计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(9):269−275 [doi: 10.15888/j.cnki.csa.009604]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于注意力特征融合的跨模态行人重识别①

邓淑雅, 李浩源

(南京信息工程大学 计算机学院、网络空间安全学院, 南京 210044)
通信作者: 邓淑雅, E-mail: dengshuya1999@163.com

摘　要: 跨模态行人重识别任务旨在匹配同一行人的可见光图像和红外图像, 在智能安全监控系统中广泛应用. 由于可见光模态和红外模态存在固有的模态差异, 给跨模态行人重识别任务在实际应用过程中带来了巨大的挑战. 为了缓解模态差异, 研究人员提出了很多有效的解决方法. 但是由于这些方法提取的是不同模态之间的特征, 彼此缺少对应的模态信息, 导致特征缺少充分的鉴别性. 为了提高模型提取特征的鉴别性, 本文提出基于注意力特征融合的跨模态行人重识别方法. 通过设计高效的特征提取网络和注意力融合模块, 并在多种损失函数的优化下, 实现不同模态信息的融合和模态对齐, 从而促进模型匹配行人准确度的提升. 实验结果表明, 本方法在多个数据集上都取得了很好的性能.

关键词: 跨模态行人重识别; 注意力机制; 特征融合; 模态差异; 模态对齐

引用格式: 邓淑雅,李浩源.基于注意力特征融合的跨模态行人重识别.计算机系统应用,2024,33(9):269–275. http://www.c-s-a.org.cn/1003-3254/9604.html

Cross-modality Person Re-identification Based on Attention Feature Fusion

DENG Shu-Ya, LI Hao-Yuan

(School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Cross-modality person re-identification is widely used in intelligent safety monitoring systems, aiming to match visible light images and infrared images of the same person. Due to the inherent modality differences between visible and infrared modalities, cross-modality person re-identification poses significant challenges in practical applications. To alleviate modality differences, researchers have proposed many effective solutions. However, existing methods extract different modality features without corresponding modality information, resulting in insufficient discriminability of the features. To improve the discriminability of the features extracted from models, this study proposes a cross-modality person re-identification method based on attention feature fusion. By designing an efficient feature extraction network and attention feature fusion module, and optimizing multiple loss functions, the fusion and alignment of different modality information can be achieved, thereby promoting the model matching accuracy for persons. Experimental results show that this method achieves great performance on multiple datasets.

Key words: cross-modality person re-identification; attention mechanism; feature fusion; modality difference; modality alignment

行人重识别 (person re-identification, ReID)[1]在智能安全监控系统中是一项至关重要的任务, 旨在将查询集中的单个行人图像与不同摄像机拍摄的图库集中的图像进行匹配. 传统的 ReID 方法侧重于匹配可见光摄像机拍摄的行人图像, 可以认定是一个在可见光场景下的单模态检索问题. 然而, 当行人出现在光线较差

① 基金项目: 江苏省研究生科研创新计划 (KYCX23_1369)
收稿时间: 2024-03-04; 修改时间: 2024-04-03; 采用时间: 2024-04-10; csa 在线出版时间: 2024-07-24
CNKI 网络首发时间: 2024-07-25

或黑暗的环境下时, 可见光摄像机通常无法提供准确的外观信息, 导致这种方法具有局限性. 为了克服这个限制, 监控系统引入了红外摄像机, 在夜间可以拍摄清晰的红外模态行人图像. 为了可以匹配可见光图像和红外图像中的同一行人, 可见光-红外的跨模态行人重识别 (visible-infrared person re-identification, VI-ReID)[2] 成为一个日益受关注的问题.

VI-ReID 主要实现在黑暗环境下捕获的红外图像和光线较好情况下拍摄的可见光图像之间的行人匹配. 通过结合可见光模态和红外模态的图片数据, VI-ReID 试图在极具挑战性的照明条件下提高行人匹配的准确性和稳健性. 但是在实际应用过程中, VI-ReID 面临着很多挑战. 一方面, 和 ReID 一样, 由于待检索的图像都是不同的摄像机在不同时间地点下拍摄的, 导致图像出现背景噪声、遮挡、行人姿态变化、拍摄异常等问题. 另一方面, VI-ReID 面临的最主要挑战是红外模态和可见光模态之间存在的模态差异.

为了缓解模态差异, 研究人员从不同方面提出了许多有效的解决方案, 这些方法主要可以分为基于表征学习、度量学习以及模态互换 3 类. 其中, 基于表征学习的方法主要研究如何设计合理的网络架构, 提取不同模态中具有鉴别性的特征, 以减少模态差异. Wu 等人[3]首次提出深度零填充 (zero-padding) 的方法学习共享模态特征. 为了进一步增强特征表示的能力, Ye 等人[4]提出了一种动态双注意聚合 (dynamic dual-attentive aggregation, DDAG) 学习方法来挖掘模态内部和跨模态图像之间的上下文线索. Wu 等人[5]提出了一种联合模态和模式对齐网络 (joint modality and pattern alignment network, MPANet) 来发现跨模态之间的细微差别. Zhang 等人[6]提出了一种多样化嵌入扩展网络 (diverse embedding expansion network, DEEN), 可以有效地生成不同的嵌入来学习特征表示. 此外, 基于度量学习的方法旨在设计一个合理的度量方法或损失函数来学习图像的相似度. Jia 等人[7]提出了一种相似性推理度量 (similarity inference metric, SIM), 通过连续相似图推理和相互最近邻推理, 利用模态内样本相似性来挖掘跨模态样本之间的相似性. Liu 等人[8]提出了一种记忆增强单向度量 (memory-augmented unidirectional metric, MAUM) 学习方法, 通过单向度量和基于记忆增强的两种设计实现跨模态关联. Miao 等人[9]通过层次特征约束 (hierarchical feature constraint, HFC) 对全局特征和局部特征进行学习, 全局特征使用知识蒸馏策略对局部特征进行监督. 还有基于模态转换的方法主要通过生成对抗性网络 (generative adversarial network, GAN) 实现可见光图像和红外图像之间模态转换, 如 alignGAN[10]、TS-GAN[11]、JSIA[12]、FMCNet[13]等, 它们在很大程度上减少了模态之间存在的差异.

尽管如此, 由于成像过程的异质性, 红外图像中的相同灰度在可见光图像中可能是完全不同的颜色, 不同模态的行人图像中彼此缺少对应的模态信息, 导致网络提取的特征缺失足够的鉴别性, 从而影响模型的性能. 为了充分融合红外模态和可见光模态的行人身份信息, 使得提取的特征具有足够的鉴别性, 促进模型实现行人匹配, 本文提出一种基于注意力特征融合的跨模态行人重识别方法. 首先以 ResNet-50[14]构建双流结构的特征提取网络, 分别提取两种模态的特定特征和共享特征. 为了提高特征的鉴别性, 本文设计了一种注意力融合模块, 通过局部注意力模块和全局注意力模块将提取特征之间的模态信息进行融合. 此外, 由于不同模态的融合特征之间仍然存在模态差异, 为了进一步减少模态差异, 本文还引入了一种最大均值差异损失优化模型, 通过减少红外模态和可见光模态特征之间的分布差异以实现模态对齐. 本文主要贡献如下.

(1) 提出一种注意力融合模块, 利用局部和全局注意力机制分别将特征中显著的身份信息提取出来, 并通过融合操作使得最终提取的特征具有充分的鉴别性.

(2) 设计一种由最大均值差异损失、身份损失和中心聚类损失联合优化模型的方法, 减少可见光模态和红外模态中同一行人的分布差异, 促进模型更准确地进行行人匹配.

(3) 本方法在 SYSU-MM01 和 RegDB 两个公开的跨模态数据集上进行了实验, 结果领先其他很多算法, 证实了有效性.

1 本文方法

为了融合红外模态和可见光模态彼此相关的身份信息, 使得模型提取的特征更具有鉴别性, 本文提出一种基于注意力特征融合的跨模态行人重识别方法. 如图 1 所示, 本文提出的方法由特征提取网络、注意力融合模块和损失优化 3 个部分组成. 首先, 将可见光图像和红外图像分别输入到特征提取网络的两个分支中,

通过特征提取网络对图像中的行人进行特征提取. 其次, 从特征提取网络获取的可见光特征和红外特征分别输入到注意力融合模块中, 通过局部注意力和全局注意力将特征中显著的行人信息进行提取, 并通过融合操作将信息进行融合, 最终可以得到更具有鉴别性的特征用于行人匹配. 最后, 为了缓解模态差异, 本文利用最大均值差异和中心聚类损失联合优化模型. 接下来对本方法中的各个部分做详细介绍.

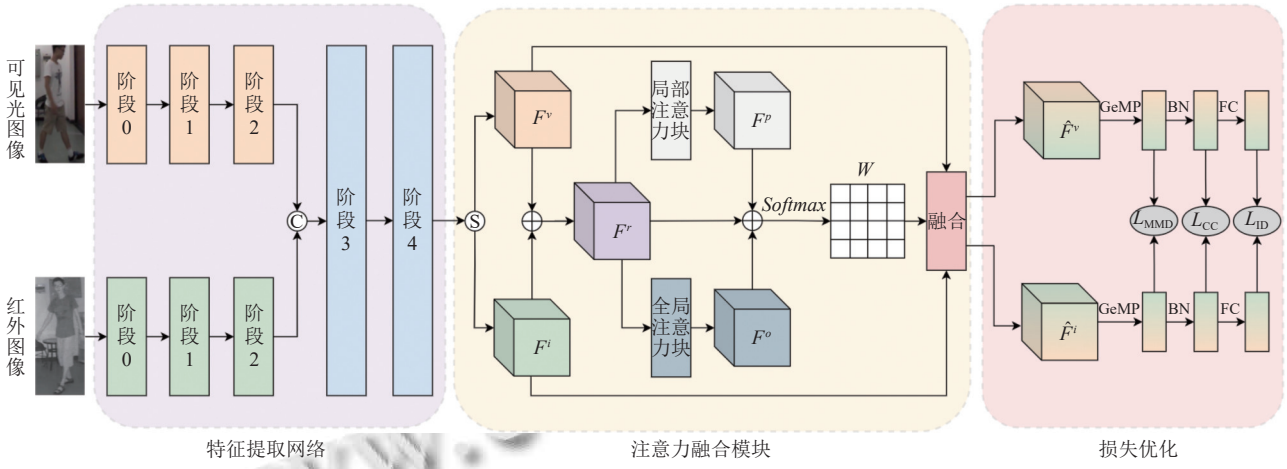


图 1 基于注意力特征融合的跨模态行人重识别方法框架

1.1 特征提取网络

本文利用 ResNet-50 为基础构建双流结构的特征提取网络, 网络共分为 5 个阶段, 每个阶段的结构和 ResNet-50 每个阶段的结构相同. 其中, 前 3 个阶段设置成参数不共享的两个分支, 分别用来提取可见光图像和红外图像的特定模态特征. 其余两个阶段设置成共享模态阶段, 主要是将两个特定模态的特征沿批次维度进行拼接形成一个整体特征, 得到的整体特征通过参数共享的阶段 3 和阶段 4 进一步提取共享模态特征, 然后得到特征提取网络最终的输出结果.

1.2 注意力融合模块

首先, 将特征提取网络最终的输出结果按批次维度重新拆分成可见光特征F^v和红外特征F^i, 即为注意力特征模块的输入. 为了接下来可以同时获取可见光模态和红外模态的信息, 将可见光特征F^v和红外特征F^i先进行简单的相加融合, 得到一个初步融合特征F^r, 具体表示为:

$$F^r = F^v \oplus F^i \tag{1}$$

其中, $\oplus$是元素级相加操作.

为了保持模型轻量化, 本文利用局部注意力机制和全局注意力机制通过改变空间池化的大小, 在不同尺度上实现通道注意, 从而提取初步融合特征F^r中可见光和红外模态共同的行人身份信息. 局部注意力块和全局注意力块的结构如图 2 所示.

如图 2(a) 所示, 局部注意力块主要由两个点卷积层, 两个批归一化层 (batch normalization, *BN*) 和一个 *ReLU* 层组成. 局部注意力块作为局部通道上下文聚合器, 只对初步融合特征F^r中每个空间位置进行通道交互作用, 得到包含局部身份信息的局部融合特征F^p. 具体可以表示为:

$$F^p = BN(Conv_{1\times1}(ReLU(BN(Conv_{1\times1}(F^r))))) \tag{2}$$

其中, $Conv_{1\times1}(\cdot)$表示卷积核大小为 1 的点卷积层, $BN(\cdot)$表示批归一化层, $ReLU(\cdot)$表示 *ReLU* 激活函数.

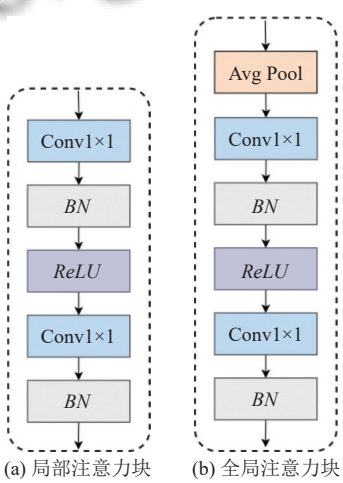


图 2 局部注意力块和全局注意力块的结构图

如图 2(b) 所示, 全局注意力块主要由一个平均池化层、两个点卷积层、两个批归一化层和一个 *ReLU*

层组成. 全局注意力块主要聚合初步融合特征F^r中的全局上下文信息, 得到包含全局身份信息的全局融合特征F^o. 具体可以表示为:

$$F^o = Avg(BN(Conv_{1\times1}(ReLU(BN(Conv_{1\times1}(F^r))))))\tag{3}$$

其中, $Avg(\cdot)$表示平均池化, $Conv_{1\times1}(\cdot)$表示卷积核大小为 1 的点卷积层, $BN(\cdot)$表示批归一化层, $ReLU(\cdot)$表示 $ReLU$ 激活函数.

接下来将初步融合特征F^r、局部融合特征F^p和全局融合特征F^o逐元素相加得到中间融合特征, 然后通过 Softmax 函数计算出中间融合特征中所包含身份信息的注意力权重 W. 根据注意力权重 W, 本文通过融合操作对可见光特征F^v和红外特征F^i中彼此缺失的模态信息进行补充. 融合的操作过程如图 3 所示.

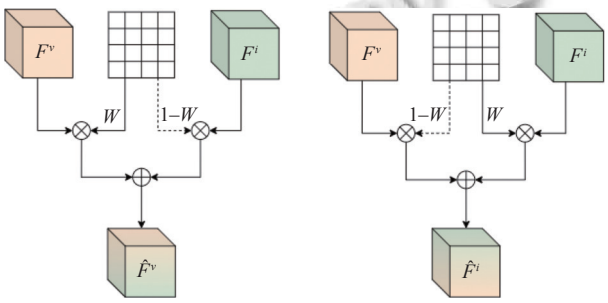


图 3 注意力融合模块中融合的具体过程

在融合过程中, 以可见光特征F^v、红外特征F^i和注意力权重 W 作为输入, 分别进行可见光特征中缺失的红外信息融合和红外特征中缺失的可见光信息融合. 以可见光特征中缺失的红外信息融合的过程为例, 将不同的权重分别和可见光特征F^v、红外特征F^i相乘, 并将相乘的结果进行相加得到最终的可见光融合特征$\hat{F}^v$, 具体的过程可以表示为:

$$\hat{F}^v = (F^v \otimes W) \oplus \left(F^i \otimes (1-W)\right)\tag{4}$$

其中, $\oplus$是元素级相加操作, $\otimes$是元素级相乘操作. 同样地, 红外特征中缺失的可见光信息融合的过程可以表示为:

$$\hat{F}^i = \left(F^i \otimes W\right) \oplus (F^v \otimes (1-W))\tag{5}$$

1.3 损失优化

本文使用最大均值差异损失L_{MMD}、身份损失L_{ID}和中心聚类损失L_{CC}对所提出的网络进行优化. 具体来说, 为了缓解模态差异, 本文引入一种最大均值差异损失L_{MMD}, 通过对融合特征进行约束, 减少可见光模态和红外模态之间的特征分布, 实现模态对齐. 最大均值差异损失L_{MMD}具体可以表示为:

$$L_{\mathrm{MMD}} = \frac{1}{P}\sum_{q=1}^{P}\left(MMD^2\left(G\left(\hat{F}_q^v\right), G\left(\hat{F}_q^i\right)\right)\right)\tag{6}$$

其中, P 为每批处理中的身份数, $G(\cdot)$表示广义平均池化 (generalized mean pooling) 操作, $\hat{F}_q^v$表示第 q 个身份行人的可见光特征分布, $\hat{F}_q^i$表示第 q 个身份行人的红外特征分布. $MMD^2(\cdot)$可以进一步表示为:

$$\begin{aligned} MMD^2\left(F^1, F^2\right) &= \left\| \frac{1}{K^v}\sum_{k^v=1}^{K^v}\phi\left(F_{k^v}^1\right) - \frac{1}{K^i}\sum_{k^i=1}^{K^i}\phi\left(F_{k^i}^2\right)\right\|^2 \\ &= \frac{1}{(K^v)^2}\sum_{k^v=1}^{K^v}\sum_{k^{v\prime}=1}^{K^v}\phi\left(F_{k^v}^1\right)^{\top}\phi\left(F_{k^{v\prime}}^1\right) \\ &\quad + \frac{1}{(K^i)^2}\sum_{k^i=1}^{K^i}\sum_{k^{i\prime}=1}^{K^i}\phi\left(F_{k^i}^2\right)^{\top}\phi\left(F_{k^{i\prime}}^2\right) \\ &\quad - \frac{2}{K^vK^i}\sum_{k^v=1}^{K^v}\sum_{k^i=1}^{K^i}\phi\left(F_{k^v}^1\right)^{\top}\phi\left(F_{k^i}^2\right) \end{aligned}\tag{7}$$

其中, F^1和F^2代指$MMD^2(\cdot)$输入的两个可见光特征和红外特征, K^v和K^i表示可见光图片和红外图片的数量, $F_{k^v}^1$和$F_{k^i}^2$分别为第k^v张可见光图片行人特征和第k^i张红外图片行人特征, $\phi(\cdot)$是将两个模态特征映射到再生核希尔伯特空间的高斯核函数.

身份损失L_{ID}主要是将图像中每个不同身份的行人视为一个类别. 在网络训练过程中, 对于给定带有标签的输入图像, 通过最小化交叉熵的方式来让模型的预测结果尽可能接近真实标签, 从而提高行人识别的准确性和性能, 计算公式如下:

$$L_{\mathrm{ID}} = -\frac{1}{P}\sum_{q=1}^{P}y_q\log\left(c_q\right)\tag{8}$$

其中, P 为每批次中的行人身份数, y_i为第 i 个身份行人的标签, c_q是对第 q 个身份行人的预测结果.

中心聚类损失L_{CC}可以将红外模态和可见光模态中相同身份行人的中心特征距离拉近, 并增加不同的身份样本之间距离. 具体可以定义为:

$$\begin{aligned} L_{\mathrm{CC}} &= \frac{1}{N}\sum_{a=1}^{N}D\left(f_a, h_{y_a}\right) \\ &\quad + \frac{2}{P(P-1)}\sum_{k=1}^{P-1}\sum_{b=k+1}^{P}\left[\rho - D\left(h_{y_k}, h_{y_b}\right)\right]_+ \end{aligned}\tag{9}$$

其中, N 为当前批次中可见光图片和红外图片的数量, f_a 为第 a 个身份行人的特征, h_{y_a}、h_{y_k} 和 h_{y_b} 分别为当前批次中标签为 y_a、y_k 和 y_b 的特征的平均值, ρ 为特征中心之间最小距离值, $D(\cdot)$ 表示欧氏距离.

最后, 本方法的总损失函数 L 可以表示为:

$$L=\alpha L_{\mathrm{MMD}}+\beta L_{\mathrm{ID}}+\gamma L_{\mathrm{CC}} \tag{10}$$

其中, α, β 和 γ 是平衡每个损失项贡献的超参数.

2 实验分析

2.1 数据集和评价指标

本文接下来进行的实验均在目前主流的跨模态行人重识别数据集 SYSU-MM01[3] 和 RegDB[15] 上进行, 这两个数据集的详情如表 1 所示. 此外, 本文的所有实验的评价指标均为累积匹配特征曲线 (cumulative matching characteristics curve, CMC) 和平均检索精度 (mean average precision, mAP).

表 1 跨模态行人重识别常用数据集对比

数据集	行人数	图像数	可见光摄像机数	红外摄像机数
SYSU-MM01	491	38271	1	1
RegDB	412	8240	4	2

2.2 实验设置

本方法实验基于 PyTorch 实现, 硬件配置环境为 NVIDIA GeForce RTX 3090 显卡、内存 24 GB 和 CPU 为 i5. 在数据预处理阶段, 将所有图像的大小调整为 288×144, 并使用各种增强策略对图像进行处理, 包括随机水平翻转、随机擦除、随机裁剪、随机旋转和随机通道增强. 在训练阶段, 随机抽取 6 个身份的行人, 每个训练批次分别选择 4 张可见光图像和 4 张红外图像. 本方法使用 SGD 优化器进行了 100 次迭代, 初始学习率设置为 0.01. 在前 16 次迭代过程中采用线性的预热策略 (warm-up), 然后在第 20 次和第 30 次逐渐衰减 5 倍, 在第 45 次和第 60 次逐渐衰减 10 倍. 此外, 总损失函数的平衡参数 α、β 和 γ 分别设置为 0.2、1 和 0.75. 在测试阶段, 只使用由特征提取网络、身份损失和中心聚类损失组成的基线来测试行人图像的匹配结果. 为了进行公平的比较, 测试时所有的超参数都和训练时保持一致.

2.3 对比实验

本文在 SYSU-MM01 和 RegDB 两个主流的跨模态行人重数据集上和现有的一些方法进行了对比实验. 结果表明本文的方法具有优秀的性能, 超过了大多数现有的方法.

- 在 SYSU-MM01 数据集上的对比实验: 如表 2 所示, 其中, 加粗为最优结果, 下划线为次优结果, 本方法的性能优于所有对比的方法. 在 All-Search 模式下, 本方法达到了 72.74% 的 Rank-1 和 67.83% 的 mAP, 分别超过 CMIT 的 1.8% 和 2.32%. 此外, 在 Indoor-Search 模式下, 本方法实现了 76.50% 的 Rank-1 和 79.44% 的 mAP. 总的来说, 实验结果证明了本方法的有效性.

表 2 在 SYSU-MM01 数据集上的对比实验 (%)

对比方法	All-Search		Indoor-Search	
	Rank-1	mAP	Rank-1	mAP
AGW (2021)[16]	47.50	47.65	54.17	62.97
TS-GAN (2021)[11]	49.80	47.40	50.40	63.10
MPANet (2021)[5]	70.58	68.24	76.74	80.95
DFLN-ViT (2022)[17]	59.84	57.70	62.13	69.03
FMCNet (2022)[13]	66.34	62.51	68.15	74.09
CMIT (2022)[18]	<u>70.94</u>	<u>65.51</u>	<u>73.28</u>	77.18
TOPLight (2023)[19]	66.76	64.01	72.89	76.70
PMT (2023)[20]	67.53	64.98	71.66	76.52
CMInfoNet (2023)[21]	67.92	63.42	73.22	<u>76.81</u>
本文方法	**72.74**	**67.83**	**76.50**	**79.44**

- 在 RegDB 数据集上的对比实验: 如表 3 所示, 其中, 加粗为最优结果,下划线为次优结果.

表 3 在 RegDB 数据集上的对比实验 (%)

对比方法	Visible to Infrared		Infrared to Visible	
	Rank-1	mAP	Rank-1	mAP
AGW (2021)[16]	70.05	66.37	—	—
MPANet (2021)[5]	83.70	80.90	82.80	80.70
SMCL (2021)[22]	83.93	79.83	83.05	78.57
SPOT (2022)[23]	80.35	72.46	79.37	72.26
FMCNet (2022)[13]	89.12	**84.43**	88.23	<u>83.86</u>
DFLN-ViT (2022)[17]	**92.10**	82.11	**91.21**	81.62
PMT (2023)[20]	84.83	76.55	84.16	75.13
TOPLight (2023)[19]	85.51	79.95	80.94	76.10
CMInfoNet (2023)[21]	86.13	82.68	84.30	79.83
本文方法	<u>89.97</u>	<u>83.56</u>	<u>88.75</u>	**84.29**

由表 3 可见, 本方法的性能超过了大多数对比方法的性能. 其中, 在 Visible to Infrared 模式下, 本方法达到了 89.97% 的 Rank-1 和 83.56% 的 mAP, 不过相较于性能最好的对比方法 DFLN-ViT 在 Rank-1 指标上低了 2.13%. 此外, 在 Infrared to Visible 模式下, 本方法实现了 88.75% 的 Rank-1 和 84.29% 的 mAP, 同样比 DFLN-ViT 在 Rank-1 上低了 2.46%. 由于本文的骨干网络使用的是 ResNet-50, 而 DFLN-ViT 方法的骨干网络使用的是 ViT, ResNet-50 相较于 ViT, 捕捉样本内不同位置和通道之间相关性的能力不足, 导致模型提

取的行人特征信息有所欠缺. 此外, DFLN-ViT 对骨干网络每层中不同粒度的特征信息进行编码融合, 并且在每个通道上细化表示, 有助于从全局角度建模通道之间的长期依赖性. 而本文方法只对单一粒度的特征进行局部和全局的注意力融合, 缺少多粒度的身份信息. 不过总的来说, 实验结果超过了大多数的方法, 本方法依旧具有一定的竞争力.

2.4 消融实验

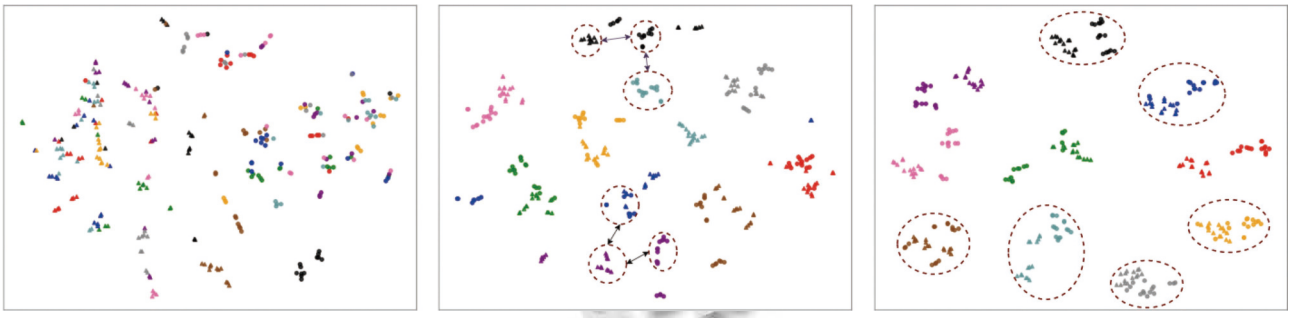
为了验证本方法中每个部分对模型性能的贡献度, 本文在 SYSU-MM01 数据集上进行了消融实验, 实验结果如表 4 所示. 其中 B 表示由特征提取网络、身份损失 L_{ID} 和中心聚类损失 L_{CC} 组成的基线, AFF 表示注意力融合模块, L_{MMD} 表示最大均值差异损失. 在基线的基础上分别加上注意力融合模块和最大均值差异损失之后, 模型的性能得到大幅度提升. 不仅如此当注意力融合模块和最大均值差异损失结合使用, 二者相互促进, 使得模型的达到了优秀效果. 总之, 每个部分对模型检测的性能都有帮助, 各个部分结合使用可以发挥出更优异的作用.

表 4 在 SYSU-MM01 数据集上的消融实验 (%)

组成部分			SYSU-MM01	
B	AFF	L_{MMD}	Rank-1	mAP
√	—	—	60.59	58.91
√	√	—	65.37	60.08
√	—	√	64.40	62.73
√	√	√	72.74	67.83

2.5 特征分布可视化

为了进一步分析本文所提出方法提取特征的分布情况, 从 SYSU-MM01 的测试集中随机抽取 10 个身份行人的图像利用 t-SNE 进行可视化实验. 对于每个身份, 随机选择 15 张可见光图像和 15 张红外图像. 从图 4 可以看出, 原始的特征分布杂乱无章. 在经过基线的作用下, 相同身份的红外特征和可见光特征逐渐靠近, 但是相同身份的特征和不同身份的特征之间的距离过于相似, 难以分辨模态距离. 在图 4(c) 中, 通过注意力融合和多种损失的共同优化下, 模型提取的特征分布形成了多个明显的簇团, 表明了相同身份行人之间的距离得到减小, 不同身份行人之间的距离得到扩大, 使得模型可以促进行人匹配.



(a) 原始特征分布 (b) 基线的特征分布 (c) 本文方法的特征分布

图 4 特征分布的可视化结果

3 结论与展望

针对现有跨模态行人重识别方法存在的特征缺失模态信息的问题, 本文提出了一种基于注意力特征融合的跨模态行人重识别方法. 首先, 构建基于 ResNet-50 的双流特征提取网络, 分别提取可见光图像和红外图像中行人的特定模态特征和共享模态特征. 然后, 设计一种注意力融合模块, 将特征提取网络提取出来的特征分割成红外特征和可见光特征并输入其中, 对两者彼此缺失的模态信息进行补偿, 使得最终的特征具有充分的鉴别性. 最后, 利用最大均值差异损失、身份损失和中心聚类损失对模型进行优化. 本文通过特征提取网络、注意力融合模块和多种损失的作用下, 提高了模型匹配行人的准确度.

参考文献

1 王琦, 刘志刚, 王淼, 等. 姿态驱动的局部特征对齐的行人重识别. 计算机系统应用, 2023, 32(4): 268–273. [doi: 10.15888/j.cnki.csa.009035]

2 陈丹, 李永忠, 于沛泽, 等. 跨模态行人重识别研究与展望. 计算机系统应用, 2020, 29(10): 20–28. [doi: 10.15888/j.cnki.csa.007621]

3 Wu AC, Zheng WS, Yu HX, *et al*. RGB-infrared cross-modality person re-identification. Proceedings of the 2017

IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 5390–5399.

4 Ye M, Shen JB, Crandall DJ, *et al*. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 229–247.

5 Wu Q, Dai PY, Chen J, *et al*. Discover cross-modality nuances for visible-infrared person re-identification. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4328–4337.

6 Zhang YK, Wang HZ. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 2153–2162.

7 Jia MX, Zhai YP, Lu SJ, *et al*. A similarity inference metric for RGB-infrared cross-modality person re-identification. Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama: IJCAI, 2020. 1026–1032.

8 Liu JL, Sun YF, Zhu F, *et al*. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 19344–19353.

9 Miao YQ, Huang NC, Ma X, *et al*. On exploring pose estimation as an auxiliary learning task for visible-infrared person re-identification. Neurocomputing, 2023, 556: 126652. [doi: 10.1016/j.neucom.2023.126652]

10 Wang GA, Zhang TZ, Cheng J, *et al*. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3622–3631.

11 Zhang ZY, Jiang S, Huang CZT, *et al*. RGB-IR cross-modality person ReID based on teacher-student GAN model. Pattern Recognition Letters, 2021, 150: 155–161. [doi: 10.1016/j.patrec.2021.07.006]

12 Wang GA, Zhang TZ, Yang Y, *et al*. Cross-modality paired-images generation for RGB-infrared person re-identification. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 12144–12151.

13 Zhang Q, Lai CZ, Liu JN, *et al*. FMCNet: Feature-level modality compensation for visible-infrared person re-identification. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 7339–7348.

14 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

15 Nguyen DT, Hong HG, Kim KW, *et al*. Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors, 2017, 17(3): 605. [doi: 10.3390/s17030605]

16 Ye M, Shen JB, Lin GJ, *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. [doi: 10.1109/TPAMI.2021.3054775]

17 Zhao JQ, Wang HZ, Zhou Y, *et al*. Spatial-channel enhanced Transformer for visible-infrared person re-identification. IEEE Transactions on Multimedia, 2023, 25: 3668–3680. [doi: 10.1109/TMM.2022.3163847]

18 Feng YJ, Yu J, Chen F, *et al*. Visible-infrared person re-identification via cross-modality interaction Transformer. IEEE Transactions on Multimedia, 2023, 25: 7647–7659. [doi: 10.1109/TMM.2022.3224663]

19 Yu H, Cheng X, Peng W. TOPLight: Lightweight neural networks with task-oriented pretraining for visible-infrared recognition. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 3541–3550

20 Lu H, Zou XZ, Zhang PP. Learning progressive modality-shared Transformer s for effective visible-infrared person re-identification. Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington: AAAI, 2023. 1835–1843.

21 Shi HC, Luo MD, Zhang XY, *et al*. Learning cross-modality information bottleneck representation for heterogeneous person re-identification. arXiv:2308.15063, 2023.

22 Wei ZY, Yang X, Wang NN, *et al*. Syncretic modality collaborative learning for visible infrared person re-identification. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 225–234.

23 Chen CQ, Ye M, Qi MB, *et al*. Structure-aware positional Transformer for visible-infrared person re-identification. IEEE Transactions on Image Processing, 2022, 31: 2352–2364. [doi: 10.1109/TIP.2022.3141868]

(校对责编: 张重毅)