面向行车视频目标实时检测的轻量级SSD网络

张琳娜[1] 陈建强[1] 陈晓玲[1] 岑翼刚[2] 阚世超[2]
(贵州大学机械工程学院 贵阳 550025)[1] (北京交通大学计算机与信息技术学院 北京 100044)[2]

摘 要 车辆和行人检测是高级辅助驾驶(ADAS)中最基本也是研究最广泛的内容,而深度学习算法是当前性能最好的目标检测算法。然而,深度学习算法的计算量非常大,通常需要高性能的GPU显卡才能快速运行。在实际使用中,目标检测算法一般要求集成到车辆硬件系统中,因此算法对硬件资源的要求不能太高。基于SSD网络,提出一种轻量级的SSD网络,用于实时目标检测。通过减小输入图像的大小以及全连接层节点数量,减少网络复杂度,提升目标实时检测速度。计算量减少将导致检测车辆和行人的准确率下降,因此提出多级损失函数监督训练方法,来解决输入图像缩小而引发的图像损失及在反向传播过程中不能有效更新VGG中浅层卷积层参数等问题。此外,提出一种基于多尺度图像分块的训练数据集扩充方法,以解决图像缩放产生的形变及图像缩小后目标可能消失的问题。实验结果表明,采用所提出的轻量级SSD网络,不但实现了笔记本电脑上的车辆和行人检测的实时性,也保持了检测准确率。对比其他目标检测算法,优化后的网络对行车视频中车辆和行人的检测速度优于其他算法,且在获得相同准确率的同时消耗的电量更少。

关键词 目标检测,深度学习,SSD,高级辅助驾驶,卷积神经网络

中图法分类号 TP391.44 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.035

Lightweight SSD Network for Real-time Object Detection in Automotive Videos

ZHANG Lin-na[1] CHEN Jian-qiang[1] CHEN Xiao-ling[1] CEN Yi-gang[2] KAN Shi-chao[2]
(School of Mechanical Engineering, Guizhou University, Guiyan 550025, China)[1]
(School of Computer Science & Information Technology, Beijing Jiaotong University, Beijing 100044, China)[2]

Abstract Vehicle and pedestrian detection are the most basic and widely studied subjectin the field of advanced driver-assistance systems (ADAS). At present, deep learning achieved the best detection performance for object detection. However, the computational cost of deep learning algorithms is very high and the algorithms often require high performance GPU. In the real applications, object detection algorithm is required to be integrated into the vehicle hardware system. So the requirement of the hardware for the algorithm can not be too high. Based on the SSD network, a lightweight SSD network was proposed for real-time objection. By resizing the input images into a smaller size and significantly reducing the node number of the fully connected layer, the network complexity could be reduced. In addition, the object detection speed was improved. A supervised training method based on the multi-stage loss function was proposed to solve the problems of image deformation and the updated parameters in the VGG low layers caused by the shrink of the input images. Furthermore, because the detection accuracy of vehicles and pedestrians would be declined after the reduction of calculations, a hierarchical image partition method was proposed to expand the training dataset, which was able to solve the object vanishing problem caused by the image shrink. Experimental results show that the proposed lightweight SSD network not only realizes real-time vehicle and pedestrian detection on a laptop, but also maintains the detection accuracy. Compared with other object detection algorithms, the optimized network achieves faster detection speed for the vehicles and pedestrians. Also, the power consuming of the laptop is reduced significantly while the detection accuracy is the same.

Keywords Object detection, Deep learning, SSD, Advanced driver-assistance systems, Convolutional neural network

到稿日期:2018-06-18 返修日期:2018-09-26 本文受贵州省自然科学基金(黔科合基础[2019]1064),国家自然科学基金(61872034),广州市科技计划项目(201804010271),广东省自然科学基金(2016A030313708)资助。

张琳娜 女,硕士,讲师,主要研究方向为计算机视觉、机械故障诊断;**陈建强** 男,硕士,副教授,主要研究方向为机械设计、故障诊断;**陈晓玲** 女,硕士,讲师,主要研究方向为机械设计、故障诊断;**岑翼刚** 男,博士,教授,主要研究方向为计算机视觉、图像处理、信号处理,E-mail:ygcen@bjtu.edu.cn(通信作者);**阚世超** 男,博士生,主要研究方向为深度学习、计算机视觉。

1 引言

目标检测是高级辅助驾驶及自动驾驶任务中的重要环节,尤其是对车辆和行人的检测。深度学习的快速发展使得实时目标检测基本得到实现,在检测准确率上逐渐逼近人眼识别准确率。然而,当前基于深度学习的目标检测大都在高性能 GPU 上实现,其实际应用存在一定困难。一方面,高性能 GPU 的成本高昂;另一方面,高性能 GPU 的电量消耗大,自动驾驶车辆上的供电系统通常无法满足该系统的耗电需求,即使采用逆变器也难以持续为支持程序运行系统的服务器供电。因此,在性能较低的 GPU 上实现实时目标检测、且尽可能多地减少计算量、保证准确率是实际应用中迫切需要解决的问题。

近年来,基于深度学习的通用目标检测框架取得了飞速发展[1-13],从早期的 R-CNN[1] 和 Fast R-CNN[2],到后来的 Faster R-CNN[3],SSD[4] 和 Yolo[5] 等,其发展特点主要是在保证目标检测准确率的情况下不断提高检测速度,或者以提高目标检测准确率为目的进行改进。为了提高目标检测的性能,Dai 等[8] 在 Faster R-CNN 的基础上提出基于区域的全卷积网络 R-FCN。Redmon 等[9] 在 Yolo 的基础上进行改进,提出能以更快的速度更好地检测更多类别的 Yolo9000。Kim 等[10] 基于 Faster R-CNN 框架,提出了轻量级网络 PVANet 实现目标检测的加速,但在 Titan X 显卡上仅获得 46ms 的检测速度。为了增强目标检测对物体形变的鲁棒性,Dai 等[11] 提出可形变的卷积网络,获得了当前最好的目标检测效果。此外,Huang 等[12] 使用一致的标准评价了当前主流的目标检测框架和网络在目标检测任务中的性能。Kang 等[13] 借鉴 Faster R-CNN 的思想提出视频目标检测算法,考虑连续多帧的视频卷积,获得了当前最好的视频目标检测性能。

从目标检测技术的发展来看,主要是通过共享卷积神经网络中的卷积计算以及减少目标候选框的数量来实现加速,且大部分目标检测技术都是在 Faster R-CNN 的基础上改进而得。此外,目标检测框架基本都是基于现有网络实现目标检测,例如 AlexNet 和 VGG[6] 等网络,并且需要依赖较大的输入图像才能获得较高的检测准确率。虽然网络的大小固定,网络的参数数量也固定,但输入图像越大,需要计算的卷积次数就越多,对单张图像检测所做的计算量更大,因而消耗的能量更多。

在高级辅助驾驶场景中,需要检测的目标一般是车辆附近的目标,相比任意场景而言,高级辅助驾驶场景中的目标大小相对固定,使得在算法优化方面和模型训练方面都具有一定的提升空间。本文针对高级辅助驾驶场景,基于 SSD 目标检测框架进行优化,提出一种轻量级的 SSD 网络结构及训练方法。提出多级损失函数监督训练,采用多尺度空间金字塔扩充训练数据集,在保证最终检测准确率的同时减少计算量,从而减少能量消耗。在实际场景检测中,最终优化所得模型在检测准确率、实时性和耗电量方面的综合性能优于现有主流模型。

2 基于 SSD 的目标检测

图 1 是基于 VGG 网络的 SSD 目标检测框架图,Data 为输入数据,Conv* 为 VGG 网络的卷积层,FC6 和 FC7 为 VGG 网络的全连接层,Conv6_*,Conv7_*,Conv8_*,Conv9_* 是在 VGG 网络之后添加的卷积层,Box_loc 是定位目标位置的坐标,Box_conf 是目标所属类别的概率,Box_priorbox 是目标预选框。

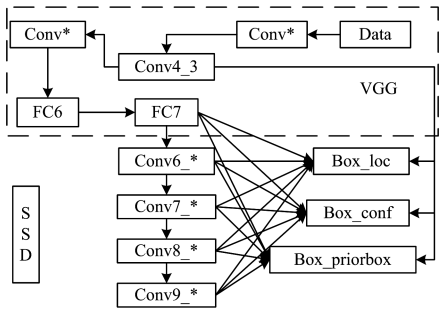

图 1 SSD 目标检测框架

Fig. 1 SSD object detection framework

图 1 中,VGG 网络的某一个中间卷积层(Conv4_3)和最后一个全连接层(FC7)的特征被直接用于最终的目标识别和定位。由于卷积神经网络主要采用卷积和池化方式,利用分层思想逐层提取图像从低级别到高级别的特征。浅层特征主要包含物体的局部特征,而高层特征主要包含物体的形状和轮廓等特征。在目标检测中,对于占图像面积比例小于阈值 α 的目标而言,若网络层级太深,池化操作过多,可能会造成关键信息的丢失,而浅层网络的特征能包含所有目标的关键信息。直接利用卷积神经网络中的某一个卷积层的特征参与最终目标识别和定位的主要优势是在阈值 α 较小时,仍然能检测并识别出目标。同时,在图 1 中还可以看到,在 VGG 网络之后,基于全连接层 FC7,又接了 4 个级联的卷积层,这 4 个卷积层的特征依然被用于参与最终的目标检测和识别。卷积神经网络中,全连接层包含的信息主要是目标的高层次特征,通常可用于进行准确的分类目标。而最后 4 个级联的卷积层主要用于从全连接层 FC7 中提取有用信息参与目标定位和识别,使得检测更加准确。

在当前基于 SSD 的目标检测中,网络输入图像的尺寸最小限定为 300×300,同时 VGG 网络最后两个全连接层 FC6 和 FC7 的节点数量都为 4 096,这需要一次性存储和计算 4 096×4 096 的权重矩阵。即检测一张图像,在全连接层需要进行1 678 万次浮点运算,大量的计算资源会被耗费在全连接层的计算上。一种可减少计算量的改进方式是使用 1×1 的卷积核进行卷积运算,代替全连接运算以减少计算量,这种方式在 SSD 框架中已经被采用。

本文主要检测高级辅助驾驶场景中的车辆和行人,在高级辅助驾驶场景下,摄像头被固定在车辆上,通常检测范围在几十米内,对比任意场景下的目标检测,该场景下的车辆和行人目标相对较大。SSD 利用卷积神经网络较浅的卷积层的特征参与目标检测和识别,因此网络中 300×300 的输入图像在

检测准确率方面基本能满足实际需求。在实际场景中，基于 300×300 的输入图像进行检测，在装配 GTx1060 显卡的笔记本电脑上能实时检测目标，且检测准确率超过 90%。然而，其耗电量非常大，在笔记本电脑电量饱和的情况下仅能运行 30 分钟。为了减少计算量，进而减少电量消耗，一方面，本文采用 200×200 的输入图像进行目标检测。较小的输入图像使得浮点计算的次数更少。假设卷积核大小为 3×3，移动步长为 1，则对一张 300×300 的特征图进行卷积需要进行约 80 万次浮点运算，而对一张 200×200 的特征图进行卷积只需要进行约 35 万次浮点运算，可减少超过一半的计算量。另一方面，本文将 VGG 网络的两个全连接层 FC6 和 FC7 的节点数量都设置为 128，该操作可减少 1024 倍的全连接层参数数量。实际场景中，通过上述改进得到的轻量级 SSD 网络目标检测框架在电量饱和的装配 GTx1060 显卡的笔记本电脑上能运行超过 1 小时，电量消耗减少了 50%以上。同时，目标检测速度也比 300×300 的输入图像快一倍以上。

上述轻量级 SSD 网络由于减小了输入图像的尺寸以及全连接层节点的数量，目标检测准确率会随之降低。为此，基于 VGG 网络，针对 200×200 的输入图像，提出了一种多尺度图像分块的数据集扩增和多级损失函数监督训练方法，使得在输入图像大小为 200×200，以及减少 1024 倍全连接层参数数量的情况下，依然与 300×300 的输入图像所得准确率一致。

3　基于多尺度图像分块的训练数据集扩增方法

多尺度图像分块[14]是一种二分图像的空间金字塔图像分块方法，能有效划分图像中的目标物体。一般而言，卷积神经网络中训练数据集的种类越多，在测试集上的鲁棒性就越强。为了有效扩充训练数据集，本文采用多尺度分块方法对图像进行分块。训练数据集图像的分辨率通常都大于 640×480，有的甚至是高分辨率图像，尺寸为 1920×1 080，如果直接将高分辨率图像缩放到 200×200，则需要将原始图像缩小 52 倍，即使是 640×480 的图像缩放到 200×200，也需要缩小 7.7 倍。在测试时使用的摄像机分辨率是固定的，本文所用的摄像机的采样图像的分辨率为 1 280×720。在自动驾驶场景中，待检测和识别的大部分车辆和行人距摄像头都较近，因而目标在图像中的面积占比较大，即使缩放后也容易区分出目标来。我们采用公开的 KITTI 车辆检测图像数据集[15]作为训练集，该数据库中图像尺寸均约为 1 224×370，且所有图像的比例也相近。如图 2 所示，如果将该数据库中的图像直接缩放为 300×300 或 200×200，则目标物体会产生较大的形变。此外，该数据集中图像目标的大小并不一致，有的可能非常小，图像缩小后目标可能会消失，这样的训练数据集对网络的训练并没有帮助。

采用金字塔式的多尺度图像分块有助于解决图像非正方形而导致的缩放形变问题。本文提出的多尺度图像分块算法的步骤如下：

(1)首先，以图像的短边边长 l 作为滑动窗口边长，以 $l/2$ 作为滑动步长，在图像上滑动，从而得到初始的 n 个矩形框，设图像长边边长为 L，则 $n=\lceil 2L/l \rceil-1$，其中$\lceil\cdot\rceil$为向上取整，如果最后一块滑动时不够步长 $l/2$，则滑到边界处停止。

(2)训练数据集的图像中目标已经被手动标记了矩形框，因此需要计算每个滑动窗口中所包含的目标的矩形框与原始图像中目标的矩形框的重合面积比例，用交并比 IOU 表示，如果 IOU 小于一定阈值，则进入下一步。否则，修正滑动块中的目标矩形框，使得矩形框刚好包含滑动块中的目标，保存该滑动块图像及其对应的目标矩形框参数，用于监督训练。如果该滑动块中没有包含目标，则直接丢弃该滑动块。

(3)调整第 2 步中 IOU 大于一定阈值但不是 1 的滑动块，扩展其边界使其包含全部目标。然后重新计算滑动块中目标矩形框与原始图像中目标矩形框的 IOU，转步骤(2)。

(4)将 l 的值减半，转步骤(1)。直到 $l<100$ 时停止迭代。

通过上述方法处理原始训练数据集后，会使得训练数据集增加，同时扩增后的训练数据集对卷积神经网络中的图像缩放不会产生较大形变，能被直接用于训练卷积神经网络和目标检测算法。由于本文中图像经多尺度空间金字塔图像分块处理，以及在 SSD 框架中使用了 VGG 网络中的 Conv4_3 层的特征图，因此在网络输入图像尺寸为 200×200 的情况下，我们的算法依然能保持高级辅助驾驶场景下的车辆和行人的检测准确率。

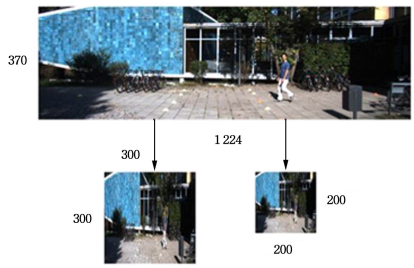

图 2　KITTI 数据库图像和缩放对比

Fig. 2　KITTI database image and scale comparison

4　多级损失函数监督训练

目标检测一般都是联合分类和定位损失共同训练网络，归纳如下：

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\beta L_{loc}(x,l,g)) \tag{1}$$

其中，$x\in\{0,1\}$，x 为 1 表示算法求得的矩形框的定位参数和对应的分类概率参与计算损失，x 为 0 表示其不参与计算损失，c 是矩形框的分类概率，x 和 c 的具体含义见式(2)，l 和 g 的具体含义见式(3)。L_{conf} 是分类损失，通常使用 softmax 损失，在图 1 中通过 Box_conf 和 Box_priorbox 的输出计算得到。softmax 损失的计算方法为：

$$L_{conf}(x,c)=-\sum_{i\in Pos}^{N}x_{ij}^{p}\log(\hat{c}_i^p)-\sum_{i\in Neg}x_{ij}^{p}\log(\hat{c}_i^o) \tag{2}$$

其中，$\hat{c}_i^p=\exp(c_i^p)/\sum_p\exp(c_i^p)$，为 Box_conf 的输出概率，$x_{ij}^p\in\{0,1\}$，取值为 1 表示 Box_priorbox 的第 i 个输出与真实标记的第 j 个类别为 p 的矩形框相匹配，否则为 0，N 用于统

计值为 1 的 Box_priorbox 的输出个数。

式(1)中,β 为定位损失的权重,通常设置为 1,L_{loc} 为网络输出的定位目标矩形框与真实标记的目标矩形框之间的损失。通常使用 smooth L_1 损失,计算方式为:

$$L_{loc}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}smooth_{L_1}(l_i^m-\hat{g}_j^m) \quad (3)$$

其中,l_i^m 为 Box_loc 的输出矩形框参数与 Box_priorbox 的输出矩形框参数之差的归一化值,而 $\hat{g}_j^m$ 为真实标记的矩形框参数与 Box_loc 的输出矩形框参数之差的归一化值,cx 和 cy 为目标矩形框的中心点坐标,w 和 h 为目标矩形框的宽和高。$smooth_{L_1}(x)$ 函数的定义为:

$$smooth_{L_1}(x)=\begin{cases}0.5x^2, & \text{if } |x|<1\\ |x|-0.5, & \text{otherwise}\end{cases} \quad (4)$$

式(4)的主要作用是使矩形框的两个坐标值之差($l_i^m-\hat{g}_j^m$)变得更加平滑,从而使得参数回归的区间更小,有利于基于梯度下降法的坐标回归。此外,由于矩形框坐标值位于区间[0,1],因此它们的残差位于区间[-1,1],平方操作和绝对值操作是将坐标值的残差规范到区间[0,1],二次项系数前的参数 0.5 是缩放因子,$|x|-0.5$ 是将残差约束到区间[0.5,1)。

当输入图像从 300×300 缩小为 200×200 后,SSD 中全连接层及其之后卷积层的特征通常会出现信息损失,因此直接在网络的最后使用损失函数计算得到的损失在反向传播过程中并不能有效更新 VGG 中较浅的卷积层的参数[16]。因此我们利用 Conv4_3 的信息直接通过式(1)计算损失,通过反向传播更新 VGG 网络中 Conv4_3 层及其之前的信息,而位于 Conv4_3 之后的层的网络参数的更新则由最终的损失进行更新,本文提出的监督损失如图 3 所示。

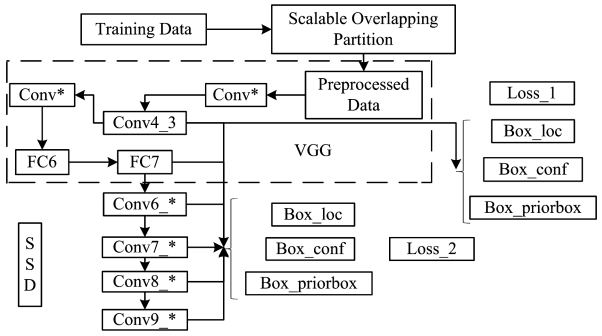

图 3 本文提出的训练思想

Fig. 3 Training ideas proposed in this paper

图 3 中 Loss_1 计算得到的损失只用于更新 Conv4_3 层及其之前的网络参数,Loss_2 计算得到的损失只用于更新 Conv4_3 层之后的网络参数。Loss_1 和 Loss_2 都是通过计算式(1)得到的。上述采用不同层的信息计算损失,用于反向传播更新网络参数的优势为:

(1)避免梯度过小和反向传播过程中梯度消失导致较浅层的参数无法更新;

(2)网络较浅层的信息通常包含图像较细节的特征,因此能利用更丰富的信息进行有效监督训练。

5 实验结果及分析

为了验证本文方法的有效性,我们通过实验对所提出的方法进行量化分析,在车上安装可变焦距摄像机采集不同道路上的视频标注后进行测试。收集的 3 段视频包括高速路、普通道路和城区道路视频。在每段视频中取 10 min 标注用于测试,被检测的目标有 car,bus,truck 和 pedestrian。摄像机采样速率为 25fps,在收集的视频中每隔 10 fps 抽取一帧标注真实矩形框。每个视频得到 1500 张测试图像,3 个视频一共 4500 张测试图像。测试样本中每一类所包含的图像数量分别为 4126,973,2341,869。实验中使用平均检测准确率[17]作为评价标准,并与 Faster R-CNN[3,18],R-FCN[8],YOLO-9000[9],SSD-300[4] 进行对比。

实验中训练集采用 KITTI 目标检测数据[9],采用提出的基于多尺度图像分块的训练数据集扩增方法处理训练集后分别用于训练 Faster R-CNN,YOLO-9000,R-FCN,SSD-300 和本文提出的 SSD-200。其中,Faster R-CNN 的训练次数为 8 万次,交替训练 2 轮,YOLO-9000 的训练次数为 12 万次,R-FCN 的训练次数为 16 万次,SSD-300 和 SSD-200 的训练次数为 24 万次。在训练时,分别对 VGG 网络的参数采用 VOC 数据集在 Faster R-CNN,YOLO-9000,R-FCN 和 SSD-300 上训练的结果进行初始化。训练平台采用配置有两块 Titan1080 显卡的服务器进行训练,测试时采用配置有一块 GTx1060 显卡的笔记本电脑进行测试,最终的测试结果如表 1 所列。表中检测速度包含了视频图像的输入、输出和显示的总时间,摄像机采集的视频帧图像尺寸为 1280×720,测试时截取图像中间 720×720 大小的图像块缩放到对应比例进行测试(Faster R-CNN 为 224×224,YOLO-9000 为 416×416,R-FCN 为 224×224,SSD-300 为 300×300,SSD-200 为 200×200)。

表 1 不同方法检测结果的对比

Table 1 Comparison of test results by different methods

方法	car /%	bus /%	truck /%	pedestrian /%	平均准确率/%	检测速度/fps
Faster R-CNN[3]	97	95	94	92	94.5	2
R-FCN[8]	97	96	95	93	95.25	3
YOLO-9000[9]	91	88	89	85	88.25	28
SSD-300[4]	95	90	91	89	91.25	15
SSD-200 本文方法	96	90	92	88	91.5	32

从表 1 可以看出,Faster R-CNN 和 R-FCN 对于每个待检测的目标在测试集上都得到高于其他方法的准确率,但是检测速度却远远比不上 SSD。在本实验使用的笔记本电脑上 1 s 才能检测 2 帧或 3 帧,这在实际应用中远远达不到实时性要求。SSD-300 对于行人的检测准确率略高于本文优化后的 SSD-200,这是因为相对于车来说,行人目标较小,较大的缩放可能会造成部分信息丢失。对比 SSD-300,本文采用的多级损失函数在一定程度上使 SSD-200 的检测准确率得到提升,其中,对 car 的检测准确率比 SSD-300 高 1%,对 truck 的

检测准确率也比 SSD-300 高 1%，最终在测试集上得到的检测准确率(91.5%)与 SSD-300(91.25%)相当。

从速度方面看，SSD-300 为 15 fps，实际应用中如果对视频每帧都进行检测，则无法满足实际应用需求，而 SSD-200 为 32 fps，能满足实际应用需求。图 4 是 SSD-200 在测试图像上的部分检测结果。虽然 YOLO-9000 也能满足实时性要求，但是速度和准确率都不及 SSD-200。

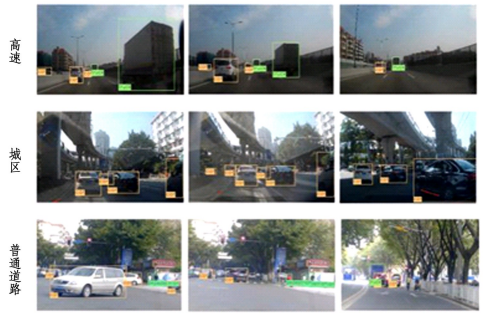

图 4　使用 200×200 的输入进行检测的部分结果

Fig. 4　Somedetection results by using 200×200 input images

结束语　本文基于 SSD 目标检测框架，对行车视频中的车辆和行人进行检测，通过优化 SSD-300，采用 200×200 的输入图像，最终在装配 GTx1060 显卡的笔记本电脑上实现了实时目标检测。提出多损失联合训练 SSD-200 以及基于多尺度图像分块的训练数据集扩增方法增加训练数据，SSD-200 获得与 SSD-300 相当的检测准确率，且检测速度比 SSD-300 快 1 倍以上，通过车载摄像头实测验证了 SSD-200 的实时性以及对高级辅助驾驶场景中的车辆和行人检测的有效性。

参 考 文 献

[1] GIRSHICKR B，DONAHUE J，DARRELL T，et al. Region-based convolutional networks for accurate object detection and segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，38(1)：142-158.

[2] GIRSHICKR B. Fast R-CNN[C]// 2015 IEEE International Conference on Computer Vision，ICCV 2015. Santiago，Chile，2015：1440-1448.

[3] REN S，HE K，GIRSHICKR B，et al. Faster R-CNN：towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(6)：1137-1149.

[4] LIU W，ANGUELOV D，ERHAN D，et al. SSD：single shot multibox detector[C]// Computer Vision-ECCV 2016-14th European Conference，Amsterdam，The Netherlands，2016：21-37.

[5] REDMON J，DIVVALAS K，GIRSHICKR B，et al. You only look once：unified，real-time object detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition，CVPR 2016. Las Vegas，NV，USA，2016：779-788.

[6] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// International Conference on Learning Representations. San Diego，USA，2015：2015-2029.

[7] CHEN S，PEI H，LAI Q，et al. Multitarget Tracking Control for Coupled Heterogeneous Inertial Agents Systems Based on Flocking Behavior[J]. IEEE Transactions on Systems Man & Cybernetics Systems，2018，PP(99)：1-7.

[8] 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 29：Annual Conference on Neural Information Processing Systems 2016. Barcelona，Spain，2016：379-387.

[9] REDMON J，FARHADI A. YOLO9000：better，faster，stronger [C]// IEEE Conference on Computer Vision and Pattern Recognition，CVPR 2017. Honolulu，HI，USA，2017：6517-6525.

[10] KIM K H，HONG S，ROH B，et al. PVANET：deep but lightweight neural networks for real-time object detection[J]. arXiv：1608.08021.

[11] DAI J，QI H，XIONG Y，et al. Deformable convolutional networks[J]. CoRR，abs/1703.06211，1(2)，3.

[12] HUANG J，RATHOD V，SUN C，et al. Speed/accuracy trade-offs for modern convolutional object detectors[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，Hawaii，USA，2017：3296-3297.

[13] KANG K，LI H，XIAO T，et al. Object detection in videos with tubelet proposal networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，Hawaii，USA，2017：889-897.

[14] KAN S C，CEN Y G，CEN Y，et al. SURF binarization and fast codebook construction for image retrieval [J]. Journal of Visual Communication & Image Representation，2017，49：104-114.

[15] GEIGER A. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// IEEE Conference on Computer Vision and Pattern Recognition，Providence，RI，USA，2012：3354-3361.

[16] YUAN Y，YANG K，ZHANG C. Hard-aware deeply cascaded embedding[C]// IEEE International Conference on Computer Vision，Venice，Italy，2017：814-823.

[17] EVERINGHAM M，GOOL L，WILLIAMS C K，et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision，2010，88(2)：303-338.

[18] LI H，HUANG Y，ZHANG Z. An improved Faster R-CNN for same object retrieval [J]. IEEE Access，2017，5：13665-13676.