软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(10):4846−4863 [doi: 10.13328/j.cnki.jos.007293] [CSTR: 32375.14.jos.007293]

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

隐式多尺度对齐与交互的文本-图像行人重识别方法*

孙 锐, 杜 云, 陈 龙, 张旭东

(合肥工业大学 计算机与信息学院, 安徽 合肥 230601)

通信作者: 杜云, E-mail: 2022171225@mail.hfut.edu.cn

摘 要: 文本-图像行人重识别旨在使用文本描述检索图像库中的目标行人, 该技术的主要挑战在于将图像和文本特征嵌入到共同的潜在空间中以实现跨模态对齐. 现有的许多工作尝试利用单独预训练的单峰模型来提取视觉和文本特征, 再利用切分或者注意力机制来获得显式的跨模态对齐. 然而, 这些显式对齐方法通常缺乏有效匹配多模态特征所需的底层对齐能力, 并且使用预设的跨模态对应关系来实现显式对齐可能会导致模态内信息失真. 提出了一种隐式多尺度对齐与交互的文本-图像行人重识别方法. 首先利用语义一致特征金字塔网络提取图像的多尺度特征, 并使用注意力权重融合包含全局和局部信息的不同尺度特征. 其次, 利用多元交互注意机制学习图像和文本之间的关联. 该机制可以有效地捕捉到不同视觉特征和文本信息之间的对应关系, 缩小模态间差距, 实现隐式多尺度语义对齐. 此外, 利用前景增强判别器来增强目标行人, 提取更纯洁的行人特征, 有助于缓解图像与文本之间的信息不平等. 在 3 个主流的文本-图像行人重识别数据集 CUHK-PEDES、ICFG-PEDES 及 RSTPReid 上的实验结果表明, 所提方法有效提升了跨模态检索性能, 比 SOTA 算法的 *Rank*-1 高出 2%–9%.

关键词: 文本-图像行人重识别; 隐式对齐; 多尺度融合; 多元交互注意力; 语义对齐

中图法分类号: TP391

中文引用格式: 孙锐, 杜云, 陈龙, 张旭东. 隐式多尺度对齐与交互的文本-图像行人重识别方法. 软件学报, 2025, 36(10): 4846–4863. http://www.jos.org.cn/1000-9825/7293.htm

英文引用格式: Sun R, Du Y, Chen L, Zhang XD. Implicit Multi-scale Alignment and Interaction for Text-image Person Re-identification Method. Ruan Jian Xue Bao/Journal of Software, 2025, 36(10): 4846–4863 (in Chinese). http://www.jos.org.cn/1000-9825/7293.htm

Implicit Multi-scale Alignment and Interaction for Text-image Person Re-identification Method

SUN Rui, DU Yun, CHEN Long, ZHANG Xu-Dong

(School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China)

Abstract: The purpose of text-image person re-identification is to employ the text description to retrieve the target persons in the image database. The main challenge of this technology is to embed image and text features into common potential space to achieve cross-modal alignment. Many existing studies try to adopt separate pre-trained unimodal models to extract visual and text features, and then employ segmentation or attention mechanisms to obtain explicit cross-modal alignment. However, these explicit alignment methods generally lack the underlying alignment ability needed to effectively match multimodal features, and the utilization of preset cross-modal correspondence to achieve explicit alignment may result in modal information distortion. An implicit multi-scale alignment and interaction for text-image person re-identification method is proposed. Firstly, the semantic consistent feature pyramid network is employed to extract multi-scale features of the images, and attention weights are adopted to fuse different scale features including global and local information. Secondly, the association between image and text is learned using a multivariate interaction attention mechanism, which can effectively capture the corresponding relationship between different visual features and text information, narrow the gap between modes, and achieve implicit multi-scale semantic alignment. Additionally, the foreground enhancement discriminator is adopted to enhance the target person and extract

* 基金项目: 国家自然科学基金 (61876057); 安徽省自然科学基金 (2208085MF158); 安徽省重点研究与开发计划 (202004d07020012)
收稿时间: 2023-10-24; 修改时间: 2024-03-22; 采用时间: 2024-09-19; jos 在线出版时间: 2025-05-14
CNKI 网络首发时间: 2025-05-15

purer person features, which is helpful for alleviating the information inequality between images and texts. Experimental results on three mainstream text-image person re-identification datasets of CUHK-PEDES, ICFG-PEDES and RSTPReid show that the proposed method effectively improves the cross-modal retrieval performance, which is 2%–9% higher than the *Rank*-1 of SOTA algorithm.

Key words: text-image person re-identification; implicit alignment; multi-scale fusion; multivariate interaction attention; semantic alignment

行人重识别 ReID (person re-identification) 是智能视频监控领域的一项基本任务. 其目的是根据给定的检索条件 (如人物图像, 相关属性或自然语言描述) 在多个非重叠相机中查询目标行人. 根据查询的模态, 行人重识别任务大致可分为基于图像的搜索[1,2]、基于属性的搜索[3,4]和基于文本的搜索[5,6]. 但是现有的行人重识别方法通常忽略了一些复杂或特殊场景下无法获得行人图像的情况. 例如一些偏远的道路没有监控探头或行人完全被遮挡[7]. 为了解决这个问题, 警方可根据目击者提供的语言描述来搜索目标行人, 即文本-图像行人重识别 TIReID (text-image person re-identification)[8−12]. 如图 1 所示. 文本-图像行人重识别会根据查询文本与图像的相似度对一个大型图像库中的所有人物图像进行排序, 选择排名靠前的人物图像作为匹配项[6]. 由于使用文本描述作为检索查询更加简单自然, 因此文本-图像行人重识别技术具有较好的应用前景.

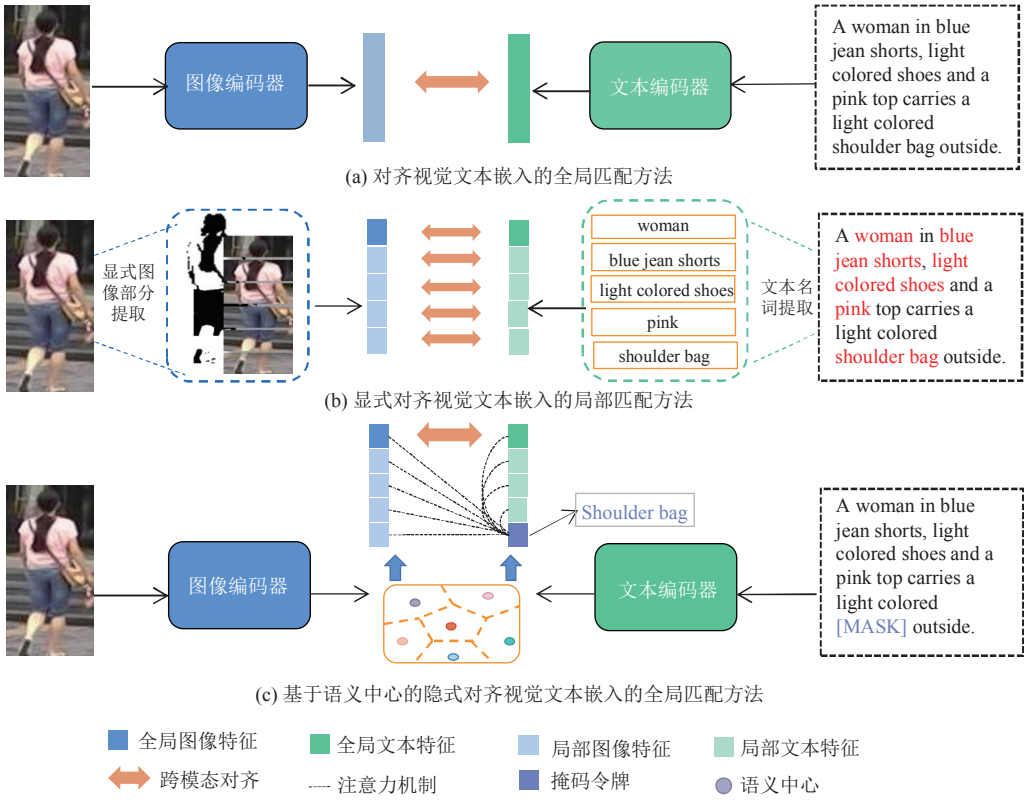


图 1　文本-图像行人重识别的检索范式演变

文本-图像行人重识别涉及两种异构模态的信息处理, 是一项具有挑战性的任务. 该任务可被视为跨模态检索的一个特定子任务[13,14]. 然而在行人重识别过程中, 由于图像可能存在遮挡[15]、背景杂波[16]和姿态干扰[17]等问题, 会使模型难以提取准确的视觉表示; 同时, 因文本描述的任意顺序和文本歧义性会增加特征对齐的不确定性, 进而导致难以实现准确的特征对齐. 此外, 不同人的图像或描述具有非常相似的高层语义, 而图像和文本之间存在显著模态差异, 导致模态间的特征差异远大于单个模态内的特征差异[6,9]. 由此可见, 文本到图像检索的核心研究问题是探索更好的方法来提取区分性和鲁棒性的多模态特征表示, 并设计更好的跨模态匹配方法将图像和文本特征嵌入在公共的潜在空间中进行 对齐. 另外, 图像和文本之间的信息是不平等的[18]: 从视频监控中捕捉到的图像

包含行人信息和背景信息, 同时, 由于相机参数和环境 (如光照条件和天气) 的差异, 采集的图像还包含一些环境因素, 导致了图像中干扰信息较多; 而文本描述通常只包含行人信息, 比如外貌, 性别, 服装及携带的物品等. 因此, 有效地分离图像中的人和背景环境信息, 并对文本特征进行适当的降噪, 是提高检索性能的关键.

近年来, 为了缩小图像和文本之间的模态差距, 人们提出了两种方法: 全局匹配方法和局部匹配方法. 全局匹配方法[6−10,19]从两种模态中提取样本的全局表示, 并设计相应的目标函数来探索共享的潜在嵌入空间, 在该空间中可以直接计算出图像-文本对的匹配分数, 如图 1(a) 所示. 这些方法通常只在网络的末端使用匹配损失来学习跨模态对齐, 缺乏中间层的充分跨模态交互, 而中间层的跨模态交互对弥合特征模态差距至关重要. 同时, 全局匹配方法无法充分挖掘图像中的局部细节. 此外, 图像中还存在一些与文本不相关的背景区域, 这些背景区域的存在以噪声的形式扩大了模态间的差距.

局部图像-文本匹配方法[5,11,12,20−22]通过构建行人身体部位与文本描述实例之间的对应关系来缩小模态差异. 它的一般过程是先显式获取图像和文本的局部表示, 然后建立它们之间的局部对应关系, 如图 1(b) 所示. 为了提取图文显式的局部特征, 常用的策略[21]是将图像分割成条带或小块, 将文本分割成单词, 然后从这些单元中提取特征表示计算局部特征. 然而, 现有的局部匹配方法复杂度高, 可能会破坏图像和文本的上下文信息或引入噪声, 从而影响随后的对齐阶段. 与全局匹配方法相比, 局部匹配方法通过细粒度的信息挖掘和模式间的信息交互提升了性能. 然而, 由于昂贵的成对图文交互操作需要较大的计算量, 局部匹配方法中的信息交互不可避免地会降低推理效率, 在实际应用中难以实现.

针对上述问题, 本文设计了一种基于隐式多尺度对齐和多元交互注意力的文本-图像行人重识别方法, 学习语义对齐的跨模态特征表示, 图 1(c) 展示了该方法的工作流程. 首先, 我们利用语义一致特征金字塔 SCFP (semantic consistent feature pyramid) 网络从图像中提取多尺度特征, 并使用注意力权重融合不同尺度的特征信息. 其次, 本文采用多元交互注意力 MIA (multivariate interaction attention) 机制捕捉视觉特征和文本信息之间的交互关系, 实现隐式多尺度对齐. 另外, 本文引入了前景增强判别器 FED (foreground-enhancing discriminator) 来增强前景, 以提取更好的行人特征, 有助于解决图像与文本之间的信息不平等问题.

本文的主要工作如下.

(1) 针对图像中全局与局部特征未有效融合问题, 本文提出了语义一致特征金字塔网络, 该网络自适应地调整不同特征图之间的权重, 将图像中的细节与整体特征有效融合, 使得最终生成的特征图包含了丰富的全局与局部信息, 从而显著提升了图像表示的表达能力.

(2) 针对显式对齐导致模态内信息失真问题, 本文提出了多元交互注意力机制学习图像和文本之间的关联, 该机制能够有效地捕捉到不同视觉特征和文本信息之间的交互关系, 从而实现隐式多尺度对齐.

(3) 针对图像与文本之间信息不平等问题, 提出了前景增强判别器模块来增强前景, 以提取更加纯净的行人特征, 保留行人身份信息的同时过滤掉环境因素, 有助于缓解图像与文本之间的信息不平等.

本文第 1 节介绍文本-图像行人重识别的相关方法和研究现状. 第 2 节介绍本文提出的基于隐式多尺度对齐和多元交互注意力的文本-图像行人重识别模型. 第 3 节介绍实验设置与结果分析. 最后总结全文.

1 相关工作

1.1 文本-图像行人重识别

文本-图像行人重识别是根据给定的文本描述查找对应的行人图像. 该任务最早由 Li 等人[6]提出. 他们提出了第 1 个基准数据集 CUHK-PEDES 并构建了一个 GNA-RNN 模型来学习文本描述和人物图像之间的亲缘关系. 后来, Li 等人[9]提出了身份感知的两阶段网络, 将模态内和模态间距离联合最小化. Sarafianos 等人[19]提出了文本-图像模态对抗性匹配方法 TIMAM (text-image modality adversarial matching), 尝试通过对抗式和交叉模式匹配目标学习模式不变特征表示. 然而, 这些方法只关注全局表示, 可能会遗漏一些独特的局部细节或噪声信息. 因此, 研究了一些局部匹配方法来克服这个问题. Wang 等人[12]通过对人体进行分段和利用 k-倒数抽样将视觉属性和文本属

性关联起来. Niu 等人[21]提出了一种多粒度图像文本对齐网络来探索不同尺度的关系. Liu 等人[23]引入了由对象属性和关系组成的文本和可视化场景图. Ding 等人[24]设计了并采用复合排名损失来克服文本描述中的方差. 最近, Gao 等人[20]尝试利用一种新型的阶梯 CNN 网络和局部约束 BERT 模型对全尺度表示进行联合对齐.

总之, 目前的研究大多侧重于局部对齐. 这些局部匹配方法都是显式地获取局部特征, 会破坏图像和文本的完整上下文信息, 并引入噪声. 本文从不同的角度研究了跨模态对齐, 在通道方向上融合所有提取的多尺度图像特征, 并与文本特征跨模态交互, 从而聚合到一组共享的语义中心点实现隐式对齐.

1.2 视觉-语言预训练

视觉-语言预训练 VLP (vision-language pre-training) 是一种通过在大规模数据集上进行预训练来学习模型参数的方法, 目的是建立视觉与语言之间的语义对应关系. 目前的 VLP 方法可以分为单流模型和双流模型两种. 在单流模型[25]中, 图像特征和语言特征被联合处理, 然后送入一个 Transformer 编码器中. 虽然单流模型已经取得了很大的成功, 但由于在训练和推理过程中需要进行交叉注意, 这不可避免地引入了延迟和大量的计算. 双流模型[26]则使用两个独立的编码器分别提取文本和视觉特征. 由于这两个编码器没有共享参数, 双流模型缺乏模拟视觉与语言之间复杂交互的能力. 目前, 视觉-语言预训练已经成为学习多模态表征的主流范式, 在视觉问答[27]等任务上显示出强大的效果, 其中最具代表性的是对比语言-图像预训练 CLIP (contrast language-image pre-training) [28]. CLIP 采用对比学习的方式, 利用自然语言监督对大量图像文本数据进行训练, 以获得高质量的视觉特征. 通过语义级别的语言监督, 视觉网络可以学习到具有丰富语义信息的视觉特征, 对跨模态任务和细粒度视觉任务有着巨大的推动作用. 一些研究工作进一步扩展了 CLIP 的应用范围, Yan 等人[29]提出了一种 CLIP 驱动的细粒度信息挖掘框架. Chen 等人[30]使用 CLIP 构建新的细粒度图像池来改善现有基准, 支持更细粒度的语义评估. Zhou 等人[31]探索如何有效地为视觉-语言模型如 CLIP 设计或生成提示, 以提高模型在特定任务上的性能. 然而, 由于 CLIP 被训练为只关注实例级的表示 (图像级、句子级), 而文本-图像行人重识别需要模型关注细粒度信息和跨模态对应, 以区分行人之间的细微差异, 而以往的工作未能直接将原始对齐的 CLIP 双编码器转换为文本到图像的人员检索. 受到这些模型的启发, 我们的研究专注于细粒度信息的挖掘和利用, 以便更精确地处理图像与文本间的微小差异. 同时, 我们通过微调策略优化了 CLIP 模型, 使其更好地适应跨模态对齐的复杂性, 从而在特定的跨模态任务 TIReID 上获得了更好的性能. 我们的工作不仅解决了单一模态数据集上 TIReID 的局限性, 还通过深入分析加深了对模型如何处理和理解跨模态信息的理解, 为未来的研究开辟了新的视角.

2 基于隐式多尺度对齐和多元交互注意力的文本-图像行人重识别

本文提出了一种基于隐式多尺度对齐和多元交互注意力的文本-图像行人重识别方法, 主要分为 3 个部分. 第 1 部分是双路径图像-文本编码器模块, 它利用图像和文本编码器提取视觉和文本特征. 第 2 部分是隐式多尺度语义对齐模块, 该模块在模态内增强行人特征的提取能力. 其中, 前景增强判别器模块对视觉特征进行处理, 从而关注行人特征并且过滤背景信息; 接着, 将过滤后的图像特征送入语义一致特征金字塔模块, 融合不同尺度的特征, 生成全局和局部信息的特征图, 该特征图将作为最终的特征向量和文本进行匹配. 第 3 部分是多元交互注意力模块, 学习图像和文本特征之间的交互关系, 缩小模态之间的差距. 此外, 通过联合优化跨模态投影匹配 (cross-modal projection matching, CMPM) 损失[10]、身份 (identification, id) 损失[7]和多样性 (diversity, div) 损失[29]来提升模型性能, 实现基于语义中心的隐式多尺度对齐. 总体网络框架如后文图 2 所示, 其中, CA 是交叉注意力 (cross attention).

2.1 双路径图像-文本编码器

以往的文本-图像行人重识别方法通常在单模态数据集上分别对图像和文本进行预训练, 这导致缺乏多模态对应信息. 受到 VLP 工作的启发, 我们的方法在 CLIP[28]的图像和文本编码器的基础上初始化. 我们使用两个独立的编码器分别从图像和文本中提取初级特征, 然后通过 Transformers 结构将它们融合. 在训练阶段, 给定训练数据

$D=\{I_i,T_i\}_{i=1}^{N}$, N 表示每批数据集中文本-图像对的总数, 每个文本-图像对由一幅图像 I 和一个相应的文本描述 T. 为了简单起见, 下文中省略下标 i.

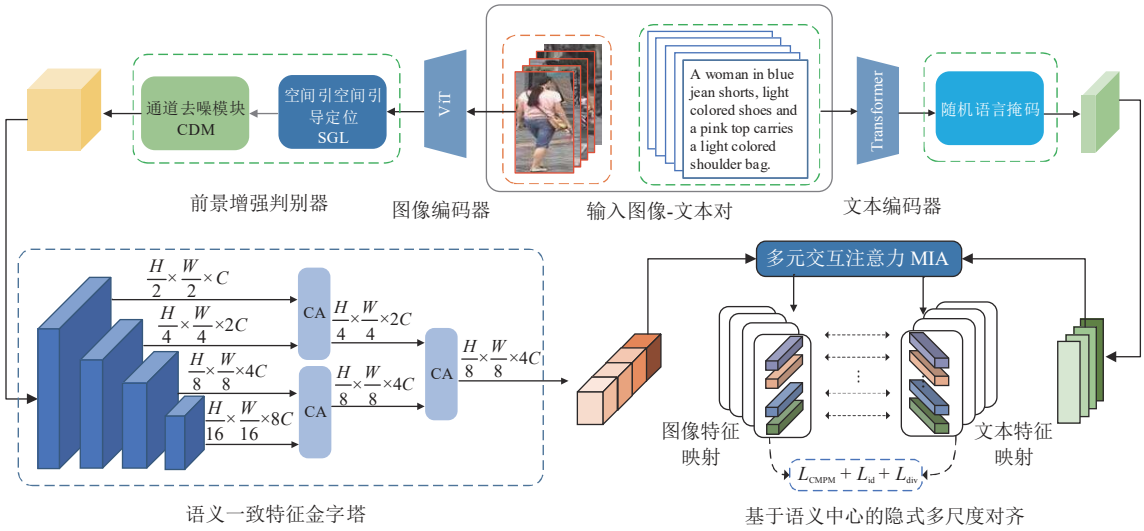


图 2 基于隐式多尺度和多元交互注意力的文本-图像行人重识别模型框架图

● 图像编码器. 给定输入图像 $I\in R^{H\times W\times C}$, 其中, H、W 和 C 分别表示上述特征映射中高度、宽度和通道的维数. 本文采用 CLIP 预训练的 ViT 作为骨干网络提取视觉特征. ViT 是一种基于自注意力机制的图像分类模型, 其核心思想是将输入图像拆分为一系列的图像块, 并将这些块转化为序列数据, 每个块被视为一个令牌 (token), 与自然语言处理中的单词类似. 最后, ViT 使用多层的自注意力机制来对这些图像块进行建模, 从而捕获它们之间的相互关系. 我们首先将 I 拆分成 $K=H\times W/P^2$ 个固定大小的网格状补丁序列, 其中, P 表示块的大小. 然后通过可训练的线性投影将补丁序列映射到 d 维嵌入, 并将可学习的 [CLS] 令牌附加到序列的开始以学习全局表示. 最后, 我们将长度为 $K+1$ 的序列送到 ViT 的编码器中. 输出表示为 $V=\{v_g,v_1,v_2,\ldots,v_k\}\in R^{(K+1)\times d}$, 其中, v_g 是输入图像的全局表示, $\{v_1,v_2,\ldots,v_k\}$ 是补丁局部表示.

● 文本编码器. 对于输入文本 T, 本文使用 CLIP 预训练的文本编码器 Transformer 提取文本表示. 具体来说, 首先使用词汇量为 49 152 的小写字节对编码 (BPE)[32]对文本 T 进行标记. 为了保证文本长度的一致性, 当文本长度大于 L 时, 选择前 L 个单词; 当文本长度小于 L 时, 在文本末尾进行零填充. 接着, 将文本序列线性投影到 d 维嵌入, 在开始处用 [CLS] 令牌填充文本令牌序列. 最后, 将长度为 $L+1$ 的序列输入到 Transformer 编码器中. 输出结果为 $T=\{t_g,t_1,t_2,\ldots,t_L\}\in R^{(L+1)\times d}$, 其中, t_g 是 [CLS] 标记的输入文本的全局特征, $\{t_1,t_2,\ldots,t_L\}$ 是单词级局部特征.

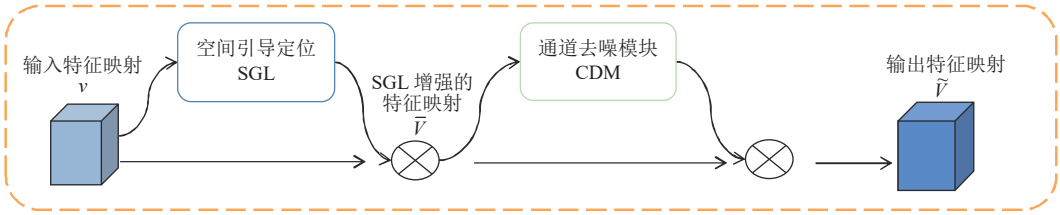
2.2 隐式多尺度对齐模块

为了计算图像和文本之间的相似性, 我们可以通过在共享的嵌入空间中显式对齐图像块序列和文本词来实现. 然而, 由于图像块和单词中存在背景噪声, 这种相似性对于 TIReID 任务是不可靠的. 因此, 我们提出了一种隐式多尺度语义对齐模块, 它能够在各个模态内挖掘更有用的匹配线索, 使得图像的多尺度特征和文本中相应的短语对齐. 具体来说, 使用前景增强判别器模块去除图像中多余的背景和环境信息, 以提高图像特征的准确性. 同时, 我们引入了随机语言掩码模块, 按比例对文本嵌入进行随机掩码, 以增加文本特征的多样性. 此外, 我们还采用语义一致特征金字塔模块从图像中提取多尺度特征, 并在通道维度上进行融合, 以充分利用图像信息.

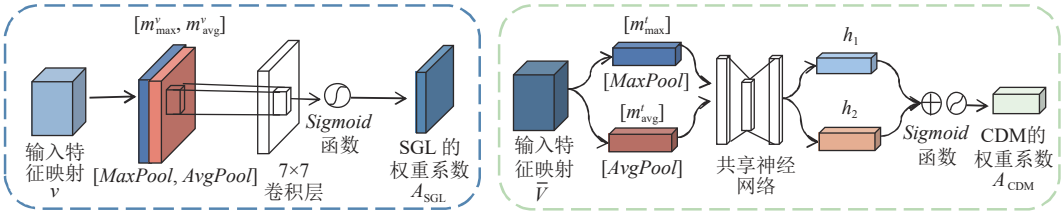
2.2.1 前景增强判别器

行人图像和对应的文本描述之间存在信息不平等的情况. 因为图像包含行人和环境信息, 如图 1(a) 所示; 而文本描述则主要涵盖人物相关特征, 例如性别、外貌、服装、动作等. 为了解决这个问题, 我们提出了前景增强判别器模块, 该模块由空间引导定位 SGL (spatial guide localization) 和通道去噪模块 CDM (channel denoising module)

这两个子模块组成, 如图 3(a) 所示. 通过去除背景和环境信息, FED 模块可以增强前景行人的表现, 在行人特征提取和信息对齐方面起到积极作用.



(a) 前景增强判别器 FED 模块总览



(b) 空间引导定位 SGL 示意图 (c) 通道去噪模块 CDM 示意图

图 3 前景增强判别器

● 空间引导定位 (SGL). 由于注意机制具有增强辨别性特征和抑制无关特征的功能, 因此被广泛应用于各种深度学习任务中, 并对图像的语义理解起到积极作用. 其中, 空间注意力可捕捉特征图内不同空间位置之间的上下文信息, 通过计算每个位置与其他所有位置之间的关系来实现. SGL 是一种利用空间位置关系来引导注意学习的方法, 它能够更准确地将注意力集中在与文本描述相关的行人区域上, 其原理是通过空间注意力来实现, 如图 3(b) 所示. 首先, 分别对来自视觉主干的嵌入序列 $V \in R^{(K+1)\times d}$ 进行通道维度的最大池化 (max pooling) 和平均池化 (average pooling), 得到通道向量 $m^v_{\max} \in R^d$ 和 $m^v_{\text{avg}} \in R^d$.

然后, 将最大池化和平均池化得到的通道向量串联在一起, 得到连接向量 $m^v \in R^{2d}$. 最后, 将连接向量依次经过一个 7×7 的卷积层和一个 *Sigmoid* 激活函数, 得到权重系数 A_{SGL}, 用于控制每个位置的重要性, 如公式 (1) 所示.

$$A_{\text{SGL}} = Sigmoid(f^{\text{Conv } 7\times 7}(m^v)) \tag{1}$$

最后, 将权重系数 A_{SGL} 与输入的图像嵌入序列 $V \in R^{(K+1)\times d}$ 相乘, 可得到 SGL 模块增强后的特征映射 $\bar{V} \in R^{(K+1)\times d}$.

● 通道去噪模块 (CDM). 通道去噪模块利用通道注意力, 旨在捕捉特征图中不同通道之间的相互依赖关系, 如图 3(c) 所示. 通过自适应地加权不同通道的重要性, CDM 使网络能够聚焦于最具信息量的通道, 减弱背景或噪声的相关通道, 使目标行人更加突出. 同时, 因为只处理感兴趣的通道, 相比于对整个图像进行全局处理, 只关注感兴趣通道可以减少计算量和内存消耗, 提高算法的效率和速度.

首先, 我们利用 SGL 增强后的特征映射 $\bar{V}$ 进行全局最大池化和全局平均池化操作, 得到两个通道向量 $m^t_{\max}$ 和 m^t_{avg}. 然后将 $m^t_{\max}$ 和 m^t_{avg} 送入共享的两层神经网络进行处理, 得到 h_1 和 h_2. 使用共享的两层神经网络既可保持模型简单性, 又可提取多维度信息并学习特征的权重, 进一步优化特征的缩放过程, 从而提升模型性能和特征表达能力.

接着, 将两个神经网络处理得到的特征向量 h_1 和 h_2 相加, 经过一个 *Sigmoid* 激活函数操作后得到权重系数向量 A_{CDM}, 其中, W_c 和 b_c 是可学习的参数, 如公式 (2) 所示.

$$A_{\text{CDM}} = Sigmoid(W_c \cdot (h_1 + h_2) + b_c) \tag{2}$$

最后, 将 SGL 增强后的特征映射 $\bar{V}$ 与对应的权重系数 A_{CDM} 逐元素相乘, 得到具有通道注意力加权的特征嵌

入序列 $\tilde{V} \in R^{(K+1)\times d}$.

2.2.2 随机语言掩码模块

给定输入文本, 我们选择以 15% 的概率随机屏蔽文本令牌, 并替换为特定的掩码符号, 比如用“[MASK]”代替. 如果特定的单词被屏蔽, 模型将尝试从其他单词中挖掘有用的行人线索. 比如图 1(c) 图片中的女士, 如果对应文本描述中突出的“a pink shirt”和“a pair of blue jean shorts”被掩码, 模型则会更加注意其他特征, 比如“a pair of gray shoes”和“shoulder bag”. 通过这种方式, 可以获得更深入的语义理解和上下文感知能力, 从而增强模型对文本-图像行人重识别任务的准确性和鲁棒性. 在训练阶段, 随机语言掩码增加了图像-文本跨模态对齐的难度, 但是在推理阶段, 有助于模型挖掘更多的语义对齐.

2.2.3 语义一致特征金字塔网络模块

传统的卷积神经网络通常在靠前的层次提取图像的低级特征, 如边缘、纹理等; 而在较深层次提取图像的高级语义特征, 如物体的形状、部件等. 这些不同层次的特征对目标检测任务都是有价值的. 然而, 如果仅在单一尺度上进行跨模态对齐, 模型可能会忽略局部组件与全局上下文之间的潜在关系, 无法满足高精度识别的要求. 为了解决这个问题, 语义一致特征金字塔网络通过提取图像的多尺度特征并利用交叉注意力进行融合, 自适应地调整特征图之间的权重, 使每个特征图都能关注到其他尺度上的重要特征.

首先, 将一维序列 $\tilde{V} \in R^{(K+1)\times d}$ 利用 *Seq2Img* 操作还原为二维图像特征 $\hat{V} \in R^{H\times W\times C}$, SCFP 采用 2×2 卷积进行降采样从输入图像 $\hat{V} \in R^{H\times W\times C}$ 中提取不同尺度的特征图. 高分辨率特征为局部特征, 包含了目标的边缘、局部大小形态等信息; 低分辨率特征为全局特征, 包含了目标的形态、大小、位置等信息. 我们设计 4 层不同分辨率的特征图层, 各层尺度和作用见表 1.

表 1 SCFP 的 4 个特征图层的尺度及功能比较

No.	尺寸	功能
0	$[H,W,C]$	原始输入图像
1	$[H/2,W/2,C]$	高分辨率的低级特征, 仅包含边缘、纹理等信息
2	$[H/4,W/4,2C]$	以提取更加丰富的降采样后特征图的特征
3	$[H/8,W/8,4C]$	尺寸适中, 既包含局部细节信息, 又有全局信息
4	$[H/16,W/16,8C]$	低分辨率但具有更丰富语义信息的全局特征, 但丢失了许多局部特征

第 1 层的通道数不变, 尺寸缩小为原图的 1/2, 仅包含边缘、纹理这样的低级特征; 第 2 层将通道数扩展为原始图像的 2 倍, 将图像尺寸缩小为原来的 1/4, 可提取更加丰富的特征; 第 3 层尺度的通道数扩展为输入图像的 4 倍, 图像尺寸缩小为原来的 1/8, 该尺度也将作为最终输出图像特征的尺度, 尺寸适中, 不会像第 4 层特征图丢失很多局部细节信息, 又不会像第 1、2 层特征图没有提取到全局信息; 第 4 层的通道数扩展为输入图像的 8 倍, 而图像尺寸缩小为原来的 1/16, 包含全局特征, 但是过于抽象丢失了许多局部特征.

然后, SCFP 将不同尺度的特征图通过交叉注意力计算进行融合. 这样可以增强各个尺度特征图之间的互补性, 使得模型能够更好地提取图像中不同尺度的语义信息. 令第 1 层和第 2 层的特征图进行一次交叉注意力计算. 首先, 利用 *Patch embedding* 将二维图像转化为一维 token 序列, 尺寸分别为 $[B,H\times W/4P,C]$ 和 $[B,H\times W/16P,2C]$, 其中, B、C、H、W 分别表示图像批大小、通道、高度和宽度的维度, 如公式 (3) 所示.

$$E_v = Patch\ embeding(\hat{V}) \tag{3}$$

然后, 将第 1 层得到的 token 序列经过一个线性层做降采样与第 2 层的 token 序列进行尺寸上的对齐, 作为注意力操作的 Q (尺寸为 $[B,H\times W/16P,2C]$), 第 2 层得到的 token 序列作为注意力的 K、V (尺寸为 $[B,H\times W/16P,2C]$). 通过一个常规的多头注意力操作 (*Multi-head attention*、*Add & Norm*、*Feed forward*、*Add & Norm*), 得到尺寸为 $[B,H\times W/16P,2C]$ 的一维 token 序列 X. 再利用 *Seq2Img* 操作还原 X 为二维图像特征 X_1, 尺寸为 $[H/4,W/4,2C]$, 包含了非常丰富的局部信息.

第 3、4 层的交叉融合操作大致同上, 略微不同之处是将第 4 层得到的 token 序列通过一个线性层做上采样

与第 3 层得到的 token 序列进行尺寸上的对齐. 融合的二维图像特征结果为 X_2, 尺寸为 $[H/8, W/8, 4C]$, 既包含了第 4 层准确的全局信息又包含了第 3 层较为丰富的局部信息.

接下来, 令 X_1 和 X_2 再进行一次交叉注意力操作, 对经过 *Patch embedding* 处理的 X_1 而得到的 token 序列需做降采样, 从而与 X_2 得到的 token 序列进行对齐. 此次融合的结果为 X^V, 尺寸为 $[H/8, W/8, 4C]$, 包含了丰富的局部和全局信息, 更好进行跨模态交互.

2.3 多元交互注意力模块

为了学习图像和文本之间的对应关系, 实现准确地语义对齐, 我们设计了一个多元交互注意力模块来整合图像和文本特征. 在这个模块中, 本文使用受多样性损失约束的多头注意[23]模块实现图像-文本对齐, 其中, Q 表示 Query, 用于查询相应的信息; K 和 V 分别是 Key 和 Value 的缩写, 作为键和值去匹配和获取相应的特征. 经过 MIA 的处理, 我们可以得到更好的图像-文本对齐效果. 如图 4 所示.

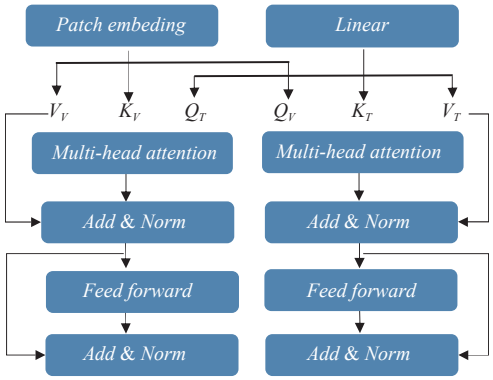


图 4 多元交互注意力模块 MIA 结构图

2.3.1 多头注意模块

对于视觉模态, 我们把 SCFP 处理得到的图像特征 $X^V \in R^{H/8,W/8,4C}$ 作为输入, 利用 *Patch embedding* 将图像的三维特征转换成一维序列矩阵 E, 如公式 (3) 所示. 接着, 通过线性投影计算出图像特征映射的 n 个 *head* 中的第 i 个中的 3 个向量 Q_V、K_V、V_V, 如公式 (4) 所示, 其中, 可训练参数矩阵 $W_i^Q, W_i^K, W_i^V \in R^{d\times d}$ 分别代表一个线性层.

$$Q_i = E_v \cdot W_i^Q,\ K_i = E_v \cdot W_i^K,\ V_i = E_v \cdot W_i^V \tag{4}$$

然后, 计算输入图像的第 i 个 *head* 的注意力权重矩阵 $head_i \in R^{(N+1)\times(N+1)}$, 如公式 (5) 所示, 其中, d_K 表示键 (K) 向量的维度, 决定了输入向量投影后的键的维度.

$$head_i(Q,K,V) = Softmax\left(\frac{Q_i \cdot K_i}{\sqrt{d_K}}\right) \cdot V_i \tag{5}$$

最后, 将 n 个 *head* 得到的注意力矩阵拼接到一起即完成多头注意力计算, 如公式 (6) 所示.

$$MHA(Q,K,V) = Concat(head_1, head_2, \ldots, head_n) \tag{6}$$

对于文本模态, 我们将文本编码器提取得到的文本特征 $X^T \in R^{(L+1)\times d}$ 输入一个线形层, 得到 Q、K、V 矩阵, 如公式 (7), 然后与图像共享相同参数的多头注意模块.

$$E_t = Linear(X^T) \tag{7}$$

2.4 交叉注意力模块

在获得图像和文本的多头注意力矩阵后, 我们需要进行交叉注意力操作. 首先, 我们分别取文本编码得到的 Q_t 矩阵与图像编码得到的 K_v、V_v 进行注意力计算, 取图像编码得到的 Q_v 矩阵与文本编码得到的 K_t、V_t 进行注意力计算, 得到注意力矩阵 $Attn_v$、$Attn_t$, 如公式 (8) 所示.

$$Attn_v = MHA(Q_t, K_v, V_v),\ Attn_t = MHA(Q_v, K_t, V_t) \tag{8}$$

在得到注意力矩阵 $Attn_v$、$Attn_t$ 后, 我们利用常规的注意力模块 (如 *Add & Norm* 和 *Feed forward*) 来融合跨模态信息. 具体而言, 我们使用文本信息作为查询矩阵 Q 对图像信息进行匹配加权操作, 使网络更加关注与文本所提及的相关特征的图像区域, 从而实现图像和文本之间的语义对齐; 同时我们也可以使用图像信息作为查询矩阵 Q, 对文本信息进行匹配加权操作, 过滤文本中的非必要的信息, 将注意力集中在与图像相关的关键词上. 通过这种跨模态交互操作, 我们能够更好地整合图像和文本信息, 利用它们之间的对应关系来获取更丰富的特征表示, 获得更准确的语义对齐结果.

虽然图像和文本的模态不同, 但它们包含的语义信息相同. 因此, 我们认为在公共的语义空间中存在一组潜在的语义中心, 其中包含了行人的语义信息, 并由不同模态共享. 本文提出的基于语义中心的多尺度对齐可自适应地选择和聚合图像和文本特征到同一主题, 并获得多个多尺度对齐的图像和文本特征. 我们通过计算特征和中心点之间的相似度, 将图像和文本特征分配给相应的语义中心. 同一行人的所有特征都向其所属的语义中心点聚集, 而不同行人的特征之间相互推远. 如图 5 所示. 例如 ID A 中分组 1 和分组 2 中的特征会向所属的中心点聚集, 且 ID A 和 ID B 的语义中心点之间的距离会加大.

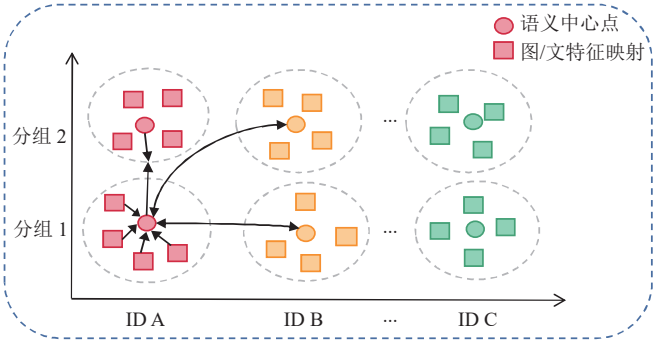


图 5 基于语义中心的多尺度对齐示意图

2.5 损失函数设计

为了消除图像和文本的模态差距从而实现隐式语义对齐, 我们引入了跨模态投影匹配损失 L_{CMPM}. 它将跨模态投影结合到 KL 散度中, 将不同模态的表示关联起来. 该算法不需要传统的双向排序丢失的三重采样和边距选择[29], 在不同批量大小的图像和文本关联中表现出很好的稳定性和优越性. 对于每个视觉表示 f_i^v, 我们假设图像-文本表示对的集合是 $\left\{\left(f_i^v, f_j^t\right), y_{i,j}\right\}_{j=1}^N$, 其中, $y_{i,j}$ 是真实匹配的标签, $y_{i,j}=1$ 表示 $\left(f_i^v, f_j^t\right)$ 是来自同一身份的匹配对, $y_{i,j}=0$ 表示非匹配对. f_i^v 和 f_j^t 是匹配对的概率可通过公式 (9) 计算.

$$p_{i,j}=\frac{\exp\left(\left(f_i^v\right)^{\mathrm{T}}\bar{f}_j^t/\tau\right)}{\sum_{k=1}^N\exp\left(\left(f_i^v\right)^{\mathrm{T}}\bar{f}_k^t/\tau\right)},\ \bar{f}_j^t=\frac{f_j^t}{\left\|f_j^t\right\|} \tag{9}$$

其中, τ 是控制概率分布峰值的温度超参数, $\bar{f}_j^t$ 表示标准化文本特征. 在几何上, $\left(f_i^v\right)^{\mathrm{T}}\bar{f}_j^t$ 表示将图像特征 f_i^v 投影到文本特征 f_j^t 上, 且 $p_{i,j}$ 可视为小批量中的一对 $\left(f_i^v, f_j^t\right)$ 标量投影的百分比. 然后, 通过公式 (10) 可计算在小批量中从图像到文本的 CMPM 损失.

$$L_{\text{CMPM}}^{v2t}=\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\varepsilon}\right),\ q_{i,j}=\frac{y_{i,j}}{\sum_{k=1}^N y_{i,k}} \tag{10}$$

其中, ε 是一个避免数值问题的小数字, N 表示小批量尺寸, $q_{i,j}$ 表示归一化的真实匹配概率. 上述过程在从图像到文本单个方向上减小了每个视觉表示与其匹配的文本表示之间的距离, 并且我们反向进行类似的过程以将每个文本表示与其匹配的视觉表示拉近. 对称地, 从文本到图像的 CMPM 损失通过在公式 (9)、(10) 中交换 f^v 和 f^t 来计算. 因此, 双向 CMPM 损失通过公式 (11) 计算.

$$L_{\mathrm{CMPM}} = L_{\mathrm{CMPM}}^{v2t} + L_{\mathrm{CMPM}}^{t2v} \tag{11}$$

同时, 考虑到多头注意模块中不同 *head* 的关注块可以捕获彼此冗余和重叠的语义, 为了充分挖掘图像和文本中的细粒度细节, 希望不同尺度的特征聚焦于不一致的信息, 我们对不同尺度的特征施加多样性约束损失 L_{div}, 避免信息冗余, 如公式 (12) 所示.

$$L_{\mathrm{div}} = \sum_{i=1}^{N}\sum_{j=1,\ i\neq j}^{N}\left(\frac{f_i^v f_j^v}{\left\|f_i^v\right\|_2\left\|f_j^v\right\|_2} + \frac{f_i^t f_j^t}{\left\|f_i^t\right\|_2\left\|f_j^t\right\|_2}\right) \tag{12}$$

此外, 我们采用身份损失 L_{id} 将行人图像或文本按身份划分为不同的群体, 保证了身份层次的匹配. 它明确地考虑了模态间的距离, 保证了同一图像/文本组的特征表示在联合嵌入空间中紧密地聚类在一起. 其中, W_{id} 是用于调整不同标签重要性的权重向量, $GN(X)$ 是通过全局规范化处理得到的归一化图像特征向量, 身份损失表示为:

$$L_{\mathrm{id}}(X) = -\log(Softmax(W_{\mathrm{id}} \times GN(X))) \tag{13}$$

通过上述跨模态投影匹配损失、多样性损失和身份损失的约束, 我们可从图文中获得不同的语义对齐感知特征. 综上, 最终的损失函数表示如下:

$$L = L_{\mathrm{CMPM}} + L_{\mathrm{div}} + L_{\mathrm{id}} \tag{14}$$

3 实验结果与分析

3.1 数据集与性能评价指标

为了验证本文方法的有效性, 我们在 3 个具有挑战性的文本到图像的人物检索数据集 CUHK-PEDES、ICFG-PEDES 及 RSTPReid 上进行了广泛的性能评估.

CUHK-PEDES[6]是第 1 个专门用于文本到图像的人检索的数据集, 如图 6 所示, 包含了 40 206 幅图像和 80 412 个文本描述, 用于 13 003 个身份. 按照官方数据分割方法, 训练集由 11 003 个身份、34 054 个图像和 68 108 个文本描述组成. 验证集包含 3 078 张图像和 6 156 个文本描述, 而测试集包含 3 074 张图像和 6 148 个文本描述, 它们都有 1 000 个标识.

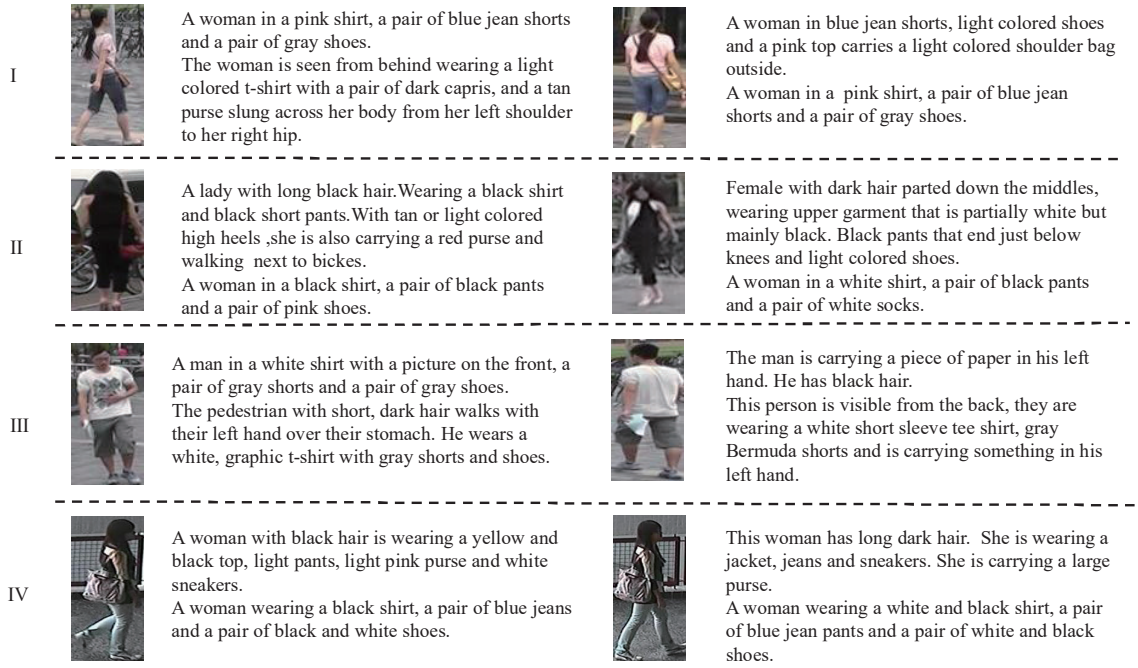


图 6 来自 CUHK-PEDES 数据集的行人图像-文本对

ICFG-PEDES[24]包含 4 102 个身份的 54 522 个图像, 包含了比 CUHK-PEDES 更多的以身份为中心和细粒度的文本描述. 每个图像只有一个对应的文本描述. 该数据集分为训练集和测试集, 训练集包含 34 674 个图像-文本对, 其中有 3 102 个标识, 而测试集包含 19 848 个图像-文本对, 用于剩余的 1 000 个标识.

RSTPReid[18]包含来自 15 个摄像头的 4 101 个身份的 20 505 张图像. 每个身份都有 5 张由不同摄像机拍摄的对应图像, 每个图像都有 2 个文本描述. 拆分官方数据之后, 分别使用 3 701、200 和 200 个身份进行训练、验证和测试. 每个句子不少于 23 个单词.

本文的评估指标采用排名前 k 命中率 *Rank-k* (k=1, 5, 10) 作为主要的评价度量. 当给定一个查询文本描述, 所有图库图像都根据其相似度值进行排名. 成功的搜索意味着在前 k 个图像中存在匹配的人物图像. 另外, 对于综合评价, 我们还采用了平均正确率均值 (*mAP*) 和平均逆负惩罚 (*mINP*) 作为另外的检索准则. *Rank-k*、*mAP* 和 *mINP* 的值越高, 性能越好.

3.2 实验配置及细节

我们使用 PyTorch 实现了所提出的模型, 在单个 RTX3090 24GB GPU 上进行了训练. 本文采用 CLIP 的图像编码器 ViT 提取视觉特征, 所有输入图像的大小均调整为 384×128, 在训练过程中采用随机水平翻转、随机填充裁剪和随机擦除等方法增强图像数据. 本文采用 CLIP 的文本编码器 Transformer 提取文本特征, 将文本序列的最大长度设置为 L=100. 图像和文本特征的嵌入维度被设置为 d=768. 对于多元交互编码器的每一层, 头的大小和个数被设置为 512 和 8. 在训练阶段, 我们采用 Adam 作为优化器, 训练 200 个周期, 批量大小为 32, 学习速率初始化为 1×10^{-5}, 余弦学习速率衰减. 初始时, 我们花了 5 个周期来预热, 将学习速率从 1×10^{-6} 线性增加到 1×10^{-5}. 对于随机初始化的模块, 我们将初始学习速率设置为 5×10^{-5}. CMPM 损失函数中的温度超参数 τ 设置为 0.02.

在测试阶段, 使用余弦距离来度量图像-文本对的相似度值. 根据文本查询, 对相似度得分进行排序, 从图像库中检索出人物图像.

3.3 与 SOTA 方法的对比分析

为了验证本方法在文本-图像行人重识别任务中的优越性, 我们将所提方法与现有主流方法在 CUHK-PEDES、ICFG-PEDES 及 RSTPReid 这 3 个数据集上进行比较, 结果如表 2–表 4 所示. 这些方法可分为两类: (1) 以 GNA-RNN[6]、IATV[9]、Dual Path[7] 等方法为代表的全局匹配方法; (2) 以 PWM+ATH[5]、GLA[22]、ViTAA[12]等方法为代表的局部匹配方法. 我们的方法在所有 3 个基准数据集上始终取得最先进的结果, 并取得了重大改进.

表 2　各方法在 CUHK-PEDES 数据集上的实验结果对比 (%)

方法	类型	来源	*Rank*-1	*Rank*-5	*Rank*-10	*mAP*	*mINP*
GNA-RNN[6]	G	CVPR 2017	19.05	—	53.64	—	—
IATV[9]	G	ICCV 2017	25.94	—	60.48	—	—
PWM+ATH[5]	L	WACV 2018	27.14	49.45	61.02	—	—
GLA[22]	L	ECCV 2018	43.58	66.93	76.26	—	—
Dual Path[7]	G	TOMM 2020	44.40	66.26	75.07	—	—
CMPM/C[10]	G	ECCV 2018	49.37	—	79.27	—	—
TIMAM[19]	G	ICCV 2019	54.51	77.56	84.78	—	—
ViTAA[12]	L	ECCV 2020	55.97	75.84	83.52	—	—
NAFS[20]	L	arXiv 2021	59.94	79.86	86.70	—	—
DSSL[18]	L	MM 2021	59.98	80.41	87.56	—	—
SSAN[24]	L	arXiv 2021	61.37	80.15	86.73	—	—
LapsCore[33]	L	ICCV 2021	63.40	—	87.80	—	—
ISANet[34]	L	TNNLS 2023	63.92	82.15	87.69	—	—
LBUL[35]	L	MM 2021	64.04	82.66	87.22	—	—
TBPS[36]	G	BMVC 2021	64.08	81.73	88.19	60.08	—

表 2 各方法在 CUHK-PEDES 数据集上的实验结果对比 (%) (续)

方法	类型	来源	*Rank*-1	*Rank*-5	*Rank*-10	*mAP*	*mINP*
SAF[37]	L	ICASSP 2022	64.13	82.62	88.40	—	—
TIPCB[38]	L	Neuro 2022	64.26	83.19	89.10	—	—
CAIBC[39]	L	MM 2022	64.43	82.87	88.37	—	—
AXM-Net[40]	L	MM 2022	64.44	80.52	86.77	—	—
LGUR[41]	L	MM 2022	65.25	83.12	89.00	—	—
IVT[42]	G	ECCVW 2022	65.59	83.11	89.21	—	—
BLIP[43]	G	ICML 2022	65.61	82.84	88.65	58.02	—
TransTPS[44]	L	TMM 2023	68.23	86.37	91.65	—	—
CFine[29]	L	TIP 2023	69.57	85.93	91.15	—	—
Ours	G	—	**73.55**	**89.30**	**93.86**	**66.28**	**50.68**

表 3 各方法在 ICFG-PEDES 数据集上的实验结果对比 (%)

方法	类型	来源	*Rank*-1	*Rank*-5	*Rank*-10	*mAP*	*mINP*
Dual Path[7]	G	TOMM 2020	38.99	59.44	68.41	—	—
CMPM/C[10]	L	ECCV 2018	43.51	65.44	74.26	—	—
ViTAA[12]	L	ECCV 2020	50.98	68.79	75.78	—	—
SSAN[24]	L	arXiv 2021	54.23	72.63	79.53	—	—
IVT[42]	G	ECCVW 2022	56.04	73.60	80.22	—	—
ISANet[34]	L	TNNLS 2023	57.73	75.42	81.72	—	—
CFine[29]	L	TIP 2023	60.83	76.55	82.42	—	—
Ours	G	—	**63.32**	**80.30**	**85.81**	**38.14**	**7.84**

表 4 各方法在 RSTPReid 数据集上的实验结果对比 (%)

方法	类型	来源	*Rank*-1	*Rank*-5	*Rank*-10	*mAP*	*mINP*
Dual Path[7]	G	TOMM 2020	38.99	59.44	68.41	—	—
CMPM/C[10]	L	ECCV 2018	43.51	65.44	74.26	—	—
ViTAA[12]	L	ECCV 2020	50.98	68.79	75.78	—	—
SSAN[24]	L	arXiv 2021	54.23	72.63	79.53	—	—
IVT[42]	G	ECCVW 2022	56.04	73.60	80.22	—	—
ISANet[34]	L	TNNLS 2023	57.73	75.42	81.72	—	—
BLIP[43]	G	ICML 2022	58.25	77.85	85.65	44.08	—
TransTPS[44]	L	TMM 2023	56.05	78.65	86.75	—	—
CFine[29]	L	TIP 2023	**60.83**	76.55	82.42	—	—
Ours	G	—	59.25	**82.40**	**88.90**	**46.80**	**24.85**

首先, 我们在最流行且广泛使用的基准 CUHK-PEDES 上评估我们的方法. 如表 2 所示, 我们的方法优于所有最先进的方法, *Rank*-1、 *mAP* 和 *mINP* 分别达到了 73.55%、66.28% 和 50.68%. 特别是, 我们的模型优于相同的基于 CLIP 的 CFine, *Rank*-1、*Rank*-5、*Rank*-10 分别提升了 3.98%、3.37%、2.71%. 这可归功于我们提出的隐式多尺度对齐模块. 不同于 CFine 的单独提取图文全局和局部特征, 我们的模型利用语义一致特征金字塔网络 SCFP 将不同尺度的特征向量沿通道方向拼接, 获得有多尺度信息的特征图. 因此, 本文在跨模态行人检索方面效率更高.

为了全面评估我们方法的泛化能力, 我们将其与 ICFG-PEDES 和 RSTPReid 两个基准数据集上的现有研究成果进行了对比, 结果如表 3 和表 4 所示. 数据显示, 我们的方法在这两个数据集上与最新的主流方法的性能接近.

在 ICFG-PEDES 数据集上, 我们的方法在 *Rank*-1 准确率达到了 63.32%, 在 *Rank*-5 和 *Rank*-10 准确率分别达到了 80.30% 和 85.81%, 而在 *mAP* 上也取得了 38.14% 的高分. 这些成绩在所有列出的最新方法中位居首位, 尤其是与 CFine 相比, 我们在 *Rank*-1 上的提升达到了显著的 2.49 个百分点, 充分证明了我们方法在这一领域具有一定的优势和竞争力.

在 RSTPReid 数据集上, 我们的方法在 *Rank*-1 准确率上同样表现良好, 达到了 59.25%, 虽略低于相同的基于 CLIP 的 CFine 模型的 60.83%, 但 *Rank*-5 和 *Rank*-10 准确率分别为 82.40% 和 88.90%, 较 CFine 显著提高了 5.85% 和 6.48%. 同时, 在 *mAP* 上也取得了令人瞩目的 46.80% 的得分. 此外, 我们在 *mINP* 指标上虽然相对较低, 仅为 24.35%, 但这个指标主要反映了模型在辨识最具挑战性的匹配样本方面的能力. 尽管如此, 我们的方法在其他关键性能指标上的显著优势, 包括 *mAP* 和 *Rank* 指标上的高分, 表明我们的模型不仅在平均性能上较好, 而且在识别最可能正确的匹配样本方面也表现优异.

综上所述, 我们的方法在所有 3 个基准数据集上的所有指标上始终达到最佳性能. 这证明了我们提出的方法的泛化性和鲁棒性.

3.4 消融实验

为了验证本文所提方法的各重要组成部分的有效性, 我们在 CUHK-PEDES 数据集上进行了广泛的消融实验. “Baseline”是指仅使用在 CLIP 上预训练的 ViT 和 Transformer 作为图像和文本编码器来提取特征, 而不添加任何模块和进一步的特征嵌入. 在实验过程中, 我们通过组合不同组件来验证模型中每个组件的贡献, 结果见表 5.

表 5　我们在 CUHK-PEDES 上所提出模型的不同组成部分的消融研究 (%)

No.	方法	组成部分			*Rank*-1	*Rank*-5	*Rank*-10
		FED	SCFP	MIA			
0	Baseline	—	—	—	68.19	86.47	91.47
1	+FED	√	—	—	69.31	86.86	91.68
2	+SCFP	—	√	—	70.16	87.15	92.04
3	+MIA	—	—	√	70.55	87.55	92.45
4	+FED+SCFP	√	√	—	71.23	88.10	92.71
5	+FED+MIA	√	—	√	71.84	88.57	93.11
6	+SCFP+MIA	—	√	√	73.09	88.93	93.42
7	Ours	√	√	√	**73.55**	**89.30**	**93.86**

注: “+”表示在基线模型(Baseline)上增量添加的组件模块

● 前景增强判别器模块 (FED). 通过比较表 5 中 No. 0 和 No. 1 的结果, 证明了添加了 FED 的模型性能得到明显提升, *Rank*-1 从 68.19% 提升到 69.31%. 这是因为相较于传统的行人图像特征提取, FED 过滤背景和环境信息并且增强行人特征, 有助于缓解图像与文本之间的信息不平等, 减少模态间差距.

● 语义一致特征金字塔网络 (SCFP). 为了验证 SCFP 的有效性, 我们比较了 No. 0 和 No. 2, 相较于 Baseline, *Rank*-1、*Rank*-5 和 *Rank*-10 分别提高了 1.97%、0.68% 和 0.57%. 如 No. 4 的结果所示, 在 No. 1 基础上添加 SCFP, 性能分别提高了 1.92%、1.24% 和 1.03%. 这表明多尺度特征融合能带来显著的性能增益, 交叉注意力通过自适应地调整特征图之间的权重, 使得处理后的特征既包含全局特征又包含局部特征, 有助于跨模态匹配图文特征.

● 多元交互注意机制 (MIA). 通过比较 No. 0 和 No. 3 的结果, 性能分别提高了 2.36%、1.08% 和 0.98%. 当在 No. 1 的基础上添加 MIA 模块时, *Rank*-1 从 69.31% 显著提升到 71.84%; 而在 SCFP 的基础上添加 MIA 的结果如 No. 6 所示, *Rank*-1、*Rank*-5 和 *Rank*-10 分别提高了 2.93%、1.78% 和 1.38%. 通过实验验证, 添加 MIA 模块的框架在图文重识别任务中表现出了更好的性能. 它能够有效地缩小不同模态之间的差距, 并增强图像和文本之间的隐式语义对齐. 这意味着模型能够更好地理解图像和文本之间的关联, 从而实现更准确的行人重识别.

通过比较 No. 0 和 No. 7 的结果, 可知直接使用 CLIP 来实现 TIReID 是不明智的, 结果并不是最优的. 当与我

们提出的 FED、SCFP 和 MIA 这 3 个模块叠加使用时, 性能分别提高了 5.36%、2.83% 和 2.39%. 综上, 本文所提出的各个模块都能够有效减少模态间差距, 对文本-图像行人重识别起积极作用.

3.5 超参数影响分析

● 温度超参数. 在损失函数中, 温度超参数 τ 可调整生成的特征向量之间的相似度, 如公式 (9) 所示. 较高值会使概率分布更平坦, 即特征向量之间的相似度更加均匀, 有利于区分不同的行人. 较低值会使概率分布更尖锐, 即特征向量之间的相似度差异更加明显, 有助于增强对于相似行人的识别能力. 本节在 CUHK-PEDES 数据集上对不同 τ 进行实验, 结果如图 7 所示. 随着 τ 从 0 增加到 0.02, *Rank*-1 和 *mAP* 的参数也在提升; 在 τ 从 0.02 到 0.05 的过程中, 模型性能逐渐下降, 在其他两个数据集上也有类似的实验结果. 因此, 本模型中 $\tau = 0.02$.

● 随机文本掩码概率. 在进行文本掩码的消融实验时, 关键在于确定掩码概率对模型性能的影响. 文本掩码概率指的是在训练过程中随机屏蔽文本输入序列中某一部分的比例. 适当的掩码概率可以迫使模型更加依赖上下文信息来预测被遮盖的词汇, 从而学习到更加丰富和鲁棒的特征表示. 合适的文本掩码概率应该平衡模型对特征差异性的敏感性与行人识别能力的增强. 较低的掩码概率可能不足以激励模型学习到足够的上下文依赖性, 而过高的掩码概率则可能导致信息缺失, 阻碍模型学习到有效的特征表示. 本节在 CUHK-PEDES 数据集上对不同的文本掩码的概率进行消融实验, 结果如图 8 所示. 随着掩码概率的增加, 模型性能经历了先上升后下降的变化. 在掩码概率较低时 (如图中的 6%、9% 和 12%), *Rank*-1 和 *mAP* 指标较低, 这可能表明模型没有足够的挑战来学习深层语义关系, 导致对不同行人的区分能力不强. 而当掩码概率增加到 15% 时, *Rank*-1 达到了峰值, 这表明适度的挑战促进了模型对于上下文和细节的学习, 从而提升了识别能力. 但是, 当掩码概率进一步提升至 18% 时, *Rank*-1 和 *mAP* 指标开始下降, 暗示过高的掩码概率可能导致信息缺失过多, 阻碍了模型从文本中学习有效的特征表示. 该模式在另外两个数据集上也有相似的趋势. 综上, 本文的随机文本掩码概率为 15%.

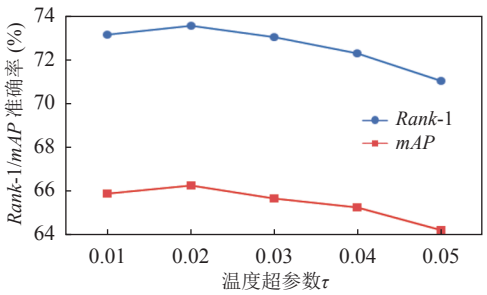


图 7 在 CUHK-PEDES 数据集上分析温度超参数 τ

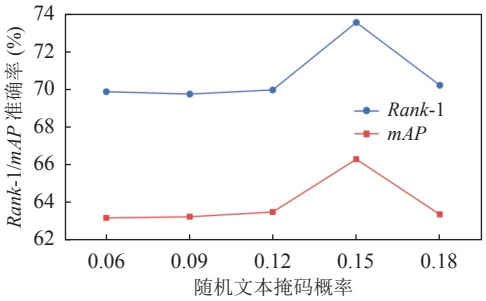


图 8 在 CUHK-PEDES 数据集上分析随机文本掩码概率

3.6 计算效率分析

在本节中, 我们对推理阶段的模型参数量和检索时间进行了细致分析. 如表 6 所示, 我们主要将模型的参数和计算成本与几个 TIReID 领域的最新方法进行比较, 例如 NAFS[20]、SSAN[24]、TBPS[14], 以及一般图文检索中的典型方法, 例如 ViLT[45]、ALBEF[46]. 由于 Transformer 的参数比 LSTM 或 BiGRU 多, 我们的模型参数量超过 SSAN、TBPS, 但我们的检索时间只有 8 s, 大幅低于上述模型在推理阶段的检索时间. 此外, 我们的方法采用了微调后的 CLIP 预训练模型来初始化参数, 通过图像编码器 ViT 和文本编码器 Transformer 分别提取视觉特征和文本特征, 总参数量达到了 194.55M, 虽然略高于仅使用 Transformer 的 SSAN 的 166.45M, 但检索时间却只占其 20%. 相较于通用的图文检索方法如 ViLT 和 ALBEF, 我们的方法在检索时间上具有明显优势. 例如, ViLT 在 CUHK-PEDES 数据集上的测试需要 103 320 s, 而我们的方法仅需 8 s. 这得益于我们的方法无需对所有可能的图文对进行编码, 而是仅提取一次特征. 综合来看, 我们的模型在保持参数量和计算成本适中的同时, 有较好的性能表现.

表 6 模型大小和计算时间的比较

方法	组成结构	参数量 (M)	检索时间 (s)
ViLT[45]	Transformer	96.50	103 320
ALBEF[46]	Transformer	209.50	12 240
NAFS[20]	ResNet+BERT	189.00	78
SSAN[24]	ResNet+LSTM	97.86	31
TBPS[36]	ResNet+BiGRU	84.83	26
IVT[42]	Transformer	166.45	42
Ours	ViT+Transformer	194.55	8

3.7 可视化分析

图 9 展示了 Baseline 和我们提出的方法的前 10 个检索结果, 其中, 匹配和不匹配的人物图像分别用红色和蓝色矩形标记. 如图所示, 我们的方法在检索结果上更加准确. 在某些 Baseline 无法检索到正确结果的情况下, 我们的方法也可以在 Rank-3 中也能找到正确的结果. 这主要得益于我们的隐式多尺度对齐及多元交互注意力模块, 它利用融合的图像多尺度特征与文本特征进行跨模态交互, 减少了模态间差距. 此外, 我们发现细粒度的判别线索 (如包、长发、鞋子等) 更能区分不同的行人, 这些线索在图 9 中绿色和橙色突出显示的文本和图像区域框中进行了说明.

图 9 Baseline 和我们的方法在 CUHK-PEDES 上对每个文本查询的前 10 个检索结果的比较

4 总 结

本文提出了一种基于隐式多尺度对齐和多元交互注意力的文本-图像行人重识别方法. 首先, 本方法利用语义一致特征金字塔 (SCFP) 提取和融合图像不同尺度特征, 来获得同时包含全局和局部信息的特征图. 其次, 使用多元交互注意力 (MIA) 学习图文特征之间的交互关系从而缩小模态间差距. 再次, 由于图像和文本之间信息的不平等, 本文提出了前景增强判别器 (FED) 来过滤背景信息并且增强前景特征. 最后, 在 3 个流行基准数据集 CUHK-PEDES、ICFG-PEDES 及 RSTPReid 上进行了消融实验以及与现有的 SOTA 方法进行比对实验, 实验结果证明了我们提出的模型框架在基于文本的人物检索方面的可行性和有效性.

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] Yang WX, Yan Y, Chen S, Zhang XK, Wang HZ. Multi-scale generative adversarial network for person re-identification under occlusion. Ruan Jian Xue Bao/Journal of Software, 2020, 31(7): 1943–1958 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5932.htm [doi: 10.13328/j.cnki.jos.005932]

[3] Su C, Zhang SL, Xing JL, Gao W, Tian Q. Deep attributes driven multi-camera person re-identification. In: Proc. of the 14th European Conf. on Computer Vision (ECCV). Amsterdam: Springer, 2016. 475–491. [doi: 10.1007/978-3-319-46475-6_30]

[4] Vaquero DA, Feris RS, Tran D, Brown L, Hampapur A, Turk M. Attribute-based people search in surveillance environment. In: Proc. of the 2009 Workshop on Applications of Computer Vision (WACV). Snowbird: IEEE, 2009. 1–8. [doi: 10.1109/WACV.2009.5403131]

[5] 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]

[6] 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]

[7] 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 (TOMM), 2020, 16(2): 51. [doi: 10.1145/3383184]

[8] 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]

[9] 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 Int'l Conf. on Computer Vision (ICCV). Venice: IEEE, 2017. 1908–1917. [doi: 10.1109/ICCV.2017.209]

[10] 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]

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

[12] 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]

[13] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015. 3128–3137. [doi: 10.1109/CVPR.2015.7298932]

[14] Lee KH, Chen X, Hua G, Hu HD, He XD. Stacked cross attention for image-text matching. In: Proc. of the 15th European Conf. on Computer Vision (ECCV). Munich: Springer, 2018. 212–228. [doi: 10.1007/978-3-030-01225-0_13]

[15] Hou RB, Ma BP, Chang H, Gu XQ, Shan SG, Chen XL. VRSTC: Occlusion-free video person re-identification. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 7176–7185. [doi: 10.1109/CVPR.2019.00735]

[16] Song CF, Huang Y, Ouyang WL, Wang L. Mask-guided contrastive attention model for person re-identification. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 1179–1188. [doi: 10.1109/CVPR.2018.00129]

[17] Sarfraz MS, Schumann A, Eberle A, Stiefelhagen R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 420–429. [doi: 10.1109/CVPR.2018.00051]

[18] 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]

[19] 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]

[20] 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.

[21] 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]

[22] 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]

[23] 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 27th ACM Int'l Conf. on Multimedia. Nice: ACM, 2019. 665–673. [doi: 10.1145/3343031.3350991]

[24] Ding ZF, Ding CX, Shao ZY, Tao DC. Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv:2107.12666, 2021.

[25] Chen YC, Li LJ, Yu LC, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu JJ. UNITER: Universal image-text representation learning. In: Proc. of the 16th European Conf. on Computer Vision (ECCV). Glasgow: Springer, 2020. 104–120. [doi: 10.1007/978-3-030-58577-8_7]

[26] Jia C, Yang YF, Xia Y, Chen YT, Parekh Z, Pham H, Le Q, Sung YH, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. 2021. 4904–4916.

[27] Antol S, Agrawal A, Lu JS, Mitchell M, Batra D, Zitnick CL, Parikh D. VQA: Visual question answering. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). Santiago: IEEE, 2015. 2425–2433. [doi: 10.1109/ICCV.2015.279]

[28] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. 2021. 8748–8763.

[29] Yan SL, Dong N, Zhang LY, Tang JH. CLIP-driven fine-grained text-image person re-identification. IEEE Trans. on Image Processing, 2023, 32: 6032–6046. [doi: 10.1109/TIP.2023.3327924]

[30] Chen WJ, Yao LL, Jin Q. Rethinking benchmarks for cross-modal image-text retrieval. In: Proc. of the 46th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Taipei: ACM, 2023. 1241–1251. [doi: 10.1145/3539618.3591758]

[31] Zhou KY, Yang JK, Loy CC, Liu ZW. Learning to prompt for vision-language models. Int'l Journal of Computer Vision, 2022, 130(9): 2337–2348. [doi: 10.1007/s11263-022-01653-1]

[32] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv:1508.07909, 2016.

[33] Wu YS, Yan ZZ, Han XG, Li GB, Zou CQ, Cui SG. LapsCore: Language-guided person search via color reasoning. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision (ICCV). Montreal: IEEE, 2021. 1604–1613. [doi: 10.1109/ICCV48922.2021.00165]

[34] Yan SL, Tang H, Zhang LY, Tang JH. Image-specific information suppression and implicit local alignment for text-based person search. IEEE Trans. on Neural Networks and Learning Systems, 2024, 35(12): 17973–17986. [doi: 10.1109/TNNLS.2023.3310118]

[35] Wang ZJ, Zhu AC, Xue JY, Wan XL, Liu C, Wang T, Li YF. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 1984–1992. [doi: 10.1145/3503161.3548166]

[36] Han X, He S, Zhang L, Xiang T. Text-based person search with limited data. arXiv:2110.10807, 2021.

[37] Li SP, Cao M, Zhang M. Learning semantic-aligned feature representation for text-based person search. In: Proc. of the 2022 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 2724–2728. [doi: 10.1109/ICASSP43922.2022.9746846]

[38] 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]

[39] Wang ZJ, Zhu AC, Xue JY, Wan XL, Liu C, Wang T, Li YF. CAIBC: Capturing all-round information beyond color for text-based person retrieval. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 5314–5322. [doi: 10.1145/3503161.3548057]

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

[41] Shao ZY, Zhang XY, Fang M, Lin ZF, Wang J, Ding CX. Learning granularity-unified representations for text-to-image person re-identification. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 5566–5574. [doi: 10.1145/3503161.3548028]

[42] Shu XJ, Wen W, Wu HQ, Chen KY, Song YR, Qiao RZ, Ren B, Wang X. See finer, see more: Implicit modality alignment for text-based person retrieval. In: Proc. of the 2022 European Conf. on Computer Vision (ECCV). Tel Aviv: Springer, 2022. 624–641. [doi: 10.1007/

978-3-031-25072-9_42]

[43] Li JN, Li DX, Xiong CM, Hoi S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proc. of the 39th Int'l Conf. on Machine Learning. 2022. 12888–12900.

[44] Bao LP, Wei LH, Zhou WG, Liu L, Xie LX, Li HQ, Tian Q. Multi-granularity matching Transformer for text-based person search. IEEE Trans. on Multimedia, 2024, 26: 4281–4293. [doi: 10.1109/TMM.2023.3321504]

[45] Kim W, Son B, Kim I. ViLT: Vision-and-language Transformer without convolution or region supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. 2021. 5583–5594.

[46] Li JN, Selvaraju RR, Gotmare AD, Joty S, Xiong CM, Hoi SCH. Align before fuse: Vision and language representation learning with momentum distillation. In: Proc. of the 35th Int'l Conf. on Neural Information Processing Systems. Curran Associates Inc., 2021. 9694–9705.

附中文参考文献:

[2] 杨婉香, 严严, 陈思, 张小康, 王菡子. 基于多尺度生成对抗网络的遮挡行人重识别方法. 软件学报, 2020, 31(7): 1943–1958. http://www.jos.org.cn/1000-9825/5932.htm [doi: 10.13328/j.cnki.jos.005932]

孙锐(1976－), 男, 博士, 教授, CCF 专业会员, 主要研究领域为机器学习, 计算机视觉.

陈龙(2000－), 男, 硕士生, 主要研究领域为图像信息处理, 计算机视觉.

杜云(1998－), 女, 硕士生, 主要研究领域为图像信息处理, 计算机视觉.

张旭东(1966－), 男, 博士, 教授, 主要研究领域为智能信息处理, 机器视觉.