

Rapider-YOLOX: 高效的轻量级目标检测网络

顾宙瑜, 於跃成, 者甜甜

(江苏科技大学计算机学院, 江苏 镇江 212000)

摘 要: 作为一种轻量级网络结构, YOLOX-Nano 具有运行速度快的优势。然而, 该网络在实际应用中仍然存在特征提取能力较弱、检测精度不足的缺陷。为此, 提出了一种综合平衡检测速度和检测精度的高效目标检测网络 Rapider-YOLOX。首先, 设计高效瓶颈模块, 以提升原始 YOLOX-Nano 模型中深度卷积模块的特征提取能力。其次, 设计软空间金字塔池化模块, 以避免原始 SPP 模块容易出现丢失部分重要信息现象, 进一步提升多尺度信息融合及通道间信息交流的能力。最后, 引入 CIoU 损失, 利用预测框与真实框的中心距离及宽高比提升预测框的位置精度。在 PASCAL VOC2007 数据集上的实验结果表明, 提出的 Rapider-YOLOX 模型的 mAP 达到 77.92%, 比原始 YOLOX-Nano 高 3.79%。此外, 在 CUDA 核心数仅为 384 的 GT1030 上, FPS 达到 45.40, 在 CPU 上 FPS 也可达到 23.94, 从而在确保网络轻量级特性的同时, 进一步提升了网络的检测精度和泛化性能。

关键词: 目标检测; 高效卷积神经网络; YOLOX-Nano; 轻量级; 高精度

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202303

Rapider-YOLOX: lightweight object detection network with high precision

GU Zhouyu, YU Yuecheng, Zhe Tiantian

College of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212000, China

Abstract: As a lightweight network structure, YOLOX-Nano has the advantage of fast running speed. However, the model still has the defects of weak feature extraction ability and insufficient detection accuracy in practical application. Therefore, an efficient object detection network Rapider-YOLOX which comprehensively balanced the detection speed and detection accuracy was proposed. Firstly, the highly efficient bottleneck module was designed to improve the feature extraction capability of depthwise convolutional blocks in the original YOLOX-Nano model. Secondly, the soft-SPP module was designed to avoid the loss of some important information in the original SPP module and improve the ability of multi-scale information fusion and information exchange between channels further. Finally, CIoU was introduced to improve the position accuracy of the prediction box by using the center distance and aspect ratio between the prediction box and the real box. The experimental results on PASCAL VOC2007 dataset showed that the mAP of Rapider-YOLOX model reached 77.92%, which was 3.79% higher than the original YOLOX-Nano. In addition, on GT1030 with only 384 CUDA cores, the FPS of the proposed method could reach 45.40. The FPS could also reach 23.94 on the CPU, which

收稿日期: 2022-09-26; 修回日期: 2022-11-17

通信作者: 於跃成, zhjyuecheng@163.com

基金项目: 国家重点研发计划资助项目 (No.2018YFC0309104); 江苏省建设系统科技项目 (No.2021JH03); 2022 年度江苏省工业和信息产业转型升级专项资金-关键核心技术 (装备) 攻关产业化项目 (No.CMHI-2022-RDG-004); “深远海智能风电运维母船研制关键技术攻关”

Foundation Items: The National Key Research and Development Program of China (No.2018YFC0309104), The Construction System Science and Technology Project of Jiangsu Province (No.2021JH03), Special Fund for Transformation and Upgrade of Jiangsu Industry and Information Industry - Key Core Technologies (Equipment) Key Industrialization Projects in 2022 (No.CMHI-2022-RDG-004): "Key Technology Research for Development of Intelligent Wind Power Operation and Maintenance Mothership in Deep Sea"

further improved detection accuracy and generalization performance of the network while ensuring the lightweight characteristics of the network.

Key words: object detection, efficient convolutional neural network, YOLOX-Nano, lightweight, high precision

0 引言

目标检测是计算机视觉^[1]领域中的主要任务之一。深度学习技术的不断发展为目标检测效果的显著提升提供了有力支撑,其已被广泛应用于人体行为检测^[2]、无人机目标检测^[3]等领域。基于深度模型的目标检测算法已逐步取代依赖手工特征的传统检测算法^[4-7]。近年来,越来越多的大型卷积神经网络相继被提出,其中二阶段目标检测的代表性算法主要包括区域卷积神经网络(region convolutional neural network, R-CNN)^[8-11]系列、基于区域的全卷积神经网络(region-based fully convolutional network, R-FCN)^[12]和 CoupleNet^[13]等,而一阶段目标检测的代表性算法主要包括 YOLO (you only look once) 系列^[14-17]、SSD (single shot multibox detector)^[18]和 RetinaNet^[19]等。

近年来,为了追求更高的检测精度,同时应对更大难度的检测任务,网络模型的规模和深度也逐渐提升。这些更加精细复杂的网络可以提取更加充分的图像特征,获取更加丰富的语义信息,并以此达到提升网络模型精度的目的。但是,这必然导致模型计算的参数量和计算复杂度急剧上升,从而限制了这些模型在计算资源有限且对检测任务有较高实时性和准确性要求的工业环境下的应用。

作为 YOLO 的改进版本,YOLOX 检测器^[20]将原有的有锚形式改为无锚形式,检测头部分采用了解耦头形式,并将 SimOTA 作为其标签分配策略,从而实现了目标检测速度和检测精度的有效平衡。YOLOX-Nano 则在 YOLOX 的基础上,采用深度可分离卷积方法进行构建。作为一种轻量级网络模型,YOLOX-Nano 的模型体积仅为 0.91 MB,计算开销极低,检测速度优于 NanoDet 等一系列轻量化模型,进一步提升了检测算法的实时性能。然而,YOLOX-Nano 大量复用深度可分离卷积,因此其仍然无法避免各种轻量级目标检测模型面临的特征提取能力弱、细节信息损失大等问题。由此可见,作为当前性能平衡方面表现突出的 YOLOX 的轻量级版本,YOLOX-Nano

检测速度快,且在同等参数数量下具有领先于其他模型的检测精度。然而,当 YOLOX-Nano 应用于工业场景时,其检测精度仍然无法满足实际使用的需求。

为此,本文以 YOLOX-Nano 网络结构为基础,首先在借鉴 ConvNeXt^[21]和 RegNet^[22]模型设计原则的基础上设计高效瓶颈(highly efficient bottleneck, HEB)模块,以便在提升传统网络特征提取能力的同时去除冗余的计算开销。其次,借助本文设计的软空间金字塔池化(soft-SPP, SSPP)模块,一方面使模型特征层在进入预测部分前能融入更多尺度的细节信息,另一方面模块采用的软池化操作(Soft Pool)^[23]优化方法可以避免重要信息在 SPP 结构中丢失,模型中融入的轻量级高效通道注意力(efficient channel attention, ECA)模块^[24]则可以进一步加强通道间的信息交流。最后,通过引入完全交并比(complete intersection over union, CIoU)优化原始模型的交并比损失函数(intersection over union loss, IoU Loss)^[25],使损失函数可以同时关注预测框和真实框之间中心位置的距离及宽高比的差异性。

1 相关介绍

作为降低网络训练和计算开销的主要方法,神经网络压缩主要分为模型剪枝^[26]、量化加速^[27]、知识蒸馏^[28]3种方法。模型剪枝方法往往通过去除网络中多余权重的方式获得更快的推理速度和更低的计算开销。量化加速方法则是以降低模型参数精度的代价换取较小的存储空间。不同于上述方法,知识蒸馏方法采用了知识迁移的思想,通过学习能力较强的大型网络辅助学习能力较差的小型网络进行训练,将大型网络中学习到的知识迁移至小型网络,以此增强小型网络的泛化能力。显然,这些方法虽然可以将现有的神经网络压缩成较小的网络,但是并未对网络结构进行有效改进,压缩后的网络模型的性能必然在很大程度上依赖预先给定的网络模型。这意味着,神经网络压缩的方式可以减小网络体积,但却无法提高模型的检测精度。

不同于神经网络压缩方法,紧凑型神经网络

主要通过设计低复杂度和低参数量的新型网络结构获得高效的网络模型。谷歌提出的 MobileNet 系列^[29-31]通过将深度可分离卷积应用于模块设计,减小了模型的参数量和计算量。旷视科技提出的 ShuffleNet^[32-33]系列则采用了分组卷积,并提出了以通道混洗的方式减少计算开销。类似地,华为提出的 GhostNet^[34]网络模型中的 Ghost 模块则能在更少的参数条件下生成更多的特征图,从而在输出特征图一定的情况下有效降低模块的参数量和计算开销。事实上,这些新颖的神经网络结构和设计方式同时兼顾了模型的速度和精度,这对于计算资源相对短缺的移动端和嵌入式应用场景来说具有很好的实用价值。

为了进一步提升网络模型的精度和速度,参考紧凑型神经网络的设计理念,Wong A 等人^[35]将 SqueezeNet^[36]中 Fire 模块与 SSD 网络模型相结合提出了 Tiny SSD 模型。类似地,Fang W 等人^[37]通过引入 Fire 模块并调整其在网络中的数量和位置,在进一步优化其连接风格的基础上提出了 Tinier-YOLO 网络模型。实验结果表明,Tiny SSD 和 Tinier-YOLO 的模型体积分别为 2.3 MB 和 8.9 MB,但是这两个模型在 VOC2007 数据集上的目标检测精度平均值(mean average precisim, mAP)却分别达到了 61.3%和 65.7%。由此可见,将高效的网络模块引入现有网络模型,进而设计相应的优化方

案,已成为提升原有模型检测精度、保持并进一步提升原有模型检测速度的有效途径。

2 模型改进方法

YOLO 系列是目前工业场景下使用最广泛的目标检测网络,而 YOLOX-Nano 则是最新的 YOLOX 系列中最轻量级的目标检测网络。与现有的其他网络相比,YOLOX-Nano 具有最少的参数量和最小的计算复杂度,同时也具有较好的检测精度。然而,在检测目标相对密集或目标相互遮挡等复杂场景下,YOLOX-Nano 往往会出现错检和漏检的情况。为此,本文在 YOLOX-Nano 的基础上提出了一种性能平衡更优的目标检测神经网络模型 Rapider-YOLOX,旨在确保模型轻量级特性的同时进一步提高模型的检测精度。Rapider-YOLOX 整体框架结构如图 1 所示。其中 Conv 表示卷积模块,BN 表示批量标准化模块,GroupConv 表示分组卷积模块,CBL 是 Conv、BN 和 LeakyReLU 激活函数 3 部分构成的组合模块,GCBL 是 GroupConv、BN 和 LeakyReLU 激活函数 3 个部分构成的组合模块,PWConv 表示点态卷积模块,Stride 表示池化步长,Add 表示求和操作,Concat 表示堆叠操作,Decoupled Head 表示解耦头,Reshape 表示铺平操作,SiLU 表示 SiLU 非线性激活函数。Reshape 用于将三维矩阵转变为二维矩阵。相较于

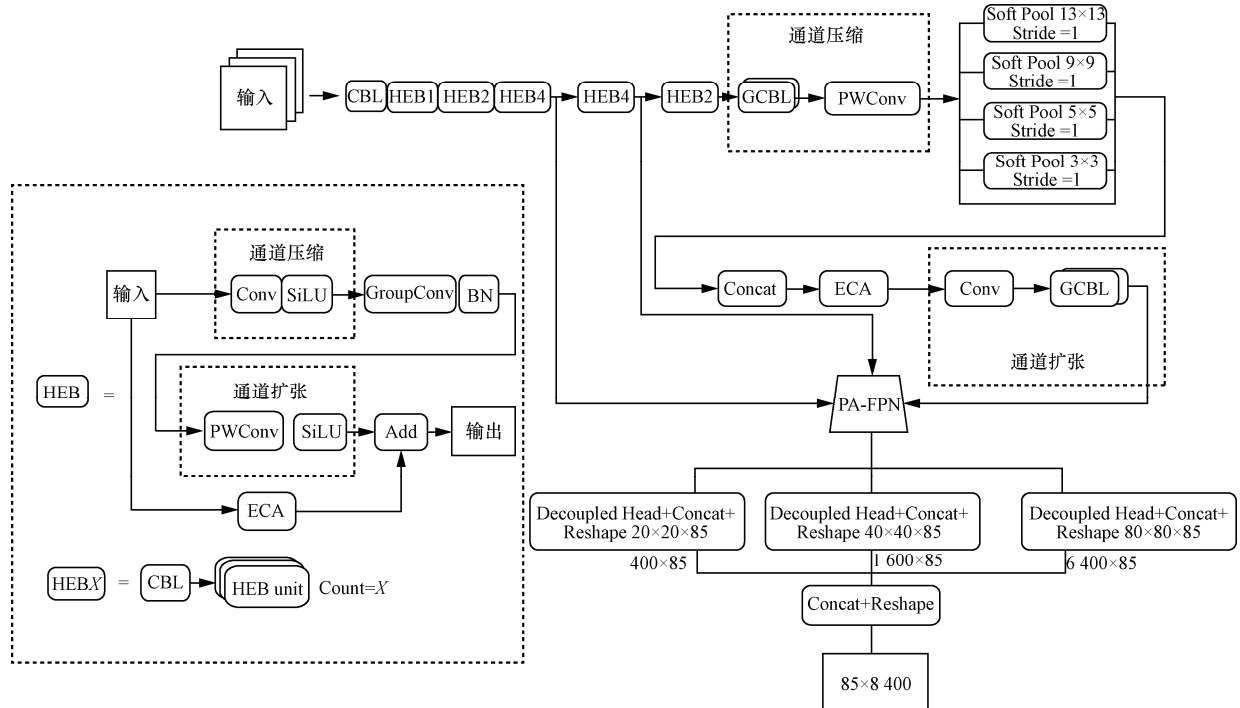


图1 Rapider-YOLOX 整体框架结构

YOLOX-Nano, Rapider-YOLOX 具有更高效的特征提取主干网络, 该主干网络由本文设计的 HEB 结构构成。图像数据经过该主干网络进行特征提取后形成的特征图被送入 SSPP, 相较于原始 SPP, 本文提出的 SSPP 使用 SoftPool 解决了原始模块仅关注输入特征图局部最大值信息而导致的部分重要信息丢失的问题, 同时融入注意力模块, 加强了不同尺度融合后特征图的通道信息交流能力。最后, 将特征图送入路径聚合特征金字塔网络 (path aggregation feature pyramid network, PA-FPN) 进行特征加强, 由解耦头构建的预测模块进行信息预测。

2.1 高效瓶颈结构设计

为了获得更小的模型体积和更少的计算量, YOLOX-Nano 中使用了大量的深度卷积来构建模型。如图 2 所示, 深度卷积 (depthwise convolution, DWConv) 是 YOLOX-Nano 的瓶颈 (bottleneck) 结构中的主要组成部分, 而且在每次卷积操作后都会统一衔接上批量标准化和激活函数。然而参考文献[22]中的大量实验表明, 在低计算量的情况下反向残差结构的使用往往会降低模型的检测精度, 而且大量使用深度卷积还会进一步降低模型的精度; 反之, 如果采用分组卷积来构建卷积模块则可以提升模型的检测性能。此外, ConvNeXt^[21]模型在微观尺度上的架构设计方面的实验结果表明, 标准化层和激活函数层的使用也会对模型的性能产生影响。参照 Vision Transformers^[38]的结构设计, ConvNeXt 模型通过实验验证了在卷积模块中使用更少的激活函数及批量标准化层的方式, 确实有助于整个网络模型精度的提升, 而在卷积模块的起始部分使用批量标准化并不能提升模型的性能。

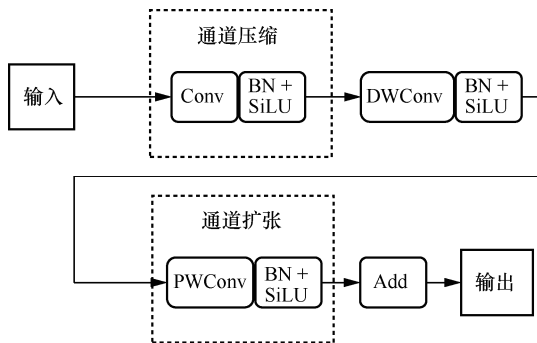


图2 YOLOX-Nano 的瓶颈结构

由此可见, YOLOX-Nano 模型中大量使用的深度卷积一方面导致较多重要信息丢失, 另一方面会产生额外的计算开销。也就是说, YOLOX-Nano 模

型中大量使用的深度卷积结构虽然有利于加快模型的正向推理速度, 但其过少的参数势必会降低模型提取特征的充分性。这也成为 YOLOX-Nano 网络特征提取能力较弱、整体性能难以满足实际应用的主要原因之一。为此, 受 RegNet 提出的设计原则^[33]及 ConvNeXt 模型中网络优化思想^[34]的启发, 本文改进了 YOLOX-Nano 网络中瓶颈的结构设计, 重新设计了在神经网络正向推理中具有更强大特征提取能力的网络模块 HEB, 并以此来构建网络模型。同时, 在卷积模块的设计过程中优化了批量标准化和激活函数的使用, 避免了每次执行卷积操作后都执行批量标准化和激活函数。

HEB 模块的结构如图 3 所示, 不同于原始 YOLOX-Nano 模型中的瓶颈结构, HEB 模块主要在模块起始部分的结构设计、卷积操作和冗余信息的处理等 3 个方面进行了改进。首先, 在设计通道压缩的卷积操作时, 考虑到在微型模块的起始部分使用批量标准化并不能有效提升网络模型的检测效果, 因此为了进一步降低计算复杂度, 在设计 HEB 模块时减去了通道压缩部分和通道扩张部分的批量标准化操作。其次, 为了提升网络模型的特征提取能力, HEB 模块采用分组卷积替换了原始瓶颈结构中的深度卷积结构。进而, 受到 ConvNeXt 模型使用更少激活函数的设计思想启发, HEB 模块在分组卷积后去除了 SiLU 激活函数。最后, 针对原 Bottleneck 结构中使用残差连接导致的低维冗余信息向高维特征层的传递问题, 将轻量级 ECA 模块引入 HEB 模块。这样, 输入特征层的通道信息被加以过滤, 实现了残差连接时输入特征层中低价值信息的自适应抑制, 加强了重要信息的特征表现, 确保了具有正向表现能力的低维信息在网络中向后传播的有效性。重新设计的 HEB 模块不但具有较低的计算复杂度, 而且进一步提升了轻量级网络模型的最终检测效果。

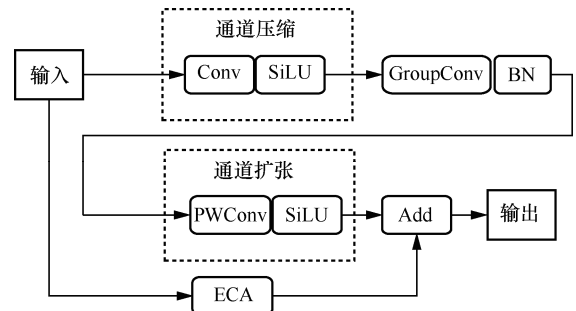


图3 HEB 模块的结构

2.2 软空间金字塔池化模块

SPP 模块是一种借鉴了空间金字塔思想的特征融合模块，其目的是融合多种尺度的特征来增强特征的表现能力。原始 SPP 结构如图 4 所示，在 YOLOV3-SPP 模型中 SPP 模块融合了 5×5 、 9×9 和 13×13 共 3 种不同感受野的特征。实验结果表明，在 YOLOV3 网络中单独添加 SPP 结构便可显著提升模型的检测精度。

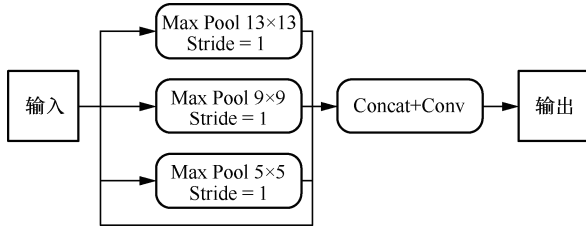


图 4 原始 SPP 结构

YOLOX-Nano 中的 SPP 模块仍然沿用了 YOLOV3-SPP 中的 SPP 结构。然而，原始 SPP 结构中采用的最大池化（Max Pool）操作仅仅关注区域内的局部最大值，完全忽略了区域内的其余信息，这必然导致压缩后的特征丢失大量的信息。为此，如图 5 所示，将 Soft Pool 操作引入 SPP 模块，同时增加提取小尺度特征信息的池化核和增强通道方向特征表现能力的 ECA 模块，设计了新的具有更好特征提取性能的特征融合结构软空间金字塔池化模块。

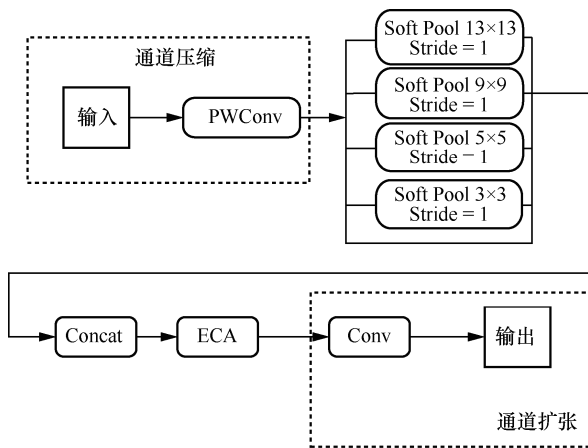


图 5 SSPP 结构

SSPP 结构中采用的池化操作方法为 Soft Pool，其本质是一种指数加权的池化方法。如式 (1) 所示，权重 ω_i 为激活值 a_i 的自然指数与池化核区域 R 中所有激活值 a_j 的自然指数之和的比值，其中 e 是

自然指数， i 是激活值 a_i 和权重 ω_i 对应的索引。这样，式 (2) 中 Soft Pool 操作的输出 $\tilde{a}$ 实际上是池化核区域 R 内所有激活值的线性加权和。

$$\omega_i = \frac{e^{a_i}}{\sum_{j \in R} e^{a_j}} \quad (1)$$

$$\tilde{a} = \sum_{i \in R} \omega_i \times a_i \quad (2)$$

由此可知，相比于 Max Pool 仅关注池化区域内的最大值，Soft Pool 在池化时考虑了区域内的所有信息。这样，SSPP 结构在池化操作的过程中，既可以避免 Max Pool 产生的重要信息丢失的问题，又可以确保具有较大激活值的特征能对输出结果产生更显著的影响。此外，相比于 Max Pool，Soft Pool 除了能够提升网络性能外，其在网络的正向推理速度方面也有着明显的优势^[23]。

如图 5 所示，为了减少额外的计算开销，在 SSPP 头部进行了通道压缩，以使后续多尺度特征融合过程在低通道维度上进行，尾部通过通道扩张还原通道维度。同时为了获得更多尺度的信息，尤其是小尺度的细节信息，在 SSPP 结构中专门添加了 3×3 大小的池化核，并将池化后的结果以原始堆叠的方式进行不同尺度的特征融合。此时，如果将多个不同尺度的池化核中的池化特征层沿通道方向堆叠在一起，并直接进行通道压缩，势必导致通道方向上的特征提取不够充分。为此，SSPP 结构在堆叠操作之后添加了一个 ECA 模块，使得堆叠后的特征层能够自适应细化通道特征，增强通道方向的特征表现能力。借助重新设计的 SSPP 结构，网络模型在进入预测之前将获得表现能力更强和细节信息更为丰富的有效特征层，并提升模型的最终预测性能。

2.3 损失函数优化

在实验分析 YOLOX-Nano 模型的预测结果时，模型的部分预测框与真实框的位置偏差较大，拟合程度偏低。其中，图 6 所示的两类情形值得关注。首先，由于 YOLOX-Nano 的 IoU Loss 以 IoU 为度量准则，仅仅考虑了预测框与真实框的交集与并集的比值。为此，当出现图 6 (a) 和图 6 (b) 所示的预测框与真实框的交集为空的情形时，IoU 的值为 0，IoU Loss 的值恒为 1。这样，既导致了网络无法收敛，也无法反映这两种情形下预测框与

真实框之间中心距的不同。事实上,在预测框与真实框不相交的情况下,预测框与真实框之间的相对位置也在一定程度上反映了预测框位置的合理性。其次,观察图 6 (c) 和图 6 (d) 可知,即使在 IoU 值相等,即预测框与真实框相对位置相同的情况下,预测框与真实框之间宽高比的差异性也直观地反映了预测框的合理性。为此,为了使预测框更加贴合目标所在真实框的位置,本文在设计 Rapider-YOLOX 模型时引入 CIoU Loss,以取代 YOLOX-Nano 模型原来采用的 IoU Loss。

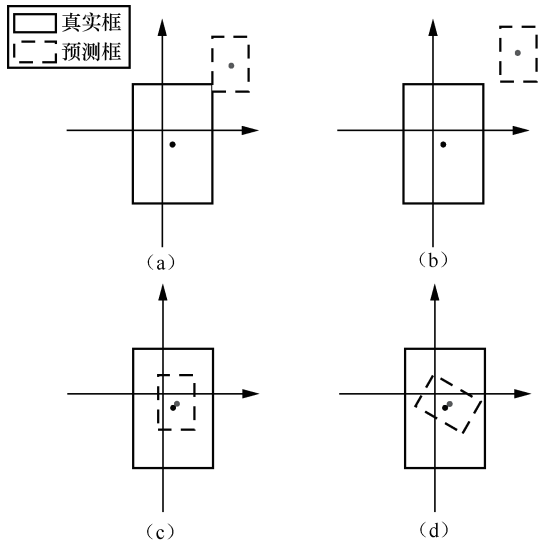


图 6 预测框与真实框相对位置示意图

不同于 IoU Loss, CIoU Loss 中增加了预测框与真实框之间的中心距和高宽比一致性的惩罚项 R_{CIoU} 。 R_{CIoU} 表达式如式 (3) 所示:

$$R_{\text{CIoU}} = \frac{\rho^2(b, b^{gt})}{c^2} + \gamma v \quad (3)$$

其中,第 1 项为距离惩罚项,第 2 项为高宽比一致性惩罚项。 b 和 b^{gt} 分别为预测框与真实框的中心点坐标, c 为包含了预测框与真实框的最小矩形的对角线长度, ρ 为预测框与真实框的中心之间的欧氏距离。式 (3) 中的高宽比一致性惩罚项 γv 的计算过程如式 (4) 所示:

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2$$

$$\gamma = \frac{v}{(1 - \text{IoU}) + v} \quad (4)$$

其中, v 是衡量高宽比一致性的因子, γ 是折中参数。

为此,综合考虑预测框与真实框之间覆盖面积、框的中心点距离和高宽比一致性的损失函数 L_{CIoU} 如式 (5) 所示:

$$L_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \gamma v \quad (5)$$

相比于 YOLOX-Nano 模型, Rapider-YOLOX 模型采用了优化后的损失函数 L_{CIoU} 。这样,在模型训练时 L_{CIoU} 不仅可以更好地引导网络向着更优的方向进行收敛,还能进一步提高预测框对真实框的拟合程度。

3 实验

3.1 数据集及实验环境

为了验证模型的有效性和普遍适用性,选取 PASCAL VOC 数据集作为实验数据集,该数据集是由 PASCAL VOC2007 和 PASCAL VOC2012 构成的混合数据集。实验时,将 VOC2007 中的 train 和 val 数据集(总计 5 011 张)与 VOC2012 中的 train 和 val 数据集(总计 11 540 张)进行合并作为训练集,并将 VOC2007 中的 test 数据集(总计 4 952 张)作为测试集。本文进行模型训练和精度指标评估的实验环境见表 1。

表 1 模型训练及精度指标评估实验环境

设备	型号
操作系统	Windows 10
处理器	Intel(R) Xeon(R) Silver 4214 CPU @ 2.20 GHz
显卡(训练)	GeForce GTX 2080TI
显卡(测试)	GeForce GT 1030
深度学习框架	Pytorch1.9.0
库版本	CUDA11.6; CUDNN7.4.1

3.2 实验结果分析

在训练阶段,本实验使用梯度下降的方式更新模型的权重。同时,为了增强网络模型的鲁棒性,实验中使用了 Mosaic 和 MixUp 两种数据增强方法。训练参数中的 batch_size 设置为 16,初始学习率为 0.01。本实验采用余弦退火衰减方法作为学习率的调整策略,使网络在训练过程中实现更有效的收敛,实验训练 500 个世代(epoch)。

本文算法与原始 YOLOX-Nano 在 500 个世代训练过程中的损失变化曲线分别如图 7 (a) 和图 7 (b) 所示。相较于原始 YOLOX-Nano,本文算法的损失

下降曲线更加光滑, 模型参数的优化过程更加平稳。此外, 从网络模型在验证集上的收敛效果来看, 原始 YOLOX-Nano 在训练 300 个 epoch 附近时, 其损失的下降便趋于平稳, 不再有明显的下降。而本文算法的损失在 300 个 epoch 时仍在持续下降, 直到迭代了 400 个 epoch 后才趋于平稳。由此可见, 相比于原始 YOLOX-Nano 模型, 本文改进后的模型在训练过程中能够更好地收敛。此外, 从图 7 可以看出, 在模型损失下降的过程中, 本文算法在验证集上的损失下降的表现明显优于 YOLOX-Nano。这说明本文算法提升了 YOLOX-Nano 的泛化性能。

Rapider-YOLOX 的检测效果如图 8 所示, 从图 8 可以观察到, 类似于当前主流目标检测网络, 对于日常生活中大部分场景中的目标检测任务,

Rapider-YOLOX 均能实现目标的准确识别。除了可以识别图中较大尺寸的马 (horse)、椅子 (chair)、车辆 (car) 等目标之外, 相比于其他轻量级网络, Rapider-YOLOX 能够更加准确地识别图中尺寸较小的人物 (person) 等目标。这是因为 Rapider-YOLOX 网络中的 HEB 模块有更好的特征提取能力, 而 SSPP 模块则有更强的特征融合能力和多尺度的识别能力。此外, 在图 9 (a) 中检测较小目标及图 9 (b) 中光线较暗的场景下, YOLOX-Nano 容易出现漏检情况, 而 Rapider-YOLOX 得益于 HEB 模块及 SSPP 模块的改进, 上述场景下的漏检现象明显减少。

为了直观地验证本文模型的有效性, 在实验中使用热力图的方式对包含位置信息的特征图进

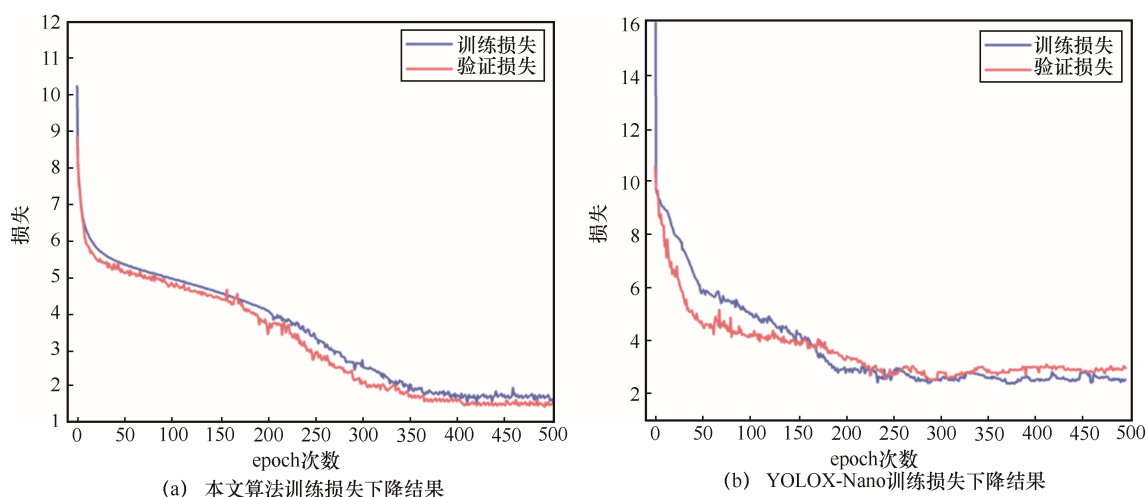


图7 本文算法与 YOLOX-Nano 训练损失下降对比



图8 Rapider-YOLOX 的检测效果

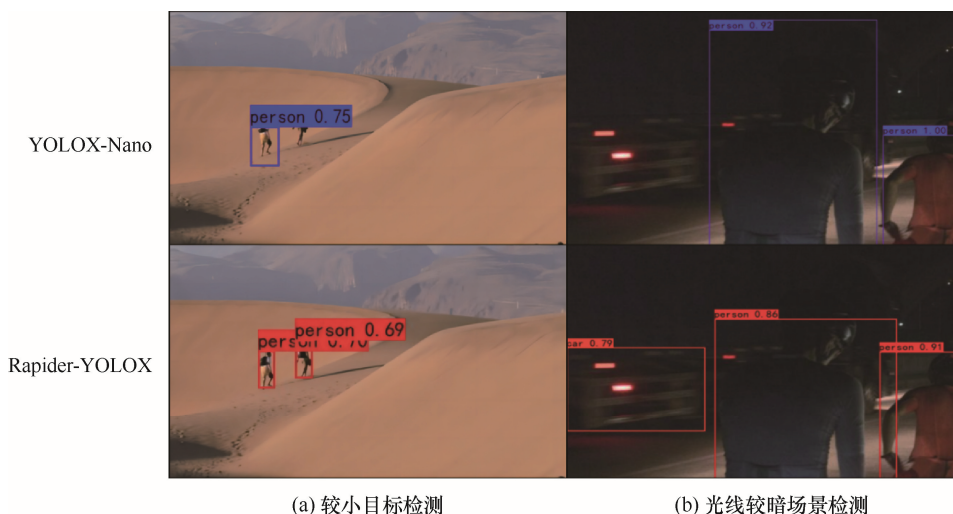


图 9 YOLOX-Nano 和 Rapider-YOLOX 检测效果

行可视化,以更直观地展示网络如何充分利用特征图及特征信息。从图 10 的实验结果可以直观地看出,图 10 (a) 组中采用 YOLOX-Nano 进行目标检测时遗漏了自行车上的人物目标,而 Rapider-YOLOX 则全面覆盖了所需检测的目标。从图 10 (b) 组对于猫的检测结果可以看出,相比于 YOLOX-Nano 网络, Rapider-YOLOX 更精准地关注到了猫的主要特征,避免了非重要特征对检测结果的干扰。由此可见,得益于本文对原始 YOLOX-Nano 网络中存在的信息丢失及特征提取能力不足等问题的改进, Rapider-YOLOX 可以在全面关注目标信息的同时更加充分地利用与目标相关程度高的特征信息。

为了进一步测试 Rapider-YOLOX 相比于

YOLOX-Nano 在性能上的提升效果,在这两个网络上分别测试了 VOC2007 的每个类的精确度 (precision, P)、召回率 (recall, R)、平均精度 (mean average precision, mAP) 以及 F1 值,评估指标见表 2。由表 2 可知,除了 sofa 类上的 AP 值, Rapider-YOLOX 在 VOC 2007 数据集的其他类别上的 AP 值均高于 YOLOX-Nano,且从表 2 列举的各项指标来看, Rapider-YOLOX 在绝大多数类别上的各项指标总体上明显优于 YOLOX-Nano。即使在 bicycle、boat 和 dining table 等极少数类别上 Rapider-YOLOX 网络的精确度或召回率略低于 YOLOX-Nano,但是其平均精度仍然高于 YOLOX-Nano。由此可见,本文对 YOLOX-Nano 网络结构的改进有效提升了网络的目标检测精度。

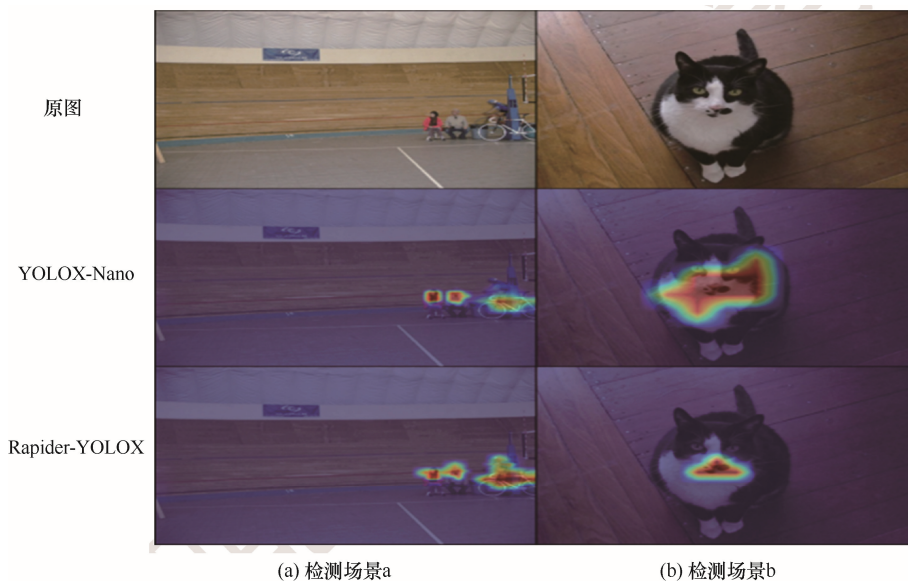


图 10 YOLOX-Nano 和 Rapider-YOLOX 热力图对比

表 2 Rapider-YOLOX 与 YOLOX-Nano 的评估指标

类别	YOLOX-Nano				Rapider-YOLOX			
	P	R	AP	F1	P	R	AP	F1
aeroplane	87.27%	80.00%	87.64%	0.83	89.38%	84.17%	90.78%	0.87
bicycle	90.68%	75.35%	82.65%	0.82	90.16%	77.46%	84.68%	0.83
bird	88.98%	64.02%	76.07%	0.74	89.76%	69.51%	80.47%	0.78
boat	77.27%	46.79%	61.75%	0.58	77.03%	52.29%	67.75%	0.62
bottle	77.14%	36.82%	49.12%	0.50	80.51%	43.18%	57.99%	0.56
bus	84.55%	75.00%	86.30%	0.79	89.29%	80.65%	89.08%	0.85
car	87.82%	69.98%	81.84%	0.78	88.40%	72.23%	84.06%	0.80
cat	81.25%	78.14%	81.56%	0.80	85.12%	78.14%	85.99%	0.81
chair	77.73%	39.90%	55.84%	0.53	82.74%	45.50%	61.47%	0.59
cow	71.64%	60.76%	63.15%	0.66	77.27%	64.56%	78.41%	0.70
dining table	73.97%	50.00%	56.98%	0.60	73.24%	48.15%	57.40%	0.58
dog	86.45%	63.36%	79.36%	0.73	86.49%	65.75%	82.25%	0.75
horse	88.50%	75.19%	87.64%	0.81	90.68%	80.45%	89.13%	0.85
motorbike	92.71%	68.99%	79.32%	0.79	93.88%	71.32%	84.15%	0.81
person	90.56%	72.53%	84.47%	0.81	91.08%	74.84%	86.16%	0.82
pottedplant	75.26%	39.89%	51.51%	0.52	86.67%	42.62%	58.74%	0.57
sheep	81.65%	70.49%	83.46%	0.76	86.71%	74.86%	85.97%	0.80
sofa	81.48%	64.71%	72.63%	0.72	73.68%	54.90%	69.85%	0.63
train	86.27%	72.73%	80.81%	0.79	91.00%	75.21%	82.60%	0.82
tvmonitor	86.79%	69.17%	80.52%	0.77	82.50%	74.44%	81.46%	0.78

Rapider-YOLOX 在全面提升 YOLOX-Nano 检测精度的同时,保留了网络轻量级的性能。为此,为了定量分析 Rapider-YOLOX 网络在低计算资源场景下的性能,在 VOC2007 数据集上对比了 Rapider-YOLOX 与当前主流轻量级模型的性能。GT1030、Titan X 及 Tesla V100 这 3 种类型 GPU 的主要参数见表 3,显然只有 384 个 CUDA 核心和 2 GB 内存容量的 GT1030,计算资源要远低于另外两种 GPU。

表 3 各推理设备配置参数详情

推理显卡	内存容量/GB	CUDA 核心数量
GT1030	2	384
Titan X	12	3 584
Tesla V100	16	5 120

Rapider-YOLOX 与当前另外 10 种主流轻量级网络的参数规模及其在 VOC2007 数据集上的测试结果见表 4。由表 4 可知,从检测速度的角度来看,

表 4 Rapider-YOLOX 与当前主流轻量级模型测试结果

网络模型	模型参数量/MB	模型复杂度/BFLOPS	平均检测精度	推理时延/ms	推理显卡型号
Tiny YOLOv2	60.5	6.97	57.1%	45.26	Tesla V100
Tiny YOLOv3	33.4	5.47	58.4%	34.30	Tesla V100
MobileNet-SSD ^[39]	22.2	4.29	72.7%	—	—
Tinier-YOLO	8.9	2.56	65.7%	39.84	Titan X
YOLO-Nano ^[40]	4.0	4.51	69.1%	29.74	Tesla V100
YOLOv5n	1.9	2.09	73.9%	27.62	GT1030
PP-PicoDet-S ^[41]	0.99	1.24	77.7%	—	GT1030
PPYOLO-Tiny ^[42]	4.20	4.95	71.3%	29.93	GT1030
PP-YOLOE-S ^[43]	7.90	7.56	78.4%	42.76	GT1030
YOLOX-Nano	0.91	1.07	74.1%	21.85	GT1030
Rapider-YOLOX	0.96	1.10	77.9%	22.02	GT1030

相比于其他网络,运行于计算资源最少的 GT1030 平台上的 Rapider-YOLOX 和 YOLOX-Nano 网络有着最快的推理速度。虽然 Rapider-YOLOX 的正向推理时延比最快的 YOLOX-Nano 慢了 0.17 ms,但是其平均精度却比 YOLOX-Nano 提升了 3.8%。从平均检测精度的角度来看,PP-YOLOE-S 的检测精度比 Rapider-YOLOX 高 0.5%,但在检测任务的正向推理上则需要更多的额外时间开销。事实上,由表 4 可知,Rapider-YOLOX 在速度和精度的平衡上有着最佳的表现。

此外,由表 4 可知,除 YOLOX-Nano 外的其他主流轻量级网络体积较大、推理时间较长,只有在计算资源十分充足的情况下才勉强达到实时的效果。而 Rapider-YOLOX 与 YOLOX-Nano 一样有着很小的参数规模,这使 Rapider-YOLOX 仅需 1.10 BFLOPS 的计算量,即使在计算资源较少的 GT1030 平台上,其每秒帧数(frames persecond, FPS)也能达到 45.41。由此可见,由于对硬件资源的需求低且目标识别时精度高,Rapider-YOLOX 网络在检测速度和检测精度的性能平衡上有着最佳的效果。这样,即使在以嵌入式设备为硬件载体的真实应用场景下,Rapider-YOLOX 也可以实时地完成检测任务。

为了验证本文模型在计算资源更加匮乏的设备上的运行情况,将训练好的模型和权重迁移至仅包含单 CPU 的主机上进行实验,同时配置相同版本的 CPU 环境。Rapider-YOLOX 在 CPU 上的检测速度见表 5, Rapider-YOLOX 在单 CPU 环境下的 FPS 值达到了 23.94,甚至超过了 Tiny YOLOv2 在 Tesla V100 上的检测速度。可见, Rapider-YOLOX 即使在没有 GPU 计算资源的单 CPU 环境下也可以满足较低的实时性需求。

表 5 Rapider-YOLOX 在 CPU 上的检测速度

网络模型	推理延迟/ms	FPS	推理硬件型号
Rapider-YOLOX	41.80	23.94	i5-12600KF
Tiny YOLOv2	45.26	22.09	Tesla V100

3.3 消融实验

本节通过消融实验验证本文网络模型中每个模块设计的有效性。采用 YOLOX-Nano 作为基线网络,并将每个模块设计依次附加在基线网络上,以验证该设计的检测效果。本实验仍然使用 PASCAL VOC2007 作

为验证集,保持训练策略不变,硬件设备为 GT1030。

消融实验结果见表 6,本文提出的网络设计模块在提升网络检测精度上有显著的效果。在使用本文提出的 HEB 和 SSPP 模块后,相较于最先进的 YOLOX-Nano 网络,本文模型的性能可以分别提升约 1.99%和 1.41%,而引入 CIOU 也可以使网络精度提升 0.99%。相比于 YOLOX-Nano 网络, Rapider-YOLOX 的精度领先 3.79%,同时对网络检测速度几乎没有影响。可见,本文提出的模型设计方法相较于原始 YOLOX-Nano 模型,在几乎不影响检测速度的条件下,增强了网络模型的特征提取能力及对特征信息充分利用的能力,确保了网络整体性能的显著提升。

表 6 消融实验结果

网络模型	mAP	FPS
YOLOX-Nano	74.13%	45.76
YOLOX-Nano +HEB	76.12%	45.16
YOLOX-Nano +SSPP	75.54%	45.96
YOLOX-Nano +CIOU	75.12%	—
Papider-YOLOX	77.92%	45.40

4 结束语

本文提出的 Rapider-YOLOX 在 PASCAL VOC2007 上的 mAP 可达 77.92%,相较于改进前的 YOLOX-Nano, Rapider-YOLOX 的平均精度提升了 3.79%。Rapider-YOLOX 网络的参数量仅为 0.96 MB,基本与 YOLOX-Nano 持平。这使 Rapider-YOLOX 不仅能在低性能的 GPU 上高效运行,甚至可以在单 CPU 上实现端到端的实时检测。实验结果表明,与目前主流的轻量化网络相比, Rapider-YOLOX 在检测精度和检测速度方面实现了最佳的性能平衡,在移动端或嵌入式平台等低计算资源的目标检测工业场景下有很好的应用前景。

然而, Rapider-YOLOX 模型的参数量相对较少,特征提取过程中的抗干扰能力有限。当场景过于复杂或者场景中干扰因素过多时,网络的目标检测性能可能会受到影响,尤其是在密集场景下小物体主要部位被遮挡, Rapider-YOLOX 的目标检测精度会有所下降。为此,在后续模型改进中可以考虑引入有助于改善小目标检测的设计思想,进一步提升网络的泛化性能,以提升网络在复杂场景中的适应性。

参考文献:

- [1] VOULODIMOS A, DOULAMIS N, DOULAMIS A, et al. Deep learning for computer vision: a brief review[J]. *Computational Intelligence and Neuroscience*, 2018: 1-13.
- [2] 张红民, 李萍萍, 房晓冰, 等. 改进 YOLOv3 网络模型的人体异常行为检测方法[J]. *计算机科学*, 2022, 49(4): 233-238.
ZHANG H M, LI P P, FANG X B, et al. Human abnormal behavior detection method based on improved YOLOv3 network model[J]. *Computer Science*, 2022, 49(4): 233-238.
- [3] 田庆, 胡蓉, 李佐勇, 等. 基于 SE-YOLOv5s 的绝缘子检测[J]. *智能科学与技术学报*, 2021, 3(3): 312-321.
TIAN Q, HU R, LI Z Y, et al. Insulator detection based on SE-YOLOv5s[J]. *Chinese Journal of Intelligent Science and Technology*, 2021, 3(3): 312-321.
- [4] VIOLA P, JONES M. Robust real-time face detection[C]//*Proceedings of the 8th IEEE International Conference on Computer Vision*. Piscataway: IEEE Press, 2002.
- [5] VEDALDI A, GULSHAN V, VARMA M, et al. Multiple kernels for object detection[C]//*Proceedings of 2009 IEEE 12th International Conference on Computer Vision*. Piscataway: IEEE Press, 2010: 606-613.
- [6] HARZALLAH H, JURIE F, SCHMID C. Combining efficient object localization and image classification[C]//*Proceedings of 2009 IEEE 12th International Conference on Computer Vision*. Piscataway: IEEE Press, 2010: 237-244.
- [7] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(7): 971-987.
- [8] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2014: 580-587.
- [9] GIRSHICK R. Fast R-CNN[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*. Piscataway: IEEE Press, 2016: 1440-1448.
- [10] REN S Q, HE K M, 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.
- [11] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Piscataway: IEEE Press, 2017: 2980-2988.
- [12] DAI J F, LI Y, HE K M, et al. R-FCN: object detection via region-based fully convolutional networks[C]//*Proceedings of the 30th International Conference on Neural Information Processing Systems*. New York: ACM Press, 2016: 379-387.
- [13] ZHU Y S, ZHAO C Y, WANG J Q, et al. CoupleNet: coupling global structure with local parts for object detection[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Piscataway: IEEE Press, 2017: 4146-4154.
- [14] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2016: 779-788.
- [15] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2017: 6517-6525.
- [16] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. *arXiv preprint*, 2018, arXiv: 1804.02767.
- [17] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[J]. *arXiv preprint*, 2020, arXiv: 2004.10934.
- [18] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//*Proceedings of European Conference on Computer Vision*. Cham: Springer, 2016: 21-37.
- [19] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Piscataway: IEEE Press, 2017: 2999-3007.
- [20] GE Z, LIU S T, WANG F, et al. YOLOX: exceeding YOLO series in 2021[J]. *arXiv preprint*, 2021, arXiv: 2107.08430.
- [21] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2022: 11966-11976.
- [22] RADOSAVOVIC I, KOSARAJU R P, GIRSHICK R, et al. Designing network design spaces[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2020: 10425-10433.
- [23] STERGIOU A, POPPE R, KALLIATAKIS G. Refining activation downsampling with SoftPool[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*. Piscataway: IEEE Press, 2022: 10337-10346.
- [24] WANG Q L, WU B G, ZHU P F, et al. ECA-net: efficient channel attention for deep convolutional neural networks[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2020: 11531-11539.
- [25] YU J H, JIANG Y N, WANG Z Y, et al. UnitBox: an advanced object detection network[C]//*Proceedings of the 24th ACM International Conference on Multimedia*. New York: ACM Press, 2016: 516-520.
- [26] LIU Z, LI J G, SHEN Z Q, et al. Learning efficient convolutional networks through network slimming[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*. Piscataway: IEEE Press, 2017: 2755-2763.
- [27] ZHANG D Q, YANG J L, YE D, et al. LQ-nets: learned quantization for highly accurate and compact deep neural networks[C]//*Proceedings of 2018 European Conference on Computer Vision*. New York: ACM Press, 2018: 373-390.
- [28] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. *arXiv preprint*, 2015, arXiv: 1503.02531.
- [29] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. *arXiv preprint*, 2017, arXiv: 1704.04861.
- [30] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2018: 4510-4520.

- [31] HOWARD A, SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 1314-1324.
- [32] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6848-6856.
- [33] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]//Proceedings of 2018 European Conference on Computer Vision. New York: ACM Press, 2018: 122-138.
- [34] HAN K, WANG Y H, TIAN Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 1577-1586.
- [35] WOMG A, SHAFIEE M J, LI F, et al. Tiny SSD: a tiny single-shot detection deep convolutional neural network for real-time embedded object detection[C]//Proceedings of 2018 15th Conference on Computer and Robot Vision. Piscataway: IEEE Press, 2018: 95-101.
- [36] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size[J]. arXiv preprint, 2016, arXiv: 1602.07360.
- [37] FANG W, WANG L, REN P M. Tinier-YOLO: a real-time object detection method for constrained environments[J]. IEEE Access, 2019, 8: 1935-1944.
- [38] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J]. arXiv preprint, 2020, arXiv: 2010.11929.
- [39] LI Y T, HUANG H S, XIE Q S, et al. Research on a surface defect detection algorithm based on MobileNet-SSD[J]. Applied Sciences, 2018, 8(9): 1678.
- [40] WONG A, FAMUORI M, SHAFIEE M J, et al. YOLO nano: a highly compact you only look once convolutional neural network for object detection[C]//Proceedings of 2019 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition.

Piscataway: IEEE Press, 2021: 22-25.

- [41] YU G H, CHANG Q Y, LYU W Y, et al. PP-PicoDet: a better real-time object detector on mobile devices[J]. arXiv preprint, 2021, arXiv: 2111.00902.
- [42] HUANG X, WANG X X, LYU W Y, et al. PP-YOLOv2: a practical object detector[J]. arXiv preprint, 2021, arXiv: 2104.10419.
- [43] XU S L, WANG X X, LYU W Y, et al. PP-YOLOE: an evolved version of YOLO[J]. arXiv preprint, 2022, arXiv: 2203.16250.

[作者简介]



顾宙瑜(1996-), 男, 江苏科技大学计算机学院硕士生, 主要研究方向为计算机视觉、深度学习。



於跃成(1971-), 男, 博士, 江苏科技大学计算机学院副教授, 主要研究方向为计算机视觉、深度学习。



者甜甜(1997-), 女, 江苏科技大学计算机学院硕士生, 主要研究方向为计算机视觉、深度学习。