目标检测模型及其优化方法综述

蒋弘毅 [1] 王永娟 [1] 康锦煜 [1]

摘 要 近年来, 基于卷积神经网络的目标检测研究发展十分迅速, 各种检测模型的改进方法层出不穷. 本文主要对近几年内目标检测领域中一些具有借鉴价值的研究工作进行了整理归纳. 首先, 对基于卷积神经网络的主要目标检测框架进行了梳理和对比. 其次, 对目标检测框架中主干网络、颈部连接层、锚点等子模块的设计优化方法进行归纳, 给出了各个模块设计优化的基本原则和思路. 接着, 在 COCO 数据集上对各类目标检测模型进行测试对比, 并根据测试结果分析总结了不同子模块对模型检测性能的影响. 最后, 对目标检测领域未来的研究方向进行了展望.

关键词 卷积神经网络, 目标检测, 子模块优化,

引用格式 蒋弘毅, 王永娟, 康锦煜. 目标检测模型及其优化方法综述. 自动化学报, 2021, **47**(6): 1232−1255

DOI 10.16383/j.aas.c190756

A Survey of Object Detection Models and Its Optimization Methods

JIANG Hong-Yi[1] WANG Yong-Juan[1] KANG Jin-Yu[1]

Abstract In recent years, research on object detection based on convolutional neural network has developed rapidly, and various improved algorithms and models have emerged one after another. This paper mainly summarizes some recent valuable research work in the field of object detection. Firstly, main object detection framework based on convolutional neural network is analyzed and compared. Secondly, the optimization methods of backbone, neck, anchors and other sub-modules are summarized, and the basic principles and ideas for the design and optimization of each modules are given. Thirdly, various objection detection models are tested and compared on the COCO dataset, effects of different sub-modules on the detector performance were analyzed according to the test results. Finally, future research direction in object detection is prospected.

Key words Convolutional neural network, object detection, sub-module optimization

Citation Jiang Hong-Yi, Wang Yong-Juan, Kang Jin-Yu. A survey of object detection models and its optimization methods. *Acta Automatica Sinica*, 2021, **47**(6): 1232−1255

目标检测是指利用计算机工具和相关算法来对现实世界中的对象进行分类和定位的一类计算机视觉技术.

传统的目标检测需要手工提取特征[1−3], 并针对特定检测对象设计和训练分类器. 这类方法难以获得鲁棒性强的特征, 对外界环境噪声十分敏感, 故在工程应用上具有较大的局限性.

现阶段, 随着深度学习技术的不断发展和硬件设施的不断进步, 基于卷积神经网络的目标检测技术发展迅速. 以区域卷积神经网络[4−6] (Region convolutinal neural network, R-CNN) 系列为代表的两阶段法, 与以单阶段检测器[7] (Single-shot detector, SSD) 和 YOLO[8−10] (You only look once) 系列为代表的单阶段法是当前基于卷积神经网络的目标检测技术的两种主流框架.

近年来, 不少学者在两种框架的基础上对内部的主干网络、锚点设计、区域特征编码等子模块进行了优化改进, 有效地提高了目标检测算法的性能. 部分学者还提出了一种基于对象关键点的目标检测框架[11−14], 并在各大数据集上取得了惊人的成绩.

本文针对近年来目标检测算法的最新研究进展, 从目标检测框架的子模块设计优化角度出发, 对该领域中一些有启发性的研究成果进行整理、归纳和分析, 并对目标检测模型的优化思路提出一些建议, 以便众多相关研究者参考和借鉴.

1 基于卷积神经网络的目标检测框架

目前, 主要使用的目标检测框架分为两阶段和单阶段两类. 两种框架在结构上 (图 1) 的最大区别是: 两阶段框架通过区域推荐网络 (Region propos-

收稿日期 2019-11-01 录用日期 2020-02-16
Manuscript received November 1, 2019; accepted February 16, 2020
军委创新项目 (18-163-11-ZT-005-032-01) 资助
Supported by Innovation Project of CMC (18-163-11-ZT-005-032-01)
本文责任编委 白翔
Recommended by Associate Editor BAI Xiang
1. 南京理工大学机械工程学院 南京 210094
1. College of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094

als, RPN) 类的子网络来辅助生成推荐框 (Proposals), 而单阶段框架直接在特征图上生成候选框.

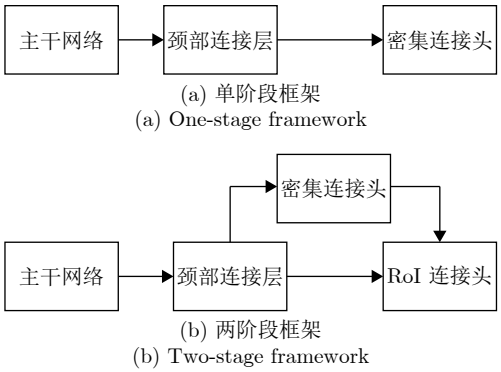


图 1　主流的目标检测框架

Fig. 1　Main object detection framework

输入的图片经过由卷积神经网络组成的主干网络 (Backbone) 后, 输出整张图片的特征图 (Feature map), 通过颈部 (Neck) 连接层对不同尺度的特征图进行融合, 以获得多尺度的特征. 此后, 单阶段方法一方面对锚点框 (Anchor box) 进行分类, 另一方面直接在不同尺度的特征图上对正样本预测边界框 (Bounding box) 的位置补偿, 最后使用非极大值抑制得到检测结果; 两阶段方法先利用区域推荐网络对锚点框进行分类和回归得到推荐框, 对其进行特征编码后再做分类和回归, 最后经过非极大值抑制完成对目标的检测.

传统上, 由于两阶段的目标检测框架相比于单阶段的目标检测框架多进行了一次分类和回归, 故在检测的准确率和召回率上都要高出较多. 相反, 单阶段目标检测框架直接在特征图上对正例锚点框进行分类和回归, 算法复杂度较小, 在检测速度上有明显优势. 但近些年来, 众多学者针对两种框架的各自缺点进行了相应改进, 使部分单阶段与两阶段模型在检测性能与速度上的差异逐渐缩小[15−17].

最近, Law 等[11] 借鉴人体关键点检测的思路, 首次提出了一种基于关键点的目标检测模型 Cornernet (图 2), 该框架与上述两类框架的最大区别是: 不再通过微调锚点框来对目标进行定位, 而是直接对目标关键点进行回归, 训练和预测边界框的位置和大小. 输入图片经过卷积网络提取特征后, 输出两组热图 (Heat map) 来预测目标边界框的左上角点和右下角点; 每个热图有 C 个 (C 为类别数) 通道, 用来预测目标的类别. 通过嵌入向量隐式表达两类角点间的距离, 完成对图上所有预测角点的两两匹配, 得到最终的目标边界框.

除了选取角点作为物体关键点进行目标检测外, Zhou 等[13] 将目标中心点作为关键点, 采用热图预测目标中心点, 并对边界框的大小直接进行回归. 与使用角点作为关键点相比, 该方法无需对关键点进行匹配分组, 检测速度有了较大提升. Yang 等[14] 通过预测不同物体的多组点对, 使模型更充分地学习到相应目标的几何、语义、姿态等利于检测任务的特征, 有效提高了模型分类与定位的能力.

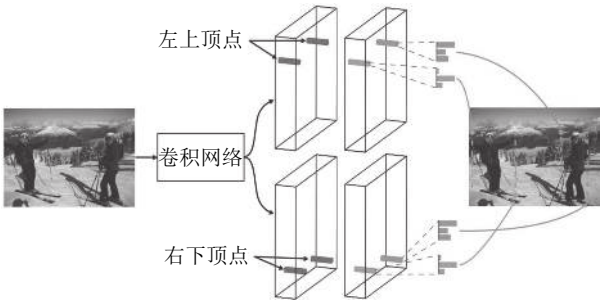


图 2　CornerNet 框架流程

Fig. 2　Overall pipeline of CornerNet

基于关键点的目标检测框架与单阶段检测框架相比, 在检测速度上几乎没有劣势, 而在检测性能上远超传统的单阶段检测框架, 甚至比没有经过优化的两阶段框架还要出色 (图 3). 同时, 其设计思路上非常直接, 得到的模型也更为简洁; 扩展性极强, 能方便地应用到三维目标检测、人体姿态估计等其他计算机视觉任务中.

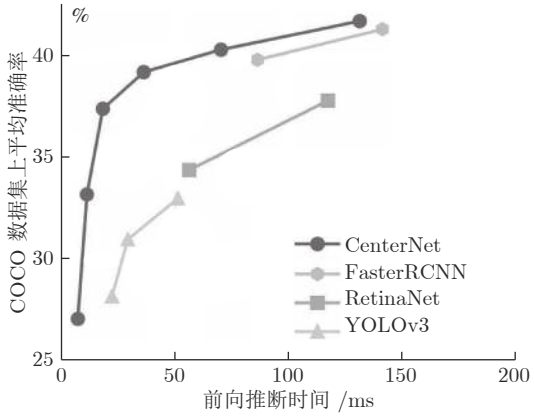


图 3　典型目标检测算法速度−准确率对比

Fig. 3　Speed-accuracy comparison of typical object detection algorithms

2　目标检测框架的子模块优化

目标检测框架一般都包含: 主干网络、颈部连接层、锚点、区域特征编码、分类与定位头部和损失函数等子模块. 此外, 不同模型还有自己独特的子模块. 对上述子模块进行合理的优化可以有效地提

高目标检测模型的性能.

2.1 主干网络与颈部连接层优化

主干网络子模块位于输入层开始到具体的下游任务层前, 用于提取目标的不同尺度特征.

早期的基于卷积神经网络的目标检测模型 (如 SSD 模型) 大多以 VGG 网络[18] 作为主干网络, 该网络通过对卷积层和池化层进行反复堆叠, 提高特征提取和语义表达能力. 然而该网络的层数仅仅只有 19 层, 提取的特征表达能力有限. 若仅通过叠加的方法加深网络层数, 则梯度在网络中传递时很容易出现消失或者爆炸, 这反而降低了网络的性能.

为解决深度网络梯度消失和爆炸的问题, He 等[19] 提出了跳连的残差 (ResNet) 网络结构 (图 4), 它将浅层的特征信息与后面层进行融合, 生成新的特征向后传递. 该方法有效保证了特征信息向深层网络中传递, 提高了深层网络的性能.

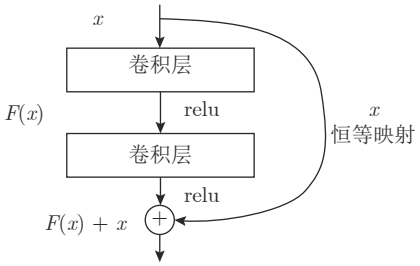


图 4 残差网络的跳连结构

Fig. 4 Shortcut structure of ResNet

采用残差结构增加网络深度虽然能够有效提高卷积网络的性能, 但带来的参数量增加也是成倍的. 为此, Xie 等[20] 将残差结构和 Inception 结构[21] 进行整合 (图 5), 同时增加网络的深度和宽度 (宽度方向参数共享), 在提高网络性能的同时有效控制了参数量的增长.

同一层特征图的不同通道蕴含的特征信息是不

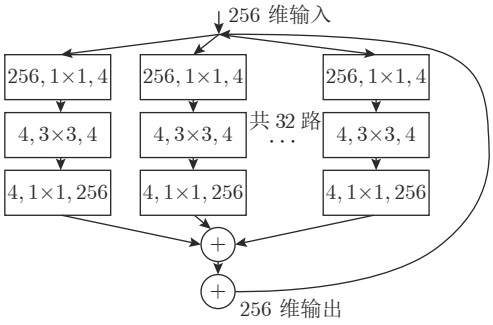


图 5 基数为 32 的 ResNeXt 块

Fig. 5 ResNeXt block with 32 cardinality

同的, Hu 等[22] 对特征图内不同通道间的特征进行融合 (图 6), 使模型学习到每个特征通道的重要程度. 这种方法显式地建立了各个通道特征间的关系, 有助于更好地提取目标特征.

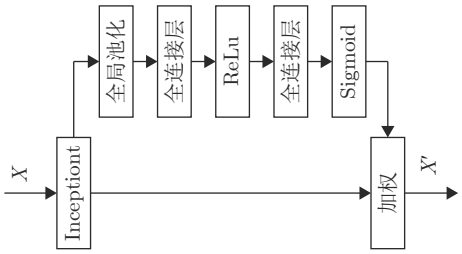


图 6 特征通道融合的 Inception 模块

Fig. 6 The schema of SE-Inception module

除了对特征通道间关系进行建模外, 特征图同一通道的不同位置上也存在着紧密联系. Wang 等[23] 显式地建立了图上任一点与全局其他位置的关系, 提出了图像自注意力机制的非局部网络 (Non-local network) (图 7). 该方法能有效地捕捉不同位置特征的空间联系, 从而增强了目标特征. Cao 等[24] 对非局部网络进行简化, 并同时考虑了特征通道间的关系, 使主干网络在提取目标特征的性能和速度上都有了提高. 相似地, Woo 等[25] 提出了通道与空间上的卷积注意力模块, 为特征图上不同通道和不同位置上的特征赋予不同的权重, 提高了网络特征提取的能力.

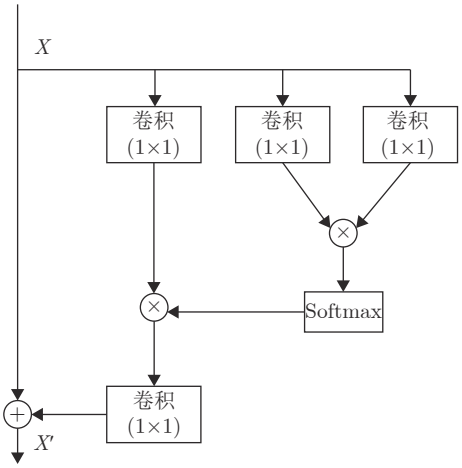


图 7 非局部网络块

Fig. 7 Block of non-local network

浅层卷积特征的感受野小, 缺少足够的语义信息, 但其通常有着较高的分辨率, 保留了较多位置信息; 深层卷积特征分辨率低, 目标位置不够精确, 但其感受野较大, 包含了丰富的语义信息.

颈部连接指将上述不同尺度的特征进行融合, 目的是生成同时具备高语义信息与精确位置信息的多尺度特征, 提高模型对不同尺度目标的检测能力.

最早将多特征融合技术应用在目标检测框架上的是特征金字塔网络 (Feature pyramid network, FPN) 模型[26]. 它采用金字塔式的层级结构将残差网络中的低分辨率特征层进行上采样, 并与相应尺度的原始特征层进行融合 (图 8), 输出信息更多、鲁棒性更强的多尺度特征.

图 8　FPN 中的金字塔结构

Fig. 8　Pyramid structure in FPN

为进一步增强融合后的多尺度特征, Pan 等[27] 在 FPN 的基础上提出了 BFP (Balanced feature pyramids) 颈部连接结构. 该算法将 FPN 提取出的多尺度特征通过线性插值和池化调整到中间尺度大小, 并对它们进行加权平均得到整合后的单一尺度特征, 再采用非局部网络[23] 加强该特征, 最后将其映射回与输入相同的尺度, 用于之后的目标检测.

FPN 的底部特征层包含较多的位置信息, Liu 等[28] 在自顶向下的特征金字塔结构的基础上, 又提出了自底向上的特征融合支路 (图 9), 将底层位置信息传递给高层特征, 进一步提高了模型的定位精度.

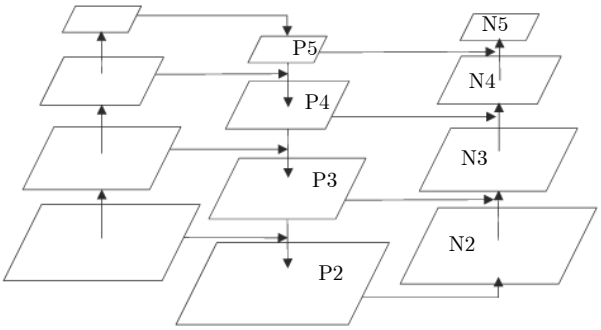


图 9　PANet 中的自底向上金字塔结构

Fig. 9　Bottom-up pyramid structure in PANet

为充分利用不同层的特征, Zhao 等[29] 对 FPN 进行堆叠, 提出了多层级的 FPN 颈部连接结构 (图 10). 它通过 U 型网络对特征进行编码–解码, 并从 SENet[22] 中受到启发, 对不同层间相同尺度的特征通道进行加权拼接, 最后利用得到的特征进行多尺度预测.

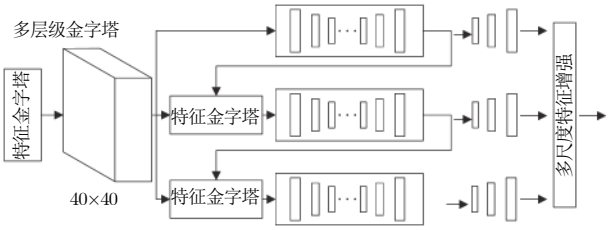


图 10　M2Det 中的多层级特征金字塔网络结构

Fig. 10　Multi-level feature pyramid network in M2Det

Tan 等[30] 认为不同尺度的特征层对目标特征的贡献应该是不同的, 并将没有进行融合的特征层从网络中剔除, 提出了加权融合的双向特征金字塔结构 (图 11), 从而进一步优化了不同特征层间的信息传递.

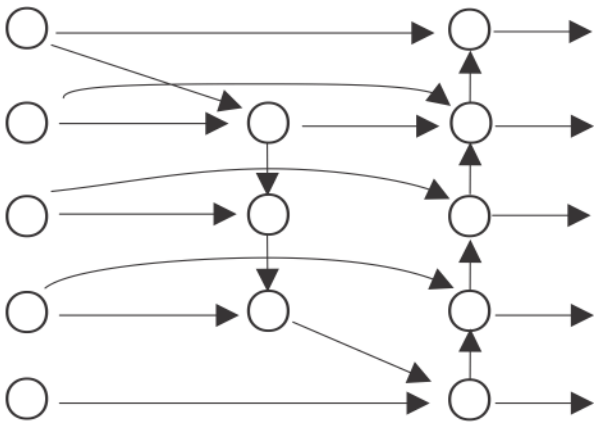


图 11　双向特征金字塔结构

Fig. 11　Framework of Bi-FPN

除了金字塔式的特征融合方法, Newell 等[31] 采用了一种 “沙漏式”的网络结构 (图 12), 它借鉴残差网络的思想, 每个经过池化后的特征层与经过上采样后相同尺度的特征层进行融合, 从而将浅层信息传递给深层特征, 提高深层特征的定位能力.

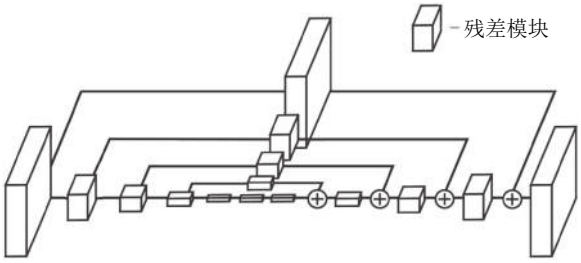


图 12　沙漏式结构的特征融合

Fig. 12　Feature fusion based on hourglass structure

无论是金字塔结构还是沙漏结构, 都是从高分辨特征图通过卷积和池化得到低分率特征图, 再从低分辨率特征图中通过上采样恢复到高分辨率特征图. 这类编码–解码 (Encoder-decoder) 的方式虽然可以很容易地实现特征融合, 但在尺度变换的过程中也无可避免地失去了部分细节, 对模型定位性能产生不利. 针对上述问题, Sun 等[32] 提出了始终保持高分辨率的高分辨率网络 (High-resolution net, HRNet). 它将高分辨率的子网络作为第一阶段, 在信息融合时不断增加低分辨率的特征层以形成更多的阶段, 并把这些不同尺度的特征层进行反复互连, 实现多尺度信息的交换, 最终输出高分辨特征图 (图 13).

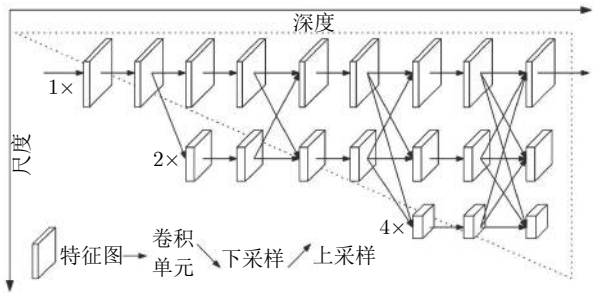


图 13 HRNet 的整体网络结构

Fig. 13 Overall network structure of HRNet

总之, 主干网络和颈部连接层的优化都是为了获得更加有利于模型分类与定位的特征, 而提取出高信息特征的关键在于:

1) 主干网络和颈部连接层应有一定的深度和宽度, 充分考虑特征图内部、不同特征图间的相互关系, 对特征图在维度、空间以及深度上进行合理建模, 从而充分地提取目标特征.

2) 为在位置与语义信息之间取得平衡, 主干网络应采用适当的下采样和上采样率. 通过使用合理的颈部连接结构或空洞卷积[33] 等方法, 保证输出的目标特征同时满足分类与定位的要求.

2.2 锚点设计的优化

锚点是在特征图上每个网格内生成的具有不同大小、比例的矩形框. 单阶段检测框架直接在锚点的基础上生成目标边界框; 两阶段检测框架通过微调正例锚点获得候选框.

最早使用锚点的模型是 Faster R-CNN[6]. 它在区域推荐子网络特征图上的每个网格中生成 3 种尺度、3 种比例共 9 个锚点框 (图 14). 根据锚点框与相对应实例框的交并比大小来确定其为正样本或负样本. 对正样本锚点框进行定位回归获得候选框, 用于进一步的定位和分类.

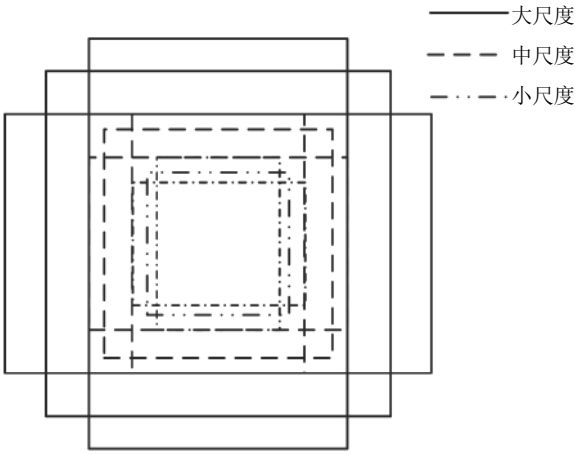


图 14 Faster R-CNN 中的锚点示意图

Fig. 14 Schematic diagram of anchors in faster R-CNN

不同尺度比例的锚点框适合检测不同物体. 为此, Zhu 等[34] 提出了基于步长缩减的锚点框设计策略. 高分辨率的特征图感受野小, 用于检测小尺度目标, 故为防止漏检现象发生, 应当缩减锚点框生成的步长来增加锚点框的密度. 低分辨率的特征图感受野大, 用于检测大尺度目标, 因而可以适当增大锚点框步长, 减少计算复杂度. Xie 等[35] 提出了一种维度可分解的区域推荐网络, 他将锚点在维度上进行分解, 使用一种锚点字符串 (Anchor string) 机制来独立地匹配目标的宽度和高度, 从而有效地解决了对比例特殊目标的检测.

Zhong 等[36] 创造性地提出通过训练让模型自动调整锚点框, 并设计了一条独立的分支用来预测和调整锚点框的形状. Wang 等[37] 在此基础上, 利用图像特征来指导锚点框的生成 (图 15), 将锚点框的生成分为位置预测和形状预测两个步骤. 通过位置和形状预测确定相应特征图的每个网格内是否存在锚点以及锚点框的长宽; 再利用可变形卷积[38−39] 对特征图进行修正, 以匹配锚点框. 这种做法大大减少了锚点的数量, 提高了锚点框对目标尺寸变化的适应性.

考虑到基于锚点的方法在分配正负样本以及处理多尺度问题上的局限性, 很多学者提出了无锚点的目标检测模型. 这类模型大多在不同尺度的特征图上进行像素级的分类和回归来代替锚点框的功能. Tian 等[40] 先计算特征图上每个点映射回原图的位置, 再根据该位置是否位于相应实例框内进行正负样本的分配, 并且定义中心度来降低实例框边缘位置预测时的分数权重, 从而抑制了低质量预测框对检测结果的影响, 提高了模型的检测性能. Kong

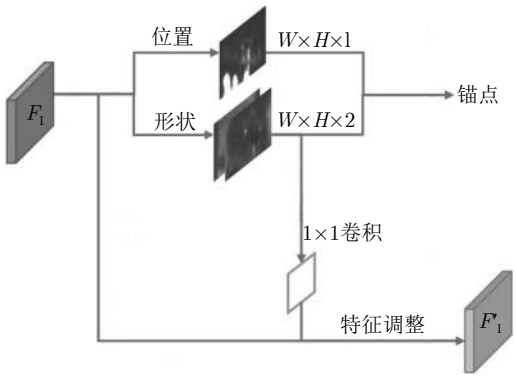


图 15　基于特征指导的锚点生成模型

Fig. 15　Anchor generation model based on feature guiding

等[41] 从人体眼球的中央凹 (Fovea) 结构中获得灵感, 通过参数调整实例框宽高来确定正负样本: 向物体中心缩放, 缩放后框内所有样本为正样本点: 向物体外扩大, 扩大化框外所有样本为负样本点 (图 16); 忽略两个边界框范围内的点. 这种做法增加了正负样本间的差异, 有利于分类问题的学习.

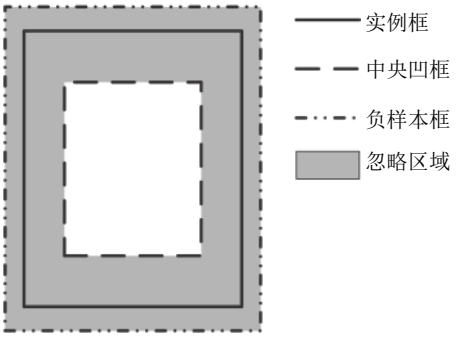


图 16　FoveaBox 模型中的标签分配

Fig. 16　Label assign in FoveaBox

利用先验尺度范围将目标分配给不同特征层的做法, 本身就是非最优的. 针对该问题, Zhu 等[42] 提出了在线特征层选择机制 (图 17): 训练阶段, 在所有尺度的特征层上进行分类和回归训练, 通过最小化焦点损失和交并比损失来选择最佳的特征层; 推理阶段, 直接选择置信度最高的特征层用于检测. 通过自动选择最佳特征, 该模型有效避免了手工选择特征层的一些弊端.

对于尺度适中的目标而言, 相邻特征层间有着相似的特性, 将其分给指定一个特征层的做法不够合理. 因而, Zhu 等[43] 又在 FSAF 模型的基础上提出了软分配的方法 (图 18), 它通过预测目标在不同尺度特征层以及同一特征层不同位置上的损失权重, 来考虑特征层之间、同一特征层上不同位置之

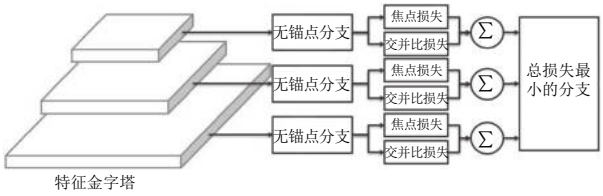


图 17　FSAF 模型的在线特征选择

Fig. 17　Online feature selection in FSFA

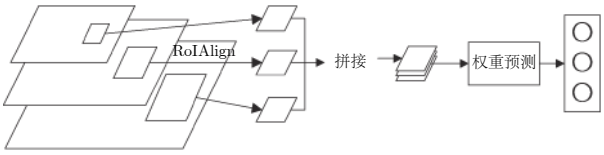


图 18　软分配的层权重预测

Fig. 18　Weights prediction for soft-selected features

间的关系, 进而计算出训练时的总损失大小.

由于正例锚点框给出了目标大致的初始位置和尺寸, 降低了模型学习分类和回归的难度, 这对于想要获得高性能的检测模型来说是有很大帮助的. 然而, 基于锚点的检测方法大多通过锚点框与相应实例框的交并比阈值来确定正负样本, 这对锚点的设计要求非常高. 相比之下, 无锚点方法对目标先验知识的要求低, 但设计过程相对繁琐. 总之, 锚点的设计及优化应遵循以下几点原则:

1) 锚点框的比例、大小范围应根据具体数据集或任务来确定. 当待检测对象的尺度变化很大时, 应采用多尺度的特征图, 设计多尺度的锚点.

2) 锚点框生成的密度应保证在不漏检的前提下尽量小, 以减少计算量. 对于小目标应增大锚点框生成的密度, 而对于大目标可以适当降低密度.

3) 锚点框在任何尺度的特征图上, 其中心都要与对应网格的中心尽可能对齐, 以保证锚点框从特征图回到原图时, 不发生位置上的偏移.

4) 对于无锚点模型, 应该重点关注如何在没有锚点的情况下更合理地分配与设置正负样本, 以及如何更有效地处理不同尺度物体带来的多尺度特征, 从而促进模型更好地从训练数据中学习.

2.3　非极大值抑制算法的优化

非极大值抑制在目标检测中是指模型在前向推理阶段, 选择置信度最高的候选框作为检测结果, 而剔除与其交并比大于阈值的周围所有候选框的一种算法. 这种方法对于同一个待检测对象, 可以排除其余非最优的候选结果, 避免出现重定位的问题.

非极大值抑制是几乎所有基于卷积神经网络的目标检测模型都使用到的方法. 经典的非极大值抑

制是直接剔除非最优的候选结果, 这种做法在图像上待检测目标是非密集的情况下有着出色的效果. 而在目标密集的场合, 目标之间相互遮挡, 属于同一类的多个目标非常靠近时, 非极大值抑制只保留置信度最高的候选框, 从而产生漏检.

为解决上述问题, Bodla 等[44] 提出软抑制 (Soft NMS) 的算法. 它对于前 n 个置信度大小的非最优候选框通过降低置信度来代替直接剔除的做法:

$$S_i = \begin{cases} S_i, & IoU(M, b_i) < N_t \\ S_i(1 - IoU(M, b_i)), & IoU(M, b_i) \geq N_t \end{cases} \tag{1}$$

其中, S 表示置信度得分, $IoU(M, b_i)$ 表示最优候选框 M 与其余某候选框 b_i 的交并比, N_t 表示进行软抑制的阈值.

从式 (1) 中可以看出, 软抑制算法认为当非最优候选框与最优候选框之间越接近, 则其越有可能是冗余的, 置信度分数也越低. 经过更新后的置信度若低于正样本置信度的阈值, 则该候选框就被剔除. 经过软抑制后的模型在密集场景下的检测召回率有了一定的提高.

候选框的置信度与交并比并非强相关, 只考虑分类置信度是片面的. He 等[45] 在软抑制算法的基础上加以改进, 同时将定位置信度融入到其中, 用来表示当前候选框与实例框重合的可信程度. 它对候选框和实例框分别进行建模, 并用 KL 散度来衡量两者分布间的距离:

$$L_{reg} = D_{KL}(P_D(x)||P_\theta(x)) \tag{2}$$

其中, $P_D(x)$ 表示以实例框中心坐标为均值的狄利克雷分布, $P_\theta(x)$ 表示已候选框中心坐标为均值, 以定位置信度为标准差的高斯分布.

该算法采用类似于集成学习的思想通过训练得到不同候选框的定位置信度, 并以其作为权重, 对所有大于非极大值抑制阈值的候选框进行加权求和, 得到最终的检测结果.

与 Softer-NMS 思想类似, Jiang 等[46] 设计了额外的分支来预测每一个候选框的交并比值大小, 再通过类似聚类的规则来更新分类置信度, 最终用更新后的置信度来完成非极大值抑制.

除上述方法外, Liu 等[47] 设计了一个仅包含卷积层和全连接层的子网络, 用来判断交并比大于阈值的非最优候选框是否和最优候选框预测的是同一个目标, 保留检测目标不同的非最优候选框, 从而有效避免了传统方法存在的弊端.

针对不同数据集和任务应当设计不同的非极大值抑制算法: 当数据集或任务的场景中目标比较稀疏, 优化的抑制算法对模型的检测性能几乎没有提升, 使用原始的非极大值抑制算法就能比较简洁地完成目标检测任务. 但在场景复杂、目标密集的情况下, 对非极大值抑制算法进行优化能够有效提升模型的检测性能.

2.4 交并比算法的优化

交并比是指两个图形的交集与并集的比值:

$$IoU = \frac{area(A) \cap area(B)}{area(A) \cup area(B)} \tag{3}$$

其中, $area(*)$ 表示图形的面积.

交并比用来衡量两个图形间的重合度. 在目标检测中, 锚点框和相应实例框的交并比决定了其是正样本还是负样本; 候选框间的交并比值决定是否进行非极大值抑制.

对于目标检测, 传统的交并比可以很好地表达两个相交矩形框间的距离. 但是, 对于不相交的矩形框, 交并比始终为 0, 无法反映它们之间的距离.

为了更一般地表达矩形框间的距离, Rezatofighi 等[48]提出了泛化交并比 (Generalized IoU) 的概念. 它对于任意两个凸形 A、B 在空间中寻找包含它们的最小凸形 C, 则泛化交并比定义为:

$$GIoU = IoU - \frac{area(C) - area(A) \cup area(B)}{area(C)} \tag{4}$$

从式 (4) 可以看出, 泛化交并比的取值范围是 $(-1, 1]$, 对于不相交的两个矩形框, 它们中心点间距离越大, 泛化交并比越小, 这种特性对降低负样本学习的难度是非常有利的.

交并比阈值决定了正负样本的划分结果. 当交并比阈值较大时, 意味着选取真实正样本的标准更为严苛, 则更少的正样本进行损失函数的计算, 这容易引起类别不平衡以及漏检的问题发生. 当交并比阈值较小时, 意味着更多的错误正样本会被当作正样本进行训练, 从而降低模型的检测性能, 为解决这一矛盾, Cai 等[49] 提出了多阶段变交并比阈值的方法, 不同阶段设置不同的交并比阈值, 满足不同阶段模型的需求 (图 19).

在前面阶段, 模型的性能较差, 需要通过不断“试错”来学习正确的分类和定位; 在后面阶段, 模型的性能有所提升, 则可以适当提高正样本的判断标准, 进一步提升检测性能.

交并比是目标检测模型中非常重要的部分, 它直接影响着模型的训练效果和检测结果. 对于交并比算法的设计和优化应注意:

1) 交并比值要充分反应候选框与实例框, 候选框与候选框之间的距离.

2) 交并比阈值应根据训练过程中模型当前性能, 以及相应任务或数据集的样本分布来确定.

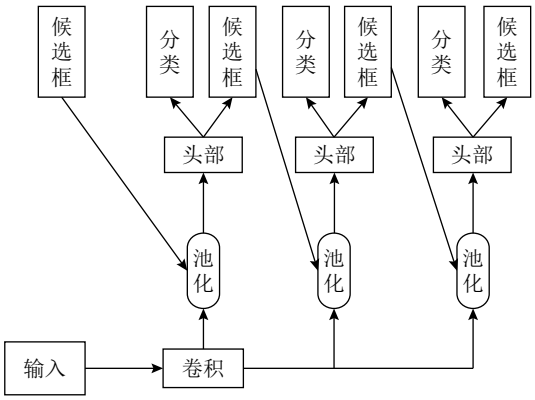


图 19　级联多阶段目标检测模型

Fig. 19　Cascade stages of object detection model

2.5　正负样本采样算法的优化

锚点框在大尺度的特征图上数量较多, 大部分负样本锚点框提供的梯度信息相近, 将它们全部用于分类和回归训练浪费计算资源和时间, 因而需要对所有锚点框进行采样, 只选择其中部分参与训练.

由于正例锚点框的数量远远小于负例锚点框, 直接在全局进行随机采样会很容易引起正负训练样本不均衡的问题. Faster R-CNN 模型[6] 对正样本和负样本分别进行随机采样, 采样比例为 1:1. 这种分类采样的方法较好地解决了类别不均衡的问题, 但没有充分利用负样本中的错误信息来帮助模型训练. SSD 模型使用了困难负样本采样的策略, 对负样本按置信度误差进行降排序, 将置信度较低的困难负样本用来更新模型.

区别于 SSD 模型, Shrivastava 等[50] 则根据输入样本的损失来在线选择困难负样本. 他们对两阶段检测框架进行扩充, 设计了另外一个 RoI 网络用来专门计算输入样本的损失, 并对输入的损失进行降排序, 选择损失最大的前 n 个负样本用于模型的训练, 用输入损失作为衡量样本学习难度标准的优势在于可以同时考虑分类和回归的困难程度. 受到上述研究的启发, Yu 等[51] 采用类似的方法对单阶段检测框架的正负样本采样进行了优化. 它直接过滤简单样本, 只对损失值最大的 k 个样本进行反向传播来更新网络参数.

困难负样本同样可以用与实例框的交并比值来表示. Pan 等[27] 提出了一种基于交并比值的分级采样方法. 它将交并比值划分为 K 个区间, 每个区间的候选采样数为 M_k, 划分保证困难负样本在每个区间均匀分布, 数量为 N, 则采样方法表示为:

$$IoU = \frac{N}{K} \times \frac{1}{M_k}, \qquad k \in [0, K) \tag{5}$$

上式采样方法有效地使参与训练的样本分布更接近于困难负样本的交并比分布, 从而提高了困难负样本被选中的概率.

正负样本采样算法设计的关键在于如何解决训练过程中正负样本数不平衡及困难负样本难以充分利用这两个问题. 针对不同数据集和任务. 上述两个问题的突出程度有所不同: 模型应用的场景越复杂, 采样带来的问题影响就越严重, 合理的采样方法能得到的模型性能收益就越大.

2.6　区域特征编码方法的优化

对于两阶段的目标检测框架, 区域特征编码是指将推荐层输出的推荐框编码成固定长度向量的过程, 其目的是便于后续全连接层或卷积层对目标特征实行进一步的分类和回归.

在 R-CNN 模型[4] 中, Girshick 等直接将区域特征从整张图片上裁剪下来, 并通过线性插值将其调整到固定尺寸后送入全连接层中, 这种方法使用的特征虽然分辨率很高, 但是计算过于耗时.

Fast R-CNN[5] 和 Faster R-CNN[6] 模型中使用了区域特征池化 (RoI pooling) 的编码方法, 将任意大小的区域特征划分为固定尺寸的网格 (图 20), 在网格内采用最大池化提取出唯一的特征传给后面的全连接层. 然而, 这种下采样的方法对小目标检测效果很不理想, 并且使得目标失去了部分的位置信息. 为此, He 等[52] 又在此基础上提出了区域特征对齐 (RoI align) 的编码方法, 用双线性插值代替了最大池化, 保证了区域特征编码的精度.

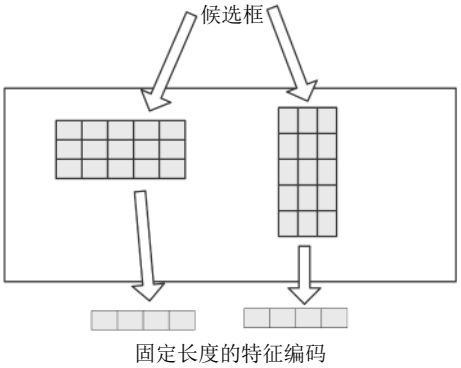


图 20　区域特征池化过程

Fig. 20　Pipeline of RoI pooling

为了更好地解决下采样带来的位置信息丢失问题, Dai 等[53] 提出了位置敏感的区域特征池化方法, 将每个网格作为特征图的一个通道, 通过最大池化的方法选择出置信度最高的通道作为目标所在的位置信息表示. 然而这种位置敏感的区域编码方法主

动放弃了全局信息, 编码后的特征缺失了空间上的关联. 针对此问题, Zhu 等[54] 提出将区域特征池化和位置敏感的区域特征池化方法进行融合, 从而得到包含全局和局部信息、鲁棒性更强的编码特征. Zhai 等[55] 在上述这些方法的基础上, 设计了一种特征选择子网络来针对不同大小和长宽比的子区域进行特征学习, 并将学习到的特征编码后送入了后续网络进行分类和回归.

区域特征编码能够为模型后续的精确分类和回归带来很大的帮助. 在设计和优化区域特征编码方法时应注意:

1) 使用下采样方法进行编码时会丢失区域特征的部分位置信息, 为保证位置信息相对完整, 应该尽量减少下采样的使用, 或在后续处理时进行相应的补偿.

2) 编码时应尽量保留区域特征内的全局信息, 避免不同区域特征间的关联缺失, 从而影响模型的检测性能.

2.7 分类与定位去冲突方法的优化

目标检测的训练是同时对物体进行分类和定位的多任务学习过程, 但由于分类任务需要位置不敏感的目标特征, 而定位任务却需目标特征对位置敏感, 这就导致模型在联合训练的过程中很难使两者同时达到最优, 最终影响模型的检测性能.

在 Fast R-CNN[5] 与 Faster R-CNN[6] 模型中, 区域特征池化操作破坏了全卷积网络的平移不变性, 从而引入了具有平移变换性的特征, 帮助模型进行更好的定位. 但是这种做法使得区域级的特征无法在后续网络中共享计算, 降低了模型训练和推理的速度. 针对此问题, R-FCN 模型[53] 创建了一个位置敏感的置信度图, 把平移变换特征引入了全卷积的网络中, 保证了特征在网络中的计算都是可以被共享的, 从而大大提高了模型的训练和推理效率.

对于单阶段框架的检测模型, 由于缺少区域特征池化等特征编码操作, 其经过直接回归后得到的候选框定位位置与用于分类的目标特征是不对齐的 (图 21). 为解决上述问题, Chen 等[56] 在单阶段模型的基础上, 利用可变形卷积[38−39] 来修正特征层的感受野, 并根据回归得到的候选框位置信息来确定卷积的补偿值, 以得到对齐后的目标特征, 从而实现了候选框位置与目标特征间的匹配.

除了上述方法外, 将分类问题与定位问题进行一定程度上的解耦后分别考虑, 同样是缓解两者冲突问题的有效手段. Cheng 等[57] 认为分类与定位任务的冲突是导致检测模型出现错误正例现象的重要原因. 对此, 他将 RPN 网络输出的推荐框映射回原

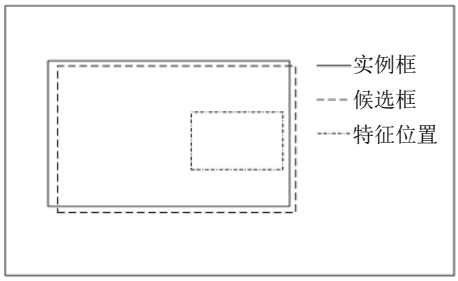


图 21 目标特征与候选框不对齐

Fig. 21 Misalignment between feature and box

图进行裁剪, 再把裁剪后的图像输入到新的 R-CNN 网络中单独进行一次分类, 得到最终的分类结果. 为解决模型两次分类导致训练和推理速度过慢的问题, Cheng 等[58] 又通过共享两次分类过程中的浅层特征计算, 来对模型进行加速. 整个模型的计算流程如图 22.

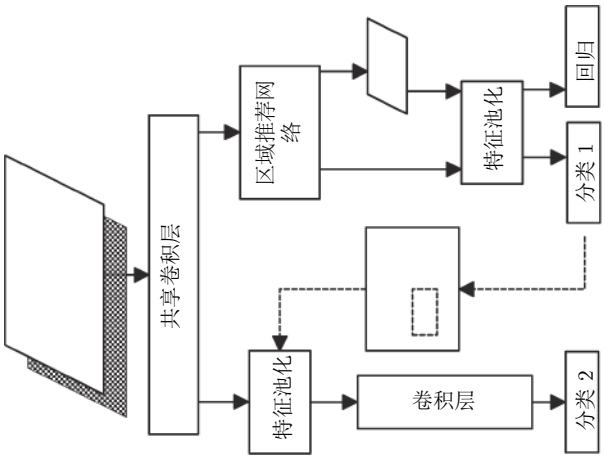


图 22 DCRv2 模型的检测流程

Fig. 22 Overall pipeline of DCRv2

模型定位性能同样受到分类任务的影响. 传统的非极大值抑制方法只根据目标的分类置信度来决定检测框的去留, 而没有直接考虑候选框的定位, 导致定位精度更高的候选框反而有可能被抑制. 对此, He 等[45] 通过衡量候选框与实例框位置分布的差异, 来重新计算每个候选框的置信度得分; Jiang 等[46] 预测每个候选框与对应实例框的交并比值作为定位置信度, 并用其来引导非极大值抑制. 这类方法在本质上通过单独考虑分类与定位对检测任务的贡献, 从而缓解了分类置信度与定位精度间不匹配的问题.

分类特征与定位特征不匹配的问题几乎贯穿目标检测的整个过程, 在设计相关解决方案时需考虑:

1) 当前检测场景或任务下, 分类与定位中哪一

个指标对模型最终性能的影响更大, 从而用不同的权值来考虑两者的重要性.

2) 在不同的检测模型中, 分类与定位产生冲突的方式有所区别, 应根据其特点来针对性地考虑缓解两者矛盾的方法.

2.8　上下文信息建模方法的设计优化

上下文信息建模是指考虑目标物体的周围环境, 通过显式地建模目标与周围环境的关系, 利用待检测物体本身之外的信息, 来帮助模型对该目标的检测. 根据利用的上下文信息范围的不同, 该方法可以被分为全局上下文建模与局部上下文建模两类.

全局上下文建模在整张图像上考虑外部上下文信息, 通常的做法是将提取到的外部特征与目标特征进行拼接, 然后送入卷积层或全连接层中进行分类与回归. 具有代表性的工作是 Bell 等[59] 提出的 ION 网络、Ouyang[60] 等提出的 DeepID 网络. 其中, ION 对内通过跳跃池化 (Skip pooling) 提取目标不同尺度的特征, 对外采用空间递归神经网络 (IRNN) 来提出目标外的上下文信息, 并将两种特征经过 L_2 归一化后拼接在一起, 送入后续的卷积层与全连接层进行分类与回归 (图 23). 而 Deepid 网络则对每一张图像学习一个类别得分, 并将其作为上下文特征与目标特征进行拼接, 送入后续的 SVM 分类器中进行分类.

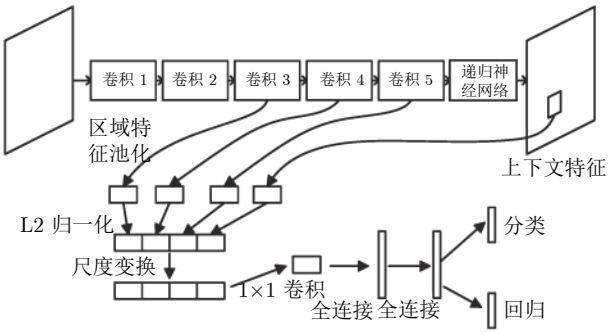


图 23　ION 网络的总体框架

Fig. 23　Pipeline of ION

局部上下文建模只考虑待检测目标与周围环境或其他目标之间的上下文关系, 并将其作为线索帮助模型进行推理. Chen 等[61] 针对非极大值抑制方法存在的问题, 提出了空间记忆网络 (Spatial memory network, SMN) 来保留和更新之前检测到的目标特征, 它在每一轮迭代中把上一次的检测结果作为先验知识输入到 RPN 网络中来提升本次检测的效果, 然后将上一记忆单元与新检测到的目标特征输入到 GRU 网络中来更新记忆单元 (图 24). 由于该方法显式地考虑了不同目标间的关系, 在后处理阶段无需再进行非极大值抑制就能得到最终的检测结果.

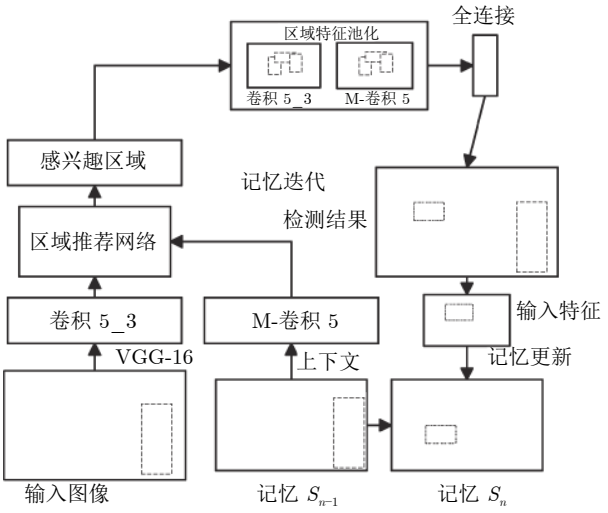


图 24　SMN 网络的记忆迭代过程

Fig. 24　Memory iterations of SMN

考虑到场景信息对不同目标的检测, 以及不同物体对某一目标的检测做出的贡献都是不同的, Liu 等[62] 提出了基于结构推理的检测网络 (Structure inference net, SIN), 它将检测任务用图结构来进行建模 (图 25), 把经过区域特征编码和全连接计算的目标特征作为图的顶点, 不同 ROI 之间的权重关系及场景信息作为图的边, 然后通过 GRU 网络 (图的边为输入特征, 图的顶点为隐藏层状态) 学习不同目标物体间的关系, 最终得到新的特征用于分类和回归.

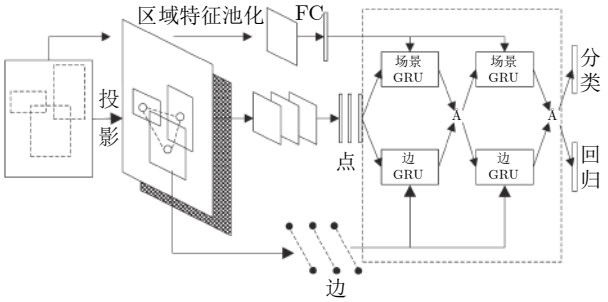


图 25　SIN 网络的检测流程

Fig. 25　Pipeline of SIN

为更加直观地建立不同目标物体间的关系, Hu 等[63] 借鉴文献 [64] 中的思想, 设计了目标关系模块 (Object relation module), 它利用图上所有目标的外观特征与几何位置特征, 来显式地计算不同

物体与待检测目标特征的权重关系, 再将其与目标原始特征进行叠加, 得到新的目标特征, 用于最终的分类与回归.

除了考虑不同目标间的上下文关系外, 部分学者还对目标内不同子区域间的关系进行了相关研究. 其中, Gidaris 等[65] 从推荐框的不同子区域 (边界区域, 上下文区域, 中心区域等) 内提取出不同特征, 并将这些特征与原始区域特征进行拼接, 得到新的目标特征. 考虑到物体间遮挡问题, Zeng 等[66] 提出了双向门卷积网络 (Gated bi-directional CNN, CBDNet), 用来在不同子区域特征间传递信息, 从而筛选出对检测有帮助的区域内上下文信息, 得到更好的目标特征.

总之, 上下文信息建模通过显式地表达不同目标与目标间, 目标与场景间的关系, 来建立相应模型对周围场景的视觉理解, 提高模型检测困难目标物体的准确率和召回率, 在建立上下文信息模型时应当考虑:

1) 上下文信息会影响目标原始的外观、几何、位置等自身特征, 并非所有场景下使用上下文信息都会对模型检测性能有提升, 需根据相应的检测任务或数据集来决定是否应用.

2) 不同对象提供的上下文信息对检测当前目标的贡献是不同的, 在利用这些上下信息进行建模时应当分别考虑.

2.9 多尺度预测方法的设计优化

多尺度预测指对于不同尺度的目标用不同分辨率的特征图进行检测. 区别于特征融合, 该预测方法在模型网络结构上表现为多分支, 每个分支的有效感受野[67] 大小不同.

在 SSD[7] 中, 模型共有 6 个不同分辨率的特征图对目标进行检测. 其中, 大分辨率的特征图有效感受野小, 用于检测小物体; 小分辨率的特征图有效感受野大, 用于检测大物体. 此后, YOLO 系列[8–10]、RetinaNet[68] 等经典的单阶段检测模型都采用了类似的预测方法.

多尺度预测在两阶段框架上同样有着较多的应用. 文献 [69−70] 都在不同分辨率的特征图上来进行候选框的选取, 以满足不同尺度大小物体的检测需求. Singh 等[71] 指出将极端尺寸的物体用于模型训练会导致其检测性能下降, 从而提出了限定训练样本尺寸的多尺度训练和预测的 SNIP (Scale normalization for image pyramid) 算法. 该算法采用图像金字塔对输入进行多尺度变换, 针对不同尺度的输入设定样本的有效尺寸范围, 使模型只从合理尺寸大小 (当前输入尺度下) 的样本中进行梯度计算和参数更新, 避免极端大小目标对模型性能的影响, 最终到达多尺度训练与预测的目的 (图 26). 但是, 由于 SNIP 模型需要在高分辨率图上进行多尺度的训练, 所以训练速度非常慢. 为此, Singh 等[72] 对模型的图片输入进行了预处理, 利用正例框选 (Positive chip selection) 的方法从原图中提取出多幅不同尺度的低分辨率子图, 用于包含原图中所有的实例. 这种方法大大减少了模型对高分辨率背景区域的计算, 提高了模型的训练速度.

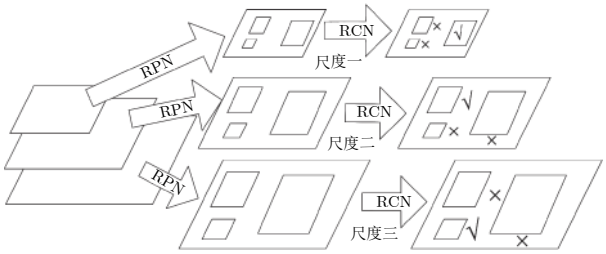


图 26 SNIP 模型的多尺度训练与预测

Fig. 26 Multi-scale training and inference of SNIP

为进一步提高模型对小目标的检测性能, Najibi 等[73] 标记与小物体 (在某一尺度下) 有交集的像素作为焦点像素 (Focus pixels), 通过连通相邻的焦点像素形成焦点区域, 并采用级联的方式训练该区域去预测小目标, 从而提高了模型对小目标检测的准确率和召回率.

多尺度预测方法的本质目的就是使得目标尺度与当前特征图的有效感受野相匹配. 为让同样结构的预测分支拥有不同感受野, Li 等[74] 借鉴了空洞空间金字塔池化(Atrous spatial pyramid pooling, ASPP)[75] 模块的设计思路, 在不同预测分支上使用膨胀率不同的空洞卷积来获得不同大小的感受野 (图 27), 并通过支路间的参数共享提升模型的训练和推理速度, 最终取得了较好的检测效果.

多尺度预测有效缓解了卷积神经网络缺少尺度

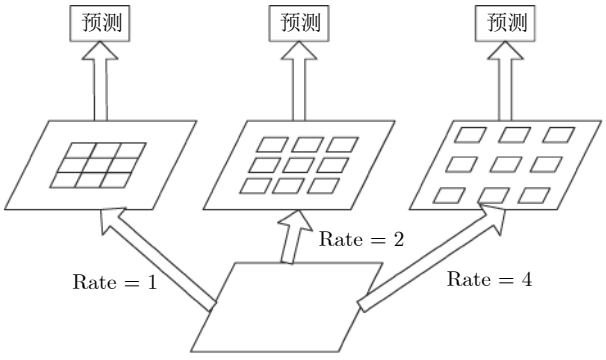


图 27 TridentNet 模型的多尺度预测

Fig. 27 Multi-scale inference of TridentNet

不变性的问题, 对检测模型性能的提升起到了关键作用. 在设计和优化多尺度预测方法时, 应当考虑:

1) 利用特征图进行多尺度预测时, 可以通过特征融合等方法来缓解浅卷积层特征提取不充分 (如 SSD) 的问题.

2) 多尺度会增加较多的计算量, 故在设计相应算法时需要考虑模型训练与推理的资源和时间消耗, 通过参数共享、转化输入等技巧来减少计算.

2.10　损失函数的设计优化

损失函数量化了检测模型在训练过程中出现的分类和定位等错误, 为模型的更新提供了方向. 大多数目标检测算法的损失函数由分类损失函数与定位损失函数的加权和求得:

$$L = L_{cls} + \lambda L_{reg} \tag{6}$$

其中, L_{cls} 表示分类损失函数, L_{reg} 表示定位损失函数, λ 表示平衡分类损失与定位损失的权重系数.

2.10.1　分类损失函数

分类损失指模型预测目标类别与实例不符带来的惩罚. 在经典的单阶段和两阶段检测框架中, 分类损失一般由二分类交叉熵表示:

$$L_{cls}(p) = -\frac{1}{N_{pos}}\sum_{i\in pos}\ln(p_i) - \frac{1}{N_{reg}}\sum_{i\in neg}\ln(1-p_i) \tag{7}$$

其中, p_i 表示由 $k+1$ 类 softmax 计算出的置信度, N_* 表示对应类别的样本数目.

从式中可以看出, 当输出的特征图某个网格上预测的正确类别置信度越低则相应的损失函数值就越高, 从而对模型参数的更新贡献就越大.

传统的二分类交叉熵没有考虑样本中存在类别不平衡等问题, 使得数量众多的简单负样本控制了模型更新的方向, 而具有更多信息的正样本和困难负样本对梯度的更新几乎没有贡献. 针对这个问题, Lin 等[68] 提出了焦点损失 (Focal loss), 它在式 (7) 基础上分别为正负样本、难易样本添加了权重:

$$L_{cls}(p,y) = \begin{cases} -\alpha(1-p)^{\gamma}\ln(p), & y=1 \\ -(1-\alpha)p^{\gamma}\ln(1-p), & \text{其他} \end{cases} \tag{8}$$

其中, α 表示正负样本间的权重, 取值为 [0, 1]. 表示难易样本间的权重, 取值为 ≥ 0 .

上式表明, 对于样本数较少、学习难度大的类别应适当增加其在损数函数中权重, 保证所有样本对模型参数更新的贡献都是相对平衡的.

焦点损失函数虽然通过调整样本权重, 有效缓解了类别不均问题, 但却没能显式地考虑样本间的关系, 并且手工设计的超参难以适用于不同的任务和数据集. 针对上述问题, Chen 等[76] 提出了基于样本置信度排序的平均准确率损失 (Average precision loss), 用样本间置信度的差值代替原本置信度:

$$x_{ij} = -(s(b_i;\theta) - s(b_j;\theta)) = -(s_i - s_j) \tag{9}$$

其中, $s(b_i;\theta)$ 表示参数 θ 为的锚点框 b_i 的置信度.

则平均准确率损失定义为:

$$L_{AP} = 1 - \frac{1}{|P|}\sum_{i\in P}\frac{1+\sum\limits_{j\in P, j\neq i}H(x_{ij})}{1+\sum\limits_{j\in P, j\neq i}H(x_{ij})+\sum\limits_{j\in N}H(x_{ij})} \tag{10}$$

其中, $H(*)$ 表示单位阶跃函数, P 表示正样本, N 表示负样本.

上式中, 当负样本的置信度得分高于正样本置信度得分时, 以正负样本置信度差值为输入的单位阶跃函数取值为 1, 从而产生损失.

除了为样本提供权重的方法外, Li 等[77] 从梯度更新角度出发, 提出了梯度均衡机制 (Gradient harmonizing mechanism). 该方法首先定义并计算了梯度模长 $g = |p - p^*|$, 其中 p 为对应类别的置信度, p^* 为对应实例标签. 如图 28 所示, 对于一个收敛的检测模型, 梯度模长很小或很大的样本数量远远多于模长中等的样本数量.

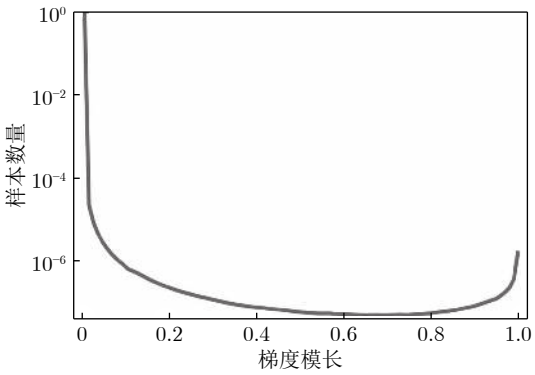


图 28　不同梯度模长的样本数量

Fig. 28　Number of samples with different gradient norm

定义区域内样本数量与区域大小比值为梯度密度, 则:

$$GD(g) = \frac{1}{l_\epsilon(g)}\sum_{k=1}^{N}\delta_\epsilon(g_k, g) \tag{11}$$

$$\delta_\epsilon(x,y)(p,y) = \begin{cases} 1, & y-\dfrac{\epsilon}{2} \leq x < y+\dfrac{\epsilon}{2} \\ 0, & \text{其他} \end{cases} \tag{12}$$

$$l_\epsilon(g) = \min(g+\frac{\epsilon}{2}, 1) - \max(g-\frac{\epsilon}{2}, 0) \tag{13}$$

其中, ϵ 表示区间长度, g_k 表示第 k 个样本的梯度.

则梯度均衡机制的分类损失定义为:

$$L_{GHM-C}=\sum_{i=1}^{N}\frac{L_{CE}(p_i-p_i^*)}{GD(g_i)} \tag{14}$$

其中, L_{CE} 表示标准的二分类交叉熵.

从上式可以看出, 对于样本多的梯度模长区间, 其在总分类损失中的权重会下降, 从而使模型更加均衡地学习不同难度的负样本, 以提高其分类性能.

2.10.2 定位损失函数

定位损失指模型预测目标位置与实例位置不重合带来的惩罚. 在经典的单阶段和两阶段检测框架中, 定位损失一般由平滑 L_1 范数来表示:

$$Smooth_{L1}=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{其他}\end{cases} \tag{15}$$

其中, x 表示候选框实际补偿与预测补偿的差.

平滑 L_1 范数的优点在于对定位误差的惩罚是线性增长的, 这有利于保持训练过程的平稳.

在目标密集的检测场景下, 模型的输出结果会很大程度上受到周围其他目标的影响干扰. 为此, Wang 等[78] 提出了排斥损失 (Repulsion loss), 迫使预测框与对应实例框靠近, 而增大其与其他实例框和预测框间的距离:

$$L=L_{Attr}+\alpha\times L_{RepGT}+\beta\times L_{RepBox} \tag{16}$$

其中, L_{Attr} 表示预测框与对应实例框间的距离损失, L_{RepGT} 表示预测框与其他实例框间的距离损失, L_{RepBox} 表示预测框与周围其他预测框间的距离损失. α, β 为实验确定的权重常数.

当预测框与对应实例框距离很近, 而与其他实例框和预测框的距离很远时, 模型计算得到的定位损失就越小, 表明当前模型检测性能越好.

通常, 模型的总损失函数是分类损失与加权定位损失的和. 然而, 这种通过加权对分类损失和定位损失的权重进行调整的做法会导致模型对定位损失大的样本更加敏感, 影响模型的性能. 针对上述问题, Pan 等[27] 提出了平衡 L_1 损失, 用于降低被过度放大的困难负样本梯度, 适当提高简单样本的梯度. 文中通过对定位损失梯度的设计来反解定位损失函数:

$$\frac{\partial L_b}{\partial x}=\begin{cases}\alpha\ln(b|x|+1), & |x|<1\\ \gamma, & \text{其他}\end{cases} \tag{17}$$

其中, α、b、γ 为实验确定的常数.

无论是 L_1 损失还是 L_2 损失, 都是将各个位置坐标分开来进行独立的预测. 这种对位置的描述方法忽略了各个点之间的联系, 因而并非一定是对位置最直观的表示. 为此, Yu 等[79] 提出了基于交并比的定位损失计算方法, 直接以预测框和相应实例框的交并比作为判断定位准确性的依据:

$$L_{IoU}=-\ln\frac{area(A)\cap area(B)}{area(A)\cup area(B)} \tag{18}$$

其中, $area(*)$ 表示区域面积.

从上式中可以看出, 若预测框与实例框交并比越大, 则损失函数的值越小, 说明模型当前的性能较好.

然而上述损失函数值在预测框和实例框交并比值很小时趋于无穷, 不利于数值稳定. Tychsen-Smith 等[80] 提出了带上界的交并比损失 (以预测框横坐标为例):

$$L_{IoU}=2L_1(1-IoU(x,b_t)) \tag{19}$$

$$IoU(x,b_t)=\max(0,\frac{w_t-2|\Delta x|}{w_t+2|\Delta x|}) \tag{20}$$

其中, b_t 表示实例框, x 表示预测框中心位置横坐标, w_t 表示实例框宽度, Δx 表示预测框与实例框中心横坐标的差.

上式损失值在预测框与实例框交并比值为 0 时是常数, 从而有效避免了数值溢出的问题.

2.10.3 其他损失函数

除了目标的类别和位置外, 一些检测模型同时需要预测其他特定模块的输出. 例如, 文献 [37] 中对锚点框生成的位置及形状进行预测, 设计了锚点框损失函数; 文献 [11] 中对特征图上所有左上角点与右下角点的配对进行预测, 设计了配对组合损失函数; 文献 [42] 中对最佳尺度的特征层进行预测, 设计了特征层选择的损失函数.

由于这类损失函数都是针对特定的目标检测算法而设计, 不具备一般性, 故在此不做具体阐述.

总之, 损失函数决定了模型参数的更新方向, 从而很大程度上影响模型最终的检测性能. 损失函数的设计优化应注意:

1) 损失函数在定义域内应是连续可微的.

2) 对于分类损失函数, 应当考虑不同类别样本对参数更新的贡献, 保证模型对各类样本学习都足够充分.

3) 对于定位损失函数, 应当选取位置表示能力较强且优化难度较低的决策变量, 并根据具体任务与数据集做合理的修正.

4) 对于端到端的训练过程, 总损失函数是由各类损失函数加权求和构成的, 其权重大小应当根据具体任务或数据集, 通过实验的方法来确定.

2.11 特定场景下的检测模型优化

在不同场景与任务下, 由于模型检测对象的差异, 相应的目标检测架构也会有较大区别.

2.11.1 自然场景下的文本检测模型优化

自然场景下的文本检测任务中, 文本框通常具有多方向、极端长宽比、形状不规则等特定问题. 因而无法直接套用通用目标检测模型.

针对上述问题, Ma 等[81] 在 Faster R-CNN 基础上, 设计了预测文本倾斜角度的 RPN 网络, 来实现多方向的文本检测. Liao 等[82] 在 SSD 模型的基础上, 通过修改卷积核尺寸以适应长文本的检测, 并将回归水平预测框修改为回归四边形框的角点, 来实现倾斜文本的检测. Zhou 等[83] 继承 DenseBox[84] 的检测思想, 在多尺度融合的特征图上进行像素级的文本检测, 预测每个像素到四个边界的距离以及旋转角度, 得到最终的检测结果.

某些情况下不同文本的间距很小, 语义分割难以将其完全分开. Deng 等[85] 提出使用实例分割来解决该问题, 它借助 FCN 网络进行像素级预测, 分别得到文本和链接的二分类, 使用正链接去连接邻近的正文本像素, 得到文本实例分割的结果. 为提高对弯曲文本的检测性能, Xie 等[86] 在 Mask R-CNN 的基础上多一个分支做文字语义分割, 并把语义分割的中间特征和检测分支特征进行融合, 用语义分割的结果作为注意力, 对实例特征再进行一次计算, 得到最终的分类得分. 这种做法显著提高了 Mask R-CNN 模型对不规则形状文本的检测精度, 取得了很好的效果. 除了自顶向下的实例分割方法外, Wang 等[87] 采用一种渐进尺度扩展网络来实现文本间紧挨情况下的检测, 它以 FPN 为基础架构, 用多支路来表示不同核的分割结果, 再通过渐进式扩展算法不断扩大文本核, 直到相邻核之间发生扩展冲突, 得到最精细的文本检测结果.

2.11.2 航拍图像下的检测模型优化

航拍图像下的目标占像素小、密集、多方向, 其数据集[88−89] 与通用场景下的图像数据差距较大, 因而通用模型在航拍图像上难以取得良好的检测效果.

为提高对航拍图像下小目标检测的召回率, Yang 等[90] 采用提高特征图分辨率 (或减少锚点框步长) 的方法使更多的小目标匹配到正样例, 并通过不同层级间的特征融合, 以及空间与通道上的注意力机制来增强不同尺度目标的特征, 来进一步区分前景与背景.

考虑到航拍角度的特殊性, 获得旋转不变的目标特征是尤为关键的. Ding 等[91] 利用一个全连接层从水平区域框中学习到旋转区域框, 并通过旋转位置敏感模块 (Rotated position sensitive RoI Align) 来从该框中获得旋转不变的目标特征, 用于后续的分类与回归.

直接回归方向边界框的角度需要一些复杂的规则来保证角度计算不出现歧义, 这加大了模型学习的难度. 对此, Qian 等[92] 提出了八参数的旋转损失, 它采用基于向量叉积 (Cross-product) 的计算方法来得到四边形边界框的回归计算顺序, 从而来消除角度计算时出现的歧义问题. Zhu 等[93] 不直接回归方向框的旋转角度, 而是用角度分别为 90°与 180°的两个不同周期的周期向量 (Periodic vectors) 来隐式地表达边界框的方向, 这种方法生成的标签数据能够省去复杂的规则描述, 从而更加简单与合理地表示了带方向的标注框.

2.11.3 遮挡环境下的行人检测模型优化

行人之间相互遮挡是密集人群检测的一大难点. 被遮挡行人的特征受到周围行人的影响, 导致检测中出现假正例与漏检的问题. 为此, 一些行人检测数据集[94−95] 提供了更有针对性的数据标注, 来帮助解决行人遮挡问题.

Pang 等[96] 利用标注中的遮挡信息设计了基于掩码的空间注意力机制模块 (Mask-guided attention), 来帮助模型更加专注于行人未被遮挡部分的特征, 从而有效缓解了周围其他特征对行人检测的干扰. Zhang 等[97] 等针对遮挡问题, 在 Faster R-CNN 模型的基础上提出用基于遮挡敏感的区域特征编码来代替原始模型中的相应操作, 它将行人分成五个部分后分别进行区域特征编码, 并根据不同部分的遮挡程度来加权组合这些特征, 得到对遮挡敏感的行人特征并用于后续的检测. Liu 等[98] 借鉴 FCN 与 DCN 的思想, 采用位置敏感的可变形卷积池化来增加模型特征编码的灵活性, 让模型更多地从行人可见部分中学习相应特征, 避免其他物体的遮挡干扰.

除了使用注意力机制外, 部分学者从损失函数的角度来解决遮挡问题. Wang 等[78] 与 Zhang 等[97] 都通过缩小候选框与相应实例框距离, 增大与实例框和候选框距离, 来缓解候选框间距离太近导致后处理困难的问题.

在密集场景下, 传统的非极大值抑制方法处理候选框会很容易出现漏检与假正例问题. 对此, Liu 等[99] 针提出了自适应的非极大值抑制算法. 它首先在两阶段模型上增加人群密度估计分支, 得到该区域的人群密度大小, 然后根据该值动态地调整非极大值抑制的阈值, 从而较好解决了遮挡下的候选框后处理问题.

综上所述, 对于特定场景下的检测模型进行优化时应注意:

1) 数据集的选择非常重要. 专业数据集往往比

通用数据集的标注信息更具针对性, 因而能更好地帮助检测模型的设计、训练与测试.

2) 不同任务对检测模型的要求不同, 应根据当前任务的特殊性与难点来针对性地优化模型.

3 实验结果对比与分析

本文在 COCO 2017 数据集上对上述各种改进的模型进行实验和比较, 结果见表 1 与表 2 (注: R-ResNet, X-ResNeXt,HR-HRNet, D-DarkNet, HG-Hourglass. ++表示使用了多尺度、水平翻转等策略).

对比表中的数据进行分析后可以看出:

1) 相同检测框架下不同主干网络的模型检测性能差别很大, 主干网络越深则相应模型检测性能越好, 颈部连接结构引入的特征融合操作, 对检测

表 1 各检测模型的性能对比

Table 1 Performance comparison of different object detection models

模型	主干网络	AP	AP_{50}	AP_{75}	$AP_{\rm S}$	$AP_{\rm M}$	$AP_{\rm L}$
Faster R-CNN	VGG-16	21.9	42.7	—	—	—	—
Faster R-CNN	R-101*	29.1	48.4	30.7	12.9	35.5	50.9
Faster R-CNN	R-101-CBAM	30.8	50.5	32.6	—	—	—
Faster R-CNN++	R-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN++	HR-W32	39.5	61.0	43.1	23.6	42.9	51.0
Faster-DCR V2	R-101	34.3	57.7	35.8	13.8	36.7	51.1
OHEM	VGG-16	22.6	42.5	22.2	5.0	23.7	37.9
SIN	VGG-16	23.2	44.5	22.0	7.3	24.5	36.3
ION	VGG-16	23.6	43.2	22.6	6.4	24.1	38.3
Mask R-CNN	R-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Mask R-CNN	HR-32	40.7	61.8	44.7	25.2	44.4	51.8
Mask R-CNN	R-101-FPN+GC	40.8	62.1	45.5	24.4	43.7	51.9
SN-Mask R-CNN	R-101-FPN	40.4	58.7	42.5	—	—	—
IN-Mask R-CNN	R-101-FPN	40.6	59.4	43.6	24.3	43.9	52.6
R-FCN	R-101	29.9	51.9	—	10.8	32.8	45.0
CoupleNet	R-101	34.4	54.8	37.2	13.4	38.1	50.8
Cascade R-CNN	R-101	42.8	62.1	46.3	23.7	45.5	55.2
Libra R-CNN	R-101-FPN	41.1	62.1	44.7	23.4	43.7	52.5
Grid R-CNN[100]	X-101*	43.2	63.0	46.6	25.1	46.5	55.2
Light-Head R-CNN	R-101	38.2	60.9	41.0	20.9	42.2	52.8
M2Det800	VGG-16	41.0	59.7	45.0	22.1	46.5	53.8
SSD512	VGG-16	28.8	48.5	30.3	10.9	31.8	43.5
GHM SSD	X-101	41.6	62.8	44.2	22.3	45.1	55.3
YOLOV3	D-53*	33.0	57.9	34.4	18.3	35.4	41.9
YOLOV3	D-53	34.3	—	36.2	—	—	—
RetinaNet	X-101-FPN	39.0	59.4	41.7	22.6	43.4	50.9
GA-RetinaNet	X-101-FPN	40.3	60.9	43.5	23.5	44.9	53.5
RefineDet512++	R-101-FPN	41.8	62.9	45.7	25.6	45.1	55.3
FCOS	X-101-FPN	42.1	62.1	45.2	25.6	44.9	52.0
FoveaBox	X-101-FPN	42.1	61.9	45.2	24.9	46.8	55.6
FSFA	X-101-FPN	42.9	63.8	46.3	26.6	46.2	52.7
CornerNet	HG-104*	40.5	56.5	43.1	19.4	42.7	53.9
ExtremeNet	HG-104	40.2	55.5	43.2	20.4	43.2	53.1
CenterNet	HG-104	42.1	61.1	45.9	24.1	45.5	52.8
RepPoints	R-101	41.0	62.9	44.3	23.6	44.1	51.7
SNIP++	R-101	43.1	65.3	48.1	26.1	45.9	55.2
SNIPER++	R-101	46.1	67.0	51.6	29.6	48.9	58.1
TridentNet	R-101	42.7	63.6	46.5	23.9	46.6	56.6

*注: R-ResNet, X-ResNeXt, HR-HRNet, D-DarkNet, HG-Hourglass. ++表示使用了多尺度、水平翻转等策略

表 2 部分检测模型的速度、显存消耗、参数量与计算量对比 (基于 Titan Xp)
Table 2 Speed, VRAM consumption, parameters and computation comparison of some object detection models (on Titan Xp)

模型	主干网络	训练速度 (s/iter)	显存消耗 (GB)	推理速度 (fps)	参数量	运算次数
Faster R-CNN++	R-101-FPN	0.465	5.7	11.9	60.52×10^6	283.14×10^9
Faster R-CNN++	HR-W32	0.593	5.9	8.5	45.0×10^6	245.3×10^9
Mask R-CNN	R-101-FPN	0.571	5.8	9.4	62.81×10^6	351.65×10^9
Mask R-CNN	x-101-FPN	0.759	7.1	8.3	63.17×10^6	355.4×10^9
Mask R-CNN	R-101-FPN+GC	0.731	7.0	8.6	82.13×10^6	352.8×10^9
R-FCN	R-101	0.400	5.6	14.6	—	—
Cascade R-CNN	R-101-FPN	0.584	6.0	10.3	87.8×10^6	310.78×10^9
Cascade R-CNN	X-101-FPN	0.770	8.4	8.9	88.16×10^6	314.53×10^9
Libra R-CNN	R-101-FPN	0.495	6.0	10.4	60.79×10^6	284.19×10^9
Grid R-CNN	X-101-FPN	1.214	6.7	10.0	82.95×10^6	409.19×10^9
M2Det800	VGG-16	—	—	11.8	—	—
SSD512	VGG-16	0.412	7.6	20.7	36.04×10^6	386.02×10^9
GHM RetinaNet	X-101-FPN	0.818	7.0	7.6	56.74×10^6	319.14×10^9
RetinaNet	X-101-FPN	0.632	6.7	9.3	56.37×10^6	319.04×10^9
GA-RetinaNet	X-101-FPN	0.870	6.7	7.5	56.01×10^6	283.13×10^9
FCOS	R-101-FPN	0.558	9.4	11.6	50.96×10^6	276.53×10^9
CornerNet	HG −104*	—	—	4.9	—	—
ExtremeNet	HG-104	—	—	3.1	—	—
CenterNet	HG-104	—	11.91	8.5	—	—
RepPoints	R-101	0.558	5.6	10.9	55.62×10^6	266.23×10^9
SNIP++	R-101	—	—	<1.0	—	—
SNIPER++	R-101	—	—	4.8	—	—
TridentNet	R-101	0.985	6.6	2.1	—	—

*注: R-ResNet, X-ResNeXt, HR-HRNet, D-DarkNet, HG-Hourglass. ++表示使用了多尺度、水平翻转等策略

效果提升明显. 使用通道、空间注意力与全局上下文后, 模型的检测准确率进一步提升. 由此表明深层、多尺度、全局的特征对模型检测起到了重要的作用.

2) 锚点的设计对模型检测性能有较大影响, 根据目标几何特征来设计或自适应生成相应尺寸的锚点框能有效提高模型的检测精度. 无锚点检测模型作为新兴研究方向也有着比较出色的性能, 但其涉及像素级的计算过程, 因而显存占用相对较多.

3) 在高交并比的评价标准下, 改进的非极大值抑制算法能使模型检测准确率有所提升. 这表明非最优的候选框中也包含对模型检测有利的信息, 对其进行充分利用可以提高模型对困难样本的检测性能.

4) 变交并比阈值能帮助模型从粗到细地调整候选框的分类与定位, 解决模型当前性能与正负样本的交并比阈值不匹配问题, 有效地提高模型整体的检测性能, 但同时也引入了更多的计算量.

5) 对正负样本进行合理采样能缓解检测中的类别不均问题, 提高模型训练的有效性, 从而帮助模型更快地学习到目标特征, 提高检测精度.

6) 对比不同区域特征编码方法可以看到, 精确的局部位置信息以及全局信息对于两阶段模型获得表达能力强的编码特征都是很重要的.

7) 通过特征对齐或者解耦的方法, 来解决分类与定位的特征不匹配问题, 可以有效避免任务间相互影响, 帮助模型更有效地进行多任务学习.

8) 上下文建模考虑全局与周围目标的信息, 来提取出包含上下文信息的目标特征, 实验结果表明该方法能够有效提高模型 (特别是当主干网络较浅时) 对困难样本的检测效果.

9) 尺度变化对模型检测性能的影响比较严重. 对不同尺度目标采用不同有效感受野大小的特征图去检测, 能显著地提高目标在实际场景下的检测性能. 但是由于这类模型需要进行不同尺度上的特征计算, 在训练/推理速度、显存消耗、计算量上有较为明显的短板.

10) 针对特定问题进行合理优化后的损失函数

可以更直接地去表达和解决目标检测中存在的正负、难易样本不均, 目标遮挡严重, 定位精度不高等难题, 从而在几乎不增加模型复杂度的情况下提高模型的检测性能.

4 展望

除了对目标检测模型的各种子模块进行优化外, 近些年来该领域也出现了一些新兴的研究方向.

4.1 基于神经算法搜索的模块优化

神经算法搜索 (NAS) 是一种在给定搜索空间中搜索最优模型架构的自动学习算法. 基于强化学习的 NAS 算法, 通过设计 RNN 控制器来进行架构搜索, 利用子模型在搜索空间中的准确度作为奖励信号, 来更新其参数. 在反复的训练过程中, 控制器就能逐渐学会生成更好的模型架构.

NAS 在计算机视觉领域最早的应用成果是 NAS-Net[101]. 它借鉴了 ResNet、Inception 等主流网络重复堆叠的思想, 通过 RNN 控制器来预测分类网络的结构, 并利用验证集上的准确率来更新 RNN 控制器的参数, 最终得到了分类网络的基本堆叠单元. 实验表明, 该方法得到的分类网络在分类性能上有较大优势.

目标检测与图像分类的任务不同, 因而 NAS-Net 架构对于检测来说不是最优的. 针对该问题, 旷视[102] 首次提出了用于自动搜索物体检测器主干网络方法. 为降低搜索过程中的时间与资源消耗, 研究人员将网络权重的训练与结构搜索进行解耦, 先在 ImageNet 数据集上进行预训练, 再在检测数据集上微调参数, 最后进行网络结构的搜索. 搜索过程采用遗传算法来进行架构的更新, 收敛后得到的 DetNAS 主干网络模型性能超越了绝大部分的手工网络结构.

最近, Google Brain 又将 NAS 应用在特征金字塔的搜索上, 得到了 NAS-FPN[103] 架构. 它利用融合单元 (Merging cells) 组建起基本的特征金字塔结构和所有可能的特征层组合, 构造出算法的搜索空间. 从最终收敛的结果来看 (图 29), 底层向上传递信息的特征融合次数较多, 这表明精确的位置信息对模型检测准确率提升的有较大作用.

为提高 NAS 方法的搜索速度和效率, 降低计算资源的消耗, Wang 等[104] 采用无锚点的 FCOS 模型作为搜索对象, 并限定了卷积操作的搜索空间, 分别对特征金字塔和预测头部的结构进行了搜索, 搜索结果表明可变形卷积与拼接操作对于提升特征金字塔的性能非常关键, 而可变形卷积+1×1 卷积的结构在预测头部上取得了性能和计算量上的最佳平衡.

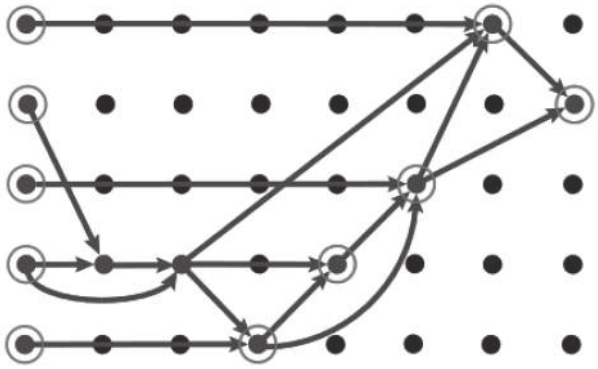


图 29 NAS 搜索收敛后的 FPN 架构

Fig. 29 NAS-FPN framework after convergence

为搜索更为复杂的目标检测模型架构, 并在更大的数据集上进行搜索验证, 降低 NAS 方法的计算和时间开销[105] 是非常必要的, 也是该研究领域未来发展的重要方向.

4.2 少样本的目标检测

目前, 主流的检测模型都是在大量数据样本下训练得到的, 这些模型在面对少样本的情况会出现严重的过拟合, 性能大大降低. 而少样本的目标检测 (Few-shot object detection) 正是针对少量训练数据提出的一类检测任务.

少样本学习在目标检测领域主要有元学习 (Meta learning) 与度量学习 (Metric learning) 两类. 元学习则注重训练模型少样本的学习能力, 使模型能够从少量样本中提取出有用的特征; 度量学习通过度量支撑样本与测试样本间的特征距离来进行目标的分类.

许多少样本学习[106−109] 模型都是在两阶段模型的基础上通过替换 RPN 网络和检测头部来实现的. 其中, Karlinsky 等[106] 在 Faster-RCNN 模型的基础上, 用子网络替换原始模型中的分类支路 (图 30), 提出了端到端的少样本学习框架, 它将区域编码后的特征送入度量嵌入模块计算出嵌入特征向量, 再计算该向量与每个类别的表征向量的距离, 来得到每个 ROI 区域的类别后验概率. 在少样本测试时, 用支撑样本计算出的表征向量替代训练过程中的表征向量, 从而获得新类别的表征, 并用于类别后验概率的计算, 得到最终的分类结果.

Fan 等[107] 提出了注意力机制的 RPN 网络 (图 31), 用来过滤与支撑样本类别不符的物体, 并为每个不同类别的支撑样本单独建立多关系头部 (Multi-relation head) 来对查询样本和支撑样本进行匹配,

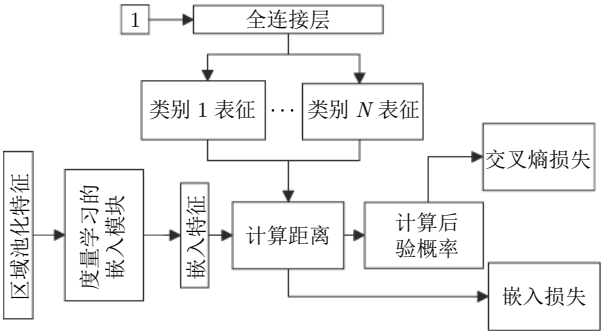


图 30　RepMet 模型的训练与推理流程

Fig. 30　Training and inference pipeline of RepMet

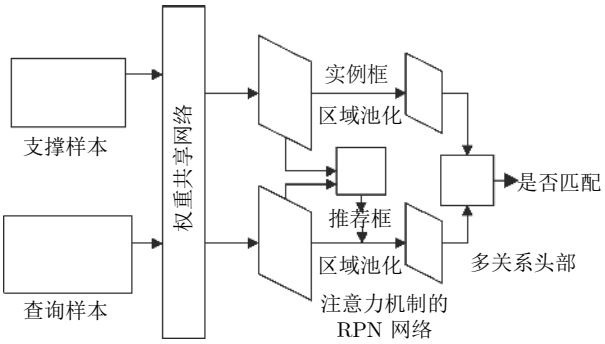


图 31　基于注意力机制 RPN 与多关系头部的少样本检测

Fig. 31　Attention-RPN and multi-relation head based few-shot detection

得到最终的检测结果.

虽然上述模型都实现了端到端的少样本检测, 但其并没考虑少样本下的定位问题. 对此, Kang 等[110] 提出了一种新的检测框架, 它包括元特征学习与特征权重调整两个部分. 给定一个查询样本和一组新类支撑样本, 特征学习器从查询样本中提取出元特征, 权重调整模块捕获支撑样本的全局特征, 并将其用于调整查询样本的元特征, 从而查询样本的元特征能够有效地使用支撑样本提供的分类与定位信息, 最终获得查询样本的分类与定位结果.

少样本检测是克服实际工程中数据样本缺乏问题的重要方法之一. 当下, 大多数相关模型在检测推理中使用的查询样本类别较少, 并对少样本下目标定位问题的研究不多, 这都是该方法在投入实际工程前需要解决的问题.

4.3　领域自适应的目标检测

目标检测通常假定训练数据与测试数据服从相同的分布, 然而在实际工程中并非总是如此, 这种分布上的不匹配会导致模型在实际场景应用中产生显著的性能下降. 领域自适应便是为了解决该问题而出现的新兴研究方向.

领域自适应的目标检测主要分为有监督与无监督两类. 有监督方法[111−113] 生成或利用少量的目标域 (Target domain) 标签来微调网络模型, 消除与源域 (Source domain) 间的差异. 无监督方法[114−116] 采用对抗训练的方式, 来训练领域判别器以最小化源域与目标域间的分布差异.

Ionue 等[111] 采用 CycleGAN[117] 将源域数据分布变换为目标域数据分布, 并在其上进行训练来对检测模型进行微调. 接着用微调后的检测器对目标域图像进行检测, 选取最高概率的结果作为目标图像的伪标注, 并用这些伪标注进一步对模型进行微调, 得到最终的检测模型.

Kim 等[113] 同时从有监督与无监督方法中受到启发, 通过 CycleGAN 生成与源域共享标签的中间域图像, 三个域图像同时输入到模型进行多类别的域判别器训练, 进而得到多个域间不变的目标特征, 用于后续的分类与回归.

Chen 等[114] 首次提出无监督的领域自适应目标检测, 它在 Faster R-CNN 模型的基础上, 同时训练图像级与实例级的域判别器 (图 32), 采用对抗训练的方法: 最小化域分类损失得到最佳域判别器, 最大化该判别器分类误差, 来对齐源域与目标域, 从而得到域不变的目标特征, 提高模型在目标域的检测性能.

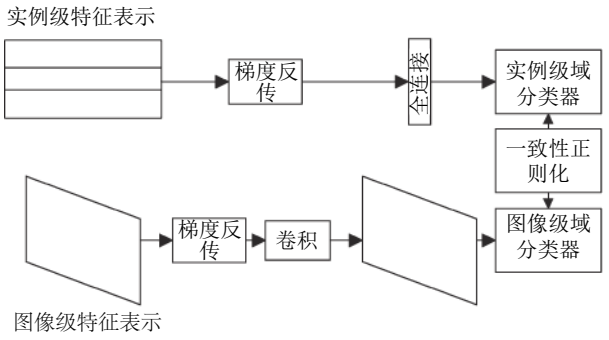


图 32　基于 Faster R-CNN 的域适应分支

Fig. 32　Domain adaptive branch based on faster R-CNN

浅层特征与高层特征包含的信息是不同的. Saito 等[115] 分别选取浅层特征与高层特征在源域与目标域上做不同程度的对齐: 底层特征使用交叉熵损失做强对齐, 高层特征使用 Focal loss 做弱对齐, 最终得到更适合于目标域检测的特征.

领域自适应的目标检测对提升模型在实际场景 (如恶劣天气条件) 下泛化能力的帮助很大, 是未来检测模型能在更一般环境下得到成功应用的关键技术.

4.4 轻量化的目标检测

为实现目标检测模型在嵌入式设备或移动端设备的落地, 减少模型推理的计算和时间开销是至关重要的.

Google 提出的 MobileNets 系列[118−120] 是专门为移动端设计的轻量级网络, 它用深度分离卷积来代替传统卷积, 并修改输入与输出层的结构, 来大大降低了模型计算量; 用倒置残差模块和线性瓶颈层来减轻低维度上非线性变换造成的信息丢失, 最终得到了在计算资源受限下具备较高检测性能的网络模型.

除了 MobileNets 系列外, Zhang 等[121−122] 设计了基于组卷积 (Group convolution) 优化的 ShuffleNet, 它针对组卷积输出的通道信息表示能力差的问题, 提出了通道间的混洗操作, 来促进不同通道间的特征信息传递, 在保证计算量几乎不变的条件下, 有效增强了每个输出通道的特征表示能力. Zhang 等[123] 提出了主从组卷积的计算方法, 通过主组卷积减少卷积计算量, 从卷积来融合不同分组间的目标特征, 并采用稀疏化卷积核、量化卷积核权重等方法来进一步压缩网络模型的大小.

Qin 等[124] 在 ShuffleNet 上进行改进, 通过增大浅层特征的通道数和感受野, 来获得更有效的目标特征, 并加入上下文信息增强模块和空间注意力模块来进一步促进多特征的融合, 从而在保证高速推理的同时提升模型的检测精度.

另外, 部分学者还通过模型剪枝的方法来降低检测推理的计算量. Zhang 等[125] 在 YOLOV3 模型的基础上, 通过对每一轮训练后的模型进行评估, 来剪枝尺度因子较小的通道, 不断降低模型的复杂度, 最终得到与原模型检测性能相近, 但推理速度更快的新模型.

在移动端与嵌入式端应用越发广泛的背景下, 降低模型计算开销以实现在这些设备上的部署是大势所趋, 也是未来目标检测领域研究的热点.

4.5 弱监督下的目标检测

实例级的数据标注是一项昂贵、费时费力的工作, 甚至在某些场合下是难以做到的. 而弱监督的目标检测从只提供图像级 (Image-Level) 的标注信息中来学习对目标的分类与定位. 由于缺少实例框标注, 弱监督下的目标检测需要根据图像特征来进行定位.

Bilen 等[126] 通过选择搜索 (Selective search) 的方法来获得大量候选框, 再利用分类阶段和检测阶段分别得到每个候选框的类别概率与每个候选框对特定目标的检测贡献率, 两者的内积作为各区域的得分, 最后根据得分来确定检测结果. 然而, WSDDN 模型的损失函数是非凸函数, 往往收敛到局部极小值, 相应的一些优化方法[127−130] 被提出用来解决该问题. Yang 等[131] 在文献 [128] 的基础上将多实例学习过程 (Multi-instance learning) 与模型训练过程连接成一个可端到端训练的网络, 并引入了空间注意力机制来获得更具有判别性的特征. Wan 等[132] 提出一种基于最小熵隐变量的弱监督模型, 它通过优化局部最小熵模型来估计伪标签 (Pseudo-objects) 和困难负例, 并充分利用这些信息来最小化学习过程中的随机性, 从而进一步提高模型的定位能力.

Zhou 等[133] 从目标的语义信息角度出发, 利用全局池化来得到每个特征层对每一类别物体的权重值, 并以此对特征图进行线性加权, 获得的高响应区域就是目标所在区域. 然而, 这种方式很容易导致模型只对目标最具有辨别力的部分进行检测, 从而降低定位的精度. 对此, Zhang 等[134] 提出了对抗互补学习的策略, 它通过交替训练两个不同的分类器, 得到互补的目标特征来进行拼接, 由此避免了上述问题, 提高了模型的定位能力. Choe 等[135] 在训练过程中随机生成掩码来遮挡目标整体范围内最显著的特征, 来引导模型学习目标完整的区域特征, 从而有效解决了目标定位误差较大的问题.

在数据标注成本越发昂贵的背景下, 弱监督目标检测的低廉成本使其受到了更多研究人员的关注. 如何进一步缩小弱监督检测模型与通用检测模型间的性能差距, 是未来该方向的研究重点.

5 结语

本文归纳分析了目标检测模型的子模块优化方法, 并对目标检测领域未来发展的方向进行了展望, 从中得出了以下结论:

1) 不同的检测场景与任务对模型性能的要求不同, 应根据具体场景与任务特点对模型做相应的改进, 并在专业数据集上对其进行训练与测试.

2) 主干网络和颈部连接层的结构优化能够抽取更好的目标特征, 因而几乎在任何场景、对任何模型的检测性能提升都是非常有利的.

3) 当检测场景中目标分布密集, 相互遮挡问题严重时, 非极大值抑制算法和交并比算法的优化能够有效缓解目标漏检的问题, 从而提升模型对密集目标的检测性能.

4) 当检测场景相对复杂, 背景嘈杂时, 正负样本采样算法和损失函数设计的优化能提升模型从各类训练样本中学习的效果, 进而提高模型对困难目

标的检测效果.

5) 目标检测领域在未来将围绕更优的检测性能与更好的工程落地两个不同方向, 从自动化、轻量化、域适应、少样本、弱监督等角度展开进一步深入研究.

References

1 Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, **60**(2): 91−110

2 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Kauai, HI, USA: IEEE, 2001. I−511−I−518

3 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Diego, CA, USA: IEEE, 2005. 886−893

4 Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2016, **38**(1): 142−158

5 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1440−1448

6 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, **39**(6): 1137−1149

7 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot MultiBox detector. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 21−37

8 Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 779−788

9 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 6517−6525

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

11 Law H, Deng J. CornerNet: Detecting objects as paired keypoints. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 765−781

12 Zhou X Y, Zhuo J C, Krähenbühl P. Bottom-up object detection by grouping extreme and center points. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 850−859

13 Zhou X Y, Wang D Q, Krähenbühl P. Objects as points. arXiv: 1904.07850, 2019

14 Yang Z, Liu S H, Hu H, Wang L W, Lin S. RepPoints: Point set representation for object detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 9656−9665

15 Zhang S F, Wen L Y, Bian X, Lei Z, Li S Z. Single-shot refinement neural network for object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 4203−4212

16 Chi C, Zhang S F, Xing J L, Lei Z, Li S Z, Zou X D. Selective refinement network for high performance face detection. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA: AAAI, 2019. 8231−8238

17 Li Z M, Peng C, Yu G, Zhang X Y, Deng Y D, Sun J. Lighthead R-CNN: In defense of two-stage object detector. arXiv: 1711.07264, 2017

18 Simonyan K, Zisseman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014

19 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 770−778

20 Xie S N, Girshick R, Dollár P, Tu Z W, He K M. Aggregated residual transformations for deep neural networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 5987−5995

21 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015. 1−9

22 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 7132−7141

23 Wang X L, Girshick R, Gupta A, He K M. Non-local neural networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 7794−7803

24 Cao Y, Xu J R, Lin S, Wei F Y, Hu H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. arXiv: 1904.11492, 2019

25 Woo S, Park J, Lee J Y, So Kweon I. CBAM: Convolutional block attention module. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 3−19

26 Lin T Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 936−944

27 Pan J M, Chen K, Shi J P, Feng H J, Ouyang W N, Lin D H. Libra R-CNN: Towards balanced learning for object detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 821−830

28 Liu S, Qi L, Qin H F, Shi J P, Jia J Y. Path aggregation network for instance segmentation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 8759−8768

29 Zhao Q J, Sheng T, Wang Y T, Tang Z, Chen Y, Cai L, et al. M2det: A single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA: AAAI, 2019. 9259−9266

30 Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection. arXiv: 1911.09070, 2020

31 Newell A, Yang K Y, Deng J. Stacked hourglass networks for human pose estimation. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). Amsterdam, The

Netherlands: Springer, Cham, 2016. 483−499

32 Sun K, Xiao B, Liu D, Wang J D. Deep high-resolution representation learning for human pose estimation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 5686−5696

33 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv: 1511.07122, 2016

34 Zhu C C, Tao R, Luu K, Savvides M. Seeing small faces from robust anchor′s perspective. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 5127−5136

35 Xie L L, Liu Y L, Jin L W, Xie Z C. DeRPN: Taking a further step toward more general object detection. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA: AAAI, 2019. 9046−9053

36 Zhong Y Y, Wang J F, Peng J, Zhang L. Anchor box optimization for object detection. arXiv: 1812.00469, 2020

37 Wang J Q, Chen K, Yang S, Loy C C, Lin D H. Region proposal by guided anchoring. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 2960−2969

38 Dai J F, Qi H Z, Xiong Y W, Zhang G D, Hu H, Wei Y C. Deformable convolutional networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 764−773

39 Zhu X Z, Hu H, Lin S, Dai J F. Deformable ConvNets V2: More deformable, better results. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 9300−9308

40 Tian Z, Shen C H, Chen H, He T. FCOS: Fully convolutional one-stage object detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 9626−9635

41 Kong T, Sun F C, Liu H P, Jiang Y N, Li L, Shi J B. FoveaBox: Beyond anchor-based object detector. arXiv: 1904.03797, 2020

42 Zhu C C, He Y H, Savvides M. Feature selective anchor-free module for single-shot object detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 840−849

43 Zhu C C, Chen F Y, Shen Z Q, Savvides M. Soft anchor-point object detection. arXiv: 1911.12448, 2020

44 Bodla N, Singh B, Chellappa R, Davis L S. Soft-NMS-improving object detection with one line of code. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICVV). Venice, Italy: IEEE, 2017. 5562−5570

45 He Y H, Zhang X Y, Savvides M, Kitani K. Softer-NMS: Rethinking bounding box regression for accurate object detection. arXiv: 1809.08545, 2019

46 Jiang B R, Luo R X, Mao J Y, Xiao T T, Jiang Y N. Acquisition of localization confidence for accurate object detection. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 816−832

47 Liu Y, Liu L Q, Rezatofighi H, Do T T, Shi Q F, Reid I. Learning pairwise relationship for multi-object detection in crowded scenes. arXiv: 1901.03796, 2019

48 Rezatofighi H, Tsoi N, Gwak J Y, Sadeghian A, Reid L, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 658−666

49 Cai Z W, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 6154−6162

50 Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 761−769

51 Yu H, Zhang Z N, Qin Z, Wu H, Li D S, Zhao J, et al. Loss Rank Mining: A general hard example mining method for real-time detectors. In: Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, Brazil: IEEE, 2018. 1−8

52 He K M, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2020, **42**(2): 386−397

53 Dai J F, Li Y, He K M, Sun J. R-FCN: Object detection via region-based fully convolutional networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS). Barcelona, Spain: Curran Associates Inc., 2016. 379−387

54 Zhu Y S, Zhao C Y, Wang J Q, Zhao X, Wu Y, Lu H Q. CoupleNet: Coupling global structure with local parts for object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 4146−4154

55 Zhai Y, Fu J J, Lu Y, Li H Q. Feature selective networks for object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 4139−4147

56 Chen Y T, Han C X, Wang N Y, Zhang Z X. Revisiting feature alignment for one-stage object detection. arXiv: 1908.01570, 2019

57 Cheng B W, Wei Y C, Shi H H, Feris R, Xiong J J, Huang T. Revisiting RCNN: On awakening the classification power of faster RCNN. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer. 2018. 473−490

58 Cheng B W, Wei Y C, Feris R, Xiong J J, Hwu W M, Huang T, et al. Decoupled classification refinement: Hard false positive suppression for object detection. arXiv: 1810.04002, 2020

59 Bell S, Zitnick C L, Bala K, Girshick R. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 2874−2883

60 Ouyang W L, Luo P, Zeng X Y, Qiu S, Tian Y L, Li H S, et al. DeepID-Net: Multi-stage and deformable deep convolutional neural networks for object detection. arXiv: 1409.3505, 2014

61 Chen X L, Gupta A. Spatial memory for context reasoning in object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 4106−4116

62 Liu Y, Wang R P, Shan S G, Chen X L. Structure inference net: Object detection using scene-level context and instance-level relationships. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 6985−6994

63 Hu H, Gu J Y, Zhang Z, Dai J F, Wei Y C. Relation networks for object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 3588−3597

64 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS). Long Beach, California, USA: Curran Associates Inc., 2017. 6000–6010

65 Gidaris S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware CNN model. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1134–1142

66 Zeng X Y, Ouyang W L, Yang B, Yan J J, Wang X G. Gated bi-directional CNN for object detection. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 354–369

67 Luo W J, Li Y J, Urtasun R, Zemek R. Understanding the effective receptive field in deep convolutional neural networks. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS). Barcelona, Spain: Curran Associates Inc., 2016. 4905–4913

68 Lin T Y, Goyal P, Girshick R, He K M, Dollár P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2999–3007

69 Cai Z W, Fan Q F, Feris R S, Vasconcelos N. A unified multi-scale deep convolutional neural network for fast object detection. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 354–370

70 Yang F, Choi W, Lin Y Q. Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 2129–2137

71 Singh B, Davis L S. An analysis of scale invariance in object detection - SNIP. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 3578–3587

72 Singh B, Najibi M, Davis L S. SNIPER: Efficient multi-scale training. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS). Montreal, Canada: Curran Associates Inc., 2018. 9333–9343

73 Najibi M, Singh B, Davis L S. AutoFocus: Efficient multi-scale inference. arXiv: 1812.01600, 2018

74 Li Y H, Chen Y T, Wang N Y, Zhang Z X. Scale-aware trident networks for object detection. arXiv: 1901.01892, 2019

75 Chen L C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017

76 Chen K, Li J G, Lin W Y, See J, Wang J, Duan L Y, et al. Towards accurate one-stage object detection with AP-Loss. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 5114–5122

77 Li B Y, Liu Y, Wang X G. Gradient harmonized single-stage detector. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA: AAAI, 2019. 8577–8584

78 Wang X L, Xiao T T, Jiang Y N, Shao S, Sun J, Shen C H. Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 7774–7783

79 Yu J H, Jiang Y N, Wang Z Y, Cao Z M, Huang T. UnitBox: An advanced object detection network. In: Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands: ACM, 2016. 516–520

80 Tychsen-Smith L, Petersson L. Improving object localization with fitness NMS and bounded IoU loss. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 6877–6885

81 Ma J Q, Shao W Y, Ye H, Wang L, Wang H, Zheng Y B, et al. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 2018, **20**(11): 3111–3122

82 Liao M H, Shi B G, Bai X. TextBoxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 2018, **27**(8): 3676–3690

83 Zhou X Y, Yao C, Wen H, Wang Y Z, Zhou S C, He W R, et al. EAST: An efficient and accurate scene text detector. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 2642–2651

84 Huang L C, Yang Y, Deng Y F, Yu Y N. DenseBox: Unifying landmark localization with end to end object detection. arXiv: 1509.04874, 2015

85 Deng D, Liu H F, Li X L, Cai D. PixelLink: Detecting scene text via instance segmentation. arXiv: 1801.01315, 2018

86 Xie E Z, Zang Y H, Shao S, Yu G, Yao C, Li G Y. Scene text detection with supervised pyramid context network. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA: AAAI, 2019. 9038–9045

87 Wang W H, Xie E Z, Li X, Hou W B, Lu T, Yu G, et al. Shape robust text detection with progressive scale expansion network. arXiv: 1903.12473, 2019

88 Xia G S, Bai X, Ding J, Zhu Z, Belongie S, Luo J B, et al. DOTA: A large-scale dataset for object detection in aerial image. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 3974–3983

89 Li K, Wan G, Cheng G, Meng L Q, Han J W. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, **159**: 296–307

90 Yang F, Fan H, Chu P, Blasch E, Ling H B. Clustered object detection in aerial images. arXiv: 1904.08008, 2019

91 Ding J, Xue N, Long Y, Xia G S, Lu Q K. Learning RoI transformer for oriented object detection in aerial images. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 2844–2853

92 Qian W, Yang X, Peng S L, Guo Y, Yan J C. Learning modulated loss for rotated object detection. arXiv: 1911.08299, 2019

93 Zhu Y X, Wu X Q, Du J. Adaptive period embedding for representing oriented objects in aerial images. arXiv: 1906.09447, 2019

94 Zhang S S, Benenson R, Schiele B. CityPersons: A diverse dataset for pedestrian detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 4457–4465

95 Shao S, Zhao Z J, Li B X, Xiao T T, Yu G, Zhang X Y, et al. CrowdHuman: A benchmark for detecting human in a crowd. arXiv: 1805.00123, 2018

96 Pang Y W, Xie J, Khan M H, Anwer R M, Khan F S, Shao L. Mask-guided attention network for occluded pedestrian detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 4966–4974

97 Zhang S F, Wen L Y, Bian X, Lei Z, Li S Z. Occlusion-aware R-

CNN: Detecting pedestrians in a crowd. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 657−674

98 Liu T R, Luo W H, Ma L, Huang J J, Stathaki T, Dai T H. Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling. arXiv: 1912.08661, 2019

99 Liu S T, Huang D, Wang Y H. Adaptive NMS: Refining pedestrian detection in a crowd. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6452−6461

100 Lu X, Li B Y, Yue Y X, Li Q Q, Yan J J. Grid R-CNN. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 7355−7364

101 Zoph B, Vasudevan V, Shlens J, Le Q V. Learning transferable architectures for scalable image recognition. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 8697−8710

102 Chen Y K, Yang T, Zhang X Y, Meng G F, Xiao X Y, Sun J. DetNAS: Backbone search for object detection. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NIPS). Vancouver, Canada: Margan Kaufmann Publishers, 2019. 6638−6648

103 Ghiasi G, Lin T Y, Le Q V. NAS-FPN: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 7029−7038

104 Wang N, Gao Y, Chen H, Wang P, Tian Z, Shen C H, et al. NAS-FCOS: Fast neural architecture search for object detection. arXiv: 1906.04423, 2020

105 Fang J M, Sun Y Z, Peng K J, Zhang Q, Li Y, Liu W Y, et al. Fast neural network adaptation via parameter remapping and architecture search. arXiv: 2001.02525, 2020

106 Karlinsky L, Shtok J, Harary S, Schwartz E, Aides A, Feris R, et al. RepMet: Representative-based metric learning for classification and few-shot object detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 5192−5201

107 Fan Q, Zhuo W, Tang C K, Tai Y W. Few-shot object detection with attention-RPN and multi-relation detector. arXiv: 1908.01998, 2020

108 Wang T, Zhang X P, Yuan L, Feng J S. Few-shot adaptive faster R-CNN. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 7166−7175

109 Yan X P, Chen Z L, Xu A N, Wang X X, Liang X D, Lin L. Meta R-CNN: Towards general solver for instance-level low-shot learning. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 9576−9585

110 Kang B Y, Liu Z, Wang X, Yu F, Feng J S, Darrell T. Few-shot object detection via feature reweighting. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 8419−8428

111 Inoue N, Furuta R, Yamasaki T, Aizawa K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 5001−5009

112 RoyChowdhury A, Chakrabarty P, Singh A, Jin S Y, Jiang H Z, Cao L L, et al. Automatic adaptation of object detectors to new domains using self-training. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 780−790

113 Kim S, Choi J, Kim T, Kim C. Self-training and adversarial background regularization for unsupervised domain adaptive one-stage object detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 6091−6100

114 Chen Y H, Li W, Sakaridis C, Dai D X, van Gool L. Domain adaptive faster R-CNN for object detection in the wild. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 3339−3348

115 Saito K, Ushiku Y, Harada T, Saenko K. Strong-Weak distribution alignment for adaptive object detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 6949−6958

116 Zhu X G, Pang J M, Yang C Y, Shi J P, Lin D H. Adapting object detectors via selective cross-domain alignment. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 687−696

117 Zhu J Y, Park T, Isola P, Efros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 2242−2251

118 Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861, 2017

119 Sandler M, Howard A, Zhu M L, Zhmoginov A, Chen L C. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 4510−4520

120 Howard A, Sandler M, Chen B, Wang W J, Chen L C, Tan M X, et al. Searching for MobileNetV3. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 1314−1324

121 Zhang X Y, Zhou X Y, Lin M X, Sun J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 6848−6856

122 Ma N N, Zhang X Y, Zheng H T, Sun J. ShuffleNet V2: Practical guidelines for efficient CNN architecture design. In: Proceedings of the 2018 European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 122−138

123 Zhang T, Qi G J, Xiao B, Wang J D. Interleaved group convolutions. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 4383−4392

124 Qin Z, Li Z M, Zhang Z N, Bao Y P, Yu G, Peng Y X, et al. ThunderNet: Towards real-time generic object detection on mobile devices. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 6717−6726

125 Zhang P Y, Zhong Y X, Li X Q. SlimYOLOv3: Narrower, faster and better for real-time UAV applications. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea: IEEE, 2019. 37−45

126 Bilen H, Vedaldi A. Weakly supervised deep detection networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV,

USA: IEEE, 2016. 2846−2854

127 Kantorov V, Oquab M, Cho M, Laptev I. ContextLocNet: Context-aware deep network models for weakly supervised localization. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 350−365

128 Tang P, Wang X G, Bai S, Shen W, Bai X, Liu W Y, et al. PCL: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2020, **42**(1): 176−191

129 Diba A, Sharma V, Pazandeh A, Pirsiavash H, van Gool L. Weakly supervised cascaded convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017. 5131−5139

130 Li Y, Liu L Q, Shen C H, van den Hengel A. Image co-localization by mimicking a good detector′s confidence score distribution. In: Proceedings of the 2016 European Conference on Computer Vision (ECCV). Amsterdam, The Netherlands: Springer, 2016. 19−34

131 Yang K, Li D S, Dou Y. Towards precise end-to-end weakly supervised object detection network. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE, 2019. 8371−8380

132 Wan F, Wei P X, Jiao J B, Han Z J, Ye Q X. Min-entropy latent model for weakly supervised object detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 1297−1306

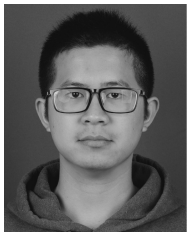
133 Zhou B L, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. 2921−2929

134 Zhang X L, Wei Y C, Feng J S, Yang Y, Huang T. Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 1325−1334

135 Choe J, Shim H. Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019. 2214−222

蒋弘毅 南京理工大学机械工程学院硕士研究生. 主要研究方向为图像处理与目标检测.
E-mail: jianghongyi_1996@163.com
(**JIANG Hong-Yi** Master student at the College of Mechanical Engineering, Nanjing University of Science and Technology. His research interest covers image processing and object detection.)

王永娟 南京理工大学机械工程学院教授. 主要研究方向复杂与智能机械系统设计. 本文通信作者.
E-mail: 13951643935@139.com
(**WANG Yong-Juan** Professor at the College of Mechanical Engineering, Nanjing University of Science and Technology. Her main research interest is design of complex and intelligent mechanical systems. Corresponding author of this paper.)



康锦煜 南京理工大学机械工程学院硕士研究生. 主要研究方向为穿戴式智能单兵系统设计.
E-mail: njust@3dgarms.com
(**KANG Jin-Yu** Master student at the College of Mechanical Engineering, Nanjing University of Science and Technology. His main research interest is design of wearable intelligent soldier system.)