

文章编号:1007-2780(2026)02-0240-13

基于 MLC-YOLO 的折纸动作关键目标检测方法

陈宇聪, 何 宏*, 李泽旭, 徐楚迪
(上海理工大学 健康科学与工程学院, 上海 200093)

摘要: 由于简易精神状态检查(CMMSE)量表中“三步指令”的人工评分存在评分者间及自身内部差异,且过程耗时,提出一种基于MLC-YOLO的目标检测方法,旨在实现该动作的自动精准检测。首先,对人体关键部位和纸张进行多级分类目标检测,采用同类极大置信度选择简化后处理,保证同类目标的唯一性;其次,设计自适应空间信道解耦模块,实现高效降采样;然后,在C3中引入幽灵卷积和小波变换,提升多尺度特征提取的效率与能力;最后,针对小目标检测层,引入空间和通道协同注意力,提高复杂场景下的准确率与召回率。实验结果表明:该方法的mAP95达61.8%,较原模型提升5.6%,参数量减少51.52%,证明了方法的有效性,为目标唯一和多级分类的目标检测问题提供了新思路,为MMSE中的“三步指令”的自动评分提供有效的目标检测方法。

关键词: 深度学习;目标检测;多级分类;轻量化网络

中图分类号: TP391.41 **文献标识码:** A **doi:** 10.37188/CJLCD.2025-0255 **CSTR:** 32172.14.CJLCD.2025-0255

MLC-YOLO based key object detection method for origami action

CHEN Yucong, HE Hong*, LI Zexu, XU Chudi

(School of Health Science and Engineering, University of Shanghai for Science and Technology,
Shanghai 200093, China)

Abstract: To address the inter-rater and intra-rater variability, as well as the time-consuming nature of manual scoring for the “three-step command” in the Mini-Mental State Examination(MMSE), an object detection method based on MLC-YOLO is proposed to achieve automatic and precise detection of this action. First, multi-level classification object detection is performed on key human body parts and paper, and a same-class maximum confidence selection strategy is adopted to simplify post-processing and ensure the uniqueness of targets of the same class. Second, an adaptive spatial-channel decoupling module is designed to achieve efficient downsampling. Then, Ghost convolution and wavelet transform are introduced into the C3 module to enhance the efficiency and capability of multi-scale feature extraction. Finally, for the small object detection layer, spatial and channel collaborative attention is introduced to improve

收稿日期:2025-12-28;修订日期:2026-01-23.

基金项目: 国家科学技术部项目(No.G2021013008);中国高校产学研创新基金(No.2023RY011);上海理工大学医工交叉重点创新项目(No.1022308502);华为AI算力加速计划

Supported by Project of the Ministry of Science and Technology of People's Republic of China(No. G2021013008); China Higher Education Industry-University-Research Innovation Fund(No.2023RY011); Key Project of Crossing Innovation Medicine and Engineering, University of Shanghai for Science and Technology(No.1022308502); Huawei AI Computing Power Acceleration Program

*通信联系人, E-mail: hehong@usst.edu.cn

precision and recall in complex scenes. Experimental results indicate that the mAP95 of the proposed method reaches 61.8%, an increase of 5.6% compared to the original model, while the parameter count is reduced by 51.52%. It proves the effectiveness of the method, offering a new approach for object detection problems involving unique targets and multi-level classification, and providing an effective object detection method for the automatic scoring of the “three-step command” in MMSE.

Key words: deep learning; object detection; multilevel classification; lightweight network

1 引言

简易精神状态检查量表(Mini-Mental State Examination, MMSE)是一种用于评估个体认知功能的简短标准化测试。目前,MMSE已广泛用于临床实践,在痴呆筛查中具有较高的敏感性^[1]。在MMSE中,语言部分的测试包括听理解、阅读理解和书写等,其中听理解又称“三步指令”,用于评估执行功能和注意力。目前,最常用的中文版MMSE是北京版和上海版。前者的“三步指令”内容为:“用右手拿着这张纸,用两只手将它对折起来,放在您的左腿上。”后者指令相似,仅最后一步为“放在您的大腿上”。该测试要求被试执行三步指令,每步正确得1分,共3分。

传统MMSE评估诊断认知障碍或认知功能下降趋势时,需患者多次接受医疗人员手动评测。不同评分者在判别纸张对折角度等细节时存在主观差异,同一评分者存在不同时期的内部差异^[2]。因此,研究一种自动评分方法可以提高评分效率和一致性。自动评分方法主要分两种:一是基于人体关键点检测,通过单人姿态估计^[3-4]和动作识别模型判断是否完成指定动作;二是基于目标检测^[5-7],检测纸张、手、大腿等目标,依据纸张形态及与手、腿的空间关系判断动作完成情况。关键点检测虽然能识别人体关键点位置,但对于复杂的操作动作与实现方式,尤其是在相对空间位置和物体操作的细节上,无法准确捕捉动作的完整性和准确性。相比之下,目标检测能够精确地定位纸张、手和大腿等目标,还能同时判断纸张的状态信息,从而更直观地判断动作是否完成。

传统目标检测需多步处理,如先进行图像预处理、特征提取,再进行目标分类和位置回归。这些步骤计算资源消耗大,无法在嵌入式设备上实时运行。自区域卷积神经网络(RCNN)被提出后,基于深度学习的目标检测成为主流。起初,采用

双阶段方法,即显式地将检测过程分为候选框选取和目标框判断,其检测精度较高,但速度较慢。随后出现检测速度更快的单阶段方法,最有代表性的是YOLO。YOLO将目标检测视为回归问题,采用单一网络结构,以整幅图像为输入直接预测特征,同时输出多个边界框和概率,一次前向传播即可完成目标的位置检测与分类。随后,YOLOv2将主干网络改为参数量更小的DarkNet-19,使用基于尺寸聚类的先验框、多尺度训练和特征融合等方法。YOLOv3则将主干网络改为Darknet-53,加深网络并以卷积层替代最大池化层,引入残差思想建立跨尺度预测模块。随着计算机视觉与光学成像技术^[8]的快速发展,YOLO系列算法在自然图像领域应用更加广泛^[9],研究者也推出了YOLO Nano^[10]和Tiny-YOLO^[11]等更轻量化的改进版本。文献[12]构建深度过参数化高效聚合网络,嵌入坐标注意力机制及采用Ghost轻量化网络,提升密集遮挡目标的识别精度与效率。针对复杂背景和外观相似等小目标检测痛点,文献[13]通过将融合全局信息和双域注意力机制等引入YOLOv8以优化性能;文献[14]基于YOLOv11嵌入多尺度卷积注意力机制,融合动态检测头等解决检测精度低、漏检误检多的问题。

尽管通用目标检测算法发展迅速,但在将其应用于MMSE“三步指令”自动化评分时,面临着两大特殊挑战。第一,通用模型擅长区分不同物体,但难以识别同一物体在不同操作阶段的细微形态变化,如纸张折叠前后的状态,而这恰恰是判断患者执行功能是否受损的关键依据。第二,为保护患者隐私,筛查视频数据不宜上传云端处理;同时,为了在社区医院广泛推广,系统必须能在低成本的嵌入式设备上实时运行。现有的SOTA模型往往参数量过大,难以满足这种实时离线的需求。因此,本文设计了一种兼顾细粒度状态识别精度与极致轻量化的检测框架,为认知障碍的自

动化、标准化和隐私化筛查提供可行的技术路径。

YOLOv5n 凭借低参量与计算量适配资源受限的嵌入式环境,拥有经长期验证的成熟部署生态以降低工程风险,且通过自适应锚框机制针对特定小样本数据实现了更优的收敛速度与小目标召回率。这使 MMSE 自动评分系统能部署在嵌入式医疗设备中,避免依赖网络,从而增强系统的稳定性,保护患者隐私,符合医疗数据保护法规^[15]。为实现 MMSE 中“三步指令”——折纸识别任务的自动评分,本文设计了一套基于 YOLOv5 的 MMSE 中“三步指令”任务自动评分方法的嵌入式系统。针对该系统,提出了 MLC-YOLO (Multi Level Classification YOLO)。

综上所述,本研究旨在解决两个核心矛盾。一是通用检测模型难以区分折纸任务中纸张状态的细微形态变化;二是数据隐私保护与设备低算力限制之间的矛盾。因此,本研究的核心目标是构建一种兼顾高细粒度识别与轻量化的目标检测框架,在低计算资源下,对折叠纸张等关键小目标实现高召回率和实时推理,从而为离线环境下的认知障碍筛查提供可靠的技术支撑。

2 方 法

YOLOv5 按模型大小分为 5 个版本,本文选用复杂度最低的 YOLOv5n 作为基准模型,参数量和计算量分别仅为 YOLOv8n 的 58.96% 和 52.44%。YOLOv5n 的网络结构分为输入端(Input)、主干网络(Backbone)、颈部网络(Neck)和头部网络(Prediction) 4 个部分。输入端接受一张 RGB 图像,经数据增强后输入主干网络。其 6.0 版本用卷积层替换切片操作,减轻计算负担;用高效快速空间金字塔池化模块取代传统模块,进一步提升。7.0 版本主干网络含 Conv、C3 和 SPPF 模块,其中 Conv 模块由卷积、批归一化和 SiLU 激活函数组成,用于特征提取;C3 模块采用残差结构和若干个迭代的瓶颈网络组成,增加网络深度和感受野;SPPF 模块作为主干网络的输出端,使用不同大小的池化核聚合不同尺度的特征信息。颈部网络采用路径聚合特征金字塔网络,引入自底向上的路径,使得底层信息更容易向上传递。头部网络负责预测多尺度的目标类别、边界框位置和置信度,结合 CIoU 计算目标损失,在验证和

推理阶段通过非极大值抑制优化预测结果。

本文以 YOLOv5n 为基础模型,提出多级分类目标检测——MLC-YOLO。首先,缩减输出端的预测头尺度,构建二级分类检测的输出结构,并改进非极大值抑制,进一步简化了推理流程。其次,提出双分支异构协同特征压缩降采样替换跨步卷积,不仅降低了计算成本,还保留了多样化信息。再者,在主干网络引入 C3Ghost,减少了生成浅层冗余特征的计算量。在颈部网络,针对图像中多样的人物服饰和背景,引入 C3WT,增强了网络在复杂场景下的特征提取能力;此外,在小目标检测层添加空间和通道协同注意力,加强模型对目标特征的关注度,从而有效解决小目标的误检和漏检问题。MMLC-YOLO 网络结构如图 1 所示,折纸动作过程关键目标检测方法整体架构如图 2 所示。

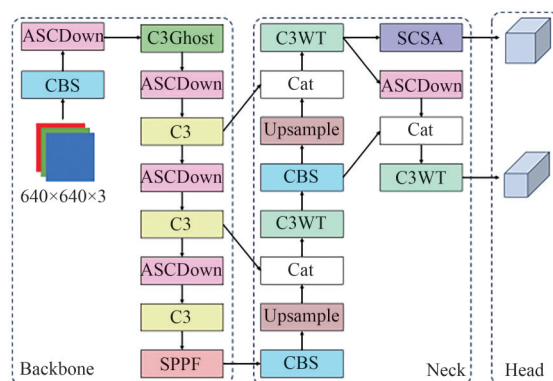


图 1 MLC-YOLO 网络结构示意图

Fig. 1 Schematic of MLC-YOLO network structure

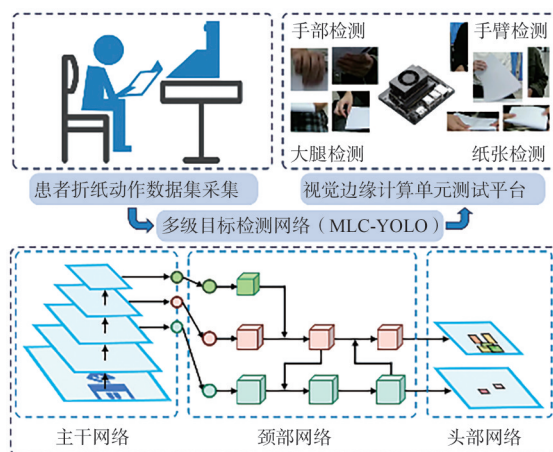


图 2 折纸动作过程关键目标检测方法整体架构

Fig. 2 Overall framework of key object detection method in origami action process

2.1 头部网络

2.1.1 增加二级类别的输出通道

YOLOv5的头部网络接受颈部网络中3个不同尺寸的张量,分别使用 1×1 的卷积操作,输出张量的宽高的下采样率分别32、16和8。输出通道数根据类别数的不同而改变,具体计算方法如式(1)所示:

$$n_o = (5 + n_c) \times n_a, \quad (1)$$

其中: n_o 为输出通道数, n_c 为类别数, n_a 为单个尺度下的锚框数量(默认为3)。

在本文的二级分类任务中,增加了子类别后输出通道数的计算方法要做出相应的改变,如式(2)所示:

$$n_o = (5 + n_c + n_s) \times n_a, \quad (2)$$

其中: n_o 为输出通道数, n_c 为一级类别数, n_s 为所有二级类别数, n_a 为单个尺度下的锚框数量(默认为3)。

2.1.2 减少输出尺度

为使模型聚焦于关键语义信息,本文对路径聚合特征金字塔网络进行改进。原结构融合3个尺度的特征图,分别负责检测大、中、小感受野下的目标,对于检测不同大小的目标有良好的效果。在人体关键部位和纸张的检测中,摄像头与患者位置相对固定,且目标在图像中占比较小,层级过深会导致细粒度信息在传递过程中逐渐丢失。对此,本文移除大目标检测层P5后,模型更聚焦在图像中小尺度的关键信息,形成专为中小目标检测的FPN-sm结构。优化后的模型不仅提升了精度,且参数量从1.78M降至1.32M,实现了显著的轻量化。

2.1.3 非极大值抑制改进

在一级类别中,所有类别都具有唯一性,即在一个画面中同一类别不应存在多个检测框。针对这个问题,需要对NMS进行改进。具体地说,NMS不再根据置信度得分最高的框与其他候选框的IoU去抑制重叠框,而是简化为对每个一级类别选择最高置信度框的同类极大置信度选择(Class-wise Max Confidence Selection, CMCS)。对本文任务而言,CMCS使得过滤后的预测框更符合实际,且极大减少了后处理时间。该方法在训练时采用一个标签对应多个预测框的分配策略,使训练过程有丰富的监督信号,同时在推理

时无需NMS,加快推理速度。

2.2 损失函数

在损失函数方面,本文在原损失函数的基础上增加了二级类别损失,改进的损失函数如式(3)所示:

$$L = \lambda_{\text{box}} \cdot l_{\text{box}} + \lambda_{\text{obj}} \cdot l_{\text{obj}} + \lambda_{\text{cls}} \cdot l_{\text{cls}} + \lambda_{\text{subcls}} \cdot l_{\text{subcls}}, \quad (3)$$

式中: λ_{box} 、 λ_{obj} 、 λ_{cls} 和 λ_{subcls} 分别是定位损失、置信度损失、一级类别损失和二级类别损失项的权重系数,用于平衡各部分的贡献; l_{box} 、 l_{obj} 、 l_{cls} 和 l_{subcls} 分别为位损失、置信度损失、一级类别损失和二级类别损失。

本文增加的二级分类损失 l_{subcls} 是基于二元交叉熵(Binary Cross-Entropy, BCE)损失计算的,该损失如式(4)和式(5)所示:

$$p_{\text{subcls}}^i = \text{Sigmoid}(x_{\text{subcls}}^i) = \frac{1}{1 + e^{-x_i}}, \quad (4)$$

$$l_{\text{subcls}} = - \sum_{i=1}^{n_s} [y_{\text{subcls}}^i \log p_{\text{subcls}}^i + (1 - y_{\text{subcls}}^i) \cdot \log(1 - p_{\text{subcls}}^i)], \quad (5)$$

式中: n_s 为二级类别数, x_{subcls}^i 为当前二级类别预测值, p_{subcls}^i 为经过激活函数后得到的当前二级类别的概率, y_{subcls}^i 为当前二级类别的标签值。

2.3 主干网络

2.3.1 ASCDown降采样模块

在YOLOv5中,使用跨步卷积进行降采样会导致细节信息丢失,尤其在特征信息少的纸张检测中,模型对细节的捕捉能力下降。本文结合YOLOv9^[16]提出的ADown与YOLOv10^[17]提出的空间通道解耦下采样SCDown,对下采样进行轻量化优化,提出ASCDown(Adaptive Spatial-channel Decoupled Down-sampling),在保留更多细节特征的同时,进一步减少降采样的参数量。

ASCDown创新地采用双分支异构协同特征压缩结构,通过差异化采样策略与深度可分离卷积优化,实现特征图的高效降维与多粒度特征融合,其结构如图3所示。图3中, k 、 s 代表卷积层或池化层的两个最核心参数。 k 代表卷积核的尺寸,例如 $k2$ 表示使用的是2像素 $\times$ 2像素的卷积核; s 代表步长,例如 $s1$ 表示卷积核每次移动1个像素。首先,输入特征图经平均池化层,对局部邻域特征进行线性平滑的同时,避免双分支特征信息偏斜。池化后的特征图沿通道维度均分为两部分,

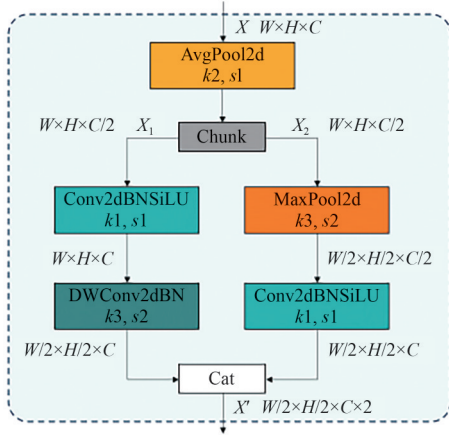


图3 ASCTDown降采样模块结构

Fig. 3 ASCTDown downsampling module structure

便于后续计算与差异化采样。

分支 X_1 的作用在于高效全局特征提取。子特征图通过 1×1 卷积调整通道数,实现轻量化特征映射;再使用深度可分离卷积,进行空间下采样,扩大感受野并减少参数冗余。

分支 X_2 的作用为增强局部显著特征。子特征图先经最大池化层,降低空间维度,提取显著局部特征,去除冗余信息;再通过 1×1 卷积调整通道数,引入非线性并促进跨通道信息交互。

分支 X_1 侧重局部到全局特征递进提取,实现显式下采样。分支 X_2 则重点关注局部细微特征的提取,隐式地压缩空间维度。二者互补减少信息丢失,避免信息单一化。最后,将双分支输出沿通道维度拼接,综合多粒度特征表达。

2.3.2 浅层特征提取轻量化

GhostConv^[18] 是一种高效轻量化卷积模块,利用轻量化的特征映射生成机制,减少冗余的计算量。网络在初期特征提取时,输入特征图尺寸较大,传统卷积的计算量大,且特征提取后存在大量冗余,在首层 C3 模块中引入 GhostConv 结构,能够有效减少冗余特征的计算量,使得网络能在不丢失深层的精细语义特征的同时,减少浅层冗余特征的计算量。其结构如图 4 所示。

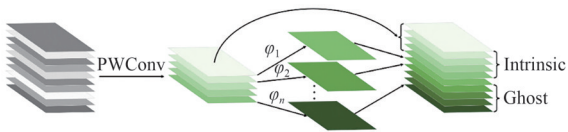


图4 Ghost卷积模块结构

Fig. 4 GhostConv module structure

这种先生成本征图 (Intrinsic), 再对本征图进行线性变换 ($\varphi_1, \varphi_2, \dots, \varphi_n$) 产生 Ghost 图, 最后将 Intrinsic 与 Ghost 拼接生成完整特征图的卷积方法, 在保证网络精度的同时有效减少了模型的复杂度。

2.3.3 C3WT 模块

卷积层对高频信息敏感,但对低频信息的捕获能力较弱^[19]。因此,传统的 C3 在纸张与人体关键部位检测中,难以捕捉形状及非纹理特征等低频信息。此外,复杂背景与患者多样服饰会降低模型的准确性和鲁棒性。处理此类复杂场景需扩大感受野以增强检测能力,但构建更大卷积核会增加计算复杂度和参数量,还会导致过度参数化,使模型在达到全局感受野前性能饱和。为克服上述挑战,本文在颈部网络引入小波卷积^[20],在融合不同层级特征时兼顾低频信息提取,只使用很少的参数量使卷积在大感受野中专注于不同频率带,避免过度参数化,从而保持整体高效性。

小波卷积是一种结合小波变换与卷积的混合域特征提取方法,其核心思想是通过小波变换的多尺度分解能力增强传统卷积的特征表示。该模块使用以下滤波器构建二维可分离滤波器:

$$f_{LL} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, f_{LH} = \begin{bmatrix} 0.5 & -0.5 \\ 0.5 & -0.5 \end{bmatrix},$$

$$f_{HL} = \begin{bmatrix} 0.5 & 0.5 \\ -0.5 & -0.5 \end{bmatrix}, f_{HH} = \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{bmatrix},$$
(6)

其中: f_{LL} 是一个低通滤波器, f_{LH} 、 f_{HL} 和 f_{HH} 是一组高通滤波器。

这些滤波器将输入特征图 X 进行小波分解,得到 4 个子带:

$$[X_{LL}, X_{LH}, X_{HL}, X_{HH}] = \text{WTConv2d}([f_{LL}, f_{LH}, f_{HL}, f_{HH}], X), \quad (7)$$

其中: X_{LL} 是 X 的低频子带,捕捉大感受野下的低频特征,通常对应于图像中的平坦区域、背景或低纹理部分,能够有效保留全局信息; X_{LH} 、 X_{HL} 和 X_{HH} 分别是其水平、垂直和对角线的高频子带,捕捉的是细节、水平和垂直边缘等高频特征。

接着,使用逐深度卷积对每个通道独立卷积,针对不同频带学习专用滤波器,既减少了参数量又避免了跨子带信息干扰。再进行通道级

缩放,平衡不同频率子带的贡献,避免单独对每个频带归一化带来的不一致性:

$$[X'_{LL}, X'_{LH}, X'_{HL}, X'_{HH}] = \alpha \cdot \text{DWConv2d}(W, [X_{LL}, X_{LH}, X_{HL}, X_{HH}]), \quad (8)$$

为了避免传统小波变换过程中递归分解所带来的复杂计算,本文在对不同频率特征图进行一次性卷积与缩放后,直接使用逆小波变换恢复原特征图 X 的尺寸。由于式(9)中滤波器的核为标准正交基,逆小波变换可以通过转置卷积实现:

$$Y = \text{IWTConv2d}([f_{LL}, f_{LH}, f_{HL}, f_{HH}], [X'_{LL}, X'_{LH}, X'_{HL}, X'_{HH}]). \quad (9)$$

最后,基于小波变换的线性性,小波变换可与卷积层进行如公式(10)所示的特征融合,而保证反向传播时无梯度畸变:

$$Z = \text{DWConv2d}(W, X) + Y, \quad (10)$$

其中: W 为 5×5 的卷积核, Z 为融合空间域与频域的混合域特征图。

C3WT模块结构如图5所示,其双路径设计在保持C3轻量化的同时,显著增强低频特征的代表能力,扩大模型感受野。其中,小波路径提取多尺度频域特征,而卷积路径捕获局部空间特征,使得该融合结构兼具局部细节与全局上下文感知能力。尤其在应对复杂服饰与背景情况下,该结构能更加精准地提取和融合多层次的信息,从而提高对服饰特征的识别能力,降低背景噪音的干扰。

2.3.4 小目标检测层引入SCSA

不同目标的图像空间特征与语义信息存在差异。人体肢体位置、姿势、角度多变,纸张则受

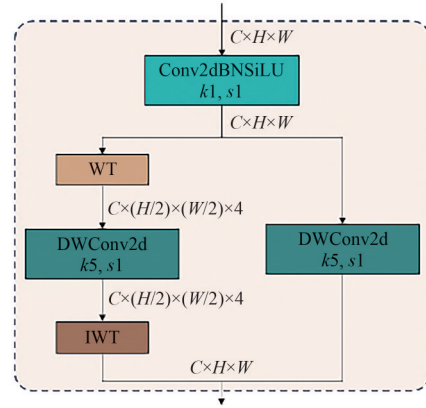


图5 小波瓶颈模块结构

Fig. 5 Wavelet bottleneck module structure

折叠状态影响外观。空间注意力机制可自动聚焦关键区域,对空间关联性强的人体姿态估计效果显著,能助力模型理解人体部位相对位置、捕捉关联;通道注意力机制通过自适应赋予不同通道权重增强关键目标特征提取,在复杂背景及不同折叠状态的纸张检测中,可有效突出重要特征区域。

对折或对折多次后的空白纸张属于特征信息少的小目标,为了提高复杂场景下模型对此类目标的关注度,在小目标检测层添加空间和通道协同注意力^[21] (Spatial and Channel Synergistic Attention, SCSA)。SCSA由两部分组成,即可共享的多语义空间注意力(SMSA)和渐进式通道自注意力(PCSA),协同增强模型对小目标的空间结构关系和特征依赖性的提取能力。

SCSA结构如图6所示,先对特征图的高、宽

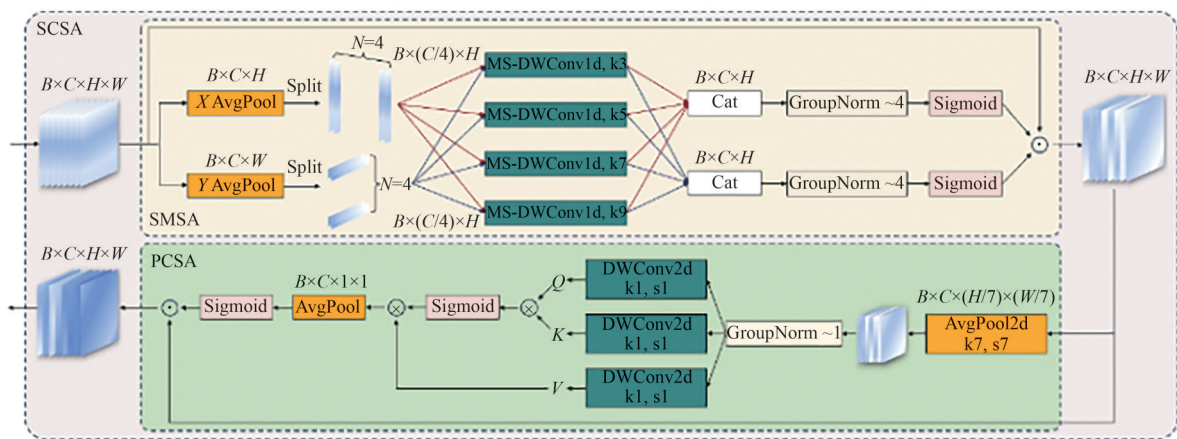


图6 SCSA模块结构

Fig. 6 SCSA module structure

方向分别平均池化,得到行、列特征序列,沿通道维度拆分为一组局部特征和3组全局特征,全局感知整体形状等高语义信息,局部捕捉细节纹理等低语义信息,再用4组深度共享1D卷积提取多尺度特征,最后拼接特征并经组归一化,通过门控激活函数生成空间注意力权重。SMSA借助多尺度、深度共享的1D卷积捕捉空间特征的多语义信息,有效整合全局上下文联系与多语义空间先验。

PCSA对SMSA输出的空间先验特征图下采样,经组归一化和3个深度可分离2D卷积得到 Q 、 K 、 V 矩阵,按式(11)计算权重,再通过全局平均池化和门控激活函数生成通道注意力权重。PCSA利用输入渐进式感知的自注意力机制计算通道相似性,减轻SMSA中不同感受野下子特征的空间语义差异。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (11)$$

SCSA有效利用多语义信息,且不会显著增加模型计算复杂度,通过多语义空间注意力引导通道特征学习,实现空间与通道特征的协同增强。

3 数据集

3.1 数据集采集

本文训练集为自建数据集,共1 620张图片。采集时长为一年,日期随机,时间为上午8时至下午20时。测试集由视频中截取三步动作过程中的图像各60张,共180张。

3.2 数据集标注

为提高标注效率,本文选用Mediapipe多媒体处理框架进行人体姿态估计,其内置多个预训练模型,广泛应用于手势识别、人脸检测、姿态估计等感知任务^[22]。本文采用Mediapipe内置的姿态估计、手部关键点检测及分割模型,对人体关键区域进行关键点检测与分割,记录每个关键区域的最小矩形框,自动生成标签文件。接着,通过LabelImg对自动标注结果进行校验,在此基础上手动进行纸张标注。最后,将标签转换为二级分类格式。在原始标签格式中,每一行表示一个标注框,格式如下:cls、 x 、 y 、 w 、 h 。其中:cls为整型数值,表示目标类别; x 、 y 、 w 和 h 为浮点型数值,分别表示归一化的标注框的中心点 x 坐标、中

心点 y 坐标、宽度和高度。目标类别表示类别的编号,由0开始编号。改进的二级分类标签格式如下:cls、subcls、 x 、 y 、 w 、 h 。其中:cls和subcls为整型数值,分别表示一级目标类别和二级目标类别;后面4位浮点型数值 x 、 y 、 w 、 h 同样表示归一化标注框信息。目标一级类别表示一级类别的编号,由0开始编号;目标二级类别表示二级类别的编号,由1开始编号,该位为0表示该标注框的一级类别下没有二级类别。本文的一级类别分为7类,其中“paper”类别下分3个二级类别,表示纸张的3种不同状态。

4 实验与分析

4.1 实验设置

4.1.1 实验环境

本文的训练实验平台采用Linux作为操作系统,具体配置如表1所示。

表 1 实验环境配置

Tab. 1 Experimental environment configuration

实验环境	配置
操作系统	Ubuntu 20.04
编程语言	Python 3.8.20
深度学习框架	Pytorch 1.12.0
CUDA	11.3
CPU	Intel® Core™ i9-10980XE
GPU	NVIDIA GeForce RTX 3090

4.1.2 参数设置

在本文任务中,多对目标为水平镜像对称关系,故不使用水平翻转。由于数据采集与测试场景的一致性,为使模型更好地学习空间关系,不使用马赛克数据增强。为确保实验的公平性并维持与原模型一致的优化动力学特性,本文遵循原框架的分类增益线性缩放原则,依据类别数量对损失权重进行自适应调整。具体而言,添加二级分类损失增益参数并设定为0.012 5,以保证每个语义类别的梯度贡献密度与基准模型(COCO, 80类)保持恒定。其余超参数及其调整策略遵循原模型,且所有实验均未使用预训练权重。

4.2 评价指标

在混淆矩阵中预测结果分为以下4种:TP(预测值为正样本,标签值也为正样本)、TN(预测值为

负样本,标签值也为负样本)、FP(预测值为正样本,标签值为负样本)和FN(预测值为负样本,标签值为正样本)。基于分类任务的预测结果,目标检测任务有一系列用于评价算法性能的指标^[23]:

交并比(IoU):检测框与标注框之间的重叠程度,用于衡量预测框的正确程度。

准确率(P):预测结果为TP的标注框数量在预测结果为阳性的标注框数量中的比例,计算公式如式(12)所示:

$$P = \frac{TP}{TP + FP} \quad (12)$$

召回率(R):预测结果为TP的标注框数量在标签值为阳性的标注框数量中的比例,计算公式如式(13)所示:

$$R = \frac{TP}{TP + FN} \quad (13)$$

平均准确率(AP):精确率-召回率(P-R)曲线下的面积。

平均准确率均值(mAP):所有类别的AP的平均值。YOLO模型中,mAP50表示IoU阈值为0.5时的mAP值,而mAP95表示在区间[0.5, 0.95]内,以0.05为步长,取10个IoU阈值,分别计算这10个IoU阈值时的mAP,再取平均值。

4.3 二级分类检测有效性实验

为验证二级分类检测的有效性,与原模型进行对比实验。由于二级分类检测中分类任务已经改变,为与原分类任务进行比较,对于父类别的评价指标将不参与全类别均值的计算,仅使用叶子节点类别(即没有子类别的类别)进行评价指标的计算。计算方法如式(14)所示:

$$M_{all} = \frac{1}{|S|} \sum_{c \in S} M_{cls}(c), \quad (14)$$

其中,S为叶子类别集合, $M_{cls}(c)$ 为c类别的评价指标, M_{all} 为全类别的评价指标。

实验结果如表2所示。二级分类检测模型的mAP95相较于原模型提高1.5%,其中对于二级类别中较难检测的“折叠纸张”和“折错纸张”类别的提升较为明显,分别从55.3%和37.3%提高至63.1%和42.6%。模型能够有效学习到细粒度的层级特征,证明了该损失函数设计的有效性。模型在更加关注二级类别的同时,对于多数其他一级类别的mAP95有所提高,使得二级分类检测模型的检测能力更强。

表2 单级分类与二级分类任务的mAP95对比

Tab. 2 Comparison of mAP95 between single level classification detection and two-level classification detection tasks

类别	单级分类 mAP/%	二级分类 mAP/%
左手	64.8	65.8
右手	67.6	63.1
左臂	49.6	52.1
右臂	55.7	54.5
左腿	55.8	56.6
右腿	52.9	54.2
完整纸张	67.2	67.5
折叠纸张	55.3	63.1
折错纸张	37.3	42.6
全部	56.2	57.7

在模型参数量方面,二级分类检测模型通过调整推理逻辑,仅改变不同检测层输出张量的大小,因此参数量仅轻微增加,约0.001M。由于推理阶段不涉及损失函数计算,且增加的参数量带来的计算负载极低,因此该改进在提升模型细粒度识别能力的同时,几乎不增加系统功耗。这一改进优化了模型推理层次,使检测层更聚焦于细致特征提取,提升了模型在复杂级别嵌套任务中的表现。

4.4 消融实验

为验证所提出改进模块的有效性,本文设计了11组消融实验,实验设计如表3所示,实验结果如表4所示。多级分类策略结果证实了该损失函数在细粒度分类上的有效性。因此,本节的消融实验均是在应用了该多级分类损失函数的基础上,进一步探讨网络结构改进模块的贡献。其中,实验1为基准模型;实验11为集成了ASC-Down、P34检测头、SCSA注意力及C3WT、C3Ghost模块的最终MLC-YOLO模型。实验2至实验10则是在基准模型之上,分别单独或组合添加上述模块的中间对照组,旨在逐一验证各改进模块对模型轻量化与检测精度的贡献。

使用FPN-sm结构和ASCDown能在保持模型性能的前提下有效降低参数量与计算量,参数量从1.778M下降至0.92M,GFLOPs从4.3G下降至3G。在小目标检测层引入SCSA,模型能更好地捕捉图像中的小纸张以及患者手部,提升小

表 3 消融实验设计
Tab. 3 Ablation experiment design

实验 编号	FPN-sm	ASCDwn	SCSA_s	C3WT	C3Ghost
1					
2	✓				
3		✓			
4			✓		
5				✓	
6					✓
7	✓	✓			
8	✓		✓		
9	✓	✓	✓		
10	✓	✓	✓	✓	
11	✓	✓	✓	✓	✓

目标的检测精度。此外,在模型颈部使用 C3WT 能增大模型感受野,增强模型的特征融合能力和上下文理解能力,在没有影响模型检测任务的同时,对细节或复杂场景时的目标定位更准确,不仅减少了计算开销,且 mAP95 提高了 1%。最后,将首个特征提取的瓶颈结构替换为更高效的 GhostConv,使模型更高效地捕捉关键特征,对遮挡、模糊目标的鲁棒性增强,召回率有 5.9% 的显著提高,进而提升模型的综合平均精度。与基准模型相比,MLC-YOLO 的 mAP50 和 mAP95 分别提高了 4.4% 和 4.1%,参数量减少了 51.52%,计算量减少了 39.53%。结果表明,本文的改进点在降低模型计算需求的基础上提高了检测精度,验证了方法的有效性。

表 4 消融实验结果
Tab. 4 Ablation experiment results

实验编号	Params/M	GFLOPs	<i>P</i> /%	<i>R</i> /%	mAP50/%	mAP95/%
1	1.778	4.3	95.9	79.9	90.5	57.7
2	1.322	3.9	90.2	82.0	90.0	58.4
3	1.237	3.2	89.9	83.3	91.4	58.8
4	1.778	4.3	92.5	82.3	91.1	59.1
5	1.593	3.8	92.5	86.2	92.5	58.3
6	1.775	4.2	92.5	84.2	91.3	57.4
7	0.920	3	92.2	84.6	91.8	59.8
8	1.323	3.9	90.8	84.9	91.5	59.3
9	0.921	3	93.3	86.4	93.1	60.1
10	0.864	2.7	93.7	86.3	93.1	61.1
11	0.862	2.6	90.9	92.3	94.9	61.8

注:粗体数字表示最优值

图 7 展示了 MLC-YOLO 和 YOLOv5n 的综合平均精度拟合曲线。可以看出,MLC-YOLO 在大约 60 个轮次(Epoch)后展现出更强的学习能力和优化潜力,最终的收敛精度较高,整体表现优于 YOLOv5n。

4.5 对比实验

鉴于目前针对 MMSE 中折纸动作自动评估的专用视觉检测模型尚属空白,且本研究的应用场景限定于低算力的嵌入式医疗设备,要求算法具备高实时性与轻量化特性,传统的基于 3D 卷积的重型动作识别网络并不适用。因此,为验证所提出模型的性能,将 MLC-YOLO 与主流通用目标检测模型在自建数据集上进行训练与测试,输入张量尺寸为 $480 \times 640 \times 3$,结果如表 5 所示。

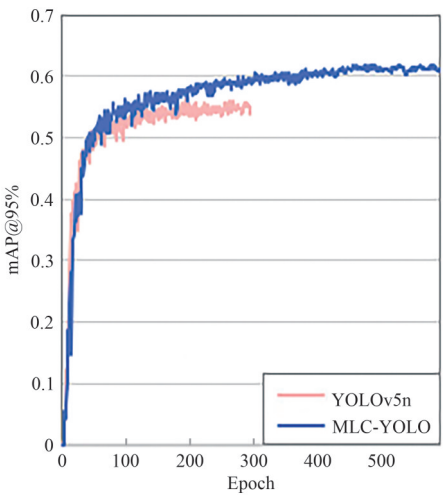


图 7 综合平均精度拟合曲线
Fig. 7 Mean average precision fitting curves

表 5 主流模型对比实验结果

Tab. 5 Comparison of experimental results of mainstream model

模型名称	Params/M	GFLOPs	P/%	R/%	mAP50/%	mAP95/%
YOLOv8n	1.851	4.2	90.4	84.5	90.3	54.2
YOLOv8s	7.347	16.5	88.9	86.0	90.9	57.8
YOLOv9t	0.975	3.5	93.1	84.9	90.0	53.0
YOLOv9s	3.863	13.5	94.7	87.0	91.9	56.9
YOLOv11n	2.178	4.8	90.7	88.7	92.0	57.3
YOLOv11s	8.646	18.6	91.2	87.0	91.9	59.0
MLC-YOLO	0.862	2.6	90.9	92.3	94.9	61.8

相较于其他主流网络,MLC-YOLO在多个关键指标上表现出色。首先,轻量化方面,参数量和计算量分别为 0.862M 和 2.6 GFLOPs,均低于其他对比模型。尽管 YOLOv9t 与 MLC-YOLO 的模型复杂度相近,但其综合平均精度表现最差。其次,MLC-YOLO 的 mAP50 和 mAP95 分别为 94.9% 和 61.8%,尤其在复杂背景、目标相互遮挡以及小目标检测场景下,表现出卓越的性能。与 MLC-YOLO 精度最接近的 YOLOv11s,其 mAP95 为 59%,位居次高,但其参数量和计算量分别为 MLC-YOLO 的 10.03 倍和 7.15 倍,导致计算资源消耗显著更大。此外,MLC-YOLO 在召

回率(92.3%)上表现突出,漏检率最低,适合医学筛查等场景。综上所述,MLC-YOLO 在 MMSE 中折纸动作关键目标的检测方面,具有极致轻量、高精度和低漏检等优点,在轻量化与效率方面显著优于其他模型,适合移动端部署。

为了更直观地展示不同模型的实际检测效果,本文选取小目标纸张、多目标遮挡和服饰与纸张颜色相近的 3 种检测场景进行比较,图 8 展现了本文模型与 YOLOv8s、YOLOv9s 和 YOLOv11s 在不同场景的检测效果,本文模型均表现出最高召回率,在多个目标的检测分数上有所提升,再次体现出 MLC-YOLO 的优越性能。

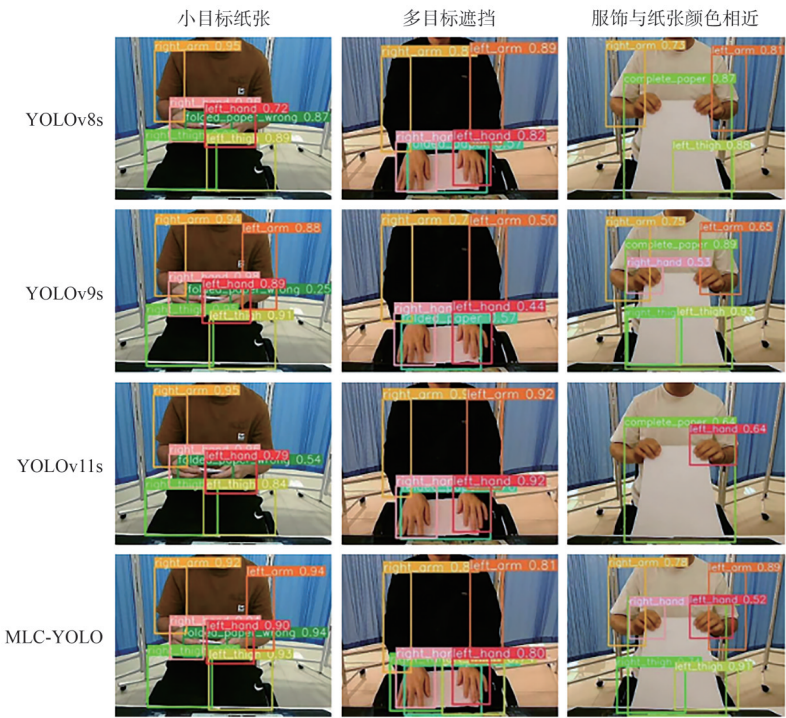


图 8 检测结果可视化
Fig. 8 Visualization of detection results

4.6 可视化结果与分析

本节对检测精度进行可视化分析,对比模型

改进前后的 P - R 曲线和热力图。 P - R 曲线展示了模型精确率与召回率间的关系,结果如图 9 所示。

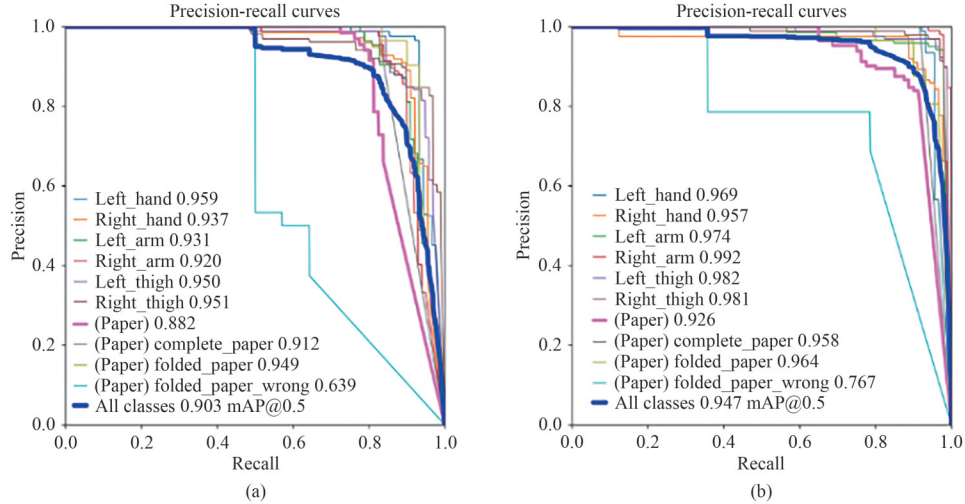


图 9 模型改进前(a)后(b)的 P - R 曲线

Fig. 9 P - R curves of the model before (a) and after (b) improvement

图 9(a)为原始二级分类模型的 P - R 曲线,可以看出“折错纸张”的精确率在召回率大于 0.5 时快速下降。图 9(b)为改进的 MLC-YOLO 模型的 P - R 曲线,改进后的模型对纹理特征较少的纸张,召回率和综合平均精度提升明显,尤其是对折多次后的小目标纸张, mAP50 从原来的 63.9% 提高至 76.7%,对于其他类别的目标精度均有小幅提高,曲线整体在更高的精确率和召回率区间延伸,展现出更优越的性能。

为进一步验证 MLC-YOLO 的性能,本文使用 XGrad-CAM 技术^[24],对比分析不同模型在小目标检测层的热力图。对比模型包括:改进前的 YOLOv5n 模型(图 10(a))、改进后的二级分类模型(图 10(b))、以及 MLC-YOLO 在添加 SCSA 注意力机制前后的热力图(图 10(c)和图 10(d))。XGrad-CAM 作为 Grad-CAM 的改进,通过计算

梯度信息生成图像的类别特征响应图,从而揭示模型在决策过程中对关键区域的关注度。图中暖色区域表示特征响应较强的区域,冷色区域表示特征响应较弱的区域。

可以看出,在背景较为复杂且存在其他人物干扰的情况下,原始模型对纸张的关注度明显不足,热力图中纸张特征响应微弱。改进后的二级分类模型则显著提升了对纸张的关注,暖色区域在纸张部分更为集中。MLC-YOLO 同时关注到了手部和纸张,但关注度相对较低。引入 SCSA 后,模型能在复杂环境下更精确地聚焦于手部和纸张边缘。该结果进一步验证了,通过将模型改进为多级分类目标检测任务、优化网络结构、改进降采样与特征提取模块,以及在小目标检测层加入注意力机制,可以有效增强模型在强背景干扰下的关键目标区域定位能力,减少误识别区域,从而提升模型性能。

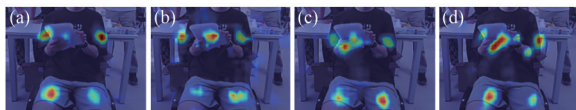


图 10 热力图对比实验。(a) 原模型;(b) 二级分类;(c) 无 SCSA;(d) MLC-YOLO 模型。

Fig. 10 Heatmap comparison experiment. (a) Original model; (b) Two-level classification; (c) Without SCSA; (d) MLC-YOLO.

5 结 论

针对 MMSE 中折纸测试任务自动评估的研究空白,以及嵌入式设备计算能力受限与传统目标检测模型性能不足的问题,本文设计了一种轻量化的多级分类目标检测模型——MLC-YOLO。该模型在 YOLOv5n 的基础上,通过一系列创新

性优化,使模型在极具轻量化的同时,显著提升了检测精度。具体而言,提出了多级分类目标检测,将同一目标的多种状态纳入考虑,使类别划分更符合实际逻辑。接着,对模型的网络架构、输出结构与后处理流程进行优化,有效提升了模型的计算效率和精度。结果表明,MLC-YOLO的mAP50和mAP95分别提升了4.4%和4.1%,同时,参数量和计算量分别减少了51.52%和39.53%,检测精度和计算效率均有较大提升。

MLC-YOLO为多级分类目标检测提供了新的技术方案,尤其在图像中各类别目标唯一的情况下,有效提高检测效率。未来的研究可以进一步探讨如何在更复杂的场景中,高效实现多级分

类下各类别最佳目标的检测,推动多级分类目标检测的广泛应用。此外,本文方法在MMSE中折纸动作关键目标检测方面具有重要意义,为人体关键部位与物体的同步检测提供了新思路,并有待引入时序特征分析,探索轻量化的时间序列模型,旨在将本文提取的细粒度空间特征与时序逻辑深度融合,构建时空联合分析框架,从而实现对连续动作的语义理解与全流程自动评分。本文弥补了MMSE中视频动作初步检测的研究空白,随着AI技术的不断发展,未来的研究有望在此基础上构建一套低成本、高隐私保护的认知障碍自动化筛查方案,推动智慧医疗技术在基层临床的广泛应用。

参 考 文 献:

- [1] 王廉源,杨毅,丛慧文,等. 不同MMSE评分下阿尔兹海默病发病风险因素的贝叶斯分位数回归联合模型分析[J]. 吉林大学学报(医学版),2023,49(2):395-401.
WANG L Y, YANG Y, CONG H W, *et al.* Bayesian quantile regression joint model analysis on risk factors of Alzheimer's disease in people with different MMSE scores [J]. *Journal of Jilin University (Medicine Edition)*, 2023, 49(2): 395-401. (in Chinese)
- [2] ANNADANAM A, KAHANA E, ANDREWS C, *et al.* Comparison of machine and human expert evaluation of capsulorrhexis creation performance through analysis of surgical video recordings [J]. *Clinical Ophthalmology*, 2024, 18: 943-950.
- [3] WANG Y T, SHEN X, HU S X, *et al.* Self-supervised transformers for unsupervised object discovery using normalized cut [C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022: 14543-14553.
- [4] MAO W A, GE Y T, SHEN C H, *et al.* TFPose: direct human pose estimation with transformers [J/OL]. *arXiv*, 2021: 2103.15320.
- [5] 吕情深. 基于目标检测的中国手语识别研究[D]. 乌鲁木齐:新疆大学,2020.
LYU Q S. Research on Chinese sign language recognition based on object detection [D]. Urumqi: Xinjiang University, 2020. (in Chinese)
- [6] 杨锋,丁之桐,邢蒙蒙,等. 深度学习的目标检测算法改进综述[J]. 计算机工程与应用,2023,59(11):1-15.
YANG F, DING Z T, XING M M, *et al.* Review of object detection algorithm improvement in deep learning [J]. *Computer Engineering and Applications*, 2023, 59(11): 1-15. (in Chinese)
- [7] 汤伟,王锦樞,张龙. 基于改进Faster R-CNN的纸病检测算法[J]. 包装工程,2023,44(21):260-266.
TANG W, WANG J Y, ZHANG L. Paper defect detection algorithm based on improved Faster R-CNN [J]. *Packaging Engineering*, 2023, 44(21): 260-266. (in Chinese)
- [8] JIA P, TIAN D P, WANG Y T, *et al.* Airborne optical imaging technology: a road map in CIOMP [J]. *Light: Science & Applications*, 2025, 14(1): 119.
- [9] 王琳毅,白静,李文静,等. YOLO系列目标检测算法研究进展[J]. 计算机工程与应用,2023,59(14):15-29.
WANG L Y, BAI J, LI W J, *et al.* Research progress of YOLO series target detection algorithms [J]. *Computer Engineering and Applications*, 2023, 59(14): 15-29. (in Chinese)
- [10] WONG A, FAMUORI M, SHAFIEE M J, *et al.* YOLO Nano: a highly compact you only look once convolutional neural network for object detection [C]//*Proceedings of 2019 Fifth Workshop on Energy Efficient Machine Learning*

- and Cognitive Computing-NeurIPS Edition (EMC2-NIPS). Vancouver: IEEE, 2019: 22-25.
- [11] KHOKHLOV I, DAVYDENKO E, OSOKIN I, *et al.* Tiny-YOLO object detection supplemented with geometrical data [C]//*Proceedings of 2020 IEEE 91st Vehicular Technology Conference*. Antwerp: IEEE, 2020: 1-5.
- [12] 陈仁祥, 邱天然, 杨黎霞, 等. 改进 YOLOv7 的服务机器人密集遮挡目标识别[J]. 光学精密工程, 2024, 32(10): 1595-1605.
- CHEN R X, QIU T R, YANG L X, *et al.* A method for dense occlusion target recognition of service robots based on improved YOLOv7 [J]. *Optics and Precision Engineering*, 2024, 32(10): 1595-1605. (in Chinese)
- [13] 林珊玲, 张雪, 陈燕, 等. 融合全局信息与双域注意力机制的光学遥感飞机目标检测[J]. 光学精密工程, 2024, 32(20): 3085-3098.
- LIN S L, ZHANG X, CHEN Y, *et al.* Integrating global information and dual-domain attention mechanism for optical remote sensing aircraft target detection [J]. *Optics and Precision Engineering*, 2024, 32(20): 3085-3098. (in Chinese)
- [14] 吕学涵, 李富, 祁铭瑞, 等. 基于改进 YOLOv11s 的无人机小目标检测算法[J]. 液晶与显示, 2025, 40(11): 1744-1756.
- LÜ X H, LI F, QI M R, *et al.* Target detection algorithm based on improved YOLOv11s UAV aerial image [J]. *Chinese Journal of Liquid Crystals and Displays*, 2025, 40(11): 1744-1756. (in Chinese)
- [15] 陈怡. 健康医疗数据共享与个人信息保护研究[J]. 情报杂志, 2023, 42(5): 192-199.
- CHEN Y. Research on healthcare data sharing and personal information protection [J]. *Journal of Intelligence*, 2023, 42(5): 192-199. (in Chinese)
- [16] WANG C Y, YE H I H, LIAO H Y M. YOLOv9: Learning what you want to learn using programmable gradient information [C]//*Proceedings of the 18th European Conference Computer Vision*. Milan: Springer, 2025: 1-21.
- [17] WANG A, CHEN H, LIU L H, *et al.* YOLOv10: Real-time end-to-end object detection [C]//*Proceedings of the 38th International Conference on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2024: 3429.
- [18] HAN K, WANG Y H, TIAN Q, *et al.* GhostNet: more features from cheap operations [C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020.
- [19] TULI S, DASGUPTA I, GRANT E, *et al.* Are convolutional neural networks or transformers more like human vision? [C]//*Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Online: Cognitive Science Society.org, 2021.
- [20] FINDER S E, AMOYAL R, TREISTER E, *et al.* Wavelet convolutions for large receptive fields [C]//*Proceedings of the 18th European Conference Computer Vision*. Milan: Springer, 2025: 363-380.
- [21] SI Y Z, XU H Y, ZHU X Z, *et al.* SCSA: exploring the synergistic effects between spatial and channel attention [J]. *Neurocomputing*, 2025, 634: 129866.
- [22] 倪广兴, 徐华, 王超. 融合改进 YOLOv5 及 Mediapipe 的手势识别研究[J]. 计算机工程与应用, 2024, 60(7): 108-118.
- NI G X, XU H, WANG C. Research on gesture recognition based on improved YOLOv5 and Mediapipe [J]. *Computer Engineering and Applications*, 2024, 60(7): 108-118. (in Chinese)
- [23] ZAIDI S S A, ANSARI M S, ASLAM A, *et al.* A survey of modern deep learning based object detection models [J]. *Digital Signal Processing*, 2022, 126: 103514.
- [24] FU R G, HU Q Y, DONG X H, *et al.* Axiom-based Grad-CAM: towards accurate visualization and explanation of CNNs [C]//*Proceedings of the 31st British Machine Vision Conference*. Online: BMVA Press, 2020.

作者简介:



陈宇聪, 男, 硕士研究生, 2023 年于绍兴文理学院获得学士学位, 主要从事机器视觉、目标检测等方面的研究。
E-mail: y2243290385@163.com



何宏, 女, 博士, 教授, 2008 年于华东理工大学获得博士学位, 主要从事医学人工智能、人机交互系统、医疗大数据分析方面的研究。E-mail: hehong@usst.edu.cn