基于改进 SSD 算法的行人检测方法

董永昌[1]，单玉刚[2+]，袁　杰[1]

（1. 新疆大学 电气工程学院，新疆 乌鲁木齐 830047；2. 湖北文理学院 教育学院，湖北 襄阳 441053）

摘　要：针对行人检测中利用 SSD 算法不易训练、检测精度低等问题提出一种改进算法。以 DenseNet 作为 SSD 的基础网络，在其后添加四层卷积层构建新的网络；为充分利用不同深度卷积层的信息，取新建网络的后四层和 DenseNet 中最后两个 Dense_Block 来提取目标框。实验结果表明，与其它算法相比，该方法对于不同场景下行人目标检测具有更强的鲁棒性，对行人的检测率超过 92%，相比改进前的算法提高 10%以上。

关键词：单次多重检测器；密集连接卷积网络；行人检测；深度学习；鲁棒性

中图法分类号：TP391.41　**文献标识号**：A　**文章编号**：1000-7024（2020）10-2921-06

doi：10.16208/j.issn1000-7024.2020.10.037

Pedestrian detection based on improved SSD

DONG Yong-chang[1], SHAN Yu-gang[2+], YUAN Jie[1]

(1. College of Electrical Engineering, Xinjiang University, Urumchi 830047, China;
2. Institute of Education, Hubei University of Arts and Science, Xiangyang 441053, China)

Abstract: To solve the problems of hard training and low detection accuracy in pedestrian detection using SSD algorithm, an improved algorithm was proposed. DenseNet was used as the basic network of SSD, and four convolution layers were added to construct a new network. To make full use of the information of convolution layers at different depths, the last four layers of the new network and the last two Dense_Blocks in DenseNet were selected to extract the target box. Experimental results show that the proposed method has better robustness and lower false detection and miss detection rate for pedestrian detection in different environments. The detection rate for pedestrians exceeds 92%, which is 10% higher than that of the traditional algorithm.

Key words: SSD; DenseNet; pedestrian detection; deep learning; robustness

0　引　言

随着深度学习在目标检测领域的表现日益突出[1]。先后出现了 R-CNN 系列[2-4]、YOLO 系列[5-7]、SSD[8]等目标检测方法。相对来说，R-CNN 方法在检测精度上更好，YOLO 方法在检测速度上表现更优。在行人检测方面：文献［9］改进 Faster R-CNN 中候选框的选取方法，使得行人检测的精度有了一定的提高。文献［10］根据行人在图像中的几何特点，对 YOLO 网络结构进行优化，取得了不错的检测效果。

SSD（single shot MultiBox detector）算法是一种以 VGG[11]为前置网络的目标检测算法。它会均匀地在图片上产生不同大小和长宽比的候选框，然后利用卷积层提取图像特征，最后是回归和分类。文献［12］对 SSD 的基础网络部分进行改进并在检测阶段对不同特征层添加缩放因子，提高了行人检测的准确性。文献［13］用 MobileNetV2 作为 SSD 模型的基础网络，有效减少了图像特征提取过程中花费的时间及运算量。DenseNet[14]是一种密集连接的卷积神经网络，比 VGG 其更易于训练，且精度更高。本文提出一种以 DenseNet 模型为前置网络的 SSD 算法。与 SSD、YOLO、Faster R-CNN 等模型进行对比，实验结果表明，改进的 SSD 模型具有更高的准确率。

1　相关模型

1.1　SSD 检测原理

原始的 SSD 以 VGG16 或 VGG19 作为其前置网络，本

收稿日期：2019-07-11；**修订日期**：2019-10-21

基金项目：国家自然科学基金项目（61863033）；湖北省教育厅科学技术研究基金项目（B2016175）；湖北文理学院博士基金项目（2015B002）

作者简介：董永昌（1992-），男，河南洛阳人，硕士研究生，研究方向为目标检测、人工智能；+通讯作者：单玉刚（1971-），男，辽宁沈阳人，博士，讲师，研究方向为目标跟踪、模式识别；袁杰（1975-），男，新疆乌鲁木齐人，博士，副教授，研究方向为机器人智能控制算法。E-mail：32748873@qq.com

文以前置网络为 VGG16 的 SSD300 模型为网络架构，其网络结构如图 1 所示。

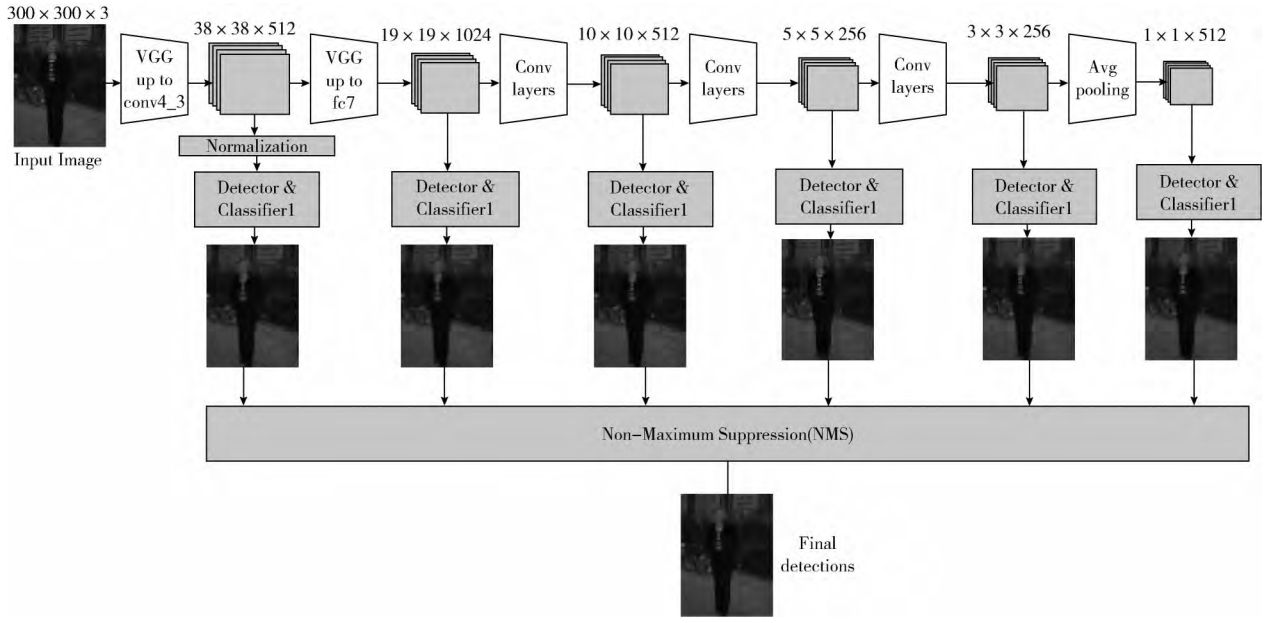

图 1 SSD 网络结构

该模型要求输入图片大小是 300 × 300，然后根据 VGG16 网络结构，将其前 5 层做为基础，将两个全连接层转换为两个卷积层，并在其后添加 3 个卷积层和一个平均池化层所组成。网络在 Conv4_3、Conv7、Conv8_2、Conv9_2、Conv10_2、Pool11 层上输出一组不同宽高比和不同大小的默认框。

默认框的生成是以特征层上每个点为中心生成一系列同心框，每个特征层上默认框的尺寸由式（1）确定

$$S_k = S_{min} + \frac{S_{max} - S_{min}}{m-1}(k-1), k \in [1, m] \quad (1)$$

式中：m 指不同预测特征层尺寸的数量，SSD300 中预测特征层尺寸的数量为 6，同时 S_{min} 为 0.2，S_{max} 为 0.9。

而更小的特征层意味着更大的感受野，更大的特征层意味着更小的感受野。因此不同特征层上相同尺寸的默认框在输入图像上的映射范围是不同的，在其默认框生成各自的置信度和位置信息之后，通过 reshape 来调整特征层为相同尺寸，之后用非极大值抑制和 Softmax 得检测结果。

1.2 DenseNet 模型

DenseNet 模型是一种用于图片分类的模型，与 VGG、ResNet[15] 和 Inception[16,17] 不同，它是一种前馈方式的密集型连接（dense connection）的网络，即

$$X_l = H_l([X_0, X_1, \cdots, X_{l-1}]) \quad (2)$$

其中，X_l 表示第 l 层的输入，[] 表示将 $X_0, X_1, \cdots, X_{l-1}$ 所有输出进行组合拼接，H 表示批量正则化、ReLU 激活函数以及 Conv(3×3) 卷积层的组合。这种连接方式加强了网络中各层之间的信息交流，有效提高了网络的训练效率，以 DenseNet-121 为例，其网络参数见表 1。

表 1 DenseNet-121 网络结构

Layers	Output size	DenseNet-121
Convolution	112×112	7×7 Conv
Pooling	56×56	3×3 Max Pool
Dense_Block0	56×56	$\begin{pmatrix}1\times1\text{ Conv}\\3\times3\text{ Conv}\end{pmatrix}\times6$
Transition Layer0	56×56 28×28	1×1 Conv 2×2 Average Pool
Dense_Block1	28×28	$\begin{pmatrix}1\times1\text{ Conv}\\3\times3\text{ Conv}\end{pmatrix}\times12$
Transition Layer1	28×28 14×14	1×1 Conv 2×2 Average Pool
Dense_Block2	14×14	$\begin{pmatrix}1\times1\text{ Conv}\\3\times3\text{ Conv}\end{pmatrix}\times24$
Transition Layer2	14×14 7×7	1×1 Conv 2×2 Average Pool
Dense_Block3	7×7	$\begin{pmatrix}1\times1\text{ Conv}\\3\times3\text{ Conv}\end{pmatrix}\times16$
Classification Layer	1×1	7×7 Global Average Pool 1000D Fully-Connected Softmax

其中，Stride 默认为 2，用 Dropout 来随机丢弃一部分特征层来减少参数。由于在 DenseNet 中需要对不同特征层进行拼接，所以要求 Dense_Block 中的特征图尺寸保持相同的尺寸，在相邻两个 Dense_Block 之间设置过渡层（transition layers）进行下采样，过渡层由批量正则化、（1

×1）的卷积层和（2×2）的平均池化层组成。在 Dense_Block 中，输入先经过 1×1 卷积核来减少特征图数量，然后作为 3×3 卷积核的输入进行运算，大大减少了计算量，过渡层通过参数缩减将输入到该层的特征图数量减小到原来的一半。

2　行人检测算法

2.1　改进 SSD 网络结构

本文用 DenseNet-121 作为 SSD 的前置网络。首层是 3×3 的卷积层，然后经过 4 个分别含有 6 层、12 层、24 层、16 层的 Dense_Block，每两个 Dense_Block 之间是非线性变换 BN+ReLU+Conv(3×3) 的组合，DenseNet 之后是 4 层包含 1×1 和 3×3 卷积层。最后，选取后两层 Dense_Block 和 4 个卷积层进行目标框的选取。新构建的网络在选取目标框的时候能够更好利用不同特征层的信息。其网络结构模型如图 2 所示。

网络中 Growth_k 为 24，Dropout_rate 为 0.5。优化函数使用 Adam（adaptive moment estimation），它综合了一阶矩估计和二阶矩估计来计算梯度下降的更新步长。相对于其它优化函数其具有效率高，实现方便，内存使用率低等优点。

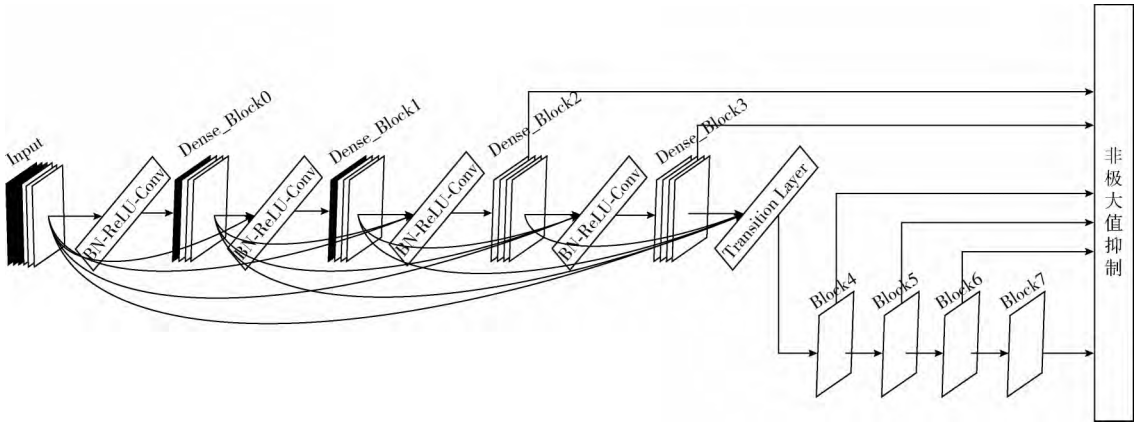

图 2　本文网络结构模型

SSD 网络结构在一定程度上利用了浅层网络的目标特征，降低了因网络太深而造成的模型参数数量暴增和梯度消失的风险。但是，更浅的特征层意味着更弱的表征能力，所以可以通过提高浅层网络的表征能力来提高网络的检测性能。本文以 DenseNet 作为 SSD 的基础网络，与传统的 SSD 模型相比，该模型不仅有效缓解了梯度消失问题，使得网络更加容易训练，又大幅度减少了网络的参数量，而且加强了特征向更深层次传播。

2.2　损失函数选择

损失函数由两部分组成：置信度损失和位置损失，如式（3）～式（5）所示

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)) \tag{3}$$

$$L_{conf}(x,c)=-\sum_{i\in Pos}^{N}x_{ij}^{p}\log(\hat{c}_i^p)-\sum_{i\in Neg}\log(\hat{c}_i^0),$$

其中

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

$$L_{loc}(x,l,g)=\sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}smooth_{L1}(l_i^m-\hat{g}_j^m) \tag{5}$$

其中，N 为匹配成功的默认边框数目，$L_{conf}(x,c)$ 代表置信度损失，$L_{loc}(x,l,g)$ 代表位置损失。用 Smooth L1 函数的好处是当预测框和真实框差距过大时避免梯度爆炸，当预测框和真实框差距过小时梯度不至于消失。

2.3　行人检测算法流程

整个算法流程分为训练模块和测试模块。在训练模块，为了使训练模型具有更强的泛化能力，防止训练模型过拟合，将原始数据集进行扩充增强。之后将数据标注成 VOC2007 数据格式用于训练、测试和验证，最后将批量输入图片送入设计网络进行训练并得到最终训练模型。在测试模块，网络会首先在输入图片中生成默认框，并判断默认框的置信度，然后根据训练模型对目标框进行位置和置信度的调整，最后在产生的 N 个检测框中利用非极大值抑制原理产生最终的检测结果。其算法流程如图 3 所示。

3　实验与分析

3.1　实验平台

仿真的实验环境为：Win10 的操作系统，英特尔 i5-9400F 处理器和 NVIDIA GeForce GTX 1070 显卡。实验采用 Python 编程语言在 Tensorflow 深度学习框架上进行模型的训练和测试。

3.2　数据集构建和模型训练

本文使用 INRIA 数据集来进行模型的训练，它包括训练样本和测试样本共 902 张（包含 3542 个行人）。图片中

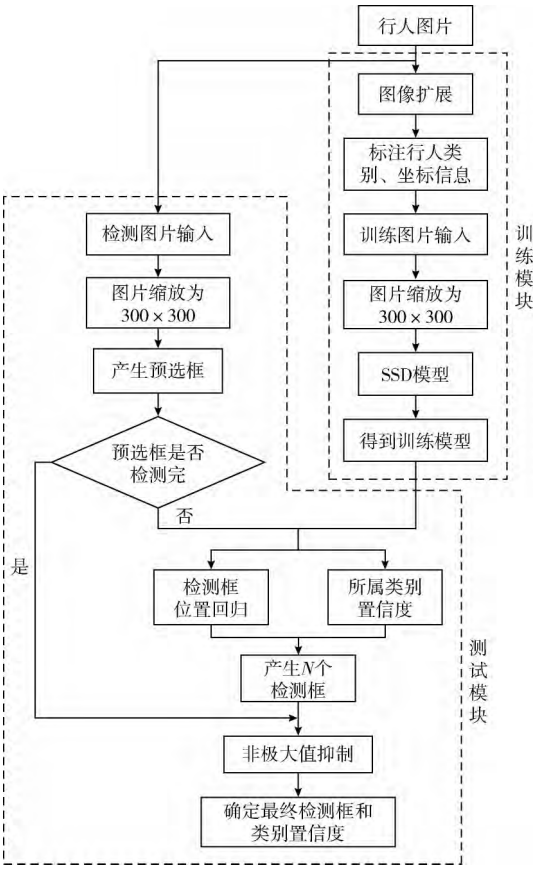

图 3 算法流程

人体有站立、坐、蹲、跳起等姿势，目标人物包含两个性别各年龄阶段，且目标相对于整幅图片尺寸大小不定。背景包括室内、野外、运动场、街道、海边等不同地方，图片高度大于 100 个像素。为了增加样本多样性，提高样本质量，有效缓解训练过程中的过拟合现象，本文在此基础上通过水平翻转、裁剪等样本增强方法将数据集扩展为 3000 张，其中 2400 张为训练样本，600 张为测试样本。每个每批次处理 16 张图片，迭代 60 000 次。训练过程中设置学习率从 0.01 以指数衰减形式进行衰减，衰减因子为 0.94，每 12 个 Epoch 衰减一次。学习率变化曲线如图 4 所示。

模型改进前后训练过程的损失值和准确度曲线如图 5 所示。

由图 5 可见，本文所提 SSD 模型比传统 SSD 模型的收敛速度更快，且改进的 SSD 模型比传统的 SSD 模型更低的损失和更高的准确率。迭代 60 000 此后，原模型的损失和准确率分别为 5.6 和 0.86；改进后模型的损失和准确率分别为 3.7 和 0.94。

3.3 评价指标

选取准确率（Precision，P）和召回率（Recall，R）作为检验网络优越性的指标，其定义为

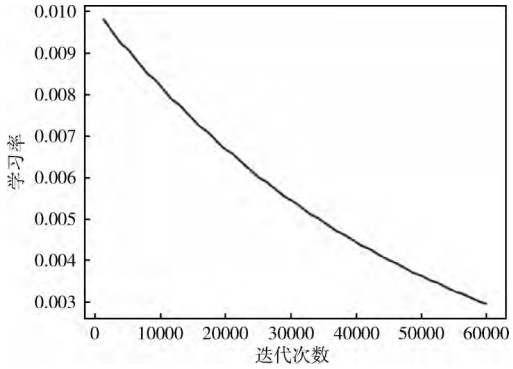

图 4 学习率变化曲线

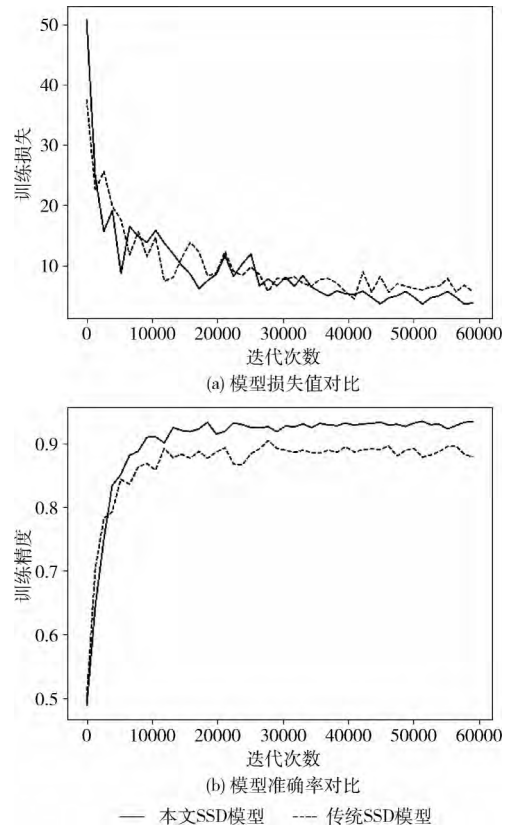

图 5 模型结果对比

$$P=\frac{T_P}{T_P+F_P}\times 100\% \tag{6}$$

$$R=\frac{T_P}{T_P+F_N}\times 100\% \tag{7}$$

式中：T_P表示正确检测到行人的数量，F_P表示误把非行人目标检测为行人目标的数量，F_N表示误把行人检测为背景的数量。

3.4 结果分析

在 INRIA 数据库中随机选出 300 张图片，其中含有 1658 个行人的图片集进行测试。将实验结果和原始 SSD 算法、Faster-RCNN 以及 YOLO_v3 进行比较，在检测精度

和目标被遮挡方面分析改进后网络的检测效果。对比结果见表 2。

表 2　不同模型检测结果对比

模型	目标总数	P/%	R/%
Faster RCNN	1658	82.30	70.06
SSD	1658	83.05	74.07
YOLO_v3	1658	87.10	78.90
本文算法	1658	92.40	85.01

从检测结果的指标来看：与 Faster RCNN、YOLO_v3 等模型相比，改进后的模型比原始模型具有更高的检测精度和召回率，本文模型在准确率上分别提高 10.1% 和 5.3%，在召回率上分别提高 14.95% 和 6.11%。故而相对于其它 3 种模型，本文模型具有更高的检测精度，具有良好的应用前景。具体检测结果如图 6 所示。

(a) Faster RCNN检测结果

(b) SSD检测结果

(c) YOLO_v3检测结果

(d) 本文算法检测结果

图 6　几种模型检测结果对比

图 6 是改进后的 SSD 网络与原始网络、Faster RCNN、YOLO_v3 的检测效果对比，从检测结果来看，Faster RCNN 检测精度较高，但是对于图 6（a）第 3 幅图中的目标人物遮挡显现的检测效果并不好；图 6（b）中原始 SSD 在对小目标的检测率上有所提高，但是对小目标的误检率也随之上升，而且对于遮挡现象也同样没有很好的检测效果；YOLO_v3 的检测结果要明显优于前两种模型，其在目标被遮挡和检测精度上有了较大的提高，但是如图 6（c）第一幅图所示，该模型对小目标的检测率却不太好；图 6（d）中经本文改进后的 SSD 模型在相同的测试图片中，其目标人物的检测置信度都较高，且对于第 3 幅图中被遮挡的目标人物的置信度为 0.916，而且对于边缘小目标的检测率也有较大的提高。相对于其它模型，本模型在保证检测精度的基础上对被遮挡目标和边缘小目标的检测率有了较大的改善，检测结果更好，具有更强的鲁棒性。

4　结束语

本文将 DenseNet 作为 SSD 模型的前置网络，使得网络的深度得到增加，在生成目标检测框的时候，不仅利用了更深特征层信息，也使得浅层网络信息得到充分利用，其结构特点也提升了梯度的反向传播，使得网络更容易训练。对由于阴影、遮挡等造成的行人检测效果不理想等问题，本模型对其性能具有较好的提升，对不同像素、不同姿态的行人检测具有较好的鲁棒性。

参考文献：

[1] ZHANG Hui，WANG Kunfeng，WANG Feiyue. Advances and perspectives on applications of deep learning in visual object detection [J]. Acta Automatica Sinica，2017，43 (8)：1289-1305 (in Chinese). [张慧，王坤峰，王飞跃. 深度学习在目标视觉检测中的应用进展与展望 [J]. 自动化学报，2017，43 (8)：1289-1305.]

[2] 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.

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

[4] 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.

[5] Redmon J，Divvala S，Girshick R，et al. You only look once：Unified，real-time object detection [C] //Computer Vision and Pattern Recognition. IEEE，2016：779-788.

[6] Redmon J，Farhadi A. YOLO9000：Better，faster，stronger [C] //IEEE Conference on Computer Vision and Pattern Recognition，2017：6517-6525.

[7] Redmon J，Farhadi A. YOLOv3：An incremental improvement [R]. [2019-06-03]. https：//arxiv.org/abs/1804.02767.

[8] Liu W，Anguelov D，Erhan D，et al. SSD：Single shot MultiBox detector [C] //European Conference on Computer Vision. Springer，2016：21-37.

[9] CHEN Enjia，TANG Xianghong，FU Bowen. Pedestrian search method based on Faster R-CNN with the integration of pedestrian detection and re-identification [J]. Journal of Computer-Aided Design & Computer Graphics，2019，31 (2)：332-339 (in Chinese). [陈恩加，唐向宏，傅博文. Faster R-

CNN 行人检测与再识别为一体的行人检索算法［J］. 计算机辅助设计与图形学学报，2019，31（2）：332-339.］

［10］GAO Zong，LI Shaobo，CHEN Jinan，et al. Pedestrian detection method based on YOLO network［J］. Computer Engineering，2018，44（5）：215-219（in Chinese）.［高宗，李少波，陈济楠，等. 基于 YOLO 网络的行人检测方法［J］. 计算机工程，2018，44（5）：215-219.］

［11］Simonyan K，Zisserman A. Very deep convolutional networks for large-scale image recognition［J］. Computer Science，2015，56（3）：1-14.

［12］XING Haoqiang，DU Zhiqi，SU Bo. Pedestrian detection method based on modified SSD［J］. Computer Engineering，2018，44（11）：228-233（in Chinese）.［邢浩强，杜志岐，苏波. 基于改进 SSD 的行人检测方法［J］. 计算机工程，2018，44（11）：228-233.］

［13］LIU Hui，ZHANG Lishuai，SHEN Yue，et al. Real-time pedestrian detection in orchard based on improved SSD［J］. Transactions of the Chinese Society for Agricultural Machinery，2019，50（4）：29-35（in Chinese）.［刘慧，张礼帅，沈跃，等. 基于改进 SSD 的果园行人实时检测方法［J］. 农业机械学报，2019，50（4）：29-35.］

［14］Huang U，Liu Z，Van Der Maaten L，et al. Densely connected convolutional networks［C］//The IEEE Conference on Computer Vision and Pattern Recognition，2017：4700-4708.

［15］He K，Zhang X，Ren S，et al. Deep residual learning for image recognition［C］//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2016：770-778.

［16］Szegedy C，Liu W，Jia Y，et al. Going deeper with convolutions［C］//Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2014：1-9.

［17］Szegedy C，Vanhoucke V，Ioffe S，et al. Rethinking the inception architecture for computer vision［C］//Conference on Computer Vision and Pattern Recognition. IEEE Computer Society，2015：2818-2826.