计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(1):153−160 [doi: 10.15888/j.cnki.csa.009741] [CSTR: 32024.14.csa.009741]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

改进 CLIP-ReID 的跨模态行人重识别①

贾军营, 杨芯茹, 杨海波, 徐　展

(沈阳工业大学 信息科学与工程学院, 沈阳 110870)
通信作者: 杨芯茹, E-mail: yangxr@smail.sut.edu.cn

摘　要: 由图像到文本的跨模态行人重识别中缩小模态间差异一直是一个主要挑战, 针对该问题, 研究了一种基于CLIP-ReID (contrastive language-image pretraining-person re-identification) 的改进方法. 引入了上下文调整网络模块和跨模态注意力机制模块. 上下文调整网络模块对图像特征进行深层次的非线性转换, 并有效地与可学习上下文向量相结合, 增强图像和文本间的语义关联性. 跨模态注意力机制模块通过对图像和文本特征进行动态加权和融合, 使得模型能够在处理一个模态的信息时考虑到另一模态, 提升模型在不同模态间的交互. 该方法分别在 MSMT17、Market1501、DukeMTMC 公共数据集上进行了评估, 实验结果在 mAP 值上分别提升了 2.2%、0.5%、0.4%; 在 R1 值上分别提升了 1.1%、0.1%、1.2%. 结果表明所提方法有效地提升了行人重识别的精度.

关键词: 行人重识别; 跨模态; 注意力机制

引用格式: 贾军营,杨芯茹,杨海波,徐展.改进 CLIP-ReID 的跨模态行人重识别.计算机系统应用,2025,34(1):153–160. http://www.c-s-a.org.cn/1003-3254/9741.html

Cross-modal Person Re-identification Based on Improved CLIP-ReID

JIA Jun-Ying, YANG Xin-Ru, YANG Hai-Bo, XU Zhan

(School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China)

Abstract: Narrowing the difference between modalities is always challenging in cross-modal person re-identification from images to texts. To address this challenge, this study proposes an improved method based on contrastive language-image pretraining-person re-identification (CLIP-ReID) by integrating a context adjustment network module and a cross-modal attention mechanism module. The former module performs a deep nonlinear transformation on image features and effectively combines with learnable context vectors to enhance the semantic relevance between images and texts. The latter module dynamically weights and fuses features from images and texts so that the model can take into account the other modality when processing the information of one modality, improving the interaction between different modalities. The method is evaluated on three public datasets. Experimental results show that the mAP on the MSMT17 dataset is increased by 2.2% and R1 is increased by 1.1%. On the Market1501 dataset, there is a 0.5% increase in mAP and a 0.1% rise in R1. The DukeMTMC dataset sees a 0.4% enhancement in mAP and a 1.2% increase in R1. The results show that the proposed method effectively improves the accuracy of person re-identification.

Key words: person re-identification; cross-modal; attention mechanism

行人重识别 (person re-identification, ReID) 技术作为计算机视觉和人工智能领域的热点之一[1], 该技术的研究和应用受到了广泛的关注. 行人重识别技术[2]是通过算法来判断在不同摄像头或不同时间段拍摄到的行

① 基金项目: 辽宁省应用基础研究项目 (2022JH2/101300243); 2022 年度沈阳市科学技术计划“揭榜挂帅”产业共性技术项目 (22-316-1-07)
收稿时间: 2024-06-24; 修改时间: 2024-07-18; 采用时间: 2024-07-25; csa 在线出版时间: 2024-11-28
CNKI 网络首发时间: 2024-11-29

人图像是否为同一个行人. 室外场景下的行人目标重识别, 因为其复杂的环境条件和高度的应用需求, 成为智能视频监控系统的关键技术之一[3], 在多个领域都有着广泛的应用前景和巨大的社会价值[4]. 例如, 在公共安全、治安管理等领域通过行人重识别技术, 可以定位和追踪特定目标, 有效防范安全威胁; 在智慧交通、零售等领域提供强有力的数据支持, 改善民众的生活质量. 然而, 室外场景的不确定性和复杂性, 如光照条件的剧烈变化[5]、不同天气状况下的视觉效果差异、遮挡[6]、行人多样姿态[7]和相似服装以及摄像头角度[8]和分辨率差异[9]等因素的影响, 为行人的识别带来挑战.

基于图像的行人重识别可以分为单模态和跨模态两大类. 单模态行人重识别技术尽管在多个应用场景中已经证明其有效性, 但仍受限于依赖单一视觉信息的本质. 文本信息能提供视觉信息难以捕捉的细节, 如衣着颜色或者行人行为等. 因此, 图像到文本的跨模态行人重识别技术旨在通过结合视觉信息和语言信息来克服仅依赖图像的传统行人重识别技术所面临的限制. 对图像中的行人信息进行文本描述, 度量图像和文本描述之间的相似性, 从而实现跨模态的识别与匹配[10]. 这种跨模态技术可以弥补单一模态难以捕捉的信息. 尽管如此, 跨模态行人重识别技术同样遇到多种挑战[11], 同单一模态一样的光照变化、遮挡等问题, 还有图像与文本在表达方式上存在的本质差异; 文本描述可能存在多样性和模糊性, 同一图像可以有多种不同的文本描述, 同样, 同一文本描述也可能对应着多个图像等. 这些都增加了匹配的难度.

综上所述, 本文的主要工作如下.

(1) 引入上下文调整网络模块, 优化和调整图像特征与文本特征之间的相互作用, 通过动态调整上下文向量来更准确地捕捉图像内容与文本描述之间的语义关联, 提升了模型对于跨模态信息的融合与理解能力.

(2) 引入跨模态注意力机制模块, 加强了模型对于图像和文本中关键信息的捕捉与融合. 通过这一机制, 模型在处理来自不同模态的输入时, 能够更加聚焦于关键信息, 从而缩小视觉信息与文本信息之间的差异, 提升模型的泛化性.

1 相关工作

1.1 单模态行人重识别方法

深度学习技术的快速发展推动许多学者将其应用于行人重识别研究, 从而取得了一些显著的成果. He 等人[12]设计了拼图模块 (jigsaw patch module, JPM) 以及侧面信息嵌入 (side information embedding, SIE), 这是一种纯基于 Transformer 结构, 增强了模型的鲁棒性和对摄像头或视角变化的适应性. Zhou 等人[13]使用对比损失, 通过判别性学习和一致性注意力正则化来增强模型在不同尺度上对于重要区域的关注能力. Hermans 等人[14]在 2017 年提出的困难三元组损失的方法, 一种改进的三元组损失函数, 每次输入选择与当前样本最不相似的正样本和最相似的负样本构建训练样本来提高模型的鲁棒性和泛化能力. Vaswani 等人[15]首次提出了 Transformer 模型, 通过自注意力机制能够动态地调整元素间的相关性权重, 更加灵活地捕获序列间的依赖性, 特别是长距离依赖. 提升了模型的精度和效率. Xu 等人[16]提出了利用视觉注意力机制的图像描述生成技术, 使得模型可以自动学习图像内容的描述, 生成的每个单词与图像不同部分的关注程度有关, 更好地理解图像并生成更准确的描述.

1.2 跨模态行人重识别方法

由文本到图像的跨模态行人重识别指的是将文本描述和图像信息相结合来对行人进行识别和匹配. 文本信息是关于行人外观特征的描述, 比如服装颜色、行为特征等信息; 而图像是指行人的图片或视频帧. 这种方法不仅依赖于视觉信息, 还结合了文本描述所提供的额外的上下文信息. Li 等人[17]提出了一个针对文本-视觉匹配问题的身份感知两阶段框架. 首先使用卷积神经网络-长短期记忆 (convolutional neural network-long short term memory, CNN-LSTM) 网络筛选容易出错的配对, 其次利用潜在的共同注意机制改进匹配结果. 增强了模型的鲁棒性. Li 等人[18]提出了一个基于 Transformer 的语言-行人的搜索框架, 通过将行人图像垂直分割成若干个区域并匹配语言描述中的词汇, 实现语言与视觉特征之间的共同注意的区域匹配. Radford 等人[19]联合训练图像编码器和文本编码器, 通过大规模的图像和文本对进行对比学习. 使得模型能够生成与图像内容相匹配的语言描述, 实现了从自然语言监督中学习可迁移的视觉模型, 也就是说模型到各种下游任务的零样本迁移. 但是在一些特定领域或细粒度的视觉任务中, 泛化能力仍然有限. Zhou 等人[20]提出了一种称为上下文优化 (context optimization, CoOp) 的方法, 将自然语言处理领域的提示学习引入到视觉领

域中. 将手动的提示模板优化为自动学习最优模板, 即使在少量标注样本的情况下, 也能有效地适配于新任务中, 目的是提高模型对不同类别的泛化能力. 但对于不同数据集的泛化性和迁移能力仍然是一个需要进一步研究的问题. Li 等人[21]提出了一种基于 CLIP 的重识别方法, 充分利用 CLIP 的跨模态能力, 设计了一种两阶段的训练策略, 解决了重识别任务中缺乏具体文本描述的问题, 在人员或车辆重识别的数据集上验证了所提策略的有效性, 但是在泛化能力等问题上还需要进一步的改进. Zhao 等人[22]提出了一个跨模式互训 (cross-modal mutual training, CMMT) 框架, 通过聚类方法为视觉和文本实例生成伪标签. CMMT 提供了一个互伪标签精炼模块, 借助一种模态的聚类结果来优化另一种模块的结果. Yao 等人[23]提出了一个大规模细粒度交互式视觉-语言预训练 (fine-grained interactive language-image pre-training, FLIP) 框架, 在保持高效的同时, 实现了图像和文本间更细粒度的交互和对齐. 综上, 相比于单模态的行人重识别方法, 有以下 3 点好处: 首先, 文本描述可以挖掘视觉图像可能被忽略或很难捕捉的细节信息, 能够获取更加丰富的信息. 其次, 利用文本描述额外提供的信息在面对如光照变化、遮挡或者摄像头角度问题等复杂环境下, 在一定程度上能够克服这些问题, 提高模型对复杂场景的适应能力. 最后, 单模态行人重识别会存在一定的局限性, 跨模态行人重识别通过融合两种模态信息, 相互弥补各自的不足, 实现更全面和准确的行人识别.

2 方法

2.1 CLIP-ReID 方法概述

CLIP 是由 OpenAI 开发的一种多模态预训练模型, 在通过大规模的图像和文本对进行预训练, 学习视觉内容和自然语言之间的共同关系表示. 对于常规的分类任务, CLIP 通过将数据集的具体标签转化为文本描述来工作, CoOp 通过固定预训练的参数不变, 为不同任务引入了一个可学习的提示. 但是在重识别任务面临着一个挑战, 标签是以索引形式存在, 而非明确的文本描述. 因此 CLIP-ReID 通过预训练一组可学习的文本来弥补文本信息的不足, 采用双阶段训练方法, 如图 1(a) 所示.

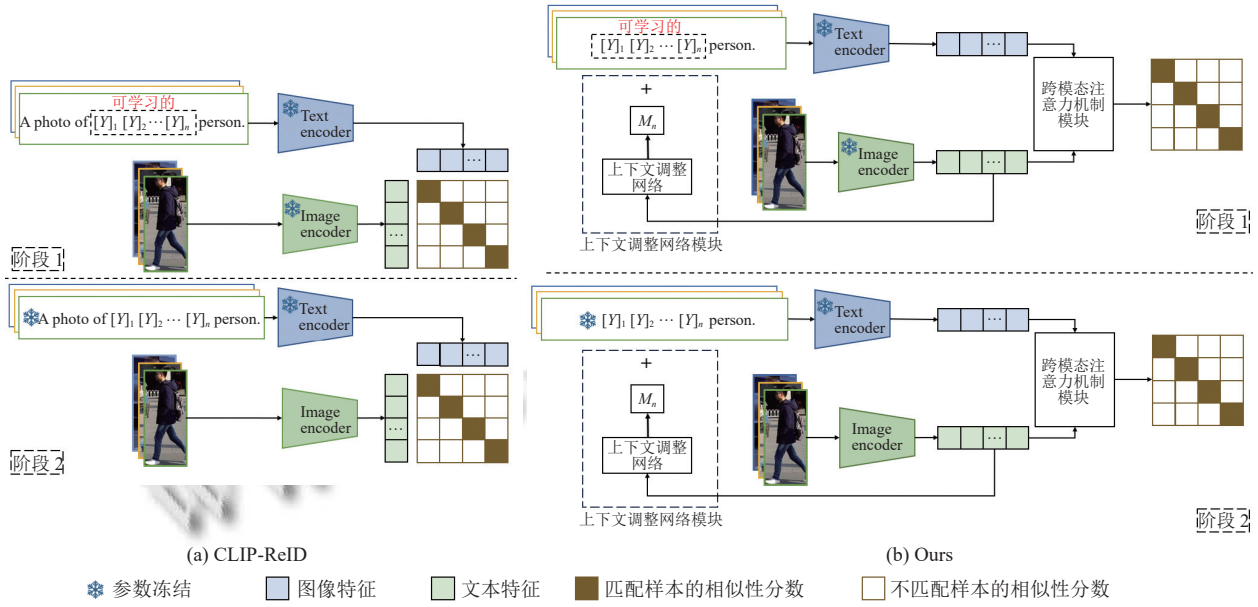

图 1 网络整体结构图

模型主要由视觉编码器和文本编码器两个部分构成, 这两部分协同工作, 同时理解图像和文本描述, 将它们映射到同一个嵌入空间中. 图像编码器可以是基于卷积神经网络的模型, 如 ResNet-50, 也可以是基于 Transformer 的模型, 如 Vision Transformer (ViT), 用于处理输入的图像并提取图像的视觉特征; 文本编码器是基于 Transformer 结构的, 用于处理对图像描述的文本, 提取文本的语义特征. 具体来说, 以“A photo of a $[Y]_1[Y]_2\cdots[Y]_n$ person.” 描述为例. 其中每个$[Y]_N$ ($N\in\{1,2,\cdots,n\}$) 是一个与词嵌入相同的可学习文本标记. n 表示可学习文本标记的数量. 将文本中每个单词转换为唯一的数字 ID, 然后每个 ID 映射为 512 维度的词

嵌入. 对于文本, 按照使用词汇量为 49 408 的小写字节对编码来标记文本. 每个文本序列以[SOS]标记开始和以[EOS]标记结尾. 通过由 8 个注意力头的 12 层模型处理后, 将[EOS]标记的输出作为文本的特征表示, 经过层归一化后通过线性变换映射到跨模态嵌入空间中. 通过对比学习, 最大化相关图像文本对的相似度, 最小化不相关图像文本的相似度. 该方法在第 1 个训练阶段固定文本编码器和图像编码器参数, 优化一组可学习的文本生成文本特征, 第 2 个训练阶段用文本特征来优化图像编码器参数.

尽管 CLIP-ReID 在利用视觉-语言模型进行图像重识别方法取得了显著进展, 但由于多模态需要处理图像和文本两种不同模态的信息, 模态间就会存在差异性问题. 因此为了缩小模态间差异, 从而提升模型的泛化性, 本文引入了上下文调整网络模块和跨模态注意力机制模块, 这两个模块加强图像特征与文本特征之间的相关性, 使得模型能够更有效地理解和利用这两种信息. 如图 1(b) 所示, 整个网络有两种类型的输入: 文本和图像, 文本输入初始化定义为“$[Y]_1[Y]_2\cdots[Y]_n$ person.”, 通过基于 Transformer 架构的文本编码器提取文本特征, 图像编码器选用基于 Vision Transformer 的架构来提取图像特征.

2.2 上下文调整网络模块

如图 2 所示, 通过引入上下文调整网络 (Context-Net) 来提升文本提示与图像特征之间的匹配度, 使得文本提示在自适应反映图像特征变化的基础上, 获取更多的信息, 能够更加准确地反映图像特征, 从而优化文本提示的生成过程, 提高匹配的精度. 具体来说, 该模块采用了一个双层全连接网络, 用于处理从图像编码器提取的特征. 首先, 该网络的第 1 个线性层将输入特征的维度从原来的 512 维扩大了 2 倍, 这一扩展过程能够捕捉到更多的图像细节, 为后续的处理提供了丰富的信息. 接着, 将扩展后的特征通过 ReLU 激活函数, 不仅为了增加非线性, 还能有效地过滤掉无关的信息, 使模型能够专注于更重要的特征, 从而增强特征表示的丰富性和表达能力. 在经过激活函数处理后, 特征进入到第 2 个线性层, 将特征维度重新压缩回原来的 512 维, 这一过程确保在扩展过程中捕捉到的丰富信息能够有效地映射回原始的特征空间. 通过这种方式, 处理后的特征既保留了扩展过程中获取的细节信息, 又与原始的特征维度保持一致. 调整后的图像特征 M 作为一个附加的输入参数用于动态调整文本提示的生成过程, 最终生成的上下文向量表示为 $C_N=[Y]_N+M$, 其中 $M=g(I_f,\gamma)$, I_f 表示图像特征, γ 表示网络的参数. 使其在自适应地反映图像内容变化的过程中具有更多的信息.

通过这种方式, 模型将调整后的图像特征引入到文本生成过程中, 提高生成文本的相关性, 文本与图像匹配的准确性, 并增强模型的适应性和鲁棒性.

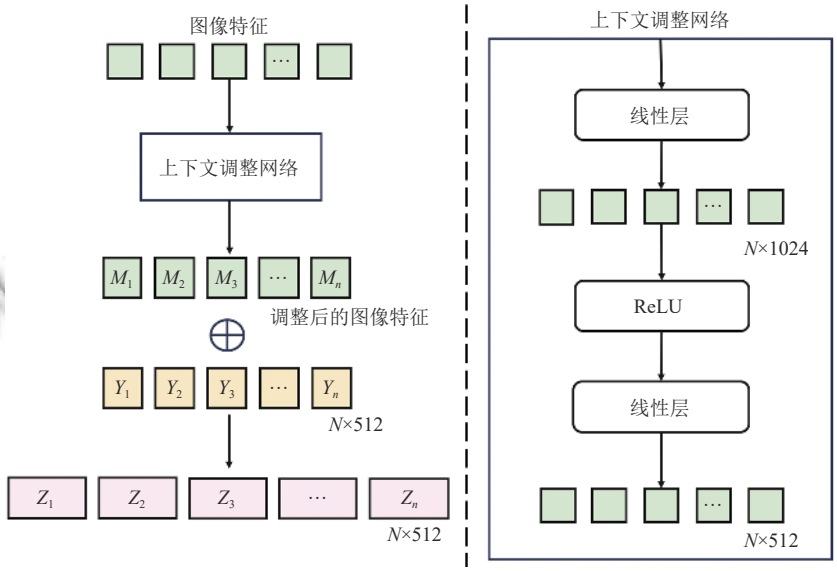

图 2 上下文调整网络模块图

2.3 跨模态注意力机制模块

如图 3 所示, 跨模态注意力机制主要是融合和增强图像和文本之间的特征表示. 该机制首先将图像特征和文本特征分别通过两个线性层映射到共同的隐藏

特征空间, 以便进行跨模态交互. 然后, 对映射后的图像特征和文本特征分别应用自注意力机制. 自注意力机制通过计算输入特征内部的相互关系, 捕捉序列中的长距离依赖关系, 使得在跨模态注意力计算之前更加适合. 接着, 计算自注意力后的图像特征和文本特征之间的点积, 生成一个注意力分数矩阵. 通过 *Softmax* 函数进行归一化处理, 允许模型动态地聚焦于文本特征与图像特征间相关的部分, 得到每个图像特征向量对所有文本特征向量的注意力分布.

随后, 对文本特征进行加权求和, 形成注意力加权的文本表示. 这些加权特征与原始图像投影特征进行结合, 形成残差连接, 生成融合了两种模态信息的输出特征. 这些特征在接下来进行相似度计算来评估图像和文本之间的匹配程度.

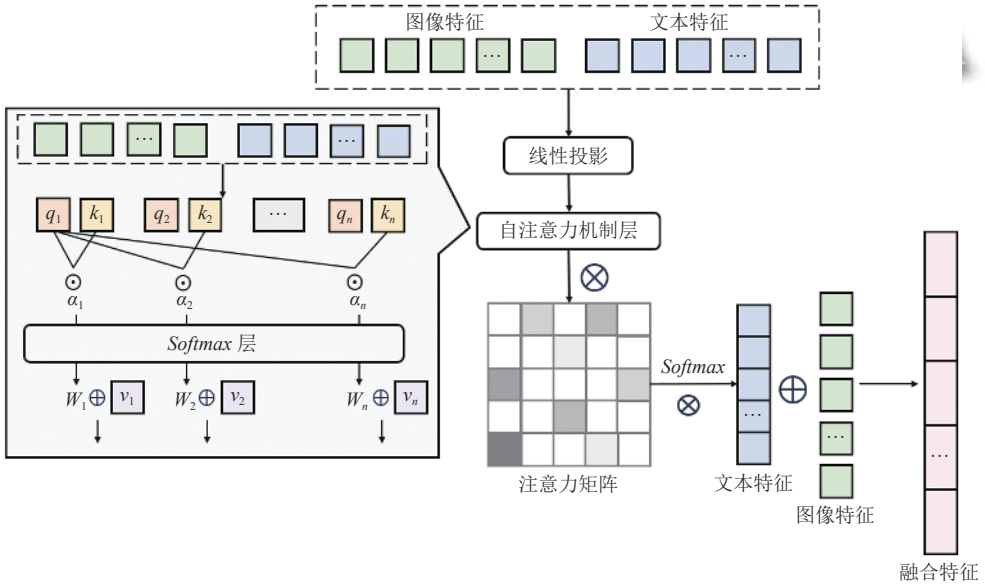

图 3 跨模态注意力机制模块图

这一过程增强图像与文本特征之间的交互, 缩小图像与文本之间模态的差异, 使得模型能够动态地捕捉与当前图像最相关的文本信息, 提高准确性. 公式如下:

$$O = W_o\left(Softmax\left((W_iA_i)\cdot(W_tA_t)^{\mathrm{T}}\right)\cdot W_tA_t + W_iA_i\right) + b_o \tag{1}$$

其中, O、i 和 t 分别表示输出特征以及输入的图像和文本特征, W_i、W_t 和 W_o 是线性变换的权重矩阵, 用于计算注意力权重, b_o 表示输出特征的偏置项. A_i 表示自注意力后的图像特征, A_t 表示自注意力后的文本特征.

2.4 损失函数

在第 1 个训练阶段, 通过自动混合精度和梯度缩放来提高训练效率和模型的稳定性. 图像编码器和文本编码器参数是不变的, 优化上下文向量 $[Y]_N$ 以及上下文调整网络的参数 θ. 使用图像到文本对比损失 L_{I2T} 和文本到图像对比损失 L_{T2I}. 给定一批大小为 N 的图像集合, i 表示一个图像的索引, $i\in\{1,\cdots,N\}$, L_{I2T} 计算公式为:

$$L_{\mathrm{I2T}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(\frac{sim(I_i,t_i)}{T}\right)}{\sum_{j=1}^{N}\exp\left(\frac{sim\left(I_i,t_j\right)}{T}\right)}\right) \tag{2}$$

其中, N 为批次中样本的数量, $sim(I_i,t_i)=I_i\cdot t_i$ 为图像特征与其对应的文本描述之间的相似度, t_j 表示批次中任意一个文本特征. T 为温度参数, 控制相似度尺度.

对于 L_{T2I} 的计算, 由于一个 ID 可能具有相同的文本描述, 因此用 T_{Zi} 表示. 一个批次中不同的图像可能属于同一个人, 因此对于同一个 ID 来说 T_{Zi} 可能对应多个正例图像. 因此 L_{T2I} 的计算公式为:

$$L_{\mathrm{T2I}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\sum_{p\in P(i)}\exp\left(\frac{sim\left(T_{Zi},I_p\right)}{T}\right)}{\sum_{j=1}^{N}\exp\left(\frac{sim\left(T_{Zi},I_j\right)}{T}\right)}\right) \tag{3}$$

其中, I_p 为与文本 T_{Zi} 相匹配的图像特征集合 $P(i)$ 中的一个图像特征. 通过最小化损失, 优化上下文向量$\{[Y]_N\}_{N=1}^{n}$以及上下文调整网络的参数 θ. 综上, 第 1 个训练阶段的总损失如下:

$$L_{\text{stage1}} = L_{\text{I2T}} + L_{\text{T2I}} \tag{4}$$

第 2 个训练阶段, 用来优化图像编码器的参数. 采用标签平滑的 ID 损失 L_{ID}, 三元组损失 L_{Tri} 以及图像到文本的交叉熵损失. L_{ID} 计算为:

$$L_{\text{ID}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{C}\tilde{y}_{n,k}\log\left(p_{n,k}\right) \tag{5}$$

其中, $\tilde{y}_{n.k} = (1-\alpha)\delta_{n,k} + \alpha$表示经过平滑处理后的目标分布. N 表示批次中样本数量, C 表示类别数量, $p_{n,k}$ 表示预测第 n 个样本属于类别 k 的概率. L_{Tri} 计算为:

$$L_{\text{Tri}} = \max\left(d_{\text{ap}} - d_{\text{an}} + m, 0\right) \tag{6}$$

其中, d_{ap} 表示正例对之间的距离, d_{an} 表示负例对之间的距离, m 表示边界值, 确保正负样本对之间有足够的距离. L_{I2Tce} 计算为:

$$L_{\text{I2Tce}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}\tilde{y}_{i,k}\log\left(\frac{\exp\left(\frac{sim(I_i,t_k)}{T}\right)}{\sum_{j=1}^{C}\exp\left(\frac{sim\left(I_i,t_j\right)}{T}\right)}\right) \tag{7}$$

在 L_{I2Tce} 中对$\tilde{y}_{n,k}$采用了标签平滑技术. 综上, 第 2 个训练阶段的总损失如下:

$$L_{\text{stage2}} = L_{\text{ID}} + L_{\text{Tri}} + L_{\text{I2Tce}} \tag{8}$$

总的来说, 整个网络结构通过两个训练阶段进行训练, 首先优化可学习上下文向量和上下文调整网络参数, 然后优化图像编码器参数. 通过对比学习来训练模型, 使得图像和文本之间能够更好地捕捉相关信息, 从而提升行人重识别的准确性和泛化性.

3 实验

3.1 数据集和评估方案

本文在行人重识别的 3 个公共数据集上进行了评估, 这些数据集包括 MSMT17[24], Market1501[25]以及 DukeMTMC-reID[26]数据集. 各数据集的详细信息如表 1 所示. 本文采用了常用的累计匹配曲线 (cumulative match characteristic, CMC) 的 Rank-1 (R1) 和平均准确率 (mean average precision, mAP) 作为评价指标. R1 表示在检索排序结果中, 查询图像的正确匹配出现在第 1 位的结果.

表 1 本文使用的数据集详细信息

Dataset	ID	Images	Camera+View
MSMT17	4101	126 441	15
Market1501	1501	32 668	6
DukeMTMC-reID	1404	36 411	8

3.2 实验细节

实验环境采用了 Windows Server 2019 操作系统, 基于 PyTorch 深度学习框架, 在 NVIDIA GeForce RTX 3090 显卡上实现模型.

所有图像大小均调整为[384, 192]. 使用 ImageNet 的均值和标准差. 在第 1 个训练阶段, 使用 Adam 作为优化器, 批次大小为 64, 初始学习率设置为 0.000 35, 为了在训练过程中平滑地调整学习率, 提高模型的稳定性, 使用余弦退火的策略. 通过线性预热方法在前 10 个 Epoch 内逐渐增加到基础学习率. 最小学习率设置为 1E–6, 权重衰减参数设为 1E–4, 此阶段的最大训练次数为 120. 优化可学习的上下文向量以及上下文调整网络参数. 第 2 个训练阶段进一步优化图像编码器参数, 继续采用 Adam 优化器, 批次大小为 64, 基础学习率降低至 0.000 005, 最大训练次数为 60. 在第 30 个和第 50 个 Epoch 时乘以衰减系数 0.1.

3.3 对比实验

如表 2 所示, 将本文的研究方法与其他现有方法在 MSMT17、Market1501 和 DukeMTMC 这 3 个公共数据集上进行比较的结果. 本文的方法在 MSMT17 公共数据集上达到了 75.6% 的 mAP 和 89.8% 的 R1, 在 Market1501 数据集上达到了 90.1% 的 mAP 和 95.6% 的 R1, 在 DukeMTMC 数据集上达到了 82.9% 的 mAP 和 91.2% 的 R1. 对于其他方法以及 CLIP-ReID 方法在 3 个公共数据集上都有着一定程度的提升.

3.4 消融实验

为了评估添加的两个模块对网络结构是有效的, 本文在 MSMT17 数据集上进行了消融实验的分析, 如表 3 所示. 首先, 跨模态注意力机制加入网络之后, 模型的 mAP 提升了 2.0%, R1 提升了 0.6%, 该模块使模型能够有效地利用不同模态间的补充信息, 加强了模态间的交互, 从而提升模型的性能. 在此基础上, 将上下文调整网络模块加入网络, mAP 又提升了 0.2%,

R1 提升了 0.5%, 该模块对图像特征进行了优化, 通过扩展特征维度、引入非线性变换等, 提高了特征的表达能力使其与上下文信息的有效结合. 这些结果证明了这两个模块的有效性.

表 2 MSMT17、Market1501 和 DukeMTTC 数据集上与其他方法对比的实验结果 (%)

Methods	MSMT17		Market1501		DukeMTMC	
	mAP	R1	mAP	R1	mAP	R1
PCB	40.4	68.2	81.6	93.8	69.2	83.3
ABD-Net	60.8	82.3	88.3	95.6	60.8	82.3
APD	61.2	82.4	89.1	95.8	81.1	90.7
TransReID	66.6	84.6	88.8	95.0	81.8	90.4
TransReID+SIE+OLP	69.4	86.2	89.5	95.2	82.6	90.7
AAformer	63.2	83.6	87.7	95.4	80.0	90.1
DCAL	64.0	83.1	87.5	94.7	80.1	89.0
CLIP-ReID	73.4	88.7	89.6	95.5	82.5	90.0
Ours	75.6	89.8	90.1	95.6	82.9	91.2

表 3 MSMT17 数据集上的消融实验 (%)

Methods	mAP	R1
CLIP-ReID	73.4	88.7
+跨模态注意力机制	75.4	89.3
+上下文调整网络模块+跨模态注意力机制模块	75.6	89.8

对文本初始化定义中可学习的模糊文本的位置以及可学习标记参数 n 的数量进行分析, 如表 4 所示. 证明可学习的模糊文本的位置位于开始, 以及 n 为 8 是最好的结果.

表 4 MSMT17 数据集上模糊文本位置和参数 n 的分析 (%)

文本初始化	n=4		n=8		n=16	
	mAP	R1	mAP	R1	mAP	R1
A photo of $[Y]_1[Y]_2\cdots[Y]_n$ person.	75.4	89.5	75.5	89.6	75.5	89.6
$[Y]_1[Y]_2\cdots[Y]_n$ person.	75.3	89.4	75.6	89.8	75.4	89.7
A photo of a person, $[Y]_1[Y]_2\cdots[Y]_n$.	75.4	89.4	75.5	89.6	75.5	89.8

3.5 可视化结果

如图 4 所示, 在 MSMT17 数据集上可视化检索结果. 图中预测正确的外框为绿色, 预测错误的外框为红色.

4 结论与展望

本文提出了一种改进的 CLIP-ReID 方法, 通过引入上下文调整网络模块和跨模态注意力机制模块, 有效缩小了图像与文本之间的模态差异. 上下文调整网络模块利用提取后的图像特征调整可学习的上下文向量, 加强了图像与文本特征的语义关联; 跨模态注意力机制模块使得模型能够更准确地理解图像内容与文本描述之间的关系, 提升模态间的信息融合能力. 实验结果表明, 所提方法在 MSMT17、Market1501 和 DukeMTMC 公共数据集上的 mAP 和 R1 值均有所提升. 本文方法可应用于视频监控的人员检索和身份验证等任务, 同时在智能安防和公共安全等领域具有广泛的应用前景. 未来的研究工作将考虑低分辨率的问题, 以进一步优化算法.

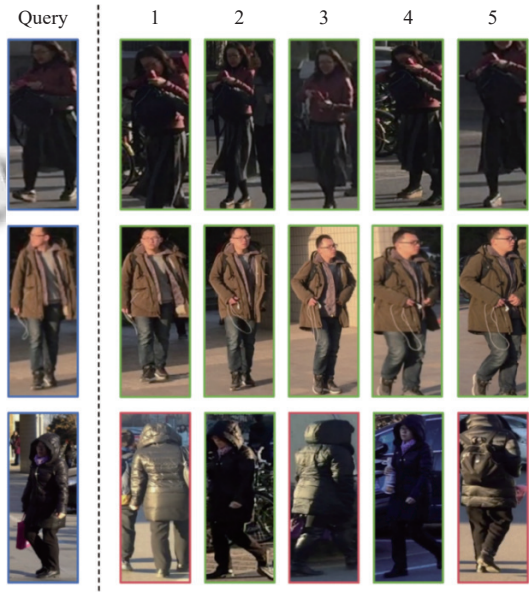

图 4 可视化检索结果图

参考文献

1 杨锋, 许玉, 尹梦晓, 等. 基于深度学习的行人重识别综述. 计算机应用, 2020, 40(5): 1243–1252. [doi: 10.11772/j.issn.1001-9081.2019091703]

2 Leng QM, Ye M, Tian Q. A survey of open-world person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(4): 1092–1108. [doi: 10.1109/TCSVT.2019.2898940]

3 胡正平, 张敏姣, 李淑芳, 等. 智能视频监控系统中行人再识别技术研究综述. 燕山大学学报, 2019, 43(5): 377–393. [doi: 10.3969/j.issn.1007-791X.2019.05.001]

4 李擎, 胡伟阳, 李江昀, 等. 基于深度学习的行人重识别方法综述. 工程科学学报, 2022, 44(5): 920–932. [doi: 10.13374/j.issn2095-9389.2020.12.22.004]

5 Huang YK, Zha ZJ, Fu XY, *et al*. Illumination-invariant person re-identification. Proceedings of the 27th ACM International Conference on Multimedia. Nice: ACM, 2019. 365–373. [doi: 10.1145/3343031.3350994]

6 Huang HJ, Li DW, Zhang Z, *et al*. Adversarially occluded

samples for person re-identification. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 5098–5107. [doi: 10.1109/CVPR.2018.00535]

7 Cho YJ, Yoon KJ. Improving person re-identification via pose-aware multi-shot matching. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1354–1362. [doi: 10.1109/CVPR.2016.151]

8 Bak S, Zaidenberg S, Boulay B, *et al*. Improving person re-identification by viewpoint cues. Proceedings of the 11th IEEE International Conference on Advanced Video and Signal Based Surveillance. Seoul: IEEE, 2014. 175–180. [doi: 10.1109/AVSS.2014.6918664]

9 Li X, Zheng WS, Wang XJ, *et al*. Multi-scale learning for low-resolution person re-identification. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3765–3773. [doi: 10.1109/ICCV.2015.429]

10 赵才荣, 齐鼎, 窦曙光, 等. 智能视频监控关键技术: 行人再识别研究综述. 中国科学: 信息科学, 2021, 51(12): 1979–2015. [doi: 10.1360/SSI-2021-0211]

11 李爽, 李华锋, 李凡. 基于互预测学习的细粒度跨模态行人重识别. 激光与光电子学进展, 2022, 59(10): 1010010. [doi: 10.3788/LOP202259.1010010]

12 He ST, Luo H, Wang PC, *et al*. TransReID: Transformer-based object re-identification. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 14993–15002. [doi: 10.1109/ICCV48922.2021.01474]

13 Zhou SP, Wang F, Huang ZY, *et al*. Discriminative feature learning with consistent attention regularization for person re-identification. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 8039–8048. [doi: 10.1109/ICCV.2019.00813]

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

15 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

16 Xu K, Ba JL, Kiros R, *et al*. Show, attend and tell: Neural image caption generation with visual attention. Proceedings of the 32nd International Conference on Machine Learning. Lille: JMLR.org, 2015. 2048–2057.

17 Li S, Xiao T, Li HS, *et al*. Identity-aware textual-visual matching with latent co-attention. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 1908–1917. [doi: 10.1109/ICCV.2017.209]

18 Li H, Xiao JM, Sun MJ, *et al*. Transformer-based language-person search with multiple region slicing. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1624–1633. [doi: 10.1109/TCSVT.2021.3073718]

19 Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8748–8763.

20 Zhou KY, Yang JK, Loy CC, *et al*. Learning to prompt for vision-language models. International Journal of Computer Vision, 2022, 130(9): 2337–2348. [doi: 10.1007/s11263-022-01653-1]

21 Li SY, Sun L, Li QL. CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 1405–1413. [doi: 10.1609/aaai.v37i1.25225]

22 Zhao SZ, Gao CX, Shao YJ, *et al*. Weakly supervised text-based person re-identification. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 11375–11384.

23 Yao LW, Huang RH, Hou L, *et al*. FILIP: Fine-grained interactive language-image pre-training. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2022.

24 Wei LH, Zhang SL, Gao W, *et al*. Person transfer GAN to bridge domain gap for person re-identification. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 79–88. [doi: 10.1109/CVPR.2018.00016]

25 Zheng L, Shen LY, Tian L, *et al*. Scalable person re-identification: A benchmark. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1116–1124.

26 Ristani E, Solera F, Zou R, *et al*. Performance measures and a data set for multi-target, multi-camera tracking. Proceedings of the 2016 European Conference on Computer Vision. Amsterdam: Springer, 2016. 17–35. [doi: 10.1007/978-3-319-48881-3_2]

(校对责编: 张重毅)