文章编号　1004-924X(2024)22-3395-14

密集连接注意力与尺度感知重组增强的人群计数

陈　永[1,2*]，董　珂[1]，安卓奥博[1]，周建宇[1]

(1. 兰州交通大学 电子与信息工程学院，甘肃 兰州 730070；
2. 甘肃省人工智能与图形图像处理工程研究中心，甘肃 兰州 730070)

摘要：针对人群计数中背景干扰、人群尺度变化剧烈，导致计数效果不佳的问题，提出了一种密集连接注意力与尺度感知重组增强的人群计数方法。首先，设计密集连接注意力机制的特征提取网络，通过使用膨胀卷积改进的VGG19网络作为模型粗特征提取网络，并嵌入密集连接双通道注意力机制，增强人群计数特征，抑制背景干扰。然后，设计尺度感知重组上采样和软掩膜特征增强及传递结构，实现从浅层到深层不同尺度人群特征信息的充分利用，克服人群尺度变化剧烈导致计数性能不佳的问题。其次，提出多分辨率融合模块，增强多分辨率信息间交互，降低不同分辨率之间的语义差距，提高人群计数的准确度。最后，在ShanghaiTech，UCF-QNRF，JHU_CROWD++等人群数据集上对比实验，结果表明所提方法均优于对比算法，相较于DM-Count人群计数算法，MAE、MSE误差分别下降了15.98%，14.52%，所提方法具有更高的计数性能。

关　键　词：人群计数；密集连接注意力；尺度感知重组；多尺度增强；多分辨率融合
中图分类号：TP391　　**文献标识码**：A　　**doi**：10.37188/OPE.20243222.3395

Crowd counting method based on dense connection attention and scale perception recombination enhancement

CHEN Yong[1,2*], DONG Ke[1], AN Zhuoaobo[1], ZHOU Jianyu[1]

(1. *School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou* 730070, *China*;
2. *Gansu Provincial Engineering Research Center for Artificial Intelligence and Graphics&Image Processing, Lanzhou* 730070, *China*)
* *Corresponding author, E-mail: edukeylab*@126.*com*

Abstract: Aiming at the problems of background interference and drastic change of crowd scale in crowd counting, which leads to poor counting effect, a crowd counting method enhanced by dense connected attention and scale perception recombination was proposed. First, a feature extraction network with dense connected attention mechanism was designed to enhance the crowd counting features and suppress the background interference by using the inflated convolutionally improved VGG19 network as the model coarse feature extraction network and embedding the dense connected dual-channel attention mechanism. Then, the scale-aware reorganized upsampling and soft mask feature enhancement and delivery structures were designed to achieve the full utilization of crowd feature information at different scales from shallow to

收稿日期：2024-04-20；**修订日期**：2024-06-18.
基金项目：国家自然科学基金(No. 62462043，No. 61963023)；兰州交通大学基础研究拔尖人才项目(No. 2022JC36)

deep, and overcome the problem of poor counting performance due to drastic changes in crowd scales. Secondly, a multi-resolution fusion module was proposed to enhance the interaction between multi-resolution information, reduce the semantic gap between different resolutions, and improve the accuracy of crowd counting. Finally, comparison experiments were conducted on ShanghaiTech, UCF-QNRF, JHU_CROWD++ and other crowd counting datasets, and the results show that the proposed method outperforms the comparison algorithms. For instance, compared with DM-Count crowd counting algorithm, the MAE and MSE error values of proposed method are reduced by 15.98% and 14.52%, respectively, and the proposed method has higher counting performance in crowd counting.

Key words: crowd counting; dense connected attention; scale perception recombination; multi-scale enhancement; multi-resolution fusion

1 引　言

随着经济的飞速发展和城市人口的高度集中,公共场所频繁发生人群聚集事故,给社会造成巨大的人员伤亡和财产损失[1]。通过人群计数合理管控人群聚集,可以有效降低人群聚集事故[2]。如何提高计数的准确性是目前研究的热点问题[3]。

目前,人群计数方法分为传统方法和深度学习方法。其中,传统人群计数方法主要包括两类:一类是基于检测[4]的方法,该类方法采用滑动检测窗遍历待计数人群图像的方法,通过统计滑动窗口中的人数,达到人群计数的目的,该类方法在稀疏场景下计数有一定的效果,但对于密集和人群尺度差异较大的场景中,滑动窗口干扰较大,其计数性能较差。另一类是基于回归[5]的方法,通过提取人群像素或边缘特征,采用回归算法得到相应的人群计数值,该类方法较基于检测的方法克服了滑动窗口干扰大的问题,但该类方法在人群密集、背景干扰等场景同样存在特征提取能力受限的问题,导致人群计数性能下降。

基于深度学习[6]的人群计数方法,因其对图像特征出色的提取能力被广泛应用于人群计数领域。Ma等[7]提出一种基于编解码器的多尺度融合人群计数模型,该方法提高了对于多尺度计数信息的利用率,但其采用原始VGG19网络作为特征提取网络,固定的卷积核大小导致人群特征提取能力受限。Zhang等[8]提出了一种三分支的多列卷积人群计数模型,通过不同尺寸卷积核堆叠的形式提高人群计数尺度覆盖率,但普通卷积受背景干扰影响较大,影响其计数结果。Li等[9]提出了一种空洞卷积(Network for Congested Scene Recognition, CSRNet)计数模型,该方法通过增大感受野来捕捉深度空间语义信息,但该方法基于单层特征信息进行尺度扩张,易造成特征信息丢失,影响人群计数性能。Zhu等[10]提出了基于视觉注意力增强的计数模型,通过注意力图引导计数估计,但该模型受双列子网络中离散感受野的影响,其对人群尺度变化适应性较差。Liu等[11]提出了一种上下文感知的人群计数模型,利用注意力机制优化尺度感知的上下文特征来提高对人群变化尺度的适应性,但该方法多通道特征图拼接时易受背景干扰,影响计数性能。Shen等[12]提出一种利用U-Net网络的像素级人群计数模型,并引入对抗损失来提高计数精度,但U-net解码器易受感受野选取影响,当感受野较大时对应的池化操作会降低计数精度。

综上所述,目前人群计数仍存在背景干扰、人群尺度变化剧烈,导致计数效果不佳的问题,因此本文提出了一种密集连接注意力与尺度感知重组增强的人群计数方法。本文所做主要工作包括:(1)设计基于密集连接注意力机制的特征提取网络,通过使用膨胀卷积来改进VGG19网络进行人群计数粗特征提取,并嵌入密集连接双通道注意力机制,增强人群计数特征,抑制背景干扰。(2)构建多尺度感知重组增强模块,设计尺度感知重组上采样和软掩膜特征增强及传递结构,实现从深层到浅层不同尺度人群特征信息的充分利用,然后在不同层次上传递尺度特征并进行聚合,克服人群尺度变化剧烈导致计数性能不佳的问题。(3)设计多分辨率融合模块,对各分辨率独立提取并融合,实现不同分辨率信息间的

相互共享，降低不同分辨率之间的语义差距，提高人群计数的准确度。最后在多个人群计数数据集上比较实验，结果表明所提方法具有更高的计数性能。

2　所提方法

2.1　整体网络

针对人群计数中背景干扰、人群尺度变化剧烈，导致计数效果不佳的问题，本文提出一种密集连接注意力与尺度感知重组增强人群计数模型。所提网络模型主要由：基于密集连接注意力机制的特征提取网络、多尺度感知重组增强模块、多分辨率融合模块及计数输出模块组成。所提方法整体模型结构，如图1所示。

模型工作时，首先将人群图像输入到基于密集连接注意力机制的特征提取网络，经过改进后VGG19的粗提取并对密集网络嵌入双通道注意力，增强了密度图的特征，抑制背景干扰。然后再进行多尺度感知重组增强，实现从深层到浅层中不同尺度下人群特征的增强。接着，利用多分辨率融合模块，将不同分辨率特征相互融合，提高计数精度。最后通过密度回归输出计数结果。

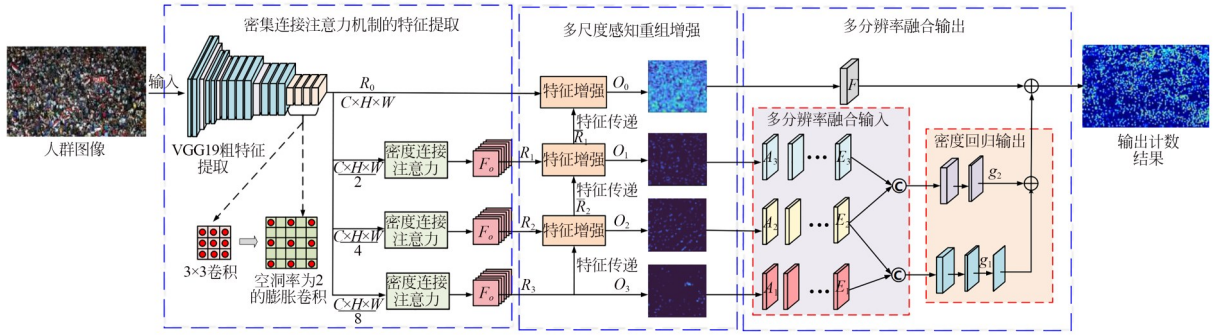


图1　网络整体结构

Fig. 1　Overall network architecture diagram

2.2　密集连接注意力机制的特征提取网络

人群计数时，易受到背景的干扰，影响计数性能[13]。因此，在图1本文所提网络模型中，特征提取网络主要由改进后的VGG19和密集连接双通道注意力网络组成。由于原始VGG19网络采用卷积核固定，导致其感受野固定且易丢失图细节，无法更好捕捉特征信息，因此本文将VGG19中3×3卷积替换成空洞率为2的膨胀卷积[14]，通过扩大感受野来实现不同尺度的粗特征提取。在此基础上，为了减少背景干扰对人群计数的影响，本文进一步设计密集连接卷积注意力模块，将提取到的人群空间分布粗提取特征图输入密集连接卷积注意力网络结构中，加强网络对于背景干扰的抑制能力，该网络结构如图2所示。

在图2中，首先通过3×3卷积获得特征图F_1，然后计算通道注意力得到注意力图F_c，该过程如(1)：

$$F_c=\sigma\Big(MLP\big(AvgPool(F_1)\big)+MLP\big(MaxPool(F_1)\big)\Big),\tag{1}$$

其中：MLP为全连接，$AvgPool$和$Maxpool$分别为平均池化和最大池化。

在得到注意力图F_c后，与原特征图F_1相乘进行加权，得到特征关联图M_c，如式(2)：

$$M_C=F_1\times F_C.\tag{2}$$

为了进一步提高所提方法对背景干扰的抑制能力，采用密集连接网络DenseNet[15]的思想，将每一层与前面所有层直接连接，扩大尺度的多样性和特征感受野，来提高空间和语义信息的捕捉能力。式(2)中特征图M_c所对应空间位置上的每个人群特征，利用最大池化与平均池化进行双池化操作，并通过7×7的卷积层，编码通道信息，得到空间注意力图F_s，如式(3)所示。在得到特征图F_s后与式(2)特征图M_c相乘，从而得到加权特征图M_s，如式(4)，通过上述操作可以得到注意力增强后的特征图M_s。

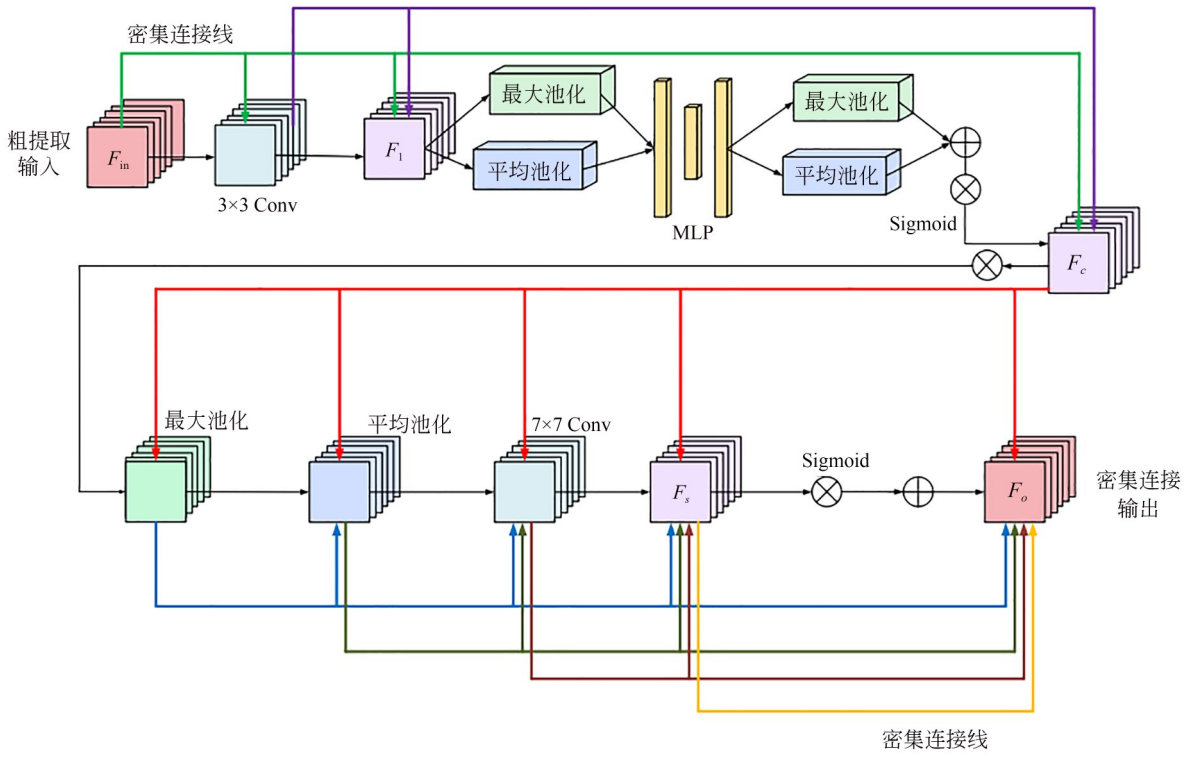


图 2　密集连接卷积注意力网络结构

Fig. 2　Densely connected convolutional attention network

$$F_s=\sigma\left(f^{7\times7}\left(\left[AvgPool(M_c);MaxPool(M_c)\right]\right)\right), \tag{3}$$

$$M_S=M_C\times F_S, \tag{4}$$

其中：σ 为 *sigmoid* 激活函数，$f^{7\times7}$ 为 7×7 卷积。

将粗提取输入特征图 F_{in} 与式(4)注意力增强后的特征图 M_s 通过密集的连接结构输出最终特征图 F_o，具体公式如式(5)：

$$F_o=F_{in}+M_s. \tag{5}$$

为了验证本文所提密集连接注意力机制的有效性，将其与改进前原始注意力进行热力图对比实验，如图 3 所示。其中，图 3(b)为原始注意力特征提取图，可以看出改进前对人群特征关注能力有限，仅能关注到少部分人群特征，人群特征缺失较多。图 3(c)为本文所提改进后密集连接注意力，可以看出采用密集连接注意力机制网络后，能有效对密集处人群进行关注，较好地抑制了背景干扰的问题，有助于提升计数的准确性，从而验证了所提方法的有效性。

(a) 人群图像
(a) Crowd image

(b) 原始注意力特征提取图
(b) Original attention feature extraction map

(c) 改进后密集连接注意力
(c) Improved dense connection attention

图 3　注意力对比热力图

Fig. 3　Attention comparison heat-map

2.3　多尺度感知重组增强模块

在完成密集连接注意力人群计数特征提取后，为了充分融合深层与浅层的特征信息，克服尺度变化对人群计数的影响，本文进一步设计了多尺度感知重组增强模块，如图 4 所示。该模块首先设计尺度感知重组上采样，提高特征图分辨率，然后使用软掩膜特征增强模块进行不同尺度特征增强及传递，实现不同尺度人群特征信息的充分利用，来解决人群计数时尺度变化过大的问题，以提高人群计数模型的性能。该模块由两部分构成：尺度感知重组上采样、软掩膜特征增强及传递。

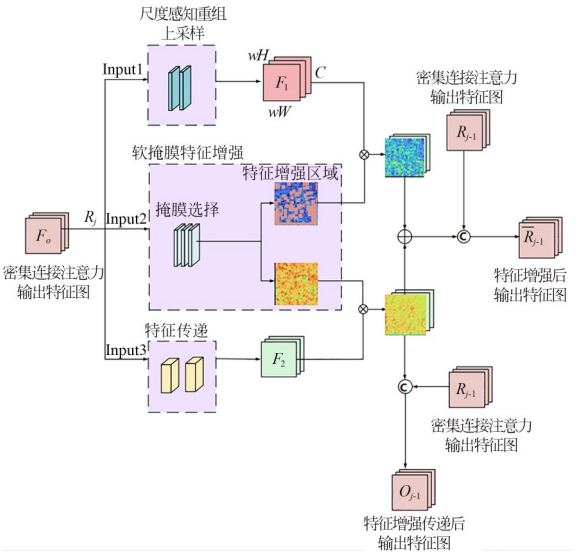


图 4　多尺度感知重组增强模块

Fig. 4　Multi-scale perception reorganization enhancement module

2.3.1　尺度感知重组上采样

在图 4 中，尺度感知重组上采样用于提高人群计数特征图分辨率，以便检测到更多的人群细节特征。传统的上采样方法仅简单地将像素复制或插值到更高尺度中，然而上述方式将会导致图像细节的丢失和模糊，不利于人群计数。因此，本文基于 CARAFE[16] 结构设计了尺度感知重组上采样模块，来提高特征图分辨率，如图 5 所示。CARAFE 是一种利用上采样核完成上采样操作的模块，其可以强化高分辨率的低层次特征图和低分辨率的高层次特征图。本文尺度感知重组模块，首先，使用 1×1 卷积层将特征通道压缩为 C_1，并使用卷积预测上采样核。在输入通道

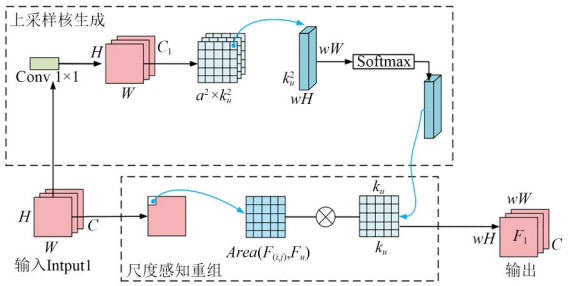


图 5　尺度感知重组上采样结构

Fig. 5　Scale-aware reorganizing the upsampling architecture

数为 C_1、输出通道数为 $\alpha^2\times k_u^2$ 的情况下，生成大小为 $wH\times wW\times k_u^2$ 的上采样核；接着，将得到的上采样核通过 sotfmax 进行归一化处理。

完成上采样操作后，继续进行尺度感知重组操作，使用加权和算子对特征图中心 k_u^2 区域和上采样预测值做点积，得到输出特征图 $\boldsymbol{F}_1(i',j')$，具体公式如式(6)所示：

$$\boldsymbol{F}_1(i',j')=\sum_{n=-r}^{r}\sum_{m=-r}^{r}Kw(Area(\boldsymbol{F}_{(i,j)},(k_u-2)))\times F(i+n,j+m),\tag{6}$$

其中：$r=\frac{k_u}{2}$，$F(i,j)$ 代表特征图中位置为 (i,j) 的特征信息；(i',j') 表示 $(\frac{i}{w},\frac{j}{w})$，$w$ 为上采样倍数，Kw 为生成的上采样核，Area(F,k)获取以 F 为中心，周围距离为 k 的领域特征信息。尺度重组后提高了输出特征图分辨率。

2.3.2　软掩膜特征增强及传递

在完成图 5 尺度感知重组上采样后，进一步采用软掩膜机制(Soft Mask Mechanism)加强对不同尺度的信息交互能力。软掩膜机制[17-18]利用软掩膜生成器增强模型对目标的关注程度，生成一组特定特征的掩膜，可以调整不同尺度特征的重要性。借助于软掩膜机制，可以更灵活地聚焦到不同尺度下的人群特征信息，以增强不同尺度的特征信息，提高人群的计数精度。

尺度感知重组上采样后，进一步采用软掩膜机制进行人群尺度特征增强，如图 6 所示。利用二进制掩膜机制对图 4 中特征图进行人群尺度特征增强，过程如式(7)所示，该过程对计数人群的特征权重 m_r 进行软掩膜调整，对于尺度区间内的掩膜权重进行区分标记，尺度区间内掩膜权重 m

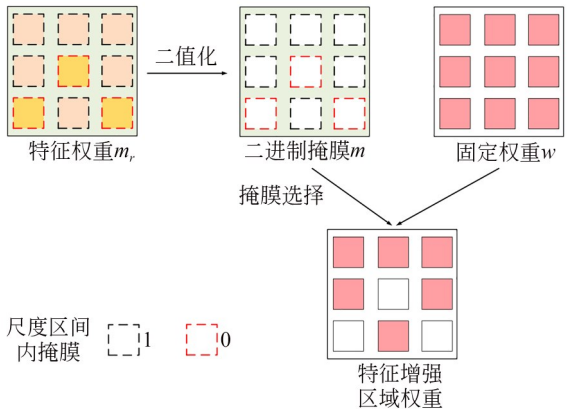


图 6　软掩膜选择机制

Fig. 6　Soft mask selection mechanism

标记为 1，而非该尺度区间掩膜权重标记为 0，即对尺度范围内的特征进行保留，而对非尺度范围内的尺度特征进行弱化。

$$m=\begin{cases}1, if\ \tau_1 \geqslant M_r \geqslant \tau_2\\ 0, otherwise\end{cases}, \tag{7}$$

其中：阈值 τ_1 为最深层尺度 $\frac{C\times W\times H}{8}$，阈值 τ_2 大小为浅层尺度 $\frac{C\times W\times H}{2}$，通过式(7)得到相应的二进制掩膜 m，然后将掩膜 m 映射到固定权重 w，从而得到软掩膜特征增强后的人群尺度特征 M_j^g，式(8)所示：

$$M_j^g=\sum_{j=0}^{N-j} mw. \tag{8}$$

通过式(8)软掩码选择机制，可以实现不同尺度特征的增强。

完成图 6 软掩膜选择后，进一步设计了特征传递模块，通过两组并列的 1×1 和 3×3 卷积在不同层次捕获特征信息，将语义信息聚合获取更丰富的人群细粒度特征。特征传递过程中，每个尺度的特征图从前一级中保留所选择的掩膜，使得不同尺度间特征目标逐步累积，最终将所有特征以最高尺度进行聚合，完成多尺度特征增强。选择与传递过程，如式(9)所示：

$$\begin{cases} l_j^s=L_1\left(M_j^g\odot E_{\theta e}\left(O_j\right), M_j^g\odot E_{\theta e}\left(O_j^g\right)\right)\\ l_j^l=L_1\left(U\left(M_{j+1}^g\right)\odot E_{\theta e}\left(O_j\right), U\left(M_{j+1}^g\right)\odot E_{\theta e}\left(O_j^g\right)\right)\end{cases}, \tag{9}$$

其中：$E_{\theta e}\left(O_j\right)$ 为预测人群计数图，$E_{\theta e}\left(O_j^g\right)$ 为真实人群计数图，l_j^s 为 R_j 级特征选择，l_j^l 为 R_j 级保留 R_{j+1} 级的特征，U 为上采样运算，L_1 为欧氏距离用来衡量预测掩码选择与真实掩码选择间人群计数图的误差，作为 R_j 级特征选择。

通过尺度感知重组上采样、软掩膜特征增强及传递操作后，最后进行特征拼接，如式(10)所示：

$$\begin{cases} O_{j-1}=C\left\{R_{j-1}, K_a\odot F_1\right\}\\ \overline{R}_{j-1}=C\left\{R_{j-1}, K_b\odot F_2+K_a\odot F_1\right\}\\ K=Soft\max\left(M_j^g\left(\overline{R}_j\right), \dim=0\right)\end{cases}, \tag{10}$$

其中：C 为特征拼接，$j=1,2,\cdots,F_2$ 为特征传递后特征图，$K\in R^{2\times h_j\times w_j}$ 是一个双通道注意力图，它沿着通道维度分为 K_a 和 K_b，$\odot$ 表示哈达玛积。R_{j-1} 和 $\overline{R}_j$ 为密集连接注意力输出的特征图，$\overline{R}_{j-1}$ 为特征选择后的输出特征图，O_{j-1} 为特征选择传递后输出的特征图。

2.4　多分辨率融合输出

在完成多尺度感知重组增强模块后，最后进行多分辨率融合输出计数结果。在一般的多分辨率融合过程中，不同分辨率层提取到的特征存在差异性，影响计数性能[19]。因此，本文设计了多分辨率融合模块，对不同分辨率特征提取并融合，达到不同分辨率间信息相互共享，克服不同分辨率之间的语义差距，以提高计数性能，具体结构如图 7 所示。

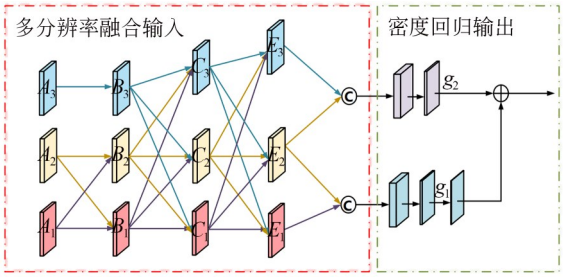


图 7　多分辨率融合结构

Fig. 7　Multi-resolution fusion structure

图 7 多分辨率融合时，分别使用空洞率大小为 1，2，3 的卷积处理输入特征图，得到 $\{A_1, A_2, A_3\}$ 特征图，而后将 A_1，A_2 特征融合，再与 A_3 进行空洞卷积，得到特征图 B_i，具体融合过程如式(11)所示：

$$B_i=\begin{cases}conv_d\Big(A_1+conv(A_2)+conv\big(cat(A_1,A_2)\big)\Big), i=1\\ conv_d\Big(A_2+conv(A_1)+conv\big(cat(A_1,A_2)\big)\Big), i=2,\\ conv_d(A_3), i=3\end{cases} \tag{11}$$

其中：$conv_d$为空洞率为d的空洞卷积，$conv$为1×1卷积，cat是串联操作。

然后，对式(11)得到的三个特征相互融合，得到输出特征图为C_i，如式(12)所示：

$$C_i=conv_d\Big(\sum_{j\neq i}conv\big(B_j\big)+conv\big(cat(B_1,B_2,B_3)\big)\Big)\quad i=1,2,3. \tag{12}$$

重复式(12)操作，将各分辨率特征融合从而实现了人群计数不同分辨率特征信息间共享。在完成不同分辨率特征融合后，最后通过密度图回归模块输出人群计数结果。将多分辨率特征融合输出特征串联后，如图7右侧融合输出，再分别使用1×1和3×3卷积得到密度图g_1和g_2。最后对g_1使用3×3卷积加权后融合到g_2中，得到最终人群计数密度图结果，从而完成人群计数结果输出。

2.5　损失函数

目前在常用的人群计数网络中，通常使用欧式距离作为损失函数，但它未考虑预测人数损失，当预测值与真实值相差较大时，会使误差变大，影响计数性能。为了克服上述欧式距离损失的不足，本文采用欧式距离损失和基于回归人数损失的联合损失函数，其中欧式距离损失$L_2(\theta)$用于测量预测密度图与真实密度图之间的差值，具体公式为式(13)：

$$L_2(\theta)=\frac{1}{2N}\left\| A_i(X_i;\theta)-B^{GT}{}_i\right\|_2^2, \tag{13}$$

其中：θ是学习参数，N是训练样本数，$A(X_i;\theta)$和B^{GT}分别代表第i张训练图像的预测密度图和真实密度图，X_i表示第i张训练图。

在欧式距离损失基础上，进一步设计基于回归人数的损失，来加强人群计数准确性，本文采用基于回归人数的损失L_{NOR}，如式(14)：

$$L_{\mathrm{NOR}}=\left\| Y_i-Y_i{}^{GT}\right\|^2, \tag{14}$$

其中：i表示第i个训练样本，Y^i表示预测人数$Y_i{}^{GT}$表示真实人数，最终损失L_{loss}由欧式距离损失$L_2(\theta)$和基于回归人数损失L_{NOR}组成，如式(15)：

$$L_{\mathrm{loss}}=\lambda_1L_2(\theta)+\lambda_2L_{\mathrm{NOR}}, \tag{15}$$

其中，λ_1和λ_2为损失函数权重系数。λ_1为欧式距离损失权重系数，在基于密度图的人群计数方法中，由于生成的密度值遵循逐像素的预测，因此输出的密度图必须包含空间相关性，以便能够呈现最近像素之间的平滑过渡，通常使用欧式距离作为损失函数，其权重系数λ_1一般设置1，用来提高密度图的计数精度[8]。但是欧式距离损失函数的主要缺点是对离群点比较敏感，因而本文为了进一步提高模型精度，在损失函数设计时，引入了回归损失，通过人群计数模型预测值与真实值的比较来进一步提高精度。在本文所提模型中，对权重λ_2取值时，在0～1间进行不同的线性取值，分别在数据集ShanghaiTech，UCF-QNRF，JHU_CROWD++进行实验，结果如表1所示。通过分析可以看出在不同数据集中进行实验时，当$\lambda_{2=}0.1$时，MAE与MSE最优，说明模型计数性能效果，因而本文损失函数中λ_2取0.1。

表1　不同数据集权重λ_2取值实验结果

Tab.1　Experimental results of different data centralization with heavy λ values

权重λ	SHA		SHB		UCF-QNRF		JHU_CROWD++	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
1.0	65.1	122.1	14.7	20.0	86.3	177.7	74.1	273.6
0.9	63.2	111.4	13.9	18.9	84.2	156.3	72.9	269.0
0.8	62.7	108.8	12.5	17.4	82.6	155.4	70.2	264.9
0.7	61.0	107.2	11.0	16.7	80.1	154.0	68.5	258.1
0.6	59.4	105.3	10.2	15.1	79.8	153.8	66.9	254.7
0.5	58.6	104.7	9.6	13.4	78.0	152.3	64.2	253.2
0.4	57.2	103.2	8.9	12.1	77.4	148.1	62.1	250.6
0.3	56.8	101.4	7.2	10.5	76.2	146.7	59.6	248.1
0.2	56.1	100.2	6.3	9.9	75.6	145.0	58.3	246.9
0.1	**55.4**	**99.6**	**5.8**	**9.3**	**74.3**	**143.1**	**57.3**	**245.3**

3　实　验

3.1　实验数据与实验环境

为验证所提方法的有效性，在人群计数公开数据集 ShanghaiTech[8]、UCF-QNRF[20] 和 JHU_CROWD++[21]分别进行对比实验，每种数据集按照 7∶3 的比例划分为训练集和测试集。进行对比实验。硬件配置为 Intel Core i9-12900K，NVIDIA RTX A5000，相同环境下进行对比实验。初始学习率为 1×10^{-5}，batch-size 为 2，使用 AdamW 优化器。

3.2　评价指标

在人群计数中，计数误差主要采用平均绝对误差（Mean Absolute Error，MAE）和均方误差（MeanSquareError，MSE）比较指标，定义如式（16）：

$$\begin{aligned} MAE&=\frac{1}{p}\sum_{i}^{p}\left|\hat{q}_i-q_i\right| \\ MSE&=\sqrt{\frac{1}{p}\sum_{i}^{p}\left(\hat{q}_i-q_i\right)^2} \end{aligned}, \tag{16}$$

其中：p 代表图像数量；$\hat{q}_i$ 和 q_i 分别表示估计人群数量和真实数量。MAE 和 MSE 越小，说明误差越小，计数性能越好。

3.3　Shanghai Tech 数据集

ShanghaiTech 是目前人群计数评估常用计数数据集，共包含 1 198 张带注释的图像，总计 330 165 人，该数据集由 SHA 和 SHB 两部分组成。其中，SHA 为网络随机选取人群图像，来源广泛，由不同相机视角下不同密度人群场景构成。SHB 为上海街道的实际监控拍摄，由分布不均匀且具有背景遮挡的人群构成，场景相对固定。首先进行该数据集下计数性能比较。

将本文方法与 MCNN[8]，CRSNet[9]，ACSCP[12]，DM-Count[22]，DCANet[23]，FIDTM[24]，MP-Count[25] 和 FFDB[26] 进行对比。表 2 为 Shanghai Tech 数据集实验评价指标对比，可以看出 MCNN 的 MAE 与 MSE 值在所有对比方法中值最大，评价值越大说明其计数性能越差，这是因为该方法使用三个分支固定感受野卷积进行计数，对于尺度变化较大的情况下，简单固定感受野卷积无法有效抑制背景干扰，导致该方法计数性能较差。在 SHA 数据集 MSE 评价时，DM-Count 略优于所提方法，这是因为 DM-Count 模型使用最优传输（Optimal Transport，OT）来计算预测密度图与实际密度图之间的相似性，对于网络图片资源随机抓取构成的 SHA 数据集，OT 算法在处理随机概率分布转换时，其准确性更高，但该方法在场景更加复杂的 SHB 数据集时，其准确性较低。综合表 2 中可以看出，所提方法相较其他算法综合计数性能更好。

为直观展示该数据集下人群计数的差异，将本文方法与 DM-Count，DCANet，FIDTM，MP-Count 和 FFDB 进行可视化比较，如图 8 所示。可以看出本文方法图 8(f) 与计数真值图 8(b) 更接近，说明本文方法相较于对比方法人群计数精度更高。

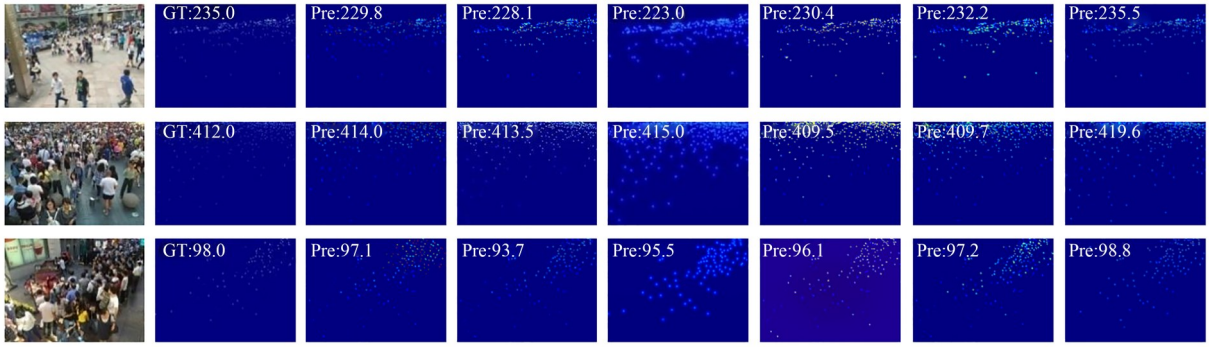


图 8　ShanghaiTech 数据集实验结果

Fig. 8　Experimental results of ShanghaiTech dataset

表 2　不同模型在 ShanghaiTech 数据集上的实验结果

Tab. 2　Experimental results of different models on ShanghaiTech

方法	SHA		SHB	
	MAE	MSE	MAE	MSE
MCNN[8]	110.5	173.0	26.8	41.3
CSRNet[9]	68.5	115.4	10.9	16.2
ACSCP[12]	75.2	102.1	17.0	27.1
DM-Count[22]	58.8	95.4	7.1	11.5
DCANet[23]	61.1	108.2	8.4	15.0
FIDTM[24]	57.4	103.9	7.3	12.0
MP-Count[25]	57.0	103.1	7.2	12.4
FFDB[26]	56.6	100.7	6.8	10.6
本文方法	**55.4**	**99.6**	**5.8**	**9.3**

3.4　UCF-QNRF 数据集

UCF-QNRF 是一个密度和背景更加复杂的大型计数数据集，其包含 1 535 张图片，包含建筑物、植物、道路等更复杂场景，其计数任务更困难。

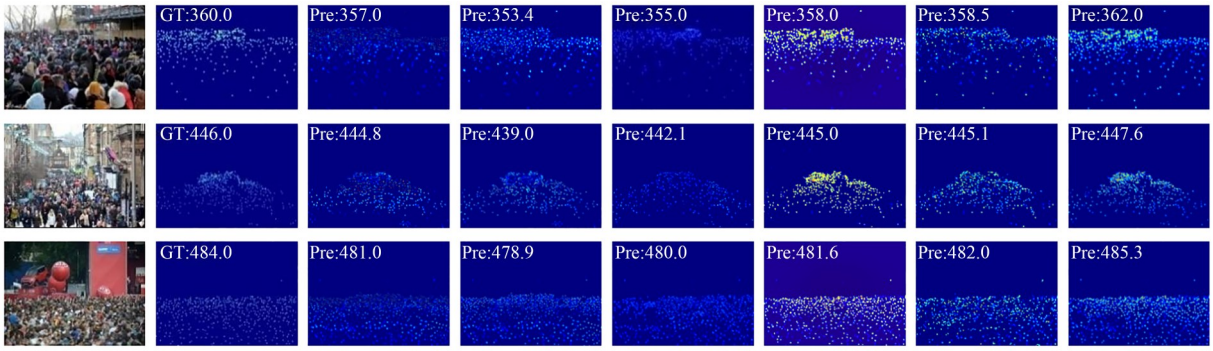
该数据集下，将本文方法与 Switch-CNN[27]，TEDNet[28]，DUBNet[29]，DM-Count[22]，DCANet[23]，FIDTM[24]，MP-Count[25] 和 FFDB[26] 更多人群计数方法进行对比。表 3 为 UCF-QNRF 数据集实验评价指标对比，可以看出 Switch-CNN 的 MAE 与 MSE 值在所比方法上最大，说明其计数性能较差，这是因为该方法利用图像块切片来进行人群密度估计，由于 UCF-QNRF 数据集包含更复杂的背景信息景，导致该模型在图像块切片搜索时易受到上述复杂背景的影响，造成其计数精确度较低。表 3 中本文所提方法相较于其他算法的误差最低，说明在 UCF-QNRF 数据集下计数性能更优。

表 3　不同模型在 UCF-QNRF 数据集上的实验结果

Tab. 3　Experimental results of different models on UCF-QNRF

方法	UCF-QNRF	
	MAE	MSE
Switch-CNN[27]	228.4	426.3
TEDNet[28]	112.9	187.7
DUBNet[29]	105.8	180.9
DM-Count[22]	85.2	148.0
DCANet[23]	99.0	177.3
FIDTM[24]	88.7	153.1
MP-Count[25]	88.5	151.3
FFDB[26]	78.1	148.6
本文方法	**74.3**	**143.1**

在此数据集上使用本文方法与 DM-Count，DCANet，FIDTM，MP-Count 和 FFDB 进行可视化比较，如图 9 所示。可以看出，本文方法在尺度变化较大、背景干扰情况下，计数依然性能更优。



(a) 输入图像 (a) Input image　(b) 真实值 (b) True value　(c) DM-Count　(d) DCANet　(e) FIDTM　(f) MP-Count　(g) FFDB　(h) 本文方法 (h) Ours

图 9　UCF-QNRF 数据集实验结果

Fig. 9　Experimental results of UCF-QNRF dataset

为更直观关注本文在密集场景人群计数中处理遮挡问题的有效性，对遮挡人群进行局部放大可视化说明。如图 10 所示(彩图见期刊电子版)，图 10(a)为原始人群图像，图 10(b)为图 10(a)中黄色人群遮挡区域的局部放大图，图 10(c)为图 10(b)局部放大图采用本文方法得到的热力图，在热力图中通过颜色的标注能够直观的展示人群分布情况。在图 10(b)局部放大图中，第一幅红框区域存在宣传牌遮挡人群的现象，第二幅红框区域存在植物遮挡人群的现象，通过采用本文所提方法后，得到对应的热力图 10(c)，可以看出，对局部遮挡情况下人群区域仍能有效关注。在图 10(d)本文方法预测的人群密度图结果中，可以看出在图 10(d)红框区域实现了有效计数，从而验证了所提方法对于遮挡场景人群计数的有效性，其原因为本文利用特征提取网络中膨胀卷积扩大了特征感受野，能够防止细节特征丢失，并采用密集连接卷积注意力模块，来提高空间和语义信息的捕获能力，增强了人群特征的提取。上述方法可以有效减少遮挡及背景干扰对人群计数的干扰，提高了人群计数能力。

(a) 人群图像
(a) Crowd image

(b) 遮挡局部放大图
(b) Enlarged image of partially occluded area

(c) 遮挡局部本文方法热力图
(c) Thermal map of occlusion in local areas using the proposed method

(d) 本文方法预测结果
(d) Ours

图 10　遮挡处理实验结果

Fig. 10　Occlusion processing experiment results

3.5　JHU_CROWD++数据集

JHU-CROWD＋＋数据集共 4 372 张图像，平均分辨率为 1 430×910，共计 151 万个注释。与现有的数据集相比，该数据集收集了不同场景、环境条件下的图像，提供了大规模密集场景中尺度变化较大、背景干扰明显等场景。

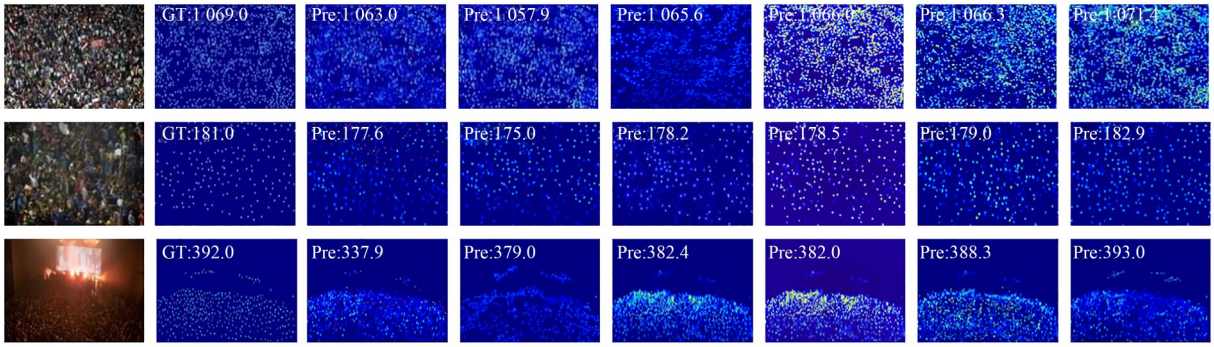
将本文方法与 MCNN[8]，CRSNet[9]，CAN[11]，DM-Count[22]，DCANet[23]，FIDTM[24]，MP-Count[25]和 FFDB[26]进行比较。表 4 为 JHU_CROWD＋＋数据集实验评价指标对比，可以看出所提模型相较于采用 T2T-ViT (token -to-token Vision Transformer)结构的 FFDB 模型 MAE 和 MSE 较小，因为 FFDB 模型将输入图像转换为标记序列并利用自注意机制捕获人群特征，但针对大规模人群数据集 JHU-CROWD＋＋基于 Transformer 结构的 FFDB 模型受自注意力机制复杂性的影响，会增加模型计算量，导致计数效果不佳。表 4 中本文所提方法 MAE 和 MSE 评价值最小，说明其计数误差更小。

表 4　不同模型在 JHU_CROWD++数据集上的实验结果

Tab. 4　Experimental results of different models on JHU_CROWD＋＋

方法	JHU_CROWD＋＋	
	MAE	MSE
MCNN[8]	188.2	482.7
CRSNet[9]	86.1	309.5
CAN[11]	99.4	313.2
DM-Count[22]	68.2	287.0
DCANet[23]	71.1	177.4
FIDTM[24]	65.8	253.1
MP-Count[25]	70.9	260.7
FFDB[26]	59.2	255.0
本文方法	**57.3**	**245.3**

在此数据集上将所提模型与 DM-Count，DCANet，FIDTM，MP-Count 和 FFDB 进行可视化比较，如图 11 所示。可以看出在 JHU＿CROWD＋＋数据集下所提模型预测人群输出值相较于其他方法与真实值误差更小，这是因为本文多尺度感知重组增强模块能有效提高密集场景下人群尺度分布不均计数性能。



(a) 输入图像 (a) Input image　(b) 真实值 (b) True value　(c) DM-Count　(d) DCANet　(e) FIDTM　(f) MP-Count　(g) FFDB　(h) 本文方法 (h) Ours

图 11　JHU＿CROWD＋＋数据集实验结果

Fig. 11　Experimental results of JHU＿CROWD＋＋ dataset

3.6　模型量化分析

为衡量本文所提模型与各种对比模型的参数量和推理速度，在 JHU＿CROWD＋＋数据集上进行参数量（Param）与浮点计算量（GFLOPs）的量化对比实验验证。其中，Param 的单位为百万（M），GFLOPs 的单位为千兆（G）。Param 越小模型复杂度越低，在训练和推理过程中减少了计算资源和时间的需求。FLOPs 值越小，模型在单位时间内执行的计算量越少，整体上计算效率更高。实验结果如表 5 所示。

对表 5 中数据进行分析发现，CSRNet 模型采用多次大卷积核进行堆叠，导致模型参数量较大，在推理时会耗费大量资源，GFLOPs 值高达 857.84 G，不利于有效提取人群特征。本文相较于其他对比算法，浮点计算量在一定程度上有所降低，说明本文方法推理速度较好。

表 5　模型量化指标对比

Tab. 5　Model quantitative index comparison

方法	JHU＿CROWD＋＋	
	Param/M	GFLOPs/G
MCNN[8]	0.13	56.21
CRSNet[9]	16.26	857.84
CAN[11]	18.1	93.58
DM-Count[22]	21.5	60.8
DCANet[23]	35.2	117.6
FIDTM[24]	56.7	70.34
MP-Count[25]	51.4	78.2
FFDB[26]	47.1	79.8
本文方法	**23.3**	**42.49**

3.7　模型泛化能力分析

为了验证本文所提模型的泛化能力，本文在 JHU＿CROWD＋＋数据集上对模型进行训练，并将训练后的模型分别部署到数据集 SHA、

表 6　模型泛化性对比

Tab. 6　Comparison of model generalization

模型部署	方法	MAE	MSE
JHU＿CROWD＋＋到 SHA	FIDTM[24]	88.6	150.0
	MP-Count[25]	86.9	145.5
	FFDB[26]	84.2	142.8
	本文方法	66.1	127.4
JHU＿CROWD＋＋到 SHB	FIDTM[24]	97.2	195.5
	MP-Count[25]	96.4	192.0
	FFDB[26]	95.9	188.1
	本文方法	84.7	155.2
JHU＿CROWD＋＋到 UCF－QNRF	FIDTM[24]	106.2	296.4
	MP-Count[25]	94.7	294.8
	FFDB[26]	89.4	291.0
	本文方法	55.1	223.8

SHB与UCF-QNRF数据集上。与FIDTM、MP-Count和FFDB模型的比较结果如表6所示。

对表6中数据分析发现，把在JHU_CROWD++数据集上训练模型分别应用到分布差异较大的新场景时，所提模型的MAE与MSE都最低。结果表明本文所提模型在跨数据集泛化时，相较于对比模型具有较好的泛化能力。

3.8 消融实验

为验证所提方法各个模块的有效性，在JHU_CROWD++数据集上进行消融实验，以特征网络Backbone为基础进行实验，分别添加本文所提模型的密集连接双通道注意力模块、尺度感知重组上采样模块、软掩膜特征增强模块、特征传递模块和多分辨率融合模块进行消融实验，分别用B_1、B_2、B_3、B_4和B_5表示。实验结果如表7所示。

表7中可以看出，采用本文粗提取网络可以在一定程度上提升计数的性能，说明该结构增强了密度图特征提取能力。接着本文所提方法在该结构的基础上，添加了密集连接双通道注意力模块，抑制复杂背景的干扰，突出人群特征，提高模型的计数能力。然后嵌入尺度感知重组上采样模块、软掩膜特征增强模块和特征传递模块，累积从浅层到深层尺度中的多尺度特征，通过该模块的MAE和MSE都取得进一步的优化，并分别降低为61.2和225.4。最后在加入多分辨率融合模块后，均有效降低了人群计数的误差，其MAE和MSE均有一定程度的减少。随着模块的逐步添加，模型的参数量和推理速度也有一定程度的增加，提高了模型的计数能力。通过消融实验充分证明了本文不同模块对网络的改进作用。

表7 消融实验指标对比

Tab. 7 Comparison of ablation experimental indicators

Backbone	模块					MAE	MSE	Param	GFLOs
	B_1	B_2	B_3	B_4	B_5				
√						86.6	314.1	8.9	22.87
√	√					82.3	301.2	10.2	25.42
√	√	√				79.1	298.3	11.6	29.7
√	√	√	√			68.4	277.5	15.8	33.1
√	√	√	√	√		61.2	255.4	18.3	39.5
√	√	√	√	√	√	57.3	245.3	23.3	42.49

4 结　论

本文提出了一种密集连接注意力与尺度感知重组增强的人群计数模型。通过构建密集连接注意力机制网络，抑制背景干扰。并通过多尺度感知重组增强模块，克服尺度变化影响计数性能问题。最后，利用多分辨率融合实现多尺度信息交互，减小语义差距，提高人群计数的准确度。实验结果表明，所提方法均优于对比算法，相较于DM-Count人群计数算法，MAE，MSE误差分别下降了15.98%，14.52%，在不同复杂场景中人群具有更高的计数性能。

参考文献：

[1] SHAO C H, SHAO P C, KUO F M. Stampede events and strategies for crowd management[J]. *Journal of Disaster Research*, 2019, 14(7): 949-958.

[2] YANG Z Y, WEN J, HUANG K D. A method of pedestrian flow monitoring based on received signal strength[J]. *EURASIP Journal on Wireless Communications and Networking*, 2022, 2022(1): 2.

[3] KHAN M A, MENOUAR H, HAMILA R. Revisiting crowd counting: state-of-the-art, trends, and future perspectives[J]. *Image and Vision Computing*, 2023, 129: 104597.

[4] LIU J, GAO C Q, MENG D Y, *et al*. DecideNet:

counting varying density crowds through attention guided detection and density estimation[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA*. IEEE, 2018: 5197-5206.

[5] LEMPITSKY V, ZISSERMAN A. Learning to count objects in images[C]. *Advances in Neural Information Processing Systems (NeurIPS). Vancouver*: MIT Press, 2010: 1324-1332.

[6] RANJAN V, SHARMA U, NGUYEN T, *et al*. Learning to count everything[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA*. IEEE, 2021: 3393-3402.

[7] MA Y M, SANCHEZ V, GUHA T. Fusioncount: efficient crowd counting via multiscale feature fusion[C]. 2022 *IEEE International Conference on Image Processing (ICIP). Bordeaux, France*. IEEE, 2022: 3256-3260.

[8] 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 (CVPR). Las Vegas, NV, USA*. IEEE, 2016: 589-597.

[9] 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, USA*. IEEE, 2018: 1091-1100.

[10] ZHU L, ZHAO Z J, LU C, *et al*. Dual path multi-scale fusion networks with attention for crowd counting[J]. *Computing Research Repository*, 2019, 1-9.

[11] LIU W Z, SALZMANN M, FUA P. Context-aware crowd counting[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA*. IEEE, 2019: 5094-5103.

[12] SHEN Z, XU Y, NI B B, *et al*. Crowd counting via adversarial cross-scale consistency pursuit[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA*. IEEE, 2018: 5245-5254.

[13] 余鹰, 李剑飞, 钱进, 等. 基于多尺度特征融合的抗背景干扰人群计数网络[J]. 模式识别与人工智能, 2022, 35(10): 915-927.
YU Y, LI J F, QIAN J, *et al*. Anti-background interference crowd counting network based on multi-scale feature fusion[J]. *Pattern Recognition and Artificial Intelligence*, 2022, 35(10): 915-927. (in Chinese)

[14] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual networks[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA*. IEEE, 2017: 636-644.

[15] DAI F, LIU H, MA Y K, *et al*. Dense scale network for crowd counting[J]. *CoRR*, 2021, 64-72.

[16] WANG J Q, CHEN K, XU R, *et al*. CARAFE: Content-aware ReAssembly of FEatures[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South). IEEE, 2019: 3007-3016.

[17] MALLYA A, LAZEBNIK S. PackNet: Adding multiple tasks to a single network by iterative pruning[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA*. IEEE, 2018: 7765-7773.

[18] MALLYA A, DAVIS D, LAZEBNIK S. Piggyback: adapting a single network to multiple tasks by learning to mask weights[C]. *European Conference on Computer Vision. Cham*: Springer, 2018: 72-88.

[19] SUN K, XIAO B, LIU D, *et al*. Deep high-resolution representation learning for human pose estimation[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA*. IEEE, 2019: 5686-5696.

[20] IDREES H, TAYYAB M, ATHREY K, *et al*. Composition loss for counting, density map estimation and localization in dense crowds[C]. *European Conference on Computer Vision. Cham*: Springer, 2018: 544-559.

[21] 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.

[22] WANG B Y, LIU H D, SAMARAS D, Minh H, *et al*. Distribution matching for crowd counting[J], *Advances in Neural Information Processing*

Systems (*NeurIPS*). *Vancouver*: *MIT Press*, 2020, 33: 1595-1607.

[23] YAN Z Y, LI P Y, WANG B, *et al*. Towards learning multi-domain crowd counting[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(11): 6544-6557.

[24] LIANG D K, XU W, ZHU Y Y, *et al*. Focal inverse distance transform maps for crowd localization [J]. *IEEE Transactions on Multimedia*, 2023, 25: 6040-6052.

[25] PENG Z X, CHAN S H G. Single domain generalization for crowd counting[C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Seattle*, *WA*, *USA*. IEEE, 2024: 28025-28034.

[26] SHI Z L, METTES P, SNOEK C G M. Focus for free in density-based counting[J]. *International Journal of Computer Vision*, 2024, 132(7): 2600-2617.

[27] SAM D B, SURYA S, BABU R V. Switching convolutional neural network for crowd counting [C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Honolulu*, *HI*, *USA*. IEEE, 2017: 4031-4039.

[28] JIANG X L, XIAO Z H, ZHANG B C, *et al*. Crowd counting and density estimation by trellis encoder-decoder networks [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *Long Beach*, *CA*, *USA*. IEEE, 2019: 6126-6135.

[29] OH M H, OLSEN P, RAMAMURTHY K N. Crowd counting with decomposed uncertainty[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(7): 11799-11806.

作者简介：

陈　永(1979—),男,甘肃武威人,教授,博士生导师,2014 毕业于兰州交通大学获得博士学位,主要从事深度学习及计算机视觉方面的研究。E-mail: edukeylab@126.com

董　珂(2000—),女,甘肃天水人,硕士研究生,2022 年于安康学院获得学士学位,主要研究方向为计算机视觉与图像处理。E-mail: 1743615325@qq.com