计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(9):253−263 [doi: 10.15888/j.cnki.csa.009930] [CSTR: 32024.14.csa.009930]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于循环卷积网络和逆变换贝叶斯损失的室内人群计数①

刘永文, 孙君宇, 凌妙根, 苏　健

(南京信息工程大学 计算机学院, 南京 210044)
通信作者: 凌妙根, E-mail: mgling@nuist.edu.cn

摘　要: 本文关注室内人群计数这一具有挑战性的任务. 在室内场景中, 人们经常聚集在一起并在有限空间内执行相似的任务. 因此, 室内人群的大多数行为都非常相似. 获取全局感受野并找出室内人群特征中的相似性是很重要的. 本文设计了一种循环卷积网络, 该网络结合了卷积神经网络和 Transformer 的优势, 以获取人群特征的局部和全局相关性. 与基于 Transformer 的方法相比, 采用了更简单且高效的循环卷积模块. 此外, 提出了一种逆变换贝叶斯损失函数, 该函数适用于具有大尺度变化的稀疏和拥挤的室内场景. 最后, 为了减轻标注偏差问题的影响, 提出了一种标签扩散策略来扩大标注区域, 假设每个原始标注点的相邻像素也有一定概率成为头部中心. 在 Class A、Class B、Canteen 和 Mall 数据集上与次优方法相比, *MAE*/*RMSE* 分别提高了 4.1%/4.4%、5.8%/8.0%、3.9%/1.6% 和 3.9%/1.6%.

关键词: 循环卷积; 贝叶斯损失; 标签扩散; 室内人群计数; 标注偏差

引用格式: 刘永文,孙君宇,凌妙根,苏健.基于循环卷积网络和逆变换贝叶斯损失的室内人群计数.计算机系统应用,2025,34(9):253–263. http://www.c-s-a.org.cn/1003-3254/9930.html

Circular Convolution Network with Inverse Transform Bayesian Loss for Indoor Crowd Counting

LIU Yong-Wen, SUN Jun-Yu, LING Miao-Gen, SU Jian

(School of Computer Science, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: This study focuses on the challenging task of indoor crowd counting. In indoor scenes, people often get together and perform similar tasks in constrained spaces. As most behaviors of indoor crowds are consequently quite similar, it is important to acquire a global receptive field and identify similarities in indoor crowd features. To address this problem, this study designs a circular convolution network, which combines the advantages of convolution neural networks and Transformer, to obtain both local and global correlations of the crowd features. Compared with the Transformer-based methods, this network adopts a much simpler and more efficient circular convolutional module. Moreover, a novel inverse-transform Bayesian loss function, which suits both sparse and crowded indoor scenes with large-scale variations, is proposed. Finally, to alleviate the influence of the annotation deviation, a label diffusion strategy that expands annotation areas by assuming adjacent pixels of each original annotation point may also potentially represent head centers. Compared with the second-best method on Class A, Class B, Canteen, and Mall datasets, this method improves *MAE*/*RMSE* by 4.1%/4.4%, 5.8%/8.0%, 3.9%/1.6%, and 3.9%/1.6%, respectively.

Key words: circular convolution; Bayesian loss; label diffusion; indoor crowd counting; annotation deviation

① 基金项目: 国家自然科学基金 (62202235)
收稿时间: 2025-01-26; 修改时间: 2025-02-17; 采用时间: 2025-03-23; csa 在线出版时间: 2025-07-25
CNKI 网络首发时间: 2025-07-28

目前大多数基于深度学习的人群计数方法主要关注室外场景. 然而, 室内人群计数同样具有巨大应用需求, 例如在公共安全, 客流管理和提升运输效率等方面. Ling 等人[1]总结了室内和室外场景在人群计数特征上的差异. 室内场景是一个相对概念, 其特点是人们通常聚集在有限空间内执行类似的任务, 人群移动较少, 人数变化缓慢, 如图 1(a) 和 (b) 所示, 分别为教室和食堂里的学生. 根据室内人群的这些特征, 可以得出结论: 大多数人群的行为非常相似. 因此, 用于捕捉人群特征相似性的全局感受野显得尤为重要. 然而, 传统的卷积神经网络 (CNN) 方法通常缺乏全局感受野. 虽然基于 vision Transformer (ViT)[2]的方法可以在图像块之间实现全局感知, 但其计算复杂度高且需要大量训练数据. 受 ParC-Net[3]启发, 本文设计了一个简单的循环卷积网络, 既能提取局部和全局特征, 又易于训练.

(a) 教室 (b) 食堂 (c) 草坪 (d) 道路

图 1 室内场景和室外场景中的人群

传统的室外人群计数方法[4,5]通常通过高斯函数对标注点进行处理, 从而生成 ground-truth 密度图, 并采用均方误差 (MSE) 损失进行监督训练. 在密集区域 (图 1(c), 草坪上的人群[6]), 每个标注点处的高斯核通常采用自适应的设置方法, 即与其到 k 个最近邻标注点的平均距离有关; 在尺度变化较小的区域 (图 1(d), 道路上的人群[7]), 高斯核则简单设为固定值. 然而, 在一些人数稀少的室内场景 (图 1(b)) 中, k 个 (如 3 个) 最近邻距离太远甚至无法找到. 因此, 自适应高斯核的方法表现不佳. 此外, 室内场景的图片通常从较近距离以较低的俯视角度拍摄, 这容易产生较大的尺度变化 (图 1 (a) 和 (b)), 使得固定高斯核不适用. 因此, 固定或自适应高斯密度图都不能很好地处理室内场景.

由于 Ma 等人[8]指出, 基于假设的高斯核而获得的密度图是不准确的, 甚至存在错误. 因此, 放弃生成高斯密度图可能更合理. Ma 等人[8]提出了一个基于点监督而非密度图监督的密度贡献概率模型来解决上述问题. 然而, 其似然函数仍依赖于高斯函数, 而根据经验设定的固定方差使其难以有效处理大尺度变化. 受 Liang 等人[9]采用焦点逆距离变换方法进行人群定位的启发, 本文提出一种逆变换贝叶斯损失函数. 该方法完全摒弃了高斯核, 适用于具有大尺度变化的稀疏场景, 而这一特点在室内场景中很常见.

由于头部中心是由人工标记的, 标注偏差是不可避免的. 因此, 为防止计数结果受到偏差影响, 本文还提出了一种标签扩散策略, 考虑了每个原始标注点的周围像素作为真实头部中心的可能性. 本文的代码可在 https://github.com/gyx-lyw/CIL 中获取.

本文的主要贡献总结如下.

1) 设计了具有全局感受野的循环卷积网络, 其结合了 CNN 和 Transformer 的优势.

2) 提出了一种基于点监督的损失函数, 消除了为室内人群计数任务选择高斯核的必要性.

3) 提出了一种标签扩散策略, 提高了对标注偏差的鲁棒性. 大量实验表明, 所提方法在所有室内数据集上都达到了先进的计数性能.

1 相关工作

1.1 室内人群计数

早期的室内人群计数方法致力于设计头部检测器[10,11]或头肩检测器[12,13]. 例如, Luo 等人[12]首先通过检测对人群进行分割并移除背景, 然后基于 K-means 聚类估计人数. 然而, 当人群严重遮挡时, 这种方法表现不佳. DigCrowd[14]利用图像中的深度信息区分近景和远景区域, 并在近景区域采用检测方法, 而在远景区域采用密度图回归方法. DecideNet[15]则从基于检测和回归的密度图中选择合适的估计. Ling 等人[1]提出了一种弱监督方法, 仅使用人群的数量信息作为监督. 该方法通过结合 $L_{2,1}$ 范数的高斯混合标签分布来建模人数标注的不确定性. 然而, 由于未使用深度学习方法, 无法直接应用于带有头部点标注的人群数据集.

1.2 室外人群计数

目前大多数方法将室外人群计数任务转化为密度图回归, 主要集中于增强 CNN 的多尺度特征提取能力. 例如多列 CNN[7]、空洞 CNN[5]、可变形 CNN[16]、残差 CNN[17]、图 CNN[18]和金字塔 CNN[19]. 然而, 上述大多数基于 CNN 的方法主要侧重于局部特征提取. Lin 等人[20]提出了一种基于 ViT[2]的可学习注意力, 用于聚合局部和全局特征. 然而, 其模型的参数量和计算量较大, 导致训练复杂度较高. TKR-Net[21]采用了多尺度高

斯平滑密度图和点图监督相结合, 由粗到精的训练方式. 引入的自适应 top-k 关系模块 (ATRM) 通过自适应滤波机制在像素之间建立 top-k 关系, 从而提升特征表示能力.

1.3 损失函数

MSE 损失广泛应用于基于密度图回归[4,5]的人群计数方法. 尽管取得了令人满意的结果, 但这些方法仍然存在一些缺陷. 首先, 由于忽略了不完美密度图带来的不利影响, 模型有时会因为使用对离群值敏感的损失函数而过拟合这些不准确的目标[8,22]. 其次, 最低训练损失并不代表最佳计数性能. 这些局限性削弱了这些方法的准确性和泛化能力. 为了解决上述问题, Wang 等人[23]采用了最优传输 (OT), 并引入了全变分 (TV) 损失来稳定 OT 的计算. 块级密度图学习[22,24,25]已在多个研究中被采用, 这些方法通过预测预定义计数区间的索引, 而不是直接预测计数值, 解决了学习目标不准确的问题. 上述方法将回归问题转化为分类问题, 其采用的交叉熵损失比 MSE 损失对噪声更具鲁棒性. 此外, P2PNet[26]和 CLTR[27]分别使用卷积和 Transformer[28]架构来预测图像中的一组头部点建议. 匈牙利算法被用来将这些建议与标注点进行匹配. 在点回归估计中使用 MSE 损失, 而在提议分类器的训练中使用交叉熵损失.

1.4 标注偏差处理

标注偏差存在于大多数人群计数数据集中, 并在生成 ground-truth 密度图时表现为噪声. 使用这些带有噪声的密度图作为学习目标可能会导致计数性能下降. 为解决这一问题, 研究者提出了多种方法. Liu 等人[16]旨在识别数据中更易受到噪声影响的特定区域. Oh 等人[29]施加了正则约束, 以防止模型对噪声数据过拟合. Jiang 等人[30]提出了通过训练合成图像来学习先验信息. 此外, 还有些方法[23,26]更新了用于将样本与其对应标签匹配的规则, 而不使用带噪声的密度图. ADSCNet[31]基于模型估计结果迭代修正标注, 但其忽略了由不准确估计引入额外噪声的问题. CHS-Net[32]提出了一种交叉头部监督框架, 该框架使用卷积头和 Transformer[28]头在噪声区域中相互监督.

2 研究方法

本节将详细说明整体框架的结构, 该框架由 3 个主要组件组成: 循环卷积网络 (CCNet)、逆变换贝叶斯损失 (ITBL) 和标签扩散策略 (LDS).

2.1 框架概述

整体框架如图 2 所示. 对于每张图像, 本文采用 VGG-19[33]提取局部特征, 提取的特征图然后被输入到改进的双分支循环卷积网络中进行全局特征提取. 随后, 将局部特征和全局特征结合, 并输入到通道注意力模块[34]中, 以关注重要通道并抑制无关通道. 之后, 添加一个回归解码器, 从通道注意力模块的输出生成密度图. 最后, 提出了一种适用于稠密和稀疏场景的逆变换贝叶斯损失函数, 用于训练整个网络, 同时采用标签扩散策略进一步提升性能.

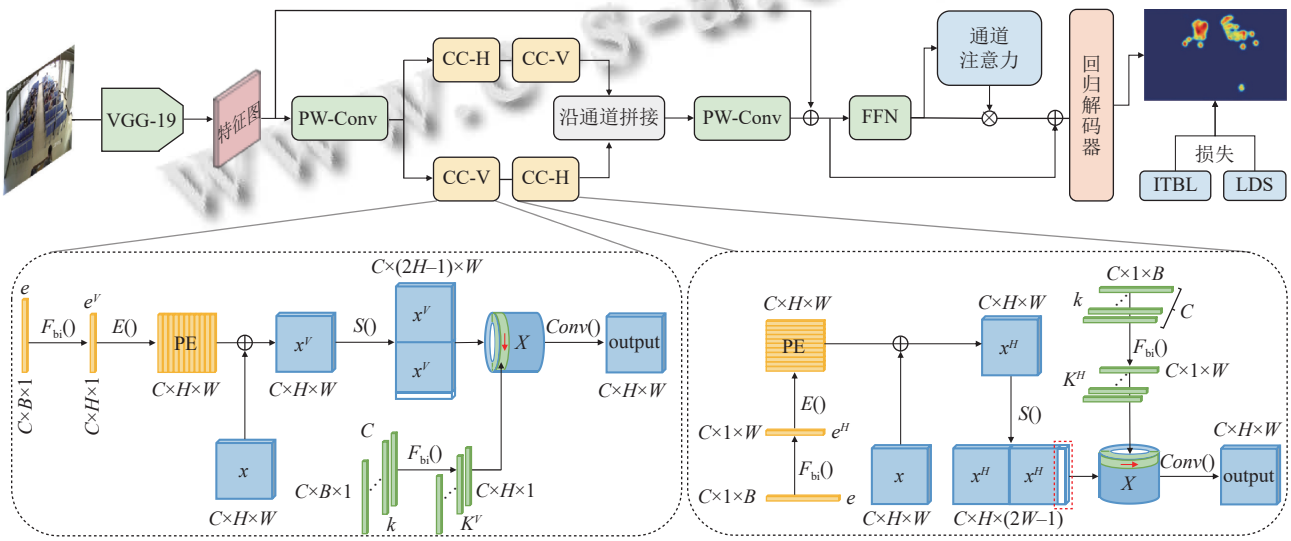


图 2 整体框架

2.2 循环卷积网络

(1) 使用骨干网络 VGG 提取局部特征.

本文采用在 ImageNet 上预训练的 VGG-19[33]作为骨干网络, 并去除其最后的池化层和全连接层. 骨干网络输出特征图的通道维度为 512.

(2) 使用循环卷积提取全局特征.

如图 2 所示, 为了减少后续循环卷积的计算复杂度, 首先通过逐点卷积将 VGG 网络输出的通道数减少到原来的 1/4. 所得到的特征图根据通道维度被平均分为两部分, 并分别送入到双分支的循环卷积 (CC) 结构中. 受位置感知循环卷积 (ParC)[3]的启发, 本文提出的循环卷积也有两种类型, 一种是水平方向的循环卷积 (CC-H), 另一种是垂直方向的循环卷积 (CC-V). 在第 1 个分支中, 特征图先经过 CC-H 再经过 CC-V 处理. 在第 2 个分支中, 特征图的处理顺序相反. CC-V 和 CC-H 的详细结构如图 2 中的虚线框所示.

以 CC-H 为例, 对于输入 $x \in R^{C\times H\times W}$, 其中 C、H 和 W 分别表示特征图的通道数、高度和宽度, 循环卷积的输出计算如下:

$$\begin{cases} e^H = F_{\text{bi}}(e),\ K^H = F_{\text{bi}}(k) \\ PE = E\left(e^H\right) \\ x^H = x + PE \\ X = S\left(x^H\right) \\ output = Conv\left(X, K^H\right) \end{cases} \tag{1}$$

其中, e 是维度为 $C\times 1\times B$ 的基础位置嵌入, k 是基础卷积核权重, 两者都是可训练的参数. 不同于 ParC 块中 k 仅包含一个卷积滤波器的情况, 本文的 k 具有 C 个滤波器, 每个滤波器的维度为 $C\times 1\times B$, 如果 k 仅有一个滤波器, 卷积操作会变成深度卷积. 实验部分表明深度卷积会破坏通道之间的交互性, 因此放弃深度卷积有助于提升模型的性能. B 是控制 e 和 k 大小的超参数, $F_{\text{bi}}()$ 是双线性插值函数, 用于调整 e 和 k 以适应网络输入不同分辨率的图像. $E()$ 是复制与拼接函数.

将输入向量 e 复制 H 次后, $E()$ 将这些复制的向量按垂直方向拼接, 生成一个 $H\times W$ 大小的实例位置嵌入 (PE) 矩阵. 通过将复制的水平位置嵌入 (PE) 添加到输入特征 x 上, CC-H 在水平方向上具有位置敏感性. 随后, $S()$ 将输入矩阵 x^H 按水平方向堆叠, 并丢弃最右边的一列 (图 2 红色虚线框标注的白色长方体), 以避免在循环卷积中出现重复计算. 为了更直观地理解循环卷积操作, 本文通过绘制一个圆柱体来描述 K^H 与 X 结合的过程. 在这里, 具有 C 个卷积滤波器的 K^H 被逐一应用于 X, 通过逆时针旋转 K^H, 获得尺寸为 $C\times H\times W$ 的输出. 每次旋转对应于一次使用 C 个滤波器的卷积操作. 由于 CC-H 的核大小为 $1\times W$ 且 W 是输入特征图的宽度, CC-H 能在水平方向提供全局感受野. 对于 CC-V, 如图 2 所示, 通过将复制的垂直位置嵌入 (PE) 添加到输入特征 x, CC-V 在垂直方向上具有位置敏感性. 由于 CC-V 的核大小为 $H\times 1$ 且 H 是输入特征的高度, CC-V 能在垂直方向提供全局感受野. 因此, 交替使用 CC-H 和 CC-V 能将输入特征的感受野扩展到全局范围.

最后, 将双分支循环卷积的输出在通道维度上拼接, 并通过逐点卷积将通道数还原为 512. 此外, 采用残差连接[35]来融合局部和全局特征, 其中全局特征的提取不会改变输入特征图的分辨率和通道数. 不同于 ParC-Net[3]使用的分叉结构来融合局部和全局特征, 为简化网络结构, 本文仅采用了一个循环卷积块.

(3) 使用通道注意力关注重要特征.

在特征前馈网络 (FFN) 模块中, 特征图经过以下处理: 1 个具有 2 个滤波器的逐点卷积层, 1 个 Hardswish 激活层, 1 个具 512 个滤波器的逐点卷积层, 以及 1 个丢弃率为 0.1 的 Dropout 层. 与 SENet[34]类似, 本文将通道注意力模块插入到框架中, 以关注重要特征. 具体来说, 输入特征图 $x_0 \in R^{C\times H\times W}$ 首先通过全局平均池化转换为 $x_1 \in R^{C\times 1\times 1}$. 然后, x_1 被用作多层感知 (MLP) 单元的输入, 生成对应的权重 $w \in R^{C\times 1\times 1}$. 最后, 原始特征图 x_0 与相应的权重 w 相乘, 以获得通道重要性特征, 并采用残差连接[35]以防止网络退化.

(4) 使用回归解码器生成密度图.

本文中的回归解码器由 1 个上采样层和 3 个带 ReLU 激活函数的卷积层组成. 上采样层使用双线性插值, 将特征图的高度和宽度变为输入图像的 1/8. 前两个卷积层的核大小为 3×3, 滤波器数量分别为 256 和 128. 最后一个卷积层是逐点卷积, 用于将通道数减少到 1.

2.3 逆变换贝叶斯损失

传统方法通常采用对异常值敏感的均方误差 (MSE) 损失进行训练, 但忽略了不准确标注对密度图质量带来的负面影响. 因此, Ma 等人[8]提出了贝叶斯损失函数, 将密度图回归问题转化为贡献概率问题. 其似然函数定义为:

$$p(x = x_m \mid z = z_n) = \exp\left(-\frac{\|x_m - y_n\|_2^2}{2\sigma^2}\right) = \exp\left(-\frac{d^2}{2\sigma^2}\right) \tag{2}$$

其中, $\{x_m \mid 1 \leqslant m \leqslant M\}$和$\{y_n \mid 1 \leqslant n \leqslant N\}$分别表示二维像素点和头部标注点, M和N分别为图像中所有像素和人群的总数. d表示每个像素点与头部标注点之间的欧几里得距离. $z_n = n$表示第 n 个人的标签. 式 (2) 计算了给定标签z_n条件下该人出现在位置x_m的概率. 然而, 其似然函数仍然依赖于具有固定方差的高斯函数. 本文认为固定方差通常是基于个人经验确定的. 此外, 如上所述, 自适应高斯方法可能不适用于某些稀疏的室内场景.

Liang 等人[9]提出使用焦点逆距离变换 (FIDT) 函数来生成 ground-truth 密度图, 其密度图在远离头部中心时减少密度值, 并将背景的大部分区域赋值为 0. FIDT 函数的输出范围是 0–1, 这似乎是另一种似然函数的选择. 然而, FIDT 仅利用每个像素点与其最近标注点之间的距离d来生成密度图. 由于标注点容易出现偏差, 而最近邻算法对噪声和异常值高度敏感, 本文放弃了最近邻算法, 提出了逆变换 (IT) 函数作为贝叶斯损失的似然函数, 其计算公式如下:

$$p(x = x_m \mid z = z_n) \overset{\text{def}}{=} F_{\text{it}}(x_m; y_n, \alpha, \beta) = \frac{1}{d^{(\alpha \times d + \beta)} + 1} \tag{3}$$

其中, 函数F_{it}被用来简化符号表示.

图 3 比较了传统高斯函数 (式 (2)) 和本文 IT 函数 (式 (3)). 可以看出, 当距离 d 为 0–5 时, IT 函数的下降速度快于方差较大的高斯函数 (σ>3). 这意味着 IT 函数在概率分布上更关注头部中心. 当距离d>5 时, IT 函数的下降速度慢于方差较大的高斯函数, 从而为周围点分配一定的概率, 以容忍标注偏差. 与方差较小的高斯函数 (如σ=1) 相比, IT 函数在距离大于 5 时提供了更丰富的概率信息. 这防止了概率分布仅限于一个狭小的区域. 因此, IT 函数比高斯函数更适合作为贝叶斯损失的似然函数.

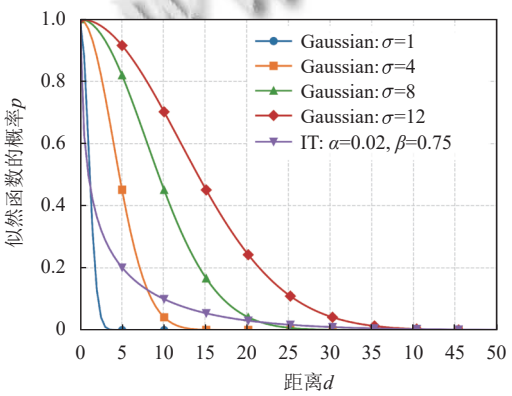


图 3 高斯函数与 IT 函数的对比

根据贝叶斯公式, 在给定位置x_m有人的条件下, 该人具有标签z_n的后验概率可以计算为:

$$p(z_n \mid x_m) = \frac{p(x_m \mid z_n)p(z_n)}{p(x_m)} = \frac{p(x_m \mid z_n)p(z_n)}{\sum_{n=1}^{N} p(x_m \mid z_n)p(z_n)} = \frac{p(x_m \mid z_n)}{\sum_{n=1}^{N} p(x_m \mid z_n)} = \frac{F_{\text{it}}(x_m; y_n, \alpha, \beta)}{\sum_{n=1}^{N} F_{\text{it}}(x_m; y_n, \alpha, \beta)} \tag{4}$$

其中, 第 3 个等号成立的前提是假设每个类别标签z_n的先验概率$p(z_n)$相等, 这一假设不失一般性. 结合后验概率$p(z_n \mid x_m)$和估计的密度图D^{est}, 本文的逆变换贝叶斯损失 (ITBL) 为:

$$\mathcal{L}_{\text{ITBL}} = \sum_{n=1}^{N} \left| 1 - \sum_{m=1}^{M} p(z_n \mid x_m) D^{\text{est}}(x_m) \right| \tag{5}$$

其中, $p(z_n \mid x_m) D^{\text{est}}(x_m)$表示像素点$x_m$对标签$z_n$的贡献计数. 所有像素点对$z_n$的累计贡献总和期望为 1, 即标签$z_n$在图像中出现的概率.

2.4 标签扩散策略

由于标注点是人工标注的, 且图片的背景环境复杂, 光照条件多变, 标注偏差不可避免. 受 Ling 等人[1]使用标签分布表示人群数量不确定性的启发, 本文提出了标签扩散策略, 为标注点周围的像素点分配一定的权重. 该策略认为, 周围的点也有一定的概率成为人的中心. 所有扩散点的权重 (包括原始点) 的总和为 1, 并且原始标注点仍然拥有最大的权重.

具体而言, 设$\left\{ y_n^k \mid 1 \leqslant n \leqslant N,\ 1 \leqslant k \leqslant K \right\}$表示第 n 个标注点的第 k 个扩散点, z_n^k表示相应的标签, 其中K是扩散像素的数量. 这里, y_n^1和z_n^1分别表示原始标注点和原始标签. 首先, 本文构造了位置x_m在给定标签z_n^k下的似然函数:

$$p\left(x = x_m \mid z = z_n^k\right) = F_{\text{it}}\left(x_m; y_n^k, \alpha, \beta\right) \tag{6}$$

然后, 计算位置x_m拥有标签z_n^k的后验概率, 表示为:

$$p\left(z_n^k \mid x_m\right) = \frac{F_{\text{it}}\left(x_m; y_n^k, \alpha, \beta\right)}{\sum_{n=1}^{N} \sum_{k=1}^{K} F_{\text{it}}\left(x_m; y_n^k, \alpha, \beta\right)} \tag{7}$$

最终, 可以得到以下损失函数:

$$\begin{cases} \mathcal{L}_{\text{ITBL+LDS}} = \sum_{n=1}^{N} \sum_{k=1}^{K} \left| \xi^k - \sum_{m=1}^{M} p\left(z_n^k \mid x_m\right) D^{\text{est}}(x_m) \right| \\ \text{s.t.} \sum_{k=1}^{K} \xi^k = 1 \end{cases} \tag{8}$$

其中, ξ^k表示每个扩散点的权重, 而$\sum_{k=1}^{K} \xi^k = 1$表示扩

散标注点的真实计数和应为 1. 实验探索了不同的标签扩散策略, 发现给扩散点分配相等的权重表现最佳, 并且扩散点应该限制在一个较小的范围内. 不同于 Ma 等人[8]提出的点监督, 本文的 ITBL+LDS 损失函数使用一个点集区域来监督训练过程, 增强了原始标注点的表征能力和模型对标注噪声的鲁棒性.

3 实验结果与分析

3.1 训练细节

本文通过随机裁剪和水平翻转来扩充训练集. 式 (3) 中的α和β分别设置为 0.02 和 0.75. 本文将式 (8) 中 K 的值设置为 4, 并将ξ^k固定为 0.2, 这意味着每个原始标注点会扩散成一个包含其本身及其上、下、左、右 5 个像素点的点集, 且这 5 个点具有相等的权重. 本文所有实验基于一个配备 Intel Core i7-12700F CPU 和 NVIDIA GTX 4090 GPU 的 PyTorch 框架, 使用 Adam 优化器, 并设置初始学习率为 5E–6 来更新参数. 批量大小和权重衰减分别设置为 1 和 1E–4.

3.2 数据集与评估指标

实验评估在 5 个室内人群计数数据集上进行: Class A[36]、Class B[1]、Bus[36]、Canteen[36]和 Mall[37]. 各数据集统计信息总结在表 1 中. 表 1 中 $Res(H\times W)$ 表示图像的分辨率 (高度×宽度). Tr 和 Tst 分别表示训练集和测试集的大小. Min、Max 和 Avg 分别表示数据集中标注的最小、最大和平均人头数量.

表 1 本文中数据集的统计信息

数据集	$Res(H\times W)$	Tr	Tst	Min	Max	Avg
Class A[36]	720×1 280	645	645	1	51	30.3
Class B[1]	576×704	2 193	2 193	24	37	31.9
Bus[36]	576×704	1 413	1 413	12	27	21.6
Canteen[36]	288×352	3 000	3 000	1	16	5.4
Mall[37]	480×640	800	1 200	13	53	31.2

● Class B 数据集. 来源于 Ling 等人[1]的研究, 但其仅包含每帧的总人数标签. 为了与基于密度图的方法进行比较, 本文人工标注了每个人的头部中心位置用于监督. 在 Class B 中, 大多数学生在教室上晚自习, 其余学生逐渐进入教室.

● Class A、Bus 和 Canteen 数据集. 同样源于 Ling 等人[36]的研究. 在 Class A 中, 首先是第 1 批学生在教室上课, 第 1 批学生下课后, 第 2 批学生进入教室上课. 因此, Class A 的人群数量范围为 1–51, 是所有数据集中人群数量和透视变化幅度最大的场景. 在 Bus 数据集中, 人群的遮挡最为严重. 而在 Canteen 数据集中, 学生在不同的队列中等候用餐.

● Mall 数据集[37]. 是一个传统的室内数据集, 采集于一个购物中心. 本文遵循其原始设置, 使用前 800 帧作为训练集, 其余 1 200 帧作为测试集.

● 评估指标. 使用平均绝对误差 (*MAE*) 和均方根误差 (*RMSE*) 来评估不同方法的计数性能, 其定义为:

$$MAE=\frac{1}{M}\sum_{i=1}^{M}\left|N_i^{\mathrm{gt}}-N_i\right| \tag{9}$$

$$RMSE=\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(N_i^{\mathrm{gt}}-N_i\right)^2} \tag{10}$$

其中, M是测试集的数量, N_i^{gt}和N_i分别表示第i张图像人数的真实值和预测值.

3.3 评估与比较

本文将所提方法与 12 个最先进方法在 Class A、Class B、Bus 和 Canteen 数据集上进行了对比. BL[5]作为本研究的基线, 本文的改进都基于此方法. 本文尝试将 ParC-Net[3]应用于室内人群计数任务, 但效果不佳. 这可能是由于 ParC-Net 旨在设计轻量级的网络, 其参数量仅接近于 5 百万, 远小于其他对比的人群计数方法 (约 16 百万–30 百万). ConvNeXt[38]是一种新提出的卷积神经网络, 其采用模块化架构, 并吸收了许多来自 Swin Transformer[39]的优秀设计. 在实验中, 本文将 ConvNeXt-T 末端的线性层替换为上采样块, 使整个下采样率保持为 8, 并使用 MSE 损失进行训练.

计数精度的结果总结于表 2. 其中 BL[8]是本文的基线, Average 表示在 4 个室内数据集上的平均结果. 表 2 中最佳性能以粗体表示, 次优性能以下划线表示. 可以看出, 在 4 个数据集的平均结果上, 本文方法在 *MAE*/*RMSE* 上分别比第 2 和第 3 优秀方法 (HMoDE[40] 和 CHS-Net[32]) 提升了 2.2%/3.5% 和 8.2%/6.2%. 除 Bus 数据集外, 本文方法在所有数据集的所有评价指标上均表现出卓越的计数性能. 具体来说, 与次优的结果相比, Class A 的 *MAE*/*RMSE* 降低了 4.1%/4.4%, Class B 降低了 5.8%/8.0%, Canteen 降低了 3.9%/1.6%. 这表明, 本文结合传统卷积归纳偏置和 Transformer[28]全局感知能力的循环卷积结构, 相较于基于卷积的方法[4,5,38]或基于 Transformer 的方法[20,32], 在 3 个室内场景中更加有效.

表 2 在 Class A, Class B, Bus 和 Canteen 上与最先进方法的比较

方法	Class A		Class B		Bus		Canteen		Average	
	MAE	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*
CSRNet[5] (CVPR 2018)	3.42	3.86	1.20	1.37	1.99	2.40	1.91	2.34	2.13	2.49
CAN[4] (CVPR 2019)	1.72	2.09	0.87	1.05	2.25	2.69	1.75	2.15	1.65	2.00
KDMG[41] (PAMI 2020)	4.17	4.48	0.94	1.14	2.43	2.92	1.66	2.10	2.30	2.66
DM-Count[23] (NeurIPS 2020)	2.86	3.49	0.74	0.96	2.37	2.95	2.37	2.82	2.09	2.56
TopoCount[42] (AAAI 2021)	2.75	3.34	1.30	1.65	2.69	3.44	1.71	2.17	2.11	2.65
P2PNet[26] (ICCV 2021)	1.63	2.08	1.31	1.71	4.28	5.26	2.64	3.30	2.47	3.09
MAN[20] (CVPR 2022)	2.35	2.72	1.15	1.35	2.75	3.28	2.54	2.87	2.20	2.56
ConvNeXt[38] (CVPR 2022)	2.71	3.29	0.83	1.04	**1.65**	**2.01**	1.55	2.02	1.69	2.09
CHS-Net[32] (ICASSP 2023)	1.46	1.80	0.80	0.94	1.81	2.19	1.82	2.19	1.47	1.78
HMoDE[40] (TIP 2023)	1.49	1.95	0.69	0.87	1.77	2.13	1.57	1.95	1.38	1.73
Gramformer[43] (AAAI 2024)	1.65	2.07	0.77	0.91	2.25	2.62	1.53	1.91	1.55	1.88
BL[8] (ICCV 2019)	2.86	3.14	1.32	1.64	3.03	3.70	2.81	3.30	2.51	2.95
Ours	**1.40**	**1.72**	**0.65**	**0.80**	1.86	2.29	**1.47**	**1.88**	**1.35**	**1.67**

在 Bus 场景中表现较差可能归因于严重的拥挤遮挡, 这容易导致较大的标注偏差. 本文的标签扩散策略把注释区域从原先的单个标注点扩大到包含其周围像素点的点集, 这虽然增强了对标注点小偏差的鲁棒性, 但在像 Bus 这种具有较大标注偏差的场景中, 进展仍然有限. 与基线方法 BL[8]相比, 本文方法显著提高了 4 个数据集的计数精度. 具体而言, Class A 上的 *MAE* 和 *RMSE* 分别提升了 51.0% 和 45.2%; Class B 上提升了 50.8% 和 51.2%; Bus 上提升了 38.6% 和 38.1%; Canteen 上提升了 47.7% 和 43.0%. 这证明了循环卷积的全局感受野、逆变换贝叶斯损失对尺度变化的鲁棒性以及标签扩散策略对标注偏差的容忍性, 在复杂的室内场景中很有效.

本文还在 Mall[37]数据集上与 11 种最新方法进行了比较, 结果如表 3 所示. 由于 Ling 等人[1]仅使用总人数进行监督, 为了公平起见, 实验中未将其与本文中的方法进行比较. 与次优方法 DecideNet[15]相比, 本文方法的 *MAE* 和 *RMSE* 分别提升了 3.9% 和 1.6%. 与 BL[8]相比, 本文方法分别将 *MAE* 和 *RMSE* 降低了 25.5% 和 24.9%. BL 与本文方法的可视化结果如图 4 所示.

表 3 在 Mall 数据集上与最先进方法的比较

方法	*MAE*	*RMSE*
CSRNet[5] (CVPR 2018)	1.72	2.17
DecideNet[15] (CVPR 2018)	1.52	1.90
KDMG[41] (PAMI 2020)	2.35	2.92
DM-Count[23] (NeurIPS 2020)	2.02	2.52
MAN[20] (CVPR 2022)	2.14	2.71
LoViTCrowd[44] (BMVC 2022)	1.66	2.10
ConvNeXt[38] (CVPR 2022)	1.62	2.02
CHS-Net[32] (ICASSP 2023)	1.71	2.18
HMoDE[40] (TIP 2023)	1.70	2.14
Gramformer[43] (AAAI 2024)	1.69	2.14
BL[8] (ICCV 2019)	1.96	2.49
Ours	**1.46**	**1.87**

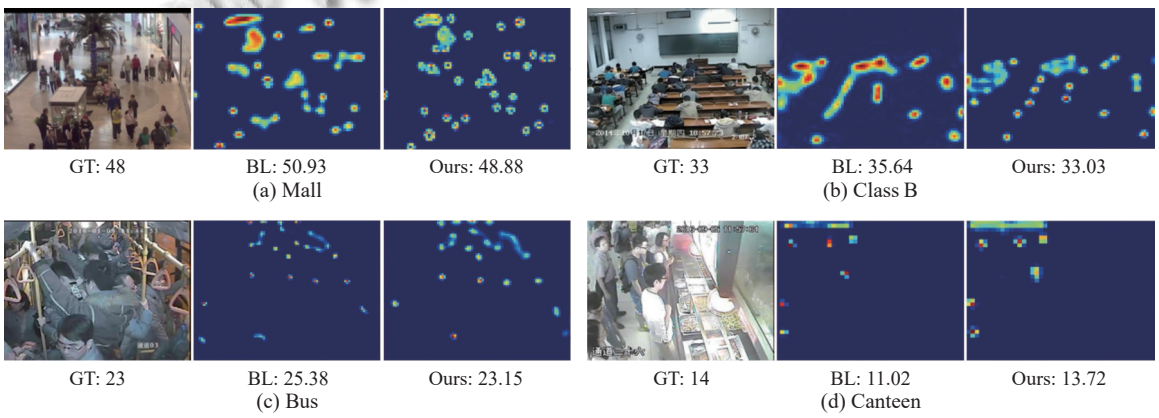
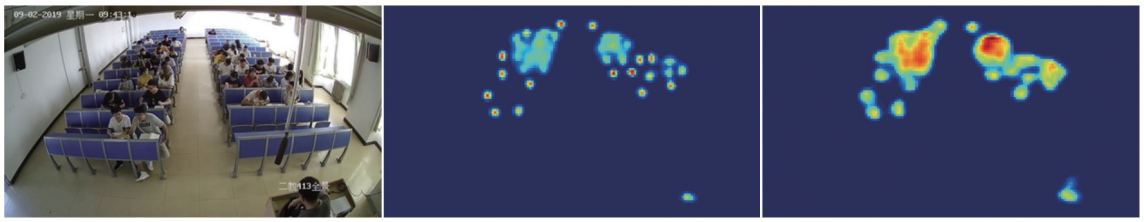


图 4 BL 和本文方法在 5 个室内人群计数数据集上的可视化结果



GT: 44 BL: 41.59 Ours: 43.73

(e) Class A

图 4 BL 和本文方法在 5 个室内人群计数数据集上的可视化结果 (续)

3.4 关键问题与讨论

(1) 消融研究

本节在 Class A 上进行消融研究, 结果见表 4. 实验从基线 BL[8] (第 1 行) 开始, 测试将贝叶斯损失替换为本文所提 ITBL 的效果 (第 2 行). 结果显示, *MAE* 和 *RMSE* 分别减少了 18.2% 和 15.3%. 这说明 ITBL 对尺度变化的鲁棒性对于增强室内场景的结果非常有用. 添加 LDS 后, ITBL 性能在 *MAE*/*RMSE* 上进一步提升了 13.2%/7.5% (第 3 行). 这表明 LDS 对标注偏差的容忍性也是有效的. 如果直接使用所提出的 CCNet 替换 BL 中的 VGG-19, *MAE* 和 *RMSE* 分别提升了 33.9% 和 29.3% (第 4 行). 这证明了 CCNet 的全局感受野对于性能提升非常有效. 随后, 在 CCNet 的基础上, 将贝叶斯损失替换为本文的 ITBL, 进一步将 *MAE* 减少了 12.2%, *RMSE* 减少了 11.7% (第 5 行). 最后, 通过引入 LDS 实现了最佳性能, 使 BL 的计数准确率在 *MAE* 和 *RMSE* 上分别提升了 51.0% 和 45.2% (最后一行).

表 4 在 Class A 上的消融研究

CCNet	ITBL	LDS	*MAE*	*RMSE*	参数量 (百万)	训练时间 (s)
—	—	—	2.86	3.14	21.50	14.68
—	√	—	2.34	2.66	21.50	14.73
—	√	√	2.03	2.46	21.50	14.75
√	—	—	1.89	2.22	21.67	15.33
√	√	—	1.66	1.96	21.67	15.38
√	√	√	**1.40**	**1.72**	21.67	15.40

本节还探索了不同方法的参数量和在 Class A 上训练每个 epoch 的平均时间. 如表 4 的最后两列所示, 与基线方法 BL[8] (第 1 行) 相比, 本文的 CCNet 仅增加了 0.17 百万个参数, 并且每个 epoch 的运行时间仅增加了 0.65 s. 此外, ITBL 和 LDS 不包含任何可学习参数. 将贝叶斯损失替换为 ITBL 和添加 LDS 分别只增加了 0.05 s 和 0.02 s 的训练时间.

(2) 循环卷积 (CC) 和位置感知循环卷积 (ParC) 的影响

ParC-Net[3]在 ParC 模块中使用深度卷积以减少参数量. 然而, 本文认为深度卷积会破坏通道间的交互. 因此, 本文在 CC-H 和 CC-V 模块中使用常规卷积代替深度卷积. 表 5 为在 Class A 数据集上进行的对比实验. 在 VGG-19+ITBL 基础上, 本文方法使用 CC 模块在 *MAE*/*RMSE* 上比 ParC 版本提高了 16.6%/18.0%. 加入 LDS 后, 本文方法使用 CC 模块在 *MAE*/*RMSE* 上仍然比 ParC 版本提高了 17.6%/16.9%. 这表明 CC 模块相较于 ParC 模块在 Class A 数据集上具有显著优势.

表 5 在 Class A 上循环卷积 (CC) 与位置感知循环卷积 (ParC) 的对比结果

方法	使用ParC		使用CC	
	MAE	*RMSE*	*MAE*	*RMSE*
VGG-19+ITBL	1.99	2.39	**1.66**	**1.96**
VGG-19+ITBL+LDS	1.70	2.07	**1.40**	**1.72**

(3) α和β的影响

α和β的值越大, 则本文 IT 函数的衰减速度越快, 这表明 IT 函数越聚焦于头部中心, 对于远离头部中心的点提供越少的概率信息. 如表 6 所示, 与 FIDT[8]类似, 本节研究了 4 组不同α和β对 ITBL 和 LDS 的影响, 实验仍然在 Class A 数据集上进行. 当α和β分别设置为 0.02 和 0.75 时, IT 函数在本文的两种方法上都取得了最佳结果, 这表明此时的衰减速度更加合适. 此外, 本文的损失函数对α和β的选取具有良好的鲁棒性, 其性能受参数变化的影响较小.

表 6 在 Class A 上α和β对损失函数的影响

α	β	CCNet+ITBL		Ours	
		MAE	*RMSE*	*MAE*	*RMSE*
0.01	0.65	**1.66**	2.00	1.45	**1.72**
0.02	0.75	**1.66**	**1.96**	**1.40**	**1.72**
0.03	0.85	1.70	2.02	1.47	1.77
0.04	0.95	1.67	1.97	1.45	1.78

(4) ξ^k的影响

本节探索了权重分配策略对 LDS 的影响. 假设仅采用原始标注点的上、 下、左、右像素点 (4 个点) 进

行标签扩散, 将最大权重分配给原始标注点, 其余权重平均分配给扩散像素. 如图 5 所示, 当所有点的权重均等分配为 0.2 时, 结果最佳. 当原始标注点的权重逐渐从 0.2 增加到 0.3 时, *MAE* 和 *RMSE* 缓慢变差. 从 0.3 增加到 0.6 时, *MAE* 和 *RMSE* 急剧增加. 这表明头部中心在相邻像素中存在的概率不应有显著差异, 均等分配权重是最佳选择.

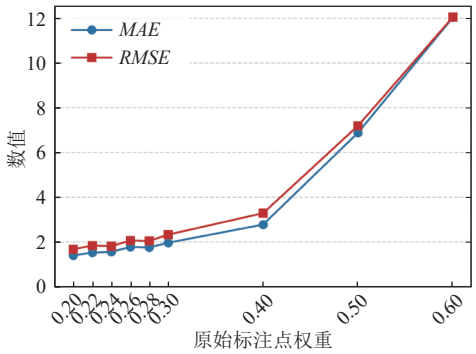


图 5 在 Class A 上不同权重分配策略对 LDS 的影响

(5) *K* 的影响

本节尝试设置不同数量的扩散像素点 *K*. 在 Class A 数据集上分别对 5、9 和 13 个扩散点进行了实验, 如表 7 所示. LDS 列的中心像素表示原始标注点, 其余像素表示用于扩散的相邻像素. 由于通过均等分配权重可以获得最佳的计数结果, 5、9 和 13 个扩散像素都被平均分配权重, 总和为 1. 从结果可以看出, 5 个扩散像素获得了最佳结果, 而增加更多扩散点会使性能变差. 这可能是因为原始头部标注点大多非常准确, 平均偏差非常小, 更大的标签扩散范围会导致性能下降.

表 7 扩散像素数量对 LDS 的影响

LDS	*K*	*MAE*	*RMSE*
	5	**1.40**	**1.72**
	9	1.61	2.01
	13	1.84	2.13

(6) 对标注偏差的鲁棒性

遵循 Ma 等人[8]的研究, 本节验证本文方法对标注偏差的鲁棒性. 为模拟手动标注的偏差, 在头部标注点中均匀引入了随机噪声, 随后在不同噪声水平下评估不同方法的性能, 本文在头部标注点中均匀添加随机噪声, 噪声范围为图像高度的百分比, 实验结果如图 6 所示. CCNet+ITBL 在所有的噪声水平下均优于 BL[8]. 这表明本文方法相比 BL 对标注偏差具有更高的鲁棒性. 当添加 LDS 后, 性能进一步提升. 例如, 在 4% 和 5% 的随机噪声下, 添加 LDS 使 *MAE*/*RMSE* 分别减少了 21.1%/18.7% 和 7.8%/6.8%. 这证明标签扩散策略在提高对标注偏差的鲁棒性方面非常有效.

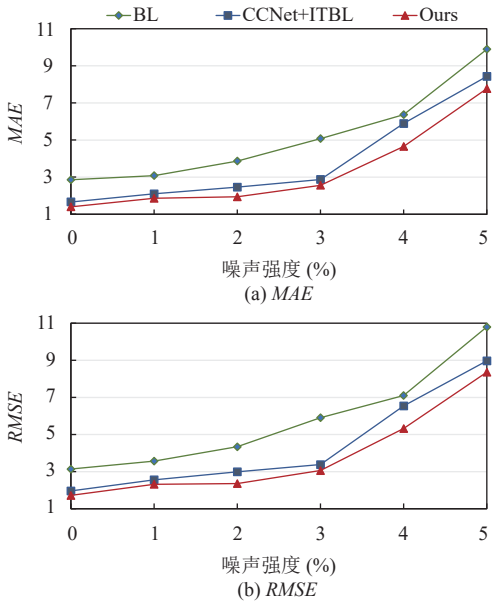


图 6 标注偏差的鲁棒性评估

(7) 跨数据集评估

本节通过在教室场景的 Class A 和 Class B 数据集上进行跨数据集实验, 以研究不同方法的泛化能力. 具体而言, 模型在一个数据集上训练, 在另一个数据集上进行测试, 且没有额外微调. 从表 8 可以看出, 本文方法在两个数据集上的泛化能力优于基线 BL[8]、基于 Transformer[28]的 MAN[20]和基于 CNN 的 ConvNeXt[38].

表 8 跨数据集评估

方法	Class A→Class B		Class B→Class A	
	MAE	*RMSE*	*MAE*	*RMSE*
BL[8]	6.66	6.97	8.96	9.15
MAN[20]	5.35	5.61	7.63	8.02
ConvNeXt[38]	3.89	4.18	6.70	7.23
Ours	**3.30**	**3.69**	**5.74**	**6.00**

(8) 运行成本评估

表 9 对不同方法的参数量进行了对比, 在分辨率为 384×384 的输入图像上计算的浮点运算数 (GFLOPs), 以及在 NVIDIA 4090 GPU 上对分辨率为 1024×1024 的 100 张图像进行推理的时间. 与基于 Transformer[28]的 MAN[20]相比, CCNet 具有更少的参数量和更短的推理时间. CCNet 的计算量略高于 MAN, 这归因于 CCNet 输出密度图的分辨率是 MAN 的 4 倍.

表 9 运行成本评估

方法	参数量 (百万)	GFLOPs	推理时间 (s)
CSRNet[5]	16.26	60.90	6.26
BL[8]	21.50	60.73	6.19
MAN[20]	30.96	59.54	6.92
Ours	21.67	60.81	6.60

4 总结与展望

本文提出了一种用于室内人群计数的循环卷积网络. 该网络结合了 CNN 和 Transformer 的优势, 捕获了人群的局部和全局相关性. 此外, 提出了一种逆变换贝叶斯损失, 适用于监督具有大规模变化的稀疏和拥挤的室内场景. 为了提高对标注偏差的鲁棒性, 提出了标签扩散策略, 以扩大标注范围. 大量实验表明, 本文的方法在准确性、鲁棒性和泛化能力上优于以前的方法.

尽管本文的方法提高了对标注偏差的容忍度, 并取得了非常好的结果, 但所有数据集仍然需要人工标注头部中心, 这既费力又费时. 考虑到室内人群移动缓慢和人群数量变化较小, 未来考虑研究半监督或无监督的方法来进行室内人群计数, 这无疑会减少对人工标注的依赖, 并更容易应用于不同的室内场景.

参考文献

1 Ling MG, Geng X. Indoor crowd counting by mixture of Gaussians label distribution learning. IEEE Transactions on Image Processing, 2019, 28(11): 5691–5701. [doi: 10.1109/TIP.2019.2922818]

2 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.

3 Zhang HK, Hu WZ, Wang XY. ParC-Net: Position aware circular convolution with merits from ConvNets and Transformer. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 613–630.

4 Liu WZ, Salzmann M, Fua P. Context-aware crowd counting. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5094–5103.

5 Li YH, Zhang XF, Chen DM. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 1091–1100.

6 Idrees H, Tayyab M, Athrey K, *et al*. Composition loss for counting, density map estimation and localization in dense crowds. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 544–559.

7 Zhang YY, Zhou DS, Chen SQ, *et al*. Single-image crowd counting via multi-column convolutional neural network. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 589–597.

8 Ma ZH, Wei X, Hong XP, *et al*. Bayesian loss for crowd count estimation with point supervision. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6141–6150.

9 Liang DK, Xu W, Zhu YY, *et al*. Focal inverse distance transform maps for crowd localization. IEEE Transactions on Multimedia, 2023, 25: 6040–6052. [doi: 10.1109/TMM.2022.3203870]

10 Stewart R, Andriluka M, Ng AY. End-to-end people detection in crowded scenes. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 2325–2333.

11 Subburaman VB, Descamps A, Carincotte C. Counting people in the crowd using a generic head detector. Proceedings of the 2012 IEEE Ninth International Conference on Advanced Video and Signal-based Surveillance. Beijing: IEEE, 2012. 470–475.

12 Luo J, Wang JQ, Xu HZ, *et al*. Real-time people counting for indoor scenes. Signal Processing, 2016, 124: 27–35. [doi: 10.1016/j.sigpro.2015.10.036]

13 Yang R, Xu HZ, Wang JQ. Robust crowd segmentation and counting in indoor scenes. Proceedings of the 22nd International Conference. Miami: Springer, 2016. 505–514.

14 Xu ML, Ge ZY, Jiang XH, *et al*. Depth information guided crowd counting for complex crowd scenes. Pattern Recognition Letters, 2019, 125: 563–569. [doi: 10.1016/j.patrec.2019.02.026]

15 Liu J, Gao CQ, Meng DY, *et al*. DecideNet: Counting varying density crowds through attention guided detection and density estimation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5197–5206.

16 Liu N, Long YC, Zou CQ, *et al*. ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3220–3229.

17 Zhao K, Liu B, Song LC, *et al*. Cascaded residual density network for crowd counting. Proceedings of the 2019 IEEE International Conference on Image Processing. Taipei: IEEE, 2019. 2199–2203.

18 Luo A, Yang F, Li X, *et al*. Hybrid graph neural networks for crowd counting. Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020. 11693–11700.

19 Chen XY, Bin YR, Sang N, *et al*. Scale pyramid network for crowd counting. Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2019. 1941–1950.
20 Lin H, Ma ZH, Ji RR, *et al*. Boosting crowd counting via multifaceted attention. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 19596–19605.
21 Dong L, Zhang HJ, Yang K, *et al*. Crowd counting by using top-k relations: A mixed ground-truth CNN framework. IEEE Transactions on Consumer Electronics, 2022, 68(3): 307–316. [doi: 10.1109/TCE.2022.3190384]
22 Wang CA, Song QY, Zhang BS, *et al*. Uniformity in heterogeneity: Diving deep into count interval partition for crowd counting. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 3214–3222.
23 Wang BY, Liu HD, Samaras D, *et al*. Distribution matching for crowd counting. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 135.
24 Liu L, Lu H, Xiong HP, *et al*. Counting objects by blockwise classification. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(10): 3513–3527. [doi: 10.1109/TCSVT.2019.2942970]
25 Xiong HP, Lu H, Liu CX, *et al*. From open set to closed set: Counting objects by spatial divide-and-conquer. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 8361–8370.
26 Song QY, Wang CA, Jiang ZK, *et al*. Rethinking counting and localization in crowds: A purely point-based framework. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 3345–3354.
27 Liang DK, Xu W, Bai X. An end-to-end Transformer model for crowd localization. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 38–54.
28 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
29 Oh MH, Olsen P, Ramamurthy KN. Crowd counting with decomposed uncertainty. Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020. 11799–11806.
30 Jiang XH, Liu H, Zhang L, *et al*. Transferring priors from virtual data for crowd counting in real world. Frontiers of Computer Science, 2022, 16(3): 163314. [doi: 10.1007/s11704-021-0387-8]
31 Bai S, He ZQ, Qiao Y, *et al*. Adaptive dilated network with self-correction supervision for counting. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4593–4602.
32 Dai ML, Huang ZZ, Gao JQ, *et al*. Cross-head supervision for crowd counting with noisy annotations. arXiv:2303.09245v1, 2023.
33 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.
34 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.
35 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
36 Ling MG, Pan TH, Ren Y, *et al*. Motional foreground attention-based video crowd counting. Pattern Recognition, 2023, 144: 109891. [doi: 10.1016/j.patcog.2023.109891]
37 Chen K, Loy CC, Gong SG, *et al*. Feature mining for localised crowd counting. British Machine Vision Conference, 2012, 1(2): 3.
38 Liu Z, Mao HZ, Wu CY, *et al*. A ConvNet for the 2020s. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11966–11976.
39 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.
40 Du ZP, Shi MJ, Deng JK, *et al*. Redesigning multi-scale neural network for crowd counting. arXiv:2208.02894, 2022.
41 Wan J, Wang QZ, Chan AB. Kernel-based density map generation for dense object counting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1357–1370. [doi: 10.1109/TPAMI.2020.3022878]
42 Abousamra S, Hoai M, Samaras D, *et al*. Localization in the crowd with topological constraints. Proceedings of the 35th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 872–881.
43 Lin H, Ma ZH, Hong XP, *et al*. Gramformer: Learning crowd counting via graph-modulated Transformer. Proceedings of the 38th AAAI Conference on Artificial Intelligence. Washington: AAAI Press, 2024. 3395–3403.
44 Tran NH, Huy TD, Duong STM, *et al*. Improving local features with relevant spatial information by vision Transformer for crowd counting. Proceedings of the 33rd British Machine Vision Conference. London: BMVA Press, 2022. 729.

(校对责编: 王欣欣)