基于混合卷积增强和内容感知注意力的跨模态行人重识别

杨真真[1]，吴心怡[1]

(1. 南京邮电大学理学院，江苏南京 210023)

摘要：跨模态行人重识别作为计算机视觉领域的研究热点，旨在解决不同成像条件下的行人匹配问题。现有研究着重于提取模态共享特征，但不能充分挖掘鉴别行人身份至关重要的细节特征。为了解决该问题，提出了一种基于混合卷积增强和内容感知注意力(Hybrid Convolutional Enhancement and Content-aware Attention, HCECA)的跨模态行人重识别方法，旨在提取更富含细节信息的行人特征。具体来说，首先在主干网络中嵌入混合卷积增强(Hybrid Convolutional Enhancement, HCE)模块，捕获更丰富的跨模态特征表示，提高特征的区分度和鲁棒性；然后，通过内容感知注意力(Content-aware Attention, CA)模块来挖掘丰富的细节信息，以提升行人特征的区分性。最后，在 SYSU-MM01 和 RegDB 数据集上进行了实验，提出的 HCECA 在 SYSU-MM01 数据集的全搜索模式下，Rank-1 和 mAP 分别达到 72.21%和 81.84%，在 RegDB 数据集上可见-红外模式下，Rank-1 和 mAP 分别达到 92.23%和 85.08%，均优于现有的跨模态行人重识别方法。

关键词：行人重识别；跨模态；注意力机制；混合卷积增强；内容感知注意力

中图分类号：TP391　　**文献标识码：**A

Hybrid Convolutional Enhancement and Content-aware Attention for Cross-modality Person Re-Identification

YANG Zhenzhen[1], WU Xinyi[1]

(1. College of Science, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China)

Abstract: Cross-modality person re-identification (Re-ID), as a research hotspot in the field of computer vision, aims to solving the challenge of matching pedestrians across varying imaging conditions. Existing methods focus on extracting modality-shared features, but they fail to fully mine the detailed features that are crucial for discriminative person identities. To address this issue, a hybrid convolutional enhancement and content-aware attention (HCECA) for cross-modality person re-identification is proposed, which aims to extract pedestrian features with more detailed information. Specifically, a hybrid convolutional enhancement (HCE) module is embedded in the backbone network to capture richer cross-modality feature representation, enhancing the distinctiveness and robustness of the features. In addition, a content-aware attention (CA) module is employed to mine rich detailed information, thereby improving the discriminability of pedestrian features. Finally, experiments are performed on the SYSU-MM01 and RegDB datasets. The proposed HCECA attains the Rank-1 accuracy of 72.21% and the mAP of 81.84% in the all-search mode on the SYSU-MM01 dataset, while achieving the Rank-1 accuracy of 92.23% and the mAP of 85.08% in the visible-infrared mode on the RegDB dataset. Both results outperform those of current cross-modality person re-identification methods.

Key words: person re-identification; cross-modality; attention mechanism; hybrid convolutional enhancement; content-aware attention

引言

随着城市化进程的加快以及公众安全意识的提升，智能监控系统在各类场所得到了广泛运用。行人重识别(Person Re-identification, Re-ID)[1][3]技术是智能监控系统的重要组成部分，旨在不同的摄像机下匹配同一身份的行人。传统的可见光摄像头在光线充足的情况下能提供高质量的图像，但是在光线不足的情况下，性能会急剧下降，造成图像的细节丢失、模糊等问题。红外摄像头能够弥补这一缺陷，它通过捕获物体放出的红外辐射来成像，不受光线的影响，能够实施全天候监控。因此可见-红外跨模态行人重识别技术顺势而生。

收稿日期：　　；**修回日期：**　　；**本刊网址：**

基金项目：国家自然科学基金(62071242, 62171232)，江苏省研究生科研与实践创新计划项目 (SJCX23_0251).

作者简介：杨真真，女，博士，副教授，yangzz@njupt.edu.cn

相对于单模态可见光行人重识别，可见-红外跨模态行人重识别不仅面临着可见光图像和红外图像之间存在的巨大模态差异，而且存在着单模态行人重识别中的行人姿态变化、背景杂波、遮挡等问题[5]，极具挑战性。为了解决上述问题，研究者对此进行了大量研究，Wu 等[6]设计了一种深度零填充方法提取模态共享特征，Wang 等[7]通过生成对抗网络生成伪图像来缩小特征的分布差异，Zhang 等[8]通过生成两种模态图像的统一中间模态图像来缩小模态差异。然而，上述方法主要集中提取模态共享特征，缩小模态差异，对细节信息挖掘不够充分，未能提取具有鉴别性的行人特征。

针对上述问题，本文提出了一种基于混合卷积增强和内容感知注意力(Hybrid Convolutional Enhancement and Content-aware Attention, HCECA)的跨模态行人重识别方法。该方法通过集成混合卷积增强(Hybrid Convolution Enhancement, HCE)模块和内容感知注意力(Content-aware Attention, CA)模块，实现对跨模态特征的高效提取。具体来说，首先采用混合卷积增强模块增强对中层特征的提取能力，捕获更丰富的跨模态特征表示，提高特征的区分度和鲁棒性。然后，采用内容感知注意力模块挖掘行人的细节信息，加强提取的行人特征的鉴别性。与现有跨模态行人重识别方法相比，提出的 HCECA 在 SYSU-MM01 数据集和 RegDB 数据集上的性能均最优，充分体现了提出方法的有效性和优越性。

2. 相关工作

2.1 可见-红外跨模态行人重识别

可见-红外跨模态行人重识别旨在匹配可见光与红外图像中的同一个行人。鉴于可见光图像和红外图像在成像机制和颜色等信息上存在着本质区别，传统的单模态行人重识别方法无法直接用于跨模态场景，这使得可见-红外跨模态行人重识别成了近几年计算机视觉领域的研究热点。Wu 等[6]提出了 SYSU-MM01 可见-红外跨模态行人重识别数据集，并设计了一种深度零填充的方法实现了可见-红外跨模态的行人匹配。Wang 等[7]提出一种端到端的对齐生成对抗网络，来实现特征的对齐，提高模型的性能。Zhang 等[8]设计了一个中间模态生成器，将可见和红外两个模态的图像投影到统一的中间模态，有效减小了两个模态之间的差异。Ye 等[9]设计了一种新的通道交换增强方法，有效提高了跨模态行人重识别的准确性。Wang 等[10]通过生成跨模态配对图像的方法，对集级和实例级特征进行约束，缩小跨模态之间的差异。Yang 等[11]从噪声标签的角度对可见-红外行人重识别展开了研究，设计了双鲁棒训练方法，有效提高了识别精度。Lu 等[12]提出了渐进模态共享 Transformer 的深度学习框架，采用两阶段学习来减小两个模态的差距。Huang 等[13]提出了多级双流模态共享特征提取子网络，在共享特征中提取模态共享外观特征和模态不变关系特征，提高了识别性能。上述研究主要集中于提取模态共享特征，减小两个模态之间的差距，但是未能充分挖掘模态内部特征。Yang 等[14]提出了一种模态共享特征协同分离方法，由一个显著响应模块和一个协同分离模块组成，以减轻模态之间的差异。该方法采用双多层感知机来分离模态共享特征和模态特定特征，虽然在一定程度上提高了特征的辨识度，但可能会导致某些细微的共享特征被误分类为特定特征，或者在分离过程中丢失。

2.2 注意力机制

注意力机制的核心在于赋予模型动态聚焦的能力，使模型能够有选择性地关注输入的不同部分，为每个部分分配不同的权重，突出关键信息。Hu 等[15]创新性地提出挤压激励(Squeeze-and-Excitation, SE)模块，通过显式建模通道之间的相互依赖关系，显著提高了模型性能。Woo 等[16]将通道注意力与空间注意力结合，设计了卷积块注意力模块(Convolutional Block Attention Module, CBAM)。Agarwal 等[17]提出跳跃注意力模块(Skip Attention Module, SAM)，使用窗口基交叉注意力机制细化像素查询，显著提高了模型精度和泛化能力。Misra

等[18]提出了一种捕获跨维度交互的三元组注意力 (Triplet Attention, TA)机制，在多个任务上的性能均得到了显著提升。Wang 等[19]提出了内容感知混合器(Content-Aware Mixer, CAMixer)，通过结合卷积和自注意力，并引入内容感知的混合策略，有效提高了图像超分辨率任务的性能和计算效率。在行人重识别领域，注意力机制也得到了广泛应用。Ye 等[20]设计了模态内注意力和跨模态注意力，提出了动态双注意力聚合特征学习，有效提高了模型性能。Zhang 等[21]提出了关系感知全局注意力(Relation-Aware Global Attention, RGA)，包括空间 RGA-S 和通道 RGA-C，挖掘全局范围内的关系信息，有效提升了模型性能。Yang 等[22]提出全局注意力，增强特征提取能力，有效提高模型的性能。Ye 等[1]在提出的基线模型中嵌入了非局部注意力机制，弥补了卷积神经网络局部感知性的局限性，有效提高了行人重识别精度。然而上述的注意力机制或是使用自注意力机制，忽略了空间通道的信息，或是只考虑全局空间或通道信息，或是强调细节特征，忽略了行人的其他有效信息。为解决上述问题，本文设计了内容感知注意力模块，旨在促使网络在捕捉全局特征的同时，深入挖掘更多细节特征，进而提升模型的整体识别性能。

3. 提出的跨模态行人重识别方法

3.1 整体框架

提出的 HCECA 的整体框架如图 1 所示，由特征提取器、HCE 和 CA 构成。在将图像输入特征提取器之前，首先，将可见光和红外两个模态的数据输入中间模态生成器(Middle Modality Generator, MMG)中，生成中间模态图像。然后，将原始图像和中间模态图像输入到双流 ResNet-50 特征提取器中，提取行人特征。其中，第一个卷积块的参数不共享，用于提取模态特定特征。后面四个残差块的参数共享，用于提取模态共享特征。在第二个残差块之后嵌入混合卷积增强(HCE)模块，增强对中层特征的提取能力，以捕获更丰富的跨模态特征表示，提高特征的区分度和鲁棒性。在第三个残差块后嵌入内容感知注意力(CA)模块，以提取细节特征，加强提取的行人特征的区分性。最后，将从特征提取器提取的特征进行水平划分成四份，将划分后的特征进行平均池化(Average Pooling, AP)，并输入到批归一化层(BatchNorm Layer, BN)，同时将水平池化后的特征使用身份损失和一致性损失进行训练，并采用三元组损失对批归一化之后的特征进行训练。

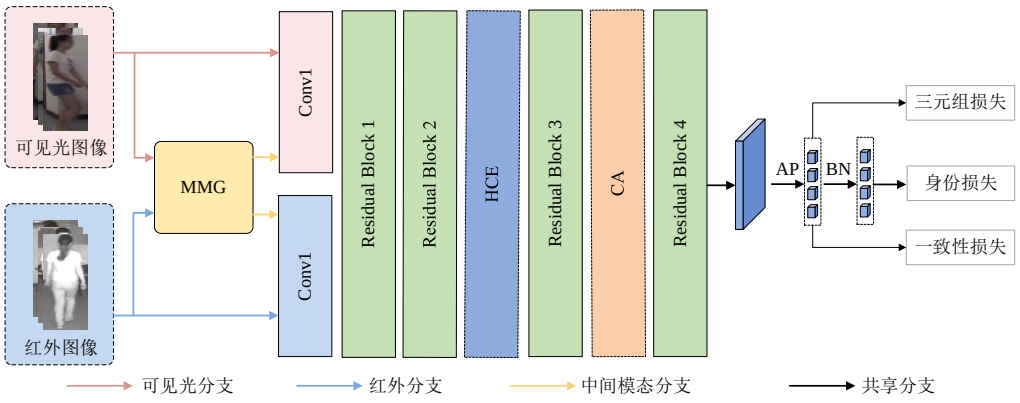


图 1 提出的 HCECA 整体框架

Figure 1 Overall framework of the proposed HCECA

3.2 混合卷积增强模块

如何有效缩小不同模态之间的特征差异，提升模型的识别性能，是跨模态行人重识别任务研究的重点。为了解决上述难题，设计高效且鲁棒的特征增强模块，成为提升跨模态行人重识别性能的关键所在。

传统的卷积神经网络在处理跨模态数据时，具有一定的局限性。单一类型的卷积操作难以全面且高效地捕获跨模态数据中的复杂特征。为了克服这个问题，受文献[23][24]启发，本文设计一种新的混合卷积增强模块，旨在进一步提炼和增强特征表示。该模块主要由多种类型的卷积操作和激活函数构成，其中卷积操作包括点卷积、深度卷积、深度可分离卷积和并行路径卷积，激活函数包括 ReLU 函数和 GeLU 函数。通过点卷积和7×7的深度卷积，可以捕获大尺度的特征，捕获丰富的上下文信息。在获得初步增强的特征之后，对其进行并行卷积操作，通过不同尺度的卷积操作，捕获多样化的特征表示。混合卷积增强模块被放置在第二层残差块后，以充分利用前期网络层提取的行人特征，并进一步增强模型对跨模态数据共享特征的捕捉能力，并减少模态噪声的干扰。具体结构如图 2 所示。

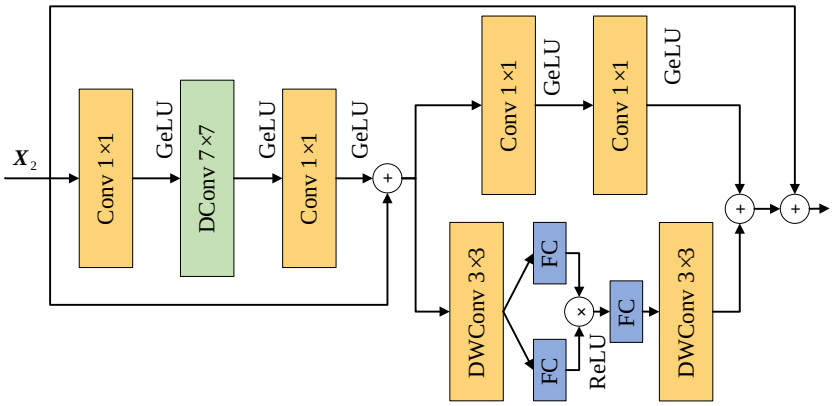


图 2 混合卷积增强模块

Figure 2 Hybrid Convolutional Enhancement Module

首先，给定的第二层残差块输出特征 $\boldsymbol{X}_2$ 通过1×1卷积和7×7的深度卷积，得到初步增强的中间特征 $\boldsymbol{X}_m$，具体过程如下：

$$\boldsymbol{X}_m=\mathrm{GeLU}(\varphi_{1\times1}(\mathrm{GeLU}(\varphi_{dc7\times7}(\mathrm{GeLU}(\varphi_{1\times1}(\boldsymbol{X}_2)))))+\boldsymbol{X}_2 \tag{1}$$

其中，$\varphi_{1\times1}$是1×1卷积操作，$\varphi_{dc7\times7}$是7×7深度卷积操作，GeLU 是激活函数。

然后，将中间特征 $\boldsymbol{X}_m$ 分别输到1×1和3×3卷积分支，通过并行的不同尺度卷积操作以增强特征表示能力。在1×1卷积分支，中间特征经过1×1卷积和 GeLU 激活函数，得到输出特征 $\boldsymbol{X}_{1\times1}$：

$$\boldsymbol{X}_{1\times1}=\mathrm{GeLU}(\varphi_{1\times1}(\mathrm{GeLU}(\varphi_{1\times1}(\boldsymbol{X}_m)))) \tag{2}$$

在3×3卷积分支，中间特征经过深度可分离卷积和全连接层，得到输出特征 $\boldsymbol{X}_{3\times3}$：

$$\boldsymbol{X}_{3\times3}=\varphi_{dw3\times3}(\mathrm{FC}_3(\mathrm{ReLU}(\mathrm{FC}_1(\varphi_{dw3\times3}(\boldsymbol{X}_m)))\cdot\mathrm{FC}_2(\varphi_{dw3\times3}(\boldsymbol{X}_m)))) \tag{3}$$

其中，$\varphi_{dw3\times3}$是3×3深度可分离卷积操作，FC 是全连接层，ReLU 是激活函数。

最后，将 $\boldsymbol{X}_{1\times1}$， $\boldsymbol{X}_{3\times3}$ 和 $\boldsymbol{X}_2$ 相加融合，得到最终增强的特征，过程如下：

$$\boldsymbol{X}_{\mathrm{E}}=\boldsymbol{X}_{1\times1}+\boldsymbol{X}_{3\times3}+\boldsymbol{X}_{2} \tag{4}$$

3.3 内容感知注意力模块

在跨模态行人重识别中，行人的细微特征对跨模态特征的匹配至关重要。传统的注意力机制，仅从单一的维度挖掘特征的重要性，难以同时捕获细节性特征和全局语义之间的关联，导致关键判别特征的丢失。为解决这一问题，受 Chen 等[25]启发，本文提出内容感知注意力模块，其结构如图 3 所示。该模块通过卷积层、AReLU、通道注意力、空间注意力和像素注意力，形成多层次的特征感知机制。首先利用卷积和 AReLU 激活函数提取特征，接着分别通过通道注意力和空间注意力，以捕获不同通道的重要性和定位特征图中的关键区域，最后采用像素注意力实现细粒度特征校准，提高对行人细节信息的捕获能力。

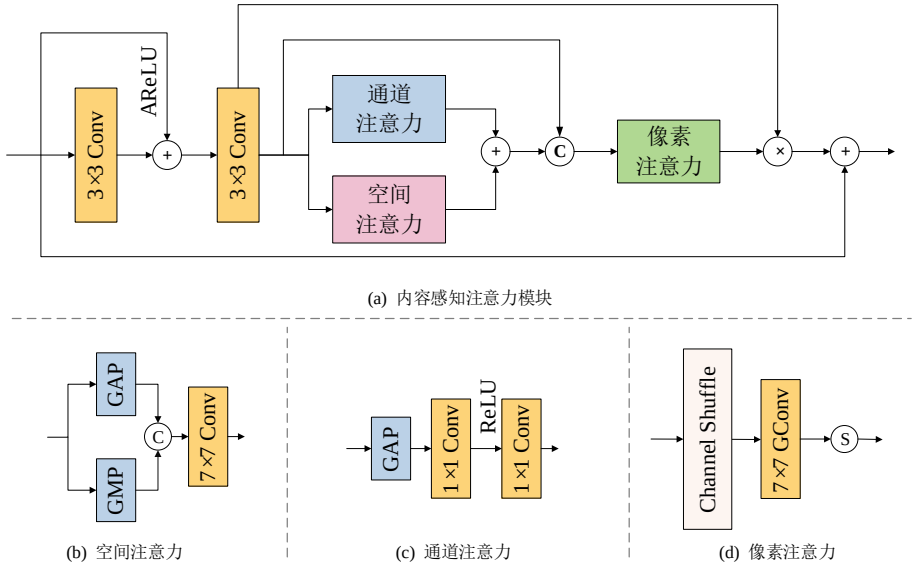


图 3 内容感知注意力模块

Figure 3 Content-aware attention module

首先，给定的第三个残差块的输出特征 $\boldsymbol{X}_{3}$ 通过两个 3×3 卷积，提取特征 $\boldsymbol{X}_{\mathrm{conv}}$：

$$\boldsymbol{X}_{\mathrm{conv}}=\varphi_{3\times3}\left(\mathrm{AReLU}\left(\varphi_{3\times3}\left(\boldsymbol{X}_{3}\right)\right)+\boldsymbol{X}_{3}\right) \tag{5}$$

其中，AReLU 是在 ReLU 的基础上进行一定的修改的注意力激活函数[26]，具体公式如下：

$$\mathrm{AReLU}\left(\varphi_{3\times3}\left(\boldsymbol{X}_{3}\right)\right)=\beta\mathrm{ReLU}\left(\varphi_{3\times3}\left(\boldsymbol{X}_{3}\right)\right)-\alpha\mathrm{ReLU}\left(-\varphi_{3\times3}\left(\boldsymbol{X}_{3}\right)\right) \tag{6}$$

其中，α 和 β 是可学习参数，是通过训练过程自动学习最优值，以调整正负激活值的重要性。具体来说，根据文献[26]，设置 α 的初始值为 0.9，并将其限制在 $[0.01,0.99]$ 范围内，以避免模型敏感度失衡。设置 β 的初始值为 2.0，并将其限制在 $(1,2)$ 范围内，以确保模型的正激活值得到合理增强。

其次，将 $\boldsymbol{X}_{\mathrm{conv}}$ 输入空间注意力分支和通道注意力分支，分别得到空间注意力特征图 $\boldsymbol{A}_{\mathrm{S}}$ 和通道注意力特征图 $\boldsymbol{A}_{\mathrm{C}}$，其格式分别如下：

$$\begin{cases} \boldsymbol{A}_{\mathrm{S}} = \varphi_{7\times7}\left(\left[\boldsymbol{X}_{\mathrm{conv}}^{\mathrm{GAP}}, \boldsymbol{X}_{\mathrm{conv}}^{\mathrm{GMP}}\right]\right) \\ \boldsymbol{A}_{\mathrm{C}} = \varphi_{1\times1}\left(\max\left(0, \varphi_{1\times1}\left(\boldsymbol{X}_{\mathrm{conv}}^{\mathrm{GAP}}\right)\right)\right) \end{cases} \tag{7}$$

其中，$\varphi_{7\times7}$ 是核为 7×7 的卷积，GAP 是全局平均池化，GMP 是全局最大池化。

再次，将空间注意力特征图 $\boldsymbol{A}_{\mathrm{S}}$ 和通道注意力特征图 $\boldsymbol{A}_{\mathrm{C}}$ 进行融合，得到如下的粗粒度的注意力特征图 $\boldsymbol{A}_{\mathrm{coarse}}$：

$$\boldsymbol{A}_{\mathrm{coarse}} = \boldsymbol{A}_{\mathrm{S}} + \boldsymbol{A}_{\mathrm{C}} \tag{8}$$

并将粗粒度的注意力特征图与 $\boldsymbol{X}_{\mathrm{conv}}$ 进行拼接：

$$\boldsymbol{X}_{\mathrm{c}} = \left[\boldsymbol{X}_{\mathrm{conv}}, \boldsymbol{A}_{\mathrm{coarse}}\right] \tag{9}$$

然后，将拼接后的特征输入到像素注意力中，得到如下的细化特征：

$$\boldsymbol{X}_{\mathrm{fined}} = \sigma\left(\varphi_{G7\times7}\left(\mathrm{CS}\left(\boldsymbol{X}_{\mathrm{c}}\right)\right)\right) \cdot \boldsymbol{X}_{\mathrm{conv}} \tag{10}$$

其中，σ 是 sigmoid 函数，$\varphi_{G7\times7}$ 是核为 7×7 的组卷积，CS 为通道混洗操作。

最后，通过一个残差连接，与 $\boldsymbol{X}_3$ 相加，得到最终的特征：

$$\boldsymbol{X}_{\mathrm{CAM}} = \boldsymbol{X}_{\mathrm{fined}} + \boldsymbol{X}_3 \tag{11}$$

3.4 损失函数

本文采用标签平滑交叉熵损失[8]、三元组损失以及一致性分布损失[8]的联合损失优化来对模型进行训练。标签平滑交叉熵损失函数是为了缓解传统交叉熵损失函数在处理 one-hot 编码标签时，可能会导致模型对训练数据过拟合的问题。标签平滑的核心思路是将原本真实标签的 one-hot 编码中值为 1 的部分替换成一个较小的值 $1-\frac{N-1}{N}\varepsilon$，同时将其余的概率值平均分配给其他所有类别，使得模型在训练过程中就不会过度关注某个特定的类别，而是对多个类别保持一定的不确定性。该损失函数具体表达式为：

$$L_{id} = \begin{cases} \sum_{i=1}^{N} -\left(1-\frac{N-1}{N}\varepsilon\right) log\left(p_i\right), & i = y \\ \sum_{i=1}^{N} -\left(\frac{1}{N}\varepsilon\right) log\left(p_i\right), & i \neq y \end{cases} \tag{12}$$

其中，N 代表训练集中行人的总数，y 为行人的身份标签，p_i 代表行人被预测为第 i 类的概率，ε 是一个常数，用于在训练集上限制模型的显著性，在实验中，ε 设置为 0.1。

分布一致性损失用来约束两个模态生成的中间模态特征 $f_{M_V}^i$，$f_{M_I}^i$ 之间的差距，其表达式如下：

$$L_{dcl}=\frac{1}{B}\sum_{1}^{B}\text{mean}\left(f_{M_V}^{i}-f_{M_I}^{i}\right) \tag{13}$$

其中，B 是训练过程中一个小批中生成的中间模态图像的数量，$\text{mean}\left(f_1-f_2\right)$ 表示特征 f_1 和 f_2 差值的平均值。通过分布一致性损失的优化，两个中间模态特征的相似度会最大化。

此外，对可见模态图像V，红外模态图像I,可见光图像生成的中间模态图像M_V,红外图像生成的中间模态图像M_I中每两两模态之间进行了三元组损失约束，以最大化类间样本距离，最小化类内样本距离。每个模态中有M张图像，其中，第 1 张到第M张是可见光图像，第$M+1$到第$2M$张是红外图像，第$2M+1$至第$3M$张是可见光图像生成的中间模态图像，第$3M+1$张至第$4M$张是红外图像生成的中间模态图像，以可见光图像与红外图像的损失为例，损失表达式如下：

$$L_{tri}^{(V,I)}=\sum_{i=1}^{M}\left[\rho+\max_{\substack{j=M+1,\ldots 2M\\ i=j}}d_{ij}-\min_{\substack{k=M+1,\ldots 2M\\ i\neq k}}d_{ik}\right]_{+} \tag{14}$$

$$L_{tri}^{(I,V)}=\sum_{i=M+1}^{2M}\left[\rho+\max_{\substack{j=1,\ldots M\\ i=j}}d_{ij}-\min_{\substack{k=1,\ldots M\\ i\neq k}}d_{ik}\right]_{+} \tag{15}$$

$$L_{tri}^{V,I}=L_{tri}^{(V,I)}+L_{tri}^{(I,V)} \tag{16}$$

其中，d_{ij}是可见光图像和红外图像之间的欧几里得距离，ρ是阈值，$\left[d\right]_{+}=max\left(d,0\right)$，$i$与$j$属于同一个行人身份，$k$与其身份不同。

总的三元组损失为：

$$L_{tri}=L_{tri}^{V,I}+L_{tri}^{V,M_I}+L_{tri}^{I,M_V}+L_{tri}^{M_I,M_V} \tag{17}$$

由分布一致性损失、身份损失、三元组损失构成的总目标优化损失如下：

$$L_{total}=L_{id}+\eta_1 L_{dcl}+\eta_2 L_{tri} \tag{18}$$

其中，η_1和η_2是权重参数，在实验中，η_1=1，η_2=0.5。

4. 实验及结果分析

4.1 数据集介绍

本文在两个广泛使用的公开数据集 SYSU-MM01[6]和 RegDB[27]上进行实验，以评估提出的 HCECA 的性能。SYSU-MM01 数据集是由四个可见光相机和两个红外相机在白天和黑夜收集的大型可见-红外行人重识别数据集，共包含了 491 个不同行人的身份识别信息，涵盖室内和室外两个不同场景。其中 395 个行人身份的图像用于训练，另外 96 个行人身份的

图像用于测试。本实验通过全搜索以及室内搜索两种模式，对提出的方法进行性能上的评估。RegDB 数据集是由可见光-红外双摄头相机拍摄，包含 412 个不同行人的身份识别信息的可见-红外跨模态行人重识别数据集，每个行人各包含 10 张可见光和红外图像。随机选取 206 个行人身份图像用于训练，另外 206 个用于测试，实验结果是十次测试结果的平均值。

4.2 评价指标

本实验所选取的性能评估指标为：标准累计匹配特性(Cumulative Matching Characteristics，CMC)曲线所得的 Rank-k 识别率，以及平均精度均值(Mean Average Precision，mAP)。其中，Rank-k 衡量的是在前 k 次检索中，正确人物图像被检索出来的概率，mAP 用于评估检索系统在图库中存在多个可能匹配项的整体性能表现。

4.3 实验设置

本文基于 Pytorch 框架实现，并利用 NVIDIA 3090 GPU 完成模型的训练过程。采用在 ImageNet 上训练的 Resnet-50 作为骨干网络，所有输入的图像统一被调整为 3×384×192，在训练时，通过应用随机水平翻转和随机擦除两种技术来增强数据。在训练中采用动态调整学习率策略，初始学习率设置为10^{-2}，在 10 轮内线性增加到10^{-1}，在第 20 轮衰减到10^{-2}，在第 60 轮衰减到10^{-3}，共训练 80 轮。使用随机梯度下降法进行优化，动量参数设置为 0.9。

4.4 对比试验

将提出的 HCECA 与现有先进的跨模态行人重识别方法进行对比，对比方法包括：AGW[1]，Zero-padding[6]，AlignGAN[7]，MMN[8]，CAJ[9]，JSIA[10]，DART[11]，PMT[12]，MTMFE[13]和 MFSCS[14]。其中，AGW，Zero-padding，MMN，DART，PMT 是在服务器上跑的结果，AlignGAN，CAJ，JSIA，MTMFE 和 MFSCS 是原论文中的结果。其实验结果如表 1 所示。

表 1 SYSU-MM01 和 RegDB 数据集上的性能对比 (%)

Table 1 Performance comparison on SYSU-MM01 and RegDB datasets (%)

数据集	SYSU-MM01				RegDB			
方法	全搜索		室内搜索		可见-红外		红外-可见	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Zero-padding[6]	16.70	19.10	21.45	32.10	18.76	19.21	17.34	18.47
JSIA[10]	38.10	36.90	43.80	52.90	48.10	48.90	48.50	49.30
AlignGAN[7]	42.40	40.70	45.90	54.30	57.90	53.60	56.30	53.40
AGW[1]	47.95	47.44	53.12	62.09	76.19	68.87	73.96	67.06
MTMFE[13]	**62.56**	**60.57**	**65.06**	**73.86**	**76.10**	**74.39**	**72.18**	**71.04**
PMT[12]	67.53	64.98	71.66	76.52	84.83	76.55	84.16	75.13
DART[11]	68.79	66.55	72.52	78.17	83.78	76.00	81.78	73.64
MMN[8]	69.63	65.72	75.96	76.99	91.68	83.97	87.30	80.27
CAJ[9]	**69.88**	**66.89**	**76.26**	**80.37**	**85.03**	**79.14**	**84.75**	**77.82**
MFSCS[14]	70.59	67.49	75.98	80.24	85.34	76.39	83.88	75.16
HCECA	**72.21**	**69.89**	**78.32**	**81.84**	**92.23**	**85.08**	**88.17**	**81.75**

如表 1 所示，本文提出的 HCECA 在 SYSU-MM01 和 RegDB 数据集上均显著优于对比方法。具体来说，在 SYSU-MM01 数据集上，HCECA 在全搜索模式下的 Rank-1 和 mAP 分别为 72.21%和 69.89%，相比于性能次优的 MFSCS，Rank-1 和 mAP 分别高出 1.62%和 2.4%，MFSCS 通过特征分离，将特征分为模态共享和模态特定，但是忽略了细节信息的捕获， 提出的 HCECA 通过混合卷积增强丰富了特征的表达能力，通过内容感知注意力挖掘行人细节

信息，提取行人的鉴别性特征，提高了模型识别性能。在室内搜索模式下，HCECA 的 Rank-1 和 mAP 分别为 78.32%和 81.84%，比 CAJ 分别高出了 2.06%和 1.47%。在 SYSU-MM01 数据集上两种搜索模式的对比实验均证明了提出的 HCECA 的优越性。

此外，为了验证 HCECA 的泛化能力和优越性，在 RegDB 数据集上也进行对比实验。在可见-红外模式下，HCECA 的 Rank-1 和 mAP 分别达到 92.23%和 85.08%，相较于次优的 MMN，Rank-1 和 mAP 分别高出 0.55%和 1.11%。在红外-可见模式下，HCECA 的 Rank-1 和 mAP 分别达到 88.17%和 81.75%，分别高出次优的 MMN 方法 0.87%和 1.48%。MMN 通过生成中间模态图像来缩小模态间差异，但是忽略了对行人鉴别性的细节信息的挖掘，提出的 HCECA 的混合卷积增强模块通过不同类型、不同大小的卷积丰富了提取特征的表达能力，而且进一步采用内容感知注意力挖掘行人的细节特征，提高模型性能。在 RegDB 数据集上两种模式下的对比实验体现了提出的 HCECA 的优越性。

上述两个数据集上的对比实验充分证明了提出的 HCECA 的优越性，并且泛化能力良好。

4.5 消融实验

为了验证所提出的 HCE 和 CA 模块的性能，在 SYSU-MM01 数据集上实施了消融实验。以 ResNet-50、中间模态生成器和损失函数组成的 MMN 网络作为基线方法，并在此基础上每次只增加一个模块，来评估 HCE 和 CA 模块的性能。构建的网络有 MMN+HCE、MMN+CA 和 MMN+HCE+CA 三种。消融实验的结果如表 2 所示：

表 2 SYSU-MM01 数据集上每个模块的性能 (%)

Table 2 Performance of each module on SYSU-MM01 dataset (%)

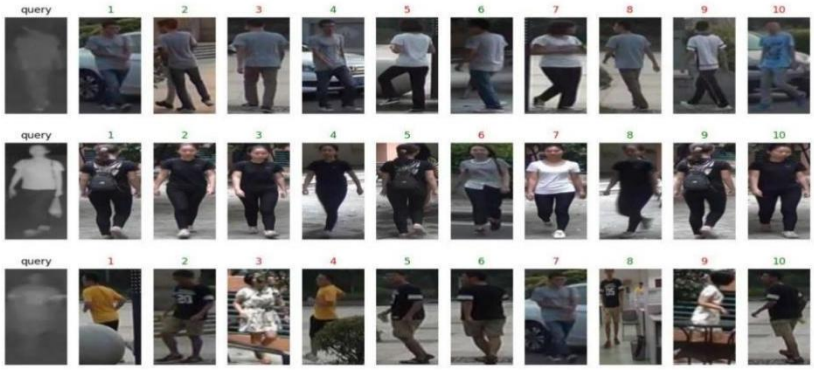
方法	全搜索模式			
	Rank-1	Rank-10	Rank-20	mAP
MMN	69.63	96.07	98.95	65.72
MMN+HCE	71.60	96.44	98.87	66.53
MMN+CA	71.69	96.36	98.89	68.93
MMN+HCE+ CA	**72.21**	**96.52**	**99.80**	**69.89**

从表 2 中可以看出，当在基线模型 MMN 上添加 HCE 时，Rank-1、Rank-10 和 mAP 分别为 71.60%、96.44%和 66.53%，相较于 MMN，分别提升了 1.97%、0.37%和 0.81%，表明 HCE 可以帮助模型提取更丰富的特征，增强特征的表达能力。Rank-20 的性能有略微下降，表明 HCE 可能对 Rank-20 这样宽松的匹配条件有所影响。当在基线模型上添加 CA 时，Rank-1、Rank-10 和 mAP 分别为 71.69%、96.36%和 68.93%，分别提升了 2.06%，0.29%和 3.21%，说明 CA 能够有效捕获行人的细节特征，提高行人特征的鉴别能力。当在基线模型上同时添加 HCE 和 CA 模块时，模型的性能得到了进一步提升。Rank-1、Rank-10、Rank-20 和 mAP 分别为 72.21%、96.52%、99.80%和 69.89%，相较于 MMN，MMN+HCE，MMN+CA，四个指标的性能均得到了提升。相比于 MMN，Rank-1、Rank-10、Rank-20 和 mAP 分别提升了 2.58%、0.45%、0.85%和 4.17%。说明 HCE 和 CA 两者共同作用，可以表现得更加出色，HCE 能帮助模型挖掘更丰富的特征，CA 使得模型提取更全面、更具鉴别性的特征。实验结果验证了 HCE 和 CA 的有效性，并且两者共同作用对模型的性能有着积极的促进作用。

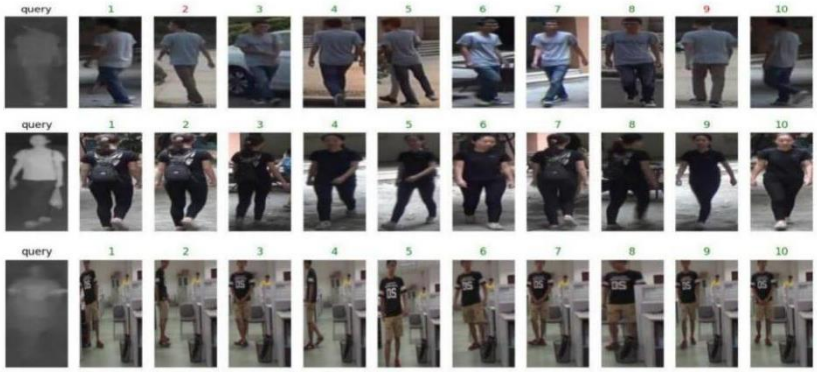
4.6 可视化分析

排序可视化：为了进一步分析提出的 HCECA 的有效性和优越性，将 HCECA 与基线方法 MMN 在 SYSU-MM01 数据集上进行排序可视化，并随机选取三组检索结果，行人检索结果如图 4 所示。图 4 第一列 query 图像为待检索的行人图像，编号为 1-10 的图像是按照余弦相似度排列的前十个检索结果。其中，正确检索的行人图像被标记为绿色，错误的检索

结果则标记为红色。图 4(a)和 4(b)分别是 MMN 和提出的 HCECA 的检索结果。



(a) MMN



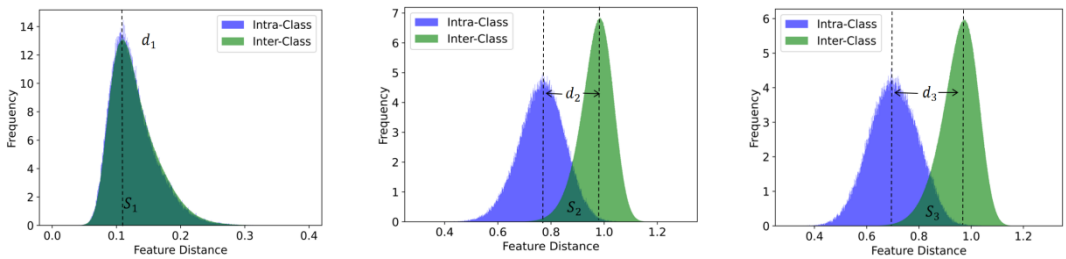
(b) HCECA

图 4　SYSU-MM01 排序结果可视化

Figure 4 Visualization of the Sorting Results for SYSU-MM01

由图 4 可知，SYSU-MM01 数据集的可见光图像和红外图像具有很大的模态差异，纹理信息、背景杂波等因素均对跨模态检索的结果造成了一定影响。从三组的检索结果看，基线 MMN 在前 10 个匹配结果中，很多图片检索错误。提出的 HCECA 在前 10 个匹配结果中，三组的检索结果只有两个是匹配错误的，HCECA 的准确率要高于 MMN。从第三组的检索结果看，HCECA 能提取到袖口图标这种细节性的特征，说明提出的 HCECA 通过引入混合卷积增强模块和内容感知注意力模块，捕获了更为丰富、更具细节信息的行人特征，从而提高了识别的准确性。

特征距离可视化：为了更进一步证明提出的 HCECA 的有效性，对类内-类间特征距离进行了可视化，结果如图 5 所示。图 5 (a)是初始化的距离，图 5 (b)是基线模型 MMN 的特征距离，图 5 (c)是提出的 HCECA 的特征距离。



(a) 初始化　　(b) MMN　　(c) HCECA

图 5 类内-类间特征距离可视化

Figure 5: Visualization of intra-class and inter-class feature distances

两个峰之间的距离 d 反映了正负样本对之间的分离程度。具体而言，d 值越大，说明正负对分离度越高，识别的结果更准确。从图 5 的特征距离可视化结果可以发现，两个峰之间的距离 $d_3 > d_2 > d_1$，说明在引入混合卷积增强模块和内容感知注意力模块后，HCECA 表现出了更强的识别能力，能够捕获并识别更具有区分性和细节性的行人信息。从而促使正负对样本的分离程度越高，进而提高了模型的识别精度。

4.7 复杂度分析

将提出的 HCECA 方法和性能较优的 MMN、PMT 以及 DART 进行复杂度分析的对比。将模型的参数量（单位/兆）作为复杂度分析的依据，在 SYSU-MM01 数据上的结果如表 3 所示。

表 3 复杂度对比结果

Table 3 Comparison results of complexity

方法	模型参数量(MB)	Rank-1	mAP
MMN	23.534	69.63	65.72
PMT	85.648	67.53	64.98
DART	23.55	68.79	66.55
HCECA	48.396	72.21	69.89

从表 3 可以看出，相比于 MMN 和 DART，提出的 HCECA 的参数量相对较高，但是相比基于 Transformer 的方法，其参数量少了将近一半。HCECA 的性能 Rank-1 和 mAP 均高于 MMN、PMT 和 DART，表明 HCECA 在处理跨模态行人重识别任务时想能的优越性，但是其复杂度较高，在未来需要对轻量化且高性能的模型进行研究。

5. 结束语

本文提出了一种基于混合卷积增强和内容感知注意力的跨模态行人重识别方法，旨在从不同模态的行人图像中提取更为精细且具有高度辨识性的特征，该网络由混合卷积增强模块和内容感知注意力模块构成。其中，混合卷积增强模块用来增强对中层特征的提取能力，捕获更丰富的跨模态特征表示。内容感知注意力模块用来提取行人的细节特征，加强提取的行人特征的区分性。实验结果显示，提出的 HCECA 方法的表现超过了当前的跨模态行人重识别方法。在未来，我们将进一步研究注意力引导的损失函数，通过设计能够直接指导注意力模块优化的损失函数，使注意力机制有效聚焦于最具区分性的特征上，以提高模型的重识别性能。

参考文献

[1] 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.

[2] 杨真真,邵静,杨永鹏,吴心怡.基于精确特征分布匹配和多域信息融合的跨域行人重识别[J]. 信号处理, 2023, 39(6):1099-1107.
Yang Zhenzhen, Shao Jing, Yang Yongpeng, Wu Xinyi. Cross-domain person re-identification based on accurate feature distribution matching and multi-domain information fusion[J]. Signal Processing, 2023, 39(6):1099-1107.

[3] Yang Z, Shao J, Yang Y. An improved CycleGAN for data augmentation in person re-identification[J]. Big Data Research, 2023, 34:1-10.

[4] 郝玲,段继忠,庞健.基于难样本混淆增强特征鲁棒性的行人重识别[J].数据采集与处理, 2022, 37(1):

122-133.

Hao Ling, Duan Jizhong, Pang Jian. Person Re-identification Based on Hard sample Confusion to Enhance Feature Robustness[J]. Journal of Data Acquisition and Processing, 2022, 37(1): 122-133.

[5] 杨真真,陈亚楠,杨永鹏,吴心怡.基于可学习掩模和位置编码的遮挡行人重识别[J].数据采集与处理, 2025, 40(01):217-229.

Yang Zhenzhen, Chen Yanan, Yang Yongpeng, Wu Xinyi. Occlusive person re-identification based on learnable mask and location encoding[J]. Journal of Data Acquisition and Processing, 2025, 40:1-14.

[6] Wu A, Zheng W S, Yu H X, et al. RGB-infrared cross-modality person re-identification[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 5380-5389.

[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/CVF International Conference on Computer Vision, 2019: 3622-3631.

[8] Zhang Y K, 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.

[9] Ye M, Ruan W, Du B, et al. Channel augmented joint learning for visible-infrared recognition [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021:13567-13576.

[10] Wang G A, Zhang T, Yang Y, et al. Cross-modality paired-images generation for RGB-infrared person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07): 12144-12151.

[11] Yang M, Huang Z, Hu P, et al. Learning with twin noisy labels for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022:14308-14317.

[12] Lu H, Zou X, Zhang P. Learning progressive modality-shared transformers for effective visible-infrared person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 1835-1843.

[13] Huang N, Liu J, Luo Y, et al. Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification[J]. Pattern Recognition, 2023, 135:109145.

[14] Yang X, Dong W, Li M, et al. Cooperative separation of modality shared-specific features for visible-infrared person re-identification[J]. IEEE Transactions on Multimedia, 2024, 26: 8172-8183.

[15] 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.

[16] Woo S, Park J, Lee JY, et al. Cbam: Convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision, 2018: 3-19.

[17] Agarwal A, Arora C. Attention everywhere: Monocular depth prediction with skip attention [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 5861-5870.

[18] Misra D, Nalamada T, Arasanipalai AU, et al. Rotate to attend: Convolutional triplet attention module[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021:3139-3148.

[19] Wang Y, Liu Y, Zhao S, et al. CAMixerSR: Only details need more attention[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 25837-25846.

[20] Ye M, Shen J, Crandall D, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]// Proceedings of the European Conference on Computer Vision, 2020: 229-247.

[21] Zhang Z, Lan C, Zeng W, et al. Relation-aware global attention for person re-identification [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 3186-3195.

[22] Yang Z, Chen Y, Yang Y, et al. Robust feature mining transformer for occluded person re-identification[J]. Digital Signal Processing, 2023, 141:1-12.

[23] Ma X, Dai X, Bai Y, et al. Rewrite the stars[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 5694-5703.

[24] Yang K, Hu T, Dai H et al. CRNet: A detail-preserving network for unified image restoration and enhancement task[EB/OL]. (2024-04-22). https://arxiv.org/abs/2404.14132.

[25] Chen Z, He Z, Lu ZM. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention[J]. IEEE Transactions on Image Processing, 2024, 33:1002-1015.

[26] Chen D, Li J, Xu K. AReLU: Attention-based rectified linear unit[EB/OL]. (2020-06-24)[2024-09-12]. https://arxiv.org/abs/2006.13858.

[27] Nguyen D, Hong H, Kim K, et al. Person recognition system based on a combination of body images from visible light and thermal cameras[J]. Sensors, 2017, 17(3): 605.

作者简介：

杨真真(1984-)，**通信作者**，女，副教授，研究方向：深度学习、计算机视觉，
E-mail: yangzz@njupt.edu.cn。

吴心怡(1999-)，女，硕士研究生，研究方向：深度学习、计算机视觉，
E-mail:1222087527@njupt.edu.cn