E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP39　　**文献标识码**:A　　**文章编号**:1006-8961(2025)06-1638-23

论文引用格式:Feng Z X, Lai J H, Yuan Z, Huang Y L and Lai P J. 2025. Advancing universal person reidentification: a survey on the applications of large-scale, self-supervised pretraining models for identifying individuals. Journal of Image and Graphics, 30(6):1638-1660(冯展祥, 赖剑煌, 袁藏, 黄宇立, 赖培杰. 2025. 走向通用行人重识别:预训练大模型技术在行人重识别的应用综述. 中国图象图形学报, 30(6):1638-1660)[DOI:10.11834/jig.240426]

走向通用行人重识别:预训练大模型技术在行人重识别的应用综述

冯展祥[1],赖剑煌[2,3,4*],袁藏[2],黄宇立[2],赖培杰[1]

1. 中山大学系统科学与工程学院,广州　510006; 2. 中山大学计算机学院,广州　510006; 3. 琶洲实验室(黄埔),广州　510000; 4. 广东省信息安全技术重点实验室,广州　510006

摘　要:行人重识别旨在对没有视野重叠覆盖的视域拍摄的行人目标进行身份匹配,是计算机视觉的研究热点,在安防监控场景有重要的研究意义和广阔的应用前景。受限于标注成本过高,行人数据集规模较小,当前行人重识别模型性能还达不到应用的水平,通用行人重识别技术还任重道远。近年来,预训练大模型引发了广泛的关注,获得了快速的发展,其核心技术在行人重识别领域获得了越来越多的应用。本文对预训练大模型技术在行人重识别的应用进行了全面的梳理回顾。首先介绍本领域的研究背景,从行人重识别的研究现状和面对的困难出发,简要阐述了预训练技术和预训练大模型的相关技术,分析预训练大模型技术在行人重识别的研究意义和应用前景。在此基础上,对基于预训练大模型的行人重识别研究进行了详细的介绍,将已有研究分为大规模自监督预训练行人重识别、预训练大模型引导的行人重识别和基于提示学习的行人重识别3类,并在多个数据集对前沿算法的效果和性能进行对比。最后,对该任务进行了总结,分析当前研究的局限,并展望未来研究的方向。整体而言,预训练大模型技术是实现通用行人重识别不可或缺的技术,当前研究还处于探索阶段,行人重识别与预训练大模型技术的结合还不够紧密,如何结合行人先验和预训练大模型技术实现通用行人重识别需要学术界和工业界共同思考和推动。

关键词:行人重识别;深度学习;自监督预训练;大模型;提示学习

Advancing universal person reidentification: a survey on the applications of large-scale, pretraining models for identifying individuals

Feng Zhanxiang[1], Lai Jianhuang[2,3,4*], Yuan Zang[2], Huang Yuli[2], Lai Peijie[1]

1. School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou　510006, China; 2. School of Computer Science, Sun Yat-sen University, Guangzhou　510006, China; 3. Pazhou Laboratory (HuangPu), Guangzhou　510000, China; 4. Guangdong Province Key Laboratory of Information Security Technology, Guangzhou　510006, China

Abstract: Person reidentification (re-id) aims to recognize target pedestrians across nonoverlapping camera views. It is a key area of focus in computer vision due to its significant research value and widespread application prospect in security sur-

收稿日期:2024-07-25;**修回日期**:2024-08-16;**预印本日期**:2024-08-23

*通信作者:赖剑煌　stsljh@mail.sysu.edu.cn

基金项目:国家自然科学基金项目(U22A2095,62076258);广东省自然科学基金(2022A1515010269)

Supported by: National Natural Science Foundation of China (U22A2095, 62076258); Natural Science Foundation of Guangdong Province, China (2022A1515010269)

veillance. In recent years, the performance of re-id techniques has seen rapid growth, with state-of-the-art (SOTA) methods outperforming human performance. Furthermore, researchers have paid increasing attention to re-id in challenging uncontrolled environments, including visible-infrared, occluded, cloth-changing, low-resolution, and aerial person re-id. Despite these advancements, the performance of re-id models remains below the desired level for practical applications for two major reasons. First, existing re-id models are trained by closed datasets with single scenarios and sufficient labeled pedestrians. This approach falls short in real-world settings characterized by diverse scenarios, varying conditions across cameras, and the high cost of obtaining labeled data, leading to inadequate performance, robustness, and generalization for actual use. Second, the expensive nature of annotation limits the scale of re-id datasets, making them significantly smaller compared to datasets for other vision tasks, such as face recognition, object recognition, and segmentation. This limitation may cause re-id models to overfit to their training images, undermining their generalizability. Consequently, reaching universal person re-id remains a significant challenge. Recently, the field of large-scale pretraining models has attracted significant attention and rapid development due to t heir critical role in enhancing person re-id techniques. In this paper, we make an overview survey on the applications of large-scale pretraining techniques for person re-id. First, we introduce the background of large-scale pretraining models. Self-supervised pretraining techniques have gained great success in natural language processing (NLP). Particularly, the Transformer structure has excelled in extracting robust NLP features, with GPT and BERT emerging as pioneering models using the Transformer to generate useful outputs for subsequent tasks. GPT3 has demonstrated that large-scale pretraining models can rival the performance of SOTA supervised models without annotations. With the successful application of GPT3, many researchers have attempted to apply self-supervised pretraining techniques to vision tasks, and some pioneering research has been conducted for vision-language cross-modal tasks. ViLBERT marked the beginning of learning the relationships between vision and language. The CLIP model shows great generalization ability for zero-shot vision tasks. Furthermore, the MAE adopts mask modeling techniques to train a pretraining model with good generalization ability. These advancements highlight that large-scale pretraining techniques leveraging vast amounts of unsupervised data, not only elevate the baseline models' performance and generalization abilities but also great promise for re-id by reducing the need for expensive labeled data gathering. Moreover, the information from large-scale pretraining models can be utilized to improve the performance of re-id models. Given that self-supervised pretraining techniques can promote re-id models, some researchers have tried pioneering efforts. Here, we introduce the existing research for large-scale pretraining re-id models, organizing the literature into three types, namely, self-supervised pretraining re-id methods, large-scale pretraining model-based re-id methods, and prompt learning-based re-id methods. We discuss the above large-scale pretraining technique-based methods and the effects and performances of SOTA methods on various benchmarks. Self-supervised pretraining re-id methods employ self-supervised pretraining techniques and large-scale unsupervised pedestrian benchmark to train a robust pretraining model, addressing the scarcity and high cost of labeled pedestrian data. Some researchers have constructed weakly supervised/unsupervised benchmarks for studying self-supervised pretraining re-id techniques. SYSU-30K is the first large-scale weakly supervised re-id dataset, which is constructed by over 30 million images and 30 000 IDs from 1 000 downloaded videos. The challenges of SYSU-30K includes low-resolution, view changes, occlusion, and changing illumination. LUPerson is the first large-scale unsupervised person benchmark, containing more than 4. 2 million unsupervised pedestrian images from 46 000 scenes and covering the challenges of illumination variations, changing resolution, and occlusion. We adopt tracking for the LUPerson dataset and construct the weakly supervised dataset LUPerson-NL, which contains more than 10 million pedestrians and 430 000 noisy identities. With the emergence of large-scale unsupervised datasets, some researchers have applied self-supervised techniques for re-id. Some studies have utilized a contrastive learning framework to learn robust re-id models from unsupervised pedestrians. The MoCo framework and catastrophic forgetting score are utilized to improve the generalization ability of re-id models. Furthermore, some studies have employed the prior knowledge of pedestrians to improve the performance of self-supervised pretraining techniques. The local structure, view information, and color information are employed to incorporate prior knowledge for pretraining re-id methods. Large-scale pretraining model-based re-id methods employ the knowledge of multimodal large-scale model and use the interaction between vision and language to improve the performance of re-id models. Given that the CLIP model has shown superior performance for zero-shot vision tasks, most of

the related studies have utilized it to learn a discriminant and robust re-id model. Llama2 is also adopted to promote re-id tasks. Prompt learning-based re-id methods introduce quick learning methods to learn a robust re-id model. First, prompt learning re-id methods utilize the relationships between text description and visual features to learn a more discriminative and robust model. We focus on employing the prompts to make the model adaptive to different environments, such that we can obtain a universal re-id model that can cope with changing environments. Experimental results show that self-supervised techniques, large-scale pretraining models, and prompt learning methods can significantly improve the performance and generalization ability of re-id models. We can achieve a more universal re-id model for unseen scenarios. Finally, we conclude the overview of the current literature, analyze the limitation of the existing literature, and discuss the potential directions for future research. In conclusion, the large-scale pretraining techniques are essential for universal re-id. Although existing research is pioneering yet nascent, with a somewhat weak connection between re-id and large-scale pretraining models, the integration of pedestrian priors and large-scale model knowledge to achieve universal re-id warrants concerted exploration and promotions from the academia and industry.

Key words: person re-identification; deep learning; self-supervised pre-train; large-scale model; prompt learning

0 引 言

行人重识别是对不同的、没有视野重叠覆盖的摄像机视域拍摄的行人目标进行身份匹配的技术，在城市安防、智慧城市、智能交通管理以及视频大数据处理等邻域具有广阔的应用前景。行人重识别是人工智能(artificial intelligence, AI)领域计算机视觉的研究热点之一(Ye等, 2022a)，过去十几年吸引了广泛的研究关注，相关技术得到长足发展，识别性能迅速增长，在公开数据库超过了人眼的识别准确率。尽管相关研究非常火热，当前行人重识别技术仍然面临着巨大的瓶颈(罗浩 等, 2019)，其性能远未达到通用行人重识别的要求。行人重识别迟迟未能落地应用有以下两方面原因：首先，当前行人重识别研究主要集中在场景单一、标签充足的封闭数据集。而在实际应用中存在场景复杂、跨域成像环境差异大、无法大量标注等问题，导致现有的方法缺乏鲁棒性、迁移性与泛化性，无法满足真实监控场景的需求。其次，行人重识别任务的标注数据规模要远远小于其他主流视觉任务。如表1所示，人脸识别开源数据集Glint360K(An等, 2022)类别数量达到36万，训练标签样本数达到1 700万，WebFace 260M(Zhu等, 2021)的训练数据更是多达400万类2.6亿幅，图像分类数据集ImageNet(Deng等, 2009)的标签分类数据达到1 400万，目标检测数据集Objects365(Shao等, 2019)标签样本超过1 000万，目标分割数据集SA-1B(Kirillov等, 2023)规模更是达到10亿。相比之下，行人重识别的主流数据集类别数不超过1万，主流的CUHK03(the Chinese University of Hong Kong 03)(Li等, 2014)和Market1501(Zheng等, 2015)都只有1 500类左右行人，标签样本规模在10万以下。当前最大规模的标注行人数据集之一MSMT17(multi-scene multi-time 17)(Wei等, 2018)的数据规模远不如2014年发布的Casia-WebFace(Yi等, 2014)的数据规模。标签行人类别数最多的数据集是Airport(Karanam等, 2019)数据集，该数据集的标签行人数量达到9 651，但是样本数只有39 902，每个行人的平均样本数只有4幅。标注数据不足是限制行人重识别模型应用的主要原因。对行人进行标注需要跨场景找到关联的行人，成本太高，因此学术界和业界缺少大规模标注行人数据库。受限于标注数据有限，已有技术泛化能力不足，距离应用还有较大的差距。行人重识别技术要落地应用，形成通用模型，关键在于如何通过有限的标签数据提高特征表达能力和泛化能力。

尽管无法获得大规模标注行人数据，但是可以从监控视频获取大规模无标签行人数据，因此可以利用大规模自监督预训练技术训练通用行人重识别模型，提升行人重识别模型的泛化性能。大规模自监督预训练技术已经在自然语言处理(natural language processing, NLP)领域取得了巨大的成功，基于自监督技术的预训练大模型是当前人工智能最引人注目的技术，被认为是推动第3代人工智能技术发展的一项重要技术(林俊安 等, 2024)。自监督预训练大模型技术出现在2018年前后，GPT(generative pre-trained Transformer)(Radford 等, 2018)和BERT (bidirectional encoder representations from

表 1　计算机视觉任务标注数据集对比

Table 1　Comparisons between annotated datasets for computer vision tasks

数据集	研究方向	样本数	类别数	发布年份
Casia-WebFace	人脸识别	494 414	10 575	2014
Glint360K	人脸识别	1 700万	36万	2020
WebFace260M	人脸识别	2.6亿	400万	2021
ImageNet	目标分类	1 400万	22 000	2009
Objects365	目标检测	1 000万	365	2019
SA-1B	目标分割	10亿	无	2023
CUHK03	行人重识别	14 096	1 467	2014
Market1501	行人重识别	32 668	1 501	2015
Airport	行人重识别	39 902	9 651	2019
MSMT17	行人重识别	126 441	4 101	2018

Transformers)(Devlin等,2019)模型成功激活了深度网络对大规模无标注数据的自监督学习能力,在GPU(general processing unit)多机多卡算力和海量无标注文本数据的双重支持下,预训练模型成为人工智能和深度学习领域的革命性突破,将模型规模和性能不断推向新的高度。在过去几年预训练大模型取得显著进步,特别是OPENAI公司在2022年11月推出的ChatGPT在短短几个月内积累了1亿用户,标志着大模型技术有了大规模商用的基础。与经典的机器学习流程相比,大规模预训练模型具有以下3方面优势:1)由于复杂的预训练目标和巨大的模型参数,大规模预训练模型可以有效地从大量标记和未标记的数据中获取知识。通过将知识存储到巨大的参数中并对特定任务进行微调,巨大参数中隐式编码的丰富知识可以使各种下游任务受益。2)大规模预训练模型是应对AI领域应用碎片化的有效方式,能够极大地降低下游任务训练、成本和门槛。大规模自监督预训练提供了更好的模型初始化,相当于一种正则化,能避免训练时候过拟合,带来更好的泛化性能,并加速对目标任务的收敛。模型只需要对少量特定任务的有标注数据进行微调即可完成下游任务学习,有标注数据的利用率高。3)大模型和海量无监督训练数据可以学习泛用性极强的模型,从而获得一个统一的通用模型,不需要针对每个任务、场景和数据集专门研制模型。

尽管预训练大模型技术在自然语言处理领域已经取得了很大的成功,但是,视觉预训练大模型的研究进展明显滞后于自然语言处理,面向生物特征提取和识别的大模型研究进展很慢,尤其是行人重识别大模型研究仍然在摸索阶段,还有许多科学问题需要解决,如何合理有效地将预训练大模型相关技术应用到行人重识别任务仍然是一个很有研究意义的开放性课题。本文将从行人重识别技术和预训练大模型的研究现状出发,结合大规模自监督预训练技术、提示学习等大模型相关技术,阐述预训练大模型技术在行人重识别任务的研究进展,分析其面临的困难和未来的发展趋势。

1　行人重识别技术的发展现状

自2005年确定行人重识别的概念以来,行人重识别吸引了大量的研究注意,并逐渐成为计算机视觉的研究热点,在过去10多年得到飞速发展,出现了许多基于机器学习和模式识别理论的行人重识别方法,其性能逐年提升。当前行人重识别方法主要分为基于手工描述子的行人重识别方法(Liu等,2014)、基于度量学习的行人重识别方法(Chen等,2016)和基于深度学习的行人重识别方法(Feng等,2020)。早期的行人重识别研究(Zheng等,2013)以手工描述子方法为主,以度量学习为辅,设计手工描述子使得特征足够鲁棒以应对拍摄视角、光照和行人姿态的变化,并通过度量学习将特征投影到判别的距离空间,从而使得行人特征具有分辨能力。随着深度网络的理论体系和结构的发展与成熟,基于深度学习的行人重识别方法已经成为了行人重识别主流方法(Zahra等,2023),并且深度行人重识别网络取得了很大的突破,多尺度深度网络结构(Qian等,2017)、深度度量学习(Li等,2022)和Transformer网络结构(Ni等,2023)等方法推动了行人重识别理论的发展和进步,使行人重识别模型的性能得到了显著的提升,在主流公开数据集,如CUHK03和Market-1501的测试结果已经超过人类(Zhang等,2017)。随着主流数据集行人重识别算法研究的深入,非可控应用环境跨模态与光照变化、遮挡、低分辨率以及航拍视角等挑战愈发突出,严重影响了行人重识别算法的应用(冯展祥 等,2020),越来越多的研究人员试图解决实际应用环境面临的挑战,实现能应对各种场景和任务的通用行人重识别模型。

1.1 可见光—红外行人重识别

红外摄像头能为克服行人光照变化问题提供可靠的支持，对全天候的监控系统至关重要。跨模态行人重识别任务带来表观特征剧变、光照变化等问题，是一个极具挑战性的研究问题（吴岸聪 等，2022）。Wu等人（2017）构建了第1个大规模可见光—红外行人重识别数据库SYSU-MM01（Sun Yat-sen University-multiple modality 01），并提出基于多模态输入结构的跨模态行人重识别算法，将不同模态的输入图像嵌入到模态特定的结构实现模态融合，学习模态鲁棒特征。Feng等人（2020）提出基于模态相关特征学习的可见光—红外行人重识别算法，通过构建模态相关网络、提取模态关联的低级特征，在深层共享网络参数、学习模态共享信息，并通过度量损失提高特征的判别能力，显著提高了跨模态行人重识别算法的识别准确率。Yang等人（2022b）提出一种基于双重噪声标签的双重鲁棒训练方法，采用一种双重鲁棒损失，包括软识别损失和自适应四元组损失，以实现对噪声注释和噪声对应的鲁棒性。Zhang等人（2022）提出特征级模态特定信息补偿框架，从已有模态的图像中生成缺失模态的图像，然后从配对图像中提取判别行人特征进行匹配。Fang等人（2023）提出语义对齐和关联推断框架，利用像素级特征与可学习的原型之间的相似性来聚合潜在的语义部分特征，并设计了一个关联推断模块，通过行人关系优化推断结果，提升模型性能。Yu等人（2023）提出模态统一网络结构，引入辅助模态模拟特定模态和模态共享表征减轻跨模态和模态内的变化，并引入身份对齐损失和模态对齐损失，缩小可见光和红外图像的分布距离，学习辨别表征。Ren和Zhang（2024）提出一种隐式判别知识学习网络挖掘和利用隐式模态判别信息，使用双流网络结构提取特定模态和模态共享特征，然后对特定模态特征提出对齐损失减少模态风格差异，同时保留身份识别的判别知识。多个公共数据集的广泛实验证明其方法的优越性。

1.2 遮挡行人重识别

在监控场景，尤其是拥挤的公共场所，行人可能被建筑物、行李以及其他行人遮挡，导致局部行人图像和关键区域信息丢失，很难从遮挡图像提取鲁棒行人特征，给识别任务带来不利的影响。Zheng等人（2013）构造了第一个遮挡行人数据库Partial REID，包含60个行人身份600幅局部遮挡的行人图像，并提出一种基于局部稀疏表示匹配和全局空间对齐匹配的模型，采用局部分块联合高斯计算相似性得分，提升了遮挡行人重识别的性能。Zhuo等人（2018）构建了Occluded-REID数据集并提出一种基于联合显著性学习的遮挡行人重识别方法，通过遮挡模拟器生成多种类型遮挡，同时添加遮挡与非遮挡分类损失实现遮挡行人与完整行人之间的显著性关注机制，降低了遮挡区域的影响。遮挡行人数据库的出现推动了相关研究的发展，Hou等人（2019）提出一个特征补完的框架，该框架包含空间模块和时间模块，利用视频帧之间的时空关联恢复特征空间中遮挡区域的语义，显著提升了遮挡行人特征的鲁棒性。Wang等人（2020）学习判别特征和人体拓扑信息的高阶关系解决行人遮挡问题，将局部特征表示为图的节点，利用图匹配策略学习节点之间的对应关系，然后将对应关系视为邻接矩阵来传递信息，抑制噪声特征的信息。Wang等人（2022）提出遮挡行人特征擦除和扩散网络，通过非行人遮挡增强生成精确的遮挡掩膜，随后通过特征扩散模块合成目标行人特征，提高对遮挡行人的感知能力。Xu等人（2022）提出特征恢复转换器结构，挖掘两幅图像的可见区域计算相似度，并设计恢复转换器来恢复完整的行人特征，解决噪声干扰和遮挡带来的行人信息丢失等问题。Huang等人（2023a）提出基于注意力图神经网络的响应和挖掘方法，利用上下文语义区分遮挡区域和可视区域，并以可视得分引导网络忽视遮挡区域，学习全局判别特征，实现可视区域响应和遮挡特征补全。Wang等人（2024）提出FCFormer（full collection former）结构，提出遮挡实例增广方法模拟真实多样的遮挡情况，然后通过共享编码器从输入对中学习配对的鉴别特征，并通过特征补全解码器从自动生成的遮挡特征中汇总可能的信息，补全特征空间遮挡区域的语义特征。FCFormer在5个数据集上进行了大量实验，证明其在遮挡数据集的卓越性能和显著优势。

1.3 低分辨率行人重识别

由于监控摄像头被布置在不同的区域，因此视频行人的成像质量和成像条件差异很大，导致不同的行人图像分辨率差异很大，低分辨率图像很难提取判别特征，识别精确明显降低（杨露露 等，2023）。当前，针对低分辨率行人重识别研究主要有行人图

像归一化预处理以及跨分辨率鲁棒行人特征提取两种。Wang等人(2018)通过使用级联的生成对抗网络结构进行行人图像超分辨率,实现从粗到细的低分辨率行人图像增强,获得细节更加逼真的高分辨率行人图像,然后用超分辨率后的图像学习行人分类器。Li等人(2018)对高低分辨率图像构建不同的字典并学习分辨率特定的投影矩阵,从而将高低分辨率图像的特征投影到一个公共的特征子空间。Cheng等人(2020)提出一种正则化方法,平衡超分辨率和行人特征提取获取对网络参数更新的权重,对深度特征进行度量学习找到对分辨率变化鲁棒的子空间,学习判别特征。Zhang等人(2021a)提出一种深度高分辨率学习框架,设计通道注意力结构,通过利用特征图的不同通道信息恢复低分辨率图像的特征,并设计孪生网络结构减少不同分辨率之间的特征分布差异。Wu等人(2023)提出一种自适应动态度量的分辨率无关框架,将来自不同分辨率的行人图像编码到特定的子空间,然后学习分辨率自适应掩码提取分辨率相关的特征,结合回归学习策略获得对分辨率鲁棒的行人特征空间,显著提升了跨分辨率行人重识别模型性能。

1.4 无人机航拍行人重识别

随着无人机技术的发展,无人机普及程度越来越高,研究人员开始将注意力放到无人机航拍的行人重识别。早期研究以数据集构建为主,Zhang等人(2021b)构建了第1个无人机航拍行人数据集PRAI-1581(person ReID in aerial imagery-1581),共有接近4万幅无人机拍摄的行人图像,拍摄高度为20～60 m之间,行人数量是1 581类,大部分图像由鸟瞰视角拍摄。Li等人(2021)构建了一个面向航拍行为识别、车辆重识别和行人重识别的大规模多源多模态航空数据集UAV-Human(unmanned aerial vehicle-human),通过可见光、红外等多种摄像机获取多源多模态的航拍行人图像及动作,并标注了行人属性和行为动作模式,包含67 428个多模态视频序列、119个动作识别的类别以及包含1 144个行人身份的41 290幅图像。Zhang等人(2023)构建了一个大规模的地面—航空行人搜索数据集G2APS(ground to aerial person search),包含2 644个行人共31 770幅图像,检测出26万个行人框。Nguyen等人(2024)构建了一个航空-地面的行人重识别数据集AG-ReID(aerial-ground person re-id),该数据集包含388个行人共21 983幅图像,每个人有15个属性标签,航拍高度从15 m到45 m不等,该工作提出一种结合特征和属性的行人重识别框架,通过知识蒸馏将属性知识传递到身份提取网络。无人机拍摄的行人图像距离更远,大部分躯干被遮挡,有效信息很少,导致传统行人重识别算法的识别准确率不高。为了解决上述挑战,当前已经出现了一些探索性的无人机航拍行人重识别研究。Chen等人(2022)提出面向无人机航拍行人的旋转不变Transformer结构,考虑地面和航空拍摄的不同视角对特征层面进行对应的旋转增强,获得更多的视角变化,并通过视角不变约束降低视角变化对特征的影响,提升了行人重识别模型对不同视角的鲁棒性。Huang等人(2024)提出一种多分辨率特征感知网络结构,通过在低分辨率图像和高分辨率图像之间建立自注意力和互注意力模块学习对不同的分辨率鲁棒的行人特征。

1.5 行人重识别技术性能和瓶颈

尽管行人重识别研究涉猎的范围很广,几乎覆盖了生活中可能遇到的所有情况,当前提出的技术仍然没有彻底解决非可控环境行人重识别的难题。随着行人重识别技术的发展,在各个领域的行人重识别模型的性能增长速度明显变低,逐渐触摸到了性能的瓶颈。总体而言,当前行人重识别模型的识别准确率还很难令人满意,主流算法在不同方向和场景的行人重识别数据集的识别性能如表2所示,包括可见光行人数据集Market1501和MSMT17、可见光—红外行人数据集SYSU-MM01、遮挡行人数据集Occluded-REID(occluded-person re-identification)、低分辨率行人数据集MLR-Market(multiple low resolution-market)、换衣行人数据集PRCC(person re-identification under clothing change)(Yang等,2021)、无人机行人数据集PRAI-1581,测试的算法都是近期发表在国际顶刊和顶会的算法,包括SOLIDER(semantic controllable self- supervised learning framework)(Chen等,2023a)、IDKL(implicit discriminative knowledge learning)(Ren和Zhang,2024)、FCFormer(Wang等,2024)、RAP(resolution-adaptive representations)(Wu等,2023)、AIM(auto-intervention model)(Yang等,2023)和RotTran(rotation invariant Transformer)(Chen等,2022)。

由表2可见,面向非可控环境闭集测试主流算

法的识别准确率基本在80%～90%，一些困难的任务识别率更低，如换衣和无人机的识别率只有57.9%和70.8%，距离实际应用的要求还有很大的距离。一个最主要的原因是收集行人标签数据的难度和成本很高，需要挖掘出现在不同视角监控摄像头的同一个行人，导致当前行人重识别的数据集规模较少，并且未来几年也很难出现大规模的标注行人数据集。因此，如何通过有限的标签数据提升行人重识别模型的泛化能力，获得识别能力更强的通用行人重识别模型，是当前行人重识别研究面临的亟待解决的瓶颈问题，自监督学习和大模型技术能为行人重识别技术破冰提供借鉴和参考。

表2　不同行人重识别任务的前沿方法性能

Table 2　The performance of advancing methods for different re-id tasks

数据集	任务	识别率/%	算法	会议/期刊
Market1501	可见光	96.7	SOLIDER	CVPR2023
MSMT17	可见光	91.7	SOLIDER	CVPR2023
SYSU-MM01	红外	81.4	IDKL	CVPR2024
Occluded-REID	遮挡	86.9	FCFormer	TMM2024
MLR-Market	低分辨率	90.1	RAP	TIP2023
PRCC	换衣	57.9	AIM	CVPR2023
PRAI-1581	无人机	70.8	RotTran	ACMMM2022

2　大规模预训练技术国内外研究现状

算法、数据和算力是人工智能的三驾马车。随着算法和算力的发展，数据的限制越发突出，高昂的标注代价限制了人工智能的应用。随着深度网络的发展，对于参数量规模庞大的神经网络，用少量标签数据训练容易产生过拟合的问题，导致模型的泛化能力较差。但是，数据标注的成本差异很大，部分任务需要专业知识进行数据标注，成本非常高，如视觉识别任务和机器翻译任务可能需要数百万标注样本的数据集，要在所有碎片化场景建立大规模标注数据集是不可能的。上述困难限制了人工智能模型的大规模应用，如何通过有限的人工标注数据构建泛化能力较强的深度模型成了落地应用的关键，预训练大模型技术就是解决上述挑战的核心。

2.1　预训练大模型技术在自然语言处理的研究

近年来，学者们关注到无标注数据的重要性，开始研究如何从大规模数据抽取信息，大规模自监督学习脱颖而出。自监督学习技术通过输入数据本身作为监督信号从无标签数据中提取领域通用知识，从而提升模型在下游应用的泛化能力，使得利用大规模无监督数据获取预训练模型成为可能，在自然语言处理任务取得了显著的进展。研究者通过在大规模无标注语料上进行自监督训练学习得到通用的语言表征，获得用于解决下游任务的泛用模型。

自监督学习和Transformer是预训练模型在NLP取得成功的关键。Transformer（Vaswani等，2017）是一种基于自注意力机制的编码器—解码器结构，能并行地建模输入序列中所有单词之间的相关性。由于其突出的性质，Transformer逐渐成为预训练大模型的标准神经结构，形成了两个里程碑式预训练模型：GPT和BERT。GPT是第1个结合Transformer和自监督预训练的大规模预训练模型，以单向Transformer解码器为骨干，采用生成式预训练和判别式微调两步训练。在预训练阶段，GPT采用无监督学习策略基于庞大的无监督语料训练一个生成式语言模型，对每个单词计算概率分布学习神经网络的预训练初始参数；在微调阶段，使用下游标注数据微调模型解决应用任务。GPT模型刷新了NLP领域的9项典型任务，效果十分惊艳。BERT是应用最广泛的预训练模型结构，采用双向深层Transformer作为主要结构。在预训练阶段，BERT设计了目标掩码语言建模，对词语进行随机掩蔽并预测掩蔽位置上的单词，从而学习双向上下文语义信息。BERT横扫了11项NLP任务，对预训练大模型产生了深远的影响。2020年，OPENAI发布GPT-3网络（Brown等，2020），是NLP大规模预训练模型的一个重要里程碑，展示了海量模型参数蕴含的潜在力量，尤其是强大的小样本学习能力。GPT-3继承了GPT的主体框架，突破了当时最大的神经网络的参数规模，参数量达到1 750亿，使用45 T数据进行训练。GPT-3显示出了极强的泛化能力，在零样本、少样本学习任务表现出了很强的泛化能力，在文本生成、自然语言推理和常识推理等任务取得了实质性进展，将NLP应用扩展到缺乏足够训练数据的领域。GPT-3在下游任务不需要精调取得了接近全微调的效果，部分任务甚至超过最好的主流监督算法，展示了大模型实现

通用人工智能任务的可能性。2022 年 11 月，OPENAI发布了ChatGPT，可以跟踪上下文对话流程生成类似人类的响应，能完成写作、编程以及问答等多项任务，两个月内全球用户量超过了1亿，已经有了大规模商用的雏形。

2.2 多模态预训练大模型技术的研究

受自监督学习技术和大模型结构在NLP任务成功的激励，部分学者尝试探索自监督学习和大模型技术在视觉任务的应用（田永林 等，2022）。最早的视觉大模型研究是关于多模态大模型的，不仅使用文本模态，还使用视觉模态等一起进行模型的预训练。ViLBERT模型（Lu 等，2019）设计了一种多模态双流模型分别预处理文本和视觉信息，并且基于共注意力层学习不同模态的联系，在下游视觉问答、视觉常识推理和指示表述任务都获得了2%～10%的精度提升。VisualBERT（Li 等，2019）扩展了 BERT 架构，Transformer层隐式对齐输入文本和图像区域中的元素，在4个下游任务上进行测试并取得了很好的泛化性能。Chen等人（2020b）提出UniT多任务多模态统一Transformer模型，能够同时解决视觉、多模态和语言等领域中的一系列任务，包括目标检测、视觉—文本推理和自然语言理解等，在7个任务上都有较强的性能。DALLE（distributed autoencoder for language and image）（Ramesh 等，2021）是第一个文本到图像的零样本预训练模型，参数规模达到百亿，通过离散自编码模型来建模图像信息特征，使用自回归Transformer建模文本特征和图像特征之间的联合分布，最大化共现概率，展示了多模态预训练模型在弥合文本描述和图像生成不同模态信息之间差距的出色能力。DALLE2（Ramesh 等，2022）首先利用先验知识从文本提取图像嵌入特征，然后通过嵌入特征利用扩散模型解码产生目标图像，展现了强大的零样本学习以及语义理解和融合能力，生成的图像逼真、细节丰富。OpenAI开发CLIP（contrastive language-image pre-training）多模态大模型（Radford 等，2021），是多模态大模型研究的里程碑工作之一。CLIP提出双塔模型结构在预训练阶段学习通用视觉语义概念，包含文本编码器和图像编码器，在互联网收集了4亿对关联的图像文本数据，并提出基于对比学习的图文预训练方法，通过判断图像和文本是否匹配进行联合训练，在下游任务取得了非常好的泛化效果。CLIP多模态预训练模型具有良好的零样本迁移性能，在20多个下游任务，包括细粒度物体分类、光学字符识别和行为识别等任务的测试性能超过了全监督主流方法的性能。多模态预训练大模型证明了可以通过视觉图象和语言文本之间的关联学习可迁移的特征，并在多个场景和任务验证了其泛化能力，形成了一批落地应用。

2.3 预训练大模型技术在计算机视觉的研究

多模态大模型需要文本提示进行辅助，并不适用于所有视觉任务，因此部分学者开展视觉预训练大模型研究。BEiT（bidirectional encoder representations from Transformers）（Bao 等，2022）是最早的视觉大规模预训练模型，将图像通过离散自编码器编解码学习隐含层特征，然后通过遮挡图象建模的方式随机遮挡40 %的图像块，预测其原始视觉特征。在图像分类和语义分割等视觉任务的实验结果表明BEiT取得了优异的泛化效果。MAE（masked autoencoder）（He 等，2022）采用图像掩码重建的方式，对输入图像块大比例随机遮挡（75%）并对遗失像素进行重建，提出非对称编解码架构，编码器基于ViT（vision Transformer）（Dosovitskiy 等，2021），解码器通过遮挡信息和隐特征进行重建，仅用于预训练阶段，采用轻量化结构，在下游分类任务带来了显著的提升。SimMIM（simple masked image modeling）（Xie 等，2022）通过遮挡图像建模来学习预训练模型的参数，对输入图像信号的一部分进行遮挡，并预测被遮挡区域的原始图像输出。对输入图像进行随机掩码（10%～70%）；编码器采用ViT提取图像块特征，解码器采用一层线性层的轻量化设计，预测目标采用损失L1函数直接回归预测原始像素RGB值。通过比以前少40倍的数据训练30亿参数的SwinV2-G模型，在ImageNet、CoCo（common objects in context）检测分割创造新记录。SAM（segment anything model）模型（Kirillov 等，2023）引入了提示学习策略，提示可以是前景/背景、粗框或掩膜以及自由文本，使用11亿分割掩码进行训练，展示了视觉通用大模型的潜力，学会了物体的一般概念，对未知的物体和不熟悉的场景也有不错的效果。SAM模型展示了强大的泛化能力，在视觉目标分割、医学图像分割和工业缺陷检测等多个场景得到了快速的落地应用。FastSAM（Zhao 等，2023）提出轻量级模型实现快速语义分割的框架，将语义分割任务分解为实例分割和提示引导两个部分，在提高模型处理速度和效率的同

时保持较高的识别精度，多个分割和检测数据集上取得了优异的性能和突出的实时性能。尽管视觉大模型的研究热火朝天，目前还存在许多问题需要解决。一方面，当前模型结构不能高效学习视觉结构和语义信息，导致视觉大模型规模远远不如语言大模型；另一方面，大多数视觉大模型技术仍然停留在预训练—微调范式，在下游任务应用仍然需要收集标注数据，距离通用人工智能的要求还非常遥远。

2.4 预训练大模型与行人重识别

回顾预训练大模型的发展历程，以下几个方面能给行人重识别的发展提供思路上的启发和技术上的支持。首先，参考语言大模型的成功经验，可以通过大规模自监督预训练技术训练预训练行人重识别模型，通过大规模的无监督行人样本提升行人重识别模型在下游的泛化能力。其次，借鉴多模态大模型的经验，可以挖掘文本信息和文本提示在行人重识别任务的应用，通过文本—视觉信息之间的关联性学习判别能力和泛化能力更强的模型。最后，考虑视觉大模型研究面临的困难和窘境，不能直接将已有的预训练大模型结构套用到行人重识别任务，而是需要结合行人先验设计模型结构，将行人重识别面临的各种挑战整合到一个统一的模型框架，从而学习一个通才的模型，获得通用行人重识别模型。

3 自监督预训练行人重识别发展现状

行人标注数据采集成本高、数据规模小是限制行人重识别模型性能和泛化能力的主要原因，随着自监督预训练大模型在自然语言处理和多模态交叉处理获得成功，部分学者开始探索自监督预训练技术在行人重识别的应用，从自然语言处理预训练的经验出发，以视觉自监督预训练框架为基础研究大规模预训练行人重识别方法。

受到自监督预训练技术在NLP领域的启发，部分研究者进行了视觉自监督预训练技术的开发，基于对比学习的框架进行自监督训练，以MoCo（He等，2020）、SimCLR（Chen等，2020a）、MAE、SimMIM等为代表，显著提升视觉模型在下游任务的泛化能力，也为自监督预训练技术在行人重识别的应用打下了基础。基于动量对比的无监督学习方法MoCo，通过对比损失从大规模无标签数据建立大型、一致的词典，使用对比损失将查询编码与字典进行匹配来训练视觉表示编码器。MoCo在分类、检测、分割和关键点检测等主流的视觉任务超越有监督预训练模型，证明了自监督学习技术在计算机视觉领域也能取得很好的效果。Hinton提出一种简单的视觉表征对比学习框架SimCLR，在对比学习框架中使用了多种数据增强的组合，在视觉特征和对比损失之间引入非线性变换提高特征质量，并提出使用更大的批数量和迭代次数进行对比学习。SimCLR在很多数据集上的实验表现了出色的性能，自监督线性分类器精度接近监督学习的ResNet-50（residual network）（He等，2016）模型。

3.1 大规模无监督行人数据集

模型预训练在行人重识别中起着重要的作用，在视觉自监督预训练技术的支持下，部分学者开始挖掘自监督预训练技术在行人重识别的应用。与其他视觉任务相比，行人重识别数据的收集和标注是非常困难的，当前公开的行人数据集在图像数量，行人数量和捕获环境等方面都非常局限。主流的标注行人数据集如表3所示，VIPeR（visual pairwise endoscopy registration）数据集（Gray等，2007）只有1 264个行人，每个人只有2幅图像，当前标注行人数据集中最大的数据集MSMT17只有12.6万图像，类别数最多的Airport有9 651人，每个人只有4幅图像。

表3 标注行人数据集和大规模行人数据集对比

Table 3 Comparisons between labeled pedestrian datasets and large-scale pedestrian datasets

数据集	标注方式	样本数	行人数量	视角数量
VIPeR	有监督	1 264	632	2
CUHK03	有监督	14 096	1 467	5
Market1501	有监督	32 668	1 501	6
Airport	有监督	39 902	9 651	6
MSMT17	有监督	126 441	4 101	15
SYSU-30K	弱监督	3 000万	3万	1 000
LUPerson	无监督	420万	20万	46 000
LUPerson-NL	无监督	1 000万	43万	46 000

已有的主流预训练方法往往使用ImageNet进行预训练，然后再进行微调。但是，ImageNet和行人数据差异是很大的，导致预训练的效果不佳，为了解决上述问题，部分研究人员开始创建大规模弱监督/无

监督行人数据集，如图 1 所示。Wang 等人（2020）创建了大规模弱监督行人数据集 SYSU-30K（Sun Yat-sen University-30K），通过网络下载 1 000 个电视节目视频，利用弱标注方式进行标注，将数据集切成 8.5 万个切片，然后标注人员记录每个切片的行人身份，获得了约 3 000 万的行人图像和超过 3 万个标注的身份，SYSU-30K 数据集包含光照变化、遮挡、低像素、俯视拍摄的摄像机和复杂背景等挑战，支持通用的、弱监督和自监督学习模式，满足不同研究和应用需求。Fu 等人（2021）构建了第 1 个大规模无监督行人数据集 LUPerson，以此解决 ImageNet 预训练模型的数据和实际行人数据差异过大的问题，推动行人重识别技术的边界，通过无监督预训练的方法，在没有人工标注的情况下也能实现高效学习和准确的识别。该数据集共包含 420 万无监督行人图像，来自 46 000 多个场景，行人数量超过 20 万，涵盖了光照变化、低分辨率和遮挡等挑战，首次将大规模无监督预训练用到行人重识别，以 MoCo 为基准进行无监督预训练，以提高学习到的特征的泛化能力，并系统地研究了数据增强和对比损失在行人重识别预训练中的作用，验证了大规模预训练对行人重识别模型的提升效果，显著提升主流模型性能，小样本、非监督迁移任务提升更加明显。Fu 等人（2022）还将多目标跟踪系统应用于 LUPerson 的原始视频，并建立带噪声标签预训练数据集 LUPerson-NL。该数据集通过多目标跟踪系统建立大规模噪声标签，为跟踪算法检测的每一个行人附上身份标签，并由此制作了一个数据规模达到千万，类别数量达到 43 万的噪声行人数据集。在此基础上，Fu 等人提出基于噪声标签的大规模预训练框架，该框架包含 3 个模块：分类监督学习模块、原型对比学习模块和标签引导的对比学习模块。通过多个数据集的实验表明，使用该数据集进行预训练的提升效果与 LUPerson 更加显著，尤其在小规模数据集优势明显。

表 3 对比了基于监督标注的行人数据集和大规模无标签行人数据集，可以看出，弱监督/无监督获得的行人数据集数据规模和行人数量要远远超过有监督的行人数据集。此外，主流的监督行人数据集

(a) SYSU-30K　　(b) LUPerson

图 1　弱监督和无监督行人数据集展示

Fig. 1　Visualization of weakly supervised and unsupervised person datasets ((a) SYSU-30K; (b) LUPerson)

只考虑单一的场景，摄像机视角数量不多，因此基于这些数据训练的模型泛化能力较差，大规模行人数据集覆盖多种场景，没有明显的数据倾向，有利于提高预训练模型泛化能力。

3.2 基于自监督预训练的行人重识别研究

数据库的出现推动了大规模预训练行人重识别研究的发展，当前自监督预训练行人重识别方法的研究可以分为两类，一类是基于传统对比学习框架的研究，一类是基于行人先验的自监督预训练框架的研究，如图2所示。早期的探索性研究往往使用面向通用视觉物体分类的自监督学习框架进行，研究者在LUPerson（Fu等，2021）和LUPerson-NL（Fu等，2022）的探索性研究就是采用了MoCo框架，以ResNet-50为基准网络。Luo等人（2021）首次探索了ViT等Transformer结构在大规模数据集的自监督预训练学习，通过引入实例正则化和快正则化来学习具有视角、姿态和光照不变性的特征。在此基础上，Luo等人（2021）还提出灾难性遗忘分数来评估预训练和微调数据之间的差距，通过对下游行人重识别数据进行采样并从预训练数据集中过滤不相关数据来选择一个更接近的相关数据子集，从而减少训练需要的时间和资源，只使用一半数据集进行训练而没有造成性能损失。Wan等人（2023）聚焦于解决多模态视频监控场景下行人重识别任务的模态偏差预训练问题，针对预训练可见光数据和实际应用的红外数据场景的差异提出一种自监督预训练模态感知多粒度学习方法MMGL（modality-aware multi-granularity learning），在多模态数据集上直接开展预训练，提出排列恢复模块学习全局模态不变表示和部分感知循环对比学习模块提升局部特征的区分能力，无须依赖外部数据集和复杂的精调技巧，训练速度更快、数据效率更高，并且表现出优秀泛化性能和跨数据集迁移性能，具有推广到更多多模态图像检索任务的潜力。Zhang等人（2024）研究从合成数据到真实数据之间的跨场景行人重识别任务，通过合成生成的方法产生大量虚拟行人图像，提出一个领域不变特征学习模块和均值教师网络训练方案，通过已有的模型提取特征，设计3个维度的自监督辅助任务：帧级别、视频级别和特征拼接级别这3个维度来预测特征分布；将基础网络训练得到的模型作为学生网络，然后通过指数移动平均方法得到教师网络，设计身份一致性损失和身份相似性损失两个自监督损失来训练学生网络。Ye等人（2022b）提出基于灾难遗忘分数的（catastrophic forgetting score，CFS）方法在LUPerson数据集得到预训练模型参数，并提出通道级自注意力模块减少特征计算开销，提出双重原型对比学习方法，聚类对比方法学习方法和难样本对比学习方法提高模型对困难样本的识别能力，在域适应任务不使用标签数据获得与主流监督算法相当的性能，通过大规模自监督学习提升无监督域适应行人重识别模型的性能。Zhu等人（2022）提出一种基于局部特征的大规模预训练行人重识别方法PASS（partially aligned spatial pyramid pooling）。此前的自监督方法直接使用图像分类的模型，没有进行框架上的调整，导致同时将局部和全局视图进行匹配，丢失大量细节信息。PASS以ViT为基准，通过生成局部特征以提供细粒度信息，将图像划分为多个局部区域，从每个区域随机裁剪得到的视图分配一个可学习的局部标记，这些标记会被添加到全局视图中。然后设计了一个知识蒸馏的框

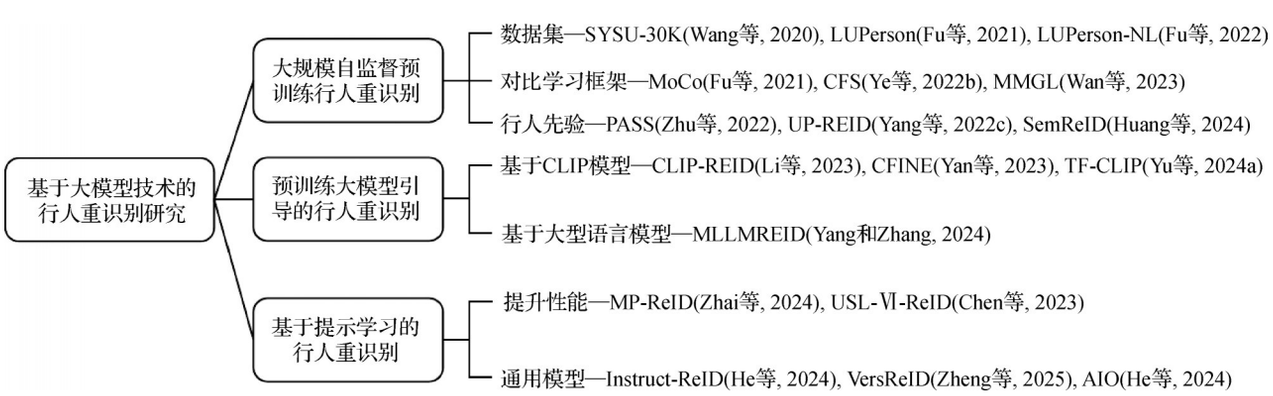

图2 基于大模型预训练技术的行人重识别研究

Fig. 2 Researches of large-scale pre-train techniques for re-id

架，所有的视图通过学生网络，只有全局视图通过教师网络，通过知识蒸馏约束教师网络和学生网络的输出，学生网络滑动平均更新教师网络的参数，并分别对全局—全局，全局—局部和身份特征之间进行对比学习，最终在多个数据集的识别实验和跨数据集迁移实验取得最好的效果，显著提升了行人模型的泛化能力。Yang等人（2022c）发现对比学习框架中的数据增强会破坏人物图像中的判别线索，忽略行人的局部特征，提出基于类内正则化的大规模预训练框架UP-REID（unsupervised pre-training framework for re-identification），引入内部身份的正则化，提高模型对于数据增强的鲁棒性，并利用人体水平对称性的先验知识提出新的困难样本挖掘策略筛选正负样本，解决了无监督预训练中的增强可能会扭曲人物图像中的判别线索的问题。UP-REID方法采用MoCo框架和一致性对比损失，相比其他预训练学习方法对基准网络带来了更显著的提升。Yang等人（2022c）研究了自监督预训练技术在地铁站行人重识别任务的应用，针对地铁站等复杂场景对行人重识别的挑战，提出一种自监督预训练方法SPLT（spindle net），首先通过一种定向标记模块模拟实际场景中的相机风格变换和交叉分辨率问题，并引入面向行人图像的可学习特征编码器，设计了两个投影层，一个用于恢复被遮挡的嵌入表示，另一个用于提取判别性行人身体部位特征，并通过知识蒸馏的方法进行参数优化：教师网络使用三元组损失进行前向传播提供指导；学生网络使用对比损失函数进行优化，学习特征，实现知识蒸馏。Huang等人（2023b）探索自监督预训练技术在换衣行人重识别的应用，针对现实场景的行人重识别任务受到粗粒度、衣服颜色干扰和预定义区域的阻碍，引入更多的语义信息来学习鲁棒的换衣行人重识别模型。具体地，受交互式语义模型启发，提出局部语义提取模块捕获细粒度、特定的语义和生物识别相关的局部语义，提升行人重识别性能。在此基础上，提出自监督预训练学习方法SemReID，通过教师学生网络、多裁剪策略以及多头自注意等结构，利用局部语义提取以人为中心的语义信息，避免衣服等干扰信息，显著提升了换衣行人重识别模型的性能。

3.3 自监督预训练行人重识别技术的意义和局限

自监督预训练行人重识别模型的实验结果如表4所示，使用大规模无标签数据预训练的模型进行微调识别准确率明显更高，Market-1501和MSMT数据集识别率最好的算法都经过LUPerson预训练。对比同一个算法，使用无监督行人数据预训练的效果比用ImageNet数据效果要强，在目标数据集微调结果明显提升，譬如MGN（multiple granularity network）网络结构，经过LUPerson预训练之后，在Market-1501数据集的R1/mAP（mean average precision）从87.5%/95.1%提升到91%/96.4%，分别提升了3.5%和1.3%。此外，许多模型不需要使用目标场景的数据进行微调就可以取得媲美监督微调的效果，譬如使用MoCov2框架预训练的ResNet网络的识别性能和MGN框架使用标签数据微调的网络性能非常接近，SPLT和VersReID训练的VIT模型更是无需目标数据微调就超过了主流的监督学习算法。

总体而言，当前已有一些自监督预训练行人重识别研究，并且已经证明自监督预训练技术对于提升行人重识别模型的性能和泛化能力是非常重要的。但是，目前基于自监督预训练技术的行人重识别研究还处于探索阶段，仍然有许多问题需要解决。首先，现在的学术界开源的大规模行人数据集很少，目前能找到的开源数据集只有SYSU-30K和LUPerson两个大规模的数据集，数据很难覆盖所有场景，譬如涉及夜间红外监控和空中航拍视角监控的行人数据就会比较少，预训练得到的模型泛用性有限。其次，当前使用的自监督预训练技术以MoCo，SimCLR等面向一般视觉任务的预训练为主，网络结构也是以ResNet和ViT等通用框架比较多，如何融合行人先验知识，如何设计适合行人结构的神经网络仍然是一个很值得探索的课题。最后，受限于应用场景行人数据较少，当前的预训练行人重识别模型只使用到了大规模数据，而没有设计大规模模型，怎样发挥大模型的规模优势，利用海量参数学习通用的行人重识别模型，也是有待开发的问题。

4 基于大模型技术的行人重识别研究

由于标注行人图像收集成本高、难度大，测试场景往往是没有标注数据的，因此对行人重识别模型的泛用性提出很高的要求，需要尽可能提高模型泛化能力，减少对标签数据的依赖，形成通用的行人重识别模型，从而在未知的场景做到即插即用。大模

表4　基于大规模无监督预训练技术的行人重识别方法实验结果

Table 4　Experimental results for re-id methods based on large-scale unsupervised pre-train techniques

/%

方法	基准网络	预训练数据	下游微调	Market-1501	MSMT17
				R1/mAP	R1/mAP
MGN	ResNet50	ImageNet	是	95.1/87.5	85.1/63.7
ABDNet	ResNet50	ImageNet	是	95.6/88.3	82.3/60.8
TransREID	VIT-B	ImageNet	是	94.8/88.2	82.5/63.6
MoCoV2	ResNet50	LUPerson	否	95.1/87.3	76.8/53.3
MoCoV2	VIT-S	ImageNet	否	72.1/63.6	36.1/19.6
MoCoV2	VIT-S	LUPerson	否	87.6/72.2	47.4/27.8
DINO+CFS	VIT-S	LUPerson	否	94.2/88.2	66.4/40.9
UP-REID	ResNet50	LUPerson	否	90.0/75.1	–/–
MAE	VIT-B	ImageNet	否	91.5/80.3	75.2/53.2
SPLT	VIT-S	LUPerson	否	95.1/89.8	67.1/42.4
VersReID	VIT-B	LUPerson	否	96.1/91.9	88.5/73.6
UP-REID	ResNet50	LUPerson	是	**97.1**/91.1	84.3/63.3
DINO+CFS	VIT-S	LUPerson	是	96.0/91.0	84.6/66.1
PASS	VIT-S	LUPerson	是	96.3/92.2	89.5/69.1
PASS	VIT-B	LUPerson	是	96.9/93.3	**89.7**/74.3
MAE	VIT-B	ImageNet	是	95.6/88.8	83.8/66.5
MGN	ResNet50	LUPerson	是	96.4/91.0	85.5/65.7
SPLT	VIT-S	LUPerson	是	96.6/92.8	89.1/75.1
SPLT	VIT-B	LUPerson	是	96.8/**93.4**	**89.7**/**75.3**

注：加粗字体表示各列最优结果，“–”表示暂无相关数据。

型通过海量的数据挖掘通用的知识，其参数蕴含了丰富的先验知识，此外，多模态大模型学习了可靠的文本—视觉特征关联，因此基于大模型学到的文本提示可以辅助视觉模型挖掘更通用的特征。由于预训练大模型的上述性质，越来越多的学者研究如何使用预训练大模型技术来提高行人重识别模型的泛化能力，学习通用的模型。当前，利用大模型的行人重识别研究主要有两种，一种是基于大模型参数的方法；另外一种是基于提示学习的方法。

4.1　预训练大模型引导的行人重识别方法

基于大模型参数的方法以大模型的参数作为初始化去提取更鲁棒的特征，由于预训练学习的参数蕴含丰富而鲁棒的知识，在下游往往能获得更好的效果。由于CLIP模型在各种下游任务（包括图像分类和分割）上表现出了优越的性能和泛化能力（Zhou等，2022），因此用于提升行人重识别模型的性能和泛化能力，简单地微调CLIP中图像编码器初始化的视觉模型，已经在各种行人重识别任务中获得了具有竞争力的性能。但是，在行人重识别任务中，标签是索引，缺乏具体的文本描述，因此无法使用预训练CLIP模型的文本—视觉先验知识。为了解决上述问题，更好地利用CLIP的先验知识，Li等人（2023）提出一个两阶段的CLIP-REID框架来促进更好的视觉表示。首先，为每个行人设置一个可学习的模板，并将它们提供给文本编码器以形成模糊的描述，以此利用CLIP的跨模态描述能力。第1阶段利用可学习模板来挖掘但又不破坏CLIP原有的特性，而在第2阶段，再根据学好的模板和文本特征来优化图像特征以此获取更好的泛化能力。具体地，在第1阶段将来自CLIP的图像和文本编码器保持固定，通

过对比损失从头开始优化模板学习过程。在第2阶段，将特定于身份的模板及其文本编码器设为静态，利用学习好的模板微调图像编码器。在下游任务中，通过设计的损失函数，图像编码器能够将数据准确地表示为特征嵌入向量。CLIP-REID在多个行人和车辆的重识别数据集中证明了能显著提高模型的识别准确率和泛化性能，并开启了利用多模态大模型参数优化行人重识别模型泛化能力的先河，在此之后出现了更多的研究。Yan等人（2023）提出CLIP驱动的细粒度文本—图像行人重识别方法，利用CLIP模型的文本—视觉先验减少文本描述和行人图像的巨大鸿沟，实现图像特征嵌入和文本空间的跨模态对齐。具体地，提出CLIP驱动的细粒度信息挖掘框架CFine（cross-feature fusion），利用CLIP在多模态预训练过程中学到的文本和视觉知识及其联系。首先，设计了一个多粒度的全局特征学习模块，利用细粒度信息挖掘模态内判别特征和模块间的对应关系，通过增强全局信息和局部信息之间的相互作用，充分挖掘每个模态内的判别局部信息，从而强调与身份相关的判别线索。然后，提出跨模态特征细化和细粒度对应模块来建立模态之间细粒度特征的关系，过滤不重要和非模态共享特征，并从粗到细挖掘跨模态对应关系。在多个基准上的广泛实验表明，CFine方法能挖掘并利用文本蕴含的信息，显著提升文本—图像行人重识别的性能，超过了其他主流方法。Yu等人（2024b）探讨了CLIP在跨模态行人重识别的应用，发现可见光行人和红外行人图像尽管在外观上存在模态差距，但是行人外观的高层次语义信息（如性别、形状和穿衣风格）在不同模态之间仍然保持一致，因此提出一个CLIP驱动的语义发现网络CSDN，通过注入高层次语义的视觉特征来弥合模态差距。该网络由模态特定提示学习器、语义信息整合和高层次语义嵌入等模块组成。考虑到语言描述中的模态差异带来的多样性，设计了双模态可学习模板，即为一个行人在不同模态上设计不同的模板以分别捕获可见光和红外图像的模态语义信息。此外，鉴于不同模态语义细节的互补性，该网络整合了来自双模态语言描述的文本特征以实现全面的语义。最后，建立了整合的文本特征与跨模态视觉特征之间的联系，将丰富的高层次语义信息嵌入到视觉表示中，从而促进视觉表示的模态不变性。通过在多个广泛使用的基准数据集上的实验评估证实了CSDN（CLIP-driven semantic discovery network）结构在现有方法中的有效性和优越性。Li等人（2024b）同样将CLIP模型用于文本行人重识别任务，引入了提示微调策略来实现域适应，在训练阶段解耦CLIP模型微调的过程，并提出一种两阶段训练方法，将域适应从任务适应中分离出来。在第1阶段，冻结了CLIP中的两个编码器，只专注于优化可学习的提示，以缓解CLIP和下游任务的原始训练数据之间的域差距。在第2阶段，保持固定的提示并微调CLIP模型，以优先捕获细粒度信息，提取更适合文本行人重识别的特征。最后，在3个广泛使用的数据集上证明了方法的有效性。与直接微调的方法相比，取得了显着的改进。Yu等人（2024a）研究如何将多模态模型学到的知识应用到基于视频的行人重识别，提出一种基于CLIP的单阶段无文本学习框架TF-CLIP（text-free CLIP）。具体地，提取了特定于身份的序列特征作为CLIP存储库，以取代文本特征。同时，设计了一个序列特定提示模块来在线更新CLIP存储库。为了捕捉时间信息，进一步提出时间记忆扩散模块，该模块由两个关键部分组成：时态记忆和记忆扩散。时态记忆模块实现序列中的帧级记忆相互通信，并根据序列中的关系提取时间信息；记忆扩散模块则进一步将时间记忆扩散到原始特征中，从而获得更稳健的序列特征。TF-CLIP在几个主流视频行人数据集的效果要明显优于其他最先进的方法。Yang和Zhang（2024）首次探索使用多模态语言大模型在行人重识别的应用，提出基于多模态大型语言模型的行人模型MLLMREID（multimodal large language model-based person re-identification），设计通用指令学习方法和基于多任务学习的同步模块确保MLLM的视觉编码器与行人重识别任务同步训练，实验证明了方法的有效性。

4.2 基于提示学习的通用行人重识别方法

提示学习是提升大语言模型和多模态大模型泛化能力的重要过程，因此研究人员将提示学习结合到行人重识别模型，基于提示学习方法利用大模型的文本—视觉关联去学习通用的知识，提升模型的泛化能力。Zhai等人（2024）探讨了在行人重识别任务将生成的细粒度多行人属性描述作为提示补充行人图像的宝贵语义信息，并结合大模型一起使用获得更准确检索结果的潜力。在提示学习和语言模型的基础上，提出多提示行人重识别的新框架MP-

ReID(multi-prompts person re-identification),以充分利用属性信息和提示特征来辅助行人重识别任务。具体来说,MP-ReID首先学习生成多种多样、信息丰富且可提示的句子来描述查询图像,包括:1)显性提示:明确提示一个人具有哪些属性,通过聚合生成模型(如ChatGPT)获得;2)隐形提示:隐含的可学习提示,用于调整/调节该人身份匹配的标准。然后,MP-ReID提出一个对齐模块以逐步融合多参数(即显性和隐性参数),并缩小跨模态差距。对于显性提示,首先对提示进行进一步处理得到显性提示特征,并与图像特征进行对齐,相似地,对隐性提示得到的特征也与图像特征进行对齐。在现有的涉及属性的行人数据集上进行大量实验的结果证明了MP-ReID框架的有效性和合理性,并显著超越了其余主流的方法。Chen等人(2023b)提出一种新的无监督可见红外人再识别提示学习框架USL-VI-ReID(unsupervised learning-visible infrared- person reidentification),利用大模型(如CLIP等)为无监督聚类学习提供文本描述提示特征作为补充的语义知识。CLIP-REID需要为每一个行人生成一个可学习的提示模板,但在无监督行人重识别中没有行人标签,所以通过聚类的方式先得到行人的伪标签,然后为每一个伪标签行人生成一个可学习的提示,然后通过训练得到符合图像的提示,并利用提示模板作为监督信息引导后续模型的训练。此外,由于聚类是在单一模态进行的,导致不同模态同一行人具有不同的标签,为了解决这个问题,提出一种记忆交换对比学习方法,首先利用原型,即每个模态下每个类别的聚类中心,通过匈牙利匹配方法找到匹配的跨模态行人原型,并交换记忆中的原型对以消除模态差异。这样,对比学习在不做任何改变的情况下就能轻松地关联跨模态信息,从而引导不同的模态逐渐走向融合。在主流的可见光—红外行人数据集的实验证明了该方法能显著提升非监督跨模态行人模型的识别率,并且在SYSU-MM01数据集的结果比最先进的对比方法提高了9%以上。

部分研究人员利用预训练模型的文本—视觉关联,通过提示学习研究通用行人重识别模型,使用单一模型同时处理多个行人任务。He等人(2024)提出基于指示学习的行人重识别框架根据视觉图像和语言描述检索来同时应对多个行人重识别任务,通过设计不同的指令去解决不同的行人重识别任务。具体地,使用经过预训练的ViT网络作为图像编码器,使用CLIP作为提示学习的编码器,通过互注意力模块建立视觉特征、提示信息和不同任务之间的联系,集成了多个主流数据集的数据形成了一种大规模的OmniReID数据集,并使用自适应三元组损失作为度量学习方法。通过在多个任务的实验表明,该模型可以使用单一的模型自适应地处理不同的行人重识别任务并且在每一个任务都取得显著的提升。Zheng等人(2025)提出VersReID框架,通过一个通用的模型同时处理多个行人重识别任务。具体地,VersReID建立了一个基于提示的两阶段孪生框架,首先利用场景标签训练一个包含丰富知识的行人重识别银行库以处理各种场景,不同场景通过特定的提示编码场景特定的知识。在第2阶段,将不同场景知识蒸馏到V-Branch模型,从而提炼出一个集成多场景知识的多功能提示,用于自适应地解决不同场景的行人重识别任务,消除在推理阶段需要场景标签的需求。此外,VersReID还通过为每个图像生成具有不同场景特性的多个增强视图模拟不同场景数据,达到数据增强的目的,在自监督学习阶段提高模型对于多场景的泛化能力。通过大量实验证明了在不需要在推理阶段手动分配场景标签的情况下,学习一个有效且多场景的通用行人重识别模型可以成功地解决不同的任务,包括一般、低分辨率、着装变化、遮挡和跨模态场景。Li等人(2024a)借鉴提示学习方法提出能够有效处理多种模态数据(包括颜色、红外、素描和文本信息)的统一多模型行人重识别框架AIO(all-in-one),它利用一个冻结的预训练大模型作为编码器,训练一个可学习的多模态特征编码器来适应不同的模态。AIO中的多种模态数据经过多模态特征编码器投射到统一空间,然后允许模态共享的冻结编码器提取所有模态的一致身份特征。此外,一个精心设计的跨模态头集合用于引导学习,包括一个传统的分类头用于学习行人特征;一个视觉引导的掩码属性建模,用于引导模型学习细粒度的人体属性特征;一个多模态特征绑定,用于对齐多模态特征。AIO是实现全能型重识别的框架,涵盖4种常用模态。在跨模态和多模态数据集的实验表明,AIO不仅能够熟练处理各种模态数据,还能在具有挑战性的环境中表现出色,在零样本和领域泛化场景中展示了卓越的性能。

4.3 基于大模型技术的通用行人重识别实验结果

基于大模型技术的通用行人重识别算法实验结果见表5，每个数据集的识别准确率用R1和mAP表示。基于CLIP模型的行人重识别研究在可见光、文本行人识别以及红外行人识别等多种挑战都显著提升了基准模型的识别性能。此外，AIO、VersREID、IRM等算法都能同时处理可见光行人、换衣行人、红外行人等多种挑战，IRM(STL)和MPDA模型是IRM(MTL)和VersREID分别使用单一数据库训练的模型。其中AIO没有使用测试场景的样本进行训练，因此总体性能偏低。没有使用大规模行人库预训练，使用单一数据进行训练的TransREID得到的模型识别率偏低。经过预训练之后只使用特定数据集微调训练的MPDA和IRM(STL)模型的识别率尽管已经达到了主流模型的识别率，但是相比使用相同基准网络和多个数据集与提示学习方法协同训练的VersReID和IRM(MTL)模型有明显差距。因此，提示学习方法不但能将多种人物的数据融合到一个模型，实现单一模型解决多个场景的问题，还能挖掘不同场景之间的行人共享知识，从而提升模型在单独场景的识别效果。

表5 基于大模型预训练技术的通用行人重识别算法实验结果

Table 5 Experimental results of universal re-id models based on large-scale pre-train techniques

/%

算法	场景标签	R1/mAP							
		Market	MSMT17	MLR-CUHK	CUHK-PEDES	PRCC	OCC-DUKE	SYSU-MM01	LLCM
ViT	需要	93.3/86.4	84.4/66.1	–/–	–/–	–/–	–/–	–/–	–/–
ViT+CLIP	需要	95.4/90.5	**89.7/75.8**	–/–	–/–	–/–	–/–	–/–	–/–
ViT	需要	–/–	–/–	–/–	57.89/–	–/–	–/–	–/–	–/–
ViT+CFine	需要	–/–	–/–	–/–	69.57/–	–/–	–/–	–/–	–/–
CLIP	需要	–/–	–/–	–/–	–/–	–/–	–/–	70.4/–	–/–
CLIP+CSDN	需要	–/–	–/–	–/–	–/–	–/–	–/–	**75.2**/–	–/–
AIO	不需要	79.5/59.9	–/–	–/–	53.4/44.4	–/–	–/–	57.6/51.9	–/–
TransREID	需要	95.0/89.7	85.8/69.4	–/–	–/–	48.5/62.2	64.2/56.6	58.8/59.5	–/–
MPDA	需要	96.1/91.9	88.6/74.1	95.8/94.9	–/–	54.4/66.6	72.7/64.0	64.6/63.5	–/–
VersReID	不需要	96.8/93.2	88.8/74.2	**97.5/98.4**	–/–	**60.7/71.4**	**75.2/66.1**	69.3/**66.9**	–/–
IRM(STL)	需要	96.2/92.3	86.2/71.9	–/–	48.1/46.0	46.0/48.1	–/–	–/–	64.9/64.5
IRM(MTL)	不需要	**96.5/93.5**	86.9/72.4	–/–	**74.2/66.5**	52.3/54.2	–/–	–/–	**65.7/67.2**

注：加粗字体表示各列最优结果，“–”表示该指标暂无相关数据。

5 结 语

5.1 总结与讨论

行人重识别是计算机视觉研究热点和智能监控领域的重要任务之一，随着相关技术的发展和社会安防的需求升级，得到了越来越多的关注。过去十数年，行人重识别技术得到了长足的发展，传统行人重识别模型性能大幅提升，并且在非可控环境的诸多任务也有深入的研究，研究深度和广度都很突出。但是，受限于行人数据收集难度大、成本高，行人重识别标注数据集的规模远远小于其他主流视觉任务，导致行人重识别迟迟未能落地应用。恰逢最近几年预训练大模型研究如火如荼，其利用大规模无监督数据学习先验知识提升模型通用泛化性能的核心思想能帮助行人重识别研究摆脱数据困境，越来越多的研究人员也注意到了这一点，开始研究如何将预训练大模型技术用于行人重识别任务。基于此，本文对预训练大模型技术在行人重识别领域的相关研究进行了全面的回顾与介绍，对前沿的算法进行了讨论，总结如下：

1）自监督预训练和大规模数据集是基础和支撑。预训练大模型在自然语言处理和多模态领域都展示了极强的泛化能力，激发了视觉研究工作者将

大规模自监督预训练技术用在视觉领域的热情，自监督预训练多年来积累的经验和算法也为研究者打下了基础，带来了启发。最早关于大规模预训练行人重识别的研究基本上是套用其他任务预训练大模型的结构，现在主流使用的预训练行人重识别模型仍然受到已有算法的影响，其结构设计仍然有GPT、BERT等自然语言处理预训练方法的影子。而大规模数据集是大规模预训练行人重识别研究的重要支撑，自LUPerson等大规模行人数据集出现以来，许多大规模预训练的研究才得以展开，并且能通过同样的开源数据集比较来判断算法的好坏。但是，当前大规模行人库的数量较少，规模也不够大，而且数据往往是从网上下载，与实际监控场景的差异较大，还无法满足碎片化场景的应用需求。上述问题导致尽管大规模自监督预训练技术能提升行人重识别算法的性能和泛化能力，但仍做不到即插即用，还有很多的研究空间。一方面，需要规模更大、覆盖范围更广，场景更接近实际应用场景的大规模行人数据集；另一方面，仍然需要研究如何更好地结合自监督预训练和行人重识别任务。

2）多模态大模型结合提示学习训练通用行人重识别模型是趋势。多模态大模型通过海量的文本—图像对训练样本学到了通用的知识和文本—图像之间的关联，CLIP等多模态大模型的应用证明了文本—图像的关联知识对识别和分类任务泛化能力的重要意义，以此形成了提示学习的范式。近年，提示学习和行人重识别框架结合已经成了一个重要的趋势，越来越多的研究人员将提示学习的范式用到行人重识别任务，并且在多个方向显著提升了模型的性能和泛化能力。随着行人重识别泛化能力的增强，近期更是涌现了许多通用行人重识别模型的研究，通过设计提示框架，将多个任务结合到一个统一的模型，显著增强了行人重识别模型的泛化性能。但是，当前研究对于提示学习研究的框架设计还比较初步，很多研究直接将多模态大模型的提示学习框架用到行人重识别任务，因此对于如何挖掘提示学习在行人重识别的应用还有很大的发展空间和潜力。

3）如何结合先验行人知识形成统一的预训练大模型框架是关键。当前基于预训练大模型技术的行人重识别研究还处于起步和探索的阶段，主流的研究思路仍然停留在将成果的预训练大模型经验用在行人重识别上，如何使用行人结构等先验知识提升大模型预训练技术的效果仍然有待开发。此外，由于实际测试环境的数据集规模较小，当前行人重识别研究在运营大模型预训练技术的时候，只有大数据，而没有大模型。当前使用大规模无监督数据训练的行人重识别模型基准网络往往是ResNet和ViT等网络，即使是使用CLIP模型参与训练的研究，在测试阶段也只使用小规模的网络结构。因此，当前尚无关于大规模行人重识别模型的研究。怎样结合行人先验知识，设计适合行人先验的网络结构，从而支撑大规模模型的训练，大模型在行人重识别任务的泛化能力如何，怎样提升其性能，也是很值得研究的内容。

5.2 未来展望

随着基于自监督预训练大模型的行人重识别技术的发展，当前的技术也存在一定的问题，从这些问题出发可以展望未来的发展趋势：

1）更大规模的无标签/合成行人数据集。大规模数据集是自监督预训练学习的根本，研究者已经通过早期的尝试验证了大规模数据集对行人重识别模型性能和泛化能力的提升。但是，当前已有的无监督预训练数据集无论在数据规模还是数据对应用的覆盖面上还有很大的发展空间。一方面，与自然语言处理、多模态大模型预训练和其他的视觉任务相比，目前的无监督行人数据只有数百万，不足以实现从量变到质变；另一方面，数据主要从网上下载，与实际应用场景差异很大，有必要从实际监控场景获取部分数据提升数据集的意义和作用。因此，当前的大规模无监督行人数据集从数据规模和场景等方面还不足以支撑通用行人重识别研究，需要更大规模的无监督数据集去支持大规模的预训练模型，嵌入更贴近实际应用场景的数据去提升模型的泛化能力，如何选择无监督行人数据集的采样场景，如何快速构建大规模数据集都是需要进一步研究的内容。此外，当前已经有虚拟合成的行人数据集，尽管由于技术原因，当前虚拟数据集与真实数据集差异很大，只能起到辅助作用，但是未来随着虚拟现实增强和虚拟生成技术的发展，虚拟人物生成的技术会越来越成熟，甚至可以从虚拟数据产生接近现实世界风格的数据，在未来也是一个可以期待的方向。怎样产生更逼真的虚拟行人图像，怎样模拟现实各种场景和挑战，怎样利用虚拟行人辅助实际场景训练，怎样从虚拟行人生成逼真多源的行人图像，都是

值得研究的课题。

2)嵌入行人先验知识的预训练/提示学习框架。当前基于预训练和提示学习框架的行人重识别技术处于探索阶段，尚未形成统一的特征提取模型和处理框架，主流的研究大多从现有的自监督技术/多模态提示学习技术出发，将已有的技术生搬硬套到行人重识别任务，很少结合行人领域知识进行研究，设计的预训练框架和提示学习框架和行人重识别任务的结合不够紧密。由于自然语言处理任务乃至于其他视觉任务和行人重识别任务有较大的差异，已有的理论方法和算法框架虽然能提供参考和启发，但是直接使用的效果不佳，甚至可能把网络结构设计的方向带偏。因此有必要研究如何嵌入行人先验知识来改进自监督预训练/提示学习的框架，可以预见未来研究如何利用行人先验知识设计预训练方法的工作将越来越多，如何把行人结构信息等嵌入到预训练模型和提示学习框架，如何利用行人先验知识设计网络结构，学习泛化能力更强的通用行人重识别模型仍然是一个开放的课题，行人结构信息、行人属性、视角信息以及姿态等多种与行人紧密相关的特征都可能被用来学习在应用监控场景鲁棒的模型。

3)基于大规模模型的通用行人重识别。当前预训练技术在行人重识别的应用只有大数据预训练而没有大模型，尽管已经有一些探索性的研究实现了在多个行人数据集和任务的识别率都比较高的算法，这些算法的基准网络模型仍然是行人重识别常见的小模型，如ResNet和ViT等。一方面，当前预训练数据集和下游场景测试集的规模较小，导致大规模模型在行人重识别任务没有用武之地。但是未来随着行人数据集规模的增长，已有的模型规模将难以处理逐渐增长的行人知识，大模型的优势会越来越明显，将来会出现只有大模型甚至是超大模型才能处理的数据和情况。考虑未来的应用和技术趋势，有必要研究实现基于大规模模型的通用行人重识别模型。另一方面，当前自监督预训练行人重识别算法使用的往往是主流的视觉模型或多模态大模型作为基准结构，并不适合作为大规模行人重识别模型基准，如何设计更适合行人重识别任务的大模型结构，如何提升大模型在行人重识别任务的泛化能力，如何实现碎片化场景通用行人重识别模型是未来重要的研究方向。

参考文献(References)

An X, Deng J K, Guo J, Feng Z Y, Zhu X H and Yang J. 2022. Killing two birds with one stone: efficient and robust training of face recognition CNNs by partial FC//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4032-4041 [DOI: 10.1109/CVPR52688.2022.00401]

Bao H B, Dong L, Piao S H and Wei F R. 2022. BEiT: BERT pre-training of image transformers//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: ICLR

Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #159

Chen S Y, Ye M and Du B. 2022. Rotation invariant transformer for recognizing object in UAVs//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 2565-2574 [DOI: 10.1145/3503161.3547799]

Chen T, Kornblith S, Norouzi M and Hinton G. 2020a. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: #149

Chen W H, Xu X Z, Jia J, Luo H, Wang Y H, Wang F, Jin R and Sun X Y. 2023a. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 15050-15061 [DOI: 10.1109/CVPR52729.2023.01445]

Chen Y C, Li L J, Yu L C, Kholy A E, Ahmed F, Gan Z, Cheng Y and Liu J J. 2020b. UNITER: universal image-TExt representation learning//Proceedings of the 16th European Conference. Glasgow, UK: Springer: 104-120 [DOI: 10.1007/978-3-030-58577-8_7]

Chen Z, Zhang Z Z, Tan X, Qu Y Y and Xie Y. 2023b. Unveiling the power of CLIP in unsupervised visible-infrared person re-identification//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 3667-3675 [DOI: 10.1145/3581783.3612050]

Chen S Z, Guo C C and Lai J H. 2016. Deep ranking for person re-identification via joint representation learning. IEEE Transactions on Image Processing, 25(5): 2353-2367 [DOI: 10.1109/TIP.2016.2545929]

Cheng Z Y, Dong Q, Gong S G and Zhu X T. 2020. Inter-task association critic for cross-resolution person re-identification//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 2602-2612 [DOI: 10.1109/CVPR42600.2020.00268]

Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR. 2009. 5206848]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Minneapolis, Minnesota, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: ICLR

Fang X Y, Yang Y and Fu Y. 2023. Visible-infrared person re-identification via semantic alignment and affinity inference//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 11236-11245 [DOI: 10.1109/ICCV51070.2023.01035]

Feng Z X, Lai J H and Xie X H. 2020. Learning modality-specific representations for visible-infrared person re-identification. IEEE Transactions on Image Processing, 29: 579-590 [DOI: 10.1109/TIP. 2019.2928126]

Feng Z X, Zhu R, Wang Y J and Lai J H. 2020. Overview of person re-identification in unconstrained environments. Acta Scientiarum Naturalium Universitatis Sunyatseni, 59(3): 1-11 (冯展祥, 朱荣, 王玉娟, 赖剑煌. 2020. 非可控环境行人再识别综述. 中山大学学报(自然科学版), 59(3): 1-11) [DOI: 10.13471/j.cnki.acta.snus.2020.03.001]

Fu D P, Chen D D, Bao J M, Yang H, Yuan L, Zhang L, Li H Q and Chen D. 2021. Unsupervised pre-training for person re-identification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14745-14754 [DOI: 10.1109/CVPR46437.2021.01451]

Fu D P, Chen D D, Yang H, Bao J M, Yuan L and Zhang L. 2022. Large-scale pre-training for person re-identification with noisy labels//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 1-11 [DOI: 10.1109/CVPR52688.2022.00251]

Gray D, Brennan S and Tao H. 2007. Evaluating appearance models for recognition, reacquisition, and tracking//10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. [s.l.]: [s.n.]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15979-15988 [DOI: 10.1109/CVPR52688.2022.01553]

He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9726-9735 [DOI: 10.1109/CVPR42600.2020.00975]

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. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He S T, Luo H, Wang P C, Wang F, Li H and Jiang W. 2021. TransReiD: transformer-based object re-identification//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14993-15002 [DOI: 10.1109/ICCV48922.2021.01474]

He W Z, Deng Y H, Tang S X, Chen Q H, Xie Q S, Wang Y Z, Bai L, Zhu F, Zhao R, Ouyang W L, Qi D L and Yan Y F. 2024. Instruct-ReID: a multi-purpose person re-identification task with instructions//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 17521-17531 [DOI: 10.1109/CVPR52733.2024.01659]

Hou R B, Ma B P, Chang H, Gu X Q, Shan S G and Chen X L. 2019. VRSTC: occlusion-free video person re-identification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7176-7185 [DOI: 10.1109/CVPR.2019.00735]

Huang M Y, Hou C P, Yang Q Y and Wang Z P. 2023a. Reasoning and tuning: graph attention network for occluded person re-identification. IEEE Transactions on Image Processing, 32: 1568-1582 [DOI: 10.1109/TIP.2023.3247159]

Huang M Y, Hou C P, Zheng X B and Wang Z P. 2024. Multi-resolution feature perception network for UAV person re-identification. Multimedia Tools and Applications, 83(23): 62559-62580 [DOI: 10.1007/s11042-023-17937-8]

Huang S Y, Zhou Y F, Prabhakar M, Liu X J, Guo Y X, Yi H R, Peng C, Chellappa R and Lau C P. 2023b. Self-supervised learning of whole and component-based semantic representations for person re-identification [EB/OL]. [2024-07-10]. http://export.arxiv.org/pdf/2311.17074v3.pdf

Karanam S, Gou M R, Wu Z Y, Rates-Borras A, Camps O and Radke

R J. 2019. A systematic evaluation and benchmark for person re-identification: features, metrics, and datasets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(3): 523-536 [DOI: 10.1109/TPAMI.2018.2807450]

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C and Gustafson L. 2023. Segment anything//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Li H, Ye M, Zhang M and Du B. 2024a. All in one framework for multimodal re-identification in the wild//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 17459-17469 [DOI: 10.1109/CVPR52733.2024.01653]

Li K, Ding Z M, Li S and Fu Y. 2018. Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI: 2331-2338 [DOI: 10.1609/aaai.v32i1.11908]

Li L H, Yatskar M, Yin D, Hsieh C J and Chang K W. 2019. VisualBERT: a simple and performant baseline for vision and language [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/1908.03557.pdf

Li M K, Li C G and Guo J. 2022. Cluster-guided asymmetric contrastive learning for unsupervised person re-identification. IEEE Transactions on Image Processing, 31: 3606-3617 [DOI: 10.1109/TIP.2022.3173163]

Li S Y, Sun L and Li Q L. 2023. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 1405-1413 [DOI: 10.1609/aaai.v37i1.25225]

Li T J, Liu J, Zhang W, Ni Y, Wang W Q and Li Z H. 2021. UAV-human: a large benchmark for human behavior understanding with unmanned aerial vehicles//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16261-16270 [DOI: 10.1109/CVPR46437.2021.01600]

Li W, Zhao R, Xiao T and Wang X G. 2014. DeepReiD: deep filter pairing neural network for person re-identification//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 152-159 [DOI: 10.1109/CVPR.2014.27]

Li W H, Tan L, Dai P Y and Zhang Y. 2024b. Prompt decoupling for text-to-image person re-identification [EB/OL]. [2024-01-04]. https://arxiv.org/pdf/2401.02173.pdf

Lin J A, Bao C Z, Dong J F, Yang X and Wang X. 2024. Multilingual text-video cross-modal retrieval model via multilingual-visual common space learning. Chinese Journal of Computers, 47(9): 2195-2210 (林俊安, 包翠竹, 董建锋, 杨勋, 王勋. 2024. 基于多语言-视觉公共空间学习的多语言文本—视频跨模态检索模型. 计算机学报, 47(9): 2195-2210) [DOI: 10.11897/SP.J.1016.2024.02195]

Liu X, Song M L, Tao D C, Zhou X C, Chen C and Bu J J. 2014. Semi-supervised coupled dictionary learning for person re-identification//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 3550-3557 [DOI: 10.1109/CVPR.2014.454]

Lu J S, Batra D, Parikh D and Lee S. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 2

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]

Luo H, Wang P C, Xu Y, Ding F, Zhou Y X, Wang F, Li H and Jin R. 2021. Self-supervised pre-training for transformer-based person re-identification [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/2111.12084.pdf

Nguyen H, Nguyen K, Sridharan S and Fookes C. 2024. AG-ReID.v2: bridging aerial and ground views for person re-identification. IEEE Transactions on Information Forensics and Security, 19: 2896-2908 [DOI: 10.1109/TIFS.2024.3353078]

Ni H, Li Y K, Gao L L, Shen H T and Song J K. 2023. Part-aware transformer for generalizable person re-identification//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 11246-11255 [DOI: 10.1109/ICCV51070.2023.01036]

Qian X L, Fu Y W, Jiang Y G, Xiang T and Xue X Y. 2017. Multi-scale deep learning architectures for person re-identification//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5409-5418 [DOI: 10.1109/ICCV.2017.577]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: ICML: 8748-8763

Radford A, Narasimhan K, Salimans T and Sutskever I. 2018. Improving language understanding by generative pre-training [EB/OL]. [2024-07-10]. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Ramesh A, Dhariwal P, Nichol A, Chu C and Chen M. 2022. Hierarchical text-conditional image generation with CLIP latents [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/2204.06125.pdf

Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M and Sutskever I. 2021. Zero-shot text-to-image generation//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: ICML: 8821-8831

Ren K J and Zhang L. 2024. Implicit discriminative knowledge learning for visible-infrared person re-identification//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 393-402 [DOI: 10.1109/CVPR52733.2024.00045]

Shao S, Li Z M, Zhang T Y, Peng C, Yu G and Zhang X. 2019. Objects365: a large-scale, high-quality dataset for object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE: 8429-8438 [DOI: 10.1109/ICCV.2019.00852]

Tian Y L, Wang Y T, Wang J G, Wang X and Wang F Y. 2022. Key problems and progress of vision Transformers: the state of the art and prospects. Acta Automatica Sinica, 48(4): 957-979 (田永林, 王雨桐, 王建功, 王晓, 王飞跃. 2022. 视觉Transformer研究的关键问题: 现状及展望. 自动化学报, 48(4): 957-979) [DOI: 10.16383/j.aas.c220027]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wan L, Jing Q Y, Sun Z Y, Zhang C, Li Z H and Chen Y. 2023. Self-supervised modality-aware multiple granularity pre-training for RGB-infrared person re-identification. IEEE Transactions on Information Forensics and Security, 18: 3044-3057 [DOI: 10.1109/TIFS.2023.3273911]

Wang G A, Yang S, Liu H Y, Wang Z C, Yang Y, Wang S L, Yu G, Zhou E J and Sun J. 2020. High-order information matters: learning relation and topology for occluded person re-identification//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6448-6457 [DOI: 10.1109/CVPR42600.2020.00648]

Wang T, Liu M Y, Liu H, Li W H, Ban M J, Guo T Y and Li Y D. 2024. Feature completion transformer for occluded person re-identification. IEEE Transactions on Multimedia, 26: 8529-8542 [DOI: 10.1109/TMM.2024.3379908]

Wang Z, Ye M, Yang F, Bai X and Satoh S. 2018. Cascaded SR-GAN for scale-adaptive low resolution person re-identification//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI: 3891-3897 [DOI: 10.24963/ijcai.2018/541]

Wang Z K, Zhu F, Tang S X, Zhao R, He L H and Song J N. 2022. Feature erasing and diffusion network for occluded person re-identification//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4744-4753 [DOI: 10.1109/CVPR52688.2022.00471]

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. Salt Lake City, USA: IEEE: 79-88 [DOI: 10.1109/CVPR.2018.00016]

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, 林城桔, 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]

Wu A C, Zheng W D, 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. Venice, Italy: IEEE: 5390-5399 [DOI: 10.1109/ICCV.2017.575]

Wu L Y, Liu L Q, Wang Y, Zhang Z, Boussaid F, Bennamoun M and Xie X H. 2023. Learning resolution-adaptive representations for cross-resolution person re-identification. IEEE Transactions on Image Processing, 32: 4800-4811 [DOI: 10.1109/TIP.2023.3305817]

Xie Z D, Zhang Z, Cao Y, Lin Y T, Bao J M, Yao Z L, Dai Q and Hu H. 2022. SimMIM: a simple framework for masked image modeling//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 9643-9653 [DOI: 10.1109/CVPR52688.2022.00943]

Xu B Q, He L X, Liang J and Sun Z N. 2022. Learning feature recovery transformer for occluded person re-identification. IEEE Transactions on Image Processing, 31: 4651-4662 [DOI: 10.1109/TIP.2022.3186759]

Yan S L, Dong N, Zhang L Y and Tang J H. 2023. CLIP-driven fine-grained text-image person re-identification. IEEE Transactions on Image Processing, 32: 6032-6046 [DOI: 10.1109/TIP.2023.3327924]

Yang E Z, Li C, Liu S Y, Liu Y X, Zhao S T and Huang N. 2022a. Self-supervised pre-training with learnable tokenizers for person re-identification in railway stations//Proceedings of the 16th IEEE International Conference on Signal Processing. Beijing, China: IEEE: 325-330 [DOI: 10.1109/ICSP56322.2022.9965305]

Yang L L, Lan L, Sun D T, Teng X, Ben X Y and Shen X B. 2023. Newly low-resolution pedestrian re-identification-relevant dataset and its benched method. Journal of Image and Graphics, 28(5): 1346-1359 (杨露露, 蓝龙, 孙冬婷, 滕霄, 贲晛烨, 沈肖波. 2023. 低分辨率行人重识别数据集及其基准方法. 中国图象图形学报, 28(5): 1346-1359) [DOI: 10.11834/jig.221082]

Yang M X, Huang Z Y, Hu P, Li T H, Lyu J C and Peng X. 2022b. Learning with twin noisy labels for visible-infrared person re-

identification//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 14288-14297 [DOI: 10.1109/CVPR52688.2022.01391]

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]

Yang S and Zhang Y F. 2024. MLLMReID: multimodal large language model-based person re-identification [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/2401.13201.pdf

Yang Z W, Lin M, Zhong X, Wu Y and Wang Z. 2023. Good is bad: causality inspired cloth-debiasing for cloth-changing person re-identification//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 1472-1481 [DOI: 10.1109/CVPR52729.2023.00148]

Yang Z Z, Jin X, Zheng K C and Zhao F. 2022c. Unleashing potential of unsupervised pre-training with intra-identity regularization for person re-identification//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 14278-14287 [DOI: 10.1109/CVPR52688.2022.01390]

Ye M, Shen J B, Lin G J, Xiang T, Shao L and Hoi S C H. 2022a. 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]

Ye Z A, Hong C Q, Zeng Z Q and Zhuang W W. 2022b. Self-supervised person re-identification with channel-wise transformer//Proceedings of 2022 IEEE International Conference on Big Data. Osaka, Japan: IEEE: 4210-4217 [DOI: 10.1109/BigData55660.2022.10020632]

Yi D, Lei Z, Liao S C and Li S Z. 2014. Learning face representation from scratch [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/1411.7923.pdf

Yu C Y, Liu X H, Wang Y Q, Zhang P P and Lu H C. 2024a. TF-CLIP: learning text-free CLIP for video-based person re-identification//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 6764-6772 [DOI: 10.1609/aaai.v38i7.28500]

Yu H, Cheng X, Peng W, Liu W H and Zhao G Y. 2023. Modality unifying network for visible-infrared person re-identification//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 11151-11161 [DOI: 10.1109/ICCV51070.2023.01027]

Yu X Y, Dong N, Zhu L H, Peng H and Tao D P. 2024b. CLIP-driven semantic discovery network for visible-infrared person re-identification [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/2401.05806.pdf

Zahra A, Perwaiz N, Shahzad M and Fraz M M. 2023. Person re-identification: a retrospective on domain specific open challenges and future trends. Pattern Recognition, 142: #109669 [DOI: 10.1016/j.patcog.2023.109669]

Zhai Y J, Zeng Y W, Huang Z Y, Qin Z, Jin X and Cao D. 2024. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 6979-6987 [DOI: 10.1609/aaai.v38i7.28524]

Zhang G Q, Ge Y, Dong Z C, Wang H, Zheng Y H and Chen S Y. 2021a. Deep high-resolution representation learning for cross-resolution person re-identification. IEEE Transactions on Image Processing, 30: 8913-8925 [DOI: 10.1109/TIP.2021.3120054]

Zhang Q, Lai C Z, Liu J N, Huang N C and Han J G. 2022. FMCNet: feature-level modality compensation for visible-infrared person re-identification//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 7339-7348 [DOI: 10.1109/CVPR52688.2022.00720]

Zhang S Z, Yang Q C, Cheng D, Xing Y H, Liang G Q, Wang P and Zhang Y N. 2023. Ground-to-aerial person search: benchmark dataset and approach//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 789-799 [DOI: 10.1145/3581783.3612105]

Zhang S Z, Zhang Q, Yang Y F, Wei X, Wang P, Jiao B L and Zhang Y N. 2021b. Person re-identification in aerial imagery. IEEE Transactions on Multimedia, 23: 281-291 [DOI: 10.1109/TMM.2020.2977528]

Zhang X, Luo H, Fan X, Xiang W L, Sun Y X, Xiao Q Q, Jiang W, Zhang C and Sun J. 2017. AlignedReiD: surpassing human-level performance in person re-identification [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/1711.08184.pdf

Zhang X Q, Feng W, Han R Z, Wang L K, Song L Q and Hou J H. 2024. Synthetic-to-real video person Re-ID [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/2402.02108.pdf

Zhao X, Ding W C, An Y Q, Du Y L, Yu T, Li M, Tang M and Wang J Q. 2023. Fast segment anything [EB/OL]. [2024-07-10]. https://arxiv.org/pdf/2306.12156.pdf

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. Santiago, Chile: IEEE: 1116-1124 [DOI: 10.1109/ICCV.2015.133]

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, Yan J K and Peng Y X. 2025. A versatile framework for multi-scene person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3): 1362-1380 [DOI: 10.1109/TPAMI.2024.3381184]

Zhou K Y, Yang J K, Loy C C and Liu Z W. 2022. Learning to prompt for vision-language models. International Journal of Computer

Vision, 130(9): 2337-2348 [DOI: 10.1007/s11263-022-01653-1]

Zhu K, Guo H Y, Yan T Y, Zhu Y S, Wang J Q and Tang M. 2022. PASS: part-aware self-supervised pre-training for person re-identification//Proceedings of the 17th European Conference. Tel Aviv, Israel: Springer: 198-214 [DOI: 10.1007/978-3-031-19781-9_12]

Zhu Z, Huang G, Deng J K, Ye Y, Huang J J and Chen X Z. 2021. WebFace260M: a benchmark unveiling the power of million-scale deep face recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10487-10497 [DOI: 10.1109/CVPR46437.2021.01035]

Zhuo J X, Chen Z Y, Lai J H and Wang G C. 2018. Occluded person re-identification//Proceedings of 2018 IEEE International Conference on Multimedia and Expo. San Diego, USA: IEEE: 1-6 [DOI: 10.1109/ICME.2018.8486568]

作者简介

冯展祥,男,副教授,主要研究方向为模式识别和计算机视觉。E-mail: fengzhx7@mail.sysu.edu.cn

赖剑煌,通信作者,男,教授,博士生导师,主要研究方向为计算机视觉与模式识别。E-mail: stsljh@mail.sysu.edu.cn

袁藏,男,硕士研究生,主要研究方向为计算机视觉。E-mail:yuanc9@mail2.sysu.edu.cn

黄宇立,男,硕士研究生,主要研究方向为计算机视觉。E-mail:huangyli23@mail2.sysu.edu.cn

赖培杰,男,本科生,主要研究方向为计算机视觉。E-mail:laipj3@mail2.sysu.edu.cn