基于双注意力机制的可见光-红外行人重识别

魏克铭[1]，韩星宇[2]，王辉[2]，范自柱[1,2]

（1.华东交通大学理学院，江西 南昌 330013；2.华东交通大学江西省先进控制与优化重点实验室，江西 南昌 330013）

摘　要：近些年可见光-红外行人重识别受到众多学者的关注，其目标是从不同模态的图像中匹配出相同身份的行人图像，但由于可见光图像和红外图像之间的巨大差异，导致可见光-红外行人重识别是一项非常具有挑战性的图像检索问题。现有的研究主要集中在通过设计网络结构提取共享特征或生成中间模态来缓解模态差异，容易受到行人之外的区域影响。为了解决此类问题，重点关注行人信息，进一步减小两种模态之间的差异，提出一种双注意力机制的网络结构用于可见光-红外行人重识别，一方面通过双注意力机制挖掘不同尺度的行人空间信息和增强局部特征的通道交互能力，另一方面利用全局分支和局部分支，学习多粒度的特征信息，使不同粒度信息可以相互补充，形成一个更具辨别性的特征。在两个公共数据集上的实验结果表明，该方法相较于基线有明显的提升，在 RegDB 数据集和 SYSU-MM01 数据集上均表现出理想的性能。

关键词：可见光-红外行人重识别；注意力机制；缓解模态差异

中图分类号：TP391　　**文献标志码：**A

Base on Dual Attention Mechanism for Visible-Infrared Person Re-Identification

Wei Keming[1], Han Xingyu[2], Wang Hui[2], Fan Zizhu[1,2]

(1. School of Science, East China Jiaotong University, Nanchang 330013, China；2. Key Laboratory of Advanced Control and Optimization of Jiangxi Province, East China Jiaotong University, Nanchang 330013, China)

Abstract: In recent years, visible-infrared person re-identification has attracted the attention of many scholars, and its goal is to match person images with the same identity from images of different modalities. Due to the huge difference between visible images and infrared images, visible-infrared person re-identification is a very challenging image retrieval problem. Existing research focuses on mitigating modal differences by designing network structures to extract shared features or generate intermediate modalities, which are susceptible to areas other than person. In order to solve such problems, focus on person information, and further reduce the difference between the two modes, a network structure of dual attention mechanism is proposed for visible-infrared person re-identification, on the one hand, through the dual attention mechanism to mine person spatial information of different scales and enhance the channel interaction ability of local features. On the other hand, the use of global branches and local branches, learn multi-granular feature information, so that different granular information can complement each other to form a more discriminating feature. Experimental results on two public datasets show that the proposed method has a significant improvement compared with the baseline, and shows ideal performance on both the RegDB dataset and the SYSU-MM01 dataset.

Key words: visible-infrared person re-identification; dual attention mechanism; mitigate modal differences

收稿日期：2023−05-10

基金项目：国家自然科学基金项目(61991401)；教育部人文社会科学研究项目(22YJCZH168)；江西省自然科学基金项目(20224BAB212014)

行人重识别主要任务是在多个不重叠的摄像机视图中匹配特定的人，在安全领域有着不可或缺的作用，近年来行人重识别一直受到广泛的关注。它的挑战主要集中在视图、姿态、光照、遮挡、背景变化等方面，为了解决这些问题，众多学者提取出了许多解决方法，取得不错的效果。这些方法主要集中在单模态的可见光行人重识别问题上，但在实际应用中，往往需要捕捉不同场景下的行人图像，特别是在夜晚光照极弱的情况下，可见光相机很难捕捉到有效的行人信息，因此可见光-红外行人重识别就引起了众多学者的注意。该领域主要研究可见光图像和红外图像之间的跨模态度量问题，以从不同模态的图像中匹配出相同的行人图像，目的是克服在复杂环境下传统行人重识别的局限性。如图 1 所示，红外图像相比于可见光图像，信息量更少、视觉效果模糊、分辨率差、对比度低，所以更难提取到有效的特征信息，常规的单模态行人重识别也不能够发挥同等的效用。

不同模态间的巨大差异导致可见光-红外行人重识别是具有挑战性的，针对模态差异，众多学者提出了一系列解决方法[1-10]。为了缓解像素级的模态差异，一些方法通常利用生成对抗网络，设计复杂的生成对抗模型[7-10]，以获得对应模态的图像，但由于红外到可见光变换的不适定性，导致生成的图像难以优化，而且不可避免地会引入噪声数据。另一方面，为了减轻特征级的模态差异，一些方法采用单流、双流或多流网络[3-6]，通过设计不同的损失函数、注意力机制等提取不同模态共享特征。然而，基于这些学习方法训练通用的网络模型，缺乏对特异性模态信息的关注度，导致关键信息丢失。

为了充分利用有价值的行人信息，减小模态间的差异，从以下几个方面出发来解决此问题：首先数据集规模有限，缺乏多样性，如果仅学习全局特征，容易导致信息丢失，而不同粒度的特征可以更有效的提取行人信息，因此引用基于部分的卷积神经网络模型(Part-based Convolutional Baseline, PCB)[11]处理不同层次的细粒度行人特征，并且保留了全局特征，在提取特征之后作为一个单独的分支进行优化，以全局优化行人特征。其次，为了进一步减小背景、光线等噪声数据的影响，受到 CCNet[12]和空洞卷积的启发引入多尺度交叉注意力机制(Multi-scale Cross Attention, MCA)，同时利用不同尺度的空洞卷积和最大池化，扩大模块的感受野，关注更多的边缘信息。最后考虑到不同通道之间的信息交互和不同层次行人特征之间的差异性，提出局部通道交互注意力机制(Part Channel-interaction Attention, PCA)，在兼顾局部特征的同时，增强不同通道间的特征交互能力。

本文的主要贡献可以总结为以下几点：

1）提出多尺度交叉注意力机制 MCA，扩大特征的感受野,关注更多的行人边缘信息；

2）提出局部通道交互注意力机制 PCA，一方面可以更好的提取局部特征，另一方面增强不同特征块间的通道交互能力；

3）同时利用全局特征和局部特征，在数据集 RegDB 和 SYSU-MM01 上均取得最优的效果。

图 1　SYSU-MM01 数据集中可见光图像和红外图像样本

Fig. 1　Visible and infrared image samples from the SYSU-MM01 dataset

1　相关工作

单模态行人重识别旨匹配不同可见光摄像机拍摄到的行人，所有的图像都来自相同的模态，而且在各大公开的单模态数据集上都取得了理想的效果。然而在实际生活场景中，由于周围环境、光照、遮挡等因素的影响，导致可见光摄像机并不能很好的捕捉行人图像，从而大大的限制了单模态行人重识别，因此近些年可见光-红外行人重识别引起众多学者的注意。另外由于不同模态的图像之间存在巨大差异，导致单模态行人重识别方法应用于可见光-红外行人重识别效果不佳，当然也有一些通用的方法，比如PCB将提取到的特征均匀分块，获取更多细粒度的信息，大大提升了模型的鲁棒性，但仍需要不断探索新的解决方法。此类行人重识别目前的方法大致可以分为两大类：1）利用通用的网络结构，通过设计损失函数、注意力机制等，提取共享特征信息；2）利用生成对抗网络或轻量级的生成器等，作模态转换，生成不同模态的图像以减少模态差异。

可见光-红外行人重识别最早是由Wu等人[13]在2017年提出，利用能够自动扩展域特定结构的深层零填充算法提取共享特征,且提供了第一个可见光-红外行人重识别数据集SYSU-MM01。Ye等人[3]提出度量学习的概念，联合优化模态特异性和模态共享矩阵来实现。将两种模态转换到一致的空间，学习共享特征。为了进一步提升性能，Ye等人[4]提出动态双注意力机制，包括类内加权注意和跨模态图注意。这些方法主要集中在扩大行人特征的类间差异上，很少有研究调查如何提高类内跨模态特征相似性。因此Zhu等人[5]提出异质中心损失（HC loss），

约束两个模态的类内中心距离，以减少类内跨模态变化。当样本存在一些异常值时，可能会破坏三元

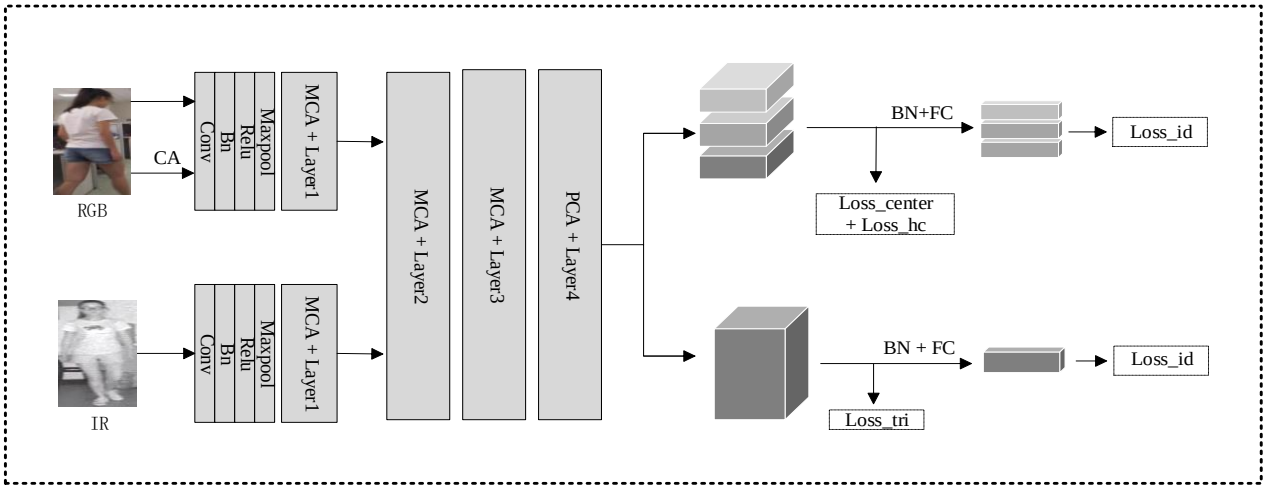

图 2　模型框架

Fig. 2　Model framework

损失已经习得的距离，因此Liu等人[6]在部分行人特征框架下，提出异质中心三元损失函数，适当放宽约束条件。但是提取模态共享特征信息会不可避免地丢失行人身份的判别性信息。另外，诸如眼镜、鞋子和衣服长度等细微但具有区别性的信息尚未得到充分探索，特别是在红外模式下，因此Wu[14]等人提出了一种联合模态和模式对齐网络，以发现不同模态的细微差别。

基于生成对抗网络的可见光-红外行人重识别可以有效拟合两种模态差异，Wang等人[7]提出了一种端到端对齐生成对抗网络，通过特征对齐和像素对齐共同弥补模态差异；Hi-CMD[8]生成网络通过生成一个新的具有不同姿势和照明的跨模态图像来学习解纠缠的表示，同时保留一个人的身份；Zhang等人[9]提出了一种新颖的特征级模态补偿网络(FMCNet)，从已有模态共享特征中直接生成缺失的模态特征；另外一些方法[10,15]通过设计轻量级的生成器生成中间模态，以缓解模态差异带来的影响。

然而，提取模态共享特征会造成特征信息损失，生成新的模态会引入额外噪声数据，而且计算量比较大，对设备要求较高。

2　方法

为了减小模态差异、背景噪声影响、增强模型的鲁棒性，基于随机通道交换[16]，提出了一种兼顾局部与全局特征的双注意力机制的网络结构用于可见光-红外行人重识别，在这一部分主要介绍网络结构的模型框架，其整体结构如图 2 所示。主要包括以下几个部分组成：(1)由 Resnet50 组成的双流骨干网络；(2)多尺度交叉注意力机制(MCA)；(3)局部通道交互注意力机制(PCA)；(4)全局特征分支及其对应损失函数；(5)局部特征分支及对应其损失函数。在测试阶段，使用全局特征分支的输出结果进行预测,局部特征分支仅在训练过程中发挥效用。

假设 $V=\{x_v^i\}_{i=1}^{N_v}$ 和 $T=\{x_t^i\}_{i=1}^{N_t}$ 分别表示数据集中的可见光图像和红外图像，其中 N_v 和 N_t 定义为相应模态的图片数量，数据集共有 $N=N_v+N_t$ 张图片，对应真实标签集 $Y=\{y_i\}_{i=1}^{N_p}$，其中 N_p 表示行人身份的数量，y_i 表示第 i 个行人标签。

2.1　模型框架

双流网络是用于可见光红外行人重识别特征提取的典型方法，而且它的有效性在众多文献中得到了有力的证明。本文利用随机通道交换、擦除[16]等做数据增强，采用在 ImageNet 上预训练的 ResNet50[17]作为骨干网络提取特征，为了保证不同模态特征的特异性，网络在第二个残差块前不共享参数；为避免噪声的影响，同时引入两个注意力机制，注意力模块在保持特征图身份识别能力的同时，减轻模态差异以及背景噪声的影响。其次，为了同时获得全局特征和局部特征，学习到细微的、具有鉴别性的特征，本文引入 PCB 特征分块机制，旨在学习不同行人图像之间的细微差别。同时将部分特征与全局特征相结合，以达到更好的效果。

2.2　多尺度交叉注意力机制(MCA)

受到 CCNet 和多尺度特征融合[18]的启发，本文提出了多尺度交叉注意力机制(MCA)：考虑到最大池化可以加强网络对显著性区域的关注度，去除背景冗余信息，但容易丢失空间分辨率，因此引入空洞卷积弥补池化的不足。同时考虑到利用不同尺度最大池化和空洞卷积可以扩大感受野，关注行人的边缘特征，获取多尺度的上下文信息，从而增强像素级的表征能力。具体流程如图 3 左侧所示，图 3 右侧展示了 MCA 中最大池化模块和空洞卷积模块的具体结构。

给定特征 $x \in R^{C\times W\times H}$，该模块首先在 x 上应用两个具有1×1滤波器的卷积层，分别生成两个特征映射 Q 和 K，其中 $\{Q,K\} \in R^{C'\times W\times H}$。$C'$ 为通道数，由于降维，通道数小于 C。然后 Q 和 K 分别经过一个多尺度最大池化块，再经过1×1滤波器的卷积

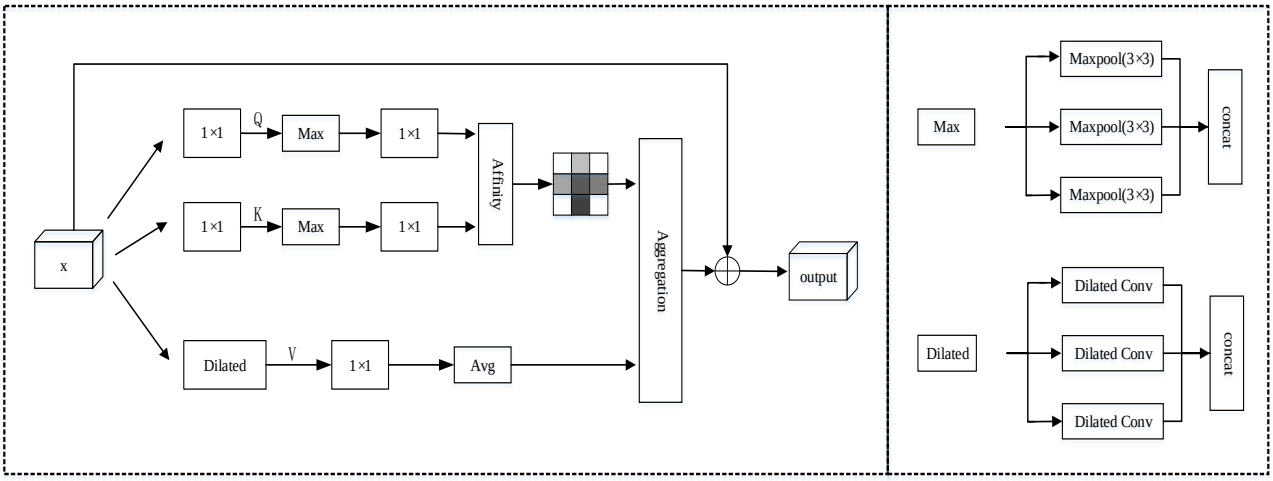

图 3　多尺度交叉注意力机制(MCA)

Fig. 3　Multi-scale Cross Attention(MCA)

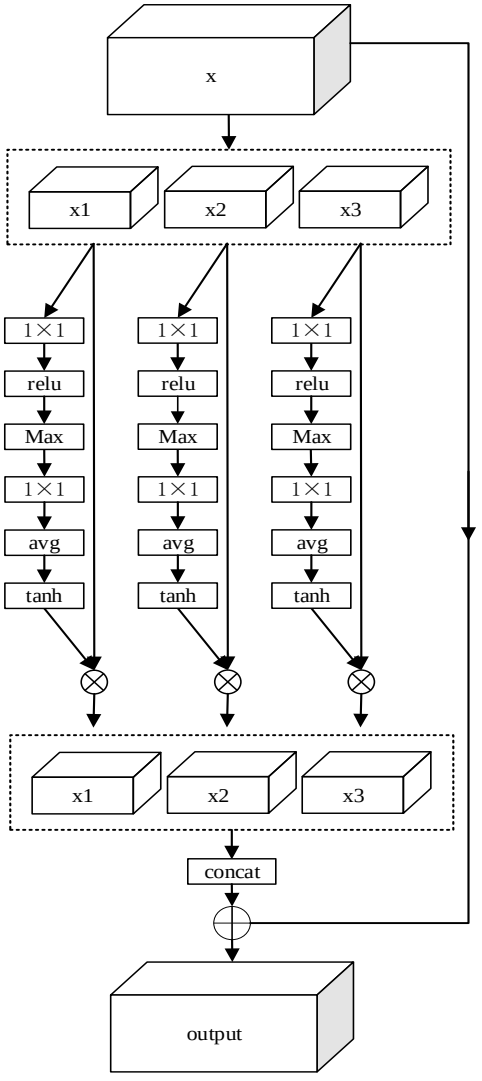

图 4　局部通道交互注意力机制(PCA)

Fig. 4　Part Channel-interaction Attention(PCA)

层得到 Q_1 和 K_1，用数学公式表示如下：

$$Q_1 = conv_2^1(\max(conv_1^1(Q))) \tag{1}$$

$$K_1 = conv_2^2(\max(conv_1^2(K))) \tag{2}$$

另外给定的输入特征 $x \in R^{C\times W\times H}$ 分别经过不同大小滤波器的空洞卷积块得到 V，而后输入1×1滤波器的卷积块和平均池化块得到 V_1：

$$V_1 = avg(conv(Dilated(x))) \tag{3}$$

Q_1 和 K_1 经过仿射变换后，与 V_1 进行聚合变换，最终与输入特征 x 求和：

$$output = x + Agg(Aff(Q_1, K_1), V_1) \tag{4}$$

其中 Aff 表示仿射变换，Agg 表示聚合变换。

2.3　局部通道注意力机制(PCA)

不同通道的特征图受到的关注度理应是不相同的，且特征图的不同层次也是如此，因此为了增强不同层次之间的通道交互能力，引入了局部通道交互注意力机制(PCA)，如图 4 所示：首先给定一个输入特征 $x \in R^{C\times W\times H}$，$x$ 在水平方向被均匀地分割成若干块，得到 $x_i, i=1,2,3$，然后分别经过两个1×1卷积块，进行缩放变换，随后对拼接的特征利用正切激活函数激活，具体过程用公式表示如下：

$$x_i' = \max(relu(conv_1^i(x_i))), i=1,2,3 \tag{5}$$

$$x_i' = \tanh(avg(conv_1^i(x_i'))), i=1,2,3 \tag{6}$$

$$x' = concat(x_1', x_2', x_3') \tag{7}$$

$$output = x + \tanh(x') \tag{8}$$

而且本文对水平分层也做了充分的实验，说明水平方向分三层在此框架下可以达到最优的效果。

2.4 损失函数

在本节中，主要介绍本文的框架中使用的损失函数，包括交叉熵损失、三元损失、聚类中心损失和中心三元损失。利用交叉熵损失和三元损失结合起来监督全局特征，利用交叉熵损失、聚类中心损失和中心三元损失作为局部分支的学习目标。

交叉熵损失函数：交叉熵损失的目标是提取特定行人身份的信息进行分类。此方法被广泛应用于行人重识别任务中，以促进模型对样本进行有效的分类。在本文中，依旧采用交叉熵损失分别优化全局特征和局部特征，以捕获每个行人不同模态的身份鉴别信息。交叉熵损失的表达式如下：

$$L_{id} = -\sum_{i=1}^{N} \log \frac{e^{W_{y_i}^T x_i}}{\sum_{k=1}^{P} e^{W_k^T x_i}} \tag{9}$$

其中 W_k 为第 k 类的权重向量，y_i 为特征 x_i 的真实身份标签，N 为批次大小，P 为训练集中的类数。

三元损失函数：对于全局特征，利用三元损失优化不同模态下不同行人图像的特征，它可以拉近不同模态相同身份的行人特征间的距离，扩大不同身份的行人间的距离，本文沿用 Ye 等人在 HCML[6] 中提出的三元损失，公式定义如下：

$$L_{tri} = \sum_{i=1}^{P} \sum_{a=1}^{K} [\rho + \max \left\| x_a^i - x_p^i \right\|_2 - \min_{i \neq j} \left\| x_a^i - x_n^j \right\|_2]_+ \tag{10}$$

其中 K 表示模态数量，P 表示行人的数量，x_a 表示锚点样本，x_p 表示正样本对，x_n 表示负样本对，ρ 是阀值参数，用以约束正负样本间的距离。

中心三元损失函数：三元损失通过锚点与所有其他样本的比较计算损失。但由于图像本身存在的一些噪声，造成局部特征可能与全局特征有很大的差异，如果存在一些异常值，可能会过于严格地约束成对距离，三元损失将不能很好的优化类内类间距离。因此，利用中心三元损失函数优化局部特征的类内与类间距离，采用每个身份的中心作为身份代理，将锚点与所有其它样本的比较替换为锚点中心与所有其它中心的比较， 具体计算公式如下：

$$\begin{aligned} L_{center} = & \sum_{i=1}^{P} [\rho + \left\| c_v^i - c_t^i \right\|_2 - \min_{i \neq j} \left\| c_v^i - c_n^j \right\|_2]_+ \\ & + \sum_{i=1}^{P} [\rho + \left\| c_t^i - c_v^i \right\|_2 - \min_{i \neq j} \left\| c_t^i - c_n^j \right\|_2]_+ \\ & + \sum_{i=1}^{P} [\rho + \left\| c_x^i - c_t^i \right\|_2 - \min_{i \neq j} \left\| c_x^i - c_n^j \right\|_2]_+ \end{aligned} \tag{11}$$

其中 c_v^i 表示可见光图像特征的聚类中心，c_t^i 表示红外图像的聚类中心，c_x^i 表示随机通道交换后得到的模态的聚类中心，c_n^i 则表示其它模态聚类中心，P 表示行人身份数量。

聚类中心损失函数：聚类中心损失通过惩罚不同模态分布的中心，优化不同模态的类内相似度，公式定义如下：

$$L_{hc} = \sum_{i=1}^{P} \left\| c_v^i - c_t^i \right\|_2 \tag{12}$$

其中 c_v^i 和 c_t^i 分别表示可见光和红外图像的聚类中心。特征聚类中心的计算方法如下：

$$c_v^i = \frac{1}{P} \sum_{i}^{P} x_v^i \tag{13}$$

$$c_t^i = \frac{1}{P} \sum_{i}^{P} x_t^i \tag{14}$$

其中 P 表示可见光或红外图像行人身份的数量。

均方差损失函数(MSE)：为进一步缩小相同身份不同模态的行人图像之间的距离，简单地应用均方差损失进行约束，公式如下：

$$L_{mse} = \sum_{i=1}^{P} \left\| x_v^i - x_t^i \right\|_2 \tag{15}$$

其中 P 表示行人的数量。

综上所述，损失函数分为全局损失和局部损失，全局损失函数定义如下：

$$L_{global} = L_{id} + L_{tri} + L_{mse} \tag{16}$$

局部损失函数定义如下：

$$L_{local} = \lambda_1 L_{id} + \lambda_2 L_{center} + \lambda_3 L_{hc} \tag{17}$$

总损失函数定义如下：

$$L = L_1 + \alpha L_2 \tag{18}$$

其中 $\alpha, \lambda_1, \lambda_2, \lambda_3$ 是超参数，用以平衡各个损失函数之间的权重。

表 1　在 RegDB 数据集上与现有的方法作对比(%)

Tab. 1　Comparison with existing methods on the RegDB dataset (%)

		Visible to infrared					Infrared to visible				
Method	Venue	Rank1	Rank10	Rank20	mAP	mINP	Rank1	Rank10	Rank20	mAP	mINP
Zero-Pad[13]	ICCV17	17.75	34.21	44.35	18.90	-	16.63	34.68	44.25	17.82	-
HCML[6]	AAAI18	24.44	47.53	56.78	20.08	-	21.70	45.02	55.58	22.24	-
HSME[23]	AAAI19	50.85	73.36	81.66	47.00	-	50.15	72.40	81.07	46.16	-
AlignGan[10]	ICCV19	57.9	-	-	53.6	-	56.30	-	-	53.40	-
X-Modal[15]	AAAI20	62.21	83.13	91.72	60.18	-	-	-	-	-	-
DDAG[3]	ECCV20	69.34	86.19	91.49	63.46	49.24	68.06	85.15	90.31	61.80	48.62
AGW[24]	TPAMI21	70.05	86.21	91.55	66.37	50.19	70.49	87.21	91.84	65.90	51.24
MCLNet[25]	ICCV21	80.31			79.83	-	73.43			69.49	-
CAJ[16]	ICCV21	85.03	95.49	97.54	79.14	65.33	84.75	95.33	97.51	77.82	61.56
MPANet[14]	CVPR21	83.7	-	-	80.9	-	82.8	-	-	80.7	-
CLMC[26]	TNNLS21	91.84	97.86	98.98	81.42	-	91.12	97.86	98.69	81.06	-
MMN[27]	ACM MM21	91.60	97.70	98.90	84.10	-	87.50	96.00	98.10	80.50	-
MID[28]	AAAI22	87.45	95.73	-	84.85	-	84.29	93.44	-	81.41	-
SPOT[29]	TIP22	80.35	93.48	96.44	72.46	56.19	79.37	92.79	96.01	72.26	56.06
MSCLNet[30]	ECCV22	84.17	-	-	80.99	-	83.86	-	-	78.31	-
MAUM[31]	CVPR22	87.87	-	-	85.09	-	86.95	-	-	84.34	
Ours	-	**95.22**	**98.54**	**99.17**	**87.70**	**74.48**	**93.67**	**98.35**	**98.83**	**86.43**	**71.68**

表 2　在 SYSU-MM01 数据集上与现有方法作对比(%)

Tab. 2　Comparison with existing methods on the SYSU-MM01 dataset (%)

		All search					Indoor search				
Method	Venue	Rank1	Rank10	Rank20	mAP	mINP	Rank1	Rank10	Rank20	mAP	mINP
Zero-Pad[13]	ICCV17	14.80	54.12	71.33	15.95	-	20.58	68.38	85.79	26.92	-
HCML[6]	AAAI18	14.32	53.16	69.17	16.16	-	24.52	73.25	86.73	30.08	-
HSME[23]	AAAI19	20.68	32.74	77.95	23.12	-	-	-	-	-	-
AlignGan[10]	ICCV19	42.4	85.00	93.70	40.7	-	45.9	87.60	94.40	54.3	-
X-Modal[15]	AAAI20	49.9	89.8	96.0	50.7	-	-			-	-
DDAG[3]	ECCV20	54.75	90.39	95.81	53.02	39.62	61.02	94.06	98.41	67.98	62.61
AGW[24]	TPAMI21	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
MCLNet[25]	ICCV21	65.40	93.33	97.14	61.98	-	72.56	96.88	99.20	76.58	-
CAJ[16]	ICCV21	69.88	95.71	98.46	66.89	53.61	76.26	97.88	99.49	80.37	76.69
MPANet[14]	CVPR21	70.58	96.21	98.80	68.24	-	76.74	98.21	99.57	80.95	-
CLMC[26]	TNNLS21	64.37	93.90	97.53	63.43	-	67.35	98.10	99.77	74.02	-
MMN[27]	ACM MM21	70.61	96.2	99.0	66.9	-	76.2	97.2	99.3	79.6	-
MID[28]	AAAI22	60.27	92.90	-	59.40	-	64.86	96.12	-	70.12	-
SPOT[29]	TIP22	65.34	92.73	97.04	62.25	48.86	69.42	96.22	99.12	74.63	70.48
CMT[30]	ECCV22	71.88	96.45	98.87	68.57	-	76.9	97.68	99.64	79.91	-
MAUM[31]	CVPR22	71.68	-	-	68.79	-	76.97	-	-	81.94	-
Ours	-	**74.18**	**96.58**	**99.20**	**70.04**	**56.97**	**79.69**	**98.12**	**99.73**	**83.08**	**79.65**

3　实验

3.1　数据集

为了评估本文提出方法的有效性，以正确图像排在候选列表前 k 个位置的概率(Rank-k)、平均准确率（mAP）和平均逆置负样本惩罚率（mINP）作为评价指标，在两个公开的数据集（SYSU-MM01 和 RegDB）上做了充分的实验。

SYSU-MM01 数据集[13]由 4 个可见光相机和 2 个红外摄像机在室内和室外拍摄而成，涉及 491 个身份。其中训练集包括 22 258 张可见光图像和 11 909 张红外图像，涉及 395 个身份。测试集包含 3 803 张用于被检索红外图像和 301 张用于检索的可见光图像，共 96 个身份。此数据集包含全局搜索和室内搜索两种测试模式。

RegDB 数据集[33]由一对对齐的可见光和热成像相机拍摄而成，包括 412 的身份的 4 120 张图片，每个身份对应 10 张可见光图像和 10 张热成像图像。此数据集被随机划分为两部分，206 个身份用于训练，其余 206 个身份用于测试。训练和测试均需基于数据集的随机划分重复进行十次实验。

3.2　参数设置

采用双注意力机制增强的双流网络，引入 PCB 模块，以 Resnet50 作为骨干提取特征，共享后三个残差块的参数。采用随机梯度下降(SGD)优化器进行训练。训练阶段，所有的可见光和红外图像的大小调整为 288×144，通过随机通道交换、擦除、翻转增强数据

集。初始学习率设置为 0.1，在前 10 个训练周期采用预热策略[26]，在第 20 个训练周期衰减为 0.01，在第 50 个训练周期衰减为 0.001。训练周期总数设置为 100。在每一个训练批次中，随机抽取 4 个行人，其中每个行人分别抽取 4 张可见光图像和 4 张红外图像，共 32 张行人图像。超参数的取值区间在$[0,1]$，根据实验结果的优劣，不断微调参数值，以取得更好的实验效果，最终总损失函数 L 的参数在 RegDB 数据集上分别设置为$\alpha=1,\lambda_1=1,\lambda_2=0.6,\lambda_3=0.6$，在 SYSU-MM01 数据集上设置$\alpha=1,\lambda_1=0.5,\lambda_2=0.1,\lambda_3=2$。

3.3 对比现有方法

在这一小节，对比了近些年来提出的可见光-红外行人重识别方法，表 1 和表 2 分别展示了在 RegDB 数据集和 SYSU-MM01 数据集上与不同方法比较的结果。在 RegDB 数据集中，可见光到红外模式下达到了 95.22%的 Rank-1，87.70%的 mAP 和 74.48%的 mINP；红外到可见光模式下达到了 93.67%的 Rank-1，86.43%的 mAP 和 71.68%的 mINP。在 SYSU-MM01 数据集中，全局搜索模式下达到了 74.18%的 Rank-1，70.04%的 mAP 和 56.97%的 mINP；室内搜索模式下达到了 79.69%的 Rank-1，83.08%的 mAP 和 79.65%的 mINP。基于所有这些评估和比较的结果，可以确认本文方法的优越性及有效性。

3.4 消融实验

在这一部分，以 RegDB 数据集为例，评估模型的有效性。

不同模块的效果：首先从模型中删除两个注意力模块、PCB 模块及其对应的损失函数，以此作为基线方法‘Base’进行比较。‘P’表示采用 PCB 模块及其对应的损失函数；‘MCA’表示采用注意力模块 MCA；‘PCA’表示采用注意力模块 PCA。具体结果如表 3 所示，从中可以清晰地看出，在采用‘MCA’时效果比‘Base’增加 2.77，比‘Base+P’增加 1.02；在采用‘PCA’时效果比‘Base’增加 2.72，比‘Base+P’增加 0.68；同时加上两个注意力机制时效果比‘Base’增加 4.56，比‘Base+P’增加 1.87。

层次划分的效果：不同的层次划分数量决定了行人局部特征通道交互的粒度。以 RegDB 数据集为例，图 5 展示了不同层次划分的效果，其中横轴 *Part* 表示分块的数量。为了保证 *Part* 的有效性，避免出现垂直方向上特征无法均分导致信息丢失的情况，所以 *Part* 在此可以取 1，2，3，6，其中 *Part* =1表示不做分块。可以观察到 *Part* =3是分层提取局部行人特征的最佳设置。

表 3 各组成部分对模型性能的影响

Tab. 3 The impact of each component on model performance

Method	RegDB visible to infrared		
	Rank1	mAP	mINP
Base	87.72	79.14	64.09
Base + MCA	90.49	81.27	66.15
Base + PCA	90.44	82.34	68.16
Base + P	93.35	84.86	70.23
Base + MCA + PCA	92.28	82.40	67.10
Base + P + MCA	94.37	85.92	72.81
Base + P + PCA	94.03	85.96	72.33
Base + P + MCA + PCA	95.22	87.70	74.48

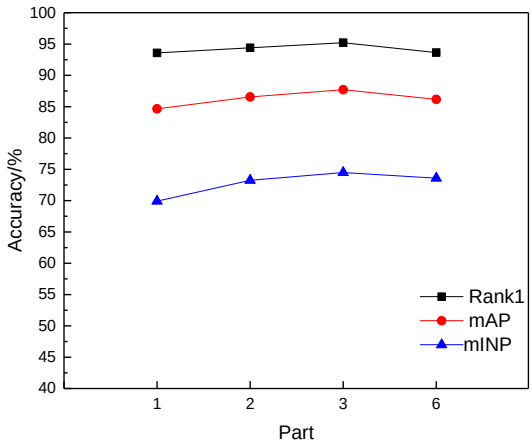

图 5 不同层次划分的效果

Fig. 5 Effects divided into different levels

利用全局特征进行预测的有效性：为了验证仅使用全局特征分支进行预测的效果，在此以 RegDB 数据集为例，对比利用全局特征和局部特征的预测结果，其中‘Local’表示仅利用局部特征进行预测，‘Global’表示仅利用全局特征进行预测，‘Global＋Local’表示联合使用全局特征和局部特征进行预测。具体结果如表 4 所示，从中可以观察到利用全局特征分支进行预测的效果最好。

表 4 不同特征进行预测的影响

Tab. 4 The impact of predictions by different features

Method	RegDB		
	visible to infrared		
	Rank1	mAP	mINP
Global	95.22	87.70	74.48
Local	94.39	87.09	74.78
Global + Local	94.73	87.51	74.45

4 结论

基于注意力机制，本文提出了一种端到端可见光-红外行人重识别模型。重点针对可见光图像和红外图像之间的模态差异，提出两种注意力机制，能够有效的提取判别性行人特征，主要结论如下。

1）提出多尺度交叉注意力机制 MCA，结合不同尺度的最大池化和空洞卷积，扩大感受野，获取多尺度的上下文信息。

2）提出局部通道交互注意力机制 PCA，增强了局部特征的通道交互能力，对不同背景和遮挡等噪声具有更强的鲁棒性。

3)通过设计网络结构，结合全局特征和局部特征，在 RegDB 数据集和 SYSU-MM01 数据集上均取得最优的效果。

参考文献：

[1] WANG X, HAN X, HUANG W, et al. Multi-similarity loss with general pair weighting for deep metric learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5022-5030.

[2] FU C, HU Y, WU X, et al. CM-NAS: Rethinking cross-modality neural architectures for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11823-11832.

[3] YE M, LAN X, LI J, et al. Hierarchical discriminative learning for visible thermal person re-identification [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018: 7501-7508.

[4] YE M, SHEN J, CRANDALL D J, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]//Proceedings of European Conference on Computer Vision. Glasgow, Britain: Springer, 2020: 229-247.

[5] ZHU Y, YANG Z, WANG L, et al. Hetero-center loss for cross-modality person re-identification[J]. Neurocomputing, 2020, 386: 97-109.

[6] LIU H, TAN X, ZHOU X. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification[J]. IEEE Transactions on Multimedia, 2020, 23: 4414-4425.

[7] WANG G, ZHANG T, CHENG J, et al. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment[C]//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea: IEEE Press, 2019: 3623-3632.

[8] CHOI S, LEE S, 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. 2020: 10257-10266.

[9] ZHANG Q, LAI C, LIU J, et al. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 7349-7358.

[10] WEI Z, YANG X, WANG N, et al. Syncretic modality collaborative learning for visible infrared person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 225-234.

[11] SUN Y, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 480-496.

[12] HUANG Z, WANG X, HUANG L, et al. Ccnet: Criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 603-612.

[13] WU A C, ZHENG W S, YU H X, et al. RGB-Infrared Cross-Modality Person Re-identification[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 5390-5399.

[14] WU, Q, DAI P, CHEN J, et al. Discover cross-modality nuances for visible-infrared person re-identification[C]. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 4330–4339.

[15] LI D, WEI X, HONG X, et al. Infrared-visible cross-modal person re-identification with an x modality[C]. Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020: 4610–4617.

[16] YE M, RUAN W, DU B, et al. Channel augmented joint learning for visible-infrared recognition[C]// International Conference on Computer Vision. 2021.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[18] 张泓，范自柱，石林瑞，等. 一种基于多尺度特征融合的人头计数检测方法研究[J]. 华东交通大学学报，2021，38（2）：115-121.

ZHANG H，FAN Z Z，SHI L R，et al. A head detection method based on multi-scale feature fusion[J]. Journal of East China Jiaotong University，2021，38（2）：115-121.

[19] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[20] CHEN Y, WAN L, LI Z, et al. Neural feature search for rgb-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 587–597.

[21] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.

[22] PARK H, LEE S, LEE J, et al. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12046-12055.

[23] HAO Y, WANG N, LI J, et al. Hsme: hypersphere manifold embedding for visible thermal person re-identification[C]//Proceedings of the 33th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2019: 8385–8392.
[24] YE M, SHEN J, LIN G, et al. Deep learning for person re-identification: A survey and outlook[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 44(6): 2872-2893.
[25] HAO X, ZHAO S, YE M, et al. Cross-modality person re-identification via modality confusion and center aggregation[C]//Proceedings of the IEEE/CVF International conference on computer vision. 2021: 16403-16412.
[26] ZHANG L, DU G, LIU F, et al. Global-local multiple granularity learning for cross-modality visible-infrared person re-identification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021.
[27] ZHANG Y, YAN Y, LU Y, et al. Towards a unified middle modality learning for visible-infrared person re-identification[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 788-796.
[28] HUANG Z, LIU J, LI L, et al. Modality-adaptive mixup and invariant decomposition for RGB-infrared person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(1): 1034-1042.
[29] CHEN C, YE M, QI M, et al. Structure-aware positional transformer for visible-infrared person re-identification[J]. IEEE Transactions on Image Processing, 2022, 31: 2352-2364.
[30] JIANG K, ZHANG T, LIU X, et al.Cross-Modality Transformer for Visible-Infrared Person Re-Identification[C]//Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV. Cham: Springer Nature Switzerland, 2022: 480-496.
[31] LIU J, SUN Y, ZHU F, et al. Learning memory-augmented unidirectional metrics for cross-modality person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 19366-19375.
[32] LUO H, GU Y, LIAO X, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//Proc- eedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2019: 0-0.
[33] 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.

第一作者：魏克铭(1998－)，男，硕士研究生，研究方向为深度学习、模式识别。
E-mail：wkmqyr@163.com。

通信作者：范自柱(1975－)，男，博士，教授，博士生导师，研究方向为模式识别、机器学习。
E-mail: zzfan3@163.com。