中国激光

基于Transformer的宫颈异常细胞自动识别方法

张峥[1]，陈明销[1]，李新宇[1]，程逸[1]，申书伟[2]，姚鹏[3*]

[1]中国科学技术大学工程科学学院精密机械与精密仪器系，安徽 合肥 230027；

[2]中国科学技术大学苏州高等研究院，江苏 苏州 215123；

[3]中国科学技术大学微电子学院，安徽 合肥 230027

摘要 宫颈异常细胞与正常细胞在形态上存在较大相似性且细胞尺寸变化较大，这使得宫颈异常细胞的精准检测变得非常困难。鉴于此，开发了一种基于Transformer模型的宫颈异常细胞自动识别模型，以帮助病理学家作出更准确的诊断。提出了两种创新性方法，一是一种改进的Transformer编码器结构，通过引入深度(DW)卷积来高效获取图像的特征，捕捉图像中的全局依赖信息；二是自适应的动态交并比(IOU)阈值，在模型训练的不同阶段使用不同的IOU阈值，实现尽可能多的有效检测，提升模型的收敛速度和检测精度。在宫颈异常细胞数据集上，通过消融实验，证明了改进的Transformer编码器和动态IOU阈值的有效性。此外，与已有的宫颈异常细胞识别方法相比，所提出的方法在平均精度指标上有明显的提高。实验结果表明，所提出的方法能够高效且准确地识别宫颈异常细胞，且能辅助病理专家提高诊断准确率和效率，具有应用到临床的潜力。

关键词 医用光学；宫颈细胞病理图像；目标检测；医学图像处理

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/CJL231261

1 引 言

宫颈癌是一种常见的肿瘤，对女性的健康构成了严重的威胁。2020年，全球有60万宫颈癌新发病例，超过34万人死于该病[1]。宫颈癌有较长的癌前阶段，这为宫颈癌的筛查和及时治疗提供了机会。目前，检测宫颈癌最常用的方法是宫颈细胞学筛查，主要通过液基薄层细胞检测(TCT)[2]进行。具体来说，在TCT过程中，对患者的宫颈细胞进行收集并在玻片上进行染色，之后在显微镜下进行目视检查和细胞病理学分析，病理学家通过对细胞类型和形态学特征如细胞核大小、核质比等作出评估，给出初步的诊断建议。然而，这种传统的宫颈细胞学筛查费时费力且容易出错[3]，同时宫颈细胞学筛查医师较为缺乏，难以满足目前宫颈癌筛查的需求[4]。随着深度学习的发展，已经有一些自动的宫颈细胞学筛查方法出现[5-6]，这些方法极大地减轻了病理学家的负担，提高了检测效率。Plissiti等[7]使用VGG-16网络对单个分离的宫颈细胞图像进行分类。Talo等[8]提出了基于DenseNet161[9]的模型，对传统的方法进行了改进。Win等[10]在深度学习的基础上，引入随机森林、支持向量机等多种传统机器学习方法进行分割和分类。Du等[11]将多细胞重叠的区域划分为多个单细胞区域，并使用Faster R-CNN模型进行检测和识别。Liang等[12]提出了一个基于YOLOv3模型的全局上下文感知框架来进行异常细胞检测。为解决数据有限的问题，一种特征比较的方法[13]利用比较检测器来进行宫颈细胞检测。这些方法大多是基于卷积神经网络(CNN)，由于CNN的卷积算子存在局部感受野受限的问题，其对图像中的全局特征以及远程依赖特征的提取能力不足，故这些方法的检测效果尚未达到临床要求[14]。

Transformer模型最初来源于自然语言处理领域，近年来，Transformer模型正在获得越来越多的关注。Vision Transformer模型[15]展现出Transformer模型在图像领域中的巨大潜力，其应用迅速被拓展到图像分类[16-17]、语义分割[18-19]、目标检测[20-21]等任务中。之后，又有一些新的改进的Transformer结构被提出，其中较为重要的工作是PVT模型[22]和Swin Transformer模型[16]。PVT模型引入了金字塔结构，将输入图像分解为不同分辨率图像，然后通过Transformer模型来处理这些图像，以捕获多尺度的信息。Swin Transformer模型是另一种视觉Transformer架构，通过滑动窗口操作来提高模型的局部性。

相比于自然图像，在宫颈细胞学图像中，因为异常细胞是由正常细胞缓慢发展变化而来的，异常细胞和正常细胞具有很高的相似性，所以异常细胞的检测变

收稿日期：2023-10-09；**修回日期**：2023-11-23；**录用日期**：2023-12-01；**网络首发日期**：2023-12-12

基金项目：安徽省自然科学基金(2308085MF219)

通信作者：*yaopeng@ustc.edu.cn

得非常困难。病理学家通常需要参照图像中的正常细胞，才能准确地分辨异常细胞。Transformer模型具有强大的全局特征以及远程依赖抽取能力，非常适用于宫颈异常细胞识别。然而，据我们所知，目前还只有少数的研究将Transformer模型应用到宫颈异常细胞识别上[23-24]。并且这些应用往往只是简单地用Transformer模型代替卷积架构，而没有针对特定任务进行特殊设计，因此很难达到较好的识别效果。

本文提出了一种新颖的基于Transformer模型的宫颈异常细胞检测方法，包含一种改进的Transformer编码器结构，该结构通过多尺度自注意力机制，能够更加有效地捕获图像中的关键信息，提升模型的检测效果。此外，我们也设计了一种能够自适应调整交并比(IOU)阈值的动态IOU阈值法，在训练的初始阶段，模型能够实现尽可能多的有效检测，而在训练的后期，模型能够过滤掉大部分的假阳性预测，从而提升模型的检测精度。实验表明本文的方法能够准确识别宫颈细胞学图像中的异常细胞并进行分类，模型性能优于一些现有的方法。

2 本文方法

2.1 模型整体架构

本文设计了一种能够高效进行全局特征提取的模型，整体架构如图1所示，采用Faster RCNN[25]作为基础架构，使用改进的Transformer结构作为骨干网络(backbone)。骨干网络所使用的Transformer模型被划分为4个阶段，分别生成不同尺度的特征图，其中C_1、C_2、C_3、C_4分别为第一、二、三、四阶段的特征维度。所有阶段共享相似的结构，每个阶段包含一个分块嵌入(Patch Embedding)模块和一个Transformer编码器模块。

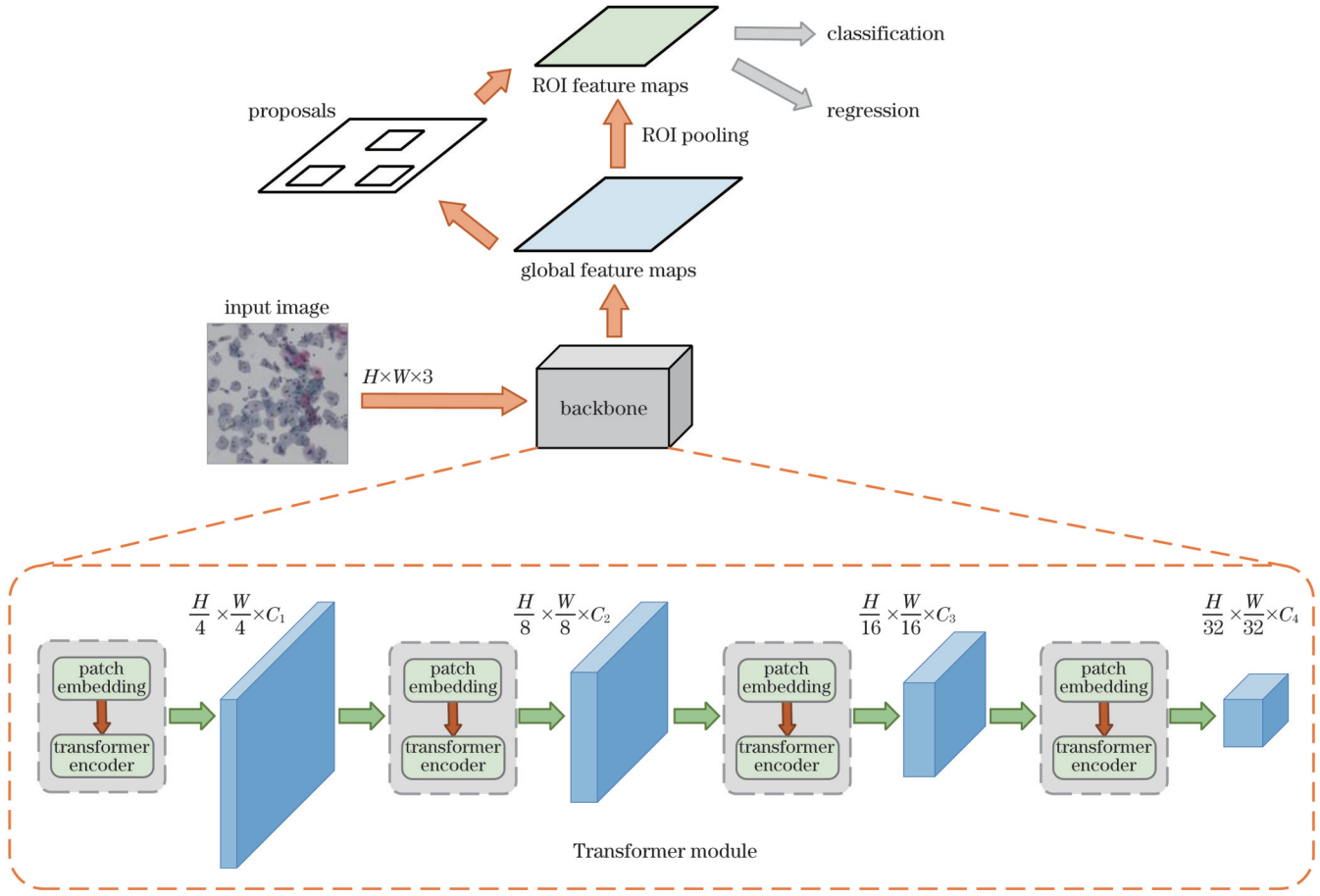

图1 模型整体架构图

Fig. 1 Overall architecture diagram of model

输入图像的尺寸是$H\times W\times 3$，首先将其划分为$4\times 4\times 3$大小的切片(patch)，为了处理patch之间的位置信息，通常需要添加位置编码，以确保模型能够理解patch之间的相对位置。位置编码是一个可学习的参数矩阵，将位置信息嵌入到patch的表示中，然后将这些patch展平成线性投影，输入到第一阶段的Patch Embedding和Transformer编码器模块中，得到$\frac{H}{4}\times\frac{W}{4}\times C_1$大小的特征图。使用前一阶段的输出作为下一阶段的输入，分别得到第二、三、四阶段的输出特征图。随着阶段数的增加，特征维度也在增加，例如，第一阶段的特征维度C_1为64，第四阶段的特征维度C_4为512。

由于目标检测可以分为分类任务和回归任务，因此模型使用交叉熵损失函数来处理分类任务，使用L1

损失(L1 Loss,L_1)函数来处理回归问题。交叉熵损失函数可以测量两个概率分布之间的差异,一般形式为

$$H(Q,P)=-\sum_{i=1}^{N}Q(i)\log\left[P(i)\right],\tag{1}$$

式中:H表示交叉熵损失函数;$Q(i)$表示真实分布中第i个类别的概率;$P(i)$表示模型生成的概率分布中第i个类别的概率;N表示类别数。交叉熵损失函数的目标是最小化模型生成的概率分布P与真实分布Q之间的差异,从而使模型更好地逼近真实标签的分布。

L1损失函数是一种用于回归问题的损失函数,它测量了预测值与实际值之间的绝对差异。L1损失函数的一般形式为

$$L_1=\sum_{j=1}^{N_1}|y_j-y_j^{\mathrm{p}}|,\tag{2}$$

式中:y_j表示第j个样本的真实标签;y_j^{p}表示模型对第j个样本的预测值;N_1表示样本总数。L1损失函数计算了每个样本的绝对误差,并将它们相加,从而得到总误差。

2.2 Transformer编码器

Transformer编码器由多个相同结构的块(block)堆叠而成,每个block包括两部分:多头自注意力层和前馈神经网络层,每个block均执行相同的操作,具体结构如图2(a)所示,其中GELU表示激活函数,FC1和FC2表示全连接层,DW Conv表示深度卷积。在多头自注意力层中,首先进行正则化处理,再将生成的查询($\boldsymbol{Q}$)、键($\boldsymbol{K}$)、值($\boldsymbol{V}$)输入到多头注意力模块中进行计算。多头自注意力层能够在不同层次和尺度上捕捉输入数据的相关性,使模型能够更好地理解输入序列的结构。而前馈神经网络层包含多个全连接层和激活函数,引入了非线性变换,这有助于模型更好地适应复杂的数据分布。前馈神经网络层充当了多头自注意力层后的特征提取和映射层,有助于生成适合特定任务的表示。

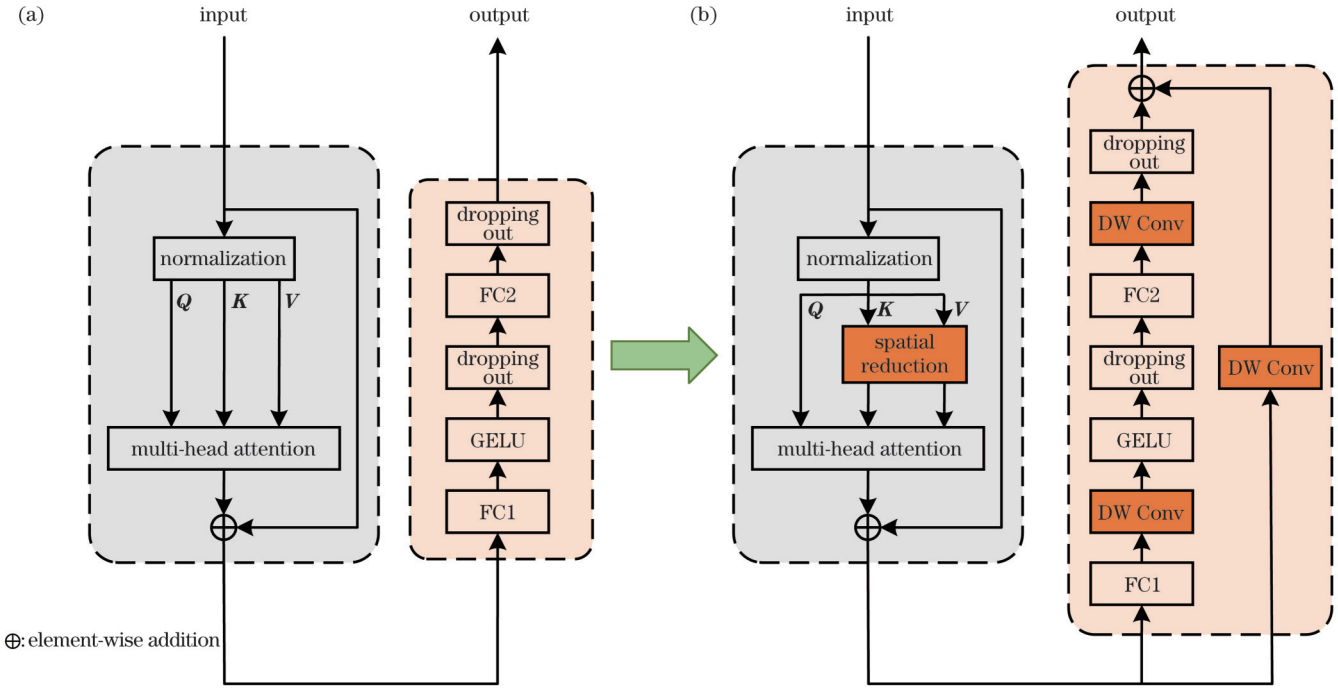

图2　Transformer编码器块结构图。(a)通用结构;(b)改进结构

Fig. 2　Architecture diagrams of transformer encoder block. (a) Generic structure; (b) improved structure

Transformer模型的$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$融合是通过计算注意力分数(A_{t})和加权和来完成的。

$$d_{\mathrm{head}}=\frac{C'}{N'},\tag{3}$$

$$A_{\mathrm{t}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{head}}}}\right)\boldsymbol{V},\tag{4}$$

式中:C'为输入特征图的通道数;N'为每个阶段注意力层的特征头数;d_{head}为每个头的维度数。

每个位置的注意力分数是通过将$\boldsymbol{Q}$与所有位置的$\boldsymbol{K}$进行点积计算得到的。然后,这些注意力分数经过Softmax函数进行归一化,得到用于加权和的权重。最后,将这些权重应用到对应位置的$\boldsymbol{V}$中,以生成最终的输出表示。融合$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的过程允许Transformer模型在输入序列中建立全局依赖性并融合上下文信息,从而更好地理解序列数据的关系和特征。这种自注意力机制使得Transformer能够处理不同长度和结构的序列数据。本文参考PVT[22]模型的设计,采用空间减少注意力(SRA)层来替代传统的多头注意力机制。SRA模块首先对输入的特征图进行通道维度上的升维操作,以得到一个更高维的特征表示。然后再使用全局池化操作对输入特征图中的$\boldsymbol{K}$和$\boldsymbol{V}$进行尺度缩减。接着,SRA模块利用全连接层分别获取$\boldsymbol{K}$和$\boldsymbol{V}$在高维特征空间中的表示。最终,通过在高维空间中计算注意力权重,并利用这些权重对$\boldsymbol{V}$进

行加权，即可得到输出特征图。相较于传统的注意力机制，SRA 模块通过空间减少注意力层在不同尺度的特征图上进行多尺度信息捕捉、特征降维和位置信息融合，以实现高效的多尺度特征表示。

本文也改进了 Transformer 编码器的前馈神经网络层的结构，使其更加适用于宫颈异常细胞的检测，修改后的结构如图 2(b)所示。在编码器中引入深度(DW)卷积。具体来说，在第一个和第二个全连接层后加入 DW 卷积以处理全连接层输出的特征。此外，前馈神经网络层的输入经 DW 卷积处理后被引到输出部分，与 Dropout 层的输出相加。DW 卷积的一个卷积核只负责处理一个通道的信息，输入特征图的通道数与用于处理的卷积核数相等，结构如图 3 所示，每个卷积核得到一个通道为 1 的特征图，再将这些特征图拼接便得到输出特征图。DW 卷积在每个输入通道上使用单独的卷积核执行卷积操作，不混合通道之间的信息，这有助于捕捉通道内的特征，而不引入通道间的关联。

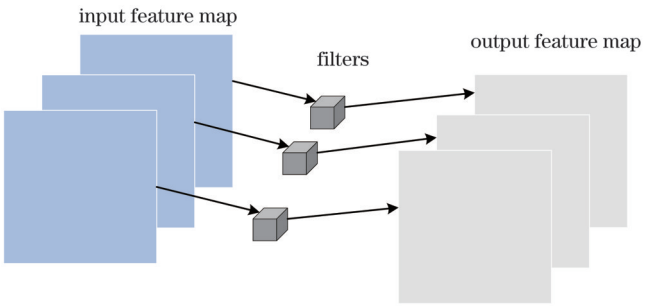

图 3　DW 卷积示意图

Fig. 3　DW convolution diagram

在前馈神经网络层中也采用了 Dropout 处理。在每个训练迭代过程中，Dropout 层随机选择一部分神经元并将其输出设置为零。这意味着在前向传播过程中，这些神经元对计算和传播梯度不起作用。Dropout 层的概率参数通常表示为 p，它确定了每个神经元被丢弃的概率。Dropout 层的核心原理在于强制模型学习多个独立的子模型。在每个训练迭代过程中都会随机丢弃不同的神经元，每个子模型都学到了输入数据的不同方面和特征，从而增加了模型的多样性。Dropout 层能够减少神经网络过拟合训练数据的倾向，使模型在陌生数据上更具泛化能力。

2.3　动态 IOU 阈值

IOU 阈值是指用于过滤检测框(bounding boxes)的一个阈值。IOU 在目标检测中起着关键作用，主要用于衡量检测框和真实标签框之间的匹配程度，通常在 0 到 1 之间，如果 IOU 接近 1，表示检测是准确的，如果 IOU 接近 0，表示检测是错误的。IOU 还用于目标检测的性能评估，在测试阶段可以使用 IOU 作为评价指标来测量检测模型的性能，例如精确率(precision)、召回率(recall)、F1 分数等都依赖于 IOU 的计算。由此可见，IOU 阈值的选择至关重要。在训练过程中，过高的 IOU 阈值可能会导致漏检，特别是在训练的起始阶段，低概率值的标签容易被忽略，导致模型收敛速度较慢。此外，过高的 IOU 阈值很容易导致小尺度目标框丢失。而过低的 IOU 阈值可能导致更多错误检测，因为即使检测框与真实目标的重叠很小，检测结果仍然被认为是有效的检测结果，所以假阳率增大，影响模型的检测效果。

本文提出了一种自适应的动态 IOU 阈值，在训练的起始阶段阈值较低，能够实现尽可能多的有效检测。而在训练的后期，较高的 IOU 阈值能够过滤掉大部分的假阳性预测，降低假阳率，提高模型的准确度。动态 IOU 阈值(ρ_t)的公式为

$$\rho_t = C + \gamma \ln\left(0.1 + \frac{t}{N_2}\right), \tag{5}$$

式中：C 和 γ 都是可以调整的参数，在实验中分别设为 0.80 和 0.25；N_2 为总训练迭代轮数(epoch)；t 为当前迭代 epoch。随着迭代次数的增加，动态 IOU 阈值 ρ_t 逐步增加。训练过程中动态 IOU 阈值的变化情况如图 4 所示。

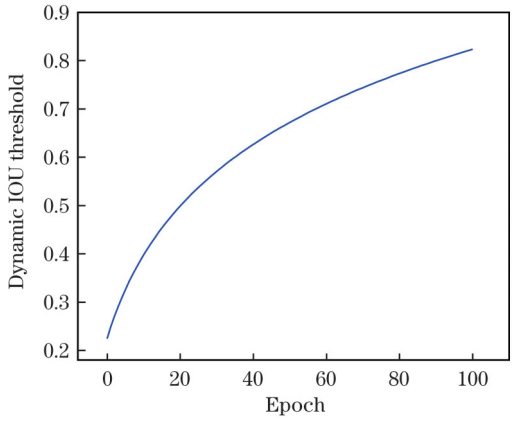

图 4　动态 IOU 阈值的变化情况

Fig. 4　Changes of dynamic IOU thresholds

3　分析与讨论

3.1　评价指标

本文采用平均精度(AP)作为评价指标。AP 是一种常用于评估目标检测和物体识别任务性能的指标，综合权衡了精确率和召回率。精确率(P_r)是指模型正确识别出的正样本占模型识别为正样本的所有样本数量的比例。召回率(R_e)则是指模型正确识别出的正样本占所有真正正样本的数量的比例。

$$P_r = \frac{T_P}{T_P + F_P}, \tag{6}$$

$$R_e = \frac{T_P}{T_P + F_N}, \tag{7}$$

式中：T_P 表示模型正确识别的正样本数量；F_P 表示模型错误识别的负样本数量；F_N 表示模型未能正确识别

的正样本数量。

AP是单个类别的精确率-召回率曲线下的面积，用于衡量模型在单个类别上的性能，范围在0到1之间。实验中也使用一些辅助性的AP指标，如AP_{50}、AP_{75}、AP_S、AP_M和AP_L。AP_{50}是指在物体检测任务中，当IOU大于等于50%时的平均精确率。AP_{75}是IoU大于等于75%时的平均精确率，这个指标是衡量模型在更高IoU阈值下的精确性。AP_S、AP_M、AP_L则是目标检测任务中小、中、大目标的AP评分。这些不同尺度的AP值可以更全面地衡量模型在不同大小物体上的性能表现。

3.2 数据集和实现细节

本文使用的全切片扫描图像(WSI)是一种高分辨率的病理组织切片的数字图像，这种图像的生成依赖于数字扫描仪和各种光学元件(如光学镜头、光学传感器和透镜)。将光聚焦在组织切片上，利用各种光学器件将玻片上的光学信息转为数字图像，保留了病理学上的详细结构，使医生能够直接观察组织和细胞的形态、结构和染色情况。

在20倍的物镜放大倍数下，通过明场扫描的方法对苏木精-伊红(H&E)染色的切片进行扫描，所获得的全切片图像是一种多分辨率的层次模型，其包含更多的信息，这是传统的光学显微镜无法实现的。

本文使用的数据集来自天池宫颈癌风险智能诊断挑战赛[26]。所有图像均为宫颈癌液基薄层细胞，所有标签均由专业医师标注[27]。该数据集包含800张WSI，其中500张为包含异常细胞的阳性WSI，我们随机选择其中的350张用来训练，75张作为验证集，剩下的作为测试集。由于WSI尺寸极大，不能直接输入到模型中，首先需要对其中包含异常细胞的局部图像进行裁剪，裁剪尺寸固定为1000 pixel×1000 pixel，从而获得模型可利用的细胞图像数据集。处理后的数据集包含27915张图片，其中23089张作为训练集，2100张作为验证集，2726张作为测试集。

由于医学领域的有标签数据相对稀缺，为了最大程度利用现有的有标签数据，在训练之前需要对输入图像进行充分的数据增强。本文主要采用两个类型的增强方法，包括形状变换和颜色变换。对于形状变换，采用旋转、镜像翻转、平移和缩放变换等操作，对输入图像的形状进行调整，增加模型对不同视角和方向图像的识别能力。对于颜色变换，主要采用亮度和对比度调整、模糊及滤波等操作，在保持数据真实性和有效性的前提下，对图像进行尽可能多的变换。这些数据增强方法在医学图像领域中尤为重要，因为医学图像往往受多种因素的影响，如设备差异、光照变化等，所以医学图像的风格差异较大。通过应用数据增强技术，可以增加医学图像数据的多样性，提高深度学习模型的鲁棒性和泛化性能。

本文代码基于PyTorch框架，使用8个NVIDIA GeForce RTX 2080 Ti显卡进行训练。优化器选择了随机梯度下降(SGD)算法，并设置动量参数为0.9，初始学习率为0.01。

3.3 对比实验

为了验证本文方法的有效性，我们在所使用的天池数据集上开展了实验，并与常见的通用目标检测方法进行了对比。为了保证对比实验的公平性，所有模型都在相同的环境中进行训练和测试，实验结果如表1所示。比较对象包含基于CNN的一阶段模型YOLOv3、RetinaNet[28]、DETR[29]、FCOS[30]和GiraffeDet[31]以及二阶段的模型Sparse R-CNN[32]、Cascade R-CNN[33]和Iter Sparse R-CNN[34]。与多种通用的目标检测模型相比，我们所提出的方法达到了最高的检测精度。Sparse R-CNN在所有通用的目标检测模型中取得了最好的效果，与Sparse R-CNN相比，本文模型在各个指标上均有提升，特别是AP和AP_{50}分别提高了3.0%和5.8%，达到了26.1%和46.8%，展现出良好的应用前景。在AP_S指标上，本文方法远超其他对比模型，证

表1 各种模型的实验结果

Table 1 Experimental results of various models

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Number of parameters /10^6
YOLOv3	18.0	35.4	16.8	0.0	6.3	25.9	61.5
RetinaNet[28]	23.5	43.7	22.5	0.3	10.6	32.2	37.9
Sparse R-CNN[32]	24.0	42.1	23.2	0.2	10.8	33.4	108.5
Cascade R-CNN[33]	22.1	39.3	21.4	0.2	7.6	32.6	69.3
DETR[29]	23.5	46.3	21.6	0.5	9.2	32.6	41.5
FCOS[30]	23.1	41.0	22.5	0.2	12.2	32.4	32.1
GiraffeDet[31]	23.9	44.3	23.1	0.3	10.9	32.6	47.8
Iter Sparse R-CNN[34]	24.8	43.8	25.3	0.7	11.7	33.0	123.4
Ours	26.1	46.8	25.5	1.8	13.5	33.5	48.3

明了所提出的模型除了能够检测正常大小的目标外，对于极小的目标，也具备一定的检测能力。这一特性在宫颈异常细胞检测中显得尤为重要，因为癌变细胞的尺度变化较大且存在部分癌变细胞尺寸很小的情况，如果模型的小目标检测性能较差则会造成严重的漏检问题。本文模型适用于各个尺寸的检测目标，具备良好的泛化性和检测能力。通过对比各个模型的参数量，可知本文模型的参数量是较少的，这表明本文模型在不大幅增加参数量的情况下可以有效地提升检测性能。

除了与通用模型进行对比外，本文模型还与专门为宫颈异常细胞检测而设计的网络 attFPN[35]进行了比较。attFPN 由两个部分组成：一是模拟病理学家阅读宫颈细胞学图像的注意力模块，能够对提取到的特征进行细化以强调或抑制某些特征；另一个模块是多尺度特征融合网络，通过融合细化后的特征，检测不同区域的宫颈癌变细胞。其检测结果如表 2 所示，结果表明：我们的方法优于 attFPN，这表明本文提出的方法在宫颈异常细胞识别上也具有优势。以上对比实验都证明，与其他检测算法相比，本文模型具有更高的检测精度，并且在小目标检测方面也表现出色。使用本文所提出的方法能够极大地提升宫颈异常细胞筛查的准确性和可靠性。

表 2　本文模型和 attFPN 的对比

Table 2　Comparison between proposed model and attFPN

Model	AP	AP_{50}	AP_{75}
attFPN	25.0	50.3	22.2
Proposed model	26.1	46.8	25.5

本文方法和其他方法的可视化结果对比如图 5 所示，取有代表性的 Sparse R-CNN 作为比较对象。从检测效果对比图可以看出，本文方法能够更加准确地检测出目标区域，和真实标签较为接近。而 Sparse R-CNN 可能会产生假阳性，将一些正常细胞检测为癌变细胞，这将对模型的准确率产生影响。此外，Sparse

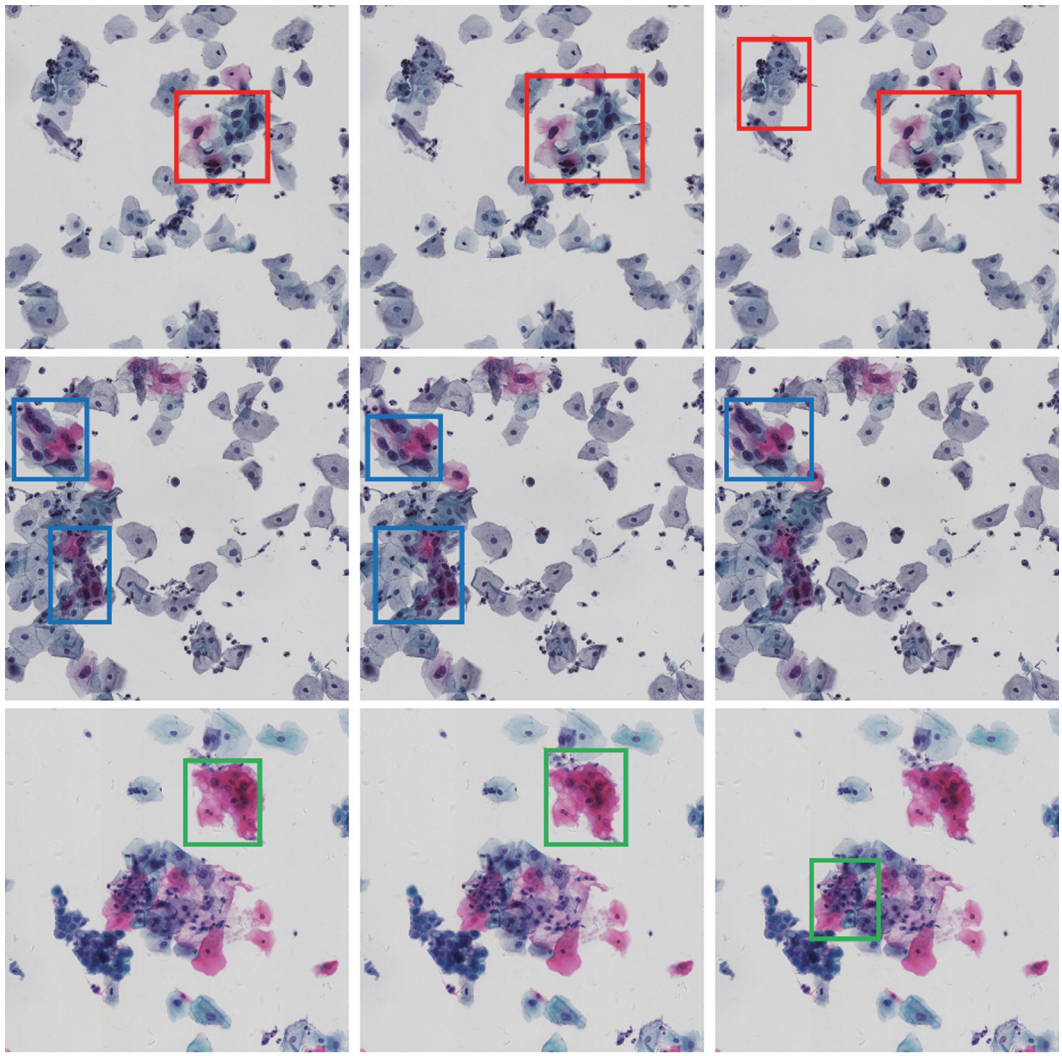

图 5　不同模型的检测效果对比。(a)真实标签；(b)所提方法；(c)Sparse R-CNN

Fig. 5　Comparison of detection effects of different models. (a) Ground truth; (b) proposed method; (c) Sparse R-CNN

R-CNN还有可能产生更多的漏检，忽略掉一部分癌变细胞，这在医学领域是难以接受的。总的来看，本文方法能够达到更好的检测精度，相比较之下具有更低的假阳率与漏检率，符合医学领域对宫颈癌变细胞自动检测模型的要求。

3.4 消融实验

消融实验主要是为了验证Transformer作为backbone时的效果，以及我们改进的Transformer编码器模块和动态IOU阈值的有效性。表3展示了分别选择Transformer和CNN(Resnet-101)作为backbone的模型的检测结果，基于Transformer的方法在各项指标上相较于基于CNN的方法均有明显提升，特别是在AP_{50}上提升了4.7%。这表明Transformer的全局特征和远程依赖信息的提取能力对于宫颈异常细胞检测至关重要，能够有效地提升检测效果。

针对改进的Transformer编码器模块和动态IOU阈值的消融实验结果如表4所示。用作对比的基线模型采用了原始的Transformer编码器和固定的IOU阈值。采用改进的Transformer编码器的方法相较于原来的模型在AP和AP_{50}上分别提高了1.8%和2.3%，在其他各项评价指标上也有更加优秀的表现。在AP和AP_{50}指标上，采用动态IOU阈值的模型相较于固定IOU阈值模型分别提高了0.6%和0.9%，这表明采用动态的IOU阈值能够帮助模型更好地学习图像中的特征。可以观察到，本文方法的AP、AP_{50}、AP_{75}相较于不采用两种新设计的原始方法分别高出2.0%、3.0%、1.7%。这些结果都表明，本文提出的Transformer编码器和动态IOU阈值能够有效提升模型对宫颈异常细胞的检测能力。

表3 不同backbone选择下的实验结果对比

Table 3 Comparison of experimental results under different backbone choices

Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
CNN(Resnet-101)	23.7	42.1	24.2	1.1	12.8	31.4
Transformer	26.1	46.8	25.5	1.8	13.5	33.5

表4 不同模块的消融实验结果

Table 4 Ablation experiment results of different modules

Improved Transformer encoder	Dynamic IOU threshold	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
		24.1	43.8	23.8	0.5	13.2	30.7
	√	24.7	44.7	25.2	0.2	12.0	32.2
√		25.9	46.1	25.3	2.5	12.2	33.8
√	√	26.1	46.8	25.5	1.8	13.5	33.5

我们也对动态IOU阈值和多个固定的IOU阈值进行了定量对比分析，固定的IOU阈值分别取0.5、0.6、0.7，结果如表5所示。可以看出：基于动态IOU阈值的模型相较于多个基于固定IOU阈值的模型，在各项指标上均表现优异。原因在于动态IOU阈值具备更好的灵活性和适应性，减少了对不太可靠检测结果的依赖，降低了误检的概率。

表5 动态IOU阈值和多个固定IOU阈值模型的对比

Table 5 Comparison among models based on dynamic IOU threshold and multiple fixed IOU thresholds

IOU threshold	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.5	23.6	45.5	22.4	0.4	12.6	31.0
0.6	24.0	42.1	23.2	0.2	10.8	33.4
0.7	24.1	43.8	23.8	0.5	13.2	30.7
Dynamic threshold	24.7	44.7	25.2	0.2	12.0	32.2

此外，我们也比较了动态IOU阈值和固定IOU阈值模型在训练过程中的损失和AP的变化，如图6所示。容易观察出，采用动态IOU阈值的模型的损失减小得更快。这是由于在训练初期，动态IOU阈值较小，能够获得更多的正确预测目标，模型可以学习到更多特征，这一特性导致其收敛速度加快。采用动态IOU阈值的模型在训练后期的平均精度也较高。这是因为在训练的中后期动态IOU阈值较高，能够有效地过滤掉错误预测，模型受假阳性预测的影响减小，因而能够获得较高的平均精度。

除了定量的比较外，也对模型的有效性进行了定性分析。本文采用Grad-CAM[36]生成热图，以此对模型的检测能力进行比对。热图可以提供关于目标位置的精细粒度信息，使模型可以更准确地理解图像中的目标分布情况。将不采用改进Transformer编码器和动态IOU阈值的原始Transformer模型和本文模型进行对比，生成的热图如图7所示。通过对比可以看出：我们所提出的模型能够将关注点更多地集中到图像中的细胞上，受背景信息的干扰较少。随着训练的进行，模型能够从细胞分布的区域学习到更多有效的特征，从而提升模型的识别效果。可以看出，所提出的模型能够更好地结合全局的上下文信息，而非仅仅关注局部信息。通过训练过程中的热图对比，我们证明了所提出的基于改进Transformer编码器和动态IOU阈值的模型的有效性。

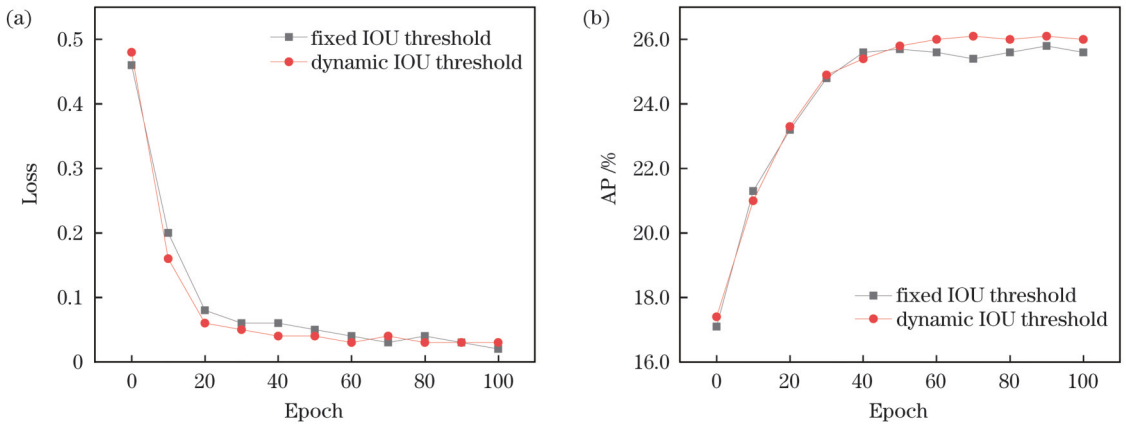

图 6　动态 IOU 阈值和固定 IOU 阈值下的损失和 AP。(a)损失;(b)AP

Fig. 6　Loss and AP under dynamic IOU threshold and fixed IOU threshold. (a) Loss; (b) AP

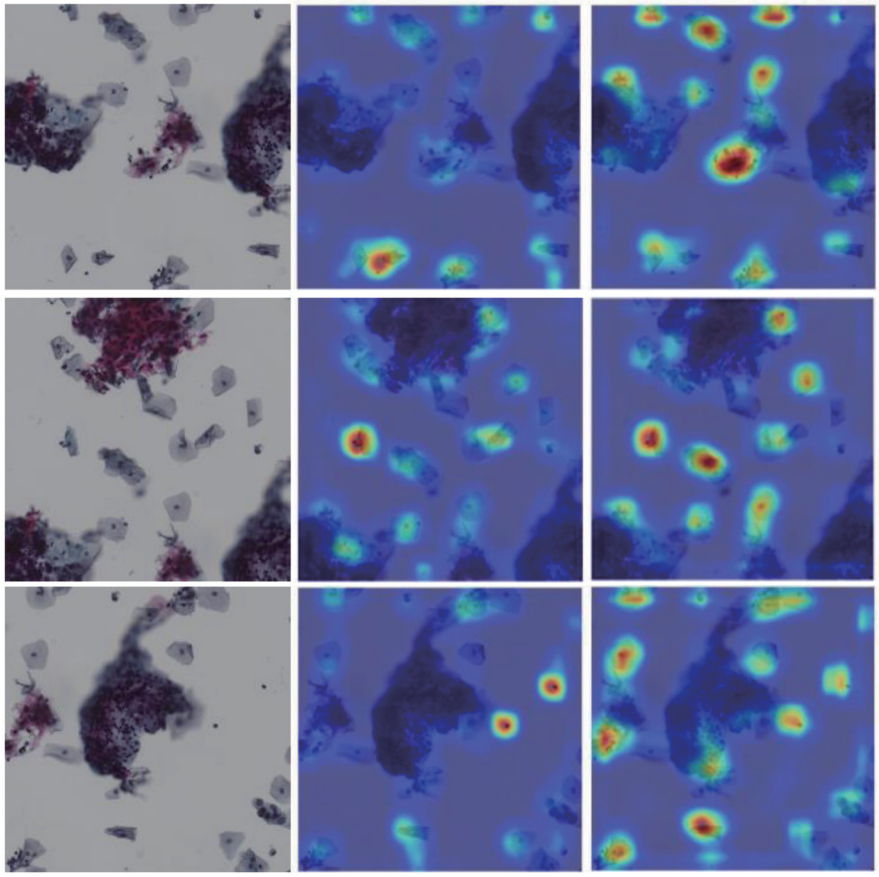

图 7　消融实验热图。(a)原始图像;(b)原始 Transformer 模型生成的热图;(c)我们的方法生成的热图

Fig. 7　Ablation experiment heatmaps. (a) Original images; (b) heatmap generated by original Transformer model; (c) heatmap generated by our method

4　结　　论

提出了一种宫颈异常细胞自动识别方法,模型基于 Transformer 架构。在此基础上提出了改进的 Transformer 编码器结构和可以动态变化的 IOU 阈值。在数据集上进行了多种对比实验,结果表明:所提出的方法在精度等各种指标上均优于已有的方法,能够实现准确的宫颈异常细胞识别。通过消融实验,证明了所提出的两个模块均能够提高模型对宫颈异常细胞的识别准确度。总的来看,所提出的方法能够极大地提升医学图像的筛查效率,节省医疗时间和资源,及时发现癌症病变,具有一定的临床和使用价值。

后续的研究将更关注数据的高效利用。医学图像领域的有标注数据集较为稀缺,这极大地限制了模型性能的提升。然而在医学领域中还存在着大量的无标签或弱标签数据,这些数据也蕴含着相当多的可以利

用的特征，因此无标签数据的利用对于医学图像领域的模型改进十分必要。后续的研究可以更多关注半监督学习和无监督学习在医学图像领域中的应用，利用这些方法提高医学图像的利用率，提升模型的检测效果，使其更好地符合临床需求。

参 考 文 献

[1] Sung H, Ferlay J, Siegel R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: A Cancer Journal for Clinicians, 2021, 71(3): 209-249.

[2] de Bekker-Grob E W, de Kok I M C M, Bulten J, et al. Liquid-based cervical cytology using ThinPrep technology: weighing the pros and cons in a cost-effectiveness analysis[J]. Cancer Causes & Control, 2012, 23(8): 1323-1331.

[3] Elsheikh T M, Austin R M, Chhieng D F, et al. American Society of Cytopathology workload recommendations for automated Pap test screening: developed by the productivity and quality assurance in the era of automated screening task force[J]. Diagnostic Cytopathology, 2013, 41(2): 174-178.

[4] 李雪, 石中月, 杨志明, 等. 人工智能辅助分析在宫颈液基薄层细胞学检查中的应用价值[J]. 首都医科大学学报, 2020, 41(3): 360-363.
Li X, Shi Z Y, Yang Z M, et al. Value about artificial intelligence-assisted liquid-based thin-layer cytology for cytology cervical cancer screening[J]. Journal of Capital Medical University, 2020, 41(3): 360-363.

[5] Chen Y F, Huang P C, Lin K C, et al. Semi-automatic segmentation and classification of pap smear cells[J]. IEEE Journal of Biomedical and Health Informatics, 2014, 18(1): 94-108.

[6] William W, Ware A, Basaza-Ejiri A H, et al. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images[J]. Computer Methods and Programs in Biomedicine, 2018, 164: 15-22.

[7] Plissiti M E, Dimitrakopoulos P, Sfikas G, et al. Sipakmed: a new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images[C]//2018 25th IEEE International Conference on Image Processing (ICIP), October 7-10, 2018, Athens, Greece. New York: IEEE Press, 2018: 3144-3148.

[8] Talo M. Diagnostic classification of cervical cell images from pap smear slides[J]. Academic Perspective Procedia, 2019, 2(3): 1043-1050.

[9] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2261-2269.

[10] Win K P, Kitjaidure Y, Hamamoto K, et al. Computer-assisted screening for cervical cancer using digital image processing of pap smear images[J]. Applied Sciences, 2020, 10(5): 1800.

[11] Du L X Y, Li Q H. Detection and classification of cervical exfoliated cells based on faster R-CNN[C]//2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), October 18-20, 2019, Jinan, China. New York: IEEE Press, 2019: 52-57.

[12] Liang Y X, Pan C L, Sun W X, et al. Global context-aware cervical cell detection with soft scale anchor matching[J]. Computer Methods and Programs in Biomedicine, 2021, 204: 106061.

[13] Liang Y X, Tang Z H, Yan M, et al. Comparison detector for cervical cell/clumps detection in the limited data scenario[J]. Neurocomputing, 2021, 437: 195-205.

[14] 辛仲宏, 雷军强, 郭城, 等. 深度学习用于宫颈癌诊疗研究进展[J]. 中国医学影像技术, 2022, 38(5): 779-782.
Xin Z H, Lei J Q, Guo C, et al. Research progresses of deep learning in diagnosis and treatment of cervical cancer[J]. Chinese Journal of Medical Imaging Technology, 2022, 38(5): 779-782.

[15] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale[EB/OL]. (2020-10-22)[2023-05-06]. https://arxiv.org/abs/2010.11929.

[16] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9992-10002.

[17] Pan X R, Ge C J, Lu R, et al. On the integration of self-attention and convolution[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 805-815.

[18] Strudel R, Garcia R, Laptev I, et al. Segmenter: transformer for semantic segmentation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 7242-7252.

[19] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), June 17-22, 2006, New York, NY, USA. New York: IEEE Press, 2006: 1735-1742.

[20] Li Y H, Mao H Z, Girshick R, et al. Exploring plain vision transformer backbones for object detection[M]//Avidan S, Brostow G, Cissé M, et al. Computer vision-ECCV 2022. Lecture notes in computer science. Cham: Springer, 2022, 13669: 280-296.

[21] Pan X R, Xia Z F, Song S J, et al. 3D object detection with pointformer[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 7459-7468.

[22] Wang W H, Xie E Z, Li X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 548-558.

[23] Liu W L, Li C, Xu N, et al. CVM-Cervix: a hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron[J]. Pattern Recognition, 2022, 130: 108829.

[24] Liu Y T, Zhao J J, Luo Q Y, et al. Automated classification of cervical lymph-node-level from ultrasound using Depthwise Separable Convolutional Swin Transformer[J]. Computers in Biology and Medicine, 2022, 148: 105821.

[25] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2016: 1440-1448.

[26] 'Dataset' [EB/OL]. [2023-05-06]. https://tianchi.aliyun.com/competition/.

[27] 梁義钦, 赵司琦, 王海涛, 等. 两阶段分析的异常簇团宫颈细胞检测方法[J]. 哈尔滨理工大学学报, 2022, 27(2): 76-84.
Liang Y Q, Zhao S Q, Wang H T, et al. Two-stage detection method for abnormal cluster cervical cells[J]. Journal of Harbin University of Science and Technology, 2022, 27(2): 76-84.

[28] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[29] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision - ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12346: 213-229.

[30] Tian Z, Shen C H, Chen H, et al. FCOS: fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 9626-9635.

[31] Jiang Y Q, Tan Z Y, Wang J Y, et al. GiraffeDet: a heavy-neck paradigm for object detection[EB/OL]. (2022-02-09)[2023-05-06]. https://arxiv.org/abs/2202.04256.

[32] Sun P Z, Zhang R F, Jiang Y, et al. Sparse R-CNN: end-to-end object detection with learnable proposals[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 14449-14458.

[33] Cai Z W, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 6154-6162.

[34] Zheng A L, Zhang Y A, Zhang X Y, et al. Progressive end-to-end object detection in crowded scenes[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 847-856.

[35] Cao L, Yang J Y, Rong Z W, et al. A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening[J]. Medical Image Analysis, 2021, 73: 102197.

[36] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 618-626.

Automatic Identification of Cervical Abnormal Cells Based on Transformer

Zhang Zheng[1], Chen Mingxiao[1], Li Xinyu[1], Chen Yi[1], Shen Shuwei[2], Yao Peng[3*]

[1]*Department of Precision Machinery and Precision Instrumentation, School of Engineering Science, University of Science and Technology of China, Hefei* 230027, *Anhui, China*;

[2]*Suzhou Advanced Research Institute, University of Science and Technology of China, Suzhou* 215123, *Jiangsu,China*;

[3]*School of Microelectronics, University of Science and Technology of China, Hefei* 230027, *Anhui, China*

Abstract

Objective Cervical cancer is one of the most common malignant tumors and poses a serious threat to human health. However, because the onset of cervical cancer is gradual, early and effective screening is crucial. Traditional screening methods rely on manual examinations by pathologists, a process that is time-consuming, labor-intensive, error-prone, and often lacks an adequate number of pathologists for cervical cytology screening, making it challenging to meet the current demands for cervical cancer screening. In recent years, several deep-learning-based methods have been developed for screening abnormal cervical cells. However, because abnormal cervical cells develop from normal cells, they exhibit morphological similarities, making differentiation challenging. Pathologists typically need to reference normal cells in images to accurately distinguish them from abnormal cells. These factors limit the accuracy of abnormal cervical cell screening. This study proposes a Transformer-based approach for abnormal cervical cell screening that leverages the powerful global feature extraction and long-range dependency capabilities of Transformer. This method effectively enhances the detection accuracy of abnormal cervical cells, improving screening efficiency and alleviating the burden on medical professionals.

Methods This study introduces a novel Transformer-based method for abnormal cervical cell detection that leverages the powerful global information extraction capabilities of Transformer to mimic the screening process of pathologists. The proposed method incorporates two innovative structures. The first is an improved Transformer encoder, which consists of multiple blocks stacked together. Each block comprises two parts: a multi-head self-attention layer and a feedforward neural network layer. The multi-head self-attention layer captures the correlation of the input data at different levels and scales, enabling the model to better understand the structure of the input sequence. The feedforward neural network layer includes multiple fully connected layers and activation functions and introduces nonlinear transformations to help the model adapt to complex data distributions. We also introduce Depthwise (DW) convolution and Dropout layers to the encoder. DW convolution layer performs convolution operations with separate kernels for each input channel, capturing features within the channels without introducing inter-channel dependencies. Dropout layer reduces the tendency of neural networks to overfit the training data, thereby enhancing the generalization of the model to unseen data. Additionally, we design a dynamic intersection-over-union (IOU) threshold method that adaptively adjusts the IOU threshold. In the initial stages of training, the model can obtain as many effective detections as possible, whereas in later stages, it can filter out most false positive predictions, thereby improving the detection accuracy of the model. Using the proposed method, the model can obtain precise information regarding the location of abnormal cells.

Results and Discussions To validate the effectiveness of our proposed method, we compare it with common general-purpose object detection methods. The average accuracy (AP) and AP_{50} of our method are 26.1% and 46.8%, respectively, surpassing those of all general object detection models (Table 1). In particular, our method outperforms other comparative models by a significant margin in AP metrics, demonstrating that our model not only detects normal-sized targets but can also detect extremely small targets. Additionally, in a comparison with attFPN, a network specifically designed for abnormal cervical cell detection, our method surpasses attFPN in terms of AP by 1.1% (Table 2). Visual inspection of the detection results reveals that our method more accurately identifies target regions with lower false-positive and false-negative rates (Fig. 5). Ablation experiments indicate that

adopting the improved Transformer encoder method increases AP and AP_{50} by 1.8% and 2.3%, respectively, compared with the original model. The use of dynamic IOU thresholds results in a 0.6% increase in AP and a 0.9% increase in AP_{50} compared with the original model (Table 4). Furthermore, a comparison between the dynamic and fixed IOU thresholds in terms of loss and AP during the training process shows that the model with dynamic IOU thresholds experiences a faster loss reduction and achieves a higher AP in the later stages of training (Fig. 6).

Conclusions This study introduces an automatic identification method for abnormal cervical cells utilizing Transformer as the backbone. We further propose an enhanced Transformer encoder structure and a dynamically adjustable IOU threshold. Various comparative experiments on datasets demonstrate that the proposed method outperforms existing approaches in terms of accuracy and other metrics, thereby achieving precise identification of abnormal cervical cells. Through ablation experiments, it is proven that both proposed modules enhance the accuracy of the model in identifying abnormal cervical cells. Overall, the proposed method significantly improves the efficiency of medical image screening, saving medical time and resources, facilitating timely detection of cancerous lesions, and presenting considerable clinical and practical value. Future research may focus on the application of semi-supervised and unsupervised learning in the field of medical imaging to enhance image utilization, improve model detection performance, and better meet clinical requirements.

Key words medical optics; cervical cytopathological images; object detection; medical image processing