模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别

石林波[1] 李华锋[1] 张亚飞[1] 谢明鸿[1]

摘 要 跨模态行人重识别方法主要通过对齐不同模态的像素分布或特征分布以缓解模态差异,却忽略具有判别性的行人细粒度信息.为了获取不受模态差异影响且更具判别性的行人特征,文中提出模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别方法.方法主要包括模态不变性特征学习模块和语义一致的细粒度信息挖掘模块,联合两个模块,使特征提取网络获取具有判别性的特征.具体地,首先利用模态不变性特征学习模块去除特征图中的模态信息,缓解模态差异.然后,使用语义一致的细粒度信息挖掘模块,对特征图分别进行通道分组和水平分块,在充分挖掘具有判别性的细粒度信息的同时实现语义对齐.实验表明,文中方法性能较优.

关键词 跨模态行人重识别, 模态差异, 细粒度信息, 语义一致性

引用格式 石林波,李华锋,张亚飞,谢明鸿.模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别.模式识别与人工智能, 2022, 35(12): 1064-1077.

DOI 10.16451/j.cnki.issn1003-6059.202212002 **中图法分类号** TP 391.41

Modal Invariance Feature Learning and Consistent Fine-Grained Information Mining Based Cross-Modal Person Re-identification

SHI Linbo[1], LI Huafeng[1], ZHANG Yafei[1], XIE Minghong[1]

ABSTRACT In the existing cross-modal person re-identification methods, modal differences are lessened by aligning features or pixel distributions of different modalities. However, the discriminative fine-grained information of pedestrians is ignored in these methods. To obtain more discriminative pedestrian features independent of modal differences, a modal invariance feature learning and consistent fine-grained information mining based cross-modal person re-identification method is proposed. The proposed method is mainly composed of two modules, modal invariance feature learning and semantically consistent fine-grained information mining. The two modules are combined to drive the feature extraction network to obtain discriminative features. Specifically, the modal invariant feature learning module is utilized to remove the modal information from the feature map to reduce the modal differences. Channel grouping and horizontal segmentation are conducted on person feature maps via the semantic consistent fine-grained information mining module. Consequently, the semantic alignment is achieved and the discriminative fine-grained information is fully mined. Experimental results show that the performance of the proposed method is significantly improved compared with the state-of-the-art cross-modal person re-identification methods.

Key Words Cross-Modal Person Re-identification, Modal Difference, Fine-Grained Information, Semantic Consistency

收稿日期:2022-08-08;录用日期:2022-11-18
Manuscript received August 8, 2022;
accepted November 18, 2022
国家自然科学基金项目(No. 62161015,62276120,61966021)资助
Supported by National Natural Science Foundation of China(No. 62161015,62276120,61966021)
本文责任编委 徐 勇
Recommended by Associate Editor XU Yong
1. 昆明理工大学 信息工程与自动化学院 昆明 650504
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504

Citation　SHI L B, LI H F, ZHANG Y F, XIE M H. Modal Invariance Feature Learning and Consistent Fine-Grained Information Mining Based Cross-Modal Person Re-identification. Pattern Recognition and Artificial Intelligence, 2022, 35(12): 1064-1077.

行人重识别[1]是判断跨相机视角拍摄的行人图像是否为同一人的技术. 受视角、姿态、光照、背景变化影响,行人重识别是一个具有挑战性的任务. 目前大多数研究[2-7]针对可见光相机捕捉的行人图像进行匹配,这是一个单模态行人重识别问题. 然而,在智能监控系统中,只有可见光摄像机是不够的. 因为当光线不足(如晚上)时,很难从可见光图像(Visible Image, VI)中提取具有判别性的行人信息. 先进的监控系统能在光照不足时自动从可见光模式切换到红外模式以捕获行人的红外图像(Infrared Image, IR),获取行人有效的外观信息.

由于成像原理不同,可见光图像和红外图像存在严重的模态差异. 因此,相比单模态行人重识别,红外-可见光跨模态行人重识别是一个极具挑战性的问题.

跨模态行人重识别的目的是将来自一个模态的查询图像(Query)与来自另一个模态的图像库(Gallery)进行匹配,在现实场景的视频监控中较重要. 相比单模态行人重识别,跨模态行人重识别面临的挑战是巨大的模态差异. 因此,如何较好地缓解模态差异是跨模态行人重识别研究的一个难点.

目前,跨模态行人重识别中缓解模态差异的方法大致分为 3 类:基于模态互转的方法[8-15]、基于度量学习的方法[16-18]和基于特征对齐的方法[19-26].

基于模态互转的方法旨在通过一种合理的方式生成当前图像在另一个模态下的图像,将跨模态行人重识别问题转化为单模态行人重识别问题,在一定程度上缓解模态间的差异. 为了进行模态互转,Wang 等[8-10]提出变分自编码器的模态互转方法,利用变分自编码器将红外图像转换为可见光图像,将可见光图像转换为红外图像. Fan 等[11]提出基于生成对抗的模态互转方法,将行人可见光图像和红外图像通过 CycleGAN(Cycle-Generative Adversarial Networks)[12]转为对应的红外图像和可见光图像,把图像统一在相同的模态下. Liu 等[13]构建频谱感知特征增强网络,将 VI 图像转换为灰度光谱图像,利用灰度光谱图像代替 VI 图像同 IR 图像进行相互检索. Li 等[14]提出 XIV-ReID(X-Infrared-Visible ReID),引入 X 模态的方法,将 VI 图像经过一个轻量型生成器转换得到 X 模态图像,这是一个介于 VI 模态和 IR 模态的中间模态,可用于缩小模态差异. Wang 等[15]提出 DPJD(Dual-Path Image Pair Joint Discriminant Model),使用模态编码器和属性编码器生成图像的模态编码和属性编码,再交换编码生成与原始图像模态不同的图像.

基于度量学习的方法关键在于如何设计合理的度量方法或损失函数,使同一行人的相同模态和不同模态图像间的距离尽可能小,不同行人的相同模态和不同模态图像间的距离尽可能大. Gao 等[16]设计 EAT(Enumerate Angular Triplet)损失和 CMKD(Cross-Modality Knowledge Distillation)损失,EAT 损失限制不同嵌入特征之间的内角,获得角度可分离的公共特征空间,CMKD 损失用于在特定的特征提取阶段结束时缩小不同模态特征之间的距离,提升跨模态行人重识别任务的有效性. Wang 等[17]提出 DPAN+CMDC Loss(Dual-Path Attention Network and Cross-Modality Dual-Constraint Loss),建立行人特征图的局部特征之间的空间依赖关系,增强网路的特征提取能力,同时,还提出跨模态双约束损失,为嵌入空间中的每个类分布添加中心和边界约束,促进类内的紧凑性,增强类间的可分离性. Hu 等[18]提出 DMiR(Adversarial Decoupling and Modality-Invariant Representation Learning),求解光谱依赖信息,优化身份信息,进一步探索跨模态行人潜在的光谱不变但具有判别性的身份表示.

基于特征对齐的方法主要思想是如何约束特征提取网络提取不同模态下图像的共有特征. 为了提取共有特征,Hao 等[19]设计双对齐(空间对齐和模态对齐)特征嵌入方法,利用行人局部特征辅助特征网络提取细粒度相机的不变信息,再引入分布损失函数和相关性损失函数,对齐 VI 模态和 IR 模态的嵌入特征. Park 等[20]提出 CMAlign,互换局部的 VI 图像特征和 IR 图像特征,约束互换后的特征在其行人类别的判断上无差别,通过这种互换思想约束网络提取两个模态的共有特征. Chen 等[21]研究神经特征搜索方法,在身份损失和三元组约束下,自动在空间和通道两个维度上选择两个模态的行人共有特征. Wu 等[22]提出双流特征提取网络(VI 模态流和 IR 模态流),先使用双流结构分别提取 VI 图像特征和 IR 图像特征,再将两个模态流的特征通过共享参数的网络获取共有特征.

综上所述,基于模态互转的方法、基于度量学习

的方法和基于特征对齐的方法旨在像素级或特征级上对齐不同模态的特征. 尽管这些方法取得不错效果,但主要关注如何缓解模态差异,未充分考虑行人的细粒度信息,因此,提取的行人特征判别性不强,效果还有待进一步改善.

为了获取更有判别性的行人信息,一些行人重识方法开始关注行人局部特征的提取. 姿态信息是定位人体不同局部部位的一个重要线索,因此近年来涌现一些基于姿态信息的行人重识别方法,利用姿态信息将输入的行人图像划分成不同部分. Su 等[27]提出 PDC(Pose-Driven Deep Convolutional Model),精准使用人体局部信息,并将整个人体和局部身体部分转换为标准化和同源状态,更好地实现特征嵌入. Zheng 等[28]通过姿态估计生成一个 PoseBox 结构,再通过仿射变换将行人与标准姿态对齐,通过 PoseBox 融合卷积神经网络(Convolutional Neural Networks, CNN)架构,减少姿态估计误差和信息丢失的影响. 此外,Tay 等[29]提出属性注意网络,将衣服颜色、头发、背包等基于物理外观的属性融入基于分类的人员重识别的框架中,增强行人特征表示. Wang 等[30]使用关节点模型提取人体 14 个局部语义特征,再利用这些局部语义特征建立关节点高阶关系信息和高阶人类拓扑关系信息,让网络学习具有鲁棒性的特征.

上述方法虽然在一定程度上挖掘行人局部信息,但存在两个主要的问题:1)需要引入外部网络或属性信息,不是端到端的学习过程,不利于在现实环境中进行部署. 2)未考虑不同模态细粒度信息语义不一致的问题,导致其无法直接应用到跨模态行人重识别任务中. 基于上述分析,在缓解模态差异的同时挖掘行人的细粒度信息,有助于网络获取更有判别性的行人特征,提升跨模态行人重识别方法的性能.

为此,本文提出模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别方法,可在挖掘细粒度信息的同时缓解模态差异. 方法的总体框架主要由模态不变性特征学习(Modal Invariance Feature Learning, MIFL)模块和语义一致的细粒度信息挖掘(Semantically Consistent Fine-Grained Information Mining, SCFIM)模块组成. MIFL 模块利用视觉 Transformer 网络[31]提取模态信息. 同时,提出模态混淆损失,利用该损失训练模态混淆分类器. 该分类器混淆模态信息和身份信息,约束特征提取网络在特征提取阶段只提取行人模态不变性特征而忽略模态信息. SCFIM 模块对特征图进行通道分组和水平分块,充分挖掘行人细粒度信息. 同时,引入语义一致性损失,约束网络提取到的不同模态同一行人的细粒度信息是语义一致的. 语义一致的行人细粒度信息更具有判别性,有助于提升网络的行人重识别性能. 在两个具有挑战性的红外-可见光跨模态行人重识别数据集(SYSU-MM01、RegDB)上的实验验证本文方法的有效性和优越性.

1 模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别方法

1.1 网络架构

在跨模态行人重识别数据集上,使用 $V=\{X_{vi}^{i}\}_{i=1}^{N_{vi}}$ 表示可见光行人图像集,$I=\{X_{ir}^{i}\}_{i=1}^{N_{ir}}$ 表示红外行人图像集,N_{vi} 表示可见光图像数量,N_{ir} 表示红外图像数量. $Y=\{y_d\}_{d=1}^{N_p}$ 表示身份标签,N_p 表示行人身份数量.

在一个批次训练数据(Batch)中,$x^{(j,i)}$ 表示第 j 个行人的第 i 个训练样本,$j=\{1,2,\cdots,B\}$,B 表示在一个批次训练数据中行人身份的个数,$i=\{1,2,\cdots,Q\}$,Q 表示在一个批次训练数据中每个行人身份抽取的样本个数.

为了缓解红外-可见光两个模态之间的差异并提取丰富的语义一致的细粒度行人特征,本文提出模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别方法,整体网络结构如图 1 所示.

本文方法主要包括模态不变性特征学习模块(MIFL)和语义一致的细粒度信息挖掘模块(SCFIM). MIFL 模块采用层数为 N 的视觉 Transformer 网络(E_T)提取模态信息,并利用模态信息训练一个模态混淆分类器. 该分类器在模态混淆损失的约束下,迫使模型学习模态不变性特征. SCFIM 模块采用 ResNet50 作为骨干网络,并将骨干网络(前 2 个块(E_L)参数不共享,后 3 个块(E_R)参数共享)提取的特征复制一份,一份经过广义平均池化(Generalized-Mean, Gem)[32]后输入 MIFL 模块,学习模态不变性特征,另一份经过通道分组和水平分块两个操作挖掘细粒度信息.

此外,引入语义对齐损失,约束网络提取到的不同模态同一行人的细粒度信息是语义一致的. 两个模块联合后训练特征提取网络,使网络提取不受模态影响、包含丰富细粒度信息、语义一致的行人特征.

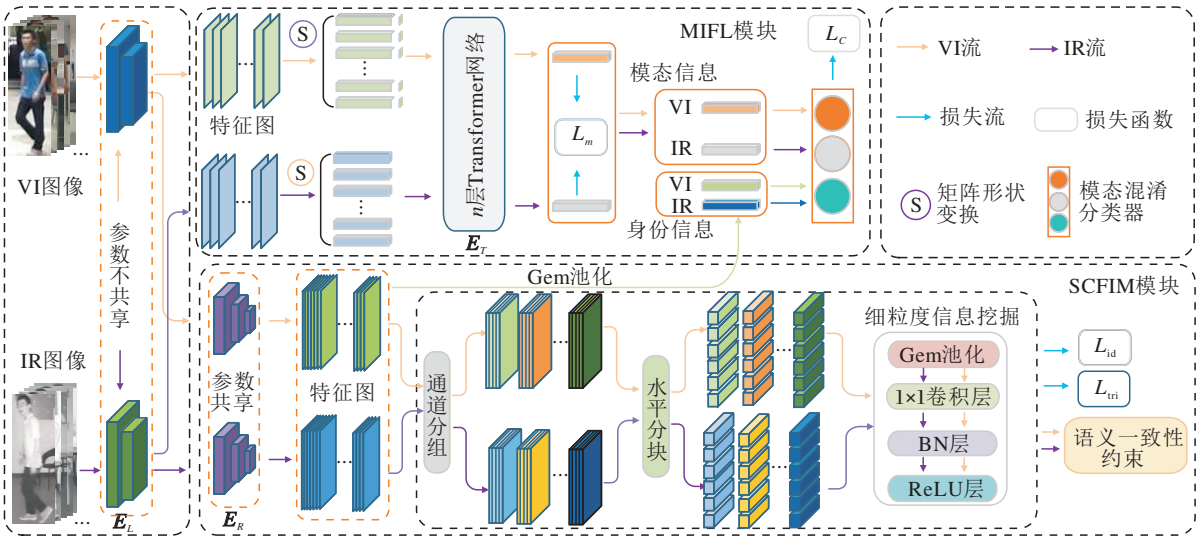


图 1　本文方法的总体框架

Fig. 1　Overall framework of the proposed method

1.2　模态不变性特征学习模块

为了缓解模态差异,本文设计模态不变性特征学习模块(MIFL).

假设输入图像为 $\boldsymbol{x}_i(i=\{vi,ir\})$,

$$\boldsymbol{M}_{vi}\in\mathbf{R}^{H_T\times W_T\times C_T},\ \boldsymbol{M}_{ir}\in\mathbf{R}^{H_T\times W_T\times C_T}$$

分别表示 $\boldsymbol{x}_{vi}$ 和 $\boldsymbol{x}_{ir}$ 经过 E_L 得到的 VI 特征图和 IR 特征图,其中,H_T、W_T 表示特征图的高、宽,C_T 表示特征图的通道. 将特征图在 $H_T\times W_T$ 维度上进行矩阵形状变换操作后得到特征

$$\boldsymbol{H}_{vi}\in\mathbf{R}^{(H_TW_T)\times C_T},\ \boldsymbol{H}_{ir}\in\mathbf{R}^{(H_TW_T)\times C_T}.$$

将 $\boldsymbol{H}_{vi}$ 和 $\boldsymbol{H}_{ir}$ 送入 E_T 学习,得到2个模态的模态信息 $\boldsymbol{m}_{vi}\in\mathbf{R}^{d_T}$ 和 $\boldsymbol{m}_{ir}\in\mathbf{R}^{d_T}$,获得模态信息的过程可表示为

$$\boldsymbol{m}_i=E_T(reshape(E_L(\boldsymbol{x}_i))),\tag{1}$$

其中,E_L 表示卷积块(ResNet50 共享参数的 2 个块),E_T 表示层数为 N 的视觉 Transformer 网络,$reshape(\cdot)$ 表示矩阵形状变换操作,$\boldsymbol{x}_i$ 表示输入图像,$\boldsymbol{m}_i$ 表示模态信息,$i=\{vi,ir\}$ 表示 VI 模态或 IR 模态. 训练时,最小化如下损失函数:

$$L_m=CE(W_m(\boldsymbol{m}_i),\boldsymbol{y}_i^m),\tag{2}$$

约束提取的模态信息. 最小化式(2)可优化 E_T 提取模态信息,其中,$W_m(\cdot)$ 表示模态分类器,$\boldsymbol{y}_i^m$ 表示模态标签,$i=\{vi,ir\}$ 表示 VI 模态或 IR 模态. 具体地,VI 模态图像的标签 $\boldsymbol{y}_{vi}^m=[1\ 0]$,IR 模态图像的标签 $\boldsymbol{y}_{ir}^m=[0\ 1]$,$CE(\cdot)$ 表示交叉熵损失函数.

得到模态信息后,利用模态信息训练模态混淆分类器,约束 E_L 和 E_R 在特征提取阶段只提取模态不变性特征,达到将模态信息从特征图中去除的目的. 具体地,图像 $\boldsymbol{x}_i$ 经过 E_L 和 E_T 得到模态信息 $\boldsymbol{m}_i$,$\boldsymbol{x}_i$ 经过 E_L 和 E_R 得到的特征图再经过 Gem 池化得到身份信息 $\boldsymbol{f}_i$. 将模态信息 $\boldsymbol{m}_i$ 和身份信息 $\boldsymbol{f}_i$ 送入模态混淆分类器,模态混淆分类器把 $\boldsymbol{m}_i$ 分到 VI 模态类或 IR 模态类,把 $\boldsymbol{f}_i$ 分到既不是 VI 模态也不是 IR 模态的第 3 个类. 在损失的约束下,E_L 和 E_R 在特征提取阶段忽略对模态信息的提取,因此 E_L 和 E_R 最终提取的是不包含模态信息的模态不变性特征. 训练时,使用如下损失函数:

$$L_c=CE(W_c(\boldsymbol{m}_i),\boldsymbol{y}_i^{m+})+CE(W_c(\boldsymbol{f}_i),\boldsymbol{y}_3^{m+}),\tag{3}$$

约束整个过程,其中,$W_c(\cdot)$ 表示模态混淆分类器,$\boldsymbol{y}_i^{m+}$ 表示模态标签,$i=\{vi,ir\}$ 表示 VI 模态或 IR 模态. 不同于式(2)的模态标签 $\boldsymbol{y}_i^m$,由于式(3)引入第 3 个类,标签信息增加一个维度. 具体地,VI 模态图像的标签形式为

$$\boldsymbol{y}_{vi}^{m+}=[1\ 0\ 0],$$

IR 模态图像的标签形式为

$$\boldsymbol{y}_{ir}^{m+}=[0\ 1\ 0],$$

$\boldsymbol{y}_3^{m+}$ 表示既不属于 VI 模态类也不属于 IR 模态类的第 3 个分类标签,标签形式为

$$\boldsymbol{y}_3^{m+}=[0\ 0\ 1].$$

1.3　语义一致的细粒度信息挖掘模块

1.3.1　细粒度信息挖掘

行人拥有的细粒度信息越丰富,就越具有判别性. 为了挖掘行人的细粒度信息,本文构建图 1 所示的细粒度信息挖掘(Fine-Grained Information Mi-

ning, FIM). 将经过 E_L 和 E_R 得到的行人特征图 $\boldsymbol{F}_i \in \mathbf{R}^{H_R\times W_R\times C_R}$ 先在通道维度上分成 L 组,得到

$$\boldsymbol{M}=\{\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_L\}\in\mathbf{R}^{H_R\times W_R\times (C_R/L)},$$

再将每组在水平维度上分成 K 块,得到

$$\boldsymbol{P}=\{\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_k\}\in\mathbf{R}^{(H_R/K)\times W_R\times (C_R/L)},$$

共得到 $L\times K$ 个细粒度块. 该过程表示如下:

$$\boldsymbol{P}=S_H(S_C(\boldsymbol{F}_i)),$$

其中,$\boldsymbol{F}_i(i=\{vi,ir\})$ 表示特征图,$S_C(\cdot)$ 表示在通道维度上的分组操作,$S_H(\cdot)$ 表示在水平维度上的分块操作.

$L\times K$ 个细粒度块经过 Gem 池化、1×1 卷积层、BN 层和 ReLU 层得到细粒度特征向量 $\boldsymbol{p}_{(l,k)}\in\mathbf{R}^{d_R}$,$l=\{1,2,\cdots,L\}$ 表示第 l 个通道分组,$k=\{1,2,\cdots,K\}$ 表示第 k 个水平块. 为了使得到的细粒度特征向量具有判别性,对其进行身份损失和中心三元组损失[33] 约束. 身份损失约束如下:

$$L_{\mathrm{id}}=-\frac{1}{BQ}\sum_{j=1}^{B}\sum_{i=1}^{Q}\sum_{l=1}^{L}\sum_{k=1}^{K}y_d\ln\left(G_{(j,i)}\left(W_{id}\left(\boldsymbol{p}_{(l,k)}^{(j,i)}\right)\right)\right),\tag{4}$$

其中,$\boldsymbol{p}_{(l,k)}^{(j,i)}$ 表示第 j 个行人的第 i 个样本的第 l 个通道分组的第 k 个水平块,$W_{id}(\cdot)$ 表示身份分类器,$G_{(j,i)}$ 表示第 j 个行人的第 i 个样本属于第 $d(d=\{1,2,\cdots,N_p\})$ 个身份的概率. 当 $d=j$ 时,$y_d=1$,当 $d\neq j$ 时,$y_d=0$.

中心三元组损失约束如下:

$$L_{\mathrm{tri}}=\sum_{i=1}^{B}\left[g+\|\boldsymbol{c}_{vi}^{i}-\boldsymbol{c}_{ir}^{i}\|_2-\min_{\substack{n\in\{vi,ir\}\\ j\neq i}}\|\boldsymbol{c}_{vi}^{i}-\boldsymbol{c}_{n}^{j}\|_2\right]_+ +\sum_{i=1}^{B}\left[g+\|\boldsymbol{c}_{ir}^{i}-\boldsymbol{c}_{vi}^{i}\|_2-\min_{\substack{n\in\{vi,ir\}\\ j\neq i}}\|\boldsymbol{c}_{ir}^{i}-\boldsymbol{c}_{n}^{j}\|_2\right]_+\tag{5}$$

其中,g 表示优化阈值,$\|\cdot\|_2$ 表示欧氏距离,

$$\boldsymbol{c}_{vi}^{i}=\frac{1}{Q}\sum_{j=1}^{Q}\boldsymbol{v}_j^i,\ \boldsymbol{c}_{ir}^{i}=\frac{1}{Q}\sum_{j=1}^{Q}\boldsymbol{t}_j^i,$$

分别表示这 B 个行人在 VI 模态和 IR 模态下的特征向量中心,$\boldsymbol{v}_j^i$ 表示 VI 模态的特征向量,$\boldsymbol{t}_j^i$ 表示 IR 模态的特征向量.

1.3.2 语义一致性约束

对每个细粒度块进行身份损失和中心三元组损失约束后,细粒度块可具有一定的判别性,但仍存在两个潜在问题:1)多个细粒度块可能表示重复的信息,2)相同行人不同模态的细粒度块表示的特征语义可能不一致. 若不解决这两个问题,会导致方法对细粒度信息挖掘不充分,对行人重识别的准确性也会产生负面影响. 为此,本文引入语义一致性约束(Semantic Consistency Constraint, SCC),结构如图 2 所示.

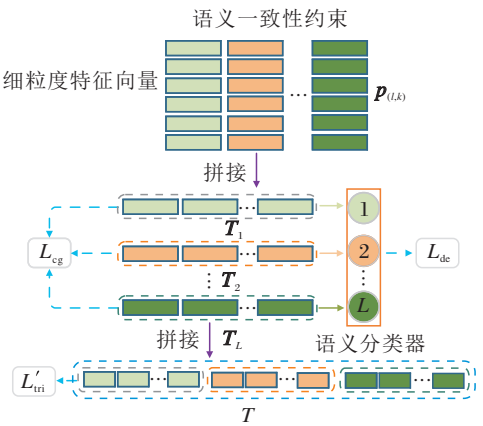


图 2 语义一致性约束结构图

Fig. 2 Semantic consistency constraint structure

拼接每个通道分组的水平块,得到

$$\boldsymbol{T}_l=[\boldsymbol{p}_{(i,1)};\boldsymbol{p}_{(i,2)};\cdots;\boldsymbol{p}_{(i,K)}]\in\mathbf{R}^{d_R},$$

再对 $\boldsymbol{T}_l(l=\{1,2,\cdots,L\})$ 进行身份损失 L_{cg}(使拼接后的水平块具有判别性)和语义分类损失 L_{de}. 在语义分类器中,设计一个固定的 L 分类标签,标签类别数与通道分组数是对应的,目的是约束相同的通道分组关注行人相同的细粒度信息. 对 $\boldsymbol{T}_l$ 进行的两个损失如下:

$$L_{\mathrm{cg}}=CE(W_{\mathrm{cg}}(\boldsymbol{T}_l),\boldsymbol{y}_d),\tag{6}$$

$$L_{\mathrm{de}}=CE(W_{\mathrm{de}}(\boldsymbol{T}_l),\boldsymbol{y}_l),\tag{7}$$

其中,$W_{\mathrm{cg}}(\cdot)$ 表示通道分组的身份分类器,$\boldsymbol{y}_d$ 表示身份标签,$W_{\mathrm{de}}(\cdot)$ 表示语义分类器,$\boldsymbol{y}_l(l=\{1,2,\cdots,L\})$ 表示一个固定的分类标签,具体的标签形式为 $\boldsymbol{y}_1=[1\ 0\ \cdots\ 0],\boldsymbol{y}_2=[0\ 1\ \cdots\ 0],\cdots,\boldsymbol{y}_L=[0\ 0\ \cdots\ 1]$.

最后,将 $\boldsymbol{T}_l(l=\{1,2,\cdots,L\})$ 按通道分组顺序拼接,表示为

$$\boldsymbol{T}=[\boldsymbol{T}_1;\boldsymbol{T}_2;\cdots;\boldsymbol{T}_L]\in\mathbf{R}^{(L\times K)\times d_R}.$$

针对 $\boldsymbol{T}$ 定义一个与式(5)定义相同的中心三元组损失,记为 L'_{tri}. 在测试阶段,$\boldsymbol{T}$ 作为行人相似性度量的信息.

1.4 总损失

综合 MIFL 模块和 SCFIM 模块,本文方法的总损失为

$$L=L_{\mathrm{id}}+L_{\mathrm{tri}}+L_{\mathrm{cg}}+\lambda_1L'_{\mathrm{tri}}+\lambda_2L_{\mathrm{de}}+\omega(L_m+L_c),\tag{8}$$

其中,λ_1、λ_2、ω 表示网络的超参数,L_{id}、L_{cg} 表示身份损失,L_{tri} 表示细粒度特征向量 $\boldsymbol{p}_{(l,k)}^{(j,i)}$ 的中心三元组

损失,$\boldsymbol{L}'_{\text{tri}}$表示 $\boldsymbol{T}$ 的中心三元组损失,L_{de} 表示语义分类损失,L_m 表示模态损失,L_c 表示模态混淆损失.

值得注意的是,损失函数 L_m 和 L_c 与其它损失函数不同,网络在最开始时不能较好地学习身份信息和模态信息,如果不对损失函数 L_m 和 L_c 的权重进行调整,方法在反向传播更新网络参数时会出现梯度爆炸情况. 因此本文提出动态训练策略,从模型训练的初始阶段到最终阶段逐渐增加损失函数 L_m 和 L_c,优化权重 ω,具体计算如下:

$$\omega = \frac{1}{1 + E(L^{t-1})},$$

其中,t 表示方法的迭代次数,$E(L^{t-1})$ 表示前一次方法迭代损失的平均值.

为了方便理解,本文方法具体步骤如算法 1 所示.

算法 1　模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别

输入　带有身份标签($Y = \{\boldsymbol{y}_d\}_{d=1}^{N_p}$) 和模态标签($\boldsymbol{y}_i, i = \{vi, ir\}$) 的 VI 图像和 IR 图像,固定的分类标签($\boldsymbol{y}_l, l = \{1,2,\cdots,L\}$)

输出　训练好的网络参数 $\boldsymbol{E}_L$ 和 $\boldsymbol{E}_R$

初始化 $\boldsymbol{E}_L, \boldsymbol{E}_T, \boldsymbol{E}_R, \boldsymbol{W}_m, \boldsymbol{W}_c, \boldsymbol{W}_{\text{id}}, \boldsymbol{W}_{\text{cg}}, \boldsymbol{W}_{\text{de}}$

for *iter* = 1,2,…,*Iteration* do

Iteration 表示模型训练时的最大迭代次数

　# 模态不变性特征学习

　最小化式(2) 更新 $\boldsymbol{E}_T$ 和 $\boldsymbol{W}_m$

　最小化式(3) 更新 $\boldsymbol{E}_L, \boldsymbol{E}_R, \boldsymbol{E}_T$ 和 $\boldsymbol{W}_c$

　# 语义一致的细粒度信息挖掘

　最小化式(4) 和式(5) 更新 $\boldsymbol{E}_L, \boldsymbol{E}_R$ 和 $\boldsymbol{W}_{id}$

　最小化式(6) 和式(7) 更新 $\boldsymbol{E}_L, \boldsymbol{E}_R, \boldsymbol{W}_{cg}$ 和 $\boldsymbol{W}_{de}$

end for

2　实验及结果分析

2.1　实验设置

本文选择 SYSU-MM01[22]、RegDB[34] 数据集进行实验.

SYSU-MM01 数据集是一个由 4 个可见光摄像机和 2 个红外光摄像机拍摄而成的数据集,有室内和室外两种环境. 训练集包含 395 个行人身份,其中可见光图像22 258 幅,红外光图像11 909 幅. 测试集包含96 个行人身份. 在测试阶段,查询图像(Query) 由 96 个行人身份的 3 903 幅红外光图像组成,图像库(Gallery) 由在 96 个行人身份中随机抽样的 301 幅单搜索 (Single-Shot) 图像或 3 010 幅多搜索 (Multi-shot) 图像组成. SYSU-M01 数据集有全搜索 (All-Search) 和室内搜索(Indoor-Search) 两种不同的测试模式. 在 All-Search 模式下, Gallery 的照片包含所有可见光摄像机拍摄的照片, 在 Indoor-Search 模式下, Gallery 的照片只包含来自室内可见光摄像机拍摄的照片. 在每个测试模式下都有 Single-Shot 和 Multi-shot 两种设置.

RegDB 数据集的图像由一个双相机系统收集而成,包括一个可见光相机和一个红外光相机,包含 412 个行人身份的 8 240 幅图像. 每个行人身份包含 10 幅可见光图像和 10 幅红外光图像. 随机选择 206 个行人身份的图像作为训练集,剩下的 206 个行人身份图像作为测试集. 测试模式包含根据可见光图像查找红外图像和根据红外图像查找可见光图像两种模式.

为了与现有方法进行公平对比,所有实验均遵循现有跨模态行人重识别方法中的常见评估设置[23,35].

在测试过程中,SYSU-M01 数据集上只有将可见光图像作为 Gallery,将红外图像作为 Query 这一种测试方式. RegDB 数据集上有两种测试方式[35]. 1)可见光图像作为 Query,红外图像作为 Gallery,即可见光图像查找红外图像(visible2infrared). 2)红外图像作为 Query,可见光图像作为 Gallery,即红外图像查找可见光图像(infrared2visible).

在 2 个公共数据集上,都采用 Rank-1,Rank-10,Rank-20 和平均精度(Mean Average Precision, mAP) 评价方法性能.

2.2　实验细节

本文与文献[23] 和文献[24] 一样,都采用在 ImageNet 数据库[36] 上进行预训练过的 ResNet50[37] 作为骨架网络. ResNet50 的 5 个块提取特征的通道数分别为 64,256,512,1 024,2 048. 提取模态特征使用层数为 N 的视觉 Transformer 网络,本文 $N = 4$. 在 SCFIM 模块中,通道分为 $L = 4$ 组,水平分为 $K = 6$ 块. 网络输入图像的尺寸统一为 288 × 144.

训练时采用随机裁剪、随机水平翻转进行数据增强,裁剪前先对图像四周使用 0 值进行扩充(扩充的具体值为 10 像素),再随机裁剪 288 × 144 的区域(如图 3 中红色虚线所示),裁剪后图像有 50% 的概率进行随机水平翻转. 数据增强的可视化过程如图 3 所示.

训练数据的批次大小(Batchsize) 设为 96,对于每个批次,在 VI 模态中随机选取 6 个行人身份,每

个行人身份选取8幅图像,IR模态进行同样选取.

本文采用Pytorch1.7学习框架,在NVIDIA RTX3090 GPU平台完成实验.采用随机梯度下降(Stochastic Gradient Descent, SGD)优化策略优化网络,动量(Momentum)设定为0.9,初始学习率设为0.01,每经过20代学习率衰减0.1,共训练60代.

具体代码发布在https://github.com/YafeiZhang-KUST/CMReID.

图3 数据增强示例

Fig. 3 Example of data enhancement

2.3 对比实验结果

本节选取如下方法进行对比实验,对比各方法在2个公共数据集上的性能,验证本文方法的有效性.具体对比方法如下.

1)基于模态互转的方法:JSIA-ReID(Joint Set-Level and Instance-Level Alignment ReID)[9], AlignGAN(Alignment GAN)[10],XIV-ReID[14],DPJD[15], Hi-CMD(Hierarchical Cross-Modality Disentanglement)[38].

2)基于度量学习的方法:EAT + CMKD[16], DPAN + CMDC Loss[17],DMiR[18],cmGAN(Cross-Modality Generative Adversarial Network)[39],CPN(Cyclic Projection Network)[40],DFLN-ViT(Discriminative Feature Learning Network Based on a Visual Transformer)[41].

3)基于特征对齐的方法:CMAlign[20], AGW(Attention Generalized Mean Pooling with Weighted Triplet Loss)[23],DDAG(Dynamic Dual-Attentive Aggregation Learning Method)[24],FBP-AL(Flexible Body Partition Model Based Adversarial Learning Method)[42],MAGC(Multi-hop Attention Graph Convolution Network)[43],CMDSF(Cross-Modality Disentanglement and Shared Feedback)[44],DML(Dual Mutual Learning)[45].

各方法在RegDB数据集上的实验结果如表1所示,表中黑体数字表示最优值,–表示原文献未提供数据.

表1 各方法在RegDB数据集上的实验结果对比

Table 1 Experimental result comparison of different methods on RegDB dataset

%

方法	visible2infrared				infrared2visible			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
AlignGAN	57.90	–	–	53.60	56.30	–	–	53.40
JSIA-ReID	48.50	–	–	49.30	48.10	–	–	48.90
XIV-ReID	–	–	–	–	62.30	–	–	60.20
DDAG	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
Hi-CMD	70.93	86.39	–	66.04	–	–	–	–
AGW	70.05	–	–	66.37	–	–	–	–
FBP-AL	73.98	89.71	93.69	68.24	70.05	89.22	93.88	66.61
CMAlign	74.17	–	–	67.64	72.43	–	–	65.46
CPN	68.59	84.81	98.33	69.20	–	–	–	–
DPJD	62.80	87.00	–	59.30	–	–	–	–
DMiR	75.79	89.86	94.18	69.97	73.93	89.87	93.98	68.22
DPAN+CMDC Loss	76.07	90.44	93.98	69.43	–	–	–	–
EAT+CMKD	79.27	–	–	77.69	80.7	–	–	79.92
MAGC	81.90	97.90	**99.20**	81.80	80.00	**98.20**	**99.60**	81.00
DML	77.60	–	–	84.30	77.00	–	–	83.60
CMDSF	84.75	92.16	96.79	77.91	–	–	–	–
DFLN-ViT	92.10	97.97	99.17	82.11	91.21	**98.20**	99.08	81.62
本文方法	**95.12**	**98.07**	99.00	**91.06**	**94.42**	97.85	98.81	**90.23**

在 visible2infrared 模式下,本文方法的 Rank-1 和 mAP(对比实验中最重要的两个性能指标) 达到最优值,比次优方法 DFLN-ViT 分别提升 3.02% 和 8.95%.

在 infrared2visible 模式下,本文方法的 Rank-1 和 mAP 也达到最优值,比 DFLN-ViT 分别提升 3.21% 和 8.61%.

为了进一步验证本文方法的有效性,在大数据集 SYSU-MM01 上进行对比实验,结果如表 2(All-Search 模式) 和表3(Indoor-Search 模式) 所示,表中黑体数字表示最优值, – 表示原文献未提供数据.

由表2 和表3 可见,本文方法在 All-Search 模式下 Rank-1 和 mAP 达到最优值,在 Indoor-Search 模式下,Single-Shot 设置下的 Rank-1 与 Multi-shot 设置下的 Rank-1 和 mAP 也达到最优值.

分析表1 ~ 表3 可知本文方法性能较优的原因如下.

1)在缓解模态差异方面,相比其它方法,本文方法创新有效. 本文利用一个视觉 Transformer 提取模态信息,再设计模态混淆分类器(两层卷积层加上一层全连接层),将模态信息和 ResNet50 提取的身份信息送入模态混淆分类器. 模态混淆分类器把模态信息分到 VI 模态类或 IR 模态类,把身份信息分到既不是 VI 模态也不是 IR 模态的第3 个类. 在损失约束下,ResNet50 在特征提取阶段忽略对模态信息的提取,因此,ResNet50 最终提取的是不包含模态信息的模态不变性特征.

2)本文的 SCFIM 模块不仅挖掘行人的细粒度信息,同时也保证挖掘信息的一致性,最终达到提升行人判别性的效果.

3)本文方法对特征不仅在水平层面上采用 PCB(Beyond Part Models)[46] 进行分块,而且还进行通道上的分组,特征划分的粒度更细,使网络提取行人更具判别性的特征.

综上所述,本文方法在两个公共数据集上取得良好效果,并且在 RegDB 数据集上的实验结果验证本文方法的优越性,在 SYSU-MM01 数据集上的实验结果验证本文方法的有效性.

表 2　各方法在 SYSU-MM01 数据集上 All-Search 模式下的实验结果对比

Table 2　Experimental results comparison of different methods in All-Search mode on SYSU-MM01 dataset

%

方法	Single-Shot				Multi-shot			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
cmGAN	26.97	67.51	80.56	27.80	21.49	72.74	85.01	22.27
AlignGAN	42.40	85.00	93.70	40.70	51.50	89.40	95.70	33.90
Hi-CMD	34.94	77.58	–	35.94	–	–	–	–
JSIA-ReID	38.10	80.70	89.90	36.90	45.10	85.70	93.80	29.50
XIV-ReID	49.92	89.79	95.96	50.73	–	–	–	–
DDAG	54.75	90.39	95.81	53.02	–	–	–	–
AGW	47.50	84.39	92.14	47.65	–	–	–	–
FBP-AL	54.14	86.04	93.03	50.20	–	–	–	–
CMAlign	55.41	–	–	54.14	–	–	–	–
CPN	57.33	92.62	97.14	56.91	63.08	93.89	97.37	50.65
MAGC	40.70	83.50	93.00	40.30	49.00	87.80	95.50	–
EAT+CMKD	43.23	**92.78**	90.91	43.09	–	–	–	–
DPJD	45.10	87.30	94.30	44.50	53.60	92.30	97.70	35.60
DMiR	50.54	88.12	94.86	49.29	–	–	–	–
DPAN+CMDC Loss	57.74	90.53	96.27	54.35	–	–	–	–
CMDSF	55.61	–	–	53.45	–	–	–	–
DML	58.40	91.20	96.90	56.10	62.20	93.40	97.80	49.60
DFLN-ViT	59.84	92.49	**97.20**	57.70	–	–	–	–
本文方法	**61.67**	91.67	96.27	**57.72**	**69.12**	**94.44**	**97.70**	**51.50**

表 3 各方法在 SYSU-MM01 数据集上 Indoor-Search 模式下的实验结果对比

Table 3 Experimental results comparison of different methods in Indoor-Search mode on SYSU-MM01 dataset

%

方法	Single-Shot				Multi-shot			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
cmGAN	31.63	77.23	89.18	42.19	37.00	80.94	92.11	32.76
AlignGAN	45.90	87.60	94.40	54.30	57.10	92.70	97.40	45.30
Hi-CMD	–	–	–	–	–	–	–	–
JSIA-ReID	43.80	86.20	94.20	52.90	52.70	91.10	96.40	42.70
XIV-ReID	–	–	–	–	–	–	–	–
DDAG	61.02	94.06	98.41	67.98	–	–	–	–
AGW	–	–	–	–	–	–	–	–
FBP-AL	58.46	–	–	66.33	–	–	–	–
CMAlign	54.17	91.14	95.98	62.97	–	–	–	–
CPN	59.30	94.46	98.38	66.70	66.26	**97.44**	**99.79**	58.51
MAGC	46.50	89.10	95.60	55.70	55.0	93.40	97.50	45.60
EAT+CMKD	50.07	90.63	96.99	58.88	–	–	–	–
DPJD	51.30	92.70	98.10	60.40	61.30	95.80	98.90	49.80
DMiR	53.92	92.50	97.09	62.49	–	–	–	–
DPAN+CMDC Loss	61.56	94.86	98.34	68.13	–	–	–	–
CMDSF	–	–	–	–	–	–	–	–
DFLN-ViT	62.13	94.83	98.24	69.03	–	–	–	–
DML	62.40	**95.20**	**98.70**	**69.50**	66.40	96.70	99.50	60.00
本文方法	**64.45**	92.29	96.34	68.97	**75.28**	95.41	98.22	**62.03**

2.4 消融实验结果

为了验证 MIFL 模块和 SCFIM 模块的有效性,将它们逐个加入基线网络中. 其中,SCFIM 模块由细粒度信息挖掘(FIM) 和语义一致性约束(SCC) 两部分组成.

本文在 SYSU-MM01、RegDB 数据集上进行消融实验. 在 SYSU-MM01 数据集上,消融实验在 All-Search 模式的 Single-Shot 设置下进行. 在 RegDB 数据集上,消融实验在 visible2infrared 模式下进行.

实验对比如下方法. 1) Baseline(基线方法). 使用 ResNet50 作为主干网络,身份损失和中心三元组损失作为损失函数. 2) Baseline+MIFL. 3) Baseline+MIFL+FIM. 4) Baseline+MIFL+FIM+SCC.

4 种方法的消融实验结果如表 4 所示. 在 SYSU-MM01 数据集上,相比 Baseline,Baseline+MIFL 的 Rank-1 和 mAP 分别提升 7.32% 和 8.03%,这意味着 MIFL 达到缓解模态差异的作用.

加入 FIM 之后,Rank-1 和 mAP 分别提升 12.19% 和 9.75%,说明 FIM 达到挖掘行人细粒度信息的目的. 加入 SCC 之后,Rank-1 和 mAP 再次提升 5.48% 和 4.8%,说明 SCC 确实起到约束行人细粒度语义一致性的作用. 在 RegDB 数据集上也能得到类似的实验结果.

总之,消融实验表明,本文方法的每个模块在缓解模态差异或提升特征的辨别性方面都起到有效作用.

表 4 各方法在 2 个数据集上的消融实验结果

Table 4 Ablation experiment results of different methods on 2 datasets

%

方法	SYSU-MM01				RegDB			
	Rank-1	Rank-10	Rank-20	mAP	Rank-1	Rank-10	Rank-20	mAP
Baseline	36.68	78.73	88.79	34.69	68.22	81.24	86.50	63.56
Baseline+MIFL	44.00	83.53	91.73	42.72	71.54	84.56	89.15	65.96
Baseline+MIFL+FIM	56.19	90.65	95.54	52.47	89.08	96.50	98.06	77.87
Baseline+MIFL+FIM+SCC	61.67	91.67	96.27	57.27	95.12	98.07	99.00	91.06

为了直观验证本文方法的有效性,利用 Grad-CAM(Visual Explanations from Deep Networks via Gradient-Based Localization)可视化方法[47],在SYSU-MM01 数据集上生成热图,具体如图 4 所示,图中颜色越深表示网络对该区域越关注.

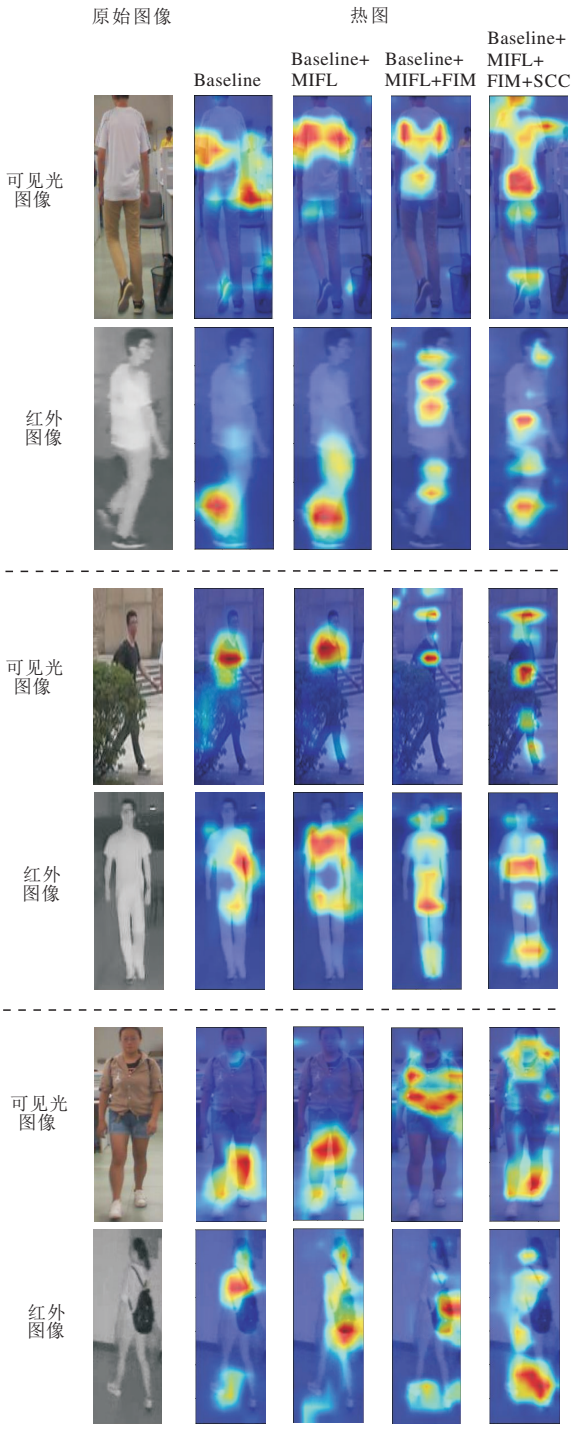


图 4　各模块生成的热图对比

Fig. 4　Comparison of heat maps generated by different modules

由图4 可看出,相比 Baseline,Baseline+MIFL 加深对模态不变性特征的学习,但此时网络只关注行人少部分区域. 加上 FIM 后,网络关注行人更多的细粒度信息,但存在多个细粒度块关注相同区域的情况和网络在两个模态下关注的行人局部区域是不同的问题,即语义不一致. 再加上 SCC 后可看出,网络关注的行人细粒度信息变多且关注的局部区域相同. 这表明加上 SCC 可解决网络在两个模态下关注信息冗余和行人语义不一致的问题.

为了进一步验证本文方法的有效性,随机选择 3 个查询实例,根据计算的余弦相识度得分,选择前 10 个检索结果,绿色框表示检索正确的图像,红色框表示检索错误的图像.

由于 SYSU-MM01 数据集上有多个检索模式,本文任选其中一个,即选择在 All-Search 模式的 Single-Shot 设置下进行,值得注意的是,该模式下正确的检索图像最多为 4 幅,具体检索结果如图 5 所示.

RegDB 数据集只有 2 种检索模式,在 2 种模式下正确检索图像最多均为 10 幅,具体检索结果分别如图 6 所示.

从图5 和图6 的可视化结果分析可得,依次加上 MIFL、FIM、SCC 模块后,检索的正确图像逐渐变多,由此再次证实本文方法的有效性.

从近期工作[48-51]中可知,行人重识别网络提取的特征图中的特征通道与行人图像的局部区域之间存在对应关系. 特征图是由输入图像或前一层的特征图利用不同卷积核进行特征映射得到的,不同的卷积核可关注行人图像不同的区域,因此特征图中不同的特征通道对行人局部区域的关注程度不同. 本文对通道进行划分,就能从特征层面关注行人的不同局部信息,因此对通道划分是合理的.

对特征进行通道和水平层面划分的优势在于 PCB 仅从人体不同空间区域的视角将身体划分成头部、躯干等不同的部分,而通道划分是从特征的视角对通道进行分组,从而使网络关注不同的人体区域. 将两者结合能更充分挖掘局部信息.

为了验证通道划分的优势,将 ResNet50 最后一层输出特征的 2 048 个特征通道分别划分为 2 组、4 组和8 组. 同时,在SYSU-MM01 数据集的 All-Search 模式下进行实验,结果如表5 所示,表中 PCB 表示仅进行水平层面的划分,CD_i 表示将特征通道划分为 $i(i=2,4,8)$ 组,黑体数字表示最优值.

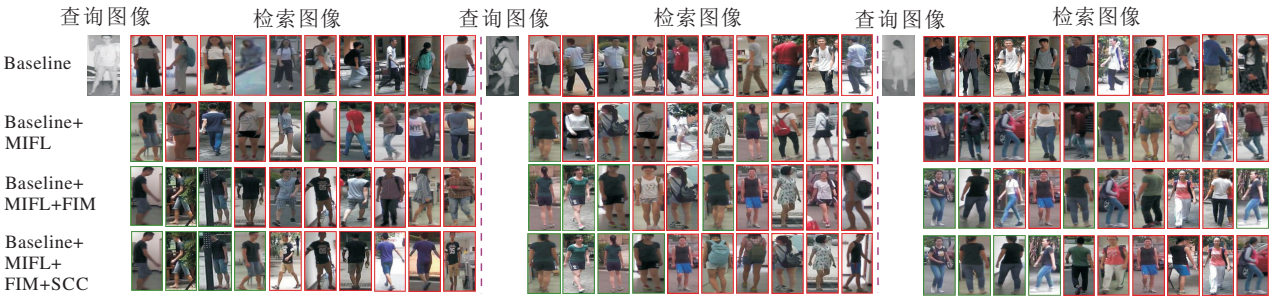
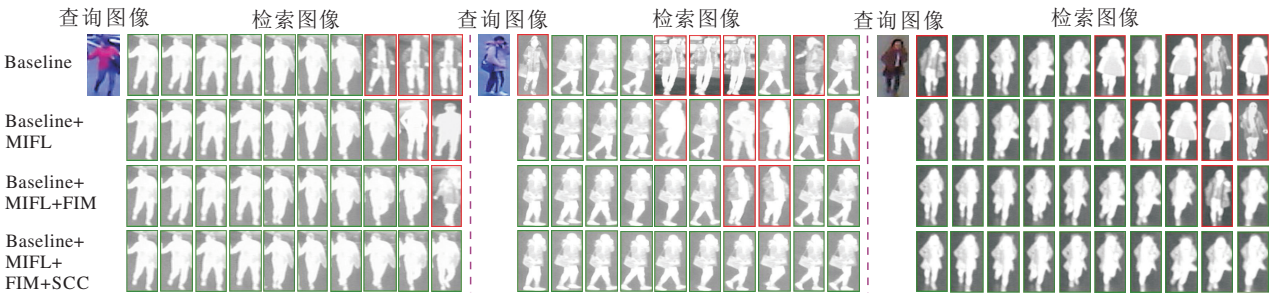
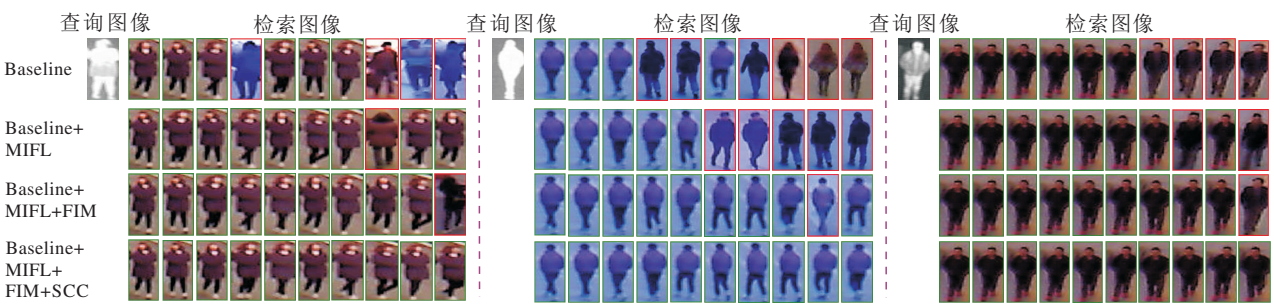


图 5 各模块在 SYSU-MM01 数据集上的检索结果

Fig. 5 Retrieval results of different modules on SYSU-MM01 dataset



(a) visible2infrared



(b) infrared2visible

图 6 各模块在 RegDB 数据集上的检索结果

Fig. 6 Retrieval results of different modules on RegDB dataset

表 5 通道划分对性能的影响

Table 5 Effect of channel segmentation on performance

通道划分	Rank-1	mAP
PCB	57.30	53.48
PCB+CD_2	59.44	56.92
PCB+CD_4	**61.67**	**57.72**
PCB+CD_8	58.19	54.70

由表 5 可见,当通道分组数为 4 时,方法性能最优. 原因是不进行通道划分或通道分组数为 2 时,网络在通道层面关注的局部区域偏少,而分组数为 8 时,会强制网络关注不重要的细粒度信息,都会造成性能的下降.

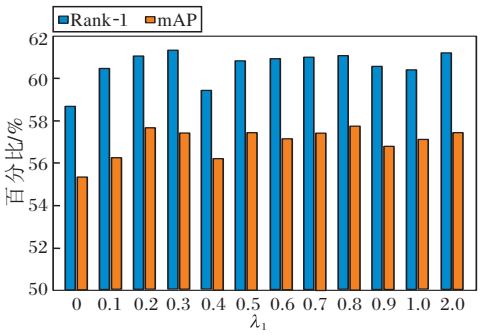
2.5 参数分析

在 SYSU-MM01 数据集的 All-Search 模式的 Single-Shot 设置下分析式(8)中 λ_1 和 λ_2 对方法性能的影响.

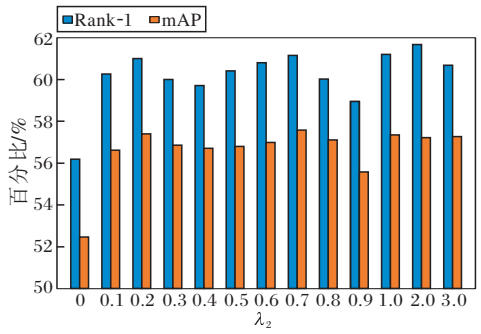
λ_1 影响三元组损失 L'_{tri} 在总损失中的占比. 设置 $\lambda_1 = 0, 0.1, 0.2, \cdots, 1.0, 2.0$, Rank-1、mAP 与 λ_1 的关系如图 7(a) 所示,当 $\lambda_1 = 0.3$ 时,Rank-1 值最高,因此,将 λ_1 设置为 0.3.

λ_2 影响语义一致性损失 L_{de} 对总损失的贡献. 设置 $\lambda_2 = 0.1, 0.2, \cdots, 1.0, 2.0, 3.0$, Rank-1、mAP

与 λ_2 的关系如图 7(b) 所示. 当 $\lambda_2=0$ 时,方法失去语义一致性损失的约束,导致性能较低,随着该损失占比逐步变大时,性能逐渐提升,当 $\lambda_2=2$ 时,性能最优. 因此,将 λ_2 设置为 2.



(a) λ_1



(b) λ_2

图 7　超参数变化对方法性能的影响

Fig. 7　Effect of variation of hyperparameters on method performance

3　结 束 语

为了解决跨模态行人重识别中存在的模态差异问题和细粒度信息挖掘不充分的问题,本文提出模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别方法,侧重于提取模态不变性的语义一致的细粒度特征. 具体地,使用模态不变性特征学习模块去除特征图中的模态信息,缓解模态差异,在使用语义一致的细粒度信息挖掘模块挖掘行人细粒度信息的同时保持挖掘的语义一致性. 在两个公共的红外-可见光跨模态数据集(SYSU-MM01 和 RegDB) 上的实验表明,本文方法性能较优. 此外,本文方法是一种端到端的网络,不需要借助例如关节点提取、GAN 这样的外部网络. 这不仅大幅降低网络的复杂性,而且还有利于在现实场景中进行部署. 今后将在提升特征的判别性方面上进行进一步探索.

参　考　文　献

[1] GONG S G, CRISTANI M, LOY C C, *et al*. The Re-identification Challenge // GONG S G, CRISTANI M, YAN S C, *et al*., eds. Person Re-identification. Berlin, Germany: Springer, 2014: 1-20.

[2] WANG J Y, ZHU X T, GONG S G, *et al*. Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-identification // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2018: 2275-2284.

[3] SONG J F, YANG Y X, SONG Y Z, *et al*. Generalizable Person Re-identification by Domain-Invariant Mapping Network // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2019: 719-728.

[4] JIN X, LAN C L, ZENG W J, *et al*. Style Normalization and Restitution for Generalizable Person Re-identification // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2020: 3140-3149.

[5] 李玲莉,谢明鸿,李 凡,等. 低秩先验引导的无监督域自适应行人重识别. 重庆大学学报, 2021, 44(11): 57-70.
(LI L L, XIE M H, LI F, *et al*. Unsupervised Domain Adaptive Person Re-identification Guided by Low-Rank Priori. Journal of Chongqing University, 2021, 44(11): 57-70.)

[6] 郑爱华,曾小强,江 波,等. 基于局部异质协同双路网络的跨模态行人重识别. 模式识别与人工智能, 2020, 33(10): 867-878.
(ZHENG A H, ZENG X Q, JIANG B, *et al*. Cross-Modal Person Re-identification Based on Local Heterogeneous Collaborative Dual-Path Network. Pattern Recognition and Artificial Intelligence, 2020, 33(10): 867-878.)

[7] 张 磊,吴晓富,张索非,等. 基于多分支协作的行人重识别网络. 模式识别与人工智能, 2021, 34(9): 853-862.
(ZHANG L, WU X F, ZHANG S F, *et al*. Multi-branch Cooperative Network for Person Re-identification. Pattern Recognition and Artificial Intelligence, 2021, 34(9): 853-862.)

[8] WANG Z X, WANG Z, ZHENG Y Q, *et al*. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-identification // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2019: 618-626.

[9] WANG G A, YANG Y, ZHANG T Z, *et al*. Cross-Modality Paired-Images Generation for RGB-Infrared Person Re-identification. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12144-12151.

[10] WANG G A, ZHANG T Z, CHENG J, *et al*. RGB-Infrared Cross-Modality Person Re-identification via Joint Pixel and Feature Alignment // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2019: 3622-3631.

[11] 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. The Visual Computer (International Journal of Computer Graphics), 2022, 38(1): 279-294.

[12] ZHU J Y, PARK T, ISOLA P, *et al*. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 2242-2251.

[13] LIU H J, MA S, XIA D X, *et al*. SFANet: A Spectrum-Aware Feature Augmentation Network for Visible-Infrared Person Reidentification. IEEE Transactions on Neural Networks and Learning Systems, 2021. DOI: 10.1109/TNNLS.2021.3105702.

[14] LI D G, WEI X, HONG X P, *et al*. Infrared-Visible Cross-Modal Person Re-identification with an X Modality. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 4610-4617.

[15] WANG Z J, LIU L, ZHANG H X. Dual-Path Image Pair Joint Discrimination for Visible-Infrared Person Re-identification. Journal of Visual Communication and Image Representation, 2022, 85. DOI: 10.1016/j.jvcir.2022.103512.

[16] GAO G W, SHAO H, WU F, *et al*. Leaning Compact and Representative Features for Cross-Modality Person Re-identification. World Wide Web, 2022, 25(4): 1649-1666.

[17] WANG C D, ZHANG C, FENG Y J, *et al*. Learning Visible Thermal Person Re-identification via Spatial Dependence and Dual-Constraint Loss. Entropy, 2022, 24(4). DOI: 10.3390/e24040443.

[18] HU W P, LIU B H, ZENG H T, *et al*. Adversarial Decoupling and Modality-Invariant Representation Learning for Visible-Infrared Person Re-identification. IEEE Transaction on Circuits and Systems for Video Technology, 2022, 32(8): 5095-5109.

[19] HAO Y, WANG N N, GAO X B, *et al*. Dual-Alignment Feature Embedding for Cross-Modality Person Re-identification // Proc of the 27th ACM International Conference on Multimedia. New York, USA: ACM, 2019: 57-65.

[20] PARK H, LEE S, LEE J, *et al*. Learning by Aligning: Visible-Infrared Person Re-identification Using Cross-Modal Correspondences // Proc of the IEEE/CVF International Conference on Computer Vision. Washington, USA: IEEE, 2021: 12026-12035.

[21] CHEN Y, WAN L, LI Z H, *et al*. Neural Feature Search for RGB-Infrared Person Re-identification // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2021: 587-597.

[22] WU A C, ZHENG W S, YU H X, *et al*. RGB-Infrared Cross-Modality Person Re-identification // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 5390-5399.

[23] YE M, SHEN J B, LIN G J, *et al*. Deep Learning for Person Re-identification: A Survey and Outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2872-2893.

[24] YE M, SHEN J B, CRANDALL D J, *et al*. Dynamic Dual-Attentive Aggregation Learning for Visible-Infrared Person Re-identification // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 229-247.

[25] YE M, LAN X Y, WANG Z, *et al*. Bi-directional Center-Constrained Top-Ranking for Visible Thermal Person Re-identification. IEEE Transactions on Information Forensics and Security, 2020, 15: 407-419.

[26] LIU H J, CHENG J, WANG W, *et al*. Enhancing the Discriminative Feature Learning for Visible-Thermal Cross-Modality Person Re-identification. Neurocomputing, 2020, 398: 11-19.

[27] SU C, LI J N, ZHANG S L, *et al*. Pose-Driven Deep Convolutional Model for Person Re-identification // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 3980-3989.

[28] ZHENG L, HUANG Y J, LU H C, *et al*. Pose-Invariant Embedding for Deep Person Re-identification. IEEE Transactions on Image Processing, 2019, 28(9): 4500-4509.

[29] TAY C P, ROY S, YAP K H. AANet: Attribute Attention Network for Person Re-identifications // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2019: 7127-7136.

[30] WANG G A, YANG S, LIU H Y, *et al*. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-identification // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2020: 6449-6458.

[31] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al*. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale[C/OL]. [2022-06-04]. https://arxiv.org/pdf/2010.11929.pdf.

[32] RADENOVIĆ F, TOLIAS G, CHUM O. Fine-Tuning CNN Image Retrieval with No Human Annotation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1655-1668.

[33] ZHU Y X, YANG Z, WANG L, *et al*. Hetero-Center Loss for Cross-Modality Person Re-identification. Neurocomputing, 2020, 386: 97-109.

[34] 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. Sensors(Basel), 2017, 17(3). DOI: 10.3390/s17030605.

[35] YE M, WANG Z, LAN X Y, *et al*. Visible Thermal Person Re-identification via Dual-Constrained Top-Ranking // Proc of the 27th International Joint Conference on Artificial Intelligence. San Francisco, USA: IJCAI, 2018: 1092-1099.

[36] DENG J, DONG W, SOCHER R, *et al*. ImageNet: A Large-Scale Hierarchical Image Database // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2009: 248-255.

[37] HE K M, ZHANG X Y, REN S Q, *et al*. Deep Residual Learning for Image Recognition // Proc of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2016: 770-778.

[38] CHOI S, LEE S, KIM Y, *et al*. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-identification // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2020: 10254-10263.

[39] DAI P Y, JI R R, WANG H B, *et al*. Cross-Modality Person Re-

identification with Generative Adversarial Training // Proc of the 27th International Joint Conference on Artificial Intelligence. San Francisco, USA: IJCAI, 2018: 677-683.

[40] ZHANG Q, LAI J H, XIE X H. Learning Modal-Invariant Angular Metric by Cyclic Projection Network for VIS-NIR Person Re-identification. IEEE Transactions on Image Processing, 2021, 30: 8019-8033.

[41] ZHAO J Q, WANG H Z, ZHOU Y, *et al*. Spatial-Channel Enhanced Transformer for Visible-Infrared Person Re-identification. IEEE Transactions on Multimedia, 2022. DOI: 10.1109/TMM.2022.3163847.

[42] WEI Z Y, YANG X, WANG N N, *et al*. Flexible Body Partition-Based Adversarial Learning for Visible Infrared Person Re-identification. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 4676-4687.

[43] GAO W B, LIU L, ZHU L, *et al*. Visible-Infrared Person Re-identification Based on Key-Point Feature Extraction and Optimization. Journal of Visual Communication and Image Representation, 2022, 85. DOI: 10.1016/j.jvcir.2022.103511.

[44] LI K F, WANG X L, LIU Y, *et al*. Cross-Modality Disentanglement and Shared Feedback Learning for Infrared-Visible Person Re-identification. Knowledge-Based Systems, 2022, 252. DOI: 10.1016/j.knosys.2022.109337.

[45] ZHANG D M, ZHANG Z Z, JU Y, *et al*. Dual Mutual Learning for Cross-Modality Person Re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5361-5373.

[46] SUN Y F, ZHENG L, YANG Y, *et al*. Beyond Part Models: Person Retrieval with Refined Part Pooling(and a Strong Convolutional Baseline) // Proc of the European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 501-518.

[47] SELVARAJU R R, COGSWELL M, DAS A, *et al*. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization // Proc of the IEEE International Conference on Computer Vision. Washington, USA: IEEE, 2017: 618-626.

[48] LI H F, CHEN Y W, TAO D P, *et al*. Attribute-Aligned Domain-Invariant Feature Learning for Unsupervised Domain Adaptation Person Re-identification. IEEE Transactions on Information Forensics and Security, 2020, 16: 1480-1494.

[49] YAO H T, ZHANG S L, HONG R C, *et al*. Deep Representation Learning with Part Loss for Person Re-identification. IEEE Transactions on Image Processing, 2019, 28(6): 2860-2871.

[50] ZHANG S S, YANG J, SCHIELE B. Occluded Pedestrian Detection through Guided Attention in CNNs // Proc of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2018: 6995-7003.

[51] DING C X, WANG K, WANG P F, *et al*. Multi-task Learning with Coarse Priors for Robust Part-Aware Person Re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(3): 1474-1488.

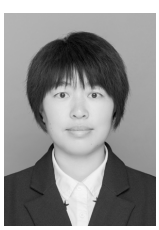
作者简介

石林波,硕士研究生,主要研究方向为计算机视觉、行人重识别. E-mail:1527467911@qq.com.

(**SHI Linbo**, master student. His research interests include computer vision and person re-identification.)

李华锋,博士,教授,主要研究方向为图像处理、计算机视觉. E-mail:hfchina99@163.com.

(**LI Huafeng**, Ph. D., professor. His research interests include image processing and computer vision.)



张亚飞(通信作者),博士,副教授,主要研究方向为图像处理、模式识别. E-mail:zyfeimail@163.com.

(**ZHANG Yafei** (Corresponding author), Ph. D., associate professor. Her research interests include image processing and pattern recognition.)

谢明鸿,博士,高级工程师,主要研究方向为计算机视觉. E-mail:minghongxie@163.com.

(**XIE Minghong**, Ph. D., senior engineer. His research interests include computer vision.)