结合单列多列神经网络的移动状态人群计数方法研究

温宇健[1,2] 郭士杰[2]

[1](复旦大学工程与应用技术研究院 上海 200433)

[2](复旦大学智能机器人教育部工程研究中心 上海 200433)

摘 要 已有人群计数方法局限于对人群的全部进行计数,在仅对人群中的移动者进行计数时准确率较低,基于注意力的多阶段深度学习框架被提出以解决这一问题。通过注意力机制适应性地在单列和多列计数网络进行选择,结合单列网络的深层特征表示能力和多列网络多尺度特征学习能力,有效提取人群中移动者的特征。实验结果表明,所提出的方法均方误差(MSE)和平均绝对误差(MAE)皆低于已有人群计数方法,能够有效提高处于移动状态的人群的计数精度。

关键词 人群计数 深度学习 单列多列网络 注意力机制

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.06.029

MOVING CROWD COUNTING BY INTERGRATING SINGLE AND MULTIPLE COLUMN NEURAL NETWORK

Wen Yujian[1,2] Guo Shijie[2]

[1](*Academy for Engineering and Technology*, *Fudan University*, *Shanghai 200433*, *China*)

[2](*Engineering Research Center of AI & Robotics*, *Ministry of Education*, *Fudan University*, *Shanghai 200433*, *China*)

Abstract Existing crowd counting methods are limited to counting the integrity of the crowd, the accuracy rate is downgraded when exclusively counting the moving people in the crowd. An attention based multi-stage deep learning framework is proposed to solve this problem. Attention module was adopted to adaptively selects both single-column and multi-column counting networks, combine the deep features of single column network and the multiple scale receptive fields of multiple column network, which effectively extracted features of the moving people. The results show that the proposed method has lower mean square error (MSE) and mean absolute error (MAE) than existing crowd counting methods. The counting accuracy of people on moving is well improved.

Keywords Crowd counting Deep learning Single and multiple column network Attention mechanism

0 引 言

人群计数是指计数一幅图片中人的数量,是计算机视觉量化理解人群信息的重要方法,多采用人群密度图(Density Map)[1-2]进行计数,根据网络框架结构,可分为两类[3]。一类为基于多列(Multi-column)的方法[4-6],通过增加网络的宽度来获取多维度的局部信息;另一类为基于单列(Single-column)的方法[7,11],通过增加网络深度来提高对高维度特征的学习能力,从而更好地估计全局人群密度图。人群计数作为获取人群信息的重要方法,在公共安全和城市智能领域受到了越来越多的关注。

在此基础上,我们提出移动人群计数问题,计数图片中处于移动状态的人群,进一步挖掘图片中的流动信息。例如,旅游景区需要知道景区内活动和休息的旅客数量,以便确定各个分区的人口流动导向;服务提供商需要知道移动中的顾客数量,以便对他们提供更为周到的服务;密集的公共区域也需要统计流动和静止的人数,以便按需要疏导人群;服务机器人需要感知

收稿日期:2021-01-07。国家重点研发计划项目(2016YFE0128700);河北省重点研发计划项目(18211816D)。**温宇健**,硕士生,主研领域:计算机视觉。**郭士杰**,教授。

人群的动态状态,以便实时做出行为决策和规划动作。

在现实人口密集的场景中,受限于遮挡、透视效果、物体形状的变化,人群计数的错误率会大大升高。移动人群计数需要进一步区分移动人群,对计数能力提出了更高的要求。为了提高对于密集场景的计数能力,算法对人物重叠和遮挡问题的处理很重要,也是最新的方法着重解决的方向。比如,Ma 等[12]提出贝叶斯损失函数,以点注释来实施更可靠的密度贡献概率模型;Wang 等[13]指出将高斯函数强加于注释会损害泛化性能,使用最优运输(OT)的非像素级对比方式来描述测量的相似性。这两种方式采用非高斯损失函数的方式缓解了人群重叠问题,降低了测量误差。但是其单一的深度结构限制了对移动人群检测的泛化能力。

本文在采用文献[13]的最优运输损失函数基础上,进一步提出一种结合多列与单列的网络构架,采用深度学习的方法,融合二者的优势,在全局密度信息中提取局部姿态信息,解决上述方法针对移动人群计数精度较低的问题。本网络由以下四段组成:(1) 使用姿势估计网络检测、提取人群移动的姿势特征,并将其输入密度图估计网络,生成初步的移动人群密度图。姿态估计网络可以检测出人体关节点信息,通过关节点的可见性与关节点间相对位置,识别人群中处于移动姿态的人。(2) 利用单列深度网络在特征表示上的优势,学习人群总体密度图,以便应对不同场景下的人群规模差异。(3) 利用多列网络结合全局和局部密度图特征。(4) 利用注意力机制分配单列和多列的权重,生成最终的移动人群密度图。

多阶段的网络模型能使网络的每个阶段单独训练,从而让网络更易于调整参数和训练。借助这一特性,本文与文献[12-13]采用相同的网络 vgg19 后,继续加深网络的整体深度,在具有更强表示能力的同时仍然使模型易于训练,进一步优化实验结果。实验结果表明,本文的方法具有更高的移动人群检测精度和更小的计数误差。

图 1　移动人群计数问题图片

1　相关概念

人群计数问题最早由行人检测问题衍化而来,并通过检测方法实现。但在人群密度高、遮挡严重的场景中,基于检测的方法存在速度慢、精度低的问题。为解决这一问题,基于密度图的方法被引入人群计数,并根据网络体系结构的不同分为基于单列和多列的方法[3],本节分别介绍这三种人群计数方法。

1.1　检测和回归

人群计数较早应用在视频行人检测问题,文献[14]提出一个检测框架,基于增强的运动特征,用一个检测器在视频序列的两个连续帧上进行扫描,从而获得行人的数量。后续工作通过不断提出鲁棒性更强的检测器来更准确地检测行人,但是检测器不能很好地应对遮挡等检测问题。基于回归的方法[15]通过区分前景和后景,并用各种方式提取多种前景特征输入到回归方程来估计人群数量。可以看出,基于检测的方法更多地关注图片的局部信息,基于回归的方法核心思想在于寻找更多的特征进行结合。前者更加适合行人检测的检测问题,后者更加适合视频等固定环境下的特征信息提取,都无法适应环境差异很大的现实场景。

1.2　多列宽度结构

Fu 等[2]第一次把卷积神经网络应用于人群计数问题,利用卷积神经网络的卷积层,池化层以及全连接层,降低了对图片以外额外信息的需求。多列卷积神经网络通过并行多个 CNN 获取多尺度信息,扩展宽度在多个接收域提取信息。比如文献[3]使用三个不同大小核心的多列结构来分别在高中低获取信息;CrowdNet[4]通过深浅两个列网络,分别获取高维度语义信息和低纬度特征信息;TDF-CNN[5]使用自底向上的网络结构来修正密度图预测。宽度网络每一列能够与其他列交互合作,获取多种类型特征。但是会带来冗余和复杂度,所以要合理控制列的数量减少重复结构。在移动人群计数问题中,人体的姿态信息能够为全局密度图提供人体的行动信息,全局特征能够为局部检测提供人群的分布信息,宽度网络能够结合两者的特征。

1.3　单列深度结构

不同于多列结构在提取广度特征时候所带来的冗余复杂度,通过增加单列网络深度也能够得到高维度的特征信息,并且形式更加紧凑简洁。由于其形式的精简和结果的准确性,单列深度学习在人群计数问题上获得了越来越多的关注。比如 CSRNet[8]利用空洞卷积层,通过扩大感受域来估计更多的多尺度上下文信息;W-net[7]使用了 U-Net[20]的结构,增加了一个

加强分支来加速深度网络的收敛速度并保持局部结构的一致性,同时使用结构相似性指数(SSIM)来估计最终密度地图;SANet[9] 和 ADCrowdNet[10] 也都分别把 Inception 和注意力机制加入到深度网络中以提高模型的准确性。

在移动人群计数问题中,深度网络间接的网络结构和强大的学习能力能够提取到移动人群的全局特征信息。

2 单列多列模型结合的移动人群计数网络

移动人群计数不仅要检测到每一个人,还要检测人的移动或静止的状态,增加了问题的复杂性。整体结构如图 2 所示,每一阶段的结构如图标注,各个阶段可以单独训练。

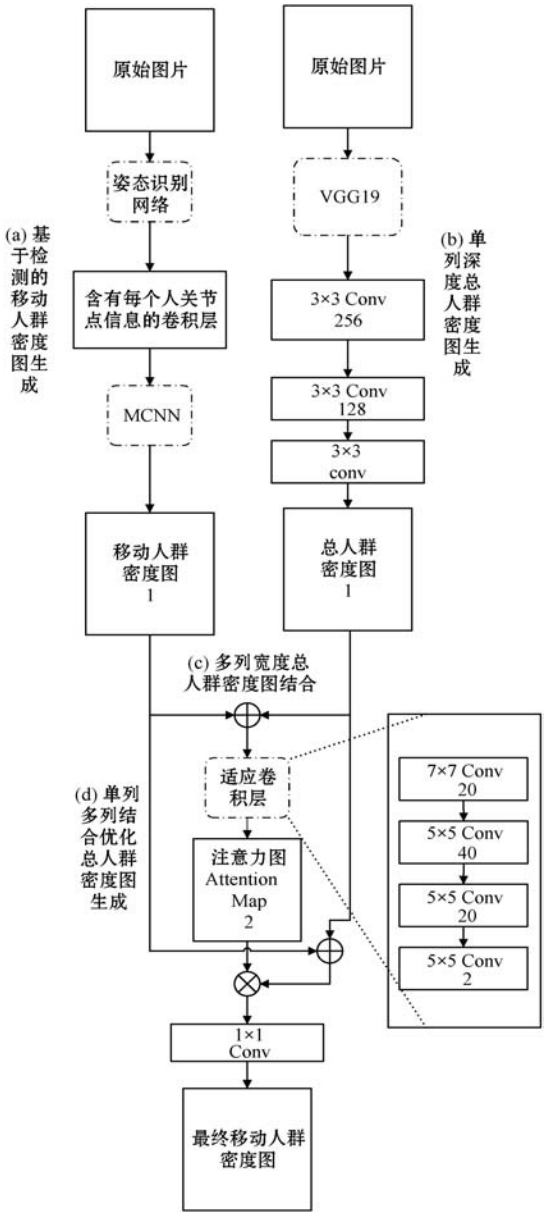

图 2 网络结构

2.1 基于检测的初步移动密度图生成

第一阶段使用人体姿态检测器检测人的移动信息,将其输入密度图生成网络预测移动人群密度图。如图 2(a)所示,我们选用自顶向上的姿态识别器 OpenPose[16],在图片人数较多时仍能够在短时间内获得人不同关节点的位置信息,从而满足高密度图中人数规模较大的情况。我们选用鼻子的关节点作为额外信息,和关节点位置信息一起输入密度图生成网络 MCNN[3] 来预测最终的人群密度图。

实验中我们发现,图片中距离很近的人往往会被同时归类于移动者,因为高斯人群密度图为了平滑而让相邻的人检测发生重叠,所以我们选用最优传输 OT 损失函数[13] 代替 MCNN 原有的像素级 MSE 损失。

最优传输(Optical Transport)代表两个分布($\boldsymbol{\mu}$,$\boldsymbol{\nu}$)之间进行转化的最小消耗 $\boldsymbol{C}$:

$$W(\boldsymbol{\mu},\boldsymbol{\nu}) = \min_{\gamma \in \Gamma} \langle \boldsymbol{C},\gamma \rangle \tag{1}$$

式中:Γ 代表所有的转化方式。由此,我们定义 OT 损失函数为密度图真实分布 $\boldsymbol{d}$ 和预测分布 $\hat{\boldsymbol{d}}$ 之间的转化消耗。

$$\ell(\boldsymbol{d},\hat{\boldsymbol{d}}) = W(\boldsymbol{d},\hat{\boldsymbol{d}}) \tag{2}$$

最优传输损失函数采用非像素级对比方式来测量两个分布之间的距离,通过消除紧邻的两个人的高斯分布像素重叠现象,有效解决了移动人群计数中相邻人同时被检测成移动状态的现象。

2.2 基于单列深度网络生成总人群密度图

第二阶段,利用结构简单但学习能力强大的单列深度网络提取高层特征,来估计总人群的密度图。大量实验[12,17] 表明,深度网络能够很好地适应不同人群规模和人群分布,获取的总人群密度可以为后续移动人群提供全局分布信息。图 2(b)为本阶段的学习模型。我们采用深度网络 VGG19[18] 作为骨干网络,将最后一个池化层和完全连接层删除后将其输出输入到两个分别具有 256 和 128 通道的 3×3 卷积层和一个 1×1 卷积层,以获得总人群密度图。

2.3 基于多列宽度网络生成最终移动人群密度图

多列网络结构通过增加并行的网络管线并在最后结合来处理不同网络管道中的有用信息,我们受文献[6]启发,采用了如图 2 网络结构(c)所示的 3 列并行的网络结构。第一阶段已经获得了基于检测的移动密度图,但是在高密度图像中表现不佳,我们通过并行第二阶段的单列深度网络管线,并在本阶段进行结合,从而把第二阶段的全局人群分布信息加入到第一阶段的

输出中,进一步改善移动行人的密度图。

2.4　单列多列网络的结合优化

本文根据单列多列网络的结合特性在文献[6]的基础上进行了优化。由于第二阶段总人群计数的网络比第一阶段初步移动人群网络更深,简单地连接这些网络会导致训练逐渐倾向于某一边,我们采用了注意力机制在两条线上进行挑选来替换原网络的决定机制,用以优化单列和多列的结合,降低网络的训练难度。

与文献[6]中的 quality-net 结构不同,我们的适应卷积层采用四层卷积神经网络,除了第一层为 7×7,其他都为5×5,我们通过注意力机制生成一个2 层注意力图,每一层分别第一阶段的初步移动人群密度图和第二阶段的总人群密度图进行对应元素逐个相乘(Element-wise Multiplication),输入到一个单个通道 1×1 过滤器以获得最终移动人群密度图。我们仍然使用图 2 网络结构中介绍的 OT 作为损失函数。本阶段使用两个单独的注意力图为多列网络提供了很大的灵活性,使其能够灵活地在全局和移动人群检测图中做出选择,获取更多的移动人群特征。

3　实　验

3.1　数据集

训练第一阶段和第三阶段时由于现有的数据集都是对图片中所有的人进行标注,没有单独对移动中的人进行标注,于是我们从公共数据集中收集了 150 幅人群图像,并对其标注进行修改使其只包含移动人群,标注的位置位于移动者的头部。图 1 是几幅数据集中的图片。

在训练第二阶段的单列深度网络时,我们使用 ShanghaiTeck 数据集[3]来进行训练。

在训练和验证期间,我们通过水平翻转进一步扩充数据。

为了进行对比实验,我们从 ShanghaiTeck 数据集以及互联网收集了 80 幅只含有移动人群的测试图像并对其进行同样的标注,总人数 766 人,最少的一幅图有 1 人,最多的一幅图 55 人。本数据集并不用作训练,只用作对比实验的测试图像。

3.2　训练步骤

第一阶段结构相对简单,层数较少,我们使用训练好的姿态估计网络 OpenPose,提取图片中人体 16 个关节点信息,输入到 MCNN 网络中。训练的过程我们使用 Adam 优化器[19]以 10^{-4}作为学习率,批量(batch)为 32 进行训练。我们用 80% 的数据用于训练,15% 的数据用于验证,5% 的数据用于测试。每次进行 1 000 次的迭代训练(epoch),保存在验证集中表现最好的模型作为本阶段的输出。

第二阶段我们使用在 ImageNet 先预训练好的骨干网络 VGG19[20],后面的卷积层使用 MSRA[21]的初始化方式。我们对单列深度网络以 10^{-5}的较小学习率进行学习,并使用 Adam 优化器进行优化。

第三和第四阶段,文献[6]指出多列网络比单列网络更容易陷入局部最小,与其使用随机初始值,对模型进行预训练可以达到更好的效果。在文献[6]中,DecideNet 使用高密度的 Part_A 作为预训练数据集,但对于移动人群计数问题,没有一个可以提供预训练的数据集,所以我们也采用 ShanghaiTeck Part_A 作为预训练数据集并随机对人群进行标注,由于 Part_A 本身多为难以分辨移动静止的高密度图,所以我们仅用来预训练使模型对整体分布有较好的感知不至于提前收敛到局部最小值。具体的参数设置同第二阶段。

3.3　性能对比

由于以往人群计数方法没有考虑针对移动人群计数的问题,无法直接和本文的结果进行比较,我们采用两种方式进行比较。

对比的模型有:MCNN[3]、BLNet[12]、SPANet[17]。其中,MCNN 作为多列网络模型的代表,BLNet 作为单列深度网络模型的代表,通过对比单个单列、单个多列、单列多列结合的方式,我们可以考量不同网络的结合方式是否有效;SPANet 是近期高效的人群计数算法,通过注意力机制和特征融合方法,为该方向的发展提供了灵感。

如表 1 所示,第一种方式我们使用类似在第三阶段预训练的策略,取其他方法在公共数据集训练好的模型,使用我们的移动人群数据集,在其模型基础上再次针对这特定任务进行训练以调整网络。基于此,我们可以考量其他方法在针对具体人群计数任务上是否具有泛化学习能力。评估方法上我们遵循传统的平均绝对误差(MAE)和均方误差(MSE)方法。

表 1　不同算法实验结果指标

算法	MAE	MSE
MCNN	21.60	30.51
BLNet	13.64	17.56
SPANet	12.97	18.29
本文第一阶段	10.44	16.57
本文方法	6.83	10.20

实验结果显示,在我们的数据集中,我们的方法取得了最好的成绩。

除此之外,我们还对比我们第一阶段单纯使用姿态检测器进行移动人群计数的结果,结果显示第一阶段有一定作用,后续阶段能够进一步降低误差,表明我们后续多阶段模型的有效性。

表 2 显示了各个现有方法对于移动人群检测的精确率和召回率,虽然各个现有方法有较好的召回率,但是精确率较低,这也解释了表 1 其他方法误差率较大的原因。其他方法对全局所有人的检测较为准确,但是无法准确地区分移动和静止的人。

表 2 不同算法对移动人群的精度和召回(%)

算法	精度	召回
MCNN	33.0	72.3
BLNet	41.1	83.7
SPANet	40.5	83.2
本文第一阶段	77.2	76.0
本文的方法	86.1	86.9

各个方法生成的移动人群密度图如图 3 所示,其中用头的位置代表人的位置。

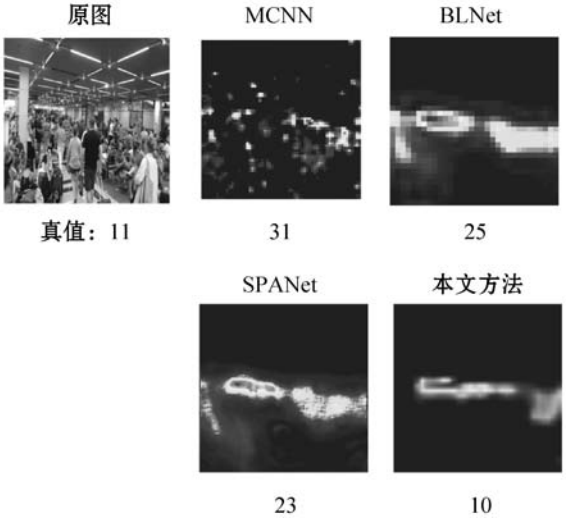

图 3 不同方法生成的移动人群密度图

可以看出,只有本文方法准确估计了移动者头的位置,其他三种方法都在不同程度上误判了两边坐着的人。

MCNN 基于多列宽度方法在四种方法中结果最差。相比于基于其他两种基于单列深度的网络,MCNN 只通过简单增加网络的列数来获取更多的特征,从而难以应对多变的环境以及复杂的任务,泛化能力较差。使用了单列深度骨干网络的 BLNet,相对于基于多列的 MCNN 方法有较好的结果,说明单列的方法受益于其简单的结构,具有相对较好的泛化能力。

第二种对比方式如表 3 所示,我们选取了 80 幅只有移动者的人群密度图,由于数据集人数偏少,结果差距较小。我们方法的误差仍然小于最新的研究结果,表明这种单列多列结合的网络结构不仅能够有效解决移动人群计数问题,在传统人群计数问题中得益于结构优势,仍然具有竞争力。后续我们也会把本文方法应用于传统人群计数,并简化网络。

表 3 不同算法在只有移动者数据集的误差

算法	MAE	MSE
MCNN	10.33	16.25
BLNet	4.20	7.98
SPANet	4.29	8.47
本文第一阶段	9.83	14.25
本文方法	4.15	7.94

4 结 语

本文针对现有人群计数方法在应用于计数处于移动状态人群时识别精度低的问题,提出一种多阶段深度学习网络框架,将姿态特征输入网络,并充分结合单列深度网络的高维度特征学习能力和多列宽度网络的在不同维度局部特征选择的灵活性,提高移动状态人群的计数精度,解决最新方法泛化能力不足的问题。同时针对高密度人群图片移动和静止者聚集在一起难以区分的问题,利用最优传输 OT 损失函数进一步提高识别精度。

实验结果显示,本文方法相对于其他最新研究结果,在区分移动和静止状态人群上具有更好的结果,降低了误差。

参 考 文 献

[1] Wang C, Zhang H, Yang L, et al. Deep people counting in extremely dense crowds[C]//23rd ACM International Conference on Multimedia,2015:1299 - 1302.

[2] Fu M, Xu P, Li X D, et al. Fast crowd density estimation with convolutional neural networks[J]. Engineering Applications of Artificial Intelligence,2015,43:81 - 88.

[3] Zhang Y, Zhou D S, Chen S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]// IEEE Conference on Computer Vision and Pattern Recognition,2016:589 - 597.

[4] Boominathan L, Kruthiventi S, Babu R V. Crowdnet: A deep convolutional network for dense crowd counting[C]//

24th ACM International Conference on Multimedia,2016:640 – 644.

[5] Sam D B, Babu R V. Top-down feedback for crowd counting convolutional neural network[EB]. arXiv:1807.08881,2018.

[6] Liu J, Gao C Q, Meng D Y, et al. DecideNet: Counting varying density crowds through attention guided detection and density estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition,2018:5197 – 5206.

[7] Valloli V K, Mehta K. W-Net: Reinforced u-net for density map estimation[EB]. arXiv:1903.11249,2019.

[8] Li Y H, Zhang X F, Chen D M. CsrNet: Dilated convolutional neural networks for understanding the highly congested scenes[C]//IEEE Conference on Computer Vision and Pattern Recognition,2018:1091 – 1100.

[9] Cao X K, Wang Z P, Zhao Y, et al. Scale aggregation network for accurate and efficient crowd counting[C]//European Conference on Computer Vision,2018:734 – 750.

[10] Liu N, Long Y C, Zou C Q, et al. AdcrowdNet: An attention-injective deformable convolutional network for crowd understanding[C]//IEEE Conference on Computer Vision and Pattern Recognition,2019:3225 – 3234.

[11] Hossain M A, Hosseinzadeh M, Chanda O, et al. Crowd counting using scale-aware attention networks[C]//IEEE Winter Conference on Applications of Computer Vision, 2019:1280 – 1288.

[12] Ma Z H, Wei X, Hong X P, et al. Bayesian loss for crowd count estimation with point supervision[C]//IEEE International Conference on Computer Vision,2019:6142 – 6151.

[13] Wang B Y, Liu H D, Samaras D, et al. Distribution matching for crowd counting[C]//34th International Conference on Neural Information Processing Systems, 2020: 1595 – 1607.

[14] Viola P, Jones M J, Snow D. Detecting pedestrians using patterns of motion and appearance[J]. International Journal of Computer Vision,2005,63(2):153 – 161.

[15] Chan A B, Liang Z S, Vasconcelos N. Privacy preserving crowd monitoring: Counting people without people models or tracking[C]//IEEE Conference on Computer Vision and Pattern Recognition,2008:1 – 7.

[16] Cao Z, Hidalgo G, Simon T, et al. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,43(1):172 – 186.

[17] Zhu L, Zhao Z J, Lu C, et al. Dual path multi-scale fusion networks with attention for crowd counting[EB]. arXiv: 1902.01115,2019.

[18] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB]. arXiv:1409.1556,2014.

[19] Kingma D P, Ba J. Adam: A method for stochastic optimization[EB]. arXiv:1412.6980,2014.

[20] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:1 – 9.

[21] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention,2015:234 – 241.

(上接第 185 页)

[11] Huang G, Liu Z, Laurens V D M, et al. Densely connected convolutional networks[C]//IEEE Conference on Computer Vision and Pattern Recognition,2017:2261 – 2269.

[12] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]//European Conference on Computer Vision,2018:3 – 19.

[13] Chaudhuri B, Demir B, Chaudhuri S, et al. Multilabel remote sensing image retrieval using a semi-supervised graph-theoretic method[J]. IEEE Transactions on Geoscience and Remote Sensing,2018,56(2):1144 – 1158.

[14] Wang J, Kumar S, Chang S F. Sequential projection learning for hashing with compact codes[C]//27th International Conference on International Conference on Machine Learning,2010:1127 – 1134.

[15] Liu W, Wang J, Ji R, et al. Supervised hashing with kernels[C]//IEEE Conference on Computer Vision and Pattern Recognition,2012:2074 – 2081.

[16] Shen F M, Shen C H, Liu W, et al. Supervised discrete hashing[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:37 – 45.

[17] Cao Y, Long M S, Liu B, et al. Deep Cauchy hashing for hamming space retrieval[C]//IEEE Conference on Computer Vision and Pattern Recognition,2018:1229 – 1237.

[18] Zhu H, Long M S, Wang J M, et al. Deep hashing network for efficient similarity retrieval[C]//30th AAAI Conference on Artificial Intelligence,2016:2415 – 2421.

[19] Liu B, Cao Y, Long M S, et al. Deep triplet quantization [C]//ACM Multimedia Conference on Multimedia, 2018: 755 – 763.

[20] Cao Y, Long M S, Wang J M, et al. Deep quantization network for efficient image retrieval[C]//30th AAAI Conference on Artificial Intelligence,2016:3457 – 3463.

[21] Moustafa M, Ahmed S, Hamed A. Learning to hash with convolutional network for multi-label remote sensing image retrieval[J]. International Journal of Intelligent Engineering and Systems,2020,13(5):539 – 547.