ISSN 1003-501X CN 51-1346/O4

光电工程

Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

多特征聚合的红外–可见光行人重识别

郑海君，葛斌，夏晨星，邬成

引用本文:

郑海君，葛斌，夏晨星，等. 多特征聚合的红外–可见光行人重识别[J]. 光电工程，2023， **50**(7): 230136.

Zheng H J, Ge B, Xia C X, et al. Infrared-visible person re-identification based on multi feature aggregation[J]. *Opto-Electron Eng*, 2023, **50**(7): 230136.

https://doi.org/10.12086/oee.2023.230136

收稿日期: 2023-06-15; 修改日期: 2023-08-10; 录用日期: 2023-08-11

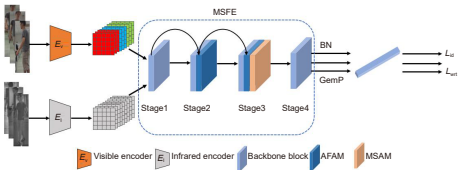
相关论文

Opto-Electronic Engineering
光 电 工 程

Article
2023 年, 第 50 卷, 第 7 期

DOI: 10.12086/oee.2023.230136

多特征聚合的红外-可见光行人重识别



郑海君[1]，葛　斌[1*]，夏晨星[1,2]，邬　成[1]

[1]安徽理工大学计算机科学与工程学院，安徽 淮南 232001；
[2]合肥综合性国家科学中心 能源研究院，安徽 合肥 230031

摘要：红外-可见光行人重识别在视频监控、智能交通、安防等领域具有广泛应用。但是不同图像模态间的差异，给该领域带来了巨大的挑战。现有方法主要集中于缓解模态间差异以获得更具鉴别性的特征，但却忽略了邻级特征之间的关系以及多尺度信息对全局特征的影响。因此，本文提出一种基于多特征聚合的红外-可见光行人重识别方法(MFANet)解决现有方法的缺陷。首先在特征提取阶段融合邻级特征，引导低级特征信息的融入，以强化高级特征，使得特征更具健壮性；然后聚合不同感受野的多尺度特征以获得丰富的上下文信息；最后，以多尺度特征作为引导，强化特征以获得更具鉴别性的特征。在 SYSU-MM01 和 RegDB 数据集上的实验结果证明了所提方法的有效性，其中 SYSU-MM01 数据集在最困难的全搜索单镜头模式下平均精度达到了 71.77%。

关键词：行人重识别；红外；多尺度；邻级特征

中图分类号：TP391　　**文献标志码：**A

郑海君，葛斌，夏晨星，等. 多特征聚合的红外-可见光行人重识别 [J]. 光电工程，2023，**50**(7): 230136
Zheng H J, Ge B, Xia C X, et al. Infrared-visible person re-identification based on multi feature aggregation[J]. *Opto-Electron Eng*, 2023, **50**(7): 230136

Infrared-visible person re-identification based on multi feature aggregation

Zheng Haijun[1], Ge Bin[1*], Xia Chenxing[1,2], Wu Cheng[1]

[1]College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, Anhui 232001, China;
[2]Institute of Energy, Hefei Comprehensive National Science Center, Hefei, Anhui 230031, China

Abstract: Infrared-visible person re-identification has been widely used in video surveillance, intelligent transportation, security, and other fields. However, due to the differences between different image modalities, it brings great challenges to this field. The existing methods mainly focus on mitigating the differences between modes to obtain more discriminating features, but ignore the relationship between adjacent features and the influence of multi-scale information on global features. Here, a infrared-visible person re-identification method (MFANet) based on multi-feature aggregation is proposed to solve the shortcomings of existing methods. Firstly, the adjacent level features are fused in the feature extraction stage, and the integration of low-level feature information is guided to strengthen the high-level features and make the features more robust. Then, the multi-scale

收稿日期：2023-06-15；修回日期：2023-08-10；录用日期：2023-08-11
基金项目：国家自然科学基金 (6210071479，62102003)；国家重大专项 (2020YFB1314103)；安徽省自然科学基金 (2108085QF258)；安徽省博士后基金 (2022B623)
*通信作者：葛斌，bge@aust.edu.cn。

features of different receptive fields of view are aggregated to obtain rich contextual information. Finally, multi-scale features are used as a guide to strengthen the features to obtain more discriminating features. Experimental results on SYSU-MM01 and RegDB datasets show the effectiveness of the proposed method, and the average accuracy of SYSU-MM01 dataset reaches 71.77% in the most difficult all-search single-shot mode.

Keywords: person re-identification; infrared; multi-scale; adjacent level features

1 引　言

智慧城市的不断发展，监控系统的完善，使得摄像头数量呈指数增长，行人重识别逐渐成为热门的研究领域。为满足公共安全需求，越来越多的红外 (IR) 摄像头集成到监控系统中，旨在增强全时段识别人员的能力。行人重识别指的是在不同的摄像头所拍摄的图像中匹配同一行人。传统行人重识别技术关注于单模态图像的检索，在红外-可见光图像集中，其性能会大大折扣，红外-可见光行人重识别技术随之兴起。行人重识别任务由于环境和行人外观的不可控性 (如遮挡、背景和噪声)，这使得精确地实现该任务具有挑战性[1-3]。基于此，红外-可见光行人重识别任务还需考虑红外图像和 RGB 图像间的模态差异，面临更大的类内变化，其更具挑战性。

现有的红外-可见光行人重识别方法主要分为两种：模态差异缓解和模态共享特征学习。生成中间模态或者目标模态是模态差异缓解常用的方法之一。目的是将跨模态检索转换成单模态检索问题，从而提高精度。Wang 等[4]采用变分编码器 (variational autoencoder，VAE) 生成目标模态图像，并利用双级差异减少方法缓解模态间差异。Zhong 等[5]提出的上色网络仅利用颜色信息而不包含整体图像，减少生成图像时信息丢失的同时生成带有颜色信息的 IR 图像。但是，由于采用的是图像生成处理方法，在训练过程中不可避免地会引入噪声，导致检索的准确率不理想。Wu 等[6]为了缓解模态间差异，提出了一种互均值学习方式，从类约束出发缓解模态间差异。Zhang[7] 等提出了一种双重相互学习方法，在跨模态和单一模态间进行相互学习，得到鉴别性特征。模态差异缓解方法专注于削弱模态之间的差异性，弱化了模态共享特征的挖掘与保留，能有效地缩小模态间的距离，但是会导致部分模态共享特征的丢失。因此，为了学习模态共享特征，Ye 等[8]提出了一种双流网络架构，该方法利用全局特征增大不同类之间的差异。随后又将非局部注意力融入双流网络中[9]，进一步增强全局特征。Chen 等[10]提出了一种基于 transformer[11] 的结构感知位置转换网络，通过利用结构和位置信息来学习模态共享特征。Lu 等[12]针对模态间差异的负面影响，以灰度图像作为辅助模态，提出了一种渐进模态共享网络，通过损失函数指导模型从模态共享特征中探索身份信息，并缓解类间差异小和类内差异大的问题，提高特征的鉴别性。这些方法能够强化全局特征并缓解模态间差异，但却忽略了不同级别特征之间信息以及多尺度信息的交互。

针对上述问题，本文提出了一种针对特征交互的多特征聚合的红外-可见光行人重识别网络 (infrared-visible person re-identification network based on multi feature aggregation，MFANet)，该模型实现了不同级别特征的交互聚合，并且利用多尺度信息进一步强化聚合特征，提升网络效率。具体地，本文设计了邻级特征聚合模块 (adjacent feature aggregation module，AFAM) 以实现相邻级别特征之间的交互聚合，提升特征的健壮性。之后，为了获得具有鲁棒性的特征，提出了多尺度聚合模块 (multi scale aggregation module，MSAM)，在挖掘多尺度特征的同时，聚合多尺度信息，并以其作为引导强化特征表示，加强模型对多尺度信息的吸纳能力。在公共数据集上进行了一系列实验，实验表明本文网络有着优越的性能，本文的主要贡献如下：

1) 设计了一种邻级特征聚合模块 (AFAM)，该模块对两个相邻级别的特征进行聚合操作，将低级特征的小感受野信息融入高级特征，进而强化特征健壮性。

2) 设计了一种多尺度聚合模块 (MSAM)，该模块对聚合特征进一步强化，主要通过多尺度信息指导多尺度特征与源特征之间的交互聚合，能使得最终特征更具鲁棒性。在 MSAM 中，设计了一个多尺度特征聚合模块 (multi scale feature aggregation module，MSFAM)，用来提取不同感受野信息的特征并进行聚合。

3) 实验结果表明，本文所提出的网络模型 (MFANet) 在 SYSU-MM01 和 RegDB 两个常用数据集上具有较优的性能。

2 本文方法

2.1 网络框架

设 $V=\{x_v^i\}_{i=1}^{N_v}$，$R=\{x_r^i\}_{i=1}^{N_r}$分别表示跨模态行人重识别数据集中的 RGB 图像集和 IR 图像集，其中N_v和N_r分别表示 RGB 和 IR 的样本图像数量，样本对应的真值标签集 $T=\{t_i\}_{i=1}^{N_z}$，其中 N_z表示人物身份类别的数量。

为了充分利用多级特征和多尺度信息，本文提出了多特征聚合的红外-可见光行人重识别方法(MFANet)，整体网络结构如图 1 所示。在 CAJ[9] 特征提取网络的模态共享特征提取阶段 (modality shared feature extraction，MSFE) 加入了邻级特征聚合模块 (AFAM) 和多尺度聚合模块 (MSAM)，以缓解卷积操作带来的空间信息丢失的情况，并利用多尺度信息强化特征的健壮性和鲁棒性。整体网络由三个阶段构成，第一个阶段为模态特定特征提取阶段，将红外和 RGB 图像x_r^i和x_v^i分别送入到对应的编码器E_I和E_v中，得到模态特定特征 f_r，f_v，此时特征包含更多的图像的颜色和纹理信息，即模态特定信息。第二阶段为模态共享特征提取阶段，将 f_r，f_v送入参数共享的特征提取结构中，获得模态共享特征。最后利用批归一化 (BN) 和 Gem 池化 (GemP) 得到最终的特征表示，并利用身份损失和加权正则化三元组损失约束特征表示。

2.2 邻级特征聚合模块 (AFAM)

低级特征感受野小，注重细节信息，包含更加丰富的空间信息。而高级特征感受野大，注重语义信息。CAJ 网络结构中引入了非局部注意力对当前级别的特征进行自我全局强化，这种方式虽然能够利用特征的全局信息，但是仅仅针对当前级别特征，无法充分利用低级特征的空间通道信息，这会导致有价值信息的丢失。为了解决这个问题，本文在模态特定特征提取结构的第二第三阶段之后加入了邻级特征聚合模块，该模块采用挖掘相邻级特征的空间和语义信息的方式指导特征聚合以强化高级特征，如图 2 所示，通过这种方式可以传递低级特征的空间通道信息，进而强化高级特征，增强重要信息的保留。

首先对低级特征 f_{i-1}进行下采样处理，再对 f_i 和高级特征 f_i 执行全局平均池化 (global average pooling, GAP)，自适应映射操作和 sigmoid 激活函数生成通道权重，并将权重与对应特征相结合强化通道信息。这种方式可以避免在对特征聚合时通道信息的丢失。具体计算方式如下：

$$f_i^s = \sigma(\varphi(GAP(f_i)))\otimes f_i\,, \tag{1}$$

$$f_{i-1}^s = \sigma(\varphi(GAP(D(f_{i-1}))))\otimes D(f_{i-1})\,, \tag{2}$$

式中：$D(\cdot)$表示下采样操作，$\sigma(\cdot)$表示 sigmoid 归一化操作，$GAP(\cdot)$表示全局平均池化操作。

随后，为了进一步聚合两级特征的空间信息，采用卷积操作将f_i^s与f_{i-1}^s的维度统一，并经过像素相加操作后，利用一个卷积层与 softmax 归一化得到结合特征的空间注意力地图 att，用于指导低级特征 f_{i-1}^s空间信息的增强，最后与f_i^s相结合得到最终的聚合特征表示$\tilde{f}_i$，具体计算方式如下：

$$att = Softmax(Conv_h^3(Conv_h^1(f_{i-1}^s)+Conv_h^2(f_i^s)))\,, \tag{3}$$

$$\tilde{f}_i = att\otimes Conv_h^4(D(f_{i-1}))+f_i\,, \tag{4}$$

式中：$Softmax(\cdot)$表示 softmax 归一化操作，$Conv_h^i$表示第 i 个通道数为 h 的卷积操作，在此，h 表示 f_i的通道数。

2.3 多尺度聚合模块 (MSAM)

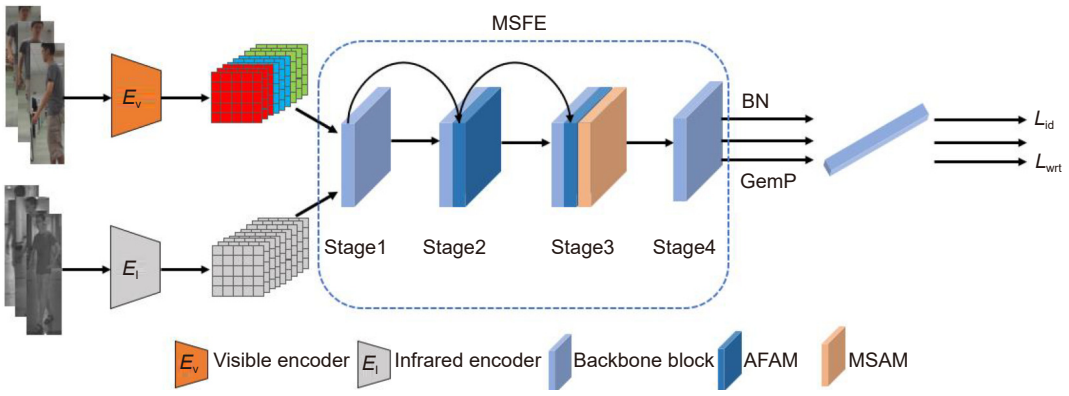


图 1 MFANet 结构图

Fig. 1 MFANet structure diagram

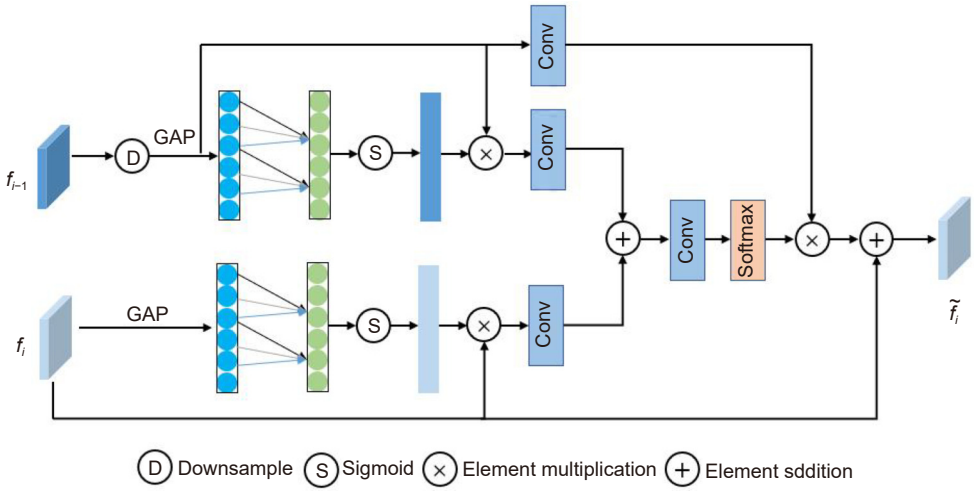


图 2 邻级特征聚合模块

Fig. 2 Adjacency feature aggregation module

由于多尺度特征包含多个层次的信息，能提供更全面和丰富的表示，因此进行多尺度聚合可以充分利用多尺度信息，增强模型的鲁棒性。文中在进行邻级特征聚合时，由于不同级别特征的感受野不同，无法有效地利用多尺度信息。因此本文在特征聚合后针对多尺度信息设计了多尺度聚合模块，如图 3 所示。

首先将特征 f_i输入到多尺度特征聚合模块中，用于将不同尺度下的空间信息的聚合，并利用聚合的空间信息引导特征 f_i的上下文信息强化，如图 4 所示。通过将不同尺度上的特征进行聚合，可以挖掘不同尺度下的人物信息和语义信息，更准确地捕获图像中上下文和细节信息，使得特征更具鉴别性。

多尺度特征聚合模块利用多个不同尺度的空洞卷积得到多尺度特征$F_{\mathrm{dc}}=\left\{f_{\mathrm{idc}}^{j}\right\}_{j=1}^{n}$，$n$ 表示多尺度特征的个数。以卷积核为 3 和 1 的两个卷积层和一个 softmax 激活层组成多尺度权重生成器G_{msw}，每个卷积层后为 ReLu 非线性激活层。将特征 f_i输入G_{msw}中得到不同尺度的特征权重，随后按通道进行分割，得到多尺度特征权重集合$W_{\mathrm{dc}}=\left\{w_{\mathrm{dc}}^{j}\right\}_{j=1}^{n}$，最后分别与对应尺度的特征加权相加得到多尺度特征表示 f_{ms}：

$$f_{\mathrm{ms}}=\sum_{j=1}^{n} f_{idc}^{j}\otimes w_{\mathrm{dc}}^{j}\,. \tag{5}$$

随后，由两个强化分支组成，一个为多尺度引导强化分支，利用多尺度信息的空间信息获取更详细的上下文信息，从而引导特征强化；另一个为特征结合强化分支，特征像素相加之后，通道的权重随之改变，因此重新对通道之间关系进行建模，从而强化结合特征。

多尺度引导强化分支以具有鉴别性的特征 f_{ms}的空间信息指导强化 f_i，f_{ms}经过一个卷积层和 Sigmoid 激活层，得到空间注意力矩阵 ***satt***，利用空间注意力矩阵 ***satt*** 指导特征 f_i的空间信息强化，具体表示

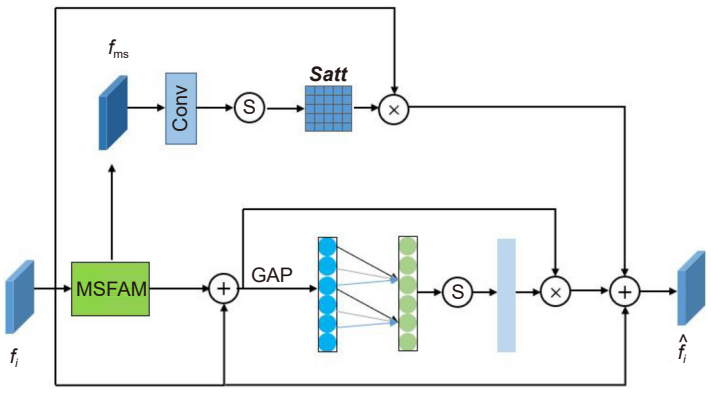


图 3 多尺度聚合模块

Fig. 3 Multi scale aggregation module

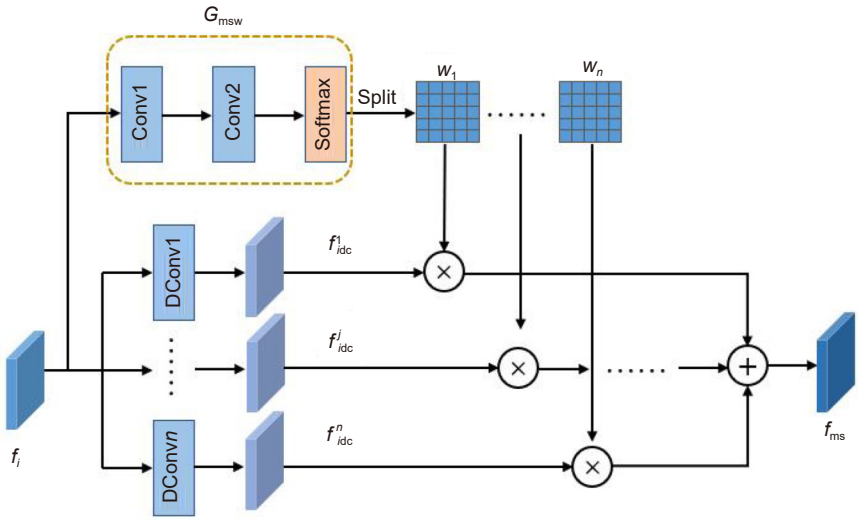


图 4 多尺度特征聚合模块

Fig. 4 Multi scale feature aggregation module

如下：

$$\boldsymbol{satt} = \sigma(Conv_c(f_{ms})), \tag{6}$$

$$f_i^1 = \boldsymbol{satt} \otimes f_i, \tag{7}$$

式中，c 表示特征 f_{ms}的通道数。

特征结合强化分支将 f_{ms}与 f_i执行像素相加操作得到结合特征 f_{ic}，为了使结合特征的信息得到充分的利用，同样采用 GAP 和自适应映射操作生成通道权重，强化结合特征的通道信息。具体表示如下：

$$catt = \sigma(\varphi(GAP(f_{ic}))), \tag{8}$$

$$f_i^2 = catt \otimes f_{ic}. \tag{9}$$

最后将两个强化特征 f_i^1， f_i^2与 f_i相结合得到最终的多尺度聚合特征 $\hat{f}_i$，可以表示为

$$\hat{f}_i = f_i^1 + f_i^2 + f_i. \tag{10}$$

2.4 Gem 池化

在行人重识别任务中，最常用的全局平均池化 (GAP) 或全局最大池化 (GMP) 都不能捕获特定区域的鉴别性特征。Gem 池化[13] 引入了参数 p，通过调整参数 p，可以实现不同程度的非线性变换，综合 GAP 和 GMP 的特点，突出关键特征。因此本文采用 Gem 池化，将三维特征图转换成一维特征图。具体计算方式如下：

$$X_p = \left(\frac{1}{|X|}\sum_{x_i \in X} x_i^p\right)^{\frac{1}{p}}, \tag{11}$$

式中，X 表示维度为三的特征图，X_p表示池化后的结果，p 为超参数，当 p=1 时，等价于全局平均池化，当 p 趋近于无穷大时，相当于全局最大池化。

2.5 损失函数

本文采用身份损失L_{id}和加权正则化三元组损失 L_{wrt}[9] 的组合优化模型。

身份损失L_{id}将具有相同身份的不同模态图像视为同一类，可以缩小类内距离，通常使用交叉熵函数，表示如下：

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N}\log(P(t_i|C(f_i|\theta))), \tag{12}$$

式中：N 表示当前批次中图像的数量，l_i表示特征 f_i的对应标签值，θ 表示分类器的参数。

加权正则化三元组损失L_{wrt}从模态内和跨模态关系中优化所有正样本对和负样本对之间的相对距离。表示如下：

$$L_{wrt} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(1+\exp\left(\sum\nolimits_{ij} w_{ij}^{o} d_{ij}^{o} - \sum\nolimits_{ik} w_{ik}^{e} d_{ik}^{e}\right)\right), \tag{13}$$

式中：(i, j, k) 表示每个锚样本在每个训练批次中的一个三元组。其中 j, k 可以来自相同模态，也可以来自不同模态。对于锚x_i，o 对应正集，e 对应负集。d_{ij}^{o}/d_{ik}^{e}表示正/负样本对的成对距离。d_{ij}表示两个样本之间的欧式距离。

最终的总损失函数为

$$L = L_{id} + L_{wrt}. \tag{14}$$

3 实验结果与分析

3.1 数据集与评价指标描述

为了验证本文算法的有效性，在两个主流公共数

据集 (SYSU-MM01 和 RegDB 数据集) 上进行了实验。

SYSU-MM01 数据集[14] 是 Wu 等人建立的第一个跨模态行人数据集，所有图像来自 4 个 RGB 摄像头和 2 个 IR 摄像头，总共包括 491 个不同的行人。训练集包含 395 人，共有 19659 张 RGB 图像和 12792 张 IR 图像，测试集包含 96 人。主要为两种检索模式：全搜索模式和室内搜索模式。对于室内搜索，只在相机 1，2，3 和 6 中搜索。

RegDB 数据集[15] 总共包含 412 个行人，每个模式下每个人都有 10 张图像。其中女性 254 人，男性 158 人。正面图包含所有的 412 人，而背面图仅包含 256 人。主要有两种检索方式：可见光-红外模式和红外-可见光模式。当一个模态样本作为图像集时，另一个模态样本则作为检索集。过程中，随机选择 206 个身份进行训练，其余 206 个身份进行测试。

评价指标主要包括：rank-1、rank-10、rank-20、mAP、mINP[16-17]。rank-i 识别率代表搜索结果中置信度最高的 i 个结果中正确结果的概率。当 i=1，10，20 时，即计算测试时前 1、10、20 张与查询库中的相似度排序后的同一类的准确率。mAP 将行人重识别看作检索任务，计算数据集中所有类的预测平均精度的均值。计算方式表示为

$$mAP = \sum_{i=1}^{l} \frac{AP_i}{l}, \tag{15}$$

式中：AP_i表示类别的平均精度，l 表示类别的个数。

mINP 用于衡量找到最难匹配样本的效率，避免 mAP 评估中容易匹配的主导地位。计算方式表示为

$$mINP = \frac{1}{n}\sum_i (1 - NP_i) = \frac{1}{n}\sum_i \frac{|G_i|}{R_i^{\text{hard}}}, \tag{16}$$

式中，NP_i表示负惩罚，用于惩罚以找到最难的正确匹配，其计算方式如下：

$$NP_i = \frac{R_i^{\text{hard}} - |G_i|}{R_i^{\text{hard}}}, \tag{17}$$

式中：R_i^{hard}表示最难匹配的排名位置，$|G_i|$表示查询 i 的正确匹配总数。NP 越小表示性能越好。

3.2 实验设置

本文实验环境为显卡 NVINIA RTX A4000 (显存 16 G)、CPU 内存 30 G，在 Ubuntu20.04LTS 的系统上采用 Python3.8 的 Pytorch 深度学习框架。训练时，采用 ImageNet 预训练的 ResNet50 骨干网络进行特征提取。在此基础上，以 CAJ 中的通道增强方法强化图像处理。批处理大小设置为 24；一个批次中行人身份数量设置为 4。本文利用常用的数据增广操作，包括随机裁剪、水平翻转、通道增强和随机噪声操作。对于通道增强方式，每张图像随机选择通道图像作为输入图像。随机噪声方式则是在图像加入局部噪声进行数据增广。训练时使用 sgd 优化器，动量值设为 0.9，权重衰减设置为 5×10^{-4}，超参数 p 设置为 3，初始学习率设置为 0.1，在 20，50 个周期衰减为 0.1 和 0.01。在前 10 个 epoch 应用热身策略。训练总迭代次数为 80[9]。

3.3 实验结果

3.3.1 结果对比

为验证提出方法对于红外-可见光行人重识别的优越性，本文将所提方法与该领域的主流方法在两个数据集上进行了比较，其结果如表 1 和表 2 所示。比较的方法有 One-stream[14]、Two-stream[14]、Zero-padding[14]、HCML[18]、BDTR[17]、D^{2}RL[4]、AlignGAN[19]、AGW[16]、Xmodal[20]、DDAG、CM-NAS[21]、CAJ[9]、MPANet[6]、MCLNet[22]、PIC[23]、DART[24]、SPOT[10]、DML[7]、PMT[12]、SFANet[25]、SIDA[26]、MTMFE[27] 等。其中 One-stream、Two-stream、Zero-padding 为 Wu 等人首次针对跨模态行人重识别任务提出的三种基础框架。D^{2}RL 利用变分编码器 (VAE) 消除模态间差异，Xmodal 利用生成对抗网络 (GAN) 生成第三种模态，将跨模态问题转换成三模态问题。DDAG 使用注意力机制来细化样本特征表示，MPANet 则利用空间注意力引导实例归一化缓解模态间差异，并采用了互均值学习的方式，DML 与 MPANet 学习方式基本相似，但网络结构不同，通过双向学习缓解模态间差异。本文所提方法则是充分利用模态特定空间的特征使得最后的特征更具鲁棒性。

根据表 1 和表 2 可以看出所提方法性能的优越性，在 SYSU-MM01 两种搜索模式下都优于基准网络。表 1 中最后一行展示了本文方法的指标性能，在全搜索单镜头模式下，本文方法在 rank-1，rank-10，rank-20，mAP 分别超出了基准网络 CAJ 方法 3.54%，0.56%，0.21%，3.11%。其中“–”表示原论文中没有报告的结果。

3.3.2 消融实验

本文在 SYSU-MM01 数据集的单镜头全搜索和室内搜索两种模式下进行了一系列消融实验，来证明

表 1 SYSU-MM01 数据集比较结果

Table 1 Comparison results on SYSU-MM01 datasets

Method	Setting									
	All search					Indoor search				
	rank-1	rank-10	rank-20	mAP	mINP	rank-1	rank-10	rank-20	mAP	mINP
One-stream[14]	12.04	49.68	66.74	13.67	-	16.94	63.55	82.10	22.95	-
Two-stream[14]	11.65	47.99	65.50	12.85	-	15.60	61.18	81.02	21.49	-
Zero-Padding[14]	14.80	54.12	71.33	15.59	-	20.58	68.38	85.79	26.92	-
HCML[18]	14.32	53.16	69.17	16.16	-	24.52	73.25	86.73	30.08	-
BDTR[17]	27.32	66.96	81.07	27.32	-	31.92	77.18	89.28	41.86	-
D^2RL[4]	28.90	70.60	82.40	29.20	-	-	-	-	-	-
AlignGAN[19]	42.03	85.25	93.73	41.48	-	45.86	90.17	95.39	55.18	-
AGW[16]	47.58	84.45	92.11	47.69	35.30	54.29	91.14	95.99	63.02	59.23
Xmodal[20]	49.92	89.79	95.96	50.73	-	-	-	-	-	-
DDAG[8]	53.61	89.17	95.30	52.02	39.62	58.37	91.92	97.42	65.44	62.61
CM-NAS[21]	62.04	92.92	97.31	60.00	-	67.03	97.02	99.34	72.97	-
CAJ[9]	68.23	95.59	98.49	65.32	53.61	74.01	97.79	**99.67**	78.52	76.79
MPANet[6]	70.07	95.39	98.39	67.07	-	76.35	97.56	99.48	80.16	-
PIC[23]	57.5	-	-	55.1	-	60.4	-	-	67.7	-
DART[24]	68.72	96.39	98.96	66.29	53.26	72.52	97.84	99.46	78.17	74.94
SPOT[10]	65.34	92.73	97.04	62.25	48.86	69.42	96.22	99.12	74.63	70.48
DML[7]	58.40	91.20	95.80	56.10	-	62.40	95.20	98.70	69.50	-
PMT[12]	67.53	95.36	98.64	64.98	51.86	71.66	96.73	99.25	76.52	72.74
SFANet[25]	65.74	92.98	97.05	60.83	-	71.60	96.60	99.45	80.05	-
SIDA[26]	68.36	95.91	98.56	64.19	-	73.28	97.35	99.52	77.49	-
MTMFE[27]	69.47	**96.42**	99.11	66.41	-	72.56	96.98	99.20	76.58	-
Ours	**71.77**	96.15	**98.7**	**68.43**	**55.21**	**78.24**	**98.23**	99.49	**81.9**	**78.44**

表 2 RegDB 数据集比较结果

Table 2 Comparison results on RegDB datasets

Method	Setting									
	Visible to infrared					Infrared to visible				
	rank-1	rank-10	rank-20	mAP	mINP	rank-1	rank-10	rank-20	mAP	mINP
Zero-Padding[14]	17.75	34.21	44.35	18.90	-	16.63	34.68	44.25	17.82	-
HCML[18]	24.44	47.53	56.78	20.08	-	21.70	45.02	55.58	22.24	-
BDTR[17]	33.56	58.61	67.43	32.76	-	32.92	58.46	68.43	31.96	-
D^2RL[4]	43.40	66.10	76.30	44.10	-	-	-	-	-	
AGW[16]	70.05	86.21	91.55	66.37	50.19	70.49	87.21	91.84	65.90	51.24
Xmodal[20]	62.21	83.13	91.72	60.18	-	-	-	-	-	-
DDAG[8]	69.34	85.77	89.98	63.19	49.24	64.77	83.85	88.90	58.54	48.62
CM-NAS[21]	84.54	95.18	**97.85**	80.32	-	82.56	94.52	**97.37**	78.31	-
MCLNet[22]	80.31	92.70	96.03	73.07	-	75.93	90.93	94.59	69.49	-
CAJ[9]	84.72	95.17	97.38	78.70	65.33	84.09	94.79	97.11	77.25	61.56
PIC[23]	83.6	-	-	79.6	-	79.5	-	-	77.4	-
DART[24]	78.23	-	-	67.04	48.36	75.04	-	-	64.38	43.32
SPOT[10]	80.35	93.48	96.44	72.46	56.19	79.37	92.79	96.01	72.26	56.06
DML[7]	77.60	-	-	**84.30**	-	77.00	-	-	**83.60**	-
PMT[12]	84.83	-	-	76.55	-	84.16	-	-	75.13	-
SFANet[25]	76.31	91.02	94.27	68.00	-	70.15	85.24	89.27	63.77	-
SIDA[26]	81.73	-	96.55	75.07	-	79.71	-	95.47	72.60	-
MTMFE[27]	85.04	94.38	97.22	82.52	-	81.11	92.35	96.19	79.59	-
Ours	**85.38**	**95.39**	97.54	79.49	**65.72**	**84.58**	**95.27**	97.23	78.02	**62.22**

MFANet 网络中各个模块的有效性。在实验中，依次删除多尺度聚合模块、邻级特征聚合模块分别训练，验证该模块是否在网络中不可或缺。

实验结果如表 3 所示，加粗数据表示最优结果。首先，将框架中的模块删除得到基准网络，即行 1；随后在其基础上加入邻级特征聚合模块 (AFAM) 或多尺度聚合模块 (MSAM)，由表中数据可得两个模块对模型都是正优化；最后将两个模块都加入基础网络中，得到最优的情况，进一步验证了模块的有效性。特别地，基准网络的 Params 为 23.540 M，flops 为 6268.596 M，本文所提网络的 Params 为 66.298 M，flops 为 15119.356 M。图 5(a-d) 表示不同设定下的类内距离和类间距离间关系，不同颜色表示不同的情形，可以看出 MFANet 的类间距离与基础网络相似，类内距离有所减小，因此，MFANet 可以有效地减小类内差异。图 5(e-h) 表示不同设定下的特征分布图，颜色相同的标志代表同一身份特征，由图可得四种情况下都可以有效地区分和聚合同一个人的特征嵌入。

3.3.3 多尺度特征聚合模块感受野分析

本节进行一系列实验来研究多尺度特征聚合模块中感受野选择对模型的影响，分别选择 5 种不同的感受野组合在 SYSU-MM01 数据集上进行实验，寻找最优解，如表 4 所示。

根据表 4 结果可以发现，当感受野过大时，由于深层特征的分辨率较小，大感受野会导致关注了更多的空余位置，导致整体模型性能较差。随着感受野的减小，性能随之提升。并且伴随着感受野数量和大小的选择进行微小下降后，性能达到最优，因此采用感受野为 (1, 3) 的方式进行多尺度特征聚合。

为了更直接地比较不同感受野下模型的性能，随机取 SYSU-MM01 数据集中的不同相机拍摄的两张图像，利用 GradCam[28] 热力图进行可视化分析，具体如图 6 所示，图中 original 列表示原始图像。(a-e) 分别表示感受野设定为 (1, 3, 5, 7)，(1, 2, 3, 4)，(1, 3, 5)，

表 3　SYSU-MM01 数据集上 4 种不同设定的消融研究

Table 3　Ablation study of four different settings on the SYSU-MM01 dataset

Settings		All search					Indoor search				
AFAM	MSAM	rank-1	rank-10	rank-20	mAP	mINP	rank-1	rank-10	rank-20	mAP	mINP
		68.23	95.59	98.49	65.32	51.90	74.01	97.79	**99.67**	78.52	74.78
√		69.30	95.69	98.41	65.95	52.14	75.27	97.84	99.48	79.51	75.79
	√	70.89	95.88	98.52	67.61	54.3	77.69	97.43	99.25	81.09	77.48
√	√	**71.77**	**96.15**	**98.70**	**68.43**	**55.21**	**78.24**	**98.23**	99.49	**81.90**	**78.44**

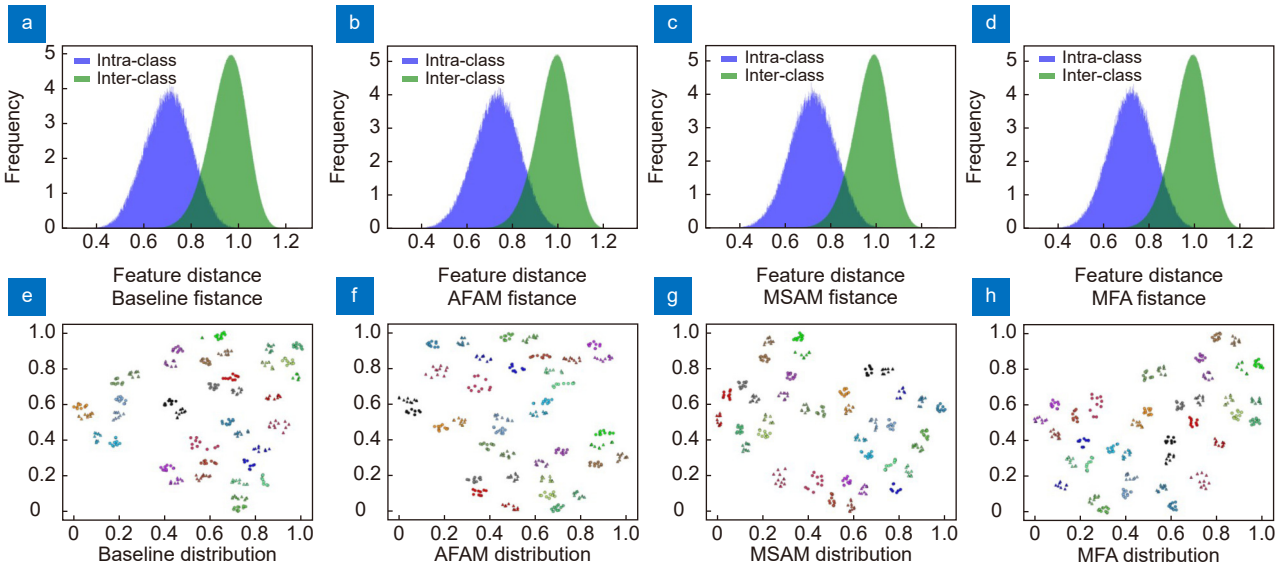


图 5　类内类间距离与特征分布图

Fig. 5　Inter-class intra-class distance and feature distribution diagram

表 4　多尺度特征聚合模块感受野分析

Table 4　Multi scale feature aggregation module receptive field analysis

Settings	Receptive field									
	All search					Indoor search				
	rank-1	rank-10	rank-20	mAP	mINP	rank-1	rank-10	rank-20	mAP	mINP
1，3，5，7	69.41	95.82	98.54	66.27	52.96	75.34	97.96	**99.66**	79.98	76.53
1，2，3，4	70.17	95.51	98.51	67.14	54.1	76.44	98.09	99.6	80.65	77.22
1，3，5	70.5	95.77	98.54	67.11	53.68	76.66	97.81	99.51	80.54	76.99
1，2，3	70.53	95.86	98.67	67.21	53.77	76.59	97.88	99.37	80.68	77.22
1，3	**71.77**	**96.15**	**98.70**	**68.43**	**55.21**	**78.24**	**98.23**	99.49	**81.90**	**78.44**

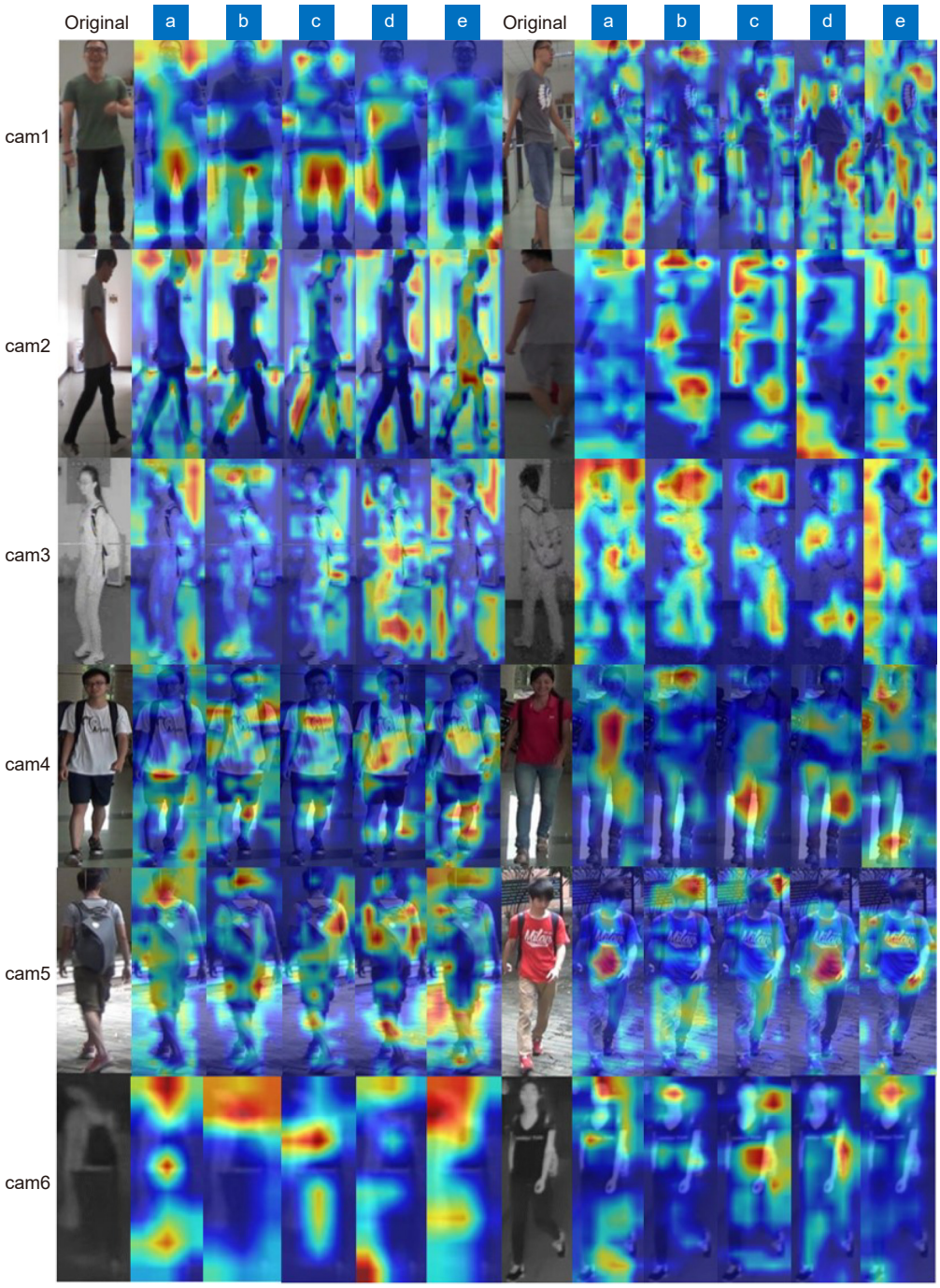


图 6　不同感受野下的热力图

Fig. 6　Heat map at different receptive fields of view

(1, 2, 3)，(1, 3) 的热力图结果。图中颜色越鲜艳、越深表示模型更关注，该部分对于最终特征表示的重要程度越高。cam3, 6 为红外相机拍摄的图像，由于红外模态信息的影响，关注了更多的背景信息。cam6 的第一张图像属于低分辨率图像，由热力图可以发现在不同感受野设定下关注信息偏向于人物的头部特征。cam1, 2, 4, 5 为 RGB 相机拍摄的图像，根据 cam4 的第一张图像和 cam5 的第二张图像的热力图可以发现 (1, 3) 设定下的模型更关注鉴别性特征 (图中对应衣服的标志特征)。相比其他感受野设定，由热力图可见在 cam1 和 cam2 中更关注人物特征。综合整体效果，相比其他感受野，(1, 3) 更具有鉴别性。

3.3.4 可视化分析

为了直观地分析本文所提方法的检索效果，在 SYSU-MM01 数据集上实现了可视化排序，随机选择了 12 张查询示例，图 7 中前三排的六个示例以 RGB 图像为检索图像，前三排的六个示例则以 IR 图像作为检索图像，显示查询示例的前 10 个检索结果，如图 7 所示，图中绿框表示匹配正确的图像，红色表示身份匹配错误。

图 7 中可以看出，SYSU-MM01 数据集中可见光行人图像的颜色信息与红外图像的差异较大，并且还伴随背景变化。因此采用多特征聚合的方式能充分利用不同级别的空间信息和上下文信息，这样能减少背景对检索任务的影响。由图 7(b) 第二行的检索结果可以发现，当红外图像包含极少的身份信息时，检索性能会偏差。同种情况下，图 7(b) 第三行的检索图像包含了背包信息，因此在其帮助下，性能也会有所增强，说明该模型有很好的鉴别性特征获取能力。而正面的 RGB 图像包含更多身份信息，因此检索性能较优。根据可视化结果可以证明所提模型的有效性。

图 7 SYSU-MM01 可视化排序结果

Fig. 7 Visual sorting results on SYSU-MM01

4 结束语

本文针对红外-可见光行人重识别提出了一种多特征聚合网络 (MFANet)，该网络主要包括双流网络主干、邻级特征聚合模块和多尺度聚合模块。其中，邻级特征聚合模块对相邻级别特征不同维度的空间信息进行聚合以强化高级特征，使得高级特征更具健壮性。多尺度聚合模块首先对不同感受野的特征进行聚合，得到多尺度特征，再将其的空间信息与源特征进行聚合，以强化特征的多尺度信息，使得特征包含更多的上下文信息。本文方法在两个数据集上进行了与现有方法的对比实验，充分验证了所提方法的先进性，具有更好的鲁棒性和鉴别性。

参考文献

[1] Liu L, Li X, Lei X M. A person re-identification method with multi-scale and multi-feature fusion[J]. *J Comput-Aided Des Comput Graphics*, 2022, **34**(12): 1868−1876.
刘丽, 李曦, 雷雪梅. 多尺度多特征融合的行人重识别模型[J]. 计算机辅助设计与图形学学报, 2022, **34**(12): 1868−1876.

[2] Shi Y X, Zhou Y. Person re-identification based on stepped feature space segmentation and local attention mechanism[J]. *J Electron Inf Technol*, 2022, **44**(1): 195−202.
石跃祥, 周玥. 基于阶梯型特征空间分割与局部注意力机制的行人重识别[J]. 电子与信息学报, 2022, **44**(1): 195−202.

[3] Wang S, Ji P, Zhang Y Z, et al. Adaptive receptive network for person re-identification[J]. *Control Decis*, 2022, **37**(1): 119−126.
王松, 纪鹏, 张云洲, 等. 自适应感受野网络的行人重识别[J]. 控制与决策, 2022, **37**(1): 119−126.

[4] 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 618–626. https://doi.org/10.1109/CVPR.2019.00071.

[5] Zhong X, Lu T Y, Huang W X, et al. Grayscale enhancement colorization network for visible-infrared person re-identification[J]. *IEEE Trans Circ Syst Video Technol*, 2022, **32**(3): 1418−1430.

[6] Wu Q, Dai P Y, Chen J, et al. Discover cross-modality nuances for visible-infrared person re-identification[C]//*Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 4330–4339. https://doi.org/10.1109/CVPR46437.2021.00431.

[7] Zhang D M, Zhang Z Z, Ju Y, et al. Dual mutual learning for cross-modality person re-identification[J]. *IEEE Trans Circ Syst Video Technol*, 2022, **32**(8): 5361−5373.

[8] 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 16th European Conference on Computer Vision*, 2020: 229–247. https://doi.org/10.1007/978-3-030-58520-4_14.

[9] Ye M, Ruan W J, Du B, et al. Channel augmented joint learning for visible-infrared recognition[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 13567–13576. https://doi.org/10.1109/ICCV48922.2021.01331.

[10] Chen C Q, Ye M, Qi M B, et al. Structure-aware positional transformer for visible-infrared person re-identification[J]. *IEEE Trans Image Process*, 2022, **31**: 2352−2364.

[11] 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*, 2017: 6000–6010.

[12] Lu H, Zou X Z, Zhang P P. Learning progressive modality-shared transformers for effective visible-infrared person re-identification[C]//*Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023: 1835–1843. https://doi.org/10.1609/aaai.v37i2.25273.

[13] Lin B B, Zhang S L, Yu X. Gait recognition via effective global-local feature representation and local temporal aggregation[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 14648–14656. https://doi.org/10.1109/ICCV48922.2021.01438.

[14] Wu A C, Zheng W S, Yu H X, et al. RGB-infrared cross-modality person re-identification[C]//*Proceedings of 2017 IEEE International Conference on Computer Vision*, 2017: 5380–5389. https://doi.org/10.1109/ICCV.2017.575.

[15] 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.

[16] Ye M, Shen J B, Lin G J, et al. Deep learning for person re-identification: a survey and outlook[J]. *IEEE Trans Pattern Anal Mach Intell*, 2022, **44**(6): 2872−2893.

[17] Ye M, Lan X Y, Wang Z, et al. Bi-directional center-constrained top-ranking for visible thermal person re-identification[J]. *IEEE Trans Inf Forensics Secur*, 2020, **15**: 407−419.

[18] Ye M, Lan X Y, Li J W, et al. Hierarchical discriminative learning for visible thermal person re-identification[C]//*Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018: 919. https://doi.org/10.1609/aaai.v32i1.12293.

[19] 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 2019 IEEE/CVF International Conference on Computer Vision*, 2019: 3623–3632. https://doi.org/10.1109/ICCV.2019.00372.

[20] 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 34th AAAI Conference on Artificial Intelligence*, 2020: 4610–4617. https://doi.org/10.1609/aaai.v34i04.5891.

[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]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 11823–11832. https://doi.org/10.1109/ICCV48922.2021.01161.

[22] Hao X, Zhao S Y, Ye M, et al. Cross-modality person re-identification via modality confusion and center aggregation[C]//*Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, 2021: 16403–16412. https://doi.org/10.1109/ICCV48922.2021.01609.

[23] Zheng X T, Chen X M, Lu X Q. Visible-infrared person re-identification via partially interactive collaboration[J]. *IEEE Trans Image Process*, 2022, **31**: 6951−6963.

[24] Yang M X, Huang Z Y, Hu P, et al. Learning with twin noisy labels for visible-infrared person re-identification[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition, 2022: 14308–14317. https://doi.org/10.1109/CVPR52688.2022.01391.

[25] Liu H J, Ma S, Xia D X, et al. SFANet: a spectrum-aware feature augmentation network for visible-infrared person reidentification[J]. *IEEE Trans Neural Netw Learn Syst*, 2023, **34**(4): 1958−1971.

[26] Gong J H, Zhao S Y, Lam K M, et al. Spectrum-irrelevant fine-grained representation for visible–infrared person re-identification[J]. *Comput Vis Image Underst*, 2023, **232**: 103703.

[27] Huang N C, Liu J N, Luo Y J, et al. Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person Re-IDentification[J]. *Pattern Recogn*, 2023, **135**: 109145.

[28] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. *Int J Comput Vis*, 2020, **128**(2): 336−359.

作者简介

郑海君 (1998-)，男，硕士，主要从事图像处理、计算机视觉等方面的研究。

E-mail: navy626@163.com

【通信作者】葛斌 (1975-)，男，博士，教授，硕士生导师，主要从事图像处理、信息安全等方面的研究。

E-mail: bge@aust.edu.cn

夏晨星 (1991-)，男，博士，副教授，硕士生导师，主要从事图像处理、计算机视觉等方面的研究。

E-mail: cxxia@aust.edu.cn

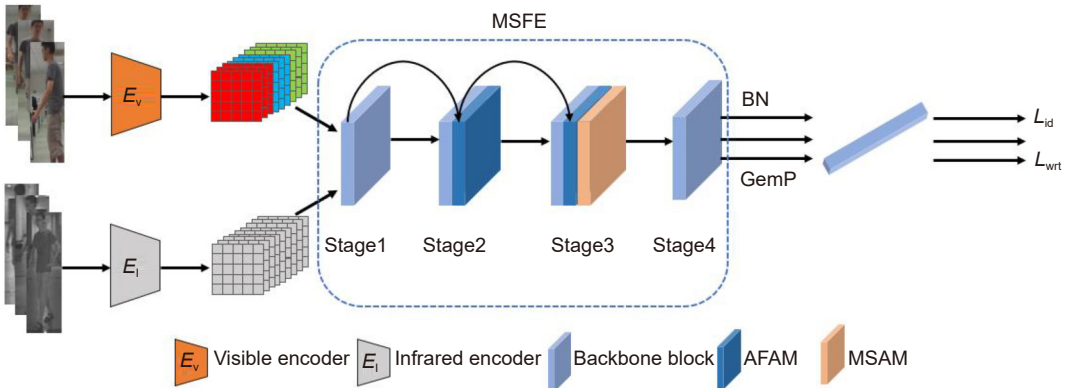
邬成 (1998-)，男，硕士，主要从事图像处理、计算机视觉等方面的研究。

E-mail: 1965596900@qq.com

扫描二维码，获取PDF全文

Infrared-visible person re-identification based on multi feature aggregation

Zheng Haijun[1], Ge Bin[1*], Xia Chenxing[1,2], Wu Cheng[1]



MFANet structure diagram

Overview: Infrared-visible person re-identification is a prominent research topic in the field of computer vision, encompassing several essential aspects. These include multi-modal perception technology, challenges in person re-identification, practical application demands, and the development of datasets and evaluation metrics. With the emergence of multi-modal perception technology, the primary objective of infrared-visible light person re-identification is to effectively fuse information from different modalities to enhance the accuracy and robustness of person re-identification. Person re-identification faces challenges such as variations in viewpoint, pose, occlusion, and lighting conditions. Furthermore, infrared-visible person re-identification poses additional challenges as a cross-modal task. This technology holds broad prospects for applications in video surveillance, security, intelligent transportation, and other related fields. Particularly, it is well-suited for person re-identification in low-light or nighttime environments. The development of relevant datasets and evaluation metrics has facilitated ongoing innovation and improvement in infrared-visible person re-identification algorithms and systems. Infrared-visible person re-identification is a research field extensively supported by various backgrounds, providing a foundation for enhancing the performance and application effectiveness of person re-identification. With the continuous exploration of researchers, the accuracy of infrared-visible person re-identification has steadily improved. However, due to the differences between different image modalities, it brings great challenges to this field. The existing methods mainly focus on mitigating the differences between modes to obtain more discriminating features, but ignore the relationship between adjacent features and the influence of multi-scale information on global features. Here, a infrared-visible person re-identification method (MFANet) based on multi-feature aggregation is proposed to solve the shortcomings of existing methods. Firstly, the adjacent level features are fused in the feature extraction stage, and the integration of low-level feature information is guided to strengthen the high-level features and make the features more robust. Then, the multi-scale features of different receptive fields of view are aggregated to obtain rich contextual information. Finally, multi-scale features are used as a guide to strengthen the features to obtain more discriminating features. Experimental results on SYSU-MM01 and RegDB datasets show the effectiveness of the proposed method, and the average accuracy of SYSU-MM01 dataset reaches 71.77% in the all-search single-shot mode and 78.24% in the indoor-search single-shot mode.

Zheng H J, Ge B, Xia C X, et al. Infrared-visible person re-identification based on multi feature aggregation[J]. *Opto-Electron Eng*, 2023, **50**(7): 230136; DOI: 10.12086/oee.2023.230136

Foundation item: Project supported by National Natural Science Foundation of China (6210071479, 62102003), National Science and Technology Major Project (2020YFB1314103), Natural Science Foundation of Anhui Province(2108085QF258) and Anhui Postdoctoral Fund (2022B623)

[1]College of Computer Science and Engineering, Anhui University of Science and Technology, Huainan, Anhui 232001, China; [2]Institute of Energy, Hefei Comprehensive National Science Center, Hefei, Anhui 230031, China

* E-mail: bge@aust.edu.cn