智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

结合多尺度特征与混淆学习的跨模态行人重识别

王路遥, 王凤随, 闫涛, 陈元妹

引用本文:
王路遥, 王凤随, 闫涛, 陈元妹. 结合多尺度特征与混淆学习的跨模态行人重识别[J]. 智能系统学报, 2024, 19(4): 898-908.
WANG Luyao, WANG Fengsui, YAN Tao, et al. Cross-modal person re-identification combining multi-scale features and confusion learning[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(4): 898-908.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202304010

DOI: 10.11992/tis.202304010
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240312.1048.002

结合多尺度特征与混淆学习的跨模态行人重识别

王路遥[1,2,3], 王凤随[1,2,3], 闫涛[1,2,3], 陈元妹[1,2,3]

（1. 安徽工程大学 电气工程学院, 安徽 芜湖 241000; 2. 安徽工程大学 检测技术与节能装置安徽省重点实验室, 安徽 芜湖 241000; 3. 安徽工程大学 高端装备先进感知与智能控制教育部重点实验室, 安徽 芜湖 241000）

摘　要: 跨模态行人重识别研究的重难点主要来自于行人图像之间巨大的模态差异和模态内差异。针对这些问题, 提出一种结合多尺度特征与混淆学习的网络结构。为实现高效的特征提取、缩小模态内差异, 将网络设计为多尺度特征互补的形式, 分别学习行人的局部细化特征与全局粗糙特征, 从细粒度和粗粒度两方面来增强网络的特征表达能力。利用混淆学习策略, 模糊网络的模态识别反馈, 挖掘稳定且有效的模态无关属性应对模态差异, 来提高特征对模态变化的鲁棒性。在大规模数据集 SYSU-MM01 的全搜索模式下该算法首位击中率和平均精度 (mean average precision, mAP) 的结果分别为 76.69% 和 72.45%, 在 RegDB 数据集的可见光到红外模式下该算法首位击中率和 mAP 的结果分别为 94.62% 和 94.60%, 优于现有的主要方法, 验证了所提方法的有效性。

关键词: 机器视觉; 行人重识别; 跨模态; 多尺度特征; 粗粒度; 细粒度; 混淆学习; 模态无关属性
中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)04−0898−11

中文引用格式: 王路遥, 王凤随, 闫涛, 等. 结合多尺度特征与混淆学习的跨模态行人重识别 [J]. 智能系统学报, 2024, 19(4): 898–908.
英文引用格式: WANG Luyao, WANG Fengsui, YAN Tao, et al. Cross-modal person re-identification combining multi-scale features and confusion learning[J]. CAAI transactions on intelligent systems, 2024, 19(4): 898–908.

Cross-modal person re-identification combining multi-scale features and confusion learning

WANG Luyao[1,2,3], WANG Fengsui[1,2,3], YAN Tao[1,2,3], CHEN Yuanmei[1,2,3]

(1. School of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China; 2. Anhui Key Laboratory of Detection Technology and Energy Saving Devices, Anhui Polytechnic University, Wuhu 241000, China; 3. Key Laboratory of Advanced Perception and Intelligent Control of High-end Equipment, Ministry of Education, Anhui Polytechnic University, Wuhu 241000, China)

Abstract: The difficulties of cross-modal person re-identification research mainly come from the huge modal differences and intra-modal differences between pedestrian images. To address these issues, a network structure combining multi-scale features with obfuscation learning is proposed. In order to achieve high-efficiency feature extraction and reduce intra-modal differences, the network is designed as a complementary form of multi-scale features to learn local refinement features and global rough features of pedestrians respectively. The feature expression ability of the network is enhanced from fine-grained and coarse-grained aspects. Confusion learning strategy is used to fuzzy the modal identification feedback of the network, and mine the stable and effective modal-independent attributes to cope with modal differences, so as to improve the robustness of features to modal changes. In the all-search mode of the large-scale data set SYSU-MM01, the results of the first hit rate and mean average precision (mAP) of the algorithm are 76.69% and 72.45%, respectively. In the Visible to Infrared mode of the RegDB data set, the results of the first hit rate and mAP of the algorithm are 94.62% and 94.60%, respectively, which are better than the main existing methods, verifying effectiveness of the proposed method.

Keywords: machine vision; person re-identification; cross-modal; multi-scale characteristics; coarse-grain; fine-grain; confusion learning; modal independent attribute

行人重识别[1-2]（person re-identification, ReID）旨在不同时间、地点跨越不同的摄像头来检索目标行人图像, 其对于智能监控领域具有重要的研究意义。随着深度学习技术[3]的发展与应用, 单

收稿日期: 2023−04−06.　**网络出版日期:** 2024−03−12.
基金项目: 安徽省自然科学基金项目（2108085MF197）; 安徽高校省级自然科学研究重点项目（KJ2019A0162）; 安徽工程大学国家自然科学基金预研项目（Xjky2022040）.
通信作者: 王凤随. E-mail: fswang@ahpu.edu.ac.cn.

模态的行人重识别[4-5]研究取得重大进展。然而，由于夜晚的能见度较低，为了保证实际的监控系统能够全天候作用，摄像机需要由可见光 (red green blue, RGB) 模式转换到红外 (infrared radiation, IR) 模式来捕获行人的外观特征。因此，RGB 图像和 IR 图像之间的跨模态行人重识别问题[6-7]被提出，且受到视觉领域的广泛关注。与单模态 ReID 相比，跨模态 ReID 研究最大的挑战是由相机原始的成像方式引起的模态差异，该差异使得可见光图像中如行人衣服颜色等重要的判别信息在红外图像中丢失，增加了识别的难度。另一方面，跨模态行人重识别还要应对传统 ReID 中行人模态内差异的挑战。

为解决以上难题，跨模态行人重识别研究重点关注如何获取具有鉴别性的跨模态共享信息，来改善网络模型的性能。如在单流网络方面，Wu 等[8]提出了深度零填充（zero-padding）的方法来训练网络，以自适应地学习模态共享特征，并且构建了一个可见光–红外的大规模行人数据集 SYSU-MM01。在双流网络方面，Ye 等[9]提出双流卷积神经网络 (two-stream CNN network, TONE) 网络框架，通过学习模态特异性信息与模态共享矩阵来分别处理交叉模态差异和模态内变化。随后，Ye 等[10]又针对跨模态行人重识别研究提出了端到端的双路径学习框架，该框架由一个用于捕获特征的双路径和用来判别特征的双向双约束顶级损失构成，以保证网络的识别率。Zhu 等[11]构建了双流的局部特征网络 (two-stream local feature network, TSLFN)，该网络具有 2 个独立的分支分别用来提取 2 种模态的特征，并且将输出的特征均匀划分为多条条纹进行局部特征学习，接着将来自不同模式的特征投射到同一子空间用共享权值的方式来学习模态共享特征。在模态转换网络方面，随着生成对抗网络 (generative adversarial network, GAN) 的发展，越来越多人应用 GAN 技术或图片风格转换的方式来解决跨模态行人重识别问题。如 Wang 等[12]使用生成对抗网络来解决跨模态差异问题，通过像素对齐模块将可见光图像合成虚拟的红外图像再与真实的红外图像进行匹配，以缓解图像的模态差异；同时考虑到行人身份信息的完整，利用联合鉴别模块优化图像和特征对之间的分布，确保网络不受新噪声的干扰，来进一步优化网络模型。Li 等[13]采用轻量级生成器生成 x 模态图像，该模态作为辅助与其他 2 个模态图像共同输入到权值共享的学习器中进行交叉模态学习，达到减少模态差异的目的。Zhang 等[14]提出了一个中间模态网络 (middle modality network, MMN)，利用非线性的中间模态发生器将 RGB 和 IR 图像有效地投射到统一的图像空间，生成中间模态图像来辅助网络减少模态差异，最后采用分布一致性损失使生成的图像分布尽可能一致。上述的研究通常仅着眼于捕获全局或局部共享特征来形成行人的特征描述符，以缩小跨模态差距，忽略了将行人的局部与全局信息相结合的作用。并且，对于使用 GAN 技术或类似生成的方式来缓解模态差异，由于生成图像质量差，往往容易在网络中引入额外的噪声，导致网络的识别率下降。而生成的图像又不能完全取代对应的缺失图像，从而带来额外的模态差异。

因此，为了克服上述方法的局限性，本文提出了一种基于多尺度特征与混淆学习的网络框架，其目的是充分获取行人图像中具有判别性且与模式无关的信息，以获得更强大的行人特征描述符。本文从全局特征与局部特征互补的角度出发，挖掘图像中有效的行人粗细粒度信息，获取丰富的鉴别性特征表示。同时，通过混淆模态识别反馈，提取与模态无关的高级语义层特征来抑制模态差异，增强表示学习对模态变化的抗干扰能力。该策略直接作用于网络中，避免了额外噪声和模态差异的影响。最后，联合损失函数来监督网络进行更新和训练，进一步提高跨模态行人重识别模型的整体性能。

1　多尺度特征及混淆学习

1.1　整体网络

网络结构如图 1 所示，主要由双流主干网络、多尺度特征互补模块、混淆学习策略 (confusion learning strategy, CLS) 和损失函数等部分组成。双流网络采用修改后的 Resnet50[15]作为骨干来提取行人图像的特征，分别输入可见光图像和红外图像在第 1 层和第 2 层进行模态特定特征提取，这两层的网络参数独立。而骨干网络的后 3 层共享权重作为特征嵌入部分，将网络提取的不同模态特征拼接后输入以提取模态共享特征，并且将最后一个卷积块的步幅由 2 改为 1，以获得较细粒度的行人特征；接着将网络分为双路径，分别聚焦于细腻和粗糙的特征区域，引入特征多样性，并采用混淆学习策略，确保优化专注于与模态无关的信息，从而提高共享特征表示的泛化能力。

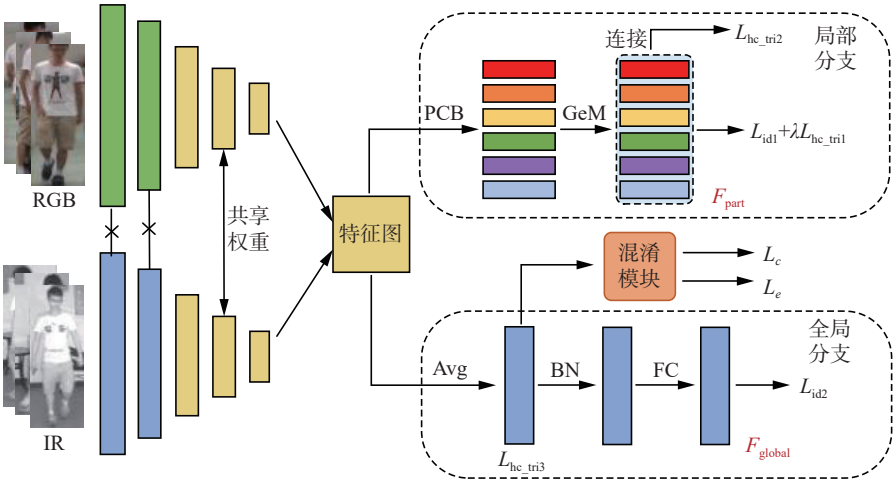


图 1　整体网络结构

Fig. 1　Overall network structure diagram

1.2　多尺度特征互补模块

身体结构是一个人的固有特征，无论图像模态如何变化，如图 2 所示人的身体结构信息都是基于模态不变的，可以作为跨模态共享的判别信息来表示一个人。因此，为提取更多具有鉴别性的特征信息，以实现缩小模态差异的目的，本文设计了多尺度特征互补模块，采用局部深度特征和全局深度特征互补的形式来学习身体的局部结构与全局结构，挖掘有效的粗细粒度信息，提高特征的判别能力。

图 2　行人粗细粒度特征划分

Fig. 2　Pedestrian coarse and fine granularity feature division

如图 1 所示，网络由可见光和红外 2 条路径组成，分别提取来自 2 种模态的图像特征，图像在经过双流网络之后转化为三维特征图。表 1 列出了 2 个分支的结构设置。对于局部特征提取分支，采用局部卷积基线网络 (part-based convolutional baseline, PCB)[16] 均匀划分策略。将三维特征图在水平方向均匀划分为 p 条，生成局部特征图以学习细粒度特征表示，引入内容粒度的多样性，其中 $p=6$。接着采用广义均值池化 (generalized-mean, GeM)[17] 将三维的部分级特征转化为一维特征向量，并且使用 1×1 卷积块降低局部特征向量的维数，得到局部特征 F_{part}。在分支的最后，将 p 块条纹特征连接用来描述人的身体结构，由异质中心三元组损失监督。对于全局深度特征提取部分，与部分级特征提取不同，该分支采用自适应平均池化 (adaptive average pooling, Avg) 来处理三维特征图，通过减小特征图大小来减少网络的计算量。然后，将特征向量输入批处理归一化 (batch normalization, BN) 层进行归一化，并且采用维数为 2 048 的全连接 (fully connected, FC) 层进行分类，将维度降至 512，可得到全局特征 F_{global}。最后，对于网络中提取的局部和全局特征采用异质中心三元组损失进行度量学习，然后采用标签平滑的识别损失对具有期望维数的全连接层输出的特征进行识别，而均匀划分的 p 块局部特征使用 p 个参数不共享的分类器。总的来说，多尺度特征互补模块使网络在捕获深层细化特征的同时也学习粗糙的全局深度特征，保持特征的多样性，以获取行人图像中完整的身体结构信息，从而增强行人的特征描述符。

表 1　局部和全局分支结构设置

Table 1　Local and global branch structure settings

分支	划分条数	特征维度	特征
局部分支	6	256	F_{part}
全局分支	1	2 048	F_{global}

1.3　混淆学习策略

由于可见光图像和红外图像来自 2 种不同的

模态，难以进行特征对齐比较。为减少模态之间的差异，受 Hao 等[18]的启发，将网络设置为忽略模态信息的结构，学习共享的特征表示。但共享的特征不等于有用的特征，故为了使网络在共有的视角下集中于收集有用的共享信息，而不过分关注琐碎的信息，本文采用一种混淆学习策略(confusion learning strategy, CLS)，利用对抗性学习的思路，混淆可见光和红外模态来欺骗网络，通过最小化不同模态间差异及最大化交叉模态相似度来实现特征对模态变化的鲁棒性。该方法直接嵌入在网络中，避免了生成图像质量差和多余噪声的影响，如图 3 所示。

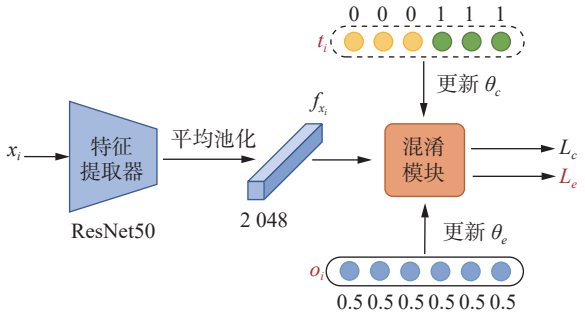


图 3　混淆学习结构

Fig. 3　Confuse learning structure

具体地来说，混淆学习策略由特征提取器和混淆模块组成。每个样本图像x_i都有一个身份标签y_i、一个真实的模态标签t_i和一个混淆的模态标签o_i。对于每个输入的图像x_i，使用一个二维向量来定义模态标签，将可见光图像和红外图像的真实模态标签分别设置为 [1,0] 和 [0,1]，而来自 2 种模态所有样本混淆的模态标签o_i设置为 [0.5, 0.5]。混淆模块由 2 个分类器构成，对输入图像的模态进行分类，其使用参数θ_c来表示。将样本图像x_i经过特征提取器得到维度为 2 048 的特征f_{x_i}输入混淆模块，并利用其得到的模态预测概率$p_c(f_{x_i})$与真实的模态标签t_i进行比较，可得到混淆模块的损失函数表达式为

$$L_c(\theta_c) = -\frac{1}{N}\sum_{i=1}^{N} t_i \log p_c(f_{x_i},\theta_c;\theta_e) \tag{1}$$

式中：N 为一批中的样本数，x_i为第 i 个输入的样本图像，θ_e和θ_c分别为学习到的特征提取器和模态分类器，$p_c(f_{x_i},\theta_c;\theta_e)$为样本被正确分类的概率。

特征提取器用来提取与模态无关且有鉴别性的特征。特征提取器为本文的 ResNet50 结构，用参数θ_e来表示。为了使特征提取器所捕获的特征达到混淆模态的作用，将特征提取器的预测概率与标签o_i进行比较，可得到：

$$L_e(\theta_e) = -\frac{1}{N}\sum_{i=1}^{N} o_i \log p_c(f_{x_i},\theta_e;\theta_c) \tag{2}$$

式中$p_c(f_{x_i},\theta_e;\theta_c)$表示提取到的样本特征能够实现模态混淆的概率，且式(1)和式(2)都使用 softmax 函数进行标准化。

在网络训练时，交替地更新θ_e和θ_c直到达到平衡，则可构成一个对抗损失，θ_e和θ_c的优化为

$$L(\theta_c,\theta_e) = L_c(\theta_c) + L_e(\theta_e)$$

$$\hat{\theta}_c = \arg\min_{\theta_c} L(\theta_c,\hat{\theta}_e)$$

$$\hat{\theta}_e = \arg\min_{\theta_e} L(\hat{\theta}_c,\theta_e)$$

优化过程中，每次仅更新一个部分。即θ_e使行人的特征分布相似来增加混淆模块的损失。而θ_c通过不断地减少混淆模块分类器的损失，使网络能够正确辨别模态。最后，通过对抗训练实现特征提取器所捕获的特征使模态分类器不能够区分图像的模态。当网络实现混淆时，L_c即可忽略。

1.4　损失函数

在本文提出的网络模型中，前部分通过多尺度特征学习和混淆学习操作得到了消除一定模态差异的特征。为了进一步处理跨模态和类内变化，保证行人图像分类的可行性和有效性，本文使用识别损失L_{id}和异质中心三元组损失$L_{\mathrm{hc_tri}}$联合监督，引导网络模型的训练。

识别损失的目的是对行人身份进行分类，通过将不同模态同一个人的特征视为同一类来整合行人的特征，学习模态特定特征来处理模态内变化。本文中采用标签平稳操作的识别损失，以防止模型训练过拟合，计算方式为

$$L_{\mathrm{id}} = \sum_{i=1}^{N} -q_i \log(p_i)$$

$$q_i = \begin{cases} 1 - \dfrac{N-1}{N}\xi, & y = i \\ \dfrac{\xi}{N}, & y \neq i \end{cases}$$

式中：N 为总训练集的身份个数；对于一个图像，y 为真实 ID 的标签；p_i为第i类的 ID 预测对数；ξ为一个常数，本文中$\xi = 0.1$，用来减少模型对训练集的置信度。

在跨模态行人重识别任务中通常采用三元组损失作为度量学习，但由于三元组损失通过锚与其他样本的比较来计算损失，具有强大的约束力，可能存在一些异常值，这将形成不利的三元组从而破坏其他成对距离。因此，本文采用异质中心三元组损失监督网络学习。该损失将三元组损失与中心损失的优势相结合，考虑使用每个人

的中心作为身份，通过替换锚点中心与其他样本的比较来放松约束，从而缓解不利的影响。故每个模态的身份特征中心的计算公式为

$$c_v^i = \frac{1}{K}\sum_{j=1}^{K} v_j^i$$

$$c_t^i = \frac{1}{K}\sum_{j=1}^{K} t_j^i$$

式中：v_j^i为第 i 个行人的第 j 张 RGB 图像，t_j^i为第 i 个行人的第 j 张 IR 图像，c_v^i为第 i 个行人 RGB 图像的中心，c_t^i为第 i 个行人 IR 图像的中心。

异质中心三元组损失不仅关注类内交叉模态紧凑型，还集中于挖掘类间可分离性特征，弥补了识别损失不能够应对跨模态变化的缺点。基于 PK 采样的方法，在每个小批中都有 P 个可见光图像的中心$\{c_v^i|i=1,2,\cdots,P\}$和 P 个红外图像的中心$\{c_t^i|i=1,2,\cdots,P\}$。因此，在跨模态行人重识别中，基于 PK 采样策略和中心距离公式，异质中心三元组损失可表述为

$$L_{\text{hc\_tri}} = \sum_{i=1}^{P}\left[\rho + \left\|c_v^i - c_t^i\right\|_2 - \min_{\substack{n\in[v,t]\\ j\neq i}}\left\|c_v^i - c_n^j\right\|_2\right]_+ + \sum_{i=1}^{P}\left[\rho + \left\|c_t^i - c_v^i\right\|_2 - \min_{\substack{n\in[v,t]\\ j\neq i}}\left\|c_t^i - c_n^j\right\|_2\right]_+$$

式中：c_n^j为第 j 个行人的难挖掘负样本中心，ρ为异质中心三元组损失的边缘值。对于每个身份，$L_{\text{hc_tri}}$只关注一个交叉模态的正对和在模态内和模态间挖掘最难的负对。

总的来说，对于局部特征分支，本文采用标签平滑的识别损失和异质中心三元组损失对细化特征进行处理，而拼接后的细化特征仅使用异质中心三元组损失监督：

$$L_{\text{part}} = \sum_{i=1}^{p}\left(L_{\text{id1}}^i + \lambda L_{\text{hc\_tri1}}^i\right) + L_{\text{hc\_tri2}}$$

式中：p 为部分级条纹特征数，λ为预定义的权衡参数。

对于全局特征分支，分别采用标签平滑的识别损失、异质中心三元组损失和混淆学习策略的对抗损失进行联合学习：

$$L_{\text{global}} = L_e + L_{\text{hc\_tri3}} + L_{\text{id2}}$$

因此，本文的总体损失函数为

$$L_{\text{all}} = L_{\text{part}} + L_{\text{global}}$$

2 实验结果与分析

2.1 数据集描述及评价标准

本文所提方法在 2 个常用的公共数据集上进行评估，包括 SYSU-MM01 和 RegDB。

SYSU-MM01 数据集由 4 个 RGB 摄像机和 4 个 IR 摄像机所捕获，分为室内和室外环境。训练集包含 395 个行人，其中 RGB 图像为 22 258 张，IR 图像为 11 909 张。测试集包含 96 个人，其中探针图像为测试集中的 3 803 张 IR 图像，图库集为 301 张 RGB 图像。在全搜索模式下，RGB 摄像机捕获的所有图像都被用作图库。在室内搜索模式下，只使用 2 个室内 RGB 摄像机捕获的图像作为图库。

RegDB 数据集由一对重叠的摄相机构建，共 4 120 张可见光图像和 4 120 张红外图像，包含 412 个行人，每个人有 10 张由可见光摄像机所捕获的 RGB 图像和 10 张由红外摄像机捕获的 IR 图像。训练时，任意选择 206 个行人身份图像，剩下 206 个行人身份图像用于测试。训练时将数据集进行 10 次自由划分，完成 10 次交叉验证，最终取平均值作为结果，使用可见光到红外和红外到可见光 2 种不同的测试设置。

在实验中，采用标准的累计匹配特征 (cumulative matching characteristics, CMC) 曲线、平均精度 (mean average precision, mAP)m_{AP} 和平均逆负惩罚 (mean of inverse negative penalty, mINP)[19]m_{INP} 作为评估标准。CMC 为第 k 次命中的概率，即根据相似度计算前 k 个结果中正确检索到图片的百分比，称为 R-k(Rank-k)。当 k=1、10、20 时计算测试集中前 1、10、20 张图片与查询集中图片相似度排序后为同一标签的准确度。mAP 表示测量所有类之间的平均检索性能，能够反应查询图片在图库图像中所有正确的图片排在检索列表前面的程度，其为多分类任务中平均精度 AP 求和后再取平均值的结果：

$$m_{\text{AP}} = \frac{\sum_{i=1}^{n} A_i}{C}$$

式中：A_i为类别的平均精度，C 为类别数。

mINP 为所有查询样本的平均逆置负样本惩罚率，表示检索最难正确匹配的工作量：

$$m_{\text{INP}} = \frac{1}{n}\sum_i (1 - N_i) = \frac{1}{n}\sum_i \frac{|G_i|}{R_i^{\text{hard}}}$$

式中：R_i^{hard}为最难匹配的排名位置，$|G_i|$为查询 i 次正确匹配的总数，N_i 为最难检索目标的匹配难度。

2.2 实验参数设置

本文实验环境：CPU 为 4 核 Intel(R) Xeon(R) Silver 4 110 CPU @ 2.10 GHz，内存 16 GB，显卡 NVIDIA GeForce RTX2080Ti(显存 11 GB)，操作系

统为 Ubuntu 16.04，深度学习框架 Pytorch1.1.0。行人图像大小为 288×144，数据增强通过对图像进行随机裁剪和翻转实现。异质中心三元组损失预定义的边缘值设置为 0.3。采样策略中，SYSU-MM01 数据集下设置了 P=6、K=8，RegDB 数据集下设置了 P=8、K=4，即在 1 个批次中，随机选取 P 个行人身份，每个身份包含 K 张 RGB 图像和 K 张 IR 图像。参数 λ 是一个预定义的权衡参数，其取值采取文献 [20] 中的设置，即在 SYSU-MM01 数据集中，λ=0.1；在 RegDB 数据集中，λ=0.2。部分级条纹特征 p 设置为 6。训练阶段 epoch 大小为 80；初始学习率为 0.1，在 20、50 个 epoch 时学习率衰减 0.1 和 0.01，在前 10 个 epoch 应用热身策略；采用动量参数为 0.9 的随机梯度下降 (stochastic gradient descent, SGD) 来优化网络。

2.3　与其他方法进行比较

为验证本文算法的有效性，本节将所提出的算法与现有的跨模态行人重识别算法进行比较。在 2 个公开数据集 SYSU-MM01 和 RegDB 的实验结果如表 2 和表 3 所示，主要包括双向双重约束排序 (bi-directional dual-constrained top-ranking, BDTR) 损失的双流网络[10]、对齐生成对抗网络 (alignment generative adversarial network, Align GAN)[12]、Xmodal[13](x modality)、双流的局部特征网络 (two-stream local feature network, TSLFN)[11]、异质中心三元组 (hetero-center triplet, HCT) 损失[20]、中间模态网络 (middle modality network, MMN)[14] 和记忆增强单向度量 (memory-augmented unidirectional metric, MAUM)[21] 等算法。

表 2　在 SYSU-MM01 数据集和其他方法对比实验结果

Table 2　Experimental results compared with other methods in the SYSU-MM01 data set　%

算法	全搜索模式					室内搜索模式				
	R-1	R-10	R-20	mAP	mINP	R-1	R-10	R-20	mAP	mINP
Zero-padding[8]	14.80	54.12	71.33	15.95	—	20.58	68.38	85.79	26.92	—
TONE[9]	12.52	50.72	68.60	14.42	—	20.82	68.86	84.46	26.38	—
HCML[9]	14.32	53.16	69.17	16.16	—	24.52	73.25	86.73	30.08	—
BDTR[10]	27.32	66.96	81.07	27.32	—	31.92	77.18	89.28	41.86	—
eBDTR[22]	27.82	67.34	81.34	28.42	—	—	—	—	—	—
D^{2}RL[23]	28.90	70.60	82.40	29.20	—	—	—	—	—	—
MAC[24]	33.26	79.04	90.09	36.22	—	33.37	82.49	93.69	44.95	—
DPMBN[25]	37.02	79.46	89.87	40.28	—	44.17	87.12	95.24	54.51	—
Align GAN[12]	42.40	85.00	93.70	40.70	—	45.90	87.60	94.40	54.30	—
LZM[26]	45.00	89.06	95.77	45.94	—	54.28	94.22	98.22	63.92	—
Xmodal[13]	49.92	89.79	95.96	50.73	—	—	—	—	—	—
DDAG[27]	54.75	90.39	95.81	53.02	—	61.02	94.06	98.41	67.98	—
TSLFN[11]	59.96	91.50	96.82	54.95	—	59.74	92.07	96.22	64.91	—
HCT[20]	61.68	93.10	97.17	57.51	39.54	63.41	91.69	95.28	68.17	64.26
DG-VAE[28]	59.49	93.77	—	58.46	—	—	—	—	—	—
cm-SSFT[29]	61.60	89.20	93.90	63.20	—	70.50	94.90	97.70	72.60	—
MMN[14]	70.60	96.20	99.00	66.90	—	76.20	97.20	99.30	79.60	—
MPANet[30]	70.58	96.21	98.80	68.24	—	76.74	98.21	99.57	80.95	—
MAUM[21]	71.68	—	—	68.79	—	76.97	—	—	81.94	—
MIFL+CFIM[31]	61.67	91.67	96.27	57.72	—	64.45	92.29	96.34	68.97	—
本文算法	**76.69**	**94.89**	**97.48**	**72.45**	**59.64**	**81.19**	**94.32**	**97.27**	**82.78**	**79.42**

表 3　在 RegDB 数据集和其他方法对比实验结果

Table 3　Experimental results compared with other methods in the RegDB dataset　%

算法	可见光到红外					红外到可见光				
	R-1	R-10	R-20	mAP	mINP	R-1	R-10	R-20	mAP	mINP
Zero-padding[8]	17.75	34.21	44.35	18.90	—	16.63	34.68	44.25	17.82	—
HCML[9]	24.44	47.53	56.78	20.08	—	21.70	45.02	55.58	22.24	—
BDTR[10]	33.56	58.61	67.43	32.76	—	32.92	58.46	68.43	31.96	—
eBDTR[22]	34.62	58.96	68.72	33.46		34.21	58.74	68.64	32.49	
MAC[24]	36.43	62.36	71.63	37.03	—	36.20	61.68	70.99	36.63	—
Align GAN[12]	57.90	—	—	53.60	—	56.30	—	—	53.40	—
Xmodal[13]	62.21	83.13	91.72	60.18	—	—	—	—	—	—
HCT[20]	91.05	97.16	98.57	83.28	68.84	89.30	96.41	98.16	81.46	64.81
cm-SSFT[29]	72.30	—	—	72.90	—	71.00	—	—	71.70	—
MMN[14]	91.60	97.70	98.90	84.10	—	87.50	96.00	98.10	80.50	—
MPANet[30]	83.70	—	—	80.90	—	82.80	—	—	80.70	—
MAUM[21]	87.87	—	—	85.09	—	86.95	—	—	84.34	—
MIFL+CFIM[31]	**95.12**	**98.07**	**99.00**	91.06	—	**94.42**	**97.85**	**98.81**	90.23	—
本文算法	94.62	97.20	98.72	**94.60**	**92.96**	92.77	96.17	98.16	**93.39**	**91.95**

从实验数据可看出，本文方法在 SYSU-MM01 数据集的全搜索设置下 R-1、mAP 和 mINP 的结果分别为 76.69%、72.45% 和 59.64%，在 RegDB 数据集的可见光到红外设置下 R-1、mAP 和 mINP 的结果分别为 94.62%、94.60% 和 92.96%，均优于对比方法，证明了本文方法的先进性。具体来看，双流卷积神经网络 (two-stream CNN network, TONE)、BDTR、双向中心约束排序 (bi-directional center-constrained top-ranking, eBDTR) 损失等以往的方法主要利用双流网络来提取红外图像和可见光图像的模态特定特征，并通过改进损失函数来约束行人特征的距离，但网络忽略了模态共享特征对处理模态间变化的重要性。本文方法将网络的主干部分划分为参数特定和参数共享的 2 个阶段，以分别提取行人的模态特异性信息和模态共享信息，为深层的网络学习特征表示奠定了基础。双水平差异减少学习 (dual-level discrepancy reduction learning, D^2RL)、Align GAN、Xmodal、MMN 等方法利用转换图片风格的方式来减轻模态差异带来的影响，但生成的图像不仅扩大了工作量还容易引入新的噪声干扰。而本文算法在不混入多余噪声的前提下利用最大–最小博弈的方法，混淆网络提取 2 种模态的特征，迫使网络专注于优化与模态无关的共享特征，同样缓解了模态差异的影响。由实验对比结果可知，与此类方法中效果最好的 MMN 算法相比，所提方法在 SYSU-MM01 数据集全搜索设置下 R-1 结果提升 6.09%，mAP 结果提升 5.55%；在 RegDB 数据集的可见光到红外设置下 R-1 结果提升 3.02%，mAP 结果提升 10.5%。TSLFN、HCT 等方法通过对行人特征均匀划分进行细化学习，来最大化跨模态特征的相似性。如 HCT算法对提取的特征进行切块处理，使网络仅关注行人图片中的细粒度信息，并在此基础上联合基于样本中心的损失函数来监督学习，通过拉近同一个人的特征距离，推远不同人的特征距离来提高行人识别的准确率。但由于网络过于依赖细节信息，而忽略了对整体特征的把握，导致模型精准度不高。本文网络将细粒度局部特征与粗粒度全局特征相结合，增加特征的丰富度来弥补这一缺陷，使模型精度得到明显提升。在 SYSU-MM01 数据集与 HCT 方法相比，全搜索设置下本文算法的 R-1、mAP 和 mINP 分别高出 15.01%、14.94% 和 20.10%；在 RegDB 数据集与 HCT 方法相比，可见光到红外设置下本文算法的 R-1、mAP 和 mINP 分别高出 3.57%、11.32% 和 24.12%。

次优的 MAUM 方法针对模态不平衡问题提出了记忆增强单向度量学习，即在 2 个单一方向学习显式的跨模态度量，并通过基于记忆的增强进一步提高网络精度。虽然该算法获得了一定的

优越性，但与本文方法相比工作量较大，且不适用于大规模数据集。

2.4　消融实验

为评估网络结构中各个模块对网络的影响，在 SYSU-MM01 数据集的 2 种模式下进行了一系列消融实验，结果如表 4 所示。具体来说 Base 表示基线，即在 HCT 网络的基础上采用 k-reciprocal[32] 重排序操作；global 表示全局特征分支；CLS 表示在网络中使用混淆学习策略。

表 4　SYSU-MM01 数据集下的消融实验

Table 4　Ablation experiments under the SYSU-MM01 data set　%

算法	全搜索模式				室内搜索模式			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
Base	71.32	92.45	96.14	66.59	76.75	92.26	95.82	78.46
Base+global	73.94	94.04	97.37	71.35	79.68	93.41	96.33	81.44
Base+global+CLS	**76.69**	**94.89**	**97.48**	**72.45**	**81.19**	**94.32**	**97.27**	**82.78**

由表 4 可知，在 Base 基础上将全局特征分支与部分级特征分支共同作用，进一步提升了网络模型的性能。即在全搜索模式下 R-1 提升了 2.62%，mAP 提升了 4.76%，在室内搜索模式下，R-1 提升了 2.93%，mAP 提升了 2.98%，可见该方法提取到不同粒度特征能够使网络学习到更多具有鉴别性信息的特征，从而获得更具鲁棒性的行人表示，有效地对行人图像进行识别分类。接着在全局特征提取部分融入混淆学习策略，使网络在全搜索模式下 R-1 和 mAP 分别提高了 2.75% 和 1.1%，在室内搜索模式下检索性能也得到进一步提高。说明网络通过最大–最小博弈的方法，能够使优化集中于与模式无关且有用的特征上，增强了表示学习对模态变化的泛化能力。总的来说与基线相比，本文方法在 SYSU-MM01 数据集的全搜索模式下 R-1 结果提升了 5.37%，mAP 结果提升了 5.86%；在室内搜索模式下 R-1 和 mAP 分别提升了 4.44% 和 4.32%。

2.5　不同池化与混淆模块位置分析

为进一步探索全局分支中不同池化方式和不同特征进行混淆学习对网络性能的影响，在 SYSU-MM01 数据集下进行了一系列实验。

图 4 为使用广义均值池化 (GeM)、最大池化 (max pooling, Max) 和自适应平均池化 (Avg) 来分别构建全局分支。由实验结果可以看出，使用自适应平均池化的网络平均精度与平均逆置负样本惩罚率最高，说明平均池化能够代替其他池化来整合全局空间信息，增强对输入空间变化的鲁棒性，从而提升网络模型的精度，故所提方法采用自适应平均池化构建全局分支。表 5 为本文分别对池化 (Pooling) 层、批量归一化 (batch normalization, BN) 层和全连接 (fully connected, FC) 层得到的行人特征进行混淆学习。具体的，将各层输出的特征作为混淆学习策略中特征提取器的输出来测试网络的性能。如表 5 所示，在全搜索模式和室内搜索模式下的各个指标都表明对 Pooling 层输出的特征进行混淆学习效果最佳，而 BN 层和 FC 层输出的特征均难以实现模态混淆的目的，侧面验证了本文选择在 ResNet50 后靠前的高级语义层进行模态混淆学习的有效性。

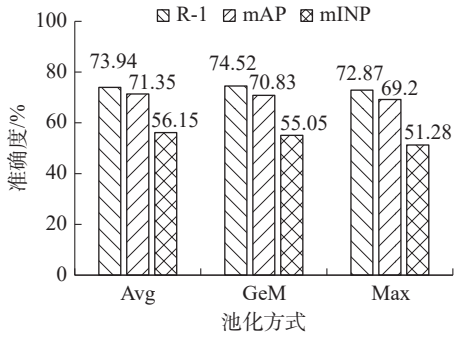


图 4　全局分支中不同池化方式的实验结果

Fig. 4　Performance experimental results of different pooling methods in the global branch

表 5　混淆模块不同位置对模型性能影响的实验结果

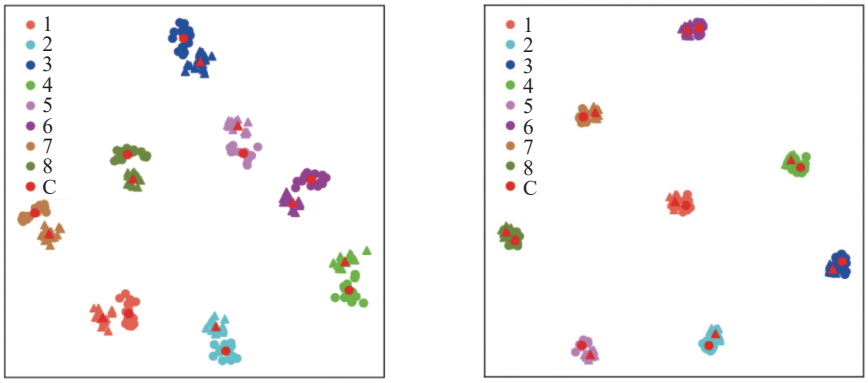
Table 5　Experimental results of the influence of different positions of obfuscation module on model performance　%

位置	全搜索模式		室内搜索模式	
	R-1	mAP	R-1	mAP
BN	73.25	70.49	79.21	81.97
FC	75.29	71.06	78.23	80.55
Pooling	**76.69**	**72.45**	**81.19**	**82.78**

2.6　t-SNE 可视化及分析

为进一步验证本文算法所捕获特征的鉴别

能力，在二维特征空间中绘制特征表示，使用 t 分布–随机邻近嵌入 (t-distributed stochastic neighbor embedding, t-SNE) 在 SYSU-MM01 数据集随机抽取 8 人，进行特征可视化。如图 5 所示，“1~8”表示从测试集中随机选出的 8 个身份，“C”表示特征中心，相同颜色的样本来自同一个人，“○”和“△”标记分别表示来自 RGB 和 IR 模式的图像。由图 5 可视化结果可以看出，与基线网络相比本文网络的特征可视化图中同类的特征分布更为紧凑，特征中心距离也更近，而不同行人的特征可分离性也更强，说明本文算法有效缩短了 2 个模式下相同身份特征之间的距离，同时扩大了类间差异，从而减少了模态间和模态内差异。



(a) 基线网络特征可视化图　　(b) 本文网络特征可视化图

图 5　SYSU-MM01 数据集特征可视化对比结果

Fig. 5　SYSU-MM01 data set feature visualization comparison results

3　结束语

本文从充分利用行人的跨模态特征和缓解模态差异的角度出发，针对跨模态行人重识别任务提出了一种基于多尺度特征与混淆学习的双流网络，该网络联合多尺度特征互补模块、混淆学习策略和多种损失函数，兼顾学习具有鉴别性和模态无关的特征。具体地说，网络通过多尺度特征互补模块同时捕获不同尺度的语义信息，丰富了行人特征图的可鉴别性，提高了特征表示的泛化能力。同时考虑到模态变化，利用混淆学习混淆不同模态的特征输入，使网络在与模式无关的视角下学习有用的特征，提高了表征对跨模态差异的抗干扰能力。最后，在损失函数的引导下，完成了网络的进一步优化。在 2 个公共数据集取得了先进的识别精度，并通过实验对比和可视化验证，证明了所提方法对行人有效特征的挖掘和识别能力。下一步工作中将考虑不同模态的特征对齐来进一步改善图像间跨模态差异，计划通过新的数据集或拓展数据集来提高网络的泛化能力。

参考文献：

[1]　宋婉茹, 赵晴晴, 陈昌红, 等. 行人重识别研究综述 [J]. 智能系统学报, 2017, 12(6): 770–780.
SONG Wanru, ZHAO Qingqing, CHEN Changhong, et al. Survey on pedestrian re-identification research[J]. CAAI transactions on intelligent systems, 2017, 12(6): 770–780.

[2]　罗浩, 姜伟, 范星, 等. 基于深度学习的行人重识别研究进展 [J]. 自动化学报, 2019, 45(11): 2032−2049.
LUO Hao, JIANG Wei, FAN Xing, et al. A survey on deep learning based person re-identification[J]. Acta automatica sinica, 2019, 45(11): 2032−2049.

[3]　刘帅师, 程曦, 郭文燕, 等. 深度学习方法研究新进展 [J]. 智能系统学报, 2016, 11(5): 567–577.
LIU Shuaishi, CHENG Xi, GUO Wenyan, et al. Progress report on new research in deep learning[J]. CAAI transactions on intelligent systems, 2016, 11(5): 567–577.

[4]　邵晓雯, 帅惠, 刘青山. 融合属性特征的行人重识别方法 [J]. 自动化学报, 2022, 48(2): 564–571.
SHAO Xiaowen, SHUAI Hui, LIU Qingshan. Person re-identification based on fused attribute features[J]. Acta automatica sinica, 2022, 48(2): 564–571.

[5]　石跃祥, 周玥. 基于阶梯型特征空间分割与局部注意力机制的行人重识别 [J]. 电子与信息学报, 2022, 44(1): 195–202.
SHI Yuexiang, ZHOU Yue. Person re-identification based on stepped feature space segmentation and local attention mechanism[J]. Journal of electronics & information technology, 2022, 44(1): 195–202.

[6]　庄建军, 庄宇辰. 一种结构化双注意力混合通道增强的跨模态行人重识别方法 [J]. 电子与信息学报, 2024,

46(2): 518–526.
ZHUANG Jianjun, ZHUANG Yuchen. A cross-modal person re-identification method based on hybrid channel augmentation with structured dual attention[J]. Journal of electronics & information technology, 2024, 46(2): 518–526.

[7] 孙锐, 张磊, 余益衡, 等. 基于局部异构聚合图卷积网络的跨模态行人重识别 [J]. 电子学报, 2023, 51(4): 810–825.
SUN Rui, ZHANG Lei, YU Yiheng, et al. Cross-modality person re-identification based on locally heterogeneous polymerization graph convolutional network[J]. Acta electronica sinica, 2023, 51(4): 810–825.

[8] WU Ancong, ZHENG Weishi, YU Hongxing, et al. RGB-infrared cross-modality person re-identification[C]//The IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5390−5399.

[9] YE Mang, LAN Xiangyuan, LI Jiawei, et al. Hierarchical discriminative learning for visible thermal person re-identification[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 7501−7508.

[10] YE Mang, WANG Zheng, LAN Xiangyuan, et al. Visible thermal person re-identification via dual-constrained top-ranking[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm: Morgan Kaufmann, 2018: 1092−1099.

[11] ZHU Yuanxin, YANG Zhao, WANG Li, et al. Hetero-center loss for cross-modality person re-identification[J]. Neurocomputing, 2020, 386: 97–109.

[12] WANG Guanan, ZHANG Tianzhu, CHENG Jian, et al. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment[C]//The IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 3622−3631.

[13] LI Diangang, WEI Xing, HONG Xiaopeng, et al. Infrared-visible cross-modal person re-identification with an x modality[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 4610−4617.

[14] ZHANG Yukang, YAN Yan, LU Yang, et al. Towards a unified middle modality learning for visible-infrared person re-identification[C]//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu: ACM, 2021: 788−796.

[15] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//The IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[16] SUN Yifan, ZHENG Liang, YANG Yi, et al. Beyond part models: person retrieval with refined part pooling[C]//Proceedings of the European Conference on Computer Vision. Munich: IEEE, 2018: 480−496.

[17] RADENOVIć F, TOLIAS G, CHUM O. Fine-tuning CNN image retrieval with no human annotation[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(7): 1655–1668.

[18] HAO Xin, ZHAO Sanyuan, YE Mang, et al. Cross-modality person re-identification via modality confusion and center aggregation[C]//The IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 16383−16392.

[19] YE Mang, SHEN Jianbing, LIN Gaojie, et al. Deep learning for person re-identification: a survey and outlook[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 44(6): 2872–2893.

[20] LIU Haijun, TAN Xiaoheng, ZHOU Xichuan. Parameter sharing exploration and hetero-center triplet loss for visible-thermal person re-identification[J]. IEEE transactions on multimedia, 2021, 23: 4414–4425.

[21] LIU Jialun, SUN Yifan, ZHU Feng, 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. New Orleans: IEEE, 2022: 19366−19375.

[22] YE Mang, LAN Xiangyan, WANG Zheng, et al. Bi-directional center-constrained top-ranking for visible thermal person re-identification[J]. IEEE transactions on information forensics and security, 2020, 15: 407–419.

[23] WANG Zhixiang, WANG Zheng, ZHENG Yinqiang, et al. Learning to reduce dual-level discrepancy for infrared-visible person re-identification[C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 618−626.

[24] WU Ancong, ZHENG Weishi, GONG Shaogang, et al. RGB-IR person re-identification by cross-modality similarity preservation[J]. International journal of computer vision, 2020, 128(6): 1765–1785.

[25] XIANG Xuezhi, LYU Ning, YU Zeting, et al. Cross-modality person re-identification based on dual-path multi-branch network[J]. IEEE sensors journal, 2019, 19(23): 11706–11713.

[26] BASARAN E, GOKMEN M, KAMASAK M E. An efficient framework for visible-infrared cross modality per-

son re-identification[J]. Signal processing: image communication, 2020, 87: 115933.

[27] YE Mang, SHEN Jianbing, CRANDALL D J, et al. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification[C]//The 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 229−247.

[28] PU Nan, CHEN Wei, LIU Yu, et al. Dual gaussian-based variational subspace disentanglement for visible-infrared person re-identification[C]//Proceedings of the 28th ACM International Conference on Multimedia. Seattle: ACM, 2020: 2149−2158.

[29] LU Yan, WU Yue, LIU Bin, et al. Cross-modality person re-identification with shared-specific feature transfer[C]//The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 13376−13386.

[30] WU Qiong, DAI Pingyang, CHEN Jie, et al. Discover cross-modality nuances for visible-infrared person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 4330−4339.

[31] 石林波, 李华锋, 张亚飞, 等. 模态不变性特征学习和一致性细粒度信息挖掘的跨模态行人重识别 [J]. 模式识别与人工智能, 2022, 35(12): 1064–1077.
SHI Linbo, LI Huafeng, ZHANG Yafei, et al. Modal invariance feature learning and consistent fine-grained information mining based cross-modal person re-identification[J]. Pattern recognition and artificial intelligence, 2022, 35(12): 1064–1077.

[32] ZHONG Zhun, ZHENG Liang, CAO Donglin, et al. Re-ranking person re-identification with k-reciprocal encoding[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3652−3661.

作者简介：

王路遥，硕士研究生，主要研究方向为深度学习、行人重识别方面的研究。E-mail: 578651059@qq.com。

王凤随，教授，博士，主要研究方向为图像与视频信息处理、视觉计算与智能分析、视觉目标跟踪、智能计算与多目标优化视频通信。主持省级自然科学研究重点项目 2 项，省级自然科学基金项目 2 项、省教育厅省级质量工程虚拟仿真实验教学项目 1 项、省教育厅质量工程“六卓越、一拔尖”卓越人才培养创新项目 1 项，获国家专利授权 16 项，发表学术论文 40 余篇。E-mail: fswang@ahpu.edu.ac.cn。

闫涛，硕士研究生，主要研究方向为深度学习、行人重识别方面的研究，发表学术论文 2 篇。E-mail: 1847026840@qq.com。