计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800200

一种基于改进三元组损失和特征融合的行人重识别方法

张新峰　宋　博
北京工业大学信息学部　北京 100124
(zxf@bjut.edu.cn)

摘　要　行人重识别旨在跨摄像头条件下,从目标数据库中检索出特定的行人目标,其在视频监控领域有重要的应用价值。目前其研究难点为样本图像类内差异大、类间差异小,因此如何设计并训练深度神经网络对行人图片提取一个判别力更强的特征成为了其关键。针对以往研究只单独进行全局特征或局部特征学习的不足,提出了一种联合全局特征和局部特征学习的网络结构,该结构能够同时提取全局特征和具有较强区分力的局部细节特征;针对每部分局部特征对行人特征描述的重要性不同,文中提出了一种局部特征的融合方式,该方法能够自适应地生成各个局部特征的权重,最后将融合后的局部特征和全局特征结合使行人特征得到更全面的表征;另外,针对以往的基于难样本挖掘的三元组损失具有优化目标模糊的特点,提出了一种改进的基于难样本挖掘的三元组损失函数。文中分别在行人重识别主流数据集 Market-1501 和 DukeMTMC-reID 上验证了所提方法的有效性,其 mAP 值分别达到了 82.16%和 74.02%,Rank-1 值分别达到了 92.75%和 86.8%。

关键词:行人重识别;检索;三元组损失;特征融合;深度学习
中图法分类号　TP391.4

A Person Re-identification Method Based on Improved Triple Loss and Feature Fusion

ZHANG Xin-feng and SONG Bo
Faculty of Information Technology,Beijing University of Technology,Beijing 100124,China

Abstract　Person re-identification aims to retrieve specific pedestrian targets from the target database under the condition of cross camera. It has important application value in the field of video surveillance. At present,the difficulty of the research is that the sample images have large intra class differences and small inter class differences. Therefore,how to design and train the deep neural network to extract a more discriminative feature from pedestrian images is the key. In this paper,we propose a network structure combining global features and local features learning,which can extract global features and local features simultaneously. In view of the different importance of each part of the local features to the description of pedestrian features,this paper proposes a fusion method of local features,which can adaptively generate the weight of each local feature. Finally,the local features and global features are combined to make the pedestrian features get more comprehensive representation. In addition,in view of the fuzzy optimization objective of the previous triple loss based on hard sample mining,this paper proposes an improved triple loss function based on hard sample mining. The effectiveness of the proposed method is verified on the mainstream person re-identification data sets Market-1501 and DukeMTMC-reID,respectively,and the mAP values are 82.16% and 74.02%,and the Rank-1 values are 92.75% and 86.8%,respectively.

Keywords　Person re-identification,Retrieval,Triplet loss,Feature fusion,Deep learning

1　引言

行人重识别旨在跨摄像头下,从目标数据库中匹配出特定的行人目标。其在平安城市和天网工程等项目中有着重要的应用,对提高刑侦效率和促进社会公共安全具有很高的应用价值。由于不同监控摄像头拍摄到的行人图片背景、姿态、视角不同,同时又有光照、遮挡等的影响,使得不同身份的行人图片很相似,而同一身份的行人图片有很大的差异,即目前行人重识别的研究难点为类内差异大、类间差异小。如何对行人图片提取一个鲁棒性更强的特征,是目前行人重识别领域的研究重点。

一个完整的行人重识别系统主要包含行人特征提取和行人特征匹配两部分。传统方法的研究思路通常是先对行人图片提取手工特征,如颜色、纹理、HOG(Histogram of Oriented Gradient)[1]、LOMO (Local Maximal Occurrence)[2-3]、SIFT (Scale Invariant Feature Transform)[4] 等,然后利用 XQDA (Cross-view Quadratic Discriminant Analysis)[2] 或者 KISSME(Keep It Simple and Straightforward Metric)[5] 等来学习一个最佳的相似性度量进行行人特征的匹配[6]。由于传统的手工特征描述能力有限,很难提取到复杂且具有判别力

到稿日期:2020-08-29　返修日期:2020-10-18　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
通信作者:宋博(bosong0812@foxmail.com)

的行人特征。随着深度学习技术的兴起，行人重识别方法受到了很大的挑战，目前基于深度学习的方法在性能上已经大大超过了传统的识别方法。

目前基于深度学习的行人重识别方法主要有表征学习和度量学习的方式[6]。表征学习的方法为训练网络学习一个有效的表征，常用的做法是把行人重识别看作一个分类问题，将行人的身份作为标签使网络进行训练，使用时则去掉分类层，把分类层前的输出作为提取到的行人图片的特征。针对整幅图片进行的表征学习提取的行人全局特征较以往手工特征在性能上有很大的提升，但仍然不具有很强的判别力，其原因主要是网络没有学习到具有很强区分力的细节特征[7]。如何让网络去关注关键的局部区域进行学习，进而提取到有效的局部特征成为了研究的热点。目前处理这个问题主要有 3 种方式：1）通过人体姿态检测模型检测出人体的各个部分，对各个部分分别进行有效的学习，之后再进行融合；2）通过额外的属性标签信息，让网络学习图片中的细节部分；3）对行人图片特征进行水平切分，分别对切割出的水平条带进行学习。这 3 种方式的第 1 种人体姿态检测模型的加入会使得网络增加更多的噪声，第 2 种加入额外的属性标签会使得工作量变得巨大，第 3 种方式简单且效果较好，是目前进行局部特征学习被广泛使用的方式。本文认为，无论是关注行人整体的全局特征还是更关注行人部分的局部特征，对行人重识别的性能都是至关重要的，通过实验对该结论进行了验证，因此提出了一个联合全局和局部特征学习的网络。其中局部特征学习的部分采用了直接对行人图片水平切分的方式，但与常规水平切分方法不同的是，本文增加了一个能自适应学习每个局部特征重要性的模块，在局部特征进行融合时基于此重要性模块进行融合，经实验证明此模块可以使得行人重识别的性能得到有效的提升，同时实验结果表明，联合全局和局部特征学习的方式比单独使用其中之一的方式在性能上更加优越。

行人重识别的度量学习方法则在不同度量损失下训练网络，使同一身份 ID（identity）的行人图片在特征空间的距离近，不同身份 ID 的行人图片距离远。三元组损失（Triplet Loss）是常被使用的度量损失函数，基于此损失函数的网络结构要求输入 3 幅图片，这 3 幅图片分别记为锚点（anchor）、正样本（positive）和负样本（negative），这样就组成了一个三元组。其中，anchor 和 postive 为一对正样本对，anchor 和 negative 为一对负样本对。网络的目标是通过训练让正样本对在特征空间的距离小于负样本对。由于从样本集中任意选取 anchor，positive，negative 组成的三元组本身已经很容易地满足损失函数的要求，网络并不能得到有效的训练[8]，因此如何选取一个有效的三元组成为此方法的关键。文献[8]提出一种难三元组（hard triplets）方式，其具体做法为：1）先从数据集中选择 P 类行人，再从每类行人选择 K 幅图片，由 $P\times K$ 幅图片组成一个 batch；2）将这个 batch 中的每幅图片都作为一个 anchor，计算并选择在特征空间中与它同一个 ID 距离最远的一幅图片（称难正样本对，hard positive）和在特征空间中与它不同 ID 距离最近的一幅图片（称难负样本对，hard negative）；3）将这个由 anchor，hard positive 和 hard negative 组成的三元组称为难三元组，利用 Triplet Loss 进行网络的训练，使得在特征空间难正样本对之间的距离小于难负样本对。上述 3 个过程也常被称为难样本的挖掘方式。以往的基于难样本挖掘的 Triplet Loss 有其局限性，如图 1 所示，本文将与 anchor 同一 ID 并在特征空间距离最远的一幅图片记为 hard positive，将与 anchor 同一 ID 并在特征空间距离最近的一幅图片记为 nearest positive，将与 anchor 不同 ID 且在特征空间距离最近的一幅图片记为 hard negative。图 1(a)为样本图片在特征空间中的一种距离表示，图 1(b)和图 1(c)为 anchor，hard positive 和 hard negative 组成的三元组，经过现有基于 Triplet Loss 的网络训练后得到的可能结果。我们发现图 1(c)的结果并不是理想的结果，因为属于同一 ID 的 anchor 和 nearest positive 在特征空间的距离大于不同 ID 的 anchor 和 hard negative 之间的距离，因此以往的基于难样本挖掘的 Triplet Loss 具有优化目标模糊的特点。针对此问题，本文提出了一种改进的基于难样本挖掘的 Triplet Loss 函数。实验结果表明，该损失函数可以使行人重识别的性能得到一定的提升。

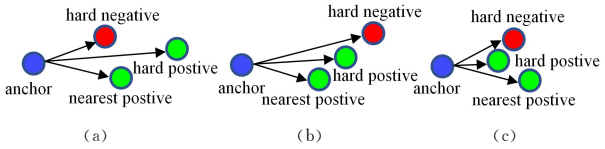

图 1　行人图片在特征空间的距离

Fig. 1　Distance of pedestrian image in feature space

综上所述，针对已有的问题，本文的主要工作如下：1）提出了一个联合全局和局部特征学习的网络，较单独使用全局或局部的方式在性能上更加优越；2）提出了一个自适应学习每个局部特征重要性的模块，对局部特征能够进行有效的融合，使对行人特征的描述更加紧凑且有效；3）提出了一个改进的基于难样本挖掘的三元组损失函数，使得网络学习到更全面的特征。

2　相关工作

由于深度神经网络技术强大的特征提取能力，人们使用这一技术进行行人重识别的研究已经成为这个领域普遍的做法。文献[9-10]首先将深度学习应用在行人重识别领域中，较同时期基于手工特征的方法在性能上有很大的提升。文献[11]把行人重识别模型的训练过程当作一个多分类问题，并使用 Resnet-50 作为主干网络对行人图片学习一个有判别力的特征表示，此网络结构作为基准模型被行人重识别领域广泛使用。为了提取到判别力更强的特征，一些关注行人关键局部区域的方法被提出。文献[12-13]先利用人体姿态检测模型检测出 14 个关键点，然后利用这些关键点提取人体具有语义的局部，如头、胳膊等，利用这些局部来训练网络以提取更具有判别力的特征，这在一定程度上解决了姿态变化的问题，但人体姿态模型的检测错误会给网络引入额外的噪声。文献[14]分别对 Market-1501 和 DukeMTMC-reID 数据集标注了行人属性，在网络中加入属性监督的信息，使模型学习到更加细粒度的特征，但是加入额外的属性标签会大大增加额外的工作量。文献[15]根据人体结构的先验知识，提出了一种直接对特征图进行水平切分的方法，用于学习每个水平条

带的局部特征，大大提高了行人重识别的性能。根据文献[15]的思路，文献[16]同样使用水平切分特征图的方式，并使用了多个支流来学习行人图片的多粒度特征，使行人重识别的性能再一次得到了提升。

损失函数作为网络训练参数更新的引导，对指引网络学习到具有判别力的行人语义特征起着关键的作用。在行人重识别领域中常被用到的损失有分类损失和度量损失。把行人重识别看作一个分类问题来设计模型时，交叉熵损失(Cross-Entropy Loss)常作为此类模型的损失函数。而用度量学习的方法进行行人重识别研究时，常被用到的损失函数有验证损失(Verification Loss)[9]、三元组损失(Triplet Loss)[17]。Verification Loss 要求输入网络一系列的样本对来使得在特征空间类内距离小、类间距离大。Triplet Loss 则通过一对正样本对和一对负样本的形式，使得在特征空间上正样本对之间的距离小于负样本对之间的距离来训练网络。文献[8]提出了一个基于难样本挖掘的 Triplet Loss 来学习行人特征的有效表达。文献[18]使用了 CrossEntropy Loss 和 Verification Loss 联合进行网络的训练，相比使用单一损失函数其在性能上有一定的提升。目前使用分类损失和度量损失一起训练网络的策略被很多文献采用[16,19-20]。

3 本文方法

针对以往的工作只单独进行全局特征和局部特征学习，不足以学习到一个具有强判别力的特征，本文提出了联合全局特征和局部特征学习的方式。由于每部分局部特征对描述行人特征的重要性不同，本文提出了一种局部特征的融合方式。针对以往 Triplet Loss 的局限性，本文通过改进 Triplet Loss，提出了一种能更好地学习训练样本特征的损失函数。

3.1 整体网络结构

图 2 给出了本文的整体网络结构，主要由特征提取主干网络(Backbone)、全局特征提取模块(I_1路)、局部特征提取模块组成(I_2路)，其中局部特征提取模块又可细分为 1 个全局特征提取支流(z_2^g路)、6 个局部特征提取支流($z_2^{p_1}$～$z_2^{p_6}$路)和 1 个局部特征融合支流(z_2^w路)。本文使用 ResNet-50[21]作为基本网络。ResNet-50 网络由 1 层卷积层(conv1)和 4 个残差模块(conv2_x～conv5_x)组成，每个残差块由多个卷积层、批量规范化层(Batch Normalization)和 ReLU 激活函数组成。

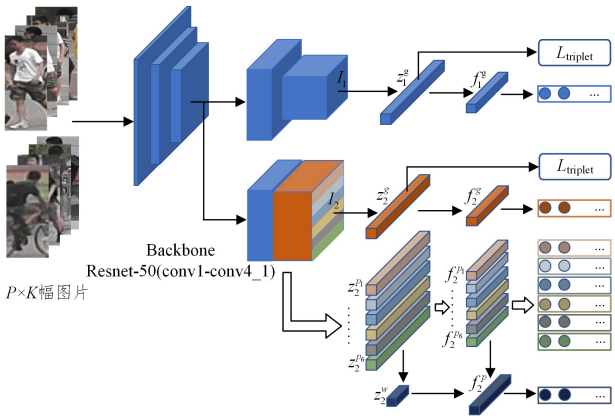

图 2 网络结构图

Fig. 2 Network architecture

为了让网络分别提取到行人图片的全局和局部特征并使其互不影响，本文对 ResNet-50 做了相应的改进。考虑到 conv4_x 层提取到的特征具有一定高层语义的特点，同时也保留了一定的细节特征，因此从 conv4_1 开始把之后的网络分别分成全局特征提取模块和局部特征提取模块，且其网络权值不共享，并使 conv4_1 之前的网络结构及参数保持不变。全局特征提取模块的 conv4_2～conv5_x 部分与原始的 ResNet-50 保持一致，对 conv5 层的输出特征图执行全局平局池化(GAP)的操作，并通过一个 1×1 的卷积、Batch Normalization 层和 ReLU 层进行降维。文献[21]提出在分类层前增加一层 BNNeck 能够提升系统的性能，因此本文采用了这种方法对降维之后的特征图增加一层 BNNeck，即对特征图进行 Batch Normaliztion 后再输入分类层。为了增加特征图的大小，使特征图中包含更多的细粒度的特征，本文去除了局部特征提取模块中 conv5 模块中的下采样操作。参考文献[16]，本文对局部特征模块增加了一条全局特征提取支流，其 conv5 后的网络结构与对全局特征提取模块的设置一样。局部特征提取支流的设置则参考了文献[15]中的方法，把 conv5 的输出特征图按从上到下平均分成 6 份，对于每份特征图均与对全局特征提取支流的参数设置一样，进行身份标签监督下的学习。局部特征融合支流通过自动学习每个局部特征支流输出所对应的权重，然后对每个局部特征以加权求和的方式融合得到融合后的行人特征，对此特征也进行身份标签监督下的学习。

在测试阶段，将全局特征提取模块提取到的全局特征、局部特征提取模块提取到的全局特征和融合后的局部特征拼接在一起，作为最终的行人特征。

3.2 局部特征融合支流

如图 2 所示，通过对局部特征提取模块(I_2路)的行人特征图进行水平分割，并经过全局平均池化和卷积等操作后，可以得到关于行人局部的 6 个特征描述，即$\{\boldsymbol{f}_2^{p_1}, \boldsymbol{f}_2^{p_2}, \cdots, \boldsymbol{f}_2^{p_6}\}$。常用的做法是直接拼接这 6 个局部特征，并将其作为最终行人特征的表达，但由于行人图片之间存在不对齐的现象[22]，这会使得每部分局部特征对整个行人特征描述的重要性不同，因此学习每个局部特征的重要性，并按其的重要性大小进行融合有利于行人特征更好的表达。图 2 中z_2^w路即为本文提出的局部特征融合支流，通过对特征图I_2进行水平分割，并进行全局平局池化操作得到的$\boldsymbol{z}_2^{p_1}, \boldsymbol{z}_2^{p_2}, \cdots, \boldsymbol{z}_2^{p_6}$使用核大小 6×1 的卷积、批量规范化层和 Sigmoid 激活函数，可以得到局部特征权重向量$\boldsymbol{z}_2^w \in \boldsymbol{R}^6$。然后，对每个局部特征以加权求和的方式进行融合，得到融合后的局部特征$\boldsymbol{f}_2^p$，过程如式(1)所示：

$$\boldsymbol{f}_2^p = \sum_{k=1}^{6} \boldsymbol{z}_2^{w_k} \boldsymbol{z}_2^{p_k} \tag{1}$$

3.3 损失函数

为了学习到更具有判别力的特征，本文在网络损失函数的使用上联合了 CrossEntropy Loss 和 Triplet Loss，用 CrossEntropy Loss 进行分类学习，用 Triplet Loss 进行度量学习，具体如下。

本文将 CrossEntropy Loss 应用在了网络结构图中的所有模块中的每个支流上。在全局特征提取模块中，对于 conv5 输出的行人特征图$\boldsymbol{I}_1 \in \boldsymbol{R}^{h \times w \times d}$，先用全局平均池化

(Global Average Pooling，GAP)对 $\boldsymbol{I}_1$ 处理得到特征 $\boldsymbol{z}_1^g \in \boldsymbol{R}^{1\times1\times d}$，随后使用 1×1 的卷积层、批量规范化层和 ReLU 激活函数对特征 z 进行降维，得到全局特征 $\boldsymbol{f}_1^g \in \boldsymbol{R}^{1\times1\times v}$，调整 $\boldsymbol{f}_1^g$ 的维度格式使 $\boldsymbol{f}_1^g \in \boldsymbol{R}^v$。局部特征模块的全局支流的设置与全局特征提取模块的设置一样，而对于 6 条局部支流的设置，则先对特征图 I_2 进行平均切分再进行全局平均池化等一系列的设置。对于局部特征融合支流，其激活函数为 Sigmoid。训练时对特征 $\boldsymbol{f}_i^l$ 使用全连接层和 Softmax 激活函数得到行人身份的分类结果，最后使用交叉熵损失函数作为目标函数。相应的过程如下：

$$\hat{p}^{(id)} = \mathrm{softmax}(\boldsymbol{w}^{(id)} \boldsymbol{f}_i^l + \boldsymbol{b}^{(id)}) \tag{2}$$

$$L_{id} = -\sum_i^N p_i{}^{(id)} \log \hat{p}_i{}^{(id)} \tag{3}$$

其中，N 为训练集中行人的类别数，$\boldsymbol{w}^{(id)} \in \boldsymbol{R}^{N\times v}$ 和 $\boldsymbol{b}^{(id)} \in \boldsymbol{R}^N$ 分别是全连接层的权重矩阵和神经元的偏置向量，$\hat{\boldsymbol{p}}^{(id)} \in \boldsymbol{R}^N$ 为输出的行人身份的预测概率，$\boldsymbol{p}^{(id)} \in \boldsymbol{R}^N$ 为一个样本 one-hot 形式的真实标签，L_{id} 为网络对于一个样本最终输出的交叉熵损失。

对于三元组损失，在网络中输入 3 幅图片 anchor，positive，negative，其中 2 幅图片为同一 ID(anchor 和 positive)组成一对正样本对，2 幅图片为不同 ID(anchor 和 negative)组成一对负样本对，这个损失函数的目标是，网络经过训练后使得同一 ID 的行人图片在特征空间的距离比不同 ID 的行人图片小。对于三元组损失函数，本文应用在了未降维的全局特征上 $(\boldsymbol{z}_1^g, \boldsymbol{z}_2^g) \in \boldsymbol{R}^{1\times1\times d}$。非难的三元组无法使网络得到有效的优化，因此本文采用了文献[8]提出的难样本对挖掘方法。对于一个 batch，随机选择 P 个行人，每个行人选择 K 幅图片，则一个 batch 的 loss 计算表达式如式(4)所示：

$$L_{\text{triplet}} = \sum_{i=1}^{P}\sum_{a=1}^{K}[\alpha + \max_{p=1,\cdots,K} D(\boldsymbol{f}_\theta(\boldsymbol{x}_a^i), \boldsymbol{f}_\theta(\boldsymbol{x}_p^i)) - \min_{\substack{j=1,\cdots,P\\ n=1,\cdots,K\\ j\neq i}} D(\boldsymbol{f}_\theta(\boldsymbol{x}_a^i), \boldsymbol{f}_\theta(\boldsymbol{x}_n^i))]_+ \tag{4}$$

其中，$D(\cdot)$ 为计算特征空间中两幅图片的欧氏距离，$\boldsymbol{f}_\theta(\boldsymbol{x}_a^i)$，$\boldsymbol{f}_\theta(\boldsymbol{x}_p^i)$，$\boldsymbol{f}_\theta(\boldsymbol{x}_n^i)$ 分别为对 anchor，postive，negative 进行特征提取得到的特征，参数 α 用于控制样本对间的相对距离。

由于常用的基于难样本挖掘的 Triplet Loss 有其局限性，网络更新后新的三元组间仍然不满足在特征空间上正样本对之间的距离小于负样本对之间距离的要求，因此本文提出了一种改进基于难样本挖掘的 Triplet Loss，我们使用了一个负样本对和两个正样本对。要求在特征空间上 anchor 与同一 ID 距离最远的样本和同一 ID 距离最近的样本都小于与 anchor 不同 ID 距离最近的样本，其表达式如式(5)所示：

$$L_{\text{triplet\_our}} = \sum_{i=1}^{P}\sum_{a=1}^{K}\{[\alpha + \max_{p=1,\cdots,K} D(\boldsymbol{f}_\theta(\boldsymbol{x}_a^i), \boldsymbol{f}_\theta(\boldsymbol{x}_p^i)) - \min_{\substack{j=1,\cdots,P\\ n=1,\cdots,K\\ j\neq i}} D(\boldsymbol{f}_\theta(\boldsymbol{x}_a^i), \boldsymbol{f}_\theta(\boldsymbol{x}_n^i))]_+ + [\beta + \min_{p=1,\cdots,K} D(\boldsymbol{f}_\theta(\boldsymbol{x}_a^i), \boldsymbol{f}_\theta(\boldsymbol{x}_p^i)) - \min_{\substack{j=1,\cdots,P\\ n=1,\cdots,K\\ j\neq i}} D(\boldsymbol{f}_\theta(\boldsymbol{x}_a^i), \boldsymbol{f}_\theta(\boldsymbol{x}_n^i))]_+\} \tag{5}$$

其中，$D(\cdot)$ 仍为计算两个向量的欧氏距离，α 和 β 为两个超参数用来控制样本对间的相对距离。

4 实验

4.1 数据集和评价标准

本文在两个常用的公开数据集上进行实验来评价本文方法，分别为 Market-1501 和 DukeMTMC-reID，下面将详细介绍这两个数据和本文所使用的评价标准。

Market-1501[23]数据集采集于清华大学的一家超市前，使用了 6 个摄像头，其中包括 5 个高分辨率摄像头和一个低分辨率摄像头。这个数据集共有 1 501 个行人 ID，其中训练集有 12 936 张图片，包含了 751 个行人 ID，测试部分中候选集包含了 750 个行人 ID 和一些干扰项的共 19 732 张图片，另外还有 3 368 张待检索的行人图片。

DukeMTMC-reID[24]采集自杜克大学内，它是行人跟踪数据集 DukeMTMC 的一个子集，使用了 8 个室外摄像头。这个数据集共包含 1 812 个行人 ID，其中有 1 402 个行人 ID 出现在两个摄像头以上。1 402 个行人 ID 中有 702 个行人 ID 共 16 522 张图像用于训练，另外，702 个行人 ID 和 408 个干扰行人共 17 661 张图片作为候选集，待检索的行人图片共 2 228 张。

本文使用累计匹配特性(Cumulative Match Characteristic，CMC)曲线和平均精度均值(Mean Average Precision，mAP)来评估本文方法的准确性，其中 CMC 主要计算击中率 Rank-1 指标。

4.2 实验设置

本文使用 pytorch 搭建整个网络，参考文献[15]，把输入图片的尺寸设置为 384×128，使用水平随机翻转图片进行数据增强。训练时使用在 ImageNet 上预训练的参数对网络进行初始化，使用随机梯度下降(SGD)优化器来进行梯度更新，其初始学习率设置为 0.01，weight_decay 为 0.000 5，momentum 为 0.9。每个 batch 输入 32 张图片，一共训练 70 轮，每隔 20 轮后所有参数的学习率调整为之前的 1/10。

4.3 实验结果分析

本节首先将本文方法与现有相关方法进行了比较，并分析了现有方法的不足。之后通过网络参数设置和网络的消融实验对本文算法进行了具体的分析。

4.3.1 与相关方法的比较和分析

表 1 列出了本文在 Market-1501 和 DukeMTMC-reID 数据集上与相关方法的比较。与传统方法 LOMO＋XQDA[2]和 LOMO＋Null Space[3]相比，我们的方法表现出了更好的性能。在网络结构方面，基于深度学习的方法(如 PIE[12]，PSE[13]和 Spindle[25])利用人体姿态模型提取了人体姿态的相关信息，并将其加入训练以期望网络能够学习到更具有判别力的细粒度特征，但人体姿态模型的检测错误会给网络引入额外的噪声，使得基于此思路的行人重识别方法的整体性能都不高[26]。APR[14]利用了额外的属性标签信息，使网络学习图片中的细节部分，但是加入额外的属性标签会使得行人标注的工作量变得巨大。PCB-RPP[15]对行人图片进行水平切分，分别对切割出的水平条带进行学习，但是没有加入全局特征。从网络损失上看，以往的 TriNet[8]使用了基于难样

本挖掘的 Triplet Loss 损失函数，但这种损失函数有其局限性。针对以上方法存在的问题，本文方法在 Market-1501 上 Rank-1 和 mAP 分别达到了 92.75%和 82.16%，较基于人体姿态模型最好的方法 PSE[13] 分别提升了 5.05%和 13.16%，超过基于属性的方法 APR[14] 5.75%和 15.26%，相比 TriNet[8] 分别提升了 7.83%和 13.02%，与 PCB-RPP[15] 相比，本文在 mAP 上高出 0.56%，在 Rank-1 上几乎一样，说明本文方法对 mAP 的影响更大，但与 PCB-RPP[15] 相比，本文通过一个局部特征融合模块融合了局部特征，使最后提取到的行人特征维度更小，能够有效减小行人特征存储时的内存消耗。另外，在 DukeMTMC-reID 数据集上本文方法仍然表现很好。

表 1 与相关方法的比较

Table 1 Comparison with related methods

（单位：%）

方法	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
LOMO+XQDA[2]	22.22	43.79	—	—
LOMO+Null Space[3]	29.87	55.43	—	—
PIE[12]	53.87	78.65	—	—
PSE[13]	69.0	87.7	62.0	79.8
Spindle[25]	—	76.9	—	—
APR[14]	66.9	87.0	55.6	73.9
TriNet[8]	69.14	84.92	—	—
PCB-RPP[15]	81.6	93.8	69.2	83.3
本文方法	82.16	92.75	74.02	86.80

4.3.2 网络参数设置

(1)不同 α 和 β

α 和 β 为本文所改进的 Triplet Loss 中的两个超参数，用于控制样本对间的相对距离。表 2 列出了不同 α，β 组合下在 Market-1501 上进行实验所得到的 mAP 和 Rank-1 值。实验结果表明，不同的 α 和 β 的组合会对结果产生一定的影响，本文在 Market-1501 数据集上，当 α 取 0.3，β 取 0.7 时能使网络取得较好的结果。

表 2 不同 α，β 组合下的性能比较

Table 2 Comparison using different combinations of α and β

（单位：%）

α	β	mAP	Rank-1
0.1	0.3	79.44	92.31
0.1	0.5	79.02	92.34
0.1	0.7	79.90	91.81
0.1	1	79.05	91.72
0.3	0.5	80.00	92.73
0.3	0.7	81.01	92.61
0.3	1	79.15	91.83
0.5	0.7	82.16	92.75
0.5	1	79.36	92.34
0.7	1	79.81	92.22

(2)不同特征维度 f

为了说明特征空间的维度对模型性能的影响，本文分别设特征维度为 512，1024 和 2048，在 Market-1501 数据集上进行实验。从表 3 中的实验结果可以看到，特征维度过大或者过小都不利于模型性能的提升。当特征维度取 1024 时，本文模型的性能指标达到较好的结果。

表 3 不同特征维度组合 f 下的性能比较

Table 3 Comparison of different feature dimension

（单位：%）

特征维度	mAP	Rank-1
512	80.58	92.19
1024	82.16	92.75
2048	80.51	92.87

4.3.3 网络消融实验

(1)网络结构的消融

表 4 列出了本文在 Market-1501 和 DukeMTMC-reID 数据集上，设置损失函数为 CrossEntropy Loss 和 Triplet Loss，特征维度为 1024 时不同网络结构的组合下的性能结果。全局特征提取模块指网络只是使用了图 2 的 I_1 路部分，没有对 I_2 路特征图进行水平条带的分割。局部特征模块指图 2 的 I_2 路部分，其后的 z_2^g 路为局部特征提取模块的全局支流，z_2^w 路为局部特征提取模块的局部特征融合支流。由表 4 可以看到，在 Market-1501 数据集上包含全局特征提取模块和局部特征提取模块的网络结构比单独只是用全局特征提取模块的网络在 Rank-1 和 mAP 上分别高出了 6.39%和 7.65%，说明在设计网络的同时提取行人图片的全局特征和局部特征是有效且有必要的。包含局部特征融合支流的全局特征提取模块和局部特征提取模块的网络结构比不包含局部特征融合支流的全局特征提取模块和局部特征提取模块的网络结构在 mAP 上高出了 0.93%，但在 Rank-1 的指标上没有明显的提升，说明局部特征融合支流对 mAP 的指标影响更大。mAP 衡量了返回的图片中与待查询图片为同一 ID 的图片靠前的位置，在一定程度上说明了局部特征融合支流的按局部特征的权重进行融合的方法能够让网络最后提取的特征更具有鲁棒性，进而让图像库中与待查询行人图片为同一 ID 的图片返回时尽可能靠前。

表 4 不同网络结构组合下的性能比较

Table 4 Performance comparison of different network structure combinations

（单位：%）

网络结构	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
全局特征提取模块	71.30	86.28	60.56	78.37
全局特征提取模块+局部特征提取模块(不含局部特征融合支流)	78.95	92.67	72.21	85.86
全局特征提取模块+局部特征提取模块(不含全局支流)	77.35	91.81	69.65	84.34
全局特征提取模块+局部特征提取模块	79.88	92.28	72.84	85.41

(2)损失函数的消融

表 5 列出了本文在 Market-1501 和 DukeMTMC-reID 数据集上，设置包含全局特征提取模块和局部特征提取模块的网络结构下，特征维度为 1024 时，使用不同网络损失函数组合得到的性能结果。从表 5 可以看到，使用 CrossEntropy Loss 和 Triplet Loss 共同训练网络比单独使用 CrossEntropy Loss 在 Market-1501 上的 Rank-1 和 mAP 分别高出了 1.01%和 3.56%，在 DukeMTMC-reID 上 Rank-1 和 mAP 上分别高出了 2.16%和 4.54%，说明联合两个 loss 训练是有效的。使用 CrossEntropy Loss 和本文改进的 Triplet Loss 比使

用 CrossEntropy Loss 和以往的 Triplet Loss 在 Market-1501 上的 Rank-1 和 mAP 分别高出了 0.47%和 2.28%，在 DukeMTMC-reID 上的 Rank-1 和 mAP 分别高出了 1.39%和 1.18%，可见本文所改进的 Triplet Loss 有效提升了模型的性能。

表 5 不同损失函数组合下的性能比较

Table 5 Performance comparison of different loss function

（单位：%）

网络结构	Market-1501		DukeMTMC-reID	
	mAP	Rank-1	mAP	Rank-1
CrossEntropy Loss	76.32	91.27	68.30	83.25
CrossEntropy Loss＋Triplet Loss	79.88	92.28	72.84	85.41
CrossEntropy Loss＋本文改进的 Triplet Loss	82.16	92.75	74.02	86.80

4.3.4 可视化分析

图 3 为对同一幅待查询图片，分别使用不同的特征检索到的前 10 幅图片。图 3 中，上面一组为只使用全局特征的方式，下面一组为使用了全局特征加融合后的局部特征，其中匹配正确的行人图片用绿色的方框标出。通过比较发现，只使用图片的全局特征的方式只关注图像的整体外观，在行人外观相似的情况下这种方式不能很好地将同一 ID 的行人检索出，而加入了融合后的包含图像细节部分的局部特征后能够有效地改善检索的结果。

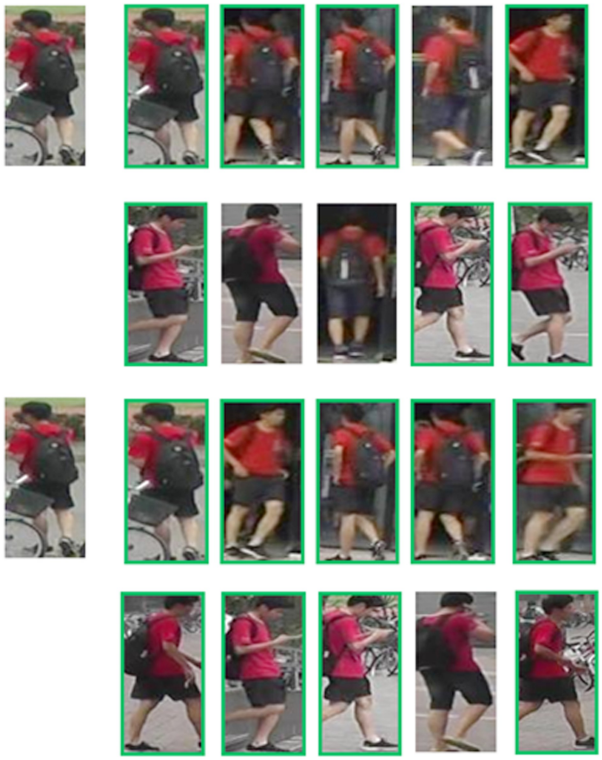

图 3 使用不同特征检索返回的前 10 幅图片（电子版为彩色）

Fig. 3 Top 10 returned images using different features

结束语 本文提出了一种联合全局特征和局部特征进行学习的网络结构，并提出了一种局部特征的融合方式，能够自适应地生成各个局部的权重，通过将融合后的局部特征和全局特征结合来共同描述行人的特征。此外，针对以往基于难样本挖掘的 Triplet Loss 具有优化目标模糊的特点，本文提出了一种改进的 Triplet Loss 函数。在 Market-1501 和 DukeMTMC-reID 上的实验的结果表明，本文方法能够有效地描述行人的特征。实验结果表明，为行人特征的描述增加像局部特征这种细粒度特征，有利于行人重识别模型性能的提升，由于多尺度特征能体现出更细粒度的部分，因此本文后续的工作将考虑如何设计网络来提取行人的多尺度特征。

参 考 文 献

[1] LIU X, SONG M, TAO D, et al. Semi-supervised coupled dictionary learning for person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 3550-3557.

[2] LIAO S, HU Y, ZHU X, et al. Person re-identification by local maximal occurrence representation and metric learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2197-2206.

[3] ZHANG L, XIANG T, GONG S. Learning a discriminative null space for person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1239-1248.

[4] ZHAO R, OUYANG W, WANG X. Learning mid-level filters for person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 144-151.

[5] KOESTINGER M, HIRZER M, WOHLHART P, et al. Large scale metric learning from equivalence constraints[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 2288-2295.

[6] LUO H, JIANG W, FAN X, et al. A survey on deep learning based person re-identiflcation [J]. Acta Automatica Sinica, 2019, 45(11): 2032-2049.

[7] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: Past, present and future[J]. arXiv: 1610.02984, 2016.

[8] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv: 1703.07737.

[9] YI D, LEI Z, LIAO S, et al. Deep metric learning for person reidentification [C]// 2014 22nd International Conference on Pattern Recognition. IEEE, 2014: 34-39.

[10] LI W, ZHAO R, XIAO T, et al. Deepreid: Deep filter pairing neural network for person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 152-159.

[11] ZHENG L, ZHANG H, SUN S, et al. Person re-identification in the wild[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1367-1376.

[12] ZHENG L, HUANG Y, LU H, et al. Pose Invariant Embedding for Deep Person Re-identification[J]. arXiv: 1701.07732, 2017.

[13] SU C, LI J, ZHANG S, et al. Pose-driven deep convolutional model for person re-identification[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 3960-3969.

[14] LIN Y, ZHENG L, ZHENG Z, et al. Improving Person Re-iden-

tification by Attribute and Identity Learning[J]. arXiv: 1703.07220, 2017.

[15] SUN Y, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 480-496.

[16] WANG G, YUAN Y, CHEN X, et al. Learning discriminative features with multiple granularities for person re-identification [C]// Proceedings of the 26th ACM International Conference on Multi-media. 2018: 274-282.

[17] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: A unified embedding for face recognition and clustering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 815-823.

[18] ZHENG Z, ZHENG L, YANG Y. A discriminatively learned cnn embedding for person reidentification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2017, 14(1): 1-20.

[19] YE M, SHEN J, LIN G, et al. Deep Learning for Person Reidentification: A Survey and Outlook[J]. arXiv: 2001. 04193, 2020.

[20] LUO H, GU Y, LIAO X, et al. Bag of tricks and a strong baseline for deep person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019: 1487-1495.

[21] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[22] ZHENG Z, ZHENG L, YANG Y. Pedestrian alignment network for large-scale person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(10): 3037-3045.

[23] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: A benchmark[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1116-1124.

[24] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]// European Conference on Computer Vision. Cham: Springer, 2016: 17-35.

[25] ZHAO H, TIAN M, SUN S, et al. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1077-1085.

[26] SHAO X, SHUAI H, LIU Q. Person re-identification method combining attribute features[J/OL]. Acta Automatica Sinica. https://doi.org/10.16383/j.aas.c190763.

ZHANG Xin-feng, born in 1974, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include image processing and machine learning.

SONG Bo, born in 1993, postgraduate. His main research interests include person re-identification and data mining.