多分支融合变分细化蒸馏的跨模态行人重识别

王路遥[1,2,3],王凤随[1,2,3],陈元妹[1,2,3]
1. 安徽工程大学 电气工程学院,安徽 芜湖 241000
2. 检测技术与节能装置安徽省重点实验室,安徽 芜湖 241000
3. 高端装备先进感知与智能控制教育部重点实验室,安徽 芜湖 241000

摘 要:目的 针对目前跨模态行人重识别研究中对行人细腻区域关注不足以及网络易受噪声影响的问题,提出一种多分支融合变分细化蒸馏学习方法。**方法** 首先,网络通过多分支聚合不同粒度的全局特征,督促深层网络学习两种模态的全局信息和细节信息,丰富行人的特征描述符;然后,结合变分细化蒸馏策略,对特征信息进行再压缩,保留与任务相关的深层信息,同时丢弃无用的干扰物;最后,将网络捕获的不同特征用多种损失函数联合监督,以提高网络对行人表征的敏感度。**结果** 所提方法在 SYSU-MM01 数据集的全搜索模式下,R-1 和 mAP 分别达到 66.93%和 65.25%;在 RegDB 数据集的可见光到红外设置下,R-1 和 mAP 分别达到 78.26%、77.83%。**结论** 通过消融实验、对比实验和可视化实验,充分验证了所提方法的有效性。

关键词:行人重识别;跨模态;多分支聚合;变分细化蒸馏;多损失
中图分类号:TP391.4 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2024.0004.010

Cross-modal Person Re-identification Based on Multi-branch Fusion Variational Refinement Distillation

WANG Luyao[1,2,3], WANG Fengsui[1,2,3], CHEN Yuanmei[1,2,3]
1. School of Electrical Engineering, Anhui Polytechnic University, Anhui Wuhu 241000, China
2. Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Anhui Wuhu 241000, China
3. Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment, Ministry of Education, Anhui Wuhu 241000, China

Abstract: Objective Aiming at the problem of insufficient attention to the delicate area of pedestrians and the vulnerability of the network to noise in the current cross-modal person re-identification research, this paper proposed a multi-branch fusion variational refinement distillation learning method. **Methods** Firstly, the network aggregated global features of different granularity through multiple branches, urging the deep network to learn the global information and details of the two modes to enrich the feature descriptors of pedestrians. Then, combined with the variational refinement distillation strategy, the feature information was recompressed, the deep information related to the task was retained, and the useless interferences were discarded. Finally, the different features captured by the network were jointly supervised by multiple loss functions to improve the sensitivity of the network to pedestrian representation. **Results** R-1 and mAP

收稿日期:2023-04-15 **修回日期:**2023-05-25 **文章编号:**1672-058X(2024)04-0077-09
基金项目:安徽省自然科学基金(2108085MF197,1708085MF154);安徽高校省级自然科学研究重点项目(KJ2019A0162);检测技术与节能装置安徽省重点实验室开放基金资助项目(DTESD2020B02);安徽工程大学国家自然科学基金预研项目(XJKY2022040);安徽高校研究生科学研究项目(YJS20210448,YJS20210449).
作者简介:王路遥(1999—),女,安徽宿州人,硕士研究生,从事计算视觉研究.
通讯作者:王凤随(1981—),男,安徽宿州人,博士,教授,硕士生导师,从事图像与视频信息处理和计算机视觉等方面的研究. Email: fswang@ahpu.edu.cn.
引用格式:王路遥,王凤随,陈元妹. 多分支融合变分细化蒸馏的跨模态行人重识别[J]. 重庆工商大学学报(自然科学版),2024,41(4):77—85.
WANG Luyao, WANG Fengsui, CHEN Yuanmei. Cross-modal person re-identification based on multi-branch fusion variational refinement distillation[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2024, 41(4): 77—85.

reached 66.93% and 65.25%, respectively, with the proposed method in the full search mode of the SYSU-MM01 dataset; the R-1 and mAP reached 78.26% and 77.83% respectively in the visible to infrared setting of the RegDB dataset. **Conclusion** Through ablation experiments, comparative experiments, and visualization experiments, the effectiveness of the proposed method is fully verified.

Keywords: person re-identification; cross-modality; multi-branch aggregation; variational refinement distillation; multiple losses

1 引 言

随着人们安全意识的不断提升,为了预防犯罪和维护公共安全,智能监控技术引起广泛关注。行人重识别(Re-Identification, Re-ID)作为当前多方位智能监控不可或缺的技术,旨在跨设备对特定行人图像进行识别。然而,大部分的 Re-ID 研究都是在白天或光线充足的场景来进行可见光(Red Green Blue, RGB)到可见光单模态下的图像搜索[1-2],但在实际智能监控中,可见光摄像机无法应对全天候的工作。为了克服这一缺点,在夜间能够自动切换到红外(Infrared Radiation, IR)模式的摄像机应运而生,跨模态行人重识别问题被提出。与单模态的 Re-ID 不同,跨模态 Re-ID 不仅要消除所收集图像之间较大的模态差异,还要解决传统行人重识别由于不同视角、遮挡等环境因素引起的行人模态内差异的问题。

为解决以上问题,现有的跨模态相关研究主要有以下两种思路:一种是通过网络捕获两种模态下的行人特征来进行行人图像匹配:如 Ye 等[3]针对图像间跨模态差距提出 TONE 双流网络训练方法,用于提取特定模态的特征,通过特征嵌入,将其投影到公共特征空间以共享参数,缩小图像间的模态差异,为后续双流网络的研究打下基础;在双路径网络基础上,Ye 等[4]又提出一种新的双向双约束顶级损失辅助端到端学习,同时处理模态内和跨模态变化,有效缓解行人图像间模态间和模态内差异;Wu 等[5]利用模态相似性作为约束,指导跨模态相似性学习,同时减轻特定模态的信息,进一步缩短了不同模态同一行人之间的距离。上述方法通过单流或双流网络提取行人单一的粗粒度全局共享特征来处理模态差异,虽缓解了部分跨模态问题,但往往缺乏对全局特征中关键细节信息的关注,使得行人的表征判别能力不足,从而导致模型识别能力不强。另一种思路主要对图像的模态进行转换或生成新的模态,以进行行人重识别:如 Dai 等[6]为解决网络提取行人判别信息的不足,在对抗性学习框架下,提出跨模态生成对抗网络(Cross-modality Generative Adversarial Network, CmGAN),分别采用生成器和鉴别器进行对抗训练,将特征映射到公共子空间中,学习模态共享特征,提高了网络识别的精度,同时为解决跨模态 Re-ID 任务提供了新方向;Wang 等[7]提出一个对齐生成对抗网络(Alignment Generative Adversarial Network, AlignGAN),利用像素对齐将 RGB 图像转化为 IR 图像再进行特征匹配学习,以缓解模态差异,同时联合判别策略最小化异构模态图像到特征之间的差异,保证身份的一致性,使基于条件 GAN 方法的能力得到进一步提升;Li 等[8]提出 x 模态,将双模态学习任务扩展到三模态联合约束,以辅助跨模态学习,该算法不仅取得了优越的效果,网络也更轻量化。这些方法虽然在一定程度上提升了跨模态行人重识别方法的性能,但网络中对图像的风格进行转换,需要引入额外的模型参与训练,不可避免地增加噪声干扰,易破坏模型的稳定性。

故针对上述问题,本文从充分利用行人间的细节信息和弱化噪声干扰的角度出发,提出了一种多分支融合变分细化蒸馏的跨模态行人重识别方法,该方法旨在学习两种模态间粗细粒度的共享全局特征和实现最小化冗余信息的影响。在框架中,包含提取行人深层信息的双重信息聚合模块(Dual Information Aggregation Module, DIAM)和最小化冗余的变分细化蒸馏(Variational Refinement Distillation, VRD)策略。最后采用多种损失函数联合的方式对网络的不同部分进行监督学习,以提高网络模型的精度。通过在两个常用数据集上进行的一系列实验,表明所提网络框架性能优于对比方法。

2 本文算法

2.1 网络总体框架

本文网络框架基于双流的 ResNet50 骨干网络提出,适用于跨模态行人重识别算法,如图 1 所示,主要包括特征提取模块、双重信息聚合模块和变分细化蒸馏策略。双重信息聚合模块(DIAM)对特征提取模块输出的特征图进行深层细化学习,挖掘全局特征中行人更细腻的特征区域,然后通过变分细化蒸馏(VRD)最大化保留已捕获的行人表征,同时减少与任务无关

的信息,来实现最优表示。

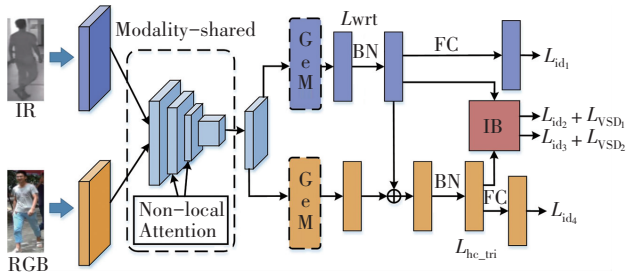

图 1　整体网络结构图

Fig. 1　Overall network structure

2.2　特征提取模块

可共享的特征提取模块包括双流的 ResNet50 网络与注意力模块。为缩小模态间的差异,将骨干网络的浅层卷积块设置为捕获特定模态的特征表示;其余 4 层为模态共享(Modality-shared)特征层,学习公共 3D 特征空间中来自异构模态的行人共享信息。由于 ResNet50 网络层数较多,为防止重要的全局语义信息丢失,在共享层的第二层和第三层中分别加入非局部注意力机制(Non-local Attention)[9],以建立两个较远距离特征像素之间的联系,增强信息共享能力(图 2),其原理如式(1)所示:

$$\boldsymbol{Z}_i=\boldsymbol{W}_Z*\varphi(\boldsymbol{X}_i)+\boldsymbol{X}_i \tag{1}$$

其中,$\boldsymbol{W}_Z$ 为待学习权重矩阵,$\varphi(\cdot)$ 为非局部学习操作,$*$ 表示卷积操作,$\boldsymbol{X}_i$ 表示输入信息,$+\boldsymbol{X}_i$ 表示残差连接。

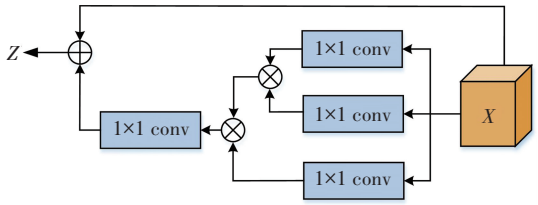

图 2　非局部注意力结构图

Fig. 2　Structure diagram of non-local attention

2.3　双重信息聚合模块

在跨模态行人重识别中,行人图像间的模态差异是影响网络性能的关键。为了减少模态差异,捕获图像的全局和关键细节信息(如行人所佩戴的眼镜、帽子及衣物上的图案等),从而获得更有判别性的特征表示就显得尤为重要。本文设计了一个双重信息聚合模块(DIAM),通过构建不同粒度的双分支对骨干网络输出的特征进行深层细化学习,以增加表征的鉴别能力,如图 3 所示。

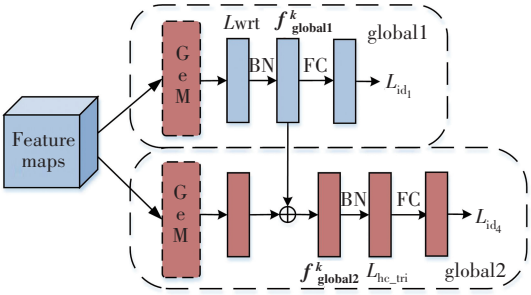

图 3　双重信息聚合模块

Fig. 3　Dual information aggregation module

DIAM 侧重于学习具有类间识别能力和类内紧凑性的高质量特征。为使网络关注行人更细腻的特征区域,在 DIAM 的双分支中采用可学习的广义均值(Generalized-Mean, GeM)池化[10]代替其他池化方式,作为细粒度检索,通过自适应的聚合特征,以产生更紧凑的图像表示,其计算如式(2)和式(3)所示:

$$\boldsymbol{f}^{(g)}=[f_1^{(g)}f_2^{(g)}\cdots f_k^{(g)}]^{\mathrm{T}} \tag{2}$$

$$f_k^{(g)}=\left(\frac{1}{|x_k|}\sum_{x\in x_k}x^{P_k}\right)^{\frac{1}{P_k}} \tag{3}$$

其中,x 为特征输入,$\boldsymbol{f}^{(g)}$ 表示对应的池化操作输出,$f_k^{(g)}$ 为特征图,P_k 表示池化参数,x_k 表示特征映射的集合。特征向量最终由每个特征图的一个值构成,其维数为 k。P_k 作为在网络反向传播中可学习的参数,当 $P_k=1$ 时,f 表示平均池化;P_k 趋近无穷大时,f 表示最大池化。

对于细粒度全局分支(global1),将 GeM 池化提取的特征输入到批处理归一化(Batch Normalization, BN)层,对特征进行归一化,得到 $\boldsymbol{f}^k_{\mathrm{global1}}$,如式(4)所示。接着将归一化后的特征送入全连接层(Fully Connected, FC),将维度从 2 048 降至 395,同时在批归一化层前后分别以不同的损失处理不同的特征向量,解决同一嵌入空间识别和度量损失不一致的问题。对于粗粒度全局分支(global2),将 GeM 池化聚合的特征与 global1 中归一化约束后的特征相结合,作为粗粒度检索得到 $\boldsymbol{f}^k_{\mathrm{global2}}$,如式(5)所示。接着将得到的特征输入到 BN 层进行归一化,最后传递给 FC 层对特征进行分类处理。DIAM 在 global1 和 global2 分支的共同作用下提取不同粒度的全局特征,避免单一分支中关键细节信息的遗漏。

$$\boldsymbol{f}^k_{\mathrm{global1}}=BN[GeM(x)] \tag{4}$$

$$\boldsymbol{f}^k_{\mathrm{global2}}=\boldsymbol{f}^k_{\mathrm{global1}}+GeM(x) \tag{5}$$

其中,x 表示特征输入,$\boldsymbol{f}^k_{\mathrm{global1}}$ 和 $\boldsymbol{f}^k_{\mathrm{global2}}$ 分别表示 global1 和 global2 中第 k 张特征图的特征向量,$GeM(\cdot)$ 为广义均值池化操作,$BN(\cdot)$ 为批量归一化操作。

2.4 变分细化蒸馏策略

为进一步提升网络模型的性能,本文将单模态中基于信息瓶颈理论的变分自蒸馏(Variational Self-Distillation, VSD)[11]损失扩展到跨模态学习,对 DIAM 提取的不同粒度信息进行再处理,设计了变分细化蒸馏(VRD)策略(图 4)。其中,信息瓶颈(Information Bottleneck, IB)为表示学习提供了一个信息理论原理,即保留所有与预测标签相关的信息,同时最小化冗余信息。虽然 IB 原理已经得到广泛应用,但其中互信息的准确估计仍然是一个挑战。VRD 策略利用 VSD 损失来拟合互信息而不估计它,以避免复杂的设计;并且通过最大化不同粒度的有用信息,同时弱化噪声的影响,来提升表征的鲁棒性。

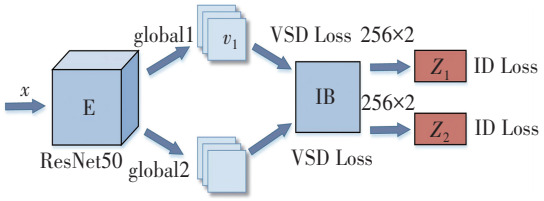

图 4 变分细化蒸馏模块

Fig. 4 Variational refinement distillation module

如图 4 所示,VRD 策略如下:首先,将两种模态的行人图像 x 输入 ResNet50 中,以学习异构模态行人的共享特征表示;然后利用 global1 和 global2 分支中 BN 层约束后的特征作为输出,以挖掘跨模态行人不同粒度的全局预测信息;接着,把不同粒度的预测信息观察值 v_1 和 v_2 输入一个共享的信息瓶颈(IB)中,结合 VSD 损失最大化保留与任务相关的深层细化信息,同时减少噪声;最后引入交叉熵损失来帮助识别行人身份,增强网络所提取行人表征的鉴别性。其中,IB 为一个多层感知器,包含两个大小分别为 1 024 和 512 的 ReLU(Rectified Linear Units)单元,其中,变分自蒸馏损失使 IB 能够掌握“表示”和“标签”之间的内在相关性,即通过最小化 $v_1(v_2)$ 中无用的信息,同时最大化 $v_1(v_2)$ 中与标签相关的信息,来保持表示 $z_1(z_2)$ 的充分性。式(6)和式(7)分别为表示 z_i 和观测 v_i 的预测信息的分布,其中 $i=1,2$。

$$P_{z_i}=p(y|z_i) \tag{6}$$

$$P_{v_i}=p(y|v_i) \tag{7}$$

x 表示输入数据,y 作为输入标签,v_i 是包含和 x 相同数量预测信息的观察值,z_i 是信息瓶颈产生的表示。

由于保持充分性的目标等价于最小化 v_i 和 z_i 预测分布之间的差异,故利用损失函数如式(8)所示,迭代的执行 L_{VSD},可实现该目标。

$$L_{\mathrm{VSD}_i}=\min_{\theta,f} E_{v_i\sim E_\theta(v_i|x)}\{E_{z_i\sim E_\varphi(z_i|v_i)}[D_{\mathrm{KL}}(P_{v_i}||P_{z_i})]\} \tag{8}$$

其中,$i=1,2$;θ 和 φ 分别表示 ResNet50 和信息瓶颈的参数;$P_{v_i}||P_{z_i}$ 表示缩小 P_{v_i} 分布与目标分布 P_{z_i} 之间的距离;D_{KL} 表示 KL 散度,对两个分布进行相似度度量;$D_{\mathrm{KL}}(P_{v_i}||P_{z_i})$ 操作为将 P_{z_i} 近似为 P_{v_i} 分布;$E[\cdot]$ 表示对分布求期望;min 表示最小化相应分布期望的相对熵。

2.5 损失函数设计

在跨模态行人重识别任务中,为进一步增强行人识别的准确度,本文使用多损失联合学习方式,利用多种损失函数的不同性质共同优化网络模型。其包括变分自蒸馏损失(L_{VSD})、交叉熵损失(L_{id})、加权三元组损失(L_{wrt})和异质中心三元组损失[12]($L_{\mathrm{hc\_tri}}$)。网络总损失 L 如式(9)所示:

$$L=L_{\mathrm{VSD}_i}+L_{\mathrm{id}_j}+L_{\mathrm{wrt}}+L_{\mathrm{hc\_tri}} \tag{9}$$

其中,$i=1,2$;$j=1,2,3,4$。交叉熵损失用来监督网络学习不随模态而改变的特定信息。它通过提取模态特定特征,以应对模态内变化,来对行人身份进行分类。L_{id} 的原理如式(10)所示:

$$L_{\mathrm{id}}=-\sum_{i=1}^{K}\ln\frac{\exp(\boldsymbol{W}_{y_i}^{\mathrm{T}}x_i+b_{y_i})}{\sum_{j=1}^{n}\exp(\boldsymbol{W}_j^{\mathrm{T}}x_i+b_j)} \tag{10}$$

其中,K 表示一个批次的大小,x_i 表示在第 y_i 类中第 i 个样本的特征,$\boldsymbol{W}_j$ 表示 $\boldsymbol{W}$ 的第 j 行参数,b 表示偏置量。

为帮助网络提高类内相似性和扩大类间差异,应对模态变化,结合加权三元组损失进行监督学习。该损失通过给正样本和负样本进行加权设计,以优化嵌入空间跨模态正负对之间的距离,达到缩小模态差异的目的。L_{wrt} 的原理如式(11)所示:

$$L_{\mathrm{wrt}}=\log\left[1+\exp\left(\sum_j w_{ij}^p d_{ij}^p-\sum_k w_{ik}^n d_{ik}^n\right)\right] \tag{11}$$

其中,i、j、k 表示每个训练批次中的难三元组。三元组是由锚点(Anchor)、正样本(Positive)和负样本(Negative)组成。对于锚点 i,P_i 和 N_i 分别为对应的正、负样本,d_{ij}^p 和 d_{ik}^n 分别表示正样本对的相对距离和负样本对的相对距离,w_{ij}^p 和 w_{ik}^n 分别表示它们的权重。

为进一步优化深度度量学习,采用异质中心三元组损失进行监督训练,同时处理跨模态差异和模态内变化。该损失利用中心距离代替样本间距离,不仅提升类内跨

模态紧凑性,学习跨模态共享特征,还使用三元组挖掘模态内和模态间的类间可分离性。故在一个批次中,各模态每个样本的中心距离如式(12)和式(13)所示:

$$c_v^i = \frac{1}{K}\sum_{j=1}^{K} v_j^i \tag{12}$$

$$c_t^i = \frac{1}{K}\sum_{j=1}^{K} t_j^i \tag{13}$$

其中,v_j^i 表示第 i 个行人的第 j 张 RGB 图像,t_j^i 表示第 i 个行人的第 j 张 IR 图像,c_v^i 表示第 i 个行人 RGB 图像的中心,c_t^i 表示第 i 个行人 IR 图像的中心。

对于每个身份,$L_{\text{hc\_tri}}$ 仅关注一个跨模态正对和挖掘在模态内与模态间最难的负对。与中心距离公式相结合,$L_{\text{hc\_tri}}$ 的原理如式(14)所示:

$$L_{\text{hc\_tri}} = \sum_{i=1}^{P}\left[m_c + \| c_v^i - c_t^i \|_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}} \| c_v^i - c_n^j \|_2\right] + \sum_{i=1}^{P}\left[m_c + \| c_t^i - c_v^i \|_2 - \min_{\substack{n\in\{v,t\}\\ j\neq i}} \| c_t^i - c_n^j \|_2\right] \tag{14}$$

其中,c_v^i 和 c_t^i 表示中心距离,来自式(11)和式(12),c_n^j 表示第 j 个行人的难负样本中心,m_c 表示异质中心三元组损失的边缘值。

3　实验结果与分析

3.1　数据集描述及评价标准

SYSU-MM01 是一个大规模跨模态 Re-ID 数据集,图像采集自在明亮和黑暗环境中工作的 6 台相机,这 6 台摄像机被分别放置在室内和室外环境中。SYSU-MM01 数据集训练时采用 395 个身份的行人图像,其中 RGB 和 IR 图像分别为 22 258 张和 11 909 张。测试时包含 96 个身份的行人图像,其中探针图像为测试集中的 3 803 张 IR 图像,图库集为 301 张 RGB 图像。本文采用全搜索和室内搜索模式进行测试,在全搜索模式下所有摄像机在该阶段使用;在室内搜索模式下只用室内摄像机来构建图库集。

RegDB 数据集由一对重叠的摄像机构建,包含 412 个行人,每个人有 10 张 RGB 图像和 10 张 IR 图像,故 RGB 图像和 IR 图像分别为 4 120 张。训练时,任意选择 206 个行人身份图像,剩下 206 个行人身份图像用于测试。训练时将数据集进行 10 次自由划分,结果取 10 次验证的平均值。测试时,采用可见光到红外和红外到可见光两种模式进行。

本文实验结果采取文献[13]中的评估指标进行评价,包括累积匹配特征(CMC)曲线,即第 n 次命中(R-n)、平均精度(mAP)和平均逆负惩罚(mINP)。

3.2　实验设置

本文实验环境:CPU 为 4 核 Intel(R) Xeon(R) Silver 4110 CPU @ 2.10 GHz,内存 16 G,显卡 NVIDIA GeForce RTX2080Ti(显存 11G),操作系统为 Ubuntu 16.04,深度学习框架 pytorch1.1.0;行人图像设置为 288×144 大小,数据增强通过对图像进行随机裁剪和翻转实现;对于采样策略,SYSU-MM01 设置了 $P=8$,$K=4$,RegDB 设置了 $P=6$,$K=4$,即在 1 个批次中,随机选取 P 个行人身份,每个身份包含 K 张 RGB 图像和 K 张 IR 图像;训练阶段 epoch 大小为 80;初始学习率为 0.1,在 20、50 个 epoch 时学习率衰减 0.1 和 0.01,在前 10 个 epoch 应用热身策略;优化器采用动量为 0.9 的 SGD 进行优化。

3.3　消融实验

本文中,在 SYSU-MM01 数据集的两种模式下进行消融实验,来评估所提网络和模块的有效性。表 1 中最后一行黑体表示本文总体框架 R-1、mAP 和 mINP 的实验结果。

如表 1 所示:DIAM 表示将网络框架设置为双重信息聚合结构;VDR1 表示仅对 global1 进行变分蒸馏;VDR2 表示仅对 global2 进行变分蒸馏;VDR 为变分细化蒸馏策略;Re-ranking 表示对网络使用 k 相互近邻算法[14]进行重排序处理。其中,Ours1 表示在 Base 基础上添加 DIAM。由实验结果可知:采用设计的 DIAM 与基线相比,模型性能得到明显提升,即在全搜索模式(室内搜索模式)下,R-1 和 mAP 分别增加了 6.16%、4.74%(3.47%、2.04%),可见 DIAM 中有效的粗细粒度信息能够被网络学习到。Ours2 表示在 Ours1 基础上对 global1 细粒度全局信息进行变分蒸馏,与 Ours1 方法相比,在全搜索模式(室内搜索模式)下 R-1 和 mAP 增加了 1.06% 和 1.45%(2.75% 和 2.78%)。Ours3 表示在 Ours1 基础上对 global2 粗粒度全局信息进行变分蒸馏,与仅有 DIAM 相比,网络精度也有一定提升。实验结果说明变分蒸馏策略无论作用在细粒度全局分支还是粗粒度全局分支,都能够削弱图像中噪声的影响。Ours4 将 DIAM 与 VRD 策略相结合,与仅添加 DIAM 相比,在全搜索模式(室内搜索模式),下 R-1 和 mAP 增加了 2.48% 和 2.33%(4.36% 和 3.94%),从而验证了变分细化蒸馏策略的重要性。最后在网络中采用 Re-ranking 重排序操作使性能得到进一步提升。

表1 SYSU-MM01数据集下的消融实验

Table 1 Ablation experiments under the SYSU-MM01 dataset

Method	DIAM	VDR1	VDR2	VRD	Re-ranking	All-search			Indoor Search		
						R-1/%	mAP/%	mINP/%	R-1/%	mAP/%	mINP/%
Base	×	×	×	×	×	47.50	47.65	35.30	54.17	62.97	59.23
Ours1	√	×	×	×	×	53.66	52.39	39.29	57.64	65.01	61.19
Ours2	√	√	×	×	×	54.72	53.84	41.14	60.39	67.79	64.20
Ours3	√	×	√	×	×	54.43	53.47	40.37	58.85	66.45	62.89
Ours4	√	×	×	√	×	56.36	54.96	42.01	62.20	69.11	65.50
Ours5	√	×	×	√	√	**66.93**	**65.25**	**53.05**	**74.72**	**78.89**	**75.89**

3.4 不同池化方式对比

如表2所示,为探索双粒度分支中不同池化方式对网络性能的影响,在SYSU-MM01数据集下设计了一系列对比实验。表2中最后一行黑体表示本文双分支选用的池化方式的实验结果。当网络仅有单分支global1时,分别采用广义平均池化和均值池化进行对比。实验发现:GeM池化相较于平均池化性能更好,说明GeM池化可捕获特定区域的鉴别特征,进行细粒度检索,故global1采用GeM池化方法。在此基础上分别对global2分支采用最大池化、平均池化和广义均值池化进行对比实验。实验发现:在双分支中都采用广义均值池化时,首张图片击中率和平均精度值最大,这说明GeM池化能够代替最大池化和平均池化,实现自适应的聚合空间信息,以此增强特征的映射能力。因此,本文采用GeM池化来构建不同粒度的双分支。

表2 不同池化方式的对比实验

Table 2 Comparative experiments of different pooling methods

Method		All-search			Indoor Search		
global1	global2	R-1/%	mAP/%	mINP/%	R-1/%	mAP/%	mINP/%
GeM	—	47.50	47.65	35.30	54.17	62.97	59.23
AVG	—	47.67	46.98	34.30	51.46	59.67	55.81
GeM	MAX	49.83	47.62	33.53	53.98	61.59	57.33
GeM	AVG	52.32	50.61	37.37	57.39	64.61	60.61
GeM	**GeM**	**53.66**	**52.39**	**39.29**	**57.64**	**65.01**	**61.19**

3.5 与其他方法进行比较

为了验证本文算法的有效性,对比实验基于两个公共数据集SYSU-MM01和RegDB,将所提算法与现有的跨模态行人重识别先进算法进行比较,主要包括TONE[3]、D^{2}RL[15]、MAC[5]、Align GAN[7]、Xmodal[8]、FBP-AL[16]、CmCEN[17]、AGW[13]等。其中"—"表示原论文中没有报告的结果,表3和表4中最后一行黑体表示本文方法分别在SYSU-MM01和RegDB数据集下R-n、mAP和mINP的实验结果。

由表3和表4可知:本文算法与基线(AGW)算法相比,在SYSU-MM01数据集全搜索模式下,R-1提升了19.43%,mAP提升了17.6%,mINP提升了17.75%;在RegDB数据集可见光到红外搜索设置下,R-1提升了8.21%,mAP提升了11.46%,mINP提升了21.6%。由于AGW算法采用的网络框架仅关注单一的全局特征,缺乏对网络细节信息的考量以及对冗余信息的处理,导致行人识别率不高。针对以上不足,本文利用不同粒度的双支路来构建网络,以提取深层全局和细节特征,增加对行人间细节信息的关注,并且对提取的全局和细节信息进行变分蒸馏,弱化冗余信息的影响,从而使网络识别精度得到提升。与Xmodal算法相比,本文算法在SYSU-MM01数据集全搜索模式下,R-1高出17.01%,mAP高出14.52%;在RegDB数据集可见光到红外搜索设置下,R-1高出16.05%,mAP高出17.65%。Xmodal算法采用x模态来辅助跨模态学习,但生成的x模态图像中可能包含新的噪声,从而影响网络模型的鲁棒性。而本文通过最大化有用信息,同时减少无关干扰的方法进行特征学习,由对比实验结果可知本文网络的鲁棒性更强。CmCEN算法与本文在同一基础上进行改进,与其相比,本文算法的效果更好,在SYSU-MM01数据集全搜索模式下,R-1高出16.84%,mAP高出15.79%,mINP高出15.88%;本文算法在RegDB数据集的两种设置下评价指标均高出CmCEN算法。CmCEN利用关键信道取代无用信道来提取更多的关键特征信息,但缺少本文算法中对行人细腻区域的关注;同时,CmCEN方法生成的额外模态特征,可能会丢弃原始特征中的有用细节信息或引入新的冗余信息,使模型识别准确度下降。本文算法利用VRD策略,在尽可能不丢失有效信息的同时弱化噪声影响,以增强表征辨别能力。

表 3　在 SYSU-MM01 数据集上和其他方法的对比实验结果

Table 3　Experimental results on SYSU-MM01 dataset compared with other methods

method	All-search					Indoor Search				
	R-1/%	R-10/%	R-20/%	mAP/%	mINP/%	R-1/%	R-10/%	R-20/%	mAP/%	mINP/%
TONE[3]	12.52	50,72	68.60	14.42	—	20.82	68.86	84.46	26.38	—
HCML[3]	14.32	53.16	69.17	16.16	—	24.52	73.25	86.73	30.08	—
BDTR[4]	27.32	66.96	81.07	27.32	—	31.92	77.18	89.28	41.86	—
D^2RL[15]	28.90	70.60	82.40	29.20	—	—	—	—	—	—
MAC[5]	33.26	79.04	90.09	36.22	—	33.37	82.49	93.69	44.95	—
DPMBN[18]	37.02	79.46	89.87	40.28	—	44.17	87.12	95.24	54.51	—
Align GAN[7]	42.40	85.00	93.70	40.70	—	45.90	87.60	94.40	54.30	—
Hi-CMD[19]	34.94	77.58	—	35.94	—	—	—	—	—	—
JSIA[20]	38.10	80.70	89.90	36.90	—	43.83	86.20	94.20	52.90	—
Xmodal[8]	49.92	89.79	95.96	50.73	—	—	—	—	—	—
DFE[21]	48.71	88.86	95.27	48.59	—	52.25	89.86	95.85	59.68	—
mtGAN-D[22]	41.10	82.40	91.90	40.50	—	—	—	—	—	—
FBP-AL[16]	54.14	86.04	93.03	50.20	—	—	—	—	—	—
CmCEN[17]	50.09	86.64	93.29	49.46	37.17	56.84	93.04	97.69	63.55	—
Baseline(AGW)[13]	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
Ours	**66.93**	**92.29**	**96.31**	**65.25**	**53.05**	**74.72**	**95.34**	**97.91**	**78.89**	**75.89**

表 4　在 RegDB 数据集上和其他方法的对比实验结果

Table 4　Comparison of experimental results in RegDB dataset with other methods

method	Visible to Infrared					Infrared to Visible				
	R-1/%	R-10/%	R-20/%	mAP/%	mINP/%	R-1/%	R-10/%	R-20/%	mAP/%	mINP/%
HCML[3]	24.44	47.53	56.78	20.08	—	21.70	45.02	55.58	22.24	—
BDTR[4]	33.56	58.61	67.43	32.76	—	32.92	58.46	68.43	31.96	—
D^2RL[15]	43.40	66.10	76.30	44.10	—	—	—	—	—	—
MAC[5]	36.43	62.36	71.63	37.03	—	36.20	61.68	70.99	36.63	—
Align GAN[7]	57.90	—	—	53.60	—	56.30	—	—	53.40	—
Xmodal[8]	62.21	83.13	91.72	60.18	—	—	—	—	—	—
DFE[21]	70.13	86.32	91.96	69.14	—	67.99	85.56	91.41	66.70	—
mtGAN-D[22]	65.60	84.20	89.60	60.00	—	65.80	85.80	91.40	59.60	—
FBP-AL[16]	73.98	89.71	93.69	68.24	—	70.05	89.22	93.88	66.61	—
CmCEN[17]	74.03	88.25	92.38	67.52	51.22	74.22	87.22	91.99	67.36	—
Baseline(AGW)[13]	70.05	86.21	91.55	66.37	50.19	70.49	87.21	91.84	65.90	51.24
Ours	**78.26**	**89.27**	**94.51**	**77.83**	**72.50**	**75.35**	**86.32**	**92.14**	**75.48**	**70.35**

3.6 可视化对比

为验证所提出方法的先进性,如图 5 所示,将基线和所提方法在 RegDB 数据集上获得的 R-10 检索结果进行可视化。其中,第一列为查询图像,分别采用可见光到红外和红外到可见光两种模式进行检索。检索到带有绿色边框的可见光图像表示与查询图像的身份相同,而带有红色边框的可见光图像表示与查询图像具有不同的身份。实验结果显示:本文方法在两种匹配模式中的表现都优于基线算法。

图 5 RegDB 数据集下基线算法与本文算法可视化对比图

Fig. 5 The visual comparison diagrams of the baseline algorithm and the algorithm in this paper under the RegDB dataset

4 结 论

本文提出一种多分支融合变分细化蒸馏的跨模态行人重识别方法。针对原始网络挖掘行人表征能力的不足和特征提取中易受图像背景杂波干扰等问题,分别提出双重信息聚合模块和变分细化蒸馏策略。前者通过捕获跨模态双粒度特征,学习行人粗糙和细腻的特征区域,增强了特征的表示能力;后者对双粒度信息进行再压缩,实现了相关信息最大化和弱化噪声的影响。最终两者相结合在多损失联合约束作用下,增强了行人检索的稳定性和准确度。通过在两个公共数据集上与目前的先进方法进行对比和消融实验,检验了所提方法的有效性。在下一步的工作中,考虑通过不同语义的信息对齐或模式对齐来缓解跨模态差异,以及如何对输入数据进行增强处理,进一步提升模型的学习能力。

参考文献(References):

[1] 徐胜军, 刘求缘, 史亚, 等. 基于多样化局部注意力网络的行人重识别[J]. 电子与信息学报, 2022, 44(1): 211—220.
XU Sheng-jun, LIU Qiu-yuan, SHI Ya, et al. Person re-identification based on diversified local attention network[J]. Journal of Electronics & Information Technology, 2022, 44(1): 211—220.

[2] 张德祥, 袁培成, 王俊. 基于多尺度批量特征丢弃网络的行人重识别研究[J]. 激光与光电子学进展, 2022, 59(12): 322—332.
ZHANG De-xiang, YUAN Pei-cheng, WANG Jun. Person reidentification based on multiscale batch feature-discarding network[J]. Laser & Optoelectronics Progress, 2022, 59(12): 322—332.

[3] YE M, LAN X, LI J, et al. Hierarchical discriminative learning for visible thermal person re-identification [C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. MenLo Park, 2018.

[4] YE M, WANG Z, LAN X, et al. Visible thermal person re-identification via dual-constrained top-ranking[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. San Framsico: Morgan Kaufmann, 2018.

[5] WU A, ZHENG W S, GONG S, et al. RGB-IR person re-identification by cross-modality similarity preservation [J]. International Journal of Computer Vision, 2020, 128(6): 1765—1785.

[6] DAI P, JI R, WANG H, et al. Cross-modality person re-identification with generative adversarial training [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. San Framsico: Morgan Kaufmann, 2018.

[7] WANG G A, ZHANG T, CHENG J, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019.

[8] 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. Menlo Park, 2020.

[9] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. NewYork: IEEE, 2018.

[10] RADENOVIC F, TOLIAS G, CHUM O. Fine-tuning CNN image retrieval with No human annotation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1655—1668.

[11] TIAN X, ZHANG Z, LIN S, et al. Farewell to mutual information: variational distillation for cross-modal person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 1522—1531.

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

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

[14] 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. NewYork: IEEE, 2017.

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

[16] WEI Z, YANG X, WANG N, et al. Flexible body partition-based adversarial learning for visible infrared person re-identification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 4676—4687.

[17] XU X, LIU S, ZHANG N, et al. Channel exchange and adversarial learning guided cross-modal person re-identification[J]. Knowledge-Based Systems, 2022, 257: 109883—109890.

[18] XIANG X, LV N, YU Z, et al. Cross-modality person re-identification based on dual-path multi-branch network [J]. IEEE Sensors Journal, 2019, 19(23): 11706—11713.

[19] CHOI S, LEE S, KIM Y, et al. Hi-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. NewYork: IEEE, 2020.

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

[21] HAO Y, WANG N, GAO X, et al. Dual-alignment feature embedding for cross-modality person re-identification [C]// Proceedings of the 27th ACM International Conference on Multimedia. Nice France; ACM, 2019.

[22] FAN X, JIANG W, LUO H, et al. Modality-transfer generative adversarial network and dual-level unified latent representation for visible thermal Person re-identification[J]. The Visual Computer, 2022, 38(1): 279—294.

责任编辑:李翠薇