基于多粒度特征融合网络的行人重识别

匡 澄,陈 莹
(江南大学轻工过程先进控制教育部重点实验室,江苏无锡 214122)

摘 要: 行人重识别旨在跨监控设备下检索出特定的行人目标.为捕捉行人图像的多粒度特征进而提高识别精度,基于OSNet基准网络提出一种多粒度特征融合网络(Multi-granularity Feature Fusion Network for Person Re-Identification, MFN)进行端对端的学习.MFN由全局分支、特征擦除分支和局部分支组成,其中特征擦除分支由双通道注意力擦除模型构成,此模型包含通道注意力擦除模块(Channel Attention-based Dropout Moudle, CDM)和空间注意力擦除模块(Spatial Attention-based Dropout Moudle, SDM). CDM对通道的注意力强度排序并擦除低注意力通道,SDM在空间维度上以一定概率擦除最具有判别力的特征,两者通过并联方式相互作用,提高模型的识别能力.全局分支采用特征金字塔结构提取多尺度特征,局部分支将特征均匀切块后级联成一个单一特征,提取关键局部信息.大量实验结果表明了本文方法的有效性,在Market1501、DukeMTMC-reID和CUHK03-Labeled(Detected)数据集上,mAP/Rank-1分别达到了90.1%/95.8%、81.8%/91.4%和80.7%/82.3%(78.7%/81.6%),大幅优于其他现有方法.

关键词: 行人重识别;多分支CNN网络;金字塔结构;特征擦除
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2021)08-1541-10
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200974

Multi-granularity Feature Fusion Network for Person Re-Identification

KUANG Cheng, CHEN Ying
(*Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi, Jiangsu* 214122)

Abstract: For the purpose of capturing the multi-granularity features and improving the recognition accuracy, a multi-granularity feature fusion network for person re-identification (MFN) is proposed based on the omist-scale network (OSNet). The MFN network is composed of a global branch, a feature dropout branch and a local branch. The feature dropout branch consists of a dual-channel attention dropout model, which includes a channel attention-based dropout moudle (CDM) and a Spatial attention-based dropout moudle (SDM). CDM sorts the attention intensity and dropouts low attention channels, and SDM dropouts the most discriminative features with a certain probability in the spatial dimension. The global branch uses the feature pyramid structure to extract multi-scale features, and the local branch employs a uniform partition strategy to produce local features which are cascaded into a single one for key local information extraction. Experiments on the large scale datasets show the effectiveness of MFN. On the Market1501, DukeMTMC-reID and CUHK03 -Labeled (Detected) datasets, mAP/Rank-1 of MFN reaches 90.1%/95.8%, 81.8%/91.4% and 80.7%/82.3% (78.7%/81.6%), which is superior to other existing methods.

Key words: person re-identification; multi-branch CNN network; pyramid structure; feature dropout

1 引言

行人重识别(Person Re-Identification)也称为行人再识别,旨在从监控环境中不同摄像头下识别同一行人的过程.近年来,行人重识别技术在公共安全、视频监控等方面有着积极的作用,具有重要的研究意义,也成为计算机视觉领域中的研究热点.但是不同图片之间存在光照、姿态、遮挡、视角等问题,且监控图片的分辨率低,导致行人重识别仍然很具有挑战性.

行人重识别的研究[1]分为传统方法和深度方法.传统的行人重识别任务包含特征提取和相似度度量两个步骤.特征提取的目的是提取具有辨别力且鲁棒性强的特征表达,如颜色、HOG、SIFT等.相似度度量的目

收稿日期:2020-09-03;修回日期:2021-01-22;责任编辑:王天慧
基金项目:国家自然科学基金(No. 61573168)

的设计度量函数类内距离越小使类间距离越大，如LMNN[2]、XQDA[3]等．然而传统方法学习能力有限，很难适应大数据量任务，目前，基于深度学习的行人重识别方法在性能上远远超过传统方法．

基于深度学习的行人重识别可以分为基于全局特征的方法和基于局部特征的方法．简单的方法是使用基于CNN方法来学习整张图像上的全局特征表示，但是这种方法会带来两个问题：(1)模型只能提取单尺度特征，这意味着会忽视一些分布在不同层上的细粒度语义信息；(2)由于检测技术的限制，行人身体部位存在着未对准的问题，影响检索结果．与全局特征模型不同，局部特征模型更加关注诸如姿态、人体部件等关键区域，基于图像切块是目前常用的提取局部特征的思路．PCB[4]模型将行人图像水平分割提取人体抽象部件；Wang等人[5]认为PCB模型忽视了整体对局部学习的影响，从而提出了多粒度模型MGN；考虑特征对齐问题，CCPAR通过挖掘行人部位特征通道间的互相关系[6]，有效优化部分特征；Zheng等人[7]提出了由粗粒度到细粒度的渐进式金字塔模型Pyramid，取得了不错的效果．但是图像切块的方法仍然存在着特征之间的对齐问题，会导致信息的丢失，同时计算量巨大，在真实的场景会带来更大的误差．

Dropout作为数据增强且避免过度拟合的方式，在训练过程中随机擦除隐藏神经元的输出，从而使神经网络学习并提取更多的特征．在输入环节，Cutout[8]和Random Erasing[9]都主张通过部分遮挡现有样本有效增强数据集．在中间特征图环节，对于一批输入张量，DropBlock[10]随机擦除每个张量的连续区域；Spatial Dropout[11]将输入特征图的整个通道随机置零．作为一种经典的Dropout方法，Batch DropBlock(BDB)随机擦除一批输入特征图的相同区域，加强次显著特征的学习，在行人重识别领域取得了巨大的成功[12]．但是BDB网络由于特征擦除的随机性，往往只能得到次优的结果，同时擦除区域的宽高比、擦除概率在很大程度上也影响着实验的精度．

针对上述方法存在的问题，本文提出基于多粒度特征融合的行人重识别方法(Multi-granularity Feature Fusion Network for Person Re-Identification，MFN)．与大多数文献使用Resnet、Vgg网络不同，本文以轻量级网络Omni-Scale Network(OSNet)[13]为基准网络，提出了由全局分支、特征擦除分支和局部切块分支组成的网络体系结构．三条分支相互作用，相互促进，提取更加丰富的多粒度特征．基于特征金字塔结构，全局分支提取不同尺度特征；特征擦除分支由双通道注意力特征擦除(Dual Channel Attention-based Dropout，DCAD)模型构成，此模型包含通道注意力擦除模块(Channel Attention-based Dropout Moudle，CDM)和空间注意力擦除模块(Spatial Attention-based Dropout Moudle，SDM)．CDM对通道的注意力强度排序并擦除低注意力通道，SDM在空间维度上自适应擦除最具有判别力的特征；局部切块分支将均匀切块后的局部特征级联进行单一的损失计算，减少因分割不均匀带来的信息损失．最后在Market1501[14]、DukeMTMC-reID[15]和CUHK03[16]数据集上的实验结果表明了本文方法的有效性，同时在mAP、Rank1两项指标上超越现有方法．

2 本文网络

在本节分3个部分介绍基于多粒度特征融合行人重识别网络(MFN)，首先介绍网络的整体架构；然后具体介绍网络的多尺度全局分支、特征擦除分支和局部切块分支；最后介绍网络中使用的损失函数．

2.1 网络结构

图1是本文的网络结构图，该网络以行人图片作为输入，使用OSNet作为主干网络提取图片的特征．网络包含1层卷积层(Conv1)、3个残差模块(Conv2~Conv4)和1个1*1卷积模块(Conv5)，每个残差模块包含多层卷积层、平均池化层、批量规范层和线性整流函数(Rectified Linear Units，ReLU)．同时，在3个残差模块两两之间加入通道注意力模块(Channel Attention Moudle，CAM)[17]和空间注意力模块(Spital Attention Moudle，SAM)[17]，在第二、三个分支卷积模块Conv5后加上SAM．将大小为256×128的图片输入网络，可以从Conv5模块输出大小为16×8的特征图．

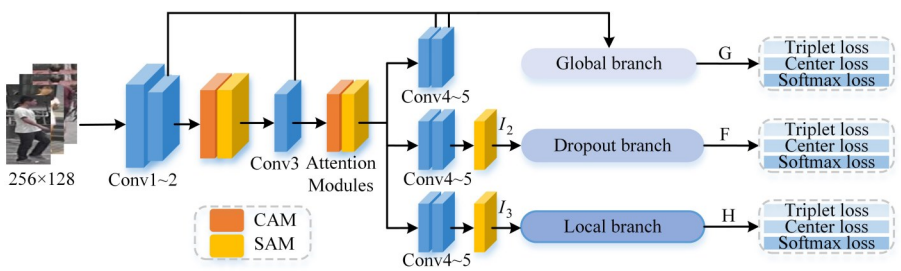


图1 网络结构图

为了减轻任务间的干扰,在第二个SAM后将网络分成3个分支,分别是全局分支、特征擦除分支和局部切块分支. 全局分支将Conv2~Conv5输出的特征图作为特征金字塔结构的输入,提取行人的多尺度特征信息;通过第三个SAM得到I_2输入特征擦除分支,该分支由双通道注意力特征擦除模型(DCAD)构成,该模型包含通道注意力擦除模块(CDM)和空间注意力擦除模块(SDM),CDM擦除注意力强度低的通道,SDM按照一定的比例擦除最显著的特征,二者强迫网络关注次显著的部分,增强感受野;局部切块分支将行人特征图片I_3分块成若干块后级联成单一特征,提取关键局部信息,提高行人重识别的准确率.

2.2　分支结构

2.2.1　多尺度全局分支

特征金字塔(Feature Pyramid)可以检测不同尺度物体,文献[18]提出了一种具有横向连接的自上而下的特征金字塔结构,用于在所有尺度上构建高级语义特征映射,在COCO[19]数据集上取得了不错的准确率. 本文将该特征金字塔模型引入行人重识别领域,在OSNet网络上提出多尺度全局分支,旨在学习多尺度语义信息,提取行人有区别性的特征,其结构图如图2所示.

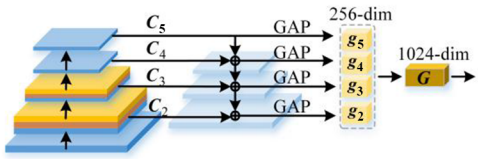


图2　多尺度全局分支结构图

全局分支从上而下构建金字塔,通过将深层的特征图上采样至前一层特征图,按元素进行加法合并后提取特征. OSNet网络一共有5层,每层的输出用$\{C_1, C_2, C_3, C_4, C_5\}$来表示($C_2$代表残差模块Conv2的输出,以此类推),除去第一层的输出外(由于第一层提取的特征数较少,故不参加上述操作),其余四层输出进行上采样(unsamlpe)和逐元素相加操作,而后使用全局平均池化(Global Average Pooling, GAP). 相应过程如下

$$C_i' = \text{unsample}\{C_i, \text{'blinear'}\}\ i = 3, 4, 5 \tag{1}$$

$$\begin{cases} g_{i-1} = \text{GAP}(C_i' \oplus C_{i-1}) & i = 3, 4, 5 \\ g_i = \text{GAP}(C_i) & i = 5 \end{cases} \tag{2}$$

其中,C_i通过双线性插值得到C_i',其长宽与特征图C_{i-1}'相同,$\oplus$表示逐元素相加. 其余层输出$\{g_2, g_3, g_4, g_5\} \in R^{256\times1\times1}$重复上述操作得到,将4张特征图级联,通过1024维的BN层、连接层得到最终特征$G \in R^{1024\times1\times1}$.

2.2.2　特征擦除分支

大量实验证明,适当使用Dropout方法可以有效避免过拟合,提高算法的准确率. 为了加强局部区域的特征学习,BDB网络(Batch DropBlock Network)[12]批量随机删除特征图的相同区域. 相比于BDB的随机性,文献[20]提出的ADL(Attention-based Dropout Layer)以一定的概率删除特征图中最显著部分,加强次显著部分的特征学习的方法则显得更为直接有效. 但是ADL模块仅仅关注在空间方向上的擦除,忽视了特征图在通道上也有很强的相关性,适当地擦除一些通道有助于网络学习到更多的细粒度特征.

对此,本文的特征擦除分支由一种轻巧而功能强大的双注意力特征擦除模型构成,如图3所示,它包含两个关键组件,即通道注意力擦除模块(CDM)和空间注意力擦除模块(SDM),两模块以并联的方式进行连接. CDM擦除掉一些重要性不高的通道,SDM根据注意区域自适应地擦除最具鉴别性的特征,两者以并联的方式连接,相互促进、相互作用,最后输出特征图F进行相应损失计算,增强网络的学习能力.

在通道注意力擦除模块中,对于第二个分支特征图$I_2 \in R^{512\times16\times8}$,使用全局平均池化(GAP)收集特征图的通道信息$S \in R^{512\times1\times1}$. 模型的判别力与每个像素的强度成正比,因此$S$可以被认为特征图在通道上注意力

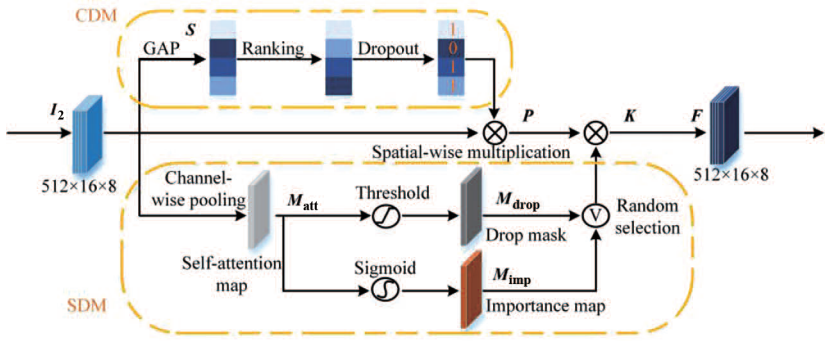


图3　特征擦除网络

强度的集合. 根据通道注意力的相对大小进行排序,排名越靠前的通道包含的具有判别力的特征就越多,因此使用二进制掩码置零排序靠后的c个通道,让网络尽可能地注意强度高的通道. 最后二进制掩码图$\boldsymbol{P}$与原始特征图$\boldsymbol{I}_2$矩阵乘后进行后续运算.

空间注意力擦除模块中,特征图$\boldsymbol{I}_2 \in \boldsymbol{R}^{512\times16\times8}$通过通道平均池化生成空间自注意力图$\boldsymbol{M}_{\text{att}}$(Self-attention map). 擦除模板$\boldsymbol{M}_{\text{drop}}$(drop mask)通过设置一个阈值$\alpha$,"惩罚"$\boldsymbol{M}_{\text{att}}$中最有判别力的部分,即当特征像素大于$\alpha$时置零,强迫网络学习次显著的特征,表达式如式(3)所示,其中(x,y)表示特征图中的像素点位置.

$$\boldsymbol{M}_{\text{drop}}^{x,y}=\begin{cases}0, & \boldsymbol{M}_{\text{att}}^{x,y}>\alpha\\ 1, & \text{otherwise}\end{cases} \tag{3}$$

虽然$\boldsymbol{M}_{\text{drop}}$的使用能够使网络学习次显著特征,但如果每次迭代中都应用了$\boldsymbol{M}_{\text{drop}}$,那么模型将不会观察到最具有判别力的特征. 因此,论文引入重要性图(Important map)$\boldsymbol{M}_{\text{imp}}$解决该问题. $\boldsymbol{M}_{\text{att}}$通过Sigmoid激活函数得到$\boldsymbol{M}_{\text{imp}}$,对于判别能力较强的区域,重要性图中的每个像素的强度接近于1,反之接近于0,进而"奖励"$\boldsymbol{M}_{\text{att}}$中具有判别力的特征. $\boldsymbol{M}_{\text{drop}}$与$\boldsymbol{M}_{\text{imp}}$以一定使用频率$\gamma$随机使用,当使用$\boldsymbol{M}_{\text{drop}}$时,就不会使用$\boldsymbol{M}_{\text{imp}}$. 本文超参数设置擦除通道数$c=10$、阈值$\alpha=0.8$、$\boldsymbol{M}_{\text{drop}}$使用频率$\gamma$=0.2,相关消融实验见3.5.6.

输入特征图$\boldsymbol{I}_2$通过CDM后与$\boldsymbol{I}_2$进行矩阵乘得到特征图$\boldsymbol{P}$,同时$\boldsymbol{I}_2$通过SDM后得到特征图$\boldsymbol{K}$,与特征图$\boldsymbol{P}$进行矩阵乘得到输出特征图$\boldsymbol{F}$. CDM与SDM通过并联相接,且仅应用于模型训练阶段. 这两个模块虽然简单,但适用性强、效果明显.

2.2.3 局部切块分支

基于特征空间均匀切块方法能够提取有判别力的行人局部特征,让模型在训练中学习局部特征之间的差异性. PCB[5]将输入图像均匀水平分割成k个部分级特征向量,随后产生k个分类器和k个ID预测损失,这样的做法存在分割匹配不均匀、训练参数大的缺点,一定程度上反而造成了有效信息的丢失. 文献[21]提出池化操作后,将均匀切块后的局部特征级联成一个单一特征,进行单一的损失计算可以提高准确率,本文采用文献[21]的方法,将特征图均匀分割成4等份,级联切块后的局部特征,旨在减少因切割带来的信息损失,提高行人重识别的准确性.

对于第三个分支特征图$\boldsymbol{I}_3 \in \boldsymbol{R}^{512\times16\times8}$,沿着垂直方向均匀切成头部、上半身、大腿、小腿四部分,如图4所示. 对切块后的4个部分进行全局平均池化(GAP)处理,得到特征$\boldsymbol{h}_1$、$\boldsymbol{h}_2$、$\boldsymbol{h}_3$、$\boldsymbol{h}_4 \in \boldsymbol{R}^{512\times1\times1}$,随后将4个特征向量连接到一个列向量中,经过一个2048维全连接层(Fully Connected, FC)得到特征$\boldsymbol{H}\in\boldsymbol{R}^{2048\times1\times1}$, $\boldsymbol{H}=[\boldsymbol{h}_1^{\text{T}},\boldsymbol{h}_2^{\text{T}},\boldsymbol{h}_3^{\text{T}},\boldsymbol{h}_4^{\text{T}}]^{\text{T}}$.

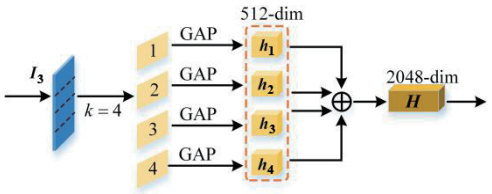


图4 切块分支结构图

2.3 损失函数

目前许多行人重识别网络联合度量损失函数和分类损失函数一起训练,共同约束特征. 论文三个分支的损失函数均为分类损失L_{id}(Softmax Loss)、中心损失L_{center}(Center Loss)[22]、三元组损失$L_{\text{soft_triplet}}$(Soft margin Triplet Loss)[23]. 为了防止训练时出现过拟合,对身份标签进行平滑操作(Label smoothing, LS)[24],LS是分类任务中防止过拟合的常用方法. 分类损失函数L_{id}为

$$L_{\text{id}}=-\sum_{i=1}^{N}q_i\log(p_i) \tag{4}$$

$$q_i=\begin{cases}1-\dfrac{N-1}{N}\varepsilon, & i=y\\ \varepsilon/N, & \text{otherwise}\end{cases} \tag{5}$$

其中,N为训练集中行人的个数,p_i为输出的行人身份的预测概率,y表示行人身份的真实标签信息. 式(5)表示对身份标签y进行LS操作,ε是一个数值较小的超参数,本文中令$\varepsilon=0.1$.

中心损失函数L_{center}为

$$L_{\text{center}}=\frac{1}{2}\sum_{i=1}^{m}\left\|f_i-C_{y_i}\right\|_2^2 \tag{6}$$

其中,m表示mini_batch,f_i表示样本i的特征,C_{y_i}表示第y_i类的特征中心.

在三元组损失函数$L_{\text{soft_triplet}}$中,图片a和图片p为一对正样本对,图片a和图片n是一对负样本对. 为了避免传统Triplet函数中超参数margin选值的困扰,本文采用文献[23]提出的soft-margin函数,具体表达式为

$$L_{\text{soft\_triplet}}=\sum_{a,p,n}\ln\left[1+\exp(D_{a,p}-D_{a,n})\right] \tag{7}$$

其中,$D_{a,p}$和$D_{a,n}$分别表示正样本对、负样本对之间的距离.

综上所述,损失函数为

$$L_{\text{total}}=\sum_i L_{\text{id}}^i+\gamma_c L_{\text{center}}^i+\gamma_t L_{\text{soft\_triplet}}^i$$
$$i\in\{\text{global、dropout、local}\} \tag{8}$$

其中,γ_t,γ_c为权重因子. 本文令γ_c=0.005,γ_t=1.0. 测试时,将全局特征$\boldsymbol{G}$、未擦除前的特征$\boldsymbol{I}_2$和切块后级联

特征 $\boldsymbol{H}$ 相连,得到行人图像总的特征 $e=[\boldsymbol{G};\ \boldsymbol{I}_2;\ \boldsymbol{H}]$,并使用欧氏距离计算特征间的距离.

3　实验

3.1　数据集

实验使用了3个主流数据集:Market1501[14]、DukeMTMC-reID[15]和CUHK03[16]. 本文在上述数据集中评估所提方法的有效性.

Market1501数据集是在清华大学校园中采集的,由6个摄像头拍摄的1 501个行人和32 668张图像组成. 其中,训练集有751个行人,包括12 936张图像,平均每个人有17.2张训练数据;测试集有750个行人,包含19 732张图像. query为测试集中随意挑选出的图像集,共3 368张.

DukeMTMC-reID数据集是跟踪数据集DukeMTMC的一个子集,由8个摄像头拍摄的1 404个行人和36 411张图像组成. 其中,训练集有702个行人,包含16 522张图像,平均每个人有23.5张训练数据;测试集有702个行人,包含17 661张图像. query和gallery分别有2 228张和16 522张.

CUHK03数据集采集于香港中文大学校园,由6个摄像头拍摄的1 467个行人和14 097张图像组成. 同时,该数据集提供了两种类型的标注:手动标记(Labeled)和DPM检测(Detected). CUHK03_Labeled数据集包含7 368张训练图像,1 400张query和5 328张gallery;CUHK03_Detected数据集包含7 365张训练图像,1 400张query和5 332张gallery.

3.2　实现细节和评估指标

实验平台的操作系统为Ubuntu16.04,使用一张NVIDIA 1080TI GPU,显存为12GB. 在Pytorch框架的基础上,使用Torchreid[25]库搭建整个网络,使用在ImageNet数据集上预训练的OSNet[13]网络作为backbone,训练时,batch_size设为64(16×4),使用Adam优化器更新梯度,初始学习率设为3.5×10^{-5},weight_decay设为0.00005. 总共训练200个epoch,在前40个epoch中,使用warm_up learning,使学习率增长至3.5×10^{-4},在100、150个epoch后,分别降至3.5×10^{-5}、3.5×10^{-6}.

在训练中,图片缩放为256×128,使用数据规范化、水平随机旋转和随机擦除作为数据增强的方法. 在测试中,图片缩放为256×128,仅仅使用数据规范化.

目前,普遍使用累积匹配特征(Cumulative Matching Characteristics, CMC)曲线和平均准确率均值(mean Average Precision, mAP)评估行人重识别模型的性能. 累积匹配特征指前K幅匹配成功的概率,本文采用第1幅就匹配成功的概率,记为Rank-1. 每个query的平均准确率是从准确率-召回率曲线计算得到的,而mAP是所有query的平均准确率的均值. 测试时,从query中选择一张图像与gallery中的所有图像匹配,计算相似度.

3.3　与最新方法的比较

本文在Market1501、DukeMTMC-reID和CUHK03数据集上,与近三年最新顶会所提方法进行比较,比较结果如表1所示. 为保证公平比较,本文实验所有方法均没有采用重排序(re-ranking)[26].

表1　不同方法在公开数据集上的性能比较　　(单位:%)

方法	Market-1501		DukeMTMC-reID		CUHK03-Labeled		CUHK03-Detected	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
HA-CNN[27](CVPR'18)	75.7	91.2	63.8	80.5	41.0	44.4	38.6	41.7
PCB[4](ECCV'18)	77.3	92.4	69.2	83.3	—	—	54.2	61.3
MGN[5](ACM MM'18)	86.9	**95.7**	78.4	88.7	67.4	68.0	66.0	66.8
Pyramid[7](CVPR'19)	88.2	**95.7**	**79.0**	89.0	76.9	78.9	74.8	78.9
MHN[28](CVPR'19)	85.0	95.1	77.2	89.1	72.4	77.2	65.4	71.7
SONA[29](ICCV'19)	**88.6**	95.6	78.1	**89.3**	**79.2**	**81.9**	**76.4**	**79.1**
ABD[17](ICCV'19)	88.3	95.6	78.59	89.0	—	—	—	—
BDB[12](ICCV'19)	86.7	95.3	76.0	89.0	76.7	79.4	73.5	76.4
Auto-ReID[30](ICCV'19)	85.1	94.5	75.1	88.5	73.0	77.9	69.3	73.3
OSNet[13](ICCV'19)	84.9	94.8	73.5	88.6	—	—	67.8	72.3
HOReID[31](CVPR'20)	84.9	94.2	75.6	86.9	—	—	—	—
SNR[32](CVPR'20)	84.7	94.4	72.9	84.4	—	—	—	—
Ours	**90.1**	**95.8**	**81.8**	**91.4**	**80.7**	**82.3**	**78.7**	**81.6**

(最好的结果标红,次之标蓝)

正如表1所示,在Market1501数据集上,在Rank-1指标上,本文方法略优于MGN[5]、Pyramid[7]达到95.8%,但是mAP远远超过前两者达到90.1%. 需要指出的是,Pyramid网络通过连接21个不同分类器训练

的局部特征得到;而MGN网络生成具有8个分支的8个特征向量,并由11个损失函数进行监督,二者的模型参数都很大. 在DukeMTMC-reID数据集上,MFN的mAP和Rank-1达到了81.8%和91.4%,大幅领先其他方法;在CUHK03-Labeled(Detected)数据集上,相比于目前精度表现最好的SONA[29],MFN在mAP和Rank-1上分别提高了1.5个百分点(2.3个百分点)、0.4个百分点(2.5个百分点). 在三个数据集中,本文模型在各种最新的方法中均获得了最好的性能,验证了MFN模型的有效性,显著地提高行人重识别的准确率.

3.4 可视化分析

图5是模型在Market1501数据集训练完成之后,使用Grad-CAM[33]得到的可视化结果. 对比图片中,第一列为输入图片,第二列和第三列分别是基准网络OSNet、MFN的热力图,图中区域颜色越高亮,表示训练中网络模型关注越多. 相比于OSNet网络只关注行人上半身,MFN关注的区域包含上下半身,基本覆盖整个行人. 不难看出,本文MFN网络感受野更大,能够提取的细粒度特征更多,同时激活区域基本覆盖整个行人区域,能够有效地减少背景的干扰,验证了本文模型的有效性.

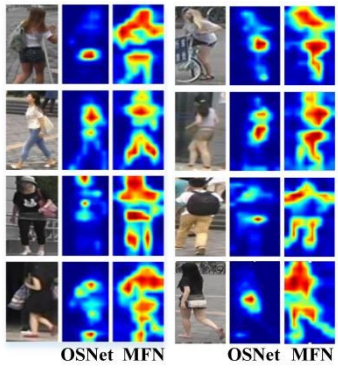
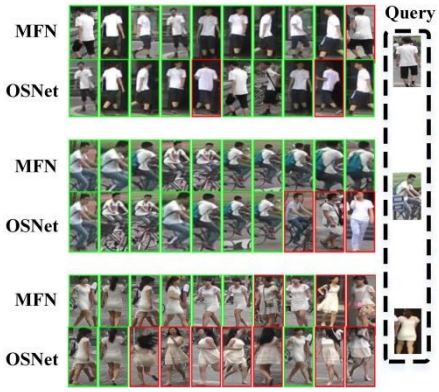


图5 Market1501数据集可视化结果



(绿色框表示识别正确,红色框表示识别错误)

图6 Market1501数据集中的识别实例

在Market1501数据集的识别实例对比如图6所示,可以更直观地从检索结果中比较两者的性能. 对于同一张query图片,MFN网络能够更好地检索出行人图片的前景和侧景,准确率更高. 以第三组对比图片为例,MFN网络在前10个匹配结果中有7个正确图像,而OSNet网络只能匹配到3个,说明MFN网络能够提取更有判别力的表征,即使不同行人穿着相似的衣服也能得到很好的识别率.

3.5 消融实验

3.5.1 不同分支对实验结果的影响

本文MFN网络由全局分支、特征擦除分支和局部切块分支三部分组成,表2显示了各个分支对实验结果的影响. 结果显示特征擦除分支对网络影响的精度最大,在单分支网络中,仅使用特征擦除分支mAP和Rank-1可以达到78.3%和88.8%,均为三条分支中性能最好. 而在双分支网络中,去除了特征擦除分支后,mAP和Rank-1为三者中最低,为80.5%和90.0%. 同时可以看出,双分支模型的识别准确率要好于单分支模型,而包含了三条分支的本文MFN模型性能优于前二者. 与基准网络OSNet相比,MFN的mAP、Rank-1提升了8.3个百分点、2.8个百分点,识别率大大提升,验证了本文模型的有效性.

表2 不同分支对实验结果的影响(DukeMTMC-reID)

Branch	mAP	Rank-1
OSNet(Baseline)	73.5	88.6
+ Global Branch	77.5	88.2
+ Feature Dropout Branch	78.3	88.8
+ Local Branch	78.0	88.5
− Global Branch	81.1	90.6
− Feature Dropout Branch	80.5	90.0
− Local Branch	80.6	90.2
MFN	**81.8**	**91.4**

(注:+表示网络中仅添加该分支;− 表示三分支网络中取消该分支)

3.5.2 分支结构对实验结果的影响

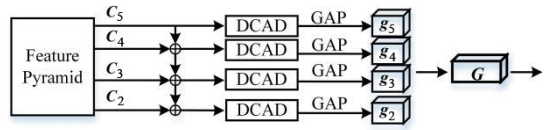
为验证本文所提三分支结构的必要性,进行分支结构消融实验. 本节去除特征擦除分支,在全局分支、局部切块分支分别使用特征擦除操作,使原网络变为双分支网络结构. 具体所比较的网络分支结构如下:

(1)分支结构_1 在该双分支网络中,保持局部分支不变,在全局分支特征金字塔结构上使用特征擦除操作,全局分支基本架构如图7(a)所示;

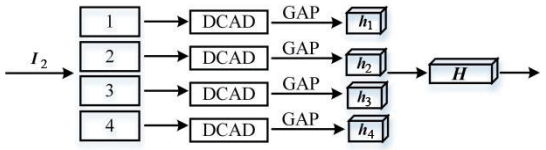
(2)分支结构_2 在该双分支网络中,保持全局分支不变,在局部分支上使用特征擦除操作,局部分支基本架构如图7(b)所示;

(3)分支结构_3 在该双分支网络中,在全局分支特征金字塔结构、局部分支上均使用特征擦除操作.

如表3所示,与双分支网络相比,本文所提出的三分支网络结构MFN具有更好的识别性能. 总结分析如下.



(*a*)　双分支网络中全局分支基本架构



(*b*)　双分支网络中局部分支基本架构

图7　双分支网络各分支架构简图

(1) 在MFN网络中,三分支网络结构的作用是提取细粒度特征,提高识别准确率. 上述的消融实验也证明了这三个分支相互协作、相互监督,对最终的性能表现都很重要.

(2) 实验证明在全局分支(局部分支)上使用特征擦除的效果并不好. 分析原因如下:对于全局分支,特征金字塔模型提取从低层到高层的多尺度特征,尤其在网络初始阶段,低层特征中具有鉴别力的信息较少,对其使用特征擦除会影响网络识别性能;对于局部分支,对切块后的特征图进行特征擦除操作,会影响切块分界处具有连续信息的特征的融合,降低最终的识别准确率.

表3　修改前后网络性能比较(DukeMTMC-reID)

Method	mAP	Rank-1
Method1	81.0	89.2
Method2	80.6	89.6
Method3	80.5	90.0
MFN	**81.8**	**91.4**

3.5.3　简单全局池化与特征金字塔结构对结果的影响

直接使用基于CNN方法提取图像的全局特征,会忽视一些分布在细粒度的语义信息. 表4更加直观地对简单全局池化和特征金字塔结构对结果的影响进行对比,在Market1501数据集上做了四组对比实验,结果显示,在mAP和Rank-1指标上,采用特征金字塔结构的方法均优于简单全局池化,以全局最大池化为例,采用特征金字塔结构在mAP、Rank-1指标上提升了0.3个百分点、0.5个百分点. 因此本文全局分支采用金字塔结构来提取不同尺度的特征的效果更好.

表4　全局分支中简单全局池化与特征金字塔结构对实验结果的影响(Market1501)

Method	mAP	Rank-1
OSNet(Baseline)	84.9	94.8
+Global-Avg-Pooling	89.9	95.5
+Global-Max-Pooling	89.6	95.4
+ Feature Pyramid(Avg-Pooling)	**90.1**	**95.8**
+ Feature Pyramid(Max-Pooling)	**90.1**	95.7

3.5.4　不同特征擦除模块对实验结果的影响

相比于BDB[16]模块擦除特征的随机性,本文双通道注意力特征擦除模块(DCAD)由通道擦除模块(CDM)、空间擦除模块(SDM)组成,以一定的比例擦除特征图中具有判别力的特征,更能够促使网络加强次显著特征的学习. 表5为增加不同特征擦除模块对实验结果的影响,不难看出,特征擦除模块提升了模型的识别精度,在CUHK03-Detected数据集上,加入CDM、SDM分别在mAP、Rank-1指标超过基准网络OSNet的结果9.0个百分点、7.3个百分点和10.1个百分点、8.1个百分点,SDM的加入更能提高模型的识别率. 同时,融合CDM和SDM的本文双通道注意力特征擦除模型(DCAD)的加入明显改善了模型的性能,在CUHK03-Labeled数据集上mAP、Rank-1能够达到80.7%、82.3%,超过了加入BDB模块的结果0.2个百分点、0.4个百分点,可见本文方法简单有效.

表5　不同特征擦除模块对实验结果的影响(CUHK03)

Method	CUHK03-Labeled		CUHK03-Detected	
	mAP	Rank-1	mAP	Rank-1
OSNet(Baseline)	—	—	67.8	72.3
+ BDB	80.5	81.9	78.8	81.2
+CDM	79.8	80.6	76.8	79.6
+SDM	80.2	81.4	77.9	80.4
+DCAD	**80.7**	**82.3**	**78.7**	**81.6**

3.5.5　CDM、SDM连接方式对实验结果的影响

如表6所示,讨论了CDM、SDM不同的连接方式对实验结果的影响. 结果显示,本文MFN网络特征擦除分支CDM+SDM并联连接,在mAP、Rank-1指标上均优于两个模块级联方法,分别达到81.8%,91.4%. 与并联方法相比,对于CDM+SDM级联方法,此时SDM的输入是CDM后的特征图,SDM的目的是加强次显著部分的特征学习,与并联方法相比,会有些许的信息损失,从而影响实验精度. 而对于SDM+CDM级联方法,先经过SDM的特征图,可能会在CDM中排名靠后从而被置零,影响识别精度. 综上所述,特征擦除分支中CDM+SDM并联连接性能更好.

3.5.6　特征擦除分支中超参数对实验结果的影响

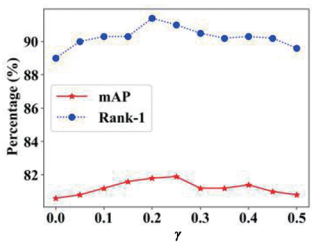
为说明超参数取值的合理性,本文在DukeMTMC-

表6 CDM和SDM连接方式对实验结果的影响(DukeMTMC-reID)

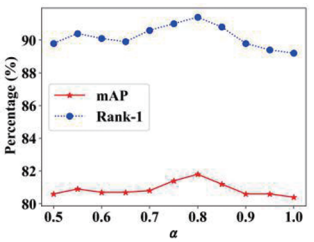
Method	mAP	Rank-1
CDM+SDM(级联)	81.4	91.0
SDM+CDM(级联)	80.9	90.8
CDM+SDM(并联)	**81.8**	**91.4**

(注:级联+号前后表示级联的前后顺序)

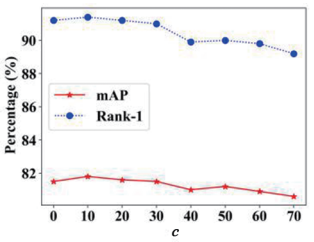
reID数据集上,对设置的超参数使用频率γ,阈值α,擦除通道数c进行消融实验. 超参数γ决定了擦除模块M_{drop}的使用频率. 如图8(a)所示,当γ取0.2时,Rank-1指标性能最好,mAP指标略低于γ=0.25时的性能,综合考虑取γ=0.2. 当特征像素大于超参数α时置零,如果α的取值太小,会导致置零的像素点过多,影响识别精度. 如图8(b)所示,取α=0.8,此时性能最好. 超参数c表示使用二进制掩码置零排序靠后的通道数,过大的c值会导致置零的通道数过多,损失有鉴别力的特征信息. 如图8(c)所示,本文取c=10.



(a) 使用频率γ的影响



(b) 阈值α的影响



(c) 擦除通道c的影响

图8 超参数对实验结果的影响(DukeMTMC-reID)

3.5.7 分块策略对实验结果的影响

表7展示了分块策略对实验结果的影响,结果显示将图片分为4等份的时候性能最好,能充分发挥局部特征的作用,有效提高行人重识别的准确率,其关键指标mAP和Rank-1在DukeMTMC-reID数据集上达到81.8%和91.4%. 通过级联合成进行单一的损失计算,减少网络参数的同时在一定程度上减少了匹配不均匀的问题,提高了识别率.

表7 分块策略对比实验(DukeMTMC-reID)

Branch	mAP	Rank-1
+3 part-level	81.5	90.8
+4 part-level	**81.8**	**91.4**
+5 part-level	81.2	90.2
+6 part-level	81.0	89.8

4 结语

本文提出基于多粒度特征融合的行人重识别方法(MFN),通过在预训练的OSNet网络上搭建全局分支、特征擦除分支和局部切块分支网络,最小化三元组损失、交叉熵损失和中心损失函数,完成表征提取. 消融实验、特征可视化结果表明,本文多分支网络结构能够很好地捕捉行人图像的多粒度特征,其中全局分支提取行人多尺度特征,特征擦除分支关注行人次显著特征并防止过拟合,切块分支进一步提取行人关键局部信息. 三条分支相互作用,极大地提高了识别精度. 后续工作将考虑分支之间的关系,进一步研究如何在简化网络模型复杂度的情况下,寻求更高重识别率的方法.

参考文献

[1] 罗浩, 姜伟, 范星, 等. 基于深度学习的行人重识别研究进展[J]. 自动化学报, 2019, 45(11): 2032 – 2049.
Luo H, Jiang W, Fan X, et al. A survey on deep learning based Person Re-Identification [J]. Acta Automatica Sinica, 2019, 45(11): 2032 – 2049. (in Chinese)

[2] Weinberger K Q, Saul L K. Fast solvers and efficient implementations for distance metric learning [A]. Proceedings of the 25th International Conference on Machine Learning [C]. Helsinki, Finland: ICML, 2008. 1160 – 1167.

[3] Liao S, Hu Y, Zhu X, et al. Person re-identification by local maximal occurrence representation and metric learning [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Boston, MA, USA: CVPR, 2015. 2197 – 2206.

[4] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [A]. Proceedings of the European Conference on Computer Vision [C]. Munich, Germany:

ECCV, 2018. 480 – 496.

[5] Wang G, Yuan Y, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification[A]. Proceedings of the 26th ACM International Conference on Multimedia [C]. Seoul, Korea: ACM, 2018. 274 – 282.

[6] 陈巧媛,陈莹.通道互注意机制下的部位对齐行人再识别[J].计算机辅助设计与图形学学报, 2020, 32(8): 1258 – 1266.
Chen Q Y, Chen Y. Correlation channel-wise based part aligned representations for person re-identification [J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(8): 1258 – 1266. (in Chinese)

[7] Zheng F, Deng C, Sun X, et al. Pyramidal person re-identification via multi-loss dynamic training[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Long Beach, CA, USA: CVPR, 2019. 8514 – 8522.

[8] DeVries T, Taylor G W. Improved regularization of convolutional neural networks with cutout[EB/OL]. https://arxiv. org/abs/1708. 04552, 2017-08-15.

[9] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[A]. Association for the Advance of Artificial Intelligence [C]. New York, USA: AAAI, 2020. 13001 – 13008.

[10] Ghiasi G, Lin T Y, Le Q V. Dropblock: A regularization method for convolutional networks [A]. Neural Information Processing Systems [C]. Montreal, Canada: NIPS, 2018. 10727 – 10737.

[11] Tompson J, Goroshin R, Jain A, et al. Efficient object localization using convolutional networks [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Boston, MA, USA: CVPR, 2015. 648 – 656.

[12] Dai Z, Chen M, Gu X, et al. Batch DropBlock network for person re-identification and beyond[A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Long Beach, CA, USA: CVPR, 2019. 3691 – 3701.

[13] Zhou K, Yang Y, Cavallaro A, et al. Omni-scale feature learning for person re-identification[A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Seoul, Korea: ICCV, 2019. 3702 – 3712.

[14] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[A]. Proceedings of The IEEE International Conference on Computer Vision[C]. Santiago, Chile: ICCV, 2015. 1116 – 1124.

[15] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking [A]. European Conference on Computer Vision[C]. Amsterdam, Netherlands: ECCV, 2016. 17 – 35.

[16] Li W, Zhao R, Xiao T, et al. Deepreid: Deep filter pairing neural network for person re-identification[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Columbus, USA: CVPR, 2014. 152 – 159.

[17] Chen T, Ding S, Xie J, et al. Abd-net: Attentive but diverse person re-identification [A]. Proceedings of the IEEE International Conference on Computer Vision[C]. Seoul, Korea: ICCV, 2019. 8351 – 8361.

[18] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. Honolulu, HI, USA: CVPR, 2017. 2117 – 2125.

[19] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[A]. European Conference on Computer Vision [C]. Zurich, Switzerland: ECCV, 2014. 740 – 755.

[20] Choe J, Shim H. Attention-based dropout layer for weakly supervised object localization[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Long Beach, CA, USA: CVPR, 2019. 2219 – 2228.

[21] Xie B, Wu X, Zhang S, et al. Learning diverse features with part-level resolution for person re-identification [EB/OL]. https: //arxiv.org/abs/2001.07442, 2020-01-21.

[22] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[A]. European Conference on Computer Vision [C]. Amsterdam, Netherlands: ECCV, 2016. 499 – 515.

[23] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[EB/OL]. https: / /arxiv. org/ abs/ 1703. 07737, 2017-03-22.

[24] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Las Vegas, NV, USA: CVPR, 2016. 2818 – 2826.

[25] Zhou K, Xiang T. Torchreid: A library for deep learning person re-identification in pytorch [EB/OL]. https: //arx-

iv. org/abs/1910.10093, 2019-10-22.

[26] Zhong Z, Zheng L, Cao D, et al. Re-ranking person re-identification with k-reciprocal encoding [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Honolulu, HI, USA: CVPR, 2017. 1318 – 1327.

[27] Li W, Zhu X, Gong S. Harmonious attention network for person re-identification [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Salt Lake City, UT, USA: CVPR, 2018. 2285 – 2294.

[28] Chen B, Deng W, Hu J. Mixed high-order attention network for person re-identification [A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Seoul, Korea: ICCV, 2019. 371 – 381.

[29] Xia B N, Gong Y, Zhang Y, et al. Second-order non-local attention networks for person re-identification [A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Long Beach, CA, USA: CVPR, 2019. 3760 – 3769.

[30] Quan R, Dong X, Wu Y, et al. Auto-reid: Searching for a part-aware convnet for person re-identification [A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Long Beach, CA, USA: CVPR, 2019. 3750 – 3759.

[31] Wang G, Yang S, Liu H, et al. High-order information matters: Learning relation and topology for occluded person re-identification [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Seattle, WA, USA: CVPR, 2020. 6449 – 6458.

[32] Jin X, Lan C, Zeng W, et al. Style normalization and restitution for generalizable person re-identification [A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. Seattle, WA, USA: CVPR, 2020. 3143 – 3152.

[33] Selvaraju R R, Cogswell M, Das A, et al. Visual explanations from deep networks via gradient-based localization [A]. Proceedings of the IEEE International Conference on Computer Vision [C]. Venice, Italy: ICCV, 2017. 618 – 626.

作者简介

匡 澄 男,1996年出生,江苏无锡人. 现为江南大学物联网工程学院硕士研究生,专业为电子通信工程,主要研究方向为行人重识别. E-mail:6191918020@stu.jiangnan.edu.cn

陈 莹(通信作者) 女,1976年12月出生,浙江丽水人. 江南大学教授,博士生导师,主要研究方向为图像处理、信息融合、模式识别. E-mail:chenying@ jiangnan.edu.cn