计算机科学
COMPUTER SCIENCE

跨模态目标重识别研究综述

崔振宇, 周嘉欢, 彭宇新

引用本文

崔振宇, 周嘉欢, 彭宇新. 跨模态目标重识别研究综述[J]. 计算机科学, 2024, 51(1): 13-25.

CUI Zhenyu, ZHOU Jiahuan, PENG Yuxin. Survey on Cross-modality Object Re-identification Research [J]. Computer Science, 2024, 51(1): 13-25.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.yg20240103

跨模态目标重识别研究综述

崔振宇　周嘉欢　彭宇新
北京大学王选计算机研究所　北京 100871
多媒体信息处理国家重点实验室　北京 100871
(cuizhenyu@stu.pku.edu.cn)

摘　要　目标重识别(ReID)技术旨在匹配不同区域摄像头在不同时间拍摄到的同一目标,其核心是通过目标间的细粒度差异实现不同目标的有效区分。因此,目标重识别技术被广泛应用于安防布控、刑侦监控等领域并发挥了重要作用。传统的目标重识别技术通常适用于光照条件良好情况下的可见光模态数据,但在处理黑夜低光照条件下的目标重识别任务时,其性能通常受到严重限制。红外摄像机因其卓越的夜视性能,通常被应用于在低光照条件下采集目标红外图像。因此,跨模态目标重识别技术旨在通过可见光图像匹配红外图像,实现全天候不间断的目标重识别。近年来,跨模态目标重识别技术取得了很大进展,然而,对于现有模型的归纳总结及深入分析仍然欠缺。为此,对跨模态目标重识别领域的相关研究和新颖方法进行了深入调研和总结,讨论了现有方法在实际场景中面临的挑战,并从模型分类和模型评价两个方面对现有方法进行归纳与分析。首先,围绕跨模态目标重识别问题的研究难点,将跨模态目标重识别分为生成式方法和非生成式方法两大类;然后,对当前跨模态重识别领域中广泛使用的评测数据集以及相关评价指标进行了综述与总结;最后,讨论了跨模态重识别领域仍然存在的挑战并对未来发展趋势进行了展望。

关键词:计算机视觉;目标重识别;跨模态;细粒度特征;表征学习
中图分类号　TP391

Survey on Cross-modality Object Re-identification Research

CUI Zhenyu,ZHOU Jiahuan and PENG Yuxin
Wangxuan Institute of Computer Technology,Peking University,Beijing 100871,China
National Key Laboratory for Multimedia Information Processing,Peking University,Beijing 100871,China

Abstract　Object re-identification(ReID) technology aims to match the same object captured by cameras across different areas at different time. The key is to distinguish different objects through fine-grained differences between different individuals,which is widely used in security control,criminal investigation and monitoring,etc. Traditional ReID technology is usually suitable for visible cameras with good lighting conditions,but its performance is severely limited under low-light conditions. The infrared camera is often used to collect infrared images of objects under low light conditions due to its outstanding night vision performance. Therefore,cross-modality object re-identification technology focuses on achieving uninterrupted object ReID across day and night from visible images to infrared images(VI-ReID),and vice versa. In recent years,VI-ReID technology has made significant progress. However,a comprehensive summary and in-depth analysis of existing models are still lacking. To this end,this paper conducts an in-depth investigation and summary of relevant research and novel methods in the field of VI-ReID. It discusses the challenges faced by existing methods in actual scenarios,and categorizes them from two aspects:model classification and model evaluation. First,focusing on the research challenges,VI-ReID is categorized into generative methods and non-generative methods. Secondly,the evaluation datasets and evaluation metrics are reviewed and summarized. Finally,the remaining challenges in VI-ReID are discussed and the future development trends are prospected.

Keywords　Computer vision,Object re-identification,Cross-modality,Fine-grained feature,Representation learning

到稿日期:2023-10-12　返修日期:2023-12-01
基金项目:国家自然科学基金(61925201,62132001)
This work was supported by the National Natural Science Foundation of China(61925201,62132001).
通信作者:彭宇新(pengyuxin@pku.edu.cn)

1 引言

随着计算机视觉领域的快速发展，视频监控技术被越来越广泛地应用于智能安防系统，并逐渐成为智慧城市发展中的重要手段之一[1-2]。同时，随着城市面积的不断扩大，成千上万的监控设备时刻记录着海量的目标数据，仅依靠人工手段无法对采集到的海量数据进行有效的目标分析[3]。为了满足日益增长的目标智能分析需求，目标重识别技术(ReID)应运而生。

如图 1 所示，在一个多摄像头分布网络[4]（见图 1(a)）中，目标重识别技术旨在利用目标的视觉信息来匹配不同地点、不同时间、不同摄像头所拍摄的同一目标(见图 1(b))。

(a) DukeMTMC[4] 相机分布网络

(b) 目标重识别任务示意图

图 1 目标重识别任务示意图(以行人为例)

Fig. 1 Illustration of ReID task(taking human as an example)

目标重识别模型通过给定的查询图像来检索数据库中的图像，从而确定查询图像中目标的身份，实现监控场景中目标的匹配与关联[5]。其核心挑战在于如何应对成像质量差异、环境视角与光照变化、目标的姿态变化，以及遮挡等信息带来的干扰，实现对目标鉴别性信息的准确提取。

近年来，深度学习技术在各个研究领域的广泛应用对 ReID 领域的发展产生了深远影响[6-9]。大量目标重识别算法[10-21]利用深度神经网络取得了很大的性能提升。然而，上述方法中绝大部分仅能在白天或光照条件良好的情况下进行目标重识别。如图 2 所示，一般常用的监控摄像头仅能采集目标的可见光模态信息，在低光照环境下，例如夜间场景，很难采集到可用的目标鉴别性信息，从而导致目标重识别算法失效。为了在低光照环境下有效地捕获目标特征，红外相机被广泛应用于弱光照条件下的图像采集[22-23]。此方法也为目标重识别算法能够跨越白天和黑夜匹配同一目标提供了可能。

为了实现跨时段、全天候的目标重识别，最新的研究开始关注从可见光图像到红外图像的跨模态目标重识别(VI-ReID)问题[24-25]。

具体而言，VI-ReID 旨在利用可见光图像匹配红外图像，实现全天候不间断的目标重识别。其核心难点在于如何克服不同模态间的差异，从而提取到目标自身的鉴别性信息。尽管近年来 VI-ReID 领域发展迅速，但仍缺少对现有 VI-ReID 方法的全面综述。本文通过梳理现有 VI-ReID 的相关理论与方法，对现有工作进行了分析、总结与展望。首先，本文对现有 VI-ReID 方法进行了分类讨论，根据算法设计理念的不同，将现有研究方法分为生成式方法[27-45]和非生成式方法[46-87]两大类。对于生成式方法，本文将其进一步划分为生成式对抗学习方法[27-35]和跨模态数据增强方法[36-45]；对于后者，本文重点综述了基于表征学习[46-74]和基于度量学习[75-87]的两类方法。其次，对 VI-ReID 领域中广泛使用的评测数据集及相关评价指标进行了归纳总结。最后，讨论了 VI-ReID 领域仍然存在的挑战性问题并对未来发展趋势进行了展望。

图 2 不同摄像头于白天/夜晚采集的图像示意图[26]

Fig. 2 Sample images collected by different cameras during the day/night[26]

2 跨模态目标重识别方法

问题定义：跨模态目标重识别旨在匹配可见光模态与红外模态图像中的同一目标。形式化如下：设$\{g_i\}_{i=1}^{N}$为一个包含L个目标共计N幅图像的数据库；$\{q_k\}_{k=1}^{M}$为包含M幅查询图像的查询集，其中g_i和q_k分别来自不同的模态。设一个跨模态目标重识别(VI-ReID)模型为$\phi(\cdot;\theta)$，VI-ReID的目的是确定查询图像q_k的身份q_k^*，即：

$$q_k^* = \arg\min_{i=1,\cdots,N} d(\phi(q_k;\theta),\phi(g_i;\theta)) \tag{1}$$

其中，$d(\cdot;\cdot)$表示通过两个表征向量间距离计算得到的相似度函数。

借鉴传统目标ReID方法，早期的VI-ReID方法[27-45]尝试利用图像生成技术将可见光图像和红外图像在数据层面进行转换，从而将跨模态目标重识别问题转化为单模态目标重识别问题，此类方法被称为生成式方法。然而，由于缺少不同模态图像间像素级别的对应关系作为生成算法的约束条件，此类方法生成的图像通常存在质量差、鉴别性低的问题。针对上述问题，近年来绝大部分VI-ReID工作尝试借助神经网络的特征提取能力，直接对齐可见光图像和红外图像[46-87]，这类方法被称为非生成式方法，其通常由表征学习方法和度量学习方法两部分组成。前者致力于如何设计更适合抽取不同模态表征的神经网络[46-74]；后者则致力于在训练阶段为神经网络设计合适的目标函数，从而提升属于同一目标表征的鉴别性和一致性[75-87]。

接下来，本章按上述分类标准详细回顾了VI-ReID领域的相关方法，本文总结的方法及其对应类型如图3所示，上述方法的核心思想、优点和不足如表1所列。

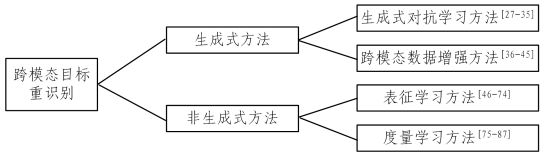

图3 对现有VI-ReID方法的分类及对应方法

Fig. 3 Categorization and corresponding methods of existing VI-ReID methods

表1 不同类型跨模态重识别方法的核心思想、优点和不足

Table1 Core methodologies, advantages and disadvantages of different cross-modality re-identification methods

方法	核心思想	优点	不足
生成式对抗学习方法[27-35]	将可见光图像和红外图像相互翻译，促进数据层面的跨模态信息对齐	方法可解释性强，便于理解	生成图像的鉴别性差，建模所需开销大
跨模态数据增强方法[36-45]	直接增强原始可见光或红外图像，促使模型适应模态的变化	方法简洁直观，生成图像鉴别性强	受限于图像自身质量，稳定性较差
表征学习方法[46-74]	设计适合抽取跨模态表征的神经网络，促进表征层面的目标特征对齐	特征提取鲁棒性强，表征能力强	忽略了不同目标的差异性，难以区分相似外观的不同目标
度量学习方法[75-87]	设计用于模型训练的目标函数，促进目标特征的鉴别性和一致性	鲁棒性强，灵活性高	引入额外超参，增加了模型的训练复杂度与开销

2.1 生成式方法

考虑到可见光模态和红外模态间显著的跨模态差异，生成式VI-ReID方法[27-45]利用图像生成技术生成不同模态的图像，从而将VI-ReID转化为单模态ReID任务。本文将从生成式对抗学习和跨模态数据增强两个方向讨论现有工作。

2.1.1 生成式对抗学习方法

如图4所示，早期的VI-ReID模型通常利用生成式对抗网络(Generative Adversarial Network，GAN)[88]将可见光图像和红外图像相互翻译，从而在数据层面实现跨模态信息对齐[27-35]。

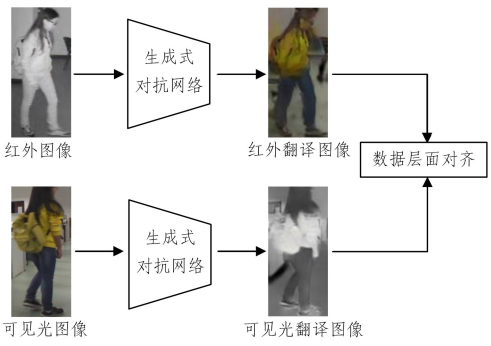

图4 生成式对抗学习方法示意图

Fig. 4 Schematic diagram of generative adversarial network-based methods

Wang等[27]提出了一种基于GAN的VI-ReID方法，构建了一个由真实红外图像和伪造红外图像组成的共享红外域，并将彩色图像翻译成伪造红外图像，最后在统一的共享红外域内提取模态不变特征，用于模型训练和推理。为了利用可见光-红外跨模态数据在物理层面的关联，Wang等[28]提出了一种基于多光谱映射的VI-ReID方法，将可见光图像和红外图像映射到一个多光谱域进行模型学习，缓解了由模态差异导致的辨识性特征提取效率低下的问题。然而，上述方法忽略了跨模态数据中所蕴含知识对图像生成的促进作用，因此Zhang等[29]提出了一种基于“教师-学生”结构的GAN模型，利用红外图像训练教师网络并将其用于指导学生网络提取图像的可见光信息，从而引导生成网络更好地利用跨模态知识生成对应模态的训练数据。然而，上述方法在实现图像生成时忽略了目标鉴别性信息和特定模态信息通常紧密纠缠在一起这一问题，不可避免地引入了特定模态的干扰信息，导致生成的图像目标质量低下，从而影响了跨模态目标重识别性能。

为了在图像生成的过程中避免特定模态信息的干扰，Choi等[30]提出了一种分层跨模态解纠缠生成模型，通过生成具有不同姿势和光照变化的不同模态图像来学习目标的解纠缠表征，并显式地提取可见光和红外图像的公共特征，增强了图像生成的鲁棒性。针对上述方法忽略了红外图像和可见光图像间由风格差异导致的生成图像质量低下的问题，

Wang 等[31]提出了一种基于循环对抗生成网络(CycleGAN)[89]的图像生成方法,通过分离模态特定信息和跨模态不变信息实现了生成数据在语义层面的对齐,其优势在于可以显式地剥离特定模态特征,从而更好地抑制模态变化对模型重识别性能的影响。然而,由于成像原理不同,可见光图像和红外图像所表达的信息存在严重差异,上述方法通常难以填补生成图像与真实图像之间的领域鸿沟,从而弱化了生成图像中目标的鉴别性信息。

为了弥补生成的可见光图像和真实可见光图像间的领域差异,Liu 等[32]提出了一种基于灰度对齐的图像生成方法。首先利用可见光图像生成灰度图像,然后利用风格迁移模型将红外图像同时转换成对应的灰度图像,通过在灰度空间的对齐降低了可见光和红外模态间的差异。为了让生成模型在图像翻译的过程中保留只存在于特定模态的鉴别性信息,Qi 等[33]提出了一种基于生成图像融合的 VI-ReID 方法,通过对比学习生成跨模态配对图像,并设计了一种基于部件的多模态特征融合模块来整合并利用两种模态的信息,提升了生成式方法的生成效果。Liu 等[34]提出了一种基于跳跃连接的生成式对抗网络,该网络通过基于注意力机制的特征融合策略,更好地利用了输入图像来提升生成图像的质量。Wei 等[35]提出了一种双向图像翻译网络,旨在通过可见光-红外图像的交互映射来促使模型生成更加逼真的图像。

综上所述,尽管基于生成式对抗学习的 VI-ReID 方法可以直观地缩小跨模态图像间的差异,但由于缺少不同模态图像间像素级别的对应关系作为生成算法的约束条件,此类方法的生成结果通常是不可靠的,其不可避免地会引入噪声信息,从而导致鉴别性特征的丢失,进而致使模型无法准确匹配同一目标。

2.1.2 跨模态数据增强方法

为了解决生成式对抗学习方法面临的难题,跨模态数据增强方法旨在直接增强原始可见光或红外图像,促使模型主动适应模态的变化。如图 5 所示,此类方法利用生成的增强后图像训练 VI-ReID 模型,有效缓解了图像翻译引入的额外噪声导致的鉴别性特征丢失的问题。

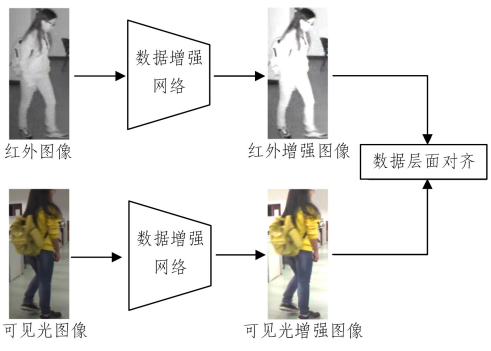

图 5 跨模态数据增强方法示意图

Fig. 5 Schematic diagram of cross-modality data augmentation methods

Li 等[37]提出了一种基于中间模态的数据增强方法。在可见光和红外模态的基础上引入了一个中间模态,然后通过对可见光图像的采样生成该中间模态的图像,并利用中间模态和原始模态的图像联合训练模型,从而弥补可见光图像和红外图像间的差距,提升了 VI-ReID 的性能。Ye 等[38]提出了一种基于通道随机拆分的数据增强方法,随机选取可见光图像的某个通道作为网络输入,并与灰度图像进行联合训练,提升了网络对可见光与灰度图像的表征能力。在此基础上,Ye 等[39]还提出了一种基于随机通道叠加的数据增强方法,从可见光图像中进行随机通道抽取,得到一种辅助灰度模态用于网络训练,该辅助模态在保留可见光结构信息的同时,又得到了近似红外图像的样式。Lu 等[40]提出了一种基于 Transformer 网络[90]的 VI-ReID 方法,将可见光图像所对应的灰度图像以图像块的形式输入 Transformer 网络,并借助自注意力机制捕获全局鉴别性信息,最终提升了 VI-ReID 的性能。然而,上述策略均只利用了可见光图像进行数据增强,缺乏对红外图像的有效利用,因此性能通常受到严重限制。

最近的研究表明,同时对可见光图像和红外图像进行双向增强可以显著提升 VI-ReID 性能。考虑到红外图像缺少丰富的颜色信息,Basaran 等[41]提出了一种基于灰度图的图像互补增强方法,将灰度图像分别与可见光图像和红外图像在通道层面进行堆叠,并送入不同的网络分支中学习局部区域的辨识性特征,丰富了生成图像的多样性。为了改善辅助模态的图像质量,Huang 等[42]提出了一种模态自适应混合策略,在可见光和红外图像之间动态生成合适的中间模态图像,通过学习跨模态图像不同区域间的一致性,动态学习得到局部自适应的线性插值策略,减小了跨模态图像之间的差异。Kim 等[43]通过混合跨模态图像中的局部视觉描述子来合成增强样本,该样本综合了来自相同/不同身份目标的正/负样本,利用所设计的基于熵的样本挖掘策略来缓解潜在错误正/负样本带来的不利影响。Lu 等[44]设计了一种基于风格转换的图像增强策略,通过生成具有一致风格的图像缩小了跨模态图像间的风格差异。在实现跨模态图像增强的同时,降低了图像增强的计算开销。

综上所述,生成式方法通常在数据层面对齐不同模态的图像,从而减小模态差异。然而,其通常会在原本 VI-ReID 模型的基础上引入一个额外的生成网络或模块,给 ReID 模型带来额外的计算开销。此外,在实际场景中,因受到数据采集、存储等因素的影响,不同模态数据质量难以保证,也会进一步限制生成模型的生成质量,导致难以有效对齐不同模态的数据,限制了 VI-ReID 的性能表现以及算法的稳定性。

2.2 非生成式方法

针对生成式方法存在的上述问题,最新的研究工作尝试利用神经网络强大的特征提取能力,直接在特征层面对齐不同模态的图像。此类方法被通常遵循以下范式:首先利用神经网络提取不同模态图像的深度表征;然后利用度量学习对齐上述表征,从而实现可见光图像与红外图像的相互匹配。此类方法称为非生成式方法[46-87]。本文从表征学习和度量学习两个角度来综述现有工作。

2.2.1 表征学习方法

表征学习方法旨在针对性地设计网络结构,增强跨模态

图像匹配的鉴别性特征提取。根据所提取的鉴别性特征类型差异，现有方法主要可分为跨模态统一表征学习和跨模态特定表征学习。

1）跨模态统一表征学习

如图 6 所示，跨模态统一表征学习方法通过提取可见光模态和红外模态中的共有特征，去除只存在于特定模态中的信息干扰，从而降低统一表征学习过程被特定模态信息影响的风险。

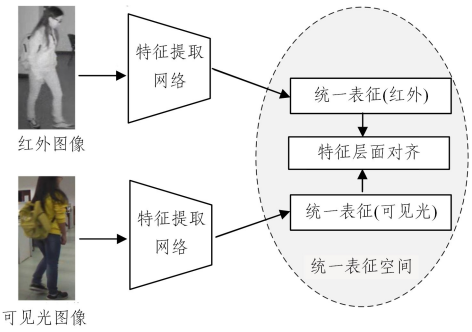

图 6 跨模态统一表征方法示意图

Fig. 6 Schematic diagram of cross-modality representation learning methods

早期的跨模态统一表征学习方法大多设计不同的神经网络结构来提取不同模态的公共信息。Ye 等[46]提出使用卷积神经网络(CNN)[91]来提取跨模态图像的公共信息，并设计了一种可泛化的平均池化层来捕获不同模态共有的判别特征。同一时期，Ye 等[47]还提出了一种基于图卷积神经网络(GNN)[92]的 VI-ReID 模型，利用图神经网络挖掘模态内局部特征和跨模态上下文特征，并提取最具鉴别性的局部特征进行聚合，增强了模型对噪声样本的鲁棒性。Liang 等[48]提出了一种基于 Transformer 网络[90]的跨模态统一表征学习网络，该网络通过显式地挖掘不同模态的公共信息，利用模态感知增强损失进一步增强了跨模态嵌入表示的鲁棒性。类似地，Chai 等[49]提出了一种基于 Transformer 网络的 VI-ReID 模型，引入了一种基于特征间最大平均差异的损失函数来处理表征分布间的跨模态差异，在减小跨模态差异的同时，促进了跨模态统一表征的学习。受到神经网络自动搜索架构的启发，Chen 等[50]提出了一种基于神经网络架构搜索的 VI-ReID 方法，设计了一种基于对称搜索的可微分特征搜索空间，可以同时将粗粒度通道信息和细粒度空间信息相结合，从而自适应地过滤背景噪声，提升了 VI-ReID 性能。为了适配跨模态视频数据，Lin 等[51]首先将不同模态的时序信息映射到一个公共空间，然后通过时序特征记忆网络从旧的视频序列中抽取具有运动不变性的特征，从而获得目标的鉴别性信息。上述以神经网络结构设计为主的统一表征学习方法在一定程度上推动了 VI-ReID 的发展，但此类方法在提取目标自身信息的同时，不可避免地会受到不同模态图像内所存在的噪声信息的干扰，抑制了目标鉴别性信息的有效提取。

针对上述难题，部分方法通过知识蒸馏[93]或注意力机制[94]筛选并过滤网络提取的原始特征，旨在筛选出所包含的有用信息，实现统一表征的建模。Tian 等[52]提出了一种基于变分自蒸馏学习的统一表征学习方法，通过变分交叉蒸馏和变分交互学习两种策略来消除特定视角信息和任务无关信息，增强了统一表征和标签信息间的内在关联，从而提高了统一表征对目标视角变化的鲁棒性。考虑到不同类别目标的统一表征中心通常包含丰富的语义信息，Cheng 等[53]提出了一种基于语义中心蒸馏的表征记忆网络，利用模态无关的语义中心进行簇间对比学习，并通过记忆的语义簇和新学到的语义簇间的蒸馏学习，提升了跨模态数据的内在关联。除了上述基于蒸馏的方法，跨模态注意力学习被认为是获取跨模态数据间一致性信息的重要策略之一。Zhang 等[54]提出了一种基于跨模态注意力学习的双重交互学习框架，该框架构建了一个跨模态特征提取分支，包含一系列由跨模态注意力机制得到的不同尺度特征图。利用上述特征图指导两个单模态特征提取分支的学习，提升了模态内统一表征的鉴别性。Feng 等[55]提出了一种基于 Transformer 网络的跨模态交互学习框架，该框架利用跨模态注意力机制提取每个图像块的表征，并与另一个模态的图像块进行注意力交互学习，提升了同一目标的类内紧凑程度和不同目标的类间分离程度。与此同时，Li 等[56]提出使用跨模态注意力机制构建一个中介分支，并利用该中介分支得到的特征，结合双向时空聚合模块，挖掘视频数据中蕴含的时空信息，减轻了噪声图像帧对最终预测结果的影响，提升了模型在不同重识别场景中的泛化性。考虑到上述基于知识蒸馏和基于注意力学习两种策略的各自优势，Wu 等[57]提出了一种结合自注意力学习和蒸馏学习的 VI-ReID 方法，首先利用自注意力学习挖掘图像中不同空间区域的特征表示，然后利用单个模态分支输出的目标分类结果指导另一个分支的学习，显著提升了跨模态图像间的类内一致性。不同于上述策略，Li 等[58]引入了反事实推理来挖掘跨模态表征中的公共信息，设计了一组同模态和跨模态特征转移模块，通过模拟检索时查询图像和数据库中图像的不平衡场景，来缓解模型测试期间由查询样本和数据库中样本数量不均衡导致的跨模态差异。此外，此方法还提出了一种对照关系干预和因果推理策略来强化样本间的拓扑结构，最终提升了同类样本间的关联的可靠性。遗憾的是，尽管上述策略充分挖掘了不同模态图像的内在信息，但却没有结合目标的先验信息。例如对于行人而言，显式地区分出不同行人的关键部位(如脸部、腿部等)，能够更准确地进行特征比对，从而提升重识别的准确性。

基于上述考虑，现有方法尝试引入外部先验知识来增强跨模态统一表征的建模能力。Zheng 等[59]引入了目标的属性信息来引导模型生成鉴别性的局部信息，同时设计了一种渐进式属性嵌入网络，通过联合属性和目标分类损失，促进网络学习与模态无关且具有局部鉴别性的表征。类似地，Feng 等[60]引入目标的轮廓信息来提升表征的辨识性，通过在两个正交子空间中去相关模态共享表征，最大化目标轮廓特征和目标类别特征间的互信息，从而增强了表征学习的多样性。为了降低模型对外部信息的依赖，Alehdaghi 等[61]将可见光图像和红外图像视为彼此的先验知识，通过构建的中间域为不同模态各自的分支提供外部监督信息，提升了 VI-ReID 的

鲁棒性。Wu 等[62]则直接对单模态特征提取网络得到的深度表征进行跨模态的转移，并将转移后的深度表征作为外部信息用于提升另一模态分支输出表征的鉴别性，实现了不同模态间表征的关联。

尽管上述跨模态统一表征学习方法可以一定程度上对齐不同模态图像的表征，但同时也削弱了仅存在于特定模态数据中的目标鉴别性信息，如颜色或图案信息。因此，跨模态统一表征学习方法在复杂场景中 VI-ReID 的鲁棒性通常有限。

2)跨模态特定表征学习

如图 7 所示，跨模态特定表征学习方法通常在统一表征的基础上，提取可见光模态和红外模态中特有的鉴别性表征来区分不同目标，从而恢复在统一表征学习中被抑制的特定模态内的鉴别性信息。

其中，一部分工作在提取不同模态特定表征后直接将它们在特征层面对齐，从而辅助目标重识别。Li 等[63]提出了一个可用于跨模态车辆重识别框架，该框架对不同模态的图像分别进行特征提取，然后基于特定频谱的视觉特征对同一目标表征进行对齐。Ye 等[64]提出了一种特定模态表征感知网络，在特征建模层面对不同模态采用具有不同参数的神经网络来建模跨模态差异，并在语义分类层面采用模态特定的分类器来建模相关模态的鉴别性信息，提升了 VI-ReID 的性能。Zhang 等[65]提了一种双路表征学习框架，该框架设计了一种空间结构保持网络将跨模态图像嵌入到公共三维表征空间中，通过设计的相关性对比网络从跨模态图像中提取对比特征，提升了跨模态表征学习的鲁棒性。Hu 等[66]提出了一种基于对抗解耦的跨模态表征学习方法，通过解耦特定模态内的目标相关信息和模态相关信息实现特征解耦，最终通过对抗学习策略抑制模态相关信息的表达，丰富了跨模态特定表征中的目标信息。Zhang 等[67]提出了一种跨模态语义一致性网络，该网络同时从细粒度通道语义和全局模态语义两方面优化跨模态特定表征的一致性，提升了同一目标所属的跨模态图像间的相关性。Wu 等[68]考虑到不同模态图像局部信息间存在的差异，为每个模态学习一组特定的局部表示，并利用该表示作为注意力区域抽取局部特征用于跨模态图像间对齐，提升了跨模态图像间潜在的语义关联性。然而，上述方法没有考虑到跨模态特定表征和跨模态统一表征间的内在关联，导致上述表征间信息存在冗余，限制了特定表征的鉴别性。

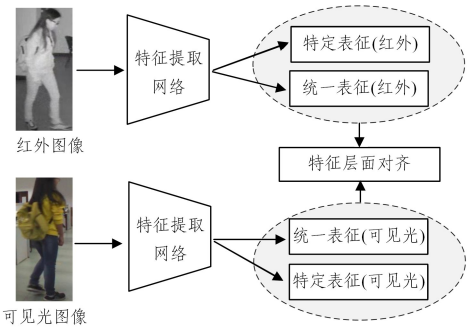

图 7 跨模态特定表征方法示意图

Fig. 7 Schematic diagram of cross-modality specific representation learning methods

针对上述问题，一些工作通过先将特定模态原始表征转换到另一模态中，再在该模态内对齐本模态特定表征和转移后特定表征，从而降低模型对齐不同模态特征的难度。Wu 等[69]提出了一种基于特定模态特征转换的循环构造网络，该网络包含一个轻量级知识获取模块，可以从模态特定表征的映射学习中捕获丰富的语义信息，并结合该语义信息指导特征转换的过程，提升了跨模态特定表征学习的鲁棒性。Pu 等[70]提出了一种基于双高斯变分自编码器的 VI-ReID 方法，根据混合高斯先验和标准高斯分布先验分别解耦不同模态的特征子空间，然后通过解纠缠转换后的跨模态鉴别性特征提高了复杂场景中目标匹配的鲁棒性。Jiang 等[71]提出了一种基于 Transformer 的跨模态表征转移网络，根据查询图像的特定表征信息，自适应地调整数据库中图像特定表征信息的均值与方差，从而实现特定模态信息的对齐。然而，这类方法通常无法解释转移后的表征中包含的信息，不可避免地会导致转移后的目标鉴别性信息的丢失，从而导致 VI-ReID 性能下降。

针对上述问题，基于特定模态特征补偿的方法通过结合跨模态统一表征和跨模态特定表征实现了更鲁棒的特征表示。Li 等[72]提出了一种特定模态表征记忆网络来补偿缺失的模态信息，并通过特征一致性、记忆代表性和结构对齐 3 种学习策略，促使记忆网络学习互补的特征表示，从而提升了 VI-ReID 性能。Lu 等[73]提出了一种基于特定-共享表征的互补学习框架，该框架首先根据模态共享表征对不同模态样本间的相似度进行关联建模，然后利用特定模态表征进一步检索潜在的正样本，显著提升了 VI-ReID 的准确性。受此启发，Zhang 等[74]提出了一种特征层面的跨模态补偿学习方法，通过对抗学习生成特定模态缺失的表征，然后将该表征与原始表征相结合进行 VI-ReID，显著提升了跨模态表征的多样性和完整性。

跨模态表征学习通常关注提取目标的外观表征，但忽略了同一目标间的相似性以及不同目标间的差异性，易导致模型难以区分具有相似外观的不同目标。

2.2.2 度量学习方法

除了表征学习方法之外，另一类重要的非生成式方法是度量学习方法。度量学习通过对比和优化不同样本表征间的距离，从而提升跨模态表征学习的有效性和鲁棒性。根据样本表征粒度的不同，现有方法主要分为基于样本个体表征的方法和基于类别中心表征的方法。

1)基于样本个体表征的度量学习

基于样本个体表征的度量学习方法通过选择并优化不同模态中的样本个体表征间的距离，针对性地缩小表征间的跨模态差异。Zhao 等[75]提出了一种五元组度量学习方法，对于每个样本而言，从两个模态的样本中分别挑选出最难优化的正/负样本形成五元组，并利用排序损失(Ranking Loss)进行模型优化，提升了度量学习算法的稳定性。类似地，Ye 等[76]提出了一种对称排序度量学习方法，从不同模态中挑选出两组相互对称的样本对，然后以排序损失联合优化上述样本对，从而缩小了不同模态间样本的差异。Jia 等[77]提出了一种

基样本间相似性推理的度量损失，该损失利用模态内样本间相似性来缓解跨模态差异，提升了跨模态图像间近邻样本的匹配程度。Kamenou 等[78]提出了一种基于模态内一致性约束的度量学习方法，首先利用三元组损失优化不同模态内的难例样本，再对齐两个模态间的其余样本，从而提升了同一类型样本表征的类内一致性。尽管上述方法可以一定程度提升样本表征间的相关性，但忽略了同一类别样本的表征中心对不同模态样本的高层次语义约束能力。

2）基于类别中心表征的度量学习

基于类别中心表征的度量学习方法通过计算并优化样本及其中心表征间的距离，使不同模态样本的表征在语义层面彼此接近，从而提升同一类别样本表征的类内一致性和类间鉴别性。

Zhang 等[79]提出了一种渐进式度量学习方法，首先将同一目标的表征聚类为多个子簇，然后将多个子簇进一步聚类为目标簇，最后通过对比学习区分不同目标所属的聚类簇，提升了度量学习的稳定性。类似地，Ye 等[80]结合样本的模态标签信息，提出了一种双阶段度量学习方法。先将同一目标的表征按不同模态进行语义中心聚类，然后利用对比学习优化同一目标的表征中心间的距离，增强了同一样本表征的内在语义关联。然而，上述多阶段度量学习策略会导致跨模态差异信息的不断累积，从而削弱了模型的度量学习能力。针对上述问题，Yu 等[81]提出了一种基于多模态语义中心对齐的度量学习方法，首先对可见光、红外及其混合模态中的样本以及对应的语义中心进行对齐，然后拉近上述模态的语义中心，从而更好地缩小了模态间的领域差异。为了进一步提升不同模态目标的鉴别性，Liu 等[82]提出了一种基于异构语义中心对齐的度量学习方法，直接促使不同目标在不同模态中的语义中心进行分离，进一步提升了不同模态目标表征的鉴别性。然而，表征学习方法通常会引入额外的超参数，增加了模型训练时的复杂度和开销。

除此之外，还有小部分方法利用其他度量学习方法对 VI-ReID 进行改进与优化。Dai 等[83]提出了一种基于对抗学习的跨模态度量学习方法，首先训练一个能区分不同模态的二分类器，迫使模型输出的结果迷惑该分类器，使其输出错误结果，从而缩小了同一类别目标样本间的跨模态差异。

3 数据集与评价指标

为了评估不同跨模态目标重识别模型的性能，高质量的公共数据集和统一的评价指标必不可少。本章介绍了现有跨模态目标重识别数据集以及对应的评价指标。

3.1 数据集介绍

近年来，随着跨模态目标重识别领域逐渐受到关注，出现了多个大型公开 VI-ReID 数据集。

RegDB 数据集[85]采集于韩国东国大学，该数据集由 1 台可见光摄像机和 1 台红外摄像机采集，包含 412 个行人共计 8240 幅图像，其中，训练集和测试集各包含 206 个行人共计4120 幅图像。测试时，包含可见光到红外图像搜索（V-I）和红外到可见光图像搜索（I-V）两种设定。其数据集样例如图 8 所示。该数据集是最早的跨模态目标重识别数据集之一，然而，由于其数据量有限且行人目标姿态变化较少，难以适用于全面评估不同 VI-ReID 算法的性能优劣。

SYSU-MM01 数据集[34]采集于中山大学。该数据集由 4 台可见光摄像机和 2 台红外摄像机进行数据采集，共包含 491 个行人共计 38271 幅图像。其中，训练集包含 395 个行人共计 34167 幅图像，测试集包含 96 个行人共计 4104 幅图像。其中摄像机 1，2，4，5 采集了可见光场景下的图像，摄像机 3 和 6 采集了红外场景下的图像。测试时，包含全部搜索模式和室内搜索模式。在此基础上，又进一步分为单图搜索（Single-shot）和多图搜索（Multi-shot）两种设定。其中，前者指数据库中每个行人仅包含一张待检索图像，后者指数据库中每个行人包含 10 张图像。其样例如图 9 所示。该数据集是目前最常用的 VI-ReID 算法评估数据集，具有数据量大、行人目标姿态丰富等特点。然而，收集该数据集时所采用的摄像机数量较少，因此无法全面反映真实条件下不同采集设备对 VI-ReID 性能的影响。

图 8 RegDB 数据集中图片样例

Fig. 8 Sample images in RegDB dataset

图 9 SYSU-MM01 数据集中图片样例

Fig. 9 Sample images in SYSU-MM01 dataset

HITSZ-VCM 数据集[51]是一个用于视频 VI-ReID 的数据集，由 12 台可自由切换的可见光摄像机和红外摄像机采集，包含 927 个行人共计 21863 个片段和 463259 幅图像。其

中，训练集和测试集分别包含 11 061 个片段和 10 802 个片段。测试时，包含可见光到红外图像搜索(V-I)和红外到可见光图像搜索(I-V)两种模式。但是，该数据集仅包含由 12 台设备采集的 12 个场景，因此其多样性有限。

RGBN300 数据集[63]是一个用于车辆 VI-ReID 的数据集，由 8 台摄像机采集，包含 300 辆车辆共计 100 250 幅图像。其中，训练集包含 150 辆车辆共计 25 200 幅图像，测试集包含 150 辆车辆共计 24 925 幅图像。然而，受限于车辆目标类型较少，该数据集在大规模检索场景中的实用性较差。

综上所述，现有 VI-ReID 数据集存在样本数量有限、场景多样性差、采集设备少等问题。因此，亟需提出更完备的数据集来促进 VI-ReID 方法评价体系的构建。

3.2 评价指标介绍

在评价指标方面，现有方法主要采用累计匹配特性(Cumulative Matching Characteristics，CMC)[95]和平均精度(mean Average Precision，mAP)[96]两种指标来评估模型性能。

CMC 评价的是模型在检索时第 K 次能正确判断出目标身份的样本数量和总样本数量的比值，因此一般用 $R@K$ 表示。其计算过程如下：

1)对于给定的 M 个查询样本$\{q_i\}_{i=1}^{M}$，计算样本 q_i 与数据库中所有 N 个样本$\{g_k\}_{k=1}^{N}$的表征间的距离，并根据该距离从小到大对数据集中的样本进行排序，得到排序后的数据库$\{g_{ak}\}_{k=1}^{N}$。

2)计算在前 K 个距离最小的样本中，是否存在与给定的查询样本具有相同身份的样本 g_i，如果存在，则命中数计为 1，否则记为 0。其计算过程如下：

$$acc_i=\begin{cases}1, & g_i\in[g_{a1},\cdots,g_{aK}]\\0, & \text{else}\end{cases} \tag{2}$$

3)重复计算并累积所有样本的命中数量，然后除以总样本数量即可得到 $R@K$ 的准确率。其计算过程如下：

$$R@K=\frac{\sum_{i=1}^{M}acc_i}{M} \tag{3}$$

CMC 值越大说明模型检索出正确结果的概率越大，模型性能越好。

mAP 评价的是在检索过程中所有类别样本检索结果的平均准确率。其计算方式为：计算给定的所有样本在检索数据库时，正确检索样本数量与正确＋错误检索样本数量之和的比值。其计算过程如下：

$$P=\frac{TP}{TP+FP} \tag{4}$$

$$R=\frac{TP}{TP+FN} \tag{5}$$

其中，TP 表示真实标签和预测标签均为同一目标的数量；FP 表示真实标签不为同一目标，而预测标签为同一目标的数量；FN 表示真实标签为同一行人，而预测标签不为同一目标的数量；P 表示精度；R 表示召回率。在此基础上，可以进一步对不同召回率下的精度进行累积，从而得到检索结果的 AP 值，其计算过程如下：

$$AP=\sum_{k=1}^{M}P(k)\Delta R(k) \tag{6}$$

最终，mAP 的计算过程是所有类别目标 AP 的平均值。因此，mAP 值越大说明模型检索出的正样本位置越靠前，模型性能越好。

4 现有方法的性能对比

本章总结了现有跨模态目标重识别方法在 3 个大规模数据集上(SYSU-MM01[34]，RegDB[97]，HITSZ-VCM[51])的实验结果。上述 3 个数据集中的实验结果分别如表 2 和表 3 所列。

表 2 不同跨模态重识别方法在 SYSU-MM01 和 RegDB 数据集上的准确率

Table 2 Accuracy of different cross-modality re-identification methods on SYSU-MM01 and RegDB datasets

(%)

方法类型		方法	SYSU-MM01								RegDB			
			全部搜索				室内搜索				V-I		I-V	
			单图搜索		多图搜索		单图搜索		多图搜索					
			R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
生成式方法	生成式对抗学习方法	D2RL[28]	28.9	29.2	—	—	—	—	—	—	43.4	44.1	—	—
		Hi-CMD[30]	34.9	35.9	—	—	—	—	—	—	70.9	66.0	—	—
		JSIA-ReID[31]	38.1	36.9	45.1	29.5	43.8	52.9	52.7	42.7	48.5	49.3	48.1	48.9
		AlignGAN[27]	42.4	40.7	51.5	33.9	45.9	54.3	57.1	45.3	57.9	53.6	56.3	53.4
		RBDF[35]	57.7	54.4	—	—	—	—	—	—	79.8	76.7	76.2	73.9
		TS-GAN[29]	58.3	55.1	55.9	39.7	62.1	71.3	59.7	50.9	—	—	—	—
		TSME[34]	64.2	61.2	70.3	54.4	64.8	71.5	76.8	65.0	87.4	76.9	86.4	75.7
		AGMNet[32]	69.6	66.1	—	—	74.7	78.3	—	—	88.4	81.5	85.3	81.2
		GC-IFS[33]	74.8	71.5	80.7	66.6	78.7	82.3	86.6	77.4	94.4	92.2	92.9	91.0
	跨模态数据增强方法	LZM[41]	48.9	49.9	—	—	54.3	63.9	—	—	57.0	58.1	—	—
		XIV[37]	49.9	50.7	—	—	—	—	—	—	62.2	60.2	—	—
		CMCEN[36]	50.1	49.5	—	—	56.8	63.6	—	—	74.0	67.5	74.2	67.4
		HAT[39]	55.3	53.9	—	—	62.1	69.4	—	—	71.8	67.6	70.0	66.3
		MID[42]	60.3	59.4	—	—	64.9	70.1	—	—	87.5	84.9	84.3	81.4
		PMT[40]	67.5	65.0	—	—	71.7	76.5	—	—	84.8	76.6	84.2	75.1
		CA[38]	69.9	66.9	—	—	76.3	80.4	—	—	85.0	79.1	84.8	77.8
		TMD[44]	73.9	67.8	—	—	81.2	78.9	—	—	93.0	84.3	87.4	81.3
		PartMix[43]	77.8	74.6	80.5	69.8	81.5	84.4	88.0	80.8	85.7	82.3	84.9	82.5

（续表）

方法类型	方法	SYSU-MM01 全部搜索 单图搜索 R@1	SYSU-MM01 全部搜索 单图搜索 mAP	SYSU-MM01 全部搜索 多图搜索 R@1	SYSU-MM01 全部搜索 多图搜索 mAP	SYSU-MM01 室内搜索 单图搜索 R@1	SYSU-MM01 室内搜索 单图搜索 mAP	SYSU-MM01 室内搜索 多图搜索 R@1	SYSU-MM01 室内搜索 多图搜索 mAP	RegDB V-I R@1	RegDB V-I mAP	RegDB I-V R@1	RegDB I-V mAP
非生成式方法	AGW[46]	47.5	47.7	—	—	54.2	63.0	—	—	70.1	66.4	70.5	65.9
共享模态表征学习方法	DDAG[47]	54.8	53.0	—	—	61.0	68.0	—	—	69.3	63.5	68.1	61.8
	NFS[50]	56.9	55.5	63.5	48.6	62.8	69.8	70.0	61.5	80.5	72.1	78.0	69.8
	DML[54]	58.4	56.1	62.2	49.6	62.4	69.5	66.4	60.0	77.6	84.3	77.0	83.6
	VSD[52]	60.0	68.9	—	—	66.1	73.0	—	—	73.2	71.6	71.8	70.1
	CMTR[48]	65.5	62.9	72.0	57.1	71.5	76.7	80.0	69.5	88.1	81.7	84.9	80.8
	MPANet[57]	70.6	68.2	75.6	62.9	76.7	81.0	84.2	75.1	83.7	80.9	82.8	80.7
	CMIT[55]	70.9	65.5	—	—	73.3	77.2	—	—	88.8	88.5	84.6	83.6
	LUPI[61]	71.1	67.6	—	—	82.4	82.7	—	—	88.0	82.7	86.8	81.3
	TransVI[49]	71.4	68.6	—	—	77.4	81.3	—	—	96.7	91.2	96.3	91.2
	CIFT[58]	74.1	74.8	79.7	75.6	81.8	85.6	88.3	86.4	92.0	92.0	90.3	90.8
	PAENet[59]	74.2	73.9	—	—	78.0	83.5	—	—	97.6	91.4	95.4	90.0
	MBCE[53]	74.8	72.0	78.4	65.8	83.5	86.1	88.4	80.6	93.2	88.3	93.4	88.0
	RMBL[62]	76.1	72.7	82.1	68.4	83.5	85.8	91.8	82.0	94.1	88.8	93.3	88.2
	SEFL[60]	77.1	72.3	—	—	82.1	83.0	—	—	92.2	86.6	91.1	85.2
特定模态表征学习方法	MAC[64]	33.3	36.2	—	—	—	—	—	—	36.4	37.0	—	—
	DSCSN+CCN[65]	35.1	37.4	—	—	—	—	—	—	60.8	60.0	—	—
	DmiR[66]	50.5	49.3	—	—	53.9	62.5	—	—	75.8	70.0	73.9	68.2
	DG-VAE[70]	59.5	58.5	—	—	—	—	—	—	73.0	71.8	—	—
	cm-SSFT[73]	61.6	63.2	63.4	62.0	70.5	72.6	73.0	72.4	72.3	72.9	71.0	71.7
	FMCNet[74]	66.3	62.5	73.4	56.1	68.2	74.1	78.9	63.8	89.1	84.4	88.4	83.9
	CMT[71]	71.9	68.6	80.2	63.1	76.9	79.9	84.9	74.1	95.2	87.3	92.0	84.5
	CycleTrans[69]	72.0	67.2	79.4	62.5	82.6	80.5	89.3	76.3	90.3	84.9	91.3	84.9
	MSMNet[72]	73.5	69.6	78.6	64.1	78.9	81.2	87.3	75.8	96.0	88.6	93.4	86.7
	DSCNet[67]	73.9	69.5	—	—	79.4	82.7	—	—	85.4	77.3	83.5	75.2
	CAL[68]	74.7	71.7	77.1	64.9	79.7	83.7	87.0	78.5	84.5	88.7	93.6	87.6
基于样本个体表征的度量学习	BDTR[76]	17.0	19.7	—	—	—	—	—	—	33.5	31.8	—	—
	HPILN[75]	41.4	43.0	47.6	36.1	45.8	56.5	53.1	47.5	—	—	—	—
	SIM[77]	60.9	56.9	—	—	—	—	—	—	74.5	75.3	75.2	78.3
基于样本中心表征的度量学习	TONE[80]	—	—	—	—	—	—	—	—	16.9	14.9	—	—
	HSME[87]	20.7	23.1	—	—	—	—	—	—	50.9	47.0	50.2	46.2
	cmGAN[83]	27.0	27.8	31.5	22.3	31.6	42.2	37.0	32.8	—	—	—	—
	DFE[86]	48.7	48.6	54.6	42.1	52.3	59.7	59.6	50.6	70.1	69.1	68.0	66.7
	HC[85]	57.0	55.0	62.1	48.0	59.7	64.9	69.8	57.8	—	—	—	—
	HCT[82]	61.7	57.5	—	—	63.4	64.3	—	—	91.1	83.3	89.3	81.5
	MAUM[84]	71.7	68.8	—	—	77.0	81.9	—	—	87.9	85.1	87.0	84.3
	MUN[81]	76.2	73.8	—	—	79.4	82.1	—	—	95.2	87.2	91.9	85.0
	MSCLNet[79]	77.0	71.6	—	—	78.5	81.2	—	—	84.2	80.1	83.9	78.3

表 3 不同跨模态重识别方法在 HITSZ-VCM 数据集上的准确率

Table 3 Accuracy of different cross-modality re-identification methods on HITSZ-VCM dataset

（%）

方法	I-V R@1	I-V R@5	I-V R@10	I-V mAP	V-I R@1	V-I R@5	V-I R@10	V-I mAP
MITML[51]	63.7	76.9	81.7	45.3	64.5	79.0	83.0	47.7
IBAN[56]	65.0	78.3	83.0	48.8	69.6	81.5	85.4	51.0
SEFL[60]	67.7	80.3	84.7	52.3	70.2	82.2	86.1	52.5

表 2 列出了现有方法在 SYSU-MM01 和 RegDB 两个图像数据集中的 VI-ReID 性能。从上述结果可以看出，在 SYSU-MM01 数据集中，性能最强的算法为 PartMix[43]，其在 SYSU-MM01 数据集的“全部搜索”设定下的整体 R@1 和 mAP 准确率达到了 77.8% 和 74.6%。此外，其他各类方法中性能最优的方法也能达到 74.0% 和 71.0% 的 R@1 和 mAP 准确率。这表明，目前在 VI-ReID 领域，各类算法都有各自的优势。在 RegDB 数据集中，性能最强的算法为 PAENet[59]，其在 V-I 和 I-V 两种设定中分别达到了 97.6% 和 95.4% 的 R@1 和 mAP 准确率。相比 RegDB 数据集，现有方法在较难的 SYSU-MM01 数据集上性能下降明显。上述结果说明，复杂场景中的 VI-ReID 仍然存在严峻挑战和改进空间，亟需具有更全面、更复杂场景的数据集来系统地评价 VI-ReID 算法性能。此外，如何将不同类型算法的优势进行整合，从而提升模型在不同场景中的泛化性能有待进一步探索和研究。

表 3 列出了现有方法在 HITSZ-VCM 数据集中的结果。该数据集是最近提出的一个基于视频的跨模态行人重识别数据集，仅有少量最新研究工作在其上进行了性能测试，现有性能仍然有较大的提升空间。这说明基于视频数据的跨模态行人重识别仍然是一个亟待解决的难题。

5 总结与展望

跨模态目标重识别旨在匹配不同模态图像中的同一目标，是实现跨时段以及全天候的目标识别、追踪的重要技术，具有重要的科学研究价值和实际应用意义。虽然现有工作取得了一定进展，但跨模态目标重识别的研究仍处在探索阶段。总的来说，跨模态目标重识别还需要重点解决以下难点：

1)数据受限场景下的跨模态目标重识别

尽管现有方法在高质量公共数据集中取得了较好的效果,但在真实场景中,往往难以收集到足够数量的高质量跨模态目标数据,过少的、低质量的样本极易导致现有方法性能急剧下降。另一方面,即便能收集到海量数据,对其进行大规模标注的挑战性和代价往往难以承受,而弱标注样本可能会进一步增加模型训练的难度。最后,即便进行了标注,也很难保证标注的质量,含有噪声的样本会导致模型难以正确关联同一目标。上述问题会严重制约现有跨模态目标重识别的性能。因此,解决小样本、弱监督以及带噪场景中的跨模态目标重识别难题是亟待解决的难题之一。

2)开放域跨模态目标重识别

现有跨模态目标重识别方法大多在封闭数据集中进行性能验证与测试,然而真实世界是一个开放场景,查询目标往往并不存在于收集的数据库中,导致现有方法无法应对上述情况,丧失判断能力。因此,现有跨模态目标重识别方法在真实场景中应用的一个关键性挑战是开放域中的跨模态目标重识别方法研究。

3)基于动态数据流的跨模态目标重识别

真实场景中的数据通常不是一次性给定的,而是以数据流的形式按不同时间、不同地点动态采集的,这要求模型在不断学习新数据中包含的新知识的同时,抵抗模型对已经学习的旧知识的遗忘。除此之外,学习后的模型如何快速部署到下游场景中也是实际中面临的挑战之一。

4)面向多模态数据的跨模态目标重识别

现有跨模态目标重识别方法大多面向图像数据,然而随着多模态数据的出现,面向视频/文本/3D等模态数据的跨模态目标重识别需求日益增加。因此,如何将现有跨模态目标重识别方法进一步扩展到更多模态的数据中,是该领域未来的研究重点之一。

5)构建符合真实场景的数据集和更全面的评价指标

现有跨模态目标重识别数据集并不能反映真实的应用场景,其问题在于采集设备少、数据量少、目标类型少,不能用于全面评估现有方法的性能。因此,有必要构建更完备的数据集以反应真实场景中的重识别需求。其次,现有跨模态目标重识别评价指标主要沿用了评价传统的重识别任务的通用指标,如何结合跨模态目标重识别任务的特性,设计更具针对性的评价指标,是未来的发展方向之一。

结束语 本文首先回顾并总结了跨模态目标重识别任务的定义、面临的挑战、方法分类以及方法的评估;然后根据现有方法设计动机的不同,从生成式和非生成式两个角度对现有跨模态目标重识别方法进行了系统的回顾和阐述;接着,整理了现有跨模态目标重识别评价数据集与方法,并在此基础上对比了现有方法在该任务上的重识别性能;最后对跨模态目标重识别存在的研究难题进行了归纳,并为未来研究进行了展望。可以看出,跨模态目标重识别具有重要的实际价值、研究意义和广阔的未来发展空间。

参 考 文 献

[1] YE Y, WANG Z, LIANG C, et al. A survey on multi-source person re-identification[J]. Acta Automatica Sinica, 2020, 46(9): 1869-1884.

[2] YANG F, XU Y, YIN M, et al. Review on deep learning-based pedestrian re-identification[J]. Journal of Computer Applications, 2020, 40(5): 1243.

[3] QI L, YU P, GAO Y. Research on weak-supervised person re-identification[J]. Journal of Software, 2020, 31(9): 2883-2902.

[4] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]// European Conference on Computer Vision. 2016: 17-35.

[5] SONG W, ZHAO Q, CHEN C, et al. Survey on pedestrian re-identification research[J]. CAAI Transaction on Intelligent Systems, 2017, 12(6): 770-780.

[6] SUN H, HE X, PENG Y. HCL: Hierarchical Consistency Learning for Webly Supervised Fine-Grained Recognition[J]. IEEE Transactions on Multimedia, 2023: 1-13. DOI: 10.1109/TMM.2023.3330076.

[7] SUN H, HE X, ZHOU J, et al. Fine-Grained Visual Prompt Learning of Vision-Language Models for Image Recognition[C]// Proceedings of the 31st ACM International Conference on Multimedia. 2023: 5828-5836.

[8] CUI Z Y, ZHOU J H, PENG Y X, et al. DCR-ReID: Deep Component Reconstruction for Cloth-Changing Person Re-Identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(8): 4415-4428.

[9] LENG J, WANG H, GAO X, et al. Where to look: Multi-granularity occlusion aware for video person re-identification[J]. Neurocomputing, 2023, 536: 137-151.

[10] ZHONG Z, ZHENG L, CAO D, et al. Re-ranking person re-identification with k-reciprocal encoding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1318-1327.

[11] 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.

[12] SUH Y, WANG J, TANG S, et al. Part-aligned bilinear representations for person re-identification[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 402-419.

[13] LI D, CHEN X, ZHANG Z, et al. Learning deep context-aware features over body and latent parts for person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 384-393.

[14] CHEN Z Q, CUI Z C, ZHANG C, et al. Dual Clustering Co-teaching with Consistent Sample Mining for Unsupervised Person Re-Identification[J]. arXiv: 2210.03339, 2023.

[15] ZHOU J, SU B, WU Y. Online joint multi-metric adaptation from frequent sharing-subset mining for person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2909-2918.

[16] ZHOU J, SU B, WU Y. Easy identification from better constraints: Multi-shot person re-identification from reference constraints[C]// Proceedings of the IEEE Conference on Computer

Vision and Pattern Recognition. 2018:5373-5381.

[17] ZHOU J,YU P,TANG W,et al. Efficient online local metric adaptation via negative samples for person re-identification [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2420-2428.

[18] YOU J,WU A,LI X,et al. Top-push video-based person re-identification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1345-1353.

[19] ZHENG L,BIE Z,SUN Y,et al. Mars:A video benchmark for large-scale person re-identification[C]// Computer Vision—ECCV 2016:14th European Conference,Amsterdam,The Netherlands,October 11-14,2016,Proceedings,Part VI 14. Springer International Publishing,2016:868-884.

[20] YU S,LI S,CHEN D,et al. Cocas:A large-scale clothes changing person dataset for re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:3400-3409.

[21] ZHENG Z,ZHENG L,YANG Y. Pedestrian alignment network for large-scale person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology,2018,29(10):3037-3045.

[22] WAXMAN A M,AGUILAR M,FAY D A,et al. Solid-state color night vision:fusion of low-light visible and thermal infrared imagery[J]. Lincoln Laboratory Journal,1998,11(1):41-60.

[23] AGUILAR M,FAY D A,ROSS W D,et al. Real-time fusion of low-light CCD and uncooled IR imagery for color night vision [C]// Enhanced and Synthetic Vision 1998. SPIE,1998:124-135.

[24] 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,2021,44(6):2872-2893.

[25] HUANG N,LIU J,MIAO Y,et al. Deep learning for visible-infrared cross-modality person re-identification:A comprehensive review[J]. Information Fusion,2023,91:396-411.

[26] WU A,ZHENG W S,YU H X,et al. RGB-infrared cross-modality person re-identification[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5380-5389.

[27] 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.

[28] WANG Z,WANG Z,ZHENG Y,et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:618-626.

[29] ZHANG Z,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.

[30] 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.

[31] WANG G A,ZHANG T,YANG Y,et al. Cross-modality paired-images generation for RGB-infrared person re-identification [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:12144-12151.

[32] 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.

[33] QI J,LIANG T F,LIU W,et al. A Generative-based Image Fusion Strategy for Visible-infrared Person Re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology,Early Access.

[34] LIU J,WANG J,HUANG N,et al. Revisiting modality-specific feature compensation for visible-infrared person re-identification [J]. IEEE Transactions on Circuits and Systems for Video Technology,2022,32(10):7226-7240.

[35] WEI Z,YANG X,WANG N,et al. RBDF:Reciprocal bidirectional framework for visible infrared person re-identification[J]. IEEE Transactions on Cybernetics,2022,52(10):10988-10998.

[36] XU X,LIU S,ZHANG N,et al. Channel exchange and adversarial learning guided cross-modal person re-identification [J]. Knowledge-Based Systems,2022,257:109883.

[37] LI D,WEI X,HONG X,et al. Infrared-visible cross-modal person re-identification with an x modality[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:4610-4617.

[38] YE M,RUAN W,DU B,et al. Channel augmented joint learning for visible-infrared recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:13567-13576.

[39] YE M,SHEN J,SHAO L. Visible-infrared person re-identification via homogeneous augmented tri-modal learning[J]. IEEE Transactions on Information Forensics and Security,2020,16:728-739.

[40] 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. 2023:1835-1843.

[41] BASARAN E,GÖKMEN M,KAMASAK M E. An efficient framework for visible-infrared cross modality person re-identification[J]. Signal Processing:Image Communication,2020,87:115933.

[42] HUANG Z,LIU J,LI L,et al. Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022:1034-1042.

[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. 2023:18621-18632.

[44] LU Z,LIN R,HU H. Tri-level Modality-information Disentanglement for Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Multimedia,Early Access.

[45] YANG B,YE M,CHEN J,et al. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification[C]// Proceedings of the 30th ACM International

Conference on Multimedia. 2022:2843-2851.
[46] 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, 2021, 44(6): 2872-2893.
[47] YE M, SHEN 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, Part XVII 16. 2020:229-247.
[48] LIANG T, JIN Y, LIU W, et al. Cross-Modality Transformer With Modality Mining for Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Multimedia, Early Access.
[49] CHAI Z, LING Y, LUO Z, et al. Dual-stream Transformer with Distribution Alignment for Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, Early Access.
[50] 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.
[51] LIN X, LI J, MA Z, et al. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:20973-20982.
[52] TIAN X, ZHANG Z, LIN S, et al. Farewell to mutual information: Variational distillation for cross-modal person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:1522-1531.
[53] CHENG D, WANG X, WANG N, et al. Cross-modality person re-identification with memory-based contrastive embedding [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023:425-432.
[54] ZHANG D, ZHANG Z, JU Y, et al. Dual mutual learning for cross-modality person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5361-5373.
[55] FENG Y, YU J, CHEN F, et al. Visible-Infrared Person Re-Identification via Cross-Modality Interaction Transformer[J]. IEEE Transactions on Multimedia, Early Access.
[56] LI H, LIU M, HU Z, et al. Intermediary-guided Bidirectional Spatial-Temporal Aggregation Network for Video-based Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, Early Access.
[57] WU Q, DAI P, CHEN J, et al. Discover cross-modality nuances for visible-infrared person re-identification[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:4330-4339.
[58] LI X, LU Y, LIU B, et al. Counterfactual Intervention Feature Transfer for Visible-Infrared Person Re-identification[C]// European Conference on Computer Vision. 2022:381-398.
[59] ZHENG A, PAN P, LI H, et al. Progressive attribute embedding for accurate cross-modality person re-id[C]// Proceedings of the 30th ACM International Conference on Multimedia. 2022:4309-4317.
[60] FENG J, WU A, ZHENG W S. 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.
[61] ALEHDAGHI M, JOSI A, CRUZ R M, et al. Visible-infrared person re-identification using privileged intermediate information[C]// European Conference on Computer Vision. 2022:720-737.
[62] WU J, LIU H, SHI W, et al. Style-Agnostic Representation Learning for Visible-Infrared Person Re-identification[J]. IEEE Transactions on Multimedia, Early Access.
[63] LI H, LI C, ZHU X, et al. Multi-spectral vehicle re-identification: A challenge[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:11345-11353.
[64] YE M, LAN X, LENG Q. Modality-aware collaborative learning for visible thermal person re-identification[C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019: 347-355.
[65] ZHANG S, YANG Y, WANG P, et al. Attend to the difference: Cross-modality person re-identification via contrastive correlation[J]. IEEE Transactions on Image Processing, 2021, 30: 8861-8872.
[66] HU W, LIU B, ZENG H, 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.
[67] ZHANG Y, KANG Y, ZHAO S, et al. Dual-Semantic Consistency Learning for Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Information Forensics and Security, 2022, 18: 1554-1565.
[68] WU J, LIU H, SU Y, et al. Learning Concordant Attention via Target-aware Alignment for Visible-Infrared Person Re-identification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023:11122-11131.
[69] WU Q, XIA J, DAI P, et al. CycleTrans: Learning Neutral yet Discriminative Features for Visible-Infrared Person Re-Identification[J]. arXiv: 2208.09844, 2022.
[70] PU N, CHEN W, LIU Y, et al. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification[C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020:2149-2158.
[71] JIANG K, ZHANG T, LIU X, et al. Cross-modality transformer for visible-infrared person re-identification[C]// European Conference on Computer Vision. 2022:480-496.
[72] LI Y, ZHANG T, LIU X, et al. Visible-Infrared Person Re-Identification With Modality-Specific Memory Network[J]. IEEE Transactions on Image Processing, 2022, 31: 7165-7178.
[73] LU Y, WU Y, LIU B, et al. Cross-modality person re-identification with shared-specific feature transfer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:13379-13389.
[74] ZHANG Q, LAI C, LIU J, 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. 2022:7349-7358.
[75] ZHAO Y B,LIN J W,XUAN Q,et al. Hpiln:a feature learning framework for cross-modality person re-identification[J]. IET Image Processing,2019,13(14):2897-2904.
[76] YE M,WANG Z,LAN X,et al. Visible thermal person re-identification via dual-constrained top-ranking[C]// IJCAI. 2018.
[77] JIA M,ZHAI Y,LU S,et al. A similarity inference metric for RGB-infrared cross-modality person re-identification[J]. arXiv:2007.01504,2020.
[78] KAMENOU E,Del RINCON J M,MILLER P,et al. Closing the domain gap for cross-modal visible-infrared vehicle re-identification[C]// 2022 26th International Conference on Pattern Recognition(ICPR). 2022:2728-2734.
[79] ZHANG Y,ZHAO S,KANG Y,et al. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification[C]// European Conference on Computer Vision. 2022:462-479.
[80] 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.
[81] YU H,CHENG X,PENG W,et al. Modality Unifying Network for Visible-Infrared Person Re-Identification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023:11185-11195.
[82] LIU H,TAN X,ZHOU X. Parameter sharing exploration and hetero-center triplet loss for visiblethermal person re-identification[J]. IEEE Transactions on Multimedia, 2020, 23: 4414-4425.
[83] DAI P,JI R,WANG H,et al. Cross-modality person re-identification with generative adversarial training[C]// IJCAI. 2018.
[84] LIU J,SUN Y,ZHU F,et al. Learning memory-augmented unidirectional metrics for cross-modality person re-identification [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:19366-19375.
[85] ZHU Y,YANG Z,WANG L,et al. Hetero-center loss for cross-modality person re-identification[J]. Neurocomputing, 2020, 386:97-109.
[86] HAO Y,WANG N,GAO X,et al. Dual-alignment feature embedding for cross-modality person re-identification[C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019:57-65.
[87] HAO Y,WANG N,LI J,et al. HSME:Hypersphere manifold embedding for visible thermal person re-identification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:8385-8392.
[88] GOODFELLOW I,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial networks[J]. Communications of the ACM, 2020,63(11):139-144.
[89] ZHU J Y,PARK T,ISOLA P,et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:2223-2232.
[90] DOSOVITSKIY A,BEYER L,KOLESNIKOV A,et al. An image is worth 16x16 words:Transformers for image recognition at scale[J]. arXiv:2010.11929,2020.
[91] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems,2017,60(6):84-90.
[92] SCARSELLI F,GORI M,TSOI A C,et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008,20(1):61-80.
[93] HINTON G,VINYALS O,DEAN J. Distilling the knowledge in a neural network[J]. arXiv:1503.02531,2015.
[94] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.
[95] MOON H,PHILLIPS P J. Computational and performance aspects of PCA-based face-recognition algorithms[J]. Perception, 2001,30(3):303-321.
[96] ZHENG L,SHEN L,TIAN L,et al. Scalable person re-identification:A benchmark[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015:1116-1124.
[97] 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.

CUI Zhenyu, born in 1995, postgraduate. His main research interests include computer vision and deep learning.

PENG Yuxin, born in 1974, Ph.D, professor. His main research interests include cross-media analysis and reasoning, image and video recognition and understanding, and computer vision.

(责任编辑:何杨)