高技术通讯 2024 年 第 34 卷 第 9 期:945-959

doi:10.3772/j.issn.1002-0470.2024.09.004

基于区域信息聚合的轻量化人群计数方法①

席志国② 刘光辉③

(西安建筑科技大学信息与控制工程学院 西安 710055)
(建筑机器人陕西省高等学校重点实验室 西安 710055)
(西安市建筑制造智动化技术重点实验室 西安 710055)

摘 要 针对高密人群图像中细节信息丢失、背景噪声易与人群特征混淆以及网络模型复杂度高等问题,本文提出一种基于区域信息聚合的轻量化人群计数方法。首先,为获取高密图像中细粒度化的多尺度特征,设计了基于通道激活的多尺度特征提取模块,此模块通过引入 Ghost 卷积构建了层间分级类残差连接结构,同时对每级特征辅以通道激活,以轻量化的方式实现了网络感受野的逐级扩张。其次,提出一种自注意力区域信息聚合模块获取不同尺度区域的特征信息,该模块通过轻量级自注意力机制分别从通道和空间维度集成区域信息,增强对人群特征的关注,从而弱化背景噪声对计数的影响。最后,考虑到原始计数损失收敛过程中的不稳定性,在 DM-Count 损失的基础上引入一种新型计数损失,提高了模型稳定性和计数敏感性,进一步提升了计数性能。在 Shanghai Tech、UCF-QNRF、JHU-CROWD + + 以及 NWPU-Crowd 这 4 个公开数据集的实验结果表明,本文所提方法对比其他主流轻量级人群计数方法有一定的提升,且模型参数量保持在较低水平。

关键词 人群计数; 区域信息聚合; 轻量化; 自注意力; 损失函数

随着我国城镇化进程加快,城镇人口数量逐年上升,导致在城市车站、大型商场、旅游景点等公共场所多发生踩踏事件,民众安全面临着极大隐患[1]。因此,人群计数在道路规划、视频监控、安防预警等领域具有潜在的应用价值[2]。

传统的人群计数方法主要分为基于检测和基于回归 2 类,但这些方法因行人遮挡、空间信息缺失及特征提取效率不足等问题[3],影响了计数性能。近年来,卷积神经网络 (convolutional neural network, CNN) 因其出色的特征提取能力在人群计数领域得到了广泛应用[4]。例如,Li 等人[5]引入连续的膨胀卷积层以增大网络感受野,从而输出了高质量密度图。Chen 等人[6]设计的模型能够跨域进行无偏差学习,增强了模型的泛化能力。

尽管上述方法的计数性能较好,但模型复杂的网络结构导致推理速度降低。为了提高计数效率,现有方法通常采用轻量级特征提取网络或对原有网络结构进行优化以实现对人群特征的高效提取。其中,Chen 等人[7]通过组卷积降低网络复杂度,利用不同组之间的信息交换增强特征的表达能力,但因多列网络具有相似的学习功能而导致结构冗余。因此,Shi 等人[8]通过不同核大小的卷积层提取多尺度特征,提高了计数效率,但其特征融合结构相对简单,致使部分细节信息丢失。Wang 等人[9]通过知识蒸馏方法将复杂模型的特性迁移到其网络架构中。然而,因链式残差池化结构倾向于保留图中显著特征,导致网络忽略局部细节信息。Gao 等人[10]通过随机高级密度分类方法来提取全局特征,但该

① 国家自然科学基金(52278125) 资助项目。
② 男,1998 年生,硕士生;研究方向:人群计数与密度估计,模型轻量化;E-mail: x1661928747@163.com。
③ 通信作者,E-mail: guanghuil@163.com。
(收稿日期:2024-03-08)

方法的密度类别数量有限,未能涵盖不同密度场景,使得细节信息的捕获能力较差。虽然上述方法实现了计数性能和计算成本之间的平衡,但在捕获人群细节信息方面仍显不足。

另一方面,背景噪声易与人群特征混淆亦是亟需解决的问题。针对此问题,He 等人[11]通过学习前、背景区域的掩码来抑制背景噪声的干扰,但静态掩码无法较好地适应场景变化,导致模型区分人群与背景的能力较弱。因此,Thanasutives 等人[12]提出一种双路径网络,通过注意力图提升密度图的质量,但注意力图只关注空间位置信息,忽略了通道信息和通道间的相关性。张宇倩等人[13]利用注意力机制对通道信息重新校准以优化特征融合过程。然而,此方法忽略了空间位置信息,导致空间上下文信息部分丢失。张世辉等人[14]通过融入通道和空间注意力机制来解决人群与背景的特征相似性问题。但在全局上下文的理解和长距离空间位置依赖关系的捕捉上仍有不足,限制了网络在全局层面上的推理能力。

基于上述分析,本文提出一种基于区域信息聚合的轻量化人群计数方法,采用 GhostNetV2 架构前 6 层作为特征提取网络以降低模型复杂度。为了保留更多的人群细节特征,构建了基于通道激活的多尺度特征提取模块(multi-scale feature extraction based on channel activation,MFEC)以轻量级的方式逐级扩大网络的感受野,从而有效捕捉细粒度化的多尺度特征。其次,在信息融合部分设计自注意力区域信息聚合模块(self-attention area information aggregation,SAIA)获取不同尺度的区域特征以提升尺度表征能力,同时通过轻量级自注意力机制从通道和空间维度融合多尺度区域特征信息,强化人群特征,从而抑制背景噪声的干扰。最后,考虑到欧氏损失函数处理离群点的效果不佳,可能会损害网络的泛化性能,引入自适应平滑 L1 损失函数(self-adjusting smooth L1 loss,SS)对 DM-Count 损失进行改进,不仅提高了模型训练时的稳定性,还增强了网络对计数的敏感度,进一步提升了计数性能。

1　人群计数模型

基于区域信息聚合的轻量化人群计数网络结构如图 1 所示,包含基础骨干网络、基于通道激活的多尺度特征提取模块、自注意力区域信息聚合模块。

首先,利用基础骨干网络 GhostNetV2[15]部分结构生成初始特征图,通过 MFEC 重构层间分级类残差连接,同时对每级特征辅以通道激活,实现网络感受野的逐级扩张以捕捉特征图中细粒度化的多尺度特征。之后利用SAIA获取不同尺度的区域特征,

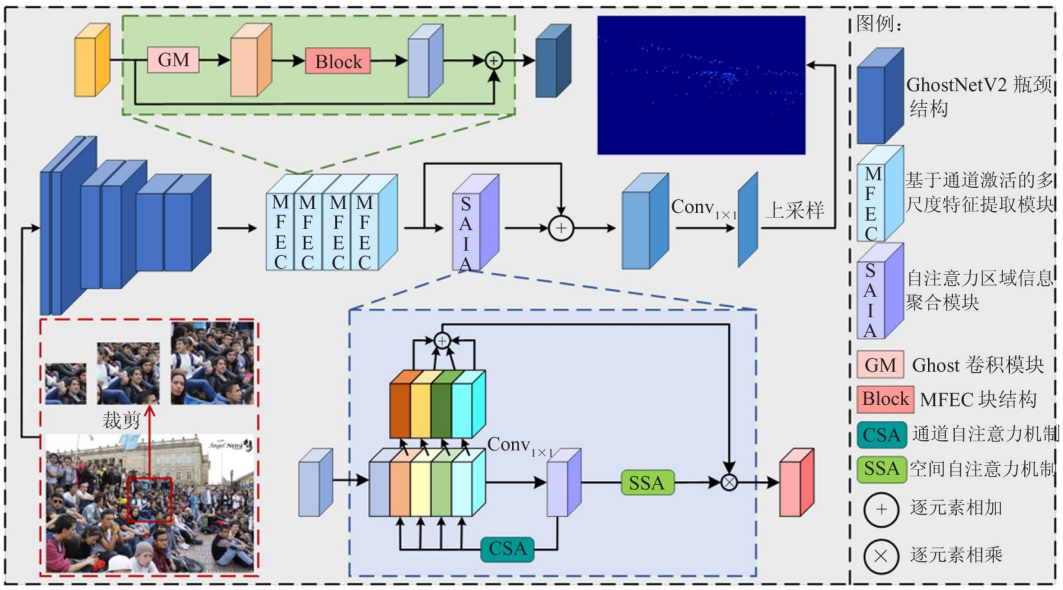


图 1　网络整体结构

通过轻量级自注意力机制集成不同尺度区域的特征信息,增强对人群特征的关注,从而弱化背景噪声的影响。最后通过逐点卷积进一步融合区域特征信息,并将其上采样至原始图像大小后输出密度图。

1.1 基于通道激活的多尺度特征提取模块

针对人群尺度变化问题,现有方法主要在不同分辨率下处理特征图以提取多尺度特征,通过直接合并方式实现多尺度表示,以此来适应图像中人群尺度变化的多样性。然而,这种多尺度特征获取方式可能会忽略细粒度层面的尺度特征,特别是在密度变化微小的高密场景中。此外,简单的特征合并方式无法充分利用多尺度特征间的相关性,导致特征融合效果不佳。Res2Net[16] 证明,其设计的块结构能够充分利用不同尺度信息,将提取的多尺度特征细粒度化。因此,本文提出一种基于通道激活的多尺度特征提取模块来捕捉人群细节信息,通过在骨架网络后嵌入 4 层 MFEC 来增强网络的细节信息提取能力,从而深入挖掘初始特征中细粒度化的多尺度特征,提升最终输出的密度图质量。

MFEC 以设计的块结构(block)为基础对原瓶颈结构进行重建,其结构如图 2 所示,分 2 阶段进行。

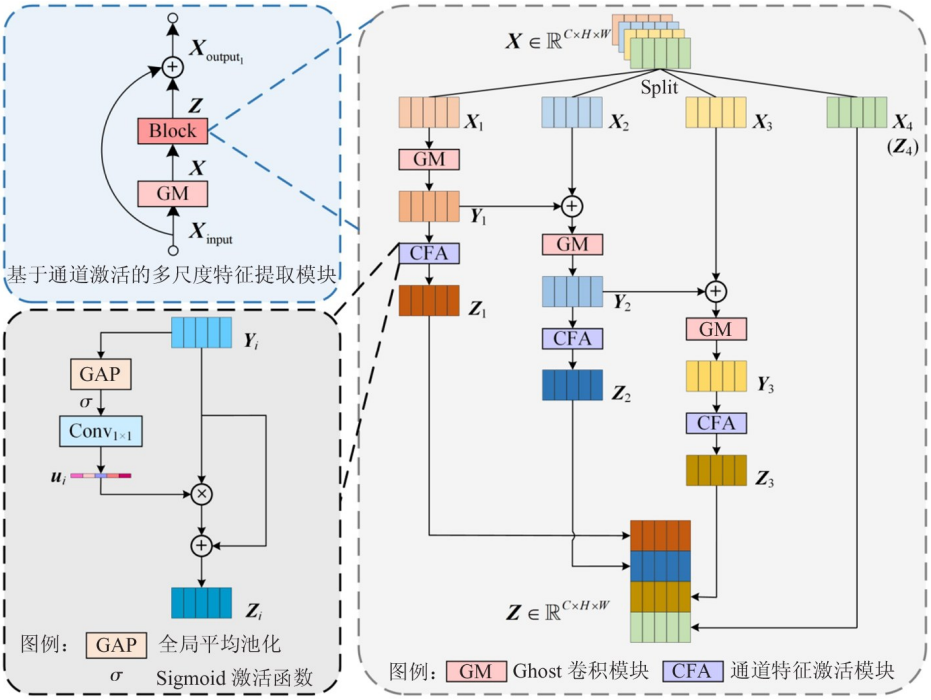


图 2 基于通道激活的多尺度特征提取模块

在第 1 阶段,引入 Ghost 卷积以轻量化的方式对输入特征图 $\boldsymbol{X}_{\text{input}} \in R^{C\times H\times W}$ 进行优化得到特征图 $\boldsymbol{X} \in R^{C\times H\times W}$,此阶段不仅降低了模型的计算复杂度,也间接增加了特征的丰富度。

进入第 2 阶段,通过 Block 模块逐级扩大网络的感受野。首先,将输入特征图 $\boldsymbol{X} \in R^{C\times H\times W}$ 在通道维度平分为 4 组特征,即 $\boldsymbol{X}_i \in R^{C'\times H\times W}(1 \leqslant i \leqslant 4)$,每组特征图的空间大小相同,其中 $C' = C/4$。对前 3 组特征分别进行 Ghost 卷积,得到中间输出 $\boldsymbol{Y}_i(1 \leqslant i \leqslant 3)$,中间输出 $\boldsymbol{Y}_i$ 可表示为

$$\boldsymbol{Y}_i = \begin{cases} G_i(\boldsymbol{X}_i) & i = 1 \\ G_i(\boldsymbol{X}_i \oplus \boldsymbol{Y}_{i-1}) & i = 2,3 \end{cases} \tag{1}$$

式中,⊕表示逐元素相加操作;$G_i(\cdot)$ 表示 Ghost 卷积操作;$G_i(\boldsymbol{X}_i \oplus \boldsymbol{Y}_{i-1})$ 表示将分组特征 $\boldsymbol{X}_i$ 与上一组的输出 $\boldsymbol{Y}_{i-1}$ 逐元素相加后执行 Ghost 卷积操作。通过上述方式,当前分组的中间输出 $\boldsymbol{Y}_i$ 融合了上组输出 $\boldsymbol{Y}_{i-1}$ 的特征信息从而实现感受野的逐级扩张。随着网络层的加深,每层特征都在聚合不同层级感受野的信息,最终使得网络输出层能够整合各层中不同大小感受野的信息,实现了对细节特征的深入

挖掘,提高了网络整体的信息提取能力。

其次,对每组的中间输出 $\boldsymbol{Y}_i$ 分别进行通道特征激活,其输出特征 $\boldsymbol{Z}_i$ 可描述为

$$\boldsymbol{Z}_i = \begin{cases} \boldsymbol{Y}_i \oplus (\boldsymbol{u}_i \otimes \boldsymbol{Y}_i) & i = 1,2,3 \\ \boldsymbol{X}_i & i = 4 \end{cases} \tag{2}$$

$$\boldsymbol{u}_i = \mathrm{Conv}_{1\times1}(\sigma(\mathrm{GAP}(\boldsymbol{Y}_i)))$$

式中,⊗表示逐元素相乘操作;GAP(·)表示全局平均池化操作;$\sigma(\cdot)$ 表示 Sigmoid 激活函数;$\mathrm{Conv}_{1\times1}(\cdot)$ 表示核大小为 1×1 的卷积。$\boldsymbol{u}_i = \mathrm{Conv}_{1\times1}(\sigma(\mathrm{GAP}(\boldsymbol{Y}_i)))$ 表示通过全局平均池化操作捕获每组中间输出 $\boldsymbol{Y}_i$ 的全局信息,同时利用 Sigmoid 激活函数和 1×1 卷积构建的特征重要性激活层学习全局信息以此对每组特征图的重要性建模,输出一个表示特征重要性的向量 $\boldsymbol{u}_i$。

然后,由 $\boldsymbol{Y}_i \oplus (\boldsymbol{u}_i \otimes \boldsymbol{Y}_i)$ 可知,将表示特征重要性的向量 $\boldsymbol{u}_i$ 与中间输出 $\boldsymbol{Y}_i$ 进行逐元素相乘,实现对原始特征中重要特征的强化,增强特征的空间相关性。再将相乘结果与 $\boldsymbol{Y}_i$ 逐元素相加,既避免通道激活响应被过度放大或抑制,又可保持原始特征信息的完整,从而进一步提高网络的特征提取能力。

最后,将每组的输出特征 $\boldsymbol{Z}_i$ 在通道维度合并,通过 1×1 卷积层获得包含每组不同尺度信息的特征 $\boldsymbol{Z}$,增强了网络的特征表达能力,特征 $\boldsymbol{Z}$ 的生成过程如下:

$$\boldsymbol{Z} = \mathrm{Conv}_{1\times1}(\boldsymbol{Z}_1 \Theta \boldsymbol{Z}_2 \Theta \boldsymbol{Z}_3 \Theta \boldsymbol{Z}_4) \tag{3}$$

式中,Θ 表示特征合并操作。此外,以跳跃连接(skip connection,SC)将输入特征 $\boldsymbol{X}_{\mathrm{input}}$ 和输出特征 $\boldsymbol{Z}$ 相连接,得到最终输出 $\boldsymbol{X}_{\mathrm{output}_1}$ 结束第 2 阶段,生成特征 $\boldsymbol{X}_{\mathrm{output}_1}$ 的具体过程可表示为

$$\boldsymbol{X}_{\mathrm{output}_1} = \boldsymbol{Z} \oplus \boldsymbol{X}_{\mathrm{input}} \tag{4}$$

引入跳跃连接的目的在于减少原始信息的丢失及避免训练中的梯度消失问题。

1.2　自注意力区域信息聚合模块

在人群计数领域,由于拥挤场景中背景噪声易与人群特征混淆导致网络模型难以区分人群与背景,极大影响了计数的准确性,应对上述问题的关键在于增强网络对人群特征的关注度,弱化背景噪声的影响。基于此,本文受文献[17]的启发设计了一种自注意力区域信息聚合模块,其结构如图 3 所示。

自注意力区域信息聚合模块通过将多个具有不同空洞率的卷积层进行稠密连接输出更密集的特征金字塔,从而打破尺度限制得到不同大小区域的特征信息。之后在轻量级自注意力机制的引导下从通道和空间维度对区域信息进行融合,使得模型能够更加准确地区分并强化人群特征,减少对无关信息与背景噪声的关注,从而在拥挤场景中减轻对行人的错误估计,其工作流程如下:

经过基础骨干网络和 MFEC 处理后,此时的特征图尺寸缩减至原始图像的 1/8。特征图 $\boldsymbol{X}_{\mathrm{output}_1}$ 上每个像素点映射回原图代表一个具体空间区域的信息聚合。首先将特征图 $\boldsymbol{X}_{\mathrm{output}_1}$ 通过卷积核大小为 3、空洞率大小为 1、2、5、9 的 4 个分支,使得模型能够在不增加参数量的情况下捕获更广泛区域的信息,从而得到不同尺度的区域特征 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$,

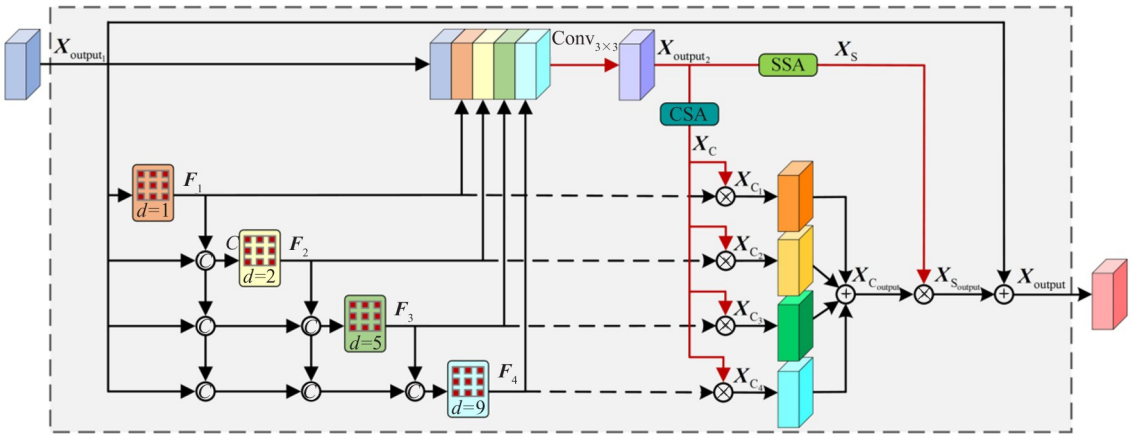


图 3　自注意力区域信息聚合模块

确保有效覆盖不同的空间区域进而加深网络对不同密度下人群特征的理解。而 4 个分支之间使用稠密连接的方式使得每个卷积层都直接与前面的所有层相连接,不仅能够从相应的分支中提取特定尺度的区域信息,同时也能够整合前面分支中的区域信息,实现信息的最大化共享与区域特征的充分融合,从而提高了特征的复用率并增强了模型对输入数据的表示能力,确保了深层网络中信息的丰富性与多样性。

然后,将 4 个分支的特征与输入特征 $\boldsymbol{X}_{\text{output}_1}$ 合并,并通过一个 3×3 大小的卷积层来融合上述特征以此获得特征图 $\boldsymbol{X}_{\text{output}_2}$:

$$\boldsymbol{X}_{\text{output}_2} = \text{Conv}_{3\times3}(\boldsymbol{X}_{\text{output}_1}\Theta\boldsymbol{F}_1\Theta\boldsymbol{F}_2\Theta\boldsymbol{F}_3\Theta\boldsymbol{F}_4) \tag{5}$$

式中,$\text{Conv}_{3\times3}(\cdot)$表示核大小为 3×3 的卷积。

最后,利用轻量级自注意力机制对特征图 $\boldsymbol{X}_{\text{output}_2}$ 从通道和空间维度建模,其结构如图 4 所示。

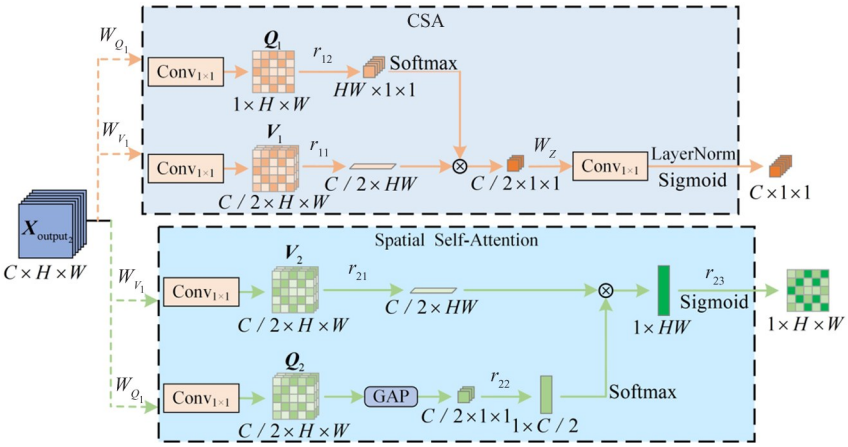


图 4　轻量级自注意力机制

对于通道自注意力(channel self-attention, CSA),首先,以输入特征图 $\boldsymbol{X}_{\text{output}_2} \in R^{C\times H\times W}$ 为例,通过 2 个 1×1 卷积层 W_{V_1} 与 W_{Q_1} 将特征图 $\boldsymbol{X}_{\text{output}_2}$ 转换成 $\boldsymbol{V}_1$ 和 $\boldsymbol{Q}_1$,接着利用整形算子 r_{11} 和 r_{12} 对 $\boldsymbol{V}_1$ 和 $\boldsymbol{Q}_1$ 进行形状变换。因为 $\boldsymbol{Q}_1$ 的通道维度被完全压缩,需要采用 Softmax 激活函数对其进行信息增强来提升注意力的动态范围,同时将 $\boldsymbol{V}_1$ 和 $\boldsymbol{Q}_1$ 进行逐元素相乘得到形状为 $C/2\times1\times1$ 的特征图。然后,经过 1×1 卷积层 W_Z 与 LayerNorm 恢复通道原本的维度 C,最终利用 Sigmoid 激活函数对特征进行动态映射后输出信息,利用通道自注意力强化输入特征 $\boldsymbol{X}_{\text{output}_2}$ 的过程可以表示为

$$\begin{aligned} &CSA(\boldsymbol{X}_{\text{output}_2}) \in R^{C\times1\times1} \\ &CSA(\boldsymbol{X}_{\text{output}_2}) = \sigma[\text{LN}(W_Z(r_{11}(W_{V_1}(\boldsymbol{X}_{\text{output}_2})) \\ &\qquad \otimes\delta(r_{12}(W_{Q_1}(\boldsymbol{X}_{\text{output}_2})))))] \end{aligned} \tag{6}$$

式中,W_{V_1} 表示通道数为 $C/2$、核大小为 1×1 的卷积;W_{Q_1} 表示通道数为 1、核大小为 1×1 的卷积;W_Z 表示通道数为 C、核大小为 1×1 的卷积;r_{11} 和 r_{12} 是改变张量形状的算子;δ 是 Softmax 激活函数;LN 是 LayerNorm 即层归一化;σ 是 Sigmoid 激活函数。

本文利用通道自注意力加强特征图 $\boldsymbol{X}_{\text{output}_2}$ 中通道间的相互作用,突出图中的人群特征,再将强化后的特征 $\boldsymbol{X}_C$ 与每个分支输出的区域特征 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$ 逐元素相乘,确保每个分支的输出能够更加准确地区分人群特征与背景噪声,从而提高模型对人群特征的识别准确性。然后,将每个分支相乘后的特征 $\boldsymbol{X}_{C_1}$、$\boldsymbol{X}_{C_2}$、$\boldsymbol{X}_{C_3}$、$\boldsymbol{X}_{C_4}$ 进行逐元素相加操作得到融合特征 $\boldsymbol{X}_{C_{\text{output}}}$,其过程可表示为

$$\begin{aligned} &\boldsymbol{X}_{C_i} = \boldsymbol{F}_i \otimes \boldsymbol{X}_C \quad (i = 1,2,3,4) \\ &\boldsymbol{X}_{C_{\text{output}}} = \boldsymbol{X}_{C_1} \oplus \boldsymbol{X}_{C_2} \oplus \boldsymbol{X}_{C_3} \oplus \boldsymbol{X}_{C_4} \end{aligned} \tag{7}$$

通过逐元素相加操作,每个分支捕捉到的关于图像不同尺度的区域特征得以融合,使得模型能够更加准确地区分不同尺度的人群特征,从而减轻对拥挤场景中行人的错误估计,也为后续结合空间维度的人群信息提供了较为细致的通道特征表示。

对于空间自注意力(spatial self-attention,SSA),首先,以输入特征图 $\boldsymbol{X}_{\text{output}_2} \in R^{C\times H\times W}$ 为例,通过 2 个 1×1 卷积层 W_{V_2} 与 W_{Q_2} 将特征图 $\boldsymbol{X}_{\text{output}_2}$ 转换成 $\boldsymbol{V}_2$ 和 $\boldsymbol{Q}_2$,对 $\boldsymbol{Q}_2$ 进行全局平均池化操作。其次,利用整形算子 r_{21} 和 r_{22} 对 $\boldsymbol{V}_2$ 和 $\boldsymbol{Q}_2$ 进行形状变换。因为 $\boldsymbol{Q}_2$ 在空间维度被完全压缩,以 Softmax 激活函数对其进行信息增强来提升注意力的动态范围,之后将 $\boldsymbol{V}_2$ 和 $\boldsymbol{Q}_2$ 进行逐元素相乘得到形状为 $1\times HW$ 的特征图。最终经过整形算子 r_{23} 恢复空间原始尺寸 $H\times W$,并利用 Sigmoid 激活函数对特征进行动态映射后输出信息,利用 SSA 强化输入特征 $\boldsymbol{X}_{\text{output}_2}$ 的过程可以表示为

$$\text{SSA}(\boldsymbol{X}_{\text{output}_2}) \in R^{1\times H\times W}$$

$$\text{SSA}(\boldsymbol{X}_{\text{output}_2}) = \sigma[r_{23}(\delta(r_{22}(\text{GAP}(\boldsymbol{W}_{Q_2}(\boldsymbol{X}_{\text{output}_2}))))) \otimes r_{21}(\boldsymbol{W}_{V_2}(\boldsymbol{X}_{\text{output}_2})))] \tag{8}$$

式中,W_{V_2} 和 W_{Q_2} 表示通道数为 $C/2$、核大小为 1×1 的卷积;r_{21}、r_{22} 和 r_{23} 是改变张量形状的算子;δ 是 Softmax 激活函数;GAP(·)是全局平均池化操作;σ 是 Sigmoid 激活函数。

本文利用空间自注意力聚焦于特征图 $\boldsymbol{X}_{\text{output}_2}$ 的空间关系,捕捉全局信息和上下文联系,增强了网络对空间结构的理解,同时将强化后的特征 $\boldsymbol{X}_S$ 与特征图 $\boldsymbol{X}_{C_{\text{output}}}$ 进行逐元素相乘操作得到特征 $\boldsymbol{X}_{S_{\text{output}}}$,其目的在于结合通道层面与空间层面的信息,从而形成更全面的特征表示,特征 $\boldsymbol{X}_{S_{\text{output}}}$ 生成过程如式(9)所示:

$$\boldsymbol{X}_{S_{\text{output}}} = \boldsymbol{X}_S \otimes \boldsymbol{X}_{C_{\text{output}}} \tag{9}$$

SAIA 模块的最终阶段通过跳跃连接将特征 $\boldsymbol{X}_{S_{\text{output}}}$ 与特征 $\boldsymbol{X}_{\text{output}_1}$ 进行逐元素相加操作得到最终特征 $\boldsymbol{X}_{\text{output}}$。引入跳跃连接的目的是保留原始输入数据的重要信息,防止在网络深层丢失,同时缓解梯度消失问题,确保网络的稳定训练,最终特征 $\boldsymbol{X}_{\text{output}}$ 可表示为

$$\boldsymbol{X}_{\text{output}} = \boldsymbol{X}_{S_{\text{output}}} \oplus \boldsymbol{X}_{\text{output}_1} \tag{10}$$

2　损失函数

虽然欧氏距离损失是人群计数领域的常用损失函数[18],但基于概率论的损失函数最近已经被证明能够显著增强网络的泛化能力,例如贝叶斯定理[19]和 Wasserstein 距离[20]。因此本文选择了 DM-Count(distribution matching for croud counting)损失函数,该损失主要通过最小化预测和真值密度图之间的 Wasserstein 距离来监督模型的训练。

DM-Count 整体损失函数包含计数损失、OT(optimal transport)损失和 TV(total variation)损失 3 部分,其中计数损失通过减小预测人数与真实人数之间的差异来提高模型的性能,计数损失可表示为

$$l_C = (z,\hat{z}) = |\ \|z\|_1 - \|\hat{z}\|_1\ | \tag{11}$$

式中,$z \in R_+^n$ 表示矢量化的真实密度图;$\hat{z} \in R_+^n$ 表示矢量化的预测密度图;$\|\cdot\|_1$ 表示向量的 L_1 范数;$\|z\|_1$ 和 $\|\hat{z}\|_1$ 分别代表实际人数和预测人数。

OT 损失基于最优传输理论,用于减小预测密度图和真实密度图之间的分布差异,从而提高模型在复杂和拥挤场景中的表现,OT 损失可定义为

$$\begin{aligned} l_{\text{OT}} &= \omega\left(\frac{z}{\|z_1\|}, \frac{\hat{z}}{\|\hat{z}\|_1}\right) \\ &= \left\langle \boldsymbol{\alpha}^*, \frac{z}{\|z\|_1}\right\rangle + \left\langle \boldsymbol{\theta}^*, \frac{\hat{z}}{\|\hat{z}\|_1}\right\rangle \end{aligned} \tag{12}$$

式中,z 和 $\hat{z}$ 表示矢量化的真实密度图和预测密度图,分别与各自人数总和相除转换为概率密度的表示,而 $\boldsymbol{\alpha}^*$ 和 $\boldsymbol{\theta}^*$ 则是式(13)的解。

$$\begin{aligned} &\omega(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\boldsymbol{\alpha},\boldsymbol{\theta}\in R^n} \langle \boldsymbol{\alpha},\boldsymbol{\mu}\rangle + \langle \boldsymbol{\theta},\boldsymbol{\nu}\rangle \\ &\text{s.t.}\quad \boldsymbol{\alpha}_i + \boldsymbol{\theta}_j \leqslant c(x_i, y_j),\ \forall i, j \end{aligned} \tag{13}$$

式中,$\boldsymbol{\mu}$ 和 $\boldsymbol{\nu}$ 分别代表 $\frac{z}{\|z\|_1}$ 和 $\frac{\hat{z}}{\|\hat{z}\|_1}$,是真实和预测密度图的概率密度分布(密度图均矢量化);$\boldsymbol{\alpha}$ 和 $\boldsymbol{\theta}$ 分别表示一组势函数,对应真实和预测密度图概率空间上的点;$c(x_i, y_j)$ 表示真实密度图位置 x_i 的点转移到预测密度图位置 y_j 的点之间的成本。此式的目的是找到 $\boldsymbol{\alpha}_i$ 和 $\boldsymbol{\theta}_j$ 的最优组合,使得整体传输成本最小化,同时满足成本函数 $c(x_i, y_j)$ 的约束条件。

TV 损失类似于重建损失,有助于平滑图像,减少预测密度图中的噪声,旨在增加训练过程中的稳定性,TV 损失可表示为

$$l_{\text{TV}} = (z,\hat{z}) = \left\| \frac{z}{\|z\|_1} - \frac{\hat{z}}{\|\hat{z}\|_1} \right\|_{\text{TV}}$$

$$= \frac{1}{2} \left\| \frac{z}{\|z\|_1} - \frac{\hat{z}}{\|\hat{z}\|_1} \right\|_1 \quad (14)$$

然而,在模型训练后期,当学习率恒定且预测与真实值非常接近时,由于计数损失在非零点的导数为常数以及零点处的不可导性,致使损失值接近最优解时在稳定值附近波动,不能平稳收敛且训练中可能出现梯度无法下降的问题。受文献[21]的启发,本文引入自适应平滑 L1 损失来解决上述问题,不仅提高了模型在训练时中的稳定性,还增强了网络对计数的敏感性,自适应平滑 L1 损失公式如下:

$$l_{SS}(x) = \begin{cases} 0.5\dfrac{x^2}{\beta} & |x| < \beta \\ |x| - 0.5\beta & \text{其他} \end{cases} \quad (15)$$

为方便表述 β, 将结合图 5 来详细解释。其中, x 代表 $\|z\|_1 - \|\hat{z}\|_1$, β 表示动态变化的控制点,其值被限制在 $[0,\hat{\beta}]$ 的区间内, $\hat{\beta}$ 被设置为 1/9。通过记录损失函数内部的滑动均值 μ_R 和滑动方差 σ_R^2 来确定 β 的值,并在动量为 0.9 的情况下利用小批量的滑动均值 μ_B 和方差 σ_B^2 来更新 μ_R 和 σ_R^2, 使得 β 成为动态变化的控制点, β 的求解公式可表示为

$$\begin{aligned} &\mu_B = \frac{1}{n}\sum_{i=1}^{n} |x_i|, \ \sigma_B^2 = \frac{1}{n}\sum_{i=1}^{n} (|x_i| - \mu_B)^2 \\ &\mu_R = \mu_R \times m + \mu_B \times (1 - m) \\ &\sigma_R^2 = \sigma_R^2 \times m + \sigma_B^2 \times (1 - m) \\ &\beta = \max(0, \min(\hat{\beta}, \mu_R - \sigma_R^2)) \end{aligned} \quad (16)$$

式中, x_i 表示 $(\|z\|_1 - \|\hat{z}\|_1)_i$, n 表示批量大小, m 表示动量。

由图5可以看出,自适应平滑L1损失曲线平滑

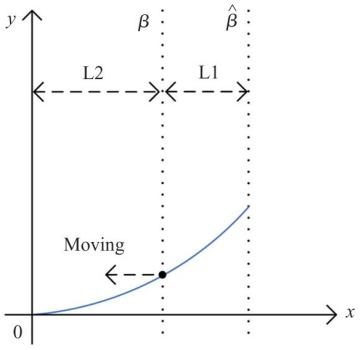


图 5　自适应平滑 L1 损失函数

且处处可导,既避免模型训练后期预测与真实人数差值较小时在稳定值附近波动而无法较好地收敛,又降低了模型训练初期预测与真实人数差值较大时对异常值的敏感性。此外,该损失还能根据实时数据的变化自动调整对不同大小损失值的响应程度,增强了网络对人群计数的敏感性,使得网络在估计人群密度时能够更敏锐地捕捉到预测值与真实值之间的微小差异,从而提高了计数精度。

最后,本文的总损失函数为

$$\begin{aligned} l(z,\hat{z}) = {} & l_{SS}(z,\hat{z}) + \lambda_1 l_{OT}(z,\hat{z}) \\ & + \lambda_2 \|z\|_1 l_{TV}(z,\hat{z}) \end{aligned} \quad (17)$$

式中, λ_1 和 λ_2 是 OT 和 TV 损失的可调权值,本文设置 $\lambda_1 = 0.1$、$\lambda_2 = 0.01$ 并用于所有数据集。为确保 TV 损失与自适应平滑 L1 损失具有相同的比例,将 TV 损失与真实人数进行相乘。

3　实验结果与分析

为验证基于区域信息聚合的轻量化人群计数方法性能,本文在 Shanghai Tech、UCF-QNRF、JHU-CROWD ++ 以及 NWPU-Crowd 这 4 个公开数据集上进行实验。其中,Shanghai Tech 和 UCF-QNRF 数据集分别发布于 2016 和 2018 年,被广泛用于评估人群计数模型,现有研究方法较为丰富;JHU-CROWD ++ 和 NWPU-Crowd 数据集的发布时间为 2020 年,其图像内的人群密度跨度更大、背景更复杂,具有更高的挑战性,且现有研究方法相对前者较少。因此,本文选择在这 4 个数据集上进行实验、对比和分析以全面评估本文方法的性能,确保了评估的公平性和深度,同时能够展示所提方法的竞争力和创新性。

本文与对比算法的实验均在 Ubuntu 系统下进行并基于 Pytorch 深度学习框架实现,GPU 型号为 RTX2080Ti。算法主干采用预训练过的 GhostNetV2 前 6 层,训练过程中选择了初始学习率为 1×10^{-4} 的 Adam 优化器,同时设置权重衰减系数为 1×10^{-4} 实现模型参数的优化。当训练的迭代次数达到 2 000次时,停止训练。此外,为了充分利用现有数据训练模型,采用数据增强方法对训练图像进行随机裁剪和水平翻转操作,以达到扩充数据集样本数

量的目的。本文将训练图像输入网络之前,分别对各组实验数据集进行不同尺寸大小的随机裁剪操作。对于 Shanghai Tech Part _ A 数据集,将随机裁剪出 256 像素 ×256 像素的子区域进行训练,对于 NWPU-Crowd 数据集,将随机裁剪出 384 像素 ×384 像素的子区域进行训练,而对于 Shanghai Tech Part _ B、UCF- QNRF 以及 JHU-CROWD ++ 数据集,则随机裁剪出 512 像素 ×512 像素的子区域进行训练。对于原始尺寸不足的数据集图像,将其等比例放大后再进行相应尺寸的裁剪处理。

3.1　评价指标

在人群计数领域中,平均绝对误差(mean absolute error, MAE)和均方根误差(root mean square error, RMSE)被广泛选用为评价指标,它们综合考虑了模型的预测准确性与鲁棒性,直观反映了模型预测与实际值之间的差异。为确保本文算法与现有算法能够进行实验对比分析,选用平均绝对误差及均方根误差作为指标评估模型的计数精度。

MAE 反映网络预测人数与真实人数之间的平均误差,强调每个样本的平均预测精度,直观反映了模型在整体数据集上的平均性能表现,从而帮助研究人员来评估模型的预测准确性,MAE 可定义为

$$f_{\mathrm{MAE}} = \frac{1}{N}\sum_{i=1}^{N} | C_i - C_i^{\mathrm{GT}} | \tag{18}$$

RMSE 则反映网络预测人数与真实人数之间的差异程度,通过加重对大误差的惩罚力度,使得其对模型预测中的大偏差更加敏感,能够捕捉到模型在极端条件下的表现差异,从而帮助研究人员来评估模型的鲁棒性,RMSE 可定义为

$$f_{\mathrm{RMSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i - C_i^{\mathrm{GT}})^2} \tag{19}$$

式中, N 为测试集包含的图像数量, C_i 和 C_i^{GT} 分别为第 i 幅测试图像的预测人数与真实人数。此外, MSE 和 RMSE 的数值越小代表着模型的性能越好。

峰值信噪比(peak signal to noise ratio, PSNR)和结构相似性指数(structural similarity index, SSIM)是评估预测与真值密度图之间相似度的常用指标,通过图像的均值、方差以及两者间的协方差来衡量 2 张图像的一致性。这 2 个指标的值越高,表明生成图像与真值图像的相似度越高,质量越好。

对于大小为 $H \times W$ 的图像 A 和噪声图像 B, PSNR 和 SSIM 计算公式如下:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_i^2}{MSE}\right) \tag{20}$$

$$SSIM(\mathrm{A},\mathrm{B}) = \frac{(2\mu_{\mathrm{A}}\mu_{\mathrm{B}} + C_1)(2\sigma_{\mathrm{AB}} + C_2)}{(\mu_{\mathrm{A}}^2 + \mu_{\mathrm{B}}^2 + C_1)(\sigma_{\mathrm{A}}^2 + \sigma_{\mathrm{B}}^2 + C_2)} \tag{21}$$

式中, MAX_i 是图片 i 可能的最大像素值, μ_{A} 和 μ_{B} 分别是图像 A 和 B 的均值, σ_{A}^2 和 σ_{B}^2 分别是图像 A 和 B 的方差, σ_{AB} 是图像 A 和 B 之间的协方差, C_1 和 C_2 是用来维持稳定性的常数,防止分母为 0。

3.2　Shanghai Tech 数据集的实验对比分析

Shanghai Tech 数据集[22]是人群计数领域中被广泛引用的基准数据集,由 1 198 幅详细标注的图像组成,总共标记了 330 165 个人头。该数据集被划分为 2 部分:Part _ A 和 Part _ B。Part _ A 数据集包含 482 幅图像,主要来源于互联网,特点是场景中人群高度拥挤。而 Part _ B 数据集则包含 716 幅图像,图像拍摄于上海的街道,特点是人群分布较为稀疏,规模变化较为明显。在实验中,选取 Part _ A 数据集的 300 幅图像和 Part _ B 数据集的 400 幅图像作为训练数据集,其余图像则分配为测试数据集。针对 Shanghai Tech 数据集单幅图像的可视化结果如图 6 所示,多种算法的性能指标结果对比如表 1 所示。在表中,本文将对比算法分为 2 类,第 1 类是拥有大量参数和复杂网络结构的算法,为表中上半部分,第 2 类是基于轻量级或简单网络结构的算法,为表中下半部分。此外,对于 2 类算法的最佳成绩和本文数据分别用黑色加粗字体表示,后文不在赘述此项。

从图 6 中测试图像的可视化结果可知,本文方法能够得到质量较高的预测密度图。具体来说,预测生成的密度图与实际场景的人群分布具有高度一致性,准确捕捉了人群的高密度区域及边界,同时在低密度区域也能维持良好的分辨率,表明了模型出色的细节表现能力。图中的预测人数结果进一步证明了所提方法的计数性能较为优秀。

由表 1 中的实验对比结果可知,本文所提方法以极少的参数量达到了与第 1 类复杂算法相近的性能水平,而与第 2 类轻量级算法的对比中,本文方法

(a) 原图　(b) 真值图　(c) 预测图

图 6　Shanghai Tech 数据集实验结果

表 1　Shanghai Tech 数据集多算法性能指标对比结果

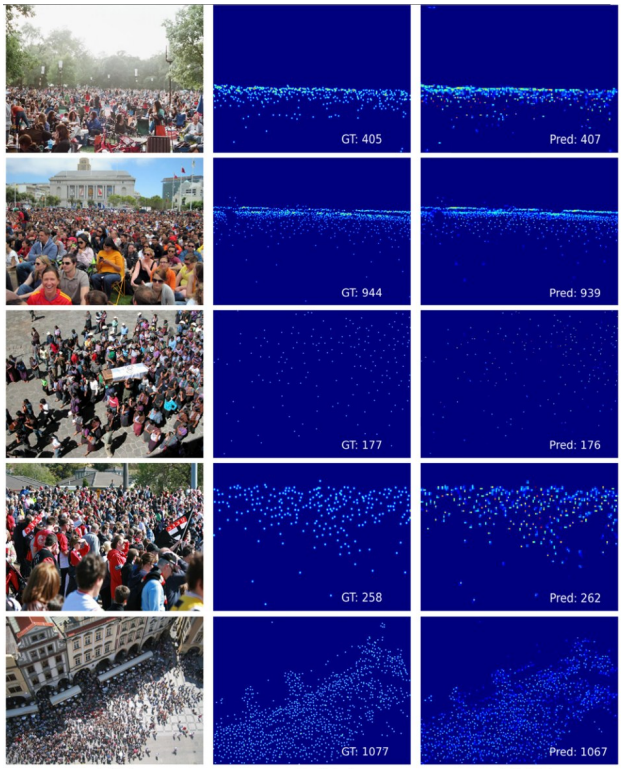
算法	Part _ A		Part _ B		参数量
	MAE	*RMSE*	*MAE*	*RMSE*	$/\times 10^6$
CSRNet[5]	68.2	115.0	10.6	16.0	16.26
CAN[4]	62.3	100.0	7.8	12.2	18.10
BL[19]	62.8	101.8	7.7	12.7	21.50
DUBNet[23]	64.6	106.8	7.7	12.5	18.05
DM-Count[20]	59.7	95.7	7.4	11.8	21.50
DKPNet[6]	**55.6**	**91.0**	**6.6**	**10.9**	30.63
MIA[3]	59.4	96.2	7.7	11.9	**13.40**
MCNN[22]	110.2	173.2	26.4	41.3	0.13
PCC-Net[10]	73.5	124.0	11.0	19.0	0.55
C-CNN[8]	88.1	141.7	14.9	22.1	**0.07**
Mobile-Count[9]	98.6	162.9	9.1	15.1	3.34
LMSFFNet[2]	85.9	139.9	9.2	15.1	4.58
本文方法	**65.4**	**102.7**	**7.5**	**12.3**	**0.33**

以较少的参数量实现了超越此类算法的性能表现。这一结果表明本文方法在参数利用效率上的显著优势,以及在人群计数任务中具有较高的计数准确度。此外,在 Part _ A 数据集,本文方法与文献[10]中的方法相比,MAE 降低了 8.1,RMSE 降低了 21.3;在 Part _ B 数据集,与文献[10]中的方法相比,MAE 降低了 3.5,RMSE 降低了 6.7,但与文献[20]中的方法相比差距甚微,这主要是由于本文方法能够有效抑制场景中的背景噪声,而 Part _ B 数据集中人群分布相对稀疏,同时伴有较多的背景噪声,因此所提方法在 Part _ B 数据集的提升效果更加明显。

3.3　UCF-QNRF 数据集的实验对比分析

近年来,研究人员对于网络模型在高密场景下的计数性能表现出越来越多的关注,为了准确评估模型的计数性能,引入 UCF-QNRF 数据集[24]进行实验。该数据集来源于拍摄的真实场景,既涵盖了建筑物、植被、天空、道路等多样化的背景环境,又包含多样的视角、密度和光照变化。UCF-QNRF 数据集具有 1 535 张图像,选取其中 1 201 张图像作为训练数据集,其余 334 张图像作为测试数据集。UCF-QNRF 数据集单幅图像的可视化结果如图 7 所示,多种算法的性能指标结果对比如表 2 所示。

由图 7 中测试图像的可视化结果可以看出,在场景复杂多变的 UCF-QNRF 数据集中,本文方法生成的预测密度图能够较为准确地反映出真实场景中



(a) 原图　(b) 真值图　(c) 预测图

图 7　UCF-QNRF 数据集实验结果

表 2　UCF-QNRF 数据集多算法性能指标对比结果

算法	*MAE*	*RMSE*	参数量/ $\times 10^6$
CSRNet[5]	120.3	208.5	16.26
CAN[4]	107.0	183.0	18.10
BL[19]	88.7	154.8	21.50
DUBNet[23]	105.6	180.5	18.05
DM-Count[20]	85.6	148.3	21.50
DKPNet[6]	**81.4**	**147.2**	30.63
MIA[3]	89.3	164.5	**13.40**
MCNN[22]	**277.0**	**426.0**	**0.13**
PCC-Net[10]	148.7	247.3	0.55
MobileCount[9]	137.8	238.2	3.34
LMSFFNet[2]	112.8	201.6	4.58
本文方法	**91.8**	**156.1**	**0.33**

的人群分布,预测人数与真实人数之间的较小差值也证明所提方法较为优秀的计数准确性。

由表 2 中的实验对比结果可知,在 UCF-QNRF 数据集中,本文所提方法与文献[9]中算法相比,MAE 下降了 46.0,RMSE 下降了 82.1,相较于第 2 类轻量级算法均有提升。而本文方法与第 1 类复杂算法中文献[3]的方法相比,MAE 和 RMSE 都相差较小,这表明本文所提方法不仅具有较好的精度,而且显著减少了网络参数量。此外,UCF-QNRF 数据集中的复杂场景导致多样的视角、密度和光照变化,而实验结果表明本文提出的自注意力区域信息聚合模块在减轻上述因素带来的影响上表现较为出色,同时在处理复杂场景时具有良好的鲁棒性。

3.4　JHU-CROWD + + 数据集的实验对比分析

近年来,提高模型在复杂背景下的计数精度逐渐成为研究热点,用于复杂场景计数性能评价的数据集 JHU-CROWD ++ [25] 随之被提出。该数据集中的图像在各种不同场景和环境天气下被收集,包括低密度、中等密度、高密度、基于恶劣天气变化和光照变化的图像以及综合图像。JHU-CROWD ++ 数据集共有 4 372 张样本图像,本文选取其中 2 272 张样本图像作为训练数据集,500 张样本图像作为验证数据集,余下 1 600 张样本图像作为测试数据集。JHU-CROWD ++ 数据集单幅图像的可视化结果如图 8 所示,多种算法的性能指标结果对比如表 3 所示。

由图 8 中测试图像的可视化结果可以看出,本文方法在面对恶劣天气变化及光照波动影响等复杂场景时依然能够实现准确的人群计数和密度图预测效果。

由表 3 中的实验对比结果可知,在 JHU-CROWD ++ 数据集中,本文所提方法与文献[11]中算法相比,MAE 下降了 1.0,RMSE 下降了 0.4,相较于第 2 类轻量级算法均有提升。虽然较文献[11]中算法相比在计数性能指标上提升幅度较小,但本文方法的参数量却大幅减少,表明本文方法在参数利用效率方面的显著优势。而本文方法与第 1 类复

(a) 原图　(b) 真值图　(c) 预测图

图 8　JHU-CROWD + + 数据集实验结果

表 3　JHU-CROWD ++ 数据集多算法性能指标对比结果

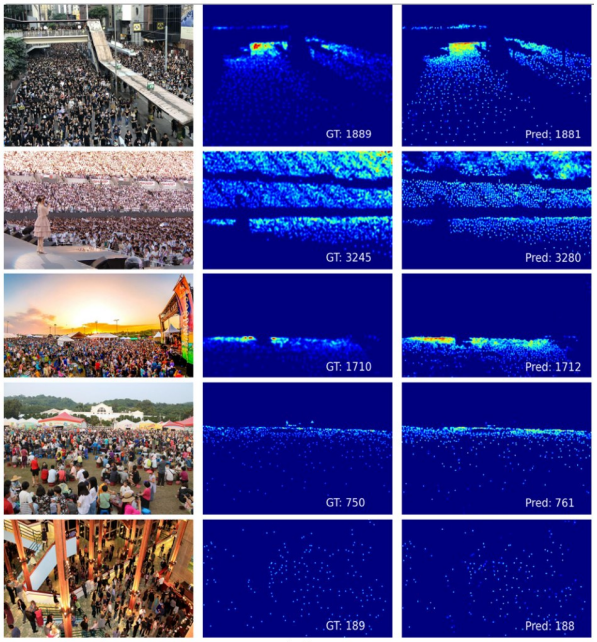
算法	*MAE*	*RMSE*	参数量/ $\times 10^6$
CSRNet[5]	85.9	309.2	**16.26**
CAN[4]	89.5	**239.3**	18.10
BL[19]	75.0	299.9	21.50
SFCN[26]	**62.9**	247.5	38.60
MCNN[22]	188.9	483.4	**0.13**
SANet[27]	91.1	**320.4**	0.91
GAENet[11]	77.9	320.9	1.97
本文方法	**76.9**	**320.5**	**0.33**

杂算法中文献[19]的方法相比,在 MAE 和 RMSE 上仅有较小差距,这表明本文所提方法不仅显著降低了网络的参数量,而且网络能够保持良好的计数精度。此外,面对 JHU-CROWD ++ 数据集中基于恶劣天气变化、光照变化等复杂场景,本文提出的自注意力区域信息聚合网络能够减轻背景噪声带来的影响,使得网络模型可以保持较高的计数精度,表明所提方法在处理复杂场景时具有较好的鲁棒性。

3.5 NWPU-Crowd 数据集的实验对比分析

NWPU-Crowd 数据集[28]包含 5 109 张图像和 2 133 238个标注对象。该数据集不仅涵盖了从极低到极高密度的人群图像,还特别包含了负样本,如无人的背景图像,以增强模型在面对非典型场景时的鲁棒性。每张图像中的标注对象数量差异较大,范围从 0 ~ 20 033 人。本文选取 NWPU-Crowd 数据集中的 3 109 张图像作为训练数据集,500 张图像作为验证数据集,其余 1 500 张图像用于测试。NWPU-Crowd 数据集单幅图像的可视化结果如图 9 所示,多种算法的性能指标结果对比如表 4 所示。

从图 9 中测试图像的可视化结果可以得出,本文方法在高密度的真实场景中也能够精确估计人群数量,同时预测出质量较高的密度图,证明了本文方法在计数性能方面的有效性。



(a) 原图　(b) 真值图　(c) 预测图

图 9　NWPU-Crowd 数据集实验结果

表 4　NWPU-Crowd 数据集多算法性能指标对比结果

算法	Val		Test		参数量
	MAE	*RMSE*	*MAE*	*RMSE*	$/\times 10^6$
CSRNet[5]	104.8	433.4	121.3	387.8	16.26
CAN[4]	93.5	489.9	106.3	386.5	18.10
BL[19]	93.6	470.3	105.4	454.2	21.50
DM-Count[20]	70.5	**357.6**	88.4	388.6	21.50
DKPNet[6]	**61.8**	438.7	**74.5**	**327.4**	30.63
MIA[3]	87.9	417.2	101.2	418.5	**13.40**
MCNN[22]	218.5	700.6	232.5	714.6	**0.13**
SANet[27]	—	—	190.6	491.4	0.91
本文方法	**74.6**	**216.8**	**96.7**	**351.3**	**0.33**

由表 4 中的实验对比结果可知,本文方法以极低的参数量实现了与第 1 类复杂算法相近的性能,而在第 2 类轻量级算法的对比中,本文方法又以较少的参数量差距达到了超越该类算法的性能水平。与文献[3]中方法相比,本文方法在 Val 验证数据集下的 MAE 下降了 13.3,RMSE 下降了 200.4,在 Test 测试数据集下的 MAE 下降了 4.5,RMSE 下降了 67.2,验证了本文方法在应对 NWPU-Crowd 数据集中负样本时能够保持较高计数精度以及参数利用方面的高效性,表明本文算法在高分辨率下不同密度的场景中具有较好的鲁棒性。

3.6 消融实验与速度对比分析

为验证所提方法中基于通道激活的多尺度特征提取模块 MFEC、自注意力区域信息聚合模块 SAIA 以及自适应平滑 L1 损失函数的有效性,本文选择在 Shanghai Tech 数据集 Part _ A 部分上仅用基础骨干网络进行消融实验,基础损失函数为 DM-Count 损失函数,将自适应平滑 L1 损失替换原有计数损失部分后组成新的联合损失用 SSL-DM(self-adjusting smooth loss for distribution matching)表示,实验结果如表 5 所示。

表 5　网络整体消融实验结果

算法	*MAE*	*RMSE*
Backbone	81.2	127.7
Backbone + 4MFEC	72.7	118.0
Backbone + 4MFEC + SAIA	68.9	104.1
Backbone + 4MFEC + SAIA + SSL-DM	65.4	102.7

由表 5 可知,在基础骨干网络后嵌入 4 层 MFEC 使得网络 MAE 下降 10.46%,RMSE 下降 7.60%,MAE 较 RMSE 提升幅度明显,这一结果证明了 MFEC 模块能够有效增强网络的细节信息提取能力,从而提升了计数精度。进一步地,将自注意力区域信息聚合模块 SAIA 引入网络后,使得网络 MAE 下降 5.20%,RMSE 下降 11.80%,RMSE 较 MAE 提升幅度明显,这一结果验证了 SAIA 模块既能够优化特征的表示,也增强了网络对于人群和背景噪声的区分能力,从而在复杂场景下具有较好的鲁棒性。把基础损失函数计数损失部分替换为自适应平滑 L1 损失函数后,网络 MAE 下降 5.10%,RMSE 下降 1.30%,可以看出新的联合损失函数 SSL-DM 不仅提高了模型的稳定性,还增强了对计数任务的敏感性,进一步提升了计数精度。由此表明 MFEC、SAIA 及新的联合损失函数 SSL-DM 均对网络的计数精度及鲁棒性有所提高。

为验证 SAIA 模块中通道自注意力 CSA 和空间自注意力 SSA 的有效性,本文在 Shanghai Tech Part _ A 数据集上对 SAIA 进行消融实验,将利用稠密连接结合的 4 个空洞卷积用 D-ASPP(densely connected atrous spatial pyramid pooling)表示,跳跃连接用 SC 表示,基础模块为骨干网络和 4 层 MFEC,基础损失函数为 DM-Count 损失函数,实验结果如表 6 所示。

由表 6 可知,在基础网络后连接无内嵌自注意力机制的 SAIA 模块即 D-ASPP 能够使得网络 MAE 下降 3.4%,RMSE 下降 3.3%。随后,在上述网络分别单独融入通道自注意力机制和空间自注意力机制,并且使用跳跃连接 SC 结束特征融合过程,两种自注意力机制均使得网络 MAE 和 RMSE 有所降低,从而验证了它们各自的有效性。此外,相较于仅融入通道或空间自注意力机制引导网络,通道和空间自注意力机制之间协同作用于网络能够进一步提升模型的计数性能,相比仅融入通道自注意力机制的网络,MAE 降低了 0.5,RMSE 降低了 7.6,而相比仅融入空间注意力机制的网络,MAE 降低了 0.8,RMSE 降低了 8.2。由此证明了 SAIA 模块中通道和空间自注意力机制是相互协同的关系,共同作用于网络,而非互相抵消或削弱彼此的效果。

表 6　SAIA 消融实验结果

算法	*MAE*	*RMSE*
Backbone + 4MFEC	72.7	118.0
Backbone + 4MFEC + D-ASPP	70.2	114.0
Backbone + 4MFEC + D-ASPP + CSA + SC	69.4	111.7
Backbone + 4MFEC + D-ASPP + SSA + SC	69.7	112.3
Backbone + 4MFEC + D-ASPP + CSA + SSA + SC	68.9	104.1

为验证本文方法在推理速度上的优越性,在 Shanghai Tech Part _ B 数据集上进行实验,以大小为 1 024 像素 × 768 像素的图像作为输入并与先进算法进行对比实验,从而对训练模型参数量大小、计算量大小及推理速度进行比较分析,证明所提算法在推理速度上的优势,使用的 GPU 型号为 GTX 1650 SUPER。为准确评估模型推理速度的性能,将测试 300 轮,每轮耗时被累加求和,随后得到总耗时的平均值作为最终推理速度的结果,实验结果如表 7 所示。

表 7　模型大小与推理速度实验结果

算法	*MAE*	*RMSE*	参数量/ $\times 10^6$	浮点操作/GFLOPs	时间/ms
CSRNet[5]	10.6	16.0	**16.26**	**325.66**	**154.8**
CAN[4]	**7.8**	**12.2**	18.10	367.58	178.0
MCNN[22]	26.4	41.3	**0.13**	21.46	28.3
SANet[27]	8.4	13.6	0.91	71.45	103.1
PCC-Net[10]	11.0	19.0	0.55	72.80	62.3
MobileCount[9]	9.1	15.1	3.34	6.31	23.8
LMSFFNet[2]	9.2	15.1	4.58	14.90	102.4
本文方法	**7.5**	**12.3**	**0.33**	**4.41**	**23.7**

根据表 7,所提方法在第 1 类复杂算法和第 2 类轻量级算法中取得了最佳的 MAE 与次佳的 RMSE,同时大幅降低了模型的参数量及计算量,参数量以 0.33×10^6 的结果取得了次佳成绩,而计算量则以 4.41 GFLOPs 的结果取得了最佳成绩。在推理时间上对比 CSRNet[5]、CAN[4]以及 SANet[27]等上述方法也以 23.7 ms 的结果取得了最佳成绩,而上述模型在推理时间上大约是本文方法的 1.0 到 7.5 倍,证明了所提方法不仅对于实时计数具有重要价值,而且在计数性能和计算成本上取得了较为优秀的平衡。

3.7 预测密度图与真值密度图的分析

由于在人群计数领域中采用 PSNR 和 SSIM 作为评价指标的方法相对较少,其中多数方法选择在 Shanghai Tech Part _ A 数据集上进行实验,为了全面、公平的比较所提方法与现有方法在密度图质量上的性能差异,同时为验证所提方法可以生成更高质量的预测密度图,本文以 PSNR 和 SSIM 作为图像相似度评价指标,并选择在具代表性的 Shanghai Tech Part _ A 数据集进行实验。表 8 列出了不同算法之间的比较结果。

表 8 Shanghai Tech Part _ A 数据集不同算法的密度图质量比较结果

算法	*PSNR*	*SSIM*
Switching-CNN[18]	21.9	0.67
CP-CNN[29]	21.7	0.72
CSRNet[5]	23.8	0.76
DADNet[30]	24.2	0.81
ADCrowdnet[31]	24.5	0.88
文献[32]方法	25.3	0.90
MCNN[22]	21.4	0.52
PCC-Net[10]	22.8	0.74
本文方法	25.4	0.86

PSNR 是一种衡量图像重建质量的常用指标,其值较高意味着重建质量较好;SSIM 是衡量 2 幅图像间相似度的指标,其值越接近 1 意味着两幅图像之间的相似度就越高,值为 1 时则表示两幅图像完全相同。根据表 8,所提方法在第 1 类复杂算法中与最优数据相比,PSNR 值略有胜出,表明所提方法能更准确地重建原始人群分布的密度图,减少了重建图像中的误差。SSIM 值则相差 0.4,这是因为文献[32]专注于弱化背景噪声,从而得到了结构相似度更高的人群密度图。但在第 2 类轻量级算法中所提方法的数据较为优秀,由此证明,所提方法对像素误差具有更低的敏感性,减少了图像重建误差,而对图像人群分布则具有较高的结构相似性,提高了人群分布的结构相似度,从而能够较好地拟合真实的人群分布,有效提升了预测密度图的质量。

4 结 论

本文构建了一种基于区域信息聚合的轻量化人群计数方法,以 GhostNetV2 部分结构作为骨干网络,显著降低了网络模型复杂度。通过设计基于通道激活的多尺度特征提取模块,有效提高了网络对人群细微尺度变化的捕捉能力以及对细节特征的提取能力。随后,设计的自注意力区域信息聚合模块打破了尺度限制,得到了不同尺度的区域特征,在轻量级自注意力机制的引导下从通道和空间维度建模并集成不同区域信息,强化了人群特征的表达,抑制了背景噪声的干扰,从而提升了模型区分人群与背景的能力,解决了人群细节特征较难提取和背景噪声和人群特征区分困难的问题。最后提出一种联合损失函数,既提高了模型的稳定性,又加强了网络对计数的敏感性,从而进一步提升了计数精度。在多个数据集上的实验结果证明:所提算法在参数量上进行了优化,不仅稳定性良好,泛化能力较强,而且模型在多元场景下对不同密度的人群都具有较高的计数性能。但所提方法仍有其局限性,在面对极端密集或稀疏的人群场景时,性能仍有待提高,如何在上述场景中进一步挖掘更多富有辨识度的细节特征和提高网络对人群与背景的区分能力是未来工作中改进的重点。

参考文献

[1] NAWARATNE R, KAHAWALA S, NGUYEN S, et al. A generative latent space approach for real-time road surveillance in smart cities[J]. IEEE Transactions on In-

dustrial Informatics, 2020,17(7):4872-4881.

[2] YI J, SHEN Z, CHEN F, et al. A lightweight multiscale feature fusion network for remote sensing object counting [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023,61:1-13.

[3] 刘光辉, 王秦蒙, 陈宣润, 等. 多元信息聚合的人群密度估计与计数[J]. 光学精密工程, 2022,30(10):1228-1239.

[4] LIU W, SALZMANN M, FUA P. Context-aware crowd counting[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:5099-5108.

[5] LI Y, ZHANG X, CHEN D. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:1091-1100.

[6] CHEN B, YAN Z, LI K, et al. Variational attention: propagating domain-specific knowledge for multi-domain learning in crowd counting [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021:16065-16075.

[7] CHEN J, XIU S, CHEN X, et al. Flounder-Net: an efficient CNN for crowd counting by aerial photography[J]. Neurocomputing, 2021,420:82-89.

[8] SHI X, LI X, WU C, et al. A real-time deep network for crowd counting[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020:2328-2332.

[9] WANG P, GAO C, WANG Y, et al. MobileCount: an efficient encoder-decoder framework for real-time crowd counting[J]. Neurocomputing, 2020,407:292-299.

[10] GAO J, WANG Q, LI X. PCC Net: perspective crowd counting via spatial convolutional network [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019,30(10):3486-3498.

[11] HE Y, ZU X, XIA Y, et al. Ghost attention efficient network for crowd counting[C]//The 7th International Conference on Computer and Communication Systems (ICCCS). Wuhan, China: IEEE, 2022:289-293.

[12] THANASUTIVES P, FUKUI K, NUMAO M, et al. Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting[C]//2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE, 2021:2382-2389.

[13] 张宇倩, 李国辉, 雷军, 等. FF-CAM: 基于通道注意机制前后端融合的人群计数[J]. 计算机学报, 2021, 44(2):304-317.

[14] 张世辉, 赵维勃, 王磊, 等. MSIANet:多尺度交互注意力人群计数网络[J]. 电子与信息学报, 2023,45(6):2236-2245.

[15] TANG Y, HAN K, GUO J, et al. GhostNetv2: enhance cheap operation with long-range attention[J]. Advances in Neural Information Processing Systems, 2022, 35: 9969-9982.

[16] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: a new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019,43(2):652-662.

[17] YANG M, YU K, ZHANG C, et al. Denseaspp for semantic segmentation in street scenes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018:3684-3692.

[18] BABU SAM D, SURYA S, VENKATESH BABU R. Switching convolutional neural network for crowd counting [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017:5744-5752.

[19] MA Z, WEI X, HONG X, et al. Bayesian loss for crowd count estimation with point supervision[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019:6142-6151.

[20] WANG B, LIU H, SAMARAS D, et al. Distribution matching for crowd counting[J]. Advances in Neural Information Processing Systems, 2020,33:1595-1607.

[21] FU C Y, SHVETS M, BERG A C. RetinaMask: learning to predict masks improves state-of-the-art single-shot detection for free[EB/OL]. (2019-01-10)[2024-03-23]. https://arxiv.org/pdf/1901.03353.

[22] ZHANG Y, ZHOU D, CHEN S, et al. Single-image crowd counting via multi-column convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:589-597.

[23] OH M, OLSEN P, RAMAMURTHY K N. Crowd counting with decomposed uncertainty[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2020:11799-11806.

[24] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]//Proceedings of the European Conference on Computer Vision. Munich, Germany: ECCV, 2018:532-546.

[25] 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.

[26] WANG Q, GAO J, LIN W, et al. Pixel-wise crowd understanding via synthetic data[J]. International Journal of Computer Vision, 2021,129(1):225-245.

[27] CAO X, WANG Z, ZHAO Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]// Proceedings of the European Conference on Computer Vision. Munich, Germany: ECCV, 2018:734-750.

[28] WANG Q, GAO J, LIN W, 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.

[29] SINDAGI V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid CNNs[C] // Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017:1861-1870.

[30] GUO D, LI K, ZHA Z J, et al. DADNet: dilated-attention-deformable convnet for crowd counting[C] // Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM, 2019:1823-1832.

[31] LIU N, LONG Y, ZOU C, et al. ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019:3225-3234.

[32] 张译, 吴秦. 特征增强损失与前景注意力人群计数网络[J]. 计算机科学, 2023,50(3):246-253.

Lightweight crowd counting method based on regional information aggregation

XI Zhiguo, LIU Guanghui

(College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055)

(Higher Education Key Laboratory of Construction Robot in Shaanxi Province, Xi'an 710055)

(Xi'an Key Laboratory of Intelligent Technology for Building Manufacturing, Xi'an 710055)

Abstract

Aiming at the problems of loss of detailed information in high-density crowd images, background noise being easily confused with crowd features, and high complexity of network models, this article proposes a lightweight crowd counting method based on regional information aggregation. First, a multi-scale feature extraction module based on channel activation is designed to obtain fine-grained multi-scale features in high-density images. This module introduces Ghost convolution to construct an inter-layer hierarchical residual connection structure, and supplementing each level of features with channel activation, achieving a gradual expansion of the network's receptive field in a lightweight manner. Secondly, a self-attention region information aggregation module is proposed to obtain feature information from regions of different scales. This module integrates region information from both channel and spatial dimensions using a lightweight self-attention mechanism, enhancing focus on crowd features to weaken the impact of background noise on counting. Finally, considering the instability in the convergence process of the original count loss, a new counting loss is introduced based on the DM-Count loss, which improves the model stability and counting sensitivity, and further improves the counting performance. Experimental results on four public data sets of Shanghai Tech, UCF-QNRF, JHU-CROWD ++ , and NWPU-Crowd show that the method proposed in this thesis has a specific improvement compared with other mainstream lightweight crowd counting methods, and the number of model parameters remains relatively low-level.

Key words: crowd counting, regional information aggregation, lightweight, self-attention, loss function