山西大学学报(自然科学版)47(2):359－368,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023044

基于特征去噪的密集人群定位方法

温超[1,2*],贺宏强[1,2]
(1.山西大学 大数据科学与产业研究院,山西 太原 030006;
2.山西大学 计算机与信息技术学院,山西 太原 030006)

摘 要:密集人群定位分析的难点之一是获取图像中个体目标的准确预测。密集人群场景条件下,前景图像小目标物体和其他物体存在相互遮挡和干扰等造成的特征噪声。采用传统人群定位方法学习得到的人头特征容易遭受特征噪声的影响,进而容易造成人头特征判别性弱和边界信息获取不准确。针对上述问题,文章提出一种特征去噪方法用于人群定位,其利用语义特征解耦的思想抑制特征噪声以增强独立人头的检测。不同于旨在改善图像视觉质量的传统像素域去噪方法,所提方法将在特征空间对多尺度特征进行去噪,促使模型学习到更多目标类特征,并抑制干扰特征。通过前景目标特征与背景特征的语义解耦,分别增强和减弱头部特征和背景特征响应,有利于改善独立个体目标的检测。实验结果表明,所提出的方法在密集人群数据集Shanghai Tech、UCF-QNRF和NWPU-Crowd上的平均$F1$值分别为81.2%,72.4%和77.1%,提高了密集人群定位的性能。

关键词:多尺度;特征去噪;背景干扰抑制;独立个体检测;语义解耦
中图分类号:TP183 **文献标志码**:A **文章编号**:0253-2395(2024)02-0359-10

Dense Crowd Localization Method Based on Feature Denoising

WEN Chao[1,2*], HE Hongqiang[1,2]
(1. *The Institute of Big Data Science and Industry, Shanxi University, Taiyuan* 030006*, China;*
2. *School of Computer and Information Technology, Shanxi University, Taiyuan* 030006*, China*)

Abstract: The difficulty of dense crowd location analysis is to obtain accurate prediction of individual targets in the image. In dense crowd scenes, small target objects and other objects in the foreground image have characteristic noises caused by mutual occlusion and interference. The human head features learned by traditional crowd location methods are easily affected by feature noise, which may lead to weak discrimination of human head features and inaccurate boundary information acquisition. To solve the above problems, a class aware feature denoising method is proposed for crowd location, which uses the idea of semantic feature decoupling to suppress feature noise and thus to enhance the detection of independent heads. Different from the traditional pixel domain denoising methods aiming at improving image visual quality, the proposed method will denoise multi-scale features in the feature space, promote the model to learn more about target features, and suppress interference features. Through the semantic decoupling of foreground target features and background features, the response of head features and background features is enhanced and weakened, respectively and the detection performance of independent individual targets can be improved. Experimental results show that the average $F1$ values of the proposed method on Shanghai Tech, UCF-QNRF and NWPU crowd dense population data sets are 81.2%, 72.4% and 77.1% respectively, which shows that it improves the performance of dense population localization.

Key words: multi-scale; feature denoising; background interference suppression; independent individual detection; semantic decoupling

收稿日期:2023-01-03;**接受日期**:2023-03-10
基金项目:国家自然科学基金(62106134);山西省留学回国人员科技活动择优资助项目(20220002);
* **通信作者**:温超(1987-),男,山西太原人,博士,副教授,研究方向为智能信号处理。E-mail:cwen@sxu.edu.cn
引文格式:温超,贺宏强.基于特征去噪的密集人群定位方法[J].山西大学学报(自然科学版),2024,47(2):359-368.
DOI:10.13451/j.sxu.ns.2023044

0 引言

人群定位是人群分析中具有挑战性的任务之一。与人群计数[1-2]不同,人群定位旨在判断区域是否存在目标个体,并提供个体目标的边界信息。人群定位以图像中实例级特征而非人群计数的图像级特征作为处理单元,能够检测场景中每个实例级个体目标而非仅仅获取图像场景人数。人群定位性能很大程度上影响了图像高层语义分析的效果,在人群异常检测,人群行为检测,安全监测等方面具有重要应用价值[3]。

得益于目标检测模型的快速发展,现有研究方法利用目标检测模型[4-5]在人群定位上取得了巨大进展,如TinyFaces[6],LSC-CNN[3]对模型提取的特征直接进行人头目标预测。但经典目标检测模型在密集人群场景下容易遭受其他物体特征干扰导致个体头部区域特征边界模糊。为应对密集人群场景,基于点监督的方法利用点标签或生成伪盒标签来训练定位器[7]。由于个体目标尺度信息和标签高度相关,因而这类方法很可能因缺乏头部尺度信息而受到干扰特征与标签伪相关的影响。同基于图像特征直接进行人群定位的目标检测方法相比,启发式定位方法通过对特征图进一步处理生成中间图,如密度图,置信图等,并利用中间图信息进行高斯回归[8-9],实例分割[10]等进行定位。得益于高分辨率像素级特征,Gao等[11]利用阈值网络对置信图编码为用于实例分割二值图进行人群定位。基于密度图焦点反距离变换的方法[12]利用个体目标间距的非线性函数在空间域分离目标区域。然而基于上述方法的密集人群图像特征图中存在大量来自相邻物体目标类内特征干扰与复杂背景的类间特征干扰,造成个体目标特征弱化以及边界模糊。这些背景与其他前景目标噪声干扰导致独立个体目标检测性能下降,成为提升人群定位性能的瓶颈之一。

为抑制背景与其他前景目标噪声干扰带来的影响,本文提出一种面向密集人群定位的类别感知特征去噪方法。典型的图像噪声有两类:(1)像素域噪声,即对象不可知,一般表现为随机特征;(2)特征域噪声,指来自相邻物体和背景的特征干扰。第二种噪声在密集人群场景图像中是无处不在并且较为显著。虽然目前关于图像去噪任务的研究很多,但很少有针对人群定位的特征去噪研究。值得注意的是,特征去噪用于下游语义任务的特征映射,而不是用于原始图像的视觉增强。本文利用特征去噪模块将目标和背景特征分别解耦到各自的通道,在基于去噪特征的置信图预测基础上,利用自适应阈值图改善置信图的区域级预测,并通过二值图预测网络输出用于个体目标检测的二值图,最后利用连通成分检测实现基于二值图的人群定位。本文工作的主要贡献在于:

(1)提出一种类别感知的特征去噪方法,利用语义分割提供的类别感知信息,引导去噪网络学习目标类和背景类的解耦特征,解决类间与类内特征干扰抑制的问题。

(2)将个体目标检测转化为二值图预测问题,在多尺度图像特征图使用特征去噪来解决个体目标特征弱化以及边界模糊的问题,引导自适应阈值网络对目标类和背景类特征进行差异化学习,实现独立个体目标的准确预测。

(3)公开数据集上的实验表明,基于特征去噪的人群定位方法能有效提升个体目标预测的$F1$指标。

1 相关工作

卷积神经网络(Convolutional Neural Network, CNN)在特征提取方面表现出了优势。大多人群分析关注图像级人群计数和区域密度估计,利用基于CNN特征提取器生成密度图并估计人数[13-17]。然而,人群计数和区域密度估计缺少对独立个体目标的预测。

1.1 人群定位

人群定位旨在准确预测个体目标,现有方法大致可以分为基于检测的方法,点监督的方法和基于启发式的方法。基于检测的方法主要在经典的目标检测模型基础上增强了对小尺度目标的检测性能。Hu等[6]提出了一种微小目标检测框架,评估了像素分辨率、目标尺度和场景背景特征对定位性能的影响。由于人脸特征具有更好的判别性,Li等[18]通过平衡采样获得更加均匀的多尺度人脸分布,并利用双金字

塔锚充分捕获上下文信息更好的定位个体目标。然而,密集场景条件下人脸细节特征难以获取,并且人头作为小尺度目标特征易受相邻目标与背景特征干扰,导致上述经典目标检测器难以在复杂密集人群场景取得令人满意的效果。

基于点监督的方法使用点级标签,有利于克服人工标注盒级注释的成本高和盒级数据集稀缺的问题。Abousamra 等[10]通过像素级扩展点级标签生成伪盒级标签,并引入拓扑约束来避免边界模糊和伪目标。Wang 等[19]提出一种用于精炼伪标签的顺序感知迭代机制,并利用优化的伪标签预测目标及其尺度。然而,基于点标签的方法难以准确刻画头部尺度信息,容易导致出现目标特征边界模糊的问题。

启发式方法利用与人头区域相关的信息增强人头特征的判别性,改善目标定位性能,具有更好的解释性。Lian 等[5]利用深度估计获取人头区域的边界信息,学习目标多尺度特征。Idrees 等[9]利用计数、密度图估计和定位任务的相关性,引导密度图的“锐化”并估计二值定位图。大多基于置信图分割的启发式方法利用固定阈值检测独立连通的成分,但这种固定阈值难以在复杂密集场景下检测独立的相邻目标。为此,Gao 等[11]利用自适应阈值网络学习更加精细的阈值,以改善独立目标分割以及检测性能。Liang 等[12]利用基于目标间距的焦点反距离变换函数在空间域分离目标区域特征。

上述方法关注密集人群特征图的后处理,缺少对特征空间存在的类间特征耦合和类内特征边界模糊的处理,因而独立个体目标的分割与检测容易遭受特征噪声的干扰。

1.2 图像特征去噪

深度学习在图像去噪方面得到了研究者的广泛关注。传统图像噪声大致包括加性白噪声、真实噪声和混合噪声[20]三种类型,均属于像素域噪声。传统图像去噪主要通过学习鲁棒特征抑制某种分布的像素域噪声实现原始图像的视觉增强效果。Guo 等[21]利用额外的噪声估计网络产生噪声图以提升主干网络去噪的鲁棒性。Zhang 等[22]利用通道依赖关系实现特征注意力,通过增强关键特征实现图像域噪声抑制。

近年来,越来越多研究关注图像去噪如何影响高层语义任务的性能。Xie 等[23]提出了一种特征去噪方法抑制像素域扰动,证明了像素域噪声可以转化为特征图扰动,并通过特征注意力提升图像质量和模型预测的鲁棒性。目前图像去噪方法通过模拟不同类型的微分滤波器,如非局部均值、双边滤波、均值滤波器和中值滤波器,对输入特征进行去噪处理,去噪过程可以表示为:

$$\mathcal{Y}=\mathcal{F}(\mathcal{X})+\mathcal{X}, \tag{1}$$

其中,$\mathcal{X},\mathcal{Y}\in\mathbb{R}^{C\times H\times W}$分别表示输入特征图和输出特征图,$\mathcal{F}(\cdot)$表示用于去噪的滤波器。由公式(1)可以看出,现有特征去噪方法关注特征图扰动和分布变化对原始图像视觉质量的影响,缺少对目标与背景特征解耦机制的研究,因而密集场景人群目标特征噪声难以从根本上抑制。

传统像素域去噪方法通过学习图像域合成噪声分布实现高分辨图像去噪。然而,合成噪声分布难以准确刻画当前图像的真实噪声特征。与传统像素域去噪方法不同,Liu 等[24]证明了语义信息对于高分辨图像去噪的重要性。受此工作启发,本文从基于多尺度空间位置的目标类特征出发,通过解耦目标类与背景类特征,抑制特征图噪声对目标区域特征干扰,实现基于特征去噪的独立个体检测。

2 基于类别感知特征去噪的密集人群定位方法

2.1 方法概述

为解决密集人群场景图像特征噪声干扰的问题,本文提出一种基于类别感知特征去噪的人群定位方法,总体架构如图 1 所示。它主要由四个模块组成:(1)利用 CNN 进行特征提取,可以从现有的特征提取器(如 VGG[25],HRNet[26])中采用不同形式的 CNN,用于生成多尺度图像特征;(2)特征去噪模块,利用增大感受野与类别感知去噪权重学习实现多尺度特征去噪,生成用于个体目标分割检测的置信图;(3)自适应阈值和二值化模块,由阈值图学习与人头检测两个部分构成,用于生成独立个

体目标的二值图。

具体而言,首先利用特征提取器获取不同尺度的图像特征,各尺度图像特征通道数均为64。在不改变特征通道维度的情况下利用特征去噪模块对各尺度特征进行去噪,将目标类和背景类特征在语义空间解耦,利用学习到的类别感知权重对目标类和背景类特征分别增强与减弱。利用上采样统一去噪后的特征图的尺度并进行通道级联拼接,然后利用1×1卷积和两层转置卷积通道降维生成单通道的置信图。自适应阈值模块利用卷积网络和置信图对拼接特征图进行重加权,生成具有空间区域注意力的单通道阈值图。二值化模块利用阈值图和置信图的大小关系生成用于预测独立个体目标的二值图。最终根据二值图通过连通成分检测生成独立个体目标方框。

2.2 密集人群图像特征域噪声

特征域噪声通常指目标之间的相互干扰和来自背景的干扰,且特征噪声对人群独立个体目标预测的负面效应较为明显,如图2所示,因此更需要在特征图而非原始输入图像上进行去噪。特征噪声对人群独立个体目标预测的负面影响主要涉及两个方面:

1)密集人群图像容易遭受目标类与背景类之间特征耦合导致的个体目标特征边界模糊的问题(参见图2的第一行)。

2)被背景包围的目标特征图响应不够突出(参见图2的第二行)。

图2展示了特征去噪之前和之后的图像特征图。从特征去噪前后特征图的变化可以看出未使用去噪方法的目标特征边界不清晰,目标与背景特征耦合在一起,而且部分目标特征响应弱。经去噪处理后目标特征边界较为清晰,与背景特征的耦合被抑制,目标特征响应增强较为明显。

2.3 特征去噪

针对密集人群场景图像特征去噪问题,本文设计了类别感知特征去噪方法。所提去噪模块可以与其他模块一起以端到端的方式学习,并对独立个体目标检测任务进行优化。为了阐述类别感知在特征图噪声抑制的重要性,将基于类别感知的特征去噪与基于自注意力的特征去噪进行比较。

为了消除特征域噪声,传统自注意力机制,如通道注意力[27]和空间注意力[28],通过重加权响应图,突出重要特征部分并抑制非目标信息的成分。基于传统自注意力特征去噪的重加权输出可以简化为以下一般形式:

$$\mathcal{Y}=\mathcal{A}(\mathcal{X})\odot\mathcal{X}=\bigcup_{c=1}^{C}W_c^{Sa}\odot(w_c^{Ca}X_c),\qquad(2)$$

其中$\mathcal{X},\mathcal{Y}\in\mathbb{R}^{C\times H\times W}$分别表示输入特征图和输出特征图,注意力函数$\mathcal{A}(\mathcal{X})$为某个注意力模块输出,$\odot$表示逐点乘积,$W_c^{Sa}$和$w_c^{Ca}$分别表示关于第$c$通道的空间权重和信道权重,$\bigcup\cdot$表示沿特征通道连接张量的级联操作。基于公式(2)的去噪方法仅在空间域中区分目标和背景之间的特征响应。换句话说,基于传统自注意力的特征去噪方法未在语义层面考虑类内人头目标

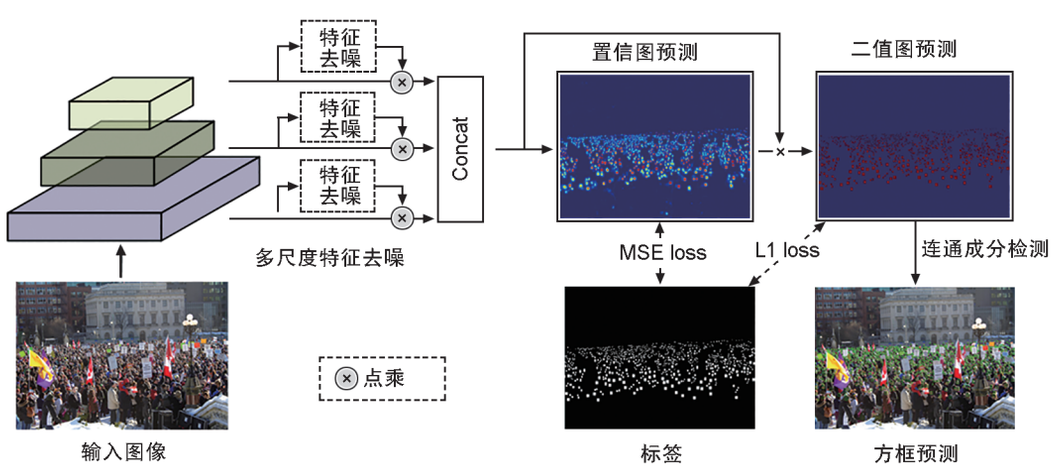
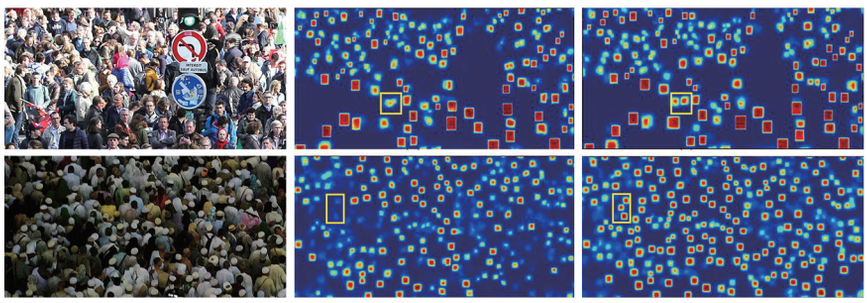


图1 模型框架图

Fig. 1 Framework of model



(a) 输入图像 (b) 未使用去噪的特征图 (c) 使用去噪后的特征图

图2 特征去噪示意图

第一行:个体目标特征边界模糊;第二行:目标特征响应弱

Fig. 2 Schematic diagram of feature denoising

The first line: individual object feature boundary blur, the second line: weak object feature response

特征干扰和类间物体特征耦合。

与基于传统自注意力的去噪模块不同,本文目标是设计一种新的特征去噪网络模块。该特征去噪模块利用目标区域的语义信息学习空间和通道两个维度的类别感知特征去噪权重,促使目标类和背景类特征的差异化学习。通过将不同类别的特征解耦到各自的通道中,增强独立个体目标在空间域的特征响应。基于类别感知特征去噪的重加权输出可以表示为如下公式:

$$\mathcal{Y}=\mathcal{A}_{FD}(\mathcal{X})\odot\mathcal{X}=\bigcup_{k}^{K}\mathcal{A}_{FD}(\mathcal{X})^{k}\odot\mathcal{X}^{k}=\bigcup_{k=1}^{K}\bigcup_{c=1}^{C_k}\tilde{W}_c^k\odot X_c^k, \tag{3}$$

其中,$\mathcal{A}_{FD}(\mathcal{X})^{k}\odot\mathcal{X}^{k}$表示第$k$类语义的部分特征,本文将目标和背景物体各归为一类,因此总类别数$K=2$。$\tilde{W}_c^k$表示第k类语义的类别感知权重。由公式(3)可以看出,注意力函数$\mathcal{A}_{FD}(\mathcal{X})$可以表示为第$k$类语义的形式,重加权输出特征可以分解为两个相互连接的部分:

$$\mathcal{Y}=\underbrace{\bigcup_{c=1}^{C_1}\tilde{W}_c^1\odot X_c^1}_{\text{目标类}}\cup\underbrace{\bigcup_{c=1}^{C_2}\tilde{W}_c^2\odot X_c^2}_{\text{背景类}}, \tag{4}$$

其中,C_1和C_2分别表示关于前景目标和背景特征的通道数,目标类的权重$\tilde{W}_c^1$可以被视为特定类图像语义分割的结果。可以看出,通过特征学习将前景目标与背景特征分别解耦到不同空间实现类别感知特征去噪。

2.4 基于特征去噪的人群定位

为克服基于自注意力的方法[27-28]难以抑制类内人头目标特征干扰和类间物体特征耦合的问题。在语义分割任务中,前景目标和背景的特征响应在空间上被分别增强和抑制,如图3所示。因此,采用语义分割网络以端到端的监督方式学习特征去噪权重,其特点是比基于自注意力的方法更加具有类别感知特性。为解决类别感知特征去噪问题,该端到端特征去噪模块在结构上主要包括两个部分:

(1)扩张卷积。为增强去噪网络上下文语义信息的感知能力,特定尺度特征图首先通过若干扩张卷积层,扩张卷积通过调整扩张率改变感受野大小。为避免扩张率设置不当会造成的“栅格化”问题[29],故采用“锯齿状”扩张率,实验设置膨胀率为1,2,5的三层扩张卷积层对特征图进行处理。

(2)类别感知去噪权重学习。通过两个1×1卷积层分支并行地处理上下文语义增强的特征图,以获得两个平行输出。支路一用于二分类语义分割预测,利用人头区域图像分割的特定二值图作为近似的真值进行监督学习;支路二输出去噪权重。两个支路的输出在语义上具有一致性,因此支路一对支路二的去噪权重学习具有引导作用。特征在语义空间进行了解耦,促使去噪权重对物体特征具有类别感知特性。

特征去噪网络结构如图3所示。其中膨胀卷积包含了三层扩张卷积,每层后利用ReLU函数进行非线性变换,在通道数不变的条件下,增加特征上下文语义信息的捕获。经扩张卷积处理的特征分别输入两个基于1×1卷积的支路:支路一用于语义分割预测,预测输出人体目标和背景物体分别看作两通道二值分割

预测图,采用二分类交叉熵损失进行监督学习;支路二用于输出64通道的特征去噪权重。最后利用去噪权重与特征图相点乘的方式,对物体特征的类别解耦得到特征去噪的特征图。

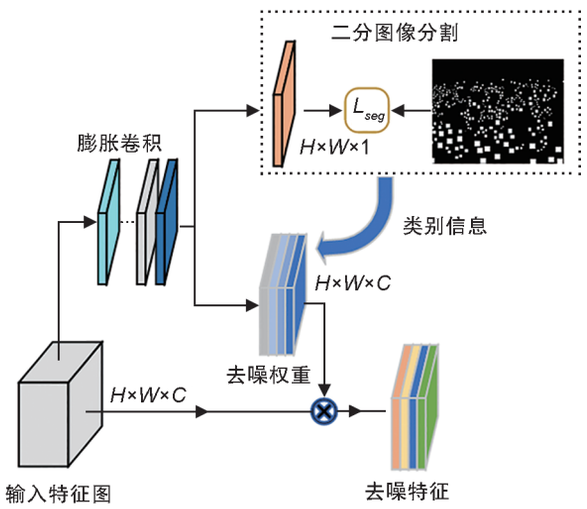


图3 特征去噪结构图

Fig. 3 Structural of feature denoising

如此,通过将目标类与背景类特征在类别空间解耦,实现在空间上增强目标区域特征响应和背景特征相应减弱,进而锐化人头目标特征边界。特征去噪模块只会少量增加网络推理阶段的运算复杂度,这促使其成为面向人群定位端到端网络的一种可插拔模块。

去噪增强后的特征有利于获取更加合理的目标置信图预测,如图2所示。在常规置信图预测基础上,采用如图1所示的二值图预测模块进一步处理,利用自适应阈值图学习重新评估置信图的区域级预测并生成新的阈值图T,进而利用二值图预测网络将置信图I编码为独立个体目标检测的二值图M,二值图编码规则如下:

$$M_{i,j}=\begin{cases}1, & I_{i,j}>T_{i,j}\\ 0, & \text{otherwise}\end{cases}, \tag{5}$$

其中,下标i,j表示二维索引。最后通过连通成分检测[30]和目标方框输出实现人群定位。

2.5 损失函数

模型在训练过程中针对主干网络目标置信图和二值图的预测任务以及特征去噪网络的语义分割任务设计了损失函数。对置信图预测任务,采用均方误差(MSE)回归损失L_{MSE},保留目标区域的预测正确的特征;在基于自适应阈值的二值图预测任务中,采用平均绝对误差(MAE)损失L_{MAE},对预测错误部分进行惩罚;在面向特征去噪的语义分割任务,采用二分类交叉熵(BCE)损失L_{BCE},用于增强目标类区域特征的判别性。因此,总损失函数L_{total}为上述损失的结合,可以表示为:

$$L_{total}=L_{MSE}+L_{MAE}+\lambda L_{BCE}, \tag{6}$$

其中,前两个损失用于主干网络个体目标高分辨特征学习,具有相同的重要性,L_{BCE}损失在主干网络的特征基础上增强目标区域特征的类别感知特性,具有对主干网络的精炼优化作用,损失权重λ经验设置为$\lambda=0.1$。

不同于传统目标检测中的方框标签,本文利用盒标签生成目标区域二值图标签,采用二值图标签进行损失计算。所提方法在主干网络的人群目标特征基础上,利用类别感知特征去噪对目标区域特征增强,抑制特征噪声对独立个体目标预测的负面影响。

3 实验

3.1 数据集

NWPU-Crowd[31]:NWPU是近年来开发的一个大规模的拥挤的人群定位数据集,它包含了丰富的场景和数据注释。由5 109张图像组成,共2 133 375个带有点和盒子的注释头。与其他真实数据集相比,它包含了不同的照明场景,具体包括正常光、极端光、暗光场景,并具有最大的密度范围。与传统数据集不同的是,NWPU数据集引入了351个负样本(即无人场景),在纹理特征方面与拥挤的人群场景相似,有效提高了人群定位模型在现实世界中的泛化效果。这些负样本具体包括动物迁徙和假的人群,其中个体的目标包括雕塑、兵马俑、2d卡通人物等非人体目标。

Shanghai Tech[32]:数据集包含了Part A和Part B两个子数据集。A数据集共482张图片,主要从互联网上随机获取的图像,其中测试集大小为182张。共241 677个实例标注。B数据集共716张图片,主要来源于上海大都市繁华街道的图像,测试集大小为316张。共88 488个实例标注。这两个子数据集之间的人群密度差异很大,使得人群个体目标的准确预测比大多数现有数据集更具挑战性。

UCF-QNRF[9]:是一个拥挤的人群数据集,由佛罗里达大学在2018年发布,共包含1 535个密集人群图像,均为高清大图,图像分辨率为2 013×2 902。其中训练集1 201张图像,测试集334张图像,共计1 251 642个注释实例。该数据集包含了多种场景、多个视角、多种光线及密度变化的大规模已标注样本。

3.2 参数设置

由于原始图像的尺寸大,且不同图像尺寸不一致,为减小模型计算量和增加模型的处理速度,将原始图像随机水平翻转,缩放0.8~1.2倍,最后在图像中随机裁剪512×1 024大小区域输入模型中。输入图像的批处理大小为16。由于数据集大小不同,在训练过程中对不同数据集训练次数设置不同,NWPU数据集设置了600个epoch,Shanghai Part A和QNRF数据集设置1 000个epoch、Part B设置2 000个epoch。骨干网络学习置信度图的学习率为2×10^{-6},阈值网络学习阈值图的初始学习率为1×10^{-6},在训练过程中学习率随着训练次数epoch的增加线性减小。模型采用Adam优化器对网络进行优化,针对骨干网络优化器权重衰减率设置为1×10^{-5}。实验使用显卡设备为两张TITAN RTX。

3.3 评价指标

主要使用判别性评价指标对本文方法进行定量分析,包括实例级精度(Precision),召回率(Recall)和$F1$值。

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}},\tag{7}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}},\tag{8}$$

$$F1=\frac{2\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}},\tag{9}$$

其中,TP表示正确预测为真样本,FP表示错误预测为真样本,FN表示错误预测为假样本,可以看出,$F1$指标兼顾了精度和召回率两个方面。

3.4 实验结果分析

3.4.1 特征去噪消融实验

本文提出的特征去噪模块由膨胀卷积(D_conv)和类别感知权重学习(W_learning)两个部分组成,表1的消融实验在NWPU验证集上分别评估了特征去噪模块中两部分对特征去噪性能的作用大小。完整特征去噪模块在原有定位模型基础上性能提升了2.0%,在分别仅添加类别感知权重学习和膨胀卷积部分的条件下,$F1$值相对于基准方法几乎没有明显改善。这说明仅仅依靠扩张卷积或者权重学习难以实现特征去噪的效果。一方面,更大范围的邻域特征学习有利于捕获更多的非局部信息,更好地刻画目标相似性特征;另一方面,类别感知权重学习有利于将不同尺度的特征沿通道维度进行类别解耦。因此,扩大感受野和权重学习均是类别感知特征去噪中必不可少的组件。

表1 特征去噪模块消融实验

Table 1 Ablation experiment of feature denoising module

methods	$F1$/Precision/Recall
Baseline	0.791/0.860/0.732
Baseline+D_conv	0.805/0.871/0.748
Baseline+W_learning	0.805/0.857/0.758
Baseline+D_conv+W_learning(Ours)	**0.811**/0.854/0.772

人群图像特征可以用于识别对最终人头判别决策有强烈影响的像素,置信度高的位置会以较高概率被判别为存在人头目标。为了展示模型在进行人头预测时是关注人头目标还是背景区域,图4展示了在稀疏、密集和黑暗场景下使用特征去噪前后的人群定位结果和特征图,其中,第四列和第五列可视化了归一化后的置信图,以显示相应位置像素判别为目标的概率,概率越大表明相应位置像素为目标的可能性越大。可以看出,通过特征去噪对目标类和背景类特征进行解耦,抑制了其他特征对目标特征的干扰,提高了独立个体目标特征的判别性,改善了真样本正确预测的概率。

图4第二列和第三列分别展示了在稀疏、密集和黑暗场景下在baseline上未使用和使用特征去噪前后的人群定位效果图。其中,绿点代表真阳性目标,红点代表假阴性目标,粉点代表假阳性目标,绿圈和红圈是由真值给定的半径画圆。相应量化指标表明特征去噪模块在假阳性和假阴性检测结果具有一定的抑制作用,提升了密集人群定位的总体性能$F1$指标。

3.4.2 去噪方法对比实验

如图5所示,根据3类评价指标,将所提方法同基线方法,如基于双边滤波(bilateral)去噪,双

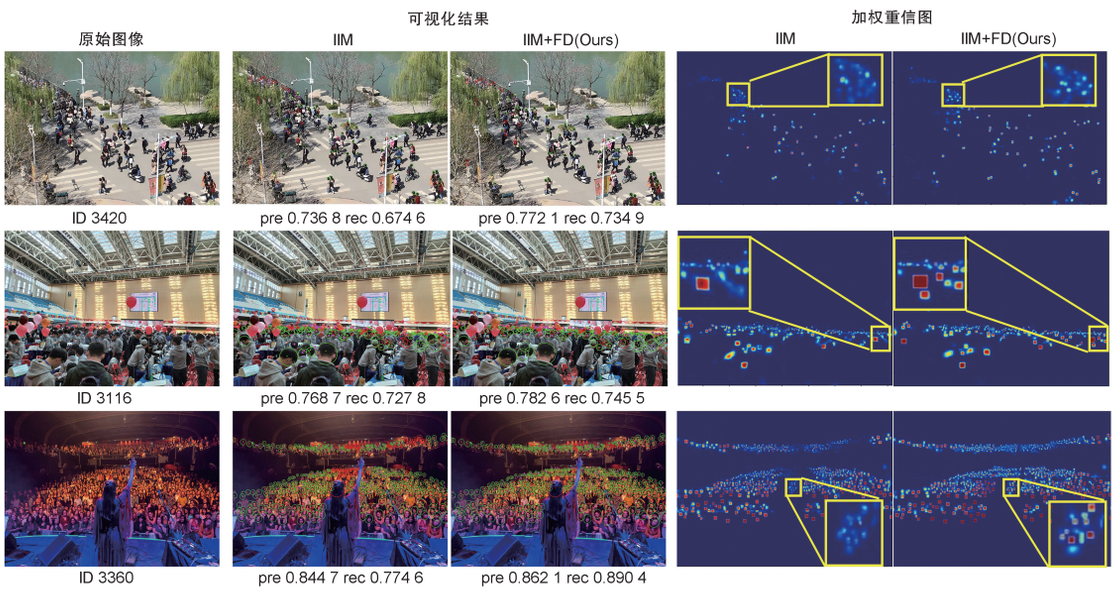


图4 使用特征去噪前后人群定位可视化结果

Fig. 4 Crowd localization visual results before and after feature denoising

边滤波结合高斯滤波去噪(bilateral +gaussian)和非局部均值结合高斯滤波去噪(nonlocal +gaussian)[23]的人群定位方法进行比较。以综合指标$F1$值为例,所提特征去噪方法与先进的像素域去噪方法在密集人群定位任务上分别提升了2.0%、1.3%、1.5%和1.5%。实验结果表明现有使用高斯处理的双边滤波和非局部均值方法均能有效的去除特征干扰提升人群定位效果,但本文使用的类别感知特征去噪方法能取得更高的综合性能。所提方法不仅具有较高的召回率,并且至少比先进的像素域去噪方法提高0.5%的$F1$值。换而言之,特征域去噪比像素域去噪更能提升高层视觉任务性能。

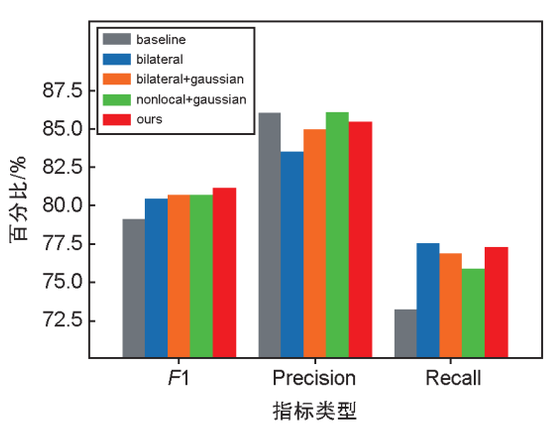


图5 图像去噪方法性能对比

Fig. 5 Performance of comparison among image denoising methods

3.4.3 人群定位对比实验

为合理验证所提基于类别感知特征去噪的人群定位方法的有效性,与现有算法分别在3种量化指标进行了对比分析。人群定位对比算法包括TinyFaces[6]、RAZ_Loc(Recurrent Attentive Zooming for Localization)[33]、LSC-CNN[3]和FIDTM(Focal Inverse Distance Transform Maps)[12],分别在NWPU-Crowd、ShanghaiTech Part A/B和UCF-QNRF公开数据集上进行验证。如表格2所示,所提方法在公开人群场景数据集上都取得最优的$F1$指标。在ShanghaiTech A和UCF-QNRF这两个密集人群数据集中,TinyFaces在召回率表现最好。通过详细的性能比较,TinyFaces、RAZ loc和FIDTM往往输出较高的召回率,但更低的精确率,而所提方法具有更为平衡的召回率和精确率。

为直观地展示定位的视觉效果,比较现有四种先进方法和本文方法,图6给出了在NWPU-Crowd验证集上的四组典型样本(无人场景,稀疏人群,黑暗环境,密集人群场景)的可视化结果。在无人场景(ID3389),所提基于类别感知特征去噪的定位方法输出零人头,明显优于TinyFaces, RAZ Loc, LSC-CNN和FIDTM。在稀疏场景(ID3329)和黑暗场景(ID3177),所提方法的$F1$指标是所有方法中最

表2 人群定位方法对比

Table 2 Comparison of crowd localization methods

Method	Backbone	ShanghaiTech Part A	ShanghaiTech Part B	UCF–QNRF	NWPU–Crowd
		*F*1/Pre/Rec/%	*F*1/Pre/Rec/%	*F*1/Pre/Rec/%	*F*1/Pre/Rec/%
TinyFaces	ResNet–101	57.3/43.1/**85.5**	71.1/64.7/79.0	49.4/36.3/**77.3**	56.7/52.9/61.1
RAZ_Loc	VGG–16	69.2/61.3/79.5	68.0/60.0/78.3	53.3/59.4/48.3	59.8/66.6/54.3
LSC–CNN	VGG–16	68.0/69.6/66.5	71.2/71.7/70.6	58.2/58.6/57.7	—/—/—
FIDTM	HRNet	64.6/58.7/71.9	74.6/79.6/70.3	60.9/51.7/73.9	71.1/67.5/**75.2**
本文方法	HRNet	**75.9**/**82.7**/70.1	**86.4**/**90.9**/**82.2**	**72.4**/75.5/69.6	**77.1**/**82.9**/72.1

好的,可以产生比其他方法更少的FP和更多的TP。在密集的人群场景(ID3308),Tiny Faces,RAZ Loc和LSC-CNN在非人群区域输出较多的FP和在密集人群区域输出较多FN。FIDTM在稀疏和密集人群场景输出较少的TP和FN。当面对极其拥挤的场景时,所提方法的*F*1和精确率指标是最好的。总体而言,所提方法能够处理各种人群场景,对无人场景的人群定位具有鲁棒性,在召回率损失较小的条件下较明显地提高了密集场景人群定位的精确率。可以看出,所提特征去噪方法在上述四类场景中均有助于改善独立个体目标检测的性能。

4 结论

本文提出一种基于特征去噪的人群定位方法,该去噪方法区别于传统面向像素域噪声的去噪方法,其通过类别特征去耦实现下游任务性能提升。类别特征学习促使神经网络模型捕获更多关于目标特征与标签之间的关联信息,并抑制复杂背景特征干扰,进而改善在密集场景目标定位性能。然而粗语义分割在微小目标空间特征获取方面准确率较低,成为类别感知权重学习性能的重要瓶颈之一,造成定位性能难以得到显著改善。未来的工作将继续围绕复杂场景下高层语义任务中的特征去噪展开更深入的研究,研究基于类别感知的特征增强方法的内在机理,进一步改善复杂场景下的独立个体目标预测性能。

参考文献:

[1] 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.

[2] 李佳倩, 严华. 基于跨列特征融合的人群计数方法[J]. 计算机科学, 2021, **48**(6): 118–124. DOI: 10.11896/jsjkx.200700107.
LI J Q, YAN H. Crowd Counting Method Based on Cross-column Features Fusion[J]. *Comput Sci*, 2021, **48**(6): 118–124. DOI: 10.11896/jsjkx.200700107.

[3] SAM D B, PERI S V, SUNDARARAMAN M N, *et al*. Locate, Size, and Count: Accurately Resolving People in Dense Crowds *via* Detection[J]. *IEEE Trans Pattern Anal Mach Intell*, 2021, **43**(8): 2739–2751. DOI: 10.1109/TPAMI.2020.2974830.

[4] STEWART R, ANDRILUKA M, NG A Y. End-to-end People Detection in Crowded Scenes[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 2325–2333. DOI: 10.1109/CVPR.2016.255.

[5] LIAN D Z, LI J, ZHENG J, *et al*. Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 1821–1830. DOI: 10.1109/CVPR.2019.00192.

[6] HU P Y, RAMANAN D. Finding Tiny Faces[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 1522–1530. DOI: 10.1109/CVPR.2017.166.

[7] LIU Y T, SHI M J, ZHAO Q J, *et al*. Point In, Box Out: Beyond Counting Persons in Crowds[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 6462–6471. DOI: 10.1109/CVPR.2019.00663.

[8] GAO J, HAN T, WANG Q, *et al*. Domain-adaptive Crowd Counting via High-quality Image Translation and Density Reconstruction"[EB/OL]. arXiv Preprint: 1912.03677, 2019.

[9] IDREES H, TAYYAB M, ATHREY K, *et al*. Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds[C]//Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part II. New York: ACM, 2018: 544–559. DOI: 10.1007/978-3-030-01216-8_33.

[10] ABOUSAMRA S, HOAI M, SAMARAS D, *et al*. Localization in the Crowd with Topological Constraints[J]. *Proc AAAI Conf Artif Intell*, 2021, **35**(2): 872–881.

DOI: 10.1609/aaai.v35i2.16170.

[11] GAO J, HAN T, WANG Q, *et al*. Learning Independent Instance Maps for Crowd Localization[J]. arXiv Preprint: 2012.04164, 2020.

[12] LIANG D K, XU W, ZHU Y Y, *et al*. Focal Inverse Distance Transform Maps for Crowd Localization[J]. *IEEE Trans Multimed*, 2022, PP(99): 1-13. DOI: 10.1109/TMM.2022.3203870.

[13] CAO X K, WANG Z P, ZHAO Y Y, *et al*. Scale Aggregation Network for Accurate and Efficient Crowd Counting[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 757-773. DOI: 10.1007/978-3-030-01228-1_45.

[14] 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: IEEE, 2018: 1091-1100. DOI: 10.1109/CVPR.2018.00120.

[15] 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: IEEE, 2020: 5094-5103. DOI: 10.1109/CVPR.2019.00524.

[16] DAI F, LIU H, MA Y K, *et al*. Dense Scale Network for Crowd Counting[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval. New York: ACM, 2021: 64-72. DOI: 10.1145/3460426.3463628.

[17] JIANG X H, ZHANG L, XU M L, *et al*. Attention Scaling for Crowd Counting[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 4705-4714. DOI: 10.1109/CVPR42600.2020.00476.

[18] LI Z, TANG X, HAN J, *et al*. PyramidBox++: High Performance Detector for Finding Tiny Face[J]. arXiv Preprint: 1904.00386, 2019.

[19] WANG Y, HOU J H, HOU X Y, *et al*. A Self-training Approach for Point-supervised Object Detection and Counting in Crowds[J]. *IEEE Trans Image Process*, 2021, **30**: 2876-2887. DOI: 10.1109/TIP.2021.3055632.

[20] TIAN C W, FEI L K, ZHENG W X, *et al*. Deep Learning on Image Denoising: an Overview[J]. *Neural Netw*, 2020, **131**: 251-275. DOI: 10.1016/j.neunet.2020.07.025.

[21] GUO S, YAN Z F, ZHANG K, *et al*. Toward Convolutional Blind Denoising of Real Photographs[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2020: 1712-1722. DOI: 10.1109/CVPR.2019.00181.

[22] ZHANG B, JIN S Y, XIA Y L, *et al*. Attention Mechanism Enhanced Kernel Prediction Networks for Denoising of Burst Images[C]/2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 2083-2087. DOI: 10.1109/ICASSP40776.2020.9053332.

[23] XIE C H, WU Y X, VAN DER MAATEN L, *et al*. Feature Denoising for Improving Adversarial Robustness [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 501-509. DOI: 10.1109/CVPR.2019.00059.

[24] LIU D, WEN B H, JIAO J B, *et al*. Connecting Image Denoising and High-level Vision Tasks *via* Deep Learning[J]. *IEEE Trans Image Process*, 2020, **29**: 3695-3706. DOI: 10.1109/TIP.2020.2964518.

[25] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-scale Image Recognition[J]. arXiv Preprint : 1409.1556, 2014.

[26] WANG J D, SUN K, CHENG T H, *et al*. Deep High-resolution Representation Learning for Visual Recognition [J]. *IEEE Trans Pattern Anal Mach Intell*, 2021, **43**(10): 3349-3364. DOI: 10.1109/TPAMI.2020.2983686.

[27] HU J, SHEN L, SUN G. Squeeze-and-excitation Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141. DOI: 10.1109/CVPR.2018.00745.

[28] WANG X L, GIRSHICK R, GUPTA A, *et al*. Non-local Neural Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 7794-7803. DOI: 10.1109/CVPR.2018.00813.

[29] WANG P Q, CHEN P F, YUAN Y, *et al*. Understanding Convolution for Semantic Segmentation[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1451-1460. DOI: 10.1109/WACV.2018.00163.

[30] APPIAH K, HUNTER A, DICKINSON P, *et al*. Accelerated Hardware Video Object Segmentation: From Foreground Detection to Connected Components Labelling[J]. *Comput Vis Image Underst*, 2010, **114**(11): 1282-1291. DOI: 10.1016/j.cviu.2010.03.021.

[31] 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.

[32] 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: IEEE, 2016: 589-597. DOI: 10.1109/CVPR.2016.70.

[33] LIU C C, WENG X Y, MU Y D. Recurrent Attentive Zooming for Joint Crowd Counting and Precise Localization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2020: 1217-1226. DOI: 10.1109/CVPR.2019.00131.