中文引用格式:冉瑞生,李进,董殊宏. 基于 Swin Transformer 的弱监督人群计数研究[J].中国安全科学学报,2023,33(3):111-117.
英文引用格式:RAN Ruisheng,LI Jin,DONG Shuhong. Research on weakly supervised crowd counting based on Swin Transformer [J].China Safety Science Journal,2023,33(3):111-117.

基于 Swin Transformer 的弱监督人群计数研究*

冉瑞生 教授, 李 进, 董殊宏
(重庆师范大学 计算机与信息科学学院,重庆 401331)

中图分类号:X915.2 **文献标志码**:A **DOI**: 10.16265/j.cnki.issn1003-3033.2023.03.0784
资助项目:重庆市技术创新与应用发展专项面上项目(cstc2020jscx-msxmX0190);重庆市教委科学技术研究重点项目(KJZD-K202100505)。

【摘 要】 为降低人群聚集引发安全事故的概率,解决完全监督方法数据标注成本高,而现有弱监督方法性能欠佳的问题,提出一种基于 Swin Transformer 的弱监督人群计数模型。首先,引入具有全局感受野且能够有效提取语义人群信息的 Transformer 模型,来应对基于卷积神经网络(CNN)的弱监督人群计数方法感受野有限、性能欠佳的问题;然后,采用具有层级设计并且拥有多尺度、层次化计算图像特征能力的 Swin Transformer 模型作为主干网络,以加强对不同尺度特征的学习,使模型能够更好地应对人群尺度变化的问题;最后,选择只需要人群数量作为监督信息的弱监督方式进行训练,避免对图像中每个人的头部进行标注这一繁琐易错的工作。结果表明:所提模型在 ShanghaiTech Part A、ShanghaiTech Part B、UCF-QNRF 数据集上的平均绝对误差依次为 66.1、8.7、97.1,均方误差依次为 106.2、14.9、165.8,在主流数据集上计数性能较好;该模型的性能优于此前的弱监督方法和部分完全监督方法。

【关键词】 Swin Transformer; 弱监督; 人群计数; 卷积神经网络(CNN); 数据集

Research on weakly supervised crowd counting based on Swin Transformer

RAN Ruisheng, LI Jin, DONG Shuhong
(College of Computer & Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: In order to reduce the probability of safety accidents caused by crowd gathering, research is carried out on the crowd counting task. For the problem of the high data labeling cost of the full supervision method and poor performance of the existing weak supervision method, a weak supervision crowd counting model based on Swin Transformer is designed. First, a Transformer model with a global receptive field and the ability to effectively extract semantic crowd information was introduced to deal with the problem of the limited receptive field and poor performance of the weakly supervised crowd counting method based on CNN. Then, a hierarchical design was adopted. The Swin Transformer model with multi-scale and hierarchical computing image features was used as the backbone network to strengthen the learning of different scale features, so that the model can better deal with the problem of crowd scale changes. Finally, the selection only needs the number of people as supervisory information. Weakly supervised training of

* 文章编号:1003-3033(2023)03-0111-07; 收稿日期:2022-10-16; 修稿日期:2023-01-09

information, avoiding the tedious and error-prone work of labeling each person´s head in the image. The results show that the average absolute error of the method in this paper on ShanghaiTech Part A, ShanghaiTech Part B, and UCF-QNRF datasets is 66.1, 8.7, and 97.1, and the mean square error is 106.2, 14.9, and 165.8, which is better than the previous weakly supervised method and partially fully supervised methods.

Keywords: Swin Transformer; weakly supervised; crowd counting; convolutional neural network(CNN); datasets

0 引 言

近年来,由公共场所的人群聚集引发的踩踏事故频发,导致了大量的人员伤亡以及公共财产损失。对公共场所的人群采取智能安全监控,可提供有效预警,预防事故的发生。智能安全监控系统的人群计数功能旨在估计图像或视频中的人数,辅助相关人员及时发现拥挤情况并制定疏散策略。此外,新型冠状病毒肺炎目前仍在全球肆虐蔓延,有效疏散人群能够降低病毒传播的风险。

人群计数方法主要有3类,基于检测的方法[1]、基于回归的方法[2]以及基于深度学习的方法[3]。当前,受益于卷积神经网络(Convolutional Neural Network,CNN)[4-6]的蓬勃发展,基于深度学习的方法已经成为人群计数领域的主流选择。ZHANG Yingying 等[3]提出多列 CNN(Multi-Column CNN,MCNN),使用三列卷积分别对人群图像进行特征计算,提取多尺度的特征信息,更好地解决了人群尺度变化问题。LI Yuhong 等[7]提出使用空洞卷积的人群计数模型(Dilated Convolutional Neural Networks for Understanding Highly Congested Scenes,CSRNet),保留更多图像细节并扩大感受野,生成了更高质量的人群密度图。MCNN、CSRNet 等人群计数方法都在完全监督下进行训练,虽然其精度在不断提高,但要求标注数据集中的每个人头部,而密集场景下目标繁多,标注工作繁琐且易错。此外,模型在每轮训练完成进行测试时,性能的评价指标只使用了人群数量,而没有使用标注的头部位置。人群数量的获取比数据标注更容易,针对现有数据集,可以通过收集环境信息获取人群数量,CHAN 等[8]提出,根据人群运动进行场景分割,并通过计算分割后的区域面积来估计人群数量。若要收集新的数据集,可采用移动人群感应技术[9]和无 GPS 能源效率的传感调度[10]等方式获取人群数量。目前,研究人员也展开了基于弱监督的人群计数研究,YANG Yifan 等[11]为减轻标注负担,提出在软标签排序网络的基础上直接对人群数量进行回归,不进行定位监督。LEI Yinjie 等[12]提出使用少量完全监督信息和大量弱监督信息进行计数。然而,目前的弱监督方法无法取得可观的性能,很难应用到实际中。Transformer[13]具有出色的性能以及全局感受野,在自然语言处理领域取得了突破性的成绩。DOSOVITSKIY 等[14]最早将 Transformer 引入到计算机视觉领域并提出了第一个视觉的 Transformer 模型 ViT(Vision Transformer),取得了优异的性能。

尽管 ViT 的动态感受野能更有效地提取语义人群信息,但该模型始终保持单一尺度,缺乏对不同尺度特征的学习,会丢失部分特征信息。鉴于此,笔者拟提出一种基于 Swin Transformer[15]的弱监督人群计数模型,利用其多尺度、层次化的计算特性来提升性能,并在主流数据集上开展试验,与部分主流方法(包括完全监督、弱监督方法)进行对比,验证该方法的科学性和有效性,旨在提高人群计数方法在弱监督下的准确性和鲁棒性,为弱监督人群计数研究提供论据。

1 Swin Transformer 简介

1.1 Swin Transformer 模型

Swin Transformer 模型计算过程为:首先,将输入图像传入核大小为 4×4 的卷积层进行下采样,完成后投影到 C 维上,得到维度为 $\frac{W}{4}\times\frac{H}{4}\times C$ 的特征向量,其中,W 和 H 为输入图像的宽度和高度,C 为特征向量的维度;然后,依次进行窗口多头自注意力(Window Multi-head Self Attention,W-MSA)模块的计算,以及移位 W-MSA(Shifted W-MSA,SW-MSA)模块的计算,得到第 1 阶段的特征向量;最后,利用图像块合并层(Patch Merging)对其进行下采样并增加维度,再重复上述操作,得到第 2、3、4 阶段的特征向量,实现在每个阶段学习到相应尺度的特征信息。

1.2 Swin Transformer 模块

图像的上下文信息对于人群计数的准确性十分

重要,而长距离的语义信息能够让模型不仅仅依赖于图像信息。文献[16]提出,自注意力机制可以有效提取语义人群信息。不同于多头自注意力(Multi-head Self Attention, MSA)模块, Swin Transformer 模块基于移位窗口构建,其结构如图 1 所示。图 1 中,2 个连续的计算模块由归一化(Layer Norm,LN)层、W-MSA 模块、SW-MSA 模块、残差连接和具有非线性 RELU 激活函数的多层感知机(Multi-layer Perceptron,MLP)组成。$\hat{z}^l$ 和 z^l 分别为 W-MSA 模块和 MLP 模块在第 l 块的输出。

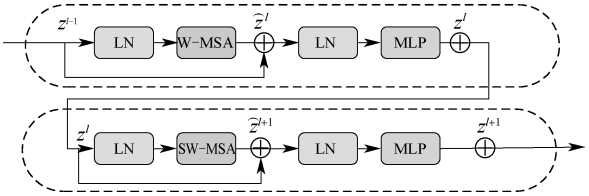

图 1　Swin Transformer 模块结构

Fig.1　Swin Transformer Block structure

W-MSA 模块和 SW-MSA 模块被分别应用在 2 个连续的模块上,基于这样的窗口划分机制,该结构可以表述为:

$$\hat{z}^l = \text{W-MSA}(\text{LN}(z^{l-1})) + z^{l-1} \quad (1)$$

$$z^l = \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l \quad (2)$$

$$\hat{z}^{l+1} = \text{SW-MSA}(\text{LN}(z^l)) + z^l \quad (3)$$

$$z^{l+1} = \text{MLP}(\text{LN}(\hat{z}^{l+1})) + \hat{z}^{l+1} \quad (4)$$

自注意力计算表述为:

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}} + B\right)\boldsymbol{V} \quad (5)$$

式中:$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为 Query、Key 和 Value 向量;d_k 为 Query 以及 Key 的维度;B 为注意力头的相对位置偏差。

2　构建人群计数模型

2.1　总体结构设计

基于 Swin Transformer 的弱监督人群计数模型由输入图像、Swin Transformer 结构、全局平均池化层以及回归模块组成。构建模型的具体步骤为:首先,将 $W\times H\times 3$ 的人群图像输入到卷积核大小为 4×4,步幅为 4 的卷积层,将其转换为 $\frac{W}{4}\times\frac{H}{4}\times 48$ 的特征向量,并经过线性嵌入层转换维度;然后,进入 Swin Transformer 模块进行 W-MSA 和 SW-MSA 的计算,得到第 1 阶段的结果;最后,通过图像块合并层(Patch Merging)对第 1 阶段的特征进行下采样转换尺度和维度,并将其作为第 2 阶段的输入。重复上述操作,依次进行第 2、3、4 阶段的计算,最终输出特征计算结果。在计算的过程中,特征向量不断变换维度,形成层次化的特征图,使模型学习到不同尺度的特征信息。层次化特征如图 2 所示,模型在每个阶段进行特征计算时都会将注意力计算限制在窗口内,结合移动窗口机制使相邻窗口间信息交互,学习当前尺度下的局部信息以及全局信息,然后通过图像块合并层转换特征的尺度和维度并传递到下一个阶段,实现对人群图像特征多尺度、层次化的计算。具体来说,图像块合并层的工作是负责下采样和增加维度,W-MSA 模块和 SW-MSA 模块的工作是负责特征计算和学习。

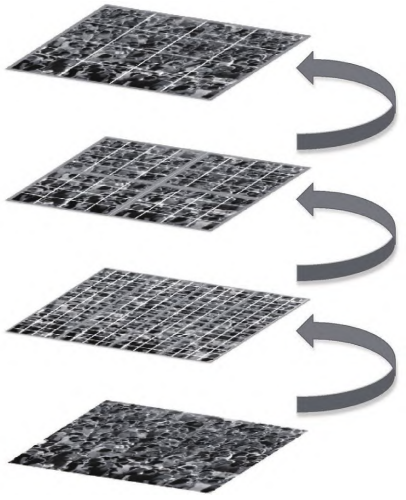

图 2　层次化特征

Fig.2　Hierarchical feature maps

将 Swin Transformer 网络提取的特征输入到全局平均池化层进行降维,再对其进行回归计算得到预测人数。模型训练只使用真实人数作为监督信息,预测人数与真实人数的误差即为模型的损失值。整体网络结构如图 3 所示,输入为一张人群图像,输出为通过模型计算得到的对应预测人数值。

2.2　全局平均池化层

将网络提取的特征序列经过全局平均池化层进行降维。文献[17]提出,全局平均池化能够减少网络的参数量,可以对模型进行正则化,防止过拟合,全局平均池化操作可以更好地整合全局空间信息,对于输入图片的空间语义提取更加鲁棒,并且还能有效的保持特征中有用的人群信息。

2.3　回归模块和损失函数

网络计算的特征通过全局平均池化降维后输入到回归模块,计算预测人数,回归模块包括 2 个

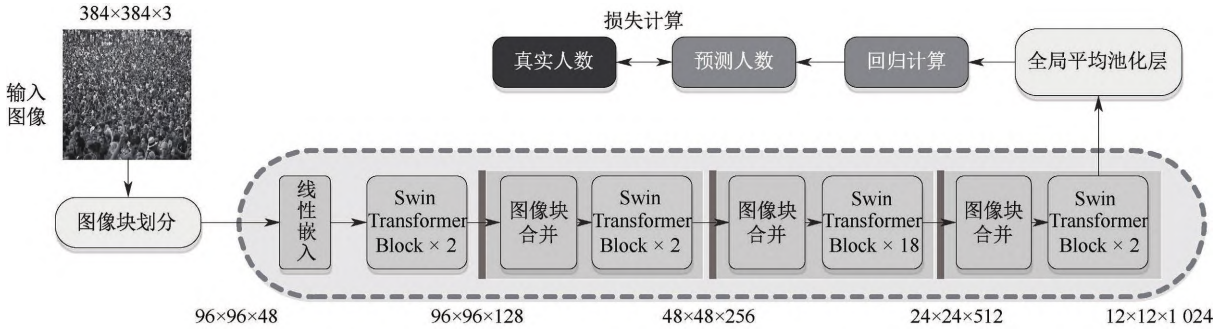

图 3　网络结构设计

Fig.3　Network structure design

ReLU 激活函数、2 个全连接层以及 1 个 Dropout 层，如图 4 所示。

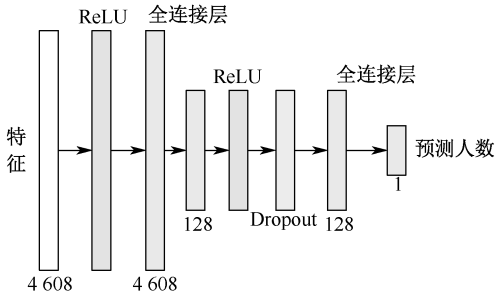

图 4　回归模块

Fig.4　Regression module

将衡量预测值与真实值之间的损失函数表述为：

$$L = \frac{1}{M}\sum_{i=1}^{M} | P_i - G_i | \tag{6}$$

式中：P_i 和 G_i 分别为第 i 张图的预测人数和对应的真实人数；M 为一次训练放入网络中图片的数量。

3　试验结果与分析

3.1　数据集和评价指标

ShanghaiTech 数据集、UCF-QNRF 数据集是目前人群计数研究领域的主流数据集，详细情况见表 1。

表 1　ShanghaiTech 数据集和 UCF-QNRF 数据集信息

Tab.1　Information of ShanghaiTech datasets and UCF-QNRF datasets

数据集	分辨率	图片数量	最多人数	最少人数	平均人数	总体人数
ShanghaiTech Part A	不统一	482	3 139	33	501	241 667
ShanghaiTech Part B	768×1 024	716	578	9	123	88 488
UCF-QNRF	不统一	1 535	12 865	49	815	1 251 642

ShanghaiTech 数据集[3]由 2 部分组成：Part A 和 Part B，数据集包含 1 198 张人群图像，共有 330 165 个注释头。A 部分包含 482 张密集人群的互联网图像，其中，训练图像 300 张、测试图像 182 张；B 部分包含 716 张固定尺寸为 768×1 024 的图像，其中，训练图像 400 张、测试图像 316 张。

UCF-QNRF 数据集[18]包含 1535 张从密集人群场景中捕获到的图像，其中，包含有 1 252 642 人，它的计数范围由 49 到 12 865，平均计数 815.4，其中训练图像 1 201 张、测试图像 334 张。

计数模型的准确率和鲁棒性的评价指标为平均绝对误差（Mean Absolute Error，MAE）、均方误差（Mean Square Error，MSE），具体定义为：

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} | R_i - R'_i | \tag{7}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (R_i - R'_i)^2} \tag{8}$$

式中：N 为样本的数量；R_i 为预测的人数；R'_i 为真实的人数。MAE 可以反映模型的准确性，值越小说明模型准确度越高，而 MSE 可以反映模型的鲁棒性，其值越小说明模型鲁棒性越好。

3.2　试验环境和参数配置

试验均在 Ubuntu 系统下，使用 24 G NVIDIA GeForce RTX 3090 GPU 显卡和 Cuda 11.1 版本的 Pytorch 框架下进行训练。数据集图像统一处理为 384×384，使用 Adam 优化器，学习率设置为 10^{-5}，Batch Size 设置为 12。将 TransCrowd[14]选作基准方法，与该方法对比的试验结果均在相同的试验参数设置下得到。

3.3　Swin Transformer 模型

选用 Swin-B[15]模型,输入图像通过 4×4 大小,步幅为 4 的卷积层进行下采样,通过该卷积层将 384×384 的输入图像快速下采样至 96×96。将图像处理为 384×384 作为网络输入,通过卷积层将 384×384×3 的图像下采样并增加维度至 96×96×128,再进行 MSA 和 SW-MSA 模块的计算,进而通过图像块合并层下采样增加为维度至 48×48×256,重复上述工作,依次变为 24×24×512 和 12×12×1 024。

3.4　全局平均池化层和回归模块的设计

采用全局平均池化层对特征进行降维,将降维后的特征送入到回归模块中计算得到预测人数。具体地,回归模块设计为 ReLU 激活函数-全连接层-ReLU 激活函数-DropOut 层-全连接层。输入图像传入网络提取的特征维度为 12×12×1 024,经过全局平均池化层将其降维至 12×12×32。再将特征展平为一维,此时维度处理为 4 608,经过 ReLU 激活函数后,送入到全连接层将维度处理为 128,再经过 ReLU 激活函数和 Dropout 函数,最后传入全连接层直接得到预测人数值。

3.5　ShanghaiTech Part A 数据集上的试验结果

Part A 数据集上的试验结果见表 2。与弱监督方法相比,文中方法比 Sorting 的 MAE 低 38.5,MSE 低 39.0;比 MATT 的 MAE 低 13.1,MSE 低 28.4;比基准方法 TransCrowd 的 MAE 和 MSE 均更低。与完全监督方法相比,文中方法的性能与 SCANet 的性能几乎持平,能够略优于 CSRNet,且大幅优于较早的完全监督方法,但略差于 BL、DM-Count 等先进完全监督方法。

表 2　ShanghaiTech Part A 数据集上的试验结果

Tab.2　Experimental results on ShanghaiTech Part A dataset

方法	标签		MAE	MSE
	位置	人数		
MCNN[3]	√	√	110.2	173.2
L2R[19]	√	√	73.6	112.0
ResNeXtFP[20]	√	√	69.3	104.7
CSRNet[7]	√	√	68.2	115.0
SCANet[6]	√	√	66.0	104.3
BL[21]	√	√	62.8	101.8
DM-Count[22]	√	√	59.7	95.7
Sorting[11]	×	√	104.6	145.2
MATT[12]	×	√	80.1	129.4
TransCrowd-G[16]	×	√	70.5	115.1
TransCrowd-T[16]	×	√	69.8	109.5
Ours	×	√	66.1	106.2

注:√表示需要相应标签;×表示无需相应标签。表 3、表 4 同。

Part A 数据集中的图像人群分布较密集且场景较复杂,说明模型在复杂场景和密集场景下能具备较好准确性和鲁棒性。

3.6　ShanghaiTech Part B 数据集上的试验结果

Part B 数据集上的试验结果见表 3。与弱监督方法相比,文中方法比 Sorting 的 MAE 低 3.5,MSE 低 6.3;比 MATT 的 MAE 低 3.0,MSE 低 2.6;比基准方法 TransCrowd 的 MAE 和 MSE 均更低。与完全监督方法相比,文中方法 SCANet 的性能几乎持平,能够略优于 CSRNet,且大幅优于较早的完全监督方法,但略差于 BL、DM-Count 等先进完全监督方法。

表 3　ShanghaiTech Part B 数据集上的试验结果

Tab.3　Experimental results on ShanghaiTech Part B dataset

方法	标签		MAE	MSE
	位置	人数		
MCNN[3]	√	√	26.4	41.3
L2R[19]	√	√	13.7	21.4
CSRNet[7]	√	√	10.6	16.0
SCANet[6]	√	√	8.4	13.9
BL[21]	√	√	7.7	12.7
DM-Count[22]	√	√	7.4	11.8
Sorting[11]	×	√	12.3	21.2
MATT[12]	×	√	11.7	17.5
TransCrowd-T[16]	×	√	10.6	23.0
TransCrowd-G[16]	×	√	9.6	16.3
Ours	×	√	8.7	14.9

Part B 数据集中的图像人群分布相对较稀疏,试验结果验证了模型在稀疏场景下也具备较好的准确性和鲁棒性。

3.7　UCF-QNRF 数据集上的试验结果

UCF-QNRF 数据集上的试验结果见表 4。文中方法在该数据集上 MAE 达到了 97.1,MSE 达到了 165.8,与弱监督方法相比,MAE 和 MSE 2 项指标均优于基准方法 Trans Crowd。与完全监督方法相比,文中方法的性能能够优于 DSSI-Net、S-DCNet 等方法,且大幅优于较早的完全监督方法。

表 4 UCF-QNRF 数据集上的试验结果

Tab.4 Experimental results on UCF-QNRF dataset

方法	标签		MAE	MSE
	位置	人数		
MCNN[3]	√	√	277.0	426.0
CL[18]	√	√	132.0	191.0
L2R[19]	√	√	124.0	196.0
TEDnet[23]	√	√	113.0	188.0
CAN[24]	√	√	107.0	183.0
S-DCNet[25]	√	√	104.4	176.1
DSSI-Net[26]	√	√	99.1	159.2
BL[21]	√	√	88.7	154.8
DM-Count[22]	√	√	85.6	148.3
Sorting[11]	×	√	—	—
MATT[12]	×	√	—	—
TransCrowd-T[16]	×	√	99.7	172.3
TransCrowd-G[16]	×	√	99.0	169.3
Ours	×	√	97.1	165.8

UCF-QNRF 数据集分辨率较高,包含不同场景、光线、视角,试验结果说明,在不同的场景和环境变化下,模型仍具备良好的性能。

3.8 模型参数量、计算量和运行时间对比

对比各个模型在 ShanghaiTech Part A 数据集上的训练模型对参数量、计算量(Floating Point Operations,FLOPs)和单张图片的运行时间,结果见表5。完全监督方法 BL、DM-Count 和弱监督方法 MATT 均未使用 Transformer 模型,参数量相对较低,但 FLOPs 相对较高。文中方法的参数量为 87.2 M,FLOPs 为 44.5 G,运行时间为 1.46 s,基准方法 TransCrowd 的参数量为 86.4 M,FLOPs 为 49.3 G,运行时间为 1.44 s。通过试验发现,文中方法在性能提升的同时,参数量相比基准方法略有增加,运行时间持平,而 FLOPs 有明显的降低。

表 5 参数量、FLOPs 及运行时间试验

Tab.5 Parameter amount, FLOPs and running time experiment table

方法	参数量	FLOPs	运行时间
BL[21]	21.5M	60.8G	—
DM-Count[22]	21.5M	60.8G	—
MATT[12]	16.3M	60.9G	—
TransCrowd[16]	86.4M	49.3G	1.44 s
Ours	87.2M	44.5G	1.46 s

3.9 消融试验

为证明全局平均池化能够提升模型性能,在 UCF_QNRF 数据集上进行消融试验,结果见表6。

表 6 在 UCF_QNRF 数据集上的消融试验

Tab.6 Ablation experiments on UCF_QNRF dataset

方法	MAE	MSE
Swin	97.8	173.3
Swin+全局平均池化层	97.1	165.8

表6中,直接通过 Swin Transformer 网络提取特征,进行回归预测,得到的 MAE 为 97.8,MSE 为 173.3。结合 Swin Transformer 提取特征,再通过全局平均池化层对其降维后回归预测得到的 MAE 为 97.1,MSE 为 165.8。MAE 减少 0.7,MSE 减少 7.5,由此可见:增加全局平均池化层后,模型的准确性和鲁棒性均有所提升。试验结果表明:加入全局平均池化层能够有效保留特征中的人群信息,能够提升网络的准确性和鲁棒性。

4 结 论

1)文中提出一种基于 Swin Transformer 的弱监督人群计数模型,该模型在 ShanghaiTech Part A、ShanghaiTech Part B、UCF-QNRF 数据集上的平均绝对误差依次为 66.1、8.7、97.1,均方误差依次为 106.2、14.9、165.8,可见:该模型在主流数据集上计数性能较好,与其他弱监督方法相比,该模型的平均完全误差和均方误差更优异;与完全监督方法相比,具有较强的竞争力。

2)文中计数方法也存在一定局限性,例如:放弃行人头部位置信息,导致失去获取人群位置以及感知各区域人群密度变化的能力。下一步研究可提出性能更好且可以适用于2种监督方式的人群计数模型,进而实现在不同的需求下,选择不同的监督方式进行人群计数。

参 考 文 献

[1] TOPKAYA I S, ERDOGAN H, PORIKLI F. Counting people by clustering person detector outputs[C]. 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2014: 313-318.

[2] CHAN A B, VASCONCELOS N. Bayesian poisson regression for crowd counting[C]. 2009 IEEE 12th International Conference on Computer Vision. IEEE, 2009: 545-551.

[3] 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 (CVPR), 2016: 589-597.
[4] 徐丹, 代勇, 纪军红. 基于卷积神经网络的驾驶人行为识别方法研究[J]. 中国安全科学学报, 2019, 29(10): 12-17.
XU Dan, DAI Yong, JI Junhong. Research on driver behavior recognition method based on convolutional neural network[J]. China Safety Science Journal, 2019, 29(10): 12-17.
[5] 熊若鑫, 宋元斌, 王宇轩,等. 基于 CNN 的 3D 姿势估计在建筑工人行为分析中的应用[J]. 中国安全科学学报, 2019, 29(7): 64-69.
XIONG Ruoxin, SONG Yuanbin, WANG Yuxuan, et al. Application of convolutional neural network-based 3D posture estimation in behavioral analysis of construction workers[J]. China Safety Science Journal, 2019, 29(7): 64-69.
[6] 吴思, 张旭光, 方银锋. 基于注意力机制的人群计数方法[J]. 中国安全科学学报, 2022, 32(1): 127-134.
WU Si, ZHANG Xuguang, FANG Yinfeng. Method of crowd counting based on attention mechanism[J]. China Safety Science Journal, 2022, 32(1): 127-134.
[7] LI Yuhong, ZHANG Xiaofan, CHEN Deming. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes [C]. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1 091-1 100.
[8] CHAN A B, LIANG Z S J, VASCONCELOS N. Privacy preserving crowd monitoring: counting people without people models or tracking[C]. 2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-7.
[9] GUO Bin, WANG Zhu, YU Zhiwen, et al. Mobile crowd sensing and computing: the review of an emerging human-powered sensing paradigm[J]. ACM Computing Surveys (CSUR), 2015, 48(1): 1-31.
[10] SHENG Xiang, TANG Jian, XIAO Xuejie, et al. Leveraging GPS-less sensing scheduling for green mobile crowd sensing[J]. IEEE Internet of Things Journal, 2014, 1(4): 328-336.
[11] YANG Yifan, LI Guorong, WU Zhe, et al. Weakly-supervised crowd counting learns from sorting rather than locations[C]. European Conference on Computer Vision. Springer, Cham, 2020: 1-17.
[12] LEI Yinjie, LIU Yan, ZHANG Pingping, et al. Towards using count-level weak supervision for crowd counting[J]. Pattern Recognition (PR), 2021, 109:DOI:10. 1016/j.patcog.2020. 107616.
[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017: 5 998-6 008.
[14] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv preprint, 2020:DOI:10.48850/arXiv.2010.11929.
[15] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10 012-10 022.
[16] LIANG Dingkang, CHEN Xiwu, XU Wei, et al. TransCrowd: weakly-supervised crowd counting with transformers[J]. Science China Information Sciences, 2022, 65(6): 1-14.
[17] LIN Min, CHEN Qiang, YAN Shuicheng. Network in network[J]. arXiv Preprint, 2013: DOI: 10. 48550/arXiv. 1312. 4400.
[18] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 532-546.
[19] LIU Xialei, VAN D W J, BAGDANOV A D. Exploiting unlabeled data in cnns by self-supervised learning to rank[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1 862-1 878.
[20] KALYANI G, JANAKIRAMAIAH B, PRASAD L V, et al. Efficient crowd counting model using feature pyramid network and ResNeXt[J]. Soft Computing, 2021, 25(15): 10 497-10 507.
[21] MA Zhiheng, WEI Xing, HONG Xiaopeng, et al. Bayesian loss for crowd count estimation with point supervision[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6 142-6 151.
[22] WANG Boyu, LIU Huidong, SAMARAS D, et al. Distribution matching for crowd counting[J]. Advances in Neural Information Processing Systems, 2020, 33: 1 595-1 607.
[23] JIANG Xiaolong, XIAO Zehao, ZHANG Baochang, et al. Crowd counting and density estimation by trellis encoder-decoder networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 6 133-6 142.
[24] LIU Weizhe, SALZMANN M, FUA P. Context-aware crowd counting[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5 099-5 108.
[25] XIONG Haipeng, LU Hao, LIU Chengxin, et al. From open set to closed set: counting objects by spatial divide-and-conquer[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 8 362-8 371.
[26] LIU Lingbo, QIU Zhilin, LI Guanbin, et al. Crowd counting with deep structured scale integration network[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1 774-1 783.

作者简介: 冉瑞生 (1976—),男,重庆人,博士,教授,主要从事机器学习、计算机视觉等方面的研究。E-mail: rshran@ cqnu.edu.cn。