计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(9): 2194−2207, 2025

渐进式认知引导的双域半监督人群计数

余 鹰[1,2] 范在昌[2] 曾康利[2] 黄晓辉[2] 苗夺谦[2,3]

[1]（山区土木工程安全与韧性全国重点实验室（华东交通大学） 南昌 330013）

[2]（华东交通大学信息与软件工程学院 南昌 330013）

[3]（同济大学电子与信息工程学院 上海 201804）

（yuyingjx@163.com）

Progressive Cognition-Guided Dual-Domain Semi-Supervised Crowd Counting

Yu Ying[1,2], Fan Zaichang[2], Zeng Kangli[2], Huang Xiaohui[2], and Miao Duoqian[2,3]

[1]（*State Key Laboratory of Safety and Resilience of Civil Engineering in Mountain Area* (*East China Jiaotong University*), *Nanchang* 330013）

[2]（*School of Information and Software Engineering*, *East China Jiaotong University*, *Nanchang* 330013）

[3]（*College of Electronics and Information Engineering* , *Tongji University*, *Shanghai* 201804）

Abstract In crowd counting tasks, the high cost of data annotation severely limits the broad application of fully supervised methods. To significantly reduce reliance on labeled data, semi-supervised counting methods leveraging large amounts of unlabeled data have emerged as the mainstream research direction. However, existing semi-supervised approaches typically rely on iteratively generating pseudo-labels for training, and their performance is significantly constrained by two types of uncertainty factors: 1) Epistemic uncertainty (model's inherent lack of knowledge), leading to unstable pseudo-label generation quality; and 2) Aleatoric uncertainty (inherent noise and ambiguity in the data), making the model susceptible to background interference and distribution bias. To address these challenges, we propose the progressive cognition-guided dual-domain network for semi-supervised crowd counting (PCDNet). It designs a progressive cognition-guided pseudo-label refinement mechanism which employs multi-granularity filtering to effectively eliminate low-quality pseudo-labels and significantly mitigate epistemic uncertainty. Furthermore, it introduces a spatial-frequency dual-domain joint counting loss. By enforcing the model to learn consistent and robust feature representations in both the spatial and frequency domains, this loss effectively constrains aleatoric uncertainty arising from data bias, enhancing the model's generalization capability for complex scenes. Extensive experiments on four mainstream crowd counting datasets demonstrate that PCDNet significantly outperforms existing semi-supervised methods, with advantages being particularly pronounced when labeled data is scarce. The results fully validate the effectiveness of the proposed method in addressing epistemic and aleatoric uncertainty, generating high-quality pseudo-labels, and enhancing model robustness and generalization performance, providing a superior solution for semi-supervised crowd counting.

Key words crowd counting; semi-supervised learning; knowledge distillation; cognitive uncertainty; dual-domain joint counting loss (DDL)

收稿日期: 2025-06-06; **修回日期**: 2025-07-14

基金项目: 国家自然科学基金项目 (62163016, 62462033); 江西省自然科学基金项目（20212ACB202001, 20242BAB25092）; 科技部高端外国专家项目 (G2023022005L); 华东交通大学山区土木工程安全与韧性全国重点实验室开放课题（HJGZ2023203）; 赣鄱俊才支持计划（20242BCE50030)

This work was supported by the National Natural Science Foundation of China (62163016, 62462033), the Natural Science Foundation of Jiangxi Province (20212ACB202001, 20242BAB25092), the Foreign Expert Project of Ministry of Science and Technology(G2023022005L), the Open Project of State Key Laboratory of Safety and Resilience of Civil Engineering in Mountain Area, East China Jiaotong University (HJGZ2023203), and the Ganpo Talent Support Plan (20242BCE50030).

摘　要　在人群计数任务中，高昂的数据标注成本严重制约了全监督方法的广泛应用. 为显著降低对标注数据的依赖，利用大量未标注数据的半监督计数方法已成为当前研究的主流方向. 然而，现有半监督方法通常依赖迭代生成伪标签进行训练，其性能深受2种不确定性因素的制约，其中认知不确定性来自模型自身对知识掌握的不足，易导致伪标签生成质量不稳定；任意不确定性源于数据固有的噪声和歧义，使得模型易受背景干扰和分布偏差的影响. 为应对上述挑战，提出渐进式认知引导的双域半监督人群计数网络（PCDNet）. 该网络设计了渐进式认知引导的伪标签精炼机制，通过多粒度筛选，有效剔除低质量伪标签，缓解认知不确定性干扰；同时，提出联合频域和空间域的双域联合计数损失（dual-domain joint counting loss, DDL），强制模型在空间域和频域学习一致且鲁棒的特征表达，有效约束由数据偏差引发的任意不确定性，增强模型对复杂场景的泛化能力. 在4个主流人群计数数据集上的大量实验表明，PCDNet显著优于现有半监督方法，尤其在标注数据稀缺时优势更为明显. 实验结果充分验证了该方法在应对认知与任意不确定性、生成高质量伪标签以及提升模型鲁棒性和泛化性能方面的有效性，为半监督人群计数提供了更优的解决方案.

关键词　人群计数；半监督学习；知识蒸馏；认知不确定性；双域联合损失

中图法分类号　TP391

DOI: 10.7544/issn1000-1239.202550453　　**CSTR:** 32373.14.issn1000-1239.202550453

人群计数旨在估计图像或视频中人群数量、密度与空间分布，是智能视频监控的核心任务之一[1]，也是后续行为分析[2]、拥塞分析[3]、异常检测[4-5]和事件检测[6]等高级视觉处理任务的基础. 鉴于其在公共安全、交通调度、人流监测与灾害应急响应中的关键作用[7-8]，人群计数技术的发展受到社会各界广泛关注.

近年来，基于深度学习的人群计数方法已成为该领域的主流，并在性能上展现出显著的技术优势[9-21]. 以Wang等人[19]提出的DMCount为例，其采用深度卷积神经网络（convolutional neural network, CNN）作为特征提取器，利用最优传输（optimal transport, OT）理论度量分布匹配度，在多个基准数据集上实现了计数性能的显著提升，充分印证了深度学习方法在该领域的技术优势. 然而，此类方法高度依赖大规模标注样本以构建高精度、高鲁棒性的模型，而在密集人群场景下，高质量标注数据的获取面临人工成本高企的现实困境. 以UCF-QNRF数据集[22]为例，其包含125万个人头实例标注，累计耗费工时达2 000小时.

为降低对高成本、高耗时人工标注的依赖，半监督人群计数方法逐渐兴起. 该类方法利用少量有标注样本训练初始模型，继而基于该模型对无标注样本进行预测并生成候选伪标签，然后从中筛选出高质量伪标签样本加入训练集，通过多轮迭代训练提升模型的计数性能和泛化能力. 由此可见，伪标签的生成质量是决定半监督人群计数模型性能的关键.

为避免低质量伪标签引入训练噪声，每轮迭代均会根据置信度阈值对伪标签进行筛选，仅保留高置信度的预测结果作为增量训练数据，以此实现训练数据集质量的动态提升. 在理想情况下，高置信度应对应高可靠性，即模型对其预测结果有高度的信任，且这种信任是基于可靠的预测结果. 然而，在实际应用中，高置信度并不总是意味着高可靠性. 如图1所示，红色框标注的背景区域由于与人群特征相似，从而生成高置信度但错误的伪标签. 产生这个问题的根源在于伪标签生成质量易受数据内在不确定性的干扰. 这种不确定性可细化为认知不确定性和任意不确定性[23]. 在半监督人群计数模型训练过程中，因自身认知局限，模型容易对错误预测赋予虚高置信度，即认知不确定性会导致模型对训练数据未覆盖的场景产生预测偏差. 与此同时，数据固有的噪声（包括标注误差、背景干扰等）所引发的任意不确定性同样会引入预测偏差. 因此，为提升训练数据质量，需针对性降低这2类不确定性的影响，避免生成“高置信度但不可靠”的伪标签.

在半监督人群计数任务中，受限于有标签训练样本的数量，模型普遍存在较高的认知不确定性. 这种不确定性在训练数据分布尚未覆盖的场景中尤为显著，表现为模型对其预测缺乏可靠的知识支撑，导致预测值与真实值之间存在偏差. 如图2中间图所示，训练初始阶段，模型对图像局部区域的认知不确定性极高，由于尚未学习到识别此类样本的关键知识，因而出现了误判甚至可能生成错误的高置信度伪标

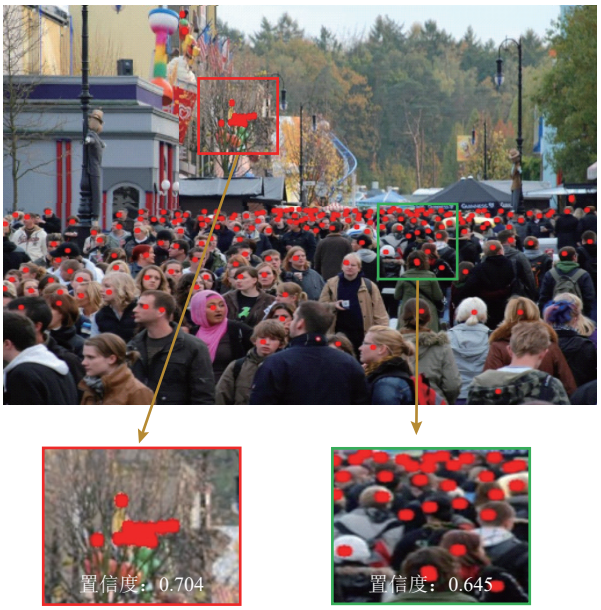


Fig. 1 Schematic diagram of pseudo-label

图 1 伪标签示意图

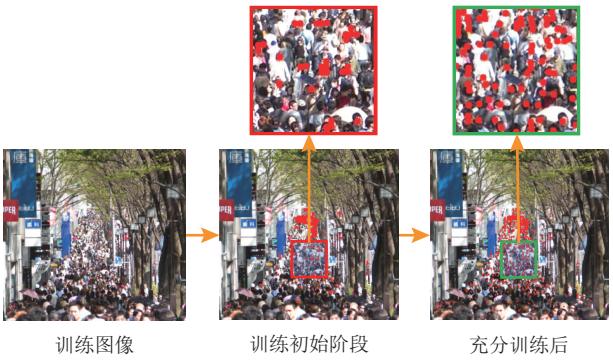


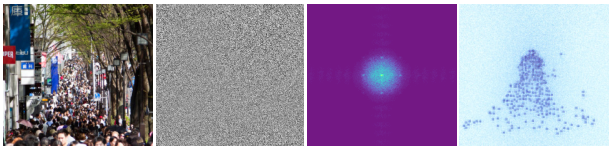
Fig. 2 Progressive cognitive learning of the model

图 2 模型的渐进式认知学习

签. 随着训练样本扩充和迭代的进行，模型积累的知识增加，认知不确定性逐渐降低. 如图 2 右侧图所示，经过充分训练，模型的认知水平显著提升，先前误判的区域最终被准确识别. 因此，要减少伪标签噪声，关键在于抑制模型在高认知不确定性区域的过度自信，确保伪标签不仅“表面合理”，而且“内在可靠”. 由此可知，半监督人群计数模型的认知过程是渐进式的，并随着训练数据的增加而持续完善.

现有方法生成伪标签时通常采用同等式过滤机制[16,24-25]，即基于全局预测置信度选取伪标签参与训练. 该类方法平等地对待图像的所有区域，默认各区域识别难度相同. 然而，在训练数据有限的情况下，对于图像中某些低频出现或者特征复杂的区域，即模型认知不确定性高的区域，模型会因缺乏足够知识而产生认知偏差. 此时，应主动放弃识别这些存在高度认知不确定性的图像块，避免其引入噪声伪标签. 鉴于此，为评估模型对图像不同区域的认知能力，引入基于图像块分析的认知评估模块(cognition assessment module, CAM). 该模块在有标签数据训练阶段，通过度量人群预测分布与真实分布之间的差异来学习图像块特征与认知不确定性之间的映射关系. 然后，在伪标签生成阶段，作为外部监督机构，对无标签图像进行认知能力评估和粗粒度筛选，把对模型来说可能存在识别困难的区域进行遮挡. 同时，进一步联合主网络的像素级细粒度筛选，构建由粗到细的多尺度筛选机制，对可能包含噪声的预测进行逐级剔除. 该机制通过渐进式地扩展模型的认知能力，有效提升伪标签的准确性和可靠性.

尽管 CAM 有效抑制认知不确定性对模型的影响，但是训练阶段模型仍面临数据固有噪声(如标注误差、背景噪声等)所引发的任意不确定性干扰. 现有方法主要通过空间域特征建模[13,19]或引入多模态数据，例如红外图像[26]来缓解此问题，却往往忽略了图像频域信息在刻画图像全局结构与抑制高频噪声方面的天然优势. 文献 [18] 指出，人群分布的空间域与频域表征存在显著差异. 如图 3 所示，图 3(a) 展示空间域人群分布，其呈现离散特性；图 3(b) 和图 3(c) 分别对应的频域相位图和振幅图则显示，人群分布在频域中高度集中，且频谱能量主要集中于低频区域，这反映了人群整体分布的结构一致性. 这一特性在图 3(d) 的相位谱重构图中得到直观印证. 因此，这种低频主导的频域特性为建模人群全局结构提供了新视角，并有助于抑制空间域中由任意不确定性引发的高频噪声干扰. 鉴于此，引入联合频域和空间域的双域联合计数损失(dual-domain joint counting loss, DDL)，旨在提升人群定位精度并抑制噪声干扰. 在该损失函数中，频域损失项约束人群分布的全局结构，抑制噪声引发的高频干扰；而空间域损失项则聚焦局部细节建模，确保人群个体的定位精度以及计数准确性. 通过协同利用空间域与频域对人群信息的互补表征，DDL 能够降低认知不确定性对模型的



(a) 人群空间分布图 (b) 频域相位图 (c) 振幅图 (d) 相位重构图

Fig. 3 Dual-domain crowd information maps

图 3 双域人群信息图

负面影响，从而生成更加鲁棒和稳定的伪标签，为后续半监督训练提供更可靠的监督信号.

综上，为提升半监督人群计数算法的性能，本文提出渐进式认知引导的双域半监督人群计数网络（progressive cognition-guided dual-domain network for semi-supervised crowd counting, PCDNet）. 该网络以教师-学生知识蒸馏框架为核心[27]，教师模型在 CAM 引导下，通过多尺度过滤机制生成高质量伪标签，有效抑制认知不确定性的干扰；学生模型则接受来自教师模型的输出，并通过 DDL 提升人群定位精度，显著降低任意不确定性的影响. 本文的主要贡献有 3 个方面：

1) 提出了端到端的 CAM. 该模块可以利用人群分布的预测值与真实值的差异来评判模型认知能力，从而引导模型从易到难，渐进式地提升认知能力，缓解认知不确定性引发的预测偏差，避免引入噪声伪标签.

2) 提出了 DDL，通过频率域与空间域的联合约束，提升人群定位精度并降低任意不确定性，从而提高模型的鲁棒性与泛化能力.

3) 在 4 个主流人群计数数据集 ShanghaiTech Part_A&Part_B[12]、UCF-QNRF[22]、JHU-CROWD++[28-29] 上与其他先进的人群计数算法进行对比实验，验证了 PCDNet 方法的有效性和优越性.

1 相关工作

1.1 有监督人群计数

早期人群计数主要依赖目标检测[30-31]或回归模型[32-34]来获取人群数量，在稀疏场景有效，但在人群密集场景下因严重遮挡、尺度变化以及背景干扰导致计数性能显著下降. 深度学习方法兴起后，主流的有监督人群计数方法大致可划分为 2 类，即基于密度图估计的方法和基于点定位的方法.

基于密度图估计的方法通过对人群密度图积分求和来估计人群数量[12,35-38]. 该类方法通常使用多列网络[12,35]、多任务学习[36]或扩张卷积[37]解决人群尺度变化问题[12,36-38]，但需要对真实点进行高斯模糊以生成监督信号，导致个体位置信息模糊. 为保留位置精度，文献 [13, 19, 20, 39] 引入 OT 理论度量人群预测分布和真实分布之间的距离，避免了传统高斯平滑过程的局限性. 然而，此类方法对数据噪声高度敏感，难以有效抑制任意不确定性引起的预测偏差.

基于点定位的方法通过回归或分类的方式预测个体坐标，例如 Song 等人[21]提出了一对一的匹配机制优化个体位置预测，以实现精确的人群计数. 与密度图方法相比，此类方法能够提供精确位置信息，在建模个体空间分布方面具有更高的灵活性和可塑性. 然而，在高密度人群场景下基于点定位的方法易因遮挡、背景干扰影响导致漏检率激增. 为缓解高密度场景下定位精度下降的问题，研究人员开始引入频域分析. Shu 等人[18,40]发现人群分布在频域呈低频主导结构，说明频域在建模全局结构和抑制局部噪声方面具有天然优势，因此利用频域信息进行全局约束，辅助空间域实现局部定位.

1.2 半监督人群计数

由于人群图像标注成本高昂，近年来的研究逐步转向半监督人群计数领域，皆在探索低成本的计数解决方案. 半监督范式的核心挑战在于如何有效利用未标记数据生成可靠伪标签，以扩充模型训练数据集. 文献 [17] 通过密度感知的特征解耦将图像区域划分为不同密度等级组，并利用对比学习强化组间特征判别性，为提升伪标签可靠性提供了新思路. Yang 等人[41]利用扩散模型对人群背景进行替换，通过生成式数据增强提升样本多样性，为模型训练提供了更丰富的学习样本，有效增强了模型对复杂场景的认知能力. 然而，扩散模型固有的幻觉特性可能导致生成内容偏离真实场景分布，进而使模型学习到虚假的判别性特征，影响预测准确性. Meng 等人[15]通过师生模型的密度图差异估计像素不确定性，并引入分割图提供场景结构约束. 该方法虽有效改善了伪标签质量，但像素级不确定性估计在人群密集区域易失效，且模型间的输出差异也会引入额外噪声. Lin 等人[16]提出的 OT-M 算法基于最优传输将预测密度图转换为伪点图，期望能够提供更强的监督. 该方法打破了基于密度图的半监督人群计数的局限性（如无法分辨出相邻 2 个过于靠近的人头），但过度依赖模型对未标记数据的预测，容易在高密集高遮挡区域出现过度自信，导致引入噪声伪标签. Li 等人[14]提出在训练阶段以预测分布和真实分布的 Wasserstein 距离[42]作为优化目标，增强模型对整体分布的感知. 尽管该方法在一定程度上考虑了背景噪声对预测结果的干扰，但对于复杂场景中人群细节的捕捉能力仍然不足，导致在背景扰动较强的情况下，伪标签的质量仍受影响. 总之，以上方法虽然在一定程度上提升了伪标签质量，但在高遮挡、密集或复杂背景等场景下，如何生成可靠的伪标签仍是极具挑战性的问题.

2 本文方法

PCDNet的整体架构如图4所示. 该网络以教师-学生知识蒸馏框架为核心，通过渐进式认知学习与双域协同约束，实现对未标记数据的高效挖掘与伪标签质量的精准把控.

PCDNet的训练过程可划分为3个递进阶段. 第1阶段基于有标签数据同步优化学生模型与CAM. 首先通过监督学习训练学生模型的基础计数能力，同时基于当前迭代轮次中主网络预测点位图与真实点位图的分布差异，驱动CAM的端到端学习. CAM通过量化主网络对局部区域的认知表现，从而捕获块级特征中潜在的隐含信息，实现对模型认知不确定性的显式建模. CAM的学习呈渐进式提升，随着训练迭代推进，主网络积累的知识愈发丰富，其对认知不确定性的评估将愈发精准.

第2阶段采用指数移动平均(EMA)策略更新教师模型参数，并由教师模型为学生模型生成稳定的训练样本. 此时CAM转为推理模式，对无标签图像进行块级认知不确定性评估，保留低认知不确定性区域并遮挡高不确定性区域，以降低输入数据的识别难度. 伪标签筛选环节构建“块-像素”多粒度筛选机制，联合教师网络输出的像素级置信度，从块级和像素级2个尺度逐级剔除噪声伪标签. 具体来说，首先由CAM定位认知困难的图像块，再基于像素置信度阈值筛选可靠样本，从而在零额外标注成本下实现伪标签质量的显著提升.

第3阶段用教师模型生成的高质量伪标签数据扩充训练集，学生模型则基于扩充数据继续训练. 为了抑制任意不确定性干扰，引入DDL，通过空间域与频率域的跨域一致性约束，强化模型对透视畸变、密集遮挡等复杂场景的适应性. 其中，空间域损失直接约束人群位置与数量等局部细节，频域损失则通过捕捉人群分布的全局结构特征(如密度波动规律、聚集模式)，补充模型对宏观人群趋势的理解. 二者联

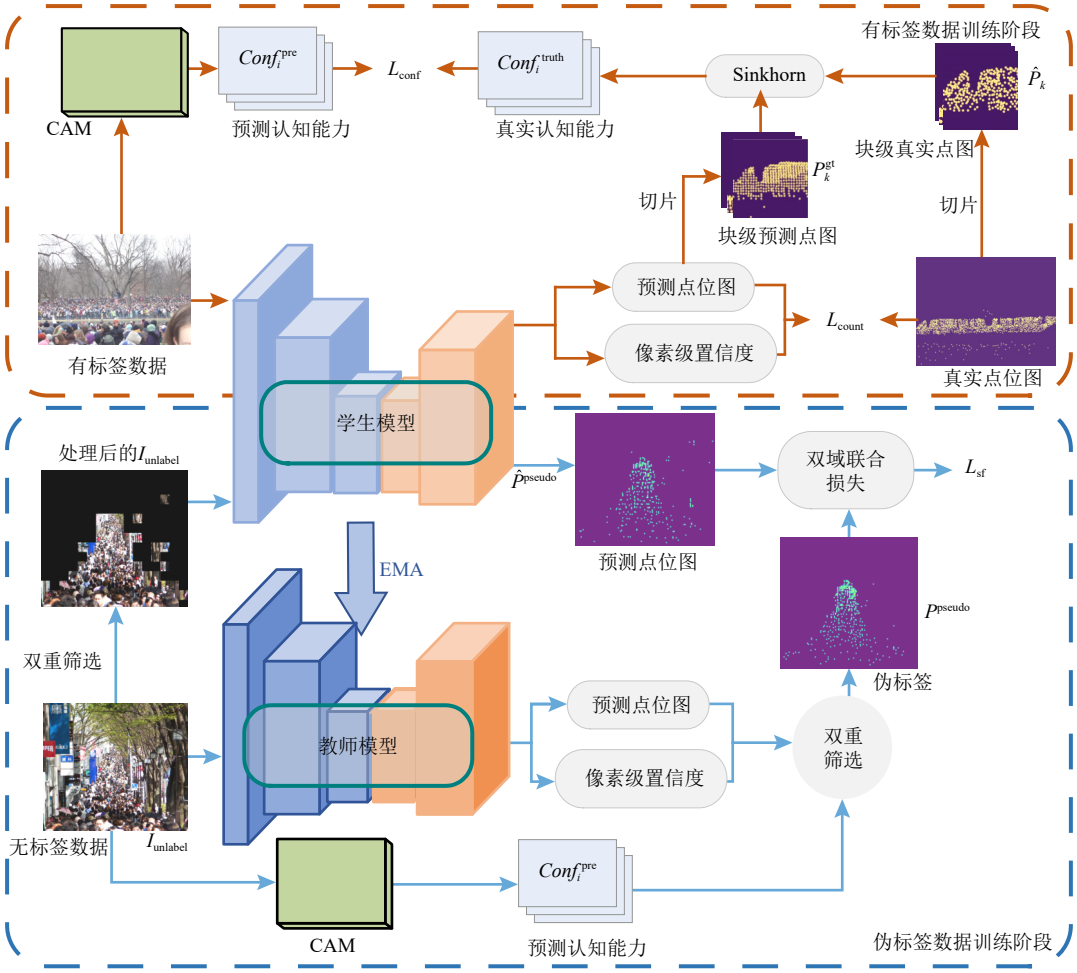


Fig. 4 System architecture diagram of PCDNet

图4 PCDNet系统架构图

合建模有效提升了模型在高密度遮挡、复杂聚集区域等挑战性场景下的计数鲁棒性，同时进一步净化伪标签生成过程.

2.1 CAM

CAM 的结构如图 5 所示，作为 PCDNet 的核心模块，其作用贯穿于模型训练全流程，并在不同阶段承担不同角色.

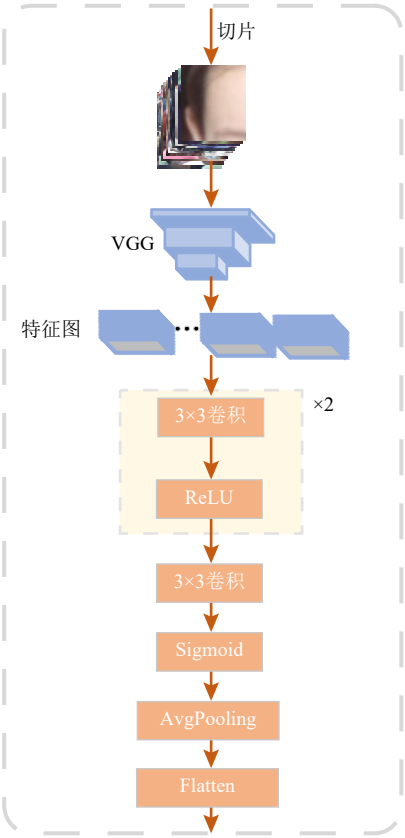


Fig. 5 Architecture of CAM module

图 5 CAM 模块结构

在训练第 1 阶段，CAM 基于有标签数据进行认知不确定性建模. 通过比对主网络预测点位图与真实点位图的分布差异，计算局部区域即图像块的认知置信度. 将模型的“隐性认知缺陷”转化为“显性不确定性分数”，为后续筛选提供量化依据，具体计算过程为：

给定一张包含 N 个个体的人群图像 $\boldsymbol{I}\in\mathbb{R}^{H\times W}$，设第 i 个个体的中心坐标为 $p_i=(x_i,y_i)$，则所有个体中心点集合可以表示为：$P^{\mathrm{gt}}=\{p_i|i\in\{1,2,\cdots,N\}\}$. 主网络对图像的预测输出包含 2 个部分，即预测个体中心点集合 $\hat{P}=\{\hat{p}_j|j\in\{1,2,\cdots,M\}\}$ 以及置信度集合 $\hat{C}=\{\hat{c}_j|j\in\{1,2,\cdots,M\}\}$，其中 M 为预测人群数量，$\hat{c}_j$ 表示点 $\hat{p}_j$ 的像素级置信度且 $\hat{c}_j\in[0,1]$. 将输入图像划分成 S 个大小为 $K\times K$ 的图像块，其中 $S=\lfloor H/K\rfloor\times\lfloor W/K\rfloor$. 对真实中心点集合 P^{gt} 和预测中心点集合 $\hat{P}$ 同步划分，同时剔除各图像块中置信度低于阈值 δ 的预测点，则第 k 个图像块 $Block_k$ 中真实中心点集合为 $P_k^{\mathrm{gt}}=\{p_i|p_i\in Block_k\}$，真实中心点个数为 N'，以及过滤后的预测中心点集合为 $\hat{P}_k=\{\hat{p}_j|\hat{p}_j\in Block_k\}$，预测中心点个数为 M'.

通过计算 P_k^{gt} 和 $\hat{P}_k$ 之间的 Sinkhorn 距离[43] 可以得到第 k 个图像块的人群真实分布和预测分布之间的差异 $L(P_k^{\mathrm{gt}},\hat{P}_k)$，计算方法如式(1)所示.

$$L(P_k^{\mathrm{gt}},\hat{P}_k)=\min_{\gamma\in U(a,b)}\left(\sum_{i\in N',j\in M'}\gamma_{ij}C_{ij}-\varepsilon H(\boldsymbol{\gamma})\right),\tag{1}$$

$$H(\boldsymbol{\gamma})=-\sum_{i\in N',j\in M'}\gamma_{ij}(\log\gamma_{ij}-1),\tag{2}$$

$$C_{ij}=\left\|p_i-\hat{p}_j\right\|_2^2,\tag{3}$$

其中 $\boldsymbol{\gamma}$ 是传输矩阵，表示人群真实点位分布迁移到预测点位分布的最优传输方案. 矩阵元素 γ_{ij} 是从源点 p_i 到目标点 $\hat{p}_j$ 的传输质量，其满足边际约束 $U(a,b)$，使得 $\boldsymbol{\gamma}\mathbf{1}_{M'}=a$ 和 $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{1}_{N'}=b$，其中 $\mathbf{1}_{N'}$ 和 $\mathbf{1}_{M'}$ 表示长度为 N' 和 M' 的全为 1 的行向量和列向量. a 和 b 分别是真实点位集合和预测点位集合的质量分布，且 $a=\sum_{i\in N'}a_i=\sum\frac{1}{N'}=1$，$b=\sum_{i\in M'}b_i=\sum\frac{1}{M'}=1$. C_{ij} 是成本矩阵，用于衡量从真实点位 p_i 到预测点位 $\hat{p}_j$ 的传输成本，此处采用欧几里得距离来度量，如式(3)所示. $-\varepsilon H(\boldsymbol{\gamma})$ 是熵正则化项，其中 $H(\boldsymbol{\gamma})$ 是传输矩阵 $\boldsymbol{\gamma}$ 的信息熵，$\boldsymbol{\gamma}$ 分布越均匀，信息熵越大，点与点之间强对齐的概率越小. 在实际的训练过程中，传输矩阵 $\boldsymbol{\gamma}$ 分布越均匀，模型的性能越好. 因为传输矩阵 $\boldsymbol{\gamma}$ 均匀分布可以避免点与点之间的强对齐，这样一个真实点能够接受多个预测点的“部分匹配”，从而增强模型的泛化能力. 熵正则化参数 ε 负责调整信息熵的影响程度，ε 越小，信息熵的影响越小，最终结果受成本矩阵的影响越大.

在 Sinkhorn 距离的迭代过程中，设核矩阵为 $K_{ij}=\exp(-C_{ij}/\varepsilon)$，$\boldsymbol{K}$ 将成本矩阵的空间距离转化为数值权重，指导模型将更多质量分配给那些空间上贴近的点对，从而实现局部人群分布的精细对齐. 令 $\boldsymbol{u}^{(0)}=\mathbf{1}_{N'}$（长度为 N' 的全 1 向量），$\boldsymbol{v}^{(0)}=\mathbf{1}_{M'}$（长度为 M' 的全 1 向量），不断迭代 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 直到其收敛，其中 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的迭代优化公式为：

$$\boldsymbol{u}^{(k+1)}=\frac{a}{\boldsymbol{K}\boldsymbol{v}^{(k)}},\tag{4}$$

$$\boldsymbol{v}^{(k+1)}=\frac{b}{\boldsymbol{K}^{\mathrm{T}}\boldsymbol{u}^{(k+1)}},\tag{5}$$

通过不断迭代，便可获得最优传输计划$\boldsymbol{\gamma}$，如式(6)所示，其中$\boldsymbol{u}^*$和$\boldsymbol{v}^*$是收敛后的向量，以及最优Sinkhorn距离$L(P_k^{\mathrm{gt}},\hat{P}_k)$.

$$\boldsymbol{\gamma}=\mathrm{diag}(\boldsymbol{u}^*)\boldsymbol{K}\mathrm{diag}(\boldsymbol{v}^*), \tag{6}$$

由于Sinkhorn距离默认人群预测数量与真实数量相等，而实际上模型很少能完全准确地预测人群实际数量，因此需要考虑点集数量不一致对置信度的影响，为此在计算第k个图像块的真实置信度$Conf_k^{\mathrm{tru}}$时，增加对P_k^{gt}和$\hat{P}_k$点集数量差异的惩罚：

$$Conf_k^{\mathrm{tru}}=\frac{1}{\exp(L(P_k^{\mathrm{gt}},\hat{P}_k)+\partial|Num(P_k^{\mathrm{gt}})-Num(\hat{P}_k)|)}, \tag{7}$$

其中，$\partial>0$是数量差异惩罚项的权重系数，用于平衡距离差异与数量差异对置信度的负面影响. 当P_k^{gt}和$\hat{P}_k$的分布越匹配且P_k^{gt}和$\hat{P}_k$的数量越接近时，置信度越高，模型的认知可靠性越高；反之置信度越低，模型的认知可靠性越低.

在获得模型真实置信度的同时，需要同步训练CAM，使之具备判别模型可靠性的能力，即CAM可以为输入图像块生成预测置信度. 计算置信度损失L_{conf}时，不仅要求置信度数值保持一致，还要求置信度排序也保持一致性，如式(8)所示. 其中，L_{reg}为回归损失，用于量化真实置信度与预测置信度的数值差异，采用均方误差(MSE)以增强对异常值的鲁棒性，其中$Conf_i^{\mathrm{pre}}$和$Conf_i^{\mathrm{tru}}$分别为第i个图像块的预测置信度和真实置信度.L_{rank}为排序一致性损失，Ω为所有满足$Conf_i^{\mathrm{tru}}>Conf_j^{\mathrm{tru}}$的样本对集合，$\alpha$为排序敏感强度. 当真实排序为$Conf_i^{\mathrm{tru}}>Conf_j^{\mathrm{tru}}$时，若预测满足$Conf_i^{\mathrm{pre}}>Conf_j^{\mathrm{pre}}$，损失趋近0；反之，若预测$Conf_i^{\mathrm{pre}}<Conf_j^{\mathrm{pre}}$，则损失显著增加.$\lambda$为排序损失权重，用于平衡数值差异与排序一致性对置信度的影响. 若更关注置信度的绝对数值，可减小λ；反之，若更关注相对顺序，则增大λ.

$$L_{\mathrm{conf}}=L_{\mathrm{reg}}+\lambda L_{\mathrm{rank}}, \tag{8}$$

$$L_{\mathrm{reg}}=\frac{1}{S}\sum_{i=1}^{S}\left|Conf_i^{\mathrm{pre}}-Conf_i^{\mathrm{tru}}\right|, \tag{9}$$

$$L_{\mathrm{rank}}=\frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\ln(1+\mathrm{e}^{-\alpha(Conf_i^{\mathrm{pre}}-Conf_j^{\mathrm{pre}})(Conf_i^{\mathrm{tru}}-Conf_j^{\mathrm{tru}})}). \tag{10}$$

在训练第2阶段，CAM将作为独立评估模块参与伪标签的生成和筛选. 它会对输入的无标签图像块进行认知不确定性打分. 保留低不确定性的图像块，即模型已“理解”的区域，作为有效无标签样本参与训练. 同时，遮挡高不确定性的图像块，即让模型“困惑”的复杂区域，避免噪声信息干扰模型学习，从而降低无标签数据的学习难度，提升训练效率.

在第3阶段双域联合优化中，CAM的不确定性信号间接引导模型聚焦于高价值样本，减少伪标签噪声对模型更新的干扰.

2.2 双域联合计数损失（DDL）

已有研究表明[18]，空间域特征虽能直接表征个体精确位置与局部密度分布，但在遮挡严重或者人群密集的场景中，其泛化能力显著受限，且难以建模全局分布规律. 与此互补的是，频域特征通过频谱分解呈现层次化特性，高频分量对应边缘细节但对噪声敏感，低频分量承载整体结构且具有强抗干扰性. 为突破单一空间域建模的局限性，本文提出了DDL机制，如图6所示，通过协同优化空间域和频域特征，使模型兼具细节定位能力与结构鲁棒性，从而提升计数性能.

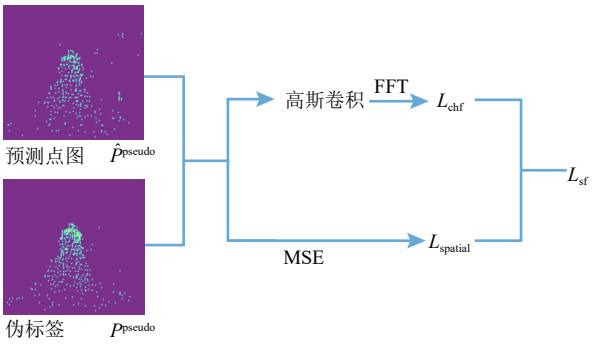


Fig. 6 Calculation process of DDL

图6 DDL的计算过程

根据卷积定理，高斯核卷积在频域等效于低通滤波器. 当人群点标注图经过高斯核卷积转换至频域后，低频分量(人群宏观结构)被保留，而高频分量(如标注误差、光照扰动以及边缘等细节)则被抑制. 值得注意的是，高斯滤波虽可增强噪声鲁棒性，但会模糊个体空间位置信息. 因此联合空间域特征可有效弥补频域转换中丢失的边缘、空间位置等细节信息，形成双域互补优势.

此外，频域中相位分量决定信号的空间结构，幅值分量反映能量. 由于高斯核傅里叶变换相位为零，高斯核卷积后相位仍继承原始点位图的相位信息，故相对位置得以保留，仅绝对位置的高频细节受损，这也再次说明频域有必要联合空间域对定位细节进行补偿.

假设无标签图像I_{unlabel}经过多粒度筛选得到伪标签$P^{\mathrm{pseudo}}=\{p_t|t\in\{1,2,\cdots,T\}\}$，而学生模型的预测结果为$\hat{P}^{\mathrm{pseudo}}=\{\hat{p}_n|n\in\{1,2,\cdots,W\}\}$. 此时空间域和频域的联合损失为：

$$L_{\mathrm{sf}} = l_{\mathrm{spatial}} + \eta \cdot l_{\mathrm{chf}}, \tag{11}$$

其中l_{spatial}为空间域损失，主要用于确保定位精准，具体计算方法为：

$$l_{\mathrm{spatial}} = \left\| P^{\mathrm{pseudo}} - \hat{P}^{\mathrm{pseudo}} \right\|_2. \tag{12}$$

l_{chf}为频域损失，负责强化结构一致性.η为频域损失权重，用于平衡空间域和频域的影响.

假设学生模型预测人群点位图经过傅里叶变换后的实部和虚部分别为Re_{pre}和IM_{pre}，伪标签的实部和虚部分别为Re_{persudo}和IM_{persudo}，则频域损失可定义为：

$$l_{\mathrm{chf}} = \left\| \sqrt{(Re_{\mathrm{pre}} - Re_{\mathrm{persudo}})^2 + (IM_{\mathrm{pre}} - IM_{\mathrm{persudo}})^2} \right\|_2. \tag{13}$$

由于高斯滤波已经滤除高频噪声，因此频域损失l_{chf}可有效约束相位与幅值差异.

3 实验与分析

3.1 实验数据集

为了评估算法的有效性，在4个主流人群计数基准数据集上进行了实验，数据集基本信息如表1所示. 这些数据集均为人群计数研究领域中的热门对比数据集，其在图像分辨率、拍摄场景以及人群密集度等方面均具有较大差异.

ShanghaiTech数据集是Zhang等人[12]于2016年提出了一个大型人群研究数据集，该数据集分为Part_A和Part_B两个子集，由1 198幅图像组成，共有330 165个头部进行了注释. ShanghaiTech Part_A包含从互联网上挑选的482张人群密度较高的照片，其中300张作为训练集，182张作为测试集；ShanghaiTech Part_B包含了从上海地区监控摄像头拍摄的716张照片，人群密度相对稀疏，其中400张作为训练集，316张作为测试集.

UCF-QNRF数据集[22]因具有场景丰富、视角多样、高密度、多尺度等特点，是一个极具挑战性的人群计数数据集. 此数据集的图像是从不同网站上收集的，共标注了1 535幅图像，其中训练集有1 201幅图像，测试集有334幅图像，标注总人数达到1 251 642.

JHU-CROWD++数据集[28-29]因具有人群密集差距大、场景多样化、极端天气环境的特点，是一个大规模无约束人群计数数据集，同时数据集在图像级别和头部级别均包含丰富的注释. 包含4 372张图像和151万个注释. 该数据集的2 272张图像用于训练，500张图像用于验证，同时数据集在图像级别和头部级别均包含丰富的注释.

训练过程中采用了[0.7, 1.3]的随机缩放策略，对输入图像进行尺度变换，并裁剪大小为128×128的图像块；同时，对这些图像块以0.5的概率进行随机翻转. 对于较大分辨率的图像数据集（例如UCF-QNRF和JHU-CROWD++），对图像进行了重新缩放处理，使其最大尺寸不超过特定长度. 具体而言，与文献[21]中的设置基本保持一致，对于UCF-QNRF和JHU-CROWD++数据集，分别将图像最大长度设置为1 408和1 430，并在保证长宽比的前提下进行等比例缩放.

为了度量模型的性能，使用平均绝对误差（mean absolute error, *MAE*）和均方误差（mean square error, *MSE*）作为评价指标. 其中，*MAE*用来评估模型的准确性，*MSE*则用来评估模型的鲁棒性. *MAE*和*MSE*的取值越小，表示模型的性能越好，其计算公式为：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|C_i^{\mathrm{P}} - C_i^{\mathrm{GT}}\right|, \tag{14}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(C_i^{\mathrm{P}} - C_i^{\mathrm{GT}}\right)^2, \tag{15}$$

其中，N表示测试数据集中的图像总数量，C_i^{P}表示第i张图片的预测人数，C_i^{GT}表示第i张图片中的真实人数.

3.2 实验设置

参与对比的模型均基于PyTorch深度学习框架实现，运行环境为Windows 10，所使用的显卡为配备了24 GB显存的NVIDIA-3090.

为确保实验结果具有较高的公平性，我们采用了VGG16模型作为特征编码器，并使用PyTorch官方提供的VGG16_bn预训练权重. 模型训练使用64×64大小的图像块进行不确定估计，基于Adam[44]方法并使用大小为8的批量进行训练，特征提取器

Table 1 Experimental Datasets

表1 实验数据集

数据集	训练集	测试集	标注数量	特点
ShanghaiTech Part_A[12]	300	182	241 677	任意视角、人群密度较高
ShanghaiTech Part_B[12]	400	316	88 488	摄像头拍摄、人群密度稀疏
UCF-QNRF[22],	1 201	334	1 251 642	人群高度密集、视角、环境、光照多样化
JHU-CROWD++[28-29]	2 272	1 600	151万	人群密度差距大、场景光照多样化、恶劣天气

参数的学习率为 1E−5，频域损失权重η根据经验得出为 0.2，以下参数和文献 [14] 保持一致，∂设为$64\sqrt{2}$，排序损失权重λ设为 0.7，像素级置信度筛选阈值δ设为 0.5，排序敏感度σ为 1，正则化参数ε与文献 [20] 一致，设为 0.01.

3.3 实验结果分析

在实验中，数据集包含有标签数据的比例分别为 5%、10% 以及 40% 这 3 种情况，其实验结果分别对应表 2、表 3 以及表 4. 从实验结果可见，在半监督人群计数任务中，随着标签量的减少，模型初始的认知能力以及后续的推理能力均会受到影响.

在 ShanghaiTech Part_A 数据集上，PCDNet 模型在 40% 和 10% 标签量下的 *MAE* 均取得了 SOTA 水平，*MSE* 则分别取得了最优和次优. 与表现最差的 L2R[46] 网络相比，PCDNet 的性能有了较大的提升，在 40% 标签量上 *MAE* 和 *MSE* 分别提升了 26.3 个百分点和 25.3 个百分点；在 10% 标签量上分别提升 20.2 个百分点和 24.2 个百分点. 但 PCDNet 模型在 5% 数据集下性能表现较差，说明 PCDNet 模型在一般人群密度场景下，如果标签量少于 5%，可能会导致模型无法学到足够的知识来生成稳定的认知能力. 总的来看，PCDNet 模型在相对密集的 ShanghaiTech Part_A 数据集上，相较于较新的模型 CU[14]、DACount[17] 和 OT-M[16] 而言，表现出了明显的优势.

在 ShanghaiTech Part_B 数据集上，PCDNet 仅在 5% 标签量时 *MAE* 指标取得了最优值，其余标签量下性能均表现欠佳. 其根本原因在于，在固定视角的稀疏场景中，含有人群的图像块中的人群数量差异不明显，且空块图像远多于其他数据集，背景特征差异不明显，此时数据多样性远少于其他数据集，PCDNet

Table 2 Experimental Results Comparison of 5% Labeled Datasets

表 2 5% 有标签数据集实验结果对比 %

方法	期刊/会议	ShanghaiTech Part_A		ShanghaiTech Part_B		UCF-QNRF		JHU-CROWD++	
		MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
MT[45]	2017 NIPS	104.7	156.9	19.3	33.2	172.4	284.9	101.5	363.5
L2R[46]	2018 CVPR	103.0	155.4	20.3	27.6	160.1	272.3	101.4	338.8
DREAM[47]	2022 TNNLS	112.7	165.7	19.8	30.3	173.4	279.8		
GP[48]	2020 ECCV	102.0	172.0	15.7	27.9	160.0	275.0		
DACount[17]	2022 ACM MM	85.4	134.5	12.6	22.8	120.2	209.3	**82.2**	**294.9**
OT-M[16]	2023 CVPR	86.0	**132.7**	12.8	**22.0**	123.6	208.9	82.7	307.8
CU[14]*	2023 ICCV	**82.2**	137.0	12.0	23.1	126.2	201.1	83.6	300.4
PCDNet（本文）		83.9	142.9	**11.2**	24.7	**116.7**	**197.4**	83.4	303.9

注：*根据公开源代码复现的结果. 黑体数值表示最优值.

Table 3 Experimental Results Comparison of 10% Labeled Datasets

表 3 10% 有标签数据集实验结果对比 %

方法	期刊/会议	ShanghaiTech Part_A		ShanghaiTech Part_B		UCF-QNRF		JHU-CROWD++	
		MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
MT[45]	2017 NIPS	319.3	94.5	15.6	24.5	156.1	145.5	250.3	90.2
L2R[46]	2018 CVPR	90.3	153.5	15.6	24.4	148.9	249.8	87.5	315.3
AL-AC[49]	2020 ECCV	87.9	139.5	13.9	26.2				
IRAST[50]	2020 ECCV	86.9	148.9	14.7	22.9				
IRAST+SPN[50]	2020 ECCV	83.9	140.1						
MTCP[51]	2023 TNNLS	81.3	130.5	14.5	22.3				
SSCC[52]	2023 TCSVT	72.5	118.2	11.7	23.8	123.7	200.9	103.1	324.7
DACount[17]	2022 ACM MM	74.9	**115.5**	11.1	19.1	109.0	187.2	75.9	282.3
OT-M[16]	2023 CVPR	82.9	116.4	**10.9**	**18.2**	109.0	184.5	75.5	287.9
CU[14]*	2023 ICCV	75.2	125.0	11.8	23.9	108.0	178.0	76.7	**280.3**
PCDNet（本文）		**72.0**	116.2	11.4	21.8	**106.5**	**173.5**	**75.3**	280.8

注：*根据公开源代码复现的结果. 黑体数值表示最优值.

Table 4 Experimental Results Comparison of 40% Labeled Datasets

表 4 40% 有标签数据集实验结果对比 %

方法	期刊/会议	ShanghaiTech Part_A		ShanghaiTech Part_B		UCF-QNRF		JHU-CROWD++	
		MAE	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*	*MAE*	*MSE*
MT[45]	2017 NIPS	88.2	151.1	15.9	25.7	147.2	249.6	121.5	388.9
L2R[46]	2018 CVPR	86.5	148.2	16.8	25.1	145.1	256.1	123.6	376.1
GP[48]	2020 ECCV	89.0				136.0			
IRAST[50]	2020 ECCV					138.9			
SUA[15]	2021 ICCV	68.5	121.9	14.1	20.6	130.3	226.3	80.7	290.8
DACount[17]	2022 ACM MM	67.5	**110.7**	9.6	14.6	91.1	**153.4**	**65.1**	**260.0**
OT-M[16]	2023 CVPR	70.7	117.0	9.0	14.2	97.2	162.5	66.9	278.0
CU[14]*	2023 ICCV	69.0	116.4	**8.6**	16.6	95.6	153.6	66.59	277.43
PCDNet（本文）		**63.7**	**110.7**	8.8	**13.8**	**90.0**	159.8	65.56	262.16

注：*根据公开源代码复现的结果. 黑体数值表示最优值.

模型仅靠自身足以学习这种简单场景，而 CAM 在这种人群稀疏场景中对模型的认知引导有限，限制了模型的自我学习能力. 随着标签量的升高，CAM 对于模型的限制影响逐渐加大，从而导致模型性能降低.

在 UCF-QNRF 数据集上，PCDNet 与近年流行的人群计数算法相比，整体性能表现突出. 例如与 DACount[17] 相比，PCDNet 在 40% 标签量时 *MAE* 指标提升 1.2 个百分点，*MSE* 降低 4.17 个百分点；与模型 CU[14] 相比，PCDNet 在 10% 标签量下 *MAE* 提升 1.38 个百分点，*MSE* 提升 2.52 个百分点. 与模型 OT-M[16] 相比，PCDNet 在 5% 标签量下 *MAE* 提升 1.43 个百分点，*MSE* 降低 1.02 个百分点.

在 UCF-QNRF 数据集上，相较于近期的模型 CU[14]、DACount[17]、OT-M[16] 来说，PCDNet 表现出了明显的优势. 随着标签量的上升，CAM 引导模型对高密集人群区域进行渐进式学习，同时 DDL 函数在这种正确的认知能力指导下，有效抑制任意不确定性的干扰，从而提升了计数精度. 由该数据集上的表现也可知，PCDNet 更适合于高密集场景的人群计数，更容易培养良好的认知评估能力.

在 JHU-CROWD++ 数据集上，PCDNet 模型在 10% 标签量以及 40% 标签量时表现次优，但在 5% 标签量时表现较差. 其原因是 PCDNet 随着标签量的减少，CAM 可训练图像块含有的极端环境数据比率逐渐降低，模型无法获得良好的认知指导能力，不能为伪标签的生产提供有效的认知指导，使模型在极端环境下预测人群分布时容易出现偏差. 随后，FSL 对这种错误的结构性分布对齐，从而进一步放大误差，导致模型性能降低. 但从整体上来看，PCDNet 在该数据集上依然取得了具有竞争性的实验效果.

总之，PCDNet 在 4 个半监督人群计数基准任务中均展现出优越的计数性能. 尽管在部分数据集上表现稍次，但在高密度及复杂场景下展现出明显优势，验证了认知评估模块以及 DDL 在模型认知能力有限时能够在复杂背景干扰下有效降低不确定性并生成可靠伪标签的能力，从而提升了整体模型的泛化性能与鲁棒性. 然而，在稀疏人群场景下，该方法的计数性能相对较弱. 究其原因，可能源于认知评估模块的设计机制，其在处理稀疏人群图像时，由于局部图像块中预测人群数量和真实人群数量的差异更为明显，使得模型认知能力容易被低估，导致该图像块被错误遮盖，模型从而被错误引导，最终影响计数性能. 这一观察揭示了当前模块在处理低密度场景时的局限性，也为后续改进提供了潜在方向.

为了形象地说明模型的计数性能，本节展示了 PCDNet 的部分预测结果，如图 7 所示. PCDNet 方法在大规模游行、集会、演唱会以及大型比赛等人群密集场景中，预测的人群分布点位图与真实的人群分布点位图相近，说明 PCDNet 具有较强的泛化能力，可以很好地满足高密度人群以及恶劣环境场景的计数需求. 并且由于基于点的人群计数算法具有较强的定位能力，也可以很好地为下游任务提供支持.

3.4 消融实验

为了验证 PCDNet 所提出模块的有效性，我们在保留了 40% 有标注数据的 Shanghai Tech Part_A 数据集上进行了消融实验，并且采用 *MAE* 与 *MSE* 这 2 种评价指标对模型性能进行评估.

消融实验采用 CU[14] 作为 Baseline，实验结果如表 5 所示. 可知，当引入 CAM 辅助模型逐步提升认知能力后，相较于 Baseline，PCDNet 的 *MAE* 和 *MSE*

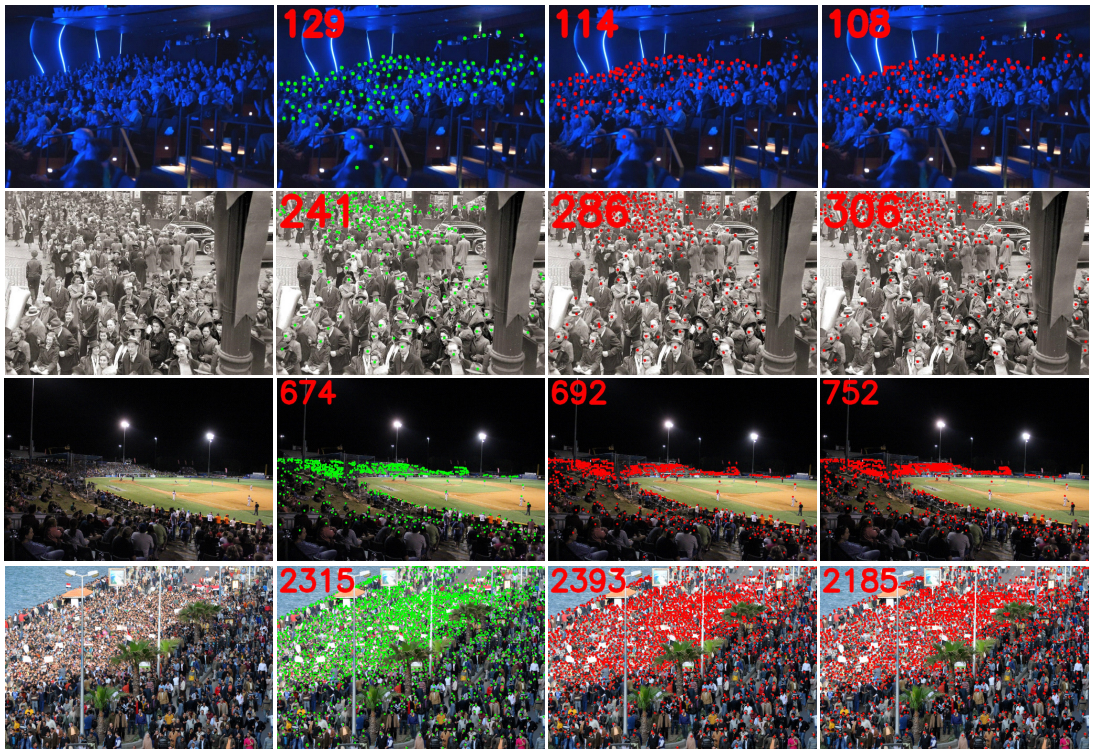


Fig. 7 Visual analysis

图 7 可视化分析

指标分别提升了 3.04 个百分点和 2.31 个百分点. 由此可见，引入 CAM 可以显著提升了模型的计数准确性及稳定性，反映出其在降低模型认知不确定性从而改善伪标签质量方面具有良好的效果.

Table 5 Validity Verification of CAM and DDL

表 5 CAM 和 DDL 有效性验证

Baseline（CU[14]）	CAM	DDL	*MAE*	*MSE*
√			69.0	116.4
√	√		66.9	113.4
√		√	66.7	119.1
√	√	√	63.7	110.7

当单独引入 DDL 时，相较于 Baseline，PCDNet 的 *MAE* 指标提升了 3.33 个百分点，*MSE* 指标下降 2.1 个百分点，说明抑制任意不确定性的影响有助于提升计数准确性. *MSE* 指标下降可能是因为 Baseline 在训练过程中学习了错误的判别信息使得模型认知出现偏差，DDL 对这种结果进行错误结构性对齐，进一步放大这种误差，导致计数性能不稳定.

当同时引入 CAM 和 DDL 时，PCDNet 的性能得到极大的提升，相较于 Baseline，*MAE* 和 *MSE* 指标分别提升了 4.49 个百分点和 7.05 个百分点，达到了最优值. 由此可见，DDL 在抑制任意不确定性方面具有明显优势，但其效能的充分释放依赖于 CAM 提供的先验筛选. 两者协同作用可在抑制不确定性的同时，有效提升模型的整体性能与鲁棒性.

为了研究频域损失权重参数η对模型性能的影响，分别取不同的η值进行实验，其结果如表 6 所示. 由表可见，PCDNet 对η较为敏感，当η=0.2 时，模型综合性能表现最好.

Table 6 Ablation Study on the Weight of Frequency-Domain Loss

表 6 频域损失权重的消融实验

η	*MAE*	*MSE*
0.1	67.7	112.0
0.2	63.7	110.7
0.3	66.3	112.2
0.4	66.3	108.9

4 结 论

本文针对半监督人群计数中的不确定性干扰问题，提出渐进式认知引导的双域半监督人群计数网络(PCDNet). 通过多粒度认知评估模块动态筛选训

练样本，显著提升伪标签质量. 同时，设计了双域联合计数损失 DDL，通过融合空间域的精确定位与频域的结构鲁棒性，有效缓解复杂背景、遮挡及光照变化引起的任意不确定性干扰. 在 4 个主流人群计数数据集上的实验表明，PCDNet 在减少标注依赖的同时，性能显著优于现有半监督方法. 然后，由于 PCDNet 算法对稀疏场景适应性仍显不足，未来将引入场景感知加权项，动态调整稀疏区域的认知置信度阈值. 此外，双域损失计算复杂度较高，未来将研究频域特征蒸馏技术，实现高精度轻量级模型部署.

作者贡献声明：余鹰负责确定选题，完成论文主体撰写和修订，并指导和督促完成相关文献资料的收集整理以及论文的写作；范在昌负责论文实验和文献收集整理；曾康利提供论文选题指导意见；黄晓辉给出论文实验部分修改意见；苗夺谦提出论文修改意见，审阅与修订了论文.

参 考 文 献

[1] Yu Ying, Zhu Huilin, Qian Jin, et al. A survey on crowd counting based on deep learning[J]. Journal of Computer Research and Development, 2021, 58(12): 2724−2747 (in Chinese)
(余鹰, 朱慧琳, 钱进, 等. 基于深度学习的人群计数研究综述[J]. 计算机研究与发展, 2021, 58(12): 2724−2747)

[2] Saxena S, Brémond F, Thonnat M, et al. Crowd behavior recognition for video surveillance[C]//Proc of the Int Conf on Advanced Concepts for Intelligent Vision Systems. Berlin: Springer, 2008: 970–981

[3] Huang Lida, Chen Tao, Wang Yan, et al. Congestion detection of pedestrians using the velocity entropy: A casestudy of love parade 2010 disaster[J]. Physica A: Statistical Mechanics and its Applications, 2015, 440: 200−209

[4] Li Weixin, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes[J], IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(1): 18−32

[5] Chaker R, AI Aghbari Z, Junejo I N. Social network model for crowd anomaly detection and localization[J]. Pattern Recognition, 2017, 61: 266−281

[6] Benabbas Y, Ihaddadene N, Djeraba C. Motion pattern extraction and event detection for automatic visual surveillance[J]. EURASIP Journal on Image and Video Processing, 2010, 2011(1): 163682

[7] Ma Zhiheng, Wei Xing, Hong Xiaopeng, et al. Bayesian loss for crowd count estimation with point supervision[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6141−6150

[8] Wang Qi, Gao Junyu, Lin Wei, et al. NWPU-Crowd: A large-scale benchmark for crowd counting and localization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(6): 2141−2149

[9] Wan Jia, Luo Wenhan, Wu Baoyuan, et al. Residual regression with semantic prior for crowd counting[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4031−4040

[10] Yang Yifan, Li Guorong, Wu Zhe, et al. Reverse perspective network for perspective-aware object counting[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 4373−4382

[11] Wang Qi, Gao Junyu, Lin Wei, et al. Learning from synthetic data for crowd counting in the wild[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 8198−8207

[12] Zhang Yingying, Zhou Desen, Chen Siqin, et al. Single-image crowd counting via multi-column convolutional neural network[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 589−597

[13] Wan Jia, Liu Ziquan, Chan A B. Chan. A generalized loss function for crowd counting and localization[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1974−1983

[14] Li Chen, Hu Xiaoling, Abousamra S, et al. Calibrating uncertainty for semi-supervised crowd counting[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 16685−16695

[15] Meng Yanda, Zhang Hongrun, Zhao Yitian, et al. Spatial uncertainty-aware semi-supervised crowd counting[C]//Proc of the IEEE Int Conf on Computer Vision . Piscataway, NJ: IEEE, 2021: 15549−15559

[16] Lin Wei, Chan A B. Optimal transport minimization: Crowd localization on density maps for semi-supervised counting[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 21663−21673

[17] Lin Hui, Ma Zhiheng, Hong Xiaopeng, et al. Semi-supervised crowd counting via density agency[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 1416−1426

[18] Shu Weibo, Wan Jia, Tan K C, et al. Crowd counting in the frequency domain[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition . Piscataway, NJ: IEEE, 2022: 19618−19627

[19] Wang Boyu, Liu Huidong, Samaras D, et al. Distribution matching for crowd counting[J]. Advances in Neural Information Processing Systems, 2020, 33: 1595−1607

[20] Lin Hui, Hong Xiaopeng, Ma Zhiheng, et al. Direct measure matching for crowd counting[C]//Proc of Int Joint Conf on Artificial Intelligence. San Francisco: Margan Kaufmann, 2021: 837−844

[21] Song Qingyu, Wang Changan, Jiang Zhengkai, et al. Rethinking counting and localization in crowds: A purely point-based framework[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 3365−3374

[22] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting,

density map estimation and localization in dense crowds[C]//Proc of the European Conf on Computer Vision (ECCV). Berlin: Springer, 2018: 532−546

[23] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision?[C]//Proc of the Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 5574−5584

[24] Zhang Shiwei, Ke Wei, Liu Shuai, et al. Boosting semi-supervised crowd counting with scale-based active learning[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 8681−8690

[25] Qian Yifei, Hong Xiaopeng, Guo Zhongliang, et al. Semi-supervised crowd counting with contextual modeling: Facilitating holistic understanding of crowd scenes[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024: 8230−8241

[26] Tang Haihan, Wang Yi, Chau L P. TAFNet: A three-stream adaptive fusion network for RGB-T crowd counting[C]//Proc of the IEEE Int Symp on Circuits and Systems. Piscataway, NJ: IEEE, 2022: 3299−3303

[27] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint, arXiv: 1503.02531, 2015

[28] Sindagi V A, Yasarla R, Patel V M. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 1221−1231

[29] 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, 2020, 44(5): 2594−2609

[30] Leibe B, Seemann E, Schiele B. Pedestrian detection in crowded scenes[C]//Proc of the 2005 IEEE Computer Society Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2005: 878−885

[31] Lempitsky V, Zisserman A. Learning to count objects in images[C]//Proc of the Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2010: 1324−1332

[32] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 2547−2554

[33] Chan A B, Vasconcelos N, Bayesian poisson regression for crowd counting[C]//Proc of the 2009 IEEE 12th Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2009: 545−551

[34] Chan A B, Vasconcelos N. Counting people with low-level features and bayesian regression[J]. IEEE Transactions on Image Processing, 2011, 21(4): 2160−2177

[35] Babu Sam D, Surya S, Venkatesh Babu R. Switching convolutional neural network for crowd counting[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 5744−5752

[36] Gao Junyu, Wang Qi, Li Xuelong. PCC Net: Perspective crowd counting via spatial convolutional network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(10): 3486−3498

[37] Li Yuhong, Zhang Xiaofan, Chen Deming. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 1091−1100

[38] Zhang Hong, Fan Zizhu, Shi Linrui, et al. A head detection method based on multi-scale feature fusion[J]. Journal of East China JiaoTong University, 2021, 38(2): 116−121(in Chinese)
(张泓, 范自柱, 石林瑞, 等. 一种基于多尺度特征融合的人头计数检测方法研究[J]. 华东交通大学学报, 2021, 38(2): 116−121)

[39] Ma Zhiheng, Wei Xing, Hong Xiaopeng, et al. Learning to count via unbalanced optimal transport[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI , 2021: 2319−2327

[40] Shu Weibo, Wan Jia, Chan A. B, Generalized characteristic function loss for crowd analysis in the frequency domain[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46: 2882−2899

[41] Yang Maochen, Li ZeKun, Zhang Jian, et al. Taste more, taste better: Diverse data and strong model boost semi-supervised crowd counting[C]// Proc of the Computer Vision and Pattern Recognition Conf. Piscataway, NJ: 2025: 24440−24451

[42] Shen Jian, Qu Yanru, Zhang Weinan, et al. Wasserstein distance guided representation learning for domain adaptation[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018: 4058−4065

[43] Cuturi M. Sinkhorn distance: Lightspeed computation of optimal transport[C]//Proc of the Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2013: 2292−2300

[44] Kingma D P, Ba J. Adam: A method for stochastic optimization[C/OL]//Proc of the 3rd Int Conf on Learning Representations (ICLR). San Diego, CA: 2015[2025-05-05]. https://arxiv.org/abs/1412.6980

[45] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results[C]//Proc of the Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 1195−1204

[46] Liu Xialei, Weijer J V, Bagdanov A D. Leveraging unlabeled data for crowd counting by learning to rank[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7661−7669

[47] Gao Jiaqi, Huang Zhizhong, Lei Yiming, et al. Deep rank-consistent pyramid model for enhanced crowd counting[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 36: 299−312

[48] Sindagi V A, Yasarla R, Babu D S, et al. Learning to count in the crowd from limited labeled data[C]//Proc of the 16th European Conf on Computer Vision(ECCV). Berlin: Springer, 2020: 212−229

[49] Zhao Zhen, Shi Miaojing, Zhao Xiaoxiao, et al. Active crowd counting with limited supervision[C]//Proc of the European Conf on Computer Vision(ECCV). Berlin: Springer, 2020: 565−581

[50] Liu Yan, Liu Lingqiao, Wang Peng, et al. Semi-supervised crowd counting via self-training on surrogate tasks[C]//Proc of the 16th

European Conf on computer Vision(ECCV) . Berlin: Springer, 2020: 242−259

[51] Zhu Pengfei, Li Jingqing, Cao Bing, et al. Multi-task credible pseudo-label learning for semi-supervised crowd counting[J]. IEEE Transations on Neural Networks and Learning Systems, 2023: 10394−10406

[52] Wang Xin, Zhan Yue, Zhao Yang, et al. Semi-supervised crowd counting with spatial temporal consistency and pseudo-label filter[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4190−4203



Yu Ying, born in 1979. PhD, professor, PhD supervisor. Her main research interests include multi-label learning, computer vision, and granular computing.

余　鹰, 1979 年生. 博士, 教授, 博士生导师. 主要研究方向为多标记学习、计算机视觉、粒计算.



Fan Zaichang, born in 2001. Master candidate. His main research interests include machine learning and computer vision.

范在昌, 2001 年生. 硕士研究生. 主要研究方向为机器学习、计算机视觉.



Zeng Kangli, born in 1992. PhD, lecturer, master supervisor. His main research interests include artificial intelligence, computer vision, image processing, and pattern recognition.

曾康利, 1992 年生. 博士, 讲师, 硕士生导师. 主要研究方向为人工智能、计算机视觉、图像处理、模式识别.



Huang Xiaohui, born in 1984. PhD, professor, PhD supervisor. His main research interests include deep learning, artificial intelligence, intelligent transportation, and remote sensing image analysis.

黄晓辉, 1984 年生. 博士, 教授, 博士生导师. 主要研究方向为深度学习、人工智能、智能交通、遥感图像分析.



Miao Duoqian, born in 1964. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include machine learning, granular computing, and pattern recognition.

苗夺谦, 1964 年生, 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为机器学习、粒计算、模式识别.