基于元权重网络的跨场景点预测人群计数方法

徐 昕[1],谭卓林[1],高陈强[1*],席 跃[2]
(1. 重庆邮电大学通信与信息工程学院,重庆 400065;2. 澳门大学科技学院,澳门 999078)

摘 要: 跨场景人群计数由于光照、尺度、拍摄角度和人群密度等因素引起的数据分布差异,导致在不同场景下的计数精度下降. 针对现有人群计数模型在跨场景应用时存在的问题,本文提出了一种基于元学习的场景感知重新加权方法. 该方法通过设计点预测计数模型直接预测每个行人的精确坐标,避免了传统密度图方法的定位模糊问题. 元权重网络从元数据中学习显式点预测损失的加权方案,通过场景感知分支将每个场景视为一个单独的学习任务,利用不同场景的内在特征实现自适应的加权方案,降低标注噪声对模型跨场景泛化能力的影响. 此外,针对现有数据集在教学领域的局限性构建了新的校园多场景人群计数数据集(Multi-Scene Crowd counting dataset,MS-Crowd),为跨场景研究提供了更全面的评估基准. 实验结果表明该方法在MS-Crowd和户外公开数据集ShanghaiTech上的平均绝对误差(Mean Absolute Error,MAE)分别降低了19.7%和10.7%,验证了方法的有效性.

关键词: 人群计数;人群定位;元学习;跨场景
基金项目: 国家重点研发计划(No.2022YFA1004100)
中图分类号: TN911.73;TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)09-3371-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250285

Cross-Scene Point Prediction Crowd Counting Method Based on Meta-Weight-Net

XU Xin[1], TAN Zhuo-lin[1], GAO Chen-qiang[1*], XI Yue[2]
(1. *School of Communications and Information Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*;
2. *Faculty of Science and Technology*, *University of Macau*, *Macao* 999078, *China*)

Abstract: Cross-scene crowd counting often suffers from degraded accuracy due to data distribution disparities caused by factors such as illumination, scale, camera angles, and crowd density. To address the limitations of existing crowd counting models in cross-scene applications, a meta-learning-based scene-aware reweighting method is proposed. Instead of relying on traditional density map approaches that suffer from localization ambiguity, the method employs a point prediction counting model to directly estimate the precise coordinates of each individual. A meta-weight network is introduced to learn an explicit weighting scheme for the point prediction loss from meta-data, while a scene-aware branch treats each scene as an independent learning task, leveraging intrinsic features across scenes to adaptively adjust the weighting scheme and mitigate the impact of annotation noise on cross-scene generalization. Furthermore, to overcome the limitations of existing datasets in educational settings, a new campus multi-scene crowd counting dataset (MS-Crowd) is constructed, providing a more comprehensive benchmark for cross-scene evaluation. Experimental results demonstrate that the proposed method reduces the mean absolute error (MAE) by 19.7% and 10.7% on the MS-Crowd and the public outdoor dataset ShanghaiTech, respectively, validating its effectiveness.

Key words: crowd counting; crowd localization; meta-learning; cross scene
Foundation Item(s): National Key Research and Development Program of China (No.2022YFA1004100)

1 引言

人群计数是指对图像中人的数量、分布情况和个体位置进行估计. 这项技术对于智慧教育、智慧城市建设、交通流量变化和公共安全管理等领域具有重要应

收稿日期:2025-04-14;录用日期:2025-08-22;责任编辑:覃怀银
*通讯作者:高陈强

用价值[1]. 在校园室外场景下,人群计数可用于分析学生的流动规律、优化校园资源配置、提升安全管理水平,并为校园突发事件的应急响应提供数据支持[2]. 在室内教学场景中,该技术可用于分析课堂出勤状况和分布情况,辅助教师调整教学策略,提高课堂互动效果. 因此,精准的人群计数在保障教学场景安全和提升教学管理智能化方面具有重要意义.

早期的人群计数方法主要依赖于目标检测框架,通过检测人体躯干或头部来获取图像中的人数[3]. 然而,在目标排列密集且存在严重遮挡[4],或背景复杂度较高时[5],检测框架的准确性往往无法得到有效保障. 为了解决这一问题,研究人员提出了基于密度图的人群计数方法. 该方法将图像转换为单通道人群密度图,通过对密度图求和得到最终的预测人数. 文献[6]通过多个分支的卷积神经网络来捕获不同尺度的特征,将各分支的预测结果合并得到最终的预测密度图. 文献[7]提出一个逆向透视网络,将图像的透视变换为同一尺度,避免复杂的多尺度分支网络架构. 文献[8]则提出了一种基于特征中心性编码的方案,针对不同的注意力图调整输入节点特征,增强网络对关键节点的敏感度. 基于密度图的方法有效地提供了人群密度的关键信息,并在拥挤场景中表现出较高的计数精度. 然而,该方法无法提供实例级的精准定位,预测结果仅能反映人群的稠密程度,这无法满足要求更高的人流监控应用需求. 尽管一些基于密度图的方法能够展示定位结果,但它们都需要复杂的后处理流程. 例如,文献[9]通过密度图中局部区域的峰值点来估计头部位置,然而在拥挤场景下很可能导致定位失败.

因此,最近基于点的方法引起了研究人员的广泛关注. 基于点的方法直接使用点标注作为训练目标,预测出每个人的坐标. 文献[10]预设一组坐标点,使用匈牙利算法将坐标点查询与真值点一对一匹配. 文献[11]则使用可训练的实例查询代替预设的锚点. 由于预测点与目标点的匹配过程不稳定,文献[12]引入辅助点指导策略,根据每个真值点的位置增加辅助正点和辅助负点来指导网络学习,为预测点的选择和优化提供有效的指导. 文献[13]提出可变的分解点查询方法以替代固定坐标点查询. 当人群密度较高时,每个查询点可动态分解为四个新点,从而实现稀疏和密集区域的动态处理.

如图1所示,由于不同场景在光照、尺度、拍摄角度和人群密度等方面存在差异,训练好的计数模型迁移至其他场景时,其计数性能通常会因为数据分布差异而显著下降. 为解决这一问题,研究者们提出了多种基于域泛化的方法,在目标场景不可知的情况下增强跨场景计数性能. 文献[14]提出了一种基于动态子域划分的域泛化框架,通过设计域不变记忆模块和域特定记忆模块,实现了图像特征的重编码与分离,从而有效区分了域不变信息和域特定信息. 文献[15]进一步引入了记忆库机制,结合内容错误掩码和注意力一致性损失来重建域不变特征,同时创新性地将图像划分为网格单元,利用分块分类作为辅助任务,有效缓解了标签模糊性问题. 尽管这些方法显著提升了模型的跨场景适应能力,但由于其均基于密度图估计框架,无法提供精确的行人位置信息,难以满足实际场景中高精度人流监控的需求.

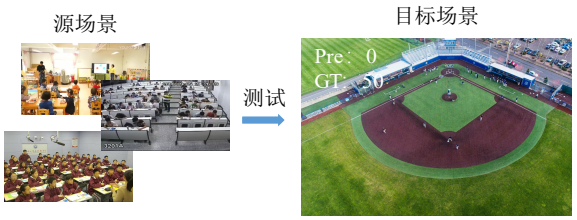


图1 跨场景人群计数示例图

为了提升人群计数模型在不同场景下的适应性,研究人员不断扩展和优化人群计数数据集. UCSD[16]数据集和Mall[17]数据集是早期的代表性数据集,主要由监控视角下的单一场景图像组成,具有较低的图像分辨率和人群密度. 随后发布的WorldExpo'10[18]数据集首次对不同场景进行了细分,但其各场景之间的差异较小,人群密度较低. 近年来,数据集的发展方向逐步转向更高密度和更高分辨率的互联网采集图像,旨在评估计数网络在高密度场景下的表现. 代表性的数据集包括NWPU-Crowd[19]、UCF-QNRF[9]、ShanghaiTech Part A[3]、JHU-CROWD++[20]. 文献[21]通过电子游戏GTA5构建了一个大型合成数据集GCC. 该数据集可直接生成精确的人物头部位置,从而规避了手工标注过程中常见的标注缺失与位置偏移问题,同时显著降低了高昂的人力成本. 尽管现有数据集已覆盖部分应用场景,但针对教学领域中的多场景复杂环境,仍难以全面满足模型泛化能力与适应性的高标准要求.

针对现有方法在跨场景人群计数中存在的局限性,本文提出了一种基于场景感知元权重网络的预测点重新加权方法. 首先,该方法针对模型在未知场景下因分布差异导致的性能下降问题,设计了一种元权重网络架构,通过场景任务感知分支为不同场景构建专属的加权函数,降低标注噪声的影响,显著提升了模型的跨场景泛化能力. 其次,该方法针对传统密度图方法定位精度不足的问题,直接预测个体坐标,并结合元数据指导的预测点加权策略,实现了更精确的人群定位. 此外,为弥补现有数据集在校园场景覆盖的不足,该方法构建了面向教学环境的多场景人群计数数据集(Multi-Scenc Crowd counting dataset,MS-Crowd),为跨场

景研究提供了更全面的评估基准.

2　本文方法

2.1　点预测人群计数框架

本文基于元权重网络的跨场景点预测人群计数模型如图2所示. 该模型以点对点人群计数网络P2P-Net[10]作为基线结构,通过对图像中每一个个体的精确定位后进行汇总,实现总人数的估计. P2P-Net是首个完全基于点注释的端到端人群计数框架. 该方法设计简洁、思路清晰,在提升计数精度的同时避免了密度图方法中繁琐的中间表示构造及复杂的后处理定位流程,因而已成为当前人群计数任务中广泛采用的重要基准方法. 其结构包括以下几个部分:以VGG-16_bn[22]为基础的特征编码主干网络、颈部网络、并行的分类分支与回归分支,以及用于预测点与真实点一对一匹配的匈牙利算法.

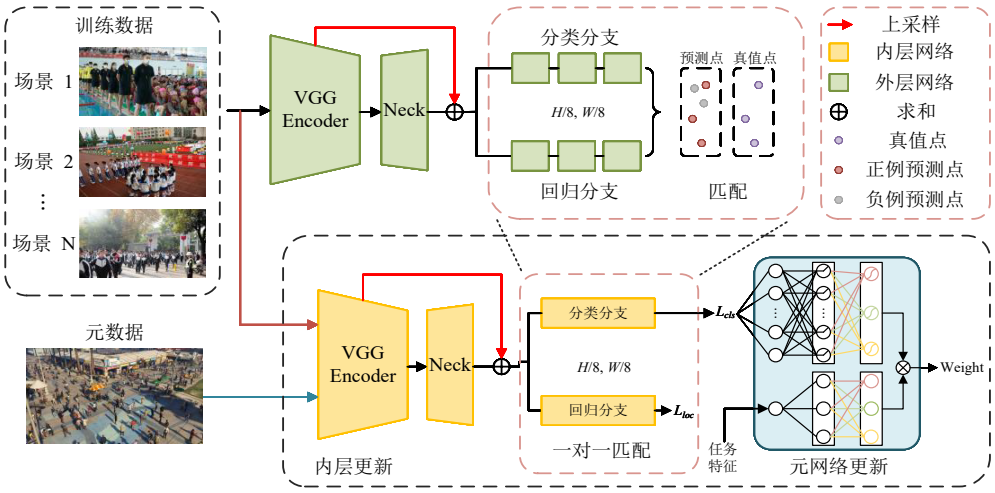


图2　基于元权重网络的跨场景点预测人群计数模型

给定一张人群数量为N的图像$I\in\mathbb{R}^{H\times W\times 3}$,$H$和$W$分别表示图像的高和宽,每个人头中心点位置使用$p_i=(x_i,y_i),i\in\{1,2,\cdots,N\}$表示. VGG-16_bn的前13个卷积层作为主干网络用于提取图像特征表示,第一阶段空间分辨率由$H\times W$降为$H/2\times W/2$,第二阶段空间分辨率降为$H/4\times W/4$,第三阶段空间分辨率降为$H/8\times W/8$,第四阶段空间分辨率降为$H/16\times W/16$. 在颈部网络阶段,第四阶段输出的特征图经过一个卷积层后,使用最近邻插值将其空间分辨率由$H/16\times W/16$上采样为$H/8\times W/8$,然后与主干网络中第三阶段卷积后的特征图逐元素相加,以捕获不同尺度的信息. 再经过一个卷积层得到特征图F以减少上采样过程中可能产生的混叠效应.

分类分支和回归分支同时接收特征提取网络输出的特征图F,并分别预测目标个体的位置和置信度分数. 由于特征图F的空间尺寸缩小为输入图像的1/8,因此特征图F中的每个特征向量对应于输入图像上的8×8像素区域. 回归分支通过预设的锚点,将预测目标个体的位置转换为相对于锚点的偏移量$(\Delta_{jx}^k,\Delta_{jy}^k)$,目标个体的位置表示为

$$\hat{x}_j=x_k+\gamma\Delta_{jx}^k,\hat{y}_j=y_k+\gamma\Delta_{jy}^k \tag{1}$$

其中,$(\hat{x}_j,\hat{y}_j)$表示回归分支的预测点坐标;(x_k,y_k)表示锚点的参考坐标;γ表示缩放因子,用于调整偏移量的范围,以保证预测的稳定性和收敛性. 锚点数量设置为一个特征向量对应4个锚点,锚点在输入图像空间中均匀分布.

分类分支和回归分支采用结构相同的网络结构,每个分支由三个卷积层组成,前两层均包括ReLU(Rectified Linear Unit)激活函数以增强非线性表达能力,输入通道数和输出通道数保持不变. 第三个卷积层用于通道降维,以匹配锚点数量. 分类分支使用Softmax函数将预测值映射至(0,1)区间生成预测点置信度分数,以衡量该位置属于目标个体类别的概率.

2.2　预测点的一对一匹配

在模型训练过程中,需要将真值点与预测点进行一对一匹配,以解决真值点应当与哪个预测点对应的问题. 当真值点附近出现距离相同的两个预测点时,置信度更高的预测点应当被视为正例预测点并逐渐实现更高的置信度;而另一个预测点将被视为负例预测点并逐渐降低其置信度. 同理,当真值点附近出现置信度相同的两个预测点时,距离更近的预测点应当被视为正例,并逐渐提高其定位精度. 为实现这一匹配过程,本文采用匈牙利算法[23]作为一对一匹配策略,未能匹配上的预测点将被视为背景类.

真值点集P与预测点集$\hat{P}$的一对一匹配表示为

$\boldsymbol{H}(P,\hat{P},Q)$,$Q$是维度为$N\times M$的预测点与真值点成对匹配成本矩阵,其中包括了点与点之间的距离和置信度分数:

$$Q\left(P,\hat{P}\right)=\left(\tau\left\|p_i-\hat{p}_j\right\|_2-\hat{c}_j\right)_{i\in N,j\in M} \tag{2}$$

其中,τ是用于平衡像素欧式距离和置信度分数影响的权重;$\|\cdot\|_2$表示第i个真值点p_i与第j个预测点$\hat{p}_j$之间的欧几里得距离;$\hat{c}_j$表示第j个预测点$\hat{p}_j$的置信度分数.

2.3 损失函数

在一对一匹配之后,每个预测点将与真值点唯一匹配,其他预测点被归类为背景,此时点定位回归分支和分类分支将被同时优化. 点定位回归分支使用欧几里得损失更新:

$$L_{\text{loc}}=\frac{1}{N}\sum_{i=1}^{N}\left\|p_i-\hat{p}_i\right\|_2^2 \tag{3}$$

其中,N表示真值点数量;$\|\cdot\|_2$表示真值点与匹配的预测点之间的欧几里得距离. 分类分支使用交叉熵损失更新:

$$L_{\text{cls}}=-\frac{1}{M}\left\{\sum_{i}^{N}\ln\hat{c}_i+\lambda_1\sum_{i=N+1}^{M}\ln\left(1-\hat{c}_i\right)\right\} \tag{4}$$

其中,M表示预测点的总数;λ_1是用于平衡正类和背景类的权重参数;当$i=1,2,\cdots,N$时$\hat{c}_i$表示第i个正类预测置信度分数;当$i=N+1,N+2,\cdots,M$时,$\hat{c}_i$表示第i个背景类的置信度分数.

损失函数L_{cro}是分类损失和回归损失的组合:

$$L_{\text{cro}}=L_{\text{cls}}+\lambda_2 L_{\text{loc}} \tag{5}$$

其中,λ_2是平衡分类损失和回归损失的权重.

2.4 元学习过程

本文将跨场景学习任务数据集设定为$D=\left\{D_a^{\text{train}},D_b^{\text{test}}\right\}(a=1,2,\cdots,K,b=1,2,\cdots,Y)$,其中$D_a^{\text{train}}$表示第$a$个训练场景,$D_b^{\text{test}}$表示第$b$个测试场景. 在域泛化设定中,训练场景$D_a^{\text{train}}$被称为源域,测试场景$D_b^{\text{test}}$被称为目标域. 由于在训练过程中测试场景$D_b^{\text{test}}$将是未知的,因此所有训练场景$D_a^{\text{train}}$与测试场景$D_b^{\text{test}}$将被指定为互不重叠的不同场景. 在测试期间,训练得到的人群计数模型将直接应用于未知的目标场景. 然而,由光照、尺度、拍摄角度和人群密度等因素引起的数据分布差异,以及标注缺失和位置偏移等噪声标注问题,均会导致模型的跨场景能力下降. 为了解决这一问题,本文使用重加权策略,构建了预测点训练损失与权重之间的自适应函数映射,以减轻手工标注产生的错标、漏标和标注偏移等标注噪声对训练产生的影响. 具体而言,将每个场景视为一个独立的学习任务,依据其内在偏差信息生成相应的样本权重. 该映射函数由场景感知元权重网络实现,并采用元学习[24]框架对其进行优化,该网络的加权分支以预测点损失值作为输入,施加权重:

$$V\left(L_i^{\text{train}}(w);\Theta\right)\in[0,1]^{K\times C} \tag{6}$$

其中,$V(\cdot)$表示元权重网络;w表示人群计数模型参数;Θ表示元权重网络的参数. 其输出包含$K\times C$个输出权重,对应于K个训练场景的图像施加K种C个不同的权重方案. $L_i^{\text{train}}(w)$表示第i张图像预测点的分类分支中,C个预测点分别计算得到的未归一化损失,用于后续加权:

$$L_i^{\text{train}}(w)=\left[\ell_{i,1},\ell_{i,2},\cdots,\ell_{i,C}\right]\in\mathbb{R}^C \tag{7}$$

为提升加权策略对不同训练任务场景的适应能力,本文进一步引入场景感知分支,通过将图像所属的训练场景作为先验信息引导加权策略的选择. 对于场景感知分支,令$S_i(i=1,2,\cdots,I)$表示第i张图像所属的训练场景,则场景感知分支可表示为$\mathcal{S}\left(S_i;\Omega\right)\in\{0,1\}^K$,其中$S_i$作为输入用以表征任务层面的特征信息. 该分支包含一个具有K个节点的隐藏层,其中每个节点对应于一个任务场景,组成场景集合$\Omega=\{\mu_k\}_{k=1}^K$. 输出为一个K维 one-hot 向量,用于表示当前图像所属的场景类别. 其中,值为1的元素位于第k个维度,对应于与输入S_i距离最近的场景μ_k. 对于已做出场景划分的数据集,本文直接按照图像所属场景S_i作为μ_k. 而对于未做出场景划分的其他数据集,本文令K=3,根据按升序排列的$\{\mu_k\}_{k=1}^K$将场景划分为小型规模、中等规模和大型规模三类. 通过该结构,可以实现对不同任务规模的有效区分,从而为加权策略提供场景感知能力. 式(6)所表示的元权重网络加权分支输出一个$K\times C$的权重矩阵,场景感知分支则根据图像所属训练场景S_i输出一个K维 one-hot 向量用于选择与该图像最匹配的一组加权向量. 最终,权重矩阵与场景感知向量通过矩阵乘操作,实现了针对特定场景的权重选择机制:

$$\mathcal{V}\left(L_i^{\text{train}}(w),S_i;\Theta,\Omega\right)=\mathcal{S}\left(S_i;\Omega\right)^{\mathrm{T}}\cdot V\left(L_i^{\text{train}}(w);\Theta\right) \tag{8}$$

通过调用高层任务特征信息,场景感知元权重网络能够积累具有相似偏差的训练任务,灵活调整损失函数的贡献,从而优化模型在不同场景下的计数表现.

场景感知元权重网络的最优参数可以通过以下双层优化问题来计算:

$$\{\Theta^*,\Omega^*\}=\arg\min_{\Theta,\Omega}\sum_{i=1}^{I}L_i^{\text{meta}}\left(w^*\left(\Theta,\Omega\right)\right) \tag{9}$$

$$w^*(\Theta,\Omega)=\arg\min_{w}\sum_{i=1}^{I}\mathcal{V}\left(L_i^{\text{train}}(w),S_i;\Theta,\Omega\right)L_i^{\text{train}}(w) \tag{10}$$

其中,$\{\Theta^*,\Omega^*\}$表示场景感知元权重网络的最优参数;

$L_i^{\text{meta}}\left(w^*(\Theta,\Omega)\right)$表示元训练损失. 求解场景参数$\Omega$具有较高的计算复杂度,难以设计高效的算法求其全局最优解,因此本文首先预设一个合理的Ω^*,在固定Ω^*的基础上优化其他参数. 为简化符号表示,后续省略场景参数Ω表示. 权重网络的参数将通过元学习[25]的方式自适应地优化. 元权重网络被设计为一个具有两个隐藏层的多层感知机(MultiLayer Perceptron,MLP)网络. 每个隐藏层包含1 000个隐藏结点,并使用ReLU激活函数以增强非线性表达能力. 输入结点数和输出结点数均定义为100,该数值依据训练集中图像平均人数的统计结果,并结合实际任务需求考虑,在保证覆盖大部分样本的同时保留一定冗余性. 输出结点使用Sigmoid激活函数使权重保持在(0,1)范围内.

为了使权重网络学习到干净无偏的元知识,需要从一个没有标注噪声的数据集$\left\{(x_i^{\text{meta}},y_i^{\text{meta}})\right\}_{i=1}^{m}$中获取元数据[24],其中$m$表示每个批次的样本数量. 考虑到手工标注的数据集将不可避免地产生标注噪声,本文选择GCC合成数据集[21]作为元数据来源计算元训练损失$L_i^{\text{meta}}(w)=l\left(y_i^{(\text{meta})},f(x_i^{(\text{meta})},w)\right)$.

更新内层元计数网络参数. 场景感知元权重网络学习过程与Model-Agnostic Meta-Learning (MAML)[26]元学习算法的优化过程相似,均采用内外双层优化框架. 首先是内层优化阶段,对计数网络进行复制得到一份元计数网络. 将训练数据作为此元计数网络输入,通过其分类分支得到预测点的分类损失. 分类损失与图像所属的场景共同构成元权重网络的输入,生成预测点损失的权重系数. 基于加权后的损失,计算其在元计数网络上的梯度,执行梯度下降算法更新元计数网络的参数:

$$\hat{w}^{(t+1)}(\Theta)=w^{(t)}-\alpha\sum_{i=1}^{n}\mathcal{V}\left(L_i^{\text{train}}\left(w^{(t)}\right),S_i;\ \Theta\right)\nabla_w L_i^{\text{train}}(w)\Big|_{w^{(t)}} \tag{11}$$

其中,$\hat{w}$表示元计数网络参数;α表示元计数网络更新步长.

更新元权重网络的参数. 随后,元数据作为元计数网络的输入,元权重网络的参数的更新过程表示为

$$\Theta^{(t+1)}=\Theta^{(t)}-\beta\frac{1}{m}\sum_{i=1}^{m}\nabla_\Theta L_i^{\text{meta}}\left(\hat{w}^{(t+1)}(\Theta)\right)\Big|_{\Theta^{(t)}} \tag{12}$$

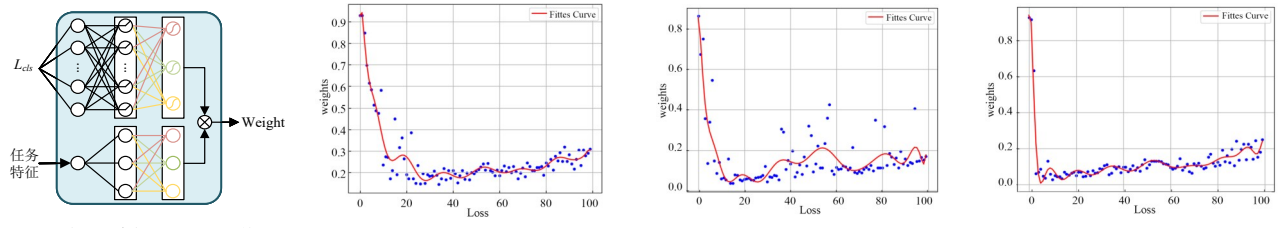
其中,β表示元权重网络更新步长.

更新外层计数网络参数. 最后,利用参数更新为$\Theta^{(t+1)}$后的元权重网络更新外层计数网络:

$$w^{(t+1)}=w^{(t)}-\alpha\sum_{i=1}^{n}\mathcal{V}\left(L_i^{\text{train}}(w^{(t)}),S_i;\ \Theta^{(t+1)}\right)\nabla_w L_i^{\text{train}}(w)\Big|_{w^{(t)}} \tag{13}$$

外层计数网络与内层计数网络共享相同的源场景数据作为输入,在此基础上,外层计数网络的损失经过更新后的元权重网络进行加权调整,从而优化外层计数网络的参数更新策略.

如图3(a)所示,场景感知元权重网络整体由两个分支组成:加权分支以预测点的训练损失为输入,输出对应的$K\times C$维加权权重矩阵;场景分支以样本所属训练场景为输入,输出其对应的K维one-hot向量用于选择权重方案. 两分支共同构成一个基于任务划分的动态加权机制:通过将样本分配至其对应的任务组,并从多个预定义的加权策略中选择最适合其任务的权重向量,从而实现按场景分配损失权重的目标. 图3(b)~图3(d)展示了不同场景下预测点损失权重与原始损失之间的关系. 横轴为按损失值升序排序后的预测点编号,纵轴为对应权重大小.



(a) 场景感知元权重网络　(b) 运动场场景样本权重直方图　(c) 放学场景样本权重直方图　(d) 游泳池场景样本权重直方图

图3　场景感知元权重网络结构及权重可视化

结果表明,模型在不同场景下均表现出对高损失样本赋予更高权重的趋势,这表明模型能够在训练过程中学习到损失结构中蕴含的难易样本信息,从而对预测不确定区域赋予更大关注度,实现动态加权学习. 网络通过训练样本的损失分布自主学习损失与权重之间的映射关系,对损失较小的预测点赋予更高的权重,以强化模型对准确预测样本的学习,同时逐步提升对损失较高预测点的权重,使模型能够更关注那些难以预测的样本. 这一策略不仅有助于模型在早期阶段充分利用高质量样本进行稳定学习,还能在后续训练过程中逐步增强对复杂场景的适应能力. 此外,由于元权重网络能够根据不同场景自适应调整损失权重分布,其生成的权重方案可以有效避免单一加权策略可能带来的过拟合问题,从而进一步提升模型在未见场景下

的泛化能力.

2.5 校园多场景人群计数数据集 MS-Crowd

现有的人群计数数据集主要通过两种方式构建:其一是在固定场景下采集不同时间段的行人图像,例如 UCSD[16]和 Mall[17]数据集分别基于校园人行道和购物中心的监控视频;其二是以高密度人群场景为目标,通过关键词(如"春运""演唱会""集会"等)在互联网上搜索筛选相关图像,典型数据集包括 ShanghaiTech Part A[3]、NWPU-Crowd[19]和 JHU-CROWD++[20]等. 然而,这些数据集对于教学领域场景上覆盖不足. 教学领域的环境具有独特的挑战性,例如多样化的教室布局、动态变化的学生密度及场景类型的多样性(如教室、运动场、图书馆、体育馆等),这些因素对人群计数模型的跨场景泛化能力提出了更高要求. 目前尚无专门针对校园场景的人群计数数据集,为此,本文整合监控视频、视频网站及互联网搜索引擎数据构建了专注于校园场景的多场景人群计数数据集 MS-Crowd,为教学环境下的人群计数任务提供新的解决思路,并为智慧教育领域的研究提供参考和借鉴. 表 1 详细对比了 MS-Crowd 与其他公开数据集的关键特性.

MS-Crowd 数据集包含 11 种校园室内外场景,共 5 408 张图像、299 030 个头部标注,人数范围为 1~2 209. 室内场景包括大学教室、中学教室、小学教室、幼儿园教室、食堂、体育馆、图书馆和演讲厅,共 8 种;室外场景包括运动场、校门和游泳池,共 3 种. 图 4 展示了 MS-Crowd 数据集教学场景的示例图,第一行从左到右依次为幼儿园教室、小学教室、中学教室和大学教室;第二行从左到右依次为校门、运动场、游泳池和另一运动场;第三行从左到右依次为演讲厅、图书馆、体育馆和食堂.

表 1 主流人群计数数据集与 MS-Crowd 数据集计数信息

数据集	图像数量/张	平均分辨率/像素	标注数量/人	最小人数/人	平均人数/人	最大人数/人	场景划分	场景类型
UCSD[16]	2 000	158 × 238	49 885	11	25	46	×	校内同一个街道
Mall[17]	2 000	320 × 240	62 325	13	—	53	×	同一个商场
ShanghaiTech[3]	1 198	598 × 868	330 165	9	275	3 139	×	上海街道和互联网图像
Crowd_Surv[27]	13 945	840 × 1 342	386 513	2	35	1 420	×	互联网图像
UCF_CC_50[28]	50	2 101 × 2 888	63 974	94	1 279	4 543	×	互联网图像
UCF-QNRF[9]	1 535	2 013 × 2 902	1 251 642	49	815	12 865	×	互联网图像
NWPU-Crowd[19]	5 109	2 190 × 3 209	2 133 375	0	418	20 033	×	互联网图像
WorldExpo'10[18]	3 980	576 × 720	199 923	1	50	253	√	世博会监控
JHU-CROWD++[20]	4 372	910 × 1 430	1 515 005	0	346	25 791	√	互联网图像
MS-Crowd	5 408	1 326 × 870	299 030	1	55	2 209	√	教学环境多场景

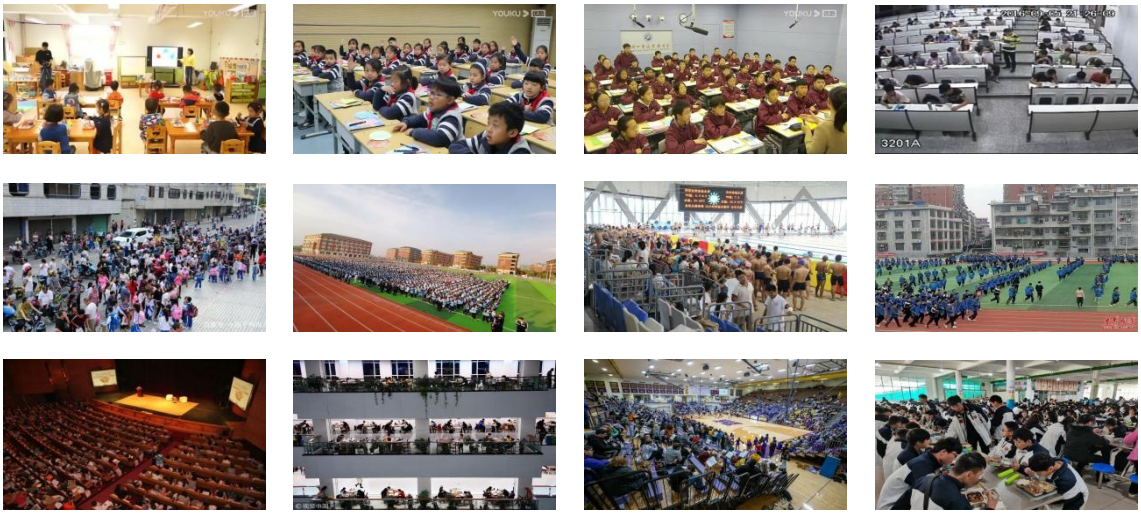


图 4 校园多场景人群统计数据集 MS-Crowd 示例图

为了更合理地评估计数模型的跨场景性能,根据场景特征和人群密度将11种场景分为三组:第一组(S组)包括幼儿园教室、小学教室、中学教室和大学教室;第二组(M组)包括食堂、体育馆、图书馆和演讲厅;第三组(L组)包括校门、游泳池和运动场. 为了保持数据来源的多样性,本文通过以下三个途径获取图像数据:监控视频、视频网站和互联网搜索引擎. 其中监控视频来源于重庆邮电大学教室,涵盖不同视角的中小型教室和大型阶梯教室. 在人工筛选过程中,重点关注高密度场景的计数难点,最终保留了797张包含20人以上的图像. 中学教室、小学教室和幼儿园教室的图像通过视频网站的公开课教学视频获取,筛选后得到2 860张适合统计人数的关键帧图像. 图5展示了场景标签的分布情况.

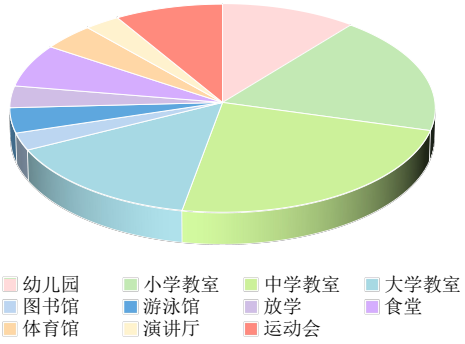


图5　MS-Crowd数据集场景标签分布情况

本文还扩展了教学场景的覆盖范围,包括图书馆、游泳馆、食堂、体育馆、演讲厅室内场景共计889张图像,以及运动场、游泳池和校门附近室外场景共计856张图像. 这些数据来源于互联网搜索引擎Bing、百度和谷歌,用于搜索的关键字确保了场景类型、地理位置、分辨率和密度等方面的多样性. 在数据预处理阶段,删除了重复及像素过低的图像. 最终,每个场景的图像按7:3的比例随机划分为训练集和测试集. 表2列出了所构建数据集的详细信息.

3　实验结果与分析

本文使用当前主流的开源方法评估MS-Crowd数据集,所选方法包括MCNN[3]、CSRNet[29]、SFCN[19]、MAN[30]、DCCUS[14]以及MPCount[15]. 评估指标使用人群计数领域广泛使用的平均绝对误差(Mean Absolute Error,MAE)和均方误差(Mean Squared Error,MSE). 平均绝对误差定义为

$$\mathrm{MAE}=\frac{1}{I_{\mathrm{test}}}\sum_{i=1}^{I_{\mathrm{test}}}\left|\hat{C}_i-C_i\right| \tag{14}$$

其中,I_{test}为测试样本的数量;$\hat{C}_i$为模型预测第i张图像的人群总数;C_i为对应的真实人数. MAE用于评估计数模型预测值与真实值之间的接近程度. MSE则注重于预测稳定性,定义为

$$\mathrm{MSE}=\sqrt{\frac{1}{I_{\mathrm{test}}}\sum_{i=1}^{I_{\mathrm{test}}}\left|\hat{C}_i-C_i\right|^2} \tag{15}$$

由于平方项的存在,MSE对较大的误差更为敏感,因此能够突出模型在某些场景下可能出现的大幅度预测偏差.

表2　多场景人群计数数据集MS-Crowd计数信息

分组	场景	平均分辨率/像素	图像数量/张	平均人数/人	最少人数/人	最多人数/人	总人数/人
S	幼儿园教室	1 324×744	585	12.6	1	33	7 375
	小学教室	1 920×1 080	1 000	33.8	7	73	33 824
	中学教室	1 920×1 080	1 275	34.0	5	67	43 384
	大学教室	704×576	797	48.9	20	106	39 005
M	食堂	1 173×806	364	68.5	20	367	24 924
	体育馆	1 478×973	227	119.0	20	805	27 021
	图书馆	1 088×746	145	39.1	20	159	5 679
	演讲厅	1 178×746	153	122.3	22	536	18 707
L	校门	1 053×847	183	54.7	20	264	10 012
	游泳池	1 190×888	210	45.1	20	512	9 491
	运动场	1 561×1 083	469	169.7	20	2 209	79 608

3.1　实验设置

硬件平台为单张RTX 3090显卡,开发环境为Ubuntu20.04操作系统、Pytorch 1.10、Python 3.6. 对输入图像使用缩放因子为[0.7,1.3]的随机缩放并保证图像的短边大于128像素. 将图像随机裁剪为128像素×128像素大小的图像块,并采用左右翻转作为训练时的数据增强策略. 使用Adam作为优化器,其初始学习率为4×10^{-5},主干网络使用ImageNet预训练权重,学习率为1×10^{-5};batch-size设为32. 回归分支中γ为100,匹配期间的权重项τ设为5×10^{-2}. 损失函数中,λ_1设为0.5,λ_2设为2×10^{-4}. 对于元权重网络的训练,同样采用Adam优化器,初始学习率设置为1×10^{-3},batch-size设置为1.

3.2　与主流算法的对比

在MS-Crowd数据集上,本文的跨场景评估设定为将一组场景作为训练场景(源场景),另外两组场景作为目标场景进行测试,以验证模型在不同场景间的泛化能力. 对于其他对比方法的实验结果,本文统一采用论文中作者提供的公开代码与原始参数设置,将训练集和测试集设置为不同场景以进行跨场景性能测试,确保实验条件的一致性与公平性. 表3展示了本文方法与现有人群计数算法的指标对比结果. 其中,"DG"表示是否属于跨场景域泛化对比方法,加粗数据表示

最优结果，下划线数据表示次优结果．表中前四行为未针对跨场景或跨域任务专门设计的经典人群计数方法，关注的是相同数据分布下的密度图预测结果：MCNN[3]方法使用了不同尺寸卷积核的多列卷积架构；CSRNet[29]方法改进了MCNN方法，将空洞卷积层作为密度生成器以扩大感受野并在不丢失分辨率的情况下提取更深层的特征；SFCN[19]设计了一个空间全卷积网络来生成密度图；MAN[30]则使用科学区域注意力来针对尺度差异和标注偏移问题．针对跨场景和跨域任务设计的最新方法为DCCUS[14]和MPCount[15]：DCCUS结合了元学习域泛化（Meta-Learning for Domain Generalization，MLDG）框架来增强模型的泛化能力；MPCount使用记忆库、划分网格预测前后景和注意力一致性损失来重建域不变特征．

表3 校园多场景MS-Crowd数据集计数性能结果

源场景→目标场景		S→M		S→L		M→L		M→S		L→M		L→S	
方法	DG	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓
MCNN[3]	×	50.4	97.4	87.9	218.0	66.1	174.2	7.0	9.4	48.0	81.9	15.4	21.9
CSRNet[29]	×	41.1	85.5	79.6	198.0	60.7	145.3	7.9	11.6	26.4	42.0	11.0	15.5
SFCN[19]	×	49.1	100.6	81.2	207.8	37.2	115.5	5.3	7.4	14.7	23.3	6.6	8.8
MAN[30]	×	43.7	78.6	87.4	206.3	36.3	100.3	**3.8**	6.1	11.6	18.7	4.4	11.6
DCCUS[14]	√	43.9	95.7	85.8	215.4	36.6	102.9	**3.8**	**5.7**	14.2	22.8	6.6	9.3
MPCount[15]	√	32.0	89.6	78.6	189.4	29.4	76.1	5.0	7.4	**10.0**	**15.6**	**3.9**	**5.9**
本文方法	√	**25.7**	**43.1**	**65.6**	**159.3**	**23.9**	**61.4**	4.4	5.8	11.9	18.1	5.6	7.8

当S组作为源场景训练，并在M组和L组作为目标场景进行测试时，本文方法在MAE和MSE两项指标上均优于所有对比方法．相较于MPCount[15]，在S→M任务上MAE降低19.7%，S→L任务上MAE降低16.5%，表明计数精度得到显著提升，验证了本文方法在跨场景任务中的有效性．当M组场景作为源场景训练，并在S组和L组作为目标场景进行测试时，本文方法在M→L任务上的MAE和MSE指标均优于所有对比方法，相较于MPCount[15]，MAE降低18.7%．当L组场景作为源场景训练，并在S组和M组作为目标场景进行测试时，本文方法取得次优结果．这可能是由于L组为室外场景，而S组和M组包含大量室内场景，场景转换导致头部尺寸变化较大，部分头部被分类为背景．相比之下，MPCount[15]采用PCM（Patch-wise Classification Map）辅助任务，将图像分为更小的块，并分别判断是否包含行人头部，从而减少像素级密度图的不确定性．本文方法尽管在部分实验配置中取得次优结果，但能够提供更精确的行人坐标信息，这在实际应用中的人流监控任务具有更大的优势．

为进一步验证方法在其他场景的有效性，本文在户外场景公开数据集ShanghaiTech上进行了跨场景实验．尽管当前有多个公开数据集可供使用，但这些数据集场景划分单一或未对场景进行细致区分，这限制了其在跨场景人群统计任务中的评估能力，因此未被纳入本次实验．ShanghaiTech数据集由两个部分组成：SHA和SHB，如图6所示．

(*a*) SHA示例图 (*b*) SHB示例图

图6 ShanghaiTech数据集示例图

SHA部分由互联网收集的密集人群图像构成,这些图像的人群密度较高,且背景较为复杂;而SHB部分则来源于上海商业街道的监控视角图像,具有较低的人群密度和相对简单的背景. SHA和SHB的场景和密度差异符合本文跨场景任务. 设SHA包含300张训练图像和182张测试图像,SHB包含400张训练图像和316张测试图像.

本文分别将SHA和SHB作为训练场景(源场景),另一部分作为测试场景,实验数据均来自论文的公开结果. 由于该数据集未提供场景标签,本文对所有训练样本的人数进行标准K均值聚类,并按照升序排列聚类中心,$\Omega=\{\mu_k\}_{k=1}^{K}$令$K$=3以获得三种不同密度的训练任务. 表4展示了本文方法与其他方法的对比结果,其中加粗数据表示最优结果,下划线数据表示次优结果. 在对比方法中额外引入了域适应方法作为参考. 域适应方法通常会使用一部分目标场景的数据,这在模型部署中可能不切实际. 在所有对比方法中,本文方法获得了最优或者次优的结果,且超越了所有基于域适应的方法,再次验证了其在跨场景任务中的有效性. 本文方法在SHB→SHA任务上,与DCCUS[14]相比,MAE降低了10.3%;与MPCount[15]相比,在MSE降低了3%. 尽管本文方法相较于MPCount在MAE上上升了9.7%,但相较于基于密度图的方法,能够在MAE上相较MPCount[15]误差上升了9.7%,但本文方法相较于基于密度图的方法能够额外提供每个行人的位置信息,在实际部署时可实现更精确的人流监控,因此在高精度应用场景中更具优势.

表4　ShanghaiTech数据集计数性能结果

源域→目标域			SHB→SHA		SHA→SHB	
方法	域适应	域泛化	MAE↓	MSE↓	MAE↓	MSE↓
MCNN[3]	×	×	221.4	357.8	85.2	142.3
DSSINet[31]	×	×	148.9	273.9	21.7	37.6
DMCount[32]	×	×	143.9	239.6	23.1	34.9
MAN[30]	×	×	133.6	255.6	22.1	32.8
FGFD[33]	√	×	120.3	202.6	12.7	23.3
RBT[34]	√	×	112.2	218.2	13.4	29.3
SE+FD[35]	√	×	129.3	187.6	16.9	24.7
C^2MoT[36]	√	×	120.7	192.0	12.4	21.1
DG-MAN[37]	×	√	130.7	225.1	17.3	28.7
DCCUS[14]	×	√	121.8	203.1	12.6	24.6
MPCount[15]	×	√	**99.6**	<u>182.9</u>	<u>11.4</u>	**19.7**
本文方法	×	√	<u>109.3</u>	**177.3**	**10.1**	<u>19.8</u>

3.3　可视化结果分析

如图7所示,本节选取表3中S组到L组的跨场景实验结果进行分析. S组为大学、中学、小学和幼儿园教室的室内场景,具有较为复杂的背景、较大的人头尺度、较小的人群密度以及相对静态的环境特征;L组为学校体育场、运动场、足球场和篮球场等室外场景,具有更高的人群密度、更强的光照变化、更丰富的尺度变化以及更加开放的场景布局. 图7(a)为最新方法基于密度图的MPCount[15]的可视化预测结果,图7(b)为基线模型的可视化预测结果,图7(c)为本文方法的可视化预测结果. 为了更直观地展示三者差异,所有可视化结果均经过放大处理. 从实验可视化结果可以看出,由于训练集中缺少俯瞰视角、暗光环境和逆光环境,基线模型在测试集中面对模糊人群时难以做出准确判断,导致对人群数量的预测精度大幅下降. 图7(b)和图7(c)中,第一行测试场景为游泳池,图像中个体佩戴泳帽和泳镜使基线模型无法正确识别和计数,而本文方法在未曾见过该场景的情况下仍能有效识别并计数. 第二行测试场景为远处的暗光人群,基线模型仅能识别近处受阳光照射的个体,而本文方法能够精确计数后排暗光区域的人群. 第三行测试场景为俯瞰视角,基线模型仅能识别出少量个体,而本文方法在视角变换的情况下依然保持较高的计数精度. 第四行测试场景为逆光复杂背景,测试图像中的个体像素占比极小,与背景难以区分. 基线模型在此场景下几乎无法识别,而本文方法能够准确识别看台上的观众并完成统计. 图7(b)第二列MPCount[15]只识别出了部分的人群,无法识别泳池中的个体,对于远处的高密度人群预测能力明显不足. 此外,MPCount[15]仅能输出表示人群分布的密度图,无法提供实例级别的精确定位. 由此可见,改进后的方法在远处模糊人群的识别方面表现出显著优势,提升了统计网络在未知场景下的泛化能力,并有效避免了传统密度图方法所存在的定位模糊问题.

3.4　消融实验

本文在MS-Crowd数据集上设计了消融实验来验证所提出的场景感知元加权方法的有效性和可靠性. 实

(*a*) MPCount (*b*) 基线模型 (*c*) 本文方法

图7 MPCount[15]、基线模型和本文方法可视化预测结果对比

验以P2PNet[10]作为基线模型,并在其训练过程中引入场景感知元权重网络方法,保持其他参数不变.对于MS-Crowd数据集以表3的S组到L组评估跨场景计数性能.通过对这两组具有显著场景差异的数据集进行对比,能够更好地评估模型在跨场景任务中的泛化能力和适应性.在MS-Crowd数据集上S组到L组的跨场景元权重网络消融实验结果如表5所示,其中加粗数据为最优结果,"任务场景"表示场景感知元权重网络设定的场景数量.

表5 跨场景元权重网络消融实验结果

方法	任务场景	MAE ↓	MSE ↓
基线模型	0	75.7	176.0
基于元权重网络	1	69.5	164.6
基于场景感知元权重网络	4	**65.6**	**159.3**

实验结果第一行为基线方法结果,第二行表明不使用场景感知分支,仅使用元加权分支,因此所有场景将通过共享同一个元权重网络对损失进行加权.可以看到基于元权重网络的方法在MAE上相比基线方法下降8.2%,MSE则下降了6.5%.实验结果有力地表明了使用元权重网络对损失权重的动态调整策略的有效性.然而,由于不同场景的数据分布存在差异,统一的元权重网络难以充分适应各个场景的偏差.因此,第三种方法为同时使用场景感知分支和元加权分支,每一个场景将独立训练一个专属的元权重网络,MAE相比基线模型下降13.3%,MSE则下降了9.5%.为了验证元数据对于本文方法性能的影响,本节提供了不同数量的元数据下在MS-Crowd数据集上S组到L组的跨场景元数据消融实验结果,如表6中"元数据"所示,其中加粗数据为最优结果.表中"元数据"表示元数据图像数量,即使用GCC合成数据集中的图像数量;"任务场景"表示训练场景数量.S组训练图像数量为2 560张,而GCC合成数据集为每个场景创建了约50张不同人群数量和光照条件的图像共计15 212张.因此,本次消融实验结合GCC数据集的划分情况,使用1%~10%左右的元数据进行实验.实验结果第二种方法的元数据使用了5个场景,每个场景4张图像,共计20张元数据,相比基线方法MAE下降了8.5%.第三种方法将元数据每个场景的图像数量提升至12张,共计60张元数据,MAE相比第二种方法下降了0.7%.第四种方法将元数据场景数量提升为80个,每个场景使用4张图像,共计320张元数据,丰富了场景类型并保持每个场景的元数据量不变,MAE相比第三种方法下降了4.7%.这表明,除了保证元数据的干净无偏性外,场景的多样性对于提升场景感知元权重网络的泛化能力同样具有重要作用.可以看出,随着元数据量和场景多样性的增加,模型的性能逐渐提升,表明元数据在提升模型跨场景适应性方面起到了关键作用.

表6 跨场景元数据消融实验结果

方法	任务场景	元数据	MAE ↓	MSE ↓
基线方法	0	0	75.7	176.0
基于场景感知元权重网络	4	20	69.3	175.5
基于场景感知元权重网络	4	60	68.8	177.1
基于场景感知元权重网络	4	320	**65.6**	**159.3**

为了验证元数据的质量对加权策略学习效果的影响,表7展示了跨场景元数据质量消融实验结果,其中加粗数据为最优结果. 在与保持元数据数量一致的前提下,将原先使用的GCC合成数据替换为训练集中的部分样本作为元数据,并分别以S、M、L为源域,在其余两个目标场景中进行测试. 实验结果表明,当使用GCC作为元数据时,模型在所有设置下均取得了最优性能,验证了高质量、干净且无偏的元数据对学习有效权重分配策略的重要性.

表7　跨场景元数据质量消融实验结果

源→目标	方法	元数据	MAE ↓	MSE ↓	MAE ↓	MSE ↓
S→M/L	本文方法	训练集	30.8	56.1	69.5	168.6
	本文方法	GCC	**25.7**	**43.1**	**65.6**	**159.3**
M→S/L	本文方法	训练集	4.5	6.8	26.6	70.5
	本文方法	GCC	**4.4**	**5.8**	**23.9**	**61.4**
L→S/M	本文方法	训练集	5.7	9.2	14.7	23.4
	本文方法	GCC	**5.6**	**7.8**	**11.9**	**18.1**

4　结论

本文针对现有人群计数模型在跨场景应用时存在的两个关键问题:一是因场景分布差异导致的性能下降问题,二是缺乏精确的行人位置信息,提出了一种基于元学习的场景感知预测点重新加权方法. 针对人群计数模型在未知场景下因分布差异导致的性能下降问题,设计了一种元权重网络架构,通过场景感知任务分支为不同场景构建专属的加权函数. 元网络从元数据中自适应地学习显式加权方案,将每个场景视为一个单独的学习任务,构造一个显式的加权函数,以预测点损失和任务特征作为输入,输出相应的样本权重. 该方法利用不同场景的内在特征实现自适应的加权方案,显著提升了模型的跨场景泛化能力. 其次,针对传统密度图方法的定位模糊问题,采用点预测计数框架直接预测行人坐标,并结合元数据指导的预测点加权策略,实现了更精确的人群定位. 此外,针对现有公开数据集在教育领域场景覆盖上的局限性,本文构建了校园环境下的多场景人群计数数据集MS-Crowd. 通过在MS-Crowd数据集和户外场景公开数据集上进行实验和可视化分析,验证了本文方法在跨场景人群计数任务中的有效性.

参考文献

[1] BAI H Y, MAO J G, GARY CHAN S H. A survey on deep learning-based single image crowd counting: Network design, loss function and supervisory signal[J]. Neurocomputing, 2022, 508: 1-18.

[2] 卢振坤, 刘胜, 钟乐, 等. 人群计数研究综述[J]. 计算机工程与应用, 2022, 58(11): 33-46.
LU Z K, LIU S, ZHONG L, et al. Survey on reaserch of crowd counting[J]. Computer Engineering and Applications, 2022, 58(11): 33-46. (in Chinese)

[3] LIN Z, DAVIS L S. Shape-based human detection and segmentation via hierarchical part-template matching[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(4): 604-618.

[4] 张智, 易华挥, 郑锦. 聚焦小目标的航拍图像目标检测算法[J]. 电子学报, 2023, 51(4): 944-955.
ZHANG Z, YI H H, ZHENG J. Focusing on small objects detector in aerial images[J]. Acta Electronica Sinica, 2023, 51(4): 944-955. (in Chinese)

[5] 钟佳平, 李云松, 谢卫莹, 等. 结合区域引导和双注意力机制的高光谱目标检测判别式学习网络[J]. 电子学报, 2024, 52(5): 1716-1729.
ZHONG J P, LI Y S, XIE W Y, et al. Region-guided and dual attention discriminative learning network for hyperspectral target detection[J]. Acta Electronica Sinica, 2024, 52(5): 1716-1729. (in Chinese)

[6] ZHANG Y Y, ZHOU D S, CHEN S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 589-597.

[7] YANG Y F, LI G R, WU Z, et al. Reverse perspective network for perspective-aware object counting[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 4373-4382.

[8] LIN H, MA Z H, HONG X P, et al. Gramformer: Learning crowd counting via graph-modulated transformer[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(4): 3395-3403.

[9] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]//Computer Vision-ECCV 2018. Cham: Springer, 2018: 544-559.

[10] SONG Q Y, WANG C G, JIANG Z K, et al. Rethinking counting and localization in crowds: A purely point-based framework[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 3345-3354.

[11] LIANG D K, XU W, BAI X. An end-to-end transformer model for crowd localization[C]//Computer Vision-ECCV 2022. Cham: Springer, 2022: 38-54.

[12] CHEN I H, CHEN W T, LIU Y W, et al. Improving point-based crowd counting and Localization based onAuxiliary

point guidance[C]//Computer Vision-ECCV 2024. Cham: Springer, 2025: 428-444.

[13] LIU C X, LU H, CAO Z G, et al. Point-query quadtree for crowd counting, localization, and more[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2024: 1676-1685.

[14] DU Z P, DENG J K, SHI M J. Domain-general crowd counting in unseen scenarios[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(1): 561-570.

[15] PENG Z X, GARY CHAN S H. Single domain generalization for crowd counting[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 28025-28034.

[16] CHAN A B, LIANG Z J, VASCONCELOS N. Privacy preserving crowd monitoring: Counting people without people models or tracking[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2008: 1-7.

[17] CHEN K, LOY C C, GONG S G, et al. Feature mining for localised crowd counting[C]//Proceedings ofthe British Machine Vision Conference 2012. Surrey: British Machine Vision Association, 2012: 21.1-21.11.

[18] ZHANG C, LI H S, WANG X G, et al. Cross-scene crowd counting via deep convolutional neural networks[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 833-841.

[19] 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.

[20] SINDAGI V A, YASARLA R, PATEL V M. JHU-CROWD: Large-scale crowd counting dataset and a benchmark method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2594-2609.

[21] WANG Q, GAO J Y, LIN W, et al. Learning from synthetic data for crowd counting in the wild[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 8190-8199.

[22] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2025-04-10]. https://arXiv.org/abs/1409.1556.

[23] KUHN H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2(1/2): 83-97.

[24] SHU J, YUAN X, MENG D Y, et al. CMW-net: Learning a class-aware sample weighting mapping for robust deep learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 11521-11539.

[25] FRANCESCHI L, FRASCONI P, SALZO S, et al. Bilevel programming for hyperparameter optimization and meta-learning[C]//the 35th International Conference on Machine Learning. Cambridge: PMLR, 2018:1568-1577.

[26] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning, 2017.

[27] YAN Z Y, YUAN Y C, ZUO W M, et al. Perspective-guided convolution networks for crowd counting[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 952-961.

[28] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. New York: ACM, 2013: 2547-2554.

[29] LI Y H, ZHANG X F, CHEN D M. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1091-1100.

[30] LIN H, MA Z H, JI R R, et al. Boosting crowd counting via multifaceted attention[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 19596-19605.

[31] LIU L B, QIU Z L, LI G B, et al. Crowd counting with deep structured scale integration network[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 1774-1783.

[32] WANG B Y, LIU H D, SAMARAS D, et al. Distribution matching for crowd counting[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 1595-1607.

[33] ZHU H L, YUAN J L, YANG Z W, et al. Fine-grained fragment diffusion for cross domain crowd counting[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 5659-5668.

[34] LIU Y T, WANG Z, SHI M J, et al. Towards unsupervised crowd counting via regression-detection bi-knowledge transfer[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 129-137.

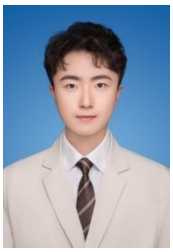
[35] HAN T, GAO J Y, YUAN Y, et al. Focus on semantic consistency for cross-domain crowd understanding[C]//ICASSP 2020 - 2020 IEEE International Conference on Acoustics,

Speech and Signal Processing. Piscataway: IEEE, 2020: 1848-1852.

[36] WU Q Q, WAN J, CHAN A B. Dynamic momentum adaptation for zero-shot cross-domain crowd counting[C]// Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 658-666.

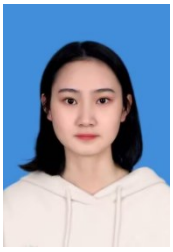
[37] MANSILLA L, ECHEVESTE R, MILONE D H, et al. Domain generalization via gradient surgery[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 6610-6618.

作者简介



徐　昕　男,1999年1月出生于四川省资阳市. 现为重庆邮电大学通信与信息工程学院硕士研究生. 主要研究方向为人群计数、计算机视觉和机器学习.

E-mail: beckhamfrog@foxmail.com



谭卓林　女,1998年1月出生于四川省达州市. 现为重庆邮电大学博士研究生. 主要研究方向为视频分析、图像处理和计算机视觉.

E-mail: tanzhuolin98@gmail.com

高陈强　男,1981年8月出生于重庆市. 现为重庆邮电大学通信与信息工程学院教授、博士生导师. 主要研究方向为图像处理、视频分析和机器学习.

E-mail: gaocq@cqupt.edu.cn

席　跃　男,2004年3月出生于重庆市. 现为澳门大学应用数学本科生. 主要研究方向为算法设计.

E-mail: DC22908@um.edu.mo