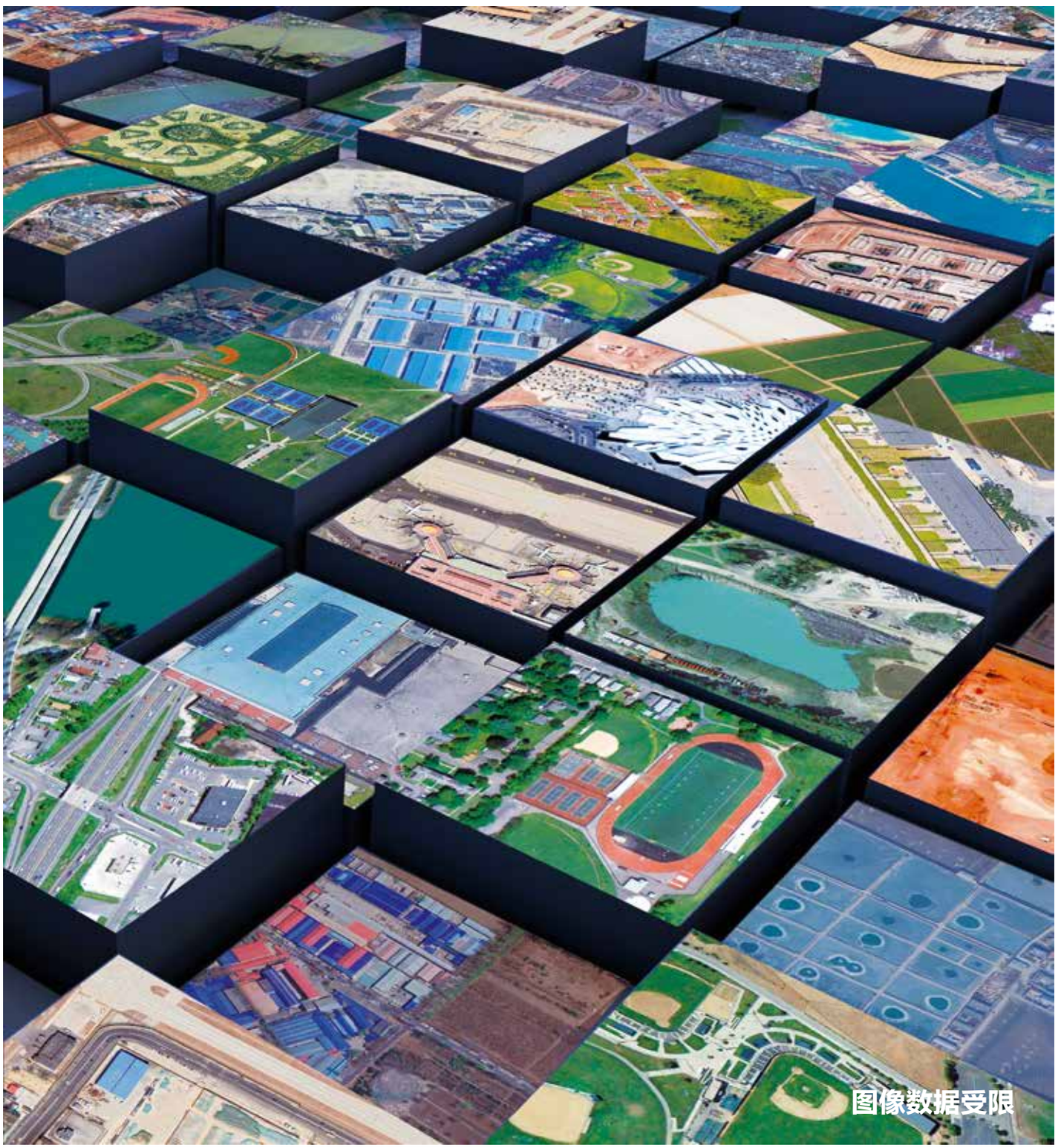
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
10
VOL.27

ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第10期（总第318期）
2022年10月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

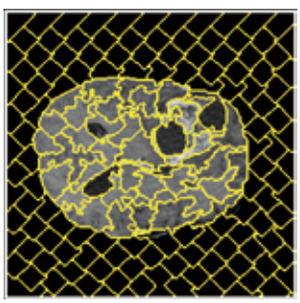
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

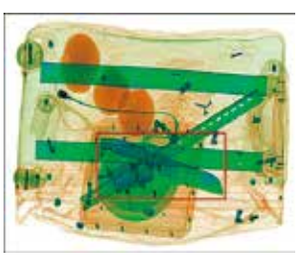
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年10月 第27卷 第10期（总第318期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

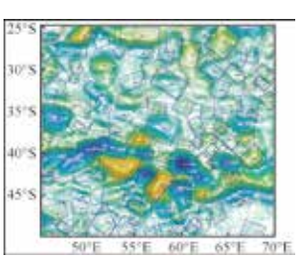
目次



MRI脑肿瘤图像的超像素/体素分割及发展现状(第2897页)



融合多尺度特征与全局上下文信息的X光违禁物品检测(第3043页)

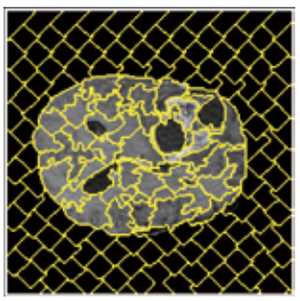


融合多尺度旋转锚机制的海洋中尺度涡自动检测(第3092页)

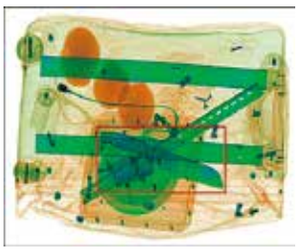
封面素材来自中国科学院空天信息创新研究院二部孙显研究员团队

Volume 27, Number 10
Published October 16, 2022

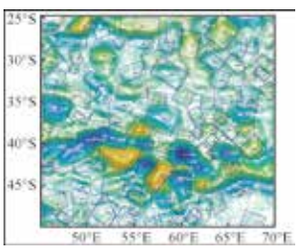
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



The review of superpixel/voxel segmentation of MRI brain tumor images(P2897)



Integrated multi-scale features and global context in x-ray detection for prohibited items(P3043)



Multi-scale rotating anchor mechanism based automatic detection of ocean mesoscale eddy(P3092)

中图法分类号:TP301.6　　**文献标识码**:A　　**文章编号**:1006-8961(2022)10-2843-17

论文引用格式:Wu A C, Lin C Z and Zheng W S. 2022. Single-modality self-supervised information mining for cross-modality person re-identification. Journal of Image and Graphics,27(10):2843-2859(吴岸聪,林城栋,郑伟诗. 2022. 面向跨模态行人重识别的单模态自监督信息挖掘. 中国图象图形学报,27(10):2843-2859)[DOI:10.11834/jig.211050]

面向跨模态行人重识别的单模态自监督信息挖掘

吴岸聪,林城栋,郑伟诗*

中山大学计算机学院,广州　510006

摘　要:**目的**　在智能监控视频分析领域中,行人重识别是跨无交叠视域的摄像头匹配行人的基础问题。在可见光图像的单模态匹配问题上,现有方法在公开标准数据集上已取得优良的性能。然而,在跨正常光照与低照度场景进行行人重识别的时候,使用可见光图像和红外图像进行跨模态匹配的效果仍不理想。研究的难点主要有两方面:1)在不同光谱范围成像的可见光图像与红外图像之间显著的视觉差异导致模态鸿沟难以消除;2)人工难以分辨跨模态图像的行人身份导致标注数据缺乏。针对以上两个问题,本文研究如何利用易于获得的有标注可见光图像辅助数据进行单模态自监督信息的挖掘,从而提供先验知识引导跨模态匹配模型的学习。**方法**　提出一种随机单通道掩膜的数据增强方法,对输入可见光图像的3个通道使用掩膜随机保留单通道的信息,使模型关注提取对光谱范围不敏感的特征。提出一种基于三通道与单通道双模型互学习的预训练与微调方法,利用三通道数据与单通道数据之间的关系挖掘与迁移鲁棒的跨光谱自监督信息,提高跨模态匹配模型的匹配能力。**结果**　跨模态行人重识别的实验在"可见光—红外"多模态行人数据集SYSU-MM01(Sun Yat-Sen University Multiple Modality 01)、RGBNT201(RGB, near infrared, thermal infrared, 201)和RegDB上进行。实验结果表明,本文方法在这3个数据集上都达到领先水平。与对比方法中的最优结果相比,在RGBNT201数据集上的平均精度均值mAP(mean average precision)有最高接近5%的提升。**结论**　提出的单模态跨光谱自监督信息挖掘方法,利用单模态可见光图像辅助数据挖掘对光谱范围变化不敏感的自监督信息,引导单模态预训练与多模态有监督微调,提高跨模态行人重识别的性能。

关键词:行人重识别;跨模态检索;红外图像;自监督学习;互学习

Single-modality self-supervised information mining for cross-modality person re-identification

Wu Ancong, Lin Chengzhi, Zheng Weishi*

School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou　510006, China

Abstract: **Objective** Urban video surveillance systems have been developing dramatically nowadays. The surveillance videos analysis is essential for security but a huge amount of labor-intensive data processing is highly time-consuming and costly. Intelligent video analysis can be as an effective way to deal with that. To analyze the concrete pedestrians' event, person

收稿日期:2021-11-04;**修回日期**:2022-05-23;**预印本日期**:2022-05-30

*通信作者:郑伟诗　zhwshi@mail.sysu.edu.cn

基金项目:国家自然科学基金青年科学基金项目(62106288);中国博士后创新人才支持计划(BX20200395);中国博士后科学基金面上资助(2021M693616)

Supported by: National Science Foundation for Young Scientists of China (62106288); China National Postdoctoral Program for Innovative Talents (BX20200395); China Postdoctoral Science Foundation (2021M693616)

re-identification is a basic issue of matching pedestrians across non-overlapping cameras views for obtaining the trajectories of persons in a camera network. The cross-camera scene variations are the key challenges for person re-identification, such as illumination, resolution, occlusions and background clutters. Thanks to the development of deep learning, single-modality visible image matching has achieved remarkable performance on benchmark datasets. However, visible image matching is not applicable in low-light scenarios like night-time outdoor scenes or dark indoor scenes. To resilient the related low-light issues, most of surveillance cameras can automatically switch to acquire near infrared images, which are visually different from visible images. When person re-identification is required for the penetration between normal-light and low-light, current person re-identification performance for cross-modality matching between visible images and infrared images cannot be satisfied. Thus, it is necessary to analyze the visible-infrared cross-modality person re-identification further. For visible-infrared cross-modality person re-identification, there are two key challenges as mentioned below: first, the spectrums and visual appearances of visible images and infrared images are significantly different. Visible images contain three channels of red (R), green (G) and blue (B) responses, while infrared images contain only one channel of near infrared responses. This leads to big modality gap. Next, lack of labeled data is still challenged based on manpower-based identification of the same pedestrian across visible image and infrared image. Current multi-modality benchmark dataset contains 500 personal identities only, which is not sufficient for training deep models. Existing visible-infrared cross-modality person re-identification methods mainly focus on bridging the modality gap. The small labeled data problem is still largely ignored by these methods. **Method** To provide prior knowledge for learning cross-modality matching model, we study self-supervised information mining on single-modality data based on auxiliary labeled visible images. First, we propose a data augmentation method called random single-channel mask. For three-channel visible images as input, random masks are applied to preserve the information of only one channel, to realize the robustness of features against spectrum change. The random single-channel mask can force the first layer of convolutional neural network to learn kernels that are specific to R, G or B channels for extracting shared appearance shape features. Furthermore, for pre-training and fine-tuning, we propose mutual learning between single-channel model and three-channel model. To mine and transfer cross-spectrum robust self-supervision information, mutual learning leverages the interrelations between single-channel data and three-channel data. We sort out that the three-channel model focuses on extracting color-sensitive features, and the single-channel model focuses on extracting color-invariant features. Transferring complementary knowledge by mutual learning improves the matching performance of the cross-modality matching model. **Result** Extensive comparative experiments were conducted on SYSU-MM01, RGBNT201 and RegDB datasets. Compared with the state-of-the-art methods, our method improve mean average precision (mAP) on RGBNT201 by 5% at most. **Conclusion** We propose a single-modality cross-spectrum self-supervised information mining method, which utilizes auxiliary single-modality visible images to mine cross-spectrum robust self-supervision information. The prior knowledge of the self-supervision information can guide single-modality pretraining and multi-modality finetuning for achieving better matching ability of the cross-modality person re-identification model.

Key words: person re-identification; cross-modality retrieval; infrared image; self-supervised learning; mutual learning

0 引 言

随着城市监控系统的完善,对监控视频进行智能分析的需求越发迫切。行人重识别作为智能监控视频分析的基础技术受到越来越多的关注。其任务是跨无交叠视域的摄像头进行行人图像的匹配。由于不同摄像头下采集的行人图像受到光照、分辨率、遮挡和背景变化等影响,这些场景因素导致的数据分布偏移是行人重识别的难点。随着深度学习的迅速发展,基于有监督学习与弱监督学习的算法,行人重识别模型在公开标准数据集上已经能达到很高的性能。

然而,现有的研究大部分集中在基于可见光图像的行人重识别。对于可见光图像无法适用的应用场景,如在跨白天和夜晚或跨室外与室内的情况下,行人图像的外观会受到显著光照变化的影响。在正常光照场景下通常使用可见光图像。而在低照度场景中由于采集的可见光图像质量退化严重,其中包含的信息不具有判别性,难以从中提取特征进行匹配。在这种情况下监控摄像头通常会切换为采集近红外图像,克服光照不足的影响。可见光图像包含

红(R)、绿(G)、蓝(B)3 个通道,而红外图像包含单个通道。由于成像原理不同,可见光图像与红外图像属于不同模态的数据,存在显著的视觉差异(如图 1 左侧),使得现有针对可见光行人图像的方法难以适用。为克服显著光照变化的影响,有必要研究“可见光—红外”跨模态行人重识别问题。

目前,“可见光—红外”跨模态行人重识别的研究主要围绕设计能有效消除模态鸿沟的跨模态匹配算法开展,但是性能仍然不理想。除了不同模态数据的显著视觉差异导致的模态鸿沟外,数据难以标注也是一个限制模型性能的重要问题。目前公开的多模态行人数据集的训练集身份总数均不超过 500,对于训练深度学习模型仍然不够。如图 1 左侧,由于在红外图像中缺失可见光图像的颜色信息,视觉模糊度高使人工观察也很难分辨行人图像是否属于同一个人,导致人工标注跨模态的样本比一般情况下标注同模态的样本耗时更长以及成本更高。

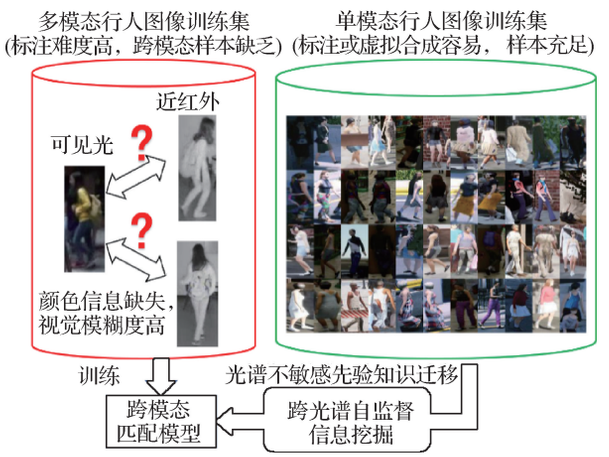


图 1　基于单模态数据辅助的跨模态行人重识别示意图

Fig. 1　Cross-modality person re-identification based on extra auxiliary single-modality data

在有标注的多模态数据量有限的情况下,从其他领域迁移对跨模态匹配有帮助的先验知识是其中一种重要的解决思路。如图 1,本文提出使用单模态可见光行人图像作为辅助,从中挖掘对光谱范围不敏感的特征,并把这种先验知识迁移到基于有限的有标注多模态训练数据学习的跨模态匹配模型中,以提高其判别能力。单模态数据的标注相比跨模态数据的标注更加容易,用于辅助的可见光行人图像的获取可以选择用人工标注,也可以选择更容易获得标签的 3 维合成虚拟行人数据。

面向跨模态行人重识别的任务,针对模态鸿沟与有标注训练数据有限的问题,本文从单模态自监督信息挖掘的角度开展研究,利用额外的单模态可见光图像作为辅助挖掘对光谱范围不敏感的特征,通过预训练模型初始化与下游任务微调把先验知识迁移到跨模态匹配模型中,提高多模态数据有限情况下的判别性能。本文的创新点如下:

1)提出一种随机单通道掩膜的数据增强方法来提取通道共享的特征,使模型对成像光谱变化不敏感;

2)提出一种基于三通道与单通道双模型互学习的预训练与微调方法,从三通道与单通道的关系中挖掘自监督信息引导模型学习鲁棒的跨模态匹配特征。

1　国内外研究现状与相关工作

1)基于可见光图像的行人重识别。近年来,行人重识别研究(罗浩 等,2019)快速发展,技术日趋成熟。现有的行人重识别方法研究主要集中在可见光图像和视频的理解上。行人重识别技术经历了从手工特征设计(Liao 等,2015)到距离度量学习(Zheng 等,2013)和端到端深度学习(Ahmed 等,2015)的快速发展。大多数现有的行人重识别研究都从可见光图像中提取视觉表观特征,然后学习计算相似度进行匹配。其难点在于姿态变化与遮挡(史维东 等,2020)、多分辨率(沈庆 等,2020)等方面。

虽然有监督学习(Sun 等,2018;Ye 等,2022)、弱监督学习(Meng 等,2021)和无监督学习(Ge 等,2020;Zheng 等,2021b;Wei 等,2018;Yu 等,2020)的方法都已经可以在基于可见光图像的行人重识别研究中取得很好的性能,但这些方法仍然未能解决开放环境中的行人重识别问题,如光照变化强烈的跨模态行人重识别场景(Wu 等,2017)、换衣行人重识别场景(Yang 等,2021)、细粒度行人重识场景(Yin 等,2020)、跨分辨率行人重识别场景(Zheng 等,2022)以及基于群体验证的场景(Zheng 等,2016)等。

2)跨模态行人重识别。为解决不同场景光照变化强烈的问题,“可见光—红外”跨模态行人重识别(陈丹 等,2020)的研究主要围绕能有效消除模态

鸿沟的跨模态匹配算法开展,但是性能仍然不理想。Wu 等人(2017)首次开展跨模态行人重识别的研究,并公开了首个包含可见光图像与红外图像的多模态行人重识别数据集。之后,跨模态行人重识别的研究逐渐开始发展。Ye 等人(2018)提出基于双流网络的方法 HCML (hierarchical cross-modality metric learning),利用双流网络消除模态差异以及通过度量学习得到更稳定的跨模态匹配。在模态鸿沟消除方法的设计上,还发展了一些代表性的方法,包括基于图像和特征联合对齐的 D^2RL(dual-level discrepancy reduction learning)(Wang 等,2019b)、基于生成对抗学习的 cmGAN(cross-modality generative adversarial network)(Dai 等,2018)与 AlignGAN(aligned genereative adversarial network)(Wang 等,2019a)、基于跨模态相似度保持的 CMSP(cross-modality similarity preservation)(Wu 等,2020)、基于第 3 模态生成的 XIV-ReID(x-infrared-visible re-identification)(Li 等,2020)和 MMG(middle modality generator)(Zhang 等,2021b)、基于多模态图像混合的 CMM(class-aware modality mix)(Ling 等,2020)、基于协同注意力机制的 CoAL(co-attentive lifting)(Wei 等,2020)、基于模态"特有—共享"特征迁移的 cmSSFT(cross-modality shared-specific feature transfer)(Lu 等,2020)、基于模态和模式联合对齐的(joint modality and pattern alignment network,MPANet)(Wu 等,2021)、基于模态混淆学习网络的方法(modality confusion learning network,MCLNet)(Hao 等,2021)、基于融合模态联合学习的方法(syncretic modality collaborative learning,SMCL)(Wei 等,2021)、基于密集关键点配对的方法(learning by aligning,LBA)(Park 等,2021)和基于多特征空间联合优化的方法(multi-feature space joint optimization,mSO)(Gao 等,2021)等。针对可见光图像与红外图像的模型设计开始引入自动机器学习的思想,发展了基于特征搜索的 NFS(neural feature search)方法(Chen 等,2021)与基于架构搜索的 CM-NAS(cross-modality neural architecture search)方法(Fu 等,2021)。Tian 等人(2021)基于信息瓶颈理论,提出变分自蒸馏的方法避免互信息的显式估计。

基于通道增强联合学习的 CAJL(channel-augmented joint learning)方法(Ye 等,2021)与跨光谱图像生成方法(Fan 等,2020)都是与本文研究高度相关的方法。它们通过分离可见光图像的 RGB 通道分别做数据增强来使数据更接红外图像的模态。一方面,在单通道图像特征提取上,它们对不同通道学习共享的卷积核,而本文使用的随机单通道掩膜相当于在卷积网络第 1 层学习通道特有的卷积核,通过识别通道特有的纹理更好地提取通道共享的外观形状特征。另一方面,本文进一步通过互学习探索了从单模态数据得到的三通道图像与单通道图像的关系来挖掘自监督信息用于跨模态匹配。除了消除模态鸿沟的研究思路,有小部分研究从单模态数据中迁移先验知识来帮助跨模态匹配的学习。Liang 等人(2021)提出一种跨模态自训练方法,利用单模态预训练的模型通过跨模态伪标签学习无监督地提高跨模态匹配的性能。与现有相关方法对比,本文从单模态自监督信息挖掘的新角度,学习有利于跨模态匹配的先验知识,克服模态鸿沟的问题。

3)互学习。深度互学习(Zhang 等,2018)是与本文密切相关的方法。其主要思想是训练多个模型,使其互为教师模型和学生模型,通过知识蒸馏(Hinton 等,2015)互相迁移学习到的知识从而提高模型的判别能力。互学习的思想在行人重识别的研究领域也受到关注。例如,在单模态无监督行人重识别方法 MMT(mutual mean-teaching)(Ge 等,2020)和跨模态有监督行人重识别方法 MPANet(Wu 等,2021)中也使用了互学习。MMT 用互学习探索当前模型与平均模型之间的关系,来获取更可靠的伪标签用于单模态无监督学习。MPANet 用互学习拉近不同模态的输出消除模态鸿沟。本文通过把单模态数据变换为三通道与单通道两个视角,并通过它们之间的互学习挖掘有助于跨模态匹配的自监督信息,不受限于以往方法训练过程中对多模态数据同时存在的要求。

2 单模态跨光谱自监督信息挖掘

2.1 问题与符号定义

为建模"可见光—红外"行人重识别问题,首先对需要使用的符号进行定义。假设在实际场景中采集和标注的多模态行人数据集表示为可见光图像集合 $\boldsymbol{D}_{\mathrm{RGB}}=\{(\boldsymbol{I}_i^{\mathrm{RGB}},y_i^{\mathrm{RGB}})\}_{i=1}^{N_{\mathrm{RGB}}}$ 与红外图像集合 $\boldsymbol{D}_{\mathrm{IR}}=\{(\boldsymbol{I}_j^{\mathrm{IR}},y_j^{\mathrm{IR}})\}_{j=1}^{N_{\mathrm{IR}}}$,其中 $\boldsymbol{I}_i^{\mathrm{RGB}}$ 表示包含红(R)、绿(G)、

蓝(B)三通道的可见光图像，$\boldsymbol{I}_j^{\mathrm{IR}}$ 表示单通道的红外图像，y_i^{RGB} 与 y_j^{IR} 表示身份标签，N_{RGB} 和 N_{IR} 表示集合的样本数量。数据集的行人身份数量通常比较有限。

通过人工标注或者 3 维合成的方式，获得大量辅助的行人可见光图像数据 $\boldsymbol{D}_{\mathrm{RGB}-A}=\{(\boldsymbol{I}_k^{\mathrm{RGB}-A},y_k^{\mathrm{RGB}-A})\}_{k=1}^{N_{\mathrm{RGB}-A}}$。

本文目标是在辅助的可见光图像数据集 $\boldsymbol{D}_{\mathrm{RGB}-A}$ 上进行预训练，挖掘对光谱范围变化具有稳定性的特征。然后在真实的多模态行人数据集 $\boldsymbol{D}_{\mathrm{RGB}}$ 和 $\boldsymbol{D}_{\mathrm{IR}}$ 上进行微调时，把预训练模型中的先验知识通过初始化参数迁移到跨模态行人匹配的下游任务上，提高模型的判别性能。

2.2　随机单通道掩膜数据增强

跨模态行人重识别问题中的模态鸿沟是由可见光图像和红外图像的成像原理不同导致的。可见光的红、绿、蓝通道与红外光的通道上的灰度值表示对物体反射不同波长光线的强度。可见光图像中的 3 个通道共同反映了可见光的颜色信息。在使用单模态可见光图像进行行人特征学习的时候，三通道共同反映的颜色信息是重要的判别性特征。然而，由于光谱范围不同，这对红外图像却无法适用。本文模型可以学习到跨光谱不变的特征。

可见光图像示例如图 2(RGB 图像)所示。分离的红、绿、蓝 3 个通道无法反映三通道图像丰富的颜色信息。当进行不同通道图像的对比时，比如第 1 个通道和第 3 个通道，发现行人的上下身衣服灰度值均发生了变化。如果要对不同通道的图像进行身份匹配，也就是跨光谱的匹配，学习的特征需要包含更加丰富的行人形状等细粒度信息。由于可见光图像与红外图像的匹配同样是一种跨光谱的匹配，假设对红、绿、蓝通道跨光谱匹配具有判别性的特征，在红外图像上也具有适用性。假设从不同通道的图像中提取共享的外观形状特征需要识别不同的纹理。在神经网络浅层使用通道特有的卷积核。

基于上述思路，对可见光图像提出一种随机单通道掩膜数据增强方法。构造 3 种通道掩膜 $\boldsymbol{m}_{\mathrm{R}}=[1,0,0]$、$\boldsymbol{m}_{\mathrm{G}}=[0,1,0]$ 与 $\boldsymbol{m}_{\mathrm{B}}=[0,0,1]$，作用是分别只保留 RGB 图像其中一个通道的信息，其他通道的灰度值都变为 0。在训练过程中，每次迭代都对当前的输入图像 $\boldsymbol{I}_i^{\mathrm{RGB}}\in\mathbf{R}^{H\times W\times 3}$ 随机选择一个通道掩膜 $\boldsymbol{m}$ 进行数据增强操作 aug，即

$$aug(\boldsymbol{I}_i^{\mathrm{RGB}},\boldsymbol{m})=concat(m_1\boldsymbol{I}_i^{\mathrm{R}},m_2\boldsymbol{I}_i^{\mathrm{G}},m_3\boldsymbol{I}_i^{\mathrm{B}}) \quad (1)$$

式中，$\boldsymbol{I}_i^{\mathrm{R}}$，$\boldsymbol{I}_i^{\mathrm{G}}$，$\boldsymbol{I}_i^{\mathrm{B}}\in\mathbf{R}^{H\times W\times 1}$ 分别是图像 $\boldsymbol{I}_i^{\mathrm{RGB}}$ 的红、绿、蓝通道，$concat(\cdot)$ 表示在通道的维度进行拼接。具体操作如图 2 所示。

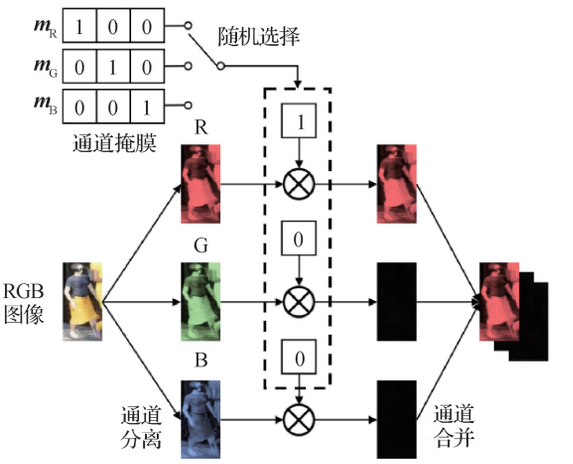


图 2　随机单通道掩膜数据增强示意图

Fig. 2　Data augmentation by random single-channel masks

由于被随机掩膜提取的输入通道之外的两个输入通道为 0，使用随机单通道掩膜相当于使卷积神经网络的第 1 层学习 R、G、B 通道特有的卷积核(Wu 等，2017)，通过在网络浅层识别通道特有的纹理更好地提取通道共享的外观形状特征。

2.3　三通道与单通道双模型互学习

2.3.1　基于单模态可见光图像的双模型预训练

1)三通道与单通道模型训练。本文基于单模态可见光图像数据 $\boldsymbol{D}_{\mathrm{RGB}-A}$ 训练对提高跨模态匹配能力有帮助的模型。基于随机单通道掩膜数据增强训练模型有利于学习到跨光谱不变的特征。对于辅助的 RGB 图像 $\boldsymbol{I}_i^{\mathrm{RGB}-A}$，根据式(1)可以得到其进行随机单通道掩膜数据增强操作(aug)后的图像 $\boldsymbol{I}_i^{\mathrm{sin}}$，即

$$\boldsymbol{I}_i^{\mathrm{sin}}=aug(\boldsymbol{I}_i^{\mathrm{RGB}-A},\boldsymbol{m}) \quad (2)$$

在可见光图像与红外图像匹配的下游任务中，特征也需要在可见光图像模态内具有判别性，所以也需要从三通道数据 $\boldsymbol{I}_i^{\mathrm{RGB}-A}$ 中学习特征。

用标签 $y_i^{\mathrm{RGB}-A}$ 作为监督信号，基于增强后的单通道数据 $\boldsymbol{I}_i^{\mathrm{sin}}$ 与三通道数据 $\boldsymbol{I}_i^{\mathrm{RGB}-A}$ 分别训练单通道特征提取模型 M_{sin} 与三通道特征提取模型 M_{RGB}，模型提取的特征分别表示为 $\boldsymbol{f}_i^{\mathrm{sin}}=M_{\mathrm{sin}}(\boldsymbol{I}_i^{\mathrm{sin}})$ 和 $\boldsymbol{f}_i^{\mathrm{RGB}}=M_{\mathrm{RGB}}(\boldsymbol{I}_i^{\mathrm{RGB}-A})$，通过分类器后分别得到分类概率向

量 $\boldsymbol{p}_i^{\text{sin}}$ 和 $\boldsymbol{p}_i^{\text{RGB}}$。用于单通道模型特征学习的判别损失函数表示为

$$L_{\text{ID-sin}} = L_{\text{cls-sin}} + L_{\text{tri-sin}} \tag{3}$$

判别损失函数 $L_{\text{ID-sin}}$ 由交叉熵分类损失 $L_{\text{cls-sin}}$ 和软间隔三元组损失 $L_{\text{tri-sin}}$（Hermans 等，2017）两部分组成。

（1）交叉熵分类损失 $L_{\text{cls-sin}}$ 表示为

$$L_{\text{cls-sin}} = -\frac{1}{N_{\text{RGB-}A}}\sum_{i=1}^{N_{\text{RGB-}A}}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}^{\text{sin}}) \tag{4}$$

式中，$p_{i,c}^{\text{sin}}$ 和 $y_{i,c}$ 分别是分类概率向量 $\boldsymbol{p}_i^{\text{sin}}$ 和样本 $\boldsymbol{I}_i^{\text{sin}}$ 的 one-hot 类别标签 $y_{i,c}$ 的第 c 个元素，C 是总类数。

（2）软间隔三元组损失 $L_{\text{tri-sin}}$ 表示为

$$L_{\text{tri-sin}} = \sum_{(a,n,p)\in \boldsymbol{Ind}_{\text{tri}}} \log(1+\exp(dist(\boldsymbol{f}_a^{\text{sin}},\boldsymbol{f}_p^{\text{sin}}) - dist(\boldsymbol{f}_a^{\text{sin}},\boldsymbol{f}_n^{\text{sin}}))) \tag{5}$$

式中，(a,p,n) 表示三元组下标集合 $\boldsymbol{Ind}_{\text{tri}}$ 中的元素。$\boldsymbol{f}_a^{\text{sin}}$，$\boldsymbol{f}_p^{\text{sin}}$，$\boldsymbol{f}_n^{\text{sin}}$ 分别表示特征三元组中的锚点、正样本和负样本。$dist$ 表示特征之间的欧氏距离，函数 $\log(1+\exp(\cdot))$ 的作用是使正负样本对的间隔软化。三元组的采样使用了难样本挖掘的策略（Hermans 等，2017），在每个批次的样本中寻找距离最近的负样本对与距离最远的正样本对。

对于三通道特征提取模型 M_{RGB}，学习的目标函数 $L_{\text{ID-RGB}}$ 可类比单通道特征提取模型 M_{sin} 的目标函数 $L_{\text{ID-sin}}$ 进行构建。

2）三通道与单通道模型互学习。三通道特征提取模型 M_{RGB} 是从 RGB 图像学习得到的。对于与行人身份相关的判别性信息，从 RGB 图像中既可以提取如色度、饱和度等与颜色相关的特征，也可以提取形状、纹理等与颜色无关的特征。虽然两种特征都包含可以区分行人的信息，但模型会趋向于学习与颜色相关的特征，可以为下游任务中可见光图像的模态内匹配提供先验知识。

单通道特征提取模型 M_{sin} 是从经过随机单通道掩膜数据增强后的单通道数据中学习得到的。由于模型输入的单通道数据中的信息是三通道 RGB 图像的子集，在缺失了不同通道组合的情况下，难以提取到色度、饱和度等与颜色相关的特征。模型会更趋向于学习形状、纹理等与颜色无关的特征。这种特征具有对光谱范围变化不敏感的特点，可以为下游任务中红外图像的模态内匹配和可见光与红外图像之间的跨模态匹配提供先验知识。

为了直观地理解三通道模型与单通道模型学习到的特征，按照上述介绍的训练方法，在 UnrealPerson（Zhang 等，2021a）数据集上训练了以 ResNet-50（He 等，2016）为骨干模型的 M_{RGB} 和 M_{sin}。然后，在 DukeMTMC（Duke multi-target multi-camera）（Ristani 等，2016）数据集的 RGB 图像上进行测试。随机选择训练集中的一幅 RGB 图像，分别根据两个模型提取的特征的欧氏距离检索训练集中最相似的图像，得到排序列表如图 3 所示。在排序列表中，为显示更多不同身份的行人，同一个身份的行人图像只保留最靠前的一幅。可以观察到，对于同一幅查询图像，三通道模型检索到的图像衣着颜色与查询图像高度相似，其中也包含与查询图像中的男性行人外观形状不同的女性行人（如图 3（a）红框中的第 3 幅和第 6 幅图中的女性发型与男性短发形状不同，第 7 幅图中的女性腿型比男性腿型细）；单通道模型检索到的图像则都是与查询图像中的男性行人外观形状接近的其他男性行人，但衣着颜色则未必相近（如图 3（b）红框中的第 5、7、9 幅图都是与查询图像的行人体型相近的短发男性，但衣服颜色不同）。观察结果与上述三通道模型和单通道模型提取特征的特点相符。

(a) 三通道模型检索排序列表

(b) 单通道模型检索排序列表

图 3　三通道模型与单通道模型检索排序列表对比

Fig. 3　Comparison of retrieval ranking list between three-channel model and single-channel model

((a) three-channel model; (b) single-channel model)

上述的三通道特征提取模型 M_{RGB} 和单通道特征提取模型 M_{sin} 是单独训练的，会受到与颜色相关特征和与颜色无关特征学习倾向性的影响。在跨模态匹配的下游任务中，期望预训练模型可以联合两种特征的学习，对模态内匹配和模态间匹配都能提供有效的先验知识。因此，提出单通道模型与双通道模型的互学习方法，使三通道模型与单通道模型互为教师模型与学生模型，互相指导和约束对方的

学习。由于分类概率 $\boldsymbol{p}_i^{\text{sin}}$ 和 $\boldsymbol{p}_i^{\text{RGB}}$ 中包含不同身份行人之间的相似性关系，参照 Zhang 等人(2018)提出的深度互学习方法，构造互学习损失函数

$$L_{\text{mu}} = \frac{1}{N_{\text{RGB}-A}} \sum_{i=1}^{N_{\text{RGB}-A}} D_{KL}(\boldsymbol{p}_i^{\text{sin}} \| \boldsymbol{p}_i^{\text{RGB}}) + D_{\text{KL}}(\boldsymbol{p}_i^{\text{RGB}} \| \boldsymbol{p}_i^{\text{sin}}) \tag{6}$$

式中，D_{KL} 表示 KL(Kullback-Leibler)散度。其目标是用 Jensen-Shannon 散度度量最小化 $\boldsymbol{p}_i^{\text{sin}}$ 和 $\boldsymbol{p}_i^{\text{RGB}}$ 两个分类概率分布之间的距离。

结合判别损失函数和互学习损失函数得到预训练的总体损失函数

$$L_{\text{pre}} = L_{\text{ID-sin}} + L_{\text{ID-RGB}} + w_{\text{mu}} L_{\text{mu}} \tag{7}$$

式中，w_{mu} 是控制互学习损失 L_{mu} 影响的权重参数。

在三通道模型与单通道模型之间进行互学习，一方面可以促进单通道模型对在缺乏颜色信息情况下容易忽略的特征的提取，另一方面可以促进三通道模型对光谱范围不敏感特征的提取。

由于单通道数据和三通道数据是从同一个数据变换得到的两个不同视角的输入，单通道模型和三通道模型的互学习可以看做是从三通道和单通道的关系中挖掘有利于跨模态匹配的自监督信息。只需作为辅助数据的单模态可见光图像，即可为下游的跨模态匹配任务提供先验知识。三通道与单通道模型互学习的示意图如图 4 所示。

2.3.2　基于多模态数据的双模型微调

在基于互学习的双模型预训练后，三通道模型 M_{RGB} 和单通道模型 M_{sin} 学习到两种不同的先验知识。虽然在互学习中两个模型的知识会相互补充，但由于输入数据的不同，两个模型仍分别侧重颜色相关特征与光谱范围不敏感特征的提取。为更有效地在下游任务中利用两种先验知识帮助有监督学习，避免预训练与微调的学习目标之间产生差异，使用与双模型预训练相同的框架(如图 4 所示)，在多模态数据上进行双模型微调。

训练的数据包括可见光图像集合 $\boldsymbol{D}_{\text{RGB}} = \{(\boldsymbol{I}_i^{\text{RGB}}, y_i^{\text{RGB}})\}_{i=1}^{N_{\text{RGB}}}$ 与红外图像集合 $\boldsymbol{D}_{\text{IR}} = \{(\boldsymbol{I}_j^{\text{IR}}, y_j^{\text{IR}})\}_{j=1}^{N_{\text{IR}}}$。与使用单模态可见光图像的预训练不同，进行微调时需要同时使用三通道可见光图像 $\boldsymbol{I}_i^{\text{RGB}}$ 与单通道红外图像 $\boldsymbol{I}_i^{\text{IR}}$ 两个模态的数据作为三通道模型 M_{RGB} 或者单通道模型 M_{sin} 的输入。其中，可见光图像 $\boldsymbol{I}_i^{\text{RGB}}$ 的输入处理方式与预训练时可见光图像的输入处理方式一致。对于单通道的红外图像 $\boldsymbol{I}_i^{\text{IR}}$，在输入三通道模型 M_{RGB} 时，把单通道复制成三通道。在预训练过程中，对三通道 RGB 图像进行随机单通道掩膜数据增强输入到单通道模型 M_{sin} 时，由于只有一个输入通道被掩膜提取出来而另外两个输入通道为 0，相当于使卷积神经网络的第 1 层学习到 R、G、B 通道特有的卷积核(Wu 等，2017)。本文假设这些通道特有卷积核均有利于在红外图像上提取跨光谱不变的特征。为了在微调过程中充分利用预训练过程中已学习到的所有通道特有卷积核器，把红外图像的单通道复制为三通道再

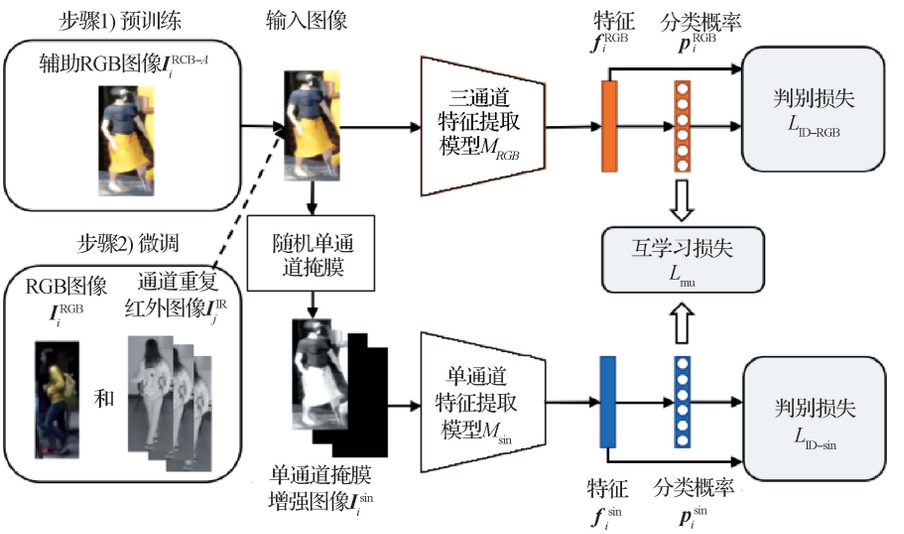


图 4　三通道与单通道双模型互学习示意图

Fig. 4　Mutual learning between three-channel model and single-channel model

进行单通道掩膜数据增强,可使红外图像分别通过不同的R、G、B通道特有的卷积核提取特征进行互学习。

微调训练的目标函数 L_{fine} 参照式(7)中的 L_{pre} 构造。与预训练的区别在于输入训练数据的不同。微调过程除了使用可见光图像,还增加了通过上述预处理转换成三通道的红外图像。

在基于多模态数据的双模型微调中,三通道模型与单通道模型的互学习起到的作用与预训练过程类似,从可见光图像的三通道数据与单通道数据之间挖掘得到的自监督信息可提供先验知识作为正则化,提高跨模态匹配特征的判别性。

2.3.3 模型推断

在完成基于三通道与单通道模型互学习的预训练与微调之后,测试阶段由于有两个不同的模型,采用不同的推断方式。

1)三通道模型推断。对于可见光图像,直接输入三通道模型提取特征。对于红外图像,把单通道复制为三通道输入三通道模型提取特征。

2)单通道模型推断。由于单通道模型训练时的输入是应用了不同通道掩膜的图像,为保持训练和测试输入的一致性,在推断之前需要对测试图像进行预处理。对于可见光图像,参照1.2节分别使用掩膜 $\boldsymbol{m}_{\text{R}}$、$\boldsymbol{m}_{\text{G}}$ 和 $\boldsymbol{m}_{\text{B}}$ 得到3幅只包含单通道信息的图像,分别输入单通道模型进行特征提取。对于红外图像,首先把单通道复制为三通道,然后采用与可见光图像相同的方式进行掩膜处理与特征提取。最后,把提取的3个特征取平均得到融合的特征。单通道模型特征提取过程如图5所示。

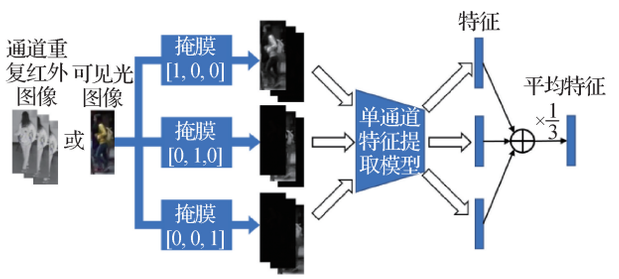


图5 单通道模型特征提取示意图

Fig. 5 Feature extraction of single-channel model

3)双模型融合推断。在计算资源允许的情况下,可以通过特征串联的方式,融合两个模型的输出作为特征。

在提取特征后,通过度量查询图像和图库图像的特征欧氏距离进行检索。

3 实 验

本文在SYSU-MM01(Wu等,2020)、RGBNT201(Zheng等,2021a)和RegDB(Nguyen等,2017)这3个多模态数据集上测试提出的基于单模态跨光谱自监督信息挖掘的预训练与微调方法,与当前先进的方法进行对比,并进行了消融实验、使用不同超参数与预训练数据集的实验。

3.1 实验设置

1)数据集。图6展示了3个数据集的一些示例图像。

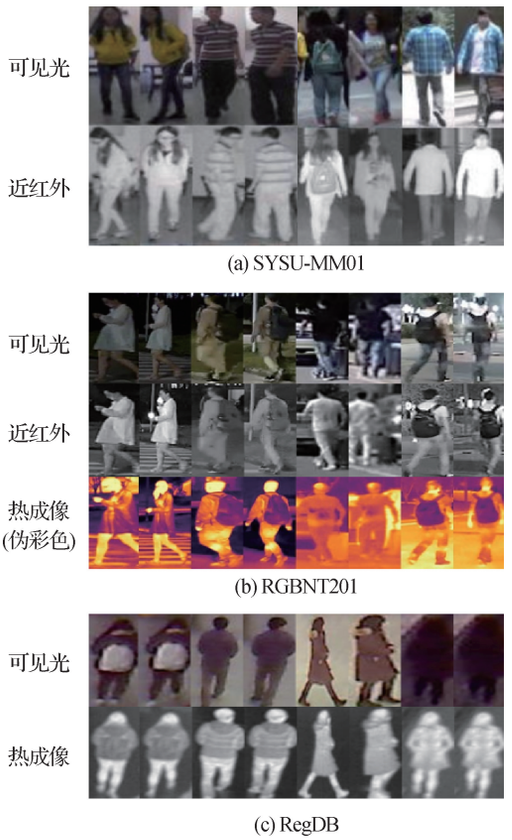


图6 实验中各个数据集的样本示例

Fig. 6 Examples of datasets used for evaluations

((a) SYSU-MM01; (b) RGBNT201; (c) RegDB)

SYSU-MM01数据集由6个摄像头拍摄,其中4个是正常光照环境下的可见光摄像头,2个是黑暗环境下的近红外摄像头。拍摄的场景包括2个室内场景与3个室外场景。不同场景下的图像有光照、背景等场景变化。行人身份总数为491个,可见光

图像数量为 30 071 幅，近红外图像数量为 15 792 幅。

RGBNT201 数据集由 4 个摄像头拍摄，其中每个摄像头拍摄了同步的可见光图像、近红外图像与热成像图像。场景变化包括有天气、光照等。行人身份总数为 201 个，可见光图像、近红外图像与热成像图像的数量均为 4 787 幅。

RegDB 数据集包含了一个可见光摄像头和一个热成像摄像头拍摄的 412 个身份的行人的 8 240 幅图像。对于每个身份，有 10 幅可见光图像和 10 幅热成像图像。

在提出方法的预训练过程中，需要辅助的有标注可见光行人图像数据集。在默认的实验设置下，采用 3 维合成的虚拟数据集 UnrealPerson（Zhang 等，2021a）作为辅助数据集。UnrealPerson 是 3 维合成的大规模数据集，包含了 3 000 个身份的 120 000 幅行人图像，无需人工进行标注。使用虚拟数据进行预训练可避免使用真实数据的隐私问题。

2）测试协议。在 SYSU-MM01 数据集上，遵循 Wu 等人（2020）对训练集和测试集中查询图像和图库图像的划分。“全搜索”表示在全部摄像头的数据组成的图库图像中的搜索实验；“室内搜索”表示在室内摄像头的数据组成的图库图像中的搜索实验，难度比“全搜索”稍低。

在 RGBNT201 数据集上，遵循了数据集提出者 Zheng 等人（2021a）的测试协议。训练集身份数为 141，测试集身份数为 30。由于 RGBNT201 上有 3 种模态的图像，跨模态匹配分为 4 种情况：“可见光—近红外”表示把近红外图像作为图库图像，把可见光图像作为查询图像；“可见光—热成像”表示把热成像图像作为图库图像，把可见光图像作为查询图像；“近红外—可见光”和“热成像—可见光”是将上述两种情况的图库图像和查询图像种类交换。

对于 RegDB 数据集，遵循了现有方法 HCML（Ye 等，2018）中使用的测试协议。一半的身份用于训练，其他身份用于测试。跨模态匹配的方式有两种：“热成像—可见光”表示把热成像图像作为查询图像，把可见光图像作为图库图像，“可见光—热成像”表示把可见光图像作为查询图像，把热成像图像作为图库图像。

实验的性能指标是通过度量查询图像和图库图像之间的相似度获得的排序列表计算得到的累积匹配特性 CMC（cumulative match characteristic）、Rank-k 正确率和平均精度均值（mean average precision，mAP），参照 Zheng 等人（2015）在 Market-1501 数据集上的计算方法。

3）实现细节。（1）基础模型。在实现中采用了 ResNet-50（He 等，2016）作为骨干模型，把输入图像的尺寸调整为 384 × 128 像素，然后在输出的特征图上分割水平条带提取特征，具体参照 MGN（multiple granularity network）（Wang 等，2018）的模型设计。分类器层参照 circle 损失（Sun 等，2020）的实现方式，其中 circle head 的间隔参数设为 0.35，特征的尺度设为 64。然后使用平滑参数为 0.1 的标签平滑操作，得到最后的分类概率。提出方法中的单通道模型和三通道模型均基于此模型实现。

（2）训练策略。由于模型参数用 ImageNet 预训练的参数初始化，输入数据的预处理先使用 ImageNet 的均值和标准差进行图像的归一化。然后应用随机水平翻转、随机裁剪、随机擦除和颜色抖动的数据增强策略。数据增强策略的参数参照 He 等人（2020）提出的 FastReID 中的策略设置。在颜色抖动策略中设置亮度变化范围为[0.8，1.2]，对比度变化范围为[0.85，1.15]。训练过程分为预训练和微调两个步骤。在预训练过程中，默认使用单模态可见光图像数据集 UnrealPerson。在微调过程中，使用需要测试的目标数据库的训练集数据。在梯度下降的的迭代过程中，使用 ADAM（adaptive moment estimation）优化器（Kingma 和 Ba，2015）。预训练步骤分双模型单独训练和双模型互学习两个阶段。第 1 阶段的迭代总数为 15 000 次，前 2 000 次迭代使用 warmup 策略使学习率从 3.5×10^{-6} 增加到 3.5×10^{-4}，之后的 7 000 次迭代保持学习率不变，最后 6 000 次迭代使用 Cosine 学习率下降策略。第 2 阶段的训练设置与第 1 阶段相同，区别在于加入了互学习损失 L_{mu}，其权重 w_{mu} 设置为 0.1。微调步骤也参照预训练步骤分两阶段进行，区别在于每阶段迭代总次数变为 8 000 次，去除了预训练步骤中保持学习率不变的 7 000 次迭代。

（3）测试过程。默认使用三通道模型进行推断。在展示的实验结果中，“三通道模型”“单通道模型”和“双模型融合”均表示提出的方法。

4）对比方法。在 SYSU-MM01 与 RegDB 数据集上，对比了当前具有代表性的、先进的跨模态行人重

识别方法,包括基于非对称建模的 Zero-Padding(Wu 等,2017)、基于图像和特征联合对齐的 D^2RL(Wang 等,2019b)、基于生成对抗学习的 AlignGAN(Wang 等,2019a)、基于跨模态相似度保持的 CMSP(Wu 等,2020)、基于第 3 模态生成的 XIV-ReID(Li 等,2020)、基于多模态图像混合的 CMM + CML(Ling 等,2020)、基于协同注意力机制的 CoAL(Wei 等,2020)、基于模态“特有—共享”特征迁移的 cm-SSFT(Lu 等,2020)、基于密集关键点配对的方法 LBA(Park 等,2021)、基于神经网络架构搜索的方法 CM-NAS(Fu 等,2021)、基于通道增强联合学习的 CAJL(Ye 等,2021)和基于模态和模式联合对齐的 MPANet(Wu 等,2021)。在 RGBNT201 数据集上,由于目前已经公开测试过的方法较少,只有 TSLFN + HC(hetero-center loss)(Zhu 等,2020)和 DDAG(dynamic dual-attentive aggregation)(Ye 等,2020)有公开报告的结果。其中,CAJL 是一种单通道数据增强方法。MPANet 应用了多模态的互学习,是当前在 SYSU-MM01 数据集上性能最高的方法。对于具有代表性的先进方法 LBA、CM-NAS、CAJL 和 MPANet,基于作者公布的代码以 UnrealPerson 数据集预训练的参数作为初始化进行实验,以保证与提本文方法使用相同的训练数据进行公平的对比。在实验结果中表示为“方法名(unreal)”。其中,CM-NAS 由于只在 SYSU-MM01 和 RegDB 两个数据集上提供了模型架构且没有公开架构搜索的代码,不进行在 RGBNT201 数据集上的实验。CM-NAS 的所有实验结果都是基于作者公开代码实现得到的。

3.2 实验结果对比与分析

1)消融实验。为说明方法各个部分的有效性,在 SYSU-MM01 上进行了如表 1 所示的消融实验。展示的结果是全搜索设置下的性能。所有实验都默认在 ImageNet 预训练参数的基础上使用 UnrealPerson 作为进一步预训练的数据集,除了实验 0 只使用 ImageNet 预训练的参数。实验 1 是使用单个模型在可见光数据上进行预训练和在多模态数据上微调的基础模型。在实验 2—实验 11,预训练互学习和微调互学习两列中的内容表示是否有使用互学习的训练策略。“无”表示只使用单个模型进行训练。“有”表示使用了“互学习类型”一列中的策略进行双模型互学习。在预训练有两个模型的情况下,微调使用的单个模型选择三通道模型。在互学习类型中,“A—B”中的 A 和 B 表示互学习的两个模型输入的数据类型。“三通道—单通道掩膜”是本文方法。

表 1 在 SYSU-MM01 数据集上的消融实验性能

Table 1 Ablation study performance on SYSU-MM01 dataset

/%

实验编号	互学习类型	预训练互学习	微调互学习	全搜索		室内搜索	
				mAP	Rank-1	mAP	Rank-1
0(ImageNet)	无	无	无	53.2	53.4	65.5	57.1
1(基础模型)	无	无	无	61.8	62.9	72.6	65.7
2	三通道—三通道	有	无	63.4	63.9	75.3	69.2
3	三通道—乱序通道	有	无	63.5	64.0	74.9	68.6
4	三通道—灰度图	有	无	63.2	63.7	75.3	68.4
5	三通道—跨光谱图像(Fan 等,2020)	有	无	63.3	64.4	74.8	68.7
6	三通道—单通道掩膜(本文)	有	无	64.3	65.0	75.8	69.9
7	三通道—三通道	有	有	63.1	64.8	75.4	69.9
8	三通道—乱序通道	有	有	66.2	67.4	77.1	71.9
9	三通道—灰度图	有	有	66.3	68.6	78.6	73.6
10	三通道—跨光谱图像(Fan 等,2020)	有	有	67.6	70.6	79.4	75.0
11(三通道模型)	三通道—单通道掩膜(本文)	有	有	**69.0**	**71.3**	**80.1**	**76.2**

注:加粗字体表示各列最优结果。

为说明随机单通道掩膜的作用,对比了 3 种引导模型学习颜色无关信息的数据增强方法,分别是使用灰度图作为输入、使用 RGB 三通道随机打乱的图像作为输入(表示为“乱序通道”)以及使用跨光谱图像生成方法(Fan 等,2020)得到的单通道 R、G、B 和灰度图像作为输入(表示为“跨光谱图像”)。

结果表明:对比实验 1 的基础模型,在实验 6 和实验 11 中提出的预训练互学习和微调互学习两个步骤都能带来显著的性能提升。对比实验 2 和实验 6 还有实验 7 和实验 11,结果表明“三通道—单通道掩膜”的互学习比“三通道—三通道”的互学习更有效,说明从三通道和单通道的关系中能学习到的对光谱范围不敏感的特征,对跨模态匹配更有帮助。对比实验 3—实验 6 还有实验 8—实验 11,结果表明使用随机单通道掩膜比使用灰度图、随机打乱通道顺序和跨光谱图像更有助于挖掘跨光谱不变的自监督信息。实验 5 和实验 10 中跨光谱图像与提出的随机单通道掩膜一样使用了分离的单通道 R、G、B 图像,但性能不如本文方法。由于随机单通道掩膜方法进一步考虑了网络第 1 层通道特有卷积核的建模,通过在网络浅层识别通道特有的纹理更有效地提取通道共享的外观形状特征。对比实验 0 和实验 1,说明了 UnrealPerson 虚拟数据的预训练对于行人重识别任务的有效性。

2)与当前先进方法的对比。在 SYSU-MM01、RGBNT201 和 RegDB 这 3 个数据集上与当前先进方法的对比结果分别如表 2—表 4 所示。

从实验结果可得,在 SYSU-MM01 上,在本文方法使用三通道模型或者单通道模型单独进行测试时,取得与 MPANet (unreal)相当的性能。在使用双模型融合的情况下,mAP 和 Rank-1 的准确率取得最优的效果。在 RGBNT201 和 RegDB 数据集上,单独使用本文方法中的三通道模态或者单通道模型对比性能排第 2 的方法均有提升。在 RGBNT201 上,在热成像图像和可见光图像的匹配实验中,Rank-1 准确率和 mAP 有接近 5% 的提升。对于 LBA (unreal)、CM-NAS (unreal)、CAJL (unreal)和 MPANet (unreal)这几种先进的方法,相比使用 ImageNet 预训练的结果,使用 UnrealPerson 进行预训练的结果在 SYSU-MM01 上提升不明显,在 RegDB 上稍有提升,但都不如使用本文双模型互学习方法。对比方法的预训练方法是使用单模态 RGB 图像直接训练单个模型,倾向于学习颜色相关的判别性先验知识,对跨模态匹配帮助不大。而提出的双模型互学习方法可在预训练与微调阶段都从单通道模型迁移跨光谱不变的判别性知识到三通道模型中,更有利于模型提高跨模态匹配的判别性能。

与 SYSU-MM01 相比,本文方法在 RGBNT201 和 RegDB 数据集上的提升更明显。在数据集的规模上,RGBNT201 的训练身份数比 SYSU-MM01 和 RegDB 都少,RegDB 的训练样本数比 SYSU-MM01 少。不同数据集上性能提升的差别说明,在目标域越缺乏监督信息的情况下,提出的自监督信息挖掘方法提供的先验知识对下游任务的提升越大。

双模型融合在大部分情况下都能相比三通道模型和单通道模型有一定的性能提升,说明两个模型在互学习后提取的特征仍具有互补性。在少数情况下,比如在 RGBNT201 上“热成像—可见光”设置下的结果,互学习有可能使双模型的知识比较完全地进行互相迁移,导致双模型互补性变弱,但是双模型融合的结果仍与最优的单模型结果相当。在计算资源允许的情况下,使用双模型提取特征进行融合以获得更好的效果。

3)预训练模型对现有方法的提升作用。为验证提出的预训练方法的通用性,基于在单模态的虚拟行人数据集 UnrealPerson(Zhang 等,2021a)上进行三通道与单通道双模型互学习得到的三通道模型参数作为初始化,使用具有代表性的先进方法 LBA、CM-NAS、CAJL 和 MPANet 进行学习(表示为“方法名 (unreal + 提出的预训练)”),并与使用 UnrealPerson 直接训练单个模型作为初始化的实验(表示为“方法名 (unreal)”)进行对比。

在 SYSU-MM01 数据集全搜索设置、RGBNT201 数据集的“热成像—可见光”设置以及 RegDB 数据集的“热成像—可见光”设置下得到的实验结果如表 5 所示。使用提出的双模型预训练方法(表示为“(unreal + 提出的预训练)”)的性能高于使用一般的单模型预训练方法(表示为“(unreal)”),表明了三通道与单通道双模型预训练的有效性。三通道模型在 SYSU-MM01 数据集上取得与对比方法相当的性能,而在样本数更加受限的 RGBNT201 数据集和 RegDB 数据集上能取得更优的性能。

表 2　在 SYSU-MM01 数据集上的跨模态匹配性能对比

Table 2　Performance comparisons for cross-modality matching on SYSU-MM01 dataset

/%

方法	全搜索			室内搜索		
	mAP	Rank-1	Rank-10	mAP	Rank-1	Rank-10
Zero-Padding（Wu 等,2017）	16.0	14.8	54.1	26.9	20.6	68.4
D^2RL（Wang 等,2019b）	29.2	28.9	70.6	–	–	–
AlignGAN（Wang 等,2019a）	40.7	42.4	85.0	54.3	45.9	87.6
CMSP（Wu 等,2020）	45.0	43.6	86.3	57.5	48.6	89.5
XIV-ReID（Li 等,2020）	50.7	49.9	89.8	–	–	–
CMM + CML（Ling 等,2020）	51.2	51.8	92.7	63.7	55.0	94.4
CoAL（Wei 等,2020）	57.2	57.2	92.3	70.8	63.9	95.4
cm-SSFT（Lu 等,2020）	63.2	61.6	89.2	72.6	70.5	94.9
LBA(Park 等,2021）	54.1	55.4	–	68.0	61.0	–
CM-NAS(Fu 等,2021）	52.4	55.0	89.7	67.3	60.7	93.8
CAJL(Ye 等,2021）	66.9	69.9	95.7	80.3	76.3	97.9
MPANet（Wu 等,2021）	68.2	70.6	96.2	81.0	76.7	**98.2**
LBA（unreal）（Park 等,2021）	54.6	55.4	92.2	68.8	60.9	96.5
CM-NAS（unreal)(Fu 等,2021）	50.6	52.4	89.2	65.3	57.7	94.4
CAJL（unreal)(Ye 等,2021）	67.8	70.5	95.5	79.9	75.9	97.3
MPANet（unreal）（Wu 等,2021）	68.3	71.3	95.7	79.7	75.5	97.9
本文（三通道模型）	69.0	71.3	96.4	80.1	76.2	97.6
本文（单通道模型）	67.8	71.0	96.5	80.0	76.3	97.5
本文（双模型融合）	**70.3**	**73.0**	**96.9**	**81.2**	**77.5**	97.8

注：“–”表示对应论文中没有该结果,加粗字体表示各列最优结果。

表 3　在 RGBNT201 数据集上的跨模态匹配性能对比

Table 3　Performance comparisons for cross-modality matching on RGBNT201 dataset

/%

方法	热成像—可见光		可见光—热成像		近红外—可见光		可见光—近红外	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
TSLFN + HC（Zhu 等,2020）	16.6	13.0	15.7	11.3	22.0	18.4	22.6	26.4
DDAG（Ye 等,2020）	17.0	15.5	18.1	18.4	30.6	34.5	29.5	35.0
LBA（unreal）（Park 等,2021）	18.6	18.1	18.4	17.2	31.4	37.3	30.7	31.9
CAJL（unreal)(Ye 等,2021）	26.8	25.1	28.7	28.0	36.8	38.4	36.6	41.1
MPANet（unreal）（Wu 等,2021）	28.8	27.6	25.0	23.4	34.5	32.1	34.2	37.8
本文（三通道模型）	33.5	**42.5**	31.7	27.4	37.0	40.1	36.7	43.0
本文（单通道模型）	**34.5**	38.0	30.2	28.2	35.6	40.3	35.1	43.0
本文（双模型融合）	34.3	39.7	**32.0**	**28.5**	**38.5**	**43.5**	**37.6**	**44.7**

注：“–”表示对应论文中没有该结果,加粗字体表示各列最优结果。

表 4　在 RegDB 数据集上的跨模态匹配性能对比

Table 4　Performance comparisons for cross-modality matching on RegDB dataset

/%

方法	热成像—可见光		可见光—热成像	
	mAP	Rank-1	mAP	Rank-1
Zero-Padding（Wu 等,2017）	17.9	16.7	18.9	17.8
D^2RL（Wang 等,2019b）	–	–	44.1	43.4
AlignGAN（Wang 等,2019a）	53.4	56.3	53.6	57.9
CMSP（Wu 等,2020）	–	–	64.5	65.1
XIV-ReID（Li 等,2020）	60.2	62.3	–	–
CMM + CML（Ling 等,2020）	60.9	59.8	–	–
CoAL（Wei 等,2020）	69.9	74.1	–	–
cm-SSFT（Lu 等,2020）	71.7	71.0	72.9	72.3
LBA(Park 等,2021)	65.5	72.4	67.6	74.2
CM-NAS(Fu 等,2021)	76.7	80.3	78.0	80.9
CAJL(Ye 等,2021)	77.8	84.8	79.1	85.0
MPANet（Wu 等,2021）	80.7	82.8	80.9	83.7
LBA（unreal）	67.0	72.1	67.7	72.8
CM-NAS（unreal）	83.3	87.3	82.0	86.8
CAJL（unreal）	76.3	84.4	78.9	87.0
MPANet（unreal）	82.2	84.7	83.4	85.2
本文（三通道）	84.8	86.6	85.1	87.0
本文（单通道）	84.2	**90.3**	85.3	88.6
本文（双模型）	**87.0**	89.9	**86.9**	**89.1**

注：“–”表示对应论文中没有该结果，加粗字体表示各列最优结果。

表 5　把本文提出的预训练应用到其他现有方法的性能

Table 5　Performances of applying the proposed pretraining method to other existing methods

/%

方法	SYSU-MM01(全搜索)		RGBNT201(热成像—可见光)		RegDB(热成像—可见光)	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
LBA（unreal）（Park 等,2021）	54.6	55.4	18.6	18.1	67.0	72.1
LBA（unreal + 提出的预训练）	58.0	59.3	20.7	18.9	67.8	73.3
CM-NAS（unreal）(Fu 等,2021）	50.6	52.4	–	–	83.3	87.3
CM-NAS（unreal + 提出的预训练）	52.8	54.3	–	–	84.5	88.4
CAJL（unreal）(Ye 等,2021）	67.8	70.5	26.8	25.1	76.3	84.4
CAJL（unreal + 提出的预训练）	68.8	71.0	29.0	31.9	77.7	85.6
MPANet（unreal）（Wu 等,2021）	68.3	71.3	28.8	27.6	82.2	84.7
MPANet（unreal + 提出的预训练）	69.3	72.4	32.9	31.9	83.8	85.9
本文（三通道模型）	69.0	71.3	33.5	**42.5**	84.8	86.6
本文（单通道模型）	67.8	71.0	**34.5**	38.0	84.2	**90.3**
本文（双模型融合）	**70.3**	**73.0**	34.3	39.7	**87.0**	89.9

注：“–”表示对应论文中没有该结果，加粗字体表示各列最优结果。

4)超参数影响分析。在本文方法中,式(7)里互学习的权重 w_{mu} 是控制互学习影响的重要参数。在图7展示了 w_{mu} 从0～0.3变化时三通道模型在SYSU-MM01上全搜索设置下的性能变化。当权重取值在0.1附近时,模型性能达到最高。当权重从0增加到0.1时,由于互学习的作用使得三通道和单通道模型学习到的知识互相迁移,挖掘的自监督信息可以提高跨模态匹配的效果。当权重从0.1向上增加时,由于互学习损失影响的增强使两个模型的输出趋向相同,难以从中挖掘两个模型输出的关系,跨模态匹配性能缓慢下降。

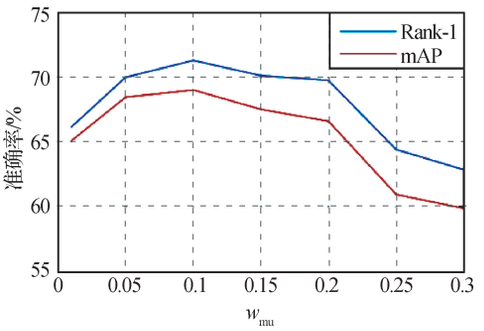


图7　互学习损失权重 w_{mu} 的影响

Fig. 7　The effect of mutual learning loss weight w_{mu}

5)不同预训练数据集的影响。为说明本文方法对不同单模态可见光预训练数据集的适用性,除了实验默认使用的UnrealPerson(Zhang等,2021a)、在虚拟行人数据集RandPerson(Wang等,2020)和真实行人数据集MSMT17(Wei等,2018)上也应用了本文预训练方法。实验结果展示了在SYSU-MM01上三通道模型的性能,如表6所示。在不同的预训练数据集上,和基础模型对比,本文方法都能在mAP和Rank-1正确率有大于6%～8%的提升,说明提出的单模态跨光谱自监督信息挖掘方法的有效性。和ImageNet预训练的结果对比,在行人数据集上的预训练效果有显著提升。

基于ResNet-50默认架构的实验访问链接https://github.com/wuancong/cjig_supplementary/blob/main/附录.pdf。

4　结　论

本文研究“可见光—红外”跨模态行人重识别,适用于跨正常光照与低照度场景进行行人匹配的情况。造成跨模态行人重识别性能不理想的难点主要是图像视觉差异导致的模态鸿沟以及标注数据缺乏。为解决这些问题,本文研究如何利用易于获得的有标注可见光图像作为辅助,挖掘单模态自监督信息来提供跨模态匹配的先验知识。主要创新点有两方面:1)提出一种随机单通道掩膜的数据增强方法,促使模型学习对光谱范围不敏感的特征;2)提出一种三通道与单通道双模型互学习的方法,从三通道数据与单通道数据的关系中挖掘跨光谱自监督信息,使这种先验知识在预训练和微调过程中在双模型之间互相迁移和补充,提高跨模态匹配模型的判别能力。在“可见光—红外”多模态行人数据集SYSU-MM01、RGBNT201和RegDB上进行的跨模态行人重识别对比实验表明,本文方法能有效地利用单模态可见光图像辅助数据挖掘对光谱范围变化不敏感的自监督信息以帮助跨模态匹配,达到当前最优的性能。

提出的互学习方法需要在训练阶段使用双模型同时进行训练,虽然测试过程可以只使用单模型,但是训练过程中的开销比一般情况下的单模型训练大一倍。进一步的工作可以考虑在互学习的框架中研究共享参数、模型压缩和知识蒸馏等新方法实现计算开销的减少。

表6　不同预训练数据集在SYSU-MM01的性能对比

Table 6　Performance comparisons on SYSU-MM01 dataset when using different pretraining datasets

/%

预训练数据集	方法	全搜索		室内搜索	
		mAP	Rank-1	mAP	Rank-1
ImageNet	基础模型	53.2	53.4	65.5	57.1
unreal Person	基础模型	61.8	62.9	72.6	65.7
	本文	69.0	71.3	80.1	76.2
RandPerson	基础模型	60.5	62.2	71.2	64.1
	本文	66.9	69.4	77.9	72.6
MSMT17	基础模型	58.1	59.0	70.2	63.2
	本文	65.5	67.8	77.0	71.9

参考文献(References)

Ahmed E, Jones M and Marks T K. 2015. An improved deep learning

architecture for person re-identification//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 3908-3916 [DOI: 10.1109/CVPR.2015.7299016]

Chen D, Li Y Z, Yu P Z and Shao C B. 2020. Research and prospect of cross modality person re-identification. Computer Systems and Applications, 29(10): 20-28 (陈丹, 李永忠, 于沛泽, 邵长斌. 2020. 跨模态行人重识别研究与展望. 计算机系统应用, 29(10): 20-28) [DOI: 10.15888/j.cnki.csa.007621]

Chen Y, Wan L, Li Z H, Jing Q Y and Sun Z Y. 2021. Neural feature search for RGB-infrared person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 587-597 [DOI: 10.1109/CVPR46437.2021.00065]

Dai P Y, Ji R R, Wang H B, Wu Q and Huang Y Y. 2018. Cross-modality person re-identification with generative adversarial training//Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI). Stockholm, Sweden: AAAI: 677-683 [DOI: 10.24963/IJCAI.2018/94]

Fan X, Luo H, Zhang C and Jiang W. 2020. Cross-spectrum dual-subspace pairing for RGB-infrared cross-modality person re-identification [EB/OL]. [2020-02-29]. https://arxiv.org/pdf/2003.00213.pdf

Fu C Y, Hu Y B, Wu X, Shi H L, Mei T and He R. 2021. CM-NAS: cross-modality neural architecture search for visible-infrared person re-identification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 11803-11812 [DOI: 10.1109/ICCV48922.2021.01161]

Gao Y J, Liang T F, Jin Y, Gu X Y, Liu W, Li Y D and Lang C Y. 2021. MSO: multi-feature space joint optimization network for RGB-infrared person re-identification//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 5257-5265 [DOI: 10.1145/3474085.3475643]

Ge Y X, Chen D P and Li H S. 2020. Mutual mean-teaching: pseudo label refinery for unsupervised domain adaptation on person re-identification//Proceedings of the 8th International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia: OpenReview.net

Hao X, Zhao S Y, Ye M and Shen J B. 2021. Cross-modality person reidentification via modality confusion and center aggregation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 16383-16392 [DOI: 10.1109/ICCV48922.2021.01609]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He L X, Liao X Y, Liu W, Liu X C, Cheng P and Mei T. 2020. FastReID: a pytorch toolbox for general instance re-identification [EB/OL]. [2020-07-15]. https://arxiv.org/pdf/2006.02631.pdf

Hermans A, Beyer L and Leibe B. 2017. In defense of the triplet loss for person re-identification [EB/OL]. [2017-11-21]. https://arxiv.org/pdf/1703.07737v2.pdf

Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network [EB/OL]. [2015-03-09]. https://arxiv.org/pdf/1503.02531.pdf

Kingma D P and Ba J. 2015. Adam: a method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego, USA: [s.n.]

Li D G, Wei X, Hong X P and Gong Y H. 2020. Infrared-visible cross-modal person re-identification with an X modality//Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI). New York, USA: AAAI: 4610-4617 [DOI: 10.1609/AAAI.V34I04.5891]

Liang W Q, Wang G C, Lai J H and Xie X H. 2021. Homogeneous-to-heterogeneous: unsupervised learning for RGB-infrared person reidentification. IEEE Transactions on Image Processing, 30: 6392-6407 [DOI: 10.1109/TIP.2021.3092578]

Liao S C, Hu Y, Zhu X Y and Li S Z. 2015. Person re-identification by local maximal occurrence representation and metric learning//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 2197-2206 [DOI: 10.1109/CVPR.2015.7298832]

Ling Y G, Zhong Z, Luo Z M, Rota P, Li S Z and Sebe N. 2020. Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 889-897 [DOI: 10.1145/3394171.3413821]

Lu Y, Wu Y, Liu B, Zhang T Z, Li B P, Chu Q and Yu N H. 2020. Cross-modality person re-identification with shared-specific feature transfer//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 13376-13386 [DOI: 10.1109/CVPR42600.2020.01339]

Luo H, Jiang W, Fan X and Zhang S P. 2019. A survey on deep learning based person re-identification. Acta Automatica Sinica, 45(11): 2032-2049 (罗浩, 姜伟, 范星, 张思朋. 2019. 基于深度学习的行人重识别研究进展. 自动化学报, 45(11): 2032-2049) [DOI: 10.16383/j.aas.c180154]

Meng J K, Zheng W S, Lai J H and Wang L. 2021. Deep graph metric learning for weakly supervised person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence: #3084613 [DOI: 10.1109/TPAMI.2021.3084613]

Nguyen D T, Hong H G, Kim K W and Park K R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors, 17(3): #605 [DOI: 10.3390/S17030605]

Park H, Lee S, Lee J and Ham B. 2021. Learning by aligning: visible-infrared person re-identification using cross-modal correspondenc-

es//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 12026-12035 [DOI: 10.1109/ICCV48922.2021.01183]

Ristani E, Solera F, Zou R, Cucchiara R and Tomasi C. 2016. Performance measures and a data set for multi-target, multi-camera tracking//Proceedings of European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 17-35 [DOI: 10.1007/978-3-319-48881-3_2]

Shen Q, Tian C, Wang J B, Jiao S S and Du L. 2020. Multi-resolution feature attention fusion method for person reidentification. Journal of Image and Graphics, 25(5): 946-955 (沈庆, 田畅, 王家宝, 焦珊珊, 杜麟. 2020. 多分辨率特征注意力融合行人再识别. 中国图象图形学报, 25(5): 946-955) [DOI: 10.11834/jig.190237]

Shi W D, Zhang Y Z, Liu S W, Zhu S D and Bao J N. 2020. Person re-identification based on deformation and occlusion mechanisms. Journal of Image and Graphics, 25(12): 2530-2540 (史维东, 张云洲, 刘双伟, 朱尚栋, 暴吉宁. 2020. 针对形变与遮挡问题的行人再识别. 中国图象图形学报, 25(12): 2530-2540) [DOI: 10.11834/jig.200016]

Sun Y F, Cheng C M, Zhang Y H, Zhang C, Zheng L, Wang Z D and Wei Y C. 2020. Circle loss: a unified perspective of pair similarity optimization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6397-6406 [DOI: 10.1109/CVPR42600.2020.00643]

Sun Y F, Zheng L, Yang Y, Tian Q and Wang S J. 2018. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 501-518 [DOI: 10.1007/978-3-030-01225-0_30]

Tian X D, Zhang Z Z, Lin S H, Qu Y Y, Xie Y and Ma L Z. 2021. Farewell to mutual information: variational distillation for cross-modal person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 1522-1531 [DOI: 10.1109/CVPR46437.2021.00157]

Wang G A, Zhang T Z, Cheng J, Liu S, Yang Y and Hou Z G. 2019a. RGB-infrared cross-modality person re-identification via joint pixel and feature alignment//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea(South): IEEE: 3622-3631 [DOI: 10.1109/ICCV.2019.00372]

Wang G S, Yuan Y F, Chen X, Li J W and Zhou X. 2018. Learning discriminative features with multiple granularities for person re-identification//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Korea (South): ACM: 274-282 [DOI: 10.1145/3240508.3240552]

Wang Y N, Liao S C and Shao L. 2020. Surpassing real-world source training data: random 3D characters for generalizable person re-identification//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 3422-3430 [DOI: 10.1145/3394171.3413815]

Wang Z X, Wang Z, Zheng Y Q, Chuang Y Y and Satoh S. 2019b. Learning to reduce dual-level discrepancy for infrared-visible person re-identification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 618-626 [DOI: 10.1109/CVPR.2019.00071]

Wei L H, Zhang S L, Gao W and Tian Q. 2018. Person transfer GAN to bridge domain gap for person re-identification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 79-88 [DOI: 10.1109/CVPR.2018.00016]

Wei X, Li D G, Hong X P, Ke W and Gong Y H. 2020. Co-attentive lifting for infrared-visible person re-identification//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1028-1037 [DOI: 10.1145/3394171.3413933]

Wei Z Y, Yang X, Wang N N and Gao X B. 2021. Syncretic modality collaborative learning for visible infrared person re-identification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 225-234 [DOI: 10.1109/ICCV48922.2021.00029]

Wu A C, Zheng W S, Gong S G and Lai J H. 2020. RGB-IR person reidentification by cross-modality similarity preservation. International Journal of Computer Vision, 128(6): 1765-1785 [DOI: 10.1007/S11263-019-01290-1]

Wu A C, Zheng W S, Yu H X, Gong S G and Lai J H. 2017. RGB-infrared cross-modality person re-identification//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5390-5399 [DOI: 10.1109/ICCV.2017.575]

Wu Q, Dai P Y, Chen J, Lin C W, Wu Y J, Huang F Y, Zhong B N and Ji R R. 2021. Discover cross-modality nuances for visible-infrared person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 4328-4337 [DOI: 10.1109/CVPR46437.2021.00431]

Yang Q Z, Wu A C and Zheng W S. 2021. Person re-identification by contour sketch under moderate clothing change. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6): 2029-2046 [DOI: 10.1109/TPAMI.2019.2960509]

Ye M, Lan X Y, Li J W and Yuen P. 2018. Hierarchical discriminative learning for visible thermal person re-identification//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 7501-7508 [DOI: 10.1609/aaai.v32i1.12293]

Ye M, Ruan W J, Du B and Shou M Z. 2021. Channel augmented joint learning for visible-infrared recognition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE: 13547-13556 [DOI: 10.1109/ICCV48922.2021.01331]

Ye M, Shen J B, Crandall D J, Shao L and Luo J B. 2020. Dynamic

dual-attentive aggregation learning for visible-infrared person re-identification//Proceedings of the 16th European Conference on Computer Vision (ECCV). Glasgow, UK: Springer: 229-247 [DOI: 10.1007/978-3-030-58520-4_14]

Ye M, Shen J B, Lin G J, Xiang T, Shao L and Hoi S C H. 2022. Deep learning for person re-identification: a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6): 2872-2893 [DOI: 10.1109/TPAMI.2021.3054775]

Yin J H, Wu A C and Zheng W S. 2020. Fine-grained person re-identification. International Journal of Computer Vision, 128(6): 1654-1672 [DOI: 10.1007/s11263-019-01259-0]

Yu H X, Wu A C and Zheng W S. 2020. Unsupervised person re-identification by deep asymmetric metric embedding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(4): 956-973 [DOI: 10.1109/TPAMI.2018.2886878]

Zhang T Y, Xie L X, Wei L H, Zhuang Z J, Zhang Y F, Li B and Tian Q. 2021a. UnrealPerson: an adaptive pipeline towards costless person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 11501-11510 [DOI: 10.1109/CVPR46437.2021.01134]

Zhang Y, Xiang T, Hospedales T M and Lu H C. 2018. Deep mutual learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 4320-4328 [DOI: 10.1109/CVPR.2018.00454]

Zhang Y K, Yan Y, Lu Y and Wang H Z. 2021b. Towards a unified middle modality learning for visible-infrared person re-identification//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: ACM: 788-796 [DOI: 10.1145/3474085.3475250]

Zheng A H, Wang Z, Chen Z H, Li C L and Tang J. 2021a. Robust multi-modality person re-identification//Proceedings of the 35th AAAI Conference on Artificial Intelligence. [s.l.]: AAAI: 3529-3537

Zheng K C, Liu W, He L X, Mei T, Luo J B and Zha Z J. 2021b. Group-aware label transfer for domain adaptive person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 5306-5315 [DOI: 10.1109/CVPR46437.2021.00527]

Zheng L, Shen L Y, Tian L, Wang S J, Wang J D and Tian Q. 2015. Scalable person re-identification: a benchmark//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1116-1124 [DOI: 10.1109/ICCV.2015.133]

Zheng W S, Hong J C, Jiao J N, Wu A C, Zhu X T, Gong S G, Qin J Y and Lai J H. 2022. Joint bilateral-resolution identity modeling for cross-resolution person re-identification. International Journal of Computer Vision, 130(1): 136-156 [DOI: 10.1007/s11263-021-01518-z]

Zheng W S, Gong S G and Xiang T. 2013. Reidentification by relative distance comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(3): 653-668 [DOI: 10.1109/TPAMI.2012.138]

Zheng W S, Gong S G and Xiang T. 2016. Towards open-world person re-identification by one-shot group-based verification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(3): 591-606 [DOI: 10.1109/TPAMI.2015.2453984]

Zhu Y X, Yang Z, Wang L, Zhao S, Hu X and Tao D P. 2020. Hetero-center loss for cross-modality person re-identification. Neurocomputing, 386: 97-109 [DOI: 0.1016/J.NEUCOM.2019.12.100]

作者简介

吴岸聪，男，博士后，主要研究方向为视频图像理解、行人识别。E-mail: wuanc@mail.sysu.edu.cn

郑伟诗，通信作者，男，教授，主要研究方向为视频图像理解与处理、行为理解、行人识别。

E-mail: zhwshi@mail.sysu.edu.cn

林城栋，男，硕士研究生，主要研究方向为行人检测与识别。

E-mail: linchzh3@mail2.sysu.edu.cn