结合语义和多层特征融合的行人检测

储 珺 [1] 束 雯 [1] 周子博 [1] 缪 君 [1] 冷 璐 [1]

摘 要 遮挡及背景中相似物干扰是行人检测准确率较低的主要原因. 针对该问题, 提出一种结合语义和多层特征融合 (Combining semantics with multi-level feature fusion, CSMFF) 的行人检测算法. 首先, 融合多个卷积层特征, 并在融合层上添加语义分割, 得到的语义特征与相应的卷积层连接作为行人位置的先验信息, 增强行人和背景的辨别性. 然后, 在初步回归的基础上构建行人二次检测模块 (Pedestrian secondary detection module, PSDM), 进一步排除误检物体. 实验结果表明, 所提算法在数据集 Caltech 和 CityPersons 上漏检率 (Miss rate, MR) 为 7.06 % 和 11.2 %. 该算法对被遮挡的行人具有强鲁棒性, 同时可方便地嵌入到其他检测框架.

关键词 行人检测, 语义分割, 特征融合, 遮挡, 二次检测

引用格式 储珺, 束雯, 周子博, 缪君, 冷璐. 结合语义和多层特征融合的行人检测. 自动化学报, 2022, **48**(1): 282−291

DOI 10.16383/j.aas.c200032

Combining Semantics With Multi-level Feature Fusion for Pedestrian Detection

CHU Jun[1] SHU Wen[1] ZHOU Zi-Bo[1] MIAO Jun[1] LENG Lu[1]

Abstract Occlusion and similar objects in the background typically degrade the accuracy of pedestrian detection. To solve the above problems, this paper proposes a pedestrian detection algorithm that combines semantics with multi-level feature fusion (CSMFF). Firstly, multi-convolutional-layer features are fused, and semantic segmentation is added to the fusion layer. The obtained semantic features are connected to the corresponding convolutional layers as the prior information of the pedestrian target location, which enhances the discrimination between pedestrian and background. Based on the preliminary regression, a pedestrian secondary detection module (PSDM) is constructed to further eliminate false positives. The experimental results show that the miss rates (MR) of the proposed algorithm on the datasets Caltech and CityPersons are 7.06 % and 11.2 %, respectively. The algorithm has strong robustness to occluded pedestrians, and can be easily embedded into other detection frameworks.

Key words Pedestrian detection, semantic segmentation, feature fusion, occlusion, secondary detection

Citation Chu Jun, Shu Wen, Zhou Zi-Bo, Miao Jun, Leng Lu. Combining semantics with multi-level feature fusion for pedestrian detection. *Acta Automatica Sinica*, 2022, **48**(1): 282−291

行人检测是目标检测领域研究最广泛的任务之一, 也一直是计算机视觉任务中的热点和难点. 行人检测任务是给出图像或视频中所有行人的位置和大小, 一般用矩形框标注. 行人检测技术可以与目标跟踪[1]、行人重识别[2] 等技术结合, 应用于汽车无人驾驶系统[3]、智能视频监控[4]、人体行为分析[5] 等领域. 在实际场景中, 由于行人与物体、行人间互相遮挡以及交通标志、橱窗中的模特等相似信息的干扰, 行人检测任务仍然存在很大的挑战[6].

行人检测是目标检测中的一种特例, 现阶段的很多行人检测算法都以目标检测框架为基础. 快速区域卷积神经网络[7] (Fast region convolutional neural network, Fast R-CNN) 和更快速区域卷积神经网络[8] (Faster region convolutional neural network, Faster R-CNN) 是目标检测[9−11] 和行人检测[12−14] 中被广泛采用的基础框架, 目前在 Caltech[15] 行人检测数据集上效果较好的算法大多是基于这两个框架. 如多尺度卷积神经网络[10] (Multi-scale convolutional neural network, MS-CNN) 和尺度感知的快速卷积神经网络[12] (Scale-aware fast region convolutional neural network, SA-FastRCNN) 分别基于 Faster R-CNN 和 Fast R-CNN 框架强调了尺度问题, 针对不同尺寸的行人特征设计了不同尺度的子网络.

收稿日期 2020-01-16 录用日期 2020-06-01
Manuscript received January 16, 2020; accepted June 1, 2020
国家自然科学基金 (62162045, 61866028), 江西省重点研发计划项目 (20192BBE50073), 研究生创新基金 (YC2018094) 资助
Supported by National Natural Science Foundation of China (62162045, 61866028), Jiangxi Key Research and Development Project (20192BBE50073), Innovation Foundation for Postgraduate (YC2018094)
本文责任编委 杨健
Recommended by Associate Editor YANG Jian
1. 江西省图像处理与模式识别重点实验室 (南昌航空大学) 南昌 330063
1. Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition (Nanchang Hangkong University), Nanchang 330063

Zhang 等[13] 证明了 Faster R-CNN 的候选区域网络 (Region proposal network, RPN) 对提取行人候选区域的有效性. 但同时也指出基于区域的卷积神经网络 (Region-based convolutional neural network, R-CNN) 在分类阶段, 由于高层卷积特征图分辨率降低, 小尺寸的行人无法得到有效的描述, 会降低检测的总体性能. 因此提出一种结合候选区域网络与决策森林 (Region proposal network + boosted forests, RPN + BF) 的算法. 该算法用 RPN 提取候选区域, 然后用决策森林对候选区域进行分类, 有效缓解了上述问题. 同样, 针对 Faster R-CNN 中小尺寸行人检测效果不佳的问题, Zhang 等[14] 提出自适应更快速区域卷积神经网络 (AdaptFasterRCNN), 通过量化 RPN 尺度、增大上采样因子、微调特征步幅、处理被忽略区域和调整损失函数的方式, 进一步提升了检测效果. Yun 等[16] 提出一种基于显著性和边界框对齐的部分卷积神经网络 (Part-level convolutional neural network, PL-CNN), 其用 RPN 提取候选区域, 对特征图中前景和背景设置不同的权重来消除背景干扰引起的误检, 有效解决了行人检测中遮挡和复杂背景干扰等问题.

目标检测算法的设计是为了更好地定位不同的对象, 检测过程中只用矩形框标注目标的位置, 通常不提供目标的边界信息. 语义分割能逐像素地定位目标的边界, 将检测和分割联合, 使用基于区域的分割方法提取特征, 自上而下地聚类计算候选区域, 能有效改进目标检测的性能[17]. Hariharan 等[18] 首次提出将分割与检测同时用于行人检测, 与文献 [17] 一样采用自上而下的分割方法, 不同的是使用多尺度组合分组[19] (Multi-scale combinatorial grouping, MCG) 作为分割的候选区域. Wang 等[20] 提出一种基于卷积神经网络的结合部件与上下文信息 (Part and context information with convolutional neural network, PCN) 的算法, 部件分支利用行人的语义信息来精准分类, 对被严重遮挡的行人具有良好的检测效果. Du 等[21] 提出深层神经网络融合 (Fused deep neural network, F-DNN) 的架构, 主要由行人候选区域生成器、分类网络和像素级别语义分割网络组成. 该算法在语义分割网络中使用掩膜增强行人特征, 降低行人检测的漏检率 (Miss rate, MR), 缺点是架构结构复杂, 提高了精度, 但牺牲了速度.

上述行人检测方法虽然添加了语义分割以解决遮挡及背景干扰等问题, 但把语义分割作为一个独立的任务来设计额外的分割网络, 计算复杂. 并且在检测过程中没有针对漏检和误检问题设计独立模块. 因此, 本文提出一种新的利用语义分割来增强检测效果的行人检测框架, 将语义分割掩膜融合到共享层, 增强行人特征, 解决行人的漏检和误检问题. 由于不增加单独的语义分割网络, 因此基本不增加模型的计算复杂度. 在 RPN 的回归分支中用 VGG-16[22] 构建一个轻量的二次检测模块, 解决前一模块初步检测的误检问题, 并且对前一次检测的结果进行二次回归.

本文的主要创新点包括:

1) 提出一种新的结合语义和多层特征融合 (Combining semantics with multi-level feature fusion, CSMFF) 的行人检测算法. 增加了行人特征增强模块 (Pedestrian feature enhancement module, PFEM) 和行人二次检测模块 (Pedestrian secondary detection module, PSDM), 将语义分割掩膜融合到共享层, 有效抑制背景信息的干扰和解决不同程度的遮挡问题, 并在此基础上通过二次检测和回归减少误检, 提高定位精度.

2) 在多层特征融合的基础上结合语义分割, 将骨干网络的浅层特征像素信息与深层特征语义信息进行融合, 有效提高了小尺寸行人的检测性能.

3) 行人特征增强模块可以很方便地嵌入到已有检测框架, 基本不增加运算复杂度.

1 本文算法

提出的 CSMFF 行人检测算法除骨干网络外由两个关键部分组成: 行人特征增强模块和行人二次检测模块.

行人特征增强模块在 Faster R-CNN 的 RPN 之前添加语义分割分支, 得到以目标框为边界的分割掩膜. 即对骨干网络采用多层特征融合, 在此基础上用 1×1 卷积实现分割. 分割时逐像素遍历图像中每个像素点, 并对每个像素点单独预测和分类, 形成语义分割掩膜. 分割掩膜通过编码得到语义信息, 映射到骨干网络的深层特征作为 RPN 的输入.

行人二次检测模块添加在 RPN 的回归分支上, 同样对多层特征融合后添加语义分割分支, 用来解决 PFEM 初步检测的误检问题, 并对初次检测结果进行二次回归. CSMFF 框架的流程如图 1 所示.

1.1 行人特征增强模块

浅层卷积产生的特征图包含更多像素信息, 有较高的空间分辨率, 行人的轮廓更加清晰, 用来定位行人会更准确. 深层卷积产生的特征图则包含更多的语义信息, 用于行人的检测会更精确. 所以文

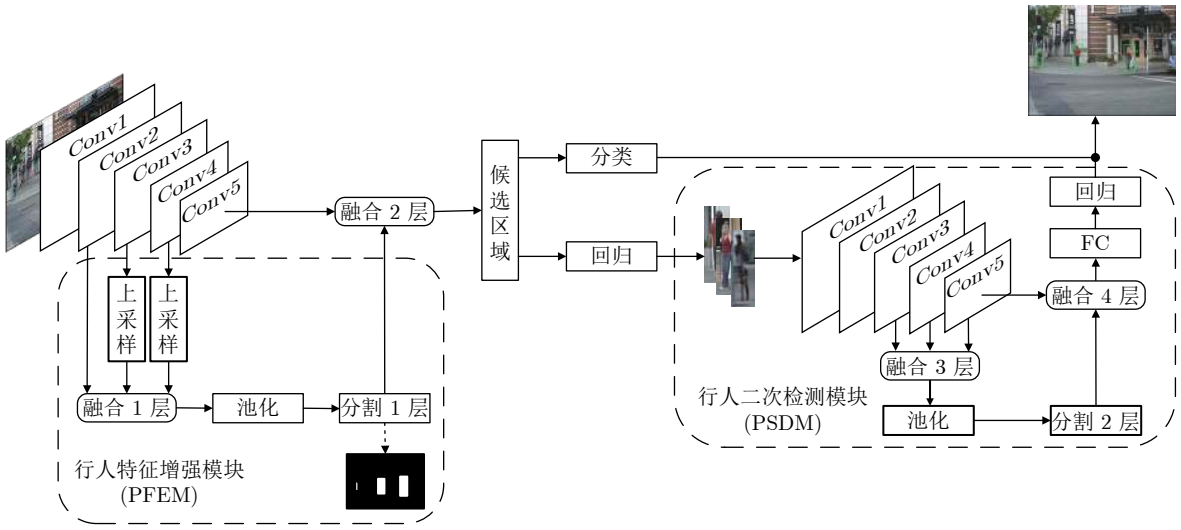


图 1 本文算法框架

Fig. 1 Overview of our proposed framework

中在分割时把多个卷积特征的融合特征作为分割的输入特征.

行人特征增强模块采用的骨干网络是 VGG-16, 用卷积的前 5 层来提取特征. 不同卷积层生成的特征图表示不同尺度的行人, 卷积层越深, 特征图的尺寸就越小, 因此, 需要对不同的卷积层采用不同的采样策略. 具体做法为: 保持 Conv2_2 层的特征图尺寸不变 (112 × 112 像素), 在 Conv3_3 层和 Conv4_3 层上分别添加一个 2 × 2 和 4 × 4 的反卷积对特征图进行上采样, 记为 Dconv3_3 和 Dconv4_3. 然后将 Dconv3_3、Dconv4_3 与 Conv2 输出的特征图进行级联, 生成多层特征融合层, 记为融合 1 层. 为获得较好的语义特征映射, 在融合 1 层上添加由 1 × 1 的卷积构成的语义分割分支, 用于预测输入图像在采样分辨率上每个像素的类别, 记为分割 1 层. 语义分割层形成的行人掩膜有效抑制了背景信息的干扰, 并且网络加深时, 语义信息会随之进入到卷积层. 利用分割的掩膜获取语义特征映射后, 将其与相应的卷积特征图连接作为行人分类的最终特征. 具体为分割 1 层与 Conv5_3 层特征映射相加连接, 记为融合 2 层, 最终输入 RPN 网络.

现阶段的行人检测数据集大都缺乏基于物体轮廓为边界的逐像素语义标注, 无法正常对其进行训练. 而且随着卷积网络的加深, 图像的尺寸在经过多个池化层后越变越小, 对于被严重遮挡的行人和小目标来说, 使用物体轮廓和目标框作为边界的逐像素标注的差异已经微乎其微. 图 2 比较了在 Conv5_3 层后分别采用基于物体轮廓和目标框为边界的逐像素分割结果. 从图 2 (b) 和图 2 (c) 来看, 两者相差不大. 并且我们的分割只是用来辅助检测, 无需分割出行人的精确形状, 所以文中选择基于目标框为边界的逐像素分割方式. 训练时利用训练数据集中行人的标注信息 (坐标、宽、高) 形成基于目标框式的分割区域, 作为行人分割的标注.

随着卷积网络的加深, 网络训练越来越困难, 收敛也越来越慢. 前期有很多方法可以解决该问题, 如修正线性单元激活函数[23] (Rectified linear unit, ReLU)、残差网络[24] (Residual network, ResNet) 以及梯度下降法[25] (Gradient descent, GD). 尽管这些方法训练神经网络非常简单高效, 但是需要人为地选择参数, 如学习率、参数初始化、权重衰减系数等, 而且这些参数的选择对于训练结果至关重要, 需要花费很多时间在参数调整上. 本文使用 BatchNorm 层[26] 来解决该问题, 具体是在 Conv2 ~ Conv5 每一个卷积层中添加 BatchNorm 层, 采用的 BatchNorm 层位于卷积层和激活层中间.

图 3 是行人特征增强模块在添加语义分割前后 Conv5_3 层的特征可视化对比, 图中差异明显的地方用矩形方框做了相应的标记. 第 1 行是 Caltech 数据集部分测试图像结果, 第 2 行是骨干网络 Conv5_3 层的特征图, 第 3 行是在 Conv5_3 层上添加本文语义分割分支的特征图, 第 4 行是融合多层特征后添加语义分割分支的特征图. 通过对比可以看出, 受背景和行人较为相似、行人被遮挡等因素影响, 行人在骨干网络 Conv5_3 层的特征不明显. 添加了本文的语义分割分支后, 行人特征增强, 但当目标比较小时, 增强效果不太明显 (第 3 行方框). 在提出的融合语义和多层卷积特征后, 行人特征增强更加明显 (第 4 行方框). 验证了 CSMFF 可

(a) CityPersons 数据集的原图
(a) An original image in the CityPersons dataset

(b) 基于物体轮廓为边界的逐像素分割结果
(b) Pixel-by-pixel segmentation result based on object contour boundary

(c) 基于目标框为边界的逐像素分割结果
(c) Pixel-by-pixel segmentation result based on object box boundary

图 2 基于目标框和物体轮廓为边界的逐像素分割结果

Fig. 2 The pixel-by-pixel segmentation results based on object box boundary and object contour boundary

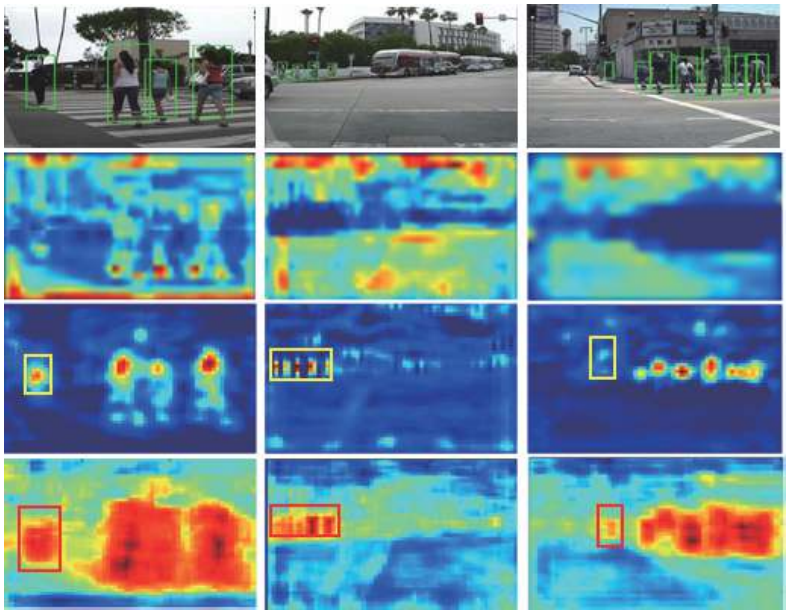


图 3 添加语义分割前后 Conv5_3 层的特征可视化对比

Fig. 3 Visual comparison of features of Conv5_3 layer before and after adding semantic segmentation

以更好地区分行人与背景区域.

1.2 PFEM 损失函数

PFEM 模块训练时的损失函数包含三个部分: 分类损失、回归损失和分割损失. 分类损失和回归损失与一般的目标检测器一致. 分割在本文是一个二分类的辅助检测工作, 与一般的分割方法不同,

此处分割损失也采用与分类损失一样的损失函数. PFEM 的总损失函数如下:

$$L_{\text{PFEM}}(\{p_i\},\{\boldsymbol{t}_i\},\{s_i\}) = \sum_i L_{\text{cls}}(p_i,\, p_i^*) + \alpha\sum_i L_{\text{reg}}(\boldsymbol{t}_i,\, \boldsymbol{t}_i^*) + \sum_i L_{\text{seg}}(s_i,\, s_i^*) \quad (1)$$

式中, α 是 PFEM 模块中回归的损失权重, 在实验中, $\alpha = 5$.

1) L_{cls} 为分类损失函数:

$$L_{\text{cls}}(p_i,\, p_i^*) = -\ln[p_i p_i^* + (1-p_i)(1-p_i^*)] \quad (2)$$

式中, p_i 表示分类时第 i 个锚 (anchor) 框为行人的概率, p_i^* 表示第 i 个标记框为行人的概率. 当第 i 个 anchor 框与标记框的交并比大于等于 0.5 时, 说明预测的是正样本, 即是行人, $p_i^* = 1$, 否则 $p_i^* = 0$.

2) L_{reg} 为回归损失函数:

$$L_{\text{reg}}(\boldsymbol{t}_i,\, \boldsymbol{t}_i^*) = \sum_{j\in\{x,\,y,\,w,\,h\}} \text{Smooth}_{L1}(t_{i,\,j} - t_{i,\,j}^*) \quad (3)$$

式中, $\boldsymbol{t}_i = [t_{i,\,x},\, t_{i,\,y},\, t_{i,\,w},\, t_{i,\,h}]$ 是一个向量, 表示第 i 个预测目标框的坐标、宽、高的偏移量, $t_{i,\,j}$ 是向量 $\boldsymbol{t}_i$ 的第 j 个元素; $\boldsymbol{t}_i^*$ 表示第 i 个真实目标框与对应 anchor 的坐标、宽、高的偏移量, $t_{i,\,j}^*$ 是向量 $\boldsymbol{t}_i^*$ 的第 j 个元素. 其中, Smooth_{L1} 函数定义如下:

$$\text{Smooth}_{L1}(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{其他} \end{cases} \quad (4)$$

3) L_{seg} 为分割损失函数:

$$L_{\text{seg}}(s_i,\, s_i^*) = -\ln[s_i s_i^* + (1-s_i)(1-s_i^*)] \quad (5)$$

式中, s_i 表示分割时第 i 个掩膜为行人的概率, s_i^* 为掩膜标记框的标签. 目标区域标记为 $s_i^* = 1$; 背景区域标记为 $s_i^* = 0$.

1.3 行人二次检测模块

PFEM 模块增加了分割分支, 将分割得到的语义信息和骨干网络 Conv5_3 层特征融合, 通过 RPN 网络提取候选区域, 再经过分类和回归得到初步的检测结果. 其中, 分割语义信息增强行人特征, 抑制背景信息, 可以减少相似背景干扰, 同时提高小目标的检测率. 但当图像中存在与行人特征相似的目标, 如停车牌、树木等, 也会出现误检. 因此在后端提出 PSDM 以进一步提高被遮挡、小尺寸等行人的分数, 提高整体的检测性能.

在二阶段目标检测框架中, 大多数后端的分类和回归采用 Faster R-CNN 后端 R-CNN 的分类部分, 但是文献 [13] 指出 Faster R-CNN 的后端会降低行人检测的精度. 通过 Caltech 数据集进行验证, 前端使用提出的 PFEM, 后端用 R-CNN 与提出的 PSDM 模块对比分类和回归的效果, 实验结果也说明后端采用 R-CNN 的结果不如 PSDM 模块. 其主要原因有两个: 1) 行人在数据集中的尺寸较小, 在行人检测 Caltech 数据集中, 大约有 88 % 的行人低于 112 × 112 像素. 对于小尺寸的行人来说, 若后端感兴趣区域池化层的输入分辨率小于其输出分辨率, 会降低提取特征的辨别力. 2) 行人检测和目标检测两者误检的针对性不同. 行人检测误检是指将背景中的相似物预测为行人, 而传统目标检测中存在多个类别, 其误检是指将一个正确的目标错误地预测为另一个目标. 而且基于目标检测中的 R-CNN 缺乏挖掘难负样本的机制, 将这样的 R-CNN 直接用于行人检测时对于被遮挡严重或者尺寸较小的行人检测效果不佳.

行人检测是一个二分类问题, 与文献 [13] 不同, 本文使用 VGG-16 中的 Conv1 ~ Conv5 层作为骨干网络构建一个单独的识别网络, 减少了计算量. 为解决感兴趣区域池化层的输入分辨率小于输出分辨率问题, 去除 Conv5_3 层后的池化层, 将输入 PFEM 之前的图像尺寸调整为 112 × 112 像素. PSDM 中仍然增加了多层特征融合层和语义信息来提高识别率. Conv4 层和 Conv5 层的深层语义特征有助于分类, Conv3 层是中间层, 存在边缘信息, 可以更好地回归, 使行人定位更精准. 因此, 行人二次检测模块分别在 Conv4_3 层、Conv5_3 层上添加一个步长为 2 的 2 × 2 卷积核和步长为 4 的 4 × 4 卷积核进行反卷积上采样, 然后与 Conv3 层输出的特征级联, 生成多层特征融合层, 记为融合 3 层. 在此基础上添加语义分割分支, 记为分割 2 层. 获得的语义特征映射后与 Conv5_3 层特征映射相加融合, 记为融合 4 层, 作为行人分类的特征. PSDM 的损失函数可表示为:

$$L_{\text{PSDM}}(\{c_i\},\{\boldsymbol{b}_i\},\{m_i\}) = \sum_i L_{\text{cls}}(c_i,\, c_i^*) + \beta\sum_i L_{\text{reg}}(\boldsymbol{b}_i,\, \boldsymbol{b}_i^*) + \sum_i L_{\text{seg}}(m_i,\, m_i^*) \quad (6)$$

其中, 分类损失、回归损失与 PFEM 一致. β 是 PSDM 模块中回归的损失权重, 在实验中, $\beta = 5$.

2 实验与结果分析

本文采用基于 ImageNet[27] 上预训练的 VGG-16 网络作为骨干网络. 实验均是在 MATLAB

2016b 环境下进行, 操作系统为 64 位的 Ubuntu 16.04, 深度学习框架为 Caffe[28]; 硬件配置为 CPU Intel Xeon(R) E5-2678 v3 @ 2.50 GHz 十二核; 内存 32 GB; GPU NVIDIA GeForce GTX 1080Ti.

2.1　实验数据

在 Caltech 和 CityPersons 数据集上进行实验, 它们是近几年使用规模最广的行人检测数据集. 文献 [14] 比较了 Caltech 和 CityPersons 不同遮挡水平下的行人分布. 从文献 [14] 可知, CityPersons 数据集行人被遮挡程度更大, Caltech 完全可见的行人超过 60 %, 而 CityPersons 则不到 30 %.

Caltech 数据集是目前规模较大的行人数据库, 采用车载摄像头拍摄 10 个小时左右, 背景主要是公路或街道, 视频的分辨率为 640 × 480 像素. 其中标注了 350 000 个矩形框, 2 300 个行人, 超过 70 % 的行人至少在一帧内出现了遮挡. 该数据集分为 11 个视频组 set00 ~ set10, 其中 set00 ~ set05 为训练集, 根据 Caltech10 × [29] 的标准对训练集中 42 782 张图像训练, 对剩余的 set06 ~ set10 中 4 024 张图像进行测试.

CityPersons 数据集是基于语义分割 Cityscapes 数据集[30] 的一个行人检测数据集, 其数据是从德国的 18 个城市, 在三个不同的季节和不同的天气条件下收集的. 该数据集总共包括 5 000 张图像 (2 975 张用于训练, 500 张用于验证, 1 525 张用于测试), 总共约有 35 000 人, 另外还有约 13 000 个未标注的区域, 图像分辨率为 2 048 × 1 024 像素. 本文对该数据集的训练和测试都是在其训练和验证集上进行.

2.2　评估标准

为验证实验的全面性, 根据官方提供的数据集评估标准, Caltech 和 CityPersons 的数据依据行人高度和被遮挡比例被划分成很多子集. 因为本文实验主要验证对被遮挡行人和小目标的检测性能, 所以只比较 Caltech 数据集中的 Reasonable、Partial、Heavy 子集和 CityPersons 数据集中的 Bare、Reasonable、Partial、Heavy 子集. 表 1、表 2 分别给出了 Caltech、CityPersons 数据集中不同遮挡情况下每个子集的划分标准.

本文实验中, 采用行人检测和目标检测领域常用的漏检率 (Miss rate, MR)、漏检率–每帧图像误检率曲线 (Miss rate-false positives per image, MR-FPPI) 及对数平均漏检率[31] (Log-average miss rate, LAMR) 作为评价指标. 其中, 漏检率是指正样本被模型预测为负样本的数目与所有正样本数目的比例; 每帧图像误检率 (False positives per image, FPPI) 是指负样本被模型预测为正样本的数目与所有样本的比例; 对数平均漏检率是 MR-FPPI 曲线在对数空间 $[10^{-2} \sim 10^{0}]$ 内均匀分布的九个点的平均值.

表 1　Caltech 数据集中部分子集的划分标准
Table 1　Evaluation settings for partial subsets of the Caltech dataset

子集	行人高度 (Height)	行人被遮挡程度 (Occlusion)
Reasonable	> 50 PXs	occ < 0.35
Partial	> 50 PXs	0.10 < occ ≤ 0.35
Heavy	> 50 PXs	0.35 < occ ≤ 0.80

表 2　CityPersons 数据集中部分子集的划分标准
Table 2　Evaluation settings for partial subsets of the CityPersons dataset

子集	行人高度 (Height)	行人被遮挡程度 (Occlusion)
Bare	> 50 PXs	occ ≤ 0.10
Reasonable	> 50 PXs	occ < 0.35
Partial	> 50 PXs	0.10 < occ ≤ 0.35
Heavy	> 50 PXs	0.35 < occ ≤ 0.80

2.3　实验结果分析

1) Caltech 数据集

为验证本算法性能, 选取了 8 种在 Caltech 数据集性能较好的、能解决不同程度遮挡行人的检测算法与 CSMFF 结果比较. 其中 AdaptFasterRCNN[14]、PCN[20]、PL-CNN[16]、MS-CNN[10]、F-DNN + SS[21] 是基于目标检测框架结合语义的算法; RPN + BF[13] 采用决策森林代替 Faster R-CNN 中的 R-CNN, 对候选区域进行分类, 有助于提升小尺寸行人的检测效果; Faster R-CNN + ATT[32] 增加了注意力机制, 在被严重遮挡行人检测上取得了最佳性能. 总体性能和运行速度如表 3 所示, 因为 PL-CNN、Faster R-CNN + ATT、AdaptFasterRCNN、PCN 原论文中没有比较检测速度, 所以表 3 中没有给出它们的检测速度.

从表 3 可以看出, CSMFF 算法在 Reasonable 和 Partial 子集上都达到了最低的漏检率, 分别比效果第二的 F-DNN + SS 算法降低了 1.12 % 和 0.75 %. Caltech 数据集的检测性能已接近饱和, 因此在 Reasonable 子集上性能的提升非常重要. 但在 Heavy 子集上效果低于 Faster R-CNN + ATT, 排在第二

表 3 在 Caltech 测试数据集上对比算法性能以及运行速度比较

Table 3 Performance and runtime comparisons of our proposed CSMFF with state-of-the-art approaches on the Caltech test dataset

方法	Reasonable MR (%)	Partial MR (%)	Heavy MR (%)	速度 (s/帧)
PL-CNN[16]	12.40	16.68	—	—
Faster R-CNN + ATT[32]	10.33	22.29	**45.18**	—
MS-CNN[10]	9.95	19.24	59.94	0.40
RPN + BF[13]	9.58	24.23	74.36	0.60
AdaptFasterRCNN[14]	9.18	26.55	57.58	—
F-DNN[21]	8.65	15.41	55.13	0.30
PCN[20]	8.45	16.09	55.81	—
F-DNN + SS[21]	8.18	15.11	53.76	2.48
CSMFF	**7.06**	**14.36**	50.62	**0.12**

位. 主要原因是卷积通道特征分别对应行人身体的不同部位, 其对行人定位非常有效. Faster R-CNN + ATT[32] 在 Faster R-CNN 中添加了一个额外的注意力机制网络, 以通道方式的注意力机制有效地利用行人身体部位与不同卷积通道的关系来处理严重遮挡模式下的行人. 虽然文献 [32] 未给出 Faster R-CNN + ATT 算法的运行速度, 但是其注意力机制的复杂度可以从其论文的描述中体现出来. 这些结果表明, 行人出现不同程度的遮挡会减少行人的有效特征, 本文设计的 PFEM 和 PSDM 可以在一定程度上增强行人的特征辨别性, 因此 CSMFF 在不同程度的遮挡情况下均具有良好的泛化能力.

表 3 也给出了各算法运行速度的比较, 本文训练和测试仅在一张 1080Ti GPU 上进行. 从表 3 可以看出, 所提出的 CSMFF 算法的运行速度约为 0.12 s/帧, 在比较的算法中是最快的, F-DNN 排在第二位.

图 4 是 CSMFF 与各种对比算法在 Caltech 数据集 Reasonable、Partial、Heavy 子集上 MR-FPPI 变化. 横坐标表示每帧图像误检率, 纵坐标表示漏检率, MR-FPPI 曲线越低, 表示行人检测算法在测试集上测试效果越好. 从图中可以观察到, CSMFF 算法的曲线在 Reasonable 和 Partial 子集上最低, 且下降很快, 取得了最佳的检测性能. 主要原因是: 1) 虽然一些算法结合了从 CityPersons 数据集训练的高精度像素级语义信息, 但是语义分割模型是独立于候选区域生成器进行训练的, 语义特征无法进入候选区域网络; 2) 本文针对前端模块产生的误检问题进行了二次检测, 提高了整体性能.

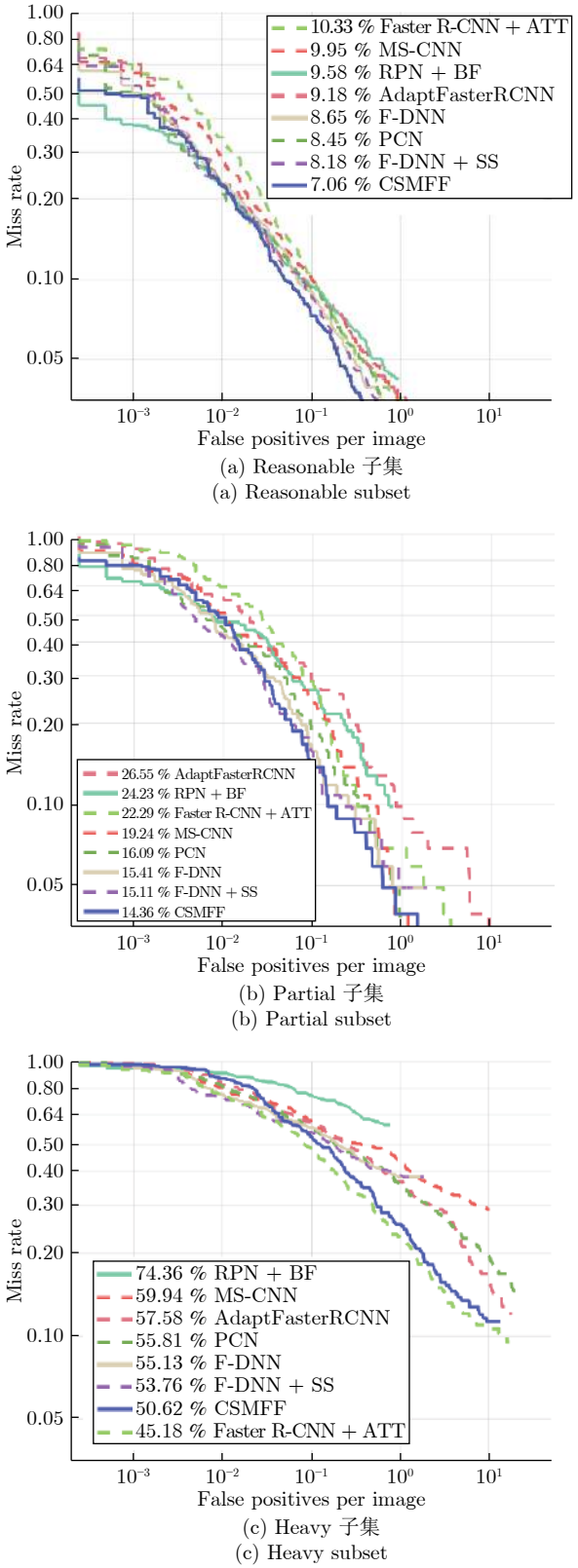


图 4 CSMFF 与各种对比算法在 Caltech 测试数据集上 MR-FPPI 变化

Fig. 4 The variations of MR-FPPI of our proposed CSMFF with state-of-the-art approaches on the Caltech test dataset

2) CityPersons 数据集

为验证算法的鲁棒性, 在 CityPersons 数据集的部分子集上也做了实验. 表 4 比较了 CityPersons 数据集上效果较好的几种行人检测方法与 CSMFF 的检测性能. 从表 4 中可以看出, CSMFF 在大部分子集上都能实现最佳检测性能, 分别在 Reasonable、Partial、Heavy 子集上实现了 11.2 %、13.4 % 和 50.1 % 的漏检率, 但在 Bare 子集上弱于 OR-CNN. 因为 Bare 子集的遮挡率不到 10 %, 在被轻度遮挡的情况下, 人体的四肢躯干完全, 人体结构信息比较清楚, OR-CNN 将人体分成 5 个部分, 利用人体结构先验信息, 所以 OR-CNN 方法在这种情况下的漏检率较低.

2.4　消融实验

之前 Zhang 等[6] 已经揭示了多层特征融合对各种尺寸行人检测的重要性, 本文在 PFEM 的 VGG-16 网络上训练了几种模型, 以此来研究哪些卷积层融合会达到最佳效果. 由于浅层的判别信息有限, 所以选取 Conv2_2 的输出作为起点. 表 5 比较了 PFEM 融合不同卷积层和完整 CSMFF 算法的结果. 结果表明较浅的特征图对定位有帮助, Conv4 和 Conv5 等较深卷积层的特征图有丰富的语义特征, 有利于小目标的检测. 最终采用 Conv2_2、Conv3_3 和 Conv4_3 层卷积融合生成多层特征层.

为证明模型的有效性, 在 Caltech 数据集上进行消融实验. 表 6 比较了 PFEM 中每个组件以及添加 PSDM 后与完整算法的对比结果. 从表 6 可以看出, 在 VGG-16 上将各层特征融合以及在此基础上添加语义分割分支时, 漏检率都有所下降, 这表明提出的 PFEM 是有效的. 针对 Faster R-CNN 的后端会降低行人检测精度的问题, 本文设计了 PSDM. 从实验结果可以很明显地看到, 对前一模块由于背景干扰和遮挡产生的误检, 进行行人二次检测和回归后, 可以提高算法整体的检测性能.

3　结论

本文提出了一种基于 Faster R-CNN 的结合语义和多层特征融合的行人检测算法. 在多层卷积特征融合基础上添加语义分割分支, 并将其结果作为行人目标特征信息, 为行人检测和背景的区分提供了更多的辨别信息. 后端在初步检测的基础上增加行人二次检测模块, 并对初步检测结果进行二次回归, 解决了前一阶段产生的误检问题. 但由于行人被严重遮挡时的可见部分很少, 造成用于训练的有效特征少, 加上行人周围大量背景等无用信息的干扰, 导致检测性能下降. 我们下一步工作拟在本文

表 4　在 CityPersons 测试数据集上不同算法性能比较

Table 4　Performance comparison of our proposed CSMFF with state-of-the-art approaches on the CityPersons test dataset

方法	骨干网络	Bare MR (%)	Reasonable MR (%)	Partial MR (%)	Heavy MR (%)
TLL[33]	ResNet-50	10.0	15.5	17.2	53.6
Repulsion Loss[34]	ResNet-50	7.6	13.2	16.8	56.9
LBST[35]	ResNet-50	—	12.8	—	53.7
CC-CNN[36]	VGG-16	8.2	11.8	14.1	—
OR-CNN[37]	VGG-16	**6.7**	12.8	15.3	55.7
Faster R-CNN[14]	VGG-16	—	15.4	—	—
CSMFF	VGG-16	7.5	**11.2**	**13.4**	**50.1**

表 5　在 Caltech 测试数据集上融合不同卷积层的性能

Table 5　Performance of fusing different convolutional layers on the Caltech test dataset

卷积层				MR (%)	
Conv2_2	Conv3_3	Conv4_3	Conv5_3	PFEM	CSMFF
√	√	√		**12.22**	**7.06**
	√	√	√	32.42	18.15
√	√	√	√	18.72	11.79

表 6　在 Caltech 数据集上测试每个组件的消融实验

Table 6　Ablation experiments for testing each component on the Caltech dataset

组件	选择			
Faster R-CNN	√			
多层特征融合		√	√	√
语义分割分支			√	√
PSDM				√
PFEM MR (%)	14.93	13.27	12.58	**12.22**
CSMFF MR (%)	12.11	9.53	8.68	**7.06**

算法框架上提出一种新型压缩激励的注意力机制网络, 可以自动选择卷积层通道中行人的语义以及有用信息, 抑制无用信息, 降低被严重遮挡行人的漏检率.

References

1 Danelljan M, Bhat G, Khan F S, Felsberg M. Atom: Accurate tracking by overlap maximization. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, California, USA: IEEE, 2019. 4660−4669

2 Li You-Jiao, Zhuo Li, Zhang Jing, Li Jia-Feng, Zhang Hui. Overview of pedestrian re-identification technology. *Acta Automatica Sinica*, 2018, **44**(9): 1554−1568
(李幼蛟, 卓力, 张菁, 李嘉锋, 张辉. 行人再识别技术综述. 自动化学报, 2018, **44**(9): 1554−1568)

3 Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, Rhode Island, USA: IEEE, 2012. 3354−3361

4 Wang Meng-Lai, Li Xiang, Chen Qi, Li Lan-Bo, Zhao Yan-Yun. CNN-based surveillance video event detection. *Acta Automatica Sinica*, 2016, **42**(6): 892−903
(王梦来, 李想, 陈奇, 李澜博, 赵衍运. 基于 CNN 的监控视频事件检测. 自动化学报, 2016, **42**(6): 892−903)

5 Kanazawa A, Black M J, Jacobs D W, Malik J. End-to-end recovery of human shape and pose. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 7122−7131

6 Zhang S, Benenson R, Omran M, Hosang J, Schiele B. How far are we from solving pedestrian detection? In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 1259−1267

7 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1440−1448

8 Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 2015 Advances in Neural Information Processing Systems (NIPS). Montreal, Quebec, Canada: MIT Press, 2015. 91−99

9 Yang F, Choi W, Lin Y. Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 2129−2137

10 Cai Z, Fan Q, Feris R S, Vasconcelos N. A unified multi-scale deep convolutional neural network for fast object detection. In: Proceedings of the 2016 European Conference on Computer Vision. Scottsdale, AZ, USA: Springer, 2016. 354−370

11 Gidaris S, Komodakis N. Object detection via a multi-region and semantic segmentation-aware CNN model. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1134−1142

12 Li J, Liang X, Shen S M, Xu T F, Feng J S, Yan S C. Scale-aware Fast R-CNN for pedestrian detection. *IEEE Transactions on Multimedia*, 2017, **20**(4): 985−996

13 Zhang L L, Lin L, Liang X D, He K M. Is Faster R-CNN doing well for pedestrian detection? In: Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, Noord-Holland, The Netherlands: IEEE, 2016. 443−457

14 Zhang S, Benenson R, Schiele B. CityPersons: A diverse dataset for pedestrian detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 3213−3221

15 Dollár P, Wojek C, Schiele B, Perona P. Pedestrian detection: A benchmark. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA: IEEE, 2009. 304−311

16 Yun I, Jung C, Wang X R, Hero A O, Kim J K. Part-level convolutional neural networks for pedestrian detection using saliency and boundary box alignment. *IEEE Access*, 2019, **7**: 23027−23037

17 Fidler S, Mottaghi R, Yuille A, Urtasun R. Bottom-up segmentation for top-down detection. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, Oregon, USA: IEEE, 2013. 3294−3301

18 Hariharan B, Arbeláez P, Girshick R, Malik J. Simultaneous detection and segmentation. In: Proceedings of the 2014 European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 297−312

19 Arbeláez P, Pont-Tuset J, Barron J T, Marques F, Malik J. Multiscale combinatorial grouping. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio, USA: IEEE, 2014. 328−335

20 Wang S G, Cheng J, Liu H J, Tang M. PCN: Part and context information for pedestrian detection with CNNs. arXiv preprint arXiv: 1804.04483, 2018.

21 Du X, El-Khamy M, Lee J, Davis L. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In: Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, California, USA: IEEE, 2017. 953−961

22 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv: 1409.1556, 2014.

23 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 2011 International Conference on Artificial Intelligence and Statistics. Espoo, Finland, German: Springer, 2011. 315−323

24 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 770−778

25 Hochreiter S, Younger A S, Conwell P R. Learning to learn using gradient descent. In: Proceedings of the 2001 International Conference on Artificial Neural Networks. Vienna, Austria, German: Springer, 2001. 87−94

26 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv: 1502.03167, 2015.

27 Deng J, Dong W, Socher R, Li L J, Li K, Li F F. Imagenet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA: IEEE, 2009. 248−255

28 Jia Y Q, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R. Caffe: Convolutional architecture for fast feature embedding.

arXiv preprint arXiv: 1408.5093, 2014.

29 Zhang S, Benenson R, Schiele B. Filtered channel features for pedestrian detection. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, Massachusetts, USA: IEEE, 2015. 1751−1760

30 Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 3213−3223

31 Dollar P, Wojek C, Schiele B, Perona P. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **34**(4): 743−761

32 Zhang S, Yang J, Schiele B. Occluded pedestrian detection through guided attention in CNNs. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 6995−7003

33 Song T, Sun L Y, Xie D, Sun H M, Pu S L. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 536−551

34 Wang X, Xiao T, Jiang Y, Shao S, Sun J, Shen C H. Repulsion loss: Detecting pedestrians in a crowd. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 7774−7783

35 Cao J L, Pang Y W, Han J G, Gao B L, Li X L. Taking a look at small-scale pedestrians and occluded pedestrians. *IEEE Transactions on Image Processing*, 2019, **29**: 3143−3152

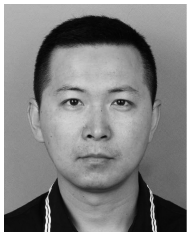
36 Zhao Y, Yuan Z J, Chen B D. Training cascade compact CNN with region-iou for accurate pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 2020, **21**(9): 3777−3787

37 Zhang S F, Wen L Y, Bian X, Lei Z, Li S Z. Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 637−653

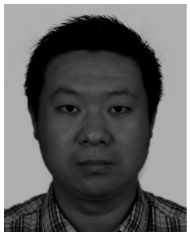
储 珺 江西省图像处理与模式识别重点实验室 (南昌航空大学) 教授. 主要研究方向为计算机视觉, 模式识别和深度学习. 本文通信作者.
E-mail: chujun99602@163.com
(**CHU Jun** Professor at Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition (Nanchang Hangkong University). Her research interest covers computer vision, pattern recognition, and deep learning. Corresponding author of this paper.)

束 雯 江西省图像处理与模式识别重点实验室 (南昌航空大学) 硕士研究生. 主要研究方向为图像处理, 计算机视觉.
E-mail: shuwen0418@163.com
(**SHU Wen** Master student at Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition (Nanchang Hangkong University). Her research interest covers image processing and computer vision.)

周子博 江西省图像处理与模式识别重点实验室 (南昌航空大学) 硕士研究生. 主要研究方向为图像处理, 计算机视觉.
E-mail: abaabc13@163.com
(**ZHOU Zi-Bo** Master student at Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition (Nanchang Hangkong University). His research interest covers image processing and computer vision.)



缪 君 江西省图像处理与模式识别重点实验室 (南昌航空大学) 副教授. 主要研究方向为计算机视觉, 3D 重建和模式识别.
E-mail: miaojun@nchu.edu.cn
(**MIAO Jun** Associate professor at Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition (Nanchang Hangkong University). His research interest covers computer vision, 3D reconstruction, and pattern recognition.)



冷 璐 江西省图像处理与模式识别重点实验室 (南昌航空大学) 副教授. 主要研究方向为计算机视觉, 模式识别和生物特征模板保护.
E-mail: leng@nchu.edu.cn
(**LENG Lu** Associate professor at Key Laboratory of Jiangxi Province for Image Processing and Pattern Recognition (Nanchang Hangkong University). His research interest covers computer vision, pattern recognition, and biometric template protection.)