软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2023,34(5):2035−2050 [doi: 10.13328/j.cnki.jos.006766]

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

# 基于虚拟属性学习的文本-图像行人检索方法*

王成济[1], 苏家威[1], 罗志明[1], 曹冬林[1], 林耀进[2,3], 李绍滋[1,3]

[1](厦门大学 信息学院, 福建 厦门 361005)

[2](闽南师范大学 计算机学院, 福建 漳州 363000)

[3](数据科学与智能应用重点实验室 (闽南师范大学), 福建 漳州 363000)

通信作者: 曹冬林, E-mail: another@xmu.edu.cn; 李绍滋, E-mail: szlig@xmu.edu.cn

**摘　要**: 文本-图像行人检索旨在从行人数据库中查找符合特定文本描述的行人图像. 近年来受到学术界和工业界的广泛关注. 该任务同时面临两个挑战: 细粒度检索以及图像与文本之间的异构鸿沟. 部分方法提出使用有监督属性学习提取属性相关特征, 在细粒度上关联图像和文本. 然而属性标签难以获取, 导致这类方法在实践中表现不佳. 如何在没有属性标注的情况下提取属性相关特征, 建立细粒度的跨模态语义关联成为亟待解决的关键问题. 为解决这个问题, 融合预训练技术提出基于虚拟属性学习的文本-图像行人检索方法, 通过无监督属性学习建立细粒度的跨模态语义关联. 第一, 基于行人属性的不变性和跨模态语义一致性提出语义引导的属性解耦方法, 所提方法利用行人的身份标签作为监督信号引导模型解耦属性相关特征. 第二, 基于属性之间的关联构建语义图提出基于语义推理的特征学习模块, 所提模块通过图模型在属性之间交换信息增强特征的跨模态识别能力. 在公开的文本-图像行人检索数据集 CUHK-PEDES 和跨模态检索数据集 Flickr30k 上与现有方法进行实验对比, 实验结果表明了所提方法的有效性.

**关键词**: 行人检索; 跨模态; 属性学习; 预训练

**中图法分类号**: TP391

中文引用格式: 王成济, 苏家威, 罗志明, 曹冬林, 林耀进, 李绍滋. 基于虚拟属性学习的文本-图像行人检索方法. 软件学报, 2023, 34(5): 2035–2050. http://www.jos.org.cn/1000-9825/6766.htm

英文引用格式: Wang CJ, Su JW, Luo ZM, Cao DL, Lin YJ, Li SZ. Text-based Person Search via Virtual Attribute Learning. Ruan Jian Xue Bao/Journal of Software, 2023, 34(5): 2035–2050 (in Chinese). http://www.jos.org.cn/1000-9825/6766.htm

## Text-based Person Search via Virtual Attribute Learning

WANG Cheng-Ji[1], SU Jia-Wei[1], LUO Zhi-Ming[1], CAO Dong-Lin[1], LIN Yao-Jin[2,3], LI Shao-Zi[1,3]

[1](School of Informatics, Xiamen University, Xiamen 361005, China)

[2](School of Computer Science, Minnan Normal University, Zhangzhou 363000, China)

[3](Key Laboratory of Data Science and Intelligence Application (Minnan Normal University), Zhangzhou 363000, China)

**Abstract**: The text-based person search aims to find the image of the target person conforming to a given text description from a person database, which has attracted the attention of researchers from academia and industry. It faces two challenges: fine-grained retrieval and a heterogeneous gap between images and texts. Some methods propose to use supervised attribute learning to obtain attribute-related features and build fine-grained associations between tests and images. The attribute annotations, however, are hard to obtain, which leads to poor performance of these methods in practice. Determining how to extract attribute-related features without attribute annotations and establish fine-grained and cross-modal semantic associations becomes a key problem to be solved. To address this issue, this study incorporates the pre-training technology and proposes a text-based person search via virtual attribute learning, which builds the cross-modal semantic

* 基金项目: 国家自然科学基金 (61876159, 62076210, 62076116)
本文由“融合预训练技术的多模态学习研究”专题特约编辑宋雪萌副教授、聂礼强教授、申恒涛教授、田奇教授、黄华教授推荐.
收稿时间: 2022-04-12; 修改时间: 2022-05-29; 采用时间: 2022-08-24; jos 在线出版时间: 2022-09-20
CNKI 网络首发时间: 2023-03-17

associations between images and texts at a fine-grained level through unsupervised attribute learning. Specifically, in view of the invariance and cross-modal consistency of pedestrian attributes, a semantics-guided attribute decoupling method is proposed, which utilizes identity labels as the supervision signal to guide the model to decouple attribute-related features. Then, a feature learning module based on semantic reasoning is presented, which utilizes the relations between attributes to construct a semantic graph. This model uses the graph model to exchange information among attributes to enhance the cross-modal identification ability of features. The proposed approach is compared with existing methods on the public text-based person search dataset CUHK-PEDES and cross-modal retrieval dataset Flickr30k, and the experimental results verify the effectiveness of the proposed approach.

**Key words**: person search; cross-modality; attribute learning; pre-training

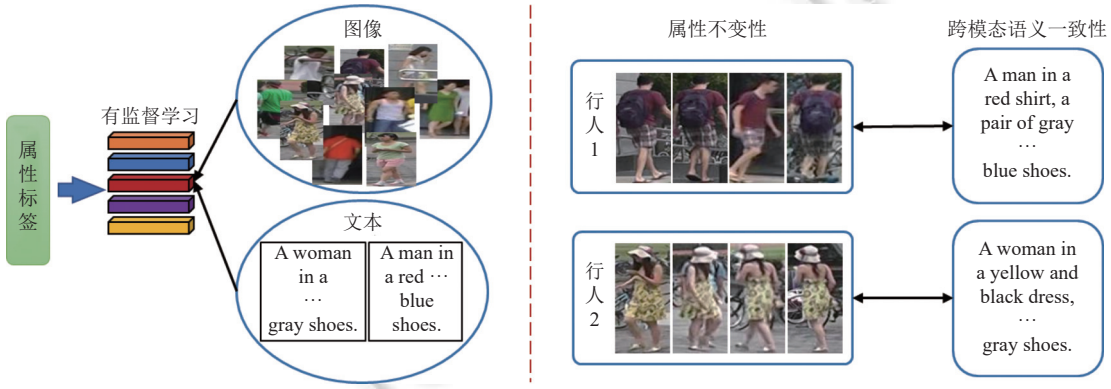
行人检索旨在从图像数据库或视频集中找到特定的行人, 达到跨时空行人跟踪的目的. 自动化地行人检索是智能安防系统的重要组成部分, 广泛应用于安防监控、群体性事件识别等. 行人检索技术是当前计算机视觉领域中具有极高研究和应用价值的前沿方向之一. 当前的行人检索方法可以根据输入数据类型的不同简要地分为两大类: 单模态行人检索[1−3]和跨模态行人检索[4−7]. 与基于素描画[4]或红外图像[5,6]的跨模态行人检索技术相比, 文本数据容易获取、适用场景多, 使得文本-图像跨模态行人检索开始受到业界的广泛关注. 文本-图像跨模态行人检索是跨媒体检索[8]和行人检索[9,10]的交叉研究领域. 如图 1 所示, 给定行人描述, 文本-图像跨模态行人检索旨在根据文本描述从图像数据库中找出目标行人. 对比其他行人检索技术, 文本-图像跨模态行人检索能够更加灵活、全面地满足复杂场景下的行人检索需求. 然而, 图像和文本的抽象层次不同, 不同类型数据间的“异构鸿沟”导致无法直接比较图像和文本的相似性. 同时算法还需要鉴别不同个体的细微差异以区分不同的行人. 这要求算法不仅要克服图像与文本间的“异构鸿沟”也要对比行人的局部细节. 因此, 文本-图像跨模态行人检索是一个极具挑战性的问题.


图 1 文本-图像跨模态行人检索示例

在大型数据集上预训练的特征提取模型隐式地学习到通用的语义知识. 研究人员提出借助预训练模型的表征能力提取多模态的行人特征表示[11−13]. 文献 [11−13] 使用在 ImageNet 数据集[14]预训练的卷积神经网络模型[15−17]提取视觉模态的行人特征表示. 文献 [13] 进一步引入预训练的基于变压器的双向编码语言模型 (bidirectional encoder representations from Transformers, BERT)[18]提取文本特征表示. 这些方法[11−13]建立一个共同子空间, 将不同类型的数据映射到这个子空间中得到统一表征. 文献 [11−13] 直接使用距离度量方法计算文本与图像之间的相似性实现跨模态行人检索. 这类融合预训练技术的文本-图像跨模态行人检索方法专注于在全局上关联文本和图像, 然而无法应对细粒度给行人检索带来的挑战. 于是, 使用属性关联文本和图像的方法被提出. 典型的方法有 AATE[19]和 CMAAM[20]. AATE[19]提出有监督属性学习获取的属性相关的行人视觉表征可以提高特征的鉴别力. CMAAM[20]则认为应该同时从图像和文本中学习属性相关的行人特征表示. 属性是行人部位、着装、性别等的抽象描述, 编码细粒度的语义信息. 文献 [19,20] 已经通过实验证明属性学习可以有效地建立细粒度的跨模态语义关联. 然而属性标签难以获取、手工标注成本高昂, 致使这类方法在实践中表现不佳. CMAAM[20]提出使用名词短语作为属性标签. 但是, 不同文本对属性的描述存在差异导致获得的属性标签不准确. 认知科学的研究成果表明, 人

脑能够自动地从输入中解耦出关键信息并关联不同输入类型的数据, 从而更加全面的认知外部世界[21]. 因此, 如何通过模拟人脑的认知过程, 在没有属性标注的情况下建立细粒度的跨模态语义关联, 是文本-图像跨模态行人检索亟待解决的关键问题.

事实上, 属性是行人固有的特征具有不变性, 且描述同一行人的不同类型的数据存在天然的语义一致性. 其中, 语义一致性是指属性的模态无关性. 因此, 利用属性的不变性和跨模态语义一致性解耦行人的属性信息, 可以得到属性相关的特征. 其次, 属性之间存在依存关系, 例如“裙子”一般与“女性”同时出现. 这表明利用属性间的关联可以更好地解耦属性信息. 此外, 单一的属性不足以区分不同的行人. 与之不同的是属性的组合携带大量可用于识别行人身份的信息. 这表明充分地建模属性的全局上下文信息可以获得更加鲁棒的跨模态行人特征表示. 图 2 是有监督属性学习和无监督属性解耦的对比.




(a) 有监督属性学习 (b) 无监督属性解耦

图 2 有监督属性学习与无监督属性解耦

然而, 现有的基于属性建立细粒度跨模态语义关联的方法依赖于属性标注, 这些方法利用有监督属性学习引导模型提取属性相关的特征. 但多模态行人数据的属性标签难以获取, 手工标注代价高等问题限制这类方法的应用. 因此, 如何在没有属性标签的情况下自动地解耦属性相关特征, 建立细粒度的跨模态语义关联. 本文提出基于虚拟属性学习的文本-图像行人检索方法, 利用行人的身份标签进行虚拟属性学习引导模型解耦属性相关特征, 通过虚拟属性建立细粒度的跨模态语义关联.

本文的主要贡献如下.

(1) 提出基于虚拟属性学习的文本-图像行人检索方法. 该方法无需标注属性标签就可以完成细粒度的跨模态关联分析与挖掘, 摆脱对属性标注的依赖.

(2) 针对没有属性标签无法提取到属性相关特征的问题, 提出语义引导的属性解耦方法. 该方法使用行人的身份标签作为监督信号引导模型解耦属性相关特征. 无需属性标签就可以充分挖掘多样的属性信息, 从而建立细粒度的跨模态语义关联.

(3) 提出基于语义推理的特征学习模块, 利用属性间的共现关系构建语义图. 一方面, 通过语义图在属性之间交换信息, 补全缺失的属性信息; 另一方面, 多层图模型可以充分挖掘属性的上下文信息, 从而提升特征的识别能力.

通过在 CUHK-PEDES 数据集上的实验设计和分析, 验证了所提出方法的有效性. 在通用跨模态检索数据集 Flickr30k 上的实验证明了所提出方法的泛化性. 本文第 1 节介绍相关工作. 第 2 节介绍提出的基于虚拟属性学习的文本-图像行人检索方法. 第 3 节为实验设置与结果分析. 第 4 节为本文结论.

## 1 相关工作

### 1.1 文本-图像跨模态行人检索

当前的文本-图像跨模态行人检索方法可以简要分为 3 类: 基于跨模态交互的方法、基于联合特征映射的方

法和基于属性学习的方法.

基于跨模态交互的方法旨在通过文本与图像的跨模态交互计算文本与图像的相似度. Li 等人[7]结合卷积神经网络 (convolutional neural network, CNN) 和递归神经网络 (RNN) 提出一个 CNN-RNN 模型. 该模型首先提取图像的全局特征, 分别计算图像特征与不同词语的相似度, 聚合图像与词语的相似度得到文本与图像的相似度. Li 等人[22]在 CNN-RNN 模型的基础上提出使用行人身份标签挖掘难分样本以增加跨模态特征的相似性和不同人特征的差异. Chen 等人[23]提出将图片均匀划分为 49 个子区域并计算每一个图像区域和不同词语的相关性, 构建文本与图像的细粒度关联. 通过对相关性取最大值得到文本与图像的相似度. Niu 等人[24]将图像纵向切分为 6 个子区域并分别计算图像区域与名词短语的相关性并根据相关性聚合图像和文本的特征得到细粒度的跨模态行人特征表示. Niu 等人[24]同时计算细粒度的和全局的图文相似度. Gao 等人[25]基于跨模态注意力设计了一个文本引导的视觉特征增强模块, 建模不同图像区域与文本的关联. Jing 等人[26]使用一个模态的局部特征引导并聚合另一个模态的局部特征, 得到双向对齐的跨模态行人局部特征表示. 基于跨模态交互的方法强制对齐所有的图像区域和文本, 这类方法认为细粒度匹配问题是解决文本-图像跨模态行人检索的关键. 然而文本只能描述有限的图像内容, 致使这类方法学习到错误的局部匹配导致错误的检索结果.

基于联合特征映射的方法将两个模态的数据映射到同一空间中. 在共享的特征空间中进行特征学习和跨模态匹配. Zhang 等人[11]分别使用卷积神经网络和双向长短记忆神经网络提取图像和文本特征表示, 提出基于跨模态交叉投影的跨模态匹配损失和身份分类损失. Zheng 等人[12]使用两个卷积神经网络分别提取图像和文本特征表示. 文献 [13,27] 提出使用对抗学习减少模态差异. 也有一些文献进一步引入局部特征学习. Jing 等人[28]认为使用图注意力建模名词短语间的关系可以提升特征的鉴别力. Liu 等人[29]使用预训练的目标检测模型获取物体的特征表示后通过图模型聚合物体的特征表示得到行人的全局特征表示. 基于联合特征映射的方法致力于从全局上建立文本和图像的跨模态语义关联, 这类方法没有建立细粒度的跨模态语义关联导致提取的特征较为粗糙, 限制了模型性能.

基于属性学习的方法利用属性的跨模态语义一致性提取跨模态的细粒度语义表征, 从而建立细粒度的跨模态语义关联, 提升模型的检索精度. 具有代表性的方法有 Aggarwal 等人[20]提出的 CMAAM 和 Wang 等人[30]提出的 ViTAA. Aggarwal 等人[20]对整个数据集中的名词短语和名词进行聚类, 将类别中心作为属性标签进行有监督属性学习提取跨模态的属性相关特征. Wang 等人[30]提出使用预训练的行人语义分割模型分割出行人不同的部位, 通过对齐行人的部位特征和名词短语获得跨模态的行人特征表示. Wang 等人[30]认为分割得到的行人部位特征代表行人的属性特征, 对齐行人的部位特征和名词短语的特征可以得到细粒度的行人特征表示. 基于属性学习的方法对属性标签或者预训练的属性相关特征提取模型的依赖导致该方法在实践中表现不佳.

基于属性学习的方法融合了基于跨模态交互的方法和基于联合特征映射的方法的优点, 能够在细粒度上关联图像和文本同时具有较快的推理速度. 针对没有属性标签无法提取属性相关特征的问题, 提出基于虚拟属性学习的文本-图像行人检索方法. 与之前的方法不同, 本文所提出的方法无需属性标签和预训练的属性相关特征提取模型. 所提出的语义引导的属性解耦方法可以自动地从文本和图像中解耦出属性相关特征, 在细粒度上建立跨模态语义关联.

### 1.2 行人属性学习及其在行人检索中的应用

属性是结构化的行人语义特征描述着装、发型、携带物体等诸多可识别的信息, 其对光照变化和视角变化都具有鲁棒性. 史金婉等人[31]认为属性编码的个性化的信息, 可以用于服装推荐. 郑鑫等人[32]在单模态行人检索中引入属性标签学习具有鉴别力的行人部位特征. 文献 [33] 认为可以使用属性检索目标行人图像. 与文本-图像跨模态行人检索一样, 使用属性检索行人图像是跨模态行人检索任务. 然而属性不具有唯一性, 使用属性作为输入可能会检索出多个表观相似的行人.

文献 [19,20] 引入有监督属性学习解决文本-图像跨模态行人检索问题. Zha 等人[19]使用手工标注的属性学习属性相关的特征, 并融合属性相关的特征和图像特征获取更具有鉴别力的跨模态行人特征表示. Zha 等人[19]只提

取图像中的属性相关特征表示, 没有提取文本的属性相关特征表示. Aggarwal 等人[20]提出使用数据集中的名词和名词短语作为属性标签. 作者对数据集中的名词和名词短语进行聚类, 将类别中心作为属性标签. 由于自然语言的多样性和复杂性, 使用聚类的方法获得的属性标签包含有大量噪声.

本文方法利用属性在细粒度上关联图像和文本提升特征的鉴别能力. 与上述方法不同, 本文方法基于属性的不变性和跨模态语义一致性, 在没有使用属性标签的前提下自动地解耦属性信息获取跨模态的属性相关特征, 提升模型检索精度.

## 2 基于虚拟属性学习的文本-图像行人检索

属性是与模态无关的细粒度语义信息. 提取行人属性相关特征能有效地建立细粒度跨模态语义关联, 提升特征的跨模态识别能力. 为摆脱对属性标签的依赖, 提出基于虚拟属性学习的文本-图像行人检索方法, 所提出方法的网络结构如图 3 所示. 针对两种模态的输入, 本文使用两个预训练模型分别提取图像和文本的特征表示. 语义引导的属性解耦使用双线性注意力 (bilinear attention, BA)[34]计算图像或文本的局部特征与不同属性的语义相关性. 根据相关性聚合特征得到属性相关的特征表示. 基于语义推理的特征学习将属性作为图的节点、属性的共现概率作为边, 使用图神经网络 (graph neural network, GNN)[35]建模属性的全局上下文, 获得具有鉴别力的跨模态行人特征表示.

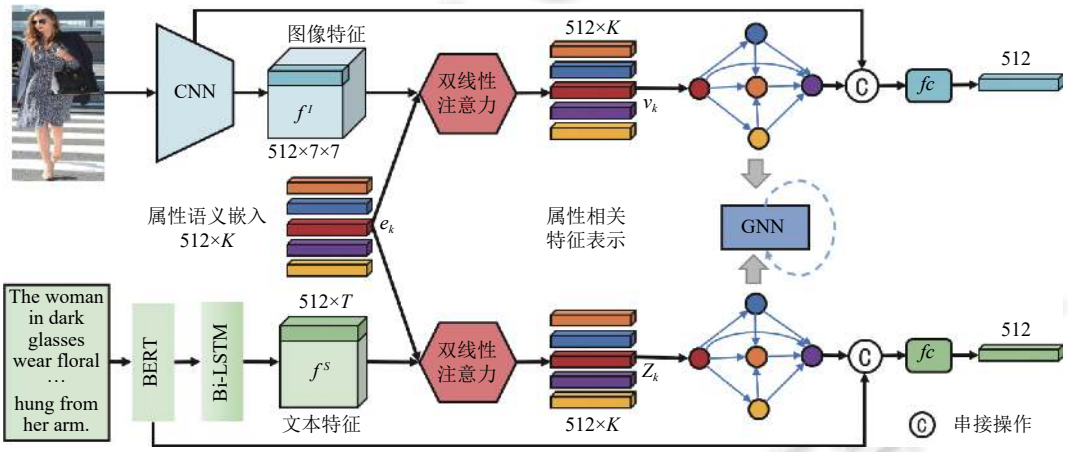



图 3 本文方法整体框架示意图

下面给出文本-图像跨模态行人检索的形式化定义. 假定数据集为 $D=\{X,S,Y\}$, 其中 $X=\{(x_i,y_i)\}_{i=1}^{N}$ 和 $S=\{(s_i,y_i)\}_{i=1}^{N}$ 分别表示图像和文本数据所构成的集合, $y_i$ 表示对应样本的身份标签. 文本是不定长序列, 用 $T$ 表示文本长度. 假定行人有 $K$ 个属性, 属性的语义嵌入表示为 $\{e_k\}_{k=1}^{K}$. 文本-图像跨模态行人检索旨在给定一条文本描述返回其所描述的行人的图像. 在训练时, 以成对的文本和图像 $(x,s,y)$ 作为输入.

### 2.1 模态相关的特征提取

如图 3 所示, 给定图像-文本对 $(x,s)$. 针对两种模态的输入分别使用两种预训练神经网络 (详细的预训练神经网络结构和抽取的特征维度见表 1) 提取模态相关的特征. 对于图像 $x$, 使用预训练的卷积神经网络 (CNN) 提取特征, 紧接着使用一个 1×1 的卷积层将图像特征映射到 $d$ 维空间中. 提取的图像特征表示为 $f^I\in\Re^{d\times W\times H}$, 其中, 特征图的大小为 $W\times H$. 对于文本 $s$, 使用 BERT 提取每个词的特征表示, 使用双向长短记忆神经网络 (bi-directional long short-term memory networks, Bi-LSTM) 建模词的上下文关系, 对 Bi-LSTM 的输出取平均获得融合上下文信息的文本特征表示, 一个全连接层将文本特征映射到 $d$ 维空间中. 提取的文本特征表示为 $f^S\in\Re^{d\times N}$, 其中, $N$ 是文本长度.

表 1 预训练神经网络结构

| 模态 | 网络 | 基本结构 | 网络深度 | 特征维度 |
|---|---|---|---|---|
| 图像 | VGG[15] | 卷积层 | 16 | 512 |
| | MobileNet[17] | 深度可分离卷积层 | 28 | 1 024 |
| | ResNet[16] | 卷积层、残差层 | 50 | 2 048 |
| 文本 | BERT[18] | 自注意力层 | 12 | 768 |

### 2.2 语义引导的属性解耦

语义引导的属性解耦模块旨在使用属性语义嵌入引导模型提取属性相关的特征表示. 属性语义嵌入编码了属性信息. 使用与模态无关的属性语义嵌入引导模型学习, 可以使得模型更加关注属性相关的图像区域或词语, 学习到跨模态的属性相关特征表示. 因此, 如何合理利用属性的不变性和跨模态语义一致性自动地解耦属性信息成为无监督属性解耦的关键问题. 受属性的跨模态一致性和行人属性的不变性的启发, 提出使用行人的身份标签作为监督信号, 引导模型进行属性解耦.

对于每一个图像区域 $(w,h)$, 使用如下双线性注意力融合图像特征和属性 $k$ 的语义嵌入, 得到融合属性信息的图像特征 $\widetilde{f}^I_{k,wh}$:

$$\widetilde{f}^I_{k,wh} = P^{\mathrm{T}}\left(\tanh\left(\left(U^{\mathrm{T}} f^I_{wh}\right)\odot\left(V^{\mathrm{T}} e_k\right)\right)\right) \tag{1}$$

其中, $\tanh(\cdot)$ 是双曲正切函数, $U\in\mathfrak{R}^{d\times d}$, $V\in\mathfrak{R}^{d\times d}$, $P\in\mathfrak{R}^{d\times 1}$ 是可学习的参数, $\odot$ 表示点乘. 之后, 使用归一化指数函数得到不同图像区域与属性 $k$ 的相关性, 图像区域 $(w,h)$ 与属性 $k$ 的相关性表示如下:

$$a_{k,wh} = Softmax\left(\widetilde{f}^I_{k,wh}\right) \tag{2}$$

最后, 使用加权求和的方式聚合所有位置的图像特征, 得到属性相关的图像特征表示. 与属性 $k$ 相关的图像特征表示如下:

$$v_k = \sum_{w,h} a_{k,wh} f^I_{wh} \tag{3}$$

重复上述过程 $K$ 次, 可以得到 $K$ 个属性相关的图像特征向量, 表示为: $\{v_1,v_2,\ldots,v_K\}$. 行人属性表示为: $Q^I=\left\{q^I_k\right\}_{k=1}^K$. 其中 $q^I_k$ 计算如下:

$$q^I_k = I\left(\tanh\left(w^{\mathrm{T}}_k v^k\right)\right) \tag{4}$$

其中, $w_k\in\mathfrak{R}^{d\times 1}$ 是属性分类器; $\tanh(\cdot)$ 是双曲正切函数; $I(\cdot)$ 是指示函数, 当括号内数值为正时输出 1, 其他输出 0. 使用归一化函数得到行人的属性分布, 表示为: $P^I=\left\{p^I_k,k=1,\ldots,K\right\}$, 其中 $p^I_k=q^I_k\Big/\sum_h q^I_h$.

同理, 对于每一个词语 $t$ 的特征 $f^S_t$. 用 $f^S_t$ 替换公式 (1)–(3) 中的 $f^I_{wh}$, 可以得到与属性 $k$ 相关的文本特征, 表示为 $z_k$. 重复上述过程 $K$ 次可以得到 $K$ 个属性相关的文本特征向量, 表示为: $\{z_1,z_2,\ldots,z_K\}$. 同样地, 可以得到属性预测 $Q^S=\left\{q^S_k\right\}_{k=1}^K$ 和行人的属性分布 $P^S$.

在得到输入的属性预测和行人的属性分布之后, 根据行人的属性不变性设计中心点损失函数, 给每一个行人提供一个属性分布中心, 使得同一行人不同样本的属性分布尽可能地靠近其属性分布中心缩小同一行人不同样本间的属性分布的距离. 采用如下损失函数约束同一行人不同样本的属性分布的距离:

$$L_c = \left\|P^I - c_y\right\|_2^2 + \left\|P^S - c_y\right\|_2^2 + \frac{\eta}{M}\sum_{h\neq y} c_y^{\mathrm{T}} c_h \tag{5}$$

其中, $y$ 是输入图像-文本对的身份标签. $c_y\in\mathfrak{R}^K$ 是可学习的特征向量, 表示行人 $y$ 的属性分布中心. 数据集中有 $M$ 个不同身份的行人. 超参数 $\eta=0.0005$. 公式 (5) 不仅约束同一行人属性分布的距离, 也要求不同行人属性分布尽可能地不相似.

本文设计基于属性的跨模态匹配损失限制身份标签相同的图像和文本的属性分布的距离小于身份标签不同

的图像和文本的距离. 属性分布的距离定义如下:

$$d\left(P^I,P^S\right)=1-\frac{\left\|P^I-P^S\right\|_1}{2} \tag{6}$$

基于属性的跨模态匹配损失定义如下:

$$L_m=\begin{cases}\left[0,\alpha-d\left(P^I,P^S\right)\right]_+\\ \left[0,d\left(P^I,P_n^S\right)-(1-\alpha)\right]_+\end{cases} \tag{7}$$

其中, $P_n^S$是不与图像$x$匹配的文本, $\alpha$限制正负样本对间的距离. 本文使用随机采样的方法生成负样本.

为保证模型学习到多样化的属性相关特征提出的基于对比学习的属性解耦损失定义如下:

$$L_d=-\frac{1}{K}\sum_{k=1}^{K}q_k^I\log\left(\frac{\exp(w_kv_k)}{\sum_h\exp(w_hv_k)}\right)+q_k^S\log\left(\frac{\exp(w_kz_k)}{\sum_h\exp(w_hz_k)}\right) \tag{8}$$

最终的属性解耦损失函数定义如下:

$$L_{att}=L_c+L_m+L_d \tag{9}$$

### 2.3 基于语义推理的特征学习

属性之间存在语义关联, 单一属性缺少上下文信息不足以区分不同的行人. 对属性之间的关联进行建模可以很好地挖掘属性的上下文信息, 提升模型的跨模态表征能力. 基于语义推理的特征学习以属性相关特征为节点、属性之间的共现概率为边, 使用图神经网络 (GNN) 构造语义图, 基于图模型在属性之间交换信息, 对属性的全局上下文信息进行建模提升特征的跨模态识别能力.

首先, 构建语义图$G=\{H,A\}$. 其中, 节点$H=\{h_1,h_2,\ldots,h_K\}$是属性相关特征, $E=\{e_{11},e_{12},\ldots,e_{1K},\ldots,e_{KK}\}$是统计的属性共现概率, 其中$e_{kh}$表示属性$k$与属性$h$同时出现的概率. $E$初始化为0. 累积每一次迭代获得的属性的共现概率将其求和作为图模型的边. 在$t$时刻, 图中节点的状态$H^t=\left\{h_1^t,h_2^t,\ldots,h_K^t\right\}$依赖于$t-1$时刻节点的状态. 根据属性$k$与其他属性的共现概率聚合属性的特征, 得到特征$a_k^t$; 其次, 使用门控注意力融合特征$a_k^t$和特征$h_h^{t-1}$更新节点的状态, 得到$t$时刻节点$k$的状态$h_k^t$. 详细计算过程表示如下:

$$\begin{cases}a_k^t=\sum_h e_{kh}h_h^{t-1}\\ z_k^t=\sigma(W^za_k^t+U^zh_k^{t-1})\\ r_k^t=\sigma(W^ra_k^t+U^rh_k^{t-1})\\ \widetilde{h}_k^t=\tanh\left(Wa_k^t+U\left(r_k^t\odot h_k^{t-1}\right)\right)\\ h_k^t=\left(1-z_k^t\right)\odot h_k^{t-1}+z_k^t\odot\widetilde{h}_k^t\end{cases} \tag{10}$$

其中, $\sigma(\cdot)$是Sigmoid函数, $\tanh(\cdot)$是双曲正切函数, $\odot$代表点乘. $a_k^t$聚合所有节点的信息, $z_k^t$和$r_k^t$决定是否要使用该信息更新当前节点的状态. 串接所有节点的特征并将其投影到$d$维空间中, $t$时刻获得的语义增强的特征表示如下:

$$o^t=fc([h_1^t\|\ldots\|h_k^t]) \tag{11}$$

其中, $fc$代表输出特征维度为$d$的全连接层, $[\cdot\|\cdot]$代表按通道串接特征. 第0时刻, 使用属性相关的图像特征$\{v_1,v_2,\ldots,v_K\}$和文本特征$\{z_1,z_2,\ldots,z_K\}$初始化图的节点. 本文的图模型有3个特征提取层. 将图像和文本特征输入语义图模型, 可以分别得到3个语义增强的图像特征$\left\{o^{1,I},o^{2,I},o^{3,I}\right\}$和3个语义增强的文本特征$\left\{o^{1,S},o^{2,S},o^{3,S}\right\}$.

为使得语义增强的图像和文本特征更好地保留行人的身份信息, 可以最大化特征与行人身份的相关性. 计算每个特征层输出的特征$o^t$与行人$y$的相关性表示为概率$p(y|o^t)$, 计算过程如下:

$$p(y|o^t)=\frac{\exp\left(w_y^{\mathrm{T}}o^t\right)}{\sum_{m=1}^{M}\exp\left(w_m^{\mathrm{T}}o^t\right)} \tag{12}$$

其中, $w_m$是行人$m$的身份分类器, 训练集种共有$M$个行人身份. 使用交叉熵函数计算身份分类损失. 基于语义推

理的特征学习损失函数表示如下:

$$L_{asr}=-\sum_{t=1}^{3}\sum\nolimits_{o^t\in\{o^{t,I},o^{t,S}\}}\log\left(p\left(y|o^t\right)\right) \tag{13}$$

### 2.4 全局特征学习

语义增强的特征可以提取丰富的细粒度语义信息, 但是仍缺少行人的全局信息. 图像和文本的全局特征分别编码图像和文本的空间分布, 携带具有鉴别力的全局信息. 将全局特征与语义增强的特征融合, 可以使得模型获得更加具有鉴别力的跨模态行人特征表示. 两个模态的行人特征表示如下:

$$\begin{cases}F^I=fc^I\left(\left[avg\_pool(f^I)\|o^{3,I}\right]\right)\\F^S=fc^S\left(\left[max\_pool(f^S)\|o^{3,S}\right]\right)\end{cases} \tag{14}$$

其中, $avg\_pool(\cdot)$ 代表全局平均池化, $max\_pool(\cdot)$ 代表全局最大池化, $[\cdot\|\cdot]$ 代表按通道串接特征, $fc^I$ 和 $fc^S$ 代表输出特征维度为 512 的全连接层.

本文使用如下三元组损失训练模型:

$$L_r=\left[\beta-\cos(F^I,F^S)+\cos(F^I,H^S)\right]_+ +\left[\beta-\cos(F^I,F^S)+\cos(H^I,F^S)\right]_+ \tag{15}$$

其中, $H^I$ 和 $H^S$ 是难分负样本, $\beta=0.2$ 限制正样本和负样本间的距离, $\cos(\cdot,\cdot)$ 代表余弦函数.

综合公式 (9)、公式 (13) 和公式 (15), 总体目标函数表示如下:

$$L=L_r+\lambda_1L_{att}+\lambda_2L_{asr} \tag{16}$$

其中, $\lambda_1$ 和 $\lambda_2$ 是损失函数的权重. 本文使用多任务学习, 同时优化 3 个损失函数.

## 3 实验结果与分析

本节先说明实验设置. 然后, 在公开数据集上与多种现有方法进行对比说明本文方法的性能. 之后, 通过销蚀实验分析本文方法各个部分的作用. 最后, 通过实验分析各个损失函数的重要性.

### 3.1 实验设置

● 数据集. 为验证本文提出方法的有效性, 我们在当前公开的大型文本-图像跨模态行人检索数据集 CUHK-PEDES[7]和跨模态检索数据集 Flickr30k[36]上进行实验. CUHK-PEDES 数据集包含 130 003 个行人身份, 总共有 40 206 张行人图像, 每张图像有两条文本描述. CUHK-PEDES 数据集分为训练集、验证集和测试集, 3 个子集中行人的身份互不重叠. 训练集中有 11 003 个行人身份、验证集和测试集各有 1 000 个行人身份. Flickr30k 数据集有 31 783 张图片, 每张图片标注 5 条文本描述, 其中 29 783 张图片用于训练、验证集和测试集各有 1 000 张图片. 两个数据集的训练集、验证集和测试集的详细划分如表 2 所示.

表 2 数据集划分

| 数据集 | CUHK-PEDES | | | Flickr30k | |
|---|---|---|---|---|---|
| | 行人身份 | 图片 | 文本描述 | 图片 | 文本描述 |
| 训练集 | 11 003 | 34 054 | 68 126 | 29 783 | 148 915 |
| 验证集 | 1 000 | 3 078 | 6 156 | 1 000 | 5 000 |
| 测试集 | 1 000 | 3 074 | 6 148 | 1 000 | 5 000 |

● 评价指标. 我们采用累计匹配特性 (cumulative matching characteristic, CMC) 评价模型的好坏. CMC 值统计的是目标图像出现在前 $K$ 个排序的检索结果中的概率, 也被叫作前 $K$ 位命中率. 以 CMC-Rank-1 为例, 如果检索出的得分最高的图像是目标图像, 则 CMC-Rank-1=1, 否则 CMC-Rank-1=0. 通常使用的评价指标是 CMC-Rank-1、CMC-Rank-5 和 CMC-Rank-10, 可简写为 Rank-1、Rank-5 和 Rank-10.

● 实现细节. 本文使用 TensorFlow 实现图 3 所示的模型, 所有实验在配备有 Intel Core i7-7700K CPU,

GeForce GTX 1080Ti 显卡, 64 位 Ubuntu 16.04 系统, 32 GB 内存的工作站上进行. 使用的预训练 BERT 模型 (bert-as-service, https://bert-as-service.readthedocs.io) 提取文本特征. 输入图像的大小为 224×224. 使用 Adam 优化器优化模型, 动量和衰减率分别设置为 0.5 和 0.000 5, 初始学习率设置为 0.000 2. 对于 CUHK-PEDES 数据集[7], 使用在 ImageNet[14]上预训练的图像分类模型提取图像特征, 每次迭代输入 32 个匹配的图像-文本对, 遍历训练集 50 次. 对于 Flickr30k 数据集[36], 使用预训练的 ResNet-152 模型[16]提取图像特征. 在训练过程中, 先固定主干网络遍历数据集 20 次, 再训练整个模型遍历数据集 15 次, 每次迭代输入 128 个匹配的图像-文本对. 设置 $\alpha = 0.5$, 属性数目 $K = 12$, 特征维度 $d = 512$.

● 测试设置. 在测试阶段, 分别提取图像特征 $F^I$ 和文本特征 $F^S$. 给定一条文本描述的特征 $F^S$, 使用余弦函数计算其与所有图像特征的相似度得分并根据得分对图像进行排序并报告排序结果.

### 3.2 与现有方法的实验结果对比

#### 3.2.1 在 CUHK-PEDES 数据集上的行人检索结果

表 3 中对比本文方法与其他方法在 CUHK-PEDES 数据集上的检索结果. 预训练列出了方法所使用的预训练图像特征提取模型. 属性表示是否提取行人属性相关特征. 文本-图像跨模态行人检索旨在使用文本检索行人图像, 其他方法没有报告图像检索文本的结果. 与其他方法一样, 本文报告并对比文本检索图像的 Rank-1、Rank-5 和 Rank-10 准确率.

表 3 本文方法与其他方法在 CUHK-PEDES 数据集上的比较结果 (%)

| 方法 | 主干网络 | 预训练 | 属性学习 | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|---|---|
| GNA-RNN (CVPR 2017)[7] | VGG-16 | CNN | × | 19.05 | — | 53.64 |
| IATV (ICCV 2017)[22] | | CNN | × | 25.94 | — | 60.48 |
| PWM-ATH (WACV 2018)[23] | | CNN | × | 27.14 | 49.45 | 61.02 |
| Dual-Path (TOMM 2020)[12] | | CNN | × | 32.15 | 54.42 | 64.30 |
| GLA (ECCV 2018)[37] | | CNN | × | 43.58 | 66.93 | 76.26 |
| GARN (TIP 2021)[28] | | CNN | × | 46.25 | 67.48 | 76.84 |
| PWA (AAAI 2020)[26] | | 姿态估计模型 | × | 47.82 | 69.83 | 78.31 |
| 本文方法 | | CNN | √ | **49.31** | **71.64** | **80.07** |
| CMPC (ECCV 2018)[11] | MobileNet | CNN | × | 49.37 | 71.69 | 79.27 |
| GARN (TIP 2021)[28] | | CNN | × | 52.75 | 74.36 | 81.85 |
| TVFR (ICMR 2021)[25] | | CNN | × | 53.87 | 75.25 | 83.47 |
| CMAAM (WACV 2020)[20] | | CNN | √ | 55.13 | 76.14 | 83.77 |
| 本文方法 | | CNN | √ | **56.17** | **77.05** | **83.74** |
| Dual-Path (TOMM 2020)[12] | ResNet-50 | CNN | × | 44.40 | 66.26 | 75.07 |
| GARN (TIP 2021)[28] | | CNN | × | 52.25 | 73.51 | 81.12 |
| AATE (TMM 2020)[19] | | CNN | √ | 52.42 | 74.98 | 82.74 |
| MIA (TIP 2020)[24] | | CNN | × | 53.10 | 75.00 | 82.90 |
| A-GANet (MM 2019)[29] | | 目标检测模型 | × | 53.14 | 74.03 | 82.95 |
| PWA (AAAI 2020)[26] | | 姿态估计模型 | × | 54.12 | 75.45 | 82.97 |
| CMKA (TIP 2021)[38] | | CNN | × | 54.69 | 73.65 | 81.86 |
| ViTAA (ECCV 2020)[30] | | 语义分割模型 | √ | 55.97 | 75.84 | 83.52 |
| 本文方法 | | CNN | √ | **57.31** | **76.95** | **84.24** |

注: “—”代表原论文没有报告此项结果, “×”代表没有使用该项, “√”代表使用该项

我们可以得到以下的观察结果.

第一, 当使用相同的主干网络时, 本文的方法在 3 个评价指标上都取得最好的结果. 与之前的方法相比, 本文

方法的检索准确率有较大幅度的提升. 首先, 在仅使用预训练的 CNN 模型的方法, 本文方法的实验结果大幅领先于 CMKA[38]. 其次, 对比使用更好的预训练模型的方法 A-GANet (预训练的目标检测模型提取图像中物体的特征)[29]、PWA (预训练的人体姿态估计模型提取行人身体关键点的特征)[26]、ViTAA (预训练的行人语义分割模型提取行人的部位特征)[30], 本文的方法同样取得更好的结果. 这说明本文所提出的方法的优越性. 本文所提出的方法能够减少对属性标注的依赖也可以避免预训练模型带来的噪声干扰.

第二, 与使用有监督属性学习的方法比较. AATE[19]和 CMAAM[20]引入有监督属性学习; ViTAA[30]使用重新训练的行人语义分割模型分割出行人的部位, 将行人的部位特征视为属性特征. 对比 AATE[19]、CMAAM[20]、ViTAA[30]和本文方法, 本文提出的基于无监督属性解耦的方法具有明显优势. 本文方法无需属性标签就能够有效地利用属性的不变性和跨模态一致性解耦属性信息, 减少对属性标签和预训练模型的依赖.

综上所述, 本文方法在文本-图像跨模态行人检索任务上表现优异, 可以: 1) 降低现有方法对属性标注的需求, 2) 避免预训练模型的不确定性带来的干扰. 本文方法改善现有方法在特征鉴别力不足的问题. 实验结果表明, 提出的无监督属性解耦可以有效地弥合图像与文本间的异构鸿沟. 对比其他使用预训练模型的方法, 例如, PWA[26]、A-GANet[29]、ViTAA [30]需要针对不同任务重新训练预训练模型, 本文的方法具有更好的通用性.

表 4 对比本文方法与 ViTAA[30]的检索结果. ViTAA 使用行人语义分割模型提取行人的属性相关特征表示. 分割结果的好坏会直接影响模型的检索结果. 观察表 4 可以发现以下结果.

表 4　本文方法与 ViTAA 模型在 CUHK-PEDES 数据集的检索结果对比

| 方法 | 查询文本 | |
|---|---|---|
| | the man is wearing a black and white striped shirt he is wearing black shorts he has on sandals | the woman is wearing a mid length purple dress with matching pull over she has short brown hair |
| 本文方法 |  |  |
| ViTAA[30] |  |  |

第一, 不准确的预训练分割模型导致错误的检索结果. 第 1 个样例检索结果中的第 2 张和第 4 张图像中的行人没有“shorts”, ViTAA 则根据腿部特征判定“wearing shorts”; 第 5 张图像中的男子的特征不明显, ViTAA 受到背景的干扰判定其“wearing a black and white striped shirt”. 以上结果由于预训练的行人分割模型在原有数据集上过拟合导致分割出的行人部位不准确, 致使 ViTAA 学习到错误的匹配. 同样的情况也能够在第 2 个样例的检索结果中发现. 本文方法仅使用预训练 CNN 模型提取特征. 预训练 CNN 模型具有良好的泛化性. 本文方法能够避免由于预训练模型的不确定性导致的噪声干扰.

第二, ViTAA 直接学习属性的跨模态匹配而忽略属性的全局上下文. 对属性的全局上下文建模可以更加全面和立体的提取行人特征表示. ViTAA 直接学习属性的跨模态匹配. 会导致当某一个属性的特征不明显时, 模型会忽略该属性导致错误的匹配结果. 提出的基于语义推理的特征学习可以充分地建模属性的全局上下文获得更加鲁棒的跨模态行人特征表示.

### 3.2.2　在 Flickr30k 数据集上的检索结果

为说明本文方法的泛化性, 表 5 对比本文方法与表 3 中的方法在 Flickr30k 数据集上的结果, 包括 CMPC[11]、CMKA[38]和 GARN[28]. 表 2 中的其他方法没有报告在 Flickr30k 数据集上的结果. 从表 5 中可以看到, 本文方法在

其中 5 个评价指标上都取得最好的结果. 图像检索文本的 Rank-1 准确率只比 GARN[28]低 0.3%. 表 5 的结果表明本文方法具有良好的泛化性能.

表 5 本文方法与其他方法在 Flickr30k 数据集上的比较结果 (%)

| 方法 | 图像检索文本 | | | 文本检索图像 | | |
|---|---|---|---|---|---|---|
| | Rank-1 | Rank-5 | Rank-10 | Rank-1 | Rank-5 | Rank-10 |
| CMPC (ECCV 2018)[11] | 49.6 | 76.8 | 86.1 | 37.3 | 65.7 | 75.5 |
| Dual-Path (TOMM 2020)[12] | 55.6 | 81.9 | 89.5 | 39.1 | 69.2 | 80.9 |
| CMKA (TIP 2021)[38] | 55.7 | 82.9 | 90.0 | 45.0 | 73.4 | 82.7 |
| GARN (TIP 2021)[28] | **60.1** | 84.6 | 90.4 | 44.2 | 71.2 | 80.3 |
| 基线方法 | 49.2 | 75.8 | 85.7 | 37.5 | 65.4 | 75.2 |
| 本文方法 | 59.8 | **84.7** | **90.6** | **45.3** | **72.1** | **81.0** |

### 3.3 模型销蚀实验分析

第一, 我们通过消减相应的模块分析不同部件 (包括语义引导的属性解耦 ATT、基于语义推理的特征学习 ASR、串接语义增强的特征和全局特征 FF 及预训练语言模型 BERT) 的贡献. 销蚀实验结果见表 6. 基准方法使用全局特征作为跨模态行人特征表示. 模型 2 将属性相关特征串接后映射到低维特征空间. 模型 3 将属性相关特征和全局特征串接后映射到低维特征空间. 模型 4 将语义增强的特征作为最终跨模态行人特征表示. 模型 5 是本文方法. 通过比较这些模型的实验结果, 我们可以得出以下结论: 1) 语义引导的属性解耦模块有效地挖掘细粒度行人特征. 与基准方法对比, 模型使用语义引导的属性解耦后 Rank-1 准确率明显提高. 图 4 展示部分虚拟属性的注意力热图可视化结果, 从图 4 中可以看出模型会自动关注具有鉴别力的行人部位并能有效地对齐图像和文本. 这说明模态无关的虚拟属性的语义嵌入可以引导模型学习到跨模态的细粒度行人特征表示, 有效地建立细粒度的跨模态关联. 2) 基于语义推理的特征学习, 基于属性构建的语义图可以有效地建模属性的全局上下文, 进一步增强特征的跨模态识别能力. 从表 6 中可以看到, 增加基于语义推理的特征学习模块后, 3 个评价指标都有明显提升. 这是因为基于语义推理的特征学习模块在属性间交换信息, 不仅可以基于语义推理补全缺失的属性信息, 还充分考虑到属性与全局语义的关联. 3) 全局特征和语义增强的特征是互补的. 语义增强的特征充分挖掘细粒度的行人语义特征. 全局特征提取了输入的空间分布, 融合全局特征和语义增强的特征可以提高特征的鉴别力和行人检索的准确率. 4) 预训练的 BERT 模型可以提供更加鲁棒的词嵌入使得模型更好、更快的收敛.

表 6 在 CUHK-PEDES 数据集上, 每种模块的销蚀实验结果 (%)

| 编号 | 方法 | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| 1 | 基准方法 | 48.52 | 71.57 | 80.36 |
| 2 | ATT | 50.52 | 72.60 | 80.70 |
| 3 | ATT+FF | 51.86 | 73.95 | 81.87 |
| 4 | ATT+ASR | 54.35 | 75.19 | 82.99 |
| 5 | ATT+ASR+FF | 55.24 | 76.13 | 83.26 |
| 6 | ATT+ASR+FF+BERT | 56.17 | 77.05 | 83.74 |

第二, 我们通过消减属性解耦模块损失函数对应项分析不同损失 (包括中心点损失函数 $L_c$, 基于属性的跨模态匹配损失 $L_m$、基于对比学习的属性解耦损失 $L_d$) 的贡献. 销蚀实验结果见表 7. 通过比较这些模型的实验结果, 我们可以得出以下结论: 1) 每个损失都对属性解耦起着正向的作用. 去掉任何一个损失, 模型的性能都会下降. 2) 同时使用 3 个损失的结果最好, 说明这 3 个损失是互补的. 它们能够相互协作使得模型能够更好地提取属性相关特征.

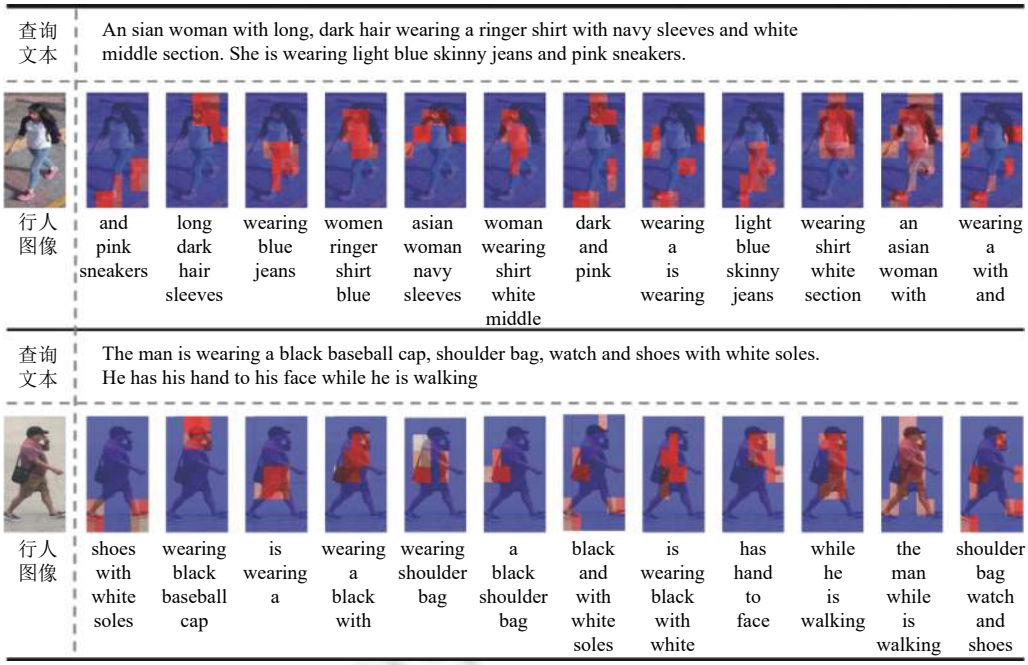



图 4 虚拟属性的注意力热图可视化结果 (红色越深表示相关性越高)

表 7 在 CUHK-PEDES 数据集上, 属性解耦模块损失函数的销蚀实验结果

| 编号 | 方法 | Rank-1 | Rank-5 | Rank-10 |
|---|---|---|---|---|
| 1 | 基线方法 | 48.52 | 71.57 | 80.36 |
| 2 | ATT (w/o $L_c$) | 49.34 | 72.01 | 80.52 |
| 3 | ATT (w/o $L_m$) | 49.25 | 72.13 | 80.54 |
| 4 | ATT (w/o $L_d$) | 49.03 | 71.95 | 80.41 |
| 5 | ATT | 50.52 | 72.60 | 80.70 |

### 3.4 超参数分析

在本实验中, 通过改变 $K$, $\alpha$, $\lambda_1$ 和 $\lambda_2$ 的值进行参数分析. 将 $K$ 的范围设定为 $[0,6,\ldots,36]$, $\alpha$ 的范围设定为 $\{0.5,0.6,0.7,0.8\}$, $\lambda_1$ 的范围设定为 $\{0.1,0.5,1,2\}$, $\lambda_2$ 的范围设定为 $\{0,0.1,0.5,1,2\}$, 并在图 5 中显示结果. 我们可以观察到: 1) 随着 $K$ 值的变化, 模型的准确率稳步上升. 当 $K \geqslant 12$ 时, 提取更多的属性会增加模型的复杂度, 模型的准确率并没有继续提高. 2) 当 $\alpha \geqslant 0.5$ 时, 基于属性的跨模态匹配损失限定匹配的图像-文本对的相似度大于不匹配的图像-文本对的相似度. 随着 $\alpha$ 的值的增加, 匹配样本对与不匹配样本对间的距离也会随之增加. 当 $\alpha = 0.5$ 时, 模型的 Rank-1 准确率最高. 3) 当 $\lambda_1 = 1.0$, $\lambda_2 = 1.0$ 时, 模型的 Rank-1 准确率达到峰值. 当它们的值过小或过大时, 准确率都会下降. 这说明模型中的 3 个损失是同等重要的. 当 $\lambda_2 = 0$ 时可以看到准确率大幅下降, 这说明该损失可以有效地增强模型的跨模态表征能力. 基于以上观察, 我们设置 $K = 12$, $\alpha = 0.5$, $\lambda_1 = 1.0$ 和 $\lambda_2 = 1.0$.

### 3.5 讨 论

本文方法具有较高的检索效率和较好的可解释性. 本文提出的方法使用向量表示图像和文本, 通过比较向量的相似度能够实现快速地检索. 虚拟属性学习可以在细粒度上建立跨模态语义关联增强方法的可解释性. 但虚拟属性不是真实的属性. 从图 4 的可视化结果可以看到, 模型会反复激活相同的区域, 这说明所学习到的虚拟属性缺乏多样性. 同时不同行人的同一个属性会激活不同的行人部位, 这说明学习到的虚拟属性缺乏一致性. 产生这种结

果的可能原因是所使用的属性语义嵌入是随机初始化的, 在没有属性标签的情况下模型会更多地关注对齐图像和文本. 本文方法无法应对多模态语义理解中的不确定性问题[39]. 多模态语义理解中有两种不确定性: 1) 模态信息的不确定性, 由于文本只能描述有限的图像内容, 这导致同一条文本描述可能存在多个对应的行人, 本文方法没有考虑到这种情况. 2) 模态间关联的不确定性, 各个模态上的信息分布是不确定的, 这导致模态间的关联是不确定的和模糊的.

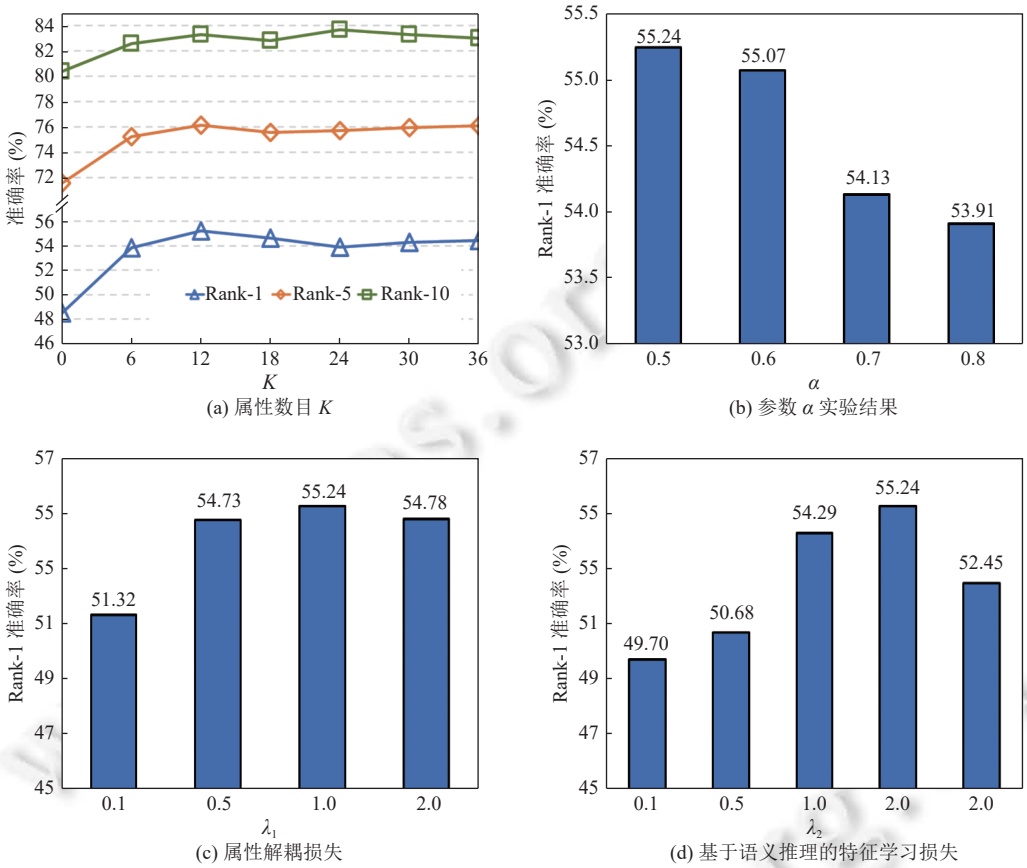



图 5 在 CUHK-PEDES 数据集, 随参数 $K$, $\alpha$, $\lambda_1$ 和 $\lambda_2$ 变化的实验结果

## 4 结 论

本文提出基于虚拟属性学习的文本-图像行人检索方法. 第一, 基于属性的不变性和语义一致性提出语义引导的属性解耦方法. 该方法可以充分地解耦出多样化的属性信息并有效地利用行人属性建立细粒度的跨模态语义关联减少不同模态的异构鸿沟. 第二, 提出的基于语义推理的特征学习模块利用属性构建的语义图模型有效地增强特征的跨模态识别能力. 所提出的方法降低了对数据的标注要求, 通过在公开的文本-图像行人检索数据集和跨模态检索数据集上的实验对比, 表明了本文方法的有效性. 本文提出的方法可以应用于智能视频监控系统中, 比如, 协助办案人员快速筛查可疑人员、在人流量较大的场所 (机场、火车站、游乐场等) 寻找走失儿童或老人等.

本文提出的基于属性学习的文本-图像行人检索方法没有考虑到属性类别的不平衡问题. 真实数据中不同的属性类之间是不平衡. 下一步工作拟引入代价敏感学习对不同属性给予不同的权重. 并尝试采用聚类分析技术对图像和文本进行聚类. 使用类别中心初始化属性的语义嵌入, 根据聚类的结果对不同属性赋予不同权重. 未来还可以围绕多模态语义理解中的不确定性问题开展研究工作.

**References:**


[1] Zheng L, Shen LY, Tian L, Wang SJ, Wang JD, Tian Q. Scalable person re-identification: A benchmark. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 1116–1124. [doi: 10.1109/ICCV.2015.133]

[2] Zhong Z, Zheng L, Cao DL, Li SZ. Re-ranking person re-identification with k-reciprocal encoding. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 3652–3661. [doi: 10.1109/CVPR.2017.389]

[3] Xiao T, Li S, Wang BC, Lin L, Wang XG. Joint detection and identification feature learning for person search. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 3376–3385. [doi: 10.1109/CVPR.2017.360]

[4] Pang L, Wang YW, Song YZ, Huang TJ, Tian YH. Cross-domain adversarial feature learning for sketch re-identification. In: Proc. of the 26th ACM Int'l Conf. on Multimedia. Seoul: ACM, 2018. 609–617. [doi: 10.1145/3240508.3240606]

[5] Wu AC, Zheng WS, Yu HX, Gong SG, Lai JH. RGB-infrared cross-modality person re-identification. In: Proc. of the 2017 IEEE Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 5390–5399. [doi: 10.1109/ICCV.2017.575]

[6] Nguyen DT, Hong HG, Kim KW, Park KR. Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors, 2017, 17(3): 605. [doi: 10.3390/s17030605]

[7] 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]

[8] Zhuo YK, Qi JW, Peng YX. Cross-media deep fine-grained correlation learning. Ruan Jian Xue Bao/Journal of Software, 2019, 30(4): 884–895 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5664.htm [doi: 10.13328/j.cnki.jos.005664]

[9] Luo H, Jiang W, Fan X, Zhang SP. A survey on deep learning based person re-identification. Acta Automatica Sinica, 2019, 45(11): 2032–2049 (in Chinese with English abstract). [doi: 10.16383/j.aas.c180154]

[10] Qi L, Yu PZ, Gao Y. Research on weak-supervised person re-identification. Ruan Jian Xue Bao/Journal of Software, 2020, 31(9): 2883−2902 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6083.htm [doi: 10.13328/j.cnki.jos.006083]

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

[12] 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): 1–23. [doi: 10.1145/3383184]

[13] Sarafianos N, Xu X, Kakadiaris I. Adversarial representation learning for text-to-image matching. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 5813–5823. [doi: 10.1109/ICCV.2019.00591]

[14] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proc. of the 2nd Int'l Conf. on Learning Representations (ICLR). San Diego: ICLR, 2015. 1–14.

[16] 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]

[17] Howard AG, Zhu ML, Chen B, Kalenichenko D, Wang WJ, Weyand T, Andreetto M, Adam H. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861, 2017.

[18] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]

[19] Zha ZJ, Liu JW, Chen D, Wu F. Adversarial attribute-text embedding for person search with natural language query. IEEE Trans. on Multimedia, 2020, 22(7): 1836–1846. [doi: 10.1109/TMM.2020.2972168]

[20] 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. 2617–2625. [doi: 10.1109/WACV45572.2020.9093640]

[21] Dayan P, Abbott LF. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. London: The MIT Press, 2005.

[22] Li S, Xiao T, Li HS, Yang W, Wang XG. Identity-aware textual-visual matching with latent co-attention. In: Proc. of the 2017 IEEE Conf. on Computer Vision. Venice: IEEE, 2017. 1890–1899. [doi: 10.1109/ICCV.2017.209]

[23] Chen TL, Xu CL, Luo JB. Improving text-based person search by spatial matching and adaptive threshold. In: Proc. of the 2018 IEEE Winter Conf. on Applications of Computer Vision (WACV). Lake Tahoe: IEEE, 2018. 1879–1887. [doi: 10.1109/WACV.2018.00208]

[24] 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]

[25] Gao LY, Niu K, Ma ZH, Jiao BL, Tan TH, Wang P. Text-guided visual feature refinement for text-based person search. In: Proc. of the 2021 Int'l Conf. on Multimedia Retrieval. Taipei: ACM, 2021. 118–126. [doi: 10.1145/3460426.3463652]

[26] 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 AAAI Conf. on Artificial Intelligence, 2020, 34(7): 11189–11196. [doi: 10.1609/aaai.v34i07.6777]

[27] Chen W, Liu Y, Bakker EM, Lew MS. Integrating information theory and adversarial learning for cross-modal retrieval. Pattern Recognition, 2021, 117: 107983. [doi: 10.1016/j.patcog.2021.107983]

[28] Jing Y, Wang W, Wang L, Tan TN. Learning aligned image-text representations using graph attentive relational network. IEEE Trans. on Image Processing, 2021, 30: 1840–1852. [doi: 10.1109/TIP.2020.3048627]

[29] Liu JW, Zha ZJ, Hong RC, Wang M, Zhang YD. Deep adversarial graph attention convolution network for text-based person search. In: Proc. of the 2019 ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 665–673. [doi: 10.1145/3343031.3350991]

[30] 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. Glasgow: Springer, 2020. 402–420. [doi: 10.1007/978-3-030-58610-2_24]

[31] Shi JW, Song XM, Liu ZX, Nie LQ. Fashion graph-enhanced personalized complementary clothing recommendation. Journal of Cyber Security, 2021, 6(5): 181–198 (in Chinese with English abstract). [doi: 10.19363/J.cnki.cn10-1380/tn.2021.09.14]

[32] Zheng X, Lin L, Ye M, Wang L, He CL. Improving person re-identification by attention and multi-attributes. Journal of Image and Graphics, 2020, 25(5): 936–945 (in Chinese with English abstract). [doi: 10.11834/jig.190185]

[33] Dong Q, Zhu XT, Gong SG. Person search by text attribute query as zero-shot learning. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Seoul: IEEE, 2019. 3652–3661. [doi: 10.1109/ICCV.2019.00375]

[34] Kim JH, Jun J, Zhang BT. Bilinear attention networks. In: Proc. of the 32nd Conf. on Neural Information Processing Systems. Montréal: NeurIPS, 2018. 1571–1581.

[35] Li YJ, Tarlow D, Brockschmidt M, Zemel RS. Gated graph sequence neural networks. In: Proc. of the 4th Int'l Conf. on Learning Representations (ICLR). San Juan: ICLR, 2016. 1–20.

[36] Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. of the Association for Computational Linguistics, 2014, 2: 67–78. [doi: 10.1162/tacl_a_00166]

[37] 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. Munich: Springer, 2018. 56–73. [doi: 10.1007/978-3-030-01270-0_4]

[38] Chen YC, Huang R, Chang H, Tan CQ, Xue T, Ma BP. Cross-modal knowledge adaptation for language-based person search. IEEE Trans. on Image Processing, 2021, 30: 4057–4069. [doi: 10.1109/TIP.2021.3068825]

[39] Xu T, Zhou PL, Chen EH. Uncertainty in multimodal semantic understanding. Communications of the CAAI, 2020, 10(9): 7–11 (in Chinese with English abstract).

**附中文参考文献**:

[8] 卓昀侃, 綦金玮, 彭宇新. 跨媒体深层细粒度关联学习方法. 软件学报, 2019, 30(4): 884–895. http://www.jos.org.cn/1000-9825/5664.htm [doi: 10.13328/j.cnki.jos.005664]

[9] 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. [doi: 10.16383/j.aas.c180154]

[10] 祁磊, 于沛泽, 高阳. 弱监督场景下的行人重识别研究综述. 软件学报, 2020, 31(9): 2883−2902. http://www.jos.org.cn/1000-9825/6083.htm [doi: 10.13328/j.cnki.jos.006083]

[31] 史金婉, 宋雪萌, 刘子鑫, 聂礼强. 基于时尚图谱增强的个性化互补服装推荐. 信息安全学报, 2021, 6(5): 181–198. [doi: 10.19363/J.cnki.cn10-1380/tn.2021.09.14]

[32] 郑鑫, 林兰, 叶茂, 王丽, 贺春林. 结合注意力机制和多属性分类的行人再识别. 中国图象图形学报, 2020, 25(5): 936–945. [doi: 10.11834/jig.190185]

[39] 徐童, 周培伦, 陈恩红. 多模态语义理解中的不确定性. 中国人工智能学会通讯, 2020, 10(9): 7–11.

**王成济**(1993－), 男, 博士生, 主要研究领域为多媒体信息检索, 机器学习.

**苏家威**(1993－), 男, 博士生, 主要研究领域为医学图像处理, 机器学习.

**罗志明**(1989－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为计算机视觉, 机器学习.

**曹冬林**(1977－), 男, 博士, 助理教授, CCF 专业会员, 主要研究领域为 Web 信息检索, 自然语言处理.

**林耀进**(1980－), 男, 博士, 教授, 主要研究领域为数据挖掘, 机器学习.

**李绍滋**(1963－), 男, 博士, 教授, CCF 高级会员, 主要研究领域为计算机视觉, 机器学习, 多媒体信息检索.