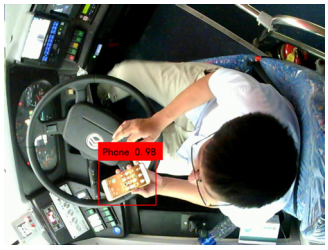
Opto-Electronic Engineering
光 电 工 程

Article

2021 年，第 48 卷，第 4 期

DOI: 10.12086/oee.2021.200325

一种用于驾驶场景下手机检测的端到端的神经网络



戴 腾[1,2]，张 珂[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： 小目标物体实时检测一直是图像处理领域中的难点。本文基于深度学习的目标检测算法，提出了一种端到端的神经网络，用于复杂驾驶场景下的手机小目标检测。首先，通过改进 YOLOv4 算法，设计了一个端到端的小目标检测网络(OMPDNet)来提取图片特征；其次，基于 K-means 算法设计了一个聚类中心更加贴切数据样本分布的聚类算法 K-means-Precise，用以生成适应于小目标数据的先验框(anchor)，从而提升网络模型的效率；最后，采用监督与弱监督方式构建了自己的数据集，并在数据集中加入负样本用于训练。在复杂的驾驶场景实验中，本文提出的 OMPDNet 算法不仅可以有效地完成驾驶员行车时使用手机的检测任务，而且对小目标检测在准确率和实时性上较当今流行算法都有一定的优势。

关键词： 目标检测；神经网络；聚类算法；监督与弱监督

中图分类号： TP181；TP391.41　　**文献标志码：** A

戴腾，张珂，尹东. 一种用于驾驶场景下手机检测的端到端的神经网络[J]. 光电工程，2021，**48**(4): 200325
Dai T, Zhang K, Yin D. An end-to-end neural network for mobile phone detection in driving scenarios[J]. *Opto-Electron Eng*, 2021, **48**(4): 200325

An end-to-end neural network for mobile phone detection in driving scenarios

Dai Teng[1,2], Zhang Ke[1,2], Yin Dong[1,2*]

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

Abstract: Real-time detection of small objects is always a difficult problem in image processing. Based on the target detection algorithm of deep learning, this paper proposed an end-to-end neural network for mobile phone small target detection in complex driving scenarios. Firstly, an end-to-end small target detection network (OMPDNet) was designed to extract image features by improving the YOLOv4 algorithm. Secondly, based on the K-means algorithm, a K-means-Precise clustering algorithm of more appropriate data samples distribution in the clustering center was designed, which was used to generate prior frames suitable for small target data, so as to improve the efficiency of the network model. Finally, we constructed our own data set with supervision and weak supervision, and added negative samples to the data set for training. In the complex driving scene experiments, the OMPDNet algorithm

收稿日期：2020-09-02； 收到修改稿日期：2020-12-21
基金项目：安徽省 2018 年度重点研究与开发计划项目(1804a09020049)
作者简介：戴腾(1996-)，男，硕士研究生，主要从事计算机视觉的研究。E-mail：daiteng@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

proposed in this paper can not only effectively complete the detection task of using mobile phone while driving, but also has certain advantages over the current popular algorithms in accuracy and real-time for small target detection.

Keywords: object detection; neural network; clustering algorithm; supervision and weak supervision

1 引 言

遵守道路交通法规、安全驾驶是驾驶员牢记于心的铁则。但司机开车时使用手机的现象却屡见不鲜，这将造成巨大的安全隐患，甚至酿成惨祸。因此，实时检测司机驾驶行为，不仅有利于交管部门的管控，而且对于减小交通事故的发生都具有重大的现实意义。

近年来，Rodríguez-Ascariz 等[1]利用电子电路射频采集来捕获使用手机所产生的功率，并将位于车内的两个天线和信号分析算法用于识别驾驶员何时使用手机。Leem SK 等[2]通过脉冲无线电超宽带雷达进行生命体征监测，提出了一种不受驾驶活动引发的运动影响的生命体征估算算法。这些方法依靠的是手机产生的信号，通过硬件传感器设备捕获并处理，从而有一定的误差。当前计算机视觉发展迅猛，应用也越来越广泛，图像和视频处理更加准确高效。Berri 等[3]构建了一个模式检测识别系统，由车内监控摄像头得到图像后提取图像特征，用具有多项式内核的支持向量机(SVM)[4]实现分类。Xiong 等[5]提出了一种基于深度学习的驾驶员手机使用率检测算法。它首先使用渐进式校准网络(PCN)[6]进行面部检测和跟踪；其次，采用卷积神经网络检测候选区域中的手机。其结果对比基于传感器和信号处理的方法，减小了环境干扰的因素，具有实时的直观反馈，在准确率上也有提升，但仍未满意。

在实际检测识别的工程应用中，不仅要求有较高的准确率和较强的鲁棒性，而且实时性也很重要。故本文基于深度学习的目标检测算法，提出了一种用于驾驶场景下手机检测的端到端的神经网络。首先，为了维持较高的准确率，同时还能保证实时性，本文改进了 YOLOv4[7]算法，设计了一个端到端的小目标检测网络(one-stage mobile phone detection network，OMPDNet)来提取图片特征；其次，基于 K-means[8]设计了一个聚类中心更加贴切样本数据分布的聚类算法 K-means-Precise，用以生成适应于小目标数据的先验框(anchor)，从而提升网络模型的效率；最后，由于公开数据集不能适用于特定的驾驶场景，本文采用监督与弱监督方式构建了自己的数据集。同时，为了解决训练时正负样本不平衡问题，在数据集中加入负样本用于训练。

2 相关工作

2.1 卷积神经网络

随着深度学习的不断发展，卷积神经网络[9]在计算机视觉中取得了巨大的成功，其卷积特性使得图像和视频的处理更加精准、高效。1998 年，LeCun 等[10]建立了一个手写体的卷积神经网络 LeNet-5，是首个卷积神经网络的现代模型。随着 2012 年 Krizhevsky 等人提出的卷积神经网络，AlexNet[11]大幅度提升了图像分类的准确率，掀开了卷积神经网络的高潮，从而产生了一系列优秀的神经网络，如 VGG[12]、MobileNets[13]、GoogleNet[14]、ResNet[15]等。这些网络为本文的研究提供了思路，具有很好的借鉴意义。

2.2 目标检测

目标检测是计算机视觉中的基础任务，其研究有两大类：一是传统的目标检测方法，如 Viola 和 Jones 提出的 Haar 分类器[16-17]，它由 Haar 特征提取、离散强分类器和强分类级联器组成，核心思想是提取人脸的 Haar 特征，可应用于人脸检测，同时级联分类器 Cascade[16]与 Dalal[18]等人提出的 HOG 特征，在行人检测方面取得了较好的结果。Felzenszwalb[19]设计了 DPM 算法实现了人体检测拓展到了物体检测。此外，便是这些算法的诸多改进和优化。另一类是基于深度学习的目标检测算法，旨在利用神经网络对图像进行特征提取，同时输出目标。其总体方向是通过搭建复杂的卷积神经网络，在庞大的数据驱动下不断地训练和优化，最终得到各项指标较好的模型。

基于深度学习的目标检测建立在基础网络上，其模型可分为两阶段(Two-stage)和一阶段(One-stage)两种类型。以 R-CNN[20]、SPPNet[21]、Fast R-CNN[22]、Faster R-CNN[23]和 R-FCN[24]为主的 Two-stage 算法其核心思想是先提取候选区域(region proposal)，然后再利用卷积神经网络进行分类，原理如图 1 所示。One-stage 算法则去掉了候选框的提取步骤，在分类的同时，直接对边界框回归，其原理如图 2。这一类算法代表主要有 YOLO 系列算法[25-27,7]、SSD[28]等。Two-stage 检测算

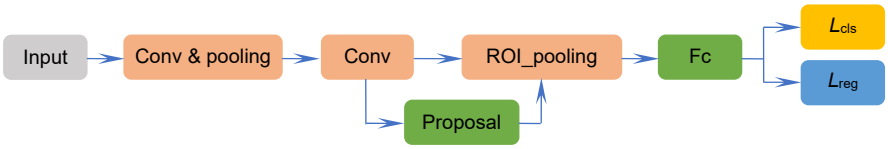


图 1 两阶段过程原理

Fig. 1 Basic two-stage process

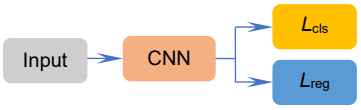


图 2 一阶段过程原理(端到端)

Fig. 2 Basic one-stage process (end to end)

法精度高，但大量的候选区域框的输入使得模型速度稍逊。One-stage 方法则在损失精度的同时提升速度，是一种严格意义上的端到端的训练，这使得训练更加简便，可应用在需实时处理视频的工程中。

2.3 YOLOv4 算法

YOLOv4 是目前官方认可的 YOLO 系列中发展的最新、最好的算法，其结构主要分为三个部分：主干网(Backbone)、颈部(Neck)和头部(Head)。

YOLOv4 的主干网是 CSPDarkNet53。它是基于 DarkNet53[27]网络并通过 CSPNet[31]方法改进得到的一种性能优异的网络。CSPNet 通过改进密集块和过渡层的信息流，避免了重复的梯度信息，从而大大减少计算量，优化了网络性能，提高推理速度和准确性。对于轻量化后的卷积神经网络，CSPNet 仍能加强网络的学习能力，保持足够的准确性。

YOLOv4 以 SPP[21]和 PANet[32]作为颈部，目的是对浅层特征进行加工和增强。SPP 层是一种多尺度特征提取层，可以产生固定大小的输出。它扩大了感受野，提高图像的尺度不变性，降低了过拟合。PANet 将高低层特征融合，减少了浅层特征传递信息的损失，从而也提高了小尺寸目标的检测。

YOLOv4 的头部沿用了 YOLOv3 结构，采用 YOLO Head 多尺度输出，在不同尺度下输出张量的深度表示边界框偏移量、类别和先验框。网络端到端的实现过程是由预设的候选框(region proposal)通过线性回归得到预测框(predict box)。其原理在于，当真实框(ground truth)和候选框之间的 IOU 很大时，反映了两者之间极为接近，此时用线性回归近似代替。

定义四种线性变换函数 $f_x(R)$，$f_y(R)$，$f_w(R)$，$f_h(R)$，如下所式：

$$\begin{cases} P_x = R_w f_x(R) + R_x \\ P_y = R_h f_y(R) + R_y \\ P_w = R_w \mathrm{e}^{f_w(R)} \\ P_h = R_h \mathrm{e}^{f_h(R)} \end{cases}, \tag{1}$$

其中：X_x 和 X_y 为窗口中心坐标，X_w 和 X_h 为窗口的宽和高(X 表示 R, P, G，分别是候选框、预测框和真实框)。

根据模式识别与机器学习[8]中的线性回归思想构建线性模型，图像由神经网络模型提取的特征 φ，通过学习参数 w_* (*表示 x,y,w,h)，得到窗口的偏移量，即上述线性变换函数 f_*：

$$f_* = w_* \varphi(R) \quad 。 \tag{2}$$

损失函数：

$$L_{\mathrm{Loss}} = \arg\min \sum_{i=0}^{N} (g_*^i - w_*'^{\mathrm{T}} \varphi(R^i))^2 + \lambda \| w_*' \|^2 \quad , \tag{3}$$

其中：

$$\begin{cases} g_x = \dfrac{G_x - R_x}{R_w} \\ g_y = \dfrac{G_y - R_y}{R_h} \\ g_w = \log(\dfrac{G_w}{R_w}) \\ g_h = \log(\dfrac{G_h}{R_h}) \end{cases}, \tag{4}$$

X^i 表示第 i 个窗口。

3 本文工作和算法

本文的工作是完成司机行车时是否使用手机的检测任务，手机在驾驶场景下属于小目标，但小目标检测却是视频图像处理中的难点。小目标是指目标尺寸的长宽是原图像尺寸的 10%以下，或者尺寸小于 32 pixels × 32 pixels 的目标，具有低分辨率的特点。由于网络模型识别的精度往往不够，处理此类问题颇为棘

手。金瑶等[29]提出了一种基于 Road_Net 卷积神经网络的检测方法，在城市道路视频中取得了较好的检测结果。Hu 等[30]为了提升小脸检测精度，围绕着尺度不变、图像分辨率和上下文三个方面做了研究，在公开人脸数据集 WIDER FACE 上，比当前的技术减少了 2 倍误差。

3.1 OMPDNet 网络架构

在复杂的驾驶场景下，由于摄像头和驾驶员操控手机的角度和位置的不同，时常会出现遮挡、尺寸小、分辨率低的情况，也就是典型的小目标检测。虽然 YOLOv4 在公开数据集中速度和精度上都有较好的表现,但是公开数据集具有数量多、种类繁杂和尺度不一的特点，是不同于本文指定的驾驶场景，此时 YOLOv4 并不能很好地直接应用。为解决此问题，本文基于 YOLOv4，提出了一种端到端的、可适应于驾驶场景下的手机小目标检测网络(one-stage mobile phone detection network，OMPDNet)。

如图 3 所示，OMPDNet 网络架构主要包括主干特征提取网络、特征增强网络和端到端的多尺度输出。输入的 RGB 三通道图片通过主干特征提取网络的一系列不同数量和大小的 Resblock_body 块，得到高度抽象的多尺度特征图(feature map)。在特征增强网络部分，高层特征图通过下采样(DownSampling)与低层特征图堆叠(concat)，而低层特征图通过上采样(UpSampling)与高层特征图堆叠，达到多尺度特征提取的目的。而最底层特征图则通过 SPP 的最大池化(max pooling)，进一步增强特征。高低层堆叠和特征增强后，最终通过端到端的多尺度输出得到预测结果 output1、output2、output3 和 output4。

由于浅层特征对小目标更加敏感，因此OMPDNet 在保持网络深度的情况下，充分利用浅层特征。以输入图片大小为 608×608 为例，在特征提取的主干网络

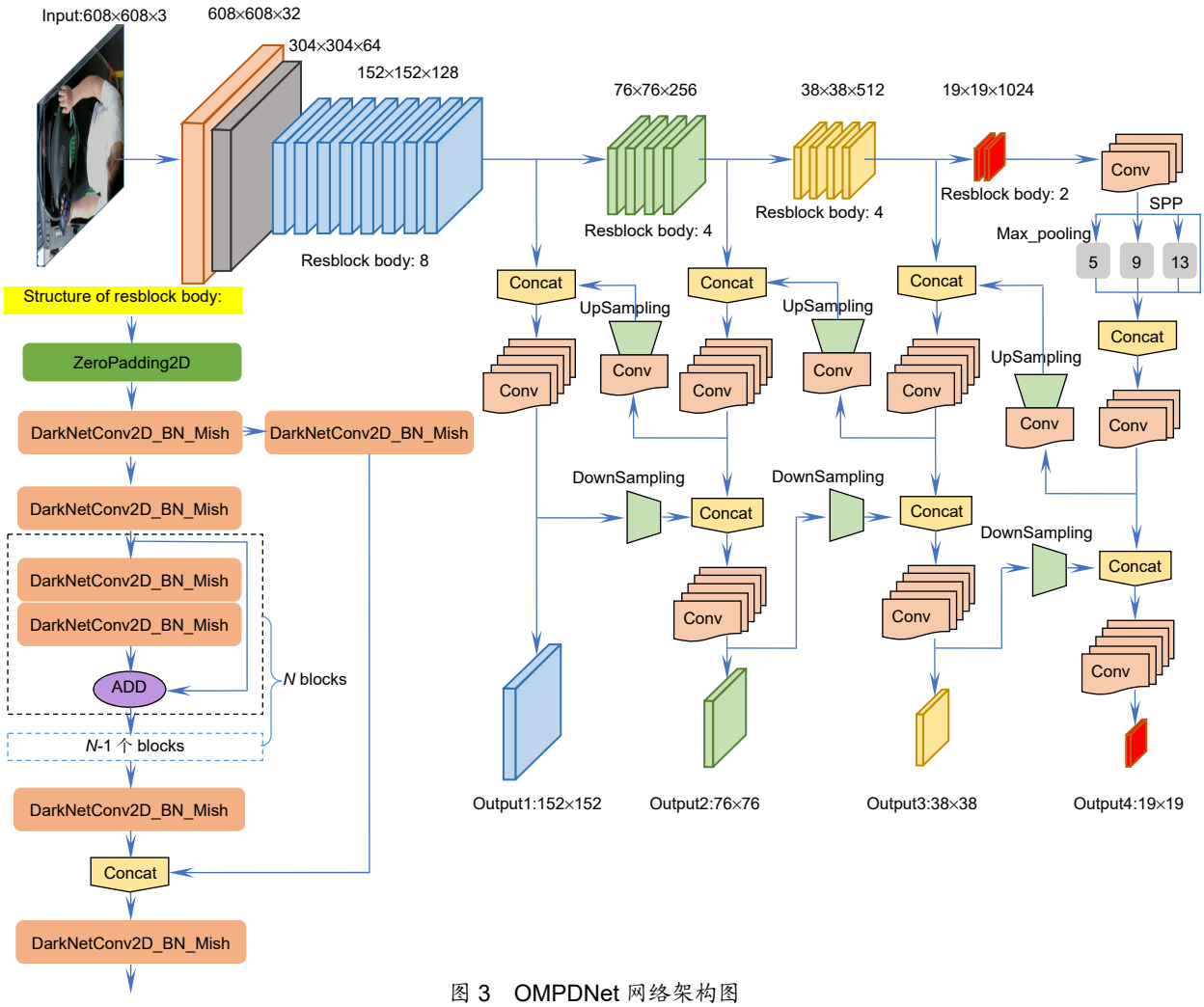


图 3 OMPDNet 网络架构图

Fig. 3 OMPDNet network architecture diagram

中，将特征尺度大小为 104×104 的浅层特征也加入多尺度融合。但 YOLOv4 中的 CSPDarkNet53 在这一特征尺度的 Resblock_body 的卷积数量只有 2，残差网络不够深，head 会过早地输出，无法充分提取特征，故 OMPDNet 通过将 Resblock_body 的卷积数量增至 8，从而克服这一问题。为了提高实时性和网络的轻量化，在保证精度不降的情况下，本文精简了网络，在 76×76、38×38 和 19×19 的特征尺度上，特征提取的 Resblock_body 的层数分别降至 4、4 和 2。

3.2 K-means-Precise 算法

设置大小合适的 Anchor，有利于提高模型的收敛速度和精度。YOLOv4 采用 K-means 算法对数据聚类获得 Anchor 的大小，此时 Anchor 对应着聚类中心。这种方法充分利用数据分布特性，比预定义更为可靠。但 K-means 算法在意的是类别的划分，而不是聚类中心。而小目标数据要求 Anchor 的跨度较小，密度精细且集中，这将导致由 K-means 生成的 Anchor 并不处于最佳位置。如图 4 可以发现，当数据中存在着少数的不可抗的噪声数据和难例数据时，虽然聚类簇没有发生变化，但聚类中心已偏移，降低了 Anchor 的精度。

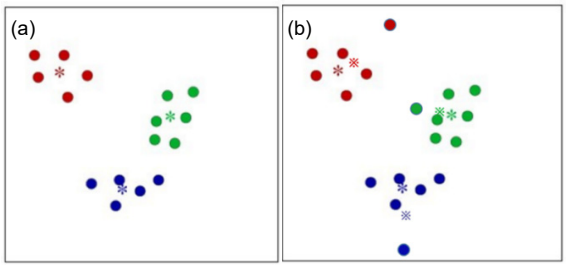


图 4 数据和难例数据聚类。
(a) 无噪声；(b) 有噪声
Fig. 4 Clustering of data and difficult data.
(a) Noise-free data; (b) Noisy data

为此，本文设计了一个使聚类中心更为精确的算法，称作 K-means-Precise。首先采用 K-means 聚类得到 k 个类别，通过设置方差阈值λ，将每个类别中偏离中心的噪声数据和难例数据移除。然后不断重复，当数据集和聚类中心都不再发生变化时，就得到了较为准确的结果。由此得到的 Anchor 更加符合数据样本分布，可用于后续的网络检测。

K-means-Precise 算法描述如下：

Input: dataset $\{x_1,x_2,\ldots,x_N\}$,number of clusters k,
variance threshold λ
Output : clusters $q\{x_i\}\in\{1,2,\ldots,k\}$
Initialize centroids $\{c_1,c_2,\ldots,c_k\}$
Repeat
  Repeat
    for i=1,2,…,N do
      $q\{x_i\}\in\leftarrow\text{argmin}_j|x_i-c_j|$
    end for
    for j=1,2,…,k do
      $c_j\leftarrow\text{mean}|x_i|q\{x_i\}=j|$
    end for
  Until centroids do not change
  for m=1,2,…,k do
    calculate $q\{x_i\}$ mean μ_i and variance σ_i
    if $\sigma_i>\lambda$ then
      remove the sample
    end if
  end for
  $\{x_1,x_2,\ldots,x_N\}\leftarrow\{x_1,x_2,\ldots,x_{N'}\}$ $N\leftarrow N'$
Until dataset and centroids do not change

3.3 OMPD Dataset 数据集

本文通过车内监控摄像头拍摄视频，并结合少量公开数据集和互联网图片等，手动制作了一个数据集(OMPDDataset)，如图 5 所示。采集数据时，为了解决正负样本不平衡问题，将负样本也加入训练中，提高模型精度和鲁棒性。在对数据集进行标注的时候，本文不完全依靠单一的人工标注位置，而采用了一种监督和弱监督结合的方式对数据进行标注。具体做法是，首先人工标注手机具体位置，然后借鉴当前行人检测较好的模型 CenterNet[33]对本文的数据预测，为检测到的行人打上标注，最后结合人工标注剔除一部分只有行人而没有手机的标注。通过人和手机的双重检测减少漏检和误检，为准确率增加一道保障，同时又减少了人工标注量。

4 实验结果与分析

4.1 实验环境

本文实验环境为 Ubuntu 16.04 系统，显卡为 1080Ti，cuda 版本为 10.0，编程语言为 python3.6，深度学习框架 pytorch1.1.0。

4.2 数据集增强

数据集 OMPDDataset 中训练集有 20000 张，测试集有 2000 张。在训练集的 20000 张图片中，10000 张为驾驶员在驾驶过程中使用手机的正样本，10000 张为驾驶员正常驾驶的负样本。在数据增强方面，本文借鉴 mixup[34]方法，基于邻域风险最小化原则，使用

图 5 数据集 OMPDDataset

Fig. 5 OMPDDataset

线性插值得到新数据，新数据的生成方式：

$$(x_n, y_n) = \lambda(x_i, y_i) + (1-\lambda)(x_j, y_j), \lambda \in (0,1) \quad , \quad (5)$$

其中：(x_n, y_n) 是插值生成的新数据，(x_i, y_i) 和 (x_j, y_j) 是在训练集数据中随机选取的两个数据。此外，还采用随机裁剪、旋转、马赛克数据增强、调整饱和度等数据增强方法对训练集进行预处理。

4.3 评价方法

本文使用召回率(Recall，简称 R)和精确率(Precision，简称 P)作为评价指标。

召回率的计算如式：

$$R = \frac{TP}{TP + FN} \quad , \quad (6)$$

其中：TP 是检测正确的正样本数量，FN 是检测错误的负样本数量。

精确率的计算式：

$$P = \frac{TP}{TP + FP} \quad , \quad (7)$$

其中 FP 是检测错误的正样本数量。

在深度学习目标检测中，仅凭精确率来评估模型性能是不够的，我们需要在精确率较高的基础上，召回率也能实现最大化。平均精确度(Average precision，简称 AP)将不同召回率下的精确率累积，反映了整体信息，其评估更为有效，计算式为

$$AP = \int_0^1 P(r)\mathrm{d}r \quad , \quad (8)$$

其中 $P(r)$是当召回率 R=r 的精确率。

本文使用平均精确率(mean average precision，简称 mAP)对结果进行评估：

$$mAP = \frac{1}{|Q_R|} \sum_{q \in Q_R} AP(q) \quad , \quad (9)$$

其中 Q_R是指模型类别。在模型速度的可行性上，则用每秒传输帧数评估。

4.4 对比实验与数据分析

本文做了两类对比实验，并利用模型评估方法来客观评估本文算法。

对比实验一：如表 1 和表 2 所示，在 OMPDNet 网络上分别单独加入 K-means-Precise 和负样本训练(negative sample training)。实验结果表明，增加 K-means-Precise 对模型的准确率会有微弱的提升，但更重要的是，它提前了 5 个周期(epoch)收敛，而将负样本加入训练其精确率和平均精确率分别提高了

表 1 K-means-Precise 的实验结果

Table 1 Experimental results of K-means-Precise

Method	*P*/%	*R*/%	*mAP*/%	Convergence time
OMPDNet	84.5	94.2	82.4	64 epoch
OMPDNet+K-means-Precise	85.7	94.3	83.2	**59 epoch**

表 2 负样本训练的实验结果

Table 2 Experimental results of negative sample training

Method	*P*/%	*R*/%	*mAP*/%
OMPDNet	84.5	94.2	82.4
OMPDNet+negative sample training	**89.2**	94.1	**86.3**

4.7%和 3.9%，由此可见，本文算法能加快模型收敛，提高模型精度。

对比实验二：选取当前具有代表性且性能优异的目标检测算法与本文提出的用于驾驶场景下的 OMPDNet 手机检测算法做对比，实验结果如表 3。本文提出的算法在速度和平均精确率都有较大的提升，而召回率相比于 YOLOv4 提高了 11.9%，达到了 96.1%，这得益于 OMPDNet 网络的多尺度改进。

再则，图 6~7 展示了本文实际检测效果图。图 6 展示了在不同的位置、不同的角度和不同的驾驶场景下，驾驶员使用手机的检测结果，均达到了实际检测要求。

表 3 五种算法的性能比较

Table 3 Performance comparisons of five algorithms

Method	*P*/%	*R*/%	*mAP*/%	Speed/(f/s)
Faster R-CNN	85.4	83.5	78.6	23.2
SSD	78.6	75.9	75.8	44.5
YOLOv3	82.3	79.4	80.1	52.4
YOLOv4	89.8	84.2	83.6	56.8
Ours	89.7	96.1	89.4	72.4

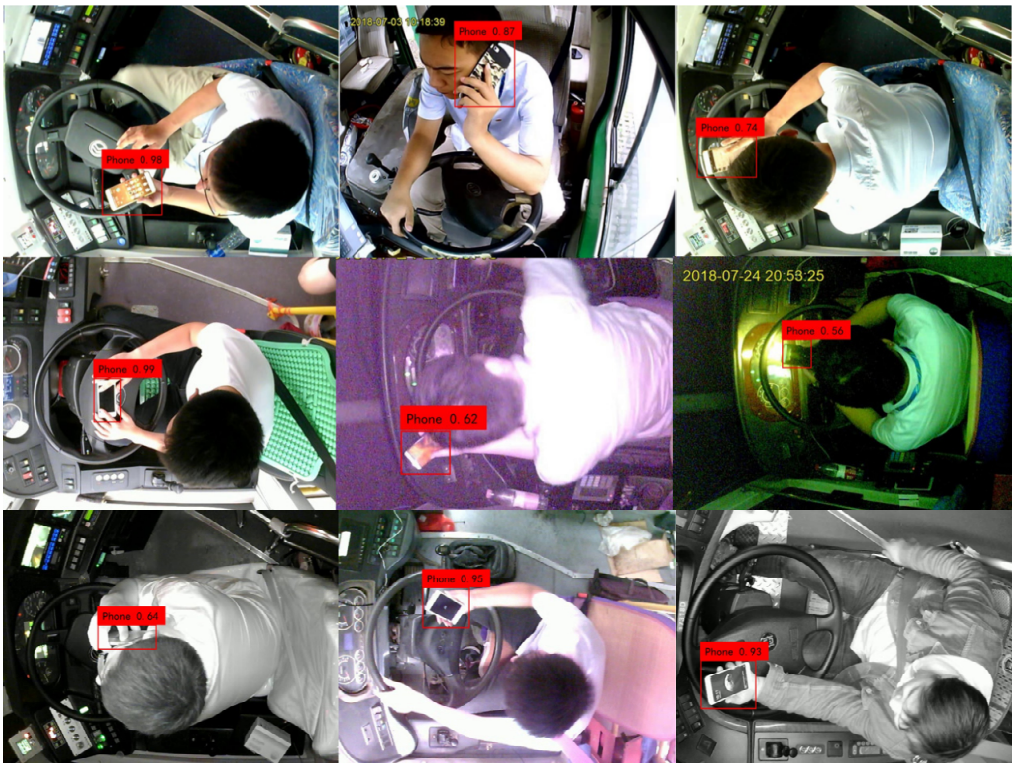


图 6 复杂的驾驶场景下的检测结果

Fig. 6 Detection results in complex driving scenarios

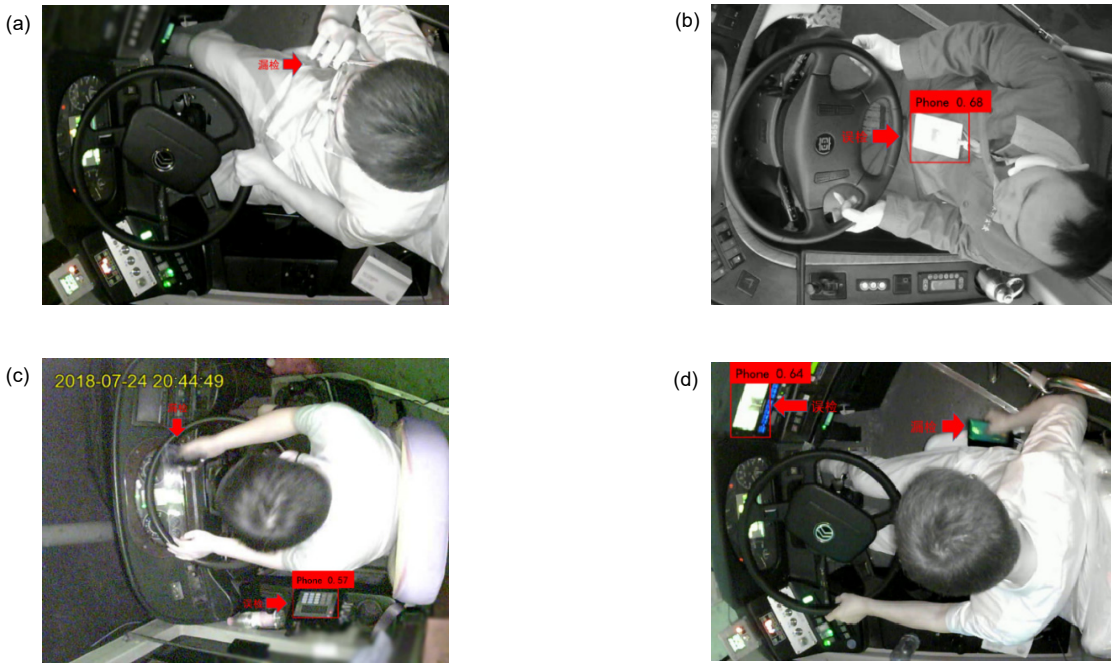


图 7 异常检测

Fig. 7 Abnormal detection

图 7 展示的是在实验过程中出现的异常检测。图 7(a)展示的是漏检，图 7(b)展示的是误检，图 7(c)和图 7(d)展示的是同时漏检和误检。出现这一类问题的原因在于图像质量差、手机被严重遮挡、光照影响以及高度相似物体的混入。这些异常检测问题后续可以通过两方面来解决和优化。一方面是通过改善外部条件，如提高图片分辨率、数据集平衡扩增、驾驶环境中减少不必要的高度相似物体；另一方面则是提高自身模型的健壮性，如困难样本挖掘、改进损失函数、优化网络等。

5 总 结

本文基于 YOLOv4 算法，提出了一种端到端的神经网络 OMPDNet 用于驾驶场景下的手机检测。同时，为了评估网络模型性能，本文制作了一个新的数据集 OMPDDataset 来进行验证。实验结果表明，本文方法在精确率、召回率和平均精确率上分别达到了 89.7%、96.1%和 89.4%，在速度上达到了每秒 72.4 帧，优于当前几种主流目标检测算法，并且更为贴切工程中的应用，能为交管部门贡献一份力量。本文算法不仅适用于手机检测，该思路亦可拓展到深度学习小目标检测的问题。在今后的工作中，将继续改进算法，泛化其性能。

参考文献

[1] Rodríguez-Ascariz J M, Boquete L, Cantos J, *et al*. Automatic system for detecting driver use of mobile phones[J]. *Transp Res C Emergi Technol*, 2011, **19**(4): 673–681. doi: 10.1016/j.trc.2010.12.002.

[2] Leem S K, Khan F, Cho S H. Vital sign monitoring and mobile phone usage detection using IR-UWB radar for intended use in car crash prevention[J]. *Sensors (Basel)*, 2017, **17**(6): 1240. doi: 10.3390/s17061240.

[3] Berri R A, Silva A G, Parpinelli R S, *et al*. A pattern recognition system for detecting use of mobile phones while driving[C]//*Proceedings of the 9th International Conference on Computer Vision Theory and Applications*, 2014: 411–418. doi: 10.5220/0004684504110418.

[4] Cortes C, Vapnik V. Support-vector networks[J]. *Mach Learn*, 1995, **20**(3): 273–297.

[5] Xiong Q F, Lin J, Wei Y, *et al*. A deep learning approach to driver distraction detection of using mobile phone[C]//*2019 IEEE Vehicle Power and Propulsion Conference*, 2019: 1–5. doi: 10.1109/VPPC46532.2019.8952474.

[6] Shi X P, Shan S G, Kan M N, *et al*. Real-time rotation-invariant face detection with progressive calibration networks[C]//*2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 2295–2303.

[7] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[Z]. arXiv: 2004.10934, 2020.

[8] Bishop C. *Pattern Recognition and Machine Learning*[M]. New York: Springer-Verlag, 2006.

[9] Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. *Biol Cybern*, 1980, **36**(4): 193–202.

[10] Lecun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning

applied to document recognition[J]. *Proc IEEE*, 1998, **86**(11): 2278–2324. doi: 10.1109/5.726791.

[11] Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks[C]//*Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//*ICLR*, 2015.

[13] Howard A G, Zhu M L, Chen B, *et al*. Mobilenets: efficient convolutional neural networks for mobile vision applications[Z]. arXiv: 1704.04861, 2017.

[14] Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions[C]//*2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015: 1–9.

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

[16] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//*Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001. doi: 10.1109/CVPR.2001.990517.

[17] Viola P, Jones M J. Robust real-time face detection[J]. *Int J Comput Vis*, 2004, **57**(2): 137–154.

[18] 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.

[19] Felzenszwalb P F, Girshick R B, McAllester D, *et al*. Object detection with discriminatively trained part-based models[J]. *IEEE Trans Pattern Anal Mach Intell*, 2010, **32**(9): 1627–1645. doi: 10.1109/TPAMI.2009.167.

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

[21] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. *IEEE Trans Pattern Anal Mach Intell*, 2015, **37**(9): 1904–1916.

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

[23] 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.

[24] Dai J F, Li Y, He K M, *et al*. R-FCN: Object detection via region-based fully convolutional networks[C]//*Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016: 379–387.

[25] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 779–788.

[26] Redmon J, Farhadi A. Yolo9000: Better, faster, stronger[C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017: 6517–6525.

[27] Redmon J, Farhadi A. YOLOv3: An incremental improvement[Z]. arXiv: 1804.02767, 2018.

[28] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot MultiBox detector[C]//*European Conference on Computer Vision*, 2016: 21–37.

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

[30] Hu P, Ramanan D. Finding tiny faces[Z]. arXiv: 1612.04402, 2016.

[31] Wang C Y, Liao H Y M, Wu Y H, *et al*. CSPNet: a new backbone that can enhance learning capability of CNN[C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020**:** 1571–1580.

[32] Liu S, Qi L, Qin H F, *et al*. Path aggregation network for instance segmentation[C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 8759–8768.

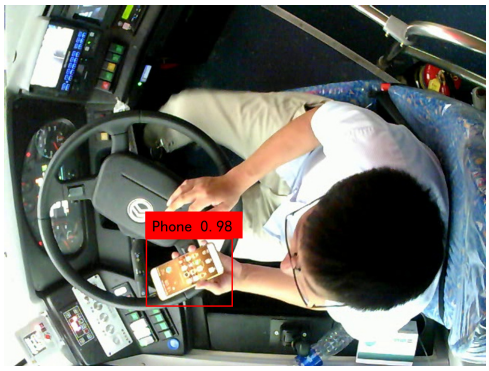
[33] Duan K W, Bai S, Xie L X, *et al*. CenterNet: keypoint triplets for object detection[C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019: 6569–6578.

[34] Zhang H Y, Cisse M, Dauphin Y N, *et al*. mixup: beyond empirical risk minimization[Z]. arXiv: 1710.09412, 2017.

An end-to-end neural network for mobile phone detection in driving scenarios

Dai Teng[1,2], Zhang Ke[1,2], Yin Dong[1,2*]

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



Detection results in complex driving scenarios

Overview: Real-time detection of small objects is always a difficult problem in the field of image processing. It has the characteristics of low resolution and difficult detection, which often leads to missed detection and false detection. In this paper, based on the deep learning target detection algorithm, an end-to-end neural network is proposed for small target detection like mobile phone in complex driving scenes. Firstly, in order to maintain a high accuracy rate and ensure real-time performance, this paper improves the YOLOv4 algorithm and designs an end-to-end small target detection network (OMPDNet) to extract image features. Secondly, setting an appropriate size of Anchor is conducive to improving the convergence speed and accuracy of the model. Meanwhile, based on K-means, this paper presents a clustering algorithm K-means-Precise, which is more suitable for the distribution of sample data. It is used to generate anchors suitable for small target data, so as to improve the efficiency of the network model. Finally, a data set (OMPD Dataset) is made by using supervision and weak supervision method to make up for the lack of public data set in specific driving scenes. It is composed of shooting videos from the in-car monitoring camera, a small number of public data sets and internet pictures. And more, in order to solve the problem of imbalance between positive and negative samples, negative samples are added to the data set for training in the paper. The experimental results on OMPD Dataset show that K-means-Precise can slightly improve the accuracy of the model. But importantly, it converges five cycles ahead of time. The overall detection of the network model is evaluated by the accuracy rate, recall rate and average accuracy rate, which are 89.7%, 96.1% and 89.4% respectively, and the speed reaches 72.4 frames per second. It shows that in the complex driving scene experiments, the OMPDNet proposed in this paper can not only effectively complete the detection task of drivers using mobile phones while driving, but also has certain advantages in accuracy and real-time performance of small target detection compared with current popular algorithms. Especially, in the practical engineering application, real-time is more important, which can recognize the behavior while driver playing mobile phone to reduce the occurrence of traffic accidents, and be benefit to the traffic management department. Our proposed method is not only suitable for mobile phone detection, but also can be extended to small target detection problems in the field of deep learning. In the future work, we will continue to improve the algorithm and generalize its performance.

Dai T, Zhang K, Yin D. An end-to-end neural network for mobile phone detection in driving scenarios[J]. *Opto-Electron Eng*, 2021, **48**(4): 200325; DOI: 10.12086/oee.2021.200325

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