z H, Zhang H, et al. Operating condition recognition based on temporal cumulative distribution function and AdaBoost-extreme learning machine in zinc flotation process. *Powder Technol*, 2022, 395: 545

[28] Zhou F Y, Jin L P, Dong J. Review of convolutional neural network. *Chin J Comput*, 2017, 40(6): 1229
(周飞燕, 金林鹏, 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 1229)

[29] Chu X X, Tian Z, Zhang B, et al. Conditional positional encodings for vision transformers [J/OL]. *arXiv preprint* (2021-02-22) [2024-05-12] https://arxiv.org/abs/2102.10882

[30] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks *via* gradient-based localization // 2017 *IEEE International Conference on Computer Vision*. Venice, 2017: 618