基于联合学习的多视角室内人员检测网络

王霞[1,2*]，张为[1,2]

[1]天津大学电气自动化与信息工程学院，天津 300072；

[2]天津大学微电子学院，天津 300072

摘要 建立了室内人员检测数据集(IHDD)，提出了基于联合学习的多视角室内人员检测网络模型(MVNN)。该模型由输入数据层、特征提取层、可变形处理层、可见性估计层、分类判别层等组成，并加入区域建议模型和多视角模型以提升算法的检测性能。在自建的 IHDD 数据集上的实验结果表明，与现有其他检测算法相比，MVNN 算法的检测率更高；在人体目标呈现多视角、多姿态及存在遮挡等困难情况下仍有不错的检测效果，具有一定的理论研究价值和实际应用价值。

关键词 图像处理；室内人员检测；卷积神经网络；多视角；联合学习；视频监控

中图分类号 TP391.4 **文献标识码** A **doi**：10.3788/AOS201939.0210002

Multi-View Indoor Human Detection Neural Network Based on Joint Learning

Wang Xia[1,2*], Zhang Wei[1,2]

[1] *School of Electrical Automation and Information Engineering, Tianjin University, Tianjin* 300072, *China*;

[2] *School of Microelectronics, Tianjin University, Tianjin* 300072, *China*

Abstract An indoor human detection dataset (IHDD) is established, and a novel multi-view indoor human detection neural network (MVNN) based on joint learning is proposed. The model consists of input data layer, feature extraction layer, deformation layer, visibility reasoning layer and classification layer, and the proposed MVNN algorithm can improve the detection performance when combined with the region proposal model and the multi-view model. Experimental results on the self-built IHDD show that compared with other existing detection algorithms, the proposed MVNN algorithm has a higher detection rate. It can still obtain good detection results even in the case of difficult situations such as various views, changing poses and occlusion for human targets, which indicates certain theoretical research value and practical value.

Key words image processing; indoor human detection; convolutional neural network; multi-view; joint learning; video surveillance

OCIS codes 100.4996; 100.5010; 100.3008

1 引 言

随着生活水平的不断提高，人们的安防意识也不断增强，视频监控设备也被广泛应用到生产和生活中，伴随计算机视觉技术的不断发展，智能视频监控应运而生。目标检测，尤其人体目标检测是智能视频监控领域中一项必不可少的研究课题，在自动驾驶、智能机器人、行为分析等方面应用广泛[1]，尤其在一些特殊场所，如值班室、消防控制室、军事哨所等，实时检测室内是否有人在岗对于保障人们的生命财产安全意义重大。

国内外有大量人体目标检测，尤其是行人检测的相关研究[2]。一般而言，人体目标检测框架主要包括特征提取和分类判别两个部分。特征提取部分可分为两大类。第一类是人工设计的传统特征，用来描述颜色、边缘、纹理、深度及梯度信息等，有时会添加运动信息、上下文信息等。最早的人工设计的传统特征以 Dalal 等[3]设计的方向梯度直方图(HOG)特征为代表，描述目标的边缘信息，可以适应目标的微小形变，成为后续研究的基础。之后有人将其与描述纹理

收稿日期：2018-08-06；**修回日期**：2018-09-03；**录用日期**：2018-09-17

基金项目：公安部技术研究计划竞争性遴选项目(2016JSYJD04-O3)、火灾调查视频图像分析关键技术研究(2017JSYJC35)

* **E-mail**：1257833489@qq.com

信息的局部二值模式(LBP)特征[4-5]、尺度不变特征变换(SIFT)特征[6]、描述输入图像矩形特征的 Haar-like 特征[7]等相结合。之后人工设计的传统特征则主要以 Dollar 等[8]提出的 ICF(Integral Channel Features)为代表,提出三种通道,即 LUV 颜色通道(L 表示亮度,U 和 V 是色度)、归一化梯度幅值通道和梯度直方图通道。后续一些方法也是基于这种思想,比如 ACF(Aggregate Channel Features)[9]、Informed Haar 特征[10]、LDCF(Locally Decorrelated Channel Features)特征[11]、Checkerboards 特征[12]等。这类提取手工设计特征的方法主要依据行人外形,并不能很好地区分具有相似外形特征的背景目标。第二类是提取神经网络特征的方法[13-14]。此外,还有将两种特征结合的方法,如 Cai 等[15]将传统特征与深度网络特征相结合以提升检测性能。分类判别方法首先从 HOG 特征+支持向量机(SVM)框架[3]中的 SVM 开始,采用线性与非线性核函数(如线性 SVM、直方图交叉核 SVM[16]、latent SVM[17]等)进行分类。后来出现了一些利用统计概率思想的分类方法,如基于提升思想的 Adaboost(Adaptive Boosting)方法等,这些方法对环境中的噪声等非目标因素较为敏感。随着相关技术的不断发展,出现了一些基于神经网络模型进行分类判别的方法,如 ChnFtrs[8]、JointDeep[18]等方法将多个模块进行集成,以实现最终的分类判别。

随着深度学习的不断发展,研究人员开始转向利用神经网络实现目标检测方法的研究,这种方法大致可以分为两大类:一类是以 R-CNN(Regions with CNN features)[19]为代表的两阶段网络模型,如 Fast R-CNN[20]、Faster R-CNN[21-22]等;另一类是单阶段模型,以 YOLO(You Only Look Once)系列[23]、SSD(Single Shot MultiBox Detector)[24]等为典型代表。这些网络模型最初都是针对二分类或多分类问题提出的,将这些网络模型用于人的检测效果并不理想,而且,其中大部分方法都是用于室外开阔环境中处于直立状态的行人检测[25],图像采集范围广、距离远,人体目标比较完整,且基本处于站姿或走动状态,呈现正面或背面视角。室内人体目标检测的情形则与之不同:采集图像的范围小,距离近,目标存在一定倾角,大部分时候仅部分人体躯干可见,基本上呈现的都是坐姿,少量为站姿,且存在正面、背面、侧面等多种视角,多种姿态变化以及不同程度的变形与遮挡,受到复杂背景等因素的干扰等。为此,研究人员提出了一些方法来解决这些问题。

针对多变外形、多姿态的问题,主要以可变形部件模型(DPM)方法[17]为代表,在整体模型的基础上集成了部件模型并考虑其变形因素,在一定程度上提升了检测性能;还有一些方法对部件的旋转、大小变化、形状等进行建模以应对更复杂的变形问题[26],但基本上都是利用传统特征。针对遮挡问题,一些方法对遮挡进行推理,以部件图像块的检测分数为基础估计部件可见性[4,27];还有一些方法利用其他手段,如进行图像分割处理或提取深度信息等辅助判断[26],有时也加入运动信息、上下文信息、帧间信息等,但都未考虑与模型其他部分之间的联系。

除了检测性能之外,检测速度也是一个重要考虑因素,主要体现在图像区域预选取阶段。最早的方法是对图像进行密集式的滑动窗口遍历搜索[28],然后提取特征进行检测,计算复杂、耗时长,对分类器性能要求高。后来,区域建议的方法陆续出现,将目标转向重点关注区域,大大降低了计算成本,提升了分类、检测等后续阶段的效率。一般可以将其分为两类:一类基于区域分组的思想,以 Selective Search[28]为代表;一类以窗口得分为基准,以 Edge Boxes[29]为代表,后者的速度更快。之后,很多网络模型开始采用这种区域建议的思想,如 DBN-Mut 模型使用 DPM 进行区域建议[30];JointDeep[18]、SDN(Switchable Deep Network)[31]采用 HOG+CSS(Color Self-Similarity)特征结合线性 SVM 的方法进行区域建议;还有不少网络模型使用 ICF 特征检测算子进行区域建议[32];Faster R-CNN[21]则基于一个网络模型实现区域建议。研究表明,除了提升算法效率之外,区域建议对检测性能也有一定的提升作用[33]。

基于以上研究基础,针对监控条件下的室内人体目标检测,本文提出了一种基于联合学习思想的多视角室内人员检测网络模型(MVNN)。该模型包括两大部分,即运用区域建议和多视角模型的预处理部分和多模块联合学习的检测网络部分。其中,区域建议部分可对输入图像进行预处理,排除大量的非目标区域;多视角模型有效提升了对多视角人体目标的召回率,提升网络的整体检测性能;多模块联合学习的检测网络可实现特征提取、可变形处理,并可对可见性进行估计,得到最终的分类结果。实验结果表明,所提出的算法模型比其他算法模型具有更高的准确率,尤其在处理变形与遮挡情况和多视角人体检测方面的性能更优。

2　基本原理

2.1　区域建议网络

分析采集的视频片段，主要对人体上半身进行标注。借鉴 Faster R-CNN[21]模型中的区域建议网络部分，网络结构（基于 VGG 模型）示意如图 1 所示。

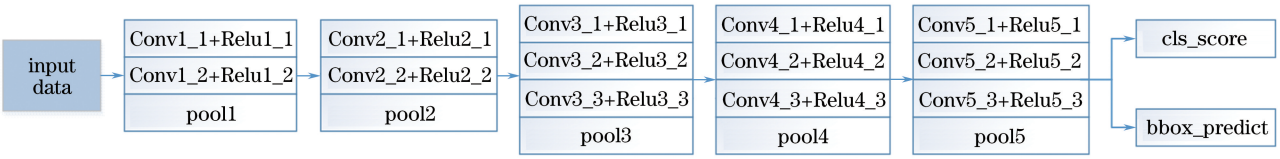

图 1　区域建议网络模型结构图

Fig. 1　Architecture of region proposal network

图 1 中，input data 表示输入图像数据；Conv 表示卷积层；Relu 为修正线性单元，是人工神经网络中常用的一种激活函数，可在一定程度上避免神经网络中容易出现的梯度消失等问题；pool 为进行 pooling 运算的池化层，常常接在卷积层之后，可以有效地减少运算参数，提升神经网络的整体运算性能；Conv1_1 与 Conv1_2 表示第一层级中的两个参数相同的卷积层，同理类推，共有 5 种卷积层，最终得到类别检测分数和目标的空间位置信息；cls_score 表示类别检测分数结果；bbox_predict 表示目标边界框坐标信息的预测结果。

该网络通过提取图像 CNN 特征预提取有较大概率存在目标对象的图像块。实验结果表明，与原算法[18]中预处理阶段的 HOG＋CSS 特征结合 SVM 的算法模型相比，使用该区域建议网络进行预提取具有更高的召回率。通常采用非最大值抑制（NMS）方法对区域建议结果进行进一步处理。依据 IoU（Intersection Over Union）指标，可用交并比 R_{IoU} 描述预测框 BB_1 与真实框 BB_2 之间的关系，即两者相交面积与集合面积的比值，计算公式为

$$R_{\mathrm{IoU}}=\frac{A_{\mathrm{area\_1}}}{A_{\mathrm{area\_2}}},\tag{1}$$

式中：$A_{\mathrm{area\_1}}$ 表示两者的相交面积；$A_{\mathrm{area\_2}}$ 表示两者的集合面积。将交并比 R_{IoU} 大于 0.5 的建议框预测为可能含有目标的预测框。设定阈值，从最高分预测框开始对所有的预测框进行 NMS 处理。最后按照分数从高到低的顺序保存这些区域建议块的坐标信息、可信度分数、类别标签及对应的图像块。

2.2　多视角模型

设计了三视角（包括正面、背面、侧面三种视角）模型。由于监控摄像头的安装位置固定，目标会存在一定倾角，同时呈现出多种姿态，尤其是侧面视角，且每种视角又存在不同程度的变形，多视角模型比单视角模型更适合这种应用环境，部分样本如图 2 所示。

图 2　多视角样本。（a）正面；（b）侧面；（c）背面

Fig. 2　Multi-view samples. (a) Frontal view; (b) profile view; (c) back view

这些多视角样本分为三类，包括正面、侧面、背面三种视角的人体目标。图 2（a）为正面视角样例，这种情况较少，因为摄像头的安装位置常在与桌椅成对角线的屋顶角落里，一般较难从正面拍摄到人脸；图 2（b）为侧面视角样例，这种情形最多；图 2（c）为背面视角样例，整个身体完全处于背面状态的情形并不多。

2.3　联合学习网络模型

联合学习网络主要包括特征提取、可变形处理、可见性估计、分类判定模块，网络模型框架如图 3 所示。该网络模型的设计借鉴 JointDeep 模型[18]（灰色部分为原有模块，蓝色部分为重新设计的模块），主要区别包括：预处理阶段的区域建议模型和多视角模型，MVNN 数据输入层以及可变形处理层。

区域建议网络和多视角模型部分主要是对大量的原始图像数据做预处理，提取最可能包含人体目

标的候选框，减少后面网络层的计算量，且保证了较高的召回率。MVNN 数据输入层的作用是将区域建议图像块做进一步处理，经过特征提取，添加适合本应用的可变形处理层，以应对遮挡及多视角、多姿态的情况，提升检测率和召回率。

在该基础模型上进行调整，使之适合本应用环境，主体网络结构及参数设置如图 4 所示。输入层样本大小为 84×84，后面部件映射层采用 15 种部件滤波器与前面特征映射层进行卷积，得到对应的部件检测分数，再经过可见性估计得到预测分数，根据预测分数判定是否为正类。

2.3.1　图像特征提取模块

将区域建议图像块作为输入，提取其颜色通道特征和边缘信息，最终以三通道图像的形式实现对图像信息的多层次表达。具体来说，第一通道是原图的 Y 通道；第二通道是由四个图像块拼接而成，包括 U 通道、V 通道、Y 通道和全零图像块；第三通道是计算前面的 U、V、Y 三通道图像块边缘信息并选择其最大值。三个通道都归一化到零均值-单位方差的分布，然后综合三通道信息描述图像，图 5 给出两个样例。

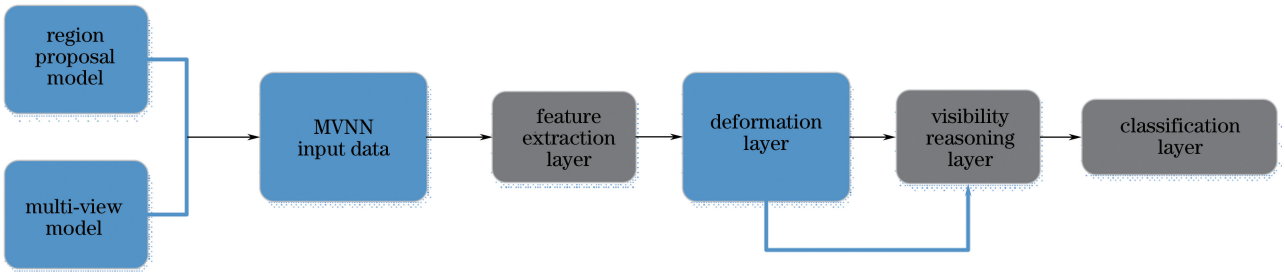

图 3　MVNN 模型框架

Fig. 3　Architecture of MVNN

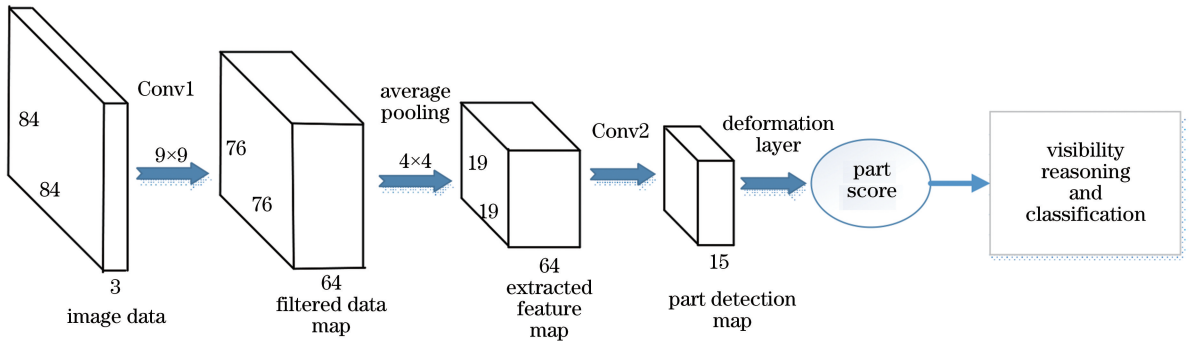

图 4　MVNN 主体检测网络模型框架

Fig. 4　Principal detection neural network model of MVNN

图 5　输入层三通道示意图。(a)样本 1；(b)样本 2

Fig. 5　Three channels of input layer. (a) Sample 1; (b) sample 2

图 5 中，对应每个样本的第一列是原图，第一通道信息并未在图中显示，第二、三列分别对应第二、三通道。此时，将其输入网络中的卷积层，进一步提取特征信息，并进行可变形处理、可见性估计及分类等。

2.3.2　目标可变形处理模块

室内环境中，由于监控摄像头安装位置固定，目标人员在监控视野中存在一定倾角，而且人的状态不定，呈现出多种视角，如背面、正面和侧面视角等，有时也会出现一些复杂情形，如人突然起身走动、弯腰捡物品、伸展双臂等，不同于通常的坐姿状态，这些情形给检测造成困难。为此，设计一种与之适应的 DPM，可以在一定程度上提高检测率。

部件模型源于 DPM[17] 的思想，主要是将原图像目标看成两个层次，即整体层次和分部件层次，综

合两层次得到最终结果，包括检测分数及位置信息。

如图 6 所示，向可变形层输入 p 个部件映射图，对应输出 p 个部件分数 $s=\{s_1,s_2,\cdots,s_p\}$，$p=15$。第 p 个部件的综合检测分数为

$$B_p=M_p+\sum_{n=1}^{N}c_{n,p}D_{n,p},\tag{2}$$

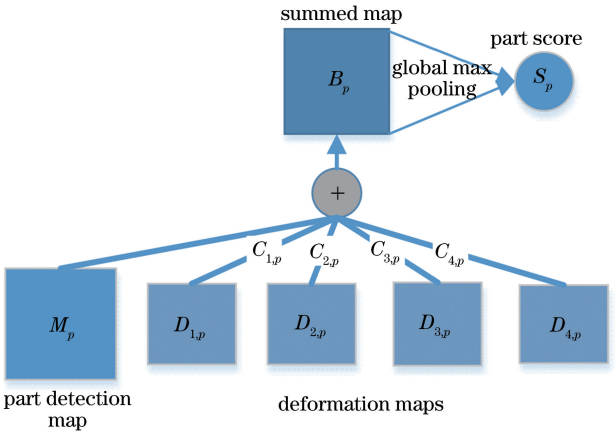

图 6　可变形层的部件分数计算模型

Fig. 6　Calculation model of part score for deformation layer

式中：M_p 为检测分数；$D_{n,p}$ 为第 p 个部件对应的第 n 个变形映射分数；$c_{n,p}$ 为第 p 个部件的权重系数，$c_{n,p}$ 和 $D_{n,p}$ 是需要学习的关键参数；N 为变形映射的个数，n 从 1 到 N 取值。第 p 个部件的最终检测分数为 s_p，由预测到的所有部件位置[用部件中心坐标(x,y)表示]上的部件综合检测分数 B_p 决定，再对 B_p 进行全局最大池化计算，得到的最大值即为最终部件检测分数 s_p，取得最大检测分数 s_p 的位置$(x,y)_p$ 即为部件预测位置，计算公式分别为

$$s_p=\max_{(x,y)}b_p^{(x,y)},\tag{3}$$

$$(x,y)_p=\underset{(x,y)}{\operatorname{argmax}}\,b_p^{(x,y)},\tag{4}$$

式中：$b_p^{(x,y)}$ 为 B_p 中对应坐标(x,y)处的部件检测分数，其计算公式为

$$b_p^{(x,y)}=m^{(x,y)}+c_1\left(x-a_x+\frac{c_3}{2c_1}\right)^2+c_2\left(y-a_y+\frac{c_4}{2c_2}\right)^2,\tag{5}$$

式中：$m^{(x,y)}$ 是 M_p 中(x,y)位置的元素，代表中心坐标位于(x,y)位置的部件初始检测分数；(a_x,a_y)是第 p 个部件的原定目标位置，通过 $c_3/2c_1$ 和 $c_4/2c_2$ 对该位置进行调整；c_1 和 c_2 决定变形损失。结合(2)式和(5)式，可得

$$B_p=M_p+c_1D_{1,p}+c_2D_{2,p}+c_3D_{3,p}+c_4D_{4,p}+c_5\cdot 1,\tag{6}$$

式中：$d_n^{(x,y)}$ 是 $D_{n,p}(n=1\sim4)$中坐标为(x,y)位置处的元素；$c_1\sim c_4$ 均是需要学习的参数，c_5 是一个定值。一些变量的具体计算式为

$$\begin{cases}b_p^{(x,y)}=m^{(x,y)}+c_1d_1^{(x,y)}+c_2d_2^{(x,y)}+c_3d_3^{(x,y)}+c_4d_4^{(x,y)}+c_5\\ d_1^{(x,y)}=(x-a_x)^2,\ d_2^{(x,y)}=(y-a_y)^2,\ d_3^{(x,y)}=x-a_x\\ d_4^{(x,y)}=y-a_y,\ c_5=c_3^2/(4c_1)+c_4^2/(4c_2)\end{cases}。\tag{7}$$

结合文献[18]中的部件模型，共设计了 15 种部件滤波器，参数设置如表 1 所示。

表 1　所提算法中部件滤波器参数

Table 1　Parameters of part filters in proposed algorithm

Parameter	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Starting line	1	1	4	4	1	1	1	4	1	1	1	1	−1	1	1
Ending line	3	3	9	9	3	9	9	9	3	9	9	9	9	9	9
Starting column	1	3	1	3	1	1	4	1	1	1	1	4	1	1	1
Ending column	3	5	3	5	5	2	5	5	5	5	2	5	5	5	5

使用 4 个变量来定义部件滤波器的对应位置，即起始行、结束行、起始列和结束列。每列代表一种部件滤波器，假设目标整体尺寸大小是 9×5，因此行的范围是 1～9 ，列的范围是 1～5。这 15 种部件滤波器分为 3 个层次，第 1～4，5～8，9～15 种部件分别为第 1，2，3 层次。高层次部件是由低层次的对应部件组成，如第 5 个部件由部件 1 和部件 2 组成，滤波器参数设置和研究目标相关。

2.3.3　可见性估计及分类模块

图像中会存在遮挡情况，如桌椅、植物、设备的遮挡，目标间的遮挡等，借鉴文献[18]，在 DPM 基础上估计可见性，可改善由遮挡导致的信息丢失问

题，提高对遮挡目标的检测率。计算该图像区域块的检测分数，估计可见性，给出最终的分类结果，基于可变形层输出的部件分数 $s=\{s_1,s_2,\cdots,s_p\}$ 进行可见性估计，计算公式为

$$\tilde{h}_j^l=\sigma(c_j^l+g_j^l s_j^l),\tag{8}$$

式中：h_j^l 表示第 l 层级的第 j 个部件的可见性，上波浪线表示预测；s_j^l 表示其检测分数；g_j^l 和 c_j^l 分别为 s_j^l 的权重和偏置项；$\sigma(t)$ 代表 sigmoid 函数，即 $\sigma(t)=[1+\exp(-t)]^{-1}$；$h^l$ 为处于第 l 层级的共 P_l 个部件的可见性，$\boldsymbol{h}^l=[h_1^l\ h_2^l\cdots h_{P_l}^l]^{\mathrm{T}}$。输入前面步骤得到的部件检测分数 s，计算部件的可见性并判断分类。对于相邻层级的部件，其可见性估计之间存在一定的转化关系：

$$\tilde{h}_j^{l+1}=\sigma(\tilde{\boldsymbol{h}}^{l\mathrm{T}}\boldsymbol{w}_{*,j}^l+c_j^{l+1}+g_j^{l+1}s_j^{l+1}),\ l=1,2,\tag{9}$$

式中：$\boldsymbol{W}^l$ 表示第 l 层的可见性估计 h^l 与 h^{l+1} 之间的转换矩阵；$\boldsymbol{w}_{*,j}^l$ 是 $\boldsymbol{W}^l$ 中的第 j 个元素，表示 $\tilde{h}_j^l$ 与 $\tilde{h}_j^{l+1}$ 之间的转化矩阵。模型预测的类别标签 $\tilde{y}$ 可表示为

$$\tilde{y}=\sigma(\tilde{h}^{3\mathrm{T}}w^{\mathrm{cls}}+b),\tag{10}$$

式中：w^{cls} 是对应 $\tilde{h}^3$ 的线性分类器，上标 cls 表示分类。同一层级的部件可见性相关，g_j^l，c_j^l，W^l，w^{cls} 和 b 为需要学习的参数。

所提模型中采用了二分类中常用的交叉熵对数损失函数，计算方法为

$$L=y_{\mathrm{gnd}}\lg\tilde{y}+(1-y_{\mathrm{gnd}})\lg(1-\tilde{y}),\tag{11}$$

式中：y_{gnd} 为部件的真实类别标签；L 为类别预测的交叉熵对数损失函数值。

根据随机梯度下降(SGD)准则进行网络训练，输入为前面得到的部件检测分数，将 SGD 训练过程中的参数计算分解为

$$\frac{\partial L}{\partial s_i^l}=\frac{\partial L}{\partial h_i^l}\frac{\partial h_i^l}{\partial s_i^l}=\frac{\partial L}{\partial h_i^l}h_i^l(1-h_i^l)g_i^l,\tag{12}$$

$$\frac{\partial L}{\partial h_i^3}=\frac{\partial L}{\partial\tilde{y}}\tilde{y}(1-\tilde{y})w_i^{\mathrm{cls}},\tag{13}$$

$$\frac{\partial L}{\partial h_i^2}=\boldsymbol{w}_{i,*}^2\left[\frac{\partial L}{\partial h^3}\odot h^3\odot(1-h^3)\right],\tag{14}$$

$$\frac{\partial L}{\partial h_i^1}=\boldsymbol{w}_{i,*}^1\left[\frac{\partial L}{\partial h^2}\odot h^2\odot(1-h^2)\right],\tag{15}$$

式中：$\odot$ 代表 $(U\odot V)_{i,j}=U_{i,j}V_{i,j}$ 的运算；$\boldsymbol{w}_{i,*}^l$ 为 $\boldsymbol{W}^l$ 的第 i 行元素；$\boldsymbol{w}_i^{\mathrm{cls}}$ 是 $\boldsymbol{w}^{\mathrm{cls}}$ 的第 i 个元素。

3　实验部分

实验平台是基于 64 位的 Ubuntu14.04 操作系统和 NVIDIA TITAN Xp GPU，采用的软件有 Matlab2014b、Python2.7，采用的深度学习框架为 tensorflow 及相关深度学习库。

3.1　性能评价指标

二分类问题中，通常采用查准率($R_{\mathrm{Precision}}$)、查全率(R_{Recall})来客观评价算法性能，就人体目标检测算法而言，将样例真实类别与算法的预测类别相比较。$R_{\mathrm{Precision}}$、R_{Recall} 计算式为

$$R_{\mathrm{Recall}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}},\tag{16}$$

$$R_{\mathrm{Precision}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}},\tag{17}$$

式中：N_{TP}、N_{FP}、N_{FN} 和 N_{TN} 分别为将人体样本正确预测为正样本数、非人样本错误预测为正样本数、人体样本错误预测为负样本数、非人样本正确预测为负样本数。R_{Recall} 反映分类器对正例的覆盖能力，$R_{\mathrm{Precision}}$ 反映分类器预测正例的准确程度。以查全率 R_{Recall} 和查准率 $R_{\mathrm{Precision}}$ 分别作为横纵坐标，绘制查准率-查全率曲线(P-R 曲线)，该曲线可以反映分类器对正例的识别准确程度和对正例的覆盖能力之间的权衡。$R_{\mathrm{Precision}}$ 和 R_{Recall} 往往此消彼长，R_{AP} 为 P-R 曲线与 x 轴围成的曲线下图形面积。该图形面积越大，分类器的效果越好。对于连续的 P-R 曲线，R_{AP} 计算方式为

$$R_{\mathrm{AP}}=\int_0^1 p(r)\mathrm{d}r,\tag{18}$$

式中：p 代表 $R_{\mathrm{Precision}}$；$R_{\mathrm{Precision}}$ 是关于 R_{Recall} 分段常数的函数。

人体目标检测领域，根据 Dollar 等[34]提出的工具箱，通常用 FPPI-MR(False Positives Per Image-Miss Rate)衡量人体目标检测算法的检测性能，把测试图片中包含人体目标的窗口切割出来，然后从不包含人体目标的测试集采样非人样本，将窗口作为测试集来评估算法的性能。相关分析表明[34]，相比 FPPW(False Positives Per Window)，FPPI 指标更合理。具体而言，给定一定数目的包含或不包含人体目标的样本图像，N_{Pos} 和 N_{Neg} 分别为正、负样本数，R_{FPPI} 表示分类器的误检率，R_{MR}(Miss Rate)与召回率 R_{Recall} 的和为 1，表示分类器的漏检率，R_{MR} 值越低，算法性能越好，计算公式分别为

$$R_{\mathrm{FPPI}}=\frac{N_{\mathrm{FP}}}{N_{\mathrm{Neg}}+N_{\mathrm{Pos}}}\times 100\%,\tag{19}$$

$$R_{\mathrm{MR}}=\frac{N_{\mathrm{FN}}}{N_{\mathrm{FN}}+N_{\mathrm{TP}}}\times 100\%。\tag{20}$$

为了验证所提算法模型对室内人员目标的检测

性能，将其与其他相关算法模型进行了对比，选择 R_{FPPI}-R_{MR} 和 R_{AP} 作为衡量多个算法性能的评价指标。

3.2　数据集建立

人体目标检测研究领域中，INRIA、Caltech、PASCAL VOC 等数据集在相关研究中使用较多，其中，PASCAL VOC 的图像标注方式及质量都较合理，常被用在一些比赛及课题研究中[35]。针对室内人员检测任务，目前尚没有公开相关专用数据集，因此，按照 PASCAL VOC 数据集的格式标准建立了一个监控环境下的室内人员检测数据集(IHDD)，数据采集自某企业值班室和工作室，对应的每种场所都在不同角落安装了摄像头，人的上半身基本可完整呈现在监控视野中。对一个月内的监控视频进行截取并采集图像，涉及白天和夜晚时段、不同性别和年龄段的人员以及正面、背面、侧面多种视角，大部分人员是坐姿及其变形状态，也有少量直立或倾斜站立的姿态。共采集了 14665 个图像样本，其中包含 8799 个训练样本，2933 个验证样本，以及 2933 个测试样本，一共标注了 17854 个人体目标，数据集构成如表 2 所示。同一种场所中的多个视频是由不同位置的摄像头采集得到的。

3.3　训练网络模型

所提网络模型包括两个部分。第一部分是区域建议网络，其训练方法和文献[21]一致，使用了 VGG-16 在 ImageNet 上的预训练模型。训练参数为：动量为 0.9，权重衰减为 0.0005，两阶段学习率分别为 0.001 和 0.0001。第二部分是联合学习的多视角检测网络，网络训练参数设置如表 3 所示。

根据学习率和模型误差间的变化关系，设定两阶段学习率分别为 0.025 和 0.0125，并将 epoch 设定为250。模型选择了交叉熵损失函数，使用SGD的方式进行训练。

表 2　室内人员检测数据集

Table 2　Indoor human detection dataset

Test environment	Total frames	Annotated humans No.
Office day 1	3912	5163
Office day 2	3157	4785
Office day 3	246	394
Duty room day 1	108	178
Duty room day 2	813	893
Duty room night 1	6063	6072
Duty room night 2	369	369
Training set	8799	10479
Validation set	2933	3701
Test set	2933	3674
Total number of samples	14665	17854

表 3　第二阶段网络模型参数设置

Table 3　Parameter setting of second-stage network model

Parameter	Epoch 1-150	Epoch 151-250
Learning rate	0.025	0.0125
Momentum	0.9	0.9
Batch size	80	80

3.4　MVNN 的切片分析

为了验证所提算法模型的有效性，下面对几个主要部分的作用单独进行研究。

3.4.1　区域建议网络

所提模型构建了区域建议网络，为说明此部分作用，与文献[18]中 HOG＋CSS＋SVM 的预处理算法模型进行了对比。由于对应源码不容易获得，此处采用具有类似检测思想的 HOG 特征结合 Adaboost 方法的模型替代该方法，测试结果如图 7 所示。

图 7　输入数据的区域建议结果比较。(a)HOG 特征结合 Adaboost 方法；(b)所提区域建议算法

Fig. 7　Comparison results of region proposal for input data. (a) HOG＋Adaboost algorithm; (b) Proposed region proposal algorithm

图 7 中共 4 个测试样例，图 7(a)对应 HOG 特征结合 Adaboost 方法，图 7(b)对应所提的构建区域建议网络的方法。分析前两个样例发现，所提算法在处理人体存在较大变形时仍有较好的检测性能，而文献[18]中的方法易出现误检，如对第一个样例中的行李包和第二个样例中的盆景出现了误检；文献[18]的方法对存在一定遮挡与变形的情形(如第三个测试样例)出现了漏检，而所提算法在该情况下仍有较好的检测效果；对于最后一个样例，两人体目标区域存在一定重叠与遮挡，文献[18]的方法只检测到一个目标，而所提算法在检测到两个人体目标的前提下，结合多视角模型，可以正确获得两个目标的检测结果。

出现上述现象的原因主要是算法特征类型不同，文献[18]利用的是传统 HOG 特征，仅提取边缘信息，对目标的表达能力有限，当出现其他特征类似的目标时容易误检，而所提算法利用深度网络特征，对目标的特征表达更充分，不易出现误检，也更易检测出同类的新测试样例。文献[18]算法的特征表达能力有限，当应对变形较大或存在遮挡的新样例时易出现漏检。

分析以上测试结果发现，所提的区域建议模型具有更高的召回率和更低的误检率，尤其在处理遮挡和多视角情形的时候检测效果更好。

3.4.2　多视角模型

视频监控下的人体目标往往不能显示出完整外形，且大部分都存在一定倾角与形变，并呈现多种视角。为此，设计了多视角模型，部分检测结果如图 8 所示。

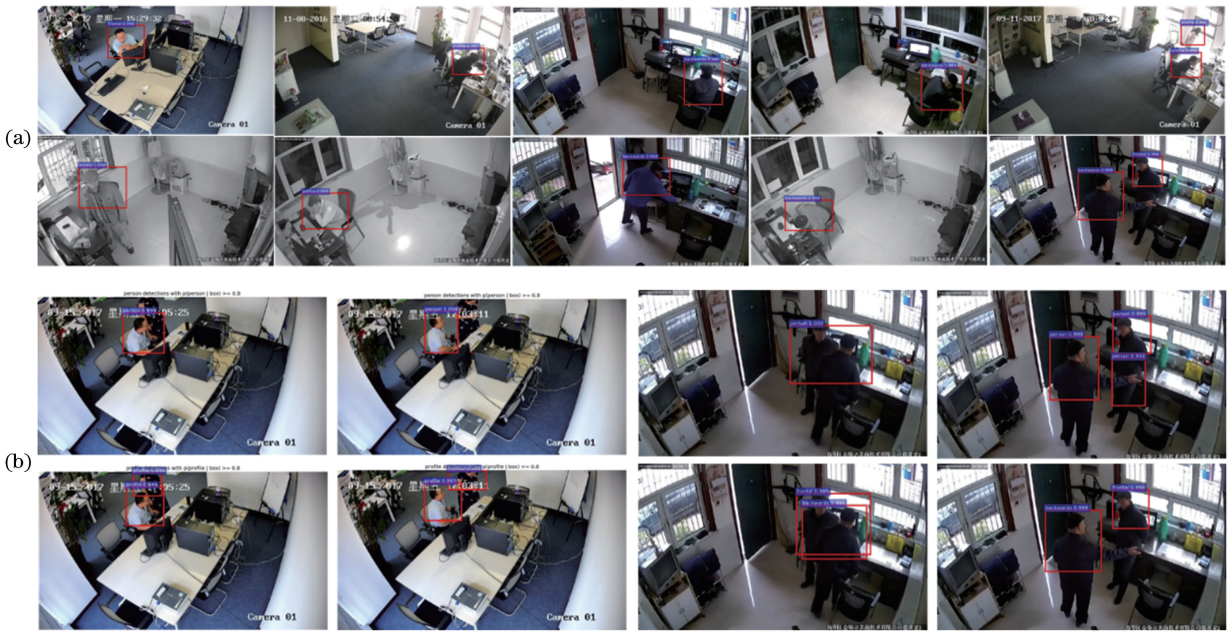

图 8　多视角模型测试结果。(a)不同视角下测试样本的检测结果；(b)单视角模型和多视角模型的检测效果对比

Fig. 8　Testing result of multi-view model. (a) Testing result of multiple views; (b) Comparison results of single-view model and multi-view model

图 8(a)所示为不同视角下测试样本的检测结果。第 1 列对应正面视角；第 2 列是侧面视角；第 3 列是背面视角，除了坐姿也有站立和行走中的情形；第 4 列也是背面视角，不过存在一定变形；第 5 列是包含两目标且可能为相同视角和不同视角的情形，分别对应了两侧面视角和正面＋背面视角。实验结果表明，多视角模型对多视角目标的检测效果较好。

图 8(b)是单视角模型和多视角模型的检测效果对比，每一列代表一个测试样本，两行分别对应单视角模型和所提多视角模型。对于前两个样本，多视角模型可将两个目标都检测出来，而单视角模型仅检出其中一个目标，漏检了存在遮挡的目标。第 3 列为两个人体目标之间存在遮挡且处于不同视角的情形，在这种情况下，单视角模型仅检测出一个目标，而多视角模型根据不同视角类型正确检测出两个目标。分析可能原因：1)该结果与对检测结果进行非最大值抑制处理有关，单视角模型会抑制属于同类且相交面积较大的检测结果；2)单视角模型只检测出其中基本未被遮挡的人体目标。第 4 列是两目标几乎无遮挡但距离较近的情形，在这种情况下，单视角模型产生误检，多视角模型检测正确，可能原因在于多视角模型学习每种视角的特征，而单视角模型综合学习各种视角的目标特征，模型特征识别能力弱，遇到相似图像易产生误检。

总的来说，针对监控环境下的室内人员检测，多视角模型比单视角模型具有更高的检测率和召回率，不易产生误检。

3.4.3　可变形处理和可见性估计

使用可变形处理可在一定程度上应对目标形变，并通过估计可见性应对遮挡，文献[18]中已说明了其对提高模型性能的重要作用。部分测试如图 9 所示，实验结果表明，对存在变形或部分遮挡的目标，所提算法仍有不错的检测性能，检测率和召回率均有提升。

4　分析与讨论

所提算法的最终定量统计结果如图 10 所示，图 10(a)中为 R_{FPPI}-R_{MR} 指标下的测试结果，可得 $R_{\mathrm{MR}}=14.66\%$，图 10(b)中 P-R 曲线下面积表示的平均准确率 $R_{\mathrm{AP}}=87.34\%$。

图 9　DPM 测试结果

Fig. 9　Testing result of DPM

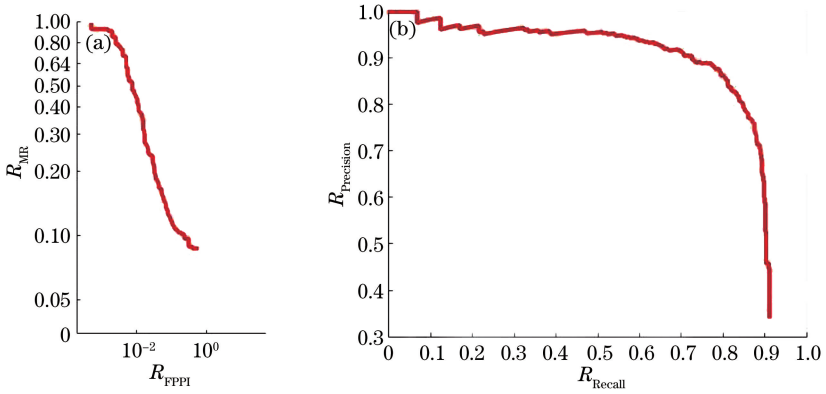

图 10　所提算法在 IHDD 上的测试结果。(a) R_{FPPI}-R_{MR} 曲线；(b) P-R 曲线

Fig. 10　Testing result of proposed algorithm on IHDD. (a) R_{FPPI}-R_{MR} curve; (b) P-R curve

为了进一步验证所提算法的有效性，将其与现有人体目标检测经典算法模型(包括文献[3]、文献[21]和文献[18]中的算法模型)进行对比，相关测试结果统计如表 4 所示。

表 4　不同算法在数据集上的定量比较

Table 4　Quantitative comparison of different algorithms on dataset

Algorithm measurements	R_{MR} (R_{FPPI}-R_{MR}) /%	R_{AP} /%
Ref. [3]	37.38	57.32
Ref. [21]	17.29	84.82
Ref. [18]	28.84	75.52
Proposed	14.66	87.34

表 4 中，文献[3]代表传统特征检测方法，即 HOG 特征结合 SVM 的方法，此处使用提升算法 Adaboost 替代 SVM，文献[21]是代表基于精度较高的神经网络检测模型 faster R-CNN 的方法，文献[18]为基于基准网络模型 JointDeep 的方法。

可以看出，对于监控环境下的室内人体目标检测，所提算法的召回率与检测率更高。分析原因如下：1)文献[3]的方法是在遍历图像的基础上提取人工设计的特征，在一定程度上表达了目标的边缘等信息，但是对于其他方面的信息(如颜色通道信息、纹理信息等)表达还不够充分，而且其他物体的边缘信息也可能和人体目标很相似，容易造成误检，通过网络学习提取的 CNN 特征则具有更丰富的表达能

力，可以提取出目标的更多代表性信息，更能做出可靠性判断。2）文献[21]的方法是采用两阶段的网络模型，通过第一阶段的区域预提取保证了较高的召回率，第二阶段在此基础上进行更精确地判断，但其针对的是通常情况下的目标分类与检测问题，没有考虑监控环境下室内人体目标的多视角、多形态、存在变形及遮挡的情形，因而检测性能并不理想。3）文献[18]基于基准网络模型，采用联合学习的方法处理目标的变形与遮挡问题，在第一阶段中采用传统目标检测方法进行区域预提取。这种方法速度快，但准确率和召回率并不高。类似地，该模型针对的是处于开阔环境中处于水平视角的行人，并不适合所提算法模型的研究环境，尤其是监控环境下室内人体目标存在一定倾角、呈现多视角、多形态、信息不完整的复杂情形。

综上所述，针对监控环境下室内人体目标检测任务，所提算法模型能够合理处理不同光照（如白天和黑夜）、多种形态、多视角、存在一定遮挡的情形，检测效果更好。

5　结　　论

针对现有室内人员检测算法在处理多视角、存在多种变形和遮挡等方面的不足，结合监控条件下室内人员检测任务的特点，搭建了一个联合学习的MVNN，集成了区域建议、多视角模型、特征提取、可变形处理、可见性估计和判别分类等部分。在自建的IHDD上进行训练和测试，结果表明，相比于传统的人员检测算法和利用经典网络模型检测的方法，所提算法通过使用区域建议网络模型显著提高了目标召回率，而多视角模型和DPM的设计则使得算法在处理多视角、存在一定变形与遮挡的情形时仍有较高的检测率和较低的误检率，可以为后续相关研究提供参考。

由于实际监控条件和室内环境的不确定性，所提算法在一些复杂环境下的人员检测性能还有提升空间，综合考虑经济条件和应用需求，以后可在更多相关场所中架设摄像头，增加样本的数量并提高样本质量，不断提升所提算法在多视角、存在变形与遮挡等复杂环境下的检测性能。

参　考　文　献

[1] Zou J H, Zhao Q C, Yang W, *et al*. Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation [J]. Energy and Buildings, 2017, 152: 385-398.

[2] Benenson R, Omran M, Hosang J, *et al*. Ten years of pedestrian detection, what have we learned? [C]. European Conference on Computer Vision, 2015: 613-627.

[3] Dalal N, Triggs B. Histograms of oriented gradients for human detection [C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 886-893.

[4] Wang X Y, Han T X, Yan S C. An HOG-LBP human detector with partial occlusion handling [C]. IEEE 12th International Conference on Computer Vision, 2009: 32-39.

[5] Zhang J G, Huang K Q, Yu Y N, *et al*. Boosted local structured HOG-LBP for object localization [C]. IEEE International Conference on Computer Vision and Pattern Recognition, 2011: 1393-1400.

[6] Lowe D G. Distinctiveimage features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[7] Lienhart R, Maydt J. An extended set of Haar-like features for rapid object detection [C]. International Conference on Image Processing, 2002: 900-903

[8] Dollar P, Tu Z W, Perona P, *et al*. Integral channel features [J]. Proceedings of the British Machine Vision Conference, 2009: 91.

[9] Dollar P, Appel R, Belongie S, *et al*. Fast feature pyramids for object detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532-1545.

[10] Zhang S S, Bauckhage C, Cremers A B. Informed haar-like features improve pedestrian detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 947-954.

[11] Nam W, Dollár P, Han J H. Local decorrelation for improved detection [J]. Advances in Neural Information Processing Systems, 2014, 1: 424-432.

[12] Zhang S S, Benenson R, Schiele B. Filtered channel features for pedestrian detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1751-1760.

[13] Sermanet P, Kavukcuoglu K, Chintala S, *et al*. Pedestrian detection with unsupervised multi-stage feature learning [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3626-3633.

[14] Hosang J, Omran M, Benenson R, *et al*. Taking a deeper look at pedestrians [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4073-4082.

[15] Cai Z W, Saberian M, Vasconcelos N. Learning complexity-aware cascades for deep pedestrian

detection[C]. IEEE International Conference on Computer Vision, 2015: 3361-3369.

[16] Maji S, Berg A C, Malik J. Classification using intersection kernel support vector machines is efficient[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2008: 4587630.

[17] Felzenszwalb P F, Girshick R B, McAllester D, *et al*. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[18] Ouyang W L, Wang X G. Joint deep learning for pedestrian detection [C]. IEEE International Conference on Computer Vision, 2013: 2056-2063.

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

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

[21] 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.

[22] Feng X Y, Mei W, Hu D S. Aerialtarget detection based on improved faster R-CNN [J]. Acta Optica Sinica, 2018, 38(6): 0615004.
冯小雨, 梅卫, 胡大帅. 基于改进 Faster R-CNN 的空中目标检测[J]. 光学学报, 2018, 38(6): 0615004.

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

[24] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot multibox detector[C]. European Conference on Computer Vision, 2016: 21-37.

[25] Liu H, Peng L, Wen J W. Multi-scale aware pedestrian detection algorithm based on improved full convolutional network[J]. Laser & Optoelectronics Progress, 2018, 55(9): 091504.
刘辉, 彭力, 闻继伟. 基于改进全卷积网络的多尺度感知行人检测算法[J]. 激光与光电子学进展, 2018, 55(9): 091504.

[26] Enzweiler M, Eigenstetter A, Schiele B, *et al*. Multi-cue pedestrian classification with partial occlusion handling [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2010: 990-997.

[27] Ouyang W L, Wang X G. A discriminative deep model for pedestrian detection with occlusion handling[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3258-3265.

[28] Uijlings J R R, van de Sande K E A, Gevers T, *et al*. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2): 154-171.

[29] Lawrence Zitnick C, Dollár P. Edge boxes: locating object proposals from edges [C]. European Conference on Computer Vision, 2014: 391-405.

[30] Ouyang W L, Wang X G. Single-pedestrian detection aided by multi-pedestrian detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3198-3205.

[31] Luo P, Tian Y L, Wang X G, *et al*. Switchable deep network for pedestrian detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2014: 899-906.

[32] Zhang S S, Benenson R, Omran M, *et al*. Towards reaching human performance in pedestrian detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 973-986.

[33] Hosang J, Benenson R, Dollar P, *et al*. What makes for effective detection proposals? [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(4): 814-830.

[34] Dollar P, Wojek C, Schiele B, *et al*. Pedestrian detection: A benchmark[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2009: 304-311.

[35] Everingham M, van 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.