DOI:10.19651/j.cnki.emt.2314263

基于 Multi-WHFPN 与 SimAM 注意力机制的版面分割*

杨陈慧[1] 周小亮[1] 张 恒[1] 孙 政[2] 业 宁[1]
(1.南京林业大学信息科学技术学院 南京 210037; 2.南京兰台信息技术有限公司 南京 210009)

摘 要:作为 OCR 的预处理工作,版面分割技术越来越受到学术界和工业界重视。针对版面分割中遇到的检测速度慢、目标区域边界不准确以及细小目标易遗漏等问题,提出了 YOLOv7-MSY 模型。此模型首先借鉴残差连接思想,提出了 Multi-WHFPN 网络结构。它采用可训练的权重参数,突出特征融合过程中特征重要性,并添加了小目标检测头,从而提升对小目标的检测性能;其次,引入 SimAM 注意力机制,可以在不增加额外参数的基础上在 3D 维度评估特征权重,以增强重要特征,抑制无效特征;最后,使用 YEIOU 来代替原模型中的定位损失函数,提升了模型的收敛速度与回归精度。在江苏省档案馆提供的数据集上进行实验对比,YOLOv7-MSY 对目标区域边界检测更加敏感,对细小目标的检测效果更好。YOLOv7-MSY 的 mAP@.5 达到了 0.871,相较于原 YOLOv7 模型提高了 7.84%。该模型的版面分割的效果优于其他类型的版面分割算法,具有良好的泛化性能,并且版面分割速度处于较高水平。

关键词:版面分割;YOLOv7-MSY;Multi-WHFPN;SimAM 注意力机制;YEIOU
中图分类号:TP391.41 **文献标识码**:A **国家标准学科分类代码**:520.6040

Layout segmentation based on Multi-WHFPN and SimAM attention mechanism

Yang Chenhui[1] Zhou Xiaoliang[1] Zhang Heng[1] Sun Zheng[2] Ye Ning[1]
(1. College of Information Science and Technology, Nanjing Forestry University, Nanjing 210037, China;
2. Nanjing Lantai Information Technology Co., Ltd., Nanjing 210009, China)

Abstract: As a pre-processing step for OCR, the layout segmentation technology is receiving increasing attention from both academic and industrial communities. To address the problems encountered in layout segmentation, such as slow detection speed, inaccurate boundary detection of target areas, and easy omission of small targets, the YOLOv7-MSY model is proposed. Firstly, the Multi-WHFPN network structure is proposed by combining the idea of residual connection, and trainable weighted parameters are introduced to highlight the importance of features and add a small target detection head to enhance small target detection. Secondly, the SimAM attention mechanism is introduced to evaluate feature weights in the 3D dimension without adding extra parameters, to enhance important features and suppress invalid features. Finally, the YEIOU is used to replace the original model's localization loss function, which improves the convergence speed and regression accuracy of the model. Experimental comparisons on the dataset provided by the Jiangsu Provincial Archives show that YOLOv7-MSY is more sensitive to boundary detection of target areas and performs better in detecting small targets. The mAP@.5 of YOLOv7-MSY reaches 0.871, which is 7.84% higher than the original YOLOv7 model. The layout segmentation effect of this model is superior to other types of layout segmentation algorithms. It has good generalization performance, and the layout segmentation speed is relatively high.

Keywords: layout segmentation; YOLOv7-MSY; Multi-WHFPN; SimAM; YEIOU

0 引 言

在人工智能日益发展的今天,文档电子化向着自动化方向发展,文字识别技术(optical character recognition, OCR)越来越被学术界和工业界重视。OCR 可以将纸张或者图片上的字符转换成计算机文字,随后可以供用户进一

收稿日期:2023-07-31
*基金项目:国家重点研发计划(2016YFD0600101)项目资助

步加工利用。由于大部分文档不是由单纯的文字组成，可能包含着大量的图片、表格等等，如果对这些文档直接使用 OCR 文字识别技术很难达到理想的效果。版面分割作为 OCR 的预处理工作，将输入的带有文字的图片按照特征分成若干个区域，随后对各个区域分别进行 OCR 文字识别，可以使得 OCR 的识别准确率获得大幅提高。版面分割模型的输出可分为两部分，第 1 部分是分割之后各个区域的坐标信息，第 2 部分是各个区域所属的分类，如标题、图片、文本、表格等。

近年来，版面分割的相关研究主要围绕目标检测算法展开。基于深度学习的目标检测算法大致可以分为两类：单阶段目标检测方法(one-stage)与双阶段目标检测方法(two-stage)。单阶段目标检测方法主要思想是利用回归将图片送入卷积神经网络，经过运算直接得出结果，代表算法有 YOLO、SSD 等；双阶段目标检测算法的主要思想是首先通过卷积神经网络得到候选区域，再基于候选区域进行后续的定位及分类工作，以 Faster-RCNN 算法为代表。在之前关于版面分割的一些工作当中：应自炉等[1]提出了一种多特征融合的卷积神经网络，选取 DeeplabV3 中串并行金字塔策略，并添加图像级特征对提取的特征做进一步优化，最后使用双线性插值法对图像进行恢复，完成对文档中目标的定位与识别，该方法对曼哈顿版面的文档效果良好，但对非曼哈顿版面文档效果欠佳。Tian 等[2]提出了 CPTN 算法，利用微分思想将 Anchor 的横向长度固定为 16 个像素点，竖直方向可调整。通过采用文本线构建策略，将大量小宽度候选框组合成文本行，从而克服通用目标检测框架对长度剧烈变化敏感度的不足，但该算法对曲线、竖直文本的检测效果较差。龚承志[3]基于 Faster-RCNN 模型，通过增加轻量化 PAFPN 模块，优化锚框大小设定，使用多尺度训练机制来提高算法在文档区域定位与分类中的精度，但该算法不是一个端到端的处理方法，需要繁琐的预处理与后处理操作。Liao 等[4]提出的 TextBoxes＋＋算法可以直接在特征图上进行分类与回归，该算法结构简单运行速度快，支持不规则四边形文本检测，但在文本字间距过大、纵向文字的情况下检测效果欠佳。2017 年发表在 CVPR 上的 EAST 算法[5]使用全卷积神经网络模型，在整个图像目标上直接回归文字目标区域的轮廓，并在每个像素位置输出密集的文字预测，但该算法对长文本区域检测效果欠佳。Liu 等[6]提出的 FOTS 算法将文本检测与识别结合到一起，检测与识别模块之间共享卷积特征，降低了模型的计算量，是一个可训练的端到端快速定向文本定位网络，该方法虽然准确率高，但召回率较低。

为了克服上述缺陷并加速版面分割速度，采用 YOLOv7-MSY 模型进行处理。首先通过改进模型中的特征金字塔层，有效降低了特征提取过程中的特征丢失，并且降低了版面分割过程中小目标的遗漏率。其次，引入注意力机制以增强包含目标信息量大的特征，并抑制无用特征。最后为了提升模型效率，增强损失函数收敛稳定性，改进了损失函数，将预测框与真实框的相似度纳入考虑，提升了模型的回归精度。

1　相关工作

1.1　YOLOv7 算法介绍

YOLOv7[7]是最新的 YOLO 系列目标检测器。它在 5～160 fps 帧率范围内的速度和精度超过了大部分目前已知的目标检测器，并且在 GPU V100 上拥有 56.8％AP，准确率居于所有已知的帧率高于 30 fps 的实时目标检测器中最高。YOLOv7 网络结构简单，主要分为输入层、主干网络、特征金字塔层及预测层。如图 1 所示。

1.2　注意力机制

当计算资源有限时，可以通过注意力机制将计算资源分配给更重要的任务[8]。通常情况下，模型所包含的参数越多则模型的表达能力就会更强，与此同时模型存储的信息量也会越大。使用注意力机制可以在输入的信息中增加对重要信息的权重，降低无效信息的权重，以达到过滤有效信息的目的，从而提高任务的处理速度与准确性。常见的有空间注意力机制与通道注意力机制。空间注意力机制工作时寻找图片中最重要的部分进行处理，通道注意力机制会根据通道的重要程度对通道进行加权，来增强或者抑制通道以提升模型效率。

通过对 YOLOv7 网络多次实验发现，在特征提取的过程当中容易忽略一些重要性信息，例如在做报纸版面分割时，容易忽略文章作者，图片题注等区域[9]。通过添加注意力机制增强了 YOLOv7 网络特征提取能力，从而提升模型检测能力。

1.3　回归损失函数

在目标检测中，回归损失函数是评估模型预测框准确性的关键因素之一。目标检测模型通常需要预测每个物体的位置和大小，以便将其从图像中定位出来。在训练期间，模型会学习如何从图像中提取特征，并在预测物体位置时调整预测框的坐标和大小[10]。

回归损失函数是模型训练期间的一个重要组成部分，它通过比较预测框与真实框之间的位置和大小差异来衡量模型的预测准确性。如果模型的预测框与真实框的位置和大小非常接近，损失函数的值将很小，反之则会很大。

在目标检测的实际应用中，回归损失函数的选择会受到多种因素的影响，例如训练集的大小、数据集的特点、模型的架构等。不同的损失函数会对模型的训练和推理产生不同的影响，因此需要在具体场景中进行选择和优化。

2　YOLOv7-MSY 模型的构建

2.1　Multi-WHFPN 网络结构

YOLOv7 模型通过 PANet 网络结构将不同尺度的特征进行融合，PANet 网络结构如图 2 所示。

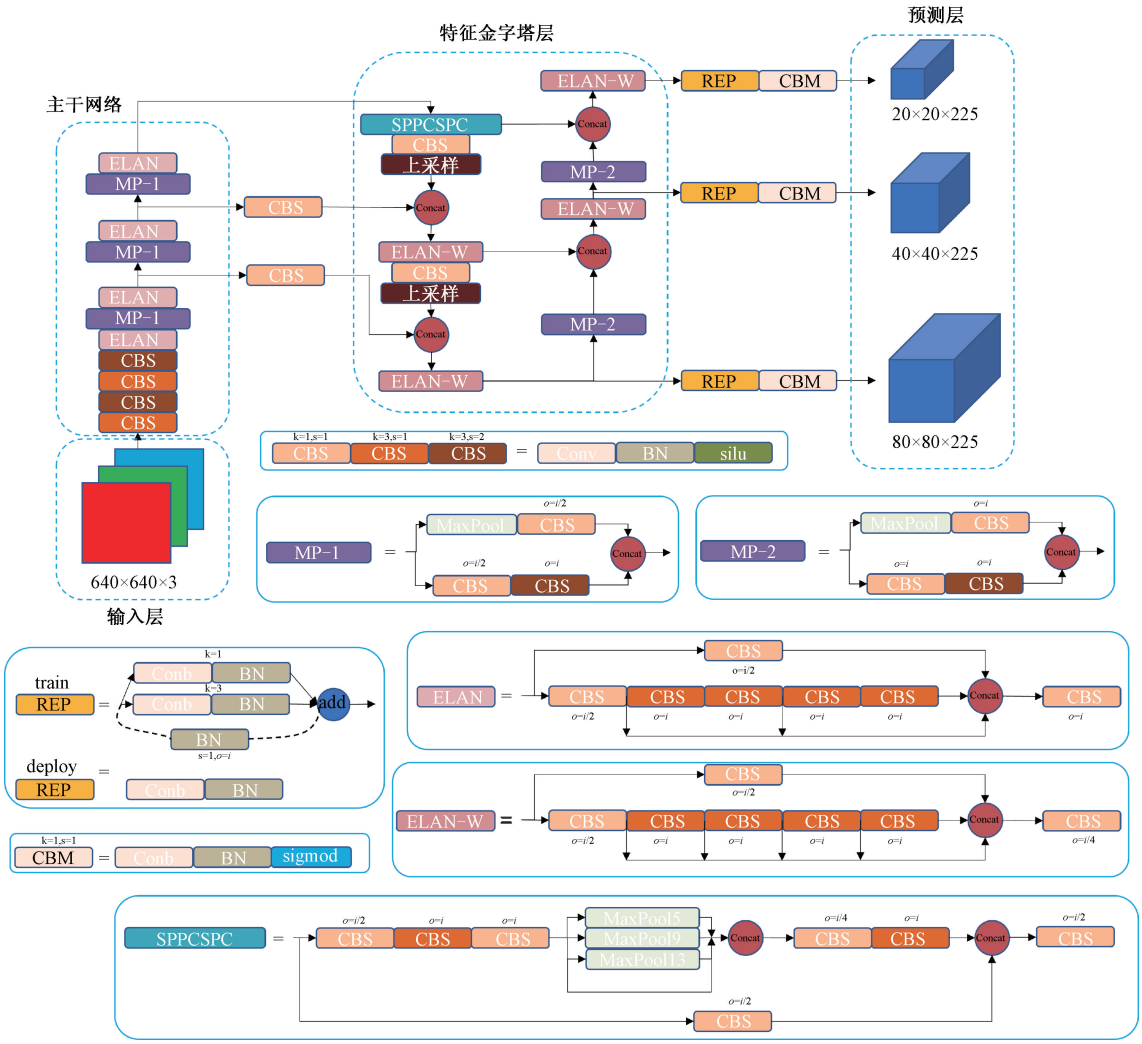

图 1 YOLOv7 结构图

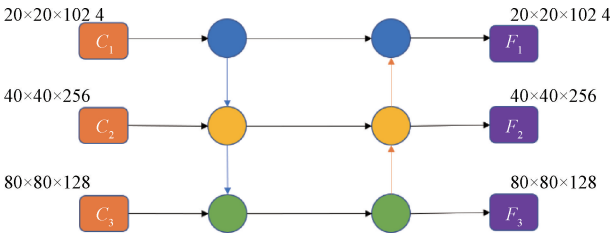

图 2 PANet 网络结构

PANet 网络结构采用固定权重的方式将不同尺度下的特征图进行直接拼接融合,这种方式不够灵活,没有考虑到不同尺度的特征图在特定情况下的重要程度,特征融合不充分。并且该结构存在只有单侧输入且没有特征融合的节点,这些节点对特征融合的影响非常小,并且它会增加额外的参数和计算量,从而导致网络变得更加复杂和低效。

针对以上问题本文提出了加权多层特征融合金字塔(WHFPN)[11],该结构删除了冗余节点,并且对特征进行加权融合,对不同特征的重要程度做出区分。此外为了减少特征丢失,借鉴残差连接的思想在 WHFPN 中添加跳跃结构。并对不同尺度特征层上的节点进行下采样操作,扩大感知野,增强获取全局信息的能力。为了实现更深层次的特征提取,提高模型的预测性能,采用了多重的 WHFPN 结构(Multi-WHFPN)。最后为了提升模型的小目标检测能力在现有的三层检测头的基础上增加一层小目标检测头,将主干网络第一次输出的 160×160 特征图与主干网络输出的 20×20 的特征图经过三次上采样得到的 160×160 特征图进行特征融合,那么可从 640×640 输入图片中检测到的最小目标由 8×8 变为 4×4,缩小为原来的四分之一。Multi-WHFPN 结构如图 3 所示。

Multi-WHFPN 的输入为 $C_1=(20,20,512)$,$C_2=(40,40,256)$,$C_3=(80,80,128)$,$C_4=(160,160,256)$,每个 Multi-WHFPN 节点都会对输入特征进行加权,并使用快速归一化方法对这些权重进行训练。

本文的 Multi-WHFPN 采用的快速归一化融合方法与

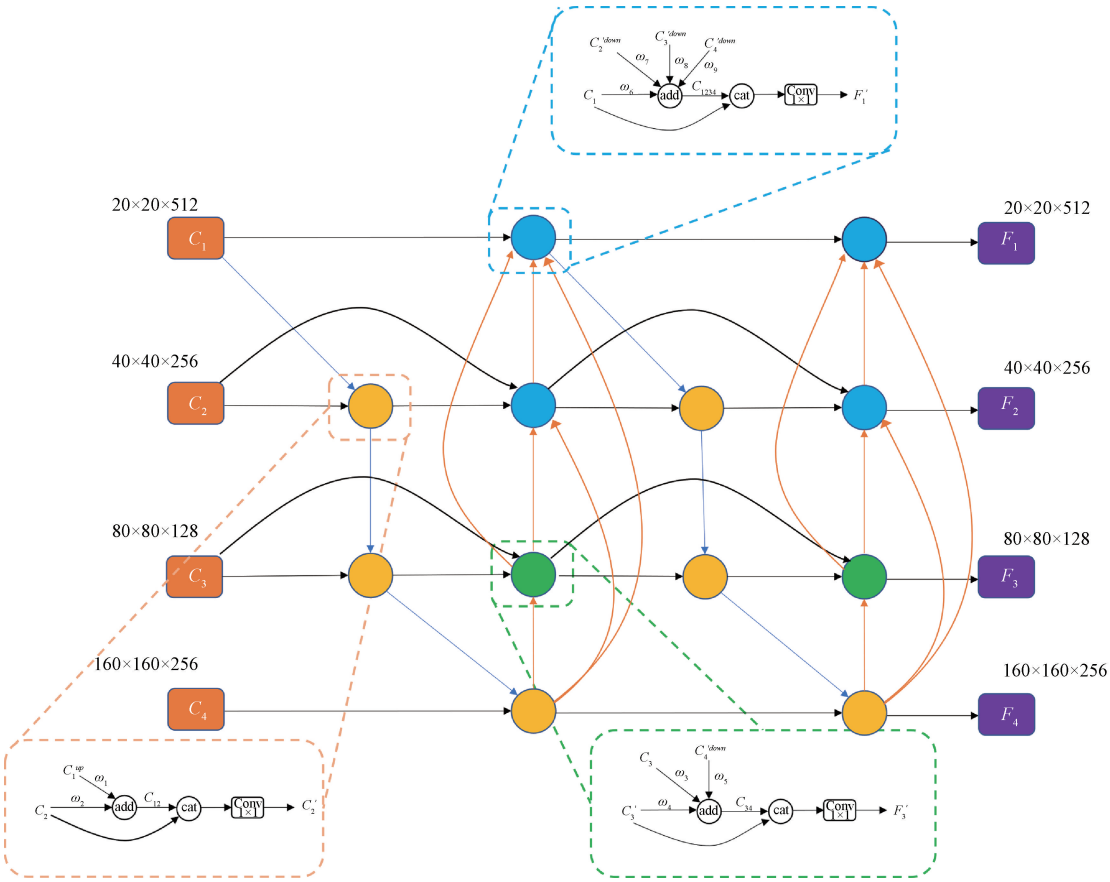

图 3　Multi-WHFPN 网络结构

Softmax 相似，同样将值域缩放到[0，1]之间，但它具有更快的训练速度和更高的效率，其公式如下[12]：

$$O=\frac{\sum_i \omega_i \cdot I_i}{\varepsilon+\sum_j \omega_j} \tag{1}$$

其中，O 是输出特征，I_i 是输入特征，ω_i 和 ω_j 都是输入特征所带的权重，ε 是一个很小的数字，防止训练时梯度变为 0，确保输出的稳定性。

图 3 中有 3 类节点，一类是双输入节点，一类是三输入节点，一类是三输入节点，分别以第 2 层的第 1 个节点、第 3 层第 2 个节点和第 1 层第 1 个节点为例。第 2 层第 1 个节点为双输入节点，输入分别是 C_2 和对 C_1 进行上采样的结果，特征融合公式为：

$$C_{12}=Conv\left(\frac{\omega_1 \cdot C_1^{up}+\omega_2 \cdot C_2}{\varepsilon+\omega_1+\omega_2}\right) \tag{2}$$

其中，C_1^{up} 是 C_1 进行上采样的结果，由于 C_1 比 C_2 尺寸小，所以要对 C_1 进行上采样操作。ω_1 是对 C_1 进行上采样结果的权重和 ω_2 是 C_2 的权重，Conv 是卷积操作。最终该节点的输出为：

$$C_2'=Conv(Concat(C_{12},C_2)) \tag{3}$$

为了在不增加太多开销的情况下融合更多的特征，此处增加一条额外的边将处于同一层的原始输入 C_2 输入到该节点，最后进行卷积操作得到该节点的输出 C_2'。

类似的可以得到第 3 层第 2 个三输入节点的特征融合结果为：

$$C_{34}=Conv\left(\frac{\omega_3 \cdot C_3+\omega_4 \cdot C_3'+\omega_5 \cdot C_4'^{down}}{\varepsilon+\omega_3+\omega_4+\omega_5}\right) \tag{4}$$

该节点的最终输出为：

$$F_3'=Conv(Concat(C_{34},C_3')) \tag{5}$$

第 1 层第 1 个四输入节点的特征融合结果为：

$$C_{1234}=Conv\left(\frac{\omega_6 \cdot C_1+\omega_7 \cdot C_2'^{down}+\omega_8 \cdot C_3'^{down}+\omega_9 \cdot C_4'^{down}}{\varepsilon+\omega_6+\omega_7+\omega_8+\omega_9}\right) \tag{6}$$

该节点的最终输出为：

$$F_1'=Conv(Concat(C_{1234},C_1)) \tag{7}$$

Multi-WHFPN 结构输出 $F_1=(20,20,512)$，$F_2=(40,40,256)$，$F_3=(80,80,128)$，$F_4=(160,160,256)$，最后再将输出结果经过 RepConv 操作后送入检测头进行检测。

2.2　SimAM 注意力机制

常见的一些注意力机制都需要增加额外的参数。SimAM 注意力机制[13]可以在不引入额外的参数的基础上在 3D 维度评估特征权重，增加 3D 注意力权重示意图如图 4 所示。通过引入 SimAM 注意力机制来增强报纸版面

分割中遇到的文本、标题、图片等目标的特征,通过最直接有效的方法对版面分割过程中检测目标的特征进行优化。

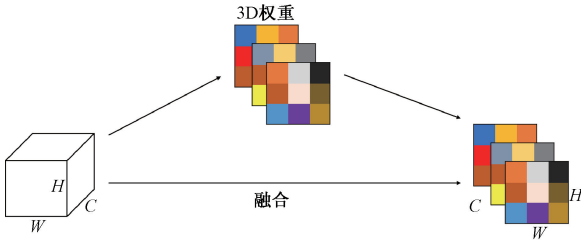

图 4　增加 3D 权重示意图

SimAM 注意力机制通过引入描述线性可分性的能量函数来体现每个神经元的重要程度,能量函数定义如下:

$$e_t(w_t,b_t,y,x_t)=\frac{1}{M-1}\sum_{i=1}^{M-1}(-1-(w_tx_i+b_t))^2+(1-(w_tt+b_t))^2+\lambda w_t^2 \tag{8}$$

其中,t 是目标神经元,x_i 是同一个通道中的其他神经元,$M=H\times W$,H 是输入图片的高度,W 是输入图片的宽度,w_t 是权重变量,b_t 是偏移变量,λ 为超参数。对上述函数的变量求偏导再代入原函数可以的到 e_t 最小值的解析解,如下:

$$e_t^*=\frac{4(\hat{\sigma}^2+\lambda)}{(t-\hat{\mu})^2+2\hat{\sigma}2+2\lambda} \tag{9}$$

其中,$\hat{\mu}=\frac{1}{M}\sum_{i=1}^{M}x_i$,$\hat{\sigma}^2=\frac{1}{M}\sum_{i=1}^{M}(x_i-\hat{\mu})^2$。上述能量函数 e_t^* 的值越小,表示目标神经元与其他神经元的线性可分性越强,该神经元也就越重要。最终 SimAM 模块优化为:

$$\tilde{X}=sigmoid(\frac{1}{E})\odot X \tag{10}$$

上式通过 $1/e_t^*$ 求得神经元的重要程度,并对其进行加权处理。添加完 SimAM 注意力机制的 YOLOv7 特征金字塔部分网络结构如图 5 所示。

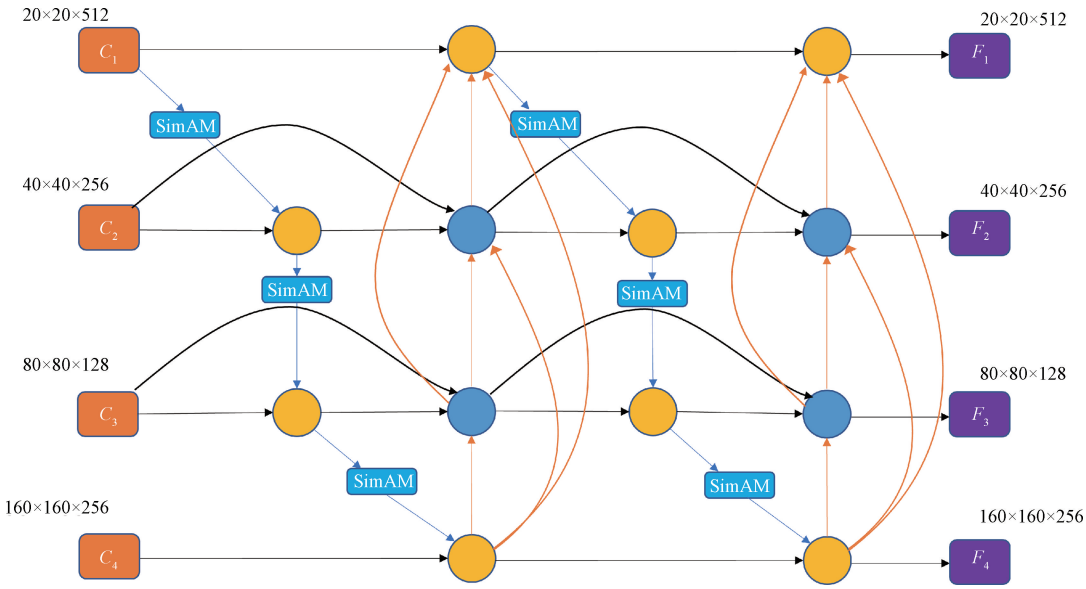

图 5　引入 SimAM 注意力机制的 Multi-WHFPN 网络

2.3　YEIOU 损失函数

IOU 损失函数考虑到了预测框与真实框的重叠占比,但忽略了两框不重合时预测框的偏移距离。随后提出的 GIOU[14] 解决了这一问题,不过当预测框在真实框内部时,预测框与真实框的差集为 0,GIOU 退化为 IOU,无法判断预测框的好坏。此后提出的 DIOU[15] 将预测框与真实框的中心点距离纳入考虑,解决了这一问题。但上述提到的几种回归损失函数都没有将预测框的长宽纳入考虑,本文基于 EIOU[16] 损失函数,设计了新的 YEIOU 损失函数作为新的预测框损失函数。

EIOU 损失函数综合考虑了预测框与真实框重叠面积、中心点的距离、长宽这 3 个重要因素,其由三部分构成,IOU 损失 L_{IOU},中心距离损失 L_{dis},高宽损失 L_{asp},其计算公式如下:

$$L_{EIOU}=L_{IOU}+L_{dis}+L_{asp}=1-IOU+\frac{\rho^2(b,b^{gt})}{(w^c)^2+(h^c)^2}+\frac{\rho^2(w,w^{gt})}{(w^c)^2}+\frac{\rho^2(h,h^{gt})}{(h^c)^2}=1-EIOU \tag{11}$$

其中,w^c 和 h^c 是能将真实框和预测框覆盖的最小矩形的宽和高,$\rho(b,b^{gt})$ 表示真实框和预测框中心点之间的欧式距离,$\rho(w,w^{gt})$ 表示真实框和预测框宽的差,$\rho(h,h^{gt})$ 表示真实框和预测框高的差 w、w^{gt}、h、h^{gt} 分别表示预测框与真实框的宽和高。

EIOU 损失函数将预测框的长宽纳入考虑,解决了 DIOU 损失函数预测框与真实框相似度较差的问题,并且将长与宽分开计算,提升了模型的收敛速度与回归精度。

原始的 EIOU 函数收敛较慢,为了进一步提升预测框损失函数的收敛速度,设计 YEIOU 如式(12)所示。

$$L_{YEIOU}=\frac{1}{\ln\pi}\times(e^{(1-EIOU)}-1) \tag{12}$$

原始的 L_{EIOU} 与 L_{YEIOU} 都是对于 EIOU 的减函数，且下限均为 0。相比之下，L_{YEIOU} 曲线的梯度比 L_{EIOU} 更大，因此收敛速度更快。此外，L_{YEIOU} 曲线的梯度随着 EIOU 值的增大而逐渐降低。当 EIOU 值趋向于 1 时，L_{YEIOU} 曲线越趋向于平缓，这表明 L_{YEIOU} 具有良好的自适应调节性能。L_{EIOU} 与 L_{YEIOU} 曲线对比图如图 6 所示。

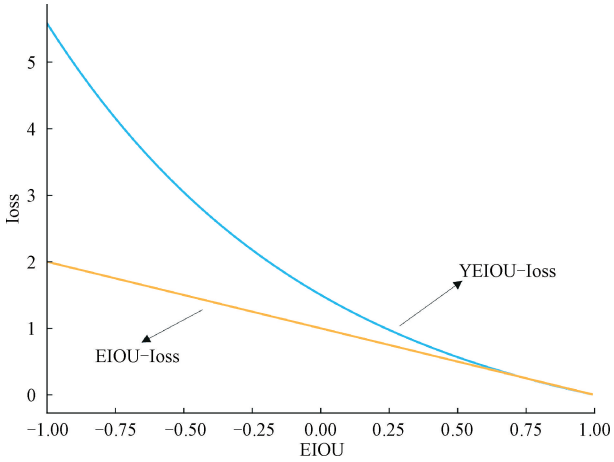

图 6　L_{EIOU} 与 L_{YEIOU} 曲线对比图

3　实验结果及分析

3.1　实验环境与参数设置

出于对实验公平性的考虑，本文采用相同的实验环境，实验的软硬件环境如表 1 所示。

表 1　软硬件环境

名称	配置
处理器	Intel(R) Core(TM) i9-10900K CPU@3.70 GHz
内存	64 G
显卡	两张 RTX 3070 Ti 8 G
深度学习框架	pytorch 1.10.2
操作系统	windows10

训练过程中采用 Adam 优化算法对模型进行优化，更多的实验训练相关参数设置如表 2 所示。

表 2　实验参数设置

训练设置	值
迭代次数(Epoch)	100
优化器(Optimizer)	Adam
输入尺寸(Input size)	640×640×3
批量大小(Batch size)	12
学习率(Learning rate)	0.001
学习率调整方案	线性学习率

3.2　实验数据集

实验数据集是江苏省档案馆提供的新锡山报纸扫描件。该数据集包括了新锡山从 2010 年～2019 年所有刊登的报纸。本文选取了其中 3 212 张报纸，报纸图片尺寸为 4 480×6 340 像素，为满足实验需要按照 7∶2∶1 的比例划分训练集、验证集和测试集。最终目标要将这些报纸划分为抬头(head)、标题(title)、正文(text)、图片(picture)四种区域。数据集图例如图 7 所示。

图 7　数据集图例

3.3　评价指标

选取目标检测中几个比较常见的指标作为评价指标：准确率(P)、召回率(R)、平均准确度(AP)、度量分数(F1)、检测速度(FPS)[17]，具体公式如下：

$$P=\frac{T_p}{T_P+F_p}\times 100\% \tag{13}$$

$$R=\frac{T_P}{T_P+F_N}\times 100\% \tag{14}$$

$$AP=\int_0^1 P(r)\mathrm{d}r \tag{15}$$

$$F_1=\frac{2P\times R}{P+R} \tag{16}$$

其中，T_P 为被正确识别的正样本的数量，F_p 为检测错误的负样本数量，F_N 为遗漏的正样本的数量。在准确率(P)和召回率(R)围成的曲线中，曲线与坐标轴围成的面积为 AP，所有类别 AP 的平均值为 mAP，本文选取 L_{IOU} 阈值为 0.5 的 mAP，即 mAP@.5 作为评价指标。

3.4　对比试验

为了验证改进算法的有效性，选择了 EAST、TextBoxes＋＋、Faster-RCNN、SSD、YOLOv5 与 YOLOv7-MSY 在相同的运行环境下进行对比。实验结果如表 3 所示。

实验结果表明，YOLOv7-MSY 除 FPS 以外的各项指标均优于其他算法。其准确率比表现最好的 TextBoxes＋＋

表 3　实验结果

模型	准确率(P)	召回率(R)	mAP@.5	F_1	检测速度/fps
EAST	0.833	0.783	—	0.807	16.8
TextBoxes++	0.861	0.742	—	0.804	11.6
Faster-RCNN	0.719	0.503	0.698	0.592	14
SSD	0.571	0.447	0.351	0.501	**91.3**
YOLOv5	0.705	0.811	0.805	0.754	64.9
YOLOv7	0.849	0.809	0.816	0.829	59.5
YOLOv7-MSY	**0.871**	**0.856**	**0.88**	**0.863**	53.8

高 1.15%，召回率比表现最好的 YOLOv5 高 5.26%，mAP@.5 比表现最好的 YOLOv7 高 7.27%，F1 比表现最好的 YOLOv7 高 3.94%，并且检测速度也处在较高水平。

准确率排名前三的非 YOLO 系列算法 EAST、TextBoxes++、Faster-RCNN 与 YOLOv7-MSY 版面分割效果对比图表 4 所示。为了提升分割效果的清晰度，我们可以利用算法得出的坐标信息来对相应区域进行颜色填充，抬头用蓝色填充、标题用橙色填充、文本用绿色填充、图片用紫色填充：

以 a 组为例，从左侧的方框中可以看出，YOLOv7-MSY 在处理纵向且间距较大的标题时表现优越，其他的版面分割算法则出现了漏检错检的情况。

表 4　版面分割效果对比

组别	YOLOv7-MSY	EAST	TextBoxes++	Faster-RCNN
a				
b				
c				

以 b 组为例，将其放大如表 5 所示。

从表 4 可以看出，Faster-RCNN 的表现比较差。与 EAST 算法相比 YOLOv7-MSY 对目标边界的识别更加准确，并且 YOLOv7-MSY 比 TextBoxes++更擅长识别小目标，可以识别出上表报纸中所有的图片下方名称，而 TextBoxes++则有一些漏识的情况。此外，YOLOv7-MSY 对报纸版面图片中的文字识别更加敏感，可以识别出上表报纸最下方图片的左上方文字，已用箭头指出，而其他 3 种算法都无法识别到。

为验证模型的泛化性能，选取人民日报、参考消息、环球时报 3 种不同类型的报纸来进行对比试验，实验结果如表 6 所示。

表 5　b 组放大图

YOLOv7-MSY	EAST
	
TextBoxes＋＋	Faster-RCNN
	

表 6　不同数据集对比试验

	人民日报			参考消息			环球时报		
	P	R	F_1	P	R	F_1	P	R	F_1
EAST	0.812	0.762	0.786	0.837	0.751	0.791	0.816	0.779	0.797
TextBoxes＋＋	0.866	0.733	0.794	0.856	0.761	0.806	0.871	0.727	0.793
Faster-RCNN	0.682	0.493	0.572	0.703	0.551	0.618	0.721	0.526	0.608
SSD	0.612	0.399	0.483	0.594	0.492	0.538	0.576	0.473	0.519
YOLOv5	0.729	0.794	0.760	0.736	0.826	0.778	0.715	0.833	0.770
YOLOv7	0.854	0.814	0.834	0.846	0.822	0.834	0.837	0.792	0.814
YOLOv7-MSY	**0.882**	**0.859**	**0.870**	**0.864**	**0.861**	**0.863**	**0.856**	**0.848**	**0.852**

可以看出 YOLOv7-MSY 模型在不同类型报纸上的准确率 P,召回率 R 以及度量分数 F1 均优于其他同类算法,YOLOv7-MSY 模型具有良好的泛化性能。

3.5 消融实验

对算法的关键模块进行了消融实验,验证了算法增加模块的有效性。包括 Multi-WHFPN、SimAM 和 YEIOU 模块,并记录了每个模块被消融后的实验结果。实验结果如表 7 所示。

实验结果表明,在逐步优化的过程中,准确率与检测速度始终维持在一个较高水平。在将原始的 PANet 网络替换为 Multi-WHFPN 后,模型准确率提升较高,但召回率有所下降。再在 Multi-WHFPN 中嵌入 SimAM 注意力机制之后,召回率与 mAP@.5 有了明显提高。最后将原有损失函数替换为 YEIOU 损失函数之后,模型的各项检测性能得到了进一步提高,并且检测速度相较于前一步提升了 13.5%。最终 YOLOv7-MSY 模型相较于原始 YOLOv7 模型检错率降低了 14.6%,召回率提高了 5.8%,mAP@.5 提升了 7.84%,F1 提高了 4.1%,检测速率与原模型处于相近水平。

表 7 消融实验结果

模型	准确率(P)	召回率(R)	mAP@.5	F_1	检测速度(FPS)
YOLOv7	0.849	0.809	0.816	0.829	**59.5**
+Multi-WHFPN	0.861	0.787	0.808	0.822	52.6
+Multi-WHFPN+SimAM	0.864	0.813	0.841	0.826	47.4
YOLOv7-MSY	**0.871**	**0.856**	**0.88**	**0.863**	53.8

YOLOv7-MSY 模型与原始 YOLOv7 模型在迭代过程中的 loss 曲线对比图如图 8 所示。

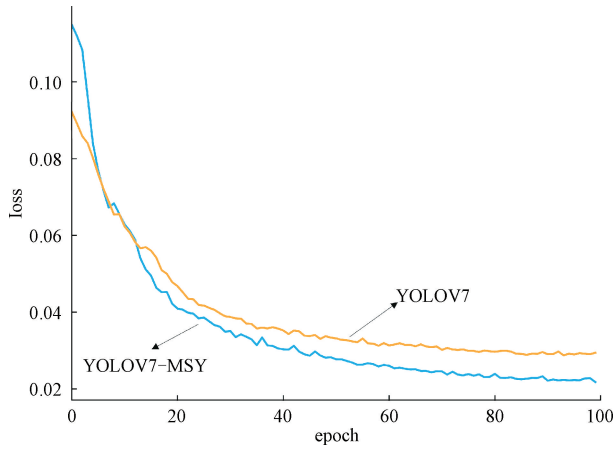

图 8 算法改进前后 loss 曲线对比图

从图中可以看出,改进后的 loss 曲线比改进前的 loss 曲线下降速度更快,并且最终的收敛值更低。这是由于 YEIOU 重新定义了预测框损失函数,新的损失函数直接预测目标框的长和宽这两个度量,并且优化了下降的梯度,从而有效加快了收敛速度。

报纸经过 YOLOv7-MSY 模型分割得到的结果与原模型结果对比如表 8 所示。

表 8 消融实验效果对比

YOLOv7-MSY	YOLOv7
	

从表 8 可以看出，在①处与③处 YOLOv7-MSY 模型比 YOLOv7 模型对目标边界的识别更加准确，这是由于由于使用了更深层的网络结构与卷积层，从而提高了模型对目标边界的感知能力，对目标区域边界检测更加敏感；从②处与④处可以看出 YOLOv7-MSY 模型比 YOLOv7 模型对小目标的检测能力更强，这是因为在模型中增加了小目标检测头，从而提高了对细节的感知能力；在⑤处 YOLOv7 遗漏了该目标，这是由于引入了 SimAM 注意力机制，增强了有效特征，从而降低了特征丢失。

4　结　　论

针对报纸版面分割，提出了 YOLOv7-MSY 模型。首先使用 Multi-WHFPN 网络结构代替原有的 PANet 网络，充分提取图片的特征，并且增加了小目标检测头，加强了小目标检测能力；引入无参 SimAM 注意力机制，在不增加额外的参数的基础上增强有效特征，减少特征丢失；将 YOLOv7 中原有的定位损失函数替换为 YEIOU 损失函数以提高预测框与真实框的相似度，并且提升了模型的收敛速度与回归精度。最后在数据集上进行测试，实验结果表明，YOLOv7-MSY 模型版面分割的效果优于原 YOLOv7 网络和其它经典版面分割网络，并且拥有较快速度。在不同数据集上进行验证，该算法均表现优异，具有良好的泛化性能，该算法有较好的版面分割应用前景。

参考文献

[1] 应自炉，赵毅鸿，宣晨，等. 多特征融合的文档图像版面分析[J]. 中国图象图形学报，2020，25(2)：311-320.

[2] TIAN Z，HUANG W，HE T，et al. Detecting text in natural image with connectionist text proposal network[C]. 14th European Conference on Computer Vision(ECCV)，2016：56-72.

[3] 龚承志. 中文合同文档版面分析关键技术研究[D]. 重庆：西南大学，2022.

[4] LIAO M，SHI B，BAI X. TextBoxes plus plus：A single-shot oriented scene text detector[J]. Ieee Transactions on Image Processing，2018，27(8)：3676-3690.

[5] ZHOU X Y，YAO C，WEN H，et al. EAST：An efficient and accurate scene text detector[C]. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)，2017：2642-2651.

[6] LIU X，LIANG D，YAN S，et al. FOTS：Fast oriented text spotting with a unified network[C]. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)，2018：5676-5685.

[7] WANG C Y，BOCHKOVSKIY A，LIAO H，et al. YOLOv7：Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)，2023：7464-7475.

[8] 朱张莉，饶元，吴渊，等. 注意力机制在深度学习中的研究进展[J]. 中文信息学报，2019，33(6)：1-11.

[9] 董刚，谢维成，黄小龙，等. 深度学习小目标检测算法综述[J]. 计算机工程与应用，2023，59(11)：16-27.

[10] 周葳楠，吴治海，张正道，等. 基于弱特征增强的轻量化小目标检测方法研究[J]. 控制与决策，2023，DOI：10.13195/j.kzyjc.2022.1432.

[11] CHEN J，MAI H，LUO L，et al. Effective feature fusion network in BIFPN for small object detection[C]. IEEE International Conference on Image Processing(ICIP)：IEEE，2021：699-703.

[12] 亢洁，王勃，刘文波，等. 融合 CAT-BiFPN 与注意力机制的航拍绝缘子多缺陷检测网络[J]. 高电压技术，2023，49(8)：3361-3376.

[13] YANG L，ZHANG R Y，LI L，et al. SimAM：A simple，parameter-free attention module for convolutional neural networks[C]. International Conference on Machine Learning(ICML)，2021：139.

[14] HOU Z，LIU X，CHEN L. Object detection algorithm for improving non-maximum suppression using GIoU[J]. IOP Conference Series Materials Science and Engineering，2020，790(1)：12062.

[15] ZHENG Z，WANG P，LIU W，et al. Distance-IoU Loss：faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2020，34(7)：12993-13000.

[16] ZHANG Y F，REN W Q，ZHANG Z，et al. Focal and efficient IOU loss for accurate bounding box regression[J]. Neurocomputing，2022，506：146-157.

[17] 邵延华，张铎，楚红雨，等. 基于深度学习的 YOLO 目标检测综述[J]. 电子与信息学报，2022，44(10)：3697-3708.

作者简介

杨陈慧，硕士研究生，从事机器学习、深度学习方面的研究。

E-mail：ych525@126.com

业宁(通信作者)，博士，博士生导师，从事机器学习、深度学习方面的研究。

E-mail：yening@njfu.edu.cn