基于图像的城市场景垃圾自动检测

魏书法[1,2] 程章林[1]

[1](中国科学院深圳先进技术研究院 深圳 518055)
[2](中国科学院大学深圳先进技术学院 深圳 518055)

摘 要 基于城市场景照片快速准确地自动检测垃圾在“智慧城管”等应用中具有重要的研究价值。城市垃圾在颜色纹理、几何形态上具有极大的多样性，甚至部分垃圾的认定具有一定的主观性，这给垃圾自动检测带来很大的挑战。文章提出了一种基于高速区域卷积神经网络的垃圾检测方法，通过使用数据融合、数据扩充、迁移学习等方法解决训练样本不足的问题，实现了城市场景图片中垃圾的自动、快速、准确检测。文章最后基于深圳市道路垃圾照片构建了一个包含多种形态类型垃圾的垃圾图片数据库，在该库中垃圾检测准确度高达 89.07%。

关键词 垃圾检测；深度学习；迁移学习
中图分类号 TG 156 **文献标志码** A

Image-Based Garbage Detection in Urban Scenes

WEI Shufa[1,2] CHENG Zhanglin[1]

[1](*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen* 518055, *China*)
[2](*Shenzhen Institutes of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen* 518055, *China*)

Abstract It is of great value to rapidly and accurately detect garbage from urban images in the application of intelligent city management. Garbage images are highly diverse in color texture and geometry; moreover, garbage recognition can be a subject matter, which poses great challenges to automatic detection of garbage. In this paper, a garbage detection method based on faster region-based convolutional neural networks was proposed. It can detect garbage from urban images with high accuracy by integrating techniques such as data fusion, data augmentation, and transfer learning. We have built an image database containing various types of garbage based on photographs taken from urban scenes in the Shenzhen city, showing a detection accuracy of 89.07%.

Keywords garbage detection; deep learning; transfer learning

收稿日期：2016-11-07 **修回日期**：2016-12-11
基金项目：深圳市基础研究项目(JCYJ20140901003938994、JCYJ20150401145529008)；国家自然科学基金项目(61379091、61602461)；863 项目(2015AA016401)
作者简介：魏书法，硕士，研究方向为计算机视觉与机器学习；程章林(通讯作者)，博士，副研究员，研究方向为可视计算，E-mail：zl.cheng@siat.ac.cn。

1 引 言

随着城市规模的扩大和人口的聚集，维护城市整洁，保持优美的市容环境变得尤为重要。市容环境成为城市整体形象的综合体现，直接影响着居民的生活品质和身体健康。而城市垃圾是影响市容环境的一个关键要素，城市路面的垃圾监测与整洁度评价成为城市相关管理部门(城管或环卫部门等)日常工作的重要内容。目前城市路面垃圾监测与评价主要靠专人人工巡查并拍照登记。巡查过程中需要人工定位无序丢弃垃圾的位置并进行拍照，之后整理归档并统计垃圾分布情况及相应的责任人和单位。这种人工的检测方法需要耗费大量的人力物力，并且受交通、天气、人员休假与工作时间等各种外界因素的影响。随着国内外各大城市“智慧城市”，尤其是“智慧城管”[1]工作的开展，如何提高城市市容环境监测评估的信息化、智能化水平成为一个重要的研究课题。对此我们开展了初步的研究，期望通过城市路口的摄像头和沿街拍摄的 360 度街景图像自动检测路面垃圾，并提出了一种基于实际拍摄的城市场景照片自动检测垃圾的方法。

基于图像的目标检测是视觉领域长期以来的热点研究问题，其中行人检测[2]、人脸检测[3]是研究的热点，在安防、自动驾驶、智能管理等领域有着广泛的应用。传统的检测方法常使用人工设计的算法[2,4]从图像中构造有代表性的特征，然后使用支持向量机、随机森林等机器学习方法对提取的特征进行分类。这一类方法的局限是提取的特征在代表性和鲁棒性方面都有欠缺，在光照、遮挡、尺度等变化因素的影响下检测精度会大大降低；另外传统方法多采用滑动窗口的方式从图像中提取感兴趣目标可能存在的区域，这种方法计算量大，使得算法的实时性成为瓶颈。深度学习即多层神经网络，是一种泛化能力强的机器学习算法。在 2013 年的大规模视觉识别挑战赛(ImageNet Large Scale Visual Recognition Competition，ILSVRC)[5]中，Krizhevsky 等[6]使用卷积神经网络取得远超第二名的成绩之后，深度学习成为研究的热点。在基于深度学习的目标检测方面，主要有两种思路。第一种是将感兴趣对象检测转化为回归问题[7]，即同时预测检测目标的位置、空间范围和类别；第二种是基于区域的检测方法[8]，即从图像的不同位置截取不同尺度的图像块，并对截取后的图像块进行分类，以图像块的位置和尺度作为感兴趣对象的位置和尺度，对图像块进行分类的结果作为感兴趣对象的类别。

垃圾检测相较于普通物体检测的一大难点在于垃圾的定义范围非常广，如在城市路面场景中，地面上的塑料袋是垃圾，一堆废弃的物料是垃圾，甚至丢弃在废弃停车场中的汽车也是垃圾。虽然他们都属于垃圾，但在颜色、纹理、几何形态上的差别非常大，跨越不同的物体类别。另外一大难点是，垃圾是一个带有主观性的概念，不同于行人、人脸等对象，有明确的定义。对于垃圾，同一个对象，在不同的情境下是否是垃圾，可能会有不同的判定结果。上述两大难点给基于图像的垃圾检测带来极大的挑战性。从公开的文献报道来看，目前国内外基于图像的城市垃圾自动检测工作极少。在我们的工作开展的同时，Mittal 等[9]也开展了基于深度学习的垃圾检测研究，他们使用全卷积神经网络作为主要工具，对影像中含有垃圾的区域进行分割，在他们自己的测试数据集上达到了 87.69% 的精度。该方法的优点是提取的垃圾区域边界较准确，但存在较多错误判断，将非垃圾区域判定为垃圾以及部分垃圾漏检，本文第 5 部分对 Mittal 方法与我们方法进行了详细对比。鉴于高速区域卷积神经网络(Faster Region-based Convolutional Neural Network，Faster-RCNN)方法[10]具有自动提取特征、泛化能力强以及准实时检测的特点，我们尝试引入 Faster-RCNN 方法自动从城市场景图像中

检测垃圾。本文的主要创新点和贡献体现在以下两方面：(1)提出了一种基于城市场景照片进行垃圾实时自动检测的方法，其中城市场景照片可以很容易地从目前城市路口固定的摄像头获取，或者采用车载移动摄像头采集的 360 度街景照片，这为“智慧城管”市容环境监测评估提供了可行的技术手段；(2)引入 Faster-RCNN 方法作为垃圾检测的核心算法，通过数据融合、数据扩充以及在 ImageNet 数据集[5]上对模型进行预训练的方法解决训练样本不足的问题，本文仅采用 200 张垃圾照片作为训练样本，在独立测试集上取得了 89.07% 的垃圾检测精度，且在普通台式机上达到每张照片 0.2 秒的准实时检测速度。

2　基于卷积神经网络的目标检测

2.1　卷积神经网络

卷积神经网络是一种受生物视觉皮层启发的机器学习模型。在 20 世纪 60 年代，Hubel 与 Wiesel[11]发现视皮层有简单细胞和复杂细胞两种神经元，不同的神经元负责不同的子区域，即感受野，简单细胞对其感受野内的线条式模式十分敏感，而复杂细胞感受野更大，对感受野内的模式具有局部不变性。Fukushima[12]在 1980 年提出了卷积神经网络(Convolutional Neural Network)的雏形 neocognitron(认知机)，与卷积神经网络不同的是，neocognitron 不强制要求权值共享。LeCun 等[13]在 1990 年奠定了现代卷积神经网络结构的基础。他们提出的网络结构被称为 LeNet-5，LeNet-5 由卷积层、下采样层和全连接层组成，对手写字体识别效果非常好，但由于计算资源与训练数据的限制，LeNet-5 在更加复杂的问题上表现差强人意。2006 年，Hinton 和 Salakhutdinov[14]在《Science》上指出“具有多个隐层的神经网络学习特征的能力更为优异，并且在训练上的复杂度可以通过逐层预训练来进行缓解”，重新使深度学习获得了关注。2012 年，Alex 等[6]在 ILSVRC 竞赛中在图像分类任务上采用卷积神经网络以压倒性的优势获得冠军，引发了深度学习研究的热潮。新的网络结构与应用，如 Network in Network(网络中的网络)[15]、空间金字塔采样网络(Spatial Pyramid Pooling-Net)[16]、全卷积神经网络(Fully Convolutional Network)[17]等被不断地提出并刷新在标准数据集上的表现，使得深度学习方法在视觉领域占据主导地位。

相较于传统的多层感知机，典型的卷积神经网络通常有卷积层、池化层、非线性层、Dropout 层(随机丢弃层)等 4 种不同类型的层。

(1)卷积层将输入的特征图与三维滤波器做卷积处理并输出特征图。滤波器的参数在训练过程中进行优化调整，被称为可学习参数。在网络的前向传播过程中，卷积层的滤波器与输入特征图的局部做内积，滤波器在输入特征图上做滑窗操作可以产生新的输出特征图。在位置(i, j)处的数据向量经过卷积层以后输出结果的计算公式为：

$$y_{ij}=f_{ks}\left(X_{s_i+\delta_i,s_j+\delta_j}\right)\ 0\leqslant\delta_i,\ \delta_j\leqslant k \qquad (1)$$

其中，k 表示卷积核的大小；s 为步长。

卷积层的滤波器可以看作是广义的线性模型，当实例的隐概念线性不可分时，线性模型并不能很好地进行特征组合与提取。为了增强卷积层的表征能力，林敏等[15]提出了 Network in Network，将卷积层的线性滤波器替换为微型的神经网络来增强卷积层的表征能力。之后 Szegedy 等[18]提出了 Inception module(感知模块)，对 Network in Network 做了改进。

(2)池化层对输入的特征图做下采样操作，减少了网络的参数，降低了网络的复杂度，并增强网络对输入特征图中平移和畸变的鲁棒性。常用的池化手段有最大池化和平均池化，前者输出

池化区域内的最大值，后者输出池化区域内的平均值。

(3)非线性层的引入使得卷积神经网络由广义线性模型变为非线性模型。经典的卷积神经网络使用 Sigmoid 函数(逻辑函数)作为非线性单元，但是由于 Sigmoid 函数存在梯度弥散的现象，后续 Nair 和 Hinton[19]使用了效果更好的修正线性单元(Rectified Linear Unit， ReLU)。在 ReLU 的基础上，陆续出现了 Leaky ReLU[20]、Parametric ReLU[21]等改进。ReLU 层的公式如下：

$$f(x)=\max(0, x) \tag{2}$$

(4)Dropout 层由 Hinton 等[22]提出，是防止过拟合的有效手段。在训练过程中，Dropout 层依照一定的概率随机抑制部分神经元，防止神经元之间相互依赖，并强制神经网络在缺失部分信息的情况下做出预测。在预测中，关闭 Dropout 层，能起到模型集成的作用，可以缓解过拟合现象。

2.2 基于区域的感兴趣目标检测

目标检测从给定的图像或视频帧中找出感兴趣目标的位置并给出类别信息，是计算机视觉中的基本问题，有着广泛的应用。

传统的目标检测方法主要分为三个阶段：首先从给定的图像中选择候选区域，然后从候选区域提取特征，接着使用分类器对提取后的特征进行分类。在经典的行人检测中[2]，首先使用滑动窗口在给定图像中选择不同位置和尺度候选区域，然后在候选区域中提取方向梯度直方图(Histogram of Oriented Gradient，HoG)特征，最后对候选区域的 HoG 特征使用支持向量机进行二分类。传统的目标检测方法主要有两个局限：一是使用滑动窗口作为候选区域会造成大量的重复计算；二是手工设计的特征在鲁棒性与预测性上都存在瓶颈。

自 Krizhevsky 等[6]在 2012 年的 ILSVRC 上取得令人瞩目的成绩之后，卷积神经网络出色的特征提取与抽象能力引起视觉界广泛的兴趣。在目标检测领域，Girshick 等[8]提出了基于区域的感兴趣目标检测(Region-based Convolutional Neural Network，R-CNN)。R-CNN 在传统的检测方法基础上做了两处改进：

(1)使用非监督的候选区域推荐算法代替滑动窗口方法

滑动窗口方法在召回率与计算量之间存在互斥的矛盾，为提高算法的召回率需要增加滑动窗口的密度，继而大大增加计算量。候选区域推荐算法根据图像中的纹理、边缘、颜色等信息对可能是候选区域的图像块进行推荐，在保证召回率的同时降低了冗余计算。

(2)使用卷积神经网络提取候选区域的特征

卷积神经网络能够从图像中提取相较于传统的手工设计的尺度不变特征提取算子(Scale Invariant Feature Transformation)、HoG 等特征描述子更鲁棒、预测性更好的特征，从而提高算法的准确率。

由于区域推荐算法推荐的区域存在很多重叠部分，而后续的特征提取过程会单独在每个候选区域中提取特征，因此对同一个图片块会提取多次特征，进行了大量重复计算。He 等[16]将金字塔采样思想引入了卷积神经网络中。不同尺寸的候选区域映射到卷积神经网络的特征图之后对应不同位置与尺度的矩形，使用金字塔采样可以将这些不同位置与尺度的矩形区域抽样为统一长度的向量，保证了卷积神经网络的全连接层在图像尺度发生变化时依然能正常工作。使用金字塔采样的快速区域卷积神经网络(Fast Region-based Convolutional Neural Network，Fast-RCNN)[23]首先在整个图像上用卷积神经网络提取特征，然后将不同的候选区域在特征图映射采样后的固定长度的向量输至回归器和分类器进行分类和回归操作。

Fast-RCNN 仍然需要使用非监督的方法进行候选区域推荐，成为制约速度与算法准确率的瓶颈。而 Faster-RCNN[10]使用有监督的方法进行候

选区域的推荐，首先固定选取特征图上的局部作为候选区域，然后将候选区域进行分类与位置回归，这里的分类是二分类问题，即此候选区域是否含有感兴趣的目标，如果置信度大于一定的阈值，那么将回归后的位置作为优化后的候选区域输入到 Fast-RCNN 中。Faster-RCNN 可以达到近似实时的速度。

3　基于卷积神经网络的垃圾检测

鉴于 Faster-RCNN 方法在目标检测方面的优异性能，我们采用 Faster-RCNN 算法作为垃圾自动检测的核心模块，整个算法流程如图 1 所示。

下面我们将结合垃圾检测中的实际问题对本文算法流程进行阐述。

3.1　原始训练数据与相似数据的获取

我们从深圳市城市管理相关部门获取含有垃圾的城市影像，并选取含有较多城市场景的数据集作为相似数据。关于数据的详细信息将在文章的第 4 部分给出。

3.2　数据融合

我们的垃圾检测任务，实质是对图像局部区域进行的二分类问题，即该局部区域是感兴趣的对象垃圾还是不感兴趣的对象背景。在城市场景下，对于背景类而言，其类内的多样性是非常大的，包括城市内常见的汽车、行人、交通标志、自行车等对象。考虑到垃圾没有确定的形状，在视觉上主要依靠纹理和颜色上的差异来区分垃圾与其他对象，而城市场景内的其他对象很容易具有类似的属性，如有行人穿着较宽松的红色或黑色衣服，在没有行人这一类别的先验知识条件下，很容易与垃圾混淆，从而降低检测的精度。对此，我们考虑到通过对背景类进行细分，降低背景类内的多样性，使得算法能够更加有效地区分垃圾与其他不感兴趣对象，从而提高算法的鲁棒性。我们尝试将视觉目标分类挑战赛 2007（The PASCAL Visual Object Classes Challenge 2007，VOC2007）[24] 数据集与已标注的城市场景内垃圾影像进行融合并对模型进行训练。

图 2 左边图片展示了由于颜色的相似，算法将在室外晾晒的寝具棉被误识别为垃圾的例子。右边图片展示了一个将纹理和色彩与典型垃圾较为相似的行人误分为垃圾的例子。如果我们在感兴趣目标类中加入行人，该对象将优先被归类为

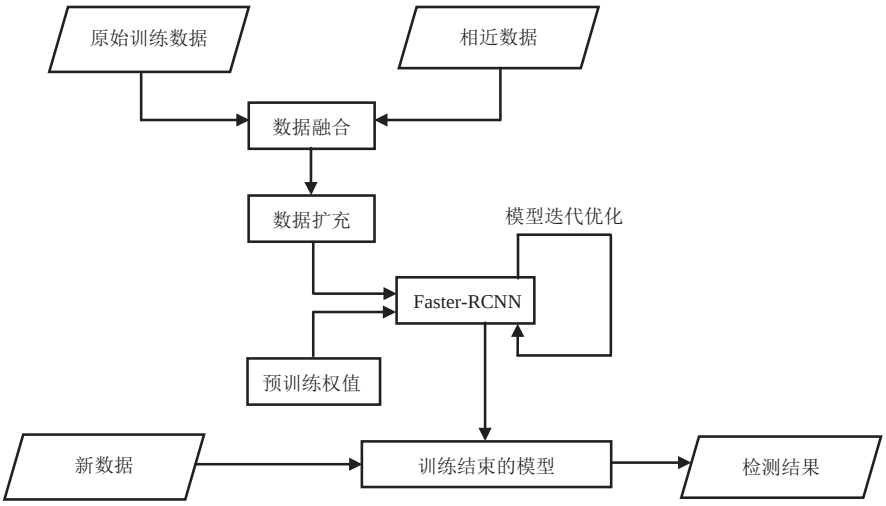

图 1　本文采用的算法流程

Fig. 1　Pipeline of garbage detection

行人。

图 3 展示了 VOC2007 数据集中的典型图片。其中，第一行从左往右图像中的主体分别是汽车和自行车，第二行从左往右图像中的主体分别是巴士和行人。将这些数据与训练数据融合，可以提高算法的泛化能力。

3.3 数据扩充

由于数据来源和标注成本的限制，我们训练数据的量并不足以用来训练一个实用的深度学习模型，而数据扩充可以在一定程度上缓解这个问题。我们将在本文第 4 部分详细叙述。

3.4 预训练权值与迁移学习

为了构造一个泛化能力强、精度高的机器学习模型，使用的数据往往要满足两个条件：首先是训练数据与测试数据独立同分布；其次是有足够多的训练样本。在实际的应用中，第二个条件往往很难满足，有时候样本很难获取，有时候对样本的标注成本很高。如何能运用已有的知识对

图 2 错误识别的例子

Fig. 2 An example of incorrect detection

图 3 VOC2007 数据集中的代表图片

Fig. 3 Typical images in VOC2007 dataset

不同但相关领域的问题进行求解成为新的研究热点，这也正是迁移学习的研究核心。迁移学习放松了传统机器学习的两个基本假设，尝试使用已经学习到的知识配合少量的领域样本甚至不使用该领域的样本来完成特定领域的机器学习问题。从使用的技术方法来看，可以将迁移学习分为基于特征选择的方法、基于权重迁移的方法和基于特征映射的方法；而从源领域和目标领域的样本是否标注以及学习目标的异同性来看，可以把迁移学习分为直推式迁移学习、归纳式迁移学习和无监督迁移学习等。当前视觉领域主流的机器学习方法是基于高性能并行计算和海量训练数据的深度学习算法，并且几乎在所有的视觉任务上都显著超越了其他方法。由于深度学习模型的高度复杂性，十分容易造成过拟合，因此使用少量的训练样本进行深度学习模型的有效训练十分困难。Girshick 等[8]指出，对深度学习模型使用有监督地预训练，然后在特定领域的任务上进行调优是一种十分有效的迁移学习方法。具体的做法是，在 ImageNet[6]数据集上对模型进行分类任务的预训练，用预训练得到的模型参数权值作为领域任务模型的初始值。在本文中，我们将使用预训练的结果作为模型的初值，以解决训练样本不足的问题。

3.5　模型的训练与测试

我们使用卷积结构特征快速镶嵌（Convolutional Architecture for Fast Feature Embedding，CAFFE）[25]作为基本框架，搭建 Faster-RCNN。在准备好训练数据之后，我们使用蔡勒&福格斯网络（Zeiler & Fergus-Net，ZF-Net）[26]作为网络的基本结构，并添加区域推荐层、金字塔采样层等作为拓展。我们使用网格搜索来确定最优的先验参数，具体的配置将在第 4 部分给出。

图 4 给出了基于 ZF-Net 的 Faster-RCNN 的网络结构。由于使用金字塔采样层对最后一层卷积层的特征图进行采样，理论上网络可以输入任意尺寸的图像。第 1 层使用 96 个尺寸为 7×7 的滤波器以 2 的间隔在输入图像上进行卷积操作并得到输出特征图（第 1 层上部），输出特征图连接 ReLU 非线性层，再经过尺寸为 3×3、间隔为 2 的最大池化层。第 2 层使用 256 个尺寸为 5×5 的滤波器以 2 的间隔在第 1 层的输出上做卷积操作，后经过 ReLU 层和最大池化层得到第 2 层的输出（第 2 层下部）。第 3 层和第 4 层的结构与前面两层类似。第 5 层的输出作为区域推荐网络的输入，区域推荐网络给出可能存在感兴趣目标区域的位置与置信程度，并将其作为金字塔采样层的输入，之后采样层根据推荐的区域从第 5 层的特征图中进行采样，并将采样结果输入区域分类器和区域位置回归器。其中，分类器给出推荐候选区域的类别，位置回归器给出候选区域内感兴趣对象修正后的位置。

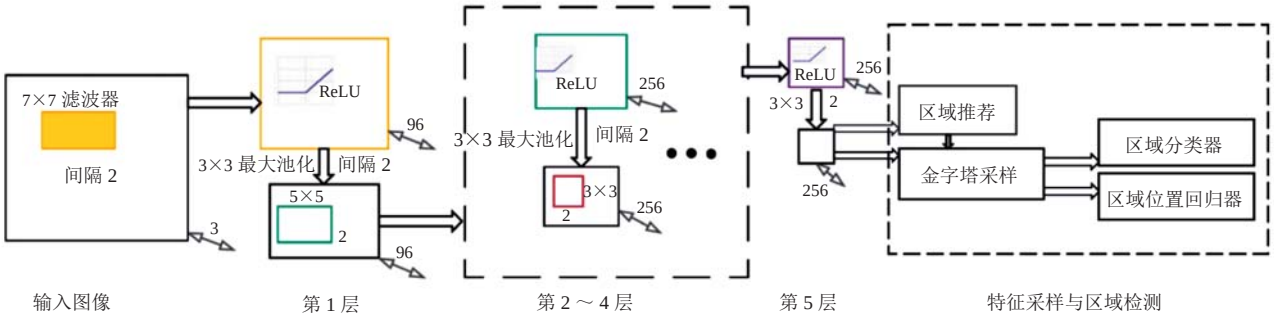

图 4　基于 ZF-Net 的高速区域卷积神经网络

Fig. 4　ZF-Net based faster region-based convolutional neural network

4 实验设置

4.1 数据集

我们采集了 372 张含有垃圾的城市影像数据，图片的平均尺寸为 420×400 像素，并将采集到的数据划分为训练集、验证集和测试集。其中训练集有 200 张影像，验证集有 72 张影像，测试集有 100 张影像。具体实验数据汇总如表 1 所示。图 5 为典型的含有垃圾的城市影像数据。

4.2 数据标定

我们按照 VOC2007 数据集规定的格式对感兴趣目标进行标注。感兴趣的目标使用矩形包围盒进行表示，并将矩形包围盒的位置作为预测的目标。

4.3 数据融合与扩充

为了提高模型的泛化能力，我们对数据进行融合与扩充。图 6 描述了数据融合与扩充的过程，我们使用 VOC2007 数据集中的影像数据与采集到的城市影像数据进行融合，并进行扩充，组成训练用的数据集。

(1) VOC2007 数据的融合

如果仅仅将垃圾作为感兴趣目标进行检测，实质上我们对每个推荐的区域进行的是一个二分类的问题，即此区域是不是垃圾。所有的其

表 1 实验数据概述

Table 1 Overview of our data

总数量（张）	训练集数量（张）	验证集数量（张）	测试集数量（张）	图片平均尺寸（像素）
372	200	72	100	420×400

图 5 含有垃圾的城市影像

Fig. 5 Images of urban scene that contains garbage

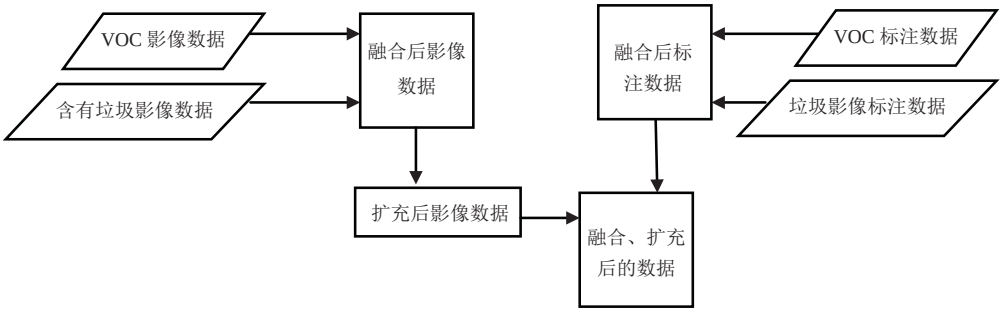

图 6 数据融合与扩充的流程

Fig. 6 Pipeline of data merge and augmentation

他类别，包括差别很大的行人、汽车、自行车等对象，都被归为背景类，这样将导致背景类的类内差距很大。VOC2007 数据集共有 5 000 张训练验证图片和 5 000 张测试图片，共有 20 类对象，其中含有城市场景中常见的对象，如行人、自行车、汽车、巴士、摩托车等。如果我们将 VOC2007 数据集与含有垃圾的城市场景数据进行融合后作为训练数据，在不增加数据标注量的同时，可以降低背景类的差异，从而增强算法的鲁棒性。数据融合的另外一个好处是可以将背景中容易与感兴趣目标混淆的部分进行区分。在我们的任务中，感兴趣的影像中的垃圾与其他对象的主要区别在颜色与纹理这两个特征上，而垃圾形状的变化非常大，容易产生误判。如当一位行人的衣服色彩变化和纹理特征与典型的垃圾的颜色变化和纹理特征比较相似，那么算法可能会将行人的衣服误分为是垃圾。但当我们引入行人作为感兴趣对象之后，由于对推荐区域的分类是互斥的，那么该区域也会以一定概率被分类为行人，从而降低该区域被分类为垃圾的概率，进而提高算法的鲁棒性。我们将在后续实验中对数据融合的作用进行验证。

(2) 融合后数据的扩充

限于客观原因，我们获取的训练数据有限，为了解决这个问题，我们使用了旋转、裁剪、颜色扰动等常见的数据扩充方法。经过扩充，能够将训练数据的数量增加 10 倍左右。在图 7 中，最左侧的图片是原始图片，之后依次是旋转、裁剪、颜色扰动后的效果。

4.4　模型的选择与模型参数的初始化

由于受硬件条件限制，我们选择 ZF-Net 作为提取图像特征的特征提取器。训练深度学习模型需要大量的数据与长时间的迭代，而我们能够获取的样本数量有限，因此使用在 ImageNet 分类任务上预训练好的模型的权值作为垃圾检测模型的权值初值，并在此基础上进行微调。

4.5　模型的先验参数设置

我们使用分阶段迭代优化的方法对模型进行训练。第一个阶段，首先训练区域推荐网络，在使用预训练的权值对模型进行初始化之后，对区域推荐网络进行端到端的训练。第二个阶段，使用第一个阶段训练的区域推荐网络的输出结果来独立训练用于检测感兴趣目标的 Fast-RCNN。第三个阶段，使用第二个阶段训练好的模型作为区域推荐网络的初始权值，再一次进行区域推荐网络的训练。最后一个阶段，固定第三个阶段训练得到的卷积层模型，仅对 Fast-RCNN 的全连接层进行微调。

我们使用网格搜索的方法来确定适用于该问题的先验参数。其中，候选的学习率分别为 0.000 8、0.001、0.012，候选的权值衰减量分别为 0.000 3、0.000 5 和 0.000 7，候选的冲量分别为 0.7、0.9 和 1.1。我们枚举候选学习率、候选权值衰减量、候选冲量的组合并依次作为模型的先验参数，记录不同先验参数下模型在验证数据上的检测精度，并选择精度最高的一组作为模型

图 7　经过扩充后的图片数据

Fig. 7　Images augmentation

的最终参数。最终，我们在该问题中选择学习率为 0.001、权值衰减量为 0.000 5 和冲量为 0.9。

5 实验结果

5.1 使用数据融合与不使用数据融合的结果对比

我们通过随机选取图片来构造不同的训练集、验证集与测试集，进行 10 次实验，并求取多次实验结果的均值与方差作为最终结果(表 2)。从表 2 可以看出，在对数据进行扩充融合之前，在原始的数据集上，平均检测精度可以达到 85.05%，方差为 0.096。将原始数据集与 VOC2007 的数据进行融合扩充之后，对垃圾的平均检测精度达到了 89.07%，方差为 0.049，相较之前提升了 4.02%，并且对两组数据进行 t 检验后 $P<0.05$，证明多元数据的融合对提高算法的鲁棒性具有一定作用。另外，本文算法在 Nvidia GTX980 GPU 上速度达到了每张 0.2 秒的速度，可以做到准实时的垃圾检测。

5.2 检测结果示例

图 8 显示了我们的检测结果，被识别为垃圾的对象被标注在矩形框之内，矩形框之上的数值代表算法对该区域确定是感兴趣对象的置信度。可以看出，对于场景中不同类型不同尺度大小的垃圾，本文算法都能够以很高的置信度标示出来。

表 2 实验结果概述

Table 2 Overview of the experimental results

测试图片数量	检测速度	数据融合之前		数据融合之后	
（张）	（秒/张）	平均检测精度（%）	检测精度方差	平均检测精度（%）	检测精度方差
100	0.2	85.05	0.096	89.07	0.049

图 8 检测结果

Fig. 8 Results of detection

5.3　与其他方法的对比

在 Mittal 的方法[9]中，垃圾检测的精度在他们构造的图像垃圾数据集(Garbage In Images，GINI)上达到了 87.69%。由于其使用的算法数据标注格式与我们的标注格式不一致，并且没有提供其算法的开源代码。因此我们采用视觉对比的方法来对两种算法的效果进行评价，具体为：从我们的数据集中选取典型的场景，同时使用本文方法和 Mittal 方法[9]进行检测，并对检测结果进行对比；从 GINI 数据集中选取典型的场景，同时使用本文方法和 Mittal 方法[9]进行检测，并对检测结果进行对比。我们的训练数据只有 200 张，对方的训练数据有 2 558 张，而在典型的城市场景上本文算法表现要优于对方。两种方法的具体对比结果如表 3 所示。

图 9 中第一行和第二行的图片是数据集 GINI 中的图片。其中，第一行是 Mittal 算法[9]检测的结果，第二行是本文算法检测的结果。可以看出，本文算法除了第一张图片中将一辆较小的汽车以较小的置信度误检为垃圾之外，并没有发生其他误检，并且定位较为准确。第三行和第四行的图片是我们的数据集中的图片。其中，第三行是使用 Mittal 方法[9]进行检测的结果，在第一张和第三张图片中都有误检产生，另外在垃圾定

表 3　两种方法的结果对比

Table 3　Results Compare of two methods

方法	训练集数量（张）	测试集数量（张）	在测试集上检测精度（%）
Mittal[9]	2 558	620	87.69
本文方法	200	100	89.07

图 9　两种算法的对比

Fig. 9　Results of two methods

位的准确度方面，本文方法的表现也更优秀。第四行是使用本文算法检测的结果。

5.4 方法的局限性

尽管在测试数据集上本文方法取得了较高的精度，但是，该方法仍有一些局限性。

(1) 由于垃圾的类内变化非常之大，并且在垃圾判别上也存在一定的主观性，导致本文方法还无法正确处理这种带有一定主观性的垃圾检测。在特定的场景，如在废旧汽车回收厂内，我们会认定破旧的汽车是垃圾，但很明显本文方法会倾向于将其分类为汽车。

图 10 展示了垃圾变化的多样性。图 11 是在测试过程中漏检的图片。由于垃圾定义的主观性，第一行第二张图片和第三张图片在其他情境下可能不会被认定为垃圾，第一行第一张图片和第二行第三张图片由于目标过小而发生漏检。

(2) 测试照片与训练集中照片的采集方式差异较大时，本文方法在垃圾检测精度上会有所下降。

图 10 变化多样的垃圾

Fig. 10 The variance of garbage

图 11 漏检的样例

Fig. 11 Examples of failed detection

图 12　在街景图像中的垃圾检测

Fig. 12　Garbage detection in street view image

在图 12 中，我们截取百度街景照片作为测试图像。由于采集设备的不同，街景图像与训练集中的图像在尺度、色彩分布方面有显著的差异。在图 12 左边的图像中，本文方法没有检测出左下角的垃圾，但在将其进行局部缩放之后，本文方法成功地检测出了垃圾。由于训练数据来源的单一性，本文方法在泛化能力上存在提升空间。

6　讨　论

本文使用基于深度学习的 Faster-RCNN 方法，并利用数据融合、样本扩充和迁移学习等多种手段，实现了小样本条件下城市场景图像中垃圾的快速、高精度检测。本文方法在普通台式机上可以达到每张 0.2 秒的准实时垃圾检测速度，检测精度达到 89.07%，可以应用于城市场景中垃圾的自动检测，极大地降低人力成本，具有较高的实用价值。当前的方案也还存在一定的局限性。首先是检测精度仍有待提升的空间，当前的精度并不能完全代替人工手动分类，在要求严格的情况下仍然需要人工对处理后的数据进行校验。另外在算法的泛化性能上，对于与训练集中照片在尺度与图像质量上相差较大的测试样本，仍然存在检测精度下降的问题。由于垃圾的类内差别很大以及存在一定的主观性判别问题，这给算法的泛化能力提出了很高的挑战。

7　结　论

本文采用基于深度学习的目标检测方法自动检测城市场景中的垃圾，并针对数据获取困难、数据标注成本高等问题，使用了数据扩充、数据融合、迁移学习等方法，最终实现城市场景中垃圾的快速高精度检测，具有较高的实用价值。目前该方法对于训练集中不同光照、不同采集设备的测试图片还存在泛化能力不够的问题。未来，我们将尝试在改进网络结构、增加训练数据的数量与来源、二维与三维信息相结合等方面对算法进行改进。

参 考 文 献

[1] 宋刚. 从数字城管到智慧城管: 创新 2.0 视野下的城市管理创新 [J]. 城市管理与科技, 2012, 14(6): 11-14.

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

[3] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features [C] // Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001: 518.

[4] Lowe DG. Object recognition from local scale-invariant features [C] // Proceedings of the International Conference on Computer Vision, 1999: 1150-1157.

[5] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[6] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks [C] // Advances in Neural Information Processing Systems, 2012: 1097-1105.

[7] Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection [C] // Advances in Neural Information Processing Systems, 2013: 2553-2561.

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

[9] Mittal G, Yagnik KB, Garg M, et al. SpotGarbage: smartphone app to detect garbage using deep learning [C] // Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 2016: 940-945.

[10] Ren S, He K, 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, 2015: 91-99.

[11] Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex [J]. The Journal of Physiology, 1968, 195(1): 215-243.

[12] Fukushima K. A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position [J]. Biological Cybernetics, 1980: 193-202.

[13] LeCun Y, Boser B, Denker JS, et al. Handwritten digit recognition with a back-propagation network [M] // Advances in Neural Information Processing Systems 2, 1990: 396-404.

[14] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507.

[15] Lin M, Chen Q, Yan SC. Network in network [J]. arXiv, 2013.

[16] He KM, Zhang XY, Ren SQ, et al. Spatial Pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9): 1904-1916.

[17] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, 79(10): 1337-1342.

[18] Szegedy C, Liu W, Jia YQ, et al. Going deeper with convolutions [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[19] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines [C] // Proceedings of the 27th International Conference on Machine Learning, 2010: 807-814.

[20] Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic models [C] // Proceedings of the 30th International Conference on Machine Learning, 2013: 1-30.

[21] He KM, Zhang XY, Ren SQ, et al. Delving deep into rectifiers: surpassing Human-level performance on imageNet classification [C] // Proceedings of the IEEE International Conference on Computer Vision, 2015: 1026-1034.

[22] Hinton GE, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3(4): 212-223.

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

[24] Everingham M, Luc V, Christopher K, et al. The pascal visual object classes (VOC) challenge [J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[25] Jia YQ, Shelhamer E, DonahueJ, et al. Caffe: convolutional architecture for fast feature embedding [C] // Proceedings of the 22nd ACM International Conference on Multimedia, 2014: 675-678.

[26] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks [C] // European Conference on Computer Vision, 2013: 818-833.