

融合多尺度特征和子空间注意力的黏膜下肿瘤检测

蔡贤杰¹, 丁德锐¹, 魏国亮², 武俊珂³

(1. 上海理工大学 光电信息与计算机工程学院, 上海 200093; 2. 上海理工大学 管理学院, 上海 200093;

3. 上海理工大学 理学院, 上海 200093)

摘要: 计算机辅助检测工具可以帮助医生减少在临床检查中漏检误检的情况, 从而提高诊断准确度, 同时减轻医生的劳动强度。针对超声胃肠镜检查中黏膜下肿瘤的定位与分类问题, 提出了一种融合多尺度特征和子空间注意力的黏膜下肿瘤检测算法(MFSA-YOLOv7t)。首先, 移除小目标预测头, 在保证精度下使网络轻量化; 然后, 基于浅层特征提出多尺度特征融合模块, 提取肿瘤细节信息; 其次, 改进上采样结构, 在保留上层信息的同时增强感受野; 最后, 引入子空间位置注意力模块, 捕获肿瘤的位置和边界特征, 进一步提升黏膜下肿瘤的检测性能。实验表明, MFSA-YOLOv7t 在平均精度均值、敏感度以及准确度上分别达到 97.32%, 96.99% 和 96.24%, 相比 YOLOv7-tiny 算法检测性能有较大的提升, 分别提高了 2.39%, 2.75% 和 2.59%。MFSA-YOLOv7t 为医生在临床检查中的辅助诊断提供更加可靠的肿瘤类型参考, 同时为黏膜下肿瘤的检测提供了一个新的思路和研究方向。

关键词: 黏膜下肿瘤; 计算机辅助检测; 多尺度特征; 上采样; 子空间注意力

中图分类号: TP 391.1

文献标志码: A

Detection method in submucosal tumor based on multi-scale feature and subspace attention fusion

CAI Xianjie¹, DING Derui¹, WEI Guoliang², WU Junke³

(1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. Business School, University of Shanghai for Science and Technology, Shanghai 200093, China; 3. College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Computer-aided detection tools can help doctors reduce the situations of missed and false detections in clinical examinations to improve diagnostic accuracy and reduce the labor intensity of doctors. Aiming at the problem of the localization and identification of submucosal tumors in ultrasound gastroscopy, a submucosal tumor detection algorithm combined with multi-scale feature and subspace attention (MFSA-YOLOv7t) was proposed. Firstly, the small target prediction head was removed to make the network lightweight while maintaining accuracy. Secondly, A multi-scale feature fusion module was proposed based on shallow features to aggregate details. Then, the algorithm improved the

收稿日期: 2022-08-30

基金项目: 国家自然科学基金资助项目 (61973219)

第一作者: 蔡贤杰 (1997-), 男, 硕士研究生。研究方向: 目标检测、医学图像处理。E-mail: 1210702806@qq.com

通信作者: 丁德锐 (1981-), 男, 教授。研究方向: 图像处理、计算机视觉。E-mail: deruiding2010@usst.edu.cn

up-sampling structure to retain the upper layer information to the greatest extent and enhance the feature perception field. Finally, the coordinate subspace attention module was introduced to capture tumor location and boundary features to improve the detection performance of submucosal tumors. Experiments show that the mAP, sensitivity, and accuracy of MFSA-YOLOv7t reach 97.32%, 96.99%, and 96.24%, respectively, compared with YOLOv7-tiny, MFSA-YOLOv7t has a great improvement in detection performance, which is improved by 2.39%, 2.75%, and 2.59%, respectively. MFSA-YOLOv7t provides a more reliable reference of tumor type for doctors in clinical examination of the auxiliary diagnosis and also provides a new idea and a research direction for detecting submucosal tumors.

Keywords: *submucosal tumor; computer-aided detection; multi-scale feature; up sample; subspace attention*

消化道黏膜下肿瘤^[1](submucosal tumor, SMT)是一类临床上罕见的, 来源于非黏膜层的消化道肿瘤性病变, 包含平滑肌瘤、脂肪瘤、胃肠道间质瘤、血管瘤及异位胰腺等。在超声内镜下, SMT主要集中在黏膜肌层、黏膜下层、固有肌层。SMT以良性病变为主, 但仍有约15%的SMT具有恶性倾向^[2]。临床上需要根据不同类型的SMT, 决定是手术切除或还是随访。

传统的人工诊断是一个完全依靠医生凭借其专业知识储备和丰富的超声影像诊断经验来判别肿瘤的过程。因此, 医生的主观想法对SMT的诊断结果影响较大, 且缺乏诊断经验的医生容易出现误诊情况。研究表明, 利用计算机辅助诊断(computer aided diagnosis, CAD)系统对超声图像进行自动且精确的定位、分割或分类, 不仅能为医生提供了更加客观的诊断参考, 也能提高医生的诊断效率和准确度^[3]。因此, 利用CAD系统对超声图像进行良恶性判别具有重要的研究价值和应用前景^[4]。目前, 在医学图像处理及CAD领域, 人工智能算法与医学临床诊断评估方法紧密地结合在一起, 形成了一个新的研究方向, 用于解决临床诊断过程中的图像降噪、目标识别、组织分割、疾病预测等问题。

近年来, 基于端到端的深度学习^[5]在医学图像领域取得了一系列的成果^[6-8]。不同于传统方法, 深度学习无需人工的目标定位、分割以及特征提取等步骤, 而是直接输入图像及标签进行自我学习, 得到一系列模型参数, 对输入的图像进行预测。人工智能技术在医学图像处理领域的主要研究场景以分类为主, 检测为分类的衍生, 在兼顾分类的同时进行目标定位。国内外研究人员基于此方向作了相关研究: Xu等^[9]以YOLOv3^[10]

为基础网络, 引入Inception结构^[11], 即多尺度卷积模块, 通过解决传统的卷积层捕获特征信息不足的问题来提高肺结节检测效果。Zhang等^[12]通过使用基于回归的目标检测方法ResYOLO, 能够有效地提取结直肠息肉的空间位置信息, 经实验证明该检测方法具有较高的检测精度, 并且可以准确定位视频帧中的结直肠息肉。余逸飞等^[13]以Faster RCNN为基础网络, 构建CBAM(convolutional block attention module)骨干模块, 以提高模型对食道癌病灶区域亮度值偏高或偏低的病灶区域检测能力。孙跃军等^[14]通过图像融合技术提升模型的泛化性, 并通过通道混合技术实现空间的信息交流, 增强模型的表征能力, 以提升检测结果。Kumar等^[15]提出了一种混合深度的ResNet脑肿瘤检测和分类算法, 引入灰度共生矩阵的相关特征用于辅助分类, 并结合迁移学习提升检测结果的准确性。马金林等^[16]使用结合注意力机制的FPN(feature pyramid networks)结构, 通过进一步融合高低层特征的语义信息提高模型的特征提取能力, 从而提高模型对肝脏肿瘤的检测效果。

作为一种相对较为罕见的肿瘤, SMT的诊断对于医生来说具有一定的挑战性。近年来, 人工智能算法在医学图像领域的应用逐渐广泛, 取得了一定的成效。

因此, 本文提出了一种基于改进单阶段网络的黏膜下肿瘤检测方法。

a. 设计了一个用于检测SMT的单阶段算法MFSA-YOLOv7t。

b. 设计了一个多尺度特征融合模块(multi-scale feature fusion module, MFFM), 通过主干网络中不同尺度特征图进行融合, 构建新的特征提取分

支, 关注肿瘤细节信息, 同时弥补后续池化的相关信息丢失。

c. 融合了一个新的上采样结构 (Fuse-up, Fup), 用于保留上文特征信息, 以此增大感受野。

d. 结合位置与通道注意力, 提出了一种子空间位置注意力模块 (coordinate subspace attention module, CSAM), 用于关注特征图的感兴趣区域, 从而提升网络的检测效果。

1 相关工作

1.1 MFSA-YOLOv7t 网络结构

通常, 深度学习需要大量的数据样本进行标记训练, 深层或复杂的网络对于大样本数据有更好的检测效果, 由于本文所用的数据集样本较少, 且图像特征只包含背景和标注样本两部分, 采取深层或复杂的网络模型容易导致过拟合并增加训练成本。综上, 考虑到本文数据样本量的局限性, 以及后续模型训练的稳定性, 本文基于 YOLOv7^[17] 网络思想, 选择 YOLOv7-tiny 网络针对本文数据集进行后续研究。

本文设计的网络 MFSA-YOLOv7t 结构如图 1 所示, 主要包括用于特征提取的主干部分 (Backbone)、用于特征融合的颈部 (Neck)、和用于预测分类的头部 (Head) 三大部分。首先将预处理后的超声图像输入网络, 然后在 Backbone 中利用两组 CBS 模块和 4 组高效聚合网络 ELAN-T 模块提取图像特征, 其次在 Neck 中经过上采样扩大感受野并拼接 MFFM, 融合浅层细节信息获得增强

的 16 倍下采样特征图, 经过注意力模块 CSAM, 得到同时含有位置和空间信息的特征图, 并依次生成 $26 \times 26 \times 256$ 、 $13 \times 13 \times 512$ 的特征图用于检测中目标和大目标, 其中每个特征图对应 3 种大小不同的锚框 (Anchor) 负责预测目标, 最后在预测模块 Predict 中利用非极大值抑制 (nonmaximum suppression, NMS) 选择最优预测结果。

1.2 预测头选择

一般用于检测的模型会设置 3 个预测头来预测大中小 3 个尺寸的目标, 同时每个预测头由 3 种不同比例的锚框负责预测不同比例的目标。相比一些经典算法, 如 Faster R-CNN^[18], SDD^[19] 等都是根据经验预设锚框的高和宽, 只适用于特定场景下的目标尺寸, 无法适应其他任务。YOLO (you only look once) 目标检测系列对数据集采用 K-means 自适应锚框聚类算法, 得到符合数据集的锚框大小。本实验利用 K-means 算法对现有肿瘤数据集聚类, 得到 3 组初始候选框, 如表 1 所示。

小目标指在 256×256 的图像中目标面积小于 80 个像素, 即小于 256×256 的 0.12% 就为小目标。由于本文设置图像大小为 416×416 , 且最小初始候选框为 37×34 , 得到目标占比为 0.72%。因此, 考虑删除小目标预测头, 以达到轻量化模型的效果。基础网络在 Neck 中首先经过 2 次上采样扩大感受野, 输出分支 Predict 用于预测小目标, 然后经过 2 次下采样分别输出分支 Predict 用于预测中目标和大目标。本文保留 YOLOv7-tiny 的主干网络, 如图 1 所示, 对 Neck 中 Concat 操作后的第二次上采样及后续部分网络模块进行移除, 将

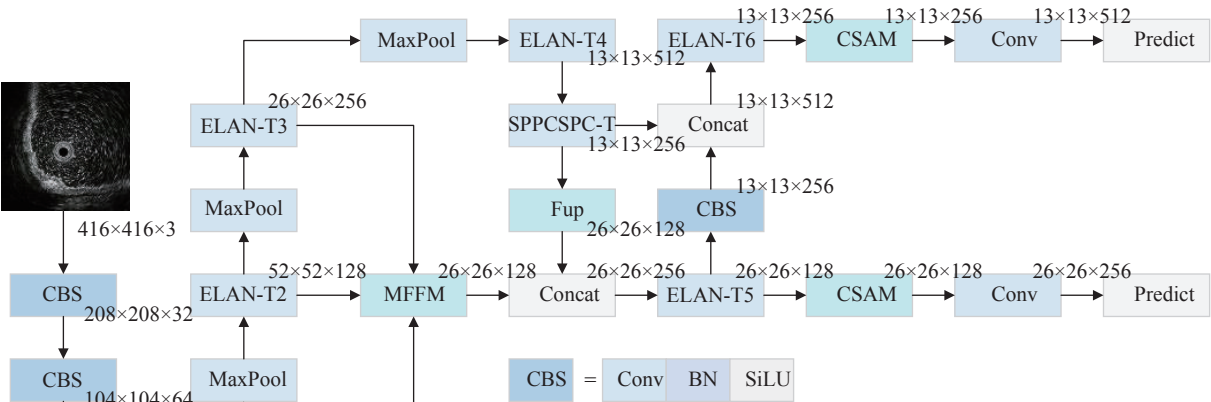


图 1 网络结构图

Fig.1 Network architecture

表 1 预测头锚框信息

Tab.1 Anchor information for prediction head

输出层	输出尺寸	初始锚框尺寸
特征输出层1	52×52	(37×34, 54×49, 64×66)
特征输出层2	26×26	(88×79, 79×113, 115×89)
特征输出层3	13×13	(108×129, 152×117, 161×170)

Concat 操作直接连接至引出中目标预测分支的 ELAN-T5 模块。通过减少预测头，不仅使模型轻量化，而且降低了模型的深度，相对于基础网络，有效地避免了过拟合现象发生，同时提升对浅层特征的学习能力。新的候选初始框如表 2 所示。

表 2 移除小目标预测头的锚框信息

Tab.2 Anchor information for removing small prediction head

输出层	输出尺寸	初始锚框尺寸
特征输出层1	26×26	(49×45, 63×64, 98×83)
特征输出层2	13×13	(83×115, 115×127, 158×160)

1.3 多尺度特征融合模块

在网络中，语义信息包括浅层语义和深层语义，浅层语义即浅层细节特征，由于经过较少的卷积操作，具有分辨率高、细节丰富的特点，如轮廓、纹理、颜色等；深层语义经过多次卷积，感受野大，对细节的感知能力较差，但一个像素点可代表一个区域，获得的全局语义信息更加丰富。考虑到前文中预测头的筛选，导致对浅层信息没有很好地利用，而融合不同尺度的特征信息是提高检测效果的一个重要方法，并且池化之前的特征图在很大程度上保留了原始图像的细节信息。因此，选择 ELAN-T1，T2，T3 这 3 个输出特征图作为特征融合的基底。此外，在 3 种不同尺度融合的过程中必不可少地会出现维度压缩，使用步长为 2 的卷积会使特征图的部分信息丢失，

从而失去表征能力。为了尽可能地保留浅层的信息，提出了一个空间维度切分至通道 (spatial to channel, STC) 的操作用来代替步长为 2 的卷积，如图 2 所示，首先将 $W \times H \times C$ 大小的特征图根据下采样倍数 S ($S=2$) 按空间维度进行划分，得到 4 个子特征图，每个特征图的尺寸为 $W/2 \times H/2 \times C$ ，然后 Concat 操作将每个特征图按通道拼接得到大小为 $W/2 \times H/2 \times 4C$ 的特征图，最后利用 1×1 卷积进行通道整合。 W ， H ， C 分别为网络层张量的宽度、高度和通道数。

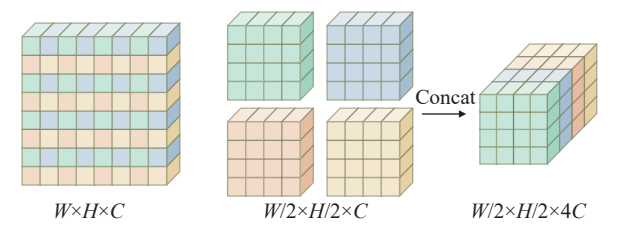


图 2 STC 流程图
Fig.2 Flow chart of STC

基于 STC 结构，本文设计的多尺度特征融合模块 MFFM 如图 3 所示，首先 ELAN-T1 模块通过 1×1 卷积压缩通道降低模型复杂度，然后利用 STC 进行维度压缩同时扩展通道，保留浅层的原始信息，再次通过 1×1 卷积压缩通道进行通道重组，同时与压缩通道的 ELAN-T2 模块拼接以融合多尺度特征信息，重复上述操作再与 ELAN-T3 模块拼接，最后利用 1×1 卷积进行通道重组。该多尺度特征融合分支包含了大量的局部细节信息，与主分支进行拼接互相补充，提升模型的检测能力。

1.4 上采样模块

上采样是一种对图像进行扩充放大并生成对应高分辨率图像的方法。本文认为上采样应具有较大的感受野，更好地利用特征及其周围的信息，且上采样应与特征图的语义信息相关。为弥补上采样过程中特征信息丢失的问题，参考

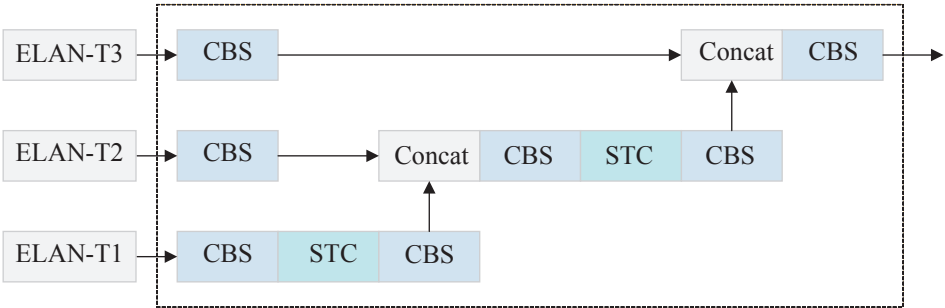


图 3 多尺度特征融合模块
Fig.3 Multi-scale feature fusion module

STC 结构, 提出了一种通道切分 (channel to spatial, CTS) 的上采样结构, CTS 是一种利用通道信息转换到空间维度的上采样方法, 上层特征图含有丰富的语义信息, 通过直接通道切片并补充到空间维度可以更好地保留图像的原始语义特征信息。

本文设计的 Fup 模块如图 4 所示。具体分为两部分, 首先对输入的特征图利用 1×1 卷积压缩通道, 减少后续计算量, 然后通过 Up 上采样输出第一特征图, 其次通过两组 CTS 模块并联输出, 然后拼接得到与第一特征图同尺寸的特征图, 并与之进行叠加操作逐元素相加, 保持通道数不变, 以增加每个通道特征图的信息量, 突出细节信息, 最后通过 1×1 卷积进行特征重组, 得到最终上采样结果。本文使用 Fup 模块替换网络中原有的上采样操作, 以提升网络模型的检测精度。

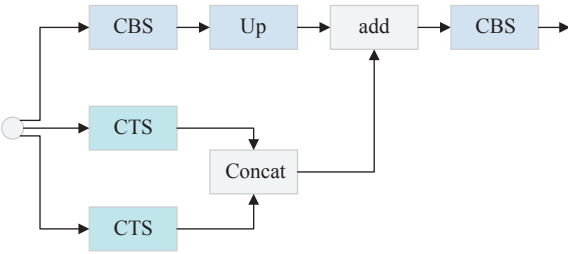


图 4 上采样模块
Fig.4 Fuse-up module

1.5 子空间注意力模块

注意力机制源于对人类视觉的研究, 通过有选择性地专注于显著区域, 从而获得更好的视觉效果。神经网络中的注意力机制指在全局信息中

重点关注当前特征图中的显著区域。因此, 引入注意力机制能够使网络在训练时抑制次要信息, 增强特征的表征能力, 提高深度学习模型的性能。

SMT 在超声内镜下观察属于消化道内壁的凸起物及其位于黏膜层内部的不同层次, 通过高低回声反馈, 网络会在目标边界处理上产生一定误差, 影响定位和分类性能。为了进一步提高模型对于肿瘤边界的检测能力, 可以引入轻量化注意力模块。为了使深层特征图包含更多的远程依赖信息, 并且尽可能地保证网络计算量, 本文参考文献 [20] 的子空间特征图划分思想, 以及文献 [21] 的位置注意力思想, 在主分支 Neck 层的 ELAN-T5 模块, ELAN-T6 模块之后引入 CSAM, 该方法仅添加少量的参数就可以计算子空间的跨通道信息交互和特征图的位置信息。为选择合适的 CSAM 添加位置以输出更好的表征能力, 本文提出了两种策略: a. 考虑双分支并行结构, 基于 MFFM 和 Fup 拼接后的特征层, 设计新的分支添加 CSAM, 再分别于 ELAN-T5 和 ELAN-T6 作叠加操作。b. 考虑深度融合特征, 直接分别在 ELAN-T5 和 ELAN-T6 之后添加 CSAM。经过实验对比, 本文最终选择策略 b 为优选策略。

CSAM 结构如图 5 所示, 通过对特征图进行分组注意力计算, 使用不同子空间之间的线性关系来整合跨通道信息, 同时将位置信息嵌入到通道注意力, 利用 2 个并行的一维特征编码, 高效地整合空间坐标信息。该方法对于复杂的肿瘤边界层次具有更好的检测效果。

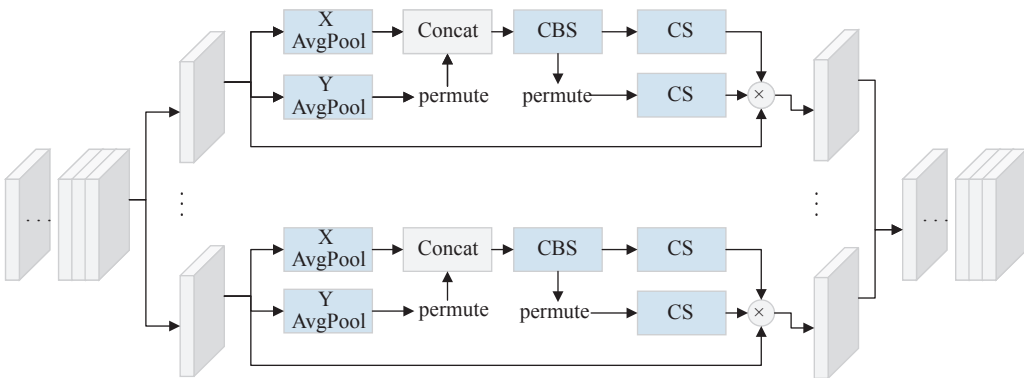


图 5 子空间位置注意力模块
Fig.5 Coordinate subspace attention module

$$F = \text{Concat}(F_1, F_2, F_3, \dots, F_g) \quad (1)$$

$$f = \text{CBS}[\text{Concat}(X \text{AvgPool}(\hat{F}_g), P[Y \text{AvgPool}(\hat{F}_g)])] \quad (2)$$

$$f_h = \sigma\{\text{Conv}[\text{split}(f_x)]\} \quad (3)$$

$$f_w = \sigma\{\text{Conv}[\text{split}(f_y)]\} \quad (4)$$

$$F_g = \hat{F}_g f_h P(f_w) \quad (5)$$

式中: F 为输出特征图; F_1, F_2, F_3, F_g 为输出子空间特征图; $Concat$ 为通道拼接操作; $Conv$ 为卷积操作; $\hat{F}_g$ 为输入子空间特征图; $XAvgPool, YAvgPool$ 分别按水平方向、垂直方向平均池化; f_h, f_w 分别为水平方向、垂直方向的注意力权重; CBS 表示卷积、归一化和 SiLU 激活函数联合操作; P 为维度交换操作; $split$ 为维度拆分操作; σ 为 sigmoid 激活函数操作。

首先如式 (2)~(4) 所示, 分别沿着水平和垂直方向进行池化, 得到一对 HC_g, WC_g 大小的具有方向感知的特征编码图, 接着按通道拼接并通过两步卷积操作进行转换, 得到 2 组注意力编码权重。然后如式 (5) 所示, 与原特征图相乘, 得到带有位置注意力的子特征图, 最终的输出 F 即为子特征图拼接得到的带有位置和通道注意力的特征图。 C_g 为子空间张量的通道数。

2 实验与结果分析

2.1 数据集预处理

实验使用的数据集来自于上海交通大学附属第六人民医院消化内镜中心, 数据涵盖不同年龄段、不同性别的病人数据, 具有良好普适性。数据集包括 109 例病人的 423 张图片, 4 种肿瘤病例。其中, 平滑肌瘤 54 例、胃肠道间质瘤 30 例、异位胰腺 10 例、脂肪瘤 15 例。数据集的原始尺寸为 764×572 , 包含机器参数、白光胃肠镜画面等无关信息, 需要选择合适的窗宽窗位之后进行数据处理, 因此, 裁剪原始图片得到最终图片尺寸为 416×416 , 如图 6 所示。由于医学数据集数量有限, 因此, 采用离线数据增强技术扩充数据集, 在原超声肿瘤数据集的基础上通过平移、旋转及镜像等方法将数据量扩充为原来的 5 倍, 最终得到 2115 张图片, 以提升模型的泛化能力, 同时在训练时使用 mosaic 数据增强进一步提升模型的泛化能力。实验按照 6:2:2 对数据集进行划分, 得到 1269 张图片进行训练、423 张图片进行验证、423 张图片进行测试。

2.2 参数设置

实验的硬件环境为 Intel® Xeon(R) Gold 5220R CPU @ 2.20GHz, NVIDIA RTX 3060 GPU, 12GB 显存。软件环境为 Ubuntu20.04, Pytorch 架构, 算法实现编程语言为 Python。

训练时设置学习率为 0.01, 权重衰减系数为

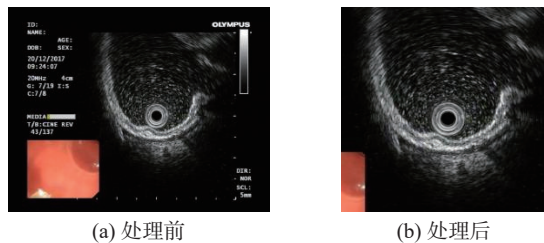


图 6 图片预处理前后比较

Fig. 6 Comparison of images before and after preprocessing

0.0005, 训练次数为 350, 批次大小取 16, 训练迭代次数为 37 600, 动量参数为 0.937。置信度阈值和 IoU 阈值均为 0.5。在训练集、验证集和测试集上的输入图片尺寸统一为 $416 \times 416 \times 3$ 。

2.3 评价指标

在医学上主要有准确度 (accuracy) A_{cc} 、敏感度 (sensitivity) S_e 、特异度 (specificity) S_p 这 3 个指标。假定 TP 表示阳性样本中正确预测为阳性的样本数目, TN 表示阴性样本中正确预测为阴性的样本数目, FP 表示阴性样本中错误预测为阳性的样本数目, FN 表示阳性样本中错误预测为阴性的样本数目。

$$A_{cc} = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

$$S_e = \frac{TP}{TP + FN} \quad (7)$$

$$S_p = \frac{TN}{TN + FP} \quad (8)$$

在目标检测模型中, 通常通过平均精度均值 (mean average precision, mAP)、召回率 (Recall) 来进行评估。mAP 为所有类的平均精度 (AP)。召回率即医学上的敏感度。目标检测中背景和未成功预测的目标都是 N , 因此, 目标检测一般不区分 TN 和 FN , 由于负样本没有标出, 不存在区分正负样本问题, 且本数据集对于每一类别只存在是肿瘤一种情况, 也不存在所谓的阴性样本, 因此, 不考虑特异度指标, 准确度公式也需重新考虑, 其中, $TP + FN$ 为正样本总量, 在目标检测中召回率仍然适用。最终, 本文选取准确度、召回率、mAP 作为本实验的评价指标, 其中, 准确度根据混淆矩阵的每个类别的 TP 之和与总样本数量的百分率作为本文的评价标准。

$$A_{cc} = \sum_{i=1}^4 TP_i / T_a \quad (9)$$

式中, T_a 表示测试集样本总量。

混淆矩阵如图 7 所示，对角线即为各类别检测出的 TP 数量。

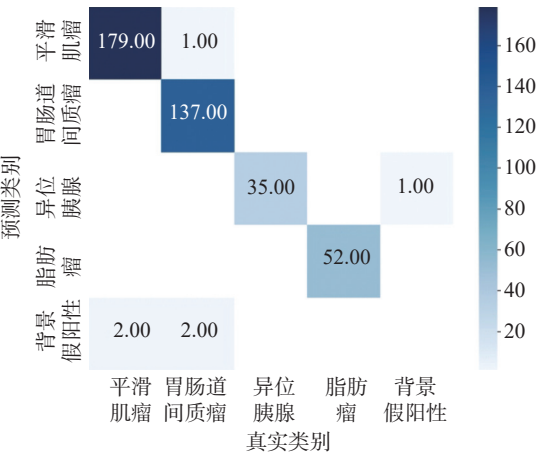


图 7 混淆矩阵
Fig.7 Confusion matrix

2.4 消融实验

为了解双预测头、多尺度特征融合、上采样以及子空间位置注意力对模型精度的影响，现介绍消融实验，主要包括 8 个部分。首先，1 组实验用于验证 YOLOv7-tiny 在 SMT 数据集上的检测性能。其次，2 组实验用于验证改进的双检测头的检测效果。接着，3，4，5 组实验分别用于验证在 2 组上添加 MFFM、Fup 以及策略 b 的 CSAM 的检测效果。然后，6 组实验用于验证在基础网络中加入 MFFM 和 Fup 模块后的检测效果。最后，7，8 组实验用于验证本文提出的 2 种策略下的 3 个模块同时加入网络的检测效果。

实验结果如表 3 所示，针对不同的实验，分别计算其检测精度、参数量 (Params) 及推理速度

FPS(frames per second)。对比 1，2 两组的检测结果发现，mAP，Recall 和准确度分别提高了 0.25%，0.16% 和 0.23%，Params 降低了 0.34 MB 且推理速度显著上升，这说明虽然检测效果提升很小，但在保证网络性能的情况下能使网络轻量化。3 组表示在 2 组基础上添加了 MFFM。对比 2，3 两组的检测结果发现，mAP，Recall 和准确度分别提高了 0.49%，0.80% 和 0.71%，这说明通过对浅层特征进行多尺度融合，使得网络保留了肿瘤更多浅层细节信息，再与深层特征拼接，提升了检测效果。4 组表示在 2 组基础上添加了 Fup 模块。对比 2，4 两组的检测结果发现，mAP，Recall 和准确度分别提高了 0.07%，0.43% 和 0.47%，这说明融合了通道信息的上采样操作在一定程度上保留了特征图上层语义信息，能够提升模型的检测效果。5 组表示在 2 组基础上添加了 CSAM。对比 2，5 两组的检测结果发现，mAP，Recall 和准确度分别提高了 0.81%，1.12% 和 1.18%，这说明在输出端加入子空间位置注意力模块，能较好地进行跨通道信息交互，同时关注肿瘤的语义特征信息，证明了该模块的有效性。6 组表示在 2 组基础上添加了 MFFM 和 Fup。对比 2，6 两组的检测结果发现，mAP，Recall 和准确度分别提高了 1.01%，1.48% 和 1.18%，说明将融合浅层特征的结果与能保留更多上层语义的上采样模块拼接能进一步提高网络的检测效果。8 组表示本文算法，对比 7 组的检测结果发现，经过深度融合后的特征具有丰富的语义信息和细节特征，因此，选择策略 b 为优选策略。对比前 6 组实验结果发现，本文所提出的网络在病灶区域上取得了最好的检测性能。

表 3 消融实验
Tab.3 Ablation studies

实验组	消融方法						平均精度均值/%	召回率/%	准确度/%	参数量/MB	推理速度
	基础网络	双预测头	MFFM	Fup	CSAM1	CSAM2					
1	√						94.93	94.24	93.65	6.02	72
2	√	√					95.18	94.40	93.88	5.68	82
3	√	√	√				95.67	95.20	94.59	5.77	78
4	√	√		√			95.25	94.83	94.35	5.72	78
5	√	√				√	95.99	95.52	95.06	5.71	63
6	√	√	√	√			96.19	95.78	95.06	5.80	76
7	√	√	√	√	√		96.97	96.63	95.53	5.99	60
8(本文)	√	√	√	√		√	97.32	96.99	96.24	5.83	62

同时相较于1组实验，mAP，Recall和准确度分别提高了2.39%，2.75%和2.59%，这说明了本文模块的有效性。此外，观察Params一栏发现，对比2组实验即双预测头网络，本文在只增加少量参数的情况下，检测结果就能有明显的提升；对比1组实验即原始基础网络，本文在使网络更加轻量化的同时，显著地提升了模型的检测效果，说明了本文模型的优越性。

为进一步体现MFSA-YOLOv7t在网络中的特征提取能力和关注全局显著信息能力，通过绘制热力图Grad-CAM(*gradient-weighted class activation mapping*)展现模块的改进优势。如图8所示，颜色越红，表示网络对于该特征越敏感。其中，图8(a)和(b)表示Fup对应的Grad-CAM，图8(c)和(d)表示MFFM对应的Grad-CAM，图8(e)和(f)表示CSAM对应的Grad-CAM，3组改进模块的Grad-CAM分别为网络中对应层次输出，且每组图示中左图对应基础网络，右图对应MFSA-YOLOv7t。如图8(a)和(b)所示，添加Fup模块使上采样操作保留更多肿瘤的外侧边界、轮廓等细节信息，弥补了上采样过程中特征信息丢失的问题。如图8(c)和(d)所示，添加MFFM并与Fup进行拼接融合操作，随着网络层次加深，使得网络在保留目标细节的同时开始关注肿瘤的语义特征。如图8(e)和(f)所示，引入注意力模块CSAM，通过跨通道信

息交互及捕捉位置边界特征，使网络重点关注目标肿瘤的语义特征，弱化次要背景信息，从而提升模型的检测能力。

2.5 对比实验

主要设计两部分实验。第一部分进行纵向对比，验证MFSA-YOLOv7t对肿瘤各类别的检测效果，第二部分将MFSA-YOLOv7t与其他比较流行的深度学习目标检测算法进行横向对比实验，验证本文算法的可行性和有效性。如表4所示，a代表平滑肌瘤，b代表胃肠道间质瘤，c代表异位胰腺，d代表脂肪瘤。从平均精度AP上看，4种肿瘤的检测指标分别提升了1.64%，2.23%，2.30%，3.29%；从Recall上看，4种肿瘤的指标分别提升了1.67%，3.62%，2.20%，3.57%。说明改进后的算法在特征提取细节信息和语义信息方面有很大的提升，通过多尺度融合提取特征图的细节信息，同时关注肿瘤的边界纹理，使得一些难以检测的、较为模糊的肿瘤被检测出来，说明本文算法的改进能有效地解决YOLOv7-tiny的漏检问题，具有很好的定位效果，证明了本文算法对于黏膜下肿瘤检测的可行性。

为进一步评价本文算法对黏膜下肿瘤的检测性能，将MFSA-YOLOv7t与其他经典算法，如Faster-rcnn，SSD，Retinanet^[22]等，以及最新算法，如VFNet^[23]，YOLOF^[24]，YOLOv5s^[25]等进行

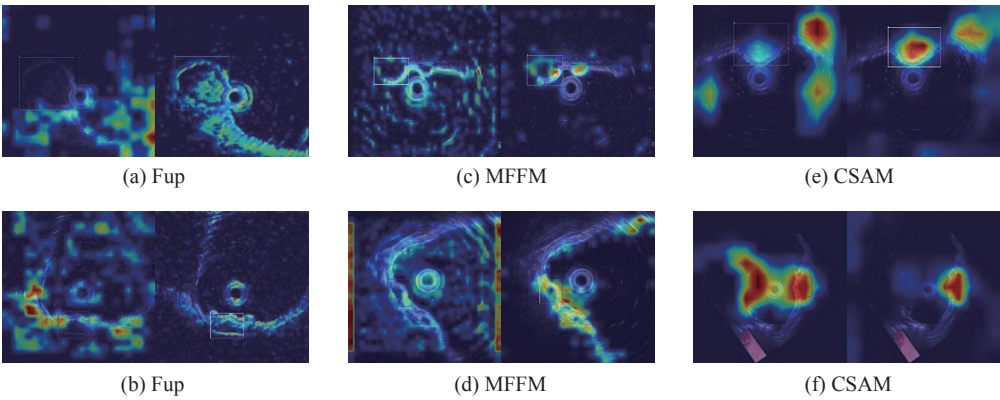


图8 不同改进的热力图对比

Fig.8 Comparison of different improved grad-CAM

表4 数据集中不同目标的检测性能对比

Tab.4 Comparison of detection performance for different targets in the dataset

算 法	平均精度				召回率				平均精度均值	召回率	准确度
	a	b	c	d	a	b	c	d			
YOLOv7-tiny	94.46	94.70	95.42	95.14	94.35	92.86	95.02	94.64	94.93	94.22	93.65
MFSA-YOLOv7t	96.10	96.85	97.72	98.63	96.02	96.48	97.22	98.21	97.32	96.99	96.24

对比。对比结果如表 5 所示, 为保证实验结果的公平性, 这些算法采用了与 MFSA-YOLOv7t 一致的输入尺寸、数据集划分比例、批次大小、初始学习率和损失函数。从表 5 中可以看出, MFSA-YOLOv7t 的 mAP 指标显著优于其他算法, 对比经典算法, 提升了 5%~6%; 对比其他最新算法, 提升了 3%~4%。在推理速度上也优于大部分算法, 且 MFSA-YOLOv7t 与 YOLOv5s 推理速度接近, 得益于 YOLOv7 模型的结构重参数化, 即使加入新的模块, 也能保持一定的推理速度。其中, 对比 MFSA-YOLOv7t 与 Retinanet, VFNet, YOLOF, YOLOv5s 和 YOLOv7-tiny, 相较于传统的 FPN 特征融合模块, 本文提出的融合多尺度浅层特征的模块能显著提升网络的检测能力。同时, 对于引入的子空间位置注意力模块, 不仅融合了含有上下文语义信息的深层特征和含有细节信息的浅层特征, 还进一步地进行跨通道融合实现信息交互, 并重点聚焦肿瘤的边界纹理、关注肿瘤位置, 增强了模型对肿瘤特征的表征能力, 说明 MFSA-YOLOv7t 对于 SMT 具有更好的特征提取能力和目标预测能力。

2.6 结果可视化

为了更加直观地评估本文算法检测性能的优越性, 对部分网络检测结果进行了可视化。如图 9 所示, MFSA-YOLOv7t 检测结果的置信度优于 YOLOv5s 和 YOLOv7-tiny, 说明本文算法能更好地关注 SMT 的病灶区域, 可以有效地检测病灶

表 5 不同算法在数据集上的 mAP 和推理速度对比
Tab.5 Comparison of different algorithm in the dataset with mAP and speed

算 法	平均精度均值/%	推理速度
Faster-rcnn	91.20	33
SSD	90.83	36
RetinaNet	90.01	35
VFNet	93.92	31
YOLOF	93.61	53
YOLOv5s	94.49	63
YOLOv7-tiny	94.93	72
MFSA-YOLOv7t	97.32	62

位置和类型。对比胃肠道间质瘤的检测结果, YOLOv5s 存在误检的情况, 而 YOLOv7-tiny 虽然能够正确检测但置信度偏低, 相比之下, MFSA-YOLOv7t 的检测结果不仅能消除误检, 还有很高的置信度, 并且检测边界也更加贴合病灶区域, 同时也能提高医生的识别可信度。对比脂肪瘤的检测结果, YOLOv5s 和 YOLOv7-tiny 均无法识别肿瘤, 而 MFSA-YOLOv7t 虽然检测的置信度偏低, 但是, 对比其他算法能够大致定位病灶位置, 且能够正确地识别出脂肪瘤。通过可视化对比, 进一步证明 MFSA-YOLOv7t 在进行检测时所提取的特征信息更加准确。另外, 通过病灶区域的可视化能有效地辅助医生在临床中进行快速的肿瘤判别和病情诊断。

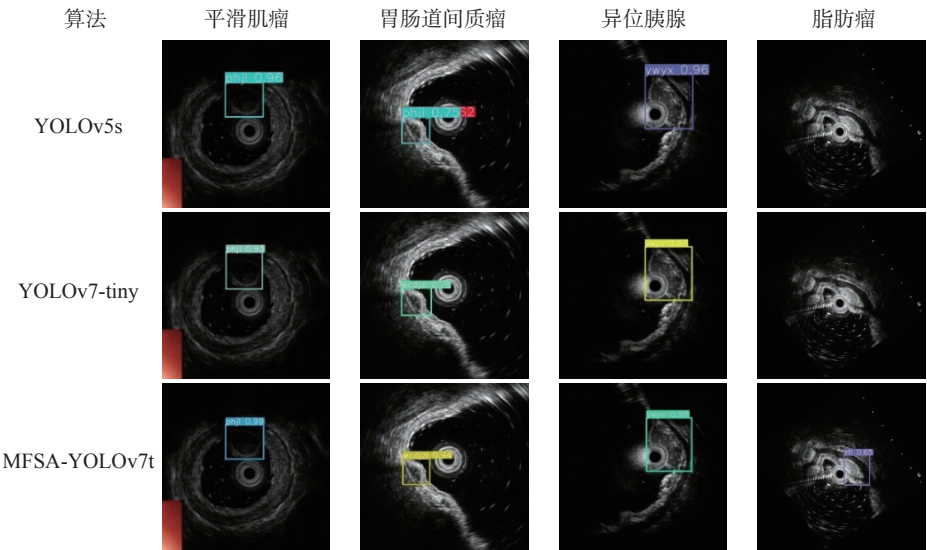


图 9 不同算法的检测结果对比

Fig.9 Comparison of detection results for different algorithms

3 结 论

针对 SMT 的检测问题,提出了一种融合多尺度特征和子空间位置注意力的黏膜下肿瘤检测算法 MFSA-YOLOv7t。首先采用 K-means 算法对 SMT 数据集进行聚类分析,获得 3 组适合该数据集的锚点,并与初始锚点对比,发现小目标检测头存在检测冗余,提出将小目标预测头移除,在保证精度的情况下减少网络参数。其次,针对主干网络存在丰富的细节信息,提出了一种多尺度特征融合模块 MFFM,用于提取浅层特征,同时引入新的上采样结构 Fup,尽可能地保留上层语义信息,并扩大感受野,将 MFFM 和 Fup 拼接,使得浅层细节信息与上层语义信息融合。最后,利用注意力机制增强特征表达能力,提出子空间位置注意力模块 CSAM,用于跨通道信息交互,同时关注肿瘤的位置边界信息,提升网络的分类和定位能力。经实验验证,改进算法提高了检测精度, mAP、敏感度和准确度分别达到 97.32%, 96.99% 和 96.24%, 相比 YOLOv7-tiny 检测性能有较大的提升,分别提高了 2.39%, 2.75% 和 2.59%。在一定程度上解决了模型误检、漏检、定位不准确等问题,取得了良好的检测性能,表明本文算法在 SMT 数据集上具有较好的应用价值。

本文所用的 SMT 数据集仅针对 4 种肿瘤进行研究,但 SMT 肿瘤还包括神经内分泌肿瘤、血管瘤和囊肿等,因此,下一步的研究目标是针对肿瘤是否存在恶性病变倾向进行判别,进一步完善黏膜下肿瘤诊断算法的功能,提高医生在临床诊断中的工作效率。

致谢 本实验数据集是由上海交通大学附属第六人民医院消化内镜中心提供,特别是宛新建主任和赵祥运医生对于本课题的支持,在此表示衷心的感谢!

参考文献:

- [1] GUO J T, LIU Z J, SUN S Y, et al. Endosonography-assisted diagnosis and therapy of gastrointestinal submucosal tumors[J]. *Endoscopic Ultrasound*, 2013, 2(3): 125–133.
- [2] POLKOWSKI M. Endoscopic ultrasound and endoscopic ultrasound-guided fine-needle biopsy for the diagnosis of malignant submucosal tumors[J]. *Endoscopy*, 2005, 37(7): 635–645.
- [3] EADIE L H, TAYLOR P, GIBSON A P. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging[J]. *European Journal of Radiology*, 2012, 81(1): e70–e76.
- [4] SHAO H Y, ZHANG Y T, XIAN M, et al. A saliency model for automated tumor detection in breast ultrasound images[C]//Proceedings of 2015 IEEE International Conference on Image Processing. Quebec City: IEEE Press, 2015: 1424–1428.
- [5] SCHMIDHUBER J. Deep learning in neural networks: an overview[J]. *Neural Networks*, 2015, 61: 85–117.
- [6] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 770–778.
- [7] SHI B G, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(11): 2298–2304.
- [8] GAO Y Q, MOSALAM K M. Deep transfer learning for image-based structural damage recognition[J]. *Computer-Aided Civil and Infrastructure Engineering*, 2018, 33(9): 748–768.
- [9] XU K J, JIANG H, TANG W F. A new object detection algorithm based on YOLOv3 for lung nodules[C]//Proceedings of the 6th International Conference on Computing and Artificial Intelligence. Tianjin: ACM, 2020: 233–239.
- [10] REDMON J, FARHADI A. YOLOv3: an incremental improvement [EB/OL]. [2018-04-08]. <http://arxiv.org/abs/1804.02767>.
- [11] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Press, 2016: 2818–2826.
- [12] ZHANG R K, ZHENG Y L, POON C C Y, et al. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker[J]. *Pattern Recognition*, 2018, 83: 209–219.
- [13] 余逸飞, 高军峰, 闵祥德, 等. 基于 CBAM Faster R-CNN 的食道癌检测 [J]. 中南民族大学学报(自然科学版), 2021, 40(6): 631–638.

- [14] 孙跃军, 屈赵燕, 李毅红. 基于改进的 Mask R-CNN 的乳腺肿瘤目标检测研究 [J]. 光学学报, 2021, 41(2): 0212004.
- [15] KUMAR K S A, PRASAD A Y, METAN J. A hybrid deep CNN-Cov-19-Res-Net Transfer learning architype for an enhanced brain tumor detection and classification scheme in medical image processing[J]. *Biomedical Signal Processing and Control*, 2022, 76: 103631.
- [16] 马金林, 毛凯绩, 马自萍, 等. 基于 ConA-FPN 的肝脏肿瘤检测算法 [J]. 计算机工程与应用, 2023, 59(2): 161–169.
- [17] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023.
- [18] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015: 91–99.
- [19] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016: 779–788.
- [20] SAINI R, JHA N K, DAS B, et al. ULSAM: ultra-lightweight subspace attention module for compact convolutional neural networks[C]//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass: IEEE, 2020: 1627–1636.
- [21] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 13713–13722.
- [22] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2980–2988.
- [23] ZHANG H Y, WANG Y, DAYOUB F, et al. VarifocalNet: an IoU-aware dense object detector[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE Press, 2021: 8514–8523.
- [24] CHEN Q, WANG Y M, YANG T, et al. You only look one-level feature[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE Press, 2021: 13039–13048.
- [25] JOCHER G, CHAURASIA A, STOKEN A, et al. Ultralytics/yolov5: v6.1-TensorRT, TensorFlow edge TPU and OpenVINO export and inference[CP/OL]. [2022-02-22]. <http://github.com/ultralytics/yolov5>.

(编辑: 石 瑛)