



深度频谱自适应调制的密集场景人群计数方法

刘 盟¹, 张凯兵^{1,2}, 孟雅蕾¹

1. 西安工程大学 计算机科学学院, 陕西 西安 710048; 2. 陕西省服装设计智能化重点实验室, 陕西 西安 710048

摘要 在人群计数领域, 由于场景中人群密度和遮挡程度高度变化, 传统方法往往难以在保持局部细节和捕捉全局上下文之间取得平衡, 导致在高拥挤度或透视畸变严重的图像上性能下降。针对这一问题, 提出一种基于深度频谱自适应调制的密集场景人群计数方法(deep spectral adaptive modulation for dense scenes, DSAMNet), 旨在同时提升密度图估计的空间分辨率与全局计数一致性。首先, 通过深度感知模块联合提取多尺度纹理特征与粗略深度估计, 并利用预测深度引导透视校正, 实现几何一致的特征采样; 然后, 引入空间自适应高频编码模块, 在编码过程中依据局部深度动态调制坐标频率, 从而增强几何敏感区域的位置表达能力, 提升网络对尺度与结构变化的适应性; 最后, 隐式密度解码模块通过跨域注意力机制融合视觉与几何表征, 并借助分层条件化多层感知机进行逐点密度回归, 实现对连续密度场的高质量重建。实验结果表明, 所提方法在多个主流人群计数数据集上取得了优异性能, 展现出在复杂透视场景下良好的鲁棒性与泛化能力。

关键词 人群计数; 深度感知; 透视校正; 自适应高频编码; 隐式密度解码

中图分类号: TP391 **DOI:** 10.16152/j.cnki.xdxbzr.2026-03-010

Crowd counting via deep spectral adaptive modulation for dense scenes

LIU Meng¹, ZHANG Kaibing^{1,2}, MENG Yalei¹

1. School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China;

2. Shaanxi Key Laboratory of Clothing Intelligence, Xi'an 710048, China

Abstract In crowd counting, traditional methods often struggle to balance preserving local details and capturing global context due to the high variability in crowd density and occlusion levels within scenes, leading to performance degradation in highly congested or severely perspective-distorted images. To address this issue, this paper proposes a dense-scene crowd counting method based on deep spectral adaptive modulation (DSAMNet), aiming to simultaneously enhance the spatial resolution of density map estimation and global counting consistency. Specifically, the method first employs a depth-aware module to jointly extract multi-scale texture features and coarse depth estimates, utilizing predicted depth to guide perspective correction and achieve geometrically consistent feature sampling. Subsequently, a spatially adaptive high-frequency encoding module is in-

收稿日期: 2025-12-20 修回日期: 2026-01-05

基金项目: 陕西省重点研发计划(2024GX-ZDCYL-02-07); 陕西省自然科学基金重点项目(2018JZ6002); 西安工程大学研究生创新基金项目(CHX2024025)。

第一作者: 刘盟, 男, 博士研究生, 从事计算机视觉研究, mengliu1007@mail.nwpu.edu.cn。

通信作者: 张凯兵, 男, 博士, 教授, 博士生导师, 从事机器学习、图像超分辨重建等研究, zhangkaibing@xpu.edu.cn。

troduced, which dynamically modulates coordinate frequencies based on local depth during the encoding process, thereby enhancing the positional representation capacity in geometrically sensitive regions and improving the network's adaptability to scale and structural variations. Finally, an implicit density decoding module integrates visual and geometric representations through a cross-domain attention mechanism and performs point-wise density regression via a hierarchically conditioned multi-layer perceptron, achieving high-quality reconstruction of continuous density fields. Experimental results demonstrate that the proposed method achieves superior performance on several mainstream crowd counting datasets, exhibiting strong robustness and generalization capability in complex perspective scenarios.

Keywords crowd counting; depth awareness; perspective correction; adaptive high-frequency encoding; implicit density decoding

密集人群计数作为智慧城市管理中的关键技术,在公共安全预警^[1]、交通流量监测^[2]等领域具有重要应用价值。传统方法基于显式密度预测范式,通过卷积神经网络直接回归离散密度图^[3],但其固定分辨率输出机制导致严重的尺度敏感性问题,难以适应复杂场景下的多尺度人群分布。

在多数传统人群计数方法中,网络的输出通常是与输入图像等比或下采样的离散密度图^[4-8]。在现实场景中,由于投影透视和相机高度不同,同一视野中人的头部尺寸会随深度显著变化,而采用统一的像素网格和统一尺度的高斯核会使其能量在多个像素上扩散,导致远处小目标产生过度平滑,近处大目标产生能量集中不准确的问题。离散化和固定核尺度导致了尺度不变性差,在密集或透视畸变严重的区域表现为计数偏差与局部密度形态失真。此外,离散密度图的生成通常依赖人工设定的核函数与参数,这些参数在不同场景、不同相机高度、视角和透视投影条件下并无统一的最优值,因而容易导致标注形式与真实尺度分布之间的不一致性。这种人为设定的监督信号在训练中会将尺度与视角差异的负担全部交给网络去补偿,增加了学习难度并降低了泛化能力。

当密度场的真实结构需要在任意空间位置以任意分辨率解析时,基于固定像素网格和固定平滑核的离散密度图无法提供足够的表达灵活性与几何一致性,进而在透视畸变、密集遮挡与尺度变化显著的场景下导致系统性能的下降。因此,突破这一瓶颈的方向就是寻找能够以连续方式表达密度场,并能自适应几何变换与视觉上下文的建模范式。

隐式神经表示(implicit neural representations, INRs)^[9-10]为解决上述问题提供了一个理论可行

的方向:将密度视为位置到标量的连续函数 $\rho(x)$,允许在任意坐标处以任意分辨率进行采样,从根本上避免像素化引起的量化与下采样误差。尽管现有 INRs 方法在图像重建任务中展现出细粒度建模优势,但直接应用于人群计数时存在本质缺陷。首先,大多数隐式表征以像素坐标及其位置编码作为主要输入^[11-15],但这种“坐标唯输入”的设计忽视了图像提供的丰富视觉上下文信息,使得隐式密度函数在面对相同坐标但不同语义结构时无法做出差异化响应,即密度场与视觉语义在模型中出现了解耦,导致在存在复杂纹理、遮挡或类别背景变化的区域中预测的密度缺乏语义引导。其次,城市监控图像中普遍存在的透视投影和摄像机姿态差异,使得像素坐标系的几何意义随场景深度而变化,若没有显式的几何先验或深度信息,坐标-密度映射在深度突变或视角变化处会产生系统性偏移,进而影响局部计数的准确性和几何一致性。再次,隐式网络常用的固定频带位置编码在面对区域内尺度与细节复杂度显著不同的图像时表现出受限性:固定的频谱设计既难以在需要高频辨识的区域学习到足够的细节,也在平坦区域浪费表示能力,因而很难兼顾全局一致性与局部细节恢复。综上,问题的根源可归结为现有隐式表示框架未能在输入层面将视觉上下文与空间几何有效地联合建模,缺乏一种机制可以在保持连续坐标表示优势的同时,按需地将语义特征、几何先验与频谱表达融合起来,以适应场景内的尺度与结构变化。

为弥补离散密度表示的尺度限制,并克服 INRs 直接迁移所面临的语义与几何失配问题,本文提出一种基于深度频谱自适应调制的密度建模范式,旨在在连续坐标回归的框架下恢复高分辨率且几何一致的密度场。为此,我们在网络设计

上引入 3 类互补机制以实现视觉语义与空间几何的协同。首先,通过一个深度感知的特征融合通道将多尺度的图像特征与自监督估计得到的深度信息结合,构建一个粗略但稳定的几何场用于引导空间上一致的语义采样与特征对齐,缓解由于纯坐标输入导致的语义脱钩问题;其次,提出一种基于深度调制的空间自适应高频编码策略,使得位置编码的频谱响应随局部几何复杂度和深度变化动态调整,在需要精细表达的区域提升频率辨识度,而在平坦区域保持平滑性,提高模型对尺度与结构变化的适应性;最后,在密度解码阶段,将几何调制后的位置编码与视觉特征在跨域注意力机制下进行联合建模,并采用分层条件化的感知机逐点回归密度值,这一设计既保持了隐式表示的连续采样能力,又通过跨域信息交互实现了局

部细节与全局一致性的协同优化。系统性的消融与基准对比表明,本文方法在保持局部密度细节的同时,显著提升了整体计数准确性与几何一致性,为隐式表示在人群理解任务中的应用提供了新的范式。

1 DSAMNet 方法

本文提出的基于深度频谱自适应调制的人群计数网络 DSAMNet 以连续密度场建模为核心,通过多阶段特征变换实现几何自适应的密度估计。DSAMNet 框架如图 1 所示,网络架构包含编码器、深度感知模块、空间自适应高频编码模块和隐式密度解码模块。

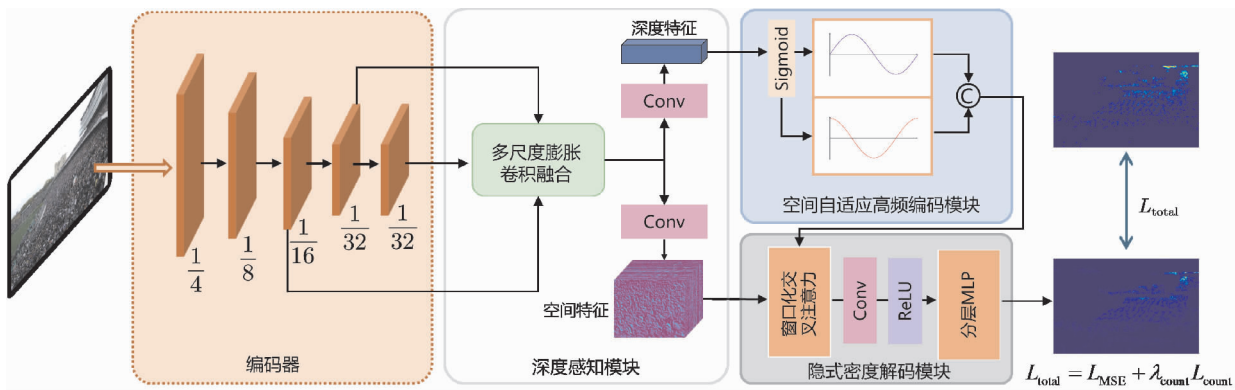


图 1 DSAMNet 整体框架

Fig. 1 The framework of DSAMNet

DSAMNet 以 RGB 图像 $I \in \mathbf{R}^{H \times W \times 3}$ 为输入,通过多尺度特征金字塔提取具有语义感知的视觉特征,同时生成描述场景几何结构的几何先验图 $D \in \mathbf{R}^{\frac{H}{4} \times \frac{W}{4} \times 1}$ 。基于归一化坐标网格 $C \in \mathbf{R}^{B \times N \times 2}$,通过可微分形变模块生成几何对齐坐标 C' ,并编码为深度感知的高频特征 $H \in \mathbf{R}^{B \times N \times 128}$ 。高频编码器将形变坐标与深度特征融合,生成几何感知的位置编码,并通过局部窗口注意力机制与视觉特征交互。最终,条件化多层感知机以视觉特征为调制条件、高频编码为查询输入,输出连续密度场 $\Phi(C')$ 。

1.1 深度感知模块

在人群拥挤的环境中,目标尺度变化剧烈,单层特征难以兼顾局部细节与全局上下文。为此,我们采用预训练 VGG-16-BN 的前 4 阶段分别抽取不同分辨率的视觉表征 $F^{(1)}$ 、 $F^{(2)}$ 、 $F^{(3)}$ 、 $F^{(4)}$,特征映射过程如下:

$$F^{(l)} = \text{VGGStage}_l(I), l = 1, 2, 3, 4 \quad (1)$$

式中: $F^{(l)}$ 具备各自的空间尺度优势,能够捕获从低层纹理到高层语义的信息;VGGStage 为编码器池化前的卷积过程。

为了统一后续处理通道数,对 $F^{(i)} (i = 2, 3, 4)$ 进行 1×1 卷积降维:

$$P^{(i)} = \text{Conv}_{1 \times 1}(F^{(i)}), i = 2, 3, 4 \quad (2)$$

接着,本文设计了不同膨胀率 $d \in \{1, 2, 4\}$ 的并行卷积分支,以扩展感受野并并行获取多尺度上下文:

$$B_d^{(i)} = \text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3, d}(P^{(i)}))) \quad (3)$$

将 3 层所有分支的输出沿通道拼接,采用 3×3 和 1×1 组合卷积融合多尺度信息:

$$F^{\text{fused}} = \text{Conv}_{1 \times 1}(\text{Conv}_{3 \times 3}([B^{(2)}; B^{(3)}; B^{(4)}])) \quad (4)$$

$$D = \sigma(\text{Conv}_{1 \times 1}(\text{ReLU}(\text{Conv}_{3 \times 3}(F^{\text{fused}})))) \quad (5)$$

模块输出包括统一尺度的融合特征 $\mathbf{F}^{\text{fused}}$ 和对应的粗略几何先验图 $\mathbf{D}$ 。该设计既保持了分辨率 $\frac{H}{4} \times \frac{W}{4}$ 上的高层语义,也引入了深度先验,为后续几何自适应提供基础。深度学习分支并无深度真值监督,而是通过透视校正与频率调制 2 条可微路径,由密度回归与全局计数损失端到端隐式驱动学习。其输出并不是严格意义上的物理深度,但具有面向计数任务的透视几何意义。

1.2 空间自适应高频编码模块

真实场景中,透视变形导致人群密度在图像平面上非线性分布,常规位置编码无法精确定位。本文提出基于深度加权的高频编码策略,提升模型对透视畸变和尺度变化区域的高频几何信息表征能力。首先,定义图像的标准化采样坐标矩阵为 $\mathbf{X} \in \mathbf{R}^{B \times N \times 2}$,其中, B 为批量大小, N 为采样点总数,坐标维度对应 $(x, y) \in [-1, 1]^2$ 。随后,模块通过双线性插值在 $\mathbf{D}$ 上采样得到每个空间点的估计深度值 $\mathbf{z}_n \in \mathbf{R}^{B \times N \times 1}$ 。

$$\mathbf{z}_n = \text{Interp}(\mathbf{D}, \mathbf{g}_n) \quad (6)$$

式中:Interp 表示上采样操作; $\mathbf{g}_n$ 表示上采样尺寸。

为模拟透视几何效应,对原始坐标进行几何校正:

$$\tilde{\mathbf{X}} = \mathbf{X} \odot (1 + \alpha \mathbf{z}_n) \quad (7)$$

式中: $\odot$ 表示逐元素乘法,该变换自动权衡远近处对象的尺度缩放。接着,引入可学习频率基向量 $\mathbf{f} \in \mathbf{R}^K$ (K 为频率带数量),并通过广播与深度引导项结合,获得调制频率矩阵:

$$\mathbf{M} = \mathbf{z}_{\text{Sigmoid}} \otimes \mathbf{f} \quad (8)$$

式中: $\mathbf{z}_{\text{Sigmoid}} = \sigma(\mathbf{z}_n)$ 为深度值经 Sigmoid 映射后的结果,以抑制极端数值; $\otimes$ 为广播相乘。基于 $\tilde{\mathbf{X}}$ 与 $\mathbf{M}$,分别沿坐标分量执行正余弦调制,形成高频编码:

$$E(\tilde{\mathbf{X}}) = [\sin(2\pi \mathbf{M} \odot \tilde{\mathbf{X}}), \cos(2\pi \mathbf{M} \odot \tilde{\mathbf{X}})] \quad (9)$$

最后,将该高维编码输入至一个线性层 $\text{Linear}(2K, \mathbf{D})$,映射为固定维度的隐式特征 $\mathbf{h} \in \mathbf{R}^{B \times N \times D}$,并作为后续跨域注意力模块的 K/V 特征。

该设计在传统位置编码基础上引入了基于深度的动态调频机制,使高频表达对几何变形更为敏感,从而在透视与密集人群场景中显著提升隐式密度场回归的精度与鲁棒性。

1.3 隐式密度解码模块

尽管深度调制的位置编码可在隐式空间中对

透视畸变进行一定校正,但若缺乏足够的语义感知与显式纹理线索,则难以恢复具有连续性的细粒度密度估计结果。因此,首先将视觉空间特征与高频几何编码特征通过窗口化多头跨注意力进行细粒度交互,设来自多尺度感知的高分辨率特征经透视采样后得到空间查询张量 $\mathbf{Q} \in \mathbf{R}^{B \times N \times D}$,高频编码输出的键值对张量均为 $\mathbf{K}, \mathbf{V} \in \mathbf{R}^{B \times N \times D}$ 。模块在每个局部窗口 w 上执行 H 头注意力:

$$\text{head}_h^w = \text{softmax}\left(\frac{\mathbf{Q}_h^w \mathbf{K}_h^w}{\sqrt{d_h}}\right) \mathbf{V}_h^w \quad (10)$$

$$\mathbf{A}^w = [\text{head}_1^w, \text{head}_2^w, \dots, \text{head}_H^w] \mathbf{W}_o \quad (11)$$

$$\mathbf{A} = \text{Concat}_w(\mathbf{A}^w) \in \mathbf{R}^{B \times N \times D} \quad (12)$$

式中: $d_h = \frac{D}{H}$; $\mathbf{W}_o \in \mathbf{R}^{D \times D}$ 为输出投影矩阵。得到

跨注意力融合特征 $\mathbf{A}$ 后,应用简单的 2 层全连接网络 $\varphi: \mathbf{R}^D \rightarrow \mathbf{R}^{2D}$,并以 ReLU 激活:

$$\mathbf{Z} = \text{ReLU}(\mathbf{A} \mathbf{W}_\varphi^{(1)} + \mathbf{b}_\varphi^{(1)}) \mathbf{W}_\varphi^{(2)} + \mathbf{b}_\varphi^{(2)} \quad (13)$$

随后,引入来自多尺度特征融合分支的条件向量 $\mathbf{C} \in \mathbf{R}^{B \times N \times C_c}$ 。并在分层条件化 MLP 中逐级对 $\mathbf{Z}$ 进行仿射调制与残差解码。具体地,令网络层数为 L ,第 l 层使用条件网络 $\psi^{(l)}: \mathbf{R}^{C_c} \rightarrow \mathbf{R}^{2D}$ 生成仿射参数:

$$[\gamma^{(l)}, \beta^{(l)}] = \psi^{(l)}(\mathbf{C}) \in \mathbf{R}^{B \times N \times 2D} \quad (14)$$

令 $\mathbf{x}^{(0)} = \mathbf{Z}$,其前向更新规则为

$$\mathbf{x}^{(l)} = \text{ReLU}(\gamma^{(l)} \odot (\mathbf{x}^{(l-1)} \mathbf{W}^{(l)} + \mathbf{b}^{(l)}) + \beta^{(l)}) + \mathbf{x}^{(l-1)} \quad (15)$$

式中: $l = 1, 2, \dots, L$; $\mathbf{W}^{(l)} \in \mathbf{R}^{2D \times 2D}$, $\mathbf{b}^{(l)} \in \mathbf{R}^{2D}$ 为各层权重与偏置; $\odot$ 表示逐元素乘法。最后通过线性变换与 Sigmoid 映射得到点密度:

$$\hat{\rho}_i = \sigma(\mathbf{x}_i^{(L)} \mathbf{w}_{\text{out}} + \mathbf{b}_{\text{out}}), \hat{\rho} \in \mathbf{R}^{B \times N \times 1} \quad (16)$$

最终,将 $\hat{\rho}$ 重塑为空间格点 (H', W') 形式,输出隐式回归的密度图。此流程通过窗口化跨注意力捕获视觉与几何高频信息的多尺度耦合,并借助分层条件化解码将丰富的上下文先验转化为精准的局部密度估计。相比于传统密度回归网络中基于卷积的逐像素预测方式,本文提出的方法不仅在密度恢复上具有更强的连续性表达能力,也展现出更高的鲁棒性与泛化性,尤其适用于透视严重、人群密度分布极不均匀的场景。

1.4 损失函数

在本模型中,联合了加权均方误差损失 (weighted MSE loss) 和全局计数损失 (count loss),同时约束密度图的局部精度与全局一致性。记网络对第 b 个样本在像素位置 i , ($i = 1, 2,$

..., N) 的预测密度为 $\hat{\rho}_{b,i}$, 对应的标注密度为 $\rho_{b,i}$, 则局部误差以加权均方误差形式定义为

$$L_{\text{MSE}} = \frac{1}{BN} \sum_{b=1}^B \sum_{i=1}^N \omega_{b,i} (\hat{\rho}_{b,i} - \rho_{b,i})^2 \tag{17}$$

$$\omega_{b,i} = \sigma(\rho_{b,i} + \varepsilon) \tag{18}$$

式中: $\sigma(\cdot)$ 为 Sigmoid 函数; $\varepsilon = 10^{-6}$, 用于防止权重萎缩至 0。这一权重能够强化稠密区域的误差惩罚, 提升模型对人群聚集处的敏感度。与此同时, 全局计数损失衡量预测密度图与真值密度图在整体人数统计上的偏差, 其表达式为

$$L_{\text{count}} = \frac{1}{B} \sum_{b=1}^B \left| \sum_{i=1}^N \hat{\rho}_{b,i} - \sum_{i=1}^N \rho_{b,i} \right| \tag{19}$$

通过对每个样本的绝对计数差取平均, 保障最终预测在总人数上与真实标注保持一致。整个网络以加权线性形式将两部分损失合并:

$$L_{\text{total}} = L_{\text{MSE}} + \lambda_{\text{count}} L_{\text{count}} \tag{20}$$

式中, 超参数 $\lambda_{\text{count}} = 0.1$ 。

2 实验结果与分析

2.1 实现配置

在本方法的训练过程中, 使用预先训练好的 VGG16 网络初始化其主干网络, 其余的参数由标准偏差为 0.01 的高斯分布初始化。使用学习率为 1×10^{-5} 的 Adam 优化器对模型进行优化。在所有实验中, 批量大小和训练 epoch 分别设置为 64 和 500。为了达到更好的性能, 在图像输入网络之前, 都经过一系列数据增强处理, 包括 [0.8, 1.2] 的随机重缩放和概率为 0.5 的水平翻转, 在概率为 0.3 的情况下, 使用参数 [0.5, 1.5] 进行 gamma 对比度变换, 以及在输入图像的随机位置裁剪一组 400×400 的图像块。本章内容所有实

验在部署了 GeForce RTX3090 GPU 和 PyTorch 的深度学习平台上实现。

2.2 数据集

在以下 4 个具有挑战性的人群计数数据集上评估本文方法: Shanghai Tech A 和 B^[16], UCF-QNRF^[17] 和 JHU-Crowd++^[18]。其中, Shanghai Tech A 和 B 是 2 个小数据集, 前者包含比后者密集得多的群体。UCF-QNRF 和 JHU-Crowd++ 是 2 个大规模数据集, 都涵盖了不同密度的范围。

1) Shanghai Tech A(简记为 SHA)由 482 张图像组成, 从互联网上随机收集。每张图像的注释范围从 33 ~ 3 139, 训练集包含 300 张图像, 剩余的 182 张图像用于测试。

2) Shanghai Tech B(简记为 SHB)包含 716 张图像, 从上海拥挤的街道收集。与 SHA 相比, 该数据集的密度较低, 图像中的人数为 9 ~ 578。训练集有 316 张图像, 其余 400 张图像用于测试。

3) UCF-QNRF(简记为 QNRF)是一个拥挤的数据集, 包含 1 535 张高分辨率图像和大约 125 万个点注释, 每张图像平均有 185 个人。训练集有 1 201 张图像, 其余的 334 张图像用于测试。

4) JHU-Crowd++ 是一个大型数据集, 包含 4 327 张图像和 151 万个注释。训练集有 2 272 张图像, 验证集有 500 张图像, 剩余的 1 600 张图像用于测试。

2.3 不同方法的对比实验

在公开的人群计数通用数据集上与其他主流人群计数方法进行对比实验, 并给出了可视化实验结果, 验证本文方法的有效性。从表 1 可以看出, 本文方法在 3 个常用数据集 SHA、SHB 和 QNRF 上均取得了比现有方法更为出色的表现。

表 1 本文方法在 Shanghai Tech 数据集与 UCF-QNRF 数据集上与其他方法的对比结果

Tab. 1 Comparison results of this method with other methods on the Shanghai Tech dataset and UCF-QNRF dataset

方法	方式	SHA		SHB		QNRF	
		MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
MCNN ^[16]	密度图	110.2	173.2	26.4	41.3	227.0	426.0
CP-CNN ^[19]	密度图	73.6	106.4	20.1	30.1	—	—
CSRNet ^[20]	密度图	68.2	115.0	10.6	16.0	118.8	204.4
CAN ^[21]	密度图	61.3	100.0	7.8	12.2	107.0	183.0
DM-Count ^[22]	密度图	59.7	95.7	7.4	11.8	85.6	148.3
SS-DCNet ^[23]	密度图	59.9	95.0	7.7	11.1	92.4	158.7

续表 1

方法	方式	SHA		SHB		QNRF	
		MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
SRRNet ^[24]	密度图	60. 8	103. 0	7. 4	13. 6	89. 5	162. 9
CAFNet ^[26]	密度图	59. 4	94. 5	6. 8	11. 2	96. 4	155. 5
DDEAR ^[27]	密度图	74. 2	123. 8	—	—	121. 0	225. 8
DLPTNet ^[25]	密度图	58. 4	95. 0	9. 3	15. 6	121. 0	225. 8
LSC-CNN ^[29]	检测框	66. 4	117. 0	8. 1	12. 7	120. 5	218. 2
TopoCount ^[28]	检测框	61. 2	104. 6	7. 8	13. 7	89. 0	159. 0
本文方法	密度图	58. 9	89. 5	7. 1	12. 8	96. 3	163. 2

注:加黑数值表示最优。

在 SHA 数据集上,本文方法的 MAE 和 MSE 分别为 58.9 和 89.5。与所列的其他密度图回归方法相比,本文方法在 MAE 方面低于 MCNN^[16]、CP-CNN^[19]、CSRNet^[20]、CAN^[21]、DM-Count^[22]、SS-DCNet^[23]、SRRNet^[24]等方法,具体降低幅度为 13%~46%,MSE 同样表现出稳定的改进,降幅为 15%~48%。MAE 与 DLPTNet^[25]的结果相差 0.5,基本处于相同水平。综合来看,本文方法在该数据集上的误差指标处于较低水平,表明其在密集场景下的鲁棒性和密度回归精度均有良好的保证。

在 SHB 数据集上,本文方法的 MAE 和 MSE 分别为 7.1 和 12.8。与 MCNN、CP-CNN、CSRNet、DM-Count 等方法相比,MAE 和 MSE 的降低幅度分别达到 65%~73%和 60%~69%;与部分近年来提出的方法如 SRRNet、DLPTNet 相比,本文方法在 MAE 上有 5%~23%的降低,同时在 MSE 上也降低了近 18%。值得注意的是,本文方法与 CAFNet^[26]和 SS-DCNet 在误差指标上接近,均处于较低水平。这说明,尽管近年来多种方法在 SHB 数据集上取得了良好的性能,本文方法依然能够实现与之相当甚至略有优越的表现,具备较强的密集场景适配能力。

在 QNRF 数据集上,本文方法的 MAE 和 MSE 分别为 96.3 和 163.2,相较于 CSRNet、CAN、SS-DCNet、SRRNet、CAFNet、DDEAR^[27]、DLPTNet 等方法,MAE 和 MSE 的降低幅度为 10%~24%。与检测式方法如 TopoCount^[28]和 LSC-CNN^[29]相比,本文方法在 MAE 和 MSE 上分别提升了 10%~20%。虽然在与部分密度图回归方法如 DM-Count 的对比中,误差略高 10%,但整体仍表现出

较低的误差水平,体现了本文方法在处理高密度、复杂场景分布方面的泛化能力。

综上所述,本文提出的方法在 3 个数据集上的表现均优于或持平于现有最优水平,特别是在高密度 QNRF 场景下的显著改进,彰显了所提方法在提升计数精度和适应复杂场景变化方面的有效性和综合优势。

为了更加直观地展示本文方法的优越性,在 Shanghai Tech A/B 数据集上分别选择了具有较大透视效应的一些场景,可视化示例分别如图 2 与图 3 所示。可以看出,在极端透视场景下,本文所提方法的预测密度图与真实密度图有着非常相似的分



注:Pre 为预测人数;GT 为真实人数。

图 2 本文方法在 SHA 数据集上预测结果可视化示例
Fig. 2 Visualization example of prediction results using this method on the SHA dataset



注:Pre 为预测人数;GT 为真实人数。

图 3 本文方法在 SHB 数据集上预测结果可视化示例

Fig. 3 Visualization example of prediction results using this method on the SHB dataset

JHU-Crowd++数据集因其高密度、复杂场景以及多样化的遮挡模式,被广泛认为是当前最具挑战性的密集人群计数数据集之一。本文在该数据集上与其他方法进行对比,结果如表 2 所示。本文方法相较于 MCNN,在 MAE 与 MSE 上分别降低了 62.3% 和 36.7%,相较于 CSRNet,误差分别减少了 16.2% 与 4.4%。与 SANet^[30]、CAN 等方法相比,本文方法分别在 MAE 上分别降低了 26.3% 与 32.4%,在 MSE 上分别降低了 12.0% 与 0.2%。相较于 SFCN^[31],本文方法在 MAE 与 MSE 上仍分别实现了 3.8% 与 3.5% 的性能增益。这些结果表明本文方法在该数据集上的量化提升,

表 3 不同模型效率的比较

Tab. 3 Comparison of efficiency across different models

方法	Params/10 ⁶	FLOPs/10 ⁹	FPS	MAE ↓	RMSE ↓
CSRNet ^[20]	16.26	182.69	66.35	68.2	115.0
CAN ^[21]	18.10	193.65	56.17	61.3	100.0
SASNet ^[32]	38.90	393.20	21.77	53.6	88.4
BL ^[33]	21.50	182.19	63.75	62.8	101.8
SRRNet ^[24]	66.14	162.09	26.93	60.8	103.0
RAQNet ^[34]	28.30	250.80	28.30	59.0	101.2
本文方法	2.42	163.13	81.27	58.9	89.5

本文对 3 类极端场景,分别为雾霾天气、暗光逆光、类似人群的纹理干扰场景进行了测试,结果

验证了经过深度感知和自适应高频编码增强后的隐式神经表示在面对遮挡、尺度变化以及背景噪声的多样性时,能够更好地适应复杂几何特征与语义上下文,实现更加鲁棒和精细的人群密度估计结果。

表 2 在 JHU-Crowd++数据集上的对比结果

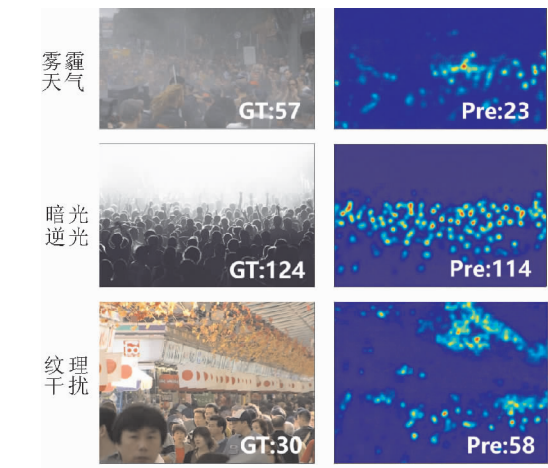
Tab. 2 Comparison results on the JHU-Crowd++ dataset

方法	方式	MAE ↓	MSE ↓
MCNN ^[16]	密度图	160.6	377.7
CSRNet ^[20]	密度图	72.2	249.9
CAN ^[21]	密度图	89.5	239.3
SANet ^[30]	密度图	82.1	271.6
SFCN ^[31]	密度图	62.9	247.5
本文方法	密度图	60.5	238.9

2.4 模型性能

为评估本文方法的实际落地潜力,对主流人群计数方法的复杂度与推理效率进行了对比,包括参数量(Params)、计算量(FLOPs)与推理速度(FPS),结果见表 3。通过引入隐式解码头采用点级两层隐藏 MLP,并通过 FiLM 方式注入条件特征与透视条件。条件调制网络以少量层级共享的方式生成 γ, β ,对中间激活进行逐元素仿射调制,避免了像素级密集卷积解码头或自注意力解码头带来的高计算开销。总体来看,本文方法在保持较强计数精度的同时,展现出更优的推理吞吐能力。需要说明的是,参数量并不能完全决定速度差异,推理效率受算子类型与访存模式等实现因素影响。因此,在同一测试环境下报告 FPS,能直接反映落地时的实际性能。综合精度与效率,本文方法在接近主流精度的基础上体现出较强的应用可行性。

如图 4 所示。



注:Pre 为预测人数;GT 为真实人数。

图 4 不同极端场景下的测试结果

Fig. 4 Test results under different extreme scenarios

在严重雾霾场景下,密度图过于平滑但整体漏计明显,性能退化最明显。这是雾霾带来的对比度下降与纹理抹平造成显著分布偏移,同时依赖图像统计线索学习到的透视/几何先验在该条件下更易失真,从而削弱透视校正与频率调制的有效性。但在暗光逆光场景下,模型依然能稳定发挥效用,这是由于暗光/逆光主要改变亮度幅度而非结构本身,多尺度融合与归一化一定程度上缓解了光照变化。在类似人群的纹理干扰场景下模型也表现出一定的退化,这是受限于频谱编码提升的是空间函数的表达上限,但不具备语义判别能力。若视觉条件特征在语义上发生混淆,隐式回归会把这种混淆更精确地表达出来,表现为密度图在纹理区域出现规律性虚假响应。

2.5 消融实验

2.5.1 轻量级主干设计

为了验证本文所提方法中各个模块的有效性,在 Shanghai Tech A 数据集上对深度感知模块(DPM)、空间自适应高频编码模块(SAHFEM)和隐式密度解码模块(IDDM)3个关键模块进行了广泛的消融研究,结果如表4所示。首先移除了空间自适应高频编码模块与隐式密度解码模块,以单独使用空间感知模块的模型作为基线模型,然后对其输出的特征图直接投影生成密度图。该基线模型在测试集上的MAE为75.3,MSE为112.1。随后,用固定频率编码方式取代了空间自适应高频编码模块,以验证其自适应几何增强能力。在该设置下,模型的MAE降低至65.6,MSE降低至105.2,相对误差分别降低了12.9%与

6.1%,证明了该模块在捕捉局部几何细节与高频纹理中的优势。

表 4 模块消融实验结果

Tab. 4 Results of module ablation experiments

DPM	SAHFEM	IDDM	MAE ↓	MSE ↓
✓	×	×	75.3	112.1
✓	✓	×	65.6	105.2
×	×	✓	64.2	105.1
✓	✓	✓	58.9	89.5

此外,本文采用不带分层条件的普通MLP解码器替代IDDM,结果表明MAE和MSE分别下降至58.9与89.5,相比前一配置额外降低了10.2%和14.9%。这一结果清晰地验证了分层条件化解码和残差解码策略在特征跨域融合与高精度细节还原方面的有效性,证明了本文所提的分层条件和残差解码策略对有效地跨域融合和精细细节恢复至关重要。为了评估深度感知模块的影响,移除了深度感知矩阵(DPM),并直接使用未处理的输入特征作为几何先验。结果显示模型的MAE和MSE分别上升至75.3和112.1,该配置较完整模型分别降低了21.8%和20.2%,凸显出深度感知特征在全局一致性和几何先验补充方面的核心作用。证明了本文所提模块化架构能够在几何建模、特征自适应增强以及隐式密度回归等方面协同作用。

2.5.2 损失函数消融实验

本文在训练阶段引入了数据增广策略以增强模型的鲁棒性,并在SHA和QNRF数据集上进行消融实验,结果如表5所示。

表 5 SHA 和 QNRF 数据集上的数据增广策略消融实验结果

Tab. 5 Experimental results of data augmentation strategy ablation on SHA and QNRF datasets

数据增广	SHA		QNRF	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓
w/o	60.8	93.4	98.2	169.8
w	58.9	89.5	96.3	163.2

在不使用数据增广策略的情况下,模型在SHA上取得的MAE和MSE分别为60.8和93.4,在QNRF数据集上分别为98.2和169.8。当使用数据增广策略后,SHA数据集的MAE和MSE分别降低至58.9和89.5,QNRF数据集的MAE

和 MSE 分别降低至 96.3 和 163.2,对应的相对误差分别降低了 3.1% 与 4.2% 以及 1.9% 与 3.9%。该结果清晰表明,数据增广策略在不同数据集上均能够有效提升模型的鲁棒性,有助于减少过拟合现象并提高重建结果的稳定性。

2.5.3 参数消融实验

本文对高频编码中的频带数量 L 与频率采样范围 $f_{\max}$ 进行细粒度分析。频带数量 L 表示编码器为每个输入坐标引入的正余弦基函数个数,即每个坐标维度将被映射为 L 组不同频率的 $\sin(\cdot)$ 与 $\cos(\cdot)$ 分量。对于二维坐标输入,未压缩的频谱编码维度与 L 呈线性关系。因此, L 越大,编码维度越高,模型能够表达更细粒度的空间变化。但与此同时,计算与显存开销也会随编码维度近似线性增长,并进一步放大后续跨域交互模块与解码器的计算负担,从而影响训练效率与推理速度。

为平衡表示能力与计算成本,本文在多个候选频带数量上进行了对比($L \in \{16, 32, 64\}$),并在不同频率采样范围下评估性能变化,结果如图 5 与图 6 所示。当 L 从 16 增加到 32 时,模型性能提升更为明显,说明增加频带能够有效增强模型对局部细节与密集区域的刻画能力;但当 L 从 32 增加到 64 时,性能增益趋于饱和,提升幅度明显减弱。其主要原因在于,在给定最大频率采样范围 $f_{\max}$ 的情况下,继续增加频带数量会使频率分布更密集并引入更多高频分量,这些分量对输入坐标扰动更敏感,容易放大噪声与采样误差,并使模型更倾向于拟合局部高频振荡而非稳定的密度变化趋势,从而导致收益边际递减。即超过一定规模后,频带数量的增加更多带来敏感性与复杂度而非有效表征能力的增长。因此,本文最终将频带数量 L 确定为 32。

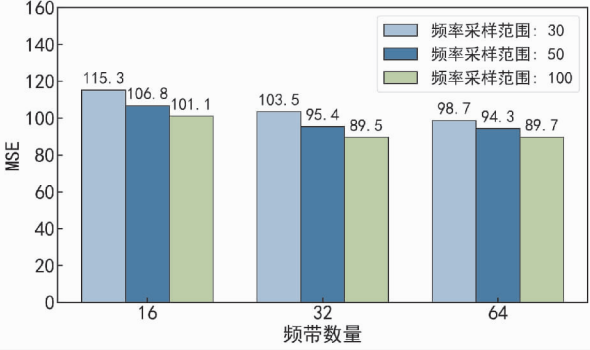


图 5 不同频率采样范围下的 MSE

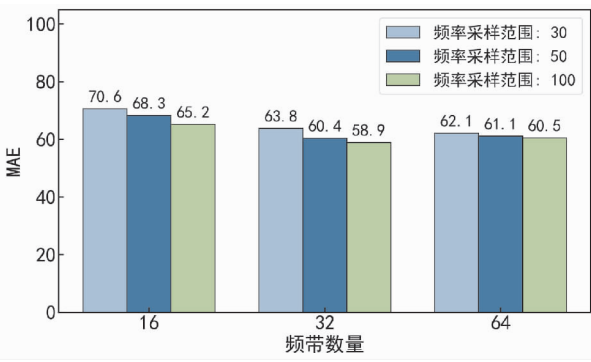


图 6 不同频率采样范围下的 MAE

在频率采样范围方面,通过设置不同的 $f_{\max}$ 验证其对模型的影响。较小的频率上界会限制编码的高频表达能力,使模型在密集区域的细节区分不足;而过大的频率上界会引入过强的高频分量,使表示对坐标扰动高度敏感,训练更不稳定,并可能造成泛化性能下降。综合图 5 与图 6 的实验结果,最终选择 $f_{\max} = 50$ 作为频率采样范围的上界,在表达能力与计算复杂度之间取得更好的折中。

为了验证全局计数损失的有效性,本文对损失项中权重 λ_{count} 设置为多个离散取值,并在相同的训练配置下分别进行实验,结果如图 7 与图 8 所示。从实验结果可见,当不加入全局计数损失时,模型的 MAE 和 MSE 分别为 62.1 与 101.3。当 λ_{count} 从 0 增加到 0.1 时,MAE 和 MSE 均显著下降,MAE 从 62.1 降至 58.9,MSE 从 101.3 降至 89.5,表明在引入一定权重时,模型性能得到了明显提升。然而,随着 λ_{count} 进一步增大至 0.2 及以上,模型性能均出现明显恶化。说明过高的权重会引发模型对该项的过拟合,从而削弱整体性能。因此,本文最终将 λ_{count} 设置为 0.1。

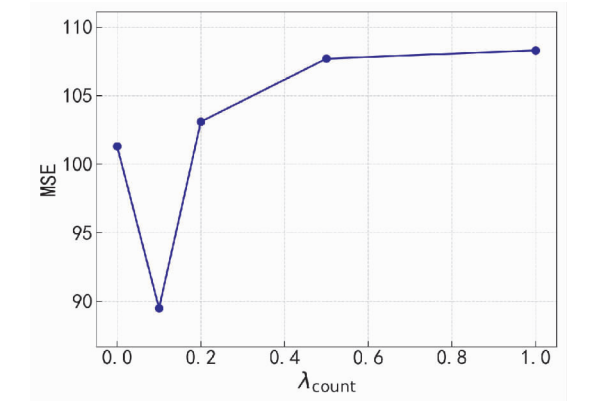


图 7 不同 λ_{count} 在 SHA 数据集上的 MSE

Fig. 5 MSE for various frequency sampling intervals

Fig. 7 MSE of different λ_{count} on SHA dataset

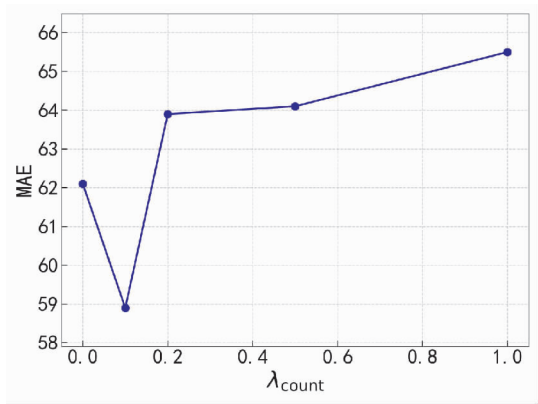


图8 不同 λ_{count} 在SHA数据集上的MAE
Fig. 8 MAE of different λ_{count} on SHA dataset

3 结论

1) 将隐式神经表示引入人群计数任务,通过轻量级多层感知机建立图像坐标到密度值的连续映射函数。该方法融合多尺度视觉特征与动态深度编码,实现分辨率无关的密度估计,有效缓解尺度敏感性与透视形变导致的密度分布偏差。

2) 利用高频编码器生成深度权重调节的位置特征,将其作为注意力机制的键值源,与视觉特征生成的查询进行局部窗口交叉计算。通过动态窗口划分策略自适应调整感受野,在降低计算复杂度的同时,实现密集区域的精细化特征对齐。

3) 提出端到端可学习的几何形变补偿框架,联合优化深度估计模块与多尺度特征金字塔,生成与图像透视变换匹配的动态采样网格。

参考文献

[1] 王怡婷. 基于深度学习的人群计数方法研究[D]. 北京:北京建筑大学,2024:1.

[2] 扶齐彦. 基于信道感知和迁移学习的复杂交通场景中密集车辆计数[D]. 南昌:南昌大学,2024:2.

[3] 陈慧. 基于多维度感知特征金字塔的人群计数算法[J]. 南通职业大学学报,2025,39(1):72-79.

CHEN H. Crowd counting algorithm based on multi-dimensional perceptual feature pyramid [J]. Journal of Nantong Vocational University, 2025, 39(1): 72-79.

[4] 杨月, 连淑娟, 马朝鹏, 等. 基于 HOD-Net 模型的教室密集场景人群计数方法[J/OL]. 计算机应用与软件. (2025-04-18) [2025-12-10]. <https://kns.cnki.net/KCMS/detail/detail.aspx?filename=JYRJ20250417001&dbname=CJFD&dbcode=CJFQ>.

YANG Y, LIAN S J, MA C P, et al. Dense crowd

counting method for classroom scenes based on HOD-Net model [J/OL]. Computer Applications and Software. (2025-04-18) [2025-12-10]. <https://kns.cnki.net/KCMS/detail/detail.aspx?filename=JYRJ20250417001&dbname=CJFD&dbcode=CJFQ>.

[5] 时东亮, 葛艳, 徐慕君. 基于多尺度特征加权融合注意力的密集人群计数算法[J]. 计算机系统应用, 2025, 34(3): 210-219.

SHI D L, GE Y, XU M J. Dense crowd counting algorithm based on multi-scale feature weighted fusion attention [J]. Computer Systems and Applications, 2025, 34(3): 210-219.

[6] 文帅, 蒋勇, 杨丹, 等. 基于多尺度注意力网络的密集人群计数[J]. 计算机应用与软件, 2025, 42(1): 130-136.

WEN S, JIANG Y, YANG D, et al. Dense crowd counting based on multi-scale attention network [J]. Computer Applications and Software, 2025, 42(1): 130-136.

[7] 郑棣文, 石洋宇, 谢承杰, 等. 多尺度感知与红外特征增强的 RGB-T 人群计数方法[J/OL]. 北京航空航天大学学报. (2024-09-19) [2025-12-10]. <https://link.cnki.net/urlid/11.2625.v.20240919.1254.003>.

ZHENG D W, SHI Y Y, XIE C J, et al. RGB-T crowd counting method with multi-scale perception and infrared feature enhancement [J/OL]. Journal of Beihang University. (2024-09-19) [2025-12-10]. <https://link.cnki.net/urlid/11.2625.v.20240919.1254.003>.

[8] 谢嘉豪. 基于多尺度超分辨率信息增强的人群计数方法[D]. 北京:北京邮电大学,2024:18-19.

[9] MOLAEI A, AMINIMEHR A, TAVAKOLI A, et al. Implicit neural representation in medical imaging: A comparative survey [C] // 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris: IEEE, 2023: 2373-2383.

[10] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: Representing scenes as neural radiance fields for view synthesis[M] // Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 405-421.

[11] PENG S Y, NIEMEYER M, MESCHEDER L, et al. Convolutional occupancy networks[C] // Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 523-540.

[12] TANCIK M, SRINIVASAN P P, MILDENHALL B, et al. Fourier features let networks learn high frequency functions in low dimensional domains [EB/OL].

- (2020-06-18) [2025-12-10]. <https://arxiv.org/abs/2006.10739>.
- [13] LIN C H, MA W C, TORRALBA A, et al. BARF: Bundle-adjusting neural radiance fields [C] // 2021 IEEE/CVF International Conference on Computer Vision. Montreal, QC: IEEE, 2021: 5721-5731.
 - [14] RAN Y, ZENG J, HE S, et al. NeurAR: Neural uncertainty for autonomous 3D reconstruction with implicit neural representations[J]. IEEE Robotics and Automation Letters, 2023, 8(2): 1125-1132.
 - [15] BARRON J T, MILDENHALL B, TANCIK M, et al. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields [C] // 2021 IEEE/CVF International Conference on Computer Vision. Montreal, QC: IEEE, 2021: 5855-5864.
 - [16] 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. Las Vegas, NV: IEEE, 2016: 589-597.
 - [17] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds [M] // Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 544-559.
 - [18] 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.
 - [19] ZHAO L, DAI H, WU Z, et al. CP-CNN: Core-periphery principle guided convolutional neural network [EB/OL]. (2023-03-27) [2025-12-10]. <https://arxiv.org/abs/2304.10515>.
 - [20] LI Y H, ZHANG X F, CHEN D M. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 1091-1100.
 - [21] LIU W Z, SALZMANN M, FUA P. Context-aware crowd counting [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 5094-5103.
 - [22] WANG B, LIU H, SAMARAS D, et al. Distribution matching for crowd counting [EB/OL]. (2020-09-28) [2025-12-10]. <https://arxiv.org/abs/2009.13077>.
 - [23] XIONG H P, LU H, LIU C X, et al. From open set to closed set: Supervised spatial divide-and-conquer for object counting [J]. International Journal of Computer Vision, 2023, 131(7): 1722-1740.
 - [24] GUO X Y, GAO M L, ZHAI W Z, et al. Scale region recognition network for object counting in intelligent transportation system [J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 15920-15929.
 - [25] CHEN J, GAO M, LI Q, et al. Privacy-aware crowd counting by decentralized learning with parallel transformers [J]. Internet of Things, 2024, 26: 101167.
 - [26] WANG T, ZHANG T, ZHANG K B, et al. Context attention fusion network for crowd counting [J]. Knowledge-Based Systems, 2023, 271: 110541.
 - [27] ZHAO Z Y, LI X L. Deformable density estimation via adaptive representation [J]. IEEE Transactions on Image Processing, 2023, 32: 1134-1144.
 - [28] ABOUSAMRA S, HOAI M, SAMARAS D, et al. Localization in the crowd with topological constraints [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 872-881.
 - [29] SAM D B, PERI S V, SUNDARARAMAN M N, et al. Locate, size, and count: Accurately resolving people in dense crowds via detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2739-2751.
 - [30] CAO X K, WANG Z P, ZHAO Y Y, et al. Scale aggregation network for accurate and efficient crowd counting [M] // Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 757-773.
 - [31] 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. Long Beach, CA: IEEE, 2019: 8190-8199.
 - [32] SONG Q Y, WANG C G, WANG Y B, et al. To choose or to fuse? Scale selection for crowd counting [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2576-2583.
 - [33] MA Z H, WEI X, HONG X P, et al. Bayesian loss for crowd count estimation with point supervision [C] // 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 6141-6150.
 - [34] CHENG J A, LI Q L, CHEN J Y, et al. Efficient vehicular counting via privacy-aware aggregation network [J]. Measurement Science and Technology, 2025, 36(2): 026213.