文章编号：2095-2163(2022)08-0008-06 中图分类号：TP391 文献标志码：A

基于改进 YOLOv3 的行人检测研究

车启谣，严运兵

（武汉科技大学 汽车与交通工程学院，武汉 430065）

摘 要：针对 YOLOv3 在检测行人时易漏检小目标以及遮挡问题，提出了一种改进的 YOLOv3 行人检测模型。改进模型采用 K-means++聚类算法取代原 K-mean 聚类算法，以减轻因初始聚类中心随机选择不当对结果所造成的误差影响；通过加入残差网络模块方法的轻量化模型，并在结构中加入 CBAM 注意力机制与 MHSA 多头自注意力机制；通过高效的分配计算资源与捕获全局信息，来提高算法的特征提取能力。实验表明，改进后的算法在 CUHK 数据集上取得了较好的效果，其中通过实验得到的 *mAP* 值为 88.20%，相对原算法提升了 17.45%，有着较好的特征提取能力，提升了检测小目标与被遮挡行人的能力，同时在检测精准度方面更优。

关键词：行人检测；YOLOv3；K-means++；残差网络模块；CBAM 注意力机制；MHSA 多头自注意力机制

Pedestrians detection based on improved YOLOv3

CHE Qiyao，YAN Yunbing

（College of Automobile and Traffic Engineering，Wuhan University of Science and Technology，Wuhan 430065，China）

【Abstract】 Aiming at the problem of YOLOv3′s easy to miss small targets and occlusion problem in pedestrian detection，an improved YOLOv3 pedestrian detection model is proposed. The improved model uses K-means ++ clustering algorithm to replace the original K-mean clustering algorithm，so as to reduce the error effect caused by improper random selection of initial clustering centers. Meanwhile，the residual network module is added into the model to lightweight the model，and CBAM attention mechanism and MHSA multi-head self-attention mechanism are added into the structure. The feature extraction ability of the algorithm is improved by efficiently allocating computing resources and capturing global information. The experiment shows that the improved algorithm has achieved good results on CUHK dataset，in which the *mAP* value obtained through the experiment is 88.20%，17.45% higher than the original algorithm. It has better feature extraction ability，improves the ability of detecting small targets and blocked pedestrians，and is better in accuracy of detection.

【Key words】 pedestrians detection；YOLOv3；K-means++；residual network module；CBAM attention mechanism；MHSA multiplex self-attention mechanism

0 引 言

近年来，行人检测技术得到快速发展并取得了一定成果，在智能人脸安检、辅助驾驶系统，以及智能网联等领域占据了重要地位。在自动驾驶领域中，行人检测技术主要采用安装在物体上的视觉传感器，对采集到的感兴趣区域进行分析处理，以完成对行人的识别。现有的行人检测技术主要通过提取人体的几何特征[1-4]与运动信息特征[5-7]来设计特征提取方法，虽在检测速度与精准度上都有所提升，但仍有一些问题未得到很好的解决。如：检测结果易受到光照变化和遮挡物的影响，降低了行人检测效果，且行人自身存在较大的形变，加大了有效特征提取的难度。此外，还存在目标检测算法复杂、对于多目标情况下检测时间较长、实时性较差、鲁棒性也难以达到要求等问题。

目前，针对行人检测问题所采用的算法可分为 2 种：一是采用人工选取目标点特征，再通过机器学习训练分类器的检测算法；二是采用深度学习训练网络模型的检测算法。

采用机器学习训练分类器的算法提出较早，并且一直处于不断的优化中。2001 年，Viola 等人[8]通过提取 Haar 特征，再对 AdaBoost 级联分类器进行训练的方法实现目标人脸检测。2005 年，Dalal 等人[9]通过提取 HOG 特征，对行人的边缘特征进行描述，并发现 HOG 描述子十分适合人的检测，但是描述子生成

基金项目：国家自然科学基金（51975428）。

作者简介：车启谣（1997-），男，硕士研究生，主要研究方向：无人驾驶视觉感知；严运兵（1968-），男，博士，教授，主要研究方向：汽车动力学及其控制、新能源汽车驱动与控制、智能驾驶人机交互与决策。

通讯作者：严运兵 Email：835209056@ .qq.com

收稿日期：2022-02-19

过程较长，实时性差，并且难以处理行人被遮挡的问题。2015 年，谭飞刚等人[10] 提出一种结合二值化 Haar 特征多部件验证的双层行人检测算法，提高了行人部分被遮挡时的检测精准度。虽然传统的机器学习算法也在持续优化与更新，但仍难以满足相应的要求。

目前，基于深度学习的检测算法在描述目标特征图时的精准检测与输出，已逐渐占据了主导地位。2014 年，Girshick 等人[11] 提出的 R-CNN 目标检测框架，主要通过选择性地搜索可能包含检测目标的候选区域，并对每个候选框进行分类，再利用 CNN 提取特征，现已成为最典型的双阶段目标检测算法。随后，针对双阶段目标检测算法空间模型规模大、测试速度慢等缺点，2015 年，Girshick 等人[12] 提出基于边界框和多任务损失分类的 Fast R-CNN 算法。2017 年，Ren 等人[13] 通过引入区域建议网络，提出了 Faster R-CNN 算法。

当前情况下，双阶段目标检测算法虽然在不断地进行优化，但是仍然难以满足目标检测算法适用场景的实时性与鲁棒性。相比较之下，基于回归分析思想的单阶段目标检测算法具有检测速度快、精度高的特点。2015 年，Redmon 等人提出了 YOLO 检测算法[14]，将特征提取、回归和分类置于单个卷积网络中，通过简化网络结构的方法，实现端到端的目标检测，但该算法对小尺度目标的检测精准度与召回率低。针对此问题，2016 年，Liu 等人[15] 采用分层提取特征的思想与目标预测机制提出了 SSD 算法。2017 年，Jeong 等人[16] 基于 SSD 算法，通过增减反卷积模块提出 R-SSD 算法。Li 等人[17] 通过融合多种特征层与特征尺度并生成特征金字塔的方法，提出了 F-SSD 算法。但是，以上算法仍然会存在对小目标检测效果欠佳、检测速度较慢的缺陷。

尽管传统的目标检测方法基本能够满足物体本身的检测要求，但效率和精准度方面却仍有不足亟待完善。为此，本文对传统 YOLOv3 检测算法提出了一些改进，并在 CUHK 数据集上对改进后算法的准确率和召回率进行了仿真验证。

1　YOLOv3 神经网络算法

自 2015 年以来，学界就陆续推出了更新换代的数个 YOLO 系列版本。YOLO 系列算法相对于 Fast R-CNN 算法来说，未将检测结果分 2 部分进行求解，而是基于回归的思想，在输出回归层直接回归出目标位置及其类别，有着更好的检测精度与检测速度。

2018 年，Joseph 等人提出 YOLOv3 算法，相对于前身对多个部分融入了改进内容。主要借用 ResNet 残差网络的思想，采用更好的基础特征提取网络 Darknet-53，和之前的网络结构相比，在一定程度上提升了检测速度，网络性能对比见表 1[18]；同时，采用多尺度融合预测的方法，共提取 3 个特征层，提升了算法对小目标的检测精度。至此，为保证每个目标的预测准确率与多目标标签分类，采用新的代价函数 *sigmoid* 替换原函数 *Softmax*。

表 1　Darknet-53 网络性能对比表

Tab. 1　Comparison table of Darknet-53 network performance

主干网络	Top-1 准确率/%	Top-5 准确率/%	浮点运算次数/(次 · s^{-1})	图像检测速度/FPS
Darknet-19	74.1	91.8	1 246	171
Darknet-53	77.2	93.8	2 457	78
ResNet-101	77.1	93.7	1 039	53
ResNet-152	77.6	93.8	1 090	37

YOLOv3 检测模型主要由骨干网络和检测网络两部分组成，其网络结构如图 1 所示[19]。图 1，采用基于残差网络思想的 Darknet-53 作为用于特征提取的主干网络。Darknet-53 模型包含 53 个卷积层和 23 个跳跃连接，具有相对 YOLOv2 模型更深的卷积层。

检测网络部分采用 Faster R-CNN 中使用的 FPN 特征金字塔结构[20]，尽可能地减少特征损失，提高检测精度。其中，共提取 3 个特征层，分别为：输出特征分辨率为 52×52 的中间层、26×26 的中下层和 13×13 的底层，3 个特征层分别针对小、中、大三种分辨率的目标对象通过检测。在获得 3 个有效特征层后通过多特征融合，并对有效特征层进行预测，得到预测结果后，再利用解码预测模块，对网络处理后的数据进行解码，由此得到最终结果。

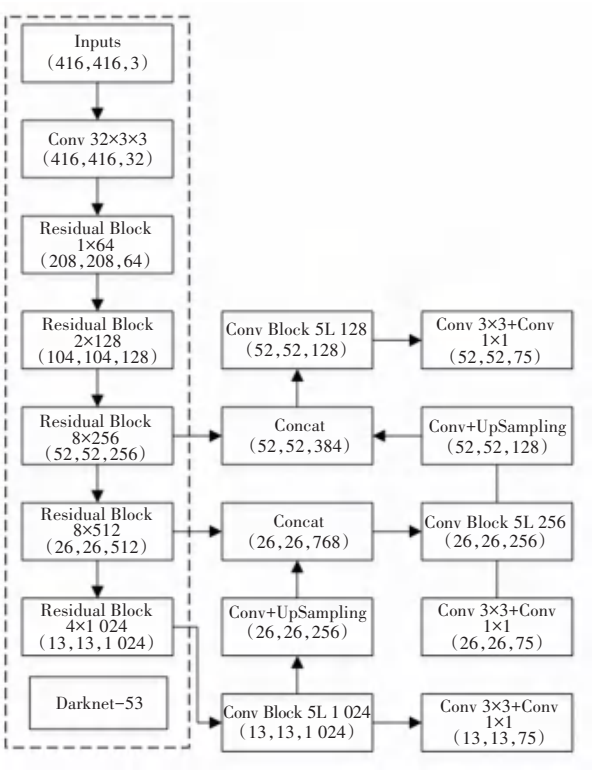

图 1 YOLOv3 模型结构

Fig. 1 YOLOv3 model structure

2 改进的 YOLOv3 神经网络算法

2.1 选用 K-means++聚类算法

YOLO 系列算法从 YOLOv3 开始,采用 9 个 anchor 进行预测,但仍然采用与 YOLOv2 相同的 K-means 聚类算法来获取 anchors 的大小。K-means 聚类算法会随机指定 K 个聚类中心(cluster)作为初始点,并将距离相近的 cluster 不断进行均值化处理,当 cluster 很小时,保存聚类来确定 anchor 的初始位置。距离相近的依据以 IOU 值进行判定,具体公式如下:

$$d(box,centroid) = 1 - IOU(box,centroid) \quad (1)$$

其中, d 为样本点到每一个 *cluster* 质心的距离; *box* 为其它边框; *centroid* 为聚类时被选作中心的边框; IOU 为目标预测框与目标标签框的交并比。

由于 K-means 聚类算法中聚类中心的随机性与离群点及孤立点的敏感性,算法的聚类效果易受到因初始值选取不当而造成的影响。此外,也会导致算法在分类时的不精准现象,出现错误分类的情况。为此,本文使用 K-means++算法取代原聚类算法,以期得到更符合样本的先验证框。

K-means++与 K-means 算法不同的是:第一步会随机选取一个 cluster 作为初始点,同时为了避免噪声,采取轮盘法选择一个新的距离较远的点,直至 K 个 cluster 被选出;此后再进行 K-means 聚类算法。尽管在初始点的选择上,K-means++算法花费了较多的时间,但实际上却减轻了初始聚类中心选择不当所造成的误差,提升了算法的计算效率。

2.2 残差网络模块

卷积神经网络深度的加深,可以提取更为丰富的特征,提升检测性能。但随着网络层数的增加,也会加重深层网络的训练负担,造成网络的性能退化等一系列问题。为了缓解网络加深后所带来的种种问题,YOLOv3 算法采用了与文献[21]中提出的 ResNet 相类似的残差网络结构。

残差网络结构通过快捷连接的方式,将每若干层中的某一层数据直接添加到后面数据层的输出部分,将中间的某一层或多层卷积层进行缩减,用来减少计算量,并降低网络深度。

YOLOv2 采用大量的 3×3 的卷积核进行卷积,而 YOLOv3 则先采用一个大小为 3×3、步长为 2 的卷积核进行卷积,将输入特征层的高和宽进行压缩,从而产生一个新的卷积层 Layer,稍后再保存 Layer 并进行一次 3×3 的卷积和 1×1 的卷积,得到的结果与 Layer 相加,便构成了残差网络结构。YOLOv3 残差网络模块结构如图 2 所示。

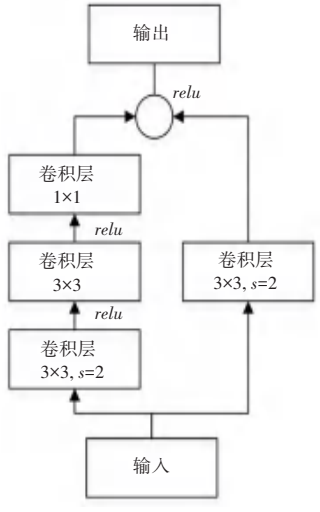

图 2 残差网络模块结构图

Fig. 2 Residual network module structure diagram

为了解决由于网络深度增大而产生的准确率下降和性能退化问题,本文在网络进行最末端的卷积处理后,加入残差网络模块来降低模型的计算量。该模块采用 3 层卷积层的网络结构,并使用 1×1 的卷积代替 3×3 卷积。改进后的 ResNet-H 结构如图 3 所示。

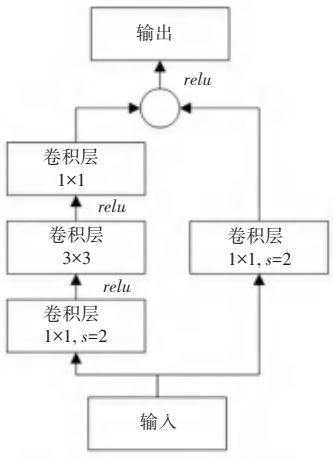

图 3　ResNet-H 模块结构图

Fig. 3　ResNet-H module structure diagram

2.3　CBAM 注意力机制

通常，人眼在接触到某一场景或客观事物时，关注点的不均匀分布会导致人的注意力朝向会转移至感兴趣的区域或者信息。通过这种选择性的视觉注意力机制，可以高效地分配注意力资源，并最终服务于人的主观意志。基于此，为了使计算机视觉在识别信息时自主学会留意关键有用的信息，研究人员通过计算概率分布的形式来展示词之间的关系，从而产生了注意力机制。

与其它机制相比，CBAM 注意力机制采用了通道注意力与空间注意力相结合的方法，通过 2 个维度依次在输入特征图中推断出特征权重，再将该权重与输入特征图进行点积，从而得到优化后的输出特征图。整体流程如图 4 所示。

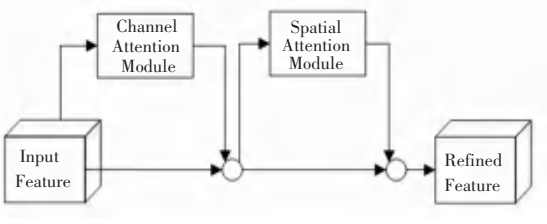

图 4　CBAM 模块的结构

Fig. 4　Structure of CBAM module

YOLOv3 检测算法的本质是将输入的图像进行编码，而后再从中解码出目标位置和类别信息进行输出[22]。在此过程中，通过在网络中加入 CBAM 注意力机制，可以使 YOLOv3 网络对行人施加较大的权重，从而提升特征提取能力。

由图 1 可知，YOLOv3 最初会提取 3 个基础特征层，每次基础特征层与其他上采样的特征层堆叠拼接后，会进行 5 次的卷积处理，此时在卷积处理中加入注意力模块。网络结构中加入的具体部位如图 5 所示。

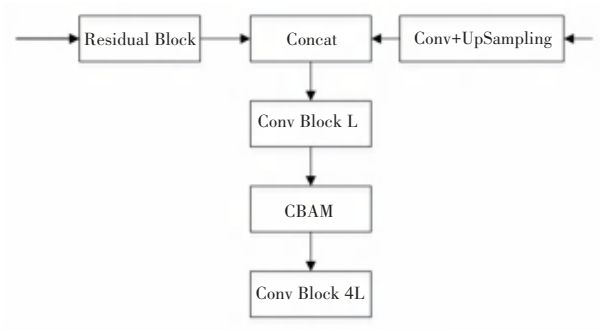

图 5　加入 CBAM 模块后的结构图

Fig. 5　Structure diagram after adding CBAM module

2.4　多头自注意力机制 MHSA

卷积神经网络 CNN 在识别特征时，通常会采用较小的卷积核来识别物体的局部特征，增加网络层数的同时，也减少了参数，但也会使得卷积层的感受野相对特征图要小上很多。尤其在行人检测中，往往需要在较大的特征图中获取行人特征，使得网络能够从目标的较大相邻区域中收集上下文信息，提取更好的行人特征。为了得到全局信息，需要拓展网络的深度，堆叠多个卷积层，为此需要消耗很多计算资源。

自注意力机制（Self-attention）在 2017 年由 Vaswani 等人[23]提出，并主要应用于学习文本表示[23]。在文本语言处理中，自注意力机制能够通过计算每个词的注意力概率来更好地捕获上下文信息，以表达词与词之间的语义关系。同时，文中还提出了多头自注意力机制（Multi-headed Self-attention），即通过多次 Self-attention 计算，将每个机制上的不同注意点权重矩阵结果进行拼接融合，即可表达出更加全面的关联程度。为此，考虑将其引入行人检测网络，由此来提取图片的全局特征。

多头自注意力机制结构如图 6 所示。该机制将 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 矩阵进行不同维度的矩阵映射，输出参数后进行储存，并将多次结果进行拼接融合后、再进行一次矩阵映射，就得到了输出结果。另外，由于使用过多的多头自注意力机制会导致计算机负荷增加，从而降低检测精度，故会在第四次的下采样后再将其加入进去。

2.5　改进后的 YOLOv3 模型结构

本文主要通过将 K-means 更改为 K-means++ 聚类算法、在网络结构中加入改进的残差网络模块、CBAM 注意力机制与 MHSA 多头自注意力机制的方法来对 YOLOv3 加以改进，改进后的 YOLOv3-i 模型结构如图 7 所示。

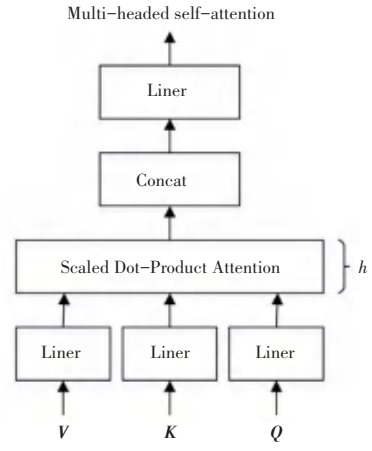

图 6 多头自注意力机制结构

Fig. 6 Multi-headed self-attention mechanism structure

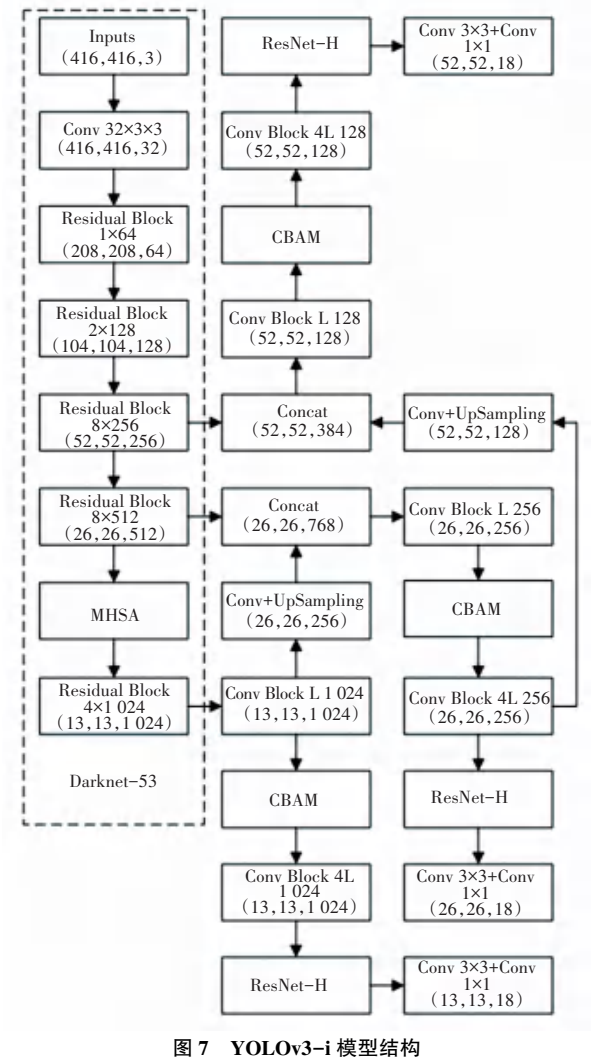

图 7 YOLOv3-i 模型结构

Fig. 7 YOLOv3-i model structure

3 验证与分析

3.1 实验环境设置

本文实验在 Windows 10 系统下进行。GPU 为 NVIDIA Tesla V100、内存 32 G、显存 32 GB；深度学习框架为 Pytorch 1.5.1 版本。

3.2 实验数据配置

实验所选取的数据集为香港中文大学(CUHK)开源行人检测数据集，共包括 1 063 张行人图片。实验开始前，先选用其中的 800 张图片进行调试训练，稍后另取各 100 张图片进行验证与测试。

训练采用的初始学习率为 0.001，学习率衰减策略为每经过一个 *epoch*，学习率降低为原来的 0.05 倍，迭代次数为 1 000 次，*Batch size* 为 45。即每个 *epoch* 在训练集中取 45 个样本进行训练，直至全部样本都遍历完成一次训练。

3.3 实验结果及分析

在相同的实验场景下，本文将基于 YOLOv3 改进的 YOLOv3-i 算法与 YOLOv3 检测算法共进行 5 组实验，旨在验证加入各个模块后，对行人检测方面的性能改进。实验明细概述如下。

(1)实验 *A*：原 YOLOv3 算法；

(2)实验 *B*：使用 K-means++改进后的 YOLOv3 算法；

(3)实验 *C*：使用 K-means++和残差网络模块改进后的 YOLOv3 算法；

(4)实验 *D*：使用 K-means++、残差网络模块、CBAM 注意力机制改进后的 YOLOv3 算法；

(5)实验 *E*：使用 K-means++、残差网络模块、CBAM 注意力机制和 MHSA 多头自注意力机制改进后的 YOLOv3 算法；

5 组实验的平均准确率 *mAP*、召回率 *Recall*、准确率 *Precision* 及调和平均值 F_1 *Score* 结果见表 2。

表 2 不同实验检测结果对比

Tab. 2 Comparison of detection results of different experiments %

名称	*mAP*	*Precision*	*Recall*	F_1 *Score*
实验 *A*	70.75	85.67	73.98	79.40
实验 *B*	76.55	87.59	77.05	82.35
实验 *C*	77.68	88.46	78.67	83.69
实验 *D*	79.35	89.97	81.44	85.50
实验 *E*	88.20	93.83	85.31	89.37

从表 2 中可以看出，本文提出的 YOLOv3-i 算法的 *mAP* 值达到 88.20%，而原算法的 *mAP* 值为 70.75%，相比而言提高了 17.45%；召回率、检测精准度等数值也均有提升。而从 *D*、*E* 两组实验可见，实验 *E* 的 *mAP* 相对提升了 8.85%，表明多头注意力机制相比于传统的卷积网络有更强的特征提取能力。

原 YOLOv3 与本文提出的 YOLOv3-i 改进算法的实际测试结果对比如图 8 所示。在小目标的检测上，经过图 8(a)与图 8(b)的对比，可以清晰地看出

改进的算法对行人目标的边缘轮廓进行了更好的刻画，提升了特征检测效果。而对于受到严重遮挡重叠的行人目标部分，图 8(d)相对图 8(c)而言仍能进行识别。由此可以表明：改进后的算法提升了对小目标以及被遮挡部位的特征提取能力，能够更加精准地发现行人检测目标。

(a) YOLOv3 检测结果　(b) YOLOv3-i 检测结果

(c) YOLOv3 检测结果　(d) YOLOv3-i 检测结果

图 8　不同实验检测效果对比图

Fig. 8　Comparison of detection effects of different experiments

4　结束语

本文以 YOLOv3 网络为基础，通过选用 K-means++聚类算法、在网络结构中加入改进的残差网络模块、CBAM 注意力机制与 MHSA 多头自注意力机制的方法，提出了改进的行人检测方法。本文算法在 CUHK 数据集上进行训练及对比测试，实验结果表明：优化后的算法有着更强的特征提取能力，较大地提升了 YOLOv3 算法对行人的检测效果。

但本文提出的方法仍然存在问题，如改进的算法在当前的训练集上会有较好的提升，但在其它数据集上，若图片中行人受到严重遮挡或距离较远时，检测效果会有所下降。其次，算法未在实际道路及场景上进行测试，后续将进一步展开研究，尝试提高算法的抗干扰和实时检测能力。

参考文献

[1] FUJIYOSHI H，LIPTON A J. Real-time human motion analysis by image skeletonization[C]// Proceedings of the 4th IEEE Workshop on Applications of Computer Vision. Princeton，NJ，USA：IEEE，2002：15-21.

[2] WANG Heng，KLASER A，SCHMID C，et al. Action recognition by dense trajectories[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Colorado Springs，Co，USA：IEEE，2011：3169-3176.

[3] GAVRILA D M.Pedestrian detection from a moving vehicle[C]// European Conference on Computer Vision. Berlin/Heidelberg：Spring-Verlag，2000：37-49.

[4] 潘锋，王宣银，王全强.智能监控中基于头肩特征的人体检测方法研究[J]. 浙江大学学报，2004，38(04)：397-401.

[5] BOBICK A，DAVIS J.An appearance-based representation of action [C]// Proceedings of the 13th International Conference on Pattern Recognition(ICPR). Vienna，Austria：IEEE，1996：307-312.

[6] 刘菲. 运动人体行为分析系统及关键技术研究[D]. 西安：西安电子科技大学，2007.

[7] 唐勇，姜昱明. 彩色图像序列中运动人体轮廓提取[J]. 计算机工程与设计，2006，27(20)：3901-3903.

[8] VIOLA P，JONES M. Rapid object detection using a boosted cascade of simple features [C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. Kauai，HI，USA：IEEE Computer Society，2001：511.

[9] DALAL N，TRIGGS B. Histograms of Oriented Gradients for Human Detection [C]//IEEE Computer Society Conference on Computer Vision Pattern Recognition. San Diego，CA，USA：IEEE Computer Society，2005：886-893.

[10] 谭飞刚，刘伟铭. 多部件验证的双层行人检测算法[J]. 华南理工大学学报（自然科学版），2015，43(01)：79-86.

[11] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// Computer Vision and Pattern Recognition. Columbus，OH，USA：IEEE，2014：580-587.

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

[13] REN S，GIRSHICK R，et al. Faster R-CNN：Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，39(6)：1137.

[14] REDMON J，DIVVALA S，GIRSHICK R，et al. You only look once：Unified，real-time object detection [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas，NV，USA：IEEE，2015：779-788.

[15] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single shot multibox detector [C]// European Conference on Computer Vision. Cham：Springer，2016：21-37.

[16] JEONG J，PARK H，KWAK N. Enhancement of SSD by concatenating feature maps for object detection[J]. arXiv preprint arXiv：1705.09587，2017.

[17] LI Z，ZHOU F. FSSD：feature fusion single shot multibox detector[J]. arXiv preprint arXiv：1712.00960，2017.

[18] 王思远，王俊杰. 基于改进 YOLOv3 算法的高密度人群目标实时检测方法研究[J]. 安全与环境工程，2019，26(05)：194-200.

[19] 黄同愿，杨雪姣，向国徽，等. 基于 YOLOV3 的改进模型在行人检测中的应用[J]. 重庆理工大学学报(自然科学)，2020，34(08)：155-164.

[20] 武明虎，黄咏曦，王娟. 基于改进 YOLOv3 的街道行人检测与跟踪方法.科学技术与工程[J]. 2021，21(17)：7230-7236.

[21] HE K，ZHANG X，REN S，er al.Deep residual learning for image recognition[C]//Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition. New York：IEEE，2016：770-778.

[22] 陈皋，王卫华，林丹丹. 基于无预训练卷积神经网络的红外车辆目标检测[J]. 红外技术，2021，43(04)：342-348.

[23] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. Long Beach，CA，USA：NIPS Foundation，2017：30.