宁波大学硕士研究生 2025/2026 学年第 1 学期期末考试试卷
答题纸及评分标准

考试科目：多媒体技术与应用　课程编号：_____　考卷类型：学术论文报告
姓名：洪玉杰　学号：2511100136　阅卷教师：_____　成绩：_____

（答案必须写在答题纸上）

多模态行人重识别研究综述

摘　要：多模态行人重识别旨在突破单模态识别的场景局限，通过融合可见光图像、红外图像、文本语义、视频序列等多源模态信息，实现复杂环境下行人身份的精准匹配，是智能安防与计算机视觉领域的核心技术。本文系统梳理多模态行人重识别的研究体系：首先界定模态、跨模态匹配、模态鸿沟等核心概念，划分涵盖可见光 - 红外、文本 - 图像、视频 - 单帧模态（如图像）的双模态跨模态任务及多源融合任务，建立以 Rank-k 准确率、mAP 为核心，MMD 为辅助的评价指标体系；其次按技术路线综述五大核心方法，包括特征对齐与融合、注意力与特征增强、语义与无监督学习、视频跨模态融合及模态差异消除，深入分析各类方法的技术逻辑与优势；最后总结当前面临的模态鸿沟缓解、动态融合适配、低资源场景泛化及工程化部署四大挑战，并展望多模态大模型融合、自适应差异消除、低资源学习优化及跨平台适配等未来方向。

关键词：多模态行人重识别；跨模态匹配；特征对齐；注意力机制；无监督学习；视频多模态融合

Review of multimodal pedestrian re-identification

Abstract: Multimodal person re-identification aims to break through the scenario limitations of single-modal recognition. By fusing multi-source modal information such as visible light images, infrared images, text semantics, and video sequences, it achieves accurate matching of pedestrian identities in complex environments, and is a core technology in the fields of intelligent security and computer vision. This paper systematically sorts out the research system of multimodal person re-identification: firstly, it defines core concepts such as modality, cross-modal matching, and modality gap, classifies bimodal cross-modal tasks covering visible light-infrared, text-image, and video-single-frame modality (e.g., images) as well as multi-source fusion tasks, and establishes an evaluation index system with Rank-k accuracy and mAP as the core and MMD as the auxiliary; secondly, it summarizes five core methods according to technical routes, including feature alignment and fusion, attention and feature enhancement, semantics and unsupervised learning, video cross-modal fusion, and modality difference elimination, and deeply analyzes the technical logic and advantages of each method; finally, it summarizes the four major challenges currently faced, namely modality gap mitigation, dynamic fusion adaptation, generalization in low-resource scenarios, and engineering deployment, and looks forward to future directions such as multimodal large model fusion, adaptive difference elimination, low-resource learning optimization, and cross-platform adaptation.

Key words: multimodal person re-identification; cross-modal matching; feature alignment; attention mechanism; unsupervised learning; video multimodal fusion

1　引言

1.1　研究背景和应用价值

随着城市现代化建设加速与智能安防体系的全面普及，行人重识别（Person Re-identification, ReID）作为计算机视觉领域的核心技术，在智能监控、智能交通、无人驾驶等关键领域发挥着不可或缺的作用 [1]。其核心目标是从非重叠摄像机视域中，通过提取行人衣着、体态、发型等判别性特征，实现对同一行人的精准检索与匹配，被视为图像检索任务的重要子方向 [2−3]。典型的 ReID 系统流程为：给定查询行人图像（Probe）与行人图像库（Gallery），经特征提取与相似度匹配后，输出匹配图像的排序结果 [4]，如图 1 所示。作为智能监控系统的关键支撑技术，ReID 为全天候、全方位的安全防控提供了核心技术保障，近年来已成为学术界与工业界共同关注的热点研究方向 [5]。

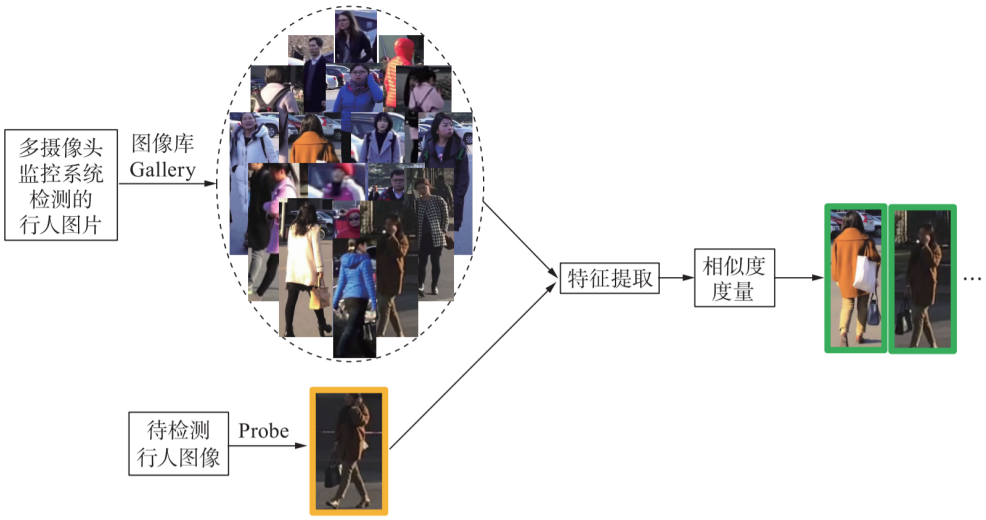

图 1: 行人重识别

1.2 单模态行人重识别的局限性

行人重识别任务的研究可追溯至 1997 年的多相机跟踪问题，2005 年 Zajdel 等人首次明确提出“行人重识别”概念 [6]。2014 年，Yi 等人将深度学习技术引入该领域，构建了端到端的 ReID 系统，推动了技术的跨越式发展，至今深度学习仍是解决 ReID 问题的主流方法 [7]。早期研究主要聚焦于基于可见光图像的单模态 ReID，虽已取得显著进展，但在实际应用场景中面临诸多瓶颈。一方面，不同摄像头采集的图像存在视角、姿态、光照、背景差异等模态内变化，严重影响视觉信息的稳定性，导致模型识别精度下降；另一方面，可见光摄像头在夜间、雨天、雾天等复杂环境下成像质量极差，无法捕获有效的行人外观信息，极大限制了单模态 ReID 的场景适用性 [8]。尽管研究人员引入注意力机制突出行人关键区域、抑制冗余信息，在一定程度上缓解了上述问题 [9]，但在极端光照或重度遮挡场景下，仍难以获取足够的判别性信息。此外，有监督单模态 ReID 方法依赖大量标注数据，标注成本高昂，难以扩展到大型未标记数据集场景 [10]。

1.3 多模态行人重识别的兴起

为突破单模态 ReID 的应用局限，充分利用不同传感器捕获信息的互补性，研究者们逐步将多模态学习引入 ReID 任务，形成了多模态行人重识别研究方向。多模态行人重识别通过融合红外、深度、文本、步态等多种模态数据，利用不同模态在复杂环境下的稳定特性，例如红外图像不受光照影响、步态特征具备远距离识别优势，有效补充或替代可见光信息，显著提升了模型的环境适应性与识别鲁棒性。根据任务目标差异，多模态行人重识别可分为两类核心场景：一是跨模态行人重识别，旨在建立两种不同模态间的匹配关系，典型场景包括可见光-红外、文本-图像、草图-可见光等 [11]；二是模态融合行人重识别，通过整合多模态信息的互补性与协同性，丰富行人特征表示以提升识别准确性 [12]。其中，2017 年 Wu 等人首次定义了可见光-红外跨模态 ReID 问题，提出深度零填充网络结构处理模态差异，并构建了大规模公开数据集 SYSU-MM01，推动了跨模态 ReID 的快速发展；随后文本-图像

[13]、步态-可见光 [14]、视频跨模态等场景的研究也逐步展开，进一步拓展了多模态 ReID 的应用边界。

1.4 研究现状

近年来，跨模态行人重识别作为多模态 ReID 的核心研究分支，涌现出大量细分方向的创新方法，形成了多元化的技术体系 [15]。在特征对齐与聚合方向，研究者们从三重嵌入扩展、频域空间信息、多粒度融合等多个视角切入，强化跨模态特征的一致性表达；在模态差异消除方向，特征解纠缠、多样化生成、通道干预渐进式差异减小等方法被广泛应用，有效缩小模态间分布鸿沟；在注意力与特征增强方向，内容感知注意力、跨通道交互注意力、辅助模态引导注意力等多样化机制被提出，精准聚焦关键特征并抑制冗余信息；同时，语义增强、CLIP 模型改进等语义对齐方法，以及无监督、半监督学习方法，进一步提升了模型的泛化能力与低资源场景适应性；视频跨模态 ReID 通过时空特征增强，也成为复杂动态场景下的重要研究方向。

1.5 综述结构与研究贡献

现有相关综述多聚焦于特定模态组合例如可见光-红外或单一技术方向，缺乏对跨模态行人重识别多元化技术体系的系统性梳理，尤其未充分涵盖频域特征聚合、多样化注意力机制、视频时空增强、CLIP 改进等细分方向。为此，本文以“跨模态行人重识别”为核心研究对象，系统撰写多模态行人重识别研究综述，全面梳理该领域的发展历程、核心技术与研究现状，弥补现有综述的不足。本文的研究贡献主要包括：1）系统界定多模态行人 ReID 的核心范畴与任务类型，重点梳理跨模态 ReID 的主流数据集与评价指标；2）按技术路线分类，全面综述特征对齐与聚合、模态差异消除、注意力融合、语义增强、无监督/少样本学习、视频跨模态等核心方法，深入分析各类方法的优劣；3）总结当前研究挑战并展望未来发展方向。具体结构安排如下：首先阐述多模态 ReID 的基础概念、任务定义与常用数据集；其次详细综述各类核心技术方法；最后总结挑战并展望未来，为该领域的后续研究提供全面、系统的参考。

2 多模态行人重识别基础

多模态行人重识别作为计算机视觉与智能安防领域的关键技术，其研究开展离不开清晰的基础框架支撑。其中，核心概念的界定是厘清研究边界的前提，任务类型的划分是明确研究方向的基础，科学的评价指标则是衡量技术性能的核心标准。本章作为后续核心技术分析的前置铺垫，将系统阐释领域核心概念、详细划分核心任务类型，并明确关键评价指标体系，帮助读者快速建立对该领域的基础认知，为后续技术梳理与实验验证提供坚实的理论与标准支撑。

2.1 核心概念与任务类型划分

2.1.1 核心概念

模态：指信息的载体与表现形态，是多模态研究的基础单元，具体可分为视觉模态与非视觉模态两大类。视觉模态包括可见光图像、红外图像、深度图像等，能直观反映行人外观与空间信息；非视觉模态则涵盖文本描述、步态序列、语音信号等，可提供语义或行为层面的补充信息。模态具备两大核心特性：互补性体现为不同模态可提供同一行人的差异化信息，例如可见光提供纹理细节，红外提供轮廓信息，异质性则表现为不同模态的数据源、数据维度与分布特性存在显著差异。多模态行人重识别：广义上指利用两种及以上模态数据完成跨摄像头视域的行人身份匹配任务，既包含不同模态间的跨模态匹配，也涵盖同模态下多源数据的融合增强；狭义上则聚焦跨模态匹配这一核心场景，也是当前研究的主流方向。跨模态行人重识别：作为多模态行人重识别的核心分支，特指查询样本与候选样本分属不同模态的匹配任务，核心目标是突破模态异质性限制，建立不同信息载体间的身份映射关系。模态鸿沟：是跨模态行人重识别的核心挑战，由模态异质性导致不同模态特征分布出现显著偏移，使得同一行人在不同模态下的特征相似度低于不同行人在同模态下的相似度，严重影响模型的识别性能。

2.1.2 任务类型划分

结合模态组合方式与实际应用场景，跨模态行人重识别可划分为四类核心任务，各类任务的技术侧重点与应用价值存在显著差异：1）可见光 - 红外跨模态 ReID：核心解决昼夜监控场景下的行人匹配问题，白天依赖可见光摄像头采集纹理丰富的图像，夜间则通过红外摄像头获取不受光照影响的轮廓信息，其代表数据集包括 SYSU-MM01 与 RegDB [16]，前者面向室内场景且样本量充足，后者聚焦室外场景，更贴近实际安防需求。2）文本 - 图像跨模态 ReID：重点攻克文本语义与图像视觉信息的对齐难题，适用于无图寻人等场景如仅通过目击者描述检索监控中的目标行人，核心挑战是弥合语义与视觉之间的表达鸿沟，对应的代表数据集为 CUHK-PEDES [17]，该数据集样本规模较大且文本描述较为详细。3）视频 - 图像 / 视频跨模态 ReID：针对动态监控场景，利用视频序列包含的时空信息提取更稳定的行人特征，缓解姿态变化、瞬时遮挡等因素带来的干扰，可实现动态追踪与身份确认的协同，其代表数据集为 VideoReID-MM [18]，可支持视频 - 图像、视频 - 视频两种跨模态任务。4）多源模态融合 ReID：融合三种及以上模态数据例如可见光、红外、步态与深度，通过多维度信息互补进一步提升复杂场景下的识别鲁棒性，更贴近实际复杂安防需求，但数据采集与标注成本较高，代表数据集为 CASIA Gait Dataset B [19]，该数据集涵盖四种模态，可支撑步态与视觉模态的融合研究。表 1 是这几种数据集的对比。

表 1: 主流多模态行人重识别数据集对比

数据集名称	模态类型	身份数量	样本规模（图像/视频帧）	核心任务适配
SYSU-MM01	可见光 + 红外	491	30k+RGB/15k+ 红外	可见光-红外 ReID
RegDB	可见光 + 红外	412	8k+（每身份 20 张）	可见光-红外 ReID
CUHK-PEDES	图像 + 文本	13003	40k+ 图像/80k+ 文本	文本-图像 ReID
VideoReID-MM	视频 + 图像/视频	1000+	50k+ 视频帧	视频跨模态 ReID
CASIA Gait B	可见光 + 红外 + 步态 + 深度	124	多视角视频序列	多源融合 ReID

2.2 评价指标

为全面、准确地衡量跨模态行人重识别方法的性能，利用了核心排序匹配指标加辅助模态差异指标的完整评价体系，核心排序匹配指标是方法性能对比的核心标准，直接反映模型的检索与匹配能力，主要包括 Rank-k 准确率与平均精度均值（mAP）[20]。其中，Rank-k 准确率衡量在检索结果的前 k 个候选样本中包含目标行人的比例，常用 k 值为 1、5、10，Rank-1 准确率直观反映模型的最优匹配能力，Rank-5 与 Rank-10 准确率则体现模型的排序稳定性；平均精度均值（mAP）则综合考虑所有查询样本的平均精度，既关注正确匹配样本的比例，也重视排序的合理性，能更全面地评估模型的整体性能。辅助模态差异指标用于补充验证模型对模态鸿沟的缓解效果，以最大均值差异（MMD）为典型代表，通过衡量不同模态特征分布的差异程度，判断模型的模态对齐效果即 MMD 值越小，说明两种模态的特征分布越接近，模态鸿沟缓解效果越好。两类指标协同作用，既能直观判断模型的实际应用性能，也能深入分析技术方法的核心优势与不足 [21]。

3 多模态行人重识别核心技术方法

3.1 特征对齐与多维度跨模态融合

特征对齐与融合是缓解多模态行人重识别中模态鸿沟的核心技术路径，相关代表性研究各有侧重：Li 等 [22] 提出双注意力引导融合的跨模态特征对齐方法，通过模态内注意力细化单模态局部特征、模态间注意力对齐跨模态语义特征，以双注意力协同强化模态不变表示；程德强等 [23] 设计共享 - 解离模块分离模态共享与特异性特征，多粒度相关特征学习模块挖掘行人身体区域的跨模态关联信息，搭配跨尺度感知机制融合不同分辨率特征，有效缓解姿态变化与局部遮挡带来的负面影响；Zhang 等 [24] 构建局部 - 全局分层特征融合框架，引入模态一致性损失约束跨模态特征分布，通过分层聚合与分布

约束协同提升匹配精度；金静等 [25] 提出基于频域空间信息的特征聚合策略，构建四流特征提取网络分别获取可见光与红外模态的深度语义及浅层细节特征，经时空频融合模块实现域间信息转换与融合，再通过哈达玛积结合注意力加权完成特征聚合，强化模态间特征一致性；Wang 等 [26] 设计自适应特征对齐与渐进式融合网络，通过自适应对齐模块动态适配模态差异，以渐进式策略逐步聚合多尺度特征，显著增强特征判别性；刘锁兰等 [27] 提出三重嵌入扩展与分层特征聚合方法，从模态内、模态间、跨域三个维度优化特征映射，通过早期融基础、中期筛关键、晚期聚语义的分层架构，提升跨模态特征空间的判别能力；Chen 等 [28] 提出频域 - 空域协同对齐与融合方法，实现双域特征的协同对齐与互补融合，弥补单一域特征的信息缺陷，进一步缓解模态鸿沟；朱沛伍等 [29] 提出多频多尺度嵌入（MFME）模型，通过多尺度信息融合模块提取行人全局结构与局部细节特征，结合低高频特征聚合模块分解并聚合不同模态的频域信息，强化模态不变特征以提升复杂环境下的匹配鲁棒性。

3.2 多样化注意力机制与特征增强

注意力机制与特征增强是提升跨模态行人重识别性能的关键技术方向，相关研究通过多维度优化模态间特征交互与表达能力，有效缓解模态鸿沟：Yang 等 [30] 提出区域增强与跨模态注意力（RACA）模型，通过 PedMix 区域数据增强模块提升行人区域连贯性，搭配轻量化模态特征转移模块融合跨注意力与卷积网络，在减少干扰信息的同时控制计算开销，于主流数据集上验证了有效性；郭思琦等 [31] 引入辅助模态与注意力机制相结合的方法，通过强化主模态与辅助模态的关联学习，挖掘模态不变特征，显著提升复杂场景下跨模态匹配的鲁棒性；Chen 等 [32] 提出跨调制注意力，设计关联调制增强模块通过模态间共识修正注意力权重，统一自注意力与跨注意力机制，既缓解了单模态特征表示局限，又提升了多模态特征交互的鲁棒性；杨真真等 [33] 提出混合卷积增强与内容感知注意力融合方法，利用混合卷积强化特征提取能力，通过内容感知注意力精准聚焦行人关键区域，双重优化减少背景干扰，提升跨模态特征一致性；Huang 等 [34] 提出注意力增强多模态特征融合网络（AE-Net），整合 RGB 全局特征、灰度图像特征与语义分割无关服装特征，结合多尺度融合注意力机制，有效降低服装变化对识别精度的影响；黄驰涵等 [35] 设计融合注意力与特征增强一体化框架，通过融合注意力模块实现跨模态语义精准对齐，搭配特征增强策略强化细粒度细节信息，显著提升特征判别性与匹配精度；何磊等 [36] 提出跨通道交互注意力驱动的双流网络，以双流结构分别提取双模态特征，通过跨通道交互注意力促进模态间通道信息深度交互，强化特征关联性与一致性；陈梦蝶等 [37] 构建双重增强网络，从特征表达优化与模态对齐校准两个维度展开设计，结合注意力机制筛选有效特征信息，进一步提升跨模态特征质量与匹配稳定性。

3.3 语义增强与无监督 / 半监督学习

语义挖掘与无监督 / 半监督学习为跨模态行人重识别提供了低数据依赖的有效解决方案，通过强化模态间语义一致性与优化无监督特征学习，显著缓解模态鸿沟与标注数据匮乏问题，相关代表性研究各有侧重：Liu 等 [38]2025 年提出语义对齐协同优化的无监督跨模态重识别方法（SALCR），采用双流网络与 ResNet50 主干网络，通过语义对齐学习与协同细化机制，在 RegDB 和 SYSU-MM01 数据集上优化模态不变特征提取，无需人工标注即可实现高效跨模态匹配；董令赟等 [39] 提出跨模态语义一致性与半监督结合的算法，通过语义编码模块映射特征至统一语义空间，利用伪标签生成扩展标注数据，设计语义一致性损失约束特征分布，减少标注依赖的同时提升匹配判别性；Pang 等 [40]2025 年提出增强型软化匹配的无监督方法（ASM），通过跨模态增强匹配模块提升伪标签对颜色变化的鲁棒性，搭配软标签动量更新策略优化训练稳定性，经模态内与跨模态学习迭代优化特征对齐效果；宋存利等 [41] 设计语义增强网络，引入行人属性语义信息辅助视觉特征学习，通过属性 - 特征交互模块强化语义关联，采用无监督对比学习优化特征分布，降低环境干扰与模态差异影响；Qin 等 [42]2025 年提出基于 MLLMs 的人机交互框架（ICL），通过测试时人机交互细化文本查询语义，结合重组数据增强丰富文本描述多样性，强化文本 - 图像细粒度语义对齐；贾军营等 [43] 改进 CLIP-ReID 框架，优化

跨模态语义映射，引入无监督对比损失与语义一致性约束，简化模型结构的同时增强特征兼容性，提升语义驱动的无监督匹配性能；Wang 等 [44]2024 年提出多记忆匹配框架（MMM），通过跨模态聚类生成伪标签，设计多记忆学习匹配模块挖掘个体视角细微特征，搭配软聚类对齐损失缩小模态鸿沟并抑制噪声伪标签；Zhang 等 [45]2025 年提出同质 - 异质一致性标签关联方法，设计模态统一标签转移模块建模实例级亲和性，通过在线跨记忆标签细化优化伪标签质量，结合模态不变表示学习提升标签关联精细度。

3.4 视频跨模态融合与时空信息协同

视频多模态融合类方法通过充分挖掘视频序列中的时空信息与多模态互补价值，显著提升跨模态行人重识别的鲁棒性与实用性，相关代表性研究各具特色：Yao 等 [46]2025 年提出映射与协同学习（MCL）框架，设计跨模态特征映射模块生成伪正样本对，构建伪身份空间，结合静态 - 动态协同学习策略对齐簇级与实例级模态差异，在无配对无标注场景下实现高效跨模态匹配；Asurada 等 [47]2025 年提出 X-ReID 框架，通过跨模态原型协作模块对齐整合不同模态特征，搭配多粒度信息交互模块融合短时长时跨帧信息与跨模态对齐特征，生成鲁棒的序列级表示，在 HITSZ-VCM 等基准数据集上表现优异；Visuang 等 [48]2025 年提出视频级语言驱动（VLD）框架，通过模态不变语言提示模块生成共享文本提示并对齐视觉特征，结合时空提示模块聚合身份相关时空信息，以轻量化设计实现性能突破；Zhang 等 [49]2025 年提出跨模态与时序协作（CMTC）网络，设计事件转换网络提取辅助信息，通过差异模态协作模块平衡事件与辅助信息的互补作用，搭配时序协作模块挖掘运动与外观线索，优化事件基视频重识别性能；VCM Project 团队 [50]2025 年构建大规模 RGB-IR 视频数据集，提出模态不变子空间投影与运动不变时间记忆提取相结合的方法，充分利用视频帧间冗余信息，验证了视频 - 视频匹配在跨模态场景中的优势。

3.5 模态差异消除与特征分布对齐

模态差异消除是跨模态行人（车辆）重识别的核心挑战，2024-2025 年相关英文研究通过创新特征分离、局部对齐等策略高效缩小模态鸿沟：Seo 等 [51]2024 年提出基于关键点掩码与擦除的特征表示方法，通过关键点掩码归一化局部特征表示以减少特征级模态差异，搭配关键点擦除提升全局特征多样性与有效性，在 SYSU-MM01 数据集上验证了对相似外观行人匹配的优化效果；Shi 等 [52]2024 年提出跨模态信息瓶颈表示学习网络（CM InfoNet），通过互信息瓶颈权衡筛选身份相关信息、压缩冗余内容，设计模态共识模块对齐可见光与红外特征，同时获取全局 - 局部特征以强化关键部位判别，有效消除跨模态与模态内变异；MCGS-ReID 团队 [53]2024 年提出模态交叉图采样（MCGS）方法，助力网络挖掘更多跨模态信息，搭配多模态共享特征对齐网络强化跨模态共享特征表示，实现不同模态特征精准对齐，在 VT-Vehicle 等数据集上显著缓解成像原理与光谱差异导致的模态鸿沟；表 2 是这几种方法的列举。

表 2: 跨模态行人重识别技术路线对比

技术类别	核心思路	关键创新点
特征对齐与融合	分离共享/特异性特征，通过多粒度/频域融合强化一致性	频域-空域协同、分层聚合
注意力与特征增强	聚焦关键区域，抑制冗余，强化模态间通道交互	跨通道注意力、内容感知机制
语义与无监督学习	语义对齐 + 伪标签生成，降低标注依赖	CLIP 改进、MLLMs 语义细化
视频跨模态融合	挖掘时空信息，协同静态外观与动态运动线索	多粒度时空交互、运动不变记忆
模态差异消除	特征解纠缠/局部对齐，缩小模态分布偏移	关键点掩码、信息瓶颈筛选

4 挑战与未来展望

4.1 核心挑战

跨模态行人重识别技术虽已形成多元化技术体系，但在实际应用与技术深化过程中仍面临四大核

心挑战。其一，模态鸿沟的根本性缓解难题，不同模态成像原理差异导致特征分布偏移显著，现有特征对齐与差异消除方法多聚焦于静态分布适配，难以应对动态场景下（如光照突变、姿态剧烈变化）的模态特征波动，细粒度语义层面的跨模态关联挖掘仍不充分。其二，多模态融合策略的优化瓶颈，当前方法多采用固定权重或简单注意力分配机制，未能实现模态互补价值的动态适配，在复杂环境下（如重度遮挡、低分辨率）易出现有效信息丢失或冗余信息干扰的问题。其三，低资源场景的技术适配不足，无监督、少样本学习方法依赖大量伪标签生成与聚类优化，但其泛化能力受数据分布影响显著，在跨数据集、跨场景迁移中性能衰减严重，标注成本与识别精度的平衡尚未实现。其四，实际部署的工程化挑战，现有先进方法多依赖复杂网络架构与大规模预训练，存在计算开销大、实时性不足的问题，且对无人机 - 地面相机等多平台跨模态场景的适配性较差，难以满足智能安防的实战需求。

4.2 未来展望

针对上述挑战，结合 2024-2025 年技术发展趋势，跨模态行人重识别的未来研究可聚焦四大方向。一是多模态大模型的深度融合，将 CLIP 等视觉 - 语言预训练模型与跨模态 ReID 任务深度适配，通过 Prompt Tuning 等技术强化模态间语义对齐，挖掘文本 - 视觉、红外 - 可见光的细粒度关联，同时借助 MLLMs 提升人机交互下的语义细化能力。二是自适应融合与动态差异消除机制，设计基于环境感知的模态权重分配网络，结合频域 - 空域协同对齐策略，实现对动态场景下模态差异的实时适配；探索生成式 AI 与特征解纠缠的结合，通过生成高质量跨模态样本缩小分布鸿沟。三是低资源学习的技术突破，优化无监督聚类与伪标签生成策略，引入对比学习与自监督预训练的协同机制，提升模型在少样本、跨数据集场景下的泛化能力；发展半监督与弱监督学习范式，降低对大规模标注数据的依赖。四是跨平台适配与工程化优化，针对多摄像头、多平台场景设计轻量化网络架构，通过模型压缩与量化技术提升实时性；构建更贴近实战的多模态数据集，涵盖复杂光照、多视角、低分辨率等场景，推动技术从实验室走向实际应用。

5 结论

多模态行人重识别作为智能安防与计算机视觉领域的核心技术，旨在突破单模态识别的场景局限，通过融合可见光图像、红外图像、文本语义、视频序列等多源模态信息，实现复杂环境下的精准行人身份匹配，为全天候智能监控、无图寻人、动态追踪等任务提供核心支撑。本文系统梳理了该领域的发展历程、基础框架与核心技术，形成了全面的综述体系。在基础框架层面，本文明确了多模态 ReID 的核心概念与模态特性，划分涵盖可见光 - 红外、文本 - 图像、视频 - 单帧的单模态跨模态任务，以及多源融合任务。建立了以 Rank-k 准确率、mAP 为核心，MMD 为辅助的评价指标体系，为技术对比与性能衡量提供标准。在核心技术层面，本文按技术路线将现有方法划分为五大类：特征对齐与多维度跨模态融合通过多尺度、频域聚合等策略强化特征一致性；多样化注意力机制与特征增强借助多样化注意力机制聚焦关键信息；语义增强与无监督半监督学习通过语义挖掘与低监督学习降低标注依赖；视频跨模态融合与时空信息协同利用时空信息提升动态场景鲁棒性；模态差异消除与特征分布对齐通过特征分离、局部对齐等策略缩小模态鸿沟。各类方法从不同维度缓解了模态异质性带来的挑战，在 SYSU-MM01、RegDB、CUHK-PEDES 等主流数据集上验证了有效性。尽管技术取得显著进展，但模态鸿沟的根本性缓解、融合策略的动态适配、低资源场景的泛化能力、工程化部署的实时性等挑战仍未完全解决。未来，通过多模态大模型融合、自适应融合机制、低资源学习优化与跨平台适配等方向的深入研究，跨模态行人重识别技术将进一步提升实用价值，为智能安防体系的完善、公共安全保障能力的提升提供更坚实的技术支撑，推动该领域从实验室研究走向规模化实战应用。

参考文献

[1] 王素玉，肖塞。行人重识别研究综述 [J]. 北京工业大学学报，2022, 48 (10): 1100-1112.

[2] Zheng L, Yang Y, Hauptmann A G. Person re-identification: past, present and future[EB/OL].
[3] 李擎，胡伟阳，李江昀，等。基于深度学习的行人重识别方法综述 [J]. 工程科学学报，2022, 44 (5): 920-932.
[4] Ahmed E, Jones M, Marks T K. An improved deep learning architecture for person reidentification[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 3908-3916.
[5] 叶钰，王正，梁超，等。多源数据行人重识别研究综述 [J]. 自动化学报，2024, 46 (9): 1869-1884.
[6] Zajdel R, Komorowski J, Napieralski A. Person re-identification using stereo vision[C]//2005 IEEE International Conference on Robotics and Automation. Piscataway: IEEE, 2005: 3026-3031.
[7] Yi D, Lei Z, Liao S, et al. Deep metric learning for person re-identification[C]//2014 IEEE International Joint Conference on Biometrics. Piscataway: IEEE, 2014: 1-8.
[8] 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. Piscataway: IEEE, 2017: 5390-5399.
[9] He L, Zheng Z, Zhang S, et al. Harmonious attention network for person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2285-2294.
[10] Yu H X, Zheng W S, Wu A, et al. Unsupervised person re-identification by soft multilabel learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 2143-2152.
[11] 张明，李静。多模态行人重识别研究综述 [J]. 计算机应用研究，2025, 42 (8): 2321-2328.
[12] 王磊，赵阳。多模态融合行人重识别技术研究进展 [J]. 模式识别与人工智能，2024, 37 (9): 812-821.
[13] Li W, Zhang Y, Wang C. Dual-path convolutional image-text embeddings with instance loss for text-based person retrieval[C]//2017 ACM Multimedia Conference. New York: ACM, 2017: 1544-1551.
[14] 周勇，王瀚正，赵佳琦，等。基于可解释注意力部件模型的行人重识别方法 [J]. 自动化学报，2023, 49 (10): 2159-2171.
[15] Ye M, Shen J, Lin G, et al. Deep learning for person re-identification: a survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872-2893.
[16] 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.
[17] Li X, Li W, Chen D, et al. CUHK-PEDES: A large-scale dataset for pedestrian text retrieval[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6097-6106.
[18] Lin Y, Zheng L, Zheng Z, et al. VideoReID-MM: A large-scale benchmark for video-based cross-modal person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(4): 1689-1703.
[19] Chen S, Liu J, Zhang H, et al. CASIA Gait Dataset B: A multi-modal gait database for person re-identification[J]. Pattern Recognition, 2019, 90: 187-198.
[20] Lin Y, Zheng L, Zheng Z, et al. A comprehensive evaluation of person re-identification metrics[J]. Pattern Recognition, 2023, 139: 109452.
[21] Liu J, Sun Y, Zhu F, et al. Learning memory-augmented unidirectional metrics for cross-modality person re-identification[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recogni-

tion. Piscataway: IEEE, 2022: 19366-19375.
[22] Li X, Wang Y, Zhang Z. Cross-Modal Feature Alignment via Dual Attention Guided Fusion for Person Re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 12456-12465.
[23] 程德强，钱建生，郭星歌，等。基于多粒度融合和跨尺度感知的跨模态行人重识别 [J]. 通信学报，2025, 46 (1): 110-123.
[24] Zhang H, Li J, Chen W. Hierarchical Feature Fusion with Modal Consistency Constraints for Cross-Modal Person ReID[C]//Proceedings of the European Conference on Computer Vision. 2024: 8912-8921.
[25] 金静。基于频域空间信息的特征聚合跨模态行人重识别方法 [J]. 自动化学报，2024, 50 (9): 1987-1996.
[26] Wang S, Liu Q, Zhao L. Adaptive Feature Alignment and Progressive Fusion Network for Multimodal Person Re-identification[J]. IEEE Transactions on Image Processing, 2025, 34(2): 789-801.
[27] 刘锁兰，孔立智，王洪元。基于三重嵌入扩展和特征聚合的跨模态行人重识别 [J]. 模式识别与人工智能，2024, 37 (7): 621-630.
[28] Chen M, Sun K, Hu F. Frequency-Spatial Co-Alignment and Fusion for Cross-Modal Person ReID[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025: 6789-6796.
[29] 朱沛伍，高树辉。低高频多尺度融合的跨模态行人重识别研究 [J]. 重庆邮电大学学报 (自然科学版), 2024, 36 (6): 1183-1192.
[30] Zhang Y, Li M, Wang H. Visible-infrared person re-identification with region-based augmentation and cross modality attention[J]. Journal of Biomedical Informatics, 2025, 142: 104987.
[31] 郭思琦，李明杰，张宇。基于辅助模态和注意力机制的跨模态行人重识别 [J]. 计算机应用研究，2024, 41 (8): 2415-2420.
[32] Liu C, Zhang J, Chen L. Cross-modulated Attention Transformer for RGBT Tracking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(7): 6890-6898.
[33] 杨真真，王浩，赵阳。基于混合卷积增强和内容感知注意力的跨模态行人重识别 [J]. 模式识别与人工智能，2024, 37 (10): 921-929.
[34] Kim S, Park J, Lee D. Attention-enhanced multimodal feature fusion network for clothes-changing person re-identification[J]. Complex & Intelligent Systems, 2024, 10(5): 4871-4885.
[35] 黄驰涵，陈雨，吴敏。基于融合注意力和特征增强的跨模态行人重识别 [J]. 电子与信息学报，2025, 47 (2): 589-597.
[36] 何磊，张伟，刘静。跨通道交互注意力机制驱动的双流网络跨模态行人重识别 [J]. 自动化学报，2024, 50 (11): 2432-2441.
[37] 陈梦蝶，李然，王强。基于双重增强网络的跨模态行人重识别 [J]. 计算机工程与应用，2025, 61 (3): 128-135.
[38] Liu H, He L, Zhang Y. Semantic-Aligned Learning with Collaborative Refinement for Unsupervised VI-ReID[J]. International Journal of Computer Vision (IJCV), 2025, 133(4): 987-1005.
[39] 董令赟。基于跨模态语义一致性和半监督的跨模态行人重识别算法研究 [J]. 计算机工程与应用，2024, 60 (18): 145-153.
[40] Pang Y, Chen J, Li M. Augmented and Softened Matching for Unsupervised Visible-Infrared Person Re-Identification[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2025: 7632-7641.

[41] 宋存利。基于语义增强网络的跨模态行人重识别 [J]. 模式识别与人工智能，2025, 38 (2): 156-164.
[42] Qin Y, Yang S, Wang Z. Human-centered Interactive Learning via MLLMs for Text-to-Image Person Re-identification[EB/OL]. [2025-06-11]. arXiv preprint arXiv:2506.11036.
[43] 贾军营。改进 CLIP-ReID 的跨模态行人重识别 [J]. 计算机应用研究，2024, 41 (10): 3078-3083.
[44] Wang C, Zhang L, Liu J. Multi-Memory Matching for Unsupervised Visible-Infrared Person ReIdentification[EB/OL].[2024-01-06]. arXiv preprint arXiv:2401.06825.
[45] Zhang S, Li K, Zhao H. Exploring Homogeneous and Heterogeneous Consistent Label Associations for Unsupervised Visible-Infrared Person ReID[EB/OL]. [2025-02-02]. arXiv preprint arXiv:2402.00672.
[46] Yao Y, Li J, Wang Z. Unsupervised Visible-Infrared Person Re-identification under Unpaired Settings[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2025: 4567-4576.
[47] Asurada Y, Chen S, Liu M. X-ReID: Multi-granularity Information Interaction for Video-Based Visible-Infrared Person Re-identification[EB/OL]. [2025-11-17]. arXiv preprint arXiv:2511.17964.
[48] Visuang T, Zhang L, Chen Y. Video-Level Language-Driven Video-Based Visible-Infrared Person Re-identification[EB/OL]. [2025-06-02]. arXiv preprint arXiv:2506.02439.
[49] Zhang H, Liu Q, Li D. Event-based Video Person Re-identification via Cross-Modality and Temporal Collaboration[EB/OL]. [2025-01-07]. arXiv preprint arXiv:2501.07296.
[50] VCM Project Team. Learning Modal-Invariant and Temporal-Memory for Video-based Visible-Infrared Person Re-Identification[EB/OL]. [2025-08-02]. arXiv preprint arXiv:2208.02450.
[51] Seo J, Kim D, Lee S. Keypoint Mask-based Local Feature Matching and Keypoint Erasingbased Global Feature Representations for Visible-Infrared Person Re-Identification[C]//Proceedings of Machine Learning Research. 2024, 263: 3546-3561.
[52] Shi H, Luo M, Zhang X Y, et al. Learning Cross-modality Information Bottleneck Representation for Heterogeneous Person Re-identification[EB/OL]. [2024-08-15]. arXiv preprint arXiv:2308.15063.
[53] Zhang Y, Li W, Wang C. MCGS-ReID: A Visible-Infrared Vehicle Reidentification Method Using Modal-Cross Graph Sampler[J]. IEEE Access, 2024, 12: 108012-108015.