◆交通场景◆

GCTR:粒度统一的跨模态文本行人检索网络模型*

覃　晓[1,2,3],张金勇[1],龚远旭[1],吴琨生[1],黄豪杰[1],淳　鑫[1],元昌安[2,3**]
(1. 南宁师范大学,广西人机交互与智能决策重点实验室,广西南宁　530100;2. 广西科学院,广西南宁　530012;3. 广西区域多源数据集成与智能处理协同创新中心,广西桂林　541004)

摘要:现有的文本行人检索网络模型在检索任务中缺乏对图文语义联系的关注,且容易忽略文本与图像特征之间的粒度差异,针对这两大问题,本研究提出一种粒度统一的跨模态文本行人检索网络模型(Granularity-unified Cross-modal Text-person Retrieval model,GCTR)。首先,GCTR 利用具备跨模态迁移知识能力的视觉语言预训练模型来获取具有基础关联性的文本和图像特征;其次,本研究提出一个跨模态粒度特征增强模块(Cross-Model Feature Enhancement module,CMFE),它利用跨模态特征增强码表(Enhanced Cross-modal Feature Codebook,ECFC)获取具有统一粒度的图像文本特征,解决了图文特征粒度差异的问题;最后,结合局部和全局的匹配损失策略完成模型的训练。GCTR 在 CUHK-PEDES、ICFG-PEDES 和 RSTPReid 3 个公开数据集上的表现均优于现有的主流模型,证明了 GCTR 在跨模态文本行人检索任务上的优越性。

关键词:跨模态检索;图文检索;行人检索;视觉语言预训练;粒度特征增强

中图分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1005-9164(2024)05-0988-14
DOI:10.13656/j.cnki.gxkx.20241127.015

行人检索任务是指通过特定人物的图像、视频或文本描述,在数据库中寻找包含对应人物的图像或视频的这一过程,该技术在公共安全领域有着广泛应用,例如,警察根据犯罪嫌疑人的照片,利用监控系统去寻找嫌疑人。然而,现实中的检索方往往仅有对目标人物的文本描述,缺乏对应的人物图片,这对行人检索技术提出了新的挑战。文本行人检索技术使得检索方可以在海量数据中迅速找到目标人员,这不仅能大幅度提升检索效率,还能节省大量人力资源,具有极高的应用价值。

文本行人检索的过程如图 1 所示,文本行人检索任务要从行人图像数据源中,找出与文本描述匹配的行人图像。相较于传统的基于图像的行人检索任务[1-3],基于文本描述去检索行人的任务因为文本的获取相对比较容易,因而更加符合现实场景。

文本行人检索任务涉及到自然语言处理和计算

收稿日期:2024-07-13　　修回日期:2024-09-25
*科技部科技创新 2030—"脑科学与类脑研究"重大项目(2021ZD0201904)和广西科技重大专项(桂科 AA22068057)资助。
【第一作者简介】
覃　晓(1973—),女,硕士,教授,主要从事自然语言处理、图像处理等研究,E-mail:7670172@qq.com。
【**通信作者简介】
元昌安(1964—),男,博士,教授,主要从事智能计算研究,E-mail:yuanchangan@126.com。
【引用本文】
覃晓,张金勇,龚远旭,等. GCTR:粒度统一的跨模态文本行人检索网络模型[J]. 广西科学,2024,31(5):988-1001.
QIN X,ZHANG J Y,GONG Y X,et al. GCTR:A Granularity-Unified Cross-Modal Text-person Retrieval Model [J]. Guangxi Sciences,2024,31(5):988-1001.

机视觉两个研究领域,且存在着不同模态特征之间难以对齐的问题。早期,为了解决这些问题,研究人员提出了基于注意力机制的文本行人检索方法[4-8],但这些方法忽略了文本和图像之间的语义联系,导致文本和图像特征的匹配效果不尽如人意。

近年来,以对比语言-图像预训练模型(Contrastive Language-Image Pre-training model,CLIP)[9]为代表的视觉语言预训练模型在挖掘图像与文本特征之间的关联性方面展现了巨大的潜力,因此,基于CLIP的文本行人检索方法逐渐引起人们的关注。Jiang等[10]利用CLIP进行随机文本掩码和隐式关系推理,Bai等[11]则通过引入基于动量的替换令牌检测和掩码语言模型进一步优化了这一过程。这些方法依赖CLIP来匹配文本与图像之间的关系,并在检索效果上超越了当前主流的单模态预训练模型,但是CLIP本身更关注行人图像与描述文本词汇之间的特征匹配,因而难以在语义特征粒度上很好地对行人图像和描述文本特征进行匹配。

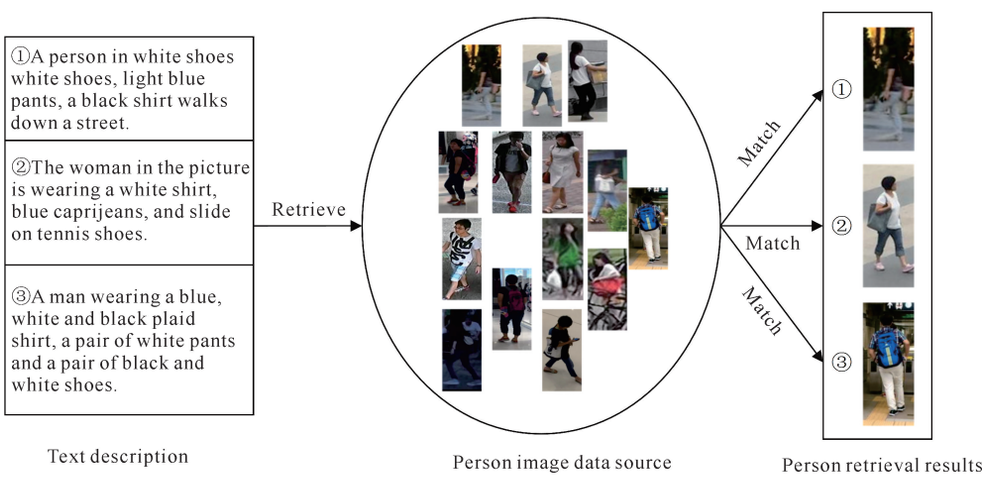


图1 文本行人检索流程图

Fig. 1 Flow chart of text-person retrieval

综上所述,针对现有跨模态文本行人检索方法对图文语义关系缺乏关注,以及忽略了文本与图像特征粒度存在差异所导致的跨模态检索复杂性的问题,本研究提出了一个粒度统一的跨模态文本行人检索网络模型(Granularity-unified Cross-modal Text-person Retrieval model,GCTR)。首先,GCTR采用CLIP提取行人图像和文本特征,对跨模态特征进行粗匹配;然后,利用跨模态特征增强模块(Cross-Modal Feature Enhancement module,CMFE)学习文本描述的语义特征,指导模型在语义粒度上对行人图像和描述文本相匹配的特征进行增强;最后,通过整合局部和全局的匹配损失策略,实现更精确的语义信息匹配。

1 相关工作

1.1 跨模态文本行人检索网络模型

基于注意力机制的方法主要通过分配权重给图像和文本之间的特征来揭示它们的图像文本关联性[4,12],从而完成行人检索任务。Zheng等[13]提出的双向注意力机制网络,可以同时处理图像对文本以及文本对图像的关注点,实现特征层面上的双向对齐,为跨模态文本行人检索领域的研究开辟了思路。

Li等[14]提出的GNA-RNN网络模型将每个单词的热编码与VGG网络[15]提取的图像特征进行拼接,并通过LSTM网络处理[16],最终利用注意力机制分配特征权重,计算文本与图像的亲和力得分。然而,Chen等[17]指出,GNA-RNN网络模型会导致某些单词的特征权重过高,因此他们通过引入自适应阈值,对权重分配单元进行改善,确保图文匹配过程中重要信息得到全面考虑。然而,这些方法仍然停留在全局层面上对特征进行处理,忽略了对局部特征的关注。

为了解决对局部细节发现不足的问题,Wang等[18]提出了ViTAA网络模型,从属性对齐的角度出发,将全局特征空间分解到对应属性的子空间,再使用对比学习的方法实现图像特征与文本属性的对齐。但Gao等[19]指出,ViTAA网络模型未能捕捉到隐藏的高级抽象语义信息,导致某些文本描述与图像区域无法正确对齐。Niu等[20]尝试通过融合图像局部特征与全局特征的语义信息来改善这一问题,但此方法主要局限于图像部位,缺乏对相关短语特征之间深层语义关联性的探索。因此,本研究将面向如何结

合全局特征与局部特征,深入挖掘图像与文本属性特征之间的语义关联性展开研究。

1.2 CLIP

CLIP[9]是一个能够实现强大的跨模态特征对齐效果的视觉语言预训练模型。CLIP 主要根据以下 4 个步骤实现跨模态语义对齐:①基于 Transformer 架构对图像和文本进行编码,获得它们的特征表达;②计算图像和文本表达之间的余弦相似度,依据匹配的图像文本对在特征空间中的相似度更高,不匹配的图像文本对相似度更小的准则学习特征相似性矩阵;③利用类别词构建文本描述模板,并利用训练好的编码器对文本描述进行编码;④利用训练好的图像编码器对图像进行编码,根据图像-文本相关矩阵,找到相似度最高的图像-文本配对特征,最终得到正确的类别预测。

CLIP 通过融合基于 Transformer 架构的图像和文本编码器,在大规模的图像文本对上应用对比学习策略,成功捕获到丰富的视觉-语言关联信息,挖掘出具有高度关联性的跨模态特征。特别是它在没有大量标注数据的情况下,仅仅通过理解文本与图像之间的关系,一样能够高效、准确地执行检索任务的能力[21-24]。Jiang 等[10]提出的 IRRA 证明了完整的 CLIP 模型可以迁移、应用于文本行人检索任务中,并通过微调模型,使得 IRRA 超越了众多先进的文本行人检索方法。这展现了 CLIP 能够迁移到其他跨模态任务上的巨大潜力。因此,本研究将 CLIP 作为特征提取骨干网络来获取文本描述和行人图像之间的特征表达。

2 方法

图 2 是本研究提出的 GCTR。首先,对于给定的一组行人图像 V 和对应的文本描述 T,使用 CLIP 的双流编码器来提取图像和文本特征。其次,为了深入挖掘文本与图像之间语义层面上统一粒度的高质量特征,本研究设计了 CMFE,用来获取在语义层面上统一粒度的文本增强特征和图像增强特征。最后,采用局部、全局匹配损失策略以及多样性的正则化约束策略来调整特征对齐损失函数,提高模型对文本特征与图像特征的对齐能力。

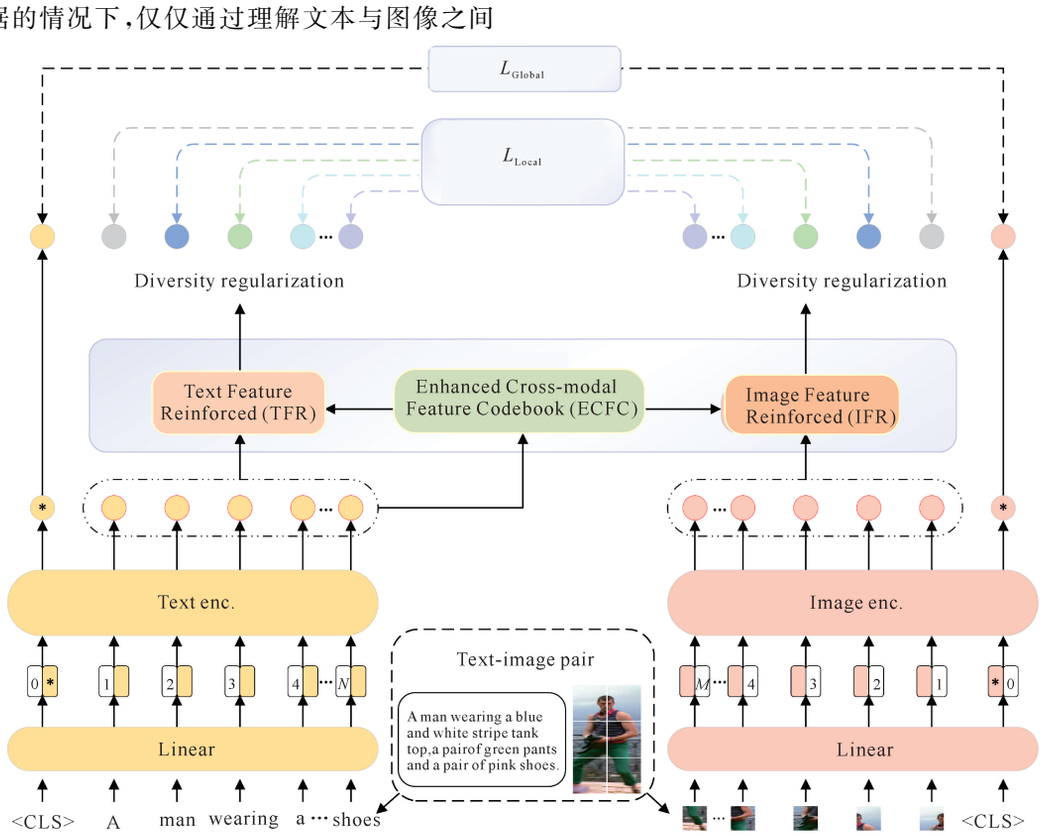
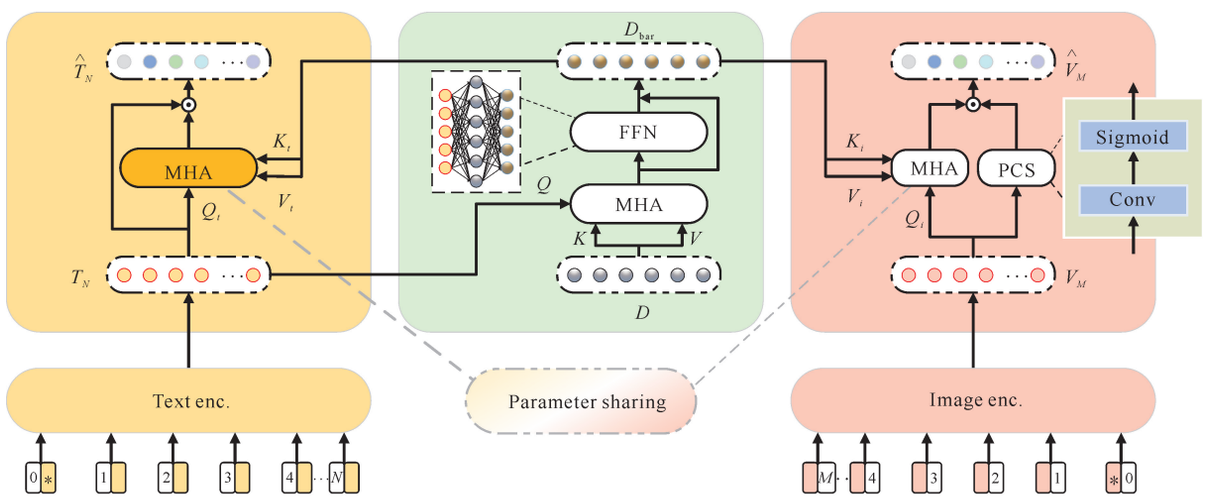


图 2 GCTR

Fig. 2 GCTR

2.1 CMFE

为了在语义层面上统一跨模态文本图像特征粒度,CMFE 接收 CLIP 获取的文本特征和图像特征,通过自注意力机制获取文本特征的深层语义信息,然后利用深层语义信息对文本特征和图像特征分别进行增强指导,从而获得在语义粒度上更加一致的跨模态特征。图 3 展示了 CMFE 的详细设计,CMFE 由 3 个模块组成,包括跨模态特征增强码表(Enhanced Cross-modal Feature Codebook,ECFC)、文本特征增强模块(Text Feature Reinforced module,TFR)和图像特征增强模块(Image Feature Reinforced module,IFR)。ECFC 能够获取文本的深层语义信息并进行文本特征重构,CMFE 利用 ECFC 重构的文本特征去指导 TFR 和 IFR,生成语义粒度匹配的文本特征和图像特征。同时,为了减轻图像背景对图像增强特征的干扰,本研究在图像增强模块中设计了专注于行人图像特征信息的行人卷积采样模块(People Convolution Sampling module,PCS)。



(a) TFR (b) ECFC (c) IFR

图 3 CMFE

Fig. 3 CMFE

2.1.1 ECFC

ECFC 如图 3(b)所示,其采用 Transformer 架构来生成文本特征增强矩阵码表 D_{bar},其中 D 是随机初始化的码表。将单词级文本序列特征 T_N 作为 Q(Query),初始码表 $D \in R^{o\times d}$ 作为 K(Key)和 V(Value),其中 N 表示单词级文本序列的特征长度,o 表示码表序列个数,d 表示维度,则有

$$Q=T_N W_q, K=DW_k, V=DW_v, \tag{1}$$

其中,W_q、W_k 和 W_v 为可学习参数,特征码表计算过程表示如下:

$$D'=\text{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d}})V, \tag{2}$$

$$\text{FFN}(D')=\text{ReLU}(D'W_1+b_1)W_2+b_2, \tag{3}$$

$$D_{\text{bar}}=\text{concat}(\text{FFN}(D'),D'), \tag{4}$$

其中,W_1、W_2 表示可学习参数,b_1、b_2 表示偏置数。首先通过自注意力机制对初始码表 D 进行处理,得到与文本语义信息加权融合后的特征码表 D',再经过 FFN 网络层后得到与原有文本特征相似的文本特征增强矩阵码表 D_{bar}。ECFC 算法流程如算法 1 所示。

算法 1 ECFC 算法流程

输入:文本特征序列 $T_N=\{t_1,t_2,\cdots,t_N\}$ 和空码表 $D=\{\}$

输出:文本特征增强矩阵码表 D_{bar}

1:初始化可学习参数 W_q、W_k 和 W_v;

2:按公式(1)计算 Q、K 和 V;

3: $D'=\text{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d}})V$;

4:初始化可学习参数 W_1、W_2 和偏置数 b_1、b_2;

5: $\text{FFN}(D')=\text{ReLU}(D'W_1+b_1)W_2+b_2$;

//计算码表预测结果

6: $D_{\text{bar}}=\text{concat}(\text{FFN}(D'),D')$; //残差连接和归一化

7: Output D_{bar}

2.1.2 TFR

TFR 通过多头注意力来融入码表的语义信息,从而获取带有词向量关联性的文本特征,过程如图 3(a)所示。TFR 使用单词级文本序列特征 T_N 作为

多头注意力的Q_t,文本特征增强矩阵码表D_{bar}作为K_t和V_t,应用多头注意力后得到特征增强后的文本特征,过程表示如下。

$$Q_t=T_N W_t^Q, K_t=D_{\text{bar}}W_t^K, V_t=D_{\text{bar}}W_t^V, \quad (5)$$

其中,W_t^Q,W_t^K,W_t^V为可学习参数矩阵。文本增强特征$\hat{T}_N$的计算过程如下。

$$\text{Attention}(Q_t,K_t,V_t)=\text{softmax}(\frac{Q_tK_t^{\text{T}}}{\sqrt{d_t}})V_t, \quad (6)$$

$$\hat{T}_N=\text{Attention}(Q_t,K_t,V_t)\odot T_N, \quad (7)$$

其中,$\odot$表示点乘运算,d_t是特征向量$\boldsymbol{V}_t$的维度。

2.1.3 IFR

IFR 的操作与 TFR 类似,如图 3(c)所示。以图像局部特征V_M作为Q_i,M表示图像序列的特征长度,使用文本特征增强矩阵码表D_{bar}作为K_i和V_i,再计算多头注意力,最后获得与文本语义粒度相统一的图像特征。为保持跨模态信息之间的关联性,CMFE 使用了多头注意力参数共享策略[25]。考虑到图像背景会对前景身份造成信息干扰,IFR 还设计了一个包含 1×1 的卷积和 Sigmoid 函数的 PCS 模块,用于增强前景身份的信息表达,以减少无关信息的噪声数据干扰。将V_M输入多头注意力模块与 PCS 模块后得到图像增强特征$\hat{V}_M$。

$$\hat{V}_M=\text{Attention}(Q_i,K_i,V_i)\odot\text{Sigmoid}(\text{Conv}(V_M)), \quad (8)$$

图像特征增强的具体流程如算法 2 所示。对比于原来的图像局部特征V_{M},增强后的$\hat{V}_M$不仅携带有与文本相同粒度的语义特征,还增强了包含行人身份的前景信息表达,这有助于文本特征与图像特征达到更好的匹配对齐效果。

算法 2 IFR 算法流程

输入:图像特征序列$V_M=\{v_1,v_2,\cdots,v_M\}$和文本特征增强矩阵码表$D_{\text{bar}}=\{d_1,d_2,\cdots,d_N\}$

输出:图像增强特征$\hat{V}_M$

1:$Q_i=V_M W_i^Q$;

2:$K_i=D_{\text{bar}}W_i^K$;

3:$V_i=D_{\text{bar}}W_i^V$;

4:$V_A=\text{softmax}(\frac{Q_iK_i^{\text{T}}}{\sqrt{d_i}})V_i$; //计算融合图像特征$V_A$

5:$F=\text{Conv}_{1\times1}(V_M)$; //压缩人物特征信息

6:$F'=\text{Sigmoid}(F)$; //计算前景人物权重概率

7:$\hat{V}_M=T_A F'$;

8: Output $\hat{V}_M$

2.2 跨模态对齐损失函数

受文献[26]的启发,本研究采用了交叉熵损失作为码表预测损失。设文本输入特征为T,真实标签向量为$\boldsymbol{y}$,码表标签预测的概率为$\hat{y}$,码表预测损失可表示为

$$L_D(T)=-\sum y\log(\hat{y})。 \quad (9)$$

对于模型匹配损失,本研究使用 CMPM+CMPC 损失策略来进行对局部特征的对齐。输入文本增强特征$\hat{T}_N$和图像增强特征$\hat{V}_M$,局部损失可表示为

$$L_{\text{Local}}=\sum(L_{\text{CMPM}}(\hat{T}_N,\hat{V}_M)+L_{\text{CMPC}}(\hat{T}_N,\hat{V}_M))。 \quad (10)$$

L_{CMPM}通过 KL 散度来最小化不同身份的文本特征与图像特征之间的相似度,最大化相同身份的文本特征与图像特征之间的相似度,加入L_{CMPC}旨在鼓励相同身份的特征表达能力,以区分不同身份的特征表达。此外,对于一批图像-文本对,还引入了交叉双向三重排序(Triplet)损失来完成对全局特征匹配损失的优化,全局损失表示如下:

$$L_{\text{Global}}=L_{\text{Triplet}}=\max(\alpha-S(V_M,T_N)+S(V_M,T_N^-),0)+\max(\alpha-S(V_M,T_N)+S(V_M^-,T_N),0), \quad (11)$$

其中,(V_M,T_N)表示匹配的图像-文本正样本对,(V_M,T_N^-),(V_M^-,T_N)表示不匹配的图像-文本负样本对,α是一个边距超参数,S表示余弦相似度计算度量。

此外,每个模态内不同的粒度特征应该聚焦于不同信息,因此,对于不同的粒度特征,可通过添加多样性约束L_{Div}来避免信息冗余,提高模型对多粒度细节的敏感性。多样性约束表示如下:

$$L_{\text{Div}}=\sum_{i=1}\sum_{j=1,i\neq j}(\frac{v_i,v_j}{\|v_i\|_2\|v_j\|_2}+\frac{t_i,t_j}{\|t_i\|_2\|t_j\|_2}), \quad (12)$$

其中,v_i和t_i分别为图像和文本在不同粒度下的特

征,通过计算它们与其他粒度特征 v_j 和 t_j 的相似性,来确保它们在信息上不相似。

整个模型的总体损失表示为

$$L = L_D + L_{\text{Local}} + L_{\text{Global}} + \lambda L_{\text{Div}}, \tag{13}$$

其中,λ 为控制训练过程中多样性损失的参数。

GCTR 模型的完整流程如算法 3 所示。

算法 3 GCTR 模型算法流程

输入:图像 V 和文本 T,参数 λ

输出:对齐后的图像特征 V' 和文本特征 T'

1: $V_M, T_N = \text{CLIP}(V, T)$; //CLIP 获取图像特征和文本特征

2: $D = \{\}$; // 初始化码表 D

3: $D_{\text{bar}} = \text{ECFC}(T_N, D)$; //使用增强码表 ECFC 获得码表 D_{bar}

4: $\hat{T}_N = \text{TFR}(T_N, D_{\text{bar}})$; //使用 TFR 模块获得增强文本特征 $\hat{T}_N$

5: $\hat{V}_M = \text{IFR}(V_M, D_{\text{bar}})$; //使用 IFR 模块获得增强图像特征 $\hat{V}_M$

6: $L_D(T) = -\sum y\log(\hat{y})$; //计算码表损失

7: $L_{\text{Global}} = L_{\text{Triplet}} = \max(\alpha - S(V_M, T_N) + S(V_M, T_N^-), 0) + \max(\alpha - S(V_M, T_N) + S(V_M^-, T_N), 0)$; //计算全局损失

8: $L_{\text{Local}} = \sum(L_{\text{CMPM}}(\hat{T}_N, \hat{V}_M) + L_{\text{CMPC}}(\hat{T}_N, \hat{V}_M))$; //计算局部对齐损失

9: $L_{\text{Div}} = \sum_{i=1}\sum_{j=1, i\neq j}\left(\frac{v_i, v_j}{\| v_i \|_2 \| v_j \|_2} + \frac{t_i, t_j}{\| t_i \|_2 \| t_j \|_2}\right)$; //进行多样性约束 L_{Div}

10: $L = L_D + L_{\text{Local}} + L_{\text{Global}} + \lambda L_{\text{Div}}$; //计算总损失

11: return V', T'

3 实验结果与分析

3.1 数据集

3.1.1 CUHK-PEDES

CUHK-PEDES 数据集[17]是一个公用的基于文本行人检索的大规模基准数据集。它包含 40 206 张图片和关于 13 003 个行人的 80 412 个文本描述,每张图片都有手工标注的 2 个文本描述,每个文本描述的平均长度不低于 23 个单词。表 1 为 CUHK-PEDES 数据集的数据分布情况。

表 1 CUHK-PEDES 数据集介绍

Table 1 Introduction of the CUHK-PEDES dataset

类型 Type	内容 Content	数量 Number
Training set	Person ID	11 003
	Image	34 054
	Textual description	68 108
Validation set	Person ID	1 000
	Image	3 078
	Textual description	6 156
Testing set	Person ID	1 000
	Image	3 074
	Textual description	6 148

3.1.2 ICFG-PEDES

ICFG-PEDES 数据集[27]从 MSMT17 数据集[28]收集、整理得到,每张图片只有 1 个文本描述,每个文本描述的平均长度为 37 个单词。它包含 54 522 张图片和 4 102 个不同身份行人。表 2 为 ICFG-PEDES 数据集的数据分布情况。

表 2 ICFG-PEDES 数据集介绍

Table 2 Introduction of the ICFG-PEDES dataset

类型 Type	内容 Content	数量 Number
Training set	Person ID	3 102
	Image	34 674
	Textual description	34 674
Testing set	Person ID	1 000
	Image	19 848
	Textual description	19 848

3.1.3 RSTPReid

RSTPReid 数据集[29]基于 MSMT17 数据集[28]构建,是专为基于文本的人物重新识别设计的一个新数据集。它包含 20 505 张图片和 4 101 个不同身份的行人,图片由 15 个不同角度的摄像头拍摄而成,每个人物都有 5 张不同摄像头拍摄的照片,每张图片都有 2 个文本描述。表 3 为 RSTPReid 数据集的数据分布情况。

3.2 评价指标

本研究使用 $Rank\text{-}K(K=1,5,10)$ 作为本研究实验的性能评价指标,它是本领域常用的性能评估指标。具体来说,在模型返回的图片排序列表中,前 K

张行人图片存在与文本描述一致的检索目标则称为 *Rank-K* 命中,*Rank-K* 指标越高说明模型的检索性能越好。

表 3　RSTPReid 数据集介绍

Table 3　Introduction of the RSTPReid dataset

类型 Type	内容 Content	数量 Number
Training set	Person ID	3 701
	Image	18 505
	Textual description	37 010
Validation set	Person ID	200
	Image	1 000
	Textual description	2 000
Testing set	Person ID	200
	Image	1 000
	Textual description	2 000

3.3　实验设置

本研究全部实验在单张 Nvidia GeForce RTX 4080 GPU 上进行,处理器使用 Intel Core i9-10980XE CPU @ 3.00 GHz×36,开发环境为 Python 3.7,PyTorch 1.13.0,Cuda 11.7,实验的具体参数设置如表 4 所示。

表 4　实验参数设置

Table 4　Experimental parameter settings

参数 Parameter	数值 Value
Image size	384×128
Visual representation	512
Textual representation	512
Margin	0.2
Batch	64
Epoch	60
Initial learning rate	1×10^{-5}
Visual sequence length	196
Textual sequence length	100
λ	0.3

3.4　对比实验

3.4.1　在 CUHK-PEDES 数据集上进行的对比实验分析

表 5 为 GCTR 与主流行人重识别模型在 CUHK-PEDES 数据集上的对比结果,可以看出 GCTR 均取得了最佳的结果。GCTR 优异的检索性能,得益于它引用了 CLIP 来构建匹配网络,使其能够更好地挖掘图像、文本的先验知识,使用 CLIP 图像和文本编码器来获取底层特征的效果,比使用单一模态的预训练特征编码器来提取跨模态特征的效果要好。在 CUHK-PEDES 数据集上,GCTR 比 CFine[39] 在 *Rank*-1、*Rank*-5、*Rank*-10 上分别提升 0.70%、2.20%、1.31%,证明了 GCTR 的跨模态特征增强模块能更好地提取到具有统一粒度的图像、文本特征,实现更精准的图像文本特征对齐。

3.4.2　在 ICFG-PEDES 数据集上进行的对比实验分析

表 6 为 GCTR 与主流行人重识别模型在 ICFG-PEDES 数据集上的对比结果。GCTR 相比于目前使用单一模态特征提取网络达到最好效果的 LGUR[37] 相比,在 *Rank*-1、*Rank*-5、*Rank*-10 上分别有 2.21%、1.84%、1.13%的提升,证明了使用 CLIP 提取模态特征的效果比使用单模态特征提取网络的效果要好。而相比于 CFine,GCTR 在 *Rank*-1、*Rank*-5、*Rank*-10 上也分别提升 0.40%、0.61%、0.27%。CFine 也受益于使用 CLIP 预训练模型,从而获得了很好的图像、文本模态特征,但是,GCTR 仍能取得比 CFine 更好效果的原因可能是 GCTR 具有挖掘局部特征信息的能力,以及统一了特征粒度。

表 5　GCTR 与主流行人重识别模型在 CUHK-PEDES 数据集上的对比结果

Table 5　Performance comparisons of GCTR with SOTA models on CUHK-PEDES dataset　　**Unit:%**

模型 Model	视觉分支 Image enc.	文本分支 Text enc.	*Rank*-1	*Rank*-5	*Rank*-10
GNA-RNN[14] (2017)	VGG	LSTM	19.05		53.65
Dual Path[30] (2020)	ResNet	ResNet	14.40	66.26	75.07
CMPM/C[26] (2018)	ResNet	LSTM	49.37	71.69	79.27
MIA[20] (2020)	VGG	GRU	53.10	75.00	82.90
PMA[6] (2020)	ResNet	LSTM	54.12	75.45	82.97
TLMAM[31] (2019)	ResNet	BERT	54.51	77.56	84.78
CMKA[32] (2021)	ResNet	LSTM	54.69	73.65	81.86

续表

Continued table

模型 Model	视觉分支 Image enc.	文本分支 Text enc.	*Rank*-1	*Rank*-5	*Rank*-10
ViTAA*[18](2020)	ResNet	LSTM	55.97	75.84	83.52
CMAAM[33](2020)	Mobilenet	LSTM	56.68	77.18	84.86
NAFS[19](2020)	ResNet	BERT	59.94	79.86	86.70
NAFS+LapsCore[34](2021)	ResNet	BERT	63.40	—	87.80
SSAN*[27](2021)	ResNet	LSTM	61.37	80.15	86.73
LBUL[35](2022)	ResNet	BERT	64.04	82.66	87.22
TIPCB[36](2022)	ResNet	BERT	64.26	83.19	89.10
LGUR*[37](2022)	DeiT-Small	BERT	64.68	83.12	89.00
CFine*[30](2020)	CLIP-Vit	BERT	69.47	85.93	91.15
CSKT*[38](2024)	CLIP-Vit	CLIP-Te	69.80	86.42	91.68
CLIP[9]	CLIP-Vit	CLIP-Te	65.73	83.27	88.95
GCTR (Ours)	CLIP-Vit	CLIP-Te	**70.17**	**88.13**	**92.46**

Note: the best results are in bold; * represents the experimental results reproduced on this dataset using open-source code of their paper.

表 6 GCTR 与主流行人重识别模型在 ICFG-PEDES 数据集上的对比结果

Table 6 Performance comparisons of GCTR with SOTA models on ICFG-PEDES dataset Unit: %

模型 Model	视觉分支 Image enc.	文本分支 Text enc.	*Rank*-1	*Rank*-5	*Rank*-10
Dual Path[30](2020)	ResNet	ResNet	38.99	59.44	68.41
CMPM/C[26](2018)	ResNet	LSTM	43.51	65.44	74.26
MIA[20](2020)	VGG	GRU	46.49	67.14	75.18
ViTAA*[18](2020)	ResNet	LSTM	50.98	68.79	75.76
SSAN*[27](2021)	ResNet	LSTM	54.23	72.63	79.53
TIPCB[36](2022)	ResNet	BERT	54.96	74.72	81.89
LGUR*[37](2022)	DeiT-Small	BERT	59.02	75.32	81.56
CSKT*[38](2024)	CLIP-Vit	CLIP-Te	58.60	**77.28**	82.56
CFine*[30](2020)	CLIP-Vit	BERT	60.83	76.55	82.42
CLIP[9]	CLIP-Vit	CLIP-Te	55.89	75.21	81.17
GCTR (Ours)	CLIP-Vit	CLIP-Te	**61.23**	77.16	**82.69**

Note: the best results are in bold; * represents the experimental results reproduced on this dataset using open-source code of their paper.

3.4.3 在 RSTPReid 数据集上进行的对比实验分析

表 7 表示为 GCTR 与主流行人重识别模型在 RSTPReid 数据集上的对比结果。在 RSTPReid 数据集上,GCTR 比 CFine 在 *Rank*-1、*Rank*-5、*Rank*-10 上分别提升 2.00%、3.85%、1.16%,对比使用单一模态预训练特征提取网络的模型 LGUR,在 *Rank*-1、*Rank*-5、*Rank*-10 上分别获得 5.85%、6.35%、3.96%的提升,这证明了对 CLIP 提取的特征使用本研究提出的跨模态粒度特征增强模块能够减小图像-文本特征的粒度差异,进而提高了图像-文本特征之间的匹配效果。对于 RSTPReid 这种具有复杂场景的数据集,GCTR 在 *Rank*-1、*Rank*-5、*Rank*-10 上的效果没有比在 CUHK-PEDES 和 ICFG-PEDES 数据集上提升的效果明显,原因可能是模型缺少对具有遮掩性人物的认知,对于复杂场景下的识别效果不足。与 CSKT[38] 相比,GCTR 在 *Rank*-1 和 *Rank*-5 上分别有 0.80%、0.25%的提升,而在 *Rank*-10 上则存在着差距,实验效果不如 CSKT 的原因可能是在关系学习上,CSKT 能够更好地学习到对于符合同一个文本描述的相似行人的信息差异,而 GCTR 则更着重

于关注图像与文本特征的粒度差异。

表 7　GCTR 与主流行人重识别模型在 RSTPReid 数据集上的对比结果

Table 7　Performance comparisons of GCTR with SOTA models on RSTPReid dataset　　**Unit:%**

模型 Model	视觉分支 Image enc.	文本分支 Text enc.	*Rank*-1	*Rank*-5	*Rank*-10
CMKA[32](2021)	ResNet	LSTM	37.60	61.15	73.55
ViTAA*[18](2020)	ResNet	LSTM	38.45	62.40	73.80
SSAN*[27](2021)	ResNet	LSTM	43.50	67.80	77.15
TIPCB[36](2022)	ResNet	BERT	45.55	68.20	77.85
LGUR[37](2022)	DeiT-Small	BERT	46.70	70.00	78.80
CFine*[39](2023)	CLIP-Vit	BERT	50.55	72.50	81.60
CSKT*[38](2024)	CLIP-Vit	CLIP-Te	51.75	76.10	**85.56**
CLIP[9]	CLIP-Vit	CLIP-Te	46.85	71.59	79.10
GCTR (Ours)	CLIP-Vit	CLIP-Te	**52.55**	**76.35**	82.76

Note: the best results are in bold; * represents the experimental results reproduced on this dataset using open source code of their paper.

3.5　消融实验

3.5.1　结构性消融实验

消融实验在 CUHK-PEDES 数据集上进行。如表 8 所示,No. 0 表示采用 ImageNet 数据集预训练的 ViT 和 Bert 作为图像和文本的特征提取基准模型,与使用 CLIP,以及增加 CMFE 中 TFR、IFR 和 PCS 模块的模型进行对比实验。

表 8　GCTR 在 CUHK-PEDES 数据集上的结构性消融实验

Table 8　Ablation experiments on each module of GCTR on CUHK-PEDES dataset　　**Unit:%**

序号 No.	模型 Model	模块 Module				*Rank*-1	*Rank*-5	*Rank*-10
		CLIP	TFR	IFR	PCS			
0	Baseline					57.36	77.49	85.98
1	Baseline +	√				65.24	83.27	89.48
2	Baseline ++	√	√	√		68.72	85.41	90.79
3	GCTR	√	√	√	√	**70.17**	**88.13**	**92.46**

Note: the best results are in bold.

首先,将 No. 0 和 No. 1 进行比较,对比基准模型,本研究模型在加入了 CLIP 后,在 *Rank*-1 上提升 7.88%。相比于单模态图像文本特征提取预训练模型,跨模态预训练模型具有更强的提取跨模态语义特征的能力,这证明了 CLIP 的有效性。对比表 8 的 No. 1 与 No. 2,CMFE 中 TFR 模块和 IFR 模块的加入使模型在 *Rank*-1 上提升 3.48%,这说明增强统一文本与图像的粒度特征,能帮助模型在文本与图像之间获取更强大的匹配联系能力。从 No. 1 与 No. 3 的比较中可以看出,在加入 PCS 模块后,完整的 CMFE 使得加入 CLIP 之后的模型在 *Rank*-1 上继续提升了 4.93%,证明了 CMFE 的有效性。从 No. 2 与 No. 3 的比较中可以发现,PCS 模块的补充让模型在 *Rank*-1 上提升了 1.45%,证明了 PCS 模块能够有效地增强行人图像中重要的前景特征信息,同时抑制背景中的噪声干扰因素,从而与文本相似特征进行更好地匹配。

3.5.2　局部-全局匹配策略实验分析

对于 *Rank*-1 的最大值与最小值,本研究针对不同损失策略,每 10 个 epoch 进行了一次准确率对比,如表 9 所示。本研究所使用的损失策略 No. 5 使得 *Rank*-1 在整体上都高于其他策略,与目前最好的策略 No. 4 相比,*Rank*-1 最低都有约 1%的提升。单纯使用 CMPM + CMPC 损失这种通用损失策略(No. 3)落后于使用 CMPM+CMPC 损失进行局部特征对齐以及 Triplet 损失进行全局特征的损失策略(No. 4),因为 CMPM+CMPC 损失更关注局部人物信息在文本描述特征上的投影匹配,对于人物与文本全局信息的样本信息配对缺乏关注,而 Triplet 损失在文献[40,41]上已经证明了其在全局样本匹配对上

具有好的效果。因此,本研究在使用 CMPM+CMPC 损失进行局部特征对齐的基础上,对全局特征使用 Triplet 损失策略,加强了模型对于全局信息的关注。

表 9　不同损失策略下的 *Rank*-1 在各个 epoch 的准确率对比

Table 9　Comparison of accuracy of *Rank*-1 at each epoch under different loss strategies　　**Unit:%**

序号 No.	损失策略 Loss strategies	值域 Range	[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60]
0	Global:CMPM Local:CMPM	Min Max	28.41 48.43	49.41 66.76	65.48 67.93	66.53 67.98	67.23 67.84	66.58 67.85
1	Global:CMPC Local:CMPC	Min Max	29.07 47.12	49.37 66.94	66.69 67.91	66.76 67.98	66.95 67.94	67.19 67.97
2	Global:Triplet Local:Triplet	Min Max	31.15 47.24	53.22 66.74	65.76 67.90	67.36 67.93	66.71 67.99	65.67 68.10
3	Global:CMPM+CMPC Local:CMPM+CMPC	Min Max	32.19 54.92	56.32 67.51	66.96 67.93	67.81 69.07	67.00 67.93	67.10 67.97
4	Global:CMPM+CMPC Local:Triplet	Min Max	33.68 57.34	59.86 67.97	67.05 68.83	67.75 68.79	68.16 69.12	68.27 69.85
5	Global:Triplet Local:CMPM+CMPC	Min Max	**34.56** **56.67**	**59.90** **69.92**	**68.47** **70.08**	**68.18** **69.78**	**68.27** **70.07**	**68.28** **70.17**

Note:the best results are in bold.

3.5.3　α 和 λ 的参数设置实验分析

图 4 和图 5 描述了在 CUHK-PEDES 数据集上进行的 α 和 λ 参数设置实验,发现边距参数 α 设置为 0.2 时实验的效果最好,多样性损失的参数 λ 设置为 0.3 时实验效果最好。

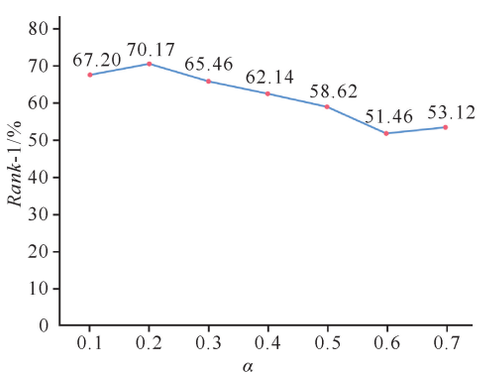


图 4　α 参数设置

Fig. 4　Diagram of α parameter setting

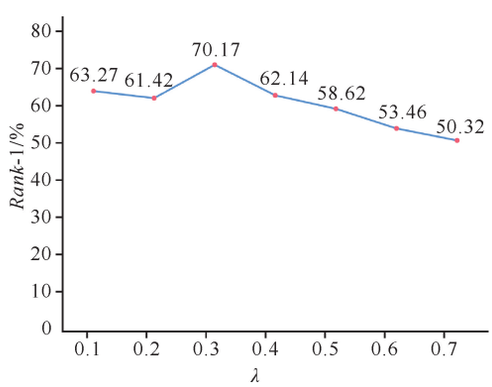


图 5　λ 参数设置

Fig. 5　Diagram of λ parameter setting

3.5.4　可视化实验分析

图 6 为模型在 CUHK-PEDES 数据集上进行检索的可视化图。图 6(a)为 3 位不同行人的查询文本。图 6(b)展示了查询文本中目标单词对应图像特征的关注热力图,可以看出本研究模型更注重于单词所匹配的局部关注点,以获取更好的检索效果。图 6(c)为不同模型实际检索到的前 10 张图片。每个样本的第一行为 GCTR 的检索结果,第二行为“基准模型+CLIP”的检索结果。绿色框图为正确匹配的行人图片,红色为错误匹配的行人图片,图片中的蓝色框和黄色框表示文本与图像的区域匹配点。从图 6(c)可以看出,本研究模型在处理目标行人图像时,能够更加精确地捕捉和匹配行人的特征细节,从而提高检索结果的准确性。相比于基准模型的检索结果,本研究模型对目标行人图像的特征关注契合度更高,检索结果更准确,这表明本研究所提出的模型具有较高的实用价值和可靠性。

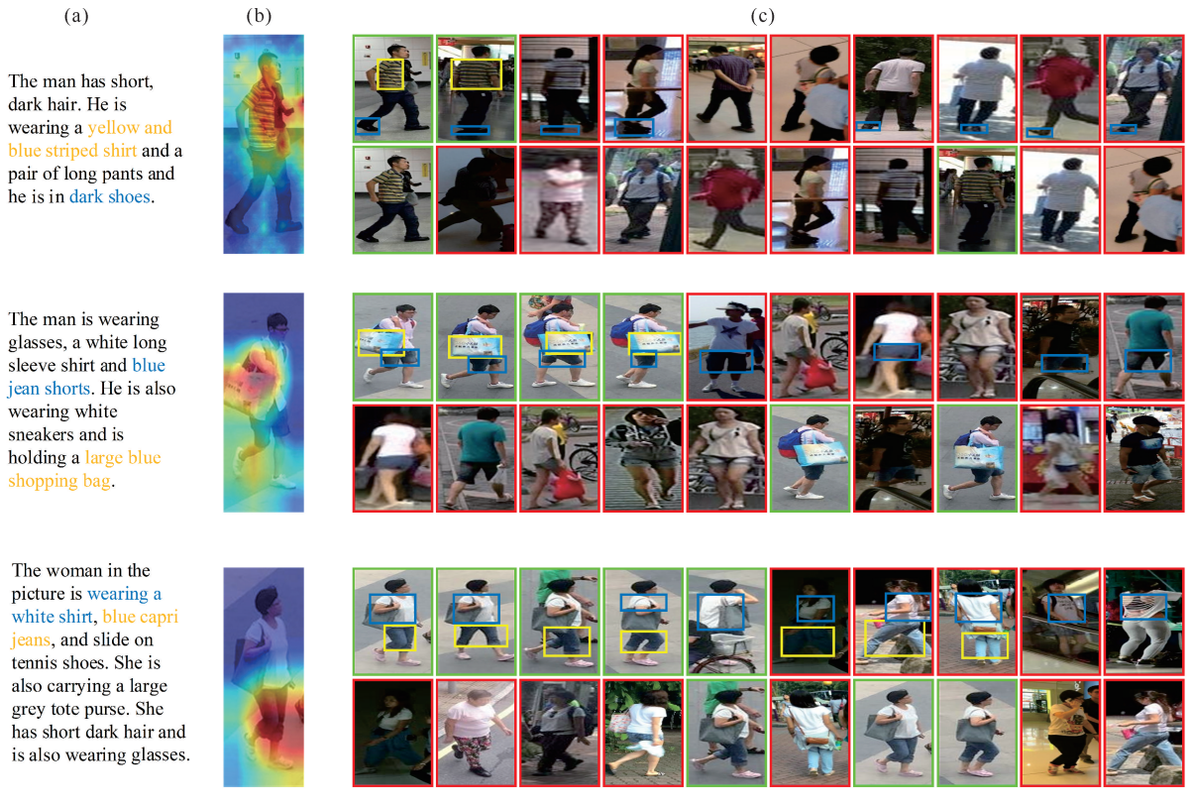


图 6　文本行人检索可视化

Fig. 6　Text-person retrieval visualization

4　结论

本研究提出了一种粒度统一的跨模态文本行人检索网络模型(GCTR),引入具备跨模态迁移知识能力的对比语言-图像预训练模型(CLIP)作为特征提取的核心网络来获取具有基础关联性的文本和图像特征,并设计了一个跨模态粒度特征增强模块(CMFE),它利用多模态特征增强码表提炼出增强的文本和图像特征,获取到具有统一粒度的图像文本特征。在图像特征增强方面,CMFE 整合了人物前景信息,以进一步提升图像特征的表达力,有效地减少了背景噪声对行人特征表达的干扰。CMFE 实现了特征粒度的统一,精确捕获了图像块与单词之间的对应关系。实验结果充分验证了所提出模型的有效性和优越性,在公共安全领域存在着广阔的应用前景。

然而,本研究的工作也存在一些局限性。首先,尽管 CMFE 能够处理复杂场景下的跨模态信息,但其泛化能力和在极端条件下的鲁棒性还有待进一步提高。在未来的工作中,将计划从提升模型对于极端场景(如低分辨率、复杂背景等)的处理能力以及对增强模型的泛化能力和鲁棒性等方面进行进一步研究和改进。

参考文献

[1] WIECZOREK M,RYCHALSKA B,DABROWSKI J. On the unreasonable effectiveness of centroids in image retrieval [C]//Neural Information Processing:28th International Conference. Berlin:Springer,2021:212-223.

[2] GONG Y P,HUANG L Q,CHEN L F. Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method [EB/OL]. (2022-06-13)[2024-05-23]. https://arxiv.org/abs/2101.08533.

[3] WANG G C,LAI J H,HUANG P G,et al. Spatial-Temporal Person Re-Identification [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York:AAAI,2019,33(1):8933-8940.

[4] 宋雨,王帮海,曹钢钢. 结合数据增强与特征融合的跨模态行人重识别[J]. 计算机工程与应用,2024,60(4):133-141.

[5] 张广耀,宋纯锋. 融合人体全身表观特征的行人头部跟踪模型[J]. 计算机应用,2023,43(5):1372-1377.

[6] JING Y,SI C Y,WANG J B,et al. Pose-guided multi-granularity attention network for text-based person search [C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York:AAAI,2020,34(7):11189-11196.

[7] 孟月波,穆思蓉,刘光辉,等.基于向量注意力机制GoogLeNet-GMP的行人重识别方法[J].计算机科学,2022,49(7):142-147.

[8] 张新峰,宋博.一种基于改进三元组损失和特征融合的行人重识别方法[J].计算机科学,2021,48(9):146-152.

[9] RADFORD A,KIM J W,HALLACY C,et al. Learning transferable visual models from natural language supervision [C]//38th International Conference on Machine Learning. Piscataway,NJ:IEEE,2021:8748-8763.

[10] JIANG D,YE M. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE,2023:2787-2797.

[11] BAI Y,CAO M,GAO D M,et al. Rasa:relation and sensitivity aware representation learning for text-based person search [EB/OL]. (2023-03-23)[2024-05-23]. https://arxiv.org/abs/2305.13653.

[12] 赵永威,李弼程,柯圣财.基于弱监督E2LSH和显著图加权的目标分类方法[J].电子与信息学报,2016,38(1):38-46.

[13] ZHENG L,SHEN L Y,TIAN L,et al. Scalable person re-identification:a benchmark [C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway,NJ:IEEE,2015:1116-1124.

[14] LI S,XIAO T,LI H S,et al. Person search with natural language description [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2017:1970-1979.

[15] SIMONYAN K. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10)[2024-05-23]. https://arxiv.org/abs/1409.1556.

[16] HOCHREITER S,SCHMIDHUBER J. Long short-term memory [J]. Neural computation,1997,9(8):1735-1780.

[17] CHEN T L,XU C L,LUO J B. Improving text-based person search by spatial matching and adaptive threshold [C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ: IEEE,2018:1879-1887.

[18] WANG Z,FANG Z Y,WANG J,et al. Vitaa:Visual-textual attributes alignment in person search by natural language [C]//Computer Vision-ECCV 2020:16th European Conference. Berlin:Springer,2020:402-420.

[19] GAO C Y,CAI G Y,JIANG X Y,et al. Contextual non-local alignment over full-scale representation for text-based person search [EB/OL]. (2021-01-08)[2024-05-23]. https://arxiv.org/abs/2101.03036.

[20] NIU K,HUANG Y,OUYANG W L,et al. Improving description-based person re-identification by multi-granularity image-text alignments [J]. IEEE Transactions on Image Processing,2020,29:5542-5556.

[21] CHEN S F,GE C J,TONG Z,et al. Adaptformer:adapting vision transformers for scalable visual recognition [J]. Advances in Neural Information Processing Systems,2022,35:16664-16678.

[22] KHATTAK M U,RASHEED H,MAAZ M,et al. MaPLe:multi-modal prompt learning [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE,2023:19113-19122.

[23] GAO P,GENG S J,ZHANG R R,et al. CLIP-adapter:better vision-language models with feature adapters [J]. International Journal of Computer Vision,2024,132(2):581-595.

[24] ZHOU K Y,YANG J K,LOY C C,et al. Learning to prompt for vision-language models [J]. International Journal of Computer Vision,2022,130(9):2337-2348.

[25] ITTI L,KOCH C,NIEBUR E. A model of saliency-based visual attention for rapid scene analysis [J]. IEEE Transactions on pattern analysis and machine intelligence,1998,20(11):1254-1259.

[26] ZHANG Y,LU H C. Deep cross-modal projection learning for image-text matching [C]//Proceedings of the European conference on computer vision (ECCV). Berlin:Springer,2018:686-701.

[27] DING Z F,DING C X,SHAO Z Y,et al. Semantically self-aligned network for text-to-image part-aware person re-identification [EB/OL]. (2021-08-09)[2024-05-23]. https://arxiv.org/abs/2107.12666.

[28] WEI L H,ZHANG S L,GAO W,et al. Person transfer GAN to bridge domain gap for person re-identification [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018:79-88.

[29] ZHU A C,WANG Z J,LI Y F,et al. DSSL:deep surroundings-person separation learning for text-based person retrieval [C]//Proceedings of the 29th ACM international conference on multimedia. New York:ACM,2021:209-217.

[30] ZHENG Z D,ZHENG L,GARRETT M,et al. Dual-path convolutional image-text embeddings with instance loss [J]. ACM Transactions on Multimedia Computing,Communications,and Applications,2020,

16(2):1-23.

[31] SARAFIANOS N,XU X,KAKADIARIS I A. Adversarial representation learning for text-to-image matching [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019:5814-5824.

[32] CHEN Y C,HUANG R,CHANG H,et al. Cross-modal knowledge adaptation for language-based person search [J]. IEEE Transactions on Image Processing, 2021,30:4057-4069.

[33] AGGARWAL S,RADHAKRISHNAN V B,CHAKRABORTY A. Text-based person search via attribute-aided matching [C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway,NJ:IEEE,2020:2617-2625.

[34] WU Y S,YAN Z Z,HAN X G,et al. LapsCore:language-guided person search via color reasoning [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 1624-1633.

[35] WANG Z J,ZHU A C,XUE J Y,et al. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold [C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022:1984-1992.

[36] CHEN Y H,ZHANG G Q,LU Y J,et al. TIPCB:a simple but effective part-based convolutional baseline for text-based person search [J]. Neurocomputing, 2022,494:171-181.

[37] SHAO Z Y,ZHANG X Y,FANG M,et al. Learning granularity-unified representations for text-to-image person re-identification [C]//Proceedings of the 30th ACM International Conference on Multimedia. New York:ACM,2022:5566-5574.

[38] LIU Y T,LI Y W,LIU Z M,et al. Clip-based synergistic knowledge transfer for text-based person retrieval [C]//ICASSP 2024—2024 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). Piscataway,NJ:IEEE,2024:7935-7939.

[39] YAN S L,DONG N,ZHANG L Y,et al. CLIP-driven fine-grained text-image person re-identification [J]. IEEE Transactions on Image Processing, 2023, 32: 6032-6046.

[40] YANG H M,ZHANG X Y,YIN F,et al. Convolutional prototype network for open set recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,44(5):2358-2370.

[41] LI S P,CAO M,ZHANG M. Learning semantic-aligned feature representation for text-based person search [C]//ICASSP 2022—2022 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). Piscataway,NJ:IEEE,2022:2724-2728.

GCTR:A Granularity-Unified Cross-Modal Text-person Retrieval Model

QIN Xiao[1,2,3],ZHANG Jinyong[1],GONG Yuanxu[1],WU Kunsheng[1],HUANG Haojie[1],CHUN Xin[1],YUAN Chang'an[2,3**]

(1. Guangxi Key Laboratory of Human-machine Interaction and Intelligent Decision,Nanning Normal University,Nanning,Guangxi,530100,China; 2. Guangxi Academy of Sciences,Nanning,Guangxi,530012,China; 3. Guangxi Regional Collaborative Innovation Center for Multi-Source Data Integration and Intelligent Processing,Guilin,Guangxi,541004,China)

Abstract:The existing text-image person retrieval tasks-lack the attention to the semantic connection between text and image and neglects the granularity difference between text and image features. In view of the problems above,a Granularity-unified Cross-Modal Text-person Retrieval model (GCTR) is proposed. Firstly, GCTR utilizes the contrastive language-image pre-training model,which possesses the cross-modal transfer

learning capability, to obtain text and image features with basic relevance. Secondly, a Cross-model Feature Enhancement module (CMFE) is proposed, which utilizes the Enhanced Cross-modal Feature Codebook (ECFC) to obtain text and image features with unified granularity, solving the problem of granularity difference between text and image features. Finally, a combination of local and global matching loss strategies is employed to optimize the model training process. The retrieval evaluation metrics achieved by GCTR on three public datasets CUHK-PEDES, ICFG-PEDES, and RSTPReid surpass those of existing state-of-the-art methods, demonstrating the effectiveness of the proposed GCTR model in cross-modal text-image person retrieval tasks.

Key words: cross-modal retrieval; text-image retrieval; person retrieval; visual language pre-training; granularity feature enhancement

责任编辑:陆　雁