第 58 卷 第 24 期/2021 年 12 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

多尺度特征融合的双通道 SSD 行人头部检测算法

周永福[1]，李文龙[1,2]，胡冉冉[2*]

[1]吉林交通职业技术学院管理工程学院，吉林 长春 130012；

[2]长春理工大学电子信息工程学院，吉林 长春 130022

摘要 针对行人头部易受光照变化和遮挡的影响而导致目标检测准确率较低的问题，提出一种基于多尺度融合的双通道 SSD(Single Shot Multibox Detector)行人头部检测算法。首先在 SSD 网络上增加一条深度通道，将带有深度信息的头部特征与 SSD 网络的特征进行融合，形成双通道 SSD 网络；然后在双通道 SSD 网络的基础上，将具有丰富语义信息的高层特征图与低层特征图进行特征融合，实现更精确的头部定位；最后重新调整 SSD 的先验框以减少 SSD 网络的计算量。实验结果表明，在光照和遮挡的情况下，相比于传统 SSD 目标检测算法，改进后算法的检测精度提高了 12.9 个百分点，其可有效解决光照变化和遮挡对行人头部检测的影响。

关键词 机器视觉；行人头部检测；SSD 网络；深度信息；多尺度特征融合

中图分类号 TP391 **文献标志码** A **doi**：10.3788/LOP202158.2415009

Two-Channel SSD Pedestrian Head Detection Algorithm Based on Multi-Scale Feature fusion

Zhou Yongfu[1], Li Wenlong[1,2], Hu Ranran[2*]

[1] *School of Management Engineering, Jilin Communications Polytechnic, Changchun, Jilin* 130012, *China*;

[2] *School of Electronic Information Engineering, Changchun University of Science and Technology, Changchun, Jilin* 130022, *China*

Abstract Aiming at the problem that pedestrian head is susceptible to illumination changes and occlusion, which leads to low target detection accuracy, a pedestrian head detection algorithm based on two-channel single shot multibox detector (SSD) with multi-scale fusion is proposed. First, a deepth channel is added to the SSD network, and the head features with depth information are fused with the features of the SSD network to form a two-channel SSD network. Then, on the basis of the two-channel SSD network, the high-level feature map with rich semantic information is fused with the low-level feature map to achieve more accurate head location. Finally, the prior frame of SSD is re-adjusted to reduce the computational complexity of the SSD network. Experimental results show that in the case of illumination and occlusion, the detection accuracy of the improved algorithm is improved by 12.9 percentage points compared with the traditional SSD target detection algorithm, and it can effectively solve the influence of illumination changes and occlusion on pedestrian head detection.

Key words machine vision; pedestrian head detection; SSD network; depth information; multi-scale feature fusion

OCIS codes 150.1135; 100.4996; 100.3008; 100.2960

1 引 言

随着计算机技术的迅速发展，行人检测在智能交通、视频监控和客流量统计等领域得到了广泛的应用。在街道、商场和车站等公共场所内，行人检测能够实现场所内人员的流量分析和预测，从而提升

收稿日期：2021-07-26；**修回日期**：2021-08-22；**录用日期**：2021-09-02

基金项目：吉林省重点科技发展计划(20180201042GX)

通信作者：*huranran111@126.com

区域的安全性和公共资源的利用率。然而，行人容易受到遮挡、光照和复杂背景等因素的影响，使得目标检测算法在应用的过程中难以获得较高的检测精度。相对于人体目标，行人头部目标受到遮挡等不利因素的影响较小，所以更加有利于对行人进行检测[1-4]。

近年来，为了得到高精度的行人检测算法，基于深度学习的行人检测网络得到了广泛的关注并相继取得了较好的研究成果。主流的基于深度学习的目标检测算法可分为两步(two-stage)目标检测算法和一步(one-stage)目标检测算法[5]，其中 two-stage 目标检测算法首先对图片上的目标进行定位以得到目标的候选区域，然后对候选区域进行识别和分类。基于深度学习的目标检测算法有 R-CNN(Region-Convolutional Neural Network)、SPP-Net(Spatial Pyramid Pooling Network)、Fast R-CNN 和 Faster R-CNN[6-9]，这些算法虽然通过选择性搜索和区域提议网络(RPN)的改进提高了目标检测的精度，但行人姿态各异且易受外界干扰，导致目标检测的准确率较低。为了提高目标检测的精度，Li 等[10]提出了一种基于尺度感知的 Fast R-CNN 行人检测方法，通过引入多个内建的子网可以从不相交的区域内检测尺度不同的行人，然后将所有子网的输出自适应地组合在一起，从而生成最终的检测结果。Xie 等[11]将反卷积集融合 Faster R-CNN(DIF R-CNN)并采用 Inception-ResNet 模型来提供丰富的特征，以提高目标检测的准确率。文献[12]提出了一种基于聚类与 Faster R-CNN 的行人头部检测算法，通过新设计的距离度量方法结合 K-means＋＋算法可以得到更准确的聚类框，并优化了非极大值抑制(NMS)算法，该算法改善了行人间彼此因遮挡而造成召回率不佳的问题。然而，以上算法的网络结构较为复杂，检测效率较低，不能够很好地满足实时性的要求。相比于 two-stage 目标检测算法，one-stage 目标检测算法受到了学者们更多的关注，其网络可直接输出目标的类别与位置信息，检测速度较快。Redmon 等[13]在 2016 年提出 YOLO(You Only Look Once)目标检测算法，通过回归的思想来处理目标的位置和类别信息，可以实现端到端的优化以及较快的检测速度。紧接着，YOLOv2[14]和 YOLOv3[15]网络又相继被提出，通过对网络参数进行修改并将多个网络的特征信息进行融合，可以减少目标特征的丢失，在保证目标检测效率的同时，大大提高了目标检测的准确率。在 YOLO 系列算法被广泛研究的同时，SSD(Single Shot MutiBox Detectior)目标检测网络[16]也得到了研究人员的关注，SSD 网络是将 Faster R-CNN 中的锚框机制和 YOLO 中的回归思想结合，利用不同输出层的多尺度特征图进行目标检测。与 YOLO 相比，SSD 更适合检测小目标。为了进一步提高 SSD 检测小目标的精度，Fu 等[17]提出了基于 DSSD(Deconvolutional SSD)的目标检测算法，该算法将 ResNet 作为基础网络，加入解卷积层来提高浅层的表征能力以提高目标检测的鲁棒性。文献[18]提出了基于 MobileNet-SSD 的目标检测算法，将卷积神经网络中的 MobileNet 与 SSD 结合并对网络结构与参数进行调整，简化了模型结构。Li 等[19]将特征金字塔结构中选定的多尺度特征层与尺度不变卷积层进行融合，可以生成一组增强的多尺度特征映射，从而提高 SSD 对小目标的检测精度。为了实现行人头部的检测，文献[20]针对性地提出了一种优化的可形变区域全卷积网络，该网络通过在线难例挖掘(OHEM)算法与软非极大值抑制(S-NMS)算法来提升检测效果。

然而，以上算法在对行人头部进行检测的过程中都是对 RGB(Red, Green, Blue)图像进行处理，而彩色图像对光照较为敏感，在光线较强或者较弱的情况下，容易丢失目标的特征信息，从而影响目标检测的精度。考虑到深度图像可以显示出不同目标物体之间的空间位置，为了克服单一图像的检测弊端，受到文献[21-23]工作的启发，在原有 SSD 检测网络的模型上增加一个深度通道网络，从而构建双通道 SSD 目标检测网络。将 RGB 图像与其对应的深度图像同时作为网络的输入，并将 RGB 图像与深度图像的特征进行融合。此外，针对头部遮挡对算法检测结果的影响，以及 SSD 网络中高层特征图的特征信息对头部检测贡献较小的情况，将高层特征图的特征融合到低层特征图上以提高目标检测能力，从而解决头部遮挡的问题，进而提高目标检测算法的精度。最后，根据行人头部的尺寸信息重新设置 SSD 网络的先验框以减少计算量。在自制的行人头部数据集上对算法进行训练和分析，验证算法在遮挡和不同光照条件下的有效性。

2 改进 SSD 的头部检测网络

2.1 SSD300 网络模型

SSD 目标检测算法通过对网络中多尺度特征

图的特征信息进行提取和检测，再通过多框预测来输出目标的位置和类别信息。SSD300 目标检测的前置网络采用 VGG16（Visual Geometry Group 16），并将最后两层的全连接层改为卷积层，然后新增 4 个卷积层。SSD300 网络的结构如图 1 所示，其中 FC 为全连接层。

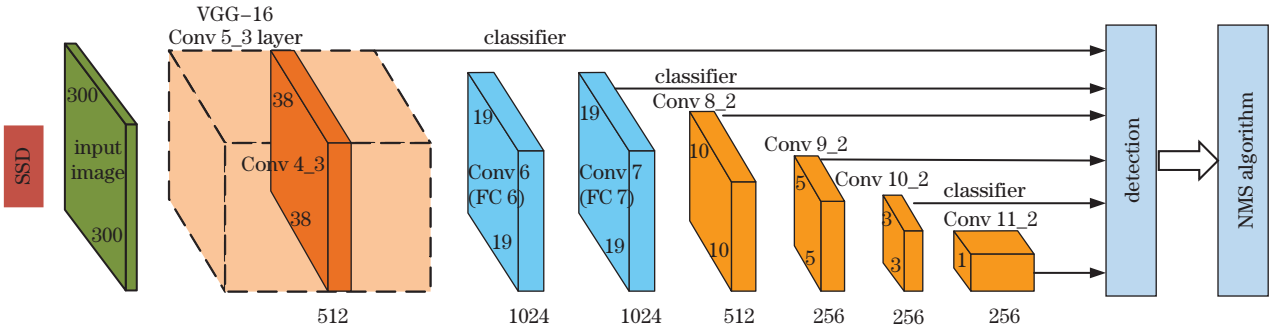


图 1　SSD300 网络的结构

Fig. 1　Structure of SSD300 network

SSD300 利用多层特征图对不同尺度的目标进行检测，然后通过 NMS 算法对所有的候选框进行预测以剔除重复候选框。与 two-stage 目标检测算法的不同之处在于，SSD 目标检测算法在检测的过程中不使用 RPN（Region Proposal Network），就可直接预测目标的类别和位置信息；针对不同的特征图，SSD 目标检测算法可以通过设置不同尺寸的先验框来匹配目标。因此，SSD 算法在检测效率和精度方面表现都较好。然而，SSD 网络以 RGB 图像作为输入，而行人头部信息对光照比较敏感，在光线较弱的条件下会影响目标检测的准确率。此外，SSD 网络中高层特征图的语义信息与低层特征图的特征信息没有得到充分利用，因此在处理遮挡目标的过程中检测效果也较差。

2.2　双通道 SSD 目标检测算法

彩色图像对光照变化较敏感，导致行人头部图像会丢失某些特征信息，而深度图像中含有行人头部的深度信息，对光照变化的鲁棒性更强。因此，在原有的 RGB 图像通道中增加一条深度图像通道，形成 RGB-D 双通道网络，网络结构如图 2 所示。首先选用两个 VGG16 子网络分别处理 RGB 图像和深度图像，然后在两个子网络中选择具有相同感受野的特征层进行层次融合，可以提取目标的彩色和深度信息，能够解决光照变化对目标检测的影响。

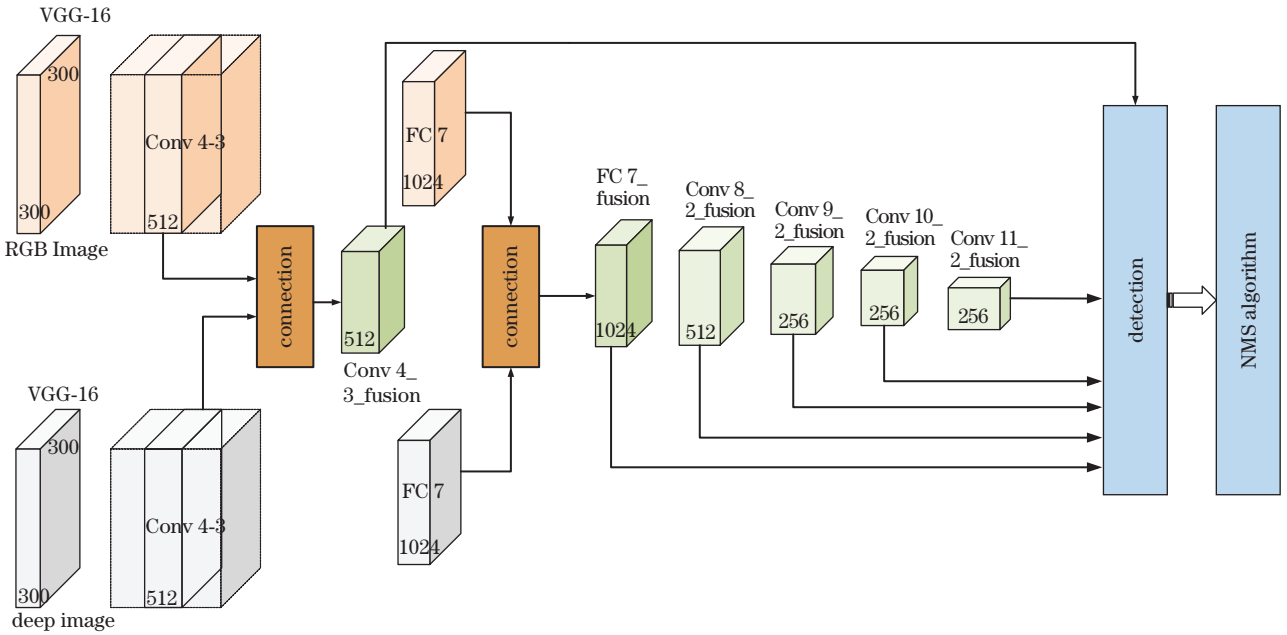


图 2　双通道网络的结构

Fig. 2　Structure of two-channel network

从图 2 可以看到，在特征融合的过程中，将 RGB 图像通道的 Conv 4_3 层和 FC 7 层分别与深度图像通道的 Conv 4_3 层和 FC 7 层通过 concat/eltwise 的融合方法进行特征融合，分别得到 Conv 4_3_fusion 层和 FC 7_fusion 层并将输出的结果作为后续特征层的输入，最后得到带有深度信息的 Conv 8_2_fusion 层、Conv 9_2_fusion 层、Conv 10_2_fusion 层和 Conv 11_2_fusion 层。若特征层间的融合方式为 concat 融合，需对融合后的特征图进行 1×1 的卷积操作，从而得到与原图像相同大小的特征图。

2.3　多尺度融合

在 SSD 网络中，低层特征图适合检测小目标，高层特征图含有丰富的语义信息，通过将高层特征图与低层特征图的融合能够增强 SSD 网络的性能[16]。因此，在双通道 SSD 网络的基础上进行了改进，首先采用反卷积来增大 Conv 4_3_fusion 层的特征图，然后对具有丰富语义信息的 Conv 10_2_fusion 层和 Conv 11_2_fusion 层的特征图通过上采样进行融合，最后再与改进的 Conv 4_3_fusion 层的特征图进行拼接以充分利用高层的语义信息，从而提高算法的检测准确率。多尺度融合的 SSD 检测模型如图 3 所示，其中 DConv 为解卷积层，MConv10_2_fusion 为高层的 Conv10_2_fusion 层与经过上采样操作后的 Conv11_2_fusion 层相融合后再进行上采样操作得到的新的卷积层，MDConv4_3_fusion 为将 DConv4_3_fusion 特征图和 MConv10_2 特征图采用 Concat 方式进行连接并进行 1×1 卷积操作得到的新的卷积层，⊕为叠加操作。

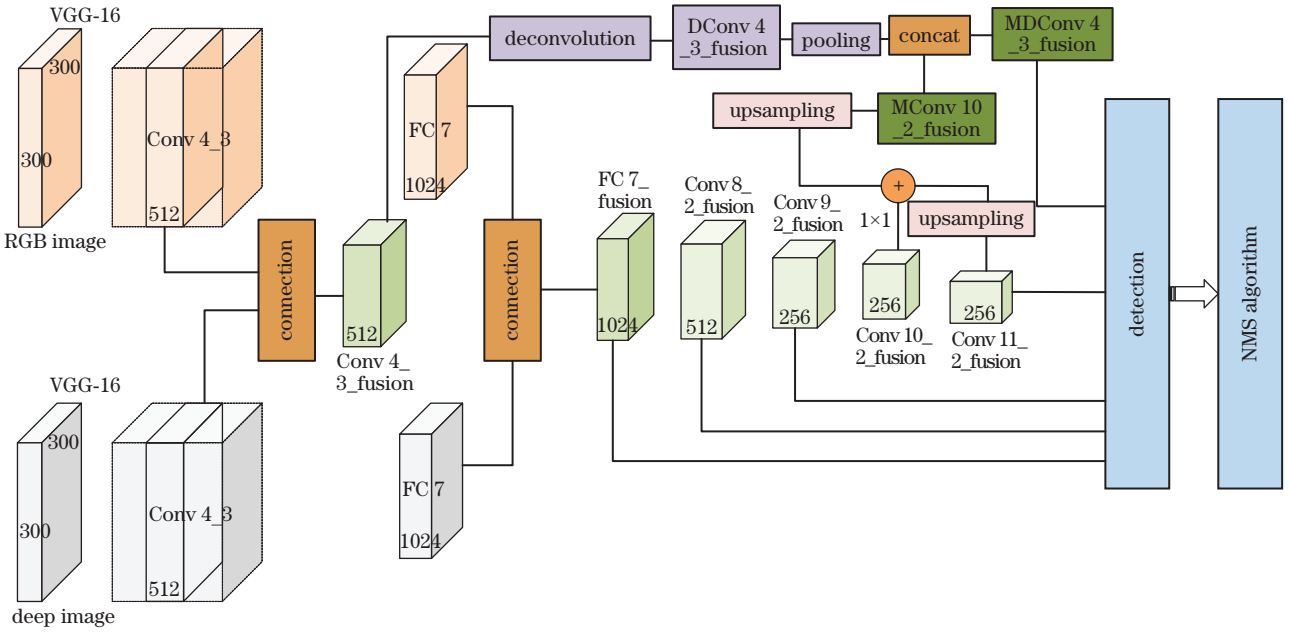


图 3　基于特征融合的改进双通道 SSD 网络的模型

Fig. 3　Improved two-channel SSD network model based on feature fusion

多尺度融合的 SSD 检测模型的具体处理步骤如下。

1）反卷积处理可增大特征区域的分辨率，扩大感受野，使其具有更加详细的边缘信息[17]。首先对 Conv 4_3_fusion 层的特征图进行反卷积操作后输入到 DConv 4_3_fusion 层，再将 DConv 4_3_fusion 层的特征图进行池化操作以减少卷积层输出的特征向量，使 DConv 4_3_fusion 层的输出特征图的大小仍为 38 pixel×38 pixel 以防止网络过拟合。使用 VGG16 卷积得到 Conv 3_3 层特征图的尺寸为 75 pixel×75 pixel，Conv 4_3 层特征图的尺寸为 38 pixel×38 pixel，反卷积输入输出公式为

$$i=(o-1)s+k'-2p, \tag{1}$$

式中：i 和 o 分别为输出和输入特征图的大小；s 为步长；k' 为卷积核大小；p 为填充。本文设置卷积步长 $s=2$，卷积核尺寸 $k'=1$，填充 $p=0$。

2）虽然高层特征图含有丰富的语义信息，但是 Conv 10_2_fusion 和 Conv 11_2_fusion 特征图的尺寸分别为 3 pixel×3 pixel 和 1 pixel×1 pixel，可融合的特征少，因此将 Conv 11_2_fusion 层的特征图与 Conv 10_2_fusion 层的特征图进行叠加操作，叠加后的特征图通过上采样操作可以得到与池化处理后的 DConv 4_3_fusion 层特征图相同尺寸的 MConv 10_2_fusion 层特征图。

3）特征融合。最后将 DConv 4_3_fusion 层特征图和 MConv 10_2 层特征图通过 concat 来连接并进行 1×1 的卷积操作，可以得到新的 MDConv 4_3_fusion 层特征图。MDConv 4_3_fusion 层可以代替原来的 Conv 4_3 层，其可以充分地学习行人头部特征以提高遮挡目标检测的准确度。

2.4 先验框

在改进的 SSD 网络模型中，MDConv 4_3_fusion、FC 7_fusion、Conv8_2_fusion、Conv 9_2_fusion、Conv10_2_fusion 和 Conv11_2_fusion 这 6 层的特征图大小分别为 38 pixel×38 pixel、19 pixel×19 pixel、10 pixel×10 pixel、5 pixel×5 pixel、3 pixel×3 pixel 和 1 pixel×1 pixel，先验框尺度的大小变换为线性变换，其公式为

$$S'_k = S_{\min} + \frac{S_{\max} - S_{\min}}{m-1}(k-1),\ k \in [1,m], \tag{2}$$

式中：S'_k 为先验框尺寸相对于原尺寸的比例；$S_{\max}$ 和 $S_{\min}$ 分别为比例的最大值和最小值；m 为特征图的个数。实验中将 $S_{\max}$ 值和 $S_{\min}$ 值分别设为 0.9 和 0.2。其中，在 MDConv 4_3 层特征图中，先验框的基础尺寸为 30 pixel。在 SSD 网络模型中，第一层、第五层和第六层先验框默认的长宽比 $a_r=\{2,1/2\}$，在每层特征图的中心点处可以分别产生 4 个默认框，第二～四层先验框默认的长宽比 $a_r=\{2,1/2,3,1/3\}$，在每层特征图的中心点处可以分别产生 6 个默认框，每个先验框的尺寸可表示为

$$W_l = S_l \cdot \sqrt{a_r}, \tag{3}$$

$$H_l = S_l / \sqrt{a_r}, \tag{4}$$

式中：W_l 和 H_l 分别为先验框的宽和高；S_l 为特征图生成小尺寸先验框的边长；l 表示特征图的层数。

当特征图的尺寸越小时，该层特征图上的预选框所检测到的目标就越大，而行人头部的尺寸较小，不会超过整个图像尺寸的一半，所以 Conv 9_2_fusion、Conv 10_2_fusion 和 Conv 11_2_fusion 这三层特征图上所产生的默认框尺寸偏大，对检测结果的贡献较小，因此将 Conv 9_2_fusion、Conv 10_2_fusion 和 Conv1 1_2_fusion 这三层去掉，可以保证在准确率不变的情况下减少网络参数，提高检测速度。改进后的 SSD 网络模型将 MDConv 4_3_fusion、FC 7_fusion 和 Conv 8_2_fusion 这三层特征图作为检测所用的特征图，不同先验框的长宽比下每层特征图的中心点处产生的默认框个数分别设置为 4、6 和 3，共产生 8242 个默认框。最终每层默认框的尺寸如表 1 所示。

表 1 每层默认框的尺寸

Table 1 Size of default box for each layer

Parameter	MDConv 4_3_fusion	FC 7_fusion	Conv 8_2_fusion
Feature map size/(pixel×pixel)	38×38	19×19	10×10
Number of default boxes	4	6	3
a_r	{1,1,2,1/2}	{1,1,2,1/2,3,1/3}	{1,2,1/2}
Small side length	30	60	111
Large side length	60	111	—

3 实验结果与分析

3.1 网络训练

由于目前没有公开的基于实际场景且垂直拍摄的行人头部数据集，为此实验采用自制的行人头部数据集(VS-Head)以满足算法验证的需求。通过在距离天花板 3.0～3.5 m 高度处安装双目相机，在实验室的电梯和业务中心大厅拍摄了不同光照、不同遮挡情况下的密集行人视频序列，视频序列共包含了三个场景，总时长约为 2000 s，据此视频序列来得到彩色行人头部图像。为了得到相应的深度图像，首先对获取到的彩色图像应用文献[24]的方法进行立体匹配操作，得到视差图的平均误匹配率为 7.03%，再将视差图转化为深度图像，部分深度图像示例如图 4 所示。

为了提升检测算法的性能，以及使图像中的特征信息得到最大化的利用，将获取的彩色图像和深度图像分别旋转 180°，旋转后再沿 x 轴和 y 轴进行翻转，最终得到 4000 张彩色图像和深度图像对，部分图像对如图 5 所示。将行人头部数据集按照 7∶3 的比例分为训练集和测试集，其中训练集包含 2800 张图片，测试集包含 1200 张图片。在网络训练的过程中，网络的优化算法采用随机梯度下降(SGD)法，学习率设为 0.0001，批处理大小为 16，输入的 RGB-D 图像尺寸为 640 pixel×480 pixel，最大迭代次数为 40000 次。训练环境为 Windows10，Caffe，GTX1080Ti 16 GB，生成的训练损失和精度曲线如图 6 所示。

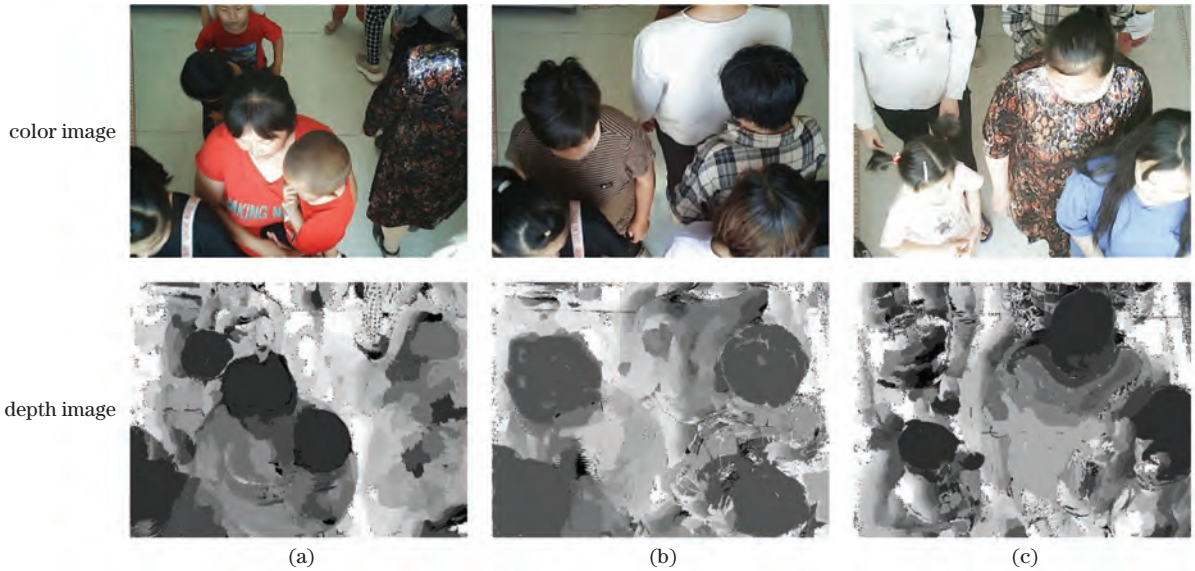


图 4　部分深度图像。(a)图像 1;(b)图像 2;(c)图像 3

Fig. 4　Partial depth images. (a) Image 1; (b) image 2; (c) image 3

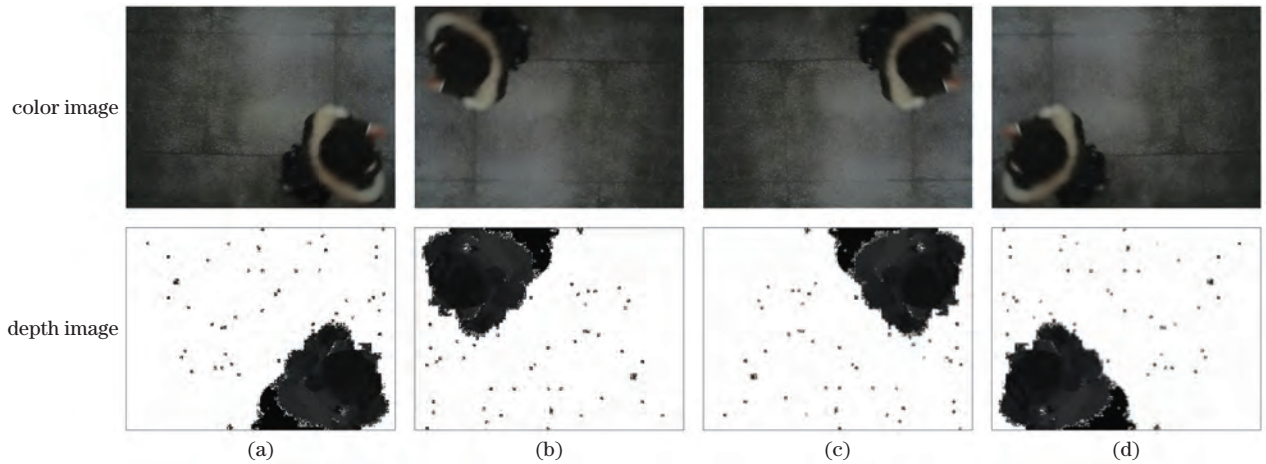


图 5　不同操作下的彩色图像和深度图像。(a)原始图像;(b)180°旋转;(c) x 轴翻转;(d) y 轴翻转

Fig. 5　Color images and depth images under different operations. (a) Original images; (b) 180 ° rotation; (c) x-axis reversal; (d) y-axis reversal

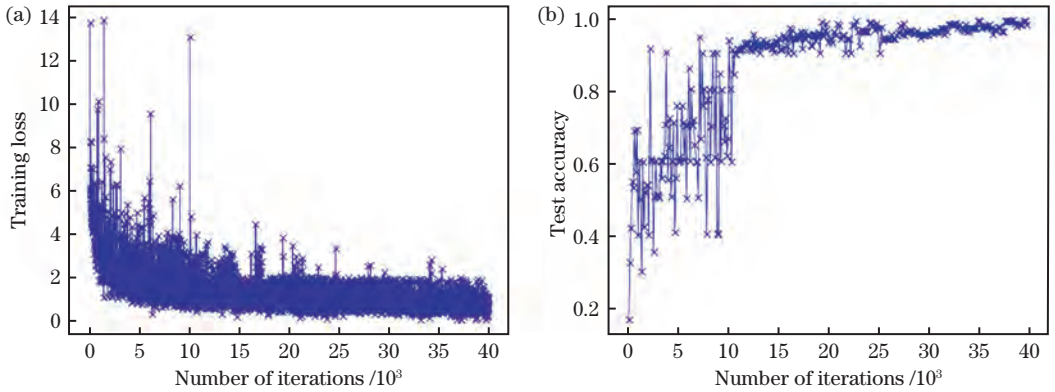


图 6　训练损失和精度曲线。(a)训练损失曲线;(b)精度曲线

Fig. 6　Training loss and accuracy curves. (a) Training loss curve; (b) accuracy curve

从图 6 可以看到,训练损失在迭代 10000～20000 次之间趋于平稳下降,在迭代 20000 次之后趋于收敛,在迭代次数达到 25000 之后,平均检测精度可达 0.978。

3.2 实验验证与分析

3.2.1 先验框实验的对比分析

为了验证简化先验框后的 SSD 网络对目标检测精度的影响，使用只有 Conv 4_3、FC 7 和 Conv 8_2 这三层作为输出的 SSD 网络与传统 SSD 网络分别在自制的数据集上进行训练和对比。实验结果如表 2 所示，其中 FPS 为每秒能够检测的帧数，mAP 为平均精度。

表 2 不同先验框网络的目标检测平均精度和速度

Table 2 Average accuracy and speed of target detection in different prior frame networks

Network	mAP/%	FPS
Simplified SSD network	83.70	33
SSD network	84.90	29

从表 2 可以看到，简化后的 SSD 网络虽然损失了一定的精度，而且与传统 SSD 网络的检测精度相差较小，但检测速度有所提升，为后续改进 SSD 网络中复杂的结构减少了一定的计算量。由此可以证明，简化后的 SSD 网络可以作为基础网络进行改进，后文的 SSD 网络均为简化先验框后的 SSD 网络。

3.2.2 双通道 SSD 网络和多尺度 SSD 网络实验分析

为了验证双通道 SSD 网络相对于传统 SSD 网络在光照变化的条件下是否有所改进，将 RGB 图像与深度图像融合的双通道 SSD 网络与传统 SSD 网络进行对比，并将 eltwise 融合方式下 RGB 图像通道与深度图像通道的权重比分别设置为 0.7∶0.3、0.5∶0.5 和 0.3∶0.7，用来验证不同融合方式对目标检测的效果，双通道网络的对比实验结果如表 3 和图 7 所示。图 7 中 RGB-SSD 和 Depth-SSD 分别表示以彩色图像和深度图像作为网络输入的 SSD 网络，Concat-SSD 表示彩色图像通道与深度图像通道以 concat 方式融合得到的双通道 SSD 网络；Three-SSD、Five-SSD 以及 Seven-SSD 表示在不同权重比下得到的双通道 SSD 网络。

表 3 双通道 SSD 网络的平均检测精度

Table 3 Average detection accuracy of two-channel SSD network　　unit: %

Fusion mode	Weight ratio	RGB image	Depth image	Two-channel SSD
Concat				89.86
Eltwise	0.7∶0.3	83.70	82.61	93.59
	0.5∶0.5			91.94
	0.3∶0.7			87.46

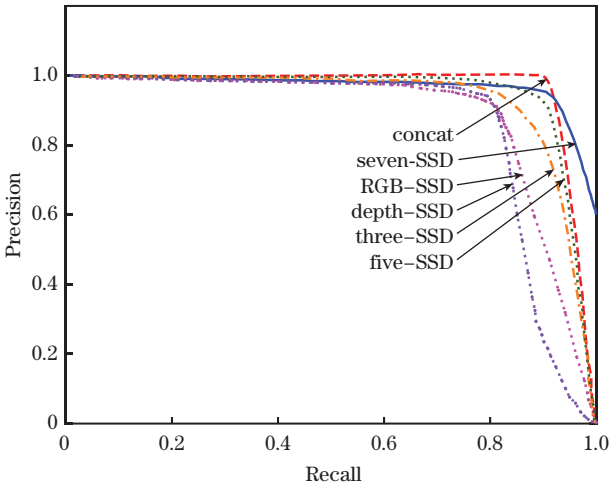


图 7 不同网络的精度与召回率的关系

Fig. 7 Relationship between precision and recall of different networks

从表 3 和图 7 可以看到，RGB 图像通道和深度图像通道所融合的双通道 SSD 网络的目标检测精度高于传统 SSD 网络。在双通道 SSD 网络中，当 RGB 图像通道与深度图像通道的权重比为 0.3∶0.7 时，concat 融合方式的准确率高于 eltwise 融合方式；然而，从整体上来讲，eltwise 融合方式的平均检测精度高于 concat 融合方式。其中，在 eltwise 融合方式下不同权重比例所获得的目标检测精度是不同的，当 RGB 图像通道与深度图像通道的权重比为 0.7∶0.3 时（如图 7 实线所示），所获得的目标检测精度最高，说明融合深度信息的双通道 SSD 网络可以提高目标检测的准确率，而且深度图像和彩色图像在目标检测过程中都发挥作用，彩色图像的作用更加明显。图 8 为不同光照条件下的部分实例检测结果。

从图 8 可以看到，SSD 网络在光照变化的情况下容易出现漏检的情况，而 RGB 图像通道和深度图像通道相融合的双通道 SSD 网络可以在明亮和较暗等光照条件下将目标信息准确地检测出来，说明融合图像的深度信息能够有效解决光照变化对目标检测的影响。

为了验证基于低层和高层特征融合的 SSD 网络结构能够有效解决遮挡所带来的问题，实验将简化先验框后的 SSD 网络结构进行多尺度融合，而且不加入深度通道。将简化先验框后的 SSD 网络在本文制作的数据集上进行训练，并与 SSD 网络和 DSSD 网络进行对比，得到的结果如表 4 所示。

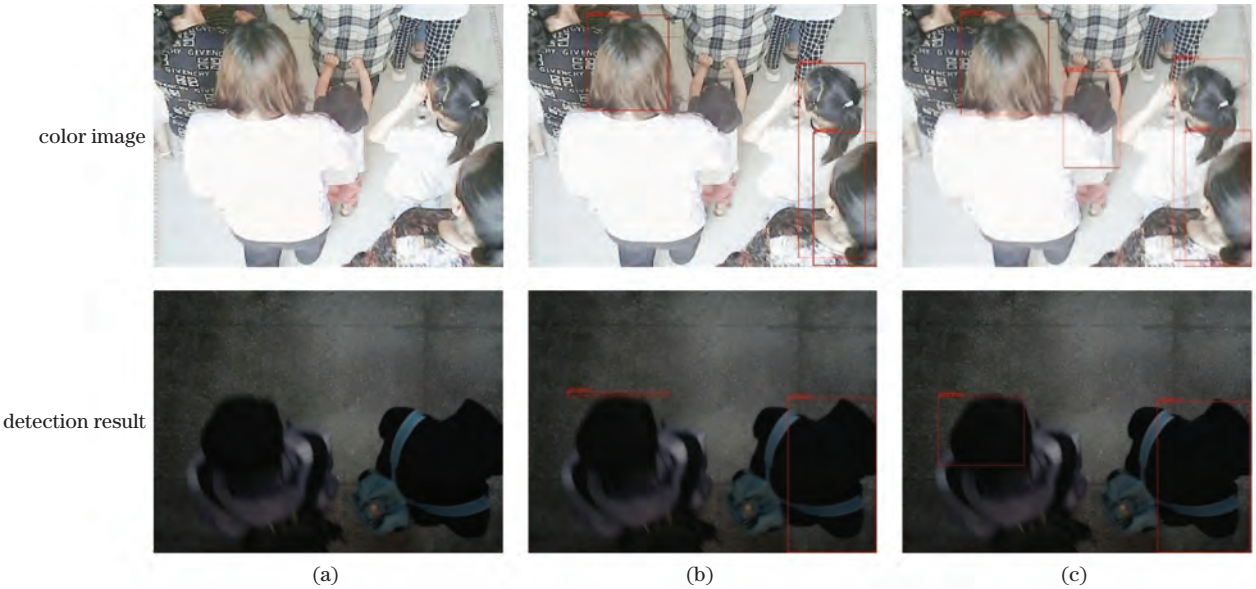


图 8　SSD 网络和双通道 SSD 网络的检测结果。(a)原图；(b)SSD 网络；(c)双通道 SSD 网络

Fig. 8　Detection results of SSD network and two-channel SSD network. (a) Original images; (b)SSD network; (c) two-channel SSD network

表 4　多尺度特征融合网络的平均检测精度

Table 4　Average detection accuracy of multi-scale feature fusion network　　unit: %

Parameter	SSD	DSSD	Multi-scale SSD
mAP	83.70	88.43	91.47

从表 4 可以看到，多尺度特征融合 SSD 网络的目标检测精度高于 SSD 网络和 DSSD 网络，说明多尺度特征融合 SSD 网络可以有效提高目标检测的精度。为了直观体现多尺度特征融合 SSD 网络对遮挡目标的表现性能，图 9 给出了部分遮挡目标的检测结果。

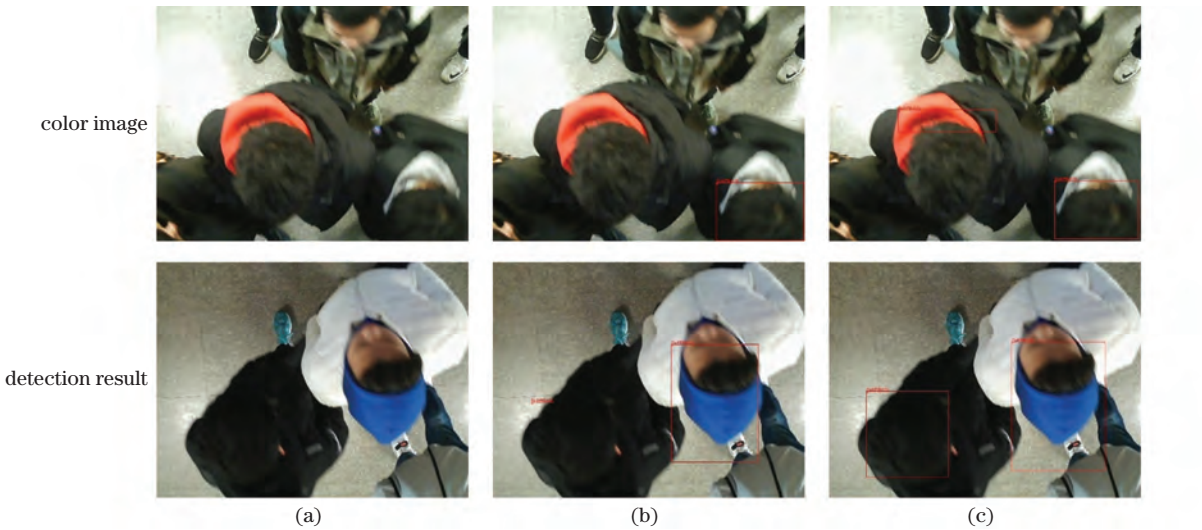


图 9　SSD 网络和多尺度特征融合的改进 SSD 网络的检测结果。(a)原图；(b) SSD 网络；(c)双通道 SSD 网络

Fig. 9　Detection results of SSD network and improved SSD network with multi-scale feature fusion. (a) Original images; (b) SSD network; (c) two-channel SSD network

从图 9 可以看到，当行人头部被部分遮挡或者只显示部分头部信息时，相较于传统 SSD 网络，多尺度融合 SSD 网络可以更准确地检测被遮挡的目标。虽然网络的检测准确率从整体上来讲不是十分理想的，但相对于传统 SSD 网络和 DSSD 网络均有明显提升，说明多尺度特征层融合的网络结构能够实现对遮挡样本的学习，提高网络的性能。

3.2.3　算法对比

为了验证基于多尺度特征融合的双通道 SSD 算法的有效性，将基于传统 SSD 网络、RGB-D＋YOLOv2 网络、RGB-D＋Faster-RCNN、MobileNet-

SSD 网络[18]、FPEF-SSD（Feature Pyramid-Enhanced Fusion-SSD）网络[19]和 SSD-Head 网络[25]的算法分别在自制的数据集上进行训练和对比，实验结果如表 5 所示，由于实验中只有行人一个类别，故 mAP 值为精度值（AP）。

根据表 2 可知，当双通道 SSD 网络模型在 eltwise 特征融合方式下 RGB 图像通道与深度图像通道的权重比为 0.7∶0.3 时，获得的目标检测精度最高，故基于多尺度特征融合的改进双通道 SSD 网络采用此种特征融合方式。

表 5　不同模型的平均检测精度

Table 5　Average detection accuracy of different models　　unit: %

Parameter	SSD	RGB-D+YOLOv2	RGB-D+Faster-RCNN	Ref. [18]	Ref. [19]	Ref. [25]	Proposed model
mAP	84.90	92.95	93.14	92.01	90.78	95.47	97.80

从表 5 可以看到，相比于传统 SSD 网络，双通道 SSD 网络在精确率上提高了 12.9 个百分点，并且优于其他先进算法。一方面，所提模型优于 RGB-D+YOLOv2 和 RGB-D+Faster-RCNN，证明其在行人头部检测的过程中具有适用性；另一方面，与文献[18-19，25]相比，所提模型具有更高的 mAP 值，证明了本文改进 SSD 网络的有效性。其中，文献[25]在本文自制的头部数据集上的检测结果仅稍逊于所提模型，主要是由于其网络结构为 RGB 单通道，这在一定意义上限制了网络的性能。此外，将所提模型分别与表 3 和表 4 的双通道 SSD 和多尺度 SSD 网络进行比较，基于多尺度特征融合的双通道 SSD 算法的目标检测精度分别提高了 4.21 个百分点和 6.33 个百分点，说明 RGB-D 双通道网络和多尺度特征在模型中为相互促进的作用，证明了在 SSD 网络中加入深度信息以及融合多尺度特征可以提高行人头部检测的精度。

为了更加直观地说明所提算法的优越性，图 10 对比了 SSD 算法和 DSSD 算法与所提算法在不同光照变化下的目标检测结果。

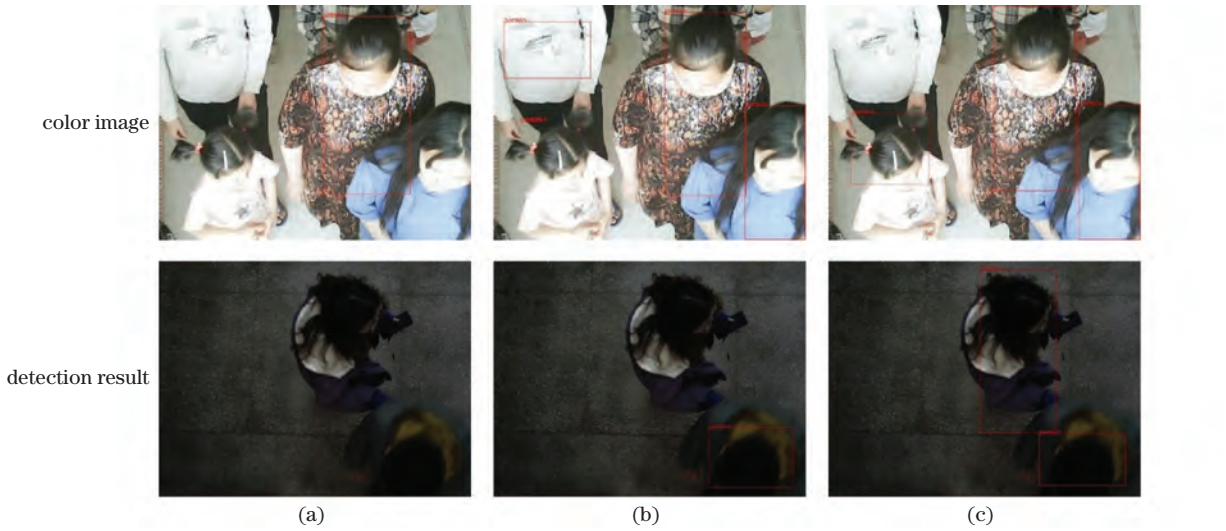


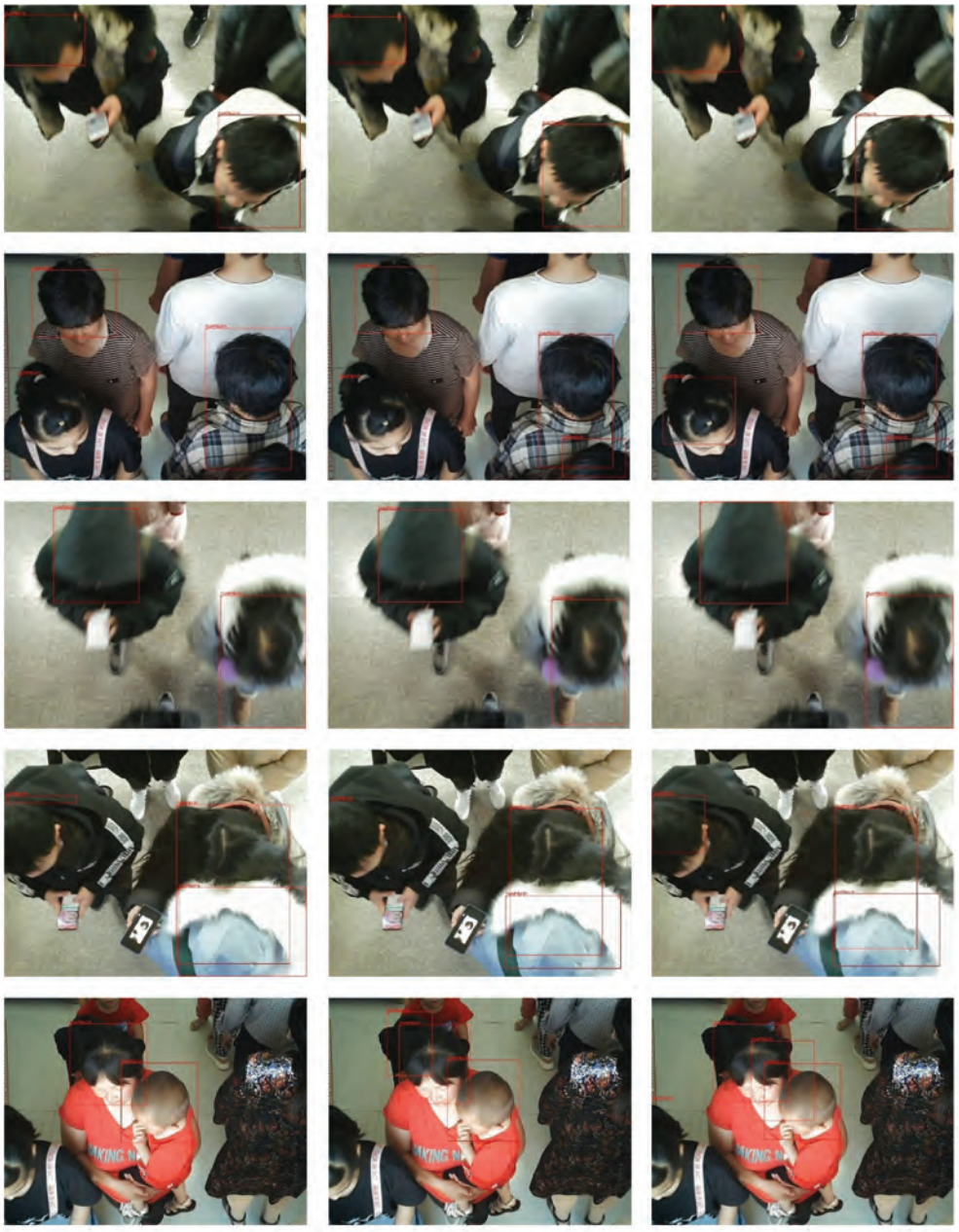
图 10　不同光照变化条件下各算法的检测结果。(a) SSD 算法；(b) DSSD 算法；(c)改进算法

Fig. 10　Detection results of each algorithm under different illumination variation conditions. (a) SSD algorithm; (b) DSSD algorithm; (c) improved algorithm

从图 10 可以看到，在过曝光和逆光的条件下，SSD 网络和 DSSD 网络出现明显的误检和漏检现象，这主要是由曝光和逆光的条件下彩色图像中的行人头部信息不突出导致的；无论是在过曝光还是逆光的环境下，所提算法都能够准确将图像中的行人头部检测出来，主要是带有深度信息的头部特征与 SSD 网络的特征的融合可以提升网络对光照变化的鲁棒性。结合表 2 可知，改进网络的目标检测精度也明显高于原始双通道 SSD 网络。此外，还给出了不同遮挡情况下的目标检测效果，如图 11 所示。其中第一张图像为几乎无遮挡情况下的目标检测效果，后 4 张图像为存在不同程度和不同类型遮挡情况下的行人头部目标检测效果。

从图 11 可以看到，在无遮挡的情况下，三种目标检测算法都能够较准确将图片中的目标检测出来，但是在遮挡的情况下，SSD 算法和 DSSD 算法均

存在明显的错检、漏检以及定位不准确的情况，这主要是由 SSD 网络中高层特征图的特征信息对头部检测贡献较小导致的；改进后的 SSD 算法可准确地将遮挡目标检测出来，这是因为改进网络通过将高层特征图的特征融合到低层特征图，所以提高了算法在遮挡条件下的目标检测精度。实验结果表明，改进后 SSD 算法的检测精度可达 97.8%，其精度高于其他目标检测算法，在光照变化和遮挡这种复杂的场景下都能够获得良好的行人头部检测结果，证明了改进后算法的有效性。



(a)　(b)　(c)

图 11　不同遮挡条件下各算法检测结果。(a) SSD 算法；(b) DSSD 算法；(c)改进算法

Fig. 11　Detection results of each algorithm under different occlusion conditions. (a) SSD algorithm; (b) DSSD algorithm; (c) improved algorithm

4　结　　论

为了解决由光照变化和遮挡造成行人头部检测精度低的问题，利用目标的深度信息与多尺度特征来改进原有的 SSD 网络模型。SSD 网络上增加一条深度通道以提取目标的深度特征信息，并将其与彩色图像的特征信息进行融合，从而形成双通道 SSD 网络，该网络可以提高光照变化情况下目标检测的准确率。同时，对低层特征图进行反卷积处理可以增强特征的分辨率，然后与含有丰富语义信息

的高层特征图进行融合，可以减少遮挡条件下目标的误检和漏检。将改进的算法与其他算法在行人头部数据集上进行训练、测试和对比，证明改进后算法的检测精度有明显提升，尤其是在光线较弱和存在遮挡的条件下，目标检测的准确率相比于其他算法提升更为显著。然而改进后算法的模型仍较复杂，计算量较大，目标检测的实时性有待进一步提升，在保证检测精度的情况下，进一步优化网络结构以提高目标检测速度是下一步主要的研究方向。

参 考 文 献

[1] Gu C, Qian W X, Chen Q, et al. Rapid head dectection method based on binocular stereo vision [J]. Chinese Journal of Lasers, 2014, 41(1): 0108001.
顾骋, 钱惟贤, 陈钱, 等. 基于双目立体视觉的快速人头检测方法[J]. 中国激光, 2014, 41(1): 0108001.

[2] Li J, Zhang F B, Wei L S, et al. Nighttime foreground pedestrian detection based on three-dimensional voxel surface model[J]. Sensors, 2017, 17(10): E2354.

[3] Zuo J, Ba Y L. Population-depth counting algorithm based on multiscale fusion [J]. Laser & Optoelectronics Progress, 2020, 57(24): 241502.
左静, 巴玉林. 基于多尺度融合的深度人群计数算法[J]. 激光与光电子学进展, 2020, 57(24): 241502.

[4] Khan S D, Basalamah S. Scale and density invariant head detection deep model for crowd counting in pedestrian crowds[J]. The Visual Computer, 2021, 37(8): 2127-2137.

[5] Khan S D, Ali Y, Zafar B, et al. Robust head detection in complex videos using two-stage deep convolution framework[J]. IEEE Access, 2020, 8: 98679-98692.

[6] 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.

[7] 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.

[8] Girshick R. Fast R-CNN [C] // 2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[9] 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.

[10] Li J N, Liang X D, Shen S M, et al. Scale-aware fast R-CNN for pedestrian detection [J]. IEEE Transactions on Multimedia, 2018, 20(4): 985-996.

[11] Xie H, Chen Y F, Shin H. Context-aware pedestrian detection especially for small-sized instances with Deconvolution Integrated Faster RCNN (DIF R-CNN)[J]. Applied Intelligence, 2019, 49(3): 1200-1211.

[12] Zhang J, Chen L, Li Z, et al. Pedestrian head detection algorithm based on clustering and Faster RCNN[J]. Journal of Northwest University (Natural Science Edition), 2020, 50(6): 971-978.
张洁, 陈莉, 李铮, 等. 基于聚类与 Faster RCNN 的行人头部检测算法[J]. 西北大学学报(自然科学版), 2020, 50(6): 971-978.

[13] 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.

[14] Liu Z M, Chen Z C, Li Z M, et al. An efficient pedestrian detection method based on YOLOv2[J]. Mathematical Problems in Engineering, 2018, 2018: 3518959.

[15] Zhang X L, Dong X P, Wei Q J, et al. Real-time object detection algorithm based on improved YOLOv3[J]. Journal of Electronic Imaging, 2019, 28(5): 053022.

[16] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M] // Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[17] Fu C Y, Liu W, Ranga A, et al. DSSD: deconvolutional single shot detector[EB/OL]. (2017-01-23) [2021-07-20]. https://arxiv.org/abs/1701.06659.

[18] Li Y T, Huang H S, Xie Q S, et al. Research on a surface defect detection algorithm based on MobileNet-SSD[J]. Applied Sciences, 2018, 8(9): 1678.

[19] Li H T, Lin K Z, Bai J X, et al. Small object detection algorithm based on feature pyramid-enhanced fusion SSD[J]. Complexity, 2019, 2019: 7297960.

[20] Ji X S, Wang H. Head detection method based on optimized deformable regional fully convolutional neutral networks [J]. Laser & Optoelectronics Progress, 2019, 56(14): 141009.
吉训生, 王昊. 基于优化可形变区域全卷积神经网络的人头检测方法[J]. 激光与光电子学进展, 2019, 56(14): 141009.

[21] Xu X Y, Li Y C, Wu G S, et al. Multi-modal deep feature learning for RGB-D object detection [J]. Pattern Recognition, 2017, 72: 300-313.

[22] Ophoff T, van Beeck K, Goedemé T. Exploring RGB+Depth fusion for real-time object detection[J]. Sensors, 2019, 19(4): 866.

[23] Luo Q H, Ma H F, Tang L, et al. 3D-SSD: learning hierarchical features from RGB-D images for amodal 3D object detection[J]. Neurocomputing, 2020, 378: 364-374.

[24] Gao M. Research on people counting algorithm based on binocular stereo vision [D]. Changchun: Changchun University of Science and Technology, 2019.
高淼. 基于双目立体视觉的客流统计算法研究[D]. 长春: 长春理工大学, 2019.

[25] Li H, Chen X Q, Shi H, et al. Pedestrian head detection method based on SSD [J]. Computer Engineering and Design, 2020, 41(3): 827-832.
李欢, 陈先桥, 施辉, 等. 基于 SSD 的行人头部检测方法[J]. 计算机工程与设计, 2020, 41(3): 827-832.