DOI:10.19651/j.cnki.emt.2517971

# 联合多尺度注意力与混合池化的手腕创伤 X 光图像检测*

林淑娟[1] 钟铭恩[1] 谭佳威[2] 范 康[2] 林志强[1]
(1.厦门理工学院机械与汽车工程学院 厦门 361024;2.厦门大学航空航天学院 厦门 361005)

**摘 要**:针对 X 光图像中的骨折、软组织肿胀、骨病变等多类创伤的辅助检测问题,提出一种基于深度卷积神经网络的目标检测算法模型 WristXNet。首先设计了多尺度注意力特征聚合模块 C2f_MSAF 来增强模型对多尺度目标的特征理解能力;其次构建了混合池化空间金字塔模块 HPSP 来增强对不同类别目标关联特征的提取能力;随后引入动态上采样模块 DySample 来进一步增强对细粒度特征的捕捉能力;最后设计了具有解耦结构的轻量化检测头 LDDHead 来提升模型计算效率。在儿童手腕创伤 X 光图像公开数据集 GRAZPEDWRI-DX 上的实验结果表明,所提算法针对 X 光图像中的 7 类常见目标的平均检测精度 mAP 取得最高值 68.5%,相比现有最优算法提升了 1.6%,且模型大小仅为 3.3 M,处理效率达到每秒 156.9 张图像,体现了良好的综合性能。

**关键词**:医学图像处理;腕骨 X 光图像;多尺度注意力;混合池化;细粒度特征
**中图分类号**:TP391.4;TN919.8 **文献标识码**:A **国家标准学科分类代码**:520.20

# Combining multi-scale attention and hybrid pooling for wrist trauma X-ray image detection

Lin Shujuan[1] Zhong Ming'en[1] Tan Jiawei[2] Fan Kang[2] Lin Zhiqiang[1]
(1. School of Mechanical and Automotive Engineering, Xiamen University of Technology, Xiamen 361024, China;
2. School of Aerospace Engineering, Xiamen University, Xiamen 361005, China)

**Abstract**: To address the challenge of assisting in the detection of multiple types of traumas, including fractures, soft tissue swelling, and bone lesions in X-ray images, a target detection algorithm model based on deep convolutional neural networks WristXNet is proposed. Firstly, a multi-scale attention feature aggregation module C2f_MSAF was designed to enhance the model's ability to understand features of multi-scale targets. Secondly, a hybrid pooling spatial pyramid module HPSP was constructed to improve the extraction of correlated features among different target categories. Subsequently, a dynamic upsampling module DySample was introduced to further enhance the capture of fine-grained features. Finally, a lightweight detection head with a decoupled structure LDDHead was developed to improve computational efficiency. Experimental results on the publicly available pediatric wrist trauma X-ray dataset GRAZPEDWRI-DX, demonstrate that the proposed algorithm achieves the highest mean average precision (mAP) of 68.5% across seven common target categories in X-ray images, surpassing the current state-of-the-art algorithm by 1.6%. Additionally, the model size is only 3.3 M, and it achieves a processing efficiency of 156.9 images per second, demonstrating excellent overall performance.

**Keywords**: medical image processing; wrist bone X-ray images; multi-scale attention; hybrid pooling; fine-grained features

## 0 引 言

X 射线在骨科应用中较为广泛,常被用于诊断骨折、骨病变、旋前肌征、软组织肿胀、骨膜反应和异物等问题。在现有临床应用中,主要依靠专业人员对 X 光图像进行解读,容易因疲劳、经验不足以及技术水平有限等因素导致误诊和漏诊,以急诊科中的骨折诊断为例错误率高达 44%[1]。因此,借助计算机进行辅助诊断具有一定的现实

收稿日期:2025-01-20
*基金项目:福建省自然科学基金(2023J011439)项目资助

需求和良好的临床应用价值。

近年来,深度学习技术的快速发展为医学影像处理提供了新的技术路径,并诞生了诸多优秀的研究成果。吴慧东等[2]提出了一种结合卷积神经网络(convolutional neural network, CNN)和 Transformer 优势的混合架构,用于在结构性核磁共振成像上的 AD 病症诊断;Shen 等[3]通过 2.5 D 分类模型实现了 CT 影像中器官损伤的高效评估。针对病理区域尺寸不均、细节特征不明显及正负样本区分度低等问题,研究者们利用注意力机制聚焦关键特征,并结合多尺度信息来提升模型对细节的敏感性,在肺炎等疾病的影像识别与定位中取得了较高的准确性[4-6]。在骨骼疾病辅助诊断方面,Dibo 等[7]结合 CNN 中的 YOLOv7 与 Swin Transformer 提出了 Deeploc 混合模型,提升了骨病理定位与分类的精度;黄泽青等[8]提出的基于 ResNet-152 的迁移学习模型在髋关节疾病 X 光图像诊断中取得了良好效果;Jabbar 等[9]设计的 RN-21CNN 在手腕裂缝分类上实现了 97%的准确度,优于 Inception V3、Vgg16、ResNet-50 和 Vgg19 等常见的迁移学习模型;Schilcher 等[10]通过结合多模态数据与深度学习技术,提高了非典型股骨骨折的检测准确性;Ju 等[11]采用 YOLOv8 算法开发了辅助诊断应用;Chien 等[12]通过注意力机制对 YOLOv8 进行了改进提出了 YOLOv8-AM 模型,显著提升了骨创伤的检测精度;Ahmed 等[13]基于 YOLOv10 算法的双标签分配系统,通过优化模型复杂度和扩展架构,展现了更高的检测效率和精准度。然而,由于 X 光图像所包含的信息量相对有限,且部分图像因患者已使用石膏进行受伤部位包裹固定而存在干扰,使得现有方法在 X 光图像特征提取的有效性上受到限制。如何更有效的理解、提取和融合 X 光图像中目标的细粒度特征是提升算法平均检测精度的一种可探索路径。

为此,本文以手腕创伤后拍摄的 X 光图像为研究对象,并围绕多尺度特征提取、目标特征融合及细粒度信息捕捉 3 个方面进行改进,提出一种针对其内部包含的 7 类典型目标(5 类典型创伤+异物+文本)的自动检测算法 WristXNet。

其中,在多尺度特征提取方面,设计了多尺度注意力特征聚合网络模块 C2f_MSAF,使得模型能够更好的捕捉不同尺度特征之间的依赖关系,增强模型对多尺度目标的特征理解和提取能力,提升鲁棒性;在目标特征融合方面,设计了混合池化空间金字塔网络模块(hybrid pooling spatial pyramid, HPSP),用来增强模型对不同类别目标关联特征的感知和融合能力;在细粒度信息捕捉方面,引入动态采样模块 (dynamic sampling module, DySample),通过精细调整采样位置提升对细粒度特征的捕捉和保存能力。此外,构建了轻量解耦检测头(lightweight decoupled detection head, LDDHead)来提高模型的计算效率。最后在儿童手腕 X 光图像公开数据集 GRAZPEDWRI-DX 上实验验证了所提算法的性能。

## 1 算法设计

### 1.1 整体结构设计

WristXNet 的整体架构如图 1 所示,由提取基础特征的 Backbone、融合不同尺度特征的 Neck 以及用于检测任务推理的 Head 3 部分组成。在 Backbone 部分引入了 C2f_MSAF 模块来增强模型对多尺度特征的理解和提取能力。针对 X 光图像中特征不明显和信息不丰富的问题,设计了混合池化空间金字塔模块 HPSP,旨在提升模型对不同类别目标关联特征的感知能力。为进一步优化特征提取和模型性能,在 Neck 部分引入动态上采样模块 DySample,通过精细调整采样位置和特征形态,增强模型对细粒度特征的捕捉。最后搭建了 LDDHead 检测头来提高模型的计算效率。在网络中,用符号©表示对各层级图像特征的拼接融合操作、基础卷积 CBS 以及特征聚合模块 C2f 的设计灵感均来自于文献[14],在此不再赘述。

### 1.2 多尺度注意力特征聚合模块

多尺度注意力特征聚合模块 C2f_MSAF 的设计思路来源于 YOLOv8 中的 C2f 模块,具体结构如图 2 所示。传统 C2f 模块首先通过基础卷积对输入特征图进行特征转换,并将其拆分为多个分支,然后利用 Bottleneck 层进一步提取特征信息,最后再将处理后的特征与原始输入特征进行拼接和压缩,从而实现跨阶段的信息融合。而在 C2f_MSAF 模块中,将 Bottleneck 层替换为多尺度注意力融合层 Bottleneck_MSAF,这一替换将基础卷积 CBS 替换为深度卷积(depthwise convolution, DWConv),DWConv 通过分解卷积操作来减少计算复杂度和参数量,引入多尺度特征聚合(multi-scale attention fusion, MSAF)模块来加强模型对细节信息和多尺度上下文的捕获能力。

具体而言,MSAF 通过空间通道注意力机制(spatial-channel attention, SCA)对输入特征进行融合,使得模型能够自适应地强调重要特征,抑制背景干扰特征,从而提高模型的表达能力和检测精度。与此同时,进一步应用多尺度特征提取模块(multi-scale feature extraction, MFE)来从不同尺度提取图像特征。通过融合 MFE 和 SCA,实现对输入特征的多尺度感知和重要性加权。

$$F_{Out} = F_{SCA} + F_{MFE} \tag{1}$$

SCA 的具体实现步骤如下:

1)采用自适应平均池化(AdaptiveAvgPool)操作对输入特征 $F \in R^{B\times C\times H\times W}$ 进行处理,实现高度和宽度的分离,分别记为 $x \in R^{B\times C\times H\times 1}$ 和 $y \in R^{B\times C\times 1\times W}$。

$$x = AdaptiveAvgPool(F) \tag{2}$$

$$y = AdaptiveAvgPool(F) \tag{3}$$

2)分别对分离后的特征 $x$ 和 $y$ 进行特征投影,并扩展到原始特征图的大小 $R^{B\times C\times H\times W}$,生成用于计算注意力的 $\boldsymbol{q}$ 和 $\boldsymbol{k}$,$\boldsymbol{v}$ 则是对原始输入特征图 $F$ 进行特征投影得到

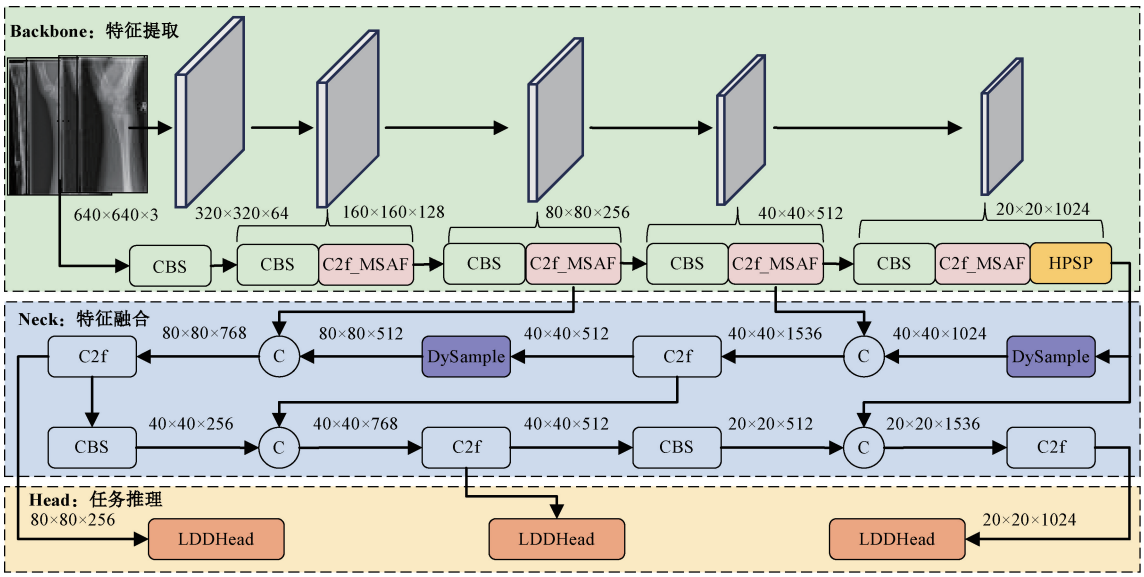


图 1　WristXNet 整体结构

Fig. 1　The overall structure of the WristXNet

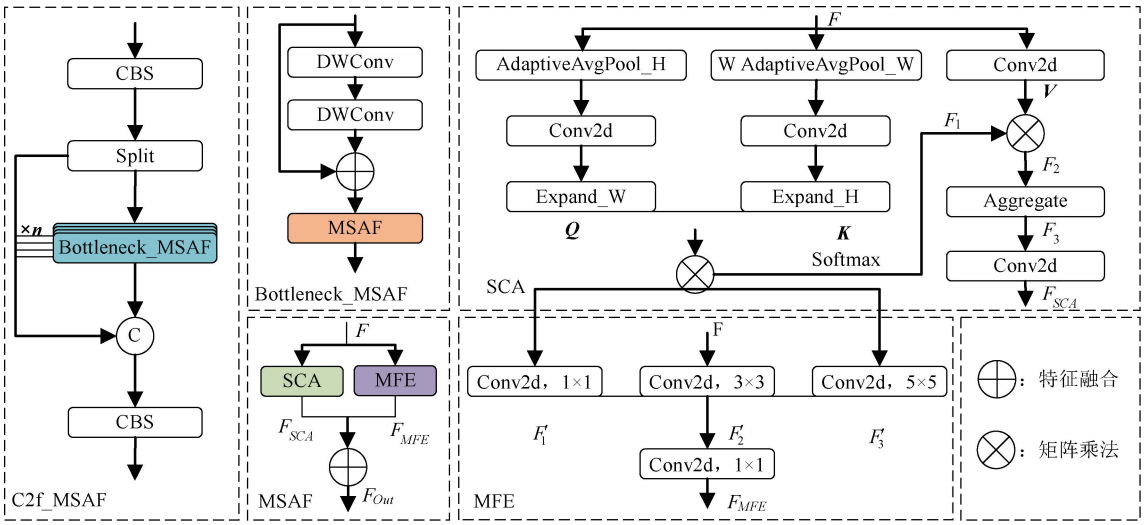


图 2　C2f_MSAF 结构

Fig. 2　The structure of the C2f_MSAF

的，根据式(4)～(8)进行空间注意力运算。

$$\boldsymbol{Q} = \text{expand}(Conv2d_{1\times1}(x)) \tag{4}$$

$$\boldsymbol{K} = \text{expand}(Conv2d_{1\times1}(y)) \tag{5}$$

$$F_1 = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{l}) \tag{6}$$

$$\boldsymbol{V} = Conv2d_{1\times1}(F) \tag{7}$$

$$F_2 = F_1\boldsymbol{V} \tag{8}$$

第 3 步，对特征进行 Aggregate 操作。首先对特征图 $F_2$ 沿空间维度计算全局均值，并通过 Softmax 函数归一化得到通道注意力权重。随后，将该权重与 $F_2$ 进行通道加权，以强化关键特征响应。最后通过 1×1 卷积实现跨通道信息融合与线性变换，输出增强后的特征图 $F_{SCA}$。

$$F_3 = Aggregate(F_2) \tag{9}$$

$$F_{SCA} = Conv2d_{1\times1}(F_3) \tag{10}$$

MFE 的具体实现步骤如下：使用 1×1、3×3 和 5×5 的卷积核，从输入特征 F 中提取多尺度信息，然后将这些特征沿通道维度拼接，并通过一个 1×1 卷积层进行特征融合。

$$F'_n = \text{Conv2d}_{n\times n}(F), (n = 1,3,5) \tag{11}$$

$$F_{MFE} = Conv2d_{1\times1}(F'_1 + F'_3 + F'_5) \tag{12}$$

C2f_MSAF 模块融合了多尺度特征和注意力机制，有利于增强模型对细节和上下文的捕获能力，减缓梯度消失问题，从而保持梯度信息传递，这些对于模型更好地适应儿童手腕创伤检测的复杂性和多样性是有益的。

### 1.3　混合池化空间金字塔

在儿童手腕创伤 X 光图像分析中，传统的特征提取方

法,如仅依赖最大池化的 SPPF(spatial pyramid pooling-fast)[15] 和 SimSPPF(simplified spatial pyramid pooling-fast)[16],可能无法充分捕捉图像的全局和局部特征。尽管最大池化能够突出显著特征,但它忽略了图像中的其他重要信息。为此,本文提出了一种新的特征提取模块 HPSP 来尝试解决该问题,具体结构如图 3 所示。HPSP 模块借鉴了 YOLOv9 中的 SPPELAN[17](spatial pyramid pooling enhanced with elan)设计思路,引入了平均池化分支,以更有效的捕捉全局背景和上下文信息,从而实现对图像特征更为全面和深入的提取。

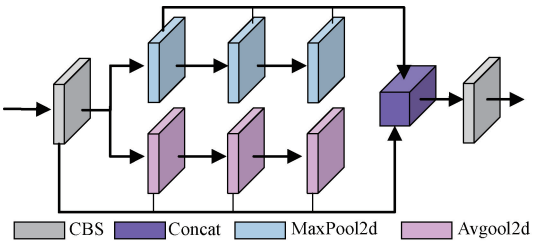


图 3　混合池化空间金字塔结构

Fig. 3　Hybrid pooling spatial pyramid structure

HPSP 的具体实现步骤如下:首先,输入特征图通过一个 1×1 的卷积层处理,将特征图的通道数减半。随后,在两个分支上分别使用最大池化和平均池化操作,双分支策略在保留全面信息的同时能够突出关键细节。并将这些不同尺度下的特征图进行拼接,有效融合全局和局部信息。最后,通过一个 1×1 的卷积层对拼接后的特征图进行处理,以整合不同分支的信息并调整输出通道数。

HPSP 通过结合最大池化和平均池化操作,促进了多尺度特征的有效融合,可为模型提供更全面且均衡的信息。

### 1.4　动态上采样器

上采样技术是目标检测中的关键环节,通过提升特征图的空间分辨率,确保检测细节的丰富性和准确性。然而,传统的上采样方法,如最近邻插值和双线性插值,往往难以满足高分辨率特征图的需求,可能导致细粒度特征的丢失。本文引入动态上采样器 DySample[18] 来尝试解决该问题,其具体结构如图 4 所示。

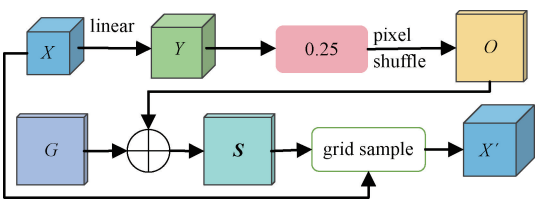


图 4　DySample 模块结构

Fig. 4　Structure of DySample module

与传统的基于卷积核的上采样方法不同,DySample 根据输入特征自适应优化采样位置,相较于固定的采样网格,更具灵活性,特别是在处理复杂或高维特征时,能够更好地捕捉图像中的细粒度特征。

DySample 的具体实现步骤:给定大小为 $c\times h\times w$ 的特征图 $X$ 和上采样比例因子 $s$。首先通过线性层将 $X$ 转换为大小为 $2s^2\times h\times w$ 的偏移量 $Y$,为了避免放大过程中出现重叠和性能下降,对 $Y$ 进行 0.25 倍的缩放并通过像素洗牌技术将其重塑为 $O$。随后,创建与 $O$ 相同尺寸的原始采样网格 $G$,将偏移量 $O$ 与原始采样网格 $G$ 相加得到采样点集 $S$。最后使用 PyTorch 内置的网格采样函数,根据采样点集 $S$ 对连续的特征图进行采样,生成大小为 $c\times sh\times sw$ 的上采样特征图 $X'$。

$$S=G+O \tag{13}$$

$$X'=\text{gridsample}(X,S) \tag{14}$$

DySample 通过动态调整采样点位置,能够更好地保留和细化图像中的信息与结构,从而有利于提升模型性能。

### 1.5　轻量解耦检测头

轻量解耦检测头 LDDHead,在保留传统检测头高效特征处理流程的基础上,引入 LightConv[14] 模块,LightConv 模块以其独特的结构,优化了特征处理过程,构建出更为丰富和精确的特征表示,为后续边界框的预测和类别分类提供了坚实的基础。其具体结构如图 5 所示。

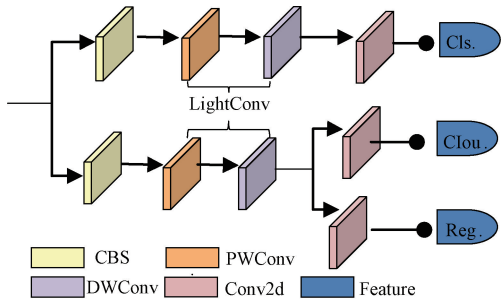


图 5　轻量解耦检测头结构

Fig. 5　Lightweight decoupled detection head structure

LDDHead 检测头由分类预测分支与边界框预测两分支组成,分别用于输出目标类别概率和负责预测边界框的偏移量,以确保分类精度和定位准确性。两分支均通过 3×3 基础卷积层对输入特征进行细化,从而增强对目标边缘和形状的感知能力。LightConv 检测头模块结合 1×1 逐点卷积(pointwise convolution, PWConv)与 3×3 深度卷积,在降低计算开销的同时保留特征细节信息,使模型能够更加精确地捕捉骨折等医学目标的细微特征。该设计在保持较浅层结构的情况下,增强了检测头对目标特征的敏感度和表达力。最终,在后处理阶段融合两分支的预测结果,形成完整的检测输出,包括目标的边界框位置、大小和类别概率。

### 1.6　损失函数

WristXNet 采用二元交叉熵损失 $L_{cls}$[15] 作为网络训练时的目标分类损失,计算原理为:

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \lg(p_i) + (1-y_i)\lg(1-p_i)] \tag{15}$$

式中：$N$ 表示样本数量，$y_i$ 表示样本的真实类别标签(0 或 1)，$p_i$ 表示网络预测的类别概率。

在对感兴趣目标进行边界框回归定位时，采用联合交并比损失 $L_{ciou}$[19]衡量预测边界框和真实边界框之间的匹配程度，使用区域分布自适应损失 $L_{Focal}$[20]来预测置信度的不确定性。

$$IOU = \frac{|A \cap B|}{|A \cup B|} \tag{16}$$

$$v = \frac{4}{\pi^2}(\arctan(\frac{w_{gt}}{h_{gt}}) - \arctan(\frac{w}{h}))^2 \tag{17}$$

$$L_{ciou} = 1 - IoU + \frac{\rho^2(A,B)}{c^2} + \alpha \times v \tag{18}$$

$$\alpha = \frac{v}{(1-IoU)+v} \tag{19}$$

$$L_{Focal} = -\frac{1}{N}\sum_{c=0}^{c=1}\sum_{n=1}^{N} g_n(c)(1-p_n(c))^2 \lg(p_n(c)) \tag{20}$$

其中，$IOU$ 表示预测边界框 $A$ 和真实边界框 $B$ 的交并比，$\alpha$ 是一个平衡参数，$v$ 是关于 $A$ 和 $B$ 的宽高比和长宽比的函数，$\rho$ 表示 $A$ 和 $B$ 的中心点之间的欧式距离，$c$ 表示包围 $A$ 和 $B$ 的最小外接框的对角线长度，$N$ 表示样本数量，$C$ 表示类别数量，$g_n(c)$是样本 $n$ 是否属于类别 $c$ 的指示函数，$p_n(c)$是模型对样本 $n$ 属于类别 $c$ 的概率。

将训练总损失定义为分类损失、边界框回归损失和焦点损失的加权和：

$$L_{total} = \alpha \times L_{cls} + \beta \times L_{ciou} + \gamma \times L_{Focal} \tag{21}$$

其中，$\alpha$、$\beta$ 和 $\gamma$ 表示平衡三者对网络影响的权重参数。为了算法性能对比的公平性，在实验中采用与 YOLOv8[14]相同的权重系数，即分别取 0.5、7.5 和 1.5。

## 2　实验与分析

### 2.1　数据集

实验采用儿童手腕创伤 X 光图像的公共数据集 GRAZPEDWRI-DX[21]，该数据集收集来自 6 091 名患者的带注释的儿科手腕创伤 X 光图像，包含 74 459 个图像标签，共计 67 771 个标记对象，涵盖 9 个类别。由于数据集中未提供明确的划分准则，本文依据患者 ID 将数据集按 7∶2∶1的比例划分为训练集、验证集和测试集，以避免同一患者的样本出现在多个子集中。为增强模型的鲁棒性，使用 OpenCV 中的 addWeighted 函数调整 X 光图像的对比度和亮度，实现训练集扩充。鉴于 GRAZPEDWRI-DX 数据集中存在 9 个标注类别的不均衡，本文借鉴 Dibo 等人在 Deeploc 研究中的方法[7]，将异物和金属合并为异物一类，骨骼异常和骨骼病变合并为骨病变一类，不同于 Deeploc 研究，本文保留骨膜反应和骨折类别的同时还保留了其他所有标注类型，以确保数据集的丰富性和多样性。具体类别为：骨病变(bone lesion，BL)、旋前肌征(pronator sign，PS)、软组织肿胀(soft tissue swelling，ST)、骨折(fracture，F)、异物(foreign body，FB)、骨膜反应(periosteal reaction，PR)和文本(text，Tx)。

### 2.2　实验条件与参数设置

实验主机操作系统为 64 位 Windows10，硬件采用 Intel Core i5-12400F CPU 和 NVIDIA GeForce RTX 3060 显卡。算法开发环境采用 Python3.9.18 和 PyTorch 深度学习框架。模型训练选用 SGD 优化器，初始学习率设置为 0.01，权重衰减率设置为 $5\times10^{-4}$，学习率衰减策略选择线性衰减。各对比模型都训练 300 轮次，批大小为 16。数据预处理时，将图像大小统一调整为 640×640，并通过随机缩放、旋转、平移、透视变换等操作来增加几何畸变，以及调整图片的亮度、对比度、饱和度等进行随机光照增强。

### 2.3　评价指标

选择平均检测精度(mean average precision，mAP)作为目标检测性能的评价指标：

$$P = \frac{TP}{TP+FP} \tag{22}$$

$$R = \frac{TP}{TP+FN} \tag{23}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{24}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{25}$$

$$F_1 = \frac{2\times P\times R}{P+R} \tag{26}$$

其中，$TP$ 指被正确预测为正样本的数量，$FP$ 指被错误预测为正样本的数量，$FN$ 指被错误预测为负样本的数量；$P$ 为准确率，$R$ 为召回率；$N$ 代表检测类别的数量，$AP_i$ 表示第 $i$ 个类别的平均精度；选取模型参数量 $N_p$ 和每秒浮点运算数 GFLOPs 来衡量模型的内存占用程度和计算复杂度，采用每秒传输帧数 FPS 衡量模型的推理速度，用 $F_1$ 评估模型性能。

### 2.4　消融实验

为探明所提出的多尺度注意力特征聚合模块 C2f_MSAF、混合池化空间金字塔 HPSP、动态上采样器 DySample 以及轻量解耦检测头 LDDHead 模块的有效性，在 GRAZPEDWRI-DX 数据集上进行了消融实验来分析其对模型性能的提升情况，具体实验结果如表 1 所示。

可以看出 C2f_MSAF 模块、HPSP 模块、DySample 模块以及 LDDHead 模块的加入对于基础网络模型的提升较为明显：(1)在 Backbone 使用 C2f_MSAF 模块，mAP 提高了 0.3%，GFLOPs 增加了 22.2%，参数量增加了 20.0%。这意味着 C2f_MSAF 模块能更好的地捕捉到骨折图像中的细粒度特征，尤其是在复杂背景下，增强了网络对关键区域的敏感性和对多尺度信息的融合能力。尽管计算

**表 1　消融实验结果**

**Table 1　Results of ablation experiment**

| 算法改进 | C2f_MSAF | HPSP | DySample | LDDHead | mAP/% | GFLOPs | $N_p$/M | FPS | $F_1$/% |
|---|---|---|---|---|---|---|---|---|---|
| 基础网络 | × | × | × | × | 65.3 | 8.1 | 3.0 | 303.9 | 64.8 |
| 改进 1 | √ | × | × | × | 65.6 | 9.9 | 3.6 | 181.3 | 64.8 |
| 改进 2 | × | √ | × | × | 66.4 | 8.0 | 2.8 | 308.0 | 64.5 |
| 改进 3 | × | × | √ | × | 65.8 | 8.1 | 3.0 | 250.8 | 64.9 |
| 改进 4 | × | × | × | √ | 65.7 | 7.0 | 2.8 | 319.0 | 62.8 |
| 本文方法 | √ | √ | √ | √ | 68.5 | 8.7 | 3.3 | 156.9 | 67.0 |

注:“×”表示未添加优化模块,“√”代表加入相应的优化模块

复杂度和模型参数有所增加,但这一增加换取了模型更好的检测精度,使得模型在实际应用中更加有效和鲁棒。(2)单独使用 HPSP 模块后,平均检测精度 mAP 提高 1.1%,参数量和每秒浮点运算数均有所下降,这表明 HPSP 模块通过提升特征融合效率,优化了计算资源的使用,使得模型在提升检测精度的同时,减少了计算开销。(3)单独使用 DySample 模块后,mAP 提高了 0.5%,$N_p$ 和 GFLOPs 无明显变化。这表明 DySample 在不增加额外的计算和存储开销,通过精细调整采样位置和特征重塑,有效提升了模型对细粒度特征的捕捉和保存能力。(4)单独使用 LDDHead 模块后,mAP 提高了 0.4%,参数量减少了 6.7%,每秒浮点运算数降低了 13.6%,模型推理速度提升至 319.0 fps。这表明 LDDHead 能够减少冗余计算和模型参数,推理更高效。(5)同时使用这 4 个改进策略后,模型的平均检测精度达到了 68.5%,且计算复杂度 GFLOPs 仅增加 7.4%,参数量增加 10.0%,模型大小仅为 3.3 M,处理效率达到了每秒 156.9 张图像,$F_1$ 得分也有较大提升,达到了 67.0%。尽管计算复杂度和参数量有所增加,但检测精度得到了有效提升,证明了这些改进能够使模型在精度与效率间达到良好的平衡。

如图 6 所示,对比了不同模型的检测效果。展示了原始标签、消融实验中基线模型与最终模型 WristXNet 在 GRAZPEDWRI-DX 数据集上的推理结果。从图 6 中可以看出:对于特征明显且易于检测的损伤类型,基线模型和最终模型都能准确识别。然而,在处理包含骨异常、软组织肿胀及多重损伤类型的复杂 X 光图像时,基线模型表现出更多的漏检和误检现象。此外,由于石膏引起的 X 光散射增加了图像分析的复杂性,使得检测任务更加困难。而 WristXNet 模型在这种复杂条件下依然能够保持稳定的检测表现,显著优于基线模型。尽管 WristXNet 模型在某些情况下未能检测到所有目标,但相较于基线模型,其在相同条件下的整体检测效果更佳,图 6 的检测效果对比清晰展示了这一优势。

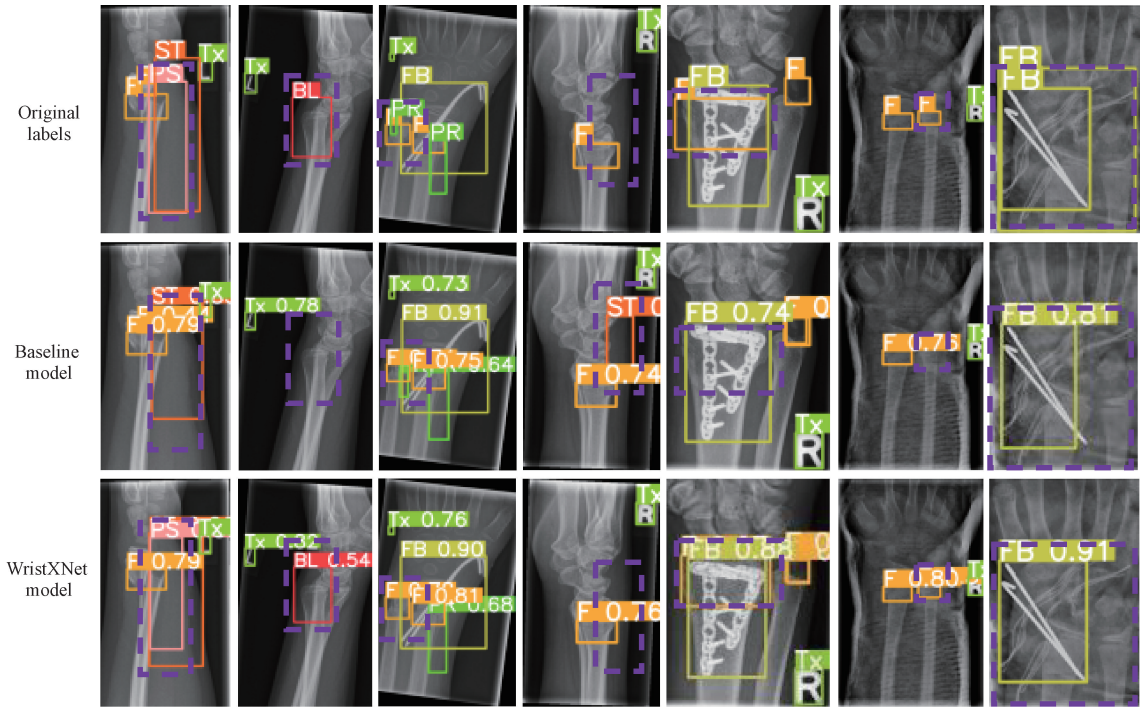


图 6　检测效果可视化对比

Fig. 6　Visual comparison of detection results

### 2.5　对比实验

1)C2f_MSAF 对比实验

表 2 展示了 C2f_MSAF 模块位置对于网络性能的影响。结果表明,仅在 Backbone 使用时,虽然 GFLOPs 和参数量有所增加,但模型的 mAP 达到了 65.6%,这一精度的提升证明了 C2f_MSAF 模块应用于 Backbone 中的有效性。相比之下,仅在 Neck 引入 C2f_MSAF 模块时虽然 GFLOPs 较低,但整体检测精度不如仅用在 Backbone 中的策略。如果在 Backbone 和 Neck 全局应用 C2f_MSAF 模块,GFLOPs 和参数量分别增加到 11.1 G 和 4.2 M,计算成本显著提高且检测精度更低。这表明全局应用 C2f_MSAF 模块并未带来预期的性能提升,反而可能引发过拟合或导致特征信息的过度提取。因此,基于计算效率和避免过拟合的考虑,本文仅在 Backbone 引入 C2f_MSAF 模块。

**表 2　C2f_MSAF 位置对网络性能的影响**

**Table 2　Effect of C2f_MSAF location on network performance**

| 位置 | mAP/% | GFLOPs | $N_p$/M | FPS | $F_1$/% |
|---|---|---|---|---|---|
| Backbone | 65.6 | 9.9 | 3.6 | 181.3 | 64.8 |
| Neck | 65.1 | 9.3 | 3.6 | 207.9 | 62.2 |
| Backbone+Neck | 64.3 | 11.1 | 4.2 | 142.1 | 62.4 |

2)与不同检测算法的性能对比

为进一步评估本文所提算法的性能,将其与现有 CNN 架构(其中 YOLOv8n[11]、YOLOv8-AM[12]、YOLOv8_GC[22]和 YOLOv10n[23]是与手腕创伤相关的研究算法)和 Transformer 架构的主流算法进行对比实验,实验结果如表 3 所示。

**表 3　不同算法模型的性能对比结果**

**Table 3　Comparison of the performance of different algorithms**

| 方法 | AP/% | | | | | | | mAP/% | GFLOPs | $N_p$/M | FPS | $F_1$/% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BL | PS | ST | F | FB | PR | Tx | | | | | |
| YOLOv5n | 5.2 | 65.6 | 33.5 | 94.1 | **94.7** | 64.1 | 99.1 | 65.2 | 7.1 | **2.5** | **305.3** | 61.3 |
| YOLOv8n[11] | 8.1 | 68.1 | 33.2 | 94.4 | **94.7** | 66.6 | 99.1 | 65.3 | 8.1 | 3.0 | 303.9 | 64.8 |
| YOLOv8_SA[12] | 8.1 | 69.3 | 35.8 | 93.0 | 94.5 | 66.2 | 99.0 | 66.6 | 8.1 | 3.0 | 270.7 | 63.4 |
| YOLOv8_ECA[12] | 9.2 | 69.1 | 35.1 | 94.5 | 93.8 | 66.3 | 99.0 | 66.7 | 8.1 | 3.0 | 299.2 | 63.6 |
| YOLOv8_GAM[12] | 5.3 | 68.6 | 35.6 | **94.7** | 94.4 | 66.6 | 99.1 | 66.3 | 9.5 | 3.7 | 168.3 | 62.5 |
| YOLOv8_ResCBAM[12] | 9.4 | 66.3 | 31.6 | 94.5 | 94.3 | 65.5 | **99.2** | 65.8 | 10.4 | 4.2 | 231.8 | 63.1 |
| YOLOv8_GC[22] | 12.2 | 68.2 | 33.8 | 94.6 | 93.5 | **66.7** | 99.0 | 66.9 | 8.1 | 3.0 | 238.2 | 64.7 |
| YOLOv10n[23] | 6.8 | 61.7 | 33.0 | **94.7** | 92.0 | 63.0 | 98.8 | 64.3 | 8.2 | 2.7 | 215.9 | 62.9 |
| YOLOv11n[24] | 1.5 | 65.8 | 33.3 | **94.7** | **94.7** | 66.5 | 99.0 | 65.1 | **6.3** | 2.6 | 258.3 | 62.3 |
| DETR[25] | 9.5 | 27.4 | 7.5 | 82.9 | 85.6 | 37.3 | 90.1 | 48.6 | 20.4 | 41.5 | 47.8 | 57.5 |
| RT-DETR[26] | 4.9 | 65.1 | 19.1 | 90.7 | 91.0 | 60.4 | 97.4 | 61.2 | 14.7 | 8.1 | 77.6 | 64.2 |
| DAB-DETR[27] | 0.6 | 25.1 | 11.5 | 86.3 | 93.3 | 39.0 | 83.5 | 47.0 | 24.3 | 43.7 | 37.7 | 54.6 |
| DINO[28] | 8.1 | 51.2 | 20.5 | 91.4 | 89.2 | 64.6 | 97.8 | 60.4 | 67.7 | 47.7 | 18.4 | 63.4 |
| WristXNet(Ours) | **12.9** | **74.0** | **37.5** | 94.4 | **94.7** | **66.7** | 99.1 | **68.5** | 8.7 | 3.3 | 156.9 | **67.0** |

可以看出:(1)本文算法在骨病变、旋前肌征和软组织肿胀分类上的表现最佳,分别比已有最佳指标提升 0.7%、4.7%和 1.7%,达到 12.9%、74.0%和 37.5%,异物和骨膜反应均与最佳指标持平,而对于骨折和文本分类,分别略低于最佳指标 0.3%和 0.1%,最终整体 mAP 达到了 68.5%,超越了现有主流算法。(2)本文算法的参数量仅为 3.3 M,与现有 CNN 架构的轻量算法处于同一数量级,但显著小于以 Transformer 为架构的各类算法,推理速度可以达到 156.9 fps,体现了良好的检测效率。(3)本文算法的 F1 得分为 67.0%,优于其他算法模型,说明本文模型在处理不平衡数据时具有更好的鲁棒性。这些结果表明,所设计的 WristXNet 算法模型在综合性能上优于现有的主流目标检测网络。

为进一步展示 WristXNet 在特征关注度方面的能力,揭示其特征提取的性能,利用 Grad-CAM 热力图进行了可视化分析,并将 mAP 取得次优的 YOLOv8_GC 作为对比参照(以下简称为对比算法),结果如图 7 所示。图 7 中暖色区域代表模型对特征的关注强度,亮度越高表示算法对该区域的关注越集中,清晰地展示了 WristXNet 与对比算法在关注区域上的差异。可以观察到:(1)在第 1、2 和 3 列中,对比算法出现了明显的误检现象。(2)在第 4 和 5 列中,对比算法未能有效的关注漏检区域,而 WristXNet 则能够精准地聚焦于这些区域。(3)在第 6 和 7 列中,尽管两种算法均能正确检测损伤且没有误检或漏检,但对比算法的特征关注存在较大偏差,而 WristXNet 则在这些区域表现出更一致和集中的特征关注。(4)在

第 8 列中,两种算法均存在漏检现象,其中对比算法存在两处漏检,而本文算法仅有一处漏检。综上,WristXNet 算法在准确性、特征关注能力以及漏检减少方面表现出显著的优势。其有效的特征关注能力使其能够精准识别复杂的损伤区域,且在特征处理的一致性和精度上优于对比算法。

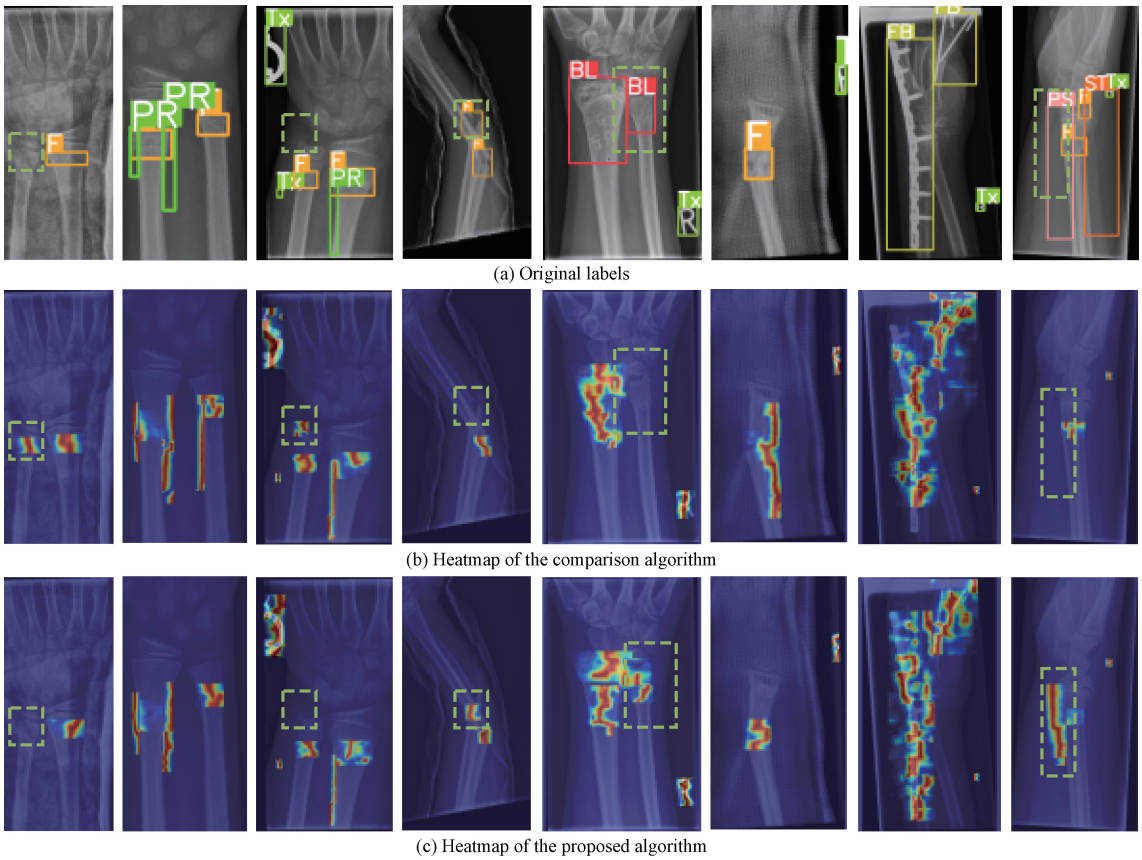


图 7　本文算法与对比算法预测结果比较

Fig. 7　Comparison of prediction results between the proposed algorithm and the comparison algorithm

## 3　结　　论

为更好的应对儿童手腕创伤 X 光图像的检测挑战,本文自主设计一种深度卷积网络模型 WristXNet。通过构建多尺度注意力特征聚合模块 C2f_MSAF、设计混合池化空间金字塔模块 HPSP、引入动态上采样模块 DySample 以及构建 LDDHead 检测头,使得 WristXNet 模型在精确检测和细粒度特征捕捉方面取得了突破。在儿童手腕创伤公开数据集 GRAZPEDWRI-DX 上的实验结果表明,WristXNet 在骨病变、旋前肌征、软组织肿胀、异物和骨膜反应等多种损伤类型的检测中均超越现有主流算法,整体检测精度达到 68.5%。WristXNet 通过有效的特征提取和准确的区域关注,适合应用于儿童手腕创伤的自动检测任务,具有潜在的临床应用价值。

## 参考文献

[1] HUSSAIN F, COOPER A, CARSON-STEVENS A, et al. Diagnostic error in the emergency department: Learning from national patient safety incident report analysis [J]. BMC Emergency Medicine, 2019, 19(1):77.

[2] 吴慧东,刘立程,潘丹. 结合视图感知 CNN 和 Transformer 的阿尔茨海默病诊断研究[J]. 电子测量技术,2025,48(1):145-153.
WU H D, LIU L CH, PAN D. Combining view-aware CNN and transformer for Alzheimer′s disease diagnosis research [J]. Electronic Measurement Technology, 2025,48(1):145-153.

[3] SHEN X R, ZHOU Y X, SHI X Y, et al. The application of deep learning in abdominal trauma diagnosis by CT imaging [J]. World Journal of Emergency Surgery, 2024, 19: 17.

[4] 周林鹏,姚剑敏,严群,等. 融合多尺度特征及注意力机制的医学图像检索[J]. 液晶与显示,2021,36(8):1174-1185.
ZHOU L P, YAO J M, YAN Q, et al. Medicalimage retrieval with multiscale features and attention mechanisms[J]. Chinese Journal of Liquid Crystals and Displays, 2021,36(8):1174-1185.

[5] 张物华,李锵,关欣. 基于多尺度卷积神经网络的 X 光

图像中肺炎病灶检测[J].激光与光电子学进展,2020,57(8):187-194.
ZHANG W H, LI Q, GUAN X. Detection of pneumonia lesions in X-Ray images based on multi-scale convolutional neural networks[J]. Laser & Optoelectronics Progress, 2020,57(8):187-194.

[6] 车翔玖,董有政.基于多尺度信息融合的图像识别改进算法[J].吉林大学学报(工学版),2020,50(5):1747-1754.
CHE X J, DONG Y ZH. Improved image recognition algorithm based on multi-scale information fusion[J]. Journal of Jilin University (Engineering and Technology Edition),2020,50(5):1747-1754.

[7] DIBO R, GALICHIN A, ASTASHEV P, et al. Deeploc: Deep learning-based bone pathology localization and classification in wrist x-ray images[C]. International Conference on Analysis of Images, Social Networks and Texts, 2023: 199-211.

[8] 黄泽青,刘予豪,方汉军,等. 基于深度迁移学习模型实现股骨头坏死与其他髋部疾病的X线片鉴别诊断[J].中华骨科杂志,2023,43(1):72-80.
HUANG Z Q, LIU Y H, FANG H J, et al. A deep transfer learning method using plain radiographs for the differential diagnosis of osteonecrosis of the femoral head with other hip diseases[J]. Chinese Journal of Orthopaedics, 2023, 43(1): 72-80.

[9] JABBAR J, HUSSAIN M, MALIK H, et al. Deep learning based classification of wrist cracks from X-ray imaging[J]. CMC-Computers Materials & Continua, 2022, 73(1): 1827-1844.

[10] SCHILCHER J, NILSSON A, ANDLID O, et al. Fusion of electronic health records and radiographic images for a mul-timodal deep learning prediction model of atypical femur fractures[J]. Computers in Biology and Medicine, 2024, 168: 107704.

[11] JU R Y, CAI W M. Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm[J]. Scientific Reports, 2023, 13(1): 20077.

[12] CHIEN C T, JU R Y, CHOU K Y, et al. YOLOv8-AM: YOLOv8 based on effective attention mechanisms for pediatric wrist fracture detection[J]. ArXiv preprint arXiv:2402.09329, 2024.

[13] AHMED A, MANAF A. Pediatric wrist fracture detection in X-rays via YOLOv10 algorithm and dual label assignment system[J]. ArXiv preprint arXiv: 2407.15689, 2024.

[14] TERVEN J, CÓRDOVA-ESPARZA D M, ROMERO-GONZÁLEZ J A. A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS[J]. Machine Learning and Knowledge Extraction, 2023, 5(4): 1680-1716.

[15] Ultralytics. (2023). YOLOv5 [EB/OL]. [2025-03-09]. https://github.com/ultralytics/yolov5.

[16] LI CH Y, LI L L, JIANG H L, et al. YOLOv6: A single-stage object detection framework for industrial applications[J]. ArXiv preprint arXiv:2209.02976, 2022.

[17] WANG C Y, YEH I H, LIAO H Y M. YOLOv9: Learning what you want to learn using programmable gradient information[C]. European Conference on Computer Vision, 2024:1-21.

[18] LIU W Z, LU H, FU H T, et al. Learning to upsample by learning to sample[C]. IEEE/CVF International Conference on Computer Vision, 2023: 6027-6037.

[19] ZHENG ZH H, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]. AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.

[20] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]. IEEE International Conference on Computer Vision, 2017: 2980-2988.

[21] NAGY E, JANISCH M, HRZIC F, et al. A pediatric wrist trauma X-ray dataset(GRAZPEDWRI-DX) for machine learning[J]. Scientific Data, 2022,9(1):222.

[22] JU R Y, CHIEN C T, LIN C M, et al. Global context modeling in YOLOv8 for pediatric wrist fracture detection[C]. 2024 International Symposium on Intelligent Signal Processing and Communication Systems(ISPACS), 2024: 1-5.

[23] WANG AO, CHEN H, LIU L H, et al. YOLOv10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[24] KHANAM R, HUSSAIN M. YOLOv11: An overview of the key architectural enhancements[J]. ArXiv preprint arXiv:2410.17725, 2024.

[25] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]. European Conference on Computer Vision, 2020: 213-229.

[26] ZHAO Y A, LYU W Y, XU SH L, et al. Detrs beat YOLOs on real-time object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.

[27] LIU SH L, LI F, ZHANG H, et al. Dab-detr: Dynamic anchor boxes are better queries for detr[J]. ArXiv preprint arXiv:2201.12329, 2022.

[28] ZHANG H, LI F, LIU SH L, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection[J]. ArXiv preprint arXiv:2203.03605, 2022.

## 作者简介

**林淑娟**,硕士研究生,主要研究方向为计算机视觉与图像处理。
E-mail:1921669437@qq.com

**钟铭恩**(通信作者),博士,教授,主要研究方向为计算机视觉和人工智能。
E-mail:zhongmingen@xmut.edu.cn