计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(7):52−62 [doi: 10.15888/j.cnki.csa.009523]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进 YOLOv7 的视频监控小目标检测①

夏　翔, 朱　明

(中国科学技术大学 信息科学技术学院, 合肥 230026)
通信作者: 夏　翔, E-mail: xiaxiang1@mail.ustc.edu.cn

摘　要: 小目标检测作为目标检测中一项极具挑战性的项目, 广泛分布于日常生活中, 在视频监控场景中, 距离摄像头约 20 m 远处的行人人脸就可以被认为是小目标. 由于人脸可能相互遮挡并容易受到噪声和天气光照条件的影响, 现有的目标检测模型在这类小目标上的性能劣于中大型目标. 针对此类问题, 本文提出了改进后的 YOLOv7 模型, 添加了高分辨率检测头, 并基于 GhostNetV2 对骨干网络进行了改造; 同时基于 BiFPN 和 SA 注意力模块替换 PANet 结构, 增强多尺度特征融合能力; 结合 Wasserstein 距离改进了原来的 *CIoU* 损失函数, 降低了小目标对锚框位置偏移的敏感性. 本文在公开数据集 VisDrone2019 以及自制的视频监控数据集上进行了对比实验. 实验表明, 本文提出的改进方法 *mAP* 指标在 VisDrone2019 数据集上提高到了 50.1%, 在自制视频监控数据集上高于现有方法 1.6 个百分点, 有效提高了小目标检测的能力, 并在 GTX1080Ti 上达到了较好的实时性.

关键词: 小目标检测; 注意力机制; 特征融合; 损失函数

引用格式: 夏翔,朱明.改进 YOLOv7 的视频监控小目标检测.计算机系统应用,2024,33(7):52–62. http://www.c-s-a.org.cn/1003-3254/9523.html

Small Target Detection in Video Surveillance Based on Improved YOLOv7

XIA Xiang, ZHU Ming

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: As a very challenging project in target detection, small target detection is widely distributed in daily life. In video surveillance scenarios, pedestrians' faces about 20 meters away from the camera can be considered small targets. Due to the possibility of mutual occlusion of faces and their susceptibility to noise and weather, lighting conditions, the performance of existing target detection models on such small targets is inferior to that on medium and large targets. To address these issues, this study proposes an improved YOLOv7 model with a high-resolution detection head and transforms the backbone network based on GhostNetV2. At the same time, the PANet structure is replaced by the BiFPN and SA attention modules combined to enhance the multi-scale feature fusion capability; the original *CIoU* loss function is improved by combining the Wasserstein distance, reducing the sensitivity of small targets to anchor frame position offset. This study conducts comparative experiments on the public dataset VisDrone2019 and a self-made video surveillance dataset. Results show that the *mAP* of the improved method proposed in this study improved to 50.1% on the VisDrone2019 dataset and is 1.6 percentage points higher than existing methods on the self-made video surveillance dataset, which effectively improves the ability of small target detection and achieves good real-time performance on the GTX1080Ti.

Key words: small target detection; attention mechanism; feature fusion; loss function

① 基金项目: 科技创新特区计划 (20-163-14-LZ-001-004-01)
收稿时间: 2023-12-14; 修改时间: 2024-01-17; 采用时间: 2024-01-29; csa 在线出版时间: 2024-05-31
CNKI 网络首发时间: 2024-06-04

1 引言

随着人工智能和深度学习的快速发展, 视频监控逐渐走向了智能化, 人们可以通过目标检测等技术对视频中出现的人员及行为进行识别, 理解和分析, 并以可视化结果呈现出来, 并对设置的报警条件及时作出反应, 实现监控场景的智能化. 但市场上已有的监控摄像头一般都架设在 4–5 m 的空中, 监控范围最远能够达到 20–30 m, 在这种情况下若有人员出现在摄像头边缘处, 其在监控视频中占据的像素点较小, 人脸所占的面积更小, 这种较小目标的检测对于视频监控的目标检测算法提出了挑战. 而自然环境中的小目标受很多不确定因素影响, 比如彼此重叠, 或是容易受到天气, 光照产生的噪点影响等. 小目标尺寸小, 特征不明显, 在检测中误检率和漏检率一般均较高, 而视频监控的智能化离不开对人脸的识别. 因此提升小目标的检测性能, 增强检测器在视频监控环境中的鲁棒性有着重要的现实意义.

目标检测中的深度学习技术通常可以分为两阶段目标检测和单阶段目标检测方法[1]. 两阶段目标检测方法检测精度更高, 但由于需要进行两阶段的处理, 检测速度较慢, 代表性的方法有源于 RCNN 的一系列模型如 Cascade R-CNN[2]等; 单阶段目标检测则同时进行分类和检测, 直接生成物体位置和类别置信度, 提高了检测速度, 但精度有所下降, 代表性的方法有源于 YOLO 的一系列算法如 YOLOv3[3]等. 基于此, 综合考虑视频监控场景下的硬件性能和实时性要求, 本文使用 YOLO 模型作为基础.

在 YOLOv3 模型之后, 出现了各种改进后的模型, 如 YOLOv5 模型, 并应用在不同的领域上. 如 Song 等人[4]将 YOLOv5 应用于道路密集车辆实时检测, 张凯祥等人[5]在 YOLOv5 的基础上添加了其他任务的检测头, 将目标检测, 车道线识别和语义分割集中到一个网络上, 搭建了多任务的自动驾驶环境感知网络. 相较于之前的 YOLO, YOLOv5 将 CBL 模块替换成 CSP 结构, 加强了网络融合特征的能力. 在后续美团提出的 YOLOv6[6]引入 RepVGG 风格的 RegBlock 替换了 YOLOv5 中的 CSPBlock, 同时调整了网络中的算子, 使网络更加硬件友好.

YOLO 模型在迭代过程中性能越来越好, 但它们在小目标上的检测性能提升并没有中大目标上提升的明显, 这是由以下几个因素决定的: 1) 特征信息少. 小目标在图像中占据的像素较少, 特征不足; 现有的目标检测器通过不断下采样进行深度特征提取, 一个占据区域为 8×8 个像素点的目标在通过步长为 2 的卷积层后会被压缩为 4×4 个像素点, 此过程中小目标的特征信息丢失严重[7]; 2) 由于小目标占据的区域较小, 在相同的 *IoU* 下, 背景样本要远多于小目标样本数; 同时在大多数数据集中, 大中目标的样本数也是大于小目标样本数的, 这就造成网络更倾向于学习占比更大的中大目标; 3) 摄像机在现实生活中容易受光照等环境影响, 产生的噪声很难和小目标区分开, 同时小目标锚框的轻微偏移也会导致较大的 *IoU* 变化, 导致小目标容易受到干扰, 增加检测难度.

针对以上小目标检测面临的挑战, 目前主流的解决方法主要有以下几个方面: 1) 增强骨干网络提取特征的能力. 而随着注意力机制在计算机视觉得到应用, 注意力机制也进入了目标检测研究者的视野[8], 比如 Liu 等人仿照 ViT 的结构提出了 ConvNeXt[9]. 2) 多尺度表示特征. 在卷积神经网络中, 深层特征分辨率低, 拥有更大的感受野和更强的语义信息; 浅层信息分辨率高, 具有丰富的纹理信息和准确的空间信息. 融合不同层之间的信息, 能够提高小目标检测精度. 3) 改进锚框回归损失函数. *IoU* 是目标预测框和真实框交集和并集的比值, 是目标检测常用的指标. Rezzatofighi 等人[10]提出 GIoU 解决了非交叉框无法回归的难题; 为了提高回归的收敛速度, Zheng 等人[11]引入两框的距离和尺度, 提出了 DIoU, 并提出了回归定位的 3 个重点因素: 重叠面积, 中心点距离和长宽比; *CIoU* 作为 DIoU 的改进版本, 它将两框的长宽比代入计算, 解决了当多个候选框中心点重合时难以找出最优预测框的问题. 4) 添加增益模块. Hu 等人提出的 SE 模块[12]挖掘通道特征之间的相互依赖关系, 自适应校正各个通道的权重参数, 让网络聚焦于更关键的特征信息. Woo 等人在空间维度上进行扩展, 设计了 CBAM 模块[13], 它由通道注意力模块 CAM 和空间注意力模块 SAM 串联而成, 提高了网络对特征信息的提取能力和强化能力. TPH-YOLOv5[14]将此模块和高分辨率分支用在了无人机小目标检测任务中, 取得了当时最好的成绩. 此外, 在原有模型的基础上增加超分辨模块也是常用的方法. Deng 等人提出的 EFPN[15]借鉴了超分辨率的思想, 将 FPN 中较高层的特征图通过超分辨率方法重建, 并与底层特征融合, 在小目标检测上达到了较好的精度.

因此, 针对小目标检测精度不高, 易受噪声影响的问题, 本文提出了一种改进后的 YOLOv7 模型, 主要贡献如下: 1) 增加了高分辨率检测分支, 增强模型对小目标检测能力. 2) 基于 ECA[16]注意力机制和 GhostNetV2[17]设计了轻量化模块, 在减少模型参数量的同时保留了提取特征的能力. 3) 在特征融合网络中结合了 BiFPN[18]和 ShuffleAttention (SA)[19]注意力机制, 融合了更多的特征. 4) 结合 Wasserstein 距离和 *IoU* 函数设计了新的损失函数, 提高了小目标检测的精度.

2 相关工作

YOLOv7 由 YOLOv4 的团队于 2022 年 7 月提出[20], 在一阶段目标检测器中性能达到了 SOTA, 且有着不同精度的模型, 所以选择它作为基础模型. 根据模型运行环境的不同, 共有 YOLOv7-Tiny, YOLOv7, YOLOv7-W6 这 3 个模型. 本文结合实际的应用场景, 选择 YOLOv7 作为基础模型.

YOLOv7 整体可以分为 3 个模块: 输入端、Backbone 网络和 Head 输出端. 其整体结构如图 1 所示.

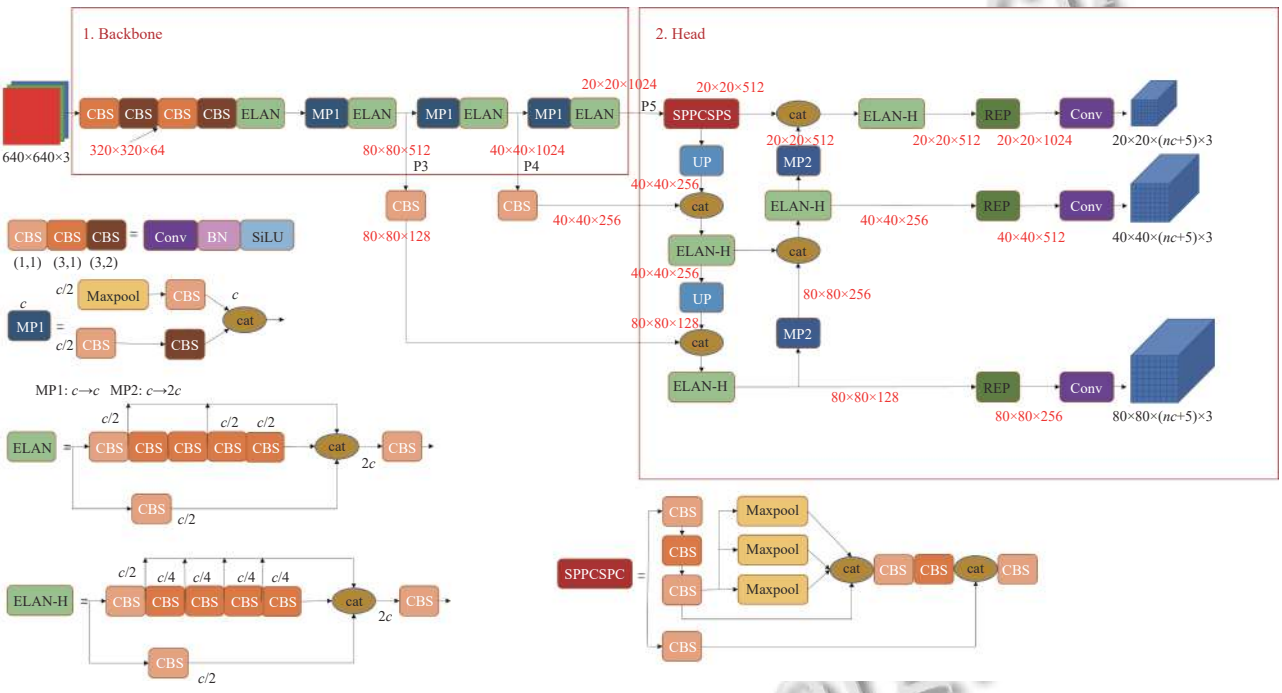


图 1 YOLOv7 结构图

输入端主要完成图像预处理工作, 即缩放图片为网络输入大小、数据归一化以及数据增强工作等, YOLOv7 用到了 Mosaic, Mixup 等数据增强方式, 主要是为了打乱图片排布, 提供更多的训练样本. Backbone 网络负责提取通用的特征表示, 其中 CBS 是使用了 BatchNorm 和 SiLU 的卷积层, 其相对 ReLU 函数在接近 0 的时候有更平滑的曲线; ELAN 及其变种 ELAN-H 则是 YOLOv7 的特有结构, 其吸取了 VoVNet[21], CSPNet[22]和 ELAN 的思想, 通过减少 DenseNet[23]的连接个数和 split 操作增加梯度信息; 并考虑到最短梯度路径的存在, 进一步使用 expand, shuffle, merge cardinality 操作在提升网络效率的同时, 减少了计算量, 使得网络可以堆叠更多 block; MP 结构则是池化层和卷积层的结合, 主要作用是实现下采样. 骨干网络设置了 3 个重复的 MP+ELAN 模块, 它们的输出将对应后面的 3 个常规检测头. 在检测头部分, 首先是 SPPCSPC 模块, 其中 SPP 的作用是能够增大感受野, 并且在模块的第 1 个分支中有 4 个不同的最大池参数, 代表着它能够处理不同尺寸的目标; CSP 模块则将特征分为两部分, 其中一部分进行常规处理, 另一部分则是 SPP 结构处理, 最后合并两个部分, 减少了一般的计算量, 在速度变快的同时还能提升精度. 接着将 FPN[24]和 PAN[25]结合起来, FPN 自顶向下把深层特征语义信息传递到浅层, PAN 自底向上将浅层的定位信息传递给深层, 在多尺度上进行预测. Head 部分大体结构上沿用了 YOLOv5 检测头, 分类和置信损失采用 BCEloss, 定位损失采用 CIoUloss, 但 YOLOv7 将模型重参数化引入到网络架构中, 在预测头上表现为 REP 模块, REP 总共有 3 个分支, 每个分

支包括一个 (或不包括) CBS 模块和 BN 模块, 训练时由 3 个分支相加输出, 而在部署时会将分支的 CBS 模块中的参数等价转换, 最后部署模型上只有一个分支, 这样既利用了多分支模型训练时性能高的优点, 在模型部署时又有单分支模型速度快, 省内存的优点.

3 算法设计和实现

本文算法在结构上分为两部分: 骨干网络, 检测头部分. 数据在训练过程中, 首先是预处理部分, 对图片进行自适应缩放以及自适应锚框计算. 骨干网络包括改进的 GhostV2, CBS, ELAN 以及 MP 模块, 骨干网络使用改进的 GhostNetV2 模块替换掉部分 Conv 模块, 在提升速度的同时帮助骨干网络提取感兴趣区域 提高骨干网络对小目标的特征提取能力; 同时在骨干网络的 P2 层额外引出了一条高分辨率的目标检测分支, 参与最后的检测任务. 在检测头部份, 以双向交叉和加权融合的方式进行特征融合, 并把特征传递到后面的检测头, 检测头输出采用改进的小目标鲁棒性更强的函数作为损失函数, 并提供 4 种检测尺度 (20×20, 40×40, 80×80, 160×160) 进行检测.

3.1 增加小目标检测头

在原始 YOLOv7 模型中, 骨干网络共有 3 次下采样, 得到 4 层特征表达 (P2, P3, P4 和 P5), 其中 Pi 表示分辨率为原始图片的$1/2^i$, 这 3 个特征图在经过特征融合网络后实现多尺度特征融合, 最后检测头在这 3 级特征图上引出的特征头上进行目标检测. 下面简称通过 Pi 层特征图引出的特征头为 Pi 层检测头. 然而, 小目标的尺寸往往处在 20 个像素左右, 在极端条件下 (如 20 m 远处的人脸), 甚至存在小于 10 像素的目标. 类似的目标在经过多次下采样后, 其在特征图中仅占 1–2 个像素, 其大部分信息已经丢失, 通过具有较高分辨率的 P3 层检测头依然难以检测.

为了在上述的小目标上同样达到较好的检测效果, 本文在 YOLOv7 模型上通过 P2 层特征引出了新的小目标检测头. P2 层检测头分辨率为 160×160 像素, 在主干网络中只进行了两次下采样, 具有更为丰富的底层特征信息. 在特征融合网络中, P2 层特征与 PAN 模块中上采样的同尺度特征通过 concat 形式融合, 最后输出高分辨率的检测结果, 使 P2 检测头能够快速地检测小目标. P2 层检测头加上其他 3 个原有检测头, 能够有效地缓解目标尺度方差带来的影响, 增加的 P2 层检测头是通过高分辨率, 低层的特征图生成的, 这个检测头增加了模型的计算量和内存开销, 但是对小目标检测能力有着不小的提升.

3.2 增加基于 GhostNetV2 的轻量化结构

在视频监控实际场景中, 一台服务器往往要同时负责多台摄像头视频的同时监控, 而在设备限制的情况下, 模型通常无法运行或只能在很低的 batchsize 下运行. 在同时监控多台设备时, 如果使用传统卷积网络, 由于模型添加了高分辨率分支, 大大增加了模型的参数和计算量, 对硬件资源的要求更高, 势必要减少模型的计算量和参数量.

GhostNet[26]是硬件友好的注意力机制模块, 其简单有效, 方便在移动设备上即插即用, Ghost module 从特征图冗余问题出发, 利用特征图的相似性, 通过少量计算产生大量特征图, 尽管 Ghost 模块可以大幅度的减少计算代价, 但是其特征的表征能力也因为卷积操作只能建模一个窗口内的局部信息而被削弱了. 为了提高 Ghost 模块捕捉空间信息的能力, 研究者们在其基础上提出了 GhostNetV2, 它提出了一种新的注意力机制 (DFC attention) 来捕获长距离的空间信息, 同时保持了计算效率. 在论文中, GhostNetV2 和其他注意力机制模块在不同数据集上比较了性能, 证明了 GhostNetV2 的有效性.

Ghost module 被设计为一种分阶段的卷积计算模块, 在少量的普通卷积得到的特征图基础上, 再通过简单的线性变换生成剩下一部分特征图, 最后将两部分特征图拼接起来得到最终的特征图. Ghost module 结构如图 2 所示. GhostBottleneck 由两个 Ghost module 组成. 第 1 个 Ghost module 作为扩展层增加通道的数量, 第 2 个 Ghost module 减少输出特征图通道的数量使其与输入通道数相配. 值得注意的是, 如果想要输出尺寸减半的特征图, 需要额外加入一层 stride 为 2 的 depth-wise convolution (分组卷积). 在 GhostNetV1 的基础上, 考虑到通过廉价操作生成特征图的操作 (1×1 卷积) 会导致这些特征图没有与空间其他像素的交互, 模块捕获空间信息的能力较弱, 华为诺亚实验室提出了硬件友好的注意力机制 DFC, GhostNetV2 及 DFC attention 的结构如图 3(a) 所示, 其是基于全连接层构建的, 不仅可以在普通硬件上快速执行, 还可以捕获远程像素之间的依赖关系. 在 self-attention 中, 计算量通常是特征图大小的二次方关系, 为了减少这部分计算

量, 将特征图中某点的注意力计算方式改为由该点所在行和列的像素参与计算得到. DFC attention 生成了具有全局感受野的注意力图, 具有捕捉长程空间信息的能力, 类似于空间注意力机制. 然而, Ghost 模块使用的分组卷积消除了通道间的相关性, 使得当前通道特征仅与自己相关, 使得模型对全局特征的提取减少. 为了改善这种情况, 本文在 GhostNetV2 Bottleneck 中添加了通道注意力模块 ECA, 其在 SE 模块的基础上避免了降维, 有效捕获了跨通道交互的信息, 是一种高效的通道注意力模块. 改进后的 EGhostBottleneck 如图 3(b) 所示.

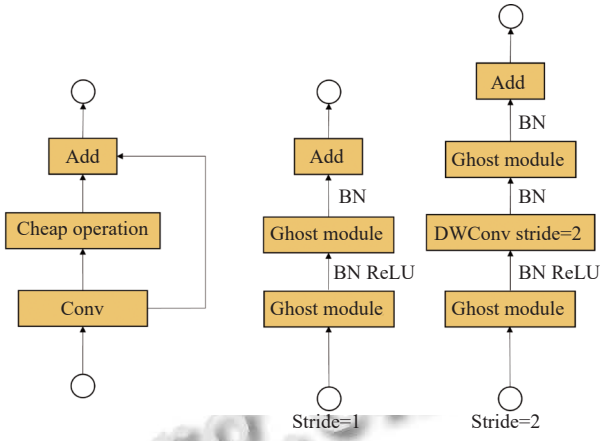


图 2 Ghost module 及 GhostV1 Bottleneck

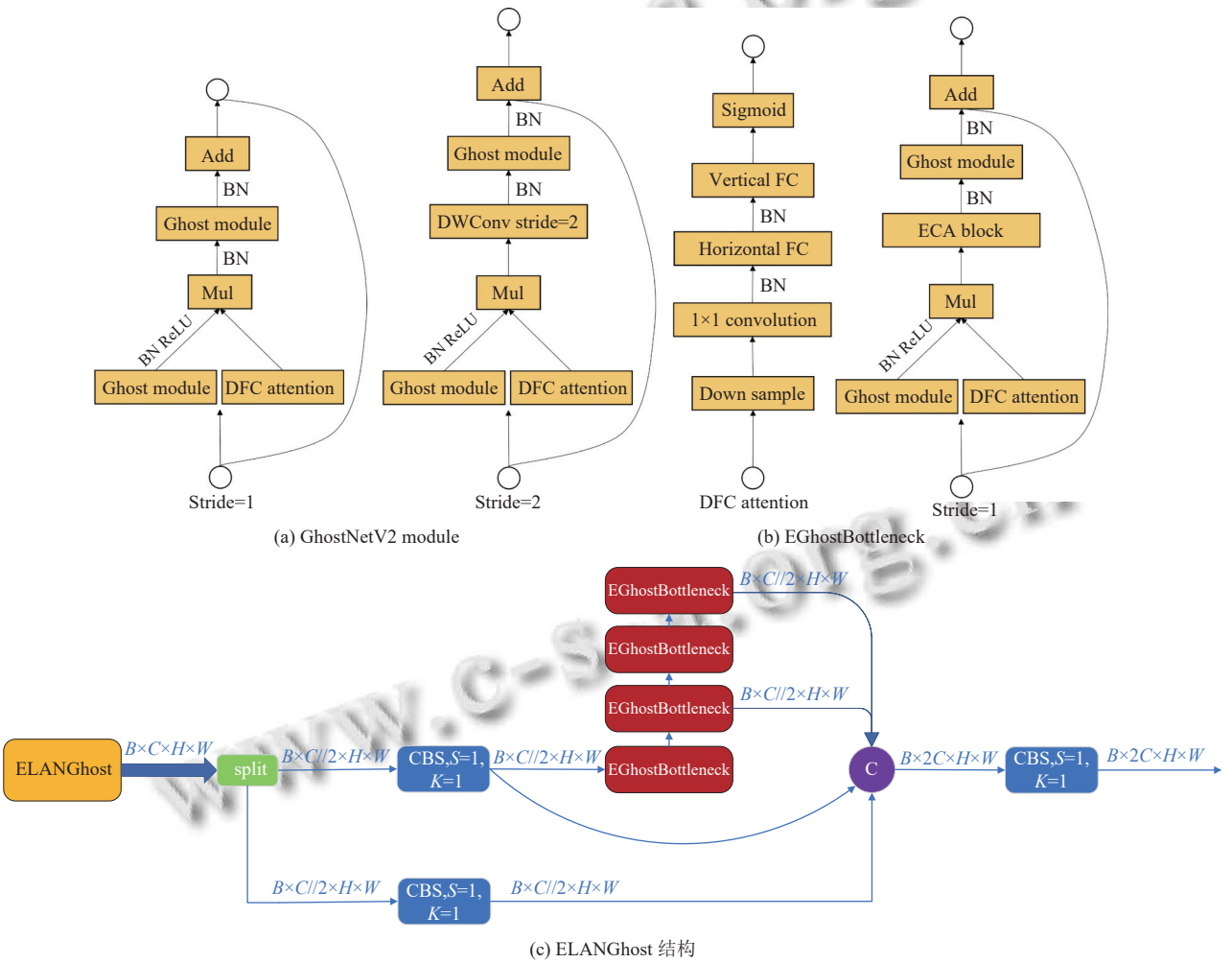


图 3 Ghost 结构示意图

ELAN 是 YOLOv7 提出的创新结构, 在 VoVNet, CSPNet 的基础上, 通过跨阶段特征融合策略和截断梯度流技术, 加入了更多分支, 减少冗余梯度信息的影响, 增强网络学习能力. 而 GhostNet 网络采用的是 MobileNetV3 的结构, 网络结构太深, 由此增加的模型参数量和计算量反而会使得模型总体的计算量上升, 达不到轻量化设计的目的. 因此, 在改进 YOLOv7 模型时, 并没有直接将 ELAN 结构用 GhostNet 简单替换, 而是设

计了 ELANGhost 结构, 将其中 3×3conv 模块替换成了 EGhostBottleneck 模块. 如图 3(c) 所示. 其中一半的特征信息通过 CBS 层和两个 EGhostBottleneck 的 3 个分支和另一半通过 1×1 卷积的特征信息进行融合, 最后将融合的特征信息通过 1×1 卷积进行调整, 模块最后输出的通道数是输入的两倍.

3.3 增加基于 SA 的特征融合模块

考虑到 Ghost 模块使用的轻量化操作减少了模型的全局特征提取能力, 同时注意力机制能够帮助模型提取特征图中感兴趣的区域, 因此在模型中增加了注意力机制, 弥补轻量化模块带来的精度损失.

SA 模块结合了空间和通道注意力机制, 特征图输入时, 首先会按照通道被分为多个组的子特征, 然后对每组子特征使用 SAUnit 进行处理, 就是将子特征拆分成两个分支, 一个分支学习空间注意力特征, 具体是通过组归一化 (GroupNorm) 获取空间维度的信息, 然后通过全连接层和 *Sigmoid* 进行增强; 另一个分支学习通道注意力特征, 为了实现轻量化的设计目标, SA 使用平均池化层将每个通道上所有特征图压缩为一个像素的特征, 然后通过全连接层和 *Sigmoid* 进行增强. 在完成两种注意力计算后, 首先通过 concat 进行融合, 然后使用通道置换操作进行不同组间的特征融合. SA 模块可以作为轻量化模块集成到 CNN 架构中, 并且可以帮助模型提取感兴趣的区域和特征, 减少无关信息对模型的影响, 提高模型的准确率.

在 YOLOv7 网络中, 将 FPN 和 PAN 结合使用进行多尺度特征融合, 目的是将浅层网络的强位置信息和深层网络的强语义信息传递给其他网络层. 为了更加充分利用骨干网络提取的底层特征, 本文将原始 FPN-PANet 换成结合了 BiFPN 和 SA 模块的网络, BiFPN 在网络中同一尺度的输入和输出节点之间添加了跳跃连接, SA 模块则被添加在 FPN 和 PAN 结构之间, 计算位置信息权重和语义信息权重, 关注重要的特征, 特征融合结构整体示意图如图 4 所示.

3.4 WDLoss 损失函数

本文的目标检测损失函数共包括 3 个部分: 分类损失 L_{class}, 边界框回归损失 L_{box} 和目标置信度损失 L_{object}, 3 类损失函数加权后构成了总的损失函数, 如下所示:

$$L_{\text{det}} = \alpha_1 L_{\text{class}} + \alpha_2 L_{\text{box}} + \alpha_3 L_{\text{object}} \tag{1}$$

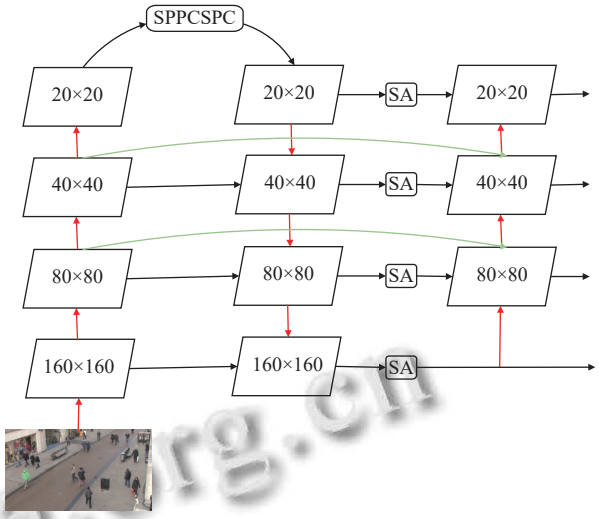


图 4 BSFPN 示意图

分类损失使用分类交叉熵损失, 计算公式如下:

$$y_i = Sigmoid(x_i) = \frac{1}{1+\mathrm{e}^{-x_i}} \tag{2}$$

$$L_{\text{class}} = -\sum_{n=1}^{N} y_i^* \log(y_i) + (1-y_i^*)\log(1-y_i) \tag{3}$$

式 (2) 和式 (3) 考虑到一个物体可能属于多个类别的问题, N 表示类别总数, x_i 表示当前类别的预测值, y_i 表示经过激活函数后的当前类别的概率, y_i^* 表示当前类别的真实值 (0 或 1).

原 YOLOv7 的回归损失采用的是 *CIoU*, 计算公式如下所示:

$$CIoU(B,B_{\text{gt}}) = IoU(B,B_{\text{gt}}) - \frac{\rho^2(B,B_{\text{gt}})}{c^2} - \beta v \tag{4}$$

$$v = \frac{4}{\pi^2}\left(\arctan\left(\frac{\omega^{\text{gt}}}{h^{\text{gt}}}\right) - \arctan\left(\frac{\omega}{h}\right)\right)^2 \tag{5}$$

$$\beta = \frac{v}{1 - IoU(B,B_{\text{gt}}) + v} \tag{6}$$

其中, $\rho(B,B_{\text{gt}})$ 为目标框和真实框中心点之间的距离, c 为包住目标框和真实框的最小外接框的对角线长度, v 用来度量目标框宽高比的一致性, β 为权重函数. 然而, 基于 *IoU* 的损失函数对于小目标的位置偏差非常敏感, 并且在用于基于锚点的检测器中会大大降低检测性能. 为了缓解这种情况, Wang 等人[27]提出了一种使用 Wasserstein 距离进行小目标检测的新评估指标. 具体来说, 就是先将边界框建模为 2D 高斯分布, 并提出了一个称为归一化 Wasserstein 距离的新度量, 以通过它们对应的高斯分布来计算它们之间的相似性.

本文针对 *IoU* 损失函数在小目标检测中存在的问题, 结合 Wasserstein 和 *IoU* 函数取代了 *CIoU* 损失函数计算边界框回归损失, 计算公式如下.

对于两个 2D 高斯分布N_a和N_b, 其二阶 Wasserstein 距离可以定义为:

$$W_2^2(\mu_1,\mu_2)=\|m_1-m_2\|_2^2+\left\|\sum_1^{\frac{1}{2}}-\sum_2^{\frac{1}{2}}\right\|_F^2 \tag{7}$$

其中, $\|\cdot\|_F$是范贝里乌斯范数.

考虑两个高斯分布的边界框$A=(cx_a,cy_a,\omega_a,h_a)$和$B=(cx_b,cy_b,\omega_b,h_b)$, 则式 (7) 可以被推广为:

$$W_2^2(N_a,N_b)=\left\|\left([cx_a,cy_a,w_a,h_a]^{\mathrm{T}},[cx_b,cy_b,w_b,h_b]^{\mathrm{T}}\right)\right\|_2^2 \tag{8}$$

然而, 式 (8) 的结果是一个距离度量, 需要通过归一化将其转化成为相似性度量 (0–1 内的值), 由此可以得到损失函数如下:

$$NWD(N_a,N_b)=\exp\left(-\frac{\sqrt{W_2^2(N_a,N_b)}}{C}\right) \tag{9}$$

通过调节C, 可以使检测器在不同数据集的边界框回归上具有更大的灵活性, 一般取C为数据集的平均绝对大小可以达到较好的效果, 并且C在一定范围内是鲁棒的, 这使得在调节超参数的时候能够更加灵活, 因此本文中将C取为数据集的平均绝对大小.

在得到 Wasserstein 距离后, 边界框回归损失的计算函数为如下:

$$L_{\mathrm{box}}=\alpha_1 CIoU(B,B_{\mathrm{gt}})+\alpha_2 NWD,\ \alpha_1+\alpha_2=1 \tag{10}$$

通过调节两个边界框损失的函数的权重α_1和α_2, 可以使检测器在不同尺寸的边界框回归方面具有更大的灵活性, $0.5<\alpha_1<1$ 时有助于提高 *IoU* 较大的目标回归精度.

3.5 基于 GhostNetV2 和注意力机制的 YOLOv7 模型

本文所提算法的整体结构如图 5 所示, 网络前 3 层仍然是卷积+BN+SiLU, 在经过一层步长为 2 的卷积得到 4 倍降采样图后, 将后续的所有非 1×1 卷积替换为 Ghost module, 所有 ELAN 结构替换为 ELAN-Ghost 结构, 并在第 1 个 ELAN 结构后新引出一条 P2 分支, 添加一条高分辨率检测头; 在后面的特征融合模块中, 在 FPN-PAN 结构中添加跳跃连接和 SA 模块, 实现 BSFPN 结构. 并在最后的模型训练过程中, 将原始的定位损失函数替换为 WDLoss 函数, 使模型训练对小目标更加鲁棒, 帮助锚框回归在特殊情况时也能提供梯度.

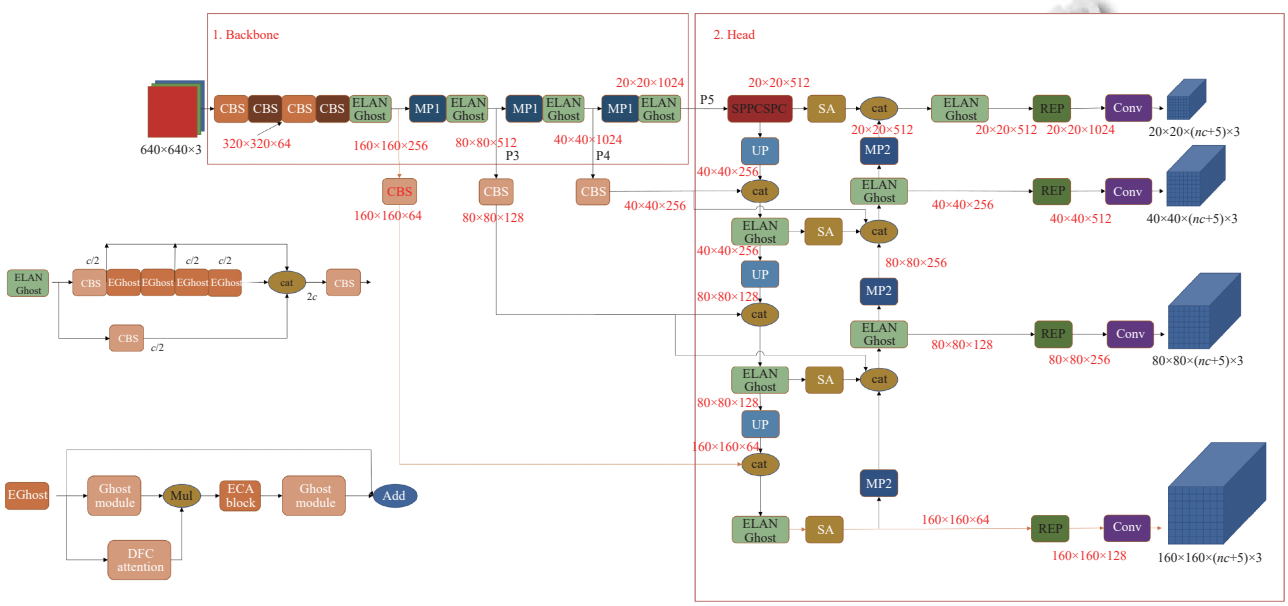


图 5 改进后的 YOLOv7 结构

4 实验分析

本文实验使用公开无人机数据集 VisDrone2019 和自制视频监控数据集评估模型性能, 并分别从数据集介绍, 实现细节, 实验结果与分析和消融实验 4 个方

面展开介绍.

4.1 数据集介绍

4.1.1 VisDrone2019 公开无人机数据集

VisDrone2019 数据集由天津大学机器学习和数据挖掘实验室 AISKYEYE 团队收集[28], 基准数据集包括 288 个视频片段、 261 908 帧和 10 209 幅静态图像. 数据集是在不同的场景、 不同的天气和光照条件下使用各种无人机平台 (即不同型号的无人机) 收集的, 共有约 260 万个标签, 包含行人, 人, 轿车, 货车, 公共汽车, 卡车, 摩托车, 自行车, 遮阳篷三轮车和三轮车 10 个类别, 里面包含大量尺寸小于 12×12 像素的小目标, 非常适合验证目标检测模型的小目标检测能力.

4.1.2 自制视频监控场景行人检测数据集

本文还采用了笔者自己采集并建立的数据集. 考虑到公开的视频监控场景数据集中较少有同时标注人体和人脸的数据集, 笔者自己使用 1920×1080 的网络摄像头拍摄了一个行人人脸检测数据集. 整个数据集在室外多个场景下拍摄行人, 采集总时长约两小时的视频, 剔除质量不高的图片后, 选取 2 000 张图像作为实验数据集, 其中训练集占比 80%, 验证集占比 10%, 测试集占比 10%.

4.2 实现细节

本文实验中最大训练次数为 300, 训练时的前 3 个 epoch 设置为 warmup, 初始的学习率设置为 0.01, 并使用 SGD 优化策略进行学习率调整, 并在最后一轮时降为 0.001, 训练时 *IoU* 阈值设定为 0.2, 训练过程中 batch-size 设置为 8, 输入图片尺寸均设置为 640×640. 在训练本文改进后的 YOLOv7 模型和其他模型时, 超参数均保持一致, 并且均不加载预训练模型. 本实验中使用的硬件配置如表 1 所示.

表 1 训练所用机器配置表

类型	型号	参数
系统	Ubuntu	16.04.2 LTS
CPU	Intel(R) Core(TM) i9-9900	8核
GPU	Nvidia GeForce GTX 1080Ti	11 GB
内存	DDR4	32 GB

4.3 实验结果

为了验证实验结果, 本文使用准确率, 平均精度 (average precision, *AP*), 平均精度均值 (mean average precision, *mAP*) 来评估模型的检测性能.

准确率的定义如下:

$$\begin{aligned} Accuracy &= \frac{2\times P\times R}{P+R}\\ P &= \frac{TP}{TP+FP}\\ R &= \frac{TP}{TP+FN} \end{aligned} \tag{11}$$

其中, FP是假阳性, FN 是假阴性, TP 是真阳性.

平均精度和平均精度均值的定义如下:

$$\begin{aligned} AP &= \int_0^1 P(R)\mathrm{d}R\\ mAP &= \frac{\sum P_A}{N_C} \end{aligned} \tag{12}$$

其中, N_C 为类别数量, P_A 为各类别的平均精度. 利用实验数据可绘制模型的 *PR* 曲线, 曲线所围面积即为 *AP*, 该指标被用来评估模型对于单个类别的目标检测性能表现, 将所有类别的 *AP* 值取平均就得到了 *mAP*, 实验中使用到了两种 *mAP*, 分别是 *mAP*@.5 和 *mAP*@.5:.95, 前者表示 *IoU* 设定为 0.5 时, 所有类别的平均精度, 后者表示 *IoU* 阈值在 (0.5, 0.95) 区间内, 步长为 0.05, 分别计算 *mAP*, 然后取平均值. *mAP* 取值在 0–1 之间, 其值越接近于 1, 表示模型的性能越好, 在不同类型标签上的检测能力越强.

检测速度常用的指标是模型每秒检测的图片数量, 单位为 f/s. 模型大小的评估可以采用计算量和参数量两个指标, 计算量是模型的浮点运算次数, 单位 GFLOPs; 参数量是模型参数的数量总和, 单位为 M.

4.3.1 VisDrone2019 数据集实验分析

为了验证本文算法在小目标上的有效性, 本文采用 VisDrone2019 数据集作为基准数据集. 为保证实验的严谨性与可对比性, 本文实验参考 VisDrone2019 数据集的实验设置, 输入图片尺寸均设置为 640×640, 且均采取相同的数据预处理方法, 并在 VisDrone2019 数据集上分别验证了 YOLOv5l, YOLOv7 与 Cascade R-CNN 的实验结果, 并与本文算法进行对比, 对比结果如表 2.

从表 2 中可以看出, 本文对标 YOLOv7, *R* 由 0.491 提升到 0.493, *mAP* 指标由 0.487 提升到 0.501, 但由于模型参数量的增加, 导致模型的推理速度略有下降, 由 73.53 f/s 降低到 64.51 f/s. 可能是数据量较小的原因, 基于 Transformer 的 RT-DETR 方法并没有取得理想中的效果. 对比 YOLOv7, 本文算法在略微降低检测速度的基础上在检测精度、泛化性能上都得到了优化.

表 2 VisDrone2019 数据集实验结果 (均为本地实验结果)

模型	*R*	*mAP*@.5	*mAP*@.5 : .95	推理速度 (f/s)
YOLOv7	0.491	0.487	0.277	73.53
Cascade R-CNN	0.411	0.403	0.258	60.54
YOLOv5l	0.407	0.417	0.25	**80.57**
YOLOv8x	0.432	0.446	0.274	66.67
RT-DETR	0.359	0.337	0.19	51.99
本文算法	**0.493**	**0.501**	**0.289**	64.51

4.3.2 模块有效性评价

为了分析各模块组合对模型的影响, 设计消融实验, 采用同样的硬件配置和超参数, 训练 300 个 epoch, 并在 VisDrone2019 数据集上训练, 取 *mAP* 最高的结果保存, 结果如表 3 所示, 可以看到, 加入 P2 高分辨率分支的效果很好, 检测精度上升了 2.1%, 但模型的参数量增加了 0.52M, 导致模型有一定程度的推理速度下降; GhostV2 轻量化模块的加入使得模型整体的参数量大幅下降, 约减少了 7M 左右的参数量, 但检测精度等指标下降了约 2 个百分点; BSFPN 的加入提升了模型的精确度和召回率, 同时 *mAP* 指标相对于 YOLOv7 提升了 1.8%, 说明在特征融合模块中加入注意力模块和跳跃连接能够帮助模型更好地融合不同层面的特征, 同时参数量也只增加了 0.5M; 损失函数的修改是能够帮助网络克服小目标的尺度敏感性问题, 但是由于 *NWD* 在损失函数中的占比较小, 因此模型的精确度和召回率仅是略有上升. 本文算法将 P2、EGhost、BSFPN 和 WDLoss 这几种思想融入 YOLOv7, 可以看到, 模型整体的准确率和召回率都有所上升, 同时 *mAP* 指标也得到了优化, 但是数据集不够大, 注意力机制未能有很好的表现, 对比只添加了 P2 的模型, 性能有轻微下降.

表 3 VisDrone2019 数据集消融实验 (均为本地实验结果)

模型	*P*	*R*	*mAP*@.5	*mAP*@.5 : .95	参数量 (M)
YOLOv7	0.576	0.491	0.487	0.277	36.9
+P2	0.578	**0.517**	**0.505**	**0.295**	37.42
+EGhost	0.548	0.473	0.462	0.262	**29.73**
+BSFPN	0.582	0.502	0.496	0.286	37.4
+WDLoss	0.582	0.498	0.493	0.28	36.9
本文算法	**0.584**	0.493	0.501	0.289	31.3

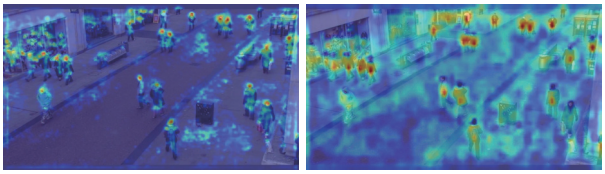
4.3.3 自制视频监控数据集实验分析

为了验证本文所提出算法的有效性, 将本文算法与经典的目标检测方法在测试集上进行实验结果对比, 如表 4 所示. 表中对比了不同网络的召回率和平均精度以及推理速度, 从表 4 中可以看出, 本文方法对比其他目标检测网络效果较好, 更有利于对视频监控场景下的行人检测.

表 4 自制数据集实验结果

模型	*P*	*R*	*mAP*@.5	*mAP*@.5 : .95
YOLOv7	0.806	0.718	0.779	0.453
Cascade R-CNN	0.794	0.716	0.766	0.422
YOLOv5l	0.797	0.696	0.757	0.414
YOLOv8x	**0.831**	0.708	0.785	0.446
本文算法	0.804	**0.732**	**0.795**	**0.455**

为了验证增加注意力模块能够帮助网络在提取特征时学习到感兴趣的区域, 将本文算法提取的特征可视化, 如图 6 所示, 分别展示了网络预测人脸和人体时哪些区域对预测结果影响最大. 从图 6 中可以看到, 本文方法提取到的高响应区域基本集中在人们认为有利于帮助判断人脸和人体的部位, 说明模型学习到了有用的特征. 由图 7 可知, 随着训练的进行, Loss 值逐渐下降, *R* 和 *mAP* 值稳步上升. 在训练进行到一定轮次时, 模型收敛, 此时 Loss, *R* 和 *mAP* 的值处于一个稳定区间内, 停止训练, 得到训练好的模型.



(a) 预测人脸时网络的高响应区域 (b) 预测人体时网络的高响应区域

图 6 本文方法热力图可视化

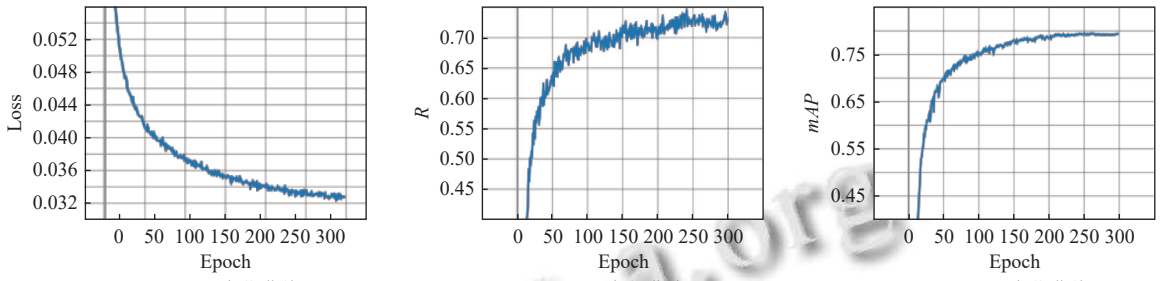
4.4 实验结果对比

为了体现本文算法的有效性, 从测试集中选取了图像进行检测, 检测结果如图 8 所示. 在图 8 中, 行人数量约为 7 位, 不同行人之间存在车辆遮挡, 行人重叠等情况, 属于较复杂监控场景. YOLOv7 在检测中出现人脸的漏检现象, 而使用本文改进的 YOLOv7 算法检测, 则检测出了所有正面的人脸, 同时两者都检测除了所有行人. 这表明本文提出的模型对于小目标有着不错的检测能力, 能够实现对视频监控画面中人体和人脸的同时检测, 能够实现对视频监控场景下行人以及人脸的有效检测.

5 结论与展望

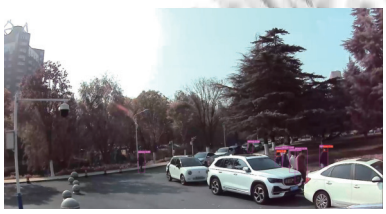
针对小目标难以检测的问题, 本文提出了一种改进的 YOLOv7 检测模型, 通过在检测头部分增加一个高分辨率检测头用于小目标检测, 有效提高了模型对于小目标的检测精度; 并在将骨干网络中的部分 Conv 替换成 EGhostBottleneck 模块, 减少网络的参数量; 在

骨干网络和检测头部分连接处, 结合 BiFPN 和 SA 改进特征融合结构, 强化底层特征利用; 并结合了 Wasserstein 距离改进了损失函数, 提升了模型在小目标上的性能. 最后, 本文在 VisDrone2019 数据集上和自制视频监控数据集上做了对比实验, 不同指标均得到了一定提升, 证明了本文方法的有效性. 不过, 由于增加了额外的检测头的同时网络结构也更加复杂, 导致模型的大小增加, 在推理速度上略低于原模型, 后续工作致力于在提升检测精度的同时减少模型推理量, 加快模型检测速度.



(a) Loss 变化曲线 (b) *R* 变化曲线 (c) *mAP* 变化曲线

图 7 训练过程曲线变化图



(a) 原 YOLOv7 检测结果

(b) 改进 YOLOv7 检测结果

图 8 室外视频监控场景

参考文献

1 Zhao ZQ, Zheng P, Xu ST, *et al*. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3212–3232. [doi: 10.1109/TNNLS.2018.2876865]

2 Cai ZW, Vasconcelos N. Cascade R-CNN: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1483–1498. [doi: 10.1109/TPAMI.2019.2956516]

3 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

4 Song XY, Gu W. Multi-objective real-time vehicle detection method based on YOLOv5. Proceedings of the 2021 International Symposium on Artificial Intelligence and Its Application on Media (ISAIAM). Xi'an: IEEE, 2021. 142–145. [doi: 10.1109/ISAIAM53259.2021.00037]

5 张凯祥, 朱明. 基于 YOLOv5 的多任务自动驾驶环境感知算法. 计算机系统应用, 2022, 31(9): 226–232. [doi: 10.15888/j.cnki.csa.008698]

6 Li CY, Li LL, Jiang HL, *et al*. YOLOv6: A single-stage object detection framework for industrial applications. arXiv: 2209.02976, 2022.

7 Zhang SF, Zhu XY, Lei Z, *et al*. S^3fd: Single shot scale-invariant face detector. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 192–201.

8 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. ICLR, 2021.

9 Liu Z, Mao HZ, Wu CY, *et al*. A ConvNet for the 2020s. Proceedings of the 2022 IEEE/CVF Conference on Computer vision and Pattern Recognition. New Orleans: IEEE, 2022. 11966–11976.

10 Rezatofighi H, Tsoi N, Gwak J, *et al*. Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 658–666.

11 Zheng ZH, Wang P, Liu W, *et al*. Distance-IoU loss: Faster and better learning for bounding box regression. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 12993–13000.

12 Hu J, Shen L, Sun G. Squeeze-and-excitation networks.

Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

13 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceeding of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19.

14 Zhu XK, Lyu SC, Wang X, *et al*. TPH-YOLOv5: Improved YOLOv5 based on Transformer prediction head for object detection on drone-captured scenarios. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021. 2778–2788.

15 Deng CF, Wang MM, Liu L, *et al*. Extended feature pyramid network for small object detection. IEEE Transactions on Multimedia, 2022, 24: 1968–1979. [doi: 10.1109/TMM.2021.3074273]

16 Wang QL, Wu BG, Zhu PF, *et al*. ECA-Net: Efficient channel attention for deep convolutional neural networks. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 11531–11539. [doi: 10.1109/CVPR42600.2020.01155]

17 Tang YH, Han K, Guo JY, *et al*. GhostNetV2: Enhance cheap operation with long-range attention. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 9969–9982.

18 Tan MX, Pang RM, Le QV. EfficientDet: Scalable and efficient object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 10778–10787. [doi: 10.1109/CVPR42600.2020.01079]

19 Zhang QL, Yang YB. SA-Net: Shuffle attention for deep convolutional neural networks. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021. 2235–2239.

20 Wang CY, Bochkovskiy A, Mark Liao HY. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 7464–7475.

21 Lee Y, Hwang JW, Lee S, *et al*. An energy and GPU-computation efficient backbone network for real-time object detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE, 2019. 752–760. [doi: 10.1109/CVPRW.2019.00103]

22 Wang CY, Mark Liao HY, Wu YH, *et al*. CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle: IEEE, 2020. 1571–1580. [doi: 10.1109/CVPRW50498.2020.00203]

23 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2261–2269.

24 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944.

25 Liu S, Qi L, Qin HF, *et al*. Path aggregation network for instance segmentation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8759–8768.

26 Han K, Wang YH, Tian Q, *et al*. GhostNet: More features from cheap operations. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 1577–1586. [doi: 10.1109/CVPR42600.2020.00165]

27 Wang JW, Xu C, Yang W, *et al*. A normalized Gaussian Wasserstein distance for tiny object detection. arXiv: 2110.13389, 2021.

28 Zhu PF, Wen LY, Du DW, *et al*. Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 7380–7399. [doi: 10.1109/TPAMI.2021.3119563]

(校对责编: 孙君艳)