

文章编号 1004-924X(2020)05-1152-13

# 基于深度卷积神经网络的目标检测研究综述

范丽丽<sup>1</sup>, 赵宏伟<sup>1,2</sup>, 赵浩宇<sup>3\*</sup>, 胡黄水<sup>4</sup>, 王 振<sup>5</sup>

- (1. 吉林大学 计算机科学与技术学院, 吉林 长春 130012;
2. 吉林大学 符号计算与知识工程教育部重点实验室, 吉林 长春 130012;
3. 吉林大学 学报(工学版)编辑部, 吉林 长春 130012;
4. 长春工业大学 计算机科学与工程学院, 吉林 长春 130012;
5. 山东理工大学 计算机科学与技术学院, 山东 淄博 255000)

**摘要:**作为计算机视觉中的基本视觉识别问题,目标检测在过去的几十年中得到了广泛地研究。目标检测旨在给定图像中找到具有准确定位的特定对象,并为每个对象分配一个对应的标签。近年来,深度卷积神经网络 DCNN(Deep Convolutional Neural Networks)凭借其特征学习和迁移学习的强大能力在图像分类方面取得了一系列突破,在目标检测方面,它越来越受到人们的重视。因此,如何将 CNN 应用于目标检测并获得更好的性能是一项重要的研究。首先回顾和介绍了几类经典的目标检测算法;然后将深度学习算法的产生过程作为切入点,以系统的方式全面概述了各种目标检测方法;最后针对目标检测和深度学习算法面临的重大挑战,讨论了一些未来的方向,以促进深度学习对目标检测的研究。

**关键词:**图像处理;深度卷积神经网络;目标检测;特征表示;深度学习

中图分类号:TP391 文献标识码:A doi:10.3788/OPE.20202805.1152

## Survey of target detection based on deep convolutional neural networks

FAN Li-li<sup>1</sup>, ZHAO Hong-wei<sup>1,2</sup>, ZHAO Hao-yu<sup>3\*</sup>, HU Huang-shui<sup>4</sup>, WANG Zhen<sup>5</sup>

- (1. College of Computer Science and Technology, Jilin University, Changchun 130012, China;
  2. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China;
  3. Editorial Department of Journal (Engineering and Technology Edition), Jilin University, Changchun 130012, China;
  4. School of Computer Science and Engineering, Changchun University of Technology, Changchun 130012, China;
  5. School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China)
- \* Corresponding author, E-mail: zhaohaoyu@jlu.edu.cn

收稿日期:2019-10-30;修订日期:2019-12-27.

基金项目:国家自然科学基金资助项目(No. 61841602);吉林省省级科技创新专项资助项目(No. 20190302026GX);吉林省高等教育学会高教科研项目资助(No. JGJX2018D10);中央高校基本科研业务基金资助项目(JLU)

**Abstract:** Object detection, which is a fundamental visual recognition problem in computer vision, has been extensively studied in the past few decades and has become one of the popular research areas in the world. The aim of object detection is to accurately locate specific objects in a given image and assign a corresponding label to each object. In recent years, Deep Convolutional Neural Networks (DCNN) have been used in a series of developments in object detection and image classification owing to their powerful capabilities of feature learning and transfer learning. It has garnered considerable attention in the field of computer vision for object detection. Therefore, the method of applying CNN in target detection to obtain better performance is an important topic for research. First, we reviewed and introduced several types of classic object detection algorithms. Next, we considered the generation process of the deep learning algorithm as a starting point, analyzed the technical ideas and key problems of DCNN in the application of target detection, and provided a comprehensive overview of various target detection methods in a systematic manner. Finally, in view of the major challenges in target detection and deep learning algorithms, we provided future development scope and direction to promote the study of target detection using deep learning.

**Key words:** image processing; deep convolutional neural network; object detection; feature representation; deep learning

## 1 引言

随着世界经济的飞速发展,目标检测已经成为社会安全、公共交通和国防军事等各个领域的基础性研究课题之一,在机器人导航<sup>[1]</sup>、航空航天<sup>[2]</sup>、工业检测<sup>[3-4]</sup>、行人追踪<sup>[5]</sup>和军事应用<sup>[6-7]</sup>等领域应用广泛。所谓的目标检测,是通过分析目标的几何特征,从图像或视频中定位感兴趣的目標,准确地判断每个目标的具体类别,并给出每个目标的边界框。然而,目标所处的环境缤纷复杂、目标本身通常具有不同的形态以及光亮、颜色和遮挡等多种因素的影响使目标检测在近年来受到广泛关注。

传统的目标检测的方法主要包括区域选择、特征提取和分类三个阶段。

(1) 区域选择是定位目标位置。在定位的过程中,因为不能确定检测目标的大小、长宽比以及位置,因此使用多尺度滑动窗口,以便尽可能找到目标的所有可能位置;

(2) 采用 SIFT (Scale-Invariant Feature Transform)<sup>[8]</sup>, Strip<sup>[9]</sup>, Haar<sup>[10]</sup> 等能提高特征的抗变形能力和表达能力的方法进行特征提取;

(3) 利用 AdaBoost<sup>[11]</sup> 和 DPM (Deformable Parts Model)<sup>[12]</sup> 等方法通过提高分类器的准确度和速度进行特征分类。

尽管传统的目标检测算法广泛应用于各个领域,但是其缺点引起人们的关注:

(1) 众多的候选窗口,不仅会产生很大的计算花销,并且还会产生许多多余的窗口;

(2) 语义鸿沟无法通过人工手动设计的描述符和经过判别训练的浅层模型组合来弥补;

(3) 分类训练效果受到提取特征的影响,大量的特征维数,不仅影响训练效果,还会产生“维数灾难”等问题;

(4) 算法可迁移性差,并且对特定任务的依赖性高,当检测目标发生较大改变时,需要设计新的算法。

为了能让计算机模拟人类提取特征的过程而让计算机更自然地获取图像信息,人们进行不断地研究和尝试,2006年,深度学习被 Hinton 提出,他利用深度神经网络从大量数据中自动地学习高层语义特征。深度学习强大的表征能力以及丰富的特征表示使目标检测的准确度获得了很大提升,推动了目标检测的发展。

卷积神经网络不仅能够提取高层特征,提高特征的表达能力,还能够将特征提取、特征选择和特征分类融合在同一个模型中,通过端到端的训练,从整体上进行功能优化,增强特征的可分性。

2014年, Ross B. Girshick 将检测抽象为两个过程,一是基于图片提出若干可能包含目标的区域(即图片的局部裁剪),二是在提出的这些区

域上运行 Alexnet 分类网络,得到目标的类别。这就是检测任务转化为区域上的分类任务的 R-CNN(Convolutional Neural Networks)框架,它在目标检测方面取得重大突破,影响着检测任务上的深度模型革命。基于卷积神经网络的目标检测成为计算机视觉领域的研究热点。

## 2 深度卷积神经网络概述

卷积神经网络是一种常用的深度学习网络框架。1959 年,Hubel & Wiesel<sup>[13]</sup>发现视觉系统中的可视皮层处理信息是分级处理的。20 世纪 90 年代,LeCun 等人<sup>[14]</sup>建立了 CNN 的现代结构,并对其进行了改进。他们设计了一种可以对手写数字进行分类的 LeNet-5 网络。与其他神经网络相同,也能使用反向传播算法<sup>[15]</sup>对数据进行训练。

卷积神经网络可以获取原始图像的有效表示,这使得它能够通过很少的预处理直接从原始像素识别视觉之上的规则。但是,训练数据的不足限制了计算机的计算能力,从而制约了 LeNet-5 在处理复杂问题方面的能力。之后,设计了多种方法来克服卷积神经网络训练深度的困难。其中 AlexNet<sup>[16]</sup>框架的提出促进了图像识别领域的发展,该框架类似于 LeNet-5,但层次结构更深。同时使用了非线性激活函数 ReLu<sup>[17]</sup>与 Dropout<sup>[18]</sup>方法,取得了卓越的效果。

AlexNet 之后, ZFNet<sup>[19]</sup>, VGGNet<sup>[20]</sup>, GoogleNet<sup>[21]</sup>和 ResNet<sup>[22]</sup>等方法相继被提出。从结构上讲,CNN 发展的一个方向是增加层数。通过增加深度,可以利用增加的非线性得到目标函数的近似结构,同时得到更好的特征。然而,这也使网络的整体复杂性和优化难度增加,并且会产生过度拟合现象。

自从 2012 年以来,多种改进 CNN 的方法被提出。ZFNet 对 AlexNet 的改进首先通过将第一层的卷积核尺寸从  $11 \times 11$  降为  $7 \times 7$ ,同时将卷积时的步长从 4 降至 2。这样扩张中间的卷积层可以获取到更多的信息。VGGNet 将网络的深度扩展到了 19 层,并且每个卷积层使用了  $3 \times 3$  这种小尺寸的卷积核,结果证明深度对网络性能有着重要影响。GoogleNet 同时增加了网络的

宽度与深度,相比于更窄更浅的网络,在计算量没有增加的同时,网络性能明显增强。

## 3 通用的目标检测方法

通用的目标检测算法旨在任意一张图片中定位和分类已有对象,并用矩形框标记,最后显示置信度。其主要包括两种类型,一种是基于锚点的目标检测,一种是基于 Anchor-free 的检测,其中基于锚点的目标检测分为 Two-stage 检测模型和 One-stage 检测模型,文中将进行一一介绍。

### 3.1 基于 Two-stage 的检测模型

#### 3.1.1 R-CNN

R-CNN<sup>[23]</sup>不仅能够提高候选边界框质量,还能提取高级特征,其工作可以分为三个阶段:生成候选区域、基于 CNN 的特征提取和分类与定位。R-CNN 采用选择搜索<sup>[24]</sup>为每个图片产生 2k 个区域提议,选择性搜索方法依赖于简单的自下而上的分组和显著性提示,从而快速提供任意大小的更准确的候选框,并有效减少了搜索空间。在分类上,R-CNN 使用区域提议进行评分,然后通过边界框回归调整评分区域,并用非极大值抑制进行过滤以产生用于保留目标位置的最终边界框。流程图如图 1 所示。

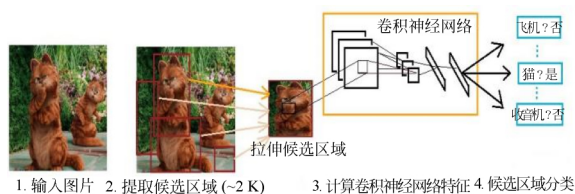


图 1 R-CNN 的网络结构

Fig. 1 Network structure of R-CNN

尽管 R-CNN 在准确程度上相比于传统方法有所提高,并且将 CNN 应用于实际目标检测中具有重要意义,但是仍然存在一些缺点。

(1)全连接层的存在使得输入图像的大小固定,直接导致每个评估区域的整个 CNN 都需要重新计算,花费大量的测试时间;

(2)R-CNN 的训练是分为多阶段完成的。首先,对提议的卷积网络进行了微调。然后,将通过微调学习的 softmax 分类器替换为 SVM。最后,训练边界框回归器;

(3)训练阶段时间长,占用空间大。从不同区域提议提取并存储在磁盘上的特征占用很大的内存。

(4)尽管选择性搜索可以生成具有较高召回率的区域提议,但是获得的区域仍然存在多余的部分,并且这个过程需要耗费很长时间。

为了解决这些问题,现在已经提出了许多方法。DeepBox<sup>[25]</sup>和SharpMask<sup>[26]</sup>试图重新排序或者完善提取特征前的区域提议以去除冗余信息,获得少数的有价值的建议。此外,还有一些改进可以解决定位不准确的问题。Saurabh Gupta等改进了具有语义丰富的图像和深度特征的RGB-D图像的目标检测<sup>[27]</sup>。目标检测器和超像素分类框架的结合在语义场景分割任务上取得了可喜的成果。

### 3.1.2 SPP-net

为了解决因为目标比例发生变化造成的目标丢失或失真问题,He等人考虑到空间金字塔匹配(SPM)<sup>[28]</sup>的理论,并提出了一种名为SPP-net<sup>[29]</sup>的新型CNN体系结构。SPM需要更精细的尺度才能将图像划分为多个部分,并将量化的局部特征聚合为中级表示形式。

目标检测中SPP-net的结构如图2所示,与R-CNN不同,目标检测中SPP-net重新使用了第五层卷积层(Conv5)的特征图,将任意大小的区域提议投影成固定长度的特征向量。这些特征图的可重复使用的可行性不仅涉及局部响应的增强,而且与其空间位置也有关系。SPP-net不

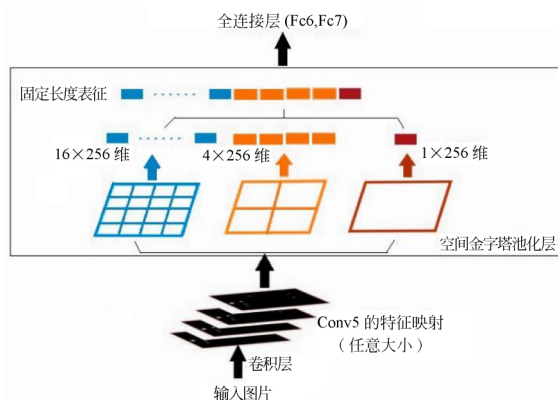


图2 卷积层和全联接层的SPP层

Fig. 2 SPP layer between convolutional layers and fully-connected layers

仅可以通过估计不同区域提议的大小来获得更好的结果,而且还可以通过共享不同区域之间的SPP层,提前计算成本来提高测试期间的效率。

### 3.1.3 Fast-R-CNN

与R-CNN相比,尽管SPP-net网络在准确性和效率上都提高了很多,但是仍然存在一些问题。SPP-net采用了跟R-CNN几乎相同的多级操作,需要额外的存储空间。除此之外,在SPP层之前的卷积层不能使用微调算法进行更新参数<sup>[29]</sup>,这些都会导致深度网络的准确性下降。

针对以上问题,Girshick<sup>[30]</sup>引入了多任务损失,并提出了一种新颖的名为Fast R-CNN的卷积神经网络。Fast R-CNN在卷积的最后一层采用感兴趣区域(Region of Interests, ROI)池化生成固定尺寸的特征图,另外,使用多任务损失函数,将边界回归加入到CNN网络中进行训练。Fast R-CNN的体系结构如图3所示,体系整个图像都经过卷积层处理后生成特征图。然后,从每个感兴趣区域的区域提议中提取一个固定长度的特征向量。感兴趣区域层是特殊的SPP层,它只有一个金字塔。它将每个特征向量送入一系列全连接层,最后分为两个同级输出层。一个输出层产生所有 $C+1$ 类( $C$ 目标类加上一个“背景”类)的softmax概率,另一输出层用四个数值对编码的边界框的位置进行编码。这些过程中的所有参数(生成区域提议的除外)都通过端到端的多任务损失函数进行了优化。

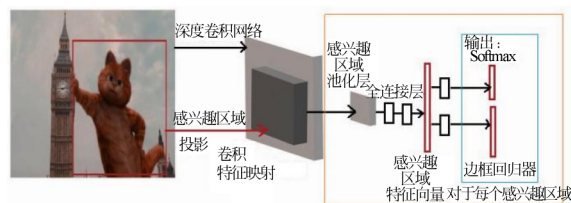


图3 Fast-R-CNN的网络结构

Fig. 3 Network structure of Fast-R-CNN

为了加快Fast R-CNN的运行速度,可以采用两个方式。第一,如果训练样本来自不同的图像,Fast R-CNN可以对小批样本进行分层采样,即首先随机采样 $N$ 个图像,然后在每个图像中采样 $R/N$ 个感兴趣区域,其中 $R$ 代表感兴趣区域的数量,在向前和向后传播中从同一图像共享计

算和内存。第二,在前向通道<sup>[29]</sup>中,计算 FC 层花费了大量时间,SVD<sup>[31]</sup>可用于压缩较大的 FC 层并加快测试过程。在 Fast R-CNN 中,无论区域提议如何产生,所有网络层的训练都可以在一个多阶段的损失下进行单阶段处理。这样可以节省更多的存储空间,并通过更合理的训练方式提高准确性和效率。

### 3.1.4 Faster-R-CNN

在目标检测中的特征提取和分类上,Fast R-CNN 已经很好地解决了很多问题,但是最新的目标检测网络仍然依赖于选择性搜索算法和 Edgebox 来生成候选区域。区域提议的计算也影响效率的提高。为了解决这个问题,Ren 等人引入了附加的区域提议网络 (Region Proposal Network, RPN)<sup>[32]</sup>,如图 4 所示,网络通过与检测网络共享全卷积特征,以几乎无代价的方式进行工作。区域提议网络是通过全卷积网络实现的,该网络具有同时预测每个位置上的目标边界和出现概率的能力。类似于<sup>[24]</sup>,区域提议网络将任意大小的图像生成一组矩形目标建议。区域提议网络在特定的卷积层上运行,与之前的层共享目标检测网络。

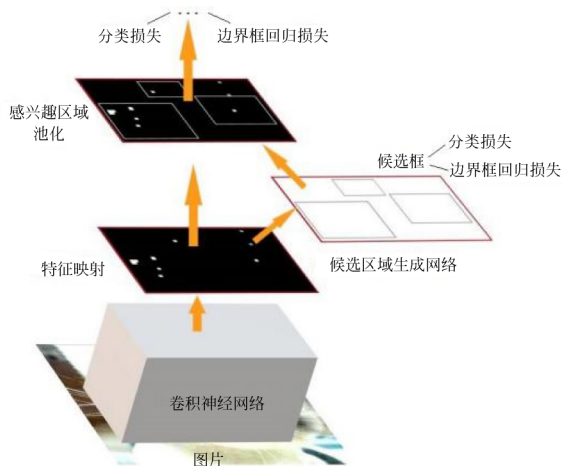


图 4 Faster R-CNN 的网络结构

Fig. 4 Network structure of faster R-CNN

在 Faster R-CNN 的作用下,基于区域提议的 CNN 体系结构可以以端到端的方式训练。此外,在 PASCAL VOC 2007 上,最新的目标检测精度,可以在 GPU 上达到 5 FPS(frame/s)的帧率。但是,更换训练算法非常耗时,并且区域

提议网络会生成类似于目标的区域,而不是目标实例,并且不擅长处理极端比例或形状的目标。

### 3.1.5 R-FCN

根据感兴趣区域池化层划分,用于目标检测的流行的深度网络<sup>[30,32]</sup>由两个子网络组成:共享的完全卷积子网(独立于感兴趣区域)和不共享的感兴趣区域子网。目前最先进的图像分类网络,例如残差网络 (ResNets)<sup>[22]</sup>和 GoogLeNets<sup>[21]</sup>,都是完全卷积的。为了适应这些架构,构建没有感兴趣区域子网的完全卷积目标检测网络,然而事实证明这种解决方案效果并不好<sup>[22]</sup>。是因为在图像分类中,图像中的一个目标的移动对分类是没有影响的,但是在候选区域中,目标的任何移动在目标检测中都是意义重大的。将感兴趣区域手动接入卷积层中会解决平移不变形问题,但是会影响其他区域。因此,Li<sup>[33]</sup>等人提出了一种基于区域的全卷积网络 (R-FCN)。其运行步骤如图 5 所示。

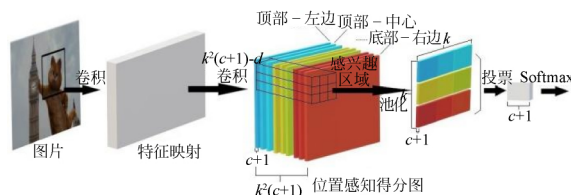


图 5 R-FCN 的网络结构

Fig. 5 Network structure of R-FCN

与 Faster R-CNN 不同的是,对于每一个类别,R-FCN 的最后一个卷积层会生成  $k^2$  个具有固定  $K \times K$  网格的位置敏感得分图,然后添加一个位置敏感感兴趣区域层进行聚合。最后,在每个感兴趣区域中,将  $k$  个位置敏感得分取平均,生成  $C+1$  维矢量,并计算跨类别的 Softmax 响应。另外附加了一个  $4k^2-d$  的卷积转换层,获取与类无关的边界框。

借助 R-FCN 可以通过共享几乎所有层来采用更强大的分类网络实现全卷积架构中的目标检测。在 PASCAL VOC 和 Microsoft COCO<sup>[34]</sup>数据集上均获得了最好的结果,测试速度为每张图像 170 ms。

### 3.2 基于 one-stage 的模型

基于 two-stage 的框架由几个阶段组成,区



域提议生成,使用 CNN 进行特征提取,分类和边界框回归,这些阶段通常是分开训练。即使在最近的端到端网络 Faster R-CNN 中,仍然需要进行替代训练才能获得区域提议网络和检测网络之间的共享卷积参数。因此,处理不同部分都需要花费的时间成为实时目标检测的瓶颈。基于全局回归/分类的一步框架,可以直接从图像像素映射到边界框坐标和分类概率,减少时间开销。本文介绍几个重要的框架,即 YOLO(You Only Look Once)和 SSD(Single Shot MultiBox Detector)<sup>[35]</sup>。

### 3.2.1 YOLO

Redmon 等人<sup>[36]</sup>提出了一个名为 YOLOv1 的框架,该框架利用最上层的整个特征图来预测多个类别和计算边界框的置信度。YOLOv1 的基本思想如图 6 所示,YOLOv1 将输入图像划分为  $S \times S$  网格,每个网格单元负责预测以该网格单元为中心的目标的边界框以及其对应的置信度分数。其中置信度表示存在目标的可能性并显示其预测的置信度。与此同时,无论边界框的数量为多少,都会在每个网格单元中预测  $C$  个条件类概率。

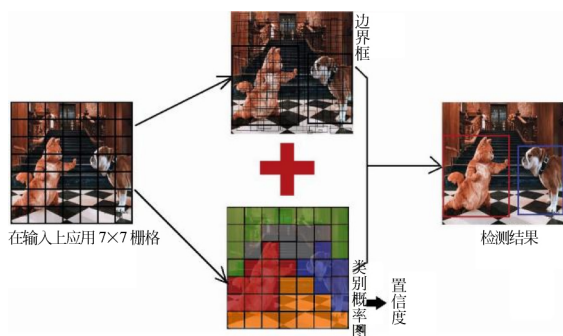


图6 YOLO模型

Fig.6 Model of YOLO

在 YOLOv1 的基础上,一种改进的版本 YOLOv2<sup>[37]</sup>被提出,YOLOv2 提出了一个新的拥有 19 个卷积层和 5 个池化层的分类主干网络,即 Darknet-19,并采用了更强大的深度卷积 ImageNet 主干框架,输入图片的分辨率由原来的  $224 \times 224$  直接设置为  $448 \times 448$ ,这使得学习到的权重对于获取微小信息更加敏感。除此之外,YOLOv2 借鉴 Faster R-CNN 中设定先验框的策略,使用全卷积网络,用 K-means 聚类算法获取先验框的宽和高,并通过预测偏移量来降低网络

训练难度。最后与批归一化、多尺度处理技术一起形成训练网络。

作为 YOLOv2 的改进版本 YOLOv3<sup>[38]</sup>,提出了更深入,更强大特征提取网络 Darknet-53。为了适应包含许多重叠标签的复杂数据集,YOLOv3 使用多标签分类。在对边界框进行预测时,YOLOv3 在 3 个不同比例的特征图进行预测,是当时速度和精度最均衡的目标检测网络。

### 3.2.2 SSD

YOLO 难以处理组合中的小物体,是因为对边界框的预测增加了空间约束<sup>[36]</sup>。同时,由于多次进行下采样,将其刻画为具有新的宽高比的目标,产生了相对粗糙的特征。

针对这些问题,Liu 等人提出了 SSD<sup>[35]</sup>,其灵感来自于 MultiBox<sup>[39]</sup>、区域提议网络<sup>[32]</sup>和多尺度表示<sup>[40]</sup>中采用的锚点。给定一个特定的特征图,SSD 利用一组具有不同长宽比和比例的默认锚点框来使边界框的输出空间离散化。为了处理不同大小的目标,将会利用多个具有不同分辨率的特征图的预测融合网络。SSD 在 VGG16 网络末端添加了几个功能层,这些功能层用于预测具有不同比例和纵横比及其相关置信度的默认框的偏移量。最终检测结果是通过对多尺度边界框上通过非极大值抑制获得的。SSD 在 PASCAL VOC 和 COCO 的准确性方面明显优于 Faster R-CNN,并且速度提高了三倍。SSD300(输入图像尺寸为  $300 \times 300$ )以 59 FPS 运行,相比于 YOLO 更加准确和高效。针对 SSD 不善于处理小目标的问题,可以通过采用更好的特征提取网络主干(例如 ResNet101),添加反卷积层以引入额外的大规模上下文<sup>[41]</sup>和设计更好的网络结构<sup>[42]</sup>。

SSD 与 YOLO 在结构上的不同在于,SSD 的特征是从不同的卷积层提取出来,进行组合再回归和分类,而 YOLO 只有一层,在 YOLO 之后的版本中也借鉴了 SSD 的这种多尺度的思想。

SSD 网络中的预选框的大小和形状不能直接通过学习获得,需要通过手动设置。而网络中的每一层使用的预选框大小和形状都不一样,调试过程依赖经验,除此之外,检测靠近彼此的多个小目标的问题仍然没有解决。SSD 使用低级特征检测小目标,但是低级特征卷积层数少,存在特征提取不充分的问题,在理论上还有很大的提升空间。

### 3.3 基于 Anchor-free 的检测模型

虽然基于 Anchor 的目标检测模型(one-stage 模型和 two-stage 模型)在效率和精度上有很大的提升,但是自身的局限性也阻碍了模型的优化创新。大多数预设的锚点是负样本,在训练过程中会导致正样本和负样本的失衡;为了实现较高的查全率,需要设置大量密集的锚点候选框,为此造成巨大的计算成本和内存需求开销;预设的锚框是人为设置的,其比例和宽高比对数据集比较敏感,很大程度上会影响检测性能。为了解决这个问题,最近提出一些 Anchor-free 检测模型。

#### 3.3.1 CornerNet<sup>[43]</sup>

CornerNet 是一种新颖的 anchor-free 网络,使用 Hourglass 网络作为主干网络,后面连接两个预测模型,分别预测左上和右下角点。在特征图的每个位置上都可以预测角度偏移、对嵌入和类热图。通过类热图可以计算出现拐角的可能性,并使用拐角偏移量来预测拐角位置。CornerNet 在 MSCOCO 数据集上获得了显著改进。

#### 3.3.2 FSAF<sup>[44]</sup>

为了让位于图像中的每个目标都能自助选择最合适的特征层,FSAF (Feature Selective Anchor-Free)在特征层中不设置 anchor,实现了 anchor-free。在训练过程中,FSAF 根据损失,自动选择最好的特征层,并将每个对象动态分配给最合适的特征层,在这个特征层进行后续的回归和预测。在预测阶段,FSAF 可以独立预测,也可以与基于锚点的方法相结合。

#### 3.3.3 ExtremeNet<sup>[45]</sup>

ExtremeNet 是在 CornerNet 上的改进,不同于 CorNet, ExtremeNet 的检测目标是 4 个极值点(最左点、最右点、最上点和最下点),而不是检测目标的左上点和右下点,这种做法避免了因为强行使用矩形包围框包围物体带来的问题。当检测到极值点之后,利用匹配算法,将所有极值点进行组合,通过验证组合中是否存在中心极值点,对所有组合示例进行筛选。该方法速度不是很快,但是标注简单,在此基础上可以进行更多的改进。

#### 3.3.4 FCOS<sup>[46]</sup>

FCOS(Fully Convolutional One-Stage)是一个基于全卷积的检测网络,它利用一个新的分支

“center-ness”来预测一个像素与对应对框中心的偏差,所得到的分数用于降低低质量检测框的权重,最后通过 NMS(Non-Maximum Suppression)将检测结果进行融合。FCOS 可以将语义分割和目标检测任务结合,除此之外,还可以对模型进行小幅度改造(例如,关键点检测)以应用于其他检索任务。

#### 3.3.5 CenterNet<sup>[47]</sup>

CenterNet 结合了基于中心的方法(中心点)和基于角点的方法(左上角点和右下角点)的思想,使网络花费小的代价便能感知物体内部信息,从而消除误检框。其作用过程为,首先通过成对的角预测边界框,然后将角点的位置映射到输入图片的对应位置,最优判断哪两个角点属于同一物体,以便形成一个检测框。该方法可以移植到其他锚点或者 anchor-free 网络上。

### 3.4 应用深度卷积神经网络算法进行目标检测存在的问题

尽管深度卷积神经网络在目标检测中发展迅速并且取得了可喜的进展,但是其本身算法的应用仍然存在挑战,本文主要划分为以下几点:

算法属于有监督的学习,在学习过程中需要大量的标记样本。从目标检测的发展状况来看,有监督的学习已经远远超过无监督学习。训练深度卷积神经网络需要大量的数据。虽然在网可以找到大量的数据,但是都杂乱无章,需要人工进行标注,这几年的标注成本逐渐升高。

训练时间过长且计算资源代价大。深度学习的参数过多,训练周期时间长,GPU 的运用加快了训练速度,缩短了训练时间,深度卷积神经网络算法随着 GPU 的运用流行起来,然而价格昂贵的 GPU 为建立大规模 GPU 集群的科研院所带来的负担。

理论研究不足。借助 GPU 的发展,深度学习使用端到端的方式优化大容量模型。近几年,度量学习、大数据、各种各样的样本选择方法、训练方法和网络结构的优化都促进了卷积神经网络的发展,然而,这些进步缺乏理论支撑,虽然有效,但是缓慢。这种没有深层次本质理解的改进,使得设计和训练卷积神经网络时,经常会出现梯度不稳定和拟合问题。卷积神经网络本身有很多超参数,这些超参数的组合是一个很大的数字,很多时候靠经验无法完成,只能凭直觉反复尝试其中

一部分组合,因此现在的卷积神经网络只能实现效果很好,但是效果和效率都不能达到最优。

## 4 目标检测数据集和评价指标

建立具有较小偏差的较大数据集对于开发高级计算机视觉算法至关重要。作为场景理解的重要组成部分,目标检测已广泛应用于现代生活的许多领域,到目前为止,许多基准数据集在目标检测中发挥了重要作用。在本节中,将展示衣蛾写常见的检测数据集以及用来衡量检测器性能的评估指标。

### 4.1 数据集

#### 4.1.1 通用的目标检测数据集

(1) Pascal 视觉挑战数据集<sup>[48]</sup>

PASCAL 视觉挑战是早起计算机视觉中最重要的竞赛之一。Pascal-VOC 的两个版本: VOC07 和 VOC12 主要用于对象检测。其中 VOC07 是包含 20 类图片的中等数据集,分为三部分,分别使用 2 501, 2 510 和 5 011 张图像进行训练、验证和测试。VOC12 与 COCO07 包含相同类别的图像,分别使用 5 717, 5 823 和 10 991 张图像进行训练、验证和测试近年来,随着一些较大数据集的发布,VOC 逐渐被取代。

(2) MSCOCO<sup>[34]</sup>是具有 80 个类别的大规模数据集,分别使用 118 287, 5 000 和 40 670 图像进行训练、验证和测试。

(3) Open Images<sup>[49]</sup>是一个包含 920 万张图像的数据集,并带有图像级标签、对象边界框和视觉关系。Open Images V5 在 1.9 M 图像上总共包含 600 万个对象类别的 1 600 万个边界框,这使其成为具有对象位置注释的最大的现有数据集。该数据集中的框是由专业注释者手工绘制的,以确保准确性和一致性,并且图像多样化,并且大多数包含有多个对象的复杂场景。

(4) ImageNet<sup>[50]</sup>包含 120 万个高分辨率图像,是一个具有 200 个类别的大型数据集,数据规模大,对象范围广,这使得使用大型训练数据训练深度模型成为可能。

4.1.2 除了一些通用的目标检测数据集,还有一些数据集是用于具体检测任务。比如,在人脸检测任务中,经常用到的数据集有 WIDER FACE 数据集<sup>[51]</sup>、FDDB 数据集<sup>[52]</sup>和 Pascal Face 数据

集<sup>[53]</sup>。在行人检测任务中,广泛使用的数据集是 Caltech 数据集<sup>[54]</sup>、CityPersons 数据集<sup>[55]</sup>、KITTI 数据集<sup>[56]</sup>、INRIA 数据集<sup>[57]</sup>和 ETH 数据集<sup>[58]</sup>。

### 4.2 评估指标

定位和分类是目标检测的两个重要任务,在物体检测评估指标下,检测准确性和定位精度是用于评估检测准确性常用的两个标准。对于定位精度,IoU 通过判断预测边界框及其对应的分配地面真值判断回归质量。对于检测准确性,常用均值平均精度(mAP)作为评估指标。

## 5 目标检测的未来方向和任务

尽管目标检测发展迅速并且检测效果显著,但是仍然存在许多未解决的问题。

### 5.1 小物体检测

在 COCO 数据集和人脸检测任务中的检测都是小物体检测,为了提高遮挡情况下小物体的定位精度,可以从以下几个方面对网络体系结构进行修改。

尺度自适应。检测目标通常以不同的比例存在,这在人脸检测和行人检测中表现尤为突出,为了增加尺度变化的鲁棒性,需要训练尺度不变、多尺度<sup>[59]</sup>或尺度自适应检测器。对于尺度不变的检测器,可以采用更强大的网络架构、负样本挖掘<sup>[60]</sup>、反向连接<sup>[61]</sup>和子类别建模<sup>[62]</sup>。对于多尺度检测器,产生多尺度特征图的 FPN<sup>[63]</sup>和利用缩小物体与大物体之间的表示差异的生成对抗网络<sup>[64-65]</sup>都弥补了特征金字塔的不足。对于尺度自适应检测器,可以将知识图谱、注意力机制、级联网络和尺度分布估计组合在一起以更好定位检测目标。

空间相关性和上下文建模。空间分布在目标检测中起着重要作用。因此,可以采用区域提议生成和网格回归来获取目标的可能位置。但是,多个提议和目标类别之间的相关性往往被忽略。除此之外,R-FCN 中的位置敏感分数图会忽略全局结构信息。为了解决这些问题,可以参考各种子集选择和顺序推理任务来获得合适的解决方案。

多任务联合优化和多模式信息融合。根据目标检测中内外部不同任务之间的相关性,许多研究人员研究了多任务联合优化<sup>[30,32]</sup>。但是除了



其中提到的任务还需要考虑目标检测中不同子任务的特征,并将多任务优化扩展到其他应用程序,例如实例分割和多目标跟踪。此外,在特定的应用中,可以将来自不同形式的信息融合在一起,以构建更具判别性的网络。

## 5.2 减轻人工标注的负担并实现实时目标检测

### 5.2.1 级联网络

级联网络往往在不同的阶段或者不同的层中构建级联检测器,浅层中不使用易于区分的示例,后期的分类器可以处理更多不易区分的样本。但在训练新网络时,级联的前段网络将被固定,阻止了不同 CNN 的优化,这表明了 CNN 级联的端到端优化的必要性。

### 5.2.2 无监督和无监督的学习

采用手动的方式标注大量的边界框非常耗时。为了减轻负担,可以整合语义先验、无监督目标发现、多实例学习和深度神经网络预测等方式,充分利用图像监督分配目标类别标签给相应的目标区域和精准的目标边界。此外,弱注释也有助于通过适量的注释工作构建高质量的检测模型。

### 5.2.3 网络优化

在特定的应用程序和平台下,通过选择最佳的检测架构<sup>[66]</sup>,在速度、内存和准确性之间取得平衡是非常重要的。然而,通过降低检测精度来让模型学习较少参数更为有意义<sup>[67]</sup>。引入更好的预训练方案、知识提炼和提示学习可以缓解这种情况。DSOD(Deeply Supervised Object)还提

## 参考文献:

- [1] KHAN A, RINNER B, CAVALLARO A. Cooperative robots to observe moving targets [J]. *IEEE Transactions on Cybernetics*, 2016, 48 (1): 187-198.
- [2] SAPUTERA Y P, WAHAB M, ESTU T T. Radar Software Development for the Surveillance of Indonesian Aerospace Sovereignty [C]. 2018 *International Conference on Electrical Engineering and Computer Science (ICECOS)*, IEEE, 2018: 189-194.
- [3] ANTON S D, SINH S, SCHOTTEN H D. Anomaly-based Intrusion Detection in Industrial Data with SVM

出可以从头开始进行训练,以弥合不同图像源和任务之间的鸿沟<sup>[42]</sup>。

## 5.3 3D 目标检测和视频目标检测

3D 目标检测。深度学习借助 3D 传感器可以更好地理解 2D 图像,并将图像级别的知识扩展到现实世界。但是,很少利用这些 3D 感知技术将 3D 边界框放置在检测到的目标周围。为了获得更好的边界标定,惯性传感器可以借助多视图表示和 3D 建议网络对深度信息进行编码。

视频目标检测。在理解不同目标的行为时,时间信息起着重要的作用。但是,准确性容易受到视频中目标外观退化的影响,并且不采用端到端的方式训练网络。为此,可以考虑时空小管、光学流和 LSTM 等方法从根本上为连续的帧之间建立目标关联。

## 6 结 论

目标检测经过几十年的研究发展,已经逐渐成为一个成熟的研究领域。凭借其强大的学习能力以及在处理遮挡、尺度转换和背景切换方面的优势,基于深度学习的目标检测已成为近年来的研究热点。本文详细介绍了基于卷积神经网络的目标检测框架,以及其在 R-CNN 上进行的不同程度的改进,最后提出了目标检测在未来的发展方向,以全面了解目标检测领域。

and Random Forests [C]. 2019 *International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, IEEE, 2019: 1-6.

- [4] 王耀东,朱力强,余祖俊,等. 用于机械系统瞬时目标的双视角高速视觉检测系统 [J]. *光学 精密工程*, 2017, 25(10): 2725-2735.  
WANG Y D, ZHU L Q, YU Z J, et al.. Two-view high speed vision system for instant object detection in mechanical system [J]. *Opt. Precision Eng.*, 2017, 25(10): 2725-2735. (in Chinese)
- [5] JIANG A Q, HUYNH D. Multiple pedestrian tracking from monocular videos in an interacting multiple model framework [J]. *IEEE Transactions on Image Processing*, 2017, 27 (3): 1361-1375.

- [6] 张小荣, 胡炳樑, 潘志斌, 等. 基于张量表示的高光谱图像目标检测 [J]. 光学 精密工程, 2019, 27(2): 488-498.  
ZHANG X R, HU B L, PAN ZH B, *et al.*. Tensor representation based target detection for hyperspectral imagery [J]. *Opt. Precision Eng.*, 27(2): 488-498. (in Chinese)
- [7] 李正周, 曹雷, 邵万兴, 等. 基于空时混沌分析的海面小弱目标检测精密工程 [J]. 光学 精密工程, 2018, 26(1): 193-199.  
LI ZH ZH, CAO L, SHAO W X, *et al.*. Detection of small target in sea clutter based on spatio-temporal chaos analysis [J]. *Opt. Precision Eng.*, 2018, 26(1): 193-199. (In Chinese)
- [8] LOWE D. Distinctive image features from scale-invariant keypoints [J]. *International Journal of Computer Vision*, 2004, 60(2): 91-110.
- [9] CAI Z W, SABERIAN M, VASCONCELOS N. Learning complexity-aware cascades for deep pedestrian detection [C]. *Proceedings of the IEEE International Conference on Computer Vision*, 2015: 3361-3369.
- [10] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features [J]. *CVPR*, 2001, 1(3): 511-518.
- [11] ZHANG C X, ZHANG J S, KIM S W. PBoost-GA: pseudo-boosting genetic algorithm for variable ranking and selection [J]. *Computational Statistics*, 2016, 31(4): 1237-1262.
- [12] PEI L, YE M, ZHAO X Z, *et al.*. Learning spatio-temporal features for action recognition from the side of the video [J]. *Signal, Image Video Processing*, 2016, 10(1): 199-206.
- [13] LECUN Y, BOTTOU L, BENGIO Y, *et al.*. Haffner, " Gradient-based learning applied to document recognition [J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.
- [14] LECUN Y, BOSER B, DENKER J, *et al.*. Handwritten digit recognition with a back-propagation network [J]. *Advances in Neural Information Processing Systems*, 1990: 396-404.
- [15] HECHT-NIELSEN. Theory of the backpropagation neural network [J]. *Neural networks for perception*; Elsevier, 1992: 65-93.
- [16] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. *Advances in neural information processing systems*, 2012: 1097-1105.
- [17] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines [C]. *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010: 807-814.
- [18] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, *et al.*. Improving neural networks by preventing co-adaptation of feature detectors [J]. *Computer Ence*, 2012, 3(4): 212-223.
- [19] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks [J]. *European conference on computer vision*, Springer, 2014: 818-833.
- [20] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. *Computer Ence*, 2014.
- [21] SZEGEDY C, LIU W, JIA Y, *et al.*. Going deeper with convolutions [C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015: 1-9.
- [22] HE K, ZHANG X Y, REN S Q, *et al.*. Deep residual learning for image recognition [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.
- [23] NORMALIZATION B. Accelerating deep network training by reducing internal covariate shift [C]. *International Conference on International Conference on Machine Learning JMLR*, 2015.
- [24] UIJLING J R, VAN DE SANDE K E, GEVERS T, *et al.*. Selective search for object recognition [J]. *International Journal of Computer Vision*, 2013, 104(2): 154-171.
- [25] KUO W, HARIHARAN B, MALIK J. Deepbox: Learning objectness with convolutional networks [C]. *Proceedings of the IEEE International Conference on Computer Vision*, 2015: 2479-2487.
- [26] PINHEIRO P O, LIN T-Y, COLLOBERT R, *et al.*. Learning to refine object segments [C]. *European Conference on Computer Vision*, Springer-

- er, 2016; 75-91.
- [27] GUPTA S, GRISHICK R, ARBELAEZ P, *et al.*. Learning rich features from RGB-D images for object detection and segmentation [C]. *European Conference on Computer Vision*, Springer, 2014; 345-360.
- [28] PERRONNIN F, SANCHEZ J, MENSINK T. Improving the fisher kernel for large-scale image classification [C]. *European Conference on Computer Vision*, Springer, 2010; 143-156.
- [29] HE K, ZHANG X Y, REN S Q, *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.
- [30] GIRSHICK R. Fast R-CNN [C]. *Proceedings of the IEEE International Conference on Computer Vision*, 2015; 1440-1448.
- [31] XUE J, LI J, GONG Y. Restructuring of deep neural network acoustic models with singular value decomposition [C]. *Interspeech*, 2013; 2365-2369.
- [32] REN S, HE K, GIRSHICK R, *et al.*. Faster R-CNN: Towards real-time object detection with region proposal networks [C]. *Advances in Neural Information Processing Systems*, 2015; 91-99.
- [33] DAI J, LI Y, HE K, *et al.*. R-FCN: Object detection via region-based fully convolutional networks [C]. *Advances in Neural Information Processing Systems*, 2016; 379-387.
- [34] LIN T-Y, MAIRE M, BELONGIE S, *et al.*. Microsoft coco: Common objects in context [C]. *European Conference on Computer Vision*, Springer, 2014; 740-755.
- [35] LIU W, ANGUELOV D, ERHAN D, *et al.*. SSD: Single shot multibox detector [C]. *European conference on computer vision*, Springer, 2016; 21-37.
- [36] REDMON J, DIVVALA S, GIRSHICK R, *et al.*. You Only Look Once: Unified, real-time object detection [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; 779-788.
- [37] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017; 7263-7271.
- [38] REDMON J, FARHADI A. YOLOV3: An incremental improvement [J]. *arXiv preprint arXiv: 1804.02767*, 2018.
- [39] ERHAN D, SZEGEDY C, TOSHEV, *et al.*. Scalable object detection using deep neural networks [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014; 2147-2154.
- [40] BELL S, LAWRENCE Z, BALA K, *et al.*. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; 2874-2883.
- [41] FU C-Y, LIU W, RANGA A, *et al.*. DSSD: Deconvolutional single shot detector [J]. *arXiv preprint arXiv: 1701.06659*, 2017.
- [42] SHEN Z, LIU Z, LI J, *et al.*. DSOD: Learning deeply supervised object detectors from scratch [C]. *Proceedings of the IEEE International Conference on Computer Vision*, 2017; 1919-1927.
- [43] LAW H, HENG J. Cornernet: Detecting objects as paired keypoints [C]. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018; 734-750.
- [44] ZHU C, HE Y, SAVVIDES M. Feature selective anchor-free module for single-shot object detection [J]. *arXiv preprint arXiv: 00621*, 2019.
- [45] ZHOU X, ZHOU J, KRAHENBUHL P. Bottom-up object detection by grouping extreme and center points [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019; 850-859.
- [46] TIAN Z, SHEN C, CHEN H, *et al.*. FCOS: Fully Convolutional One-Stage Object Detection [J]. *arXiv preprint arXiv:01355*, 2019.
- [47] DUAN K, BAI S, XIE L, *et al.*. Centernet: Key-point triplets for object detection [C]. *Proceedings of the IEEE International Conference on Comput-*

- er Vision*, 2019; 6569-6578.
- [48] EVERINGHAM M, WAN G, WILLIAMS C, *et al.*. The pascal visual object classes (voc) challenge [J]. *International Journal of Computer Vision*, 2010, 88(2): 303-338.
- [49] KUZNETSOVA A, ROM H, ALLDRIN N, *et al.*. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale [J]. *arXiv preprint arXiv:1811.00982*, 2018.
- [50] DENG J, DONG W, SOCHER R, *et al.*. Imagenet: A large-scale hierarchical image database [C]. *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009; 248-255.
- [51] YANG S, LUO P, LOY C-C, *et al.*. Wider face: A face detection benchmark [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; 5525-5533.
- [52] JAIN V, LEARNED-MILLER E. FDDB: A benchmark for face detection in unconstrained settings [C]. *Computer Science*, 2010.
- [53] FELZENSZWALB P, GIRSHICK R, MCALLESTER D, *et al.*. Discriminatively trained mixtures of deformable part models [J]. *PASCAL VOC Challenge*, 2008.
- [54] DOLLAR P, WOJEK C, SCHIELE B, *et al.*. Pedestrian detection: An evaluation of the state of the art [J]. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2011, 34(4): 743-761.
- [55] ZHANG S, BENENSON R, SCHIELE B. Citypersons: A diverse dataset for pedestrian detection [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017; 3213-3221.
- [56] GEIGER A, LENZ P, STILLER C, *et al.*. Vision meets robotics: The KITTI dataset [J]. *The International Journal of Robotics Research*, 2013, 32(11): 1231-1237.
- [57] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, 1: 886-893.
- [58] ESS A, LEIBE B, VANGOOL L. Depth and appearance for mobile scene analysis [C]. *2007 IEEE 11th International Conference on Computer Vision*, IEEE, 2007; 1-8.
- [59] 刘晓, 崔光照, 李正周, 等. 基于视觉系统分层的小目标运动检测 [J]. *光学精密工程*, 2019, 27(10): 2251-2262.
- LIU X, CUI G ZH, LI ZH ZH, *et al.*. Small target motion detection based on layering of vision system [J]. *Opt. Precision Eng.*, 2019, 27(10): 2251-2262. (in Chinese)
- [60] SHRIVASTAVA A, GUPTA A, GIRSHICK R. Training region-based object detectors with online hard example mining [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; 761-769.
- [61] KONG T, SUN F, YAO A, *et al.*. Ron: Reverse connection with objectness prior networks for object detection [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017; 5936-5944.
- [62] XIANG Y, CHOI W, LIN Y, *et al.*. Subcategory-aware convolutional neural networks for object proposals and detection [C]. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2017; 924-933.
- [63] LIN T-Y, DOLLAR P, GIRSHICK R, *et al.*. Feature pyramid networks for object detection [C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017; 2117-2125.
- [64] GOODFELLOW L, POUGET-ABADIE J, MIRZA M, *et al.*. Generative adversarial nets [C]. *Advances in Neural Information Processing Systems*, 2014; 2672-2680.
- [65] 梁浩, 刘克俭, 刘康, 等. 引入再检测机制的孪生神经网络目标跟踪 [J]. *光学精密工程*, 2019, 27(7): 1621-1631.
- LIANG H, LIU K J, LIU K, *et al.*. Target tracking in twin neural networks with re-detection mechanism [J]. *Opt. Precision Eng.*, 2019, 27(7): 1621-1631. (in Chinese)
- [66] HUANG J, RATHOD V, SUN C, *et al.*. Speed/

accuracy trade-offs for modern convolutional object detectors [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 7310-7311.

[67] TOME D, BONDI L, BAROFFIO L, *et al.*. Re-

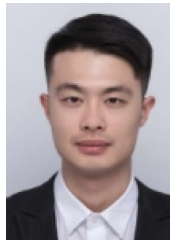
duced memory region based deep Convolutional Neural Network detection [C]. 2016 *IEEE 6th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, IEEE, 2016: 15-19.

#### 作者简介:



范丽丽(1991—),女,山东烟台人,博士研究生,2016 年于吉林大学获得硕士学位,主要从事计算机视觉及人工智能方面的研究。E-mail: llfan18@mails.jlu.edu.cn

#### 通讯作者:



赵浩宇(1991—),男,吉林长春人,硕士,2016 年于吉林大学获得硕士学位,主要从事智能信息系统与科技信息传播方面的研究。通讯作者, E-mail: zhaohaoyu@jlu.edu.cn

#### 导师简介:



赵宏伟(1962—),男,吉林长春人,教授,博导,1993 年于吉林大学(原吉林工业大学)获得硕士学位,2001 年于吉林大学获得博士学位,主要从事图像理解,认知计算,嵌入式系统与物联网技术、汽车电子和智能信息系统方面的研究。E-mail: zhaohw@jlu.edu.cn