王鑫, 辛国江, 张杨,等. 基于 SA-YOLO X 的齿痕舌与裂纹舌综合检测模型[J]. 智能计算机与应用,2025,15(7):37-41. DOI:10.20169/j.issn.2095-2163.250705

基于 SA-YOLO X 的齿痕舌与裂纹舌综合检测模型

王 鑫, 辛国江, 张 杨, 朱 磊, 刘嵘澂
(湖南中医药大学 信息科学与工程学院, 长沙 410208)

摘 要: 舌诊是中医体系中的重要组成部分。针对传统中医舌象识别受医生个人因素影响,主观性强,当前舌象检测网络多用于提取单一目标的现状以及移动端应用的需求,本文提出一种开放环境下齿痕舌与裂纹舌的综合检测模型 SA-YOLO X(SE & ASFF based on YOLO X)。在主干网络引入挤压和激励注意力机制(Squeeze-and-Excitation,SE);在颈部网络添加自适应空间特征融合(Adaptively Spatial Feature Fusion,ASFF)。实验结果表明,本文提出的网络在保持高每秒帧数(Frames Per Second,FPS)的同时均值平均精度(mean Average Precision, mAP)值提升至 80.99%,相较于 YOLO X 提升了 9.26%,与双阶段目标检测网络相比具有更快的检测速度。

关键词: 裂纹舌; 齿痕舌; 挤压和激励; 自适应空间特征融合
中图分类号: TP391 **文献标志码**: A **文章编号**: 2095-2163(2025)07-0037-05

Comprehensive detection model of tooth mark tongue and crack tongue based on SA-YOLO X

WANG Xin, XIN Guojiang, ZHANG Yang, ZHU Lei, LIU Rongcheng
(School of Information Science and Engineering, Hunan University of Chinese Medicine, Changsha 410208, China)

Abstract: Tongue diagnosis is an important part of Chinese medicine system. In view of the fact that tongue image recognition in traditional Chinese medicine is highly subjective due to the personal factors of doctors, the current tongue image detection network is mostly used to extract a single target, and the demand of mobile applications, a comprehensive detection network SE & ASFF based on YOLO X(SA-YOLO X) in an open environment is proposed. The Squeeze-and-Excitation mechanism (SE) was introduced in the backbone network, and Adaptively Spatial Feature Fusion (ASFF) was added to the neck network. After experimental verification, the mean Average Precision (mAP) value of the network in this thesis increased to 80.99% while maintaining a high number of Frames Per Second (FPS), which increased by 9.26% compared with YOLO X. Compared with the two-stage target detection network, the detection speed is faster.

Key words: cracked tongue; dentate tongue; squeeze-and-excitation; Adaptively Spatial Feature Fusion

0 引 言

舌诊是中医体系中的重要组成部分,是中医诊断疾病的常用方法,医生通过观察舌形、舌色等特征了解身体机能与病理变化。中医体系中,舌形主要包括舌体的胖瘦、齿痕、老嫩、点刺及裂纹等。相比舌色与舌苔,舌形受外界的影响小,而且形态特征短期内变化较小,具有更强的客观性,在中医临床实践中具有重要研究价值[1]。传统舌诊通过医生目测观察,根据经验辨析舌象,诊断结果受医生的知识水平、思维能力和诊断技能的限制,同时也受到光照等外界条件的影响,缺乏一套客观评价标准,制约了舌诊的研究与应用。

齿痕舌与裂纹舌是舌诊的重要指标。齿痕舌临床主脾虚、湿盛之证,舌体边缘的牙齿印痕迹称为齿痕舌,多由舌体胖大受牙齿边缘压迫所致,是异常舌

基金项目: 国家级大学生创新训练项目(2022 批次); 湖南省一流本科课程(2021-896); 湖南省教改课题(HNJG-2021-0584)。
作者简介: 王 鑫(1997—),男,硕士研究生,主要研究方向:医学图像处理; 张 杨(1999—),男,硕士研究生,主要研究方向:医学图像处理; 朱 磊(1997—),男,硕士研究生,主要研究方向:医学图像处理; 刘嵘澂(2002—),女,本科生,主要研究方向:医学图像处理。
通信作者: 辛国江(1979—),男,博士,副教授,主要研究方向:医学图像处理。Email:lovesin_guojiang@126.com。
收稿日期: 2023-12-06

形的一种。齿痕舌识别的重点在于边缘特征的提取,以及对齿痕数和齿痕深度的测量。裂纹舌主阴虚、热盛之证,舌面间数量不等、深浅不一、形状各异的裂纹称为裂纹舌,精准捕捉裂纹所在区域并测量其形态与数量是裂纹舌客观化研究的重要内容。杨佳欣等[2]采用 Graham 扫描法提取齿痕凹陷特征,用支持向量机(SVM)算法判断齿痕有无,用道格拉斯—普朗克算法计算齿痕数量,在齿痕舌的综合检测上具有较好的性能;王鹏等[3]合理利用舌面三维点云数据,通过将三维点云处理技术与传统中医经验融合,提出了基于扩展快速点特征颜色直方图(Fast Point Feature&Color Histogram, FPFCH)特征值的欧式聚类舌体分割算法和基于法线区域分割的舌裂纹提取算法;Li 等[4]提出一种基于宽线检测器的统计形状特征(Wide Line Detector based statistical shape Feature, WLDF)来识别舌裂纹,并将 WLDF 特征输入到支持向量机中训练,最终取得 95%的准确率。机器学习具有可解释性且对数据量需求较少,但其性能高度依赖于特征的选择与设计,人工参与度高,且在模型复杂时容易过拟合,在新数据上的泛化能力差。随着计算设备性能提升和深度学习在医疗领域中的广泛应用,基于深度学习的舌特征检测也成为了一个研究热点。颜建军等[5]为解决齿痕舌识别准确率低的问题,提出一种基于二级分类器的齿痕舌分类模型,首先利用 DeepLabV3+分割模型剔除舌图像背景,再用随机森林分类齿痕舌,模型最终取得 93%的准确率;王一丁等[6]针对舌图像质量差异大、背景复杂且患者舌头颜色纹理存在差异的现状,提出一种基于深度学习的 、面向小目标数据集的舌裂分割算法,通过在 U-Net 中引入 SE 模块和 Focal Loss 函数,实现对舌裂纹的精准分割;LI M Y 等[7]提出了一种改进的 U-Net 网络,用于裂纹舌的语义分割,通过将 Global Convolution Network 模块引入 U-Net 的编码器部分,解决了编码器相对简单,无法提取相对抽象的高级语义特征的问题;刘梦等[8]基于 Faster R-CNN 深度学习和微调的研究方法对舌象局部特征进行提取,构建了齿痕舌与裂纹舌的特征综合提取模型,实现了局部特征的一体化识别。

尽管舌象识别取得了一定进展,但目前舌象的特征提取多集中于单一的特征。中医舌象的研究仍存在数据获取困难、样本类间不平衡,图像处理算法繁杂、迁移能力较差等问题。随着硬件性能提升,移动设备的广泛应用,图像识别、语音识别等技术在移动设备上的使用率大幅上升,在保证模型精度的前提下对检测速度同样具有较高的要求。

本文针对双阶段算法识别速度慢和单阶段算法检测精度低的问题,提出 SA-YOLO X 齿痕舌与裂纹舌综合检测模型。通过在特征提取网络中增加挤压和激励(SE)注意力模块,在 YOLO Head 模块中引入自适应空间特征融合(ASFF),在仅损失些许速度的情况下检测精度大幅提升。

1 基础理论

目标检测是计算机视觉领域的一个重要课题,早期的目标检测方法通常采用区域选择、特征手工提取和分类回归 3 个阶段实现,存在主观性强、泛化能力差等缺点。卷积神经网络(Convolutional Neural Networks,CNN)是深度学习在图像处理领域的主流模型,目前基于卷积神经网络的目标检测方法分为双阶段(two-stage)算法和单阶段(one-stage)算法。双阶段目标检测算法包含两个关键步骤即提取候选区域和对候选区域进行分类。常见的双阶段算法有 RCNN (Region - based Convolutional Neural Networks)、Fast R-CNN 和 Faster R-CNN,通常情况下双阶段算法具有更高的准确度。单阶段算法通过密集地在输入图像上设置多个锚框来预测目标的位置与类别,只需要一次前向传播过程即可同时完成回归和分类任务。常见的单阶段算法有 YOLO(You Only Look Once)和 SSD (Single Shot MultiBox Detector)。单阶段算法具有较高的实时性,适用于移动设备等对检测速度要求较高的场景。

YOLO X 最初由 GE Z 等[9]于 2021 年提出,与之前版本的 YOLO 结构相似,整体分为 3 个部分:主干特征提取网络、加强特征提取网络和分类器与回归器。主干提取网络通过逐层卷积与池化操作提取图像中的特征,构成特征层,主干网络最终输出 3 个特征层用于下一步网络构建;特征增强网络对主干部分获得的 3 个特征层信息进行融合,从而增强特征信息的表达;分类器与回归器将特征层视作特征点的集合,判断特征点是否有物体与之对应,每个特征点具有数个特征通道。

YOLO X 的主干特征提取网络为 CSPDarknet(Cross Stage Partial Darknet, CSPDarknet),主干部分沿用 YOLOv5 的 Focus 模块,其实现原理如图 1 所示。通过每隔一个像素取一个值,获得 4 个独立的特征层,4 个特征层沿通道方向堆叠,将宽高信息集中到通道信息中,保留图像信息的同时减少参数计算。

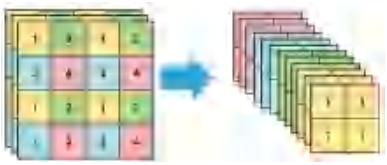

图 1　Focus 模块实现原理

Fig. 1　Focus module implementation principle

YOLO X 的加强特征提取网络为路径聚合特征金字塔(Path Aggregation Feature Pyramid Network，PAFPN)，结构如图 2 所示。dark3、dark4、dark5 是 CSPDarknet 下 3 层获得的有效特征层，在 PAFPN 中高层的特征信息先通过上采样进行传递融合，然后采用下采样的融合方式获得加强特征并传入 Yolo Head。concat 结构能有效解决 FPN 中浅层特征在传递过程中的信息丢失。

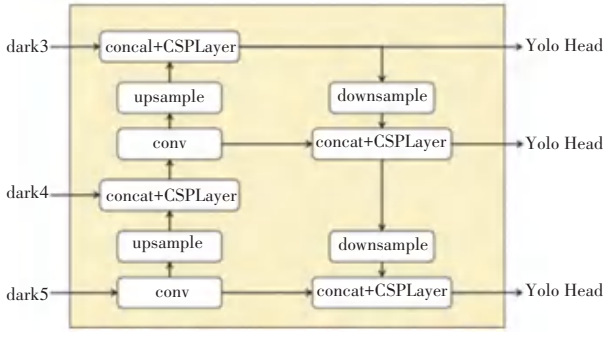

图 2　路径聚合特征金字塔结构

Fig. 2　Path aggregation feature pyramid network structure

YOLO X 的 Yolo Head 与以前版本的 YOLO 不同，将分类和回归分开实现，在最后预测时再重新整合。解耦头结构如图 3 所示，对于每一个特征层，均可以获得 3 个预测结果：Cls 用于判断每个特征点所包含的物体种类；Reg 用于判断每个特征点的回归参数，经过调整获得预测框；Obj 用于判断每一个特征点是否包含目标；最后将输出特征图以 Reg、Obj、Cls 的顺序沿通道方向拼接。

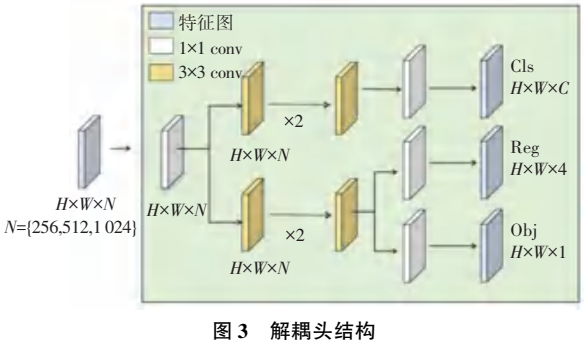

图 3　解耦头结构

Fig. 3　Decoupled head structure

YOLO X 中物体的真实框由其中的特征点来预测。正样本的特征点具有两个特点：在真实框中，距离物体中心在一定半径内。损失函数用来表示预测结果与真实结果的差异。损失函数由 3 部分组成：$Loss_{Reg}$、$Loss_{Obj}$、$Loss_{Cls}$。Reg 部分：取出每个真实框对应特征点的预测框，利用真实框与预测框计算 IOU 损失，构成 $Loss_{Reg}$；Obj 部分：真实框对应的特征点为正样本，剩余的特征点为负样本，根据正负样本和特征点是否包含物体的预测结果计算交叉熵损失，构成 $Loss_{Obj}$；Cls 部分：获取真实框对应的特征点后，取出特征点的种类预测结果，根据真实框的种类和特征点的种类预测结果计算交叉熵损失，构成 $Loss_{Cls}$。

2　SA-YOLO X 网络结构

本文在 YOLO X 的基础上，引入 SE 注意力模块和 ASFF 特征融合机制构建了一种用于实现舌图像齿痕与裂纹检测的特征提取网络模型，网络结构如图 4 所示，输入图片的尺寸为 640×640，图像经过特征提取网络，提取出特征并最终获得 3 个有效特征层，3 个特征层由 SE 注意力模块增强后通过上采样与下采样构建 PAFPN 进一步强化特征信息；ASFF 将增强的不同尺度特征信息自适应融合并配置权重后分别传入不同尺度的 Yolo Head 中，得到目标预测结果。

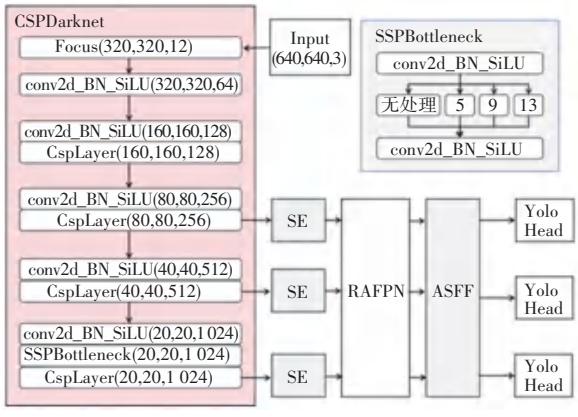

图 4　SA-YOLO X 网络结构

Fig. 4　SA-YOLO X network architecture

2.1　SE 注意力机制

SE(Squeeze-and-Excitation)注意力机制是一种用于增强卷积神经网络表达能力的方法[10]，结构如图 5 所示。SE 包含挤压与激励两部分：在挤压阶段通过全局平均池化将每个通道的特征图转换为单一的值；在激励阶段使用全连接层和激活函数学习通道权重。通道权重与原始特征图相乘，得到重新分配权重的特征图。引入注意力机制，模型能够自适应地选择和加权特征图，抑制特征图中的无用信息，进而提高模型的特征表达能力。

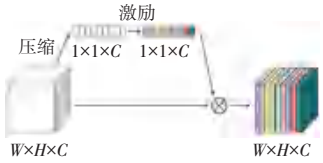

图 5 挤压激励注意力模块的结构

Fig. 5 Squeeze-and-Excitation Attention Module structure

2.2 自适应空间特征融合

自适应特征融合(Adaptive Spatial Feature Fusion,ASFF)用于目标检测中的特征融合[11]。在卷积神经网络中,特征金字塔(Feature Pyramid Networks,FPN)是实现目标尺度不变性的有效方法[12]。对于 YOLO、SSD 等单阶段目标检测网络,不同尺度特征之间的不一致性是限制其性能的主要因素。大目标通常与深层特征图相关联、小目标与浅层特征图相关联,当图像中同时存在尺度相差较大的目标时,检测目标在不同层级的特征图间存在冲突,干扰模型在训练期间的梯度计算,导致特征金字塔的有效性降低。

ASFF 能自适应地学习每个尺度特征图的空间融合权重,其结构如图 6 所示。特征融合过程如公式(1)所示:

$$y^l = \alpha^l \cdot x^{1\to l} + \beta^l \cdot x^{2\to l} + \gamma^l \cdot x^{3\to l} \quad (1)$$

其中,$x^{1\to l}$、$x^{2\to l}$、$x^{3\to l}$表示为 feat1、feat2、feat3 到第 l 层的映射,不同层特征图通过上采样或下采样调整尺寸与通道数;α^l、β^l、γ^l 为对应权重参数,由各层特征图经过 1×1 卷积获得,并通过 Softmax 将范围限制在[0,1]且和为 1。权重与相应特征图相乘,实现自适应融合。

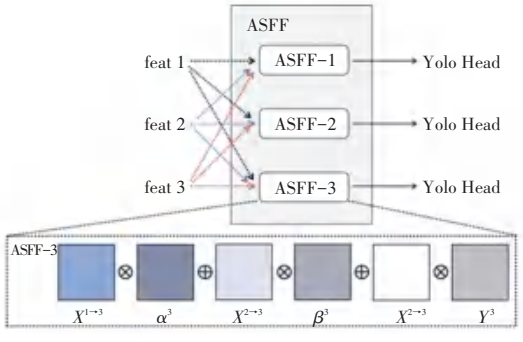

图 6 自适应空间特征融合结构

Fig. 6 Structure of Adaptive Spatial Feature Fusion

3 实验分析

3.1 数据集制作

本实验舌象数据集样本初始容量为 765 例,包含齿痕舌、裂纹舌两类,按照 1∶3 的比例进行数据增强将样本扩充至 2 295 例,用 LabelImg 软件对数据进行标注。预处理后的舌图像如图 7 所示。

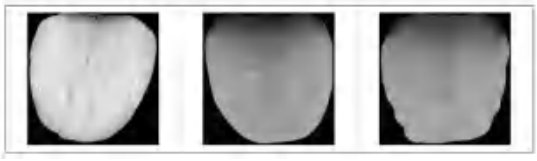

图 7 预处理后的舌象数据

Fig. 7 Tongue image data after preprocessing

3.2 模型训练

舌象数据集按 8∶1∶1 的比例划分为训练集、验证集、测试集,训练集和验证集完成模型学习、模型参数优化,测试集验证模型的性能。

模型超参数设置如下:(1)初始学习率为 0.001,学习率衰减采用余弦退火算法;(2)优化器为 Adam;(3)训练批次 epoch 为 200;(4)批大小 batchsize 为 8。

3.3 横向对比

本次实验采用均值平均精度(mean Average Precision,*mAP*)、每秒帧数(Frames Per Second,*FPS*)作为评价指标,并与 Faster R-CNN、YOLOv5、YOLOv7 作性能对比,如图 8 所示。左侧坐标轴表示平均精准度(Average Precision,*AP*);右侧坐标轴表示每秒帧率(Frame Per Second,*FPS*),即每秒可以处理的图片数量,当 *FPS* > 30 时具备实时性;*AP* - C 表示裂纹舌的平均精准度,*AP* - D 表示齿痕舌的平均精准度,*AP* 的计算如下式即对 pr 曲线取积分求准确度的平均值。*mAP* 为 *AP* - C 与 *AP* - D 的平均值。

$$AP = \int_0^1 p(r)\,\mathrm{d}r \quad (2)$$

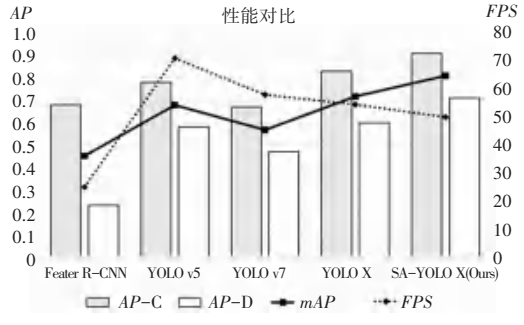

图 8 模型性能对比

Fig. 8 Model performance comparison

相较于双阶段的 Faster R-CNN,单阶段的 YOLO 算法具有更快的处理速度,且由于检测目标的复杂度较低,参数量较大的 Faster R-CNN 易过拟合,在测试集中 YOLO 的 *mAP* 比 Faster R-CNN 更高,且类间差距更小。裂纹舌特征明显、覆盖范围大

且多集中于图像中央；齿痕舌区域较小，分布于舌体边缘，检测难度较大，各算法裂纹舌的检测准确度均高于齿痕舌。在本次检测任务中，YOLO X 与 YOLOv5 的检测效果更好，同时 YOLOv5 具有最快的检测速度，SA-YOLO X 的检测精度最高，比 YOLOv5 和 YOLO X 高约 10%，且检测速度表现良好。

3.4　消融实验

SA-YOLO X 综合检测模型在 YOLO X 的基础上引入 SE 注意力机制和 ASFF 特征融合机制，为了系统性地判断各改进机制的作用，本文设计了消融实验，结果见表 1，可见新模块的引入均会导致检测速度下降，但均能达到实时检测的标准，符合移动设备应用要求；ASFF 的引入，舌体特征检测精度上提高 4%，表明 YOLO X 在本次检测任务中不同的特征层间存在冲突，而 ASFF 有效地将各层特征图中的信息融合；引入 SE 模块后 *mAP* 相较于只引入 ASFF 的模型上升 5.26%并超过 80%，而 *FPS* 仅损失 2%，表明 SE 注意力机制在该模型中成功抑制无用信息，为后续的 PAFPN 输入更优质的特征信息，提高了网络的特征检测能力。

表 1　消融实验

Table 1　Ablation experiment

模型	*mAP*/%	*FPS*
YOLO X	71.73	54.48
YOLOX+ASFF	75.73	51.95
YOLO X+ASFF+SE	80.99	49.95

本文 SA-YOLO X 模型的检测结果如图 9 所示。从实验结果看，该模型不受齿痕、裂纹的位置和形态影响，能准确检测出不同尺度的目标。裂纹舌的检测效果更优，置信度均在 0.75 以上；舌体边缘不同形态的齿痕均能有效检测，检测出的齿痕满足病理判断的要求，但齿痕舌检测的置信度低于裂纹舌且存在部分漏检现象，是后续研究的重点内容。

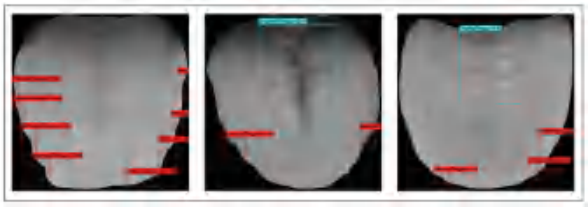

图 9　SA-YOLO X 的检测结果

Fig. 9　Detection result for SA-YOLO

4　结束语

本文针对传统中医舌象识别主观性强和易受光照环境影响的问题，以及当前移动端应用部署的趋势，提出一种引入 SE 注意力机制和 ASFF 特征融合机制的 SA-YOLO X 模型，在开放环境下获取的舌图像检测任务中平均检测精度为 80.99%，相较 YOLO X 提升 9.26%，*FPS* 为 49.95。与其他深度学习模型相比，该模型具备一定检测效率的同时在舌象形态特征的综合识别具有良好表现，齿痕舌与裂纹舌检测的 *mAP* 相差较小；齿痕舌与裂纹舌的某些特征在检测过程中存在冲突，影响模型的识别性能，SE 注意力机制的引入，能有效抑制无用信息，提升 RAFPN 特征加强网络在模型中的性能表现；ASFF 特征融合机制能有效缓解模型中的特征冲突，自适应地学习不同特征层的权重。在齿痕舌的检测中存在部分漏检，模型对齿痕舌的特征捕捉不全，仍需要后续改进，进一步缓解不同尺度目标间特征学习的冲突，并探究舌象的舌色对舌形特征检测的影响。

参考文献

[1] 夏雨墨，王庆盛，冯晓，等. 舌形特征的提取与分析技术研究及其在临床诊断中的应用进展[J]. 世界中医药，2023，18(14)：2059-2063.

[2] 杨佳欣，韩东，董新明，等. 基于形态特征提取的中医齿痕舌客观化研究[J]. 激光与光电子学进展，2022，59(11)：365-373.

[3] 王鹏，杨文超，孙长库，等. 舌面彩色三维点云的舌体分割及舌裂纹提取[J]. 红外与激光工程，2017，46(S1)：88-95.

[4] LI X，WANG D，CUI Q. WLDF：Effective statistical shape feature for cracked tongue recognition[J]. Journal of Electrical Engineering & Technology，2017，12(1)：420-427.

[5] 颜建军，李东旭，郭睿，等. 基于二级分类器的齿痕舌分类模型研究[J]. 中华中医药杂志，2022，37(4)：2181-2185.

[6] 王一丁，孙常浩，崔家礼，等. 基于深度学习的舌裂分割算法研究[J]. 世界科学技术-中医药现代化，2021，23(9)：3065-3073.

[7] LI M Y，ZHU D J，XU W，et al. Application of u-net with global convolution network module in computer-aided tongue diagnosis[J]. Journal of Healthcare Engineering，2021，2021：5853128. DOI：10.1155/2021/5853128.

[8] 刘梦，王曦廷，周璐，等. 基于深度学习与迁移学习的中医舌象提取识别研究[J]. 中医杂志，2019，60(10)：835-840.

[9] GE Z，LIU S，WANG F，et al. Yolox：Exceeding yolo series in 2021[J]. arXiv preprint arXiv，2107.08430，2021.

[10] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2018：7132-7141.

[11] LIU S，HUANG D，WANG Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv，1911.09516，2019.

[12] LIN T Y，DOLLÁR P，GIRSHICK R，et al. Feature pyramid networks for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2017：2117-2125.