第 60 卷 第 14 期/2023 年 7 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于通道重组和注意力机制的跨模态行人重识别

霍东东，杜海顺*

河南大学人工智能学院，河南 郑州 450046

摘要 近年来，跨模态行人重识别逐渐成为了计算机视觉领域的热门研究方向之一。然而，在跨模态行人重识别任务中，高效地提取行人特征，进一步实现图像之间的交互融合、挖掘行人图像之间的潜在关系是至关重要的。为了解决这一问题，提出一种基于通道分组重组和注意力机制的双流网络来提取两种模态之间更加稳定且丰富的特征。具体地：首先在主干网络中嵌入模态内特征通道分组重组模块以提取跨模态图像的共享特征，实现模态信息的交互融合；然后，通过聚合特征注意力机制及跨模态自适应图结构来挖掘不同模态行人图像之间的潜在关系，提取更具判别力的局部特征。在主流数据集SYSU-MM01、RegDB上进行的大量实验结果表明，所提算法在多个数据集上具有较好的泛化能力，与现有的主要算法相比，跨模态行人重识别精度达到较高的水准。

关键词 图像处理；跨模态；行人重识别；通道分组重组；注意力机制

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP221850

Cross-Modal Person Re-Identification Based on Channel Reorganization and Attention Mechanism

Huo Dongdong, Du Haishun*

School of Artificial Intelligence, Henan University, Zhengzhou 450046, *Henan, China*

Abstract In recent years, cross-modal pedestrian re-identification has gradually become one of the hotspots in the field of computer vision. However, it is crucial to effectively extract pedestrian features, further realize the interactive fusion of photos, and mine any potential relationships between pedestrian images while performing cross-modal pedestrian re-identification. To address this issue, a dual stream network based on channel grouping reorganization and attention mechanisms is proposed to extract more stable and rich features between the two modes. Specifically, to extract the shared characteristics of cross-modal images and to achieve the interactive fusion of modal information, the intra-modal feature channel grouping rearrangement module (ICGR) was inserted in the backbone network. Furthermore, to extract additional distinct local features, the possible association between pedestrian images captured using various modes was mined using the aggregated feature attention mechanism and cross-modal adaptive graph structure. A large number of experimental results on mainstream datasets such as SYSU-MM01 and RegDB demonstrate that the proposed algorithm has good generalization ability on multiple datasets. The cross-modal pedestrian re-identification algorithm achieves higher accuracy compared with the existing main algorithms.

Key words image processing; cross-modal; person re-identification; channel grouping reorganization; attention mechanism

1 引 言

行人重识别是指给定某监控场景下的特定行人图像，利用计算机视觉与机器学习等技术来检索跨摄像头或跨时间域下的同一身份行人图像[1]。随着时代的发展，人们的安全意识水平随之提高，特别是近年来，城市监控网络不断完善，行人重识别技术被广泛应用于智能视频监控领域[2]。因此，行人重识别技术引起了众多学者们的关注并成为了计算机视觉领域热门的研究方向之一。

然而，在实际的监控系统中，特别是在光照不足的条件下，摄像机通常需要从可见光模式切换到红外模式来捕获行人的有效外观信息[3-5]。由于广泛的应用需求，基于可见光图像和红外图像的跨模态行人重识

收稿日期：2022-06-15；**修回日期**：2022-08-02；**录用日期**：2022-08-29；**网络首发日期**：2022-09-10

基金项目：河南省自然科学基金(202300410093)

通信作者：*jddhs@henu.edu.cn

别应运而生，并成为近年来业内的一个重要关注点，其任务是对可见光状态下和红外状态下的行人进行匹配[6]。

相对于单一模态行人重识别，跨模态行人重识别除了要面对行人图像因角度变化、姿势变化等因素造成的外观差异挑战，还要关注可见光图像和红外图像成像过程中所产生的模态差异[7]。上述并存的差异使得可见光红外跨模态行人重识别更具有挑战性。最初，Wu等[8]提出一种基于深度零填充的方法将可见光红外两种模态以参数共享的方式进行训练从而完成跨模态行人重识别。后来，多数方法利用双流网络结构来学习共享特征。此外，伴随着生成对抗网络（GAN）的快速发展，CycleGAN、PNGAN、FDGAN等方法相继应用于可见光图像与红外图像跨模态行人重识别中[9-11]。然而，现有的学习方法主要关注如何尽可能地缩小两种异质模态之间的差异，对于各个模态共享信息的挖掘和利用还不够充分，难以快速有效地获取行人特征。

常用的提取特征方式通过卷积操作提取特征。在传统卷积中，每一个输出通道都与输入通道相连接，通道之间采用稠密连接，该方法计算复杂度高且易生成多余的参数。Ye等[12]采用传统卷积操作，虽然识别率能够达到一定的标准，但是卷积计算效率未有进一步的提升。分组卷积可以有效提高模型训练效率且拥有较强特征表示能力，然而分组卷积中不同组的特征图之间无信息交流，降低了网络的特征提取能力。此外，现有的卷积操作多在空间上进行特征融合，很少关注通道上的特征融合，忽略了通道上特征融合的重要性。

针对以上问题，本文提出一种基于通道重组和注意力机制的双流网络模型（DCA-Net）。DCA-Net不仅能够实现图像之间的交互融合，提高网络效率，还能够进一步挖掘行人身体不同部位之间的联系，获取更加丰富的行人特征信息。具体地：首先设计模态内特征通道分组重组模块（ICGR）以实现图像之间的交互融合，提升模型提取行人图像不同模态之间共享特征的能力，提高模型训练效率；然后采用聚合特征注意力机制（AFA）挖掘每个模态中行人身体不同位置之间的联系，加强模型对行人信息的挖掘能力；最后使用跨模态自适应图结构（CGSA），结合跨模态两种模式的结构关系来加强特征表示。所提网络能够提取行人两个模态之间的共享特征，挖掘不同模态行人图像之间的潜在关系，加强模型对行人图像信息的挖掘能力。实验结果表明，所提DCA-Net在SYSU-MM01和RegDB两个常用基准数据库上的性能表现优于大多数先进的跨模态行人重识别网络。

2 相关工作

2.1 单模态行人重识别

单模态下的可见光行人重识别可以分为基于表征学习和基于度量学习两类，以解决行人的外观变化、光照变化等[13]问题。基于表征学习的方法主要通过提取图像的颜色、纹理等信息来实现行人重识别，主要采用方向梯度直方图特征（HOG）[14]、尺度不变特征变换特征（SIFT）[15]等方法以人工方式提取行人特征。基于度量学习的核心思想是使同一类样本间的距离最小，使各类样本间的距离最大，如LMNN[16]、XQDA[17]等。Li等[18]首次采用深度学习的方法来解决行人重识别问题，并取得了显著的效果。与传统方法相比，基于深度学习的方法[19-21]可以有效提高行人重识别的准确率。但由于不同模态间存在较大的异质性，目前的单模态行人重识别方法不能直接应用于跨模态行人重识别。

2.2 跨模态行人重识别

跨模态行人重识别旨在解决不同模态图像之间的行人匹配问题，例如可见光图像与红外图像的匹配、图像与文本的匹配、可见光图像与素描图像的匹配等。

近年来，基于可见光红外图像跨模态行人重识别逐渐成为热点研究方向。Wu等[8]提出“可见光-红外”跨模态行人重识别的数据集——SYSU-MM01，并采用深度零填充的方法缓解跨模态数据信息之间的错位问题，获得了较好的识别精度。Ye等[1]利用双流神经网络提取不同模态之间的特征，将其映射到相同特征空间中，并采用对比损失函数来约束不同模态数据分布之间的一致性。Wang等[22]提出一种将可见光图像和红外图像互相转换的方法以减小模态间的差异。Wang等[11]提出像素对齐的方法来缓解模态差异问题，并提出联合判别策略来保持对齐过程中的身份一致性。Hao等[23]采用一种端到端的双流超球面流形嵌入模型来约束模态内和模态间的变化。Zhu等[24]提出异质中心损失，对两个异质模态之间的类内中心距离进行约束，从而监督网络学习跨模态图像间的不变信息，减小了类内交叉模态的变化。Lu等[25]提出跨模态共享及非共享特征转移算法，采用多流结构的基线网络来提取特征。为了减少两种模态图像差异，Dai等[26]首次在跨模态行人重识别中引入GAN，提出一种全新的跨模态生成对抗网络（cmGAN），从两种模态中学习具有判别力的特征。随着GAN的发展，基于GAN的跨模态行人重识别方法有效地实现了图像风格的转变，进一步减小了模态差异，保留身份一致性。然而，上述的研究往往着眼于学习全局特征，忽略了行人之间局部共享特征和同一行人不同模态图像之间的潜在关系，对模态内部信息挖掘不够充分。此外，这些方法还容易引入噪声从而影响跨模态行人重识别的准确率。

2.3 注意力机制

注意力机制广泛应用于语音识别、自然语言处理及图像识别等各种机器学习任务中。在神经网络中，注意力机制可以引导网络关注获取输入的某些部分，或者赋予输入部分不同的权重，以增强数据特征表示。

Wang 等[27]提出一种非局部神经网络(non-local),通过建模像素之间的相互关系来捕获特征的长距离依赖,并以像素间关系的重要程度作为权重以进一步地提取显著性特征。Hu 等[28]提出压缩和激励(SE)模块,首次关注模型通道层面的依赖关系,通过对各通道添加注意力权重提高网络表达能力。Fu 等[29]提出双注意力网络(DANet),基于自注意力机制来分别捕获空间维度和通道维度中的特征依赖关系。Wang 等[30]通过对 SE 模块的改进,设计了一种能够解决错位问题,并对差异性局部特征定位的注意力模块。Gui 等[31]提出无参数的空间注意力(SA)模块,该模块通过对不同空间位置赋予不同的权重,从而获得更具表现力的特征。然而,这些注意力机制对于较大的模态差异和噪声会着重强化某一特定的信息从而忽略行人的其他特征。

为了解决以上问题,设计了注意力模块 AFA 来引导网络挖掘更丰富的局部特征,从而提升模型的泛化能力。该注意力模块还能够和 CGSA 互相配合,减小不同模态图像之间特征差异,提升模型在跨模态行人重识别中的性能。

3　基本原理

DCA-Net 的主体结构如图 1 所示。该网络主要由主干网络 ResNet-50、ICGR、AFA、CGSA 构成。其中:主干网络用于提取行人图像的特征;ICGR 被嵌入 ResNet-50 共享层中以更高效地提取跨模态图像的共享特征;AFA 和 CGSA 用来获取行人图像更丰富的局部特征,实现模态信息的交互融合。

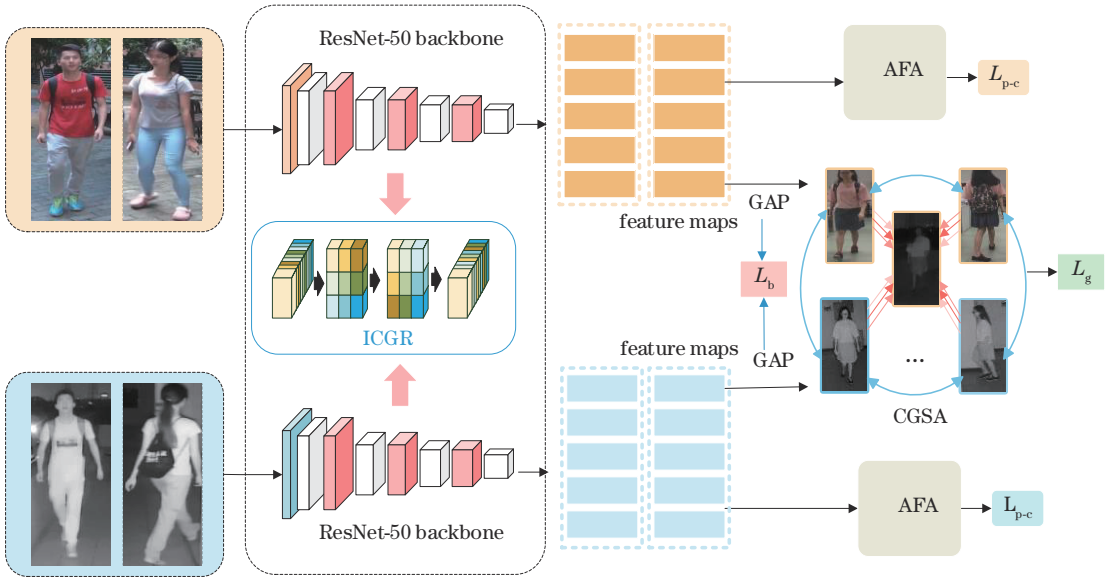

图 1　DCA-Net 整体框架

Fig. 1　Overall framework of DCA-Net

3.1　跨模态行人重识别基础网络框架

随着深度学习的发展,学者们提出了很多有效的网络模型,如 GoogleNet、VGG 和 ResNet 等[32]。由于 ResNet 不仅能够有效地解决梯度爆炸和梯度消失的问题,还可以提取具有丰富高层语义信息和判别力的特征,因此选择 ResNet-50 作为主干网络。为了获取两种不同模态图像的独有特征,选取 ResNet-50 网络的第一层卷积块作为独有特征提取网络。记行人的可见光图像为 $\boldsymbol{x}_{\mathrm{rgb}}$,红外图像为 $\boldsymbol{x}_{\mathrm{ir}}$,通过可见光模态独有特征提取卷积块 $\mathrm{Conv}_{\mathrm{rgb}}$ 和红外模态独有特征提取卷积块 $\mathrm{Conv}_{\mathrm{ir}}$ 分别提取可见光图像特征 $\boldsymbol{X}_{\mathrm{rgb}}$ 和红外图像特征 $\boldsymbol{X}_{\mathrm{ir}}$。

3.2　特征通道分组重组模块

由于跨模态数据之间包含模态共享和模态特有信息,因此跨模态行人重识别任务期望学习到更多模态共享特征信息。

为了解决上述问题,设计了模态内特征通道分组重组模块。首先对输入特征 $\boldsymbol{X}_{\mathrm{in}}$ 进行分组卷积操作,在减少参数量的同时增加相邻特征图之间的相关性,然后对输出特征进行通道重组以融合多通道特征信息,获取更具有判别力的特征图,最后得到整个模块的输出特征 $\boldsymbol{X}_{\mathrm{out}}$。

分组卷积不同于一般意义上的标准卷积,分组卷积需要先对输入的特征图进行分组,然后逐个进行卷积,最后对卷积得到的特征进行拼接作为分组卷积的输出。分组卷积的计算公式为

$$\boldsymbol{y}_i = f\left(\sum_{i=1}^{N/G} \boldsymbol{W}_j \otimes \boldsymbol{x} + \boldsymbol{b}_j\right), j = 1, 2, \cdots, \tag{1}$$

式中:$\boldsymbol{y}_i$ 表示每个分组卷积输出的特征;$\boldsymbol{x}$ 表示输入特征;$\boldsymbol{W}_j$、$\boldsymbol{b}_j$ 分别表示每个分组卷积的权值和偏置值;f 表示 ReLU 激活函数;N 表示输出的特征个数;G 表示分组数。

由于分组卷积中,每一组都独立进行卷积运算,这

种方式会导致各组之间无法进行信息交流。通道重组操作通过对分组卷积之后的特征图以均匀打乱的方式进行特征图重新组合，从而确保分组卷积层之后的网络层输入来自不同的组。因为通道重组融合了多通道的特征信息，所以输出的特征图更具有判别力。

具体地：首先进行第1次特征维度重塑，将输入特征按照通道维度拆分为[n, m]两个维度；随后进行转置操作得到[m, n]；最后进行第2次特征维度重塑得到1个通道维度为$n\times m$的特征。通道重组可描述为

$$F(\boldsymbol{x}')=F_{r2}\left\{F_{T}\left[F_{r1}(\boldsymbol{x}')\right]\right\}, \tag{2}$$

式中：$\boldsymbol{x}'$表示输入特征；F_{r1}代表第1次特征维度重塑操作；F_{T}代表转置操作；F_{r2}代表第2次特征维度重塑操作。

所设计的特征通道分组重组模块如图2所示。具体地，该模块由N个分组卷积层及通道重组操作构成。该模块以提取到的可见光图像特征$\boldsymbol{X}_{rgb}$或红外图像特征$\boldsymbol{X}_{ir}$作为输入$\boldsymbol{X}_{in}$。首先将输入特征$\boldsymbol{X}_{in}$均分为N组进行卷积，得到N个输出，记为$\boldsymbol{y}_1$, $\boldsymbol{y}_2$, …, $\boldsymbol{y}_n$，然后将其沿通道方向进行拼接得到特征图$\overline{\boldsymbol{X}}_{in}$。将$\overline{\boldsymbol{X}}_{in}$进行第1次特征维度重塑，随后进行转置操作，最后进行第2次特征维度重塑，得到模块输出$\boldsymbol{X}_{out}$。

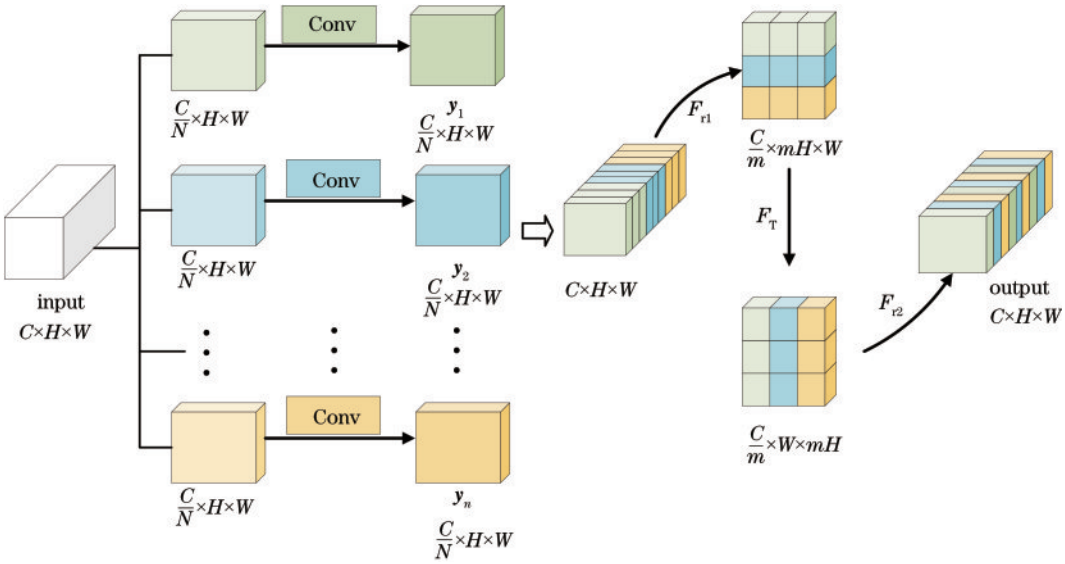

图2　模态内特征通道分组重组模块

Fig. 2　Intra-modal feature channel grouping and reorganization module

3.3　注意力机制

3.3.1　聚合特征注意力机制

在跨模态行人重识别任务中，注意力机制能够有效抑制无关的背景信息，关注行人特征。然而，当两种模态之间图像差异过大时，仅仅通过一个注意力模块难以挖掘更加具有判别力的局部特征，从而影响跨模态行人重识别的精度。

为解决以上问题，提出聚合特征注意力模块，其结构如图3所示。该注意力机制分为两个分支，其中，一个分支包含通道注意力模块(CAM)和空间注意力模

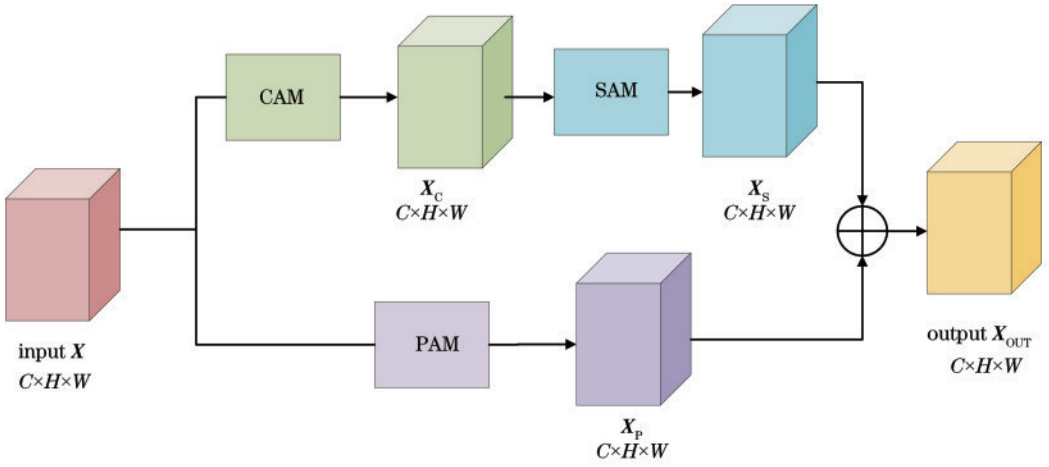

图3　聚合特征注意力机制模块

Fig. 3　Aggregated feature attention mechanism module

块(SAM),另一个分支为位置注意力模块(PAM)。给定输入特征$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$,首先通过第一个分支中的通道注意力模块得到特征映射$\boldsymbol{X}_{\mathrm{C}}$,然后通过空间注意力模块得到特征映射$\boldsymbol{X}_{\mathrm{S}}$;另一分支中,通过位置注意力模块得到特征映射$\boldsymbol{X}_{\mathrm{P}}$;最后将上述所得到的特征映射$\boldsymbol{X}_{\mathrm{S}}$、$\boldsymbol{X}_{\mathrm{P}}$相加。最终特征映射为

$$\boldsymbol{X}_{\mathrm{OUT}}=\boldsymbol{X}_{\mathrm{S}}+\boldsymbol{X}_{\mathrm{P}}。\tag{3}$$

聚合特征注意力模块以多模块及多层级的嵌入方式协同三种注意力模块,捕获多尺度上下文信息,进而保留显著细节特征,增强特征间的相互依赖性,挖掘行人图像之间的潜在关系,提取到更具有判别力的行人特征,提高模型对行人特征的敏感度,进而提高模型的识别性能。

3.3.2　通道注意力模块

通道注意力能够通过赋予每个通道不同的权重系数以增强具有显著性特征的通道和抑制不重要的通道。为此,设计了一个通道注意力模块来构建通道之间的相互关联并获取各通道信息的重要程度。具体如图4所示,该通道注意力模块由2个池化层(全局平均池化和全局最大池化)、2个全连接层和1个Sigmoid层构成。给定输入特征映射$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$,其中,$C$表示通道数,$H$、$W$表示特征映射的高和宽。该通道注意力模块生成的通道注意特征图$\boldsymbol{A}_{\mathrm{C}}\in\mathbf{R}^{C\times 1}$的表达式为

$$\boldsymbol{A}_{\mathrm{C}}=\sigma\left[\boldsymbol{W}_1\boldsymbol{X}_{\mathrm{MP}};\ \boldsymbol{W}_2\boldsymbol{X}_{\mathrm{AP}}\right],\tag{4}$$

式中:[•;•]表示沿通道拼接操作;$\sigma(\cdot)$表示Sigmoid函数;$\boldsymbol{X}_{\mathrm{MP}}$、$\boldsymbol{X}_{\mathrm{AP}}$分别表示经过全局最大池化和全局平均池化处理后的特征映射;$\boldsymbol{W}_1\in\mathbf{R}^{\frac{C}{r}\times C}$、$\boldsymbol{W}_2\in\mathbf{R}^{\frac{C}{r}\times C}$分别表示全连接层的参数,其中,$r$表示降维比。

在得到通道注意力特征图$\boldsymbol{A}_{\mathrm{C}}$后,将输入特征映射与$\boldsymbol{A}_{\mathrm{C}}$相乘得到该通道注意力模块最终输出特征映射$\boldsymbol{X}_{\mathrm{C}}=\boldsymbol{X}\otimes\boldsymbol{A}_{\mathrm{C}}$,$\otimes$表示对应元素相乘。

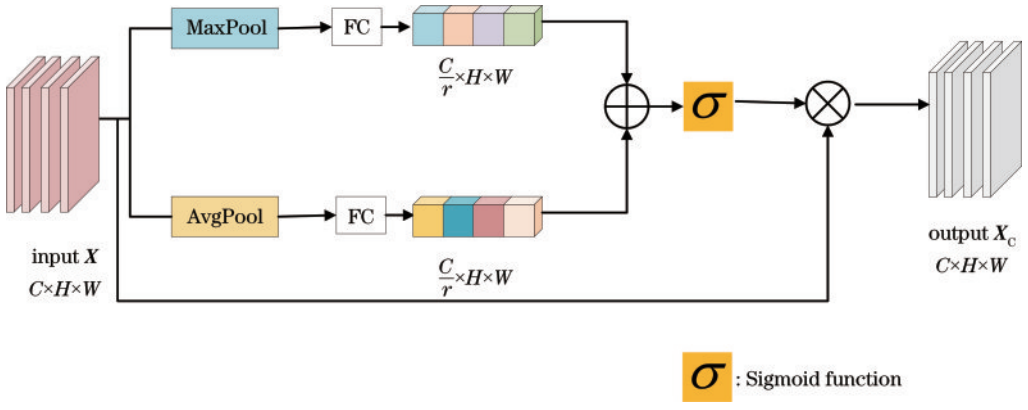

图4　通道注意力模块

Fig. 4　Channel attention module

3.3.3　空间注意力模块

空间注意力特征能够使网络在关注行人图像中最显著区域特征的同时,抑制背景干扰信息。所提空间注意力模块包含两个池化层(全局最大池化和全局平均池化)、一个卷积核大小为1×1的卷积层和一个Sigmoid层,如图5所示。

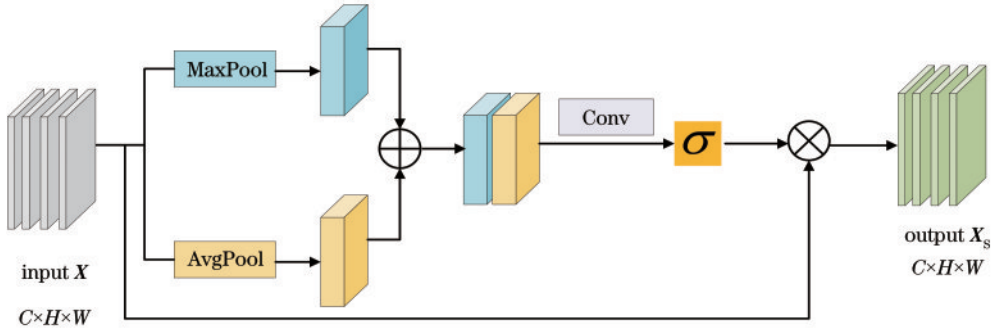

图5　空间注意力模块

Fig. 5　Spatial attention module

给定一个特征映射$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$,空间注意力模块产生的空间注意力特征图$\boldsymbol{A}_{\mathrm{S}}\in\mathbf{R}^{2\times H\times W}$的表达式为

$$\boldsymbol{A}_{\mathrm{S}}=\sigma\left\{\varphi\left[\boldsymbol{X}_{\mathrm{MP}};\ \boldsymbol{X}_{\mathrm{AP}}\right]\right\},\tag{5}$$

式中:$\varphi(\cdot)$表示卷积核为1×1的卷积运算。

在得到空间注意力特征图$\boldsymbol{A}_{\mathrm{S}}$后，将输入特征映射与$\boldsymbol{A}_{\mathrm{S}}$相乘得到该空间注意力模块最终输出特征映射$\boldsymbol{X}_{\mathrm{S}}$。

3.3.4　位置注意力模块

位置注意力特征能够对差异性特征信息进行合理约束，提升网络挖掘相似特征信息的能力，从而获得更丰富的局部特征。为此设计了一个位置注意力模块来加强模型对行人图像信息的挖掘能力。

位置注意力模块如图 6 所示，该模块包含 3 个卷积核大小为 1×1 的卷积层，即 $m(\cdot)$、$n(\cdot)$、$o(\cdot)$，1 个归一化层(BN)，一个可学习权重向量$\boldsymbol{W}^{p}$。给定一个特征映射$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$，位置注意力特征图$\boldsymbol{A}_{\mathrm{P}}$如下所示：

$$\begin{cases}\boldsymbol{a}_{i,j}=\dfrac{\exp\left[m\left(\boldsymbol{x}_i^{p}\right)^{\mathrm{T}}n\left(\boldsymbol{x}_j^{p}\right)\right]}{\sum_{i=1}^{N}\exp\left[m\left(\boldsymbol{x}_i^{p}\right)^{\mathrm{T}}n\left(\boldsymbol{x}_j^{p}\right)\right]},\\ \boldsymbol{A}_{\mathrm{P}}=\boldsymbol{W}^{p}\left[\boldsymbol{a}_i^{p}*o\left(\boldsymbol{x}_i^{p}\right)\right]\end{cases}\tag{6}$$

式中：$\boldsymbol{a}_{i,j}$是指 i 位置对 j 位置的影响；$m(\boldsymbol{x}_j^{p})$、$n(\boldsymbol{x}_i^{p})$、$o(\boldsymbol{x}_i^{p})$分别表示将特征映射$\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$划分为 p 个非重叠部分后通过卷积的特征图；$m(\boldsymbol{x}_j^{p})$与$n(\boldsymbol{x}_i^{p})$相乘得到局部注意力特征图$\boldsymbol{a}_i^{p}$；$\boldsymbol{W}^{p}$代表不同部分的可学习权重向量。

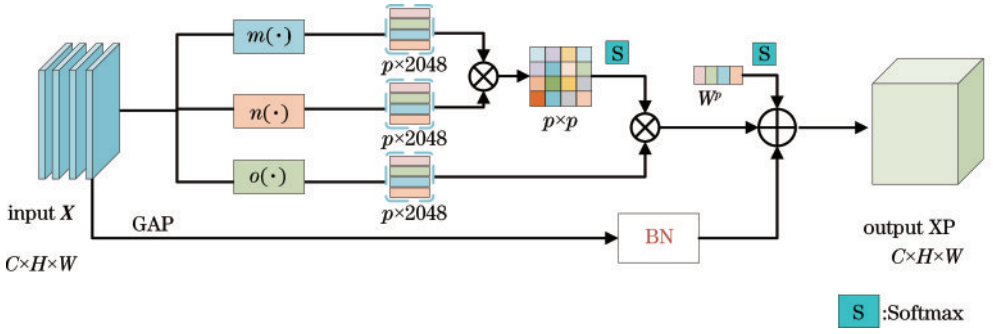

图 6　位置注意力模块

Fig. 6　Position attention module

在得到位置注意力图$\boldsymbol{A}_{\mathrm{P}}$后，将输入特征图经过全局自适应池化、归一化操作后与$\boldsymbol{A}_{\mathrm{P}}$相加，该位置注意力模块最终的输出特征映射为

$$\boldsymbol{X}_{\mathrm{P}}=B(\boldsymbol{x}^{o})+\boldsymbol{A}_{\mathrm{P}},\tag{7}$$

式中：$\boldsymbol{x}^{o}$代表输入特征图$\boldsymbol{X}$的全局自适应池化输出；$B(\cdot)$为批归一化操作。

3.3.5　跨模态自适应图结构

可见光-红外模态图像对视觉差异较大，如图 7 所示，导致模型无法充分学习到辨别性行人特征，破坏了优化过程。引入跨模态自适应图结构用于跨模态行人

图 7　可见光图像与红外图像对比

Fig. 7　Comparison of visible images and infrared images

重识别问题，通过学习两种模态图像之间的结构关系，以加强特征表示。主要思想是属于同一身份的不同模态图像的特征表示是互利的[12]。

图注意力可以衡量单节点 i 对另一模态中节点 j 的重要性。用池化层的输出 $\boldsymbol{X}^o=\left\{\boldsymbol{x}_k^o\in\mathbf{R}^{C\times 1}\right\}_{k=1}^{K}$ 表示输入节点特征。图关注系数 $\boldsymbol{\alpha}_{i,j}^g\in[0,1]^{K\times K}$ 的表达式为

$$\boldsymbol{\alpha}_{i,j}^g=\frac{\exp\left\{\Gamma\left\{\left[h(\boldsymbol{x}_i^o),h(\boldsymbol{x}_j^o)\right]\cdot\boldsymbol{w}^g\right\}\right\}}{\sum_{\forall\boldsymbol{A}^g(i,k)>0}\exp\left\{\Gamma\left\{\left[h(\boldsymbol{x}_i^o),h(\boldsymbol{x}_j^o)\right]\cdot\boldsymbol{w}^g\right\}\right\}},\tag{8}$$

式中：$\Gamma(\cdot)$ 代表 Leaky ReLU 操作；$[\cdot,\cdot]$ 表示连词运算；$h(\cdot)$ 表示将输入节点特征维度 C 缩减为 d 的变换矩阵，实验中设 d 为 256；$\boldsymbol{w}^g\in\mathbf{R}^{2d\times 1}$ 表示可学习的权重向量，用来衡量不同特征维度在串联特征中的重要性；$\boldsymbol{A}^g$ 为规范化邻接矩阵的无向图。

在每一输入批次中随机选取 N 个行人样本，对每个行人样本选取 M 张可见光图像和 M 张红外图像，从而在每一训练批次中都生成 $K(2\times N\times M)$ 个图像。图结构用归一化邻接矩阵表示：

$$\begin{cases}\boldsymbol{A}^g=\boldsymbol{A}_i^g+\|_K\\ \boldsymbol{A}_i^g(i,j)=\boldsymbol{l}_i*\boldsymbol{l}_j\end{cases},\tag{9}$$

式中：$\boldsymbol{l}_i$ 与 $\boldsymbol{l}_j$ 分别为图节点与相应的独热编码；$\|_K$ 是由其自身构成的矩阵，表示各节点都与其自相连。

将具有相同身份的上下文信息和跨模态的图像之间的关系结合起来，可以提高特征的表达。通过学习两个不同模态间的关系，结合跨模态两种模式的结构关系来加强特征表示，最后输出特征表示为

$$\boldsymbol{X}^{g'}=\{\boldsymbol{x}_i^{g'}\in\mathbf{R}^{C\times 1}\}_{i=1}^{m},\tag{10}$$

式中：m 表示当前输入批次样本数；C 表示最后一个池化层输出的特征维度。

3.4 损失函数

3.4.1 跨模态三元组损失

在所提方法中，每一批的网络训练过程包含 N 个不同身份的行人，每个行人包含 M 个可见光图像和 M 个红外图像，所以每一批次总计 $2NM$ 个样本。针对每个样本 $\boldsymbol{X}$，选择另一种模态的正样本 $\boldsymbol{X}_{i\text{-}\mathrm{p}}$，相同模态的负样本 $\boldsymbol{X}_{i\text{-}\mathrm{N}}$ 组成跨模态三元组[12]。例如：对于可见光图像特征 $\boldsymbol{T}'_{\mathrm{RGB}}$，选择红外模态的正样本 $\boldsymbol{T}'_{\mathrm{IR\text{-}P}}$，可见光模态的负样本 $\boldsymbol{T}'_{\mathrm{RGB\text{-}N}}$ 组成跨模态三元组；对于红外图像特征 $\boldsymbol{T}'_{\mathrm{IR}}$，选择可见光模态的正样本 $\boldsymbol{T}'_{\mathrm{RGB\text{-}P}}$，红外模态的负样本 $\boldsymbol{T}'_{\mathrm{IR\text{-}N}}$ 组成跨模态三元组。跨模态三元组损失的表达式为

$$\begin{aligned}L_{\mathrm{tri}}=&\max\left[D(\boldsymbol{T}'_{\mathrm{RGB}},\boldsymbol{T}'_{\mathrm{IR\text{-}P}})-D(\boldsymbol{T}'_{\mathrm{RGB}},\boldsymbol{T}'_{\mathrm{RGB\text{-}N}})+\right.\\&\left.V_{\mathrm{margin}}\right]+\max\left[D(\boldsymbol{T}'_{\mathrm{IR}},\boldsymbol{T}'_{\mathrm{RGB\text{-}P}})-D(\boldsymbol{T}'_{\mathrm{IR}},\boldsymbol{T}'_{\mathrm{IR\text{-}N}})+\right.\\&\left.V_{\mathrm{margin}}\right],\end{aligned}\tag{11}$$

式中：距离度量 D 选择欧氏距离；V_{margin} 为阈值。

3.4.2 身份损失

身份损失针对属于同一类别的样本图像，可以引导网络模型学习到更相似的行人特征图，在样本图像的语义特征空间减小类内距离。身份损失使用交叉熵函数，具体形式如下：

$$L_{\mathrm{id}}(x_i)=y_i\log(p_i)+(1-y_i)\log(1-p_i),\tag{12}$$

式中：对于每个样本 x_i，其对应的行人身份标签为 y_i；使用分类器网络预测样本 x_i 属于该身份类别的概率为 p_i。

跨模态三元组损失 L_{tri} 优化了两种模式下不同人物图像之间的三重态关系，身份损失 L_{id} 引导网络模型学习到更相似的行人特征图。所提学习判别特征部分包括跨模态三元组损失和身份损失，该部分的损失为

$$L_{\mathrm{b}}=L_{\mathrm{tri}}+L_{\mathrm{id}}。\tag{13}$$

3.4.3 动态多注意聚合学习

为更好地解决跨模态行人重识别任务中行人图像特征差异较大的问题，需将以上提出的聚合特征注意力机制和跨模态自适应图结构整合。由于这两个部分侧重于不同学习目标，若将其损失函数进行简单组合用于监督网络训练，跨模态自适应图结构约束模块部分将会十分不稳定。

为了解决上述问题，采用动态多注意聚合学习的方法。具体实现如下，将整个联合学习框架分解为两个不同的任务，分别作用于模态内聚合特征学习损失 L_{p} 和跨模态自适应图结构模块 L_{g} 两部分。其中，L_{p} 是学习目标损失 L_{b} 和聚合特征注意力机制损失 $L_{\mathrm{p\text{-}c}}$ 的组合：

$$L_{\mathrm{p\text{-}c}}=-\frac{1}{N}\sum_{i=1}^{N}y_i\log\left[p(y_i|x_i^*)\right],\tag{14}$$

$$L_{\mathrm{p}}=L_{\mathrm{b}}+L_{\mathrm{p\text{-}c}},\tag{15}$$

式中：N 代表每一批次的图片数量；p 表示特征被正确分类的概率；y_i 表示输出的图片特征；x_i^* 表示输入的图片特征。

为了引导跨模态自适应图结构的学习，选择负对数似然损失作为跨模态自适应图结构约束模块部分的损失表示：

$$L_{\mathrm{g}}=-\sum_{i}^{M}\log\left[\mathrm{Softmax}(x_i^{g'})\right],\tag{16}$$

式中：$x_i^{g'}$ 为通过图卷积操作后的输出特征。

受多任务学习的启发[12]，动态多注意力聚合学习策略实质上是将 L_{p} 看作主要损失，逐步增加学习损失 L_{g}。其主要原因是，在初期训练阶段能够更简单地学习到作用于图像级部分聚合特征。在经过监督网络学习一段时间后，引入跨模态间的全局特征学习损失 L_{g} 可以对网络进一步优化，而不会导致过分剧烈的震荡。动态多注意力聚合学习损失为

$$L_e=\frac{1}{1+E(L_{\mathrm{p}}^{e-1})}L_{\mathrm{g}}^e+L_{\mathrm{p}}^e,\tag{17}$$

式中：e为训练次数；$E(L_p^{e-1})$代表前一个训练轮次的平均损失值；L_p^e代表当前轮次模态内聚合特征学习损失值；L_g^e代表当前轮次跨模态自适应图结构约束数值。

学习判别特征L_b及动态多注意聚合损失L_e之和即为所提网络的总体损失函数：

$$L = L_b + L_e。 \tag{18}$$

4 实验与结果分析

4.1 实验设置

所提算法在PyTorch框架上实现，使用NVIDIA 3090 GPU进行模型训练。采用ResNet-50[32]作为骨干网络进行特征提取，将输入图像大小调整到288×144，然后采用模态内特征通道分组重组模块，对于注意力机制部分采用文献[12]中同样的设置。采用随机梯度下降优化算法，将动量参数设置为0.3，将两个数据集的初始学习率都设置为0.1，学习率在第20轮次时衰减到0.01，在第50轮次时衰减到0.001，在两个数据集上各有80个训练轮次。

4.2 数据集和评价标准

RegDB数据库[6]包含412个不同身份行人，针对每个行人身份采集10张可见光图像及10张红外图像，训练集和测试集分别包含206个行人，4120张图像。对于RegDB数据库，常规的测试方法利用可见光进行检索，将红外图像作为待检索图像[1]。

SYSU-MM01数据集[8]是跨模态行人重识别的公认权威数据集，包含287628张可见光图像和15792张红外图像，共计491个行人信息。SYSU-MM01数据集分为训练集和测试集，分别包含395个和96个行人，训练集中有22258张可见光图像和11909张红外图像；测试集中有3803张可查询图像和随机选取的301张可见光图像作为图库集。根据其标准评估协议，数据集包括all-search模式和indoor-search检索模式。

使用累计匹配特性(CMC)曲线中的Rank-1识别率、Rank-10识别率和Rank-20识别率作为评价指标[1]。此外，还采用均值平均精度(mAP)作为评价指标。CMC统计在前k次检索结果中出现正确的人物图像的概率，mAP衡量图库集中出现多个匹配图像时的检索性能。

4.3 与其他算法的比较

将所提算法与现有的跨模态行人重识别算法进行比较，在SYSU-MM01数据集上的两种查询模式的实验结果如表1所示，可以看出，所提算法在性能相较于现有算法有着一定程度的提高。DCA-Net在具有挑战性的SYSU-MM01数据集全局查询模式下实现了59.23%的Rank-1精度和56.55%的mAP准确率。特别地，相比于先进方法DDAG，在all-search下，DCA-Net的Rank-1精度和mAP分别提高了4.48个百分点和3.53个百分点。这进一步验证了所提解决方案的有效性。

表1 DCA-Net和目前先进方法在SYSU-MM01数据集上的性能比较

Table 1 Performance comparison of DCA-Net and current state-of-the-art methods on the SYSU-MM01 dataset

Setting	all-search				indoor-search			
Method	r=1	r=10	r=20	mAP /%	r=1	r=10	r=20	mAP /%
HOG[14]	2.76	18.30	31.90	4.24	3.22	24.70	44.50	7.25
BDTR[33]	17.01	55.43	71.96	19.66				
HSME[23]	20.68	32.74	77.95	23.12				
D2RL[22]	28.90	70.60	82.40	29.20				
MAC[34]	33.26	79.04	90.09	36.22	36.43	62.36	71.63	37.03
MSR[35]	37.35	83.40	93.34	38.11	39.64	89.29	97.66	50.88
AlignGAN[11]	42.40	85.00	93.70	40.70	45.90	87.60	94.40	54.30
cmGAN[26]	26.97	67.51	80.56	31.49	31.63	77.23	89.18	42.19
HPILN[36]	41.36	84.78	94.31	42.95	45.77	91.82	98.46	56.52
LZM[37]	45.00	89.06	95.77	45.94	49.66	92.47	97.15	59.81
AGW[1]	47.50	84.39	92.14	47.65	54.17	91.14	95.98	62.97
X-modal[38]	49.92	89.79	95.96	50.73				
DDAG[12]	54.75	90.39	95.81	53.02	61.02	94.06	**98.41**	67.98
Proposed method	**59.23**	**91.83**	**96.63**	**56.55**	**63.22**	**94.39**	**98.20**	**69.54**

在RegDB数据集上的实验结果如表2所示，所提模型在两种查询设置中都获得了较高的性能，对于可见光到红外查询设置，Rank-1和mAP的数值分别为78.16%和71.18%。

4.4 消融实验

4.4.1 模块有效性分析

为了评估DCA-Net中每个组件的有效性，在SYSU-MM01数据集上进行了消融实验。具体地，以

表 2 DCA-Net 和目前先进方法在 RegDB 数据集上的性能比较

Table 2 Performance comparison of DCA-Net and current state-of-the-art methods on RegDB dataset

Setting	Visible to thermal				Thermal to visible			
Method	r=1	r=10	r=20	mAP /%	r=1	r=10	r=20	mAP /%
HCML[24]	24.44	47.53	56.78	20.08	21.70	45.02	55.58	22.24
BDTR[33]	33.56	58.61	67.43	32.76	32.92	58.46	68.43	31.96
D2RL[22]	43.40	66.10	76.30	44.10				
HSME[23]	50.85	73.36	81.66	47.00	50.15	72.40	81.07	46.16
MAC[39]	36.43	62.36	71.63	37.03	36.20	61.68	70.99	36.63
MSR[35]	48.43	70.32	79.95	48.67				
EDFL[40]	52.58	72.10	81.47	52.98	51.89	72.09	81.04	52.13
AlignGAN[11]	57.90			53.60	56.30			53.40
LZM[37]	57.03	76.10	84.34	58.06				
X-modal[38]	62.21	83.13	91.72	60.18				
AGW[1]	70.05	86.21	91.55	66.37	70.49	87.12	91.84	65.90
DDAG[12]	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
Proposed method	**78.16**	**91.75**	**94.66**	**71.18**	**77.62**	**91.60**	**94.47**	**70.56**

由主干网络 ResNet-50、损失函数组成的网络模型作为基线(Baseline),并在此基础上构建了以下 3 个网络:1) Baseline+CGSA;2) Baseline+CGSA+AFA;3)Baseline+CGSA+AFA+ICGR。消融实验结果如表 3 所示。从表 3 可以看出,在 SYSU-MM01 数据集上,与 Baseline 相比,Baseline+CGSA 的 Rank-1 精度提升了 2.57 个百分点,mAP 提升了 2.09 个百分点,这说明 CGSA 能够有效提高基线网络的性能。在 SYSU-MM01 数据集上,与 Baseline+CGSA 相比,Baseline+CGSA+AFA 的 Rank-1 精度提升了 6.98 个百分点,mAP 提升了 4.69 个百分点,这充分说明 AFA 模块能够有效提升网络性能。与 Baseline+CGSA+AFA 相比,Baseline+CGSA+AFA+ICGR 的 Rank-1 精度提升了 1.50 个百分点,mAP 提升 2.13 个百分点,这充分说明 ICGR 模块能够进一步提升网络的性能。

表 3 在 SYSU-MM01 数据集上的消融实验研究

Table 3 Experimental study of ablation on SYSU-MM01 dataset

unit: %

Baseline	CGSA	AFA	ICGR	SYSU-MM01	
				Rank-1	mAP
√				48.18	47.64
√	√			50.75	49.73
√	√	√		57.73	54.42
√	√	√	√	59.23	56.55

4.4.2 ICGR 插入位置有效性探究

模态内特征通道分组重组模块通过提取不同模态样本的共享特征,以缓解跨模态差异。为了验证模态内特征通道分组重组模块的有效性,在 SYSU-MM01 数据集上进行了实验。首先,将保留主干网络、注意力机制、损失函数的网络模型设置为基线模型(Baseline)。然后分成以下 4 组实验:1)不使用 ICGR 模块(对应 Baseline);2)在 ResNet-50 的第 2 个卷积块后插入 ICGR 模块;3)在 ResNet-50 的第 2、3 个卷积块后插入 ICGR 模块;4)在 ResNet-50 的第 2、3、4 个卷积块后插入 ICGR 模块。

所有实验采用相同的参数设置,对应方法的实验结果如表 4 所示,使用 Rank-1 和 mAP 作为评价指标。

表 4 ICGR 插入位置在 SYSU-MM01 在数据集下的实验结果

Table 4 Experimental results of ICGR inserted different position under SYSU-MM01 dataset

unit:%

Baseline	Conv2	Conv3	Conv4	SYSU-MM01	
				Rank-1	mAP
√				57.73	54.42
√	√			58.27	54.73
√	√	√		59.19	56.19
√	√	√	√	59.23	56.55

从表 4 可以看出,在 SYSU-MM01 数据集上,当没有插入模态内特征通道分组重组模块时,方法 1 基线模型在 all-search 模式下获得 57.73% 的 Rank-1 准确率和 54.42% 的 mAP。当插入模态内特征通道分组重组模块后,方法 2、方法 3、方法 4 的性能得到明显提升,这是因为模态内特征通道分组重组模块通过获取两种模态的语义信息,可以将一种模态的图像语义信息融合至其他模态特征中,提高局部特征敏感度,从而提高单一模态图像特征的语义丰富度,缓解跨模态差异。方法 4 提升了 1.5 个百分点的 Rank-1 准确率和 2.13 个百分点的 mAP,这是因为方法 2、方法 3 中,模态内特征通道分组重组模块插入较低层的网络中,此时网络学习到的低级信息相对较多,容易导致通道重

组后的图像表征包含大量的混杂信息，影响最终的高层语义匹配。而方法4将模态内特征通道分组重组模块插入高层网络中，此时网络能更好地利用高级语义信息，可以有效降低跨模态图像之间的差别，提高特征匹配精度。因此，在网络高层更容易获得语义信息，从而缓解跨模态差异，提高检索性能。

4.5 损失函数的有效性

探究了不同损失函数对模型性能的影响，从而验证所提方法的有效性，结果如表5所示。可以看出：仅采用身份损失 L_{id} 时，准确率较低；当加入跨模态三元组 L_{tri} 后，类间距离扩大，准确率有一定的提升；当再次加入动态多注意聚合损失 L_e 后，模态间的差距较小，模型的准确率进一步提升。

表5 不同损失函数对模型性能的影响

Table 5 Effect of different loss functions on model performance

unit: %

Loss function	SYSU-MM01		RegDB	
	Rank-1	mAP	Rank-1	mAP
L_{id}	56.89	54.75	70.63	62.03
$L_{tri}+L_{id}$	57.73	54.42	72.18	66.03
$L_e+L_{tri}+L_{id}$	59.23	56.55	78.16	71.18

4.6 复杂度分析

还比较了所提DCA-Net与AGW[1]、DDAG[12]的计算时间和参数量，结果如表6所示，其中，时间代表的是每训练一个epoch所用的时间。可以看出，相对于AGW模型和DDAG模型，所提方法并未引入额外较大的计算开销。具体地：相对于AGW模型，所提方法训练时间与其相近，而训练模型所占内存有一定增量，但相较于模型性能的提升，不足100 MB的模型内存增量是合理的；相对于DDAG模型，所提方法训练时间更短，训练模型所占内存只增加了1.52 MB，几乎可以忽略不计。

表6 模型复杂性分析

Table 6 Model complexity analysis

Model	Model memory /MB	Training time /s
AGW	273	234.33
DDAG	362.48	299.82
DCA-Net	364	237.07

5 结 论

提出一个基于通道重组和注意力机制的双流网络DCA-Net，用来提取不同模态行人图像之间更加稳定的共享特征和更加丰富的行人特征信息。DCA-Net包括模态内特征通道分组重组模块、聚合特征注意力机制和跨模态自适应图结构。模态内特征通道分组重组模块用于提取不同模态图像的共享特征；聚合特征注意力机制用以挖掘行人图像之间的潜在关系；跨模态自适应图结构用于加强特征表示。大量实验结果表明，所提DCA-Net跨模态行人重识别精度达到了目前先进水平。

参 考 文 献

[1] Ye M, Shen J B, Lin G J, et al. Deep learning for person re-identification: a survey and outlook[EB/OL]. (2020-01-13)[2021-05-05]. https://arxiv.org/abs/2001.04193.

[2] 刘莎, 党建武, 王松, 等. 结合一阶和二阶空间信息的行人重识别[J]. 激光与光电子学进展, 2021, 58(2): 0215005.
Liu S, Dang J W, Wang S, et al. Person re-identification based on first-order and second-order spatial information [J]. Laser & Optoelectronics Progress, 2021, 58(2): 0215005.

[3] 李爽, 李华锋, 李凡. 基于互预测学习的细粒度跨模态行人重识别[J]. 激光与光电子学进展, 2022, 59(10): 1010010.
Li S, Li H F, Li F. Fine-grained cross-modality person re-identification based on mutual prediction learning[J]. Laser & Optoelectronics Progress, 2022, 59(10): 1010010.

[4] 王凤随, 刘芙蓉, 陈金刚, 等. 融合注意力机制的多损失联合跨模态行人重识别方法[J]. 激光与光电子学进展, 2022, 59(8): 0810010.
Wang F S, Liu F R, Chen J G, et al. Multi-loss joint cross-modality person re-identification method integrating attention mechanism[J]. Laser & Optoelectronics Progress, 2022, 59(8): 0810010.

[5] Tian Y M, Li Q, Wang D, et al. Robust joint learning network: improved deep representation learning for person re-identification[J]. Multimedia Tools and Applications, 2019, 78(17): 24187-24203.

[6] 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.

[7] Mudunuri S P, Venkataramanan S, Biswas S. Dictionary alignment with re-ranking for low-resolution NIR-VIS face recognition[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(4): 886-896.

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

[9] Zhong X, Lu T Y, Huang W X, et al. Visible-infrared person re-identification via colorization-based Siamese generative adversarial network[C]//ICMR '20: Proceedings of the 2020 International Conference on Multimedia Retrieval, June 8-11, 2020, Dublin, Ireland. New York: ACM Press, 2020: 421-427.

[10] Wang G A, Zhang T Z, Yang Y, et al. Cross-modality paired-images generation for RGB-infrared person re-identification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12144-12151.

[11] Wang G A, Zhang T Z, Cheng J, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 3622-3631.

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

[13] Zhang W, He X Y, Lu W Z, et al. Feature aggregation with reinforcement learning for video-based person re-identification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(12): 3847-3852.

[14] Oreifej O, Mehran R, Shah M. Human identity recognition in aerial images[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 709-716.

[15] Jüngling K, Bodensteiner C, Arens M. Person re-identification in multi-camera networks[C]//CVPR 2011 Workshops, June 20-25, 2011, Colorado Springs, CO, USA. New York: IEEE Press, 2011: 55-61.

[16] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of Machine Learning Research, 2009, 10: 207-244.

[17] Liao S C, Hu Y, Zhu X Y, et al. Person re-identification by Local Maximal Occurrence representation and metric learning[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 2197-2206.

[18] Li W, Zhao R, Xiao T, et al. DeepReID: deep filter pairing neural network for person re-identification[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 152-159.

[19] Wang J Y, Zhu X T, Gong S G, et al. Transferable joint attribute-identity deep learning for unsupervised person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 2275-2284.

[20] Zhang X, Luo H, Fan X, et al. AlignedReID: surpassing human-level performance in person re-identification[EB/OL]. (2017-11-22)[2022-02-04]. https://arxiv.org/abs/1711.08184.

[21] Zheng F, Deng C, Sun X, et al. Pyramidal person re-IDentification via multi-loss dynamic training[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 8506-8514.

[22] Wang Z X, Wang Z, Zheng Y Q, et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 618-626.

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

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

[25] Lu Y, Wu Y, Liu B, et al. Cross-modality person re-identification with shared-specific feature transfer[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 13376-13386.

[26] Dai P Y, Ji R R, Wang H B, et al. Cross-modality person re-identification with generative adversarial training[C]//IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence, July 13-19, 2018, Stockholm, Sweden. New York: ACM Press, 2018: 677-683.

[27] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[28] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[29] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 3141-3149.

[30] Wang C, Zhang Q, Huang C, et al. Mancs: a multi-task attentional network with curriculum sampling for person re-identification[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11208: 384-400.

[31] Gui S J, Zhu Y, Qin X X, et al. Learning multi-level domain invariant features for sketch re-identification[J]. Neurocomputing, 2020, 403: 294-303.

[32] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[33] Ye M, Wang Z, Lan X Y, et al. Visible thermal person re-identification via dual-constrained top-ranking[C]//IJCAI'18: Proceedings of the 27th International Joint Conference on Artificial Intelligence, July 13-19, 2018, Stockholm, Sweden. New York: ACM Press, 2018: 1092-1099.

[34] Ye M, Lan X Y, Leng Q M. Modality-aware

collaborative learning for visible thermal person re-identification[C]//MM '19: Proceedings of the 27th ACM International Conference on Multimedia, October 21-25, 2019, Nice, France. New York: ACM Press, 2019: 347-355.

[35] Feng Z X, Lai J H, Xie X H. Learning modality-specific representations for visible-infrared person re-identification [J]. IEEE Transactions on Image Processing, 2020, 29: 579-590.

[36] Lin J W, Li H. HPILN: a feature learning framework for cross-modality person re-identification[EB/OL]. (2019-06-07)[2021-05-06]. https://arxiv.org/abs/1906.03142.

[37] Basaran E, Gökmen M, Kamasak M E. An efficient framework for visible-infrared cross modality person re-identification[J]. Signal Processing: Image Communication, 2020, 87: 115933.

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

[39] Ye M, Lan X Y, Leng Q M, et al. Cross-modality person re-identification via modality-aware collaborative ensemble learning[J]. IEEE Transactions on Image Processing, 2020, 29: 9387-9399.

[40] Liu H J, Cheng J, Wang W, et al. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification[J]. Neurocomputing, 2020, 398: 11-19.