DOI:10.13878/j.cnki.jnuist.20240413002

张国庆[1]　杨珊[1]　汪海蕊[2]　王准[2]　杨艳[1]　周洁琼[1]

基于深度学习的多模态行人重识别综述

摘要

行人重识别(Re-ID)旨在跨像机检索同一目标行人,它是智能视频监控领域的一项关键技术.由于监控场景的复杂性,单模态行人重识别在低光、雾天等极端情况下的适用性较差.因实际应用的需要以及深度学习的快速发展,基于深度学习的多模态行人重识别受到了广泛的关注.本文针对近年来多模态行人重识别的发展脉络进行综述:阐述了传统单模态行人重识别方法存在的不足;归纳了多模态行人重识别的常见应用场景及其优势,以及各数据集的构成;重点分析了各种场景下多模态行人重识别的相关方法及其分类,并探讨了当前研究的热点和挑战;最后,讨论了多模态行人重识别的未来发展趋势及其潜在应用价值.

关键词

深度学习;神经网络;行人重识别;多模态

中图分类号 TP391.41

文献标志码 A

收稿日期 2024-04-13

资助项目 国家自然科学基金(62172231);江苏省自然科学基金(BK20220107)

作者简介

张国庆,男,博士,教授,主要从事复杂场景行人检索与重识别、目标检测、图像分类等研究.guoqingzhang@nuist.edu.cn

1　南京信息工程大学 计算机学院,南京,210044

2　南京信息工程大学 软件学院,南京,210044

0　引言

随着城市现代化建设的加速发展和智能安防的普及,行人重识别(Person Re-identification,Re-ID)作为智能视频分析领域的重要技术,具有不可或缺的作用.Re-ID 任务的核心目标是在不同的监控摄像机下提取行人图像衣着、配饰、体态等特征实现对同一行人的识别.由于实际应用的需要,Re-ID 技术得到了越来越多的重视,并在智能安防、智能交通、无人驾驶等领域[1]得到了广泛应用.然而,由于监控场景的复杂性以及多样性,采集的行人图像面临光照、视角、姿势变化,以及遮挡等因素的影响,存在巨大的模态内差异,行人重识别仍然具有一定的挑战性.

为了应对上述挑战,近年来,研究人员专注于缓解外在因素引起的类内变化,通过优化特征提取算法和增强数据预处理技术,显著地减少了由光照、视角和姿势变化等导致的识别误差.然而,可见光图像在弱光、雨天、雾天等复杂条件下的成像质量较差,限制了可见光单模态行人重识别在极端场景下的适用能力.考虑到不同传感器捕获的目标行人信息具有互补性,学者们开始探索其他非可见光模态信息的利用[2-4],例如红外图像、草图、文本描述等,以最大化其互补优势,此类模态信息能够在不同环境条件下提供稳定的特征信息,有效补充或替代可见光图像,提升模型整体的识别性能.

多模态行人重识别可以分为两种:1)跨模态行人重识别;2)模态融合行人重识别.具体来说,跨模态行人重识别旨在确立两种不同模态之间的有效匹配关系,例如从可见光到红外图像[5]、从草图到可见光图像[6]、从文本到图像[7]等,而模态融合行人重识别旨在充分利用多种模态信息之间的互补性和协同性,融合不同角度的行人判别性信息以丰富行人的特征表示,进而实现更全面、更准确的识别.

随着多模态行人重识别任务的飞速发展,研究者提出了多种涉及多模态信息利用的应用场景及方法[8-9].本文将重点介绍基于深度学习的多模态行人重识别方法,首先阐述其相关知识,随后介绍各种应用场景中的常用方法和相关数据集,最后是总结和展望.

1　相关知识

1.1　单模态行人重识别

行人重识别作为智能图像处理领域的一个重要研究课题,其核

心目标是跨摄像机追踪和识别行人.由于监控场景的复杂性,行人重识别技术仍面临许多亟待解决的难题[10]:1)由于监控系统中硬件设备及其部署的远近距离不同,获取的行人图像可能存在分辨率不同的情况(图1a);2)由于行人图像或视频的拍摄角度和地点不同,会产生光照、姿势、视角变化,对行人的外观特征产生巨大影响(图1b—d);3)真实场景中行人处于移动状态,不可避免地会出现行人部位遮挡(图1e);4)某些复杂场景会导致检测框不准确(图1f).

针对上述问题,当前的方法主要侧重于表示学习[11-18]和度量学习[19-22].表示学习侧重于学习如何从原始数据中提取更具有表现力的行人特征表示.Chen 等[11]提出一种空间和通道分区表示网络(SCR),在金字塔多分支架构中将特征图按通道分为相互关联的特征组,并联合全局特征构建具有识别性和概括性的特征表示.Chen 等[13]引入一种自我批评注意力学习方法,批评者衡量注意力质量并提供强大的监督信号来指导特征学习过程;此外,批评者模型通过估计注意力图的质量,有助于解释学习过程中注意力机制的有效性.度量学习侧重于设计一个合适的度量或距离函数,以便在特征空间中更好地衡量各样本的相似度或差异性,常用的度量函数有身份损失函数、三元组损失函数等.Yi 等[19]使用"暹罗"深度神经网络联合学习颜色特征、纹理特征,使用二项式偏差来评估相似性和标签之间的成本.Sikdar 等[21]提出一种批量自适应三元组损失函数,使最硬样本的权重根据其与锚的距离自适应调整,较好地克服了图像尺度对模型效果的影响.

然而,这些模型通常针对单一可见光模态设计,当涉及具有较大模态差异的多模态重识别任务时,这些方法一般不适用.

1.2 多模态行人重识别

不同于单模态行人重识别,多模态行人重识别考虑到可见光图像在低光、雾天等极端场景中成像质量差的特殊情况.本文将多模态行人重识别分为两种:1)跨模态行人重识别;2)模态融合行人重识别.

当前常见的跨模态行人重识别任务有可见光红

图1 行人重识别难点

Fig. 1 Challenges in person Re-ID

外行人重识别(Visible-Infrared Person Re-identification,VIReID)、草图到可见光图像行人重识别(Sketch ReID)、文本到可见光图像行人重识别(Text-to-Image Person Re-identification,TIReID)等.

行人红外图像可以反映目标物体的热分布情况,这种热分布信息在恶劣天气或者复杂环境的情况下(如夜间、弱光条件下、雨天、雾天、遮挡等)也能提供鲁棒性较强的行人信息.但由于可见光与红外图像之间存在较大的模态差异,除固有的模态内变化外,可见光红外行人重识别还需应对模态间差异带来的匹配困难问题.Wu 等[23]首先为行人重识别构建了一个名为 SYSU-MM01 的大规模可见光红外数据集,并引入了一种深度零填充方法自动推进网络中的特定领域节点进行跨模态对齐.最近,学者们提出了基于模态不变特征学习的可见光红外行人重识别,将各模态的特征映射到共享的特征空间中[24-25],此种方法注重学习模态间的共享特征,不能充分利用各模态的特定线索.Zhang 等[26]提出一种师生对抗模型(TS-GAN)将现有的可见光数据生成伪红外表示,以减少跨模态变化并指导模型提取模态特定特征.

行人的文本描述通常包含有关行人外貌特征和环境的描述,例如行人衣着颜色、发型、面部特征、所处地点等.在很多刑事案件中,工作人员可以根据证人的自然语言描述直接搜索目标行人图像,即文本到图像行人重识别任务.Li 等[2]首先提出用自然语言描述搜索目标行人的问题.Zhou 等[27]通过使用注意力机制引导模型对齐图像和文本模态的行人表示,充分利用行人的识别性信息.最近,Shao 等[28]分析了视觉和文本模态之间的粒度差异,提出一种基于文本行人重识别的粒度统一表示学习方法,缓解了文本图像信息粒度不统一的问题.

另一方面,还可以根据文字描述制作行人草图进行间接行人搜索.行人草图描述通常包含对行人轮廓的简略描绘,强调身体的整体形状、轮廓和姿势等基本外观特征.此外,草图还可以强调一些特殊的标志或细节,如发型、眼镜、服装上的图案和手部动作等.Pang 等[29]首先提出使用专业草图作为查询在 RGB 图库中搜索目标人物,设计了跨模态对抗性学习方法来挖掘模态不变的特征表示.Chen 等[3]提出一种新颖的非对称解耦方案解决草图和 RGB 模态之间的信息不对称问题.

考虑到跨模态行人重识别利用的模态信息有限,而模态融合行人重识别联合利用多种模态信息,最大化模态间的互补优势,获得更加全面的行人表示,提高模型在复杂场景中的性能.为了探索草图模态和文本模态特征间的互补性,Zhai 等[30]引入一种多模态行人重识别任务,它将草图和文本模态结合起来作为检索查询.为了解决描述性行人重识别模态缺失问题,Chen 等[31]首次提出了用描述性行人重识别来研究模态不可知的重识别任务,联合训练文本到 RGB、草图到 RGB、文本和草图到 RGB 三个任务,集成跨模态和多模态任务学习.

然而,当前的研究更多地侧重于处理两种模态的数据,涉及三种及以上模态数据处理的方法较少,多模态行人重识别的更多应用场景及方法有待深入探索和拓展,从而更好地适应实际的应用需求.

2 基于深度学习的常用方法

深度学习的兴起显著推动了行人重识别领域的进步,尤其是与神经网络相关的一系列算法[32-34]表现出色,研究人员开始关注基于深度学习的多模态行人重识别任务,其中最具代表性的是跨模态行人重识别.本章将首先介绍不同场景的跨模态行人重识别方法,随后论述模态融合行人重识别相关方法,最后对现有的多模态学习架构做分类概括.

2.1 跨模态数据对齐

根据查询和目标数据的模态,跨模态行人重识别可分为可见光到红外、草图到可见光图像和文本到图像行人重识别等.由于不同模态信息的数据分布不同,不同模态间存在巨大的模态差异,因此,跨模态特征对齐存在极大困难.为充分学习不同模态的相关匹配信息,需设计合适的特征提取和匹配模型.本节对可见光到红外、草图到可见光图像和文本到图像三种跨模态行人重识别方法及其分类做详细介绍.由于可见光到红外行人重识别和草图到可见光图像行人重识别方法分类类似,故将其同时进行可视化(图 2).

2.1.1 可见光红外行人重识别

可见光红外行人重识别旨在利用一种模态组成的查询集与另一模态组成的候选集进行匹配,例如可见光到红外图像匹配和红外到可见光图像匹配.该任务主要有两方面挑战:模态内变化和模态间差异.模态内变化主要是由同一模态下光照、视角、行人姿势等条件的变化引起,而模态间差异是由可见光和红外图像间的特征分布造成的.具体来说,可见

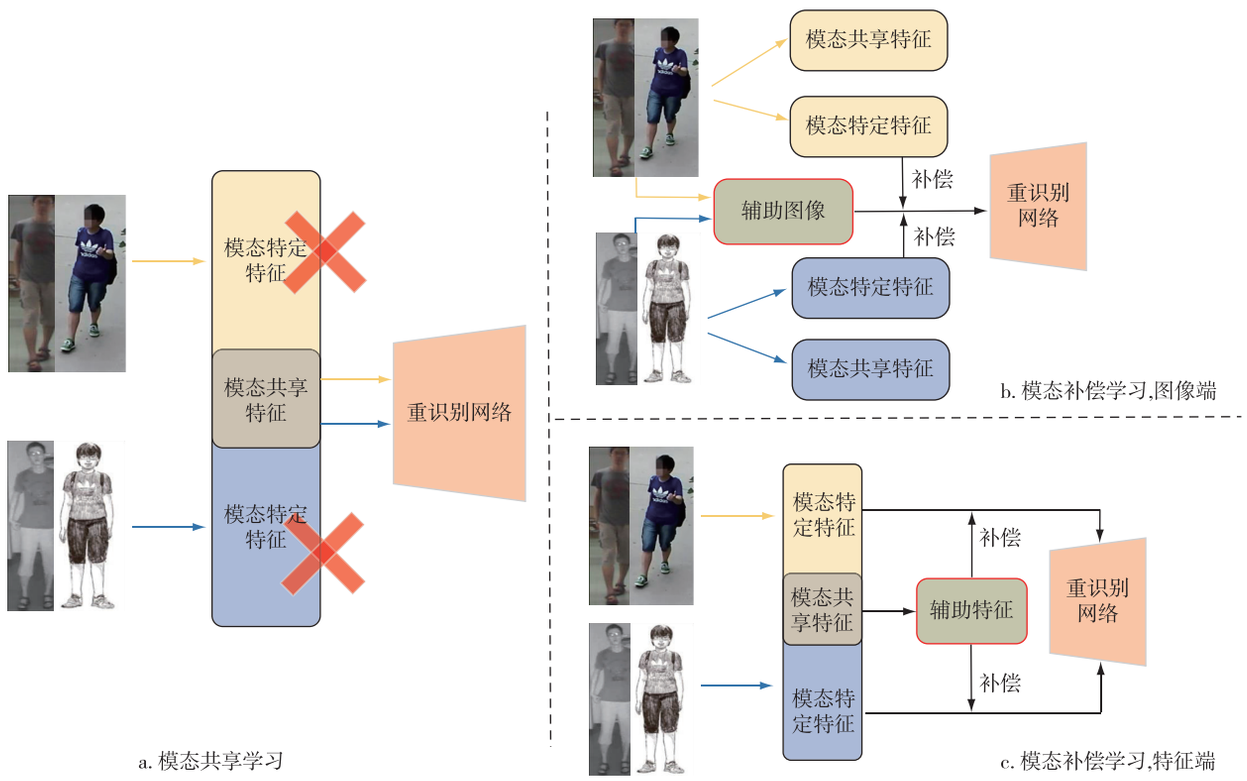


图 2　可见光红外行人重识别和草图到可见光行人重识别方法分类

Fig. 2　Classifications of VIReID and Sketch ReID

光和红外图像的成像原理不同,前者通常具有三个通道,主要包含色彩和纹路等丰富的视觉信息,而后者只有一个通道,提供轮廓和热力信息,细节信息较少,因此,二者的特征分布各异,进而造成同一身份模态间差异大于不同身份之间的模态内差异.

现有的方法可以分为基于模态共享特征学习[4,35-36]和基于模态补偿学习[37-39]两种.模态共享特征学习侧重于将不同模态特征投影到共享空间中,进而学习共享特征(图 2a).而基于模态补偿学习侧重于对现有的模态做图像或特征补偿,从多角度获取行人判别性信息,以获得更加全面的行人表示(图 2b、c).

基于模态共享特征学习的方法试图提取可见光图像与红外图像的共享特征,进而获取有判别性的行人表示(图 2a).常用方法有探索多级特征[4,35,40-42]、挖掘全局和局部信息[43-44]、采用频度信息[45-46]、解耦[47]等.Xiang 等[35]在双流网络中采用多粒度网络,以多级特征方式提取具有判别性的共享特征.Lu 等[48]提出一种共享特定特征转移(cm-SSFT)算法,在特征提取器上添加对抗和重建模块,对不同模态样本之间的亲和力建模以传播信息,充分利用了每个样本之间的丰富关系.Wei 等[42]提出一种基于灵活身体分区模型的对抗性学习方法(FBP-AL),可以根据特征图中最大响应的位置灵活地对特征图进行聚类,达到行人图像特征自动区分部分表示的效果.最近,有研究人员对图像使用傅里叶变换提取图像的频度信息和相位信息,进而利用其对齐图像的风格和语义.如 Zhang 等[45]提出一种新颖的频域细微差别挖掘(FDNM)方法,通过幅度引导相位模块和幅度细微差别挖掘模块探索频域可见光与红外图像间的细微差别,从而有效地减少频域的模态差异.此外,Hu 等[47]从信息解耦角度创新性地提出一种新颖的对抗性解耦和模态不变表示学习(DMiR)方法,该模型使用身份网络和领域网络,通过对抗解耦过程将输入特征分别解耦为身份相关特征和领域相关特征,进而进行特征表示与对齐.

基于模态共享特征学习的方法目前已经取得了很大的进步,但这些模型不可避免地会丢弃大量与个人相关的模态特定信息,阻碍行人的部分判别性信息充分利用,因此引入基于模态补偿学习的方法.基于模态补偿学习的方法试图充分利用模态共享特征学习过程中忽略的各模态特定的特征,通过联合共享特征和易忽略的特定特征获得更全面的行人表示.根据信息补偿的方式可将基于模态补偿学习分

为图像端补偿[5,37-38,49]和特征端补偿[39,41,50-51]两种(图2b、c).

图像端补偿通常使用GAN网络或变分自动编码器等技术将不同模态图像转化成同一模态,包括可见光模态转化为红外模态、红外模态转化为可见光模态以及生成中间模态.Wang等[37]结合像素对齐和特征对齐,提出一种新的可见光红外行人重识别对齐生成对抗网络(AlignGAN),由可见光图像生成伪红外图像后,通过对抗网络减少跨模态和模态内的变化,并捕获身份一致特征.Dai等[38]设计了一种新颖的CE2L模型,通过模态转换操作将红外图像转换为可见光图像后,使用特征提取模块和特征学习模块来提取它们的判别特征.Hu等[49]提出一种新颖的对抗性解耦相关网络(ADCNet),该网络通过特征解耦网络提取模态共享表示并结合各模态特定信息生成可见光红外图像对,进而通过二阶非局部操作来细化身份一致性信息.Zhang等[5]将可见光图像转换为灰度图像以减轻可见光和红外图像间的视觉差异,提取灰度图像与红外图像共享特征后,通过双重注意力特征增强模块从共享特征中挖掘更多有用的上下文信息,以缩短模态内的类间距离.

与图像端补偿不同,特征端补偿通常是在初步特征提取后进行特征交互达到特征补偿的效果,进而执行对齐操作.Yu等[39]提出一种新颖的模态统一网络(MUN),通过跨模态学习器和模态内学习器生成一种强大的中间特征,随后用身份对齐损失和模态对齐损失约束网络跨三种模态对齐身份.Zhang等[51]引入一种特征级模态补偿网络(FMCNet),利用模态共享特征生成另一模态的特定特征,进而融合共享特征与特定特征达到特征补偿的目的.Feng等[50]提出跨模态交互Transformer(CMIT)框架,利用不同模态的CLS标签之间的特征交互实现特征补偿.

2.1.2 草图到可见光图像行人重识别

草图到可见光图像行人重识别是指利用手绘的行人草图实现与行人可见光图像的匹配.草图通常是由用户根据语言描述手绘而成,包含人物轮廓信息,缺乏颜色与纹理信息.该任务的难点主要在于不同绘画者手绘的草图通常是抽象和风格各异的,与真实摄像机捕获的可见光图像具有显著的模态差异.

早期研究者针对基于草图的图像检索任务提出了许多方法[52-54],包括注意力机制、图神经网络等.与可见光红外行人重识别类似,在草图到可见光图像的行人识别中,现有方法可以分为基于模态共享特征学习和基于模态补偿学习两种.

基于模态共享特征学习的方法一般使用双流网络,采用额外的局部特征学习分支,这些分支可以促进模型更好地关注不同模态的相关信息,学习有判别性的特征表示[29,55-63](图2a).Pang等[29]提出一种对抗特征学习机制,通过跨模态对抗特征学习框架来过滤低级干扰特征并保留高级语义信息,学习身份特征和模态不变特征.但该方法丢失了部分有利于行人识别的模态特定信息,并且没有考虑联合优化草图和可见光图像特征表达学习.Lin等[57]针对草图描述的主观性设计非局部融合模块和属性对齐模块融合草图主观性,并利用属性作为隐式掩码来对齐跨模态特征,达到引入客观性的效果.Yang等[58]针对适度服装变化行人重识别任务提出一种可学习的空间极坐标变换模型来自动选择相对不变和有区别的局部草图曲线特征,并引入角度特定提取器来对每个角度条纹的特征图通道之间的相互依赖性进行建模,以探索细粒度的角度特定特征.Chen等[59]从光谱角度入手提出一种跨谱图像生成(CSIG)方法,使用双流特征提取器在多种光谱图像上训练,迫使网络挖掘多频谱图像的共享特征.Zhu等[63]提出一种新颖的跨模态注意力(CDA),使模型更多地关注可见光图像中与草图相关的区域,有效地减小了两个模态之间的差距.

基于模态补偿学习的方法试图利用现有的模态生成辅助信息在共享特征的基础上补充更多的特定信息,进而获得更全面的行人表示[3,6,64](图2b、c).Chen等[3]基于Transformer提出SketchTrans模型解决了草图识别任务,借助动态生成的辅助草图模态特征与原始可见光图像特征之间的关系进行非对称解耦学习,并通过知识迁移将草图特征表示转换为可见光图像特征表示进行模态间优化对齐.SketchTrans模型也被应用到行人重识别领域[6],通过非对称解耦后做信息补偿进行跨模态对齐;此外,还提出模态感知原型对比学习方法拉近不同模态之间的距离.上述方法忽略了每种模态内的显著差异,因此,Liu等[64]提出一种辅助学习网络,联合草图、生成的辅助模态和可见光图像三种模态数据,通过模态交互注意力模块迫使学习到的表示分布在不同模态和每种模态内保持不变,以达到特征对齐的目的.

2.1.3　文本到图像行人重识别

文本到图像行人重识别旨在利用给定的文本描述与图像中的行人进行准确关联和匹配,主要有以下挑战:

1)图像特征和文本特征之间模态差异的问题:图像是通过像素点的颜色和位置来表示的,而文本描述是通过字词的排列组合来表达的,所以图像和文本描述的特征分布是不同的.

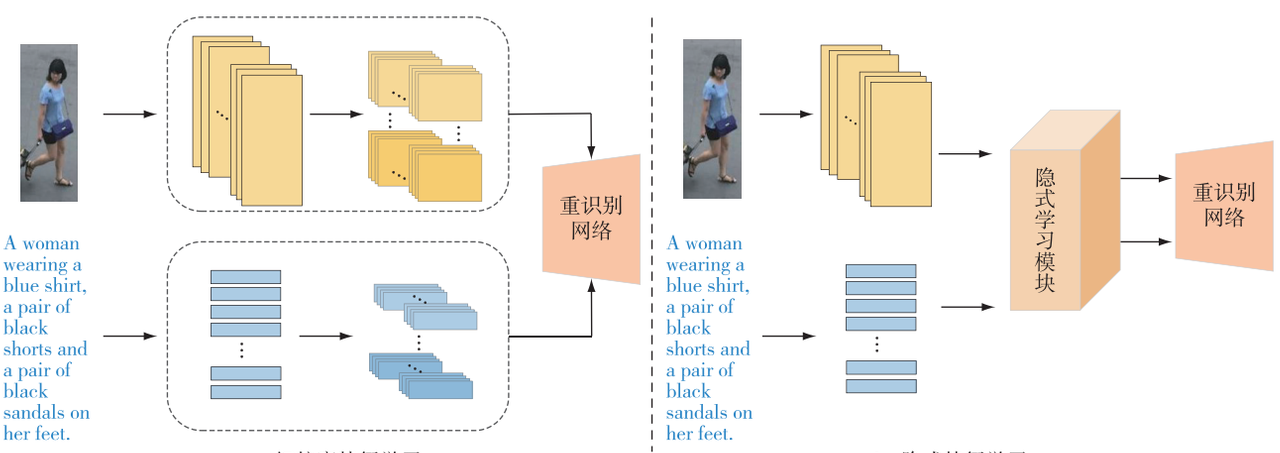
2)文本描述质量和形式的变化剧烈问题:文本描述通常是灵活的、具有主观性的,这可能导致同一图像有不同的文本描述方式.

3)文本描述和图像之间信息不平衡问题:相对于文本描述,图像信息可能会包含更多的环境或行人特殊情况信息.

早期的工作[2,65]利用 VGG 和 LSTM 来学习视觉文本模态的表示,并使用匹配损失来对齐它们.后来的工作使用 ResNet50/101[32] 和 BERT[66] 改进特征提取主干,并设计新颖的跨模态匹配损失[67-69]在联合嵌入空间中对齐图像与文本特征.最近的工作则广泛采用额外的局部特征学习分支[67,70-71],这些分支充分利用了身体部位和文本短语等信息进行特征对齐.Chen 等[70]提出一个部分卷积基线(TiPCB),在视觉主干后应用 PCB 模型[12]进行图像局部特征提取.Ding 等[67]在局部分支引入单词注意力模块引导模型关注与局部图像相关的单词,同时引入多视图非局部网络(MV-NLN)解决模态间的特征对齐问题.

考虑图像和文本描述信息之间的粒度不同,一些工作从细粒度角度进行特征学习[7,14,72-73](图 3a).Niu 等[7]提出一种多粒度图像文本对齐(MIA)模型,分层次地进行全局-全局、全局-局部和局部-局部三种不同粒度的对齐,以缓解跨模态粒度不同的问题.Yan 等[72]提出一种 CLIP 驱动的细粒度信息挖掘框架(CFine),设计了跨粒度特征细化模块和细粒度对应发现模块,以在不同粒度建立跨模态对应,确保局部补丁/单词的可靠性.Chen 等[14]在现有的 TiPCB 模型中引入文本部分感知匹配(TPM)模块,使模型从视觉和文本部分感知方面挖掘更全面的局部感知信息,进而将得到视觉/文本局部特征和视觉/文本局部感知特征进行多级特征融合,以进行后续的匹配.姜定等[73]借助 CLIP 模型的跨模态文本图像对齐的能力,提出仅使用全局特征的 Transformer 网络,并提出温度缩放跨模态投影匹配损失,以约束模型进行细粒度的语义特征对齐.

考虑显式生成的局部部分可能会产生上下文缺乏和噪声引入的问题,一些工作通过注意力机制或引入可学习的数据单元等方法隐式地进行局部特征学习[28,74-77](图 3b).Shao 等[28]针对模态间特征粒度的差异,引入一种新颖的学习粒度统一表示(LGUR)框架,通过引入多模态共享字典和一组可学习的原型构建两个模块达到粒度统一对齐的效果.Yan 等[75]提出一种高效联合多级对齐网络(MA-Net),通过关系引导的注意力和通道注意力获得增强特征后,引入可学习的语义主题中心隐式地进行局部对齐.Gao 等[77]针对文本图像之间信息不平衡的问题提出一种文本引导去噪和对齐(TGDA)模型,利用一个可学习的原型隐式地指导图像特征突出行人部分,进而利用偏差感知注意力从局部对齐



a. 细粒度特征学习　　b. 隐式特征学习

图 3　文本到图像行人重识别方法分类

Fig. 3　Classification of TIReID

文本图像特征.有些学者利用最近的视觉语言模型强大的多模态表示理解能力来构建相关模块以降低模型复杂度,Jiang 等[76]证明了 CLIP 模型可以轻松地转移到文本到图像的人物检索任务,提出一种跨模态隐式关系推理和对齐框架(IRRA)来学习更具辨别力的图像文本嵌入.

2.2 多模态数据融合与泛化

目前,多模态行人重识别任务以两种模态之间进行跨模态匹配为主,许多方法生成辅助模态并利用辅助模态信息联合训练网络[6,9,39],直接结合多种模态数据进行训练的行人重识别方法[8,30-31]较少,亟待研究者探索.联合使用多种模态信息需要设计合适的融合模型和训练策略,以确保捕获的行人特征既保留关键信息又不受模态间噪声的干扰.

利用多模态数据获取行人表征常见的方法是根据已有的数据生成辅助模态并利用辅助模态信息联合训练网络.Ye 等[9]针对可见光红外行人重识别提出了同质增强三模态(HAT)学习方法,从均匀的可见光图像生成辅助灰度图像联合训练,从多模态分类和多视图检索角度处理三模态特征学习问题.Chen 等[6]基于 Transformer 提出的 SketchTrans 模型解决了草图识别任务,借助草图动态生成网络生成辅助草图与原始可见光图像进行非对称解耦学习,随后通过知识迁移将草图特征表示转换为可见光图像特征表示进行信息融合,并进行模态间优化对齐.

生成辅助模态进行多模态联合训练需引入现有的生成模型,进而增加模型的复杂度与计算量,一些方法利用现有的多种模态的数据直接进行数据融合与泛化[8,30-31].为了探索草图模态和文本模态间的互补性,Zhai 等[30]首先提出使用草图和文本模态作为查询来实现多模态行人重识别,借助生成对抗性网络分别将视觉空间和描述性空间的结构信息与内容信息分离,进而将分离的信息进行交叉融合以缩小视觉特征与描述特征之间的差异.随后,为了解决描述性行人重识别模态不可知和模态缺失问题,Chen 等[31]提出一种统一行人重识别架构(UNIReID),联合训练文本到 RGB、草图到 RGB 以及文本和草图到 RGB 识别三个任务,以集成跨模态和多模态任务学习.

最近,一些研究者从开放场景的角度在多种场景的数据集上对模型进行联合训练[78-80].He 等[78]提出统一的指令行人重识别,在多种场景的训练数据上通过场景指令联合训练模型,使其能够通过查询图像和多模态指令解决常见的 6 个行人重识别任务,即传统行人重识别、换衣行人重识别、基于衣服模板的换衣行人重识别、语言指令行人重识别、可见光红外行人重识别和文本到图像行人重识别,为多场景行人重识别任务作出了巨大贡献.Zhang 等[80]为开放世界行人重识别构建了一个名为 OWD 的大规模、多样化、跨时空数据集,并在此基础上提出潜在域扩展方法(LDE),通过解耦和域扩展模块开发模型的泛化能力.Wei 等[79]考虑到在光照条件弱的情况下进行换衣的场景,构建了一个名为 NEU-VICC 的可见光红外换衣数据集,并在此基础上提出一种语义约束换衣增强网络(SC3ANet),对可见光和红外图像分别进行换衣操作之后,引入双粒度约束损失模块指导细粒度特征学习.

2.3 多模态学习架构分类

分析前两节不同场景的多模态行人重识别相关方法,可以将多模态学习架构分为两类:辅助模态学习和模态融合学习(图 4).

如图 4a 所示,辅助模态学习侧重于对现有的模态做图像或特征补偿,并通过特征交互以获得更全面的行人表示[3,37,39,61].如 Jiang 等[61]提出一种新颖的跨模态转换器(CMT),引入模态级对齐模块,通过 Transformer 编码器-解码器架构来补偿模态特定信息的缺失.

而如图 4b 所示,辅助模态学习侧重于利用现有的多种模态的数据直接进行数据融合与泛化[8,30-31].如 Wang 等[8]基于 Transformer 提出 TOP-ReID 模型,通过循环标记排列模块联合利用 RGB 图像、近红外图像和热红外图像的互补信息,以对齐不同光谱的空间特征,同时可以促进不同光谱信息感知其他光谱的局部细节.

3 数据集及评价指标

3.1 常用数据集

目前,常用的多模态数据集涉及可见光红外数据集、草图到可见光图像数据集和文本图像数据集,为促进多模态行人重识别的发展,仍需要探索更多模态和更大规模的多模态数据集.本节对常见数据集做详细介绍.

3.1.1 可见光红外数据集

为了评估可见光红外行人重识别方法的性能,常用的包含可见光图像和红外图像的公开基准数据集有三个,分别是 SYSU-MM01[23]、RegDB[81]和 LLCM[4],汇总信息如表 1 所示.

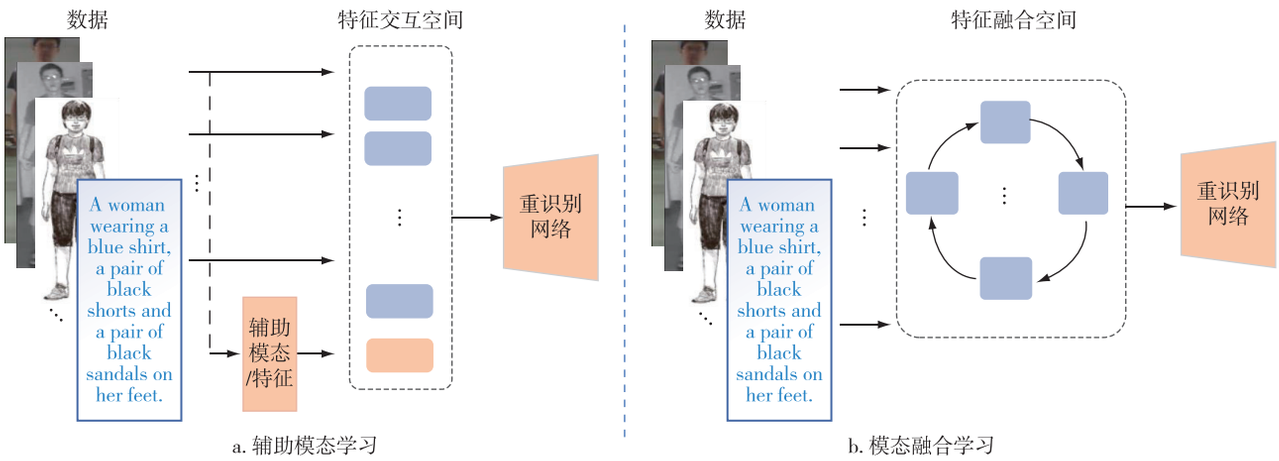


图 4　多模态方法分类

Fig. 4　Classification of multi-modal Re-ID

1)SYSU-MM01 数据集共包含 491 个行人身份,其中 296 个身份用于训练,其余 195 个身份用于验证和测试.训练时,应用训练集中 296 人的所有图像,在测试阶段,来自 RGB 相机的样本用于候选集,来自红外相机的样本用于查询集.数据集设计了全搜索模式和室内搜索模式 2 种测试模式.全搜索模式中,RGB 摄像机 1、2、4 和 5 用于候选集,红外摄像机 3 和 6 用于查询集;室内搜索模式中,RGB 摄像机 1 和 2 用于候选集,红外摄像机 3 和 6 用于查询集.

2)RegDB 数据集使用双摄像机捕获了 412 个人在没有任何指令的情况下移动的图像,它包含 412 个行人身份的 8 240 张图像,每个行人对应的可见光和红外图像分别有 10 张.412 人中,女性 254 人,男性 158 人,其中 156 人是从正面拍摄的,另外 256 人是从背面拍摄的,每个人的 10 张图像在身体姿势、光照条件等方面都存在差异.该数据集被随机分成两部分用于训练和测试,每部分包含 206 个身份,有可见光到红外、红外到可见光 2 种测试模式.

3)LLCM 数据集是一个在低光环境下收集的跨模态数据集,利用部署在弱光环境中的 9 个摄像机白天捕获可见光图像,夜间捕获红外图像,数据集收集时间为 100 d,考虑了气候条件和衣服变化.按照大约 2∶1的比例将数据集划分为 2 个部分分别用于训练和测试,训练集包含 713 个身份的 30 921 个边界框(16 946 个边界框来自可见光模态,13 975 个边界框来自红外模态),测试集包含 351 个身份的 15 846 个边界框(8 680 个边界框来自可见光模态,7 166 个边界框来自红外模态).与 RegDB 类似,在测试阶段,该数据集也分为可见光到红外、红外到可见光 2 种测试模式.此外,也同样计算 10 次测试结果的平均值作为最终结果,以获得稳定的测试效果.

表 1　可见光红外数据集细节

Table 1　Details of VIReID datasets

数据集	行人/个	相机/个	可见光图像/张	红外图像/张
SYSU-MM01[23]	491	6	287 628	15 792
RegDB[81]	412	2	4 120	4 120
LLCM[4]	1 064	9	25 626	21 141

3.1.2　草图到图像数据集

为了评估草图到图像行人重识别方法的性能,常用的包含草图图像和可见光图像的公开基准数据集有 PKU-Sketch[29]、Market-Sketch-1K[57],汇总信息如表 2 所示.

1)PKU-Sketch 数据集是第 1 个草图到可见光图像数据集.该数据集由 200 个行人组成,每个身份都有来自 2 个不同相机的 2 张可见光图像和 1 张草图,草图由 5 位绘图人员完成,并从每个绘图人员绘画的图像中随机选择 3/4 的行人图像进行训练,1/4 的行人图像进行测试,以消除绘画风格的影响,总体而言,有 150 人进行训练,50 人进行测试.

2)Market-Sketch-1K 数据集基于 Market-1501 数据集构建,从 Market-1501 的训练集中选择 498 个身份,从查询集中选择 498 个身份,充当 Market-Sketch-1K 数据集的可见光图像部分,每个身份的行人草图由 6 位绘图人员绘制,共包含 996 个身份的 4 763 个草图图像和 1 501 个身份的 32 668 张可见光图像,与 PKU-Sketch 数据集相比具有规模大、多视角和多风格的特点.

表 2　草图到可见光图像数据集细节

Table 2　Details of Sketch ReID datasets

数据集	行人/个	绘图人员/个	RGB 图像/张	草图/张
PKU-Sketch[29]	200	5	400	200
Market-Sketch-1K[57]	996	6	32 668	4 763

3.1.3　文本到图像数据集

为了评估文本到图像行人重识别方法的性能,常用的包含图像和文本描述的公开基准数据集有三个,分别是 CUHK-PEDES[2]、ICFG-PEDES[67] 和 RSTPReID[82],汇总信息如表 3 所示.

表 3　文本图像数据集细节

Table 3　Details of TIReID datasets

数据集	行人/个	图像/张	文本/个
CUHK-PEDES[2]	13 003	40 206	80 412
ICFG-PEDES[67]	4 102	54 522	54 522
RSTPReID[82]	4 101	20 505	41 010

1) CUHK-PEDES 数据集由香港中文大学于 2017 年提出,是第 1 个文本图像基准数据集,包含详细的自然语言描述和来自各种来源的人物样本.数据集分为 3 个子集用于训练、验证和测试,且没有相同的人员身份重叠,训练集由 13 003 个身份、40 206 张图像和 80 412 个句子描述组成,验证集和测试集分别包含 3 078 和 3 074 张图像,且都有 1 000 个行人,测试数据的图像和文字描述分别构成候选集和查询集.

2) ICFG-PEDE 数据集由华南理工大学于 2021 年提出,与 CUHK-PEDES 相比,该数据集包含更多关注身份和更细致的文本描述.该数据集共包含 4 102 个身份的 54 522 张行人图像,所有图像均来自 MSMT17 数据集[83],每张图像有一个文本描述,每个描述平均包含 37.2 个单词,共包含 5 554 个唯一单词.该数据集分为 2 个子集用于训练和测试,前者包含 3 102 人的 34 674 个图像文本对,而后者包含其余的 1 000 人的 19 848 个图像文本对.

3) RSTPReID 数据集是南京工业大学于 2021 年提出,基于 MSMT17[83] 构建,包含来自 15 个摄像机的 4 101 个人的 20 505 张图像,每个行人有 5 张不同相机拍摄的对应图像,每张图像都附有 2 段文字描述.对于数据划分,分别使用 3 701、200 和 200 个身份进行训练、验证和测试,每段描述不少于 23 个词,丢弃出现次数少于 2 次的单词后,单词数量为 2 204.

3.2　评价指标

模型训练后需要一个统一的评价指标来衡量方法的准确度好坏,目前常用的评价指标有:平均精度(mAP)和标准累积匹配特征(CMC),下面将分别对其进行介绍:

1)平均精度(mAP)是对多个查询的性能的平均度量.首先,计算每个查询的 AP 值,即对单个查询的命中概率的平均值,然后取所有查询的 AP 的平均值得 mAP.mAP 综合了多个查询的性能,可以更全面地评估模型在整个数据集上的表现.实际应用中,Rank-K,即查询图像的正确匹配出现在检索结果的前 K 个候选集中的概率,通常与 mAP 一起使用,以全面评估模型的性能.

2)标准累积匹配特征(CMC)曲线显示的是在前 K 个检索结果中是否包含真正匹配的样本,其中,K 从 1 开始逐渐增加.曲线横坐标表示排名,纵坐标表示在前 K 个结果中包含真正匹配的概率.CMC 曲线越高,表示在前 K 个结果中包含真正匹配的概率越大.CMC 曲线直观地表示了在前 K 个结果中包含真正匹配的概率,易于理解,但 CMC 的结果会受到 K 值选择的影响,因此,在使用 CMC 时需要选择合适的 K 值,通常会结合其他指标一起考虑.

总的来说,mAP 提供了一个全面的性能评估,反映了整个数据集上的平均表现,而标准累积匹配特征(CMC)曲线则提供了在前 N 次检索中成功匹配的概率,更直观地展示了不同检索次数下的性能.但这两种评价指标的计算方式都是计算对应查询排名的离散值.近年来,一些新颖的评价指标开始从相似度的连续值入手,计算相应相似度的平均值并取得了优异的效果.通过结合这些传统和新兴的评价方法,我们能够更全面和准确地评估行人重识别模型的性能.

4　问题及发展趋势

4.1　存在的问题

本文概述了多模态行人重识别的最新发展,总结了广泛采用的方法、可用的数据集,并对现有技术进行了比较.多模态行人重识别是一个活跃且有前途的研究领域,具有广泛的潜在应用价值,但许多问题仍然存在:

1)缺乏大规模多模态数据集.对于多模态行人重识别的各种应用场景,现有的数据集大多是涉及两种模态的跨模态数据集,且该类数据集还存在选

择有限、不同模态数据量不平衡、场景单一等问题,缺乏大规模的、包含多模态多场景的数据集限制了深度学习模型在真实多模态场景中的训练和泛化能力,这对提升模型性能和实际应用效果提出了严峻挑战.

2)泛化能力不足.当前的多模态行人重识别模型主要是针对特定场景数据集中的行人进行特征学习与对齐,在面对新的、未见过的数据时,其适应能力和表现能力仍有待提高.这一问题在实际应用中尤为明显,因为真实世界中的行人数据具有高度的多样性和复杂性,而现有模型在训练过程中往往缺乏对这些多样性和变化性的有效处理,导致其在新的环境或条件下的识别效果降低.

3)缺乏更多标准的评估指标.多模态行人重识别的评估涉及到多个模态的信息,当前多模态行人重识别任务的评估仍是在具体的应用情况及其对应的数据集下进行的,缺乏多样的标准评估指标,这不仅影响了模型开发和优化的进程,也给学术研究和工业应用带来了不便.具体来说,现有评估指标主要集中在精度、召回率等传统指标上,未能充分考虑多模态数据的特性和要求,如跨模态识别任务中,如何有效地评估不同模态之间的匹配效果,如何量化模型在处理模态差异时的表现,这些都是需要深入研究的问题.

总体来说,尽管多模态行人重识别领域已经取得了显著的进展,但仍有许多关键问题亟待解决,只有在数据集建设、模型泛化能力提升和评估标准完善等方面取得突破,才能真正推动这一技术在实际应用中的广泛落地和发展.

4.2 发展趋势

在深度多模态行人重识别领域,为了提升方法的性能和实用性,未来的发展趋势主要涉及以下几个方面:

1)行人特征方面.研究者应该对行人的本质特征进行深入探索,关注何种特征最能代表行人的判别性,而不仅仅专注于提高特征匹配的准确度.这包括对行人不同模态下的特征进行深度挖掘和分析,以及这些特征在不同环境条件下的表现,通过深入理解和提取这些本质特征,可以构建更加稳定和鲁棒的特征表示,提升模型的准确率和可靠性.

2)技术演进方面.随着大语言模型(如 GPT-4、BERT 等)的迅速发展,研究者应该探索如何利用这些大模型推动多模态行人重识别任务充分利用先验信息,促进模型鲁棒性的发展.例如,大语言模型可以提供丰富的上下文信息和先验知识,这些信息可以帮助多模态行人重识别模型更好地理解和解释图像中的细节,从而提升模型的整体性能.

3)应用场景方面.通过联合建模以及拓展技术应用场景实现对多模态数据、多任务协同和多场景分析的统一任务处理,也是未来的趋势之一.例如,在智能城市建设中,可以将多模态行人重识别技术应用于交通管理、公共安全监控、智能安防等多个场景,实现对不同数据源的综合分析和处理,提高系统的整体效率和智能化水平.

4)模型参数方面.多模态行人重识别模型通常包含更多参数,为提升实用性,轻量级模型的开发也尤为重要.研究者可以通过模型压缩、剪枝、量化等技术手段,优化模型的结构和参数,开发出性能优异且计算成本低的轻量级模型,这不仅可以降低计算资源,提高模型的运行效率,还可以在资源受限的环境中实现高效的行人重识别.

总体而言,未来深度多模态行人重识别的发展将围绕提升特征表示的区分性和鲁棒性、融合先进技术、扩展应用场景以及优化模型参数等方面展开.通过这些努力,将不断推动该领域的技术进步,满足实际应用需求,推动社会发展和智能化进程.

5 总结和展望

本文针对多模态行人重识别领域的研究现状,对深度多模态行人重识别方法从不同应用场景进行了归纳和总结.首先介绍多模态行人重识别的基础概念和相关知识,描述该技术在各种实际情况下的应用情况,随后系统地介绍当前各种应用场景中的常用方法和模型,以及常见的相关数据集和评价指标,并总结了多模态行人重识别当前存在的亟待解决的问题和未来发展趋势,为初学者了解多模态行人重识别常见分类与方法提供了有效途径,为研究者提供了快速归纳现有方法的视角以及值得探索的方向.虽然经过多年的发展,多模态任务取得了一定的成就,但随着深度学习技术的不断进步,对于模型轻量化、无监督学习、多任务统一学习等方向的研究需求逐渐凸显.期待未来出现更加全面、高效、鲁棒的行人重识别方法,以满足不断增长的实际应用需求,为社会发展带来更大的推动力,在公共安全、智能城市建设等领域发挥更重要的作用.

参考文献
References

[1] 何智敏,许佳云.基于深度学习的行人重识别算法研究进展[J].智能制造,2023(3):80-83
HE Zhimin, XU Jiayun. Research progress of pedestrian re-recognition algorithm based on deep learning[J]. Intelligent Manufacturing, 2023(3):80-83

[2] 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). July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017:5187-5196

[3] Chen C Q, Ye M, Qi M B, et al. Sketch transformer: asymmetrical disentanglement learning from dynamic synthesis [C]//Proceedings of the 30th ACM International Conference on Multimedia. October 10-14, 2022, Lisboa, Portugal. ACM, 2022:4012-4020

[4] Zhang Y K, Wang H Z. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-22, 2023, Vancouver, BC, Canada. IEEE, 2023:2153-2162

[5] Zhang G Q, Zhang Y Y, Zhang H W, et al. Learning dual attention enhancement feature for visible-infrared person re-identification[J]. Journal of Visual Communication and Image Representation, 2024, 99:104076

[6] Chen C Q, Ye M, Qi M B, et al. SketchTrans: disentangled prototype learning with transformer for sketch-photo recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5):2950-2964

[7] 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: a Publication of the IEEE Signal Processing Society, 2020, 29:5542-5556

[8] Wang Y H, Liu X H, Zhang P P, et al. TOP-ReID: multi-spectral object re-identification with token permutation [J]. arXiv e-Print, 2023, arXiv:2312.09612

[9] Ye M, Shen J B, Shao L. Visible-infrared person re-identification via homogeneous augmented tri-modal learning [J]. IEEE Transactions on Information Forensics and Security, 2021, 16:728-739

[10] Wei W Y, Yang W Z, Zuo E G, et al. Person re-identification based on deep learning: an overview[J]. Journal of Visual Communication and Image Representation, 2022, 82:103418

[11] Chen H, Lagadec B, Bremond F. Learning discriminative and generalizable representations by spatial-channel partition for person re-identification[C]//2020 IEEE Winter Conference on Applications of Computer Vision (WACV). March 1-5, 2020, Snowmass, CO, USA. IEEE, 2020:2472-2481

[12] Sun Y F, Zheng L, Yang Y, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 501-518

[13] Chen G Y, Lin C Z, Ren L L, et al. Self-critical attention learning for person re-identification[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). October 27-November 3, 2019, Seoul, Korea (South). IEEE, 2019:9636-9645

[14] Chen Y H, Zhang G Q, Zhang H W, et al. Multi-level part-aware feature disentangling for text-based person search[C]//2023 IEEE International Conference on Multimedia and Expo (ICME). July 10-14, 2023, Brisbane, Australia. IEEE, 2023:2801-2806

[15] Zhang G Q, Liu J, Chen Y H, et al. Multi-biometric unified network for cloth-changing person re-identification[J]. IEEE Transactions on Image Processing, 2023, 32:4555-4566

[16] Zhang G Q, Ge Y, Dong Z C, et al. Deep high-resolution representation learning for cross-resolution person re-identification[J]. IEEE Transactions on Image Processing, 2021, 30:8913-8925

[17] Zhang G Q, Zhang H W, Lin W S, 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

[18] Zhang G Q, Luo Z Y, Chen Y H, et al. Illumination unification for person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10):6766-6777

[19] Yi D, Lei Z, Liao S C, et al. Deep metric learning for person re-identification[C]//2014 22nd International Conference on Pattern Recognition. August 24-28, 2014, Stockholm, Sweden. IEEE, 2014:34-39

[20] Sarafianos N, Xu X, Kakadiaris I. Adversarial representation learning for text-to-image matching[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). October 27-November 3, 2019, Seoul, Korea (South). IEEE, 2019:5813-5823

[21] Sikdar A, Chowdhury A S. Scale-invariant batch-adaptive residual learning for person re-identification[J]. Pattern Recognition Letters, 2020, 129:279-286

[22] Zhang H W, Zhang G Q, Chen Y H, et al. Global relation-aware contrast learning for unsupervised person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(12):8599-610

[23] Wu A C, Zheng W S, Yu H X, et al. RGB-infrared cross-modality person re-identification[C]//2017 IEEE International Conference on Computer Vision (ICCV). October 22-29, 2017, Venice, Italy. IEEE, 2017:5390-5399

[24] Feng Z X, Lai J H, Xie X H. Learning modality-specific representations for visible-infrared person re-identification[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2019, 29:579-590

[25] Hao Y, Wang N N, Li J, et al. HSME: hypersphere manifold embedding for visible thermal person re-identification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1):8385-8392

[26] Zhang Z Y, Jiang S, Huang C, et al. RGB-IR cross-

modality person ReID based on teacher-student GAN model [J]. Pattern Recognition Letters, 2021, 150: 155-161

[27] Zhou J F,Huang B G,Fan W J,et al.Text-based person search via local-relational-global fine grained alignment [J].Knowledge-Based Systems,2023,262:110253

[28] 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.October 10-14,2022,Lisboa,Portugal.ACM,2022:5566-5574

[29] Pang L,Wang Y W,Song Y Z,et al.Cross-domain adversarial feature learning for sketch re-identification[C]//Proceedings of the 26th ACM International Conference on Multimedia.October 22-26,2018,Seoul,Republic of Korea.ACM,2018:609-617

[30] Zhai Y J,Zeng Y W,Cao D,et al.TriReID:towards multi-modal person re-identification via descriptive fusion model[C]//Proceedings of the 2022 International Conference on Multimedia Retrieval.June 27-30,2022,Newark,NJ,USA.ACM,2022:63-71

[31] Chen C Q,Ye M,Jiang D.Towards modality-agnostic person re-identification with descriptive query[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 18-22, 2023, Vancouver, BC,Canada.IEEE,2023:15128-15137

[32] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).June 27-30,2016,Las Vegas,NV,USA.IEEE, 2016:770-778

[33] Simonyan K,Zisserman A.Very deep convolutional networks for large-scale image recognition [J]. arXiv e-Print,2014,arXiv:1409.1556

[34] Krizhevsky A,Sutskever I,Hinton G E.ImageNet classification with deep convolutional neural networks[J].Communications of the ACM,2017,60(6):84-90

[35] Xiang X Z,Lv N,Yu Z T,et al.Cross-modality person re-identification based on dual-path multi-branch network [J].IEEE Sensors Journal,2019,19(23):11706-11713

[36] Zhang G Q,Zhang Y Y,Chen Y H,et al.Multi-granularity feature utilization network for cross-modality visible-infrared person re-identification[J].Soft Computing,2023:10:1-4

[37] Wang G A, Zhang T Z, Cheng J, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV).October 27-November 3, 2019, Seoul, Korea (South). IEEE, 2019: 3622-3631

[38] Dai H P,Xie Q,Ma Y C,et al.RGB-infrared person re-identification via image modality conversion[C]//2020 25th International Conference on Pattern Recognition (ICPR).January 10-15,2021,Milan,Italy.IEEE,2021: 592-598

[39] Yu H,Cheng X,Peng W,et al.Modality unifying network for visible-infrared person re-identification[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV).September 30-October 7,2023,Paris,France. IEEE,2023:11151-11161

[40] Ye M, Shen J B, Crandall D J, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[M]//Computer Vision-ECCV 2020. Cham:Springer International Publishing,2020:229-247

[41] Cheng D,Li X H,Qi M B,et al.Exploring cross-modality commonalities via dual-stream multi-branch network for infrared-visible person re-identification [J]. IEEE Access,2020,8:12824-12834

[42] Wei Z Y, Yang X, Wang N N, et al. Flexible body partition-based adversarial learning for visible infrared person re-identification[J].IEEE Transactions on Neural Networks and Learning Systems,2022,33(9):4676-4687

[43] Kim M, Kim S, Park J, et al. PartMix: regularization strategy to learn part discovery for visible-infrared person re-identification [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-22, 2023, Vancouver, Canada. IEEE, 2023: 18621-18632

[44] Wu Z S, Ye M. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).June 18-22, 2023, Vancouver, BC, Canada. IEEE, 2023: 9548-9558

[45] Zhang Y K, Lu Y, Yan Y, et al. Frequency domain nuances mining for visible-infrared person re-identification[J].arXiv e-Print,2024,arXiv:2401.02162

[46] Li Y L,Zhang T Z,Zhang Y D.Frequency domain modality-invariant feature learning for visible-infrared person re-identification [J]. arXiv e-Print, 2024, arXiv:2401.01839

[47] Hu W P, Liu B H, Zeng H T, et al. Adversarial decoupling and modality-invariant representation learning for visible-infrared person re-identification [J]. IEEE Transactions on Circuits and Systems for Video Technology,2022,32(8):5095-5109

[48] Lu Y,Wu Y,Liu B,et al.Cross-modality person re-identification with shared-specific feature transfer[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 16-18, 2020, Seattle, WA, USA.IEEE,2020:13376-13386

[49] Hu B Y,Liu J W,Zha Z J.Adversarial disentanglement and correlation network for rgb-infrared person re-identification[C]//2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, China. IEEE, 2021:1-6

[50] Feng Y J,Yu J,Chen F,et al.Visible-infrared person re-identification via cross-modality interaction transformer [J]. IEEE Transactions on Multimedia, 2023, 25: 7647-7659

[51] Zhang Q, Lai C Z, Liu J N, et al. FMCNet: feature-level modality compensation for visible-infrared person re-identification [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).June 21-24, 2022, New Orleans, LA, USA. IEEE, 2022: 7339-7348

[52] Yu Q,Liu F,Song Y Z,et al.Sketch me that shoe[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. June 26 - July 1, 2016, Las Vegas,Nevada,USA.IEEE,2016:799-807
[53] Song J F,Yu Q,Song Y Z,et al.Deep spatial-semantic attention for fine-grained sketch-based image retrieval[C]//2017 IEEE International Conference on Computer Vision (ICCV). October 22 - 29, 2017, Venice, Italy. IEEE,2017:5552-5561
[54] Pang K Y,Song Y Z,Xiang T,et al.Cross-domain generative learning for fine-grained sketch-based image retrieval[C]//The 28th British Machine Vision Conference.April 9-September 17,2017,London,UK.2017:1-12
[55] Gui S J, Zhu Y, Qin X X, et al. Learning multi-level domain invariant features for sketch re-identification[J]. Neurocomputing,2020,403:294-303
[56] Yang F, Wu Y, Wang Z, et al. Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval[J].IEEE Transactions on Multimedia,2021,23:2347-2360
[57] Lin K J,Wang Z X,Wang Z,et al.Beyond domain gap:exploiting subjectivity in sketch-based person retrieval[C]//Proceedings of the 31st ACM International Conference on Multimedia.October 29-November 3,2023,Ottawa,ON,Canada.ACM,2023:2078-2089
[58] Yang Q Z,Wu A C,Zheng W S.Person re-identification by contour sketch under moderate clothing change[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,43(6):2029-2046
[59] Chen Q S, Quan Z Z, Zhao K, et al. A cross-modality sketch person re-identification model based on cross-spectrum image generation[C]//International Forum on Digital TV and Wireless Multimedia Communications.December 9-10,2022,Singapore.Springer,2022:312-324
[60] Wang Z,Wang Z X,Zheng Y Q,et al.Beyond intra-modality:a survey of heterogeneous person re-identification[J].arXiv e-Print,2019,arXiv:1905. 10048
[61] Jiang K Z,Zhang T Z,Liu X,et al.Cross-modality transformer for visible-infrared person re-identification[C]//Computer Vision - ECCV 2022: 17th European Conference.October 23-27,2022,Tel Aviv,Israel.ACM,2022:480-496
[62] Zhang Y F,Wang Y Z,Li H F,et al.Cross-compatible embedding and semantic consistent feature construction for sketch re-identification[C]//Proceedings of the 30th ACM International Conference on Multimedia. October 10-14,2022,Lisboa,Portugal.ACM,2022:3347-3355
[63] Zhu F Y,Zhu Y,Jiang X B,et al.Cross-domain attention and center loss for sketch re-identification[J]. IEEE Transactions on Information Forensics and Security, 2022,17:3421-3432
[64] Liu X Y, Cheng X, Chen H Y, et al. Differentiable auxiliary learning for sketch re-identification[J]. Proceedings of the AAAI Conference on Artificial Intelligence,2024,38(4):3747-3755
[65] 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).March 12-15,2018,Lake Tahoe, NV,USA.IEEE,2018:1879-1887
[66] Devlin J,Chang M W,Lee K,et al.BERT:pre-training of deep bidirectional transformers for language understanding[J].arXiv e-Print,2018,arXiv:1810.04805
[67] Ding Z F,Ding C X,Shao Z Y,et al.Semantically self-aligned network for text-to-image part-aware person re-identification[J].arXiv e-Print,2021,arXiv:2107. 12666
[68] Wei D L,Zhang S P,Yang T,et al.Calibrating cross-modal features for text-based person searching[J].arXiv e-Print,2023,arXiv:2304. 02278
[69] Zhang Y,Lu H.Deep cross-modal projection learning for image-text matching[C]//Proceedings of the European Conference on Computer Vision (ECCV).Sepember 8-14,2018,Munich,Germany.Springer,2018:686-701
[70] 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
[71] Bird S.NLTK:the natural language toolkit[C]//Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. July 17 - 18, 2006, Sydney, Australia. ACM, 2006:69-72
[72] 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
[73] 姜定,叶茫.面向跨模态文本到图像行人重识别的Transformer网络[J].中国图象图形学报,2023,28(5):1384-1395
JIANG Ding, YE Mang. Transformer network for cross-modal text-to-image person re-identification[J]. Journal of Image and Graphics,2023,28(5):1384-1395
[74] Li S Y,Sun L,Li Q L.CLIP-ReID:exploiting vision-language model for image re-identification without concrete text labels[J].Proceedings of the AAAI Conference on Artificial Intelligence,2023,37(1):1405-1413
[75] Yan S L,Tang H,Zhang L Y,et al.Image-specific information suppression and implicit local alignment for text-based person search[J].IEEE Transactions on Neural Networks and Learning Systems,2023,PP(99):1-14
[76] Jiang D, Ye M. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-22,2023,Vancouver,Canada.IEEE,2023:2787-2797
[77] Gao L Y,Niu K,Jiao B L,et al.Addressing information inequality for text-based person search via pedestrian-centric visual denoising and bias-aware alignments[J]. IEEE Transactions on Circuits and Systems for Video Technology,2023,33(12):7884-7899
[78] He W Z,Deng Y H,Tang S X,et al.Instruct-ReID:a multi-purpose person re-identification task with instructions[J].arXiv e-Print,2023,arXiv:2306. 07520
[79] Wei X B,Song K C,Yang W K,et al.A visible-infrared clothes-changing dataset for person re-identification in natural scene[J].Neurocomputing,2024,569:127110
[80] Zhang L,Fu X W,Huang F X,et al.An open-world,diverse,cross-spatial-temporal benchmark for dynamic wild person re-identification [J]. arXiv e-Print, 2024,

arXiv:2403.15119
[81] Nguyen D T, Hong H G, Kim K W, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors,2017,17(3):605
[82] 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.October 20-24,2021,Virtual Event,China.ACM,2021:209-217
[83] 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.June 18-22,2018,Salt Lake City,UT,USA.IEEE,2018:79-88

Multi-modal person re-identification based on deep learning:a review

ZHANG Guoqing[1] YANG Shan[1] WANG Hairui[2] WANG Zhun[2] YANG Yan[1] ZHOU Jieqiong[1]
1 School of Computer Science,Nanjing University of Information Science & Technology,Nanjing 210044,China
2 School of Software,Nanjing University of Information Science & Technology,Nanjing 210044,China

Abstract Person re-identification (Re-ID),which involves retrieving the same person across cameras,is a key technology in the field of intelligent video surveillance.However,due to the complexity of surveillance scenarios,traditional single-modal approaches encounter limitations in extreme conditions such as low lighting and foggy days. Given the practical demands and the swift advancement in deep learning,multi-modal person Re-ID based on deep learning has received widespread attention.This article provides a review of the progress in multi-modal person Re-ID based on deep learning in recent years,elaborates on the shortcomings of traditional single-modal approaches and summarizes the common application scenarios and advantages of multi-modal person Re-ID,as well as the composition of various datasets.The article also highlights the relevant methods and classification of multi-modal person Re-ID across diverse scenarios,exploring current research hotspots and challenges.Finally,it discusses the future development trends and potential applications of multi-modal person Re-ID.

Key words deep learning; neural network; person re-identification (Re-ID); multi-modal