CN 51-1346/O4 ISSN 1003-501X (印刷版) ISSN 2094-4019 (网络版)

光电工程
Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

GLCrowd: 基于全局-局部注意力的弱监督密集场景人群计数模型

张红民，田钱前，颜鼎鼎，卜令宇

引用本文:

张红民，田钱前，颜鼎鼎，等. GLCrowd: 基于全局-局部注意力的弱监督密集场景人群计数模型[J]. 光电工程，2024， **51**(10): 240174.

Zhang H M, Tian Q Q, Yan D D, et al. GLCrowd: a weakly supervised global-local attention model for congested crowd counting[J]. *Opto-Electron Eng*, 2024, **51**(10): 240174.

https://doi.org/10.12086/oee.2024.240174

收稿日期: 2024-07-24; 修改日期: 2024-10-07; 录用日期: 2024-10-08

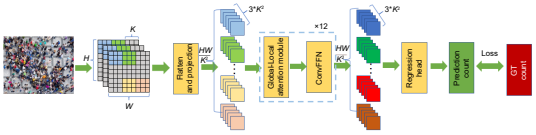
Opto-Electronic Engineering
光 电 工 程

Article

2024 年, 第 51 卷, 第 10 期

DOI: 10.12086/oee.2024.240174 CSTR: 32245.14.oee.2024.240174

GLCrowd：基于全局-局部注意力的弱监督密集场景人群计数模型



张红民*，田钱前，颜鼎鼎，卜令宇

重庆理工大学电气与电子工程学院，重庆 400054

摘要：针对人群计数在密集场景下存在背景复杂、尺度变化大等问题，提出了一种结合全局-局部注意力的弱监督密集场景人群计数模型——GLCrowd。首先，设计了一种结合深度卷积的局部注意力模块，通过上下文权重增强局部特征，同时结合特征权重共享获得高频局部信息。其次，利用 Vision Transformer (ViT) 的自注意力机制捕获低频全局信息。最后，将全局与局部注意力有效融合，并通过回归令牌来完成计数。在 Shanghai Tech PartA、Shanghai Tech PartB、UCF-QNRF 以及 UCF_CC_50 数据集上进行了模型测试，MAE 分别达到了 64.884、8.958、95.523、209.660，MSE 分别达到了 104.411、16.202、173.453、282.217。结果表明，提出的 GLCrowd 网络模型在密集场景下的人群计数中具有较好的性能。

关键词：人群计数；Vision Transformer；全局-局部注意力；弱监督学习

中图分类号：TP391 **文献标志码：**A

张红民，田钱前，颜鼎鼎，等. GLCrowd：基于全局-局部注意力的弱监督密集场景人群计数模型 [J]. 光电工程，2024，**51**(10): 240174

Zhang H M, Tian Q Q, Yan D D, et al. GLCrowd: a weakly supervised global-local attention model for congested crowd counting[J]. *Opto-Electron Eng*, 2024, **51**(10): 240174

GLCrowd: a weakly supervised global-local attention model for congested crowd counting

Zhang Hongmin*, Tian Qianqian, Yan Dingding, Bu Lingyu

School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, China

Abstract: To address the challenges of crowd counting in dense scenes, such as complex backgrounds and scale variations, we propose a weakly supervised crowd counting model for dense scenes, named GLCrowd, which integrates global and local attention mechanisms. First, we design a local attention module combined with deep convolution to enhance local features through context weights while leveraging feature weight sharing to capture high-frequency local information. Second, the Vision Transformer (ViT) self-attention mechanism is used to capture low-frequency global information. Finally, the global and local attention mechanisms are effectively fused, and

收稿日期：2024-07-24；修回日期：2024-10-07；录用日期：2024-10-08

基金项目：重庆市自然科学基金面上项目 (cstc2021 jcyj-msxmX0525，CSTB2022NSCQ-MSX0786，CSTB2023NSCQ-MSX0911)；重庆市教委科学技术研究项目 (KJQN202201109)

*通信作者：张红民，hmzhang@cqut.edu.cn。

counting is accomplished through a regression token. The model was tested on the Shanghai Tech Part A, Shanghai Tech Part B, UCF-QNRF, and UCF_CC_50 datasets, achieving MAE values of 64.884, 8.958, 95.523, and 209.660, and MSE values of 104.411, 16.202, 173.453, and 282.217, respectively. The results demonstrate that the proposed GLCrowd model exhibits strong performance in crowd counting within dense scenes.

Keywords: crowd counting; Vision Transformer; global-local attention; weakly supervised learning

1 引　言

近年来，人群计数技术[1]在公共安全、城市规划、广告营销、旅游管理等多个方面得到了广泛的应用。典型的人群计数模型[2-5]主要以卷积神经网络(convolution neural network，CNN)为主干，利用回归密度图来预测总人数[6]。然而，因 CNN 存在全局感知能力有限、特征提取能力不足及尺度变化敏感等问题，计数性能受到较多限制。为缓解此问题，国内外学者设计了多尺度机制，比如金字塔网络[7]、多列网络等，通过引入直观的局部结构归纳偏差[8]，使得模型能够适应不同大小的目标，提升特征提取的有效性和计数精度。

虽然 Transformer[9]模型在自然语言处理和计算机视觉等领域中表现展现出显著的性能优势。但是，直到 ViT[10]通过引入图像分块机制作为局部结构的归纳偏置， Transformer 在视觉任务中才达到与 CNN 相匹敌的性能水平。到目前为止，基于 ViT 的人群计数研究尚处于起步阶段[11-13]。ViT 强调全局信息的处理，在捕捉局部细节上尚不如传统的 CNN，因此在密集场景下，其计数能力仍受到限制。研究者尝试引入多尺度金字塔结构[14]或结合 CNN[15]来增强 ViT 的局部特征提取能力，然而多尺度金字塔结构在复杂背景和遮挡较多的密集场景人群计数中仍存在细节信息丢失的情形；而与 CNN 结合的方法也还未能充分发挥 Transformer 在全局信息处理上的优势。

基于此，本文提出了一种结合全局-局部注意力的弱监督人群计数方法。首先，在全局信息处理方面，通过 ViT 的自注意力 (self-attention) 机制有效捕获低频全局信息。其次，为了精细处理局部信息，提出了一种局部增强的自注意力机制，从而充分利用共享权重和上下文感知权重对局部感知的优势。此外，采用比普通局部自注意力更强的非线性方法来生成上下文感知权重，有效地捕获高频局部信息。最后，通过整合全局和局部注意力机制，并利用回归令牌 (regression token) 来获取精确的密集场景人群计数信息。

2 相关研究

2.1 弱监督人群计数

现有的人群计数方法被分为全监督学习[16-18]、弱监督学习[19-21]和无监督学习[22-24]。全监督学习常用于已经标注了每个图像或视频帧中人群数量的情况。通过配对的输入特征和精确的人群数量标签来进行预测和计数。然而其需要大量的标注数据和专业的标注技巧，费时费力。无监督学习大幅度减少了对大量精确标注数据的需求，并且在未知的场景中拥有更好的泛化能力，但也面临着算法复杂、性能不稳定、精确度受限等问题。因此，研究者开始聚焦弱监督人群计数方法研究，以兼顾全监督的准确性与无监督的数据利用效率。

其中，CrowdGraph[25]通过动态图卷积主干、多尺度膨胀图卷积模块和回归头完成从图形到计数。Yang 等[26]基于所提出的软标签排序网络，直接将图像映射到没有点级标注的人群上。Liu 等[27]利用每个图像中个体之间的相似性直接监督未标记区域。CLFormer[28]利用 Transformer 提取全局信息，并交替训练回归和定位分支，以生成高质量伪点级注释。OT-M[20]通过最优运输最小化算法生成硬伪标签，从而提升人群定位和半监督计数的性能，同时采用置信度加权策略增强监督效果。

2.2 Transformer

因 Transformer 比 CNN 更擅长捕捉全局特征和长距离，而被用于提高人群计数性能。例如，Transcrowd[29]重新定义了弱监督的人群计数问题，采用 Transformer 从序列到计数的角度进行解读。BCCTrans[30]引入了一个全局上下文可学习令牌来指导计数任务。CCTrans[31]结合了金字塔 Transformer 和多尺度回归头来实现全监督和弱监督人群计数。SFSL[32]引入人的可学习无偏特征估计，利用特征相似度对人群计数进行回归来缓解局部监督不足。DCSwinTrans[33]通过采用扩张型 Swin Transformer 骨

干增强范围上下文信息，并配备特征金字塔网络解码器来实现人群即时定位。MAN[11] 通过增强区域关注来提高 Transformer 的全局关注以缓解人群图像中尺度变换问题。JCTNet[15] 在卷积神经网络的高层特征上引入 Transformer 结构，并针对计数任务进行回归。HANet[34] 在平行空间注意和通道注意中引入尺度变换。CrowdFormer[35] 通过重叠模拟人类自下而上的视觉感知机制，考虑了不同的目标比例。CrowdMLP[5] 提出了一个多粒度多层 (MLP) 回归器来扩大接受域，并提出了一个分裂计数来解耦空间约束。Gramformer[13] 通过图形调制 Transformer 学习人群计数。

尽管人群计数方法已取得了进展，但针对密集场景的人群计数，现有方法仍难以平衡全局上下文感知与局部细节捕捉的关系。因此，本文提出了一种结合全局与局部注意力的弱监督方法，引入了局部增强注意力机制，并与 ViT 的全局注意力相结合，有效提升了密集场景中的计数精度。

3 GLCrowd 网络模型

提出的全局-局部注意力相结合的弱监督密集人群计数模型 (weakly supervised global-local attention model for congested crowd counting, GLCrowd) 的整体框架如图 1 所示。该框架首先对输入的图像进行预处理，将其分成 16×16 的小块并平坦化为向量序列。接着，通过设计的全局-局部注意力机制有效提取关键信息，该机制结合了对整体场景的广泛理解与对关键细节的精确捕捉。然后，通过回归令牌聚合特征，最终利用回归头生成人群数量预测。

3.1 数据准备

为确保数据质量的一致性，采取一种标准化的图像预处理流程。具体步骤如下。

首先，图像尺寸的调整基于其宽度和高度的比较。图像的宽度 H 大于或等于其高度 W，则调整图像的宽度至 1152 pixels，同时按照原始宽高比例调整高度，确保图像比例不变。反之，高度则调整至 1152 pixels，宽度按比例缩放。此步骤确保在不失去重要视觉信息的前提下，调整图像至一个标准化的分辨率。

其次，为了提升在不同局部图像特征上的适应性和鲁棒性，调整后的图像及其对应的密度图被进一步裁剪成多个 384 pixel×384 pixel 的块，供后续的模型训练和测试使用。

3.2 图像块嵌入

将输入 RGB 图像 $I \in \mathbb{R}^{3\times W\times H}$ 划分为 $\frac{H}{K}\times\frac{W}{K}$ 个补丁的网格，每一个补丁的尺寸为 $K\times K\times 3$，将每个补丁展平且合并为一个序列 $x=\left\{x_i\in\mathbb{R}^{K^2\times 3}|i=1,\cdots,N\right\}$，其中 (H,W) 为输入 RGB 图像的分辨率，N 为序列长度 $N=HW/K^2$。然后，通过一个可学习线性投影将 x 映射到一个潜在的 D 维嵌入特征中，即 $f:x\to e=\{e_i\in\mathbb{R}^D|i=1,\cdots,N\}$。最后，为编码补丁的空间位置信息，每个补丁添加了一个标准可学习的 $1D$ 位置嵌入 $\{P_i\in\mathbb{R}^D|i=1,\cdots,N\}$：

$$Z_0=[e_1+p_1;\ e_2+p_2;\ \cdots;\ e_N+p_N]\ , \tag{1}$$

其中：Z_0 表示第一层 Transformer 的输入。

3.3 全局-局部注意力模块

全局-局部注意力模块的整体框图如图 2 所示。模块被设计为包含两个互补的分支：全局分支和局部分支。全局分支主要负责捕获图像中的低频全局信息，用于理解图像的整体上下文和背景布局。局部分支专注于处理图像中的高频局部信息，便于识别图像中的具体对象和细节特征。通过两者结合，GLCrowd 网络模型能够有效地捕获高频和低频信息，从而实现对复杂场景的全面理解和精准分析。

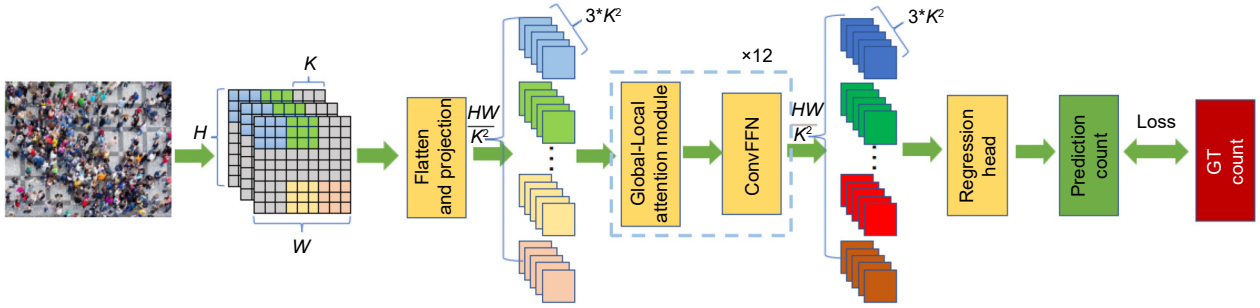


图 1 GLCrowd 网络模型结构

Fig. 1 GLCrowd network structure

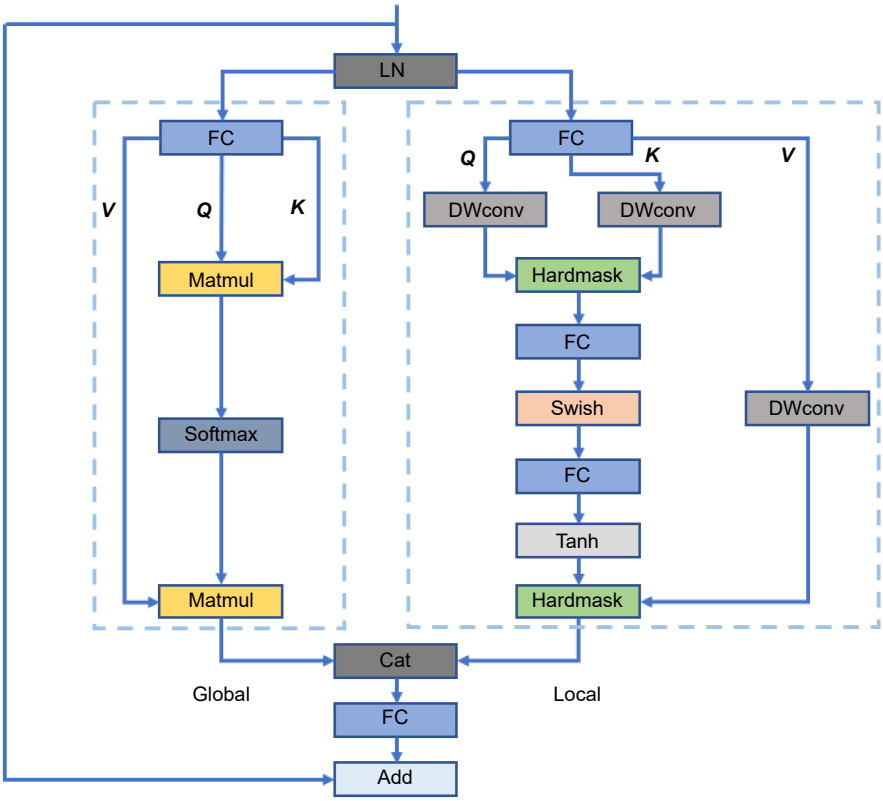


图 2 全局-局部注意力模块

Fig. 2 Global-local attention module

3.3.1 全局注意力

全局注意力主要利用 ViT 的 L 层多头自注意力 (MSA) 和多层感知器 (MLP) 获得全局注意力。对于每一层 l 处理前采用归一化 (layer normalization，LN) 以稳定训练过程并加速收敛，处理后通过残差连接，其输出结果为

$$Z'_l = MSA(LN(Z_{l-1})) + Z_{l-1}\,, \tag{2}$$

$$Z_l = MLP(LN(Z'_l)) + Z'_l\,, \tag{3}$$

其中：Z_l表示第 l 层的输入。MLP包含两个具有GELU激活函数的线性层，第一个线性层将特征嵌入维度从 D 扩展为 $4D$，增强模型的表示能力；第二个线性层将嵌入维度从 $4D$ 压缩回 D，完成一次特征转换。MSA拥有 m 个独立自注意力 (SA) 模型的扩展，在每个独立的SA中，输入由查询 (query,$\boldsymbol{Q}$)、键 (key,$\boldsymbol{K}$) 和值 (value,$\boldsymbol{V}$) 组成。

$$\boldsymbol{Q} = \boldsymbol{Z}_{l-1}\boldsymbol{W}_Q,\ \boldsymbol{K} = \boldsymbol{Z}_{l-1}\boldsymbol{W}_K,\ \boldsymbol{V} = \boldsymbol{Z}_{l-1}\boldsymbol{W}_V\,, \tag{4}$$

$$\boldsymbol{X}_{\text{global}} = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{D}}\right)\boldsymbol{V}\,, \tag{5}$$

其中：$\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{D\times\frac{D}{m}}$ 为三个可学习矩阵。$\sqrt{D}$提供适当的归一化。

3.3.2 局部注意力

提出的 GLCrowd 网络模型中的局部注意力与传统卷积以及标准自注意力的对比如图 3 所示，三种操作均使用残差连接。传统卷积操作通过固定大小的卷积核在局部区域内提取特征。标准自注意力机制通过上下文权重计算相邻令牌的相似度分数，以提取高频局部表示。

局部注意力首先通过线性变换得到 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，这与标准注意力相似：

$$\boldsymbol{Q},\ \boldsymbol{K},\ \boldsymbol{V} = FC(\boldsymbol{X}_{\text{in}})\,, \tag{6}$$

其中：$\boldsymbol{X}_{\text{in}}$表示输入，$FC$表示全连接层。在线性变换以后，对于 $\boldsymbol{V}$，使用具有共享权值的深度卷积 (depthwise convolution, DWconv) 操作以进行局部特征聚合处理，公式如下：

$$\boldsymbol{V}_{\text{s}} = DWconv(\boldsymbol{V})\,. \tag{7}$$

利用 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 结合起来生成上下文感知权值。不同于标准自注意力的方法，首先使用两个 DWconv 分别聚合 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的本地信息，然后 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 做哈达玛积，通过一系列变换生成注意力图，最后与 $\boldsymbol{V}$ 相乘得到注意力分数。整个过程总结如下：

$$\begin{aligned}\boldsymbol{Q}_l &= DWconv(\boldsymbol{Q}),\\ \boldsymbol{K}_l &= DWconv(\boldsymbol{K}),\\ Attn_t &= FC(Swish(FC(\boldsymbol{O}_l \Theta \boldsymbol{K}_l))),\\ \boldsymbol{X}_{\text{local}} &= \text{Tanh}\left(\frac{Attn_t}{\sqrt{d}}\right)\Theta \boldsymbol{V}_{\text{S}},\end{aligned} \tag{8}$$

其中：d 表示令牌通道数，Θ表示哈达玛积，用Tanh函数代替$softmax$算子，使模型能利用更强的非线性处理以获得更高质量的上下文权重。

局部分支与传统卷积相比，上下文感知权重的利用使模型在局部感知过程中能够更好地适应输入内容；与标准自注意力相比，共享权值的引入使 GLCrowd 模型能够更好地处理高频信息，从而提高性能。

3.3.3 注意力融合

GLCrowd 模型采用一种简洁的策略来融合局部分支和全局分支，即在通道维度上将两者的输出进行拼接，形成一个包含所有特征的张量，使得 GLCrowd 网络可以同时学习由不同注意力模块捕获的不同特征或信息，再通过一个全连接层输出：

$$\boldsymbol{X}_{\text{out}} = FC(Concat(\boldsymbol{X}_{\text{local}}, \boldsymbol{X}_{\text{global}})), \tag{9}$$

其中：$Concat$ 表示在通道维度上的拼接操作，X_{out}表示局部分支和全局分支融合的输出结果。

然后，将该输出 Drop Path 正则化与原始输入相加，形成第一层的残差连接。对该结果应用层标准化和 ConvFFN 模块，再次通过 Drop Path 正则化后，与前一阶段的输出相加，形成第二层的残差连接。这种设计不仅增强了网络捕捉复杂特征的能力，而且还利用残差连接和 Drop Path 提高了训练的稳定性和模型的泛化能力，使得深层网络能够有效地训练，从而提升了模型的性能和可靠性。

3.4 卷积前馈网络 (ConvFFN)

为了将局部信息整合到全连接前馈网络 (FFN)，在传统全连接层的基础上引入 DWconv，其结构如图 4 所示，这种设计允许 ConvFFN 在保持对全局信息敏感的同时，增强对图像局部特征的处理能力。此外，由于深度卷积的结构特性，ConvFFN 能够直接在网络内部进行下采样，从而无需额外的模块来减少特征维度。ConvFFN 通过一个 1×1 卷积层初步变换输入特征的通道数，接着用 GELU 激活函数添加非线性，然后通过组内独立的深度卷积层 (DWconv) 深入提取局部空间特征。此后，重新通过 1×1 卷积调整通道数到输出规模，并在两个关键位置应用 Dropout 减少过拟合以增强模型的泛化能力。具体公式为

$$\begin{aligned}X_1 &= GELU(W_{\text{fc1}} * X)\\ X_2 &= Dropout(Dwconv(X_1))\\ X_3 &= Dropout(W_{\text{fc2}} * X_2),\end{aligned} \tag{10}$$

其中：X_3表示残差连接前的输入，W_{fc1}、W_{fc2}表示第一个和第二个1×1卷积层的权重。

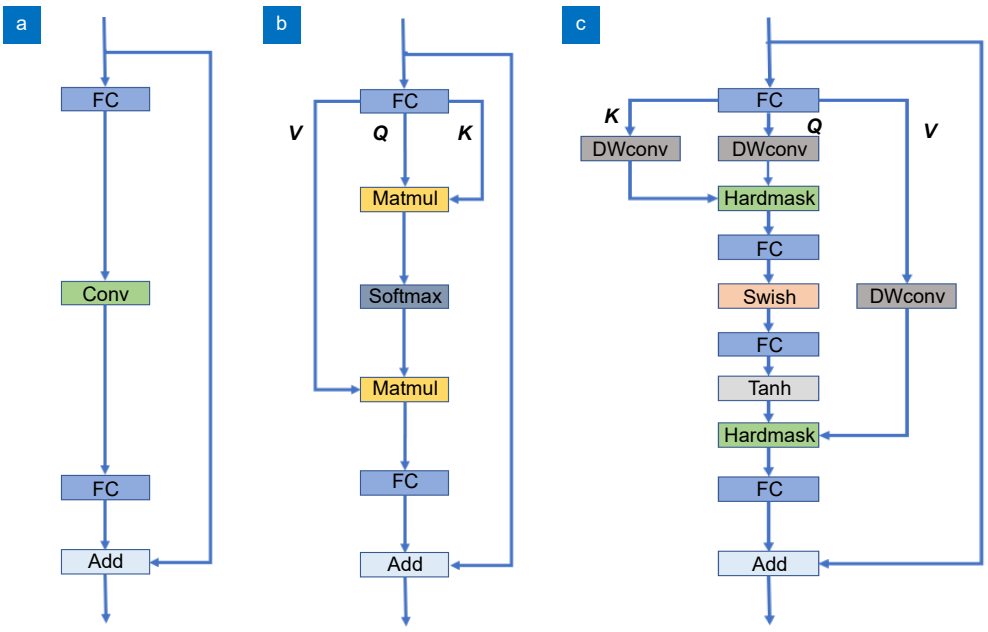


图 3 不同方法的比较。(a) 传统卷积; (b) 标准自注意力; (c) 局部注意力

Fig. 3 Comparison of different methods. (a) Traditional convolution; (b) Standard self-attention; (c) Local attention

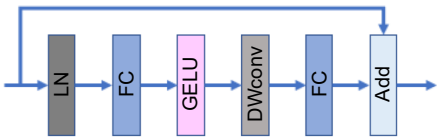


图 4 卷积前馈网络模块

Fig. 4 ConvFFN moudle

3.5 回归令牌

在输入序列Z_0前加入一个带位置嵌入的回归令牌 cls_token，这种结构负责回归令牌和补丁令牌之间的信息传播，使回归令牌包含整体语义信息。再通过由 ReLU 激活层、Dropout 层和一个线性层组成的回归头以完成人群计数任务，并利用L_1作为损失函数来度量预测人数和真实人数之间的差值。

$$L_1 = \frac{1}{M}\sum_{i=1}^{M}|P_i - T_i|\,, \tag{11}$$

其中：M表示训练图像的批处理大小，P_i和T_i分别表示第 i 个图像的预测人数和真实人数。

4 实 验

4.1 评价指标

本文使用的评价指标为平均绝对误差 (mean absolute error, MAE) 和均方误差 (mean squared error, MSE)。MAE 和 MSE 的定义见式 (12) 和式 (13)。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_i - T_i|\,, \tag{12}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - T_i)^2}, \tag{13}$$

其中：MAE表示预测人数与真实人数之间的平均误差，提供了对模型整体误差水平的衡量，MSE表示预测人数与真实人数之间的均方误差，提供了对模型误差变异度的衡量。N表示测试集中图像总数，P_i和T_i分别表示第 i 个图像的预测人数和真实人数。

4.2 实验配置

为验证所提出的模型效果，本文在 Shanghai Tech[36]、UCF-QNRF[37]、UCF_CC_50[38] 三个数据集进行模型实验。数据集的详细信息如表 1 所示。

1) Shanghai Tech 数据集：共有 1198 张图像，330165 个人头标注，分为 Part A 和 Part B 两部分。其中 Part A 有 482 张图像，包含了多样化的背景和环境。Part B 包含 716 张图像，人群密度相对稀疏，但仍涵盖了从稀疏到中等密度的人群分布，且摄像头视角的透视效果导致远近人物的比例存在显著变化。

2) UCF-QNRF 数据集：共有 1535 张高分辨率图像，总标注超过一百万个人头。拥有高分辨率图片、广泛的场景覆盖、极度拥挤的人群、精确的人头标注等特点。

3) UCF_CC_50 数据集：共有 50 张高质量图像，63974 个人头标注，是目前平均密度最高的人群数据集之一。由于该数据集的样本量较少，且图像间的人数差异极大，因此即使是主流方法也难以在该数据集上取得理想成绩[39]。

表 1 数据集信息

Table 1 Information of datasets

	Shanghai Tech		UCF-QNRF	UCF_CC_50
	Part A	Part B		
图像数量	482	716	1535	50
平均尺寸	589×868	768×1024	2013×2902	2101×2888
平均人数	501	123.6	1279	1278
最大人数	3139	578	12865	4543
总人数	241677	88488	1251642	63974

表 2 实验环境信息

Table 2 Information of experimental environment

配置	参数
操作系统	Ubuntu 20.04.3 LTS (GNU/Linux 5.15.0-97-generic x86_64)
显卡型号	NVIDIA GeForce RTX 4090(×1)
显存大小	24 G

表 3 部分实验参数数据

Table 3 Data of partial experimental parameters

实验参数	具体参数
权重衰减	5×10^{-4}
训练总周期数	1200
优化器动量	0.95
学习率	1×10^{-5}

表 2 列举了实验环境的各项参数；表 3 列举了部分实验参数。

4.3 方法对比实验结果

为了验证提出的 GLCrowd 网络模型的有效性，本文将其与当前主流的弱监督人群计数模型进行了比较分析。表 4、表 5 及表 6 分别展示了数据集 Shanghai Tech、UCF-QNRF 以及 UCF_CC_50 的实验结果，其中，最优结果用加粗倾斜表示，次优结果用下划线表示。在 Shanghai Tech 数据集的 Part A 部分，相比于次优方法 Transcrowd_gap 模型，MAE 提升了 1.2，MSE 提升了 0.7。虽然 OT_M 模型在数据集 Shanghai Tech Part B 显示出略微优势，但究其原因是 OT_M 模型在利用弱监督方法进行人群计数时，采用点标注提供准确的人头信息，其最优结果是在标注比例占总图片 40% 的情况下获得的，而提出的 GLCrowd 模型采用级标注，仅提供了图片总人数。尽管如此，GLCrowd 模型在 Shanghai Tech Part B 数据集上的表现与 OT_M 模型几乎持平，显示出强劲的竞争力，位居第二。

同时，在 UCF_QNRF 数据集的实验中，GLCrowd 模型比 Transcrowd_gap 模型在 MAE 上提升了 1.7。在 UCF_CC_50 数据集中，GLCrowd 模型与 PC-Net 模型相比在 MAE 上提升了 7.6，并且在 MSE 上大幅度提升了 27.5。

此外，对 GLCrowd 模型开展了可视化实验，部分可视化结果如图 5 所示。所选的五幅原始图像 (original picture) 包含稀疏人群场景和密集人群场景，第一行图片表示稀疏人群图像；第二行图片提供了基础的密集人群图像；第三行图片在密集人群的基础上增加了更复杂的背景信息，这些复杂背景包括不规则的建筑场景以及多样的颜色变化；第四行图片引入了由摄像机镜头远近变化引起的人物尺寸大小不一的问题。第五行图像则集合了密集人群、复杂背景以及尺度变化三个问题。通过对所选的五幅原始图像的 Attention weight 和 Attention map 分析，提出的 GLCrowd 模型能够成功聚焦地在人群所在区域，并

表 4 Shanghai Tech 数据集实验结果

Table 4 Experimental results on Shanghai Tech dataset

算法模型	Part A		Part B	
	MAE	MSE	MAE	MSE
HLNet[40]	71.5	108.6	11.3	20
Transcrowd_gap[29]	<u>66.1</u>	<u>105.1</u>	9.3	<u>16.1</u>
Transcrowd_token[29]	69.0	116.5	10.6	19.7
MATT[41]	80.1	129.4	11.7	17.5
OT_M[20]	70.7	114.5	**8.1**	**13.1**
SUA_crowd[42]	68.5	121.9	14.1	20.6
PDDNet[43]	72.6	112.2	10.3	17.0
GLCrowd	**64.887**	**104.411**	<u>8.958</u>	16.202

表 5 UCF_QNRF 数据集实验结果

Table 5 Experimental results on UCF_QNRF dataset

算法模型	MAE	MSE
HLNet[40]	100.4	182.6
Transcrowd_token[29]	98.9	176.1
Transcrowd_gap[29]	<u>97.2</u>	<u>168.5</u>
OT_M[20]	100.6	**167.6**
SUA_crowd[42]	130.3	226.3
PDDNet[43]	130.2	246.6
GLCrowd	**95.523**	173.453

表 6 UCF_CC_50 数据集实验结果

Table 6 Experimental results on UCF_CC_50 dataset

算法模型	MAE	MSE
MATT[41]	355.0	550.0
CCTrans[31]	245.0	343.0
SSGP_Crowd[44]	355.0	505.0
SE Cycle GAN[45]	373.4	528.8
CDPL_crowd[46]	336.5	486.1
Transcrowd[29]	272.2	395.3
CrowdFormer[47]	218.8	330.4
PC-Net[48]	<u>217.3</u>	<u>309.7</u>
GLCrowd	**209.660**	**282.217**

且人群越密集时分配更多的 Attention weight，从而使预测人数 (Pre) 与真实人数 (GT) 相接近。其中，在第二行和第四行图片的无复杂背景的密集场景中，GLCrowd 模型因关注局部区域导致预测人数略高于真实人数；而在第三行和第五行图片的复杂背景场景中，复杂环境信息干扰注意力分配，导致少量人物未被准确识别，造成预测人数略低于真实人数。尽管存在细微差异，GLCrowd 模型仍有效缓解了在人群计数中因背景复杂和尺度变化带来的挑战。

对三个数据集的实验结果进行分析可知，针对密集场景中的人群计数性能较好，尤其在 UCF_CC_50 和 Shanghai Tech Part A 数据集上表现出色，同时，在稀疏或密度跨度较大的数据集 (如 Shanghai Tech Part B 和 UCF_QNRF) 中同样取得了与其他主流模型相媲美的效果。这表明 GLCrowd 模型不仅在处理密集场景方面具有优势，还具备一定的泛化能力，能够应对不同密度条件下的人群计数任务。

4.4 消融实验

为研究局部注意力、ConvFFN 以及回归令牌的有效性，在数据集 Shanghai Tech 和 UCF_CC_50 中

图 5 部分可视化结果

Fig. 5 Partial visualization results

开展了消融实验。第一组实验专注于测试回归令牌在全局注意力机制下的作用，旨在证明回归令牌在整体特征聚合和精确度提升方面的贡献。第二组实验在第一组实验的基础上加入局部注意力机制，旨在展示局部注意力在增强模型对细节特征捕捉能力的有效性，尤其是在人群密集区域的表现。第三组实验在第二组实验的基础上进一步引入 ConvFFN 模块，检验了 ConvFFN 在增强模型处理复杂场景时的能力，特别是在提供上下文信息和增强模型整体理解能力方面的贡献。实验结果如表 7 所示，表中数据详细展示了各组件对改善模型在人群计数精度上的具体影响。以第一组为参照，首先在 UCF_CC_50 数据集中添加局部注意力，MAE 和 MSE 分别下降 1.62% 和 6.92%；再引入 ConvFFN，MAE 和 MSE 分别进一步下降 11.25% 和 7.51%，总体下降 12.69% 和 13.90%。数据集 Shanghai Tech 呈现相同的下降趋势，证明了各个模块组合能较好地提升模型的整体性能。

5 总 结

为了提高在密集人群场景中的计数准确性，同时降低对大量标注数据的依赖，本文提出了基于全局-局部注意力的弱监督人群计数模型 GLCrowd。该模型的全局注意力能够有效理解场景的整体布局和背景，为整体人群分布提供上下文信息；局部注意力则有助于模型理解密集区域的精细特征，增强对局部人群动态的理解和计数准确性。实验结果表明，提出的 GLCrowd 网络模型在 Shanghai Tech、UCF-QNRF 以及 UCF_CC_50 三个人群计数数据集上取得了较好的人群计数结果。

参考文献

[1] Tian Y Y, Deng M L, Gao H, et al. Review of crowd counting algorithms based on deep learning[J]. *Electron Meas Technol*, 2022, **45**(7): 152−159.
田月媛, 邓淼磊, 高辉, 等. 基于深度学习的人群计数算法综述[J]. 电子测量技术, 2022, **45**(7): 152−159.

[2] 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]. *Int J Comput Vis*, 2023, **131**(7): 1722−1740.

[3] Wu S K, Yang F Y. Boosting detection in crowd analysis via underutilized output features[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 15609–15618.
https://doi.org/10.1109/CVPR52729.2023.01498.

[4] Yu Y, Cai Z, Miao D Q, et al. An interactive network based on transformer for multimodal crowd counting[J]. *Appl Intell*, 2023, **53**(19): 22602−22614.

[5] Wang M J, Zhou J, Cai H, et al. CrowdMLP: weakly-supervised crowd counting via multi-granularity MLP[J]. *Pattern Recognit*, 2023, **144**: 109830.

[6] Lu Z K, Liu S, Zhong L, et al. Survey on reaserch of crowd counting[J]. *Comput Eng Appl*, 2022, **58**(11): 33−46.
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. d.

[7] Guo A X, Xia Y F, Wang D W, et al. A multi-scale crowd counting algorithm with removing background interference[J]. *Comput Eng*, 2022, **48**(5): 251−257.
郭爱心, 夏殷锋, 王大为, 等. 一种抗背景干扰的多尺度人群计数算法[J]. 计算机工程, 2022, **48**(5): 251−257.

[8] Zhang Q M, Xu Y F, Zhang J, et al. ViTAEv2: vision transformer advanced by exploring inductive bias for image recognition and beyond[J]. *Int J Comput Vis*, 2023, **131**(5): 1141−1162.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6000–6010.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//*Proceedings of the 9th International Conference on Learning Representations*, 2021.

[11] Lin H, Ma Z H, Ji R R, et al. Boosting crowd counting via multifaceted attention[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 19628–19637.
https://doi.org/10.1109/CVPR52688.2022.01901.

[12] Liang D K, Xu W, Bai X. An end-to-end transformer model for crowd localization[C]//*Proceedings of the 17th European Conference on Computer Vision*, 2022: 38–54.
https://doi.org/10.1007/978-3-031-19769-7_3.

[13] Lin H, Ma Z H, Hong X P, et al. Gramformer: learning crowd counting via graph-modulated transformer[C]//*Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024.
https://doi.org/10.1609/aaai.v38i4.28126.

[14] Li B, Zhang Y, Xu H H, et al. CCST: crowd counting with swin transformer[J]. *Vis Comput*, 2023, **39**(7): 2671−2682.

[15] Wang F S, Liu K, Long F, et al. Joint CNN and Transformer

表 7 消融实验对比结果

Table 7 Comparative results of ablation experiments

	局部注意力	ConvFFN	回归令牌	Part A		Part B		UCF_CC_50	
				MAE	MSE	MAE	MSE	MAE	MSE
第一组	×	×	√	75.260	125.441	10.003	19.315	240.120	327.796
第二组	√	×	√	67.403	108.298	9.428	18.446	236.233	305.122
第三组	√	√	√	**64.887**	**104.411**	**8.955**	**16.202**	**209.660**	**282.217**

Network via weakly supervised Learning for efficient crowd counting[Z]. arXiv: 2203.06388, 2022. https://arxiv.org/abs/2203.06388.

[16] Wang C A, Song Q Y, Zhang B S, et al. Uniformity in heterogeneity: diving deep into count interval partition for crowd counting[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 3234–3242. https://doi.org/10.1109/ICCV48922.2021.00322.

[17] Song Q Y, Wang C A, Jiang Z K, et al. Rethinking counting and localization in crowds: a purely point-based framework[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 3365–3374. https://doi.org/10.1109/ICCV48922.2021.00335.

[18] Shivapuja S V, Khamkar M P, Bajaj D, et al. Wisdom of (binned) crowds: a bayesian stratification paradigm for crowd counting[C]//*Proceedings of the 29th ACM International Conference on Multimedia*, 2021: 3574–3582. https://doi.org/10.1145/3474085.3475522.

[19] Wang X Y, Zhang B F, Yu L M, et al. Hunting sparsity: density-guided contrastive learning for semi-supervised semantic segmentation[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 3114–3123. https://doi.org/10.1109/CVPR52729.2023.00304.

[20] Lin W, Chan A B. Optimal transport minimization: crowd localization on density maps for semi-supervised counting[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 21663–21673. https://doi.org/10.1109/CVPR52729.2023.02075.

[21] Gao H, Zhao W J, Zhang D X, et al. Application of improved transformer based on weakly supervised in crowd localization and crowd counting[J]. *Sci Rep*, 2023, **13**(1): 1144.

[22] Liu Y T, Wang Z, Shi M J, et al. Discovering regression-detection bi-knowledge transfer for unsupervised cross-domain crowd counting[J]. *Neurocomputing*, 2022, **494**: 418−431.

[23] Xu C F, Liang D K, Xu Y C, et al. AutoScale: learning to scale for crowd counting[J]. *Int J Comput Vis*, 2022, **130**(2): 405−434.

[24] Liang D K, Xie J H, Zou Z K, et al. CrowdCLIP: unsupervised crowd counting via vision-language model[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023: 2893–2903. https://doi.org/10.1109/CVPR52729.2023.00283.

[25] Zhang C Y, Zhang Y, Li B, et al. CrowdGraph: weakly supervised crowd counting via pure graph neural network[J]. *ACM Trans Multimedia Comput, Commun Appl*, 2024, **20**(5): 135.

[26] Yang Y F, Li G R, Wu Z, et al. Weakly-supervised crowd counting learns from sorting rather than locations[C]//*Proceedings of the Computer Vision–ECCV 2020: 16th European Conference*, 2020: 1–17. https://doi.org/10.1007/978-3-030-58598-3_1.

[27] Liu Y T, Ren S C, Chai L Y, et al. Reducing spatial labeling redundancy for active semi-supervised crowd counting[J]. *IEEE Trans Pattern Anal Mach Intell*, 2022, **45**(7): 9248−9255.

[28] Deng M F, Zhao H L, Gao M. CLFormer: a unified transformer-based framework for weakly supervised crowd counting and localization[J]. *Vis Comput*, 2024, **40**(2): 1053−1067.

[29] Liang D K, Chen X W, Xu W, et al. TransCrowd: weakly-supervised crowd counting with transformers[J]. *Sci China Inf Sci*, 2022, **65**(6): 160104.

[30] Sun G L, Liu Y, Probst T, et al. Rethinking global context in crowd counting[Z]. arXiv: 2105.10926, 2021. https://arxiv.org/abs/2105.10926.

[31] Tian Y, Chu X X, Wang H P. CCTrans: simplifying and improving crowd counting with transformer[Z]. arXiv: 2109.14483, 2021. https://arxiv.org/abs/2109.14483.

[32] Chen X S, Lu H T. Reinforcing local feature representation for weakly-supervised dense crowd counting[Z]. arXiv: 2202.10681, 2022. https://arxiv.org/abs/2202.10681v1.

[33] Gao J Y, Gong M G, Li X L. Congested crowd instance localization with dilated convolutional swin transformer[J]. *Neurocomputing*, 2022, **513**: 94−103.

[34] Wang F S, Sang J, Wu Z Y, et al. Hybrid attention network based on progressive embedding scale-context for crowd counting[J]. *Inf Sci*, 2022, **591**: 306−318.

[35] Yang S P, Guo W Y, Ren Y H. CrowdFormer: an overlap patching vision transformer for top-down crowd counting[C]//*Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022: 23–29. https://doi.org/10.24963/ijcai.2022/215.

[36] Zhang Y Y, Zhou D S, Chen S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 589–597. https://doi.org/10.1109/CVPR.2016.70.

[37] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 532–546. https://doi.org/10.1007/978-3-030-01216-8_33.

[38] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//*Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013: 2547–2554. https://doi.org/10.1109/CVPR.2013.329.

[39] Patwal A, Diwakar M, Tripathi V, et al. Crowd counting analysis using deep learning: a critical review[J]. *Proc Comput Sci*, 2023, **218**: 2448−2458.

[40] Chen Y Q, Zhao H L, Gao M, et al. A weakly supervised hybrid lightweight network for efficient crowd counting[J]. *Electronics*, 2024, **13**(4): 723.

[41] Lei Y J, Liu Y, Zhang P P, et al. Towards using count-level weak supervision for crowd counting[J]. *Pattern Recognit*, 2021, **109**: 107616.

[42] Meng Y D, Zhang H R, Zhao Y T, et al. Spatial uncertainty-aware semi-supervised crowd counting[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 15549–15559. https://doi.org/10.1109/ICCV48922.2021.01526.

[43] Liang L J, Zhao H L, Zhou F B, et al. PDDNet: lightweight congested crowd counting via pyramid depth-wise dilated convolution[J]. *Appl Intell*, 2023, **53**(9): 10472−10484.

[44] Sindagi V A, Yasarla R, Babu D S, et al. Learning to count in the crowd from limited labeled data[C]//*Proceedings of the Computer Vision–ECCV 2020: 16th European Conference*, 2020: 212–229. https://doi.org/10.1007/978-3-030-58621-8_13.

[45] Wang Q, Gao J Y, Lin W, et al. Learning from synthetic data for crowd counting in the wild[C]//*Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 8198–8207. https://doi.org/10.1109/CVPR.2019.00839.

[46] Liu W Z, Durasov N, Fua P. Leveraging self-supervision for cross-domain crowd counting[C]//*Proceedings of 2022*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5341–5352. https://doi.org/10.1109/CVPR52688.2022.00527.

[47] Savner S S, Kanhangad V. CrowdFormer: weakly-supervised crowd counting with improved generalizability[J]. *J Vis Commun Image Representation*, 2023, **94**: 103853.

[48] Li Y C, Jia R S, Hu Y X, et al. A weakly-supervised crowd density estimation method based on two-stage linear feature calibration[J]. *IEEE/CAA J Autom Sin*, 2024, **11**(4): 965−981.

作者简介

【通信作者】张红民 (1970-)，男，博士，教授，主要研究方向为图像处理与模式识别。

E-mail：hmzhang@cqut.edu.cn

田钱前 (1999-)，女，硕士研究生，主要研究方向为图像处理、深度学习。

E-mail：qianqiantian@stu.cqut.edu.cn

颜鼎鼎 (2000-)，女，硕士研究生，主要研究方向为图像处理、计算机视觉。

E-mail：ydd0010@stu.cqut.edu.cn

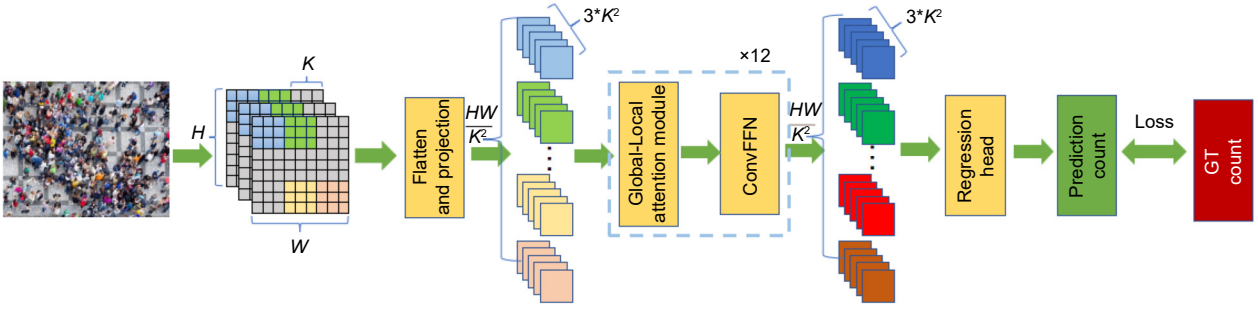
卜令宇 (1999-)，男，硕士研究生，主要研究方向为图像处理。

E-mail：bulingyuuuu@163.com

扫描二维码，获取PDF全文

GLCrowd: a weakly supervised global-local attention model for congested crowd counting

Zhang Hongmin*, Tian Qianqian, Yan Dingding, Bu Lingyu



GLCrowd network structure

Overview: Crowd counting aims to estimate the number of people in an image using computer algorithms and has wide applications in public safety, urban planning, advertising, and tourism management. However, counting in dense crowds still faces significant challenges due to complex backgrounds and scale variations. Vision Transformer (ViT) offers superior information processing capabilities compared to convolutional neural networks (CNNs), but it falls short in detail capture compared to traditional CNNs, limiting its performance in dense scenarios. To address this issue and reduce reliance on extensive annotated data, this paper proposes a weakly supervised crowd counting method based on a Transformer with a combination of global and local attention mechanisms. The approach aims to leverage the complementary strengths of both global and local attention to enhance crowd counting performance.

First, the Vision Transformer's self-attention captures global features, focusing on low-frequency global information to understand the overall context and background layout of the image. Next, local attention captures high-frequency local information, using depthwise convolution (DWconv) to aggregate local features from the values (V). Depthwise convolution is also applied to queries (Q) and keys (K), and attention maps are generated through the Hadamard product and Tanh function. This approach aims to achieve higher-quality contextual weights with stronger nonlinearity for identifying specific objects and detailed features in the image. Global and local attention outputs are then concatenated along the channel dimension. To integrate local information into the feed-forward network (FFN), the paper replaces the original FFN in ViT with a feed-forward network that includes DWconv. Finally, a regression head is used to complete the crowd counting task.

The proposed model was validated on the Shanghai Tech, UCF-QNRF, and UCF_CC_50 datasets and compared with current mainstream weakly supervised crowd counting models. On the Shanghai Tech Part A dataset, the proposed model improved MAE by 1.2 and MSE by 0.7 compared to the second-best model, Transcrowd_gap. Although the OT_M model showed slight advantages in Shanghai Tech Part B, this is mainly because OT_M uses point annotations. The proposed model uses coarse annotations that only provide the total number of people in the image. On the UCF_QNRF dataset, the proposed model improved MAE by 1.7 compared to Transcrowd_gap. On the UCF_CC_50 dataset, the proposed model achieved a 9.2 improvement in MAE and a significant 48.2 improvement in MSE compared to the CrowdFormer model. These results demonstrate that the proposed network model outperforms other mainstream models in weakly supervised crowd counting.

Zhang H M, Tian Q Q, Yan D D, et al. GLCrowd: a weakly supervised global-local attention model for congested crowd counting[J]. *Opto-Electron Eng*, 2024, **51**(10): 240174; DOI: 10.12086/oee.2024.240174

Foundation item: Chongqing Natural Science Foundation Top Project (cstc2021 jcyj-msxmX0525, CSTB2022NSCQ-MSX0786, CSTB2023NSCQ-MSX0911); Science and Technology Research Project of Chongqing Municipal Education Commission (KJQN202201109)

School of Electrical and Electronic Engineering, Chongqing University of Technology, Chongqing 400054, China

* E-mail: hmzhang@cqut.edu.cn