智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

比例融合与多层规模感知的人群计数方法

孟月波, 张娅琳, 王宙

引用本文:
孟月波,张娅琳,王宙. 比例融合与多层规模感知的人群计数方法[J]. 智能系统学报, 2024, 19(2): 307–315.
MENG Yuebo, ZHANG Yalin, WANG Zhou. Crowd counting method based on proportion fusion and multilayer scale–aware[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 307–315.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202208048

DOI: 10.11992/tis.202208048
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20231116.1447.010

比例融合与多层规模感知的人群计数方法

孟月波，张娅琳，王宙
（西安建筑科技大学 信息与控制工程学院，陕西 西安 710055）

摘　要：针对密集场景下人群图像拍摄视角或距离多变造成的多尺度特征获取不足、融合不佳和全局特征利用不充分等问题，提出一种比例融合与多层规模感知的人群计数网络。首先采用骨干网络 VGG16 提取人群密度初始特征；其次，设计多层规模感知模块，获得人群多尺度信息的丰富表达；再次，提出比例融合策略，根据卷积层捕获的特征权重重构多尺度信息，提取显著性人群特征；最后，采用卷积回归策略进行密度图的回归。同时，提出一种局部一致性损失函数，通过区域化密度图的方式增强生成密度图与真实密度图的相似度，提高计数性能。在多个人群数据集上的试验结果表明，所提模型优于近年人群计数的先进方法，且在车辆计数上有较好推广性。

关键词：人群密度估计与计数；卷积神经网络；多层规模感知；比例融合；局部一致性损失；密度图回归；多尺度信息；空洞卷积
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673−4785(2024)02−0307−09

中文引用格式：孟月波，张娅琳，王宙. 比例融合与多层规模感知的人群计数方法 [J]. 智能系统学报, 2024, 19(2): 307–315.
英文引用格式：MENG Yuebo, ZHANG Yalin, WANG Zhou. Crowd counting method based on proportion fusion and multilayer scale-aware[J]. CAAI transactions on intelligent systems, 2024, 19(2): 307–315.

Crowd counting method based on proportion fusion and multilayer scale-aware

MENG Yuebo, ZHANG Yalin, WANG Zhou
(College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: To deal with the problems of insufficient multiscale feature acquisition, poor fusion, and insufficient utilization of global features as a result of the changing view angles or distances of crowd images in dense scenes, we propose a crowd counting network based on proportion fusion and multilayer scale-aware. First, the backbone network VGG16 is employed to extract the initial characteristics of the population density. Subsequently, a multilayer scale-aware module is developed to acquire a rich expression of multiscale information from the crowd. Afterward, a proportional fusion strategy is designed to reconstruct the multiscale information based on the feature weights captured by the convolution layer and extract the significant crowd features. Lastly, convolution regression is utilized to regress the density map. Concurrently, a local consistency loss function is proposed, which improves the similarity between the generated density map and the real density map by regionalizing the density map and enhances the counting performance. The results of the experiments on multiple population datasets exhibit that the model proposed here surpasses the existing state-of-the-art methods of population density counting and has good generalization in vehicle counting.

Keywords: crowd density estimation and counting; convolutional neural network; multilayer scale-aware; proportional fusion; local consistency loss; density map regression; multiscale information; dilated convolution

随着城市化进程的加快以及人均消费水平的提高，城市人口数量和旅游人员总数急剧增加，局部区域的人群数量超出安全等级等状况时有发生，给治安管理、交通调度和疫情防控等方面带

收稿日期：2022−08−30.　**网络出版日期**：2023−11−16.
基金项目：陕西省重点研发计划项目（2021SF-429）.
通信作者：孟月波. E-mail: mengyuebo@163.com.

来了极大地挑战。充分地利用公共场所的视频监控设备来准确地进行人群密度估计与计数可以在一定方面减轻工作人员的负担，同时提高管理的效率。

传统的人群计数方法通过检测整体或部分身体的方式预估人群总数[1]，但密集场景下的严重遮挡导致计数性能低下；基于密度图回归的方法是目前实现人群计数任务的主流方式，不仅能够准确地预估人群数量而且能够直观反映人员的位置信息与聚集程度[2]。随着深度学习的迅猛发展，卷积神经网络（convolutional neural network，CNN）在人群计数任务上取得了极大地突破[3]，但现有 CNN 人群密度估计模型在距离多变和多视角场景下仍然存在较大的误差。为改善这些问题，人群密度估计与计数的网络框架多设计为多分支和多尺度的结构。

Zhang 等[4] 首次提出了人群计数多分支概念，通过设计 3 个卷积分支获取不同尺度的感受野，降低视角变换带来的影响，计数精度提升显著；在此基础上，Sam 等[5] 增加了 Switch 模块，利用密度分级策略来更好地回归密度图，进一步提升计数性能。但这类简单并联多个分支的结构训练成本较高，且忽略了分支间的信息互补。孟月波等[6] 采用带有空洞卷积的多分支卷积编码结构，将多尺度特征映射到同一空间维度，大幅度提升了训练效率。为增强分支间的信息交互，Shen 等[7] 采用对抗网络的方式生成 2 个密度生成器分支，一个作为密度图的最终预测结果，另一个通过网络对抗方式改善第一个分支的预测精度；Jiang 等[8] 在多分支任务中融入注意力思想，设计了注意力密度网络和注意力扩展网络，注意力密度网络旨在提升各密度等级分支的特征捕获能力，注意力扩展网络则着重于提高密度图的生成精度，这一方法虽然计数结果表现较优，但是网络的训练需提供额外的标注，进一步增加了多分支结构的训练成本。

多尺度的结构不需引入额外的分支网络，采用主干网络串联多尺度特征提取器的方式缓解多尺度特征不足，是近年来的研究热点。Li 等[9] 提出的 CSRNet 采用加大计数骨干网络的特征提取深度的方式强化网络的特征提取，通过空洞卷积的简单串联，在保持高分辨率的同时预测出较为优秀的密度图，但此方法只基于单层特征信息进行尺度扩张，易造成特征信息丢失。为充分利用多层特征信息，Xu 等[10] 引入规模学习思想，从多层特征中学习图像块需要缩放的比例因子，通过比例缩放实现头像块规模一致，从而缓解网络获取人群信息的压力；Jiang 等[11] 设计了多个解码器，通过特征交错连接的方式分层聚合不同编码阶段特征，获得人群图像的多尺度表达；Liu 等[12] 通过自适应编码人群密度所需的上下文信息，优化透视现象下的网络特征提取能力，但此类方法因人群全局特征的过量导入增加了计算量，降低了人群计数的总运算效率。为降低网络运算成本，Liu 等[13] 提出了局部密度计数方法，通过利用不同卷积特征的上下文和多尺度信息回归实现了局部图块的准确计数；但局部密度的回归方式会造成部分位置信息缺失，影响密度图质量，同时网络获取的多尺度特征利用不佳，致使网络挖掘的人群特征不能更好地应对拍摄距离和视角多变场景下人群尺度变化。

基于上述分析，本研究提出了一种比例融合与多层规模感知的人群计数网络，具体工作为：1）设计了多层规模感知模块，通过多列化的视野扩张编码多层特征信息，提高网络提取不同尺度语义信息的能力，相较于文献 [9,12-13]，该模块避免了多尺度模块的重复使用，在缩减参数的同时保证多尺度特征的获取效率；2）提出了比例融合策略，通过预测特征层权重指导多尺度特征集成，获得人群的显著性特征，相较于文献 [10-12]，该策略以自适应关注核心尺度的方式提升网络对多视角人群的预测性能，更好地应对视角、尺度多变的人群计数场景；3）为提高网络对局部人群数量的敏感程度，提出了一个新型的局部一致性损失函数用于网络的训练。本研究试验主要在 Shanghai Tech、UCF_CC_50 和 UCF-QNRF 等主流人群数据集上完成，试验结果表明，本研究方法较近年先进方法相比人群密度计数准确度较高，具有一定的优越性。

1　比例融合与多层规模感知网络

本研究网络结构如图 1 所示，主要由骨干特征提取模块、多规模感知模块、比例融合模块、密度回归卷积模块 4 个部分构成。

1.1　骨干特征提取模块

VGG16[14] 网络采用小卷积核堆叠串联方式[15]，在增加网络深度提升学习能力的同时拥有适中的参数量，网络计算代价较小，是人群密度估计与计数骨干网络的较优选择。基于此，本研究利用 VGG16 网络的前 4 个卷积层和 3 个池化层提取人群初始特征 F_i，i=1,2,3,4，分别对应 4 个卷积层的输出。将 F_1–F_3 定义为浅层初始特征，F_4 定义为深层初始特征。

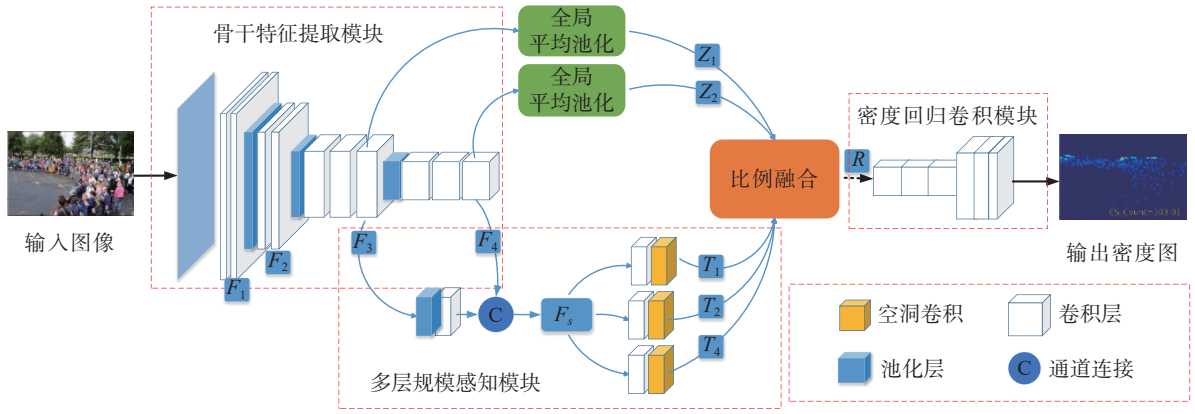


图 1　网络整体结构

Fig. 1　Overall structure of the network

1.2　多层规模感知模块

现有多尺度人群计数网络多忽略浅层初始特征，仅利用深层初始特征用于密度预测，造成浅层信息丢失，降低了网络对人群密集区域的感知能力。基于此，本研究设计了多层规模感知模块，通过上下文信息融合的方式增强人群密集区域信息，提高网络对密集人群的检测能力。

$$F_s = \Phi(F_3)\mathrm{U}F_4 \tag{1}$$

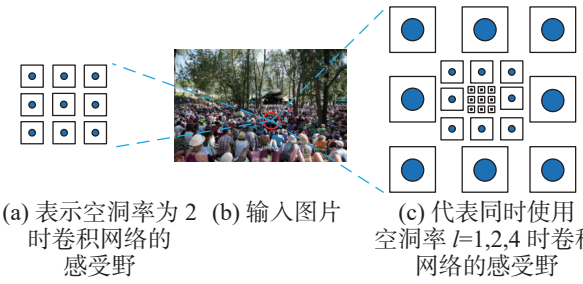
式中：$\Phi(\cdot)$ 包含 2×2 最大池化以及 1×1 卷积操作；U 代表通道连接。

首先，采用式 (1) 将骨干特征提取模块获得的浅层初始特征 F_3、深层初始特征 F_4 编码到同一维度空间，再通过通道连接获得高维人群初始特征 (F_s)，丰富人群密集区域的特征描述。

$$T_l = \xi_l(\mathrm{Cov}(F_s)) \tag{2}$$

式中：$\mathrm{Cov}(\cdot)$ 表示 1×1 卷积操作；$\xi_l(\cdot)$ 代表空洞率为 l 的空洞卷积。本研究选用 3 列空洞卷积表达尺度特征，空洞率组合通过试验方式探索得出。人群主流数据集 Shanghai Tech[4] 的试验结果表明，空洞率 (l) 依次设置为 l=1,2,4 的 3 通道组合可获得最佳多尺度特征。

然后，设计多列空洞卷积[16] 结构，采用式 (2) 将 F_s 映射到多尺度空间，通过感受野适当的多列化提升多尺度人群特征捕获的丰富性，得到人群多尺度描述 (T_l)。多列化卷积不同空洞率下的感受野变化效果如图 2 所示。



(a) 表示空洞率为 2 时卷积网络的感受野　(b) 输入图片　(c) 代表同时使用空洞率 l=1,2,4 时卷积网络的感受野

图 2　不同空洞率下的感受野变化

Fig. 2　Changes of receptive field under different atrous rates

1.3　比例融合模块

多尺度特征信息的综合表达充分与否是影响人群密度估计准确性的又一关键。现有方法多采用通道叠加方式完成多个尺度特征的融合，网络平等地对待每一尺度特征。但是，对于具有拍摄视角多变特点的人群图像来说，不同图像尺度信息分布不均，各尺度特征的重要程度不同[17]，简单通道叠加造成多尺度特征综合表达效果较差。

$$Z_j = \frac{1}{C_{j+2}\times H_{j+2}\times W_{j+2}}\sum_{k=1}^{C_{j+2}}\sum_{m=1}^{H_{j+2}}\sum_{n=1}^{W_{j+2}} f_k(m,n) \tag{3}$$

$$R^* = \left(\frac{Z_1}{Z_1+Z_2}\varphi(T_1,T_2)\right)T_2 + \left(\frac{Z_2}{Z_1+Z_2}\varphi(T_1,T_2)\right)T_1 \tag{4}$$

$$R = \left(\frac{Z_1}{Z_1+Z_2}\varphi(R^*,T_4)\right)R^* + \left(\frac{Z_2}{Z_1+Z_2}\varphi(R^*,T_4)\right)T_4 \tag{5}$$

式中：当 j=1 时，C_{j+2}、H_{j+2}、W_{j+2} 代表初始特征 F_3 的通道数、高度以及宽度；j=2 时，C_{j+2}、H_{j+2}、W_{j+2} 代表初始特征 F_4 的通道数、高度以及宽度；$f_k(m,n)$ 表示初始特征 F_i 第 k 个特征通道中，第 (m,n) 位置处的特征值；$\varphi(\cdot)$ 为比较函数，用于判别 T_1、T_2、T_4 的共性信息，将 2 个大小相等的特征矩阵逐元素进行比较，如果同位置的 2 个张量相同则返回 1，否则返回 0。

本研究设计了图 3 的比例融合结构，通过自适应关注核心尺度的方式提升网络显著性人群特征 (R) 的描述能力。首先，根据初始特征 F_3、F_4，采用式 (3) 获得融合比例系数 Z_j，j=1,2；然后，利用 Z_j 预测多尺度特征 T_1、T_2、T_4 的重要程度，采用式 (4)、式 (5) 实现尺度特征的分阶段融合，获得显著性人群特征 R。

1.4　密度回归模块

密度回归模块利用提取人群显著性特征 R 进

行密度图回归，回归网络设计为 6 层 3×3 卷积网络串联，通道数依次设置为 512、512、512、256、128、64。受特征提取阶段池化层的影响，用于回归的特征分辨率降低，预测的密度图分辨率为原始图像的 1/8。

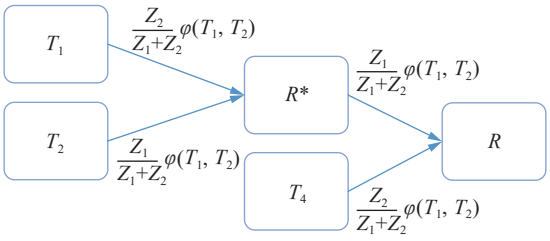


图 3　比例融合结构

Fig. 3　Proportional fusion structure

$$D_n=\sum_{i=1}^{K}\delta(R_n-R_{ni})\otimes G_{\sigma(R_{ni})} \tag{6}$$

式中：函数$\delta(R_n-R_{ni})$表示人头标记点图像 R_n 中第 i个坐标为R_{ni}的人头标记点的密度平滑区域，区域大小与积分为 1 的自适应高斯滤波器$G\sigma(R_{ni})$一致；K 为图像中人头标记点总数；⊗表示卷积运算。滤波器大小$\sigma(R_{ni})=\beta\bar{d}_i$，$\bar{d}_i$表示标记点$R_{ni}$与其最近的 Z 个人头之间的平均距离。

假设共有 N 幅大小各为 $H_n\times W_n$ 的训练图像 P_n，对应人头点真值标记图为 R_n，n=1, 2, …, N，采用如式 (6) 所示密度图生成方法[6] 生成密度图 D_n。

2　损失函数

人群计数任务广泛采用损失函数：

$$L=\frac{1}{N}\sum_{n=1}^{N}\left(\sum_{h=1}^{H_n}\sum_{w=1}^{W_n}|D'_n(h,w)-D_n(h,w)|\right)^2 \tag{7}$$

式中：$D'_n(h,w)$、$D_n(h,w)$ 分别表示第 n 幅预测密度图 D'_n、密度真值图 D_n 中坐标为 (h,w) 像素的人群密度值。

由于图像透视现象以及拍摄视角多样等因素的影响，在密度图生成过程中，部分人头标注R_{ni}采用的滤波器$\sigma(R_{ni})$大小并不准确，使得这部分人头标注生成的密度图在描述范围上存在一定程度的误差。

基于此，本研究提出一种局部一致性损失函数L'用于网络的训练，以缓解这一问题。具体公式为

$$L'=\frac{1}{N}\sum_{n=1}^{N}\sum_{s=0}^{s=f_n}\left|\sum_{h=s\times Q}^{(s+1)\times Q}\sum_{w=s\times Q}^{(s+1)\times Q}D'_{n(s)}(h,w)-\sum_{h=s\times Q}^{(s+1)\times Q}\sum_{w=s\times Q}^{(s+1)\times Q}D_{n(s)}(h,w)\right| \tag{8}$$

$$f_n=\max\left(\frac{H_n}{Q},\frac{W_n}{Q}\right) \tag{9}$$

式中：$D'_{n(s)}(h,w)$、$D_{n(s)}(h,w)$分别表示第 n 幅训练图像的第 s 个图像块中坐标为 (h,w) 像素的密度预测值和真值；max(·) 表示取 2 个数的最大整数。

先将 N 幅 $H_n\times W_n$ 大小的输入图像按照 $Q\times Q$ 的大小分割成 $m\times n$ 个小块，再将每一个小区域的总体计数值来表示对应区域的总人数。网络通过学习 $Q\times Q$ 区域内的人群特征表述，弱化高斯滤波平滑范围不准确带来的密度图错误信息的影响，使得局部计数图对应区域的人群表述更接近真实密度图，不仅有利于网络的训练而且提高了网络的计数性能。通过试验方式测得 Q=64 时效果最佳。

3　试验及分析

3.1　硬件环境及参数设置

本研究试验硬件配置为 4 卡 GPU RTX2080Ti，64G RAM，算法采用 Ubuntu 系统基于 Ptorch 框架实现，试验运行环境为 CUDA10.0+anaconda3+Python3.7+Pytorch1.2。使用 VGG16 的预训练模型来初始化骨干特征提取网络，剩余的卷积层均采用标准偏差为 0.01 的随机高斯初始化。模型训练时使用 Stochastic Gradient Descent (SGD) 优化器优化模型参数，学习率设置为 10^{-7}，训练总数为 400 次。

3.2　数据集与评价指标

本研究选择人群密度估计与计数领域常用数据集 Shanghai Tech[4]、UCF_CC_50[18] 和 UCF-QNRF[19] 开展算法对比试验，并利用 VisDrone-2019[20] 目标检测数据集中的车辆图像对本研究方法的拓展应用可能性进行了评估。为与对比方法保持一致，选用平均绝对误差 (MAE) 和均方误差 (MSE) 作为评价指标[21]：

$$E_{\mathrm{MA}}=\frac{1}{N_t}\sum_{\tau=1}^{N_t}\left|T_\tau-T_\tau^{\mathrm{GT}}\right| \tag{10}$$

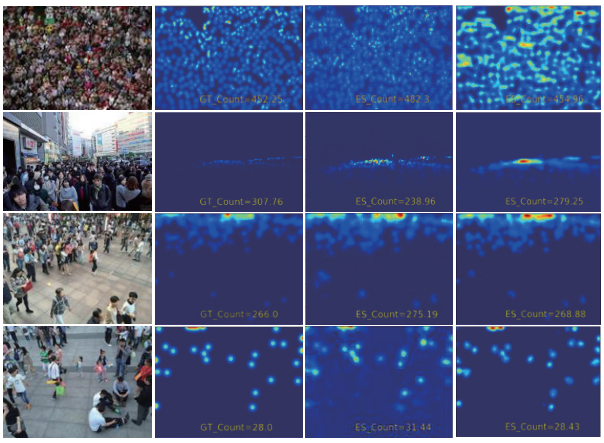
$$E_{\mathrm{MS}}=\sqrt{\frac{1}{N_t}\sum_{\tau=1}^{N_t}\left|T_\tau-T_\tau^{\mathrm{GT}}\right|^2} \tag{11}$$

式中：N_t 表示用于测试的图像总数；T_τ、T_τ^{GT}分别表示第τ幅图像的预测密度图和真值密度图包含的人员总数。

3.3　Shanghai Tech 数据集试验及试验结果分析

Shanghai Tech 数据集由 2 部分组成：Part A 部分和 Part B 部分，将 330 165 个人头注释标记到 1 198 幅图片中。Part A 部分从因特网上收集得到，人群高度拥挤的场景较多，拍摄视角多样，由 300 幅训练图像和 182 幅测试图像组成；Part B 部

分为购物街道下同一视角拍摄的图像，人群分布较为稀疏，包含训练图像 400 幅，测试图像 316 幅。在该数据集上，本研究方法与 CSRNet[9] 在单幅图像上的密度预估对比如图 4 所示，多方法对比试验结果如表 1 所示。



(a) 输入图像　(b) 手工标注　(c) CSRNet (d) 本研究方法

图 4　Shanghai Tech 数据集试验结果

Fig. 4　Experimental results of Shanghai Tech dataset

表 1　Shanghai Tech 数据集多方法性能对比试验结果

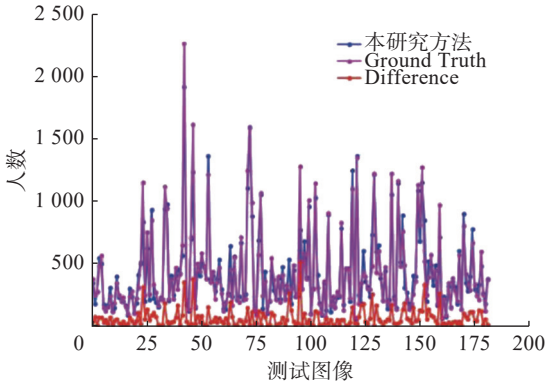
Table 1　Experimental results of multi-method performance comparison on Shanghai Tech dataset

方法	Part A		Part B	
	MAE	MSE	MAE	MSE
MCNN[4]	110.20	173.20	26.40	41.30
Switch-CNN[5]	90.40	135.00	21.60	33.40
ACSCP[7]	75.70	102.70	17.20	27.40
FF-CAM[22]	71.00	109.80	10.30	15.80
SPN[10]	70.00	106.30	9.10	14.60
MRM[13]	69.52	110.23	8.96	13.51
CSRNet[9]	68.20	115.00	10.60	16.00
DSPNet[23]	68.20	107.80	8.90	14.00
MFEM[24]	67.00	112.60	9.90	14.60
TriangleNet[25]	66.40	113.80	11.30	18.30
TEDNet[11]	64.20	109.10	8.20	12.80
CAN[12]	62.30	100.00	**7.80**	**12.20**
本研究方法0	**61.75**	**97.46**	8.30	13.60

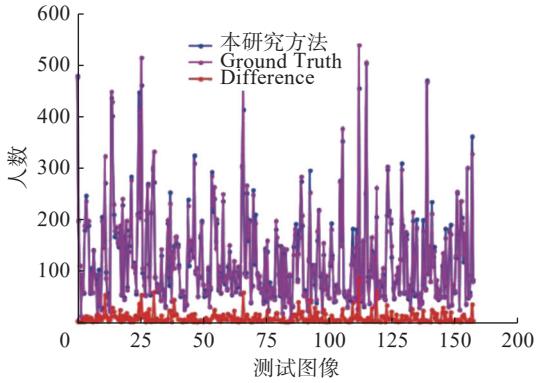
图 4(a) 为 Shanghai Tech 数据集 Test 部分的图像，前两幅来自 Part A 部分，后两幅来自 Part B 部分。可以看出，本研究方法在 Part A、Part B 部分均得到较为优秀的人群计数准确度以及密度图预测质量。在人群较为密集的 Part A 部分，生成的密度图真值表征有提升空间，本研究方法预测的密度图能更真实地反映人群密度的分布情况。

表 1 试验结果表明，在 Part A 部分，本研究方法达到了最优的计数性能，相较于 CSRNet，MAE 降低了 6.45，MSE 降低了 17.54，说明本研究方法具有良好的多尺度特征丰富以及缓解视角变化的能力。在单一视角的 Part B 部分，本研究方法计数准确性优于大部分已有算法，但较 TEDNet[11] 以及 CAN[12] 方法略显不足，这是由于该方法更多关注于上下文信息的获取对人群密度造成的影响，故在单视角下效果更优，而本研究设计的多层感知与比例融合模块主要关注多视角场景下信息的融合，在多视角的自适应能力优于单视角场景。

为了进一步验证算法性能，图 5 给出数据集所有图像测试误差结果。Difference 代表真实人数与预测人数之间的误差。从图 5 中可以看出，本研究方法在整个数据集上单幅图像误差值均保持较低水平，说明本研究方法鲁棒性较高。



(a) Shanghai Tech Part A 数据集



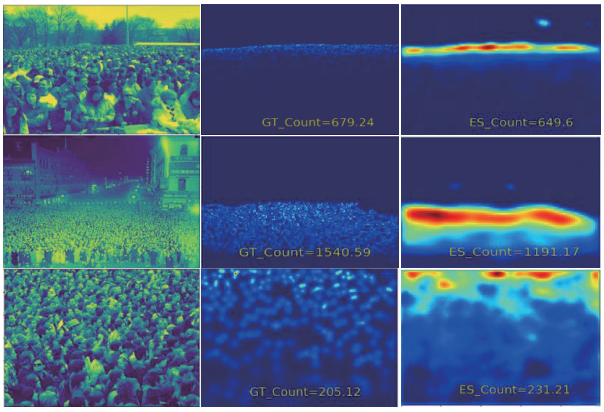
(b) Shanghai Tech Part B 数据集

图 5　Shanghai Tech 数据集单幅图像误差结果示意

Fig. 5　Schematic diagram of error results of single image of Shanghai Tech dataset

3.4　UCF_CC_50 数据集试验及试验结果分析

UCF_CC_50 数据集仅有 50 幅图像，但单幅图像上最多包含 4633 个人头标注，平均每幅图像包含 1 279 个人头标注。在该数据集上密度预测的准确性可以充分体现网络对高密度人群图像估计的性能。测试数据集中单幅图像的密度预估结果如图 6 所示，多方法性能对比如表 2 所示，数据集所有图像测试误差结果如图 7 所示。



(a) 输入图像　(b) 手工标注　(c) 本研究算法

图 6　UCF_CC_50 数据集试验结果

Fig. 6　Experimental results of UCF_CC_50 dataset

表 2　UCF_CC_50 和 UCF-QNRF 多方法性能对比试验结果

Table 2　Experiment results of multi-method performance comparison on UCF_ CC_ 50 and UCF-QNRF

方法	UCF_CC_50		UCF-QNRF	
	MAE	MSE	MAE	MSE
MCNN[4]	377.60	509.1	277.0	426.0
Switch-CNN[5]	318.10	439.2	228.0	445.0
ACSCP[7]	291.00	404.6	—	—
CSRNet[9]	266.10	397.5	—	—
TEDNet[11]	249.40	354.5	113.0	188.0
FF-CAM[22]	246.80	322.2	114.5	200.5
DSPNet[23]	243.30	307.6	107.5	182.7
TriangleNet[24]	240.70	357.0	—	—
MFEM[25]	226.50	305.6	107.3	181.8
CAN[12]	212.20	243.7	107.0	183.0
SPN[10]	204.70	340.4	110.3	184.6
本研究方法	**203.07**	**241.0**	**104.6**	**182.6**

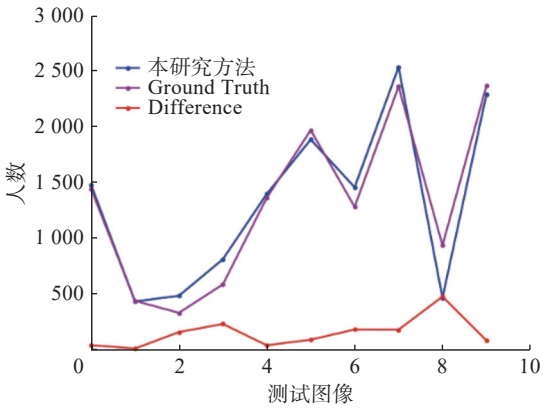


图 7　UCF_CC_50 数据集单幅图像误差结果示意

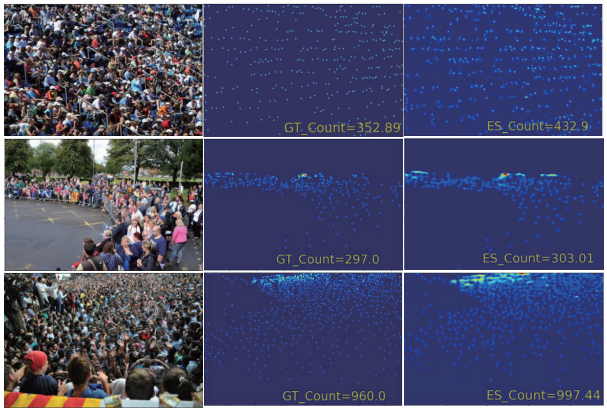
Fig. 7　Schematic diagram of error results of single image of UCF_CC_50 dataset

为与对比方法保持一致，UCF_CC_50 数据集试验采用 5 折交叉验证[2]，表 2 试验结果为 5 次测试结果的平均值。试验数据表明，在训练样本较少的情况下，本研究方法在计数的准确性和鲁棒性依然占优，性能较佳；同时，本研究方法单幅图像误差值均较低，可以有效应对密集场景。

3.5　UCF-QNRF 数据集试验及试验结果分析

UCF-QNRF 数据集于 2018 年提出，具有清晰度高、样本尺寸差异大和人群密集程度高等特点，是人群计数领域中最具挑战的数据集，由 1 201 幅训练图像和 334 幅测试图像组成，单幅图像最多包含 12865 个人头标注，平均每幅图像包含 815 个人头标注。

为优化网络训练，训练图像分辨率统一调整至 1 024×1 024，测试时使用原始尺寸测试图像。部分单幅图像密度预测结果如图 8 所示，多算法对比试验结果详见表 2。可以看出，本研究方法在 MAE 和 MSE 上均占优，可以有效对应稀疏和密集场景，且在同一幅图像中人群分布变化较大场景中也取得较好的预测结果。



(a) 输入图像　(b) 手工标注　(c) 本研究算法

图 8　UCF-QNRF 数据集试验结果

Fig. 8　Experimental results of UCF-QNRF dataset

由于 UCF-QNRF 数据集图像较多，将单幅图像预测人数与真实人数误差结果以图 9 所示密度等级柱状图方式进行对比描述，同一密级水平下柱状图的差值为预测误差值。将测试数据集划分为 8 个密度等级，相邻密度等级之间相差 500 人。试验结果显示，本研究方法前 5 个密度等级的柱状图相距较小，误差较小，预测精度优于密度等级 6~8，在超高密集场景下的人群计数性能仍有待进一步提高。

3.6　VisDrone2019 数据集试验和试验结果分析

人群与车辆计数均为复杂背景、视角多变场景下的目标计数任务，同时存在密集遮挡的情况。为进一步探索本研究方法对应于复杂背景、视角多变场景计数的能力，本研究也采用车辆数据集进行了一定的试验。

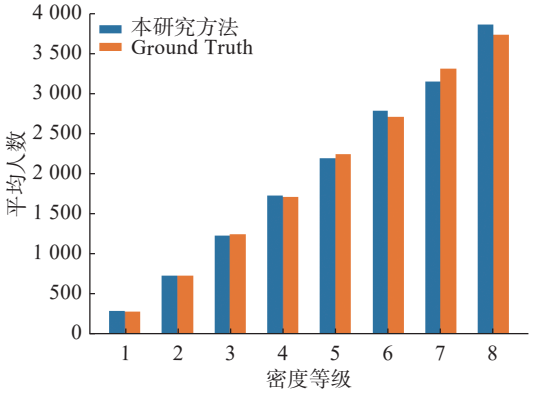
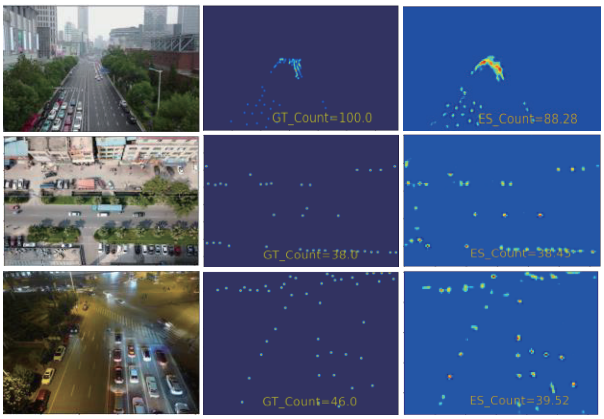


图 9　UCF-QNRF 数据集单幅图像误差结果示意

Fig. 9　Schematic diagram of error results of single image of UCF-QNRF dataset

VisDrone2019 数据集由多种无人机平台在各种天气和光照条件下采集，数据集包含车辆等多种物体，拍摄环境和背景复杂，并且同时拥有稀疏和密集的场景。本研究选择该数据集进行车辆密度估计与计数试验，验证本研究方法的可推广性。

选择 VisDrone2019 数据集中含车辆图像 2646 幅，采用随机划分方式，1 853 幅图像用于训练，793 幅图像用于测试。图 10、表 3 所示的车辆密度与计数预测结果表明，本研究方法计数精度较高，可以较好地应对多种拍摄环境、拍摄视角以及车辆复杂分布场景。



(a) 输入图像　(b) 手工标注　(c) 本研究算法

图 10　VisDrone2019 数据集试验结果

Fig. 10　Experimental results of VisDrone2019 dataset

表 3　在 VisDrone2019 数据集多方法性能对比试验结果

Table 3　Experiment results of multi-method performance comparison on VisDrone2019

方法	MAE	MSE
MCNN[4]	14.9	21.6
CSRNet[9]	9.8	14.6
本研究方法	**8.8**	**13.6**

单幅图像预测误差结果如图 11 所示，测试数据集被划分为 5 个密度等级，相邻密度等级之间相差 50 人。由试验结果可以看出，5 个密度等级下的预测人数和真实人数柱状图均相差较小，说明本研究方法即使在密度等级为 5 的情况下仍然得到准确的计数结果，证明本研究方法可以有效推广至车辆计数方面。

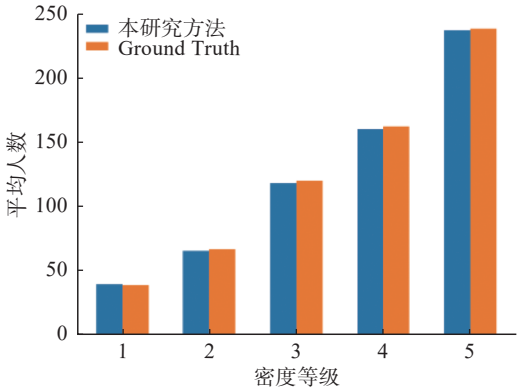


图 11　VisDrone2019 数据集单幅图像误差结果示意

Fig. 11　Schematic diagram of error results of single image of VisDrone2019 dataset

3.7　消融研究

网络整体框架包含多个模块，为验证所提模块的有效性，采用与 3.1 小节描述一致的试验硬件环境，在 Shanghai Tech 数据集下对各个模块进行消融研究。

多层规模感知模块的多种空洞卷组合消融试验结果如表 4 所示，表中空洞卷积的参数表示为“Conv-(卷积核大小)-(过滤器数量)-(空洞率)”。可以看出，Net B 结构最为合理，因此，本研究多层感知模块最终选择 Net B 结构，前序试验结果均在采用该结构开展的试验中获得。

表 4　多层规模感知模块消融试验

Table 4　Ablation Experiment of multi-layer scale sensing module

类型	Net A		Net B		Net C	
结构	Conv 3-512-1 Conv 3-512-2 Conv 3-512-3		Conv 3-512-1 Conv 3-512-2 Conv 3-512-4		Conv 3-512-2 Conv 3-512-3 Conv 3-512-4	
数据集	Part A	Part B	Part A	Part B	Part A	Part B
MAE	62.62	8.47	61.75	8.30	62.57	8.38
MSE	101.66	13.78	97.46	13.60	97.76	13.65

计数网络框架与损失函数消融试验结果如表 5 所示。结果表明，Model B 在仅改变损失函数的情况下，MAE 和 MSE 较 Model A 下降了 2.05 和 5.74，通过区域化密度图的方式增强生成密度图与真实密度图的相似度，提高计数性能；Model C 较 Model B 引入多规模感知模块，通过上下文信息融合的方式增强人群密集区域信息，降低 MAE 和 MSE；Model D 较 Model C 加入

比例融合模块，以自适应关注核心尺度的方式提升了网络对多视角人群的预测性能，从而使得 MAE 下降了 2.85、MSE 下降了 8.06，同时 Model D 网络达到了最优的计数性能。综上，本研究所提比例融合与多层规模感知的人群计数网络，分别从多尺度特征获取不足、融合不佳和全局特征利用不充分等方面有效提升密集场景下人群计数精确度。

表 5　计数网络与损失函数消融试验

Table 5　Counting network and loss function ablation experiment

网络	Model A		Model B		Model C		Model D	
结构	VGG16 密度回归 MSE LOSS — —		VGG16 密度回归 局部一致性损失 — —		VGG16 多规模感知 密度回归 局部一致性损失 —		VGG16 多规模感知 比例融合 密度回归 局部一致性损失	
参数量	16263489		16263489		25048897		25048897	
数据集	Part A	Part B	Part A	Part B	Part A	Part B	Part A	Part B
MAE	69.70	10.6	67.65	8.96	64.65	8.55	61.75	8.3
MSE	116.0	16.0	110.26	13.5	105.52	13.8	97.46	13.6

4　结束语

本研究提出一种比例融合与多层规模感知的人群计数网络降低拍摄视角多变带来的影响，并通过新型的局部一致性损失函数提高网络对局部人群数量的敏感程度，提升人群密度估计与计数的准确性。试验结果表明，本研究方法在 Shanghai Tech 的 Part A 部分，相较于 CSRNet，MAE 降低了 6.45，MSE 降低了 17.54，同时在 UCF_CC _50 和 UCF-QNRF 数据集上表现出较好的计数性能，在 VisDrone2019 车辆数据集上 MAE 达到 8.8，MSE 为 13.6，证明了本研究方法的可推广性。

综上所述，本研究方法在拍摄视角和距离多变场景中的计数能力有较大提升，但人群极度密集下的严重遮挡仍是影响人群准确计数的关键问题，这也是后续工作的重点。

参考文献：

[1] CHEN Ke, LOY C C, GONG Shaogang, et al. Feature mining for localised crowd counting[C]//Proceedings of the British Machine Vision Conference 2012. Surrey. British Machine Vision Association, 2012: 1−11.

[2] 向飞宇, 张秀伟. 基于卷积神经网络的人群计数算法研究 [J]. 计算机技术与发展, 2021, 31(7): 42–46.
XIANG Feiyu, ZHANG Xiuwei. Research on crowd counting algorithm based on convolution neural network[J]. Computer technology and development, 2021, 31(7): 42–46.

[3] SOURTZINOS P, VELASTIN S A, JARA M, et al. People counting in videos by fusing temporal cues from spatial context-aware convolutional neural networks[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 655−667.

[4] ZHANG Yingying, ZHOU Desen, CHEN Siqin, et al. Single-image crowd counting via multi-column convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 589−597.

[5] 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. Honolulu: IEEE, 2017: 4031−4039.

[6] 孟月波, 纪拓, 刘光辉, 等. 编码−解码多尺度卷积神经网络人群计数方法 [J]. 西安交通大学学报, 2020, 54(5): 149–157.
MENG Yuebo, JI Tuo, LIU Guanghui, et al. Encoding-decoding multi-scale convolutional neural network for crowd counting[J]. Journal of Xi'an Jiaotong university, 2020, 54(5): 149–157.

[7] SHEN Zan, XU Yi, NI Bingbing, et al. Crowd counting via adversarial cross-scale consistency pursuit[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5245−5254.

[8] JIANG Xiaoheng, ZHANG Li, XU Mingliang, et al. Attention scaling for crowd counting[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 4705−4714.

[9] LI Yuhong, ZHANG Xiaofan, CHEN Deming. 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: IEEE, 2018: 1091−1100.

[10] XU Chenfeng, QIU Kai, FU Jianlong, et al. Learn to scale:

generating multipolar normalized density maps for crowd counting[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2020: 8381−8389.

[11] JIANG Xiaolong, XIAO Zehao, ZHANG Baochang, et al. Crowd counting and density estimation by trellis encoder-decoder networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 6126−6135.

[12] LIU Weizhe, SALZMANN M, FUA P. Context-aware crowd counting[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 5094−5103.

[13] LIU Xiyang, YANG Jie, DING Wenrui, et al. Adaptive mixture regression network with local counting map for crowd counting[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 241−257.

[14] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. 3rd international conference on learning representations, ICLR 2015-conference track proceedings, 2015: 1−14.

[15] SINDAGI V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid CNNs[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 1879−1888.

[16] 刘万军, 佟畅, 曲海成. 空洞卷积与注意力融合的对抗式图像阴影去除算法 [J]. 智能系统学报, 2021, 16(6): 1081–1089.
LIU Wanjun, TONG Chang, QU Haicheng. An antagonistic image shadow removal algorithm based on dilated convolution and attention mechanism[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1081–1089.

[17] ZHANG Cong, LI Hongsheng, WANG Xiaogang, et al. Cross-scene crowd counting via deep convolutional neural networks[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 833−841.

[18] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013: 2547−2554.

[19] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 544−559.

[20] ZHU Pengfei, WEN Longyin, DU Dawei, et al. Vision Meets Drones: Past, Present and Future[EB/OL]. (2020−01−16)[2022−01−01]. https://arxiv.org/pdf/2001.06303.pdf.

[21] XIONG Feng, SHI Xingjian, YEUNG D Y. Spatiotemporal modeling for crowd counting in videos[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5161−5169.

[22] 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.
ZHANG Yuqian, LI Guohui, LEI Jun, et al. FF-CAM: crowd counting based on frontend-backend fusion through channel-attention mechanism[J]. Chinese journal of computers, 2021, 44(2): 304–317.

[23] ZENG Xin, WU Yunpeng, HU Shizhe, et al. DSPNet: deep scale purifier network for dense crowd counting[J]. Expert systems with applications, 2020, 141: 112977.

[24] 翁佳鑫, 仝明磊. 基于 Triangle Net 的密集人群计数 [J]. 科技创新与应用, 2021(9): 38–40,44.
WENG Jiaxin, TONG Minglei. Dense crowd counting based on Triangle Net[J]. Technology innovation and application, 2021(9): 38–40,44.

[25] ZHANG Jun, LIU Jiaze, WANG Zhizhong. Convolutional neural network for crowd counting on metro platforms[J]. Symmetry, 2021, 13(4): 703.

作者简介：

孟月波，教授，博士生导师，博士，主要研究方向为机器视觉信息处理与分析、建筑智能化。近年来主持/参与国家自然科学基金项目、国家重点研发计划项目、陕西省基础研究项目和陕西省重点研发项目 10 项。发表学术论文 30 余篇。 E-mail：mengyuebo@163.com。

张娅琳，硕士研究生，主要研究方向为计算机视觉理解、建筑智能化技术。E-mail：1243697118@qq.com。

王宙，硕士研究生，主要研究方向为深度学习、计算机视觉。E-mail：1119307454@qq.com。