研究与开发

改进 YOLOv4 算法的复杂视觉场景行人检测方法

康帅[1]，章坚武[1]，朱尊杰[1]，童国锋[2]

（1. 杭州电子科技大学，浙江 杭州 310018；2. 绍兴供电公司柯桥供电分公司，浙江 绍兴 330600）

摘　要：复杂视觉场景下存在过暗或者过曝的光照、恶劣的天气、严重遮挡、行人尺寸差别大以及图像模糊等问题，大大增加了行人检测的难度。因此，针对复杂视觉场景下行人检测准确度低、漏检严重的问题，提出了改进的 YOLOv4 算法以增强复杂视觉场景下的行人检测效果。首先，构建复杂视觉场景下的行人数据集。然后，在主干网中加入混合空洞卷积，提高网络对行人特征的提取能力。最后，提出空间锯齿空洞卷积结构，代替空间金字塔池化结构，获取更多细节特征。实验表明，在本文构建的行人数据集上，改进后的 YOLOv4 算法的平均精度（average precision，AP）达到了 90.08%，相比原 YOLOv4 算法提高了 7.2%，对数平均漏检率（log-average miss rate，LAMR）降低了 13.69%。

关键词：复杂视觉场景；YOLOv4；混合空洞卷积；空间锯齿空洞卷积

中图分类号：TP393

文献标识码：A

doi: 10.11959/j.issn.1000−0801.2021198

An improved YOLOv4 algorithm for pedestrian detection in complex visual scenes

KANG Shuai[1], ZHANG Jianwu[1], ZHU Zunjie[1], TONG Guofeng[2]

1. Hangzhou Dianzi University, Hangzhou 310018, China

2. Keqiao Branch, Shaoxing Power Supply Company, Shaoxing 330600, China

Abstract: At present, the difficulty of pedestrian detection has been dramatically increased because of some problems, such as the dark or exposed illumination, bad weather, serious occlusion, large difference size of pedestrians and blurred images in complex visual scenes. Therefore, an improved YOLOv4 algorithm was proposed, which improved the detection performance of pedestrian detection in complex visual scenes, aiming at the problems of low accuracy and highly missed detection rate. Firstly, the self-annotation data set pedetrian were constructed. Secondly, the hybrid dilated convolution (HDC) was added into the backbone network to improve the ability of pedestrian feature extraction. Finally, in order to obtain more detailed feature, the spatial jagged dilated convolution (SJDC) structure was proposed to replace the spatial pyramid pooling structure. The experimental results show that the average precision (AP) of the proposed algorithm can achieve 90.08%. The proposed algorithm can substantially improve AP by 7.2%, and the log-average miss rate (LAMR) reduce by 13.69% compared with the original YOLOv4 algorithm.

Key words: complex visual scenes, YOLOv4, hybrid dilated convolution, spatial jagged dilated convolution

---

收稿日期：2021−03−17；**修回日期：**2021−08−13

通信作者：章坚武，jwzhang@hdu.edu.cn

基金项目：国家自然科学基金资助项目（No.U1866209，No.61772162）

Foundation Items: The National Natural Science Foundation of China (No.U1866209, No.61772162)

1　引言

行人检测是近年来计算机视觉领域的研究热点，同时也是目标检测领域中的难点。其目的是识别和定位图像中存在的行人，在许多领域中都有广泛的应用[1]。交通安全方面，无人驾驶汽车通过提前检测到行人及时避让来避免交通事故的发生；安防保护方面，通过行人检测来防止可疑人员进入；公共场所管理方面，通过行人检测统计人流量数据，优化人力物力等资源的分配[2]。

传统检测方法利用的大部分特征都是通过手工设计的。Viola 和 Jones[3]在 2001 年采用 Haar+AdaBoost 的人脸检测框架进行行人检测。Dalal 和 Triggs[4]在 2005 年提出了 HOG+SVM 的框架来检测行人。随后研究者将局部二值模式特征与方向梯度直方图（histogram of oriented gradient，HOG）相结合，采用随机森林作分类，构建了能够处理部分遮挡问题的行人检测模型[5]。2014 年有学者结合了积分直方图、梯度直方图和 Haar-like 等的特征，将聚合通道特征[6]用于行人检测。虽然上述传统的检测方法对行人检测都有了不同程度的改进，但是检测效果的好坏取决于设计者的先验知识，并没有很好地利用大数据的优势，对于复杂场景下行人目标多样性变化的情况算法鲁棒性较差。

随着深度学习的发展，越来越多的研究者开始使用深度学习进行目标检测。基于深度学习的检测方法比传统方法的检测性能更好，可以自主地从训练的数据集中学习不同层次的特征。该方法主要可以分成两类：第一类是基于候选区域的方法，代表性的有 2014 年提出的基于区域的卷积神经网络（R-CNN）[7]目标检测框架，以及在此基础上改进的快速区域卷积神经网络（Fast R-CNN）[8]、更快速区域卷积神经网络（Faster R-CNN）[9]。基于候选区域的检测方法虽然准确率有所提高，但是检测速度慢，很难达到实时性的要求。第二类是基于回归的方法，如 YOLO[10]、RetinaNet[11]、SSD[12]等方法，这类方法不再使用区域候选网络，而将检测转化为回归问题，提高了网络的检测速度，但是没有候选区域方法的检测效果好。YOLOv4 作为 YOLO 系列算法中较新的模型，在许多方面进行了优化，实现了速度与精度的最佳平衡[13]。为了使 YOLOv4 算法更适用于特定目标的检测，研究人员往往对通用的目标检测算法进行改进。李彬等[14]针对航空发动机部件表面特征不明显的问题，巧妙增加小卷积层来改进 YOLOv4 算法，提高了网络对部件表面缺陷检测的精度。Lee 等[15]提出了一种两阶段时尚服装检测方法（YOLOv4 two-phase detection，YOLOv4-TPD）对服装图像进行检测，效果优于原始网络模型。Wu 等[16]为使采摘机器人能够在复杂背景下快速准确地检测苹果，从数据增强和 YOLOv4 骨干网两方面进行改进，在降低网络计算复杂度的同时提高了检测性能。Fu 等[17]针对自建海洋目标数据集检测效果差的问题，在 YOLOv4 模型中加入卷积注意力模块，对海洋小目标、多目标有了更好的检测效果。本文检测的目标是行人，主要针对现实复杂视觉场景中检测精度不高、行人漏检严重的问题，对原始 YOLOv4[18]算法做了改进。

现实场景中行人检测的难点：复杂的光照情况，过暗或者过曝导致无法检测到行人；不同的天气情况（如暴雨、雾霾等）影响视线，对检测存在干扰；拥挤场景中，行人存在不同程度的遮挡，导致漏检严重；行人姿态多变，大小尺寸差别大；图像模糊、分辨率低等使检测不准确。

为了提高复杂视觉场景下行人的检测性能，本文采用改进的 YOLOv4 算法进行行人检测的研究。首先，收集不同复杂环境情况下的图片来构建行人数据集（包括复杂光照、不同天气、不同程度遮挡、不同视角、低分辨率和模糊图片等）；其次，采用先验框和聚类中心的交并比

（intersection over union，IOU）作为 *K*-means[19]算法参数，对不同尺度的锚框重新聚类，获得先验框；再次，在原来的 CSPDarkNet53 特征提取网络中加入混合空洞卷积（hybrid dilated convolution，HDC）[20]模块，提高行人检测的准确度。同时提出空间锯齿空洞卷积（spatial jagged dilated convolution，SJDC）结构，代替原算法中的空间金字塔池化（spatial pyramid pooling，SPP）[21]结构，降低复杂视觉场景下行人检测的漏检率；最后，使用测试集检测本文中两种不同改进后算法的性能，并与参考文献[14]改进的 YOLOv4 算法、原 YOLOv4 算法以及常用的 YOLOv3[22]算法进行对比。

2 YOLOv4 目标检测算法

YOLOv4 算法在之前的算法基础上增加了许多优化方法，包括骨干网络、激活函数、损失函数、网络训练、数据处理等方面的优化。YOLOv4 的网络结构主要有 3 个部分：骨干网络（CSPDarkNet53）、颈部（neck）、头部（head）[18]。

- YOLOv4 骨干网络采用的 CSPDarkNet53 是 DarkNet53 的改进版本，不仅解决了梯度消失的问题，增强了模型的学习能力，还减少了网络参数，降低了模型训练的代价。CSPDarkNet53 中使用的激活函数是 Mish 激活函数，与传统的负值硬零边界的 ReLU 相比，Mish 函数使网络具有更好的泛化性和准确性。
- 颈部采用路径聚合网络 PANet[23]，相比于 YOLOv3 中的特征金字塔网络 FPN 增加了一个自底向上的特征金字塔，从主干网络的不同层聚合不同的特征，增强网络的特征提取能力。同时引入 SPP 作为附加模块，增大了网络的感受野，进一步将重要的上下文特征信息提取出来。
- YOLOv4 网络的头部还是沿用 YOLOv3。损失函数方面，不再使用 YOLOv3 的均方误差（mean square error，MSE）作为回归框预测误差，而是使用 CIOU（complete intersection over union）误差，具体如式（1）～式（3）：

$$L_{\mathrm{CIOU}} = 1 - \mathrm{IOU}_{(a,b)} + \frac{\rho^2(a_{\mathrm{ctr}}, b_{\mathrm{ctr}})}{d^2} + \alpha\nu \tag{1}$$

$$\alpha = \frac{\nu}{(1 - \mathrm{IOU}_{(a,b)}) + \nu} \tag{2}$$

$$\nu = \frac{4}{\pi^2}\left(\arctan\frac{w_{\mathrm{gt}}}{h_{\mathrm{gt}}} - \arctan\frac{w}{h}\right)^2 \tag{3}$$

其中，$\mathrm{IOU}_{(a,b)}$ 是真实框和预测框的交并比，$\rho^2(a_{\mathrm{ctr}}, b_{\mathrm{ctr}})$ 是真实框和预测框中心点的欧氏距离，d 是包含真实框和预测框的最小框的对角线距离。w_{gt}、w、h_{gt}、h 分别是真实框和预测框的宽、高。

对于网络训练的正则化方式，采用 Dropblock 的方式线性增加特征图部分区域清零的比率，效果比传统的 Dropout 更好。对于数据集小、中、大目标数目不均衡的问题，YOLOv4 引入了 Mosaci 数据增强的方式，这种方式随机使用 4 张图片进行缩放、拼接，极大地丰富了数据集。

3 对 YOLOv4 算法的改进

3.1 改进后的网络框架

与原 YOLOv4 网络不同的是，本文在主干网 CSPDarkNet53 中加入 HDC 模块，提高了网络的特征提取能力。同时，提出 SJDC 结构代替 SPP 结构，获取更多的细节特征。改进后的 YOLOv4 网络结构如图 1 所示。

3.2 HDC 网络结构

YOLOv4 特征提取网络在检测时，浅层的网络特征为模型提供目标的位置信息，深层的网络特征为模型提供语义信息。浅层特征包含更多的位置信息，但感受野较深层特征更小一些。传统的网络中一般采用池化来增加感受野。但是采用池化的方式存在一个很大的缺点：在增加感受野的同时，往往会降低特征图的分辨率。

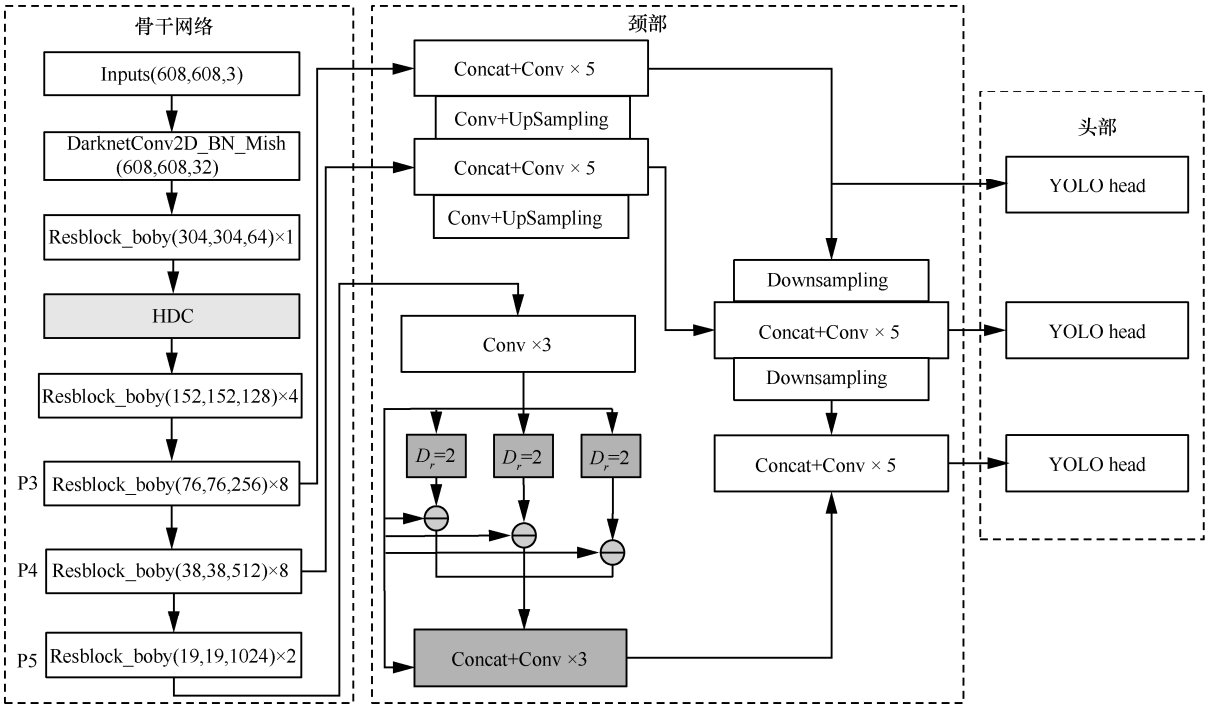


图 1　改进后的 YOLOv4 网络结构

为了更好地提升模型的检测能力，引入了空洞卷积（dilated convolution，DC）[24]，通过空洞卷积来增大浅层特征图的感受野，捕获更多尺度的上下文信息。空洞卷积与普通卷积的不同点在于卷积核中填充了 0。根据空洞率的不同，填充 0 的个数也不同，感受野也不同，从而获得更多不同尺度的特征信息。空洞卷积的感受野与卷积核的计算方法为：

$$x_n = x_k + (x_k - 1) \times (D_r - 1) \tag{4}$$

其中，x_n 是空洞卷积的卷积核大小，x_k 是原卷积核大小，D_r 为空洞率。

$$y_m = y_{m-1} + [(x_m - 1) \times \prod_{i=1}^{m-1} s_i] \tag{5}$$

其中，y_m 为第 m 层每个点的感受野，y_{m-1} 是第 $m-1$ 层每个点的感受野，x_m 是第 m 层卷积的卷积核大小，s_i 是第 i 层卷积的步长。

因为空洞卷积有着类似棋盘的计算方式，单加入一个尺度的空洞卷积会使局部的信息丢失，并且远距离得到的信息之间没有关联性，如图 2 所示。因此本文在 CSPDarkNet53 中引入 HDC 模块，将不同空洞率的空洞卷积进行融合。在不增加训练参数以及损失特征图分辨率的前提下，增大特征层的感受野，使特征提取网络能够获得更丰富的特征信息，增强对行人目标的检测效果。本文设计的 HDC 模块如图 3 所示。

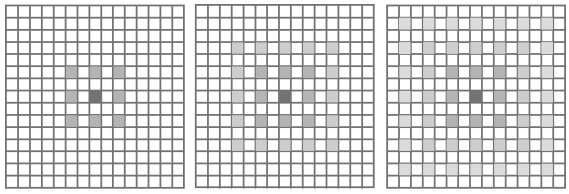


图 2　空洞率为 2 的空洞卷积

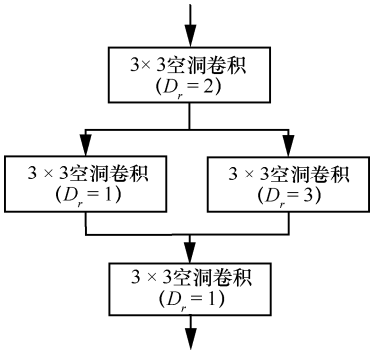


图 3　HDC 模块

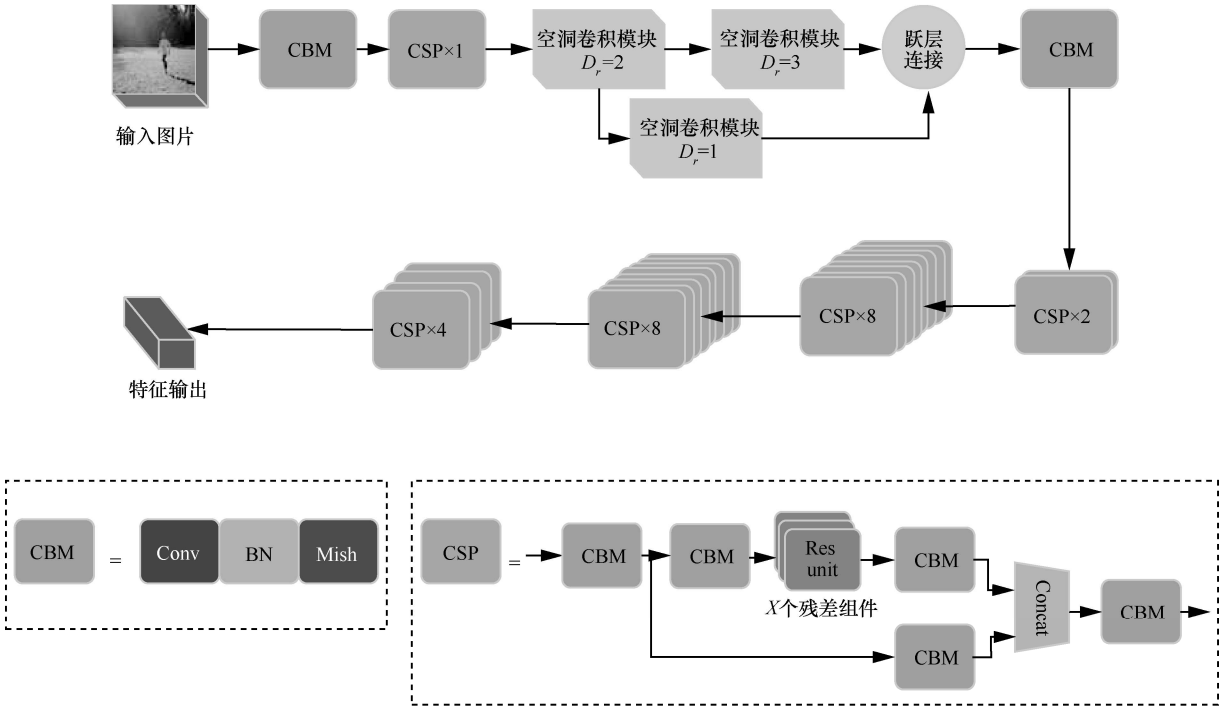


图 4　引入 HDC 后的骨干网结构

当空洞系数为 D_r 时，即在卷积核中加入（D_r–1）个空洞。特征图输入空洞卷积模块中，采用 3×3 的卷积核，当 D_r=2 时，感受野变为 7×7。再进行 D_r=3 的空洞卷积将感受野变为 19×19。不同层的特征进行直接融合，以获得更加广泛的图像特征，增强特征提取网络对于行人目标的特征提取能力。改进后的骨干网如图 4 所示。

3.3　SJDC 网络结构

YOLOv4 中从主干网 CSPDarkNet53 输出的特征层经过 3 次卷积后进入 SPP 结构。在 SPP 结构中利用 4 个不同尺度的最大池化 1×1、5×5、9×9、13×13 进行处理，增加感受野，分离出显著的上下文特征。但是这样会损失掉一些细节特征信息，使模型的漏检率过高，漏检的行人目标过多。SPP 的结构如图 5 所示。

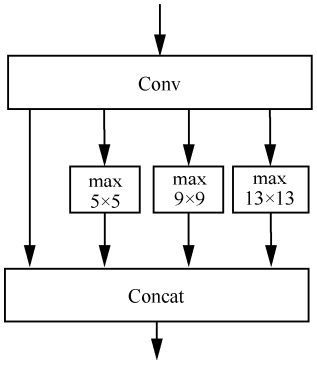


图 5　SPP 的结构

本文提出 SJDC 结构代替 SPP 结构，在保持特征图大小不变的情况下，使网络能获得更多的特征信息提高目标检测的效果。SJDC 网络的设计思想类似 SPP，区别在于 SJDC 将输入的特征层分为 4 路，其中 3 个分别进行系数为 2、5、7 的锯齿状的空洞卷积变换，获得不同尺度的特征信息。

同时采用 Mish 作为激活函数，表达式见式（6），这种自正则的非单调性平滑的激活函数可以使模型得到更好的准确性和泛化性。

$$f(x)=x\times\tanh(\log(1+e^{x})) \tag{6}$$

考虑到如果直接将得到的多尺度信息进行融合会有许多重复的特征信息，增大了计算的成本，因此本文提出的 SJDC 结构在信息融合之前进行了一个冗余信息剔除的操作，剔除后进行融合得

到特征的提取结果。这种网络结构的变化有效地降低了网络的漏检率，使复杂视觉场景下漏检的行人人数大大减少。SJDC 结构如图 6 所示。

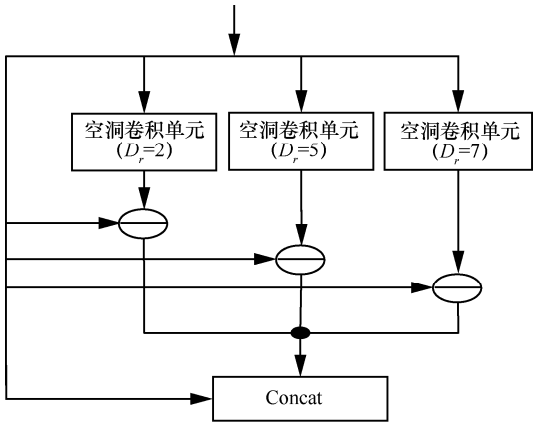


图 6　SJDC 结构

3.4　*K*-means 聚类

考虑到初始的先验框是在 VOC 数据集的标注框上聚类得到的，该数据集的物体种类有 20 种，而且物体长与宽的比例差别很大，与本文使用的行人数据集中的实际边界框存在很大的差别。如果直接使用 YOLOv4 的先验框会造成很大的漏检率。因此，本文采用 *K*-means 聚类算法重新对训练集中标注的行人尺寸进行聚类分析，获取最优的聚类结果。本文自建的数据集中行人的尺寸大小差距大，使用框与框中心点的欧氏距离作为候选框和真实框之间的误差会影响大尺寸标签的聚类结果，所以采用先验框和聚类中心的交并比 IOU 作为 *K*-means 聚类的参数，反映真实框和候选框之间的差距。计算聚类距离的公式如下：

$$d(\text{box},\text{center}) = 1 - \text{IOU}(\text{box},\text{center}) \qquad (7)$$

其中，box 为样本的先验框，center 为行人类的聚类中心[25]。经过实验，得到本文改进的 YOLOv4 算法候选框的初始大小为[11 29, 16 47, 24 62, 31 97, 52 119, 66 204, 109 249, 184 316, 309 381]。将宽和高归一化到 0～1，得到聚类结果如图 7 所示。

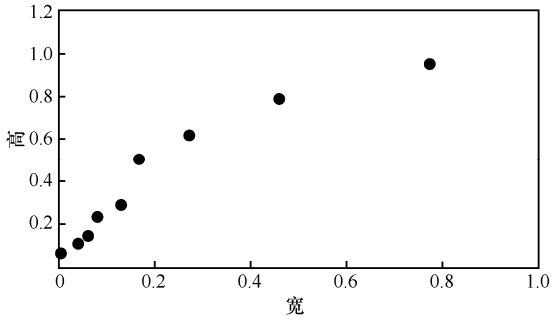


图 7　聚类结果分布

4　模型训练与调整

4.1　数据来源

目前目标检测常用的是 VOC2007 和 VOC2012[26]数据集，其中包含 20 个类别，从中挑选出包含行人目标的图片。为了更符合真实复杂场景，从 USC Pedestrian Detection Test Set、Caltech Pedestrian Detection Benchmark[27] 以及 CUHK Occusion Pedstrian[28]3 个公开行人数据集中挑选出部分图片加入本实验数据集中进行扩充。USC Pedestrian Detection Test Set 行人数据库包括各种视角的行人、行人之间相互遮挡以及无相互遮挡的情况。Caltech Pedestrian Detection Benchmark 行人数据库是车载摄像头视角下的数据集，不仅存在不同状况的光照变化，同时行人之间也有不同程度的遮挡。CUHK Occusion Pedstrian 数据集可用于拥挤场景中的行人目标检测。对数据集进行筛选，构建出本实验所需的行人数据集，共 5 329 张。由于公开数据集中的标注存在漏标和标注不符合预期的情况，因此，使用 labelimge 工具对数据集重新进行标注获得新的行人数据集。

4.2　数据集增强

本文所选的数据集中考虑了视角、遮挡、光照，但是并没有将复杂的天气因素考虑进去，如雾霾、暴雨等，这些因素也会影响模型的检测效果。为了更好地适应真实场景，本文从数据集中随机挑选出 500 张图片进行加雾、500 张图片进行加雨实现数据增强。图 8（a）、图 8（c）是未处

理的图片，图 8（b）是加雾处理后的图片，图 8（d）是加雨处理后的图片。将处理过后的图片加入数据集中形成包含 6 329 张行人图片的新数据集。行人数据集中各种情况分布如图 9 所示。将该数据集以 8:2 的比例划分，其中训练集包括 5 063 张图片，测试集包括 1 266 张图片。

图 8　加雾加雨处理图片对比

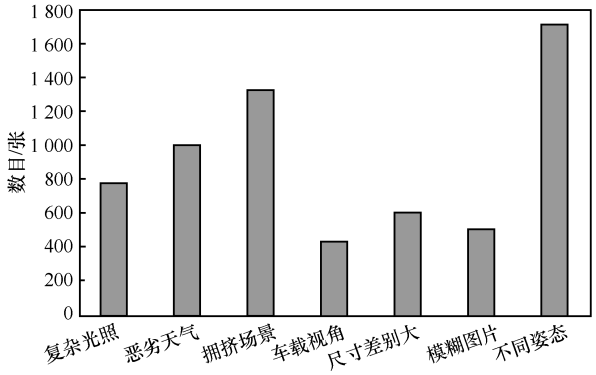


图 9　自建行人数据集中复杂视觉场景分布情况

4.3　模型的训练参数设置调整

本文在 Ubuntu16.04 操作系统进行实验，计算机内存为 24 GB、显卡为 TAITAN，采取从头开始训练的方式。每个网络训练 150 epochs，输入分辨率为 608 dpi×608 dpi，根据网络参数和 GPU 内存大小调整 batch_size 为 8。为防止过拟合，采用余弦退火衰减来实现训练过程中学习率的变化。将原始的学习率设置为 0.000 1，学习率的最高值设为 0.001，当学习率线性增长到最高值后保持一段时间不变，然后将模拟余弦函数的下降趋势进行衰减。余弦退火的衰减式如式（8）：

$$\lambda_t = \lambda_{\min+} \frac{1}{2}(\lambda_{\max} - \lambda_{\min})(1+\cos(\frac{T_{\text{cur}}}{T_i}\pi)) \quad (8)$$

其中，$\lambda_{\max}$ 和 $\lambda_{\min}$ 表示学习率的最大值和最小值，T_{cur} 表示当前执行了多少 epoch，T_i 表示模型训练的总 epoch 数目。

5　实验结果与分析

5.1　模型评估指标

本文采用对数平均漏检率（log-average miss rate，LAMR）、平均精度 $\text{AP}_{\text{person}}$（average precision）对目标检测网络进行比较。

（1）对数平均漏检率 LAMR

$$\text{LAMR} = \exp\{\frac{1}{9}\sum_{j=1}^{9}\ln[\text{missrate}(10^{-2.25+0.25\text{j}})]\} \quad (9)$$

LAMR 可以表明数据集中测试集的漏检情况。LAMR 越大，漏检行人数量越多，LAMR 越小，漏检行人数量越少。

（2）平均精度 $\text{AP}_{\text{person}}$

$$\text{AP}_{\text{person}} = \frac{\sum \frac{N(\text{TP})_{\text{person}}}{N(\text{Object})_{\text{person}}}}{N(\text{Images})_{\text{person}}} \quad (10)$$

用 $\text{AP}_{\text{person}}$ 来表示行人类检测的平均精度。$N(\text{TP})_{\text{person}}$ 表示正确检测出行人的数量，$N(\text{Object})_{\text{person}}$ 表示该图片中实际的行人数量，$N(\text{Images})_{\text{person}}$ 表示测试图片中含有的行人数量。$\text{AP}_{\text{person}}$ 越大，模型的误检和漏检就越少。

5.2　不同算法的检测效果对比

5.2.1　HDC-YOLOv4

针对复杂视觉场景下，行人检测中模型检测出的目标准确度不高的问题，本文在骨干网中加入 HDC 网络提高目标置信度。使用一阶目标检测领域常用的网络 YOLOv3、YOLOv4 原网络与本文加入 HDC 之后的 YOLOv4 进行对比。分别设置阈值 IOU 为 0.5、0.6、0.7、0.8，得到相应的 $\text{AP}_{\text{person}}$ 值，见表 1。相比于原来的 YOLOv4 模型，改进 HDC-YOLOv4 在阈值为 0.5、

0.6、0.7、0.8 时，AP_{person} 值分别提高了 2.07%、2.50%、2.98%、2.44%。在不同阈值的情况下，相比 YOLOv3 以及原来的网络都有了提高。主要原因是加入 HDC 后更有利于行人特征的提取，能够提高该算法对复杂场景下行人特征的识别度。

表 1　3 种网络模型在不同阈值下的 AP_{person} 对比

阈值	YOLOv3	YOLOv4	YOLOv5
0.5	79.16%	81.46%	83.53%
0.6	67.89%	71.55%	74.05%
0.7	44.93%	49.37%	52.35%
0.8	18.36%	20.62%	23.06%

为验证其有效性，将 IOU 设置为 0.5，从测试集中选取 2 个特殊情况下的图片进行测试，图 10 和图 11 是 3 个检测网络的对比。商场服装店前的模糊图片检测效果对比如图 10 所示，图片比较模糊且行人与店内悬挂的衣服有很大的相似性，导致 YOLOv3 和 YOLOv4 都产生了误检，本文改进的 HDC-YOLOv4 在复杂环境下精确地检测出了右侧的 2 个行人，并且置信度较高。雨天图片过暗检测效果对比如图 11 所示，图片光照过暗并且做了模拟下雨的图像增强操作，对视线产生了干扰，行人几乎与背景融为一体很难被识别检测到，YOLOv3 和 YOLOv4 均无法检测出该复杂场景下的行人，改进的 HDC-YOLOv4 检测出了 1 个行人，但距离较远、目标更小、更模糊的也未能检测出。因此，本文改进的算法仍然有提升的空间。总体看来，在一些特殊复杂场景下，加入 HDC 的 YOLOv4 网络对行人检测的准确度都有很大的提高。

(a)YOLOv3　(b)YOLOv4　(c)HDC-YOLOv4

图 10　服装店前的模糊图片检测效果对比

(a)YOLOv3　(b)YOLOv4　(c)HDC-YOLOv4

图 11　雨天图片过暗检测效果对比

5.2.2　SJDC-YOLOv4

除了检测准确度不高的问题以外，复杂视觉场景下行人检测还存在严重的漏检问题，本文提出 SJDC 结构代替 SPP 结构来解决行人的漏检问题。在本文制作的数据集上使用 YOLOv3、YOLOv4 和改进 SJDC-YOLOv4 模型进行对比，实验结果见表 2。结果表明，对于复杂情况下的行人检测，加入 SJDC 结构改进后的 SJDC-YOLOv4 比 YOLOv3、YOLOv4 两个模型的 LAMR 都低，相比于 YOLOv4 原来网络 LAMR 降低了 9.7%，行人目标的漏检情况有了很大的改善。

表 2　3 种网络在性能检测漏检情况上的对比

模型	Epoch	LAMR	AP_{person}
YOLOv3	150	42.42%	79.16%
YOLOv4	150	40.03%	82.88%
SJDC-YOLOv4	150	30.33%	86.90%

从数据集中分别选取行人存在严重遮挡和光照严重不足的特殊场景的图片，分别采用 YOLOv3、YOLOv4 和 SJDC-YOLOv4 模型进行检测，结果如图 12 和图 13 所示。图 12 中行人之间存在严重遮挡并且服装非常相似，存在“双重错觉”效应，容易产生漏检。YOLOv3 只检测到了 8 人（漏检 4 人），YOLOv4 检测到了 9 人（漏检 3 人），SJDC-YOLOv4 几乎全部检测出来（漏检 1 人）。从右边数第 2、4、6 个行人由于衣服颜色与遮挡者的完全一样，只露出很小部分的头部，YOLOv3 和 YOLOv4 对这种情况存在严重的漏检，改进 SJDC-YOLOv4 检测出了其中的 2 人。光照不足情况检测效果对比如图 13 所示，图 13 中右侧的阴影区中光线较暗，存在建筑物的遮挡，

其中一个行人距离较远，尺寸很小并且几乎与背景融为一体很难被检测到。YOLOv3 只检测到最左边的行人，YOLOv4 检测到了 2 个行人。SJDC-YOLOv4 不仅能够检测到距离较近的两人，而且还检测到了图中与背景几乎融为一体的小尺寸行人。从图 13 中的结果可以看出，SJDC-YOLOv4 在复杂情况下，能够有效减少网络的漏检情况。

(a)YOLOv3 (b)YOLOv4 (c)SJDC-YOLOv4

图 12 行人重叠遮挡检测效果对比

(a)YOLOv3 (b)YOLOv4 (c)SJDC-YOLOv4

图 13 光照不足情况检测效果对比

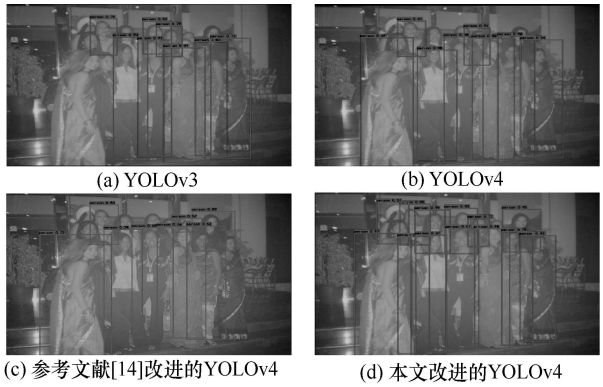
5.2.3 改进的 YOLOv4 算法与其他算法比较

本文针对复杂视觉场景下行人检测的准确度不高、漏检严重 2 个问题进行改进，在 YOLOv4 主干网 CSPDarkNet53 中加入 DHC，以及提出 SJDC 结构来代替 SPP 提升模型的检测性能。将本文改进的检测模型与原 YOLOv4 模型、参考文献[14]改进的 YOLOv4 模型以及其他典型的一阶检测模型 YOLOv3 网络进行对比，使用自建行人数据集进行实验，具体结果见表 3。从测试集中选取行人存在严重遮挡的图片并模拟雾霾天气进行加雾处理，将处理后的图片分别采用 YOLOv3、YOLOv4、参考文献[14]改进的 YOLOv4 模型以及本文改进后的 YOLOv4 模型进行检测，结果如图 14～图 15 所示。

通过表 3 可知，在自建的行人数据集包含多种复杂环境的情况下，改进后的 YOLOv4 模型比原 YOLOv4 模型的 AP_{person} 的值高了 7.20%，LAMR 降低了 13.69%。与参考文献[14]中改进的 YOLOv4 算法比较 AP_{person} 提高了 6.08%，LAMR 降低了 10.64%。说明本文的改进措施可以有效地提高模型对复杂视觉场景下行人的检测性能。

表 3 4 种网络在复杂视觉场景下的行人检测效果对比

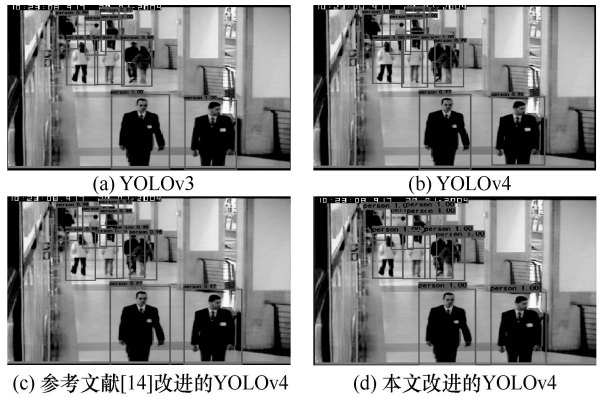
模型	LAMR	AP_{person}
YOLOv3	42.42%	79.16%
YOLOv4	40.03%	82.88%
参考文献[14]改进的 YOLOv4	36.80%	84.00%
本文改进的 YOLOv4	26.34%	90.08%



(a) YOLOv3 (b) YOLOv4

(c) 参考文献[14]改进的YOLOv4 (d) 本文改进的YOLOv4

图 14 雾霾天气行人严重重叠检测效果对比



(a) YOLOv3 (b) YOLOv4

(c) 参考文献[14]改进的YOLOv4 (d) 本文改进的YOLOv4

图 15 模糊图片下行人尺寸差距大且存在遮挡情况检测效果对比

在图 14 中，首先，加雾处理对行人检测造成了一定的困难；其次，图片中的行人存在非常严重的遮挡，这种场景下很容易检测不准确，产生漏检。从图 14 可看出，本文改进后的 YOLOv4 在检测的准确度和漏检情况上的表现都明显优于其他 3 个网络。图 15 中行人尺寸大小差距大，远距离的行人目标尺寸小且非常模糊，同时被前面的行人遮挡，在该场景下，改进后的 YOLOv4 检测出了全部的行人，并且准确度较高。YOLOv3 网络存在漏检，

YOLOv4 原网络和参考文献[14]中改进的 YOLOv4 网络虽然检测到了全部的行人，但是检测出的行人置信度没有改进后的 YOLOv4 高。

为验证改进后的网络对非复杂视觉场景的行人检测是否有效，本文另外选取了 2 000 张非复杂场景下的行人图片制作成数据集，并在 YOLOv3、YOLOv4、参考文献[14]改进的 YOLOv4 模型以及本文改进后的 YOLOv4 模型上重新训练并进行实验对比，结果见表 4。

表 4　4 种网络在非复杂视觉场景下的行人检测效果对比

模型	LAMR	AP_{person}
YOLOv3	27.17%	86.36%
YOLOv4	8.63%	95.89%
参考文献[14]改进的 YOLOv4	7.07%	96.22%
本文改进的 YOLOv4	5.70%	97.00%

由表 4 可知，在非复杂视觉场景下，本文改进的 YOLOv4 检测效果仍比其他 3 个网络好，但是相较于复杂视觉场景下 AP_{person} 的提升和 LAMR 的降低没有那么明显，也更加说明了本文改进的网络对复杂视觉场景行人检测的有效性。

6　结束语

本文提出了一种基于改进 YOLOv4 算法的复杂视觉场景行人检测方法。使用先验框和聚类中心的交并比作为 *K*-means 算法参数对行人数据进行聚类得到先验框，在主干网 CSPDarkNet53 中加入 HDC 结构，提高网络对复杂视觉环境下行人特征的提取能力，增加模型检测的准确度。同时，提出 SJDC 网络模块，用来代替 SPP 结构，获取更多的细节特征信息，降低模型漏检率。实验结果表明，改进后的方法在复杂环境下具有较好的检测效果，行人的平均精度 AP_{person} 达到了 90.08%，LAMR 相比原网络降低了 13.69%，FPS 为 25.84 f/s。改进后的 YOLOv4 网络在满足实时条件的情况下，检测准确度有所提高，漏检率大幅降低。

参考文献：

[1] LIN C Z, LU J W, WANG G, et al. Graininess-aware deep feature learning for robust pedestrian detection[J]. IEEE Transactions on Image Processing, 2020(29): 3820-3834.

[2] 王霞, 张为. 基于联合学习的多视角室内人员检测网络[J]. 光学学报, 2019, 39(2): 78-88.
WANG X, ZHANG W. Multi-view indoor human detection neural network based on joint learning[J]. Acta Optica Sinica, 2019, 39(2): 78-88.

[3] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Piscataway: IEEE Press, 2001.

[4] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Piscataway: IEEE Press, 2005.

[5] WANG X Y, HAN T X, YAN S C. An HOG-LBP human detector with partial occlusion handling[C]//Proceedings of 2009 IEEE 12th International Conference on Computer Vision. Piscataway: IEEE Press, 2009: 32-39.

[6] DOLLÁR P, APPEL R, BELONGIE S, et al. Fast feature Pyramids for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532-1545.

[7] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of 27th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.

[8] GIRSHICK R. Fast R-CNN[C]//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 1440-1448.

[9] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 779-788.

[11] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2999-3007.

[12] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[M]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[13] WU D H, LV S C, JIANG M, et al. Using channel prun-

ing-based YOLOv4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments[J]. Computers and Electronics in Agriculture, 2020, 178: 105742.

[14] 李彬, 汪诚, 吴静, 等. 改进 YOLOv4 算法的航空发动机部件表面缺陷检测[J]. 激光与光电子学进展, 2021: 1-17.
LI B, WANG C, WU J, et al. Surface defect detection of aero-engine components based on improved YOLOv4 algorithm[J]. Laser & Optoelectronics Progress, 2021: 1-17.

[15] LEE C H, LIN C W. A two-phase fashion apparel detection method based on YOLOv4[J]. Applied Sciences, 2021, 11(9): 3782.

[16] WU L, MA J, ZHAO Y H, et al. Apple detection in complex scene using the improved YOLOv4 model[J]. Agronomy, 2021, 11(3): 476.

[17] FU H X, SONG G Q, WANG Y C. Improved YOLOv4 marine target detection combined with CBAM[J]. Symmetry, 2021, 13(4): 623.

[18] BOCHKOVSKIY A, WANG C Y, LIAO H Y MARK. YOLOv4: optimal speed and accuracy of object detection[EB]. 2020.

[19] 李玉, 甄畅, 石雪, 等. 基于熵加权*K*-means全局信息聚类的高光谱图像分类[J]. 中国图象图形学报, 2019, 24(4): 630-638.
LI Y, ZHEN C, SHI X, et al. Hyperspectral image classification algorithm based on entropy weighted *K*-means with global information[J]. Journal of Image and Graphics, 2019, 24(4): 630-638.

[20] WANG P Q, CHEN P F, YUAN Y, et al. Understanding convolution for semantic segmentation[C]//Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2018: 1451-1460.

[21] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[22] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB]. 2018.

[23] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8759-8768.

[24] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 636-644.

[25] 赵朵朵, 章坚武, 傅剑峰. 基于深度学习的实时人流统计方法研究[J]. 传感技术学报, 2020, 33(8): 1161-1168.
ZHAO D D, ZHANG J W, FU J F. Research on real-time statistics of people flow based on deep learnning[J]. Chinese Journal of Sensors and Actuators, 2020, 33(8): 1161-1168.

[26] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[27] DOLLAR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: an evaluation of the state of the art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 743-761.

[28] OUYANG W L, WANG X G. A discriminative deep model for pedestrian detection with occlusion handling[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 3258-3265.

[作者简介]

康帅（1996− ），女，杭州电子科技大学通信工程学院硕士生，主要研究方向为计算机视觉与人工智能等。

章坚武（1961− ），男，博士，杭州电子科技大学通信工程学院教授、博士生导师，主要研究方向为移动通信、多媒体信号处理与人工智能、通信网络与信息安全。

朱尊杰（1994−），男，杭州电子科技大学通信工程学院讲师，主要研究方向为机器人导航定位、场景三维重建、三维模型语义理解与交互等。

童国锋（1968− ），男，绍兴供电公司柯桥供电分公司总经理、高级工程师，主要研究方向为继电保护、配网。