山西大学学报(自然科学版)49(1):100—107,2026
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2024056

基于局部信息编码特征金字塔的轻量多类目标计数网络

魏祥一,张莉*
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

摘 要:针对现有目标计数算法抗背景杂波能力差以及对高度遮挡、尺度变化大的目标计数准确率低的问题,提出一种基于局部信息编码特征金字塔的多类目标计数网络模型。所提算法利用卷积神经网络中特征图的冗余性来搭建轻量级的骨干网络,同时添加带有局部信息编码机制的特征金字塔模块学习目标的局部特征,最后使用卷积层组成的回归头和分类头分别进行目标的数量预测和位置预测。为实现多类目标计数任务,将已有人群计数数据集ShanghaiTech和车辆计数数据集CARPK(Car Parking)的训练集混合并对之训练;为与已有方法进行对比,分别在这两类目标数据集的测试集上进行测试,并以平均绝对误差和均方误差作为计数评价指标。实验结果证明所提出算法能进行多类目标计数且在目标计数上的表现优于其他方法。

关键词:多类目标计数;局部特征编码;卷积神经网络;轻量级骨干网络
中图分类号:TP183 **文献标志码**:A **文章编号**:0253-2395(2026)01-0100-08

Lightweight Multi-class Target Counting Network Based on Feature Pyramid with Local Information Encoding

WEI Xiangyi, ZHANG Li*
(*School of Computer Science and Technology, Soochow University, Suzhou* 215006, *China*)

Abstract: Addressing the limitations of existing object counting algorithms, which struggle with background clutter and exhibit low accuracy when dealing with heavily occluded or significantly varying object scales, we propose a novel lightweight multi-class object counting network based on feature pyramid with local information encoding (FPLE-MOCN). This model leverages the redundancy of feature maps in convolutional neural networks to construct an efficient and rapid lightweight backbone network. Additionally, a feature pyramid module with a local information encoding mechanism is introduced to capture the local features of targets. Finally, regression and classification heads composed of convolutional layers are employed for predicting the number and the location of objects at the same time. To achieve multi-class object counting, we combine the training sets of the existing crowd counting dataset (ShanghaiTech) and the vehicle counting dataset (CARPK) for training. For comparison with existing methods, we evaluate our model on the test sets of both datasets separately and use both mean absolute error and mean squared error as evaluation metrics for counting. Experimental results demonstrate that FPLE-MOCN can perform multi-class object counting and outperforms other methods in terms of counting accuracy.

Key words: multi-class object counting; local feature coding; convolutional neural networks; lightweight backbone network

收稿日期:2024-02-20;**修回日期**:2024-03-27
基金项目:江苏省高校自然科学研究基金资助项目(19KJA550002)
作者简介:魏祥一(1999—),男,山东济宁人,硕士研究生,研究方向为计算机视觉。E-mail:20214227040@stu.suda.edu.cn
* **通信作者**:张莉(ZHANG Li),E-mail:zhangliml@suda.edu.cn
引文格式:魏祥一,张莉. 基于局部信息编码特征金字塔的轻量多类目标计数网络[J]. 山西大学学报(自然科学版),2026,49(1):100-107. DOI:10.13451/j.sxu.ns.2024056.

0 引言

目标计数是计算机视觉领域的一个重要下游任务,也是异常检测[1]、公共安全管理[2]、人群分析[3]、人群定位[4]等应用的基础任务,近年来因其广泛的应用场景和实用价值而受到研究人员的关注。目标计数的任务是估计场景中目标的数量,研究重点是在密集场景中的准确计数。然而现在大多数研究都是针对单类别的目标计数,如人群计数[5]、车辆计数[6]等,随着应用需求的增长,单类别目标计数已渐渐难以满足城市公共安全管理的需求,例如在城市交通的监控场景中,汽车和行人往往同时出现,单类别目标计数方法的作用在这种场景下比较有限。能够同时对多类目标计数的方法已成为迫切的需要。

随着深度学习的发展,在目标计数领域已经涌现出很多优秀的成果。目前,深度学习中主流的计数方法主要有基于检测的方法[7]、基于密度图的方法[5]和基于定位点的方法[8]。起初,Leibe等[7]使用匹配目标完整特征的方法进行检测,但在高度遮挡的场景下这种方法的表现不佳。因此,Li等[9]开始利用目标部分特征来进行匹配。但是基于检测的方法在面对背景难以区分、目标遮挡严重、密集复杂的场景时会出现漏检严重的现象,导致计数准确率大大降低。

基于密度图的方法通常会使用高斯核生成原始图像的密度图,再使用模型建立原始图像和密度图之间的映射,用预测图像中目标的密度的方式进行计数。Zhang等[5]首先提出在人群计数数据集上使用几何自适应核生成密度图进行人群计数,取得了优秀的表现。当前车辆计数任务也普遍使用基于密度图的方法[6]。基于密度图的算法面对密集和遮挡的场景依然能够准确计数,但实际上这是一种违背视觉直觉的方法,对目标的分布并不敏感,并且模型在学习过程中容易受到背景杂波的影响。

基于定位点的方法令模型直接学习标注点周围的特征,这就避免了模型受背景杂波的影响,巧妙地解决了背景多变、场景复杂、图像分辨率低造成模型计数性能下降的问题。Song等[8]搭建了一个基于定位点方法的人群计数框架,这是一种符合视觉直觉的计数方法,在计数和定位实验中取得了非常优越的成果。但是基于定位点的方法对硬件环境的需求较高,且当场景较稀疏或目标尺度变化大时这种算法的表现会受到影响。在车辆计数中,有些汽车目标仅在图像中显示出部分车轮或车尾,基于定位点的方法很难准确对这种目标进行计数。

针对以上描述的基于定位点的方法存在的问题,本研究提出了一种新的定位点方法,称为基于局部信息编码特征金字塔的多类目标计数网络模型。该方法利用卷积神经网络的冗余性,仅对部分特征图进行卷积操作和跳跃连接,从而搭建了一个轻量级骨干网络。我们在ImageNet数据集[10]上预训练,在保证模型计数性能的同时,极大减少了不规则的内存访问和参数量。另外,本研究还构建了局部信息编码特征金字塔模块,使模型不仅关注定位点周围的特征,还会对目标的局部特征进行编码,从而对高度遮挡的目标进行有效计数。该模块可以在一定程度上解决由目标尺度变化大、外形差异大、遮挡范围大导致的低准确率问题。本研究将提出的方法在ShanghaiTech[5]和CAR-PK[11]数据集的混合训练集上进行,并分别在它们的测试集上进行测试。实验结果验证了本研究方法能够对人群和车辆两种目标进行同时计数,并且均取得最优秀的表现。鉴于在多类目标计数中类间差异也是影响模型计数性能的因素,我们认为增加对上下文特征信息互补的利用或者使用注意力机制或许对提高模型的计数性能有帮助。

1 本文网络

图1展示了本文网络的整体架构,主要包含轻量级骨干网络、局部信息编码的特征金字塔模块、分类回归头模块三个部分。

1.1 轻量级骨干网络

基于定位点的目标计数方法在输入阶段需要对输入点进行一对一匹配,这个过程非常消耗内存。受到部分卷积神经网络[12]的启发,本研究设计了一个轻量级骨干网络。如图1所示,该骨干网络由四个阶段块组成,每个阶段均包含轻量级卷积块。除了第一个阶段块外,每个阶段前有一个融合层。融合层是普通的卷

积层,卷积核大小为2×2,步长为2,用于空间下采样和通道数扩展。第一个阶段块前使用嵌入层,嵌入层卷积核大小为4×4,步长为4的卷积层。最后两个阶段块会消耗较少的内存访问,往往有更高的每秒浮点运算次数。因而,我们应该在前两个阶段块中使用少量的轻量级卷积块,在后两个阶段块中分配更多的轻量级卷积块。具体地,在本研究网络中阶段块一的轻量级卷积块数量为2,阶段块二的轻量级卷积块数量为3,阶段块三和阶段块四的轻量级卷积块数量均为4。输入特征图在经过四个阶段块后,将进行一次全局池化操作和一个1×1卷积层,相比于平均池化和最大池化,全局池化可以有效减少模型的参数量,并降低过拟合的风险。

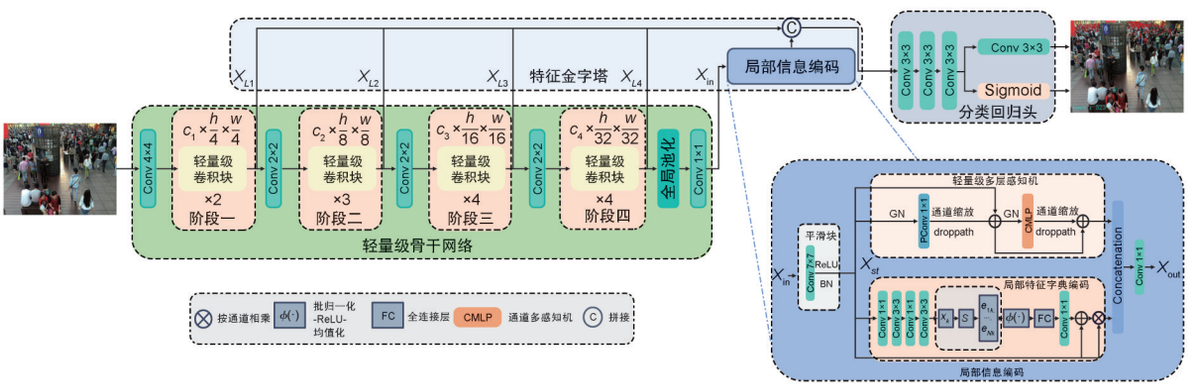


图1 本文网络的整体架构

Fig. 1 The overall architecture of the network in this paper

如前所述,轻量级卷积块是轻量级骨干网络基本组成单位,其结构展示在图2中,其中“Conv”表示卷积层,“PConv”表示部分卷积层,“3×3”表示卷积核尺寸为3×3,“ReLU”表示线性整流函数(Linear Rectification Function, ReLU)[13],“BN”表示批归一化(Batch Normalization, BN),“⊕”表示加和操作,即按通道对应元素相加。值得注意的是,部分卷积层只对输入特征图的一部分进行一般卷积操作,而将其他未经卷积操作的特征图按通道与经卷积操作后的特征图拼接到一起。当进行卷积操作的特征图的占比为1/4时,部分卷积层的参数量仅为一般卷积层的1/16。在本文中,部分卷积层进行卷积操作的特征图数均为输入通道数的1/4。

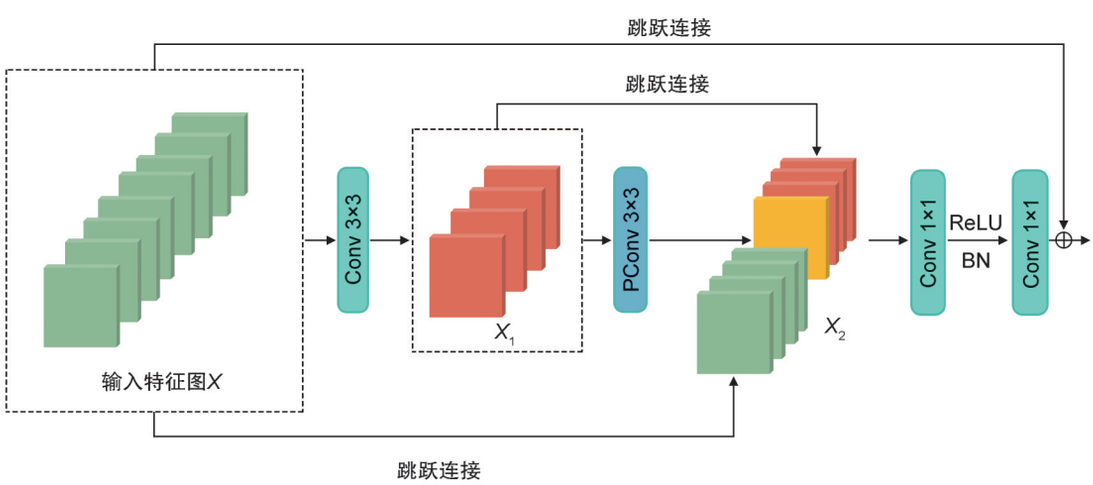


图2 轻量级卷积块结构示意

Fig. 2 Schematic of lightweight convolutional block structure

在轻量级卷积块中,令输入特征图为X,其维度表示为$[C,H,W]$,即C为特征图的通道数,H和W分别为特征图的高和宽。首先,特征图为X经过一个卷积核尺寸为3×3的卷积

层,将通道数压缩为原通道数的一半,得到特征图X_1。这一步的目的是对特征图进行加深并去噪,使得特征图质量更高。随后,对特征图进行部分卷积操作,会对通道总数1/4的特征图进行普通卷积运算。经过部分卷积之后,得到形状为$[C/8,H,W]$的特征图,然后与X_1中未经部分卷积的特征图和输入特征图X中一半的特征图进行跳跃连接,得到形状与原始特征图一致的特征图X_2。此后,X_2将经过两个倒置的1×1卷积层组成的残差块,在两个1×1卷积层间进行ReLU激活函数和批归一化处理。

对于高性能神经网络来说,归一化和激活层是不可或缺的。然而,特征在经过归一化和激活函数后会损失一部分多样性。本研究在轻量级卷积块中只将激活层和归一化放在每个块中间倒置的1×1卷积层之中,以保持特征多样性并实现较低的延迟。批归一化的好处是可以通过结构重参数化的方式合并到相邻的卷积层中,便于快速推理。由图1可以观察到,与Chen等[12]设计的部分卷积神经网络块的中间特征图相比,特征图X_2携带了多层次的信息,并且拥有更深的总深度和平均深度。这意味着特征图X_2的质量更高,能够让模型取得更佳的性能,同时模型的参数量并未增加。

基于密度图的方法学习的是原始图像到密度图之间的映射。在复杂背景中,目标与背景之间的边界可能模糊不清,使得从图像中准确提取目标特征变得困难。这可能导致密度图估计不准确,从而影响目标计数的准确性。本文方法在数据输入阶段需要对样本和标注点进行一对一匹配,而且模型学习的是标注点周围的特征。当模型的感兴趣目标为多个类别时,模型同样可以对其进行无差别的学习,从而能在模型输出阶段对预测图像中识别出的所有感兴趣目标进行估计和计数。相比于基于密度图的方法,本文方法会在很大程度上减少对背景特征的学习,从而大大减少背景杂波带来的影响。

1.2 局部信息编码特征金字塔和分类回归头

本研究设计了局部信息编码特征金字塔模块,能使得模型关注目标的局部特征,并能充分利用骨干网络产生的多尺度特征图,以提高模型对目标尺度变化的适应能力。局部信息编码特征金字塔模块由两部分组成,一是特征金字塔部分,二是局部信息编码部分。

轻量级骨干网络对输入图像提取出四级特征$X_{Li}(i=0,1,2,3)$,其空间尺寸分别为输入图像的1/4,1/8,1/16和1/32。特征金字塔部分的作用是通过在不同的图像尺度上提取特征,以捕捉不同大小和分辨率的目标信息。

如图1所示,局部信息编码部分的输入特征图为轻量级骨干网络的输出,即X_{in}。局部信息编码部分首先将X_{in}由携带ReLU激活函数和批归一化的通道为256的7×7卷积层组成的平滑块得到X_{st},随后X_{st}进入并行的两个阶段,这两个阶段分别是轻量级多层感知机和局部特征字典编码。轻量级多层感知机阶段由两个残差模块组成:部分卷积模块和通道MLP(Multilayer Perceptron,MLP)模块,其中通道MLP模块的输入是部分卷积模块的输出和X_{st}的加和。这两个模块都经过了通道缩放和DropPath操作以提高特征的泛化性和鲁棒性。

局部特征字典编码的灵感来源于人脸识别中的字典算法[14-15],是一个具有固有字典的编码器,由一个固有的码本和一组可学习的视觉中心比例因子组成。固有码本可表示为$b_1,b_2,\cdots,b_K$,其中$K=H\times W$是输入特征的总维度;可学习的视觉中心缩放因子表示为$S=\{s_1,s_2,\cdots,s_K\}$。具体来说,经过平滑块的特征X_{st}首先被一组由1×1卷积、3×3卷积、1×1卷积组成的卷积层组合编码,然后对经过编码的特征进行平滑块处理,该块由带有批归一化层的3×3卷积和ReLU激活函数组成。经过以上步骤,将编码后的特征输入码本中。对于整幅图像的第k个码,使用缩放因子s_K和码本b_K来使编码后的特征映射到相应的位置信息。将两个阶段的输出进行拼接,将拼接的特征进行1×1卷积下采样到256的通道大小,完成局部信息编码的操作。经过局部信息编码的操作,模型能够对目标的局部信息建立感知,从而增加模型对目标整体和角落区域的特征捕捉能力。因而,当模型感兴趣目标在图像中呈现出高度遮挡时,仍然可以通过识别被遮挡目标的

局部特征进行准确计数。另外,特征金字塔部分对图像不同尺度特征进行聚合,从而提高了模型应对目标尺度变化的鲁棒性,减少多类目标计数中目标的较大尺度变化为模型计数性能带来的影响。

最终局部信息编码的结果X_{out}与从骨干网络提取的四级特征进行拼接,得到局部信息编码特征金字塔的输出,并送入分类回归头。分类回归头由三层3×3的卷积层以及两个并行的3×3卷积层和Sigmoid函数组成,回归头的作用是预测目标数量,分类头的作用是预测目标坐标。

1.3 损失函数

本研究对计数任务设计了损失函数。由于本研究是面向多类别目标的计数模型,所以计数任务的损失函数考虑了类别不平衡的因素,提出一个可以动态调整类损失权重的类别自适应权值调整损失函数,其表达式如下:

$$L_{\text{cout}}=\frac{1}{2N}\sum_{j=1}^{M}\sum_{i=1}^{N}g_{ij}\left\|\hat{D}_j(X_i)-D_{ij}\right\|_2^2, \tag{1}$$

其中L_{cout}是计数损失函数,M是类别数量,N是测试样本的数量,$\hat{D}_j(X_i)$表示模型对第i张图像第j类的预测计数值,D_{ij}表示第i张图像第j类的真实计数值,g_{ij}是第i个训练样本中类别j的权重,与绝对百分比误差有关,定义为:

$$g_{ij}=\left(1-\frac{1}{\log_2\left(\frac{|\hat{T}_{ij}-T_{ij}|}{T_{ij}+10^{-6}}\right)}\right)^{\gamma}, \tag{2}$$

其中$\widehat{T_{ij}}$和T_{ij}分别表示第i个训练样本中类别j的预测值和实际计数真实值,γ的作用是调节对损失贡献低的样本在总损失中的权重。根据文献[16],训练过程中经验性地设置$\gamma=1.5$。

2 实验结果

2.1 实验设置

实验环境:操作系统为Ubuntu18.04,显卡配置为NVIDIA GeForce RTX 3090,显存为24 GiB;软件平台为Python3.7,CUDA 10.2,Pytorch 1.10.2。

本研究在ShanghaiTech人群计数数据集和车辆计数数据集CARPK(Car Parking)上进行实验。ShanghaiTech数据集分为PartA和PartB两部分,Part A包含了482张图像,人群非常密集。Part B包含了716张图像,人群分布相对稀疏。CARPK是一个车辆检测和计数数据集,包含无人机航拍的989张训练图像和459张测试图像,共有89 777个汽车注释实例。本研究算法在ShanghaiTech和CARPK的混合训练集上进行训练,分别在这两个数据集的测试集上进行测试,以验证本研究算法的多类目标计数性能。

本研究采用随机裁剪和翻转的方式进行数据增强。具体操作为首先将样本随机裁剪为原尺寸的9/16,然后对裁剪后的图像块作随机垂直和水平的翻转。学习率初始设置为0.001,对所有实验模型使用适应性矩估计(Adaptive Moment Estimation, Adam)算法[17]进行优化。

常用的目标计数性能指标有平均绝对误差(Mean Absolute Error, MAE)和均方误差(Mean Squared Error, MSE),具体表达式如下:

$$E_{\text{MAE}_j}=\frac{1}{N}\Sigma_{i=1}^{N}|\hat{T}_{ij}-T_{ij}|, \tag{3}$$

$$E_{\text{MSE}_j}=\sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(\hat{T}_{ij}-T_{ij})^2}。 \tag{4}$$

2.2 对比实验结果及分析

为了评估所提算法在人群计数任务上的有效性,在ShanghaiTech数据集上与目前主流的人群计数算法进行对比,包括multi-column convolutional neural network (MCNN)[5], congested scene recognition network (CSRNet)[18], point to point network (P2PNet)[8], weakly-supervised crowd counting with transformers (TransCrowd)[19], distribution matching for crowd counting (DM-Count)[20]和segmentation guided attention network (SGANet)[21]。人群计数的对比实验结果如表1所示。

由表1可知:在ShanghaiTech PartA数据集上,本研究所提出的算法取得最优的MAE和MSE,比次优方法P2PNet分别提升了1.06和0.06;在ShanghaiTech PartB数据集上,本研究算法取得了最好的MAE指标,比次优方法P2PNet提高0.02,并在MSE指标上取得了次优表现。综上,本研究所提出的算法在ShanghaiTech数据集的四项对比中取得三个最优表现,

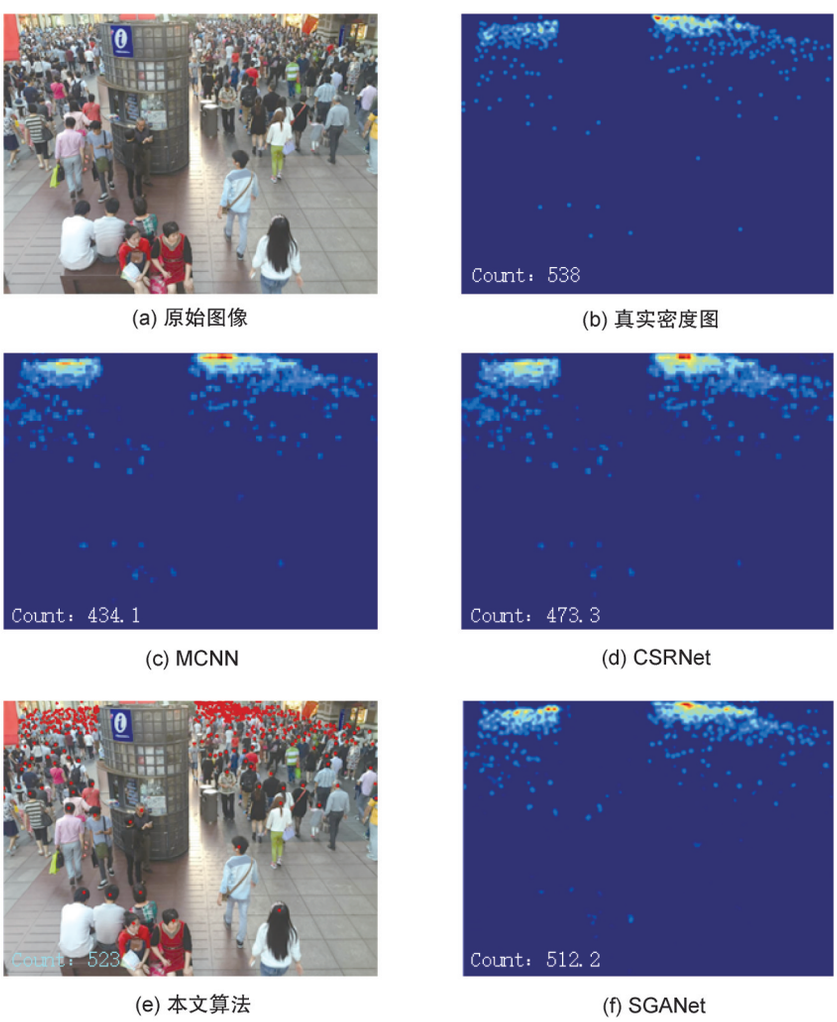
表1 在ShanghaiTech数据集上的对比结果

Table 1 Comparison results on the ShanghaiTech dataset

方法	ShanghaiTech PartA		ShanghaiTech PartB	
	MAE	MSE	MAE	MSE
MCNN	110.20	173.20	26.40	41.30
CSRNet	68.20	115.00	10.60	16.00
P2PNet	52.74	85.06	6.25	**9.90**
TransCrowd	66.10	105.10	9.30	16.10
DM-Count	59.70	95.70	7.40	11.80
SGANet	57.60	101.10	6.60	10.20
本研究算法	**51.66**	**85.00**	**6.23**	10.13

注:加粗表示最优结果。

一个次优表现。实验结果验证了本研究所提方法在人群计数任务上的先进性和有效性。图3展示了在ShanghaiTech Part B数据集上某张图像的计数情况,其中图3(a)是原图,图3(b)是真实密度图和真实计数结果,图3(c)是MCNN的预测密度图和计数结果,图3(d)是CSRNet预测密度图及计数结果,图3(f)是SGANet预测密度图及计数结果,图3(e)是本文算法的预测及计数结果(红色点为预测点)。由图3可见,本研究算法的预测结果很接近真值。



(a) 原始图像 (b) 真实密度图 (c) MCNN (d) CSRNet (e) 本文算法 (f) SGANet

图3 人群计数结果对比

Fig. 3 Comparison of crowd counting results and generated density maps

为了验证本文算法的车辆计数性能,在CARPK数据集上与基于密度图的计数方法MCNN[5]、compact convolutional neural network(C-CNN)[22]和基于检测的方法RetinaNet[16]、faster region-based convolutional neural network(Faster R-CNN)[23]进行对比,结果如表2所示。

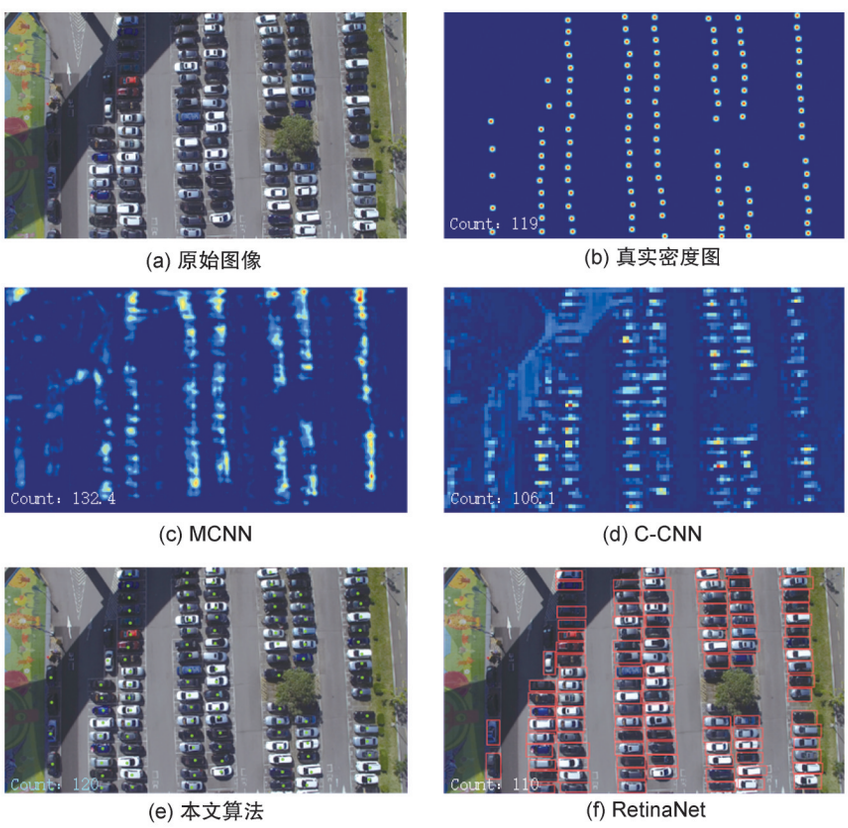
由表2可知,本研究算法在CARPK数据集上均获得了最优的MAE、MSE,比次优的MCNN分别提升了9.69的MAE和10.89的

表2 在CARPK数据集上的对比结果

Table 2 Comparison results on the CARPK dataset

方法	CARPK	
	MAE	MSE
MCNN	15.70	18.68
C-CNN	34.50	37.64
RetinaNet	16.62	22.30
Faster R-CNN	24.32	37.62
本研究算法	**6.01**	**7.79**

MSE。这充分证明了本研究算法在车辆计数上具有先进的性能。图4展示了在CARPK数据集上某张图像的计数情况,其中图4(a)—图4(f)分别是原图、真实密度图及计数结果、MCNN的预测密度图及计数结果、C-CNN的预测密度图及计数结果、本文算法的预测结果(绿色点为预测点)、RetinaNet预测结果(红色框为预测候选框)。



(a) 原始图像 (b) 真实密度图

(c) MCNN (d) C-CNN

(e) 本文算法 (f) RetinaNet

图4 车辆计数结果对比

Fig. 4 Comparison of vehicle counting results and generated density maps

3 结论

本研究提出一种能在复杂背景、高度遮挡场景下准确对人群和车辆多类目标计数的卷积神经网络算法。该算法的特点在于针对卷积神经网络特征图冗余的问题设计了参数量更少但性能更优越的骨干网络和增强模型处理多尺度信息和高度遮挡场景计数能力的局部信息编码特征金字塔模块。在ShanghaiTech人群计数数据集和CARPK车辆计数数据集上,本算法与主流算法的对比实验结果显示本算法仅在一个指标上取得次优结果,其他结果均为最优表现。在多类目标计数任务中,类间差异也是影响计数准确率的因素,后续研究将寻求在模型中添加注意力机制或充分利用上下文特征、构建更优秀的网络结构以及改进卷积算子来进一步提高模型的计数性能。

参考文献:

[1] MANJU D, RADHA V. A Survey on Human Activity Prediction Techniques[J]. *Int J Adv Technol Eng Explor*, 2018, **5**(47): 400-406. DOI: 10.19101/ijatee.2018.547006.

[2] KARPAGAVALLI P, RAMPRASAD A V. Estimating the Density of the People and Counting the Number of People in a Crowd Environment for Human Safety[C]//

2013 International Conference on Communication and Signal Processing. New York: IEEE, 2013: 663−667. DOI: 10.1109/iccsp.2013.6577138.

[3] WAN J, KUMAR N S, CHAN A B. Fine-grained Crowd Counting[J]. *IEEE Trans Image Process*, 2021, **30**: 2114−2126. DOI: 10.1109/TIP.2021.3049938.

[4] WANG Q, GAO J Y, LIN W, *et al*. NWPU-crowd: A Large-scale Benchmark for Crowd Counting and Localization[J]. *IEEE Trans Pattern Anal Mach Intell*, 2021, **43**(6): 2141−2149. DOI: 10.1109/TPAMI.2020.3013269.

[5] 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). New York: IEEE, 2016: 589−597. DOI: 10.1109/CVPR.2016.70.

[6] ZHANG S H, WU G H, COSTEIRA J P, *et al*. FCN-RLSTM: Deep Spatio-temporal Neural Networks for Vehicle Counting in City Cameras[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 3687−3696. DOI: 10.1109/ICCV.2017.396.

[7] LEIBE B, SEEMANN E, SCHIELE B. Pedestrian Detection in Crowded Scenes[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). New York: IEEE, 2005: 878−885. DOI: 10.1109/CVPR.2005.272.

[8] SONG Q Y, WANG C G, JIANG Z K, *et al*. Rethinking Counting and Localization in Crowds: A Purely Point-based Framework[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). New York: IEEE, 2021: 3345−3354. DOI: 10.1109/ICCV48922.2021.00335.

[9] LI M, ZHANG Z X, HUANG K Q, *et al*. Estimating the Number of People in Crowded Scenes by MID Based Foreground Segmentation and Head-shoulder Detection [C]//2008 19th International Conference on Pattern Recognition. New York: IEEE, 2008: 1−4. DOI: 10.1109/ICPR.2008.4761705.

[10] DENG J, DONG W, SOCHER R, *et al*. ImageNet: A Large-scale Hierarchical Image Database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 248−255. DOI: 10.1109/CVPR.2009.5206848.

[11] HSIEH M R, LIN Y L, HSU W H. Drone-based Object Counting by Spatially Regularized Regional Proposal Network[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 4165−4173. DOI: 10.1109/ICCV.2017.446.

[12] CHEN J R, KAO S H, HE H, *et al*. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 12021−12031. DOI: 10.1109/CVPR52729.2023.01157.

[13] NAIR V, HINTON G E. Rectified Linear Units Improve Restricted Boltzmann Machines[C]//Proceedings of the 27th international conference on machine learning (ICML−10). Madison: Omnipress, 2010: 807−814. https://icml.cc/Conferences/2010/papers/432.pdf.

[14] LV S L, LIANG J Z, DI L, *et al*. A Probabilistic Collaborative Dictionary Learning-based Approach for Face Recognition[J]. *IET Image Process*, 2021, **15**(4): 868−884. DOI: 10.1049/ipr2.12068.

[15] CHEN S, LAI X, YAN Y, *et al*. Learning an Attention-Aware Parallel Sharing Network for Facial Attribute Recognition [J]. *J Vis Commun Image Represent*, 2023, **90**: 103745. DOI: 10.1016/j.jvcir.2022.103745.

[16] LIN T Y, GOYAL P, GIRSHICK R, *et al*. Focal Loss for Dense Object Detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). New York: IEEE, 2017: 2999−3007. DOI: 10.1109/ICCV.2017.324.

[17] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[C]//2014 International Conference on Learning Representations. New York: Curran Associates. DOI: 10.48550/arXiv.1412.6980.

[18] 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. New York: IEEE, 2018: 1091−1100. DOI: 10.1109/CVPR.2018.00120.

[19] 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. DOI: 10.1007/s11432-021-3445-y.

[20] WANG B, LIU H, SAMARAS D, *et al*. Distribution Matching for Crowd Counting[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc., 2020, 33: 1595−1607. DOI: https://doi.org/10.48550/arXiv.2009.13077.

[21] WANG Q, BRECKON T P. Crowd Counting via Segmentation Guided Attention Networks and Curriculum Loss[J]. *IEEE Trans Intell Transp Syst*, 2022, **23**(9): 15233−15243. DOI: 10.1109/TITS.2021.3138896.

[22] SHI X W, LI X, WU C L, *et al*. A Real-time Deep Network for Crowd Counting[C]//ICASSP 2020−2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: IEEE, 2020: 2328−2332. DOI: 10.1109/ICASSP40776.2020.9053780.

[23] REN S Q, HE K M, GIRSHICK R, *et al*. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks[J]. *IEEE Trans Pattern Anal Mach Intell*, 2017, **39**(6): 1137−1149. DOI: 10.1109/TPAMI.2016.2577031.