基于局部异构聚合图卷积网络的跨模态行人重识别

孙 锐[1,2],张 磊[1,2],余益衡[1,2],张旭东[1,2]
(1. 合肥工业大学计算机与信息学院,安徽合肥 230601;2. 工业安全与应急技术安徽省重点实验室,安徽合肥 230009)

摘 要: 由于构建全天候视频监控系统的需要,基于可见光与红外的跨模态行人重识别问题受到学术界的广泛关注. 因为类内变化和类间差异的影响,可见光与红外行人重识别是一项具有挑战性的任务. 现有的工作主要集中在可见光-红外图像转换或跨模态的全局共享特征学习,而身体部位的局部特征和这些特征之间的结构关系在很大程度上被忽略了. 我们认为局部关键点之间的图结构关系在模态内与模态间的变化是相对稳定的,充分挖掘与表示这种结构信息有助于解决跨模态行人重识别问题. 本文提出了一种基于局部异构聚合图卷积网络的跨模态行人重识别方法,采用关键点提取网络提取图像的局部关键点特征,并构建了一种新颖的图卷积网络建模人体各部位之间的结构关系. 该网络通过图内卷积层表征局部特征的高阶结构关系信息,提取具有辨别力的局部特征. 网络中的跨图卷积层使两个异构图结构之间可以传递差异性特征,有助于减弱模态差异的影响. 针对异构图结构的图匹配问题,设计了一种跨模态排列损失以更好地测度图结构的距离. 本文方法在主流跨模态数据集 RegDB 和 SYSU-MM01 上的 mAP/Rank-1 为 80.78%/80.55% 和 67.92%/66.49%,比 VDCM 算法的 Rank-1 分数高出 7.58% 和 1.87%.

关键词: 行人重识别;跨模态;异构聚合;图卷积网络;关键点提取网络
基金项目: 国家自然科学基金面上项目(No.61471154,No.61876057);安徽省重点研发计划-科技强警专项项目(No.202004d07020012)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2023)04-0810-16
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220011

Cross-Modality Person Re-identification Based on Locally Heterogeneous Polymerization Graph Convolutional Network

SUN Rui[1,2], ZHANG Lei[1,2], YU Yi-heng[1,2], ZHANG Xu-dong[1,2]
(1. *School of Computer Science and Information Engineering*, *Hefei University of Technology*, *Hefei*, *Anhui* 230601, *China*;
2. *Anhui Province Key Laboratory of Industry Safety and Emergency Technology*, *Hefei*, *Anhui* 230009, *China*)

Abstract: The research of cross-modality person re-identification based on visible-infrared has attracted widespread attention from the academia due to the need to build an all-day video surveillance system. Visible-infrared person re-identification is a challenging task due to intra-class variation and cross-modality discrepancy. Existing work focused on visible-infrared modal transformations or global shared feature learning across modalities, while local features of body parts and the structural relationships between these features have been largely ignored. We consider that the graph structure relationship between local key-points is relatively stable within and between modality variations, and fully mining and representing this structural information can help solve the cross-modal person re-identification problem. Therefore, this paper proposes a cross-modal person re-identification method based on local heterogeneous polymerization graph convolutional networks. A key-points extraction network is used to extract the local key-points' features of the image, and then a novel graph convolutional network is constructed to model the structural relationships between various parts of the human body. The network characterizes the higher-order structural relationship information of local features through the intra-graph convolutional layer, and finally extracts discriminative local features. The cross-graph convolutional layer in the network enables the transfer of discriminative features between two heterogeneous graph structures, which helps to reduce the effect of modal differences. Finally, a cross-modality permutation loss is designed to better measure the distance of graph structures for the graph matching problem of heterogeneous graph structures. The mAP/Rank-1 of our method on the mainstream cross-modal datasets RegDB and SYSU-MM01 is 80.78%/80.55% and 67.92%/66.49%, which is 7.58% and 1.87% higher than the Rank-1

收稿日期:2021-12-29;修回日期:2022-05-13;责任编辑:朱梅玉

scores of the VDCM algorithm.

Key words:　person re-identification; cross-modality; hetero-polymerization; graph convolutional network; key-points extraction network

Foundation Item(s):　National Natural Science Foundation of China (No.61471154, No.61876057); Key Research and Development Plan of Anhui Province (No.202004d07020012)

1　引言

近年来,随着城市视频监控网络的不断完善,行人重识别技术由于其巨大的应用潜力而受到越来越多的关注. 行人重识别(Person Re-identification,Re-ID)是一种跨摄像机的图像检索任务,其目的是从不相交摄像机采集的图像库中检索给定查询的人员. 早期的研究工作主要集中在手工特征构建或相似度度量学习. 近几年随着深度学习的发展,基于深度学习的行人重识别方法在性能上大大超过了传统方法[1]. 许多研究采取深度度量学习[2~5],或者使用卷积神经网络提取具有区分度的特征[5~8],随着生成对抗网络(Generative Adversarial Networks,GAN)的新进展,另一种主流方法是将GAN作为一种风格转换器,以增强训练数据并提高模型的辨别能力[9~13].

现有的行人重识别方法主要是处理单模态的可见光图像,也称作同构行人重识别. 但在真实复杂场景中,即在明亮和黑暗的交叉光照环境中捕捉人物图像,这些方法的效果会显著降低. 同时,可见光相机不能在夜间工作. 幸运的是,一些新型监视设备,如可见光-红外双模摄像机(RGB-IR Dual-mode Cameras),在较差的照明条件下仍然可以捕捉人的外观特征. 这引起了工业界和学术界对可见光-红外(RGB-IR)异构匹配的广泛研究兴趣. 与同构行人重识别不同,跨模态异构场景下两种模态间图像以及单一模态内人的外观特征均存在较大差异. 例如,可见光图像包含了一些像颜色这样的鉴别线索,而这些信息在红外图像中缺失[1]. 可见光图像与红外图像的差异如图1所示. 因此,跨模态异构行人重识别更具有挑战性.

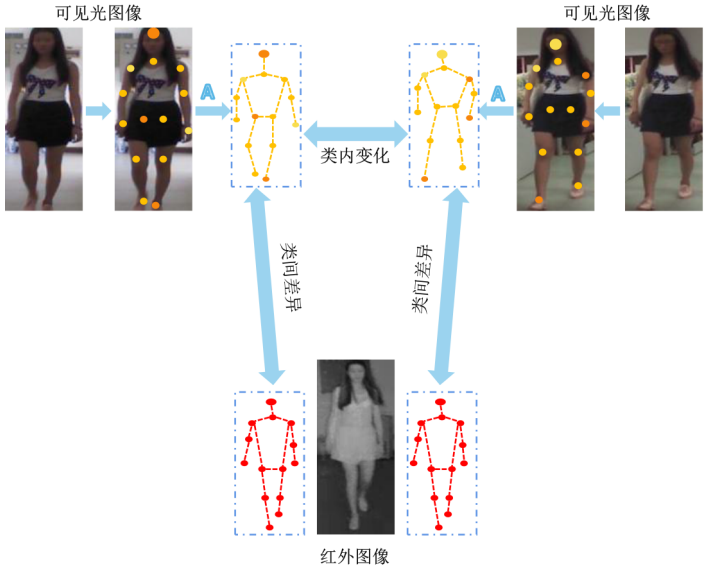

图1　类内变化和类间差异的图示

最近,许多研究采用了两类典型的方法来解决上述跨模态行人重识别的挑战. 第一类方法试图通过特征级约束(如对齐图像的特征分布)来减少跨模态差异[14~19]. 其目标是最大化具有相同身份特征(类内特征)的相似性,并最小化具有不同身份特征(类间特征)的相似性. Wu等[14]提出了一种基于深度零填充的方式将两种模态以参数共享的方法进行训练来解决跨模态问题. Ye等[16]设计了一种双流网络来学习多模态共享特征,同时利用双向双约束顶级损失来约束模态间和模态内的差异. Zhu等[19]设计了双流局部特征网络,为了改进类内跨模态相似性,提出异质中心损失,将两个异质模态中心之间的距离拉近,提高模态间特征的相似度.

第二类方法是在输入级使用GAN将图像从一种模态转换到另一种模态[20~23],同时尽可能地保存身份信息. Dai等[20]设计一个基于判别器的生成对抗训练模型,从不同的模态中学习具有判别力的特征. 为了减少模态差异,Wang等[21]提出一种将红外图像和可见光图

像进行相互转换的方法,统一不同模态的图像表示,并结合特征级子网络,获得更具有辨别力的特征. Wang等[22]认为红外图像比彩色图像的识别效果高,将彩色图像全部转换为红外图像并用于跨模态行人重识别中. 这两类方法主要关注于减少模态间的差异,然而在单一RGB或IR模态中仍然存在外观差异的挑战,包括背景杂乱、视点变化、遮挡等.

与以往的可见光-红外重识别方法不同,我们建议分别处理外观差异和模态差异,而不是同时以特征嵌入的方式处理混合差异. 虽然背景、视角和遮挡等类内变化较大,但相同行人中的局部特征之间存在内在联系,这种结构关系在同一行人中变化较小,而在不同行人中变化较大,并且这种结构关系在模态变化之间是相对稳定的. 从这种观察角度出发,我们提出了局部异构聚合图卷积网络的跨模态行人重识别方法. 首先采用姿态估计的方法,提取行人的局部关键点信息,构成特殊的人体图结构. 随后,设计了从局部到全局区分特征的自适应图卷积层来充分表示这种结构关系. 采用图结构来表示行人特征的优点在于,行人的不同部位包含不同的分辨力信息,提取关键点信息构成图结构可以有效减弱背景差异、视点变化和遮挡等. 即使人体的某些部位受到干扰,图结构仍然可以利用图内卷积层从未被干扰的区域捕获有用的信息. 而传统的共享特征提取网络或者基于可见光-红外的生成对抗网络难以减弱干扰. 此外,构建图结构可以更好地建模特征关系,对于图像差异较大的跨模态数据集来说,建模图结构关系可以拉进图像之间的距离. 具体而言,局部特征经过方向自适应图内卷积(directional adaptive Self-Graph Convolution Layer,SGCL)细化了有辨别力的特征,细化的局部聚合特征考虑了不同身体部位之间的高阶结构关系信息. 然后,局部特征再与全局特征整合,进一步增强了特征的分辨性. 这样,我们既考虑粗粒度整体外观信息又考虑细粒度局部结构信息,从整体和局部两个角度来判断一个人身体不同部位的重要性. 这也与人们在寻找判别性线索时的感知相一致:先整体进行比较,再观察细节,最后确定其重要性[24].

上述过程只独立地处理每个行人样本,忽略了可见光和红外异构图像之间的模态差异. 因此,我们设计了一种新颖且有效的模态匹配跨图卷积层(modality matching Cross-Graph Convolution Layer,CGCL),以获得最优的模内和模间的对应关系. 每个样本都包含其模态的特异性信息,可组成异构图结构对,利用模态匹配跨图卷积层将信息传播到其邻居. 该方案可以弥补同一行人不同图像中异构信息的缺失,进一步缩小两种模态间的隔阂. 最后,利用图相似性中的节点到节点的对应关系作为监督信息来测度RGB和IR图结构之间的距离. 实验结果表明,在SYSU-MM01[14]和RegDB[23]两个广泛使用的跨模态重识别数据集上,本文方法的性能显著超出现有方法.

综上所述,本文的主要贡献包括以下3点.

(1)提出一种基于局部异构融合图卷积网络的跨模态行人重识别方法,该方法结合整体语义和身体各部位高阶拓扑关系信息,构建人体拓扑图结构,并将模态内结构信息和模态间关系信息嵌入到节点的向量中,实现精确的样本级异构聚合.

(2)提出一种方向自适应图内卷积层(SGCL),促进语义区域的有意义信息的传递,抑制异常值等无意义区域的信息传递. 同时设计了一种模态匹配跨图卷积层(CGCL),用于学习可见光图与红外图之间的特征对齐,减弱了模态的差异.

(3)设计跨模态排列损失来对齐跨模态两个图结构之间的距离,利用异构图结构之间节点到节点的对应关系,即指派矩阵作为端到端的监督信息,有效地对跨模态中异构图结构进行距离测度. 实验表明,在多种损失函数的联合监督下,该方法可达到局部异构聚合的目的.

2 相关工作

2.1 RGB-RGB单模态行人重识别

传统的RGB-RGB单模态行人重识别是指仅采用单模态RGB图像的行人重识别,又称为同构行人重识别,用以解决行人在非重叠摄像头下的匹配问题. 在人的图像中存在较大的类内变化和较小的类间变化,如何从行人图像中学习有效特征是单模态面临的挑战. 为了应对上述挑战,已有许多深度行人重识别方法被提出[5,25~35]. 一些工作把目标放在部分特征学习上[25,26],专注于强大的网络结构来对齐身体部位[5,33]. 其他方法尝试设计新的损失函数在度量空间中约束特征分布,其中包括对比损失[2]、三元组损失[3]、四元组损失[26]. 最近的基于图的方法[27~31]考虑了样本对之间的关联性. 然而,这些方法只适用于单模态的再识别,因为跨模态图像不仅存在类内变化,还有较大的类间差异.

2.2 RGB-IR跨模态行人重识别

跨模态行人重识别的巨大差异不仅来自于图像外观的变化(类内变化),更来自于可见光图像和红外图像之间的跨模态变化(类间差异),因此也被称为异构行人重识别. 目前关于跨模态行人重识别的研究主要可以归纳为两类方法. 第一类方法试图在表示空间中对齐训练样本的特征分布[14~19]. Wu等[14]的工作重点是如何设计单流网络,例如改进特定域节点的深度零填充网络. Ye等[16]设计了基于双向双约束top-ranking损失的跨模态双流网络. Hao等[18]设计了一个超球流形

嵌入模型来学习不同模态下有辨别力的表征. 第二种方法是利用跨模态生成对抗网络(GAN)将行人图像风格从一种模态转移到另一种模态[20~23,32]. Dai 等[20]收集了一个新的热图像数据集,并提出了一个用于可见光-热图像转换的ThermalGAN框架. Wang等[22]进一步考虑双层差异,并使用双向cycleGAN生成未标记的图像作为数据增强. 在文献[23]中,作者提出了一种生成式对抗训练方法来联合识别身份和模态. Choi等[32]提出了一种分层解纠缠方法,利用跨模态图像中的姿态和光照不变的特征,同时分离身份判别因子和身份排除因子.

2.3　基于局部特征的行人重识别

众多学者[27,33~35]在行人重识别研究中开始探索图像的局部特征,通过结合身体部分区域信息,即加入局部空间信息来提高行人表示性. 局部CNN方法[33]提出将主干网络水平拆分为多个部分,并在每一层纳入局部区域信息. 在水平拆分的基础上,多粒度网络MGN[34]将网络分成三个分支,其中第一个分支用于提取全局特征,另外两个分支学习行人的局部表示. SPReID[35]提出将人体语义解析与行人重识别结合起来的建议. SPReID方法通过骨架关键点提取网络提取14个人体关键点,之后整合7个人体结构ROI,得到一个融合全局特征和多尺度局部特征的行人重识别特征. 然而,上述方法都是独立提取局部特征,忽略了身体不同部位之间的关系. Wang等人[27]为解决行人重识别中的遮挡问题,提取人体骨架关节点,并考虑遮挡中的节点间拓扑关系,提出新的对齐策略来减弱遮挡带来的影响. 在本文中,我们不仅利用不同身体部位之间的结构关系来增强局部信息的可辨别性,还利用不同模态中相同身份之间的整体性关系来增强模态感知力.

2.4　基于图卷积的行人重识别

近几年,图卷积网络(GCN)由于其强大的关系建模能力,已经成功应用于行人重识别领域[28~31]. 具体而言,Shen 等人[28]引入了相似性引导图神经网络(SGGNN),并利用探针-图库对之间的关系以端到端的方式更新关系特征. Yan 等[29]提出了一个上下文实例扩展模块,并构建了一个图学习框架,有效地利用上下文对更新目标相似度,提高了学习特征的区分度. 然而,这些基于图像的方法忽略了不同身体部位之间的关系. 特别地,Wu 等人[30]提出了自适应图表示学习方案,该方案利用姿态对齐连接和特征亲和连接来实现相关区域特征之间的语义交互. Yang 等[31]提出了一种时空图卷积网络(STGCN),它综合考虑了时间和结构的关系,以提取鲁棒的时空信息. 以上是基于图卷积网络学习相关的特征表示,但仅用来解决单模态任务.

与上述方法不同,我们采用新的图卷积网络来解决跨模态行人重识别的类间差异问题. 和传统的图卷积网络相比,我们的图卷积网络由图内卷积层和跨图卷积层组合而成,图内卷积层挖掘人体不同关键点的高阶关系信息,跨图卷积层缩小了不同模态的类别差异,最终由Sinkhorn算法求解图对应关系.

3　局部异构聚合图卷积网络框架

本文提出的局部异构聚合图卷积网络的框架由一阶关键点提取模块、高阶图卷积模块和图匹配及损失模块组成. 首先,关键点提取模块用于提取人体骨架的局部关键点特征,把每个关键点特征当作图节点,这些关键点构成天然的行人姿态图结构. 第二阶段,将图结构作为输入,通过高阶嵌入图卷积层进行聚合. 最后使用图匹配模块精确求解匹配结果,定义跨模态排列损失监督匹配信息. 这三个模块以端到端的方式进行联合训练. 整体网络框架如图2所示.

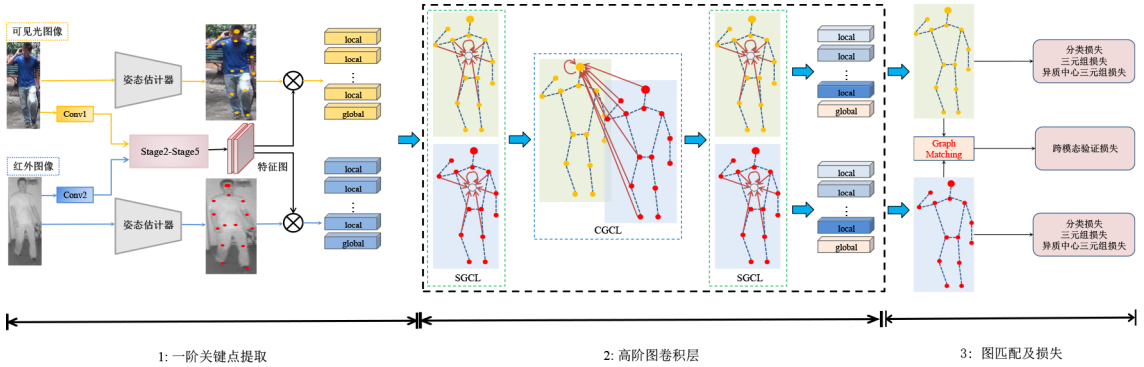

图2　本文方法网络框图

3.1　一阶关键点提取

3.1.1　关键点提取模块的架构

关键点提取模块的目标是提取关键点区域的一阶语义特征,其必要性有两点:首先,基于局部特征单模态行人重识别已被证明是非常有效的[5];其次,在跨模态行人重识别中,由于异构数据其特征相似度低,全局

特征的差异性巨大,因此提取局部特征是必要的. 根据上述想法,并受到最近在行人重识别[2,6,26,33]和人类关键点预测[36,37]方面的发展的启发,本文采用提取关键点的方法获取局部特征点. 本文采用2D人体姿态估计网络[38](deep High Resolution Network,HRNet)提取不同关键点的局部特征. 具体来说,给定一幅行人图像$\boldsymbol{x}^{\mathrm{RGB}}$和$\boldsymbol{x}^{\mathrm{IR}}$,通过HRNet可得到其关键点热度图$\boldsymbol{m}_{\mathrm{kp}}^{\mathrm{RGB}}$和$\boldsymbol{m}_{\mathrm{kp}}^{\mathrm{IR}}$.

关于提取一阶语义特征,可能会出现一个问题:我们采用的人体姿态估计网络通常使用可见光样本进行训练,它能否适用于红外图像? 为了回答这个问题,我们检查了RGB和IR图像中关键点提取的可视化结果(图3),并观察到我们的姿态估计网络可以在跨模态场景下有效工作. 图3(a)是随机抽取RegDB图像可视化的结果,图3(b)和(c)是SYSU-MM01数据集可视化的结果. 一个可能的原因是:行人地标的预测过程更多地依赖于模态共享的信息,如身体结构和形状,而不是颜色等模态特有的线索. 但在某些情况下,如图3(c)所示,行人的姿态、外观、背景等会产生巨大差异,甚至存在部分遮挡等干扰. 而如果仅使用传统双路网络,效果会显著下降. 所以我们通过构建图结构,采用图内卷积层加强与身份相关的结构关系信息,同时抑制被遮挡部分等无意义信息,来减弱这些干扰对跨模态行人重识别的影响.

(a) RegDB展示 (b) SYSU-MM01展示一 (c) SYSU-MM01展示二

图3 在跨模态数据集中姿态估计网络的可视化结果

此外,双流特征提取网络是可见光-红外跨模态人再识别中常用的全局特征提取方法[39],由特征提取和特征嵌入两部分组成. 特征提取器的目标是从两个异构模态中学习模态特有信息,而特征嵌入则侧重于学习多模态共享特征,通过将这些模态特定特征投影到模态共享的公共特征空间中进行特征学习. 在现有文献中,特征嵌入通常由一些共享的全连接层计算,特征提取器是一个设计良好的卷积神经网络. 然而这种特征嵌入难以捕获深层次共享特征模式. 我们先将$\boldsymbol{x}^{\mathrm{RGB}}$和$\boldsymbol{x}^{\mathrm{IR}}$送入参数不共享的卷积层$\mathrm{Conv}_1$和$\mathrm{Conv}_2$,以捕获特殊模态的低级特征模式. 然后选择共享特征提取器,即修正后的残差网络ResNet50[40],选用后4个板块Stage2-Stage5,将特征转换为一个共同的表示空间,以获得模态共享的全局特征$\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{RGB}}$和$\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{IR}}$,即

$$\begin{aligned}\boldsymbol{m}_{\mathrm{kp}}^{\mathrm{RGB}}&=\mathrm{HRnet}\left(\boldsymbol{x}^{\mathrm{RGB}}\right)\\ \boldsymbol{m}_{\mathrm{kp}}^{\mathrm{IR}}&=\mathrm{HRnet}\left(\boldsymbol{x}^{\mathrm{IR}}\right)\end{aligned}\tag{1}$$

$$\begin{aligned}\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{RGB}}&=\mathrm{ResNet}\left(\mathrm{Conv}_1\left(\boldsymbol{x}^{\mathrm{RGB}}\right)\right)\\ \boldsymbol{m}_{\mathrm{gr}}^{\mathrm{IR}}&=\mathrm{ResNet}\left(\mathrm{Conv}_2\left(\boldsymbol{x}^{\mathrm{IR}}\right)\right)\end{aligned}\tag{2}$$

通过哈达玛积($\boxtimes$)和全局平均池化操作($g(\cdot)$),我们可以得到一组关键点区域的语义局部特征和一个全局特征,即

$$\begin{aligned}\boldsymbol{V}_l^{\mathrm{RGB}}&=\left\{\boldsymbol{v}_k^{\mathrm{RGB}}\right\}_{k=1}^{K}=g\left(\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{RGB}}\boxtimes\boldsymbol{m}_{\mathrm{kp}}^{\mathrm{RGB}}\right)\\ \boldsymbol{V}_l^{\mathrm{IR}}&=\left\{\boldsymbol{v}_k^{\mathrm{IR}}\right\}_{k=1}^{K}=g\left(\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{IR}}\boxtimes\boldsymbol{m}_{\mathrm{kp}}^{\mathrm{IR}}\right)\end{aligned}\tag{3}$$

$$\begin{aligned}\boldsymbol{V}_g^{\mathrm{RGB}}&=\boldsymbol{v}_{K+1}^{\mathrm{RGB}}=g\left(\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{RGB}}\right)\\ \boldsymbol{V}_g^{\mathrm{IR}}&=\boldsymbol{v}_{K+1}^{\mathrm{IR}}=g\left(\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{IR}}\right)\end{aligned}\tag{4}$$

$$\boldsymbol{V}^E=\left(\boldsymbol{V}_l^{\mathrm{RGB/IR}},\boldsymbol{V}_g^{\mathrm{RGB/IR}}\right)=\left\{\boldsymbol{v}_k^E\right\}_{k=1}^{K+1}\tag{5}$$

其中,K是关键点的数量,$\boldsymbol{v}_k\in\mathbb{R}^C$,$C$为通道数. 此外,为防止噪声和异常值干扰特征热图,在使用前对$\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{RGB}}$和$\boldsymbol{m}_{\mathrm{gr}}^{\mathrm{IR}}$使用softmax函数归一化处理,归一化在公式中未表达.

3.1.2 关键点提取模块损失函数

三元组损失首先在FaceNet[41]中提出,然后通过难样本挖掘三元组[3]来改进. 其核心思想是通过随机抽样P个身份以及每个身份的K张图像,得到一个含有P个K张图像的小批量. 对于每一个小批量,选择批次中最难正样本和最难负的样本组成三元组,计算难样本挖掘三元组损失如下:

$$L_{\mathrm{bh\_tri}}(\boldsymbol{X})=\overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{\text{all anchors}}\left[\rho+\overbrace{\max_{p=1,2,\cdots,K}\left\|\boldsymbol{x}_a^i-\boldsymbol{x}_p^i\right\|_2}^{\text{hardest positive}}-\overbrace{\min_{\substack{j=1,2,\cdots,P\\ n=1,2,\cdots,K\\ j\neq i}}\left\|\boldsymbol{x}_a^i-\boldsymbol{x}_n^j\right\|_2}^{\text{hardest negative}}\right]_+\tag{6}$$

其中,$[x]_+=\max(x,0)$;$\|\boldsymbol{x}_a^i-\boldsymbol{x}_p^i\|_2$表示样本点$\boldsymbol{x}_a^i$与相同ID内正样本$\boldsymbol{x}_p^i$的欧氏距离;$\|\boldsymbol{x}_a^i-\boldsymbol{x}_p^j\|_2$表示样本点$\boldsymbol{x}_a^i$与不同ID内负样本$\boldsymbol{x}_n^j$的欧氏距离;max表示距离最大值,min表示距离最小值;ρ是超参数,用于控制样本对间的相对距离.

基于难样本挖掘表示的三元组损失是通过锚点与

所有其他样本的比较计算的. 这是一个强约束,如果存在一些严重干扰的样本,可能对成对距离的约束太过严格,会形成不利的三元值来破坏其他成对距离. 因此,我们考虑采用每个人的中心作为身份锚点,将锚点与所有其他样本的比较改为锚点中心与所有其他样本中心的比较.

$$\begin{aligned}\boldsymbol{c}_v^i&=\frac{1}{M}\sum_{j=1}^{M}\boldsymbol{v}_j^i\\ \boldsymbol{c}_t^i&=\frac{1}{M}\sum_{j=1}^{M}\boldsymbol{t}_j^i\end{aligned}\tag{7}$$

式(7)为一个批次异质中心的定义,其中,$\boldsymbol{v}_j^i$表示小批量中第i个人的第j张可见光图像特征,$\boldsymbol{t}_j^i$对应红外图像特征. 跨模态三元组损失的目标是将相同身份标签的中心从不同模态彼此拉近(类内紧致),而来自不同类的特性彼此远离(类间分离). 假设随机抽样N个身份形成批次,然后从每个身份的可见光和红外图像各随机抽样M张图像,得到一个含有N个$2M$张图像的批量. 可以将异质中心三元组损失定义为

$$\begin{aligned}L_{\text{hc\_tri}}(\boldsymbol{C})=&\sum_{i=1}^{N}\left[\rho+\left\|\boldsymbol{c}_v^i-\boldsymbol{c}_t^i\right\|_2-\min_{\substack{n\in\{v,t\}\\ j\neq i}}\left\|\boldsymbol{c}_v^i-\boldsymbol{c}_n^j\right\|_2\right]_+\\&+\sum_{i=1}^{N}\left[\rho+\left\|\boldsymbol{c}_t^i-\boldsymbol{c}_v^i\right\|_2-\min_{\substack{n\in\{v,t\}\\ j\neq i}}\left\|\boldsymbol{c}_t^i-\boldsymbol{c}_n^j\right\|_2\right]_+\end{aligned}\tag{8}$$

其中,$\boldsymbol{C}$是批量的特征中心,包含可见光特征中心$\{\boldsymbol{c}_v^i|i=1,2,\cdots,N\}$和红外特征中心$\{\boldsymbol{c}_t^i|i=1,2,\cdots,N\}$,对于每一个行人身份,$L_{\text{hc_tri}}$只计算一个跨模态的正样本对,以及一个在模态内和模态间挖掘最困难的负样本对.

为保证模态内特征的紧致性和模态间特征的可分辨性,本阶段联合异质中心损失以及传统的分类损失和三元组损失作为一阶关键点提取模块优化的目标,总损失L_E为

$$\begin{aligned}L_E&=\frac{1}{K+1}\sum_{k=1}^{K+1}\beta_k\left[L_{\text{cls}}\left(\boldsymbol{v}_k^{\text{R,I}}\right)+\left(1-\lambda_1\right)\times L_{\text{tri}}\left(\boldsymbol{v}_k^{\text{R,I}}\right)+\lambda_1\times L_{\text{hc\_tri}}\left(\boldsymbol{v}_k^{\text{R,I}}\right)\right]\\&=\frac{1}{K+1}\sum_{k=1}^{K+1}\beta_k\left[-\log p_{\boldsymbol{v}_k^{\text{R,I}}}+\left(1-\lambda_1\right)\times\left[\alpha+d_{\boldsymbol{v}_{\text{ak}}^{\text{R,I}},\boldsymbol{v}_{\text{pk}}^{\text{R,I}}}-d_{\boldsymbol{v}_{\text{ak}}^{\text{R,I}},\boldsymbol{v}_{\text{nk}}^{\text{R,I}}}\right]_+\!+\lambda_1\times L_{\text{hc\_tri}}\left(\boldsymbol{v}_k^{\text{R,I}}\right)\right]\end{aligned}\tag{9}$$

其中,$\beta_k=\max(\boldsymbol{m}_{\text{kp}}[k])\in[0,1]$是第$k$个关键点置信因子,而$\beta_{K+1}=1$指全局特征置信因子;$p_{\boldsymbol{v}_k^{\text{R,I}}}$是分类器预测的特征$\boldsymbol{v}_k^{\text{R,I}}$属于其真实身份的概率;$\alpha$是裕度;$d_{\boldsymbol{v}_{\text{ak}}^{\text{R,I}},\boldsymbol{v}_{\text{pk}}^{\text{R,I}}}$是来自相同身份的正样本对$(\boldsymbol{v}_{\text{ak}}^{\text{R,I}},\boldsymbol{v}_{\text{pk}}^{\text{R,I}})$之间的距离;$d_{\boldsymbol{v}_{\text{ak}}^{\text{R,I}},\boldsymbol{v}_{\text{pk}}^{\text{R,I}}}$是来自不同身份的正样本对$(\boldsymbol{v}_{\text{ak}}^{\text{R,I}},\boldsymbol{v}_{\text{pk}}^{\text{R,I}})$之间的距离;$\lambda_1$是平衡各损失的超参数,针对不同局部特征的分类器不共享.

3.2　高阶嵌入图卷积层

虽然我们有不同关键点区域的一阶语义信息,但由于可见光和红外图像属于两种不同模态的异质数据,其模态特征差异很大,而且还要面临在单模态行人重识别中仍然存在的挑战,包括背景杂乱、视点变化、姿势变化等,因此,有必要挖掘更有鉴别力的特征. 我们转向图卷积网络(GCN)方法[42,43],并尝试建模高阶关系信息. 在GCN中,将不同关键点区域的语义特征视为节点. 通过节点间的消息传递,不仅可以考虑一阶语义信息(局部关键点特征),还可以考虑高阶关系信息(边缘特征). 受图嵌入方法的启发[43~45],本文考虑图卷积网络来解决跨模态中的两类关键问题. 我们提出了一种方向自适应图内卷积(SGCL)层来动态地学习消息传递的方向和程度,用以促进鲁棒的模态不变语义特征的消息传递. 此外,在图内卷积的基础上进一步加入跨图卷积步骤,其中一个模态的局部特征可以从包含相似特征节点的另一个模态图中聚合. 希望通过对应节点交叉嵌入更好地融合两个异构图之间的信息. 最后,通过不同图卷积层的内部衔接,构成端到端可学习的图卷积网络.

3.2.1　方向自适应图内卷积层

一个简单的图卷积层有两个输入,分别为图的邻接矩阵$\boldsymbol{A}$和所有节点的特征$\boldsymbol{X}$. 输出计算公式为

$$\boldsymbol{O}=\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}\tag{10}$$

其中,$\hat{\boldsymbol{A}}$是对$\boldsymbol{A}$归一化的结果;$\boldsymbol{W}$是网络可学习的超参数.

我们对简单图卷积层进行改进,希望根据输入特征自适应学习邻接矩阵(节点的连接程度). 假设获取到K个局部特征,其中可能会有被干扰了甚至对判别产生负面影响的低价值特征,但高价值的特征会比低价值的特征更接近全局特征. 按照这种思路,我们设计了方向自适应图内卷积(SGCL)层,其输入是全局特征$\boldsymbol{V}_g$和K个局部特征$\boldsymbol{V}_l$,以及一个预定义的图(邻接矩阵是$\boldsymbol{A}$). 利用局部特征$\boldsymbol{V}_l$和全局特征$\boldsymbol{V}_g$的差异,动态更新图中所有对应边权值,得到$\boldsymbol{A}_{\text{adap}}$. 然后通过$\boldsymbol{V}_l$和$\boldsymbol{A}_{\text{adap}}$的Hadamard积可以得到一个简单的图卷积. 为了稳定训练,再将输入的局部特征$\boldsymbol{V}_l$融合到SGCL层的输出中,构成如同ResNet[38]的残差结构. 整体流程如下算法1所示.

SGCL层可以表示如下:

$$\boldsymbol{V}^{\text{out}}=\left[f_1(\boldsymbol{A}_{\text{adap}}\otimes\boldsymbol{V}_l^{\text{in}})+f_2(\boldsymbol{V}_l^{\text{in}}),\boldsymbol{V}_g^{\text{in}}\right]\tag{11}$$

其中,f_1和f_2是两个不共享的全连接层.

事实上,SGCL层算法根据图内各节点之间的关联

算法1 图内卷积层(SGCL)

Input: key-points features $\boldsymbol{V}_l^k \in \mathbb{R}_1^{K\times C}, k\in\{0,1,2,\cdots,K\}$;
resnet feature maps $\boldsymbol{V}_g \in \mathbb{R}_2^{1\times C}$;
1. //Initialization: $\boldsymbol{A}_{\text{adap}}$ as zero matrix; $\boldsymbol{A}_{\text{lim}}$ as pre-defined adjacent matrix $\boldsymbol{A}$;
2. $\boldsymbol{A}_{\text{adap}}\leftarrow\boldsymbol{0}^{K\times K}$; $\boldsymbol{A}_{\text{lim}}\leftarrow\boldsymbol{A}$; $\boldsymbol{V}_g^2\leftarrow\boldsymbol{V}_g$; $\boldsymbol{V}_l\leftarrow\boldsymbol{V}_l^1$;
3. for $k\leftarrow\{2,3,\cdots,K\}$ do
4. $\quad\boldsymbol{V}_g^k\leftarrow\text{CONC}(\boldsymbol{V}_g,\boldsymbol{V}_g^k)$;
5. $\quad\boldsymbol{V}_l\leftarrow\text{CONC}(\boldsymbol{V}_l,\boldsymbol{V}_l^k)$;
6. $\boldsymbol{A}_{\text{adap}}\leftarrow\text{ABF}(\boldsymbol{V}_g-\boldsymbol{V}_l)$;
7. $V\leftarrow(\boldsymbol{A}_{\text{adap}}\boxtimes\boldsymbol{A}_{\text{lim}})\otimes\boldsymbol{V}_l^k$;
8. $\boldsymbol{V}_{\text{out}}^k\leftarrow f(V)+f(\boldsymbol{V}_l^k), k\in\{0,1,2,\cdots,K\}$; $\boldsymbol{V}_{\text{out}}^{K+1}\leftarrow\boldsymbol{V}_g$;
Output: fusion features $\{\boldsymbol{V}_{\text{out}}^k,\boldsymbol{V}_{\text{out}}^{K+1}\}$, $k\in\{0,1,2,\cdots,K\}$.

程度重构局部特征,通过聚合高阶拓扑关系信息使网络增强身份相关特征,以减弱干扰的影响,从而得到更具有辨别力的特征. 给定一副图像$\boldsymbol{x}$,可以通过式(5)得到其语义特征$\boldsymbol{V}^E=\{\boldsymbol{v}_k^E\}_{k=1}^{K+1}$. 则其图内卷积特征$\boldsymbol{V}^S=\{\boldsymbol{v}_k^S\}_{k=1}^{K+1}$可表示为

$$\boldsymbol{V}^S=F_S\left(\boldsymbol{V}^E\right) \tag{12}$$

3.2.2 模态匹配跨图卷积层

解决跨模态行人重识别问题的关键是聚合两种模态的共同特征,减少不同模态之间的差异. 在图内卷积层中,我们获得了两种异构模态的图结构. 对于跨模态问题,可以将特有信息从一个模态聚合到另一个模态,以补偿缺乏的异构模态特征. 所提出的模态匹配跨图卷积层(CGCL)旨在通过聚合两种模态的对应节点特征来增强模态感知力. CGCL层将相同行人不同模态的子图两两配对作为嵌入特征的输入,子图由SGCL层生成,包括一阶关键点特征和高阶拓扑结构特征.

我们通过嵌入异构模态的特征来补偿模态间的差异,CGCL层的计算过程如算法2所示. 其输入是两组异构图特征,输出为模态匹配后的两组图特征,匹配过程限制在相同身份内. 首先,将两组特征$\boldsymbol{V}_1^{\text{in}}\in\mathbb{R}^{(K+1)\times C^{\text{in}}}$和$\boldsymbol{V}_2^{\text{in}}\in\mathbb{R}^{(K+1)\times C^{\text{in}}}$嵌入到一个全连接层和一个ReLU激活层的隐藏空间,得到两组隐藏特征$\boldsymbol{V}_1^{\text{fr}}\in\mathbb{R}^{(K+1)\times C^{\text{out}}}$和$\boldsymbol{V}_2^{\text{fr}}\in\mathbb{R}^{(K+1)\times C^{\text{out}}}$. 其次,通过式(18)对$\boldsymbol{V}_1^{\text{fr}}$和$\boldsymbol{V}_2^{\text{fr}}$进行图匹配,得到指派矩阵$\boldsymbol{U}^{(K+1)\times(K+1)}$,这里$\boldsymbol{U}(i,j)$表示$\boldsymbol{V}_{1\text{i}}^{\text{fr}}$和$\boldsymbol{V}_{2\text{j}}^{\text{fr}}$之间的对应关系,第$K$+1行表示图$\boldsymbol{G}_1$与图$\boldsymbol{G}_2$全局特征的匹配关系. 根据$\boldsymbol{U}$将可匹配的对应关系从一种模态聚合到另一种模态. 其中,CONC表示通道维度拼接,FR表示全连接层和ReLU激活. 具体细节如算法2所示.

最后,输出如式(13)所示:

$$\begin{aligned}\boldsymbol{V}_1^{\text{out}}&=f\left(\left[\boldsymbol{V}_1^{\text{fr}},\boldsymbol{U}\otimes\boldsymbol{V}_2^{\text{fr}}\right]\right)+\boldsymbol{V}_1^{\text{fr}}\\ \boldsymbol{V}_2^{\text{out}}&=f\left(\left[\boldsymbol{V}_2^{\text{fr}},\boldsymbol{U}\otimes\boldsymbol{V}_1^{\text{fr}}\right]\right)+\boldsymbol{V}_2^{\text{fr}}\end{aligned} \tag{13}$$

算法2 跨图卷积层(CGCL)

Input: $(k-1)$-th layer features $\{\boldsymbol{V}_{1\text{in}}^{(k)},\boldsymbol{V}_{2\text{in}}^{(k)}\}\boldsymbol{V}_{1\text{in}}^{(k)},\boldsymbol{V}_{2\text{in}}^{(k)}\in\mathbb{R}^{N\times C}$;
1. //initialization: $\boldsymbol{U}_{(K+1)\times(K+1)}$ as zero matrix;
2. $\boldsymbol{U}\leftarrow\boldsymbol{0}^{(K+1)\times(K+1)}$;
3. //build $\boldsymbol{U}$ from $\{\boldsymbol{V}_{1\text{in}}^{(k-1)},\boldsymbol{V}_{2\text{in}}^{(k-1)}\}$;
4. $\boldsymbol{V}_1^{\text{fr}}\leftarrow\text{FR}(\boldsymbol{V}_{1\text{in}}^{(k-1)})$; $\boldsymbol{V}_2^{\text{fr}}\leftarrow\text{FR}(\boldsymbol{V}_{2\text{in}}^{(k-1)})$;
5. $\boldsymbol{U}\leftarrow\text{GM}\{\boldsymbol{V}_1^{\text{fr}},\boldsymbol{V}_2^{\text{fr}}\}$;
6. $\boldsymbol{V}_1^c\leftarrow\text{CONC}\{\boldsymbol{V}_1^{\text{fr}},\boldsymbol{U}\otimes\boldsymbol{V}_2^{\text{fr}}\}$;
7. $\boldsymbol{V}_2^c\leftarrow\text{CONC}\{\boldsymbol{V}_2^{\text{fr}},\boldsymbol{U}^{\text{T}}\otimes\boldsymbol{V}_1^{\text{fr}}\}$;
8. $\boldsymbol{V}_{1\text{out}}^{(k)}\leftarrow\text{FR}(\boldsymbol{V}_1^c)+\boldsymbol{V}_1^{\text{fr}}$;
9. $\boldsymbol{V}_{2\text{out}}^{(k)}\leftarrow\text{FR}(\boldsymbol{V}_2^c)+\boldsymbol{V}_2^{\text{fr}}$;
Output: k-th cross-graph convolution features
$\{\boldsymbol{V}_{1\text{out}}^{(k)},\boldsymbol{V}_{2\text{out}}^{(k)}\}$ $\boldsymbol{V}_{1\text{out}}^{(k)},\boldsymbol{V}_{2\text{out}}^{(k)}\in\mathbb{R}^{N\times C}$

指派矩阵$\boldsymbol{U}$给出可见光图和红外图的匹配分数,包含不同模态的关系信息,以$\boldsymbol{U}$作为指导可聚合模态差异信息. 此外,由于配对过程具有随机性,跨图卷积层能更广泛地聚合异构模态信息. 给定一对跨模态图像$(\boldsymbol{x}_1,\boldsymbol{x}_2)$,可以通过式(11)得到它们的图内关系聚合特征$(\boldsymbol{V}_1^S,\boldsymbol{V}_2^S)$,则它们的跨图模态聚合特征$(\boldsymbol{V}_1^C,\boldsymbol{V}_2^C)$可用式(14)表示. 将跨图卷积层(CGCL)与图内卷积层(SGCL)级联,构成S-C-S图卷积网络. 由式(5)可以得到其输入特征对$(\boldsymbol{V}_1^E,\boldsymbol{V}_2^E)$,则图卷积网络的输出表示为

$$\boldsymbol{V}^C=F_C\left(\boldsymbol{V}^S\right) \tag{14}$$

$$\left(\boldsymbol{V}_1^{\text{GCN}},\boldsymbol{V}_2^{\text{GCN}}\right)=F_S\left\{F_C\left[F_S\left(\boldsymbol{V}_1^E,\boldsymbol{V}_2^E\right)\right]\right\} \tag{15}$$

高阶嵌入图卷积网络损失函数L_{GCN}延续使用分类损失、硬挖掘三元组损失和异质中心损失. 此外,还设计跨模态排列损失用于图匹配,更多细节在2.3节表述.

$$\begin{aligned}L_{\text{GCN}}=\frac{1}{K+1}\sum_{k=1}^{K+1}\beta_k\Big[&L_{\text{cls}}\left(\boldsymbol{v}_k^{\text{GCN}}\right)+\left(1-\lambda_2\right)\times L_{\text{tri}}\left(\boldsymbol{v}_k^{\text{GCN}}\right)\\&+\lambda_2\times L_{\text{hc\_tri}}\left(\boldsymbol{v}_k^{\text{GCN}}\right)\Big]\end{aligned} \tag{16}$$

其中,$L_{\text{cls}}(\cdot)$,$L_{\text{tri}}(\cdot)$和$L_{\text{hc_tri}}(\cdot)$的定义可以在式(8)和式(9)中找到;β_k是第k个关键点的置信因子.

3.3 图匹配及损失

3.3.1 图匹配算法

图匹配(graph matching)试图在两个或多个图(graph)结构之间,建立节点与节点的对应关系. 在计算机视觉领域,图匹配算法通常被用于求解多个图像(image)之间,关键点到关键点的匹配关系,给定来自图像$\boldsymbol{x}_1$和$\boldsymbol{x}_2$的两个图$\boldsymbol{G}_1=\left(\boldsymbol{V}_1,\boldsymbol{E}_1\right)$和$\boldsymbol{G}_2=\left(\boldsymbol{V}_2,\boldsymbol{E}_2\right)$,则图匹配旨在学习一个指派矩阵$\boldsymbol{U}\in[0,1]^{K\times K}$. 让$\boldsymbol{U}\in[0,1]^{KK\times 1}$成为一个指示向量,使得$\boldsymbol{U}_{ia}$表示$\boldsymbol{v}_{1i}$和$\boldsymbol{v}_{2a}$之间的匹配程度.

跨模态图匹配算法主要由三步组成:(1)确定待匹配对象,将待匹配图组成可见光-红外图对;(2)亲和度

度量,把图相似度建模为亲和度度量矩阵 $\boldsymbol{M}$;(3)Sinkhorn求解,将求解最佳匹配矩阵 $\boldsymbol{U}$ 看作线性指派问题,采用Sinkhorn算法精确求解,具体算法框架如图4所示.

亲和矩阵 $\boldsymbol{M}$ 由输入图像计算出的节点和边缘特征参数化得到,表示图像的深度特征层次. $\boldsymbol{M}\in\mathbb{R}^{KK\times KK}$ 是平方对称正定矩阵,$\boldsymbol{M}_{ia,jb}$ 度量每对边 $(i,j)\in\boldsymbol{E}_1$ 与 $(a,b)\in\boldsymbol{E}_2$ 的匹配程度. $\boldsymbol{M}$ 的计算方法如下:

$$\boldsymbol{M}_{ia,jb}=\exp\left(\frac{\boldsymbol{V}_{1i}^{\mathrm{GCN}^{\mathrm{T}}}\boldsymbol{A}\boldsymbol{V}_{2j}^{\mathrm{GCN}}}{\tau}\right) \tag{17}$$

对于不形成边的节点对,它们在矩阵中的对应项设置为0. 因此,$\boldsymbol{M}$ 的对角线项包含节点到节点的得分,而非对角线项包含边到边的得分. 因此,最佳指派矩阵 $\boldsymbol{U}^*$ 可以表述为

$$\boldsymbol{U}^*=\arg\max_{\boldsymbol{U}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{U},\mathrm{s.t.}\|\boldsymbol{U}\|=1 \tag{18}$$

由于图卷积层将图结构特征嵌入到了节点的特征向量中,因此式(17)得到的亲和矩阵规模是线性的[43]. 由式(18)组成的图匹配问题可以被公式化为二次指派问题. 可采用Sinkhorn算法[46]在端到端的框架中精确求解. 优化过程由幂迭代和双随机化操作构成. 由于优化过程只包含了乘、除操作,Sinkhorn算法完全可微,故能够被用于端到端的深度学习训练中. 因此,可以在我们的模型中使用自适应矩估计来优化 $\boldsymbol{U}$.

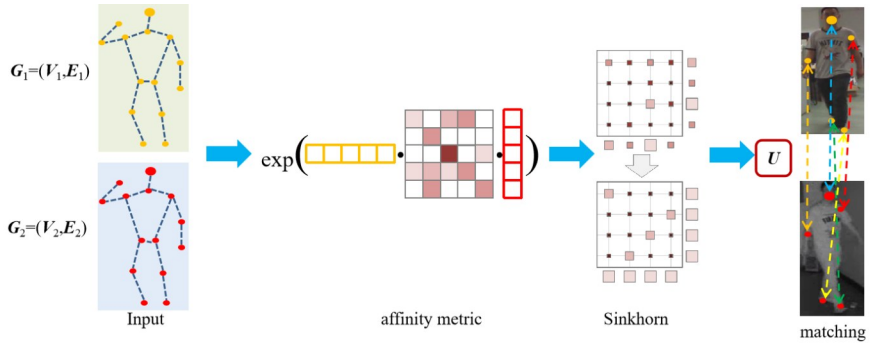

图4　图匹配算法

对于跨模态图像匹配,我们的目标是使正特征对(相同身份可见光与红外图像)的置信程度尽可能大,而负特征对(不同身份可见光与红外图像)的置信程度尽可能接近0. 可以通过跨模态排列损失函数 L_V 实现,计算方法可表示如下:

$$L_V=-\sum_{i,j}\left(\boldsymbol{U}_{i,j}^{\mathrm{gt}}\log\boldsymbol{U}_{i,j}+\left(1-\boldsymbol{U}_{i,j}^{\mathrm{gt}}\right)\log\left(1-\boldsymbol{U}_{i,j}\right)\right) \tag{19}$$

其中,$\boldsymbol{U}_{i,j}^{\mathrm{gt}}$ 是基准真值矩阵,批处理时由输入标签生成,如果 i,j 是同一个行人图像,令 $\boldsymbol{U}_{i,j}^{\mathrm{gt}}=1$,否则 $\boldsymbol{U}_{i,j}^{\mathrm{gt}}=0$.

3.3.2　总损失及相似度

在训练阶段,局部异构聚合图卷积网络的总体损失函数可表示为

$$L=L_E+\mu_1\times L_{\mathrm{GCN}}+\mu_2\times L_V \tag{20}$$

其中 μ_* 为对应模块的权重值,通过最小化 L 来端到端的训练我们的网络.

对于相似度,给定可见光图像 $\boldsymbol{x}_1$ 和红外图像 $\boldsymbol{x}_2$,通过关键点提取模块可以得到它们的一阶特征 $\boldsymbol{V}_1^E=\{\boldsymbol{v}_{1k}^E\}_{k=1}^{K+1}$ 和 $\boldsymbol{V}_2^E=\{\boldsymbol{v}_{2k}^E\}_{k=1}^{K+1}$,并用余弦距离计算它们的一阶相似性,公式定义如下:

$$S_{x1,x2}^E=\frac{1}{K+1}\sum_{k=1}^{K+1}\sqrt{\beta_{1k}\beta_{2k}}\,\mathrm{cosine}(\boldsymbol{v}_{1k}^E,\boldsymbol{v}_{2k}^E) \tag{21}$$

在得到图卷积特征对 $(\boldsymbol{V}_1^{\mathrm{GCN}},\boldsymbol{V}_2^{\mathrm{GCN}})$ 后,高阶相似度的计算方式如下:

$$S_{x1,x2}^{\mathrm{GCN}}=\sigma\left(f_s\left(-\left|\boldsymbol{V}_1^{\mathrm{GCN}}-\boldsymbol{V}_2^{\mathrm{GCN}}\right|\right)\right) \tag{22}$$

其中,$|\cdot|$ 为单元级的绝对值运算,f_s 为 $\boldsymbol{C}^{\mathrm{T}}$ 到1的全连接层,σ 为sigmoid激活函数.

最终的相似度可以通过两种相似度的结合来计算,引入平衡相似度的超参数 ω,最终计算方法如式(23)所示:

$$S=(1-\omega)S_{x1,x2}^E+\omega S_{x1,x2}^{\mathrm{GCN}} \tag{23}$$

4　实验及结果分析

4.1　数据集和评价方法

为验证本文方法的有效性,在跨模态行人重识别主流数据集RegDB[23]和SYSU-MM01[14]上评估我们的模型.

RegDB是由双摄像机系统采集的小型数据集,采用一台可见光摄像机和一台热敏摄像机拍摄而成. 这个数据集总共包含412个身份,其中每个身份有10个可见光图像和10个红外图像. 按照文献[16]的评估协议,随机选取206个身份(2 060张图像)用于训练,其余206个身份(2 060张图像)用于测试. 在测试阶段,有两种检索模式. 将可见光图像作为检索图像,同时将红外模态的图片作为被检索图像,称为可见光检

索模式(visible to thermal). 而将红外图像作为检索图像称为红外检索模式(thermal to visible). 最终结果均为10次实验的平均结果. 图5展示了两个数据集随机抽取的一些样本,用来说明两个数据集拍摄图像的差异.

SYSU-MM01数据集是跨模态行人重识别的公认权威数据集. 由中山大学校园内的4个普通RGB摄像机和2个IR摄像机所采集. SYSU-MM01数据集包含491个行人,每个行人出现在两个以上不同的相机中. 数据集共有287 628张RGB图像和15 792张红外图像,图像示例如图5右侧所示. 训练集有395个行人共32 451张图像,其中RGB图像19 659张,红外图像12 792张. 测试集包含96个行人,其中3 803幅红外图像当作被检索图像,随机抽取301幅RGB图像当作检索图像. 由于随机选择图像当作检索图像的原因,在测试阶段采用10次随机实验的平均值作为最终结果. 根据其标准评估协议,数据集包括All-Search检索模式和Indoor-Search检索模式. 对于All-Search模式,可见光相机1,2,4和5用于被检索图像集,红外相机3和6用于检索图像集. 对于Indoor-Search模式,可见光摄像机1和2(不包括室外摄像机4和5)用于被检索图像集,红外摄像机3和6用于检索图像集. 同时,遵循文献[14]的验证协议,在单幅图像命中(single-shot)设置下进行10次随机实验,把平均值作为最终结果.

图5 RegDB和SYSU-MM01数据集中的示例图像

为了评价方法的性能,使用累积匹配特性(Cumulative Matching Characteristics, CMC)、平均精度均值(Mean Average Precision, mAP)和平均反向负惩罚(mean of Inverse Negative Penalty, mINP)作为评估指标. 给定一幅被检索的行人图像,CMC中的Rank-k测量的是在前k个检索结果中出现正确跨模态行人图像的概率. mAP评估算法的平均检索性能.

此外,本文引入新的评估指标mINP[2],最早在文献[2]中提出. 对于一个真实的身份识别系统,通过算法一般会返回一个检索到的排名列表,供人工进一步调查. 目标人物不应该在从多个摄像机检索到的排名榜中被忽视,因此最难正确匹配的排名位置决定了检查人员的工作量. 在实际应用场景中,所有正确的匹配项都应具有低Rank值. 而目前广泛使用的CMC和mAP指标不能评估这一特性. 本文引入mINP,从多方面评估模型的可靠性. mINP用来衡量Re-ID算法找到最难匹配样本的效率,定义为

$$\mathrm{mINP}=\frac{1}{n}\sum_i\left(1-\mathrm{NP}_i\right)=\frac{1}{n}\sum_i\frac{|G_i|}{R_i^{\mathrm{hard}}} \tag{24}$$

其中,R_i^{hard}表示最难匹配样本的排名位置,$|G_i|$表示表示查询i次正确匹配总数.

4.2 实验细节

4.2.1 网络细节

我们采用在ImageNet上预训练的ResNet50[40]作为双流特征提取的基线网络,并删除其全局平均池化(GAP)层和全连接层. 需要注意的是,参数在第一个卷积块中是不同的,而后四个卷积块对于每个模态是共享的. 对于分类器,参照文献[47],使用一个批归一化层(BN neck)和一个全连接层,后面连接一个softmax函数. 训练时,在分类损失和三元组损失之间添加一个批归一化层[48]. 对于人体关键点检测模型,我们使用在COCO数据集[49]上预先训练的HR-Net[38],这是一个先进的2D关键点检测模型. 该模型预测了17个关键点,并对头部区域的所有关键点进行整合,最终得到K=14个关键点,包括头部、肩膀、肘部、手腕、臀部、膝盖和脚踝.

4.2.2 训练细节

我们使用Pytorch[50]来实现我们的框架. 在训练阶段,图像的大小调整为256×128,填充10像素,本文选择图像随机水平翻转和随机擦除[51]作为图像增强. 采样阶段,从训练集中随机选取N个身份标签,然后随机选取对应身份的M个可见光图像及M个红外图像,则每个训练批次大小为N×2M. 对于RegDB数据集,本文设定N=8,M=4. 对于SYSU-MM01数据集,本文设定N=6,M=6. 训练过程分为两个阶段. 第一阶段是训练2.1节中的关键点提取网络. 在这一阶段中,初始学习率为3.5×10^{-4},总共训练20轮. 第二阶段训练整体网络,加载第一阶段训练好的网络参数,之后训练图卷积、图匹配网络,以端到端的方式联合训练100轮,批大小为$2NM$,初始学习率为3.5×10^{-4},学习率在30和70轮时衰减为原来的0.1倍. 优化器采用的是Adam. 当数据集为RegDB时,超参数设定为:$\lambda_1=\lambda_2=0.3,\mu_1=2.0,\mu_2=1.0$. 当数据集为SYSU-MM01时,超参数设定为:$\lambda_1=\lambda_2=0.6,\mu_1=1.0,\mu_2=1.0$. 设计代码使用pytorch 1.6.0框架,python版本为3.7,训练过程使用NVIDIA GTX1080Ti显卡加速.

4.3　与主流方法的对比分析

为验证本方法对跨模态行人重识别的优越性,本文将所提方法与近几年该领域的主流方法在RegDB和SYSU-MM01两个数据集上进行了比较,其结果如表1和表2所示. 实验中选择的对比方法如下:双向双重限制的排序算法即BDTR[17]和eBDTR[17],双重差异减小算法D2RL[16],对齐的生成对抗网络算法AlignGAN[20],非局部注意块和权重正则化基线AGW[2],X模态辅助学习算法Xmodal[52],动态双注意力聚合学习算法DDAG[53],双模态特征混合算法CAFM[15],密集对齐算法LbA[54],模态融合及中心聚合算法MCLNet[55],跨模态变分蒸馏方法VDCM[56],双模对齐方法MPANet[57].

在RegDB数据集中,由于拍摄原因,可见光图像与红外图像的姿态差别较小,且行人数量和图像数量具有较好的对称性,故RegDB数据集的跨模态行人重识别难度较低. 本文在RegDB数据集上进行2种不同搜索模式下的实验:可见光图像搜索红外图像(Visible to Thermal,V to T)模式、红外图像搜索可见光图像(Thermal to Visible,T to V)模式. 各方法在两种不同匹配模式下的实验结果如表1所示,加粗数据表示最优结果.

表1　各方法在RegDB数据集上的实验结果对比　　单位:%

方法	Visible to Thermal					Thermal to Visible				
	Rank-1	Rank-10	Rank-20	mAP	mINP	Rank-1	Rank-10	Rank-20	mAP	mINP
BDTR[17](TIFS 19)	33.56	58.61	67.43	32.76	—	32.92	58.46	68.43	31.96	—
eBDTR[17](AAAI 19)	34.62	58.96	68.72	33.46	—	34.21	58.74	68.64	32.49	—
D2RL[16](CVPR 19)	43.40	66.10	76.30	44.10	—	—	—	—	—	—
AlignGAN[20](ICCV 19)	57.90	—	—	53.60	—	56.30	—	—	53.40	—
Xmodal[52](AAAI 20)	62.21	83.13	91.72	60.18	—	—	—	—	—	—
DDAG[53](ECCV 20)	69.34	86.19	91.49	63.46	—	68.06	85.15	90.31	61.80	—
AGW[2](TPAMI 21)	70.05	—	—	66.37	50.19	70.49	87.12	91.84	65.90	51.24
LbA[54](ICCV 21)	74.17	87.66	—	67.64	—	72.43	87.37	—	65.46	—
VDCM[56](CVPR 21)	73.20	—	—	71.60	—	71.80	—	—	70.10	—
CAFM[15](IEEESPL 21)	78.62	91.63	96.32	71.30	—	—	—	—	—	—
MCLNet[55](ICCV 21)	80.31	92.70	96.03	73.07	57.39	75.93	90.93	94.59	69.49	52.63
MPANet[57](CVPR 21)	79.27	**98.79**	**99.81**	77.61	—	79.03	**99.22**	**100.00**	77.45	—
本文方法	**80.78**	94.89	97.60	**80.55**	**65.76**	**80.48**	90.33	94.80	**79.26**	**61.90**

从表1可看出,本文方法在不同查询设置中的性能优于其他大部分方法. 对于Visible to Thermal模式,本文方法的Rank-1达到80.78%,mAP达到80.55%,分别比VDCM方法高出7.58%和8.95%,与MCLNet相比高出0.47%和7.48%. 本文方法的Rank-1和mAP均优于SOTA(MPANet),比其高出1.51%和2.94%. 对于Thermal to Visible的表现,本文方法的Rank-1为80.48%,mAP为79.26%. 这表明我们的图卷积模型对不同的评估模式是鲁棒的. 特别地,本文方法在评估指标mINP上比MCLNet表现更优越,高出8.37%. 这说明,在本文方法的检索结果中,全部正确样本更集中在排序较低的位置. 本文方法能以更快的速度检索到最难正样本,其人工排查干预的代价更小,故本文方法在检索交叉模态行人的效率更高.

SYSU-MM01数据集为一个大型的跨模态行人重识别数据集,红外图像由近红外相机拍摄,符合大多数的实际场景条件. 与文献[14]的评测协议保持一致,本文使用行人的红外图像检索可见光图像,在单幅图像命中(Single-Shot)设置下进行10次随机实验,最终实验结果是10次实验的平均值. 各方法在SYSU-MM01数据集上的对比结果如表2所示,加粗数据表示最优结果.

从表2可看出,本文方法在三种的评估指标方面领先大部分主流方法. 具体来说,在All-Search模式下,相比于使用度量学习的BDTR,DDAG,LbA等方法,及同时使用度量学习和生成对抗策略的D2RL,AlignGAN,CAFM和Xmodal等,本文方法有更高的识别性能,Rank1值和mAP值分别提高33.5%和31.65%(BDTR),6.07%和5.95%(DDAG),5.41%和4.83%(LbA),31.92%和29.77%(D2RL),18.42%和18.27%(AlignGAN),10.9%和8.24%(Xmodal). 同时,本文方法的Rank-1值明显优于AGW,新的评估指标mINP比AGW提高9.28%. 相比SOTA方法MPANet,本文方法在Rank-20和mAP上高出0.15%和1.76%,但Rank-1,Rank-10低于SOTA,这表明本文方法在度量学习上有待改进. 在Indoor-Search模式下本文方法的Rank-1达到67.92%,mAP达到66.49%,mINP达到62.92%. 与MCLNet方法相比,本文方法在Rank-10,Rank-20上更优,但Rank-1和mAP略低于该方法. 其主要原因是选取的基线模型不同以及该方法加入相机学习策略,但该方法在基线AGW上提升5.58%的Rank-1分数(3.15%的mAP分

表2 各方法在SYSU-MM01数据集上的实验结果对比 单位:%

方法	All-Search					Indoor-Search				
	Rank-1	Rank-10	Rank-20	mAP	mINP	Rank-1	Rank-10	Rank-20	mAP	mINP
BDTR[17](TIFS 19)	27.32	66.96	81.07	27.32	—	31.92	77.18	89.28	41.86	—
D2RL[16](CVPR 19)	28.90	70.60	82.40	29.20	—	—	—	—	—	—
AlignGAN[20](ICCV 19)	42.40	85.00	93.70	40.70	—	45.90	87.60	94.40	54.30	—
Xmodal[52](AAAI 20)	49.92	89.79	95.96	50.73	—	—	—	—	—	—
AGW[2](TPAMI 21)	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
DDAG[53](ECCV 20)	54.75	90.39	95.81	53.02	—	61.20	94.06	98.41	67.98	—
CAFM[15](IEEESPL 21)	55.23	90.21	95.66	52.57	—	61.26	94.19	98.24	67.94	—
LbA[54](ICCV 21)	55.41	91.12	—	54.14	—	58.46	94.13	—	66.33	—
VDCM[56](CVPR 21)	60.02	94.18	98.14	58.80	—	66.05	96.59	99.38	72.98	—
MCLNet[55](ICCV 21)	65.40	93.33	97.14	**61.98**	**47.39**	72.56	96.98	99.20	**76.58**	**72.10**
MPANet[57](CVPR 21)	**70.28**	**96.02**	98.73	57.21	—	**75.34**	**97.87**	99.05	67.85	—
本文方法	60.82	94.36	**98.88**	58.97	44.58	67.92	97.01	**99.34**	66.49	62.92

数),而本文方法总提升10.69%的Rank-1分数(13.75%的mAP分数). 上述比较表明,本文提出的联合模态内及模态间图卷积学习方法,学习到更具有区分度的特征,与SOTA相比相当具有竞争力.

4.4 消融实验

消融实验是深度学习研究中确定某种方法是否有效的最直接方式. 我们在表3中的7种不同设定下进行了广泛的实验,以评估我们提出的方法的每个组成部分:(1)关键点提取模块的有效性,(2)图内卷积层的有效性,(3)跨图卷积层的有效性,(4)图匹配模块的有效性. 所有实验均在RegDB数据集上进行,有两种评估模式,其余设定保持不变.

实验结果如表3所示,加粗数据表示最优结果. 首先,我们删除了框架中所有模块,将框架降级为单模态常用的双流特征提取网络,其中只有全局特征V_g可用,如表3索引1所示,单模态双流特征提取网络比当前的一些跨模态方法获得了更好的结果. 这表明,从单模态Re-ID中获得的一些训练技巧有助于提升方法的性能. 接下来将姿态估计网络合并到基于基线模型的训练过程中,即组成了一阶关键点提取模块,这两种评估模式的mAP分数分别提高了6.32%和8.81%. 上述提升表明,学习从整体到局部的精细化特征有利于跨模态行人重识别. 当图内卷积层被纳入训练过程时,也可以观察到类似的结果,这也证明了图内卷积层的有效性. 如表3的索引4和5所示,当我们再加入跨图卷积层时,性能会进一步提高. 然后,我们扩展了图卷积层,采用S-C-S结构的图卷积网络,发现网络性能有更进一步提升. 这表明,我们提出的图卷积网络是有用的. 并且这些组成部分是相互有利的. 最后,在索引6和7中,我们比较了图匹配模块对模型的影响. 加入图匹配后,Rank-1进一步提高0.5%,达到了最佳性能. 这充分证明,我们的方法是有效的,各个不同组成部分之间是相互协同的.

此外,我们进一步分析所提出的模型. 在这一部分中,我们进一步分析了关键点置信因子(Key-point Confidence-factors,CONF)、置信因子的归一化(Normalization of Key-point Confidence-factors,NORM)、方向自适应图内卷积(SGCL)层和模态匹配跨图卷积(CGCL)层,它们是形成一阶语义模块和高阶图卷积模块的关键组成部分. 首先,我们评测关键点置信因子的影响. 设置关键点置信因子的原因是,经过预训练的人体关键点检测模型AP为92.7%,在检测中仍可能出现错误

表3 RegDB数据集上7种不同设定的消融研究

Index	Settings						V to T		T to V	
	Base	KE	SGCL1	CGCL	SGCL2	GM	Rank-1/%	mAP/%	Rank-1/%	mAP/%
1	√	×	×	×	×	×	70.67	66.80	69.79	63.54
2	√	√	×	×	×	×	75.84	73.12	74.92	72.35
3	√	√	√	×	×	×	77.64	76.73	77.33	75.89
4	√	√	×	√	×	×	77.82	75.44	76.87	74.18
5	√	√	√	√	×	×	79.56	79.05	79.28	77.72
6	√	√	√	√	√	×	80.28	79.97	80.01	78.65
7	√	√	√	√	√	√	**80.78**	**80.55**	**80.48**	**79.26**

点(概率很小),有些检测扭曲、错位和模糊的关键点往往起到消极作用,而模态不变特征才是重要线索. 因此,在使用人体关键点检测网络时需想办法减弱检测异常带来的影响,于是我们加入可学习的置信因子策略,可根据检测差异自适应学习K个关键点最佳权重. 再通过跨模态数据集训练20轮,已基本满足实验需求.

采用关键点置信因子和未采用关键点置信因子的对比实验如表4索引1和2所示,加粗数据表示最优结果. 表4索引1直接使用原始的置信权重,而在表4索引2中展示了使用置信因子但未进行归一化的效果. 其次,去掉SGCL的时候,在式(11)用一个固定的初始邻接矩阵代替$\boldsymbol{A}_{\text{adap}}$,邻接矩阵和人的拓扑结构一致. 因此,SGCL层退化为普通的GCN层,不能抑制无用噪声信息. 随后,去掉CGCL的同时,在式(13)用全连接矩阵代替$\boldsymbol{U}$. 也就是说,图$\boldsymbol{G}_1$的每个节点都连接到图$\boldsymbol{G}_2$的所有节点. CGCL层不包含用于模态对准的高阶跨模态关系信息,退化为特征叠加模块. 实验结果如表4所示. 当移除CONF,NORM,SGCL或CGCL时,性能显著下降了2.91%,1.29%,1.80%和2.48%的Rank-1分数,mAP的值也有所下降. 实验结果表明了不同组件CONF,NORM,SGCL和CGCL对于模型性能有重要影响.

表4　分析不同的组件对性能的影响

Index	CONF	NORM	SGCL	CGCL	Rank-1/%	mAP/%
1	×	×	√	√	77.87	76.76
2	√	×	√	√	79.49	79.38
3	√	√	×	√	78.98	76.94
4	√	√	√	×	78.30	78.23
5	√	√	√	√	**80.78**	**80.55**

在分别得到行人的可见光模态和近红外模态的K个局部特征和一个全局特征后,图卷积网络结构的选择是本文研究的重点. 如图2所示,以局部特征及全局特征作为输入,通过将图内卷积层和跨图卷积层组合,可得到多样化的高阶嵌入图卷积网络. 为了确定网络的最佳结构,比较了几种不同级别的级联组合策略. 实验采用的组合方式如下:C-S(跨图卷积-图内卷积双层级联),S-C(图内卷积-跨图卷积双层级联),S-C-S(图内卷积-跨图卷积-图内卷积三层级联),C-S-C(跨图卷积-图内卷积-跨图卷积三层级联)和S-C-S-C(图内卷积-跨图卷积-图内卷积-跨图卷积四层级联).

选择在SYSU-MM01数据集的全局搜索模式下分析网络结构. 图6为相同条件下不同网络融合方式的Rank1值、mAP值和mINP值的结果对比. 由图可见,第3种S-C-S式网络融合方式的效果最佳. 原因在于这种融合方式恰好过滤了干扰因素对图卷积过程的负面影响,第二次图内卷积进一步拉进相邻节点的聚合特征,使模态间聚合了更有意义的特征. 第2种S-C方法效果次佳,虽然Rank-1性能接近,但是mAP及mINP的性能远低于S-C-S结构. 第1种C-S结构首先混合不同模态中各个局部特征的信息,使在学习图内局部特征的过程中引入过多的差异信息,因此性能反而下降. 此外,过多的直接级联,会大大增加网络的收敛难度,使特征学习过程受到破坏,网络性能严重下降. 根据对实验结果的分析,本文采用第3种S-C-S结构作为高阶图卷积网络的网络构建策略.

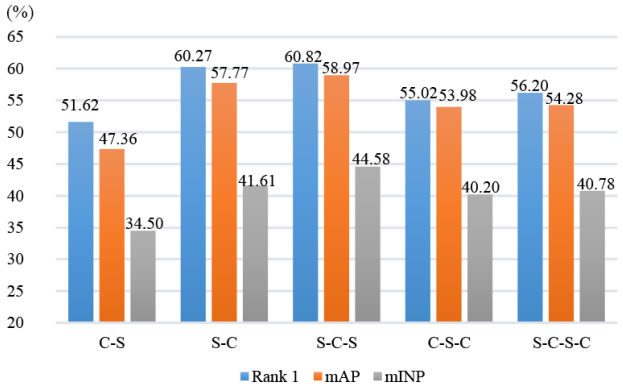

图6　在SYSU-MM01全局搜索模式下,分析不同图卷积模式的性能

4.5　超参数分析

超参数N和M分别决定训练批次中身份和样本数量. N和M对模型的影响主要有两点. 第一,在图卷积网络层中,不同行人的RGB特征与IR特征要经过跨图卷积层聚合,NM数量决定匹配矩阵$\boldsymbol{U}$的大小. 第二,行人数量N和图像数$2M$会对难样本挖掘难度和模态中心位置产生影响. 为测量采样策略对模型性能的影响,在RegDB和SYSU-MM01数据集上对不同采样策略进行实验. 为公平比较,批处理大小K小于72,而K,N和M满足条件K=2M×N. 但由于RegDB数据集只有10副RGB图像和IR图像,故对于RegDB数据集的M≤10. 图7展示了跨模态行人重识别的性能随着行人数量及图像数量的变化曲线. 可以观察到在RegDB数据集上,随着N从4增加到8,重识别性能持续上升,其中参数设定为(8,4)时的mAP比(4,8),(6,6)高出9.66%,1.39%. 当N过大时,性能开始下降. 也就是说,在N=8和M=4的情况下达到最佳精度. 而SYSU-MM01数据集在N=6,M=6时才能达到最佳效果. 产生这种现象的主要原因是,当N过小时,难样本挖掘的难度减小,而且模态中心较模糊,使网络难以学习特征空间的正确映射. 当N太大时,相同行人的图像数量会减少,RGB和IR图像关联度不够,使模型在图卷积时难以聚合相同身份的特征,不利于两种模态差异性特征的嵌入.

随后,我们评估了相似度度量超参数ω. 如图8所示,随着ω的增加,Rank-1和mAP迅速提升,然而精度

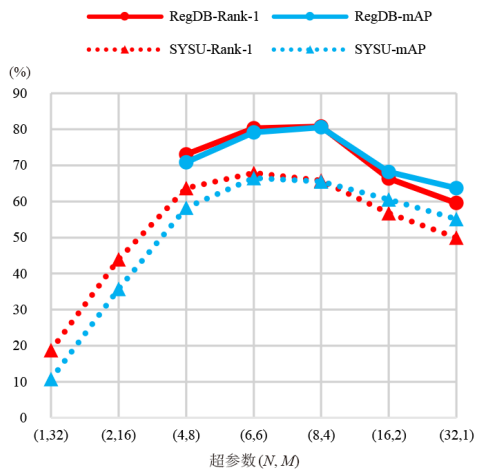

图7 在RegDB、SYSU-MM01数据集上超参数N和M对模型性能的影响

并不总是随着ω增加，当ω大于0.8时，不论怎么调整，模型的性能依然下降. 在RegDB数据集上也能观察到相同的结果. 尽管性能随不同的参数值而变化，我们的模型依然稳定地优于当前主流方法. 实验说明，一阶关键点提取模块与图卷积模块是互利的. 最佳设定为ω=0.8.

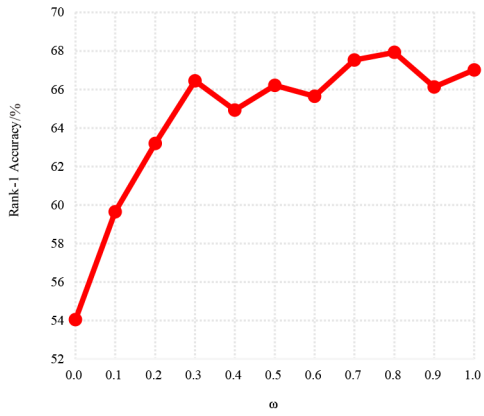

图8 在SYSU-MM01室内搜索模式下分析超参数ω

4.6 可视化分析

为了直观分析本文所提方法的重识别效果，我们可视化了RegDB和SYSU-MM01数据集上不同检索模式下12个随机选择的查询示例的前5个检索结果. 图9中的前两行是RegDB数据集的检索结果，图9中的后两行是SYSU-MM01数据集的检索结果. 首列中的图像是查询图像，检索到的图像按照相似性得分的降序从左到右排序. 图中绿框为检索正确的样本，红框为检索错误的样本.

如图9的前两行所示，RegDB数据集中可见光行人图像的衣服颜色与红外模态虽然差异较大，但行人的动作或姿态是相同的. 这需要模型更加关注行人的动作、姿态和一些细节纹理特征. 而本文利用关键点提取的方法，更加注重行人动作上的细节信息. 从检索结果可见，基于关键点提取的模型是非常有效的.

从图9中后两行可见，行人的共有特征，如头发、衣服标志依然会成为识别过程中信息匹配的关键，而这些模态共有特征可能对正确结果的判别有所帮助. 而SYSU-MM01数据集的行人背景变化较大，要求网络应该更高效关注行人细节以及它们之间的关联性，减少背景的注意力. 所以，采取图内卷积的方式，促进模态内的信息传递，对提取更有辨别力的特征十分重要. 此外，数据集中存在部分外观差异较大，行人的身体被部分遮挡等干扰图像，网络应更多地关注行人中非干扰部分，抛弃干扰部分的特征. 由图9第三行可以看出，网络对多幅遮挡和背景杂散图像进行了正确的检索，表明基于图卷积的模型可有效抑制干扰信息. 最后，由于使用联合特征来建模两种模式中相邻节点的拓扑关系，因此即使使用红外图像作为查询集，也可以获得较高的精度. 可视化结果证明了我们提出的模型的优越性.

图9 本文方法检索结果的可视化

5 结论

本文针对可见光和红外图像之间存在的类内变化和模态差异的问题，提出一种局部异构聚合图卷积网络，它由一阶关键点提取模块、高阶嵌入图卷积模块和图匹配及损失模块组成. 一阶关键点提取模块用来构建图结构，高阶嵌入图卷积模块中的图内卷积层通过聚合每个模态中不同身体特征点之间的关系，获得更具辨别力的信息. 跨图卷积层在匹配矩阵的指导下嵌入异构模态的关系信息，从而减少模态差异. 同时设计跨模态排列损失，通过多元化和多阶段的损失设计策略进一步优化模型性能. 大量实验证明，本文方法在两个公开的跨模态行人重识别数据集RegDB和SYSU-MM01上明显优于主流方法. 后续的工作将考虑在关键点提取及图匹配上展开，进一步研究最优的图构建策略，以及可见光-红外异构图的匹配问题，希望在低复杂性的前提下提高重识别率.

参考文献

[1] 罗浩, 姜伟, 范星, 等. 基于深度学习的行人重识别研究进展[J]. 自动化学报, 2019, 45(11): 2032-2049.
LUO H, JIANG W, FAN X, et al. A survey on deep learning based person re-identification[J]. Acta Automatica Sinica, 2019, 45(11): 2032-2049. (in Chinese)

[2] YE M, SHEN J B, LIN G J, et al. Deep learning for person re-identification: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872-2893.

[3] ALEXANDER H, LUCAS B, BASTIAN L. In defense of the triplet loss for person re-identification[EB/OL]. (2017-03-22)[2021-12]. https://arxiv.org/abs/1703.07737.

[4] ZHENG Z D, ZHENG L, YANG Y. A discriminatively learned CNN embedding for person reidentification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2018, 14(1): 1-20.

[5] LI W, ZHU X T, GONG S G. Person re-identification by deep joint learning of multi-loss classification[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne: AAAI Press, 2017: 2194-2200.

[6] SUN Y F, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//European Conference on Computer Vision - ECCV 2018. Munich: Springer, 2018: 501-518.

[7] TANG Z, NAPHADE M, LIU M Y, et al. CityFlow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 8789-8798.

[8] WU Y, LIN Y T, DONG X Y, et al. Progressive learning for person re-identification with one example[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2019: 2872-2881.

[9] ZHENG Z D, ZHENG L, YANG Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 3774-3782.

[10] GE Y X, LI Z W, ZHAO H Y, et al. FD-GAN: Pose-guided feature distilling GAN for robust person re-identification[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018: 1230-1241.

[11] LIU J X, NI B B, YAN Y C, et al. Pose transferrable person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4099-4108.

[12] QIAN X L, FU Y W, XIANG T, et al. Pose-normalized image generation for person re-identification[C]//European Conference on Computer Vision - ECCV 2018. Munich: Springer, 2018: 661-678.

[13] ZHENG Z D, YANG X D, YU Z D, et al. Joint discriminative and generative learning for person re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 2133-2142.

[14] WU A C, ZHENG W S, YU H X, et al. RGB-infrared cross-modality person re-identification[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 5390-5399.

[15] KONG J, HE Q B, JIANG M, et al. Dynamic center aggregation loss with mixed modality for visible-infrared person re-identification[J]. IEEE Signal Processing Letters, 2021, 28: 2003-2007.

[16] YE M, WANG Z, LAN X Y, et al. Visible thermal person re-identification via dual-constrained top-ranking[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: AAAI Press, 2018: 1092-1099.

[17] YE M, LAN X Y, WANG Z, et al. Bi-directional center-constrained top-ranking for visible thermal person re-identification[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 407-419.

[18] HAO Y, WANG N N, LI J, et al. HSME: Hypersphere manifold embedding for visible thermal person re-identification[C]//Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019: 8385-8392.

[19] ZHU Y X, YANG Z, WANG L, et al. Hetero-Center loss for cross-modality person re-identification[J]. Neurocomputing, 2020, 386: 97-109.

[20] DAI P Y, JI R R, WANG H B, et al. Cross-modality person re-identification with generative adversarial training[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: AAAI Press, 2018: 677-683.

[21] WANG Z X, WANG Z, ZHENG Y Q, et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 618-626.

[22] WANG G A, ZHANG T Z, YANG Y, et al. Cross-modal-

ity paired-images generation for RGB-infrared person re-identification[C]//Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence and the Thirty-Second Conference on Innovative Applications of Artificial Intelligence and the Tenth Symposium on Educational Advances in Artificial Intelligence. New York: AAAI Press, 2020: 12144-12151.

[23] NGUYEN D T, HONG H G, KIM K W, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors (Basel, Switzerland), 2017, 17(3): 605.

[24] 孙锐, 张磊, 余益衡, 等. 一种基于异构融合图卷积网络的跨模态行人重识别方法: CN113989851A[P]. 2022-01-28.
SUN R, ZHANG L, YU Y H, et al. Cross-modal pedestrian re-identification method based on heterogeneous fusion graph convolutional network: CN113989851A[P]. 2022-01-28. (in Chinese)

[25] HE L X, LIANG J, LI H Q, et al. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7073-7082.

[26] ZHENG L, YANG Y, ALEXANDER G H. Person re-identification: Past, present and future[EB/OL]. (2016-10-10)[2021-12]. https://arxiv.org/abs/1610.02984.

[27] WANG G A, YANG S, LIU H Y, et al. High-order information matters: Learning relation and topology for occluded person re-identification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 6448-6457.

[28] SHEN Y T, LI H S, YI S, et al. Person re--identification with deep similarity-guided graph neural network[C]//European Conference on Computer Vision - ECCV 2018. Munich: Springer, 2018: 508-526.

[29] YAN Y C, ZHANG Q, NI B B, et al. Learning context graph for person search[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 2153-2162.

[30] WU Y M, BOURAHLA O E F, LI X, et al. Adaptive graph representation learning for video person re-identification[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2020, PP: 10.1109/TIP.2020.3001693.

[31] YANG J R, ZHENG W S, YANG Q Z, et al. Spatial-temporal graph convolutional network for video-based person re-identification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 3286-3296.

[32] CHOI S, LEE S M, KIM Y, et al. Hi-CMD: Hierarchical cross-modality disentanglement for visible-infrared person re-identification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 10254-10263.

[33] YANG J W, SHEN X, TIAN X M, et al. Local convolutional neural networks for person re-identification[C]//Proceedings of the 26th ACM International Conference on Multimedia. Seoul: ACM, 2018: 1074-1082.

[34] WANG G S, YUAN Y F, CHEN X, et al. Learning discriminative features with multiple granularities for person re-identification[C]//Proceedings of the 26th ACM International conference on Multimedia. Seoul: ACM, 2018: 274-282.

[35] KALAYEH M M, BASARAN E, GÖKMEN M, et al. Human semantic parsing for person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1062-1071.

[36] FU Y, WEI Y C, ZHOU Y Q, et al. Horizontal pyramid matching for person re-identification[C]//Proceedings of The Thirty-Third AAAI Conference on Artificial Intelligence and the Thirty-First Conference on Innovative Applications of Artificial Intelligence and the Ninth Symposium on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019: 8295-8302.

[37] CAO Z, HIDALGO G, SIMON T, et al. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172-186.

[38] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 5686-5696.

[39] LIU H J, TAN X H, ZHOU X C. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification[J]. IEEE Transactions on Multimedia, 2021, 23: 4414-4425.

[40] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 770-778.

[41] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: A unified embedding for face recognition and clustering[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 815-823.

[42] BATTAGLIA P W, HAMRICK J B, BAPST V, et al. Relational inductive biases, deep learning, and graph networks[EB/OL]. (2018-06-04)[2021-12]. https://arxiv.org/abs/1806.01261.
[43] WANG R Z, YAN J C, YANG X K. Combinatorial learning of robust deep graph matching: An embedding based approach[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. DOI: 10.1109/TPAMI.2020.3005590.
[44] THOMAS N K, WELLING M. Semi-supervised classifi-cation with graph convolutional networks[EB/OL]. (2016-09-09)[2021-12]. https://arxiv.org/abs/1609.02907.
[45] ZANFIR A, SMINCHISESCU C. Deep learning of graph matching[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2684-2693.
[46] SINKHORN R. A relationship between arbitrary positive matrices and doubly stochastic matrices[J]. The Annals of Mathematical Statistics, 1964, 35(2): 876-879.
[47] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille: JMLR.org, 2015: 448-456.
[48] LUO H, GU Y Z, LIAO X Y, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE, 2019: 1487-1495.
[49] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//European Conference on Computer Vision - ECCV 2014. Zurich: Springer, 2014: 740-755.
[50] PASZKE A, GROSS S, CHINTALA S, et al. Automatic differentiation in pytorch[C]//Proceedings of the Conference and Workshop on Neural Information Processing Systems. California: NIPS, 2017: 820-828.
[51] ZHONG Z, ZHENG L, KANG G L, et al. Random erasing data augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020: 13001-13008.
[52] LI D G, WEI X, HONG X P, et al. Infrared-visible cross-modal person re-identification with an X modality[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020: 4610-4617.
[53] YE M, SHEN J B, CRANDALL D J, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]//European Conference on Computer Vision - ECCV 2020. Glasgow: Springer, 2020: 229-247.
[54] PARK H, LEE S, LEE J, et al. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 12026-12035.
[55] HAO X, ZHAO S Y, YE M, et al. Cross-modality person re-identification via modality confusion and center aggregation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 16383-16392.
[56] TIAN X D, ZHANG Z Z, LIN S H, et al. Farewell to mutual information: Variational distillation for cross-modal person re-identification[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 1522-1531.
[57] WU Q, DAI P Y, CHEN J, et al. Discover cross-modality nuances for visible-infrared person re-identification[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 4328-4337.

作者简介

孙　锐　男,1976年出生于安徽省. 现为合肥工业大学计算机与信息学院教授. 主要研究方向为计算机视觉、机器学习. 中国电子学会会员编号:E1900005402S.
E-mail: sunrui@hfut.edu.cn

张　磊(通讯作者)　男,1997年出生于安徽省. 现为合肥工业大学计算机与信息学院硕士研究生. 主要研究方向为图像信息处理、计算机视觉.
E-mail: 2020171121@mail.hfut.edu.cn

余益衡　男,1997年出生于浙江省. 现为合肥工业大学计算机与信息学院硕士研究生. 主要研究方向为图像信息处理、计算机视觉.
E-mail: 2020111040@mail.hfut.edu.cn

张旭东　男,1966年出生于安徽省. 现为合肥工业大学计算机与信息学院教授. 主要研究方向为智能信息处理、机器视觉.
E-mail: xudong@hfut.edu.cn