DOI:10.13878/j.cnki.jnuist.20240330001

黄驰涵[1]　沈肖波[2]

基于融合注意力和特征增强的跨模态行人重识别

摘要

跨模态行人重识别是一项具有挑战性的任务,目的是在可见光和红外模式之间匹配行人图像,以便在犯罪调查和智能视频监控应用中发挥重要作用.为了解决跨模态行人重识别任务中对细粒度特征提取能力不强的问题,本文提出一种基于融合注意力和特征增强的行人重识别模型.首先,利用自动数据增强技术缓解不同摄像机的视角、尺度差异,并基于交叉注意力多尺度 Vision Transformer,通过处理多尺度特征生成具有更强区分性的特征表示;接着,提出通道注意力和空间注意力机制,在融合可见光和红外图像特征时学习对区分特征重要的信息;最后,设计损失函数,采用基于自适应权重的难三元组损失,增强了每个样本之间的相关性,提高了可见光和红外图像对不同行人的识别能力.在 SYSU-MM01 和 RegDB 数据集上进行大量实验,结果表明,本文提出方法的 mAP 分别达到了 68.05% 和 85.19%,相较之前的工作性能有所提升,且通过消融实验和对比分析验证了本文模型的先进性和有效性.

关键词

行人重识别;跨模态;交叉注意力;特征提取;多尺度

中图分类号 TP391.41

文献标志码 A

收稿日期 2024-03-30

资助项目 国家自然科学基金(62176126);江苏省自然科学基金优秀青年基金(BK20230095)

作者简介

黄驰涵,男,主要研究方向为深度学习、对抗攻击.huangchihan@njust.edu.cn

沈肖波(通信作者),男,博士,教授,主要研究方向为模式识别、机器学习、计算机视觉.njust.shenxiaobo@gmail.com

1 南京理工大学 设计艺术与传媒学院,南京,210094

2 南京理工大学 计算机科学与工程学院,南京,210094

0 引言

行人重识别(Re-Identification,Re-ID)是智能监控系统中最重要的部分之一,它可以在不同的摄像头视图之间识别行人.行人重识别在许多视频任务中都具有实际应用,包括法证搜索[1]、多摄像头跟踪[2]、门禁控制[3]和体育分析[4].它还被应用于服务机器人和人机交互,老年人监控和协助执行个性化任务等[5].然而,由于观察角度、照明强度、姿势、遮挡和背景杂乱等变化,行人重识别在计算机视觉领域中仍然是一个具有挑战性的任务.

以往大多数的 Re-ID 任务都是在白天或是可见光(RGB)充足的情况下进行单模态的识别,但在夜间,监控摄像头很难利用可见光谱的摄像头来进行识别[6].故目前的监控摄像头在夜间能转换为红外(IR)模式,而 IR 图像存在缺少颜色信息的重要问题,这就需要 Re-ID 能够适用于跨模态的行人检索.

为了解决跨模态行人重识别问题,目前已有多种方法,主要包括两种思路:第一种是利用网络捕获两种模态下的行人特征来进行行人图像匹配[7];第二种是对图像的模态进行转换或生成新的模态来进行行人重识别.对于第一种思路,Yuan 等[8]引入了并行的多流分类器,通过使每个流中的分类器关注不同的特征维数,以确保特征提取器的类内一致性.但是其并未考虑对原始图像进行数据增强,处理数据时也较为困难.Chen 等[9]引入一个新的特征搜索空间,并提出一种自动选择通道和空间维度中身份信息的特征选择方法.但是其直接将特征映射到公共特征空间来缩小模态差异,使得一些重要的行人判别特征丢失,影响模型的性能.对于第二种思路, Choi 等[10]提出一种层次跨模态解耦(Hi-CMD)模型,改变光照属性和行人的姿态使得编码器能够提取到更具有判别性的特征.但是训练一个好的生成器和判别器需要花费大量的计算资源,且在利用生成对抗网络的同时势必引入一些噪声,影响模型的稳定性. Liu 等[11]采用对齐灰度模态(AGM)将可见红外双模学习重新表述为灰度-灰度单模学习问题,在图像空间中显著减少了模态差异.Xia 等[12]在真实图像上训练图像模态转换(IMT)网络,并生成目标模态样本,以扩大训练数据集的大小并增加其多样性,同时将源图像和模态传递的图像组合训练 Re-ID-CNN 模型,以提高跨模态检索性能.但是这些模型网络中需要进行图像风格转换,不可避免地会增加噪声干扰,影响模型的稳定性,使得

生成的图像并不可靠,且这些模型高度依赖训练样本,很难应用于大规模监测场景.

针对以往研究中对细粒度信息提取能力不强的问题,本文提出了基于交叉注意力多尺度残差 Vision Transformer (ViT)的特征提取器,它能够处理多尺度特征,生成区分性较强的特征表示.提取到可见光和红外图像的特征后,采用通道和空间注意力机制来融合不同模态的特征.最后,利用平滑的标签损失和自适应权重的难三元组损失对训练过程联合监督.

本文的贡献可以概括如下:1)提出一种基于交叉注意力多尺度残差 ViT 框架,专注于提取判别性和鲁棒性更强的特征;2)在融合可见光和红外图像特征时,在网络中使用通道和空间注意力来学习对特征区分重要的信息;3)对难三元组损失进行自适应权重的改进,增强了每个样本之间的相关性;4)在 SYSU-MM01 和 RegDB 数据集上进行评估实验,结果表明本文提出的方法具有良好的性能.

1 模型设计

行人重识别是计算机视觉领域中的一个关键任务,旨在从不同的摄像头视角中识别并匹配同一行人的图像.该任务通常涉及两个主要组成部分:查询(query)和图库(gallery).查询是指用户提供的用于检索的行人图像,而图库是已存储的、需要与查询图像进行比较的行人图像集合.整个网络的输入是从不同摄像头捕获的行人图像,输出是与查询图像相匹配的行人图像的排名列表.

本节将介绍如图 1 所示的基于融合注意力和特征增强的行人重识别模型框架.首先,设计了交叉注意力多尺度 ViT,并对 IR 和 RGB 图像以及它们的灰度图分别使用 CrossViT 提取特征,形成不同的特征集.接着,设计了通道和空间注意力模块,利用通道注意力(Channel Attention, CA)和空间注意力(Spatial Attention,SA)来突出输入的 RGB-IR 图像对的有意义的信息.最后,对标签损失和难三元组损失进行改进,使用标签平滑和自适应权重实现联合监督.以下将对各个模块进行详细介绍.

1.1 自动数据增强

自动数据增强(AutoAugment)[13]于 2018 年被提出,通过强化学习来搜索数据增强策略.但是其搜索空间过大,使得实验时间过长、计算成本过高,在实际中难以直接应用.故提出了 Trivial Augment (TA)[14],它不需要采用 AutoAugment 类似的代理任务,直接采用简单的网格搜索即可获得更好的效果.

本文使用 Trivial Augment 的流程如下:在给定一组图像和一组数据增强操作 A 的情况下,每次随机选择一张图像 I,并在 A 中随机选择一个数据增强操作以及它的增强幅度.在这里,数据增强操作一般是较为基础的图像处理方法,例如裁剪、翻转、旋转、锐化、对比度增强等.这样的数据增强方法能够在不引入额外噪声的同时使模型获得更多样化、更丰富

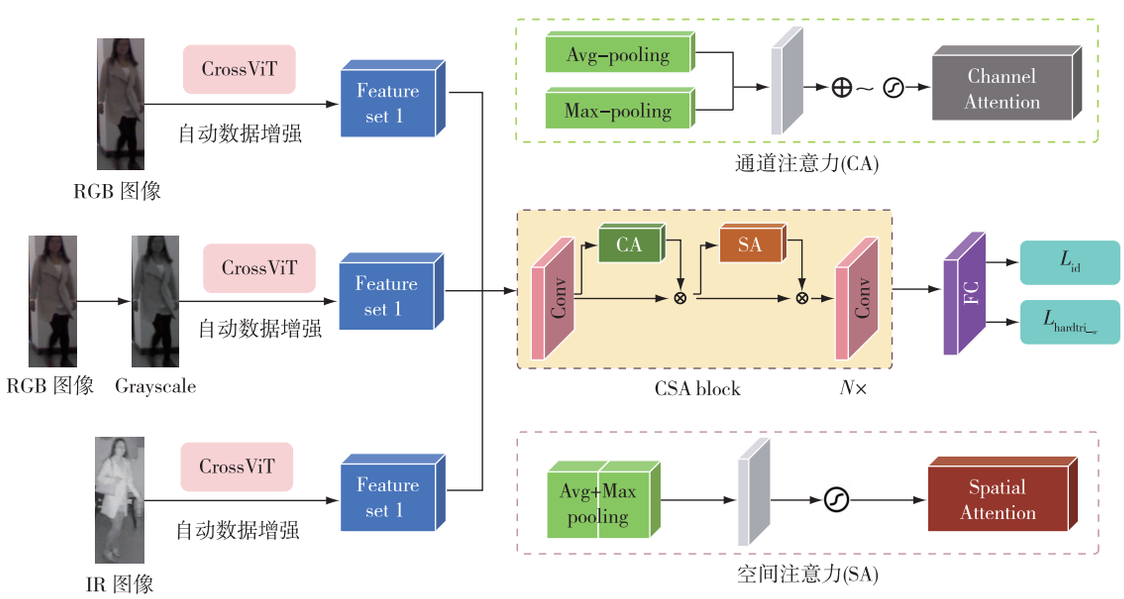

图 1 基于融合注意力和特征增强的行人重识别模型结构

Fig. 1 Model structure for person Re-ID based on fused attention and feature enhancement

的数据,从而提升模型的鲁棒性和泛化能力.

1.2 CrossViT

跨注意力机制确保模型根据不同模态的信息关注到图像中的重要区域,多尺度特征在图像内捕获不同空间分辨率或尺度的信息.图像中的行人可能会因视角或尺度差异而改变大小,而不同尺度的特征恰能捕获细粒度和粗粒度的结构.

多尺度特征有助于处理图像内行人大小和形状的变化.本文在标准的 Vision Transformer (ViT)[15] 中引入了多尺度残差连接,以分析不同空间分辨率下的特征,处理由于不同成像条件而引起的外观变化.CrossViT 有效地整合了可见光和红外模态的信息,捕获了空间关系,且考虑了全局和局部语境.它能够处理多尺度特征,生成区分性较强的特征表示.

1.2.1 ViT

ViT 是一种基于 Transformer 的神经网络模型,最初被设计用于解决图像分类任务.它将图像视为一系列补丁,并利用 Transformer 架构对这些补丁进行处理.由于 ViT 采用了自注意力机制,因此 Transformer 编码器需要包含位置信息.为此,ViT 将位置嵌入到每个含有 CLS 的令牌中.该网络的编码器由一系列块组成,每个块都包含多头自注意力和一个前馈层网络(FFN).FFN 由两层多层感知器组成,其中隐藏层包含扩展比率,并在后续层中采用 GELU 激活函数.每个块都采用层归一化进行处理,以生成残差特征.记 ViT 的输入为 z_0,第 l 个块的公式如下:

$$z_0 = [x_{\text{CLS}};x_{\text{patch}}] + E_{\text{pos}}, \tag{1}$$

$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}, \tag{2}$$

$$z_l = \text{MLP}(\text{LN}(z'_l)) + z'_l, \tag{3}$$

$$y = \text{LN}(z_l^0). \tag{4}$$

其中:x_{CLS} 和 x_{patch} 分别表示类别和补丁令牌;E_{pos} 表示位置嵌入令牌;MSA 为多头注意力机制;MLP 为多层感知机;LN 为层归一化.由于 x_{CLS} 令牌综合了所有补丁令牌的信息,而不同补丁大小之间的差异将会影响模型的性能,因此,本文采用具有交叉注意力的多尺度 ViT 来应对这种复杂性.

1.2.2 交叉注意力多尺度特征

图 2 展示了使用交叉注意力 ViT 学习多尺度特征的框架.该模型由 L 个多尺度 Transformer 编码器组成,每个编码器中都包括一个粗粒度分支 C 和一个细粒度分支 F.粗粒度分支负责提取较为粗糙的特征,具有更多的编码器和嵌入维度,而细粒度分支则专注于提取更加细致的特征,具有较少的编码器和嵌入维度.在这两个分支的每个令牌中都需要添加位置嵌入.在多次应用粗粒度分支和细粒度分支之后,通过多尺度 Transformer 编码器进行处理,并最终利用 CLS 令牌进行分类.该网络的关键在于通过交叉注意力机制进行特征处理,从而实现多尺度特征的学习.在交叉注意力机制中,取一个分支的 CLS 令牌,在另一个分支中取补丁令牌并将它们融合在一起.

CLS 令牌能够传递不同分支的补丁令牌信息,并在融合后返回到原始分支.因此,补丁令牌的信息已经被 CLS 令牌学习,CLS 令牌与不同分支的补丁令牌进行交互,以获取多尺度特征的信息.融合后,CLS 令牌将获得用于后续编码器的补丁令牌,从而提升了信息交换传递的性能.在编码器上融合 CLS

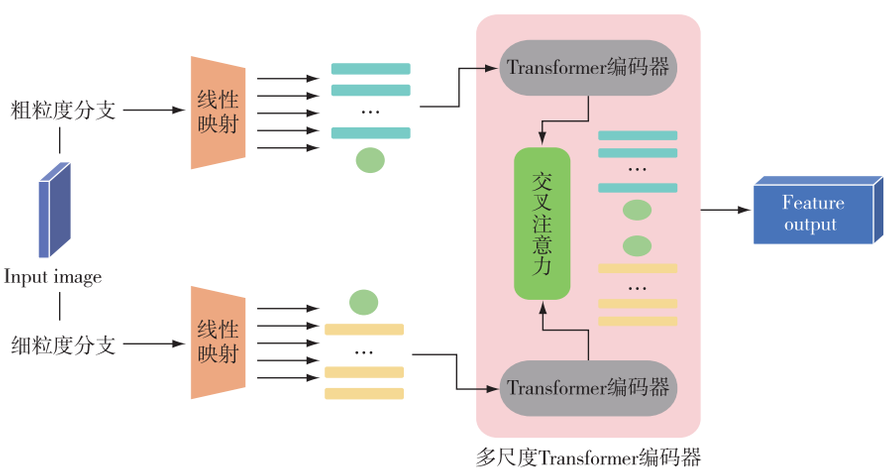

图 2 交叉注意力 ViT 网络结构

Fig. 2 Cross attention ViT network structure

和补丁令牌的公式如下:

$$z^{C}=[P^{C}(I_{CLS}^{C};I_{patch}^{F})]. \tag{5}$$

其中:$P^{C}(\cdot)$ 为投影函数,用于对齐维度;I_{CLS}^{C} 和 I_{patch}^{F} 分别表示粗粒度 CLS 令牌和细粒度补丁令牌,上标 C 和 F 分别表示粗粒度和细粒度分支.接着在 z^{C} 和 I_{CLS}^{C} 之间使用交叉注意力,其表达式为

$$\mathrm{Cross}A(z^{C})=\mathrm{softmax}\left(\frac{qk^{T}}{\sqrt{\frac{e}{h}}}\right)\cdot(z^{C}V_{v}). \tag{6}$$

其中:$q=I_{CLS}^{C}V_{q}$;$k=z^{C}V_{k}$;V_{q}、V_{k} 和 V_{v} 为可学习的参数;e 和 h 分别表示嵌入令牌的大小以及注意力头的总数.最后,CrossViT 的输出结果通过层归一化与残差特征集融合,其表达式为

$$z_{CLS}^{C}=P^{C}(I_{CLS}^{C})+\mathrm{MultiCross}A(\mathrm{LN}[z^{C}]), \tag{7}$$

$$Z^{C}=[\mathrm{BP}^{C}(z_{CLS}^{C};I_{patch}^{F})]. \tag{8}$$

其中:$\mathrm{BP}^{C}(\cdot)$ 表示用于对齐维度的反投影函数.该模型能够检索特征并根据获取的 RGB 图像来精准识别行人.

1.3 通道及空间注意力模块

1.3.1 通道注意力

在通过 CrossViT 获得行人图像的特征图 M 后,本文设计了一个通道注意力模块,旨在生成一个加权图,以跨特征图进行加权计算,使模型更加关注更重要的通道.嵌入通道注意模块的网络可以表示为

$$M'=A_{c}(M)\otimes M. \tag{9}$$

其中:⊗表示加权算子;$A_{c}(M)$ 表示通道注意力模块;M' 表示通道注意特征.

通道注意力模块通过选择 RGB-IR 特征图中更具意义的通道来优化特征的通道关系.一种常见的特征图聚合方法是利用平均池化学习输入对象的范围.为了更好地挑选出具有区分性特征并保留更多纹理信息,本文进一步引入最大池化[16]操作.平均池化和最大池化均被前向传播到卷积块以生成通道注意力图,最终通过逐元素求和来合并输出特征,其表达式如下:

$$A_{c}(M)=M_{Avg}+M_{Max}=\sigma(C_{2}(\mathrm{ReLU}(C_{1}\mathrm{Avg}(M)))+C_{2}(\mathrm{ReLU}(C_{1}\mathrm{Max}(M)))). \tag{10}$$

其中:M 为不同层后图像的特征;σ 为 sigmoid 函数;C_{1} 和 C_{2} 为两个不同的卷积层.

1.3.2 空间注意力

为了捕捉特征的空间关系,本文进一步采用空间注意力模块来强调特征信息,作为通道注意力的补充信息.嵌入空间注意力模块的网络可以表示为

$$M''=A_{s}(M')\otimes M'. \tag{11}$$

其中:M'' 表示最终的注意力特征;$A_{s}(M')$ 表示空间注意力模块.

空间注意力模块的设计与通道注意力模块有所不同.在空间注意力模块中,首先对输入特征 E 进行平均池化和最大池化,得到两个全局空间描述符.具体来说,平均池化和最大池化都将输入特征图 E 的每个通道从原始的 $H\times W$ 维度降维到 1×1,其中,H 和 W 分别表示特征图的高度和宽度.这两个 $1\times 1\times C$ 维的特征图分别对应平均池化和最大池化的结果,这里的 C 为通道数.得到的平均池化和最大池化结果被拼接在一起,形成一个 $1\times 1\times 2C$ 的特征向量,它随后被送入一个卷积层,减少参数数量并生成最终的空间注意力图.空间注意力的表达式为

$$A_{s}(M)=M_{Avg,Max}=\sigma(\mathrm{Conv}(\mathrm{Avg}(M)+\mathrm{Max}(M))). \tag{12}$$

1.4 损失函数设计

1.4.1 标签平滑交叉熵损失

为了增加网络分类的精度,交叉熵损失将行人的特征分为不同的类别,以便更好地捕获身份相关的信息.Radenovic 等[17]使用带有标签平滑的交叉熵损失来防止网络的过拟合且能提升模型的泛化能力,本文将该种方法引入损失函数中.对于一张图像,记 y 为其真实标签,p_{i} 为模型对其分到第 i 类的概率,则身份损失 L_{id} 的表达式为

$$L_{id}=\sum_{i=1}^{N}-q_{i}\log(p_{i}),\quad \text{s.t. } q_{i}=\begin{cases}1-\dfrac{N-1}{N}\xi, & y=i;\\ \dfrac{\xi}{N}, & y\neq i.\end{cases} \tag{13}$$

这里的 N 代表训练集中的不同身份数量,而 ξ 是一个常数,它在模型训练中用来调整对样本标签的依赖程度,有助于提升模型在面对错误标注时的修正能力.

1.4.2 自适应权重的难三元组损失

三元组损失[18]是一种用于训练人脸识别或行人重识别模型的损失函数,通过最小化锚点与正样本之间的距离并最大化锚点与负样本之间的距离,以实现特征嵌入的优化.为了防止模型陷入局部最优,人们一般会采用约束性能更好的难三元组损失作为度量损失,其表达式为

$$L_{\text{trihard}} = \sum_{i}^{N} [\max\{d(x_a,x_p)\} - \min\{d(x_a,x_n)\} + \alpha]_+. \tag{14}$$

其中:α 为边界值;P 为行人类别总数;N 表示一个批次的大小;x_a、x_p、x_n 分别表示锚点和正负样本;$d(\cdot)$ 为欧氏距离.

难三元组损失仅考虑极端样本,而样本的缺失部分会降低置信水平.本文设计了一种具有自适应权重的三元组损失类型,其中,自适应权重是基于样本之间的距离计算的.对于正样本对,距离和权重具有正相关性;对于负样本对,距离和权重之间存在明显的负相关性.其表达式如下:

$$L_{\text{trihard}_w} = \sum_{i}^{N} [\sum \omega_p(d(x_a,x_p)) - \sum \omega_n(d(x_a,x_n)) + \alpha]. \tag{15}$$

其中:$\omega_p = \dfrac{e^{d(x_a,x_p)}}{\sum\limits_{x\in P(a)} e^{d(x_a,x_p)}}$、$\omega_n = \dfrac{e^{-d(x_a,x_n)}}{\sum\limits_{x\in N(a)} e^{-d(x_a,x_n)}}$ 分别表示正负样本的自适应权重.

2 实验与分析

2.1 数据集与实验参数

本实验选用 SYSU-MM01[19] 多模态行人重识别数据集.该数据集包含了来自室内和室外 6 台摄像机的 491 个行人的 287 628 张 RGB 图像和 15 792 张红外图像.训练集包含 395 个身份的 22 258 张 RGB 图像和 11 909 张红外图像,测试集则包含 96 个身份的 301 张 RGB 图像和 3 803 张红外图像.SYSU-MM01 提供了 2 种搜索模式,分别为全搜索(all search)和室内搜索(indoor search).前者将可见光相机 1、2、4、5 的图像作为 gallery 集,将红外相机 3、6 的图像作为 query 集;而后者则将可见光相机 1、2 的图像作为 gallery 集,将红外相机 3、6 的图像作为 query 集.

RegDB[20] 是一个用于行人重识别的数据集,它模拟了室外环境下的行人检测和识别任务.RegDB 数据集包含了来自 2 个不同摄像头的 412 个行人的图像序列,分别为可见光和红外图像,每个人包含 10 张可见光和 10 张红外图像.RegDB 数据集的评估模式是可见光到红外(V to I)和红外到可见光(I to V).

本实验使用 Windows 11 操作系统,采用 PyTorch 1.13.1 + cu117 深度学习框架,PyCharm 17.4.0.1 作为编辑器.Python 版本为 3.9,模型训练过程中利用 NVIDIA GeForce RTX 3050 图形处理器.采用累积匹配特性(Cumulative Matching Characteristics,CMC)和平均准确率 mAP 作为评估指标,其中,CMC 反映了在不同排名下的命中率.

在实验过程中设定了一系列超参数:批量大小为 16,行人图像尺寸为 288×144 像素,图像增强中随机擦除的概率为 0.5;训练轮数设定为 100,优化器为 Adam 优化器[21],权重衰减设置为 5×10^{-4},动量为 0.9;学习率在前 20 轮设定为 0.1,在第 21~第 50 轮设定为 0.01,在第 51~第 100 轮设定为 0.001;在自适应权重的难三元组损失中,α 值设定为 1.

2.2 模型对比

2.2.1 本文提出模型与先进算法的对比

为了验证本文所提出模型的优越性,将本文提出模型的性能与该领域的先进算法比较,包括 IMT[12]、GPFF[22]、PDRNet[23]、AGM[11]、FMCNet[24]、MSO[25]、PMT[26]、NFS[9] 和 TVTR[27].为了更加全面地比较,本文将上述方法在 SYSU-MM01 的 all search 和 indoor search 两种模式下进行实验,对比结果如表 1 所示.从与其他先进算法的对比来看,本文提出的算法在 all search 模式的 Rank-1、Rank-10、Rank-20 和 mAP 均较高,分别达到了 70.71%、96.63%、98.83%和 68.05%,比其他先进算法至少高出 1.08、0.36、0.01 和 1.94 个百分点,这说明本文提出模型的鲁棒性强,检索和匹配性能较好.然而,在 indoor search 模式中,本文提出算法的 Rank-10 为 98.49%,不如 PDRNet 模型,比其低了 0.47 个百分点,这反映出本文提出的模型对某些室内场景下的变化不够敏感,因此这成为下一阶段的工作重心.

本文方法在 RegDB 数据集上与上述先进算法的对比结果如表 2 所示,其中 V to I 表示可见光到红外模态的检索,I to V 表示红外到可见光模态的检索.从与其他先进算法的对比来看,本文提出的算法在多数评价指标上达到了较好的性能,Rank-1 和 mAP 都达到了 SOTA,但是在部分 Rank-10 和 Rank-20 中没有 GPFF 模型表现得好,这可能是因为模型在不同域间的泛化性能以及数据规模迁移能力不强.这也是下一阶段需要解决的问题.

2.2.2 可视化对比

图 3 可视化了通过 CSA 模块增强的注意力特征图与基线模型的对比.从特征图的直观比较中可以观察到,本文模型能够更有效地聚焦行人的关键生物特征,如头部、肩部和腿部区域.这种精细的注意

表 1　在 SYSU-MM01 数据集上与其他先进算法的对比

Table 1　Comparison with other advanced algorithms on SYSU-MM01 dataset　%

方法	全搜索				室内搜索			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
IMT	56.62	90.26	95.59	57.47	68.72	94.61	97.42	75.11
NFS	56.91	91.34	96.52	55.45	62.79	96.53	99.07	69.79
MSO	58.70	92.06	97.20	56.42	63.09	96.61		70.31
GPFF	63.83	93.63	97.67	59.62	66.67	93.39	95.97	70.75
TVTR	65.30			64.15	72.21			77.94
FMCNet	66.34			62.51	68.15			74.09
PMT	67.53	95.36	98.64	64.98	71.66	96.73	99.25	76.52
PDRNet	68.48	95.74	98.73	64.96	74.52	**98.96**	99.43	78.80
AGM	69.63	96.27	98.82	66.11	74.68	97.51	99.14	78.30
本文	**70.71**	**96.63**	**98.83**	**68.05**	**76.58**	98.49	**99.61**	**80.36**

表 2　在 RegDB 数据集上与其他先进算法的对比

Table 2　Comparison with other advanced algorithms on RegDB dataset

方法	V to I				I to V			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
IMT	78.30	92.38	95.05	70.37	75.22	91.27	93.19	67.28
NFS	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
MSO	73.60	88.60		66.90	74.60	88.70		67.50
GPFF	91.22	**97.30**	**98.79**	84.33	89.89	96.82	**99.73**	82.26
TVTR	84.10			79.50	83.70			78.00
FMCNet	89.12			74.43	88.38			83.86
PMT	84.83			76.55	84.16			75.13
PDRNet	83.98	93.05	97.05	78.24	82.06	91.49	95.56	77.26
AGM	88.40	95.10	96.94	81.45	85.34	94.56	97.48	81.19
本文	**92.43**	96.22	97.24	**85.19**	**89.95**	**97.08**	99.16	**84.34**

力定位表明,本文模型对行人的关键识别信息有更高的灵敏度.此外,与基线模型相比,本文模型在背景噪声抑制方面展现出更优的性能.在多个测试场景中,基线模型的注意力图往往包括大量非目标区域的活跃响应,它们通常是行人后面的背景或其他干扰元素.本文模型则通过引入改进的空间注意力机制,有效地减少了对这些无关区域的关注,从而增强了模型对实际行人目标的聚焦能力.

为了直观地验证本文所提出模型的先进性,在 SYSU-MM01 数据集上进行了检索结果的可视化,如图 4 所示.将红外模态图像作为 query,可见光模态图像作为 gallery,绿色外框代表检索正确,红色外框代表检索错误.从图 4 中可以直观地感受到本文提出的基于融合注意力和特征增强的行人重识别模型的检索能力远高于基线模型,检索图像和待检索图像的匹配度较高,能够更加准确地识别行人身份.

2.3　消融实验

2.3.1　模块消融实验

为了验证本文提出模型中各个模块的有效性,对模块的有效性进行消融实验分析,结果如表 3 所示,其中,TA 表示自动数据增强,CSA 表示通道和空间注意力机制,L_{trihard_w}表示自适应权重的难三元组损失.本文模型在文献[20]的基础上进行优化,故选取其作为基线模型.在 all search 模式下,最终本文提出模型的 Rank-1、Rank-10、Rank-20 和 mAP 分别比基线模型高 5.83、2.98、1.62 和 3.59 个百分点,证明本文模型的有效性.在使用 TA 模块后,模型性能有小幅度的提升,说明自动数据增强确实能够缓解一部分视角、尺度差异,提升模型鲁棒性.进一步添加了 CrossViT 模块后,模型性能有了巨大提升,说明 CrossViT 能够在融合并丰富不同模态特征的同时捕获细粒度信息和高级语义表征.在加入 CSA 模块后,模型性能有一定的提升,说明 CSA 模块增强了网络

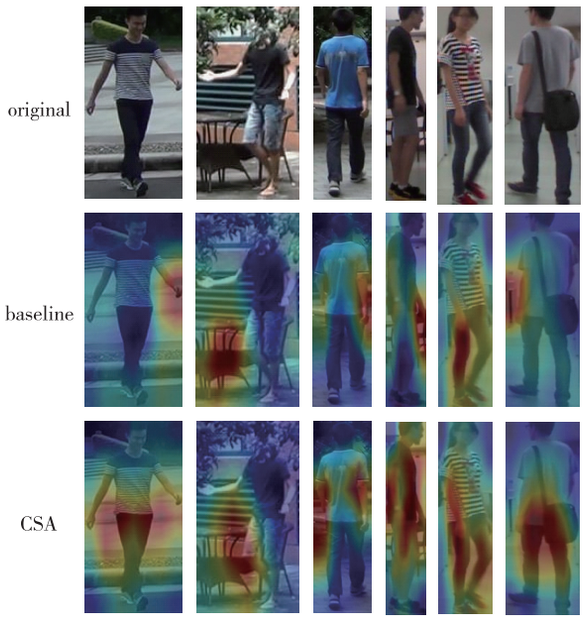

图 3　注意力特征图可视化

Fig. 3　Visualization of attention feature maps

捕获不同模态的通道和空间特征的能力.在加入L_{trihard_w}后模型性能有一定的提升,说明自适应权重的难三元组损失能够有效地约束模型学习方向.

2.3.2　损失函数消融实验

在损失函数设计中,本文提出了自适应权重的难三元组损失.为了验证其相较其他损失函数的优越性,将其与常见的中心损失L_{center}[28]和难三元组损失L_{trihard}[16]进行对比消融实验,结果如表 4 所示.可以看出,自适应权重的难三元组损失在各个评价指标上均高于中心损失和难三元组损失.在 SYSU-MM01 数据集上,Rank-1 和 mAP 在 all search 模式下至少提高了 0.91 和 0.31 个百分点,indoor search 模式下至少提高了 0.62 和 0.58 个百分点.在 RegDB 数据集上,Rank-1 和 mAP 在 V to I 模式下至少提高了 0.78 和 0.47 个百分点,在 I to V 模式下至少提高了 0.47 和 0.29 个百分点.对比实验证明自适应权重的难三元组损失通过分配不同权重来提高每个样本

图 4　在 SYSU-MM01 数据集上的 top-10 检测结果示例

Fig. 4　Examples of top-10 retrieval results on SYSU-MM01 dataset

表 3　模块有效性消融分析

Table 3　Module validity ablation analysis　%

方法				全搜索				室内搜索			
TA	CrossViT	CSA	L_{trihard_w}	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
baseline				64.88	93.65	97.21	64.46	72.17	95.86	98.42	76.37
√				65.14	93.87	97.35	64.99	72.74	96.07	98.68	76.89
	√			67.83	95.46	98.14	65.70	74.11	97.35	99.26	78.19
		√		67.15	95.08	98.02	65.41	73.92	97.06	99.14	77.83
√	√			68.20	95.94	98.36	65.95	74.68	97.79	99.39	78.52
√	√	√		68.93	96.32	98.52	67.26	75.21	98.13	99.50	79.03
√	√	√	√	70.71	96.63	98.83	68.05	76.58	98.49	99.61	80.36

表 4　不同损失函数的对比消融实验

Table 4　Comparative ablation results of different loss functions　%

损失函数	SYSU-MM01				RegDB			
	全搜索		室内搜索		V to I		I to V	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
L_{center}	69.13	67.44	75.60	79.25	91.33	84.59	89.36	83.92
L_{trihard}	69.80	67.74	75.96	79.78	91.65	84.72	89.48	84.05
L_{trihard_w}	70.71	68.05	76.58	80.36	92.43	85.19	89.95	84.34

的相关性,能够改善度量学习方法的性能并加速网络收敛速度.

3　总结

针对以往研究中对细粒度特征提取能力不强的问题,本文提出一种基于融合注意力和特征增强的跨模态行人重识别模型.本文的主要工作如下:1)利用自动数据增强对原始数据进行增强,增加模型的鲁棒性,缓解视角、尺度等差异;2)提出了基于交叉注意力多尺度 CrossViT 特征提取模型,能够在融合不同模态特征的同时捕获来自网络不同层次的信息,并获取细粒度信息和高级语义表征;3)提出了 CSA 模块,同时关注通道和空间级特征,在融合可见光和红外图像特征时学习对区分特征重要的信息;4)提出了自适应权重的难三元组损失,增强了样本之间的相关性,提高了可见光和红外图像对不同行人的识别能力.

参考文献

References

[1] Bhardwaj S, Dave M. Enhanced neural network-based attack investigation framework for network forensics: identification, detection, and analysis of the attack[J]. Computers & Security, 2023, 135: 103521

[2] Zhu J L, Li Q L, Gao C B, et al. Camera-aware re-identification feature for multi-target multi-camera tracking[J]. Image and Vision Computing, 2024, 142: 104889

[3] Zennayi Y, Benaissa S, Derrouz H, et al. Unauthorized access detection system to the equipments in a room based on the persons identification by face recognition[J]. Engineering Applications of Artificial Intelligence, 2023, 124: 106637

[4] Eli M B, Lidor R, Lath F, et al. The feudal glove of talent-selection decisions in sport-strengthening the link between subjective and objective assessments[J]. Asia Journal of Sport and Exercise Psychology, 2024, 4(1): 1-6

[5] Coşar S, Bellotto N. Human re-identification with a robot thermal camera using entropy-based sampling[J]. Journal of Intelligent & Robotic Systems, 2019, 98: 85-102

[6] Fu D P, Chen D D, Bao J M, et al. Unsupervised pre-training for person re-identification[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 20-25, 2021, Nashville, TN, USA. IEEE, 2021: 14750-14759

[7] 周非,舒浩峰,白梦林,等.生成对抗网络协同角度异构中心三元组损失的跨模态行人重识别[J].电子学报,2023,51(7):1803-1811
ZHOU Fei, SHU Haofeng, BAI Menglin, et al. Cross-modal person re-identification based on generative adversarial network coordinated with angle based heterogeneous center triplet loss[J]. Acta Electronica Sinica, 2023, 51(7): 1803-1811

[8] Yuan B W, Chen B R, Tan Z Y, et al. Unbiased feature enhancement framework for cross-modality person re-identification[J]. Multimedia Systems, 2022, 28(3): 749-759

[9] Chen Y,Wan L,Li Z H,et al.Neural feature search for RGB-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 20 - 25, 2021, Nashville,TN,USA.IEEE,2021:587-597

[10] Choi S,Lee S,Kim Y,et al.Hi-CMD:hierarchical cross-modality disentanglement for visible-infrared person re-identification [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).June 13 - 19, 2020, Seattle, WA, USA. IEEE, 2020: 10257-10266

[11] Liu H J, Xia D X, Jiang W. Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification [J]. IEEE Journal of Selected Topics in Signal Processing, 2023,17(3):545-559

[12] Xia D X,Liu H J,Xu L L,et al.Visible-infrared person re-identification with data augmentation via cycle-consistent adversarial network [J]. Neurocomputing, 2021,443:35-46

[13] Cubuk E D, Zoph B, Mané D, et al. AutoAugment: learning augmentation strategies from data [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).June 15-20,2019,Long Beach, CA,USA.IEEE,2019:113-123

[14] Müller S G, Hutter F. TrivialAugment: tuning-free yet state-of-the-art data augmentation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). October 10 - 17, 2021, Montreal, QC, Canada. IEEE, 2021:754-762

[15] Dosovitskiy A,Beyer L,Kolesnikov A,et al.An image is worth 16×16 words:transformers for image recognition at scale[J].arXiv e-Print,2020,arXiv:2010.11929

[16] Woo S,Park J,Lee J Y,et al.CBAM:convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV),2018:3-19

[17] Radenovic F,Tolias G,Chum O.Fine-tuning CNN image retrieval with no human annotation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,41(7):1655-1668

[18] Liu H J, Tan X H, Zhou X C. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification[J].IEEE Transactions on Multimedia,2080,23:4414-4425

[19] Wu A C,Zheng W S,Yu H X,et al.RGB-infrared cross-modality person re-identification [C]// IEEE International Conference on Computer Vision. October 22-29,2017,Venice,Italy.IEEE,2017:5380-5389

[20] Ye M,Shen J B,Lin G J,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

[21] Kingma D P,Ba J.Adam:a method for stochastic optimization[J].arXiv e-Print,2014,arXiv:1412.6980

[22] Wang X J,Cordova R S.Global and part feature fusion for cross-modality person re-identification[J].IEEE Access, 2038,10:122038-122046

[23] 刘志刚,常乐乐,赵宜珺,等.基于通道干预渐进式差异减小网络的跨模态行人重识别[J/OL].计算机辅助设计与图形学学报,2024:1-11.[2024-03-14].https://kns.cnki.net/kcms/detail/11.2925.TP.20240314.1047.012.html
LIU Zhigang, CHANG Lele, ZHAO Yijun, et al. Progressive difference reduction network with channel intervention for visible-infrared re-identification [J/OL]. Journal of Computer-Aided Design & Computer Graphics, 2024: 1-11. [2024-03-14]. https://kns.cnki.net/kcms/detail/11.2925.TP.20240314.1047.012.html

[24] Zhang Q,Lai C Z,Liu J N,et al.FMCNet:feature-level modality compensation for visible-infrared person re-identification [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).June 19-20,2022,Long Beach, CA,USA.IEEE,2022:7349-7358

[25] Gao Y J,Liang T F,Jin Y,et al.MSO:multi-feature space joint optimization network for RGB-infrared person re-identification[C]//Proceedings of the 29th ACM International Conference on Multimedia.New York,NY,USA. ACM,2021:5257-5265

[26] Lu H, Zou X Z, Zhang P P. Learning progressive modality-shared transformers for effective visible-infrared person re-identification [J]. Proceedings of the 37th AAAI Conference on Artificial Intelligence, 2023, 37(2):1835-1843

[27] Yang B,Chen J,Ye M.Top-K visual tokens transformer: selecting tokens for visible-infrared person re-identification[C]// 2023 IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). June 4-10,2023,Rhodes Island,Greece.IEEE,2023:1-5

[28] Luo H,Gu Y Z,Liao X Y,et al.Bag of tricks and a strong baseline for deep person re-identification [C]/2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).June 16-17,2019, Long Beach,CA,USA.IEEE,2019:1487-1495

Cross-modal person re-identification based on fused attention and feature enhancement

HUANG Chihan[1] SHEN Xiaobo[2]

1 School of Design Art and Media,Nanjing University of Science & Technology,Nanjing 210094,China

2 School of Computer Science and Engineering,Nanjing University of Science & Technology,Nanjing 210094,China

Abstract RGB-Infrared person re-identification (Re-ID) is a challenging task which aims to match person images between visible and infrared modalities,playing a crucial role in criminal investigation and intelligent video surveillance.To address the weak feature extraction capability for fine-grained features in current cross-modal person Re-ID tasks,this paper proposes a person re-identification model based on fused attention and feature enhancement.First,automatic data augmentation techniques are employed to mitigate the differences in perspectives and scales among different cameras,and a cross-attention multi-scale Vision Transformer is proposed to generate more discriminative feature representations by processing multi-scale features.Then the channel attention and spatial attention mechanisms are introduced to learn information important for distinguishing features when fusing visible and infrared image features.Finally,a loss function is designed,which adopts the adaptive weight based hard triplet loss,to enhance the correlation between each sample and improve the capability of identifying different persons from visible and infrared images.Extensive experiments conducted on the SYSU-MM01 and RegDB datasets show that the proposed approach achieves mAP of 68.05% and 85.19%,respectively,outperforming many state-of-the-art approaches.Moreover,ablation experiments and comparative analysis validate the superiority and effectiveness of the proposed model.

Key words person re-identification (Re-ID); cross-modal; cross attention; feature extraction; multi-scale