软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2024,35(10):4766−4780 [doi: 10.13328/j.cnki.jos.006993]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 模态间关系促进的行人检索方法*

李 博[1], 张飞飞[1], 徐常胜[2]

[1](天津理工大学 计算机科学与工程学院, 天津 300384)

[2](多模态人工智能系统全国重点实验室 (中国科学院 自动化研究所), 北京 100190)

通信作者: 徐常胜, E-mail: csxu@nlpr.ia.ac.cn

**摘 要**: 基于文本描述的行人检索是一个新兴的跨模态检索子任务, 由传统行人重识别任务衍生而来, 对公共安全以及人员追踪具有重要意义. 相比于单模态图像检索的行人重识别任务, 基于文本描述的行人检索解决了实际应用中缺少查询图像的问题, 其主要挑战在于该任务结合了视觉内容和文本描述两种不同模态的数据, 要求模型同时具有图像理解能力和文本语义学习能力. 为了缩小行人图像和文本描述的模态间语义鸿沟, 传统的基于文本描述的行人检索方法多是对提取的图像和文本特征进行机械地分割, 只关注于跨模态信息的语义对齐, 忽略了图像和文本模态内部的潜在联系, 导致模态间细粒度匹配的不准确. 为了解决上述问题, 提出模态间关系促进的行人检索方法, 首先利用注意力机制分别构建模态内自注意力矩阵和跨模态注意力矩阵, 并将注意力矩阵看作不同特征序列间的响应值分布. 然后, 分别使用两种不同的矩阵构建方法重构模态内自注意力矩阵和跨模态注意力矩阵. 其中自注意力矩阵的重构利用模态内逐元素重构的方式可以很好地挖掘模态内部的潜在联系, 而跨模态注意力矩阵的重构用模态间整体重构矩阵的方法, 以跨模态信息为桥梁, 可充分挖掘模态间的潜在信息, 缩小语义鸿沟. 最后, 用基于任务的跨模态投影匹配损失和 *KL* 散度损失联合约束模型优化, 达到模态间信息相互促进的效果. 在基于文本描述的行人检索公开数据库 CUHK-PEDES 上进行了定量以及检索结果的可视化, 均表明所提方法可取得目前最优的效果.

**关键词**: 行人检索; 跨模态任务; 文本语义学习; 关系对齐; 注意力机制

**中图法分类号**: TP18

中文引用格式: 李博, 张飞飞, 徐常胜. 模态间关系促进的行人检索方法. 软件学报, 2024, 35(10): 4766–4780. http://www.jos.org.cn/1000-9825/6993.htm

英文引用格式: Li B, Zhang FF, Xu CS. Cross-modal Person Retrieval Method Based on Relation Alignment. Ruan Jian Xue Bao/Journal of Software, 2024, 35(10): 4766–4780 (in Chinese). http://www.jos.org.cn/1000-9825/6993.htm

## Cross-modal Person Retrieval Method Based on Relation Alignment

LI Bo[1], ZHANG Fei-Fei[1], XU Chang-Sheng[2]

[1](School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China)

[2](State Key Laboratory of Multimodal Artificial Intelligence Systems (Institute of Automation, Chinese Academy of Sciences), Beijing 100190, China)

**Abstract**: Text-based person retrieval is a developing downstream task of cross-modal retrieval and derives from conventional person re-identification, which plays a vital role in public safety and person search. In view of the problem of lacking query images in traditional person re-identification, the main challenge of this task is that it combines two different modalities and requires that the model have the capability of learning both image content and textual semantics. To narrow the semantic gap between pedestrian images and text

* 基金项目: 国家重点研发计划 (2018AAA0102200); 国家自然科学基金 (62036012, 62002355, 61720106006, 62102415, 62106262, 62072455, 62202331, 62206200); 天津市自然科学基金 (22JCYBJC00030); 北京市自然科学基金 (L201001, 4222039)
收稿时间: 2022-11-18; 修改时间: 2023-02-28; 采用时间: 2023-06-15; jos 在线出版时间: 2023-11-15
CNKI 网络首发时间: 2023-11-16

descriptions, the traditional methods usually split image features and text features mechanically and only focus on cross-modal alignment, which ignores the potential relations between the person image and description and leads to inaccurate cross-modal alignment. To address the above issues, this study proposes a novel relation alignment-based cross-modal person retrieval network. First, the attention mechanism is used to construct the self-attention matrix and the cross-modal attention matrix, in which the attention matrix is regarded as the distribution of response values between different feature sequences. Then, two different matrix construction methods are used to reconstruct the intra-modal attention matrix and the cross-modal attention matrix respectively. Among them, the element-by-element reconstruction of the intra-modal attention matrix can well excavate the potential relationships of intra-modal. Moreover, by taking the cross-modal information as a bridge, the holistic reconstruction of the cross-modal attention matrix can fully excavate the potential information from different modalities and narrow the semantic gap. Finally, the model is jointly trained with a cross-modal projection matching loss and a *KL* divergence loss, which helps achieve the mutual promotion between modalities. Quantitative and qualitative results on a public text-based person search dataset (CUHK-PEDES) demonstrate that the proposed method performs favorably against state-of-the-art text-based person search methods.

**Key words**: person retrieval; cross-modal task; textual semantic learning; relation alignment; attention mechanism

随着社会经济的发展, 计算机技术的进步和硬件能力的不断提升, 互联网时代产生的信息量呈井喷式增长, 而频繁的信息交互导致数据模态的多元化, 推动了跨模态学习任务[1−4]的蓬勃发展. 与此同时, 在高新科技助力于智慧城市建设的时代背景下, 社会公共安全领域对视频监控的需求日益增加. 自 AlexNet[5]于 2012 年夺得 ImageNet[6]图像识别大赛冠军之后, 卷积神经网络 (convolutional neural network, CNN) 开始被广泛地应用到图像领域相关的任务之中. 利用海量视频监控所捕获的图像数据和基于 CNN 的深度学习模型相结合, 使得行人重识别领域[7−10]得到了飞速发展.

行人重识别是一个跨非重叠摄像头的图像检索任务, 在社会安防领域具有广泛的应用场景. 但由于摄像头安放位置、拍摄角度的不同, 所拍摄的行人图像质量往往参差不齐、行人姿态差异较大, 导致传统行人重识别任务面临严峻的挑战. 虽然目前的深度学习模型在行人重识别的几个公开数据库上[11−13], 取得了与人类识别结果不相上下的性能表现, 但在实际应用中, 传统行人重识别还往往面临缺少查询图像的情况, 而文本描述相对图像更加方便获取, 于是基于文本描述的行人检索应运而生. 基于文本内容的行人检索[1]作为跨模态检索领域的一个子任务, 由行人重识别衍生而来, 其目标是在给定的检索条件下, 从数据库中找到与文本描述相对应的目标人物图像. 相比于传统的单模态图像检索的行人重识别任务, 基于文本的行人检索适用于缺少可用查询图像的情况, 不仅文本描述更加容易获得, 而且应用上更加灵活且贴合实际场景. 例如在警方追寻疑犯时, 目击证人并不能提供清晰有效、可供检索的疑犯相片, 只能用口头描述的形式提供犯罪嫌疑人的外貌特征, 此时可将目击证人的语言描述转为文本记录作为依据, 从而在视频监控中追踪到符合所描述特征的犯罪嫌疑人去向. 近年来, 行人检索因其在公共安全和视频监控领域中的广泛应用而受到越来越多的关注, 为寻找犯罪嫌疑人和失踪人员提供了巨大帮助. 尽管目前基于深度学习的方法取得了具有一定性能的检索结果, 但由于行人图像和文本描述模态间的语义差别较大, 基于文本描述的行人检索仍是一个具有挑战性的问题, 要求模型同时具有图像理解能力和文本语义学习能力, 并能充分挖掘模态间及模态内部的潜在联系.

作为一个跨模态语义理解的检索任务, 基于文本描述的行人检索数据库中的每幅图像只包含一个特定的行人, 而不像传统跨模态检索任务或图像文本匹配任务中图像可能包含多个对象类别. 同时, 文本描述查询中提供了相关行人的更多详细信息, 而不是粗略地概括图像中的对象. 因此, 由于任务的特殊性, 许多在一般跨模态检索基准, 如 Microsoft COCO (MSCOCO)[14]和 Flickr30k[15]上所提出的方法难以进行适用和推广. 基于文本描述的行人检索面临的首要挑战是图像和文本的模态间语义鸿沟, 即数据库中两种模态数据的特点相差甚远. 比如, 行人图像的特征结构往往相对固定, 数据库中的行人在图像中都是站立或行走的, 身体各个部位在图像中的位置都是相对固定的, 导致图像特征的信息结构相对固定; 而文本描述的表现形式多变, 在语法、词汇等方面具有很大的不确定性, 并且可以从任意位置、多个角度对图片内容进行描述, 不具备固定的信息结构和规律性. 其次, 现有的方法[16−19]对提取的图像特征和文本特征进行机械地分割, 只关注于跨模态信息的语义对齐和实例级别的特征匹配, 忽略了图像和文本模态内部的潜在联系, 破坏了原有的信息结构, 导致模态间细粒度匹配的不准确.

为了解决上述挑战, 传统的基于文本描述的行人检索方法按照研究方法的不同可分为基于网络结构设计的方法[1,20,21]、基于属性的方法[18,19,22,23]和基于注意力机制的方法[16,17]. 其中基于网络结构设计的方法通过不同特征提取网络的组合、子网络的堆叠、分支结构等设计进行模态间的实例级别匹配, 提升检索性能. 比如, Li 等人[1]开创了基于文本描述的行人检索任务的先河, 提出了基于门控神经注意力机制的文本与图像间相似性学习的方法, 该方法通过网络结构设计创新性地融合了图像和文本特征, 但仅用门控机制控制模型对于细粒度的学习, 取得的效果差强人意. Zheng 等人[20]则首次使用双分支卷积神经网络来分别提取文本和图像特征, 利用卷积块的堆叠和平均池化策略, 并提出了一个新的分类损失用于挖掘不同模态内的细微差异, 得到了研究人员的广泛关注. 但基于网络结构设计的方法要么直接将整个的图像和文本描述输入网络, 忽略了细粒度特征, 要么将图像进行分割, 破坏了输入数据的原始信息结构, 而且无法解决背景噪音对模型的干扰. 为了解决背景环境对检索结果所带来的影响, 基于属性的行人检索方法将行人图像分为多个属性, 极大降低了背景噪音的干扰, 比如 Wang 等人[18]从属性对齐的角度出发, 首次将对比学习应用到行人检索任务中, 减少了背景噪音对检索所带来的干扰, 但此方法对图像特征进行处理的本质还是将行人图像分割为多个属性, 没有解决其所造成的图像内部原始信息被破坏的问题, 丢失了语义信息. 而基于注意力机制的方法往往通过特征增强获得具有良好表示能力的特征表示, 通过多尺度的特征匹配, 缩小语义鸿沟. 注意力机制在跨模态检索领域最早应用于视觉问答中[24−28]. 例如, 为了解决细粒度推理不准确的问题, Yang 等人[26]提出了一种堆叠注意力网络, 通过递归关注与问题相关的图像区域来细化联合特征, 从而提高问答的准确性, 该网络是注意力机制在多模态领域的首次应用. 而在基于文本描述的行人检索任务上, Jing 等人[29]提出注意力机制结合姿态检测器的方法, 以获得更多姿态相关的细粒度信息, 但这种方法过于依赖姿态检测器的性能. 此外, Niu 等人[17]首次提出了一个多粒度的图像与文本对齐的方法, 不需要复杂的预处理即可进行端到端的训练. 但该方法对图像和文本特征进行机械地分割且不加以任何限制, 破坏了原始的信息结构和固有的潜在联系, 造成细粒度信息丢失.

为了解决上述问题, 本文提出了一个模态间关系促进的行人检索模型, 对于提取到的视觉和文本特征序列, 首先利用注意力机制分别构建模态内自注意力矩阵和跨模态注意力矩阵. 不同于以前的方法将注意力机制用于特征增强, 本方法将注意力矩阵看作不同特征序列间的响应值分布, 矩阵中的每一个元素都表示一个对应关系, 以便挖掘潜在联系. 然后, 分别使用两种不同的矩阵构建方法重构模态内自注意力矩阵和跨模态注意力矩阵. 其中模态内逐元素重构自注意力矩阵通过元素提取的方式, 可以很好地挖掘模态内部的潜在联系, 构建出与原始自注意力矩阵维度相同的矩阵分布. 而跨模态注意力矩阵的重构利用了矩阵相乘的特性, 以跨模态信息为桥梁, 构建出与原始跨模态注意力矩阵维度相同的矩阵分布, 可充分挖掘模态间的潜在语义信息, 缩小语义鸿沟. 最后, 用 Kullback-Leibler (*KL*) 散度损失和基于任务的跨模态投影匹配 (cross-modal projection matching, CMPM) 损失联合约束模型优化. 其中 *KL* 散度用于约束原始注意力矩阵和重构矩阵的矩阵分布, 跨模态投影匹配损失用于约束不同模态的多尺度特征表示. 模态间关系促进的行人检索方法通过注意力机制充分挖掘自然语言描述和行人图像模态内、模态间潜在联系, 从而可以更加有效且精准地建立文本与图像内容之间对应的匹配映射关系, 减小模态间差异. 最后, 本研究在公开数据库 CUHK-PEDES[1]上进行了实验.

本文第 1 节介绍基于文本描述行人检索的相关方法和研究现状. 第 2 节详细介绍了研究方法及其组成模块. 第 3 节进行实验验证与分析所提模型的有效性. 最后总结全文.

## 1 相关工作

### 1.1 单模态行人重识别

行人重识别[30−32]是视频监控领域的一项重要研究内容, 这一领域的研究有助于快速准确地检索大型图像或者视频数据. 最近, 行人重识别越来越受到学术界和工业界的广泛关注, 其中最先进的方法主要依赖新兴的深度学习技术. Ye 等人[33]将行人重识别领域分为封闭世界和开放世界两个主要的研究方向. 封闭世界中的数据模态是单一的, 需要额外生成行人目标检测边界框, 数据有足量的标注且都默认正确, 同时查询默认存在于数据库中. 在封闭

世界, 全局特征表示学习 PersonNet[34]最先利用深度神经网络提取全局特征向量来表示行人图片. 而这之后, 注意力机制的应用大大提升了特征表示学习的效果. 为解决行人错位问题, Li 等人[35]通过联合软注意力和硬区域注意力, 增强特征表示对错位情况的鲁棒性; 此外, Wang 等人[36]提出完全注意力模块, 创建通道方向和空间方向的注意信息, 以挖掘用于单模态行人重识别的行人. 另一方面, 为解决多人图像的特征提取问题, Chen 等人[37]提出了一种上下文感知的特征学习方法, 结合序列内和序列间的注意力进行特征对齐和细化. 但全局表示通常对局部特征的错位变换不具有鲁棒性, 导致局部特征的对齐出现错误. 为解决此类问题, 出现了根据姿态或水平切分的局部特征表示学习[38~40], 并将全局特征和局部特征进行结合, 以更好地实现特征匹配和对齐. 虽然封闭世界的检索任务取得了诸多成果, 但开放世界往往更加贴合实际应用场景, 因为开放世界中的数据模态是多样的, 通过原始数据直接进行端到端的行人重识别, 标注不完全且有噪音, 甚至查询可能不存在于数据库中. 基于文本描述的行人检索任务同时结合了视觉内容和文本描述两种模态的数据, 是属于开放世界中的一个子任务, 解决了单模态行人重识别中缺少查询图像的问题.

### 1.2 基于文本描述的行人检索

相比于封闭世界内的单模态行人重识别任务, 基于文本的行人检索的应用场景更加灵活且贴合实际. 比如在疑犯追踪时, 警方往往不具备疑犯的相关照片, 这时可通过目击证人对疑犯外形的描述, 追寻疑犯踪迹. 按照研究方法的不同, 可将现有的基于文本的行人检索方法分为基于网络结构设计的方法、基于属性的方法和基于注意力机制的方法. 接下来, 将对不同的方法进行详细介绍.

● 基于网络结构设计的行人检索方法. 为了对特定任务取得更好的效果, 深度学习网络结构的设计策略包括不同特征提取网络的组合、子网络的堆叠、分支结构等设计方法. Liu 等人[41]通过抑制模态差异来解决跨模态行人重识别问题, 提出了单向度量和基于记忆的增强方法, 用代理的方式缓解了中继效应, 从而进一步增强了跨模态关联. Yu 等人[42]提出了基于级联 Transformer 的端到端行人检索框架, 利用 Transformer 能够处理行人不同尺度、姿态和视角的变化, 但将其应用于跨模态行人检索问题中, 有待进一步研究. 针对基于自然语言描述的行人检索这一问题, Zheng 等人[20]首次使用双分支卷积神经网络结合平均池化策略来提取文本和图像特征, 并提出了一个新的分类损失用于挖掘同模态内的细微差异. 最近, Chen 等人[21]在双路特征提取的基础上, 结合多尺度特征匹配的思想, 利用残差网络[43]的堆叠和预训练好的 BERT[44], 得到了一个高性能的基线. 但传统基于网络结构设计的方法往往只关注模态间的实例匹配, 对于分割的特征不加以任何限制, 导致模态内的原始信息结构被破坏, 从而语义信息丢失, 影响检索效果. 本文提出的模态间关系促进的行人检索方法, 不同于 Chen 等人[21]提出的方法, 通过两种不同的矩阵构建方式, 分别挖掘模态内部和模态之间的潜在联系, 用注意力矩阵表示特征序列间的响应关系, 最后通过 *KL* 散度结合跨模态投影匹配损失对模型进行优化, 有效减少了原始语义信息的丢失, 提升了检索精度.

● 基于属性的行人检索方法. 在基于网络结构设计的行人检索方法的基础之上, 基于属性方法的核心思想认为, 背景环境对于基于文本描述的行人检索往往起到负面影响, 故采用合适方法消除背景噪音对于跨模态行人检索至关重要. 以往的方法对于图像直接采用全局特征表示, 不仅容易受到背景噪音的影响, 而且容易忽略图像细粒度特征. 为了解决这些问题, Wang 等人[22]首次引入属性的概念, 提出了一种语义自对齐网络, 设计了一个多视图网络来捕获人身体各个部位之间的关系, 利用文本对图像的描述提供额外的监督. 而 Chen 等人[23]提出在训练阶段利用文本信息作为属性来辅助图片进行特征的学习. 类似地, Aggarwai 等人[19]从文本语料库中挖掘属性标注行人图片, 通过学习属性驱动空间和类信息驱动空间来获得检索结果. 最近, Wang 等人[18]从属性对齐的角度出发, 首次将对比学习应用到行人检索任务中, 用一个轻量级的辅助属性分割层将行人的特征空间分割成与属性相对应的子空间, 然后通过对比损失将图像特征与文本属性进行语义对齐, 使模型更好地学习属性短语与视觉区域的对应关系. 以上基于属性的检索方法虽然减少了背景噪音对匹配结果的影响, 实现了细粒度的特征匹配, 但基于属性的检索方法本质还是对图像进行分割, 不可避免地破坏了图像内部的上下文语义信息, 对于跨模态任务, 这是得不偿失的. 本文所提出的模态间关系促进的行人检索方法通过 *KL* 散度测量文本和视觉内部关系之间的距离来量化关系的一致性, 充分挖掘模态内部潜在的上下文语义信息, 避免了由于属性分割造成的上下文语义信息丢失的情

况, 取得了不错的检索结果.

● 基于注意力机制的行人检索方法. 基于注意力机制的行人检索方法自提出以来就受到了研究人员的广泛关注, 其具体做法往往是通过注意力机制对提取到的特征进行特征增强, 然后将多模态的特征进行细粒度的交互, 实现图像与文本的特征对齐[1,16,17]. 例如, Niu 等人[17]首次提出了一个多粒度的图像与文本对齐的方法, 提取的图像和文本特征分别都有各自的全局和局部特征表示, 4 种特征组合成 4 种不同的对齐方式, 不需要复杂的预处理即可进行端到端的训练. 而 Zheng 等人[16]利用硬注意力机制从图像和文本中自适应地选择语义相关度强的图像区域和单词短语, 进行细粒度对齐和相似度计算, 此方法可以很好地融合不同模态强相关的特征, 缓解匹配冗余问题. Gao 等人[45]提出了一种能够多尺度自适应对齐图像和文本特征的方法, 通过上下文非局部注意力机制来挖掘多尺度的语义对应关系. 不同于基于跨模态注意力机制的方法, Farooq 等人[46]提出上下文语义对齐的共享注意力模块, 结合模态内注意力机制, 学习隐式特征映射关系. 但是, 现有的基于注意力机制的方法都只关注模态间的实例特征匹配, 忽略了模态内部的潜在上下文信息. 为了解决上述问题, 本文把注意力矩阵看作不同特征序列间的响应值分布, 并通过模态内逐元素重构矩阵和模态间整体重构矩阵两种不同的矩阵构建方式, 分别重构模态内自注意力矩阵和模态间注意力矩阵, 最后通过 *KL* 散度约束原矩阵与重构矩阵的特征分布. 实验结果证明, 这种把注意力矩阵看作不同特征序列间的响应值分布的方式, 比传统利用注意力矩阵进行特征增强的方法更加关注语义之间的潜在联系, 同时充分利用了模态内和模态间的特征关系, 得到了更好的效果.

## 2 模态间关系促进的行人检索方法

本文提出的模态间关系促进的行人检索方法总体流程如图 1 所示. 具体来说, 对于提取的图像和文本特征序列, 首先利用注意力机制分别生成各自模态的自注意力矩阵和跨模态注意力矩阵, 然后分别通过模态内 (intra modal, IM) 逐元素重构矩阵和模态间 (cross modal, CM) 整体重构矩阵两种不同的矩阵构建方法分别对模态内和模态间的矩阵进行重构, 从而充分挖掘模态内部和模态间的潜在信息. 将重构后的矩阵与原矩阵用 *KL* 散度进行约束, 相互校准两种模态内部的自注意力矩阵, 使得特征序列间的潜在关系趋于相似的分布, 达到缩小图像和文本信息的语义鸿沟的目的, 从而得到高鲁棒性、多尺度的图像和文本特征, 使得模型能够得到更加精准的检索结果. 在本章节, 首先将分别介绍图像和文本的特征提取模块, 然后对矩阵构建模块中的模态内逐元素重构矩阵和模态间整体重构矩阵进行详细描述. 最后, 给出了模型所采用的损失函数.

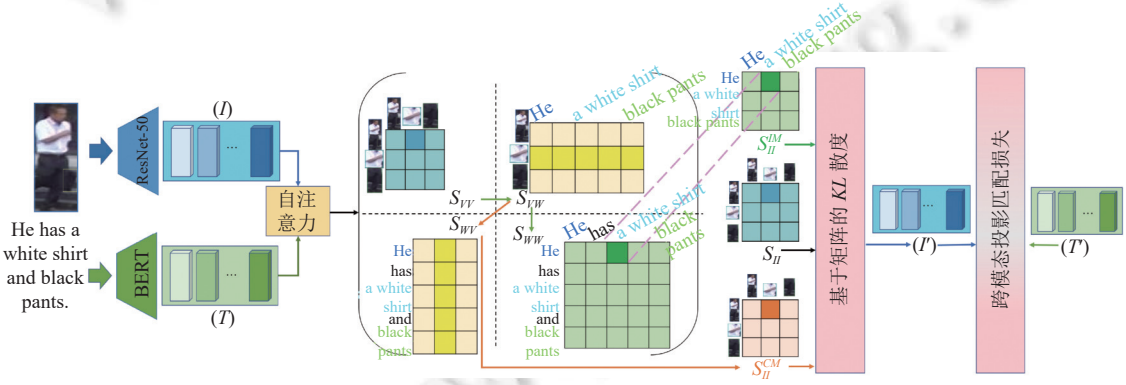



图 1 模态间关系促进的行人检索模型框架图

### 2.1 特征提取模块

● 图像特征提取. 参考 Chen 等人[21]的方法作为基线, 提取多尺度的图像特征. 具体来说, 在训练阶段, 假设训练数据为 $D=\{I_i,T_i\}_{i=1}^{N}$, 其中 $N$ 表示每个批次中图像文本对的数量, 每一对包含一张行人图像和对应的文本描述. 本文所提方法采用 ResNet-50[43]作为提取图像特征的主干网络, 由 4 个残差块组成, 获得多个尺度的局部特征表

示, 并采用 PCB (part-based convolutional baseline)[47]策略, 将高层特征均匀分割成 $F$ 个局部特征, 之后采取全局最大池化获得全局特征表示, 用 $v^g$ 表示全局特征, 与局部特征表示进行连接, 得到图像特征序列 $I=\{v^1,\ldots,v^F,v^g\}$, $I\in R^{n\times d}$, 其中 $n=F+1$, $d$ 代表特征维度.

● 文本特征提取. 最近, BERT 模型由于其易于使用、稳定性强等优点, 得到了研究人员的广泛关注. 文本特征提取模块通过在大型语料库上预训练的 BERT[44]提取单词特征表示并学习单词之间的上下文关系. 具体来说, 将每个文本描述 $T$ 分解成一个单词列表, 在每个句子的开头和结尾分别插入 [CLS] 和 [SEP]. 为确保文本长度的一致性, 当文本长度大于 $L$ 时, 舍弃掉 $L$ 之后的内容, 当文本长度小于 $L$ 时, 在文本末尾用 0 填充至长度 $L$. 然后将每个标记后的文本描述输入到经过预处理并固定参数的 BERT 模型中, 以提取单词特征表示 $t\in R^{L\times C}$, 其中 $C$ 表示单词特征维度. 之后, 将单词特征 $t$ 从 $t\in R^{L\times C}$ 延展至 $t\in R^{1\times L\times C}$, 其中 1 看作卷积输入的高度, $L$ 为宽度, $C$ 为通道尺寸. 最后通过多分支结构将低级文本特征映射为与图像特征相同的维度, 将全局文本特征 $t^g$ 与局部特征进行连接, 得到 $T=\{t^1,\ldots,t^F,t^g\}$, $T\in R^{n\times d}$.

### 2.2 矩阵重构模块

自注意力机制可以有效地捕捉文本及图像内部信息间存在的潜在联系. 不同于以往利用注意力机制对特征进行增强的方法, 本文所提方法将注意力机制矩阵中的元素看作不同特征序列之间的响应值. 为了使不同模态的特征可以更好地匹配和对齐, 在缩小模态间语义鸿沟的同时充分挖掘特征内部的潜在联系, 具体来说, 本文提出了两种矩阵构建方法, 用于构建和原始注意力矩阵维度相同的特征矩阵表示. 首先, 对于特征提取模块中得到的图像特征序列 $I\in R^{n\times d}$ 和文本特征序列 $T\in R^{n\times d}$ 来说, 通过注意力机制分别求得模态间和模态内的自注意力矩阵, 具体公式如下:

$$Att(Q,K,V)=\sigma(QK^{\top})V=\sigma(S)V \tag{1}$$

$$Q=XW^{Q} \tag{2}$$

$$K=XW^{K} \tag{3}$$

其中, $\sigma$ 为按行的 $Softmax$ 操作, $S$ 作为矩阵表示向量间的相似度. $X$ 分别代入图像特征 $I\in R^{n\times d}$ 和文本特征 $T\in R^{n\times d}$, $W^{Q}$ 和 $W^{K}$ 均为可学习的参数.

$X$ 表示图像特征序列 $I\in R^{n\times d}$ 或文本特征序列 $T\in R^{n\times d}$. 具体来说: 1) 当图像特征序列 $I$ 作为公式 (1) 中的 $Q$、$K$、$V$ 时, 得到关于图像的模态内自注意力矩阵 $S_{II}$; 2) 当文本特征序列 $T$ 作为公式 (1) 中的 $Q$、$K$、$V$ 时, 得到关于文本的模态内自注意力矩阵 $S_{TT}$; 3) 当图像特征序列 $I$ 作为公式 (1) 中的 $Q$、$V$, 文本特征序列 $T$ 作为 $K$ 时, 得到图像对于文本的跨模态注意力矩阵 $S_{IT}$; 4) 当文本特征序列 $T$ 作为公式 (1) 中的 $Q$、$V$, 图像特征序列 $I$ 作为 $K$ 时, 得到文本对于图像的跨模态注意力矩阵 $S_{TI}$. 其中, 图像的模态内自注意力矩阵 $S_{II}$ 表示图像内部的特征关系分布, 而文本的模态内自注意力矩阵 $S_{TT}$ 表示文本内部的特征关系分布. 以图像的模态内自注意力矩阵 $S_{II}$ 为例, 矩阵的每一行代表当前某一部分的图像特征序列与图像所有图像特征序列的响应关系, 元素所对应的响应值越大, 代表当前区域和图像特征序列越相关. 另一方面, 跨模态的注意力机制在面对不同数据模态的情况下, 用于更好地挖掘语义信息. 图像对于文本的跨模态注意力矩阵 $S_{IT}$ 表示图像特征对于文本特征的关系分布, 而文本对于图像的跨模态注意力矩阵 $S_{TI}$ 表示文本特征对于图像特征的关系分布. 以 $S_{IT}$ 为例, 矩阵的每一行表示每一个图像特征序列对于所有文本特征序列的相关性.

● 模态内逐元素重构矩阵. 模态内逐元素重构矩阵的整体思路如下: 以图像为例, 自注意力矩阵 $S_{II}$ 的每一个元素所代表的响应关系, 都理应可以在文本自注意力矩阵中找到对应的元素值, 以这种方式逐元素构建出与视觉自注意力矩阵维度相同的一个矩阵, 即 $S_{II}^{IM}$. 具体来说, 首先将视觉序列 $I\in R^{n\times d}$ 的输入作为锚点, 对于跨模态注意力矩阵 $S_{IT}$ 中每一行的最大值, 即为与当前视觉元素最相关的文本元素. 通过这种方式, 对于矩阵 $S_{IT}$ 中的任意两行所对应的最大元素, 都可以在 $S_{TT}$ 矩阵中找到两元素的对应位置, 将其作为新构造矩阵的元素, 逐元素生成重构矩阵. 即对于视觉部分的每一对关系, 都找到了与之对应的文本关系. 算法具体流程如算法流程图 1 所示.

通过模态内逐元素重构矩阵, 所重构出的矩阵充分挖掘并利用了图像和文本模态内部的潜在联系, 将重构后

的矩阵 $S_{II}^{IM}$ 与原矩阵 $S_{II}$ 通过 $KL$ 散度进行约束, 可以让两个矩阵的分布趋于相似, 使模态间的关系充分接近, 从而缩小语义鸿沟. 对于文本序列 $T \in R^{n \times d}$ 作为锚点时, 采用对称的操作, 具体算法流程如算法 1 所示. 其中 $Ext(S_{TT}, T_{i^*}, T_{j^*})$ 和 $Ext(S_{II}, I_{i^*}, I_{j^*})$ 分别表示从文本模态内自注意力矩阵 $S_{TT}$ 和图像的模态内自注意力矩阵 $S_{II}$ 矩阵中提取对应的第 $i^*$ 行第 $j^*$ 列元素.

**算法 1**. 模态内逐元素重构矩阵.

输入: 模态内自注意力矩阵 $S_{II}$, $S_{TT}$;
输出: 矩阵 $S_{II}^{IM}$, $S_{TT}^{IM}$.

**FOR** $i = 1$ **to** $n$ **DO**
  $i^* \leftarrow \text{argmax} S_{IT}[i:]$
  **FOR** $j = 1$ **to** $n$ **DO**
    $j^* \leftarrow \text{argmax} S_{IT}[j:]$
    $S_{II}^{IM}[i, j] \leftarrow Ext(S_{TT}, T_{i^*}, T_{j^*})$
  **END FOR**
**END FOR**
**FOR** $i = 1$ **to** $n$ **do**
  $i^* \leftarrow \text{argmax} S_{TI}[i:]$
  **FOR** $j = 1$ **to** $n$ **do**
    $j^* \leftarrow \text{argmax} S_{TI}[j:]$
    $S_{TT}^{IM}[i, j] \leftarrow Ext(S_{II}, I_{i^*}, I_{j^*})$
  **END FOR**
**END FOR**

● 模态间整体重构矩阵. 模态间整体重构矩阵方法如图 2 所示, 通过矩阵相乘的特性, 图像对于文本的跨模态注意力矩阵 $S_{IT}$ 中的每一行, 都表示当前行所对应的图像特征与所有文本描述的相关程度. 类似地, 文本对于图像的跨模态注意力矩阵 $S_{TI}$ 中的每一列, 都表示当前列所对应的图像特征与所有文本描述的相关程度. 公式表示如下:

$$S_{II}^{CM} = S_{IT} \cdot S_{TI} \tag{4}$$

$$S_{TT}^{CM} = S_{TI} \cdot S_{IT} \tag{5}$$

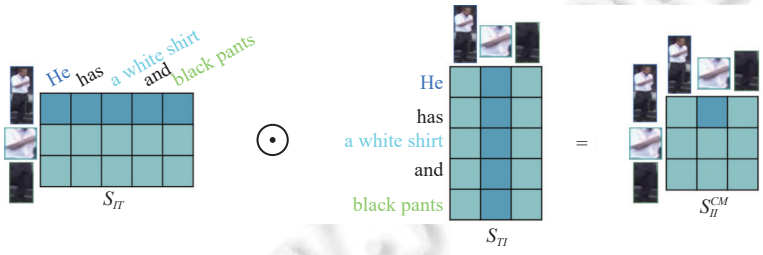



图 2 模态间整体重构矩阵方法示意图

和原始的图像模态内自注意力矩阵 $S_{II}$ 相比, $S_{II}^{CM}$ 矩阵也可以用来表示图像特征序列之间的关系. 但是不同于模态内自注意力矩阵 $S_{II}$, 矩阵 $S_{II}^{CM}$ 是充分挖掘了视觉模态间潜在联系所得到的关系矩阵. 如此重构矩阵的优势在于, 将跨模态注意力矩阵 $S_{IT}$ 和 $S_{TI}$ 直接相乘, 可以自然地得到与原始自注意力矩阵维度相同的矩阵 $S_{II}^{CM}$, 并且矩阵 $S_{II}^{CM}$ 中横纵坐标所对应的每一个元素都代表图像模态内部的潜在联系. 这种方式以跨模态注意力矩阵作为信息交互的桥梁, 可以充分利用模态间注意力, 挖掘出模态内部的潜在联系. 同样地, 对文本特征序列的跨模态注意力

矩阵采用对称的操作, 可以得到与文本特征序列维度相同的关系矩阵 $S_{TT}^{CM}$.

### 2.3 损失函数

最后使用 *KL* 散度结合跨模态投影匹配[48]损失对所提模型进行优化. 其中 *KL* 散度可以衡量生成的矩阵分布与数据原分布之间的差异, 而 CMPM 损失可以最小化两个模态的特征投影分布.

● *KL* 散度. *KL* 散度用于比较两个矩阵分布的接近程度, 给定两个维度相同的矩阵 $A$ 和 $B$, 基于矩阵的 *KL* 散度[49]可用于测量矩阵 $A$ 和 $B$ 之间的距离:

$$M\text{-}KL(A,B)=\sum_{i}^{N}KL(A_i\|B_i)+KL(B_i\|A_i) \tag{6}$$

其中, $A_i$ 和 $B_i$ 表示矩阵 $A$ 和矩阵 $B$ 的第 $i$ 个行向量, *M-KL* 表示基于矩阵的 *KL* 散度. 利用上述公式分别约束图像特征自注意力矩阵 $S_{II}$ 与模态内逐元素重构矩阵模块得到的矩阵 $S_{II}^{CM}$ 和文本特征自注意力矩阵 $S_{TT}$ 与模态内逐元素重构矩阵模块得到的矩阵 $S_{TT}^{IM}$. 通过这种方式, 拉近原矩阵和模态内逐元素重构矩阵方法所得矩阵间的距离, 使其分布趋于相似, 从而充分挖掘潜在语义联系, 具体损失函数如下:

$$L_1=M\text{-}KL\left(\sigma(S_{II}),\sigma\left(S_{II}^{IM}\right)\right)+M\text{-}KL\left(\sigma(S_{TT}),\sigma\left(S_{TT}^{IM}\right)\right) \tag{7}$$

其中, $\sigma$ 表示 *Softmax* 操作. 同理, 在利用公式 (7) 计算损失的同时, 利用基于矩阵的 *KL* 散度分别约束图像的模态内自注意力矩阵 $S_{II}$ 与模态间整体重构矩阵模块得到的矩阵 $S_{II}^{CM}$ 和文本的模态内自注意力矩阵 $S_{TT}$ 与模态间整体重构矩阵模块得到的矩阵 $S_{TT}^{CM}$. 从而直接缩小模态间的矩阵分布距离, 以达到缩小语义鸿沟的目的, 具体损失函数如下:

$$L_2=M\text{-}KL(\sigma(S_{II}),\sigma(S_{II}^{CM}))+M\text{-}KL(\sigma(S_{TT}),\sigma(S_{TT}^{CM})) \tag{8}$$

● 跨模态投影匹配损失. 利用跨模态投影匹配损失函数[48], 对不同模态的多尺度特征进行约束. 一个批次中给定 $N$ 个图片文本对, 对于图片 $I_i$ 可构建图片文本对 $\{(I_i,T_j),y_{i,j}\}_{j=1}^{N}$, 其中 $y_{i,j}=1$ 表示 $I_i$ 和 $T_j$ 是同一个 id 类. 相反, $y_{i,j}=0$ 表示图像和文本不匹配. 全局图像特征 $v^g\in I_i$ 和全局文本特征 $t^g\in T_j$ 的匹配概率如公式 (9) 所示:

$$p_{i,j}^{g}=\frac{\exp((v_i^g)^{\top}\bar{t}_j^g)}{\sum_{k=1}^{N}\exp((v_i^g)^{\top}\bar{t}_k^g)}\quad \text{s.t. } \bar{t}_j^g=\frac{t_j^g}{\|t_j^g\|} \tag{9}$$

其中, $p_{i,j}^{g}$ 是一个批次对中 $(I_i,T_j)_{j=1}^{N}$, 全局特征的标量投影比例, 当图像特征与文本特征越相似, 前者到后者的标量投影越大. 由于每个批次中, 图像特征 $I_i$ 可能有多个匹配的文本描述, 所以将真实匹配概率归一化为:

$$q_{i,j}=\frac{y_{i,j}}{\sum_{k=1}^{N}y_{i,k}} \tag{10}$$

公式 (9) 为图像到文本方向的标量投影匹配概率, 为使不同模态间的匹配特征表示更加接近, 损失函数中还添加了图像到文本方向计算的对称操作. 全局特征的双向 CMPM 损失计算如公式 (11) 所示:

$$L_{\text{CMPM}}^{g}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(p_{i,j}^{g}\log\left(\frac{p_{i,j}^{g}}{q_{i,j}+\epsilon}\right)+p_{j,i}^{g}\log\left(\frac{p_{j,i}}{q_{j,i}^{g}+\epsilon}\right)\right) \tag{11}$$

其中, $p_{i,j}^{g}$ 表示全局特征图像到文本方向投影的匹配概率. $p_{j,i}^{g}$ 表示全局特征文本到图像方向投影的匹配概率. $\epsilon$ 为一个很小的数值, 用于防止出现分母为零的情况.

对于 $F$ 对图像文本局部特征的匹配概率如公式 (12) 所示:

$$p_{i,j}^{l}=\sum_{f=1}^{F}\frac{\exp((v_i^f)^{\top}\bar{t}_j^f)}{\sum_{k=1}^{N}\exp((v_i^f)^{\top}\bar{t}_k^f)}\quad \text{s.t. } \bar{t}_j^f=\frac{t_j^f}{\|t_j^f\|} \tag{12}$$

其中, $p_{i,j}^{l}$ 是在一个批次对中 $(I_i,T_j)_{j=1}^{N}$, 所有局部特征的标量投影的匹配概率之和.

由公式 (10) 和公式 (12) 可以得到局部特征的双向 CMPM 损失计算如公式 (13) 所示:

$$L_{\text{CMPM}}^{l}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(p_{i,j}^{l}\log\left(\frac{p_{i,j}^{l}}{q_{i,j}+\epsilon}\right)+p_{j,i}^{l}\log\left(\frac{p_{j,i}^{l}}{q_{j,i}+\epsilon}\right)\right) \tag{13}$$

CMPM 最终目标函数中包含以上两个尺度的跨模态匹配. 其中全局级的 CMPM 损失确保了所得到的最终特征表示具有更强的跨模态语义理解能力. 局部级的 CMPM 损失可以实现图像和文本之间的局部对齐. 通过这种方式, 图像文本特征表示的匹配度可以逐渐提高. CMPM 的总体计算公式如公式 (14) 所示:

$$L_{\mathrm{CMPM}} = L^{g}_{\mathrm{CMPM}} + L^{l}_{\mathrm{CMPM}} \tag{14}$$

结合基于矩阵的 $KL$ 散度约束, 模型的最终损失函数如公式 (15) 所示:

$$L = \lambda_1 L_1 + \lambda_2 L_2 + L_{\mathrm{CMPM}} \tag{15}$$

其中, $\lambda_1$, $\lambda_2$ 是平衡化参数.

## 3 实验分析

### 3.1 实验数据

为了验证所提模态间关系促进的行人检索方法的有效性. 本文在公开的基于文本描述的行人检索数据库 CUHK-PEDES[21]上进行验证.

CUHK-PEDES 数据库[21]为基于自然语言描述的跨模态行人检索任务常用数据库, 由香港中文大学首次提出. 其中的行人图片来自 5 个不同的行人重识别数据库 (CUHK03[13], Market-1501[11], SSM[50], VIPER[51], CUHK01[52]), 共 13 003 个行人, 图片 40 206 张. 由于各个数据库中图片背景、光照、拍摄角度的不同, 保证了行人图像的多样性. 数据库的划分采用原始的划分策略, 其中训练集包含 34 054 张图片, 11 003 个 id 和 68 126 条文本描述, 验证集包含 3 078 张图片, 1 000 个 id 和 6 158 条文本描述; 测试集包含 3 074 张图片, 1 000 个 id 和 6 156 条文本描述. 平均每张图片大概对应 2 条文本描述, 每条文本描述平均大约 23 个单词, 整个数据库共 9 408 个不同的单词. CUHK-PEDES[21]是目前基于文本行人检索任务的最常用的数据库.

### 3.2 实验设置

本方法以 Chen 等人[21]提出的 TIPCB 为基线, 分别用在 ImageNet[6]上预训练好的 ResNet-50[43]和大型语料库上预训练的 BERT 模型提取图像和文本特征. 所有输入图像的大小调整为 384×128, 文本长度统一为 64. 每个批次包含 56 个图像文本对. 在训练阶段, 使用 Adam 优化模型, 对于 CUHK-PEDES 数据库, 学习率设置为 0.000 9, 50 轮之后下降 0.1, 权重衰减设置为 $4\times10^{-5}$, 模型共训练 80 轮. 平衡化参数 $\lambda_1$, $\lambda_2$ 均设置为 1. 此外, 采用水平翻转进行数据增强, 每个图像有 50% 的几率随机翻转. 实验在单个 GTX3090GPU 上使用 PyTorch 进行. 本文采用 Top-$k$ 标准评估指标来评估模型性能, $k$ 分别取值 1, 5 和 10. 对于给定文本描述, 筛选出数据库中符合描述的图像, 并根据相似性值进行排序. 如果在前 $k$ 个图像中存在匹配的人物图像, 表示检索成功.

### 3.3 实验方法与现有方法的对比实验

如表 1 所示, 将本文所提方法与 CNN-RNN[53]、NeuralTalk[2]、GLA[23]、ViTAA[18]、CMAAM[19]、MIA[17]、HGAN[16]、Dual Path[20]、CMPM+CMPC[48]、TIPCB[21]、GNA-RNN[1]和 DSSL[54]等方法进行对比. 其中, GLA[23]、ViTAA[18]和 CMAAM[19]属于基于属性的方法; MIA[17]和 HGAN[16]为基于注意力机制的方法; Dual Path[20]、CMPM+CMPC[48]和 TIPCB[21]为基于网络结构设计的方法; GNA-RNN[1]和 DSSL[54]属于网络结构设计与注意力机制相结合的方法. 此外, 还将本方法与视觉语义嵌入的方法 CNN-RNN[53]和图像字幕生成领域的方法 NeruralTalk[2]进行对比. 具体来说, 基于属性的方法中, GLA[23]提出在训练阶段利用文本信息作为属性来辅助图片进行特征的学习. ViTAA[18]从属性对齐的角度出发, 用一个轻量级的辅助属性分割层将行人的特征空间分割成与属性相对应的子空间, 然后通过对比损失使模型更好地学习属性短语与视觉区域的对应关系. CMAAM[19]从文本语料库中挖掘属性标注行人图片, 通过学习属性驱动空间和类信息驱动空间来获得检索结果. 基于注意力机制的方法中, MIA[17]首次提出了一个多粒度的图像与文本对齐的方法, 不需要复杂的预处理即可进行端到端的训练. HGAN[16]利用硬注意力机制从图像和文本中自适应地选择语义相关度强的图像区域和单词短语, 很好地融合了不同模态强相关的特征. NAFS[45]提出了一种能够多尺度自适应对齐图像和文本特征的方法, 通过上下文非局部注意力机制来挖掘尽

可能多尺度的语义对应关系. AXM-Net[46]提出上下文语义对齐的共享注意力模块, 结合模态内注意力机制, 学习隐式特征映射关系. 基于网络结构设计的方法中, Dual Path[20]首次使用双分支卷积神经网络结合平均池化策略来提取文本和图像特征, 并提出了一个新的分类损失用于挖掘同模态内的细微差异. CMPM+CMPC[48]提出了跨模态投影匹配损失和跨模态投影分类损失, 前者最小化两个模态特征投影分布的 *KL* 散度; 后者对模态 *A* 在模态 *B* 上的投影特征进行分类, 进一步增强模态之间的契合度. TIPCB[21]在双路特征提取的基础上, 结合多尺度特征匹配的思想, 利用残差网络[43]的堆叠和预训练好的 BERT[44], 得到了一个高性能的基线. 而基于网络结构设计和注意力机制的方法中, GNA-RNN[1]基于门控神经注意力学习文本与图像间相似性, 用门控机制控制模型对于模态间细粒度关系的学习. DSSL[54]提出了一种深度环境行人分离学习模型和新的 RSTPReid 数据库, 其中环境行人分离与融合机制在信息正交约束下, 通过行人特征的分离与跨模态融合重构匹配, 实现了环境行人的剥离与匹配. 另外, 视觉语义嵌入方法 CNN-RNN[53]提出了一种用于零样本学习进行图像到文本检索的方法. 图像字幕生成方法 NerualTalk[2]提出基于循环神经网络生成描述图像的自然语句.

表 1　与现有的行人检索方法在 CUHK-PEDES 上的比较结果 (%)

| 方法 | Top-1 | Top-5 | Top-10 |
|---|---|---|---|
| CNN-RNN[53] | 8.07 | — | 32.47 |
| NeuralTalk[2] | 13.66 | — | 41.72 |
| GLA[23] | 43.58 | 66.93 | 76.26 |
| ViTAA[18] | 55.97 | 75.84 | 83.52 |
| CMAAM[19] | 56.68 | 77.18 | 84.86 |
| MIA[17] | 53.10 | 75.00 | 82.90 |
| HGAN[16] | 59.00 | 79.49 | 86.62 |
| NAFS[45] | 59.94 | 79.86 | 86.70 |
| AXM-Net[46] | 61.90 | 79.40 | 85.75 |
| Dual Path[20] | 44.40 | 66.26 | 75.07 |
| CMPM+CMPC[48] | 49.37 | — | 79.27 |
| TIPCB[21] | 62.44 | 82.06 | 88.23 |
| GNA-RNN[1] | 19.05 | — | 56.64 |
| DSSL[54] | 59.98 | 80.41 | 87.56 |
| 本文方法 | **63.23** | **82.76** | **88.97** |

注: —表示未给出结果

实验表明, 本文所提出的模态间关系促进的行人检索方法在公开的基于文本的行人检索数据库 CUHK-PEDES 上, Top-1, Top-5 以及 Top-10 结果分别达到了 63.23%, 82.76% 和 88.95%, 比现有的最好方法 AXM-Net[46]分别提高了 1.33%, 3.36% 和 3.22%. 本文所提方法在基于网络结构设计和注意力机制的基础上, 比传统基于属性的最优方法 CMAAM[19]在 Top-1, Top-5 以及 Top-10 结果分别提升了 6.55%, 5.58% 和 4.11%, 这是因为基于属性的方法对图像进行分割, 不可避免地破坏了图像内部的上下文语义信息, 而模态间关系促进的行人检索方法对于多尺度特征加以限制, 通过注意力机制将响应分布用矩阵表示, 再通过 *KL* 散度进行模型优化, 减少了潜在上下文语义信息的丢失; 而相比基于注意力机制的方法, 本文的方法区别于传统利用注意力机制进行特征增强的做法, 将注意力矩阵看成模态间、模态内关系的分布, 通过模态内逐元素重构矩阵和模态间整体重构矩阵的方法, 分别重构模态内和模态间的注意力矩阵, 使模型可以学习到更多潜在联系, 更大限度地保留了原始信息结构, 因此比基于注意力机制的最优方法 AXM-Net[44]在 Top-1, Top-5 以及 Top-10 结果分别提升了 1.33%, 3.36% 和 3.22%; 与基于网络结构设计的方法相比, 本文提出的方法不仅关注模态间的实例匹配, 同时对于实例级别的特征加以限制, 同时挖掘到了更多模态内部的潜在联系, 比基于网络结构设计的最优方法 TIPCB[21]在 Top-1, Top-5 以及 Top-10 结果分别提

升了 0.79%, 0.70% 和 0.74%; 相比于同时利用网络结构设计和注意力机制的方法 DSSL[54], 所提方法的好处在于避免了将行人和背景进行分割所造成的信息丢失, 保留了原始信息结构, 在 Top-1, Top-5 以及 Top-10 结果分别提升了 3.25%, 3.27% 和 2.35%.

### 3.4 消融实验

本文在 CUHK-PEDES 数据库上进行了消融实验以验证所提方法的有效性, 实验结果如表 2 所示. 通过比较有或没有特定模块的模型性能, 我们可以得知各模块对模型性能的贡献, 实验设计了如下模型的变体.

表 2 在 CUHK-PEDES 数据集上, 各模块的消融实验结果 (%)

| 方法 | IM | CM | Top-1 | Top-5 | Top-10 |
|---|---|---|---|---|---|
| Baseline | — | — | 62.44 | 82.06 | 88.23 |
| B+IM | √ | — | 63.20 | 82.34 | 88.70 |
| B+CM | — | √ | 63.18 | 82.75 | 88.97 |
| 本文方法 | √ | √ | **63.23** | **82.76** | **88.97** |

● Baseline: 基线模型为 Chen 等人[21]提出的 TIPCB, 分别用 ImageNet[6]上预训练好的 ResNet-50[41]和大型语料库上预训练的 BERT 模型提取图像和文本的特征表示. 最后, 利用跨模态投影匹配损失来计算文本查询和目标特征之间的匹配分数.

● B+CM: 该变体表示仅对模型添加模态间整体重构矩阵模块, 而不使用模态内逐元素重构矩阵模块. 此变体将原始的模态内自注意力矩阵和模态间整体重构矩阵所构建出的矩阵用 *KL* 散度进行约束. 通过与基线进行比较, 可以评估模态间整体重构矩阵模块的效果.

● B+IM: 该变体表示仅对模型添加模态内逐元素重构矩阵模块, 而不使用模态间整体重构矩阵模块. 该变体将原始的模态内自注意力矩阵和模态内逐元素重构矩阵所构建出的矩阵用 *KL* 散度进行约束. 通过与基线进行比较, 可以评估模态内逐元素重构矩阵模块的效果.

通过表 2 可以观察到, 在仅保持基线模型的情况下, Top-1, Top-5 和 Top-10 准确度分别为 62.44%, 82.06% 和 88.23%. 当添加 IM 模块时, Top-1 指标提升 0.76%, 这是因为 IM 模块通过模态内自注意力重构矩阵的方式, 可以充分挖掘并利用图像和文本模态内部的潜在联系, 故可以带来检索精度的全面提升. 而当增加 CM 模块时, Top-1 精度与单独使用 IM 模块持平, 但 Top-5 与 Top-10 精度分别提升 0.41% 和 0.27%, 说明 CM 模块通过模态间整体重构矩阵的方式, 所重构出的矩阵对于模型全面理解跨模态语义信息具有重要意义, 可以帮助模型检索出更多符合文本描述的行人图像. 当同时使用 IM 和 CM 模块时, 模型性能在 Top-1, Top-5 和 Top-10 均达到最佳, 证明 IM 和 CM 模块可以同时利用模态内和模态间的潜在联系, 得到模态间关系促进的效果.

### 3.5 行人检索结果展示

为了验证所提方法在基于文本描述行人检索方面的效果, 后文图 3 展示了对于输入文本描述的行人检索结果. 绿色边框下的图像表示检索符合结果, 红色表示检索与真实答案不相符, 检索结果给出了最符合文本描述的前 5 张行人图像. 由图 3 可以看出, 即使是红色边框下的不匹配人物图像的行人, 也与目标人物的外观非常相似, 可见本文所提方法可以根据文本描述有效检索数据库中符合描述的行人图像.

## 4 总 结

本文提出了一种模态间关系促进的行人检索方法, 用跨模态检索损失约束不同模态间的特征, 使模型更好地建模跨模态语义关系. 同时通过两种矩阵构建方法构建注意力矩阵, 分别用于表示模态内特征的关系和模态间特征的关系, 通过拉近相对应的矩阵间的距离, 使模型不同模态间的关系相互促进. 实验验证了本文提出的模态间关系促进的行人检索方法具有优越的性能, 提取的特征具有良好的检索表现, 在基于文本的行人检索任务上取得了最好的性能表现. 可以预见, 随着跨模态任务和行人检索应用的普及, 未来会有越来越多的科研工作聚焦于跨模态

检索任务, 如视频检索等. 希望本文可以为相关工作提供一个充分的基准性能和完善的验证实验设置, 对后续的相关研究有所启发.


图 3 行人检索在 CUHK-PEDES 数据库上的结果示例, 检索结果上方代表查询文本

**References:**

[1] Li S, Xiao T, Li HS, Zhou BL, Yue DY, Wang XG. Person search with natural language description. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 5187–5196. [doi: 10.1109/CVPR.2017.551]

[2] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 3156–3164. [doi: 10.1109/CVPR.2015.7298935]

[3] Wu Q, Teney D, Wang P, Shen CH, Dick A, van den Hengel A. Visual question answering: A survey of methods and datasets. Computer Vision and Image Understanding, 2017, 163: 21–40. [doi: 10.1016/j.cviu.2017.05.001]

[4] Wang KY, He R, Wang L, Wang W, Tan TN. Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2010–2023. [doi: 10.1109/TPAMI.2015.2505311]

[5] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Communications of the ACM, 2017, 60(6): 84–90. [doi: 10.1145/3065386]

[6] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[7] Wieczorek M, Rychalska B, Dabrowski J. On the unreasonable effectiveness of centroids in image retrieval. In: Proc. of the 28th Int'l Conf. on Neural Information Processing (NIPS). Sanur: Springer, 2021. 212–223. [doi: 10.1007/978-3-030-92273-3_18]

[8] Fu DP, Chen DD, Bao JM, Yang H, Yuan L, Zhang L, Li HQ, Chen D. Unsupervised pre-training for person re-identification. In: Proc. of

the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 14745–14754. [doi: 10.1109/CVPR46437.2021.01451]

[9] Wang GC, Lai JH, Huang PG, Xie XH. Spatial-temporal person re-identification. Proc. of the 2019 AAAI Conf. on Artificial Intelligence, 2019, 33(1): 8933–8940. [doi: 10.1609/aaai.v33i01.33018933]

[10] Zhu ZH, Jiang XY, Zheng F, Guo XW, Huang FY, Sun X, Zheng WS. Viewpoint-aware loss with angular regularization for person re-identification. Proc. of the 2020 AAAI Conf. on Artificial Intelligence, 2020, 34(7): 13114–13121. [doi: 10.1609/aaai.v34i07.7014]

[11] Masson H, Bhuiyan A, Nguyen-Meidine LT, Javan M, Siva P, Ben Ayed I, Granger E. A survey of pruning methods for efficient person re-identification across domains. arXiv:1907.02547, 2021.

[12] Wu L, Wang Y, Gao JB, Wang M, Zha ZJ, Tao DC. Deep coattention-based comparator for relative representation learning in person re-identification. IEEE Trans. on Neural Networks and Learning Systems, 2021, 32(2): 722–735. [doi: 10.1109/TNNLS.2020.2979190]

[13] Yang F, Yan K, Lu SJ, Jia HZ, Xie XD, Gao W. Attention driven person re-identification. Pattern Recognition, 2019, 86: 143–155. [doi: 10.1016/j.patcog.2018.08.015]

[14] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Proc. of the 13th European Conf. on Computer Vision. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]

[15] Plummer BA, Wang LW, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 2641–2649. [doi: 10.1109/ICCV.2015.303]

[16] Zheng KC, Liu W, Liu JW, Zha ZJ, Mei T. Hierarchical gumbel attention network for text-based person search. In: Proc. of the 28th ACM Int'l Conf. on Multimedia. Seattle: ACM, 2020. 3441–3449. [doi: 10.1145/3394171.3413864]

[17] Niu K, Huang Y, Ouyang WL, Wang L. Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Trans. on Image Processing, 2020, 29: 5542–5556. [doi: 10.1109/TIP.2020.2984883]

[18] Wang Z, Fang ZY, Wang J, Yang YZ. ViTAA: Visual-textual attributes alignment in person search by natural language. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 402–420. [doi: 10.1007/978-3-030-58610-2_24]

[19] Aggarwal S, Babu RV, Chakraborty A. Text-based person search via attribute-aided matching. In: Proc. of the 2020 IEEE Winter Conf. on Applications of Computer Vision (WACV). Snowmass: IEEE, 2020. 2606–2614. [doi: 10.1109/WACV45572.2020.9093640]

[20] Zheng ZD, Zheng L, Garrett M, Yang Y, Xu ML, Shen YD. Dual-path convolutional image-text embeddings with instance loss. ACM Trans. on Multimedia Computing, Communications, and Applications, 2020, 16(2): 51. [doi: 10.1145/3383184]

[21] Chen YH, Zhang GQ, Lu YJ, Wang ZX, Zheng YH. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. Neurocomputing, 2022, 494: 171–181. [doi: 10.1016/j.neucom.2022.04.081]

[22] Wang JY, Zhu XT, Gong SG, Li W. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 2275–2284. [doi: 10.1109/CVPR.2018.00242]

[23] Chen DP, Li HS, Liu XH, Shen YT, Shao J, Yuan ZJ, Wang XG. Improving deep visual representation for person re-identification by global and local image-language association. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 56–73. [doi: 10.1007/978-3-030-01270-0_4]

[24] Ren MY, Kiros R, Zemel RS. Exploring models and data for image question answering. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 2953–2961.

[25] Noh H, Seo PH, Han B. Image question answering using convolutional neural network with dynamic parameter prediction. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 30–38. [doi: 10.1109/CVPR.2016.11]

[26] Yang ZC, He XD, Gao JF, Deng L, Smola A. Stacked attention networks for image question answering. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 21–29. [doi: 10.1109/CVPR.2016.10]

[27] Saito K, Shin A, Ushiku Y, Harada T. DualNet: Domain-invariant network for visual question answering. In: Proc. of the 2017 IEEE Int'l Conf. on Multimedia and Expo (ICME). Hong Kong: IEEE, 2017. 829–834. [doi: 10.1109/ICME.2017.8019436]

[28] Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv:1606.01847, 2016.

[29] Jing Y, Si CY, Wang JB, Wang W, Wang L, Tan TN. Pose-guided multi-granularity attention network for text-based person search. Proc. of the 2020 AAAI Conf. on Artificial Intelligence, 2020, 34(7): 11189–11196. [doi: 10.1609/aaai.v34i07.6777]

[30] Yi D, Lei Z, Liao SC, Li SZ. Deep metric learning for person re-identification. In: Proc. of the 22nd Int'l Conf. on Pattern Recognition. Stockholm: IEEE, 2014. 34–39. [doi: 10.1109/ICPR.2014.16]

[31] Hou RB, Ma BP, Chang H, Gu XQ, Shan SG, Chen XL. Interaction-and-aggregation network for person re-identification. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 9309–9318. [doi: 10.1109/CVPR.2019.00954]

[32] Xia BN, Gong Y, Zhang YZ, Poellabauer C. Second-order non-local attention networks for person re-identification. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 3759–3768. [doi: 10.1109/ICCV.2019.00386]

[33] Ye M, Shen JB, Lin GJ, Xiang T, Shao L, Hoi SCH. Deep learning for person re-identification: A survey and outlook. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872–2893. [doi: 10.1109/TPAMI.2021.3054775]

[34] Wu L, Shen CH, van den Hengel A. PersonNet: Person re-identification with deep convolutional neural networks. arXiv:1601.07255, 2016.

[35] Li W, Zhu XT, Gong SG. Harmonious attention network for person re-identification. arXiv:1802.08122, 2018.

[36] Wang C, Zhang Q, Huang C, Liu WY, Wang XG. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 384–400. [doi: 10.1007/978-3-030-01225-0_23]

[37] Chen GY, Lin CZ, Ren LL, Lu JW, Zhou J. Self-critical attention learning for person re-identification. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 96376–9645. [doi: 10.1109/ICCV.2019.00973]

[38] Cheng D, Gong YH, Zhou SP, Wang JJ, Zhang NN. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 1335–1344. [doi: 10.1109/CVPR.2016.149]

[39] Li DW, Chen XT, Zhang Z, Huang KQ. Learning deep context-aware features over body and latent parts for person re-identification. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 7398–7407. [doi: 10.1109/CVPR.2017.782]

[40] Zhao LM, Li X, Zhuang YT, Wang JD. Deeply-learned part-aligned representations for person re-identification. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 3239–3248. [doi: 10.1109/ICCV.2017.349]

[41] Liu JL, Sun YF, Zhu F, Pei HB, Yang Y, Li WH. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 19344–19353. [doi: 10.1109/CVPR52688.2022.01876]

[42] Yu R, Du DW, LaLonde R, Davila D, Funk C, Hoogs A, Clipp B. Cascade transformers for end-to-end person search. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 7257–7266. [doi: 10.1109/CVPR52688.2022.00712]

[43] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[44] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, 2019.

[45] Gao CY, Cai GY, Jiang XY, Zheng F, Zhang J, Gong YF, Peng P, Guo XW, Sun X. Contextual non-local alignment over full-scale representation for text-based person search. arXiv:2101.03036, 2021.

[46] Farooq A, Awais M, Kittler J, Khalid SS. AXM-Net: Implicit cross-modal feature alignment for person re-identification. Proc. of the 2022 AAAI Conf. on Artificial Intelligence, 2022, 36(4): 4477–4485. [doi: 10.1609/aaai.v36i4.20370]

[47] Sun YF, Zheng L, Yang Y, Tian Q, Wang SJ. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 501–518. [doi: 10.1007/978-3-030-01225-0_30]

[48] Zhang Y, Lu HC. Deep cross-modal projection learning for image-text matching. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 707–723. [doi: 10.1007/978-3-030-01246-5_42]

[49] Ren SH, Lin JY, Zhao GX, Men R, Yang A, Zhou JR, Sun X, Yang HX. Learning relation alignment for calibrated cross-modal retrieval. arXiv:2105.13868, 2021.

[50] Xiao T, Li S, Wang BC, Lin L, Wang XG. End-to-end deep learning for person search. arXiv:1604.01850, 2017.

[51] Cho YJ, Yoon KJ. PaMM: Pose-aware multi-shot matching for improving person re-identification. IEEE Trans. on Image Processing, 2018, 27(8): 3739–3752. [doi: 10.1109/TIP.2018.2815840]

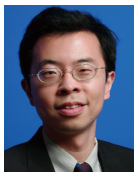
[52] Li W, Zhao R, Wang XG. Human reidentification with transferred metric learning. In: Proc. of the 11th Asian Conf. on Computer Vision. Daejeon: Springer, 2012. 31–44. [doi: 10.1007/978-3-642-37331-2_3]

[53] Reed S, Akata Z, Lee H, Schiele B. Learning deep representations of fine-grained visual descriptions. In: Proc. of the 2016 IEEE Conf. on

Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016. 49–58. [doi: 10.1109/CVPR.2016.13]

[54] Zhu AC, Wang ZJ, Li YF, Wan XL, Jin J, Wang T, Hu FQ, Hua G. DSSL: Deep surroundings-person separation learning for text-based person retrieval. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. ACM, 2021. 209–217. [doi: 10.1145/3474085.3475369]

**李博**(1997－), 男, 硕士生, 主要研究领域为多媒体分析.



**徐常胜**(1969－), 男, 博士, 研究员, 博士生导师, CCF 杰出会员, 主要研究领域为多媒体分析, 计算机视觉, 模式识别, 图像处理.

**张飞飞**(1989－), 女, 博士, 教授, CCF 专业会员, 主要研究领域为多媒体分析, 计算机视觉, 模式识别, 图像处理.