文章编号：2096-1618(2024)04-0422-08

自注意力结合上下文解耦的交通车辆检测

孙光灵[1,2]， 周云龙[1]

(1. 安徽建筑大学电子与信息工程学院，安徽 合肥 230601；2. 合肥工业大学智能互联系统安徽省实验室，安徽 合肥 230009)

摘要：为应对车流量、时间地点和天气等因素给交通车辆检测带来的挑战，提出基于 YOLOv5s 模型的新算法。该改进模型适用于各种交通场景，其改进如下：在特征串联阶段引入高效的二维局部特征叠加自注意力(ELFSa)，以增强模型对目标的感知能力；将 YOLOv5s 的检测头替换为简易特定于任务的上下文解耦(S-TSCODE)，以实现分类和定位子任务之间的完美平衡，从而改善模型的收敛；为减少运算负担，模型中的大于 3×3 的部分卷积操作被替换成了 GSConv。实验结果显示，改进的 YOLOv5s 在各方面均有提升，其中 $mAP_{@0.5}$ 为98.9%，$mAP_{@0.5:0.95}$ 为87.0%，分别提升0.1% 和1.5%。针对各种复杂的交通场景，所提出的方法增强了车辆检测的性能和鲁棒性。

关 键 词：目标检测；自注意力；解耦检测头；轻量化卷积；YOLOv5s

中图分类号：TP391.4 **文献标志码**：A

doi：10.16836/j.cnki.jcuit.2024.04.005

0 引言

在这个科技高度发展时代，交通问题成为人们日常出行面对的挑战，尤其是交通堵塞和早晚高峰等问题。通过目标检测技术为城市智能交通系统[1]提供各种道路交通异常信息，有助于减缓常见的交通问题。然而，经过调查学习后发现，在实际车辆检测[2]过程中，由于车流量庞大，导致一些车辆被其他车辆遮挡，增加了检测难度。此外，随着地点和时间的变化，光照条件对于车辆检测造成极大影响，加上不同的天气条件，如雨雪等，也会给车辆检测带来困难。为解决上述环境变化的问题，需要收集涵盖不同时间段和天气环境的多样化数据。

Cheng 等[3]和 Zhang 等[4]提出一种用于探测图像中车辆目标的技术。该方法应用边缘运算器来计算图像的边缘信息，借助车辆尾部的对称性这一局部特征来实现。Girshick 等[5]提出的 Fast R-CNN 是一种双阶段检测器，将分类和定位任务分隔开。这种算法通过增加预处理时间来换取检测精度。为获得高精度的目标检测器，在 Cascade R-CNN[6]的研究中，为持续优化检测结果，首次引入一种通过级联多个基于卷积神经网络而形成的检测网络。

然而交通车辆检测需要较快的检测速度，以上几种方法无法达到即时性要求。因此模型选用单阶段目标检测器 YOLOv5。随着 ViT[7]的发展，注意力机制在目标检测领域的应用越来越广泛。因此使用 2D 局部特征叠加自注意力(local feature superimposed self-attention，LFSa)[8]和特定于任务的上下文解耦(task-specific context decoupling，TSCODE)[9]对 YOLOv5s 模型进行改进，以求在交通车辆检测实现更良好的效果。

1 YOLOv5 模型介绍

YOLOv5 是 Uitralytics LLC 公司发布的一种有代表性的单阶段检测算法。如图 1 所示，模型主要包括主干(BackBone)、中间颈部(Neck)和检测头(Head)。

主干网络采用 CSPDarknet53 网络，将原始的输入图像转化为多层特征图。CBS 模块负责对输入进行 2 倍下采样。C3 模块的功能在于扩展网络的深度和感受野，以此提升特征提取的性能。SPPF 在主干网络中作为空间金字塔池化模块，能把局部特征和全局特征融合起来，而无须调整特征图的尺寸。

在颈部应用路径聚合网络，经过上采样和下采样操作后会形成不同尺度和不同特征信息的特征图，这些特征图通过通道拼接进行多尺度特征融合，构建形如金字塔的特征结构。自下向上则通过卷积层融合来自不同层次的特征图。通过上采样和通道信息融合获取更粗粒度的特征信息是自顶向下的作用。检测头有 3 个尺度的输出，每一个输出涵盖了分类和边界框的信息，并使用非极大值抑制在输出的边界框中抑制重叠的框，只保留最具有代表性的边界框。

收稿日期：2023-08-15

基金项目：国家自然科学基金资助项目(62001004)；安徽省高校协同创新项目(GXXT-2021-024)；2023 年安徽省住房城乡建设科学技术计划资助项目(2023-YF058、2023-YF113)

通信作者：孙光灵. E-mail：sunguangling@163.com

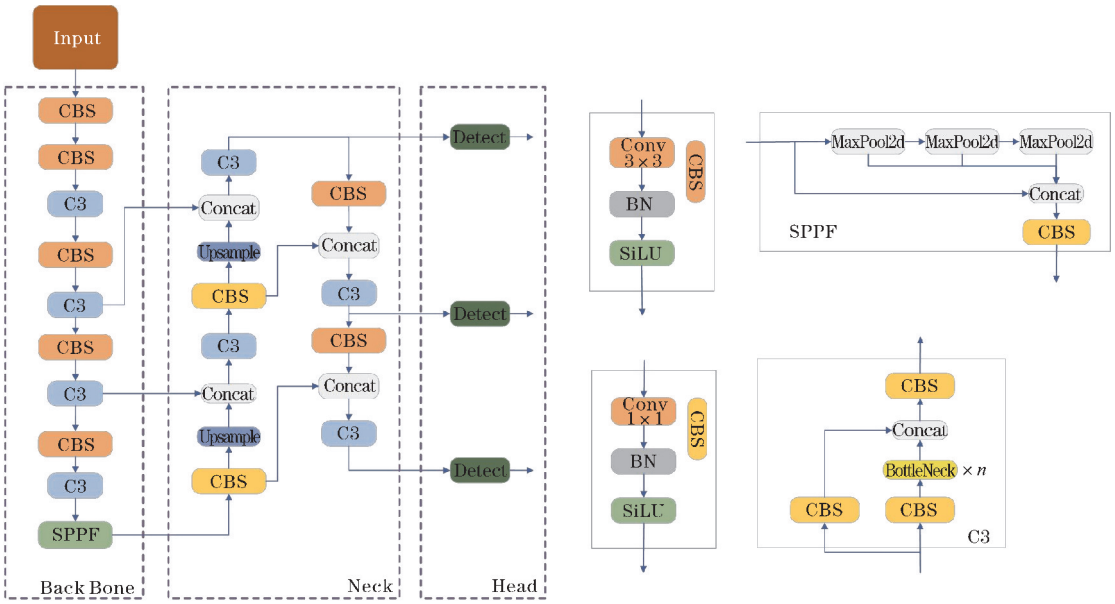


图 1　YOLOv5 网络结构图

2　YOLOv5s 改进

2.1　GSConv

一味地通过增加模型参数量无法建立优秀的模型,而轻量化设计[10]能有效缓解当前高昂的计算成本,提高 YOLOv5s 的推理速度。

虽然标准卷积(standard convolution,SC)可以用于模型表达能力,但在适应不同尺度的输入和捕捉空间关系方面仍有限制。深度可分离卷积(depthwise separable convolution,DSC)[11]是通过对每一个通道单独卷积从而减少参数量,但有特征表示能力的限制。DSC 对输入尺度变化敏感,并且无法直接捕捉全局上下文信息。

在卷积网络中,递送图像必须在主干中由空间信息逐步向通道转换(图 2)。然而,每次特征图的空间压缩和通道扩展都会导致部分特征信息的丢失。

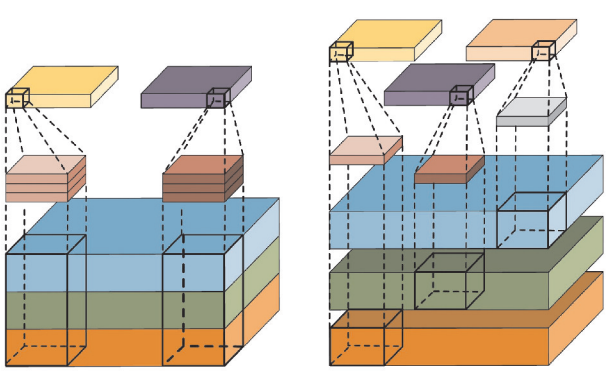


图 2　SC 和 DSC 卷积操作示意图

SC 在计算时极大地保留了每个通道之间的连接,有助于更好地提取语义信息。相比之下,DSC 则完全切断了通道之间的连接,虽然减少了大部分参数量和浮点运算量,却导致了提取语义信息的效果不佳。

尽管 DSC 在计算资源限制下表现出轻量化优势,但其局限性也需被熟知,特别是在需要较高语义信息的任务中,可能无法达到与标准卷积相当的性能。

模型引入 GSConv[12],旨在降低模型计算复杂度的同时,实现类似于 SC 的效果。GSConv 通过混合使用 SC、DSC 和 shuffle[13]实现接近 SC 的卷积效果。如图 3 所示,GSConv 使用 shuffle 技术将 SC 生成的信息渗透到最终输出的每一个部分。

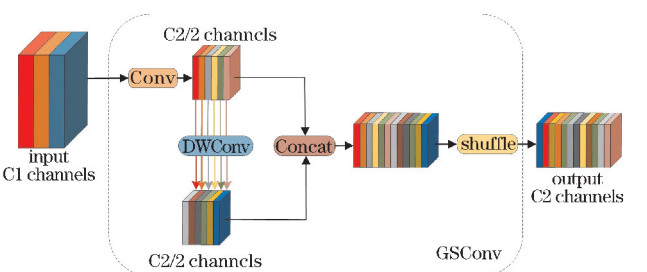


图 3　GSConv 模块结构图

在 GSConv 的结构中,其中一半通道使用 SC 进行卷积操作,以保留通道之间的特征联系,而另一半通道使用 DSC 避免参数冗余。单独加入 GSConv 对模型性能有小幅度的提升。

2.2　2D 局部特征叠加自注意力

在 YOLOv5s 颈部结构中,使用了流行的特征金字塔网络(feature pyramid network,FPN)+PAN[14]结构。该结构的特征串联阶段用于连接来自主干网络的多个尺度特征图的输出,涵盖多种至关重要的特征信息。

在此阶段(图 4)添加挤压-激发网络[15]或卷积块注意力模块[16]后,发现模型的精度并没有太大提高。也尝试在这里添加多头自注意力模块,但是复杂的模块给模型带来沉重的负担,使其复杂度急剧上升,从而导致模型的检测速度急剧下降,无法满足实时检测交通车辆的要求。为克服这些挑战,引入了高效的局部特征叠加自注意力(efficient local feature superimposed self-attention,ELFSa)。

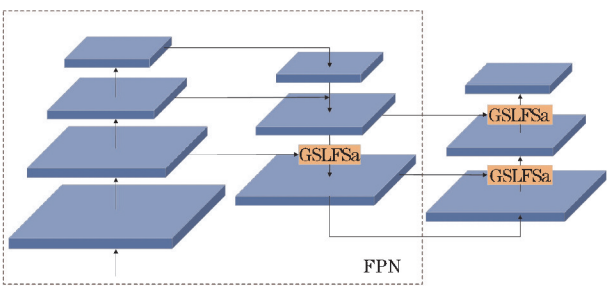


图 4 在特征串联阶段加入 ELFSa

自注意力模块通常一次性计算全部通道的特征图,导致极大的参数量和浮点运算量,对于实时检测模型的推理速度有严重的影响。与自注意力模块的计算不同,LFSa 只计算每个通道的特征图中行或列的局部特征,经过权重叠加后实现自注意力模块的功能。

根据图 5 的描述,假设 $\boldsymbol{X}\in \mathrm{R}^{C\times H\times W}$ 为输入特征图,首先使用一个 3×3 的 GSConv 对输入特征图 $\boldsymbol{X}$ 提取特征,获取高级语义信息。因为后续的卷积操作中使用 1×1 和 7×7 的尺寸,这使得模型传递的细节特征在此处有一定程度的丢失。选择在输入特征图 $\boldsymbol{X}$ 时加入一个 3×3 的 GSconv,加强细节特征的提取。再经过 3 个 1×1 卷积后得到 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 3 个特征图充当注意力的计算因子,$\boldsymbol{Q}_i$、$\boldsymbol{K}_i$、$\boldsymbol{V}_i$ 来自特征图的同一个通道。ELFSa 行和列的注意力计算公式:

$$\boldsymbol{F}_i^{\mathrm{row}}(\boldsymbol{X})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i(\boldsymbol{X})\boldsymbol{K}_i^{\mathrm{T}}(\boldsymbol{X})}{\sqrt{H}}\right)\boldsymbol{V}_i(\boldsymbol{X})$$

$$\boldsymbol{F}_i^{\mathrm{col}}(\boldsymbol{X})=\left(\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_i^{\mathrm{T}}(\boldsymbol{X})\boldsymbol{K}_i(\boldsymbol{X})}{\sqrt{H}}\right)\boldsymbol{V}_i^{\mathrm{T}}(\boldsymbol{X})\right)^{\mathrm{T}}$$

式中 $\boldsymbol{F}_i^{\mathrm{row}}(\boldsymbol{X})$ 和 $\boldsymbol{F}_i^{\mathrm{col}}(\boldsymbol{X})$ 表示特征图的一个通道的行或列的加权特征,对于一个通道的特征图是全局的,对于所有通道来说是局部的。

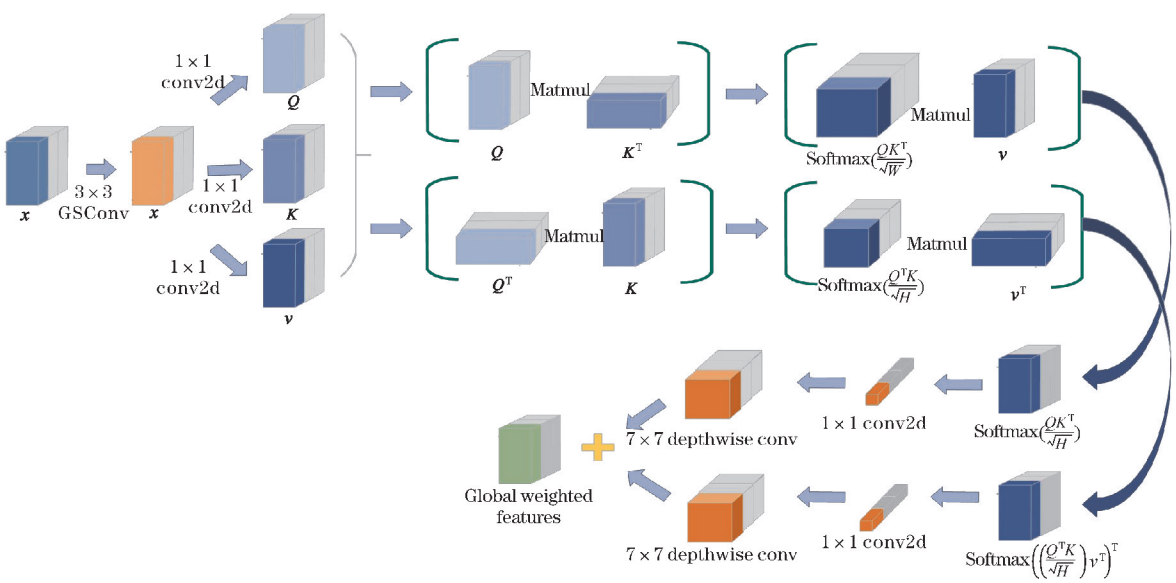


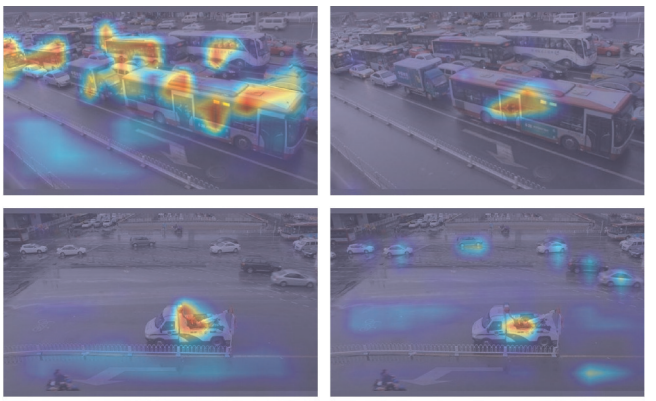
图 5 ELFSa 结构图

在 $\boldsymbol{F}_i^{\mathrm{row}}(\boldsymbol{X})$ 和 $\boldsymbol{F}_i^{\mathrm{col}}(\boldsymbol{X})$ 处理这两个行和列的加权特征时,首先进行了自注意力操作,但由于该操作仅对一个通道的特征图,导致通道之间的信息联系中断。为解决这一问题,通过使用 1×1 卷积分别打通行和列上的加权特征的通道联系,从而提取通道维度中 $\boldsymbol{F}_i^{\mathrm{row}}(\boldsymbol{X})$ 和 $\boldsymbol{F}_i^{\mathrm{col}}(\boldsymbol{X})$ 的通道特征。通过在一个通道上的特征进行自注意力操作,实现了全局的注意力机制。对于一个通道这是全局的,使用 1×1 卷积获取通道特征,这使得通过局部感受野得到了整个图片的全局加权特征。当然,使用 1×1 卷积得到的感受野具有局限性,因此在 1×1 卷积后使用 7×7 的卷积来对每个通道的特征图扩充感受野。使用行和列的全加权特征表示全局特征映射:

$$\mathrm{ELFSa}=\boldsymbol{X}+\mathrm{DWConv}(\mathrm{Conv}(\boldsymbol{F}^{\mathrm{row}}(\boldsymbol{X})))+\mathrm{DWConv}(\mathrm{Conv}(\boldsymbol{F}^{\mathrm{col}}(\boldsymbol{X})))$$

图 6(a)是 YOLOv5s 加入了 ELFSa 的检测热力图,图 6(b)是原始的 YOLOv5s。第一行是对公交车的检测,理想的热力图应覆盖所有公交车,然而 YOLOv5s 的热力图只集中在最近的公交车上,相比加入 ELFSa 后,其特征注意力有局限性。第二行是对其

他车辆的检测,在 YOLOv5s 的检测中,虽然检测到了其他车辆,但仍关注了汽车的特征。而加入 ELFSa 的检测模型则更专注、更精准地检测其他车辆。



(a) YOLOv5s+GSLFSa　　(b) YOLOv5s

图 6　YOLOv5s 加入 ELFSa 和原始模型的热力图比较

2.3　简易特定于任务的上下文解耦

YOLOv5 检测头使用单一卷积实现分类和定位任务,这两个任务在目标检测中是最主要的子任务。然而,由于任务目的性不同,对于特征的上下文需求也不同。定位任务需要更多具有边界感知的特征以准确定位边界框,而对象分类则更偏好更多语义上下文的特征。

YOLOv5 的检测头导致模型耦合性较高,对于处理不同大小的目标具有一定的局限性,这是因为分类和定位任务无法获取其所需要精确的上下文信息。YOLOX[17] 的解耦头可更好地同时处理来自不同层级特征图信息,从而处理不同尺寸的目标。但 YOLOX 的解耦头仍应用于相同的输入特征,导致分类和定位两个子任务之间的平衡不够完美。

因此,在 YOLOv5s 模型中引入 TSCODE,进一步解开两个子任务的特征编码,生成分别适用于分类和定位特征信息。在 TSCODE 中,通过 3 层特征融合生成了符合分类任务的语义上下文信息编码(semantic context encoding,SCE)和符合定位任务的细节保留编码(detail-preserving encoding,DPE),并将其输入到检测头部进行后续处理。

尽管引入 TSCODE 可以提升性能,但同时也增加了模型的复杂度,使 YOLOv5s 模型的参数量和浮点计算量大幅增加。因此对 TSCODE 进行简化和降参是必要的。TSCODE 将 3 层特征融合后生成的 SCE 和 DPE 输入到检测头部,以用于分类和定位任务(图 7)。

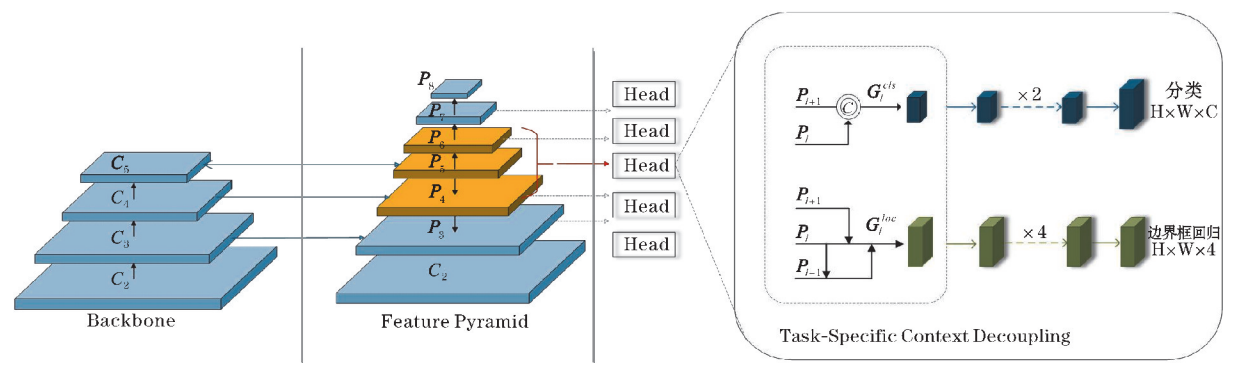
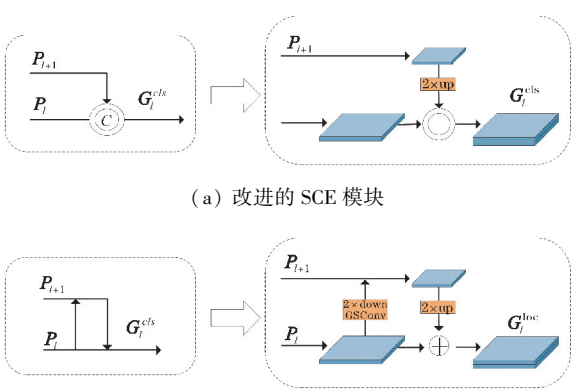


图 7　TSCODE 原版的结构示意图

相对于原版 SCE,新方法将 $\boldsymbol{P}_{l+1}$ 层进行 2 倍上采样与 $\boldsymbol{P}_l$ 层通道拼接(图 8(a))。通过这种方式将低级语义特征融合到上级语义特征中,为分类子任务提供更多需要的语义上下文信息。

与分类任务不同,定位是一个更细粒度的任务,需要更多的纹理细节和边界信息来进行精准的定位。在 DPE 改进中(图 8(b)),将输入选择为两层特征,即 $\boldsymbol{P}_l$ 和 $\boldsymbol{P}_{l+1}$。将 $\boldsymbol{P}_l$ 层进行 2 倍下采样,然后将一些能够提供更多物体视角的特征融合到 $\boldsymbol{P}_{l+1}$ 中,再将 $\boldsymbol{P}_{l+1}$ 进行 2 倍的上采样与 $\boldsymbol{P}_l$ 加权相加。这样的设计使得 $\boldsymbol{P}_l$ 增强了物体的视角,并获得了 $\boldsymbol{P}_{l+1}$ 提供的细节和边缘特征。简易特定于任务的上下文解耦(simple task-specific context decoupling,S-TSCODE)通过舍弃相对复杂的结构,实现了良好的效果,并减轻了整体模型的负担。



(a) 改进的 SCE 模块

(b) 改进的 DPE 模块

图 8　改进后的 SCE 和 DPE 模块结构

主流的解耦头通常是基于相同的输入 $\boldsymbol{P}_l$ 来最小化分类和定位损失,但由于来自相同的特征,分类和定位之间的冲突在 $\boldsymbol{P}_l$ 中施加了相反的上下文偏好,导致

这两个任务之间的不完美平衡。

$$L=L_{\text{cls}}(F_c(\boldsymbol{P}_l),c)+L_{\text{loc}}(F_r(\boldsymbol{P}_l),B)$$

为解决该问题,S-TSCODE 通过 SCE 和 DPE 模块将特征划分为 $\boldsymbol{G}_l^{\text{cls}}$ 和 $\boldsymbol{G}_l^{\text{loc}}$,为两个子任务提供适合的、特定的特征输入,完美解决该问题

$$L=L_{\text{cls}}(F_c(\boldsymbol{G}_l^{\text{cls}}),c)+L_{\text{loc}}(F_r(\boldsymbol{G}_l^{\text{loc}}),B)$$

式中,$F_c(\cdot)$和 $F_r(\cdot)$分别是分类和定位的分支,c 代表类标签,B 代表边界框。

$$F_c(\cdot)=\{f_{\text{cls}}(\cdot),C(\cdot)\}$$

$$F_r(\cdot)=\{f_{\text{loc}}(\cdot),R(\cdot)\}$$

式中,$f_{\text{cls}}(\cdot)$和 $f_{\text{loc}}(\cdot)$是用于分类和定位的特征投影函数,$C(\cdot)$和 $R(\cdot)$分别是分类和定位的最后一层,将特征解码为分类得分和边界框位置信息。

3 实验

3.1 实验环境与评价指标

实验在 Windows11 系统上进行,使用 PyCharm 2022.3.2(Professional Edition)配置 Anaconda 的虚拟环境进行实验。具体实验环境详见表 1。

表 1 实验环境具体配置

名称	参数
GPU	Nvidia 3060-12G
CPU	AMD Ryzen 5 5600 6 核 12 线
实验平台	PyCharm 2022.3.2(Professional Edition)
环境配置	Anaconda
Pytorch	2.0.0
CUDA	11.8
cuDNN	8700
Python 解释器	3.9.16

为准确验证改进模型的性能,在改进的 YOLOv5s 模型与其他目标检测模型对比时,保证同一实验环境。选用精确率(Precision)、召回率(Recall)、$\text{mAP}_{@0.5}$(IoU 阈值为 50%)、$\text{mAP}_{@0.5:0.95}$(IoU 阈值在 50% ~95%)作为模型检测性能的评价指标。具体计算公式如下:

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

式中:TP 表示成功识别正样本的数量;FP 表示误判,将负样本误分类为正样本的数量;FN 与 FP 相反,是将正样本错误分类为负样本的数量。

3.2 数据集

实验使用 PASCAL VOC[18] 和 UA-DETRAC[19] 两个数据集。VOC 数据集是一个广泛用于目标检测和图像分割算法评估的数据集,最初由英国牛津大学的视觉几何组创建。实验使用的 VOC(2007+2012)数据集训练集有 16551 张图片,校验集有 4952 张。

UA-DETRAC 是车辆检测和跟踪的大规模数据集,数据集手动标注了北京和天津过街天桥(京津冀地区)拍摄的图片,共计 121 万目标对象边框和 8250 个车辆。该数据集涵盖了多云、夜间、晴天和雨天 4 种天气在不同时间地点的轿车、公交车、厢式货车和其他类型车辆的图片。由于 UA-DETRAC 数据集较为庞大,因此使用 Python 脚本对视频每隔 10 帧提取 1 张图片,并从 VOC 格式的 xml 标签文件读取目标类被和边界框坐标信息,将坐标信息归一化转为 YOLO 格式的 txt 标签文件。最终通过脚本得到训练集和测试集,分别有 9801 张和 2144 张图片。

3.3 实验与结果分析

3.3.1 YOLOv5s 改进与原版结果分析

改进模型与原始模型相比,两个数据集均有提升(表 2),这证明了改进模型的泛化能力优越和具有较好的鲁棒性。在 VOC 数据集上,改进模型 P 提升 2.1%,$\text{mAP}_{@0.5}$ 提升1.6%,$\text{mAP}_{@0.5:0.95}$ 提升4.9%,其中 $\text{mAP}_{@0.5:0.95}$的提升最为显著,这使改进模型具有更准确的目标定位和更高的目标检测率。在 YOLOv5s 检测结果极高的前提下,改进模型仍能有全方位的提升,其中精确率提升0.5%,召回率提升0.6%,$\text{mAP}_{@0.5}$ 提升0.1%,$\text{mAP}_{@0.5:0.95}$提升1.5%。由此可见,将改进模型应用到交通汽车检测具有可靠性。

表 2 不同算法在 VOC 和 UA-DETRAC 数据集上的实验结果 单位:%

模型	PASCAL VOC				UA-DETRAC			
	精确率	召回率	$\text{mAP}_{@0.5}$	$\text{mAP}_{@0.5:0.95}$	精确率	召回率	$\text{mAP}_{@0.5}$	$\text{mAP}_{@0.5:0.95}$
YOLOv5s	80.4	78.6	83.2	59.8	97.1	96.4	98.8	85.5
OUR	82.5	78.6	84.8	64.7	97.6	97.0	98.9	87.0

分别在雨天、夜晚、晴天场景下,使用改进模型和 YOLOv5s 检测结果见图 9。雨天对照原图可以看到,YOLOv5s 出现错误检测和检测遗漏的问题。YOLOv5s 将公交车尾部检测为汽车;图的左上角有 4 辆汽车,只

检测出 3 辆。在夜晚场景下,同样出现了检测遗漏的问题。在晴天,改进模型具有更优越的检测结果。综合比较,改进模型检测的边界框更合适,检测结果更精准,且仍能检测出被遮挡的车辆,未曾出现错误检测和检测遗漏的问题。

(a) 原图　　(b) YOLOv5s改进　　(c) YOLOv5s原版

图 9　改进模型与原始模型的交通车辆检测对比

3.3.2　局部自注意力分析

在前面提及 YOLOv5s 的特征串联阶段是至关重要的,因此在该阶段加入了 SE、CBAM 和 LFSa(ELFSa 的原始模块)来进行对比实验。

从表 3 可以看到 SE 模块加入后模型性能不升反降,CBAM 提升幅度小。LFSa 在该阶段表现良好,但与 ELFSa 相比仍有改进空间。ELFSa 在高级语义信息的提取更加强大,使精确率提升了1.6%,$mAP_{@0.5}$提升了0.9%,$mAP_{@0.5:0.95}$提升了3.0%。

表 3　在特征串联阶段添加不同注意力对比结果　　单位:%

模型	PASCAL VOC			
	精确率	召回率	$mAP_{@0.5}$	$mAP_{@0.5:0.95}$
YOLOv5s	80.4	78.6	83.2	59.8
YOLOv5s+SE	80.2	78.3	82.8	58.6
YOLOv5s+CBAM	80.2	78.2	83.5	59.7
YOLOv5s+LFSa	80.2	78.8	83.3	61.5
YOLOv5s+ ELFSa	82.0	78.5	84.1	62.8

3.3.3　解耦检测头对比分析

在实验过程中加入目前相对熟知的检测头,分别是 YOLOX 的解耦头、ASFF[20] 检测头和 TSCODE(原版)。通过这 3 个检测头的实验结果充分体现了 S-TSCODE 的优越性。

由表 4 可知,S-TSCODE 相对比较于 TSCODE,$mAP_{@0.5}$ 提升 0.9%、$mAP_{@0.5:0.95}$ 提升 1.1%。与 YOLOv5s 比较,$mAP_{@0.5}$ 提升1.1%、$mAP_{@0.5:0.95}$ 提升 3.5%。S-TSCODE 具有更简洁的 SCE 和 DPE 模块,相比原始模块结构更简洁,但是拥有更优秀的检测结果。这证明了 S-TSCODE 能够更好的使分类和定位两个子任务获取其所需要的精确上下文信息。

表 4　不同检测头的对比实验结果　　单位:%

模型	PASCAL VOC			
	精确率	召回率	$mAP_{@0.5}$	$mAP_{@0.5:0.95}$
YOLOv5s	80.4	78.6	83.2	59.8
YOLOv5s+DecoulpedHead	81.5	78.0	83.4	61.5
YOLOv5s+ASFF	80.6	78.2	83.4	60.4
YOLOv5s+TSCODE	82.7	77.7	83.4	62.2
YOLOv5s+ S-TSCODE	82.0	78.7	84.3	63.3

3.3.4　消融实验

为验证改进模型添加的各模块的有效性,通过控制变量的方法在 PASCAL VOC 和 UA-DETRAC 数据集上进行多组实验来验证。

由表5、6 可以得出,添加 GSConv 主要起到降低模型计算复杂度的作用,对模型仅有小幅度提升。加入 ELFSa 和 S-TSCODE 后,模型的 4 个评价指标都有较大幅度提升。通过 ELFSa 提高模型的特征表示能力并具有更优秀的泛化能力,使用 S-TSCODE 对特征进行解耦,给予分类和定位任务精确的上下文信息,完美平衡这两个极其重要的子任务。在 UA-DETRAC 数据集上,原版模型具有极高的检测效率的前提下,通过 3 个模块的改进模型,在精度上仍然有较高的提升。

表 5 改进模型在 PASCAL VOC 上进行消融实验

Num	GSConv	ELFSa	S-TSCODE	精确率/%	召回率/%	$mAP_{@0.5}$/%	$mAP_{@0.5:0.95}$/%
0				80.4	78.6	83.2	59.8
1	√			81.3	78.3	83.7	61.4
2	√	√		82.0	78.9	84.4	63.0
3	√	√	√	82.5	78.6	84.8	64.7

表 6 改进模型在 UA-DETRAC 上进行消融实验

Num	GSConv	ELFSa	S-TSCODE	精确率/%	召回率/%	$mAP_{@0.5}$/%	$mAP_{@0.5:0.95}$/%
0				97.1	96.4	98.8	85.5
1	√			97.8	96.8	98.9	86.6
2	√	√		97.7	97.0	98.9	86.4
3	√	√	√	97.6	97.0	98.9	87.0

4 结束语

针对天气、场景、车流量过大导致车辆重叠遮挡等问题,改进算法在不同时间段、不同地点、不同天气状况下仍能发挥优秀的检测效率,且能避免错检和漏检,同时能检测到被遮挡的车辆:(1)在特征串联阶段加入 ELFSa 提高了特征表达能力,使模型重点关注检测目标;(2)使用 S-TSCODE 替换 YOLOv5s 检测头,解决了模型耦合性较高的问题,完美平衡分类和定位任务;(3)引用 GSConv 替换 3×3 尺寸或更大尺寸的卷积核,降低模型的计算复杂度,保证模型的推理速度。

在 VOC 数据集上,改进模型相较于原始模型具有更优秀的检测结果,精确率提升2.1%,$mAP_{@0.5}$ 提升1.6%,$mAP_{@0.5:0.95}$提升4.9%;在 UA-DETRAC 数据集上,P 提升0.5%,召回率提升0.6%,$mAP_{@0.5}$ 提升0.1%,$mAP_{@0.5:0.95}$提升1.5%。实验数据证明,改进算法具有良好的泛化能力和鲁棒性,但是 ELFSa 的注意力仍然有分散到其他目标,而且改进算法的车辆检测只有 4 种类型车辆。因此今后将继续研究更加集中的注意力模块,并将改进算法应用到车辆类别更广泛的数据集。

致谢:感谢合肥工业大学智能互联系统安徽省实验室开放基金(PA2021AKSK0107)对本文的资助

参考文献:

[1] 陆化普,李瑞敏.城市智能交通系统的发展现状与趋势[J].工程研究-跨学科视野中的工程,2014,6(1):6-19.

[2] 张富凯,杨峰,李策.基于改进 YOLOv3 的快速车辆检测方法[J].计算机工程与应用,2019,55(2):12-20.

[3] Cheng H Y,Weng C C,Chen Y Y. Vehicle detection in aerial surveillance using dynamic Bayesian networks[J]. IEEE transactions on image processing,2011,21(4):2152-2159.

[4] Zhang J,Guo X,Zhang C,et al. A vehicle detection and shadow elimination method based on greyscale information,edge information,and prior knowledge [J]. Computers & Electrical Engineering,2021,94:107366.

[5] Girshick R. Fast r-cnn [C]. Proceedings of the IEEE international conference on computer vision. 2015:1440-1448.

[6] Cai Z,Vasconcelos N. Cascade r-cnn: Delving into high quality object detection [C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:6154-6162.

[7] Dosovitskiy A,Beyer L,Kolesnikov A,et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2010.11929,2020.

[8] Li W,Huang L. YOLOSA: Object detection based on 2D local feature superimposed self-attention [J]. arXiv preprint arXiv:2206.11825,2022.

[9] Zhuang J,Qin Z,Yu H,et al. Task-Specific Context Decoupling for ObjectDetection[J]. arXiv preprint arXiv:2303.01047,2023.

[10] 毕鹏程,罗健欣,陈卫卫.轻量化卷积神经网络技术研究[J].计算机工程与应用,2019,55(16):25-35.

[11] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:1251-1258.

[12] Li H,Li J,Wei H,et al. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles[J]. arXiv preprint arXiv:2206.02424,2022.

[13] Krizhevsky A,Sutskever I,Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems,2012,25.

[14] Liu S,Qi L,Qin H,et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:8759-8768.

[15] Hu J,Shen L,Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:7132-7141.

[16] Woo S,Park J,Lee J Y,et al. Cbam:Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV). 2018:3-19.

[17] Ge Z,Liu S,Wang F,et al. Yolox:Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430,2021.

[18] Everingham M,Gool L V,Williams C K I,et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision,2010,88(2):303-338.

[19] Wen L,Du D,Cai Z,et al. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking[J]. ComputerVision and Image Understanding,2020,193:102907.

[20] Liu S,Huang D,Wang Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv:1911.09516,2019.

Traffic Vehicle Detection based on Self-Attention Combined with Context Decoupling

SUN Guangling[1,2], ZHOU Yunlong[1]

(1. School of Electronic and Information Engineering,Anhui Jianzhu University,Hefei 230601,China;2. Anhui Laboratory of Intelligent Interconnection System,Hefei University of Technology,Hefei 230009,China)

Abstract:To address the challenges arising from factors like traffic flow,time,place,and weather on traffic vehicle detection,A novel algorithm,which builds upon the YOLOv5s model,has been introduced. This enhanced model demonstrates adaptability across diverse scenarios of transportation. The improvements are as follows: the incorporation of an efficient 2D local feature superimposed self-attention (ELFSa) during the feature series stage,aiming to enrich the model's object perception capabilities;replace the detection head of YOLOv5s with a simple task-specific context decoupling (STSCODE) to achieve a perfect balance between classification and localization subtasks Balance,thereby improving the convergence of the model;To mitigate computational complexity,certain convolutions within the model with dimensions of 3×3 or larger are substituted with GSConv. The experimental findings demonstrate that the enhanced YOLOv5s model exhibits improvements across all aspects,particularly in terms of $mAP_{@0.5}$ is 98.9%,and $mAP_{@0.5:0.95}$ is 87.0%,which are respectively increased by 0.1% and 1.5%. Addressing a wide range of intricate traffic scenarios,the suggested methodology enhanced the performance and robustness of vehicle detection.

Keywords:object detection;self-attention;decoupled head;lightweight convolution;YOLOv5s