

中山大学硕士学位论文

中文题目： 基于模态不变特征学习的近红外-可见光人脸识别研究

英文题目： Research of NIR-VIS Face Recognition Based on Modality-invariant Feature Learning

专 业： 电子信息

学位申请人： 杨奕铭

导 师 姓 名： 胡海峰 教授

论文答辩委员会主席： _____

成员： _____

二〇二六年九月六日

原创性及学位论文使用授权声明

本人郑重声明: 所呈交的学位论文, 是本人在导师的指导下, 独立进行研究工作所取得的成果。除文中已经注明引用的内容外, 本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体, 均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名:

日期: 年 月 日

本人完全了解中山大学有关保留、使用学位论文的规定, 即: 学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版; 有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅; 有权将学位论文的内容编入有关数据库进行检索; 可以采用复印、缩印或其他方法保存学位论文; 可以为建立了馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

保密论文保密期满后, 适用本声明。

学位论文作者签名:

日期: 年 月 日

导师签名:

日期: 年 月 日

论文题目： 基于模态不变特征学习的近红外-可见光人脸识别研究

专业： 电子信息

硕士生： 杨奕铭

指导教师： 胡海峰 教授

摘 要

近红外-可见光 (Near-InfraRed and VISible, NIR-VIS) 人脸识别能够精确匹配同一身份的 NIR 人脸图像与 VIS 人脸图像。作为异构人脸识别的典型子任务，NIR-VIS 人脸识别能够克服低光条件下人脸识别的瓶颈。因此，本项技术在公共安全、犯罪调查、多媒体信息检索等领域中都有广泛的研究和应用前景。然而，面对不断涌现的社会挑战，NIR-VIS 人脸识别仍然存在待解决的问题。首先，由于成像原理和相机参数设置的不同，不同模态人脸图像之间存在明显的模态差异。其次，异质人脸图像在图像质量上参差不齐，具体表现在姿态、遮挡、表情、照明等附加的表观差异，也被称为剩余差异。最后，从大量跨模态人脸数据中获取注释信息是耗时且昂贵的，注释问题直接限制了数据集规模和导致过拟合问题，这不利于有监督 NIR-VIS 人脸识别方法在更大规模真实场景中的扩展性和可行性。针对上述问题，本文的主要工作和贡献总结如下：

- 1) **融合空间学习方法:** 针对模态差异问题，提出了一种新颖的融合空间学习方法。首先，融合模态生成器通过浅层 CNN 的通道级卷积将 NIR 和 VIS 图像合成为融合图像。特别是，它能够很好地保留图像中的判别性结构信息，并且可以通过自监督方式进一步提高图像的质量。其次，模态对抗融合空间学习通过融合模态对抗学习策略将 NIR 和 VIS 图像投影到融合空间。采用融合模态引导学习的目标函数，有效减少 NIR-VIS 人脸中存在的模态差异。最后，由 NIR-融合、融合-融合、VIS-融合一致性构成的融合分布一致性可以增强类内紧凑性，有助于学习判别性表示。实验分析验证了所提出的方法能够有效解决 NIR-VIS 人脸识别任务中存在的模态差异问题并提高模型性能。
- 2) **中立人脸学习和渐进融合合成方法:** 针对模态差异和剩余差异问题，提出了一种中立人脸学习和渐进融合合成方法。首先，身份相关中立人脸学习模块通过引导网络学习同模态的中立人脸特征表示，将跨模态图像映射到共同的中立表示子空间以解决模态差异和剩余差异的挑战。其次，属性渐进融合模块以渐进的方式进

行身份相关特征、模态相关特征、剩余特征的解耦和融合，进一步净化身份相关的特征。在三个具有挑战性的 NIR-VIS 数据库上的广泛实验验证了所提出方法的有效性。

- 3) **鲁棒跨模态伪标签和对比学习方法:** 针对获取注释困难问题，提出了鲁棒跨模态伪标签和对比学习方法网络，实现无监督设置下的 NIR-VIS 异构人脸识别。首先，基于 NIR 集群的伪标签共享模块估计准确的伪标签作为训练所需的监督信号。其次，模态域特定集群对比学习增强了模态内特征的区分能力。最后，跨模态域集群对比学习引导网络学习可靠且模态无关的特征表示。与有监督方法相比，本方法实现了媲美有监督方法的性能。这些结果表明所提出的方法更适用于大规模、复杂且缺乏手工注释的实际场景。
- 4) **同质到异质学习和剩余不变增强方法:** 针对模态差异、剩余差异、注释问题，提出了同质到异质学习和剩余不变增强方法。首先，模态-对抗对比学习充分利用模态对比学习和对抗学习的协同作用，缩小模态之间的分布差距，并在每个模态内学习有区别性的表示，即使模态之间的身份关系不明确，也能有效减小模态差异。其次，基于 K-相互编码的跨模态标记被引入作为伪标签估计方法，通过探索 K-相互最近邻将跨模态特征分组成集群以分配伪标签。接着，精细跨模态对比学习采用模态不变的平均初始化和动态关注加权策略，有效提取模态不变的全局特征。最后，考虑到剩余差异影响使用全局特征计算相似性的精度，剩余不变表示增强在跨模态人脸中挖掘局部特征进行精细化的匹配。与有监督方法相比，所提出的方法展现出很大的潜力，不仅能够减低标签采集需求和成本以提供更大的便利性和实用性，还能够达到媲美有监督方法的性能。

[关键词] 近红外-可见光人脸识别; 异构人脸识别; 模态不变特征; 无监督学习

Title: Research of NIR-VIS Face Recognition Based on Modality-invariant Feature Learning

Major: Electronic Information

Name: Yiming Yang

Supervisor: Prof. Haifeng Hu

ABSTRACT

The Near-Infrared and Visible (NIR-VIS) face recognition is able to accurately match NIR face images with VIS face images of the same identity. As a typical sub-task of heterogeneous face recognition, NIR-VIS face recognition is able to overcome the bottleneck of face recognition under low-light conditions. Therefore, it has broad research and application prospects in areas such as public safety, criminal investigation, and multimedia information retrieval. However, in the face of emerging social challenges, NIR-VIS face recognition still faces unresolved issues. Firstly, due to different imaging principles and camera parameter settings, there are significant modality differences between different modal face images. Secondly, heterogeneous face images have uneven image quality, manifested in apparent differences in pose, occlusion, expression, lighting, etc., known as residual differences. Lastly, obtaining annotation information from a large amount of cross-modal face data is time-consuming and expensive, and labeling issues directly limit the dataset's scale, leading to over-fitting problems, which in turn limit the scalability and feasibility of supervised NIR-VIS face recognition methods in larger real-world scenarios. In response to the above issues, the main work and contributions of this paper are summarized as follows:

- 1) **Syncretic space learning method:** Considering modality differences, this thesis proposes a novel syncretic space learning method. First, syncretic modality generator synthesizes NIR and VIS images into syncretic images by channel-level convolution with a shallow CNN. In particular, the discriminative structural information in the images is well preserved and the quality can be further improved in a self-supervised manner. Second, modality-adversarial syncretic space learning projects NIR and VIS images into the syncretic space by a syncretic-modality adversarial learning strategy with syncretic pattern guided objective, so the modal gap existing in NIR-VIS faces can be effectively reduced.

Finally, the syncretic distribution consistency constructed by NIR-syncretic, syncretic-syncretic, and VIS-syncretic consistency can enhance the intra-class compactness, which is effective for learning discriminative representations. Experimental analysis verifies that the proposed method can effectively address the modality gap issue in NIR-VIS face recognition tasks and improve model performance.

- 2) **Neutral face learning and progressive fusion synthesis method:** Aiming at modality differences and residual differences, this thesis proposes a neutral face learning and progressive fusion synthesis method. Firstly, identity-related neutral face learning guides the network to learn homogeneous neutral face feature representations, which tackles the challenge of modality and residual differences by mapping cross-modal images to a common neutral representation subspace. Besides, attribute progressive fusion is presented to perform the disentanglement and reintegration of identity-related features, modality-related features and residual features in a progressive fusion manner, which helps to further purify identity-related features. Extensive experiments on three challenging HFR databases verify the effectiveness of the proposed method.
- 3) **Robust cross-domain pseudo-labeling and contrastive learning method:** Aiming at the difficulty of obtaining annotations, this thesis proposes a robust cross-domain pseudo-labeling and contrastive learning method, which achieves NIR-VIS heterogeneous face recognition in an unsupervised setting. Firstly, NIR cluster-based pseudo labels sharing estimates accurate pseudo labels for guiding the network to learn domain-invariant representations. Secondly, domain-specific cluster contrastive learning enhances the discriminativeness of individual features within the modality. Finally, inter-domain cluster contrastive learning guides the network to learn robust and modality-independent representations. In comparison with supervised methods, our approach achieves comparable performance, demonstrating that our method is more suitable for real scenarios that are large, complex, and lack manual annotation.
- 4) **Homogeneous-to-heterogeneous learning and residual-invariant enhancement method:** Considering modality differences, residual differences and annotation difficulty, this the-

sis proposes homogeneous-to-heterogeneous learning and residual-invariant enhancement method. Firstly, modality-adversarial contrastive learning leverages the collaborative power of modality contrastive learning and adversarial learning to narrow the distribution gap between modalities and learn discriminative representations within each modality, effectively reducing modal differences even when identity relationships between modalities are unclear. K-reciprocal-encoding-based cross-modal labeling is introduced as robust pseudo label estimation to group cross-modal features into clusters by exploring k-reciprocal nearest neighbors. Additionally, refined cross-modal contrastive learning is developed with modality-invariant averaging initialization and dynamic focus weighting strategies to effectively extract modality-invariant features. Finally, considering that residual variations can affect the accuracy of similarity computation using global features, residual-invariant representations enhancement mines partial features under the cross-modal face for robust matching. Compared to supervised methods, the proposed method demonstrates significant potential. It not only reduces the need for labels collection and costs to provide greater convenience and practicality but also achieves performance comparable to supervised methods.

[Keywords] NIR-VIS face recognition, Heterogeneous face recognition, Modality-invariant feature, Unsupervised learning

目录

摘要	I
ABSTRACT	III
第 1 章 绪论	1
1.1 研究背景和意义	1
1.2 NIR-VIS 人脸识别框架	5
1.3 国内外研究现状	6
1.4 NIR-VIS 人脸数据集	14
1.5 NIR-VIS 人脸识别性能评估指标	17
1.6 论文研究内容	18
1.7 论文组织架构	20
第 2 章 融合空间学习方法	23
2.1 引言	23
2.2 融合空间学习模型	26
2.3 实验结果与分析	34
2.4 本章小结	43
第 3 章 中立人脸学习和渐进融合合成方法	45
3.1 引言	45
3.2 中立人脸学习和渐进融合合成模型	48
3.3 实验结果与分析	56
3.4 本章小结	64

第 4 章 鲁棒跨模态伪标签和对比学习方法	65
4.1 引言	65
4.2 鲁棒跨模态伪标签和对比学习模型	69
4.3 实验结果与分析	77
4.4 本章小结	85
第 5 章 同质到异质学习和剩余不变增强方法	87
5.1 引言	87
5.2 同质到异质学习和剩余不变增强模型	92
5.3 实验结果与分析	100
5.4 本章小结	108
第 6 章 总结与展望	111
6.1 工作总结	111
6.2 研究展望	113
参考文献	115
攻读硕士学位期间相关的科研成果目录	131
致谢	133
中山大学学位论文版权使用授权书	135

第 1 章 绪论

1.1 研究背景和意义

随着信息化时代到来，许多重要的安全和隐私领域都需要准确确认个人身份。生物特征因其不可复制性，在学术和工业界广受关注。目前的生物特征识别方法可分为几个关键方向，包括人脸、静脉、虹膜、视网膜、指纹、DNA 等。其中人脸识别 (Face Recognition, FR) 技术具有最广泛的应用背景^[1-4]。这是因为人脸识别技术有着无需接触、安全可靠、易于采集等特点，因此适用于各种不可控的安全监控场景下^[5-8]。随着科技的进步，计算机的算力和数据集规模迅速提高，人脸识别技术也随着科技浪潮取得重大突破，越来越被大众认可和接受^[9,10]。

为了适应日新月异的人脸识别技术，国家相关部门出台了一系列政策，引导和规范人脸识别技术的应用。国务院于 2017 年 7 月印发并实施的《新一代人工智能发展规划》强调大力发展人脸识别等人工智能新兴产业。2021 年 11 月正式施行的《中华人民共和国个人信息保护法》促进了人脸等个人信息的规范使用。2022 年 11 月，国家标准 GB/T41772-2022《信息技术生物特征识别人脸识别系统技术要求》对人脸识别技术实现的行业标准有重要的指导意义。2023 年 8 月，国家互联网信息办公室起草的《人脸识别技术应用安全管理规定(试行)(征求意见稿)》对人脸的使用和存储方法进行了规定。此外，为促进人脸识别相关技术的研究，多项科研基金被投入到了相关课题研究，包括国家自然科学基金、国家重点研发计划、国家科技支撑计划等。在以上相关政策和基金的支持下，人脸识别技术被广泛应用于人们的日常生活中。支付宝、微信、银联等陆续为自家移动支付平台推出刷脸支付功能。小米、华为等知名科技公司推出人脸识别解锁的智能门锁等产品，用于打造安全智慧生态家居。各大手机厂商将人脸解锁作为手机解锁的主推手段，其中苹果在疫情期间推出了全新的口罩人脸解锁功能。在安防领域，人脸识别技术展现出了重大作用。2019 年，广西阳朔警方在“天网”平台的人脸识别技术帮助下，成功锁定了逃亡 25 年的犯罪嫌疑人。在备受瞩目的劳荣枝案中，公安部门通过大规模人像比对最终将嫌疑人抓获。腾讯优图实验室推出的跨年龄人脸识别技术，仅凭

几张婴幼儿时期的模糊照片就帮助 4 位走失 10 年的孩子与父母团聚。



图 1-1 不同类别人脸示例图片，图片来自文献^[11,12]

随着低成本、高分辨率的可见光 (VISible, VIS) 相机日益普及，当前人脸识别技术重点关注于可见光人脸的识别，即同构 (质) 人脸识别，又称为同模态人脸识别。随着科技与社会的不断进步，人们对人脸识别的需求日益多样化，单一模态的人脸识别已难以满足需求。为了克服这一限制，异构 (质) 人脸识别 (Heterogeneous Face Recognition, HFR) 应运而生，其目的在于匹配不同模态的人脸图像，也被称为跨模态人脸识别。如图 1-1 所示，常见的异构人脸识别可大致分为五个类别。近红外-可见光 (Near InfraRed-VISible, NIR-VIS) 人脸识别技术在昏暗环境中显示出巨大潜力。在黑暗场景中，由于光照不足，可见光摄像头难以捕捉清晰的人脸图像。一些监控摄像头能够智能地在黑暗环境中切换到近红外模式，获取细节丰富且可辨认的近红外图像。此时，近红外图像可以用于与可见光图像进行比对来确认身份。素描-可见光 (Sketch-VIS) 人脸识别常用于刑侦场景。在一些案件中，犯罪嫌疑人未被摄像头捕捉到，需要请专业法医素描师根据受害人口述绘制大致的脸部特征形成素描图片。刑侦人员可使用素描图片缩小人员筛查范围。热红外-可见光 (Thermal InfraRed-VISible, TIR-VIS) 人脸识别与 NIR-VIS 人脸识别类似。热辐射不受环境光照的影响，在较暗环境中，可以获取适用于夜间人脸识别的热红外人脸图像。传统的二维人脸识别系统易受表情、姿态、遮挡等因素影响，三维-二维 (3D-2D) 人脸识别可用于获取姿态和遮挡鲁棒的三维几何表示以供人脸识别。传统的人脸识别方法在应对漫画-可见光人脸识别中极端的类内变化时常遇到困难。相比之下，人眼在处

理远离正常人脸分布的漫画时表现出显著的识别能力，强调了人们在漫画中更容易辨认熟悉的面部特征。因此，在漫画-可见光人脸识别领域，采用模仿人类对漫画理解的编码方式有助于人脸识别系统的完善^[11]。

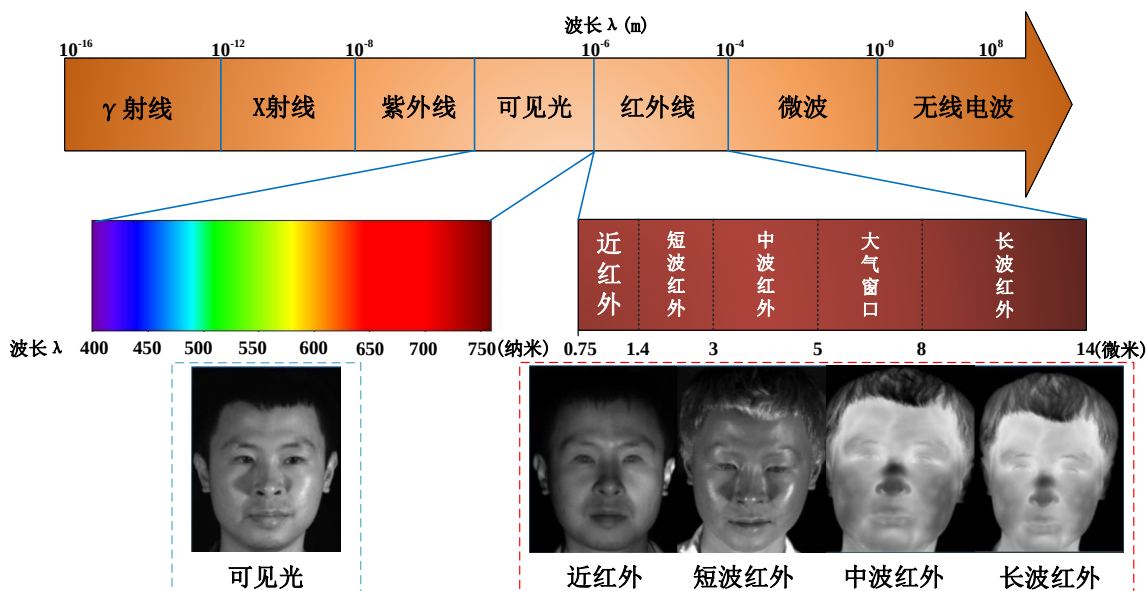


图 1-2 电磁波谱，根据波长的不同，人脸在可见光谱和红外光谱中的成像有着明显的差异，图片来自文献^[11,12]



图 1-3 在不同角度光照照射下的可见光(上)和近红外(下)人脸成像，图片来源于文献^[13]

近年来，国家对于公共安全的危机预防重视程度越来越高，许多公共场合都需要 24 小时稳定运作的人脸识别系统。人脸在近红外设备中的成像很好地弥补了可见光人脸识别在夜间的缺陷，NIR-VIS 人脸识别逐渐成为最具有代表性和最受关注的人脸识别技术之一^[14,15]。具体来说，红外成像系统通过观测物体表面反射或发出的辐射或热量来

获取成像^[16-18]。如图1-2所示, 根据电磁波谱的波长, 可以观察到可见光的波长大约在400nm到750nm之间, 红外光谱则在0.75 μm 到14 μm 之间。其中, 红外光谱可以根据波长进一步细分为近红外、短波红外 (Short-Wave InfraRed, SWIR)、中波红外 (Middle-Wave InfraRed, MWIR) 以及长波红外 (Long-Wave InfraRed, LWIR)。具体地, 近红外的波长在0.75-1.4 μm 之间, 短波红外的波长在1.4-3 μm 之间, 中波红外的波长在3-5 μm 之间, 长波红外的波长在8-14 μm 之间。根据红外光谱的特性, 这四者可以分为有源红外 (Active Infrared) 和无源红外 (Passive Infrared)。有源红外需要借助外部光源照射材料表面才能获得成像, 因此需要不可见的红外光来照射场景, 近红外和短波红外属于有源红外的范畴。无源红外包含波长大于3 μm 的中波红外和长波红外。这一波段的红外线具有明显的发射性, 可以设计热传感器来捕捉由于面部血管系统引起的面部皮肤组织的温度变化^[16]。然而, 由于中波红外和长波红外的波长范围远离可见光波段, 在成像结果上与可见光图片有着显著的不同, 皮肤表面的细节纹理几乎无法分辨。与此相反, 近红外图像与可见光图像在肉眼的视觉感官上最为相似。此外, 近红外图像对光照变化和环境温度变化不敏感, 更适合用于各种开放、不可控的场景。如图1-3所示, 在不同角度的极端光照条件下, 可见光人脸图像中有着非常明显的明暗变化, 这将使得人脸上的纹理出现巨大差异和不连贯性, 在这种情况下可见光人脸识别可能会出现明显的性能退化。与此不同的是, 在近红外成像中, 人脸的纹理和结构信息被很好地保留了下来, 并且几乎不受外界光源的影响, 最大程度上获取到了清晰可辨的人脸图像。

鉴于近红外对人脸的高质量成像结果, 在许多公共场所, 如小区、银行和公园里的监控摄像头都配备了近红外功能, 确保夜间监控效果。公安机关在调查监控系统时, 需要将近红外人脸图片与二代身份证人脸图像库中的可见光人脸图像进行比对, 确定昏暗场景下嫌疑人的身份。在刑事侦查中, 识别嫌疑人的身份并及时抓捕, 可以更好地保障公众的生命安全。人工进行近红外和可见光人脸的比对需要具有多年经验的公安民警进行判断, 这不仅效率低下还容易延误案件破解的时机。若将可见光人脸识别系统直接应用于 NIR-VIS 人脸识别中会导致严重的性能下降, 出现大量的误判和漏判。这是因为可见光人脸识别系统缺乏跨模态的泛化能力, 无法有效弥补不同模态间的差异^[19,20]。因此, 对 NIR-VIS 人脸识别方法进行研究具有重要意义, 并逐渐成为当前计算机视觉领域的研究热点^[14,21,22]。由于 NIR-VIS 人脸识别涉及图像增强、特征提取、网络设计、损失

设计、域适应和无监督学习等多个技术难题，通过深入研究 NIR-VIS 人脸识别可为其他跨模态任务，如文本-图像检索、跨模态行人重识别、跨年龄人脸识别等提供一定参考价值^[23-26]。

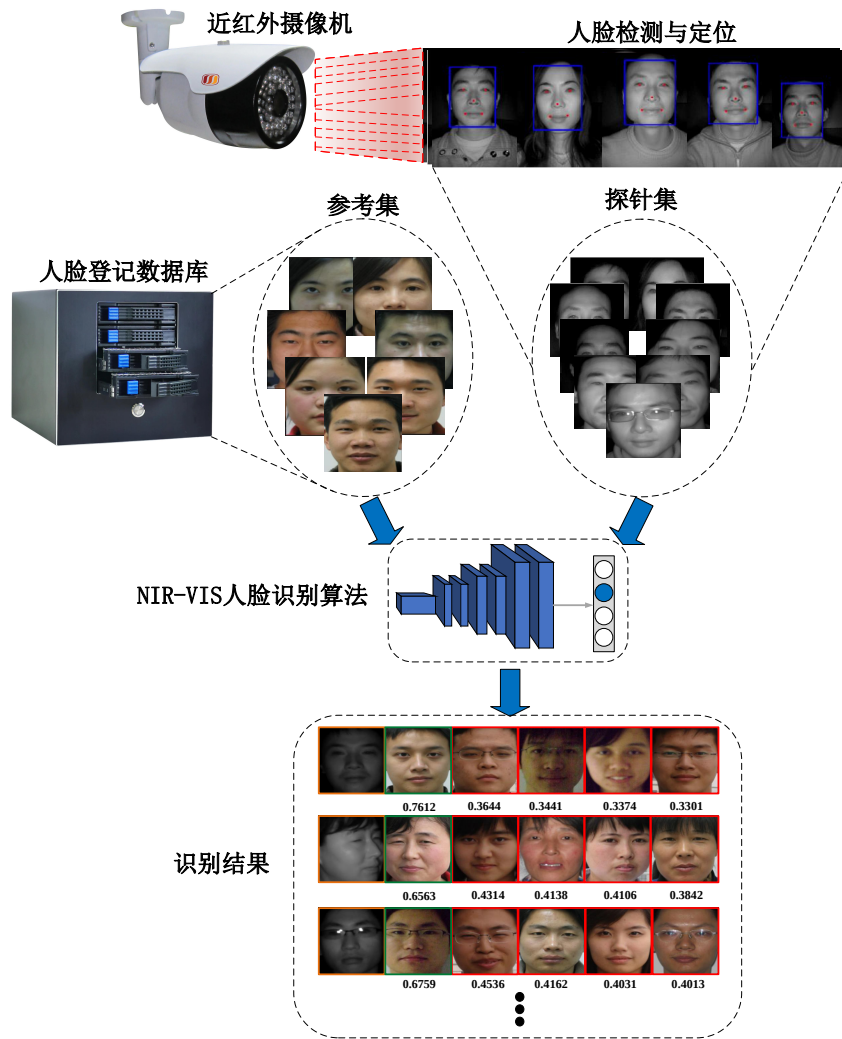


图 1-4 NIR-VIS 人脸识别流程框架

1.2 NIR-VIS 人脸识别框架

NIR-VIS 人脸识别常见框架如图1-4所示，主要包括人脸的检测与定位以及 NIR-VIS 人脸识别算法。首先，通过近红外摄像机捕获到视频流数据，并将其分割为逐帧图像。随后，利用人脸检测算法自动定位图像帧中的人脸位置，对人脸区域进行关键点定位对齐，可以裁剪掉无效干扰区域，得到尺寸统一的近红外人脸图像。这些近红外人脸图像

通常被称为探针集 (Probe set), 因为检测到的人脸可能出现在视频的多帧中, 每个被检测到的人员在探针集中存在多张近红外人脸图像。参考集 (Gallery set) 来自于人脸登记数据库, 如二代身份证人脸图像库, 通常每位个体只有一张或者少量可见光图片登记在库中。将参考集和探针集输入 NIR-VIS 人脸识别网络进行特征提取, 计算每个探针集特征向量与参考集特征向量之间的相似度, 然后根据相似度排序结果和设定的阈值, 最终得出识别结果。

NIR-VIS 人脸识别中的关键在于消除模态差异, 一个直观的思路是在每个关键步骤中引入改进策略来有效消除模态差异。举例来说, 在提取人脸图像时, 可以采用合成或者颜色空间转换方法, 在像素级别上减少模态差异^[27]。在网络结构方面, 可以设计有效的网络层和损失函数, 学习模态不变的映射函数, 从而消除模态差异^[28]。最后, 在匹配阶段, 引入新的度量策略, 结合全局和局部度量方法, 可获得更细致、精确的匹配结果^[29]。通过这些改进策略, 可以有效降低 NIR-VIS 人脸识别中的模态差异。

1.3 国内外研究现状

近年来, 人脸识别技术取得了突破性进展^[7,30,31], 而跨模态人脸识别也在这一背景下取得了长足进步。当前的研究主要集中在如何消除模态差异这一关键问题上, 其核心思想是在特征空间中缩小不同模态数据的分布差异, 提取具有身份鉴别力的特征。跨模态人脸识别可以细分为三类方法: 基于模态不变描述子的方法、基于合成的方法以及基于子空间学习的方法。这些方法为解决 NIR-VIS 人脸识别中的模态差异问题提供了重要的思路。

1.3.1 基于模态不变描述子的方法

基于模态不变描述子的方法核心思想是设计一种对模态差异不敏感、但关注身份信息描述子的方法。这些描述子可用于对异构图片进行编码, 获得特征表示进行匹配。传统的手工描述子包括高斯-拉普拉斯滤波 (Laplacian of Gaussian, LoG)^[32]、局部二进制模式 (Local Binary Pattern, LBP)^[33]、定向梯度直方图 (Histograms of Oriented Gradients, HOG)^[34]、尺度不变特征变换 (Scale Invariant Feature Transform, SIFT)^[35]、高斯差分 (Dif-

ference of Gaussian, DoG)^[36] 等。但它们无法很好地从异构人脸中提取出模态不变的通用表示。针对这一问题, Zhang 等人^[37] 基于耦合信息论编码设计了描述符, 将跨模态图像的局部结构量化为离散编码以提取模态不敏感的特征。Liao 等人^[38] 结合 LBP 算子和 DoG 算子对局部图像结构进行编码, 以获得判别性的局部特征进行匹配。Klare 等人^[39] 采用 SIFT 和多尺度局部二进制模式 (Multiscale Local Binary Pattern, MLBP) 从跨模态图像中提取部分特征, 然后利用判别分析技术来优化特征空间。Liu 等人^[40] 提出使用多尺度 HoG 描述子来提取对光源不敏感的特征集, 并提出融合策略在匹配过程中提高模型性能。Yi 等人^[41] 采用 Gabor 滤波器和受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 来提取异构人脸中的共享表示, 并使用主成分分析 (Principal Component Analysis, PCA) 减少人脸中的异构性和冗余性。Zhu 等人^[32] 提出使用三步转导 (Transduction) 的方法匹配 NIR 和 VIS 人脸, 包括 Log-DoG 滤波、局部编码、均匀特征归一化。文献^[42] 提出了一种基于原型随机子空间的异构人脸识别方法, 其核心在于使用多种图像滤波器分别对探针集和参考集的图像进行滤波, 然后测量探针集和参考集的核相似性。Bhatt 等人^[43] 认为局部纹理存在鉴别性信息, 因此使用拉普拉斯金字塔来提取图像中结构边缘等高频信息, 然后使用加权的卡方距离计算多层拉普拉斯金字塔的分数。Yi 等人^[44] 提出二值化的 LoG 描述子提取人脸有效特征, 使用模板匹配方法对描述子处理后的图片进行小区域的匹配克服异质人脸局部遮挡问题。受到范德华力原理的启发, 文献^[45] 提出局部力模式 (Local Force Pattern, LFP) 描述子。该描述子可以根据像素间的关系来保留反射率分量同时消除图像亮度分量的影响, 局部人脸信息可以得到很好的保留。Roy 等人^[46] 引入局部重力角的概念设计描述子提取光照不变的特征。

自 2015 年以来, 深度学习 (Deep learning) 被引入至跨模态人脸识别任务中, 基于模态不变描述子的方法也逐渐与深度学习相结合。Peng 等人^[47] 提出一种深度局部描述符的学习方法, 使用深度卷积网络从人脸关键点附近的多个局部块中提取特征表示用于学习描述符, 然后设计损失减小跨模态的类内变化。Lu 等人^[48] 考虑到量化二进制码可以抑制表情和照明引起的人脸图像变化, 提出了一种同时进行局部二进制特征学习和编码 (Simultaneous Local Binary Feature Learning and Encoding, SLBFLE) 方法。文献^[49] 提出紧凑二进制人脸描述符特征学习方法, 以无监督的方式将人脸信息表达为低维的二进制向量中, 然后汇聚到直方图特征中形成完整的人脸特征表示用于匹配。Bhattacharya

等人^[50]提出了基于哈希编码的描述子,该描述子在编码过程有突出人脸反射变量的能力,提取出的特征编码用于学习空间映射使得相同身份的样本距离更近。为了阐述深度编码的机理,Roy 等人^[51]提出了一种可解释局部频率二进制模式 (interpretable Local Frequency Binary Pattern, LFrBP) 特征学习方法,并通过浅层深度卷积网络对 LFrBP 特征进行微调,克服跨模态变化。

由于手动设计描述子需要引入大量的先验知识,并且特征的紧凑性不尽人意,模态不变描述子方法在当前主流技术中难以推广。一些描述子在不同模态的跨模态泛化能力和学习能力较差,只适用特定跨模态场景。总的来说,描述子方法大多属于早期阶段的工作,为理解跨模态差异机理奠定了基础。

1.3.2 基于合成的方法

合成方法的核心思想相对直观,它将一个模态的人脸图像合成到另一个模态中,或将来自两个不同模态的图像统一合成到特定的模态中,从而在相同模态域内进行匹配。早在 2003 年,香港中文大学的 Tang 等人^[52]将异构人脸合成视为线性转换过程,采用本征变换对形状和细节信息进行分别转换,并设计了一个结合纹理和形状特征的贝叶斯分类器,实现跨模态人脸识别。Wang 等人^[53]则开发了一种马尔可夫随机场模型,用于在不同尺度上合成局部人脸。Liu 等人^[54]利用局部保持方法实现了素描-可见光人脸的转换,并提出了核非线性判别分析 (Kernel-based Nonlinear Discriminant Analysis, KNDA) 匹配人脸。为了解决合成过程中高频信息丢失的问题,Wang 等人^[55]提出了稀疏特征选择和支持向量回归框架。Peng 等人^[56]利用 DoG、中心环绕分裂归一化 (Center-Surround Divisive Normalization, CSDN)、高斯平滑滤波处理图像提取多重特征表示,然后基于马尔可夫网络来交替优化合成多个局部区域。文献^[57]提出了一种基于转导的方法,优化模型在测试集上的性能。Chang 等人^[58]提出了一种基于多元输出回归的人脸素描合成方法,通过学习局部区域的回归函数和采用相关向量回归机的方法解决了回归问题中的不适应性。Zhang 等人^[59]提出了一种基于支持向量回归的素描照片合成方法,以提升合成图像的高频信息。Xiao 等人^[60]采用嵌入式隐式马尔可夫模型生成跨模态人脸,其优势在于只需少量样本进行学习即可得到高质量的合成结果。Zhong 等人^[61]在隐式马尔可夫模型基础上引入了选择性组合策略,以获得更精细的人脸合成质量。文献^[62]引

入了人脸结构先验, 采用 DoG 描述子处理光照变化, 并提出两种平滑项测量马尔可夫随机场网络上相邻的图像块。Li 等人^[63] 为了满足实时性的人脸合成要求并减少信息丢失, 在混合空间中结合光照不敏感的描述子实现跨模态人脸合成。Liu 等人^[64] 提出了贝叶斯张量推理方法, 捕捉两个模态的子空间信息内在关联以学习两个模态间的图像级映射。Chang 等人^[65] 从稀疏信号表示中得到启发, 在跨模态人脸局部区域提取稀疏的表示, 并通过 L1 范数优化计算稀疏表示系数合成人脸。Gao 等人^[66] 引入了邻域选择机制增强稀疏表示模型, 学习跨模态局部区域的映射, 并应用基于局部区域导数的稀疏表示方法以提高合成照片和素描的质量。为了解决伪影和合成图像中鉴别信息丢失的问题, Ji 等人^[67] 探索了多种回归方法, 如最小二乘法、岭回归法、k 近邻等, 学习跨模态映射关系, 并研究了局部重叠部分大小对合成效果的影响。文献^[68] 将跨模态人脸合成视为同时学习字典和映射函数的问题。为了解决合成的人脸存在错位问题, Zhao 等人^[69] 提出了一种自对准生成器对准两种模态之间的数据分布, 接着使用了一种多尺度的鉴别器获得高质量的合成图像。

随着学术界全面拥抱深度学习的时代到来, 对抗生成网络 (Generative Adversarial Networks, GANs) 的提出极大地推动了跨模态人脸合成的研究。He 等人^[27] 引入了一种跨模态人脸合成网络, 其中包含纹理修复组件和姿态校正组件。纹理修复组件用于生成逼真的纹理, 而姿态校正组件则用于生成正面姿势图像。为解决 NIR-VIS 数据集中的姿态不匹配问题, Yu 等人^[70] 提出了保持姿态的跨模态人脸幻觉框架, 生成像素级保持姿态的 NIR-VIS 人脸对。Duan 等人^[71] 调整了异构图像的 UV 贴图, 实现了光谱转换, 生成具有逼真纹理的 VIS 人脸。为了改进将合成任务建模为“一对一”问题的思路, 近期的方法将合成过程视为“一对多”问题来解决。Luo 等人^[72] 提出了一种风格转换器模块, 实现具有多种参考风格的身份保持图像的生成。Yang 等人^[28] 设计了一个人脸合成模块, 将身份和风格属性随机组合, 丰富异构图像的多样性。文献^[73] 提出了一种对偶变分生成 (Dual Variational Generation, DVG) 框架, 通过从噪声中采样来生成大规模异构数据。Fu 等人^[74] 使用 3D 人脸模型和概率像素损失, 在参数较少的情况下实现姿态和模态的转换。Liu 等人^[75] 提出了异构人脸可解释解纠缠表示 (Heterogeneous Face Interpretable Disentangled Representation, HFIDR), 分别探索了异构人脸识别和生成的身份和模态表示, 并验证了可以交换这两种表示生成不同模态下的人脸相片。Song 等人^[76] 提出了双

路径对抗框架，在图像和特征层面进行对抗学习消除模态差异。具体而言，在图像层面衡量了全局和局部的纹理差异，而在特征层面测量了全局和局部的异构分布差异。文献^[77]借鉴了 Cycle-GAN^[78]的思想，采用循环损失、全局强度损失、局部纹理损失更好地捕捉 NIR 和 VIS 域之间的差异。Pang 等人^[79]以学习原型的方式，合成了不受人脸姿态、遮挡、表情等干扰的正面姿态异构人脸图片。文献^[80]针对人脸合成所需的成对图片难以获取的问题，提出了半监督跨模态合成方法。Zhang 等人^[81]提出了一种少样本训练框架，可以实现跨模态人脸合成。该方法引入了协同对比损失，包括潜在增广对比正则化和空间相关对比正则化，共同保证异构人脸图像的多样性和同一性。Wu 等人^[82]提出了一种示例引导的跨模态人脸幻觉网络，通过模态和结构编码器进行互信息最小化和最大化，获得分解的模态和结构表示，随后进行特征表示的重组合成高保真的跨模态人脸图片。Yu 等人采用自动变分编码器 (Variational AutoEncoder, VAE) 将可见光图像的光谱信息与输入近红外图像的内容信息相结合，生成有助于跨模态人脸识别的逼真图像。Mei 等人^[83]充分利用 VIS 人脸图像中丰富的面部先验信息，在联邦学习的框架下促成了不会数据泄露的多端联合训练。

虽然基于合成的方法在视觉上对齐了两个模态的分布，但还面临着一些挑战：1) 为保证合成效果，GANs 引入了庞大的参数量，导致额外的计算开销。这使得基于合成的方法在对实时性有要求的场景中可能不太适用。2) 在人脸合成过程中，如何抽离高层的属性信息，例如身份和模态信息，用于解耦和匹配，仍然是待解决的问题。3) 将 VIS 模态图片合成为其他模态图片较为简单。反之，将其他模态图片合成为 VIS 图片的合成难度较高。这是因为 VIS 人脸图片是三通道图片而其他模态图片大多为单通道图片。目前的方法通常将 NIR 人脸合成为 VIS 人脸，以减小模态差异，这一过程可能引入额外噪声，导致身份相关信息的丢失。

1.3.3 基于子空间学习的方法

基于子空间学习的方法目的在于将跨模态图像统一投影到一个公共子空间中，然后在特征层面增加类间距离和减小类内距离，实现有效的身份分类。Lin 等人^[84]首次提出公共判别特征提取方法，并且制定了类内紧凑度和类间离散度的学习目标将跨模态图片投影到公共子空间中作比较。Lei 等人^[44]提出了耦合谱回归的子空间学习方法，相比于

描述子方法大大降低了计算复杂度和提升了泛化性。文献^[85]改进了 Lei 等人的框架，提出了改进的耦合谱回归算法，充分探索两个模态间的分布关系以及样本中的判别信息。Gong 等人^[86]提出了多特征规范相关分析方法，目的在于将同一身份的图像投影到子空间中并使得它们的相关性最大化。Kan 等人^[87]提出了多视图判别分析方法，联合多个模态的特征进行分析，并引入了瑞利熵求解不同模态间的子空间变换关系。Mignon 等人^[88]在公共子空间中提出了新的度量学习方法，当与正样本之间的距离大于阈值和负样本对距离小于阈值时，则对其进行约束。为了解决基于欧氏距离和马氏距离度量受到模态差异影响的问题，Huo 等人^[89]运用三元组损失保持类间特征的界限以及类内简单样本的界限，提高可分性和泛化性。Huang 等人^[90]提出判别谱回归模型，将子空间学习分解为最小二乘问题，并在此基础上添加了正则化项保持类内与类间的关系。Jin 等人^[91]挖掘跨模态人脸中的判别结构，并使用基于裕度的度量学习方法对人脸匹配问题进行求解。

由于深度学习在计算机视觉领域大放异彩，近年来许多工作都采用深度学习的方式研究跨模态人脸识别问题。通常来说，这些方法构建卷积网络作为投影至公共子空间的媒介，并通过学习的方式调整卷积网络的参数实现有效的分类。Liu 等人^[92]提出了一种转移学习 (Transfer learning) 方法，基于卷积网络，在大型人脸数据集上进行预训练，然后在跨模态人脸数据上进行微调。这一策略在多个数据集上都取得了大幅度的性能提升。基于解耦的方法是子空间学习中净化身份特征的常用手段。He 等人^[93]将 NIR 和 VIS 人脸投影到三个正交的子空间中，并引入 Wasserstein 距离缩小模态间的差异。Hu 等人^[94]将跨模态人脸建模为两个因素，即模态私有因素和身份相关因素，并提出了一个跨领域因素分离模块消除 NIR-VIS 人脸匹配中的因子耦合干扰。Wu 等人^[95]提出了一种解纠缠变分表示框架解决 NIR-VIS 表示解纠缠和过拟合问题。Yang 等人^[96]引入了对偶去相关网络，以促进身份特征对外部变化的独立性。Hu 等人^[97]在网络中插入解纠缠频谱变化子模块，并提出逐步正交解耦的方式，使身份相关信息与模态相关信息无关。文献^[98]提出将跨模态人脸分解为多种属性，接着引入了跨模态裕度损失保证身份相关特征的学习。此外，还提出了双重对抗性纠缠变化和深度去相关模块，消除身份无关变量的影响。为了利用跨模态人脸中的语义信息，一些方法使用结构、纹理信息辅助跨模态人脸识别。Peng 等人^[99]引入了一种由软语义级特征和轮廓级特征组成的软语义

表示进行跨模态人脸识别。Cho 等人^[100]提出了一个关系图模块探索跨模态人脸中的全局信息，并减少纹理信息的依赖性。Yu 等人^[101]提出了内在结构图对齐模块，其目的在于构建人脸内部结构图以精准对齐跨模态人脸中的局部区域。Mudunuri 等人^[102]提出了一种用于匹配低分辨率 NIR 和 VIS 人脸图像的正交字典对齐方法。在文献^[29]中，引入了一种重排序 (Re-ranking) 技术探索异构人脸匹配中的有价值信息。Pereira 等人^[103]引入模态特定单元在主干网络中分别学习模态特定的低级特征，并且在无需重新训练的前提下就可以共享模态不变高级特征。文献^[104]对模态间的深层非线性映射进行建模设计网络，在前馈深度神经网络及其有效的设计步骤下，映射的学习过程无需大量数据就能够提取身份特征。Fondje 等人^[105]提出模态与姿态不变框架，其核心在于引入一个子网络桥接特征中的模态和姿态差异。Wu 等人^[106]通过构建大量的跨模态图像对之间的关系以增强了模态间和类内样本的关联，并削弱了模态间样本和类间样本的关联。文献^[107]提出了子类异质性感知损失，该损失同时考虑了 NIR、VIS、高分辨率、低分辨率图像之间的关系，形成了有序的子空间。Hu 等人^[108]提出了模态不变损失以增加类间跨模态扩展性和降低类内跨模态分离性。

相比于模态不变描述子和合成的方法，子空间学习方法有着部署简单、网络轻便等优点。因此，基于子空间学习的方法适合部署到对实时性有要求的场景中。然而，子空间学习方法仍存在一些不足：1) 对标签的依赖性强，损失的设计需要身份标签作为监督信息。在现实场景中，获取跨模态的同一个体的身份标签是十分困难的，因为标注是一项费时费力的工作。使用少量标签进行训练容易出现过拟合、性能退化等问题；2) 特征的学习和解耦是一个不断降维的过程，如何最大限度地保留身份信息并减少身份无关信息的干扰，是值得考虑的问题。

1.3.4 面临的难题与挑战

在过去的二十年中，跨模态人脸识别取得了突破性的进展。随着深度学习方法的引入和设备算力的不断增强，跨模态人脸识别已经从学术界的理论研究成功地转变为工业界实际应用。面对社会发展和不断涌现的现实挑战，NIR-VIS 人脸识别作为跨模态人脸识别的代表性任务，其需求持续增加，因此在金融、安防、执法等领域得到广泛应用。然而，NIR-VIS 人脸识别技术仍然面临一些问题和挑战：



图 1-5 模态差异和剩余差异示意图，样本图片来自于 CASIA NIR-VIS 2.0 数据集^[12] 和 LAMP-HQ 数据集^[14]。第一行和第二行分别是 NIR 和 VIS 人脸

(1) 模态差异 (Modality differences) 问题：由于成像原理和相机参数设置的不同，不同模态人脸图像之间存在显著的外观差异，例如纹理细节和边缘轮廓等。图1-5展示了模态差异导致的成像上的不同。导致经卷积网络提取的特征在数据分布上存在明显差异。若直接采用同质人脸识别算法 (如常见的二维可见光人脸识别算法) 进行人脸匹配，识别性能将受到严重影响。现有 NIR-VIS 人脸识别方法在克服模态差异问题时，往往只考虑增加类间距离、减小类内距离这一简单的前提，而忽略了对模态信息的抽离与解耦。目前仅有少数方法专注于这一问题，但在解耦模态信息时可能会导致身份相关信息的损失。因此，如何有效减小 NIR-VIS 人脸图像的模态差异且增强身份相关的表示仍是亟待解决的难题。

(2) 剩余差异 (Residual differences) 问题：NIR-VIS 人脸识别中，人脸图像往往在开放世界中进行采集获得的。这些图像除了包含模态信息差异外，在图像质量上残次不齐，具体表现在姿态、遮挡、表情、照明和上述情况多种组合所附加的差异化信息。这些与身份无关、模态无关的表现差异称为剩余差异。剩余差异使得类内特征表示的紧凑度差，进一步增加了 NIR-VIS 人脸匹配的难度。图1-5展示了剩余差异导致的类内变化。目前主流方法只考虑到了身份信息 and 模态信息的影响，在设计网络和目标函数时并没有将剩余差异考虑在内。因此，如何进一步提纯身份相关特征表示，使其不受剩余差异的干扰，也是亟待解决的难题。

(3) 无监督学习 (Unsupervised learning) 问题：卷积神经网络需要大量样本进行学习，来获取出色的分类能力。在训练中，由标签提供的监督信号指明网络哪些跨模态图像属于同一身份，这对于构建有效损失提升模型性能至关重要。此外，数据集规模也是提升网络性能的关键。目前，可见光人脸识别技术的成功部分得益于大型人脸数据集，例如

MS-Celeb-1M^[109]、Webface^[110]等。在 NIR-VIS 人脸识别领域,现有的 NIR-VIS 人脸数据集中的训练样本数量非常有限,缺乏如此庞大的可用数据集。此外,从大量原始跨模态人脸数据中获取标注信息是耗时且昂贵的,因为需要对不同跨模态图像进行人工标注。标签问题直接限制了数据集规模,可能导致过拟合问题,并降低了有监督 NIR-VIS 人脸识别方法在更大规模真实场景中的可扩展性^[95]。无监督方法在图像分类、目标检测、行人重识别等任务中展现出令人鼓舞的表现,这是因为它们无需手动标注的标签即可训练得到接近有监督方法的性能水平。受此启发,无监督 NIR-VIS 人脸识别 (Unsupervised NIR-VIS Heterogeneous Face Recognition, NIR-VIS-UHFR) 成为解决标注不足问题的有效途径。如何设计有效方法来挖掘跨模态身份的关联关系,并将其用于为网络提供训练所需的监督信号,是目前 NIR-VIS 人脸识别尚未解决的难题之一。

1.4 NIR-VIS 人脸数据集

为了直观测评所提出的模型在不同场景下的性能,目前的研究方法在几个主流的数据集上采用相同的验证设置进行测试,方便在多种评价指标上公平比较并展示模型的性能。

1.4.1 CASIA NIR-VIS 2.0

CASIA NIR-VIS 2.0^[12] 是由中国科学院自动化所 (Chinese Academy of Sciences Institute of Automation, CASIA) 采集的大型 NIR-VIS 人脸数据集。该数据集包括在模拟的不可控环境中拍摄的 NIR-VIS 图像,在这种环境中,大的模态和剩余变化使来自同一个人的图像看起来不相似。采集时间分别于 2007 年春季、2009 年夏季、2009 年秋季和 2010 年夏季。该数据集一共包含 725 名受试者,且每一个受试者都有 1-22 张 VIS 和 5-50 张 NIR 人脸图像。需要注意的是,通常每个身份的 NIR 图像会多于 VIS 图像两到三倍。数据集包含十个评估协议^[12],训练实例和测试实例之间没有交集。图1-6展示了 CASIA NIR-VIS 2.0 数据集中经过人脸对齐并剪裁后的图像。



图 1-6 CASIA NIR-VIS 2.0 数据集^[12] 样本图片示例。第一行是 NIR 图像，第二行是 VIS 图像

1.4.2 Oulu-CASIA NIR-VIS

Oulu-CASIA NIR-VIS 数据集^[11] 由芬兰奥卢大学 (University of Oulu) 和 CASIA 共同采集的小型数据集。该数据集包含 80 个受试者，其中有 50 名来自奥卢大学，30 名来自 CASIA。测试协议中，40 名受试者被平均分为训练集和测试集。训练集和测试集有同等数量的 1920 张人脸图片，并且每个受试者的 NIR-VIS 图像的数量是均衡的，分别有 48 个 NIR 图像和 48 个 VIS 图像。特别地，NIR-VIS 图像中有各种类型的剩余信息，如愤怒、厌恶、幸福、悲伤等，也有不同的照明条件，如弱光照、普通室内光照以及暗黑场景。测试集和训练集中采用的身份是不同的，因此没有交集。图1-7展示了 Oulu-CASIA NIR-VIS 数据集中经过人脸对齐并剪裁后的异质图像。



图 1-7 Oulu-CASIA NIR-VIS 数据集^[11] 样本图片示例。第一行是 NIR 图像，第二行是 VIS 图像

1.4.3 BUAA NIR-VIS

BUAA NIR-VIS 数据集^[12] 由北京航空航天大学构建，在室内场景中收集了来自于 150 个受试者的 NIR-VIS 图片。每个受试者有 9 张 NIR 人脸图片和 9 张 VIS 人脸图片，

这些人脸图片具有不同的表情和姿态。在评估协议中，训练集包含 50 个身份，共有 900 张图片，测试集包含 100 个身份共有 1800 张图片。此外，训练集和测试集中不包含相同的身份，因此不存在交集。图1-8展示了 BUAA NIR-VIS 数据集中经过人脸对齐并剪裁后的 NIR-VIS 图像。



图 1-8 BUAA NIR-VIS 数据集^[112] 样本图片示例。第一行是 NIR 图像，第二行是 VIS 图像

1.4.4 LAMP-HQ NIR-VIS

LAMP-HQ NIR-VIS 数据集^[14] 英文全称为 Large-Scale Multi-Pose High-Quality NIR-VIS database，是由 CASIA 在 2020 年采集而成，具有高质量和多姿态的特点。该数据集是目前最大的开源 NIR-VIS 人脸数据库，包含来自 573 名受试者的 16828 张 VIS 图像和 56788 张 NIR 图像。每个受试者的 NIR 和 VIS 图像的数量是不平衡的，NIR 图像比 VIS 图像多 2-3 倍。异质图像有着极具挑战的表情、眼镜遮挡、人脸姿态变化等特点。LAMP-HQ NIR-VIS 数据集提供了 10 个官方协议以供平等地验证算法的性能。在每个协议中，训练集包括 300 名受试者，而测试集包括 273 名受试者。图1-9展示了 LAMP-HQ NIR-VIS 数据集中经过人脸对齐并剪裁后的 NIR-VIS 图像。

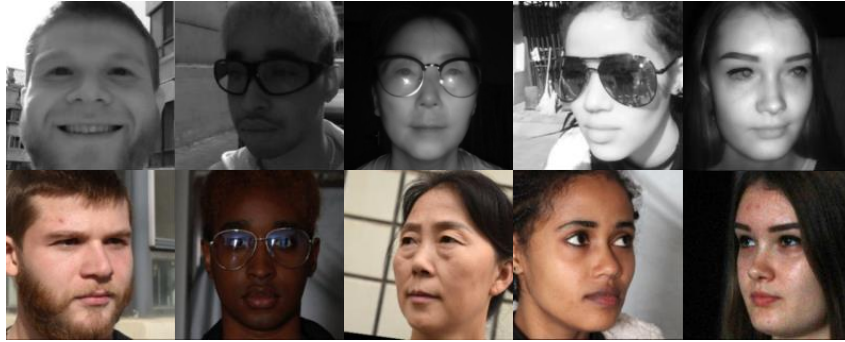


图 1-9 LAMP-HQ NIR-VIS 数据集^[14] 样本图片示例。第一行是 NIR 图像，第二行是 VIS 图像

1.5 NIR-VIS 人脸识别性能评估指标

目前用于评估网络 NIR-VIS 人脸识别性能的常见指标包括: Rank-1、VR@FAR=1%、VR@FAR=0.1%、VR@FAR=0.01% 等。具体而言, Rank-1 指的是在度量探针集与参考集距离时, 正样本在相似度排名中排在第一的概率。另外, VR@FAR=1%、VR@FAR=0.1%、VR@FAR=0.01% 指标与受试者工作特征 (Receiver Operating Characteristic, ROC) 曲线相关, 可直观地展现系统在不同阈值下的性能。在受试者工作特征曲线的坐标系中, 横坐标表示错误接受率 (False Accept Rate, FAR), 纵坐标表示验证率 (Verification Rate, VR)。因此, VR@FAR=1% 表示在 FAR 为 1% 时对应的 VR 值。VR@FAR=0.1% 和 VR@FAR=0.01% 的定义与 VR@FAR=1% 类似, 只是横坐标的取值稍有不同。

表 1-1 本文涉及的工作对应的成果

研究方法	发表的期刊
融合空间学习方法	ACM Trans. on Multimedia Computing, Communications, and Applications (一作)
中立人脸学习和渐进融合合成方法	IEEE Trans. on Circuits and Systems for Video Technology (一作)
鲁棒跨模态伪标签和对比学习方法	IEEE Trans. on Image Processing (一作)
同质到异质学习和剩余不变增强方法	IEEE Trans. on Information Forensics and Security (一作)

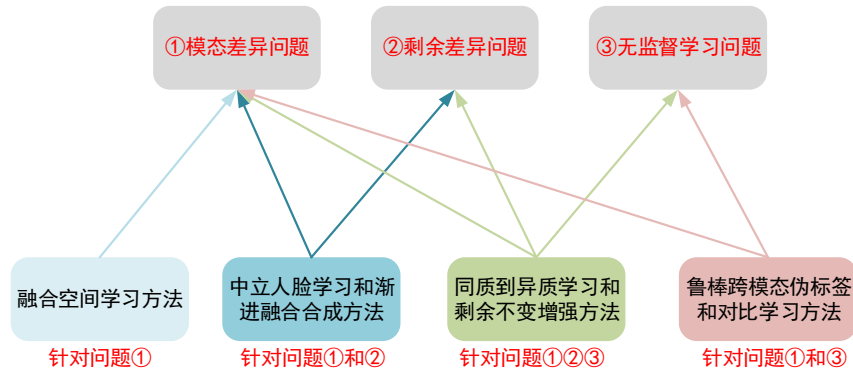


图 1-10 本文各个工作与研究的关键问题之间的关系

1.6 论文研究内容

本论文针对当前 NIR-VIS 人脸识别领域存在的模态差异、剩余差异及无监督学习等问题展开深入研究，旨在进一步提升 NIR-VIS 人脸识别的性能，同时增强其应用的便利性和可拓展性。在文章的结构安排上，本文首先介绍了跨模态人脸识别研究的背景与意义，接着阐述了 NIR-VIS 人脸识别框架、国内外研究现状、当前所面临的难题和挑战，数据集以及 NIR-VIS 人脸识别验证的方法。随后，针对目前存在的挑战性问题，本文提出了多种研究方法，并进行了深入的分析。如表1-1所示，本文所涉及的研究工作已经在多个国际知名期刊上发表。这些文章均针对所研究的关键性问题，具体对应关系如图1-10中所示。本文涉及的四项工作的创新点总结如下：

- 1) **融合空间学习方法**: 针对模态差异问题，本文设计了融合空间学习网络 (Syncretic Space Learning network, SSL)。网络集成融合模态生成器 (Syncretic Modality Generator, SMG)、模态对抗融合空间学习 (Modality-adversarial Syncretic space Learning, MSL)、融合分布一致性 (Syncretic Distribution Consistency, SDC) 到端到端的框架中，旨在通过生成融合模态来缩小模态之间的差异。本方法的创新点总结如下: (1) 为了在图像级别上缩小模态差距，引入了 SMG，用于将 NIR 和 VIS 图像合成到保留面部结构且模态变化较小的融合图像。(2) MSL 采用了以融合模态为目标的融合模态对抗学习策略，将 NIR 和 VIS 图像投影到融合空间中。(3) SDC 包含三种分布一致性，即 NIR-融合一致性、融合-融合一致性、VIS-融合一致性，增强特征的类内紧凑性。(4) 在 CASIA NIR-VIS 2.0^[12]、Oulu-CASIA NIR-VIS^[111]、LAMP-HQ NIR-VIS^[14] 三个具有挑战性的数据库上进行了大量实验，结果显示了 SSL 方法的

有效性。

- 2) **中立人脸学习和渐进融合合成方法:** 针对模态差异与剩余差异问题, 设计了中立人脸学习和渐进融合合成网络 (Neutral face Learning and Progressive Fusion synthesis network, NLPF)。模型将身份相关中立人脸学习 (Identity-related Neutral face Learning, INL) 和属性渐进融合 (Attribute Progressive Fusion, APF) 集成到一个联合框架中, 通过对抗生成的方法分离身份、模态、剩余相关特征, 并以中立人脸作为监督信息提取鉴别性的中立身份特征。本方法的创新点包括: (1) 所提出的 INL 模块指导网络学习同质的中立人脸表示, 消除模态差异和剩余差异。在重构同质中立人脸的过程中, 身份相关特征被映射到一个共同的中立表示子空间。(2) APF 对身份、模态、剩余特征进行解耦与重组。通过渐进融合过程促进跨模态人脸属性的分解, 有助于促进身份相关特征的净化。(3) 在具有挑战性的 CASIA NIR-VIS 2.0^[12]、BUAA NIR-VIS^[11]、LAMP-HQ NIR-VIS^[14] 数据集上验证了 NLPF 方法的有效性和鲁棒性。
- 3) **鲁棒跨模态伪标签和对比学习方法:** 针对模态差异和无监督学习问题, 本文提出了鲁棒跨模态伪标签和对比学习方法 (Robust cross-domain Pseudo-labeling and Contrastive learning, RPC)。这一方法包含三个核心模块: 基于 NIR 集群的伪标签共享 (NIR cluster-based Pseudo labels Sharing, NPS) 模块、模态域特定集群对比学习 (Domain-specific cluster Contrastive Learning, DCL) 模块、跨模态域集群对比学习 (Inter-domain cluster Contrastive Learning, ICL) 模块, 通过无监督学习降低对标签的依赖。该方法的创新点如下: (1) 所提出的方法能够生成高质量的伪标签, 并且在性能上与有监督方法相媲美。RPC 取得了超过 99% 的标签估计准确率, 并在多个 NIR-VIS 数据集上展现出领先的性能。(2) NPS 探索鲁棒的 NIR 集群, 并以跨模态域检索的方式与 VIS 模态共享可靠的标签信息, 从而建立跨模态域身份特征的内在关系。(3) DCL 执行模态域内对比损失计算和动态记忆更新, 学习模态域内紧凑且具有区分性的表示。(4) ICL 通过基于混合模态域记忆的对比学习, 引导实例级特征学习到鲁棒且与模态域无关的人脸表征。
- 4) **同质到异质学习和剩余不变增强方法:** 针对模态差异、剩余差异、无监督学习问题, 本文提出了同质到异质学习和剩余不变增强网络 (homogeneous-to-Heterogeneous

learning and Residual-invariant Enhancement, HERE)。具体来说,该方法将模态-对抗对比学习 (Modality-Adversarial Contrastive Learning, MACL)、基于 K-相互编码的跨模态标记 (K-reciprocal-Encoding-based Cross-modal Labeling, KECL)、精细跨模态对比学习 (Refined cross-modal Contrastive Learning, RCL)、剩余不变表示增强 (Residual-invariant Representations Enhancement, RRE) 纳入一个端到端的框架中。本方法的创新点可以归纳如下: (1) MACL 通过模态对比学习和对抗学习的协同作用,缩小两种模态分布、学习模态内部可区分表示,有效地降低了模态差异。(2) KECL 通过探索跨模态 K-相互编码来解决为 NIR-VIS 数据分配全局共同伪标签的挑战。(3) RCL 利用模态不变的平均初始化和动态关注加权策略,探索模态不变的表示。(4) RRE 挖掘并优化局部线索,并将其与全局特征结合,增强模型对剩余变化的鲁棒性。(5) 在具有挑战性的 CASIA NIR-VIS 2.0^[12]、BUAA NIR-VIS^[111]、LAMP-HQ NIR-VIS^[14] 数据集上验证了 HERE 方法的有竞争力性能。

1.7 论文组织架构

图1-11展示了本文整体结构,共分为六章,具体每章的主题和内容介绍如下:

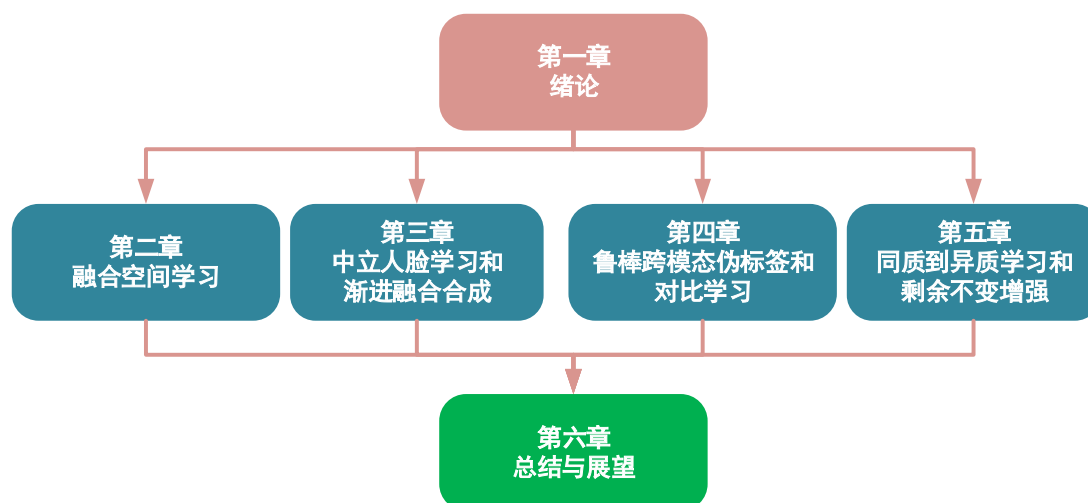


图 1-11 本文整体结构

第一章: 绪论部分介绍了人脸识别及跨模态人脸识别的背景和重要性、NIR-VIS 人脸识别框架、国内外研究现状、现有方法所面临的挑战与难题、NIR-VIS 人脸识别数据集和验证方法,并阐述了本文的研究内容,包括各个方法所涉及的问题和贡献点,最后

对本文的结构和章节安排进行了说明。

第二章：针对模态差异问题，本章提出了融合空间学习方法。首先概述了融合空间学习的总体流程，然后详细介绍了各个模块，包括融合模态生成器、模态对抗融合空间学习、融合分布一致性，最后对各模块特性及整体框架在多个数据集上的性能进行了分析。

第三章：为了解决模态差异与剩余差异问题，本章设计了中立人脸学习和渐进融合合成方法。首先描述了中立人脸学习和渐进融合合成网络的总体框架，接着详细阐述了身份相关的中性人脸学习、属性渐进融合模块的机制和实现细节，最后在定性和定量实验中验证了本方法在多个 NIR-VIS 数据集上的有效性。

第四章：为了解决模态差异和无监督学习问题，本章提出了鲁棒跨模态伪标签和对比学习方法。首先，介绍了该方法的总体流程，然后详细说明了基于 NIR 集群的伪标签共享、模态域特定集群对比学习、模态域集群对比学习模块的动机和技术流程，并在多个数据集上验证了伪标签质量和整体模型性能。

第五章：为了应对模态差异、剩余差异、无监督学习问题，本章提出了同质到异质学习和剩余不变增强方法。首先对总体框架进行了介绍，随后详细阐述了模态-对抗对比学习、基于 K-相互编码的跨模态标记、精细跨模态对比学习、剩余不变表示增强模块的动机和算法细节。最后，在多个具有挑战性的 NIR-VIS 人脸数据集上验证了本方法的有效性。

第六章：总结与展望，对本文提出的工作和贡献进行总结，并展望未来的研究方向。

第 2 章 融合空间学习方法

模态差异在跨模态图像上可以直观地表现为纹理和颜色上的不同,这使得在特征空间中 NIR 和 VIS 特征之间出现很大的模态差异。由于模态差异的存在,模型无法准确地进行跨模态人脸匹配。当前的基于合成的方法在像素层面上进行模态转换消除模态差异,但是面临参数量过大和合成质量差等挑战。为了解决这一问题,本章提出了一种新的融合空间学习 (Syncretic Space Learning network, SSL) 模型,包含三个关键模块:融合模态生成器 (Syncretic Modality Generator, SMG)、模态对抗融合空间学习 (Modality-adversarial Syncretic space Learning, MSL)、融合分布一致性 (Syncretic Distribution Consistency, SDC)。首先,融合模态生成器利用浅层卷积神经网络将 NIR 和 VIS 图像合成为融合图像。该过程采用通道级卷积处理图片,保留了具有辨识性的结构信息,并以自监督的方式进一步提升了图像质量,使得异质图像能够自适应学习转化到模态差异较小的融合模态空间。其次,模态对抗融合空间学习模块通过对抗学习的策略将 NIR 和 VIS 图像映射到融合空间中。这种利用融合模态作为目标的学习方法,有效地减少了 NIR-VIS 人脸图像中的模态差异。最后,由 NIR-融合一致性 (NIR-Syncretic Consistency)、融合-融合一致性 (Syncretic-Syncretic Consistency) 以及 VIS-融合一致性 (VIS-Syncretic Consistency) 构建的融合分布一致性有助于增强类内紧凑性和学习到鉴别的身份特征。在 CASIA NIR-VIS 2.0^[12]、Oulu-CASIA NIR-VIS^[111]、LAMP-HQ NIR-VIS^[14] 数据库上的大量实验表明了 SSL 方法的有效性。这一研究成果已在 ACM Transactions on Multimedia Computing, Communications, and Applications 正式发表。

2.1 引言

在低光照环境下,特别是夜晚,可见光传感器往往难以捕捉清晰的人脸图像,这会导致传统可见光人脸识别系统的性能下降。近年来,越来越多的安全场景采用能够在低光照环境下自动切换到 NIR 模式的监控摄像头,用于获取清晰的 NIR 人脸图像。由于存在模态差异,NIR 与 VIS 人脸之间的匹配仍然是一个充满挑战的问题。

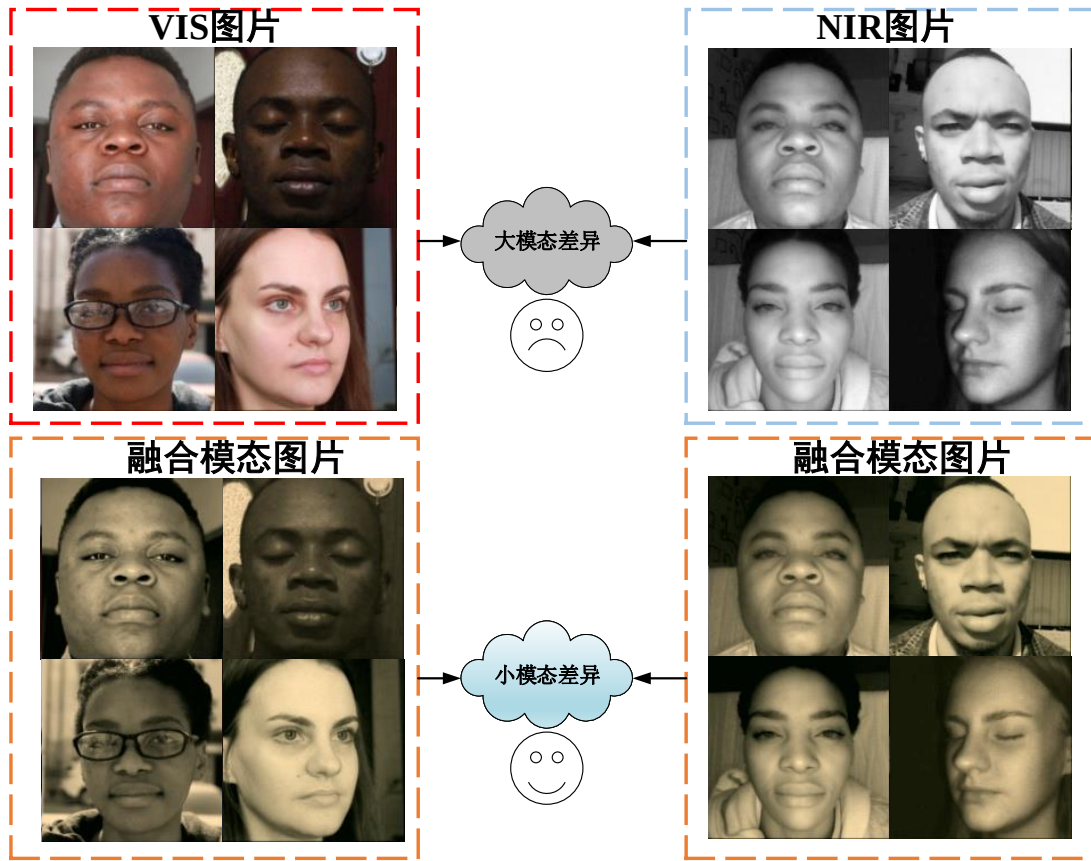


图 2-1 模态差异展示。NIR 和 VIS 图像之间存在较大的模态差异。本章旨在将 NIR 和 VIS 图像合成为融合图像，并将 NIR、VIS、融合图像投影到融合空间中进行有效分类。生成的融合图像具有较小的模态差异，并保留了大部分的结构信息。

为了解决 NIR-VIS 异构人脸识别问题，学术界已经提出了许多有效的方法来消除模态差异。现有方法可分为三种策略，包括跨模态合成、模态不变描述子以及子空间学习方法。跨模态合成方法旨在通过生成对抗网络 (GANs) 将 NIR 图像转换为保留身份信息的 VIS 图像^[27,71-73]。模态不变描述子的主要思想是基于先验知识设计手工描述子，用于提取对光谱信息不敏感且具有区分性的异构人脸特征^[47,48,113,114]。子空间学习方法目的在于学习一种映射，将 NIR 和 VIS 人脸图像投影到一个公共子空间中，并调整特征分布进行有效分类^[94,96,99,108]。

最近，一些跨模态合成方法^[70,115,116] 尝试将 NIR 图像转换为 VIS 图像，并使用未经过微调的 VIS 人脸识别网络对生成的 VIS 人脸与查询集中的 VIS 人脸进行匹配。然而，这些跨模态合成方法面临一些挑战^[72,74]：1) **身份信息丢失**。训练 GANs 在特征图的缩小和放大过程中，合成图像的某些身份信息会丢失。2) **推理效率**。GANs 引入了具有大参数的额外模块，如生成器 G 和鉴别器 D ，不适用于实时人脸识别系统。3) **感知差异**。

不同摄像头之间的感知差异导致 NIR 图像和对应的 VIS 图像之间在纹理和结构上存在显著差异。因此，生成的图像存在分辨率低或者伪影等问题。

为解决这些问题，本章提出融合空间学习 (Syncretic Space Learning network, SSL) 方法合成 NIR 和 VIS 图像为融合图像，并将 NIR、VIS、融合图像投影到融合空间中以实现有效分类。其中包括融合模态生成器 (Syncretic Modality Generator, SMG)、模态对抗融合空间学习 (Modality-adversarial Syncretic space Learning, MSL)、融合分布一致性 (Syncretic Distribution Consistency, SDC) 三个模块。与现有的合成方法不同，SSL 方法假设存在一个既不属于 NIR 模态也不属于 VIS 模态的融合图像空间。通过将跨模态图像转换为这个融合模态，能够有效地消除了模态差异并提取具有区分性的表示。首先，SMG 将 NIR 和 VIS 图像合成为融合模态。在网络优化过程中，以自监督学习方式进一步提升了图像质量，生成减小模态差异的融合图像。如图2-1所示，具有较大模态差异的 NIR 和 VIS 图像被转换为具有小模态差异的融合图像。其次，引入 MSL 通过对抗学习策略将 NIR 和 VIS 图像映射到融合空间中。特别地，与对抗学习方法^[117-119]不同，SSL 方法通过一种新颖的融合模态对抗学习策略，在融合空间中学习到了模态不变特征。最后，在 SDC 中，三类内约束一致性 (即 NIR-融合一致性、融合-融合一致性、VIS-融合一致性) 有效提升了类内特征的紧凑性。总之，本章的主要贡献如下：

- 提出的 SSL 网络将 SMG、MSL、SDC 融合到一个统一框架中，通过生成融合模态提高模型性能。
- 为了在图像层级上缩小模态差异，引入了 SMG 将 NIR 和 VIS 图像合成为保持面部结构的融合图像。
- MSL 通过融合模态对抗学习策略将 NIR 和 VIS 图像投影到融合空间，提高了跨模态检索性能。
- SDC 中由 NIR-融合一致性、融合-融合一致性、VIS-融合一致性构成的类内一致性约束了融合空间中特征的分布，以增强特征的紧凑性。
- 在 CASIA NIR-VIS 2.0^[12]、Oulu-CASIA NIR-VIS^[111]、LAMP-HQ NIR-VIS^[14] 数据库上的大量实验表明了 SSL 方法的有效性。

2.2 融合空间学习模型

表 2-1 SSL 中符号的含义。

符号	含义
X^N	NIR 模态下的图像
X^V	VIS 模态下的图像
X^{SN}	由 NIR 数据生成的融合图像
X^{SV}	由 VIS 数据生成的融合图像
l_i^N / l_{mi}^N	NIR 图像 X_i^N 的身份/模态标签
l_i^V / l_{mi}^V	VIS 图像 X_i^V 的身份/模态标签
l_i^S / l_{mi}^S	融合图像 X_i^{SN}, X_i^{SV} 的身份/模态标签
W^N / W^V	模态特定的全连接层结构
W^I	模态共享的全连接层结构
W_{id}^N / W_{id}^V	模态特定身份分类器的参数
W_{id}	模态共享的身份分类器的参数
W_m	模态特定模态分类器的参数

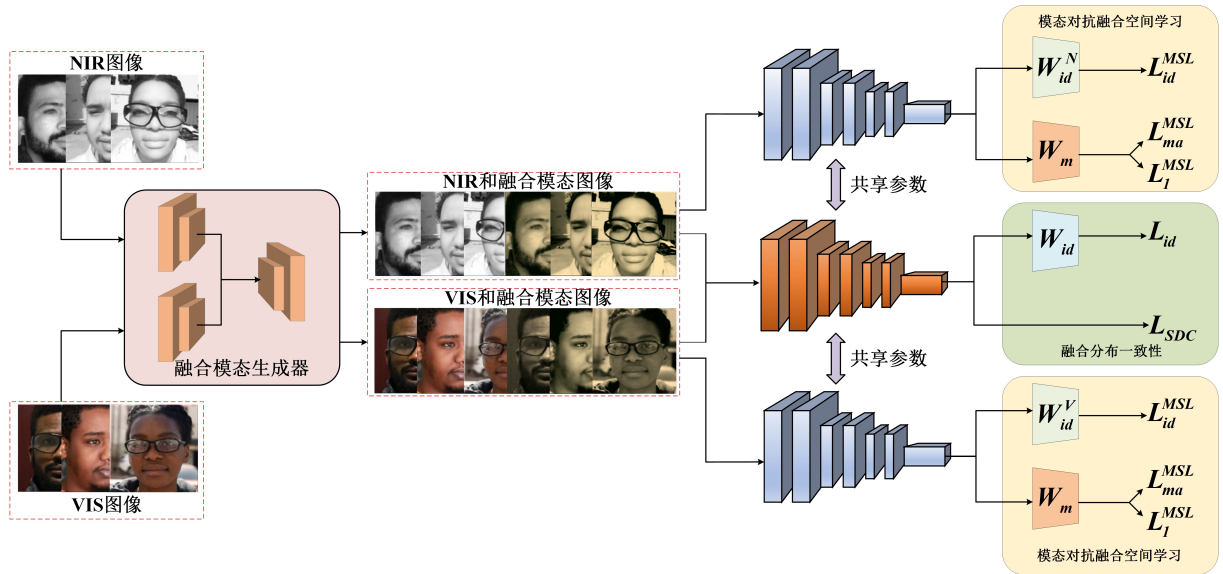


图 2-2 SSL 模型的流程图。SSL 模型由三个部分组成，融合模态生成器 SMG、模态对抗融合空间学习 MSL、融合分布一致性 SDC

在本节中，将详细介绍 SSL 网络。如图2-2所示，SSL 方法包括三个有效模块，分别是融合模态生成器 SMG、模态对抗融合空间学习 MSL、融合分布一致性 SDC。首先，融合模态生成器采用轻量化的卷积结构，生成了模态差异小的融合模态图片。其次，模态对抗融合空间学习结合了对抗学习和分类损失，对齐了 NIR、VIS、融合模态图片的特征分布。最后，融合分布一致性通过引入三种模态间的距离约束保持了类内的一致

性。

2.2.1 问题建模

NIR-VIS 异构人脸数据库包括 NIR 人脸数据集 $X^N = \{(x_i^N, l_i^N, l_{mi}^N)\}_{i=1}^{n^N}$ 和 VIS 人脸数据集 $X^V = \{(x_i^V, l_i^V, l_{mi}^V)\}_{i=1}^{n^V}$ ，其中 x_i^N 和 x_i^V 表示图像数据， $l_i^N, l_i^V \in \{1, 2, \dots, c\}$ 代表身份标签， $l_{mi}^N \in NIR$ 以及 $l_{mi}^V \in VIS$ 表示模态标签， n^N 以及 n^V 分别表示 NIR 和 VIS 人脸图像数据的数量。

一些基于深度学习的人脸识别方法，如 WCNN^[93]、DFD^[94]、DADRD^[98]、FAN^[120]，通过在多分支网络中引入解耦技术来学习判别性表示。与他们不同的是，本方法在每个分支中引入融合模态进行学习。此外，当前的跨模态合成方法通过视觉转换方法将 NIR 模态转换为 VIS 模态图像，然后在相同的模态中进行人脸匹配。然而，使用基于参数量庞大的 GANs^[70,115,116] 进行视觉转换导致训练难度高和推理效率低的问题。与这些方法不同的是，本方法采用轻量化的卷积同时合成 NIR 和 VIS 图像为模态无关的融合图像，并考虑了一种新颖的融合空间学习方法消除模态差异。具体而言，SSL 的网络设计中引入了一个三支网络结构，包括两个模态特定分支和一个模态共享分支，有效处理 NIR、VIS、融合图像并学习模态不变表示。这三个分支的映射可以定义如下：

$$\begin{aligned} y^I &= f(X; \Theta, W^I) \\ y^N &= f(X; \Theta, W^N) \\ y^V &= f(X; \Theta, W^V) \end{aligned} \quad (2-1)$$

其中 $y^s (s \in I, N, V)$ 表示提取的特征， X 表示输入数据， Θ 表示主干网络参数， W^N 以及 W^V 表示模态特定的全连接结构， W^I 表示模态共享的全连接结构。

2.2.2 融合模态生成器

消除模态差异并学习一个模态无关的共享子空间对于跨模态人脸识别任务至关重要。现有的跨模态合成方法^[70,115,116]在消除图像级别的模态差异方面存在技术瓶颈。一些在跨模态行人重识别任务中的优秀工作可作为解决上述问题的参考，例如灰度转换^[121]、

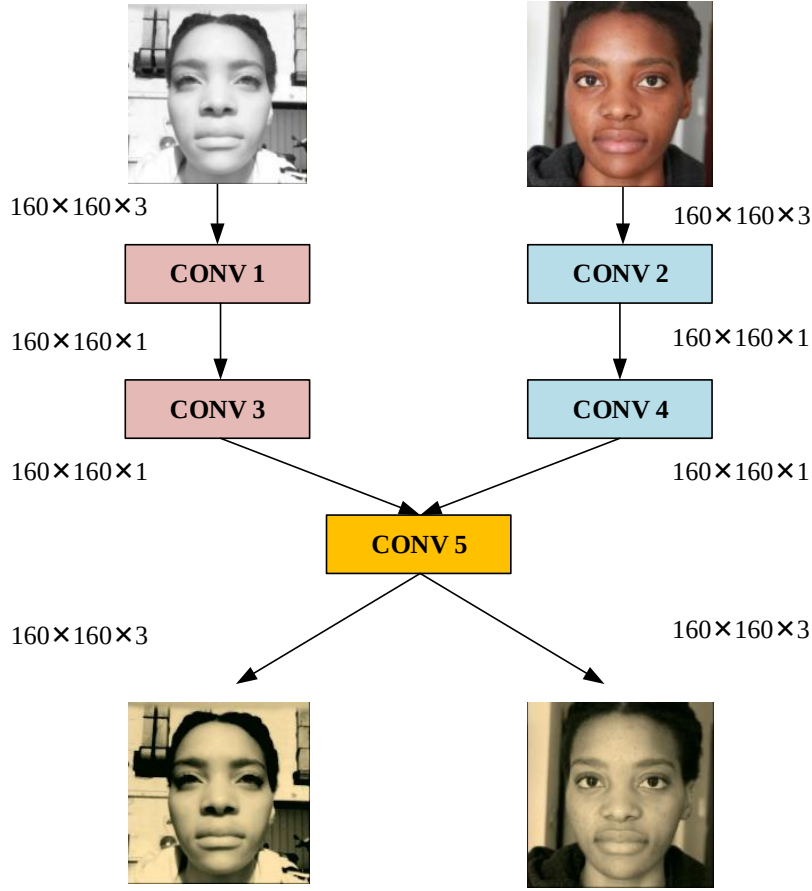


图 2-3 融合模态生成器 SMG 结构

X 模态^[122]、中间模态学习^[123]。受这些工作启发,本节设计了融合模态生成器 SMG 来消除模态差异。SMG 的结构显示在图2-3中。从图中可以看出 SMG 的优势在于其只包含少量卷积层的简单轻巧结构。具体而言,编码器(包括 CONV1、CONV2、CONV3 和 CONV4)执行通道级卷积操作,以尽量不破坏图像中的结构信息。NIR 图像 $X^N \in \mathbb{R}^{H \times W \times 3}$ 和 VIS 图像 $X^V \in \mathbb{R}^{H \times W \times 3}$ 被映射到特征图 $M \in \mathbb{R}^{H \times W \times 1}$ 。然后,解码器(即 CONV5)将特征图 $M \in \mathbb{R}^{H \times W \times 1}$ 投影到具有三通道且能自适应学习颜色模式的融合图像 $X^{SN} \in \mathbb{R}^{H \times W \times 3}$ 和 $X^{SV} \in \mathbb{R}^{H \times W \times 3}$ 中。通过上述过程,可以有效保留了与身份信息相关的图像结构信息,并通过可训练的网络结构将异构信息转换为共享的融合模态。在网络的优化过程中,SMG 以自监督方式学习新的共享颜色模式,并自适应地提高融合图像的质量,以减小模态差异。生成器处理图片的前向过程可以写成:

$$\begin{aligned} X^{SV} &= f(X^V; \Theta^{SMG}) \\ X^{SN} &= f(X^N; \Theta^{SMG}) \end{aligned} \quad (2-2)$$

其中 X^{SV} 表示由 VIS 图像生成的融合图像, X^{SN} 表示由 NIR 图像生成的融合图像, Θ^{SMG} 表示 SMG 网络的参数。融合图像的身份标签和模态标签分别表示为 l_i^S 和 l_{mi}^S 。图2-4展示了一些生成的融合图像。可以清晰地观察到, 尽管最初由 SMG 生成的融合图像 (详见图2-4 (a) 和 (b) 的第二行) 类似于灰度图像。若没有引入可以自适应训练的过程, 面部结构和纹理信息并不会得到很好的保留。显然, 使用这样的图像用于训练学习身份信息是不合适的。对此, 引入了一个可训练的 SMG, 允许异构图像自适应地学习到一个新的共享融合模态 (见图2-4 (a) 和 (b) 的第三行)。这在很大程度上保留了大部分结构信息, 有助于身份信息的学习。所提出 SMG 的详细网络结构显示在表2-2中。与侧重于模态无关高级特征空间学习的子空间方法以及侧重于模态间图像级映射的跨模态合成方法不同。通过引入 SMG, SSL 假设在 NIR 和 VIS 图像之间存在模态无关的融合图像空间, 通过在融合空间内引导学习过程, 可以有效解决模态差异问题。

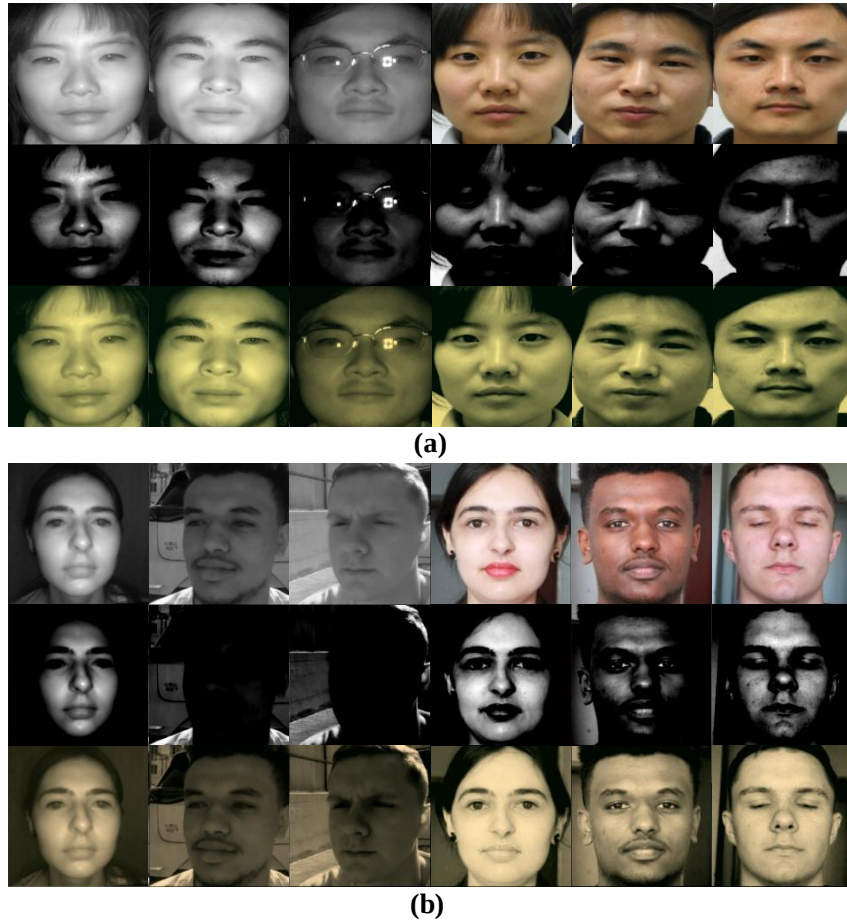


图 2-4 SMG 生成的融合图像示例。(a) 显示从 CASIA NIR-VIS 2.0^[12] 中选择的样本。(b) 显示从 LAMP-HQ NIR-VIS^[14] 中选择的样本。原始的 NIR-VIS 图像显示在第一行。从 SMG 中最初生成的融合图像显示在第二行。最后一行展示了经过网络优化过程后从 SMG 中获得的融合模态图像

表 2-2 融合模态生成器 SMG 中的卷积层设置

名称	卷积核大小/步长/填充	输出尺度
图像输入	-	160x160x3
Conv 1	1x1 / 1 / 0	160x160x1
Conv 2	1x1 / 1 / 0	160x160x1
Conv 3	1x1 / 1 / 0	160x160x1
Conv 4	1x1 / 1 / 0	160x160x1
Conv 5	1x1 / 1 / 0	160x160x3

2.2.3 模态对抗融合空间学习

最近，对抗学习策略展示了其引导网络提取模态不变特征表示的优势^[117-119]。在模态对抗融合空间学习 MSL 中，通过一种融合模态对抗学习策略，并以融合模态作为引导学习的目标，将 NIR 和 VIS 图像投影到融合空间中。

图2-5总结了近期发表的方法^[117-119]中的对抗学习思想以及本章提出的对抗学习方法。对抗学习 (Adversarial learning) 方法类似于 GANs 的原理。例如，GANs 采用鉴别器与生成器进行极小-极大博弈，其中鉴别器区分重建图像的真实性，生成器生成更真实的图像以迷惑鉴别器。在对抗学习策略中，可以采用模态分类器来区分特征属于哪个模态。然后，为主干网络设置不同或相反的学习目标。这样，主干网络提取的特征将混淆分类器，使得分类器无法分类特征属于哪个模态。在这种情况下，可以认为提取的特征是模态不变的。如图2-5 (a) 所示，Dai 等人^[117]采用模态分类器将特征分为 NIR 和 VIS 模态，并以相反的分类目标优化主干网络。在图2-5 (b) 中，Hao 等人^[118]训练主干网络的目的是使模态分类器能够以相同概率分类 NIR 或 VIS 模态。如图2-5 (c) 所示，Lin 等人^[119]在训练模态分类器时引入了模态不变类别。与上述方法不同的是，SSL 方法假设融合图像来自于同一个模态空间，它们的特征是模态不变的，将 NIR 和 VIS 图像映射到融合空间可以获得模态不变特征。为了将 NIR 和 VIS 图像投影到融合空间，首先采用常用的交叉熵损失来训练模态特定的分支，形式如下：

$$\begin{aligned}
 L_{id}^{MSL} &= - \sum_i^{2 \times n^V} l_i^k \log(\hat{l}_i^k) - \sum_i^{2 \times n^N} l_i^\delta \log(\hat{l}_i^\delta) \\
 \hat{l}_i^k &= f(y_i^V; W_{id}^V) \\
 \hat{l}_i^\delta &= f(y_i^N; W_{id}^N) \quad (k \in \{V, S\}, \delta \in \{N, S\})
 \end{aligned} \tag{2-3}$$

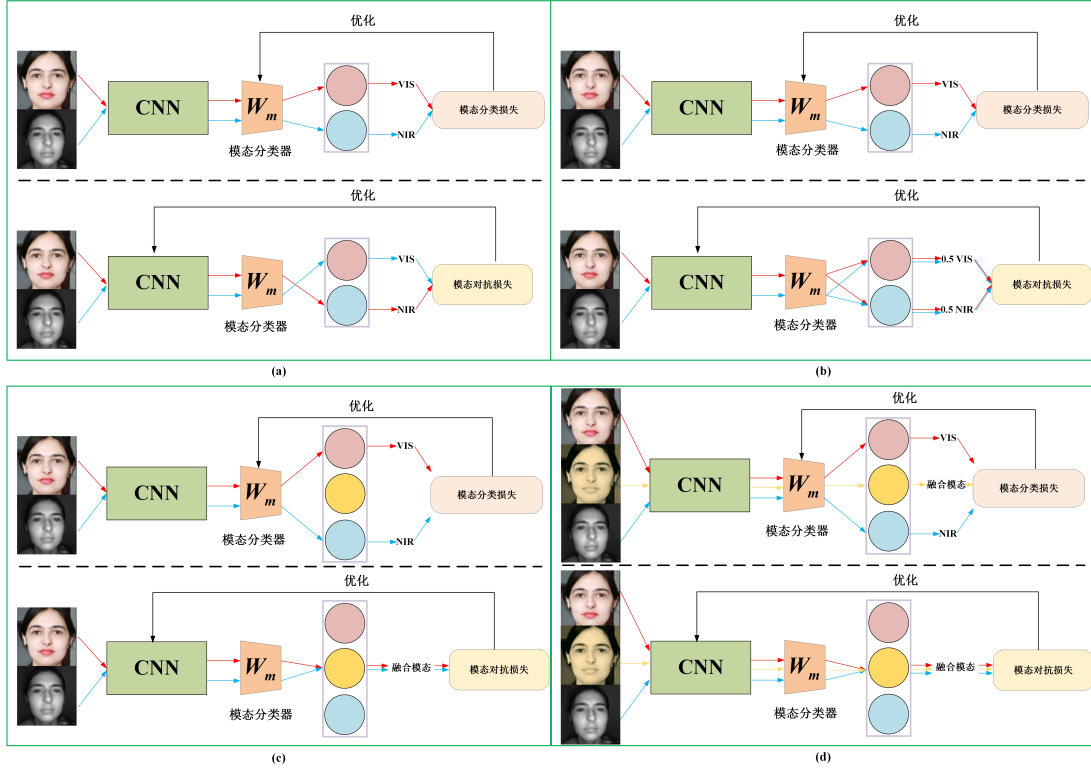


图 2-5 对抗学习中不同策略的示例。(a) cmGAN^[117]; (b) MCLNet^[118]; (c) MITML^[119]; (d) SSL 方法

其中, $y_i^V = f(X^s; \Theta, W^V)(s \in V, SV)$ 表示在使用 VIS 图像 X^V 和融合图像 X^{SV} 作为输入时, 从 VIS 模态特定分支提取的特征, $y_i^N = f(X^s; \Theta, W^N)(s \in N, SN)$ 表示在使用 NIR 图像 X^N 和融合图像 X^{SN} 作为输入时, 从 NIR 模态特定分支提取的特征。 W_{id}^V 和 W_{id}^N 表示模态特定身份分类器的参数, $\hat{l}_i^k$ 和 $\hat{l}_i^\delta$ 表示身份分类器预测的身份标签向量。

引入模态分类器来确定特征属于哪个模态。具体而言, 采用交叉熵损失作为监督目标, 引导模态分类器准确地将输入特征分类为 NIR、VIS 和融合模态类别。因此, 优化模态分类器的函数可以表示为:

$$L_1^{MSL}(W_m) = - \sum_i^{2 \times n^V} l_{mi}^k \log(\hat{l}_{mi}^k) - \sum_i^{2 \times n^N} l_{mi}^\delta \log(\hat{l}_{mi}^\delta) \quad (2-4)$$

$$\hat{l}_{mi}^k = f(y_i^V; W_m)$$

$$\hat{l}_{mi}^\delta = f(y_i^N; W_m) \quad (k \in \{V, S\}, \delta \in \{N, S\})$$

其中 W_m 表示模态特定模态分类器的参数, $\hat{l}_{mi}^k$ 和 $\hat{l}_{mi}^\delta$ 表示预测的模态标签向量。

为了将提取到的 NIR、VIS、融合模态特征统一分类为融合模态, 设计模态对抗目

标函数，具体可以表达为：

$$L_{ma}^{MSL} = - \sum_i^{2 \times n^V} l_{mi}^S \log(\hat{l}_{mi}^k) - \sum_i^{2 \times n^N} l_{mi}^S \log(\hat{l}_{mi}^\delta) \quad (2-5)$$

根据公式 (2-3) 和公式 (2-5)，优化主干网络和身份分类器的函数可以表示为：

$$L_2^{MSL}(\Theta^{SMG}, \Theta, W^N, W^V, W_{id}^N, W_{id}^V) = \lambda_1 L_{id}^{MSL} + \lambda_2 L_{ma}^{MSL} \quad (2-6)$$

其中 λ_1 和 λ_2 是 L_{id}^{MSL} 和 L_{ma}^{MSL} 的权重。在优化中，公式 (2-4) 和公式 (2-6) 交替优化。如图2-5 (d) 所示，在对抗过程中，模态分类器试图将 NIR、VIS、融合图像分类到各自的模态类别中，而主干网络试图将这三种图像统一分类为融合模态。MSL 中的联合损失可以表示为：

$$L_{MSL} = L_1^{MSL} + L_2^{MSL} \quad (2-7)$$

融合模态对抗学习策略以融合图像作为引导，确保 NIR 和 VIS 图像被投影到一个共同的融合空间。

2.2.4 融合分布一致性

为了进一步增强类内特征的紧凑性，在模态共享分支中引入了一种融合分布一致性 SDC 策略。首先，采用交叉熵损失来训练模态共享分支，具体表达如下：

$$L_{id} = - \sum_i^{2 \times (n^V + n^N)} l_i^k \log(\hat{l}_i^k) \quad (2-8)$$

$$\hat{l}_i^k = f(y_i^I; W_{id}) \quad (k \in \{V, N, S\})$$

其中，模态共享分支以 NIR 图像 X^N 、VIS 图像 X^V 、生成的融合图像 $\{X^{SV}, X^{SN}\}$ 作为输入， W_{id} 表示模态共享身份分类器的参数。

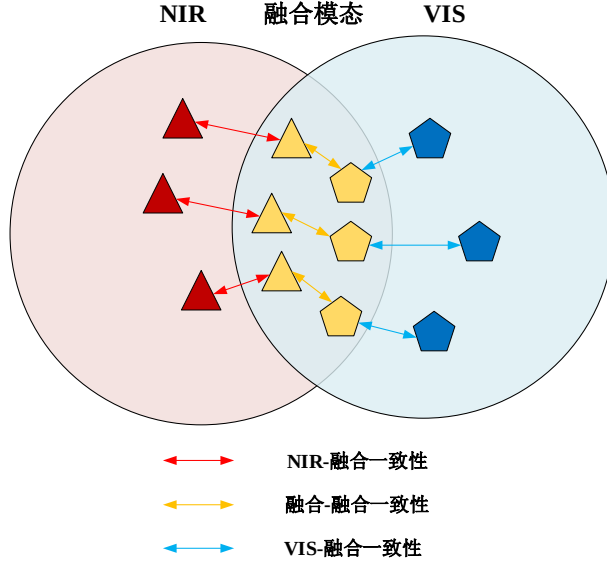


图 2-6 融合分布一致性 SDC 的示意图。其中包含 NIR-融合一致性、融合-融合一致性和 VIS-融合一致性

SDC 包含 NIR-融合一致性、融合-融合一致性和 VIS-融合一致性，可以表示为：

$$L_{SDC} = \kappa_1 l_{ns} + \kappa_2 l_{ss} + \kappa_3 l_{vs} \quad (2-9)$$

其中，

$$\begin{aligned}
 l_{ns} &= \frac{1}{c} \sum_i^c \|m_i^N - m_i^{SN}\|_2^2 \\
 l_{ss} &= \frac{1}{c} \sum_i^c \|m_i^{SN} - m_i^{SV}\|_2^2 \\
 l_{vs} &= \frac{1}{c} \sum_i^c \|m_i^V - m_i^{SV}\|_2^2
 \end{aligned}$$

其中， $m_i^N = \frac{1}{n_i^N} \sum_{k=1}^{n_i^N} \mathbf{1}\{l_k^N = i\} f(X_k^N; \Theta, W^I)$ 表示 NIR 人脸第 i 个身份的均值向量， $m_i^V = \frac{1}{n_i^V} \sum_{k=1}^{n_i^V} \mathbf{1}\{l_k^V = i\} f(X_k^V; \Theta, W^I)$ 表示 VIS 人脸第 i 个身份的均值向量， $m_i^{SN} = \frac{1}{n_i^N} \sum_{k=1}^{n_i^N} \mathbf{1}\{l_k^S = i\} f(X_k^{SN}; \Theta, W^I)$ 表示从 NIR 数据生成的融合图像第 i 个身份的均值向量， $m_i^{SV} = \frac{1}{n_i^V} \sum_{k=1}^{n_i^V} \mathbf{1}\{l_k^S = i\} f(X_k^{SV}; \Theta, W^I)$ 表示从 VIS 数据生成的融合图像第 i 个身份的均值向量， $\|\cdot\|_2$ 表示矩阵的 L_2 范数。

如图2-6所示，SDC 使用三个类内约束以增强类内紧凑性。并且有效确保了 NIR-融

算法 2-1: 融合空间学习 SSL 方法

Input: 训练集 X^N and X^V , 迭代次数 $iter$

Output: 网络参数 Θ, W^I

- 1: 用 MS-Celeb-1M 数据库预训练网络, 初始化 Θ 和 W^I ;
 - 2: **for** $i = 1, 2, \dots, iter$ **do**
 - 3: 更新 W_m 通过最小化 L_1^{MSL} ;
 - 4: 更新 $\Theta^{SMG}, \Theta, W^N, W^V, W_{id}^N, W_{id}^V$ 通过最小化 L_2^{MSL} ;
 - 5: 更新 $\Theta^{SMG}, \Theta, W^I, W_{id}$ 通过最小化 L_{id} ;
 - 6: 更新 $\Theta^{SMG}, \Theta, W^I$ 通过最小化 L_{SDC} ;
 - 7: **end for**
 - 8: **Return** Θ and W^I
-

合、融合-融合和 VIS-融合的特征分布一致性, 有助于学习既紧凑又具有区分性的表示。

2.2.5 模型优化

SSL 网络将融合模态生成器、模态对抗融合空间学习、融合分布一致性结合到统一的框架中, 其最终目标函数可以表示为:

$$\mathcal{L} = \varphi_1 L_{id} + \varphi_2 L_{MSL} + \varphi_3 L_{SDC} \quad (2-10)$$

模型的优化细节在算法 2-1 中进行了详细阐述, SSL 网络中的参数包括 Θ^{SMG} 、 Θ 、 W^N 、 W^V 、 W^I 、 W_m 、 W_{id}^N 、 W_{id}^V 、 W_{id} , 网络参数采用循环迭代的方式进行更新。

2.3 实验结果与分析

在本节中, 在三个具有挑战性的 NIR-VIS 数据集上验证 SSL 网络的性能, 包括 CASIA NIR-VIS 2.0^[12]、Oulu-CASIA NIR-VIS^[11]、LAMP-HQ NIR-VIS^[14]。首先, 简要介绍用于评估的数据集和实施细节。其次, 在 NIR-VIS 数据库上进行 SSL 的各种性能分析。最后, 将 SSL 模型与主流的方法进行比较, 以展示 SSL 的先进性能。

2.3.1 实验数据集

本方法采用的三个主流数据集简要概括如下：

CASIA NIR-VIS 2.0 dataset^[12] 是当前流行且具有挑战性的 NIR-VIS 人脸数据集之一。它包含超过 17,000 张图像，充分涵盖了光照、表情、姿势、遮挡、距离等剩余变化。训练集和测试集分别包含 357 个和 358 个身份类别。参考文献^[12] 中的十个评估协议 (Protocols) 来评估 SSL 模型。具体而言，本文基于第一个协议进行消融研究和各种性能分析，使用十个评估协议来展现整体性能。

Oulu-CASIA NIR-VIS dataset^[111] 是常用的 NIR-VIS 数据集之一。它包含 80 个受试者的 7,680 张图像，其中 50 个受试者来自奥卢大学，其余来自 CASIA。图像中的人脸具有六种表情 (惊讶、悲伤、愤怒、厌恶、快乐、恐惧) 和三种光照条件 (包括弱光、正常室内光照、黑暗场景)。这些变化因素进一步增加了数据集的挑战性。实验时，参考文献^[93] 中的评估协议，并且训练集和测试集之间没有交集。

LAMP-HQ NIR-VIS dataset^[14] 是近几年开源且最具挑战性的 NIR-VIS 数据集。该数据集包含来自 573 个受试者的 56,788 张 NIR 图像和 16,828 张 VIS 图像。数据集中的图像在面部表情、光照条件、遮挡方面有显著的不同。使用十个官方协议用于比较模型性能，并且训练集和测试集之间没有相同的身份。

2.3.2 实验设置

SSL 网络在 Tensorflow 上进行实现，使用的 GPU 为 TITAN XP。所有输入图像都通过 MTCNN^[124] 被裁剪到 160×160 的尺寸。 y^I 、 y^N 、 y^V 的维度设置为 256。采用预训练的 ResNet50-CBAM^[125,126] 作为 SSL 的主干网络 $\{\Theta^B, W^B\}$ 。按照文献^[93,94] 中的方法，首先，使用 MS-Celeb-1M^[109] 通过 softmax 损失函数对主干网络参数 $\{\Theta^B, W^B\}$ 进行预训练。然后，在另一个具有挑战性的 VIS 数据集 LFW^[127] 上进行测试时，预训练网络能获得 99.5% 的准确率。模态共享 $\{\Theta, W^I\}$ 和模态特定分支 $\{\Theta, W^N, W^V\}$ 是由 $\{\Theta^B, W^B\}$ 初始化的。在优化过程中，使用 RMSPROP 作为网络的优化器。在测试阶段，不引入 SMG 模块，而是直接采用模型提取的特征来进行匹配。初始学习率设置为 0.01，每 4 个 epoch 的衰减因子为 0.9。超参数 λ_1 、 λ_2 、 κ_1 、 κ_2 、 κ_3 、 φ_1 、 φ_2 、 φ_3 设置为 1.0、0.01、

1.0、0.1、1.0、1.0、1.0、0.1 以获得令人满意的性能。批处理大小和迭代次数分别设置为 64 和 600。

2.3.3 模块性能分析

在这一小节,进行了 SSL 的各种性质分析,包括消融实验和可视化分析,以展示每个组件的性能和 SSL 模型的有效性。

表 2-3 所提出的模型在 CASIA NIR-VIS 2.0 数据库上的消融实验

方法	Rank-1 (%)	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
BL	89.5	81.2	57.9
$BL + CE$	98.9	98.7	95.4
$BL + CE + L_{MSL}$	99.4	99.0	96.0
$BL + CE + L_{SDC}$	99.2	99.0	96.2
$BL + CE + L_{MSL} + L_{SDC}$	99.5	99.1	96.3

消融实验: 在两个具有挑战性的数据集上进行了大量的实验,即 CASIA NIR-VIS 2.0 和 Oulu-CASIA NIR-VIS,以验证 SSL 中各模块的有效性。表格 2-3 和表格 2-4 中展示了消融实验的结果。本节使用在 MS-Celeb-1M 上预训练的 ResNet50-CBAM 网络作为基线 (BL) 方法。BL+ L_{id} 表示使用损失函数 L_{id} 在 NIR-VIS 数据集上对预训练的网络进行微调。通过将每个目标函数 (即 L_{id} , L_{MSL} 、 L_{SDC}) 叠加到模型训练中,可以分析 SSL 中每个组件的影响。

首先,在 CASIA NIR-VIS 2.0 数据集 (第一个协议) 上,BL 表现出了不理想的性能,表现在 Rank-1 为 89.5%,VR@FAR=0.01% 为 57.9%。这表明在 VIS 数据集上预训练的模型无法很好地进行 NIR-VIS 人脸匹配。然后,使用损失函数 L_{id} 在 NIR-VIS 数据集上进行微调时,BL+ L_{id} 在 Rank-1 上比 BL 表现更好,提高了 9.4%,在 VR@FAR=0.01% 上提高了 37.5%。这说明预训练网络在 NIR-VIS 数据集上进行微调,可以在一定程度上减小模态差异对模型性能的影响。其次,BL+CE+ L_{MSL} 在 Rank-1 上相比 BL+CE 提高了 0.5%,在 VR@FAR=0.01% 上提高了 0.6%,表明 MSL 能够将 NIR 和 VIS 图像投影到一个共同的融合空间,并有效减小模态差异。BL+CE+ L_{SDC} 在 Rank-1 上相比 BL+CE 提高了 0.3%,在 VR@FAR=0.01% 上提高了 0.8%。结果表明 SDC 可以通过探索跨模态特征分布之间的一致性来增强身份判别性表示。最后,与 BL+CE+ L_{MSL} 相

比, $BL+CE+L_{MSL}+L_{SDC}$ 在 Rank-1 上提高了 0.1%, 在 $VR@FAR=0.01\%$ 上提高了 0.3%。同时, $BL+CE+L_{MSL}+L_{SDC}$ 在 Rank-1 和 $VR@FAR=0.01\%$ 上分别超过了 $BL+CE+L_{SDC}$ 0.3% 和 0.1%。这表明将 MSL 和 SDC 结合训练模型能够获得更好的性能。

表 2-4 所提出的模型在 Oulu-CASIA NIR-VIS 数据库上的消融实验

方法	$VR@FAR=1\%$ (%)	$VR@FAR=0.1\%$ (%)	$VR@FAR=0.01\%$ (%)
BL	92.7	81.4	66.4
$BL + CE$	97.5	90.8	82.5
$BL + CE + L_{MSL}$	97.9	93.2	85.4
$BL + CE + L_{SDC}$	97.8	93.0	85.1
$BL + CE + L_{MSL} + L_{SDC}$	98.1	93.3	85.6

此外, 如表2-4所示, Oulu-CASIA NIR-VIS 数据集上也可以观察到和 CASIA NIR-VIS 2.0 中类似的结果: (1) $BL+CE$ 在 $VR@FAR=0.1\%$ 和 $VR@FAR=0.01\%$ 上分别比 BL 提高了 9.4% 和 16.1%。(2) 与 $BL+CE$ 相比, $BL+CE+L_{MSL}$ 在 $VR@FAR=0.1\%$ 和 $VR@FAR=0.01\%$ 上取得更好的性能, 分别提高了 2.4% 和 2.9%。(3) $BL+CE+L_{SDC}$ 在 $VR@FAR=0.1\%$ 和 $VR@FAR=0.01\%$ 上比 $BL+CE$ 分别提高了 2.2% 和 2.6%。(4) $BL+CE+L_{MSL}+L_{SDC}$ 在 $VR@FAR=0.1\%$ 和 $VR@FAR=0.01\%$ 上分别超过了 $BL+CE+L_{MSL}$ 0.1% 和 0.3% 以及 $BL+CE+L_{SDC}$ 0.2% 和 0.5%。这表明通过融合空间学习可以进一步提高跨模态匹配性能。

SDC 分析: 在表格2-5中, 展示了 SDC 中 l_{ns} 、 l_{ss} 、 l_{vs} 的有效性, 即验证了三种情况: $BL+CE+L_{SDC}(w/o \ l_{ns})$ 、 $BL+CE+L_{SDC}(w/o \ l_{ss})$ 和 $BL+CE+L_{SDC}(w/o \ l_{vs})$ 。结果表明, $BL+CE+L_{SDC}$ 在 $VR@FAR=0.01\%$ 和 $VR@FAR=0.001\%$ 方面优于 $BL+CE+L_{SDC}(w/o \ l_{ns})$, 分别提高了 0.5% 和 2.3%。类似地, $BL+CE+L_{SDC}$ 的性能在 $VR@FAR=0.01\%$ 和 $VR@FAR=0.001\%$ 方面优于 $BL+CE+L_{SDC}(w/o \ l_{ss})$, 分别提高了 0.3% 和 1.5%。此外, $BL+CE+L_{SDC}$ 在 $VR@FAR=0.01\%$ 和 $VR@FAR=0.001\%$ 方面优于 $BL+CE+L_{SDC}(w/o \ l_{vs})$, 分别提高了 0.4% 和 1.4%。最终, 通过引入 SDC 中的三个类内约束, SSL 获得了令人满意的性能, 其中 $VR@FAR=0.01\%$ 为 96.2% 以及 $VR@FAR=0.001\%$ 为 84.2%。

SMG 分析: 对模态特定编码层和模态共享解码层的数量进行消融实验, 结果如表2-6所示。这里, n 和 m 分别表示模态特定编码层和模态共享解码层中的卷积层数。从结

表 2-5 SDC 在 CASIA NIR-VIS 2.0 数据集上的消融实验分析

方法	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)	VR@FAR=0.001%(%)
$BL + CE + L_{SDC}(w/o \ l_{ns})$	98.8	95.7	81.9
$BL + CE + L_{SDC}(w/o \ l_{ss})$	98.8	95.9	82.7
$BL + CE + L_{SDC}(w/o \ l_{vs})$	98.9	95.8	82.8
$BL + CE + L_{SDC}$	99.0	96.2	84.2

果可以看出, 增加 n 会在模型性能上带来一些改善, 当 n 等于 4 时性能趋于饱和。此外, 随着 m 的增加, 模型性能略有提升, 最佳性能在 $m=4$ 时实现。然而, 当进一步增加 m 到 8 时, 模型性能开始下降。总体而言, 在参数成本和模型性能之间的权衡考虑下, 默认设置 ($n=2$ 和 $m=1$) 可以获得令人满意的结果。为了展示人脸合成网络 SMG 的效率, 可以将其的参数和浮点运算数 (FLoating-point of OPerations, FLOPs) 与 CFC^[27] 和 SAFE^[74] 进行了比较。图像级生成器的参数和 FLOPs 显示在表2-7中。CFC 通过其精心设计的人脸转换网络, 能够将具有多种姿势的 NIR 图像有效地合成为正面姿势的 VIS 图像。然而, CFC 的人脸转换网络依赖于一个庞大的生成器。因此, CFC 的人脸转换网络可能会遇到低推理效率的问题, 这体现在其参数/FLOPs 高达 33.49M/36.26G。此外, SAFE 利用 3D 对齐的人脸作为辅助, 并引入像素级概率损失来训练用于轻量级 GANs。与 CFC 相比, SAFE 实现了更低的参数/FLOPs, 为 0.26M/0.57G。最后, SMG 的参数和 FLOPS 分别为 11 和 358,400, 相较于 CFC 和 SAFE 的人脸转换网络, 具有更高的推理效率。

表 2-6 SMG 在 CASIA NIR-VIS 2.0 上不同设置分析

n	m	Rank-1 (%)
1	1	99.3
2	1	99.5
4	1	99.5
8	1	99.5
2	2	99.6
2	4	99.6
2	8	99.4

表 2-7 SSL、CFC^[27] 和 SAFE^[74] 的人脸变换网络的参数和浮点运算数 FLOPs

Method	Parameters	FLOPs
CFC	33.49M	36.26G
SAFE	0.26M	0.57G
SSL	11	358,400

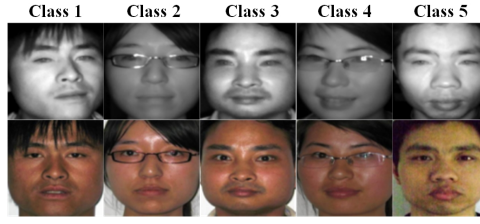


图 2-7 CASIA NIR-VIS 2.0 数据集上选择了五个主体进行特征分布的可视化分析

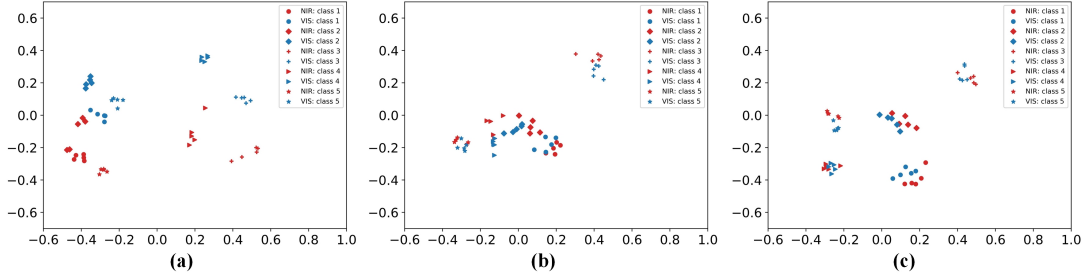


图 2-8 特征分布的可视化结果如下: (a) 是基线模型 (BL) 的分布; (b) 是在 MITML^[119] 中使用 BL+CE+ 对抗学习策略的分布; (c) 是在 SSL 中使用 BL+CE+MAL 的分布

MAL 分析: 进行可视化实验来验证 MAL 的有效性。图 2-7 展示了 CASIA NIR-VIS 2.0 数据集中五个身份的特征分布的可视化分析。将 BL+CE+MAL 与 MITML^[119] 中的 BL+CE+ 对抗学习策略的特征分布进行了比较。从图 2-5 (c) 和 2-5 (d) 可以看出, MITML 和 SSL 的对抗学习策略之间的区别在于 MITML 缺乏足够的引导使得分类器获得模态无关类别的信息。相反, 本章的假设明确地使分类器学习到融合模态代表模态无关的类别, 有助于主干网络学习到模态无关的特征。此外, 图 2-8 (a) 说明了 BL 方法中提取的特征存在明显的模态差异。在 MITML 和 SSL 中, CE 和对抗学习策略使得模态差异得到了有效减小。MAL 中的明确了模态无关学习目标, 这有助于学习更具有判别性的特征。

可视化分析: 为了更好地展示 SSL 网络的鲁棒性, 在 LAMP-HQ 数据集上进行了类内和类间相似性结果的分布以及可视化匹配结果。基线模型 BL 和 SSL 提取的特征相似性分布的可视化如图2-9所示。负样本对之间的特征相似性分布以红色显示, 正样本对之间的特征相似性分布以绿色显示。可以观察到 SSL 网络能够有效增加类内相似性并减小类间相似性, 从而实现对 NIR-VIS 人脸的有效识别。此外, 在图2-10中展示了正确分类的结果, 其中正确匹配的样本对之间的相似度分数明显高于其他结果。表明 SSL 模型对于一些复杂情况 (如轻微眼镜遮挡、视角轻微变化、光照条件、年龄等) 具有良好的泛化性。图2-10中也展示了一些错误匹配的结果, 可以发现极端的遮挡和姿态是导致错

误匹配的主要原因。

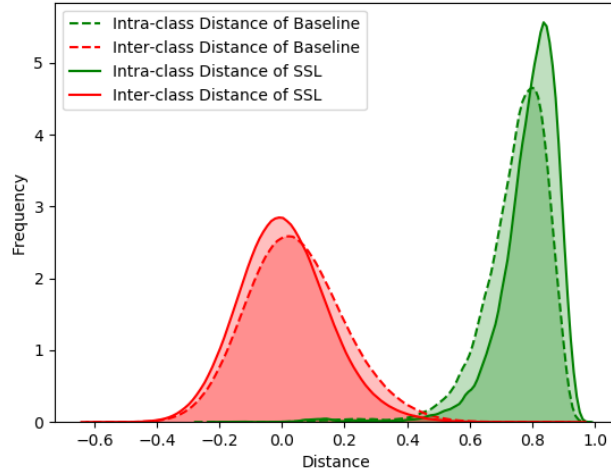


图 2-9 特征相似性分布的可视化。红色表示类间相似性，绿色表示类内相似性



图 2-10 在 LAMP-HQ NIR-VIS 数据集上，NIR 图像与 VIS 图像之间的可视化匹配结果。‘GT’ 表示正确的匹配图像

2.3.4 与主流的方法对比

在本节中，SSL 方法与多个流行的跨模态人脸识别方法进行比较，以验证 SSL 在三个主流数据集上具有竞争性的识别性能。

CASIA NIR-VIS 2.0 实验结果: 表格2-8展示了 SSL 与主流方法的比较结果。首先，SSL 优于传统的基于 CNN 的方法，包括 Light CNN-29^[129] 和 TRIVET^[92]。例如，与 Light

表 2-8 不同方法在 CASIA NIR-VIS 2.0 数据库上的结果

方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
TRIVET ^[92] (ICB 2016)	95.7 ± 0.5	98.1 ± 0.3	91.0 ± 1.3
IDR ^[128] (AAAI 2017)	97.3 ± 0.4	98.9 ± 0.3	95.7 ± 0.7
Light CNN-29 ^[129] (TIFS 2018)	96.7 ± 0.2	-	94.8 ± 0.4
ADFL ^[76] (AAAI 2018)	98.1 ± 0.4	99.1 ± 0.2	97.1 ± 0.5
CDL ^[130] (AAAI 2018)	98.6 ± 0.2	-	98.3 ± 0.5
WCNN ^[93] (TPAMI 2018)	98.7 ± 0.3	99.5 ± 0.1	98.4 ± 0.4
PCFH ^[70] (IJCAI 2019)	98.8 ± 0.3	99.6 ± 0.1	97.7 ± 0.3
PACH ^[71] (CVPR 2020)	98.9 ± 0.2	99.6 ± 0.1	98.3 ± 0.2
CFC ^[27] (TPAMI 2020)	98.6 ± 0.1	99.2 ± 0.1	97.3 ± 0.2
SDDL ^[106] (PR 2021)	92.1 ± 0.7	-	88.7 ± 1.0
ADCANs ^[131] (TMM 2021)	99.1 ± 0.2	99.6 ± 0.1	98.5 ± 0.2
VSANet ^[14] (IJCV 2021)	99.2 ± 0.2	99.6 ± 0.1	98.3 ± 0.2
EGCH ^[82] (ICPR 2021)	98.8 ± 0.2	99.6 ± 0.0	97.7 ± 0.2
MMTN ^[72] (TIFS 2022)	99.4 ± 0.0	99.8 ± 0.0	98.7 ± 0.0
SAFE ^[74] (TPAMI 2022)	98.9 ± 0.1	99.6 ± 0.1	98.6 ± 0.1
SSL (本章方法)	99.3 ± 0.2	99.9 ± 0.0	99.0 ± 0.2

CNN-29 相比, SSL 在 Rank-1 上提高了 2.6%, 在 VR@FAR=0.1% 上提高了 4.2%。这些结果表明, 融合模态对抗学习策略对于跨模态人脸识别任务是有效的。SSL 优于近几年发布的方法, 如 IDR^[128]、ADFL^[76]、CDL^[130]、WCNN^[93]、PCFH^[70]、PACH^[71]、CFC^[27]、SDDL^[106]、ADCANs^[131]、VSANet^[14]、EGCH^[82]、MMTN^[72]、SAFE^[74]。例如, SSL 在 VR@FAR=1% 上超过 VSANet 0.3%, 在 VR@FAR=0.1% 上超过 0.7%。SSL 在 CASIA NIR-VIS 2.0 数据集上的卓越性能表明, 学习融合空间可以有效提高模型性能。

Oulu-CASIA NIR-VIS 实验结果: 比较的结果显示在表2-9中。与经典的 CNN 方法(包括 LightCNN-29^[129] 和 TRIVET^[92]) 相比, SSL 取得了明显的优势。与主流的深度学习方法相比, SSL 取得了一些改进。具体而言, SSL 在 VR@FAR=0.1% 上的表现优于 SAFE^[74], 提高了 1.5%。结果表明, 通过生成融合图像来探索一个融合模态的空间, SSL 能够提取模态不变表示。

LAMP-HQ NIR-VIS 实验结果: LAMP-HQ NIR-VIS 数据库是一个近几年发布的数据集, 因此许多方法并没有报告在这个数据集上的性能。SSL 方法与一些方法进行比较, 包括 Light CNN-29^[129]、ADFL^[76]、PCFH^[70]、PACH^[71]、CFC^[27]、ADCANs^[131]、VSANet^[14]、MMTN^[72]。如表2-10所示, SSL 获得了最好的结果。例如, SSL 的性能优于 ADCANs, 在 Rank-1 上提升了 1.0%, 在 VR@FAR=1% 上提升了 0.7%, 在 VR@FAR=0.1% 上提升了 2.2%。SSL 在最具挑战性的数据集上优越的性能说明该模型能够在融合空间中学到更多与身份相关的信息, 并减少模态干扰。

表 2-9 不同方法在 Oulu-CASIA NIR-VIS 数据库上的结果

方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
TRIVET ^[92] (ICB 2016)	92.2	67.9	33.6
IDR ^[128] (AAAI 2017)	94.3	73.4	46.2
Light CNN-29 ^[129] (TIFS 2018)	96.7	92.4	65.1
ADFL ^[76] (AAAI 2018)	95.5	83.0	60.7
CDL ^[130] (AAAI 2018)	94.3	81.6	53.9
WCNN ^[93] (TPAMI 2018)	98.0	81.5	54.6
PCFH ^[70] (IJCAI 2019)	100	97.7	86.6
PACH ^[71] (CVPR 2020)	100	97.9	88.2
CFC ^[27] (TPAMI 2020)	99.9	98.1	90.7
ADCANs ^[131] (TMM 2021)	99.8	93.2	78.9
VSANet ^[14] (IJCV 2021)	100	97.7	89.0
EGCH ^[82] (ICPR 2021)	100	98.0	89.5
MMTN ^[72] (TIFS 2022)	100	99.2	89.0
SAFE ^[74] (TPAMI 2022)	100	98.3	91.8
SSL (本章方法)	100	98.1	93.3

表 2-10 不同方法在 LAMP-HQ NIR-VIS 数据库上的结果

方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
Light CNN-29 ^[129] (TIFS 2018)	94.6 ± 0.3	92.5 ± 0.6	75.6 ± 1.9
ADFL ^[76] (AAAI 2018)	95.1 ± 0.5	92.1 ± 0.9	73.3 ± 2.2
PCFH ^[70] (IJCAI 2019)	95.3 ± 0.5	92.9 ± 0.6	75.1 ± 1.8
PACH ^[71] (CVPR 2020)	95.4 ± 0.5	93.1 ± 0.4	75.3 ± 1.7
ADCANs ^[131] (TMM 2021)	97.8 ± 0.3	98.8 ± 0.3	96.1 ± 0.4
VSANet ^[14] (IJCV 2021)	97.3 ± 0.2	94.8 ± 0.7	78.2 ± 3.0
MMTN ^[72] (TIFS 2022)	97.7 ± 0.0	95.7 ± 0.6	79.2 ± 0.0
SSL (本章方法)	98.8 ± 0.3	99.5 ± 0.3	98.3 ± 0.3

2.3.5 通用性分析

为了进一步验证 SSL 方法在其他跨模态任务中的通用性 (Generalizability), 按照文献^[94]中的协议, 在具有挑战性的 VIS-Thermal Tufts Face^[132] 数据库上进行实验。热红外图像是使用长波红外光谱拍摄的, 其包含的纹理信息与可见光图像相比较少。训练集包括从 50 名受试者中收集的 1,490 张可见光图像和 450 张热红外图像。另一方面, 测试集包含另一部分的 63 名受试者中拍摄得到的 1,916 张可见光图像和 530 张热红外图像。从表2-11中显示的结果中, 可以得出以下结论。首先, $BL+CE$ 在 Rank-1 和 VR@FAR=1% 方面分别超过了 BL 46.0% 和 48.9%。此外, $BL+CE+L_{MSL}$ 在 Rank-1 和 VR@FAR=1% 方面的性能均高于 $BL+CE$, 分别提高了约 2.1% 和 2.6%。 $BL+CE+L_{SDC}$ 在 Rank-1 和 VR@FAR=1% 方面的性能也优于 $BL+CE$, 分别提高了约 4.0% 和 5.4%。最终, SSL 获得了最佳性能, Rank-1 为 77.3%, VR@FAR=1% 为 60.0%。这些结果表明, 通过将

跨模态模态图像投影到一个融合空间中，可以有效地学习模态不变的表示。其次，与 LightCNN 相比，SSL 在 Rank-1 和 VR@FAR=1% 方面的提升高达 27.1% 和 30.2%，表明 SSL 可以学到更具判别力的特征。以上结果表明，即使面对更大的模态差异，SSL 也表现出良好的性能，验证了 SSL 在其他模态人脸匹配任务上的通用性。

表 2-11 VIS-Thermal Tufts Face 数据库上通用性分析

Method	Rank-1 (%)	VR@FAR=1%(%)
LightCNN ^[129]	50.2	29.8
<i>BL</i>	25.7	2.4
<i>BL + CE</i>	71.7	51.3
<i>BL + CE + L_{MSL}</i>	73.8	53.9
<i>BL + CE + L_{SDC}</i>	75.7	56.7
SSL (本章方法)	77.3	60.0

2.4 本章小结

针对模态差异问题，本章提出了一种新颖的融合空间学习 SSL 方法，它包括三个关键模块：融合模态生成器 SMG，模态对抗融合空间学习 MSL、融合分布一致性 SDC。首先，SMG 通过浅层 CNN 的通道级卷积将 NIR 和 VIS 图像合成为融合图像。特别是，它能够很好地保留图像中的判别性结构信息，并且可以通过自监督方式进一步提高图像的质量。其次，MSL 通过融合模态对抗学习策略将 NIR 和 VIS 图像投影到融合空间，采用融合模态引导学习的目标函数，有效减少 NIR-VIS 人脸中存在的模态差异。最后，由 NIR-融合，融合-融合和 VIS-融合一致性构成的 SDC 可以增强类内紧凑性，有助于学习判别性表示。因此，SSL 方法巧妙地将 SMG、MSL、SDC 集成到一个联合框架中，以有效解决 NIR-VIS 人脸识别任务中存在的模态差异问题并提高模型性能。

第 3 章 中立人脸学习和渐进融合合成方法

NIR-VIS 人脸中不仅具有模态差异，还有明显的表观差异，被称为剩余差异，这两种差异都为 NIR-VIS 人脸识别带来一定的挑战。模态差异指的是匹配的图像位于不同的模态中，而剩余差异意味着匹配的图像在姿态、表情、光照等方面存在显著差异。这两种情况都给 NIR-VIS 人脸匹配带来了挑战。在上一章中，本文已经针对模态差异问题提出了有效的解决方法，但是现有方法很少针对剩余差异提出有效的解决方案。为了同时消除模态差异和剩余差异，本章提出了一种新颖的中立人脸学习和渐进融合合成 (Neutral face Learning and Progressive Fusion synthesis, NLPF) 网络，以解耦异质人脸的潜在属性并学习中立人脸表示。NLPF 巧妙地将身份相关中立人脸学习 (Identity-related Neutral face Learning, INL) 和属性渐进融合 (Attribute Progressive Fusion, APF) 整合到一个联合框架中。首先，INL 通过引导网络学习同模态的中立人脸特征表示，它将跨模态图像映射到共同的中立表示子空间来解决模态差异和剩余差异的挑战。此外，APF 以一种渐进融合的方式进行身份相关特征、模态相关特征、剩余特征的解耦和重新融合，有助于进一步提纯身份相关特征。在三个主流的 NIR-VIS 数据集上进行了全面评估，验证了 NLPF 模型的鲁棒性和有效性。特别是，NLPF 在具挑战性的 NIR-VIS 数据集 LAMP-HQ NIR-VIS^[14] 上取得了具有竞争力的识别性能。这一研究成果已在 IEEE Transactions on Circuits and Systems for Video Technology 正式发表。

3.1 引言

最近，在多媒体信息检索、安全系统和刑事调查等领域，往往需要在低光照场景中部署人脸识别系统^[44,66,133,134]。然而，在低光照场景中，可见光传感器难以获取清晰可辨的人脸图像，而近红外传感器可以收集清晰 NIR 人脸图像。由于在不同的光谱波段获取了人脸图像，导致在跨模态人脸匹配时存在较大的模态差异。此外，除了模态差异，剩余差异导致的面部变化进一步增加了类内差异，为 NIR-VIS 人脸匹配带来了不小的挑战。因此，作为异构人脸识别中最具代表性的任务之一，NIR 和 VIS 人脸识别受到了

越来越多的关注。

为了减小模态差异并提高识别性能,学术界已经提出了大量方法来尝试解决跨模态人脸匹配问题。Yang 等人^[28]将异质人脸解耦为与身份无关的表示和与身份相关的表示。并提出了一个人脸合成模块,随机重组这两种表示生成异质人脸,以丰富异质人脸图像的属性多样性。Liu 等人^[75]利用身份信息部分、模态部分、基于高斯噪声的剩余部分来表征异质人脸潜在空间中的主要属性特征。接着,通过交换模态部分来实现模态之间的转换,从而将 NIR 图像转换为保持身份的 VIS 图像。如图3-1所示, NIR-VIS 人脸具有模态差异和剩余差异的特点。上述提到的方法可以解决模态差异问题,但忽略了剩余差异问题。最近,一些工作使用中立人脸来解决剩余差异问题^[79,135]。中立人脸的身份信息不受多样面部变化的污染,使其成为解决剩余差异问题的可行解决方案。Pang 等人^[79]提出了一个双向异质原型学习框架,通过使用中立人脸作为监督信号来解决剩余差异问题。Hu 等人^[135]旨在消除类内变化,并提出了一个双重面部对齐学习网络,在特征级和图像级学习中立表示。通过这些工作可以观察到,学习中立人脸表示具有解决剩余差异问题和消除类内差异的潜力。但这些方法只考虑了中立人脸对齐,而忽略了非中立人脸中有价值的特征解耦。

受到上述分析的启发,本章提出了一种新颖的中立人脸学习和渐进融合合成 (Neutral face Learning and Progressive Fusion synthesis, NLPF) 网络。NLPF 引入了两个有效的模块,包括身份相关中立人脸学习 (Identity-related Neutral face Learning, INL) 和属性渐进融合 (Attribute Progressive Fusion, APF)。NLPF 从以下两个角度解决了模态差异和剩余差异问题,并获取了有辨识性的表示: **1) 引导网络学习同模态的中立人脸表示,从而有效提取身份相关的鲁棒表示。2) 通过在图像层面和特征层面之间进行转换,解耦和重新融合面部属性特征,以提纯身份相关信息。**具体而言, NLPF 方法将异质人脸建模为三个因素,即身份相关因素、模态相关因素、剩余因素。身份相关因素直接反映辨识性信息,用于进行鲁棒的人脸匹配。模态相关因素表示 NIR 与 VIS 人脸的光谱信息。剩余因素是由遮挡、表情、光照、姿态等引起的明显面部变化。INL 引导网络学习同模态的中立人脸特征表示,消除了模态差异并减小了由剩余差异引入的类内距离。一方面,身份相关表示投影到共同的中立表示子空间,以获取有辨识性的特征。另一方面,异质问题被转化为同质问题,从而降低了特征空间中人脸匹配的难度。在 APF 中,异质图

像本身被作为监督信号，在渐进融合的过程中解耦和重新融合身份相关特征、模态相关特征、剩余特征。特别地，在渐进融合过程中可以进一步提纯身份相关特征。



图 3-1 除了巨大的模态差异之外，非中立的异质人脸还附加了由姿态、遮挡、表情、光照引起的剩余信息，这进一步增加了 NIR-VIS 人脸匹配的难度。相反，在中立人脸中可以提取清晰的面部身份信息。数据来源于 CASIA NIR-VIS 2.0 数据集^[12]

综上所述，本章的贡献可以总结为四个方面：

- NLPF 模型将 INL 和 APF 结合到一个联合框架中，该联合方案能够解决模态差异和剩余差异问题，同时解耦面部属性表示以提纯身份相关表示。
- 引入的 INL 模块通过引导网络学习同质中立人脸表示，消除了模态差异和剩余差异。在重建同质中立人脸的过程中，身份相关特征被映射到共同的中立表示子空间。
- APF 通过渐进融合过程执行身份相关特征、模态相关特征、剩余特征的解耦和重新融合。所提出的渐进融合过程促进了面部属性的解耦，有助于进一步提纯身份相关特征。
- 在三个具有挑战性的数据集上进行了全面评估，即 CASIA NIR-VIS 2.0^[12]、BUAA NIR-VIS^[112]、LAMP-HQ NIR-VIS^[14]，验证了 NLPF 方法的有效性和鲁棒性。

3.2 中立人脸学习和渐进融合合成模型

在本节中，详细介绍了提出的中立人脸学习和渐进融合合成 NLPF 网络。图3-2展示了 NLPF 模型的框架。NLPF 模型包括两个有效的模块，即身份相关中立人脸学习 INL 和属性渐进融合 APF。NLPF 模型的目标是学习同模态下的中立人脸表示，并解耦面部属性的表示提纯身份特征。为了实现上述目标，INL 通过同质中立人脸重建以尽可能地缩小模态差异，并通过去除表情、姿态、遮挡和光照引起的剩余信息来缩小类内差异。此外，APF 通过身份相关特征、模态相关特征、剩余特征的解耦与重新融合来重建异质人脸，以获取具有区分性的身份相关表示。NLPF 中的一些主要符号在表3-1中显示。

表 3-1 NLPF 中符号的含义。

符号	含义
x_i^N	NIR 模态下的图像
x_i^V	VIS 模态下的图像
l_i^N / l_{mi}^N	NIR 图像 x_i^N 的身份/模态标签
l_i^V / l_{mi}^V	VIS 图像 x_i^V 的身份/模态标签
X_V^{neu}, X_N^{neu}	中立人脸图片
$\tilde{X}_V^{neu}$	生成的中立人脸图片
$\tilde{X}_V, \tilde{X}_N$	生成的异质人脸图片
D	鉴别器
G^I, G^M, G^R, G^N	子解码器
t^I	身份相关特征
t^M	模态相关特征
t^R	剩余特征

3.2.1 问题建模

网络以异质人脸数据集 $X = \{X_N, X_V\}$ 作为训练数据。 $X_N = \{(x_i^N, l_i^N, l_{mi}^N)\}_{i=1}^{n^N}$ 和 $X_V = \{(x_i^V, l_i^V, l_{mi}^V)\}_{i=1}^{n^V}$ 表示 NIR-VIS 数据集，其中 x_i^N 和 x_i^V 表示图像数据。 $l_i^N \in \{1, 2, \dots, c\}$ 和 $l_i^V \in \{1, 2, \dots, c\}$ 代表身份相关标签， $l_{mi}^N \in \{NIR\}$ 和 $l_{mi}^V \in \{VIS\}$ 代表模态相关标签。 n^N 和 n^V 分别表示 NIR 和 VIS 数据的数量。接着，从数据集中选择出了中立人脸 $\{X_V^{neu}, X_N^{neu}\}$ 。具体而言，使用基于 VGG-19 的面部表情识别网络^①和面部姿态估计器网络^②来初始选择具有中立表情和正脸姿态的人脸。然后进行手动过滤，以

^① <https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch>

^② <https://github.com/Merle314/FacePoseEstimation>

找到每个人的唯一的 VIS 和 NIR 中立人脸。在消除模态差异的过程中，NLPF 统一将人脸图像转换为 VIS 模态，不使用 NIR 中立人脸 X_N^{neu} 作为监督信号。通过减小模态差异和剩余变化，NLPF 可以实现准确的 NIR-VIS 人脸匹配。

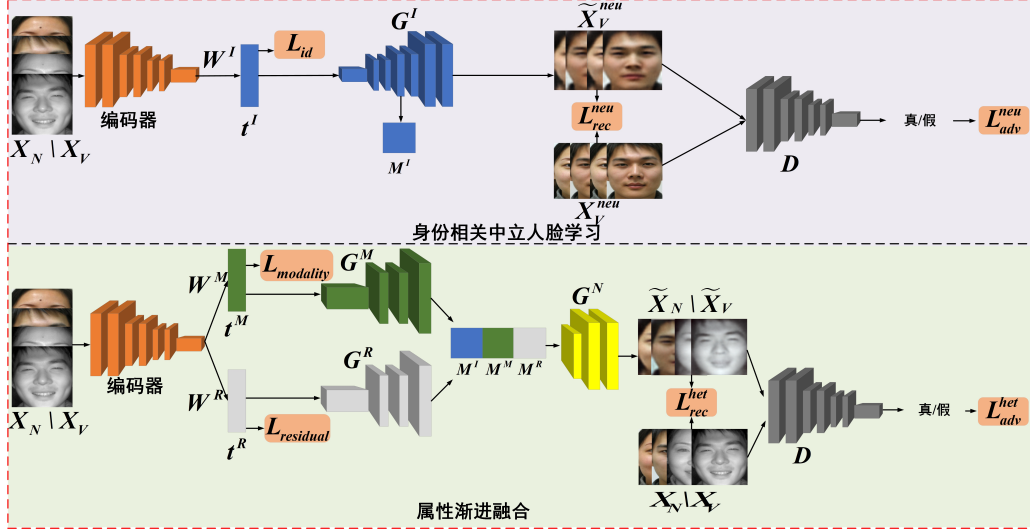


图 3-2 NLPF 模型框架。NLPF 网络包括两个有效模块，即身份相关中立人脸学习 INL 和属性渐进融合 APF。这一联合方案能够解决模态差异和剩余差异问题，并对面部属性表示进行建模与解耦，以提纯身份相关的表征

3.2.2 身份相关中立人脸学习

许多现有的模型^[76,93,130]已经提出了有效的算法来减小模态差异。在文献^[76]中，提出了特征空间对抗学习减小 NIR 和 VIS 特征之间的分布差异。在文献^[130]中，提出了跨模态排名缩小 NIR 和 VIS 特征在一个共同子空间中的差异。然而，它们忽略了剩余差异。实际上，剩余差异可能导致类内距离大于类间距离，从而降低模型性能。先前的研究^[79,135]已经证明，学习中立人脸表示可以消除类内差异。为了解决模态差异和剩余差异，NLPF 考虑使用 VIS 中立人脸来学习具有区分性的身份特征。

为了从图像中提取异质人脸的身份相关信息，引入了身份相关的全连接结构来学习身份相关特征。对于 VIS 图像和 NIR 图像 $X \in \{X_N, X_V\}$ ，身份相关表示 t^I 的形式如下：

$$t^I = f(X; \theta, W^I) \quad (3-1)$$

其中， $f(\cdot)$ 表示网络映射函数， θ 表示主干网络的参数， W^I 表示身份相关的全连接结

构。具体而言,采用 softmax 函数作为监督损失,并使用身份相关标签作为监督信号。详细的目标函数可以表示为:

$$\begin{aligned} L_{id} &= \text{softmax}(t^I; \theta, W^I, W_{id}^I) \\ &= -\frac{1}{n^V + n^N} \sum_{j=1}^{n^V + n^N} \sum_{i=1}^c \mathbf{1}\{l_j^M = i\} \log \widehat{p}_{ij} \end{aligned} \quad (3-2)$$

其中, W_{id}^I 代表身份相关的分类器层, $\mathbf{1}\{\cdot\}$ 是指示函数, $l_j^M \in \{l_j^N, l_j^V\}$ 表示身份标签, $\widehat{p}_{ij}$ 表示身份相关分类器的预测概率。

为了将跨模态图像映射到一个共同的中立表示子空间中,学习同模态的中立身份相关表示,本章设计了一个解码器 G^I ,用于从异质人脸的身份相关特征中进行上采样,以重建同模态的中立人脸。使用 VIS 中立人脸 X_V^{neu} 作为监督信号,与重建的中立人脸 $\tilde{X}_V^{neu}$ 进行人脸对齐。参考文献^[73,75]中的方法,使用 L_1 范数进行图像重建。人脸对齐的重构目标可以表示为:

$$L_{rec}^{neu} = \|X_V^{neu} - \tilde{X}_V^{neu}\|_1 \quad (3-3)$$

其中 $\|\cdot\|_1$ 表示矩阵 L_1 范数, $\tilde{X}_V^{neu}$ 代表了重建的 VIS 中立人脸。同时,图像的纹理细节作为高频信息直接影响重建图像的质量。为了获得质量高、纹理细节精致的图像,采用了一个多尺度鉴别器^[78] D 与 G^I 进行最小最大博弈。 D 的目标是区分重建图像 $\tilde{X}_V^{neu}$ 的真实性,而 G^I 生成更为真实的图像以迷惑 D 。这个对抗过程表达如下:

$$L_{adv}^{neu} = \mathbb{E}[\log(D(X_V^{neu}))] + \mathbb{E}[\log(1 - D(G^I(t^I)))] \quad (3-4)$$

其中 $G^I(t^I) = \tilde{X}_V^{neu}$ 表示以 t^I 为输入,通过 G^I 重建的中立 VIS 人脸图像。

最终, INL 中的目标损失可以写成:

$$L^{INL} = L_{id} + L_{rec}^{neu} + L_{adv}^{neu} \quad (3-5)$$

其中, L_{id} 的目标是提取与身份相关的特征,而 L_{rec}^{neu} 通过将跨模态图像映射到共同的中

立表示子空间，消除了身份相关特征中模态差异和剩余差异。最后， L_{adv}^{neu} 通过对抗过程丰富了重建图像的纹理细节。

3.2.3 属性渐进融合

解耦身份相关和身份无关的特征有助于提高识别性能^[75,93,98,108]。在本小节中，假设在 NIR-VIS 人脸中，身份和模态是普遍存在的属性，与前两者无关的其他表征属性统称为剩余属性。NLPF 在特征级别从 NIR-VIS 人脸中提取这些属性，并将其用于恢复 NIR-VIS 人脸来实现解耦过程。因此，NLPF 以一种渐进融合的方式执行与身份相关的特征、模态相关的特征、剩余特征的提取和重新融合。通过这种方式，NLPF 可以巧妙地分离这三个特征，以有效提纯身份信息。

为了解耦与身份相关的特征之外的其他潜在面部属性，NLPF 引入了模态相关的全连接结构来提取模态相关的特征，具体形式如下：

$$t^M = f(X; \theta, W^M) \quad (3-6)$$

其中， W^M 代表模态相关的全连接结构。具体而言，采用 softmax 函数，并使用模态标签作为监督信号，具体如下：

$$\begin{aligned} L_{modality} &= softmax(t^M; \theta, W^M, W_{modality}^M) \\ &= -\frac{1}{n^V} \sum_{j=1}^{n^V} \sum_{i \in \{VIS\}} \mathbf{1}\{l_{mj}^V = i\} \log \hat{p}_j^M \\ &\quad - \frac{1}{n^N} \sum_{j=1}^{n^N} \sum_{i \in \{NIR\}} \mathbf{1}\{l_{mj}^N = i\} \log \hat{p}_j^M \end{aligned} \quad (3-7)$$

其中， $W_{modality}^M$ 代表模态相关的分类器层， $\hat{p}_j^M$ 表示模态相关分类器的预测概率。

不同于提取身份相关和模态相关特征的方法，数据集中没有直接与剩余信息相关的标签。剩余特征无法通过使用给定的标签信息作为监督信号以优化分类损失函数的方式来提取。受文献^[98]的启发，NLPF 以对抗的方式从主干网络中提取剩余特征，其形式如下：

$$t^R = f(X; \theta, W^R) \quad (3-8)$$

其中, W^R 代表剩余全连接结构。目标函数可以表示为:

$$L_{residual} = \frac{1}{c} \sum_{r=1}^c \frac{1}{n_r^V} \sum_{i=1}^{n_r^V} \|f(x_i^V; \theta, W^R) - m_r^V\|_2^2 + \frac{1}{c} \sum_{r=1}^c \frac{1}{n_r^N} \sum_{i=1}^{n_r^N} \|f(x_i^N; \theta, W^R) - m_r^N\|_2^2 \quad (3-9)$$

其中, n_r^V 表示第 r 类中 VIS 人脸图像的数量, n_r^N 分别表示第 r 类中 NIR 人脸图像的数量。 $m_r^V = \frac{1}{n_r^V} \sum_{i=1}^{n_r^V} \mathbf{1}\{l_i^V = r\} f(x_i^V; \theta, W^R)$ 表示第 r 类别中 VIS 人脸图像的均值向量, $m_r^N = \frac{1}{n_r^N} \sum_{i=1}^{n_r^N} \mathbf{1}\{l_i^N = r\} f(x_i^N; \theta, W^R)$ 表示第 r 类别中 NIR 人脸图像的均值向量。

损失函数中的对抗过程表达如下:

$$\theta, W^R = \arg \min_{\theta} \max_{W^R} L_{residual} \quad (3-10)$$

在最大化 $L_{residual}$ 时, 剩余全连接结构能够学习剩余信息。此外, 在最小化 $L_{residual}$ 时, 剩余信息将在主干网络中被消除。

属性解耦目标在于减小身份相关表示与其他属性表示之间的相关性。对此, NLPF 引入了一个解耦目标函数, 以进一步减小参数 $W_i^I \in \mathbb{R}^{d_0}$ 、 $W_i^M \in \mathbb{R}^{d_0}$ 、 $W_i^R \in \mathbb{R}^{d_0}$ 之间的相关性。这三个参数是映射矩阵 $W^I \in \mathbb{R}^{d_0 \times d_1}$ 、 $W^M \in \mathbb{R}^{d_0 \times d_1}$ 、 $W^R \in \mathbb{R}^{d_0 \times d_1}$ 的映射向量。解耦函数的目标如下所示:

$$\begin{aligned} \cos(W_i^I, W_i^M) &= \frac{W_i^{IT} W_i^M}{\|W_i^I\| \|W_i^M\|} = 0 \quad (i = 1, 2, \dots, d_1) \\ \cos(W_i^M, W_i^R) &= \frac{W_i^{MT} W_i^R}{\|W_i^M\| \|W_i^R\|} = 0 \quad (i = 1, 2, \dots, d_1) \\ \cos(W_i^I, W_i^R) &= \frac{W_i^{IT} W_i^R}{\|W_i^I\| \|W_i^R\|} = 0 \quad (i = 1, 2, \dots, d_1) \end{aligned} \quad (3-11)$$

其中 $\cos(\cdot, \cdot)$ 表示余弦相似度函数。因此, 解耦损失可以表示为:

$$L_{dec} = \|\cos(W_i^I, W_i^M)\|_F^2 + \|\cos(W_i^M, W_i^R)\|_F^2 + \|\cos(W_i^I, W_i^R)\|_F^2 \quad (i = 1, 2, \dots, d_1) \quad (3-12)$$

其中, $\|\cdot\|_F^2$ 表示 Frobenius 范数。理想情况下, L_{dec} 中的每一项都应趋近于 0, 这意味

着 W_i^I 、 W_i^M 和 W_i^R 应该彼此正交。

根据公式3-7、3-9、3-12，属性解耦的损失函数 L_{attr} 总结如下：

$$L_{attr} = L_{modality} + L_{residual} + L_{dec} \quad (3-13)$$

如图3-3所示，本章提出了一个渐进融合合成网络，以重新整合面部属性，即身份相关的特征、模态相关的特征、剩余特征。渐进融合合成网络是一个由四个子解码器结构组成的三支网络，包括 G^I 、 G^M 、 G^R 、 G^N 。FC 层表示全连接层。DeConv 层包含反卷积层、批归一化层和 ReLU 激活层。Conv 包括卷积层和批归一化层。具体而言， G^I 从身份相关的特征中上采样以重建 VIS 模态的中立人脸。同时，在 G^I 的中间层，可以获得身份相关的特征图 $M^I \in 80 \times 80 \times 64$ 。 G^M 和 G^R 从与模态相关的特征和剩余特征中上采样，以获得模态相关的特征图 $M^M \in 80 \times 80 \times 64$ 和剩余特征图 $M^R \in 80 \times 80 \times 64$ 。然后，这三个特征图被级联作为 G^N 的输入。 G^N 重新整合面部属性信息并重建异构人脸。

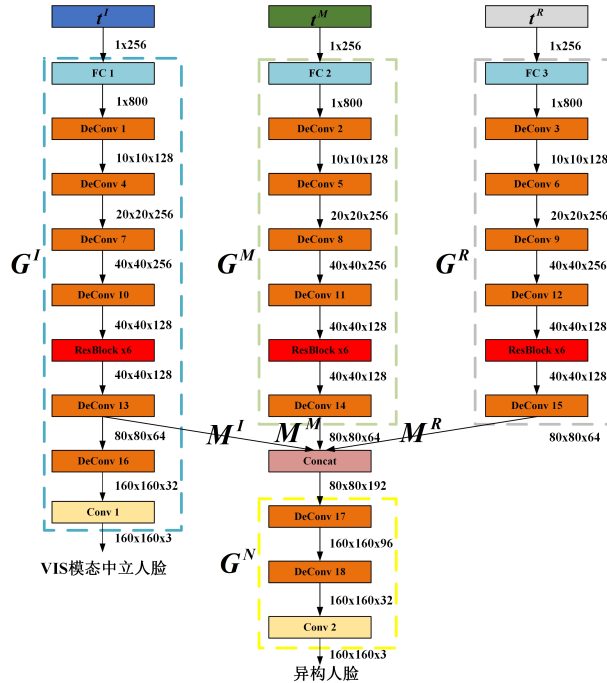


图 3-3 渐进融合合成网络的结构。渐进融合合成网络包括四个子解码器，将面部属性从特征级别重新整合到图像级别

异构人脸重建的目标函数可以表达为：

$$L_{rec}^{het} = \|\tilde{X}_V - X_V\|_1 + \|\tilde{X}_N - X_N\|_1 \quad (3-14)$$

其中， $\tilde{X}_V$ 代表 VIS 模态中重建的异构人脸， $\tilde{X}_N$ 代表 NIR 模态中重建的异构人脸。

在异构人脸重建过程中的对抗损失可以表达为：

$$L_{adv}^{het} = \mathbb{E}[\log(D(X))] + \mathbb{E}[\log(1 - D(G^N(\tilde{G}^I(t^I) + G^M(t^M) + G^R(t^R)))] \quad (3-15)$$

其中， $X \in \{X_N, X_V\}$ ，而 $G^N(\tilde{G}^I(t^I) + G^M(t^M) + G^R(t^R)) = \tilde{X}_N \setminus \tilde{X}_V$ 表示通过渐进融合合成网络将面部属性特征合成为异构人脸。 $\tilde{G}^I$ 表示部分的 G^I 的网络结构。在梯度反向传播过程中，损失函数 L_{rec}^{het} 、 L_{adv}^{het} 和 $L_{modality}$ 对主干网络参数不提供梯度，以免影响主干网络提取具有辨识性的身份相关表征的能力。通过这种方式，NLPF 不仅实现了异构人脸属性解耦的目的，还达到了净化身份相关表征的目的。

最后，在渐进融合合成网络中，目标损失可以表达为：

$$L^{APF} = L_{attr} + L_{rec}^{het} + L_{adv}^{het} \quad (3-16)$$

其中， L_{attr} 的目标是分解异构人脸的潜在属性，而 L_{rec}^{het} 和 L_{adv}^{het} 则尝试重新整合面部属性。上述过程促进了面部属性的解耦，并获得了只与身份相关的表征。

3.2.4 模型优化

提出的 NLPF 方法由两个关键模块组成，即 INL 和 APF。NLPF 中的最终目标函数可以表示为：

$$\mathcal{L} = \lambda_1 L^{INL} + \lambda_2 L^{APF} \quad (3-17)$$

其中， λ_1 和 λ_2 表示损失的权重。算法 3-1 展示了该模型的优化细节，NLPF 网络中的参数包括 θ 、 θ_D 、 θ_{G^I} 、 θ_{G^M} 、 θ_{G^R} 、 θ_{G^N} 、 W^I 、 W^M 、 W^R 、 W_{id}^I 以及 $W_{modality}^M$ ，网络参数

算法 3-1: 中立人脸学习和渐进融合合成 NLPF 方法

Input: 训练集 X_N and X_V , 迭代次数 $iter$

Output: 网络参数 θ, W^I

- 1: 使用 MS-Celeb-1M 数据库对网络进行预训练, 初始化 θ 和 W^I ;
 - 2: **for** $i = 1, 2, \dots, iter$ **do**
 - 3: 更新 θ, W^I, W_{id}^I 通过最小化 L_{id} ;
 - 4: 更新 $W^M, W_{modality}^M$ 通过最小化 $L_{modality}$;
 - 5: 更新 θ 通过最小化 $L_{residual}$;
 - 6: 更新 W^R 通过最小化 $L_{residual}$;
 - 7: 更新 W^I, W^M, W^R 通过最小化 L_{dec} ;
 - 8: 更新 θ, θ_{G^I} 通过最小化 L_{rec}^{neu} ;
 - 9: 更新 θ_{G^I}, θ_D 通过最小化 L_{adv}^{neu} ;
 - 10: 更新 $\theta_{G^I}, \theta_{G^M}, \theta_{G^R}, \theta_{G^N}$ 通过最小化 L_{rec}^{het} ;
 - 11: 更新 $\theta_{G^I}, \theta_{G^M}, \theta_{G^R}, \theta_{G^N}, \theta_D$ 通过最小化 L_{adv}^{het} ;
 - 12: **end for**
 - 13: **Return** θ and W^I
-

通过循环迭代方式进行更新。

3.3 实验结果与分析

在本节中,为了展示 NLPF 方法的性能,使用了三个主流的 NIR-VIS 数据库 CASIA NIR-VIS 2.0^[12]、BUAA NIR-VIS^[112]、LAMP-HQ NIR-VIS^[14] 来获得大量的实验结果。本节主要进行两个验证。首先,验证了 NLPF 模型中 INL 模块和 APF 模块的有效性。其次,展示了 NLPF 与 ADFL^[76]、CDL^[130]、WCNN^[93]、PCFH^[70]、PACH^[71]、EGCH^[82]、ADCANs^[131]、SAFE^[74]、MMTN^[72] 等最主流方法之间的比较结果,以验证 NLPF 具有竞争力的识别性能。

3.3.1 实验数据集

本方法采用的三个主流数据集简要概括如下:

CASIA NIR-VIS 2.0 dataset^[12] 包含 725 个身份,拥有超过 17,000 张近红外和可见光人脸图像。图像中的人脸表现出丰富的面部变化,包括遮挡、表情、姿态等,使其成为一个具有挑战性的数据集。具体而言,该数据集包含十个评估协议^[12],训练实例与测试实例之间没有交集。本章参考评估协议中的实验设置,以验证 NLPF 模型的有效性。

BUAA NIR-VIS dataset^[112] 是一个小规模的数据集,包括 150 个具有明显面部变化的身份个体。每个个体有 18 张图像(包括 9 张近红外图像和 9 张可见光图像)。本章参考文献^[93]中的验证设置,其中训练集(50 个身份,900 张图像)与测试集(100 个身份,1,800 张图像)之间没有交集。

LAMP-HQ NIR-VIS dataset^[14] 是一个于 2021 年发布的新型大规模 NIR-VIS 数据库。该数据库包含超过 70,000 张近红外-可见光图像,具有丰富的面部表情、不同的光照环境、复杂的背景干扰以及各种遮挡。本章参考文献^[14]中的工作,采用了第一个实验协议进行验证,并与主流的方法进行比较。

3.3.2 实验设置

NLPF 模型是在 TensorFlow 上使用 TITAN Xp GPU 进行编程实现的。如图3-4所示,受文献^[136]启发,构建了一个名为 SLnet 的轻量级网络作为主干网络,用于提取特征表示。具体而言,SLnet 包括一个 Stem 块、10 个 Inception-ResNet-A 块、一个 Reduction-A

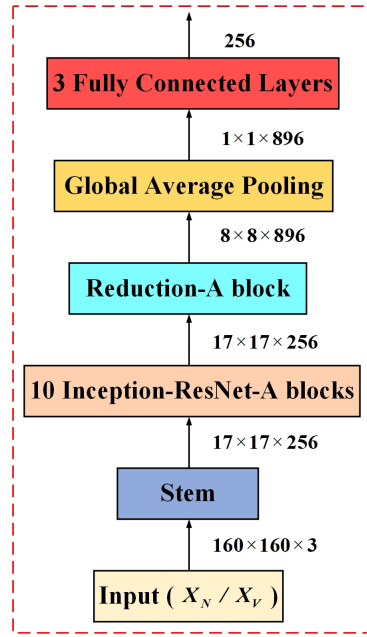


图 3-4 SLnet 网络结构

表 3-2 鉴别器的网络结构

名称	卷积核大小/步长/填充	输出尺度
图像输入	-	160x160x3
Conv	4x4 / 2 / 1	80x80x64
Conv	4x4 / 2 / 1	40x40x128
Conv	4x4 / 2 / 1	20x20x256
Conv	4x4 / 2 / 1	10x10x512
Conv	4x4 / 2 / 1	10x10x512
Conv	4x4 / 1 / 1	10x10x1

块、一个全局平均池化层和 3 个全连接层。属性特征 (包括 t^I, t^M, t^R) 设置为 256 维。NLPF 在 INL 和 APF 中都采用了基于多尺度的鉴别器^[78]，鉴别器的结构如表3-2所示。在渐进融合合成网络中的 ResBlock 结构^[78,125] 如表3-3所示，其中 RefPed 表示使用输入边界的反射对输入张量进行填充。

首先，在 MS-Celeb-1M^[109] 上对主干网络进行预训练，其中使用 MTCNN^[124] 将图像裁剪为 160×160 。在 LFW^[127] 上评估主干网络，取得了令人满意的 99.2% 准确率。在微调阶段，数据库中的近红外-可见光人脸图像通过 MTCNN^[124] 裁剪为 160×160 ，以进一步调整网络参数。训练的批量大小设置为 32。模型的超参数 λ_1 和 λ_2 都设置为 1。

表 3-3 ResBlock 的网络结构

名称	卷积核大小/步长/填充	输出尺度
RefPed	-	42x42x128
Conv	3x3 / 3 / 0	40x40x128
RefPed	-	42x42x128
Conv	3x3 / 2 / 0	40x40x128
Skip Connection	-	40x40x128

表 3-4 CASIA NIR-VIS 2.0 数据库上的消融实验

方法	Rank-1(%)	VR@FAR=1%(%)	VR@FAR=0.1%(%)
<i>BL</i>	91.6	96.6	84.4
<i>BL+INL</i>	99.2	99.7	98.5
<i>BL+APF</i>	96.6	98.4	89.0
<i>BL+INL+APF</i>	99.5	99.9	99.0

3.3.3 模块性能分析

为了展示 INL 模块和 APF 模块在 NLPF 模型中的贡献，在 NIR-VIS 数据集上对模型进行了深入的评估。

消融实验: 在表3-4和3-5中，展示了在 CASIA NIR-VIS 2.0 和 BUAA NIR-VIS 数据集上的消融实验结果。使用四个评估指标 (包括 Rank-1、VR@FAR=1%、VR@FAR=0.1%、VR@FAR=0.01%) 来衡量每个模块的贡献。基线方法 (*BL*) 表示在 MS-Celeb-1M 数据集上预训练的主干网络 SLnet。消融实验验证了四种情况，即 *BL*、*BL+INL*、*BL+APF*、*BL+INL+APF*。在 CASIA NIR-VIS 2.0 的评估中，参考第一个协议^[12] 中的实验设置。根据实验结果，可以得出以下分析结论。

首先，*BL* 方法未能在 CASIA NIR-VIS 2.0 上取得令人满意的识别结果，Rank-1 仅为 91.6%，而在 BUAA NIR-VIS 上 VR@FAR=0.01% 仅为 47.1%。以上结果表明，仅依赖可见光数据集对网络进行预训练并不适用于跨模态人脸识别任务。

表 3-5 BUAA NIR-VIS 数据库上的消融实验

方法	Rank-1(%)	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
<i>BL</i>	95.7	83.8	47.1
<i>BL+INL</i>	99.8	99.2	98.0
<i>BL+APF</i>	99.7	98.1	93.1
<i>BL+INL+APF</i>	100	99.8	99.0

相比 BL 方法, $BL + INL$ 方法在 CASIA NIR-VIS 2.0 数据集上, Rank-1 提高了 7.6%, $VR@FAR=1\%$ 提高了 3.1%, $VR@FAR=0.1\%$ 提高了 14.1%。此外, 在 BUAA NIR-VIS 数据集上, Rank-1 提高了 4.1%, $VR@FAR=0.1\%$ 提高了 15.4%, $VR@FAR=0.01\%$ 提高了 50.9%。可以得出结论, 对 NIR-VIS 数据集进行微调并学习中立人脸表征有助于提高识别性能。

相比 BL 方法, $BL + APF$ 方法在 CASIA NIR-VIS 2.0 数据集上, Rank-1 提高了 5.0%, $VR@FAR=1\%$ 提高了 1.8%, $VR@FAR=0.1\%$ 提高了 4.6%。此外, 在 BUAA NIR-VIS 数据集上, Rank-1 提高了 4.0%, $VR@FAR=0.1\%$ 提高了 14.3%, $VR@FAR=0.01\%$ 提高了 46.0%。从结果中可以观察到, APF 能够解耦异构数据中的潜在属性, 因此有助于提升模型性能。

最后, $BL+INL+APF$ 方法优于 $BL+INL$ 方法, Rank-1 提高了 0.3%, $VR@FAR=1\%$ 提高了 0.2%, $VR@FAR=0.1\%$ 提高了 0.5%。同样, $BL + INL + APF$ 方法在 BUAA NIR-VIS 数据集上也有明显的性能提升, $VR@FAR=0.1\%$ 提高了 0.6%, $VR@FAR=0.01\%$ 提高了 1.0%。

表 3-6 在 CASIA NIR-VIS 2.0 数据集上进行的 NLPF 消融研究的定量结果

方法	$VR@FAR=0.01\%$ (%)
BL	64.6
w/o L_{id}	86.7
w/o $L_{modality}$	96.8
w/o $L_{residual}$	96.3
w/o L_{dec}	96.7
w/o L_{rec}^{neu}	96.1
w/o L_{adv}^{neu}	96.2
w/o L_{attr}	95.6
w/o L_{rec}^{het}	95.6
w/o L_{adv}^{het}	96.8
NLPF (本章方法)	97.0

定性和定量分析: 为了更好地理解 NLPF 中损失函数对模型性能和图像合成的影响, 在表3-6和图3-5中展示了定性和定量的分析结果。首先, 由于 L_{id} 有助于网络学习具有区分性的表示, NLPF 在 $VR@FAR=0.01\%$ 上的性能比未使用 L_{id} 的情况更好, 提高了 10.3%。其次, $L_{modality}$ 有助于解耦与身份表示相关的模态特征。在 $VR@FAR=0.01\%$ 上, NLPF 的性能比未使用 $L_{modality}$ 的情况更好, 提高了 0.2%。接着, $L_{residual}$ 有助于异构人

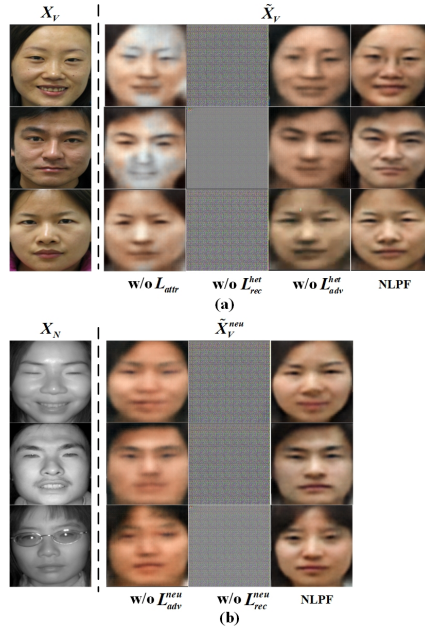


图 3-5 NLPF 在 CASIA NIR-VIS 2.0 数据集上进行的消融研究的定性结果。(a) 是 APF 的合成结果。第一列是从 CASIA NIR-VIS 2.0 中选择的 VIS 图像。最右侧列的图像是由去除 APF 中损失函数 L_{attr} 、 L_{rec}^{het} 、 L_{rec}^{adv} 的消融版本合成的。(b) 是 INL 的合成结果。第一列是从 CASIA NIR-VIS 2.0 中选择的 NIR 图像。最右侧列的图像是由去除 INL 中损失函数 L_{adv}^{neu} 和 L_{rec}^{neu} 的消融版本合成的结果

脸中的剩余特征解耦。在 $VR@FAR=0.01\%$ 上, NLPF 的性能比未使用 $L_{residual}$ 的情况更好, 提高了 0.7%。 L_{dec} 减少了高层次属性特征之间的相关性。因此, 在 $VR@FAR=0.01\%$ 上, NLPF 的性能比未使用 L_{dec} 的情况更好, 提高了 0.3%。 L_{rec}^{neu} 能够将异构人脸合成为 VIS 中立人脸并学习 VIS 中立表示。因此, 在 $VR@FAR=0.01\%$ 上, NLPF 的性能比未使用 L_{rec}^{neu} 的情况更好, 提高了 0.9%。 L_{adv}^{neu} 影响重建人脸中的高频信息。因此, NLPF 在 $VR@FAR=0.01\%$ 上的性能比未使用 L_{adv}^{neu} 的情况更好, 提高了 0.8%。此外, 由于缺乏身份特征和其他潜在属性的解耦, 不使用 L_{attr} 会导致合成的缺陷。例如, 生成的可见光人脸可能受到近红外模态的伪影和不一致的纹理的影响。在 $VR@FAR=0.01\%$ 上, NLPF 的性能比未使用 L_{attr} 的情况, 提高了 1.4%。若不使用 L_{rec}^{het} , NLPF 无法在像素级别上使用异构人脸作为监督信号, 因此无法从特征级别到图像级别重构异构人脸。在 $VR@FAR=0.01\%$ 上, NLPF 的性能比不使用 L_{rec}^{het} 的情况更好, 提高了 1.4%。最后, L_{adv}^{het} 可以增强图片中的细节, 这些细节这有助于人脸身份确定。这导致 NLPF 在 $VR@FAR=0.01\%$ 上的性能比不使用 L_{adv}^{het} 的情况更好, 提高了 0.2%。以上结果验证了 NLPF 中所有提出的损失函数对模型性能的益处。联合方案学习中立人脸表示, 并进一

步解耦潜在属性以获得具有区分性的表示。

人脸图像合成分析：为了展示 INL 中的合成效果，图3-6展示了在 CASIA NIR-VIS 2.0 上生成的同模态中立人脸的结果。可以明显看到，INL 学习到了中立人脸信息，并有效消除了由遮挡、姿态、表情、光照等引起的剩余信息。而 NIR-VIS 非中立人脸被统一转换为 VIS 中立人脸，这可以消除异构非中立人脸的模态差异和剩余差异。

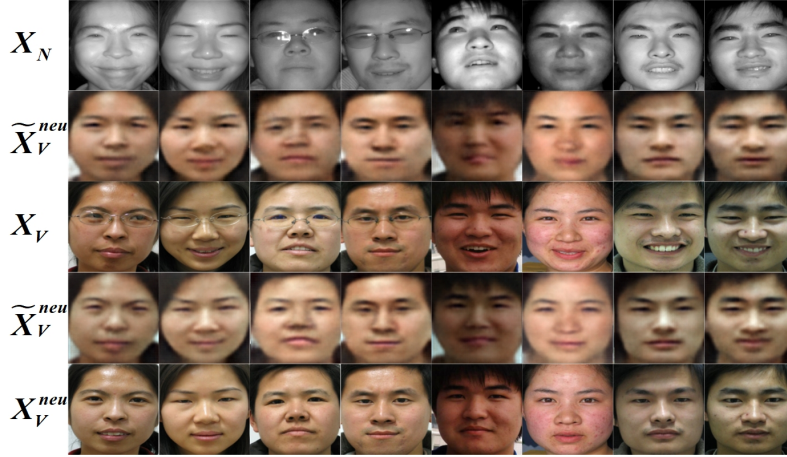


图 3-6 INL 模块的生成结果。异构人脸图像来自 CASIA NIR-VIS 2.0 数据集。从上到下：输入的 NIR 人脸、以 NIR 图片为输入的输出 VIS 中立人脸、输入的 VIS 人脸、以 VIS 人脸为输入的输出 VIS 中性人脸，以及原始的 VIS 中性人脸。

可视化分析：图3-7中展示了探针样本和查询样本之间的近红外-可见光人脸匹配结果。余弦相似度显示在图像底部，用于衡量特征之间的距离。为了展示 NLPF 模型消除剩余信息的能力，展示了六种情况的结果，即光照、姿态、遮挡、遮挡 & 表情、姿态 & 表情、姿态 & 遮挡。结果表明 Ground Truth (GT) 的置信度明显优于其他人员，这体现在 GT 的置信度通常比第二名高出 0.2。同时，在近红外探测和可见光库之间存在明显的面部差异时，NLPF 模型仍然具有良好的识别性能。

INL 与 Arcface 的对比分析：为验证 INL 的有效性，在 CASIA NIR-VIS 2.0 数据集上比较了 INL 和 Arcface^[31] 的结果，如表 3-7 所示。在 Rank-1 方面，INL+APF 的性能优于 Arcface+APF 0.4%，在 VR@FAR=0.1% 方面优于 Arcface+APF 0.5%，在 VR@FAR=0.01% 方面优于 Arcface+APF 2.2%。

表 3-7 在 CASIA NIR-VIS 2.0 数据集上对比 INL 损失和 Arcface 损失^[31]

方法	Rank-1(%)	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
Arcface+APF	99.1	98.5	94.8
INL+APF	99.5	99.0	97.0

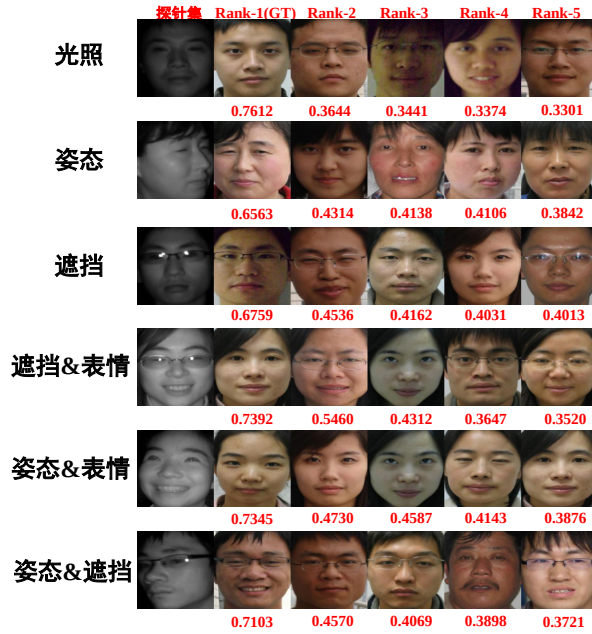


图 3-7 六个情况下进行 NIR-VIS 人脸匹配的可视化检索结果，即光照、姿态、遮挡、遮挡 & 表情、姿态 & 表情、姿态 & 遮挡。图片下的数字为余弦相似度

$L_{residual}$ 与 center loss 的对比分析：为验证 $L_{residual}$ 的有效性，将 $L_{residual}$ 替换为 center loss^[137]，以在 CASIA NIR-VIS 2.0 数据集上展示比较结果。结果显示在表3-8中。使用 $L_{residual}$ 的 NLPF 在 Rank-1 方面优于使用 center loss 的 NLPF 0.2%，在 VR@FAR=0.1% 方面优于 center loss 的 NLPF 0.4%，在 VR@FAR=0.01% 方面优于 center loss 的 NLPF 0.9%。结果表明， $L_{residual}$ 能够更有效地提取剩余特征，有助于属性解耦。

表 3-8 在 CASIA NIR-VIS 2.0 数据库上对比 $L_{residual}$ 和 center loss^[137]

方法	Rank-1(%)	VR@FAR=0.1%(%)	VR@FAR=0.01%(%)
NLPF (center loss)	99.3	98.6	96.1
NLPF ($L_{residual}$)	99.5	99.0	97.0

3.3.4 与其他主流的方法对比

将 NLPF 模型与主流的跨模态人脸识别方法进行比较，如 ADFL^[76]、CDL^[130]、WCNN^[93]、PCFH^[70]、PACH^[71]、EGCH^[82]、ADCANs^[131]、SAFE^[74]、MMTN^[72]，以验证 NLPF 在识别性能上的优势。

CASIA NIR-VIS 2.0 实验结果：在 CASIA NIR-VIS 2.0 数据集中，如表3-9所示，将 NLPF 与其他方法进行了比较。根据结果，可以得出以下结论。NLPF 模型优于传统的深

度跨模态人脸识别方法 (即 Light CNN-29、TRIVET)。具体而言, NLPF 方法的性能超过了 Light CNN-29, Light CNN-29 是深度传统跨模态人脸识别方法中最具代表性的方法之一, 在 Rank-1 准确度上提高了 3%, 在 $VR@FAR=0.1\%$ 上提高了 4.7%。其次, 与最近的跨模态人脸识别方法相比, 包括 IDR、ADFL、CDL、WCNN、PCFH、PACH、EGCH、ADCANs、SAFE、MMTN, NLPF 实现了具有 99.7% 的 Rank-1 和 99.9% 的 $VR@FAR=1\%$ 以及 99.5% 的 $VR@FAR=0.1\%$ 的有竞争力的测试性能。以上结果表明了 NLPF 方法能够有效消除模态差异以及剩余信息, 并增强与身份相关的特征学习。

表 3-9 不同方法在 CASIA NIR-VIS 2.0 数据库上的结果

Method	Rank-1 (%)	$VR@FAR = 1\%$ (%)	$VR@FAR = 0.1\%$ (%)
TRIVET ^[92] (ICB 2016)	95.7 $\pm$ 0.5	98.1 $\pm$ 0.3	91.0 $\pm$ 1.3
IDR ^[128] (AAAI 2017)	97.3 $\pm$ 0.4	98.9 $\pm$ 0.3	95.7 $\pm$ 0.7
Light CNN-29 ^[129] (TIFS 2018)	96.7 $\pm$ 0.2	-	94.8 $\pm$ 0.4
ADFL ^[76] (AAAI 2018)	98.1 $\pm$ 0.4	99.1 $\pm$ 0.2	97.1 $\pm$ 0.5
CDL ^[130] (AAAI 2018)	98.6 $\pm$ 0.2	-	98.3 $\pm$ 0.5
WCNN ^[93] (TPAMI 2018)	98.7 $\pm$ 0.3	99.5 $\pm$ 0.1	98.4 $\pm$ 0.4
PCFH ^[70] (IJCAI 2019)	98.8 $\pm$ 0.3	99.6 $\pm$ 0.1	97.7 $\pm$ 0.3
PACH ^[71] (CVPR 2020)	98.9 $\pm$ 0.2	99.6 $\pm$ 0.1	98.3 $\pm$ 0.2
CFC ^[27] (TPAMI 2020)	98.6 $\pm$ 0.1	99.2 $\pm$ 0.1	97.3 $\pm$ 0.2
SDDL ^[106] (PR 2021)	92.1 $\pm$ 0.7	-	88.7 $\pm$ 1.0
ADCANs ^[131] (TMM 2021)	99.1 $\pm$ 0.2	99.6 $\pm$ 0.1	98.5 $\pm$ 0.2
VSANet ^[14] (IJCV 2021)	99.2 $\pm$ 0.2	99.6 $\pm$ 0.1	98.3 $\pm$ 0.2
EGCH ^[82] (ICPR 2021)	98.8 $\pm$ 0.2	99.6 $\pm$ 0.0	97.7 $\pm$ 0.2
MMTN ^[72] (TIFS 2022)	99.4 $\pm$ 0.0	99.8 $\pm$ 0.0	98.7 $\pm$ 0.0
SAFE ^[74] (TPAMI 2022)	98.9 $\pm$ 0.1	99.6 $\pm$ 0.1	98.6 $\pm$ 0.1
NLPF (本章方法)	99.7 $\pm$ 0.2	99.9 $\pm$ 0.0	99.5 $\pm$ 0.3

BUAA NIR-VIS 实验结果: 将 NLPF 方法与各种跨模态人脸识别模型进行比较, 包括 TRIVET^[92]、IDR^[128]、LightCNN-29^[129]、ADFL^[76]、CDL^[130]、WCNN^[93]、PCFH^[70]、PACH^[71]、CFC^[27]、EGCH^[82]、ADCANs^[131]、SAFE^[74]、MMTN^[72]。在 BUAA NIR-VIS 数据库中, 结果如表3-10所示。可以观察到 NLPF 模型在 Rank-1 为 100%, $VR@FAR=1\%$ 为 100% 以及 $VR@FAR=0.1\%$ 为 99.8%。它明显比所展示的深度跨模态人脸识别方法更好。

LAMP-HQ NIR-VIS 实验结果: 由于 LAMP-HQ 是一个新的、最具挑战性的 NIR-VIS 数据集, 大多数的方法没有在该数据集上报告性能。使用 LAMP-HQ NIR-VIS 第一个协议验证了 NLPF 模型的实验结果与一些主流深度学习方法进行比较, 例如 LightCNN-29^[129]、ADFL^[76]、PCFH^[70]、PACH^[71]、ADCANs^[131]、VSANet^[14]、MMTN^[72]。结果如表3-11所示。特别是, NLPF 方法在 Rank-1 上优于 ADCANs 方法, 其在 Rank-1 上提高

表 3-10 不同方法在 BUAA NIR-VIS 数据库上的结果

Method	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
TRIVET ^[92] (ICB 2016)	93.9	93.0	80.9
IDR ^[128] (AAAI 2017)	94.3	93.4	84.7
Light CNN-29 ^[129] (TIFS 2018)	96.5	95.4	86.7
ADFL ^[76] (AAAI 2018)	95.2	95.3	88.0
CDL ^[130] (AAAI 2018)	96.9	95.9	90.1
WCNN ^[93] (TPAMI 2018)	97.4	96.0	91.9
PCFH ^[70] (IJCAI 2019)	98.4	97.9	92.4
PACH ^[71] (CVPR 2020)	98.6	98.0	93.5
CFC ^[27] (TPAMI 2020)	99.7	98.7	97.8
ADCANs ^[131] (TMM 2021)	99.8	99.7	98.4
VSANet ^[14] (IJCV 2021)	98.8	98.3	93.4
EGCH ^[82] (ICPR 2021)	98.1	92.7	
MMTN ^[72] (TIFS 2022)	99.2	99.5	97.2
SAFE ^[74] (TPAMI 2022)	99.4	99.0	98.0
NLPF (本章方法)	100	100	99.8

了 1.0%，在 VR@FAR=0.1% 上提高了 2.2%。因此，NLPF 在复杂环境中表现出稳健且有效的 NIR-VIS 人脸匹配性能。

表 3-11 不同方法在 LAMP-HQ NIR-VIS 数据库上的结果

Method	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
Light CNN-29 ^[129] (TIFS 2018)	94.6 ± 0.3	92.5 ± 0.6	75.6 ± 1.9
ADFL ^[76] (AAAI 2018)	95.1 ± 0.5	92.1 ± 0.9	73.3 ± 2.2
PCFH ^[70] (IJCAI 2019)	95.3 ± 0.5	92.9 ± 0.6	75.1 ± 1.8
PACH ^[71] (CVPR 2020)	95.4 ± 0.5	93.1 ± 0.4	75.3 ± 1.7
ADCANs ^[131] (TMM 2021)	97.8 ± 0.3	98.8 ± 0.3	96.1 ± 0.4
VSANet ^[14] (IJCV 2021)	97.3 ± 0.2	94.8 ± 0.7	78.2 ± 3.0
MMTN ^[72] (TIFS 2022)	97.7 ± 0.0	95.7 ± 0.6	79.2 ± 0.0
NLPF (本章方法)	98.8 ± 0.3	99.2 ± 0.3	98.3 ± 0.4

3.4 本章小结

在本章中，提出了一种新颖的中立人脸学习和渐进融合合成 NLPF 方法。NLPF 方法由两个关键组件组成，即与身份相关中立人脸学习 INL 和属性渐进融合 APF。首先，INL 通过引导网络学习同模态的中立人脸特征表示，抑制了模态和剩余差异。具体而言，NLPF 通过将跨模态图像映射到共同的中立表示子空间来解决模态差异和剩余差异的挑战。此外，APF 以渐进融合的方式进行与身份相关的特征、模态相关的特征、剩余特征的解耦和融合，进一步净化身份相关的特征。在三个具有挑战性的 NIR-VIS 数据库上的大量实验验证了 NLPF 方法的有效性和鲁棒性。

第 4 章 鲁棒跨模态伪标签和对比学习方法

在上两章中，针对模态差异和剩余差异，本文提出了有效的 SSL 和 NLPF 方法。然而，这些方法严重依赖精心标注的训练数据，为大量异构人脸图像注释身份标签是耗时且昂贵的，这限制了 NIR-VIS 人脸识别系统在更大规模的实际场景中进行应用。为了减少对手动注释的依赖，本章尝试以无监督的方式实现 NIR-VIS 人脸识别，提出了一种新颖的鲁棒跨模态伪标签和对比学习 (Robust Cross-domain Pseudo-labeling and Contrastive Learning, RPC) 网络。该网络具有三个主要模块，即基于 NIR 集群的伪标签共享 (NIR cluster-based Pseudo labels Sharing, NPS)、模态域特定集群对比学习 (Domain-specific cluster Contrastive Learning, DCL)、跨模态域集群对比学习 (Inter-domain cluster Contrastive Learning, ICL)。首先，NPS 用于获得准确的 NIR 集群以生成伪标签，并与 VIS 模态共享可靠的标签知识。其次，DCL 学习模态域内紧凑而有辨别力的表示。最后，ICL 动态地优化跨模态身份关联，引导实例特征学习鲁棒且与模态无关的表示。对本方法进行了大量实验，实验结果表明其在伪标签分配方面达到了 99% 以上的准确性，同时在四个主流 NIR-VIS 数据集上的获得的优越性能。这一研究成果已在 IEEE Transactions on Image Processing 正式发表。

4.1 引言

NIR-VIS 人脸识别的目的在于准确匹配同一身份的 NIR 人脸图像与 VIS 人脸图像^[128,129]。作为异构人脸识别的代表性子任务，NIR-VIS 人脸识别是解决低光条件下人脸识别瓶颈的有效方法。因此，许多涉及公共安全、犯罪调查、多媒体信息检索等领域的应用和研究都围绕着 NIR-VIS 人脸匹配展开^[75,97,138]。

近年来，针对 NIR-VIS 人脸匹配中存在的模态差异，学术界中提出了基于模态不变描述子的方法、基于合成的方法、基于子空间学习的方法。当前的异构人脸识别方法取得了显著的成功，前提是准确的身份标签可用于指导网络学习识别能力。然而，对来自不同传感器的大量异构人脸图像进行身份标签注释是耗时且昂贵的。这明显限制

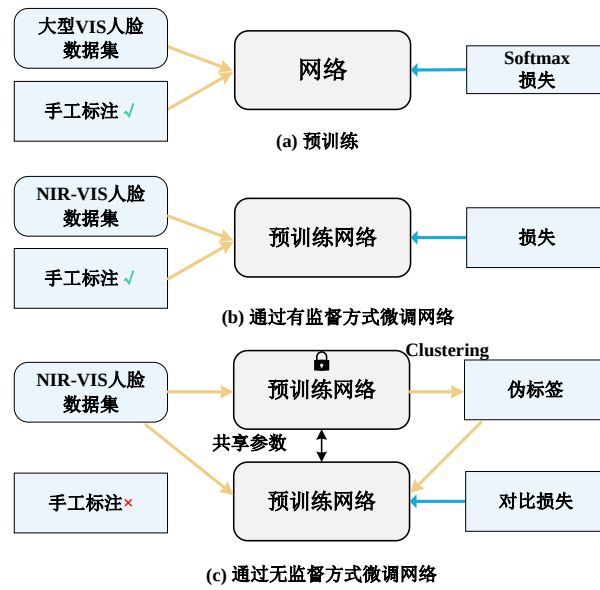


图 4-1 有监督方法与本子方法的区别示意图: (a) 表示模型的预训练, (b) 表示有监督方法的训练过程, 而 (c) 则表示无监督 RPC 的训练过程

了 NIR-VIS 人脸识别系统在更大规模实际场景中的应用。当前, 无监督学习方法已经在行人重识别任务中得到了成功应用^[139-141], 可以减轻了对昂贵数据注释的需求。但是在跨模态人脸识别任务中, 学术界对无监督学习方面关注甚少。本章尝试实现无监督的 NIR-VIS 异构人脸识别 (Unsupervised NIR-VIS Heterogeneous Face Recognition (NIR-VIS-UHFR)) 以突破获取标签注释困难的问题。具体而言, 有监督方法和本章方法的学习过程如图4-1所示。参考大多数跨模态人脸识别方法的常规做法^[72,93,108], 在大规模 VIS 人脸数据集上预训练网络。然后, 使用对比损失和算法估计的伪标签而非手工标签对网络进行微调。其中, 伪标签是通过将 NIR-VIS 人脸数据特征进行聚类而获得的。

无监督域自适应 (Unsupervised Domain Adaptation, UDA) 是通过对一个在有标签源域上训练的模型进行泛化, 以适应无标签的目标域。最近, 为了解决行人重新识别的 UDA 问题, 学者们提出了许多方法, 具体而言可分为两类: 领域转换方法^[139-143] 和基于聚类的方法^[144-146]。其中, 领域转换方法旨在通过迁移学习的方式最小化源域图像与目标域图像之间的领域差异。一些优秀的工作, 如 SPGAN^[139], SBSGAN^[140], FD-GAN^[141], CR-GAN^[142], IPES-GAN^[143], 从目标域传输低级特征 (包括背景、相机风格、光照等) 到源域。对于高级特征 (包括性别、服装颜色、发型等), 一些方法^[147-149] 假设这些特征模式在领域之间是共享的。一些研究采用聚类算法, 如 DBSCAN^[150] 或 K-means^[151], 将特征分组为簇。然后, 根据簇标识为每个图像预测身份, 并使用伪标签以类似有监督的

方式微调网络^[144-146]。然而，由于以下因素，NIR-VIS-UHFR 任务与 UDA 行人重识别任务存在较大差异：1) UDA 是一个同质的身份识别问题，而 NIR-VIS-UHFR 是一个异质的身份识别问题。在模型训练过程中需要消除模态差异，获取模态不变的判别特征。2) NIR-VIS-UHFR 任务的主干网络在 VIS 模态上进行了预训练，但在 NIR 和 VIS 模态上进行微调。3) 在 UDA 行人重识别任务中，聚类算法被应用于同质的训练数据。但在 NIR-VIS-UHFR 任务中，跨模态特征分布中存在巨大的模态差异。直接在异构数据上应用聚类算法会导致出现大量错误的聚类结果 4) 大多数 UDA 方法的评估方式都在封闭集设置下进行，而 NIR-VIS-UHFR 的评估方式则在更为困难的开放集设置下进行。因此，UDA 不能简单地被 NIR-VIS-UHFR 任务直接借鉴使用。

基于上述分析，NIR-VIS-UHFR 任务中的挑战可以总结为两个方面：1) **伪标签分配**。由于存在巨大的模态差异，建立 NIR 和 VIS 模态之间的关联是为 NIR-VIS 人脸图像分配伪标签的一大挑战。2) **模型优化**。在训练初期，由于模型性能不佳容易分配不准确的噪声标签。如何减轻噪声标签对模型优化的影响，并消除模态差异、逐渐提升模型性能以及伪标签分配的准确性是无监督学习的核心问题。针对这些问题，本章提出了一种新颖用于 NIR-VIS-UHFR 的鲁棒跨模态伪标签和对比学习 (Robust Cross-domain Pseudo-labeling and Contrastive Learning, RPC) 方法。如图4-2所示，RPC 将基于 NIR 集群的伪标签共享 (NIR cluster-based Pseudo labels Sharing, NPS)、模态域特定集群对比学习 (Domain-specific cluster Contrastive Learning, DCL)、跨模态域集群对比学习 (Inter-domain cluster Contrastive Learning, ICL) 结合到一个端到端的框架中。为了解决当前聚类方法仅适用于同质数据而无法为异质数据分配准确伪标签的问题，NPS 促进了模态间的标签共享，建立了可靠的跨模态关联。DCL 通过在模态特定集群层面进行对比学习和动态内存更新来学习模态内紧凑而有辨别力的表示。最后，ICL 动态地优化跨模态的关联关系，引导实例特征学习到强大且与模态无关的表示。通过 NPS、DCL、ICL，所提出的无监督方案可以有效地生成高质量的伪标签，并为 NIR-VIS-UHFR 任务学习到模态无关的表示。

总的来说，本章的贡献总结如下：

- 提出了一个无监督联合学习框架，它能够生成高质量的伪标签并达到与有监督方法的性能相媲美。这体现在伪标签分配的准确率超过 99% 以及在多个 NIR-VIS 数

数据集上获得先进性能。

- 基于 NIR 集群的伪标签共享 NPS 在 NIR 模态下获得可靠的集群，并与 VIS 模态共享可靠的标签信息，建立跨模态身份特征的内在关联。
- 模态域特定集群对比学习 DCL 通过模态内进行对比学习和动态内存更新，学习模态内紧凑而有辨别力的表示。
- 跨模态域集群对比学习 ICL 通过模态不变记忆库的对比学习，引导实例特征学习模态无关的表示。

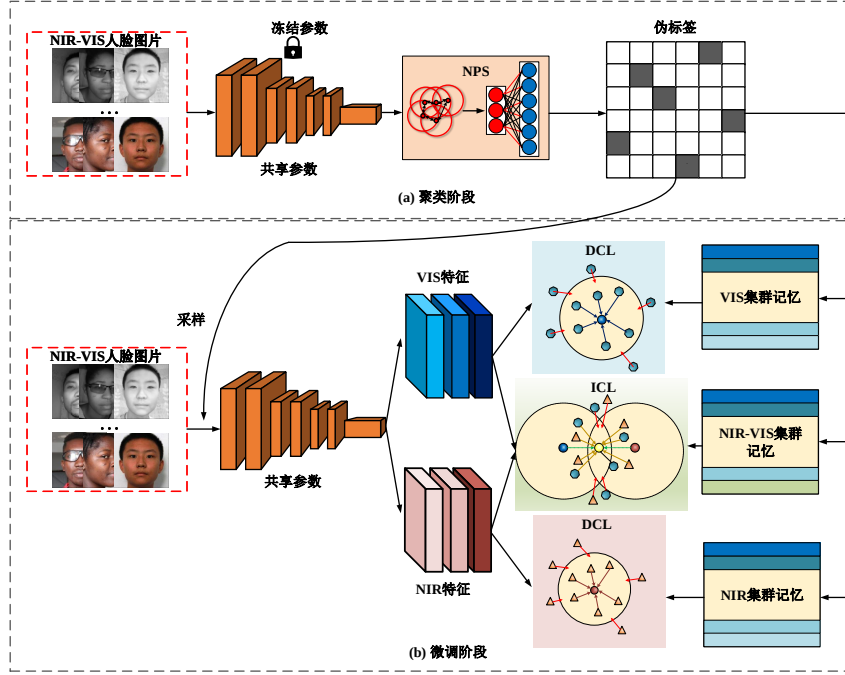


图 4-2 RPC 模型的流程图。RPC 网络包括三个关键模块，包括基于近红外集群的伪标签共享 NPS、模态特定集群对比学习 DCL、跨模态集群对比学习 ICL。该联合方案能够估计可靠的伪标签并学习模态无关的表示

4.2 鲁棒跨模态伪标签和对比学习模型

在本节中，提出鲁棒跨模态伪标签和对比学习模型，在更具挑战性的无监督设置中实现异质人脸识别。图4-2展示了 RPC 模型的框架。RPC 模型的目标是生成高质量的伪标签并学习模态无关的紧凑表示。为了实现这一目标，首先通过基于 NIR 集群的伪标签共享 NPS 来估计准确的伪标签，指导网络学习模态不变的表示。其次，提出模态域特定集群对比学习 DCL 来增强模态内的特征区分度。最后，设计了跨模态域集群对比学习 ICL 来引导网络修正错误的跨模态关联以及学习模态无关的表示。

4.2.1 问题建模

在 NIR-VIS-UHFR 任务的设置中，网络在 MS-Celeb-1M^[109] 数据集上进行预训练，然后在无标签的 VIS 训练数据 D_V 和无标签的 NIR 训练数据 D_N 上进行微调。无标签的 VIS 训练数据包括 n^V 个样本，可以表示为 $D_V = \{x_i^V\}_{i=1}^{n^V}$ 。无标签的 NIR 训练数据包括 n^N 个样本，可以表示为 $D_N = \{x_i^N\}_{i=1}^{n^N}$ 。在这项工作中，参考“聚类后微调”的

机制，充分利用跨模态数据的内在关系，生成可靠的伪标签 $\{\hat{y}_i^N, \hat{y}_i^V\}$ 并以类似有监督的方式微调网络。

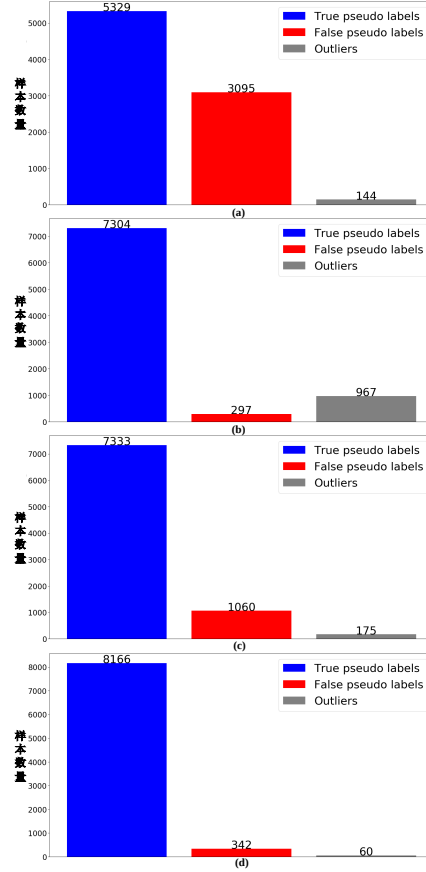


图 4-3 不同使用 DBSCAN 策略在 CASIA NIR-VIS 2.0 数据集上生成伪标签的分析。(a) 直接在整个数据集上使用 DBSCAN，(b) 分别在 NIR 数据和 VIS 数据上使用 DBSCAN，通过相似性计算合并 NIR 和 VIS 聚类，(c) 基于 VIS 集群的伪标签共享，(d) 基于 NIR 集群的伪标签共享，即在 RPC 网络中的伪标签估计方法

4.2.2 基于 NIR 集群的伪标签共享

为了获得伪标签并微调网络，使用聚类算法 DBSCAN^[150] 来估计伪标签。如图4-3 (a) 所示，直接在整个跨模态数据集上使用 DBSCAN 不能生成准确的伪标签，这体现在 3,095 个错误的伪标签的结果。如图4-3 (b) 所示，在每个模态内分别对数据进行聚类。然后，基于模态之间的集群相似性进行一对一的关联。如果最相似的集群已经关联，则选择下一个最相似的集群。然而，由于 NIR 和 VIS 集群的数量不相同，未关联的集群被指定为离群点 (Outliers)，而离群点没有对应的分类标签。大多离群点是难样本，简单丢弃这些离群点可能不利于模型的泛化能力提升。

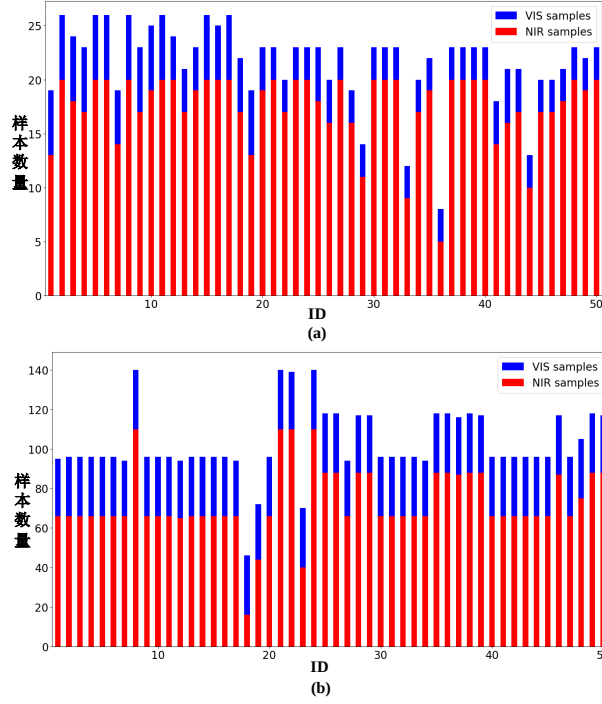


图 4-4 在 CASIA NIR-VIS 2.0 和 LAMP-HQ 数据集中选择的 50 个 ID 中，NIR 样本和 VIS 样本的分布情况。(a) 是 CASIA NIR-VIS 2.0 数据集的分布，(b) 是 LAMP-HQ NIR-VIS 数据集的分布

为了解决上述问题，一个直观的想法是仅对一个模态内的样本进行聚类。然后共享标签知识给另一个模态的样本。从图4-4 可以看出，在大规模 NIR-VIS 数据库中 (即 CASIA NIR-VIS 2.0^[12] 和 LAMP-HQ NIR-VIS^[14])，在选定的 50 个 ID 中，NIR 样本的数量通常是 VIS 样本数量的三倍。事实上，这符合实际情况，因为每个个体通常只在查询集中有少数已注册的 VIS 图像，而 NIR 摄像头会在开放场景中捕获到多个连续帧的 NIR 图像。因此，数量更多的 NIR 模态图像适合进行聚类，然后与 VIS 样本共享标签知识。

在上述分析的启发下，本文提出了基于 NIR 集群的伪标签共享 NPS 来对跨模态数据进行聚类。图4-5展示了 NPS 的伪标签估计过程。首先使用 DBSCAN 聚类算法为 NIR 数据 D_N 估计伪标签。具体而言，每个 NIR 数据的伪标签采用以下形式：

$$\hat{y}^N = \mathbf{DBSCAN}(F(x^N; \Theta)) \quad (4-1)$$

其中， $F(\cdot; \Theta)$ 表示由参数 Θ 学习到的网络映射函数， $\mathbf{DBSCAN}(\cdot)$ 表示的是 DBSCAN 聚类操作。 $\hat{y}^N \in \{1, 2, \dots, I^N\}$ 是 NIR 数据的伪标签，其中 I^N 表示 NIR 估计集群的数

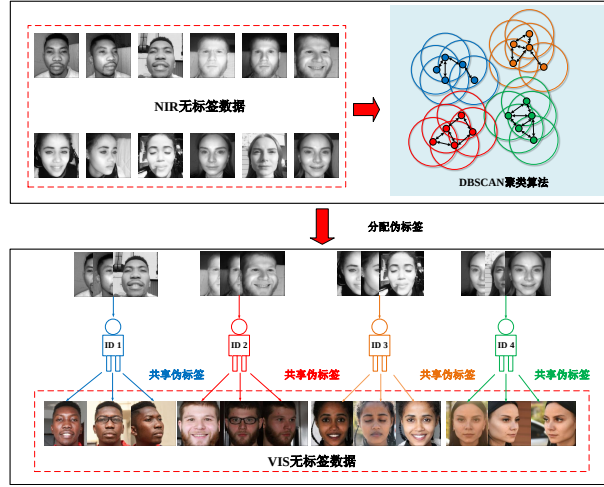


图 4-5 在 NPS 中伪标签估计的图示。首先使用 DBSCAN 聚类算法获取 NIR 数据的准确聚类中心，然后以跨模态检索的方式将可靠的标签知识与 VIS 模态共享

量。

通过计算每个 NIR 集群的特征均值来获得 NIR 聚类中心，具体表达如下：

$$c_k^N = \frac{1}{n_k^N} \sum_j \mathbf{1}\{\hat{y}_j^N = k\} f_j^N \quad (k = 1, 2, \dots, I^N) \quad (4-2)$$

其中， c_k^N 表示第 k 个集群的聚类中心，而 n_k^N 表示第 k 个聚类中的实例数。 $\mathbf{1}\{\cdot\}$ 是指示函数， $f_j^N = F(x_j^N; \Theta)$ 是 NIR 实例特征。

在 NIR 模态中获得聚类中心后，将 NIR 模态中的标签知识与未标记的 VIS 模态共享，以关联两个模态之间的身份信息。关联过程以一种跨模态检索的方式进行。具体而言，计算 f_j^V 与 NIR 聚类中心 c_k^N 之间的余弦相似度：

$$S_{j,k} = \{sim(f_j^V, c_k^N) |_{k=1}^{I^N}\} \quad (j = 1, 2, \dots, n^V) \quad (4-3)$$

其中， $f_j^V = F(x_j^V; \Theta)$ 表示 VIS 实例特征， $S_{j,k}$ 是 f_j^V 与所有 NIR 聚类中心 c_k^N 之间相似度的集合。

高度相似的 NIR 聚类中心可能包含与 VIS 实例相同的身份信息。根据 $S_{j,k}$ 中的相

似性排序结果，选择相似性分数最高的 NIR 聚类的标签与每个 VIS 样本共享：

$$k^* = \arg \max_k S_{j,k} \quad (4-4)$$

$$\hat{Y}_{k^*}^N \rightarrow \hat{y}_j^V \quad (j = 1, 2, \dots, n^V)$$

其中，变量 k^* 是当 $S_{j,k}$ 达到最大值时的 k 的索引， $\hat{Y}_{k^*}^N$ 表示第 k^* 个 NIR 集群的唯一伪标签。重复执行公式4-3和4-4，直到所有 VIS 数据都从具有最高相似度的 NIR 聚类中共享伪标签。在此之后，NIR 数据 $\hat{D}_N = \{x_i^N, \hat{y}_i^N\}_{i=1}^{n^N}$ 和 VIS 数据 $\hat{D}_V = \{x_i^V, \hat{y}_i^V\}_{i=1}^{n^V}$ 共享一个共同的身份标签空间。图 4-3 (c) 显示了基于 VIS 集群伪标签共享方法的聚类结果，图 4-3 (d) 展示了 NPS 的聚类结果。与基于 VIS 集群伪标签共享方法相比，NPS 获得了更准确的伪标签，有助于模型学习模态不变的表示。在下一小节中，这些伪标签被用作监督信号，通过对网络进行类似于有监督学习的微调过程，逐步提高伪标签分配的准确性和模型的识别性能。

4.2.3 模态域特定集群对比学习

近期的研究^[144,152-155]表明，基于记忆的学习方法能够提高特征的判别能力，同时避免嘈杂标签在后续训练步骤中的传播效应。文献^[144]指出了集群大小不平衡分布导致的偏置更新问题。如图 4-6 所示，大规模的 NIR-VIS 数据集存在集群大小不平衡的问题。采用基于实例的记忆学习方法可能加剧偏置更新，因为每次迭代的批处理大小是有限的，对于较小的集群，更新速度比较大集群要快。因此，在本小节中，采用基于集群记忆的方法来学习模态特定的紧凑表示。为了适应跨模态识别任务，本方法构建了三个记忆 (Memories)，包括两个模态特定的记忆和一个模态共享的记忆，有助于消除模态差异。首先，在每个训练循环的开始阶段，构建 NIR 聚类记忆 $\mathcal{M}^N$ 和 VIS 聚类记忆 $\mathcal{M}^V$ 来存储模态特定的集群级特征。NIR 聚类记忆 $\mathcal{M}^N$ 有 I^N 个槽位，VIS 聚类记忆 $\mathcal{M}^V$ 有 I^V 个槽位。NIR 聚类级特征的计算参考公式 4-2，而 VIS 聚类级特征的计算形式如下：

$$c_k^V = \frac{1}{n_k^V} \sum_j \mathbf{1}\{\hat{y}_j^V = k\} f_j^V \quad (k = 1, 2, \dots, I^V) \quad (4-5)$$

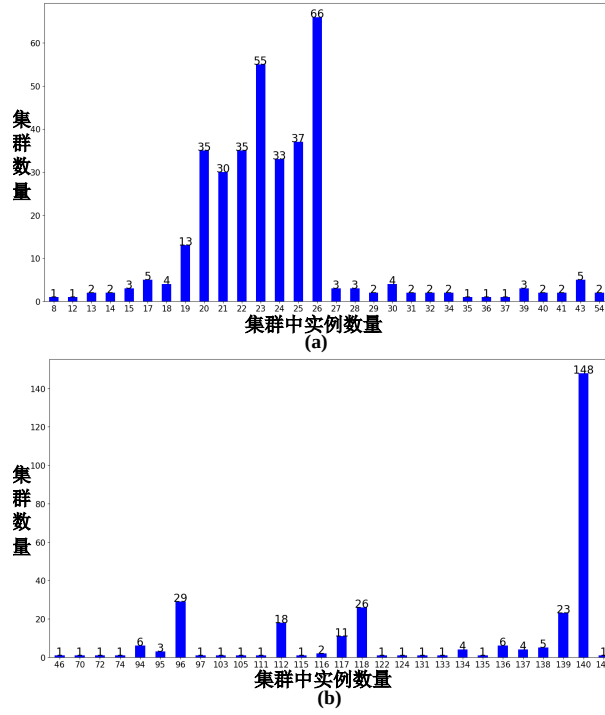


图 4-6 NIR-VIS 数据集中集群大小不平衡的分布情况。(a) 是 CASIA NIR-VIS 2.0 中集群大小的分布, (b) 是 LAMP-HQ NIR-VIS 中集群大小的分布

其中, c_k^V 表示第 k 个集群的 VIS 集群中心, n_k^V 表示第 k 个集群中的实例数量。然后, 将集群中心存储到记忆中:

$$\begin{aligned}\mathcal{M}_k^N &\leftarrow c_k^N \quad (k = 1, 2, \dots, I^N) \\ \mathcal{M}_k^V &\leftarrow c_k^V \quad (k = 1, 2, \dots, I^V)\end{aligned}\tag{4-6}$$

在每次迭代中, 根据伪标签选择 $P \times Z$ 个实例特征, 形成 NIR 小批量 q^N 和 VIS 小批量 q^V , 其中 P 表示选择的身份数量, Z 表示每个身份的实例数量。

为了在 NIR 模态中学习紧凑而具有判别性的特征, 设计了基于 NIR 集群的对比学习目标:

$$L_{NCC} = - \sum_{i=1}^{P \times Z} \log \frac{\exp(q_i^N \cdot \mathcal{M}_{y_i^N}^N / \tau)}{\sum_{k=0}^{I^N} \exp(q_i^N \cdot \mathcal{M}_k^N / \tau)}\tag{4-7}$$

其中, y_i^N 是实例特征 q_i^N 的类别, τ 是温度系数, 用于控制对难负样本的关注程度。类

似地，也为 VIS 模态设计了基于 VIS 集群的对比学习目标：

$$L_{VCC} = - \sum_{i=1}^{P \times Z} \log \frac{\exp(q_i^V \cdot \mathcal{M}_{y_i^V}^V / \tau)}{\sum_{k=0}^{I^V} \exp(q_i^V \cdot \mathcal{M}_k^V / \tau)} \quad (4-8)$$

其中， y_i^V 是实例特征 q_i^V 的类别。

在网络更新的每个迭代之后，记忆中属于 q^N 和 q^V 的集群中心会进行更新：

$$\begin{aligned} \mathcal{M}_{y_i^N}^N &\leftarrow \alpha \mathcal{M}_{y_i^N}^N + (1 - \alpha) q_i^N \quad (i = 1, 2, \dots, P \times Z) \\ \mathcal{M}_{y_i^V}^V &\leftarrow \alpha \mathcal{M}_{y_i^V}^V + (1 - \alpha) q_i^V \quad (i = 1, 2, \dots, P \times Z) \end{aligned} \quad (4-9)$$

其中， α 表示动量更新因子。通过上述的集群级记忆学习，所有实例对集群中心的更新贡献相等，避免了次优的学习目标。DCL 的目标损失公式为：

$$L_{DCL} = \lambda_1 L_{NCC} + \lambda_2 L_{VCC} \quad (4-10)$$

其中， λ_1 和 λ_2 分别是 L_{NCC} 和 L_{VCC} 的权衡权重。 L_{DCL} 增强了模态内特征的紧凑程度。

4.2.4 跨模态域集群对比学习

为了进一步提高模型性能和伪标签分配的准确性，本节提出了跨模态域集群对比学习 ICL，以学习模态不变的表示，从而动态调整记忆和优化目标。同样，基于集群的记忆学习方案可以用于避免由于数据集分布不平衡而导致性能下降^[144]。

首先，在每个训练循环的开始阶段，为了获得跨模态数据的内在关系，构建了一个 NIR-VIS 集群记忆 $\mathcal{M}^{N,V}$ ，用于存储跨模态的集群中心，集群中心采用以下形式进行构建：

$$c_k = \frac{1}{n_k^V + n_k^N} \left\{ \sum_j \mathbf{1}\{\hat{y}_j^V = k\} f_j^V + \sum_i \mathbf{1}\{\hat{y}_i^N = k\} f_i^N \right\} \quad (k = 1, 2, \dots, I) \quad (4-11)$$

其中， I 是两个模态的共同身份类别， c_k 表示第 k 个集群中心。然后，将这些集群中心

算法 4-1: 鲁棒跨模态伪标签和对比学习 RPC 方法

Input: 无标签的 NIR 训练集 D^N , 无标签的 VIS 训练集 D^V , epochs 数量 $epoch$, 迭代次数 $iter$
Output: 网络参数 Θ

```

1: for  $\kappa = 1, 2, \dots, epoch$  do
2:   生成伪标签标注的数据  $\{\hat{D}_N, \hat{D}_V\}$  通过式子 (4-1)-(4-4);
3:   构建模态特定记忆  $\{\mathcal{M}^N, \mathcal{M}^V\}$  通过式子 (4-2), (4-5), (4-6);
4:   构建跨模态记忆  $\{\mathcal{M}^{N,V}\}$  通过式子 (4-11)-(4-12);
5:   for  $\pi = 1, 2, \dots, iter$  do
6:     更新  $\Theta$  通过最小化式子 (4-10);
7:     更新记忆  $\{\mathcal{M}^N, \mathcal{M}^V\}$  式子 (4-9);
8:     更新  $\Theta$  通过最小化式子 (4-13);
9:     更新记忆  $\{\mathcal{M}^{N,V}\}$  式子 (4-14);
10:  end for
11: end for
12: Return  $\Theta$ 
    
```

存储在 NIR-VIS 集群记忆 (NIR-VIS cluster memory) 中:

$$\mathcal{M}_k^{N,V} \leftarrow c_k \quad (k = 1, 2, \dots, I) \quad (4-12)$$

在每次迭代中, 基于伪标签选择 $P \times \frac{Z}{2}$ NIR 实例特征和 $P \times \frac{Z}{2}$ VIS 实例特征, 形成 NIR-VIS 训练批次 $q^{N,V}$ 。

然后, 采用混合模态记忆对比学习, 以学习模态无关的表示。其表述如下:

$$L_{ICL} = - \sum_{i=1}^{P \times Z} \log \frac{\exp(q_i^{N,V} \cdot \mathcal{M}_{y_i}^{N,V} / \tau)}{\sum_{k=0}^I \exp(q_i^{N,V} \cdot \mathcal{M}_k^{N,V} / \tau)} \quad (4-13)$$

其中, y_i 是跨模态实例特征 $q_i^{N,V}$ 的类别。

每次迭代后, 使用跨模态实例特征 $q_i^{N,V}$ 更新其对应所在的记忆 $\mathcal{M}^{N,V}$:

$$\mathcal{M}_{y_i}^{N,V} \leftarrow \alpha \mathcal{M}_{y_i}^{N,V} + (1 - \alpha) q_i^{N,V} \quad (i = 1, 2, \dots, P \times Z) \quad (4-14)$$

根据公式4-13, 通过最小化 L_{ICL} , 可以优化跨模态内在关系, 并引导实例级特征学习稳健且模态不变的表示。

4.2.5 模型优化

RPC 网络将 NPS、DCL、ICL 整合到一个联合框架中，以生成高质量的伪标签，并学习具有判别性和模态不变的特征，很好地解决了 NIR-VIS-UHFR 任务的难题。RPC 的最终目标函数表达式如下：

$$L = \lambda_3 L_{DCL} + \lambda_4 L_{ICL} \quad (4-15)$$

其中， λ_3 和 λ_4 分别是 L_{DCL} 和 L_{ICL} 的权重。RPC 方法的训练步骤总结如算法 4-1 所示，其中 $epoch$ 表示训练的轮数， κ 表示轮数的索引， $iter$ 表示带有索引 π 的训练迭代次数。

4.3 实验结果与分析

为验证无监督 RPC 方法的有效性，使用 RPC 在几个主流 NIR-VIS 人脸数据集上进行了大量实验，包括 CASIA NIR-VIS 2.0^[12]、BUAA NIR-VIS^[112]、Oulu-CASIA NIR-VIS^[111]、LAMP-HQ NIR-VIS^[14] 数据集。首先，进行消融研究以验证 RPC 方法中的每个组件，即 NPS、DCL、ICL，对模型性能的贡献。此外，将 RPC 网络与主流的有监督方法进行了比较，包括 SAFE^[74]、MMTN^[72]、VSANet^[14]、EGCH^[82]、ADCANs^[131]、CFC^[27]、PACH^[71]、PCFH^[70]、ADFL^[76]、WCNN^[93]、CDL^[130]、LightCNN-29^[129]、IDR^[128]、TRIVET^[92]，以验证 RPC 方法的有效性。

4.3.1 实验数据集

为了进行公平比较，RPC 方法和其他主流的有监督 NIR-VIS 人脸识别方法采用相同的测试设置。在训练过程中，RPC 网络不使用任何手工标注的标签。

CASIA NIR-VIS 2.0 dataset^[12] 作为最受欢迎的人 NIR-VIS 人脸数据集之一，包含近 18,000 张图像。其中的图片来源于 725 个受试者并具有丰富的面部变化。特别是，每个身份的 NIR 图像数量通常高于 VIS 图像数量。根据数据集的官方协议，参考文献^[93,94]中的工作，确保测试集 (包括 358 个身份) 和训练集 (包括 357 个身份) 不交叉包含部分相同内容。与有监督方法不同，NIR-VIS-UHFR 的训练集是在没有身份标签的情况下构

建的。为保证公平比较，测试集与有监督的设置相同。

BUAA NIR-VIS dataset^[112] 是一个常用的异构人脸数据集。它包含 150 个身份，每个身份对应 9 对具有不同视角和表情的 NIR 和 VIS 图像。训练集包含与 50 个身份相关的 900 个实例，测试集包含来自 100 个身份的 1,800 个实例。在 NIR-VIS-UHFR 的训练设置中，所有训练标签被删除。

Oulu-CASIA NIR-VIS dataset^[111] 数据集包含 80 个不同身份个体，具有六种表情(愤怒、厌恶、恐惧、快乐、悲伤、惊讶)和三种光照条件(正常室内照明、弱光、暗黑)。参考此数据集的原始协议，选择了 20 个身份的 1,920 张图像作为训练集，并选择了另外 20 个身份的相同数量的图像作为测试集。同样地，在无监督设置下训练集中将不包含手工标签。

LAMP-HQ NIR-VIS dataset^[14] 是目前最新和最大的 NIR-VIS 数据集，分别包含 56,788 张 NIR 图像和 16,828 张 VIS 图像。特别地，每个身份的 NIR 图像数量和 VIS 图像数量是不相等的。跨模态数据包含丰富的表情、角度、光照的变化。参考官方协议^[14]，验证时进行了 10 次实验，并且测试集与训练集之间没有重叠。此外，训练集的身份标签完全被丢弃。

4.3.2 实验设置

RPC 模型使用深度学习框架 TensorFlow 实现，并在 TITAN Xp GPU 上进行训练。一采用 LWNet 的轻量级网络构建作为主干网络 $F(\cdot; \Theta)$ 。如图 4-7 所示，LWNet 的网络结构由一个 Stem 块、12 个 Inception-ResNet-A 块、一个 Reduction-A 块、3 个 Inception-ResNet-B 块^[30,136]、一个全局平均池化层、一个全连接层构成。

首先，LWNet 网络在 MS-Celeb-1M 数据集^[109] 上进行预训练，其中图像使用 MTCNN 网络^[124] 裁剪到 160×160 。然后，在 LFW 数据集^[127] 上测试预训练网络，并获得 99.3% 的准确率。在预训练网络的基础上，使用所有丢弃标签的 NIR-VIS 人脸数据集 $\{x_i^N, x_i^V\}$ 对网络进行微调，图像同样被裁剪到 160×160 。采用 DBSCAN 生成伪标签，其中邻域半径设置为 0.6，每个集群中的最小实例数设置为 4。训练批次大小 $P \times K$ 和每个 epoch 的训练批数 $iter$ 分别设置为 64 和 300。选定的身份数量 P 和每个身份的实例特征数量 K 分别设置为 4 和 16。损失函数的学习率设置为 0.002，每 1200 个训练批次衰减一次，

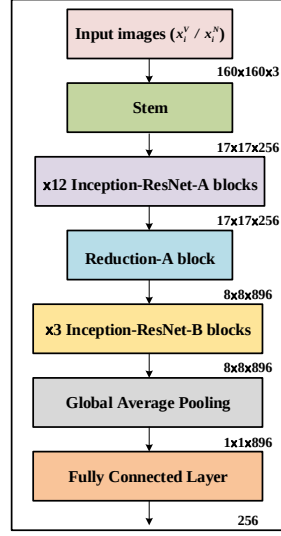


图 4-7 LWNNet 的网络结构

衰减因子为 0.95。权重 λ_1 、 λ_2 、 λ_3 、 λ_4 都设置为 1.0。超参数 α 和 τ 分别设置为 0.1 和 0.12。

4.3.3 模块性能分析

在本节进行大量实验，以验证每个模块 (即 NPS、DCL、ICL) 的有效性。由于 CASIA NIR-VIS 2.0 和 LAMP-HQ NIR-VIS 数据集中存在多个协议，参考最近的工作^[72,74] 统一在这两个数据集中都使用第一个协议，验证 RPC 方法的各种性质。

消融实验: 如表4-1所示，为了评估 RPC 网络中每个组件的有效性，在三个 NIR-VIS 数据库 (即 CASIA NIR-VIS 2.0、LAMP-HQ NIR-VIS、Oulu-CASIA NIR-VIS) 中进行了消融研究。首先采用预训练的 LWNNet 模型作为基线，并将其命名为 BL。通过将模块叠加到 BL 中，可以得到三种组合，即 BL、BL+NPS+DCL、BL+NPS+DCL+ICL。在 CASIA NIR-VIS 2.0 中，BL+NPS+DCL 在 Rank-1 上的表现比 BL 提高了 4.7% (98.2% – 93.5%)，在 VR@FAR=0.1% 上提高了 10.9% (96.9% – 86.0%)，表明 DCL 能够通过 NPS 估计的高质量伪标签学习具有紧凑性和区分性的模态域内表示。BL+NPS+DCL+ICL 相较于 BL+NPS+DCL 获得了更好的性能，Rank-1 提高了 1.5% (99.7% – 98.2%)，在 VR@FAR=0.1% 上提高了 2.7% (99.6% – 96.9%)，这表明 ICL 能够关联 NIR 和 VIS 数据之间的内在关系，从而学习模态域不变的表示。在 LAMP-HQ NIR-VIS 和 Oulu-CASIA NIR-VIS 数据库中也可以得出类似的结论。在 Oulu-CASIA NIR-VIS 中，BL+NPS+DCL

表 4-1 CASIA NIR-VIS 2.0、LAMP-HQ NIR-VIS、Oulu-CASIA NIR-VIS 数据库上的消融实验

方法	CASIA NIR-VIS 2.0		LAMP-HQ NIR-VIS		Oulu-CASIA NIR-VIS	
	Rank-1	VR@FAR = 0.1%	Rank-1	VR@FAR = 0.1%	VR@FAR = 1%	VR@FAR = 0.1%
BL	93.5	86.0	93.6	66.7	91.7	78.5
BL+NPS+DCL	98.2	96.9	98.4	97.4	98.3	91.4
BL+NPS+DCL+ICL	99.7	99.6	98.9	98.4	98.5	94.4

在 VR@FAR=1% 上超过 BL 了 6.6% (98.3% – 91.7%), 在 VR@FAR=0.1% 上超过 BL 了 12.9% (91.4% – 78.5%)。在 LAMP-HQ NIR-VIS 中, BL+NPS+DCL 在 Rank-1 上的表现优于 BL, 提高了 4.8% (98.4% – 93.6%), 在 VR@FAR=0.1% 上提高了 30.7% (97.4% – 66.7%)。这些结果表明, 在具有大量类内变化的数据集 LAMP-HQ NIR-VIS, DCL 仍然能够学习 NIR 和 VIS 模态域的紧凑类内身份表示。此外, BL+NPS+DCL+ICL 在 LAMP-HQ NIR-VIS 中的表现优于 BL+NPS+DCL, Rank-1 提高了 0.5% (98.9% – 98.4%), 在 VR@FAR=0.1% 上提高了 1.0% (98.4% – 97.4%), 以及在 Oulu-CASIA NIR-VIS 中 VR@FAR=1% 上提高了 0.2% (98.5% – 98.3%), 在 VR@FAR=0.1% 上提高了 3.0% (94.4% – 91.4%), 这验证了 ICL 能够进一步学习模态域不变和有区分性的表示。

主干网络分析: 表4-2展示了主流的方法和 RPC 方法的主干网络分析。LWNet 包含 11 亿次 Mult-Adds 和 550 万个参数, 而 LightCNN-29 包含 38 亿次 Mult-Adds 和 1260 万个参数。因此, LightCNN-29 中的模型参数量是 LWNet 的两倍多。与使用 LWNet 作为主干相比, 使用 LightCNN-29 的 RPC 性能更好, 但代价是付出额外的参数量。此外, RPC 使用 LightCNN-29 作为主干网络的性能与其他使用 LightCNN-29 主干网络主流的方法相当, 验证了无监督 RPC 方法的先进性。

表 4-2 RPC 和其他主流的方法在 CASIA NIR-VIS 2.0、LAMP-HQ NIR-VIS 的主干网络性能

主干网络	方法	CASIA NIR-VIS 2.0		LAMP-HQ NIR-VIS	
		Rank-1	VR@FAR=0.1%	Rank-1	VR@FAR=0.1%
LWNet	RPC	99.7 ± 0.0	99.6 ± 0.2	98.9 ± 0.2	98.7 ± 0.3
	PCFH	98.8 ± 0.3	97.7 ± 0.3	95.3 ± 0.5	75.1 ± 1.8
LightCNN-29	PACH	98.9 ± 0.2	98.3 ± 0.2	95.4 ± 0.5	75.3 ± 1.7
	VSANet	99.2 ± 0.1	98.2 ± 0.2	97.3 ± 0.2	78.2 ± 3.0
	MMTN	99.4 ± 0.0	98.7 ± 0.0	97.7 ± 0.0	79.2 ± 0.3
	RPC	99.8 ± 0.1	99.7 ± 0.1	99.0 ± 0.2	98.8 ± 0.3

NPS 的有效性: 为了验证 NPS 生成的伪标签的质量, 本研究展示了伪标签的准确性分析。图4-8 (a) 和 (b) 展示了真实标签数、离群点数、正确伪标签、错误伪标签的数量。从结果中可以观察到, 随着 epochs 数从 0 增加到 50, 正确伪标签的数量逐渐增加, 而离群点和错误伪标签的数量逐渐减少。这表明 RPC 方法可以降低初始嘈杂伪标签的影

表 4-3 在 NIR-VIS 人脸数据集中伪标签估计的结果

数据集	离群点	正确伪标签	错误伪标签	真实集群数量	预测的集群数量
CASIA NIR-VIS 2.0	24	8544	0	357	357
LAMP-HQ NIR-VIS	64	38195	64	300	300
Oulu-CASIA NIR-VIS	0	1920	0	20	20
BUAA NIR-VIS	0	900	0	50	50

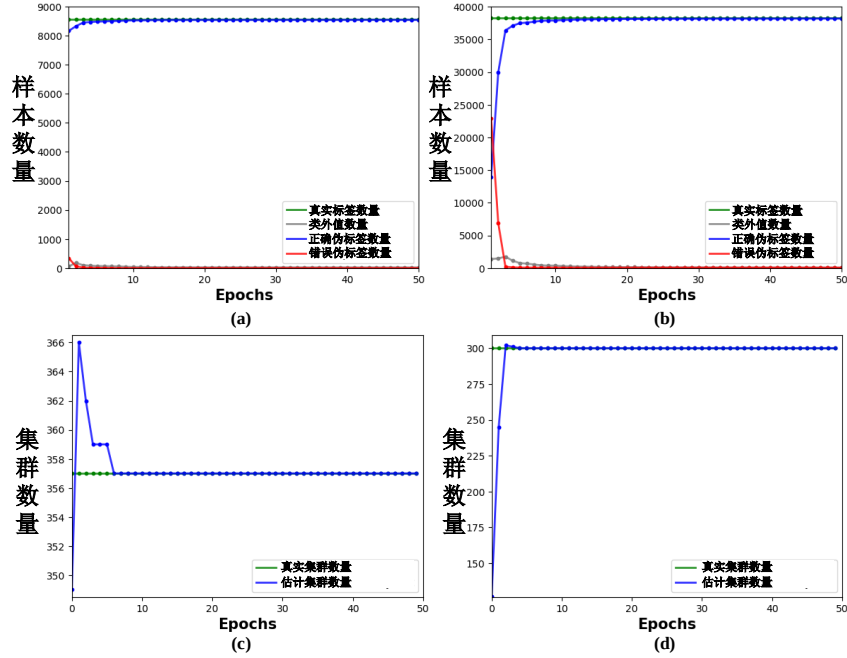


图 4-8 RPC 在 CASIA NIR-VIS 2.0(a) 和 LAMP-HQ NIR-VIS(b) 上生成的伪标签质量，以及 RPC 在 CASIA NIR-VIS 2.0(c) 和 LAMP-HQ NIR-VIS(d) 上估计的集群数量

响，逐渐提高特征表示的鉴别性，并纠正错误的关联关系。图4-8 (c) 和 (d) 显示了 RPC 在 epochs 数从 0 到 50 的情况下估计的集群数。通过加入 DCL 和 ICL，NPS 逐渐获得跨模态数据中的正确关联，以生成可靠的集群中心。表4-3显示了在四个主流 NIR-VIS 数据集上的伪标签估计结果。结果表明，在小规模数据集，即 Oulu-CASIA NIR-VIS 和 BUAA NIR-VIS 上，可以实现 100% 的伪标签估计准确率，这表明 RPC 方法可以生成可靠和有区分度的集群。在大规模数据集，即 CASIA NIR-VIS 2.0 和 LAMP-HQ NIR-VIS 上，RPC 可以实现超过 99% 的伪标签估计准确率，并能够利用大多数的难样本 (即只有少数离群点)。这些结果意味着无监督方案有可能减少手工标注的依赖。以无监督的方式可以生成可靠的标签，并在 NIR-VIS 人脸识别中获得令人满意的性能，这为 NIR-VIS 人脸识别带来了一种新的有效方案。

可视化分析：图 4-9 展示了基线模型和 RPC 方法在 CASIA NIR-VIS 2.0 和 LAMP-HQ 数据集上的类内相似度和类间相似度分布。在 CASIA NIR-VIS 2.0 和 LAMP-HQ

NIR-VIS 数据集中，可以发现到 RPC 网络能够有效增加类内相似度并降低类间相似度，从而学习有序的子空间。图 4-10 (a) 显示了预训练网络提取的特征的初始分布，其中存在较大的模态差异。而图 4-10 (b) 显示了 RPC 最终微调的结果。在没有任何手工标签的情况下，所提出的 NPS、DCL、ICL 模块能够获得跨模态数据的内在关联，消除模态差异并学习具有辨识力的身份特征。

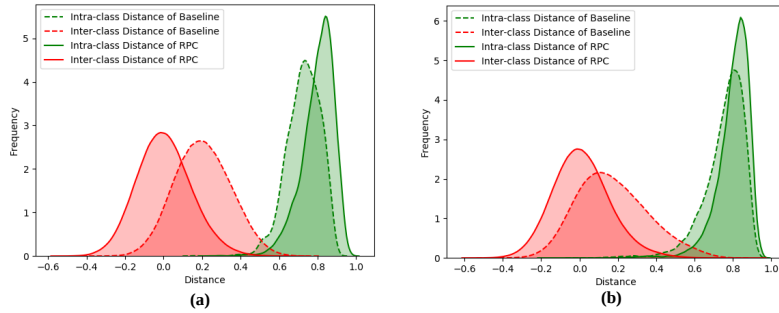


图 4-9 基线模型和 RPC 方法在 CASIA NIR-VIS 2.0 和 LAMP-HQ 数据集上的类内距离和类间距离分布。(a) 显示了在 CASIA NIR-VIS 2.0 上的类内和类间距离分布，(b) 显示了在 LAMP-HQ NIR-VIS 上的类内和类间距离分布。绿色表示类内距离，红色表示类间距离

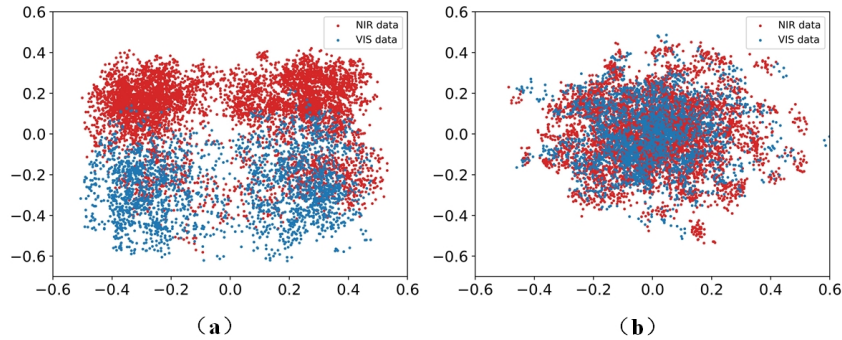


图 4-10 特征分布示意图。(a) 展示了在 CASIA NIR-VIS 2.0 数据集上由预训练网络提取的特征的初始分布。(b) 是 RPC 最终微调的结果

4.3.4 与主流的方法对比

本节，将无监督 RPC 方法与有监督方法进行性能比较。

CASIA NIR-VIS 2.0 实验结果: 对比结果显示在表4-4中。提出的 RPC 方法比一些方法表现更优异，比如 TRIVET^[92]、IDR^[128]、LightCNN-29^[129]、CDL^[130]、WCNN^[93]、ADFL^[76]、SDDL^[106]、CFC^[27]。体现在，RPC 在 Rank-1 为 99.7%，VR@FAR=1% 为 99.9%，VR@FAR=0.1% 为 99.6% 的情况下胜过了 WCNN 模型。此外，与一些有竞争力的方法

相比, 如 PCFH^[70]、PACH^[71]、ADCANs^[131]、VSANet^[14]、MMTN^[72]、SAFE^[74], RPC 能取得与有监督方法相当甚至更优的性能。结果表明, 即使 RPC 舍弃了所有的真实标签, 所提出的 NPS 模块仍然能够生成高质量的伪标签, 以引导 DCL 和 ICL 学习模态不变和紧凑的表示。

表 4-4 不同方法在 CASIA NIR-VIS 2.0 数据库上的结果

种类	方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
有监督	TRIVET ^[92] (ICB 2016)	95.7 ± 0.5	98.1 ± 0.3	91.0 ± 1.3
	IDR ^[128] (AAAI 2017)	97.3 ± 0.4	98.9 ± 0.3	95.7 ± 0.7
	LightCNN-29 ^[129] (TIFS 2018)	96.7 ± 0.2	98.5 ± 0.6	94.8 ± 0.4
	ADFL ^[76] (AAAI 2018)	98.1 ± 0.4	99.1 ± 0.2	97.1 ± 0.5
	CDL ^[130] (AAAI 2018)	98.6 ± 0.2	99.5 ± 0.1	98.3 ± 0.1
	WCNN ^[93] (TPAMI 2018)	98.7 ± 0.3	99.5 ± 0.1	98.4 ± 0.4
	PCFH ^[70] (IJCAI 2019)	98.8 ± 0.3	99.6 ± 0.1	97.7 ± 0.3
	PACH ^[71] (CVPR 2020)	98.9 ± 0.2	99.6 ± 0.1	98.3 ± 0.2
	CFC ^[27] (TPAMI 2020)	98.6 ± 0.1	99.2 ± 0.1	97.3 ± 0.2
	SDDL ^[106] (PR 2021)	92.1 ± 0.7	-	88.7 ± 1.0
	ADCANs ^[131] (TMM 2021)	99.1 ± 0.2	99.6 ± 0.1	98.5 ± 0.2
	VSANet ^[14] (IJCV 2021)	99.2 ± 0.1	99.7 ± 0.0	98.2 ± 0.2
	EGCH ^[82] (ICPR 2021)	98.8 ± 0.2	99.6 ± 0.0	97.7 ± 0.2
	MMTN ^[72] (TIFS 2022)	99.4 ± 0.0	99.8 ± 0.0	98.7 ± 0.0
	SAFE ^[74] (TPAMI 2022)	98.9 ± 0.1	99.6 ± 0.1	98.6 ± 0.1
无监督	RPC(本章方法)	99.7 ± 0.0	99.9 ± 0.0	99.6 ± 0.2

Oulu-CASIA NIR-VIS 实验结果: 无监督 RPC 方法与主流的有监督方法的比较结果显示在表4-5中。RPC 方法在多个指标上达到满意的结果: 在 Rank-1 为 100%, VR@FAR=1% 为 98.5%, VR@FAR=0.1% 为 94.4%。RPC 方法与有监督方法性能相当甚至更优, 如 PCFH^[70]、PACH^[71]、VSANet^[14]、MMTN^[72]、SAFE^[74]。这表明了即使不使用真实标签, RPC 能够有效解决模态差异问题。

表 4-5 不同方法在 Oulu-CASIA NIR-VIS 数据库上的结果

种类	方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
有监督	TRIVET ^[92] (ICB 2016)	92.2	67.9	33.6
	IDR ^[128] (AAAI 2017)	94.3	73.4	46.2
	LightCNN-29 ^[129] (TIFS 2018)	96.7	92.4	65.1
	ADFL ^[76] (AAAI 2018)	95.5	83.0	60.7
	CDL ^[130] (AAAI 2018)	94.3	81.6	53.9
	WCNN ^[93] (TPAMI 2018)	98.0	81.5	54.6
	PCFH ^[70] (IJCAI 2019)	100	97.7	86.6
	PACH ^[71] (CVPR 2020)	100	97.9	88.2
	CFC ^[27] (TPAMI 2020)	99.9	98.1	90.7
	ADCANs ^[131] (TMM 2021)	99.8	93.2	78.9
	VSANet ^[14] (IJCV 2021)	100	97.7	89.0
	EGCH ^[82] (ICPR 2021)	100	98.0	89.5
	MMTN ^[72] (TIFS 2022)	100	99.2	89.0
	SAFE ^[74] (TPAMI 2022)	100	98.3	91.8
无监督	RPC(本章方法)	100	98.5	94.4

BUAA NIR-VIS 实验结果: 如表4-6所示, 无监督 RPC 方法与主流的有监督模型进行了比较。首先, RPC 模型大幅超越了一些对比方法, 包括 TRIVET^[92]、IDR^[128]、

LightCNN^[129]、CDL^[130]、WCNN^[93]、ADFL^[76]、PCFH^[70]、PACH^[71]。例如, RPC 在 Rank-1 方面超过 PACH, 达到 1.4% (100% – 98.6%), 在 VR@FAR=1% 方面超过 1.9% (99.9% – 98.0%)。其次, RPC 与当前主流的有监督方法, 如 CFC^[27]、ADCANs^[131]、VSANet^[14]、MMTN^[72]、EGCH^[82]、SAFE^[74] 的性能相媲美, 在 Rank-1 为 100%, VR@FAR=1% 为 99.9%, VR@FAR=0.1% 为 99.8%。结果表明, 无监督的 RPC 降低对手工标签的依赖。然后, 提出的 ICL 和 DCL 能够逐渐优化网络, 在存在噪声标签的情况下实现满意的性能。

表 4-6 不同方法在 BUAA NIR-VIS 数据库上的结果

种类	方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
有监督	TRIVET ^[92] (ICB 2016)	93.9	93.0	80.9
	IDR ^[128] (AAAI 2017)	94.3	93.4	84.7
	LightCNN-29 ^[129] (TIFS 2018)	96.5	95.4	86.7
	ADFL ^[76] (AAAI 2018)	95.2	95.3	88.0
	CDL ^[130] (AAAI 2018)	96.9	95.9	90.1
	WCNN ^[93] (TPAMI 2018)	97.4	96.0	91.9
	PCFH ^[70] (IJCAI 2019)	98.4	97.9	92.4
	PACH ^[71] (CVPR 2020)	98.6	98.0	93.5
	CFC ^[27] (TPAMI 2020)	99.7	98.7	97.8
	ADCANs ^[131] (TMM 2021)	99.8	99.7	98.4
	VSANet ^[14] (IJCV 2021)	98.8	98.3	93.4
	EGCH ^[82] (ICPR 2021)	98.1	92.7	
	MMTN ^[72] (TIFS 2022)	99.2	99.5	97.2
	SAFE ^[74] (TPAMI 2022)	99.4	99.0	98.0
无监督	RPC(本章方法)	100	99.9	99.8

LAMP-HQ NIR-VIS 实验结果: 表4-7展示了无监督 RPC 与主流的有监督方法的比较结果。由于 LAMP-HQ NIR-VIS 数据集是当前最新的 NIR-VIS 人脸数据集, 一些主流的有监督方法并未给出该数据集上的性能报告。因此, 只能与部分有监督方法进行了比较, 包括 LightCNN-29^[129]、ADFL^[76]、PCFH^[70]、PACH^[71]、ADCANs^[131]、VSANet^[14]、MMTN^[72]。结果表明, RPC 方法获得了最佳性能。例如, RPC 在 Rank-1 方面比 MMTN 表现更好, 提升了 1.2% (98.9%–97.7%), 在 VR@FAR=1% 方面提升了 3.7% (99.4%–95.7%), 在 VR@FAR=0.1% 方面提升了 19.5% (98.7% – 79.2%)。上述结果表明即使在大规模且最具挑战性的 NIR-VIS 数据集上不使用真实标签, RPC 方法仍然能够通过生成正确的伪标签作为监督信号来学习模态不变的表示, 这对 NIR-VIS 人脸识别方法的设计产生积极影响。

表 4-7 不同方法在 LAMP-HQ NIR-VIS 数据库上的结果

种类	方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
有监督	LightCNN-29 ^[129] (TIFS 2018)	94.6 ± 0.3	92.5 ± 0.6	75.6 ± 1.9
	ADFL ^[76] (AAAI 2018)	95.1 ± 0.5	92.1 ± 0.9	73.3 ± 2.2
	PCFH ^[70] (IJCAI 2019)	95.3 ± 0.5	92.9 ± 0.6	75.1 ± 1.8
	PACH ^[71] (CVPR 2020)	95.4 ± 0.5	93.1 ± 0.4	75.3 ± 1.7
	ADCANs ^[131] (TMM 2021)	97.8 ± 0.3	98.8 ± 0.3	96.1 ± 0.4
	VSANet ^[14] (IJCV 2021)	97.3 ± 0.2	94.8 ± 0.7	78.2 ± 3.0
	MMTN ^[72] (TIFS 2022)	97.7 ± 0.0	95.7 ± 0.6	79.2 ± 0.0
无监督	RPC(本章方法)	98.9 ± 0.2	99.4 ± 0.2	98.7 ± 0.3

4.3.5 通用性分析

为了进一步验证 RPC 方法在其他跨模态人脸数据上的有效性, 使用 CUHK 学生人脸数据库的素描-VIS 子数据库进行了实验^[156]。按照先前方法^[157,158] 的实验设置, 该数据库分为 88 个身份用于训练和 100 个身份用于测试。由于每个身份只有一对 VIS-素描图像, 本节采用了随机水平翻转、随机擦除、随机裁剪等数据增强方法来生成更多数据用于聚类。最终, 每个身份在训练集中有 220 张 VIS 图片和 220 张素描图片。如表4-8所示, BL+NPS+DCL 的准确性比 BL 提高了 6.0%。此外, BL+NPS+DCL+ICL 在准确性上超过了 BL+NPS+DCL, 提高了高达 25%。总体而言, 上述结果验证 RPC 方法在 CUHK 学生人脸数据库的素描-VIS 子数据库上表现出有效性, 因此表明了 RPC 在素描-VIS 跨模态人脸识别任务中依然通用。

表 4-8 素描-VIS CUHK 学生数据集上的准确率

方法	准确率 (%)
BL	68.0
BL+NPS+DCL	74.0
BL+NPS+DCL+ICL	99.0

4.4 本章小结

本文提出了鲁棒跨模态伪标签和对比学习 RPC 网络, 实现无监督设置下的 NIR-VIS 异构人脸识别。所提出的 RPC 将基于 NIR 集群的伪标签共享 NPS、模态域特定集群对比学习 DCL、跨模态域集群对比学习 ICL 结合成一个统一的联合框架中, 获得跨模态数据中的内在关联并学习模态不变的表示。具体而言, 首先使用 NPS 来估计准确的伪标签作为监督信号。其次, DCL 增强了模态内特征的区分能力。最后, ICL 引导网络学

习可靠且模态无关的特征表示。与有监督方法相比，RPC 实现了媲美有监督方法的性能。这些结果表明 RPC 适用于大规模、复杂且缺乏手工注释的实际场景。

第 5 章 同质到异质学习和剩余不变增强方法

在前几章, 本文针对模态差异、剩余差异、无监督学习问题分别提出了 SSL、NLPP、RPC 几个有效的解决方法。然而, 目前还没有方法同时考虑这三个有挑战的问题来提升 NIR-VIS 人脸识别框架的准确性和便利性。本章同时考虑了这三个有挑战的问题, 设计了无监督方法, 在不依赖标签信息的情况下消除 NIR-VIS 人脸中的模态差异和剩余差异。为此, 提出了一种新颖的同质到异质学习和剩余不变增强 (homogeneous-to-Heterogeneous learning and Residual-invariant Enhancement, HERE) 网络。顾名思义, HERE 的优化遵循“同质到异质学习”策略, 充分探索不同模态之间的互补和共同语义信息。在同质学习阶段, 模态-对抗对比学习 (Modality-Adversarial Contrastive Learning, MACL) 充分利用对比学习和对抗学习的协同作用。一方面, MACL 分别学习 NIR 和 VIS 数据的紧凑且具有区分性的模态内部表示。另一方面, 即使在没有跨模态身份标注的情况下, MACL 也能对齐 NIR-VIS 数据在共享特征空间中的分布, 有效降低模态差异。在异质学习阶段, 基于 K-相互编码的跨模态标记 (K-reciprocal-Encoding-based Cross-modal Labeling, KECL) 被引入用于伪标签估计, 充分挖掘跨模态关系, 并将跨模态特征分组到集群中。通过 KECL 提供的伪标签, 精细跨模态对比学习 (Refined cross-modal Contrastive Learning, RCL) 模块在对比学习的基础上引入了模态不变的平均初始化和动态关注加权策略, 学习模态不变特征。最后, 剩余不变表示增强 (Residual-invariant Representations Enhancement, RRE) 在跨模态人脸中挖掘局部特征以进行精细化的匹配。与有监督方法相比, 无监督 HERE 在多个数据集上表现出有竞争力的性能。此外, 本章的方法可以降低数据采集要求和成本, 具有更大的可扩展性和实用性。这一研究成果已在 IEEE Transactions on Information Forensics and Security 正式发表。

5.1 引言

为了实现全天候的人脸识别, 现代监控摄像头配备了自动切换到近红外模式的功能, 使其能够在夜间捕捉近红外人脸图像。由于 NIR-VIS 人脸图像在不同波长范围内

获取,即使从相同的视点捕获,它们也呈现出显著的外观差异,这使得 NIR-VIS 人脸识别成为一个具有挑战性的问题。因此,作为异构人脸识别中最具代表性的任务之一,NIR-VIS 人脸识别引起了越来越多的关注,并在安全监控、刑事调查、多媒体信息检索中发挥着重要作用^[72,97,128]。

当前的异构人脸识别方法严重依赖于手工标注,以身份标签作为监督信号减小模态差异。然而,从大量原始跨模态人脸数据中获取注释信息是耗时且昂贵的,从不同跨模态图像中进行手工标注是具有挑战的。这一标注问题直接限制了数据集的规模,导致当前跨模态人脸数据集中样本和身份类别有限。这可能导致过拟合问题,不利于有监督异构人脸识别方法在更大规模的实际场景中进行部署^[93,95,159]。为解决这一难题,本章研究一个更具挑战性的无监督学习框架,称为无监督 NIR-VIS 异构人脸识别 (Unsupervised NIR-VIS Heterogeneous Face Recognition, NIR-VIS-UHFR),其目标是从未标记的 NIR-VIS 训练数据中学习模态不变表示。

目前,有监督的 NIR-VIS 人脸识别方法已经取得了一定的成功^[14,73,98],但 NIR-VIS-UHFR 仍然是一个相对未被深入研究的领域。一个相关的领域是伪标签学习,它提供了无需身份标签就可以训练的便利性。然而,伪标签中存在的错误阻碍了无监督方法的性能。为解决这一问题,许多工作致力于提高伪标签估计的准确性,并为多个领域设计了抗噪声学习目标,如图像分类^[160]、目标检测^[161]、行人重识别^[144]、深度估计^[162]等。在这些领域中,自监督方法^[163]、基于聚类的方法^[144]、知识蒸馏方法^[162]、对抗学习方法^[143]等均给出了振奋人心的结果。然而,这些伪标签学习方法是为同质数据设计的,不适用于更具挑战性的异质数据。

针对跨模态数据的无监督学习,许多方法尝试通过利用不同模态的互补信息来提升聚类性能。Alwassel 等人^[164]提出了一种跨模态深度聚类方法,通过在音频模态聚类作为视频模态的监督信号,有助于利用两个模态之间的语义相关性来减小差异。Liu 等人^[165]对单一模态特征进行聚类,并最大化跨模态表示之间的相互信息,减少来自视觉和文本数据的域漂移。为了保留共同的跨模态信息并减少不必要的模态内信息,Mao 等人^[166]提出了一种用于跨模态聚类的深度互信息优化方法。为了解决高维表示中缺少层次和结构信息的问题,Gao 等人^[167]通过深度典型相关分析实现了跨模态聚类。虽然这些方法提供了一些设计思路,但由于以下原因不能直接应用它们到 NIR-VIS 人脸识别

任务中：(1) 这些方法是为图像到文本、视频到文本、音频到视频等特定任务设计的，与 NIR-VIS 人脸识别任务有根本性的不同。(2) 如图5-1所示，除了模态差异之外，姿态、遮挡、表情、光照等剩余信息存在，跨模态人脸图像呈现出较大的类内变化^[97]。因此，相同身份的实例特征难以聚类到同一类别，并且在匹配过程中容易出现错误匹配。

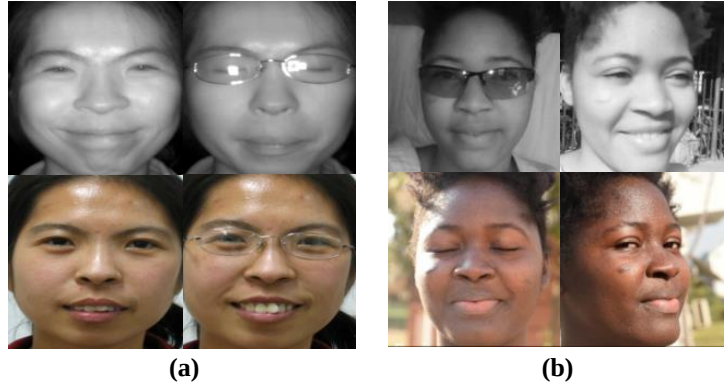


图 5-1 NIR-VIS 数据呈现出显著的类内变化，主要是由模态变化和剩余变化引起的。(a) 来自 CASIA NIR-VIS 2.0 数据集 (b) 来自 LAMP-HQ NIR-VIS 数据集

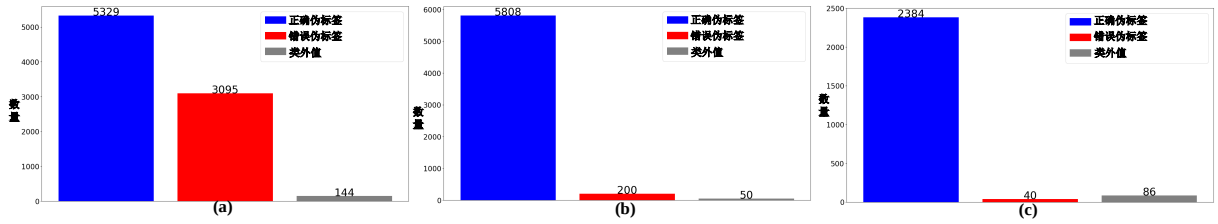


图 5-2 对 CASIA NIR-VIS 2.0 初始分配的伪标签进行准确性分析。(a) 在整个数据集上使用 DBSCAN。(b) 在近红外数据上使用 DBSCAN。(c) 在可见光数据上使用 DBSCAN

基于以上分析，来自多个任务的伪标签学习和无监督的跨模态检索方法可以作为强有力的参考，但仍需要专门针对 NIR-VIS 异构人脸识别设计无监督学习方案。本文提出了用于 NIR-VIS-UHFR 任务的同质到异质学习和剩余不变增强 (homogeneous-to-Heterogeneous learning and Residual-invariant Enhancement, HERE) 网络。为了学习不同模态之间的互补和共同语义信息，HERE 采用了“同质到异质学习”策略，包括四个模块：模态-对抗对比学习 (Modality-Adversarial Contrastive Learning, MACL)、基于 K-相互编码的跨模态标记 (K-reciprocal-Encoding-based Cross-modal Labeling, KECL)、精细跨模态对比学习 (Refined cross-modal Contrastive Learning, RCL)、剩余不变表示增强 (Residual-invariant Representations Enhancement, RRE)。如图5-2 (b) - (c) 所示，可以观察到 DBSCAN 算法^[150] 能够为从预训练网络中提取的同模态特征分配高度准确的伪标签。如图5-3所

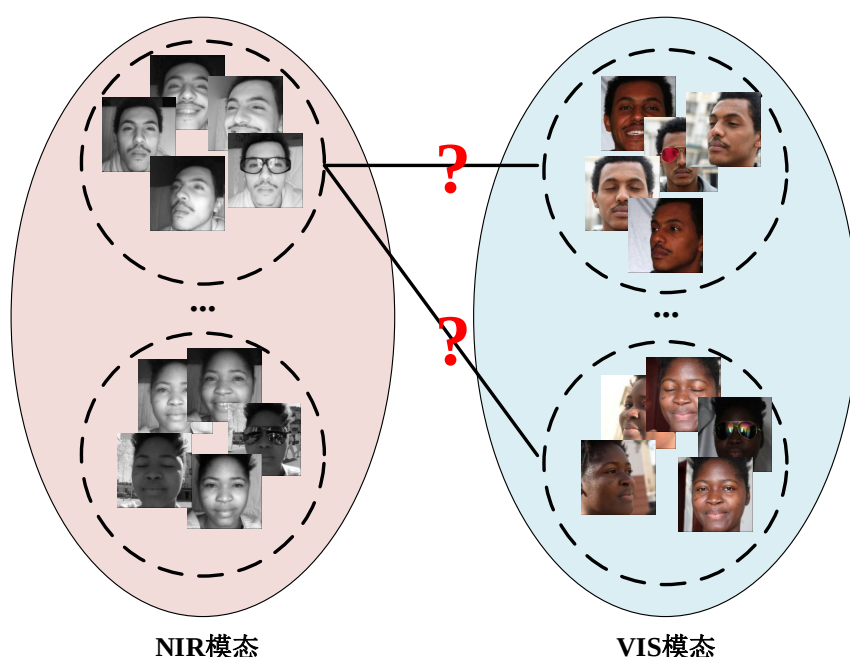


图 5-3 在执行模态内聚类后，模态之间的身份关系还是不明确的，即称为身份孤立

示，使用 DBSCAN^[150] 分别进行模态内聚类时，近红外和可见光数据被隔离开来，意味着它们不共享统一的标签域，这使得无法完全弥合模态差异。为解决这个问题，MACL 将模态对比学习与对抗学习结合起来，即使两种模态之间的身份关系不明确的情况下也能有效减小模态差异。一方面，模态对比学习分别学习近红外和可见光数据的紧凑且具有区分性的模态内表示。另一方面，对抗学习确保 NIR-VIS 数据符合共享特征空间中的分布。此外，通过引入 KECL，可以克服为 NIR-VIS 数据分配统一伪标签的挑战。然后，RCL 采用模态不变的平均初始化和动态关注加权策略，借助伪标签以类似有监督的方式学习模态不变表示。最后，RRE 目的在于提高模型对剩余变化的鲁棒性，挖掘和优化局部的特征，并将其与全局特征结合，实现精细的跨模态检索。HERE 方法的主要贡献可以总结如下：

- 提出了一种新颖的同质到异质学习和剩余不变增强 HERE 网络。该网络在无监督学习框架内集成了模态差异消除和剩余不变表示增强的能力，并在多个 NIR-VIS 数据集上展示了与最近有监督学习方法相媲美的性能。
- 模态-对抗对比学习 MACL 充分发挥了对比学习和对抗学习的协同作用。即使在两种模态之间的身份关系不明确的情况下，依然能够缩小两种模态分布并学习模态内有区别的特征表示。

- 基于 K-相互编码的跨模态标记 KECL 通过挖掘跨模态 K-相互最近邻关系，有效解决为 NIR-VIS 数据分配统一伪标签的挑战。
- 精细跨模态对比学习 RCL 采用模态不变的平均初始化和动态关注加权策略，利用由 KECL 估计的伪标签以类似有监督的方式学习模态不变表示。
- 剩余不变表示增强 RRE 探索并优化局部特征，并将其与全局特征结合，提高模型对剩余变化的鲁棒性。

5.2 同质到异质学习和剩余不变增强模型

同质到异质学习和剩余不变增强 HERE 网络结构如图5-4所示。首先，在同质学习过程中，模态-对抗对比学习 MACL 用于学习具有模态内紧凑性和跨模态分布对齐一致性的 NIR 和 VIS 人脸表示。为了充分利用全局跨模态关系，还设计了基于 K-相互编码的跨模态标记 KECL 来解决为跨模态数据分配统一标签的挑战。然后，精细跨模态对比学习 RCL 采用模态不变的平均初始化和动态关注加权策略消除模态差异。最后，剩余不变表示增强 RRE 学习 NIR-VIS 人脸的全局和局部线索，增强剩余不变表示。HERE 方法的详细内容将在以下子节中进行详细描述。

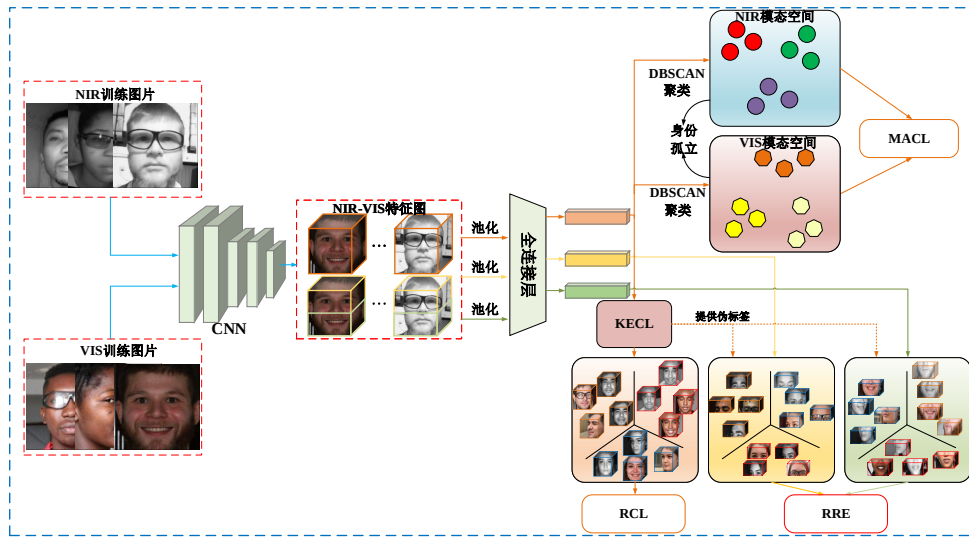


图 5-4 HERE 模型的流程图。HERE 网络包括四个关键模块，包括模态-对抗对比学习 MACL、基于 K-相互编码的跨模态标记 KECL、精细跨模态对比学习 RCL、剩余不变表示增强 RRE。本联合方案在更具挑战性的无监督设置中提高了模型的模态泛化能力和对剩余变化的鲁棒性

5.2.1 问题建模

在 NIR-VIS-UHFR 中，采用大规模的可见光人脸数据集 (Ms-celeb-1m)^[109] 对网络进行预训练，获得初始的识别能力。然后，网络在无标签的 NIR-VIS 人脸数据集上进行微调。在微调阶段，训练数据包括 NIR 数据 $X^N = \{x_1^N, x_2^N, \dots, x_{n^N}^N\}$ 和 VIS 数据 $X^V = \{x_1^V, x_2^V, \dots, x_{n^V}^V\}$ 。 n^N 和 n^V 分别表示 NIR 和 VIS 数据的数量。值得注意的是，所有身份标签都被去除。网络的优化在同质学习和异质学习之间交替进行。

5.2.2 模态-对抗对比学习

如图5-2 (a) 所示, 在大多数情况下, 由于模态差异, 初始估计的伪标签通常是不准确的。然而, 可以观察到模态内聚类错误明显减少。仅依赖模态内聚类结果会导致两种模态之间的身份关系未知, 不利于模态差异的消除。为了利用模态内共享的相似语义信息并抑制模态差异, 本文在同质学习阶段中提出了模态-对抗对比学习 MACL。MACL 结合了对比学习和对抗学习的优点, 调整两种模态的特征分布并学习它们的模态内紧凑表示。MACL 的流程如图5-5所示。

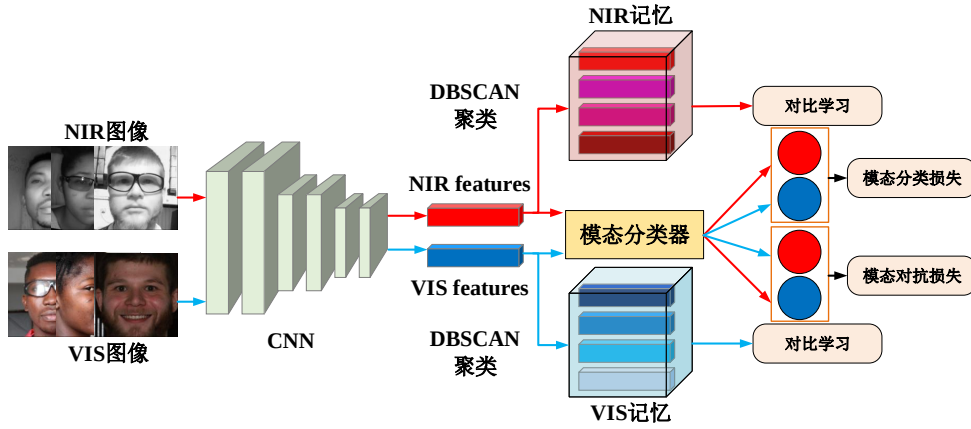


图 5-5 模态-对抗对比学习的示意图

首先, 为每个模态特征估计伪标签:

$$\begin{aligned}\hat{y}^N &= \text{DBSCAN}(f^N) \\ \hat{y}^V &= \text{DBSCAN}(f^V)\end{aligned}\tag{5-1}$$

其中, $f^{N(V)} = f(x^{N(V)}; \Theta)$ 表示提取的特征, $f(\cdot; \Theta)$ 表示特征提取器, Θ 是其参数, **DBSCAN** 表示 DBSCAN 聚类算法^[150]。 $\hat{y}_i^N \in \{1, 2, \dots, I^N\}$ 和 $\hat{y}_i^V \in \{1, 2, \dots, I^V\}$ 分别是 NIR 和 VIS 数据的伪标签, 其中 I^N 表示 NIR 模态中估计的集群数, I^V 表示 VIS 模态中估计的集群数。

为了在优化过程中保持在先前时期学到的独特集群特征表示, 在每个模态初始化一个模态特定的记忆 (Modality-specific memory), 在每个训练时期之前存储集群的特征:

$$\phi_k^N = \frac{1}{n_k^N} \sum_j \mathbf{1}\{\hat{y}_j^N = k\} f_j^N \quad (k = 1, 2, \dots, I^N)\tag{5-2}$$

$$\phi_k^V = \frac{1}{n_k^V} \sum_j \mathbf{1}\{\hat{y}_j^V = k\} f_j^V \quad (k = 1, 2, \dots, I^V) \quad (5-3)$$

其中, $\phi_k^{N(V)}$ 表示第 k 个集群的表示, $n_k^{N(V)}$ 表示 NIR 或 VIS 模态中第 k 个集群中的实例数, $\mathbf{1}\{\cdot\}$ 是指示函数。

采用对比损失^[144]来学习每个模态的紧凑表示, 具体表达如下:

$$L_N = - \sum_{i=1}^{P \times Z} \log \frac{\exp(q_i^N \cdot \phi_+^N / \tau)}{\sum_{k=0}^{I^N} \exp(q_i^N \cdot \phi_k^N / \tau)} \quad (5-4)$$

$$L_V = - \sum_{i=1}^{P \times Z} \log \frac{\exp(q_i^V \cdot \phi_+^V / \tau)}{\sum_{k=0}^{I^V} \exp(q_i^V \cdot \phi_k^V / \tau)} \quad (5-5)$$

其中, q_i^N 和 q_i^V 分别表示在训练批量中的 NIR 或 VIS 实例特征, P 和 Z 分别表示选定的身份数量和每个身份的实例特征数量。 $\phi_+^{N(V)}$ 表示查询实例 $q_i^{N(V)}$ 的集群表示。 τ 是温度超参数, 用于调节损失对于负样本关注程度。引入 τ 增强了对难负样本的惩罚, 增大正样本和负样本对的概率之间的差异, 并促进学习紧凑的嵌入空间^[168,169]。

在每个批次训练之后, 采用动量更新策略来更新模态特定的记忆:

$$\begin{aligned} \phi_+^N &\leftarrow \lambda \phi_+^N + (1 - \lambda) q_i^N \quad (i = 1, 2, \dots, P \times Z) \\ \phi_+^V &\leftarrow \lambda \phi_+^V + (1 - \lambda) q_i^V \quad (i = 1, 2, \dots, P \times Z) \end{aligned} \quad (5-6)$$

其中, λ 表示动量更新因子。

上述对比损失只能学习紧凑的模态内分布, 无法有效减小跨模态差异。对抗学习策略可以调整跨模态分布, 有助于进一步提升模型性能。因此, 它已成功应用于一些方法, 例如 DADRD^[98], DAN^[170] MCLNet^[118]。受此启发下, 对抗学习被引入在无监督学习框架中, 减小两个模态之间的分布差异。由于提取的 NIR-VIS 特征携带有模态信息, 它们呈现出不同的特征分布。对抗学习的目标是优化主干网络的参数, 使其在忽略模态信息的情况下调整 NIR-VIS 数据的特征分布。为了实现这一目标, 首先引入一个模态分类

器，用于对提取的特征进行分类，确定它们属于哪个模态。模态分类损失可以表示为：

$$L_{mc}(\theta_m) = -\frac{1}{n^V} \sum_{i=1}^{n^V} t_i^V \cdot \log p_m(f_i^V; \theta, \theta_m) - \frac{1}{n^N} \sum_{i=1}^{n^N} t_i^N \cdot \log p_m(f_i^N; \theta, \theta_m) \quad (5-7)$$

其中，定义 t_i^N 表示数据属于 NIR 模态， t_i^V 表示数据属于 VIS 模态。 θ_m 表示模态分类器的参数， $p_m(f_i^N; \theta, \theta_m)$ 和 $p_m(f_i^V; \theta, \theta_m)$ 分别表示 NIR 或 VIS 样本的模态分类概率。特征提取器的目标与模态分类器相反，因为，它的目标是忽略从提取的特征中获取的模态信息，使得模态分类器难以区分它们。特征提取器的模态对抗损失可以表示为：

$$L_{ma}(\theta) = -\frac{1}{n^V} \sum_{i=1}^{n^V} t_i^N \cdot \log p_m(f_i^V; \theta, \theta_m) - \frac{1}{n^N} \sum_{i=1}^{n^N} t_i^V \cdot \log p_m(f_i^N; \theta, \theta_m) \quad (5-8)$$

其中， $L_{mc}(\theta_m)$ 和 $L_{ma}(\theta)$ 交替更新参数 θ_m 和 θ ，在对抗过程中减小两个模态之间的分布差异。

最后，MACL 的损失函数包括四个部分：

$$L_{MACL} = \alpha_1 L_N + \alpha_2 L_V + \alpha_3 L_{mc} + \alpha_4 L_{ma} \quad (5-9)$$

其中， α_1 、 α_2 、 α_3 、 α_4 分别表示四个损失的权重。通过最小化 L_{MACL} ，MACL 在模态对比学习和对抗学习的协同作用下，有效减小两个模态之间的分布差异，并学习了每个模态内的判别性表示。通过这种方式，在跨模态关系不明确的情况下，也可以取得明显的性能改善。

5.2.3 基于 K-相互编码的跨模态标记

为无标签的数据集分配伪标签并以类似有监督的方式训练网络是无监督学习的主要策略^[144,162]。然而，由于不同模态之间存在巨大的模态差异，分配正确伪标签概率不

高，这可能对模型优化产生不利影响。为解决这个问题，在异质学习阶段，设计了计算相互最近邻对跨模态数据进行编码的方法，该方法为每个实例特征融合了相互邻域的信息生成编码向量。然后，利用这些编码向量作为聚类的输入，为跨模态数据生成高质量的伪标签，有利于解决由模态和剩余差异导致的伪标签不准确的问题。

首先，将训练实例的特征收集到一个集合中 $[f_1^V, f_2^V, \dots, f_{n_V}^V, f_1^N, f_2^N, \dots, f_{n_N}^N]$ 。为避免混淆，这里统一将 NIR 和 VIS 特征表示为 f_i ， n 是总的 NIR 和 VIS 样本数量。然后，将每个特征作为探针特征，与其他特征一起计算 K-相互编码向量^[171]：

$$\mathcal{K}_i = [D_{i,1}, D_{i,2}, \dots, D_{i,n}] \quad (5-10)$$

$$D_{i,j} = \begin{cases} \exp(-\mathcal{M}(f_i, f_j)) & \text{if } f_j \in R(f_i, k_1) \\ 0 & \text{otherwise.} \end{cases} \quad (5-11)$$

其中， $\mathcal{M}(f_i, f_j)$ 表示 f_i 和 f_j 之间的马氏距离， $R(f_i, k_1)$ 表示 f_i 的 K-相互最近邻， k_1 是调整最近邻数量的超参数。

基于如下事实，上述的 K-相互编码向量不能关联跨模态的同一身份的实例：

(1) 由于模态差异的存在，VIS 实例的 K 最近邻主要由相同模态的 VIS 实例组成，只有很小一部分是 NIR 实例。类似地，NIR 实例的 K 最近邻主要由相同模态的 NIR 实例组成，只有很小一部分是 VIS 实例。对于跨模态人脸识别任务，应更注重跨模态关联的身份标签，促进跨模态表征学习。

(2) 实例之间的 K-相互最近邻关系本身不足以关联具有相同身份的实例，特别是当它们之间存在较大的剩余差异时。

为解决这些问题，探针实例的 K-相互最近邻编码应该通过其可靠的 K-相互最近邻 $f_j \in R(f_i, k_1)$ 中的 k_2 个样本的 K-相互最近邻编码来扩展，可以表示为：

$$\tilde{\mathcal{K}}_i = \frac{1}{k_2} \sum_{j=1}^{k_2} \mathcal{K}_j \quad f_j \in R(f_i, k_1) \quad (5-12)$$

其中， k_2 的值设置为小于 k_1 。扩展后的 $\tilde{\mathcal{K}}_i$ 连接了拓展前 $\mathcal{K}_i$ 中不是 f_i 的 K-相互最近邻的相同身份跨模态实例，使其成为 K-相互编码的更为可靠的表示。随后，计算探针实

例与其他实例之间的 Jaccard 距离：

$$d_J(f_i, f_j) = 1 - \frac{\sum_{\pi=1}^n \min(D_{i,\pi}, D_{j,\pi})}{\sum_{\pi=1}^n \max(D_{i,\pi}, D_{j,\pi})} \quad (5-13)$$

$$D_J(f_i) = [d_J(f_i, f_1), d_J(f_i, f_2), \dots, d_J(f_i, f_n)] \quad (5-14)$$

其中, $d_J(f_i, f_j)$ 表示样本特征 f_i 和 f_j 之间的 Jaccard 距离, D_J 代表 f_i 的距离向量。 $\min$ 和 $\max$ 对两个输入向量进行逐元素的最小和最大筛选操作。

然后, 对距离矩阵 $[D_J(f_1), D_J(f_2), \dots, D_J(f_n)]^\top$ 使用 DBSCAN 聚类算法来分配伪标签 $\hat{y} = \mathbf{DBSCAN}(D_J)$ 。

值得注意的是, k_2 的选择极大地影响聚类效果, 本文将在实验部分展示不同 k_2 设置对结果的影响。

5.2.4 精细跨模态对比学习

在本小节, 通过使用 KECL 提供的跨模态标签来细化对比学习, 学习模态不变的代表, 从而消除模态差异。

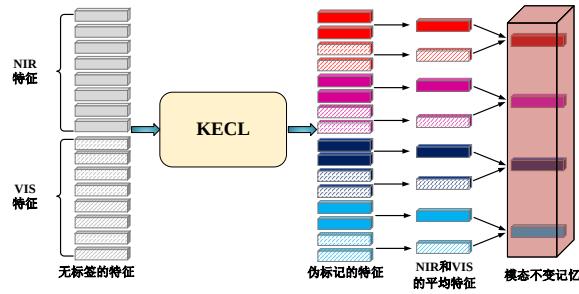


图 5-6 伪标签估计和模态不变记忆初始化的示意图。首先, 基于 K 互补编码的跨模态标签生成 (KECL) 为 NIR 和 VIS 特征提供伪标签。接下来, 采用模态不变平均初始化策略来计算 NIR 和 VIS 特征, 形成模态聚类表示

首先, 如图 5-6 所示, 提出了模态不变的平均初始化 (Modality-invariant average initialization) 策略来构建一个模态不变的记忆 (Modality-invariant memory), 用于存储每个集群的模态不变表示。在大规模的 NIR-VIS 人脸数据集中, 如 CASIA NIR-VIS 2.0^[12] 和 LAMP-HQ NIR-VIS^[14], NIR 样本的数量通常是 VIS 样本数量的数倍。因此, 在表示

每个集群的表征时，简单地对实例特征进行平均可能会在学习目标中引入模态偏差。因此，引入了模态不变均值初始化策略，具体表示如下：

$$\phi_k = \frac{1}{2} \left\{ \frac{1}{n_k^N} \sum_j \mathbf{1}\{\hat{y}_j = k\} f_j^N + \frac{1}{n_k^V} \sum_j \mathbf{1}\{\hat{y}_j = k\} f_j^V \right\} \quad (k = 1, 2, \dots, I) \quad (5-15)$$

其中， ϕ_k 表示模态不变记忆中的第 k 个集群的表示， n_k^N 表示由 KECL 估计的第 k 个集群中的 NIR 样本数量， n_k^V 表示第 k 个集群中的 VIS 样本数量。

在上一章中，跨模态域集群对比学习^[144] 可以表达为：

$$L_c = - \sum_{i=1}^{P \times Z} \log \frac{\exp(q_i \cdot \phi_+ / \tau)}{\sum_{k=0}^I \exp(q_i \cdot \phi_k / \tau)} \quad (5-16)$$

其中，查询 q 包含 $P \times \frac{Z}{2}$ 个 NIR 实例和 $P \times \frac{Z}{2}$ 个 VIS 实例， ϕ_+ 表示它们的所属集群表示。为了增强对比损失区分难易样本的能力，对公式 5-16 的改进，引入了动态关注加权：

$$L_{RCL} = - \sum_{i=1}^{P \times Z} \omega(\beta) \log \frac{\exp(q_i \cdot \phi_+ / \tau)}{\sum_{k=0}^I \exp(q_i \cdot \phi_k / \tau)} \quad (5-17)$$

$$\omega(\beta) = \left(1 - \frac{\exp(q_i \cdot \phi_+ / \tau)}{\sum_{k=0}^I \exp(q_i \cdot \phi_k / \tau)} \right)^\beta \quad (5-18)$$

当 q_i 是一个简单样本时，动态关注因子会接近零，因为 q_i 和其所属集群表示 ϕ_+ 非常相似。因此，对于这样的样本，精细跨模态对比学习 RCL 损失不需要过多优化，而是分配一个较小的损失。相反，对于难样本，分配的损失应该增加。

在每个小批量训练之后，使用动量更新策略更新模态不变记忆：

$$\phi_+ \leftarrow \lambda \phi_+ + (1 - \lambda) q_i \quad (i = 1, 2, \dots, P \times Z) \quad (5-19)$$

通过最小化损失 L_{RCL} ，网络可以学习到提取模态不变的全局特征的能力。

5.2.5 剩余不变表示增强

如图 5-1 所示，剩余信息的存在通常导致类内的距离比类间的距离更大。事实上，显著的面部变化主要体现在两个主要部分：眼部周围的上半部分（包括眼睛运动、佩戴眼镜等）和嘴巴周围的下半部分（包括嘴巴张合、表情等）。为了解决这个问题，一种直观的方法是将特征图分为两个部分：上半部分和下半部分。然后可以分别对同一人的上半部分和下半部分进行对齐，使它们更加接近。全局和局部特征可以表达为：

$$\begin{aligned} f_i &= Fc(Dropout(Pool(F_i))) \\ f_i^u &= Fc(Dropout(Pool(F_i^u))) \\ f_i^l &= Fc(Dropout(Pool(F_i^l))) \end{aligned} \quad (5-20)$$

其中， $F_i \in \mathbb{R}^{H \times W \times C}$ 表示全局特征图， $F_i^u \in \mathbb{R}^{\frac{H}{2} \times W \times C}$ 和 $F_i^l \in \mathbb{R}^{\frac{H}{2} \times W \times C}$ 分别表示上半部分和下半部分的特征图。 f_i^u 和 f_i^l 分别代表上半部分和下半部分的局部特征。 $Pool$ 表示全局平均池化层， $Dropout$ 代表 dropout 层， Fc 表示全连接层。剩余不变表示增强 RRE 的损失函数可以表达为：

$$\begin{aligned} L_{RRE} = & - \sum_{i=1}^{P \times Z} \omega(\beta) \log \frac{\exp(q_i^u \cdot \phi_+^u / \tau)}{\sum_{k=0}^I \exp(q_i^u \cdot \phi_k^u / \tau)} \\ & - \sum_{i=1}^{P \times Z} \omega(\beta) \log \frac{\exp(q_i^l \cdot \phi_+^l / \tau)}{\sum_{k=0}^I \exp(q_i^l \cdot \phi_k^l / \tau)} \end{aligned} \quad (5-21)$$

其中， q_i^u 和 q_i^l 分别表示上半部分和下半部分特征的探针， ϕ_+^u 以及 ϕ_+^l 分别表示它们的所属的集群表示。

对于上半部分和下半部分特征的记忆，其初始化和更新过程均类似地参考公式 5-15 和 5-20。在测试阶段，将全局特征和两个局部特征进行拼接来计算余弦相似度，以实现鲁棒的跨模态人脸匹配。

算法 5-1: 同质到异质学习和剩余不变增强 HERE 方法

Input: 无标签 NIR 数据 X^N , 无标签 VIS 数据 X^V , epochs 数量 $epoch$, 迭代次数 $iter$.

Output: 主干网络参数 θ .

```

1: for  $\kappa = 1, 2, \dots, epoch$  do
2:   生成模态特定伪标签  $\{\hat{y}^N, \hat{y}^V\}$  通过式子 (5-1);
3:   生成跨模态统一伪标签  $\{\hat{y}\}$  通过式子 (5-10)-(5-14);
4:   初始化记忆  $\{\phi^N, \phi^V, \phi, \phi^u, \phi^l\}$  通过式子 (5-2), (5-3), (5-15);
5:   for  $\pi = 1, 2, \dots, iter$  do
6:     更新  $\theta_m$  通过最小化式子 (5-7);
7:     更新  $\theta$  通过最小化式子 (5-4), (5-5), (5-8);
8:     更新记忆  $\{\phi^N, \phi^V\}$  通过式子 (5-6);
9:     更新  $\theta$  通过最小化式子 (5-17), (5-18);
10:    更新记忆  $\{\phi\}$  通过式子 (5-19);
11:    更新  $\theta$  通过最小化式子 (5-21);
12:    更新记忆  $\{\phi^u, \phi^l\}$  通过式子 (5-19);
13:   end for
14: end for
15: Return  $\theta$ 
    
```

5.2.6 模型优化

所提出的 HERE 网络将 MACL、KECL、RCL、RRE 整合到一个联合框架中，以学习 NIR-VIS-UHFR 的模态不变和剩余不变特征。HERE 的总目标函数可以表达为：

$$\mathcal{L} = \varphi_1 L_{MACL} + \varphi_2 L_{RCL} + \varphi_3 L_{RRE} \quad (5-22)$$

其中， φ_1 、 φ_2 、 φ_3 分别是每个损失函数的权重。HERE 方法的优化过程总结在算法 5-1 中，其中 $epoch$ 表示训练周期数， κ 表示周期索引， $iter$ 表示训练迭代次数。

5.3 实验结果与分析

在本节，验证了同质到异质学习和剩余不变增强 HERE 方法的有效性，并在四个具有挑战性的 NIR-VIS 人脸数据集上提供了丰富的实验结果，包括 CASIA NIR-VIS 2.0^[12]、BUAA NIR-VIS^[112]、Oulu-CASIA NIR-VIS^[111]、LAMP-HQ NIR-VIS^[14]。首先，将 HERE 与最近发表的主流有监督方法进行比较，以展示其有竞争性的性能。此外，还围绕着每个模块进行了深入分析，以展示每个模块的积极影响。

表 5-1 LWnet 的网络结构

名称	数量	输出尺寸
输入图片	64	$160 \times 160 \times 3$
Stem 模块	1	$17 \times 17 \times 256$
Inception-ResNet-A 模块	12	$17 \times 17 \times 256$
Reduction-A 模块	1	$8 \times 8 \times 896$
Inception-ResNet-B 模块	3	$8 \times 8 \times 896$
全局平均池化	1	$1 \times 1 \times 896$
全连接层	1	256

5.3.1 实验数据集

CASIA NIR-VIS 2.0 dataset^[12] 是一个经常被用于评估方法性能的数据集。它包含模拟不受控环境中拍摄的 NIR-VIS 人脸图像，其中大的模态和剩余变化使得来自同一人的图像差异很大。本文采用了文献^[93,98]中使用的评估设置，包括 10 次协议，每个协议中约有 8,600 张训练图像。但在无监督设置中，训练数据的标签被丢弃。在测试阶段，跨模态检索涉及来自 358 个身份的 6,000 张 NIR 探针图像和 358 张 VIS 查询图像。重要的是，在训练集和测试集之间没有重叠的身份样本。

BUAA NIR-VIS dataset^[112] 是一个经常被采用的 NIR-VIS 人脸数据集。参考文献^[73,93]中的评估设置，训练集包含来自 50 个身份的约 1,200 张图像，测试集包含来自 100 个身份的 1,300 张图像。在训练集中，本方法舍弃所有训练标签。在测试阶段，查询集中每个身份只有一张 VIS 图像，而在探针集中有多张 NIR 图像。

Oulu-CASIA NIR-VIS dataset^[111] 是一个包含 80 个身份的小规模数据集。本方法参考文献^[93]中的协议，将 80 个身份平均分为训练集和测试集。每个身份的 NIR-VIS 图像数量均衡，各有 48 张 NIR 和 48 张 VIS 图像。特别地，在 NIR-VIS 图像中有各种表情，如愤怒、厌恶、快乐、悲伤等。在训练阶段，本方法不需要身份标签，并分别通过 DBSCAN 和 KECL 估算 NIR-VIS 数据的模态特定和跨模态伪标签。

LAMP-HQ NIR-VIS dataset^[14] 是具有高质量和多姿态 NIR 和 VIS 图像的 NIR-VIS 数据集，其中包含来自 573 个受试者的 16,828 张 VIS 图像和 56,788 张 NIR 图像。每个身份的 NIR 和 VIS 图像数量不平衡，NIR 图像比 VIS 图像多几倍。本文按照 LAMP-HQ 数据集提供的 10 个官方协议验证了 HERE 方法的有效性。在每个协议中，训练集包含 300 个身份，而测试集包含 273 个身份。

表 5-2 CASIA NIR-VIS 2.0 数据集上的消融实验

BL	MCL	MACL	方法			CASIA NIR-VIS 2.0		
			D-RCL	K-RCL	RRE	Rank-1	VR@FAR=0.1%	VR@FAR=0.01%
✓						93.5	86.0	62.0
	✓					98.4	96.2	89.2
		✓				99.1	98.4	93.8
			✓			99.0	98.5	93.6
			✓		✓	99.2	98.6	94.2
		✓		✓	✓	99.8	99.7	98.7

表 5-3 Oulu-CASIA NIR-VIS 数据集上的消融实验

BL	MCL	MACL	方法			Oulu-CASIA NIR-VIS		
			D-RCL	K-RCL	RRE	VR@FAR=1%	VR@FAR=0.1%	VR@FAR=0.01%
✓						91.7	78.5	68.2
	✓					97.2	87.4	75.1
		✓				98.2	88.7	80.0
			✓			97.2	90.4	82.2
			✓		✓	98.0	91.0	83.5
		✓		✓	✓	99.1	94.6	90.7

5.3.2 实验设置

HERE 网络构建是在基于 tensorflow 实现的，GPU 使用了单卡 TITAN Xp。轻量化特征提取器 LWnet 作为主干网络。LWnet 仅包含 11 亿次乘加运算和 550 万个参数。LWnet 的网络结构显示在表5-1中，包括一个 Stem 块，12 个 Inception-ResNet-A 块，一个 Reduction-A 块，3 个 Inception-ResNet-B 块，一个全局平均池化层、一个全连接层^[30,136]。

首先，在 MS-Celeb-1M 数据集^[109] 上对 LWnet 进行预训练。图像数据由 MTCNN 网络^[124] 裁剪到 160×160 。预训练的 LWnet 在 LFW 数据集^[127] 上达到了 99.3% 的准确率。其次，使用 NIR-VIS 人脸数据集对网络进行进一步的微调。具体而言，与有监督学习方法不同，训练数据的所有身份标签都不可用。每个 epoch 的训练迭代次数 $iter$ 和批量大小 $P \times Z$ 设置为 200 和 64。 P 和 K 分别设置为 4 和 16。学习率设置为 0.002，每 800 个训练迭代减小 0.95 倍。超参数 α_1 、 α_2 、 α_3 、 α_4 、 φ_1 、 φ_2 、 φ_3 、 τ 、 λ 、 β 分别设置为 1.0、1.0、1.0、0.01、1.0、1.0、0.01、0.1、0.1、0.5。引入 RRE 模块后，本方法将全局特征和两个局部特征连接起来，以实现鲁棒的跨模态人脸匹配。

5.3.3 模块性能分析

在本小节，在三个主流数据集 (即 CASIA NIR-VIS 2.0、Oulu-CASIA NIR-VIS 和 LAMP-HQ NIR-VIS) 上验证了 HERE 方法中每个模块的贡献。

表 5-4 LAMP-HQ NIR-VIS 数据集上的消融实验

BL	MCL	MACL	方法			LAMP-HQ NIR-VIS		
			D-RCL	K-RCL	RRE	Rank-1	VR@FAR=0.1%	VR@FAR=0.01%
✓						93.6	66.7	44.4
	✓					97.8	95.8	84.9
		✓				98.1	96.5	89.2
			✓			98.4	96.4	91.3
			✓		✓	98.6	97.1	92.8
		✓		✓	✓	98.9	98.7	96.3

消融实验: 如表5-2、5-3、5-4所示, 在 CASIA NIR-VIS 2.0、Oulu-CASIA NIR-VIS、LAMP-HQ NIR-VIS 数据库上验证了 HERE 方法的主要模块的贡献。预训练 LWnet 在实验中作为基准方法 BL。MCL 表示没有对抗学习损失 L_{mc} 和 L_{ma} 的 MACL。K-RCL 表示 RCL 的伪标签监督由 KECL 提供, 而 D-RCL 表示直接在整个跨模态数据集上采用 DBSCAN 来提供标签监督。

首先, 三个表格的第一行表明 BL 方法获得了较差的性能, 并说明了在大规模 VIS 数据集上预训练网络无法有效解决由于显著的模态差异, 而导致性能下降。其次, 三个表格的第二行和第三行展示了对抗学习损失 L_{mc} 和 L_{ma} 的贡献。结果显示 MACL 在 CASIA NIR-VIS 2.0 上的 Rank-1 提高了 0.7%, 在 VR@FAR=0.1% 上提高了 4.6%; 在 Oulu-CASIA NIR-VIS 上的 Rank-1 提高了 1.0%, 在 VR@FAR=0.1% 上提高了 4.9%; 在 LAMP-HQ NIR-VIS 上的 Rank-1 提高了 0.3%, 在 VR@FAR=0.1% 上提高了 4.3%。这表明, 即使 VIS 和 NIR 模态之间的身份关系不明确, MACL 仍然能够有效减小模态差异。接着, 三个表格的第四行和第五行展示了 RRE 的有效性, 反映在 CASIA NIR-VIS 2.0 上的 Rank-1 提高了 0.2%, 在 VR@FAR=0.1% 上提高了 0.6%; 在 Oulu-CASIA NIR-VIS 上的 Rank-1 提高了 0.8%, 在 VR@FAR=0.1% 上提高了 1.3%; 在 LAMP-HQ NIR-VIS 上的 Rank-1 提高了 0.2%, 在 VR@FAR=0.1% 上提高了 1.5%。最后, 三个表格的第六行方法通过将 MACL、KECL、RCL、RRE 整合到统一的框架中, 优化了模型性能, 并展示了 KECL 的有效性。

DBSCAN 在 MACL 中的有效性: 图5-7展示了 DBSCAN 在 MACL 中生成的伪标签的质量。可以发现, 绝大多数未标记的样本得到了有效利用, 而在两种模态中离群点 (Outliers) 的数量少于 100。在 HERE 方法中, 有效的目标函数使得正确伪标签的数量趋于与真实的标签数量。这些结果表明, DBSCAN 能够分别为 NIR 和 VIS 数据分配准确的伪标签。

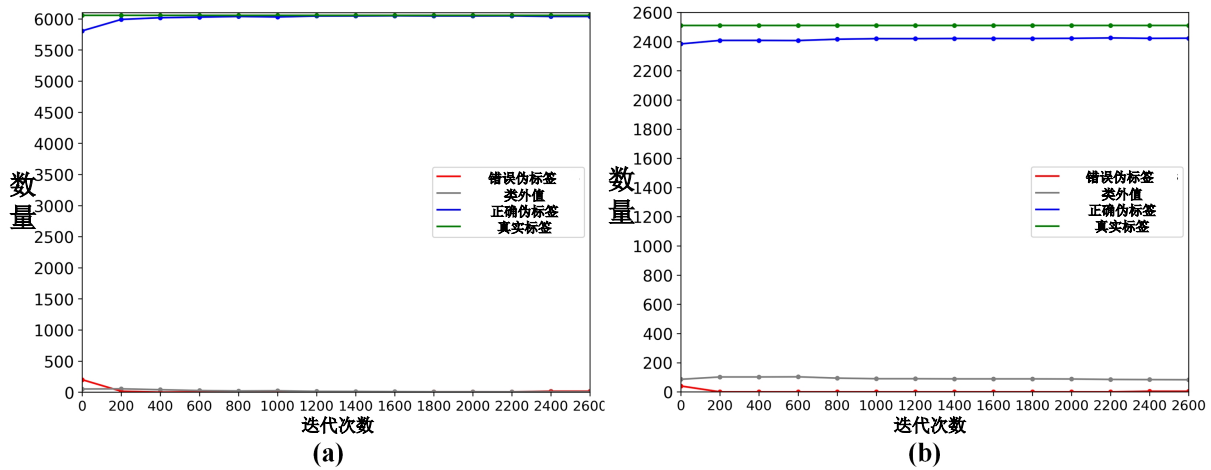


图 5-7 DBSCAN 在 MACL 中生成的伪标签质量。未标记的数据来自 CASIA NIR-VIS 2.0 数据集。图 (a) 展示了 NIR 数据的伪标签质量，而图 (b) 展示了 VIS 数据的伪标签质量

表 5-5 使用 CASIA NIR-VIS 2.0 和 LAMP-HQ NIR-VIS 数据集对 HERE 中的对抗学习方法消融研究。w/o adv 表示该方法没有引入对抗学习

方法	CASIA NIR-VIS 2.0		LAMP-HQ NIR-VIS	
	Rank-1	VR@FAR=0.01%	VR@FAR=0.1%	VR@FAR=0.01%
MACL (w/o adv)	98.4	89.2	95.8	84.9
MACL	99.1	93.8	96.5	89.2
HERE (w/o adv)	99.6	98.4	98.4	95.6
HERE	99.8	98.7	98.7	96.3

对抗学习的有效性：为验证对抗学习在 HERE 框架中的贡献，在两个数据集上进行了消融研究。结果显示在表5-5中。当将 MACL 叠加到基准方法上时，对抗学习可以在 VR@FAR=0.01% 上提高 4%-5% 的性能。相比 HERE (w/o adv)，对抗学习给 HERE 带来了 0.2%-0.7% 的性能提升。这些结果表明对抗学习可以缩小两种模态的分布，从而促进性能的提升。

RRE 的有效性：为了更深入地了解 RRE 的有效性，本文展示了特征相似性的可视化结果。图5-8展示了使用全局特征导致错误匹配的结果。引入 RRE 后，错误示例的全局特征和局部特征之间的相似性降低，而正确示例之间的相似性增加。有趣的是，一些剩余信息导致的错误匹配，这对于人眼来说都是难以区分的，但本章提出的方法仍然可以正确匹配。这归因于 RRE 强大的整合全局和局部线索的能力，有助于在结构上有效区分正样本和负样本，从而减少由剩余信息导致的错误匹配。特别是当拼接局部和全局特征时，正样本间的配对变得更有优势，从而得到更加鲁棒的匹配结果。

模态不变的平均初始化和动态关注加权策略有效性：表5-6展示了模态不变的平均初始化和动态关注加权策略在 LAMP-HQ NIR-VIS 数据集中 RCL 中的贡献。可以观

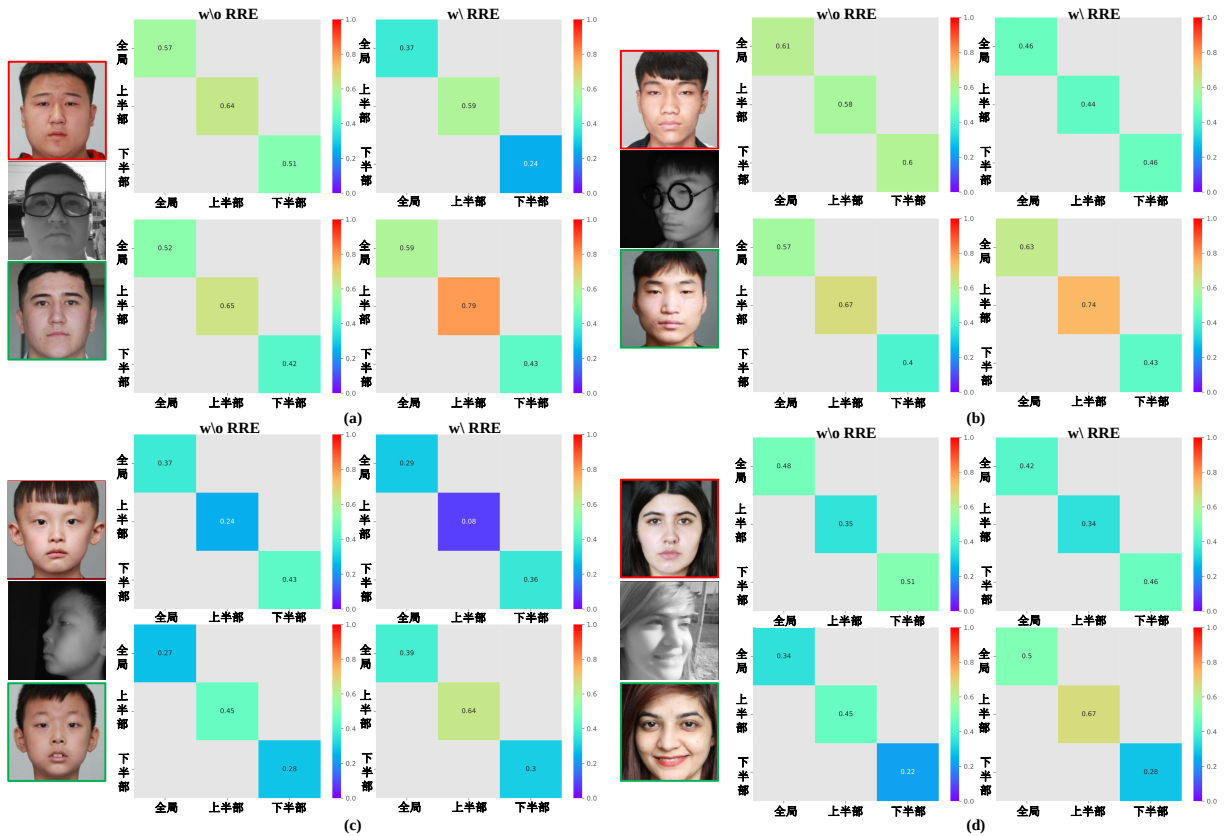


图 5-8 在没有 RRE (w/o RRE) 的消融版本和有 RRE (w/ RRE) 的情况下，来自 LAMP-HQ 数据集第一个协议测试集的图像对之间的全局和局部特征的余弦相似度。红框表示错误匹配的示例，绿框表示正确匹配的结果

表 5-6 模态不变的平均初始化和动态关注加权策略的消融实验

Method	VR@FAR=0.01%
HERE w/o (模态不变的平均初始化)	95.1
HERE w/o (动态关注加权策略)	95.7
HERE	96.3

察到 HERE 在 VR@FAR=0.1% 上的性能比没有动态关注加权策略的 HERE 好，提高了 1.2%，而在 VR@FAR=0.1% 上的表现也优于没有模态不变的平均初始化的 HERE，提高了 0.6%。

对 KECL 设置的分析：图5-9展示了由预训练模型估计的伪标签质量，其中伪标签是在不同设置下由 KECL 估计的。随着 k_2 的增加，特征的 K-相互编码通过合并其 k_2 最近邻的附加编码信息而扩展。这有助于链接更多的跨模态实例，并且减轻剩余信息的污染。当 k_2 等于 25 时，伪标签估计获得最佳的准确性。然而，当 k_2 进一步增加时，会引入了噪声实例，导致错误匹配增加。当不扩展 k-相互编码时，由于模态差异和剩余差异，每个特征的 k-相互编码相对孤立。无法找到与相同身份特征中的共性，许多实例被

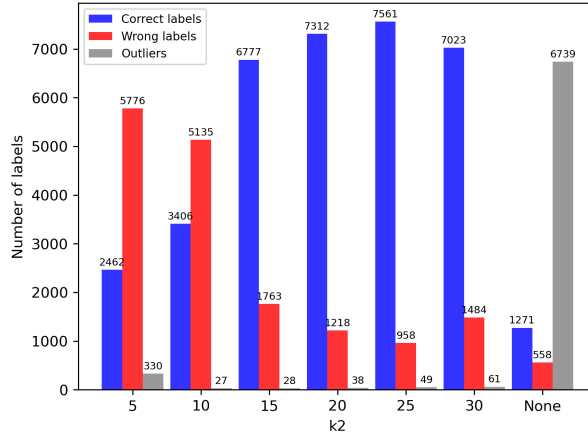


图 5-9 在 CASIA NIR-VIS 2.0 上, KECL 使用预训练网络的伪标签估计的结果。 k_1 的默认设置为 30, 'None' 表示实例的 K-相互编码将不会进行扩展

归类为离群点。

表 5-7 不同方法在 CASIA NIR-VIS 2.0 数据库上的结果

种类	方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
有监督	TRIVET ^[92] (ICB 2016)	95.7 ± 0.5	98.1 ± 0.3	91.0 ± 1.3
	IDR ^[128] (AAAI 2017)	97.3 ± 0.4	98.9 ± 0.3	95.7 ± 0.7
	LightCNN-29 ^[129] (TIFS 2018)	96.7 ± 0.2	98.5 ± 0.6	94.8 ± 0.4
	ADFL ^[76] (AAAI 2018)	98.1 ± 0.4	99.1 ± 0.2	97.1 ± 0.5
	CDL ^[130] (AAAI 2018)	98.6 ± 0.2	99.5 ± 0.1	98.3 ± 0.1
	WCNN ^[93] (TPAMI 2018)	98.7 ± 0.3	99.5 ± 0.1	98.4 ± 0.4
	PCFH ^[70] (IJCAI 2019)	98.8 ± 0.3	99.6 ± 0.1	97.7 ± 0.3
	PACH ^[71] (CVPR 2020)	98.9 ± 0.2	99.6 ± 0.1	98.3 ± 0.2
	CFC ^[27] (TPAMI 2020)	98.6 ± 0.1	99.2 ± 0.1	97.3 ± 0.2
	SDDL ^[106] (PR 2021)	92.1 ± 0.7	-	88.7 ± 1.0
	ADCANs ^[131] (TMM 2021)	99.1 ± 0.2	99.6 ± 0.1	98.5 ± 0.2
	VSANet ^[14] (IJCV 2021)	99.2 ± 0.1	99.7 ± 0.0	98.2 ± 0.2
	EGCH ^[82] (ICPR 2021)	98.8 ± 0.2	99.6 ± 0.0	97.7 ± 0.2
	MMTN ^[72] (TIFS 2022)	99.4 ± 0.0	99.8 ± 0.0	98.7 ± 0.0
	SAFE ^[74] (TPAMI 2022)	98.9 ± 0.1	99.6 ± 0.1	98.6 ± 0.1
无监督	HERE (本章方法)	99.8 ± 0.1	99.9 ± 0.0	99.7 ± 0.2

5.3.4 与主流的方法对比

在本小节中, 将无监督的 HERE 方法与主流的有监督人脸识别方法进行比较。

CASIA NIR-VIS 2.0 实验结果: 表5-7展示了无监督方法 HERE 在 CASIA NIR-VIS 2.0 数据集上与主流的有监督方法的性能比较。首先, HERE 方法与多个主流方法 (如 TRIVET^[92]、IDR^[128]、LightCNN-29^[129]、CDL^[130]、WCNN^[93]、PCFH^[70]、PACH^[71]) 相比表现出了性能优势。例如, 在 Rank-1 上, HERE 优于 PACH 0.9%, 在 VR@FAR=0.1% 上优于 PACH 1.4%。然后, 与最近发表的方法相比, HERE 在 Rank-1 准确性和验证率上取得了有竞争力的性能, 与主流的有监督方法性能相媲美。例如, 在 Rank-1 上, HERE

超过 MMTN 0.4%。这些结果表明无监督方法在大规模场景中具有更大的便利性和有竞争力的性能。

表 5-8 不同方法在 Oulu-CASIA NIR-VIS 数据库上的结果

种类	方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
有监督	TRIVET ^[92] (ICB 2016)	92.2	67.9	33.6
	IDR ^[128] (AAAI 2017)	94.3	73.4	46.2
	LightCNN-29 ^[129] (TIFS 2018)	96.7	92.4	65.1
	ADFL ^[76] (AAAI 2018)	95.5	83.0	60.7
	CDL ^[130] (AAAI 2018)	94.3	81.6	53.9
	WCNN ^[93] (TPAMI 2018)	98.0	81.5	54.6
	PCFH ^[70] (IJCAI 2019)	100	97.7	86.6
	PACH ^[71] (CVPR 2020)	100	97.9	88.2
	CFC ^[27] (TPAMI 2020)	99.9	98.1	90.7
	ADCANs ^[131] (TMM 2021)	99.8	93.2	78.9
	VSANet ^[14] (IJCV 2021)	100	97.7	89.0
	EGCH ^[82] (ICPR 2021)	100	98.0	89.5
	MMTN ^[72] (TIFS 2022)	100	99.2	89.0
	SAFE ^[74] (TPAMI 2022)	100	98.3	91.8
无监督	HERE (本章方法)	100	99.1	94.6

Oulu-CASIA NIR-VIS 实验结果: 表5-8报告了在 Oulu-CASIA NIR-VIS 数据集上主流的有监督方法和无监督 HERE 方法的结果。HERE 达到了和有监督方法相当的性能,并在关键指标 VR@FAR=1% 和 VR@FAR=0.1% 上分别为 99.1% 和 94.6% 的情况下略优于其他方法。结果表明 NIR-VIS-UHFR 方法可以减轻对手工标签的依赖,并在无监督下取得满意的性能。

表 5-9 不同方法在 BUAA NIR-VIS 数据库上的结果

种类	方法	Rank-1 (%)	VR@FAR = 1%(%)	VR@FAR = 0.1%(%)
有监督	TRIVET ^[92] (ICB 2016)	93.9	93.0	80.9
	IDR ^[128] (AAAI 2017)	94.3	93.4	84.7
	LightCNN-29 ^[129] (TIFS 2018)	96.5	95.4	86.7
	ADFL ^[76] (AAAI 2018)	95.2	95.3	88.0
	CDL ^[130] (AAAI 2018)	96.9	95.9	90.1
	WCNN ^[93] (TPAMI 2018)	97.4	96.0	91.9
	PCFH ^[70] (IJCAI 2019)	98.4	97.9	92.4
	PACH ^[71] (CVPR 2020)	98.6	98.0	93.5
	CFC ^[27] (TPAMI 2020)	99.7	98.7	97.8
	ADCANs ^[131] (TMM 2021)	99.8	99.7	98.4
	VSANet ^[14] (IJCV 2021)	98.8	98.3	93.4
	EGCH ^[82] (ICPR 2021)	98.1	92.7	
	MMTN ^[72] (TIFS 2022)	99.2	99.5	97.2
	SAFE ^[74] (TPAMI 2022)	99.4	99.0	98.0
无监督	HERE (本章方法)	100	100	99.9

BUAA NIR-VIS 实验结果: 无监督 HERE 方法与主流的有监督方法进行比较,结果显示在表5-9中。显然,HERE 的性能优于 TRIVET^[92]、IDR^[128]、LightCNN-29^[129]、CDL^[130]、WCNN^[93]、PCFH^[70]、PACH^[71] 等方法。与 PACH 相比,HERE 模型在 Rank-1 准确性上取得了明显的提高,从 98.6% 提高到 100%,而在 VR@FAR=0.1% 上从 93.5%

提高到 99.9%。与最近发表的有监督方法相比，HERE 也取得了改进，在 $VR@FAR=1\%$ 上超过 ADCANs 0.3%， $VR@FAR=0.1\%$ 超过 ADCANs 1.5%。结果验证了 HERE 能够通过 MACL、KECL、RCL、RRE 生成高质量的伪标签，并学习模态不变的表示。

表 5-10 不同方法在 LAMP-HQ NIR-VIS 数据库上的结果

种类	方法	Rank-1 (%)	$VR@FAR = 1\%$ (%)	$VR@FAR = 0.1\%$ (%)
有监督	LightCNN-29 ^[129] (TIFS 2018)	94.6 $\pm$ 0.3	92.5 $\pm$ 0.6	75.6 $\pm$ 1.9
	ADFL ^[76] (AAAI 2018)	95.1 $\pm$ 0.5	92.1 $\pm$ 0.9	73.3 $\pm$ 2.2
	PCFH ^[70] (IJCAI 2019)	95.3 $\pm$ 0.5	92.9 $\pm$ 0.6	75.1 $\pm$ 1.8
	PACH ^[71] (CVPR 2020)	95.4 $\pm$ 0.5	93.1 $\pm$ 0.4	75.3 $\pm$ 1.7
	ADCANs ^[131] (TMM 2021)	97.8 $\pm$ 0.3	98.8 $\pm$ 0.3	96.1 $\pm$ 0.4
	VSANet ^[14] (IJCV 2021)	97.3 $\pm$ 0.2	94.8 $\pm$ 0.7	78.2 $\pm$ 3.0
	MMTN ^[72] (TIFS 2022)	97.7 $\pm$ 0.0	95.7 $\pm$ 0.6	79.2 $\pm$ 0.0
无监督	HERE (本章方法)	98.9 $\pm$ 0.3	99.4 $\pm$ 0.2	98.8 $\pm$ 0.3

LAMP-HQ NIR-VIS 实验结果: 由于 LAMP-HQ NIR-VIS 是一个新发布的数据集，大多数主流的方法没有报告它们在这个数据集上的性能。在表5-10中总结了部分模型的结果。在最具挑战性的 NIR-VIS 数据集上，HERE 方法取得了满意的结果。将 Rank-1 准确率从 97.8% 提高到 98.9%， $VR@FAR=1\%$ 从 98.8% 提高到 99.4%， $VR@FAR=0.1\%$ 从 96.1% 提高到 98.8%。即使是最具挑战性的大规模数据集，HERE 模型依然能够有效获得跨模态关系，在全局和局部两个层面消除模态和剩余差异，并在不可控的场景中实现鲁棒的跨模态人脸识别。

本节将无监督的 HERE 方法与主流的有监督方法进行比较。实验结果表明，尽管缺乏手工标注，HERE 方法在常用数据集上实现了与有监督方法相似的性能，并在某些指标上超过了最近的方法。这一成就可以归因于所设计的伪标签估计方法以及在全局和局部两个方面达到解决模态和剩余差异的学习目标。这为跨模态人脸识别任务提供了新的方案，并可以给予目前方法更大的便利性和实用性，同时降低了数据采集的要求和成本。

5.4 本章小结

本章提出了一种新颖的同质到异质学习和剩余不变增强 HERE 方法。遵循“同质到异质学习”策略，HERE 在无监督条件中学习模态不变和剩余不变的表示。在同质学习阶段，模态-对抗对比学习 MACL 充分利用对比学习和对抗学习的协同作用，缩小模态之间的分布差异，并在每个模态内学习判别性的表示。在异质学习阶段，基于 K-相互

编码的跨模态标记 KECL 被引入作为伪标签估计方法，通过探索 K-相互最近邻将跨模态特征分组成集群以分配伪标签。此外，精细跨模态对比学习 RCL 采用模态不变的平均初始化和动态关注加权策略，有效提取模态不变的全局特征。最后，考虑到剩余变化可能影响全局特征相似度的计算精度，剩余不变表示增强 RRE 在跨模态人脸中挖掘局部特征进行精细化的匹配。与有监督方法相比，HERE 不仅能够降低数据采集需求和成本以提供更大的便利性和实用性，还能够学习到媲美有监督方法的性能。

第 6 章 总结与展望

6.1 工作总结

NIR-VIS 人脸识别的主要目标是精确匹配同一身份的 NIR 人脸图像与 VIS 人脸图像。作为异构人脸识别的典型子任务，NIR-VIS 人脸识别能够克服低光照条件下传统人脸识别存在的瓶颈问题。因此，在公共安全、犯罪调查、多媒体信息检索等领域的应用中，NIR-VIS 人脸识别已逐渐成为一名关键技术。然而，在面对不断涌现的社会发展挑战，NIR-VIS 人脸识别技术仍然面临一些问题和挑战：(1) 模态差异问题：由于成像原理和相机参数设置的不同，不同模态人脸图像之间存在明显的外观差异，例如纹理细节和边缘轮廓等。(2) 剩余差异问题：异质人脸图像在图像质量上参差不齐，具体表现在姿态、遮挡、表情、照明等差异化信息。(3) 无监督学习问题：从大量原始跨模态人脸数据中获取注释信息是耗时且昂贵的，获取注释问题直接限制了数据集规模，还可能导致过拟合问题，最终降低了有监督 NIR-VIS 人脸识别方法在更大规模真实场景中的扩展性和可行性。针对上述问题，本论文展开了四项与之相关的研究，具体的贡献和创新点总结如下：

- 1) **融合空间学习方法**: 针对模态差异问题，本研究引入一个轻量化的 CNN，在保持图像结构的前提下将 NIR-VIS 图像投影至模态差异最小化的融合空间中解决模态差异问题。具体的，融合空间学习方法包含以下创新点：(1) 设计了融合模态生成器来将 NIR 和 VIS 图像合成为保留面部结构以及模态差异不明显的融合图像。(2) 模态对抗融合空间学习以融合模态为目标设计了对抗学习策略，将 NIR 和 VIS 图像投影到融合空间中。(3) 融合分布一致性中设计了三种类内约束，有效地增强了特征的紧凑性。(4) 在 CASIA NIR-VIS 2.0、Oulu-CASIA NIR-VIS、LAMP-HQ NIR-VIS 数据库上的广泛实验展示了本方法的有效性。这一项研究工作已在 ACM Transactions on Multimedia Computing, Communications, and Applications 正式发表。
- 2) **中立人脸学习和渐进融合合成方法**: 考虑到目前的方法很少对模态和剩余信息同时进行建模并解耦，本项研究工作引导网络学习同模态的中立人脸表示，从而有

效提取身份相关的鲁棒表示。此外，本方法通过异构人脸图片在图像层面和特征层面之间进行转换，以解耦和重新融合面部属性特征来提纯身份相关信息。具体而言，中立人脸学习和渐进融合合成方法的创新点可以归纳为几个方面：(1) 身份相关中立人脸学习通过引导网络学习同质中立人脸表示，消除了模态差异和剩余差异。在重建同质中立人脸的过程中，身份相关特征被映射到共同的中立表示子空间。(2) 属性渐进融合对身份相关特征、模态相关特征、剩余特征进行解耦和重新融合以进一步提纯身份相关特征。(3) 在三个具有挑战性的数据集上进行了全面评估，包括 CASIA NIR-VIS 2.0^[12]、BUAA NIR-VIS^[112]、LAMP-HQ NIR-VIS^[14]，验证了所提出方法的有效性和鲁棒性。这一研究成果已在 IEEE Transactions on Circuits and Systems for Video Technology 正式发表。

- 3) **鲁棒跨模态伪标签和对比学习方法:** 当前的 NIR-VIS 人脸识别方法取得了显著的成功，前提是身份标签可用于指导网络实现准确的身份分类。然而，对来自不同传感器的大量异构人脸图像进行身份标签注释是耗时且昂贵的。这个难题明显限制了 NIR-VIS 人脸识别系统在更大规模实际场景中的应用。针对此问题，本项研究工作提出通过无监督学习的角度来实现异构人脸识别。然而，在异构数据中学习模态不变特征还需要解决跨模态伪标签关联和抗噪声学习两大技术难题。针对上述问题，本文提出了鲁棒跨模态伪标签和对比学习方法，本方法的贡献和创新点总结如下：(1) 基于 NIR 集群的伪标签共享在 NIR 模态下探索可靠的集群，并与 VIS 模态共享可靠的标签信息，建立跨模态身份关联。(2) 模态域特定集群对比学习通过模态内进行对比学习和动态内存更新，学习模态内紧凑而有辨别力的表示。(3) 跨模态域集群对比学习通过模态不变记忆的对比学习来引导实例特征学习模态无关的表示。(4) 进行了大量实验证明了本项研究工作在伪标签分配方面达到了 99% 以上的准确性以及在四个主流 NIR-VIS 数据集上媲美有监督方法的性能。这一研究成果已在 IEEE Transactions on Image Processing 正式发表。

- 4) **同质到异质学习和剩余不变增强方法:** 针对跨模态数据的无监督学习，许多方法尝试通过利用不同模态的互补信息来提升聚类性能。但这些方法大多是为图像到文本、视频到文本和音频到视频等特定任务设计的，与 NIR-VIS 人脸识别任务有根本的不同。此外，模态差异和剩余差异的存在，跨模态人脸图像呈现出较大的

类内变化，相同身份的实例特征难以聚类到同一类别。针对上述问题，提出了同质到异质学习和剩余不变增强方法，本方法的创新点和贡献如下所述：(1) 模态-对抗对比学习充分发挥了对比学习和对抗学习的协同作用。(2) 基于 K-相互编码的跨模态标记通过探索跨模态 K-相互最近邻关系，有效解决了为 NIR-VIS 数据分配统一伪标签的挑战。(3) 精细跨模态对比学习采用模态不变的平均初始化和动态关注加权策略探索模态不变的表示。(4) 剩余不变表示增强学习并优化局部特征，并将其与全局特征结合，以提高模型对剩余变化的鲁棒性。与有监督方法相比，无监督 HERE 在多个数据集上表现出有竞争力的性能。此外，本项研究工作能够有效地降低数据采集要求和成本，具有更大的可扩展性和实用性。这一研究成果已在 IEEE Transactions on Information Forensics and Security 正式发表。

6.2 研究展望

本文针对获取 NIR-VIS 人脸的模态不变特征进行研究，并深入解决剩余差异和无监督学习等更加复杂的问题。在此基础上，提出了融合空间学习方法、中立人脸学习和渐进融合合成方法、鲁棒跨模态伪标签和对比学习方法、同质到异质学习和剩余不变增强方法多种创新方案。通过对各项研究进行对比分析和模块分析，验证了这些方法能获得满意的结果。

深度学习领域近年来涌现出一系列现象级技术，如扩散模型 (Diffusion Model)^[172] 和视觉-语言大模型 (Large Vision-Language Model)^[173] 在多个关键任务中展现了巨大的潜力。鉴于此，结合目前的研究基础，将这些先进技术引入 NIR-VIS 人脸识别领域有助于克服各种挑战。此外，在实际应用中多种模态域人脸匹配需要训练多个单独网络，这造成了极大的算力负担。因此，本文的未来研究展望，可以总结如下：

- 1) 扩散模型建模：目前在许多领域，扩散模型在合成结果方面展现出比 GANs 更为逼真的效果。这是因为扩散模型具有易于训练和调节、生成图像更加稳定、抗攻击能力、可解释性强等优点^[174]。如何设计扩散模型，使其能够学会捕捉两种模态的基本分布，生成视觉逼真且保持身份特征的人脸图像，是一个值得深入研究的主题。
- 2) 大模型学习：近两年来，大模型在自然语言处理^[175]、图像识别^[176]、语音识别^[177]

等领域取得了显著的优势。大模型具有强大的泛化能力，能够从大量数据中学习内在的规律和特点，在不同场景中展现出卓越的适应性和鲁棒性。因此，设计大模型微调方法，将大模型的泛化能力应用于跨模态人脸识别任务，是提升跨模态人脸识别性能的可行途径。

- 3) 更多模态域识别验证分析：本论文讨论的方法主要针对 NIR-VIS 人脸识别模态进行设计，并通过通用性分析验证了其在素描-VIS 人脸识别上的通用能力。然而，在实际应用中，可能需要借助多个模态的人脸信息进行匹配。因此，未来的研究可以在本论文的基础上进行改进，使得一个模型能够完成多种模态的匹配。

参考文献

- [1] Hu W, Yang Y, Hu H. Pseudo Label Association and Prototype-Based Invariant Learning for Semi-Supervised NIR-VIS Face Recognition[J]. IEEE Transactions on Image Processing, 2024, 33 : 1448 – 1463.
- [2] George A, Marcel S. From Modalities to Styles: Rethinking the Domain Gap in Heterogeneous Face Recognition[J/OL]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2024. <http://dx.doi.org/10.1109/TBIOM.2024.3365350>.
- [3] Niu X, Li E, Liu J, et al. Mind the Gap: Learning Modality-Agnostic Representations with a Cross-Modality UNet[J/OL]. IEEE Transactions on Image Processing, 2024. <http://dx.doi.org/10.1109/TIP.2023.3348656>.
- [4] Kumar V, Dhivya A, Haripriya B, et al. Advancing Heterogeneous Face Recognition Through Convolutional Neural Networks DenseNet[C] // Proceedings of the International Conference on Electronics, Communication and Aerospace Technology (ICECA). 2023 : 1165 – 1169.
- [5] Zhao W, Chellappa R, Phillips P J, et al. Face recognition: A literature survey[J]. ACM computing surveys, 2003, 35(4) : 399 – 458.
- [6] Kortli Y, Jridi M, Al Falou A, et al. Face recognition systems: A survey[J]. Sensors, 2020, 20(2) : 342 – 378.
- [7] Parkhi O, Vedaldi A, Zisserman A. Deep face recognition[C] // British Machine Vision Conference(BMVC). 2015 : 1 – 12.
- [8] Adjabi I, Ouahabi A, Benzaoui A, et al. Past, present, and future of face recognition: A review[J]. Electronics, 2020, 9(8) : 1188 – 1240.
- [9] 邹章晨. “十四五”期间人脸识别市场发展前景及趋势分析[J]. 中国安防, 2021, 4 : 60 – 64.
- [10] 沈泽. 应用场景日渐繁荣发展前景更为广阔——2021 年人脸识别市场发展现状与趋势解析 [J]. 中国安防, 2021, 3 : 60 – 63.
- [11] Ouyang S, Hospedales T, Song Y-Z, et al. A survey on heterogeneous face recognition: Sketch, infra-red, 3D and low-resolution[J]. Image and Vision Computing, 2016, 56 : 28 – 48.

- [12] Li S, Yi D, Lei Z, et al. The CASIA NIR-VIS 2.0 face database[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2013 : 348–353.
- [13] Li S, Chu R, Liao S, et al. Illumination invariant face recognition using near-infrared images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(4) : 627–639.
- [14] Yu A, Wu H, Huang H, et al. LAMP-HQ: A large-scale multi-pose high-quality database and benchmark for NIR-VIS face recognition[J]. International Journal of Computer Vision (IJCV), 2021, 129 : 1467–1483.
- [15] 孙歆钰, 陈良哲, 张洋硕. 多特征融合的近红外与可见光异质人脸识别 [J]. 激光杂志, 2022, 43 : 66–70.
- [16] Buddharaju P, Pavlidis I T, Tsiamyrtzis P, et al. Physiology-based face recognition in the thermal infrared spectrum[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(4) : 613–626.
- [17] Kong S G, Heo J, Abidi B R, et al. Recent advances in visual and infrared face recognition—a review[J]. Computer Vision and Image Understanding, 2005, 97(1) : 103–135.
- [18] Bhowmik M, Saha K, Majumder S, et al. Thermal infrared face recognition a biometric identification technique for robust security system[J]. Reviews, refinements and new ideas in face recognition, 2011, 7 : 113–138.
- [19] 胡麟苗, 张[国], 楼晨风. 基于内容特征提取的短波红外-可见光人脸识别 [J]. 光学精密工程, 2021, 29 : 160–171.
- [20] 张帅, 谢志华, 牛杰一, et al. 基于对抗判别域适应的近红外与可见光异质人脸识别 [J]. 南京师大学报 (自然科学版), 2020, 43 : 95–103.
- [21] 张典, 汪海涛, 姜瑛, et al. 基于轻量网络的近红外光和可见光融合的异质人脸识别 [J]. 小型微型计算机系统, 2020, 41 : 807–811.
- [22] 史秋阳, 李浩, 石凤蕊, et al. 基于近红外结构光与可见光的三维人脸识别算法 [J]. 无线互联科技, 2019, 16 : 124–125.
- [23] Anghelone D, Chen C, Ross A, et al. Beyond the visible: A survey on cross-spectral face recognition[J]. arXiv preprint arXiv:2201.04435, 2022.

- [24] Chen C, Ye M, Jiang D. Towards modality-agnostic person re-identification with descriptive query[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2023 : 15128 – 15137.
- [25] Zhao J, Cheng Y, Cheng Y, et al. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition[C] // Proceedings of the AAAI conference on artificial intelligence. 2019 : 9251 – 9258.
- [26] 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.
- [27] He R, Cao J, Song L, et al. Adversarial cross-spectral face completion for NIR-VIS face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(5) : 1025 – 1037.
- [28] Yang Z, Liang J, Fu C, et al. Heterogeneous Face Recognition via Face Synthesis With Identity-Attribute Disentanglement[J]. IEEE Transactions on Information Forensics and Security, 2022, 17 : 1344 – 1358.
- [29] Peng C, Wang N, Li J, et al. Re-ranking high-dimensional deep local representation for NIR-VIS face recognition[J]. IEEE Transactions on Image Processing, 2019, 28(9) : 4553 – 4565.
- [30] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015 : 815 – 823.
- [31] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019 : 4690 – 4699.
- [32] Zhu J, Zheng W, Lai J, et al. Matching NIR face to VIS face using transduction[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(3) : 501 – 514.
- [33] Ahonen T, Hadid A, Pietikainen M. Face description with local binary patterns: Application to face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(12) : 2037 – 2041.

- [34] Klare B, Jain A K. Heterogeneous face recognition: Matching NIR to VIS light images[C] //Proceedings of the IEEE International Conference on Pattern Recognition (ICPR). 2010 : 1513 – 1516.
- [35] Lowe D. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2) : 91 – 110.
- [36] SHENGLiao, Yi D, Lei Z, et al. Heterogeneous face recognition from local structures of normalized appearance[C] //Proceedings of the International Conference on Biometrics (ICB). 2009 : 209 – 218.
- [37] Zhang W, Wang X, Tang X. Coupled information-theoretic encoding for face photo-sketch recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2011 : 513 – 520.
- [38] Liao S, Yi D, Lei Z, et al. Heterogeneous face recognition from local structures of normalized appearance[C] // Proceedings of International Conference on Biometrics (ICB). 2009 : 209 – 218.
- [39] Klare B, Li Z, Jain A K. Matching forensic sketches to mug shot photos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(3) : 639 – 646.
- [40] Liu S, Yi D, Lei Z, et al. Heterogeneous face image matching using multi-scale features[C] // Proceedings of the International Conference on Biometrics (ICB). 2012 : 79 – 84.
- [41] Yi D, Lei Z, Li S. Shared representation learning for heterogenous face recognition[C] // IEEE international conference and workshops on automatic face and gesture recognition (FG). 2015 : 1 – 7.
- [42] Klare B, Jain A K. Heterogeneous face recognition using kernel prototype similarities[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(6) : 1410 – 1422.
- [43] Bhatt H S, Bharadwaj S, Singh R, et al. On matching sketches with digital face images[C] //Proceedings of IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). 2010 : 1 – 7.
- [44] Lei Z, Li S. Coupled spectral regression for matching heterogeneous faces[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICPR). 2009 : 1123 – 1128.
- [45] Bhattacharya S, Nainala G S, Rooj S, et al. Local force pattern (LFP): Descriptor for heterogeneous face recognition[J]. Pattern Recognition Letters, 2019, 125 : 63 – 70.

- [46] Roy H, Bhattacharjee D. Local-gravity-face (LG-face) for illumination-invariant and heterogeneous face recognition[J]. IEEE Transactions on Information Forensics and Security, 2016, 11(7): 1412–1424.
- [47] Peng C, Wang N, Li J, et al. DLFace: Deep local descriptor for cross-modality face recognition[J]. Pattern Recognition, 2018, 90: 161–171.
- [48] Lu J, Liong V E, Zhou J. Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(8): 1979–1993.
- [49] Lu J, Liong V E, Zhou X, et al. Learning compact binary face descriptor for face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(10): 2041–2056.
- [50] Bhattacharya S. Linear Cross-Modal Hash Encoding (LCMHE) for visual and near-infrared face recognition[J]. IEEE Sensors Letters, 2021, 5(4): 1–4.
- [51] Roy H, Bhattacharjee D, Krejcar O. Interpretable Local Frequency Binary Pattern (LFrBP) Based Joint Continual Learning Network for Heterogeneous Face Recognition[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2125–2136.
- [52] Tang X, Wang X. Face sketch synthesis and recognition[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2003: 687–694.
- [53] Wang X, Tang X. Face photo-sketch synthesis and recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(11): 1955–1967.
- [54] Liu Q, Tang X, Jin H, et al. A nonlinear approach for face sketch synthesis and recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005: 1005–1010.
- [55] Wang N, Li J, Tao D, et al. Heterogeneous image transformation[J]. Pattern Recognition Letters, 2013, 34(1): 77–84.
- [56] Peng C, Gao X, Wang N, et al. Multiple representations-based face sketch-photo synthesis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2015, 27(11): 2201–2215.
- [57] Wang N, Tao D, Gao X, et al. Transductive face sketch-photo synthesis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(9): 1364–1376.

- [58] Chang L, Zhou M, Deng X, et al. Face sketch synthesis via multivariate output regression[C] // Proceedings of the international conference on Human-computer interaction: design and development approaches (HCI). 2011 : 555 – 561.
- [59] Zhang J, Wang N, Gao X, et al. Face sketch-photo synthesis based on support vector regression[C] // Proceedings of the IEEE International Conference on Image Processing (ICIP). 2011 : 1125 – 1128.
- [60] Xiao B, Gao X, Tao D, et al. A new approach for face recognition by sketches in photos[J]. Signal Processing, 2009, 89(8) : 1576 – 1588.
- [61] Zhong J, Gao X, Tian C. Face sketch synthesis using E-HMM and selective ensemble[C] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2007 : 458 – 488.
- [62] Zhang W, Wang X, Tang X. Lighting and pose robust face sketch synthesis[C] // Proceedings of the European Conference on Computer Vision (ECCV). 2010 : 420 – 433.
- [63] Li Y-H, Savvides M, Bhagavatula V. Illumination tolerant face recognition using a novel face from sketch synthesis approach and advanced correlation filters[C] // Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP). 2006 : 357 – 360.
- [64] Liu W, Tang X, Liu J. Bayesian Tensor Inference for Sketch-Based Facial Photo Hallucination.[C] // Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2007 : 2141 – 2146.
- [65] Chang L, Zhou M, Han Y, et al. Face sketch synthesis via sparse representation[C] // Proceedings of the International Conference on Pattern Recognition (ICPR). 2010 : 2146 – 2149.
- [66] Gao X, Wang N, Tao D, et al. Face sketch-photo synthesis and retrieval using sparse representation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(8) : 1213 – 1226.
- [67] Ji N, Chai X, Shan S, et al. Local regression model for automatic face sketch generation[C] // Proceedings of the International Conference on Image and Graphics (ICIG). 2011 : 412 – 417.
- [68] Wang S, Zhang L, Liang Y, et al. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2012 : 2216 – 2223.
- [69] Zhao P, Zhang F, Wei J, et al. SADG: Self-aligned dual NIR-VIS generation for heterogeneous face recognition[J]. Applied Sciences, 2021, 11(3) : 987 – 1002.

- [70] Yu J, Cao J, Li Y, et al. Pose-preserving Cross Spectral Face Hallucination.[C] //Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2019 : 1018–1024.
- [71] Duan B, Fu C, Li Y, et al. Cross-spectral face hallucination via disentangling independent factors[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2020 : 7930–7938.
- [72] Luo M, Wu H, Huang H, et al. Memory-modulated transformer network for heterogeneous face recognition[J]. IEEE Transactions on Information Forensics and Security, 2022, 17 : 2095–2109.
- [73] Fu C, Wu X, Hu Y, et al. Dvg-face: Dual variational generation for heterogeneous face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(6) : 2938–2952.
- [74] Fu C, Zhou X, He W, et al. Towards Lightweight Pixel-Wise Hallucination for Heterogeneous Face Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7) : 9135–9148.
- [75] Liu D, Gao X, Peng C, et al. Heterogeneous face interpretable disentangled representation for joint face recognition and synthesis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(10) : 5611–5625.
- [76] Song L, Zhang M, Wu X, et al. Adversarial discriminative heterogeneous face recognition[C] //Proceedings of the AAAI conference on artificial intelligence. 2021 : 7355–7362.
- [77] He R, Li Y, Wu X, et al. Coupled adversarial learning for semi-supervised heterogeneous face recognition[J]. Pattern Recognition, 2021, 110 : 107618–107628.
- [78] 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 (ICCV). 2017 : 2223–2232.
- [79] Pang M, Wang B, Chen S, et al. Cross-domain Prototype Learning from Contaminated Faces via Disentangling Latent Factors[C] //Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM). 2022 : 4369–4373.
- [80] Yu S, Han H, Shan S, et al. CMOS-GAN: Semi-Supervised Generative Adversarial Model for Cross-Modality Face Image Synthesis[J]. IEEE Transactions on Image Processing, 2022, 32 : 144–158.

- [81] Zhang Y, Hu W, Sun B, et al. CoConGAN: Cooperative contrastive learning for few-shot cross-domain heterogeneous face translation[J]. *Neural Computing and Applications*, 2023, 35(20) : 15019 – 15032.
- [82] Wu H, Huang H, Yu A, et al. Exemplar guided cross-spectral face hallucination via mutual information disentanglement[C] // *Proceedings of the International Conference on Pattern Recognition (ICPR)*. 2021 : 4206 – 4212.
- [83] Mei Y, Guo P, Patel V M. Escaping data scarcity for high-resolution heterogeneous face hallucination[C] // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 : 18676 – 18686.
- [84] Lin D, Tang X. Inter-modality face recognition[C] // *Proceedings of the European Conference on Computer Vision (ECCV)*. 2006 : 13 – 26.
- [85] Lei Z, Zhou C, Yi D, et al. An improved coupled spectral regression for heterogeneous face recognition[C] // *Proceedings of the International Conference on Biometrics (ICB)*. 2012 : 7 – 12.
- [86] Gong D, Li Z, Liu J, et al. Multi-feature canonical correlation analysis for face photo-sketch image retrieval[C] // *Proceedings of the ACM international conference on Multimedia (MM)*. 2013 : 617 – 620.
- [87] Kan M, Shan S, Zhang H, et al. Multi-view discriminant analysis[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 38(1) : 188 – 194.
- [88] Mignon A, Jurie F. CMML: A new metric learning approach for cross modal matching[C] // *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 2012 : 1 – 14.
- [89] Huo J, Gao Y, Shi Y, et al. Heterogeneous face recognition by margin-based cross-modality metric learning[J]. *IEEE Transactions on Cybernetics*, 2017, 48(6) : 1814 – 1826.
- [90] Huang X, Lei Z, Fan M, et al. Regularized discriminative spectral regression method for heterogeneous face matching[J]. *IEEE Transactions on Image Processing*, 2012, 22(1) : 353 – 362.
- [91] Jin Y, Lu J, Ruan Q. Large margin coupled feature learning for cross-modal face recognition[C] // *Proceedings of the International Conference on Biometrics (ICB)*. 2015 : 286 – 292.
- [92] Liu X, Song L, Wu X, et al. Transferring deep representation for NIR-VIS heterogeneous face recognition[C] // *Proceedings of the International Conference on Biometrics (ICB)*. 2016 : 1 – 8.

- [93] He R, Wu X, Sun Z, et al. Wasserstein CNN: Learning invariant features for NIR-VIS face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(7) : 1761 – 1773.
- [94] Hu W, Hu H. Domain-Private Factor Detachment Network for NIR-VIS Face Recognition[J]. IEEE Transactions on Information Forensics and Security, 2022, 17 : 1435 – 1449.
- [95] Wu X, Huang H, Patel V M, et al. Disentangled variational representation for heterogeneous face recognition[C] // Proceedings of the AAAI conference on artificial intelligence. 2019 : 9005 – 9012.
- [96] Yang S, Yang X, Lin Y, et al. Heterogeneous Face Recognition with Attention-guided Feature Disentangling[C] // Proceedings of the ACM International Conference on Multimedia (MM). 2021 : 4137 – 4145.
- [97] Hu W, Hu H. Disentangled spectrum variations networks for NIR–VIS face recognition[J]. IEEE Transactions on Multimedia, 2019, 22(5) : 1234 – 1248.
- [98] Hu W, Hu H. Dual adversarial disentanglement and deep representation decorrelation for NIR-VIS face recognition[J]. IEEE Transactions on Information Forensics and Security, 2020, 16 : 70 – 85.
- [99] Peng C, Wang N, Li J, et al. Soft semantic representation for cross-domain face recognition[J]. IEEE Transactions on Information Forensics and Security, 2020, 16 : 346 – 360.
- [100] Cho M, Kim T, Kim I-J, et al. Relational deep feature learning for heterogeneous face recognition[J]. IEEE Transactions on Information Forensics and Security, 2020, 16 : 376 – 388.
- [101] Yu J, Feng Y. An Intrinsic Structured Graph Alignment Module With Modality-Invariant Representations for NIR-VIS Face Recognition[J]. IEEE Signal Processing Letters, 2022, 29 : 1017 – 1021.
- [102] Mudunuri S, Venkataramanan S, Biswas S. Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(4) : 886 – 896.
- [103] de Freitas Pereira T, Anjos A, Marcel S. Heterogeneous face recognition using domain specific units[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(7) : 1803 – 1816.
- [104] Sarfraz M, Stiefelhagen R. Deep perceptual mapping for cross-modal face recognition[J]. International Journal of Computer Vision (IJCV), 2017, 122 : 426 – 438.

- [105] Fondje C, Hu S, Riggan B S. Learning Domain and Pose Invariance for Thermal-to-Visible Face Recognition[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2022, 5(1): 15 – 28.
- [106] Wu F, Jing X-Y, Feng Y, et al. Spectrum-aware discriminative deep feature learning for multi-spectral face recognition[J]. Pattern Recognition, 2021, 111 : 107632 – 107644.
- [107] Ghosh S, Singh R, Vatsa M. Subclass heterogeneity aware loss for cross-spectral cross-resolution face recognition[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2020, 2(3): 245 – 256.
- [108] Hu W, Hu H. Orthogonal modality disentanglement and representation alignment network for NIR-VIS face recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(6): 3630 – 3643.
- [109] Guo Y, Zhang L, Hu Y, et al. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition[C] // Proceedings of the European conference on computer vision (ECCV). 2016 : 87 – 102.
- [110] Zhu Z, Huang G, Deng J, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021 : 10492 – 10502.
- [111] Chen J, Yi D, Yang J, et al. Learning mappings for face synthesis from near infrared to visual light images[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2009 : 156 – 163.
- [112] Huang D, Sun J, Wang Y. The BUAA-VisNir face database instructions[J]. Tech. Rep. IRIP-TR-12-FR-001, Beihang University, Beijing, China, 2012 : 1 – 5.
- [113] Gong D, Li Z, Huang W, et al. Heterogeneous face recognition: A common encoding feature discriminant approach[J]. IEEE Transactions on Image Processing, 2017, 26(5) : 2079 – 2089.
- [114] Liu D, Li J, Wang N, et al. Composite components-based face sketch recognition[J]. Neurocomputing, 2018, 302 : 46 – 54.
- [115] Di X, Zhang H, Patel V M. Polarimetric thermal to visible face verification via attribute preserved synthesis[C] // IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS). 2018 : 1 – 10.

- [116] Zhang H, Riggan B S, Hu S, et al. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks[J]. International Journal of Computer Vision (IJCV), 2019, 127 : 845 – 862.
- [117] Dai P, Ji R, Wang H, et al. Cross-modality person re-identification with generative adversarial training[C] // Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). 2018 : 677 – 683.
- [118] Hao X, Zhao S, Ye M, et al. Cross-modality person re-identification via modality confusion and center aggregation[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2021 : 16403 – 16412.
- [119] 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 Conference on Computer Vision and Pattern Recognition (CVPR). 2022 : 20973 – 20982.
- [120] Xu Y, Zhang L, Duan Q. Domain Private and Agnostic Feature for Modality Adaptive Face Recognition[C] // Proceedings of the IEEE International Joint Conference on Biometrics (IJCB). 2020 : 1 – 9.
- [121] 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 (ICCV). 2017 : 5380 – 5389.
- [122] 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.
- [123] Zhang Y, Yan Y, Lu Y, et al. Towards a unified middle modality learning for visible-infrared person re-identification[C] // Proceedings of the ACM International Conference on Multimedia (MM). 2021 : 788 – 796.
- [124] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE signal processing letters, 2016, 23(10) : 1499 – 1503.
- [125] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 : 770 – 778.
- [126] Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module[C] // Proceedings of the European conference on computer vision (ECCV). 2018 : 3 – 19.

- [127] Huang G, Mattar M, Berg T, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments[C] // Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition. 2008 : 1 – 15.
- [128] He R, Wu X, Sun Z, et al. Learning invariant deep representation for nir-vis face recognition[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2017 : 2000 – 2006.
- [129] Wu X, He R, Sun Z, et al. A light CNN for deep face representation with noisy labels[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(11) : 2884 – 2896.
- [130] Wu X, Song L, He R, et al. Coupled deep learning for heterogeneous face recognition[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2018 : 1679 – 1686.
- [131] Hu W, Hu H. Adversarial disentanglement spectrum variations and cross-modality attention networks for NIR-VIS face recognition[J]. IEEE Transactions on Multimedia, 2020, 23 : 145 – 160.
- [132] Panetta K, Wan Q, Agaian S, et al. A comprehensive database for benchmarking imaging systems[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 42(3) : 509 – 520.
- [133] Wang N, Gao X, Sun L, et al. Anchored neighborhood index for face sketch synthesis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 28(9) : 2154 – 2163.
- [134] Peng C, Gao X, Wang N, et al. Superpixel-based face sketch-photo synthesis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 27(2) : 288 – 299.
- [135] Hu W, Yan W, Hu H. Dual face alignment learning network for NIR-VIS face recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(4) : 2411 – 2424.
- [136] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2017 : 4278 – 4284.
- [137] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition[C] // Proceedings of the European conference on computer vision (ECCV). 2016 : 499 – 515.
- [138] Luo M, Ma X, Li Z, et al. Partial NIR-VIS heterogeneous face recognition with automatic saliency search[J]. IEEE Transactions on Information Forensics and Security, 2021, 16 : 5003 – 5017.

- [139] Deng W, Zheng L, Ye Q, et al. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018 : 994 – 1003.
- [140] Huang Y, Wu Q, Xu J, et al. SBSGAN: Suppression of inter-domain background shift for person re-identification[C] //Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019 : 9527 – 9536.
- [141] Ge Y, Li Z, Zhao H, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification[J]. Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2018 : 1222 – 1233.
- [142] Chen Y, Zhu X, Gong S. Instance-guided context rendering for cross-domain person re-identification[C] //Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019 : 232 – 242.
- [143] VERMA A, SUBRAMANYAM A V, WANG Z, et al. Unsupervised domain adaptation for person re-identification via individual-preserving and environmental-switching cyclic generation[J]. IEEE Transactions on Multimedia, 2021, 25 : 364 – 377.
- [144] Dai Z, Wang G, Yuan W, et al. Cluster contrast for unsupervised person re-identification[C] // Proceedings of the Asian Conference on Computer Vision (ACCV). 2022 : 1142 – 1160.
- [145] Ge Y, Zhu F, Chen D, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id[J]. Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2020 : 11309 – 11321.
- [146] Cheng D, Zhou J, Wang N, et al. Hybrid dynamic contrast and probability distillation for unsupervised person re-id[J]. IEEE Transactions on Image Processing, 2022, 31 : 3334 – 3346.
- [147] Lin S, Li H, Li C-T, et al. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification[J]. arXiv preprint arXiv:1807.01440, 2018.
- [148] Wang J, Zhu X, Gong S, et al. Transferable joint attribute-identity deep learning for unsupervised person re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018 : 2275 – 2284.

- [149] Yang F, Yan K, Lu S, et al. Part-aware progressive unsupervised domain adaptation for person re-identification[J]. IEEE Transactions on Multimedia, 2020, 23 : 1681 – 1695.
- [150] Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C] // Knowledge Discovery and Data Mining. 1996 : 226 – 231.
- [151] MacQueen J. Classification and analysis of multivariate observations[C] // Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability. 1967 : 281 – 297.
- [152] Feng H, Chen M, Hu J, et al. Complementary pseudo labels for unsupervised domain adaptation on person re-identification[J]. IEEE Transactions on Image Processing, 2021, 30 : 2898 – 2907.
- [153] Si T, He F, Zhang Z, et al. Hybrid contrastive learning for unsupervised person re-identification[J]. IEEE Transactions on Multimedia, 2023, 25 : 4323 – 4334.
- [154] Zhou Y, Wang Y, Chau L-P. Moving towards centers: Re-ranking with attention and memory for re-identification[J]. IEEE Transactions on Multimedia, 2023, 25 : 3456 – 3468.
- [155] Liu T, Lin Y, Du B. Unsupervised person re-identification with stochastic training strategy[J]. IEEE Transactions on Image Processing, 2022, 31 : 4240 – 4250.
- [156] Tang X, Wang X. Face photo recognition using sketch[C] // Proceedings of IEEE International Conference on Image Processing : Vol 1. 2002 : 257 – 260.
- [157] Zhu M, Wang N, Gao X, et al. Face Photo-Sketch Synthesis via Knowledge Transfer[C] // Proceedings of International Joint Conference on Artificial Intelligence (IJCAI). 2019 : 1048 – 1054.
- [158] Cao B, Wang Q, Zhu P, et al. Multi-view knowledge ensemble with frequency consistency for cross-domain face translation[J/OL]. IEEE Transactions on Neural Networks and Learning Systems, 2023. <http://dx.doi.org/10.1109/TNNLS.2023.3236486>.
- [159] Sun Z, Fu C, Luo M, et al. Self-Augmented Heterogeneous Face Recognition[C] // IEEE International Joint Conference on Biometrics (IJCBI). 2021 : 1 – 8.
- [160] Wu H, Prasad S. Semi-supervised deep learning using pseudo labels for hyperspectral image classification[J]. IEEE Transactions on Image Processing, 2017, 27(3) : 1259 – 1270.
- [161] Yan P, Li G, Xie Y, et al. Semi-supervised video salient object detection using pseudo-labels[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019 : 7284 – 7293.

- [162] Petrovai A, Nedeveschi S. Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2022 : 1578 – 1588.
- [163] Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019 : 3828 – 3838.
- [164] Alwassel H, Mahajan D, Korbar B, et al. Self-supervised learning by cross-modal audio-video clustering[J]. Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2020, 33 : 9758 – 9770.
- [165] Liu Y, Chen Q, Albanie S. Adaptive cross-modal prototypes for cross-domain visual-language retrieval[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021 : 14954 – 14964.
- [166] Mao Y, Yan X, Guo Q, et al. Deep mutual information maximin for cross-modal clustering[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2021 : 8893 – 8901.
- [167] Gao Q, Lian H, Wang Q, et al. Cross-modal subspace clustering via deep canonical correlation analysis[C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2020 : 3938 – 3945.
- [168] Wang F, Liu H. Understanding the behaviour of contrastive loss[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021 : 2495 – 2504.
- [169] Wang T, Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere[C] // Proceedings of International Conference on Machine Learning (ICML). 2020 : 9929 – 9939.
- [170] Fu X, Huang F, Zhou Y, et al. Cross-Modal Cross-Domain Dual Alignment Network for RGB-Infrared Person Re-Identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6874 – 6887.
- [171] 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 (CVPR). 2017 : 1318 – 1327.
- [172] Peebles W, Xie S. Scalable diffusion models with transformers[C] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2023 : 4195 – 4205.

- [173] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C] // Proceedings of the International Conference on Machine Learning (ICML). 2021 : 8748–8763.
- [174] Croitoru F-A, Hondru V, Ionescu R, et al. Diffusion Models in Vision: A Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9) : 10850–10869.
- [175] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
- [176] Sun Z, Fang Y, Wu T, et al. Alpha-CLIP: A clip model focusing on wherever you want[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR). 2024.
- [177] Fathullah Y, Wu C, Lakomkin E, et al. Prompting large language models with speech recognition abilities[J]. arXiv preprint arXiv:2307.11795, 2023.

攻读硕士学位期间相关的科研成果目录

1. 发表论文

- [1] **Yiming Yang**, Weipeng Hu, Haifeng Hu. Syncretic Space Learning Network for NIR-VIS Face Recognition[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 20(1): 1-25. (已发表, JCRQ1、中科院三区, 对应本文第二章内容)
- [2] **Yiming Yang**, Weipeng Hu, Haifeng Hu. Neutral Face Learning and Progressive Fusion Synthesis Network for NIR-VIS Face Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(10):5750-5763. (已发表, 中科院一区、Top 期刊, 对应本文第三章内容)
- [3] **Yiming Yang**, Weipeng Hu, Haiqi Lin, Haifeng Hu. Robust Cross-Domain Pseudo-Labeling and Contrastive Learning for Unsupervised Domain Adaptation NIR-VIS Face Recognition[J]. IEEE Transactions on Image Processing, 2023, 32:5231-5244. (已发表, 中科院一区、Top 期刊, 对应本文第四章内容)
- [4] **Yiming Yang**, Weipeng Hu, Haifeng Hu. Unsupervised NIR-VIS Face Recognition via Homogeneous-to-Heterogeneous Learning and Residual-Invariant Enhancement[J]. IEEE Transactions on Information Forensics and Security, 2023, 19:2112-2126. (已发表, 中科院一区、Top 期刊, 对应本文第五章内容)
- [5] **Yiming Yang**, Weipeng Hu, Haiqi Lin, Haifeng Hu. Progressive Cross-modal Association Learning for Unsupervised Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Information Forensics and Security, 2023. (已投稿, 中科院一区、Top 期刊)
- [6] **Yiming Yang**, Weipeng Hu, Haifeng Hu. Dynamic Modality-Camera Invariant Clustering for Unsupervised Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Image Processing, 2023. (已投稿, 中科院一区、Top 期刊)

- [7] Weipeng Hu, **Yiming Yang**, Haifeng Hu. Pseudo Label Association and Prototype-based Invariant Learning for semi-supervised NIR-VIS Face Recognition[J]. IEEE Transactions on Image Processing, 2024, 33:1448-1463. (已发表, 中科院一区、Top 期刊)
- [8] Qiaolin He, **Yiming Yang**, Haifeng Hu. Part-based Bi-directional Enhancement Learning for Unsupervised Visible-Infrared Re-identification[C]. Proceedings of the European conference on computer vision (ECCV), 2024. (已投稿)
- [9] Weipeng Hu, Yap-Peng Tan, Haifeng Hu, **Yiming Yang**, Jiun Tian Hoe. Motion-Guided Token Prioritization Transformer for Exo-to-Ego Cross-view Video Generation[J]. IEEE Transactions on multimedia, 2023. (已投稿, 中科院一区、Top 期刊)

2. 奖项荣誉

- [1] 2022-2023 学年度硕士研究生**国家奖学金** (2023 年 11 月)
- [2] 2023 年中山大学**院级优秀共产党员**

致谢

%

杨奕铭

2026 年 9 月 6 日

中山大学学位论文版权使用授权书

本学位论文作者及指导教师完全了解“中山大学硕士、博士（硕士）学位论文版权使用规定”，同意中山大学保留并向国家有关部门或机构送交学位论文的复印件和电子版，允许论文被查阅和借阅。本人授权中山大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，也可采用影印、缩印或扫描等复制手段保存和汇编学位论文。

作者签名：

导师签名：

年 月 日

