CN 51-1346/O4 ISSN 1003-501X (印刷版) ISSN 2094-4019 (网络版)

光电工程
Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

四流输入引导的特征互补可见光–红外行人重识别

葛斌，许诺，夏晨星，郑海君

引用本文:

葛斌，许诺，夏晨星，等. 四流输入引导的特征互补可见光–红外行人重识别[J]. 光电工程，2024， **51**(9): 240119.

Ge B, Xu N, Xia C X, et al. Quadrupl-stream input-guided feature complementary visible-infrared person re-identification[J]. *Opto-Electron Eng*, 2024, **51**(9): 240119.

https://doi.org/10.12086/oee.2024.240119

收稿日期: 2024-05-23; 修改日期: 2024-08-16; 录用日期: 2024-08-18

相关论文

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
光电工程 2023, **50**(7): 230136 doi: 10.12086/oee.2023.230136

基于红外和可见光模态的随机融合特征金子塔行人重识别
汪荣贵，王静，杨娟，薛丽霞
光电工程 2020, **47**(12): 190669 doi: 10.12086/oee.2020.190669

基于多分区注意力的行人重识别方法
薛丽霞，朱正发，汪荣贵，杨娟
光电工程 2020, **47**(11): 190628 doi: 10.12086/oee.2020.190628

软多标签和深度特征融合的无监督行人重识别
张宝华，朱思雨，吕晓琪，谷宇，王月明，刘新，任彦，李建军，张明
光电工程 2020, **47**(12): 190636 doi: 10.12086/oee.2020.190636

更多相关论文见光电期刊集群网站

OEE 光电工程 Opto-Electronic Engineering
http://cn.oejournal.org/oee

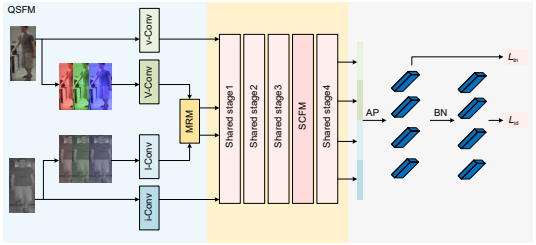
OE_Journal
Website

Opto-Electronic Engineering
光 电 工 程

Article
2024 年, 第 51 卷, 第 9 期

DOI: 10.12086/oee.2024.240119 CSTR: 32245.14.oee.2024.240119

四流输入引导的特征互补可见光-红外行人重识别



葛 斌[1,2*]，许 诺[1]，夏晨星[1]，郑海君[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： 目前可见光-红外行人重识别研究侧重于通过注意力机制提取模态共享显著性特征来最小化模态差异。然而，这类方法仅关注行人最显著特征，无法充分利用模态信息。针对此问题，本文提出了一种四流输入引导的特征互补网络 (QFCNet)。首先在模态特定特征提取阶段设计了四流特征提取和融合模块，通过增加两流输入，缓解模态间颜色差异，丰富模态的语义信息，进一步促进多维特征融合；其次设计了一个次显著特征互补模块，通过反转操作补充全局特征中被注意力机制忽略的行人细节信息，强化行人鉴别性特征。在 SYSU-MM01, RegDB 两个公开数据集上的实验数据表明了此方法的先进性，其中在 SYSU-MM01 的全搜索模式中 rank-1 和 mAP 值达到了 76.12% 和 71.51%。

关键词： 跨模态；行人重识别；红外；数据增强；注意力机制

中图分类号： TP391 **文献标志码：** A

葛斌，许诺，夏晨星，等. 四流输入引导的特征互补可见光-红外行人重识别 [J]. 光电工程，2024，**51**(9): 240119
Ge B, Xu N, Xia C X, et al. Quadrupl-stream input-guided feature complementary visible-infrared person re-identification[J]. *Opto-Electron Eng*, 2024, **51**(9): 240119

Quadrupl-stream input-guided feature complementary visible-infrared person re-identification

Ge Bin[1,2*], Xu Nuo[1], Xia Chenxing[1], Zheng Haijun[1]

[1]School 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: Current visible-infrared person re-identification research focuses on extracting modal shared saliency features through the attention mechanism to minimize modal differences. However, these methods only focus on the most salient features of pedestrians, and cannot make full use of modal information. To solve this problem, a quadrupl-stream input-guided feature complementary network (QFCNet) is proposed in this paper. Firstly, a quadrupl-stream feature extraction and fusion module is designed in the mode-specific feature extraction stage. By adding two data enhancement inputs, the color differences between modalities are alleviated, the semantic

收稿日期：2024-05-23；修回日期：2024-08-16；录用日期：2024-08-18
基金项目：国家自然科学基金资助项目 (62102003)；安徽省自然科学基金资助项目 (2108085QF258)； 安徽省博士后基金资助项目 (2022B623)
*通信作者：葛斌，bge@aust.edu.cn。

information of the modalities is enriched and the multi-dimensional feature fusion is further promoted. Secondly, a sub-salient feature complementation module is designed to supplement the pedestrian detail information ignored by the attention mechanism in the global feature through the inversion operation, to strengthen the pedestrian discriminative features. The experimental results on two public datasets SYSU-MM01 and RegDB show the superiority of this method. In the full search mode of SYSU-MM01, the rank-1 and mAP values reach 76.12% and 71.51%, respectively.

Keywords: cross-modal; person re-identification; infrared; data augmentation; attention mechanism

1 引　言

行人重识别[1-3](Person re-identification, ReID) 是一种旨在检索和识别非重叠摄像机拍摄图像中行人个体的技术。随着人们对于公共安全的日益重视，全天候的监控与检索系统在计算机视觉领域受到了极大的关注。然而，现有的大多数模型专注于可见光相机捕获的单模态 RGB 图像的匹配任务。但在较差的光照条件下，可见光摄像机无法捕捉到足够的信息。因此，越来越多的监控系统开始整合红外摄像机，以便匹配可见光和红外摄像机捕获的行人图像，实现全天候监控。由于在检索过程中需要将与身份相关的红外信息与可见光信息相结合，可见光-红外行人重识别 (visible-infrared person re-identification, VI-ReID) 技术应运而生。然而，红外光散射会造成图像中常见的颜色、纹理等细节信息的丢失，从而导致可见光和红外模态之间的差异。此外，在相同的模态中，摄像机视角、行人服装变化和场景遮挡也给 VI-ReID 领域带来了一定的挑战。

现有的 VI-ReID 方法主要集中在两个方面：1) 通过最大化模态不变性来解决跨模态问题；2) 生成中间或目标图像，将跨模态匹配问题转化为模态内匹配任务。

第一种方法旨在找到模态共享的特征空间，在该空间中学习不随模态改变的特征表示。为此，Zheng 等[4] 使用邻级特征之间的融合策略来强化高级特征，通过多尺度信息引导学习鉴别性特征。Zhang 等[5] 在共享特征提取阶段通过聚合来自不同阶段的特征以挖掘不同的通道和空间特征表示，并在特征嵌入空间生成多样化的样本，以解决训练数据不足的问题。Wu 等[6] 使用模态缓解和模式对齐联合提取特征，在互均值学习方式的指导下有选择地使用实例归一化，在减少模态差异的同时保持身份信息，发现跨模态细微差别。Zhang 等[7] 从细粒度通道间语义和综合多式语义两个方面优化通道一致性，实现跨模态语义对齐。Ye 等[8] 提出一种动态双注意聚合学习方法来挖掘模态内局部和跨模态图级上下文线索。虽然神经网络有良好的特征提取能力，但模态不变性往往难以保证特征的质量，这导致了人物图像表示中的间接信息丢失。

第二种方法通过 GAN[9-11] 的方式将图像从一种模态转移到另一种模态，或创建中间模态图像进行训练，在图像层面弥合两种模态之间的差异。Ye 等[12] 利用生成的灰度图像作为辅助模态训练更合适的网络。Zhang 等[13] 将单模态特征分解为模态专用特征和模态共享特征，然后从现有的模态共享特征中生成那些缺失的模态特定特征。Lu 等[14] 提出了一种动态中间模态学习策略，捕获具有跨模态一致性的模态不变特征。Li 等[15] 设计了一个轻量级网络来学习可见光和红外图像之间的中间模态，而不是专注于将可见光图像直接映射到红外图像。虽然这类方法可以有效地处理可见光和红外图像之间的差异。但在训练过程中其计算复杂度高且不可避免地引入噪声，从而影响训练的稳定性，使得生成图像的质量难以保证。

针对上述问题，本文提出了一种四流输入引导的特征互补网络 (quadruple-stream input-guided feature complementary networks, QFCNet)，该模型专注于行人全局鉴别性特征提取，使网络输出的行人特征更全面，从而提高行人检索效率。具体地，本文设计了四流特征提取和融合模块 (quadruple stream feature extraction and fusion module, QSFM) 和次显著特征互补模块 (sub-critical features complementary module, SFCM)。四流特征提取和融合模块 (QSFM) 将原始输入通过数据增强操作生成的图片作为两个全新的输入流并对新的输入流进一步细化，对其重新加权，去除背景噪声，然后和原始输入一起作为网络的四流输入，在测试阶段分别将属于可见光和红外模态的两流输入融合为两类。次显著特征互补模块 (SFCM) 首先通过

自注意力机制获得行人的跨模态特征，然后通过反转操作获得行人的互补细微特征，最后将二者融合，获得行人丰富的全局显著性特征。

QFCNet 网络不仅可以增强行人用于匹配的鉴别性特征，还能有效地缓解由于成像原理不同导致的模态差异。该网络的主要贡献如下：

1) 提出了一个四流特征提取和融合模块，能够在通道和全局级别提取更广泛的语义信息，促进特征不同维度语义信息的融合，探索跨模态特征之间的隐式相关性，并提取综合模态特征。

2) 设计了次显著特征互补模块，通过自注意力机制引导网络增强行人特征的学习并通过反转操作挖掘被自注意力机制所忽略的次显著特征，从而增强行人表征的全面性和显著性。

3) 通过对 SYSU-MM01、RegDB 数据集的广泛评估，本文提出的 QFCNet 在 VI-RelD 任务中优于当前主流方法。

2 本文方法

2.1 网络架构

设$V=\{x_i^{\mathrm{v}}|x_i^{\mathrm{v}}\in V\}$, $R=\{x_i^{\mathrm{r}}|x_i^{\mathrm{r}}\in R\}$分别为跨模态行人重识别任务数据集中的 RGB 图像集和 IR 图像集，其对应的真值标签表示为$Y_{\mathrm{v}}=\{y_{x_i^{\mathrm{v}}}|x_i^{\mathrm{v}}\in V\}$，$Y_{\mathrm{r}}=\{y_{x_i^{\mathrm{r}}}|x_i^{\mathrm{r}}\in R\}$，为了方便表示，将$y_{x_i^{\mathrm{v}}}$记作$y_i^{\mathrm{v}}$，$y_{x_i^{\mathrm{r}}}$记作$y_i^{\mathrm{r}}$。VI-ReID 将相同身份的可见光图像$x_i^{\mathrm{v}}$与红外图像$x_i^{\mathrm{r}}$相互匹配。因此，VI-ReID 的优化目标是使可见光图像x_i^{v}与红外图像x_i^{r}在属于同一行人身份时的映射相似性最大化，并使具有不同身份图像之间有区别。基准模型 AGW[16] 采用双流网络作为特征提取器。用于特征提取的骨干网络由两部分组成，即模态共享权值θ_w和模态特定提取器θ_{v}，θ_{r}。输入可见光和红外图像后，利用模态特定层θ_{v}，θ_{r}分别提取可见光和红外模态表示。然后应用权值共享机制获得模态共享特征。为了提取行人全面的鉴别性特征，提高行人匹配的效率，本文提出了四流输入引导的特征互补网络 (QFCNet)，网络整体架构如图 1 所示。在基准网络的输入阶段，设计了四流特征提取和融合模块 (QSFM)，扩充可见光和红外模态的语义信息。在模态共享特征提取阶段，设计了次显著特征互补模块 (SFCM) 以增强特征的全面显著性表达。在特征约束阶段，使用加权正则化三元组损失和身份损失对输出的特征进行约束。

2.2 四流特征提取和融合模块 (QSFM)

可见光图像和红外图像的成像原理不同，因而两个模态之间存在很大的模态差异。RGB 图像显示颜色、纹理和其他细粒度特征等细节。相比之下，红外图像是单通道的，故而缺乏此类信息。然而，RGB 图像的每一个通道强调的内在特征各有不同，但 R/G/B 通道之间具有显著的相关性，用来表示整体的实例语义特征，比如身体的姿势和轮廓，而不是仅仅关注颜色或细节等信息。通道语义本质上表示与身份相关的细粒度和多样化的信息。而红外图像是根据物体表面的辐射捕获的，因此不能将其视为由三个通道组成的普通图像。然而，现有的大多数方法主要是在实例级上减少模态差异。他们将红外图像视为由三个

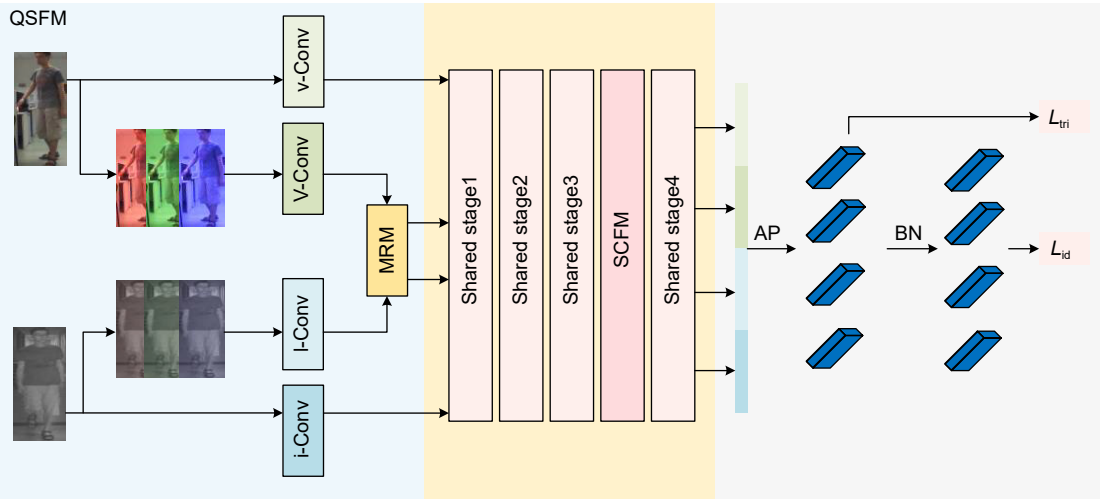


图 1 QFCNet 结构图

Fig. 1 QFCNet structure diagram

通道组成的普通图像，故而在处理过程中忽略了模态之间的语义相关性。为了实现 IR 图像和 RGB 图像之间的语义对齐，我们将 IR 图像扩展为三通道并提供通道信息来增强红外模态。这使得具有三个通道的 IR 图像与 RGB 图像相似，并且每个通道包含不同的信息。对于 RGB 图像，我们采用随机通道交换策略[17]来选择通道语义信息，主要思想是随机选择一个通道 (R/G/B) 来替换其他通道。因此，我们获得了强调不同语义信息的数据集。将 RGB 数据集中的图像进行扩充，得到一组通道级别的图像$\{x_i^{\mathrm{v_c}}\}_{i=1}^N$。同样地，IR 数据集的通道级图像为$\{x_i^{\mathrm{r_c}}\}_{i=1}^N$。它们与原始输入图像$\{x_i^{\mathrm{v_g}}\}_{i=1}^N$和$\{x_i^{\mathrm{r_g}}\}_{i=1}^N$一起组成四流输入。将他们设为$\{X_i^{\mathrm{V_g}}, X_i^{\mathrm{V_c}}, X_i^{\mathrm{R_g}}, X_i^{\mathrm{R_c}}\}$，其中$X_i \in \mathbb{R}^{[B,C,H,W]}$，$i$ 表示行人身份。V 和 R 分别表示可见光模态和红外模态，N 表示身份的总数。B 表示批处理大小，C 表示图像中的通道数，H 和 W 表示图像的高度和宽度。

将数据增强后的图像与原始图像一起作为四流输入，输入到模态特定特征提取器中，与以往使用数据增强的双流网络不同的是，我们保留了数据增强产生的模态特定的语义信息，而不是仅仅在语义维度上提取模态共享特征。它在不共享参数的情况下提取增强图像和原始图像中包含不同语义维度信息的特征。具体地，给定一组输入图像$[X_i^{\mathrm{V_g}}, X_i^{\mathrm{V_c}}, X_i^{\mathrm{R_g}}, X_i^{\mathrm{R_c}}]$，我们使用四流特征提取器 E 进行特征提取，如下所示：

$$[F_i^{\mathrm{V_g}}, F_i^{\mathrm{V_c}}, F_i^{\mathrm{R_g}}, F_i^{\mathrm{R_c}}] = E([X_i^{\mathrm{V_g}}, X_i^{\mathrm{V_c}}, X_i^{\mathrm{R_g}}, X_i^{\mathrm{R_c}}]) . \tag{1}$$

经过数据增强的图像，每个通道包含不同的细节特征。为了凸显行人信息，弱化背景信息影响的同时保留完整的语义特征，因此在进行四流特征融合之前，将新增的两流输入特征输入到模态重加权缓解模块 (modality reweight and alleviation module) 中细化特征。具体流程如图 2 所示。

模态重加权缓解模块主要具备以下功能：1) 探索 RGB 和 IR 图像显著性区域并选择性地融合；2) 在通道维度和空间维度聚合两个模态的特征，过滤出只与模态有关的真正有价值的信息。该模块由两个子模块组成：模态重加权模块 (modality reweight module, MRM) 和模态缓解模块 (modality alleviation module, MAM)。

模态重加权模块 (MRM) 通过强调模态特定行人信息，抑制复杂背景信息，从而在模态特定特征提取阶段处理好模态级的关系。模态重加权模块的内在思想是：通过有颜色信息的 RGB 图像去加强 IR 图像，突出行人信息，弱化背景信息。首先，对网络 stage0 的输出通过逐元素相乘进行融合，逐元素相乘操作可以突出多模态之间对应元素之间共同的显著性信息；然后用模态原始信息减去融合特征得到只与模态相关的特征，具体操作如下：

$$E_{\mathrm{v}} = S_0(RGB) , \tag{2}$$

$$E_{\mathrm{r}} = S_0(IR) , \tag{3}$$

$$U = E_{\mathrm{v}} \times E_{\mathrm{r}} , \tag{4}$$

$$U_{\mathrm{v}} = E_{\mathrm{v}} - U , \tag{5}$$

$$U_{\mathrm{r}} = E_{\mathrm{r}} - U , \tag{6}$$

$$F = [U_{\mathrm{v}}, U_{\mathrm{r}}] , \tag{7}$$

其中：S_0代表 stage0，×表示逐元素相乘，U_{v}和U_{r}为只属于 RGB 模态和 IR 模态的特征，$[\cdot,\cdot]$代表通道维度的连接操作，$F \in \mathbb{R}^{2C\times H\times W}$是在通道维度连接$U_{\mathrm{v}}$和$U_{\mathrm{r}}$。为了探索 RGB 模态和 IR 模态之间的相互依赖关系，本文通过卷积操作来聚合 RGB 图像和 IR 图像在通道之间的特征依赖。

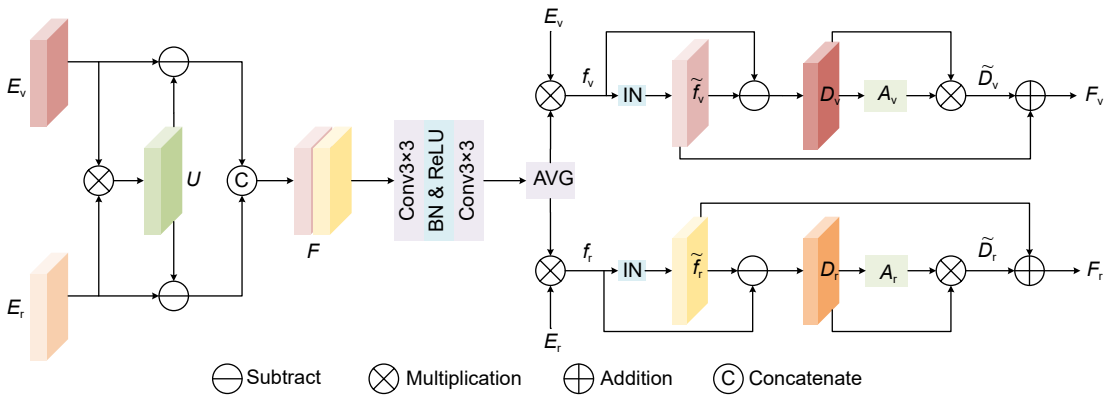


图 2　模态重加权恢复模块

Fig. 2　Modal reweighted recovery module

$$A = Conv_{3\times3}(W(Conv_{3\times3}(F)))\,, \tag{8}$$

其中：W表示在两个 3×3 卷积中间插入批归一化 (BN, batch normalization) 和ReLU激活函数的可学习参数。然后使用一个全局平均池化层 (GAP) 来聚合空间维度的模态特征，再通过 Sigmoid 函数得到模态权重，最后对相应的模态进行加权来强化模态特征。

$$w_{\rm v}, w_{\rm r} = \delta(GAP(A))\,, \tag{9}$$

$$f_{\rm v} = w_{\rm v} \times E_{\rm v}\,, \tag{10}$$

$$f_{\rm r} = w_{\rm r} \times E_{\rm r}\,, \tag{11}$$

其中：δ代表 sigmoid 函数，$f_{\rm v}$和$f_{\rm r}$为加权后的特征。

为了在模态特定特征提取阶段实现模态之间的归一化，我们对上述重加权后的特征输入到模态缓解模块 (MAM) 中使用实例正则化 (IN)[18] 来减少模态之间的差异，学习模态无关但身份相关特征。虽然直接使用 IN 可以缓解模态差异，但在这同时它也会去除一些行人的显著性信息。为了保证行人信息的完整性，我们使用通道注意力引导的 IN，从被删除的信息中学习模态无关但身份相关特征，并将其复原到模型中，从而确保行人信息的完整性与鉴别性。首先使用两个实例正则化层对$f_{\rm v}$和$f_{\rm r}$进行模态之间的归一化。对于可见光模态，实例正则化为

$$\widetilde{f}_{\rm v} = IN(f_{\rm v}) = \gamma_{\rm v}\left(\frac{f_{\rm v} - \mu(f_{\rm v})}{\sigma(f_{\rm v})}\right) + \beta_{\rm v}\,. \tag{12}$$

对于红外模态：

$$\widetilde{f}_{\rm r} = IN(f_{\rm r}) = \gamma_{\rm r}\left(\frac{f_{\rm r} - \mu(f_{\rm r})}{\sigma(f_{\rm r})}\right) + \beta_{\rm r}\,, \tag{13}$$

其中：$\gamma_*, \beta_*\ *\in\{x|{\rm v},{\rm r}\}$代表从网络中学习到的参数，$\mu(\cdot)$，$\sigma(\cdot)$为特征的均值和标准差。加权后的特征$f_{\rm v}$(或$f_{\rm r}$) 与归一化的特征$\widetilde{f}_{\rm v}$(或$\widetilde{f}_{\rm r}$) 之间的差为$D_{\rm v}$(或$D_{\rm r}$)，其表示使用 IN 层后被丢失的鉴别性模态信息，可以由如下公式表示：

$$D_{\rm v} = f_{\rm v} - \widetilde{f}_{\rm v}\,, \tag{14}$$

$$D_{\rm r} = f_{\rm r} - \widetilde{f}_{\rm r}\,. \tag{15}$$

$D_{\rm v}$和$D_{\rm r}$可以表示模态信息。RGB 和 IR 图像的主要模态差异存在于通道之间，所以为了将去除的模态信息中的行人鉴别性信息恢复到网络中，本文使用通道注意力 SENet[19] 提取模态信息中突出的行人特征同时抑制干扰信息，得到通道注意力矩阵$\boldsymbol{att}_{\rm v}$，$\boldsymbol{att}_{\rm r}$。然后通过通道注意力来提取模态无关但身份相关的特征$\widetilde{D}_{\rm v}$和$\widetilde{D}_{\rm r}$，具体计算过程如下：

$$\widetilde{D}_{\rm v} = D_{\rm v} \times \boldsymbol{att}_{\rm v}(D_{\rm v})\,, \tag{16}$$

$$\widetilde{D}_{\rm r} = D_{\rm r} \times \boldsymbol{att}_{\rm r}(D_{\rm r})\,. \tag{17}$$

最后将模态无关但身份相关的特征还原行人特征中，如下所示：

$$F_{\rm v} = \widetilde{f}_{\rm v} + \widetilde{D}_{\rm v}\,, \tag{18}$$

$$F_{\rm r} = \widetilde{f}_{\rm r} + \widetilde{D}_{\rm r}\,, \tag{19}$$

其中：$F_{\rm v}$和$F_{\rm r}$为模态重加权缓解模块的输出，将其与原始输入融合进一步优化网络。

在训练阶段为了使模型学习更多不同的模态特定特征，而不是专注于模态共享特征，因此本文不对四路特征进行融合。在测试阶段，将四路输入提取的四种独立的特征融合为可见光和红外两类。具体地，对于提取的四路特征$[F_i^{\rm V_g}, F_i^{\rm V_c}, F_i^{\rm R_g}, F_i^{\rm R_c}]$，有：

$$F_i^{\rm V_m} = \alpha \cdot F_i^{\rm V_g} \oplus (1-\alpha) \cdot F_i^{\rm V_c}\,, \tag{20}$$

$$F_i^{\rm R_m} = \alpha \cdot F_i^{\rm R_g} \oplus (1-\alpha) \cdot F_i^{\rm R_c}\,, \tag{21}$$

其中：$F_i^{\rm V_m}$，$F_i^{\rm R_m}$表示融合后的可见光和红外模态，α设置为 0.5。所以在训练时，四路特征提取器的输出为$[F_i^{\rm V_g}, F_i^{\rm V_c}, F_i^{\rm R_g}, F_i^{\rm R_c}]$，测试时的输出为$[F_i^{\rm V_m}, F_i^{\rm R_m}]$。

2.3 次显著特征互补模块 (SFCM)

注意力机制在跨模态行人重识别中广泛使用，主要用来解决一些复杂场景中的遮挡和视觉模糊等问题。尽管基于注意力机制的方法在解决行人重识别任务中取得了显著的表现，但仍然缺乏对全局行人图像的感知能力，从而忽视一些微小但对行人匹配至关重要的次显著信息，诸如衣服长度、鞋子、背包等信息。除此之外，由于数据集是在行人移动的场景中收集的，行人背景时刻在变，十分复杂，此时使用注意力机制可能会导致网络学习错误的信息，从而导致 VI-ReID 的识别能力下降。为了提高网络对于行人判别性微小信息的学习能力，增强其对整个行人的全局信息的感知，抑制干扰的背景信息，我们设计了次显著特征互补模块，如图 3 所示。

次显著特征互补模块 (SFCM) 在模型的通道维度上推导出两个互补的特征，其中一个特征专注于行人的全局显著性特征，另一个特征专注于在使用自注意力机制后被忽略的微小但具有判别性的次显著特征。具体地，首先对于 stage3 输出的特征 $\boldsymbol{F}$ 进行全局平均池化 (global average pooling, GAP)，获得全局上下文行人特征向量$\boldsymbol{F}_{\rm g}$。

为了提高模型的泛化能力，使网络能够有效地学

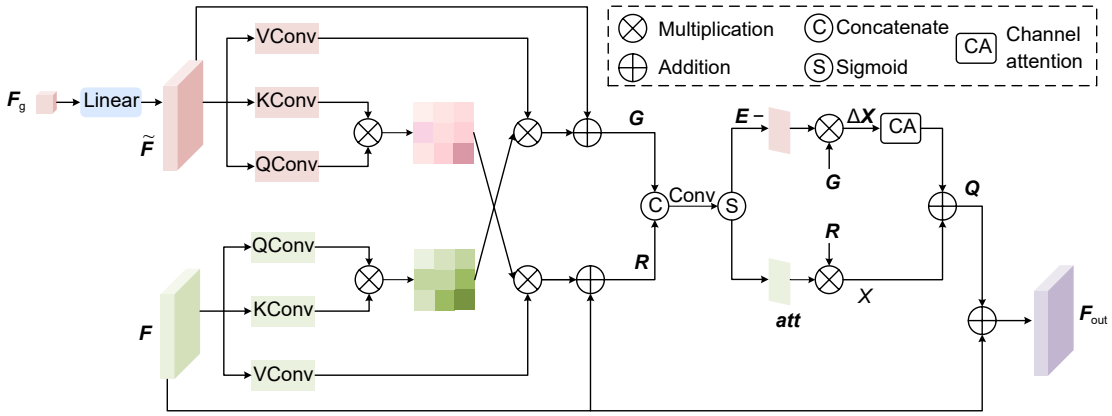


图 3 次显著特征互补模块

Fig. 3 Sub-critical features complementary module

习行人的鉴别性特征，本文对经过全局平均池化的全局特征向量$\boldsymbol{F}_{\mathrm{g}}$进行线性投影操作，然后将其维度扩展为和特征 $\boldsymbol{F}$ 相同的大小，将其记作$\widetilde{\boldsymbol{F}} \in \mathbb{R}^{C\times H\times W}$，以便进行下一步计算。其中：$C$，$H$，$W$ 分别为特征的通道数、长度和宽度。接着，进行自注意力的计算，受传统的 Non-local[20] 的启发，以原始特征 F 为例，首先将特征图输入到三个 1×1 卷积 KConn，QConv，VConv，生成三个紧凑的嵌入特征$\boldsymbol{W}_{\mathrm{k}}$，$\boldsymbol{W}_{\mathrm{q}}$，$\boldsymbol{W}_{\mathrm{v}}$。特征$\widetilde{F}$的输出为$\widetilde{\boldsymbol{W}}_{\mathrm{k}}$，$\widetilde{\boldsymbol{W}}_{\mathrm{q}}$，$\widetilde{\boldsymbol{W}}_{\mathrm{v}}$。然后通过矩阵乘法和 softmax 函数来计算相似度矩阵$\boldsymbol{M} \in \mathbb{R}^{C\times C}$和$\widetilde{\boldsymbol{M}} \in \mathbb{R}^{C\times C}$，具体计算方法如下：

$$\boldsymbol{M} = \boldsymbol{W}_{\mathrm{k}}^{\mathrm{T}} \otimes \boldsymbol{W}_{\mathrm{q}}\,, \tag{22}$$

$$\widetilde{\boldsymbol{M}} = \widetilde{\boldsymbol{W}}_{\mathrm{k}}^{\mathrm{T}} \otimes \widetilde{\boldsymbol{W}}_{\mathrm{q}}\,. \tag{23}$$

接着将特征 $\boldsymbol{F}$ 和$\widetilde{\boldsymbol{F}}$生成的相似度矩阵交换与$W_{\mathrm{v}}$和$\widetilde{\boldsymbol{W}}_{\mathrm{v}}$结合，得到强调全局特征的 $\boldsymbol{R}$ 和关注次显著特征的 $\boldsymbol{G}$，计算方法如下：

$$\boldsymbol{G} = \hat{\boldsymbol{F}} + \left(\widetilde{\boldsymbol{W}}_{\mathrm{q}} \otimes softmax\left(\boldsymbol{W}_{\mathrm{k}}^{\mathrm{T}} \otimes \boldsymbol{W}_{\mathrm{q}}\right)\right), \tag{24}$$

$$\boldsymbol{R} = \boldsymbol{F} + \left(\boldsymbol{W}_{\mathrm{v}} \otimes softmax\left(\widetilde{\boldsymbol{W}}_{\mathrm{k}}^{\mathrm{T}} \otimes \widetilde{\boldsymbol{W}}_{\mathrm{q}}\right)\right). \tag{25}$$

将 $\boldsymbol{R}$ 和 $\boldsymbol{G}$ 在通道级别连接起来，经过一个卷积层和一个 sigmoid 激活函数，得到空间注意力矩阵。然后，用矩阵 $\boldsymbol{E}$ ($\boldsymbol{E}$ 中所有元素均为 1) 减去空间注意力矩阵，得到互补的空间注意力权值矩阵。反转操作可以帮助网络关注被注意力机制忽略的微小的关键信息，还可以抑制干扰的背景信息。其中，$\boldsymbol{att}$ 为全局显著特征的空间权重矩阵，$\boldsymbol{E}$-$\boldsymbol{att}$ 为通过注意力机制忽略的次显著信息的空间权重矩阵。我们通过得到的权重矩阵指导特征 $\boldsymbol{G}$ 和特征 $\boldsymbol{R}$ 强化空间信息。在全局特征的指导下，空间权重特征图$\boldsymbol{att} \in \mathbb{R}^{1\times H\times W}$，由以下公式计算：

$$\boldsymbol{att} = \sigma(Conv([\boldsymbol{G}, \boldsymbol{R}]))\,, \tag{26}$$

其中：[·,·]表示连接操作，$Conv$表示 1*1 的卷积层，σ表示 sigmoid 激活函数。

接下来，我们通过得到的一对互补权重空间矩阵分别对全局特征 R 和其互补特征 G 进行加权操作，得到经过空间信息指导后的全局显著性特征 $\boldsymbol{X}$ 和次显著特征$\Delta\boldsymbol{X}$。计算方法如下：

$$\Delta\boldsymbol{X} = (\boldsymbol{E} - \boldsymbol{att}) \times G\,, \tag{27}$$

$$\boldsymbol{X} = \boldsymbol{att} \times R\,, \tag{28}$$

其中：×表示逐元素相乘，$\boldsymbol{X}$, $\Delta\boldsymbol{X} \in \mathbb{R}^{C\times H\times W}$。

经过空间加权后的次显著特征$\Delta\boldsymbol{X}$中缺少通道显著性特征的学习，所以我们设计了通道注意力 (CA)，如图 4 所示，其在抑制无关通道信息的同时，还能为每个通道分配重要性权重，突出通道显著性信息，帮助模型学习完整的次显著信息。具体地，$\Delta\boldsymbol{X}$作为输入信息，将经过平均池化和最大池化获得的特征在通道维度上进行连接，然后通过 MLP 操作，得到通道的权重值，接着对$\Delta\boldsymbol{X}$进行加权，具体计算方式如下：

$$CA = MLP[Avg(\Delta\boldsymbol{X}), Max(\Delta\boldsymbol{X})]\,, \tag{29}$$

$$\Delta\boldsymbol{X}_{\mathrm{c}} = CA(\Delta\boldsymbol{X})\,, \tag{30}$$

其中：[·,·]表示连接操作，Avg为自适应平均池化，Max为自适应最大池化。

接下来，为了将模型学习过程中忽略掉的次显著特征恢复到模型中，本文将全局显著性特征向量 $\boldsymbol{X}$ 和经过通道加权后的次显著特征向量$\Delta\boldsymbol{X}_{\mathrm{c}}$进行融合。从而增强模型对于行人整体特征的感知能力，提高行人的匹配效率。具体地，首先将特征 $\boldsymbol{X}$ 和$\Delta\boldsymbol{X}_{\mathrm{c}}$进行逐元

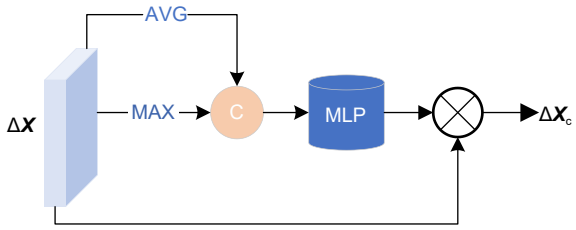


图 4 通道注意力

Fig. 4 Channel attention

素相加，将特征提取过程中忽略掉的微小但有助于特征匹配的行人信息恢复到全局特征中，推理出行人图像完整的信息。然后再通过残差连接将原始特征融合。由下列公式计算：

$$\boldsymbol{Q}=\boldsymbol{X}+\Delta\boldsymbol{X}_{\mathrm{c}}\,, \tag{31}$$

$$\boldsymbol{F}_{\mathrm{out}}=\boldsymbol{Q}+\boldsymbol{F}\,, \tag{32}$$

其中：+表示逐元素相加操作。通过融合操作输出的特征中既包含了行人的显著性信息，又结合了次显著特征，增加了行人识别的可能性。

2.4 多损失联合优化

为了保证模型性能，我们在模型输出特征之后使用基准模型的身份损失L_{id}[16]和加权正则化三元组损失L_{wrt}[16]联合优化模型。

身份损失将属于同一身份的不同模态图像视为同一类，可以在特征空间缩小类内距离，在本文中使用交叉熵函数，具体形式如下：

$$L_{\mathrm{id}}(x_i)=y_i\log(p_i)+(1-y_i)\log(1-p_i)\,, \tag{33}$$

式中：x_i为样本集中的一个样本，y_i为x_i对应的标签值，p_i为使用分类器预测x_i属于该样本身份的概率。

加权正则化三元组损失 L_{wrt} 从模态内和跨模态两个方面优化正样本对，负样本对和锚点之间的三重态关系和相对距离。表示如下：

$$L_{\mathrm{wrt}}(i)=\log(1+\exp(\sum_{j}w_{ij}^{p}d_{ij}^{p}-\sum_{ik}w_{ik}^{n}d_{ik}^{n}))\,, \tag{34}$$

$$w_{ij}^{p}=\frac{\exp(d_{ij}^{p})}{\sum_{d_{ij}^{p}\in P_i}\exp(d_{ij}^{p})},w_{ik}^{n}=\frac{\exp(-d_{ik}^{n})}{\sum_{d_{ik}^{n}\in N_i}\exp(-d_{ik}^{n})}\,, \tag{35}$$

式中：(L,f,k)表示训练批次中的一个三元组。d_{ij}表示样本之间的欧式距离。对于锚点 i，P_i是对应的正集合，N_i 是负集合。上述加权正则化继承了正负样本对之间相对距离优化的优点，但它避免了引入任何额外的边缘参数。

总的损失函数为

$$L=L_{\mathrm{id}}+L_{\mathrm{wrt}}\,. \tag{36}$$

3 实验结果与分析

3.1 1 数据集与评价指标

所提方法主要在SYSU -MM01[21]数据集和RegDB[22]数据集上进行，同时使用 rank-k 识别率和 mAP 平均精度均值作为模型的评价指标。

SYSU-MM01 是最大的 VI-ReID 数据集，该数据集包含 395 个不同身份的行人，其中训练集包括由 4 台摄像机获取的 22258 张可见光图像和 2 台摄像机获取的 11909 张近红外图像。测试集包含 96 个行人。该数据集的测试模式包含全搜索模式 (all search) 和室内搜索模式 (indoor search)，其中全搜索模式更具挑战性。

RegDB 数据集包括 412 个不同身份的行人，每个行人有 10 个可见光图像和 10 个远红外图像。根据现有的 VI-ReID 设置，随机抽取 206 个身份进行训练，其余 206 个身份用于测试。训练/测试分别过程重复十次。通过将查询设置为可见光图像 (查询) 更改为红外图像 (图库) 来评估性能。

本实验模型是在 PyTorch 库中采用 python 3.8 深度学习框架实现，并在单个 NVIDIA RTX A4000 进行训练。本文选择 AGW[16]作为基准网络。输入图像大小调整为 3×384×192。在训练阶段的每个批次中，模型随机从 6 个行人身份中选择 4 张 VIS 图像和 4 张 IR 图像。该模型使用 SGD 优化器，动量为 0.9，在训练过程中，初始学习率为 0.01，10 次训练后使用预热策略将学习率增加到 0.1，在第 20 个 epoch 时学习率衰减到 0.01，在第 90 个 epoch 时学习率进一步衰减到 0.001，本模型一共迭代 100 次。

3.2 实验结果对比

本节将所提方法与现有的可见光-红外行人重识别领域的主流方法在两个公共数据集上做比较，最终结果如表 1 和表 2 所示。

在 SYSU-MM01 数据集上的全搜索和室内搜索模式的结果如表 1 所示，可以看出，本文方法的各项指标均有一定的提升。在全搜索模式中，QFCNet 的 rank-1 准确率和 mAP 值分别达到了 76.12%, 71.51%，较次优的 DSCNet 分别提高了 2.23% 和 2.04%。在室内搜索模式中，其 rank-1 准确率和 mAP 值分别达到了 80.43%，83.61%，比 DSCNet 分别提高了 1.08% 和 0.96%。表中“-”为原论文未提及的指标。

RegDB 数据集中的结果如表 2 所示。在可见光

表 1 SYSU-MM01 数据集比较结果/%

Table 1 Comparison results on SYSU-MM01 dataset/%

Method	Publish	Setting							
		All search				Indoor search			
		rank-1	rank-10	rank-20	mAP	rank-1	rank-10	rank-20	mAP
AlignGAN[23]	ICCV19	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3
D^2RL[24]	CVPR 19	28.90	70.60	82.40	29.20	-	-	-	-
X-Modality[15]	AAAI 20	49.9	89.8	96	50.7	-	-	-	-
DDAG[8]	ECCV 20	53.61	89.17	95.3	52.02	58.37	91.92	97.42	65.44
Hi-CMD[9]	CVPR 20	34.9	77.6	-	35.9	-	-	-	-
JSIA-ReID[25]	AAAI20	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9
MPANet[5]	CVPR 21	70.6	96.2	98.8	68.2	76.2	97.2	99.3	76.9
AGW[16]	TPAMI 21	47.58	84.45	92.11	47.69	54.29	91.14	95.99	63.02
CAJ[17]	CVPR 21	69.9	95.7	98.5	66.9	76.3	97.9	99.5	80.4
MCLNet[26]	ICCV 21	65.4	93.3	97.1	62.0	72.6	97.0	99.2	76.6
DSCNet[7]	TIFS 22	73.89	96.27	98.84	69.47	79.35	98.32	**99.77**	82.65
FMCNet[9]	CVPR 22	66.3	-	-	62.5	68.2	-	-	74.1
PIC[27]	TIP 22	57.5	-	-	55.1	60.4	-	-	67.7
PMT[28]	AAAI 23	67.53	95.36	98.64	51.86	71.66	96.73	99.25	76.52
MTMFE[29]	PR23	62.56	93.85	97.63	60.57	65.06	95.17	98.17	73.86
AGMNet[30]	J-STSP23	69.63	96.27	98.82	66.11	74.68	97.51	99.14	78.30
CSVI[31]	IF24	70.13	96.15	98.79	65.32	71.00	96.96	98.99	75.21
TMD[32]	TMM24	68.18	93.08	96.84	63.96	76.31	97.28	98.91	74.52
QFCNet	Ours	**76.12**	**97.23**	**99.14**	**71.51**	**80.43**	**98.75**	99.58	**83.61**

表 2 RegDB 数据集比较结果/%

Table 2 Comparison results on RegDB dataset/%

Method	Publish	Setting							
		Visible to infrared				Infrared to visible			
		rank-1	rank-10	rank-20	mAP	rank-1	rank-10	rank-20	mAP
AlignGAN[23]	ICCV19	57.9	-	-	53.6	56.3	-	-	53.40
D^2RL[24]	CVPR 19	43.40	66.10	76.30	44.10	-	-	-	-
X-Modality[15]	AAAI 20	62.21	83.13	91.72	60.18	-	-	-	-
DDAG[8]	ECCV 20	69.34	85.77	89.98	63.19	64.77	83.85	88.90	58.54
Hi-CMD[9]	CVPR 20	70.9	86.4	-	66.0	-	-	-	-
JSIA-ReID[25]	AAAI20	48.1	-	-	48.9	48.5	-	-	49.3
MPANet[5]	CVPR 21	83.7	-	-	80.9	82.8	-	-	80.7
AGW[16]	TPAMI 21	70.05	86.21	91.55	66.37	70.49	87.21	91.84	65.9
CAJ[17]	CVPR 21	84.72	95.17	97.38	78.70	84.09	94.79	97.11	77.25
MCLNet[26]	ICCV 21	80.31	92.70	96.03	73.07	75.93	90.93	94.59	69.49
DSCNet[7]	TIFS 22	85.39	-	-	77.3	83.5	-	-	75.19
FMCNet[9]	CVPR 22	89.12	-	-	84.43	88.38	-	-	83.86
PIC[27]	TIP 22	83.6	-	-	79.6	79.5	-	-	77.4
PMT[28]	PR23	76.10	88.86	92.41	74.39	72.18	87.06	92.38	71.04
MTMFE[29]	AAAI 23	84.83	-	-	76.55	84.16	-	-	75.13
AGMNet[30]	J-STSP23	88.40	95.10	96.94	81.45	85.34	94.56	97.48	81.19
CSVI[31]	IF24	91.41	97.72	98.92	85.14	90.06	97.46	98.74	83.86
TMD[32]	TMM24	87.04	95.49	97.57	81.19	83.54	94.56	96.84	77.92
QFCNet	Ours	**93.28**	**97.94**	**99.04**	**88.89**	**92.35**	**97.72**	**99.14**	**88.10**

到红外 (visible to infrared) 的检索模式下，QFCNet 的 rank-1 准确率和 mAP 值达到了 93.28%、88.89%；比次优的 FMCNet 分别提高了 4.16% 和 4.46%。在红外到可见光 (infrared to visible) 的检索模式下，QFCNet 的 rank-1 准确率和 mAP 值分别达到了 92.35%、88.10% 。

3.3 消融实验

为了评估 QFCNet 每个模块的有效性，我们在 SYSU-MM01 数据集的两种模式下进行了一系列的实验。第一行为基准网络，其由 ResNet-50 主干网络和损失函数组成。本文在基准网络的基础上，逐个加入各模块进行训练。实验结果如表 3 所示，表中数据表明在基准网络上添加每一个模块，模型的精度都有一定的提升，这也验证了所提模块在网络中均是不可或缺的。最后将 QSFM 和 SFCM 一起整合加入基准网络，得到最优结果，用加粗数据表示，进一步说明该模型的优越性。

3.4 SFCM 插入位置有效性探究

次显著特征互补模块通过提取被注意力机制忽视的次显著信息，以补充行人的全局关键特征。为了探究在 Resnet-50 网络的 stage0 至 stage4 后插入次显著特征互补模块的有效性，本文在 SYSU-MM01 的单镜头全搜索模式下进行了一系列实验。实验结果如表 4 所示。以下实验的参数设置相同，评价指标为 rank-1 和 mAP。

根据表 4 中的实验结果，我们发现在 stage0-stage3 之后插入 SFCM 模块，模型性能都有所提升，这是因为次显著特征互补模块将注意力机制忽略的次显著特征恢复到模型中，挖掘了充分的图像显著特征，为行人匹配提供了更多可用信息。在低级别网络中插入 SFCM 模块，虽然性能有所提升，但低级别网络学习到的特征包含大量嘈杂信息，不利于高级别的行人语义匹配。然而在 stage4 后插入 SFCM 模块，性能较插入在 stage2 和 satge3 后的性能有所下降，这是由于高级别的网络虽然能提取高级别语义信息，但是随着网络的逐渐深入，参数量会急剧增加，模型复杂度提高，所以训练和匹配效率随之下降。观察表中的实验结果，在 stage3 之后插入 SFCM 模块的性能最优，因此将 SFCM 模块放入 stage3 之后。

3.5 可视化分析

3.5.1 模态差异分析

为了验证所提方法可以减小模态差异，我们在 SYSU-MM01 数据集中随机选取 20 个对象在二维空间中使用 t-SNE[33] 方法进行特征分布可视化，可视化结果如图 5(a, b) 所示。其中三角形元素和圆形元素分别代表可见光模态和红外模态，同一个颜色的所有元

表 3　SYSU-MM01 数据集上的消融实验/%

Table 3　Ablation experiments on the SYSU-MM01 dataset/%

Setting		All search				Indoor search			
QSFM	SFCM	rank-1	rank-10	rank-20	mAP	rank-1	rank-10	rank-20	mAP
		47.58	84.45	92.11	47.69	54.29	91.14	95.99	63.02
√		73.21	96.60	98.90	70.35	79.63	98.20	99.62	82.81
	√	72.91	96.46	99.06	69.84	79.88	98.17	99.43	82.90
√	√	**76.12**	**97.23**	**99.14**	**71.51**	**80.43**	**98.75**	**99.58**	**83.61**

表 4　SFCM 插入位置在 SYSU-MM01 数据集下的实验结果/%

Table 4　Experimental results of SFCM insertion positions under SYSU-MM01 dataset/%

Method	SYSU-MM01	
	rank-1	mAP
Baseline	47.58	47.69
在stage0后插入SFCM	70.35	67.42
在stage1后插入SFCM	71.75	67.66
在stage2后插入SFCM	72.60	69.68
在stage3后插入 SFCM	**72.91**	**69.84**
在stage4后插入SFCM	70.65	68.78

素代表的都是同一个行人。图 5(a) 为基准网络的特征分布，图 5(b) 为本文方法的特征分布。观察图 5 可以发现本文方法在图中的聚类效果更明显，这说明我们所提方法可以有效地聚合和区分同一个人的嵌入特征，减少模态间的差异。图 5(c, d) 为基准网络和本文网络的类内类间距离。图中蓝色代表类内分布，绿色代表类间分布。图中垂线为类内和类间距离的平均值。基准网络的类内类间距离为δ_0，本文方法的类内类间距离为δ_1，$\delta_0 < \delta_1$，这表明与基准网络相比，QFCNet 的类内距离明显减小。因此，QFCNet 可以有效地减少可见光图像和红外图像之间的模态差异。

3.5.2 热力图可视化

为了验证本文所提方法的有效性，我们在 SYSU-MM01 数据集随机选取 5 个行人的可见光图像和红外图像并使用 GradCAM[34] 进行热力图可视化。可视化结果如图 6 所示。其中，图 6(a) 为随机选取的行人图像，图 6(b) 为本文方法的热力图，图 6(c) 则为基准网络的热力图，图中每列为同一张图片。热力图代表网络的注意力分配，趋近于 H，热力图颜色越暖越鲜艳，说明模型对其关注度高，该部分特征对于最终的匹配更关键，而热力图颜色趋近 L，则说明该部分网络的关注度较低。通过观察图 6(c) 可以发现，基准网络的热力图主要关注行人的局部信息，而且当图像背景较为复杂时，热力图会错误地关注到背景信息，这对于行人匹配造成了影响。而观察图 6(b) 可以发现本文方法克服了背景信息的干扰，使网络在关注全局特征的同时考虑到次显著性信息，如 person1，person2 红外图像中的背带信息、person3 的裙子长短信息和 person4，person5 的衣服图案信息。虽然红外图像缺少一些模态信息，但是从热力图来看，本文方法在红外模态下依然可以关注行人的显著性区域，这对于行人匹配至关重要。综上所述，我们设计的 QFCNet 可以从数据集中提取较多利于行人匹配的显著性信息，可以进一步提高模型识别效率。

3.5.3 可视化排序

为了进一步展示 QFCNet 的有效性，我们在 SYSU-MM01 数据集上随机选取 10 个行人图片，左边一列查询图像为 IR 模态，在 VIS 模态中检索匹配的行人图像，右边一列则相反。图 7 展示了 rank-10 检索结果。其中，对于每个检索示例，绿色框表示与

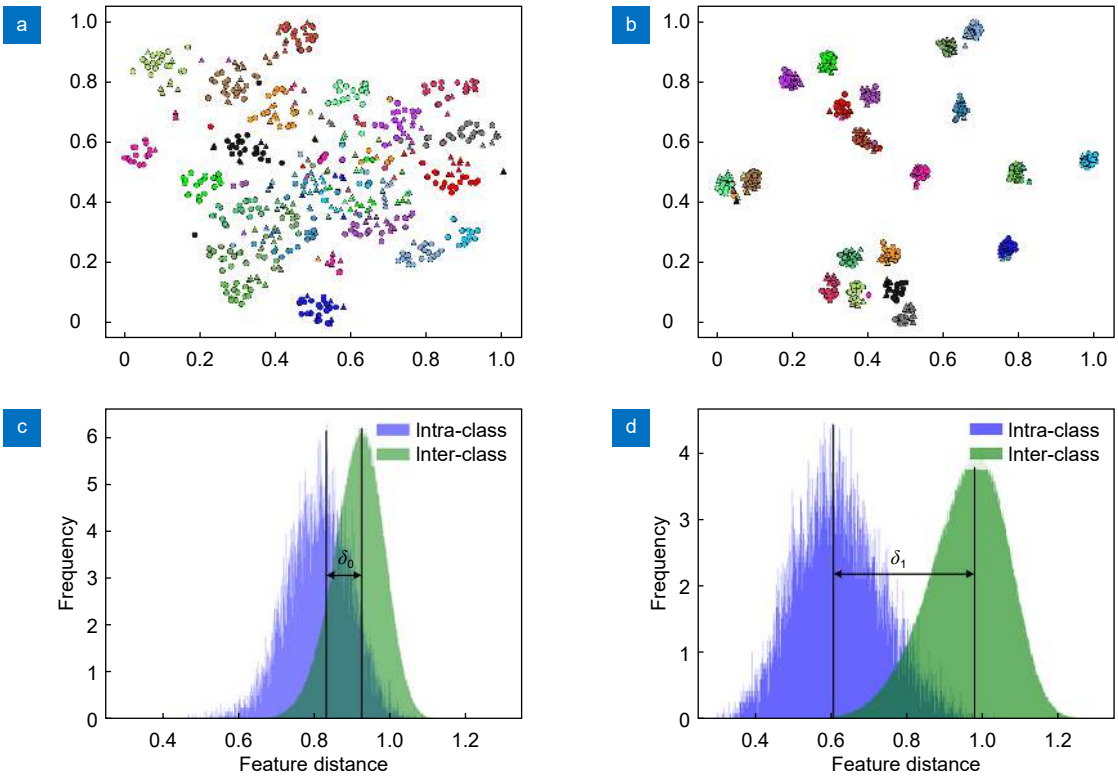


图 5 特征分布图和类内类间距离。(a) 基准特征分布图；(b) QFCNet 特征分布图；(c) Baseline 类内类间距离；(d) QFCNet 类内类间距离

Fig. 5 Feature distribution diagram and intra-class and inter-class distances. (a) Baseline feature distribution diagram; (b) QFCNet feature distribution diagram; (c) Baseline intra-class and inter-class distances; (d) QFCNet intra-class and inter-class distances

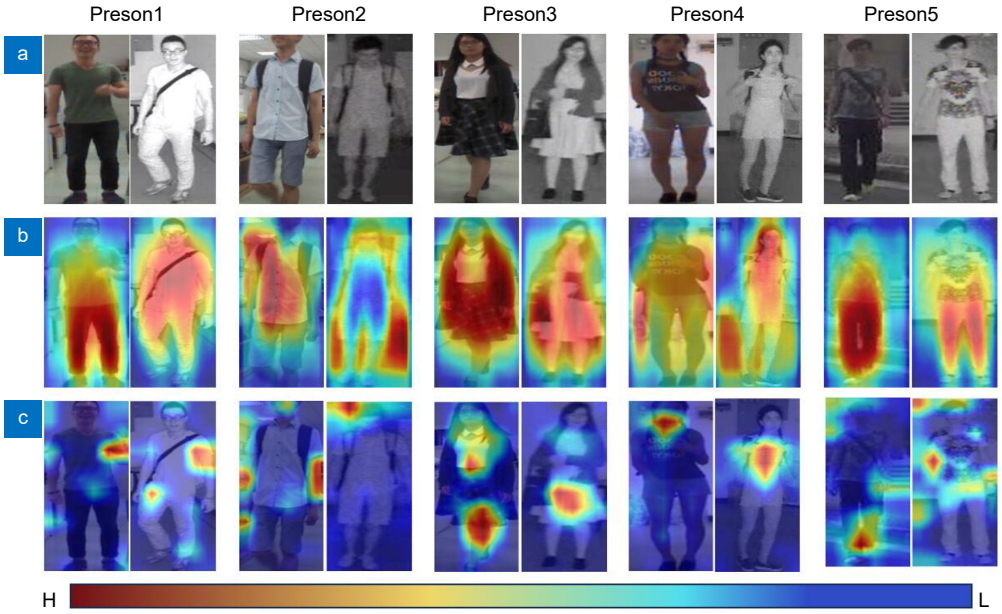


图 6 热力图可视化

Fig. 6 Visualization of the heat map

图 7 SYSU-MM01 可视化排序结果

Fig. 7 Visual sorting results on SYSU-MM01

给定的查询对象匹配正确，而红色框表示匹配有误。从查询的结果可以看出，QFCNet 可以使匹配正确的图像排在前几位，有效地提高排名结果。此外，对于一些有微小特征的图像 (如：背包、衣服图案等信息)，在红外模态下进行匹配时，仍然有较好的性能。

4 结 论

本文提出了一种四流输入引导的特征互补可见光-红外行人重识别网络 (QFCNet)，用来提取两个模态丰富的行人共享显著性特征表达，缓解模态差异。该网络主要包括四流特征提取和融合模块 (QSFM) 和模态次显著特征互补模块 (SFCM)。四流特定特征提取和融合模块将数据增强后的数据集与原始数据一起作为网络的输入，在通道维度减少模态差异，丰富特征的语义信息。模态次显著特征互补模块学习行人判别性微小信息，增强网络对整个行人的全局信息的感知能力。本文方法在两个公开数据集上的实验结果表明了所提方法的优越性。

利益冲突: 所有作者声明无利益冲突

参考文献

[1] 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.

[2] 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.

[3] Cheng S Y, Chen Y. Camera-aware unsupervised person re-identification method guided by pseudo-label refinement[J]. *Opto-Electron Eng*, 2023, **50**(12): 230239.
程思雨, 陈莹. 伪标签细化引导的相机感知无监督行人重识别方法[J]. 光电工程, 2023, **50**(12): 230239.

[4] 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.
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.

[5] Zhang Y K, Wang H Z. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification[C]//*Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 2023: 2153–2162. https://doi.org/10.1109/CVPR52729.2023.00214.

[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*, Nashville, 2021: 4330–4339. https://doi.org/10.1109/CVPR46437.2021.00431.

[7] Zhang Y Y, Kang Y H, Zhao S Y, et al. Dual-semantic consistency learning for visible-infrared person re-identification[J]. *IEEE Trans Inf Forensics Secur*, 2022, **18**: 1554−1565.

[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*, Glasgow, 2020: 229–247. https://doi.org/10.1007/978-3-030-58520-4_14.

[9] Choi S, Lee S, Kim Y, et al. Hi-CMD: hierarchical cross-modality disentanglement for visible-infrared person re-identification[C]//*Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 2020: 10257–10266. https://doi.org/10.1109/CVPR42600.2020.01027.

[10] Zhang Y K, Yan Y, Lu Y, et al. Towards a unified middle modality learning for visible-infrared person re-identification[C]//*New York*: *Proceedings of the 29th ACM International Conference on Multimedia*, 2021: 788–796. https://doi.org/10.1145/3474085.3475250.

[11] Ma L, Guan Z B, Dai X G, et al. A cross-modality person re-identification method based on joint middle modality and representation learning[J]. *Electronics*, 2023, **12**(12): 2687.

[12] Ye M, Shen J B, Shao L. Visible-infrared person re-identification via homogeneous augmented tri-modal learning[J]. *IEEE Trans Inf Forensics Secur*, 2021, **16**: 728−739.

[13] Zhang Q, Lai C Z, Liu J N, et al. FMCNet: feature-level modality compensation for visible-infrared person re-identification[C]//*Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 2022: 7349–7358. https://doi.org/10.1109/CVPR52688.2022.00720.

[14] Lu J, Zhang S S, Chen M D, et al. Cross-modality person re-identification based on intermediate modal generation[J]. *Opt Lasers Eng*, 2024, **177**: 108117.

[15] 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*, New York, 2020: 4610–4617. https://doi.org/10.1609/aaai.v34i04.5891.

[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, 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*, Montreal, 2021: 13567–13576. https://doi.org/10.1109/ICCV48922.2021.01331.

[18] Pan X G, Luo P, Shi J P, et al. Two at once: enhancing learning and generalization capacities via IBN-net[C]//*Proceedings of the 15th European Conference on Computer Vision*, Munich, 2018: 464–479. https://doi.org/10.1007/978-3-030-01225-0_29.

[19] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018: 7132–7141. https://doi.org/10.1109/CVPR.2018.00745.

[20] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//*Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018: 7794–7803. https://doi.org/10.1109/CVPR.2018.00813.

[21] 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*, Venice, 2017: 5380–5389. https://doi.org/10.1109/ICCV.2017.575.

[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] 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*, Seoul, 2019: 3623–3632. https://doi.org/10.1109/ICCV.2019.00372.

[24] 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*, Long Beach, 2019: 618–626. https://doi.org/10.1109/CVPR.2019.00071.

[25] Wang G A, Zhang T Z, Yang Y, et al. Cross-modality paired-images generation for RGB-infrared person re-identification[C]//*Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, 2020: 12144–12151. https://doi.org/10.1609/aaai.v34i07.6894.

[26] 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, Montreal, 2021: 16403–16412. https://doi.org/10.1109/ICCV48922.2021.01609.

[27] 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.

[28] 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*, Washington, 2023: 1835–1843. https://doi.org/10.1609/aaai.v37i2.25273.

[29] 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 Recognit*, 2023, **135**: 109145.

[30] Liu H J, Xia D X, Jiang W. Towards homogeneous modality learning and multi-granularity information exploration for visible-infrared person re-identification[J]. *IEEE J Sel Top Signal Process*, 2023, **17**(3): 545−559.

[31] Huang N C, Xing B C, Zhang Q, et al. Co-segmentation assisted cross-modality person re-identification[J]. *Inf Fusion*, 2024, **104**: 102194.

[32] Lu Z F, Lin R H, Hu H F. Tri-level modality-information disentanglement for visible-infrared person re-identification[J]. *IEEE Trans Multimedia*, 2024, **26**: 2700−2714.

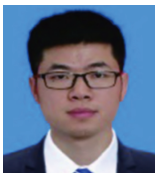
[33] van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. *J Mach Learn Res*, 2008, **9**(86): 2579−2605.

[34] 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

【通信作者】葛斌 (1973-)，男，博士，教授，研究方向为图像处理、信息安全。

E-mail:bge@aust.edu.cn



夏晨星 (1991-)，男，博士，副教授，研究方向为图像处理、计算机视觉。

E-mail：cxxia@aust.edu.cn

许诺 (2000-)，女，硕士研究生，研究方向为图像处理、计算机视觉。

E-mail：551168574@qq.com

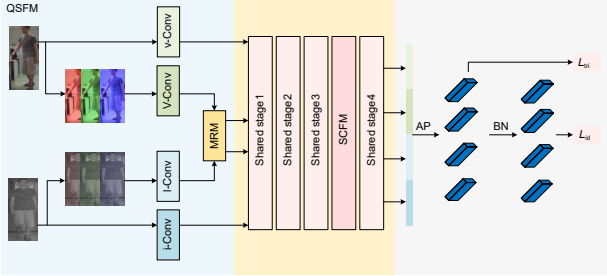
郑海君 (1998-)，男，硕士研究生，研究方向为图像处理、计算机视觉。

E-mail：navy626@163.com

扫描二维码，获取PDF全文

Quadrupl-stream input-guided feature complementary visible-infrared person re-identification

Ge Bin[1,2*], Xu Nuo[1], Xia Chenxing[1], Zheng Haijun[1]



QFCNet structure diagram

Overview: Cross-modal person re-identification is the task of identifying individuals from images of different modalities under non-overlapping camera angles, which has a wide range of practical applications. Different from the previous VV-ReID(visible-visible person reidentification), VI-ReID(visible-infrared person reidentification) aims at image matching between visible and infrared modalities. Due to the imaging differences between visible and infrared cameras, there are huge modal differences between cross-modal images, and traditional person re-identification methods are difficult to apply to this scenario. In view of this situation, it is particularly important to study the pedestrian matching between visible and infrared images. How to realize the mutual recognition between visible and infrared pedestrian images efficiently and accurately has a very great practical value for improving the level of social management, preventing crime, maintaining national security, and so on. Similarly, cross-modal person re-identification technology also involves many challenges. Not only intra-modal variations such as viewpoint, pose, and low resolution need to be considered, but also inter-modal differences caused by different image channel information need to be addressed. Existing VI-ReID methods mainly focus on two aspects: (1) solving cross-modal problems by maximizing modal invariance; (2) Generate intermediate or target images, and transform the cross-modal matching problem into an intra-modal matching task. The first method makes it difficult to guarantee the quality of the modal invariant features, which leads to the loss of indirect information in the image representation of people. The second method inevitably introduces noise, which affects the stability of training and makes the quality of generated images difficult to guarantee. Current visible-infrared person re-identification research focuses on extracting modal shared saliency features through the attention mechanism to minimize modal differences. However, these methods only focus on the most salient features of pedestrians, and cannot make full use of modal information. To solve this problem, this paper proposes a quadrupl-stream input-guided feature complementary method based on deep learning, which can effectively alleviate the differences between modalities while retaining useful structural information. Firstly, a quadrupl-stream feature extraction and fusion module is designed in the mode-specific feature extraction stage. By adding two data enhancement inputs, the semantic information of the modalities is enriched and the multi-dimensional feature fusion is further promoted. Secondly, a sub-salient feature complementation module is designed to supplement the pedestrian detail information ignored by the attention mechanism in the global feature through the inversion operation. The experimental results on two public datasets SYSU-MM01 and RegDB show the superiority of this method. In the full search mode of SYSU-MM01, the rank-1 and mAP values reach 76.12% and 71.51%, respectively.

Ge B, Xu N, Xia C X, et al. Quadrupl-stream input-guided feature complementary visible-infrared person re-identification[J]. *Opto-Electron Eng*, 2024, **51**(9): 240119; DOI: 10.12086/oee.2024.240119

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

[1]School 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