生成对抗网络协同角度异构中心三元组损失的跨模态行人重识别

周 非[1,2],舒浩峰[1],白梦林[1],王锦华[1]
(1. 重庆邮电大学通信与信息工程学院,重庆 400065;2. 泛在感知与互联重庆市重点实验室,重庆 400065)

摘 要: 基于红外与可见光域之间的跨模态行人重识别对于夜间场景监控极为重要,但由于红外图像和可见光图像的数据分布存在较大差异,使得模型很难提取到同一行人在不同模态下的模态不变特征. 本文针对现有跨模态行人重识别算法中存在的数据集样本数量较少问题以及不同模态图像之间存在较大跨模态差异问题,提出了一种新颖的生成对抗网络来生成与原始图像相似的匹配图像,在对跨模态行人数据集进行增广的同时减少跨模态差异;为减少跨模态差异和模态内差异,本文采用了双流网络来提取更具鉴别性特征,并提出了角度异构中心三元组损失对正负样本在特征空间中夹角进行约束,提升其在特征空间中的聚类效果. 本文在SYSU-MM01和RegDB数据集上进行实验验证,结果表明本文所提出的生成匹配图像方法能够有效降低不同模态图像之间的跨模态差异,同时角度异构中心三元组损失使得特征空间中的嵌入特征具有角度辨别性,从而提升模型的分类能力. 在SYSU-MM01数据集中,本文方法相较于最新算法在Rank-1和mAP分别提升了5.71%和8.18%,证实了文中方法的有效性.

关键词: 行人重识别;跨模态;生成对抗网络;双流网络;三元组损失
基金项目: 国家自然科学基金(No.62271096)
中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2023)07-1803-09
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220587

Cross-Modal Person Re-Identification Based on Generative Adversarial Network Coordinated with Angle Based Heterogeneous Center Triplet Loss

ZHOU Fei[1,2], SHU Hao-feng[1], BAI Meng-lin[1], WANG Jin-hua[1]
(1. *School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*; 2. *Chongqing Key Laboratory of Ubiquitous Sensing and Networking, Chongqing* 400065, *China*)

Abstract: Cross-modal person re-identification based on infrared and visible images is very important for night scene monitoring, but due to the large difference in the data distribution of infrared images and visible images, it is difficult for the model to extract the modal-invariant features of the same pedestrian in different modal. Aiming at the problem of the small number of dataset samples and the large cross-modal difference between different modal images in the existing cross-modal person re-identification methods, this paper proposes a generative adversarial network to generate matching images which are similar to the original images which will augment the cross-modal person dataset while reducing cross-modal differences. To further reduce cross-modal differences and intra-modal differences, this paper utilizes a two stream network to extract discriminative features. Meanwhile to improve the positive and negative sample pairs' clustering effect in the feature space, an angle-based heterogeneous center triplet loss is proposed to constrain the angle between those sample pairs. Experiments are performed on the SYSU-MM01 and RegDB datasets. The results show that the proposed method for generating matching images can effectively reduce the cross-modal differences between images of different modalities. At the same time, the angle-based heterogeneous center triplet loss makes embedding features in feature space are angle-discriminative, thus improving the model's classification ability. Results on the SYSU-MM01 dataset show that Rank-1 and mAP have increased by 5.71% and 8.18% respectively, compared with the latest methods, confirming the effectiveness of our

收稿日期:2022-05-23;修回日期:2023-02-22;责任编辑:李勇锋

method.

Key words: person re-identification; cross modal; generative adversarial network; two-stream network; triplet loss
Foundation Item(s): National Natural Science Foundation of China (No.62271096)

1 引言

随着多年以来科学技术的不断发展,城市监控网络愈发完善. 利用监控画面对行人进行检索已经成为了公安机关搜捕犯人的一个常用的技术手段. 而行人重识别技术[1]就是利用计算机视觉和机器学习技术来对某个监控场景下的特定行人进行跨摄像头或跨时间域的检索,所以在近年来引起了学术界的广泛关注,同时也出现了较多研究成果.

现阶段行人重识别算法的研究大多都是利用深度学习的方法对可见光图像进行检索,且已经取得了较为令人满意的实验结果,例如在文献[2]中提出的基于视角感知损失的可见光行人重识别方法已经能够在Market1501数据集[3]上达到95.43%的平均精度(mean Average Precision,mAP).

为实现全时段监控,大多数监控摄像头会在低光照条件(如夜间)下自动由可见光(RGB)模式切换到近红外(Near InfraRed,NIR)模式. 红外相机将接收到的被测目标的红外辐射信号转化为红外热像图;而RGB相机则是根据波长将可见光分解为三个通道(红色、绿色和蓝色)进行成像. 由于这两类相机的成像原理完全不同,这就使得这两种模态图像的数据分布具有较大差异(即模态差异),导致很难使用红外图像与RGB图像进行匹配. 所以跨模态行人重识别任务不仅要解决传统行人重识别任务中常见的模态内差异问题,还需要对红外图像与可见光图像之间的模态间差异问题进行解决. 为解决模态差异问题,学术界主要提出了两种思路:一种是采用双流网络来学习多模态图像的模态特定特征和模态共享特征[4],另一种是通过利用生成对抗网络(Generative Adversarial Networks,GANs)来生成配对图像[5]或多谱图像[6]的方式来消除模态差异.

此外,大多数行人重识别算法都是同时利用交叉熵损失和三元组损失[7]来对网络进行训练,从而达到减少类内距离,增大类间距离的目的. 而传统的跨模态难三元组损失是强约束条件,当图片中存在离群值(如错误标签)时,它反而会破坏其他已经训练好的成对距离[8]. 此外,它无法约束正负样本对在特征空间中的夹角,从而导致在测试阶段模型不能依据角度来划分不同的行人标签[9].

根据上述思路,本文在网络结构和损失函数两个角度进行了创新:(1)在网络结构上,本文提出了一个融合生成对抗网络的双流网络来对多模态图像特征进行学习;(2)在损失函数上,本文提出了一个新颖的基于角度的异构中心三元组损失,利用中心三元组损失来减少异常值对成对距离的影响,并引入角度对正负样本在特征空间中的夹角进行约束,提升其在特征空间中的聚类效果,从而改善网络性能.

2 提出的方法

由于模态差异很大程度上影响了跨模态行人重识别算法的性能,所以本文研究重点是在最小跨模态差异的同时学习到具有鉴别能力的行人特征,并利用提取到的特征对模型进行优化. 本节先会对本文的网络框架进行介绍,随后将重点对本文提出的基于角度的异构中心三元组损失进行介绍.

2.1 网络框架

如图1所示,本文提出了一个融合生成对抗网络的双流网络来对多模态图像特征进行学习,该网络主要包含两个模块,分别是利用生成对抗网络MatchGAN来生成跨模态匹配图像的图像生成模块和由双流网络构成的特征提取模块.

2.2 MatchGAN

模态差异由模态内差异和模态间差异组成,模态内差异主要指由于视角、姿势等不同从而造成同一行人两幅图像间具有较大差异的情况,而模态间差异是由于红外相机和RGB相机成像原理不同,使得同一行人的这两种模态图像的数据分布具有较大差异. 而MatchGAN则是通过产生姿势相似图像的方法来减少同一行人图像的模态内差异,从而达到减少模态差异的目的,让网络模型能够更专注于对模态间差异进行学习优化.

如图2(a)所示,每列图片为同一行人在不同光照和拍摄角度下采集到的图片,可以很明显地观察到未匹配跨模态图像之间存在着较大的模态内差异(如姿势等). 为解决这个问题,本文采用MatchGAN模块来对未匹配图像进行修正. 通过生成匹配图像的方式,如图2(b)所示,使得生成的跨模态图片与原始图片具有相同的姿势信息和着装风格,从而在对数据集增广的同时降低跨模态图像之间的模态内差异.

2.2.1 匹配图像生成流程

本文采用的MatchGan网络结构如图3所示,采用三个编码器分别对同一行人不同模态图片的模态特定特征和模态不变特征进行学习.

如式(1)所示,将RGB图片X_{rgb}通过RGB特征编码器$E^{\mathrm{s}}_{\mathrm{rgb}}$来学习RGB模态特定特征$f^{\mathrm{s}}_{\mathrm{rgb}}$(即行人服装颜色、

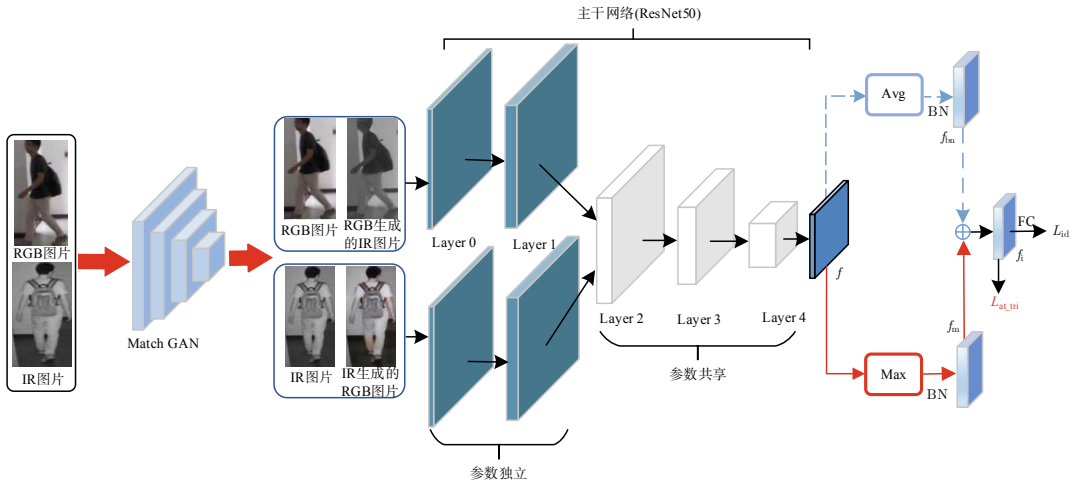


图 1　生成对抗网络融合双流网络模型框架图

(a) 未匹配图像　　(b) 生成的匹配图像

图 2　采用 MatchGAN 生成匹配图像

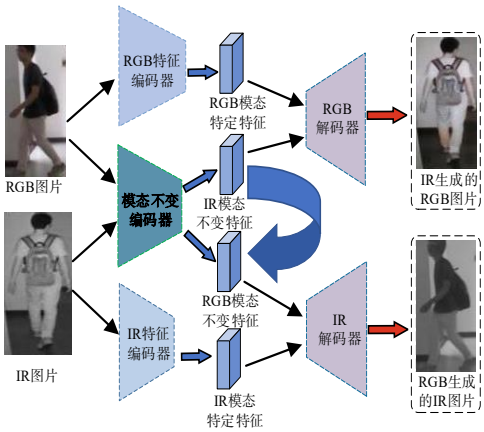


图 3　MatchGAN 结构图

纹理信息等),并通过模态不变编码器 E^{i} 来学习模态不变特征 f_{rgb}^{i}(即行人姿势、体型等);同理如式(2)所示,IR(InfraRed)图片 X_{ir} 依次通过 IR 特征编码器 E_{ir}^{s} 和模态不变编码器 E^{i} 分别得到 IR 模态特定特征 f_{ir}^{s} 以及模态不变特征 f_{ir}^{i}.

$$f_{rgb}^{s}=E_{rgb}^{s}(X_{rgb}),\ f_{rgb}^{i}=E^{i}(X_{rgb}) \tag{1}$$

$$f_{ir}^{s}=E_{ir}^{s}(X_{ir}),\ f_{ir}^{i}=E^{i}(X_{ir}) \tag{2}$$

在保持模态特定特征 f_{rgb}^{s}、f_{ir}^{s} 不变的前提下,解码器 D_{rgb} 和 D_{ir} 通过交换不同模态图片的模态不变特征(即 f_{rgb}^{i} 与 f_{ir}^{i})的方式来生成匹配图片,如式(3)所示. 这样生成的 RGB 图像 X_{fake_rgb} 不会改变原始 IR 图像的行人姿势、体型等模态不变信息,并具有与原始 RGB 图像相似的服装颜色、纹理信息等模态特定特征.

$$X_{fake\_rgb}=D_{rgb}(f_{ir}^{i},f_{rgb}^{s}),X_{fake\_ir}=D_{ir}(f_{rgb}^{i},f_{ir}^{s}) \tag{3}$$

2.2.2　重建损失

为改善网络生成图像的质量,本文采用了三个重建损失(reconstruction loss). 为保证分离出来的特征能够重建它们的原始图像,本文采用 L_{recon}^{same} 来对同模态图片重建进行优化,如式(4)所示,式中 $\|\cdot\|$ 代表 L1 距离. 这个损失函数在 GAN 网络的正则化中起着重要的作用,能够有效防止网络过拟合[10].

$$L_{recon}^{same}=\left\|X_{rgb}-D_{rgb}(f_{rgb}^{i},f_{rgb}^{s})\right\|_{1}+\left\|X_{ir}-D_{ir}(f_{ir}^{i},f_{ir}^{s})\right\|_{1} \tag{4}$$

但式(4)无法监督跨模态匹配图像的生成,所以为了保证生成的 RGB 图像 X_{fake_rgb} 与原始 IR 图像的模态不变特征 f_{ir}^{i} 以及原始 RGB 图像的模态特定特征 f_{rgb}^{s} 之间的相关性,本文采用 cycle loss[11] 来对生成图像进行约束,如式(5)所示. L_{cycle} 保证了生成图像能够重新转换为原始图像,进一步对生成图像进行限制,解决了生成图像不匹配问题.

$$L_{cycle}=\left\|X_{rgb}-\tilde{X}_{rgb}\right\|_{1}+\left\|X_{ir}-\tilde{X}_{ir}\right\|_{1} \tag{5}$$

式中:$\tilde{X}_{rgb}=D_{rgb}(\tilde{f}_{ir}^{i},\tilde{f}_{rgb}^{s})$,同样 $\tilde{X}_{ir}=D_{ir}(\tilde{f}_{rgb}^{i},\tilde{f}_{ir}^{s})$;$\tilde{f}_{rgb}^{s}$,$\tilde{f}_{rgb}^{i}$ 分别代表从 X_{fake_rgb} 中提取到的 RGB 模态特定特征

和模态不变特征，$\tilde{f}_{ir}^{i}$、$\tilde{f}_{ir}^{s}$同理.

为让生成图像更接近真实图像，本文引入了两个鉴别器Dis_{rgb}和Dis_{ir}. 以RGB图像鉴别器Dis_{rgb}为例，它在生成RGB图像X_{fake_rgb}和原始RGB图像X_{rgb}中试图分辨出生成图片和原始图片，并通过式(6)计算得到对抗损失(adversarial loss)[12]对生成器进行优化；而生成器D_{rgb}则需要生成一个更真实的RGB图片来欺骗鉴别器. 通过生成器和鉴别器相互对抗的方式让生成图像更加接近真实图像.

$$\begin{aligned} L_{GAN}^{rgb} &= E\left[\lg\left(\mathrm{Dis}_{rgb}(X_{rgb})\right)+\lg\left(1-\mathrm{Dis}_{rgb}(X_{fake\_rgb})\right)\right] \\ L_{GAN}^{ir} &= E\left[\lg\left(\mathrm{Dis}_{ir}(X_{ir})\right)+\lg\left(1-\mathrm{Dis}_{ir}(X_{fake\_ir})\right)\right] \end{aligned} \tag{6}$$

最后得到的重建损失L如式(7)所示，本文根据前人经验以及实验结果，将λ_{recon}和λ_{GAN}设为1，λ_{cycle}设为10.

$$L=\lambda_{cycle}L_{cycle}+\lambda_{GAN}L_{GAN}+\lambda_{recon}L_{recon} \tag{7}$$

2.3 双流网络

为提取到对模态变化鲁棒的特征，本文采用了双流网络结构，如图1所示. 由于跨模态差异主要存在于浅层特征[13]，同时依照前人经验[14]，本文将ResNet50[15]网络的浅层卷积模块layer 0和第一个残差卷积块layer 1参数独立，作为特定模态特征提取模块，分别学习不同模态的浅层特征. 而剩下的三个残差卷积块layer 2，layer 3和layer 4则是参数共享的，在同一个特征空间中学习模态共享的特征表示.

由于最后提取到的深层次特征f往往尺寸较小，并且包含了较为有效的高级语义信息，所以一般采用平均池化的方式对特征图的全局信息进行聚合，以保留其鉴别信息的完整性. 而最大池化如式(8)所示，与平均池化不同，它可以保留特征图中最为显著最有辨识度的信息.

$$S_{ij}=\underset{i=1,j=1}{\mathrm{MAX}}\left(f_{ij}\right)+b_2 \tag{8}$$

式中S为得到的子采样特征图，移动步长为c，池化域为$c\times c$矩阵，偏置为b_2.

所以本文结合两者特性，如图1所示，将提取到的特征f按照蓝色虚线通过平均池化层，随后对其进行批归一化操作(batch normalize)[16]，在加快网络收敛速度的同时防止网络过拟合，得到特征f_{bn}. 紧接着按照红色实线，将原始特征f通过最大池化层得到特征f_m，随后按照逐位相加(point-wise addition)的方式将特征f_{bn}与特征f_m进行特征融合，并进行批归一化得到最终的双池化融合特征f_i，在保留特征图鉴别信息完整性的同时，凸显出特征中最具有鉴别性部分.

为了让模型能有较好的特征学习能力，现阶段的行人重识别网络大多都是采用联合分类损失和度量损失的方法对网络进行训练，共同约束特征. 所以本文也采用联合训练的方法，利用特征f_i来计算角度异构中心三元组损失L_{AC_tri}，作为度量损失对类内差异进行约束，并结合交叉熵损失作为分类损失对类间差异进行约束，共同对网络进行训练.

2.3.1 基于角度的异构中心三元组损失

传统三元组损失[16]的基本思想是正样本对之间的距离加上一个预先设定好的边界值后依然要小于其与负样本之间的距离. 在这个思想的基础上，为了避免模型陷入局部最优，提升模型的泛化能力，研究者往往采用基于难样本挖掘的三元组损失(即难三元组损失)作为度量损失，如式(9)所示.

$$L_{Hard\_tri}=\sum_{i=1}^{P}\sum_{a=1}^{K}\left[\rho+\max_{p=1,2,\cdots,K}D(x_a^i,x_p^i)-\min_{\substack{n=1,2,\cdots,K\\ j=1,2,\cdots,P\\ i\neq j}}D(x_a^i,x_n^j)\right]_+ \tag{9}$$

式中P为行人类别数，K为每个行人提取的图片数量，x_a^i为锚点，x_p^i、x_n^i分别代表正样本和负样本，ρ是设定的边界值，$D(\cdot)$代表计算欧式距离.

但难三元组损失需要将锚点与batch中其他所有图片计算得到，计算开销较大；同时它也是一个强约束条件，当数据集中存在离群值时(例如错误标签)，反而会破坏其它训练好的成对距离. 针对这一问题，本文在传统中心损失的基础上进行改进，并结合了难三元组损失的思想，提出了异构中心三元组损失L_{center_tri}. 传统中心损失[16]如式(10)所示，通过约束每个行人特征f_i到其对应的类别中心c_i之间距离，从而达到提升类内紧凑性的目的. 式中B代表每批数据中图片数量，c_i代表行人i的类别中心，f_i代表特征f所属行人类别为i.

$$L_{center}=\frac{1}{2}\sum_{i=1}^{B}\left\| f_i-c_i\right\|_2^2 \tag{10}$$

而在跨模态行人重识别任务中，在训练阶段每次迭代随机采样P个行人，并对每个模态分别提取K张图片，共$2P\times K$张图片. 每批数据可以利用式(11)分别得到P个行人的RGB模态类别中心C_v^i以及P个IR模态类别中心C_t^i，式中c_j^i表示第i个行人的第j张RGB图片，同理t_j^i代表第i个行人的第j张IR图片.

$$C_v^i=\frac{1}{K}\sum_{j=1}^{K}v_j^i \quad , \quad C_t^i=\frac{1}{K}\sum_{j=1}^{K}t_j^i \tag{11}$$

与传统中心损失不同，异构中心三元组损失利用锚点的不同模态类别中心组成正样本对，并在其他所有类别中心中挖掘出最难负样本，如式(12)所示. 利用类别中心之间比较的方式来代替锚点与所有样本的比

较,这样在降低模型计算量同时降低离群值对模型性能的影响.

$$L_{\text{center\_tri}}=\sum_{i=1}^{P}\left[\rho+D\left(C_{\text{v}}^{i},C_{\text{t}}^{i}\right)-\min_{j\neq i}D\left(C_{\text{v}}^{i},C_{\text{v/t}}^{j}\right)\right]_{+}+\sum_{i=1}^{P}\left[\rho+D\left(C_{\text{t}}^{i},C_{\text{v}}^{i}\right)-\min_{j\neq i}D\left(C_{\text{t}}^{i},C_{\text{v/t}}^{j}\right)\right]_{+}\tag{12}$$

异构中心三元组损失所使用的欧式距离虽然能够较好地在特征空间中对正负样本对之间的距离进行优化(即 $d_1+\rho<d_2$),但是它潜在地在算法设计上存在着缺陷,即它不能有效地约束特征向量之间的夹角,这样会导致特征向量在特征空间中的夹角是不确定的,所以可能会出现锚点与正样本之间的夹角 α 大于与负样本之间夹角 β 的情况,如图4所示.

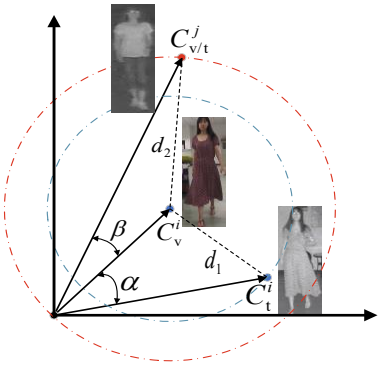


图4　正负样本在特征空间中分布情况示意图

而在训练阶段,特征空间中的嵌入特征具有角度辨别性是非常重要的. 为得到特征向量对每个类别的预测得分,最后一个全连接层会在特征向量 $\{f_i,i\in[1,N]\}$ 与不同类别的权重向量 $\{\omega_j,j\in[1,C]\}$ 之间计算点积,如式(13)所示. 而为避免对某些类别存在先验偏差,权重向量 ω_j 的大小应彼此接近,所以每个特征向量 f_i 在全连接层中对每个类别的预测得分一定程度上就取决于嵌入特征向量的夹角 $\theta(i,j)$.

$$f_i\cdot\omega_j=\left|f_i\right|\left|\omega_j\right|\cos\left(\theta(i,j)\right)\tag{13}$$

为解决上述问题,本文提出了基于角度的异构中心三元组损失 $L_{\text{AC\_tri}}$. 由于余弦距离如式(14)所示,关注重点在向量之间的夹角,因此它对特征向量在特征空间中学习合适的方向有很强的约束效果.

$$C(X,Y)=1-\cos(X,Y)=1-\frac{X\cdot Y}{|X||Y|}\tag{14}$$

所以本文尝试用余弦距离去代替三元组损失中常用的欧式距离,从而对特征向量之间的夹角进行约束. 所以将式(14)带入式(12),化简后得到 $L_{\text{AC\_tri}}$,如式(15)所示.

$$L_{\text{AC\_tri}}=\sum_{i=1}^{P}\left[\rho-\cos\left(C_{\text{v}}^{i},C_{\text{t}}^{i}\right)+\max_{j\neq i}\cos\left(C_{\text{v}}^{i},C_{\text{v/t}}^{j}\right)\right]_{+}+\sum_{i=1}^{P}\left[\rho-\cos\left(C_{\text{t}}^{i},C_{\text{v}}^{i}\right)+\max_{j\neq i}\cos\left(C_{\text{t}}^{i},C_{\text{v/t}}^{j}\right)\right]_{+}\tag{15}$$

式(15)所得到的角度异构中心三元组损失存在的问题是:$L_{\text{AC\_tri}}$ 会使得 $\cos\left(C_{\text{v}}^{i},C_{\text{v/t}}^{j}\right)$ 趋于-1,即为负相关,而训练目标是锚点与负样本不相关,所以增加了 $[\cdot]_+$ 函数对 $\cos\left(C_{\text{v}}^{i},C_{\text{v/t}}^{j}\right)$ 进行限制,使其趋于0;同时考虑到余弦函数的取值范围为 $[-1,1]$,所以本文将 ρ 设置为1来简化参数;并且在实验时,由于 $L_{\text{AC\_tri}}$ 数值太小,会使得模型训练速度过慢,导致网络无法正常收敛. 所以为加快模型训练速度,本文将指数函数 $y=\mathrm{e}^{x}$ 引入了到式(15)中,最终得到角度异构中心三元组损失 $L_{\text{AC\_tri}}$,如式(16)所示. 式中,λ_1、λ_2 为权重系数,这里本文按照经验值,它们都设为0.1.

$$L_{\text{AC\_tri}}=\lambda_1\sum_{i=1}^{P}\mathrm{e}^{\left(\left[\max\limits_{j\neq i}\cos\left(C_{\text{t}}^{i},C_{\text{v/t}}^{j}\right)\right]_{+}-\cos\left(C_{\text{v}}^{i},C_{\text{t}}^{i}\right)+1\right)_{+}}+\lambda_2\sum_{i=1}^{P}\mathrm{e}^{\left(\left[\max\limits_{j\neq i}\cos\left(C_{\text{t}}^{i},C_{\text{v/t}}^{j}\right)\right]_{+}-\cos\left(C_{\text{t}}^{i},C_{\text{v}}^{i}\right)+1\right)_{+}}\tag{16}$$

2.3.2　交叉熵损失

为了确保模型分类的准确性,交叉熵损失通过将不同行人划为不同类别的方式来整合身份信息. 本文采用了文献[16]中提出的结合标签平滑的交叉熵损失,如式(17)所示. 将模型输出通过全连接层对其所属类别进行预测,其中 p_i 表示模型预测其为类别 i 的可能性,N 代表训练样本中行人类别的数量.

$$L_{\text{ID}}=\sum_{i=1}^{N}-q_i\lg(p_i)\begin{cases}q_i=1-\dfrac{N-1}{N}\xi, y\neq i\\ q_i=\dfrac{\xi}{N}, y=i\end{cases}\tag{17}$$

同时为防止模型过拟合,提升模型的泛化能力[17],q_i 由式(17)计算得到,ξ 为超参数,本文将其设置为0.1. 它在训练时即假设标签可能存在错误,避免"过分"相信训练样本的标签,提升了模型在遇到错误标签时的修正能力.

3　实验结果与分析

3.1　实验设置

本文在常用的SYSU-MM01[13]以及RegDB[18]可见光—红外行人重识别数据集上进行了评估实验. SYSU-MM01数据集包含了491个行人由4个可见光相机拍摄的287 628幅可见光行人图像以及2个红外相机拍摄的15 792幅红外行人图像,将其中的395个行人图像作为训练集用来对模型进行训练,剩下的96个行人的所有图像作为测试集对模型进行测试. SYSU-MM01数据集包含all-search模式和indoor-search模式两种模式,在all-search模式中,将采用可见光相机1、2、4、5作为gal-

lery集,红外相机3和6作为query集;而在indoor-search模式中则仅将室内可见光相机1和2作为gallery集,红外相机3和6作为query集. 在实验中,本文均采用sing-shot模式进行检索,即在测试阶段gallery中每个行人仅包含一张图片. RegDB数据集对412个行人分别采集了10幅可见光图像和10幅红外图像,共8 240幅图像,随机划分一半作为训练集对模型进行训练,另一半则作为测试集. 本文使用可见光图像进行检索,将红外图像作为待检索图像.

在训练阶段,本文采用在ImageNet预训练后的ResNet50作为主干网络. 在实验中,图像大小设定为288×144,并使用了水平翻转以及随机擦除[19]的方式来进行图像增强. 网络共训练660个周期(epoch),在前600个周期仅对图像生成模块MatchGAN进行训练,此时损失函数为$L_{\mathrm{MG}}=10\times L_{\mathrm{cycle}}+L_{\mathrm{GAN}}+L_{\mathrm{recon}}$. 后60个周期将对图像生成模块和特征提取模块同时训练,此时图像生成模块的损失函数依然为L_{MG},特征提取模块的损失函数为$L=L_{\mathrm{AC_tri}}+L_{\mathrm{ID}}$,$L_{\mathrm{MG}}$和$L$都仅对各自模块的参数进行优化. 每批数据随机采样6个行人,每个行人每个模态提取8张图片,共96张图片. MatchGAN采用ADAM优化器对网络进行优化,学习率设置为0.000 1. 特征提取模块采用的随机梯度下降算法来对模型进行优化,初始学习率设置为0.1,到第620个周期时会衰减为0.01,到第650个周期时会衰减为0.001.

在测试阶段,本文采用了累积匹配曲线中的Rank-1、Rank-10以及Rank-20识别率和平均精度(mAP)作为性能评价指标. 考虑到测试图像选取的随机性,实验结果可能会出现偏差,所以在测试阶段,本文对模型进行了十次重复实验,取平均值作为最终测试结果,从而提升测试精度的稳定性.

3.2 对比实验

本文提出的方法在SYSU-MM01数据集上与最新的10种方法进行了比较,包括BDTR[4],eBDTR[4],D2RL[6],Hi-CMD[10],AlignGAN[20],JSIA[5],AGW[21],DDAG[22],LbA[23],SFANET[24]. 为更全面地进行比较,本文将上述方法均在SYSU-MM01数据集的all-search和indoor-search两种模式下进行实验,实验结果如表1所示,表中r=1、r=10以及r=20分别代表Rank-1,Rank-10,Rank-20,mAP为平均精度.

由表1可以看出,在SYSU-MM01数据集的两种模式下,相较于采用单流网络的D2RL[6],采用双流网络的DDAG[22]、SFANET[24]以及本文提出的方法性能表现更好,说明双流网络能够加强模型对于不同模态图像特征的挖掘能力,学习到更具判别性特征. 同时与基于生成对抗网络的Hi-CMD[10]、AlignGAN[20]和JSIA[5]算法相比,本文提出的方法效果更好,说明本文将双流网络与生成对抗网络相融合的方法能够在更大程度上减少不同模态图片之间的差异,提升网络性能. 并且与最为先进的SFANET[24]算法相比,本文提出的方法在all-search模式下,Rank-1提升了5.71%,并且在mAP上提升了8.18%,在indoor-search模式下,Rank-1提升了6.49%,mAP也略有提升. 通过对比上述实验结果可以证实本文提出方法的有效性.

表1 在SYSU-MM01数据集上与其他先进算法的准确率对比 单位:%

方法	All-search single-shot				Indoor-search single-shot			
	r=1	r=10	r=20	mAP	r=1	r=10	r=20	mAP
BDTR[4] (IJCAI-18)	27.32	66.96	81.07	27.32	31.92	77.18	89.28	41.86
eBDTR[4](IJCAI 18)	27.86	67.32	81.34	28.42	32.46	77.42	89.62	42.46
D^{2}RL[6](CVPR-19)	28.90	70.60	82.40	29.40	—	—	—	—
Hi-CMD[10](CVPR-20)	34.94	77.58	—	35.94	—	—	—	—
AlignGAN[20](ICCV-19)	42.40	85.00	93.70	40.70	45.90	87.60	94.40	45.30
JSIA[5](AAAI-20)	38.10	80.70	89.90	36.90	43.80	86.2 0	94.20	52.90
AGW[21](IEEE-20)	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
DDAG[22](ECCV-20)	54.75	90.39	95.81	53.02	61.02	94.06	98.41	67.98
LbA[23](ICCV-21)	55.41	—	—	54.14	58.46	—	—	66.33
SFANET[24](IEEE-21)	60.45	91.80	95.16	53.87	64.80	94.67	98.07	75.16
本文方法	**66.16**	**94.35**	**97.95**	**62.05**	**71.29**	**97.87**	**99.50**	**76.26**

为进一步验证提出方法的有效性,本文进行了消融实验,如表2所示. 通过实验对比了模型在不同情况下的性能表现. 其中“基线”表示仅使用双流网络时取得的实验结果;“基线+Center_tri”、“基线+AC_tri”分别表示将双流网络的难三元组损失替换为异构中心三元组损失、角度异构中心三元组损失后取得的实验结果;“基线+MatchGAN”表示使用MatchGAN对数据增广后,双流网络取得的实验结果;而“基线+MatchGAN+

AC_tri”则表示使用MatchGAN对数据增广,并采用角度异构中心三元组损失对网络进行优化取得的实验结果.

表2　在SYSU-MM01数据集下进行消融实验　　单位:%

方法	All-search single-shot		
	r=1	r=10	mAP
基线	57.27	90.01	55.81
基线+Center_tri	59.51	88.04	54.61
基线+MatchGAN	60.53	90.98	57.83
基线+AC_tri	64.74	93.58	59.90
基线+MatchGAN+AC_tri	**66.16**	**94.35**	**62.05**

通过消融实验可以看出,本文采用的双流网络在SYSU数据集上的all-search模式下rank-1能够达到57.27%,已经超出了现阶段的许多文章[10,20],这表明了本文特征融合得到的双池化融合特征更加丰富,更具鉴别性. 随后将基线中的难三元组损失分别用异构中心三元组损失和角度异构中心三元组损失代替后,rank-1分别提升了2.24%和3.92%. 但是在替换为异构中心三元组损失后,模型在rank-10和mAP上出现了负增长的情况,这说明采用欧氏距离优化的模型,正负样本在特征空间中的聚类效果并不好,网络仅能匹配到较为容易的正样本,当正样本中存在明显的视角、姿势等变化时,不能保持较高的识别准确度. 而采用角度异构中心三元组损失通过约束特征向量在特征空间中的方向,优化了特征向量在特征空间中的聚类效果,使得模型能够更好地判断行人特征所属类别,提升模型识别精度.

同时在基线上加上MatchGAN后,模型在rank-1和mAP上分别有3.26%和2.02%的提升,说明其生成的匹配图像能够对数据集进行增强并有效降低图像间的模态差异. 最后使用图像生成模块对数据增广,并采用角度异构中心三元组损失对模型进行优化取得了最优性能,证明了本文框架的有效性.

为进一步证明本文提出的匹配图像生成算法能够在对跨模态行人数据集进行增广的同时降低图像间的模态差异,本文在DGTL[25]、DDAG[22]以及AWG[21]方法上引入了MatchGAN模块,并在SYSU-MM01数据集上进行了实验,如表3所示.

可以看到,在引入MatchGAN方法对跨模态数据集进行增广后,在上述三个基线中,rank-1准确率分别提升了3.02%,3.40%以及3.32%,同时mAP值也分别提升了2.33%,2.62%以及2.13%. 实验结果证明了本文提出的MatchGAN生成匹配图像的方法能够降低图像间的模态差异,让模型能够更容易提取到对模态变化鲁棒的特征,从而提升模型的分类能力和性能表现;并且它还能够集成到现有基线中,对跨模态数据集进行数据增强,丰富训练样本的多样性,降低模型的过拟合风险.

表3　在不同基线上通过MatchGAN数据增广后实验结果对比

单位:%

Baseline	All-search single-shot		
	r=1	r=10	mAP
DGTL	52.72	84.93	49.60
DGTL+MatchGAN	55.74	87.18	51.93
DDAG	54.75	90.39	53.02
DDAG+MatchGAN	58.15	91.81	55.64
AWG	47.50	84.39	47.75
AWG+MatchGAN	50.82	86.92	49.89

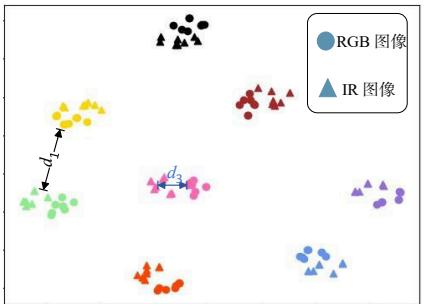
为了更直观的体现采用不同的三元组损失对网络进行优化时,正负样本在特征空间中聚类效果的差异,本文分别在采用难三元组损失和角度异构中心三元组损失训练好的模型中使用k-means算法对输入8类行人图像进行聚类,其中每个行人有8张RGB图像,8张IR图像. 随后将它们的聚类结果在特征空间中采用T-SNE算法进行可视化,得到的聚类效果如图5所示,图中圆形代表RGB图片,三角形代表红外图片,而不同颜色则代表不同的行人类别.

对比图5(a)、(b),可以明显地观察到在图5(b)中采用角度异构中心三元组损失对模型进行优化后,不同行人图像在特征空间中的类间距离(例如d_2)要明显大于其在图5(a)中的类间距离(例如d_1). 这一现象说明前者能够让不同行人在特征空间中更具区分性,从而让模型在测试阶段更容易判断输入行人图片所属类别,提升行人重识别准确度.

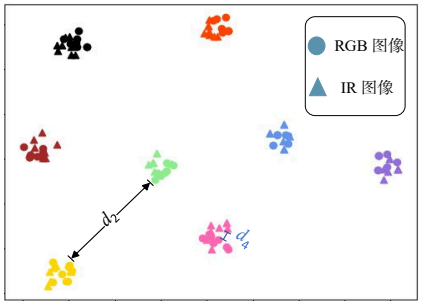
进一步观察图5(b),能够发现每类行人样本的聚类效果相较于图5(a)都更加紧凑,说明采用角度异构中心三元组损失对模型进行优化后,能够有效降低同一行人图片之间的类内距离;同时相较于图5(a),同一行人不同模态图像之间的距离(例如d_3与d_4)被大幅度减少了,这一现象表明角度异构中心三元组损失通过角度来对特征进行约束的方式,能够有效减少不同模态图像之间的跨模态差异对模型分类效果造成的影响.

为进一步验证本文提出方法的有效性,本文在RegDB数据集上将提出的方法与其他先进方法进行了比较,结果如表4所示.

通过对比表4中的实验结果,可以看出本文提出的方法在RegDB数据集上依然保持了较好的性能表现,rank-1准确率能够到达81.17%,这说明本文提出方法具有较好的泛化能力,能够更好的学习到对模态变化鲁棒的特征,让模型能够更容易找到与之匹配的跨模态图像.



(*a*) 采用传统三元组损失的聚类效果示意图



(*b*) 采用角度异构中心三元组损失的聚类效果示意图

图5 特征向量在特征空间中的聚类效果可视化结果

表4 在RegDB数据集上的对比实验结果 单位:%

方法	可见光-红外			
	r=1	*r*=10	*r*=20	mAP
eBDTR[4]	34.62	58.96	68.72	33.46
D^2RL[6]	43.40	66.10	76.30	44.10
Hi-CMD[10]	70.93	86.39	—	66.04
AlignGAN[20]	57.90	—	—	53.60
JSIA[5]	48.50	—	—	49.30
AGW[21]	70.05	86.21	91.55	66.37
DDAG[22]	69.34	86.19	91.49	63.46
SFANET[24]	76.31	91.02	94.27	68.00
本文方法	**81.17**	**93.87**	**97.65**	**70.13**

4 结语

针对跨模态行人重识别中的可见光到红外行人重识别问题,本文提出了一个生成对抗网络协同角度异构中心三元组损失的行人重识别算法.首先利用MatchGAN模块在对数据集进行增广的同时降低图像之间的跨模态差异,随后使用以ResNet50为主干网络的双流网络来提取图像特征,并采用本文提出的角度异构中心三元组损失与交叉熵损失对模型进行联合训练,在降低网络对离群点敏感度的同时提升正负样本在特征空间中的聚类效果,从而改善网络性能.大量对比实验表明本文方法在识别精度上有较明显提升,证明了本文提出算法的有效性.

参考文献

[1] 叶钰, 王正, 梁超, 等. 多源数据行人重识别研究综述[J]. 自动化学报, 2020, 46(9): 1869-1884.
YE Y, WANG Z, LIANG C, et al. A survey on multi-source person re-identification[J]. Acta Automatica Sinica, 2020, 46(9): 1869-1884. (in Chinese)

[2] ZHU Z H, JIANG X Y, ZHENG F, et al. Viewpoint-aware loss with angular regularization for person re-identification [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13114-13121.

[3] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: A benchmark[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2016: 1116-1124.

[4] YE M, WANG Z, LAN X Y, et al. Visible thermal person re-identification via dual-constrained top-ranking[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2018: 2.

[5] WANG G A, ZHANG T Z, YANG Y, et al. Cross-modality paired-images generation for RGB-infrared person re-identification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12144-12151.

[6] 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). Piscataway: IEEE, 2020: 618-626.

[7] CHENG D, GONG Y H, ZHOU S P, et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 1335-1344.

[8] LIU H J, CHENG J, WANG W, et al. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification[J]. Neurocomputing, 2020, 398: 11-19.

[9] YE H R, LIU H, MENG F Y, et al. Bi-directional exponential angular triplet loss for RGB-infrared person re-identification[J]. IEEE Transactions on Image Processing, 2021, 30: 1583-1595.

[10] 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). Pisca-

taway: IEEE, 2020: 10254-10263.
[11] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2242-2251.
[12] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2014: 2672-2680.
[13] 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). Piscataway: IEEE, 2017: 5390-5399.
[14] 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.
[15] 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). Piscataway: IEEE, 2016: 770-778.
[16] 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). Piscataway: IEEE, 2020: 1487-1495.
[17] MÜLLER R, KORNBLITH S, HINTON G. When does label smoothing help? [J]. Advances in Neural Information Processing Systems, 2019, 32: 4694-4703.
[18] NGUYEN D, HONG H, KIM K, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17 (3): 605.
[19] ZHONG Z, ZHENG L A, KANG G L, et al. Random erasing data augmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13001-13008.
[20] WANG G A, ZHANG T Z, CHENG J, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 3622-3631.
[21] 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.
[22] YE M, SHEN J B, CRANDALL D J, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]//Computer Vision—ECCV 2020. Cham: Springer International Publishing, 2020: 229-247.
[23] 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). Piscataway: IEEE, 2022: 12026-12035.
[24] LIU H J, MA S, XIA D X, et al. SFANet: A spectrum-aware feature augmentation network for visible-infrared person reidentification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(4): 1958-1971.
[25] LIU H J, CHAI Y X, TAN X H, et al. Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification[J]. IEEE Signal Processing Letters, 2021, 28: 653-657.

作者简介

周　非　男，1977 年 6 月出生，湖北浠水人．重庆邮电大学教授，博士生导师，主要研究方向为信息与信号处理、机器视觉、信息安全．
E-mail: zhoufei@cqupt.edu.cn

舒浩峰（通讯作者）　男，1998 年 11 月出生于重庆市，重庆邮电大学硕士研究生，主要研究方向为行人重识别．
E-mail: s200131214@stu.cqupt.edu.cn