融合属性特征的行人重识别方法

邵晓雯 [1] 帅 惠 [1] 刘青山 [1]

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. 由于不同的行人可能具有相似的外观, 因此要求行人重识别模型能够捕捉到充足的细粒度特征. 本文提出一种融合属性特征的行人重识别的深度网络方法, 将行人重识别和属性识别集成在分类网络中, 进行端到端的多任务学习. 此外, 对于每张输入图片, 网络自适应地生成对应于每个属性的权重, 并将所有属性的特征以加权求和的方式结合起来, 与全局特征一起用于行人重识别任务. 全局特征关注行人的整体外观, 而属性特征关注细节区域, 两者相互补充可以对行人进行更全面的描述. 在行人重识别的主流数据集 DukeMTMC-reID 和 Market-1501 上的实验结果表明了本文方法的有效性, 平均精度均值 (Mean average precision, mAP) 分别达到了 74.2% 和 83.5%, Rank-1 值分别达到了 87.1% 和 93.6%. 此外, 在这两个数据集上的属性识别也得到了比较好的结果.

关键词 行人重识别, 属性识别, 深度学习, 自适应权重

引用格式 邵晓雯, 帅惠, 刘青山. 融合属性特征的行人重识别方法. 自动化学报, 2022, **48**(2): 564−571

DOI 10.16383/j.aas.c190763

Person Re-identification Based on Fused Attribute Features

SHAO Xiao-Wen[1] SHUAI Hui[1] LIU Qing-Shan[1]

Abstract Person re-identification aims to retrieve specific pedestrian target across surveillance devices. Since different pedestrians may have a similar appearance, it requires the person re-identification model to capture sufficient fine-grained features. This paper proposes a new deep network method for person re-identification based on fused attribute features. Person re-identification and attribute recognition are integrated into the classification network for end-to-end multi-task learning. In addition, for each input image, the network adaptively generates weights corresponding to each attribute. Features of all attributes are combined in a weighted summation way, and together with the global feature for person re-identification task. Global feature focuses on the overall appearance of the pedestrian, while the attribute feature focuses on the detail area, and they can complement each other to provide a more comprehensive description of the pedestrian. Experimental results on the DukeMTMC-reID and Market-1501 datasets, two popular datasets of person re-identification, show the effectiveness of the proposed method. The mean average precision (mAP) values reach 74.2% and 83.5% respectively, and the Rank-1 values reach 87.1% and 93.6% respectively. In addition, attribute recognition on these two datasets also achieved better results.

Key words Person re-identification, attribute recognition, deep learning, adaptive weight

Citation Shao Xiao-Wen, Shuai Hui, Liu Qing-Shan. Person re-identification based on fused attribute features. *Acta Automatica Sinica*, 2022, **48**(2): 564−571

行人重识别指跨监控设备下的行人检索问题, 在公共安全、智能监控等领域具有广泛的应用. 具体而言, 给定一张行人图片, 行人重识别用来在其他摄像头拍摄的大型图片库中找出该行人的图片. 由于监控图片的分辨率低, 且不同的图片之间存在光照、姿态、摄像头视角等方面的差异, 行人重识别目前仍是一个很有挑战性的问题.

收稿日期 2019-11-04 录用日期 2020-04-06
Manuscript received November 4, 2019; accepted April 6, 2020
国家自然科学基金 (61532009, 61825601) 资助
Supported by National Natural Science Foundation of China (61532009, 61825601)
本文责任编委 赖剑煌
Recommended by Associate Editor LAI Jian-Huang
1. 南京信息工程大学自动化学院江苏省大数据分析技术重点实验室 南京 210044
1. Jiangsu Key Laboratory of Big Data Analysis Technology, School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044

1 相关研究

早期行人重识别的研究思路通常是先对行人图片提取手工特征, 如颜色直方图、方向梯度直方图 (Histogram of oriented gradient, HOG)[1] 等, 然后使用相似性度量方法, 如大边界最近邻算法 (Large margin nearest neighbor, LMNN)[2]、交叉二次判别分析算法 (Cross-view quadratic discriminant analysis, XQDA)[3] 等来学习度量矩阵. 为了克服光照、成像条件等因素影响, 采用多特征分析是常用的一种方式[4−6]. 随着深度学习技术的兴起, 深度学习广泛应用于行人重识别任务中. 目前, 基于深度学习的行人重识别方法在性能上大大超过了传统方法[7],

主要有如下两个原因: 1) 手工设计的特征描述能力有限, 而深度学习使用深度卷积神经网络可以自动学习出更复杂的特征; 2) 深度学习可以将特征提取和相似性度量联合在一起, 实现端到端的学习, 从而得到全局最优解.

目前基于深度学习的行人重识别方法主要分为度量学习和表征学习方法[8]. 度量学习通过设计不同的度量损失来约束特征空间, 使得同一个行人的不同图片在特征空间上距离很近, 而不同行人的距离很远, 如三元组损失 (Triplet loss)[9]、四元组损失 (Quadruplet loss)[10] 和群组相似性学习 (Group similarity learning)[11] 等方法. 这类方法的关键在于样本对的选取, 由于大量样本对简单易于区分, 随机采样将会导致网络的泛化能力有限, 因而需要挑选出一些难样本对进行训练. Zhu 等[12] 对困难和简单的负样本设计不同的目标函数来学习距离度量方法, 以充分利用负样本中的信息. 相对于表征学习, 度量学习的训练时间更长, 收敛也更困难. 因此, 表征学习方法得到了更加广泛的研究.

表征学习方法在训练网络时将行人重识别当作身份分类任务来学习行人特征, 关键问题是如何设计网络以学习到更具有判别力的特征. Sun 等[13] 根据人体结构的先验知识, 在垂直方向上对特征图均匀分块, 然后提取每个区域的局部特征. 还有一些方法利用额外的语义信息, 例如骨骼关键点、分割结果等, 定位行人的各个部位. Su 等[14] 借助关键点检测模型对人体区域定位、裁剪、归一化后, 拼接成新的图片作为网络的输入. Sarfraz 等[15] 将行人 14 个关键点的位置响应图和原图片一起输入到网络中, 让网络自动地学习对齐. Kalayeh 等[16] 在 LIP (Look into person)[17] 数据集上训练人体解析模型来预测 4 个人体部位和背景, 然后在特征图上提取这些部位的特征.

由于不同的行人可能具有相似的外观, 而同一个行人在不同的环境下存在很大差异, 只从全局外观的角度无法进行正确匹配. 行人的属性, 例如性别、是否背包、头发长短等, 包含丰富的语义信息, 可以为行人重识别提供关键的判别线索. 早期的研究中, Layne 等[18] 手工标注了 15 种语义属性来描述行人, 包括性别、服装种类、是否携带物品等, 并使用支持向量机 (Support vector machine, SVM) 训练属性分类器, 最后与底层特征融合得到行人图像的最终特征描述. 随着深度学习的广泛应用, Zhu 等[19] 在一个卷积神经网络中同时预测多个属性, 在 PETA (Pedestrian attribute)[20] 数据集上的属性识别性能明显优于基于 SVM 的方法. Schumann 等[21] 先在 PETA 数据集上训练属性识别模型, 然后在行人重识别模型中利用属性预测的结果, 使得网络可以学习到与属性互补的特征. 该方法分开训练两个网络, 无法充分利用属性标签和身份标签, 导致行人重识别的性能比较低. Lin 等[22] 在行人重识别数据集 DukeMTMC-reID[23] 和 Market1501[24] 上标注了行人属性, 并提出 APR (Attribute-person recognition) 模型实现行人重识别和属性识别的多任务学习, 同时将属性预测的结果和全局特征一起用于行人重识别任务. 该方法使用属性的预测结果, 当属性识别错误时, 会给行人重识别引入噪声. Tay 等[25] 提出了 AANet (Attribute attention network), 将行人属性和属性的激活区域图集成到分类网络中来解决行人重识别问题, 得到了比较好的检索结果. 上述方法同等对待所有属性, 忽略了每个属性对每张图片的重要性是不同的.

针对以上问题, 本文提出了融合属性特征的行人重识别方法, 主要工作如下: 1) 将行人重识别和属性识别集成到分类网络中进行端到端的学习; 2) 为了减小属性识别错误对行人重识别的影响, 从特征的角度利用属性信息; 3) 自适应地生成对应于每个属性的权重, 并将所有属性特征以加权求和的方式结合起来, 与全局特征一起用于行人重识别任务. 在 DukeMTMC-reID 和 Market-1501 数据集上的实验结果表明了本文方法的有效性.

2　融合属性特征的行人重识别模型

图 1 为本文的网络结构图, 前半部分为提取图片特征的主干网络, 后半部分分为身份分类、属性识别和属性特征融合三个分支. 身份分类分支对行人的全局特征进行身份的预测; 属性识别分支用来预测行人的各个属性; 属性特征融合分支首先以自适应的方式对属性特征加权求和, 得到融合后的属性特征, 然后对该特征进行身份预测.

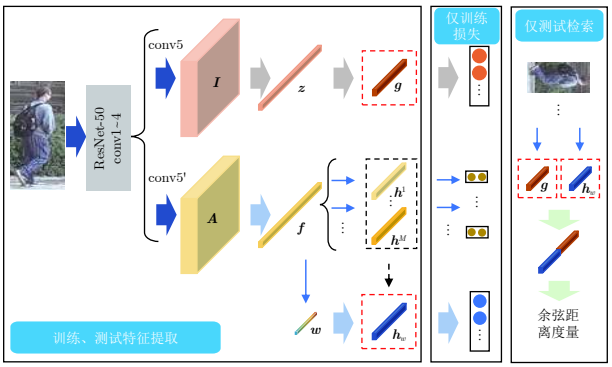


图 1　网络结构示意图

Fig. 1　Schematic diagram of network structure

2.1　主干网络结构

使用 ResNet-50 作为主干网络提取图片的特

征. ResNet-50 包含 1 层卷积层 (conv1) 和 4 个残差模块 (conv2∼conv5), 每个残差模块包含多层卷积层、批量规范化层和线性整流激活函数 (Rectified linear units, ReLU). 文献 [26] 提出多个相关性低甚至相反的任务一起学习时, 在共享参数上会产生相互竞争甚至相反的梯度方向, 从而影响所有任务的学习. 为了减轻任务间的干扰, 在 ResNet-50 的第 4 个模块 conv4 后将网络分成两个分支, 分别学习行人的全局特征和属性特征, 即两个分支中 conv5 模块的参数不共享. 根据文献 [13], 本文去除了两个分支的 conv5 模块中的下采样操作, 以增加特征图的大小、丰富特征的粒度. 将大小为 256×128 像素的图片输入网络时, 可以从 conv5 模块输出大小为 16×8 的特征图.

设 $S=\{(x_1,y_1,\boldsymbol{a}_1),\cdots,(x_n,y_n,\boldsymbol{a}_n)\}$ 为训练数据集, 其中 n 是图片的张数, x_i 表示第 i 张图片, $y_i\in\{1,2,\cdots,N\}$ 表示该图片的身份标签, N 是训练集中行人的个数, $\boldsymbol{a}_i=\left[a_i^1,a_i^2,\cdots,a_i^M\right]\in\mathbf{R}^M$ 表示这张图片的属性标签, M 是属性的个数, 对于 DukeMTMC-reID 和 Market-1501 数据集, M 分别是 10 和 12, $a_i^j\in\left\{1,2,\cdots,C^j\right\}$ 指这张图片的第 j 个属性的标签, C^j 表示第 j 个属性的类别个数. 如图 1 所示, 对于 $(x,y,\boldsymbol{a})\in S$, 将图片 x 输入到网络, 可以分别得到对应于身份分类的特征图 $I\in\mathbf{R}^{h\times w\times d}$ 和属性识别的特征图 $A\in\mathbf{R}^{h\times w\times d}$.

2.2 身份分类

对于身份特征图 $I\in\mathbf{R}^{h\times w\times d}$, 先用全局平均池化 (Global average pooling, GAP) 对 I 处理得到特征 $\boldsymbol{z}\in\mathbf{R}^d$, 随后使用全连接 (Fully connected, FC) 层、批量规范化层和 ReLU 激活函数对特征 $\boldsymbol{z}$ 进行降维, 得到全局特征 $\boldsymbol{g}\in\mathbf{R}^v$. 训练时对特征 $\boldsymbol{g}$ 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. 为了防止训练时出现过拟合的问题, 对身份标签进行平滑操作 (Label smoothing, LS)[27], LS 是分类任务中防止过拟合的常用方法. 相应的过程如下:

$$\hat{\boldsymbol{p}}^{(\mathrm{id})}=\mathrm{softmax}\left(W^{(\mathrm{id})}\boldsymbol{g}+\boldsymbol{b}^{(\mathrm{id})}\right) \tag{1}$$

$$p_i^{(\mathrm{id})}=\begin{cases}1-\dfrac{N-1}{N}\varepsilon, & i=y\\ \dfrac{\varepsilon}{N}, & i\neq y\end{cases} \tag{2}$$

$$L_{\mathrm{id}}=-\sum_{i=1}^{N}p_i^{(\mathrm{id})}\ln\hat{p}_i^{(\mathrm{id})} \tag{3}$$

其中, N 为训练集中行人的个数, $W^{(\mathrm{id})}\in\mathbf{R}^{N\times v}$ 和 $\boldsymbol{b}^{(\mathrm{id})}\in\mathbf{R}^N$ 分别是全连接层的权重矩阵和偏差向量, $\hat{\boldsymbol{p}}^{(\mathrm{id})}\in\mathbf{R}^N$ 为输出的行人身份的预测概率. 式 (2) 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, 文中 $\varepsilon=0.1$. L_{id} 为网络的身份分类损失.

2.3 属性识别

在属性识别分支中, 与身份分类类似, 先用 GAP 对属性特征图 $A\in\mathbf{R}^{h\times w\times d}$ 处理, 得到特征 $\boldsymbol{f}\in\mathbf{R}^d$, 然后使用 M 层全连接层对特征 $\boldsymbol{f}$ 进行提取, 得到 M 个属性特征 $\{\boldsymbol{h}^1,\boldsymbol{h}^2,\cdots,\boldsymbol{h}^M\}$. 对于每一个属性特征 $\boldsymbol{h}^k\in\mathbf{R}^v$, 使用全连接层和 Softmax 激活函数得到对应的属性分类结果, 最后使用交叉熵损失作为目标函数. 相应的过程如下:

$$\hat{\boldsymbol{p}}^k=\mathrm{softmax}\left(W^k\boldsymbol{h}^k+\boldsymbol{b}^k\right) \tag{4}$$

$$p_j^k=\begin{cases}1, & j=a^k\\ 0, & j\neq a^k\end{cases} \tag{5}$$

$$L_k=-\sum_{j=1}^{C^k}p_j^k\ln\hat{p}_j^k \tag{6}$$

对于第 k 个属性, C^k 表示它的类别个数, $W^k\in\mathbf{R}^{C^k\times v}$ 和 $\boldsymbol{b}^k\in\mathbf{R}^{C^k}$ 分别是对应的全连接层的权重矩阵和偏差向量, $\hat{\boldsymbol{p}}^k\in\mathbf{R}^{C^k}$ 为该属性的预测结果, L_k 为第 k 个属性的分类损失.

由于属性各个类别的样本比例不平衡, 并且为了降低大量简单样本在训练中所占的权重, 对于每个属性使用加权的焦点损失 (Focal loss) 函数[28], 更改后的属性损失函数如下:

$$L_k=-\frac{1}{C^k}\sum_{j=1}^{C^k}\frac{1}{w_j^k}\left(1-\hat{p}_j^k\right)^{\gamma}p_j^k\ln\hat{p}_j^k \tag{7}$$

$$L_{\mathrm{att}}=\sum_{k=1}^{M}L_k \tag{8}$$

其中, w_j^k 是第 k 个属性的类别 j 在训练集中所占的比例, $\gamma>0$ 使得训练更关注于易于错分的样本, 文中 $\gamma=2$, L_{att} 为总的属性损失.

2.4 属性特征的融合

如果直接应用属性的预测结果, 当属性预测错误时, 很容易给行人重识别任务引入噪声, 因此从特征的角度对属性加以利用. 属性特征更关注于行人图片的某个区域, 因而可以融合所有属性的特征和全局特征互相补充. 直接想法是将提取到的 M 个属性特征 $\{\boldsymbol{h}^1,\boldsymbol{h}^2,\cdots,\boldsymbol{h}^M\}$ 以相加或相连等方式进行融合, 但是对于每张图片, 每个属性的重要性是不同的, 如果简单地对每个属性分配相同的权重, 最终可能会降低属性信息带来的益处. 因此, 对于

每张图片, 网络都会自适应地生成每个属性对应的权重, 用来融合属性特征. 具体方法如下: 对于图片 x 得到的特征 $\boldsymbol{f}\in\mathbf{R}^d$, 首先使用一层全连接层和 Sigmoid 激活函数得到对应于每个属性特征的权重, 具体表示为

$$\boldsymbol{w}=\text{sigmoid}\left(W^{(\text{weight})}\boldsymbol{f}+\boldsymbol{b}^{(\text{weight})}\right) \tag{9}$$

其中, $W^{(\text{weight})}\in\mathbf{R}^{M\times d}$ 和 $\boldsymbol{b}^{(\text{weight})}\in\mathbf{R}^M$ 分别表示全连接层的权重矩阵和偏差向量, 得到的属性权重向量 $\boldsymbol{w}=[w_1,w_2,\cdots,w_M]$, $w_i\in(0,1)$. 然后对每个属性特征以加权求和的方式, 即 $\boldsymbol{h}_w=\sum_{k=1}^{M}w_k\boldsymbol{h}^k$, 得到融合后的属性特征 $\boldsymbol{h}_w\in\mathbf{R}^v$. 随后以额外监督的方式对特征 $\boldsymbol{h}_w$ 进行行人的身份分类, 具体与上述的分类过程相同, 使用全连接层和 Softmax 激活函数得到分类结果 $\hat{\boldsymbol{p}}^{(\text{local})}$, 最后根据身份标签使用带有 LS 的交叉熵损失函数作为优化目标, 得到损失函数 L_{local}. 训练时 L_{local} 可以监督属性权重和属性特征的生成, L_{local} 表示为

$$L_{\text{local}}=-\sum_{i=1}^{N}p_i^{(\text{id})}\ln\hat{p}_i^{(\text{local})} \tag{10}$$

其中, $p_i^{(\text{id})}$为式 (2) 对身份标签进行 LS 操作得到的结果.

总的损失函数包括全局特征、属性融合特征的身份分类损失和属性分类损失, 具体表示为

$$L=L_{\text{id}}+\alpha L_{\text{local}}+\beta L_{\text{att}} \tag{11}$$

其中, α 和 β 是平衡三个损失的权重因子. 测试时, 将全局特征 $\boldsymbol{g}$ 和融合后的属性特征 $\boldsymbol{h}_w$ 相连得到行人总的特征 $\boldsymbol{e}=[\boldsymbol{g};\boldsymbol{h}_w]\in\mathbf{R}^{2v}$, 并使用余弦距离计算特征间的距离, 排序后得到检索结果.

3　实验结果与分析

本文的实验基于行人重识别的主流数据集 DukeMTMC-reID 和 Market-1501 进行评测, 并与 PCB-RPP (Part-based convolutional baseline and refined part pooling)、PDC (Pose-driven deep convolutional)、ACRN (Attribute-complementary re-id net) 等相关方法进行了对比. 使用 Pytorch 框架搭建整个网络, 图片的大小缩放为 256 × 128 像素, 仅使用水平随机翻转作为数据增强的方法. 训练时使用 Adam 优化器更新梯度, 初始学习率设为 0.0003, weight_decay 为 0.0005. batch size 设为 32, 总共训练 60 个 epoch, 每隔 20 个 epoch, 所有参数的学习率降为之前的 0.1 倍. 使用 ImageNet[29] 上的预训练参数对网络初始化, 在前 10 个 epoch 中, 使用 ImageNet 初始化的参数保持不变, 仅更新随机初始化的参数. 降维之后的特征维度为 512, 即 $v=512$.

3.1　数据集和评估指标

DukeMTMC-reID 数据集共有 1 404 个行人, 分为训练集 16 522 张图片包含 702 个行人, 测试集 17 661 张图片包含另外的 702 个行人和 408 个干扰行人, 另外有 2 228 张图片作为待检索的行人图片. Lin 等[22] 对数据集中的每个行人标注了 23 个属性, 本文使用了所有的属性, 并把 8 个上衣颜色的属性作为 1 个类别数是 8 的属性, 同样将 7 个下衣颜色的属性作为 1 个类别数是 7 的属性, 最后得到 10 个属性.

Market-1501 数据集共有 1 501 个行人, 分为训练集 12 936 张图片包含 751 个行人, 测试集 19 732 张图片包含 750 个行人, 另外还有 3 368 张图片作为待检索的行人图片. Lin 等[22] 对数据集中的每个行人标注了 27 个属性, 本文使用了所有的属性, 对上衣颜色和下衣颜色采取上述的组合方式, 最后得到 12 个属性.

对于行人重识别任务, 使用标准的评估指标: 平均精度均值 (Mean average precision, mAP) 和累计匹配特性 (Cumulative match characteristic, CMC) 曲线. 对于属性识别任务, 本文对每个属性使用分类准确率进行评估, 同时计算了所有属性的平均分类准确率.

3.2　与其他方法的比较

表 1 是本文在 DukeMTMC-reID 和 Market-1501 数据集上与当前相关方法的比较. PCB-RPP 对特征图均匀分块, 未考虑行人图片没有对齐的情形, 而且没有去除背景的干扰. PDC 和 PSE (Pose-sensitive embedding) 利用额外的姿态估计模型, SPReID (Semantic parsing for re-identification) 利用额外的人体解析模型, 来定位行人的各个部位, 这种方法由于不能端到端地学习, 训练好的部件定

表 1　与相关方法的性能比较 (%)
Table 1　Performance comparison with related methods (%)

方法	DukeMTMC-reID		Market-1501	
	mAP	Rank-1	mAP	Rank-1
PCB-RPP[13]	69.2	83.3	81.6	93.8
PDC[14]	—	—	63.4	84.4
PSE[15]	62.0	79.8	69.0	87.7
SPReID[16]	71.0	84.4	81.3	92.5
ACRN[21]	52.0	72.6	62.6	83.6
APR[22]	55.6	73.9	66.9	87.0
AANet-50[25]	72.6	86.4	82.5	**93.9**
本文	**74.2**	**87.1**	**83.5**	93.6

位模型在行人图片上定位错误时将会引入噪声, 最终影响行人重识别的结果. 表 1 中的下面 3 种方法利用属性标签辅助行人重识别, ACRN 在属性数据集 PETA 上训练属性识别模型, APR 和 AANet-50 没有考虑行人重识别和属性识别之间的关系, 直接使用同一个网络提取两个任务的特征. 此外, 它们对所有属性同等对待, 忽略了各个属性对行人描述的重要性是不同的. 在考虑了以上问题后, 本文的方法在 DukeMTMC-reID 上, mAP 和 Rank-1 值分别达到了 74.2% 和 87.1%, 超过了 AANet-50 的结果 1.6% 和 0.7%, 在 Market-1501 上, mAP 值超过 AANet-50 1.0%, Rank-1 值降低 0.3%, 可见我们的方法对 mAP 影响更大. 而且相比于 AANet-50 使用 CAM (Class activation maps)[30] 定位属性激活区域, 本文的方法更加简单有效.

3.3 本文方法分析

表 2 是 DukeMTMC-reID 上使用不同损失函数得到的检索结果, 使用 L_{id} 相当于只训练身份分类的单支网络, 可以作为基准模型, $L_{\text{id}}+\beta L_{\text{att}}$ 指身份分类和属性识别的多任务学习网络, $L_{\text{id}}+\alpha L_{\text{local}}+\beta L_{\text{att}}$ 指在多任务网络的基础上, 加入了对属性特征进行身份监督的任务, 即本文使用的网络. 由表 2 可以得出以下结论: 基准模型得到了比较好的结果, mAP 和 Rank-1 值分别达到 70.5%和 84.1%; 多任务网络比基准模型在 mAP 和 Rank-1 值上分别提高 0.6% 和 1.7%, 说明加入的属性识别对行人重识别起到了促进的作用, 此时网络不光要正确预测行人的身份, 还需要预测出各个属性, 从而提高了网络的泛化性能; 当利用属性特征时, mAP 和 Rank-1 值分别进一步提高 3.1% 和 1.3%, 说明融合的属性特征可以补充全局特征, 最终形成了更具有判别力的特征. 另外, $L_{\text{id}}+\alpha L_{\text{local}}+\beta L^{*}_{\text{att}}$ 指身份分类和属性识别双支网络中 conv5 模块的参数共享时的结果, 相比于不共享时, mAP 和 Rank-1 值分别降低 1.2% 和 1.1%, 说明这两个任务在提取特征的目标上并不完全相同, 参数共享的模型使得两者在训练时相互影响, 导致网络无法收敛至最优解. 本文使用的网络结构减小了属性识别对行人重识别的干扰, 并以自适应的方式利用属性特征, 最终有效提高了行人的检索结果. L_{id} (no LS) 是在基准模型中没有对身份标签采用 LS 平滑操作的结果, 相比基准模型, mAP 和 Rank-1 值分别降低 3.7% 和 0.8%, 可见 LS 操作有效提升了模型的性能.

表 2 使用不同损失函数的性能比较 (%)
Table 2 Performance comparison using different loss functions (%)

损失函数	mAP	Rank-1
L_{id}	70.5	84.1
$L_{\text{id}}+\beta L_{\text{att}}$	71.1	85.8
$L_{\text{id}}+\alpha L_{\text{local}}+\beta L_{\text{att}}$	74.2	87.1
$L_{\text{id}}+\alpha L_{\text{local}}+\beta L^{*}_{\text{att}}$	73.0	86.0
L_{id} (no LS)	66.8	83.3

表 2 是对属性特征使用不同融合方式的性能比较. 将所有属性的特征相加后, 直接引入到第一支网络中与全局特征相连, 对特征 $\hat{\boldsymbol{g}}=[\boldsymbol{g};\sum_{k=1}^{M}\boldsymbol{h}^k]$ 进行身份分类, 而不进行额外的监督, 该方法记为 var1. 在 var1 的基础上, 对所有属性特征以自适应加权求和的方式进行融合, 而不是直接相加, 即 $\hat{\boldsymbol{g}}=[\boldsymbol{g};\boldsymbol{h}_w]=[\boldsymbol{g};\sum_{k=1}^{M}w_k\boldsymbol{h}^k]$, 该方法记为 var2. 第 3 种方法与本文方法类似, 区别是生成权重的方式不同, 从全局的角度生成对应于每个属性的权重, 记为 var3, 具体如下: 将第一支网络中的全局特征 $\boldsymbol{z}\in\mathbf{R}^d$ 作为输入, 使用全连接层和 Sigmoid 激活函数输出权重 $\boldsymbol{w}$. 由表 3 可知, var2 相比 var1, mAP 和 Rank-1 值分别提高 0.9% 和 1.1%, 可见自适应地对每个属性赋予不同的权重是有作用的, 对于每张图片, 网络可以自动调整各个属性的重要性. 我们的结果相比于 var2, mAP 和 Rank-1 值分别提高 1.7% 和 1.4%, 说明对融合后的属性特征进行额外的身份分类任务可以进一步提升性能, 主要有两个原因: 1) 可以对属性特征和属性权重有更强的监督信息; 2) 由于没有加入到第一支网络中, 从而不会干扰全局特征的学习. 我们的方法相比于 var3, mAP 和 Rank-1 值分别提高 0.1% 和 0.8%, 说明从属性的角度生成权重, 这种类似于自注意力机制的方法可以得到更好的结果. 另外从直觉上讲, 将属性的特征作为输入, 输出对应于每个属性的权重, 这种方式也更加合理.

表 3 的最后两行表示在训练完最终模型后, 分别使用全局特征 $\boldsymbol{g}$ 和属性融合特征 $\boldsymbol{h}_w$ 进行检索的结果. 当只使用 $\boldsymbol{g}$ 测试时, mAP 值为 72.4%, Rank-1 值为 86.1%, 只使用 $\boldsymbol{h}_w$ 测试时, mAP 和 Rank-1

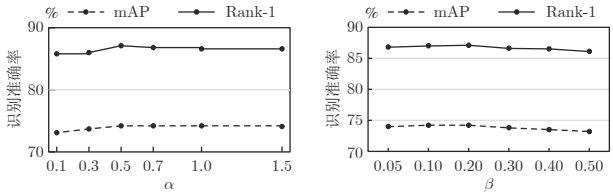
表 3 使用不同特征融合方式的性能比较 (%)
Table 3 Performance comparison using different feature fusion methods (%)

方法	特征维度	mAP	Rank-1
var1	1 024	71.6	84.6
var2	1 024	72.5	85.7
var3	1 024	74.1	86.3
本文	1 024	**74.2**	**87.1**
特征 $\boldsymbol{g}$	512	72.4	86.1
特征 $\boldsymbol{h}_w$	512	71.8	85.1

值分别为 71.8% 和 85.1%, 两者均超过了基准模型的结果. 而将 $\boldsymbol{g}$ 和 $\boldsymbol{h}_w$ 相连后测试时, mAP 和 Rank-1 分别达到了 74.2% 和 87.1%, 可见全局特征和属性特征相互补充, 可以对行人进行更全面的描述.

3.4　网络参数设置

图 2 是在 DukeMTMC-reID 上分别设置不同的 α 和 β 得到的结果, 图 2(a) 中 $\beta = 0.2$, 图 2(b) 中 $\alpha = 0.5$. 由图 2(a) 可知, 对属性融合特征进行过多或者过少的监督, 效果都有所降低, 当 α 为 0.5 时, 可以得到最好的结果. 由图 2(b) 可知, 当 β 取值比较小时, 结果有所提高, 同时为了不影响属性识别的准确率, 将 β 取为 0.2. 在所有实验中, 对 α 和 β 均进行如上设置.



(a) 参数 α　　(b) 参数 β

(a) Parameter α　　(b) Parameter β

图 2　设置不同的 α 和 β 的结果

Fig. 2　Results setting different α and β

3.5　可视化分析

图 3 是模型训练完成之后, 使用 Grad-CAM (Gradient-weighted class activation mapping)[31] 得到的各个属性的可视化结果, 10 个属性依次为 gender, hat, boots, length of upper-body clothing, backpack, handbag, bag, color of shoes, color of upper-body clothing, color of lower-body clothing. 可视化结果下方的数字表示网络生成的对应于该属性的权重. 上方是 DukeMTMC-reID 检索库中的一张图片, 下方的图片是由属性融合特征 $\boldsymbol{h}_w$ 检索出的第 1 张图片, 匹配正确, 即这两张图片属于同一个行人.

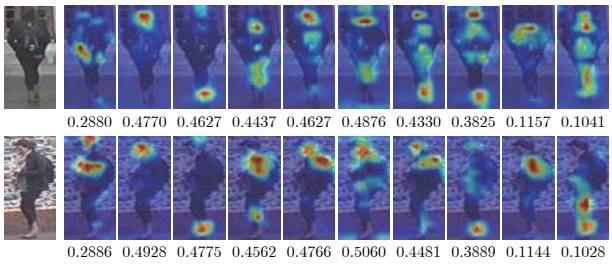


图 3　各个属性的可视化结果及对应的权重值

Fig. 3　Visualization result and corresponding weight value of each attribute

由可视化结果可以看出, 每个属性的激活区域基本都是落在行人区域内, 可见利用属性的特征可以减少背景的干扰. 此外, 帽子、靴子、鞋子颜色、上衣颜色、下衣颜色等属性的激活范围基本符合对应的属性区域. 由生成的权重值可知, 对于这两张图片, 帽子、靴子、背包等属性的重要性很大. 这两张图片由于姿态、光线等差异, 在外观上并不相似, 但网络通过自适应地融合属性特征, 关注头部、脚部、书包等区域, 最终可以正确检索出来.

图 4 是使用不同特征检索到的图片, 其中匹配错误的样本用粗线条的框表示. 对于每个行人, 后面三行分别是使用全局特征 $\boldsymbol{g}$、属性融合特征 $\boldsymbol{h}_w$ 和总的特征 $\boldsymbol{e}$ 得到的检索结果. 由第 1 个行人的结果可知, 全局特征只关注上衣和裤子, 找出的 10 张图片中只有 5 张匹配正确, 而融合特征包含对于这个行人很重要的书包信息, 检索出了 10 张正确的图片. 对于第 2 个行人, 融合特征通过帽子这个属性, 正确找出了被遮挡的图片. 相比于关注整体外观的全局特征, 融合后的属性特征包含很多细节信息, 这对于区分外观相似的行人是很重要的. 对于

图 4　使用不同特征检索到的图片

Fig. 4　Images retrieved by different features

表 4 DukeMTMC-reID 上属性识别的准确率 (%)
Table 4 Accuracy of attribute recognition on DukeMTMC-reID (%)

方法	gender	hat	boots	l.up	b.pack	h.bag	bag	c.shoes	c.up	c.low	平均值
APR[22]	84.2	87.6	87.5	88.4	75.8	93.4	82.9	89.7	74.2	**69.9**	83.4
B2	85.94	89.75	**89.92**	88.64	84.35	**93.61**	**83.06**	90.70	74.62	66.81	84.74
本文	**86.07**	**90.74**	89.72	**88.78**	**84.47**	93.28	82.04	**91.79**	**75.53**	68.14	**85.06**

表 5 Market-1501 上属性识别的准确率 (%)
Table 5 Accuracy of attribute recognition on Market-1501 (%)

方法	gender	age	hair	l.slv	l.low	s.clth	b.pack	h.bag	bag	hat	c.up	c.low	平均值
APR[22]	88.9	**88.6**	84.4	93.6	93.7	92.8	84.9	**90.4**	76.4	97.1	74.0	**73.8**	86.6
AANet-50[25]	92.31	88.21	86.58	**94.45**	94.24	**94.83**	87.77	89.61	**79.72**	98.01	**77.08**	70.81	87.80
B2	93.05	85.82	88.97	93.56	94.18	94.53	88.80	88.08	79.24	98.28	74.95	68.64	87.34
本文	**93.69**	86.16	**89.00**	94.11	**94.80**	94.81	**89.33**	88.68	79.03	**98.33**	76.52	70.68	**87.93**

第 3 个行人, 全局特征检索出的多为行人的背面图片, 而融合特征检索出许多侧面图片, 这两个结果中均有错误, 但当这两个特征相连后, 可以找出 10 张正确的包含各个视角的图片, 说明全局特征和属性特征包含不相同的信息, 可以互相补充促进最后的检索结果.

3.6 属性识别准确率

表 4 和表 5 分别表示 DukeMTMC-reID 和 Market-1501 上各个属性的识别准确率, Avg 指所有属性的平均准确率, B2 表示只训练属性识别的单支网络. 由表中结果可知, 属性识别和行人重识别的多任务网络促进了属性识别的过程, 本文方法相比于 B2, 平均准确率分别提高了 0.32% 和 0.59%. 此外, 与 APR、AANet-50 的结果比较也体现了本文方法的竞争力.

4 结束语

针对行人外观存在类内差异大、类间差异小的问题, 本文提出了一种融合属性特征的行人重识别的深度网络方法. 实验结果表明, 该方法能够通过加入的属性信息丰富行人的特征描述, 提升识别性能. 后续工作将考虑属性之间的依赖关系, 进一步研究如何在行人重识别任务中更好地利用属性标签, 实现行人共有属性的特征匹配.

References

1 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 886−893

2 Weinberger K Q, Saul L K. Fast solvers and efficient implementations for distance metric learning. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008. 1160−1167

3 Liao S C, Hu Y, Zhu X Y, Li S Z. Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 2197−2206

4 Ma F, Jing X Y, Zhu X K, Tang Z M, Peng Z P. True-color and grayscale video person re-identification. *IEEE Transactions on Information Forensics and Security*, 2020, **15**: 115−129

5 Ma F, Zhu X K, Zhang X Y, Yang L, Zuo M, Jing X Y. Low illumination person re-identification. *Multimedia Tools and Applications*, 2019, **78**(1): 337−362

6 Ma F, Zhu X K, Liu Q L, Song C F, Jing X Y, Ye D P. Multi-view coupled dictionary learning for person re-identification. *Neurocomputing*, 2019, **348**: 16−26

7 Luo Hao, Jiang Wei, Fan Xing, Zhang Si-Peng. A survey on deep learning based person re-identification. *Acta Automatica Sinica*, 2019, **45**(11): 2032−2049
(罗浩, 姜伟, 范星, 张思朋. 基于深度学习的行人重识别研究进展. 自动化学报, 2019, **45**(11): 2032−2049)

8 Jing X Y, Zhu X K, Wu F, Hu R M, You X G, Wang Y H, et al. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. *IEEE Transactions on Image Processing*, 2017, **26**(3): 1363−1378

9 Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv preprint, arXiv: 1703.07737, 2017.

10 Chen W H, Chen X T, Zhang J G, Huang K Q. Beyond triplet loss: A deep quadruplet network for person reidentification. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 403−412

11 Chen D P, Xu D, Li H S, Sebe N, Wang X G. Group consistent similarity learning via deep crf for person reidentification. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 8649−8658

12 Zhu X K, Jing X Y, Zhang F, Zhang X Y, You X G, Cui X. Distance learning by mining hard and easy negative samples for person re-identification. *Pattern Recognition*, 2019, **95**: 211−222

13 Sun Y F, Zheng L, Yang Y, Tian Q, Wang S J. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the 2018 European Conference on Computer Vision. Cham: Springer, 2018. 480−496

14 Su C, Li J N, Zhang S L, Xing J L, Gao W, Tian Q. Pose-driven deep convolutional model for person re-identification. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 3960−3969

15 Sarfraz M S, Schumann A, Eberle A, Stiefelhagen R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 420−429

16 Kalayeh M M, Basaran E, Gökmen M, Kamasak M E, Shah M. Human semantic parsing for person re-identification. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 1062−1071

17 Gong K, Liang X D, Zhang D Y, Shen X H, Lin L. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 932−940

18 Layne R, Hospedales T, Gong S G. Person re-identification by attributes. In: Proceedings of the 23rd British Machine Vision Conference. Guildford, UK: BMVA, 2012.

19 Zhu J Q, Liao S C, Lei Z, Li S Z. Multi-label convolutional neural network based pedestrian attribute classification. *Image and Vision Computing*, 2017, **58**: 224−229

20 Deng Y B, Luo P, Loy C C, Tang X O. Pedestrian attribute recognition at far distance. In: Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, Florida, USA: ACM, 2014. 789−792

21 Schumann A, Stiefelhagen R. Person re-identification by deep learning attribute-complementary information. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, Hawaii, USA: IEEE, 2017. 20−28

22 Lin Y T, Zheng L, Zheng Z D, Wu Y, Hu Z L, Yan C G, et al. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 2019, **95**: 151−161

23 Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C. Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of the 2016 European Conference on Computer Vision Workshops. Cham: Springer, 2016. 17−35

24 Zheng L, Shen L Y, Tian L, Wang S J, Wang J D, Tian Q. Scalable person re-identification: A benchmark. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1116−1124

25 Tay C P, Roy S, Yap K H. AANet: Attribute attention network for person re-identifications. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, California, USA: IEEE, 2019. 7134−7143

26 Zhao X Y, Li H X, Shen X H, Liang X D, Wu Y. A modulation module for multi-task learning with applications in image retrieval. In: Proceedings of the 2018 European Conference on Computer Vision. Cham: Springer, 2018. 401−416

27 Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 2818−2826

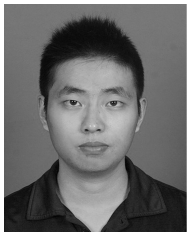
28 Lin T Y, Goyal P, Girshick R, He K M, Dollar P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2980−2988

29 Deng J, Dong W, Socher R, Li L J, Li K, Li F F. Imagenet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, Florida, USA: IEEE, 2009. 248−255

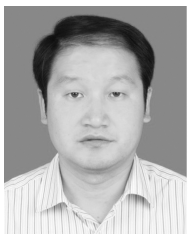
30 Zhou B L, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 2921−2929

31 Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 618−626

邵晓雯 南京信息工程大学自动化学院硕士研究生. 2018 年获得南京信息工程大学电子与信息工程学院学士学位. 主要研究方向为计算机视觉, 行人重识别.
E-mail: xiaowen_shao@nuist.edu.cn
(**SHAO Xiao-Wen** Master student at the School of Automation, Nanjing University of Information Science and Technology. She received her bachelor degree from the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology in 2018. Her research interest covers computer vision and person re-identification.)



帅 惠 南京信息工程大学博士研究生. 2018 年获得南京信息工程大学信息与控制学院硕士学位. 主要研究方向为目标检测, 3D 场景解析.
E-mail: huishuai13@163.com
(**SHUAI Hui** Ph.D. candidate at Nanjing University of Information Science and Technology. He received his master degree from the School of Information and Control, Nanjing University of Information Science and Technology in 2018. His research interest covers object detection and 3D scene analysis.)



刘青山 南京信息工程大学自动化学院院长, 教授. 2003 年获得中国科学院自动化研究所博士学位. 主要研究方向为图像理解, 模式识别, 机器学习. 本文通信作者.
E-mail: qsliu@nuist.edu.cn
(**LIU Qing-Shan** Dean and professor of the School of Automation, Nanjing University of Information Science and Technology. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2003. His research interest covers image understanding, pattern recognition, and machine learning. Corresponding author of this paper.)