基于 YOLOv3 的红外行人小目标检测技术研究

李慕锴[1,2]，张 涛[1]，崔文楠[1]
（1. 中国科学院 上海技术物理研究所 上海 200083；2. 中国科学院大学，北京 100049）

摘要：针对红外图像中行人小目标检测识别率低、虚警率高的问题，研究了当下效果最好的 YOLOv3 目标检测算法，在其基础上进行优化，提出了一种满足实时性要求的行人小目标检测算法。基于 YOLOv3 中分类准确率仍有不足的情况，借鉴 SENet 中对特征进行权重重标定的思路，将 SE block 引入 YOLOv3 中，提升了网络的特征描述能力。通过对自行收集实际复杂场景下的红外图像进行目标检测，试验验证了算法的可行性，实验结果表明本文提出的改进网络拥有更高的准确率和更低的虚警率，同时保持了原有算法的实时性。

关键词：行人检测；红外小目标；深度学习；卷积神经网络
中图分类号：TP391.41；TJ765.3 **文献标识码：**A **文章编号：**1001-8891(2020)02-0176-06

Research of Infrared Small Pedestrian Target Detection Based on YOLOv3

LI Mukai[1,2]，ZHANG Tao[1]，CUI Wennan[1]
(1. *Shanghai Institute of Technical Physics*, *CAS*, *Shanghai* 200083, *China*;
2. *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract：To solve the problem of low recognition rate and high false alarm rate in the study of small pedestrian target detection in infrared image, this paper studies YOLOv3, one of the best target detection algorithms, and based on it proposes a small pedestrian detection algorithm that meets real-time requirements. Based on the fact that the classification accuracy is still insufficient in YOLOv3, this article studies the idea of feature reweighting from SENet, and introduces the SE block into YOLOv3, which improves the feature modeling ability of the network. The feasibility of the algorithm is verified by experiments with infrared images collected in actual complex scenes. The experiment results show that the improved network has higher accuracy and lower false alarm rate in small pedestrian detection task, and the algorithm maintains real-time characteristics of the original algorithm.

Key words：pedestrian detection, infrared small target, deep learning, CNN

0 引言

行人检测是图像处理研究中的经典课题，其研究成果在视频监控、地区侦查、人体行为理解、遇难目标搜救等领域都有诸多应用。随着近年来计算机视觉、机器学习和深度学习等新技术的突破，可见光图像中的行人检测技术已经逐渐发展成熟，出现了许多具有高可用性的方法。然而可见光相机的工作依赖于白天或者其他光照充足的条件，无法满足很多夜间场景下的监控需求，其工作的可持续性存在问题。红外相机基于目标对红外光的反射和目标自身的热辐射进行成像，受光照强度条件的影响很小，可以覆盖大多数夜间的场景，在白天也有很好的工作能力，因此红外相机能够更好满足持续工作的需求。并且随着红外成像系统价格的逐年降低，红外相机越来越成为各类监控系统中的重要组成部分，而红外图像中的行人检测技术问题也成为计算机视觉研究中的重点课题。

与可见光图像相比，红外图像仅有一个颜色通道，提供的信息更少，并且红外图像往往有分辨率低、物体边缘模糊、含有噪声、对比度较低等问题，使得红外图像中能够提取到的特征信息减少。红外图像中目标往往具有较高的亮度，特征更加明显。传统的行人检测方法主要是使用人为设计的特征提取器，如 Haar[1]、histogram of oriented gridients（HOG）[2]、

收稿日期：2018-12-10；修订日期：2020-01-10.
作者简介：李慕锴（1994-），男，江西人，硕士研究生，研究方向为红外图像处理、目标检测。E-mail：clovisr@163.com。

aggregate channel features（ACF）[3]等，来提取图像中行人目标的特征，然后再通过滑动窗口的方法对图像的局部提取特征，最后通过 support vector machine（SVM）[4]、adaboost 等分类器来判断区域是否有目标。深度学习将图像领域中各个问题的处理精度都提升到了一个更高的水平，在目标检测领域，主要分为两类方法，一类通过区域打分来预测目标区域，然后通过神经网络来对区域进行分类，这类方法以 R-CNN[5] 系列为代表，包括后续的 fast R-CNN[6]、faster R-CNN[7]以及 single shot multibox detector（SSD）[8] 等；另一类方法通过回归得出目标区域再通过神经网络分类，这类方法以 YOLO[9]，YOLO9000[10] 和 YOLOv3[11]为代表，这一系列的算法都在红外图像处理中有很多应用。

现有的检测算法中，以深度学习的目标检测算法最为优秀，不过 SSD、R-CNN 系列的网络复杂度过高，即使使用运算速度非常高的 GPU 也仍然运行缓慢，而 YOLO 系列的方法解决了网络复杂度过高的问题，在主流的 GPU 上算法的运行速度达到 60 fps 以上，能够满足实时性要求。本次研究中就以增强了小目标检测能力的 YOLOv3 为主要网络，通过对网络进行改进，进一步增强了特征描述能力，使其能够在实际的红外小目标处理问题中得到应用。

1 原理简介

1.1 YOLOv3 算法简介

YOLO 目标检测算法是 Redmon 等[9]在 CVPR2016 上提出的一种全新的端到端目标检测算法。与同期的 fast R-CNN，faster R-CNN 等算法使用区域建议网络预测目标可能的位置不同，YOLO 直接一次回归得出所有目标的可能位置，虽然定位精度有所降低，但是大幅度地提升了算法的时间效率，得到了具有高实时性的目标检测方法。经过近几年的改良，Redmon 等[10-11]在 YOLO 的基础上又提出了 YOLO9000、YOLOv3 目标检测算法，到 YOLOv3 其检测精度已经超过 faster R-CNN，与精度最高的 Retina net 基本持平，在保持高精度的同时，YOLOv3 的速度比其他算法要高 3 倍以上，是目前目标检测领域的最优秀的算法之一。

YOLOv3 在目标位置预测方面引入了 faster R-CNN 中使用锚点框（anchor box）的思想，在每一个特征图上预测 3 个锚点框。对于一幅输入图像，YOLOv3 算法将其分成 13×13 块，在每一个小块上预测 3 个目标的边界框，并且 YOLOv3 引入了多尺度融合的方法，对图像在 3 个尺度上进行目标边界框的预测，从而大幅提升了小目标检测的精度。目标边界框参数的计算如图 1 所示[11]。

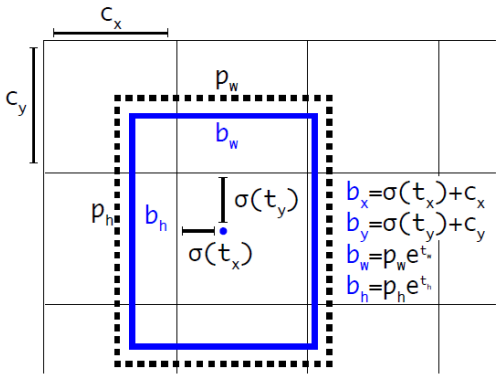


图 1 YOLOv3 边界框计算

Fig.1 YOLOv3 bounding box calculation

注：b_w 和 b_h 为边界框宽高，而 p_w 和 p_h 为锚点框的预测结果宽高。σ是 sigmoid 函数，t_x、t_y、t_w、t_h 为网络预测的目标中心坐标和宽高，c_x、c_y 为当前网格左上角坐标。

Note: b_w and b_h are the width and height of bounding box, p_w and p_h are the width and height of anchor box. σ is sigmoid function, t_x, t_y, t_w, t_h are coordinate and size of the network prediction, c_x and c_y are the coordinate of current cell's left-top corner

YOLOv3 在目标的分类上使用了比之前深度更大的神经网络，其网络中大量 3×3 和 1×1 的卷积核保证了良好的特征提取，使用多尺度预测提升了小目标的检测精度。在深度学习领域，更深的网络意味着可以提取更为复杂的特征，然而随着网络深度加大会出现训练难度加大，准确率下降的问题，Resnet 很好地解决了这个难题。YOLOv3 借鉴 Resnet 的思想，引入多个 Resnet 模块，设计了一个新的层数更多并且分类准确率更高的网络，由于其包含 53 个卷积层，作者将其命名为 darknet-53，其结构如图 2[11]。

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

图2 YOLOv3 网络结构 Fig.2 YOLOv3 network structure

1.2 SENet 简介

Squeeze-and-Excitation Networks[12] 由 Momenta 公司的 Jie Hu 等人提出，是一种能够显著提高网络性能的新型网络模型。目前在提升网络性能方面已经有大量的前人工作，有从统计角度出发的方法，例如 dropout 通过随机减少网络间的连接来减少过拟合；有从空间维度层面寻找提升的方法，例如 Inception 结构嵌入多尺度信息，聚合多种不同感受野上的特征来获得性能提升。而 SENet 从前人很少考虑到的特征通道间的关系出发，提出了一种特征重标定策略，这种策略通过显示建模特征通道间的相互依赖关系实现，可以通过学习来获取到每个特征通道的重要程度，然后根据这个主要程度来提升重要特征的权重并抑制不重要的特征。

SENet 中包含两个关键操作，压缩（Squeeze）和激励（Excitation），其主要流程如图 3[12]，其中 F_{tr} 和 F_{sq} 为压缩操作，F_{ex} 为激活操作，X 为输入，U 为中间变换结果，H、W、C 为网络的宽高和层数。压缩操作顺着空间维度来对提取到的特征进行压缩，将每个二维的特征通道换算为一个实数，这个实数在某种程度上会具有全局感受野，并且输出的维度和输入的特征通道数相匹配，它表征着在特征通道上响应的全局分布，而且使得靠近输入的层也可以获得全局的感受野。激励操作类似于循环神经网络中的门的机制，通过学习参数 w 来为每个特征通道生成权重，它可以通过两个全连接层实现，学习得到的参数 w 即表征了每个特征通道的重要性。最后的操作是权重重标定（Reweight），它将之前学习到的每个特征通道的权重归一化，然后通过乘法加权到原来的特种通道上，即完成了每个特征通道的重要性的标定。SENet 可以很方便地插入在 Resnet 之后，得到一个 SE-Resnet 模块，如图 4[12]所示。经过作者的多番验证，在不同规模的 Resnet 上引入 SENet 后，均能够大幅提升网络的准确率，并且作者依靠 SENet 赢得了 ImageNet 2017 图像分类任务的冠军。

2 改进 YOLOv3 网络

YOLOv3 在当前各类目标检测任务中已经取得了非常优越的效果，不过算法仍然有很多改进的空间，尤其对于小目标方面。在实际的红外行人小目标数据中，直接使用 YOLOv3 对数据进行训练，最后得到的模型具有良好的召回率，但是准确率不够。为了得到一个具有实时性，同时目标检测的准确率和虚警率都良好的算法模型，以 YOLOv3 为基础网络，结合 SENet 以提升分类网络的准确率是一个可行的思路。

根据 SENet 的思路，对网络进行改进一般有几种方式，一种是直接在卷积层后面直接加 SENet 模块，这种方法对所有网络都通用，但是由于现在的网络中都含有大量卷积层并且参数量巨大，这样添加 SENet 模块增加的参数量大，且需要大量实验来确定在哪些卷积层后面加入新模块。一种是用加入了 SENet 的模块替换原有网络中的 inception 或者 residual 层，这类方法替换位置较为明确，需要反复实验的可能性较小，并且作者也积累了一定经验。在 YOLOv3 中含有较多的 Residual 层，于是对网络的改进采取引入 SE-Resnet 模块的方法。

SE-Resnet 模块中，用 Global average pooling 层做压缩操作，将每个特征通道变换成一个实数值，使 C 个特征图最后变成一个 $1\times1\times C$ 的实数序列。被处理的多个特征图可以被解释为从图像中提取到的局部特征描述子的集合，每个特征图无法利用到其他特征图的上下文信息。使用 Global average pooling 可以使其拥有全局的感受野，从而让低层网络也能利用全局信息。

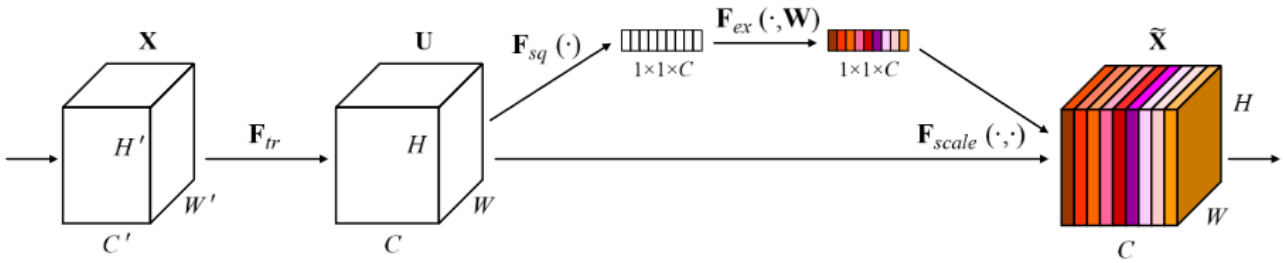


图 3 SENet 工作流程 Fig.3 SENet workflow

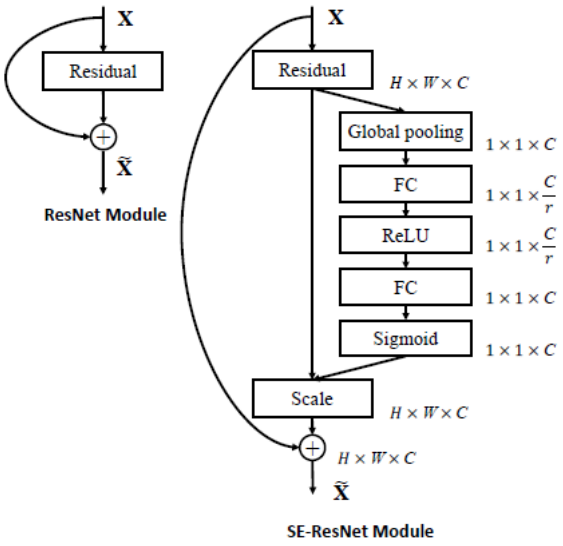


图 4 SE-Resnet 模块 Fig.4 SE-Resnet module

激活操作是 SENet 中用于捕获特征通道重要性和依赖性的关键操作，对于它的实现原作者使用了两个全连接层（full connected layer）结合 ReLU 函数去建模各个通道之间的相关性，并且其输出的权重数与输入的特征数相同。为了减少参数并且增强泛化能力，第一个全连接层将参数降维 r 倍，这里 r 取值为 16，然后经过一个 ReLU 后再经过一个全连接层升维到原来的维数。第二个全连接层后使用 sigmoid 激活函数作为阈值门限，得到了一个 $1\times1\times C$ 的序列，即每个特征通道的权重。最后将权重直接用乘法叠加到开始的特征通道上，即完成了所有特征通道的权重重标定。

引入 SENet 的 SE-Resnet 模块可以简化表示为一个 Residual 模块下添加了一个 SE 模块，如图 5[12]所示。

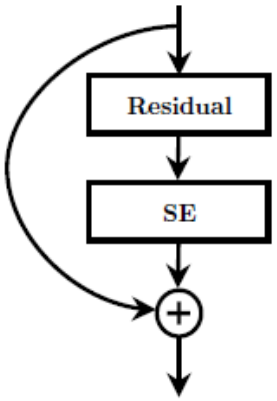


图 5 SE-Resnet 模块简化示意图

Fig.5 SE-Resnet module simplified diagram

SENet 模块的激活操作的实现中包含两个全连接层，全连接层的参数量相对其他类型的网络层是最大的，因此添加过多的 SENet 模块将会导致网络参数规模增大，影响目标检测算法的时间效率。根据原作者的经验，添加在网络末端的 SE 模块对准确率的影响较小，所以末端的几个 Residual 块不做处理。YOLOv3 包含 23 个 Residual 块，从减少模型参数量优化避免增加太多算法运行时间的角度考虑，只对每组卷积和残差层的最后一个残差层进行替换，于是改进后的网络结构如图 6 所示。

	Type	Filter	Size	Output
	Convolutional	32	3×3	256×256
	Convolutional	64	3×3/2	128×128
1×	Convolutional	32	1×1	
	Convolutional	64	3×3	
	Residual			128×128
	SE block			128×128
	Convolutional	128	3×3/2	128×128
2×	Convolutional	64	1×1	
	Convolutional	128	3×3	
	Residual			64×64
	SE block			64×64
	Convolutional	256	3×3/2	32×32
8×	Convolutional	128	1×1	
	Convolutional	256	3×3	
	Residual			32×32
	SE block			32×32
	Convolutional	512	3×3/2	16×16
8×	Convolutional	256	1×1	
	Convolutional	512	3×3	
	Residual			16×16
	SE block			16×16
	Convolutional	1024	3×3/2	16×16
4×	Convolutional	512	1×1	
	Convolutional	1024	3×3	
	Residual			8×8
	Avgpool		Global	
	Connected		1000	
	Softmax			

图 6 改进后网络结构

Fig.6 Network structure after improvement

3 实验过程与结果

3.1 数据收集与处理

实验使用焦距 20 mm，波段 8～12 μm 的长波红外热像仪在 50 m 的高度拍摄了 570 张单场景红外行人图像。数据拍摄地点在城市中，拍摄目标主体为从楼顶斜视的城市街道，因此数据集中的场景包含城市道路、建筑物、树木等，背景非常复杂。数据集图像中行人目标很小，在图像中的矩形框大小约为 13×8 个像素，形态特征较少，用传统特征提取方法将很难提取到有效特征，适合用深度学习方法进行目标检测。

数据中只对行人的目标进行了标注，为了能够提升目标检测的性能，提高泛化性，对一部分受到遮挡的行人目标也进行了标注，希望最后得到的模型能够应对一定程度的目标遮挡。由于图像数量较少，考虑对数据集进行数据增强，YOLOv3 在训练过程中有多尺度训练的部分，因此数据增强时不需要做尺度缩放，只使用翻转、加噪、随机光照改变等方法，数据增强后得到 2280 张图像，采集的红外图示例如图 7。

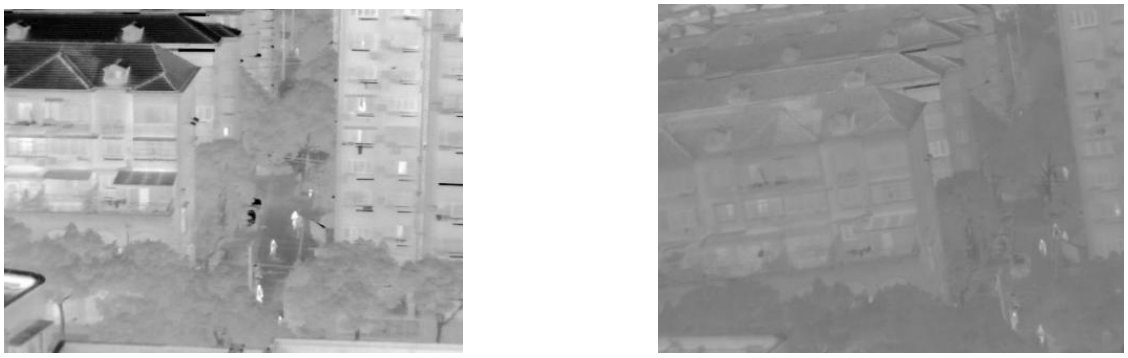


图 7 采集的红外图像示例 Fig.7 Infrared image example collected for experiment

3.2 模型训练

实验平台使用 Linux 16.04 LTS 系统，CPU i7 8700 k，GPU 为 NVIDIA GTX1080 8 G，16 G 内存。模型训练主要思路是使用已经在大规模数据集上训练好的模型进行 fine-tune，在新数据集上继续训练模型。以 YOLO 原作者在 COCO 和 VOC 上训练好的 darknet53 模型为基础模型，随机选取自建数据集中的 1710 张图像作为训练集，其余的 570 张图像为测试集，训练时初始学习率为 0.001，衰减系数为 0.0005，对于 YOLOv3 原网络和改进后的网络都进行训练。

3.3 实验结果

由于本次实验中的数据集只有一类目标，采用召回率（recall）和准确率（precision）作为模型的评价标准，其中准确率为网络预测的所有目标中真目标的比例，表征此网络的分类准确率；召回率为网络预测成功的真目标数与实际存在的真目标数的比值，表征此网络的查全率；以目标交并比（IOU，intersection over union）大于 0.5 为真目标，IOU 为预测目标矩形框和目标标签矩形重叠区域面积占二者并集面积的比值。

$$\mathrm{AP} = \frac{\mathrm{tp}}{n}$$

$$\mathrm{recall} = \frac{\mathrm{tp}}{\mathrm{tp+fn}} = \frac{\mathrm{tp}}{n'}$$

式中：tp 为网络预测出的真目标数；fn 为未能成功预测出的真目标数；n 为预测的总数；n'为标签目标数。

在训练好的模型上，用 570 张图像的测试集进行验证。YOLOv3 原网络和改进后网络的准确率和召回率对比如表 1 所示。

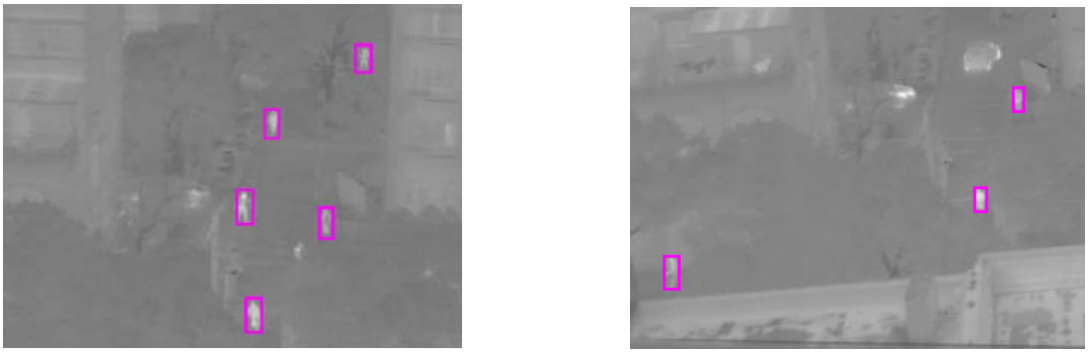
表 1 主要指标对比

Table 1 Comparison of primary specifications

Network	Precision	Recall
YOLOv3	81.33%	82.19%
SE-YOLOv3	85.89%	83.97%

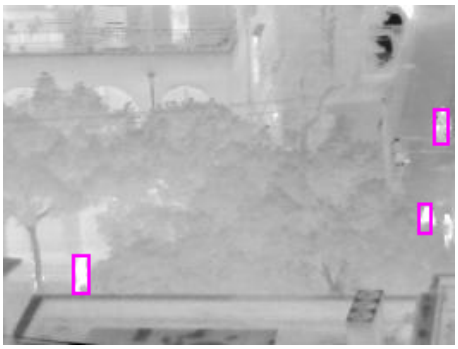
网络在测试图像中的检测效果如图 8 所示，可以看到红外图像中黯淡模糊的行人目标能够被检测出来，并且部分被遮挡的目标也有较好的检测能力。

从表 1 可以看到改进后的网络在两项主要指标上都优于原网络，由于 SENet 的特征权重重标定，增强了重要特征对分类结果的影响，抑制了非重要特征，使网络的特征描述能力进一步增强，最终令网络的召回率和准确率都得到提升。算法运行时间方面，在 GTX1080 显卡，CUDA9.0 运行环境下，570 张测试图片 YOLOv3 计算了 10.77 s，SE-YOLOv3 计算了 11.15 s，都在 50 fps 以上，网络增加的 SE block 带来的额外计算时间较少。



(a) Detect result of sample 1 (b) Detect result of sample 2

(d) Detect result of sample 3



(d) Detect result of sample 4

图 8 检测结果示例 Fig.8 Detect results of samples

4 结语

文章研究了当前主流的深度学习目标检测方法，以 YOLOv3 网络为基础，学习了 SENet 对特征进行权重重标定的思路，将 SE block 引入到 YOLOv3 网络中，得到了召回率和准确率都更高的新网络，并且保持了原有的高实时性。对实际收集的复杂红外图像进行试验，新网络取得了良好的行人小目标检测效果。

参考文献：

[1] Viola Paul, Jones M J. Robust teal-time face detection[J]. *Journal of Computer Vision*, 2004, **57** (2): 137-154.

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//*IEEE Conference of Computer Vision and Pattern Recognition*, 2005, **1**: 886-893.

[3] Dollár Piotr, Wojek Christian, Schiele Bernt, et al. Pedestrian detection: an evaluation of the state of the art[C]//*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **34**: 743-61(10.1109/ TPAMI. 2011.155).

[4] CHEN P H, LIN C J, Schölkopf B. A tutorial on ν-support vector machines[J]. *Appl. Stoch. Models. Bus. Ind*., 2005, **21**: 111-136.

[5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 580-587.

[6] Girshick Ross. Fast r-cnn[C]// *IEEE International Conference on Computer Vision* (ICCV), 2015: (DOI**:** 10.1109/ICCV.2015.169).

[7] REN S, HE K, Girshick R, et al. Faster R-CNN: towards real- time object detection with region proposal networks[C]//*Proceedings of International Conference on Neural Information Processing Systems*, 2015: 91-99.

[8] LIU W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//*Proceedings of European Conference on Computer Vision*, 2016: 21-37.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real- time object detectiono[C]//*Proceedings of CVPR*, 2015: 779-788.

[10] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger[C]// *Proceedings of CVPR*, 2016: .

[11] Redmon J, Farhadi, A.. YOLOv3: an incremental improvement[Z/OL][2018-04]. https://www.researchgate.net/publiccation/ 324387691_YOLOv3_An_Incremental_Improvement.

[12] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//*IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018: (DOI: 10.1109/TPAMI.2019.2913372).