arXiv:2302.04607v1 [cs.CV] 9 Feb 2023

面向弱监督单步行人搜索的深度图像内对比学习

Jiabei Wang、Yanwei Pang，*IEEE* 高级会员、Jiale Cao、Hanqing Sun、Zhuang Shao 和 Xuelong Li，*IEEE* 会士

摘要—弱监督行人搜索旨在仅使用行人边界框标注，联合执行行人检测与重识别（re-id）。近期，对比学习首次应用于弱监督行人搜索，常见的两种对比策略是基于记忆的对比和图像内对比。我们认为，当前的图像内对比较为浅层，容易受到空间层级变化和遮挡层级变化的影响。本文提出一种基于孪生网络的新型深度图像内对比学习方法。其包含两个关键模块：空间不变对比（SIC）和遮挡不变对比（OIC）。SIC 在孪生网络的两个分支之间执行多对一对比，并在孪生网络的一个分支中执行密集预测对比。借助这些多对一对比和密集对比，SIC 倾向于学习具有判别性的尺度不变和位置不变特征，从而解决空间层级变化问题。OIC 通过掩码策略增强特征一致性，以学习遮挡不变特征。我们分别在两个行人搜索数据集 CUHK-SYSU 和 PRW 上进行了大量实验。我们的方法在弱监督单步行人搜索方法中实现了最先进性能。我们希望，这种简单的图像内对比学习能够为弱监督行人搜索提供更多范式。源代码可在 https://github.com/jiabeiwangTJU/DICL 获取。

索引词——**弱监督行人搜索、对比学习、空间不变、遮挡不变。**

一、引言

PERSON 行人搜索旨在从一组图库图像中搜索给定的查询行人，这被视为行人检测与重识别（再识别） [5], [6], [7]的联合任务。目前，大多数现有的行人搜索方法 [2], [8], [9]均采用全监督方式，基于行人边界框标注和身份标注开发。与行人检测的边界框标注相比，行人重识别的身份标注成本高昂且耗时。一方面，这很困难，而且在不同场景中获取大量身份标注的成本很高。CUHK-SYSU和PRW数据集中，分别有近72.7%和17.4%的行人边界框没有身份标注 [10], 另一方面，准确标注不同场景中身份相同的行人十分耗时，尤其是在行人的外观、尺度变化较大或存在遮挡等情况时。因此，有必要探索仅使用边界框标注的行人搜索，即弱监督行人搜索。

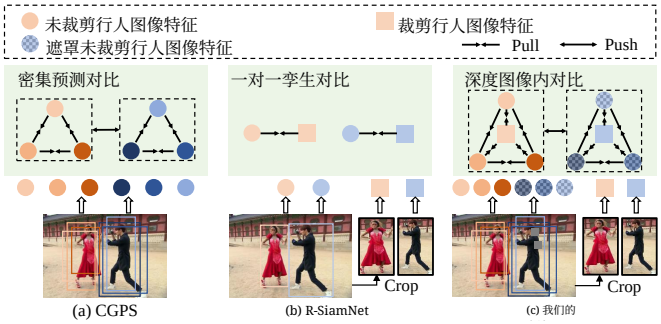

图1. 不同弱监督单步方法的图像内对比策略。CGPS（a）使用单一网络，以整幅图像作为输入，执行密集预测对比。R-SiamNet（b）使用孪生网络，以整幅图像和裁剪后的行人作为输入，执行一对一真实标注对比。与这些浅层图像内对比策略相比，我们的方法（c）利用深度图像内对比学习。我们在孪生网络的两个分支之间进行多对一孪生对比，并在孪生网络的一个分支中进行密集预测对比。此外，我们通过在孪生网络的一个分支中随机遮掩部分行人，执行遮挡不变对比。

近期，已有若干基于对比学习的单步方法被提出，用于执行弱监督行人搜索。这些方法采用了两种对比学习策略，包括基于记忆的对比和图像内对比[1]。基于记忆的对比在所有训练图像上构建簇级重识别记忆库，并利用它来监督当前图像中的重识别特征学习。图像内对比旨在利用图像内有用的对比信息进行重识别特征学习。我们致力于改进图像内对比部分。图1展示了两种典型弱监督单步方法 CGPS [11]和 R-SiamNet [10]所采用的图像内对比。如图1(a)所示，CGPS 使用单一网络探索密集预测对比。它使属于同一人物的重识别特征彼此接近，并将属于不同人物的特征彼此推远。如图1(b)所示，R-SiamNet采用孪生网络进行一对一孪生对比。它在两个分支中提取真实标注对应的两个重识别特征，并使这两个有上下文或无上下文的重识别特征保持一致。这些方法在弱监督单步行人搜索方面取得了初步成效。然而，它

J. Wang、Y. Pang、J. Cao 和 H. Sun 任职于天津大学电气与信息工程学院，中国天津 300072（电子邮箱：{jiabeiwang、pyw、connor、hqSun}@tju.edu.cn）。Z. Shao 任职于英国考文垂 CV4 7AL 的华威大学华威制造集团（电子邮箱：Zhuang.Shao@warwick.ac.uk）。X. Li 任职于西北工业大学计算机科学学院和光学图像分析与学习中心（OPTIMAL），中国西安 710072（电子邮箱：li@nwpu.edu.cn）。

[1]图像内对比在训练过程中也可能指批次内对比。为简便起见，此处使用图像内对比这一说法。

与全监督单步行人搜索方法相比，仍有很大差距 [12], [13]。

我们认为，关键原因之一是信息挖掘不足且图像内对比较浅，因而无法有效应对行人实例的空间层级和遮挡层级变化。不同图像中检测到的行人，其尺度可能存在较大差异，或边界框定位不准确（空间层级变化）。此外，检测到的行人有时会被周围物体遮挡（遮挡层级变化）。因此，由于这些空间层级和遮挡层级变化，跨图像匹配同一行人的重识别特征变得困难。

为解决空间层级和遮挡层级变化所引发的问题，我们提出了一种新颖的深度图像内对比学习方法（DICL），充分利用图像中每个行人实例的实例级信息（图1(c)）。DICL基于孪生网络构建，包含两个新颖的对比模块：空间不变对比（SIC）和遮挡不变对比（OIC）。在孪生网络中，搜索分支从整幅图像中提取重识别特征，而实例分支从裁剪后的真实标注中提取重识别特征。SIC不仅在孪生网络的两个分支之间实现多对一特征一致性，还在搜索分支中进行密集预测对比。借助多对一和密集对比，SIC能够学习尺度不变和位置不变特征，以应对空间层级变化。为应对遮挡层级变化，OIC在搜索分支中采用掩码策略，增强两个分支之间的重识别特征一致性。在两个经典行人搜索数据集CUHK-SYSU [5] 和PRW [14]上的实验结果表明，所提出的DICL有效且优越。我们的贡献总结如下：

- 我们提出一种通过深度图像内对比学习（DICL）实现的新型弱监督单步行人搜索框架。无需身份标注，DICL 便能充分挖掘图像内信息，以应对空间层级和遮挡层级变化。
- 我们设计了两个图像内对比模块：空间不变对比（SIC）和遮挡不变对比（OIC）。SIC 通过多对一孪生对比和密集预测对比，学习尺度不变和位置不变特征。OIC 则通过搜索分支中的掩码策略，学习遮挡不变特征。
- 我们的方法在 CUHK-SYSU 和 PRW 上取得了优异结果，优于以往的弱监督单阶段方法。在 CUHK-SYSU 测试集上，其平均精度均值为 87.4%，Top-1准确率为 88.8%。在 PRW 测试集上，其平均精度均值为 35.5%，Top-1准确率为 80.9%。

II. 相关工作

本节首先介绍人物搜索，包括全监督和弱监督人物搜索。随后，我们简要回顾无监督行人重识别（re-id）和对比学习。

A. 行人搜索

人物搜索是行人检测与行人重识别的联合任务，在视频监控中发挥着重要作用[15], [16], [3], [4]。现有的人物搜索方法可分为两大类：两阶段方法和单阶段方法。两阶段方法 [17], [18]先检测行人，再在裁剪后的人物图像中识别行人，将两个子任务分别执行。Zheng 等人 [14] 探索了用于人物搜索的不同检测器和再识别方法。Chen等人 [17], [2] 认为，行人前景区域对再识别更为重要，并提出了一种掩码引导的双流 CNN 模型，以突出前景信息。Yao 等人 [1] 提出在相似度计算中考虑目标性和排斥信息。Wang 等人 [18], [19]采用身份引导查询检测器生成类似查询的候选区域，并设计了检测结果自适应再识别模型，以提升再识别效果。

单阶段方法 [12], [13] 在统一的端到端框架中执行行人检测和行人重识别。与两阶段方法相比，单阶段方法较为简单，但也更具挑战性，因为需要兼顾行人检测和行人重识别的不同目标。大多数现有单阶段方法都基于 Faster R-CNN 目标检测框架 [20]–[21]。Xiao 等 [5] 最初尝试在 R-CNN 头部网络后添加一个全连接层，并生成用于在线实例匹配的重识别特征。Chen 等[22] 提出在极坐标系中分解检测特征和重识别特征。Li 等 [23]采用级联头部网络进行检测和重识别：先执行检测，再进行重识别。Han 等[24], [25] 提出仅使用感兴趣区域特征进行重识别，从而解耦检测与重识别。Dong 等 [26]提出了一个双向交互网络，在训练期间同时以整幅图像和裁剪后的实例图像作为输入，并利用来自行人图像块的精准人体外观信息来区分身份。Han 等 [27] 将检测、重识别和部件分类结合起来，以提升行人搜索效果。此外，还有一些其他方法专注于改进在线实例匹配损失，例如 HOIM [13]和 OIM ++ [28]。Yan 等 [12] 没有采用基于锚框的 Faster R-CNN 检测器，而是提出采用无锚框 FCOS 检测器 [29] 进行无锚框行人搜索，从而避免人工设计锚框。近期的两项工作 PSTR [9]和 COAT [30]采用基于 Transformer 的网络进行行人搜索。

弱监督行人搜索： 近年来，一些方法专注于弱监督行人搜索。例如，两步法 CGUA [31]提出了一种上下文引导聚类算法和非配对辅助记忆，以利用大量非配对行人。单步法 CGPS [11]使用单一网络，利用检测上下文进行密集预测对比。R-SiamNet [10] 使用孪生网络和稀疏真实标注进行图像内对比。我们认为，单步方法中的这些图像内对比较为浅层，未能充分利用图像内信息。

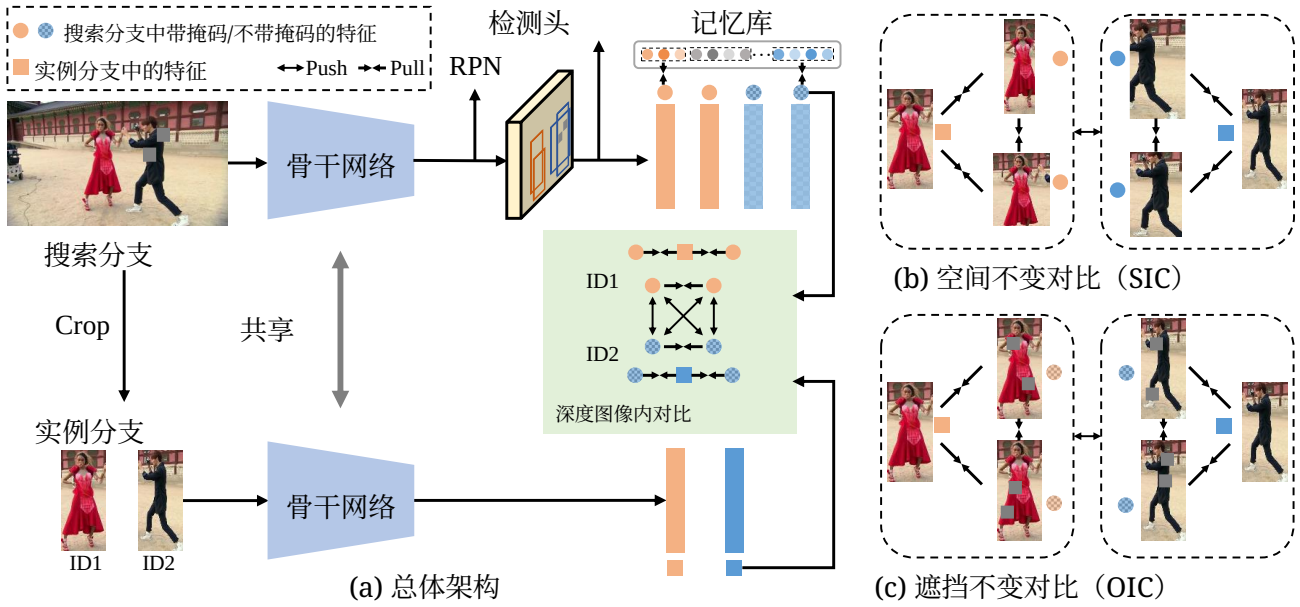

图2. 我们提出的基于深度图像内对比学习（DICL）方法的架构（a）。该方法采用孪生网络，包含搜索分支和实例分支。DICL通过两个新颖模块执行图像内对比：空间不变对比（SIC）和遮挡不变对比（OIC）。SIC（b）在孪生网络的两个分支中执行多对一对比，并对图像的所有预测进行密集对比。OIC（c）利用掩码策略增强特征一致性。

B. 无监督行人重识别

无监督行人重识别旨在从无标签裁剪图像中学习判别性特征，以识别不同的行人。大多数基于深度学习的方法专注于无监督域适应（UDA）任务，该任务要求源域中有或没有身份标签，且目标域中没有身份标签。基于伪标签的方法使用聚类算法对实例特征进行分类，例如 k均值、DBSCAN [32],和 Finch [33]。例如，BUC [34] 将每张图像视为一个身份，并对特征嵌入进行自底向上聚类，以减少类别数量。SPCL[35] 提出了一种自步对比学习策略，并采用新颖的聚类可靠性准则来生成更可靠的聚类簇。一些方法利用跨域迁移学习[36], [37]。DAPS [38] 提出了一种无监督域自适应行人搜索范式。

C. 对比学习

对比学习从正样本对或负样本对中学习特征表示。SimCLR [39] 引入了一个可学习的非线性变换头，以较大的批量大小和较长的训练步数训练两条增强路径。BYOL [40]提出了在线网络和目标网络这两个交互网络，仅使用正样本对进行缓慢更新。SimSiam [41] 提出了一种简单的孪生网络，无需负样本对、大批量或动量编码器，并使用停止梯度操作来防止坍塌。ContrastiveCrop [42] 是一种即插即用的裁剪模块，用于对比裁剪对，可避免假阳性对和外观相似对。

III. 我们的方法

本节首先介绍动机，其次展示所提出方法的总体架构，然后描述两个新颖的模块，最后介绍训练与推理过程。

A. 动机

对比学习最初被用于弱监督单步行人搜索 [11], [10], , 包括基于记忆的对比和图像内对比。然而，其性能仍远远落后于全监督单步行人搜索。我们认为，其中一个关键原因是，这些方法没有充分挖掘图像内信息，以应对行人实例的空间层级变化和遮挡层级变化。例如，在不同场景下检测到的行人可能具有不同的尺度，或边界框定位不准确，这被称为空间层级变化。遮挡层级变化是指检测到的行人有时会被其他物体遮挡。因此，由于这些空间层级变化和遮挡层级变化，要在不同图像中准确匹配同一行人的重识别特征变得十分困难。

为了弥合弱监督单阶段方法与全监督单阶段方法之间的性能差距，我们提出了一种新颖的基于深度图像内对比学习的方法（DICL），在考虑行人实例的空间层级变化和遮挡层级变化的情况下，充分利用每张行人图像的实例级信息。

B. 总体架构

图2给出了我们的DICL总体架构。我们采用包含搜索分支和实例分支的孪生网络。搜索分支提取整幅图像的特征，用于检测和再识别，而实例分支

提取从整幅图像中裁剪并调整为固定大小的真实标注行人的重识别特征。特征提取基于Faster R-CNN [20]完成。与原始Faster R-CNN不同，我们进行了如下修改。我们使用骨干网络最后两层的融合特征图来进行区域提议网络（RPN）和RoI头网络的处理。RoI头网络由一个 3×3可变形层 [43], 、两个 3×3 卷积层和三个并行的全连接层组成。在搜索分支中，我们首先使用RoIAlign层提取候选区域的特征，然后利用RoI头网络执行分类、回归和重识别特征生成。在实例分支中，我们移除RoI层，直接使用RoI头网络提取重识别特征。

基于搜索分支和实例分支生成的重识别特征，我们在训练期间进行图像内对比学习和基于记忆库的对比学习。对于图像内对比学习，我们提出了一种新颖的深度图像内对比学习策略，其中包括一个空间不变对比模块和一个遮挡不变对比模块，具体如下。对于基于记忆库的对比学习，我们使用 SPCL[35], 构建一个簇级重识别记忆库，并像 [11]一样监督重识别特征学习。该记忆库使用两个分支的平均重识别特征进行初始化。

C. 空间不变对比

同一个人在不同摄像机下的尺度会有所变化。此外，检测无法保证像真实标注那样准确地进行边界框定位。因此，不同图像中检测到的人物可能具有不同的尺度，或边界框定位不准确。我们将这种现象称为空间层级变化。空间层级变化会影响跨图像的准确重识别匹配。为应对这种空间层级变化，我们提出了空间不变对比（SIC）模块，该模块在孪生网络的两个分支之间执行多对一孪生对比，并在搜索分支中进行密集预测对比。

多对一孪生对比旨在确保搜索分支中同一人物的多个预测重识别特征与实例分支中一个真实标注的重识别特征保持一致。当对应的预测边界框与行人真实边界框的IoU重叠度高于给定阈值时，我们将这些预测重识别特征分配给该人物。假设一张图像中有 n^p 个预测和 n^g 个真实标注，搜索分支中的重识别特征表示为 f_i、 $i = 1$、.....、 n^p，实例分支中的重识别特征表示为 f_j^g、 $j = 1$、.....、 n^g。我们要求搜索分支中同一人物的重识别特征与对应真实标注的重识别特征保持一致，其中一个预测的损失写作

$$L_{mto}(i) = 1 - s(f_i, f_{j(i)}^g), i = 1, ..., n^p, \tag{1}$$

其中， s 表示余弦距离相似度，而 j（i）表示与第 j个真实标注对应的第 i个预测重识别特征。密集预测对比旨在使同一人物的重识别特征彼此更接近，并将不同人物的重识别特征彼此推远。密集对比的损失通过困难样本挖掘策略 [44],下的三元组损失来表示，其形式如下：

$$L_{tri}(i) = m + \max_{f \in f_i^+} d(f_i, f) - \min_{f \in f_i^-} d(f_i, f), \tag{2}$$

其中， d 是距离度量函数， f_i^+ 表示与预测的 f_i 具有相同身份的正样本集， f_i^- 表示与预测的 f_i 具有不同身份的负样本集， m 是距离间隔参数。为公平比较，多对一孪生对比采用了 CGPS 和 R-SiamNet 中使用的类似损失，并未探讨如何设计更好的对比损失，例如 NCE。

多对一孪生对比能够约束不同尺度和边界框位置的行人重识别特征彼此保持一致，而密集预测对比则使同一人物的重识别特征彼此更加紧凑，同时让不同 ID 的特征彼此分离。因此，我们的 SIC 倾向于学习具有判别性的尺度不变和位置不变特征，即使检测到的边界框不准确或尺度变化较大，也能实现准确的重识别匹配。

D. 遮挡不变对比

遮挡在日常生活中十分常见，会对行人再识别产生负面影响。我们将其称为遮挡层级变化。为应对这种遮挡层级变化，我们提出了一种遮挡不变对比（OIC）模块，用于对掩蔽重识别特征与未掩蔽重识别特征进行对比。该模块包含三种不同的掩蔽策略：在搜索分支上掩蔽、在实例分支上掩蔽，以及同时在搜索分支和实例分支上掩蔽。我们发现，在搜索分支上进行掩蔽的效果更好，如表 V 所示。原因可能是，未经过掩蔽的实例分支能够提供一致且无遮挡的特征，从而引导搜索分支学习遮挡不变的重识别特征。

给定图像中的一个人物，我们将人物边界框划分为 14×6 个网格，并使用像素均值随机擦除部分网格。我们以 50% 的概率对每个人物执行掩码操作，且最多掩码 2 个网格。我们约束搜索分支中掩蔽行人的重识别特征与实例搜索中的未掩蔽重识别特征保持一致，其中一次预测的损失表示为

$$L_o(i) = 1 - s(f_i^o, f_{j(i)}^g), , i = 1, ..., n^o, \tag{3}$$

其中， o 表示掩蔽行人的预测结果， n^o表示来自被遮挡行人的预测数量。OIC 模块基于 SIC 模块运行。当二者同时执行时，OIC 模块还会在掩码或未掩码实例之间使用三元组损失。

与 CGPS 的比较 [11] 和 R-SiamNet [10]: 我们提出的 IDCL 与 CGPS 和 R-SiamNet 相关，但存在显著差异，并作出了重要贡献。(i) 我们的主要贡献是提出了一种深度图像内对比学习（DICL）框架。DICL 提供了两种不同的对比策略，包括空间不变对比（SIC）和遮挡不变对比（OIC）。表 VIII 中的实验结果

算法1： 深度图像内对比学习
用于弱监督行人搜索的算法

输入： 训练图像；
边界框标注。
输出： 搜索分支的模型。
对每个训练轮次**执行**
a:使用以下信息提取所有训练图像的特征
边界框标注。
b:更新记忆库并进行聚类。
对每个小批量**执行**
1: 在部分图像上随机添加掩码块
搜索分支中输入图像的行人区域。
2: 生成搜索分支的重识别特征，并
使用孪生网络生成实例分支的重识别特征。
3: 根据公式 4 计算对比重识别损失和
检测损失。
4: 更新记忆库特征。
end
end

表明，我们的 DICL 显著优于 CGPS 和 R-SiamNet。在 PRW 上，平均精度均值分别提升了 19.3% 和 14.3%，Top-1准确率分别提升了 12.9% 和 7.5%。我们希望这种简单而有效的 DICL 能成为强有力的基线，助力这一开放研究领域的后续研究。(ii) 我们 DICL 中的 SIC 是一种深度空间不变对比策略。相比之下，CGPS 和 R-SiamNet 相对浅层，未能充分利用实例空间信息。例如，CGPS 在搜索分支中进行密集对比时会受到未裁剪上下文干扰，而 R-SiamNet 采用一对一真实标注孪生对比，忽略了边界框定位不准确的影响。表 III 中的实验结果表明，即使只是简单地结合 CGPS 和 R-SiamNet（即在搜索分支中进行一对一真实标注孪生对比 + 密集对比），其效果也不及我们的 SIC（平均精度均值：31.4% vs 33.2%，Top-1准确率：77.3% vs 79.9%）。(iii) 我们的 SIC 是一个统一的空间不变对比学习框架，CGPS 和 R-SiamNet 中的对比策略都可以视为其特例。除了 SIC，我们还进一步提出了遮挡不变对比（OIC）。

E. 训练与推理

在训练过程中，我们通过检测所有训练图像，并使用两个分支的平均重识别特征为每个人物生成重识别特征，从而构建初始化重识别记忆库。在训练开始时，我们对记忆库执行聚类。在每次迭代中，聚类级记忆库用于基于记忆库的对比学习，以监督重识别特征学习，如 [12]。除了基于记忆的对比之外，我们还使用孪生网络进行深度图内对比学习。在搜索分支中，我们随机地在部分行人区域上添加掩码，并执行前向传播以进行检测和重识别特征生成。我们保留所有 n 预测边界框的重识别特征，其中有 n^p 个未进行图像掩码的重识别特征，以及 n^o 个进行了图像掩码的重识别特征。在实例分支中，我们将真实标注图像区域裁剪并调整为固定尺寸，然后执行前向传播以生成重识别特征。利用两个分支中的重识别特征，我们执行空间不变对比和遮挡不变对比。我们的方法的对比损失可写为

$$L_c = \frac{1}{n^p}\sum_i L^i_{mto} + \frac{1}{n^o}\sum_i L^i_o + \frac{1}{n}\sum_i L^i_{tri} + L_{oim}, \quad (4)$$

其中，L_{oim} 是用于基于记忆对比的 OIM 损失。因此，总损失可表示为 $L_{all} = L_c + L_{det}$，其中 L_{det}是检测损失。我们在算法 1 中总结了用于弱监督行人搜索的 DICL。

推理时，我们仅保留孪生网络中的搜索分支，用于检测和生成重识别特征。我们首先利用搜索分支分别对查询图像和一组图库图像进行检测并生成重识别特征。之后，将与真实标注边界框重叠度最高的重识别特征设为每个人物的重识别特征。最后，我们计算查询行人与图库图像中各人物的重识别特征相似度，并根据相似度分数选出匹配的人物。

IV. 实验

A. 数据集与设置

CUHK-SYSU [5] 是一个大规模行人搜索数据集，包含街景和电影截图。该数据集包含 18,184 张图像、96,143 个带注释的行人边界框以及 8,432 个行人身份。训练集包含 11,206 张图像、55,272 个行人边界框和 5,532 个行人身份，而测试集包含 6,978 张图像、40,871 个行人边界框和 2,900 个行人身份。推理时，不同查询对应不同的图库图像，图库图像的数量范围为 50 至 4,000。除非特别说明，与大多数行人搜索方法一样，默认图库规模设为 100。

PRW [14] 采集于清华大学，使用了部署在不同位置的六台摄像机。其中五台摄像机的分辨率为 $1,080 \times 1,920$ 像素，另一台为 576×720 像素。该数据集包含 11,816 张图像、43,110 个行人边界框和 932 个行人身份。训练集包含 5,704 张图像、18,048 个行人边界框和 482 个行人身份。测试集包含 6,112 张图像、25,062 个行人边界框和 450 个行人身份。

评估协议。我们采用行人搜索中广泛使用的两种标准评估指标，包括平均精度均值（mAP）和 Top-1准确率。

B. 实现细节

我们基于开源库 mmde- tection [45]实现了我们的方法。我们采用在大规模 ImageNet-1K数据集 [46], 上预训练的 ResNet50，并使用最后两个层的融合特征图作为区域提议网络和 RoI头网络的输入。融合特征图的分辨率为 $H/16 \times W/16 \times 1024$ 像素。搜索分支使用 RoIAlign 生成分辨率为 $14 \times 6 \times 1024$ 像素的特征图。实例分支将行人图像裁剪并缩放至固定尺寸 224×96 像素，因此实例分支中的融合特征图分辨率为$14 \times 6 \times 1024$像素。最终的全连接层生成

表 I 我们的 DICL 中不同模块（包括空间不变对比（SIC）和遮挡不变对比（OIC））对 PRW 测试集上弱监督行人搜索的影响。

方法（弱监督）		再识别	
SIC	OIC	mAP	Top-1准确率
		25.6	73.2
✓		33.2	79.9
✓	✓	**35.5**	**80.9**

表 II
我们提出的 DICL 中不同模块（包括空间不变对比（SIC）和遮挡不变对比（OIC））对 PRW 测试集上全监督行人搜索的影响。

方法（全监督）		再识别	
SIC	OIC	mAP	Top-1准确率
		43.6	78.8
✓		48.8	82.2
✓	✓	**49.6**	**82.4**

用于重识别特征的256-d 维向量。我们采用SPCL [35] 中使用的聚类超参数来生成聚类级别的再识别记忆库。此外，我们利用图像中的人物属于不同身份这一先验，在进行聚类时使用该先验 [31], [11]。

我们的实验在配备SGD优化器的单块 NVIDIA GeForce RTX 3090 GPU 上进行。小批量设置为四张图像，图像尺寸固定为1，500×900 像素。总共训练26个轮次，初始学习率设置为0.001，并在训练轮次16和22之后分别衰减为原来的0.1倍。推理过程中，将图像调整为1，500×900像素。

C. 消融研究

在此，我们在 PRW数据集上进行消融研究，以证明所提出的 DICL 方法的有效性。

DICL 中不同模块的影响表 I 展示了深度图像内对比学习（DICL）在 PRW 测试集上的弱监督行人搜索任务中的效果。DICL 包含空间不变对比（SIC）模块和遮挡不变对比（OIC）模块。基线（a）仅使用真实边界框进行图像内对比学习，如 [10] 所示。在基线中加入 SIC 模块后，平均精度均值提升了 7.6%，Top-1准确率提升了 6.7%。在此基础上加入 OIC 模块后，平均精度均值进一步提升了 2.3%，Top-1准确率提升了 1.0%。总体而言，DICL 使平均精度均值提升了 9.9%，Top-1准确率提升了 7.7%。这表明，深度图像内对比学习对弱监督行人搜索有效。

此外，我们在表 II 中展示了 DICL 对全监督行人搜索的影响。我们根据身份标注构建再识别记忆库，而非使用无监督聚类。我们的 DICL 达到 49.6% 的平均精度均值和 82.4% 的 Top-1准确率，超过基线

表 III 我们的空间不变对比（SIC）模块中整合多对一孪生对比和密集对比的影响。

Num.	多对一	密集	mAP	Top-1准确率
(a)			25.6	73.2
(b)	✓		32.2	79.6
(c)		✓	31.4	77.3
(d)	✓	✓	**33.2**	**79.9**

表IV 我们的空间不变对比（SIC）模块中不同孪生对比设置的影响。真实标注表示人物的真实标注边界框，而正样本表示属于人物的所有预测。

Num.	搜索分支	实例分支	mAP	Top-1准确率
(a)	真实标注	真实标注	25.6	73.2
(b)	正样本	正样本	31.2	79.7
(c)	正样本	真实标注	**33.2**	**79.9**

6.0% 和 3.6%。相关方法 BINet [26] 也采用孪生框架进行全监督行人搜索，其平均精度均值为 45.3%，Top-1准确率为 81.7%。因此，我们的 DICL 在平均精度均值上超过 BINet 4.3%，在 Top-1准确率上超过 0.7%。这证明了我们的深度图像内对比学习在全监督行人搜索中的有效性。由于本文主要关注弱监督设置，因此我们没有在表 VIII 中加入全监督结果。同时，我们的方法在弱监督行人搜索上取得了更大的提升，这表明图像内对比对于弱监督行人搜索更为重要。

SIC 中不同对比方法的影响 我们在表 III 中展示了整合多对一孪生对比和密集预测对比的影响。多对一孪生对比表示搜索分支中人物的多个预测与实例分支中该人物的一个真实标注预测之间的对比。密集预测对比表示搜索分支中所有预测之间的对比。当在两个分支之间整合单一的多对一孪生对比（b），或在搜索分支中整合密集预测对比（c）时，性能均能得到显著提升。当同时整合这两种对比方法（d）时，可取得最佳性能。此外，方法（c）可视为 CGPS 与 R-SiamNet 的组合，其性能劣于我们的方法（d）。因此，我们的 SIC 方法优于简单组合 CGPS 和 R-SiamNet 这两种弱监督方法。

表IV展示了不同孪生对比设置的影响。当我们对真实标注执行一对一孪生对比时，如 [10], it (a)，平均精度均值为25.6%，Top-1准确率为73.2%。此外，我们对所有正样本执行一对一孪生对比，it (b) 的平均精度均值为31.2%，Top-1准确率为79.7%。通过使用更多样化的训练样本，该方法显著优于基线。当我们在所有正样本及其对应的真实标注之间执行多对一孪生对比时，it (c) 达到最佳性能。与策略 (b) 相比，策略 (c) 更适合学习尺度-

表 V 掩码策略对遮挡不变对比（OIC）模块中不同分支的影响。

Num.	掩码策略	mAP	Top-1准确率
(a)	None	33.2	79.9
(b)	搜索分支	**35.5**	**80.9**
(c)	实例分支	33.1	79.5
(d)	两个分支	32.7	80.1

表 VI 掩码网格对我们的遮挡不变对比（OIC）模块的影响。

网格数量	0	2	4	6
mAP	33.2	35.5	35.4	33.0
Top-1准确率	79.9	80.9	80.3	80.0

不变且定位不变的重识别特征。

OIC中的掩码策略的影响表 V展示了掩码策略对孪生网络不同分支的影响。仅在搜索分支上进行掩码时，性能最佳。在实例分支上或两个分支上同时采用掩蔽策略，都无法提升平均精度均值。我们认为，这是因为实例分支旨在引导重识别特征保持一致，而在实例分支上采用掩码策略会使一致特征难以学习。我们还展示了掩码网格的影响，见表VI。我们观察到，掩蔽2个网格时性能最佳。

为展示我们的 OIC 处理遮挡变化的能力，我们通过随机掩码测试图像中 20% 的像素生成了一个掩码测试集。表 VII 比较了在原始测试集和掩码测试集上的性能。使用掩码测试集时，所有方法的性能均有所下降。我们观察到，使用我们的 OIC 模块时性能下降较小。这表明，我们的遮挡不变对比能够更好地处理遮挡变化。

D. 与最先进方法的比较

我们将所提出的 DICL 与最先进的方法进行比较，包括全监督和弱监督方法。这些方法可分为两阶段方法和单阶段方法。

在 CUHK-SYSU 上的比较在表 VIII 中，我们将 DICL 与一些最先进的方法进行比较，图库规模为 100，其中包括全监督和弱监督方法。我们的方法 DICL 达到 87.4% 的平均精度均值和 88.8% 的 Top-1准确率，优于当前的弱监督单步行人搜索方法。CGPS [11] 的平均精度均值为 80.0%，Top-1准确率为 82.3%；R-SiamNet [10] 的平均精度均值为 86.0%，Top-1准确率为 87.1%。我们提出的方法在平均精度均值上比 CGPS 高 7.4%，在 Top-1准确率上高 6.5%；在平均精度均值上比 R-SiamNet 高 1.4%，在 Top-1准确率上高 1.7%。此外，我们的方法优于一些全监督单阶段方法，例如 RCAA [52]和 CTXG[53]。弱监督两阶段 CGUA [31] 的平均精度均值为 91.0%，Top-1准确率为 92.2%，其使用

表 VII 原始和掩码 PRW 测试集上有无遮挡不变对比（OIC）模块的比较。

测试集	不使用 OIC		使用 OIC	
	平均精度均值	Top-1准确率	平均精度均值	Top-1准确率
原始	33.2	79.9	35.5	80.9
掩码	15.0	60.2	20.9	67.6
△（丢弃）	18.2	19.7	14.6	13.3

表 VIII CUHK-SYSU 和 PRW 测试集上全监督与弱监督方法的比较。

	方法	CUHK-SYSU		PRW	
		mAP	Top-1准确率	mAP	Top-1准确率
Two-step	全监督：				
	IDE [14]	-	-	20.5	48.3
	MGTS [17]	83.0	83.7	32.6	72.1
	CLSA [47]	87.2	88.5	38.7	65.0
	RDLR [48]	93.0	94.2	42.9	70.2
	IGPN [49]	90.3	91.4	47.2	87.0
	TCTS [18]	93.9	95.1	46.8	87.5
	弱监督：				
	CGUA [31]	91.0	92.2	42.7	86.9
One-step	全监督：				
	OIM [5]	75.5	78.7	21.3	49.4
	IAN [50]	76.3	80.1	23.0	61.9
	NPSM [51]	77.9	81.2	24.2	53.1
	RCAA [52]	79.3	81.3	-	-
	CTXG [53]	84.1	86.5	33.4	73.6
	QEEPS [54]	88.9	89.1	37.1	76.7
	HOIM [13]	89.7	90.8	39.8	80.4
	BINet [26]	90.0	90.7	45.3	81.7
	NAE [22]	91.5	92.4	43.3	80.9
	PGA [55]	90.2	91.8	42.5	83.5
	AlignPS [12]	93.1	93.4	45.9	81.9
	DMRNet [24]	93.2	94.2	46.9	83.3
	AGWF [27]	93.3	94.2	53.3	87.7
	SeqNet [23]	93.8	94.6	46.7	83.4
	BUFF [56]	91.6	92.2	44.9	86.3
	OIMNet++[28]	93.1	94.1	47.7	84.8
	PSTR [9]	93.5	95.0	49.5	87.8
	COAT [30]	94.2	94.7	53.3	87.4
	弱监督：				
	CGPS [11]	80.0	82.3	16.2	68.0
	R-SiamNet [10]	86.0	87.1	21.2	73.4
	DICL（我们的方法）	**87.4**	**88.8**	**35.5**	**80.9**

使用两个独立网络分别进行检测和再识别，并侧重于基于记忆的对比。与弱监督两阶段 CGUA 相比，我们的单阶段 DICL 仅使用单一网络进行检测和再识别，并侧重于深层图像内对比。因此，我们的 DICL 是一个用于端到端行人搜索的简单框架。

图 5 进一步展示了 [50, 100, 500, 1000, 2000, 4000] 在不同图库规模下的性能。结果表明，我们的方法优于弱监督单步方法，尤其是在图库规模较大时，这说明我们的方法能够在

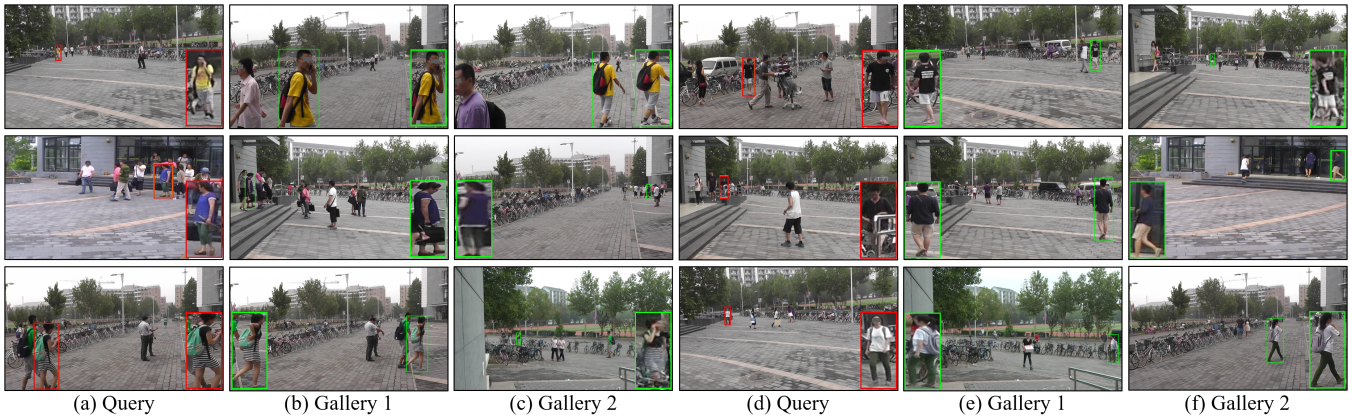

图3. 我们的方法在 PRW 测试集上的定性结果，其中红框表示查询目标，绿框表示图库图像中的搜索结果。我们的方法能够找到具有不同视角和尺度的查询行人。



图4. 我们的方法在 CUHK-SYSU 测试集上的定性结果，其中红框表示查询目标，绿框表示图库图像中的搜索结果。我们的方法能够在不同场景中匹配查询行人。

规模大且具有挑战性的场景中搜索目标。

在 PRW 上的比较我们还在表 VIII 中将我们的方法与 PRW 测试集上的先进方法进行了比较。与 CUHK-SYSU 相比，PRW 测试集的图库规模更大（6,112），更具挑战性。我们的方法取得了 35.5% 的平均精度均值和 80.9% 的 Top-1准确率，在平均精度均值上比 CGPS 高 19.3%，在 Top-1准确率上高 12.9%；在平均精度均值上比 R-SiamNet 高 14.3%，在 Top-1 准确率上高 7.5%。此外，我们的方法在平均精度均值上分别比全监督方法 MGTS [52]和 CTXG [53] 高 2.9% 和 2.1%。

定性结果 图 3 和图 4 展示了我们提出的 DICL 在 PRW 和 CUHK-SYSU 测试集上的一些定性结果。我们的方法能够成功检测并识别出不同图库图像中的查询行人，这些行人的姿态、尺度、背景和可见性各不相同。例如，如图 3 最后一行所示，女孩的位置、尺度和可见部分都发生了很大变化。我们还发现，如图 6 所示，当行人外观相似、遮挡严重或受到附近行人干扰时，我们的方法偶尔会遇到困难。因此，未来有必要解决这些问题。

V. 结论

本文提出了一种用于弱监督单步行人搜索的新型深度图像内对比学习（DICL）框架。我们采用了孪生网络

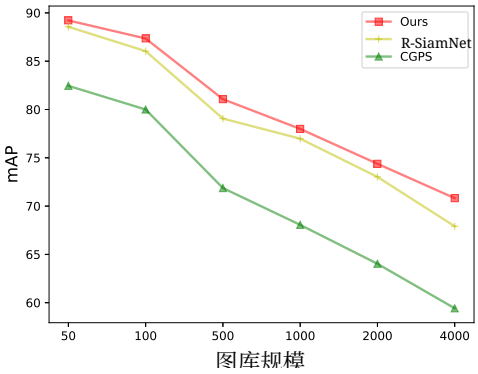

图 5. CUHK-SYSU 测试集上不同图库规模下各种弱监督单步方法的比较。

通过搜索分支和实例分支学习具有判别性的特征，以应对空间层面和遮挡层面的变化。具体而言，我们引入了两种新颖的图像内对比模块：空间不变对比（SIC）和遮挡不变对比（OIC）。SIC 在孪生网络的两个分支之间进行多对一孪生对比，并对搜索分支中的所有预测进行密集对比，从而学习具有判别性的尺度不变和位置不变的重识别特征。OIC 在



图6. 我们的方法的一些失败案例。我们发现，在严重遮挡、极低光照条件等情况下，检测和识别人物仍然具有挑战性。

搜索分支中执行掩码策略，并提升特征对遮挡变化的一致性。在两个数据集上进行的大量实验验证了我们方法的有效性。

目前，我们的方法主要关注深度图像内对比，并未考虑改进基于记忆的对比。未来，我们将探索如何改进用于基于记忆的对比的无监督聚类。

参考文献

[1] H. Yao 和 C. Xu，《用于行人搜索的联合行人客观性与排斥》，*IEEE*图像处理汇刊，第30卷，696页，2021年 685页。–。[2] D. Chen、S. Zhang、W. Ouyang、J. Yang 和 Y. Tai，《基于分离建模和掩码引导双流CNN模型的行人搜索》，*IEEE*图像处理汇刊，第29卷，第5669–4682页，2020年。[3] C. Gao、G. Cai、X. Jiang、F. Zheng、J. Zhang、Y. Gong、F. Lin、X. Sun 和 X. Bai，《基于条件特征学习的文本行人搜索变换器》，*IEEE*图像处理汇刊，第31卷，第6097–6108页，2022年。[4] Y. Chen、R. Huang、H. Chang、C. Tan、T. Xue 和 B. Ma，《面向语言行人搜索的跨模态知识适应》，*IEEE*图像处理汇刊，第30卷，第4057–4069页，2021年。[5] T. Xiao、S. Li、B. Wang、L. Lin 和 X. Wang，《用于行人搜索的联合检测与识别特征学习》，载于*IEEE*计算机视觉与模式识别会议，2017年，第3376–3385页。[6] J. Cao、Y. Pang、J. Xie、F. S. Khan 和 L. Shao，《从手工特征到深度特征的行人检测：综述》，*IEEE*模式分析与机器智能汇刊，第44卷，4934页，2022年 9，4913号，第。–。[7] M. Ye、J. Shen、G. Lin、T. Xiang、L. Shao 和 S. C. H. Hoi，《行人重识别的深度学习：综述与展望》，*IEEE*模式分析与机器智能汇刊，第44卷，第6期，第2872–2893页，2022年。[8] C. Zhao、Z. Chen、S. Dou、Z. Qu、J. Yao、J. Wu、和 D. Miao，《面向抗噪行人搜索的上下文感知特征学习》，*IEEE*视频技术电路与系统汇刊，第32卷，第10期，第7047–7060页，2022年。[9] J. Cao、Y. Pang、R. M. Anwer、H. Cholakkal、J. Xie、M. Shah 和 F. S. Khan，《PSTR：基于变换器的端到端单步行人搜索》，载于*IEEE*计算机视觉与模式识别会议，2022年，第9458–9467页。[10] C. Han、K. Su、D. Yu、Z. Yuan、C. Gao、N. Sang、Y. Yang 和 C. Wang，《基于区域孪生网络的弱监督行人搜索》，载于*IEEE*国际计算机视觉会议，2021年，第11 986–11 995页。[11] Y. Yan、J. Li、S. Liao、J. Qin、B. Ni、X. Yang 和 L. Shao，《探索弱监督行人搜索的视觉上下文》，载于*AAAI*人工智能会议，2022年，第3027–3035页。[12] Y. Yan、J. Li、J. Qin、S. Bai、S. Liao、L. Liu、F. Zhu 和 L. Shao，《无锚框行人搜索》，载于*IEEE*计算机视觉与模式识别会议，2021年，第7690–7699页。[13] D. Chen、S. Zhang、W. Ouyang、J. Yang 和 B. Schiele，《用于行人搜索的分层在线实例匹配》，载于*AAAI*人工智能会议，2020年，第10 518–10 525页。[14] L. Zheng、H. Zhang、S. Sun、M. Chandraker、Y. Yang 和 Q. Tian，《野外环境中的行人重识别》，载于IEEE计算机视觉与模式识别会议，2017年，第3346–3355页。[15] X. Ke、H. Liu、W. Guo、B. Chen、Y. Cai 和 W. Chen，《用于高效行人搜索的联合样本增强与实例敏感特征学习》，IEEE视频技术电路与系统汇刊，第32卷，第11期，第7924–7937页，2022年。[16] W. Yang、H. Huang、X. Chen 和 K. Huang，《面向实用行人搜索的自底向上前景感知特征融合》，IEEE视频技术电路与系统汇刊，第32卷，274页，2022年 1，262号，第。–。[17] D. Chen、S. Zhang、W. Ouyang、J. Yang 和 Y. Tai，《基于掩码引导双流CNN模型的行人搜索》，载于欧洲计算机视觉会议，2018年，第764–781页。[18] C. Wang、B. Ma、H. Chang、S. Shan 和 X. Chen，《TCTS：用于行人搜索的任务一致性两阶段框架》，载于IEEE计算机视觉与模式识别会议，2020年，第11 949–11 958页。[19] C. Wang、B. Ma、H. Chang、S. Shan 和 X. Chen，《基于双向任务一致性学习模型的行人搜索》，IEEE多媒体汇刊，2021年。[20] S. Ren、K. He、R. B. Girshick 和 J. Sun，《Faster R-CNN：借助区域建议网络实现实时目标检测》，IEEE模式分析与机器智能汇刊，第39卷，第6期，第1137–1149页，2017年。[21] T.-Y. Lin、P. Doll´ar、R. Girshick、K. He、B. Hariharan 和 S. Belongie，《用于目标检测的特征金字塔网络》，载于IEEE计算机视觉与模式识别会议，2017年，第2117–2125页。[22] D. Chen、S. Zhang、J. Yang 和 B. Schiele，《用于高效行人搜索的范数感知嵌入》，载于IEEE计算机视觉与模式识别会议，2020年，第12 612–12 621页。[23] Z. Li 和 D. Miao，《用于高效行人搜索的序列端到端网络》，载于AAAI人工智能会议，2021年，第2011–2019页。[24] C. Han、Z. Zheng、C. Gao、N. Sang 和 Y. Yang，《解耦与记忆增强网络：迈向有效的一步行人搜索特征学习》，载于AAAI人工智能会议，2021年，第1505–1512页。[25] C. Han、Z. Zheng、K. Su、D. Yu、Z. Yuan、C. Gao、N. Sang 和 Y. Yang，《DMRNet++：通过解耦网络和增强配对样本学习判别性特征，实现一步行人搜索》，IEEE模式分析与机器智能汇刊，2022年。[26] W. Dong、Z. Zhang、C. Song 和 T. Tan，《用于行人搜索的双向交互网络》，载于IEEE计算机视觉与模式识别会议，2020年，第2836–2845页。[27] B.-J. Han、K. Ko 和 J.-Y. Sim，《利用自适应梯度传播的端到端可训练三叉戟行人搜索网络》，载于IEEE国际计算机视觉会议，2021年，第925–933页。[28] S. Lee、Y. Oh、D. Baek、J. Lee 和 B. Ham，《Oimnet++：用于行人搜索的原型归一化与定位感知学习》，载于欧洲计算机视觉会议，2022年，第621–637页。[29] Z. Tian、C. Shen、H. Chen 和 T. He，《FCOS：全卷积单阶段目标检测》，载于IEEE国际计算机视觉会议，2019年，第9626–9635页。[30] R. Yu、D. Du、R. LaLonde、D. Davila、C. Funk、A. Hoogs 和 B. Clipp，《用于端到端行人搜索的级联Transformer》，载于IEEE计算机视觉与模式识别会议，2022年，第7267–7276页。[31] C. Jia、M. Luo、C. Yan、X. Chang 和 Q. Zheng，《CGUA：上下文引导与非配对辅助弱监督行人搜索》，arXiv:2203.14307，2022年。[32] M. Ester、H.-P. Kriegel、J. Sander 和 X. Xu，《一种在大型空间数据库中发现含噪聚类簇的基于密度的算法》，载于知识发现与数据挖掘国际会议，1996年，第226–231页。[33] M. S. Sarfraz、V. Sharma 和 R. Stiefelhagen，《利用第一邻居关系进行高效的无参数聚类》，载于IEEE计算机视觉与模式识别会议，2019年，第8934–8943页。[34] Y. Lin、X. Dong、L. Zheng、Y. Yan 和 Y. Yang，《一种用于无监督行人重识别的自底向上聚类方法》，载于AAAI人工智能会议，第33卷，2019年，第8738–8745页。[35] Y. Ge、F. Zhu、D. Chen、R. Zhao 和 H. Li，《利用混合记忆进行域自适应目标重识别的自步对比学习》，载于神经信息处理系统国际会议，2020年，第11 309–11 321页。[36] W. Deng、L. Zheng、Q. Ye、G. Kang、Y. Yang 和 J. Jiao，《通过保留自相似性和域不相似性实现行人重识别的图像到图像域适应》，载于IEEE计算机视觉与模式识别会议

，2018，第994–1003页。

[37] J. Liu, Z.-J. Zha, D. Chen, R. Hong, 和 M. Wang, 《用于跨域行人重识别的自适应迁移网络》，载于*IEEE*计算机视觉与模式识别会议，2019年，第7202–7211页。 [38] J. Li, Y. Yan, G. Wang, F. Yu, Q. Jia, 和 S. Ding, 《域自适应行人搜索》，载于欧洲计算机视觉会议，2022年。 [39] T. Chen, S. Kornblith, M. Norouzi, 和 G. Hinton, 《视觉表征对比学习的简单框架》，载于国际机器学习会议，2020年，第1575–1585页。 [40] J.-B. Grill, F. A. Florian Strub, C. Tallec, P. Richemond, E. Buchatskaya, B. A. P. Carl Doersch, Z. Guo, M. G. Azar, K. K. Bilal Piot, R. Munos, 和 M. Valko, 《自举潜在表征——一种自监督学习的新方法》，载于神经信息处理系统国际会议，2020年，第21 271–21 284页。 [41] X. Chen 和 K. He, 《探索简单的孪生表征学习》，载于*IEEE*计算机视觉与模式识别会议，2021年，第15 745–53页。 [42] X. Peng, K. Wang, Z. Zhu, M. Wang, 和 Y. You, 《为孪生表征学习设计更优的对比视图》，载于*IEEE*计算机视觉与模式识别会议，2022年，第16 031–16 040页。 [43] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, 和 Y. Wei, 《可变形卷积网络》，载于*IEEE*国际计算机视觉会议，2017年，第764–773页。 [44] A. Hermans, L. Beyer, 和 B. Leibe, 《为行人重识别中的三元组损失辩护》，*arXiv:1703.07737*，2017年。 [45] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, C. Z. Dazhi Cheng, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, YueWu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, 和 D. Lin, 《Mmdetection：OpenMMLab检测工具箱与基准》，2019年。[46] J. Deng, W. Dong, R. Socher, L. Li, K. Li, 和 F. Li, 《ImageNet：大型层次化图像数据库》，载于*IEEE*计算机视觉与模式识别会议，2009年，第248–255页。 [47] X. Lan, X. Zhu, 和 S. Gong, 《通过多尺度匹配进行行人搜索》，载于欧洲计算机视觉会议，第11205卷，2018年，第553–569页。[48] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, 和 N. Sang, 《再识别驱动的行人搜索定位细化》，载于*IEEE*国际计算机视觉会议，2019年，第9813–9822页。 [49] W. Dong, Z. Zhang, C. Song, 和 T. Tan, 《用于行人搜索的实例引导提议网络》，载于*IEEE*计算机视觉与模式识别会议，2020年，第2582–2591页。 [50] J. Xiao, Y. Xie, T. Tillo, K. Huang, Y. Wei, 和 J. Feng, 《IAN：用于行人搜索的个体聚合网络》，模式识别，第87卷，第332–340页，2019年。 [51] H. Liu, J. Feng, Z. Jie, J. Karlekar, B. Zhao, M. Qi, J. Jiang, 和 S. Yan, 《神经行人搜索机》，载于*IEEE*国际计算机视觉会议，2017年，第493–501页。 [52] X. Chang, P. Huang, Y. Shen, X. Liang, Y. Yang, 和 A. G. Hauptmann, 《RCAA：用于行人搜索的关系上下文感知智能体》，载于欧洲计算机视觉会议，2018年，第86–102页。 [53] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, 和 X. Yang, 《学习用于行人搜索的上下文图》，载于*IEEE*计算机视觉与模式识别会议，2019年，第2158–2167页。 [54] B. Munjal, S. Amin, F. Tombari, 和 F. Galasso, 《查询引导的端到端行人搜索》，载于*IEEE*计算机视觉与模式识别会议，2019年，第811–820页。 [55] H. Kim, S. Joung, I.-J. Kim, 和 K. Sohn, 《用于行人搜索的原型引导显著性特征学习》，载于*IEEE*计算机视觉与模式识别会议，2021年，第4865–4874页。 [56] W. Yang, H. Huang, X. Chen, 和 K. Huang, 《面向实用行人搜索的自底向上前景感知特征融合》，*IEEE*视频技术电路与系统汇刊，第21卷，274，2022年1，262第。– 。