DOI:10.19651/j.cnki.emt.2209992

一种融合卷积与 transformer 的级联包检测方法

罗晓霞[1] 蒋 磊[1] 蔡院强[2]
(1.西安科技大学计算机科学与技术学院 西安 710054; 2.北京邮电大学网络与交换技术国家重点实验室 北京 100876)

摘 要: 为解决目前包检测算法检测类别单一、准确度较低、复杂目标难以检测等问题,研究了一种融合卷积与 transformer 的级联包检测方法,CT-CBDet。首先,设计了 deformable conformer 作为骨干网络进行特征提取,其在 transformer 与卷积双网络融合的基础上利用可形变卷积和空间金字塔池化模块实现几何特征变换与多尺度特征融合,以强化针对复杂特征的建模能力;然后,提出一种基于 anchor 统计特征的自适应正负样本选择的区域建议网络,以平衡不同尺度目标样本正负选择的公平性,增强模型的训练稳定性;最后,利用多阶段损失对模型的级联检测组件进行端到端训练。结果表明,该方法相较于基准方法 Cascade RCNN 平均精度值提高了 5.8%,小尺度目标检测精度提高了 10.9%。可见 CT-CBDet 可有效完成复杂场景下的包检测任务。

关键词: 包检测;级联架构;特征融合;自适应区域建议网络;deformable conformer
中图分类号: TP391 **文献标识码**: A **国家标准学科分类代码**: 520.6040

Cascade bag detection method combining convolution and transformer

Luo Xiaoxia[1] Jiang Lei[1] Cai Yuanqiang[2]
(1. College of Computer Science and Technology, Xi'an University of Science and Technology, Xi'an 710054, China;
2. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: In order to solve the problems of single detection category, low detection accuracy and difficult detection of complex objects, a cascaded bag detection method integrating convolution and transformer is studied. CT-CBDet First, a deformable conformer is designed as a backbone network for feature extraction, which uses deformable convolution and spatial pyramid pooling modules to achieve geometric feature transformation and multi-scale feature fusion on the basis of the fusion of transformer and convolutional double network. feature modeling ability; then, a region proposal network with adaptive positive and negative sample selection based on anchor statistical features is proposed to balance the fairness of positive and negative selection of object samples at different scales and enhance the training stability of the model; finally, the cascade detection component of the model is trained end-to-end using multi-stage loss. The results show that the method improves the mAP by 5.8% and the small-scale object detection accuracy by 10.9% compared to the baseline method Cascade RCNN. It can be seen that CT-CBDet can effectively perform the bag detection task in complex scenes.

Keywords: bag detection; cascade architecture; feature fusion; adaptive region proposal network; deformable conformer

0 引 言

包作为一件物品被广泛使用,包检测在计算机视觉领域备受关注。一方面,包检测可以为其他视觉任务提供辅助作用,如在行人属性识别[1-2]任务中,包检测用于判断行人是否携带包;在行人重识别[3]任务中,包检测可以为重识别提供特征依据;包检测还在计算机视觉中以子任务的形式出现任务,如步态识别[4-5]中,带包行走作为一种行人状态出现在步态识别数据集 CASIA Dataset B 中;包通常也作为常规目标检测的一个子类别出现在检测任务中。另一方面,包检测还可以在实际的社会生产环境尤其是智能视频监控系统中起到至关重要的作用,如在公安智能监控系统中,包检测对于罪犯信息提取起着重要作用,可以通过监控摄像机捕捉的画面信息,利用包检测技术可以更加细化对相关的人员的语义描述为嫌疑人排查、目标跟踪提供辅助[6]。此外,在诸如交通枢纽或商场等人口密集地区的安

收稿日期:2022-05-18

防系统中，可以通过对无人看管包的检测辅助潜在危险物品排查或及时检测乘客遗留行李以减少人员损失[7]。

包检测算法大致分为两类：传统的图像处理方法和基于深度学习的方法。Chang 等[8]提出一种协同相机对系统并结合基于霍夫变换和边缘检测方法定义几何特征的背包检测算法，其摒弃了从人体轮廓中获取形状信息的方法。Lkhagvasurena 等[9]提出了一种利用模糊规则在人体轮廓内部或附近寻找包区域的检测算法。包检测可认为是目标检测的子任务，随着深度学习的发展，基于目标检测技术的包检测方法逐步取代了传统方法，如 Du 等[10]提出一种两阶段的行人图像包检测算法，其先使用 SVF 算法检测图像是否含有包，然后使用选择性搜索和卷积神经网络定位包区域；过斌[11]提出了基于 YOLO-G 的分布式检测模型完成了遗留包的检测；冯勇等[12]利用差分法和 YOLO 模型实现了列车遗留行李检测。张俊为[13]提出一种改进 YOLOv2 的方法，基于 PETS2006 实现了行李箱的检测。Dogariu 等[14]基于 Mask RCNN 和实现了手提箱和背包的检测。虽然上述方法都可以在一定程度上完成包检测任务，但也存在着一些不足。如传统方法难以提取图像深层的语义特征，从而对于遮挡、小尺度等特征不明显的目标难以识别且定位不准确，另一方面，由于其依赖于手工特征，从而难以满足多类别包检测，算法检测类别较为单一，仅涉及到单类包的检测（如双肩包、挎包、手推车）。而目前流行的目标检测算法，虽然通过深度神经网络强化了对于图像深层语义特征的提取能力，但也加剧了由于数据集的不确定性造成模型训练不稳定的问题，同时也面临着建模等复杂特征的挑战，对于模糊、形变、光照变化等具有复杂特征的目标检测能力不足。因此，为增强模型针对复杂特征的建模能力，提高模型训练稳定性，解决现实复杂场景中包检测问题，现研究了一种融合卷积与 transformer 的级联包检测方法（convolution-transformer cascade bag detector，CT-CBDet），主要工作如下。

1）为了增强模型的几何特征变换能力，设计了 deformable conformer 网络来提高关键信息的特征提取能力，并通过空间金字塔池化模块（spatial pyramid pooling，SPP）完成特征融合，丰富复杂场景下语义特征表示，进一步增强全局特征建模能力。

2）基于训练过程中的样本统计特征设计了一种的具有自适应 anchor 正负样本选择能力的区域建议网络，均标区域建议网络（region proposal network with mean and standard deviation，MSD-RPN）。其可以更好地平衡不同尺度目标正负样本选择的公平性，从而完成高质量区域建议框的获取，提高检测准确度。

3）与主流的各类检测算法（two-stage、one-stage、anchor-based 和 anchor-free）进行多维度实验比较，结果表明，使用所提出的算法检测准确度更高，可以较好地完成复杂场景包检测任务。同时，也通过多角度的消融实验来验证各改进点对于包检测的不同影响。

1　包检测模型设计

为解决现实复杂场景中包检测问题，基于 Cascade RCNN[15]架构提出一种融合卷积与 transformer 的级联包检测方法，CT-CBDet，如图 1 所示，首先基于 deformable conformer 和特征金字塔[16]（feature pyramid networks，FPN）进行特征提取；然后，通过 MSD-RPN 使用自适应 anchor 正负样本选择策略获得高质量区域建议框；最后，在级联多级检测器上进行目标定位和分类。

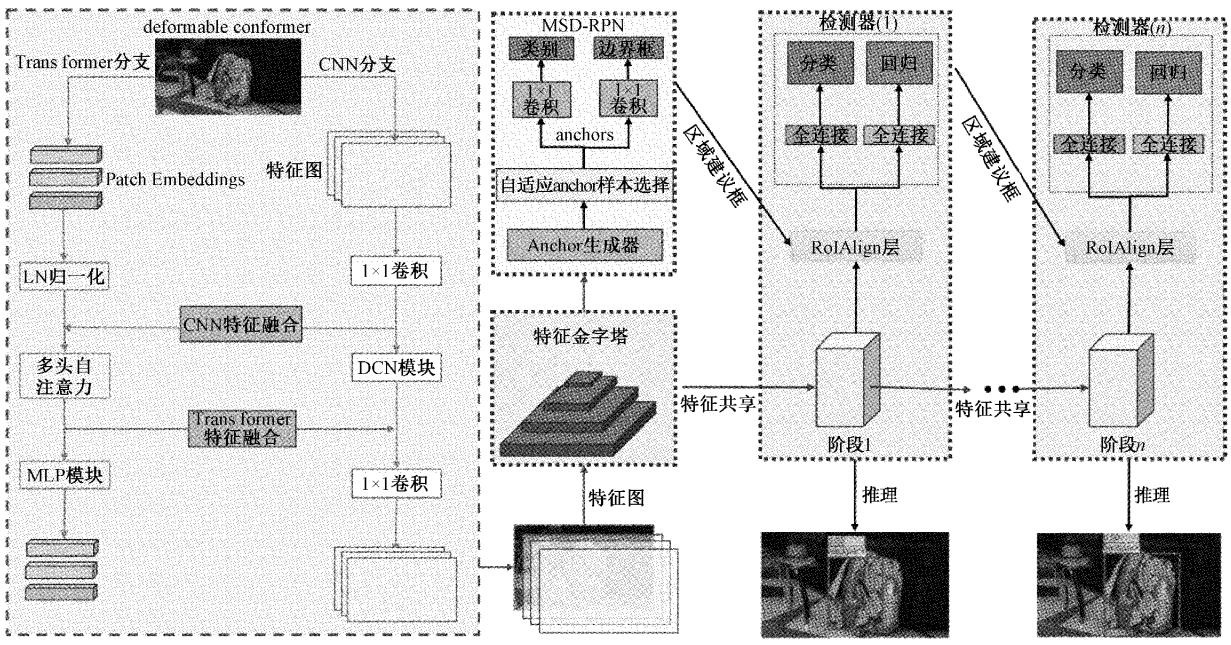


图 1　CT-CBDet 网络结构图

1.1 deformable conformer

由于卷积网络对图像具有较高的特征敏感性,可以较好地提取图像中的局部特征信息,却难以对全局特征信息进行建模。transformer 模型(如 Swin[17],PVT[18])由于自注意力机制,使其更加擅长于捕获图像中的全局特征信息。但是却难以收敛,较为依赖大规模的训练数据集。基于上述题,部分学者提出了卷积和 transformer 相结合的网络模型从而进行局部与全局信息的特征交互与建模,如 Mobile-Former[19],conformer[20]。虽然通过卷积与 transformer 的并行结合可以在一定程度上改善图像中局部信息与全局信息的交互问题,但是仍然缺少能够处理不同目标几何变换的机制。另一方面,在 conformer 模型中,特征的交互依赖于特征耦合单元,transformer 分支能够建模的深层特征信息来源于卷积分支,由于卷积特征提取的局限性仍然会限制 transformer 分支的全局特征建模能力。基于上述两个问题,本文基于 conformer 模型提出了 deformable conformer,如图 2 所示,由卷积分支和 transformer 分支并行组织的双重网络结构,并通过特征耦合单元完成两个网络分支的信息交流。在 deformable conformer 的卷积分支中利用可形变卷积[21](deformable convolutional networks, DCN)实现特征空间的几何变换机制并通过 SPP 模块丰富特征的语义表示,进一步增强模型对全局信息以及局部特征信息的特征建模。

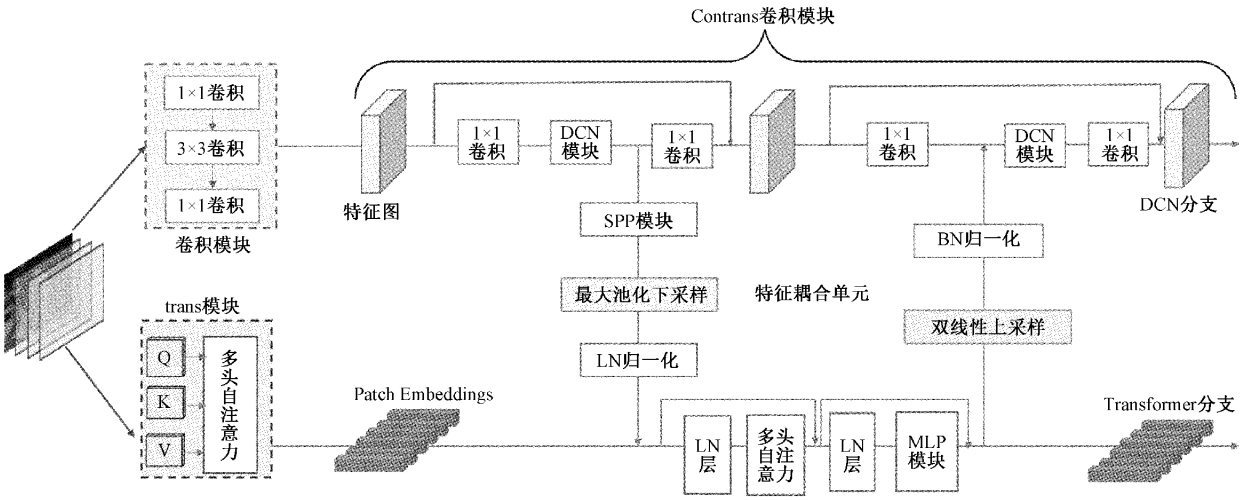


图 2 deformable conformer 网络结构图

1)基于可形变卷积的几何变换机制

本文通过可形变卷积来解决目标的几何变换问题,通过预测卷积核的形变偏移量来调整卷积核对于关键像素点的采样位置,从而捕捉不同形变目标的空间变换。因此在 conformer 的卷积分支中引入了可形变卷积来提高模型对不同尺度目标的建模能力,对于不同位置特征值计算如式(1)所示。

$$y(p_0)=\sum_{p_n\in R}w(p_n)x(p_0+p_n+\Delta P_n) \tag{1}$$

式中:p_0 为特征采样位置,Δp 表示偏移量,p_n 是对卷积采样位置 R 的枚举,w 为特征权重,x 为特征值表达式。

本文在 conformer 的卷积分支中,使用可形变卷积替代了原始的 3×3 卷积。如图 3 所示,通过 3×3×18 的卷积计算偏移量,最终将偏移量映射到原特征图和 3×3 的卷积上产生形变卷积。在卷积分支中加上可形变偏移量后,使得卷积核的大小和位置可以适应几何变换,从而能根据当前图像的特征内容进行动态调整。由此,在 deformable conformer 中,卷积分支可以根据图像特征内容自适应改变卷积核的位置及大小,从而适应图像中的不同形状及大小目标的几何特征变换,增强模型几何特征变换适应能力,强化关键信息的特征表示。

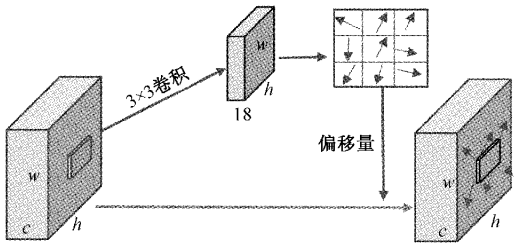


图 3 DCN 结构示意图

2)基于 SPP 的特征融合机制

由于卷积分支与 transformer 分支的特征交互依赖于特征耦合单元,即在一个卷积模块后,先使用下采样进行特征对齐再融合 transformer 分支的语义特征。因此,transformer 分支能够建模的深层语义特征极大地依赖于卷积分支所能融合的特征信息。然而,由于卷积特征提取的局部性仍然会限制 transformer 的全局特征提取能力。

基于上述问题,受 YOLOv3-SPP[22]启发,本文在卷积分支特征耦合前加入 SPP 模块以提取不同尺度下的空间语义特征,在融合多尺度空间特征后经过特征耦合单元与

transformer 分支进行特征交互，从而极大程度完成卷积分支的局部特征与 transformer 分支的全局特征融合，全面提升模型的特征建模能力。SPP 结构如图 4 所示，在一个卷积模块完成特征提取后先通过 1×1 卷积进行通道降维，然后使用三个不同步距的最大池化层提取不同尺度的特征信息，接下来再将经过最大池化采样的三个特征层与原始特征图进行叠加，最后再通过 1×1 卷积进行特征对齐。实现了大步距下提取的全局信息与小步距下提取的局部信息的特征融合。由此，经过 SPP 模块得到的特征图不仅保留了原始的特征语义信息还融合了多尺度的空间特征，最终通过特征耦合单元与 transformer 分支进行特征交互可以进一步强化 transformer 分支的全局空间感知能力。

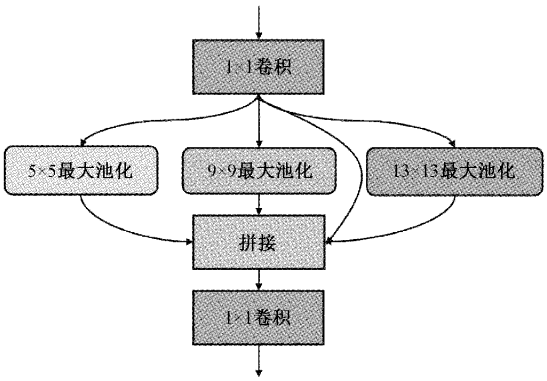


图 4　SPP 模块结构

3）基于双线性上采样的特征耦合单元

由在 conformer 网络中，由于 transformer 分支与卷积分支特征维度的不对等性，因此从 transformer 分支到卷积分支的特征交互依赖于特征耦合单元的上采样机制，在原始网络中采用最近邻差值算法实现 transformer 模块特征的上采样，而最近邻差值的上采样方法往往会导致图像边缘的不连续，会损失较多的图像特征。另一方面，由于在真实的自然场景中，往往存在着诸如多尺度、不均匀光照、相似纹理、部分遮罩、模糊、形变等复杂特征，最近邻差值的上采样方法更会加剧小尺度目标的特征损失。因此，本文使用双线性差值算法进行图像上采样完成特征对齐，从而避免采样后出现的灰度不连续，以保留更多的特征。

1.2　MSD-RPN

Faster RCNN[23] 提出了区域建议网络（region proposal network，RPN）获取区域建议框，提高了模型的检测效率与准确度。在 RPN 中，基于 anchor 与 ground-truth box 的 IoU 的正负样本匹配方式往往会更加倾向于为尺度较大的目标分配较多的正样本，从而更倾向于对大目标所匹配到的 anchor 进行训练，从而导致模型对小尺度目标检测能力不足。另一方面，区域建议网络中 anchor 的预设比例一定程度上决定了区域建议框的质量，因此，为了使区域建议网络达到更好的表现就需要根据数据集特征定义多个不同长宽比例的 anchor 以获取高质量的区域建议框，从而造成模型的不稳定性。

基于上述两个问题，受到 Zhang 等[24] 启发，基于 anchor 的统计特征设计了 MSD-RPN 以实现 anchor 正负样本的自适应分配。其步骤如下：

步骤 1）先选取距离每个 ground-truth box 最近的 9 个 anchor，采用欧式距离进行度量；

步骤 2）计算 ground-truth box 与其候选 anchor 的 IoU 值，将正负样本匹配阈值确定为 IoU 值的均值和标准差和；

步骤 3）将中心点落在 ground-truth box 内，且与该 ground-truth box 的 IoU 值不小于阀值 anchor 确定为正样本，其余则为负样本。

MSD-RPN 避免了网络模型超参数设置，并且由于 anchor 与 ground-truth box 的均值反应了二者匹配的总体情况，标准差反应了 anchor 与 ground-truth box 匹配程度的离散程度，即差异性，根据样本的统计特征进行自适应分配，从而可以适应不同 anchor 条件下的分配，样本分配阈值会随着 anchor 整体情况自适应变化，高质量的 anchor 会产生较高的阈值，低质量的 anchor 会产生较低的阈值，避免了传统分配模式下低质量 anchor 正样本不足，高质量 anchor 负样本不足的缺点。同时，MSD-RPN 本分配方式对 anchor 的预设比例并不敏感，可以保证在不设置多个 anchor 尺度的情况下仍可以实现高质量 anchor 的筛选，不会倾向于为大目标分配更多的正样本。另一方面，MSD-RPN 将 anchor 的中心点限制在 ground-truth box 内，保证具有较大的 IoU 值，以产生高质量的区域建议框。

2　数据集及评价标准

在与视觉任务相关的大多数数据集中，如目标检测（如 MS-COCO[25]）、行人属性识别（如 PA-100k[26]）、步态识别（如 CASIA Dataset B）和遗留物检测（如 PETS 2007），包通常作为辅助属性出现在这些数据集中。目前，在包检测任务研究中，大多使用自建数据集或基于其他任务数据集再标注进行研究，本实验采取自建数据集的方式。

2.1　数据集

为了保证数据集一定程度的多样性与的复杂性，本实验从网络中收集候车厅、地铁站、街道、街道夜景、商场五个不同复杂程度的自然场景视频数据，采用关键帧提取技术转换为图像数据，然后将图像中的人脸区域加入高斯模糊进行隐私匿名化，接下来采用 LabelImg 开源标注软件对日常生活中的四类包（即，背包、挎包、手提袋和拉杆箱）进行位置和类别的标注，最终数据规模达到 10 000 张图像，其中正样本图像包含 6 505 张，标注实例为 21 457 个，数据集中包含背包 7 676 个，挎包 7 136 个，拉杆箱 1 564 个以及手提袋 5 081 个。

2.2 评价标准

本实验数据集采用 MS-COCO 数据集的评价标准,使用四类指标(平均精度、跨尺度的平均精度、平均召回率和跨尺度的平均召回率)来表征该检测器的准确度。平均精度有三种度量(即 AP、AP_{50} 和 AP_{75})。AP 的值为 10 个 IoU 阈值(即 IoU∈[0.50,0.95],步长为 0.05)下计算得到的 mAP 的平均值。同理,AP_{75}/AP_{50} 为 IoU 阈值取 0.75/0.5 的计算的 mAP 值。

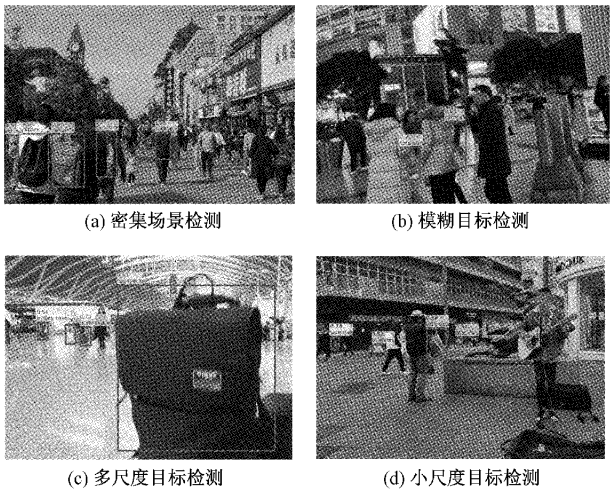
3 实验与结果

3.1 实验环境与参数

本实验所对比的方法均是通过 PyTorch 平台实现,所有的实验均是在 CPU 为 Intel(R) Xeon(R) CPU E5-1650 v3@3.50 GHz,GPU 为 GeForce RTX 3090 下进行训练与测试。对于 CT-CBDet 算法,MSD-RPN 在训练阶段每张图片产生 2 000 个区域建议框,并使用阈值为 0.7 进行非极大值抑制处理。在推理阶段,每张图片产生 1 000 个区域建议框且采用阈值为 0.7 进行非极大值抑制。整个网络采用随机梯度下降进行训练,动量值为 0.9,权重衰减为 0.000 1,进行 12 个 epoch,学习率为 0.02。采用 Warm up 训练方案,初始的学习率为原定学习率的 1/3,并在初始的 500 次迭代中线性增加,在第 8 和 11 个 epoch 时降低学习率。

3.2 实验结果分析

CT-CBDet 应用于不同复杂程度场景下的检测结果如图 5 所示,可以看出,算法可以准确识别不同类别的包;同时,对于不同复杂场景的目标都做到了较精准的识别。



(a) 密集场景检测 (b) 模糊目标检测

(c) 多尺度目标检测 (d) 小尺度目标检测

图 5 CT-CBDet 在不同场景下的检测结果

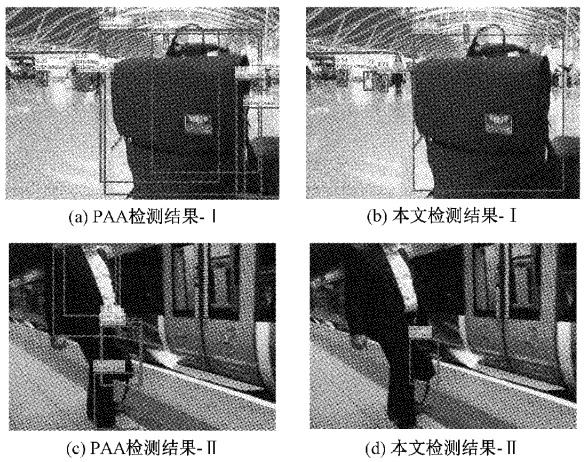
CT-CBDet 与经典检测方法的量化结果对比见表 1 所示,相较于基准算法 Cascade RCNN,在 3 个准确度指标上分别提高了 5.8%、7.3%和 7.1%。对于不同尺度目标的检测效果,也均得到了一定的提高,其中小尺度目标检测精度提高最为明显,达到了 23.3%,相较于 Cascade RCNN 提高了 10.9%。同时,在不同尺度的目标上实现了召回率的整体改善,其中小尺度目标的改进最为显着,与 Cascade RCNN 相比提高了 12.4%;相较于其他的两阶段算法,CT-CBDet 在各个指标上均有所提升,尤其是在 AP_{75} 指标上,改进的方法相较于基于 Swin 的 Cascade RCNN 提高了 5.8%,由此可见 CT-CBDet 可以实现更高精度的目标定位。值得注意的是,CT-CBDet 在小尺度目标的检测准确度和召回率上均得到了大幅度的提高,这是由于其卷积分支与 transformer 分支的共同作用,使模型同时兼顾了 Resnet[27] 局部建模与 Swin transformer 全局建模的优点。

表 1 实验结果

方法	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR_s	AR_m	AR_l
Faster RCNN	0.347	0.690	0.299	0.140	0.244	0.415	0.225	0.350	0.500
Grid RCNN	0.351	0.700	0.306	0.143	0.245	0.414	0.264	0.376	0.512
Cascade RCNN	*0.352*	*0.681*	*0.327*	*0.124*	*0.247*	*0.423*	*0.200*	*0.344*	*0.518*
Swin-Cascade	0.364	0.692	0.340	0.121	0.245	0.435	0.196	0.324	0.504
Retinanet	0.254	0.561	0.183	0.048	0.145	0.319	0.133	0.336	0.484
YOLOv3-SPP	0.278	0.599	0.217	0.073	0.139	0.368	0.134	0.235	0.459
FCOS	0.340	0.694	0.290	0.110	0.236	0.408	0.201	0.402	0.538
ATSS	0.363	0.714	0.326	0.130	0.254	0.429	0.220	0.424	0.562
PAA	0.371	0.719	0.338	0.147	0.252	0.439	0.267	**0.425**	**0.571**
CT-CBDet	**0.410**	**0.754**	**0.398**	**0.233**	**0.301**	**0.462**	**0.324**	0.402	0.540

同时,本实验也比较了经典的一阶段算法(即,YOLO v3-SPP、Retinanet[28]、ATSS、FCOS[29] 和 PAA[30])。CT-CBDet 在各个指标上均大幅度领先于其他的一阶段算法,并且在 AP_{75} 指标上的领先最为明显,相较于 YOLOv3-SPP 提高了 21.5%,相较于 PAA 提高了 6%由此可见,CT-CBDet 同样保持着高精度定位的优势。值得注意的是,所提算法在小尺度目标检测方面也保持着大幅度的领先,相较于 YOLOv3 SPP 和 PAA 分别提高了 16.0%和 8.6%;相较于 anchor free 方法 FCOS 领先了 12.3%。

此外，CT-CBDet 在中等尺度和大尺度目标的召回率指标上低于 PAA 检测器 2.3%和 3.1%，但是对于不同尺度目标的准确度指标上都存在着一定程度的领先，在召回率更低的情况下实现了更高的检测准确度，可见，CT-CBDet 可以更精准地识别目标，实现高质量特征建模。另一方面，本实验在推理阶段，设置 MSD-RPN 的 NMS 阈值为 0.7，PAA 的 NMS 阈值设置为 0.6，然而，如图 6 所示，PAA 检测器却存在非常严重地误检。由此可见，通过高斯混合模型的概率样本分配方式会使得模型训练不稳定，而在 MSD-RPN 中，即使采用了较高的 NMS 阈值，但仍然保留了高质量区域建议框，从而实现了精准的目标定位与回归。



(a) PAA检测结果-Ⅰ　(b) 本文检测结果-Ⅰ

(c) PAA检测结果-Ⅱ　(d) 本文检测结果-Ⅱ

图 6　PAA 与本文检测效果对比图

3.3　消融实验

CT-CBDet 是在 Cascade RCNN 基准算法上引入了自适应正负样本选择的区域建议网络，MSD-RPN 和基于 DCN、SPP 改进 conformer 的骨干网络 deformable conformer，为了进一步验证各改进成分的有效性及对包检测的不同影响，本实验对其进行组合实验，结果如表 2 所示，通过不同改进组件的引入，检测准确度会得到不同程度的提升。

表 2　消融实验

方法	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AR_s	AR_m	AR_l
Cascade RCNN	0.352	0.681	0.327	0.124	0.247	0.423	0.200	0.344	0.518
＋MSD-RPN	0.359	0.693	0.330	0.125	0.252	0.423	0.223	0.349	0.501
＋MSD-RPN＋conformer	0.397	0.734	0.386	0.191	0.295	0.456	0.263	0.391	0.532
＋MSD-RPN＋conformer＋SPP	0.398	0.740	0.385	0.199	0.296	0.459	0.299	0.396	0.535
＋MSD-RPN＋conformer＋DCN	0.408	0.747	**0.405**	0.224	0.299	**0.469**	0.283	0.395	0.539
＋MSD-RPN＋deformable-conformer	**0.410**	**0.754**	0.398	**0.233**	**0.301**	0.462	**0.324**	**0.402**	**0.540**

1）MSD-RPN

首先使用了 MSD-RPN 替代原始的 RPN，准确度指标上均得到了一定的提升，AP 指标提升了 0.7%，AP_{50} 指标提升了 1.2%。在自适应样本选择策略的作用下，小尺度目标的召回率提升了 2.3%，大尺度目标的召回率降低了 0.1%。虽然大尺度目标的召回率有所下降，但是检测精度保持不变。由此可见，MSD-RPN 可以在保证大尺度目标检测精度的同时提高小尺度目标的召回率。同时，也可以说明 MSD-RPN 为小尺度目标带来了更公平的 anchor 分配方案，不会倾向于为大尺度目标分配更多的正样本。

此外，在 MSD-RPN 中，仅设置了 anchor 比例为 1∶1，结果如表 3 所示，准确度与设置 3 种比例基本一致。由此可见，MSD-RPN 对于 anchor 的预设比例并不敏感，可以避免针对数据集目标的不同特征设计 anchor 比例，从而可以进一步增强模型的稳定性。

表 3　anchor 比例组合

比例组合			AP	AP_{50}	AP_{75}
1∶1	1∶2	2∶1			
√	√	√	0.359	0.693	0.330
√			0.359	0.692	0.330

2）MSD-RPN＋conformer

在引入 conformer 后，所有的检测指标均得到了大幅度的提升，尤其是小目标检测精度，提升了 6.6%。由此可见，基于卷积和 transformer 的双网络模型相较于卷积模型有更强的特征提取能力。并且由于特征耦合单元使得卷积提取的局部信息和 transformer 提取的全局特征信息可以得到充分交互，从而利于上下文特征建模。

3）MSD-RPN＋conformer＋SPP

在 conformer 的卷积分支向 transformer 分支进行特征耦合前加入 SPP 模块后，大部分指标均得到了一定的提升。其中，小尺度目标的召回率提升了 3.6%，由此可见，基于 SPP 的多尺度空间感知，极大程度完成卷积分支的局部特征与 transformer 分支的全局信息融合，对于强化局部特征信息表示起到了一定的积极作用。

此外，本实验还验证了基于空洞卷积的空间池化金字塔（atrous spatial pyramid pooling，ASPP）对检测结果的影响，结果如表 4 所示。ASPP 致力于感受野的提升，增强卷积模型的全局建模能力，但从结果看出，空洞卷积会丢失局部特征信息，从而会抑制 transformer 分支的全局建模能力，导致对小尺度目标的检测影响较大。

表 4 ASPP 与 SPP 的比较

方法	AP	AP_{50}	AP_{75}	APs	AP_m	AP_l
SPP	0.398	0.740	0.385	**0.199**	0.296	0.459
ASPP	0.390	0.726	0.375	0.161	0.290	0.456

4)MSD-RPN+conformer+DCN

在 conformer 的卷积分支中引入 DCN 模块后,对于不同尺度目标的检测准确度均得到了一定的提升,再结合 MSD-RPN 的样本选择机制,使得这个组合对于可以更加完成目标的高精度定位,其 AP_{75} 指标达到了 40.5%,相较于 conformer 提高了 1.9%。另一方面,对比表 2 的 4、5 行可看出,DCN 相较于 SPP 具有更强的特征建模能力,尤其是在小目标检测方面,准确度高于 SPP 组合 2.5%。

5)MSD-RPN+conformer+SPP+DCN

当 SPP 和 DCN 全部引入后,大部分指标均相较于其他组合均有一定的提升,尤其是在小尺度目标方面。由此可见,DCN 可以对 SPP 的多尺度特征融合起到一定的促进作用,使得 SPP 和 DCN 的组合可以得到更好的特征建模效果。另一方面,对比表 2 的最后两行可以看出,AP_{75} 和 AP_l 指标有所下降,这是由于 SPP 模块的多尺度的特征融合会在一定程度上增加大尺度目标的特征噪声,从而对目标高精度定位起到了一定的负面作用。

此外,由于 DCN 相较于原始 3×3 卷积会带来额外的性能消耗,因此,本实验对于可形变卷积在 Resnet 分支的不同阶段的使用位置进行了相关探索,结果见表 5 所示,从表中可以看出,DCN 在后三个阶段使用可以较好地取得效率与精度间的平衡。

表 5 DCN 在不同阶段使用的参数量

各阶段的组合				GFlops	参数量/MB	AP
S1	S2	S3	S4			
			√	301.64	80.89	0.399
		√	√	293.20	81.05	0.403
	√	√	√	285.75	81.68	0.409
√	√	√	√	261.92	84.53	0.410

3.4 实验结论

本实验在多个量化指标上对比了 CT-CBDet 与经典方法的准确度表现,结果表明,所提算法在多个指标上均实现了一定程度的领先;同时,通过算法在不同场景的应用结果可以看出,所提算法可以较好地实现不同复杂程度场景下各类别包的检测。

此外,从消融实验中可以看出:MSD-RPN 可有效提高小尺度目标的检测能力且对于 anchor 的预设比例不敏感;SPP 可以在一定程度上可以提高模型的特征表达能力,但在多尺度特征融合的同时也会带来一定的特征噪声;DCN 可以全方面增强模型在复杂场景下的特征建模能力,提高不同尺度目标的检测精度。并且,DCN 与 SPP 的组合可以实现小目标检测的最佳效果。

4 结 论

本文提出了一种融合卷积与 transformer 的级联检测模型,CT-CBDet。该模型包含一个新的骨干网络,deformable conformer 和区域建议网络,MSD-RPN。deformable conformer 在完成局部与全局特征信息交互的同时还实现了多尺度特征融合机制与特征几何变换机制,可以较好提取不同尺度目标的特征信息,尤其是对于小尺度目标的特征建模起到了至关重要作用。MSD-RPN 可以实现高质量的区域建议框的获取,较好地平衡不同尺度目标样本选择的公平性。实验结果表明,CT-CBDet 检测效果优秀,相较于其他经典算法具有更高的检测精度。

同时,本文所提的算法可以较好地完成复杂场景下的包检测任务,可满足自然场景下包检测的工程需求。下一步将轻量化模型,减少模型参数。

参考文献

[1] 张再腾,张荣芬,刘宇红. 基于多尺度残差注意网络的轻量级行人属性识别算法[J/OL]. 控制与决策,2022,37(10):10.

[2] HE K, WANG Z, FU Y, et al. Adaptively weighted multi-task deep network for person attribute classification[C]. ACM International Conference on Multimedia(MM), 2017: 1636-1644.

[3] 姚品,万旺根. 基于深度学习和属性特征的行人再识别算法[J]. 电子测量技术,2020,344(12): 70-74.

[4] 高经纬,马超,姚杰,等. 基于机器学习的人体步态检测智能识别算法研究[J]. 电子测量与仪器学报,2021,243(3): 49-55.

[5] 贾晓辉,王涛,刘今越,等. 基于人体模型映射的步态识别及意图感知方法[J]. 仪器仪表学报,2020,41(12): 236-244.

[6] 卫建华,刘润利,许佳豪,等. 基于 PYNQ 框架的人体目标跟踪系统[J]. 国外电子测量技术,2021,325(12): 89-95.

[7] 刘德健,吴金勇,王一科,等. 遗留物的检测方法及系统[P]. 广东:CN102663346A,2012-09-12.

[8] CHANG R, CHUA T W, LEMAN K, et al. Automatic cooperative camera system for real-time bag detection in visual surveillance [C]. International Conference on Distributed Smart Cameras (ICDSC), 2013: 1-6.

[9] LKHAGVASURENA U, KARUNGARUA S, TERADAA K. Real time backpack detection in visual surveillance based on verticals contour analysis[C]. International Conference on Industrial Application

Engineering(ICIAE), 2018.
[10] DU Y, AI H, LAO S. A two-stage approach for bag detection in pedestrian images[C]. Asian Conference on Computer Vision(ACCV), 2014: 507-521.
[11] 过斌. 基于深度学习的分布式遗留物检测[D]. 南昌: 南昌大学, 2021.
[12] 冯勇, 许华胜, 段旺旺, 等. 计算机视觉技术在列车行李遗留检测上的应用[J]. 铁道车辆, 2020, 58(9): 15-17.
[13] 张俊为. 基于改进 YOLO v2 网络的遗留物检测算法研究[D]. 杭州:浙江理工大学, 2018.
[14] DOGARIU M, STEFAN L D, CONSTAIN M G, et al. Human-object interaction: Application to abandoned luggage detection in video surveillance scenarios [C]. International Conference on Communications(COMM), 2020: 157-160.
[15] CAI Z, VASCONCELOS N. Cascade R-CNN: high quality object detection and instance segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI), 2019, 43(5): 1483-1498.
[16] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017: 2117-2125.
[17] LIU Z, LIN Y, CAO Y. Swin transformer: Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE International Conference on Computer Vision(ICCV), 2021: 10012-10022.
[18] WANG W, XIE E, LI X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions [C]. IEEE International Conference on Computer Vision (ICCV), 2021: 568-578.
[19] CHEM Y, DAI X, CHEM D, et al. Mobile-former: Bridging mobilenet and transformer [J]. ArXiv Preprint,2021, ArXiv:2108.05895.
[20] PENG Z, HUANG W, GU S, et al. Conformer: Local features coupling global representations for visual recognition[C]. IEEE International Conference on Computer Vision(ICCV), 2021: 367-376.
[21] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks [C]. IEEE international Conference on Computer Vision (CVPR), 2017: 764-773.
[22] REDMON J, FARHADI A. Yolov3: An incremental improvement [J]. ArXiv Preprint, 2018, ArXiv:1804.02767.
[23] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in Neural Information Processing Systems(NIPS), 2015, 28.
[24] ZHANG S, CHI C, YAO Y, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection [C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2020: 9759-9768.
[25] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context [C]. European Conference on Computer Vision (ECCV), 2014: 740-755.
[26] LIU X, ZHAO H, TIAN M, et al. Hydraplus-net: Attentive deep features for pedestrian analysis[C]. IEEE International Conference on Computer Vision (ICCV), 2017: 350-359.
[27] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2016: 770-778.
[28] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI), 2020, 42(2), 318-327.
[29] TIAN Z, SHEN C, CHEN H, et al. FCOS: Fully convolutional one-stage object detection [C]. IEEE International Conference on Computer Vision(ICCV), 2019: 9626-9635.
[30] KIM K, LEE H S. Probabilistic anchor assignment with iou prediction for object detection[C]. European Conference on Computer Vision (ECCV), 2020: 355-371.

作者简介

罗晓霞,教授,主要从事图像处理与识别、大数据技术方面研究。
E-mail:1536531696@qq.com

蒋磊,硕士研究生,主要从事图像处理与目标检测方面研究。
E-mail:20208088024@stu.xust.edu.cn

蔡院强(通信作者),博士,主要从事图像处理与识别、文字检测方面研究。
E-mail:caiyuanqiang@bupt.edu.cn