第 61 卷 第 10 期/2024 年 5 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于掩模重构与动态注意力的跨模态行人重识别

张阔*，范馨月，李嘉辉，张干

重庆邮电大学通信与信息工程学院，重庆 400065

摘要 跨模态行人重识别是一项具有挑战性的行人检索任务。现有研究侧重于通过提取模态共享特征来减小模态间差异，忽视了对模态内差异和背景干扰的处理。为此，提出了一种掩模重构与动态注意力(MRDA)网络，该网络通过重构人体区域特征来消除背景杂波的影响，从而增强网络对背景变化的鲁棒性。此外，该网络结合了动态注意力机制，以过滤无关信息，动态挖掘并增强具有辨别力的特征表示，消除模态内差异的影响。实验结果显示：该网络在SYSU-MM01数据集的all-search模式下的第一个检索结果匹配成功的概率(Rank-1)和均值平均精度(mAP)分别达到70.55%和63.89%；在RegDB数据集的visible-to-infrared检索模式下的Rank-1和mAP分别达到91.80%和82.08%。在公共数据集上验证了所提方法的有效性。

关键词 行人重识别；跨模态；掩模重构；双流网络；动态注意力

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP231742

Cross-Modal Person Re-Identification Based on Mask Reconstruction with Dynamic Attention

Zhang Kuo*, Fan Xinyue, Li Jiahui, Zhang Gan

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, *China*

Abstract Cross-modal person re-identification is a challenging pedestrian retrieval task. Existing research focuses on reducing inter-modal differences by extracting modal shared features, while ignoring the processing of intra-modal differences and background interference. In this regard, a mask reconstruction and dynamic attention (MRDA) network is proposed to eliminate the influence of background clutter by reconstructing the features of human body regions, thereby enhancing the robustness of the network on background changes. In addition, the dynamic attention mechanism is combined to filter irrelevant information, dynamically mine and enhance the discriminating feature representations, and eliminate the influence of intra-modal differences. The experimental results show that the probability the first search result matches successfully (Rank-1) and mean average precision (mAP) in the all-search mode of the SYSU-MM01 dataset reach 70.55% and 63.89%, respectively. The Rank-1 and mAP in the visible-to-infrared retrieval mode of the RegDB dataset reach 91.80% and 82.08%, respectively. The effectiveness of the proposed method is verified on the public datasets.

Key words person re-identification; cross-modality; mask reconstruction; two-stream network; dynamic attention

1 引 言

近年来，随着计算机视觉技术的不断发展，行人重识别(Re-ID)[1]在视频监控、安保和智慧城市等领域得到越来越多的关注和应用，其目标是在多个不重叠的相机之间检索同一身份的行人。可见光相机在夜间的成像能力不足，在黑暗条件下无法收集具有区别性的图像。为解决这一问题，跨模态行人重识别使用红外相机来拍摄夜间图像，通过对两种相机收集的图像进行匹配，实现跨可见光模态和红外模态的行人检索。但由于不断变化的相机和背景环境，模态差异和背景杂波对跨模态匹配影响较大。此外，由于遮挡、光照和姿态变化等因素的存在，网络容易提取到过多无关信息，从而无法进行准确的跨模态匹配。

收稿日期：2023-07-17；**修回日期**：2023-09-12；**录用日期**：2023-10-09；**网络首发日期**：2023-10-23

基金项目：国家自然科学基金(62271096)

通信作者：*s210101189@stu.cqupt.edu.cn

针对以上问题，Wu 等[2]提出了一种单流深度零填充网络来提取模态共享特征，缓解了跨模态差异。Gao 等[3]提出了一种新的交叉模态知识蒸馏（CMKD）损失，在特征提取阶段缩小了特定模态特征间的差异。Sun 等[4]提出了一种水平划分网络 PCB（part-based convolutional baseline），将特征水平划分为若干个局部特征，隐式地利用行人身体部位来削弱背景杂波的影响。Dai 等[5]提出了一种跨模态生成对抗网络（cmGAN），从两个不同的模态图像中生成交叉模态图像，将所生成的图像和真实的图像组合以产生混合的多光谱图像，并结合身份损失和跨模态三元组损失来最小化类间差异并最大化跨模态相似性。Zhao 等[6]提出了 Spindle Net，该网络分别捕获来自行人不同身体区域的语义特征，将来自不同语义区域的特征进行合并，有效地保留了判别特征。这些方法通常侧重于通过对卷积网络提取到的特征进行处理来缓解背景杂波的影响，忽视了网络源头对输入图像的处理，不能较好地解决遮挡、光照和姿态变化等因素引起的模态内差异问题，同时特征划分和对齐会破坏图像的原有结构，从而产生额外的噪声。

本文提出了一种基于掩模重构与动态注意力（MRDA）的网络，旨在消除背景干扰影响并动态提取判别特征。由于相机环境不断变化，图像背景复杂多样，因此提出掩模重构模块（MRM），在图像输入端构造掩模，通过掩模重构人体区域和背景区域，以此来消除背景杂波的影响，增强网络对背景变化的鲁棒性。然后将来自不同模态的特征对齐投影到公共子空间中。针对遮挡、光照和姿态变化等因素引起的模态内差异较大的问题，设计了一种动态注意力模块（DAM），动态挖掘并增强人体占比权重较大的区域，以此获得具有辨别力的判别特征。在两个公共数据集上验证了 MRDA 网络的优越性。在 SYSU-MM01 数据集的 all-search 模式下，与当前性能表现较好的 MCLNet[7]相比，该网络的第一个检索结果匹配成功的概率（Rank-1）和均值平均精度（mAP）分别高出 5.15 和 1.91 个百分点。在 RegDB 数据集的 visible-to-infrared 检索模式下，该网络的 Rank-1 和 mAP 比 MCLNet 分别高出 11.49 和 9.01 个百分点，证明了相对于当前跨模态行人重识别方法，该网络具有较大优势。

2　方　　法

2.1　网络总体框架

可见光相机和红外相机的光谱特性及拍摄环境不同，导致模态差异和背景杂波的影响较大，并且低分辨率、遮挡和人体姿态不同等原因会使网络无法提取到具有辨别力的特征，因此，提出了一种适用于跨模态的 MRDA 网络，其网络结构如图 1 所示，包括掩模重构模块、特征提取模块和动态注意力模块。

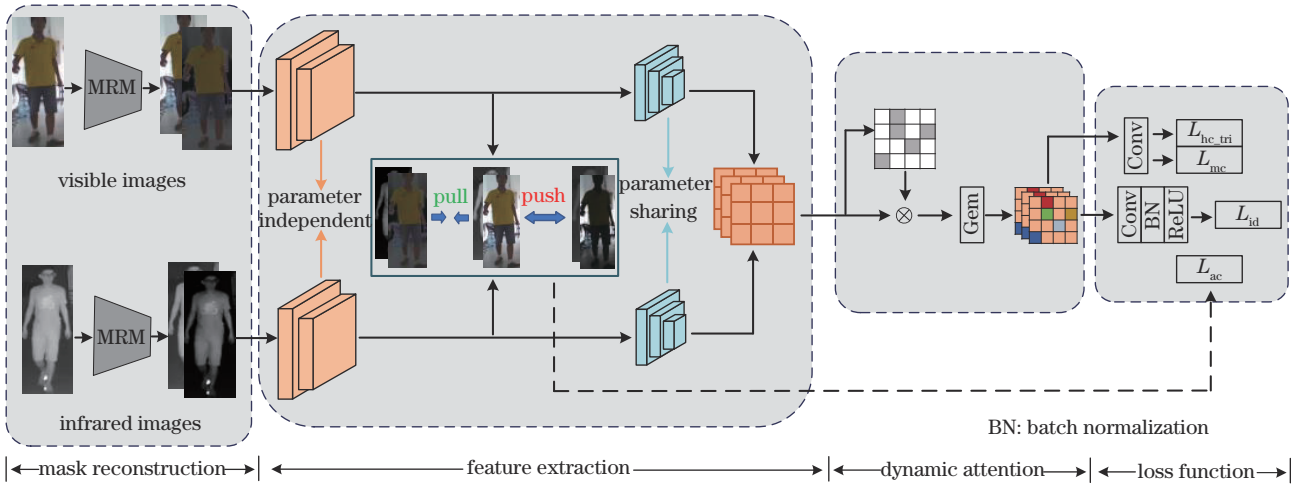

图 1　MRDA 网络结构

Fig. 1　MRDA network structure

2.2　掩模重构模块

随着深度学习在图像分割领域的应用迅速发展，出现了全卷积网络（FCN）[8]、条件随机区域（CRF）算法[9]、基于区域的卷积神经网络（Mask R-CNN）[10]以及大规模人类分割数据集[11]，它们可以较好地获得人体掩模。所提掩模重构模块通过最大-最小归一化来计算人体掩模，相对于复杂的 FCN 和 Mask R-CNN 网络结构来说简单且易于实现，无须进行复杂的模型训练或使用大规模数据集；相对于 CRF 等迭代算法来说，该模块计算复杂度较低、耗费时间短、所需计算资源更少。所提掩模重构模块通过在输入端生成掩模来重新构造图像的人体区域和背景区域，并在浅层卷积层约束原始图像特征与人体区域特征之间的距离，以缓解背景杂波的影响，其结构如图 2 所示。

在数据加载时，随机抽取 K 个行人的 M 张图像组成一个训练批次，其大小 $N=K\times M$，该集合表示为 $\boldsymbol{x}^{\mathrm{v}}=\{\boldsymbol{x}_i^{\mathrm{v}}|i=1,2,\cdots,N\}$，其中，$\boldsymbol{x}_i^{\mathrm{v}}=\{\boldsymbol{x}_i^{\mathrm{v}}|\boldsymbol{y}_i\}$ 表示训练批次中第 i 幅可见光图像，$\boldsymbol{y}_i$ 表示类别标签。同理，红

外图像集合表示为$\boldsymbol{x}^{\mathrm{r}}=\{\boldsymbol{x}_i^{\mathrm{r}}|i=1,2,\cdots,N\}$，其中，$\boldsymbol{x}_i^{\mathrm{r}}=\{\boldsymbol{x}_i^{\mathrm{r}}|\boldsymbol{y}_i\}$表示训练批次中第$i$幅红外图像。由于人体区域和背景区域像素值存在一定差别，因此通过最大-最小归一化来计算人体掩模，表示为

$$\boldsymbol{M}_{\mathrm{mask}}^{+}=\left\|\frac{\boldsymbol{x}-\min(\boldsymbol{x})}{\max(\boldsymbol{x})-\min(\boldsymbol{x})}\right\|_2,\tag{1}$$

式中：$\boldsymbol{x}$表示原始输入图像；$\boldsymbol{M}_{\mathrm{mask}}^{+}$表示人体掩模；$\|\cdot\|_2$表示L2范数。然后生成背景掩模$\boldsymbol{M}_{\mathrm{mask}}^{-}$来构建背景区域。为确保$\boldsymbol{M}_{\mathrm{mask}}^{+}$和$\boldsymbol{M}_{\mathrm{mask}}^{-}$能够构成成对关系，应满足成对约束条件：

$$\boldsymbol{M}_{\mathrm{mask}}^{+}+\boldsymbol{M}_{\mathrm{mask}}^{-}=1。\tag{2}$$

将这对掩模应用于原始输入图像$\boldsymbol{x}$，以生成人体区域和背景区域，表示为

$$\boldsymbol{x}_{\mathrm{M}^+}=\boldsymbol{M}_{\mathrm{mask}}^{+}\otimes\boldsymbol{x},\tag{3}$$

$$\boldsymbol{x}_{\mathrm{M}^-}=\boldsymbol{M}_{\mathrm{mask}}^{-}\otimes\boldsymbol{x},\tag{4}$$

式中：$\boldsymbol{x}_{\mathrm{M}^+}$和$\boldsymbol{x}_{\mathrm{M}^-}$分别表示人体区域和背景区域；$\otimes$表示空间加权操作。如图2所示，通过人体掩模重构的人体区域$\boldsymbol{x}_{\mathrm{M}^+}$保留了人体区域的全部信息，削弱了背景信息的表示。网络的权重共享结构可以捕获到更多的共享信息。在ImageNet[12]上的预训练权重通常对低级特征（如颜色或纹理）具有更强的依赖性，直接使用预训练模型可能会使网络关注到背景区域的颜色和纹理信息。因此采用将人体区域图像和原始图像合并输入的策略，通过人体区域$\boldsymbol{x}_{\mathrm{M}^+}$使网络学习到更多的人体特征，有效地减轻了来自背景区域的负面影响。

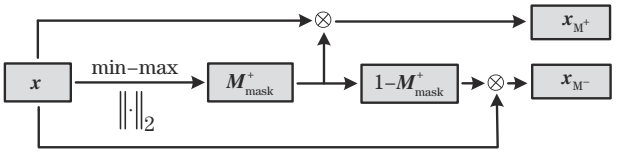

图2　掩模重构模块的结构

Fig. 2　Structure of the mask reconstruction module

2.3　特征提取模块

将ResNet50[13]作为骨干网络（ResNet50具有较深的网络结构，能够提取丰富的语义信息，有助于区分不同行人的特征），通过引入残差结构[13]来解决深层网络训练中的梯度消失问题。相对于ResNet18和ResNet34等低层次的网络结构，ResNet50拥有更多的层级和参数，可以提取更丰富和多样化的特征；相对于ResNet101和ResNet152等更深层次的网络结构，ResNet50具有较少的参数量，可以更快地进行训练和推理，并且网络层次过深会产生过拟合的风险，使网络模型的泛化性能下降。

采用双流网络的结构，如图1所示，网络由可见光路径和红外路径组成。将ResNet50作为骨干网络，其中：ResNet50的浅层卷积模块stage 0和stage 1参数不共享；深层卷积模块stage 2～stage 4的参数共享。两路径的具体操作为

$$\boldsymbol{f}^{\mathrm{v}}=\varphi_{\mathrm{v}}\left[\mathrm{cat}\left(\boldsymbol{x}^{\mathrm{v}},\boldsymbol{x}_{\mathrm{M}^+}^{\mathrm{v}}\right)\right],\tag{5}$$

$$\boldsymbol{f}^{\mathrm{r}}=\varphi_{\mathrm{r}}\left[\mathrm{cat}\left(\boldsymbol{x}^{\mathrm{r}},\boldsymbol{x}_{\mathrm{M}^+}^{\mathrm{r}}\right)\right],\tag{6}$$

式中：cat(•)函数将原始图像和重构的人体区域图像合并输入到网络中；φ_{v}和φ_{r}分别表示可见光和红外模态特征提取函数，用来提取可见光模态特定特征$\boldsymbol{f}^{\mathrm{v}}\in\mathbb{R}^{C\times H\times W}$和红外模态特定特征$\boldsymbol{f}^{\mathrm{r}}\in\mathbb{R}^{C\times H\times W}$，其中，$C$、$H$、$W$分别为特征的通道数、高和宽。由于在没有约束的情况下不能保证网络能够正确地学习到人体区域，因此，引入区域中心损失来引导人体区域特征的学习。区域中心损失表示为

$$L_{\mathrm{ac}}=\sum_{i=1}^{N}\left\|\boldsymbol{f}^{\mathrm{v,original}}-\boldsymbol{f}^{\mathrm{v,M}^+}\right\|_2+\sum_{i=1}^{N}\left\|\boldsymbol{f}^{\mathrm{r,original}}-\boldsymbol{f}^{\mathrm{r,M}^+}\right\|_2+\sum_{i=1}^{N}\left\|\boldsymbol{f}^{\mathrm{v,original}}-\boldsymbol{f}^{\mathrm{r,original}}\right\|_2,\tag{7}$$

式中：$\boldsymbol{f}^{\mathrm{v,original}}$和$\boldsymbol{f}^{\mathrm{r,original}}$是浅层卷积块提取的原始行人特征；$\boldsymbol{f}^{\mathrm{v,M}^+}$和$\boldsymbol{f}^{\mathrm{r,M}^+}$是浅层卷积块提取的人体区域特征。

如图3所示，通过约束L_{ac}将原始图像特征与人体区域特征相互拉近。由于人体特征$\boldsymbol{f}^{\mathrm{v,M}^+}$和背景区域特征$\boldsymbol{f}^{\mathrm{v,M}^-}$呈负相关关系，因此推远原始图像特征与背景区域特征的距离，以保证网络学习到更多的人体区域特征，从而消除背景区域的影响，增强网络对背景杂波的抗干扰能力，使网络对人体区域具有更强的感知能力。同时，约束可见光模态与红外模态的人体区域特征之间的距离，避免在重构时产生过大的模态差异。最后将$\boldsymbol{f}^{\mathrm{v}}$和$\boldsymbol{f}^{\mathrm{r}}$合并输入到共享卷积块中，表示为

$$\boldsymbol{f}=\varphi_{\mathrm{v,r}}\left[\mathrm{cat}\left(\boldsymbol{f}^{\mathrm{v}},\boldsymbol{f}^{\mathrm{r}}\right)\right],\tag{8}$$

式中，$\varphi_{\mathrm{v,r}}$表示模态共享特征提取函数。将合并后的两个模态特定特征投影到公共特征空间中，从而学习不同模态间共同的特征表示$\boldsymbol{f}\in\mathbb{R}^{C\times H\times W}$。

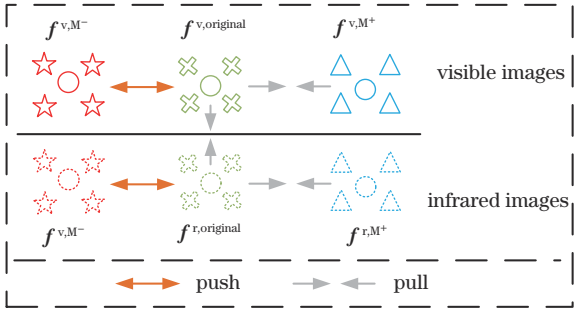

图3　区域中心损失示意图

Fig. 3　Schematic diagram of regional center loss

2.4　动态注意力模块

由于遮挡、光照和姿态变化等模态内差异因素的影响，网络难以提取到具有辨别力的重要特征。对此，文献[14-15]提出利用辅助姿态估计器和语义解析技术来对齐图像间的语义相关区域。但这些方法需要额外的数据集，并且计算开销大。文献[16]和文献[17]

分别提出瓶颈注意力模块(BAM)和卷积块注意力模块(CBAM),以不同的方式将空间注意力和通道注意力相结合,在空间和通道两个维度计算注意力权重,以获取全局与局部之间的联系,在全局中增强局部特征。文献[18]提出模态内加权局部注意力模块(IWPAM),以学习局部聚合特征,自适应地为不同特征分配不同权重。但这些方法均需要对全部特征进行运算,这会导致网络计算过多无用信息,浪费计算资源。对此,本研究提出了动态注意力模块,动态地寻找全局中权重较大的局部区域,挖掘并增强具有辨别力的特征。动态注意力模块的结构如图 4 所示。

首先将 ResNet50 网络提取到的特征 $\boldsymbol{f}$ 通过 1×1 的卷积层($\mathrm{Conv}_{1\times1}$)线性映射为 $\boldsymbol{Q}$、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 三个特征,表示为

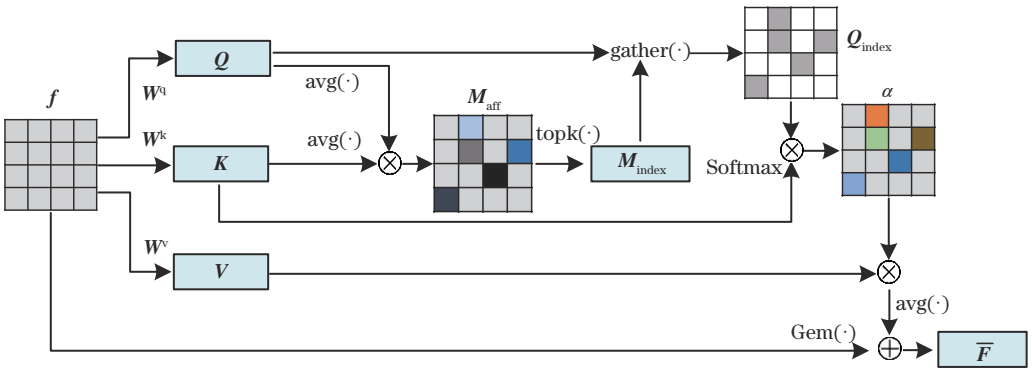

图 4　动态注意力模块的结构

Fig. 4　Dynamic attention module structure

$$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}=\mathrm{Conv}_{1\times1}(\boldsymbol{f}),\tag{9}$$

通过求 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 不同区域的均值,得到区域级特征 $\boldsymbol{Q}_{\mathrm{avg}}$ 和 $\boldsymbol{K}_{\mathrm{avg}}$;通过 $\boldsymbol{Q}_{\mathrm{avg}}$ 与 $\boldsymbol{K}_{\mathrm{avg}}$ 的矩阵运算得到区域亲和力矩阵 $\boldsymbol{M}_{\mathrm{aff}}$,表示为

$$\boldsymbol{Q}_{\mathrm{avg}}=\mathrm{avg}(\boldsymbol{Q}),\ \boldsymbol{K}_{\mathrm{avg}}=\mathrm{avg}(\boldsymbol{K}),\tag{10}$$

$$\boldsymbol{M}_{\mathrm{aff}}=\boldsymbol{Q}_{\mathrm{avg}}\cdot\boldsymbol{K}_{\mathrm{avg}}^{\mathrm{T}},\tag{11}$$

式中,avg(•)表示水平平均池化函数,用于求区域均值。矩阵 $\boldsymbol{M}_{\mathrm{aff}}$ 中的区域值衡量了特征 $\boldsymbol{f}$ 中不同区域的重要程度。保留其中前 k 个区域,得到区域索引矩阵 $\boldsymbol{M}_{\mathrm{index}}$,表示为

$$\boldsymbol{M}_{\mathrm{index}}=\mathrm{topk}(\boldsymbol{M}_{\mathrm{aff}}),\tag{12}$$

式中:$\boldsymbol{M}_{\mathrm{index}}$ 为区域索引矩阵,表示特征 $\boldsymbol{f}$ 中重要区域的位置索引;topk(•)表示保留特征中前 k 个最大元素的值和索引。通过位置索引提取 $\boldsymbol{Q}$ 中的 k 个区域,表示为

$$\boldsymbol{Q}_{\mathrm{index}}=\mathrm{gather}(\boldsymbol{Q},\boldsymbol{M}_{\mathrm{index}}),\tag{13}$$

式中,gather(•)表示按照索引矩阵从 $\boldsymbol{Q}$ 中选取指定位置的元素。$\boldsymbol{Q}_{\mathrm{index}}$ 过滤掉了大部分不相干信息,只保留了小部分重要的人体区域信息,以消除无关信息的干扰,提取更有辨别力的特征。

如图 5 所示,将 $\boldsymbol{Q}_{\mathrm{index}}$ 和 $\boldsymbol{K}$ 进行矩阵运算得到 k 个指定区域的注意力系数 α。将 α 与 $\boldsymbol{V}$ 进行加权求和得到注意力特征图,表示为

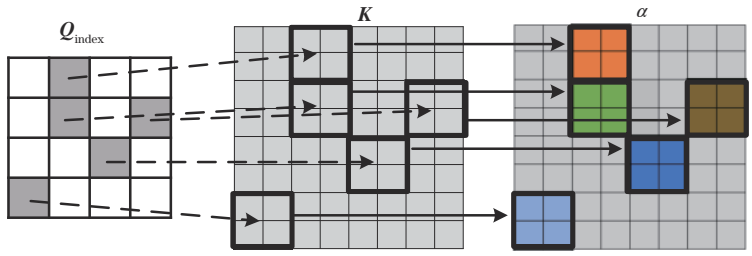

图 5　注意力系数获取示意图

Fig. 5　Schematic diagram of attention coefficient acquisition

$$\alpha=\mathrm{Softmax}(\boldsymbol{Q}_{\mathrm{index}}\cdot\boldsymbol{K}^{\mathrm{T}}),\tag{14}$$

$$\bar{\boldsymbol{f}}=\alpha\cdot\boldsymbol{V}^{\mathrm{T}},\tag{15}$$

式中:α 为归一化后的注意力系数;Softmax(•)表示 Softmax 函数;$\bar{\boldsymbol{f}}\in\mathbb{R}^{C\times H\times W}$ 表示特定区域的注意力特征图。最后,通过残差结构将注意力图 $\bar{\boldsymbol{f}}$ 叠加在原始特征 $\boldsymbol{f}$ 之上,表示为

$$\bar{\boldsymbol{F}}=\mathrm{Gem}(\boldsymbol{f})\oplus\mathrm{avg}(\bar{\boldsymbol{f}}),\tag{16}$$

式中:$\bar{\boldsymbol{F}}$ 是通过动态注意力模块增强后的特征(残差结构能够使深层网络更加稳定,避免出现梯度消失和梯度爆炸问题);⊕ 表示叠加操作;Gem(•)表示广义平均池化(generalized-mean)[19]操作,可有效降低数据维度,提高计算效率。给定一个三维特征 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$,则 Gem 操作表示为

$$\boldsymbol{x}_{\mathrm{Gem}}=\left(\frac{1}{|\boldsymbol{X}|}\sum_{\boldsymbol{x}_i\in\boldsymbol{X}}\boldsymbol{x}_i^{b}\right)^{\frac{1}{b}},\tag{17}$$

式中：$\boldsymbol{x}_{\mathrm{Gem}}$是广义平均池化后的特征；$b$是超参数。当$b\to\infty$时，Gem近似于最大池化操作；当$b\to1$时，Gem近似于平均池化操作。

与其他注意力模块的对比结果如表1所示。用BAM、CBAM和IWPAM替换所提动态注意力模块DAM并进行实验，从实验结果可以看出，所提模块的Rank-1和mAP高于BAM、CBAM和IWPAM，说明所提动态注意力模块能够有效获取具有辨别力的特征，验证了其有效性。

表1　DAM与其他注意力模块的对比

Table 1　Comparison of DAM and other attention modules

Module	Rank-1 /%	mAP /%
DAM	70.55	63.89
BAM	66.63	62.17
CBAM	66.74	60.95
IWPAM	67.79	61.29

2.5　目标函数

使用交叉熵损失和中心三元组损失[19]对MRDA网络进行约束，以优化实例级上的特征分布；通过标签平滑运算[20]防止模型训练过拟合。给定一个图像，y表示图像标签，p_i表示预测结果，则交叉熵损失表示为

$$L_{\mathrm{id}}=\sum_{i=1}^{N}-q_i\ln\left[\mathrm{Softmax}\left(p_i\right)\right]$$

$$\text{s.t }q_i=\begin{cases}1-\dfrac{N-1}{N}\xi, & y=i\\ \dfrac{\xi}{N}, & y\neq i\end{cases},\tag{18}$$

式中：N是训练集中的总类别数；ξ是一个常数，可以促使模型对训练集数据产生一定误差，提高模型的泛化能力。中心三元组损失的目的是促使来自同一类的特征中心相互接近（类内紧致）而来自不同类的特征中心相互远离（类间分离）。中心三元组损失表示为

$$\boldsymbol{C}_i^{\mathrm{v}}=\frac{1}{M}\sum_{j=1}^{M}\bar{\boldsymbol{F}}_i^{\mathrm{v},j},\quad \boldsymbol{C}_i^{\mathrm{r}}=\frac{1}{M}\sum_{j=1}^{M}\bar{\boldsymbol{F}}_i^{\mathrm{r},j},\tag{19}$$

$$L_{\mathrm{hc\_tri}}=\sum_{i=1}^{K}\left[\rho+\left\|\boldsymbol{C}_i^{\mathrm{v}}-\boldsymbol{C}_i^{\mathrm{r}}\right\|_2-\min_{\substack{n\in\{\mathrm{v},\mathrm{r}\}\\ j\neq i}}\left\|\boldsymbol{C}_i^{\mathrm{v}}-\boldsymbol{C}_j^{(n)}\right\|_2\right]_+ +$$

$$\sum_{i=1}^{K}\left[\rho+\left\|\boldsymbol{C}_i^{\mathrm{r}}-\boldsymbol{C}_i^{\mathrm{v}}\right\|_2-\min_{\substack{n\in\{\mathrm{v},\mathrm{r}\}\\ j\neq i}}\left\|\boldsymbol{C}_i^{\mathrm{r}}-\boldsymbol{C}_j^{(n)}\right\|_2\right]_+,\tag{20}$$

式中：$\bar{\boldsymbol{F}}_i^{\mathrm{v},j}$和$\bar{\boldsymbol{F}}_i^{\mathrm{r},j}$分别表示一个训练批次里第$i$个行人（$i=1,2,\cdots,K$）的第$j$张可见光和红外图像特征（$j=1,2,\cdots,M$）；$\boldsymbol{C}_i^{\mathrm{v}}$和$\boldsymbol{C}_i^{\mathrm{r}}$分别表示第$i$个行人可见光和红外模态特征类中心；$\boldsymbol{C}_j^{(n)}$表示跨模态负样本特征类中心；$\rho$为一个阈值，表示正样本距离与负样本距离之间的最小差异。$L_{\mathrm{hc_tri}}$关注每个行人的跨模态正样本特征中心和跨模态最难的负样本特征中心，可有效缩短类内距离，最大化类间差异。为消除不同模态之间的语义信息差异，保持模态间的同一性，缩小模态间差异，利用相关一致性损失[21]来约束不同模态间的距离，具体表示为

$$\boldsymbol{G}^{\mathrm{v}}=\frac{\bar{\boldsymbol{F}}^{\mathrm{v}}\cdot\left(\bar{\boldsymbol{F}}^{\mathrm{v}}\right)^{\mathrm{T}}}{\left\|\bar{\boldsymbol{F}}^{\mathrm{v}}\cdot\left(\bar{\boldsymbol{F}}^{\mathrm{v}}\right)^{\mathrm{T}}\right\|_2},\boldsymbol{G}^{\mathrm{r}}=\frac{\bar{\boldsymbol{F}}^{\mathrm{r}}\cdot\left(\bar{\boldsymbol{F}}^{\mathrm{r}}\right)^{\mathrm{T}}}{\left\|\bar{\boldsymbol{F}}^{\mathrm{r}}\cdot\left(\bar{\boldsymbol{F}}^{\mathrm{r}}\right)^{\mathrm{T}}\right\|_2},\tag{21}$$

$$L_{\mathrm{mc}}=\left\|\boldsymbol{G}^{\mathrm{v}}-\boldsymbol{G}^{\mathrm{r}}\right\|_2,\tag{22}$$

式中，$\boldsymbol{G}$是反映特征之间相似性的互相关性矩阵。通过优化L_{mc}拉近两个模态的距离，缓解模态差异。最终，总体损失表示为

$$L=L_{\mathrm{id}}+L_{\mathrm{hc\_tri}}+L_{\mathrm{mc}}+L_{\mathrm{ac}}。\tag{23}$$

3　实验结果与分析

3.1　数据集和评估协议

SYSU-MM01数据集[2]是中山大学提出的大规模跨模态行人重识别数据集，共采集了491个行人的图像，其中包含由4个可见光相机拍摄的287628幅可见光行人图像以及2个红外相机拍摄的15792幅红外行人图像。训练集有395个行人图像，测试集有96个行人图像。SYSU-MM01数据集包含all-search和indoor-search两种模式。在all-search模式下，gallery集由4个可见光相机拍摄的图像组成，query集由2个红外相机拍摄的图像组成；在indoor-search模式下，gallery集只包含2个可见光相机拍摄的图像，query集包含2个红外相机拍摄的图像。测试阶段包括mutil-shot和single-shot两种检索模式，在mutil-shot模式下的gallery集中，每个行人包含多张图像，而在sing-shot模式下的gallery集中，每个行人仅包含一张图像。对于all-search和indoor-search两种模式，均采用single-shot模式设置进行实验。

RegDB数据集[22]是一个小规模数据集，包含一个可见光相机和一个远红外相机采集的412个行人的图像，每个行人都有10张可见光图像和10张红外图像。为确保实验的科学性，利用10次交叉验证法将数据集进行随机划分，训练集和测试集各包含2060张可见光图像和2060张红外图像。测试阶段包括visible-to-infrared和infrared-to-visible两种检索模式，研究采用visible-to-infrared模式进行实验。

利用累积匹配特征（CMC）[23]和mAP作为评估指标，其中CMC指前R个检索结果中匹配成功的概率，用Rank-R表示。

3.2　实验设置

实验环境配置为GeForce RTX 2080TI、CUDA10和PyTorch1.7。在开始训练前，batch size设置为36，包含3个行人的18张可见光图像和18张红外图像。

对每一个行人，随机选择6张可见光图像和6张红外图像。在训练阶段，使用随机梯度下降(SGD)算法进行优化，动量参数设置为0.9。采用预热学习率策略，学习率初始值为0.1，经过20和50个epoch后分别衰减至0.10和0.01。

3.3　消融实验

在SYSU-MM01数据集的all-search模式下进行消融实验，以验证MRDA网络及所提模块的有效性。

1)特征提取模块。将输入数据进行预处理后输入到特征提取模块中，以此作为实验的基线网络(baseline)。如表2所示，baseline的Rank-1和mAP分别为63.21%和59.87%，说明特征提取模块能够较好地提取到行人特征。

表2　在SYSU-MM01数据集all-search single-shot模式下的实验结果

Table 2　Experimental results in all-search single-shot mode of the SYSU-MM01 dataset

Baseline	MRM	DAM	Rank-1 /%	mAP /%
√			63.21	59.87
√	√		68.39	61.82
√		√	68.47	63.09
√	√	√	70.55	63.89

2)掩模重构模块。如表2所示，将掩模重构模块加入baseline进行实验，Rank-1和mAP分别为68.39%和61.82%，较baseline分别提高了5.18和1.95个百分点，说明掩模重构模块能够很好地消除背景杂波的影响，使网络对人体区域具有更强的感知能力。

3)动态注意力模块。如表2所示，baseline加动态注意力模块的Rank-1和mAP分别为68.47%和63.09%，较baseline分别提高了5.26和3.22个百分点，说明动态注意力模块能较好地挖掘重要特征并消除无关信息的干扰，有效地提高模型的精度。

综上所述，各个模块都能有效地帮助模型提升其性能，掩模重构模块与动态注意力模块在协同工作时效果最好，Rank-1和mAP达到了70.55%和63.89%，较baseline分别提升了7.34和4.02个百分点。说明MRDA网络能够有效地消除光照、背景及姿态变化等因素的影响，缩小模态差异。

为探索在双流网络中进行掩模重构约束学习的具体位置，在ResNet50的若干个stage之后分别进行实验，以实现更好的掩模重构约束学习，构建具有独立参数的模态特定特征提取器和具有共享参数的模态共享特征提取器。

对卷积模块stage 0到stage 4后的特征分别进行实验，实验结果如图6所示。从曲线变化趋势可以看出：在stage 1之后进行掩模重构约束学习，网络的性能最好；若在浅层stage 0后进行约束学习，可能导致网络不能充分地提取到人体躯体特征；相反，若在较深层进行约束学习，网络提取到的背景区域过多，不能较好地约束原始图像特征和身体区域特征之间的距离。

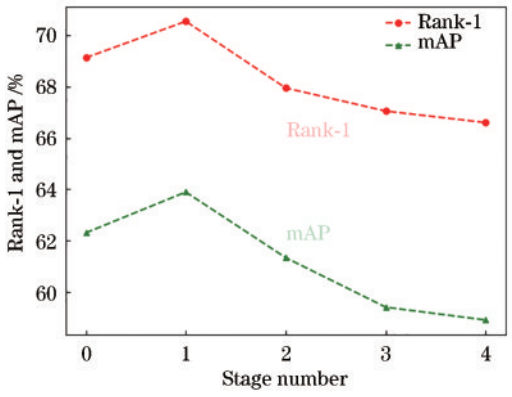

图6　在不同stage进行约束学习的实验结果

Fig. 6　Experimental results of constraint learning at different stages

3.4　可视化分析

为进一步分析MRDA网络的有效性，在SYSU-MM01数据集上对特征分布进行可视化实验。通过T-SNE算法[24]将高维特征向量转化为二维特征向量，图7(a)、(b)分别为baseline和MRDA网络的特征可视化结果，其中不同颜色表示不同的身份标签。与baseline相比，MRDA网络提取的特征能更好地聚集在一起，不同身份之间的边界更加明显，说明MRDA网络能更好地进行特征分类，更具有辨别力。图7(c)、(d)分别为baseline和MRDA网络特征距离的可视化结果。与baseline相比，MRDA网络的类内和类间特征距离平均值有所降低，说明MRDA网络能够有效地降低模态差异，提高类内相似度。

为了验证MRDA网络能挖掘并增强重要判别特征，通过Grad-CAM[25]绘制热力图来可视化图像特征(热力图反映了网络所关注的区域)。图8(a)、(b)分别为输入图像和MRDA网络的可视化结果。可以观察到，在MRDA网络中，人脸、胸口和腿部等身体部位被绘制为高亮，表示网络聚焦增强的身体区域，说明网络可以较准确地挖掘重要的判别特征，同时表现出对光照、遮挡和姿势变化等因素较强的抗干扰能力。

为验证网络的检索能力，将检索结果进行可视化，图9(a)、(b)分别为baseline和MRDA网络的检索结果。在SYSU-MM01数据集上将红外图像作为query集，可见光图像作为gallery集，检索对应的top-10重排序图像。其中，绿色实线框标记的图像表示正确检索的图像，红色虚线框标记的图像表示错误检索的图像。从检索结果可以看出，MRDA网络能够正确地区分行人身份，检索图像与待检索图像具有较高的匹配度。

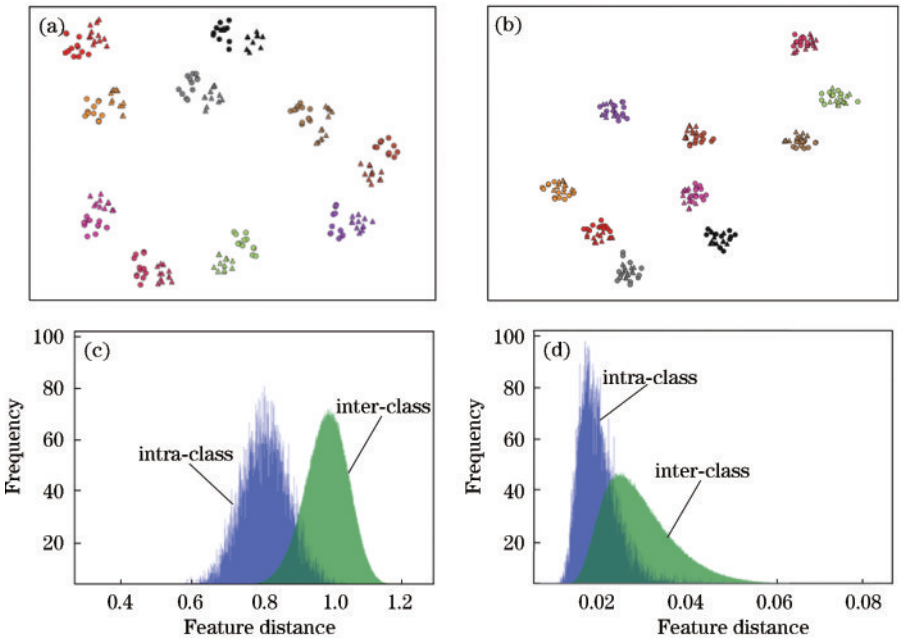

图 7　特征分布可视化。(a)Baseline 特征降维结果;(b) MRDA 网络特征降维结果;(c) baseline 特征距离分布;(d) MRDA 网络特征距离分布

Fig. 7　Feature distribution visualization. (a) Baseline feature dimensionality reduction results; (b) MRDA network feature dimensionality reduction results; (c) baseline feature distance distribution; (d) MRDA network feature distance distribution

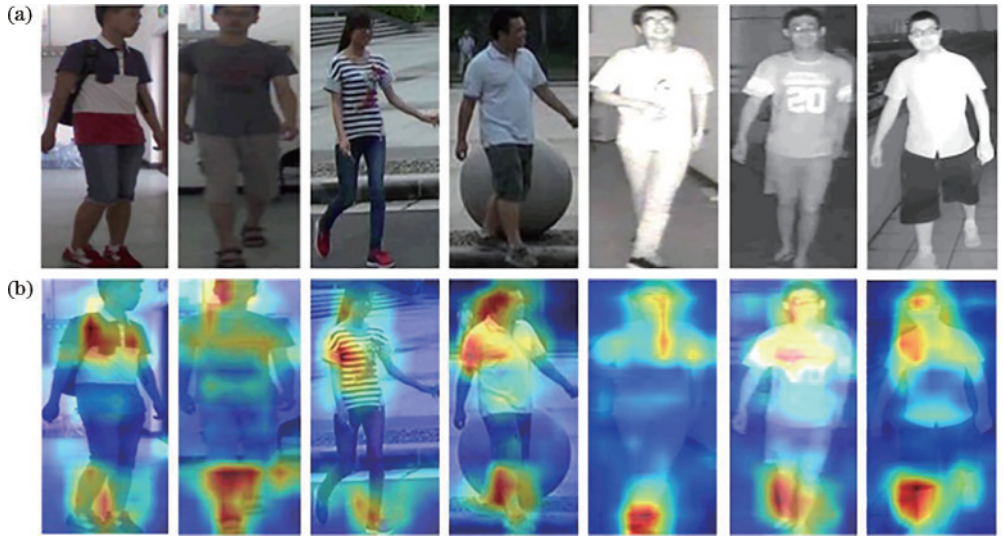

图 8　Grad-CAM 可视化结果。(a)输入图像;(b) MRDA 网络可视化结果

Fig. 8　Grad-CAM visualization results. (a) Input images; (b) MRDA network visualization results

3.5　与其他方法的对比

为验证 MRDA 网络的先进性,在 SYSU-MM01 和 RegDB 两个公共数据集上与近几年研究出的网络进行对比。加入对比的网络有 Zero-Pad[2]、cmGAN[5]、D2RL[26]、Hi-CMD[27]、DDAG[18]、AGW[1]、MCLNet[7]、DML[28]、MAUMG[29]和 MSFF[30]。表 3 和表 4 分别为在 SYSU-MM01 数据集的 all-search 和 indoor-search 两种模式下的实验结果。可以看出:所提网络与现有技术相比具有较强的竞争力,在 all-search 模式下 Rank-1 和 mAP 分别为 70.55% 和 63.89%;在 indoor-search 搜索模式下 Rank-1 和 mAP 分别为 72.69% 和 77.14%。与 AGW 和 MCLNet 等主要研究提取模态不变特征的网络相比,MRDA 网络致力于消除背景干扰因素的影响,提取具有辨别力的人体特征。与 cmGAN 和 Hi-CMD 等相比,MRDA 网络不需要耗费时间和空间来生成图像,避免了噪声的引入。与 DDAG 相同的是 MRDA 网络使用了注意力机制,与之不同的是 MRDA 网络可动态挖掘重要的身体区域而不需要进行全局计算,计算开销更少。在 RegDB 数据集上的实验结果如表 5 所示,MRDA 网络在 visible-to-infrared 检索模式下有较好的表现,Rank-1 和 mAP 分别为 91.80% 和 82.08%。

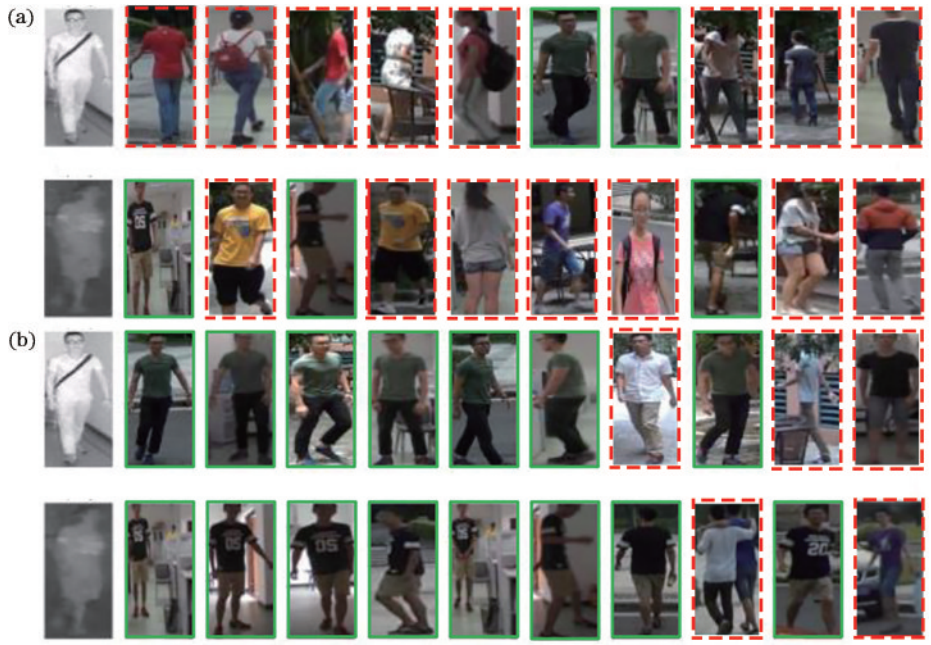

图 9 在 SYSU-MM01 数据集上获得的 top-10 检索结果。(a) Baseline 检索结果;(b) MRDA 网络检索结果

Fig. 9 Top-10 retrieval results obtained on the SYSU-MM01 dataset. (a) Baseline retrieval results; (b) MRDA network retrieval results

表 3 在 SYSU-MM01 数据集 all-search 模式下与其他网络的对比结果

Table 3 Comparison results with other networks in the all-search mode of the SYSU-MM01 dataset

Network	Rank-1 /%	Rank-10 /%	Rank-20 /%	mAP /%
Zero-Pad[2]	14.80	54.12	71.33	15.95
cmGAN[5]	26.97	67.51	80.56	27.80
D2RL[26]	28.90	70.60	82.40	29.20
Hi-CMD[27]	34.94	77.58	—	35.94
DDAG[18]	54.75	90.39	95.81	53.02
AGW[1]	47.50	84.39	92.14	47.65
MCLNet[7]	65.40	93.33	97.14	61.98
DML[28]	58.40	91.20	95.80	56.10
MAUMG[29]	61.59	—	—	59.96
MSFF[30]	62.93	93.68	97.67	60.62
MRDA	70.55	94.90	98.53	63.89

表 4 在 SYSU-MM01 数据集 indoor-search 模式下与其他网络的对比结果

Table 4 Comparison results with other networks in the indoor-search mode of the SYSU-MM01 dataset

Network	Rank-1 /%	Rank-10 /%	Rank-20 /%	mAP /%
Zero-Pad[2]	20.58	68.38	85.79	26.92
cmGAN[5]	31.63	77.23	89.18	42.19
DDAG[18]	61.02	94.06	98.41	67.98
AGW[1]	54.17	91.14	95.98	62.97
MCLNet[7]	72.56	96.88	99.20	76.58
DML[28]	62.40	95.20	98.70	69.50
MAUMG[29]	67.07	—	—	73.58
MSFF[30]	68.09	95.71	98.22	54.51
MRDA	72.69	97.15	98.73	77.14

表 5 在 RegDB 数据集上 visible-to-infrared 检索模式下与其他网络的对比结果

Table 5 Comparison results with other networks in the visible-to-infrared mode of the RegDB dataset

Network	Rank-1 /%	Rank-10 /%	Rank-20 /%	mAP /%
Zero-Pad[2]	17.75	34.21	44.35	18.90
cmGAN[5]	24.44	47.53	56.78	20.08
D2RL[26]	43.40	66.10	76.30	44.10
Hi-CMD[27]	34.94	77.58	—	35.94
DDAG[18]	72.37	—	—	69.09
AGW[1]	70.05	86.21	91.55	66.37
MCLNet[7]	80.31	92.70	96.03	73.07
DML[28]	77.60	—	—	84.30
MAUMG[29]	83.39	—	—	78.75
MSFF[30]	78.06	91.36	96.12	72.43
MRDA	91.80	97.46	98.67	82.08

4 结 论

针对跨模态行人重识别存在的模态内差异和背景干扰问题,提出了一种新颖的 MRDA 网络,该网络致力于通过掩模重构人体区域特征,动态提取具有辨别力的重要特征。一方面,通过掩模重构约束学习促使网络学习到更多的身体区域特征,减小背景区域的影响。另一方面,动态地寻找网络所关注的区域,挖掘并增强具有辨别力的特征。在两个公共数据集上进行消融实验和对比实验,实验结果验证了 MRDA 网络的先进性以及网络中各组成部分的有效性。下一步工作将对来自不同模态的特征进行特征匹配和对齐,进一步提升模型性能。

参 考 文 献

[1] 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.

[2] 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), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 5390-5399.

[3] Gao G W, Shao H, Wu F, et al. Leaning compact and representative features for cross-modality person re-identification[J]. World Wide Web, 2022, 25(4): 1649-1666.

[4] Sun Y F, Zheng L, Yang Y, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline) [M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11208: 501-518.

[5] Dai P Y, Ji R R, Wang H B, et al. Cross-modality person re-identification with generative adversarial training[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, July 13-19, 2018, Stockholm, Sweden. New York: ACM Press, 2018: 677-683.

[6] Zhao H Y, Tian M Q, Sun S Y, et al. Spindle Net: person re-identification with human body region guided feature decomposition and fusion[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 907-915.

[7] Hao X, Zhao S Y, Ye M, et al. Cross-modality person re-identification via modality confusion and center aggregation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 16383-16392.

[8] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence, May 24, 2016, New York: IEEE Press, 2016: 640-651.

[9] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[10] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2980-2988.

[11] Wu Z F, Huang Y Z, Yu Y N, et al. Early hierarchical contexts learned by convolutional networks for image segmentation[C]//2014 22nd International Conference on Pattern Recognition, August 24-28, 2014, Stockholm, Sweden. New York: IEEE Press, 2014: 1538-1543.

[12] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 248-255.

[13] 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), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[14] Miao J X, Wu Y, Liu P, et al. Pose-guided feature alignment for occluded person re-identification[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2020: 542-551.

[15] Kalayeh M M, Basaran E, Gökmen M, et al. Human semantic parsing for person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1062-1071.

[16] Park J, Woo S, Lee J Y, et al. BAM: bottleneck attention module[EB/OL]. (2018-07-18) [2023-06-19]. https://arxiv.org/abs/1807.06514.

[17] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[18] Ye M, Shen J B, Crandall D J, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12362: 229-247.

[19] 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.

[20] 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), June 16-17, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 1487-1495.

[21] Fu C Y, Hu Y B, Wu X, et al. CM-NAS: cross-modality neural architecture search for visible-infrared person re-identification[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 11803-11812.

[22] Nguyen D T, Hong H G, Kim K W, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.

[23] Moon H, Phillips P J. Computational and performance aspects of PCA-based face-recognition algorithms[J]. Perception, 2001, 30(3): 303-321.

[24] Laurens V D M, Hinton G. Visualizing data using T-

SNE[J]. Journal of Machine Learning Research, 2008, 9(2605): 2579-2605.

[25] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 618-626.

[26] 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), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 618-626.

[27] 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), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10254-10263.

[28] Zhang D M, Zhang Z Z, Ju Y, et al. Dual mutual learning for cross-modality person re-identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5361-5373.

[29] Liu J L, Sun Y F, Zhu F, et al. Learning memory-augmented unidirectional metrics for cross-modality person re-identification[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 19344-19353.

[30] 王凤随, 闫涛, 刘芙蓉, 等. 融合子空间共享特征的多尺度跨模态行人重识别方法[J]. 电子与信息学报, 2023, 45(1): 325-334.
Wang F S, Yan T, Liu F R, et al. Multi-scale cross-modal pedestrian re-recognition method integrating subspace sharing features[J]. Journal of Electronics & Information Technology, 2023, 45(1): 325-334.