第 58 卷 第 6 期/2021 年 3 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于 SSD 的行人鞋子检测算法

耿鹏志，杨智雄，张家钧，唐云祁*

中国人民公安大学侦查学院，北京 100038

摘要 基于鞋样的视频追踪技术是公安机关刑事侦查的一种常用技战法，在公安实战中起到巨大的作用，然而该项技术大量依赖于人工筛选，工作量大且效率低，容易出现漏检的状况。鉴于此，提出一种基于 SSD(Single Shot MultiBox Detector)模型的鞋子自动检测算法，实现对行人鞋子的自动检测与定位。首先对 SSD 模型的结构和先验框参数进行设计，使其符合鞋子检测的实际应用。然后采用调节网络参数的方法提高网络的检测性能和稳定性，完善适用于鞋子检测的网络模型和方法，最终得到准确且高效的单类别鞋子检测网络。最后在课题组前期构建的鞋样本数据库中进行性能评价。实验结果表明，所提算法的平均精度达到 0.891。

关键词 图像处理；鞋子检测；卷积神经网络；SSD；视频侦查

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.0610009

Pedestrian Shoes Detection Algorithm Based on SSD

Geng Pengzhi, Yang Zhixiong, Zhang Jiajun, Tang Yunqi*

School of Criminal investigation, People's Public Security University of China, Beijing 100038, *China*

Abstract Video tracking technology based on shoe patterns is commonly used by public security organizations in criminal investigations and is crucial in actual public security combat. However, this technology relies heavily on manual screening with a heavy workload and low efficiency, and it is prone to missed inspections. In view of this, an automatic shoe detection algorithm based on Single Shot MultiBox Detector (SSD) model is proposed herein to achieve automatic detection and positioning of pedestrian shoes. First, the structure of the SSD model and the parameters of the prior frame are designed to meet the practical application of shoe detection. Then, the method for adjusting network parameters is used to improve the detection performance and stability of the network and the network model and method suitable for shoe detection are improved. Thus, an accurate and efficient single-category shoe detection network is obtained. Finally, performance evaluation in the shoe sample database constructed by the research group in the early stage is conducted. Experimental results show that the average accuracy of the proposed algorithm is 0.891.

Key words image processing; shoes detection; convolutional neural network; Single Shot MultiBox Detector; video investigation

OCIS codes 100.4996; 100.3008; 100.5010

1 引 言

基于鞋样的视频追踪技术是公安机关刑事侦查常用的技战法之一，该技战法通过获取嫌疑人在犯罪现场中遗留的鞋印来识别嫌疑人在作案过程中所穿的鞋型，从而得到该鞋型的外观图像，在案发地周边的监控视频中依据所得的外观图像来检索嫌疑人影像。一般情况下，嫌疑人可能会对面部和穿着进行一定程度的伪装以逃避侦查，但是鲜有人对所穿鞋型进行伪装。基于鞋样的视频追踪技术的实战效

收稿日期：2020-06-08；**修回日期**：2020-07-02；**录用日期**：2020-08-31

基金项目：国家重点研发计划(2017YFC0822003)、中国人民公安大学“公共安全行为科学研究与技术创新专项”项目

* **E-mail**：tangyunqi@ppsuc.edu.cn

果极佳[1]，成功案例甚多。2015 年 1 月 15 日，在我国南部某地发生的两起技术开锁入室盗窃案中，现场的鞋印成为侦查破案的唯一线索，技术员通过对网上购物平台与全国公安机关鞋样本查询系统来识别和确定了两起案件中犯罪嫌疑人所穿的鞋型，进而结合时空信息从犯罪现场的周边监控视频中锁定了嫌疑人[2]。在上述案例中，侦查人员通过现场的鞋印实现了从物理空间到电磁空间的跨空间检索，有效缩小了排查范围，从而快速获得嫌疑人的影像信息。现阶段基于鞋样的视频追踪技术都是依靠人眼观察来实现的，工作量大，效率较低，实际工作中很有可能会出现遗漏或者错过最佳侦查时机的情况[3]。

目前我国警力严重不足，公安机关亟需基于鞋样的智能化视频追踪技术来实现从犯罪现场的鞋印到鞋子的图像，再到监控影像的全流程自动化比对检索。针对聚焦智能化视频追踪技术中的行人鞋子检测问题，本文提出一种基于 SSD（Single Shot MultiBox Detector）[4]的鞋子检测方法。首先针对鞋子的检测特性改进 SSD 模型结构，将深层的特征融合到浅层的特征中，扩大浅层特征的感受野；根据鞋子检测的实际情况设计先验框参数，进而细化网络参数，提高网络的检测性能和稳定性，得到准确且高效的单类别鞋子检测网络和方法。该方法可实现对行人穿着鞋子的自动检测，为基于视频图像的鞋型自动识别工作打下基础。

本文的主要贡献如下：构建一个行人鞋子检测的数据集，可用于鞋子检测及鞋型识别等研究工作；将目标检测应用于鞋子检测，提出一种基于 SSD 的鞋子检测方法，采用特征融合的方法对网络结构进行设计，同时根据行人鞋子的先验信息对先验框参数进行调整，提高网络的检测性能和稳定性。

2 相关工作

本文的相关研究工作主要应用在模式识别领域和公安领域。在公安领域，袁楚平等[3]将现场提取的图片与全国公安数据库进行对比查询，匹配到了两种足迹可能对应的鞋型，并结合相应的时空节点从视频监控中锁定了嫌疑人；许磊等[5]采用模拟实验的方法确定了视频中的可疑鞋与模拟鞋属于同一类型，实验得到的同一认定结论无疑为转变侦查方向提供了依据。但是袁楚平等[3]与许磊等[5]的实验需要依靠人工筛选、特征标示、拼接比较和重合比较等方法，这会耗费大量的时间和精力，效率较低，在实际办案中可能因为无法及时得到结果而延误时机。杨孟京等[6]将卷积神经网络运用到了鞋型的分类识别中，获得了很理想的结果，但是实验数据是由实验者模拟目标检测的过程手动切割而得。

鞋子检测在本质上属于模式识别领域中的目标检测问题。目标检测是一个对图片中的目标进行定位和分类的过程。传统的目标检测过程是通过手动选取特征[7]并结合滑动窗口对目标进行检测，但性能和准确度均不太理想。近年来，随着深度学习技术的发展和相应硬件设备性能的提升，目标检测算法取得了很大的进展。目前，比较流行的算法有两类。其中基于双阶段的目标检测算法通过 RPN（Region Proposal Network）[8]实现了端到端的目标检测，该算法的准确度较高，但速度慢，代表的模型有 RCNN（Region-Convolutional Neural Network）[9]、SPP-Net（Spatial Pyramid Pooling Network）[10]和 Faster-RCNN[8]等；基于单阶段的目标检测算法的大部分工作都是由神经网络来完成的，其直接使用 CNN 来卷积特征以回归物体的类别概率和位置坐标值，因此该算法的准确度低，但速度较快，主要模型有 YOLO（You Only Look Once）[11]和 SSD[5]等，这些模型在 PASCAL VOC 数据集[12]上均有着很好的检测效果。实验采用的 SSD 模型是一种基于单阶段的方法，其改进的模型有 FSSD（Feature Fusion SSD）[13]、CSSD（Context-aware SSD）[14]、DSSD（Deconvolutional SSD）[15]和 RSSD（Rainbow SSD）[16]等。SSD 模型结合了 YOLO 和 Faster R-CNN 的优点，可以在不同尺度的特征图上生成不同长宽比的锚框并将其作为先验框。YOLO 是在全连接层之后对目标进行检测，而 SSD 直接采用卷积对目标进行检测。使用的是全卷积结构相比于使用全连接层的 Faster R-CNN 模型和 YOLO 模型，参数大大减少，运算速度加快，并且可以输出任意尺度的图片。在视频目标的检测过程中，行人检测技术较为成熟[17]，并且也被应用到多个领域[18-19]，但却缺乏针对行人鞋子图像的检测算法，为此提出一种基于 SSD300-V 的单类别检测算法。

3 面向行人鞋子检测的 SSD 模型

3.1 SSD 模型

SSD 是由 Liu 等[4]提出的，在检测的速度和精度上都有不错的效果，是目前目标检测较为主流的方法之一。SSD 使用 VGG-16（Visual Geometry Group-16）[20]网络作为基础网络，并在此基础上增加了额外的卷积层，最终形成用于检测 6 个不同尺

度的特征层，分别为 Conv4_3、Conv7、Conv8_2、Conv9_2、Conv10_2 和 Conv11_2，之后对每个特征层使用两组卷积核进行卷积，分别用于定位和回归检测。SSD 经过 CNN 训练后可以得到一定数量的默认框并对其进行评分，采用非极大抑制（NMS）算法产生最后的预测结果，这些默认框带有偏移量和目标类别置信度的信息。SSD 的网络结构模型如图 1 所示，其中 Fc 为全连接层。

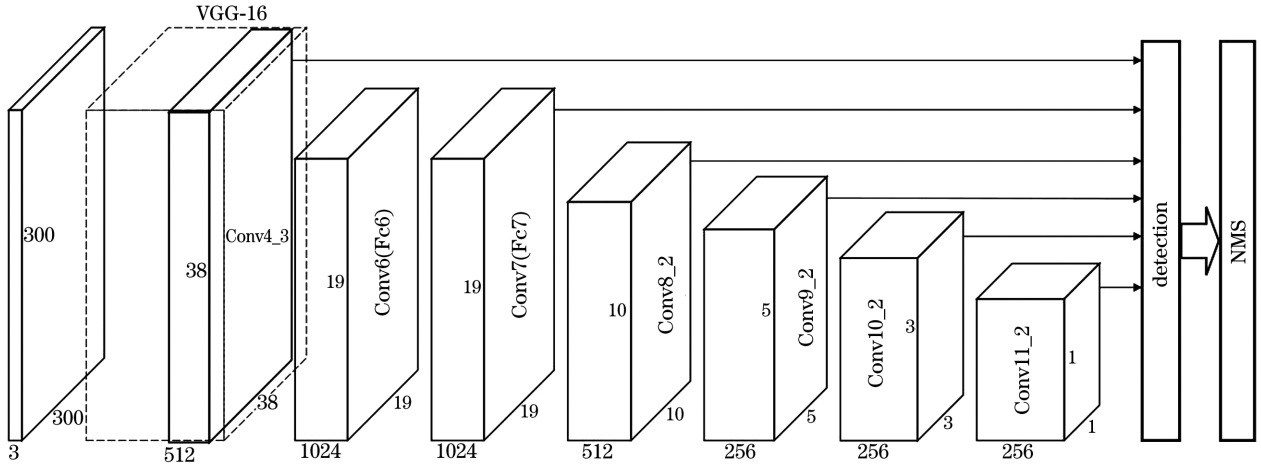

图 1　SSD 网络结构模型

Fig. 1　SSD network structure model

SSD 借鉴了 Faster R-CNN 中锚框的理念生成先验框，采用多尺度方法设置了 6 个特征响应图，并在特征图上设置长宽比不同的先验框，SSD 模型先验框的设置参数主要有尺寸和长宽比两个。先验框的尺寸是随机设定的，假设有 m 个特征层用来预测，那么其尺寸可表示为

$$S_k = S_{\min} + \frac{S_{\max} - S_{\min}}{m-1}(k-1), \tag{1}$$

式中：$k=1,2,3,4,5,6$；S_k 表示其他层默认框的尺寸；$S_{\min}$ 表示第一层默认框的尺寸，值为 0.2；$S_{\max}$ 表示第 6 层默认框的尺寸，值为 0.9。SSD 默认框的长宽比 $a_r=(1,2,3,1/2,1/3)$，其中 $r=0,1,2,3,4$。默认框的宽 $w_k^a = S_k\sqrt{a_r}$ 和高 $h_k^a = S_k/\sqrt{a_r}$，此外纵横比为 1 的 $S_k = \sqrt{S_k S_{k+1}}$。每个特征图有尺寸和种类不同的默认框，所以基本上可以覆盖到图像中各种形状和尺寸的目标。

SSD 的网络训练函数与 Faster R-CNN 类似，是一个多任务的损失函数，其包括置信度损失函数（L_{conf}）和位置损失函数（L_{loc}）两部分，表达式为

$$L(x,c,l,g) = \frac{1}{N}\left[L_{\text{conf}}(x,c) + \alpha L_{\text{loc}}(x,l,g)\right], \tag{2}$$

式中：N 表示先验框中正样本的个数，当 $N=0$ 时，则认为损失函数值为 0；x 表示目标；c 表示置信度；l 表示预测框；g 表示真实框；α 为调整 L_{loc} 的权重比例系数，默认值为 1。对于位置损失函数，使用位置回归函数 Smooth L1 损失函数来计算[10]。

对于置信度损失函数，使用 Softmax 损失函数来计算，表达式为

$$L_{\text{conf}}(x,c) = -\sum_{i\in P_{\text{positive}}}^{N} x_{ij}^{p}\ln(\hat{c}_i^{p}) - \sum_{i\in N_{\text{negative}}}^{N}\ln(\hat{c}_i^{0}), \tag{3}$$

其中

$$\hat{c}_i^{p} = \frac{\exp(\hat{c}_i^{p})}{\sum_{p}\exp(\hat{c}_i^{p})}, \tag{4}$$

式中：P_{positive} 表示正样本；N_{negative} 表示负样本；$\hat{c}_i^{p}$ 表示第 i 个预测边界框对于类别 p 的预测概率；$\hat{c}_i^{0}$ 表示负样本的损失，即类别为背景的损失。

3.2　面向行人鞋子检测的 SSD 模型设计

SSD 使用的特征层和先验框的设置都是基于 PASCAL VOC[12] 和 COCO[21] 等常见的公开数据集，所以并不适用于行人鞋子的检测。为了适应小目标的检测，需要对网络结构进行调整，同时根据鞋子检测的特点，对先验框的设置进行改进，构建面向行人鞋子检测的 SSD300-V 模型，使其可以在较少的训练资源下，加快训练收敛速度，从而提高检测精度。

3.2.1　网络结构设计

原始 SSD 模型是对不同尺寸的物体进行检测。由文献[8,22]可知，特征层次越低，保留的图像细节越多，越容易检测小物体，特征层次越高，图像的语

义信息越丰富，对大物体更敏感。行人鞋子的检测属于小目标检测，所以根据实验结果对网络模型结构进行优化，去掉后 4 层特征层，只保留 Conv4_3 和 Conv7 这两个效果最佳的特征层并进行检测。

针对原始 SSD 模型对于小目标检测效果不佳的问题，DSSD[15]将特征提取网络替换成具有更强特征提取能力的 ResNet[23]，并使用反卷积来提升分辨率，RSSD[16]通过融合不同特征来提高特征提取的能力，但是这些模型的参数量大，不适合视频中鞋子的实时检测。为了进一步提高检测效果，借鉴了 FSSD[13]和 FFSSD(Feature-Fused SSD)[24]的思想，在 SSD 的基础上将深层特征融合到浅层特征中，从而扩大浅层特征的感受野。将 Conv5_3 的特征融合到 Conv4_3 中的过程：首先将原始 SSD 模型中的 Conv5_3 标准化，接着对其进行双线性插值上采样处理后与 Conv4_3 层进行特征融合，最后使用卷积处理得到与 Conv4_3 维度一致的特征图 P 1，融合过程如图 2 所示。最终选取 P 1 层和原 Conv7(Fc7)层作为特征提取层，网络结构模型如图 3 所示，其中⊕为特征融合操作，l 为第 l 个小卷积层。

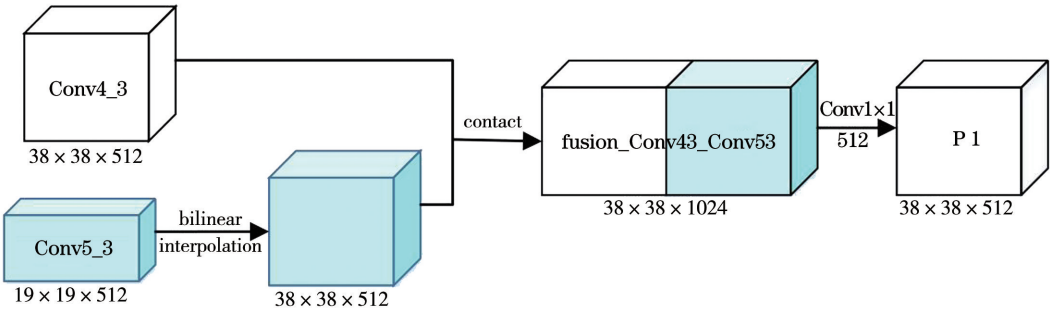

图 2　目标特征层的融合过程

Fig. 2　Fusion process of target feature layer

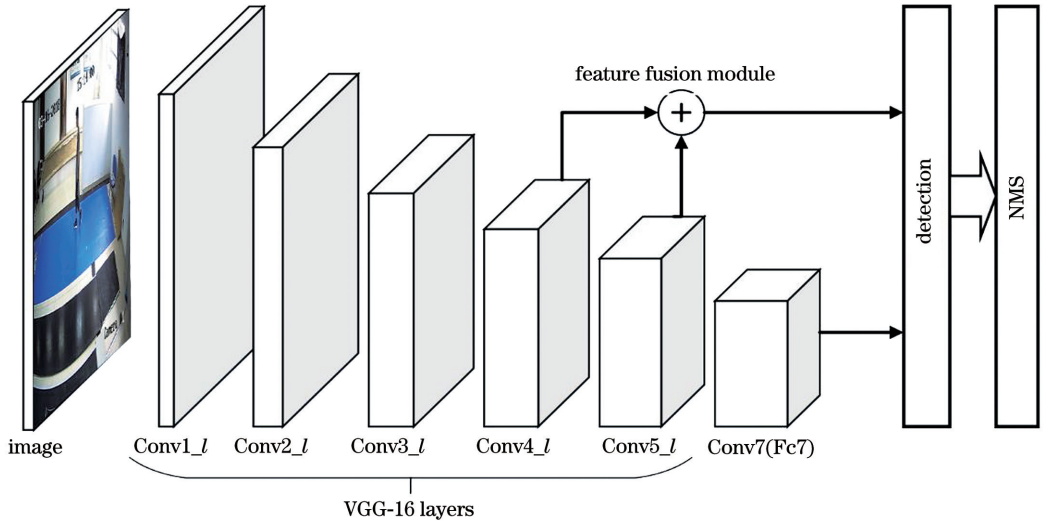

图 3　所提的网络结构

Fig. 3　Proposed network structure

3.2.2　先验框设计

为了准确检测不同尺寸的物体，原始 SSD 模型的先验框长宽比一般取(1，2，3，1/2，1/3)。但针对行人鞋子这一具体目标，原始 SSD 模型的部分长宽比不利于此任务的检测。行人在行走的过程中，鞋子会随着腿的运动以脚踝为支点进行一定角度的摆动。当鞋的方向平行于地面时，综合鞋跟及鞋脖子的因素，鞋的长宽比一般约为 4/3；当鞋的方向与地面之间的夹角为 45°时，综合鞋跟及鞋脖子的因素，鞋的长宽比一般约为 1；当鞋的方向垂直于地面时，综合鞋跟及鞋脖子的因素，鞋的长宽比一般约为 3/4。实验采用 K 均值(K-means)聚类方法对目标统计框的数据进行统计分析，发现目标框的长宽比集中在 [3/4，4/3] 区域，目标框的尺寸集中在 [15，40] 区域。因此，将 SSD 模型中先验框的长宽比设为(1，3/4，4/3)，P 1 和 Conv7(Fc7)层使用固

定的先验框尺寸范围为[15,40]。

4　实验及结果分析

4.1　实验数据

实验使用杨孟京等[6]采集的视频数据。足迹实验中,采用视频监控记录的方法对 8 名志愿者(2 名女性,6 名男性)所穿的 50 双鞋子进行视频采集,采集过程中控制角度、光照和志愿者所行走的路线等变量,以期尽可能模拟实际的情况。采集的视频尺寸为 1920 pixel×1080 pixel,每个视频时长约为 80 s。选用课题组前期构建的部分视频数据进行分帧处理,可以得到 1261 张图片,共制作标签数据 2470 个,训练集与测试集中的目标和图像数量如表 1 所示。

表 1　不同数据集中的目标和图像数量

Table 1　Number of objects and images in different datasets

Name	Training dataset		Test dataset	
	Object	Image	Object	Image
Shoe	2223	1134	247	127

制作标签数据的具体过程:对采集的视频进行分帧处理,调整其尺寸为 300 pixel×300 pixel 的统一格式,接着使用 LabelImg 软件对视频中的鞋子图像进行手工标注,最后制成 PASCAL VOC 数据集的格式并保存到指定的文件夹中,标签命名为“shoe”。在标注的过程中应当注意尽量保证每次选取的框可以紧密地覆盖住鞋子,但不要框选到图片的边缘位置,标注过程如图 4 所示。

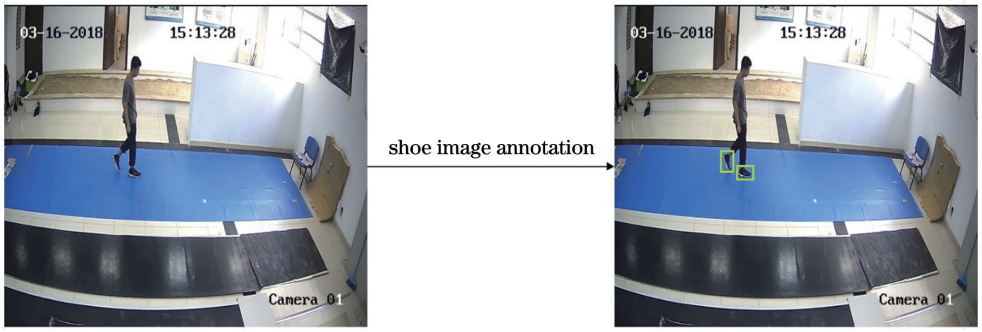

图 4　图像标注过程

Fig. 4　Image annotation process

4.2　实验环境的配置

实验的硬件配置为 Intel(R) Xeon(R) CPU E5-2650 @ 2.00 GHz,内存为 256 GB,GPU 为 Tesla K40c。软件配置为 Linux、CUDA9.0.176、CUDNN7.5.1 和 OpenCV3.4.1。深度学习算法框架为 PyTorch。

基于 PyTorch 搭建实验环境,为了加快模型的训练速度,在 ImageNet 上预训练模型。模型在训练过程中,前 65000 次迭代不使用学习率调整策略,默认的学习率为 0.001,之后的迭代过程使用 SGD (Stochastic Gradient Descent)学习率调整策略。超参数的设置:梯度下降速率为 0.001,权重衰减为 0.0005,冲量为 0.9,共训练 2×10^4 次。

4.3　评价指标

使用 AP(Average Precision)和画面每秒传输帧数(FPS)两个评价指标对图像进行评价。目标检测过程中得到的检测结果是一个带有标签的框,检测正确性的度量标准为交并比(IOU),其是模型所预测的检测框和真实检测框的交集和并集之间的比例。设定 IOU 的阈值,大于这一阈值的则被认为是正确检测的样本,小于阈值的则被认为是负样本。实验中判别模型的标准主要由准确率(P)、召回率(R)和 AP 来构成,表达式为

$$P=\frac{x_{\mathrm{TP}}}{x_{\mathrm{TP}}+x_{\mathrm{FP}}},\tag{5}$$

$$R=\frac{x_{\mathrm{TP}}}{x_{\mathrm{TP}}+x_{\mathrm{FN}}},\tag{6}$$

$$x_{\mathrm{AP}}=\frac{\sum P}{N_{\mathrm{Total}}},\tag{7}$$

式中:x_{TP} 表示假阳性样本;x_{FP} 表示真阳性样本;x_{FN} 表示假阴性样本;N_{Total} 表示总样本数。FPS 简单来说指的是在 1 s 的时间内识别的图像数(帧数)。

4.4　实验结果分析

4.4.1　不同特征图对于模型检测效果的影响

由于不同的特征层感受野不同,所以对于大、小目标的感受能力不同,为此在原始 SSD 模型上使用不同的特征层进行对比实验,用来探究不同特征层对鞋子检测的效果,结果如表 2 所示。

表 2 不同特征层的检测效果

Table 2 Detection effect of different feature layers

Conv4_3	Conv7	Conv8_2	Conv9_2	Conv10_2	Conv11_2	AP	FPS
√	√	√	√	√	√	0.831	31
√	√	√	√	√		0.834	31
√	√	√	√			0.829	32
√	√	√				**0.837**	33
√	√					0.830	35
√						0.811	**38**

从表 2 可以看到，当特征层组合包含 Conv4_3 和 Conv7 时，精度相差较小，当只有 Conv4_3 层时，精度为 0.811，精度较低，原因在于数据集中的标注框尺寸大于该特征层先验框的尺寸，从而导致精度下降；Conv8_2、Conv9_2、Conv10_2 和 Conv11_2 层对鞋子检测的作用较小，所以在设计模型时将其删去。

4.4.2 不同检测网络对于检测效果的影响

RFB(Receptive Field Block)[25]和 FSSD[13]都是 SSD 模型的改进模型，改进后的模型在小目标检测中具有较好的效果。FSSD 借鉴了 FPN 的思想对 SSD 模型进行改进，其增加了一个特征融合模块，就是将不同层、不同尺度的特征层进行融合后再进行检测。RFB 是在 SSD 模型的基础上进行改进，整体与 SSD 结构相差较小，主要是在特征提取网络上增加了一个 RFB 模块。上述两个模型对检测行人鞋子等小目标问题上有着很好的效果。实验在 SSD512-VGG、RFB-VGG、FSSD-VGG 和 SSD300-VGG 模型上使用原始 SSD 模型的先验框设置，在保证其他条件相同的情况下，将实验的图像尺寸改为 512 pixel×512 pixel，并在 SSD512-VGG 模型上进行对比实验，之后选用尺寸为 300 pixel×300 pixel 的图像在 RFB-VGG、FSSD-VGG 和 SSD300-VGG 模型上进行对比实验，结果如表 3 所示。

表 3 不同检测网络的检测效果

Table 3 Detection effect of different detection networks

Model	AP	FPS
SSD300-VGG	0.831	25
SSD512-VGG	0.863	11
FSSD-VGG	0.866	20
RFB-VGG	0.862	11
SSD300-V	**0.891**	**33**

从表 3 可以看到，SSD512-VGG 模型的检测效果高于 SSD300-VGG 模型，AP 值提升 0.032，说明对于鞋子这种小目标，高层特征图未包含足够的信息，提高图像的分辨率可以提升网络的检测精度，原因在于分辨率的增加可以使图像中鞋子的尺寸也相应增大，进而提高检测精度，这与文献[26]的结果一致，但 SSD512-VGG 的检测速度低于 SSD300-VGG，原因在于输入图像的数据量增大，使得网络处理速度变慢；FSSD-VGG 和 RFB-VGG 的 AP 值分别为 0.866 和 0.862，比 SSD300-VGG 模型提高约为 0.030；设计的 SSD300-V 模型的检测精度比 FSSD-VGG 和 RFB-VGG 模型分别提高 0.025 和 0.029，原因在于该模型采用特征融合方法并对先验框进行设置，改善用于鞋子检测的特征图质量，同时在检测速度方面，该模型有一定的优越性。

4.4.3 不同特征提取网络对于检测效果的影响

EfficientNet_b3[27]和 MobileNet-V2[28]是应用比较广泛的特征提取网络，两者都是轻量级网络，参数量少，适合部署等工程应用，而 VGG 网络的参数量较大，实时性差，在保证其他条件不变的情况下，更换了特征提取网络并与所提模型进行对比，结果如表 4 所示。

表 4 不同提取网络的检测效果

Table 4 Detection effect of different extraction networks

Model	AP	FPS
SSD-VGG	0.831	25
SSD-EfficientNet_b3	0.862	51
SSD-MobileNet_V2	0.774	**91**
SSD300-V	**0.891**	33

从表 4 可以看到，MobileNet-V2 是轻量级网络，模型参数量小，所以检测精度较低，只有 0.774，

低于 SSD-VGG 模型，但在检测速度方面是 SSD-VGG 模型的 3 倍，这有利于在实际检测环境中的部署；EfficientNet_b3 作为深度学习网络，其在测试精度和速度方面都有较好的效果，检测速度是 SSD-VGG 的 2 倍，但检测精度仅为 0.862，比 SSD-VGG 模型高 0.031，不如所提模型，说明所提模型在检测精度方面有一定优势。

综上所述，设计的 SSD 模型在测试集上最终的 AP 值达 0.891，检测精度优于其他模型，图 5 为实验检测结果。

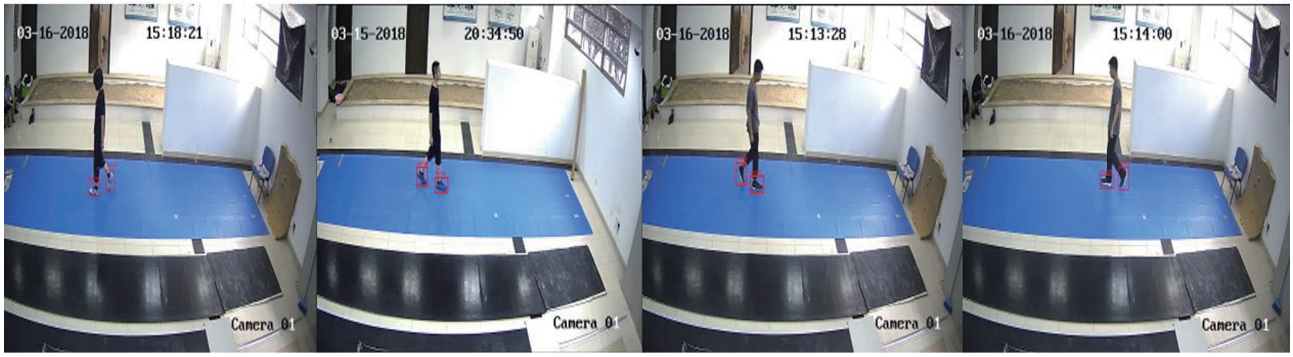

图 5 实验检测结果

Fig. 5 Experimental test results

5 结 论

提出一种基于 SSD300-V 模型的鞋子图像检测方法，该方法可以对监控视频中的鞋子图像进行检测。采用特征融合的方法对原始 SSD-VGG 网络模型进行设计，进而提高检测精度。采用 K-means 方法对鞋子目标框进行聚类，进而设置先验框的长宽比和尺寸等参数，进而提高检测精度。同时又与其他改进的 SSD 模型和不同的特征提取网络进行对比实验，突出所提模型的优越性。

所提模型虽然在该数据集上得到很好的精度，但是所使用的数据存在背景单一和实验人数不足等问题，实际案件中会存在人数多、背景复杂和光照强度变化等情况，这些实际因素会导致模型的误检率增加，精度降低，所以该模型仅适合本文的数据集，使得模型的泛化能力较低，但仍具有较大的改进空间。下一步将扩充实验的数据集，增加数据集的多样性，使其更贴近实际案件。同时使用的原始 SSD 目标检测模型也存在一定的局限性，所以下一步工作将尝试特征融合等技术以便设计出更符合视频中嫌疑人鞋子的检测模型。

参 考 文 献

[1] Sun Y H. Insight on comprehensive application of various detection means into video tracking [J]. Forensic Science and Technology, 2019(3): 257-260.
孙熠赫. 论视频追踪中多种侦查手段的综合运用[J]. 刑事技术, 2019(3): 257-260.

[2] Nong D S. Based on the footprint of the scene, expand video investigation [J]. Legal System and Society, 2015(22): 255-256.
农达史. 立足现场足迹、拓展视频侦查[J]. 法制与社会, 2015(22): 255-256.

[3] Yuan C P, Yu S W. A preliminary study on the application of footprint analysis in video investigation work [J]. Guangdong Public Security Science and Technology, 2017, 25(2): 61-63, 74.
袁楚平, 余尚伟. 足迹分析在视频侦查工作中的运用初探[J]. 广东公安科技, 2017, 25(2): 61-63, 74.

[4] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector [EB/OL]. (2016-12-29) [2020-06-07]. https://arxiv.org/abs/1512.02325.

[5] Xu L, Li Z H, Li Z G, et al. A murder case investigated and solved by applying the simulation experiment into the collected video [J]. Forensic Science and Technology, 2018(4): 330-333.
许磊, 黎智辉, 李志刚, 等. 视频侦查模拟实验在案件侦破中的应用[J]. 刑事技术, 2018(4): 330-333.

[6] Yang M J, Tang Y Q, Jiang X J. Novel shoe type recognition method based on convolutional neural network [J]. Laser & Optoelectronics Progress, 2019, 56(19): 191505.
杨孟京, 唐云祁, 姜晓佳. 基于卷积神经网络的鞋型识别方法[J]. 激光与光电子学进展, 2019, 56(19): 191505.

[7] 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), June 20-25, 2005, San Diego, CA, USA. New York: IEEE Press, 2005: 886-893.

[8] Ren S Q, He K M, 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, 2017, 39(6): 1137-1149.

[9] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[10] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[12] Everingham M, Gool L, Williams C K I, et al. The pascal visual object classes (VOC) challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[13] Li Z X, Zhou F Q. FSSD: feature fusion single shot multibox detector[EB/OL]. (2018-05-17)[2020-06-07]. https://arxiv.org/abs/1712.00960.

[14] Xiang W, Zhang D Q, Yu H, et al. Context-aware single-shot detector [C] // 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE Press, 2018: 1784-1793.

[15] Fu C Y, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector[EB/OL]. (2017-01-23)[2020-06-07]. https://arxiv.org/abs/1701.06659.

[16] Jeong J, Park H, Kwak N. Enhancement of SSD by concatenating feature maps for object detection[C]// Procedings of the British Machine Vision Conference 2017, September 4-7, 2017, London, UK. Blue Mountains: British Machine Vision Association, 2017.

[17] Shen Y J, Hao Z H, Wang P F, et al. A novel human detection approach based on depth map via kinect[C] // 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 535-541.

[18] Zhang H Y, Wang S N, Hu W B. Improved method for estimating number of people based on convolution neural network [J]. Laser & Optoelectronics Progress, 2018, 55(12): 121503.
张红颖, 王赛男, 胡文博. 改进的基于卷积神经网络的人数估计方法[J]. 激光与光电子学进展, 2018, 55(12): 121503.

[19] Ma Y J, Li X Y, Song X F. Traffic sign recognition based on improved deep convolution neural network [J]. Laser & Optoelectronics Progress, 2018, 55(12): 121009.
马永杰, 李雪燕, 宋晓凤. 基于改进深度卷积神经网络的交通标志识别[J]. 激光与光电子学进展, 2018, 55(12): 121009.

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2020-06-07]. https://arxiv.org/abs/1409.1556.

[21] Caesar H, Uijlings J, Ferrari V. COCO-stuff: thing and stuff classes in context[C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1209-1218.

[22] Liu W, Rabinovich A, Berg A C. Parsenet: looking wider to see better[EB/OL]. (2015-11-19)[2020-06-07]. https://arxiv.org/abs/1506.04579.

[23] 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 (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[24] Cao G M, Xie X M, Yang W Z, et al. Feature-fused SSD: fast detection for small objects[J]. Proceedings of SPIE, 2018, 1061: 106151E.

[25] Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11215: 404-419.

[26] Huang J, Rathod V, Sun C, et al. Speed/accuracy trade-offs for modern convolutional object detectors [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3296-3297.

[27] Tan M X, Le Q V. EfficientNet: rethinking model scaling for convolutional neural networks[EB/OL]. (2019-05-28)[2020-06-07]. https://arxiv.org/abs/1905.11946.

[28] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2019-04-17)[2020-06-07]. https://arxiv.org/abs/1704.04861.