基于关系挖掘的跨模态行人重识别

金昌胜，王海瑞
（昆明理工大学信息工程与自动化学院，昆明，650500）

摘要　基于文本的行人重识别模型通常依赖于全局特征对齐和局部特征对齐，但模态间和模态内的相关信息常被忽略。提出了一种基于关系挖掘的跨模态行人重识别方法，该方法包括双流主干网络、负相似度挖掘模块、关系编码器。首先，通过双流主干网络实现了全局和局部特征对齐；其次，通过负相似度挖掘模块提升了图像-文本对特征辨别的细粒度；最后，通过关系编码器模块分别学习图像和文本中隐含的关系信息，实现关系级别的特征对齐。在 CUHK-PEDES 数据集和 ICFG-PEDES 数据集上的实验结果证明，文中方法能够达到较高的识别精度。

关键词　行人重识别；多粒度图像；文本对齐；关系特征融合；卷积神经网络；全局特征；局部特征
DOI　10.3969/j.issn.2097-1915.2024.01.016
中图分类号　TP391.41　　**文献标志码**　A　　**文章编号**　2097-1915(2024)01-0106-09

A Cross-Modal Person Re-Identification Based on Relationship Mining

JIN Changsheng, WANG Hairui
(School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China)

Abstract　Aimed at the problems that while paying attention to the text-based person re-identification models often relying on global and local feature in alignment, very often the inter--modal and intra-modal correlations are in negative, a cross-modal pedestrian re-identification method is proposed based on relationship mining. The method includes a dual-stream network backbone, negative similarity mining module, and relationship encoder module. Firstly, the global and the local feature are in alignment through the dual-stream network backbone. Secondly the granularity of feature discrimination is enhanced by using the negative similarity mining module, and the similar incorrect results are filtered out. Finally, the relationship encoder module is utilized for respectively learning the latent relationship information in both the image and text, achieving relationship-level feature alignment. The experimental results on the CUHK-PEDES dataset and the ICFG-PEDES dataset show that this method achieves recognition accuracy higher.

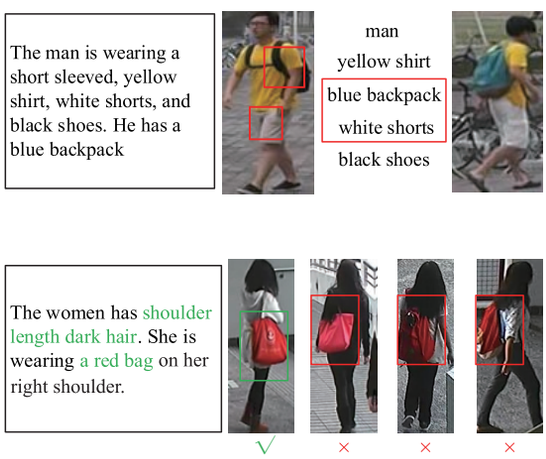
Key words　person re-identification; multi-granularity image; text alignments; relationship feature fusion; convolutional neural network; global feature; local feature

收稿日期：2023-04-27
基金项目：国家自然科学基金(61863016)
作者简介：金昌胜(1998－)，男，重庆人，硕士生，研究方向为深度学习、行人重识别。E-mail：changsheng@stu.kust.edu.cn
通信作者：王海瑞(1969－)，男，云南昆明人，教授，博士，研究方向为媒体智能技术、网络控制技术、嵌入式应用等。E-mail：hrwang88@163.com

引用格式：金昌胜，王海瑞. 基于关系挖掘的跨模态行人重识别[J]. 空军工程大学学报，2024，25(1)：106-114. JIN Changsheng, WANG Hairui. A Cross-Modal Person Re-Identification Based on Relationship Mining[J]. Journal of Air Force Engineering University, 2024, 25(1): 106-114.

基于文本的行人重识别(text-based person re-identification)是跨模态行人重识别的重要方向，它根据给定的文本描述从大型人物图像数据库中识别目标人物图像。在处理难以获得合适的目标人物照片的场景中寻找嫌疑人或寻找走失老人与儿童等问题时，这种方法非常有用。

早期的研究[1-4]一般采用卷积神经网络和递归神经网络将图像和文字分别编码为全局特征，然后计算特征距离作为其相似度。然而，一方面由于遮挡、穿着相似和视角差异等因素，更加稳健的视觉特征难以被提取；另一方面，不同图像或文本描述的相似性很高，会导致模态间差异远大于模态内差异。为了学习更有细粒度和判别性的特征，一些局部对齐的模型来匹配图像和文字描述的方法被提出[5-11]，这些方法表明准确提取和匹配局部特征可以提高模型的性能，但是，大多数方法都使计算复杂度大大提高，并忽略了模态间和模态内的相关信息。例如穿着相似的人容易和同一段文本描述相匹配，因此必须强调图像-文本间不匹配的关系信息，以降低负图像-文本对的整体相似性。如图 1(a)所示，2 张图像都能正确匹配“男性”“黄色短袖”“黑色鞋子”等词汇，但是实际上左图并非目标图像，因此应更加关注匹配错误的区域，如“蓝色双肩包”和“白色短裤”(图 1(a)中用中红色虚线框标识)。此外，图像和文本自身蕴含的关系信息对模型性能有影响，如图 1(b)所示，文本描述中的“右肩膀”和“红色背包”所蕴含的关系信息可以帮助模型很好地过滤掉“背包在左肩”或者“背着双肩包”等图像。



(a)模态间关系信息　　(b)模态内关系信息

图 1　关系信息对 ReID 的影响

针对目前基于文本的行人重识别中缺少模态内和模态间的关系信息挖掘的问题，本文提出了一种基于关系挖掘的跨模态行人重识别模型。该模型在全局特征对齐和局部特征对齐的基础上，通过负相似度挖掘实现更有细粒度的模态间特征辨别，从而过滤掉相似却错误的识别结果，最后通过特征关系编码器学习图像和文本中隐含的关系信息，实现关系级别的特征对齐。该模型在基于文本的行人重识别大型数据集 CUHK-PEDES 和 ICFG-PEDES 上均取得了较高的识别精度。

1　相关工作

目前，主流的跨模态检索算法的基本思想是从不同模态中提取有效特征来表示挖掘跨模态数据之间的相关性。早期研究[12]将深度神经网络与典型关联分析(CCA)相结合，提出深度典型关联分析(Deep CCA)来实现不同模态之间复杂的非线性变换关系；文献[13]为了充分利用训练数据的监督信息，同时设计了多个深度网络，形成层次化网络结构，通过约束模态内和模态间的相关性来学习图像和文本的表示；文献[14]设计改进的三元组损失函数用来监督训练过程；文献[15]中检测图像中的显著区域并计算每个区域与文本描述词之间的相似度以实现跨模态局部对齐；文献[16]进一步使用注意力机制来增强图像区域和文本词之间的相关性挖掘；文献[17]针对少样本场景，提出了一种跨模态记忆网络来实现跨模态检索；文献[18]为了解决跨模态训练数据不足的问题，结合对抗学习和知识迁移技术，实现了从单模态数据到跨模态数据的大规模数据迁移。上述方法虽然实现了全局或局部关系挖掘，但缺乏对模态间负面关系信息和模态内关系信息的挖掘和利用。

基于文本的行人重识别最早由 Li 等[1]提出，提出用 GNA-RNN 模型计算每个图像文本对之间的似度，并收集了一个名为 CUHK-PEDES 的大规模人物描述数据集。文献[19]提出了一种深度对抗图卷积网络通过图卷积操作学习图像区域和文本描述词之间的关系，有效地提高了跨模态表示的辨别力。文献[20]提出了一种 DSSL 模型，明确分离环境信息和人物信息，从而获得更高的检索精度；NAFS[6]使用阶梯式 CNN 和局部约束 BERT 在全尺度特征表示上进行联合对齐；ViTAA[8]从属性对齐的角度将图像和文本分解为属性组件，并使用细粒度匹配策略将身份与多个属性线索对齐，极大地提高了模型性能。然而由于对齐策略复杂、计算量巨大，这些方法仍无法简洁高效地处理基于文本的人物重识别问题。

2　建立模型

为了挖掘利用模态内和模态间的关系信息，并

设计简单高效的网络，本文提出了基于关系挖掘的行人重识别模型，包含 3 个组件：①双流主干网络：分别提取图像和文本的多尺度特征；②负相似度挖掘模块：增强图像-文本对中不匹配区域的关注度；③特征关系编码器：学习图像和文本中隐含的关系信息。整体结构如图 2 所示。

本文模型首先采用预训练的 ResNet50 和 BERT[21]模型分别对图像和文字进行全局特征提取；其次，利用 PCB[22]的分割策略分别对 CNN 输出的特征图和经过多分支残差组合得到的文本特征图进行水平分割，进而提取局部特征；然后通过负相似度挖掘以捕获更具有细粒度的图像-文本关系；最后，通过关系编码器获得图像和文本的关系信息，实现关系级别的特征对齐。

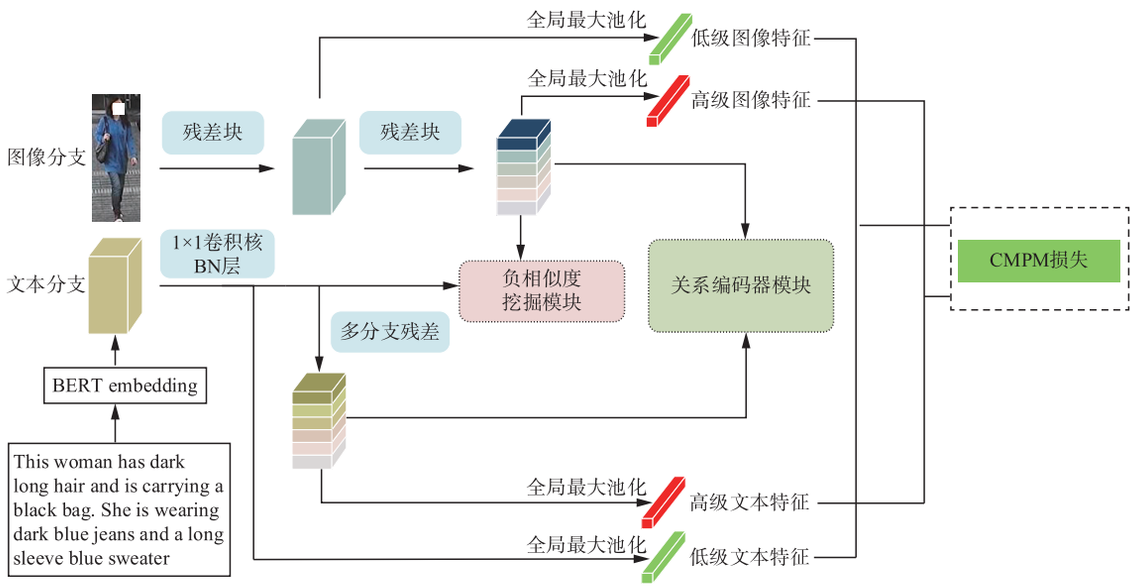


图 2 基于关系挖掘的行人重识别模型

2.1 多尺度特征

2.1.1 全局特征

对于全局图像特征，首先将所有图像调整为相同大小，其次，由于 ResNet50 原网络最后 2 个残差块可以捕获不同层次的视觉特征，本文采用在 ImageNet 上预训练的 ResNet50 网络分别获得完整的低级图像特征与高级图像特征，最后，在上述特征上应用全局最大池化操作分别捕获低级图像特征 I_{gl} 和高级图像特征 I_{gh}。

对于全局文本特征，首先采用在大型语料库训练好的 BERT 模型提取文本基础特征，然后将提取的特征经过一个 1×1 的卷积核、BN 层，最后通过全局最大池化操作捕获低级文本特征 T_{gl}。值得注意的是，在训练之前，为了确保文本长度的一致性，当文本长度大于 L 时，本文选择前 L 个标记，当文本长度小于 L 时，在文本末尾用零填充，并且在每个句子的开头和结尾插入[CLS]和[SEP]。而在训练时，BERT 参数会被固定，这种方法一方面可以有效利用 BERT 强大的语言建模能力，另一方面可以有效减少训练模型的时间消耗。

为了捕获高级文本特征，与 TIPCB[23]类似，本文所提的模型通过多分支残差卷积模块隐式提取与图像区域相对应的文本局部特征，文本特征每经过一层残差结构就会生成一级部分级特征，将所有生成的部分级文本特征进行拼接，再应用全局最大池化操作得到最终的高级文本特征 T_{gh}。具体的，多分支残差卷积模块由 6 层残差结构组成，每层残差结构由 3 组瓶颈层组成，第 1 组瓶颈层和第 3 组瓶颈层由 1×1 的卷积核和 BN 层组成，第 2 组瓶颈层由 1×3 的卷积核和 BN 层组成。

2.1.2 局部特征

受到 PCB[22]的启发，本文采用分割策略对经过双流网络得到的高级图像特征和高级文本特征进行水平分割，局部图像特征为：

$$f_p^I=\{f_{p1}^I,f_{p2}^I,\cdots,f_{pK}^I\} \tag{1}$$

局部文字特征为：

$$f_p^T=\{f_{p1}^T,f_{p2}^T,\cdots,f_{pK}^T\} \tag{2}$$

式中：K 为水平切割条数。文中 K 取 6。

2.2 负相似度挖掘

如前文所述，负相似度挖掘的目标是为了降低负图像-文本对的整体相似度，以有效的方式突出不匹配的图像-文本对对模型匹配结果的影响。

如图 3 所示，为了通过相似度计算挖掘不匹配的图像-文本区域，首先通过 1×1 的卷积层将图像特征 $I=\{i_k\}_{k=1}^K$ 和文本特征 $T=\{t_j\}_{j=1}^N$ 映射到公共特征空间，然后计算不同区域的匹配相似度，其计算式为：

$$s_{k,j}=\frac{\boldsymbol{\theta}(i_k)^{\mathrm{T}}\varphi(t_j)}{\|\boldsymbol{\theta}(i_k)\|\ \|\varphi(t_j)\|},k\in[1,K],j\in[1,N] \quad (3)$$

式中：$\boldsymbol{\theta}(i_k)=W_\theta i_k$，$\varphi(t_j)=W_\varphi t_j$。

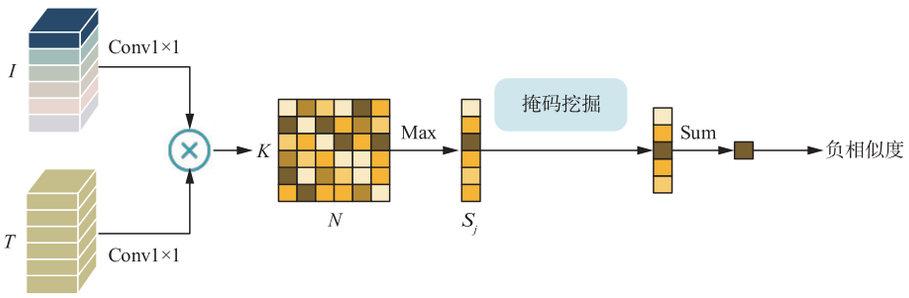


图 3　负相似度挖掘

基于文本的行人重识别，图像区域几乎都可以被文本匹配到，因此对相似度矩阵进行最大池化操作 $s_j=\max(s_{k,j})$ 搜寻与所有图像区域匹配相似度最低的文本区域，以此证明该文本与任何图像区域都不匹配，最后，为了增强判断的准确性，通过掩码挖掘过滤掉正相似度，并通过 Sum 求得最终的负相似度，其计算式为：

$$S_{\mathrm{neg}}=\sum_{j=1}^{N}M_{\mathrm{mining}}(s_j) \quad (4)$$

式中：M_{mining} 表示输入为正数时，输出为 0；输入为负数时，输出保持不变。

2.3　特征关系编码器

关系编码器可以隐式捕获图像和文本的关系信息，从而实现关系级别的特征对齐。

其模型如图 4 所示，首先将由 K 个局部特征 $p_{k=1}^{K}$ 组成的单模态特征 f 水平拼接得到 $\boldsymbol{f}_{\mathrm{hc}}$。（由于图像特征与文本特征都是经过关系编码器模块处理，故下文省略特征图下标 I 和 T）

$$\boldsymbol{f}_{\mathrm{hc}}=\begin{pmatrix}p_1 & p_1 & \cdots & p_1\\ p_2 & p_2 & \cdots & p_2\\ \vdots & \vdots & & \vdots\\ p_K & p_K & \cdots & p_K\end{pmatrix}=[f,f,\cdots,f] \quad (5)$$

然后通过下列转置相加计算初步构建 2 个局部区域之间的关系特征：

$$\boldsymbol{f}_{\mathrm{pc}}=\boldsymbol{f}_{\mathrm{hc}}+\boldsymbol{f}_{\mathrm{hc}}^{\mathrm{T}},\boldsymbol{f}_{\mathrm{pc}}\in\boldsymbol{R}^{C_1\times K\times K} \quad (6)$$

为了挖掘更加细腻的相邻区域的关系信息，如图 4 中的相邻关系分支所示，本文构建了 1×1 的卷积层和 ReLu 激活函数组成的组合层，将 $\boldsymbol{f}_{\mathrm{pc}}$ 通过 N 层组合层（图 4 中 $N=3$，得到低级的相邻关系特征 $f_{\mathrm{l\_ad\_r}}\in\boldsymbol{R}^{C_{\mathrm{r}}\times K\times K}$；最后，通过与(6)式类似的计算获得最终的相邻关系特征：

$$\boldsymbol{f}_{\mathrm{ad\_r}}=\{r_{i,j}^{\mathrm{ad}}\}_{i=1,j=1}^{K,K}=\boldsymbol{f}_{\mathrm{l\_ad\_r}}+\boldsymbol{f}_{\mathrm{l\_ad\_r}}^{\mathrm{T}} \quad (7)$$

式中：$\boldsymbol{f}_{\mathrm{ad\_r}}\in R^{C_{\mathrm{r}}\times K\times K}$；$\boldsymbol{f}_{\mathrm{ad\_r}}$ 为相邻关系特征；$r_{i,j}^{\mathrm{ad}}$ 为相邻区域的关系表示。

为了挖掘更加细腻的全局区域的关系信息，如图 4 中全局关系分支所示，本文采用与相邻关系分支类似的结构，获得最终的全局关系特征可以表示为：

$$\boldsymbol{f}_{\mathrm{gr}}=\{r_{i,j}^{\mathrm{g}}\}_{i=1,j=1}^{K,K}=\boldsymbol{f}_{\mathrm{l\_gr}}+\boldsymbol{f}_{\mathrm{lgr}}^{\mathrm{T}} \quad (8)$$

与式(7)相似，其中 $\boldsymbol{f}_{gr}\in\boldsymbol{R}^{C_{\mathrm{r}}\times K\times K}$，$f_{\mathrm{gr}}$ 为全局关系特征，$r_{i,j}^{\mathrm{g}}$ 代表全局区域的关系表示，$f_{\mathrm{l\_gr}}$ 代表经过组合层之后得到的低级全局关系特征；值得注意的是，全局关系和相邻关系的不同之处在于全局关系分支中的组合层使用的是 3×3 的空洞卷积，空洞卷积可以扩大感受野，更好地捕获多尺度上下文关系信息，同时可以很好地降低计算复杂度。

最后，将经过上下 2 路分支分别获得的相邻关系特征 $\boldsymbol{f}_{\mathrm{adr}}$ 与全局关系特征 $\boldsymbol{f}_{\mathrm{gr}}$ 进行拼接操作得到最终的关系特征 F_{r}，其计算式为：

$$F_{\mathrm{r}}=[\boldsymbol{f}_{\mathrm{adr}},\boldsymbol{f}_{\mathrm{gr}}] \quad (9)$$

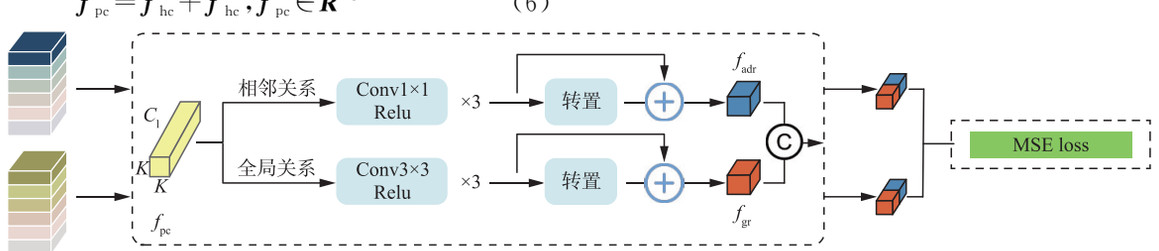


图 4　特征关系编码器

2.4　损失函数

多个研究证明，不同粒度的特征对齐可以有效的减少图像和文本之间的特征差异性。受到相关研究的启发，本文在低级、高级和局部级特征上选择跨模态投影匹配(CMPM)损失[24]监督网络训练；在负相似度挖掘模块中，本文采用排序损失约束模态间差异、降低负样本相似度；在特征关系编码器中，采用 MSE 损失实现关系级别的特征对齐。

2.4.1 CMPM 损失

对于全局特征而言，假设输入的图像-文本对数量为 N，将全局图像特征 I 与全局文本特征 T 组合起来得到图像-文本对(由于全局低级特征与全局高级特征对齐方式类似，故省略其区分下标)，其表达式为：

$$\{(\boldsymbol{I}_i, T_j), y_{i,j}\}_{j=1}^{N} \tag{10}$$

式中：$y_{i,j}$ 表示第 i 个图像特征 $\boldsymbol{I}_i$ 与第 j 个文本特征 T_j 匹配情况，其匹配的概率为：

$$p_{i,j} = \frac{\exp((\boldsymbol{I}_i)^{\mathrm{T}}\overline{T_j})}{\sum_{k=1}^{N}\exp((\boldsymbol{I}_i)^{\mathrm{T}}\overline{T_k})}, \overline{T_j} = \frac{T_j}{\|T_j\|} \tag{11}$$

由此，可以计算出正确匹配图片 $\boldsymbol{I}_i$ 的损失值为：

$$L_i^I = \frac{1}{N}\sum_{j=1}^{N} p_{i,j}\log\left(\frac{p_{i,j}}{q_{i,j}+\varepsilon}\right), q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^{N} y_{i,k}} \tag{12}$$

式中：$q_{i,j}$ 为归一化的真实匹配概率，为了避免数值问题，添加极小数 ε 在分母之中。于是，图像到文本的 CMPM 损失可以计算为：

$$L_{i2t} = \frac{1}{N}\sum_{i=1}^{N} L_i^I \tag{13}$$

同理可以得出文字到图像的 CMPM 损失 L_{t2i}。故全局 CMPM 损失为：

$$L_{\mathrm{CMPM}} = L_{i2t} + L_{t2i} \tag{14}$$

其中，全局 CMPM 损失可分为全局高级 CMPM 损失 $L_{\mathrm{CMPM}}^{\mathrm{gh}}$ 和全局低级 CMPM 损失 $L_{\mathrm{CMPM}}^{\mathrm{gl}}$。

对于局部特征而言，首先计算出图像与文本分割后对应区域的 CMPM 损失，然后计算总的局部 CMPM 损失：

$$L_{\mathrm{CMPM}}^{p} = \sum_{k=1}^{K} L_{\mathrm{CMPM}}^{p_k} \tag{15}$$

式中：K 为水平切割的条数。

综上，最终的 CMPM 损失为：

$$L_{\mathrm{CMPM}} = L_{\mathrm{CMPM}}^{\mathrm{gh}} + L_{\mathrm{CMPM}}^{\mathrm{gl}} + L_{\mathrm{CMPM}}^{p} \tag{16}$$

2.4.2 Ranking 损失

为了抑制模型对错误匹配结果的相似区域的关注度，本文采用排序损失。具体而言，首先计算图像与文本的局部相似度：

$$s_1 = \frac{{I_1}^T T_1}{\|I_1\|\|T_1\|} \tag{17}$$

其次，由于在第 2.2 节已经计算过样本负相似度 S_{neg}，故局部特征的排序损失为：

$$\begin{aligned} L_{\mathrm{Ranking}} = \max\ & (\alpha - s_{1\_\mathrm{neg}}(I_+, T_+) + \\ & s_{1\_\mathrm{neg}}(I_+, T_-), 0) + \\ \max(\alpha & - s_{1\_\mathrm{neg}}(I_+, T_+) + \\ & s_{1\_\mathrm{neg}}(I_+, T_-), 0) \end{aligned} \tag{18}$$

式中：$s_{1\_\mathrm{neg}} = s_1 + s_{\mathrm{neg}}$，$\alpha$ 代表排序损失的边界值，(I_+, T_+) 代表匹配的图像-文本对，(I_+, T_-) 或 (I_-, T_+) 代表不匹配的图像-文本对。

2.4.3 MSE 损失

对于关系特征而言，MSE 损失函数可以缩小模态间关系特征的差异，其计算如下：

$$L_{\mathrm{r}} = \frac{1}{K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}(r_{i,j}^{I} - r_{i,j}^{T})^2 \tag{19}$$

式中：$r_{i,j}^{I}$ 为经过关系编码器模块之后的图像中任意 2 个区域的关系，$r_{i,j}^{T}$ 为经过关系编码器模块之后的文本中任意 2 个区域的关系。

2.4.4 最终损失

通过前面的计算，分别得到了 CMPM 损失、Ranking 损失和 MSE 损失。故最终损失函数为：

$$L = L_{\mathrm{CMPM}} + L_{\mathrm{Ranking}} + L_{\mathrm{r}} \tag{20}$$

式中：L_{CMPM} 为多尺度 CMPM 损失之和。

3 实验

3.1 数据集

CUHK-PEDES[6]最早是唯一用于基于文本的行人重识别的大型数据集，现在也是主流的数据集。该数据集包含 13 003 名不同行人的 40 206 张图像，每个行人图像都用 2 个可变长度的描述性句子进行注释。在实验中，本文采用与文献[6]中相同的数据集拆分方法，得到一个包含来自 11 003 个不同行人的 34 054 张图像的训练集，一个包含来自 1 000 个不同行人的 3 078 张图像的验证集，其余 3 074 张图像用作测试集。

ICFG-PEDES[25]数据集是一个新收集的数据集，它包含 4 102 人的 54 522 张图，每个图像仅用 1 个文本描述。与 CUHK-PEDES 相比，ICFG-PEDES 拥有细粒度更高的文本描述。ICFG-PEDES 数据集被分为训练集和测试集，分别使用 3 102 人的 34 674 个图像-文本对和其余 1 000 人的 19 848 个图像-文本对。

3.2 评价指标

为了评估行人重识别模型的性能，本文采用了经典评价指标累计匹配曲线(cumulative matching characteristic，CMC)和平均精度(mean average precision，mAP)。rank-N 指模型在一系列结果中前 N 个包含正确行人的概率；mAP 则表示正确结果在结果排序中的前列程度。通过综合使用这 2 个指标，可以更全面地测量模型性能。

3.3 实验设置

训练过程在基于文本的跨模态行人重识别数据

集 CUHK-PEDES 和 ICFG-PEDES 上进行。在图像分支上采用预训练的 ResNet50 提取视觉特征，在文本分支上采用预训练的 BERT 模型。输入图像的尺寸统一调整为 384×128，文本长度统一为 64。预训练的 ResNet50 模型和本模型其他参数一起更新，而预训练的 BERT 参数则被冻结。设置局部特征分块数 $K=6$，设置关系编码器模块中的组合层个数 $N=3$，设置排序损失中的 $\alpha=0.2$。在训练过程中训练批次设置为 32，选择 Adam 优化器来优化模型，训练 90 个 epoch，学习率在开始训练时设置为 3×10^{-3}，每 30 个 epoch 衰减到原来的 1/3。本文模型都是在基于深度学习的框架 PyTorch 下实现，实验设备为单块显存为 12 GB 的 NVIDIA GeForce GTX 3060 GPU。

3.4　实验结果分析

在 CUHK-PEDES 数据集上将本文模型与其他模型进行比较。主流模型可以大致分为：①全局匹配方法，如 GNA-RNN[1]、IATV[26]、Dual Path[2] 和 CMPM-CMPC[24]；② 全局-局部匹配方法，如 PMA-VGG[7]、PMA-ResNet[7]、MIA[9]、ViTAA[8]、NAFS[6]、TIPCB[23]；③ 其他方法，如 CAIBC[27]、AXM-Net[28] 和 TFAF[29]。实验结果具体如表 1 所示，通过分析可知：

表 1　在 CUHK-PEDES 数据集上与其他方法比较

方法	类型	图像特征	文本特征	rank-1	rank-5	rank-10
GNA-RNN[1]	全局	VGG	LSTM	19.05		53.64
IATV[26]	全局	VGG	LSTM	25.94		60.48
DualPath[2]	全局	ResNet50	TEXT CNN	44.40	66.26	75.07
CMPM-CMPC[24]	全局	ResNet152	LSTM	49.37		79.27
PMA-VGG[7]	全局＋局部	VGG	LSTM	47.02	68.54	78.06
PMA-ResNet[7]	全局＋局部	ResNet50	LSTM	53.81	73.54	81.23
MIA[9]	全局＋局部	VGG	GRU	48.00	70.70	79.30
MIA[9]	全局＋局部	ResNet50	GRU	53.10	75.00	82.90
ViTAA[8]	全局＋局部	ResNet50	LSTM	55.97	75.84	83.52
NAFS[6]	全局＋局部	ResNet50	BERT	59.94	79.86	86.70
TIPCB[23]	全局＋局部	ResNet50	LSTM	60.82	80.88	87.74
TIPCB[23]	全局＋局部	ResNet50	BERT	63.63	82.82	89.01
CAIBC[27]		ResNet50	BERT	64.43	82.87	88.37
AXM-Net[28]			BERT	64.44	80.52	86.77
TFAF[29]		Transformer	BERT	65.69	84.75	89.93
本文方法	全局＋局部＋关系	ResNet50	BERT	66.37	85.46	90.78

1）使用全局特征加上局部特征的多尺度匹配方法，相比于仅使用全局匹配方法，能够捕获更具有细粒度的特征，从而达到更好的识别效果。

2）优秀的模态特征提取方法对模型性能有显著提升。例如 MIA 方法在从使用 VGG 提取图像特征到使用 ResNet50 提取图像特征后，rank-1 精度从 48.00%提升到 53.10%；同样的，TIPCB 从使用 LSTM 提取文本特征到使用 BERT 提取文本特征后，rank-1 精度从 60.82%提升到 63.63%。

3）本文提出的模型应用了全局、局部、关系的特征对齐机制，并且采用了负相似度挖掘的方法实现更有细粒度的关系挖掘，通过在 CUHK-PEDES 数据集上实验，实现了较高的识别精度提升，rank-1、rank-5、rank-10 分别达到了 66.37%、85.46%、90.78%。相比于图像特征提取使用了金字塔视觉 Transfomer 的 TFAF[29]，本文模型在 rank-1、rank-5、rank-10 精度上仍旧提升了 0.68%、1.71%、1.85%。

为了验证模型的泛化性，本文还在 ICFG-PEDES 数据集上进行了实验，实验结果如表 2 所示，其实验结果与 2022 年的工作 IVT[31] 相比，rank-1、rank-5、rank-10 精度分别提升了 0.58%、2.29%、2.81%。

表 2　在 ICFG-PEDES 数据集上不同方法试验结果对比

方法	年份	rank-1	rank-5	rank-10
DualPath[2]	2020	38.99	59.44	68.41
CMPM/C[24]	2018	43.51	65.44	74.26
MIA[9]	2020	46.49	67.14	75.18
SCAN[30]	2018	50.05	69.65	77.21
ViTAA[8]	2020	50.98	68.79	75.78
SSAN[25]	2021	54.23	72.63	79.53
TIPCB[23]	2022	54.96	74.72	81.89
IVT[31]	2022	56.04	73.60	80.22
本文方法	2023	56.62	75.89	83.03

3.5 消融实验

为了进一步验证本文提出模型的有效性，实验均在最常用的 CUHK-PEDES 数据集上进行。

首先，在整体模型上分别删除局部特征对齐模块、负相似度挖掘模块和关系编码器模块，其他参数保持不变，实验结果如表 3 所示，其结果说明：

1）局部特征对齐可以有效提升识别精度：在只使用全局特征对齐的情况下，rank-1 和 mAP 只有 56.24%与 48.45%，而加上局部特征对齐之后，rank-1 和 mAP 分别提升了 3.67%和 3.00%达到了 59.91%和 51.45%。

2）负相似度挖掘和关系编码器的有效性。在使用全局特征和局部特征对齐的基础之上，本文提出的负相似度挖掘模块和关系编码器模块分别在 rank-1 和 mAP 上提升了 4.03%、4.98% 和 2.34%、2.46%。

3）负相似度挖掘和关系编码器可以很好地配合实现对齐特征。在同时采用负相似度挖掘和关系编码器的情况下 rank-1 和 mAP 分别提升了 6.46%和 5.25%。负相似度挖掘和关系编码器的共同使用可以有效提升识别精度，这是由于负相似度挖掘可以很好地探索模态间的关系信息，使模型关注模态间不匹配的区域，从而过滤掉相似却错误的结果，而关系编码器可以很好地探索模态内的关系信息，实现更细腻的关系级别的特征对齐。

表 3 模型不同模块对实验结果的影响 %

L_{CMPM}^{p}	L_{Ranking}	L_{r}	rank-1	rank-5	rank-10	mAP
			56.24	77.85	86.24	48.45
√			59.91	80.81	87.18	51.45
√		√	62.25	82.96	89.43	53.91
√	√		63.94	83.49	89.86	56.43
√	√	√	66.37	85.46	90.78	56.70

其次，为了验证关系编码器的相邻关系分支和全局关系分支对模型整体性能的影响，分别采用相邻、全局和相邻-全局对同样的数据集进行训练和测试。实验结果如表 4 所示，单独使用相邻关系挖掘或者全局关系挖掘，rank-1 精度分别为 64.27%和 64.53%，而同时使用相邻关系挖掘和全局关系挖掘，rank-1 和 mAP 分别达到 66.37%和 56.70%。由此可见，同时使用相邻关系挖掘和全局关系挖掘可以更全面精确地捕获模态内的关系信息，识别效果更好。

表 4 关系编码器中上下分支对模型性能影响 %

相邻	全局	rank-1	rank-5	rank-10	mAP
√		64.27	83.50	89.75	55.44
	√	64.53	84.59	90.40	56.36
√	√	66.37	85.46	90.78	56.70

同时，为了验证关系编码器中的组合层个数对于模型性能的影响，本文将 $N\in\{1,2,3,4,5,6\}$ 对 CUHK-PEDES 数据集进行训练和测试，实验结果如图 5 所示。从图 5 的实验结果可以得出：模型的性能受组合层数量影响较为明显，组合层数量过低或过高都会影响模型性能。当 $N=1$ 和 $N=6$ 时，rank-1 的精度只有 65.56%和 65.59%；而当 $N=3$ 时，模型性能较好，rank-1 的精度为 66.37%。综合考虑模型识别精度和复杂度，本文模型中 N 取值为 3。

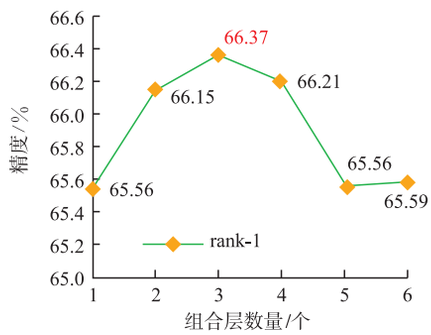


图 5 组合层数量 N 对模型性能影响

最后，对于局部特征分块数 K，为了验证不同的分割粒度对于模型性能的影响，本文将 $K\in\{1,2,3,4,5,6\}$ 对相同的数据集进行训练和测试，实验结果如图 6 所示，其表明当 $K=6$ 时，模型性能最好；当 $K=1$ 时，算法接近于全局特征匹配模型，性能显著下降；当 K 过大时，分割细粒度过高，无法捕获完整的局部特征，性能同样有所下降。综上，本文模型中的 K 取 6，这与 PCB 中的实验结果也是一致的。

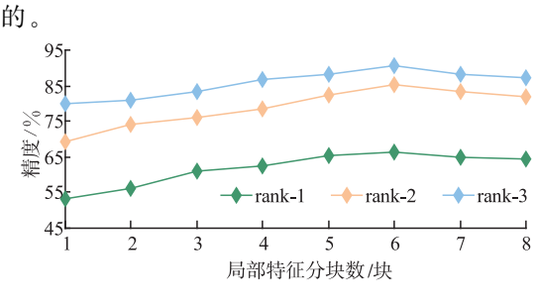


图 6 局部特征分块数 K 对模型性能影响

4 结语

为了捕获模态内相关信息、缩小模态间差异，本文提出了一种基于关系挖掘的跨模态行人重识别方法，其中包含双流主干网络、负相似度挖掘模块、关系编码器 3 个模块。其中，双流主干网络通过残差块的结构捕获多尺度特征；负相似度挖掘模块挖掘图像-文本不匹配的关系信息，降低负样本整体相似度；特征关系编码器捕获图像以及文本模态内关系信息实现更细腻的关系特征对齐。实验结果表明，本文提出的模型有着不复杂的结构和良好的识别精

度。如何更加简单高效的提取模态内特征和缩小模态间差异,以及模态内的关系信息对于跨模态问题中遮挡、背景干扰和姿态变化等问题是否有改善,都是今后重点研究的方向。

参考文献

[1] 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). Honolulu, HI: IEEE, 2017: 1970-1979.

[2] 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): 51.

[3] NIU K, HUANG Y, WANG L. Fusing Two Directions in Cross-Domain Adaption for Real Life Person Search by Language [C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South): IEEE, 2019: 1815-1818.

[4] WANG YY, BO C J, WANG D, et al. Language Person Search with Mutually Connected Classification Loss [C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, 2019: 2057-2061.

[5] 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). Lake Tahoe, NV:IEEE, 2018: 1879-1887.

[6] GAO C, CAI G Y, JIANG X Y, et al. Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search [J]. ArXiv Preprint : 2101.03036, 2021.

[7] JING Y, SI C Y, WANG J B, et al. Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34 (7): 11189-11196.

[8] WANG Z, FANG Z Y, WANG J, et al. ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language[C]//European Conference on Computer Vision. Cham: Springer, 2020: 402-420.

[9] 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.

[10] CHEN D P, LI H S, LIU X H, et al. Improving Deep Visual Representation for Person Re-Identification by Global and Local Image-Language Association [C]// Proceedings of the 15th European Conference on Computer Vision. Munich:Springer, 2018: 56-73.

[11] AGGARWAL S, BABU R V, CHAKRABORTY A. Text-Based Person Search via Attribute-Aided Matching [C]//2020 IEEE Winter Conference on Applications of Computer Vision. Snowmass Village, CO: IEEE, 2020: 2617-2625.

[12] ANDREW G, ARORA R, BILMES J, et al. Deep Canonical Correlation Analysis [J]. 30thInternational Conference on Machine Learning, 2013(PART 3): 2284-2292.

[13] PENG Y X, HUANG X, QI J W. Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks [C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16). New York: AAAI, 2016:3846-3853.

[14] FAGHRI F, FLEET D J, KIROS J R, et al. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives [EB/OL]. (2018-07-29) [2023-04-01]. https://doi.org/10.48550/arXiv.1707.05612.

[15] JIANG X Y, WU F, LI X, et al. Deep Compositional Cross-Modal Learning to Rank via Local-Global Alignment [C]// Proceedings of the 23rd ACM International Conference on Multimedia. Brisbane, Australia: ACM, 2015: 69-78.

[16] YU J, LU Y H, ZHANG W F, et al. Learning Cross-Modal Correlations by Exploring Inter-Word Semantics and Stacked Co-Attention [J]. Pattern Recognition Letters, 2020, 130: 189-198.

[17] HUANG Y, WANG L. ACMM: Aligned Cross-Modal Memory for Few-Shot Image and Sentence Matching [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 5774-5783.

[18] HUANG X, PENG Y X, YUAN M K. MHTN: Modal-Adversarial Hybrid Transfer Network for Cross-Modal Retrieval [J]. IEEE Transactions on Cybernetics, 2020, 50(3): 1047-1059.

[19] LIU J W, ZHA Z J, HONG R C, et al. Deep Adversarial Graph Attention Convolution Network for Text-Based Person Search [C]//Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM, 2019: 665-673.

[20] 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.

[21] DEVLIN J, CHANG M W, LEE K. et al. BERT:

Pre-Training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. (2019-05-24) [2023-04-01]. https://doi.org/10.48550/arXiv.1810.04805.

[22] SUN Y F, ZHENG L, YANG Y, et al. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline) [C]// Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]:Springer, 2018: 480-496.

[23] 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.

[24] ZHANG Y, LU H C. Deep Cross-Modal Projection Learning for Image-Text Matching [C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]:Springer, 2018: 686-701.

[25] 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)[2023-04-01]. https://doi.org/10.48550/arXiv.2107.12666.

[26] LI S, XIAO T, LI H S, et al. Identity-Aware Textual-Visual Matching with Latent Co-Attention [C]// 2017 IEEE International Conference on Computer Vision (ICCV). Venice:IEEE, 2017: 1890-1899.

[27] WANG Z J, ZHU A C, XUE J Y, et al. CAIBC: Capturing All-round Information Beyond Color for Text-Based Person Retrieval [C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 5314-5322.

[28] FAROOQ A, AWAIS M, KITTLER J, et al. AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-Identification [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(4): 4477-4485.

[29] LI S Z, LU A D, HUANG Y, et al. Joint Token and Feature Alignment Framework for Text-Based Person Search [J]. IEEE Signal Processing Letters, 2022, 29: 2238-2242.

[30] LEE K H, CHEN X, HUA G, et al. Stacked Cross Attention for Image-Text Matching [C]// Proceedings of the European Conference on Computer Vision. Munich,Germany:ACM, 2018: 212-228.

[31] SHU X J, WEN W, WU H Q, et al. See Finer, See More: Implicit Modality Alignment For Text-Based Person Retrieval [M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2023: 624-641.

[32] 罗浩. 深度学习时代的行人重识别技术 [J]. 人工智能, 2019, 6(2): 40-49.

[33] 朱宽堂, 张建勋, 谭暑秋. 基于全局特征和多种局部特征的行人重识别 [J]. 微电子学与计算机, 2022, 39(2): 43-50.

(编辑:徐楠楠)