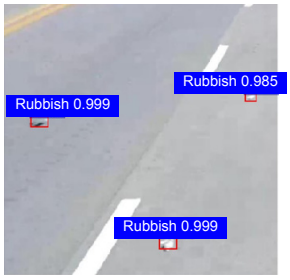
Opto-Electronic Engineering
光 电 工 程

Article
2019 年，第 46 卷，第 9 期

DOI: 10.12086/oee.2019.190053

城市道路视频中小像素目标检测



金　瑶[1,2]，张　锐[1,2]，尹　东[1,2*]

[1] 中国科学技术大学信息科学技术学院，安徽 合肥 230027；
[2] 中国科学院电磁空间信息重点实验室，安徽 合肥 230027

摘要：视频图像中的小像素目标难以检测。针对城市道路视频中的小像素目标，本文提出了一种改进 YOLOv3 的卷积神经网络 Road_Net 检测方法。首先，基于改进的 YOLOv3，设计了一种新的卷积神经网络 Road_Net；其次，针对小像素目标检测更依赖于浅层特征，采用了 4 个尺度检测方法。最后，结合改进的 M-Softer-NMS 算法来进一步提高图像中目标的检测精度。为了验证所提出算法的有效性，本文收集并标注了用于城市道路小像素目标物体检测的数据集 Road-garbage Dataset，实验结果表明，本文算法能有效地检测出诸如纸屑、石块等在视频中相对于路面的较小像素目标。

关键词：视频图像；小像素目标；卷积神经网络
中图分类号：TB872；TP391.4　　**文献标志码：**A
引用格式：金瑶，张锐，尹东. 城市道路视频中小像素目标检测[J]. 光电工程，2019，**46**(9): 190053

Object detection for small pixel in urban roads videos

Jin Yao[1,2], Zhang Rui[1,2], Yin Dong[1,2*]

[1]College of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China;
[2]Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei, Anhui 230027, China

Abstract: Small pixel targets in video images are difficult to detect. Aiming at the small pixel target in urban road video, this paper proposed a novel detection method named Road_Net based on the YOLOv3 convolutional neural network. Firstly, based on the improved YOLOv3, a new convolutional neural network Road_Net is designed. Secondly, for small pixel target detection depending on shallow level features, a detection method of 4 scales is adopted. Finally, combined with the improved M-Softer-NMS algorithm, it gets higher detection accuracy of the target in the image. In order to verify the effectiveness of the proposed algorithm, this paper collects and labels the data set named Road-garbage Dataset for small pixel target object detection on urban roads. The experimental results show that the algorithm can effectively detect objects such as paper scraps and stones, which are smaller pixel targets in the video relative to the road surface.

Keywords: video image; smaller pixel object; convolutional neural network
Citation: Jin Y, Zhang R, Yin D. Object detection for small pixel in urban roads videos[J]. *Opto-Electronic Engineering*, 2019, **46**(9): 190053

收稿日期：2019-01-30；　收到修改稿日期：2019-04-08
基金项目：2018 年度安徽省重点研究和开发计划项目(1804a09020049)
作者简介：金瑶(1995-)，女，硕士研究生，主要从事计算机视觉的研究。E-mail：joye@mail.ustc.edu.cn
通信作者：尹东(1965-)，男，副教授，主要从事图像处理的研究。E-mail：yindong@ustc.edu.cn

1 引 言

在城市道路上如何快速准确地发现和检测各种废弃物，已成为环境部门亟待解决的问题之一。随着视频图像处理技术的发展，使用摄像头记录路面状况，并通过图像处理方法分析和检测废弃物已成为可能。

物体检测是图像处理中重要的研究方向之一。传统方法利用尺度不变特征变换[1-2]、定向梯度直方图[3]、局部二值模式[4]来提取特征并使用 SVM[5]、KNN、随机森林[6]等进行分类，利用传统算法如 MIT[7]、粒子滤波[8]等实现目标检测。而如今，随着卷积神经网络(convolution neural network，CNN)的快速发展，已广泛应用于人脸识别[7]和视频分析[9-10]等领域中，而且在目标检测方面也取得了较好的表现。基于 CNN 的第一个物体检测算法是 Girshick 等提出的 R-CNN[11]，它首次提出将选择性搜索与 CNN 结合在分类任务中。Fast R-CNN[12]通过使用区域共享深度特征映射的算法来降低操作成本。Faster R-CNN[13]引入了区域提议网络(region proposals network，RPN)，其与检测共享全图像卷积特征，从而实现了几乎无成本的区域提议(region proposals)。基于 Fast R-CNN，Shrivastava[14]提出了一种 online hard example mining(OHEM)算法，用于训练基于区域的 ConvNet 探测器，该算法简单但令人惊讶地有效。与这些工作不同，Redmon[15]提出了一个 You Only Look Once(YOLO)框架，将对象检测作为一个回归问题，在空间上分离边界框和关联类别概率。YOLO 在各个领域应用时相对于 RCNN 具有更好的泛化性。

为了实现实时高效的城市道路上的废弃物检测，本文提出了改进的 YOLOv3 算法，完成了道路上小像素目标物体的检测。主要贡献如下：

1) 提出了一个基于 YOLOv3 的卷积神经网络 Road_Net，保证了算法有较好的实时性和较高的检测精度。

2) 小像素目标检测更依赖于浅层特征，故对多尺度进行改进，将 3 个尺度检测增至 4 个尺度检测来提高检测的正确率。

3) 结合改进的 M-Softer-NMS 算法，进一步提高了目标的检测准确率。

2 相关理论

2.1 目标检测

目标检测通常包含三个部分：选择检测窗口、特征提取和分类。

1) 检测窗口

基本方法是遍历整个图像以获得可能的 box 位置和多尺度搜索的缩放图像。目前流行的检测窗口方法主要是选择性搜索[16]和 Edge Box[17]，它们是基于颜色聚类和边缘聚类，并且还具有精度高、耗时少的特点。

2) 特征提取

定向梯度直方图(histogram of oriented gradient，HOG)使用直方图统计来编码对象的边缘。

第一，它规范化图像的颜色空间，如式(1)所示：

$$Y(x,y)=I(x,y)^{\gamma} \quad , \tag{1}$$

其中：$Y(x,y)$ 是位置 (x,y) 的输出值；$I(x,y)$ 为 0.3R+0.59G+0.11B，γ =0.5。

第二，计算梯度 G 和梯度方向 θ：

$$G(x,y)=\sqrt{G_x(x,y)^2+G_y(x,y)^2} \quad , \tag{2}$$

$$\theta(x,y)=\arctan\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \quad , \tag{3}$$

其中：$G_x(x,y)$ 表示水平梯度，为 $I(x+1,y)-I(x-1,y)$；$G_y(x,y)$ 表示垂直梯度，为 $I(x,y+1)-I(x,y-1)$。

最后，图像被分成若干个单元格，每个单元格包含不重叠的 8 pixels×8 pixels。一个 blob 包含相邻的 4 个单元格，在每个 blob 上，L2 范数用于标准化特征向量。HOG 的特征是每个 blob 的特征向量的组合。

3) 分类器

支持向量机(support vector machine，SVM)在高空间构造超平面或超平面集，可用于分类。对于线性 SVM，给定形式(x_i，y_i)的 n 点的训练数据集，其中 y_i 是 1 或-1，表示点 x_i 所属的类，x_i 是 p 维向量，可以将超平面写为满足式(4)的点 x 的集合。

$$\boldsymbol{w}\cdot\boldsymbol{x}-b=0 \quad , \tag{4}$$

其中：$\boldsymbol{w}$ 是超平面的法向量，b 确定偏移量。

然后，根据式(5)最小化$||\boldsymbol{w}||$，

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x}-b)\geq 1 \ \text{ for } \ i=1,\cdots,n \quad , \tag{5}$$

其中：$\boldsymbol{w}$ 和 b 确定分类器，即 $\boldsymbol{x}\rightarrow \mathrm{sgn}(\boldsymbol{w}\cdot\boldsymbol{x}-b)$。

2.2 YOLOv3

YOLOv3 是目标检测中最好的检测框架之一。YOLOv3 的 Darknet53 一方面基本采用全卷积，另一方面又引入了残差结构，使得训练深层网络难度大大减小，精度提升比较明显。在这个网络结构中，使用的是步长为 2 的卷积来进行降采样。同时，网络中使用了上采样、route 层、YOLO 层，并且在 32 倍降采样、16 倍降采样、8 倍降采样时进行检测，把三次下

采样的特征图拼接在一起，这样通过不同尺度的特征图进行融合，让网络同时学习深层和浅层特征，获得更好的表达效果。

与之前的 YOLO 版本一样，YOLOv3 继续采用先验框(anchor box)机制，利用 K-means 聚类获得先验框的大小，并且重新定义了 K-means 聚类方法中距离的判定，如式(6)。

$$d(box,centroid)=1-IOU(box,centroid) \tag{6}$$

作者通过 PASCAL VOC 数据集得到的先验框大小分别为(10, 13)，(16, 30)，(33, 23)，(30, 61)，(62, 45)、(59, 119)，(116, 90)，(156, 198)，(373, 326)共九个先验框，每种尺度预测三个先验框。

YOLOv3 对每个边界框来预测四个坐标值(t_x ， t_y ， t_w ， t_h)， t_x ， t_y ， t_w ， t_h 分别代表边界框的中心点横纵坐标和 bounding box 的长宽。YOLOv3 通过相对网格坐标来预测目标框的中心位置，如图 1 所示。

对于预测的每个网格根据图像左上角的偏移(c_x ， c_y)，以及 bounding box 的宽和高 p_w、p_h 可以对 bounding box 按式(7)进行预测：

$$b_x=\sigma(t_x)+c_x \ ,\ b_y=\sigma(t_y)+c_y \ ,$$
$$b_w=p_w e^{t_w} \ ,\ b_h=p_h e^{t_h} \ 。 \tag{7}$$

YOLOv3 在目标检测方面虽取得了不错的效果，但却不是完全适用道路检测视频画面中小像素目标检测。因为对于小像素目标检测，往往更依赖浅层特征，但是原网络结构中经过 53 个卷积层进行特征提取后，高分辨率的浅层特征则很少被利用，导致对应特征图上的特征往往难以得到充分训练。由于网络层数过大，原网络显得过于复杂和冗余，过多的参数会带来训练的难度、增大对数据集的要求以及降低了检测速度，在城市道路的小物体检测中，准确度和实时性都受到了挑战。

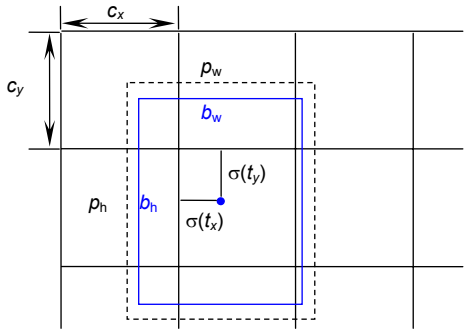


图 1 目标框位置预测

Fig. 1 Predicting target box position

3 本文算法

3.1 Road_Net

本文工作是检测城市道路上的废弃小像素目标物体，如石块、落叶、纸屑等，这些目标物体具有小尺寸和低分辨率的特点，因此现有算法无法充分提取其特征。在保持较高的检测率以及减少参数的情况下，本文借鉴 Darknet53 算法提出了一种运算复杂度相对更低的卷积神经网络 Road_Net 作为特征提取网络。由于使用非线性激活函数会在一定程度上破坏图像信息，故本文在低维度的卷积层中即在第一、二个卷积层上使用线性激活函数，如图 2 所示。

	Type	Outputs	Filters	Size
	Conv	416×416	32	3×3/1
	Conv	208×208	64	3×3/2
3×	Residual Block 1	208×208	32 64	1×1/1 3×3/1
	Conv	104×104	128	3×3/2
3×	Residual Block 2	104×104	64 128	1×1/1 3×3/1
	Conv	52×52	256	3×3/2
3×	Residual Block 3	52×52	128 256	1×1/1 3×3/1
	Conv	26×26	512	3×3/2
3×	Residual Block 4	26×26	256 512	1×1/1 3×3/1
	Conv	13×13	1024	3×3/2
3×	Residual Block 5	13×13	512 1024	1×1/1 3×3/1

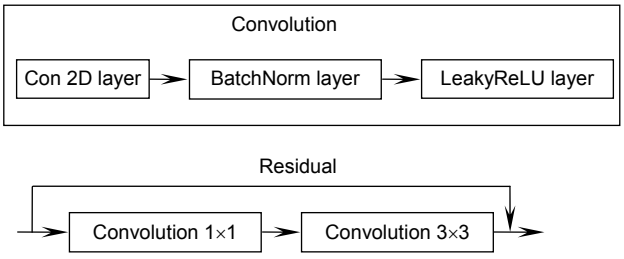


图 2 Road_Net 架构图

Fig. 2 Road_Net network architecture diagram

这里借鉴了深度残差网络[18]的思想，通过快捷连接(shortcut)有效地解决了梯度消失的问题，强化了特征的传递，有效地复用了卷积神经网络的特征，减少了参数数量，从而减少了计算量。相对于本文的小数据集而言，这可以很好地解决过拟合的问题。Road_Net 网络中使用了 5 个 residual block 模块，其中 residual block 是在前一个 residual 后经过一个步长为 2 的下采样，再经过 1×1、3×3 的卷积后进行快捷连接。以这种方式使卷积神经网络的模块间连接得到增强，减少跨模块间的特征传递损失和增强特征复用。

3.2 多尺度检测改进

YOLOv3 引入了 FPN[19]的思想，融合高分辨率的浅层特征和高语义信息的高层特征，在三个不同尺度的特征上检测物体。针对本文背景下的目标特点，将原有的 3 个尺度检测扩展为 4 个尺度检测[20]，这样可以在较大的特征图上给目标分配更为准确的框。利用 K-means 聚类获得锚(anchor)的大小，并取 12 个锚点框，分别为(12, 16)，(16, 24)，(21, 32)，(24, 41)，(24, 51)，(32, 51)，(28, 62)，(39, 64)，(35, 74)，(44, 87)，(53, 105)，(64, 135)。故在每个尺度上的每一个单元格要借助 3 个锚点框(anchor)来预测三个边界框。

本文提出进行 4 个尺度检测，即将原先的在 13×13、26×26、52×52 这三个尺度上再增加一个 104×104 尺度的检测，并且将不同尺度的特征进行融合后再进行预测，增强各个尺度特征层的语义信息。图 3 为本文提出的多尺度检测模块，其中的 Convolutional set 分别为 1×1、3×3、1×1 的卷积，upsampling 是步长为 2 的上采样过程，Concatenate 是将经过上采样的高维特征与低维特征进行拼接，之后经过 Convolutional set 等一系列卷积输出预测结果。

3.3 改进的 Softer-NMS 算法

绝大部分目标检测方法，最后都要用到非极大值抑制(non maximum suppression，NMS)算法进行处理。首先是将检测框按得分排序，然后保留得分最高的框，同时删除与该框重叠面积大于一定比例的其它框，该过程被不断的递归应用于其余检测框。如果一个物体处于预设的重叠阈值之内，则可能会导致检测不到该待检测物体。而 Soft-NMS[21]算法中的连续函数对非最大检测框的检测分数进行衰减而非彻底移除，Softer-NMS[22]是在 Soft-NMS 后进一步改进，提出了一种新的包围框回归的损失函数(KL loss)，用来同时学习包围框变换和定位置信度。KL loss 即为最小化包围

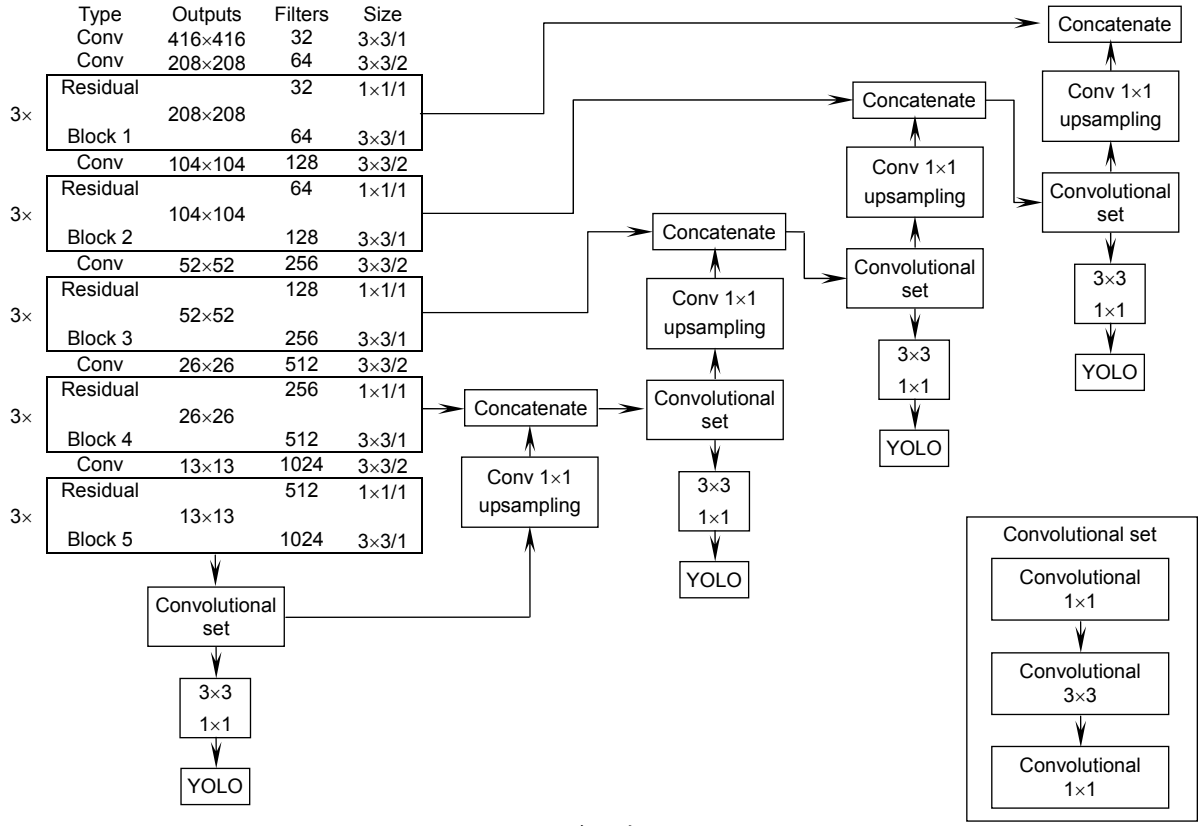


图 3 多尺度检测

Fig. 3 Multi-scale detection

框预测的高斯分布 $P_\theta(x)$和 groundtruth 的 δ分布 $P_{\rm D}(x)$ 的 KL 散度，如式(8)。

$$P_\theta(x)=\frac{1}{2\pi\sigma^2}\exp\left(-\frac{(x-x_e)^2}{2\sigma^2}\right)\ ,\ P_{\rm D}=\delta(x-x_{\rm g})\ ,$$
$$L_{\rm reg}=D_{\rm KL}(P_{\rm D}(x)||P_\theta(x))\ 。 \tag{8}$$

直观上解释，KL Loss 使得包围框预测呈高斯分布，且与 groundtruth 相近，将包围框预测的标准差看作置信度。当高斯函数的 σ 趋于 0 的时候是非常像 δ 分布的，因此在训练的过程中，通过 KL 散度优化两者的距离，调整置信度 σ 和 $x_{\rm e}$，使网络输出的分布尽量输出高置信度，而且接近 groundtruth 的偏移框。

本文改进的 Softer-NMS 伪算法 M-Softer-NMS 如算法 1。

其中，B 是初始的检测框，S 是对应于初始检测框的检测分数，C 是对应的方差，$N_{\rm t}$ 是 M-Softer-NMS 的阈值。

4 实验结果与分析

本文实验环境为：Inter(R) Core(TM)i7-4790 CPU @3.60 Hz，16 G 内存，Nvidia Geforce GTX1080，Ubuntu16.04，64 位操作系统，采用 Darknet 框架。使用精度、召回率与 AP 来评价模型的性能：

精度: $P=\frac{T_{\rm P}}{F_{\rm P}+T_{\rm P}}$；召回率: $R=\frac{T_{\rm P}}{T_{\rm P}+F_{\rm N}}$；平均精度: $AP=\int_0^1 P(r){\rm d}r$。其中：$T_{\rm P}$ 为真正例，$F_{\rm P}$ 为假正例，$F_{\rm N}$ 为假负例。

4.1 实验数据集

本文的数据集 Road-garbageDataset 为某城市的几条主干道，选取 1200 张不同地区的不同主干道路上的目标(如落叶、石块等)进行标注，这些图像由安装在车辆上的 HD 摄像机拍摄，将图像从 2014×1536 裁剪为 624×624。随机选取其中的 800 张作为训练集、200 张作为验证集、200 张作为测试集。

4.2 训练方法

在训练阶段，动量(momentum)为 0.9、衰变值(decay)为 0.0005、批尺寸(batch_size)为 64，使用小批量随机梯度下降进行优化，初始学习率为 0.001，整个过程的学习率为 10^{-3}、10^{-2}、10^{-4}，分别对应于前 10000 次、前 10000~30000 次、前 30000~45000 次。采用数据增强，包括调整饱和度、曝光度、随机裁剪等来增加训练样本。

4.3 检测结果定量评估

将本文的算法与 FasterR-CNN、YOLOv2、YOLOv3 进行定量评估，图像在训练前会处理为 416×416 大小，性能比较结果见表 1。在相同的测试集下，本文提出的 Road_Net 算法相比于 FasterR-CNN、YOLOv2 在精度、召回率、AP 上都提升了约 9 个百分点，相比于 YOLOv3 也分别提升了 1.6%、7.2%、4.3%，相比于 YOLOv3 在速度上(帧数/秒)提升了 43%。本文在

算法 1：M-Softer-NMS

Input: $B=\{b_1, .., b_N\}$, $S=\{s_1, .., s_N\}$, $C=\{\sigma^2_1, .., \sigma^2_N\}$, $N_{\rm t}$

Output : D，S

```
Begin:
D ← {}                                  //初始化 D
while B != empty do
    m←argmax S                          //取出最高的得分
    M←b_m                               //取出得分最高对应的检测框
    D←D∪M                               //更新 D
    B←B–M                               //更新 B
    for b_i in B do
      idx←IOU(M, B)≥N_t                 //取出 IOU 值大于阈值 N_t 的下标
      M ← B[idx]/C[idx]/sum(1/C[idx])   //按方差的倒数加权去和得到新的检测框
    end for
  endwhile
 return D, S                            //返回检测框和对应的分数
end
```

Road_Net 上再结合 Softer-NMS 算法进行测试，虽然在速度上略有下降，但是在召回率、*AP* 分别提升了 5.4%、2.4%，证明了本文提出的 Road_Net+M-Softer-NMS 在保证了实时性的情况下，在检测的召回率、平均精度上都有较大的提升。

表 1　5 种算法的性能对比

Table 1　Performance comparison of five algorithms

Method	*P*/%	*R*/%	*AP*/%	速度/(f/s)
Faster R-CNN	89.63	70.5	70.65	21.6
YOLOv2	86.45	63.18	71.53	44.2
YOLOv3	92.56	78.5	75.64	33.2
Road_Net	94.18	85.71	79.97	58.7
Road_Net +M-Softer-NMS	95.29	91.12	82.41	57.9

4.4　Road-garbage 的检测结果

图 4 显示了一些测试图像和检测结果，图 5 展示了一些异常检测结果。结果表明本文方法在废弃物检测上表现更好并具有发展潜力。

在图 4 中，红色框表示目标位置的预测，蓝色数字是检测目标的置信分数。从图 4 可知，置信度均超过了 96%，说明本文算法可以有效地检测道路上的小像素目标物体。

图 5(a)中的警示帽误检了；图 5(b)漏检了，是因为物体颜色和地面太近；图 5(c)虽然已检测到物体，但未标全，是因为超出了小像素目标尺寸范围；图 5(d)只正确检测出了一个物体，漏了上面一个。总之，检测异常的原因是：首先，不稳定和多样化的道路目标形状使其难以检测，这可能导致错过不常见的目标；其次，具有强泛化能力的网络模型需要很大的训练集，而目前的 Road-garbageDataset 数据量不够充足。此外，由于本文的图像是从真实场景中收集的，其包含不同的照明度、背景和干扰物体，从而引起误检。

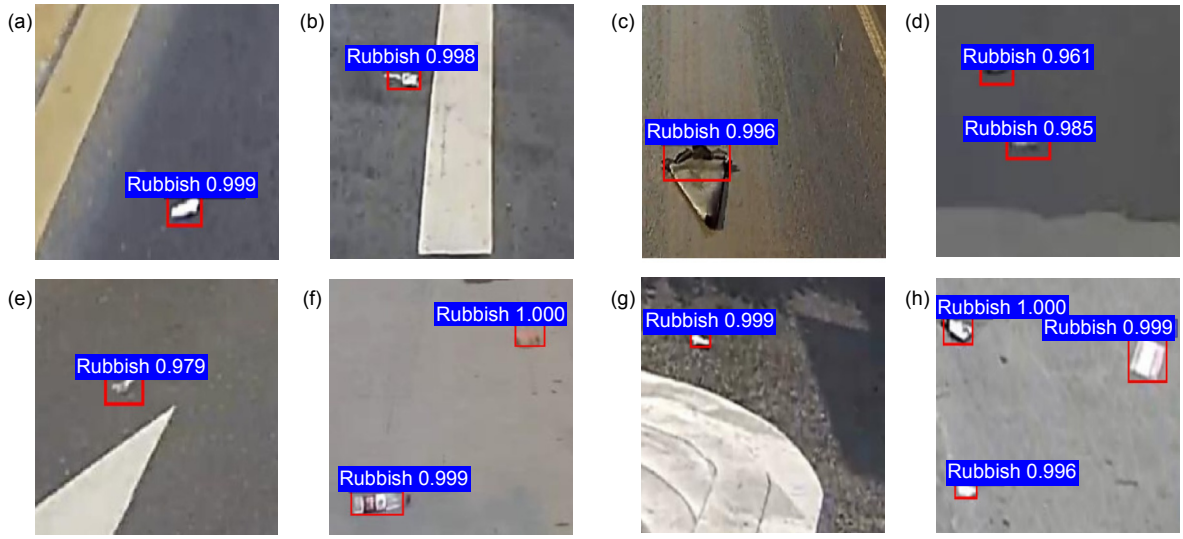


图 4　测试图像和检测结果

Fig. 4　Testing images and detection results

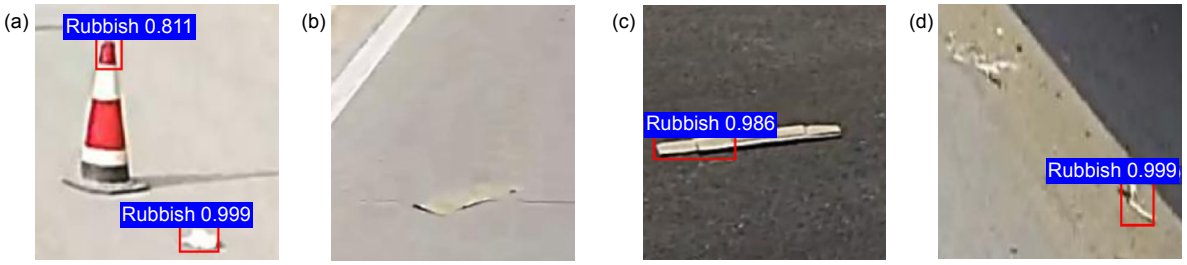


图 5　异常检测示例

Fig. 5　Examples for anomaly detection

5 结 论

本文主要基于 YOLOv3 对网络结构和对多尺度检测进行改进并结合了 M-Softer-NMS 提出了一种城市道路小像素目标检测的实时检测算法，自行设计并标注了一个城市道路上存在的小像素目标的训练、验证与测试集。实验结果表明，本文提出的新的卷积神经网络 Road_Net 对城市道路上的小像素目标检测具有很好的鲁棒性，在达到实时性检测 57.9 f/s 的同时，精度、召回率与 *AP* 分别能到达 95.29%、91.12%和 82.41%。在接下来的工作中，将继续改进网络并优化算法，以获得更高的精确性和更低的时间成本，还会继续采集更多的真实场景图像以扩大现有的数据集，以便更好地为环保部门应用。

参考文献

[1] Lowe D G. Object recognition from local scale-invariant features[C]//The *Proceedings of the 7th IEEE International Conference on Computer Vision*, 1999, **2**: 1150–1157.

[2] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. *International Journal of Computer Vision*, 2004, **60**(2): 91–110.

[3] 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*, 2005, **1**: 886–893.

[4] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**(7): 971–987.

[5] Cortes C, Vapnik V. Support-vector networks[J]. *Machine Learning*, 1995, **20**(3): 273–297.

[6] Ho T K. Random decision forests[C]//*Proceedings of the 3rd International Conference on Document Analysis and Recognition*, 1995, **1**: 278–282.

[7] Luo Z J, Zeng G Q. Space objects detection in video satellite images using improved MTI algorithm[J]. *Opto-Electronic Engineering*, 2018, **45**(8): 180048.
罗振杰, 曾国强. 基于改进 MTI 算法的视频图像空间目标检测[J]. 光电工程, 2018, **45**(8): 180048.

[8] Fan X S, Xu Z Y, Zhang J L. Dim small target tracking based on improved particle filter[J]. *Opto-Electronic Engineering*, 2018, **45**(8): 170569.
樊香所, 徐智勇, 张建林. 改进粒子滤波的弱小目标跟踪[J]. 光电工程, 2018, **45**(8): 170569.

[9] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering[C]//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 815–823.

[10] Wang X H, Gao L L, Wang P, *et al*. Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length[J]. *IEEE Transactions on Multimedia*, 2018, **20**(3): 634–644.

[11] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//*Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 580–587.

[12] Girshick R. Fast R-CNN[C]//*Proceedings of 2015 IEEE International Conference on Computer Vision*, 2015: 1440–1448.

[13] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[C]//*Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015: 91–99.

[14] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 761–769.

[15] 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*, 2016: 779–788.

[16] Uijlings J R R, Van De Sande K E A, Gevers T, *et al*. Selective search for object recognition[J]. *International Journal of Computer Vision*, 2013, **104**(2): 154–171.

[17] Zitnick C L, Dollár P. Edge boxes: locating object proposals from edges[C]//*Proceedings of the 13th European Conference on Computer Vision*, 2014: 391–405.

[18] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.

[19] Lin T Y, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 936–944.

[20] Dai W C, Jin L X, Li G N, *et al*. Real-time airplane detection algorithm in remote-sensing images based on improved YOLOv3[J]. *Opto-Electronic Engineering*, 2018, **45**(12): 180350.
戴伟聪, 金龙旭, 李国宁, 等. 遥感图像中飞机的改进 YOLOv3 实时检测算法[J]. 光电工程, 2018, **45**(12): 180350.

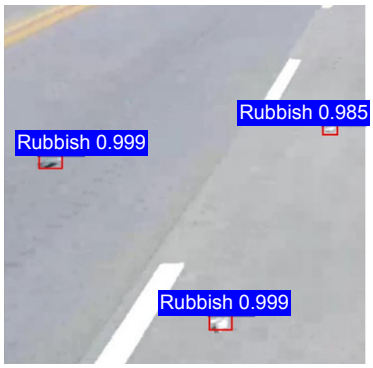
[21] Bodla N, Singh B, Chellappa R, *et al*. Soft-NMS—improving object detection with one line of code[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 5562–5570.

[22] He Y H, Zhang X Y, Savvides M, *et al*. Softer-NMS: rethinking bounding box regression for accurate object detection[J]. arXiv:1809.08545v1[cs.CV], 2018.

Object detection for small pixel in urban roads videos

Jin Yao[1,2], Zhang Rui[1,2], Yin Dong[1,2*]

[1]College of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, China;
[2]Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei, Anhui 230027, China



An illustration of detection result for small pixel targets

Overview: Small pixel target detection is a kind of difficult program. Existing object detection benchmarks and methods mainly focus on standard detection task. However, these ways cannot get good performance on low-pixel ratio object detection, which has a few pixel in high resolution images. And the early target detection frameworks such as R-CNN, YOLO series are not very good for small pixel target detection. In order to solve this problem, this paper proposes an improved YOLOv3 network and the algorithm using M-Softer-NMS to improve the detection ability of small targets. Firstly, Road_Net convolutional neural network is proposed. YOLOv3's Darknet53 network is too complicated and redundant. What's more, too many parameters will bring difficulty in training, increase the requirements on the dataset, and reduce the speed of detection, which will not achieve better real-time performance. Accuracy and real-time performance are challenging in small object detection on urban roads. Therefore, we proposed a convolutional neural network Road_Net with relatively low computational complexity as a feature extraction network. Secondly, a detection method of 4 scales is used to more fully use shallow level features. In view of the fact that the targets in this context are mostly small pixel targets, the original three scale detections are extended to four scale detections, and the larger feature maps are assigned to the smaller pixel targets with more accurate anchor frames. Finally, M-Softer-NMS algorithm is used to further improve the detection accuracy of the target in the image. Softer-NMS is further improved after Soft-NMS. A new loss function (KL Loss) for bounding box regression is proposed to learn the bounding box transformation and positional reliability at the same time. Combined with the characteristics of small pixel targets in this paper, the M-softer-NMS algorithm for this paper is proposed based on softer-NMS. In order to verify the effectiveness of the algorithm, we collected and labeled the data set named Road-garbage Dataset for the detection of small pixel target objects on the road. The Dataset is based on several main roads in a certain city and selects 1200 different main roads in different regions. The experimental results show that the accuracy, recall rate and AP can reach 95.29%, 91.12% and 82.41% respectively, while real-time detection is 57.9 f/s. In the next work, we will continue to improve the network and optimize the algorithm for higher accuracy and lower time cost, and continue to capture and use our more realistic scene images to expand our dataset for better application.

Citation: Jin Y, Zhang R, Yin D. Object detection for small pixel in urban roads videos[J]. *Opto-Electronic Engineering*, 2019, **46**(9): 190053

Supported by 2018 Anhui Key Research and Development Plan Project (1804a09020049)
* E-mail: yindong@ustc.edu.cn