**基于辅助模态和注意力机制的跨模态行人重识别**

郭思琦，谭台哲，陈梓骏

（广东工业大学计算机学院电子信息系 广州 510000）

**摘要：**现有的跨模态行人重识别方法大多选择模态互转或直接将特征映射到共同特征空间，主要关注于缓解模态差异，而忽略同一模态下的类间差异，这不利于模型学习到具有辨别性的身份特征。本文提出一种基于改进的多头自注意力机制的非共享双流网络，分别提取可见光图像和红外图像的特征，并对特征做水平划分，使用局部特征来计算身份损失以及异质中心损失，利用多头自注意力机制根据像素本身的内容及其相对于相邻像素的空间位置来计算像素的注意力图，增强模型提取的信息容量，进而学习到同模态下不同类之间具有辨别性的身份特征。同时使用辅助模态来减缓可见光图像和红外图像在颜色信息上的差异，进一步减小模态差异。在SYSU-MM01和RegDB数据集上的mAP分别达到60.36%和76.10%，实验表明该方法的有效性。

**关键词：**跨模态；行人重识别；辅助模态；多头自注意力机制

Cross-modality person re-identification model based on auxiliary modal and attention mechanism

GUO SiQi, TAN TaiZhe,Chen Zi Jun

(Department of Electronic Information, School of Computer, Guangdong University of TechnologyGuangzhou Guangzhou 510000)

**Abstract** Most of the existing cross modal pedestrian re identification methods choose modal transformation or directly map features to the common feature space, mainly focusing on mitigating modal differences, while ignoring the differences between classes under the same mode, which is not conducive to the model learning to identify identity features. This paper proposes a non shared dual stream network based on the improved multi head self attention mechanism, which extracts the features of visible and infrared images respectively, divides the features horizontally, uses local features to calculate the identity loss and heterocenter loss, uses the multi head self attention mechanism to calculate the pixel's attention map according to the content of the pixel itself and its spatial position relative to adjacent pixels, and enhances the information capacity extracted from the model, Then we can learn the distinctive identity characteristics between different kinds under the same mode. At the same time, auxiliary modes are used to reduce the difference in color information between visible and infrared images, and further reduce the modal difference. The mAP on SYSU-MM01 and RegDB datasets reaches 60.36% and 76.10% respectively. The experiment shows that the method is effective.

**Key Words** Cross-Modality; Person Re-identification; Auxiliary Modal; Multi-Head Self-Attention

1 引 言

行人重识别是图像检索的一个子问题，主要是实现从一个摄像机捕获行人图片到其他设备下中搜索具有相同ID图片的问题。早期的传统方法偏重于设计更适合的手工特征，同时在度量阶段，构建学习更有效的相似度度量。近些年来，基于深度学习方法的行人重识别方法在性能上远超于传统基于人工设计特征的方法，在现有的公开数据集上平均精度均值已经达到了九十多，但这些数据集均为可见光图像，模态单一，使用夜间拍摄到的红外图像进行检索时效果较差。为了实现全天候监控，基于可见光——红外图像跨模态行人重识别被提出，使不同模态的行人图像能够相互匹配，充分利用多种摄像头拍摄到的监控数据。

解决跨模态行人重识别的关键在于学习两种模态的共享特征, 减小不同模态之间的差异。起初有基于表征学习和基于度量学习这两种方法，基于表征学习方法是设计合理的网络架构，提取两种模态图像共享的具有鲁棒性和鉴别性的特征，缩小模态间存在的差异性。基于度量学习方法旨在设计合理的度量方法或损失函数，学习一个映射空间，使得不同模态的类内距离小于类间距离。2017年，Wu等人[1]首次提出基于可见光——红外线图像的跨模态行人重识别问题，并设计一种深度补零方法来训练单流网络，让网络根据输入的图像自动选择特定节点进行跨模态匹配。为了减少两种模态之间的差距，Ye等人[3]提出一种双流CNN网络(TONE)来学习两种异质模态的多模态共享特征表示。考虑到使用全局特征会忽略细节信息，Zhu等人[6]设计了双流局部特征网络(TSLFN)来学习跨模态行人重识别的局部特征表示,并联合异质中心损失和身份损失来训练网络，缩小每个类内两个模态特征分布之间的距离。Park等人[16]考虑到局部特征会受行人错位、遮挡等因素影响，提出特征对齐来促使模型提取鲁棒性特征。Yin等人[17]认为单使用局部特征获取的信息量少，引入了局部特征分块加权求和再和全局特征相加计算损失。随着生成对抗网络兴起, 出现了基于模态互转的学习方法，即将两种模态的图像数据转换成同一模态数据，减少两种模态的差异。Wang等人[7]提出一种双级差异减少方法(D2RL)，利用GAN将RGB(IR)图像生成其对应的IR(RGB)图像, 形成统一的多光谱图像，减少模态间差异。Li等人[8]提出生成一个中间模态，将可见光图像和红外图像之间的巨大差异缩小为两种模态和中间模态之间的较小差异。

现有方法大多采用深层网络参数共享的双流网络来提取两种模态的共享特征，但这不利于提取同模态下不同类之间的辨别性特征。使用局部特征虽然会使模型注意到一些细节信息，但会造成空间信息的缺失。为此，本文设计了参数独立的双流网络来分别提取可见光图像和红外图像特征，融合改进后的自注意力机制增强模型提取特征的信息容量。采用灰度化的可见光图像作为辅助模态，减缓可见光图像和红外图像在色彩上的差异。使用身份损失和异质中心损失来训练模型，提高类内不同模态的特征相似度。

2 算法原理

2.1 双流网络结构

本文算法框架图如图1所示，该网络采用改进后的ResNet50网络作为两个支路的骨干网络，两个支路参数相互独立，主要包括以下三个部分：

（1）将每一批次的可见光图像灰度化，得到对应的辅助模态图像。用这两种模态图像一起训练可见光图像特征提取分支，以此来减缓可见光图像和红外图像之间在颜色信息上的差异。（2）在ResNet50网络第二个卷积模块后嵌入融合位置编码的多头自注意力机制(MHSA),考虑像素的空间位置，增强特征提取的信息容量。对输出的全局特征进行水平分割，得到的局部特征中融合了更多全局信息。（3）使用身份损失和异质中心损失对网络进行联合监督学习，提高类内不同模态的特征相似度，增大类间距离。

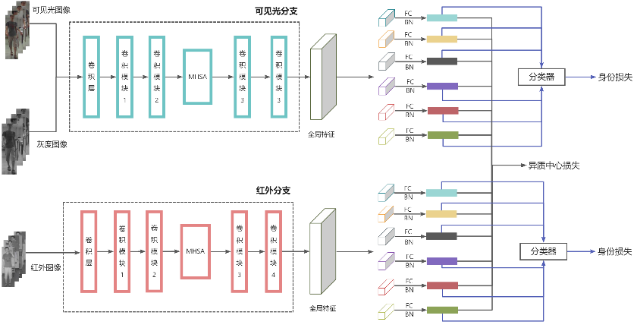


图1 算法总体框架图

2.2 辅助模态

可见光图像包含丰富的颜色信息，红外图像则包含更多的热信息，如图2所示。使用可见光图像灰度化得到的图像作为辅助模态，该灰度图像在图像结构上与可见光图像保持一致，在颜色信息上与红外图像差异较小。将可见光图像和灰度图像一起训练可见光图像特征提取分支，可以抑制网络学习特征在颜色信息上的过拟合，进而减小可见光图像与红外图像的模态差异。



图2 SYSU-MM01数据集中的示例

2.3 注意力机制

使用局部特征来计算损失会使模型缺少空间信息的提取，采用融入相对位置编码的多头自注意力机制（MHSA）[18]可以引导模型在提取特征时关注空间信息，如图3所示。

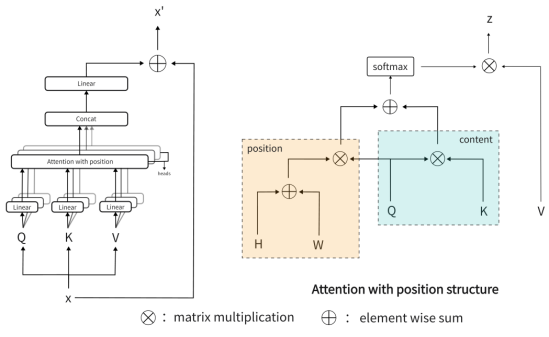


图3 多头自注意力机制结构

在MHSA中，输入图像特征x通过线性变换得到相应的*Q*、*K*、*V*（*Q*为查询矩阵，*K*为键矩阵，*V*为值矩阵），将其映射到不同的子空间，每头（即子空间）各自进行引入二维相对位置编码的自注意力变换，考虑像素的空间位置，根据像素本身的内容及其相对于相邻像素的空间位置来计算像素的注意力图。模型在不同的表示子空间里学习相关信息，联合这些学习到的信息，增强模型的表征学习能力。每头引入二维相对位置编码的自注意力变换和多头注意力计算过程见式(1)-(2)：

(1)

(2)

其中，*P*为相对位置矩阵，为相似度通过缩放因子,为最后一层全连接层的权重。

2.4 损失函数

为了扩大类间距离同时提高类内不同模态特征相似度，本文采用交叉熵损失和异质中心损失联合训练模型。将模型提取出的全局特征水平分割为*K*个水平部分，对每个部分应用全局平均池化得到*K*个列向量， 采用1×1卷积层对列向量进行降维得到局部特征。使用分类器对可见光图像、灰度图像以及红外图像对应的局部特征进行身份分类，计算交叉熵损失作为身份损失（*Lid*），让模型学习在每个类别特有的特征同时减少对颜色信息的关注，增大类间距离。*Lid*可表示为

（3）

其中*N*为批次大小，*K*为划分的局部特征数量。为第*i*张可见光图像的标签（即行人身份），为分类器对第*i*张可见光图像的第*j*个局部特征分类的概率结果。为第*i*张灰度图像的标签（即行人身份），为分类器对第*i*张灰度图像的第*j*个局部特征分类的概率结果。为第*i*张红外图像的标签（即行人身份），为分类器对第*i*张红外图像的第*j*个局部特征分类的概率结果。

使用可见光图像、红外图像对应的局部特征计算异质中心损失函数（*LHC*）[6]通过约束两个特征中心的差值来提高类内不同模态特征相似度。*Lid*可表示为

（4）

其中*L*为每一批次的类别数（行人身份个数），为第*i*类中可见光图像局部特征中心，为第*i*类中红外图像局部特征中心，两个特征中心的计算可表示为

，

（5）

其中*T*为每一批次每一类别的可见光图像（红外图像）数，*K*为划分的局部特征数量。为第*i*类中可见光图像的第*j*个局部特征，为第*i*类中红外图像的第*j*个局部特征。

最终模型整体损失函数*L*可表示为

（6）

其中是平衡两个损失函数的超参数，是异质中心损失在整体损失中的权重。

3 实验结果与分析

3.1 数据集和评价指标

本文在SYSU-MM01[1]和RegDB[2]两个公开数据集进行实验，使用Rank-n、mAP作为评价指标。

SYSU-MM01数据集包括由2台红外相机和4台可见光相机拍摄到的491个行人、303420张图像。其中三个相机采集室内图像，三个相机采集室外图像，使其能更好的模拟真实场景。使用296个身份进行训练，99个身份进行验证，96个身份进行测试。在测试阶段，使用可见光图像作为图像数据库，红外图像作为待检索的目标图像库。

RegDB数据集使用双可见光和热成像相机同时捕捉人体的可见光和红外图像，捕获了一个412人的图像数据库，对于每个人，捕捉了10张可见光图像和相应的10张红外图像。每个身份的红外图像和可见光图像的相机视角、天气状况、姿态都是一一对应的，相对SYSU-MM01数据集而言，降低了跨模态行人重识别任务难度。使用206个身份进行训练，206个身份进行测试。在测试阶段，使用可见光图像作为图像数据库，红外图像作为待检索的目标图像库。

Rank-n是指搜索结果最靠前的 n 张图片中包含待检索身份图像的概率，mAP是均值平均精度，是拿所有类别的平均精度求和除以所有类别，反应了模型检测行人的准确度。

3.2 实验细节

本文基于PyTorch 1.9.0学习库，显卡为NVIDIA RTX 3090，处理器为AMD EPYC 7543，内存为80GB的实验环境进行实验。为了减少样本不均衡带来的影响，每一批次随机选择8个行人，每个行人随机选择4张可见光图像和4张红外图像，并利用可见光图像生成对应的灰度图像，即每一批次共有96张图像。对图像做预处理，将其重置为288144尺寸，并以0.5概率对图像做水平翻转操作。选用SGD优化器优化模型，设置损失函数的超参数为0.5，学习率初始值为110-2。

3.3 实验结果

3.3.1 方法比较

为验证本文所提算法的有效性，将本文所提算法与跨模态重识别算法领域近几年主流的算法进行对比，实验结果如表1和表二所示，其中表1为在SYSU-MM01数据集采用全搜索模式的对比实验结果，表2为在RegDB数据集采用RGB-IR模式的对比结果。

表1 在SYSU-MM01数据集上的对比实验结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Single Shot | | | | Multi-shot | | | |
| Rank-1 | Rank-10 | Rank-20 | mAP | Rank-1 | Rank-10 | Rank-20 | mAP |
| Zero-padding[1] | 14.80 | 54.12 | 71.33 | 15.95 | 19.13 | 61.40 | 78.41 | 10.89 |
| BDTR[4] | 27.32 | 66.96 | 81.07 | 27.32 | - | - | - | - |
| cmGAN[5] | 26.97 | 67.51 | 80.56 | 27.80 | 31.49 | 72.74 | 85.01 | 22.27 |
| D2RL[7] | 28.9 | 70.6 | 82.4 | 29.2 | - | - | - | - |
| CPGM[10] | 38.10 | 80.70 | 89.90 | 36.90 | 45.10 | 85.70 | 93.80 | 29.50 |
| CDP+DHSM[11] | 38.0 | 82.3 | 91.7 | 38.4 | - | - | - | - |
| DPMBN[8] | 37.02 | 79.46 | 89.87 | 40.28 | - | - | - | - |
| MSPAC-MeCen[15] | 46.62 | 87.59 | 95.77 | 47.26 | 47.57 | 87.64 | 96.11 | 38.53 |
| MACE[13] | 51.64 | 87.25 | 94.44 | 50.11 | - | - | - | - |
| XCMG[9] | 49.92 | 89.79 | 95.96 | 50.73 | - | - | - | - |
| DDAG[12] | 54.75 | 89.26 | 95.25 | 53.02 | - | - | - | - |
| TSLFN[6] | 56.96 | 91.50 | 96.82 | 54.95 | 62.09 | 93.74 | 97.85 | 48.02 |
| TS-GAN[14] | 58.3 | 87.8 | 94.1 | 55.1 | 55.9 | 91.2 | 96.6 | 39.7 |
| DF2AM[19] | 56.93 | 90.80 | 96.11 | 55.10 | - | - | - | - |
| NFS[20] | 56.91 | 91.34 | 96.52 | 55.45 | 63.51 | 94.42 | 97.81 | 48.56 |
| ours | 62.20 | 95.33 | 98.45 | 60.36 | 67.15 | 97.45 | 99.43 | 53.43 |

表2 在RegDB数据集上的对比实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Rank-1 | Rank-10 | Rank-20 | mAP |
| Zero-padding[1] | 17.75 | 54.12 | 71.33 | 15.95 |
| BDTR[4] | 24.44 | 47.53 | 56.78 | 20.88 |
| D2RL[7] | 43.4 | 66.1 | 76.3 | 44.1 |
| CPGM[10] | 48.5 | - | - | 49.3 |
| MSPAC-MeCen[15] | 49.61 | 72.28 | 80.63 | 53.64 |
| CDP+DHSM[11] | 65.3 | 84.5 | 91.0 | 62.1 |
| MACE[13] | 72.37 | 88.40 | 93.59 | 69.09 |
| DDAG[12] | 69.34 | 86.19 | 91.49 | 63.46 |
| DF2AM[19] | 73.06 | 87.96 | 91.51 | 67.81 |
| XCMG[9] | 62.21 | 83.13 | 91.72 | 60.18 |
| NFS[20] | 80.54 | 91.96 | 95.07 | 72.10 |
| TSLFN[6] | 84.81 | 95.68 | 98.01 | 73.12 |
| ours | 87.23 | 96.41 | 98.40 | 76.10 |

Zero-padding[1]和D2RL[7]均为单流网络结构，D2RL[7]进一步利用生成对抗网络将跨模态问题转为单模态识别问题，提升了网络性能。BDTR[4]、DPMBN[8]、MSPAC-MeCen[15]、TSLFN[6]采用双流网络，将可见光图像和红外图像分别输入到网络中进行特征提取。MACE[13]、DDAG[12]、DF2AM[19]采用双流共享网络，提升网络提取不同模态共享特征的能力。CDP+DHSM[11]随机选择一个光谱作为中间模态，XCMG[9]根据可见光图像和红外图像生成中间模态来辅助模型学习不同模态共享特征，NFS[20]受NAS方法启发，让模型实现特征的自动选择。本文所提出的方法选择的是参数不共享的双流网络，融入MHSA让模型进行特征学习时考虑到空间信息，并在训练的时候加上辅助模态，抑制模型在可见光图像上对颜色信息的提取，进一步减小模态差异。对比XCMG[9]方法，在SYSU-MM01数据集上Rank-1和mAP均提高了12.28%和9.63%，在RegDB数据集上Rank-1和mAP均提高了25.02%和15.92%。对比最新方法NFS[20]，在SYSU-MM01数据集上Rank-1和mAP均提高了5.29%和4.91%，在RegDB数据集上Rank-1和mAP均提高了6.69%和4%。

3.3.2 消融实验

为验证本文所提算法关键部分的有效性，在SYSU-MM01数据集下分别对辅助模态（AM）模块和MHSA模块进行实验，实验结果如表3所示。直接使用基础双流网络进行测试，其Rank-1和mAP分别达到56.96%和54.95%。在双流网络中融入多头自注意力机制后分别提高到了60.24%和58.54%，在训练的时候加上辅助模态后分别提高到了62.20%和60.36%。

表3 在SYSU-MM01数据集上的消融实验结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Single Shot | | | | Multi-shot | | | |
| Rank-1 | Rank-10 | Rank-20 | mAP | Rank-1 | Rank-10 | Rank-20 | mAP |
| baseline | 56.96 | 91.50 | 96.82 | 54.95 | 62.09 | 93.74 | 97.85 | 48.02 |
| Baseline+MHSA | 60.24 | 93.47 | 97.79 | 58.54 | 65.14 | 95.15 | 98.43 | 51.46 |
| Baseline+MHSA+AM | 62.20 | 95.33 | 98.45 | 60.36 | 67.15 | 97.45 | 99.43 | 53.43 |

4 结束语

针对跨模态行人重识别问题，本文提出了基于辅助模态和注意力机制的双流跨模态行人重识别算法。在训练模型的时候加入辅助模态，促使模型在提取可见光图像特征时不会过多注意颜色信息，缩小两个模态之间的差异。考虑到局部特征对于空间信息的确实，在双流网络中融入多头注意力机制，增强模型对空间信息的提取。最后使用身份损失和异质中心损失联合训练模型，增大类间距离，提高类内两个模态特征相似度，实验结果表明了该算法的有效性。

**参考文献**

1. 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.
2. 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.
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, 32(1).
4. Ye M, Wang Z, Lan X, et al. Visible thermal person re-identification via dual-constrained top-ranking[C]//IJCAI. 2018, 1: 2.
5. Dai P, Ji R, Wang H, et al. Cross-modality person re-identification with generative adversarial training[C]//IJCAI. 2018, 1(3): 6.
6. Zhu Y, Yang Z, Wang L, et al. Hetero-Center Loss for Cross-Modality Person Re-Identification[J]. arXiv e-prints, 2019: arXiv: 1910.09830.
7. Wang Z, Wang Z, Zheng Y, et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 618-626.
8. Xiang X, Lv N, Yu Z, et al. Cross-modality person re-identification based on dual-path multi-branch network[J]. IEEE Sensors Journal, 2019, 19(23): 11706-11713.
9. 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. 2020, 34(04): 4610-4617.
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. Fan X, Luo H, Zhang C, et al. Cross-spectrum dual-subspace pairing for RGB-infrared cross-modality person re-identification[J]. arXiv preprint arXiv:2003.00213, 2020.
12. Ye M, Shen J, J Crandall D, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]//European Conference on Computer Vision. Springer, Cham, 2020: 229-247.
13. Ye M, Lan X, Leng Q, et al. Cross-modality person re-identification via modality-aware collaborative ensemble learning[J]. IEEE Transactions on Image Processing, 2020, 29: 9387-9399.
14. Zhang Z, Jiang S, Huang C, et al. RGB-IR cross-modality person ReID based on teacher-student GAN model[J]. Pattern Recognition Letters, 2021, 150: 155-161.
15. Zhang C, Liu H, Guo W, et al. Multi-scale cascading network with compact feature learning for RGB-infrared person re-identification[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 8679-8686.
16. 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.
17. Yin J, Ma Z, Xie J, et al. DF^ 2AM: Dual-level Feature Fusion and Affinity Modeling for RGB-Infrared Cross-modality Person Re-identification[J]. arXiv preprint arXiv:2104.00226, 2021.
18. Srinivas A, Lin T Y, Parmar N, et al. Bottleneck transformers for visual recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 16519-16529.
19. Yin J, Ma Z, Xie J, et al. DF^ 2AM: Dual-level Feature Fusion and Affinity Modeling for RGB-Infrared Cross-modality Person Re-identification[J]. arXiv preprint arXiv:2104.00226, 2021.
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.

作者简介：郭思琦，女，硕士研究生，研究方向：跨模态行人重识别；谭台哲，男，博士，副教授，研究方向：图像处理与计算机视觉、机器学习与大数据处理、区块链技术；陈梓骏，男，硕士研究生，研究方向：图像处理与计算机视觉、显著物体检测

作者联系方式

通信地址：广东省广州市番禺区大学城外环西路100号广东工业大学生活西区

邮政编码：510006

联系电话：13433617006

电子邮件：1303259387@qq.com