JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2020
05
VOL.25
ISSN1006-8961
CN11-3758/TB

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第25卷第5期（总第289期）
2020年5月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 顾行发
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Gu Xingfa
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2020年5月 第25卷 第5期（总第289期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

多相图像分割Vese-Chan模型连续最大流方法(第0926页)

保持区域平滑的3维动画形状高效编辑(第1019页)

树形结构卷积神经网络优化的城区遥感图像语义分割(第1043页)

封面图片由武汉大学张良培教授团队提供

Volume 25, Number 5
Published May 16, 2020

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Continuous max-flow method for multiphase segmentation Vese-Chan model(P0926)

Efficient 3D animation shape manipulation with region smoothness preservation(P1019)

Semantic segmentation of urban remote sensing image based on optimized tree structure convolutional neural network (P1043)

E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2020)05-0946-10

论文引用格式:Shen Q, Tian C, Wang J B, Jiao S S and Du L. 2020. Multi-resolution feature attention fusion method for person re-identification. Journal of Image and Graphics,25(05):0946-0955(沈庆,田畅,王家宝,焦珊珊,杜麟. 2020. 多分辨率特征注意力融合行人再识别. 中国图象图形学报,25(05):0946-0955)[DOI:10.11834/jig.190237]

多分辨率特征注意力融合行人再识别

沈庆,田畅,王家宝,焦珊珊,杜麟

陆军工程大学,南京　210007

摘　要:**目的**　行人再识别是实现跨摄像头识别同一行人的关键技术,面临外观、光照、姿态、背景等问题,其中区别行人个体差异的核心是行人整体和局部特征的表征。为了高效地表征行人,提出一种多分辨率特征注意力融合的行人再识别方法。**方法**　借助注意力机制,基于主干网络 HRNet(high-resolution network),通过交错卷积构建4个不同的分支来抽取多分辨率行人图像特征,既对行人不同粒度特征进行抽取,也对不同分支特征进行交互,对行人进行高效的特征表示。**结果**　在 Market1501、CUHK03 以及 DukeMTMC-ReID 这3个数据集上验证了所提方法的有效性,rank1 分别达到95.3%、72.8%、90.5%,mAP(mean average precision)分别达到89.2%、70.4%、81.5%。在 Market1501 与 DukeMTMC-ReID 两个数据集上实验结果超越了当前最好表现。**结论**　本文方法着重提升网络提取特征的能力,得到强有力的特征表示,可用于行人再识别、图像分类和目标检测等与特征提取相关的计算机视觉任务,显著提升行人再识别的准确性。

关键词:HRNet;交错卷积;注意力机制;多分辨率特征表示;特征融合;行人再识别

Multi-resolution feature attention fusion method for person re-identification

Shen Qing, Tian Chang, Wang Jiabao, Jiao Shanshan, Du Lin

Army Engineering University of PLA, Nanjing　210007, China

Abstract: **Objective** Person re-identification (ReID) is a computer vision task of re-identifying a queried person across non-overlapping surveillance camera views developed at different locations by matching images of the person. Given the fundamental problem of intelligent surveillance analysis, person ReID has attracted increasing interest in recent years among computer vision and pattern recognition research communities. Although great progress has been made in person ReID, it still has challenges, such as occlusion, illumination, pose variance, and background clutter. The key to solving these difficulties is to efficiently design a convolutional neural network (CNN) architecture that can extract discriminative feature representations. Specifically, the architecture should be capable of compacting the "intraclass" variation (obtained from the same individual) and separating the "interclass" variation (obtained from different individuals). The algorithm process of person ReID mainly includes two stages, namely, feature extraction and distance measurement. Most contemporary studies have focused on feature extraction because a good feature can effectively distinguish different persons. Thus, the designed CNN network needs to have good representation for the global and local features of different individuals. To fully mine the information contained in the image, we fuse the features of the same image at different resolutions to obtain a stronger fea-

收稿日期:2019-06-06;**修回日期**:2019-10-23;**预印本日期**:2019-10-30

ture representation and develop a multi-resolution feature attention fusion method for person ReID. **Method** At present, mainstream person ReID methods are based on classical networks such as ResNet and VGG (visual geometry group) -Net. The main characteristic of these networks is that the resolution of the feature maps becomes increasingly smaller as the network continuously deepens. Moreover, their high-level features contain sufficient semantic information but lack spatial information. However, for tasks involving person ReID, the spatial information of an individual is necessary. The high-resolution network (HRNet) is a multi-branch network that can maintain high-resolution representations throughout the whole process. HRNet is constructed by interleaving convolutions, which are helpful for obtaining different granularity features. It also helps in the information exchange among different branches. HRNet can output four different resolution feature representations. In this study, we first evaluate the performance of different resolution feature representations. Results show that the performance of these feature representations is not consistent on different datasets. Therefore, we propose an attention module to fuse the different resolution feature representations. The attention module can generate four weights, which add up to 1. The different resolution feature representations can be updated according to different weights. The final feature representation is the accumulation of the four different updated features. **Result** Experiments are conducted on three ReID datasets: Market1501, CUHK03, and DukeMTMC-ReID. Results indicate that our method pushes the performance to an exceptional level compared with most existing methods. Rank-1 accuracy is achieved with the following percentages: 95.6%, 72.8%, and 90.5%. Moreover, mAP (mean average precision) scores of 89.2%, 70.4%, and 81.5% are obtained on Market1501, CUHK03, and DukeMTMC-ReID datasets, respectively. Our method achieves state-of-the-art results on the DukeMTMC-ReID dataset and yields competitive performance with the state-of-the-art methods on Market1501 and CUHK03 datasets. The mAP score of our method is also the highest on the Market1501 dataset. In the ablation study, we evaluate the influence of three situations on the performance of our model, namely, the attention module at different locations, the images with different resolutions, and the weights of different normalization methods. Results show that the behind attention mechanism is better than the front attention mechanism. In addition, the image resolution has little influence on the performance, and the Sigmoid normalization method outperforms the softmax normalization method. **Conclusion** In this study, we proposed a multi-resolution attention fused method for person ReID. HRNet is an original network used to extract coarse-grain and fine-grain features, which are helpful for person ReID. Through ablation study, we found that the performance of different resolution feature representations is not consistent on different datasets. Thus, we proposed an attention module to fuse the different resolution features. The attention module outputs four weights to represent the importance of the different resolution features. The fused feature was obtained by accumulating the updated features. The experiments were conducted on Market1501, CUHK03, and DukeMTMC-ReID datasets. Results showed that our method outperforms several state-of-the-art person ReID approaches and that the attention fused method improves the performance.

Key words: high-resolution network (HRNet); interleaving convolution; attention mechanism; multi-resolution feature representation; feature fusion; person re-identification

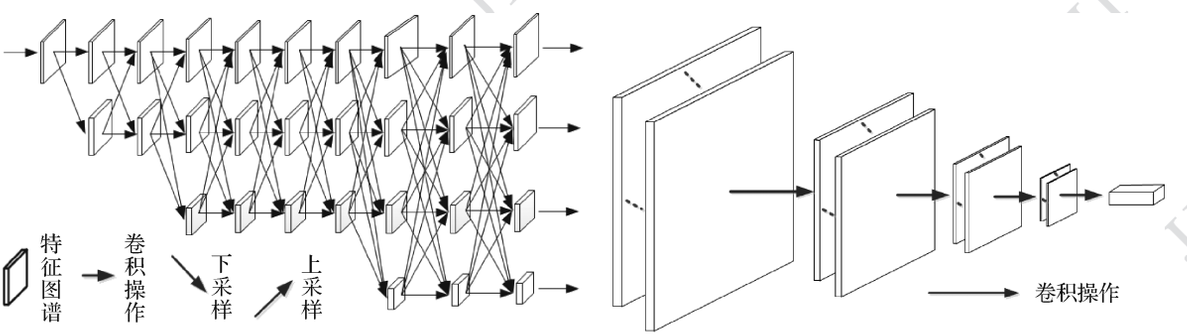
0 引　言

行人再识别聚焦于对不同摄像机中同一行人的匹配。即给定某一摄像机中的一幅行人图像，通过行人再识别找到该行人在其他摄像机中的图像。行人再识别在疑犯追踪、寻找走失老人儿童等方面有着重要的积极作用，成为近年来计算机视觉中的研究热点。

当前的行人再识别网络多以 ResNet(He 等，2016)，VGG(visual geometry group)-Net(Simonyan 和 Zisserman，2014)，DenseNet(Huang 等，2017)，GoogleNet(Szegedy 等，2015)等作为骨干网络进行研究，这些网络已经在 ImageNet 数据集上进行了实验，有效性得到了一定的验证，在计算机视觉的各个方面得到广泛应用。利用这些网络已有的预训练权重，可以显著减少模型的训练时间，快速验证相关想法。这些网络的共同特点是只有一个主分支，沿着主分支，特征的分辨率不断降低，提取的图像特征聚焦于高级的语义特征。Fu 等人(2018)的研究结果表明，通过组合神经网络不同层的特征可以取得更好的结果。但是由于浅层特征与深层特征之间固有的差异，使得直接组合这些特征带来的提升相对有限。

2018 年，人体姿态估计中提出了一种新的网络

HRNet(high-resolution network)(Sun 等,2019),通过交错卷积并行维持多个分辨率的特征,使得不同分支之间的特征可以有效进行信息的交互融合,进而提取更全面的特征,在姿态估计中取得了很好的效果。HRNet 网络强调网络的宽度,同一分支的分辨率不随深度变化而改变,而传统网络随着深度的加深,分辨率不断降低,二者的结构对比如图 1 所示。



(a) HRNet 网络结构　　(b) 传统网络结构

图 1　HRNet 与传统网络的结构对比

Fig. 1　The structure of the HRNet and traditional networks ((a) HRNet structure; (b) traditional network)

在图像识别中,基于注意力机制的研究也得到了发展。注意力机制的应用始于自然语言处理的机器翻译(Vaswani 等,2017),此前神经网络处理分析句子时是按序进行,但是翻译并不是按序一一对应。引入注意力机制后,可以使网络抓住句子的关键信息,取得了不错的效果。后来注意力机制用于图像处理(Xu 等,2015),出现了基于特征通道的注意力机制 SENet(squeeze-and-excitation network)(Hu 等,2018)、基于通道和空间特征的注意力机制 BAM(bottleneck attention module)(Park 等,2018),以及改进版的 CBAM(convolutional block attention module)(Woo 等,2018)。SENet 通过显式地建立模型,定义通道间的关系,加强有用信息,压缩无用信息。在 SENet 的基础上,SKNet(selective kernel network)(Li 等,2019)通过引入不同大小的卷积核提取不同尺度上的特征,然后将特征融合后得到不同卷积核对应的特征权重,最后对特征进行加和的融合操作。可以看出,注意力机制在计算机视觉方面也取得了一定的效果。

在当前的行人再识别中,特征提取已经由全局特征发展到全局特征与局部特征相结合。全局特征的颗粒度较粗,对于外观相似的两个人,全局特征不易区分,需要局部特征进一步区分。局部特征关注行人的局部细节,而细节更加具有区分性。Lin 等人(2019)通过属性标签这个局部特征来训练分类器;PCB(part-based convolutional baseline)(Sun 等,2018)在特征上划分得到多个局部特征;MGN(multiple granularity network)(Wang 等,2018)将特征进行不同大小的划分来获得全局和局部特征;Tian 等(2018)通过人体姿态估计以及人体解析等方法,划分得到行人的各个部分,进而进行分类器的训练。但上述方法存在一些缺点:在原图上进行划分过于复杂;在最后的特征上进行划分,特征的多样性不够充分。

HRNet 网络同时输出多种分辨率的特征,不同分辨率的特征有不同的颗粒度,关注原图像上不同尺度的区域,这些特征之间存在一定的互补性。通过组合这些特征,可以得到更好的行人特征表示。

借助 HRNet 强大的特征表示能力,本文将 HRNet 的 4 个分支输出的不同分辨率的特征通过一个注意力模块进行特征融合,得到更加高效的行人再识别特征表示。

1　多分辨率特征注意力融合方法

本文将 HRNet 作为行人再识别的骨干网络,用于提取行人特征。HRNet 有 4 个分支,输出 4 种不同分辨率、不同通道数的特征。4 个输出分支由上到下依次为第 1、2、3、4 分支。假设输入图像的分辨率为 256×128 像素,各分支输出的通道数和分辨率如表 1 所示。

表 1　HRNet 的 4 个输出分支的通道数与分辨率

Table 1　The channels and resolution of the four branches of the HRNet

分支	通道数/个	分辨率/像素
1	32	64 × 32
2	64	32 × 16
3	128	16 × 8
4	256	8 × 4

1.1　不同分辨率输出特征评测

HRNet 最先用于姿态估计。在姿态估计中，为了使特征保持较高的分辨率，原始 HRNet 只使用了具有较高分辨率的第 1 分支，其他的分支都没有使用。为了评测 HRNet 的 4 个分支输出特征之间的差异，本文设计了一个评测网络用于评测各分支特征在行人再识别中的表现，如图 2 所示。

图 2 中，P 和 K 表示 P 个行人，每个行人 K 幅图像组成一个batchsize输入网格。对于分支的输出特征 $\boldsymbol{f}_s$，先通过一个 1 × 1 的卷积模块将特征通道数增加到 2 048，然后通过一个全局最大池化模块（global max pooling，GMP）得到特征 $\boldsymbol{f}_t$，该特征用于计算 triplet loss 和 center loss。然后通过一个 BN 模块，得到特征 $\boldsymbol{f}_i$。最后接一个分类器，用于进行有监督的分类训练。

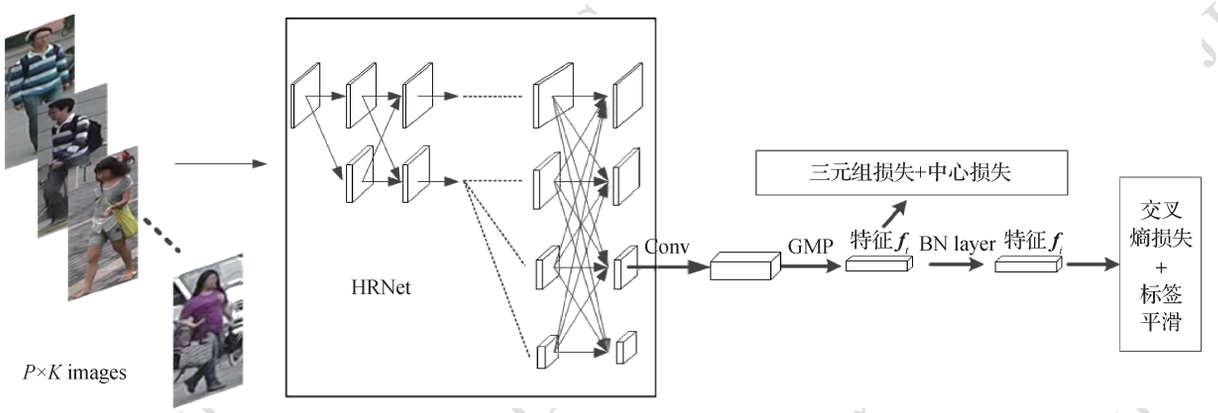


图 2　基于 HRNet 的单个分支行人再识别网络

Fig. 2　Single branch of HRNet for person re-identification

rank1 和 mAP（mean average precision）是常用的评测性能指标。rank1 表示检索图像中，第 1 幅就是该行人图像的概率，mAP 表示检索图像中，检索找回所有标签图像的准确率在整个数据集上的平均精度。实验结果如表 2 所示，从表 2 可以看出，在3 个数据集上，各分支输出特征在行人再识别上的表现存在一定的差异性。对于 Market1501 与 DukeMTMC-ReID 两个数据集，第 3 分支效果最好；对于 CUHK03 数据集，第 4 分支表现最好。由于第 1 分支到第 4 分支分辨率递减，特征抽象能力逐渐递增，第 3 分支和第 4 分支相对具有更好的语义表征能力。值得注意的是，虽然第 1 分支和第 2 分支的性能不如第 3 分支和第 4 分支，但是性能差异较小。

为了探究不同分支输出特征之间的差异，采用 Grad-CAM（gradient-based class activation mapping）（Selvaraju 等，2017）可视化特征的热力图。Grad-CAM 能够对某一层的输出特征可视化其在原图像上的位置响应强度，对应表明原图像不同位置像素对特征所做的贡献。本文选择比较 4 个分支经过 GMP 后得到的特征 $\boldsymbol{f}_t$ 在原图上的响应。图 3 展示了在 Market1501 数据集中，1 个行人图像对 4 个分支输出特征可视化的热力图。

由图3可以发现，随着通道数增加、分辨率降

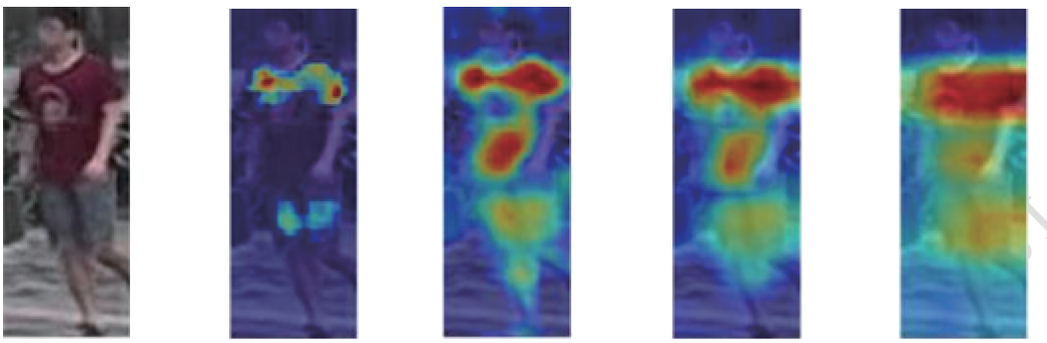
表 2　评测 HRNet 各分支性能

Table 2　Evaluate the persormance of all branches of the HRNet

/%

分支	Market1501		CUHK03		DukeMTMC-ReID	
	rank1	mAP	rank1	mAP	rank1	mAP
1	94.6	87.7	67.9	65.5	89.4	80.4
2	94.8	87.9	67.6	65.7	89.7	**80.8**
3	**94.9**	**88.1**	67.6	66.1	**89.9**	80.4
4	94.3	87.5	**70.1**	**67.0**	88.6	79.5

注：加粗字体为每列最优值。



(a) 原始图像　(b) 第1分支　(c) 第2分支　(d) 第3分支　(e) 第4分支

图 3　HRNet 的 4 个分支输出的热力图

Fig. 3　Heatmap of the four branch of HRNet

((a) original image; (b) the first branch; (c) the second branch; (d) the third branch; (e) the fourth branch)

低,分支的输出特征越来越倾向于高层的语义特征,分支的输出特征关注的区域范围也在不断增加。

1.2　基于 HRNet 的多分支注意力融合网络

鉴于 HRNet 输出的不同分辨率特征在不同数据集上表现不一致,本文借鉴注意力机制,设计了一个融合不同分辨率特征的注意力模块,将 4 个分支输出的特征进行特征融合,各分支的权重由注意力模块学习得到,无需人工设定。对于 4 个分支的输出特征,首先进行 1×1 的卷积操作将通道数增加到 2 048 维,以增加特征的丰富性,增强特征的分辨能力,接着通过一个全局最大池化(GMP),将特征大小统一为[N, 2 048],然后将 4 个分支的输出特征一起送入注意力模块,由注意力模块得到各个分支的权重,并根据各分支的权重对分支特征进行加权融合,得到最终的行人特征表示。HRNet 各分支的输出特征为$\boldsymbol{f}_s$, $s=1,2,3,4$。4 个分支的特征进行合并后的特征

$$\boldsymbol{f}_w = C_t(R_s(\boldsymbol{f}_s)),\quad s=1,2,3,4 \tag{1}$$

式中,$R_s(\cdot)$ 表示 reshape 操作,用于将输入特征由[N,2 048,1,1]变换为[N,2 048,1];$C_t(\cdot)$ 表示将变换后的特征在最后一个通道上进行连接,得到$\boldsymbol{f}_w$,大小为[N,2 048,4]。之后将$\boldsymbol{f}_w$在通道维度上进行 1 维卷积,将通道数由 2 048 下降为 1,消去通道维度,此时的输出大小为[N, 4]。接着使用 Sigmoid 操作将特征映射到(0,1)区间,并通过归一化操作得到归一化的权重值,4 个分支的权重之和为 1。单个分支的权重

$$w = N(S(C_v(\boldsymbol{f}_w))) \tag{2}$$

式中,$C_v(\cdot)$ 表示 1 维的卷积操作,卷积操作的输入通道数为 2 048,输出通道数为 1,卷积核大小为 3,步长为 1;$S(\cdot)$表示 Sigmoid 操作,$N(\cdot)$表示对特征进行权重归一化操作。

通过注意力模块得到 4 个分支的权重后,分别与各个分支的输入特征相乘,得到更新后的分支特征。将更新后的分支特征进行加权融合,得到最后的行人特征描述,即

$$\boldsymbol{F}_u = \sum_{i=1}^{4} \boldsymbol{f}_i \times w_i \tag{3}$$

本文设计的多分辨率注意力融合网络结构和注意力模块结构如图 4 和图 5 所示。

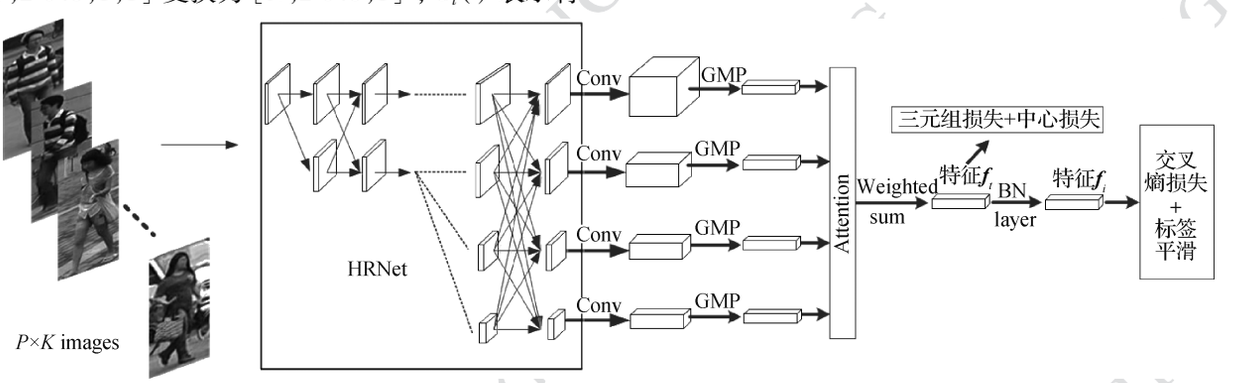


图 4　多分辨率注意力融合网络结构

Fig. 4　Proposed multi-resolution attention fused network

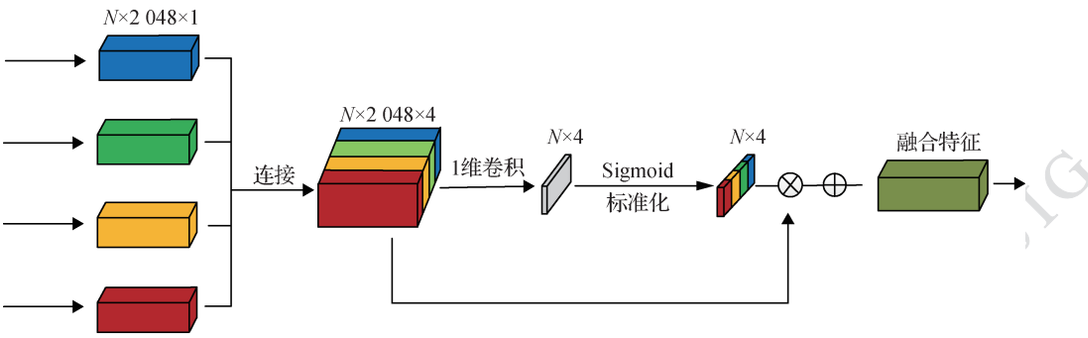


图 5　注意力模块结构

Fig. 5　Structure of the proposed attention module

2　实　验

实验借鉴 Luo 等人(2019)的训练策略并均以此为基础进行,以实现公平的对比分析。包括:1)在训练阶段,学习率的变化使用 WarmUp 方式,使网络有更好的表现;2)对训练的数据集进行概率为 0.5 的随机擦除,以达到数据增强的目的;3)标签平滑,用于降低模型过拟合的风险,提高模型的泛化性能;4)BNNeck(batch normalization neck),对特征进行 1 维的归一化,主要解决交叉熵损失与三元组损失两者度量不一致导致的总的 loss 震荡问题;5)增加中心损失,约束类内特征的相对距离。此外,为避免偶然性,所有实验都是重复 3 次取均值,以保证结果的准确性与客观性。

2.1　实验相关设置

实验使用的输入图像大小为 256 × 128 像素,batchsize 的大小为 64,由 16 个人,每个人 4 幅图像构成,用于计算 triplet loss。另外,使用了基于标签平滑的交叉熵损失和中心损失。实验中使用的损失函数为

$$L = L_{\text{Softmax}} + L_{\text{Triplet}} + \beta L_{\text{Center}} \tag{4}$$

式中,β 是用于平衡中心损失的调整参数,设置为 0.000 5。学习率的变化如图 6 所示。优化器使用 Adam,triplet loss 的 MARGIN 设置为 0.3,共训练 120 轮。

2.2　特征维度的影响评测

本文使用的 HRNet 网络 4 个分支的输出通道数分别为 32,64,128,256。为验证卷积后不同通道数量对实验结果的影响,通过 4 个 1 × 1 的卷积增加特征通道数,将输出通道数分别设置为 256,512,1 024,2 048,4 096,在效果较好的第 3 分支上进行多组实验,结果如表 3 所示。

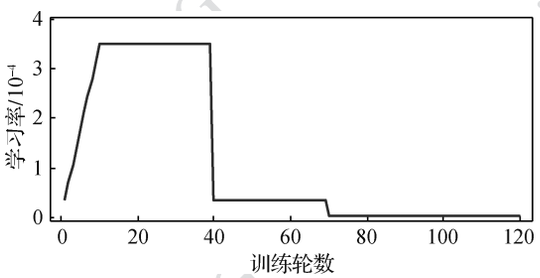


图 6　学习率变化

Fig. 6　Change of the learning rate

表 3　不同的特征维度对结果的影响

Table 3　The influence of different feature dimension

/%

卷积通道数量	Market1501		DukeMTMC-ReID	
	rank1	mAP	rank1	mAP
256	93.9	85.1	88.7	77.3
512	94.1	86.4	89.4	78.9
1 024	95.0	87.6	89.9	80.0
2 048	95.1	88.1	90.0	80.6
4 096	**95.2**	**88.6**	**90.1**	**81.3**

注:加粗字体为每列最优值。

由表 3 可以发现,随着通道数的增加,再识别的 rank1 和 mAP 性能指标逐步增加。因为通道数的增加使得最终提取的特征数量增加,特征表征能力不断增强,但也导致计算量与存储空间增加。在实验中,当通道数增加到 4 096 时,效果提升较小,但计算时间大大增加。折中考虑,本文通道数选择为 2 048。

2.3　注意力模块有效性评测

为验证注意力模块的有效性,将通过注意力模块学习得到分支权重的方法与使用单一分支及 4 个

分支取权重(实验中均手工设置为 0.25)进行特征加权融合的方法对比,结果如表 4 所示。

表 4 注意力模块实验结果

Table 4 Result of the attention module

/%

方法	Market1501		CUHK03		DukeMTMC-ReID	
	rank1	mAP	rank1	mAP	rank1	mAp
单一分支	94.9	88.1	70.1	67.0	89.7	80.8
4 分支等权重加权融合	95.2	88.9	70.6	68.8	89.6	81.2
注意力模块	**95.3**	**89.2**	**72.8**	**70.4**	**90.5**	**81.5**

注:加粗字体为每列最优值。

从表 4 可以看出,本文注意力模块的结果优于单一分支和手工设置权重加权融合的结果。4 分支等权重加权融合的结果好于单一分支但弱于注意力加权融合,说明 4 分支融合是有效的,但是需要学习到合适的权重才能取得最好结果。从结果上看,相对于单一分支的最好结果,使用注意力模块后,在 Market1501, CUHK03 和 DukeMTMC-ReID 数据集上,rank1 分别提升了 0.4%,2.7% 和 0.8%;mAP 分别提升了 1%,3.4% 和 0.5%。相对于手工设置等权重融合的结果,rank1 分别提升了 0.1%,2.2% 和 0.9%;mAP 分别提升了 0.2%,1.6% 和 0.3%。实验表明本文提出的注意力模块方法是有效的。

通过注意力模块,避免了手工设置权重的烦琐过程。此外,每幅图像自身的特征不同,4 个分支的权重分配也应该存在差异。而手工设置对数据集中的每幅图像,4 个分支都是相同权重,显然不是最好方式。而引入注意力模块,可以让网络根据每幅图像自身的特征,学习最好的加权权重,使得各个分支中具有强分辨力的特征,能够在最终特征中得到加强,进而加权融合得到最好特征。

图 7 是网络训练后选取的部分图像学习到的权重。可以发现,不同图像的分支权重间趋势上大体相同,但存在差异,进一步验证了注意力模块的有效性。

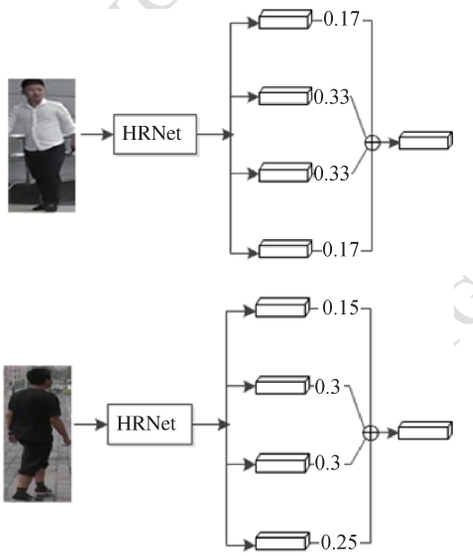


图 7 不同图像学习到的权重也不相同

Fig. 7 Different images learn different weights

2.4 与当前流行方法的比较

在 Market1501、CUHK03 和 DukeMTMC 数据集上,将本文方法及使用 re-ranking(Zhong 等,2017)后的结果与当前行人再识别中的一些流行方法进行对比实验,其中 CUHK03 数据集使用的是检测器标注的子集,结果如表 5 所示。从表 5 可以看出,Pyramid 方法(Zheng 等,2018)在 CUHK03 数据集上的结果非常高,但这是通过连接 21 个不同分类器训练的局部特征得到的。而本文方法只训练 1 个分类器,在 Market1501 数据集上的 mAP 达到了 89.2%,rank1 的准确率基本持平;在 DukeMTMC-ReID 数据集上,rank1 和 mAP 分别达到 90.5% 和 81.5%,在 CUHK03 上的结果也比较好。使用 re-ranking 后,本文方法在 3 个数据集上的 rank1 和 mAP 指标均为最优。

2.5 注意力模块的位置影响评测

注意力机制可分为前注意力机制和后注意力机制,如图 8 所示。前者先将 HRNet 网络的第 1、2、4 个分支通过卷积操作,使输出通道数和分辨率与第 3 个分支相同,通过全局最大池化模块(GMP)操作,各个分支的输出大小为[N,128],然后连接注意力模块,

表 5　本文方法与其他方法的对比

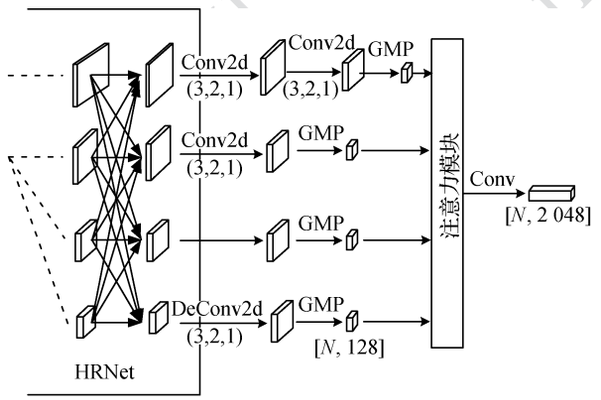
Table 5　Comparison between other methods and ours

/%

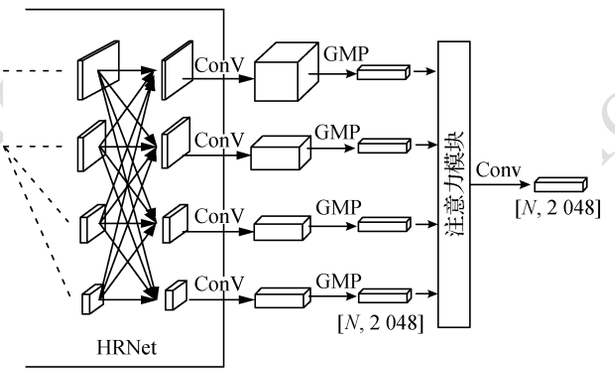
方法	特征数量	Market1501		CUHK03		DukeMTMC-ReID	
		rank1	mAP	rank1	mAP	rank1	mAP
PCB(Sun 等,2018)	6	93.8	81.6	–	–	83.3	69.2
Pyramid(Zheng 等,2018)	21	*95.7*	88.2	*78.9*	*74.8*	89.0	79.0
BEF((batch feature ereasing)Dai 等,2018)	2	94.5	85.0	74.4	70.8	88.7	75.8
MGN(Wang 等, 2018)	8	*95.7*	86.9	68.0	66.0	88.7	78.4
Baseline(Luo 等,2019)	1	94.5	85.9	–	–	86.4	76.4
本文	1	95.3	*89.2*	72.8	70.4	*90.5*	*81.5*
本文 + re-ranking	1	**96.0**	**94.8**	**80.1**	**82.5**	**91.5**	**91.5**

注:加粗字体和斜体分别表示每列最优和次优值,“–”表示未进行实验。

在注意力模块后通过 1×1 卷积操作将通道数变为 2 048。后者先将各个分支的通道数通过 1×1 卷积变为 2 048,然后连接注意力模块。两者的主要区别是注意力模块是在 1×1 的卷积操作之前还是之后。



(a) 前注意力机制



(b) 后注意力机制

图 8　两种注意力机制

Fig. 8　Two different attention mechanisms

((a) front attention mechanism;

(b) behind attention mechanism)

实验在 Market1501 数据集上进行,结果如表 6 所示,可以看出,后注意力机制的效果更好。这是因为后注意力机制是先将 4 个分支的输出特征进行通道数的上采样,获得各自分辨率上丰富的特征信息后,再用于注意力特征融合,因而可以得到更好的结果。

表 6　注意力模块的不同位置得到的结果

Table 6　Results of different position of attention modules

/%

位置	Market1501		DukeMTMC-ReID	
	rank1	mAP	rank1	mAP
前注意力(平均)	95.1	88.6	89.7	80.5
前注意力(最好)	95.2	88.7	90.0	80.6
后注意力(平均)	95.2	88.9	90.3	81.5
后注意力(最好)	**95.3**	**89.2**	**90.5**	**81.7**

注:加粗字体为每列最优值。

2.6　图像分辨率的影响

在 Market1501 数据集上,选择后注意力机制进行实验,研究不同图像分辨率对结果的影响。卷积输出通道数设置为 2 048,结果如表 7 所示。可以看出,图像分辨率对结果没有太大影响。

2.7　权重归一化方法比较

Softmax(Zhang 等,2019)和 Sigmoid(Hu 等,2018)两种方法都可以将特征归一化得到最后权重,在 Market1501 和 DukeMTMC 数据集上分别进行实验,结果如表 8 所示。可以看出,Sigmoid 进行权重归一化的效果好于 Softmax 方法。

表 7　图像分辨率对结果的影响
Table 7　Influence of the image resolution

/%

分辨率/像素	Market1501		DukeMTMC-ReID	
	rank1	mAP	rank1	mAP
256 × 128	**95.3**	89.2	**90.5**	81.5
224 × 224	95.1	88.4	90.0	80.7
384 × 128	95.1	89.1	90.3	81.4
384 × 192	95.1	**89.2**	90.4	**82.0**

注：加粗字体为每列最优值。

表 8　使用 Softmax 和 Sigmoid 进行权重归一化的结果
Table 8　Results of weight normalization by Sigmoid and Softmax

/%

权重归一化方式	Market1501		DukeMTMC-ReID	
	rank1	mAP	rank1	mAP
Softmax	95.0	88.7	90.3	81.1
Sigmoid	**95.3**	**89.2**	**90.5**	**81.5**

注：加粗字体为每列最优值。

3　结　论

基于 HRNet 网络和注意力机制提出了一种多分辨率特征注意力融合的行人再识别方法。HRNet 网络通过交错卷积能够输出多分辨率的特征。本文对不同分支的特征进行评测，发现不同分支的输出特征在不同数据集上表现不一致，因此借助注意力机制，引入一个注意力模块用于融合不同分支的特征。在 Market1501、CUHK03 和 DukeMTMC-ReID 数据集上的实验表明，使用注意力模块后，mAP 在 Market1501 上达到了 89.2% 的最好结果，rank1 在 DukeMTMC-ReID 上达到了 91.5% 的最好结果。并且对网络的相关参数进行实验，得到了各个数据集上的最优参数配置。

在接下来的工作中，将对 HRNet 网络本身进行研究，深入探索该网络中提出的交错卷积的优势，并与传统网络进行对比，尝试更改网络结构重新进行训练，探索不同分支、不同数量的模块对网络性能的影响。

参考文献(References)

Dai Z Z, Chen M Q, Gu X D, Zhu S Y and Tan P. 2018. Batch dropblock network for person re-identification and beyond [EB/OL]. (2018-11-17) [2019-09-20]. https://arxiv.org/pdf/1811.07130.pdf

Fu X Y, Qi Q, Huang Y, Ding X H, Wu F and Paisley J. 2018. A deep tree-structured fusion model for single image deraining [EB/OL]. (2018-11-21) [2019-09-20]. https://arxiv.org/pdf/1811.08632.pdf

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. LasVegas, USA: IEEE: 770-778 [DOI: 10.1109/cvpr.2016.90]

Hu J, Shen L, Albanie S, Sun G and Wu E H. 2018. Squeeze-and-excitation networks [EB/OL]. (2018-07-24) [2019-09-20]. https://arxiv.org/pdf/1709.01507.pdf

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 2261-2269 [DOI: 10.1109/cvpr.2017.243]

Li X, Wang W H, Hu X L and Yang J. 2019. Selective kernel networks [EB/OL]. (2019-03-15) [2019-09-20]. https://arxiv.org/pdf/1903.06586.pdf

Lin Y T, Zheng L, Zheng Z D, Wu Y, Hu Z L, Yan C G and Yang Y. 2019. Improving person re-identification by attribute and identity learning. Pattern Recognition, 95(1): 151-161 [DOI: 10.1016/j.patcog.2019.06.006]

Luo H, Gu Y Z, Liao X Y, Lai S Q and Jiang W. 2019. Bag of tricks and a strong baseline for deep person re-identification [EB/OL]. (2019-03-17) [2019-09-20]. https://arxiv.org/pdf/1903.07071.pdf

Park J, Woo S, Lee J Y and Kweon I S. 2018. Bam: bottleneck attention module [EB/OL]. (2018-07-17) [2019-09-20]. https://arxiv.org/pdf/1807.06514.pdf

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74]

Simonyan K and Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014-09-04) [2019-09-20]. https://arxiv.org/pdf/1409.1556.pdf

Sun K, Xiao B, Liu D and Wang J D. 2019. Deep high-resolution representation learning for human pose estimation [EB/OL]. (2019-02-25) [2019-09-20]. https://arxiv.org/pdf/1902.09212.pdf

Sun Y F, Zheng L, Yang Y, Tian Q and Wang S J. 2018. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 501-518 [DOI: 10.1007/978-3-030-01225-0_30]

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan

D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 1-9 [DOI: 10.1109/cvpr.2015.7298594]

Tian M Q, Yi S, Li H S, Li S H, Zhang X S, Shi J P, Yan J J and Wang X G. 2018. Eliminating background-bias for robust person reidentification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake, USA: IEEE: 5794-5803 [DOI: 10.1109/CVPR.2018.00607]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I. 2017. Attention is all you need [EB/OL]. (2017-06-12) [2019-09-20]. https://arxiv.org/pdf/1706.03762.pdf

Wang G S, Yuan Y F, Chen X, Li J W and Zhou X. 2018. Learning discriminative features with multiple granularities for person re-identification//Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM: 274-282 [DOI: 10.1145/3240508.3240552]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R and Bengio Y. 2015. Show, attend and tell: neural image caption generation with visual attention [EB/OL]. (2015-02-10) [2019-09-20]. https://arxiv.org/pdf/1502.03044.pdf

Zhang W, Hu S N, Liu K and Zha Z J. 2019. Learning compact appearance representation for video-based person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 29(8): 2442-2452 [DOI: 10.1109/TCSVT.2018.2865749]

Zheng F, Deng C, Sun X, Jiang X Y, Guo X W, Yu Z Q, Huang F Y and Ji R R. 2018. Pyramidal person re-identification via multi-loss dynamic training [EB/OL]. (2018-10-29) [2019-09-20]. https://arxiv.org/pdf/1810.12193.pdf

Zhong Z, Zheng L, Cao D L and Li S Z. 2017. Re-ranking person re-identification with k-reciprocal encoding//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 3652-3661 [DOI: 10.1109/CVPR.2017.389]

作者简介

沈庆，1994 年生，男，硕士研究生，主要研究方向为人体再识别。
E-mail：ericsqxd@163.com

田畅，男，教授，主要研究方向为计算机网络、图像处理。
E-mail：tianchang_cce@163.com

王家宝，男，讲师，主要研究方向为人体再识别。
E-mail：jiabao_1108@163.com

焦珊珊，女，博士研究生，主要研究方向为人体再识别。
E-mail：597241031@qq.com

杜麟，男，博士研究生，主要研究方向为人体再识别。
E-mail：du_ming_lin@126.com