计算机科学
COMPUTER SCIENCE

感受野扩展与多分支聚合的目标检测方法

阙越, 甘梦晗, 刘志伟

引用本文

阙越, 甘梦晗, 刘志伟. 感受野扩展与多分支聚合的目标检测方法[J]. 计算机科学, 2024, 51(6A): 230600151-6.

QUE Yue, GAN Menghan, LIU Zhiwei. Object Detection with Receptive Field Expansion and Multi-branch Aggregation [J]. Computer Science, 2024, 51(6A): 230600151-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600151

感受野扩展与多分支聚合的目标检测方法

阙 越 甘梦晗 刘志伟
华东交通大学信息工程学院 南昌 330013
(qki120@163.com)

摘 要 目标检测旨在实现对图像中目标的精确识别和定位,是计算机视觉中一个重要的研究领域。基于深度学习的目标检测已取得长足的发展,但依然存在不足之处。大的下采样系数带来的语义信息有利于图像分类,但下采样过程中不可避免地会造成信息损失,导致模型特征提取不充分,从而检测准确性下降。针对上述问题,提出一种感受野增强与多分支聚合模型用于目标检测。首先,设计感受野增强模块,以扩大主干网络的感受野。该模块可以获取目标上下文线索,且不改变特征的空间分辨率,可以缓解下采样过程中目标信息丢失问题。然后,为了充分利用卷积神经网络的局部性以及自注意力机制的长距离特征依赖特性,构建感受野扩展复合主干网络,以保留局部特征以及提高模型的全局特征感知能力。最后,提出多分支聚合检测头网络,在3个预测分支之间形成信息流动,融合分支之间的特征信息,以提高模型检测能力。在MS COCO数据集上进行了验证实验,结果表明所提模型的平均精度优于多种主流目标检测模型。

关键词:目标检测;自注意力机制;感受野扩展;特征融合;解耦检测头
中图分类号 TP391.4

Object Detection with Receptive Field Expansion and Multi-branch Aggregation

QUE Yue,GAN Menghan and LIU Zhiwei
School of Information Engineering,East China Jiaotong University,Nanchang 330013,China

Abstract Object detection aims to achieve accurate recognition and localization of objects in images and is an important research area in computer vision. Deep learning-based object detection has made great progress,but there are still shortcomings. The semantic information brought by large down-sampling coefficients is beneficial to image classification,but the down-sampling process inevitably brings information loss,resulting in insufficient model feature extraction and thus a decrease in detection accuracy. To address these problems,this paper proposes a receptive field enhancement and multi-branch aggregation network for object detection. First,the receptive field enhancement module is designed to expand the receptive field of the backbone network. This module can acquire object context cues and can alleviate the problem of object information loss during down-sampling because it does not change the feature spatial resolution. Then,in order to take full advantage of the localization of convolutional neural networks and the long-range feature-dependent property of the self-attention mechanism,the receptive field expanding composite backbone network is constructed to retain local features as well as to improve the global feature perception capability of the model. Finally,a multi-branch aggregation detection head network is proposed to form information flow between three prediction branches and fuse feature information between branches to improve the detection capability of the model. Validation experiments are carried out on MS COCO datasets,and the results show that the average accuracy of the proposed model is better than that of many mainstream object detection models.

Keywords Object detection,Self-attention mechanism,Receptive field expansion,Feature fusion,Decoupled head

1 引言

目标检测是计算机视觉领域的一项经典任务,旨在检测和识别图像的潜在对象。近年来,随着深度学习技术的发展,目标检测的性能取得了显著进步,成为计算机视觉领域的一个热点方向。基于深度学习的目标检测替代人工对图像的处理,节约了大量的人力物力。因此,目标检测技术在日常生活中得到了广泛应用,例如自动驾驶、异物检测、损伤检测等。图像中通常具有不同视角和比例的物体,还有不同的光线以及复杂的背景。因此,对图像提取高质量特征是目标检测模型最为重要的一步。在深度网络中使用较大的下采样率可以获取更大的感受野,得到更多的语义信息。但是,在下采样操作过程中目标信息不可避免地会受到损失,导致检测模型对图像特征提取不充分,从而降低检测性能。

针对上述问题,本文提出一种感受野增强与多分支聚合模型。首先,针对卷积神经网络(CNN)的局部感知特性导致

基金项目:国家自然科学基金(62362032);江西省自然科学基金(20232BAB212011)
This work was supported by the National Natural Science Foundation of China(62362032) and Natural Science Foundation of Jiangxi Province,China(20232BAB212011).
通信作者:刘志伟(zwliu1982@hotmail.com)

获取全局信息的能力有限，提出一种感受野增强模块，在不改变特征空间分辨率的前提下，提高模型获取大感受野特征的能力。其次，基于自注意力机制模型（Transformer）的长距离特征依赖能被级联的自注意力模块所获取，能够很好地提取全局特征，近年来受到计算机视觉领域的广泛关注。结合 CNN 和 Transformer 模型，能够同时保留局部特征和全局特征。因此，本文通过感受野增强模块、Transformer 模块与源主干网络 CSPDarkNet[1] 联合构建一种感受野扩展复合主干网络（Receptive Expanding Composite Network，RECNet），提高模型的特征提取能力。最后，提出一种多分支聚合检测头架构，增加分类、定位和置信度 3 个分支之间的信息交流，以进一步提高目标检测准确性。本文的主要贡献总结如下：

（1）设计了感受野增强模块，在不改变特征分辨率的前提下获取大感受野特征；

（2）构建了感受野扩展复合主干网络，结合感受野增强模块和 Transformer 模块，提高主干网络的特征提取能力；

（3）提出了多分支聚合检测头网络，促进预测分支间的信息流通，提高检测准确性。

本文第二章介绍了相关技术及其相关工作；第三章对本文提出的感受野增强与多分支聚合网络进行了详细的阐述；第四章通过实验对比分析所提出模块的有效性，并与现有工作进行比较分析；最后总结全文。

2 相关工作

2.1 基于感受野扩展的目标检测

大感受野特征具有丰富的语义信息和上下文信息，有利于提升目标检测网络性能。2015 年，He 等[2] 提出空间金字塔池化层结构，由不同区域大小的池化获取多尺度的信息。然而，特征信息在降低特征空间分辨率过程中不可避免地会出现损失。使用大的卷积核[1,3] 来扩大局部计算范围，能够获取大感受野，但会带来较大的资源消耗。还可以利用空洞卷积来扩大感受野。2018 年，Liu 等[4] 提出感受野模块网络，融合多分支特征，构成感受野空间阵列。2019 年，Li 等[5] 提出多分支架构，在不同空洞卷积扩展率下训练不同尺度的样本。2021 年，Chen 等[6] 扩展了主干网络最后一层的感受野，通过不同扩展率的空洞卷积形成的感受野覆盖不同尺度的目标。然而部分小尺度目标的信息在深层特征层可能已经丢失。

本文基于空洞卷积构建感受野增强模块，并结合深度可分离卷积减少参数量。其可实现不改变特征空间分辨率，保留目标的上下文线索，提高特征提取能力。

2.2 基于耦头与解耦头的目标检测

检测头又称为解码器，用于预测图像中目标类别和定位。检测器的分类和边界框回归共享同一个头，称为耦头；检测器的分类器和边界框回归放在不同的头上，称为解耦头。

耦头广泛地应用于 RCNN 系列[7-8] 以及 YOLO 系列[1,3,9]。Fast RCNN[7] 是最早提出共享头结构的工作，模型检测速度有很大的提升。Song 等[10] 提出分类和回归所关注的内容不同，分类任务关注提取的特征与所预测类别的相近程度，回归任务关注真实框与预测边界框位置，耦头结构具有一定的局限性。因此提出解耦头结构来提升模型检测效果，在不同分支上进行分类和回归任务。近年来，解耦头结构被广泛应用于各类目标检测器中。Wu 等[11] 提出一个双头结构，全连接头用于分类任务，卷积头用于定位任务。YOLOX[12] 提出两个平行分支用于分类与回归。

但是，现有解耦头结构的分支之间缺乏信息交流，本文提出的多分支聚合检测头网络分为分类、回归、置信度 3 条分支，并通过自顶向下和自底向上的方式增加分支间的信息交流，提出一种新的解耦头结构。

3 方法介绍

3.1 双残差感受野增强模块

双残差感受野增强模块用于扩大主干网络的感受野。如图 1 所示，模块中使用深度可分离卷积，以降低模块的参数量和计算复杂度，并且设置不同扩展率的空洞卷积，保证高层特征像素映射在底层上的信息是连续的。模块中每一个深度可分离卷积和空洞卷积进行组合，相邻组合使用长残差连接构建恒等映射，并且特征进行交叉拼接，以提高获取特征的灵活性，在一定程度上也能缓解梯度消失或过拟合问题。

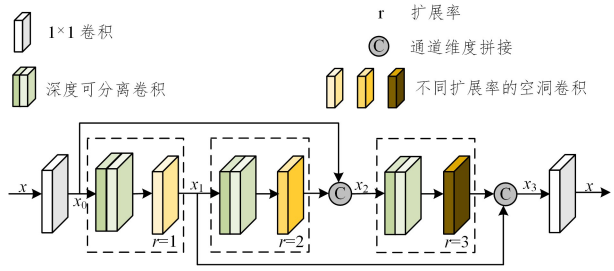


图 1 感受野增强模块

Fig. 1 Receptive field enhancement module

如图 1 所示，输入特征 x 首先由一个 1×1 卷积调整通道数为输入通道数的一半，以降低计算量，其数学表达如下：

$$x_0 = Swish(BN(conv(x))) \tag{1}$$

其中，$Swish$ 是一种自门控激活函数[13]，BN 是批归一化，$conv$ 为卷积。再将处理后的特征 x_0 输入到第一个深度可分离卷积 $DSConv$[14] 与空洞卷积 $DilatedConv$ 的组合中得到特征 x_1，并通过残差连接将特征 x_0 与第二个组合的输出特征在通道维度上进行拼接（Concat）获得特征 x_2，其数学表达如下：

$$x_1 = DilatedConv(DSConv(x_0)) \tag{2}$$

$$x_2 = Concat(DilatedConv(DSConv(x_1)), x_0) \tag{3}$$

而后再经过第三个组合，并再次通过残差连接将特征 x_1 与第三个组合的输出特征在通道维度上拼接获得特征 x_3。最后将由一个 1×1 卷积调整通道数，作为模块最后的输出 x，其数学表达为：

$$x_3 = Concat(DilatedConv(DSConv(x_2)), x_1) \tag{4}$$

$$x = Swish(BN(conv(x_3))) \tag{5}$$

双残差感受野增强模块在不改变特征分辨率的前提下扩大感受野，将其放置在高层特征层能够获取更多的语义信息，提高对应特征层上目标检测性能，也能保留小目标的上下文线索，同时降低计算量。

3.2 感受野扩展复合主干网络

CNN 提取的是局部区域的图像特征，并且已有工作[5,15] 证明 CNN 提取的有效感受野比理论上小，通过加深神经网络来建立图像特征之间的全局关系很困难。为了有效扩大模型感受野，在主干网络中加入设计的感受野增强模块，使模型

能更好获取上下文线索和语义信息，从而提高网络的特征处理以及信息获取能力。此外，长距离的特征依赖对模型关注感兴趣区域以及忽略噪声是非常重要的。因此在主干网络中引入 Transformer 模块，以克服卷积运算的局部性限制。但是，Transformer 与 CNN 相比缺乏平移不变性和局域性等归纳偏好，Transformer 不能很好地泛化，在训练时则需要大量的数据量或合理的训练方式。并且，图像在 Transformer 中的自注意力模块和多层感知器模块的积累下进行处理，将导致计算量快速增加，尤其是在浅层特征层上，特征图越大，计算复杂度越高。因而，在高层特征层使用 Transformer 模块，构建特征远程依赖关系网络，提高模型处理全局信息的能力，并且高层特征层比低层特征层的空间特征小，能够降低模型计算量。

本文模型以 YOLOX-S 为基线模型，其中构建复合主干网络的结构如图 2 所示，在原设计基础上增加感受野增强模块及 Transformer 模块。主干网络中使用 Focus 调整输入图像空间分辨率和通道数大小。模型的阶段 1 和阶段 2 属于低层特征层，特征分辨率大，包含丰富的细节信息和位置信息，利于小尺度目标的检测。利用卷积神经网络的局部性，感受野较小，不会因注入过多的背景信息而湮没小目标特征。因此，直接使用原始的阶段 1 和阶段 2 设计。阶段 3 和阶段 4 属于高层特征层，特征分辨率低，包含丰富的上下文信息和语义信息，对目标的识别十分有利。因此，高层特征层需能捕获长距离的特征依赖，提高模型的全局感知能力。在模型的阶段 3 和阶段 4 均增加了设计的感受野增强模块和 Transformer 模块，目的在于放大高层特征层的感受野以及获取全局特征。由于 Transformer 和 CNN 处理特征不一致，在数据输入 Transformer 模块前使用 flatten 函数及 permute 函数将其调整为适应于 Transformer 的特征，最后使用 reshape 函数和 permute 函数将输出数据调整为适应于 CNN 的特征，以此灵活处理 Transformer 与 CNN 之间的特征尺寸变化。

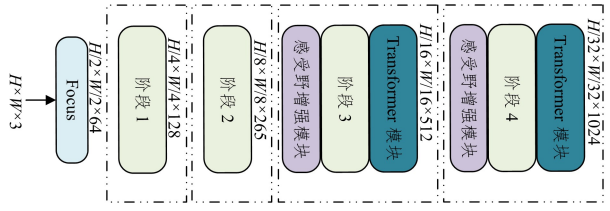


图 2　主干网络结构

Fig. 2　Backbone network architecture

感受野扩展复合主干网络用于保留局部特征和提高全局感知能力，降低训练难度。为了使主干网络能够获取更大的感受野，提出了感受野增强模块，其能够在保持特征分辨不变的情况下，获取丰富的语义信息和挖掘上下文线索。总体上，感受野扩展复合主干网络具有更好的特征提取能力，可避免重要信息丢失。

3.3　多分支聚合检测头网络

过去在 RCNN 系列的工作中使用耦合的检测头，即分类和定位的共享头，预测特征层要同时预测类别和边界框回归。由于分类与定位所关注的信息是不一致的，使用同一个预测特征层检测效果不理想，而解耦的检测头将预测类别分数、边界框回归分为两条独立的分支，分支中包含一系列的卷积操作，能够提升网络的收敛速度与检测效果。为了提升目标检测性能，本文提出多分支聚合检测头，如图 3 所示。在多分支聚合检测头网络中建立 3 个分支，分别对应分类、边界框回归和置信度，3 个分支的参数是独立不共享的。在 3 个预测分支之间进行自顶向下和自底向上的信息交流，以提高模型的检测能力。此外，对于不同特征层采用不同的检测头，即模型中的多分支聚合检测头的参数是不共享的。

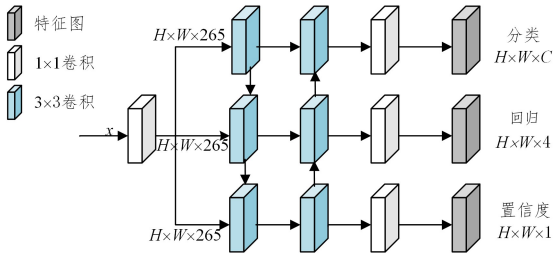


图 3　多分支聚合检测头

Fig. 3　Multibranch aggregate detection head

在多分支聚合检测头中，首先通过一个 1×1 的卷积调整输入特征 x 的通道数，调整后的输出分别为 $P_{\text{cls}}1$，$P_{\text{reg}}1$，$P_{\text{iou}}1$，其数学表达式如下：

$$P_{\text{cls}}1=Swish(BN(conv(x))) \tag{6}$$

$$P_{\text{reg}}1=Swish(BN(conv(x))) \tag{7}$$

$$P_{\text{iou}}1=Swish(BN(conv(x))) \tag{8}$$

然后 3 个分支将继续通过一个 3×3 卷积，并且将特征从分类分支向边界框回归分支和置信度分支依次向下传递，完成一次信息交流后得到特征分别是 $P_{\text{cls}}2$，$P_{\text{reg}}2$，$P_{\text{iou}}2$，该过程的数学表达为：

$$P_{\text{cls}}2=Swish(BN(conv(P_{\text{cls}}1))) \tag{9}$$

$$P_{\text{reg}}2=Swish(BN(conv(P_{\text{reg}}1)))+P_{\text{cls}}2 \tag{10}$$

$$P_{\text{iou}}2=Swish(BN(conv(P_{iou}1)))+P_{\text{reg}}2 \tag{11}$$

接着将 3 个分支将再通过一个 3 ×3 卷积，并且将特征从置信度分支向边界框回归分支和分类分支依次向上传递，完成信息交流后获取的特征分别为 $P_{\text{cls}}3$，$P_{\text{reg}}3$，$P_{\text{iou}}3$，数学表达如下：

$$P_{\text{iou}}3=Swish(BN(conv(P_{\text{iou}}2))) \tag{12}$$

$$P_{\text{reg}}3=Swish(BN(conv(P_{\text{reg}}2)))+P_{\text{iou}}3 \tag{13}$$

$$P_{\text{cls}}3=Swish(BN(conv(P_{\text{cls}}2)))+P_{\text{reg}}3 \tag{14}$$

由此完成了特征自顶向下和自底向上的信息交流，在多分支聚合网络中通过融合特征以提升模型的检测性能。最后，3 个分支分别通过一个卷积层对类别分数、边界框和置信度进行预测，数学表达式如下：

$$P_{\text{cls}}=conv(P_{\text{cls}}3) \tag{15}$$

$$P_{\text{reg}}=conv(P_{\text{reg}}3) \tag{16}$$

$$P_{\text{iou}}=conv(P_{\text{iou}}3) \tag{17}$$

其中，P_{cls}，P_{reg}，P_{iou} 分别代表最后的分类、回归和置信度的预测输出。多分支聚合检测头融合了 3 个预测分支上的特征信息，进行了分支之间的特征交流，以提升模型检测性能。

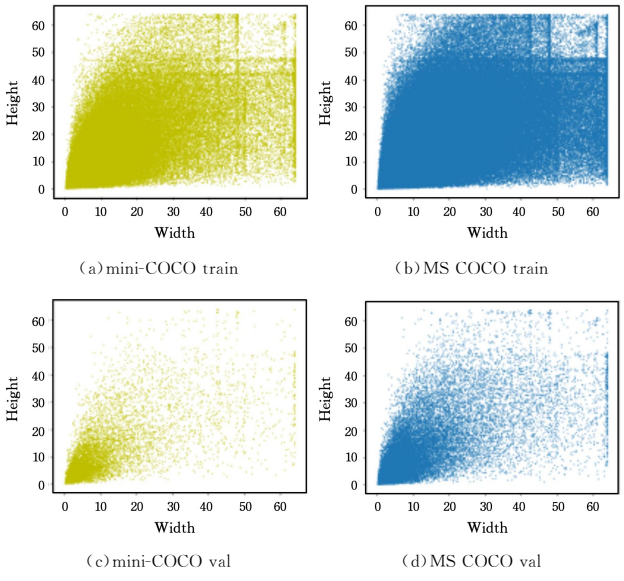
4　实验与分析

4.1　数据集

本文使用 MS COCO 数据集[16] 进行验证实验。COCO 数据集中的图像主要来自生活场景，背景比较复杂。数据集中共有 80 个类别，每张图像平均包含 3.5 个类别和 7.7 个实例。其中使用 118000 张图像进行训练(train 2017)并使用

5000 张图像进行验证(val 2017)。

为了进行快速消融实验研究,在保证小、中、大主体总比例的前提下,从 MS COCO 数据集中随机抽取图像,在随机选择图像时,随机数是均匀分布的,以确保与原始数据的分布一致。取 MS COCO 数据集的 25% 作为 mini-COCO 数据集[17]。训练数据集包含 29 000 张图像,验证数据集包含 1 250 张图像。图 4 给出了 mini-COCO 和 MS COCO 数据集所有实例的宽度和高度分布。mini-COCO 数据集与 MS COCO 数据集的数据分布相同,能够保证实验的可靠性。



(a) mini-COCO train (b) MS COCO train

(c) mini-COCO val (d) MS COCO val

图 4 数据集数据分布对比

Fig. 4 Dataset data distribution comparison

本文模型通过 MS COCO 数据集的评估指标进行评估。使用 10 个 IoU 阈值(0.50:0.05:0.95)计算平均精度 AP;AP_{50}使用 0.5 的 IoU 阈值计算,AP_{75}使用 0.75 的 IoU 阈值计算。本文通过 AP,AP_{50},AP_{75}评估所提模型的检测性能。

4.2 实验设置

本文实验在 MMDetection[18]平台上建立,基于 Pytorch 1.8.1,CUDA 11.1。输入图片大小为(640,640),训练 300 轮,进行 5 轮的预热。使用 4×GPU(NVIDIA RTX A5000),每次输入 64 张图片,进行模型训练与验证。使用 8 ×GPU (NVIDIA RTX A5000),每次输入 128 张图片,消融研究感受野对模型性能影响和所提模块的有效性。模型使用 SGD 优化器,动量设置为 0.9,以保证模型在训练过程中的稳定性,达到避免振荡和摆脱局部约束的效果。权重衰减被设置为 0.0005,以避免因权重过大而出现过度拟合。学习率设置为 0.01,其余设置与 YOLOX[12]模型训练设置保持一致。在后续实验说明中,如果没有特殊指定,则默认使用以上设置。

4.3 实验对比与分析

本小节介绍模型在 MS COCO 数据集上训练及验证结果。如表 1 所列,在 MS COCO val 2017 数据集上比较了不同目标检测方法的平均精度。本文模型以 YOLOX-S[12]为基线模型,YOLOX-S 的 AP 达到 40.3%,本文模型的 AP 提高了 3.4%。TridentNet[15]和 YOLOF[6]也是基于空洞卷积的方法,本文模型通过在主干网络的高层特征增加感受野扩展模块来获取大感受野特征,相较于 TridentNet,AP 提高了 6.0%,相比 YOLOF,AP 提高了 6.2%。实验表明,本文模型比其他基于空洞卷积的方法表现更好。与基于 Transformer 的模型 UP-DETR[19]和 DETR[20]相比,本文方法的目标检测效果有明显提升,并且模型的参数量也少于其他基于 Transformer 的模型。此外,本文模型是基于无锚框机制,相较于先进的无锚框目标检测模型 FCOS[21],Reppoint[22],CornerNet[23]等,在平均精度 AP 上均有所提升。本文模型与 Reppoint 相比,AP 有 3.2%的提升。本文模型还与一些其他的先进工作进行了比较,通过数据对比,本文模型在平均精度上具有明显优势。

表 1 不同方法在 COCO val 上比较

Table 1 Comparison of different methods on COCO val

模型	主干网络	AP	AP_{50}	AP_{75}	Params/M
TridentNet[5]	ResNet50	37.7	57.4	40.6	33.57
YOLOF[6]	ResNet50	37.5	57.0	40.4	43.88
UP-DETR[19]	ResNet50	40.5	60.8	42.6	—
DETR[20]	ResNet50	40.1	60.6	42.0	41.30
ABFPN[24]	ResNet50	38.6	61.3	—	24.4
QueryDet[25]	ResNet-50	39.3	59.9	42.0	24.9
RetinaNet[26]	ResNet101	38.6	57.8	41.1	56.74
Faster RCNN[8]	ResNet101	39.4	60.1	43.1	84.36
Cascade RCNN[27]	ResNet101	42.0	60.4	45.7	88.16
FCOS[21]	ResNet101	40.6	60.2	43.5	50.96
Reppoint[22]	ResNet101	40.5	61.3	43.5	55.62
DETR	TNT-S[28]	38.2	58.9	39.4	39.00
DETR	PVT-M[29]	36.4	57.9	37.2	57.00
RetinaNet	PVTv2-b1[30]	41.1	62.0	43.8	23.75
YOLOX[12]	CSPDarknet	40.3	59.1	43.4	8.97
CornerNet[23]	Hourglass104	40.6	56.6	43.2	201.04
REMBA	RECNet	**43.7**	**62.5**	**47.0**	17.49

如图 5 所示,选取了数据集中具有密集目标和不同尺度目标的图像为本文模型和现有工作的检测效果对比图。从左向右依次为本文模型、PVTv2-b1[30]、Reppoint 和 YOLOX-S。图上显示了检测目标的检测框、目标分类以及置信度,其中为了更方便观察目标检测效果,图片部分区域使用矩形框标记并且放大。在这 3 组图中,本文模型相较于其他模型在检测目标上获得了更高的置信度,检测框更趋近于目标。实验结果表明本文模型在目标检测效果上具有一定的优势。

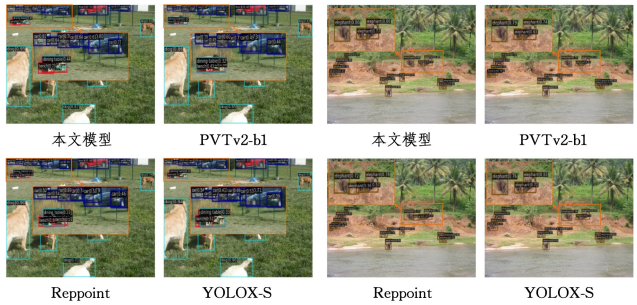


图 5 本文模型与现有工作在 COCO 数据集上的检测效果比较

Fig. 5 Detection effect comparison of the proposed model and existing works on COCO dataset

4.4 消融实验

在本小节中,为了更好地理解所提方法,在 mini-COCO 数据集上验证放大感受野以及所提方法对模型性能的影响。

(1)感受野对模型性能的影响

TridentNet[5]指出,感受野越大,对大目标的检测效果就越好,反对小目标的检测效果就越好。为此,模型在深层使用所提出的感受野扩展模块来扩大感受野,获取上下文信息和

语义信息，通过目标检测颈部网络对特征进行处理，将深层特征相融合至浅层特征中，丰富浅层特征信息，从而提高目标检测效果。对此设计 3 个实验来验证不同感受野对模型性能的影响。如表 2 第二行所列，在主干网络阶段 4 增加感受野增强模块，比基线模型 YOLOX-S 的 AP 增加了 0.8%。如表 2 第三行所列，继续在主干网络阶段 3 增加感受野扩展模块，AP 提高了 0.5%。从实验结果可以看出，通过扩大感受野可以提升模型的检测能力。

表 2　感受野增强模块对模型性能的影响

Table 2　Effect of receptive field enhancement module on model performance

阶段 4	阶段 3	AP	AP_{50}	AP_{75}
×	×	28.8	47.6	30.7
√	×	29.6	48.6	30.9
√	√	**30.1**	**49.5**	**31.7**

(2)模块有效性验证

该实验是对感受野扩展复合主干网络和特征增强检测头网络的有效性进行消融分析。具体来说，分成 3 个组成模块进行有效性实验：Transformer 模块、感受野增强模块和特征增强检测头网络。如表 3 所列，Transformer 具有长距离特征依赖，能够获取全局信息，对模型性能提高产生积极影响。在基线模型上增加 Transformer 模块对目标检测性能有一定的提升。感受野扩展模块在不改变特征空间分辨率以及不增加额外参数情况下扩大感受野，以保留目标上下文线索，以及获取更多的语义信息。增加感受野扩展模块使模型性能 AP 提升了 1.1 %。多分支聚合检测头网络设计分类、边界框回归以及置信度 3 个预测分支，并促进 3 条分支之间的信息交流。将基线模型检测头替换为多分支聚合检测头，网络目标检测性能 AP 提高了 1.2 %。以上实验结果表明，Transformer 模块、感受野增强模块和多分支聚合检测探头网络对提高检测模型性能是有效的，相比基线模型检测效果更好。

表 3　在 mini-COCO 数据集上验证模块的有效性

Table 3　Verify module validity on mini-COCO dataset

模型	AP	AP_{50}	AP_{75}	Flops/(GFLOPs)	Params
基线模型	28.8	47.6	30.7	13.32	8.97×10^6
＋ Transformer 模块	28.9	47.9	31.1	14.58	12.22×10^6
＋感受野增强模块	30.0	48.1	31.8	16.54	15.27×10^6
＋多分支聚合检测头	**31.2**	**50.0**	**32.9**	**22.74**	$\mathbf{17.49\times10^6}$

结束语　本文利用特征融合方法与感受野增强方法构建了一个感受野扩展和多分支聚合网络用于目标检测。在本文模型中，考虑到不同大小目标所需的特征不同，将 Transformer 和感受野增强模块设置在高层特征层。所构建的感受野扩展复合主干网络融合了 CNN 的局部性、平移不变性以及 Transformer 捕获长距离特征依赖的能力，很好地保留了模型局部特征，并提高模型全局感知能力。在主干网络中插入感受野增强模块，能够在不改变特征空间分辨率的前提下，增大模型感受野，从而缓解下采样扩展感受野带来的信息损失，并能获取到更多的语义信息和上下文线索。本文还提出了一个多分支聚合检测头网络，通过不同预测分支之间自顶向下和自底向上的特征融合来完成分支之间的信息交流，以提升模型的检测性能。在 MS COCO 数据集上训练及验证模型性能，本文模型的平均精度优于多种主流目标检测模型。

参 考 文 献

[1] BOCHKOVSKIY A，WANG C，LIAO H，et al. Yolov4：Optimal Speed and Accuracy of Object Detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online，2021：13029-13038.

[2] HE K，ZHANG X，REN S，et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，37(9)：1904-1916.

[3] WANG C，BOCHKOVSKIY A，LIAO H，et al. YOLOv7：Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors[J]. arXiv：2207.02696，2022.

[4] LIU S，HUANG D，et al. Receptive Field Block Net for Accurate and Fast Object Detection[C]// European Conference on Computer Vision. Munich，Germany，2018：385-400.

[5] LI Y，CHEN Y，WANG N，et al. Scale-Aware Trident Networks for Object Detection[C]// IEEE/CVF International Conference on Computer Vision. Seoul，Korea(South)，2019：6054-6063.

[6] CHEN Q，WANG Y，YANG T，et al. You Only Look One-Level Feature[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online，2021：13039-13048.

[7] GIRSHICK R. Fast R-CNN[C]// IEEE International Conference on Computer Vision. Santiago，Chile，2015：1440-1448.

[8] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(6)：1137-1149.

[9] REDMON J，FARHADI A，et al. Yolov3：An Incremental Improvement[J]. arXiv：1804.02767，2018.

[10] SONG G，LIU Y，WANG X，et al. Revisiting The Sibling Head in Object Detector[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，WA，USA，2020：11563-11572.

[11] WU Y，CHEN Y，YUAN L，et al. Rethinking Classification and Localization for Object Detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，WA，USA，2020：10186-10195.

[12] GE Z，LIU S，WANG F，et al. Yolox：Exceeding Yolo Series in 2021[J]. arXiv：2107.08430，2021.

[13] RAMACHANDRAN P，ZOPH B，LE Q，et al. Searching for Activation Functions[J]. arXiv：1710.05941，2017.

[14] CHOLLET F. Xception：Deep Learning with Depthwise Separable Convolutions[C]// IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，HI，USA，2017：1251-1258.

[15] DING X，ZHANG X，HAN J，et al. Scaling Up Your Kernels to 31x31：Revisiting Large Kernel Design in CNNs[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，Louisiana，2022：11963-11975.

[16] LIN T，MAIRE M，BELONGIE S，et al. Microsoft Coco：Common Objects in Context[C]// European Conference on Computer Vision. Zurich，Switzerland，2014：740-755.

[17] SAMET N，HICSONMEZ S，AKBAS E，et al. HoughNet：Integrating Near and Long-Range Evidence for Bottom-Up Object Detection[C]// European Conference on Computer Vision. Glasgow，US，2020：406-423.

[18] CHEN K, WANG J, PANG J, et al. Mmdetection: Open Mmlab Detection Toolbox and Benchmark [J]. arXiv: 1906. 07155, 2019.

[19] DAI Z, CAI B, LIN Y, et al. Up-Detr: Unsupervised Pre-training for Object Detection with Transformers[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1601-1610.

[20] CARION N, MASSA F, SYNNAEVE G, et al. End-to-End Object Detection with Transformers[C]// European Conference on Computer Vision. Glasgow, US, 2020: 213-229.

[21] TIAN Z, SHEN C, CHEN H, et al. Fcos: Fully Convolutional One-Stage Object Detection [C] // IEEE/CVF International Confe-rence on Computer Vision. Seoul, Korea (South), 2019: 9627-9636.

[22] YANG Z, LIU S, HU H, et al. Reppoints: Point Set Representation for Object Detection[C]// IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South), 2019: 9657-9666.

[23] LAW H, DENG J. Cornernet: Detecting Objects as Paired Keypoints[C]// European Conference on Computer Vision. Munich, Germany, 2018: 734-750.

[24] ZENG N, WU P, WANG Z, et al. A Small-Sized Object Detection Oriented Multi-Scale Feature Fusion Approach with Application to Defect Detection[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-14.

[25] YANG C, HUANG Z, WANG N, et al. QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA, 2022: 13668-13677.

[26] LIN T, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection[C]// IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2980-2988.

[27] CAI Z, VASCONCELOS N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation[J]. IEEE Transactions Pn pattern Analysis and Machine Intelligence, 2019, 43 (5): 1483-1498.

[28] HAN K, XIAO A, WU E, et al. Transformer in Transformer [J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.

[29] WANG W, XIE E, LI X, et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions [C]// IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 568-578.

[30] WANG W, XIE E, LI X. Pvtv2: Improved Baselines with Pyramid Vision Transformer[J]. Computational Visual Media, 2022, 8(3): 415-424.

QUE Yue, born in 1991, Ph.D, lecture, is a member of CCF (No. P2963M). His main research interests include computer vision and deep learning.

LIU Zhiwei, born in 1982, Ph.D, professor. His main research interests include target-aware imaging and high-performance computing.