高低密度多维视角多元信息融合人群计数方法

孟月波[1,2], 陈宣润[1], 刘光辉[1†], 徐胜军[1,2], 李彤月[3]

(1. 西安建筑科技大学 信息与控制工程学院，西安 710055；2. 人工智能与数字经济广东省实验室，广州 510000；3. 中国人民解放军军事科学院，北京 100091)

摘 要: 针对人群密度在二维图像中随图像视角变化呈现较大差异、特征空间多尺度信息丢失等问题,提出一种多维视角多元信息融合(MDPMIF)的人群密度估计方法. 首先,由“上-左-右-下”的方向对视角变化进行信息编码,通过递进聚合方式捕获深层次全局上下文信息,同步提取多维度视角的尺度关系特征;然后,设计联合学习策略获取全局尺度关系特征,并将全局上下文表达、全局尺度关系特征集成,得到更全面的视角变换描述;最后,采用语义嵌入方式实现高、低阶特征相互补充,增强输出密度图的质量. 同时,真实场景下的人群聚集模式存在差异,单纯密度图方法易对图像中的低聚集部分造成人群计数高估,基于此,提出一种高低密度多维视角多元信息融合人群计数网络. 设计高低密度区分策略对MDPMIF输出进行高低密度区域自适应划分,高密区域保持MDPMIF网络估计结果,低密区域采用检测方法实现人群计数修正,提高模型的鲁棒性. 实验结果表明,所提出方法的性能优于对比方法.

关键词: 人群计数；视角变化；高低密度区分；特征融合；上下文信息；尺度信息

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0520

引用格式: 孟月波,陈宣润,刘光辉,等. 高低密度多维视角多元信息融合人群计数方法[J]. 控制与决策, 2023, 38(1): 181-189.

High and low density multi-dimension perspective multivariate information fusion crowd counting method

MENG Yue-bo[1,2], *CHEN Xuan-run*[1], *LIU Guang-hui*[1†], *XU Sheng-jun*[1,2], *LI Tong-yue*[3]

(1. College of Information and Control Engineering，Xi'an University of Architecture and Technology，Xi'an 710055，China；2. Guangzhou Artificial Intelligence and Digital Economy Laboratory，Guangzhou 510000，China；3. PLA Academy of Military Sciences，Beijing 100091，China)

Abstract: A crowd density estimation method with multi-dimensional perspective multivariate information fusion（MDPMIF）is proposed for the problems that crowd density in two-dimensional images presents large differences with image viewpoint changes and multi-scale information loss in feature space. Firstly, the information of perspective change is encoded from ‘up-left-right-bottom’ direction, and the deep global contextual information is captured by progressive aggregation, and the scale relationship features of multi-dimensional perspective are extracted simultaneously. After that, a joint learning strategy is designed to obtain global scale relationship features and integrate global contextual expressions and global scale relationship features to obtain a more comprehensive description of perspective transformation. Finally, semantic embedding is used to realize the high and low order features to complement each other and enhance the quality of the output density map. Meanwhile, there are differences in crowd aggregation patterns in real scenes, and the simple density map method is prone to overestimate crowd counts for the low aggregation part of the image. Based on this, a high and low density multi-dimensional perspective multivariate information fusion crowd counting network（HLMMNet）is proposed on the basis of the MDPMIF network. A high and low density differentiation strategy is designed to adaptively divide the MDPMIF output into high and low density regions, keeping the MDPMIF network estimation results in the high density regions and using detection methods to achieve crowd counting correction in the low density regions, improving the robustness of the model. The experimental results show that the performance of this method is superior to other comparative methods.

Keywords: crowd counting；perspective changes；high and sparse density differentiation；feature fusion；context information；scale information

收稿日期: 2021-03-29；**录用日期:** 2021-09-28.

基金项目: 陕西省自然科学基础研究计划面上项目(2020JM-473, 2020JM-472)；陕西省重点研发计划项目(2021SF-429).

责任编委: 胡清华.

†通讯作者. E-mail: guanghuil@163.com.

0 引 言

随着国民经济迅猛发展及城市化进度不断加快[1],城市人口数量急剧增加,人们因各种原因会聚集在不同的场景下,易造成交通拥堵、人员踩踏等不安全事故. 因此,人群密度估计与计数在公共安全、城市规划等诸多领域具有较高的应用价值[2-3].

通过学习图像的局部特征与其相应密度图之间的映射回归人群人数是近年来主要采用的人群计数方法[4-8],其不仅能够回归人员数量,还能反映人群密度的分布信息,适用于更高层的认知任务. 得益于深度学习技术的发展,基于卷积神经网络(convolutional neural network, CNN)的人群密度估计与计数方法受到广泛关注并取得了显著成效[9-13].

日常生活中,大多数计数场景具有空间域、时间域人群分布变化大的特点[14].

1) 在空间域方面,拍摄场景、相机拍摄角度不同造成了图像视角变化复杂,图像会产生不同的透视现象,使得真实场景中的相同人口分布在图像的不同区域会呈现不同的分布尺度,导致图像与真实场景相比丢失较多的特征信息,人群计数精度下降. 为解决该问题,Zhang等[15]提出一种多列CNN并行结构,通过不同大小的卷积核处理输入图像中不同大小的头部,提取图像的多尺度信息;文献[8, 16-17]分别从提升多列结构特征提取能力、降低多列结构复杂度、改善多列特征融合质量等方面,对多列CNN并行结构人群计数方法加以改进,一定程度上缓解了图像的视角变换问题. 但是,上述多列结构的每一列具有相似的学习功能,视角变化的多样与不确定性使得多列卷积核大小难以匹配或适用于所有的视角情况,造成多列结构中的每列网络只能处理对应大小的头部,而在其他尺度大小头部上的性能急剧下降,导致图像上下文信息丢失较多,所生成的密度图质量不高. 补充网络上下文信息,降低特征缺失是降低视角变换影响的另一手段,文献[18-19]通过构建全局、局部空间金字塔结构获取图像全局、局部上下文信息;Sheng等[20]利用相邻斑块空间金字塔思想将上下文信息引入网络中. Li等[12]采用空洞卷积衔接方式扩大感受野,以增强全局上下文信息提取能力. 但对于人群视角变化复杂的场景,现有方法空间上下文信息提取能力不足,密度估计效果有待进一步提升. 本文认为,从多个视角层面出发提升多尺度信息、空间上下文信息挖掘深度并实现两类信息互补,是弱化视角影响、提高网络特征提取能力的有效手段.

2) 在时间域方面,人员的随机流动使得图像的同一区域会随着时间的不同表现出不同的人群分布状态,图像的不同区域在同一时刻的人群聚集度存在很大差异. 尽管采用密度图回归人员数量对人群稠密区域计数优势明显,但由于背景复杂、相似度高等因素的影响,其对人群稀疏区域易造成计数高估. “检测方式+密度图方法”是解决此问题的一个有效思路. 文献[14]通过高斯滤波器将人员检测结果转为密度图形式,利用注意力引导的置信度评价模块将其与密度图方法得到的密度图进行比较,选择置信度高的区域融合为最终的密度图. 但是,该方法的置信度评价模块将多层特征信息加和送入注意力机制,互相之间会因检测与密度的差异导致融合判别的精确度下降,且密度估计模块未考虑尺度信息,网络鲁棒性有待进一步提高.

基于上述分析,本文提出一种多维视角多元信息融合的人群密度估计方法(multi-dimension perspective and multivariate information fusion, MDPMIF),采用多维视角思想由“上-左-右-下”方向对视角变化进行信息编码,通过递进聚合方式捕获深层次全局上下文信息,并同步提取多维度视角的尺度关系特征. 同时,设计联合学习策略,实现全局上下文信息与多尺度关系特征相互补充,增强输出密度图的质量. 进一步地,在MDPMIF网络的基础上,提出一种高低密度多维视角多元信息融合人群计数网络(high and low density and multi-dimension perspective multivariate information fusion crowd counting network, HLMMNet),设计高低密度区分策略(density domain grade, DDG),实现单幅图像高,低密度区域自适应划分,高密区域保持MDPMIF网络估计结果,低密区域采用检测方法实现人群计数修正,提升人群计数精度.

1 高低密度多维视角多元信息融合人群计数网络

HLMMNet网络的基本结构如图1所示,包括多维视角多元信息融合人群密度估计网络MDPMIF、高低密度区分策略DDG以及人群计数模块.

1.1 多维视角多元信息融合网络

MDPMIF网络架构主要包括骨架网络、递进聚合(progressive aggregation, PA)模块、空洞卷积模块以及语义嵌入特征融合模块,具体如图2所示.

1.1.1 密度图生成

要完成人群计数任务,需要获得输入图像的人群密度图. 密度图是以特殊高亮的形式显示人群分布

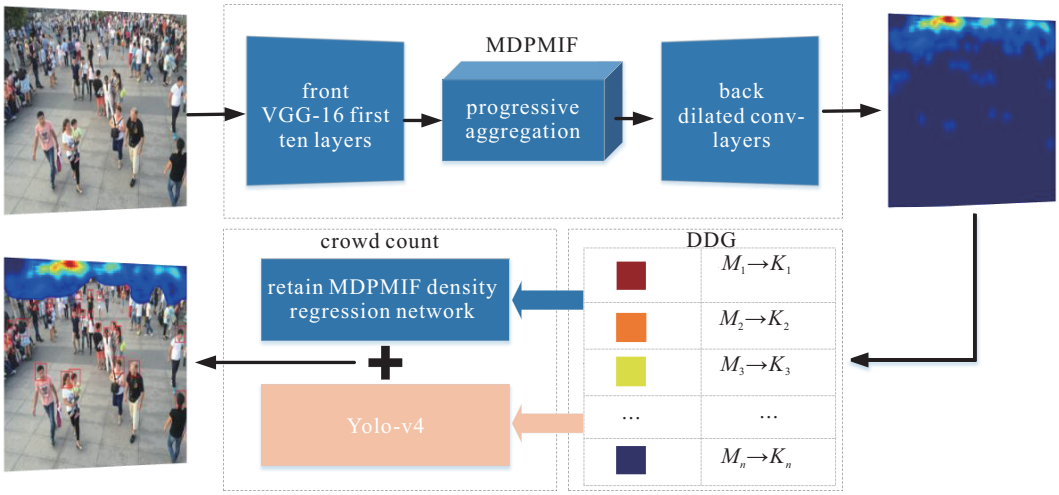


图 1　HLMMNet 网络结构

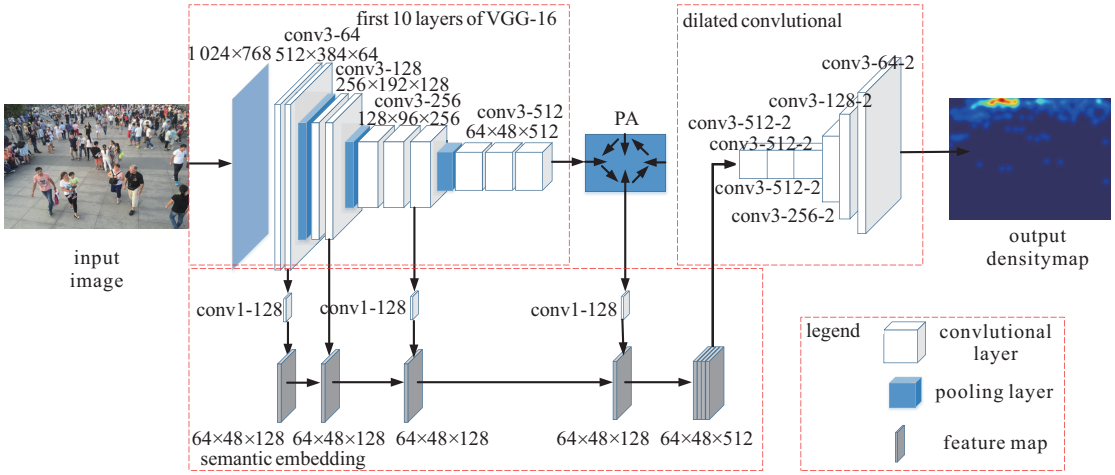


图 2　MDPMIF 网络结构

位置的图像,从空间上反映了人群分布的疏密程度,其积分值代表图像中的人数. 密度图生成如下:

$$D=\sum_{i=1}^{N}\delta(\boldsymbol{x}-\boldsymbol{x}_i)\otimes G_{\sigma(\boldsymbol{x}_i)}. \tag{1}$$

其中:函数$\delta(\boldsymbol{x}-\boldsymbol{x}_i)$表示人头标记点图像$\boldsymbol{x}$中第$i$个坐标为$\boldsymbol{x}_i$的人头标记点的密度平滑区域,区域大小与积分为1的自适应高斯滤波器$G_{\sigma(\boldsymbol{x}_i)}$一致;$N$为图像中人头标记点总数;$\otimes$表示卷积运算. 为避免漏检,采用自适应高斯滤波器$G_{\sigma(\boldsymbol{x}_i)}$与$\delta(\boldsymbol{x}-\boldsymbol{x}_i)$进行卷积,其中滤波器大小$\sigma(\boldsymbol{x}_i)=\beta\bar{d}_i$,$\bar{d}_i$表示标记点$\boldsymbol{x}_i$与其最近的$K$个人头之间的平均距离. 可以看出,密度图$D$的各像素值取值范围为[0, 1]. 将密度图$D$的各像素值规范化至[0, 255](记作规范化密度图D'),并通过JET颜色映射,密度图以特殊高亮形式显示. 人员聚集度越高,密度图对应区域颜色越亮丽;人员越分散,对应区域颜色越暗.

1.1.2　骨架网络与空洞卷积结构

VGG-16[21]由牛津大学计算机视觉实验室提出,设计之初是为解决ImageNet中的1 000类图像分类和定位问题. 因其网络架构对大部分物体的特征提取提供了较为合适的感受野且易于捕获细节信息,有效改善了密度图估计精度[12],目前已成为人群密度估计常用骨架网络之一. 本文亦是如此,选择VGG-16前10层作为MDPMIF的骨架网络,保留VGG-16强大的传输学习能力,为后续模块衔接及特征融合奠定基础. 在骨架网络后端,反卷积层的复杂性和执行延迟会对网络执行速度有拖累. 空洞卷积通过在标准卷积核内设置0值区域的方式扩大了感受野,能够保证在相同输出分辨率的前提下提取更深层次的显著性信息,降低网络复杂性. 因此,本文在MDPMIF网络末端采用空洞卷积替代反卷积层操作,具体计算为

$$(\boldsymbol{a}*_l\boldsymbol{w})(i)=\sum_{k=1}^{K}\boldsymbol{a}[i+kl]\boldsymbol{w}[k]. \tag{2}$$

其中:$\boldsymbol{w}$为卷积核;k为卷积核尺寸;$\boldsymbol{w}[k]$为大小是k的卷积核;$\boldsymbol{a}[i]$为第i个输入;$*_l$为空洞卷积运算;l为扩张率,描述卷积核处理数据时采样的步幅,调整l可自适应调整感受野大小. 本文采用膨胀率为2、卷积核为3×3的空洞卷积,输入与输出具有相同的维度,

最终输出密度图.

1.1.3 递进聚合模块

不同视角维度下人群透视现象呈现差异,图像蕴含的上下文表达、尺度层级关系均不相同. 本文设计的递进聚合PA模块从“上、左、右、下”4个视角维度出发,递进聚合多视角变化的上下文表达,捕获全局空间上下文信息. 利用不同规格卷积核、空洞卷积核的组合模拟人眼成像,分别提取4个视角维度的多尺度特征,采用联合学习策略生成全局尺度关系描述;后经1×1卷积实现全局上下文表达、全局尺度关系信息集成,得到更全面的视角变换感知,提升MDPMIF网络最终输出密度图质量. PA模块结构如图3所示.

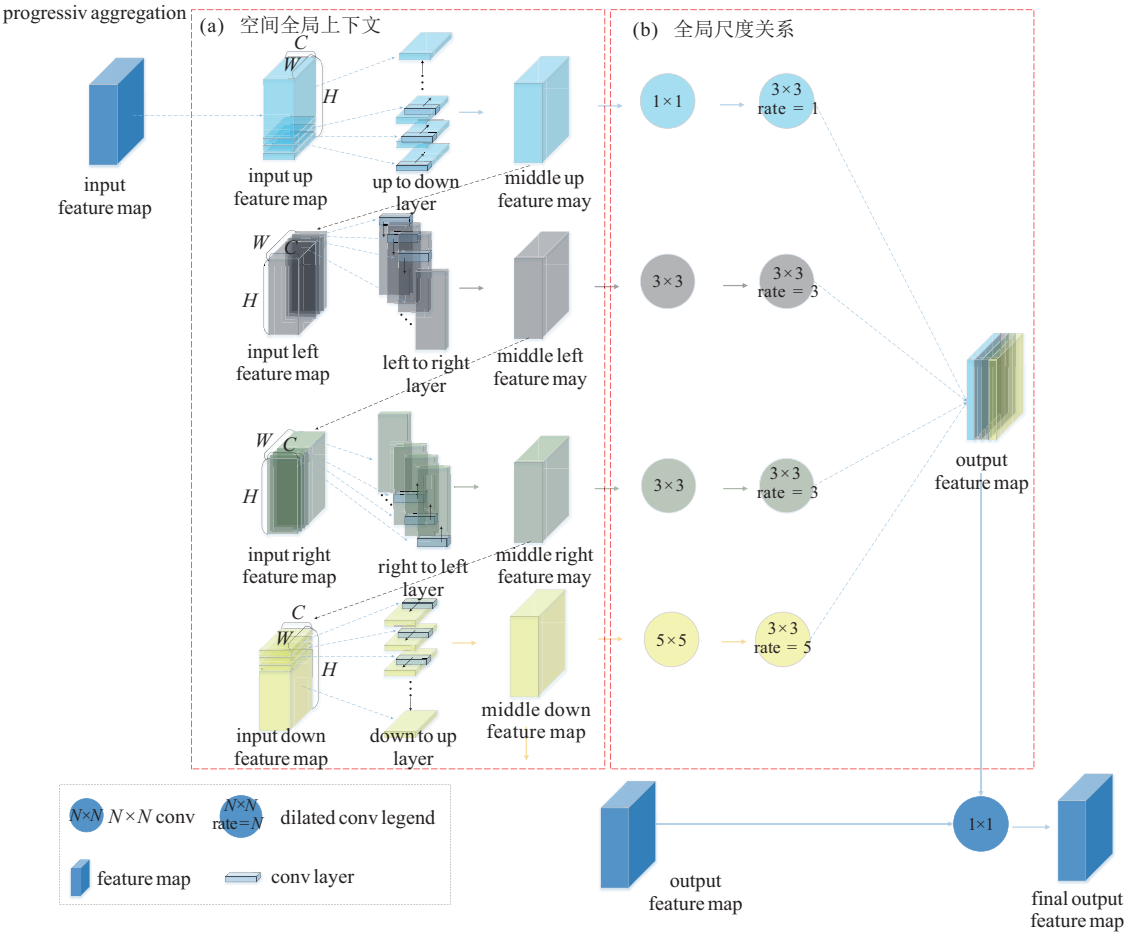


图 3 PA模块示意图

以视角维度“上”为例,具体说明PA模块不同视角维度上下文表达的递进聚合过程. 如图3(a)所示,输入特征图F的大小为$C \times H \times W$. 其中:C表示上一卷积层的卷积核个数,H表示特征图的高度,W表示特征图的宽度. 将特征图F的高度H均分为N份,即将特征图F分为N个大小为$C \times W \times (H/N)$的特征块,用F_N^i表示第i个特征块,$i \in [1, N]$. 视角维度“上”的卷积层由$C \times (H/N)$大小的卷积核c及ReLU激活函数组成. 按照下式进行由上至下(up to down)卷积过程计算:

$$\mathrm{FD}_N^i = \begin{cases} L(F_N^i), \ i = 1; \\ L(F_N^i + \mathrm{FD}_N^{i-1}), \ i = 2, 3, \ldots, N. \end{cases} \tag{3}$$

其中:$L(\cdot)$表示进行“上”视角维度卷积层(Conv+ReLU)运算. 将特征块F_N^1送入“上”视角维度卷积层后,生成一个与F_N^1同样大小的特征块,记作FD_N^1;将FD_N^1与F_N^2加和送入“上”视角维度卷积层得到FD_N^2;经过不断迭代后,输出第N个特征块FD_N^N;最后,将$\mathrm{FD}_N^1, \ldots, \mathrm{FD}_N^i, \ldots, \mathrm{FD}_N^N$连接起来,生成“上”视角维度层输出的特征图FD,其大小为$C \times H \times W$,与特征图F尺寸一致. 对于“左、右、下”3个视角维度的操作,除滑动方向不同外,计算与“上”视角维度类似.

在图3(b)结构的“上”维度视角,真实场景同等密度人群呈现“稀疏大头部至稠密小头部”样透视现象,PA模块的信息编码过程趋向关注“稀疏大头部”,因此,该维度的多尺度关系变化较小,可通过擅长反映细微特征的小感受野(1×1卷积核及rate = 1的空洞卷积核组合)捕获对应的尺度层级特征;“下”维度

视角呈现“稠密小头部至稀疏大头部”样透视现象,信息编码过程趋向关注“稠密小头部”,则该维度的多尺度关系变化偏大,选择大感受野(5×5卷积核及rate = 5的空洞卷积核组合)易于提取更丰富的尺度层级特征;“左”和“右”维度视角,“稠密小头部”与“稀疏大头部”同步编码且包含较多边缘人群信息,可视作多尺度特征变化中等,利用中等阈值感受野(3×3卷积核及rate×3的空洞卷积核组合)感知尺度层级特征. 最后,通过如下所示的联合学习策略生成全局尺度关系描述$U(U = m_4)$:

$$m_l = \mathrm{Upsample}(m_l) + F(m_{l-1}, m_l). \tag{4}$$

其中:$l = [1,2,3,4]$,分别表示“下、右、左、上”视角维度;m_l为l视角维度的多尺度特征,m_0为元素全1的m_1同阶矩阵;$F(m_{l-1}, m_l)$的函数功能为矩阵逐元素乘操作.

1.1.4 语义嵌入特征融合模块

骨架网络输出图像仅有输入端的1/8,图像在此过程中会丢失部分细节特征. 常规方法一般采用concat或add将多特征图进行直接加和,但其利用低级特征作为上采样高级特征的残差时,低级特征包含语义信息较少,不足以复原较多语义信息,导致特征图质量降低. 基于此,本文设计语义嵌入特征融合策略,采用高级特征双线性插值上采样与低级特征逐元素相乘的方式克服常规特征融合方法带来的问题. 语义嵌入特征融合策略将网络前端不同层、不同大小的特征与PA模块的输出融合,在此过程中,不同层、不同尺寸的特征分别使用的卷积核将通道尺寸减少至相同的维度,利用上采样统一多组特征图尺寸,进行语义嵌入融合后送至网络后端,在补充低层特征信息的同时增加网络可学习特征信息量,提升最终输出密度图的质量.

1.2 高低密度区分策略DDG与人群计数

由密度图生成规则可知,密度图区域颜色不同,其实质是相同像素面积对应人员数量不同. 本文根据像素值大小,将规范化密度图D'均分为n个等级,记作$M_1, M_2, \ldots, M_i, \ldots, M_n, i = 1,2,\ldots,n$. M_n人群密度等级最高,M_1人群密度等级最低. 经过多次实验,取$n = (255+1)/4 = 64$,即像素值为[0,3]区间的像素被划分为M_1等级,像素值为[4,7]区间的像素被划分为M_2等级,依此类推.

等级M_i包含N_{M_i}个像素在规范化密度图D'中,M_i中任意像素值$d'_{M_i,j} \in \left[(i-1)\times\dfrac{256}{n}, i\times\dfrac{256}{n}-1\right]$,$j = 1,2,\ldots,N_{M_i}$. 将属于$M_i$等级的$N_{M_i}$个像素由规范化密度图$D'$映射回归至密度图$D$,映射回归后的像素值记作$d_{M_i,j}$,则$d_{M_i,j} \in [0,1]$且$d_{M_1,j} < d_{M_2,j} < \ldots < d_{M_i,j} < \ldots < d_{M_n,N_{M_n}}$. 同时,对$M_i$等级的$N_{M_i}$个像素$d_{M_i,j}$做积分,即可得到人员数量$K_i, M_1, M_2, \ldots, M_i, \ldots, M_n$等级依次对应人数$K_1, K_2, \ldots, K_i, \ldots, K_n$.

采用下式描述当前图像中各密度等级与最高密度等级之间的差距,从而对MDPMIF网络输出的预测密度图进行图像分割:

$$\frac{M_i}{M_n} = \frac{\max(d_{M_i,j})}{\max(d_{M_n,1})} > \delta \Rightarrow \begin{cases} M_i\text{为高密度区域, 真;} \\ M_i\text{为低密度区域, 假;} \end{cases} i \in [1, n-1]. \tag{5}$$

将大于阈值边界δ的划分为一类,定义为高密度区域;小于阈值边界的部分划分为另一类,定义为低密度区域. 高密度区域保持MDPMIF网络密度图回归计数结果,低密度区域采用Yolo-v4[22-23]进行人头检测得到稀疏区域人群计数结果,二者之和作为最终计数结果.

高低密度区域划分问题可视为分割问题,选用Dice损失函数对式(5)的边界阈值δ进行训练,获取最佳δ值. Dice损失函数计算为

$$L_D = 1 - \frac{2|X \bigcap Y| + 1}{|X| + |Y| + 1}. \tag{6}$$

其中:X为分割图像真值,是手工标注图像;Y为预测分割图像,即执行DDG高低密度区分策略后的分割图像. 因分母重复计算X和Y,此处分子补充系数2;另外,分子分母同时加1,避免$|X|$和$|Y|$都为零时分母为0,减少过拟合.

1.3 损失函数

HLMMNet网络采用如下式(7)所示的损失函数,综合如下式(8)所示的MDPMIF网络损失L_M、式(6)所示的DDG高低密度划分损失L_D以及人员检测部分损失L_C. L_C采用Yolo-v4模型损失计算方法,具体见文献[22],此处不再赘述.

$$L = L_M + L_D + L_C, \tag{7}$$

$$L_M = \frac{1}{2N}\sum_{i=1}^{N} \|Z(X_i) - Z_i^{\mathrm{GT}}\|_2^2. \tag{8}$$

其中:Z_i^{GT}为输入图像的真值,$Z(X_i)$为MDPMIF输出结果,N为当前训练的图像数量.

总损失L采用随机梯度下降法进行优化. 在每次迭代中,L_M、L_D和L_C的梯度被交替计算并用于更新相应的参数,初始化DDG和MDPMIF的学习速率为10^{-4}.

2 实验分析

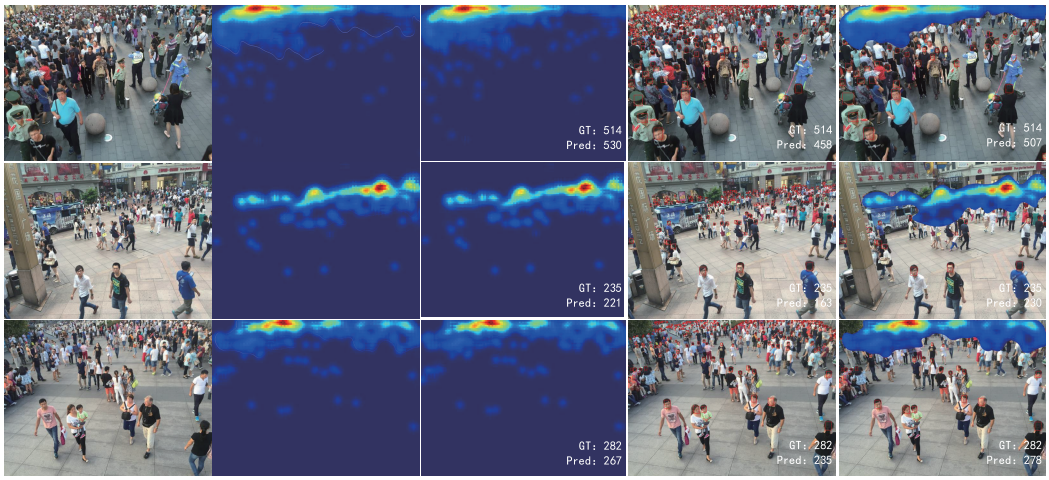
利用 Shanghai Tech[15]、Mall[24]、Wordexpo'10[25] 主流数据集及一个自建数据集开展实验,对比算法选用 MCNN[15]、Switch-CNN[8]、MSCNN[16]、CSRNet[12]、DecideNet[14]、ACSCP[26]、D-ConvNet-v1[27]、IG-CNN[28]、SCAR[29] 等近几年先进的人群计数方法.

所有实验均在 Ubuntu 系统下进行,GPU 型号为 TitanV 及 V100,环境配置为 CUDA9.0 + anaconda3 + python3 + Pytorch1.7.0. 为使模型充分训练,采用数据增强方法对样本图像随机进行裁剪、旋转、放缩等操作,扩充数据集样本数量,增强 CNN 模型的鲁棒性. 此外,将实验所用数据集进行人头检测框标注,用于对 Yolo-v4 进行预训练,预训练参数用于 HLMMNet 方法的最终训练.

现有人群密度计数方法均采用平均绝对误差 (mean absolute error, MAE) 和均方误差 (mean square error, MSE) 作为评价指标. 为了较好地进行实验对比分析,与前述对比算法保持一致,使用平均绝对误差、均方误差作为评价指标.

2.1 Shanghai Tech 数据集实验与分析

Shanghai Tech 数据集共包含 1 198 幅图像,共计 330 165 个已标记人头,是目前已知标记人数最多的数据集. 数据集共分为两部分 part_A 和 part_B. part_A 包含 482 幅图像,来源于互联网,均为高拥挤程度图像;part_B 包含 718 幅图像,来源于上海的街道,每幅图像中既包含拥挤区域也包含人群稀疏区域. 由于 part_A 不契合本文高低密度网络 HLMMNet 的关注点,本文只使用 Shanghai Tech 数据集的 part_B 部分进行 HLMMNet 实验. part_A 中,350 幅图像用于训练,其余 132 张用于测试;part_B 中,400 幅图像用于训练,其余 318 张用于测试. Shanghai Tech part_B 数据集单幅图像的 MDPMIF、HLMMNet、Yolo-v4 实验结果如图 4 所示,part_A 和 part_B 的多算法性能指标结果对比如表 1 所示.



(a) 原图 (b) MDPMIF 真值 (c) MDPMIF 预测 (d) Yolo-v4 预测 (e) HLMMNet 结果

图 4 HLMMNeT 网络测试结果

表 1 Shanghai Tech 算法性能对比

method	part_A		part_B	
	MAE	MSE	MAE	MSE
Yolo-v4[22]	–	–	21.4	31.3
MCNN[15]	110.2	173.2	26.4	41.3
Switch-CNN[8]	90.4	135.0	21.6	33.4
MSCNN[16]	83.8	127.4	17.7	30.2
DecideNet[14]	–	–	21.5	31.9
CSRNet[12]	68.2	115.0	10.6	16.0
ACSCP[26]	75.7	102.7	17.2	27.4
D-ConvNet-v1[27]	73.5	112.3	18.7	26.0
IG-CNN[28]	72.5	118.2	13.6	21.1
SCAR[29]	66.3	114.1	9.5	15.2
DUBNet[30]	64.6	106.8	7.7	12.5
MDPMIF	62.9	103.2	8.1	13.2
HLMMNet	–	–	7.4	11.9

注: 文献 [14] 重点关注高低密度的区分,其原文仅对 part_B 进行了实验.

由实验结果可知,在 part_A 部分,纯密图方法 MDPMIF 优于对比方法,相较于文献 [30],MDPMIF 的 MAE 降低了 1.7,MSE 降低了 3.6,表明多维度视角方式提取图像上下文信息与尺度特征能够较好地改善视角变换给人群计数带来的影响. 在 part_B 部分,MDPMIF 优于绝大多数对比方法,由于文献 [30] 将多模型输出结果利用统计学的方法取均值,在一定程度上降低了图像中稀疏区域因背景干扰带来的人头数量估计的不确定性,其性能指标较 MDPMIF 略有优势,也表明了多维度视角方式对密集程度较高区域更为有效. 检测 + 密度图方法 HLMMNet 在 part_B 部分的 MAE 与 MSE 分别达到了 7.4 和 11.9,表明其准确度和鲁棒性较对比方法均有较大提升;且

HLMMNet方法的精度不仅优于先进的对比方法,也优于MDPMIF,验证了本文高低密度区域自适应划分策略的有效性,结合密度图与检测方法优势,改善了全密度算法对于人员稀疏区域造成的计数高估,弥补了全人头框检测算法对于高密人群由于遮挡而产生的误检、漏检问题.

2.2　Mall数据集实验与分析

Mall数据集是使用安装在购物中心的监视摄像机收集到的数据集,该数据集具有人群密度变化大、活动模式多、 透视畸变以及遮挡严重的特点. 实验选用1 600幅图像作为训练集,400幅图像作为测试集. Mall数据集算法性能对比结果如表2所示. 由表2可知,MDPMIF、HLMMNet在Mall数据集中表现较好. 与文献[29]相比,MDPMIF模型MAE降低了0.01,但MSE升高了0.04;HLMMNet模型MAE降低了0.22,MSE降低了0.19. Mall数据集来源于室内商场场景,其人群密集区域的密集程度较弱,稀疏人群的稀疏程度较高,基于DDG策略的“密度图+检测”思想对此类特点场景精度的提升帮助更为明显.

表2　Mall数据集算法性能对比

方法	MAE	MSE	方法	MAE	MSE
Yolo-v4[22]	1.87	2.19	E3D[31]	1.64	2.13
MCNN[15]	2.21	7.33	ACSCP[26]	1.70	2.35
Switch-CNN[8]	2.01	6.25	IG-CNN[28]	1.65	2.14
MSCNN[16]	2.12	7.04	SCAR[29]	1.59	2.01
DecideNet[14]	1.52	1.90	MDPMIF	1.58	2.05
CSRNet[12]	1.71	2.06	HLMMNet	1.37	1.82

2.3　Worldexpo'10数据集实验与分析

Worldexpo'10的10个数据集是从2010年世博会103个不同场景中收集的1 132个带注释的视频序列. 这些数据共有3 980帧,统一归一化至576×720大小. 本文用于训练的图像尺寸是144×144,训练部分使用3 380帧,剩余数据用于测试. 测试场景提供了$S_1 \sim S_5$共计5个感兴趣区域(region of interest, ROI),采用文献[25]预训练方法完成ROI区域识别,分别统计5个ROI区域内人员数量. 实验结果性能比较如表3所示.

Worldexpo'10数据集5个场景差异较大,S_1、S_5场景相对稀疏程度高,人员数量少;S_4场景存在较多遮挡物,背景复杂程度高;S_2、S_3场景人员拥挤程度相对较高,背景干扰较小. 由表3可见,不同方法对特定的场景各有优势,本文提出的方法在Worldexpo'10数据集所有5个场景中的Avg-MAE达到7.9和7.6,相较对比方法,HLMMNet性能最佳,MDPMIF方法

表3　Worldexpo'10数据集算法性能对比

方法	MAE					
	S_1	S_2	S_3	S_4	S_5	Avg
MCNN[15]	3.4	20.6	12.9	13.0	8.1	11.6
MSCNN[16]	7.8	15.4	14.9	11.8	5.8	11.7
Switch-CNN[8]	4.4	15.7	10.0	11.0	5.9	9.4
DecideNet[14]	2.0	13.14	8.9	17.4	4.75	9.23
CSRNet[12]	2.9	11.5	8.6	16.6	3.4	8.6
ACSCP[26]	2.8	14.05	9.6	8.1	2.9	7.7
D-ConvNet-v1[27]	1.9	12.1	20.7	8.3	2.6	9.1
IG-CNN[28]	2.7	13.9	9.9	12.8	3.1	8.5
SCAR[29]	2.6	12.9	9.6	13.9	2.7	8.3
MDPMIF	3.0	11.2	9.3	13.5	3.2	7.9
HLMMNet	2.5	10.4	8.9	13.1	3.17	7.6

优于绝大多数对比方法. 文献[26]利用GAN与对抗性一致损失弱化了背景干扰,在S_4场景的计数精度优势显著,使得MDPMIF方法的Avg-MAE略逊0.1;引入本文高低密区分策略后,有效改善了稀疏区域检测精度,最终Avg-MAE较文献[26]降低了0.1,验证了HLMMNet模型的有效性.

2.4　自建数据集实验与分析

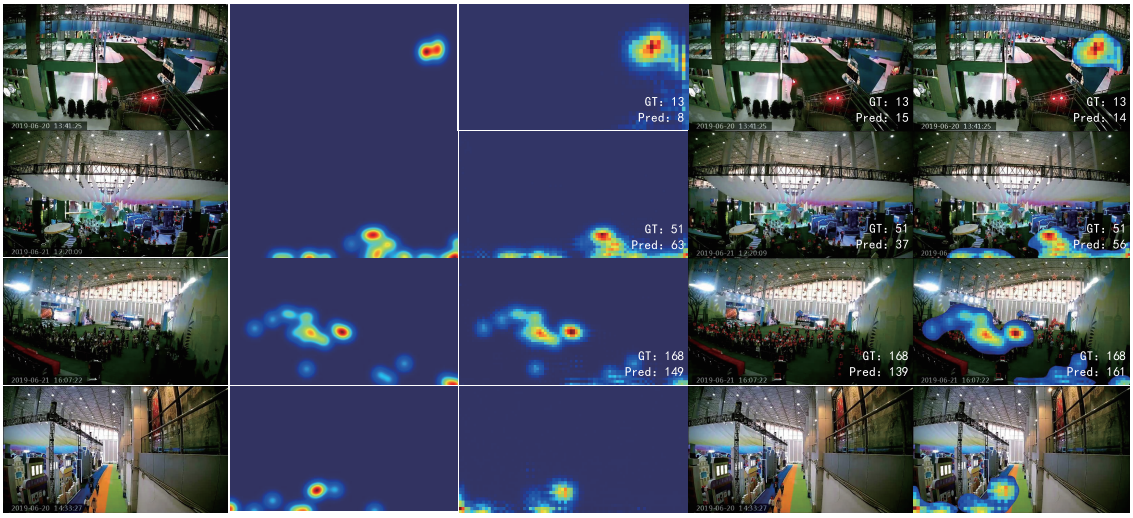
由于现有人群数据集多为室外场景,Mall数据集虽是室内场景人群数据集,但其缺少场景变化,背景单一. 本文构建了一个展厅数据集,具有背景复杂多变、人员流动性大的特点. 该数据集来源于青海测控测量技术及应用展览会展厅内4个不同场景(东扶、西扶、主会场和通道). 针对4个场景,分别采集多天多时段(9:00～16:00)视频信息,每个场景选取500幅图像,选取共计2 000幅图像构成本文自建数据集. 数据集图像未进行预处理,保留了原始图像所具有的噪声以及亮度变化干扰. 实验选用其中的1 600幅(4种场景各400幅)训练网络模型,其余400幅(4种场景各100幅)进行测试. 自建数据集单幅图像实验结果如图5所示,多算法性能对比如表4所示.

表4　自建数据集算法性能对比

方法	MAE	MSE	方法	MAE	MSE
MCNN[15]	34.5	57.1	PGCNet[32]	10.4	15.3
Switch-CNN[8]	31.3	47.9	IG-CNN[28]	10.0	14.9
MSCNN[16]	26.8	42.3	MDPMIF	9.3	14.6
CSRNet[12]	15.8	25.4	HLMMNet	8.2	13.7
SCAR[29]	13.4	19.8			

2.5　算法速度对比分析

为全面评估本文方法,从训练模型参数量大小、模型运行速度两方面进行对比实验,实验结果如表5所示.



(a) 原图 (b) MDPMIF真值 (c) MDPMIF预测 (d) Yolo-v4预测 (e) HLMMNet结果

图5 自建数据集HLMMNet网络实验结果

表5 算法速度对比

方法	大小 (MB)	测试图像平均运行速度/s			
		Shanghai Tech-B	Mall	自建	Worldexpo'10
MCNN[15]	19.2	2.31	0.32	0.35	1.15
Switch-CNN[8]	32.2	2.71	0.43	0.45	1.35
MSCNN[16]	22.2	2.34	0.32	0.31	1.14
ACSCP[26]	21.6	2.15	0.30	0.34	1.05
CSRNet[12]	16.26	1.97	0.26	0.21	0.93
IG-CNN[28]	23.5	2.30	0.26	0.27	1.12
SCAR[29]	21.8	2.24	0.31	0.31	1.27
MDPMIF	21.4	2.00	0.27	0.23	0.94
HLMMNet	28.7	2.32	0.33	0.32	1.12

由表5可见: Switch-CNN模型网络结构最大,运行速度也最慢; CSRNet模型较小,其单列结构运行速度较快; MCNN模型采用多列结构, MSCNN模型使用尺寸较大的卷积核,导致模型参数量较高,运行速度较慢. 相较而言,纯密度MDPMIF方法在保证高精度的同时,模型体积较小,且运行速度较快,分析原因有以下3点: 1) MDPMIF采用了空洞卷积思想,在扩大感受野的基础上大大减少了参数量; 2) 使用VGG-16单列结构,无多列冗余; 3) PA模块尺度信息提取轻量化,提升了运行速度. 为提高准确度, HLMMNet链接了Yolo-v4检测模型,使得模型总体积有所增大,运行速度也有所牺牲. 如何轻量化密度与检测模型是本文后续改进的重点.

3 结 论

本文提出了一种新的人群计数网络HLMMNet,在密度图的基础上利用"Switch"思想,通过对密度图进行稀疏与稠密区域的划分,有效结合检测与密度图方法的优势,达到更好的估计人数效果. HLMMNet网络由多维视角多元信息融合网络(MDPMIF)、高低密度区分策略(DDG)以及人群计数3部分组成. MDPMIF网络由"上-左-右-下"方向对视角变化进行信息编码,捕获深层次全局上下文信息以及全局尺度关系描述;将全局上下文信息、全局尺度关系描述与低阶特征融合,增强输出密度图的质量. 实验结果表明,所提出的HLMMNet网络较其他先进对比方法性能有较大幅度的提升.

参考文献(References)

[1] Li X, Chen M, Nie F, et al. A multiview-based parameter free framework for group detection[C]. Proceedings of the AAAI Conference on Artificial Intelligence. New York, 2017: 4147-4153.

[2] Coar S, Donatiello G, Bogorny V, et al. Toward abnormal trajectory and event detection in video surveillance[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 683-695.

[3] Liu C Y, Wang Q, Bi X J. Multi-target and small-scale vehicle target detection method[J]. Control and Decision, 2021, 36(11): 2707-2712.

[4] Chen M L, Wang Q, Li X L. Patch-based topic model for group detection[J]. Science China Information Sciences, 2017, 60(11): 1-7.

[5] Wang Q, Gao J Y, Lin W, et al. NWPU-crowd: A large-scale benchmark for crowd counting and localization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(6): 2141-2149.

[6] Ma Z H, Wei X, Hong X P, et al. Bayesian loss for crowd count estimation with point supervision[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 6141-6150.

[7] Yu J, Hong C Q, Rui Y, et al. Multitask autoencoder model for recovering human poses[J]. IEEE Transactions on Industrial Electronics, 2018, 65(6): 5060-5068.

[8] Sam D B, Surya S, Babu R V. Switching convolutional neural network for crowd counting[C]. IEEE Conference

on Computer Vision and Pattern Recognition. Honolulu, 2017: 4031-4039.

[9] Zhao L, He Z H, Cao W M, et al. Real-time moving object segmentation and classification from HEVC compressed surveillance video[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(6): 1346-1357.

[10] Kang D, Ma Z, Chan A B. Beyond counting: Comparisons of density maps for crowd analysis tasks—counting, detection, and tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(5): 1408-1422.

[11] Gao J Y, Wang Q, Li X L. PCC net: Perspective crowd counting via spatial convolutional network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019(99): 1.

[12] Li Y H, Zhang X F, Chen D M. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 1091-1100.

[13] Wang L Y, Yin B Q, Tang X, et al. Removing background interference for crowd counting via de-background detail convolutional network[J]. Neurocomputing, 2019, 332: 360-371.

[14] Liu J, Gao C Q, Meng D Y, et al. DecideNet: Counting varying density crowds through attention guided detection and density estimation[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 5197-5206.

[15] Zhang Y Y, Zhou D S, Chen S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]. IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 589-597.

[16] Zeng L K, Xu X M, Cai B L, et al. Multi-scale convolutional neural networks for crowd counting[C]. IEEE International Conference on Image Processing. Beijing, 2017: 465-469.

[17] Meng Y B, Ji T, Liu G H, et al. Encoding-decoding multi-scale convolutional neural network for crowd counting[J]. Journal of Xi'an Jiaotong University, 2020, 54(5): 149-157.

[18] Sindagi V A, Patel V M. CNN-Based cascaded multi-task learning of high-level prior and density estimation for crowd counting[C]. The 14th IEEE International Conference on Advanced Video and Signal Based Surveillance. Lecce, 2017: 1-6.

[19] Sindagi V A, Patel V M. Generating high-quality crowd density maps using contextual pyramid CNNs[C]. IEEE International Conference on Computer Vision. Venice, 2017: 1879-1888.

[20] Sheng B Y, Shen C H, Lin G S, et al. Crowd counting via weighted VLAD on a dense attribute feature map[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(8): 1788-1797.

[21] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. 2014, arXiv: 1409.1556.

[22] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J/OL]. 2020, arXiv: 2004.10934.

[23] Li H J, Wang H Y, Li Y, et al. An object detector based on visual feature region proposal[J]. Control and Decision, 2020, 35(6): 1323-1328.

[24] Chen K, Loy C C, Gong S G, et al. Feature mining for localised crowd counting[C]. Proceedings of the British Machine Vision Conference. Surrey, 2012: 3-27.

[25] Zhang C, Li H S, Wang X G, et al. Cross-scene crowd counting via deep convolutional neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. Boston, 2015: 833-841.

[26] Shen Z, Xu Y, Ni B B, et al. Crowd counting via adversarial cross-scale consistency pursuit[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 5245-5254.

[27] Shi Z L, Zhang L, Liu Y, et al. Crowd counting with deep negative correlation learning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 5382-5390.

[28] Sam D B, Sajjan N N, Babu R V, et al. Divide and grow: Capturing huge diversity in crowd images with incrementally growing CNN[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 3618-3626.

[29] Gao J Y, Wang Q, Yuan Y. SCAR: Spatial-channel-wise attention regression networks for crowd counting[J]. Neurocomputing, 2019, 363: 1-8.

[30] Oh M H, Olsen P, Ramamurthy K N. Crowd counting with decomposed uncertainty[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11799-11806.

[31] Zou Z K, Shao H L, Qu X Y, et al. Enhanced 3D convolutional networks for crowd counting[J/OL]. 2019, arXiv: 1908.04121.

[32] Yan Z Y, Yuan Y C, Zuo W M, et al. Perspective-guided convolution networks for crowd counting[C]. IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 952-961.

作者简介

孟月波(1979–), 女, 副教授, 博士, 从事智能感知、理解与智动化系统以及建筑智能化技术等研究, E-mail: mengyuebo@163.com;

陈宣润(1995–), 男, 硕士生,从事深度学习、计算机视觉的研究, E-mail: 544268224@qq.com;

刘光辉(1976–), 男, 副教授, 博士, 从事智能环境感知与调控等研究, E-mail: guanghuil@163.com;

徐胜军(1976–), 男, 副教授, 博士, 从事人工智能与智动化系统、模型仿真智能控制理论等研究, E-mail: duplin@sina.com;

李彤月(1994–), 女, 博士生,从事计算机视觉、强化学习的研究, E-mail: litongyue@163.com.

(责任编辑: 郑晓蕾)