DOI:10.20079/j.issn.1001-893x.240812001

基于金字塔增强与跨语义交互的轻量图像目标检测网络*

陆 蔚

(江苏信息职业技术学院 物联网工程学院,江苏 无锡 214153)

摘 要:近年来,轻量化目标检测领域取得了显著进展。然而,现有主流方法缺乏多尺度语义信息的提取,且忽略了深层语义特征与浅层细节特征之间的关系。针对上述缺陷,提出了金字塔池化多尺度增强网络(Pyramid Pooling Enhanced Multi-scale Network,PPMENet),通过设计一个高效金字塔池化模块(Efficient Pyramid Pooling Block,EPPB)来提取多尺度深层语义信息,以加强模型的特征表达能力。另一方面,设计了跨语义交互注意力模块(Cross Semantic Level Interaction Attention Module,CSIAM)以增强不同语义特征之间的联系。MS COCO 2017 测试集的实验结果表明,PPMENet 取得了 28.0% 平均精度,模型大小仅有 2.16×10^6,GFLOPs 为 0.97,并获得了 218 frame/s 的推理速度。与其他方法相比,PPMENet 在精度和执行效率间取得了较好的平衡。

关键词:实时图像目标检测;轻量级网络;多尺度特征提取;注意力机制;特征融合

开放科学(资源服务)标识码(OSID):

中图分类号:TN957.52 **文献标志码**:A **文章编号**:1001-893X(2025)11-1798-08

Pyramid-enhancedand Cross-semantic Interaction Network for Lightweight Real-time Image Object Detection

LU Wei

(School of Internet of Things Engineering, Jiangsu Vocational College of Information Technology, Wuxi 214153, China)

Abstract: Recently, with the development of deep learning, the field of lightweight object detection has witnessed significant progress. However, mainstream lightweight detectors ignore the extraction of multi-scale semantic information. In addition, these approaches ignore the relationship between deep semantic features and shallow detail features. To relieve above shortcomings, a Pyramid Pooling Enhanced Multi-scale Network (PPMENet) is proposed and an Efficient Pyramid Pooling Block (EPPB) is designed to extract multi-scale deep semantic information, strengthening the feature expression ability of the model. On the other hand, a Cross Semantic Level Interaction Attention Module (CSIAM) is designed to enhance information interaction between features at different semantic levels. Experimental results on the MS COCO 2017 test set show that PPMENet gets 28.0% average precision, only with 2.16×10^6 model size and 0.97GFLOPs, and achieves inference speed of 218 frame/s. Compared with other methods, PPMENet realizes a good balance between detection accuracy and model execution efficiency.

Key words: real-time image object detection; lightweight network; multi-scale feature extraction; attention mechanism; feature fusion

* **收稿日期**:2024-08-12;**修回日期**:2024-10-13
基金项目:江苏省高校“青蓝工程”优秀教学团队资助(苏教师函〔2021〕11 号);物联网应用技术职业教育教师教学创新团队资助(苏教办师函〔2021〕23 号)
通信作者:陆蔚 Email:4137701@qq.com

0　引　言

目标检测作为计算机视觉中一项富有挑战的任务,旨在定位图像中感兴趣的目标,预测其类别和边框坐标,在辅助驾驶[1]、智慧交通[2]等视觉任务中扮演着重要的角色。近年来,随着在边缘设备上运行目标检测模型需求的提升,轻量化实时目标检测算法得到了研究人员的广泛关注,一系列轻量化实时目标检测器[3-9]被提出,尽管通过削减模型的卷积层数或使用新的卷积算子有效降低了模型的参数量和计算量,但是忽略了对图像中的多尺度语义特征的提取,这可能使得轻量化检测器的性能受限。

除了压缩检测器的骨干网,另一些方法则致力于改进轻量化检测器颈部网络的特征融合方式。尽管文献[7-11]模型的特征融合方式消耗较少的计算量和参数量,但是简单的元素加融合忽略了不同语义级别的特征间的关系,这可能使得检测器的特征表达能力受限。

针对上述问题,本文提出了金字塔池化多尺度增强网络(Pyramid Pooling Enhanced Multi-scale Network,PPMENet)以实现轻量化实时目标检测。与其他在骨干网中仅使用单个尺度的卷积核来提取目标特征的检测器[3-6]不同,本文提出了高效金字塔池化模块(Efficient Pyramid Pooling Block,EPPB),使用多个不同尺度的池化核,以较小的计算量来捕获多个尺度的金字塔特征,并将其以自顶向下的方式逐级融合以编码多尺度特征,有效增强模型容量和多尺度特征建模能力。在颈部网络中,与其他使用简单的元素加操作[7-11]来融合不同分辨率特征的检测器不同,本文提出了跨语义交互注意力模块(Cross Semantic Level Interaction Attention Module,CSIAM),以自顶而下的方式,以包含丰富上下文的低分辨率深层特征作为引导,分别利用空间注意力和通道注意力机制来促进跨尺度特征间的信息交流,加强高分辨率浅层细节特征的语义信息,使得不同分辨率的特征能够得到充分利用和融合。

1　PPMENet 检测模型

1.1　PPMENet 网络架构

PPMENet 的整体结构如图 1 所示,由骨干网、颈部网络和检测头 3 个部分组成。

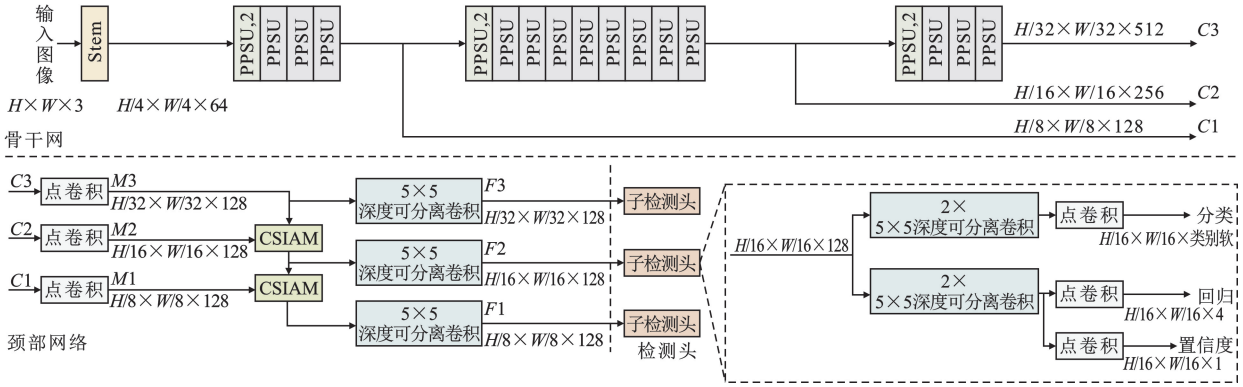

图 1　PPMENet 的整体结构

PPMENet 由起始层[5](Stem)与多个 PPSU 模块堆叠而成,拥有 3 个阶段,每个阶段拥有 4、8、4 个骨干网构造块 PPSU,逐步生成分辨率为原始输入图像 1/4、1/8、1/16、1/32 的特征图。与其他检测器类似[7,11],为了能有效检测不同尺度的目标,骨干网 3 个阶段所生成的不同的分辨率输出 $C1$、$C2$、$C3$,维度分别为 $H/8\times W/8\times 128$、$H/16\times W/16\times 256$、$H/32\times W/32\times 512$。这 3 个输出将接着被送入颈部网络,以对这些跨分辨率特征进行融合加权。其中 PPMENet 骨干网的核心构成模块 PPSU 的结构如图 2 所示,图 2(a)为 PPSU 的步长为 1 的版本。输入特征首先经过通道分裂操作[5],其中一个低维特征维持不变,另一个低维特征分别经过输入通道与输出通道均为 $C/2$ 的点卷积,5×5 逐深度卷积以提取局部特征。接着使用 EPPB 模块,利用模块内部的多个不同尺寸与感受野的池化核来提取多尺度特征,并将这些特征以自顶向下的方式进行融合以编码多尺度上下文特征,编码后的特征经过输入通道与输出通道均为 $C/2$ 的点卷积后,与另一个分支的低维特征进行拼接[5],将这两个特征沿着通道维度进行拼接合并,生成维度为 $H\times W\times C$ 的通道拼接特征。接着使用通道混洗操作[5]来加强不同通道之间的联系。图 2(b)为 PPSU 的带步长版本,与图 2(a)类似,不同之处在于两条分支均使用了步长为 2

的 5×5 逐深度卷积,将特征维度由 $H\times W\times C/2$ 降至 $H/2\times W/2\times C/2$。此外,在进行通道拼接操作前,两个分支均使用输入通道为 $C/2$,输出通道为 C 的点卷积,以增加对应分支特征的通道维度。

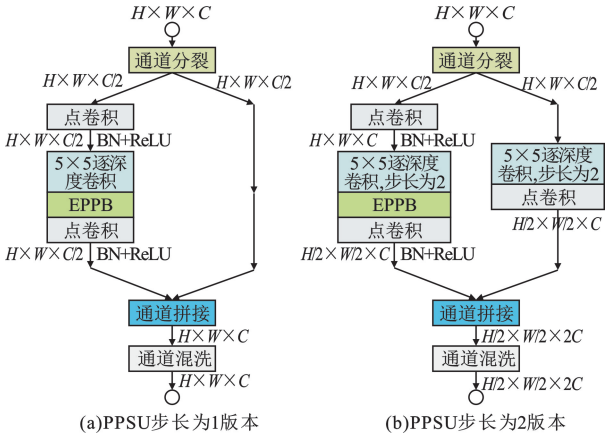

图 2 PPSU 两个版本的详细结构

PPMENet 颈部网络与 FPN 类似,自顶向下将跨尺度特征进行融合,但与基于 FPN 及其衍生模型中的元素加融合操作[7-11]不同,本文设计了 CSIAM 模块来挖掘低分辨率深层语义与高分辨率浅层细节间的关系。具体来说,骨干网的输出 $\{C1,C2,C3\}$ 作为颈部网络的输入,与特征金字塔网络(Feature Pyramid Network,FPN)相同,为了将这 3 个特征压缩至同一通道维度,颈部网络使用了 3 个不同的 1×1 点卷积,卷积的输入通道分别为 C、$2C$、$4C$,输出通道则均为 C,以此生成 3 个通道相同的多分辨率特征 $M1$、$M2$、$M3$,维度分别为 $H/8\times W/8\times C$、$H/16\times W/16\times C$、$H/32\times W/32\times C$,其中 C 为 128。接着特征 $M1$、$M2$、$M3$ 自顶向下进行融合,邻接的特征将分别被送入两个 CSIAM,通过深层语义引导浅层细节,利用 CSIAM 内部的空间注意力和通道注意力,将不同语义的跨分辨率特征进行加权融合,生成颈部网络的输出 $F1$、$F2$、$F3$,维度分别为 $H/8\times W/8\times 128$、$H/16\times W/16\times 128$、$H/32\times W/32\times 128$。

检测头包含 3 个结构相同的子网络,每个子网络分别由 2 个 5×5 深度可分离卷积和 3 个 1×1 点卷积构成。颈部网络的 3 个输出 $F1$、$F2$、$F3$ 分别被送入不同的子检测头网络,进行类别、边框和交并比的预测,最后与其他常见的检测器相同[6-10],使用非极大值抑制[12]来消除重叠的冗余预测框,生成最终的检测结果。

1.2 高效金字塔池化模块

高效金字塔池化模块具体结构如图 3 所示,给定输入特征 $\boldsymbol{F}_{\text{in}}\in\mathbb{R}^{H\times W\times C}$,其中 H、W、C 分别代表输入特征的长、宽与通道数量。为了减少计算量,$\boldsymbol{F}_{\text{in}}$ 首先经过一个点卷积来减少特征通道数量,获得 $\boldsymbol{F}'_{\text{in}}\in\mathbb{R}^{H\times W\times C/4}$。接着,为了获得层次化的多尺度特征表示,$\boldsymbol{F}'_{\text{in}}$ 并行地经过 3 个池化核尺寸各异的金字塔池化操作 $\text{Pool}_i(\cdot)$,生成分辨率不同的多尺度特征 $\boldsymbol{F}^i_{\text{pool}}$:

$$\boldsymbol{F}^i_{\text{pool}}=\text{Pool}_i(\boldsymbol{F}'_{\text{in}}),i\in\{1,2,3,4\} \tag{1}$$

式中:$i=1$ 时池化核尺寸为 $H\times W$,以捕获特征的全局信息[13];i 由 2~4 对应的池化核尺寸分别为 9×9、7×7、5×5,在保证特征中心对称与池化核尺寸小于输入特征分辨率的前提下,获得多个感受野的特征表示。输出特征 $\boldsymbol{F}^i_{\text{pool}}$ 分辨率为输入特征 $\boldsymbol{F}_{\text{in}}$ 的 $1/H\times W$、1/16、1/9 以及 1/4。

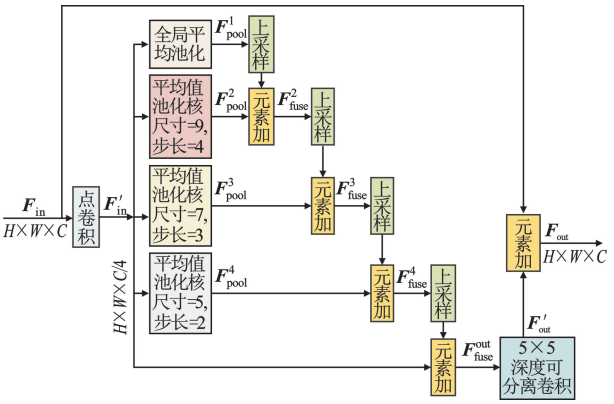

图 3 EPPB 详细结构

当生成多尺度的层次化特征 $\boldsymbol{F}^i_{\text{pool}}$ 后,EPPB 将会自顶向下地对这些多尺度特征进行逐级融合,首先将低分辨率特征 $\boldsymbol{F}^1_{\text{pool}}$ 通过最邻近插值上采样[8]操作 $\text{Upsample}(\cdot)$,接着分辨率放大后的特征与邻接的高分辨特征 $\boldsymbol{F}^{i+1}_{\text{pool}}$ 相加,生成融合后的特征 $\boldsymbol{F}^i_{\text{fuse}}$,并向下逐级融合:

$$\boldsymbol{F}^i_{\text{fuse}}=\boldsymbol{F}^i_{\text{pool}},i=1 \tag{2}$$

$$\boldsymbol{F}^{i+1}_{\text{fuse}}=\text{Upsample}(\boldsymbol{F}^i_{\text{fuse}})+\boldsymbol{F}^{i+1}_{\text{pool}},i\in\{1,2,3\} \tag{3}$$

接着,对融合后的底层特征 $\boldsymbol{F}^4_{\text{fuse}}$ 进行最邻近插值上采样并与 $\boldsymbol{F}'_{\text{in}}$ 相加,生成 $\boldsymbol{F}^{\text{out}}_{\text{fuse}}\in\mathbb{R}^{H\times W\times C/4}$。为了恢复特征的通道维度并减少上采样的混叠效应[8],$\boldsymbol{F}^{\text{out}}_{\text{fuse}}$ 经过一个 5×5 深度可分离卷积[4] $\text{DW}_{5\times5}(\cdot)$,生成 $\boldsymbol{F}'_{\text{out}}\in\mathbb{R}^{H\times W\times C}$:

$$\boldsymbol{F}'_{\text{out}}=\text{DW}_{5\times5}(\text{Upsample}(\boldsymbol{F}^4_{\text{fuse}})+\boldsymbol{F}'_{\text{in}}) \tag{4}$$

最后,为了在输出与输入间建立残差连接,维度恢复的特征 $\boldsymbol{F}'_{\text{out}}$ 与输入特征 $\boldsymbol{F}_{\text{in}}$ 进行相加,生成 EPPB 的最终输出 $\boldsymbol{F}_{\text{out}}\in\mathbb{R}^{H\times W\times C}$:

$$\boldsymbol{F}_{\text{out}}=\boldsymbol{F}'_{\text{out}}+\boldsymbol{F}_{\text{in}} \tag{5}$$

EPPB 拥有较低的计算成本(因篇幅所限,EPPB 计算量推导公式请扫描本文 OSID 码,在“本文开放的科学数据与内容”中查看),其计算量 $\text{Cost}_{\text{EPPB}}\approx 11.9HWC+0.25HWC^2$,与 3×3 深度可分离卷积的计算量 $\text{Cost}_{\text{DW3×3}}=9HWC+HWC^2$ 相比,由于通常 $C>>11.9$,因此 $\dfrac{\text{Cost}_{\text{EPPB}}}{\text{Cost}_{\text{DW3×3}}}=\dfrac{11.9+0.25C}{9+C}\approx 0.25$,EPPB 的计算量仅为 3×3 深度可分离卷积的 0.25,因此将 EPPB 模块插入网络不会带来大幅计算量的提升。

1.3　跨语义交互注意力模块

如图 4 所示,本文提出了跨语义交互注意力模块,以在检测器的颈部网络捕获深层语义与浅层细节之间的关系。更具体来说,CSIAM 主要由两个模块组成,空间注意力模块(CSIAM-SA)和通道注意力模块(CSIAM-CA)。

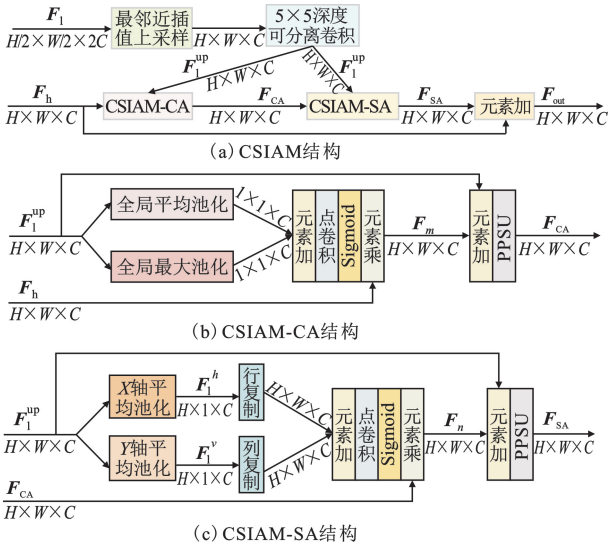

图 4　CSIAM 的详细结构

如图 4(a)所示,给定输入特征 $\boldsymbol{F}_{\text{h}}\in\mathbb{R}^{H\times W\times C}$ 为高分辨率细节特征,$\boldsymbol{F}_{\text{l}}\in\mathbb{R}^{H/2\times W/2\times C}$ 为低分辨率语义特征,其中低分辨率语义特征 $\boldsymbol{F}_{\text{l}}$ 经过最邻近插值上采样操作 Upsample(·),使得其分辨率与 $\boldsymbol{F}_{\text{h}}$ 相同。接着与文献[8]类似,最邻近插值上采样后的特征经过一个 5×5 深度可分离卷积[4] $\text{DW}_{5\times5}(\cdot)$,以恢复采样后丢失的部分细节,生成 $\boldsymbol{F}_{\text{l}}^{\text{up}}\in\mathbb{R}^{H\times W\times C}$:

$$\boldsymbol{F}_{\text{l}}^{\text{up}}=\text{DW}_{5\times5}(\text{Upsample}(\boldsymbol{F}_{\text{l}})) \tag{6}$$

在图 4(b)所示通道注意力 CSIAM -CA 模块中,输入特征 $\boldsymbol{F}_{\text{l}}^{\text{up}}$ 首先分别经过全局平均池化操作 GAP(·)和全局最大值池化操作 GMP(·),通过这两个全局池化操作来编码特征 $\boldsymbol{F}_{\text{l}}^{\text{up}}$ 的全局上下文信息[13]。接着编码后的全局上下文信息被依次送入元素加操作、点卷积操作 $\text{Conv}_{1\times1}(\cdot)$ 和 Sigmoid 激活[13]操作 $\sigma(\cdot)$,生成归一化的通道注意力权重图,然后通道注意力图与高分辨率细节特征 $\boldsymbol{F}_{\text{h}}$ 进行逐元素乘操作,生成跨语义通道交互特征 $\boldsymbol{F}_m\in\mathbb{R}^{H\times W\times C}$:

$$\boldsymbol{F}_m=\sigma(\text{Conv}_{1\times1}(\text{GAP}(\boldsymbol{F}_{\text{l}}^{\text{up}})+\text{GMP}(\boldsymbol{F}_{\text{l}}^{\text{up}})))\odot\boldsymbol{F}_{\text{h}} \tag{7}$$

式中:⊙为逐元素乘法操作。接下来,加权后的特征 $\boldsymbol{F}_m$ 与输入特征 $\boldsymbol{F}_{\text{l}}^{\text{up}}$ 进行相加以实现残差连接,相加后的特征被送入图 2(a)所示的 PPSU 模块 PPSU(·)以进行局部特征细化,生成通道注意力 CSIAM-CA 的输出 $\boldsymbol{F}_{\text{CA}}\in\mathbb{R}^{H\times W\times C}$:

$$\boldsymbol{F}_{\text{CA}}=\text{PPSU}(\text{F}_m+\boldsymbol{F}_{\text{l}}^{\text{up}}) \tag{8}$$

为了高效捕获跨语义空间关系,如图 4(c)所示,CSIAM -SA 模块从 X 轴和 Y 轴两个方向来捕获特征的空间关系。输入特征 $\boldsymbol{F}_{\text{l}}^{\text{up}}$ 首先经过 X 轴平均池化操作和 Y 轴平均池化操作,获得包含水平方向与垂直方向的全局特征 $\boldsymbol{F}_{\text{l}}^{h}\in\mathbb{R}^{H\times1\times C}$ 和 $\boldsymbol{F}_{\text{l}}^{v}\in\mathbb{R}^{1\times W\times C}$:

$$\begin{cases}\boldsymbol{F}_{\text{l}}^{h}(y=\dfrac{1}{W})\sum_{x=1}^{W}\boldsymbol{F}_{\text{l}}^{\text{up}}(x,y)\\ \boldsymbol{F}_{\text{l}}^{v}(x)=\dfrac{1}{H}\sum_{y=1}^{H}\boldsymbol{F}_{\text{l}}^{\text{up}}(x,y)\end{cases} \tag{9}$$

式中:H、W 为特征图的高度与宽度;x、y 分别为特征图上任意一点的横坐标与纵坐标。接着将 $\boldsymbol{F}_{\text{l}}^{h}\in\mathbb{R}^{H\times1\times C}$ 沿着空间维度的 X 轴逐个复制 W 份,生成行复制后的特征,维度为 $H\times W\times C$。同理,对列特征 $\boldsymbol{F}_{\text{l}}^{v}\in\mathbb{R}^{1\times W\times C}$ 沿着空间维度的 Y 轴逐一复制 H 份,生成列复制后的特征,维度与 $\boldsymbol{F}_{\text{l}}^{h}$ 行复制后的特征相同,也为 $H\times W\times C$,使得 $\boldsymbol{F}_{\text{l}}^{h}$ 与 $\boldsymbol{F}_{\text{l}}^{v}$ 复制后的特征维度相同,方便两者进行接下来的相加操作。接着,相加后的特征图分别经过点卷积操作 $\text{Conv}_{1\times1}(\cdot)$ 和 Sigmoid 激活操作 $\sigma(\cdot)$,生成归一化的空间注意力权重图,并与输入特征 $\boldsymbol{F}_{\text{CA}}$ 相乘,得到 $\boldsymbol{F}_n\in\mathbb{R}^{H\times W\times C}$:

$$\boldsymbol{F}_n=(\sigma(\text{Conv}_{1\times1}(\boldsymbol{F}_{\text{l}}^{h}+\boldsymbol{F}_{\text{l}}^{v})))\odot\boldsymbol{F}_{\text{CA}} \tag{10}$$

接下来,加权后的特征 $\boldsymbol{F}_n\in\mathbb{R}^{H\times W\times C}$ 与输入特征 $\boldsymbol{F}_{\text{l}}^{\text{up}}$ 进行相加,并送入 PPSU 模块 PPSU(·),生成空间注意力 SA 模块的输出 $\boldsymbol{F}_{\text{SA}}\in\mathbb{R}^{H\times W\times C}$:

$$\boldsymbol{F}_{\text{SA}}=\text{PPSU}(\boldsymbol{F}_n+\boldsymbol{F}_{\text{l}}^{\text{up}}) \tag{11}$$

CSIAM-CA 与 CSIAM-SA 拥有较低的计算量(因篇幅所限,CAIAM-CA 和 CSIAM-SA 计算量推导

公式请扫描本文 OSID 码,在“本文开放的科学数据与内容”中查看),两者计算量分别为

$$\mathrm{Cost}_{\mathrm{csiam\text{-}ca}}=28.4HWC+0.75HWC^2+C^2+C \quad (12)$$

$$\mathrm{Cost}_{\mathrm{csiam\text{-}sa}}=29.4HWC+1.25HWC^2 \quad (13)$$

由上述计算量分析可知,CSIAM-CA 和 CSIAM-SA 的计算量与 5×5 深度可分离卷积的计算量 $\mathrm{Cost}_{\mathrm{DW5\times5}}=25HWC+HWC^2$ 相近,将这两个模块加入网络中并不会带来大幅计算量提升。

2 实验与分析

2.1 数据集和评价指标

为了验证 PPMENet 的有效性,本文在目标检测权威数据集 MS COCO 2017[14] 和 Pascal VOC 数据集[15] 上进行了系统性对比实验和消融实验。MS COCO 2017 数据集中的目标检测任务拥有 80 个类别,包含训练集、验证集和测试集,这 3 个子集分别拥有 118 287、5 000、40 670 张图片。Pascal VOC 数据集由 VOC 2007 和 VOC 2012 构成,包含 20 个目标类别,训练验证集总共包含 16 651 张图片,测试集共包含 4 952 张图片。本文提出的模型在上述两个数据集的训练集上进行训练,系统性对比实验在测试集上进行,消融实验在验证集上进行。

为了与其他轻量化检测器进行公平比较,本文采用平均精度(Average Precision,AP)来衡量检测器的性能,通过计算不同阈值下的精度(Precision)和召回率(Recall)来评估一个目标检测模型的精度。其中,类别 i 的 AP_i 为精度(P)-召回率(R)曲线下的面积,即 $\mathrm{AP}_i=\int_0^1 P(R)\mathrm{d}R$;mAP 则是对所有类别的 AP_i 求均值 $\mathrm{mAP}=\frac{\sum_{i=1}^{N}\mathrm{AP}_i}{N}$,$N$ 代表数据集的类别总数。在 Pascal VOC 数据集中,类别 N 为 20,mAP 的交并比阈值为 0.5。而在 MS COCO2017 数据集中,类别 N 为 80,评价指标主要有 AP、AP^{50} 和 AP^{75},其中 AP 指标通过交并比阈值从[0.5,0.95]、步长为 0.05 这 10 个 mAP 的值计算均值而获得。AP^{50} 与 AP^{75} 则代表交并比阈值为 0.5 与 0.75 时所计算得到的平均精度。此外,本文还从模型参数量、模型浮点计算量 FLOPs 以及推理速度这 3 个方面来评估 PPMENet 的执行效率。

2.2 实验细节和训练参数对比

PPMENet 模型训练和测试的机器为一台搭载单块 RTX 2080Ti GPU 的工作站,CPU 为 Intel i7-8565U,内存为 32 GB,硬盘为 2 TB,CUDA 版本为 11.1,深度学习框架为 Pytorch1.8,操作系统为 Ubuntu18.04,Python 版本为 3.7。PPMENet 在使用混合精度训练技术[16] 后,在 MS COCO 2017 数据集的批图像数量为 92,使用的优化器为 SGD,训练周期为 320,包含 5 个热身周期;学习率从 0 逐渐增加到 0.015,接着学习率与文献[17]相同,逐步衰减至 0.000 75;动量和权重衰减分别设置为 0.9 和 5×10^{-4}。为了和主流方法进行公平对比,PPMENet 仅使用 SSD 中的基本数据增强策略[18]。模型训练和测试的输入图像分辨率为 320 pixel×320 pixel,与 YOLO 系列类似[19-20],进行推理速度测速前,使用 TensorRT-8.0 推理框架对模型进行加速。模型在 Pascal VOC 数据集上采用的训练和数据增强策略与其在 MS COCO 上相同,预先在 MS COCO 上进行 320 轮训练,接着在 Pascal VOC 训练集上使用 0.001 的学习率进行 12 轮微调。

PPMENet 使用的损失函数与文献[17]相同。此外,本文还提供了 PPMENet 和主流方法在 MS COCO 数据集中训练参数、损失函数,以及数据增强手段的差异,以证明本文方法和主流方法的对比是在一个较为公平的基准下进行的。因篇幅受限,相关内容请扫描本文 OSID 码,在“本文开放的科学数据与内容”中查看。

2.3 系统对比试验

2.3.1 MS COCO 2017 数据集上的系统对比实验

为了评估本文所提出 PPMENet,本节选取了一系列先进轻量级实时目标检测器,包括 Tiny-YOLO 系列[6,17,19]、Tiny-DSOS[7] 以及 SSDLite[3,21] 系列,在 MS COCO 2017 测试集上与 PPMENet 进行了系统性的对比实验。如表 1 所示,PPMENet 获得了 28.0% AP,仅有 9.7×10^8 的浮点计算量和 2.16×10^6 的模型参数,并达到 218 frame/s 的实时推理速度,证明了 PPMENet 兼顾检测性能和执行效率,在两者之间取得了令人满意的平衡。与表 1 中所示的先进轻量级实时检测器相比,PPMENet 在性能方面处于领先地位,其 AP 仅低于文献[19],但是文献[19]的参数量和计算量分别为 PPMENet 的 3 倍和 14 倍。在模型大小方面,PPMENet 的参数量相较于 Tiny-YOLOV4 和 MobileViT-XS-SSDLite 分别降低了 3.94×10^6 和 5.4×10^5,但由于 PPMENet 在骨干网中使用了 EPPB 加强了多尺度特征的提取,因此相比这些骨干网中仅使用单个尺度卷积核提取特征的检测器,本文方法拥有着更高的性能。虽然 Nano-

YOLOX 的参数量仅为 PPMENet 的 42%,但是其 AP 也比 PPMENet 大幅降低了 2.7%。在浮点计算量方面,PPMENet 拥有着较低的计算成本,分别比 Nano-YOLOX、Tiny-DSOD、CSL-YOLO 降低了 10%、15%、44% 的 GFLOPs。由于 PPMENet 在颈部网络使用了 CSIAM 加强了不同特征之间的关系,因此性能分别比上述 3 个基于 FPN 架构的检测模型高了 2.7% AP、4.8% AP、3.5% AP。虽然文献[21]的 FLOPs 仅为 PPMENet 的 82%,但是其 AP 也较本方法急剧降低了 5.9%。

表 1　PPMENet 在 MS COCO 2017 测试集上与其他轻量级检测器的性能对比

模型	FLOPs/10^9	参数量/10^6	AP/%	AP_{50}/%	AP_{75}/%	推理速度/(frame/s)
MobileNetV3-SSDLite [3]	**0.8**	4.3	22.1	—	—	—
MobileViT-XS-SSDLite [21]	—	2.7	24.8	—	—	76
SSD[18]	38.6	34.3	25.5	43.6	36.2	—
Tiny-DSOD[7]	1.12	1.15	23.2	40.4	22.8	105
Tiny-YOLOV4[6]	3.45	6.1	21.7	40.2	—	**371**
Nano YOLOX[17]	1.08	**0.91**	25.3	—	—	—
Tiny-YOLOX[17]	6.45	5.06	31.8	49.0	33.8	—
CSL-YOLO[11]	1.4	3.2	24.5	44.0	24.2	—
Tiny-YOLOV7[19]	13.8	6.3	**38.7**	**56.7**	**41.7**	273
PPMENet	0.97	2.16	28.0	44.4	29.4	218

图 5 展示了 PPMENet 和 Tiny-DSOD、SSD、Tiny-YOLOX 在 MS COCO 测试集的检测结果对比。图 5 第一列的图片中,人、餐桌、椅子尺度各异,可以看到,由于缺乏多尺度信息的提取,Tiny-DSOD 和 SSD 对餐桌、背包、碗,这 3 种大小不同的物体存在漏检的问题,性能最佳的 Tiny-YOLOX 同样漏检了图片右下角的背包,PPMENet 由于使用了 EPPB 来捕获多个尺度的信息,很好地检测出了第一列样本中的各个大小不同的物体。值得注意的是,由于 PPMENet 输入图像分辨率低于 Tiny-YOLOX,因此对餐桌的边框定位精度方面差于该方法。图 5 第二列左上角的飞机、第三列的足球和第四列右上角的伞,由于图片中存在遮挡或物体处于图像边界,目标仅露出部分区域,因此 SSD 和 Tiny-DSOD 对这 3 种情形存在漏检,而 PPMENet 由于加强了多尺度特征的提取,并使用 CSIAM 加强不同尺度特征的联系,对多尺度特征的利用更加充分,因此成功检测出了仅露出部分区域的飞机、足球和伞。但是由于池化金字塔可能会引入图片背景区域的无效特征,因此 PPMENet 对第二列左上角的飞机还存在了冗余的检测框。综合以上定性结果的比较可知,相比于 Tiny-YOLOX,PPMENet 存在更多的误检和定位不准的情形,而相较于 SSD 和 Tiny-DSOD,PPMENet 漏检率更低,这也与表 1 中的定量结果相符。此外,本文还提供了 PPMENet 对于存在光照、遮挡、背景干扰、小物体影响样本的可视化检测结果,因篇幅所限,具体内容请扫描本文 OSID 码,在“本文开放的科学数据与内容”中查看。

图 5　PPMENet 在 MS COCO 测试集中和其他方法的检测结果对比

2.3.2　Pascal VOC 数据集上的系统对比实验

如表 2 所示,PPMENet 在 Pascal VOC 2007 测试集上获得了 78.2% 的 mAP,模型大小仅有 2.14×10^6,GFLOPs 为 0.95。由于 Pascal VOC 类别数量为 MS COCO 的 1/4,且图片中目标实例不多,因此 PPMENet 的推理速度较 COCO 数据集增加了 15%,达到了 252 frame/s。与其他检测器相比,虽然 SSD 拥有更多的参数量和计算量,但是由于其忽略了多尺度特征的捕获以及没有对不同分辨率特征进行融合,因此其 mAP 比 PPMENet 降低了 1.7%。

虽然 Tiny-DSOS 使用了 FPN 来融合不同分辨率的特征,但是其缺乏骨干网中多尺度信息的提取,因此其 mAP 较 PPMENet 降低了 6. 1%。与 ThunderNet 相比,虽然 PPMENet 检测性能较之下降了 0. 4%,但是 PPMENet 拥有更低的浮点计算量与实时推理速度。此外,本文还提供了 PPMENet 在 Pascal VOC 测试集中可视化检测结果,因篇幅所限,具体内容请扫描本文 OSID 码,在“本文开放的科学数据与内容”中查看。

表 2 PPMENet 在 Pascal VOC 2007 测试集上与其他检测器的性能对比

模型	GFLOPs	参数量/10^6	mAP/%	速度/(frame/s)
SSD[18]	35. 3	26. 29	76. 5	46
Tiny-DSOD[7]	1. 10	—	72. 1	105
ThunderNet[4]	1. 30	—	78. 6	214
PPMENet	0. 95	2. 14	78. 2	252

2.4 消融实验

为了验证本文所提出的 EPPB 模块和 CSIAM 模块的有效性,本小节在 MS COCO 2017 验证集进行了详细的消融实验,以不包含 EPPB 和 CSIAM 的 PPMENet 作为基线模型,并逐步向其中添加了 EPPB、CSIAM-CA 和 CSIAM-SA,以验证每个模块的效果。如表 3 所示,由于忽略了多尺度上下文和跨语义特征间的关系,因此基线模型只获得了 25. 2% AP。当在骨干网的每一个 PPSU 模块中插入 EPPB 后,检测性能大幅提升了 1. 6% AP,达到了 26. 8% AP,这意味着加强多尺度信息提取对检测器的性能提升有很大的帮助,同时模型参数量和浮点计算量仅分别增加了 2.3×10^5 和 6×10^7。通过在检测器颈部建立跨语义特征的通道(CSIAM-CA)和空间(CSIAM -SA)之间的关系,在加入 EPPB 的基础上,在检测器的颈部网络中插入 CSIAM-CA 和 CSIAM -SA 分别获得了 0. 5% AP 和 0. 4% AP 的增益。而在使用 EPPB 的基础上,在颈部网络中插入 CSIAM-CA 和 CSIAM-SA 的组合,模型获得了 0. 9% AP 的增益,达到 27. 7% AP,说明两者联合使用能获得更佳的检测效果。上述实验表明了本文提出的各个模块都在不断提高整体网络的性能,证实了这些模块的有效性。此外,本文还提供了依次将 EPPB 和 CSIAM 插入至基线模型后,骨干网与颈部网络输出特征图的变化,因篇幅所限,具体内容请扫描本文 OSID 码,在“本文开放的科学数据与内容”中查看。

表 3 本文所提出的模块消融实验结果

方法	EPPB	CSIAM-CA	CSIAM-SA	参数量/10^6	GFLOPs	AP/%
PPMENet	×	×	×	1. 78	0. 85	25. 2
	√	×	×	2. 01	0. 91	26. 8
	√	√	×	2. 07	0. 94	27. 3
	√	×	√	2. 10	0. 95	27. 2
	√	√	√	2. 16	0. 97	27. 7

3 结束语

本文提出了一种轻量化实时目标检测模型 PPMENet,旨在改善现有轻量化检测器中的多尺度特征收集和跨尺度特征融合方式,同时维持较少的计算量和参数量。在骨干网中,本文提出的 EPPB 使用池化金字塔结构来捕获特征的多尺度上下文,有效增强了轻量化骨干网的模型容量,提升了检测精度。在颈部网络,本文设计了 CSIAM 模块来建立不同分辨率和语义级别特征之间的关系,高效融合加权了多尺度特征,并增强了检测器的特征表达能力。实验结果表明,本文提出的 EPPB 模块和 CSIAM 模块,不仅轻量高效,而且有效提升了检测性能,具有一定的实际应用价值。

尽管如此,本文提出的 PPMENet 当前仅局限于目标检测任务,下一步将会迁移其至语义分割、目标跟踪等其他视觉任务,以测试其泛化能力。此外,在未来还会使用预训练、先进数据增强[14]等技术来进一步提升检测器的性能。

参考文献:

[1] 李翠锦,瞿中. 复杂交通环境下多层交叉融合多目标检测[J]. 电讯技术,2023,63(9):1291-1299.

[2] 杨艳红,钟宝江,徐云龙,等. 改进的 SSD 算法在智慧交通中的应用[J]. 电讯技术,2022,62(2):259-265.

[3] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1314-1324.

[4] QIN Z, LI Z M, ZHANG Z N, et al. ThunderNet: towards real-time generic object detection on mobile devices[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6717-6726.

[5] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]//2018 European Confererue on Computer Vision. Cham: Springer, 2018: 122-138.

[6] WANG C Y, BOCHKOVSKIY A, LIAO H M. Scaled-

YOLOv4:scaling cross stage partial network[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville:IEEE,2021:13024-13033.
[7] LI Y,LI J,LIN W,et al. Tiny-DSOD:lightweight object detection for resource-restricted usages[C]//The 29th British Machine Vision Conference. Newcastle:ACM,2018:59-70.
[8] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu:IEEE,2017:936-944.
[9] TANG Q K,LI J,SHI Z P,et al. LightDet:a lightweight and accurate object detection network[C]//2020 IEEE International Conference on Acoustics,Speech and Signal Processing. Barcelona:IEEE,2020:2243-2247.
[10] TAN M X,PANG R M,LE Q V. EfficientDet:scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle:IEEE,2020:10778-10787.
[11] ZHANG Y M,LEE C C,HSIEH J W,et al. CSL-YOLO:a cross-stage lightweight object detector with low FLOPs[C]//2022 IEEE International Symposium on Circuits and Systems. Austin:IEEE,2022:2730-2734.
[12] NEUBECK A,VAN G. Efficient non-maximum suppression[C]//The 18th International Conference on Pattern Recognition. Hong Kong,China:IEEE,2006:850-855.
[13] HU J,SHEN L,SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018:7132-7141.
[14] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO:common objects in context[C]//2014 European Conference on Computer Vision. Cham: Springer,2014:740-755.
[15] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes(VOC) challenge[J]. International Journal of Computer Vision,2010,88(2):303-338.
[16] MICIKEVICIUS P,NARANH S,ALBEN J,et al. Mixed Precision Training[C]//The 6th International Conference on Learning Representations. Vancouver:IEEE,2018:1086-1097.
[17] GE Z,LIU S,WANG F,et al. YOLOx:exceeding YOLO series in 2021[EB/OL]. (2021-08-06)[2024-08-20]. https://arxiv.org/abs/2107.08430.
[18] LIU W,ANGUELOV D,ERHAN D,et al. SSD:single shot MultiBox detector[C]//The 14th European Conference on Computer Vision. Cham:Springer,2016:21-37.
[19] WANG C Y,BOCHKOVSKIY A,LIAO H M. YOLOv7:trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver:IEEE,2023:7464-7475.
[20] WANG C Y,YEH I H,LIAO H Y M. YOLOv9:learning what you want to learn using programmable gradient information[EB/OL]. (2024-02-21)[2024-08-20]. https://arxiv.org/abs/2402.13616.
[21] MEHTA S, RASTEGARI M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer[C]//The 10th International Conference on Learning Representations. Washington DC:IEEE,2022:3421-3446.

作者简介:

陆　蔚　女,1977 年生于江苏无锡,2010 年获工程硕士学位,现为副教授、高级工程师,主要研究方向为计算机视觉、图像处理。