面向跨模态行人重识别的双向动态交互网络

郑爱华[1,2,3], 冯孟雅[1,3], 李成龙[1,2,3]*, 汤进[1,3], 罗斌[1,3]

[1] (多模态认知计算安徽省重点实验室 合肥 230601)
[2] (安徽大学人工智能学院 合肥 230601)
[3] (安徽大学计算机科学与技术学院 合肥 230601)
(chenglongli@ahu.edu.cn)

摘 要: 为了解决当前跨模态行人重识别算法因采用权值共享的卷积核而造成模型针对不同输入动态调整能力差, 以及现有方法因仅使用高层粗分辨率的语义特征而造成信息丢失的问题, 提出一种双向动态交互网络的跨模态行人重识别方法. 首先通过双流网络分别提取不同模态各个残差块后的全局特征; 然后根据不同模态的全局内容动态地生成定制化卷积核, 提取模态特有信息, 并将其作为模态互补信息在模态间进行双向传递以缓解模态异质性; 最后对各层不同分辨率的特征进行相关性建模, 联合学习跨层的多分辨率特征以获取更具有判别性和鲁棒性的特征表示. 在 SYSU-MM01 和 RegDB 跨模态行人重识别数据集上的实验结果表明, 所提方法在第一命中率(R1)分别高于当前最好方法 4.70%和 2.12%; 在平均检索精度(mAP)上分别高于当前最好方法 4.30%和 2.67%, 验证了该方法的有效性.

关键词: 行人重识别; 跨模态; 动态卷积; 跨层多分辨率; 卷积神经网络
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023.19280

Bi-Directional Dynamic Interaction Network for Cross-Modality Person Re-Identification

Zheng Aihua[1,2,3], Feng Mengya[1,3], Li Chenglong[1,2,3]*, Tang Jin[1,3], and Luo Bin[1,3]

[1] (*Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Hefei* 230601)
[2] (*School of Artificial Intelligence, Anhui University, Hefei* 230601)
[3] (*School of Computer Science and Technology, Anhui University, Hefei* 230601)

Abstract: Current cross-modality person re-identification methods mainly use weight-sharing convolution kernels, which leads to poor dynamic adjustment ability of the model for different inputs. Meanwhile, they mainly use high-level coarse-resolution semantic features, which leads to great information loss. Therefore, this paper proposes a bi-directional dynamic interaction network for cross-modality person re-identification. Firstly, the global feature of different modalities after each residual block is extracted by the dual-flow network. Secondly, according to the global content of different modalities, it dynamically generates a customized convolution kernels to extract the modality-specific characteristics, followed by the integration of modality-complementary characteristics transferring between modalities to alleviate heterogeneity. Finally, the characteristics of different resolutions of each layer are modified to boost a more discriminative and robust

收稿日期: 2021-08-03; 修回日期: 2022-04-06. **基金项目**: 国家自然科学基金(61976002, 62076003); 安徽省高校协调创新项目(GXXT-2021-038, GXXT-2019-025); 安徽省重点研究与开发计划(202104d07020008); 安徽省高等学校省级自然科学研究项目(KJ2020A0033). **郑爱华**(1985—), 女, 博士, 副教授, 博士生导师, CCF 会员, 主要研究方向为人工智能、模式识别、计算机视觉、智能视频分析; **冯孟雅**(1995—), 女, 硕士研究生, 主要研究方向为跨模态行人重识别; **李成龙**(1988—), 男, 博士, 副教授, 硕士生导师, 论文通信作者, CCF 会员, 主要研究方向为视觉计算与学习、结构建模与深度学习、智能系统与应用; **汤进**(1976—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为计算机视觉、深度学习、多媒体大数据处理; **罗斌**(1963—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为数字图像处理、模式识别、计算机应用技术.

characteristic representation. Experimental results on two benchmark RGB-infrared person Re-ID datasets, SYSU- MM01 and RegDB demonstrate the effectiveness of the proposed method, which outperforms the state-of- the-art methods by 4.70% and 2.12% on R1 accuracy respectively, while 4.30% and 2.67% on mAP.

Key words: person re-identification; cross-modality; dynamic convolution; cross-layer multi-resolution; convolutional neural network

给定一幅监控行人图像，利用计算机视觉技术判断其是否出现于跨设备监控视域，该任务被称为行人重识别或行人再识别. 行人重识别被认为是图像检索的子问题，旨在弥补固定摄像头的视觉局限，可与行人检测、行人跟踪等任务相结合，应用于视频监控、公共安全、智慧城市等领域. 然而，由于跨相机行人姿态、光照、遮挡、视角变化等影响，使得类间差异小，类内差异大，成为行人重识别的核心挑战.

现有的行人重识别方法主要致力于基于可见光图像的单模态行人重识别问题，主流工作包括特征学习[1-2]和度量学习[3]. 虽然这些方法在行人重识别任务上取得了巨大进步，但是可见光相机在弱光照条件下成像的局限性限制了行人重识别在夜晚、恶劣天气等复杂场景的应用，而此类场景是安防监控的关键. 近年来，红外相机因可以克服可见光相机在弱光照条件下的成像局限性，被广泛应用于现实监控环境中. 许多监控相机在弱光条件下自动切换到夜间红外监控模式，因此，基于可见光-红外跨模态行人重识别技术也得到了广泛关注. 与传统的单模态行人重识别任务相比，跨模态行人重识别可以在夜晚、室内、恶劣天气等弱光照条件下呈现清晰的成像. 除了单模态行人重识别问题中存在的跨相机外观变化挑战外，跨模态行人重识别还面临着由可见光和红外相机成像差异导致的模态异质性挑战. 为了解决模态异质性带来的挑战，已有方法可归纳为 2 类：模态共享特征学习和模态特定特征补偿. 模态共享特征学习旨在将任意模态的特征嵌入同一特征空间，可见光图像的颜色、红外图像的热性等模态特定特征作为冗余信息被剔除；模态特定特征补偿旨在将模态特定特征作为补偿信息，弥补跨模态匹配过程中丢失的模态特定信息.

虽然现有方法在一定程度上解决了可见光与红外图像的模态异质性带来的挑战，但其卷积参数均在空间域共享. 权值共享无法针对不同模态的输入图像做出动态调整，无法更好地捕捉更具表达能力的模态特定特征. 针对这一问题，本文提出双向动态交互模块，利用特征图的全局内容学习一组权值；再与基础卷积核进行聚合，生成特定于输入的定制化卷积核；提取不同模态图像的特定信息并作为模态互补信息，在模态间双向传递. 为了使提取的特定信息具有身份鉴别性，使用身份损失进行约束，通过双向传递操作将模态特定信息在模态间相互传递以缓解模态差异.

现有的可见光-红外跨模态行人重识别方法仅利用网络学习到的高层特征用于行人图像匹配和检索，而这种特征的分辨率较低，往往丢失过多的细节空间信息. 卷积神经网络的高层特征主要编码语义信息，尽管语义信息在行人重识别及其他视觉问题上都发挥重要作用[4]，却严重丢弃了颜色和纹理等浅层信息，如行人服饰图案结构或背包、鞋帽配饰的特定形状等，而这些细节信息是重识别的关键线索. 针对该问题，本文提出跨层多分辨率特征融合模块，对不同残差块卷积网络层上的特征进行相关性建模，并通过可学习的权重矩阵线性加权得到最终的融合特征. 由于不同残差块卷积网络层的特征分辨率不同，且浅层及中层的高分辨率特征包含丰富的细节信息，高层的粗分辨率特征包含语义信息，因此通过跨层联合学习高分辨率和粗分辨率信息，能够有效地融合多分辨率的特征、增强特征的表达能力、提高模型在复杂场景中的准确率. 多层特征融合的思想已有效地应用在行人重识别领域中，常见的 2 类融合方法——直接融合[5-6]和多分类器[7]存在以下问题：直接融合. 由于跨模态任务中的模态异质性会引入噪声，因此影响模型性能. 多分类器方法. 需训练多个分类器，将成倍地增加训练成本；且由于浅层特征语义信息不足，强行在不同层建立身份的映射关系会增加噪声. 本文提出的跨层多分辨特征融合模块通过对不同残差块的不同分辨率特征的相关性建模，从而抑制噪声的影响，使之更适用于跨模态行人重识别，且无需训练多个分类器.

综上所述，本文提出一个双向动态交互网络，

包括 3 个组件: (1) 双流主干网络. 分别提取可见光模态和红外模态各个残差块不同分辨率的特征; (2) 双向动态交互模块. 针对不同模态的输入图像, 动态地生成定制化卷积核, 提取特定的信息, 并将其在模态间进行双向传递; (3) 跨层多分辨率特征融合模块. 对不同残差块卷积网络层上的不同分辨率的有效嵌入进行相关性建模, 得到最终的融合特征.

1　相关工作

1.1　行人重识别

行人重识别[8]的目的是给定一幅目标行人查询图像, 在匹配库中搜索与目标图像身份相同的行人图像. 通常将行人重识别任务构建为多类分类问题, 其中相同身份的行人图像属于同一类别. 目前的研究工作主要集中于利用深度学习方法来获取更具有判别能力的特征表示[9], 包括: (1) 基于身体部位的局部特征学习方法[10]. 利用深度网络模型更好地发现、对齐和表示身体部位. (2) 基于度量学习的方法. 设计合适的损失函数, 如对比损失[11]、三元组损失[12]等. (3) 基于解耦的方法[13-14]. 将每个样本分割成与身份相关和与身份无关的特征, 学习没有冗余信息的特征表示. (4) 基于对抗生成方法[15-16]. 解决不同摄像机拍摄图像及不同域图像的差异问题. (5) 基于注意力机制方法[17]. 通过引入注意力机制学习更鲁棒的特征表示, 取得了较好的效果. 上述方法都是单独处理每个样本, 忽略了行人和图像之间的关系. 最近也有一些工作利用自注意力[18]和图学习[19]的方法建立样本对之间的关系模型. Luo 等[20]使用谱特征变换来融合不同身份样本之间的特征; Shen 等[21-22]提出了组一致性约束条件和相似性引导图神经网络来融合不同样本的残差特征, 从而获得更鲁棒的表示. 为了解决监督问题中大量人工数据标注的弊端, 无监督行人重识别问题得到了广泛的发展[23-24]. 此外, 还有一些方法用于解决行人重识别的域适应问题[25-26]. 现有的行人重识别方法集中于可见光相机间的单模态行人重识别, 但可见光相机对光照的依赖性使其在弱光照条件下成像不清晰, 在复杂光照条件下成像差异大, 限制其在弱光照和复杂环境中的应用.

1.2　可见光-红外跨模态行人重识别

可见光-红外跨模态行人重识别[27-28]的目标是匹配同一个行人在不重叠相机下的可见光图像和红外图像, 其方法可划分为 2 类: 模态共享特征学习和模态特定特征补偿.

(1) 模态共享特征学习方法. Wu 等[29]首先创建 SYSU-MM01 数据集用于可见光-红外行人重识别的评估; Ye 等[30-31]使用双流网络结构提取不同模态的行人特征, 再利用双向限制排序损失约束共享特征嵌入, 进一步缩小 2 种模态数据间的差异; 随后, Ye 等[32]又提出基于模态感知的协同学习方法, 用于解决跨模态差异的问题. 为了提取模态不变特征, Hao 等[33]提出基于空间对齐和模态对齐的 2 种特征对齐方式来提取模态不变特征; Hao 等[34]又提出了一种具有分类约束和识别约束的双流超球面流形嵌入网络, 在超球面上约束模内变化和模态差异. 为了使共享特征中没有冗余信息, Dai 等[35]提出一种生成式对抗训练方法用于共享特征学习; Choi 等[36]提出分层的跨模态解纠缠算法, 自动地从可见光图像和红外图像中将包含身份的特征和不包含身份的特征分离出来, 最终只使用包含身份的特征进行重识别. 为了解决模态差异, Li 等[37]引入辅助 X 模态, 并将可见光-红外 2 个模态的任务重新定义为 X-可见光-红外的三模态学习任务; Jia 等[38]提出一种用于跨模态行人重识别的相似性度量方法. 上述方法都只关注共享特征的学习, 忽略了特定特征的作用.

(2) 模态特定特征补偿方法. 针对模态共享特征学习方法中存在的问题, 研究人员侧重于跨模态生成, 提出一系列基于模态特定特征补偿的方法. Kniaz 等[39]利用可见光图像生成对应的红外图像; Wang 等[40]提出基于双向循环生成的双级差异减少学习, 以缩小不同模态之间的差距; Lu 等[41]提出一种共享与特定特征变换算法, 同时使用共享特征和特定特征并进行相互转化, 达到目前较为先进的水平.

除了解决单模态行人重识别任务中光线变化、视角不同、姿势不同等挑战外, 跨模态行人重识别还需解决由不同光谱相机拍摄的图像带来的模态异质性挑战. 尽管上述方法在一定程度上解决了模态异质性带来的挑战, 但其卷积参数在空间域上权值共享; 针对行人重识别问题, 全局共享的加权向量无法较好地解决相同身份行人的局部差异问题. 并且现有工作主要使用高层信息作为最终的识别特征, 忽略了浅层及中间层不同分辨率特征对重识别的重要性, 限制特征表达的能力.

2 本文方法

为了学习不同模态图像的特有信息，并将其作为互补信息在模态间进行相互传递，缓解模态差异，同时充分挖掘多分辨率特征，本文提出一个用于跨模态行人重识别的双向动态交互网络，网络框架如图 1 所示. 该框架包括 3 个组件: (1) 双流主干网络; (2) 双向动态交互模块; (3) 跨层多分辨率特征融合模块. 首先使用 ResNet50[42]作为主干网络，获取各个残差块不同分辨率的特征及特征图; 然后在双向动态交互模块中使用卷积核预测网络利用全局内容预测一组权重系数，并与基础卷积核进行组合，生成特定于输入的定制化卷积核，利用生成的卷积核对原始的特征图进行特征提取，并将特征在 2 个模态间双向传递; 采用身份损失约束使提取的特征具有判别力; 最后在跨层多分辨率特征融合模块中使用多分辨率注意模块对相同维度的不同分辨率特征的相关性进行建模，得到注意力图，利用可学习权重对各个残差块不同分辨率的特征进行线性加权，得到最终的特征表示. 整个网络以端到端的方式进行训练.

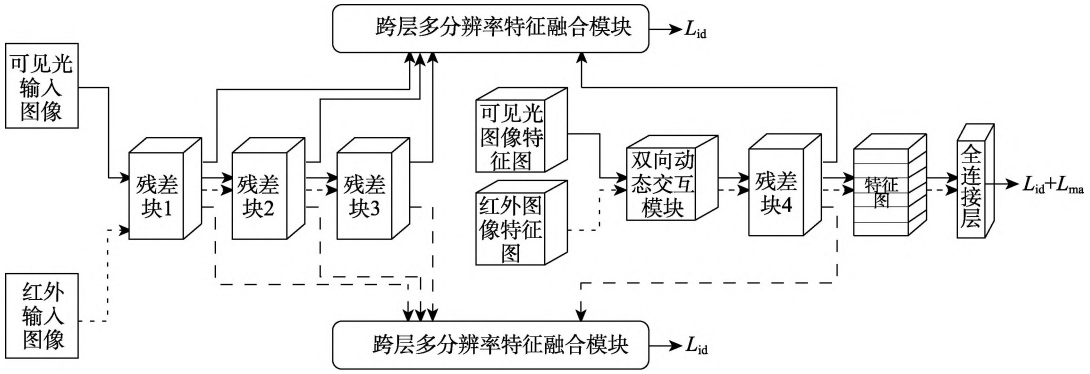

图 1 双向动态交互网络的结构示意图

2.1 双流主干网络

本文采用双流结构网络分别提取可见光和红外模态各个残差块不同分辨率特征，每个分支中可采用任何为图像分类而设计的深度卷积神经网络作为主干. 本文以 ResNet50 为主干网络，由 4 个残差块组成: 前 3 个残差块网络参数不共享，用来提取模态特有特征; 第 4 个残差块参数共享，共享层将特定于模态的信息嵌入公共空间中，学习多模态共享空间，以弥补 2 种异构模态之间的差距. 对模态嵌入特征分块划分，一共分为 P 块，以充分学习局部特征表示.

2.2 双向动态交互模块

预训练的 ResNet50 提取特征后，卷积神经网络将图像编码为特征图. 特征图可粗略捕获输入图像的语义特征，根据全局内容对特征图线性加权，增强特征的语义表达能力. 然而，目前主流的卷积操作采用卷积核参数共享方式. 由于不同图像的特征不同，在图像间共享卷积核不能高效地提取其特有特征，因此，本文通过设计定制化卷积核来专注不同图像的特性. 与卷积核参数共享方法不同，为了挖掘更多的视觉特征，本文侧重于使用空间维度上的多个卷积核，利用语义信息的多样性，根据输入图像动态生成每个区域对应的卷积核，提取其特定的细节信息; 同时，将一个模态特定的细节信息作为另一个模态的互补信息，其中，双向设置可实现可见光模态和红外模态之间特定细节信息的双向传递以缓解模态差异. 双向动态交互模块结构示意图如图 2 所示，分为 4 个步骤: (1) 卷积核预测; (2) 卷积核生成; (3) 动态卷积操作; (4) 双向交互. 网络训练阶段可简述为将可见光和红外模态的特征图输入共享的卷积核预测网络，预测一组自适应权值 $\boldsymbol{\rho}=\{\rho_1,\rho_2,\cdots,\rho_k\}$，再与基础卷积核聚合，得到特定于输入的卷积滤波器，下面详细介绍这 4 个步骤.

(1) 卷积核预测. 卷积核预测网络根据原始的特征图预测一组权重系数. 与文献[43]不同，本文不利用生成网络直接生成卷积核. 首先定义传统的或静态的感知机的表达式为 $y=G\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}\right)$; 其中，$\boldsymbol{W}$ 表示权重向量，$\boldsymbol{b}$ 表示偏置向量. 动态卷积的动态性通过学习将输入映射到卷积核的函数实现. 如图 2 所示，网络分别以可见光模态和红外模态的特征图 $\boldsymbol{X}^m\in\mathbb{R}^{h\times w\times c}$ 作为输入，其中，$m\in\{V,I\}$，V 和 I 分别表示可见光模态和红外模

态; h, w, c 分别表示特征图的高度、宽度和通道数. 2 个模态的特征图分别经过共享的卷积核预测网络为各自生成一组预测权重 $\{\rho_k(\boldsymbol{X}_m)\}$, 权重不是固定的, 依赖于输入 $\boldsymbol{X}_m$. 通过聚合 k 个线性函数 $\{\tilde{\boldsymbol{W}}_k^{\mathrm{T}}\boldsymbol{x}+\tilde{\boldsymbol{b}}_k\}$ 来定义动态感知机, 公式为

$$y = G\left(\tilde{\boldsymbol{W}}^{\mathrm{T}}(\boldsymbol{x})\boldsymbol{x}+\tilde{\boldsymbol{b}}(\boldsymbol{x})\right) \tag{1}$$

其中, $\tilde{\boldsymbol{W}}(\boldsymbol{x})=\sum_{k=1}^{K}\rho_k(\boldsymbol{x})\tilde{\boldsymbol{W}}_k$, $\tilde{\boldsymbol{b}}(\boldsymbol{x})=\sum_{k=1}^{K}\rho_k(\boldsymbol{x})\tilde{\boldsymbol{b}}_k$, ρ_k 表示第 k 个线性函数的权值. 卷积核预测网络由 1 个全局平均池化、2 个全连接层和 1 个 softmax 激活函数构成. 因为卷积核预测网络依赖于输入, 因此卷积核预测网络对 2 个模态共享.

(2) 卷积核生成. 在卷积核预测网络中学习一组预测权值后对基础卷积核进行聚合, 得到应用于特征图的卷积核, 这里卷积核也不固定, 它依赖于输入的特征图. 基础卷积核由 k 个具有不同大小的传统卷积核组成, 包括 1×1, 3×3, 5×5 的卷积核. 各个卷积核未进行线性无关约束, 但通过不同尺寸的卷积核可利用不同的感知域实现多样性. 利用卷积核预测和卷积核生成能够增强网络捕获不同图像特性的能力, 根据输入的特征动态生成卷积核, 每个卷积核的关注可根据输入的特性自动调整, 以捕获不同模态不同图像特定的信息.

(3) 动态卷积操作. 经过卷积核预测及卷积核生成模块后, 网络学到基于输入图像的卷积核, 对特征图线性加权, 以实现提取特定信息的目的.

(4) 双向交互. 经过上述 3 个步骤后已提取特定于不同模态不同图像的信息, 将其在模态间双向传递, 从而达到互补的目的.

2.3　跨层多分辨率特征融合模块

双流的主干网络 ResNet50 包含 4 个不同残差块, 每个残差块特征分辨率减半. 该子网先利用多分辨率注意力模块学习各残差模块不同分辨率特征的注意力图, 再利用一组可学习权重对特征线性加权, 网络结构如图 3 所示.

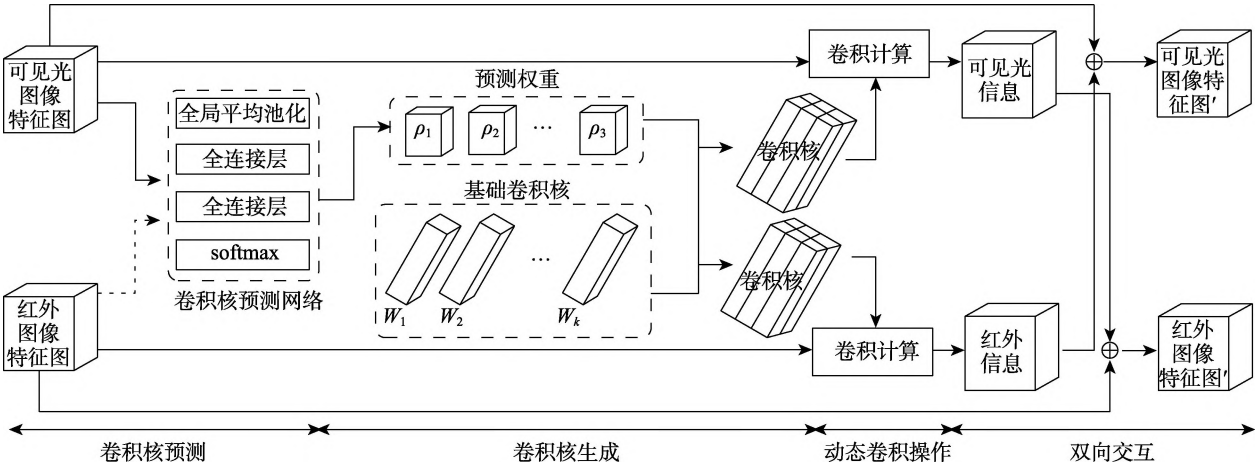

图 2　双向动态交互模块结构示意图

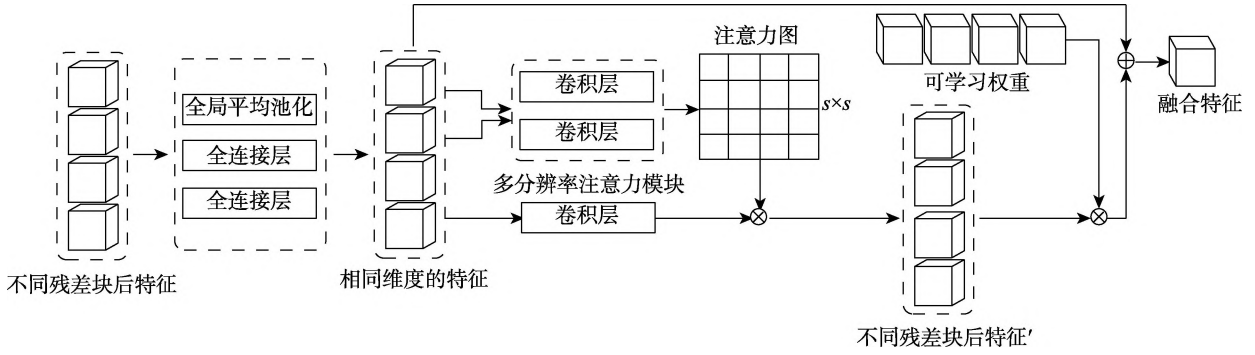

图 3　跨层多分辨率特征融合模块结构示图

(1) 多分辨率注意力模块. 跨层多分辨率特征融合模块的输入是每个残差块的特征, 计算不同残差块后的不同分辨率特征之间的相关性可得到注意力图, 跨层联合学习, 增强特征表示能力. 双流主干网络提取每个残差块的特征图, 对其使用全局平均池化和 2 个全连接层得到不同残差块的

不同分辨率特征，全连接层操作将所有的特征嵌入相同维度. 使用 $\boldsymbol{X}^s=\{\boldsymbol{x}_i^s\in{}^{C\times1}\}_{i=1}^s$ 表示每个残差块的特征，其中，C 表示通道数，s 表示主干网络的不同残差块个数. 多分辨率注意模块的输入为不同残差块的特征，该模块由 2 个 1×1 的卷积组成，通过计算不同残差块特征之间的相似性得到相似性矩阵. 不同残差块特征之间的注意力表示为

$$\alpha_{i,j}^s=\frac{f\left(\boldsymbol{x}_i^s,\boldsymbol{x}_j^s\right)}{\sum\limits_{\forall j}f\left(\boldsymbol{x}_i^s,\boldsymbol{x}_j^s\right)} \tag{2}$$

其中，$\alpha_{i,j}^s\in[0,1]^{s\times s}$；$f\left(\boldsymbol{x}_i^s,\boldsymbol{x}_j^s\right)$ 表示 2 个不同分辨率特征之间的相似度，$\boldsymbol{x}_i^s$ 表示第 i 个残差模块后的特征，$\boldsymbol{x}_j^s$ 表示第 j 个残差模块后的特征，s 表示残差块个数. 为了增强识别能力，使用指数函数放大关系，从而增大不同分辨率特征之间的差异，公式为

$$f\left(\boldsymbol{x}_i^s,\boldsymbol{x}_j^s\right)=\exp\left(u\left(\boldsymbol{x}_i^s\right)^{\mathrm{T}}v\left(\boldsymbol{x}_j^s\right)\right) \tag{3}$$

其中，$u\left(\boldsymbol{x}_i^s\right)=\boldsymbol{W}_u\boldsymbol{x}_i^s$ 和 $v\left(\boldsymbol{x}_j^s\right)=\boldsymbol{W}_v\boldsymbol{x}_j^s$ 分别表示经过多分辨率注意力模块 2 个 1×1 的卷积操作后的结果，$\boldsymbol{W}_u$ 和 $\boldsymbol{W}_v$ 分别表示卷积核的参数矩阵. 注意力图的尺寸为 $s\times s$，比像素级注意力图的尺寸 $hw\times hw$ 小得多；同时，该模块对人物图像不同分辨率特征的噪声区域和局部杂波具有较好的鲁棒性. 如图 3 所示，将输入特征送入另一个 1×1 的卷积层中提取嵌入特征，与多分辨率注意模块的输出进行内积操作得到注意力增强的特征，公式为

$$\bar{\boldsymbol{x}}_i^s=\boldsymbol{a}_i^s*z\left(\boldsymbol{x}_i^s\right) \tag{4}$$

其中，$\boldsymbol{a}_i^s\in A^s=\left\{\alpha_{i,j}^s\right\}^{s\times s}$ 表示不同分辨率的注意力图. 因此，细化的多分辨率特征考虑不同分辨率特征之间的关系. 最后，利用一组可学习的权重 $\boldsymbol{w}=\{w_1,w_2,\cdots,w_n\}$ 对不同分辨率特征线性加权，得到最终的融合特征（n 表示不同分辨率特征个数，本文中取 n=4），公式为

$$F_{\text{fusion}}=\sum_{i=1}^n w_i\bar{\boldsymbol{x}}_i^s \tag{5}$$

2.4 损失函数

本文利用行人图像的真值标签监督学习特征表示. 受文献[44]启发，本文采用模态对齐损失 L_{ma} 和身份损失 L_{id} 联合监督网络训练，传统的身份损失只能扩大类间差异，而类内相似度对跨模态行人重识别具有十分重要的作用. 此外，2 种模态的数据分布存在较大差异，直接约束模态分布的距离比较困难，但数据分布的中心一定程度上反映了特征分布信息. 为了解决这个问题，本文通过拉近类内模态分布的中心距离，达到提高类内跨模态特征相似度的目的. 通过 L_{ma} 和 L_{id} 达到扩大类间差异和拉近类内模态相似性 2 个重要目标.

L_{ma} 计算公式为

$$L_{\text{ma}}=\sum_{i=1}^U\left[\left\|\boldsymbol{c}_{i,1}-\boldsymbol{c}_{i,2}\right\|_2^2\right] \tag{6}$$

其中，U 表示类的数量；$\boldsymbol{c}_{i,1},\boldsymbol{c}_{i,2}$ 分别表示第 i 类可见光模态和红外模态特征的平均嵌入中心，$\boldsymbol{c}_{i,1}=\frac{1}{M}\sum\limits_{j=1}^M\boldsymbol{x}_{i,1,j}$，$\boldsymbol{c}_{i,2}=\frac{1}{N}\sum\limits_{j=1}^N\boldsymbol{x}_{i,2,j}$，$M$ 和 N 分别表示第 i 类可见光图像和红外图像的数量，$\boldsymbol{x}_{i,1,j}$ 和 $\boldsymbol{x}_{i,2,j}$ 分别表示第 i 类中第 j 个可见光图像和红外图像特征. L_{id} 公式为

$$L_{\text{id}}=-\sum_{i=1}^B\log\frac{\mathrm{e}^{\boldsymbol{w}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_{y_i}}}{\sum\limits_{j=1}^n\mathrm{e}^{\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}_j}} \tag{7}$$

其中，B 表示批处理数，$\boldsymbol{x}_i$ 表示第 i 个样本所提取属于 y_i 类的特征，$\boldsymbol{W}_j$ 表示第 j 列的权重，$\boldsymbol{b}$ 是偏差项.

在跨层多分辨率特征融合模块中，只有在每个特征向量对重识别具有足够的鉴别能力时，考虑多分辨特征之间的关系及融合多分辨率特征才有意义. 因此，对融合特征使用 L_{id} 进行监督. 最终总损失函数为

$$L=L_{\text{id}}+\lambda L_{\text{ma}} \tag{8}$$

3 实验结果及分析

为了验证本文方法及其各组成部分的有效性，在跨模态行人重识别的 2 个公开数据集 SYSU-MM01 和 RegDB[45]上与当前主流的跨模态行人重识别的方法进行对比.

3.1 实验设置

3.1.1 数据集

SYSU-MM01 是跨模态行人重识别任务的大规模公开数据集，其中的图像由 4 个可见光摄像机和 2 个红外摄像机分别在室内和室外环境中采集. 训练集 395 个行人，包含 22 258 幅可见光图像，11 909 幅红外图像；测试集 96 个行人，包含 3 803 幅红外图像作为查询图像，随机选取 301/3 010 幅

(一幅/多幅)可见光图像作为匹配集. 采用室内搜索和全搜索 2 种评估模式.

RegDB 由双摄像头系统采集, 共 412 个行人, 8 240 幅图像. 其中, 206 人用于训练, 206 人用于测试; 每个行人包括 10 幅不同的红外图像和 10 幅不同的可见光图像. 采用可见光图像搜索热红外图像和热红外图像搜索可见光图像 2 种评估模式. 为了获得统计上稳定的结果, 实验结果均为 10 次实验的平均结果.

3.1.2　模型的训练与测试

本文以端到端的方式训练整个网络, 训练时, 将可见光图像和红外图像对作为输入, 利用 ResNet50 分别提取可见光和红外图像来自不同残差块的不同分辨率特征, 双向动态交互模块对网络第 3 个残差块的特征图进行动态卷积操作, 并在模态间双向传递提取的特征. 跨层特征融合模块融合来自不同残差块的不同分辨率特征. 测试时, 利用单一可见光(或红外)图像作为查询图像搜索对应的红外(或可见光)图像.

3.1.3　评价指标

所有实验均遵循现有可见光-红外跨模态行人重识别方法的标准评估方案. 查询图像和匹配集图像来自不同的模态, 采用累积匹配曲线(cumulative match characteristic, CMC)和平均检索精度(mean average precision, mAP)作为评价指标. CMC 曲线是一种 Rank-k 的击中概率, mAP 则评价算法的平均检索性能.

3.1.4　实验细节

训练过程在跨模态行人重识别的 2 个公开数据集 SYSU-MM01 和 RegDB 上进行. 本文使用预训练的 ResNet50 为主体框架, 将行人图像调整为 288×144 像素, 使用随机裁剪和随机水平翻转作为数据增强. 训练过程中, 批大小设置为 32, 设超参数 $\lambda = 0.5$. 优化器使用随机梯度下降(stochastic gradient descent, SGD)算法, 动量 $M = 0.9$. 主干的输出特征等分为 $P = 6$ 个块, 并将特征维数降至 512. 预训练网络的初始学习率为 0.001, 分类器的初始学习率为 0.01, 数据集经过训练迭代 30 轮后, 所有网络的学习率降为原来的 1/10. 使用 Pytorch 1.1.0 框架实现代码, 共训练 60 轮, 在训练过程中本文使用 GTX1080Ti 显卡加速.

3.2　与主流方法对比结果

SYSU-MM01 数据集包括 2 种模态数据: 可见光模态和近红外模态. 在此数据集上, 本文使用行人的红外图像搜索可见光图像.

RegDB 数据集存在 2 种搜索模式: 可见光图像搜索红外图像和红外图像搜索可见光图像.

3.2.1　在 SYSU-MM01 数据集上的实验结果

在 SYSU-MM01 数据集上的对比结果如表 1 所示, 其中, R1, R10, R20 分别表示 Rank-1, Rank-10 和 Rank-20 准确率. 可以看出: (1) 由于双流网络可同时学习模态特定和模态共享特征, 双流网络的动态双注意力聚合学习[46](dynamic dual-attentive aggregation learning, DDAG)和本文方法具有较大优势; (2) 由于本文方法不仅考虑模态异质性对跨模态重识别的影响, 同时引入跨层多分辨特征融合模块进一步解决类内变化, 因此无论是 R1 还是 mAP 均具有显著提升; (3) 与同时学习模态共享信息和特定信息的方法(cross-modality shared-specific feature transfer, cm-SSFT)相比, 在各种设置下本文方法均具有明显优势. 虽然在全搜索且每个 ID 取多幅图像(all-search, multi-shot)设置下, 本文方法的 mAP 较 cm-SSFT 方法略显劣势, 但 R1 提升了 9.2%, 其主要原因是本文未挖掘相同 ID 不同图像间的联系; 此外, 本文方法训练过程简单, 无需进行对抗训练和构图. 为了公平起见, 本文对比 cm-SSFT 方法测试过程中未使用辅助数据集下的实验结果.

3.2.2　在 RegDB 数据集上的实验结果

在 RegDB 数据集上的对比结果如表 2 所示. 可以看出: (1) 无论是可见光图像搜索红外图像还是红外图像搜索可见光图像, 本文方法相较主流方法均有显著提升; (2) 与 SYSU-MM01 数据集相比, 由于 RegDB 数据集是由双摄像机系统相机采集, 数据变化小, 重识别挑战小, 本文方法在此数据集上提升较高.

3.3　消融实验

为了评估双向动态交互模块和跨层多分辨率特征融合模块 2 个组件的有效性, 在 SYSU-MM01 数据集全搜索模式下, 通过定量分析来评估各个组件的有效性, 结果如表 3 所示.

双向动态交互模块是由卷积核生成和双向交互 2 个子部件组成的. 下面验证这 2 个子部件的必要性. (1) 直接将 2 个模态的特征双向交互, 通过对比表 3 中情况 1 与情况 2 可以看出, 直接将 2 个模态的特征进行双向交互会大大降低模型精度, 这是由于模态异质性会引入大量噪声; 但通过对比情况 4 与情况 5, 可以验证卷积核生成后, 再进行双向交互来缓解模态差异的必要性. (2) 通过对比表 3 中的情况 2 与情况 5, 可以验证卷积核生成

表 1 在 SYSU-MM01 数据集上与主流方法的对比结果 %

方法	全搜索								室内搜索							
	单镜头				多镜头				单镜头				多镜头			
	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG[47]	2.8	18.3	31.9	4.2	3.8	22.8	37.6	2.2	3.2	24.7	44.5	7.3	4.8	29.2	49.4	3.5
LOMO[48]	3.6	23.2	37.3	4.5	4.7	28.2	43.1	2.3	5.8	34.4	54.9	10.2	7.4	40.4	60.3	5.6
Zero-Padding[29]	14.8	54.1	71.3	15.9	19.1	61.4	78.4	10.9	20.6	68.4	85.8	26.9	24.4	75.9	91.3	18.6
HCML[30]	14.3	53.2	69.2	16.2												
BDTR[31]	17.0	55.4	72.0	19.7												
D-HSME[34]	20.7	62.8	78.0	23.2												
IPVT+MSR[49]	23.2	51.2	61.7	22.5												
cmGAN[35]	27.0	67.5	80.6	27.8	31.5	72.7	85.0	22.3	31.6	77.2	89.2	42.2	37.0	80.9	92.1	32.8
D^{2}RL[40]	28.9	70.6	82.4	29.2												
DGD+MSR[50]	37.4	83.4	93.3	38.1	43.9	86.9	95.7	30.5	39.6	89.3	97.7	50.9	46.6	93.6	98.8	40.1
JSIA-ReID[51]	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
AlignGAN[27]	42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	35.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
XIV-ReID[37]	49.9	89.8	96.0	50.7												
DDAG[46]	54.8	90.4	95.8	53.0					61.2	94.1	98.4	68.0				
TSLFN+HC[44]	57.0	91.5	96.8	55.0	62.1	93.7	97.9	48.0	59.7	92.1	96.2	64.9	69.8	95.9	98.9	57.8
cm-SSFT[41]	47.7			54.1	57.4			**59.1**								
本文	**61.7**	**94.0**	**98.0**	**59.3**	**66.6**	**96.3**	**98.8**	52.5	**64.5**	**96.1**	**98.9**	**70.8**	**75.8**	**98.8**	**99.7**	**64.3**

注. 加粗和加下划线数值分别表示排名最优和次优结果.

表 2 在 RegDB 数据集上与主流方法的对比结果 %

方法	可见光搜索红外图像				红外搜索可见光图像			
	R1	R10	R20	mAP	R1	R10	R20	mAP
HOG[47]	13.5	33.2	43.7	10.3				
MLBP[52]	2.0	7.3	10.9	6.8				
LOMO[48]	0.9	2.5	4.1	2.3				
GSM[53]	17.3	34.5	45.3	15.1				
One-stream[29]	13.1	33.0	42.5	14.0				
Two-stream[29]	12.4	30.4	41.0	13.4				
TONE[30]	16.9	34.0	44.1	14.9	13.9	30.1	40.1	17.0
Zero-Padding[29]	17.8	34.2	44.4	18.9	16.6	34.7	44.3	17.8
BDTR[31]	33.5	58.4	67.5	31.8	32.7	58.0	68.9	31.1
D-HSME[34]	50.9	73.4	81.7	47.0	50.2	72.4	81.1	46.2
D^{2}RL[40]	43.4	66.1	76.3	44.1				
AlignGAN[27]	57.9			53.6	56.3			53.4
XIV-ReID[37]	62.2	83.1	91.7	60.2				
DDAG[46]	69.3	86.2	91.5	63.5	68.1	85.2	90.3	61.8
TSLFN+HC[44]	83.0	96.1	98.0	72.0	80.9	95.4	96.9	71.4
cm-SSFT[41]	65.4			65.6				
本文	**85.1**	**96.2**	**98.3**	**74.7**	**83.3**	**95.9**	**98.0**	**72.9**

注. 加粗和加下划线数值分别表示排名最优和次优结果.

表 3　SYSU-MM01 全搜索模式下的消融实验　%

情况	卷积核生成	双向交互	跨层多分辨率特征融合模块	R1	mAP
1	×	×	×	56.3	55.2
2	×	√	×	23.7	25.2
3	×	√	√	32.5	33.6
4	√	×	×	58.9	56.9
5	√	√	×	59.8	58.0
6	×	×	√	60.4	59.0
7	√	×	√	60.5	58.8
8	√	√	√	61.7	59.3

部件的必要性. (3) 通过对比表 3 中的情况 1 与情况 4, 可进一步验证卷积核生成部件的有效性. (4) 通过对比表 3 中的情况 5 与情况 1, 可以验证双向动态交互整个模块的有效性. 为了验证跨层多分辨率特征融合模块的有效性, 首先联合执行双向交互和跨层多分辨率特征融合模块, 实验结果如表 3 中情况 3 所示, 通过与表 3 中情况 2 对比, 可以验证跨层多分辨率特征融合模块的有效性; 其次, 对比表 3 中的情况 5 与情况 8, 进一步验证跨层多分辨率特征融合模块的有效性.

3.4　其他实验结果与分析

3.4.1　不同网络层使用双向动态交互模块有效性

为了进一步验证双向动态交互模块的有效性, 在主干网络 ResNet50 的 4 个不同层(残差块)分别引入双向动态交互模块进行测试. 表 4 所示为在 SYSU-MM01 数据集全搜索模式下的实验结果. 实验结果表明, 图像的底层特征(网络浅层部分提取的特征)与特征在图像中的位置无关, 而高层特征一般与位置相关[54]. 由表 4 可以看出, (1) 双向动态交互模块在 ResNet50 前 2 层没有明显优势, 因此本文针对网络前 2 层采用权值共享的方式, 可进一步减少参数量; (2) 由于网络中高层特征信息更加抽象, 且一般与位置相关, 双向动态交互模块在网络的第 3 层后提升效果最优. 因此, 本文将其位置固定在网络的第 3 层后.

3.4.2　特征融合方式影响

为了进一步验证跨层多分辨率特征融合的有效性, 分别与直接融合各层不同分辨率特征和利用各层不同分辨率特征的多分类器方式进行对比, 在 SYSU-MM01 数据集全搜索模式下的实验结果如表 5 所示. 可以看出, 相较于直接融合的方式, 跨层多分辨率特征融合在 R1 和 mAP 这 2 项指标中分别提高 2.5%和 1.9%, 原因是跨模态行人重识别任务中的模态异质性会引入噪声, 从而影响模型性能. 虽然多分类器的方式优越于直接融合, 但训练多个分类器将成倍地增加训练成本; 由于浅层特征语义信息不足, 强行在不同层建立身份的映射关系会增加噪声, 且 R1 和 mAP 这 2 项指标均逊色于本文跨层多分辨率特征融合模块 1.7%. 综上所述, 本文的跨层多分辨率特征融合模块能够提取更加丰富的行人特征, 且冗余信息少.

表 4　在不同层后使用双向动态交互模块结果　%

层	R1	mAP
骨干网络	56.3	55.2
第 1 层	56.5	55.6
第 2 层	56.8	55.4
第 3 层	59.9	58.0
第 4 层	59.2	57.5

表 5　各层不同分辨率特征融合方式结果　%

方法	R1	mAP
骨干网络	56.3	55.2
直接融合[5-6]	57.9	57.1
多分类器[7]	58.7	57.3
跨层多分辨率特征融合模块	60.4	59.0

3.4.3　λ 的影响

为了研究模态对齐损失在总体损失函数中权重的影响, 在 SYSU-MM01 数据集全搜索模式下对式(8)中的 λ 进行分析, 结果如表 6 所示. 可以看出, 本文方法受 λ 影响相对稳定, 这是由于提出的双向动态交互模块通过模态间的局部信息双向传递实现互补, 弥补了骨干网络中共享信息不足的缺点; 而骨干网络在 $\lambda>0.5$ 后模型准确率大幅下降, 原因是模态对齐损失主要依赖于共享信息, 由于骨干网络局部特征共享信息有限, 过度依赖会限制模型性能造成过拟合.

表 6　SYSU-MM01 数据集 λ 对模型的影响　%

方法	λ	R1	mAP
骨干网络	0.1	52.6	51.0
	0.3	56.0	55.0
	0.5	56.3	55.2
	0.7	27.7	30.6
	0.9	1.1	3.0
本文	0.1	57.8	55.7
	0.3	60.2	57.9
	0.5	61.7	59.3
	0.7	58.4	57.4
	0.9	51.8	52.2

4 结 语

本文提出一种面向跨模态行人重识别的双向动态交互网络方法, 包括双流主干网络、双向动态交互模块和跨层多分辨率特征融合模块 3 个组件. 其中, 双向动态交互模块利用特征图的全局内容学习不同空间位置的权值信息, 生成定制化的卷积核, 从而捕获不同模态图像的特有信息, 并将此信息作为互补信息在 2 个模态间进行双向传递以缓解模态差异; 跨层多分辨率特征融合模块对卷积网络层的有效嵌入进行建模, 以应对更加复杂的挑战. 实验结果表明, 本文方法性能显著. 相对于单模态行人重识别, 跨模态行人重识别面临着更大的挑战, 如何缓解模态差异以及建模更丰富的视觉特征是跨模态匹配的关键问题, 也是未来工作的重点.

参考文献(References):

[1] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[OL]. [2021-08-03]. https://arxiv.org/pdf/1703.07737.pdf

[2] Suh Y, Wang J D, Tang S Y, *et al*. Part-aligned bilinear representations for person re-identification[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 418-437

[3] Köstinger M, Hirzer M, Wohlhart P, *et al*. Large scale metric learning from equivalence constraints[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2012: 2288-2295

[4] Hariharan B, Arbeláez P, Girshick R, *et al*. Hypercolumns for object segmentation and fine-grained localization[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 447-456

[5] Wang Y, Wang L Q, You Y R, *et al*. Resource aware person re-identification across multiple resolutions[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8042-8051

[6] Qi L, Huo J, Wang L, *et al*. A mask based deep ranking neural network for person retrieval[C] //Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). Los Alamitos: IEEE Computer Society Press, 2019: 496-501

[7] Lan X, Zhu X T, Gong S G. Person search by multi-scale matching[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 553-569

[8] Zheng L, Yang Y, Hauptmann A G. Person re-identification: past, present and future[OL]. [2021-08-03]. https://arxiv.org/pdf/1610.02984.pdf

[9] Fang P F, Zhou J M, Roy S, *et al*. Bilinear attention networks for person retrieval[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 8029-8038

[10] Sun Y F, Zheng L, Yang Y, *et al*. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 501-518

[11] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2016: 791-808

[12] Li Hao, Tang Min, Lin Jianwu, *et al*. Cross-modality person re-identification framework based on improved hard triplet loss[J]. Computer Science, 2020, 47(10): 180-186(in Chinese)
(李灏, 唐敏, 林建武, 等. 基于改进困难三元组损失的跨模态行人重识别框架[J]. 计算机科学, 2020, 47(10): 180-186)

[13] Eom C, Ham B. Learning disentangled representation for robust person re-identification[C] //Proceedings of the Advances in Neural Information Processing Systems. New York: ACM Press, 2019: 5297-5308

[14] Zheng Z D, Yang X D, Yu Z D, *et al*. Joint discriminative and generative learning for person re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 2133-2142

[15] Yang Wanxiang, Yan Yan, Chen Si, *et al*. Multi-scale generative adversarial network for person re-identification under occlusion[J]. Journal of Software, 2020, 31(7): 1943-1958(in Chinese)
(杨婉香, 严严, 陈思, 等. 基于多尺度生成对抗网络的遮挡行人重识别方法[J]. 软件学报, 2020, 31(7): 1943-1958)

[16] Qiu Yaoru, Sun Weijun, Huang Yonghui, *et al*. Person re-identification method based on GAN uniting with spatial-temporal pattern[J]. Journal of Computer Applications, 2020, 40(9): 2493-2498(in Chinese)
(邱耀儒, 孙为军, 黄永慧, 等. 基于生成对抗网络联合时空模型的行人重识别方法[J]. 计算机应用, 2020, 40(9): 2493-2498)

[17] Mnih V, Heess N, Graves A, *et al*. Recurrent models of visual attention[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM Press, 2014: 2204-2212

[18] Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010

[19] Bai S, Bai X, Tian Q. Scalable person re-identification on supervised smoothed manifold[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 3356-3365

[20] Luo C C, Chen Y T, Wang N Y, *et al*. Spectral feature transformation for person re-identification[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 4975-4984

[21] Shen Y T, Li H S, Yi S, *et al*. Person re-identification with deep similarity-guided graph neural network[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 486-504

[22] Shen Y T, Li H S, Xiao T, *et al*. Deep group-shuffling random walk for person re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2265-2274

[23] Lv J M, Chen W H, Li Q, *et al*. Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7948-7956

[24] Wu J L, Liu H, Yang Y, *et al*. Unsupervised graph association for person re-identification[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 8320-8329

[25] Wei L H, Zhang S L, Gao W, *et al*. Person transfer GAN to bridge domain gap for person re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 79-88

[26] Deng W J, Zheng L, Ye Q X, *et al*. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 994-1003

[27] Wang G A, Zhang T Z, Cheng J, *et al*. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019:3622-3631

[28] Chen Dan, Li Yongzhong, Yu Peize, *et al*. Research and prospect of cross modality person re-identification[J]. Computer Systems & Applications, 2020, 29(10): 20-28(in Chinese)
(陈丹, 李永忠, 于沛泽, 等. 跨模态行人重识别研究与展望[J]. 计算机系统应用, 2020, 29(10): 20-28)

[29] Wu A C, Zheng W S, Yu H X, *et al*. RGB-infrared cross-modality person re-identification[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 5390-5399

[30] Ye M, Lan X Y, Li J W, *et al*. Hierarchical discriminative learning for visible thermal person re-identification[C] //Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018, 32(1): 7501-7506

[31] Ye M, Wang Z, Lan X Y, *et al*. Visible thermal person re-identification via dual-constrained top-ranking[C] //Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2018: 1092-1099

[32] Ye M, Lan X Y, Leng Q M. Modality-aware collaborative learning for visible thermal person re-identification[C] //Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press, 2019: 347-355

[33] Hao Y, Wang N N, Gao X B, *et al*. Dual-alignment feature embedding for cross-modality person re-identification[C] //Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press, 2019: 57-65

[34] Hao Y, Wang N N, Li J, *et al*. HSME: hypersphere manifold embedding for visible thermal person re-identification[J]. Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33: 8385-8392

[35] Dai P Y, Ji R R, Wang H B, *et al*. Cross-modality person reidentification with generative adversarial training[C] //Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York: ACM Press, 2018: 677-683

[36] Choi S, Lee S M, Kim Y, *et al*. Hi-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10254-10263

[37] Li D G, Wei X, Hong X P, *et al*. Infrared-visible cross-modal person re-identification with an X modality[C] //Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34(4): 4610-4617

[38] Jia M X, Zhai Y P, Lu S J, *et al*. A similarity inference metric for RGB-infrared cross-modality person re-identification[OL]. [2021-08-03]. https://arxiv.org/pdf/2007.01504.pdf

[39] Kniaz V V, Knyaz V A, Hladůvka J, *et al*. ThermalGAN: multimodal color-to-thermal image translation for person re-identification in multispectral dataset[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 606-624

[40] Wang Z X, Wang Z, Zheng Y Q, *et al*. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 618-626

[41] Lu Y, Wu Y, Liu B, *et al*. Cross-modality person re-identification with shared-specific feature transfer[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 13376-13386

[42] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[43] Jia X, De Brabandere B, Tuytelaars T, *et al*. Dynamic filter networks[C] //Proceedings of the 30th Conference on Neural Information Processing Systems. New York: ACM Press, 2016: 667-675

[44] Zhu Y X, Yang Z, Wang L, *et al*. Hetero-center loss for cross-modality person re-identification[J]. Neurocomputing, 2020, 386: 97-109

[45] 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): 605-605

[46] Ye M, Shen J B, Crandall D J, *et al*. Dynamic dual-attentive

aggregation learning for visible-infrared person re-identification[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2020: 229-247

[47] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2005: 886-893

[48] Liao S C, Hu Y, Zhu X Y, *et al*. Person re-identification by local maximal occurrence representation and metric learning[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press , 2015: 2197-2206

[49] Kang J K, Hoang T M, Park K R. Person re-identification between visible and thermal camera images based on deep residual CNN using single input[J]. IEEE Access, 2019, 7: 57972- 57984

[50] Feng Z X, Lai J H, Xie X H. Learning modality-specific representations for visible-infrared person re-identification[J]. IEEE Transactions on Image Processing, 2019, 29: 579-590

[51] 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

[52] Liao S C, Li S Z. Efficient PSD constrained asymmetric metric learning for person re-identification[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 3685-3693

[53] Lin L, Wang G R, Zuo W M, *et al*. Cross-domain visual matching via generalized similarity measure and feature learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1089-1102

[54] Sabour S, Frosst N, Hinton G E, *et al*. Dynamic routing between capsules[C] //Proceedings of the Advances in Neural Information Processing Systems. New York: ACM Press, 2017: 3856-3866