文章编号：1001-9014(2024)06-0872-08 DOI：10. 11972/j. issn. 1001-9014. 2024. 06. 018

基于人体关键点引导注意力的红外-可见光行人重识别

于 鹏[1]，田小建[2]，齐 楠[1]，朴 燕[1*]
（1. 长春理工大学，电子信息工程学院，吉林 长春 130022；
2. 吉林大学，电子科学与工程学院，吉林 长春 130012）

摘要：行人重识别是从多个数据源中检索出指定目标的任务。红外(IR)和可见光(VIS)的图像差距较大，可见光和红外图像跨模态检索是主要挑战之一。为了能在弱光或夜间也具备相同的检索能力，需要结合红外图像的跨模态模型实现判断。本文提出一个通过人体关键点引导注意力的新方法，通过关键点引导将全局特征拆分为局部特征，再用生成的局部掩码重新训练原模型，强化对不同局部信息的注意力。使用这个方法，模型可以更好地理解和利用图像中的关键部位，从而提升行人重识别任务的准确率。

关 键 词：人工智能；行人重识别；红外；注意力；自监督
中图分类号：TP391.4 **文献标识码**：A

Attention guided by human keypoint for infrared-visible person re-identification

YU Peng[1], TIAN Xiao-Jian[2], QI Nan[1], PIAO Yan[1*]
(1. School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China;
2. College of Electronic Science and Engineering, Jilin University, Changchun 130012, China)

Abstract: Person re-identification is the task of retrieving a specified target from multiple data sources. The difference between infrared (IR) and visible light (VIS) images is large, and cross-modal retrieval of visible light and infrared images is one of the main challenges. In order to have the same retrieval ability even in low light or at night, the judgment needs to be achieved by combining cross-modal modeling of infrared images. In this paper, we propose a new method of guiding attention through human keypoints, where global features are split into local features by keypoint guidance, and then the original model is retrained with the generated local masks to strengthen the attention to different local information. Using this method, the model can better understand and utilize the key regions in the image, thus improving the accuracy of the person re-identification task.

Key words: artificial intelligence, person re-identification, infrared, attention, self-supervised

引言

行人重识别(re-ID)是从多个数据源中检索出指定目标的子检索任务。随着深度学习领域的快速发展和更多场景的数据集发布，基于深度神经网络的行人重识别方法取得了巨大的进展，可见光-红外行人重识别(RGB-IR Person ReID)是在行人重识别任务基础上，使用了红外图像去检索或被检索的任务。现有的多数工作都是通过手动设计的特征选择模块来实现性能提高的，目前已有模型主要是利用骨干网提取特征，再从特征中提取所需的关

收稿日期：2024-08-09，**修回日期**：2024-09-10 **Received date**：2024-08-09，**Revised date**：2024-09-10
基金项目：吉林省自然科学基金(20210101180JC)，吉林省科技厅科技发展计划项目(20180623039TC)。
作者简介(Biography)：于鹏(1988-)，男，吉林省长春市人，博士研究生，主要研究领域为机器视觉。E-mail：yup1212@mails. cust. edu. cn
*通讯作者(Corresponding author)：E-mail：piaoyan@cust. edu. cn

键信息。由于可见光和红外图像的波长范围不同、通道数不同、需要使用不同的预处理方法,同时也存在行人遮挡及背景干扰等问题,检索任务的源图像和目标图像差别较大,效果很好的单模态方法很难适应多模态图像间的相互检索,所以跨模态的红外行人重识别任务具有挑战性。

由于跨模态特征差别大,常见的一种思路是将多种不同模态的图像映射到同一个特征空间中,通过训练最终输出为相同的类别。相比于同源数据,由于初始特征差别较大,特征对齐也变得更为重要。

使用特征金字塔可以有效定位原图特征位置。CCVID[1]使用了一种基于衣服的对抗性损失函数(Clothes-based Adversarial Loss)(CAL),根据数据集的特征设计的专用特征来提取模型GCP[2]。RGA[3]使用注意力机制可以有效的提升结果。Jiaxu Miao[4]等人的结合PCB和人体关键点检测的方法,进一步提升了PCB[5]的准确率。

在重识别任务中,普遍存在检索数据集太少、实践时泛用性与训练时不一致以及数据难以获取的问题,数据量的不足,使跨模态时检索难度加大。目前的人工智能模型不像人脑那样直接利用已经训练好的脑功能区处理新问题,一些相关尝试例如,Zhu, K.[6]提出了一种针对Transformer自动对齐转换器(AAformer)来自适应地将图像的“片标记”分组为不同的子集。所有任务都是完全独立的学习过程,面向过程变成转为面向对象编程,将一个完整的任务拆解成最小的语义“类”,用对象组成功能块,从而实现复用迁移等操作。

使用有监督训练往往比无监督训练的准确率更高,但有监督模型的迁移能力不足,遇到新场景往往需要重新学习,有些模型是针对数据特化的模型。随着更多的模型被训练出来,对已有模型的再利用的迁移学习会更有发展潜力。Transformer的方法和PCB方法都证明了分区域学习的有效性,然而目前很多方法并不使用已有训练好的模型,而是重新训练。虽然有很多有效的方法,但一直存在一个问题,即无监督很多是在相同的源域和目标域进行训练和测试,实际场景被查询的人既不在训练数据集中也不在测试数据集中。

如图1所示,由于红外图像数据集中不一定是可见光摄像机和红外摄像机同时拍摄,所以跨模态模型在融合时存在严重的特征不对齐。

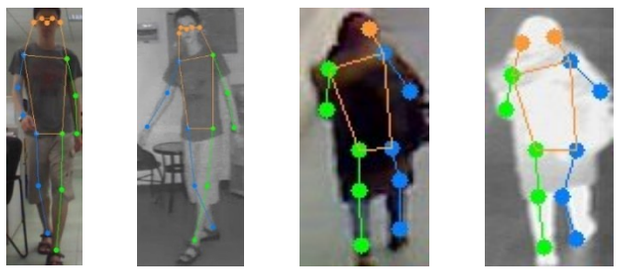

图1　红外数据集上人体关键点示意图

Fig. 1　Schematic representation of key points of the human body on the infrared dataset

Ci Yuanzheng[7]提出以人为中心的感知统一模型(UniHCP)。UniHCP统一并同时处理五个不同的以人为本的任务,即姿态估计、语义部分分割、行人检测、ReID和人属性识别。为了解决上述问题,常见的方法有调整模型使之特征对齐,例如,Che J.[8]的模型利用姿态实现特征对齐。Gu J.[9]提出了一个空间对齐模块(SAM),以增强模型面对不同来源图像的注意力一致性。Pan H.[10]提出了一种多头全注意力池化(MHFAPool)方法。Zhang G.[11]先设计了一个基于时间的相机对比学习(TCCL)模块,来促进模型学习,去掉背景信息。知识蒸馏方法[12]是从一个较大的模型(教师网络)的知识迁移到一个较小的模型(学生网络)上,从而实现剔除不需要的特征使模型轻量。 虽然目的是为了轻量化,但经过人们的实践发现也适用于大多数任务。Liu X.[13]引入了一种自蒸馏训练策略,将视频级知识转化为帧级功能,以实现更高的准确性和更高的效率,一些使用了transformer的重识别模型利用调整输入使图片输入更适配。Wang M.[14]设计了相机内和相机间对比学习组件,有效地学习相机内和跨相机的ID识别能力。Dou Z.[15]提出一种寻求身份的自监督表征学习(ISR)方法,ISR旨在学习属于同一身份的帧间图像的相似表示。Chen Z[16]提出一种新的双聚类协同教学(DCCT)框架,使用两组伪标签来训练两个网络。Feng J.[17]提出一种VI-ReID的形状擦除特征学习范式,通过正交分解将形状擦除的特征与形状相关的特征进行分离。Shimaa Saber[18]等人的方法表明,注意力引导的空间区域能更好地提升准确率。

一种分层跨模态的方法[19](Hierarchical Cross-Modality Disentanglement ,Hi-CMD),引入了一个识别的人图像生成网络和一个层次特征学习模块。

在保留人身份的同时，生成不同姿态和照明的交叉模态图像来学习解纠缠表示，以减少内部和交叉模态性的差异。动态双注意聚合[19]（DDAG：dynamic dual-attentive aggregation）方法，用IWPA（intra-modality weighted-part aggregation）挖掘每个模态内的上下文部分关系，来学习有区别的部分聚合特征的部分聚合，从而增强对噪声样本的鲁棒性。但通过特征映射到公共特征空间，来缩小模态差异并不总能提升性能。Yehansen Chen[21]等人设计了神经特征搜索（NFS：Neural Feature Search），结合了双级特征搜索空间和可微搜索策略自适应地过滤背景噪声。Lu Hu[22]等人提出了渐进模态共享变压器（PMT：Progressive Modality-shared Transformer），以减少模态间隙的负面影响，优化了全局特征却忽略了利用多尺度信息。Ye Mang[23]等人提出了一种通过联合优化模态特定指标和模态共享度量的层次交叉模态匹配模型（HCML：Hierarchical Cross-modality Metric Learning）。Wang Guan[24]等人提出的AlignGAN方法，利用GAN生成图像和利用像素对齐特征，从而实现性能的提升。

针对以上问题，本文提出了一种基于红外图像跨模态的行人重识别模型。

1. 提出一种从模型中取模型的方法。利用多级模型构建关系型数，将模型按类别的属性拆分成更小子任务，方便从模型中获得语义信息，拆分成的属性模型是模型的子任务。

2. 提出一种外源引导的注意力机制，更多地关注指定的区域，使模型可以从更多的数据中提取信息。

1 本文方法

深度学习是端到端的模型，这种现象不利于理解黑箱，由于目前没有非常通用且有效的分析出每层语义的方法，所以不利于理解和调整模型。以往的行人重识别模型中，对整体判断时输入的像素信息在局部有特征提取器黑箱中完成，要实现获得模块化的语义模型并让模型表达概念间依赖关系可理解模型内部，可以通过拆解任务解决查看黑箱的问题。本文提出一种将原始语义保留传递的模型，使端到端的深度模型变成端到子模型再到端的模型，模型首先映射成一个中间状态，这个状态描述了局部信息，然后根据每个局部信息的判断再进行整体判断。

为了提取一组标签，将一个重识别任务分成两部分学习，第一部分找到每个局部特征的位置和最佳区域，第二部分将第一部分的编码结果当作包含语义信息的词汇编码。受到transformer多头自注意力的启发，构造一个关键点引导的局部注意力机制对词汇间关系的学习能力，最后解码为一个描述全局的信息，已经训练好的其他有监督模型当作教师模型，实现对属性语义模型教学，用给模型提供信息的模型来生成标签的模型。由于知识蒸馏不但可以减少训练参数，同时也能保持甚至获得更好的效果（performance），所以，借鉴知识蒸馏的方法构造一个拆分模型的方法充分利用了已有模型。通过对局部特征的拆解，将局部特征也纳入到检索任务，从而实现对物体子属性的查询。例如，检索与目标人物相同的头、上肢和下肢的照片。

模型的训练分两个部分，第一个部分是构造局部掩码生成器（Local Mask Generator，LMG）引导拆分教师模型训练的局部特征，第二部分是融合局部特征检索器。

行人重识别模型应用于目标检测的下一步是在一帧视频里选出人的区域，利用已有的前置加工数据不会增加计算量，并且由于关键点检测模型和图像分割任务较为相似，所以前置的目标检测模型或姿态模型常常既可以做图像分割又可以检测人体关键点。例如，分割模型和姿态估计模型，姿态估计模型的重点和难点是定位人体关键点，可以使用姿态估计模型增强识别模型。本文选择用YOLO获得17个关键点特征。

1.1 拆分模型

减少参数量增加局部输出，将行人重识别拆解为局部肢体重识别。由于局部图片较少，所以我们使用已有的模型生成需要的非人工标签的数据。局部掩码生成器是一个向模型学习的模型，是利用神经网络权重实现的。因为YOLO输出的人体关键点的范围比较小，不能直接适用于行人重识别，所以设计一个结构和ResNet类似的局部掩码生成器置于resnet和YOLO之间（如图2），使得人体关键点更适配重识别模型，提取特征时，使用两个独立的多尺度卷积进行提取。

输入的图像分别送入已训练过的行人重识别模型当做教师模型，同时将图像送入人体关键点检测模型获得关键点信息和分割信息。选择用COCO与训练的模型可获得17个人体部位的位置信息，编号从0到16对应的部位分别是：鼻子、左眼、右眼、

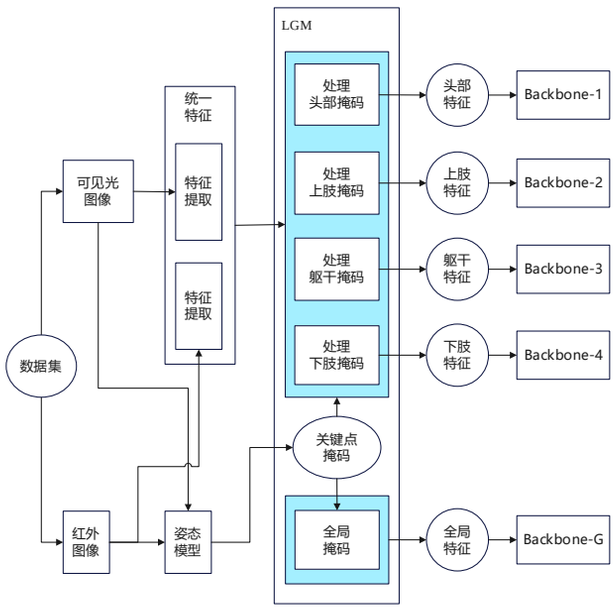

图2　LMG结构图

Fig. 2　LMG structure

左耳、右耳、左肩、右肩、左肘、右肘、左腕、右手腕、左髋关节、右髋关节、左膝、右膝、左脚踝、右脚踝。由于图像分辨率不高会出现识别不准确,将对应的通道按(0,1,2,3,4)、(5,6,7,8,9,10)、(5,6,11,12)、(11,12,13,14,15,16)重新分组,分别是头、上肢、上身和下肢,使用四个独立的卷积神经网络ResNet50记作 backbone-N (N∈{1,2,3,4})分别学习这4个局部特征。

局部掩码生成器用来修改输入数据,由于局部掩码生成器目的是让模型注意到人体关键点及临近像素的范围较小,所以局部掩码生成器通过两次池化后再使用双线性插值上采样(Bilinear interpolation)实现扩大区域,不关注部分被掩码置0,最后特征融合时使用分组1×1卷积齐通道数并将上采样调整到和LMG-layer1一致。

让输入经过统一特征模块后,再计算一次BN层有助于提取到更好的反例信息防止过拟合,按通道拼接特征输入到Backbone、Backbone-N和Backbone-G(如图3),Resnet50的特征提取由conv1和4个layer组成。而Backbone-N和Backbone-G的只有4个layer组成,在LMG中使用一个conv1共享低层特征,并且随着模型传导到高层空间,信息会被隐藏不利于调整控制空间特征,所以,将获得的局部掩码和全局掩码乘以每一个backbone N的layer1输入信息。

通过调节温度、调整知识蒸馏类别的输出值,可防止训练后学生模型的特征分布过于集中。所以,计算获得ID前的教师模型、拆解的特征重新获得各个局部特征权重的学生模型,只学部分的学生模型,都需要使用如下公式(1):

$$q_i = \frac{\exp(x_i/T)}{\sum_j \exp(x_j/T)} \quad . \quad (1)$$

重新调节权重,可以使注意力均摊到不同部位,用姿态模型进行预处理,为了可复用模型,激活原始模型后利用伪标签进行自监督学习。

1.1.1　损失函数

LMGNet模型输出的局部特征用1×1卷积调整通道后使用三元损失函数学习教师模型的特征,如公式(2):

$$\mathcal{L}_{tri} = \frac{1}{N}\sum_{i=1}^{N}\left[\left\| f_i(\boldsymbol{M}(\boldsymbol{x})) - f_i^{+}(\boldsymbol{M}(\boldsymbol{x}))\right\|\right.$$

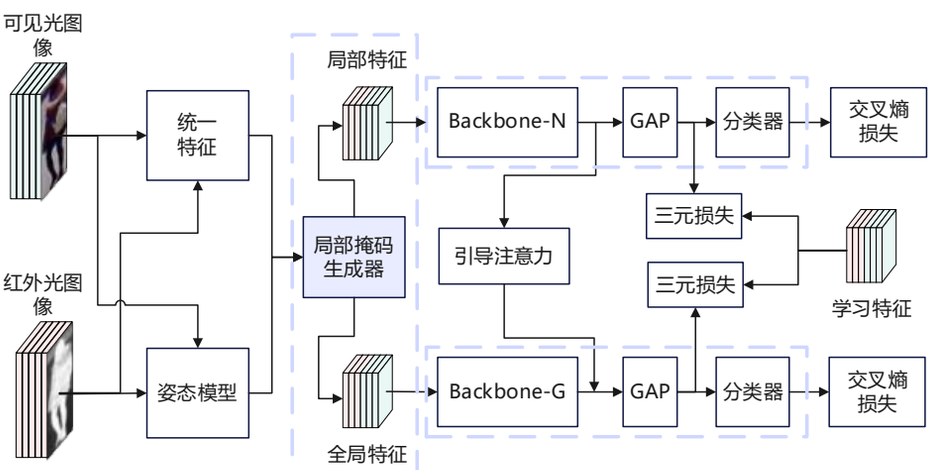

图3　用LMG拆分全局特征的方法

Fig. 3　Splitting of global features by LMG

$$-\left\| f_i(\boldsymbol{M}(\boldsymbol{x})) - f_i^-(\boldsymbol{M}(\boldsymbol{x})) \right\| + \boldsymbol{m} \Big] \quad (2)$$

$\mathcal{L}_{tri}$表示三元损失函数，f_i是预测值，f_i^+和f_i^-分别是正样本和负样本。正样本和负样本之间的距离m设置为0.3，‖*‖是欧氏距离，M表示LMG输出掩码的函数。

交叉熵损失通常用于多分类如公式(3)：

$$\mathcal{L}_{id} = \frac{1}{N}\sum_i \sum_{c=1}^{C} f_{ic}(\boldsymbol{M}(\boldsymbol{x}))\log(\boldsymbol{p}_{ic}) \quad (3)$$

使用交叉熵损失函数分类不同的id，Backbone-N和Backbone-G的输入乘以掩码后再正向传导。C是分类的总数，$f_{ic}(M(x))$的输出是类别标签，p是预测的类别，i是样本编号，c表示类别编号即行人id号。

掩码生成器使用深度可分离卷积将输入拆分为4组的特征，LMG-conv1输出的4个通道特征与从YOLO获得人体关键点位置的特征，使用均方差函数计算出loss值如公式(4)：

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \quad . \quad (4)$$

张量尺寸为(-1,17,128,64)，y_i是特征上的所有点，利用类似特征金字塔的结构让局部掩码生成器把backbone的特征当做数据输入。

整个损失函数可以表示为公式(5)：

$$\mathcal{L} = \sum_{i=0}^{4}(\lambda_i \mathcal{L}_{id_i} + \lambda_2 \mathcal{L}_{tri_i}) \quad , \quad (5)$$

其中i是每个Backbone-N的编号，当i=0时是Backbone-G。

1.2 全局模型

1.2.1 引导注意力

现有的多数模型都不是对父类和子类进行学习，一些利用在模型中区分子属性的多见于横向分割的重识别模型。试图解决局部特征的方法有很多，比如注意力机制如图4。但是注意力机制的引导完全是从训练实现的，而特征的定向分割也是基于确定数据输入的，这导致不能调整灵活的调整模型。全局池化，导致模型过于关注特定领域的信息，导致模型在源域上过度拟合。

用Other-Attention(OA)融合多局部特征为注意力，引导全局特征提取，如公式(6)：

$$\mathbf{OA}(\mathbf{L},\mathbf{W},\mathbf{G}) = \boldsymbol{softmax}(\boldsymbol{Conv}\left(\frac{\boldsymbol{B}_N(\mathbf{L})\mathbf{M}}{\sqrt{\boldsymbol{d}}}\right))\boldsymbol{B}_c(\boldsymbol{G}) \quad (6)$$

为了在空间上融合局部特征，先将4个不同部位输出的X(batch-size, C,H,W)拼接为L，L表示一组局部特征(Local features)，局部特征和全局特征

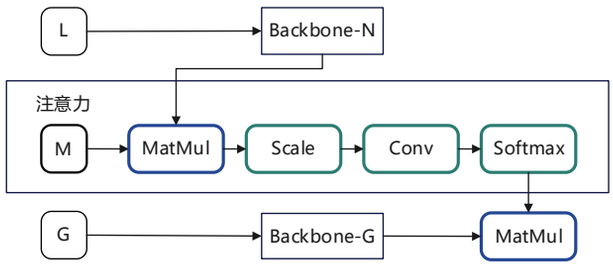

图4 引导局部注意力

Fig. 4 Guiding localized attention

在语义上是组合关系，4个局部注意力分别训练，这里用表示为一个(batch-size,4×C,H×W)尺寸的张量。所以，用M(Matrix)调整每个局部特征的位置。M是尺寸为($H\times W$,$H\times W$)，d=$H\times W$，LM输出的结果除以$\sqrt{d}$后，用Conv卷积计算实现按比例合并为全局特征。经过softmax函数后矩阵乘法乘以G，G表示全局特征(Global Features)(batch-size,C,$H\times W$)。从源数据提取的特征L和G，利用注意力融合为一个全局特征，从而使Backbone-G在排除环境干扰的基础上更关注到局部重要位置的特征。使用了局部注意力和整体注意力，都是使用了相同的输入数据提取特征，使用了相同的教师模型指导。

2 实验

2.1 数据集

为了分析本文方法的有效性，在两个经常被人使用的红外行人重识别数据集SYSU-MM01和RegDB上进行实验，来测试和评估本文提出的LMGNet的效果，并使用LMGNet与同类型的最先进的其他方法进行比较。

SYSU-MM01[25]数据集包含491个行人分别在室内和室外图像，由4个可见光相机拍摄的287 628幅可见光行人可见光图像以及2个红外相机拍摄的15 792幅红外行人图像。训练集包含395个身份的22 258张可见光图像和11 909张红外图像，测试集则包含96个身份的301张可见光图像和3 803张红外图像。SYSU-MM01数据集提供了全搜索(all-search)和室内搜索(indoor-search)两种模式。把相机编号为1、2、4、5的可见光图像当作gallery集，相机编号3、6的红外图像当作query集，则称为全搜索模式。把相机编号为1、2的可见光图像当作gallery集，相机编号3、6的红外图像当作query集，则称为室内搜索模式。测试时，根据每个行人包含图像的数量也可分为两种模式，gallery集中每

个行人包含多张图像则是 mutil-shot 模式;gallery 集中每个行人仅包含一张图像则是 sing-shot 模式。

RegDB[25]是由一个可见光相机和一个远红外相机采集的 412 个行人的图像的数据集,每个行人各有 10 张可见光图像和 10 张红外图像,有可见到红外(visible-to-infrared)和红外到可见 (infrared-to-visible)两种检索模式。

2.2　实验设置

实验的硬件使用了 GeForce RTX 3090TI、CUDA11, 框架使用 PyTorch1.7,学生模型的 Backbone 使用了 ImageNet 的预训练权重 FastReID 的 ResNet50、torchreid(框架是 OSNet 模型的文章的一部分)。开始时学习速率 0.01,优化器使用 Adam 训练 120 个 epoch,每 20 个 epoch 衰减 10,输入图像的大小调整为 256×128。在使用掩码修改演示图像后,在非关注去随机添加随机遮挡和噪声,使用的数据增强包括随机翻转、随机裁剪。

2.3　评估指标

本文采用行人重识别常见的评级标准对模型性能的评估,分别是累积匹配特征 (cumulative matching characteristic, CMC) 和平均精度(mean aver-age precision, mAP) 作为评估指标,CMC 指标报告了在给定一个文本描述作为查询时,在对可能性排序的前 k 个中找到至少一个匹配的人图像的概率, 在后续表中用 Rank-k 表示。

2.4　对比实验结果

为验证 LMG 模型的先进性 ,在 SYSU-MM01 和 RegDB 两个公共数据集上与近几年研究出的红外行人重识别模型 Zero-Pad、Hi-CMD、DDAG、AGW、NFS、PMT、HCML 和 AlignGAN 进行了对比。

从在 SYSU-MM01 数据集上测试的 all-search 和 indoor-search 两种模式的实验结果(表 1)可以看出,有较好的效果。在 all-search 模式下,Rank-1 和 mAP 分别为 71.42% 和 65.89%,相比其他模型中最高的 Rank-1 和 mAP,分别提升了 3.89% 和 0.91 %。;在 indoor-search 搜索模式下, Rank-1 和 mAP 分别为 78.88% 和 81.65%,相比其他模型中最高的 Rank-1 和 mAP,分别提升了 7.22 % 和 11.86%。

从在 RegDB 数据集上测试的可检查红外和红外查可见两种模式的实验结果(表 2)可以看出,有一定的提升。在可见查红外模式下,Rank-1 和 mAP 分别为 89.52% 和 80.89%,相比其他模型中最高的 Rank-1 和 mAP,分别提升了 4.69 % 和 4.34 %;在红外查可见搜索模式下,Rank-1 和 mAP 分别为 88.11% 和 82.57%,相比其他模型中最高的 Rank-1 和 mAP,分别提升了 3.95% 和 7.44%

2.5　消融实验

经过伪标签重新训练局部特征获得的模型更专注于局部信息,本文的模型可以调节局部特征是否开启。

为了验证我们提出的模型中组件的有效性,在 SYSU-MM01 数据集上进行了消融实验。通过取消的部分组件验证,整体设置不变,采用 DEEN [27](Diverse Embedding Expansion Network)模型预训练的 ResNet50 作为我们的基线模型(如表 3),第一种情况 baseline + Backbone-G 与基线相比,提升得并不多。

局部注意力模块,将掩模重构模块加入 baseline 进行实验,Rank-1 和 mAP 分别为 70.19% 和 64.99%,较 baseline 分别提高了 5.49 和 2.99 个百分点,说明局部注意力模块能够很好地消除无关特征

表 1　在 SYSU-MM01 数据集上测试结果对比

Table 1　Comparison of test results on SYSU-MM01 dataset

方法名	全查模式				室外模式			
	rank-1/%	rank-10/%	rank-20/%	mAP/%	rank-1/%	rank-10/%	rank-20/%	mAP/%
Zero-Pad [25]	14.8	54.12	71.33	15.59	20.58	68.38	85.79	26.92
Hi-CMD [19]	34.94	77.58	-	35.94	-	-	-	-
DDAG [20]	54.75	90.36	95.81	53.02	61.02	94.06	98.41	67.98
AGW [26]	47.5	84.39	92.14	47.65	54.17	91.14	95.98	62.97
NFS [21]	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
PMT [22]	67.53	95.36	98.64	64.98	71.66	-	-	69.5
HCML [23]	14.32	53.16	69.17	16.16	24.52	73.25	86.73	30.08
AlignGAN [24]	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.30
本文方法	71.42	95.61	98.86	65.89	78.88	98.31	99.1	81.65

表2 在 RegDB 数据集上测试结果对比

Table 2 Comparison of test results on RegDB dataset

方法名	可见查红外				红外查可见			
	rank-1/%	rank-10/%	rank-20/%	mAP/%	rank-1/%	rank-10/%	rank-20/%	mAP/%
Zero-Pad	17.75	34.21	44.35	18.9	-	-	44.25	17.82
Hi-CMD	70.93	86.39	-	66.04	-	-	-	-
DDAG	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.8
AGW	70.05	87.28	92.04	66.37	68.83	83.69	88.35	64.45
NFS	80.54	91.96	95.07	72.1	77.95	16.63	34.68	69.79
PMT	84.83	-	-	76.55	84.16	-	-	75.13
HCML	24.44	47.53	56.78	20.08	21.7	45.02	55.58	22.24
AlignGAN	57.9	-	-	53.6	56.3	-	-	53.4
本文方法	89.52	92.31	95.34	80.89	88.11	92.23	97.67	82.57

表3 有无LMG的性能比较

Table 3 Performance comparison with and without LMG

设置			SYSU-MM01	
基线	LMG	全局	mAP/%	Rank-1/%
✓			62.0	64.7
✓	✓		64.99	70.19
✓		✓	64.45	67.28
✓	✓	✓	65.89	71.42

的影响，使模型对人体区域具有更准确的识别能力。

全局模块直连后，Rank-1和mAP分别为67.28%和62.45%，较baseline分别提高了2.45%和2.58%，说明全局模块是有效的且对模型有着积极的影响。

总体上看，通过局部注意力模块引导和全局模块的融合，最终使模型的Rank-1和mAP分别提升至71.42%和65.89%，Rank-1和mAP较baseline分别提高了6.72%和3.89%。

2.6 可视化

在本小节中，将分析方法中每个组成部分的有效性。为了证明局部特征提取的有效性，输入SYSU-MM01数据集的图片后，分别输出了4个局部特征的热度图（如图5），图5(a)打印了原人体17个关键点的红外图，将其他通道置零，局部掩码生成器修改模型最终的注意力；图5(b)是头部热图；图5(c)是上肢热图；图5(d)是上身热图；图5(e)是下肢热图。从图中可看出，本文方法可以利用关键点模型实现由人类意愿决定重识别模型关注的那部分信息。

3 结论

针对红外光场景下的行人重识别存在的跨模态的差异和环境特征干扰等问题，提出了一种简单而有效的新方法LMGNet。为了实现有目的且有效的引导注意力，此方法通过学习另一个模型的输出，对不同尺度的全局范围结构信息进行建模、获得掩码，以提高模型推理的准确性。由于生成了具有语义信息特征，所以利用这些信息可以通过关系型数据库实现对特定目标的检索，而不仅仅是完全使用神经网络模型。通过指定模型的通道来学习指定的身体部位，增加了模型的可解释性，并且只比较存在相同部位的通道的距离，以提高精度。与同类的先进的方法相比较，在数据集SYSU-MM01上全查模式的mAP和Rank-1，分别提升了3.89%和0.91%。拆分局部后可以有效对齐局部特征，从

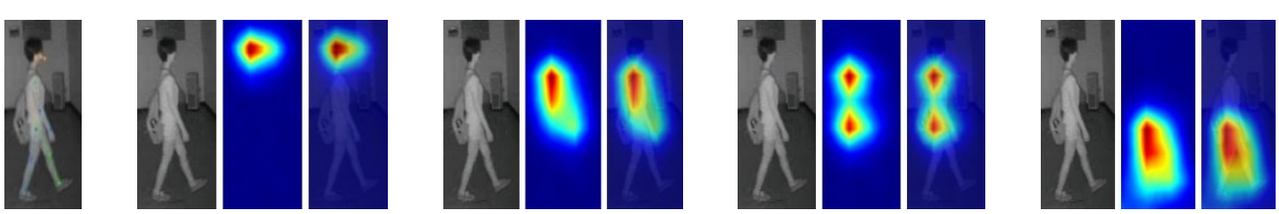

图5 模型关注的局部对比图

Fig. 5 Comparison of the local areas of interest of the model

而正确引导模型的注意力,实验表明了本文方法的有效性。

References

[1] Gu X, Chang H, Ma B, *et al*. Clothes-Changing Person Re-identification with RGB Modality Only[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1060-1069.

[2] Park H, Ham B. Relation network for person re-identification[C]. Proceedings of the AAAI conference on artificial intelligence. 2020, **34**(07): 11839-11847.

[3] Zhang Z, Lan C, Zeng W, *et al*. Relation-Aware Global Attention for Person Re-Identification[C]. Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2020: 3186-3195.

[4] Miao J, Wu Y, Yang Y. Identifying Visible Parts via Pose Estimation for Occluded Person Re-Identification [J]. *IEEE Transactions on Neural Networks and Learning Systems*. 2022, **33**(9): 4624-4634.

[5] Sun Y, 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). 2018: 480-496.

[6] Zhu K, Guo H, Zhang S, *et al*. AAformer: Auto-Aligned Transformer for Person Re-Identification[J]. *IEEE Transactions on Neural Networks and Learning Systems*. 2023.

[7] Ci Y, Wang Y, Chen M, *et al*. UniHCP: A Unified Model for Human-Centric Perceptions [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 17840-17852.

[8] Che J, Zhang Y, Yang Q, *et al*. Research on person re-identification based on posture guidance and feature alignment[J]. *Multimedia Systems*. 2023, **29**(2): 763-770.

[9] Gu J, Wang K, Luo H, *et al*. MSINet: Twins Contrastive Search of Multi-Scale Interaction for Object ReID [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19243-19253.

[10] Pan H, Chen Y, He Z. Multi-granularity graph pooling for video-based person re-identification [J]. *Neural Networks*. 2023, **160**: 22-33.

[11] Zhang G, Zhang H, Lin W, *et al*. Camera Contrast Learning for Unsupervised Person Re-Identification [J]. *IEEE Transactions on Circuits and Systems for Video Technology*. 2023, **33**(8): 4096-4107.

[12] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network [J]. *arXiv preprint* arXiv: 1503.02531. 2015.

[13] Liu X, Yu C, Zhang P, *et al*. Deeply Coupled Convolution - Transformer With Spatial - Temporal Complementary Learning for Video-Based Person Re-Identification [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2023: 1-11.

[14] Wang M, Lai B, Huang J, *et al*. Camera-aware Proxies for Unsupervised Person Re-Identification[C]. Proceedings of the AAAI conference on artificial intelligence. 2021, **35**(4): 2764-2772.

[15] Dou Z, Wang Z, Li Y, *et al*. Identity-Seeking Self-Supervised Representation Learning for Generalizable Person Re-identification[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2023: 15847-15858.

[16] Chen Z, Cui Z, Zhang C, *et al*. Dual Clustering Co-Teaching With Consistent Sample Mining for Unsupervised Person Re-Identification[J]. *IEEE Transactions on Circuits and Systems for Video Technology*. 2023, **33**(10): 5908-5920.

[17] Feng J, Wu A, Zhen W. Shape-Erased Feature Learning for Visible-Infrared Person Re-Identification [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 22752-22761.

[18] Saber S, Meshoul S, Amin K, *et al*. A Multi-Attention Approach for Person Re-Identification Using Deep Learning [J]. *Sensors*. 2023, **23**(7): 3678.

[19] Choi S, Lee S, Kim Y, *et al*. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification [C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10257-10266.

[20] Ye M, Shen J, J. Crandall D, *et al*. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-identification[C]. Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23 - 28, 2020, Proceedings, Part XVII 16. Springer International Publishing, 2020: 229-247.

[21] Chen Y, Wan L, Li Z, *et al*. Neural Feature Search for RGB-Infrared Person Re-Identification[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 587-597.

[22] Lu H, Zou X, Zhang P. Learning Progressive Modality-Shared Transformers for Effective Visible-Infrared Person Re-identification [C]. Proceedings of the AAAI conference on artificial intelligence. 2023, **37**(2): 1835-1843.

[23] Ye M, Lan X, Li J, *et al*. Hierarchical discriminative learning for visible thermal person re-identification [C]. Proceedings of the AAAI conference on artificial intelligence. 2018, **32**(1).

[24] Wang G, Zhang T, Cheng J, *et al*. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 3623-3632.

[25] Wu A, Zheng W, Yu H, *et al*. RGB-Infrared Cross-Modality Person Re-identification [C]. Proceedings of the IEEE international conference on computer vision. 2017: 5380-5389.

[26] Ye M, Shen J, Lin G, *et al*. Deep Learning for Person Re-Identification: A Survey and Outlook [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2022, **44**(6): 2872-2893.

[27] Zhang Y, Wang H. Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification [J]. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023: 2153-2162.