基于可解释注意力部件模型的行人重识别方法

周 勇 [1, 2] 王瀚正 [1, 2] 赵佳琦 [1, 2] 陈 莹 [1, 2] 姚 睿 [1, 2] 陈思霖 [1, 2]

摘 要 大多数行人重识别 (Person re-identification, ReID) 方法仅将注意力机制作为提取显著特征的辅助手段, 缺少网络对行人图像关注程度的量化研究. 基于此, 提出一种可解释注意力部件模型 (Interpretable attention part model, IAPM). 该模型有 3 个优点: 1) 利用注意力掩码提取部件特征, 解决部件不对齐问题; 2) 为了根据部件的显著性程度生成可解释权重, 设计可解释权重生成模块 (Interpretable weight generation module, IWM); 3) 提出显著部件三元损失 (Salient part triplet loss, SPTL) 用于 IWM 的训练, 提高识别精度和可解释性. 在 3 个主流数据集上进行实验, 验证所提出的方法优于现有行人重识别方法. 最后通过一项人群主观测评比较 IWM 生成可解释权重的相对大小与人类直观判断得分, 证明本方法具有良好的可解释性.

关键词 行人重识别, 注意力机制, 可解释深度学习, 部件模型

引用格式 周勇, 王瀚正, 赵佳琦, 陈莹, 姚睿, 陈思霖. 基于可解释注意力部件模型的行人重识别方法. 自动化学报, 2023, **49**(10): 2159−2171

DOI 10.16383/j.aas.c200493

Interpretable Attention Part Model for Person Re-identification

ZHOU Yong[1, 2] WANG Han-Zheng[1, 2] ZHAO Jia-Qi[1, 2] CHEN Ying[1, 2] YAO Rui[1, 2] CHEN Si-Lin[1, 2]

Abstract Most person re-identification (ReID) methods only use the attention mechanism as an auxiliary method to extract salient features, and lack of quantitative research on the attention degree of person images on the network. Based on this, this paper proposes an interpretable attention part model (IAPM). The model has three advantages: 1) Using the attention mask to extract component features for solving the problem of component misalignment; 2) To generate interpretable weights based on the significance of the components, we devise the interpretable weight generation module (IWM); 3) Salient part triple loss (SPTL) for IWM is proposed to further improve recognition accuracy and interpretability. A series of experiments are carried out on three mainstream datasets, and demonstrate that our method is superior to the state-of-the-art methods. Finally, a crowd subjective test is used to compare the relative size of the interpretable weights generated by IWM and human intuitive judgment scores, which proves that the method has good interpretability.

Key words Person re-identification (ReID), attention mechanism, interpretable deep learning, part model

Citation Zhou Yong, Wang Han-Zheng, Zhao Jia-Qi, Chen Ying, Yao Rui, Chen Si-Lin. Interpretable attention part model for person re-identification. *Acta Automatica Sinica*, 2023, **49**(10): 2159−2171

行人重识别 (Person re-identification, ReID) 旨在通过非重叠视角域多视图下判断行人是否为同一目标, 属于图像检索的子问题[1−2]. 对于一个包含目标行人的查询图像和图像集, 行人重识别技术会根据与查询图像的相似度对来自图像集的图像排名, 进而找到同一目标, 减少人力、物力在图像序列中搜索的消耗. 行人重识别技术可以与行人检测、行人跟踪技术相结合, 在视频监控、安检、刑事侦查等方面有着广泛应用[3], 因此进行行人重识别研究具有较高的理论意义和实际价值. 但是, 人类可以解释事物的来龙去脉, 行人重识别任务用到的深度神经网络却不能做到. 深度学习所用到的架构很大程度上依靠大量的经验和技巧来设定, 通过梯度下降算法[4] 来优化模型参数, 这一学习过程犹如“黑盒子”[5]. 基于深度学习模型的行人重识别研究存在可解释性较弱的问题, 而且模型预测结果缺乏符合人类逻辑的解释.

近年来, 很多学者使用的注意力机制在图像显

收稿日期 2020-07-06 录用日期 2020-10-19
Manuscript received July 6, 2020; accepted October 19, 2020
国家自然科学基金 (61806206, U1610124, 61772530, 61773383), 江苏省自然科学基金 (BK20180639, BK20171192), 江苏省六大人才高峰计划 (2015-DZXX-010) 资助
Supported by National Natural Science Foundation of China (61806206, U1610124, 61772530, 61773383), Natural Science Foundation of Jiangsu Province (BK20180639, BK20171192), and the Six Talent Peaks Project in Jiangsu Province (2015-DZXX-010)
本文责任编委 黄庆明
Recommended by Associate Editor HUANG Qing-Ming
1. 中国矿业大学计算机科学与技术学院 徐州 221116 2. 矿山数字化教育部工程研究中心 徐州 221116
1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116 2. Engineering Research Center of Mine Digitization of Ministry of Education, Xuzhou 221116

著特征提取上展现出了强大的能力, 可以利用人类视觉机制对模型进行直观解释, 在一定程度上增加了行人重识别模型的可解释性. 其主要方法分为两个方面, 一方面为基于部件模型的注意力机制[6−8], 用来学习身体部件的判别性特征; 另一方面为前景注意力机制[9−12], 使用行人掩码以一种有监督的方式驱使注意力. 前者往往对输入图像进行分割[7], 或使用姿态评估作为辅助[13], 能够有效地提取部件的判别性特征, 但由于行人的形态动作不一, 会导致部件分割不对齐现象, 影响模型性能, 且对整体图像分割容易引入复杂背景噪声; 后者能够帮助低层网络关注于前景区域, 因此更容易学习到判别性的特征表示. 但由于输入图像的分辨率较低, 行人掩码的质量往往较差, 容易造成对底层网络的误导[12]. 更好的做法是将前景注意力和判别性特征学习融合到端到端的网络, 二者可以在训练过程中实现互补.

上述方法均利用注意力机制, 学习行人的显著性特征, 提高行人重识别模型性能. 但现有基于注意力机制的行人重识别方法存在两点不足: 首先, 注意力机制仅作为网络提取显著特征的辅助手段, 无法体现网络自身对区域是否显著的判断; 其次, 大多数方法只是通过可视化注意力掩码[13−14] 和热值图[15] 来证明其提出的注意力模块的有效性, 缺少行人图像对网络输出结果影响的量化研究.

本文基于上述两点不足, 提出了一种基于可解释注意力部件模型 (Interpretable attention part model, IAPM) 的行人重识别方法. 本方法受到文献 [16] 启发, 利用注意力机制实现行人部件特征的提取, 特别地, 可以根据部件特征的显著性来生成可解释权重, 以此作为行人重识别模型对于行人部件的显著性判断, 从而获取行人部件引起模型注意的程度, 提高深度学习模型的可解释性.

本文的主要贡献包括以下方面:

1) 提出一种基于可解释注意力部件模型的行人重识别方法, 该方法可以通过注意力机制实现灵活提取人体部件特征, 特别地, 可以依照部件的显著性程度生成可解释权重, 量化人体部件在深度学习模型训练过程中的作用, 从而提高行人重识别模型的可解释性.

2) 提出一种新的可解释权重生成模块 (Interpretable weight generation module, IWM), 设计新的显著部件三元损失 (Salient part triplet loss, SPTL) 端到端地自适应训练来提高模型表征能力及可解释性.

3) 在 Market-1501、CUHK03 及 DukeMTMC-ReID 数据集上进行实验验证, 分别达到了 95.2%、72.6%、88.0% 的 Rank-1 准确率, 高于基线论文及大多数现有方法. 本文还进行了一项人群主观测评, 将主观测评结果与生成的可解释权重对比, 证明本方法具有良好的可解释性.

本文结构安排如下: 第 1 节介绍可解释深度学习及行人重识别的相关工作; 第 2 节介绍本文提出的基于可解释注意力部件模型的行人重识别方法; 第 3 节给出实验设置与实验结果分析; 第 4 节总结本文工作并对未来工作进行展望.

1 相关工作

1.1 可解释深度学习

近年来, 深度学习高速发展, 但其模型内部的运行规律, 如隐含层卷积核的特定激活情况、模型做出决策的直接依据等仍属未知. 尽管如此, 人们依靠大量工程经验, 建立模型, 初始化参数, 并使用大量标注数据, 依然可以得到一个特定场景下表现优异的深度学习模型, 这也促使人们开始探索深度学习模型内部的运作机制. 许多研究人员将深度学习模型与人类认知相结合, 以找到二者的共通之处. 目前针对深度学习可解释领域的研究主要有以下 4 个方面:

1) 可视化卷积神经网络

研究人员通过计算图像所对应神经元的梯度、偏导数以及输出热值图、类激活映射等方法, 可以很好地将神经网络可视化, 将卷积核与人类感知的可视语义概念联系起来, 直接观察得到图像分类的主要依据区域, 对模型的输出进行解释. 文献 [17] 通过局部重新分配策略将预测 $f(x)$ 反向传播, 直到将相关得分 R_i 分配到每一个输入变量 (如像素). 在图像级别上, 通过这种方法可以得到图像分类的主要依据区域. 文献 [18] 针对使用全局均值池化的分类网络, 将最后分类得分对应的全连接层中的权重取出, 计算全局均值池化之前张量各通道的加权和, 与原图像进行对照, 即可寻找出分类结果的主要依据.

2) 网络结构与语义信息的对应

Szegedy 等[19] 发现深层次的神经网络中, 语义信息与深层网络结构的整体有关. 文献 [20] 进行了网络内卷积核与可视语义概念对应的研究, 使用双线性插值在每个卷积单元对应的激活映射进行上采样, 挑选出高于阈值的激活区域, 计算与语义概念注释之间的交并比, 由此得到卷积核与可视语义概念的对应.

3) 卷积神经网络的缺陷及优化

如果一个深度学习模型具有可解释性, 那么所有参数对于实验结果的影响应该是较清晰的, 这样就可以根据输出, 对算法及模型内部参数进行高效率的改良. 因此深度网络模型的可解释性对于模型的优化有着重要意义. 文献 [21] 中提出了一种视频字幕生成的可解释性方法, 该方法可以将神经元与视频的主题联系起来. 在神经网络输出字幕丢失了某些主题时, 可以直接找到与该主题相关联的神经元, 增加其对该主题的平均激活, 进而对网络微调, 保证输出不再丢失主题.

4) 可解释性模块的引入

与上述方法不同, 此方法并不是在预训练的网络中进行可解释的尝试, 而是在网络中加入可解释模块共同训练, 使网络的隐含层不再是一个“黑盒子”. 文献 [22] 为神经网络中每个卷积核增加了损失, 使得训练之后的卷积核对应特定的目标部件, 将卷积核的特征对应加入到“端到端”训练过程中, 可以不使用人类标记指导来完成可解释学习, 得到高层卷积核中对应的特定语义概念.

1.2 行人重识别

行人重识别作为一个图像检索的子问题, 旨在预测两幅行人图像是否属于同一行人. 随着深度学习的发展, 行人重识别问题的研究达到了前所未有的高度, 利用卷积神经网络 (Convolutional neural network, CNN) 可以实现行人特征的自动提取, 行人重识别模型性能得到有效提升.

基于深度学习的行人重识别方法, 可以按照学习方式分为两类: 基于表征学习的方法[23−26] 和基于度量学习的方法[1, 27−28]. 基于表征学习的行人重识别方法并没有把比较两个行人的相似度作为研究目标, 而是将行人重识别问题当作一个分类问题来看待, 将一幅行人图像输入到网络中提取特征, 将经过全局池化的特征向量送入全连接层, 最后连接 softmax 层, 由 softmax 激活函数得到每张图像的身份预测, 具有相同预测结果的两个行人即判定为同一行人. 文献 [23−24] 将每一个行人的身份当作分类问题的标签, 用来对 CNN 进行训练. 文献 [25] 引入行人属性标签计算属性损失, 和行人身份损失结合起来训练, 增强网络的泛化能力. 文献 [26] 提出在主干网络后增加验证子网络和分类子网络, 同时使用验证损失和分类损失对整个模型进行训练, 得到了较好的结果.

基于度量学习的行人重识别方法通过 CNN 将行人特征映射到特征空间中, 比较特征向量在特征空间中的距离 (例如欧氏距离或者余弦距离). 在训练过程中, 通过优化各种度量损失, 得到一个图像与特征向量的最佳映射关系, 使得在同一个特征空间中, 相同身份的行人特征向量有着尽可能小的距离, 不同身份的行人特征向量有着尽可能大的距离. 文献 [1] 使用了余弦相似度和二项式偏差来进行度量学习. 文献 [27] 采用一种孪生网络结构, 并使用对比损失来对网络模型进行优化. 文献 [28] 对三元损失进行了改进, 提出了批量难三元组损失 (Batch hard triplet loss), 使用距离最远的正例样本对和距离最近的负例样本对进行模型的优化.

鲁棒的特征表示对于解决行人重识别问题来说至关重要, 研究者们通常设计注意力模块提取显著性特征. 如前景掩码广泛用于引导网络注意行人身体的区域[11−12]. 文献 [29] 设计了空间约束网络 (Spatial transformer network, STN) 以提取行人局部特征. 文献 [13] 通过行人姿势信息生成的注意力掩码提取行人局部特征, 并能有效处理遮挡问题. 文献 [30] 提出了一个双线性的注意力网络, 使用双线性池化来提取逐对的局部信息. 文献 [31] 使用长短期记忆网络 (Long short term memory network, LSTM)[32] 构建了一个注意力模型, 用来提取图像的显著特征.

通过以上方法可以发现, 目前研究者们使用的注意力机制, 大多数作为提取图像显著特征的辅助手段, 无法体现网络自身对局部区域是否显著的判断. 此外, 虽然有些方法[13−14] 通过可视化注意力掩码和热值图对注意力模型进行直观解释, 但缺少行人图像对网络输出结果影响的量化研究, 存在可解释性较弱的问题. 本文基于以上两点, 提出可解释注意力部件模型 IAPM, 该模型将人体部件在深度学习模型训练过程中的作用量化, 以此作为网络自身对特征显著程度的判断, 提高行人重识别模型的可解释性.

2 基于可解释注意力部件模型的行人重识别方法

本文基于注意力掩码提取人体部件特征的模型 EANet[16], 针对其可解释性差的问题, 设计可解释注意力部件模型 IAPM. IAPM 的整体结构如图 1 所示. 该模型包括注意力部件对齐池化 (Part aligned pool, PAP) 模块和可解释权重生成模块. 为了增强部件对齐池化的规范性和严整性, 增加了一个局部分割约束 (Part segmentation, PS), 减少人体部件之间的重叠特征. 在本节中, 首先介绍本文基线模型 EANet 中的 PAP 模块和 PS 模块, 之后介绍本文提出的可解释性方法.

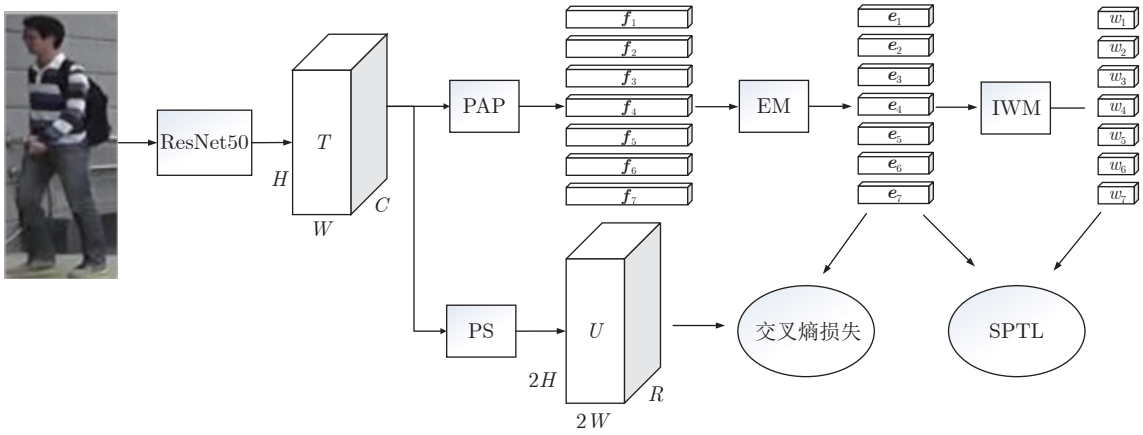


图 1 IAPM 整体结构

Fig. 1 Structure of IAPM

2.1 PAP 模块与 PS 模块

本文使用 ResNet50[33] 作为主干网络, 将尺寸为 384×128 像素的行人图像 x 输入到 ResNet50 中, 得到 $C \times H \times W$ 的三维张量 T, 其中 H 和 W 表示张量每个通道的高和宽, 分别为 24 和 8; C 表示张量的通道数, 为 2048.

PAP 模块主要实现人体特征的横向分割, Huang 等[16] 在 COCO (Common object in context) 数据集上预训练了一个关键点检测模型, 用来预测行人图像中行人身体的 17 个关键点, 从而定位出 9 个人体部件, 包括头、上躯干、下躯干、大腿、小腿、脚、上半身、下半身、全身. 在本文方法中, 人体部件个数 P 设置为 7, 从上至下依次为头、上躯干、下躯干、大腿、小腿、脚、全身七个部件, 如图 2 所示.

图 2 横向分割示意图

Fig. 2 Schematic diagram of horizontal split

根据这些部件在 ResNet50 输出张量中的对应位置, 生成部件分割注意力掩码 $M_i \in \mathbf{R}^{C\times H\times W}$, 其中 $i \in [1, P]$, M_i 表示第 i 个部件的注意力掩码. 部件对应位置元素设为 1, 其他位置设为 0. 张量 T 经过 PAP 模块, 得到横向分割的 P 个部分的特征向量为

$$\boldsymbol{f}_i = maxpool\left(M_i \otimes T\right) \tag{1}$$

其中, $\boldsymbol{f}_i \in \mathbf{R}^C$, $i \in [1, P]$, $maxpool$ 代表全局池化操作, $\otimes$ 表示逐元素相乘.

PAP 模块将张量 T 横向分割成 P 个部分, 经过全局池化得到 P 个部件的特征向量 $\boldsymbol{f}_1, \cdots, \boldsymbol{f}_P$. 将每一个特征向量输入到嵌入层 (Embedding layer, EM), 使每个部件特征向量长度由 2 048 降至 256. 得到的输出向量为

$$\boldsymbol{e}_i = g_i\left(\boldsymbol{f}_i\right) \tag{2}$$

其中, $\boldsymbol{e}_i \in \mathbf{R}^d$, d 表示 256, g_i 表示 EM 对第 i 个部件进行的全连接操作.

行人的身份损失 L^{ID} 采用交叉熵损失. 假设训练集包含 K 个行人身份, 给定一张标签为 y 的输入图像 x, 将图像 x 第 i 个部件的特征向量 $\boldsymbol{e}_i$ 输入到分类层进行一次全连接操作, 得到预测向量 $z_i = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_k] \in \mathbf{R}^K$. 经过 softmax 函数处理, 得到图像 x 中行人第 i 个部件属于第 k $(k \in 1, 2, 3, \cdots, K)$ 个行人身份的概率, 即

$$\bar{y}_k = \frac{\exp\left(\boldsymbol{z}_k\right)}{\sum\limits_{j=1}^{K}\exp\left(\boldsymbol{z}_j\right)} \tag{3}$$

第 i 个部件的身份损失为

$$L_i^{\mathrm{ID}} = -\sum_{k=1}^{K} y\log\bar{y}_k \tag{4}$$

各部件的身份损失之和即为该行人的身份损失 L^{ID}. 由于遮挡和摄像头视角的影响, 一些行人只有部分身体呈现在图像中, 因此引入一个可视度得分 v_i 来表示身体部件是否出现在图像中: $v_i = 1$ 表示身体部件 i 出现在图像中; $v_i = 0$ 表示身体部件 i 不可见. 在进行部件对齐池化时, 将不可见部件的特

征向量设置为零向量, 在计算身份损失时, 仅由可见区域产生损失. 行人身份损失函数定义为

$$L^{\mathrm{ID}}=\sum_{i=1}^{P}L_i^{\mathrm{ID}}v_i \tag{5}$$

其中, $v_i\in\{0,1\}$, 表示第 i 个部件的可视度, P 为人体部件总数.

在实验中发现, 通过 PAP 模块提取出的相邻部件之间的相似度较高, 即便模型提取到了多个具有判别性的特征, 这对部件对齐的效果仍然有影响. 为了降低不同部件之间的冗余度, 在 ResNet50 的 conv5 的特征图, 即张量 T 上增加 PS 模块来强化部件对齐池化效果.

PS 模块由一个步长为 2 的 3×3 的反卷积层及 1 个 1×1 的卷积层组成, 反卷积层用于上采样, 1×1 卷积层用于逐像素的分类预测. 分类类别包括 8 类, 即: 背景、头、躯干、前臂、后臂、大腿、小腿、脚. 尺寸为 $C\times H\times W$ 的张量 T 经过反卷积层之后, 得到尺寸为 $d\times2H\times2W$ 的中间张量. 将中间张量输入到 1×1 的卷积层中, 得到尺寸为 $R\times2H\times2W$ 的预测结果 U, 其中, R 表示类别总数, 设置为 8, 8 个通道代表 8 个类的分类结果. 需要注意的是, PAP 模块水平提取的人体部件特征, 最后用于行人相似度的计算. 而 PS 模块进行的部件分类预测, 仅为了增强 PAP 模块提取部件特征的规范性和严整性, 并未实际进行分割, 其输出张量 U 不作为行人相似度的判断依据.

训练 PS 模块使用的监督信号, 是使用 COCO 数据集预训练的部件分割模型在行人数据集上生成的伪标签. 部分伪标签如图 3 所示.

图 3　PS 模块使用的伪标签[16]

Fig. 3　Pseudo-labels used by PS[16]

张量 T 经过 PS 模块得到预测 U 之后, 计算其交叉熵损失作为部分分割损失 L^{PS}. 部分分割损失的计算式为

$$L^{\mathrm{PS}}=\frac{1}{R}\sum_{r=1}^{R}L_r^{\mathrm{PS}} \tag{6}$$

其中, R 表示部分的总数 (包括背景), 设置为 8, 与基线论文相同. L_r^{PS} 表示第 r 个部分所有像素点的交叉熵损失的均值. 取均值的原因在于避免某些部分面积过大导致其损失占比过多, 忽略头、脚这些面积小但仍含有判别性信息的部分.

2.2　IWM 模块

基线模型利用注意力机制灵活提取人体部件特征, 解决固定分割部件方法[8] 存在的不对齐问题. 但由于深度学习网络具有“黑盒子” 模型特点, 无法获取网络内部对每个部件显著程度的判断, 整个行人重识别模型的可解释性较差. 针对上述问题, 设计一种可以依照部件显著性程度来生成可解释权重的注意力权重生成模块, 结构如图 4 所示.

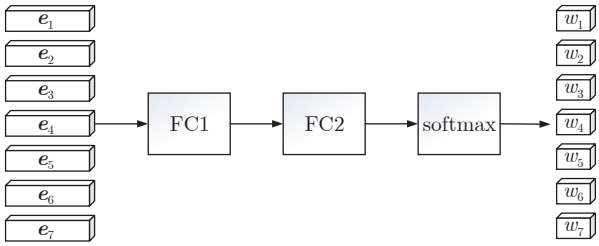


图 4　注意力权重生成模块结构

Fig. 4　Structure of IWM

IWM 由两个全连接层 FC1、FC2 以及一个 softmax 层组成. IWM 将 P 个人体部件堆叠之后的特征矩阵作为输入, 最终得到每个部件的可解释权重.

为了提升网络性能, 优化 IWM 的权重生成能力, 本文在批量难三元损失[31] 的基础上, 提出一种新的显著部件三元损失用于 IWM 的训练. SPTL 改变原有批量难三元损失中正负样本对距离的计算方式: 计算两幅图像相同部件之间的 L_2 距离, 与两部件经过 IWM 生成的权重相乘得到部件之间的权重距离, 如式 (7) 和式 (8) 所示.

$$d_{a,pos|i}=\max_{pos=1,\cdots,m}w_{a|i}w_{pos|i}||\boldsymbol{e}_{a|i}-\boldsymbol{e}_{pos|i}||_2 \tag{7}$$

$$d_{a,neg|i}=\min_{neg=1,\cdots,(m-1)n}w_{a|i}w_{neg|i}||\boldsymbol{e}_{a|i}-\boldsymbol{e}_{neg|i}||_2 \tag{8}$$

其中, $\boldsymbol{e}_{a|i}$, $\boldsymbol{e}_{pos|i}$, $\boldsymbol{e}_{neg|i}$ 分别表示锚点图像、正例图像以及负例图像第 i 个部件的特征向量; $w_{a|i}$, $w_{pos|i}$, $w_{neg|i}$ 分别表示锚点图像、正例图像以及负例图像第 i 个部件经过 IWM 生成的权重; $d_{a,pos|i}$ 和 $d_{a,neg|i}$ 分别表示锚点图像与正例图像、锚点图像与负例图像之间的权重距离. 将这个距离作为难负样

本挖掘依据. 对每一个部件进行损失的计算, 如式 (9) 所示.

$$L_i^{\mathrm{SPTL}} = \max\left(0,\, d_{a,\,pos|i} - d_{a,\,neg|i} + \alpha\right) \tag{9}$$

其中, α 为人为设定的参数, 如果正样本对距离与负样本对距离相差小于 α, 则会产生损失.

所有部件损失的和作为最终的显著部件三元损失 L^{SPTL}, 如式 (10) 所示.

$$L^{\mathrm{SPTL}} = \sum_{i=1}^{P} L_i^{\mathrm{SPTL}} \tag{10}$$

使用 SPTL 对 IWM 进行自适应训练, 若某部件对应的三元组内正负样本对距离易于改变, 即易于优化显著部件三元损失, IWM 将对该部件生成较大权重. 本文提出的可解释模型将易于优化 SPTL 的部件作为显著性部件, 通过训练赋予其可解释性, 从而使行人重识别模型对行人图像显著性的判断可见, 提高深度学习模型的可解释性.

此外, 在三元损失的计算中, 往往考虑的是正负样本之间的距离大小, 没有考虑优化每个类别内的距离. 例如对于正负样本对距离 0.3 和 0.5, 以及正负样本对距离 1.3 和 1.5, 损失均为 0.2, 但第 2 种情况下正样本对之间的距离更大, 所以对整个数据集来说无法保证正样本对之间的距离尽可能小. 因此, 本文使用中心损失[34] 来同时学习优化每个类别在特征空间中的中心位置以及每个特征到对应的类别中心位置的距离, 从而弥补上述三元损失的不足. 具体形式为

$$L^{\mathrm{C}} = \frac{1}{2}\sum_{j=1}^{B} ||\boldsymbol{f}_j - c_{y_j}||_2^2 \tag{11}$$

其中, y_j 表示第 j 幅图像的标签, c_{y_j} 表示标签 y_j 对应的中心, $\boldsymbol{f}_j$ 表示一个训练批次中第 j 幅行人图像的特征, B 为 Batchsize, 即一次迭代训练使用图像的数量.

基于以上损失函数, 可解释注意力部件模型的总损失函数可以表示为

$$L = L^{\mathrm{ID}} + L^{\mathrm{PS}} + \lambda L^{\mathrm{SPTL}} + \beta L^{\mathrm{C}} \tag{12}$$

其中, L^{ID} 代表身份损失, L^{SPTL} 代表显著部件三元损失, L^{PS} 代表部分分割损失, L^{C} 代表中心损失. L^{ID} 、L^{PS} 的系数及 β 均按照文献 [16] 设置为 1 及 0.0005, λ 根据实验结果设定为 1, 实验细节在第 3 节具体描述.

3 实验设置及实验结果

本节首先介绍实验设置和数据集及评价标准; 其次将本文提出的方法与本文的基线模型及现有的先进方法在性能上进行比较实验; 然后对本文提出的方法进行多组消融实验; 最后将网络输出的可解释权重与主观测评结果进行比较.

3.1 实验设置

本节实验使用的软硬件环境见表 1.

表 1 实验环境

Table 1 Experimental environment

软硬件环境	配置
实验平台	Pytorch
显卡	NVIDIA Tesla P100
内存	40 GB
显存	16 GB

本节实验的参数设置见表 2.

表 2 实验参数

Table 2 Experimental parameters

实验参数	参数数值
输入图像尺寸 (像素)	384 × 128
迭代次数	100
优化器	SGD
动量因子	0.9
权重衰减系数	5×10^{-4}
Batchsize	128
显著部件三元损失 α	1.2

网络中的 ResNet50 初始学习率为 0.0001, 在经过 10 次迭代后, 学习率由 0.0001 线性增加到 0.01, 并且在 50 以及 80 次迭代时, 降为原来的 1/10. 网络中的 EM 以及 IWM 初始学习率为 0.0002, 经过 10 次迭代后, 学习率由 0.0002 线性增加到 0.02, 并且在 50 以及 80 次迭代时, 降为原来的 1/10.

3.2 数据集及评价标准

Market-1501[35] 数据集中的图像包括 1 501 个行人, 总共 32 668 幅图像, 由 6 个摄像头采集获得. 751 个人的 12 936 幅图像用来进行训练, 平均每人有 17.2 幅训练图像; 750 个人的 19 732 幅图像用来进行测试, 平均每人有 26.3 幅测试图像.

DukeMTMC-reID[36] 提供了一个由 8 个摄像机拍摄得到的行人图像集, 包括 1 404 个不同身份的行人, 训练集由 1 404 中的 702 个人的 16 522 幅图像构成, 测试集由另外 702 个人的 17 661 幅图像构成.

CUHK03[37] 是在香港中文大学校园中采集的, 数据集由 1 467 个行人的 14 097 幅图像构成, 平均每人 9.6 幅训练图像.

本节实验中, 计算经过 EM 层之后得到的各部件特征向量之间的欧氏距离之和, 作为行人图像之间的相似度度量. 采用的评价标准为累积匹配特性曲线 (Cumulative match characteristic, CMC) 在第一匹配率的值 (记为 Rank-1) 和平均准确率 (Mean average precision, mAP).

3.3　对比实验

1) 与基线模型对比

将本文提出的方法与 EANet 在上述 3 个主流数据集上进行性能对比. 主要评价指标为 Rank-1 以及 mAP. 所有实验结果均在单查询样本及没有进行重新排序的情况得到. 实验结果如表 3 所示 (表 3 中数据为 Rank-1 值, 括号内数据为 mAP 值).

表 3　与 EANet 的性能对比 (%)
Table 3　Performance comparison with EANet (%)

方法	数据集		
	Market-1501	DukeMTMC-reID	CUHK03
PAP-6P	94.3 (84.3)	85.6 (72.4)	68.1 (62.4)
PAP	94.5 (84.9)	86.1 (73.3)	72.0 (66.2)
PAP-S-PS	94.6 (85.6)	87.5 (74.6)	72.5 (66.8)
IAPM-6P (本文)	95.0 (85.3)	86.9 (74.3)	72.5 (65.2)
IAPM-9P (本文)	95.1 (86.0)	87.9 (75.6)	**72.6 (67.4)**
IAPM (本文)	**95.2 (86.3)**	**88.0 (75.7)**	**72.6** (67.2)

PAP-6P、PAP 分别指的是 EANet 中使用 6 个及 9 个人体部件, 且只使用 L^{ID} 训练的单域模型; PAP-S-PS 指的是 EANet 使用 9 个部件且使用 L^{ID}、IAPM-6P 和 L^{PS} 训练的单域模型; IAPM、IAPM-6P 和 IAPM-9P 指的是本文使用 7 个、6 个和 9 个部件, 且使用总损失函数 L 训练的模型.

IAPM 在 3 个主流数据集上的 Rank-1 较 EANet 中单域表现最好的模型 (PAP-S-PS) 分别提升了 0.6%, 0.5%, 0.1%; 在 mAP 上分别提升了 0.6%, 1.1%, 0.4%. 为了与 PAP-S-PS 进行公平对比, 使用与其相同的 9 个部件进行实验. 在 3 个主流数据集上, 使用 9 个部件的模型 (IAPM-9P) 得到的结果与 PAP-S-PS 相比, Rank-1 分别提升了 0.5%, 0.4%, 0.1%; mAP 分别提升了 0.4%, 1.0%, 0.6%. 为了与 PAP-6P 进行公平对比, 使用与其相同的 6 个部件进行实验. 在 3 个主流数据集上, 使用 6 个部件的模型 (IAPM-6P) 得到的结果与 PAP-6P 相比, Rank-1 分别提升了 0.7%, 1.3%, 4.4%; mAP 分别提升了 1.0%, 1.9%, 2.8%.

2) 与其他方法对比

为了验证本文提出的可解释注意力部件模型的性能, 在主流数据集上与近年来提出的行人重识别方法进行对比, 主要评价指标为 Rank-1 以及 mAP. 所有实验结果均在单查询样本及没有进行重新排序的情况得到. 实验结果如表 4 所示 (表 4 中数据为 Rank-1 值, 括号内数据为 mAP 值).

表 4　与其他方法的性能对比 (%)
Table 4　Performance comparison with other methods (%)

方法	数据集		
	Market-1501	DukeMTMC-reID	CUHK03
Verif-Identify[38]	79.5 (59.9)	68.9 (49.3)	—
MSCAN[29]	80.8 (57.5)	—	—
MGCAM[12]	83.8 (74.3)	—	50.1 (50.2)
Part-Aligned[39]	91.7 (79.6)	84.4 (69.3)	—
SPReID[40]	92.5 (81.3)	84.4 (71.0)	—
AlignedReID[41]	91.8 (79.3)	—	—
Deep-Person[42]	92.3 (79.6)	80.9 (64.8)	—
PCB[7]	85.3 (68.5)	73.2 (52.8)	43.8 (38.9)
PCB + RPP[7]	93.8 (81.6)	83.3 (69.2)	63.7 (57.5)
HA-CNN[43]	91.2 (75.7)	80.5 (63.8)	44.4 (41.0)
Mancs[44]	93.1 (82.3)	84.9 (71.8)	69.0 (63.9)
P^2-Net [45]	95.1 (85.6)	86.5 (73.1)	**74.9** (**68.9**)
M^3 + ResNet50[46]	**95.4** (82.6)	84.7 (68.5)	66.9 (60.7)
IAPM (本文)	95.2 (**86.3**)	**88.0** (**75.7**)	72.6 (67.2)

注: "—" 表示文献中没有提供相应数据.

本文提出的方法在 Market-1501 数据集中的 Rank-1 达到 95.2%, mAP 达到 86.3%; 在 DukeMTMC-reID 数据集中的 Rank-1 达到 88.0%, mAP 达到 75.7%; 在 CUHK03 数据集中的 Rank-1 达到 72.6%, mAP 达到 67.2%. 可以看出, 在 Rank-1 及 mAP 两项主要评价指标上, 本文方法均高于近年来提出的大多数行人重识别方法.

3.4　消融实验

为了验证本文提出的可解释注意力部件模型各组成部分的有效性, 本文在 Market-1501 数据集上设计了多组消融实验, 包括验证 IWM 与中心损失函数的有效性, 分析部件个数对模型性能的影响, 以及分析 SPTL 中 α 及 λ 对实验结果的影响.

1) IWM 与中心损失函数的有效性

由第 3.3 节实验结果可以看到, 本文模型在行人重识别精度上可以达到较好的效果. 为进一步验证可解释权重生成模块的有效性, 从 IAPM 中移除

该模块作为原始模型进行实验. 仅使用身份损失函数对原始模型进行训练. 之后在此基础上依次增加IWM、SPTL 和中心损失函数. 实验结果如表 5 所示.

表 5 消融实验 1

Table 5 Ablation experiment 1

模型	Rank-1 (%)	mAP (%)
原始模型	92.4	80.5
原始模型 + IWM + SPTL	95.0	86.1
原始模型 + IWM + SPTL + 中心损失	**95.2**	**86.3**

注: 加粗字体表示各列最优结果.

由表 5 可以看到, 使用基线模型进行实验, Rank-1 和 mAP 分别为 92.4% 和 80.5%; 增加可解释权重生成模块之后, Rank-1 和 mAP 分别增加到了 95.0% 和 86.1%; 在此基础上增加中心损失后, Rank-1 和 mAP 分别增加到了 95.2% 和 86.3%. 以上实验结果说明, 可解释权重模块及中心损失对模型性能具有提升效果.

2) 人体部件个数对模型性能的影响

为了探究人体部件的个数对模型性能的影响, 在 Market-1501 数据集上, 使用不同的部件个数进行实验, 实验结果如表 6 所示.

表 6 消融实验 2

Table 6 Ablation experiment 2

人体部件个数	Rank-1 (%)	mAP (%)
6	95.0	85.3
7	**95.2**	**86.3**
9	95.1	86.0

注: 加粗字体表示各列最优结果.

人体部件个数在本次实验中分别设置为 6, 7, 9, 其中 6 个身体部件包括头、上躯干、下躯干、大腿、小腿、脚; 7 个身体部件包括头、上躯干、下躯干、大腿、小腿、脚、全身; 9 个身体部件包括头、上躯干、下躯干、大腿、小腿、脚、上半身、下半身、全身. 使用 6 个部件进行实验时, Rank-1 和 mAP 分别为 95.0% 及 85.3%; 使用 7 个部件进行实验时, Rank-1 和 mAP 分别为 95.2% 及 86.3%; 使用 9 个部件进行实验时, Rank-1 和 mAP 分别为 95.1% 及 86.0%. 使用 7 个和 9 个部件得到的实验结果, 高于使用 6 个部件得到的实验结果, 说明将全局或较大尺度特征作为局部特征的补充, 对网络模型性能的提升有一定的帮助. 使用 7 个部件得到的实验结果高于使用 9 个部件得到的实验结果, 说明使用全局特征作为局部特征的补充对本方法来说已足够, 如果增加较大尺度的特征 (上半身或下半身特征), 会造成部件特征的重叠, 无法使网络模型对相互独立的人体部件做出显著性判断.

3) 参数 α 对 SPTL 的影响

三元损失中的 α 对模型的性能同样起到非常重要的作用, 因此本节使用 4 个不同 α 的显著部件三元损失, 对 7 个人体部件的 IAPM 在 Market-1501 上进行实验, α 分别选为 0.1, 0.5, 0.8, 1.0, 1.2, 1.5, 2.0, 5.0, 10.0, 实验结果如表 7 所示.

表 7 消融实验 3

Table 7 Ablation experiment 3

α	Rank-1 (%)	mAP (%)
0.1	94.4	85.2
0.5	94.5	85.3
0.8	94.8	85.7
1.0	94.7	85.6
1.2	**95.2**	**86.3**
1.5	94.6	85.6
2.0	94.7	85.3
5.0	93.5	83.5
10.0	93.3	81.0

注: 加粗字体表示各列最优结果.

可以看出, α 选取 1.2 时, 得到最高的 Rank-1 和 mAP. 当 α 选择较小时 ($\alpha = 0.8$), 会导致正负样本对的距离无法有效拉大, 当 α 选择较大时 ($\alpha = 1.5$), 三元组中正负样本对之间的距离被过度拉大, 会导致不同三元组样本之间的距离难以控制. 容易造成三元组内的正负样本对之间距离相差很大, 而三元组之间的样本的距离很近的结果, 这同样会导致网络模型性能下降.

为了体现每个 α 对正负样本之间距离的优化效果, 选取 α 的 4 个取值, 绘制正负样本对距离的折线图, 如图 5 和图 6 所示.

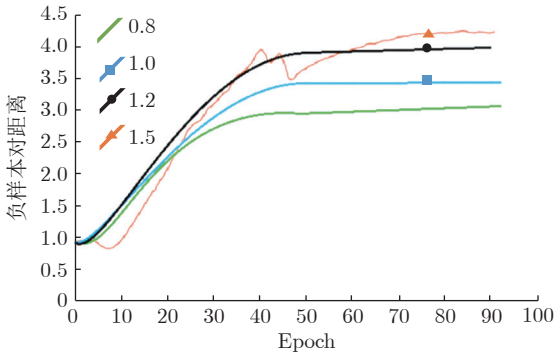


图 5 负样本对距离变化图

Fig. 5 Negative sample pair distance graph

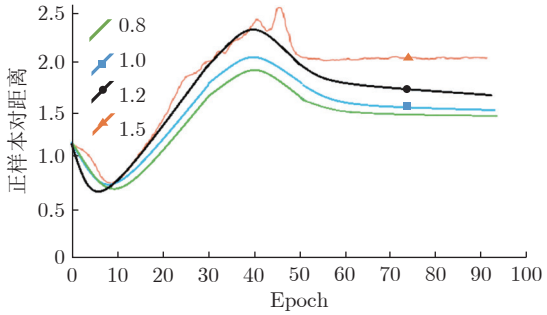


图 6　正样本对距离变化图

Fig. 6　Positive sample pair distance graph

由图 5 可以看出, 当 α 选取为 1.5 时, 负样本对之间距离的优化过程有较多起伏, 说明在 α 选取较大时, 模型需要尽可能将正负样本之间的距离进一步拉大, 这就需要在特征空间中进行较多尝试, 最终才能达到较理想的状态; 同时也可以看到, 随着 α 的不断增大, 锚定图片与负样本图片之间的距离不断拉大, 说明 SPTL 有效地进行了特征空间中的特征向量之间距离的优化. 我们还可以发现, 50 次迭代之后, 负样本对之间的距离基本上不会有较大变化, 所以选择在 50 次迭代后进行第 1 次学习率的衰减, 继续训练至 80 次迭代后进行第 2 次学习率的衰减, 然后进行最后的 20 次迭代.

由图 6 可以看出, 当 α 选择为 1.5 时, 正样本对距离的优化效果较差, 当选择其他三种 α 时, 可以使正样本对之间的距离有效缩小.

除以上实验外, 本节还将每个 α 对应的 SPTL 损失进行对比, 对比曲线图如图 7 所示.

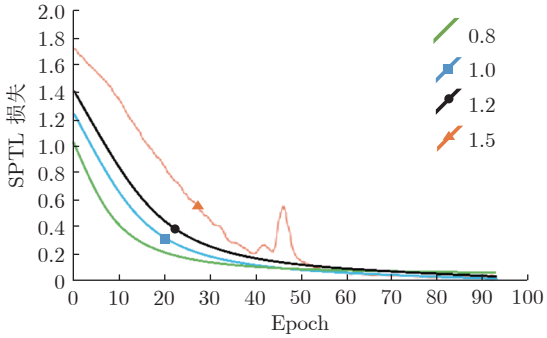


图 7　SPTL 损失曲线图

Fig. 7　SPTL loss curve graph

由图 7 可以看出, 当 α 为 1.5 时, SPTL 损失的收敛过程相对不稳定, 模型需要不断调整映射关系来满足正负样本对之间的距离要求. 当 α 选取较小时, SPTL 损失可以较好地收敛.

4) λ 对模型性能的影响

λ 旨在平衡 L^{SPTL} 与其他损失函数的重要性. 为了探究 λ 对模型性能的影响, 在 Market-1501 数据集上, 使用不同的 λ 进行实验, 实验结果如表 8 所示.

表 8　消融实验 4

Table 8　Ablation experiment 4

λ	Rank-1 (%)	mAP (%)
0.2	94.4	85.4
0.4	94.8	85.4
0.6	94.4	85.1
0.8	94.8	85.7
1.0	**95.2**	**86.3**

注: 加粗字体表示各列最优结果.

λ 在本次实验中分别设置为 0.2, 0.4, 0.6, 0.8, 1.0. λ 设置较小时, 会减弱 L^{SPTL} 的影响, 降低模型性能. 当选取为 1.0 时, Rank-1 和 mAP 分别为 95.2% 及 86.3%, 网络模型的性能可以达到最优.

3.5　可解释效果展示

除了在主流数据集上的识别准确率的提高外, 本文另一贡献是通过 IWM 生成的权重来反映部件的显著程度, 从而提高模型的可解释性. 通过以下可解释生成效果的展示以及与人群主观测评结果的对比, 证明提出的方法是具有可解释性的.

1) IWM 权重生成效果展示

从 Market-1501 和 DukeMTMC-reID 两个数据集中选取 5 幅图像, 利用本文提出的可解释模型得到的权重结果展示如图 8 所示. 图 8(a) 和图 8(b) 选自 Market-1501 数据集, 图 8(c)、图 8(d) 和图 8(e)

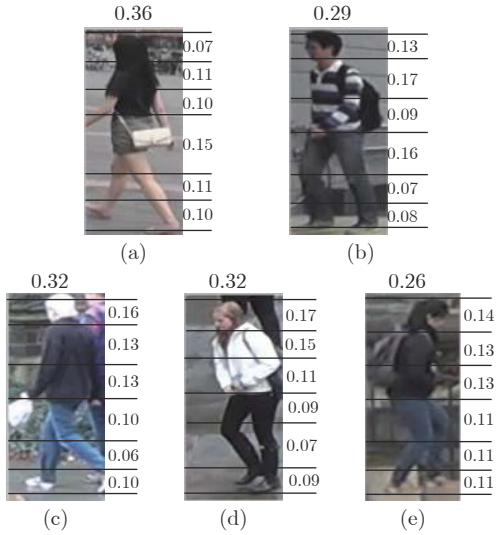


图 8　可解释权重展示

Fig. 8　The display of interpretable weights

选自 DukeMTMC-reID 数据集.

图 8 中右侧及上侧数值为显著性模型生成的 7 个部件的可解释权重. 其中图像右侧数值从上至下依次表示头、上躯干、下躯干、大腿、小腿、脚 6 个部件, 图像上端数值代表的是全局特征 (整幅图像得到的特征) 表示的第 7 个部件的可解释权重. 数值越大表示在训练过程中, 深度学习模型认为这一部件的判别力越强, 通过这一部件可以更有效地将不同身份的行人区分开来.

在使用测试集所有图像生成的可解释权重中, 每幅图像的第 7 个部件 (全局特征) 权重大于任意一个局部人体部件的可解释权重, 说明网络认为关注整体的全局特征与单个关注细节的人体部件特征相比, 判别性更强. 而第 7 个部件权重小于其他 6 个部件权重之和, 一方面说明局部的身体部件同样具有判别性较强的特征, 使用部件特征处理行人重识别任务仍可以获得较好的效果[34]; 另一方面说明全局特征可以作为局部特征的有效补充, 二者可以组成更加鲁棒的特征表示, 进一步提高行人重识别精度.

图 8(a) 中, 短裤对应的第 4 个部件的可解释权重较除整体外的其他 5 个部件高, 这与人类直观的反应相一致; 图 8(b) 中, 行人条纹上衣对应的可解释权重相对较高, 这也是与人类的直观反应相一致. 值得注意的是, Market-1501 这个数据集是 2015 年夏天在清华大学校园内采集的, 男生和女生身着短裤的居多, 而短裤往往颜色鲜明, 所以经过 Market-1501 数据集训练的部件可解释权重模型, 对于大腿这个部件尤为敏感, 这也是为什么在图 8(b) 中, 大腿部件同样会出现较高权重的原因.

在美国杜克大学冬天采集的数据集中, 因为冬天下身服装多为深色, 判别性不强, 所以并没有出现像 Market-1501 数据集中那样对于大腿部件的较高响应. 对于图 8(c), 网络将注意力集中在白色帽子对应的第 1 个部件上, 注意力权重较高; 对于图 8(d), 网络将注意力集中在白色的羽绒服, 对应着第 2 个和第 3 个部件; 图 8(e) 由于该行人的服装整体颜色较暗, 并没有特征显著的区域, 因此除全身以外的 6 个部件的特征所占权重几乎相同.

2) 人群主观测评结果

为了体现本文可解释模型生成权重的相对大小与人类直观判断的一致性, 本部分进行了一项问卷调查, 作为主观评测依据. 测评样本采用与前面实验相同的 5 幅图像, 邀请 50 位在校大学生进行问卷调查, 对 5 幅图像中的 6 个行人部件 (头、上躯干、下躯干、大腿、小腿、脚) 进行选择打分. 打分等级分别为: 很明显、较明显、一般、较不明显、不明显, 分别对应 5 分、4 分、3 分、2 分、1 分, 用来表示测试者对行人部件显著性的判断. 如果测试者认为头部更能引起测试者的注意, 那么他会在头部对应的选项中选择“很明显”, 对应的显著得分为 5 分.

将每幅行人图像同一部件的显著投票得分 (5 个选项的得分之和) 累加并除以投票总人数来计算该部件平均得分, 并用该部件平均得分除以总分 ($5 \times 6 = 30$), 从而得到人类主观显著性判断相对得分 (以下简述为相对得分), 表示该部件相对于该图像其他部件的显著程度, 得分较高的部件表示受到了测试者较多的注意, 即对应着显著性较高的人体部分. 本文之所以使用投票平均得分除以总分来计算相对得分, 而不是使用投票平均得分除以 6 个部件的投票总得分, 是因为前者对所有图像都除以固定的总分 (30 分), 不仅可以体现出某部件相对于同一行人其他部件的显著性 (进行同一行人部件之间相对得分的比较), 还可以直接通过相对得分, 比较不同行人部件之间的显著性. 主观测评阶段的相对得分展示如图 9 所示 (行人图像左侧为投票平均得分, 右侧为相对得分).

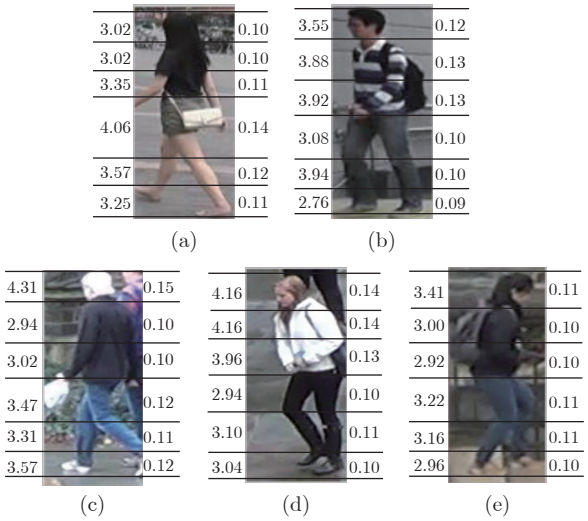


图 9 主观测评结果

Fig. 9 The display of subjective evaluation results

图 9(a) 中, 测试者大多数认为第 4 个部件 (大腿) 容易引起注意, 因此得到了最高的投票得分, 平均分为 4.06, 相对得分为 0.14, 在所有部件中得分最高. 图 9(b) 中, 该行人的条纹上衣吸引了最多的测试者的注意. 其条纹上衣对应的第 2 个和第 3 个部件的平均得分分别为 3.88 和 3.92, 相对得分都为最高的 0.13. 后 3 幅图像选择于 DukeMTMC-reID 数据集, 图 9(c) 中, 测试者认为行人白色的帽子最具有判别性, 在参与测试的 50 个测试者中, 有 28 个测试者对于头部这个部件选择了“很明显”,

有 15 个测试者选择了“较明显”, 平均得分为 4.31, 相对得分为最高的 0.15. 图 9(d) 中, 测试者认为行人头发的颜色以及白色的羽绒服最具有判别性, 第 1 个部件对应着头部, 有 22 位测试者选择“很明显”, 有 18 位测试者选择“较明显”, 平均得分为 4.16, 相对得分为 0.14; 第 2 个和第 3 个部件对应的是白色的羽绒服, 分别有 23 位及 18 位测试者选择“很明显”, 平均得分为 4.16 及 3.96, 相对得分分别为 0.14 及 0.13. 图 9(e) 中, 由于该行人的服装整体颜色较暗, 并没有特征显著的区域, 测试者的结果也显示, 大多数测试者对于每个部件选择“一般”或者“较不明显”, 部件整体的平均得分相比于其他行人较低.

3) 可解释权重与主观测评结果对比

由于全局部件权重比任何一个局部部件的可解释权重大的特殊性, 以及在下文人群主观测评中额外加入完整图像对人的主观判断造成的影响 (完整图像与局部部件的显著性不便于直观比较), 所以下文进行的可解释权重生成和主观测评结果的对照仅考虑前 6 个部件, 这样可以通过权重与测评结果部件之间的相对大小, 得出显著性模型可解释权重与人群主观评价的一致性. 比较结果如图 10 所示, 左侧数值为可解释注意力部件模型生成的可解释权重, 右侧数值为主观测评得到的相对得分.

可以看到, 通过本文显著性模型生成的可解释权重与人群主观评测结果基本一致. 图 10(a) 中模型与测试者的注意力均集中在腰部至大腿之间, 也就是第 4 个部件; 图 10(b) 中模型与测试者的注意力均集中在上衣, 对应着第 2 个和第 3 个部件, 唯一不同的是, 因为清华大学数据集中行人大腿部分裸露较多及短裤颜色鲜艳等自身数据集的特点, 会使模型对腰部至大腿这一部件有较高的响应. 图 10(c) 中模型与测试者的注意力均集中在白色的帽子, 模型输出的第 1 个部件的权重最高, 与人群主观测试结果一致. 图 10(d) 中模型与测试者的注意力均集中在上半身, 对应着该行人的金色的头发以及白色的羽绒服. 图 10(e) 中的行人因为衣服整体颜色较暗, 无明显的高判别性的特征, 因此人群主观测评结果显示, 人们认为各部件之间显著程度相似且显著得分较低, 同时网络模型输出的可解释权重之间相差无几, 表示模型认为行人中没有具有高判别性的部件, 与人群主观测评结果基本一致. 由此证明本文提出的部件显著性模型输出的可解释权重与人类对于显著性的认知基本相同, 赋予了深度学习网络在训练过程中的可解释性, 帮助我们更好地理解网络模型对于行人图像的认知和判断.

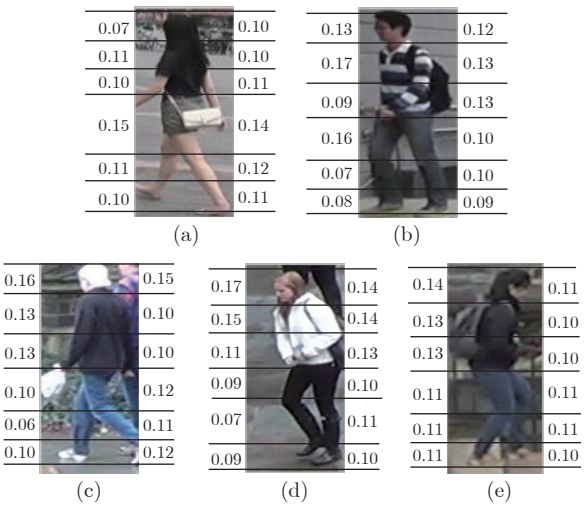


图 10　可解释权重与主观测评结果对比

Fig. 10　Comparison of interpretable weights and subjective evaluation results

4　结束语

本文详细介绍了一种基于可解释注意力部件模型的行人重识别方法, 该方法可以根据部件特征的显著性程度生成可解释权重, 获得行人重识别模型对行人图像显著性的判断, 提高深度学习模型的可解释性. 实验结果验证了本文方法的有效性. 在未来的工作中尝试使用孪生网络来获取属于同一行人身份的特征区域依据, 进一步提高行人重识别模型的可解释性.

References

1　Yi D, Lei Z, Liao S C, Li S Z. Deep metric learning for person re-identification. In: Proceedings of the 22nd IEEE International Conference on Pattern Recognition. Stockholm, Sweden: IEEE, 2014. 34–39

2　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 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 2197–2206

3　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)

4　Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**(6088): 533–536

5　Wu Fei, Liao Bin-Bing, Han Ya-Hong. Interpretability for deep learning. *Aero Weaponry*, 2019, **26**(1): 43–50
(吴飞, 廖彬兵, 韩亚洪. 深度学习的可解释性. 航空兵器, 2019, **26**(1): 43–50)

6　Chen W H, Chen X T, Zhang J G, Huang K Q. A multi-task deep network for person re-identification. In: Proceedings of the 31st Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2017. 3988–3994

7　Sun Y G, 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 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 480−496

8 Zhou S P, Wang J J, Wang J Y, Gong Y H, Zheng N N. Point to set similarity based deep feature learning for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 5028−5037

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

10 Zhao L M, Li X, Zhuang Y T, Wang J D. Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 3239−3248

11 Zhou S P, Wang J J, Meng D Y, Liang Y D, Gong Y H, Zheng N N. Discriminative feature learning with foreground attention for person re-identification. *IEEE Transactions on Image Processing*, 2019, **28**(9): 4671−4684

12 Song C F, Huang Y, Ouyang W L, Wang L. Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1179−1188

13 Xu J, Zhao R, Zhu F, Wang H M, Ouyang W L. Attention-aware compositional network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 2119−2128

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

15 Zhou S P, Wang F, Huang Z Y, Wang J J. Discriminative feature learning with consistent attention regularization for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 8039−8048

16 Huang H J, Yang W J, Chen X T, Zhao X, Huang K Q, Lin J B, et al. EANet: Enhancing alignment for cross-domain person re-identification [Online], available: http://arxiv.org/abs/1812.11369, October 21, 2020

17 Bach S, Binder A, Montavon G, Klauschen F, Muller K, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 2015, **10**(7): Article No. e0130140

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

19 Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J, et al. Intriguing properties of neural networks [Online], available: http://arxiv.org/abs/1312.6199, October 21, 2020

20 Bau D, Zhou B, Khosla A, Oliva A, Torralba A. Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 3319−3327

21 Dong Y P, Su H, Zhu J, Zhang B. Improving interpretability of deep neural networks with semantic information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 975−983

22 Zhang Q S, Wu Y N, Zhu S C. Interpretable convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 8827−8836

23 Zheng L, Yang Y, Hauptmann A G. Person re-identification: Past, present and future [Online], available: http://arxiv.org/abs/1610.02984, October 21, 2020

24 Zheng L, Zhang H H, Sun S Y, Chandraker M, Yang Y, Tian Qi. Person re-identification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 3346−3355

25 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

26 Geng M Y, Wang Y W, Xiang T, Tian Y H. Deep transfer learning for person re-identification [Online], available: http://arxiv.org/abs/1611.05244, October 21, 2020

27 Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification. In: Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 791−808

28 Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification [Online], available: http://arxiv.org/abs/1703.07737, October 21, 2020

29 Li D W, Chen X T, Zhang Z, Huang K Q. Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 7398−7407

30 Fang P F, Zhou J M, Roy S K, Petersson L, Harandi M. Bilinear attention networks for person retrieval. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 8029−8038

31 Liu H, Feng J S, Qi M B, Jiang J G, Yan S C. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017, **26**(7): 3492−3506

32 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, **9**(8): 1735−1780

33 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 770−778

34 Wen Y D, Zhang K P, Li Z F, Qiao Y. A discriminative feature learning approach for deep face recognition. In: Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 499−515

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

36 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 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 17−35

37 Li W, Zhao R, Xiao T, Wang X G. Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014. 152−159

38 Zheng Z D, Zheng L, Yang Y. A discriminatively learned CNN embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2018, **14**(1): Article No. 13

39 Suh Y, Wang J, Tang S, Mei T, Lee K M. Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018. 418−437

40 Kalayeh M M, Basaran E, Gökmen M, Kamasak M E, Shah M. Human semantic parsing for person re-identification. In: Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1062–1071

41 Zhang X, Luo H, Fan X, Xiang W L, Sun Y X, Xiao Q Q, et al. Alignedreid: Surpassing human-level performance in person re-identification [Online], available: http://arxiv.org/abs/1711.08184, October 21, 2020

42 Bai X, Yang M K, Huang T T, Dou Z Y, Yu R, Xu Y C. Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognition*, 2020, **98**: Article No. 107036

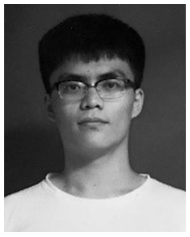
43 Li W, Zhu X T, Gong S G. Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 2285–2294

44 Wang C, Zhang Q, Huang C, Liu W Y, Wang X G. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018. 384–400

45 Guo J Y, Yuan Y H, Huang L, Zhang C, Yao J G, Han K. Beyond human parts: Dual part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 3641–3650

46 Zhou J H, Su B, Wu Y. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event: IEEE, 2020. 2909–2918

周 勇 中国矿业大学计算机科学与技术学院教授. 主要研究方向为数据挖掘, 机器学习和人工智能.
E-mail: yzhou@cumt.edu.cn
(**ZHOU Yong** Professor at the School of Computer Science and Technology, China University of Mining and Technology. His research interest covers data mining, machine learning, and artificial intelligence.)



王瀚正 中国矿业大学计算机科学与技术学院硕士研究生. 主要研究方向为计算机视觉, 图像处理, 行人重识别. E-mail: hzwang@cumt.edu.cn
(**WANG Han-Zheng** Master student at the School of Computer Science and Technology, China University of Mining and Technology. His research interest covers computer vision, image processing, and person re-identification.)

赵佳琦 中国矿业大学计算机科学与技术学院副教授. 主要研究方向为多目标优化, 深度学习, 图像处理. 本文通信作者.
E-mail: jiaqizhao88@126.com
(**ZHAO Jia-Qi** Associate professor at the School of Computer Science and Technology, China University of Mining and Technology. His research interest covers multiobjective optimization, deep learning, and image processing. Corresponding author of this paper.)

陈 莹 中国矿业大学计算机科学与技术学院博士研究生. 主要研究方向为计算机视觉, 图像处理, 行人重识别. E-mail: cheny@cumt.edu.cn
(**CHEN Ying** Ph.D. candidate at the School of Computer Science and Technology, China University of Mining and Technology. Her research interest covers computer vision, image processing, and person re-identification.)

姚 睿 中国矿业大学计算机科学与技术学院副教授. 主要研究方向为计算机视觉, 机器学习.
E-mail: ruiyao@cumt.edu.cn
(**YAO Rui** Associate professor at the School of Computer Science and Technology, China University of Mining and Technology. His research interest covers computer vision and machine learning.)

陈思霖 中国矿业大学计算机科学与技术学院硕士研究生. 主要研究方向为计算机视觉, 图像处理, 目标检测.
E-mail: silin.chen@cumt.edu.cn
(**CHEN Si-Lin** Master student at the School of Computer Science and Technology, China University of Mining and Technology. His research interest covers computer vision, image processing, and objective detection.)