网络出版时间：2019-01-22 09:19:42
网络出版地址：http://kns.cnki.net/kcms/detail/11.2109.TP.20190122.0919.007.html

基于深度学习的行人重识别研究进展

罗 浩 [1] 姜 伟 [1] 范 星 [1] 张思朋 [1]

摘 要 行人重识别是计算机视觉领域近年来非常热的一个研究课题, 可以被视为图像检索的一个子问题, 其目标是给定一个监控行人图像检索跨设备下的该行人图像. 传统的方法依赖手工特征, 不能适应数据量很大的复杂环境. 近年来随着深度学习的发展, 大量基于深度学习的行人重识别方法被提出. 本文先简单介绍了该问题的定义及传统方法的局限, 并列举了一些适用于深度学习方法的行人重识别数据集. 此外我们详细地总结了一些比较典型的基于深度学习的行人重识别方法, 并比较了部分算法在 Market1501 数据集上的性能表现. 最后我们对该问题未来的研究方向做了一个展望.

关键词 行人重识别, 深度学习, 计算机视觉, 综述, 卷积神经网络

引用格式 罗浩, 姜伟, 范星, 张思朋. 基于深度学习的行人重识别研究进展. 自动化学报, 201X, **XX**(X): X−X

DOI 10.16383/j.aas.c180154

A Survey on Deep Learning Based Person Re-Identification

LUO Hao[1] JIANG Wei[1] FAN Xing[1] ZHANG Si-Peng[1]

Abstract Person re-identification (ReID) is a popular research topic in computer vision. It aims to retrieve the given pedestrian image across the device, which can be regarded as a sub-problem of image retrieval. The traditional methods rely on hand-crafted features and can not adapt to the complicated environment with large amount of data. In recent years, with the development of deep learning, a large number of ReID methods based on deep learning have been proposed. This paper briefly introduces the definition of the problem and the limitations of the traditional methods, and then lists some popular databases suitable for deep learning. What's more, we summarize some typical deep learning based methods in detail, and compare the performance of some algorithms on Market1501. Finally, we make a prospect for the future research direction of person ReID.

Key words Person re-identification, deep learning, computer vision, review, convolutional neural networks

Citation Luo Hao, Jiang Wei, Fan Xing, Zhang Si-Peng. A survey on deep learning based person re-identification. *Acta Automatica Sinica*, 201X, **XX**(X): X−X

行人重识别 (Person re-identification) 也称行人再识别, 被广泛认为是一个图像检索的子问题, 是利用计算机视觉技术判断图像或者视频中是否存在特定行人的技术, 即给定一个监控行人图像检索跨设备下的该行人图像. 行人重识别技术可以弥补目前固定摄像头的视觉局限, 并可与行人检测、行人跟踪技术相结合, 应用于视频监控、智能安防等领域.

在深度学习技术出现之前, 早期的行人重识别研究主要集中于如何手工设计更好的视觉特征和如何学习更好的相似度度量. 近几年随着深度学习的发展, 深度学习技术在行人重识别任务上得到了广泛的应用. 和传统方法不同, 深度学习方法可以自动提取较好的行人图像特征, 同时学习得到较好的相似度度量. 当然深度学习相关的行人重识别方法也经历了一个从简单到复杂的发展过程. 起初研究者主要关注用网络学习单帧图片的全局特征, 根据损失类型的不同可以分为表征学习 (Representation learning) 和度量学习 (Metric learning) 方法. 而单帧图片的全局特征遇到性能瓶颈之后, 研究者引入局部特征和序列特征进一步发展行人重识别研究. 最近因为生成对抗网络 (Generative adversarial nets, GAN)[1] 的逐渐成熟, 一些基于 GAN 的行人重识别研究工作表明: GAN 在扩充数据集、解决图片间的偏差等问题上也有不错的效果. 虽然目前大量工作仍然是属于监督学习 (Supervised learning) 的范畴, 但是迁移学习、半监督学习和无监督学习也同样是一个值得研究的方向.

本文分析了近几年深度学习相关的方法在行人重识别问题上的发展, 归纳整合了该领域的一些优秀算法, 并探讨了未来可能的研究焦点.

收稿日期 2018-03-19 录用日期 2018-09-14
Manuscript received March 19, 2018; accepted September 14, 2018
国家自然科学基金 (61375049, 61633019), 浙江省基础公益研究计划项目 (LGF18F030002) 资助
Supported by National Natural Science Foundation of China (61375049,61633019), Zhejiang Basic Public Welfare Research Project (LGF18F030002)
本文责任编委 赖剑煌
Recommended by Associate Editor LAI Jian-Huang
1. 浙江大学智能系统与控制研究所 杭州 310027
1. Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027

本文剩余内容安排如下: 第 1 节简要回顾了行人重识别的发展历程. 第 2 节介绍一些常见的行人重识别数据集, 并分析一下主要的研究难点. 第 3 节着重介绍近几年比较典型的基于深度学习的行人重识别方法, 并按照一定的发展历程进行归纳整合. 第 4 节我们展示目前一些方法在主流数据集上的性能表现. 第 5 节简要探讨一下未来可能的研究焦点.

1 行人重识别发展简述

行人重识别可以应用到刑事侦查、视频监控、行为理解等多个方面, 但据我们所知, 其在学术界的研究最先追溯到跨摄像头多目标跟踪 (Multi-target multi-camera tracking, MTMC tracking) 问题上. 早在 2005 年, 文献 [2] 探讨了在跨摄像头系统中, 当目标行人在某个相机视野中丢失之后如何将其轨迹在其他相机视野下再次关联起来的问题. 该文献利用一个贝叶斯网络根据行人特征 (颜色、时空线索) 的相似度将行人轨迹关联起来. 而如何提取行人特征以及如何进行特征相似度度量就是行人重识别需要解决的核心问题, 也可以合称为行人跨摄像头检索. 因此行人重识别被研究者从 MTMC 跟踪问题里抽取出来, 作为一个独立的研究课题. 行人重识别领域知名学者郑良博士在论文 [3] 中将行人重识别系统总结为行人检测加上行人重识别, 如图 1 所示. 随着深度学习的发展, 行人检测技术已逐渐成熟, 本文不再做具体阐述. 目前大部分数据集直接将检测出来的行人图片作为训练集和测试集, 并且剔除了一些遮挡较严重的低质量图片. 行人重识别技术将行人检测结果作为先验知识, 直接对行人图片进行跨摄像头检索.

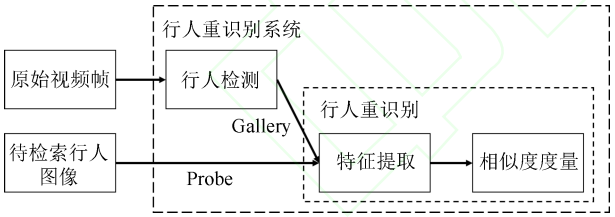


图 1 行人重识别系统

Fig. 1 Person ReID system

行人重识别任务主要包含特征提取和相似度度量两个步骤. 传统的方法思路为手工图像特征, 例如颜色、HOG (Histogram of oriented gradient)[4]、SIFT (Scale invariant feature transform)[5]、LOMO (Local Maximal Occurrence) 等. 之后, 利用 XQDA (Cross-view Quadratic Discriminant Analysis)[6] 或者 KISSME (Keep It Simple and Straightforward Metric Learning)[7] 来学习最佳的相似度度量. 然而, 传统的手工特征描述能力有限, 很难适应复杂场景下的大数据量任务. 并且, 在数据量较大的情形下, 传统的度量学习方法求解也会变得非常困难.

近年来, 以卷积神经网络为代表的深度学习在计算机视觉领域取得了极大的成功, 在多项任务上都击败传统的方法, 甚至一定程度上超越了人类的水平[8−9]. 在行人重识别问题上, 基于深度学习的方法可以自动学习出复杂的特征描述, 并且用简单的欧式距离进行相似度度量便可以取得很好的性能. 换句话说, 深度学习可以端对端地实现行人重识别任务, 这使得任务变得更加简单. 目前, 基于深度学习的行人重识别方法已经在性能上大大超越了传统的方法. 这些优势使得深度学习在行人重识别领域变得流行, 大量相关研究工作发表在高水平的会议或者期刊上, 行人重识别的研究也进入了一个新的阶段.

2 相关数据集介绍

由于 CNN 网络的训练依赖大量训练数据, 所以行人重识别研究从传统的手工特征 (Hand-crafted feature) 方法发展为如今深度学习自提特征的方法, 离不开大规模数据集的发展. 近年来, 涌现出了越来越多的大规模行人重识别数据集, 数据集特点也各自不同, 这也反映了该领域蓬勃的发展趋势和巨大的现实需求.

目前常用于深度学习方法的行人重识别数据集有:

1) VIPeR[10] 数据集是早期的一个小型行人重识别数据集, 图像来自 2 个摄像头. 该数据集总共包含 632 个行人的 1 264 张图片, 每个行人有两张不同摄像头拍摄的图片. 数据集随机分为相等的两部分, 一部分作为训练集, 一部分作为测试集. 由于采集时间较早, 该数据集的图像分辨率非常低, 所以识别难度较大.

2) PRID2011[11] 是 2011 年提出的一个数据集, 图像来自于 2 个不同的摄像头. 该数据集总共包含 934 个行人的 24 541 张行人图片, 所有的检测框都是人工手动提取. 图像大小的分辨率统一为 128×64 的分辨率.

3) CUHK03[12] 在香港中文大学采集, 图像来自 2 个不同的摄像头. 该数据集提供机器自动检测和手动检测两个数据集. 其中检测数据集包含一些检测误差, 更接近实际情况. 数据集总共包括 1 467 个行人的 14 097 张图片, 平均每个人有 9.6 张训练数据.

4) Market1501[13] 是在清华大学校园中采集, 图像来自 6 个不同的摄像头, 其中有一个摄像头为低分辨率. 同时该数据集提供训练集和测试集. 训练

集包含 12 936 张图像, 测试集包含 19 732 张图像. 图像由检测器自动检测并切割, 所以包含一些检测误差 (接近实际使用情况). 训练数据中一共有 751 人, 测试集中有 750 人. 所以在训练集中, 平均每类 (每个人) 有 17.2 张训练数据.

5) CUHK-SYSU[14] 是香港中文大学和中山大学一起收集的数据集. 该数据集的特点是提供整个完整的图片, 而不像其他大部分数据集一样只提供自动或者手动提取边框 (bounding box) 的行人图片, 图片来源于电影和电视. 该数据集总共包括 18 184 张完整图片, 内含 8 432 个行人的 99 809 张行人图片. 其中训练集有 11 206 张完整图片, 包含 5 532 个行人. 测试集有 6 978 张完整图片, 包含 2 900 个行人.

(6) MARS[15] 数据集是 Market1501 的扩展. 该数据集的图像由检测器自动切割, 包含了行人图像的整个跟踪序列 (tracklet). MARS 总共提供 1 261 个行人的 20 715 个图像序列, 和 Market1501 一样来自同样的 6 个摄像头. 和其他单帧图像数据集不一样的地方是, MARS 是提供序列信息的大规模行人重识别数据集. 特别注意的是, MARS 和 Market1501 的训练集和测试集存在重叠, 因此不能够混在一起训练网络.

(7) DukeMTMC-reID[16] 在杜克大学内采集, 图像来自 8 个不同摄像头, 行人图像的边框由人工标注完成. 该数据集提供训练集和测试集. 训练集包含 16 522 张图像, 测试集包含 17 661 张图像. 训练数据中一共有 702 人, 平均每个人有 23.5 张训练数据. 该数据集是 ICCV2017 会议之前最大的行人重识别数据集, 并且提供了行人属性 (性别/长短袖/是否背包等) 的标注.

除了以上几个已经开源的常用数据集以外, 目前还有几个比较新的数据集, 其中比较典型的有: 1) 中山大学采集的红外 ReID 数据集 SYSU-MM01[17], 可以实现夜间的行人重识别. 2) 北航大学等采集的 LPW 数据集[18], 包含 2 731 个行人的 7 694 个轨迹序列, 总共有 56 万多张图片, 该数据集的特点是有多个独立的场景, 每个场景都可以作为一个独立的数据集, 训练集和测试集按照场景分开, 因此更加接近真实使用情况. 3) 北京大学采集的 MSMT17 数据集[19], 包含室内室外 15 个相机的 12 万多张行人图片, 有 4 千多个行人 ID, 是目前最大的单帧 ReID 数据集. 4) 北京大学和微软研究院联合采集的 LVreID 数据集[20], 包含室内室外 15 个相机的 3 千多个行人 ID 的序列图片, 总共 14 943 个序列的 3 百多万张图片, 尚未开放下载链接.

以上数据集的细节可以在表 1 中查阅, 其中大部分数据集使用 Deformable Part-based Model(DPM) 或者手动标注的方法[21] 检测行人, 两个还未开放下载的同源数据集 MSMT17 和 LVreID 使用了最新的 Faster RCNN 检测器[22], MARS 在提取序列的时候还辅助了 Generalized Maximum Multi Clique problem (GMMCP) 跟踪器[23]. 几乎目前主流的数据集都使用累计匹配 (Cumulative Match Characteristics, CMC) 曲线和平均准确度 (Mean Average Precision, mAP) 准确度评估. 由于 ReID 的数据集数目繁多, 本文也只能列举一些比较常用的典型数据集, 更多数据集的信息可以查阅文献[24].

图 2 展示了一些行人重识别数据集的图片, 从图中可以看出, 行人重识别是一个非常有挑战性的问题. 其中最主要的难点主要有: 不同行人之间的外观可能高度相似, 而相同的行人在不同的时空下姿态也可能不同, 行人主体遭遇遮挡以及不同相机拍摄的光线条件差异等. 这些难点也使得行人重识别和一般的图像检索问题有所不同, 目前深度学习的方法除了扩大训练数据和改善网络结构以外, 也会针对于这些难点设计专用于 ReID 任务的算法.

表 1 典型行人重识别数据集

Table 1 Typical ReID datasets

数据集	发布时间	ID 数	图片数	序列数	室内相机	室外相机	检测器	评估
ViPeR	2007	632	1264	×	0	2	手动	CMC
PRID2011	2011	934	24 541	400	0	2	手动	CMC
CUHK03	2014	1 467	13 164	×	10	0	手动 +DPM	CMC+mAP
Market1501	2015	1 501	32 217	×	0	6	手动 +DPM	CMC+mAP
CUHK-SYSU	2016	8 432	99 809	×	0	0	DPM	CMC+mAP
MARS	2016	1 261	1 119 003	20 715	0	6	DPM+GMMCP	CMC+mAP
DukeMTMC-reID	2017	1 812	36 441	×	0	8	手动	CMC+mAP
SYSU-MM01	2017	491	287 628	×	3	3	未知	CMC+mAP
LPW	2018	2 731	590 000+	7 694	0	11	手动 +DPM	CMC+mAP
MSMT17	2018	4 101	126 441	×	3	12	Faster RCNN	CMC+mAP
LVreID	即将发布	3 772	2 989 436	14 943	3	12	Faster RCNN	CMC+mAP

图 2 行人重识别数据集图片及难点示例

Fig. 2 The examples of images and challenge of person ReID datasets

3 行人重识别深度学习方法

本小节总结概述基于深度学习的行人重识别方法. 该类方法根据训练损失可以分为基于表征学习和度量学习, 根据特征是否考虑局部特征可以分为基于全局特征和基于局部特征, 根据数据不同可以分为基于单帧图像和基于视频序列的方法. 此外, 还有一类基于 GAN 的方法利用 GAN 生成数据来解决一些行人重识别的难点. 在本小节的最后, 我们还总结概述了一下这些方法的优缺点以及如何结合这些方法来实现一个更好的行人重识别算法.

3.1 基于表征学习的方法

基于表征学习 (Representation learning) 的方法是一类非常常用的行人重识别方法[3,25−31]. 虽然行人重识别的最终目标是为了学习出两张图片之间的相似度, 但是表征学习的方法并没有直接在训练网络的时候考虑图片间的相似度, 而把行人重识别任务当做分类 (Classification) 问题或者验证 (Verification) 问题来看待. 这类方法的特点就是网络的最后一层全连接 (Fully connected, FC) 层输出的并不是最终使用的图像特征向量, 而是经过一个 Softmax 激活函数来计算表征学习损失, 前一层 (倒数第二层)FC 层通常为特征向量层. 具体言之, 分类问题是指利用行人的 ID 或者属性等作为训练标签来训练模型, 每次只需要输入一张图片; 验证问题是指输入一对 (两张) 行人图片, 让网络来学习这两张图片是否属于同一个行人.

分类网络常用的两种损失分别是行人 ID 损失 (Identification loss) 和属性损失 (Attribute loss). 文献[3,29] 将每一个行人当做分类问题的一个类别, 用行人的 ID 作为训练数据的标签来训练 CNN 网络, 这个网络损失被称为 ID 损失, 而这种网络被称为 IDE(ID embedding) 网络. IDE 网络是行人重识别领域非常重要的 baseline 基准. 假设训练集拥有 K 个行人的 n 张图片, 将图片 x 输入 IDE 网络 f, 网络最后一层输出该图片的 ID 预测向量 $z = [z_1, z_2, \cdots, z_k] \in R^K$. 因此, 图片 x 属于第 $k, k \in 1, 2, 3, \cdots, K$ 个行人 ID 的概率为 $p(k|x) = \frac{exp(z_k)}{\sum_{i=1}^{K} exp(z_i)}$. 为了方便描述, 本文此后忽略 k 和 x 的相关性, 默认用 $p(k)$ 来代表 $p(k|x)$. 于是 IDE 网络的 ID 损失为:

$$L_{ID}(f, x) = -\sum_{k=1}^{K} q(k) \log p(k) \tag{1}$$

其中 $q(k)$ 通过图片 x 的 ID 标签的到, 若图片 x 的 ID 标签为 y, 则 $q(k) = 1, y = k$, 而对于任何的 $y \neq k$ 都有 $q(k) = 0$.

后来部分研究者认为, 光靠行人的 ID 信息不足以学习出一个泛化能力足够强的模型. 因此, 他们利用了额外标注的行人图片的属性信息, 例如性别、头发、衣着等属性, 通过引入行人属性标签计算属性损失. 训练好的网络不但要准确地预测出行人 ID, 还要预测出各项行人属性, 这大大增加了网络的泛化能力, 多数论文也显示这种方法是有效的[3,26−27]. 图 3 是其中一个示例, 从图中可以看出, 网络输出的特征后面引出两个分支. 一个分支用于计算 ID 损失 L_{ID}, 此分支和上文一致; 另一个分支用于计算属性损失 L_{Att}. 假设图片 x 有 M 个属性标注, 我们针对于其中每一个属性计算一个损失. 若某个属性共有 m 种类型, 类似地我们可以计算图片 x 属于第 $j, j = 1, 2, 3, \cdots, m$ 的概率 $p(j|x) = \frac{exp(z_k)}{\sum_{j=1}^{m} exp(z_j)}$, 为了方便描述同样记作为 $p(j)$. 因此该属性的属性损失为:

$$L_{Att}(f, x) = -\sum_{j=1}^{m} q(j) \log p(j) \tag{2}$$

同理, $q(j)$ 是根据图片 x 的该属性标注 y_m 得到, 若 $y_m = j$, 则 $q(j) = 1$, 而对于任何的 $y_m \neq j$ 都有 $q(j) = 0$. 最终网络的总损失由 ID 损失和 M 个属性损失组成, 即:

$$L = \lambda L_{ID} + \frac{1}{M} \sum_{i=1}^{M} L_{Att}^{i} \tag{3}$$

其中 λ 是平衡两个损失的权重因子, L_{Att}^{i} 是第 i 个属性的损失值. 该网络提取的图像特征个不仅用于预测行人的 ID 信息, 还用于预测各项行人属性. 通过结合 ID 损失和属性损失能够提高网络的泛化能力.

验证网络是另外一种常用于行人重识别任务的表征学习方法[25,31]. 和分类网络不同之处在于, 验证网络每次需要输入两张图片, 这两张图片经过一个共享的 CNN 网络, 将网络输出的两个特征向量融合起来输入到一个只有两个神经元的 FC 层, 来预测这两幅图片是否属于同一个行人. 因此, 验证网络本质上是一个多输入单输出的二分类网络. 通常, 仅仅使用验证损失训练网络是非常低效的, 所以验证损失会与 ID 损失一起使用来训练网络. 图 4 是一个使用融合验证损失和 ID 损失的行人重识别网络. 网络输入为若干对行人图片, 包括分类子网络 (Classification Subnet) 和验证子网络 (Verification Subnet). 分类子网络对图片进行 ID 预测, 根据预测的 ID 来计算 ID 损失 L_{ID}, 这部分和前文一致. 验证子网络融合两张图片的特征, 判断这两张图片是否属于同一个行人, 该子网络实质上等于一个二分类网络. 假设网络输入一对图像对 $X = \{x_a, x_b\}$, 他们的 ID 标签分别为 y_a 和 y_b. 网络输出一个 2 维的向量 v, 则验证损失为:

$$L_V = -\sum_{i=1}^{2} y(i) \log v(i) \tag{4}$$

若 $y_a = y_b$ 则 $y = \{1, 0\}$, 反之若 $y_a \neq y_b$ 则 $y = \{0, 1\}$. 最终网络的总损失为 $L = L_{ID} + L_V$. 经过足够数据的训练, 在推理阶段再次输入一张测试图片, 网络将自动提取出一个特征, 这个特征用于行人重识别任务.

3.2　基于度量学习的方法

度量学习 (Metric learning) 是广泛用于图像检索领域的一种方法. 不同于表征学习, 度量学习旨在通过网络学习出两张图片的相似度. 在行人重识别问题上, 表现为同一行人的不同图片间的相似度大于不同行人的不同图片. 具体为, 定义一个映射 $f(x) : \mathbf{R}^F \rightarrow \mathbf{R}^D$, 将图片从原始域映射到特征域, 之后再定义一个距离度量函数 $D(x, y) : \mathbf{R}^D \times \mathbf{R}^D \rightarrow \mathbf{R}$, 来计算两个特征向量之

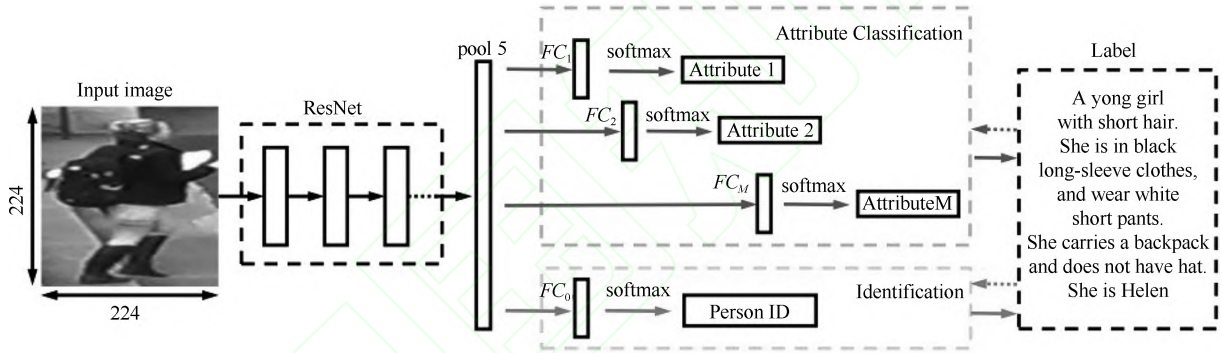


图 3　结合 ID 损失和属性损失网络示例[26]

Fig. 3　The example network with identification loss and attribute loss

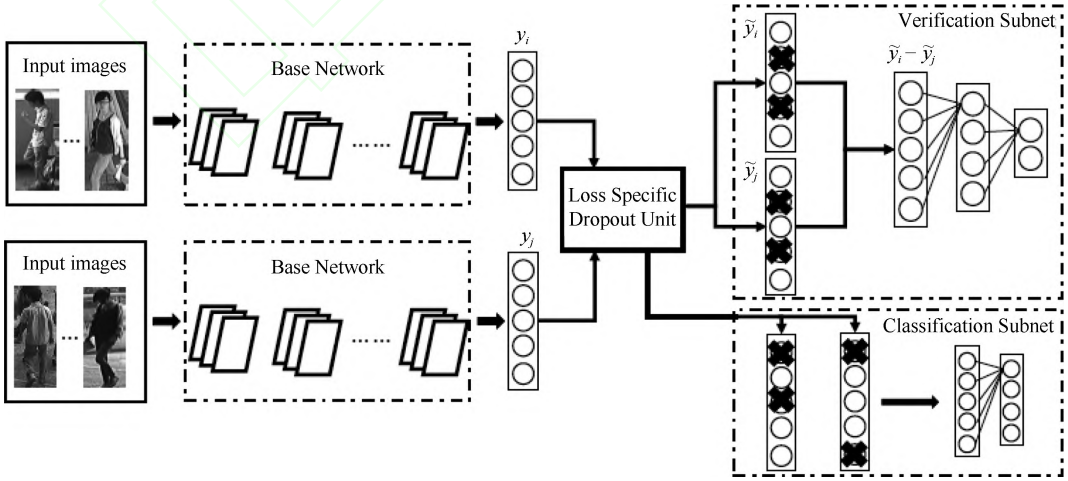


图 4　结合验证损失和 ID 损失网络示例[25]

Fig. 4　The example network with verification loss and identification loss

间的距离. 最后通过最小化网络的度量损失, 来寻找一个最优的映射 $f(x)$, 使得相同行人两张图片 (正样本对) 的距离尽可能小, 不同行人两张图片 (负样本对) 的距离尽可能大. 而这个映射 $f(x)$, 就是我们训练得到的深度卷积网络.

为了实现端对端训练, 如今深度度量学习方法和传统的度量学习方法相比已经有所变化. 关于包含 XQDA、KISSME 等在内的传统度量学习方法到深度学习度量方法的过渡与集成时期的研究, 可以查阅论文[32]. 本文着重讨论近几年基于深度学习的度量学习方法研究.

常用的度量学习损失方法包括对比损失 (Contrastive loss)[33−35]、三元组损失 (Triplet loss)[36−39]、四元组损失 (Quadruplet loss)[40]. 首先, 假如有两张输入图片 I_1 和 I_2, 通过网络的前向传播我们可以得到它们 (归一化后) 的特征向量 f_{I_1} 和 f_{I_2}. 之后我们需要定义一个距离度量函数, 这个函数并不唯一, 只要能够在特征空间描述特征向量的相似度/差异度的函数均可以作为距离度量函数. 然而, 为了实现端对端 (End-to-end) 训练的网络, 度量函数尽可能连续可导, 通常我们使用归一化特征的欧式距离或者特征的余弦距离作为度量函数, 即两张图片在特征空间的距离定义为:

$$\begin{aligned} d_{I_1,I_2} &= ||\boldsymbol{f_{I_1}} - \boldsymbol{f_{I_2}}||_2 \\ d_{I_1,I_2} &= 1 - \frac{\boldsymbol{f_{I_1}} \cdot \boldsymbol{f_{I_2}}}{||\boldsymbol{f_{I_1}}||_2||\boldsymbol{f_{I_2}}||_2}. \end{aligned} \tag{5}$$

当然曼哈顿距离、汉明距离、马氏距离等距离也可以作为度量学习的距离度量函数, 本文对此不做过多讨论.

对比损失用于训练孪生网络 (Siamese network). 孪生网络的输入为一对 (两张) 图片 I_a 和 I_b, 这两张图片可以为同一行人, 也可以为不同行人. 每一对训练图片都有一个标签 y, 其中 $y = 1$ 表示两张图片属于同一个行人 (正样本对), 反之 $y = 0$ 表示它们属于不同行人 (负样本对). 之后, 对比损失函数写作:

$$L_c = yd^2_{I_a,I_b} + (1-y)(\alpha - d_{I_a,I_b})^2_+ \tag{6}$$

其中 $(z)_+ = max(z, 0)$, α 是根据实际需求设置的训练阈值参数.

三元组损失是一种被广泛应用的度量学习损失, 之后的大量度量学习方法也是基于三元组损失演变而来. 顾名思义, 三元组损失需要三张输入图片. 和对比损失不同, 一个输入的三元组 (Triplet) 包括一对正样本对和一对负样本对. 三张图片分别命名为固定图片 (Anchor) a, 正样本图片 (Positive) p 和负样本图片 (Negative) n. 图片 a 和图片 p 为一对正样本对, 图片 a 和图片 n 为一对负样本对. 则三元组损失表示为:

$$L_t = (d_{a,p} - d_{a,n} + \alpha)_+ \tag{7}$$

文献 [36] 认为公式 (7) 只考虑正负样本对之间的相对距离, 而并没有考虑正样本对之间的绝对距离, 为此提出改进三元组损失 (Improved triplet loss):

$$L_it = d_{a,p} + (d_{a,p} - d_{a,n} + \alpha)_+ \tag{8}$$

公式 (8) 添加 $d_{a,p}$ 项, 保证网络不仅能够在特征空间把正负样本推开, 也能保证正样本对之间的距离很近.

四元组损失是三元组损失的另一个改进版本. 顾名思义, 四元组 (Quadruplet) 需要四张输入图片, 和三元组不同的是多了一张负样本图片. 即四张图片为固定图片 a, 正样本图片 p, 负样本图片 1 $n1$ 和负样本图片 2 $n2$. 其中 $n1$ 和 $n2$ 是两张不同行人 ID 的图片. 则, 四元组损失表示为:

$$L_q = (d_{a,p} - d_{a,n1} + \alpha)_+ + (d_{a,p} - d_{n1,n2} + \beta)_+ \tag{9}$$

其中 α 和 β 是手动设置的正常数, 通常设置 β 小于 α, 前一项称为强推动, 后一项称为弱推动. 其中前一项和三元组损失一样, 只考虑正负样本间的相对距离, 共享了固定图片 a. 因此在推开负样本对 a 和 $n1$ 的同时, 也会直接影响 a 的特征, 造成正样本对 a 和 p 的距离不好控制. 改进三元组损失通过直接约束 a 和 p 之间的距离来解决这个问题. 而四元组通过引入第二项弱推动实现, 添加的第二项中负样本对和正样本对不共享 ID, 所以考虑的是正负样本间的绝对距离, 在推开负样本对的同时不会太过直接影响 a 的特征. 因此, 四元组损失通常能让模型学习到更好的表征.

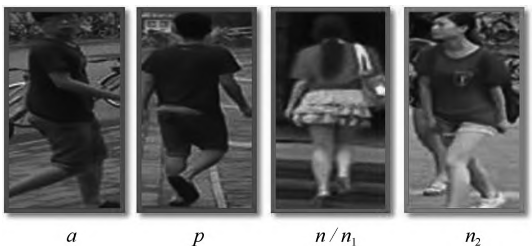


图 5 度量学习方法样本符号示例图

Fig. 5 The sample's label of metric learning

以上度量学习方法样本示例如图 5 所示, 这些方法在计算度量损失时, 样本对都是从训练集中随机挑选. 随机挑选样本对的方法可能经常挑选出一些容易识别的样本对组成训练批量 (Batch), 使得网络泛化能力受限. 为此, 部分学者提出了难样本采样 (Hard sample mining) 的方法, 来挑选出难样

本对训练网络[37,41]. 常用的思路是挑选出一个训练 Batch 中特征向量距离比较大 (非常不像) 的正样本对和特征向量距离比较小 (非常像) 的负样本对来训练网络. 难样本采样技术可以明显改进度量学习方法的性能, 加快网络的收敛, 并且可以很方便地在原有度量学习方法上进行扩展, 是目前广泛采用的一种技术.

度量学习可以近似看作为样本在特征空间进行聚类, 表征学习可以近似看作为学习样本在特征空间的分界面. 正样本距离拉近的过程使得类内距离缩小, 负样本距离推开的过程使得类间距离增大, 最终收敛时样本在特征空间呈现聚类效应. 度量学习和表征学习相比, 优势在于网络末尾不需要接一个分类的全连接层, 因此对于训练集的行人 ID 数量并不敏感, 可以应用于训练超大规模数据集的网络. 总体而言, 度量学习比表征学习使用的更加广泛, 性能表现也略微优于表征学习. 但是目前行人重识别的数据集规模还依然有限, 表征学习的方法也依然得到使用, 而同时融合度量学习和表征学习训练网络的思路也在逐渐变得流行.

3.3　基于局部特征的方法

从网络的训练损失函数上进行分类可以分成表征学习和度量学习, 相关方法前文已经介绍. 另一个角度, 从抽取图像特征进行分类, 行人重识别的方法可以分为基于全局特征 (Global feature) 和基于局部特征 (Local feature) 的方法. 全局特征比较简单, 是指让网络对整幅图像提取一个特征, 这个特征不考虑一些局部信息. 正常的卷积网络提取的都是全局特征, 因此在此不做赘述. 然而, 随着行人数据集越来越复杂, 仅仅使用全局特征并不能达到性能要求, 因此提取更加复杂的局部特征成为一个研究热点. 局部特征是指让手动或者自动地让网络去关注关键的局部区域, 然后提取这些区域的局部特征. 常用的提取局部特征的思路主要有图像切块、利用骨架关键点定位以及行人前景分割等等.

图片切块是一种很常见的提取局部特征方式[34,42−43]. 因为人体结构的特殊性, 通常研究者会将图片从上到下均分为几等份 (头部、上身、腿部等). 图 6 是图片切块的一个典型示例, 网络采用的是经典的孪生网络, 损失函数为度量学习的对比损失, 输入的两幅图片均分为若干等分. 之后, 被分割好的若干块图像块按照顺序送到一个长短时记忆网络 (Long short term memory network, LSTM), 最后的特征融合了所有图像块的局部特征. 图片切换方法的缺点在于对图像对齐的要求比较高, 如果两幅图像没有上下对齐, 那么很可能出现头和上身对比的现象, 反而使得模型判断错误. 因此 Zhang 等人设计了一种动态对齐网络 AlignedReID[43], 可以在不需要额外信息的情况下实现图片块从上到下的自动对准.

利用人体姿态关键点进行局部特征对齐是另外一种常见的方法[44−46]. 一些论文利用一些先验知识先将行人进行对齐, 这些先验知识主要是预训练的人体姿态 (Pose) 和骨架关键点模型. 其中, CVPR17 的 Spindle Net[45] 是该类方法的一个典型代表. Spindle Net 网络如图 7 所示, 首先通过骨架关键点提取的网络提取 14 个人体关键点, 之后利用这些关键点提取 7 个人体结构 ROI. 这 7 个 ROI 区域和原始图片进入同一个 CNN 网络提取特征. 原始图片经过完整的 CNN 得到一个全局特征, 三个大区域经过 FEN-C2 和 FEN-C3 子网络得到三个局部特征, 四个四肢区域经过 FEN-C3 子网络得到四个局部特征. 之后这 8 个特征按照图示的方式在不同的尺度进行联结, 最终得到一个融合全局特征和多个尺度局部特征的行人重识别特征.

与 Spindle Net 不同, 论文 [46] 先用姿态估计的模型估计出行人的关键点, 然后用仿射变换使得相同的关键点对齐. 论文 [44] 提出了一种全局 - 局部对齐特征描述子 (Global-Local-Alignment Descriptor, GLAD). GLAD 利用提取的人体关键点把图片分为头部、上身和下身三个部分. 之后将整图

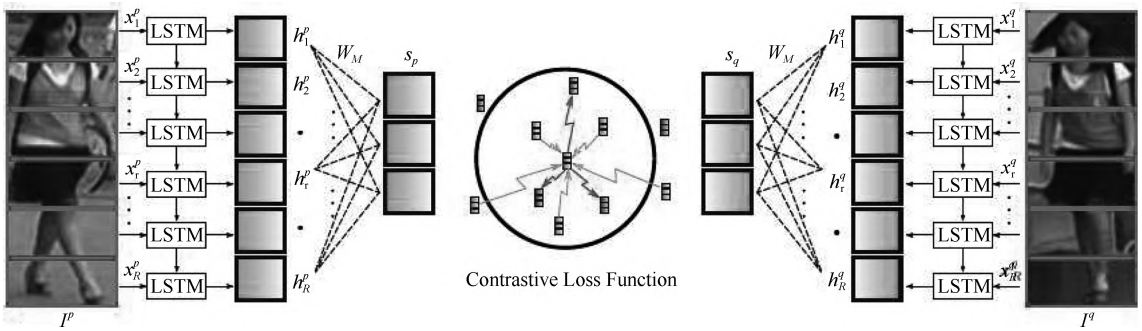


图 6　利用图片切块提取局部特征示例[34]

Fig. 6　The example of extracting local features with image blocks

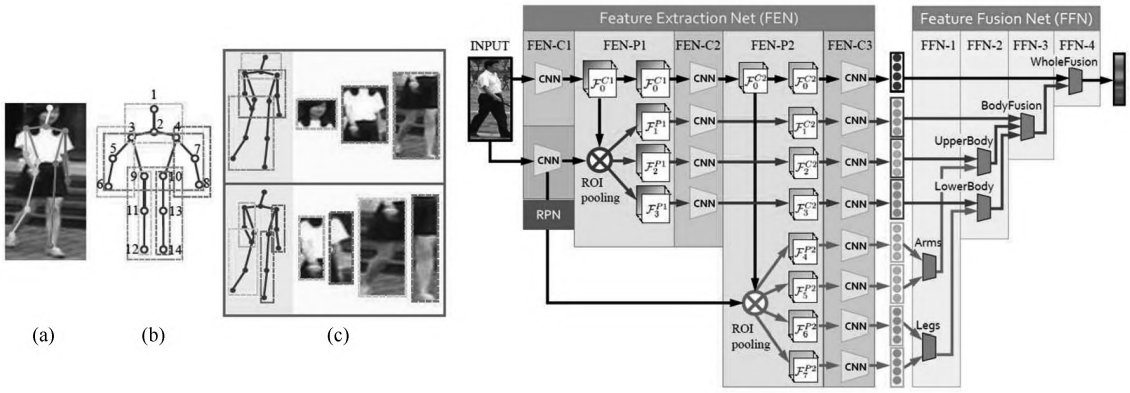


图 7 利用姿态点提取局部特征示例[45]

Fig. 7 The example of extracting local features with pose points

和三个局部图片一起输入到一个参数共享 CNN 网络中, 最后提取的特征融合了全局和局部的特征. 为了适应不同分辨率大小的图片输入, 网络利用全局平均池化 (Global average pooling, GAP) 来提取各自的特征. 和 Spindle Net 略微不同的是四个输入图片各自计算对应的损失, 而不是融合为一个特征计算一个总的损失.

行人的局部特征在最近逐渐被证明是一种有效的特征, 可以一定程度上解决行人姿态多样化的问题. 因此, 融合全局和局部特征在行人重识别领域也渐渐变得流行, 图片切块的方法简单但是需要图片比较规范化, 利用姿态点信息比较精确但是需要额外的姿态估计模型. 高效且低耗的局部特征提取模型依然是该领域一个值得研究的切入点.

3.4 基于视频序列的方法

目前主流的行人重识别方法大部分是基于单帧图像的, 然而单帧图像给予的信息终究是有限的. 此外, 单帧的方法要求图像质量很高, 这对于相机的布置和使用的场景是一个非常大的限制, 因此研究基于序列的方法便显得十分重要. 基于单帧图像的 ReID 方法可以通过一个简单方法扩展到视频序列, 即用所有序列图像特征向量的平均池化或者最大池化作为该序列的最终特征. 但是仍然有很多工作在研究如何更好地利用视频序列来进行行人重识别[47−58]. 这类方法除了考虑了图像的内容信息, 还会考虑: (1) 帧与帧之间的运动信息; (2) 更好的特征融合; (3) 对图像帧进行质量判断等. 总体来说, 基于序列的方法核心思想为通过融合更多的信息来解决图像噪声较大、背景复杂等一系列质量不佳的问题. 本小节将会着重介绍几个典型方法, 以点带面的形式来总结该类方法.

融合图像内容信息和运动信息是一种常见的思路, 因为运动信息里面可能包含了步态等信息辅助识别任务, 最早的序列类方法的关注点就在于运动信息上[49,51,56]. 主要思想是利用 CNN 来提取空间特征的同时利用递归循环网络 (Recurrent neural networks, RNN) 来提取时序 (运动) 特征. 典型代表是累计运动背景网络 (Accumulative motion context network, AMOC)[49]. AMOC 输入的包括原始的图像序列和提取的光流序列 (运动特征). 其核心思想在于网络除了要提取序列图像的特征, 还要提取运动光流的运动特征, 其网络结构图如图 8 所示. AMOC 拥有空间信息网络 (Spatial network, Spat Nets) 和运动信息网络 (Motion network, Moti Nets) 两个子网络. 图像序列的每一帧图像都被输入到 Spat Nets 来提取图像的全局内容特征. 而相邻的两帧将会送到 Moti Nets 来提取光流图特征. 之后空间特征和光流特征融合后输入到一个 RNN 来提取时序特征. 通过 AMOC 网络, 每个图像序列都能被提取出一个融合了内容信息、运动信息的特征. 网络采用了分类损失和对比损失来训练模型. 融合了运动信息的序列图像特征能够提高行人重识别的准确度.

序列图像每一帧都可以提取一个特征, 通常每一帧贡献的信息是不同的, 因此如何更好地融合每一帧的特征也是一个研究热点[47−48]. 该类方法的一个代表工作是 DFGP(Deep Feature Guided Pooling)[48]. DFGP 先用一个深度学习模型对每一帧提取一个深度特征, 之后用平均池化得到序列图像的平均特征, 这与大部分工作一致. 之后 DFGP 提出一个最稳定帧算法 (Maximally Stable Video Frame, MSVF). MSVF 通过计算每一帧图像特征与平均特征之间的距离, 挑出距离最小的那一帧为该序列的最稳定帧. 如果某一帧与最稳定帧越接近, 则给予该帧越大的权重. 实现方式为计算每一帧的特征与最稳定帧特征的距离, 距离越近权重最大, 只要满足所有帧的权重和为 1 即可. DFGP 是一种手

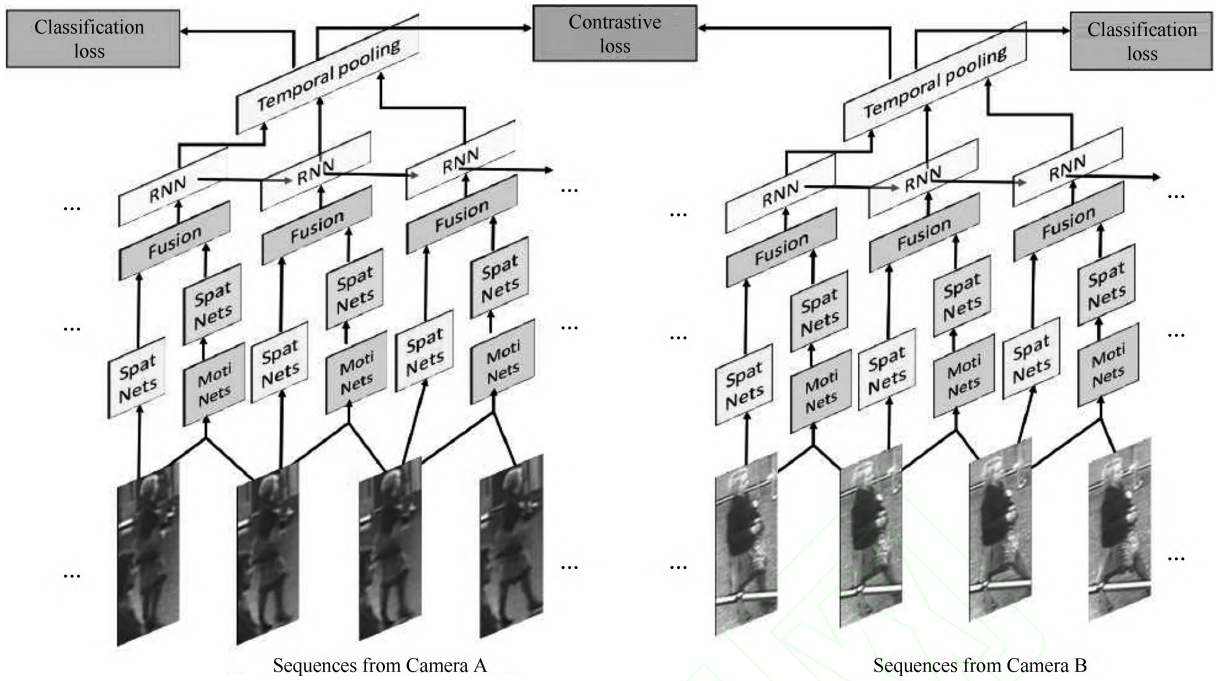


图 8 融合内容信息和运动信息的 AMOC 网络[49]

Fig. 8 The AMOC network which fusions context information and motion information

动实现特征融合的方法, 当前另一种流行的思路是利用深度学习的注意力机制, 来自动地给每一帧图像赋予一个权重[47]. 当然这个权重是由网络自动学习出来, 解释性不如手动提取方法.

序列方法另外一个思路就是对图像帧进行质量判断, 往往序列中并不是每一帧图像都是完整的高质量图像, 遮挡、姿态、分辨率等因素都是可能造成图像不佳的因素. 因此保留更多的高质量图像的特征便显得比较重要, RQEN(Region-based Quality Estimation Network) 就是一个对遮挡图像进行质量判断的工作[18]. RQEN 认为在遮挡较严重的情况下, 如果用一般的平均池化会造成遮挡区域的特征丢失很多. 而 RQEN 以姿态估计点为先验知识, 对每帧进行一个质量判断, 姿态完整的图像被认为是高质量的图像, 反之姿态不完整的图像即存在遮挡的图像是低质量的图像. 将这个先验结果输入到网络, 诱导网络学习更多高质量图像帧的信息, 给高质量图像帧打上高权重, 然后对特征图进行一个线性叠加. 图 9 显示了平均池化和 RQEN 方法的注意力图, 可以看出在存在遮挡的情况下, 平均池化在遮挡区域会丢失很多信息, 而融合质量判断的 RQEN 网络依然可以得到较好的结果.

基于视频序列的行人重识别技术是该领域未来急需解决的一个问题. 总体而言, 和单帧方法相比, 序列方法无论是从思路的多样性上, 还是从结果性能上, 都还存在一定的差距.

图 9 RQEN 与平均池化注意力图对比[18]

Fig. 9 The attention maps of RQEN and average pooling

3.5 基于 GAN 的方法

GAN 在近几年得到了蓬勃的发展, 其中一个应用就是图片生成. 深度学习的方法需要依赖大量训练数据, 而目前行人重识别的数据集总体来说规模还是比较小. 因此, 利用 GAN 来做行人重识别任务逐渐开始变得流行. 传统的 GAN 生成图片是随机的, 后来发表的 CycleGAN[59], DualGAN[60] 和 DiscoGAN[61] 实现了图片风格的转换, 更加进一步地促进了 GAN 在行人重识别领域的应用.

第一篇引入 GAN 做 ReID 的论文 [19] 发表在 ICCV17 会议上, 论文使用传统的 GAN 随机生成行人图片, 因此生成的图片是不可控的, 仅仅作为 IDE 网络训练数据的增广, 提高 IDE 网络的性能. 为了解决这个问题, 一些文献 [19, 62 – 65] 使用 (改进的)CycleGAN 来进行两个域的行人图片转换, 从而减小图片间的风格差异. 虽然每个算法使用 CycleGAN 的细节各自不同, 但是流程可以统一概括.

图 10 显示了利用 CycleGAN 将图片 1 从风格 A 转换为风格 B 的网络训练流程图, 网络输入两张不同风格的图片. 生成器 G_{AB} 将图片从风格 A 转化为风格 B, 而 G_{BA} 将图片从风格 B 转化为风格 A, 判别器 D_B 用来判断生成图片是否接近真实的风格 B. 图 10 只展示了 CycleGAN 的从 A → B 风格转换, 实际的 CycleGAN 是对称结构, 而 B → A 方向的转换在此不做赘述. 通过最小化判别损失和重建 L2 损失, CycleGAN 的生成器和判别器不断对抗互相提高直至收敛. 推理阶段只需要给对应的生成器输入一张图片, 便可以将图片从一个风格转换为另外一个风格. 与传统的 GAN 网络不同, CycleGAN 一个非常好的优点是生成的图片保留了原始图片的 ID 信息.

由于相机的光线、角度等可能不同, 不同相机拍摄的图片存在风格偏差, Zheng 等[65] 使用 CycleGAN 来实现相机风格的迁移, 从而减小相机间的风格偏差. 另外, 由于不同数据集之间存在场景域之间的偏差, 通常在一个数据集训练的网络在另外一个数据集上性能不会很好. Person Transfer GAN (PTGAN)[19] 和 Similarity Preserving GAN (SPGAN)[62] 分别改进 CycleGAN 来实现图片数据域之间的转换. PTGAN 主要思想为保持生成图片的行人前景尽可能不变, 而图片背景为目标域的风格, 因此采用一个前景分割网络先前景区域分割出来, 在训练网络时加入前景约束, 尽可能多的保护行人外观信息. SPGAN 和 PTGAN 类似, 也是利用 CycleGAN 实现数据集风格的转换, 与之不同的是 SPGAN 将 CycleGAN 与孪生网络结合, 生成图片的同时加入了 ReID 模型的约束, 使之更加适应任务需求. Pose-normalization GAN (PNGAN)[64] 是另外一篇非常典型的工作, 行人重识别任务其中一个难点就是行人的姿态存在偏差, 而 PGNAN 的主要动机就是解决姿态偏差. 为了实现行人姿态的迁移, PNGAN 将 InfoGAN[66] 和人体姿态点估计模型结合起来, 加入了姿态点损失约束, 使得生成图片的行人姿态与期望姿态尽可能一致. 之后为了消除姿态偏差, EPSAN 定义了八个姿态模板, 每一张行人图片都转化为 8 张固定姿态的图片, 在进行重识别任务时融合这 9 张图片的特征, 达到消除姿态偏差的目的. 据我们所知, PNGAN 在 Market1501 和 DukeMTMC-reID 数据集上达到了目前最高的 Rank1 准确度. 这些典型 GAN 网络生成的图片如图 11 所示, 根据目标的不同, GAN 网络图片生成的外观细节和侧重点也各自不同, 这就是这一类方法的特点所在.

表 2 基于 GAN 网络的方法比较

Table 2 The comparison of GAN based methods

算法	GAN	CycleGAN	PTGAN	SPGAN	PNGAN
基础	GAN	CycleGAN	CycleGAN	CycleGAN	InfoGAN
额外	标签平滑	标签平滑	前景分割	孪生网络	姿态估计
目标	数据增广	相机偏差	数据域偏差	数据域偏差	姿态偏差

3.6 各方法总结比较

前文按照分类介绍一些基于深度学习的行人重识别方法, 本小节将对这些方法进行总结与比较. 基于 GAN 的方法更多是作为一种图像增广或者解决图像域偏差的技术而较为独立.

从训练深度网络的角度, 我们将从三个角度来分析: 表征学习与度量学习、全局特征与局部特征、单帧图像与视频序列. 如表格 3 所示, 前文提到的代表算法所对应的类型都已标记出. 有的方法只使用

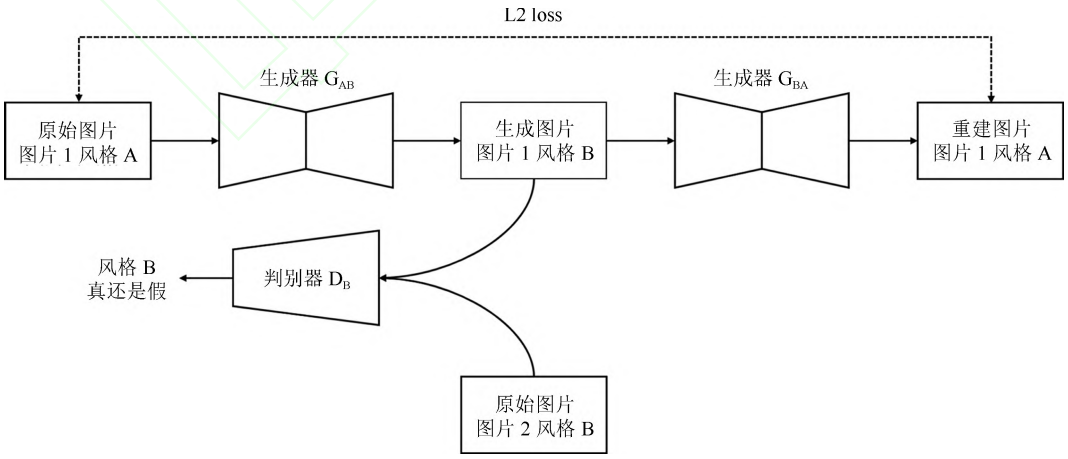


图 10 CycleGAN 进行图片风格转换流程图 (A → B)

Fig. 10 The pipeline of image style transfer using CycleGAN (A → B)

表 3 基于深度学习的行人重识别方法总结比较

Table 3 Comparison of deep learning based ReID methods

方法	表征学习	度量学习	全局特征	局部特征	单帧图像	视频序列
IDE Net[29]	√		√		√	
TriHard[37]		√	√		√	
QuadLoss[40]		√	√		√	
LSTM Siamese[34]		√		√	√	
Gate Reid[33]		√		√	√	
Spindle Net[45]		√	√	√	√	
GLAD [44]	√		√	√	√	
AlignedReID[43]	√	√	√	√	√	
AMOC[49]	√	√	√			√
RQEN[43]	√	√	√	√		√

图 11 GAN 网络生成行人图片示例

Fig. 11 The examples of pedestrian images generated by GAN

了一种类型的损失函数或者特征类型, 而有的方法融合了多种损失函数或者特征类型来达到更高的性能水平. 具体细节我们将在后文详细讨论.

3.6.1 基于表征学习与度量学习的方法

按照网络训练损失分类, 行人重识别的方法可以分为表征学习和度量学习两类. 表征学习的优点在于数据集量不大的时候收敛容易, 模型训练鲁棒性强, 训练时间短. 然而表征学习是将每一个 ID 的行人图片当做一个类别, 当 ID 数量增加到百万、千万甚至更多的时候, 网络最后一层是一个维度非常高的全连接层, 使得网络参数量巨大并且收敛困难. 由于直接计算特征之间的距离, 度量学习的优点在于可以很方便的扩展到新的数据集, 不需要根据 ID 数量来调整网络的结构, 此外也可以非常好的适应 ID 数目巨大的训练数据. 然而, 度量学习相对来说收敛困难, 需要比较丰富的训练经验来调整网络参数, 另外收敛训练时间也比表征学习要长.

表征学习和度量学习拥有各自的优缺点, 目前学术界和工业界逐渐开始联合两种学习损失[18,43,49]. 联合的方式也比较直接, 在传统度量学习方法的基础上, 在特征层后面再添加一个全连接层进行 ID 分类学习. 网络同时优化表征学习损失和度量学习损失, 来共同优化特征层. 如图 8 所示, AMOC 同时联合了 ID 损失和对比损失, 特征层之后分出了两个分支分别优化表征学习损失和度量学习损失.

3.6.2 基于全局特征与局部特征的方法

按照网络输出特征类型, 行人重识别方法可以分为基于全局特征与局部特征的方法. 全局特征一般是卷积网络的特征图直接通过一个全局池化层得到, 推理阶段计算快速, 适合于需要帧率较高的实际应用. 然而由于全局池化层会使得图像的空间特征信息丢失, 因此在姿态不对齐、行人图片不完整、只有局部细节不相似等情况下, 全局特征容易出现误识别. 而局部特征的优点在于可以一定程度上解决这些问题, 当然局部特征也有它自己的缺点. 对于分块的局部特征优点在于不需要引入额外的计算量, 但是通常并不能特别好的解决姿态不对齐的问题. 而利用姿态点估计模型估计出行人的姿态点, 然后再进行局部特征匹配可以较好的解决姿态不对齐的问题, 但是却需要一个额外的姿态点模型. 总体来说, 全局特征和局部特征是两个比较互补的特征类型, 通常不会单独使用局部特征. 广义上讲, 分块局部特征把所有的分块特征融合起来也包含了全局图像信息. 因此在不考虑推理阶段计算耗时的前提下, 融合全局特征和局部特征是目前一种提高网络性能非常常用的手段.

目前融合全局特征和局部特征常用的思路是对于全局模块和局部模块分别提取特征, 之后再将全局特征和局部特征拼接在一起作为最终的特征. 如图 7 所示, Spindle Net 就提取了全局特征和 7 个不同尺度的局部特征, 然后融合成最终的图像特征用于进行最后的相似度度量. AlignedReID 给出了另外一种融合方法, 即分别计算两幅图像全局特征距离和局部特征距离, 然后加权求和作为最终两幅图像在特征空间的距离. RQEN 则是利用一个姿态点模型来估计行人的可视性部分, 然后融合多帧信息得到一个较好的最终特征, 也可以看做是一个全局特征和局部特征融合的过程. 如何探究更好的全局特征和局部特征方法也是行人重识别未来一个重要的研究分支.

3.6.3 基于单帧图像与视频序列的方法

按照网络输入数据, 行人重识别方法可以分为基于单帧图像与视频序列的方法. 这两类方法并没有太多重合的地方, 只是针对于不同的应用选择不同类型的网络输入. 基于单帧图像的方法训练简单, 使用方便, 推理阶段耗时时间短. 然而它的缺点在于单帧图像信息有限, 对于图像质量要求较高, 一旦出现检测框错误或者行人遮挡等情况, 算法效果会大幅度下降. 基于视频序列的方法可以解决单帧图像信息不足的缺点, 并且可以融入运动信息加强鲁棒性, 然而由于每次要处理多张图像, 因此计算效率较低. 当然基于视频序列的方法大部分都是单帧图像方法的扩展延伸, 因此发展单帧图像的方法对于发展视频序列的方法也是有益的.

4 典型算法比较

本小节我们通过比较一些典型算法的性能, 来回顾一下近几年行人重识别方法的发展趋势. 由于行人重识别相关数据集非常多, 我们无法展示每个数据集的结果. 考虑到算法的优劣一般与数据集相关性不是特别大, 而多数论文都会在 Market1501 上做评测, 因此我们选择 Market1501 数据集作为示例数据集.

本文挑选了近几年比较有代表性的十余种基于深度学习的行人重识别方法, 首先挑选顶级会议上发表且在当时准确度较高的方法. 然后根据第 3 章节挑选出准确度较高且方便归类的预印版文献方法. 为了方便比较, 我们还挑选了一个可以代表最高准确度的传统方法和无监督学习方法作为参考基准.

表 4 总结比较了这十余种深度学习方法的性能、基本网络, 并简单描述了算法特性和发表状况. 行人重识别最主要的两个性能指标是一选准确率 (rank-1) 和平均准确率 (mAP). 表格给出的结果均由论文中给出, 一些代表性的算法没有在 Market1501 上进行评测因此没有展示. 除非算法本身是基于重排序的研究, 否则本文默认都是使用欧式距离的无重排序结果. 为了方便比较, 表格第一行给出了比较好的传统方法的结果, 作为传统方法的基准. 第二部分是强监督的深度学习方法, 可以看出基准网络里面 ResNet50[69]、GoogleNet[70] 和自搭的 CNN 网络使用较多, 损失函数方面分类损失和度量损失均可以取得很好的性能. 在已经发表的方法中, DML[28]、CamStyle[65]、GLAD[44] 均取得了接近 90.0 % 的一选准确率. 而在还未接受的预印版文献里, AlignedReID[43]、PNGAN[64] 已经超越了 90.0 % 的一选准确率, 代表了目前行人重识别领域 Market1501 数据集的最高准确度. 在第三部分的无监督学习方法方面, 大部分无监督学习方法都还是基于传统特征的研究. 而 CVPR2018 刚接收的 SPGAN[62] 是比较具有代表性的基于深度学习的无监督行人重识别方法, 同时在 Market1501 数据集上也击败了目前已有的无监督学习方法.

总的看来, 基于深度学习的行人重识别方法近几年来发展迅速, 每年以大概 15 % 的一选准确率速度在增长, 并且各种方法百花齐放, 并没有哪种方法相比于其他方法存在巨大的优势. 而趋势方面, 方法从早期的单网络单损失逐渐发展为现在的多损失多网络以及多尺度多特征的融合, 即一个由简到繁的发展过程. 表 4 给出的均是基于单帧图像方法的结果, 而基于视频序列的方法目前还没有特别多代表性的方法和结果, 因此本文不再做整理.

5 挑战与未来

5.1 目前的重要挑战

行人重识别虽然近几年取得了高速的发展, 然而目前依然面临着许多挑战. 目前学术界已存的数据集是清理之后的高质量图像, 然而在真实场景下行人重识别会遇到跨视角造成的姿态多变、分辨率变化、行人遮挡以及图像域变化等问题. 这些问题逐渐受到学者的重视, 本小节将会简单介绍一些克服这些挑战的代表性.

1) 跨视角造成的姿态多变问题: 由于不同摄像头架设的角度、位置不一, 拍摄图片中的行人姿态也十分多变. 目前已经有不少代表性的工作从不同角度上来解决这个问题, 而这些方法主要是依靠一个预训练的姿态模型来实现姿态的对齐. 除了 3.3 小节中介绍的 GLAD 和 SpindleNet 等工作以外, CVPR2018 提出的姿态敏感嵌入方法 (Pose-Sensitive Embedding, PSE)[71]. 如图 12 所示, PSE 利用一个预训练的姿态模型估计行人的姿态点, 然

后将姿态点信息输入到网络, 网络的视角分支会估计行人的朝向及其概率. 另一方面, PSE 的特征分支分别得到前向、背向和侧向三个视角的特征图, 之后与估计的视角概率加权得到最终的全局特征. 通过使用对齐后的全局特征, 可以更好地处理视角多变的行人图片.

2) 行人图片分辨率变化: 由于摄像头中目标拍摄距离不一致, 拍摄的行人图片分辨率也不一样. 目前专门解决这个问题的方法较少, 论文[72] 提出了一个新的图像超分辨和行人身份识别联合学习 (Super-resolution and identity joint learning, SING) 的方法. 如图 13 所示, SING 通过联合学习

表 4　典型行人重识别方法在 Market1501 上性能比较

Table 4　Comparison of the performance of typical ReID methods on Market1501

方法	rank-1	mAP	损失函数	基础网络	简单描述	发表
LOMO+XQDA[7]	43.8	22.2			传统方法基准	CVPR2015
LSTM Siamese[34]	61.6	35.3	对比损失	LSTM	图像分块 + 孪生网络	ECCV2016
Gate Reid[33]	65.9	39.6	对比损失	CNN	孪生网络 + 多尺度全局特征个	ECCV2016
Spindle Net[45]	76.9	-	分类损失	CNN	姿态对齐 +IDE	CVPR2017
GAN[39]	78.1	56.2	分类损失	Resnet50	GAN+IDE+ 数据增广	ICCV2017
Part-Aligned[67]	81.0	63.4	三元组损失	GoogleNet	姿态对齐 + 度量学习	ICCV2017
Deep Transfer[25]	83.7	65.5	分类损失	GoogleNet	ID 损失 + 验证损失 + 迁移学习	Arxiv2016
TriHard[37]	84.9	69.1	三元组损失	Resnet50	难样本挖掘 + 三元组损失	Arxiv2017
DML[28]	87.7	68.8	分类损失	MobileNets[?]	IDE+ 互学习	CVPR2018
CamStyle[65]	88.1	68.7	分类损失	Resnet50	CycleGAN+IDE+ 相机偏差	CVPR2018
GLAD [44]	89.9	73.9	分类损失	GoogleNet	姿态对齐 + 特征融合 + 重检索	ACMMM2017
AlignedReID[43]	91.8	79.8	三元组损失	Resnet50	难样本挖掘 + 图片切块 + 自动对齐 + 互学习	Arxiv2017
PNGAN[64]	95.5	89.9	分类损失	Resnet50	InfoGAN+ 姿态估计 +IDE+ 属性损失	Arxiv2017
SPGAN[62]	51.5	22.8	分类损失	Resnet50	CycleGAN+IDE+ 数据域偏差 + 无监督	CVPR2018

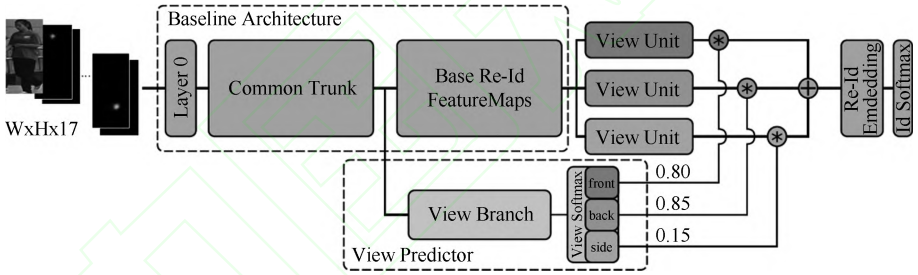


图 12　PSE 网络流程图[71]

Fig. 12　The pipeline of PSE network

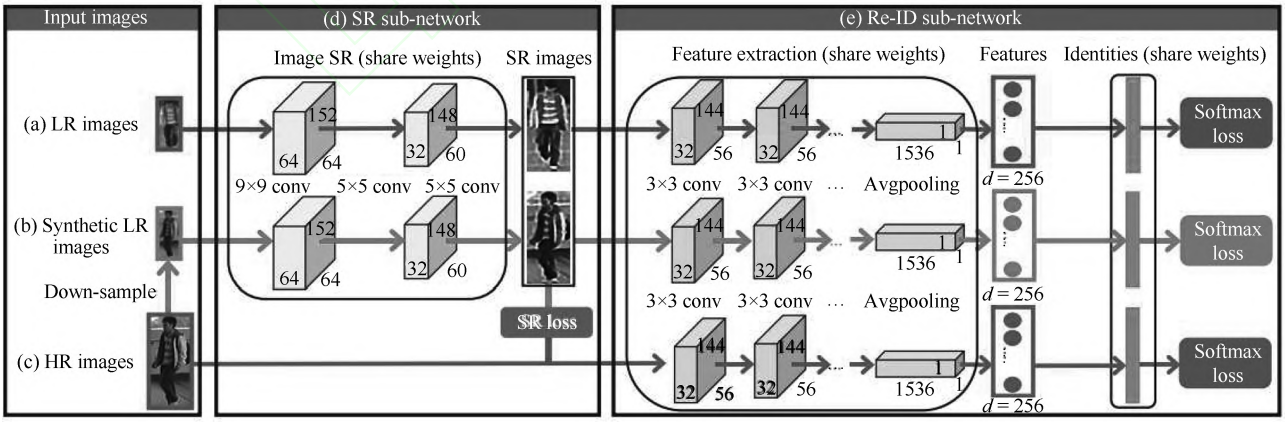


图 13　图像超分辨和行人身份识别联合学习模型示意图[72]

Fig. 13　Illustration of model structure of image super resolution and person identity joint learning

图片的超分辨率和行人重识别的问题, 既能够提升低分辨率图片的分辨率, 又能提高低分辨率图片行人重识别任务的准确度. 为了得到低分辨率的图片, SING 先用高分辨率图片降采样得到一批低分辨率图片. 之后, 网络优化联合学习图像超分辨的重构损失和行人身份识别损失函数. 低分辨率图片经过网络高分辨率处理后再进行特征提取, 而正常分辨率图像则是直接进行特征提取. 由于不同分辨率的图片经过不同的方式提取特征, 因此 SING 网络能够较好的应对分辨率变化的问题.

3) 行人图片遮挡问题: 目前学术界的行人重识别数据集大多数清洗过的高质量图像. 然而在真实的使用场景, 行人经常会被移动目标或者静态物体所遮挡, 造成行人图片的不完整. 由于失去了部分行人特征而引入了很多干扰特征, 使得很多基于全局特征的行人重识别算法效果大大下降. 为了解决这个问题, 一个思路是利用行人姿态模型来估计行人图像的可视部分, 然后对可视部分进行局部特征提取、融合[18]. 而 CVPR2018 的论文[73] 提出深度空间特征重建方法 (Deep Spatial feature Reconstruction, DSR) 来进行不完整图片和完整图片的匹配. 如图 14 所示, DSR 利用一个训练好的 ReID 网络对图片进行特征提取, 并且不对原图进行尺度变换的操作. 不同尺寸的图片经过网络后得到不同尺寸大小的特征图, 而两个不同尺寸的特征图并不能直接地进行相似度计算. 为了解决这个问题, DSR 利用空间特征重建 (Spatial Feature Reconstruction) 的方法计算出两幅特征图之间的稀疏表达系数. 完整图片的特征图经过乘以稀疏表达系数便可以与不完整图片的特征图进行欧氏距离的度量. 从而实现不同尺寸图片的特征图相似度的计算.

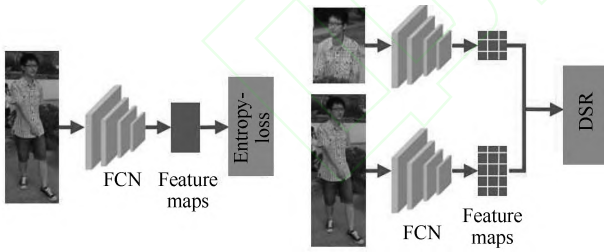


图 14 深度空间特征重建方法示意图[73]

Fig. 14 Illustration of Deep Spatial feature Reconstruction Method

4) 图像域变化的跨模态重识别. 图像域的变化是行人重识别应用上非常普遍的一个挑战. 图像域变化的类型也多种多样, 例如不同相机、不同天气、不同时间、不同城市拍摄的图像风格均可能不同. 此外, 夜晚 RGB 相机也会失效, 使用红外相机拍摄的图片没有颜色信息, 因此 RGB 图片与红外图片的行人重识别也是个典型的跨模态问题. 目前基于 GAN 网络生成图像来解决图像域偏差是一个很流行的思路, 例如前文介绍的 CamStyle 解决不同相机的图像域问题, PTGAN 解决不同城市的图像域问题. 而 RGB 与红外图片域的跨模态重识别问题逐渐开始受到关注, ICCV17 接受的一篇论文[17] 提出了深度零填充模型 (Deep zero padding model) 首次利用深度网络来解决这一问题. 如图 15 所示, 该方法的核心思想是在网络输入图片的时候, 对于不同域的图片在不同的通道上用零填充. 零填充通道记录了图像来源于哪个图像域的信息, 促使深度网络根据图像域来自适应提取不同的特征, 从而实现更好的跨模态行人重识别.

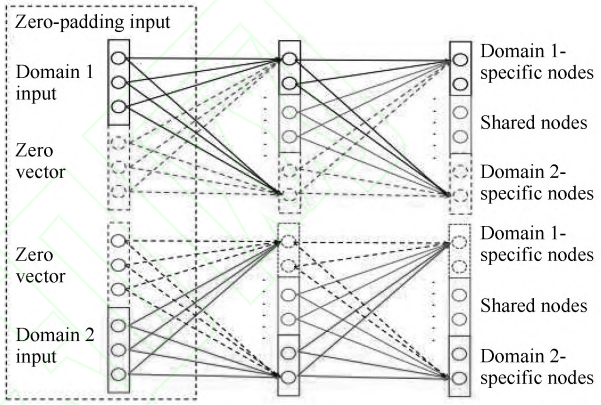


图 15 深度零填充模型详解[17]

Fig. 15 Explanation of deep zero padding model

5.2 未来的研究方向

随着深度学习的快速发展, 近几年行人重识别的发展也非常迅速. 在最流行的数据集 Market1501、CUHK03、DukeMTMC-ReID 上, 最高的一选 (rank-1) 准确率都达到了 90 % ∼ 95 %. 虽然这个准确度离人脸识别还有一定距离, 但是已经有了超越人类水平的趋势[43]. 不过行人重识别技术要从学术研究走向成熟应用, 依然还有一些需要解决的问题. 关于未来的研究方向, 我们认为可以从以下几个方面进行考虑, 并且给出了一些可能的解决思路:

1) 构建更加适应真实环境的高质量标准数据库: 目前最大的行人重识别数据集依然在几千个行人 ID 的程度, 这与人脸的百万级数据库还有着一定差距. 并且目前的数据集场景丰富度也不够高, 通常就是在一个环境下一个较短时间间隔采集的数据. 目前行人重识别数据集之间的偏差依然很大, 而不同地域、空间、时间环境下的行人着装也各有不同, 一个数据集训练的网络在另外一个数据集下性能会下降相当多. 只有足够大的高质量标准数据集的出现, 才能更好地证明算法的鲁棒性. 目前 MSMT17

数据集[19] 和 LVreID 数据集[20] 将行人重识别的数据集进一步扩大. 除此之外, 一些基于 GAN 的方法[19,62−65] 也能够生成一些接近真实场景下的图片, 来解决目前数据量较少的问题.

2) 半监督、无监督和迁移学习的方法: 采集的数据终究是有限的, 而标注数据的成本代价也很高. 因此半监督和无监督学习的方法虽然在性能上可能比不上监督学习方法, 但是性价比很高. 迁移学习也是一个值得研究的方向. 行人重识别技术的应用场景是无限的, 针对于每一个使用场景都训练一个专用模型是非常低效的. 如何通过迁移学习的方法将一个场景训练的模型适应新的场景是一个有价值的研究问题. 而半监督、无监督以及迁移学习的深度学习行人重识别技术已经有一些研究工作[25,50,74], 不过还有很大的提升空间.

3) 构造更加强大的特征: 提高行人重识别的性能主要是从特征提取和图像检索两个角度切入. 一些重排序技术可以用消耗时间为代价提高检索准确度[75], 而一个好的特征可以更加经济提高性能. 具体而言, 行人重识别任务要想构造一个更好的特征, 需要网络能够关注到更加关键的局部信息, 即更加合理的局部特征. 而利用更加丰富的序列特征也是构造特征的一个思路.

4) 丰富场景下的行人重识别: 目前行人重识别数据集以视野广阔的室外场景为主, 几个包含室内场景的数据集也能够保证行人是完整的全身. 但是在一些场景下, 例如无人超市、商场、地铁内等, 会存在非常多的半身图片. 而半身 - 半身、全身 - 半身的“部分”行人重识别技术便显得非常重要, 第一篇研究该问题的深度学习论文已经被 CVPR2018 会议接受[73]. 而夜间光照不佳情况下的行人重识别也是一个值得研究的问题. 目前的主流思路还是用红外相机在黑暗条件下采集图片, 随之引申出来的是红外行人重识别. 红外图片几乎只有轮廓, 失去了颜色信息给重识别任务带来了非常大的挑战. 除了以上举得例子, 其他场景的一些跨域行人重识别也值得关注.

5) 深度网络的可解释性: 虽然深度学习的方法在行人重识别任务上取得了很好的性能, 但是在准确度不断被刷高的背后, 很少有研究工作表明哪些信息对行人的识别影响更大. 无论是全局特征还是局部特征, 单帧图像还是序列图像, 我们都在设计更加合理的网络结构或者网络损失来学习更加有效的特征. 然而, 到底是颜色信息还是轮廓信息对识别影响更大, 或者姿态如何对齐、光线如何矫正给性能提升更大我们都不甚明确. 随着深度学习可视化技术的提升, 行人重识别网络的可解释性会在将来取得突破.

6) 行人重识别与行人检测、行人跟踪的结合: 目前大部分的方法是在已经检测出行人的先验条件下进行的. 但是这需要一个非常鲁邦的行人检测模型, 如果行人重识别与行人检测结合起来, 则更加符合实际的应用需求. 这方面的研究工作很少, ICCV2017 的一篇工作可以给予一定启示[76]. 此外, 行人重识别最直接的一个应用便是跨摄像头多目标跟踪 (Multi-target Multi-camera tracking, MTMC tracking). 因此融合行人重识别和 MTMC 跟踪的问题也是行人重识别研究未来的一个延伸.

6 结束语

行人重识别是计算机视觉领域的一个热门研究可以, 而深度学习的发展极大地促进了该领域的研究. 近几年的顶级会议 ICCV、CVPR 和 ECCV 上, 每年都有十篇以上的行人重识别研究发表, 并且绝大部分都是基于深度学习的工作. 本文总结了近年来基于深度学习的行人重识别方法, 从表征学习、度量学习、局部特征、视频序列和 GAN 网络为切入点, 进行了详细的讨论, 并展望了该领域未来可能的研究方向.

注意, 支持此研究的基金项目请直接标注在首页页脚处, 不用在此再次致谢.

References

1 Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS). Montréal, Canada: MIT Press, 2014. 2672−2680

2 Zajdel W, Zivkovic Z, Krose B J A. Keeping track of humans: have I seen this person before? In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation. Barcelona, Spain: IEEE, 2005. 2081−2086

3 Zheng L, Yang Y, Hauptmann A G. Person re-identification: past, present and future. arXiv preprint arXiv:1610.02984, 2016.

4 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 886−893

5 Lowe D G. Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE, 1999. 1150−1157

6 Köstinger M, Hirzer M, Wohlhart P, Both P M, Bischof H. Large scale metric learning from equivalence constraints. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 2288−2295

7 Liao S C, Hu Y, Zhu X Y, Li S Z. Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 2197−2206

8 He K M, Zhang X Y, Ren S Q, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1026−1034

9 Lu C C, Tang X O. Surpassing human-level face verification performance on LFW with Gaussian face. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas, USA: AAAI, 2015. 3811−3819

10 Gray D, Brennan S, Tao H. Evaluating appearance models for recognition, reacquisition, and tracking. In: Proceedings of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). Rio de Janeiro: IEEE, 2007. 1−7

11 Hirzer M, Beleznai C, Roth P M, Bischof H. Person re-identification by descriptive and discriminative classification. In: Proceedings of Scandinavian Conference on Image Analysis. Berlin, Heidelberg: Springer, 2011. 91−102

12 Li W, Zhao R, Xiao T, Wang X G. DeepReID: deep filter pairing neural network for person re-identification. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 152−159

13 Zheng L, Shen L Y, Tian L, Wang S J, Wang J D, Tian Q. Scalable person re-identification: a benchmark. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1116−1124

14 Xiao T, Li S, Wang B C, Lin L, Wang X G. End-to-end deep learning for person search. arXiv preprint arXiv:1604.01850, 2016.

15 Zheng L, Bie Z, Sun Y F, Wang J D, Su C, Wang S J, et al. Mars: a video benchmark for large-scale person re-identification. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2016. 868−884

16 Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C. Performance measures and a data set for multi-target, multi-camera tracking. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2016. 17−35

17 Wu A C, Zheng W S, Yu H X, Gong S G, Lai J H. RGB-infrared cross-modality person re-identification. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 5390−5399

18 Song G, Leng B, Liu Y, Hetang C, Cai S F. Region-based quality estimation network for large-scale person re-identification. In: Proceedings of Association for the Advancement of Artificial Intelligence. New Orleans: AAAI, 2018.

19 Wei L H, Zhang S L, Gao W, Tian Q. Person transfer GAN to bridge domain gap for person re-identification. arXiv:1711.08565, 2018.

20 Li J N, Zhang S L, Wang J D, Gao W, Tian Q. LVreID: person re-identification with long sequence videos. arXiv preprint arXiv:1712.07286, 2017.

21 Felzenszwalb P F, Girshick R B, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(9): 1627−1645

22 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS). Montreal, Quebec, Canada: MIT Press, 2015. 91−99

23 Dehghan A, Modiri Assari S, Shah M. GMMCP tracker: globally optimal generalized maximum multi clique problem for multiple object tracking. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 4091−4099

24 Karanam S, Gou M R, Wu Z Y, Rates-Borras A, Camps O, Radke R J. A Systematic evaluation and benchmark for person re-identification: features, metrics, and datasets. arXiv preprint arXiv:1605.09653, 2016.

25 Geng M Y, Wang Y W, Xiang T, Tian Y H. Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244, 2016.

26 Lin Y T, Zheng L, Zheng Z D, Wu Y, Yang Y. Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220, 2017.

27 Matsukawa T, Suzuki E. Person re-identification using CNN features learned from combination of attributes. In: Proceedings of the 23rd International Conference on Pattern Recognition. Cancun, Mexico: IEEE, 2016. 2428−2433

28 Zhang Y, Xiang T, Hospedales T M, Lu H C. Deep mutual learning. arXiv:1706.00384, 2017.

29 Zheng L, Zhang H H, Sun S Y, Chandraker M, Yang Y, Tian Q. Person re-identification in the wild. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 3346−3355

30 Zheng Z D, Zheng L, Yang Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 3774−3782

31 Zheng Z D, Zheng L, Yang Y. A discriminatively learned CNN embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2018, **14**(1): Article No. 13

32 Shi H L, Yang Y, Zhu X Y, Liao S C, Lei Z, Zheng W S, et al. Embedding deep metric for person re-identification: a study against large variations. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2016. 732−748

33 Varior R R, Haloi M, Wang G. Gated Siamese convolutional neural network architecture for human re-identification. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2016. 791−808

34 Varior R R, Shuai B, Lu J W, Xu D, Wang G. A Siamese long short-term memory architecture for human re-identification. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2016. 135−153

35 Wang Y C, Chen Z Z, Wu F, Wang G. Person re-identification with cascaded pairwise convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, Utah, USA: IEEE, 2018. 1470−1478

36 Cheng D, Gong Y H, Zhou S P, Wang J J, Zheng N N. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 1335−1344

37 Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.

38 Liu H, Feng J S, Qi M B, Jiang J G, Yan S C. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017, **26**(7): 3492−3506

39 Ristani E, Tomasi C. Features for multi-target multi-camera tracking and re-identification. arXiv:1803.10859, 2018.

40 Chen W H, Chen X T, Zhang J G, Huang K Q. Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 1320−1329

41 Xiao Q Q, Luo H, Zhang C. Margin sample mining loss: a deep learning based method for person re-identification. arXiv preprint arXiv:1710.00478, 2017.

42 Liu H, Feng J S, Qi M B, Jiang J G, Yan S C. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017, **26**(7): 3492−3506

43 Xiao Q Q, Cao K L, Chen H N, Peng F Y, Zhang C. Cross domain knowledge transfer for person re-identification. arXiv preprint arXiv:1611.06026, 2016.

44 Zhang X, Luo H, Fan X, Xiang W L, Sun Y X, Xiao Q Q, et al. AlignedReID: surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184, 2017.

45 Zhao H Y, Tian M Q, Sun S Y, Shao J, Yan J J, Yi S, et al. Spindle net: person re-identification with human body region guided feature decomposition and fusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 907−915

46 Zheng L, Huang Y J, Lu H C, Yang Y. Pose invariant embedding for deep person re-identification. arXiv preprint arXiv:1701.07732, 2017.

47 Dai J, Zhang P P, Lu H C, Wang H Y. Video person re-identification by temporal residual learning. arXiv preprint arXiv:1802.07918, 2018.

48 Li Y J, Zhuo L, Li J F, Zhang J, Liang X, Tian Q. Video-based person re-identification by deep feature guided pooling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA: IEEE, 2017. 1454−1461

49 Liu H, Jie Z Q, Jayashree K, Qi M B, Jiang J G, Yan S C, et al. Video-based person re-identification with accumulative motion context. *IEEE Transactions on Circuits and Systems for Video Technology*, to be published

50 Ma X L, Zhu X T, Gong S G, Xie X D, Hu J M, Lam K M, et al. Person re-identification by unsupervised video matching. *Pattern Recognition*, 2017, **65**: 197−210

51 McLaughlin N, Martinez del Rincon J, Miller P. Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 1325−1334

52 Wang T Q, Gong S G, Zhu X T, Wang S J. Person re-identification by video ranking. In: Proceedings of European Conference on Computer Vision. Cham: Springer, 2014. 688−703

53 Wang T Q, Gong S G, Zhu X T, Wang S J. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(12): 2501−2514

54 You J J, Wu A C, Li X, Zheng W S. Top-push video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 1345−1353

55 Zhang D Y, Wu W X, Cheng H, Zhang R M, Dong Z J, Cai Z Q. Image-to-video person re-identification with temporally memorized similarity learning. *IEEE Transactions on Circuits and Systems for Video Technology*, to be published

56 Zhang W, Ma B P, Liu K, Huang R. Video-based pedestrian re-identification by adaptive spatio-temporal appearance model. *IEEE Transactions on Image Processing*, 2017, **26**(4): 2042−2054

57 Zhao R, Ouyang W L, Wang X G. Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(2): 356−370

58 Wu Y, Lin Y T, Dong X Y, Yan Y, Ouyang W L, Yang Y. Exploit the unknown gradually: one-shot video-based person re-identification by stepwise learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 5177−5186

59 Zhu J Y, Park T, Isola P, Efros A A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2242−2251

60 Yi Z L, Zhang H, Tan P, Gong M L. DualGAN: unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2868−2876

61 Kim T, Cha M, Kim H, Lee J K, Kim J. Learning to discover cross-domain relations with generative adversarial networks. arXiv:1703.05192, 2017.

62 Deng W J, Zheng L, Ye Q X, Kang G L, Yang Y, Jiao J B. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. arXiv:1711.07027, 2018.

63 Huang Y, Xu J S, Wu Q, Zheng Z D, Zhang Z X, Zhang J. Multi-pseudo regularized label for generated data in person re-identification. arXiv preprint arXiv:1801.06742, 2018.

64 Qian X L, Fu Y W, Xiang T, Wang W X, Qiu J, Wu Y, et al. Pose-normalized image generation for person re-identification. arXiv preprint arXiv:1712.02225, 2018.

65 Zhong Z, Zheng L, Zheng Z D, Li S Z, Yang Y. Camera style adaptation for person re-identification. arXiv:1711.10295, 2018.

66 Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. arXiv:1606.03657, 2016.

67 Zhao L M, Li X, Zhuang Y T, Wang J D. Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 3239−3248

68 Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, et al. MobileNets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.

69 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 770−778

70 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, Massachusetts, USA: IEEE, 2015. 1−9

71 Sarfraz M S, Schumann A, Eberle A, Stiefelhagen R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. arXiv:1711.10378, 2018.

72 Jiao J N, Zheng W S, Wu A C, Zhu X T, Gong S G. Deep low-resolution person re-identification. In: Proceedings of Association for the Advancement of Artificial Intelligence. New Orleans: AAAI, 2018.

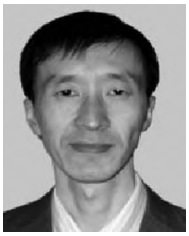
73 He L X, Liang J, Li H Q, Sun Z. Deep spatial feature reconstruction for partial person re-identification: alignment-free approach. arXiv:1801.00881, 2018.

74 Fan H H, Zheng L, Yang Y. Unsupervised person re-identification: clustering and fine-tuning. arXiv preprint arXiv:1705.10444, 2017.

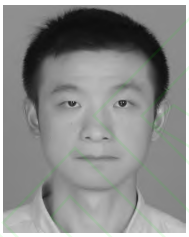
75 Zhong Z, Zheng L, Cao D L, Li S Z. Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, Hawaii, USA: IEEE, 2017. 3652−3661

76 Liu H, Feng J S, Jie Z Q, Jayashree K, Zhao B, Qi M B, et al. Neural person search machines. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 493−501

罗 浩 浙江大学控制科学与工程学院智能系统与控制研究所博士研究生. 2011 年获得浙江大学控制科学与工程学士学位. 主要研究方向为行人重识别、多目标跟踪、深度学习、计算机视觉方向.
E-mail: haoluocsc@zju.edu.cn
(**LUO Hao** Ph. D. candidate at the Institute of Cyber-Systems and Control, College of Control Science and Engineering, Zhejiang University. He received his bachelor degree from College of Control Science and Engineering, Zhejiang University in 2011. His research interest covers person re-identification, multi-target multi-camera tracking, deep learning and computer vision.)



姜 伟 于 2005 年获得日本东京工业大学博士学位. 现为浙江大学控制科学与工程学院智能系统与控制研究所副教授. 主要研究方向为机器视觉、计算机图形学、机器学习. 本文通信作者.
E-mail: jiangwei_zju@zju.edu.cn
(**JIANG Wei** He received his Ph.D. degree in Dept. of Systems and Control Engineering, School of Engineering from Tokyo institute of technology in 2005. Currently, he is a associate professor at the Institute of Cyber-Systems and Control, College of Control Science and Engineering, Zhejiang University. His research interests include machine vision, computer graphics, and machine learning. Corresponding author of this paper.)



范 星 浙江大学控制科学与工程学院博士研究生. 2011 年获得浙江大学控制科学与工程学士学位. 主要研究方向为行人重识别.
E-mail: xfanplus@zju.edu.cn
(**FAN Xing** Ph. D. candidate at the College of Control Science and Engineering, Zhejiang University. He received his bachelor degree from College of Control Science and Engineering, Zhejiang University in 2011. His research interest covers person re-identification.)

张思朋 浙江大学控制科学与工程学院研究生. 2012 年获得浙江大学控制科学与工程学士学位. 主要研究方向为计算机视觉、行人重识别.
E-mail: zhangsipeng@zju.edu.cn
(**ZHANG Si-Peng** Graduate student at college of control science and engineering, Zhejiang university. She received her bachelor degree from College of Control Science and Engineering, Zhejiang University in 2012. Her research interest covers computer vision and person re-identification.)