Computer Science and Application 计算机科学与应用, 2025, 15(10), 221-231
Published Online October 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1510262

Hans汉斯

基于多层级语义信息融合的区域一致性半监督人群计数研究

郭禹辰，宋蔓蔓

河北金融学院河北省科技金融重点实验室，河北 保定

收稿日期：2025年9月15日；录用日期：2025年10月17日；发布日期：2025年10月28日

摘　要

面向智慧城市场景的人群计数问题，本文提出一种基于多层级语义特征提取网络的半监督人群计数框架。所提方法以ResNet-34为骨干网络，设计自上而下的多层级语义增强模块，在不同层级间通过门控进行语义信息增强与轻量级特征融合，抑制背景噪声、突出前景人群信息。在无标注样本上引入区域一致性约束，设计半监督框架，提升模型的泛化能力。训练阶段采用先监督训练预热，而后交替迭代更新的策略。在ShanghaiTech数据集和UCF_CC_50数据集上进行了训练和测试，实验结果显示，该框架在ShanghaiTech数据集A和B部分上的MAE分别为65.4和9.2，在UCF_CC_50数据集上的MAE为201.2，算法在不同密度人群与复杂背景场景下均表现出了较好的识别精度，可以为基于视觉的人群计数任务提供高效的解决方案。

关键词

人群计数，多层级语义增强，半监督框架

Research on Regional Consistency Semi-Supervised Crowd Counting Based on Multi-Level Semantic Information Fusion

Yuchen Guo, Manman Song

Science and Technology Finance Key Laboratory of Hebei Province, Hebei Finance University, Baoding Hebei

Received: September 15, 2025; accepted: October 17, 2025; published: October 28, 2025

Abstract

Aiming at the problem of crowd counting in smart city scenes, this paper proposes a semi-supervised

文章引用: 郭禹辰, 宋蔓蔓. 基于多层级语义信息融合的区域一致性半监督人群计数研究[J]. 计算机科学与应用, 2025, 15(10): 221-231. DOI: 10.12677/csa.2025.1510262

crowd counting framework based on multi-level semantic feature extraction network. The proposed method uses ResNet-34 as the backbone network and designs a top-down multi-level semantic enhancement module. Semantic information enhancement and lightweight fusion are performed through gating between different levels to suppress background noise and highlight foreground crowd information. By introducing regional consistency constraints on unlabeled samples, a semi-supervised framework is designed to improve the model's generalization ability. In the training stage, the strategy of supervised training preheating and then alternating iterative updating is adopted. The framework is trained and tested on the ShanghaiTech dataset and the UCF_CC_50 dataset. The experimental results show that the MAE of the framework on the A and B parts of the ShanghaiTech dataset is 65.4 and 9.2, respectively, and the MAE on the UCF_CC_50 dataset is 201.2. The algorithm shows good recognition accuracy in different crowd densities and complex background scenarios. This algorithm can provide an efficient solution for vision-based crowd counting tasks.

Keywords

Crowd Counting, Multi-Level Semantic Enhancement, Semi-Supervised Framework

1. 引言

中国城市化进程已进入快速发展阶段，《中华人民共和国国民经济和社会发展第十四个五年规划纲要》提出2025年中国将进一步加强城市智慧化发展，推动“智慧城市”建设[1]。人群管理、城市安全、交通流量的实时监控成为城市治理的组成部分。人群计数任务能够在提高城市安全管理效率、减少安全事故和提升应急响应能力方面发挥积极作用，涉及安全监控、交通流量管理和社会活动分析等多个领域，为公共安全和应急响应提供有力支持[2]。因此，基于视觉的智能化人群计数技术，作为一种高效、低成本的解决方案，具有研究意义和应用前景。

传统的人工特征提取方法难以满足复杂环境中的高精度需求[3]，深度学习技术的迅速发展为视觉感知任务带来了显著的突破，推动图像识别和理解技术的进步。针对人群计数问题，深度学习方法通过自动学习数据中的特征，有效克服传统方法的局限。Zhang等人提出多列卷积神经网络，分别对不同尺度人群特征进行提取，较好地解决了人群尺度变化问题[4]。为了进一步提升特征提取能力，人群计数模型引入空洞卷积，提升模型的感受野，获取高质量图像底层细节实现人群密度特征提取。但是，现有的人群计数算法仍然面临挑战[5]。首先，图像的背景信息复杂，存在人群密度分布不均等问题，极大地影响了人群计数性能。其次，数据集的多样性和复杂性使得模型在不同场景中的泛化能力不足，当前大多数方法过于依赖标注数据，导致在真实应用中需要大量标注数据的支持。因此，基于注意力机制和层级特征融合的人群计数方法用来解决背景干扰问题[6]。Jiang等人通过级联多任务学习网络，同时进行密度等级和密度图生成任务，有效提取高级语义信息，排除底层特征中的背景干扰，实现高质量密度图的生成[7]。Pan等人将识别稀疏和密集人群作为任务，通过引入注意力机制的方法自适应划分人群密度[8]。Zhang等人利用自注意力机制的特点，结合人群任务中像素之间的相互关系，考虑不同位置特征之间的相关性，通过结合全局注意力和局部注意力机制的优缺点，获取相邻像素和不同位置像素的特征关系，实现不同密度的人群计数[9]。半监督算法通常用来解决训练数据量小导致的过拟合问题，可以利用无标签数据来提升模型的泛化能力[10]。余鹰等人提出一种渐进式认知引导的双域半监督人群计数网络，通过设计一种

伪标签认知机制，筛选高质量的伪标签降低不确定性干扰，利用频域、空间域的联合机制设计计数损失，约束数据偏差引发的计数不确定性，有效增强了模型的泛化能力[11]。王鑫等人基于一致性正则化的半监督方法，通过设计语义扰动的先验知识，增强无标签数据的多样性，引入跨网络的双向统计参数更新机制，有效提升了模型的收敛速度[12]。

综上所述，现有研究在特征提取、多尺度建模、注意力机制以及半监督学习等方面有较多研究，但人群计数任务仍存在诸多不足：一是面对复杂多变的场景时，模型对背景噪声的鲁棒性和对密度变化的适应性依然有限；二是对高质量标注数据的依赖仍然严重，制约了模型在真实应用中的可扩展性。因此，在有限标注的条件下，充分挖掘无标注数据潜在价值，提升模型在多场景下的泛化能力成为关键问题。本文基于点监督方式提出一种基于多层级信息融合的趋于一致性半监督计数网络，由两个分支构成，首先，设计了一种层级间特征信息增强模块，有效提取多尺度信息，聚合不同层级间的特征；其次，根据区域一致性，设计了一种半监督计数框架，与监督分支共享模型参数。在提高计数精度的同时，降低计算成本，增强算法的应用性和实用性。

2. 人群计数数据集

人群计数任务中，数据集的形式与标注方式直接影响模型的设计与训练效果。现有的主流人群计数数据集大多采用密度图标注形式，在图像中的每一个标注点位置生成一个高斯核，将所有高斯核叠加得到密度图。通过对密度图进行积分，可以得到与图像中人数一致的总和。为构建点监督的人群计数框架，需要将原有的密度图形式转换为点坐标形式。直接读取数据集中提供的标注文件，将其作为训练输入的监督信号。在 ShanghaiTech 数据集中，每个人头的标注点原本用于生成密度图，而在点监督场景下，这些标注点被直接保留为坐标，模型的输出可以与真实点进行匹配，如图 1 所示为密度图形式和点集标注形式。

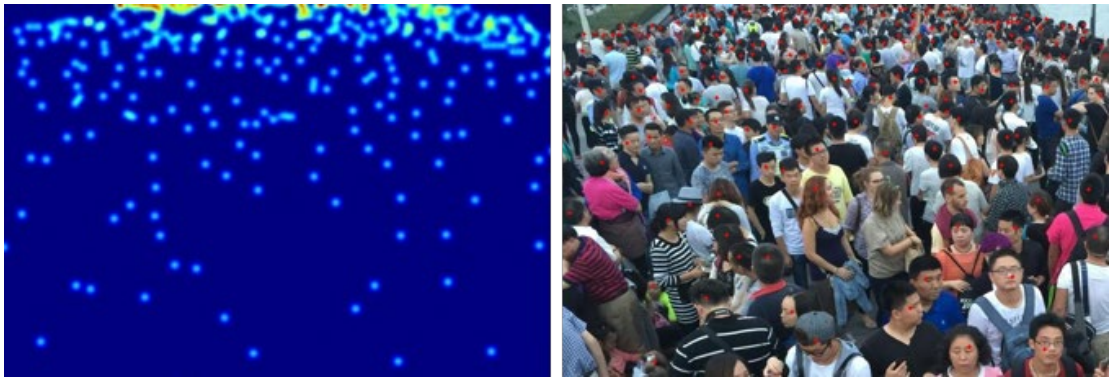


Figure 1. Density map and point supervision
图 1. 密度图与点监督

ShanghaiTech 数据集是人群计数领域最具代表性的数据集之一，包含 1198 张图像，涵盖多种复杂场景和人群密度分布，分为两部分：Part A 和 Part B。Part A 从互联网上收集，主要是高密度人群场景，共 482 张图像，其中 300 张用于训练，182 张用于测试，人数从几百到上千不等，适合密集人群的计数问题[4]。Part B 共 716 张图像，其中 400 张用于训练，316 张用于测试。相比 Part A，Part B 的人群密度较低，接近日常监控和城市交通管理的场景。UCF-CC-50 数据集是高密度小样本数据集，包含 50 张真实场景图像，人数范围由 94 到 4543、平均约 1279 人，总计约 6.4 万个点标注。图像来源涵盖集会、马拉松、演唱会、体育场等复杂场景，遮挡与尺度变化显著[13]。由于样本量极少，通常采用 5 折交叉验证；该数据

集更适合用于检验模型在极端拥挤场景下的鲁棒性，常作为补充测试集而非主要训练集。

3. 人群计数框架

为提升人群计数模型的精度与泛化能力，本文设计一种共享多层级语义特征提取网络的半监督框架。以 ResNet-34 结合层级语义增强模块构成网络结构，对有标注和未标注的数据样本共用同一模型。如图 2 所示，骨干网络设置两路训练分支，其一为点监督计数分支，采用点集预测并通过匈牙利匹配算法建立对应关系，以点作为监督，直接预测人头坐标和人数；其二为区域一致性分支，针对无标注样本以阈值判定得到有效预测点，随机还分若干互不重叠的子区域，多区域联合计数与原图人数相等为标准构建无标注分支损失，与点监督分支构成联合优化框架。在训练策略上，首先用标注数据预热获得稳定的特征提取能力，随后两分支交替混合迭代更新，共享骨干参数在两分支间同步学习。网络设计在结构上利用层级与多尺度上下文增强表达，引入区域一致性设计，同时提升计数精度与跨场景泛化能力，如图 2 所示为半监督人群计数框架图。

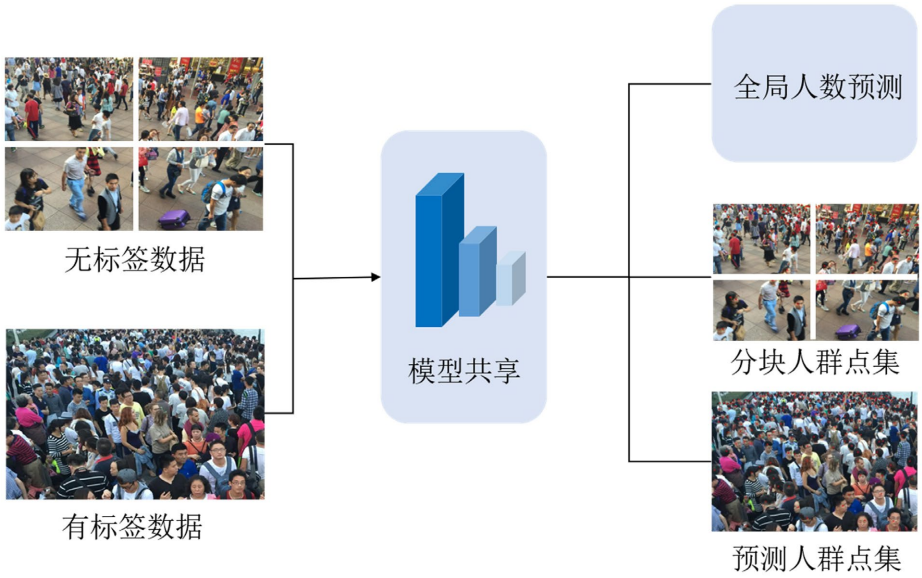


Figure 2. Diagram of the semi-supervised crowd counting framework
图 2. 半监督人群计数框架图

3.1. 点监督人群计数网络

传统人群计数方法大多依赖密度图监督，其核心是将标注点 $\{x_i\}_{i=1}^{N}$ 转换为二维高斯核分布并叠加，得到密度图 $D(x)$，通过约束预测密度 $\hat{D}(x)$ 与真实密度 $D(x)$的差异进行拟合，如式(1)所示：

$$L_{den}=\frac{1}{2}\sum_{x\in I}\left\|\hat{D}(x)-D(x)\right\|_2^2 \tag{1}$$

其中，I 表示图像数据。

然而，密度图的方式对高斯核参数存在依赖，无法保证预测点与真实点一一对应。P2PNet 将人群计数建模为了集合的预测任务，直接从图像中预测出一组点坐标集合，$\hat{Y}=\{\hat{y}_j\}_{j=1}^{M}$，并与真实标注集合 $Y=\{y_i\}_{j=1}^{N}$ 进行匹配，网络 f_θ 将输入图像 I 映射为一组二维坐标预测[14]。整体框架可表示为式(2)：

$$I\xrightarrow{f_\theta}\hat{Y} \tag{2}$$

网络 f_θ 将输入图像 I 映射为一组二维坐标预测，随后，通过匈牙利匹配算法建立预测点与真实点的对应关系，并以点到点的距离作为监督信号，如式(3)所示：

$$L_{p2p} = \frac{1}{N}\sum_{i=1}^{N}\left\| y_i - \hat{y}_{\pi(i)} \right\|_2^2 \tag{3}$$

因此，这种点到点的监督方式不仅避免了人为设计密度图核函数的限制，而且能够显示输出人头位置的集合，使计数与定位功能同时得到了优化，在保持技术准确性的同时，显著提升了定位精度和结果的可解释性，为后续半监督与注意力机制框架的结合提供了建模基础。

3.2. 多层级语义特征提取网络

为兼顾精度与效率，如图 3 所示，本文模型采用 ResNet-34 作为主干网络，在网络的不同阶段提取多尺度特征，记为$\{C_2, C_3, C_4\}$自顶向下逐层级进行融合，最终通过 ASPP (Atrous Spatial Pyramid Pooling)进行多尺度上下文聚合并输出用于计数得到最终结果的高级语义特征。具体而言，先由最高层级的语义特征 C_4 为起点，得到特征图 $T4$，随后按照层级进行自顶向下的信息传递与图像细节恢复，如式(4)所示：

$$T_1 = M\left(Concat\left(Up\left(T_{l+1}\right), C_l\right)\right) \tag{4}$$

其中，Up 表示上采样操作，$Concat$ 为通道拼接操作，M 为语义增强模块。

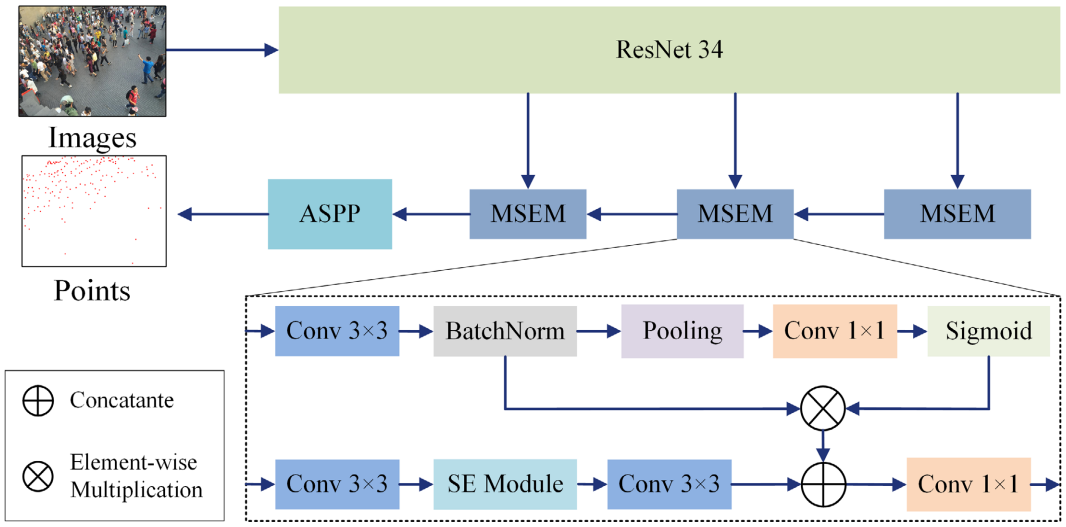


Figure 3. Diagram of the crowd counting model
图 3. 人群计数模型图

经过三次不同层级融合后得到最终特征图后送入 ASPP 模块中，利用 ASPP 模块中空洞卷积的不同采样率获取更大的感受野与多尺度上下文信息，由 ASPP 模块输出的特征图接入后续计数预测头，完成对存在不同尺度人群的图像的统一表征。多层级语义增强模块(Multi-level Semantic Enhancement Module, MSEM)包含两个分支，通过门控操作进行特征融合。设由主干网络输入的特征为 X，先经 3 × 3 卷积与 BatchNorm 提取局部语义信息，再输入到池化操作和 1 × 1 卷积得到通道压缩后的特征表示，最后通过 Sigmoid 生成空间门控图 A，逐元素相乘进行门控操作，最终得到特征图 F，如式(5)所示。另一分支以 3 × 3 卷积、SE 注意力模块和 1 × 1 卷积形成空间语义特征兼具纹理信息的表征。

$$F = A \otimes BN\left(Conv_{3\times3}\left(X\right)\right) \tag{5}$$

然后，与下分支特征进行通道级的拼接操作，通过 1×1 卷积完成轻量融合。多层级语义增强模块在两个分支中同时引入空间注意和通道注意力，自适应学习权重抑制背景干扰，突出人群的前景信息，拼接操作和 1×1 的卷积保证了跨层信息在融合时不消耗过多计算量。在多层级语义增强模块中，以自顶向下的层级设计，网络深层可以使强语义信息和浅层细节对齐并整合，ASPP 在网络末端进一步补充上下文。

3.3. 区域一致性半监督框架

为有效在有限点标注下挖掘无标注数据的信息，本文提出区域一致性半监督框架。整体思路是用一个点预测网络直接输出候选人头点的二维坐标和置信度，随后通过阈值计数规则进行人数统计，只要预测置信度不小于某个阈值，即判定为目标，预测点之间通过半径非极大抑制进行去重后得到最终预测集合。由此得到全图的人数计数 C 如式(6)：

$$\hat{C}(I)=\sum\left\{j\,|\,\hat{p}\geq r\right\} \tag{6}$$

其中，$\hat{y}_j$ 为坐标，$\hat{p}_j$ 为置信度

在此基础上，将输入图像随机划分为若干个互不重叠的子区域，使其在训练过程中进行随机化，保持多尺度采样，对子区域内的有效预测点同样以阈值规则计数[15]。对于每个无标注图像，核心思想是把全图人数转化为等于若干个互不重叠图像子区域人数之和，通过这一守恒规律，转化为可微的训练约束，并与点监督主任务端到端联合优化。设点监督网络 f_θ 输入图像 I 后输出预测点集如式(7)：

$$\hat{S}=f_\theta(I)=\left\{\left(\hat{y}_j,\hat{p}_j\right)\right\}_{j=1}^{M} \tag{7}$$

在模型进行训练过程中，对每个无标注样本随机采样一组无重叠区域划分 R 如式(8)：

$$R=\left(R_m\right)_{m=1}^{PQ} \tag{8}$$

得到子图人数集合如式(9)：

$$\hat{C}(R)=\sum\left\{j\,|\,\hat{p}\geq r,\hat{p}\geq R_m\right\} \tag{9}$$

在标记数据学习阶段，采用多层级提取网络作为半监督框架的主干计数网络，对已标注的人群图像数据集进行监督训练优化模型参数，标注数据的训练阶段与未标注数据集共享同一网络模型，因此，标注数据能够为后续的无标注学习提供稳定的指导。

对于无标注图像，依据整图人数应等于若干个不重叠子区域人数之和的规律，构造一个无需标注的代理任务：在同一张图上划分若干子块，分别估计整图与各子块的人数，并以两者的差异作为无标注损失，从而用数据自身的分布约束网络。该损失直接由全局计数和子块计数求和两部分组成输入，模型在无标注样本上进行优化。整体训练采用交替式迭代，首先进行一定次数的监督学习，使模型获得基本的特征提取能力，然后监督学习与无标注学习循环进行，逐步提升对未标注图像的利用效率与模型性能。

4. 实验与结果分析

4.1. 实验设置

所有实验均在配备 NVIDIA RTX 3060 GPU 设备上进行完成。代码框架选用 PyTorch，实验中的数据集，初始时随机抽取其中的 20%作为有标签数据，其余的 80%作为无标签数据集，初始阶段仅使用有标签数据训练生成初始模型。图像分块数选用 patch 为 4，训练轮次 1000 轮。

4.2. 评价指标

为评估提出的半监督人群计数框架性能，与多数研究保持一致，采用平均绝对误差(Mean Absolute Error, MAE)和均方误差(Mean Squared Error, MSE)评价模型性能。MAE 和 MSE 分别代表的是模型预测人数与实际人数的误差的绝对值，反映了模型的准确程度，MSE 反映了模型的鲁棒性和泛化能力。两项指标的值越小，说明模型的准确率越高、泛化能力越好。两项指标计算公式如式(10)和式(11)。

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|C_i-C_i^{GT}\right| \tag{10}$$

$$\mathrm{MSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|C_i-C_i^{GT}\right|^2} \tag{11}$$

式中，N 表示测试集中测试图像的数量，C_i 和 C_i^{GT} 分别表示网络预测的人群总数和真实人群密度图中的人群总数。

4.3. 实验结果与分析

4.3.1. 实验结果对比

为全面评估本文所提算法的有效性，在 ShanghaiTech Part A、B 和 UCF_CC_50 数据集上与四种具有代表性的计数方法进行对比，分别是 MCNN [4]、CSRNet [16]、SANet [17]和半监督的 L2R [18]算法，选用的评价指标为 MAE 和 MSE。由表 1 结果可知，在 ShanghaiTech Part A 数据集上，MCNN 算法的 MAE 为 110.2，CSRNet 算法的 MAE 为 68.2，SANet 算法的 MAE 为 62.3，本文方法达到了 66.4；在 ShanghaiTech Part B 上，几种对比方法的 MAE 分别为 26.4、10.6、8.4、13.7 和 9.2。同时，在 UCF_CC_50 数据集上，本文所提模型的 MAE 为 201.2，CSRNet 的 MAE 为 248.2，明显低于其他模型的 MAE 结果，并具备较好的泛化能力。综合三个数据集的计数结果，可以发现，在 ShanghaiTech Part B 部分上，SANet 的 MAE 优于本文所提模型结果，考虑原因是层级间特征融合模块主要解决背景复杂问题，注重于前景信息。

由表 1 结果可以看到，本文所提方法在几个数据集上均取得了较低的 MAE 值，相比于基础模型监督模型和弱监督模型 L2R 来说，在几个数据集上的相对误差下降了。从数据集的场景属性来看，本文方法在高密度与复杂背景下的优势明显，主要原因是多层级语义增强模块与网络末端的 ASPP 模块为模型带来了多尺度特征的提取能力，能够在密集区域保持对前景信息的有效提取。在低密度场景下，如 ShanghaiTech Part B 数据集中，模型的 MAE 进一步降低，可以有效抑制背景误检。区域一致性半监督框架在小样本数据集 UCF_CC_50 上也取得了较好的效果，显著提升了模型识别的精度。总体而言，共享的多层级特征骨干网络在半监督框架下可以提取语义信息与多尺度上下文信息，增强模型对前景信息的特征提取能力，降低了计数误差。

Table 1. Comparison of crowd counting algorithm results on different data sets
表 1. 不同数据集上人群计数算法结果对比

算法	SHTech Part A		SHTech Part B		UCF_CC_50	
	MAE	MSE	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3	377.6	509.1
CSRNet	68.2	115.0	10.6	16.0	248.2	332.5
SANet	67.0	104.5	8.4	13.6	258.4	334.9
L2R	72.0	106.6	13.7	21.4	337.6	434.3
Ours	65.4	108.3	9.2	12.6	201.2	294.3

在数据集 ShanghaiTech Part A 和 Part B 的测试集部分上进行了测试，可视化结果如图 4 所示，其中，结果的第一行是 Part A 部分，第二行是 Part B 部分，结果中的数字为预测人数和实际人数。上排 Part A 的三幅高密度场景分别得到 285 (297)、362 (382)、144 (153)的结果。预测点在拥挤区域内分布均匀、边界处无明显错检、漏检，说明多层级语义增强与 ASPP 的尺度建模能够在强透视与遮挡下保持稳定响应，误差主要出现在极小头部与严重遮挡的远景区域，出现一定的错检现象对有效计数影响较小。下排 Part B 的三幅低密度的日常监控场景识别结果分别为 104 (107)、53 (55)、91 (90)，在橱窗人像、车辆等易混淆背景附近几乎无误检。预测点位置与头部中心高度一致，结果中平均绝对误差小。总体而言，模型在高密度场景的误差来源主要为极端遮挡与远景微小目标，表现为轻微低估；在低密度场景中则凭借区域一致性学习带来的背景抑制能力，实现了一对一计数与定位，预测与标注较为一致。

Figure 4. Crowd counting result diagram of ShanghaiTech Part A and Part B data set
图 4. ShanghaiTech Part A 和 Part B 数据集人群计数结果图

对高密度场景且数据样本较小的 UCF_CC_50 数据集进行了测试，测试结果如图 5 所示。红色预测点在主要人群区域的覆盖较为密集，对天空、楼体与地面等背景几乎无误检，表明网络保持稳定的检测性能。四个高密度检测结果分别为：2458 (2364)、922 (1050)、837 (890)、496 (484)。体现了模型在极端密集场景下对主体人群区域形成细粒度，显示出语义门控与通道注意对噪声的屏蔽。另外，根据实验结果图可以看出，计数稳定、重复计数少。模型在高密度、强遮挡条件下可以做到低误检和尺度适应，验证了模型的精度和鲁棒性。

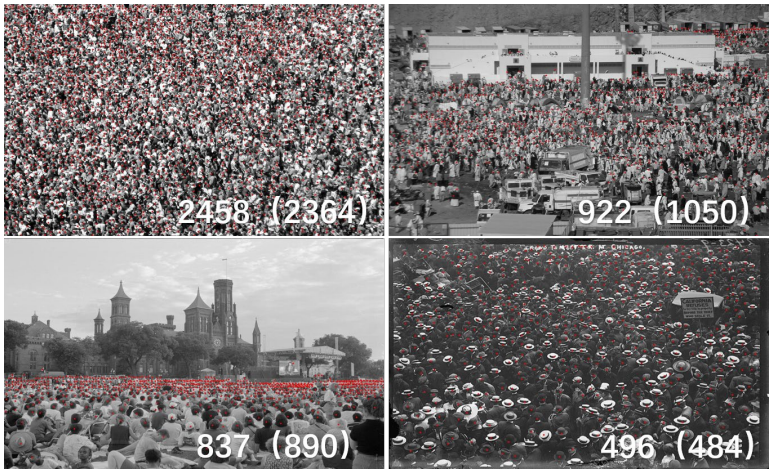


Figure 5. Crowd counting result diagram of UCF_CC_50 data set
图 5. UCF_CC_50 数据集人群计数结果图

为进一步验证模型性能，按人群密度将每个测试集划分为稀疏、中等和密集三个子集，测试模型在不同密度子集上的性能，分析对于不同密度人群的识别性能。在固定参数的条件下，我们以单图人数的33%、66%分位将 ShanghaiTech 的测试集划分为稀疏、中等和密集三档并统计 MAE、MSE，测试结果如表 2 所示。Part A 数据子集随密度上升，误差由 64.37、65.72 到 66.40 误差逐步增大，相对地，B 部分总体稀疏、背景占比大，随着人头密度提高，会出现更多有效子块，一致性守恒方法在这些子块上提供了更强的结构约束，抑制背景误检，误差随密度增加而下降，总体来看，模型在背景复杂但密度升高的城市场景受益更显著。

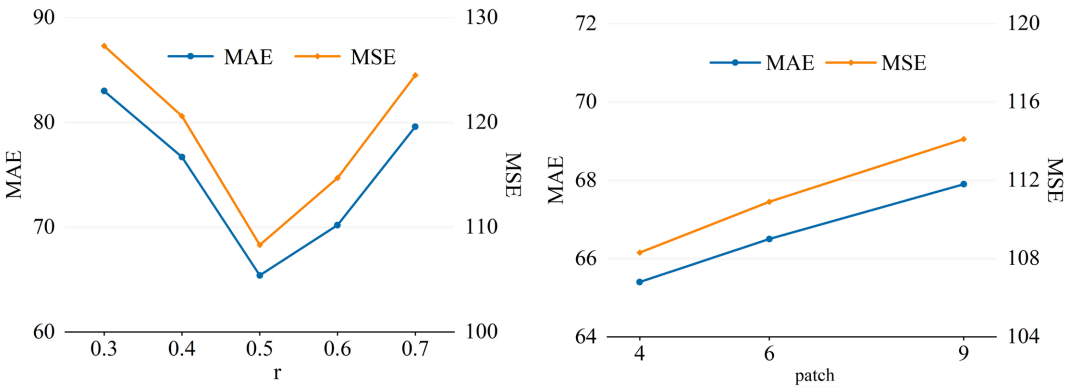
Table 2. Comparison of crowd counting algorithm results on different subsets
表 2. 不同子集上人群计数算法结果对比

子集	SHTech Part A		SHTech Part B	
	MAE	MSE	MAE	MSE
稀疏	64.37	106.3	9.42	13.1
中等	65.72	110.4	8.94	12.3
密集	66.40	112.1	9.23	12.8

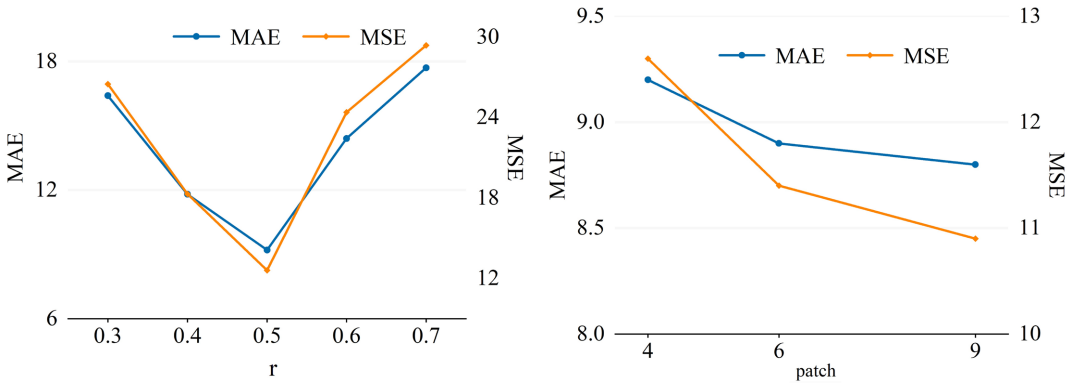
4.3.2. 超参数分析

为评估区域一致性半监督框架中关键超参数对性能的影响，我们在 ShanghaiTech 数据集的测试集上分别对置信度阈值 r 和分块数 *patch* 进行了多次训练和测试，并在其余训练设置保持不变的条件下观察 MAE 和 MSE 的变化趋势。置信度阈值 r 分别取 0.3、0.4、0.5、0.6 和 0.7，分块数 *patch* 分别取 4、6 和 9，测试结果如图 6 所示。

对于置信度阈值来说，在数据集的 A 和 B 部分上，均呈现清晰的“U”形趋势，当阈值从 0.3 提升至 0.5 时，MAE、MSE 下降，升至 0.6 和 0.7 后又升高。原因在于低阈值虽然提高了召回，但同时生成了更多的噪声点，导致计数被放大，而阈值过高使标签与有效预测点过于稀疏，可能造成漏检。介于两端之间的中等阈值约为 0.5 时能够在精度之上取得一定的平衡，因此两数据集在该点附近达到误差最低。对于分块数的影响，在数据集 A 部分中，细分后的每格有较多目标，人头贴近网格边界时，预测与伪标签容易被切割到相邻子块，可能导致重复计数，子块样本量下降带来计数方差上升。于是随着 *patch* 4 提高到 6 和 9，MAE 与 MSE 同步上升，其中 MSE 上升更快，说明大的局部误差增多。在数据集 B 部分中，由于目标间距大，同一人被切割开的概率低，因此考虑切分主要可能对背景误差抑制较多，因而 MAE 和 MSE 随 *patch* 增大而下降了。



(a) r 对 SHTA 数据集的影响　　(b) *patch* 对 SHTA 数据集的影响



(c) *r* 对 SHTB 数据集的影响　　(d) *patch* 对 SHTB 数据集的影响

Figure 6. Influence of hyperparameters on SHTech dataset
图 6. 超参数在 SHTech 数据集上测试的影响

5. 结论

针对智慧城市场景中密度变化剧烈、遮挡严重与背景干扰强的问题，本文构建基于多层级语义信息融合的区域一致性半监督人群计数框架。方法以 ResNet-34 为骨干网络，设计自上向下的多层级语义增强模块与末端 ASPP 多尺度上下文聚合，在不同层级之间实现轻量级融合，抑制背景噪声。在此基础上，引入区域一致性约束，训练上采用监督预热、无监督交替迭代更新机制。实验结果表明，所提框架在 ShanghaiTech Part A、Part B 与 UCF_CC_50 数据集上取得了较好表现，在高密度与复杂背景场景中可以有效识别人群区域。但是，研究仍存在一定的局限性，极端远景的微小目标与透视带来的漏检仍未完全解决，后续工作将向跨尺度特征提取和伪标签质量估计方向进行研究，进一步提升模型的精度和泛化能力。

基金项目

河北省社会科学基金项目“转型金融对河北省高碳企业低碳技术创新的驱动机制研究”(HB24ERJ027)。

参考文献

[1] 中华人民共和国国务院. 中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要[EB/OL]. 中国政府网. http://www.gov.cn/xinwen/2021-03/13/content_5592681.html, 2021-03-13.

[2] 陈冲, 白硕, 黄丽达, 等. 基于视频分析的人群密集场所客流监控预警研究[J]. 中国安全生产科学技术, 2020, 16(4): 143-148.

[3] 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.

[4] Zhang, Y., Zhou, D., Chen, S., Gao, S. and Ma, Y. (2016) Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 589-597. https://doi.org/10.1109/cvpr.2016.70

[5] 徐松, 许文利. 基于卷积神经网络的人群计数研究综述[J]. 科技与创新, 2025(10): 183-186.

[6] 陈永, 董珂, 安卓奥博, 等. 密集连接注意力与尺度感知重组增强的人群计数[J]. 光学精密工程, 2024, 32(22): 3395-3408.

[7] Jiang, X., Zhang, L., Zhang, T., Lv, P., Zhou, B., Pang, Y., *et al*. (2020) Density-Aware Multi-Task Learning for Crowd Counting. *IEEE Transactions on Multimedia*, **23**, 443-453. https://doi.org/10.1109/tmm.2020.2980945

[8] Pan, X., Mo, H., Zhou, Z. and Wu, W. (2020) Attention Guided Region Division for Crowd Counting. 2020 *IEEE International Conference on Acoustics*, *Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 2568-2572.

[9] Zhang, A., Shen, J., Xiao, Z., Zhu, F., Zhen, X., Cao, X., *et al*. (2019) Relational Attention Network for Crowd Counting. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 6787-6796.

https://doi.org/10.1109/iccv.2019.00689

[10] Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q. and Sebe, N. (2020) Weakly-Supervised Crowd Counting Learns from Sorting Rather than Locations. In: *Lecture Notes in Computer Science*, Springer, 1-17. https://doi.org/10.1007/978-3-030-58598-3_1

[11] 余鹰, 范在昌, 曾康利, 等. 渐进式认知引导的双域半监督人群计数[J]. 计算机研究与发展, 2025, 62(9): 2194-2207.

[12] 王鑫. 有限标注数据的复杂场景人群计数方法研究[D]: [博士学位论文]. 北京: 北京交通大学, 2024.

[13] Idrees, H., Saleemi, I., Seibert, C. and Shah, M. (2013) Multi-Source Multi-Scale Counting in Extremely Dense Crowd Images. 2013 *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 23-28 June 2013, 2547-2554. https://doi.org/10.1109/cvpr.2013.329

[14] Song, Q., Wang, C., Jiang, Z., Wang, Y., Tai, Y., Wang, C., *et al*. (2021) Rethinking Counting and Localization in Crowds: A Purely Point-Based Framework. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 3345-3354. https://doi.org/10.1109/iccv48922.2021.00335

[15] 彭思凡. 基于密度估计的人群计数方法研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2022.

[16] Li, Y., Zhang, X. and Chen, D. (2018) CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1091-1100. https://doi.org/10.1109/cvpr.2018.00120

[17] Cao, X., Wang, Z., Zhao, Y. and Su, F. (2018) Scale Aggregation Network for Accurate and Efficient Crowd Counting. In: *Lecture Notes in Computer Science*, Springer, 757-773. https://doi.org/10.1007/978-3-030-01228-1_45

[18] Liu, X., van de Weijer, J. and Bagdanov, A.D. (2018) Leveraging Unlabeled Data for Crowd Counting by Learning to Rank. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7661-7669. https://doi.org/10.1109/cvpr.2018.00799