

基于 Ghost 卷积和 YOLOv5s 网络的服装检测^①



李 雪, 吴圣明, 马丽丽, 陈金广

(西安工程大学 计算机科学学院, 西安 710048)

通信作者: 陈金广, E-mail: xacjg@163.com

摘 要: 为了降低服装目标检测模型的参数量和浮点型计算量, 提出一种改进的轻量级服装目标检测模型——G-YOLOv5s. 首先使用 Ghost 卷积重构 YOLOv5s 的主干网络; 然后使用 DeepFashion2 数据集中的部分数据进行模型训练和验证; 最后将训练好的模型用于服装图像的目标检测. 实验结果表明, G-YOLOv5s 的 *mAP* 达到 71.7%, 模型体积为 9.09 MB, 浮点型计算量为 9.8 G FLOPs, 与改进前的 YOLOv5s 网络相比, 模型体积压缩了 34.8%, 计算量减少了 41.3%, 精度仅下降 1.3%, 方便部署在资源有限的设备中使用.

关键词: 服装图像; 目标检测; YOLOv5s; DeepFashion2; Ghost 卷积; 轻量级; 深度学习

引用格式: 李雪, 吴圣明, 马丽丽, 陈金广. 基于 Ghost 卷积和 YOLOv5s 网络的服装检测. 计算机系统应用, 2022, 31(7): 203–209. <http://www.c-s-a.org.cn/1003-3254/8621.html>

Clothes Detection Using Ghost Convolution and YOLOv5s Network

LI Xue, WU Sheng-Ming, MA Li-Li, CHEN Jin-Guang

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract: To reduce the number of parameters and floating points operations of the object detection model for clothes, we propose an improved object detection model for lightweight clothes, namely G-YOLOv5s. First, the Ghost convolution is used to reconstruct the backbone network of YOLOv5s, and then the data in the DeepFashion2 dataset is employed for model training and validation. Finally, the trained model is applied to the detection of clothes images. The experimental results show that the G-YOLOv5s algorithm achieves the mean average precision (*mAP*) of 71.7%, with a model volume of 9.09 MB and the floating point operations of 9.8 G FLOPs. Compared with those of YOLOv5s, the model volume of G-YOLOv5s is compressed by 34.8%, and the floating point operations are reduced by 41.3%, with an *mAP* drop of only 1.3%. Moreover, it is convenient for deployment in equipment with limited resources.

Key words: clothes image; object detection; YOLOv5s; DeepFashion2; Ghost convolution; lightweight; deep learning

根据国家统计局的统计数据显示^[1], 2020 年我国电子商务平台交易额达到 37.2 万亿元, 按同比口径计算, 比去年增长 4.5%. 消费需求不断释放, 新消费模式拉动网络消费快速增长. 服装商品交易是电商平台交易的重要组成部分, 随着线上消费快速发展, 电商平台的服装图像数据呈指数增长, 同时这些服装图像种类繁多、样式复杂, 如何通过目标检测技术准确判断图像中每种服装的类别, 并定位出服装的具体位置, 为顾

客检索、推荐出相似的商品, 提升购买体验, 成为当前目标检测技术在服装领域的研究热点之一^[2].

传统的机器学习方法如 HOG、SIFT 等, 依赖人工设计的方法提取目标特征, 鲁棒性较差; 基于深度学习的目标检测技术能够在较大规模的数据集上自动提取图像特征, 减少人工干预, 同时提高检测准确率, 因此受到广泛的关注. 基于候选框的两阶段目标检测算法, 如 Faster-RCNN^[3]、Mask-RCNN^[4] 等, 检测精度高但

① 基金项目: 陕西省教育厅科研计划 (21JP049); 西安工程大学研究生创新基金 (chx2021026); 大学生创新创业训练计划 (S202110709112)

收稿时间: 2021-10-17; 修改时间: 2021-11-17; 采用时间: 2021-12-20; csa 在线出版时间: 2022-05-30

检测效率较低;基于回归的一阶段的目标检测算法,如SSD^[5]、YOLO^[6,7]等,检测速度快,检测精度较高,具有很强的实用性。但是以上模型训练对计算资源要求较高,训练所产生的权重文件通常高达几百兆,难以部署在资源有限的设备中使用。

为了降低模型的参数量和计算量,使得模型能够部署在资源有限的设备中,本文结合YOLOv5s网络构建一个轻量级的服装目标检测模型——G-YOLOv5s,该模型使用Ghost卷积^[8]重构YOLOv5s的主干网络,Ghost卷积首先使用传统卷积生成少量的原始特征图,然后再利用原始特征图经过线性操作生成多个Ghost

特征图,减少了生成相似特征的卷积操作,使得模型精度在基本无损失的情况下,有效降低服装检测模型对计算资源的占用。

1 YOLOv5s 网络

2020年Ultralytics公司提出YOLOv5系列算法,分别为YOLOv5s、YOLOv5m、YOLOv5l以及YOLOv5x,网络结构的深度、宽度以及所需的计算资源依次递增。YOLOv5s整体网络架构如图1所示,从图中可以看出,该网络结构大致可分为输入端、主干网络、Neck网络以及预测层4个部分,下面逐一进行介绍。

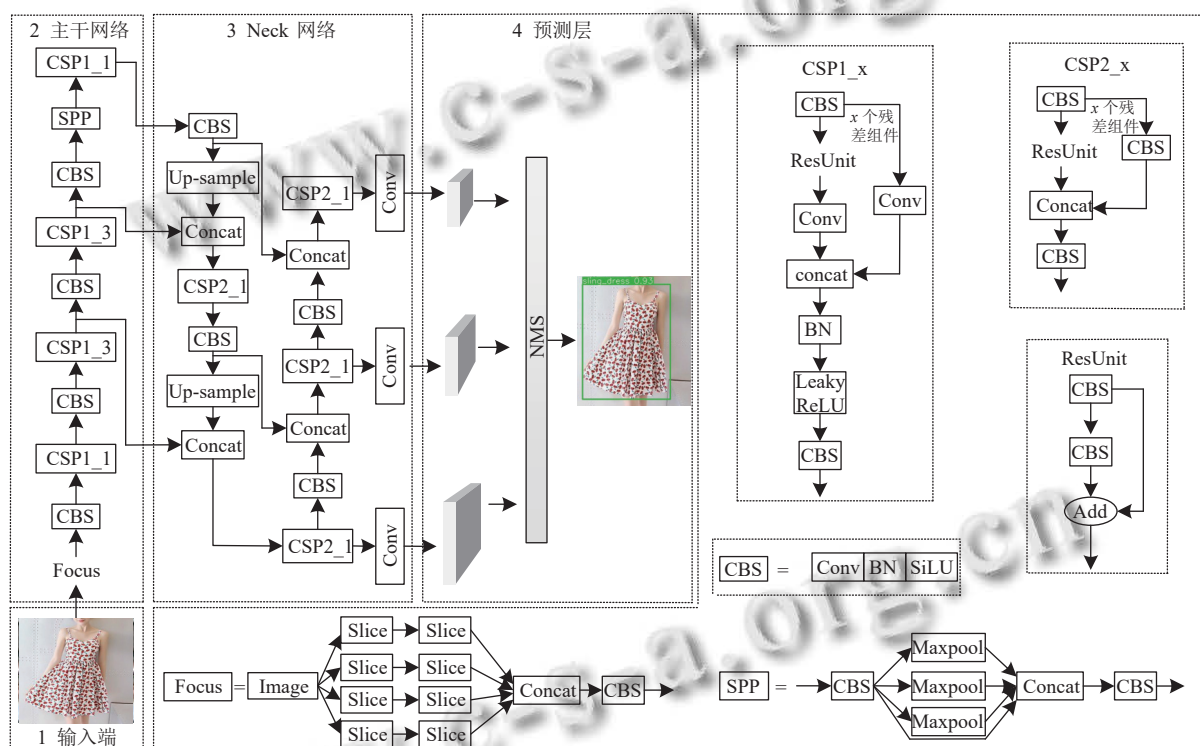


图1 YOLOv5s 网络结构

第1部分是输入端,使用了Mosaic图像增强、自适应图片缩放以及自适应锚框计算这3种图像处理方法。Mosaic图像增强方法通过随机缩放、剪裁和排布等操作将4张图片拼接成一张图像,然后输入网络进行训练,这种图像增强方式可增加数据的多样性以及训练目标个数。自适应锚框计算方法可根据不同数据集的特点重新计算初始锚框大小,使得初始锚框更加适合不同的数据集。在模型推理阶段使用自适应图片缩放方法为待检测图像填充最小的灰度值,极大减少了图像的冗余信息,提高了模型的推理速度。

第2部分是主干网络,引入空间金字塔池化模块(spatial pyramid pooling, SPP)^[9],实现局部特征和全局特征的信息融合;根据跨阶段局部连接网络(cross stage partial network, CSPNet)^[10]的设计思想,YOLOv5s构造了两种CSP结构,如图1所示,分别为CSP1_x和CSP2_x,其中CSP1_x添加在主干网络中,CSP2_x则添加在Neck网络部分,两种CSP结构都使用了残差结构,能够在尽量不增加计算复杂度的情况下提高网络的特征提取能力。

第3部分是Neck网络,受特征金字塔网络(feature

pyramid networks, FPN)^[11] 和路径聚合网络 (path aggregation network, PANet)^[12] 的启发, 将 Neck 网络构造造成 FPN+PAN 结构. FPN 结构自顶向下, 其后再添加一个自底向上的金字塔, 该金字塔由两个 PAN 结构组成. Neck 网络能够同时融合浅层和深层特征信息, 有效提升检测器的性能.

第4部分是预测层. 使用 CIOU loss (complete intersection over union loss)^[13] 作为边界框的损失函数, 采用标准的非极大值抑制操作 (non-maximum suppression, NMS) 滤除多余的预测框, 得到最终的模型预测结果.

2 主干网络的改进

2.1 模型的参数量和计算量分析

传统卷积和 Ghost 卷积操作如图2所示. Ghost 卷积将传统卷积分成两个步骤执行, 第1步使用少量传统卷积生成 m 个原始特征图; 第2步利用 m 个原始特征图经过线性运算再生成 s 个 Ghost 特征图, 经过上述两步操作, Ghost 卷积最终输出 $n = m \cdot s$ 个特征图. 同样在输出 n 个特征图的情况下, 分别使用传统卷积和 Ghost 卷积网络参数量分别为 p_1 、 p_2 .

$$p_1 = n \cdot c \cdot k \cdot k \quad (1)$$

$$p_2 = \frac{n}{s} \cdot c \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot d \cdot d \quad (2)$$

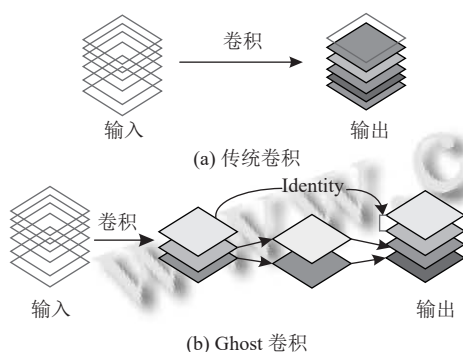


图2 传统卷积和 Ghost 卷积

两者参数量之比如式(3)所示:

$$\frac{p_1}{p_2} = \frac{n \cdot c \cdot k \cdot k}{\frac{n}{s} \cdot c \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot d \cdot d} \approx \frac{s \cdot c}{s+c-1} \approx s \quad (3)$$

使用传统卷积和 Ghost 卷积的模型浮点型计算量分别为 q_1 、 q_2 .

$$q_1 = n \cdot h' \cdot w' \cdot c \cdot k \cdot k \quad (4)$$

$$q_2 = \frac{n}{s} \cdot h' \cdot w' \cdot c \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot h' \cdot w' \cdot d \cdot d \quad (5)$$

两者浮点型计算量之比如式(6)所示:

$$\begin{aligned} \frac{q_1}{q_2} &= \frac{n \cdot h' \cdot w' \cdot c \cdot k \cdot k}{\frac{n}{s} \cdot h' \cdot w' \cdot c \cdot k \cdot k + (s-1) \cdot \frac{n}{s} \cdot h' \cdot w' \cdot d \cdot d} \\ &\approx \frac{s \cdot c}{s+c-1} \approx s \end{aligned} \quad (6)$$

其中, c 表示输入图像的通道数, $k \cdot k$ 表示传统卷积操作的卷积核大小, h' 、 w' 分别表示 Ghost 卷积生成的原始特征图的高和宽, $d \cdot d$ 为线性操作的卷积核大小, 且 $s \ll c$. 由式(3)、式(6)可以看出, 当 k 与 d 大小相等时, 使用 Ghost 卷积进行特征提取所占用的参数量和计算量约为传统卷积的 $1/s$.

利用 Ghost 卷积构成 Ghost 瓶颈结构 (G-bneck), 如图3所示, G-bneck 类似于 ResNet 结构^[14]. 在图3中, 第1个 Ghost 卷积作为扩展层, 用于增加特征通道数, 第2个 Ghost 卷积则用于减少通道数, 再经过 shortcut 连接后输出.

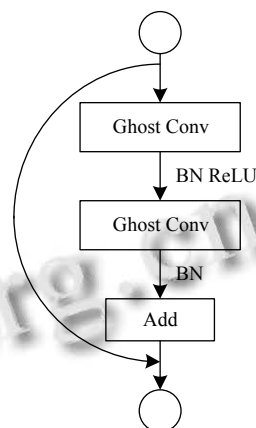


图3 Ghost 瓶颈结构

2.2 主干网络的重构

通过上文对 Ghost 卷积计算的分析, 若将该卷积应用在 YOLOv5s 中可进一步降低模型的参数量和复杂度, 所以本文以 YOLOv5s 为基本模型, 在其主干网络中, 用 Ghost 卷积替换传统卷积 (图1中 CBS 结构), 用 G-bneck 替换 CSP1_x 结构, Neck 网络层和预测层保持原结构不变. 基于 G-YOLOv5s 网络的服装检测流程如图4所示.

图4中, 步骤①—③表示模型的训练阶段, 将服装图像数据送入 G-YOLOv5s 网络中进行训练, 并将结果最优的模型权重保存下来; 步骤④—⑦是模型的验证阶

段,使用验证集数据来评估模型的好坏,若模型训练结果较差,则尝试调整网络参数后重新进行训练;步骤⑧–⑩是模型检测阶段,使用评估结果良好的模型进行服装图片检测,并输出检测后的结果。

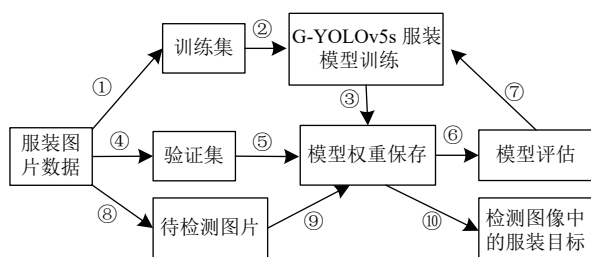


图4 基于 G-YOLOv5s 网络的服装检测流程

3 模型训练及验证

3.1 数据集及评价指标

DeepFashion2^[15]数据集总共包含约30万张图片,13种服装类别,类别名称分别为:短袖衫(short sleeve shirt)、长袖衫(long sleeve shirt)、短袖外衫(short sleeve outwear)、长袖外套(long sleeve outwear)、背心(vest)、吊带(sling)、短裤(shorts)、长裤(trousers)、半身裙(skirt)、短袖连衣裙(short sleeve dress)、长袖连衣裙(long sleeve dress)、无袖连衣裙(vest dress)、吊带裙(sling dress)。

在数据准备过程中发现,DeepFashion2数据集存在严重的数据分布不平衡问题,例如在训练集中,包含“短袖衫”这一类别的图片高达数万多张,而包含“短袖外衫”这一类别的图片仅有几百张。服装类别不同,数据量差异较大,为了减轻对模型训练的影响,在实验前将DeepFashion2数据集进行如下处理:

首先将训练集图片按照类别进行分类;然后在分好的每类的图片数据中按照如下规则进行随机抽取:对于数据量较大的服装类别,采用较小的比例进行数据抽取;对于数据量适中的类别,增大随机抽取的比例,此外,由于“短袖外衫”图片数量过少,因此保留全部数据。数据抽取完成之后再次混合;最后将混合好的训练集图像对应的标签文件转成网络能够识别的YOLO格式,得到处理完成的训练集数据。验证集同样进行如上处理。最终得到带有标签文件的训练集图片14 746张、验证集图片7 763张。处理好的训练集中,服装各类别目标数量分布如图5所示。

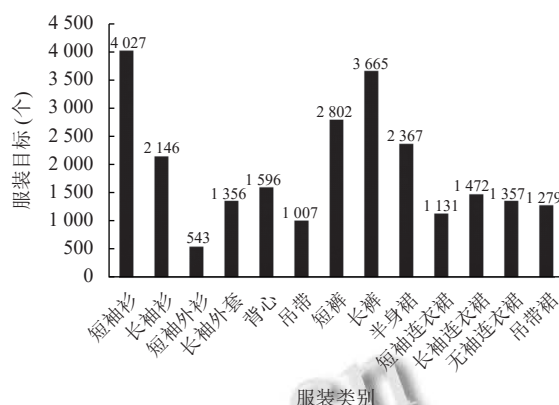


图5 训练集中13种类别分布

采用的模型评价指标有平均精度(average precision, AP)和平均精度均值(mean average precision, mAP)。 AP 表示模型对某个类的检测精度,值越大,表示模型对某个类的检测精度越高; mAP 表示模型对数据集所有类别的平均检测精度,值越大则模型对数据集整体类别的检测效果越好,评价指标计算公式如式(7)–式(10)所示。

$$AP = \int_0^1 P(R) dR \quad (7)$$

$$mAP = \frac{1}{N} \sum_{i=0}^{N-1} AP_i \quad (8)$$

其中, N 为数据集中所有的类别数,在本文中 N 等于13, $P(R)$ 表示分别以精确率(precision, P)和召回率(recall, R)为横纵坐标所构成的函数。 P 和 R 的计算公式如下:

$$P = \frac{TP}{TP + FP} \quad (9)$$

$$R = \frac{TP}{TP + FN} \quad (10)$$

其中, TP 表示将正例预测为正例的个数, FP 表示将反例预测为正例的个数, FN 表示将反例预测为反例的个数。

3.2 G-YOLOv5s 网络训练

模型损失函数由边界框回归损失(bounding box regression score)、置信度损失(objectness score)以及分类概率损失(class probability score)3部分组成,G-YOLOv5s使用二值交叉熵损失函数计算类别概率损失和置信度损失。采用CIoU loss计算边界框损失,因为该损失计算方式同时考虑了预测框与真实框的重叠

面积、两者中心点间距离以及框的长宽比等因素,能够加快模型的收敛速度,计算公式如式(11)–式(13)。

$$L_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{\text{gt}})}{c^2} + \alpha \nu \quad (11)$$

$$\nu = \frac{4}{\pi^2} \left(\arctan \frac{w^{\text{gt}}}{h^{\text{gt}}} - \arctan \frac{w}{h} \right)^2 \quad (12)$$

$$\alpha = \frac{\nu}{(1 - \text{IoU}) + \nu} \quad (13)$$

其中, $\rho(\mathbf{b}, \mathbf{b}^{\text{gt}})$ 表示预测框和真实框两个中心点间的欧式距离, c 表示两框之间的最小外接矩形的对角线长度, $\frac{w^{\text{gt}}}{h^{\text{gt}}}$ 和 $\frac{w}{h}$ 分别表示真实框和预测框各自的宽高比, IoU

是两框之间的交并比。

实验的软硬件平台以及参数设置如下: Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz; GPU: TITAN XP; 加速环境: CUDA 9.2; 操作系统: Ubuntu 16.04; 深度学习框架: PyTorch 1.7.1; 语言环境: Python 3.8. batch_size 为 16; 初始学习率为 0.01; 权重衰减系数为 0.000 5; 使用随机梯度下降法 (SGD) 进行优化; 动量等于 0.937; 总训练轮数为 300 轮。

根据上述参数设置, 分别对改进前和改进后的 YOLOv5s 和 YOLOv5l 网络进行了对比实验。4 种网络的训练损失 (loss) 以及 mAP 变化曲线如图 6 所示。

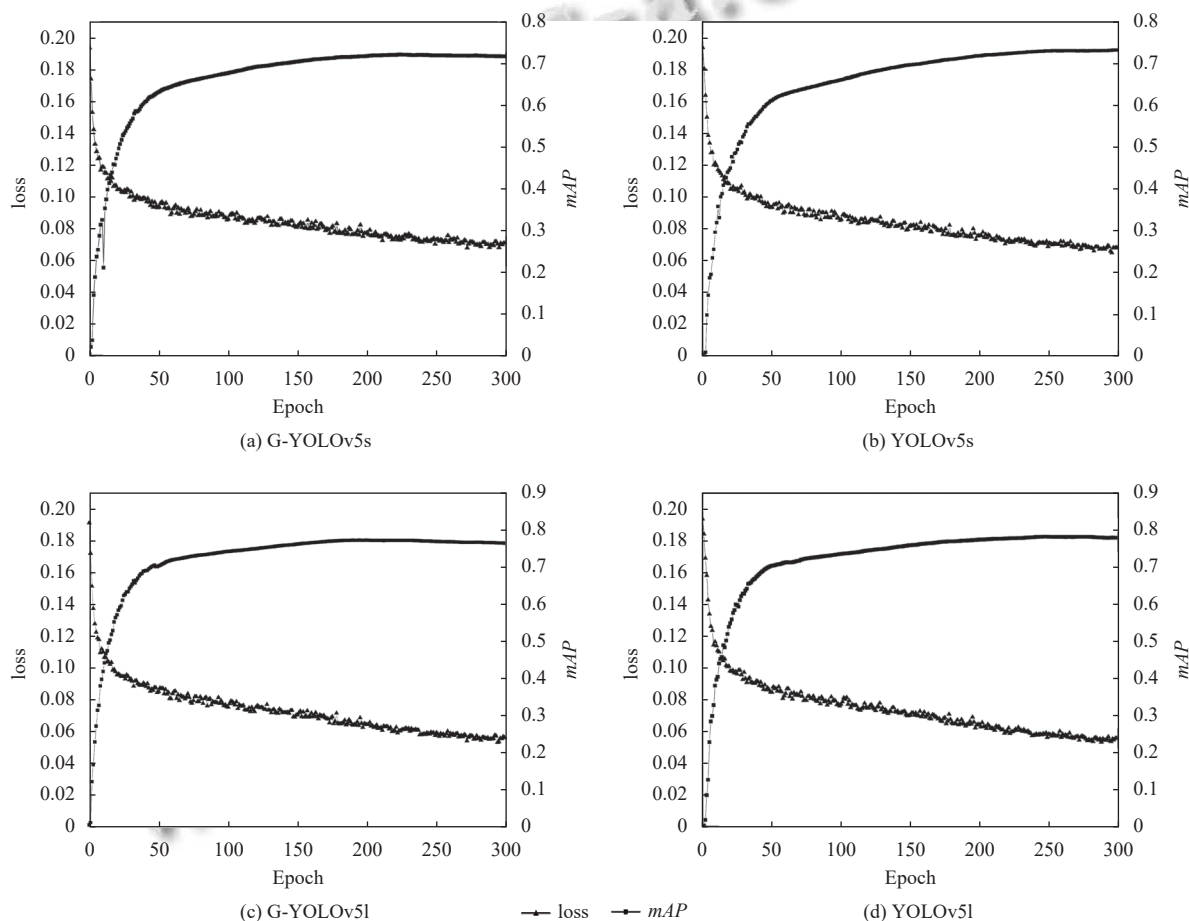


图6 网络 loss 和 mAP 随训练轮数的变化图

由图 6 可以看出, 无论是 YOLOv5s 还是 YOLOv5l, 改进前后, 网络的 loss 和 mAP 曲线变化不大, 说明引入 Ghost 卷积后, 原网络性能基本不受影响。另外, 改进前后 YOLOv5l 的 mAP 高于 YOLOv5s, 这是因为改进前后 YOLOv5l 的网络结构始终大于 YOLOv5s, 提

取的特征信息也更加丰富, mAP 会有所提升, 但是相应的, 网络参数量、浮点型计算量以及运行时间也会高于 YOLOv5s。

3.3 服装检测结果与分析

使用 G-YOLOv5s 模型对 13 种服装种类进行检

测,得到的平均精度值 (AP) 如表 1 所示。

由表 1 中可以看出, G-YOLOv5s 模型对“短袖衫”“短裤”以及“长裤”这 3 类服装的 AP 值均达到 87% 以上,检测效果较好;但对于“短袖外衫”的检测, AP 仅为 46.9%,分析原因可能是:该类别服装图片数量较少(如图 5 所示),导致 G-YOLOv5s 对“短袖外衫”这一类别的训练不足,因此检测结果略差。

表 1 每类服装的平均精度 (%)

类别	AP	类别	AP
短袖衫	89.7	长裤	92.5
长袖衫	72.2	半身裙	78.5
短袖外衫	46.9	短袖连衣裙	58.2
长袖外衫	81.1	长袖连衣裙	61.2
背心	79.8	无袖连衣裙	65.7
吊带	52.0	吊带裙	67.0
短裤	87.8		

表 2 中对比了 4 种模型的 mAP 、模型体积、浮点运算量以及在 CPU 上的推理时间,所有实验的输入图像尺寸均为 640×640 像素大小,结果如表 2 所示。

表 2 算法实验结果对比

模型	mAP (%)	模型体积 (MB)	浮点型 计算量 (G)	速度 (CPU/ms)
G-YOLOv5s	71.7	9.09	9.8	54.9
YOLOv5s	73.0	13.94	16.7	71.3
G-YOLOv5l	77.7	55.31	61.2	246.9
YOLOv5l	78.2	90.17	115.2	298.7

由表 2 可以看出, G-YOLOv5s 与 YOLOv5s 相比,虽然 mAP 有略微下降,但模型体积压缩了 34.8%,浮点运算量减少了 41.3%;G-YOLOv5l 与 YOLOv5l 相比, mAP 下降了 0.5%,但模型体积和浮点型运算量也下降了近一半左右。实验结果证明,使用 Ghost 卷积能够有效减少模型参数量以及浮点型运算量,从而降低对资源的占用。并且与其它 3 种网络相比, G-YOLOv5s 模型最为轻量、在 CPU 设备上检测一张图片用时最短,因此更适合部署在资源有限的设备上使用。

使用 G-YOLOv5s 模型进行服装检测,效果如图 7 所示,其中图 7(a) 和图 7(b) 分别为卖家秀和买家秀图片的检测效果。

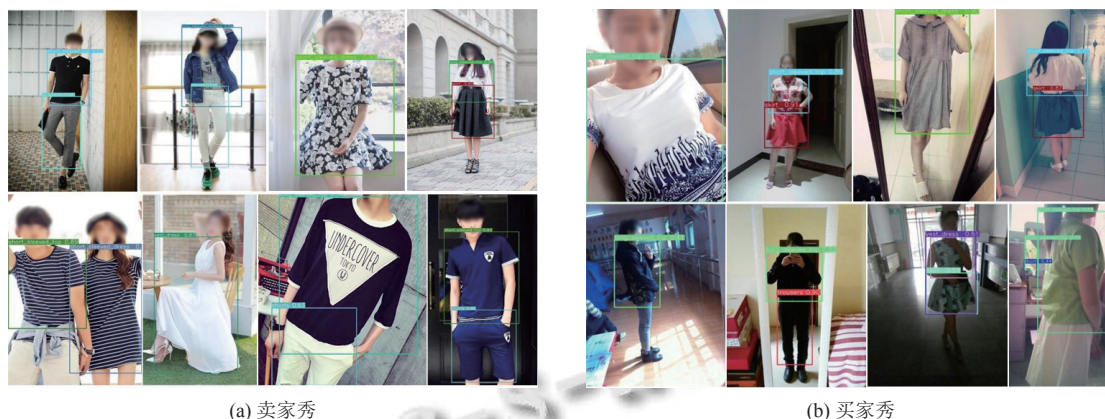


图 7 买家秀和卖家秀检测效果

对比观察可知,图 7(a) 的检测效果普遍优于图 7(b),这是因为图 7(a) 中图片背景较为简单且检测目标比较突出, G-YOLOv5s 模型对此表现较好;而在图 7(b) 中,部分图片由于光线不足、拍摄环境复杂等问题,模型检测能力有所降低,出现漏检、误检的情况(图 7(b) 中第 2 行检测图片所示)。

4 结论

为了降低服装检测模型参数量和浮点型运算量较大的问题,提出一种使用 Ghost 卷积构建 YOLOv5s 主干网络的轻量级服装检测方法,使用处理过的数据集

对 G-YOLOv5s 网络进行训练和验证。实验结果表明, G-YOLOv5s 算法对遮挡范围较小、背景不是特别复杂的服装目标检测准确率高,并且该模型权重较小、浮点型计算量较低,是一种可行的、有效的服装检测方法,可在资源有限的设备中使用,下一步将着重关注复杂背景下服装检测方法的研究。

参考文献

- 1 国家统计局. 国家统计局统计科学研究所所长阎海琪解读 2020 年我国经济发展新动能指数. http://www.stats.gov.cn/tjsj/sjjd/202107/t20210726_1819836.html. (2021-07-26).

- 2 林碧琚, 耿增民, 洪颖, 等. 卷积神经网络在纺织及服装图像领域的应用. 北京服装学院学报 (自然科学版), 2021, 41(1): 92–99, 108.
- 3 Ren SQ, He KM, Girshick R, *et al.* Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137–1149. [doi: [10.1109/TPAMI.2016.2577031](https://doi.org/10.1109/TPAMI.2016.2577031)]
- 4 He KM, Gkioxari G, Dollár P, *et al.* Mask R-CNN. *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, 2017. 2980–2988.
- 5 Liu W, Anguelov D, Erhan D, *et al.* SSD: Single shot multibox detector. *Proceedings of the 14th European Conference on Computer Vision*. Amsterdam: Springer, 2016. 21–37.
- 6 Redmon J, Farhadi A. YOLOv3: An incremental improvement. *arXiv: 1804.02767*, 2018.
- 7 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. *arXiv: 2004.10934*, 2020.
- 8 Han K, Wang YH, Tian Q, *et al.* GhostNet: More features from cheap operations. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle: IEEE, 2020. 1577–1586.
- 9 He KM, Zhang XY, Ren SQ, *et al.* Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904–1916. [doi: [10.1109/TPAMI.2015.2389824](https://doi.org/10.1109/TPAMI.2015.2389824)]
- 10 Wang CY, Liao HYM, Wu YH, *et al.* CSPNet: A new backbone that can enhance learning capability of CNN. *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle: IEEE, 2020. 1571–1580.
- 11 Lin TY, Dollár P, Girshick R, *et al.* Feature pyramid networks for object detection. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu: IEEE, 2017. 936–944.
- 12 Liu S, Qi L, Qin HF, *et al.* Path aggregation network for instance segmentation. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018. 8759–8768.
- 13 Zheng ZH, Wang P, Liu W, *et al.* Distance-IoU loss: Faster and better learning for bounding box regression. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York: AAAI, 2020. 12993–13000.
- 14 He KM, Zhang XY, Ren SQ, *et al.* Deep residual learning for image recognition. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE, 2016. 770–778.
- 15 Ge YY, Zhang RM, Wang XG, *et al.* DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach: IEEE, 2019. 5332–5340.

(校对责编: 孙君艳)