行人搜索算法综述

李位星，张 瑜，贾普阳，高 琪，潘 峰*

（北京理工大学 自动化学院，北京 100081）

摘要 随着深度学习技术的快速发展，行人搜索算法的研究得到大量学者的关注。行人搜索是在行人检测和行人重识别任务的基础上在图像中寻找特定目标行人。该文对近年来行人搜索任务相关研究进展进行了详细梳理。按照模型网络结构和损失函数两方面对现有方法进行分析和总结。依据卷积神经网络和 Transformer 两类不同的技术路线，重点阐述各自代表性方法的主要研究工作；并按照传统损失函数、OIM 损失函数及混合损失函数对行人搜索采用的训练损失函数进行详细总结。此外，总结了行人搜索任务领域常用的公开数据集，比较和分析了主要算法在相应数据集上的性能表现。最后总结了行人搜索任务的未来研究方向。

关键词 行人搜索；卷积神经网络；Transformer；损失函数；深度学习

中图分类号 TP391 **文献标志码** A DOI 10.12178/1001-0548.2023260

Person Search Algorithm: A Survey

LI Weixing, ZHANG Yu, JIA Puyang, GAO Qi, and PAN Feng*

(School of Automation, Beijing Institute of Technology, Beijing 100081, China)

Abstract In recent years, with the rapid development of deep learning technology, the research of person search algorithms has attracted a lot of scholars' interest. Person search is to find specific target person in images based on person detection and person re-identification tasks. In this paper, we review the recent research progress on person search task in detail. The existing methods are analyzed and summarized in terms of model network structures and loss functions. According to the two different technical routes of convolutional neural network and Transformer, the main research work of their respective representative methods is focused on. According to the traditional loss function, OIM loss function, and mixed loss function, the training loss functions used in person search are summarized. In addition, the public data sets commonly used in the field of person search are summarized, and the performances of the main algorithms on the corresponding data sets are compared and analyzed. Finally, we summarize the future research directions of person search task.

Key words person search; convolutional neural networks; transformer; loss function; deep learning

行人重识别（Person Re-Identification, ReID）在计算机视觉领域已受到广泛关注并取得一定研究进展[1-3]。然而，现实场景中对特定行人的搜索是基于复杂的全景图像，而不是裁剪好的、尺寸相同且对齐的行人图像。因此，行人重识别技术无法直接应用于现实场景。行人搜索任务是直接从全景监控视频或图像中准确地找到目标行人，结合行人检测和行人重识别技术，在全景图像中对行人进行定位、特征提取，然后与目标行人特征进行匹配以确定行人身份信息。

基于视频的安全防范系统中摄像机的广泛使用为行人搜索技术的现实应用提供了基础条件。在城市行人搜索技术可协助警方快速地识别和追踪犯罪嫌疑人、寻找失踪人口，避免人工审查时间长、效率低的弊端；在智能交通领域，行人搜索技术可实现行人流量和拥堵情况的实时监控和分析，以及提高自动驾驶车辆的安全性和可靠性，准确地检测和识别周围行人，避免事故的发生。此外，行人搜索技术在教育、健康医疗和金融安全等领域可帮助工作者对行人活动进行分析，以便制定个性化的应对方案。

近年来，基于深度学习的行人搜索方法取得了

收稿日期：2023−10−18；修回日期：2024−04−24
基金项目：国家自然科学基金面上项目（61973036）
作者简介：李位星，高级实验师，主要从事深度学习与计算机视觉、目标检测与跟踪方面的研究。
*通信作者 E-mail: andropanfeng@126.com

显著发展，但仍面临行人遮挡、多尺度、位置、视角变化等挑战。此外，行人搜索需要处理大规模的图像数据，这对算法实时性能和硬件算力提出了更高要求。因此，如何提高算法的鲁棒性、快速性和准确性是行人搜索技术亟待解决的问题。本文将梳理近年来行人搜索方法的研究进展和技术思路，为该领域的发展及后续学者提供一些帮助。

文献 [4] 对 2014—2019 年的行人搜索方法相关文献资料进行了归纳总结，重点阐述了行人重识别的研究进展。文献 [5] 从深度特征表示学习、深度度量学习以及查询指导检测的角度对基于图像的行人搜索和基于文本的行人搜索进行了梳理分析。但未对重点行人搜索算法网络架构进行深入剖析。随着近年来行人搜索方向的发展，对该方向的研究进展进行全面系统的分析梳理、归纳总结和展望具有重要的意义。

本文从行人搜索方法采用的网络模型结构出发，将其分为基于卷积神经网络（Convolutional Neural Networks, CNN）和基于 Transformer 的行人搜索方法两大类进行阐述；并着重总结行人搜索采用的损失函数。

1　行人搜索算法模型结构

从计算机视觉角度看，行人搜索问题是跨多个摄像机的行人检索问题。一般情况下，行人搜索的实现流程如图 1 所示。

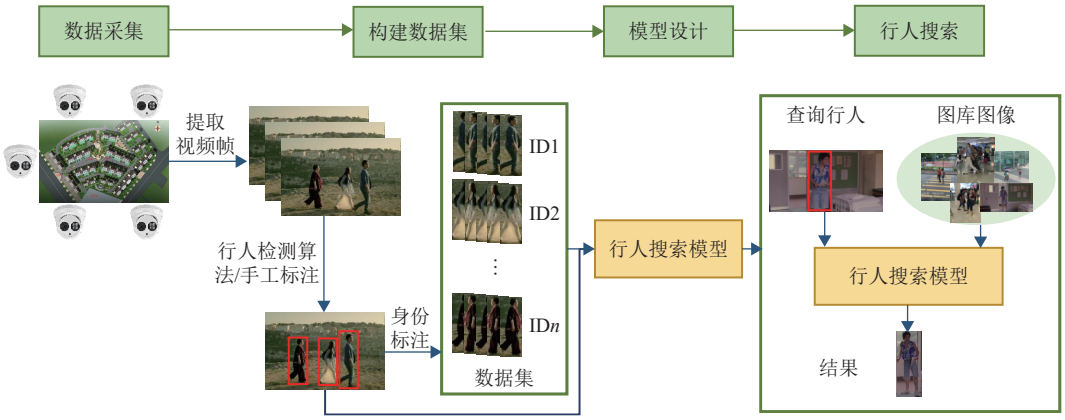

图 1　行人搜索流程框图

其主要分为 4 个步骤。

1）数据采集：从不同环境下的多个摄像机中获取原始图像或视频数据。

2）构建数据集：截取视频帧生成图像。采用行人检测算法或人工裁剪生成行人边界框。将出现在不同图像中的同一行人进行关联，赋予身份 ID。

3）模型设计：用标注的行人图像训练行人搜索模型。该模型的设计主要集中在网络结构及训练损失函数。网络结构主要用于提取行人区分性特征，损失函数则是优化网络参数和提升模型性能。

4）行人搜索：应用训练模型实现行人搜索。给定查询行人和图库图像，使用训练完成的行人搜索模型提取特征表示。对查询图像特征与图库图像特征进行匹配，得到行人搜索结果。

1.1　基于 CNN 的行人搜索方法

按照是否将行人搜索分为独立的行人检测和行人重识别，该类方法又可分为基于 Faster RCNN[6] 及基于行人检测和行人重识别两阶段的方法。基于 Faster RCNN 的行人搜索方法联合训练行人检测和行人重识别模型，给定一个未裁剪的图像，模型输出行人边界框坐标和相应的行人身份特征。该方法计算量小，速度快，但行人检测与行人重识别之间的联合优化存在矛盾。行人检测将背景与行人进行分类，学习的是所有行人的共性；而行人重识别区分不同的行人，学习的是不同行人之间的个性。基于两阶段的行人搜索方法独立训练行人检测和行人重识别。首先训练一个检测器从图像场景中检测行人边界框，其次基于边界框裁剪行人，用于训练 ReID 识别器进行目标行人匹配。该方法避免了任务联合间的矛盾，但独立训练两个网络产生了较高的计算成本，降低了网络推理速度。

1.1.1　基于 Faster RCNN 的行人搜索方法

文献 [7] 以 Faster RCNN 作为行人搜索的网络结构，提出在线实例匹配（Online Instance Matching, OIM）方法。如图 2 所示，该方法包含了主干网络（Backbone）、区域生成网络（Region Proposal

Network, RPN）、感兴趣区域（Region of Interest, RoI）池化层、识别网络、分类回归层。

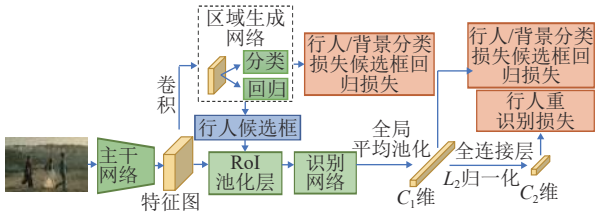

图 2 在线实例匹配网络结构示意图

主干网络：提取图像特征以生成特征图，通常选择 ResNet50[8] 中的 Conv1-Conv4 作为主干网络。

区域生成网络：初步获取候选框在原图的位置及类别。首先给定一个坐标位置的基础框，在该基础上进行面积的缩放及长宽比例的改变生成 N 个锚框，然后对这些锚框进行分类和回归，最终通过非极大抑制算法得到行人候选框。

RoI 池化层：将不同大小的候选框映射到特征图上，并生成固定大小的特征表示。

识别网络：进一步提取候选框特征。该部分通常为 ResNet50 中的 Conv5，然后经过全局平均池化，得到C_1维特征。

分类回归层：由于初步得到的候选框可能会存在错检、错位等情况，因此再次利用 Softmax 分类器和线性回归进行候选框的行人/背景分类和回归。同时将C_1维的特征向量投影到 L2 归一化的C_2维特征子空间判断行人身份。

后续学者受 OIM 启发，将基于 Faster RCNN 的网络结构进行改进，主要包括孪生网络、迁移学习、属性模块、特征表示改进、级联结构、弱监督/无监督这 6 类。

1） 孪生网络

孪生网络由两个对称的神经网络组成，具有相同的权重和架构。行人搜索中孪生网络两个分支的输入为同一行人的不同图像，如图 3 所示。文献 [9] 提出 Integration Net（I-Net），各分支均以 Faster RCNN 为基础，从两个全连接层提取的特征存储在含有一个正例对和多个负例对的在线存储库中进行识别。文献 [10] 文献改进 I-Net，每个分支均独立地进行检测任务，以获得更精确的候选框，然后再将特征融合进行重识别。

与上述孪生网络两个分支之间互相独立不同，文献 [11] 在孪生网络中引入查询行人信息，提出查询指导端到端行人搜索（Query-Guided End-to-End Person Search, QEEPS）。主干网络为查询指导孪生 SENet（Query-Guided Siamese Squeeze-and-Excitation Network, QSSE-Net），利用 SENet[12] 在残差块中连接查询行人特征和图库图像特征，计算权值向量并重加权每个通道的特征，扩展特征通道之间的交互。查询指导区域生成网络（Query-Guided Region Proposal Network, QRPN）包括查询指导的通道级注意力机制和标准的 RPN，采用 SENet 重加权图库图像的特征，并将其传递到 RPN 中，RPN 从重加权的图像特征中提取具有查询行人相似性得分的行人候选框。文献 [13] 将该方法扩展到 few-shot 细粒度分类任务，证明了该方法的有效性。但由于要对每个查询行人重新计算候选框，且每层网络结构都需要一对图像的相互作用，计算复杂度较高。

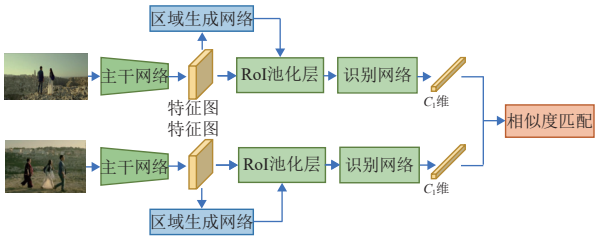

图 3 孪生网络结构示意图

文献 [14] 提出双向交互网络（Bi-Directional Interaction Network, BINet）。以场景图像和 RPN 生成的候选框经裁剪调整为固定大小后的行人图像作为孪生网络两个分支的输入，在裁剪行人的指导下关注场景中的行人，减轻冗余的上下文信息。

2） 迁移学习

迁移学习指从一个领域学习到的知识迁移到另一个领域。文献 [15] 提出 ReID 子任务主导的迁移学习（Subtask-Dominated Transfer Learning, STL）。将实例图像输入到 ReID 模型进行预训练，用重采样缓解源域和目标域的数据分布差异，利用预训练 ReID 的主干权值初始化行人搜索模型。有效改善行人搜索数据中不平衡的行人身份长尾分布对模型的影响。

文献 [16] 用与学生网络骨干相同的 Faster RCNN 行人检测器监督在线实例匹配模型，并利用预先计算好的特征查找表减轻行人重识别任务。文献 [17] 将重识别部分全连接层操作后得到的特征图和查询表中相应的类向量乘积作为类激活图，将其作为 ReID 模型中的知识监督行人搜索模型的训练。

行人搜索是行人重识别的扩展任务，有研究者

实验发现行人搜索的瓶颈在于 ReID 头，因此仅利用性能较好的 ReID 教师网络指导行人搜索模型，如图 4 所示。文献 [18] 在 ReID 头中分别采用概率感知、pair-wise 关系感知和 triplet-wise 关系感知 3 种知识蒸馏方法。文献 [19] 提出教师指导分离网络（Teacher-Guided Disentangling Networks, TDN），在 ReID 分支与教师网络间引入知识迁移桥缓解教师模型与 ReID 分支间的尺度差异。

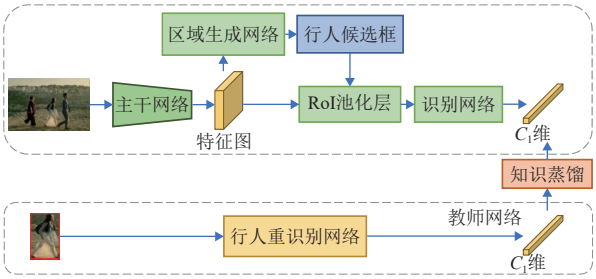

图 4　ReID 知识蒸馏网络结构示意图

3） 属性模块

现实场景中行人和背景存在错位、遮挡、背景杂乱等情况干扰行人搜索，引入行人身体部位特征、行人语义属性等可增强行人的特征表示以应对上述挑战。

文献 [20] 关注行人上半身、躯干、下半身及整体。在识别网络后，对输入的查询−图库对中检测出的行人进行全局平均池化和 part-based 操作。该模型虽然结合行人的部分特征并用图匹配法获得较好性能，但由于模型要在整个图像集中检测每个行人并与查询图像中行人进行比较，因此搜索速度慢且需要大量数据训练 part-based 模块。文献 [21] 提出行人身体感知网络（Body Perception of Person, BPNet），将行人划分为上半身、躯干和头肩 3 个矩形框，利用关键点引导的学习算法提取和融合这些区域的特征，有效解决身体空间错位问题。

文献 [22] 提出由检测、重识别和身体部分分类分支组成的端到端三分支网络，其中身体部分分类网络将裁剪后的图像水平平均分割成 4 个部分，分别代表不同的视觉结构。文献 [23] 引入行人语义属性，将行人语义分割为头部、上衣、下衣、鞋子和包 5 部分，提出多属性增强模型（Multi-Attribute Enhancement, MAE），利用属性特征融合模块对属性特征进行重加权，得到对齐且细粒度的特征。

遮挡可能会导致分割不准确而影响行人身体部位特征的失真，且检测的候选框只包含行人的部分身体，使全局特征退化为部分特征，而将部分特征与目标行人的全局特征进行匹配是不合理的。文献 [24] 提出部分对齐网络（Align-to-Part Network, APNet），补齐行人空缺部位或删减多余部位，再连接区域特征提取器，选取相互可见的部分特征进行相似度计算，丢弃遮挡或噪声区域的部分特征。

为使网络更多地关注前景信息，忽略背景偏差，文献 [25] 提出自下而上融合网络，在主干网络的 Conv3 处加入由 1×1 卷积和 Sigmoid 激活层组成的前景分割头，输出单通道的前景概率图，并将其与各层特征图乘积相加，生成前景感知特征图。文献 [26] 联合优化行人检测、重识别和行人分割进行行人搜索任务，利用分割掩膜引导特征提取网络学习丰富的前景信息。

4） 特征表示改进

为解决行人检测和行人重识别联合优化间的矛盾，文献 [27] 提出 Norm-Aware Embedding（NAE）模型，旨在平衡行人检测和行人重识别。如式 (1) 所示，将行人特征向量映射到极坐标中。

$$\boldsymbol{x} = r\cdot\boldsymbol{\theta} = \sqrt{\boldsymbol{x}_1^2+\boldsymbol{x}_2^2+\cdots+\boldsymbol{x}_{C_2}^2}\cdot\frac{\boldsymbol{x}}{\sqrt{\boldsymbol{x}_1^2+\boldsymbol{x}_2^2+\cdots+\boldsymbol{x}_{C_2}^2}} \tag{1}$$

$$\tilde{r} = \sigma\left(\frac{r-\mathbb{E}[r]}{\sqrt{\mathrm{Var}[r]+\varepsilon}}\gamma+\beta\right) \tag{2}$$

式中，$r\in[0,+\infty)$；$\boldsymbol{x}_i^2(i=1,2,\cdots,C_2)$为第$i$维特征向量的平方；$\boldsymbol{x}$为$C_2$维特征向量；$\tilde{r}\in[0,1]$；$\gamma$为缩放因子；$\beta$为偏移因子。

该模型用$\tilde{r}$区分行人和背景，接近 0 为背景，接近 1 为行人，用角度$\boldsymbol{\theta}$区分不同的行人，如图 5a 所示。NAE 中每个候选框的卷积特征通过全局平均池化分解成一个向量，丢失空间信息，因此进一步将 NAE 扩展到像素级—NAE+，降低错位对重识别的影响。如图 5b 所示，首先对$C_1\times k\times k$的特征图卷积得到$C_2\times k\times k$的特征图，然后将特征图每个位置的C_2维向量归一化并缩放为范数感知嵌入，同时获得每个位置的角度。每个位置的映射范数作为一个空间注意力，在匹配时校准每个像素的重要性。NAE 在没有添加参数的情况下达到较高的准确率，文献 [28-32] 在此基础上先后提出新方法。

只有特征以 0 为中心时，L2 归一化才有效。因此采用 BatchNorm 使特征分布在[0,1]之间，但

行人搜索存在长尾分布，使 BatchNorm 易偏向主导身份特征，削弱类间可分性。且 L2 归一化将所有元素归一化到相同尺度，可能会导致一些重要特征丢失或压缩。文献 [33] 提出原型归一化，用身份原型校准特征分布，为每个身份都分配相似的角度空间。

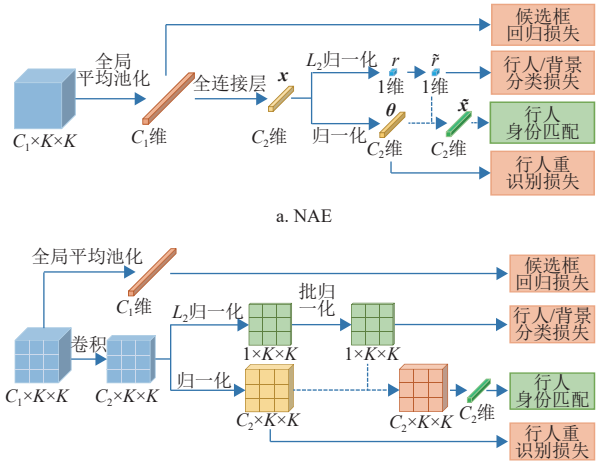

图 5 多任务头网络结构示意图

5） 级联结构

基于 Faster RCNN 的行人搜索方法提取的行人候选框质量较低，影响后续重识别的精度。文献 [29] 提出序列端到端网络（Sequential End-to-End Network, SeqNet），如图 6 所示。将检测与重识别看作是渐进的过程，在 NAE 基础上，添加额外的 Faster RCNN 头增强 RPN 提升候选框质量，序列处理两个子任务。

后续在此基础上，针对行人搜索中的问题进行相应改进。文献 [30] 利用行人同行者信息，提出全局−局部上下文网络（Global-Local Context Network, GLCNet），在特征图和最后经识别网络处理后的候选框特征后分别连接全局上下文编码器和局部组上下文编码器，分别生成 128 维特征，与 SeqNet 的特征聚合，输入到自注意力块，实现最终通道的特征交互。文献 [34] 提出注意感知关系混合器（Attention-Aware Relation Mixer, ARM），在 SeqNet 的 RoI 池化层后引入注意力感知关系融合网络，捕获 RoI 内不同局部区域间的关系，对抗行人外观变形和遮挡。文献 [35] 提出 SeqNeXt，将 SeqNet 的主干网络替换为 ConvNeXt[36]，在 RPN 后连接两个分支，第一个分支将生成的候选框输入到 RoI Align，第二个分支将特征图输入到识别网络中获得整幅图像的特征，最后将两个分支的特征输入到图库过滤网络，学习场景图像与行人的相似度分数。文献 [37] 提出场景上下文增强网络，在 SeqNet 上引入场景分支，将特征图输入到级联的通道注意力和空间注意力中，得到具有自我注意力模块的场景特征。文献 [38] 用 Inception 卷积替换 SeqNet 中 ResNet 的卷积，并在主干网络添加特征融合模块，动态增强特征图的感受野并获得多层次特征表示。

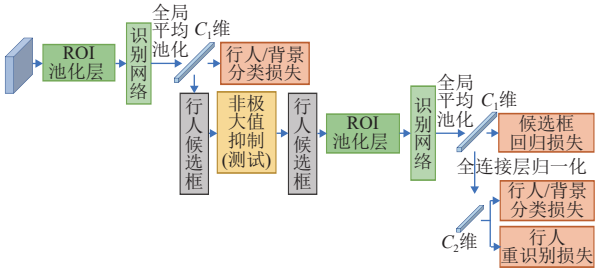

图 6 SeqNet 网络结构示意图

6） 弱监督/无监督

行人身份标注耗费大量的人力物力，因此研究者基于唯一性和共现性，即一张图像中不可能出现相同身份的行人及在同一图像中共同出现的多个行人很可能出现在其他图像中的原则，提出弱监督的行人搜索方法，如图 7 所示。

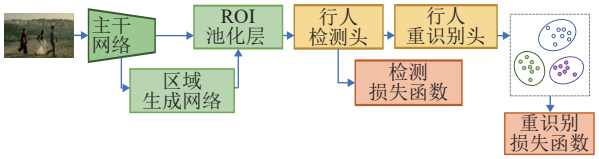

图 7 弱监督网络结构示意图

弱监督网络结构使用 Faster RCNN 对行人进行检测，将检测到的行人候选框输入到弱监督行人重识别进行处理。困难正样本和困难负样本会导致不同身份的行人之间具有较高的相似度，文献 [39] 提出基于唯一性的困难负样本挖掘方法和基于共现性的困难正样本挖掘方法。当同一幅图像中行人具有较高相似度时，则基于唯一性删除正样本集中的困难负样本；由于正样本较少，根据共现性不断迭代挖掘困难正样本，增加正样本对。文献 [40] 提出上下文指导的行人搜索（Context-Guided Person Search, CGPS），研究检测、记忆和场景的上下文信息，检测行人和背景信息。记忆上下文信息将困难负样本特征进行存储并评估其累积置信度，提出困难负样本抽样策略。场景上下文信息基于唯一性的约束在内存特征中生成聚类结果。

文献 [41] 将现有弱监督网络作为主分支，引入与主分支主干网络参数、结构相同的网络作为多尺度样本分支，提出自相似性驱动的尺度不变学习弱监督行人搜索。并基于同一行人在不同尺度上相似的自相似先验，设计多尺度样本分支，获得同一行人不同尺度的特征，指导主分支学习尺度不变特征。最后基于动态阈值获取实例伪标签。

研究者将对比学习引入行人搜索中，旨在通过学习行人之间的相似性和差异性，提高行人搜索的准确性和鲁棒性。文献 [42] 利用区域孪生网络（Region Siamese Networks, R-SiamNets）中的实例级一致性学习鼓励上下文不变表示，同时引入聚类级对比学习，将最近的实例聚合在一起，并将来自不同聚类的实例分开，将聚类生成的伪标签用于对比学习，迭代地应用非参数聚类。文献 [43] 针对行人搜索中图像尺度不一致、候选框位置不精确及遮挡问题，提出一种孪生结构的弱监督行人搜索方法—深度图像内对比学习（Deep Intra-Image Contrastive Learning, DICL）。该网络由搜索分支和实例分支构成。搜索分支输入场景图像利用 Faster RCNN 检测行人，并提取行人特征存储在记忆库中。实例分支提取行人真实框信息，与存储在记忆库中检测出的行人特征进行对比。空间不变对比学习来学习尺度不变特征和位置不变特征解决空间变化问题，遮挡不变对比学习来对搜索分支的行人候选框采用遮挡策略，并与没有遮挡的实例特征执行对比增强特征一致性，学习遮挡不变特征。

文献 [44] 提出无监督的行人搜索方法—域自适应行人搜索（Domain Adaptive Person Search, DAPS），将目标域图像输入到源域训练好的模型中，生成伪候选框，使用真实框和伪候选框进行域自适应训练，并使用动态聚类策略生成伪标签进行行人搜索。文献 [45] 使用大规模开放世界用户生成视频内容（UGC videos）训练行人搜索模型。在特征级别，学习域不变表示，引入特定域的批归一化和通道级身份相关特征去相关策略，提高模型对不同域的适应性。在数据级别，识别并解决 UGC 数据中的典型噪声源，通过软约束、伪标签生成和帧间增强策略提高模型泛化能力。文献 [46] 通过结合源域和目标域的表示来生成自适应混合域表示，设计域自适应混合器增强模型从标记源域到未标记目标域的知识迁移能力，鼓励域混合以最小化两个极端域之间的距离。文献 [47] 使用自动标注的非真实数据集进行训练，并泛化到任意未见过的真实数据集。为了缩小两个数据集间的域差距，估算每个行人实例的保真度，利用这些信息来适应性地训练端到端网络。

1.1.2　基于两阶段的行人搜索方法

基于两阶段的行人搜索方法级联行人检测器和 ReID 特征提取器，并单独训练。行人检测器获取行人候选框，然后裁剪并调整到固定的分辨率，最后输入到 ReID 模型中进行识别。

行人检测作为行人搜索的子任务，主要目的是在给定图像中准确地检测出行人位置。常用行人检测方法主要分为单阶段行人检测和两阶段行人检测。单阶段行人检测方法如 SSD[48]、YOLO 系列[49-52] 等直接在图像中预测行人位置和类别，两阶段行人检测方法如 R-CNN[53]、Faster RCNN[6] 等首先生成候选框，然后再对候选框进行分类和定位。

文献 [54] 首次将 3 种行人检测器与 4 种行人重识别方法分别组合，为后续研究人员提供单独训练行人检测和行人重识别的研究思路。现介绍基于两阶段的行人搜索技术中典型的方法。

文献 [55] 提出掩码指导的双流卷积神经网络（Mask-Guided Two-Steam CNN Model, MGTS），如图 8 所示。Faster RCNN 作为行人检测器，独立建模前景行人和原始图像。在行人重识别部分，带有背景的行人图像和前景图像分别经过两个结构相同但参数不共享的 ResNet50 的 Conv1-Conv5 分支，将两条路径的特征连接并由 SEBlock 重加权，最终将特征投影到 L2 归一化的子空间进行重识别。

针对行人搜索中多尺度问题，文献 [56] 在重识别部分提出跨层次语义对齐（Cross-Level Semantic Alignment, CLSA）。行人检测器为 Faster RCNN，检测出的行人候选框输入到识别网络，在前向传播过程中计算行人边界框的 CLSA 特征，分别在 Conv3、Conv4、Conv5 执行平均池化、全连接层、批归一化，ReLU 激活函数得到低级、中级和高级的语义信息。然后将所有金字塔层次的特征向量连接起来，作为行人重识别的最终表示。文献 [57] 提出实例指导候选框网络（Instance Guided Proposal Network, IGPN），检测阶段采用孪生区域建议网络，引入查询行人信息，学习查询行人与候选框之间的相似性，根据相似度得分来减少候选框。重识别部分分别设计局部关系块实现查询行人信息和同一场景中候选框对间的关系，输出每个候选框的相似度得分；全局关系块用于描述查询信息

与场景之间的关系。文献 [58] 也将查询行人引入网络中，提出 Task-Consist Two-Stage（TCTS）方法，基于 Faster RCNN 设计检测器，学习一个辅助的身份分支计算候选框的查询相似度得分，候选框输入到检测分支和身份分支，为重识别阶段生成更准确的类查询候选框和更少的非查询候选框。RPN 生成的候选框输入到识别器，预测检测到的行人身份标签，并利用这些行人身份构建一个更实际的混合训练集，提高重识别阶段对不准确检测的鲁棒性。

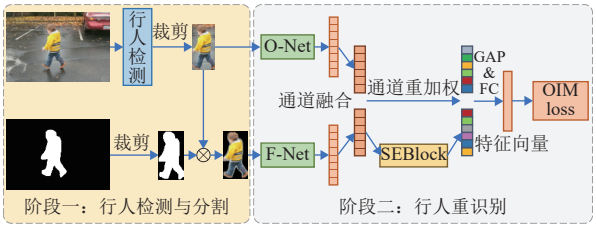

图 8 掩码指导的双流卷积神经网络结构示意图

单独训练行人检测和重识别两个子任务耗时较长，为平衡运行时间和性能表现，文献 [59] 将 SSD 作为行人检测器，在行人重识别部分，从场景图像中裁剪的行人图像连接空间学习模型，将全局特征图分成 6 部分，然后输入到全局平均池化层得到空间特征。文献 [60] 提出实时行人搜索的多任务联合框架。将 YOLOv5 集成 Ghost 和 SE 模块用于行人检测。重识别部分设计含有 ResNet18、ResNet34、ResNet50 的模型自适应体系结构，根据不同的人数选择不同的网络，平衡准确率与速度。

文献 [61] 设计样本增强和实例敏感框架（Sample-Enhanced and Instance-Sensitive, SEIE），关注行人搜索数据集中样本特异性对训练细粒度 ReID 模型的影响。在检测阶段，YOLOv5 作为检测器，提出样本增强组合提高候选框的质量和数量。在重识别阶段，以 OSNet[62] 为主干，提取行人特征。

场景图像中有很多未配对的行人，直接将其作为独立身份处理会产生长尾效应，而完全丢弃会导致严重的信息丢失。文献 [63] 引入上下文引导聚类算法（Context-Guided Cluster, CGC）和非配对辅助记忆单元（Unpaired-Assisted Memory, UAM），提出 CGUA（Context-Guided and Unpaired-Assisted）。将场景图像和行人特征输入到顺序连接的 CGC 和 UAM。CGC 采用混合相似度充分利用场景图像中的上下文信息，在聚类中增加额外的约束条件过滤聚类结果，UAM 旨在利用场景图像中大量未配对行人，将其特征存储在未配对记忆库中，帮助重识别模型学习更多区分性特征远离配对行人特征。

基于 CNN 的行人搜索代表性方法如表 1 所示，对各方法的优缺点进行了评述。

表 1 基于 CNN 的行人搜索代表性方法总结

方法分类	方法名称	年份/年	优缺点
孪生网络	I-Net[9]	2018	推理阶段需要进行两次前向传播，计算复杂度和推理时间增加
	QEEPS[11]	2019	开创性引入查询行人信息，识别精度提高。但对每个查询行人都要重新计算候选框及查询行人重要性，参数增多，计算复杂度较高，难以应用到实际中。查询行人与图库行人视觉差别较大时，查询行人的指导一定程度上会失效且存在误检、漏检
迁移学习/知识蒸馏	BINet[14]	2020	准确性受图库图像中裁剪行人准确性的制衡
	STL[15]	2021	增强模型泛化能力，但子任务设计不当会导致迁移学习效果不佳
	DKD[18]	2021	模型的准确性和鲁棒性受限于预训练的分类器和类别标签
属性模块	CTXG[20]	2019	适应各种行人搜索场景，具备一定扩展性，但计算复杂度高
	BPNet[21]	2020	结合行人检测、实例分割、关键点检测和重识别实现行人搜索，解决背景噪声和特征错位
	APNet[24]	2020	有效解决局部遮挡，匹配效率提高。但极端遮挡或严重变形的情况导致部分匹配方法失效
	AGWF[22]	2021	自适应梯度传播平衡子任务优化，三分支网络利用不同任务的相关性和互补性，提高行人搜索准确性。但同时增加模型复杂度，需要更多计算资源和时间
特征表示改进	NAE[27]	2020	方法实现简单，可与现有其他行人搜索框架相容。但范数感知会损失细节信息，且倾向主导身份特征，类间可分性减弱
	OIMNet++[33]	2022	平衡类间特征表示，紧致类内特征，减少身份之间信息混淆
级联结构	SeqNet[29]	2021	级联两个Faster RCNN头，生成更准确的候选框，实现简单且实用
	GLCNet[30]	2021	结合全局上下文信息和局部上下文信息，提高搜索准确性。但网络结构复杂，训练时间长
	SeqNeXt[35]	2023	快速过滤与目标行人不匹配的图像，提高搜索效率，但若目标行人和候选行人的相似性小或相似行人干扰，会导致过滤结果不准确，产生漏检和误检
无监督/弱监督	R-SiamNets[42]	2021	更好地捕捉行人局部特征，提高区分度。不同图像的区域对齐对孪生网络发起挑战，影响特征匹配，且增加计算复杂度和存储需求

续表

方法分类	方法名称	年份/年	优缺点
无监督/弱监督	CGPS[40]	2022	降低数据标注工作量和成本，但受到标注不准确和缺失的限制
	DAPS[44]	2022	适应不同领域之间的数据分布差异，泛化能力强。减少目标领域标注成本和工作量。源域和目标域行人标签的差异性导致域自适应过程中标签不一致或错匹配情况。领域间的数据对齐和特征转换增加计算资源和时间成本
基于两阶段的行人搜索方法	MGTS[55]	2018	特征信息丰富，避免背景噪声，提高搜索性能。但分割准确性限制算法的准确性
	CLSA[56]	2018	适应不同尺度行人搜索，但准确率较低
	IGPN+PCB[57]	2020	候选框更精确，搜索范围缩小。但IGPN基于实例引导，若存在多个实例，IGPN效率低下
	TCTS[58]	2020	定位识别更精确，但独立检测和识别导致计算量大

1.2　基于 Transformer 的行人搜索方法

Transformer 利用注意力机制建模特征间的依赖关系，同时考虑全局信息和局部信息，在目标检测、图像分割等计算机视觉任务中取得了显著成果[64-66]。行人搜索中引入 Transformer，能够提高行人搜索的准确率和鲁棒性。

文献 [67] 提出注意力上下文感知嵌入（Attention Context-Aware Embedding, ACAE），采用 DETR 模型[68] 检测行人，顺序连接 ResNet50 特征提取器及带有 Transformer 解码器的注意力上下文感知嵌入头，从图像内和图像间的单个特征中聚合上下文信息。文献 [69] 提出 PSTR，将行人搜索视为序列预测问题，检测图像中所有行人并提取特征。在特征图后引入含检测编码器–解码器和 ReID 解码器的 PSS 模块，检测编码器–解码器由 3 个级联的编码器和解码器构成，ReID 解码器用多层监督与共享解码器提取行人 ReID 特征，并引入部分注意力块捕获行人部位关系。为解决现有方法忽略图像上下文信息的问题，文献 [70] 采用 Swin Transformer 作为主干网络，自适应学习图像全局上下文，局部感知增强模块使用 Swin Sparse Transformer 专注于学习与行人相关的局部上下文信息，增强模型对行人姿态变化的适应性。

为解决行人尺度、姿势变化和遮挡导致的性能下降问题，文献 [32, 71] 基于 Cascade RCNN，引入 Transformer 模块处理每个阶段的尺度变化，如图 9 所示。具体地，ResNet50 作为主干网络，用 RPN 生成候选框，对每个候选框应用 RoI Align 操作，得到候选框特征图，最后进行三级级联操作获得行人位置及身份信息。文献 [71] 提出级联遮挡注意力 Transformer（Cascade Occluded Attention Transformer, COAT）。对每个阶段候选框特征图的通道切片处理，并交换不同候选框的通道级特征，模拟现实中的遮挡情况。然后通过 3 个独立的全连接层转换为 query 矩阵、key 矩阵和 value 矩阵，连接多头自注意力和前馈网络组成遮挡注意力机制，最后将 n 个尺度的输出连接到原始特征图，进行下一阶段的行人检测和重识别。在 COAT 基础上，文献 [32] 在卷积 Transformer 模块中引入由一个卷积编码块、两个层归一化、两个线性层、尺度调节器和 MLP 组成的尺度增强 Transformer，自适应调整不同尺度特征的重要性权重以捕捉不同尺度的行人信息。

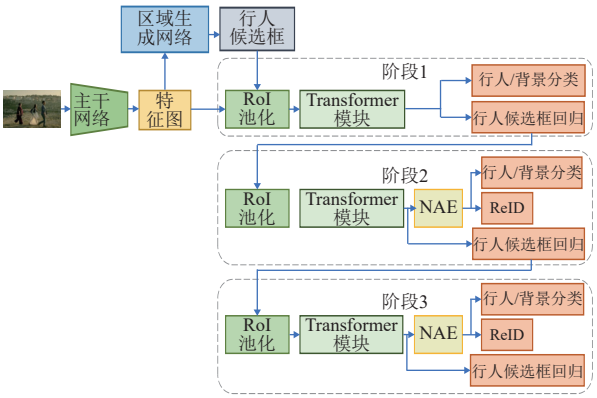

图 9　级联 Transformer 网络结构示意图

为缓解行人检测和重识别间优化的矛盾，文献 [72] 提出 Sequential Transformer（SeqTR）。用检测 Transformer 和重识别 Transformer 顺序实现行人搜索。主干网络 ResNet50 的 Conv3-Conv5 提取图像特征，检测 Transformer 选择可变形 DETR[73] 预测行人候选框。重识别 Transformer 由一个自注意力层和 K 个交叉注意力层组成，其中自注意力层包括多头注意力模块和层归一化，交叉注意力层包含一个可变形的注意力模块和层归一化学习行人特征嵌入。

除以上基于 Faster RCNN、基于两阶段的行人搜索方法及基于 Transformer 的行人搜索方法外，还有一些新颖的行人搜索方法。

根据人类神经系统对复杂视觉信息处理原理即当人看到并记住目标行人外观后，再次寻找该行人

时，通常会逐步缩小搜索区域，并在小范围内与记忆的目标行人进行细节匹配，文献 [74] 提出神经行人搜索机制（Neural Person Search Machines, NPSM），其包含初始记忆组件和神经搜索网络。初始记忆组件提取目标行人特征指导神经搜索网络在场景图像中递归地搜索行人。神经搜索网络运用卷积替换 LSTM 中全连接的乘法运算，在每次递归时选择性地聚焦与查询行人特征相似的区域，丢弃不相关区域。递归过程中，模型将置信度最大的作为最终搜索结果。

文献 [75] 提出无锚框行人搜索方法——特征对齐行人搜索（Feature-Aligned Person Search Network, AlignPS）网络，降低计算开销。该模型基于 FCOS[76]，对齐特征聚合模块利用可变形卷积和特征融合重构 FPN，克服重构特征的多尺度、区域不对准问题。对齐特征聚合模块输出的特征经 3×3 可分离卷积生成 ReID 特征，利用 FCOS 检测头进行检测和重识别。

由于检测和重识别两个子任务的优化目标会相互干扰，因此研究人员在主干网络后解耦检测和重识别任务。文献 [77] 提出解耦和记忆增强网络（Decoupled and Memory-Reinforced Network, DMRNet），检测分支利用全连接层进行行人/背景分类和边界框回归。重识别分支采用真实框提取 RoI 特征用于重识别训练，减少检测和重识别间的依赖。之后又改进 DMRNet，提出了 DMRNet++[78]，通过在网络中加入基于点的空间采样，产生细粒度特征。文献 [79] 提出视觉共享和表征独立（Vision Shared and Representation Isolated, VSRI）解耦两个子任务，提出多层特征融合（Multi-Level Feature Fusion, MLFF）模块，从主干网络获取不同特征。采用多头自注意力机制构建注意力驱动提取模块，采用级联卷积单元设计特征分解和级联集成模块，便于 MLFF 获得更多行人识别表示。

文献 [80] 将扩散模型应用到行人搜索研究中，提出 PSDiff（Person Search Framework Based on the Diffusion Model），将扩散过程设计为正向过程和逆向过程两个阶段。在正向过程中模型通过逐步添加噪声将清晰的行人特征向噪声空间推进。在逆向过程中通过学习到的去噪网络逐步引导噪声数据恢复到原始的、清晰的行人特征。从噪声到清晰的迭代恢复过程，提高了对检测误差的鲁棒性，增强了模型对行人身份的判别能力。文献 [81] 为行人搜索设计混合预训练框架，利用丰富的行人检测和行人重识别的两个子任务数据集进行行人搜索的全任务预训练，提出混合学习范式处理具有不同监督类型的数据，并通过内部任务对齐模块减少有限资源下的域差异，提高预训练模型的泛化能力。

基于 Transformer 的行人搜索和其他典型方法如表 2 所示，总结了各方法的优缺点。

表 2　基于 Transformer 的行人搜索代表性方法总结

方法分类	方法名称	年份/年	优缺点
基于Transformer的行人搜索方法	PSTR[69]	2022	获取全局和局部图像特征，但计算复杂性较高，不适用于实时应用
	COAT[71]	2022	多阶段逐步优化检测和识别，减少误检和误识别。但每个阶段都需要进行自注意力计算，增加模型的计算复杂度，级联多个检测器，对小尺度行人的检测效果较差，且训练时间和推理时间较长
基于LSTM的行人搜索方法	NPSM[74]	2017	递归搜索提高模型的计算复杂度，搜索时间长
无锚框行人搜索	AlignPS[75]	2021	无须预定义锚框，训练过程简化，但对尺度变化大的行人存在定位不准确或漏检，对于遮挡问题可能会错误合并多个行人
任务解耦的行人搜索	DMRNet[77]	2021	独立优化子任务，搜索效果提升，但计算量增大
	VSRI[79]	2022	保持任务独立性，提高泛化能力，但独立表示学习面临外观变化大的挑战

1.3　损失函数设计

行人搜索中的损失函数为行人检测损失函数和行人重识别损失函数。其中，行人检测损失函数为行人/背景分类损失和行人候选框回归损失，行人重识别损失函数用于识别不同的行人身份，拉近同一身份的行人特征，远离不同身份的行人特征。

1.3.1　行人检测损失函数

行人检测中，行人/背景分类损失函数的目的是区分图像中的行人和背景。在行人搜索中该部分损失函数大多数使用交叉熵损失函数，文献 [28, 82] 使用焦点损失[83] 作为分类损失。交叉熵损失函数在二分类中，对每个类别预测到的概率为：

$$L=\frac{1}{N}\sum_{i}-[y_i\log p_i+(1-y_i)\log(1-p_i)] \tag{3}$$

式中，N为类别数；y_i为样本i的类别标签；p_i为样

本i预测为行人的概率。

行人候选框回归损失函数用来最小化行人候选框和真实框的差值，将候选框精确定位到图像中的行人，使用 Smooth L1 损失函数，如：

$$\text{smooth}_{\text{L1}}(x)=\begin{cases}0.5x^2 & |x|<1\\ |x|-0.5 & \text{其他}\end{cases} \tag{4}$$

式中，x为候选框与真实框间的差值。

Smooth L1 损失函数的优点是当误差较小时，梯度接近于 MSE 损失函数的梯度，加快了模型收敛速度；当误差较大时，其梯度与 MSE 损失函数的梯度相比更小，增强了模型的鲁棒性。

1.3.2　行人重识别损失函数

行人重识别损失函数对图像中的行人进行身份分类，以便能够在不同的图像中匹配同一身份的行人。在行人搜索中，常用的行人重识别损失函数为在线实例匹配（Online Instance Matching, OIM）损失函数、传统损失函数以及混合损失函数。

1） OIM 损失函数及改进

OIM 损失函数是行人搜索中广泛使用的一种损失函数。文献 [7] 充分利用标记行人身份和未标记行人身份提出了 OIM 损失函数。

在行人检测任务中，会产生标记身份、未标记身份和背景杂波 3 种候选框。假设在训练集中有 L 个不同的目标行人，当检测出的候选框能够与目标行人进行匹配时，称其为标记身份实例，并为其分配一个类别 id；无法与目标行人进行匹配的行人候选框则为未标记身份实例；其余被检测出的候选框则为背景杂波。OIM 损失函数只考虑标记身份候选框和未标记身份候选框。

为最小化同一行人之间的特征差异，同时最大化不同行人之间的差异，需存储记忆所有行人的特征。用$\boldsymbol{x}\in\mathbb{R}^D$表示一个批次内标记身份的特征，其中，$D$是特征维度，建立查询表（Lookup Table, LUT）$\boldsymbol{V}\in\mathbb{R}^{D\times L}$来存储所有标记身份的特征。前向传播时，通过$\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}$计算小批样本与所有标记身份之间的余弦相似度。反向传播时，若目标行人的 id 为t，则通过式 (5) 更新 LUT 的第t列：

$$\boldsymbol{v}_t=\gamma\boldsymbol{v}_t+(1-\gamma)\boldsymbol{x} \tag{5}$$

式中，$\gamma\in[0,1]$表示动量更新参数；$\boldsymbol{v}_t$为单位 L2 范数。

建立循环队列（Circular Queue, CQ）来存储最近小批量处理中出现的未标记身份的特征。用$\boldsymbol{U}\in\mathbb{R}^{\text{DM}}$表示循环队列中的特征，$Q$为队列大小，用$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x}$计算其与小批样本的余弦相似度。每次迭代之后，将新的特征向量推入队列中，同时弹出过时的特征向量以保持队列大小不变。

基于两种数据，定义$\boldsymbol{x}$被识别为类i的概率为：

$$p_i=\frac{\exp(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{x}/\tau)}{\sum_{j=1}^{L}\exp(\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{x}/\tau)+\sum_{k=1}^{Q}\exp(\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{x}/\tau)} \tag{6}$$

式中，$\boldsymbol{v}_i^{\mathrm{T}}$为 LUT 中第$i$个标记身份的行人特征；$\boldsymbol{v}_j^{\mathrm{T}}$为 LUT 中第$j$个标记身份的行人特征；$\boldsymbol{u}_k^{\mathrm{T}}$为 CQ 中第$k$个未标记身份的行人特征；$\tau$为调节平衡程度的温度参数。

同样地，在循环队列中被识别为第i个未标记身份的概率为：

$$q_i=\frac{\exp(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{x}/\tau)}{\sum_{j=1}^{L}\exp(\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{x}/\tau)+\sum_{k=1}^{Q}\exp(\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{x}/\tau)} \tag{7}$$

式中，$\boldsymbol{u}_i^{\mathrm{T}}$为 LUT 中第$i$个未标记身份行人特征。

OIM 最大化目标行人的期望对数似然：

$$L=\boldsymbol{E}_x[\log p_t] \tag{8}$$

OIM 损失函数将小批样本与所有标记身份和未标记身份的行人特征进行了比较，驱使特征向量与目标特征向量相似，远离其他特征向量。文献 [84] 认为在 NMS 之后，图像中仍有很多被视为有标记身份的候选框与真实框有不同程度的重叠，其中部分候选框带有背景信息，使用这些更新 LUT 可能会破坏合并的特征。因此只合并真实框特征，确保 LUT 的鲁棒性。由于行人身份数量大，但每个标记身份的实例却很少，仅通过标记身份学习行人特征具有一定难度。文献 [85] 将文献 [86] 的思想引入 OIM 损失函数中，改善网络在不同训练迭代中更新 LUT 和 CQ 对特征学习的影响。文献 [87] 提出噪声抑制 OIM（Noise-Suppression OIM, NR-OIM）损失函数，遏制人工标注过程中引入的人为噪声。在第一轮训练中使用 OIM 损失函数发现训练集中潜在的错误标注行人，第二次训练时使用 NR-OIM 抑制噪声。下面重点介绍对 OIM 改进较大的方法。

文献 [88] 利用未标记身份信息增强特征，提出实例增强损失函数（Instance Enhancing Loss, IEL）。部分未标记身份与标记身份外观相似，计算每个未标记身份特征与标记身份特征中心之间的距离，选

择性地将未标记身份标注为与标记身份相同的身份，利用这些未标记身份与标记身份训练行人搜索网络。但是该方法将困难负样本认定为正样本，会使网络难以区分与查询行人相似的其他行人。

受 IEL 启发及现有局限性，文献 [89] 提出基于动态模拟的在线实例匹配（Dynamic Imposter Based Online Instance Matching, DI-OIM），根据同一张图像中行人具有不同身份的原则，为未标记行人提供动态伪标记，区分不同标记身份行人，且使未标记身份行人彼此远离。不同图像包含不同的行人，每次迭代的伪标记行人特征都是动态变化的。

OIM 损失将同一身份的行人特征集成形成该身份行人的原型。考虑新输入的特征可能会存在遮挡或背景杂波等噪声，干扰原型的优化，减少原型之间的类间差异。文献 [90] 提出原型指导动量更新的 OIM 损失函数，将动量定义为目标和正原型对、目标和最难负原型对之间余弦相似性的比率。当目标接近最难的负原型时，会给目标特征分配低动量，以防止原型接近最难的负特征。

RPN 会生成错位的行人候选框，而 OIM 损失函数认为所有正样本候选框都有助于学习行人特征，并以固定的动量更新 LUT，但并不是所有候选框都能相同程度地促进特征学习。因此，文献 [33] 提出定位感知在线实例匹配（Localization-Aware Online Instance Matching, LOIM）损失函数，为每个候选框分配动量值。根据 IoU 值为候选框分配动量，即γ= 1−IoU自适应的更新 LUT 中的每个特征。

2） 传统损失函数

除 OIM 损失函数外，有少数研究人员使用传统损失函数进行行人搜索。为缓解同一行人的不同外观变化的影响，文献 [91] 引入中心损失，增强特征类内紧致性。中心损失跟踪所有类的特征中心，并且这些特征中心根据最近观察到的类样本不断更新。文献 [92] 将 triplet loss 作为行人搜索中识别行人身份的损失函数。因行人搜索中输入图像较大，导致 batch 的大小受到显存的限制，近年来，TripHard loss 被广泛应用于行人重识别[93-94] 中，但是在行人搜索任务中，只有少数高分辨率图像可以输入到网络中进行训练，检测到的候选框不够多，且被检测到的行人往往没有正对，因此文献 [95] 提出 proxy triplet loss，建立 proxy table 存储由锚框、正样本、负样本组成的 triplets。文献 [96] 将 triplet loss 与中心损失结合可以较好地监督行人重识别。

3） 混合损失函数

部分研究者将 OIM 损失函数与其他损失函数相结合，文献 [97] 提出 Improved Online Instance Matching（IOIM）损失函数，使用不同权重系数加强标记身份特征的分布和软化未标记身份的分布，加大标记身份的差异，最小化未标记身份的差异，并将 IOIM 与中心损失结合进一步最小化识别特征的类内距离。文献 [98] 提出在线实例聚合匹配（Online Instance Aggregation Matching, OIAM）损失函数，结合 OIM 损失和中心损失，有效解决类别多但同类样本少的问题。同时，利用中心损失辅助训练，跟踪所有类的特征中心，可以减少类内差异，减轻过拟合。文献 [99] 将 OIM 损失函数与 Softmax 损失函数相结合，提出多重损失函数训练网络。文献 [100] 提出分层在线实例匹配（Hierarchical Online Instance Matching, HOIM）损失函数，由二元交叉熵损失函数和 OIM 损失函数构成，利用检测和重识别之间的层次关系来指导网络学习。对无标记身份的行人选择性记忆刷新，只有当一个新的特征嵌入的重要性权重大于现有特征嵌入的最小值时，新的特征嵌入才会进入缓冲区，同时弹出重要性小的特征嵌入。文献 [101] 认为 OIM 损失函数对正样本和负样本的关注不平衡，通过将 OIM 损失函数与 triplet loss 结合，挖掘困难负样本弥补 OIM 损失的不平衡问题。

2 数据集及评估

2.1 数据集

目前行人搜索数据集有 CUHK-SYSU[7]、PRW[54]、LSPS[24] 和 PSM[102]。表 3 总结了行人搜索数据集的统计信息，从表中可以看出：1）数据集的规模在增加。深度学习方法得益于更多的训练样本，但同时也增加了标注难度。2）摄像机数量的增加。更加贴近现实场景，接近大规模相机网络，也给模型的泛化能力带来挑战。

表 3 常用行人搜索数据集及其参数

数据集	参数				
	帧数/帧	人数/人	边界框数/个	相机数/个	检测器
CUHK-SYSU	18 184	8 432	96 143	—	手工
PRW	11 816	932	34 304	6	手工
LSPS	51 836	4 067	60 433	17	Faster R-CNN
PSW	80 983	9 415	107 865	—	手工/Faster RCNN

CUHK-SYSU 数据集：来源于城市街道场景和

包含行人的电影片段。共 18 184 张图像，96 143 个行人边界框，8 432 个行人身份。将数据集划分为训练集和测试集，训练集包含 11 206 张图像和 5 532 个身份，测试集包含 6 978 张图像和 2 900 个身份。

PRW 数据集：由 6 台摄像机采集的 10 h 视频，人工标注 11 816 帧，共 43 110 个行人边界框，34 034 个行人的身份 ID 范围为 0～932，其余的均标为-2，表示不确定的行人。训练集共有 5 704 张图像和 482 个身份，测试集包含 6 112 张图像和 450 个身份。

LSPS 数据集：由部署在室内和室外的 17 个摄像机采集，利用 Faster RCNN 对边界框进行检测，共收集 51 836 帧，60 433 个行人边界框和 4 067 个身份。

PSM 数据集：涵盖从亚洲到欧洲的超过 8 个国家的 1 236 部电影，场景包含白天和夜晚。该数据集在标注行人整体的同时也标注了行人的头部。共包含 80 983 帧，9 415 个身份，手动标注了 107 865 个人体边界框和 107 865 个头部边界框。

2.2　评估指标

累积匹配特征（CMC Top-k）和平均精度均值（mAP）是行人搜索的主要评价指标。行人搜索方法中多使用 Top-1 作为评价指标，即每次模型分类输出的 N 个相加为 1 的结果中概率最大的种类预测正确的概率。

2.3　性能分析

按照分类标准，表 4～表 6 归纳汇总了典型行人搜索方法在基准数据集上的性能表现。从网络结构及损失函数对方法性能进行分析。表 4 中基于两阶段的行人搜索的方法类别分别表示行人检测器和重识别器。

表 4　典型行人搜索方法在基准数据集上的实验结果（基于 Faster RCNN 的行人搜索方法）

方法	年份/年	方法类别	重识别损失函数	CUHK-SYSU		PRW	
				mAP/%	top-1	mAP/%	top-1
OIM[7]	2017	Faster RCNN（Baseline）	OIM	75.50	78.70	21.30	49.90
HOIM[100]	2020	Faster RCNN	HOIM	89.70	90.80	39.80	80.40
文献[82]	2022	Faster RCNN	焦点损失	94.80	95.50	47.6	81.6
QEEPS[11]	2019	孪生网络	OIM	84.40	84.40	37.10	76.70
BINet[14]	2020	孪生网络	OIM	90.00	90.70	45.30	81.70
KD-QEEPS[16]	2019	知识蒸馏	OIM	85.00	85.50	-	-
STL[15]	2021	迁移学习	交叉熵损失	93.10	94.20	61.60	90.50
TDN[19]	2022	知识蒸馏	交叉熵损失	94.90	96.30	**70.20**	**93.50**
CTXG[20]	2019	属性模块	OIM	84.10	86.50	33.40	73.60
APNet[24]	2020	属性模块	OIM	88.90	89.30	41.90	81.40
BUFF[25]	2020	属性模块	OIM	90.60	91.60	42.20	81.00
AGWF[22]	2021	属性模块	OIM	93.30	94.20	53.30	87.70
NAE+[27]	2020	特征表示改进	OIM	92.10	92.90	44.00	81.10
OIMNet++[33]	2022	特征表示改进	LOIM	93.10	94.10	47.70	84.80
SeqNet[29]	2021	级联结构	OIM	94.80	95.70	47.60	87.60
PS-ARM[34]	2022	级联结构	OIM	95.20	96.10	52.60	88.10
SeqNeXt[35]	2022	级联结构	OIM	**96.40**	**97.00**	58.30	92.40
R-SiamNets[42]	2021	弱监督	实例一致损失，实例间相似性一致损失	86.19	87.59	25.53	76.03
CGPS[40]	2022	弱监督	OIM	80.00	82.30	16.20	68.00
DAPS[44]	2022	无监督	OIM	77.60	79.60	34.70	90.90
DDAM-PS[46]	2024	无监督	Bridge loss&disparity loss	79.50	81.30	36.70	81.20

表 5　典型行人搜索方法在基准数据集上的实验结果（基于两阶段的行人搜索方法）

方法	年份/年	方法类别	重识别损失函数	CUHK-SYSU		PRW	
				mAP/%	top-1	mAP/%	top-1
MGTS[55]	2018	Faster RCNN, FCIS	OIM	83.00	83.70	32.60	72.10
CLSA[56]	2018	Faster RCNN, ResNet50	交叉熵损失	87.20	88.50	38.70	65.00
IGPN+PCB[57]	2020	Siamese RPN, ResNet50	—	90.30	91.40	**47.20**	87.00
TCTS[58]	2020	Faster RCNN, ResNet50	IDGQ损失	**93.90**	**95.10**	46.80	**87.50**
CGUA[63]	2022	Faster RCNN, ResNet50	PCC&UCC	91.00	92.20	42.70	86.90

表 6 典型行人搜索方法在基准数据集上的实验结果（基于 Transformer 的行人搜索方法&其他方法）

方法	年份/年	方法类别	重识别损失函数	CUHK-SYSU		PRW	
				mAP/%	top-1	mAP/%	top-1
PSTR[69]	2022	Transformer	OIM	95.20	96.20	56.50	**89.70**
COAT[71]	2022	Transformer	OIM	94.80	95.20	54.00	89.10
NPSM[74]	2017	Conv-LSTM	Softmax	77.90	81.20	24.20	53.10
AlignPS[75]	2021	无锚框	TOIM	94.00	94.50	46.10	82.10
VSRI[79]	2022	任务解耦	OIM	94.10	94.70	54.70	89.10
DMRNet++[78]	2023	任务解耦	UCL	94.40	95.70	51.00	86.80
PSDiff[80]	2023	扩散模型	L2 loss	**95.70**	**96.30**	**57.10**	88.10

从网络结构上看，行人搜索任务的解决方法呈多样化趋势发展。最常用的行人搜索方法以 Faster RCNN 为网络结构。基于两阶段的行人搜索方法由于计算量大、识别速度慢、准确率低，逐渐淡出研究人员的关注。随着 Transformer 在计算机视觉领域的成功应用，基于 Transformer 的行人搜索方法开始发展并取得良好的性能。除此之外，有学者提出如任务解耦、扩散模型的行人搜索方法并取得了不错的效果。

由表 4～表 6 可知，对基于 Faster RCNN 的行人搜索方法的改进，识别准确率均得到不同程度的提升。初期，通过引入孪生网络用查询行人信息指导行人搜索，提升搜索准确率，但计算复杂度较大。知识蒸馏技术的应用在提升算法性能的同时又减少了模型的计算负担和训练时间。文献 [19] 在 PRW 数据集上展现了卓越的性能。属性模块旨在提取行人的局部特征和语义属性，弥补行人遮挡、错位等带来的信息丢失。文献 [27] 解决了行人检测和重识别优化间的矛盾，且该方法可以融合到目前已有的基于 CNN 的行人搜索方法中，实现简单、适配性高。级联结构在 NAE 的基础上通过获取更精确的候选框，进一步提升了算法性能，为后续的研究提供了参考模型。弱监督/无监督的提出释放了人工标注的压力，mAP 在 CUHK-SYSU 数据集上达到 80% 以上，但与有监督的行人搜索仍然存在一定差距。基于 Faster RCNN 的行人搜索方法在 CUHK-SYSU 数据集上 mAP 大多可达到 90% 以上，基本可以满足行人搜索任务需求。

基于 Transformer 的行人搜索方法在 CUHK-SYSU 数据集上的 mAP 均达到 90% 以上，该方法有望在未来进一步提升搜索精度，并为实际应用带来可能。其他方法的提出为行人搜索的解决方法打开了新思路，丰富了行人搜索的研究方向。

损失函数指导模型训练和优化。由表 4～表 6 可知，行人搜索中传统损失函数的应用较少，最常用的是 OIM 损失函数，后续会改进 OIM 损失函数以提升模型性能和鲁棒性。

综上，基于图像的行人搜索方法需要同时考虑网络结构和损失函数的设计。网络结构应适应行人搜索任务面临的困难特点，提取具有判别性和鲁棒性的行人特征。损失函数要结合网络结构特点和数据集特点，并进行实验验证以获得最佳性能。

3 结束语

基于图像的行人搜索方法在近几年取得了显著进展，然而行人搜索任务存在场景复杂、尺度变化及检测识别不一致等挑战，使该任务应用到实际仍存在一定困难。根据已有行人搜索的研究及存在的问题，本文总结了以下 4 点行人搜索的未来研究方向。

1） 多模态行人搜索

当前大部分的行人搜索研究是基于图像或基于文本的单模态方法。在现实场景中，对目标行人进行搜索时，通常会有目击行人的描述及监控摄像头中目标行人图像。当行人图像因遮挡、错位等无法获得全身图像时，描述行人的文本或语音可以补充更多目标行人信息。研究多模态行人搜索的前提是生成多模态数据集，可以在当前已有的公开数据集上人工或利用文本生成网络添加对行人穿着、发型等语义描述。因此，如何利用图像信息和文本或语音信息并将其有效融合是未来多模态行人搜索值得研究的方向。

2） 网络模型推理的实时性

目前大多数行人搜索方法的准确度已达到 90% 以上，但仍无法应用到实际场景中。其中最重要的因素是行人搜索算法模型的实时性难以满足实际需求，还没有对行人搜索方法实时性方面的研究。尤其对基于 Transformer 的行人搜索方法在精度方面展现了显著的发展潜力，但 Transformer 模型计算量大，工程化应用困难。因此，在网络结构方面通过网络剪枝、量化等模型压缩技术，或深

入研究知识蒸馏等技术或许可以提高模型推理速度，达到实时性的需求。

3） 复杂场景下的行人搜索

对行人搜索的研究与多样的现实应用场景存在两个差异。一是现有公开数据集假设行人短期内外观不会改变，排除了行人易装的情况。而在搜索犯罪嫌疑人等特定应用下，目标行人极有可能易装。且在采集待搜索数据的过程中，行人也有可能改变穿着。目前，文献 [102] 研究了融合行人面部信息的行人搜索方法。二是夜间行人场景缺失。目标行人夜间活动时的图像为单通道的红外夜视图像，如何平衡白天三通道的 RGB 图像与夜间单通道热成像图像之间的差异是具有挑战的问题。

4） 无监督行人搜索

无监督的行人搜索可以解放对行人边界框和行人身份标注的依赖。该方法可以扩大行人搜索数据集规模，使行人搜索适应多变的应用场景成为可能。本文研究发现无监督行人搜索的研究尚有发挥空间，需要研究人员进一步探索和创新，以提高行人搜索的准确性、鲁棒性和扩展性，适应各种复杂的应用场景。

随着学者们对行人搜索技术的关注日益增加，本文围绕基于图像的行人搜索方法，详细梳理了该领域近年来相关的文献，从研究背景、方法分类、性能评估等方面进行了详细阐述。行人搜索在生活中应用广泛，但将学术成果应用到实际场景中仍有较大壁垒。因此，我们不仅要突破学术研究上的性能表现，还要将其赋予更多的应用场景。

参 考 文 献

[1] LI X L, LU Y, LIU B, et al. Counterfactual intervention feature transfer for Visible-Infrared person re-identification[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 381-398.

[2] FU D P, CHEN D D, YANG H, et al. Large-scale Pre-training for person re-identification with noisy labels[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 1-11.

[3] GU A Q, CHANG H, MA B P, et al. Cloths-changing person re-identification with RGB modality only[C]//Proc of the 2022 IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE, 2022: 1060-1069.

[4] ISLAM K. Person search: New paradigm of person re-identification: A survey and outlook of recent works[J]. Image and Vision Computing, 2020, 101: 103970.

[5] LIN X T, REN P Z, XIAO Y, et al. Person search challenges and solutions: A survey[C]//Proceedings of the 30th International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 4500-4507.

[6] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[7] XIAO T, LI S, WANG B C, et al. Joint detection and identification feature learning for person search[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 3376-3385.

[8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 770-778.

[9] HE Z W, ZHANG L. End-to-end detection and re-identification integrated net for person search[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 349-364.

[10] ZHANG L, HE Z W, YANG Y, et al. Tasks integrated networks: Joint detection and retrieval for image search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 456-473.

[11] MUNJAL B, AMIN S, TOMBARI F, et al. Query-guided end-to-end person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 811-820.

[12] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 7132-7141.

[13] MUNJAL B, FLABOREA A, AMIN S, et al. Query-guided networks for few-shot fine-grained classification and person search[J]. Pattern Recognition, 2023, 133: 109049.

[14] DONG W K, ZHANG Z X, SONG C F, et al. Bi-directional interaction network for person search[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 2836-2845.

[15] LIU C, YANG H, ZHOU Q, et al. Subtask-dominated transfer learning for long-tail person search[EB/OL]. [2023-03-21]. https://doi.org/10.48550/arXiv:2112.00527.

[16] MUNJAL B, GALASSO F, AMIN S. Knowledge distillation for end-to-end person search[C]//Proc of the 30th British Machine Vision Conference. Cardiff, UK: BMVA Press, 2019: 216.

[17] LI R L, ZHANG Y Z, ZHU S D, et al. Person search via class activation map transferring[J]. Multimedia Tools and Applications, 2021, 80(16): 24271-24286.

[18] ZHANG X Y, WANG X L, BIAN J W, et al. Diverse knowledge distillation for end-to-end person search[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(4): 3412-3420.

[19] LIU C, YANG H, ZHOU Q, et al. Making person search enjoy the merits of person re-identification[J]. Pattern Recognition, 2022, 127: 108654.

[20] YAN Y C, ZHANG Q, NI B B, et al. Learning context graph for person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 2153-2162.
[21] TIAN K, HUANG H J, YE Y, et al. End-to-end thorough body perception for person search[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12079-12086.
[22] HAN B J, KO K, SIM J Y. End-to-end trainable trident person search network using adaptive gradient propagation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2021: 905-913.
[23] CHEN L Q, XIE W, TU Z G, et al. Multi-attribute enhancement network for person search[C]//Proceedings of the International Joint Conference on Neural Networks. New York: IEEE, 2021: 1-8.
[24] ZHONG Y J, WANG X Y, ZHANG S L. Robust partial matching for person search in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 6826-6834.
[25] YANG W J, LI D W, CHEN X T, et al. Bottom-up foreground-aware feature fusion for person search[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 3404-3412.
[26] ZHENG D Y, XIAO J M, HUANG K Z, et al. Segmentation mask guided end-to-end person search[J]. Signal Processing: Image Communication, 2020, 86: 115876.
[27] CHEN D, ZHANG S S, YANG J, et al. Norm-aware embedding for efficient person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 12612-12621.
[28] LYU N, XIANG X Z, WANG X Y, et al. Stable and effective one-step method for person search[C]// 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2021: 1960-1964.
[29] LI Z J, MIAO D Q. Sequential end-to-end network for efficient person search[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2011-2019.
[30] QIN J, ZHENG P, YAN Y C, et al. Movienet-PS: A large-scale person search dataset in the wild[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2023: 1-5.
[31] TIAN Y L, CHEN D, LIU Y N, et al. Grouped adaptive loss weighting for person search[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 6774-6782.
[32] FIAZ M, CHOLAKKAL H, ANWER R M, et al. SAT: Scale-augmented transformer for person search[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2023: 4809-4818.
[33] LEE S, OH Y, BAEK D, et al. OIMNet++: Prototypical normalization and localization-aware learning for person search[C]//European Conference on Computer Vision. Cham: Springer, 2022: 621-637.
[34] FIAZ M, CHOLAKKAL H, NARAYAN S, et al. PS-ARM: An end-to-end attention-aware relation mixer network for person search[C]//Asian Conference on Computer Vision. Cham: Springer, 2023: 234-250.
[35] JAFFE L, ZAKHOR A. Gallery filter network for person search[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2023: 1684-1693.
[36] LIU Z, MAO H Z, WU C Y, et al. A ConvNet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 11966-11976.
[37] MA M Y, YIN H J. Scene context enhanced network for person search[C]//Proceedings of the IEEE International Conference on Image Processing. New York: IEEE, 2022: 2541-2545.
[38] OUYANG H, ZENG J X, LENG L. Inception convolution and feature fusion for person search[J]. Sensors, 2023, 23(4): 1984.
[39] HAN B J, KO K, SIM J Y. Context-aware unsupervised clustering for person search[C]//Proc of the 32nd British Machine Vision Conference. Virtual Event: BMVA Press, 2021: 386.
[40] YAN Y C, LI J P, LIAO S C, et al. Exploring visual context for weakly supervised person search[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 3027-3035.
[41] WANG B Z, YANG Y, WU J L, et al. Self-similarity driven scale-invariant learning for weakly supervised person search[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 1813-1822.
[42] HAN C C, SU K, YU D D, et al. Weakly supervised person search with region siamese networks[C]//Proc of the 2021 IEEE International Conference on Computer Vision. Montreal, QC: IEEE, 2021: 11986-11995.
[43] WANG J B, PANG Y W, CAO J L, et al. Deep intra-image contrastive learning for weakly supervised one-step person search[J]. Pattern Recognition, 2024, 147: 110047.
[44] LI J J, YAN Y C, WANG G S, et al. Domain adaptive person search[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 302-318.
[45] LI J J, WANG G S, YAN Y C, et al. Generalizable person search on open-world user-generated video content[EB/OL]. [2023-06-12]. https://doi.org/10.48550/arXiv.2310.10068.
[46] ALMANSOORI M K, FIAZ M, CHOLAKKAL H. DDAM-PS: Diligent domain adaptive mixer for person search[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE, 2024: 6674-6683.
[47] OH M, KIM D, SIM J Y. Domain generalizable person search using unreal dataset[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(5): 4361-4368.
[48] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot MultiBox detector[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016:

21-37.
[49] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 779-788.
[50] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. [2023-02-21]. http://arxiv.org/abs/1804.02767v1.
[51] LI C Y, LI L L, JIANG H L, et al. YOLOv6: A single-stage object detection framework for industrial applications[EB/OL]. [2023-04-22]. https://doi.org/10.48550/arXiv.2209.02976.
[52] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 7464-7475.
[53] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 580-587.
[54] ZHENG L, ZHANG H H, SUN S Y, et al. Person re-identification in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 3346-3355.
[55] CHEN D, ZHANG S S, OUYANG W L, et al. Person search via a mask-guided two-stream CNN model[M]// Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 764-781.
[56] LAN X, ZHU X T, GONG S G. Person search by multi-scale matching[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 553-569.
[57] DONG W K, ZHANG Z X, SONG C F, et al. Instance guided proposal network for person search[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 2582-2591.
[58] WANG C, MA B P, CHANG H, et al. TCTS: A task-consistent two-stage framework for person search[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 11949-11958.
[59] LI J H, LIANG F H, LI Y X, et al. Fast person search pipeline[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. New York: IEEE, 2019: 1114-1119.
[60] LI Y, YIN K N, LIANG J, et al. A multitask joint framework for real-time person search[J]. Multimedia Systems, 2023, 29(1): 211-222.
[61] KE X, LIU H, GUO W Z, et al. Joint sample enhancement and instance-sensitive feature learning for efficient person search[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7924-7937.
[62] ZHOU K Y, YANG Y X, CAVALLARO A, et al. Omni-scale feature learning for person re-identification [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 3701-3711.
[63] JIA C Y, LUO M N, YAN C X, et al. CGUA: Context-guided and unpaired-assisted weakly supervised person search[EB/OL]. [2023-04-12]. https://doi.org/10.48550/arXiv.2203.14307.
[64] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformer for image recognition at scale[EB/OL]. [2023-07-09]. https://doi.org/10.48550/arXiv.2010.11929.
[65] ZHENG S X, LU J C, ZHAO H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 6877-6886.
[66] WANG Y K, YE T Q, CAO L L, et al. Bridged transformer for vision and point cloud 3D object detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 12104-12113.
[67] CHEN S H, ZHUANG Y Q, LI B X. Learning context-aware embedding for person search[EB/OL]. [2023-03-21]. https://doi.org/10.48550/arXiv.2111.14316.
[68] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 213-229.
[69] CAO J L, PANG Y W, ANWER R M, et al. PSTR: End-to-end one-step person search with transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 9448-9457.
[70] LYU N, XIANG X Z, WANG X Y, et al. Global-aware and local-aware enhancement network for person search[J]. Computer Vision and Image Understanding, 2023, 236: 103804.
[71] YU R, DU D W, LALONDE R, et al. Cascade transformers for end-to-end person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 7257-7266.
[72] CHEN L, XU J H. Sequential transformer for end-to-end person search[M]//Lecture Notes in Computer Science. Singapore: Springer Nature Singapore, 2023: 226-238.
[73] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: Deformable transformers for end-to-end object detection [C]//The 9th International Conference on Learning Representation. [S.L.]: ICLR, 2021: 1-16.
[74] LIU H, FENG J S, JIE Z Q, et al. Neural person search machines[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2017: 493-501.
[75] YAN Y C, LI J P, QIN J, et al. Anchor-free person search[C]//Proc of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Virtual Event: IEEE, 2021: 7690-7699.
[76] TIAN Z, SHEN C H, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2019: 9626-9635.
[77] HAN C C, ZHENG Z D, GAO C X, et al. Decoupled and

memory-reinforced networks: Towards effective feature learning for one-step person search[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1505-1512.
[78] HAN C C, ZHENG Z D, SU K, et al. DMRNet++: Learning discriminative features with decoupled networks and enriched pairs for one-step person search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7319-7337.
[79] LIU Y, LI Y P, KONG C Y, et al. Vision shared and representation isolated network for person search[C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2022: 1216-1222.
[80] JIA C Y, LUO M N, DANG Z H, et al. PSDiff: Diffusion model for person search with interative and collaborative[EB/OL]. [2023-02-22]. https://doi.org/10.48550/arXiv.2309.11125.
[81] TIAN Y L, CHEN D, LIU Y N, et al. Divide and conquer: Hybrid pre-training for person search[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6): 5224-5232.
[82] LV N, XIANG X Z, WANG X Y, et al. Efficient person search via learning-to-normalize deep representation[J]. Neurocomputing, 2022, 495: 169-177.
[83] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.
[84] LI L Q, YANG H, CHEN L. Spatial invariant person search network[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018: 122-133.
[85] LU Y, HONG Z R, LIU B, et al. Dhff: Robust multi-scale person search by dynamic hierarchical feature fusion [C]//Proceedings of the IEEE International Conference on Image Processing. New York: IEEE, 2019: 3935-3939.
[86] DING S Y, LIN L, WANG G R, et al. Deep feature learning with relative distance comparison for person re-identification[J]. Pattern Recognition, 2015, 48(10): 2993-3003.
[87] ZHAO C R, CHEN Z C, DOU S G, et al. Context-aware feature learning for noise robust person search[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 7047-7060.
[88] SHI W, LIU H, MENG F Y, et al. Instance enhancing loss: Deep identity-sensitive feature embedding for person search[C]//Proceedings of the 25th IEEE International Conference on Image Processing. New York: IEEE, 2018: 4108-4112.
[89] DAI J, ZHANG P P, LU H C, et al. Dynamic imposter based online instance matching for person search[J]. Pattern Recognition, 2020, 100: 107120.
[90] KIM H, JOUNG S, KIM I J, et al. Prototype-guided saliency feature learning for person search[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 4863-4872.
[91] XIAO J M, XIE Y C, TILLO T, et al. IAN: The individual aggregation network for person search[J]. Pattern Recognition, 2019, 87: 332-340.
[92] LOESCH A, RABARISOA J, AUDIGIER R. End-to-end person search sequentially trained on aggregated dataset [C]//Proceedings of the IEEE International Conference on Image Processing. New York: IEEE, 2019: 4574-4578.
[93] WANG C, ZHANG Q, HUANG C, et al. Mancs: A multi-task attentional network with curriculum sampling for person re-identification[C]//European Conference on Computer Vision. Cham: Springer, 2018: 384-400.
[94] WANG G S, YUAN Y F, CHEN X, et al. Learning discriminative features with multiple granularities for person re-identification[C]//Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM, 2018: 274-282.
[95] HAN C C, YE J C, ZHONG Y S, et al. Re-ID driven localization refinement for person search[C]//Proc of the 2019 IEEE International Conference on Computer Vision. Seoul Korea: IEEE, 2019: 9813-9822.
[96] 孙杰, 吴绍鑫, 王学军, 等. 基于 Sophon SC5+芯片构架的行人搜索算法与优化[J]. 计算机应用, 2023, 43(3): 744-751.
SUN J, WU S X, WANG X J, et al. Efficient person search algorithm and optimization with Sophon SC5+ chip architecture[J]. Journal of Computer Applications, 2023, 43(3): 744-751.
[97] LIU H, SHI W, HUANG W P, et al. A discriminatively learned feature embedding based on multi-loss fusion for person search[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2018: 1668-1672.
[98] GAO C Y, YAO R, ZHAO J Q, et al. Structure-aware person search with self-attention and online instance aggregation matching[J]. Neurocomputing, 2019, 369: 29-38.
[99] ZHAI S L, LIU S Q, WANG X, et al. FMT: Fusing multi-task convolutional neural network for person search[J]. Multimedia Tools and Applications, 2019, 78(22): 31605-31616.
[100] CHEN D, ZHANG S S, OUYANG W, et al. Hierarchical online instance matching for person search[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 10518-10525.
[101] YANG Y Z, ZHANG X, GENG Q C, et al. Multi-task feature decomposition based marginal distribution for person search[C]//Proceedings of the IEEE International Conference on Multimedia and Expo. New York: IEEE, 2022: 1-6.
[102] SHU X J, TAO Y S, QIAO R Z, et al. Head and body: Unified detector and graph network for person search in media[EB/OL]. [2023-3-11]. https://doi.org/10.48550/arXiv.2111.13888.

编 辑 税 红