

分类号: _____

单位代码: _____

学 号: _____

浙江大学

博士学位论文开题报告



中文论文题目: 基于行人重识别的跨摄像头
多目标跟踪方法研究

英文论文题目: Study on multi-target multi-camera tracking
based on person re-identification

姓名: 罗浩

导师: 姜伟

专业: 控制科学与工程

学号: 11532034

学院: 控制学院

报告日期 2017年11月

摘 要

关键词：行人重识别，跨摄像头多目标跟踪，深度学习，卷积神经网络

目 次

摘要	I
目次	
1 研究意义与研究背景	1
2 研究现状与文献综述	3
2.1 行人重识别	3
2.1.1 相关数据集	3
2.1.2 准确度评估准则	5
2.1.3 基于表征学习的方法	7
2.1.4 基于度量学习的方法	8
2.1.5 基于局部特征的方法	11
2.1.6 基于视频序列的方法	14
2.2 跨摄像头多目标跟踪	16
2.2.1 相关数据集	16
2.2.2 准确度评估准则	19
2.2.3 表观模型	21
2.2.4 基于相关滤波的多目标跟踪	22
2.2.5 基于代价函数的多目标跟踪	23
2.2.6 跨摄像头匹配	23
3 研究内容与技术路线	24
3.1 研究内容	24
3.2 技术路线	25
3.2.1 基于深度学习的行人重识别技术	25
3.2.2 基于行人重识别特征的跨摄像头多目标跟踪技术	26
4 现有成果与研究计划	27
4.1 现有成果	27

4.1.1	边界样本挖掘损失	27
4.1.2	最短路径距离	30
4.1.3	基于度量学习的互学习方法	33
4.1.4	行人重识别的人类准确度评估	36
4.1.5	与现有方法结果对比	37
4.2	成果作品	40
4.3	研究计划	40
	参考文献	41

1 研究意义与研究背景

随着经济和社会的发展，监控视频已经广泛运用于安防、商业、工业生产以及智能机器人等各个领域。据统计世界上硬盘有一半左右用于存储监控视频，可见监控视频在我们的日常生活中占据着很重要的地位。而监控视频里面最主要的关注对象就是行人，理解行人的行为对于违章判断、刑事侦查以及危险预警等都有着非常重要的意义。而如何识别、定位以及跟踪行人是理解行人行为的前提，因此行人重识别(Person re-identification, person ReID)和跨摄像头多目标跟踪(Multi-target multi-camera tracking, MTMC tracking)是实现以上目标的第一步。

目前在跟踪问题的研究上，大部分工作都集中在单摄像头的单目标跟踪。单摄像头的单目标跟踪虽然并没有完全解决，但是已经拥有了非常多不错的成果，也在一些商业化的产品中得到了应用。基于单目标跟踪方法的扩展，单摄像头的多目标跟踪问题也有一些非常经典的研究成果，虽然算法的性能还没有完全达到商业化落地的程度，但是学术界在以一个非常良好的速度持续推进研究。然而单摄像头跟踪问题上，无论是单目标还是多目标都很少出现在视野中丢失的现象，主要的困难也是遮挡、姿态以及光纤等。但是单摄像头的视野很局限，此外也无法对目标进行三维的跟踪，而多摄像头的跟踪系统能够很好地克服单摄像头跟踪系统中的这些不足之处。因此，跨摄像头的监控视频系统正在逐渐得到关注，并开始得到学者们的深入研究。但是跨摄像头跟踪不仅要面临单摄像头跟踪所需要面临的问题，更重要的是需要面对跟踪目标在一个摄像头中消失，之后可能会在其他摄像头中出现的问题。而把从一个摄像头视野中丢失的行人目标在其他摄像头拍摄的视频中找回来的过程就称为行人重识别。所以行人重识别是跨摄像头多目标跟踪的基础。

行人重识别最直接、最典型的应用就是跨摄像头的多目标跟踪，但是作为一个独立的研究课题，还拥有很多有价值的应用场景。行人重识别是人脸识别无法使用的场景下的替代品，因此人脸识别可以使用的场景理论上行人重识别都可以得到应用，而人脸识别是目前我们生活中几乎离不开的应用，所以行人重识别的应用前景是值得期待的。对于人脸识别，目前大部分的学术界以及著名的企业所能达到的技术界限是人脸照片需要不低于 32×32 分辨率，如果要实现高精度的人脸识别，通常需要 96×96 像素的人脸正面图片。

但是这种要求在监控视频中是很难满足的，为了得到足够大的监控视野，监控摄像头所能拍摄的人脸图片一般只有十至二十个像素左右。即使存在比较高清的人脸图片，也很难保证人脸是正对相机拍摄，所以在监控视频中人脸识别技术的使用非常受限。而在人脸识别技术无法使用的场景下，行人重识别技术就是识别人体目标的替代品技术。所以作为一个独立的研究课题，抛开跨摄像头多目标跟踪，行人重识别也有非常大的研究价值。对于跨摄像头多目标跟踪，传统的方法是基于一些统计方法的时空数据关联技术。但是这种方法非常依赖统计的先验知识，可以一定概率上关联多个摄像头中的目标，但是效果的局限性也是显而易见的。所以基于行人重识别的跨摄像头匹配如今成为了跨摄像头多目标跟踪最主流的解决思路。而在单摄像头的目标跟踪问题上，行人重识别并非唯一的解决思路，但是在行人跟踪问题上是一种非常好的外观特征(Appearance feature)。综上，一个好的行人重识别方法对于跨摄像头多目标跟踪问题而言有这非常重要的意义。

另外，在目前学术界的研究之中，学者大部分将行人重识别和多目标跟踪作为独立的两个研究课题分别研究。然而这两个课题其实是相辅相成的，行人重识别是跨摄像头多目标跟踪的基础，而行人的跟踪序列也能弥补单帧图像信息的局限辅助行人重识别的研究。目前行人重识别的方法大部分还是依赖单帧图像，但是单帧图像的信息终究有局限，在遇到遮挡等情况下基本无法解决，此时序列信息中往往可以找到至少一帧优质的图像来进行重识别。此外，当目标外观十分接近的时候，通过图像内容特征也许并不能区别目标，此时序列的运动信息就可能发挥其作用。目前，基于视频序列的行人重识别研究也逐渐开始受到关注。但是跟踪序列的标注代价十分昂贵，目前行人重识别的行人框标注基本都是有计算机自主完成，这得益于行人检测算法的发展。而行人跟踪序列基本还只能靠人工标注，所以一个好的跟踪检测模型也是十分必要的。因此本文把两个课题结合起来，有着十分大的研究意义。

2 研究现状与文献综述

本章节主要介绍本课题相关的研究现状，包括行人重识别(Person re-identification, person ReID)和跨摄像头多目标跟踪(Multi-target multi-camera tracking, MTMC tracking)两个部分。在本章节将会分别介绍这两个子课题相关的数据集和现有算法。其中行人重识别着重介绍近几年深度卷积神经网络相关的方法，而跨摄像头多目标跟踪将会着重介绍基于行人重识别的方法。

2.1 行人重识别

行人重识别也称行人再识别，是利用计算机视觉技术判断图像或者视频序列中是否存在特定行人的技术。广泛被认为是一个图像检索的子问题。给定一个监控行人图像，检索跨设备下的该行人图像。旨在弥补目前固定的摄像头的视觉局限，并可与行人检测/行人跟踪技术相结合，可广泛应用于智能视频监控、智能安保等领域。

而对于跨摄像头多目标跟踪问题，当一个行人目标在其中一个摄像头视野中消失后，要把该行人在其他摄像头中再次识别出来，这就是典型的行人重识别问题。也就是说，行人重识别技术是跨摄像头多目标跟踪的基础。因此，在本小节将会先介绍现有的行人重识别相关的数据集、准确度评估准则和一些现有的主流方法。

2.1.1 相关数据集

行人重识别相关的数据集总共有十几个，在早年深度学习还未出现的时候，那时的数据集图片数量还比较少。随着深度学习的诞生，行人重识别问题对数据量的要求大大增加，本小节将介绍几个适用深度学习的大规模行人识别数据集。

- Market1501

Market1501^[1]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像素。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，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。

- MARS

MARS (Motion Analysis and Re-identification Set)^[2]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1,267个行人的20,478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。

- CUHK03

CUHK03^[3]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。

- CUHK-SYSU

CUHK-SYSU^[4]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。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。

- DukeMTMC-reID

DukeMTMC-reID^[5]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。该数据集是目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。

- VIPeR

VIPeR^[6]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，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。

- PRID2011

PRID2011^[7]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，所以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。

以上是目前行人重识别研究中主要运用的数据集。由于行人重识别图片采自于不同摄像头，所以会出现光照、行人姿态、拍摄角度、遮挡、图像模糊等问题，造成同一行人的图片在不同摄像头中表现差异很大。如图2-1所示，上一排与下一排为同一个行人在两个不同摄像头拍摄的图片。可以看出，第一列存在遮挡现象，第二列至第四列存在拍摄角度、姿态等的巨大差异，第五列由于拍摄距离不同造成行人占图像比例大小差异很大，而最后一列是典型的摄像头分辨率不同而造成的图像差异。正式因为各种因素造成的图像差异，使得行人重识别很难通过手动提取特征就达到很好的识别效果，需要通过一定手段来学习到非常鲁棒的图像特征。

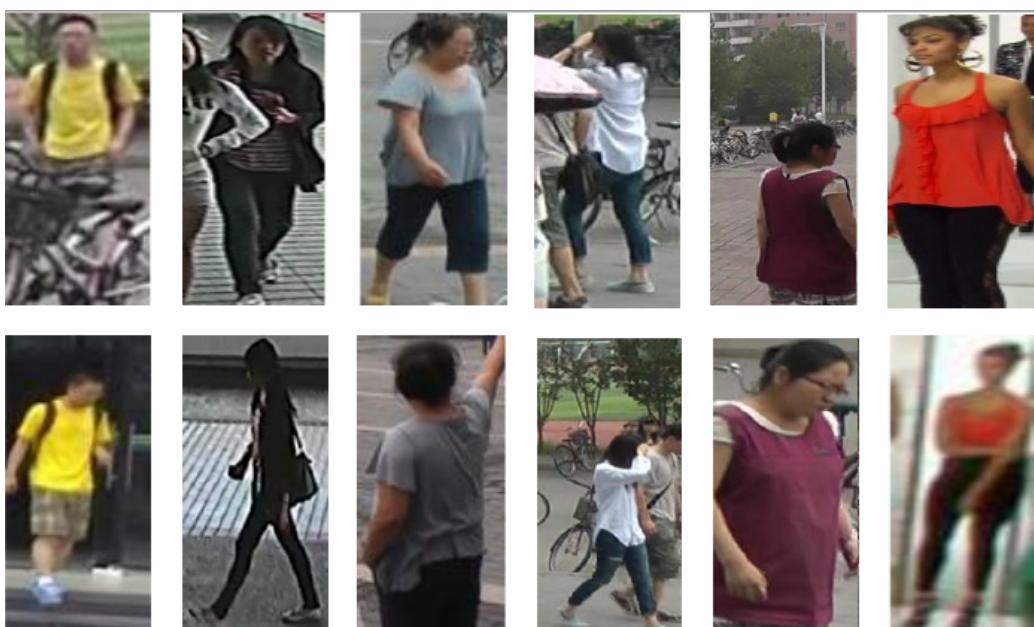


图 2-1 行人重识别数据集图片示例

2.1.2 准确度评估准则

为了评估行人重识别算法的优劣，需要统一一些评价准则。在学术论文中，通常大家默认选择累计匹配(Cumulative Match Characteristics, CMC)曲线和平均准确度(Mean Average

Precision, mAP)来作为评价准则。CMC和mAP是检索问题中常用的评价准则，在介绍它们之前，我们先介绍一些要用到常用术语。

- **query:** 指测试集中的待检索库,包含图片的数目为 N_q 。
- **gallery:** 指测试集中的搜索库。
- **probe:** 指query中的某张待检索的图片，测试时需要将gallery中和probe为同一行人的图片全部检索出来。

(1) CMC曲线

CMC曲线主要用于计算rank-k的击中概率，在行人重识别、人脸识别领域使用较多。针对query集中的一张带检索的probe图片，返回gallery的一系列排好序的结果，排序按照相似度排序。越靠前的结果表示和probe图片越相似，在行人重识别领域也等同于和probe是同一个人的概率越高。在测试阶段，需要排除gallery集中和probe处于同一摄像头的图片，防止其参与检索排序。我们设 $index_{probe}$ 表示和gallery和probe为相同行人的最靠前的排序结果。最后rank-k准确度 $A(rank-k)$ 可以表示为：

$$A(rank-k) = \frac{\sum_{probe \in query} f_{CMC}(index_{probe}, k)}{N_q} \quad (2-1)$$

其中：

$$f_{CMC}(index_{probe}, k) = \begin{cases} 0 & index_{probe} > k \\ 1 & index_{probe} \leq k \end{cases} \quad (2-2)$$

在实际使用中，为了减少计算量，通常我们比较关心rank-1,rank-5,rank-10,rank-20等准确度。

(2) mAP

mAP是另外一种重要的评价指标。CMC曲线通常只关心检索库中最靠前的正样本排序，而mAP由gallery中所有正样本的排序结果决定，所以通常能够更加鲁邦地反映模型的性能。计算mAP需要以下三步：

- (1) Precision: 对于query中的某一张probe图片，返回了gallery的一系列排序结果，考虑前 n 个查询结果， $P(n)=$ 前 n 个结果中与probe图片是相同行人的数目/ n ；
- (2) Average Precision: 对于query的第 K 个probe图片，记录排序结果中所有M个正样本排序结果的集合 $\{i_1, i_2, \dots, i_M\}$ ，计算它们的平均Precision，即 $AP_K = \sum P(i)/M$ ，其中 $i \in \{i_1, i_2, \dots, i_M\}$ ；

(3) Mean Average Precision (mAP): 所有 N_q 张probe图片的Average Precision 的平均值, 即 $mAP = \sum_K AP_K/N$ 。

2.1.3 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^[8-11]。这主要得益于深度学习, 尤其是卷积神经网络(Convolutional neural network, CNN)^[12]的快速发展。由于CNN可以自动从原始的图像数据中根据任务需求自动提取出表征特征(Representation), 所以有些研究者把行人重识别问题看做分类(Classification/Identification)问题或者验证(Verification)问题。分类问题是利用行人的ID或者属性等作为训练标签来训练模型。验证问题是输入一对(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。

论文^[8]利用Classification/Identification loss和verification loss来训练网络, 其网络示意图如图2-2所示。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来计算分类误差损失。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。经过足够数据的训练, 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。

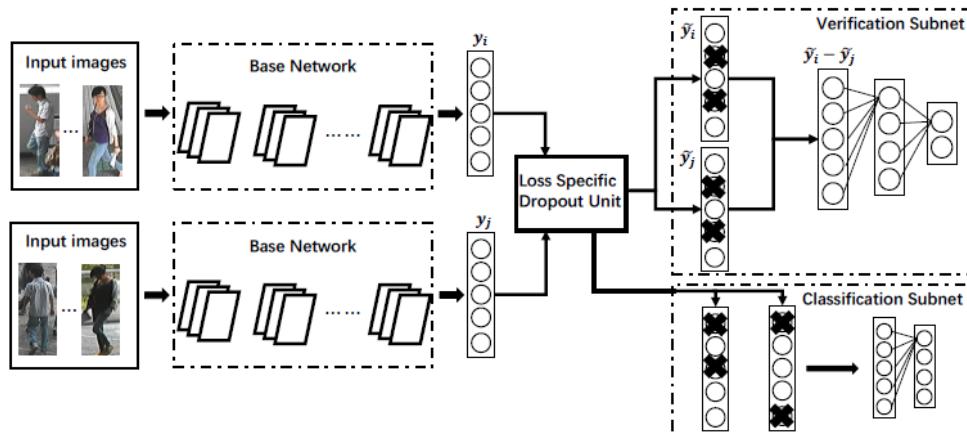


Figure 1. The proposed deep Re-ID network architecture.

图 2-2 结合分类损失和验证损失训练ReID网络示意图

论文^[9-11]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。在这些工作中, 它们额外标注了行人图片的属性特征, 例如性别、头发、衣着等属性。通过引入行人属性标签, 模型不但要准确地预测出行人ID, 还要预测出各项正确的行人属性, 这大大增加了模型的泛化能力, 多数论文也显示这种方法是有效的。图2-3是其中一个示例, 从

图中可以看出，网络输出的特征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。

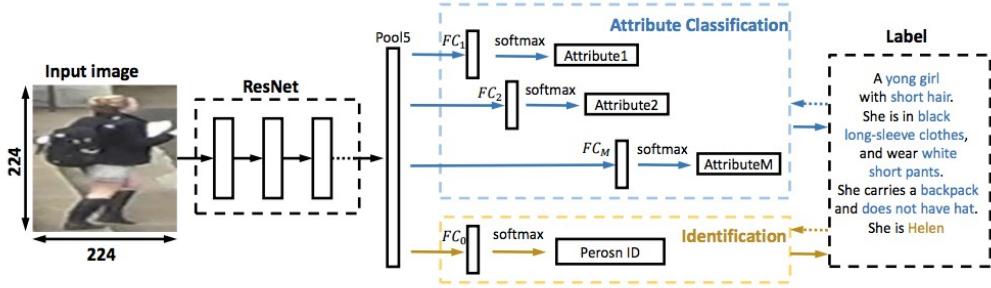


Figure 2. An overview of the APR network. During training, it predicts M attribute labels and an ID label. The weighted sum of the individual losses is back propagated. During testing, we extract the Pool5 (ResNet-50) or FC7 (CaffeNet) descriptors for retrieval.

图 2-3 结合行人ID标注和行人属性训练ReID网络示例

2.1.4 基于度量学习的方法

度量学习(Metric learning)是广泛用于图像检索利的一种方法。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，具体为同一行人的不同图片相似度大于不同行人的不同图片。最后网络的损失函数使得相同行人图片(正样本对)的距离尽可能小，不同行人图片(负样本对)的距离尽可能大。常用的度量学习损失方法有对比损失(Contrastive loss)^[13]、三元组损失(Triplet loss)^[14-16]、四元组损失(Quadruplet loss)^[17]。首先，假如有两张输入图片 I_1 和 I_2 ，通过网络的前馈我们可以得到它们归一化后的特征向量 f_{I_1} 和 f_{I_2} 。我们定义这两张图片特征向量的欧式距离为：

$$d_{I_1, I_2} = \|f_{I_1} - f_{I_2}\|_2 \quad (2-3)$$

(1) 对比损失(Contrastive loss)

对比损失用于训练孪生网络(Siamese network)，其结构图如图2-4所示。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：

$$L_c = y d_{I_a, I_b}^2 + (1 - y)(\alpha - d_{I_a, I_b})_+^2 \quad (2-4)$$

其中 $(z)_+$ 表示 $\max(z, 0)$ ， α 是根据实际需求设计的阈值参数。为了最小化损失函数，当网络输入一对正样本对， $d(I_a, I_b)$ 会逐渐变小，即相同ID的行人图片会逐渐在特征空间形成聚类。反之，当网络输入一对负样本对时， $d(I_a, I_b)$ 会逐渐变大直到超过设定的 α 。

通过最小化 L_c , 最后可以使得正样本对之间的距离逐渐变下, 负样本对之间的距离逐渐变大, 从而满足行人重识别任务的需要。

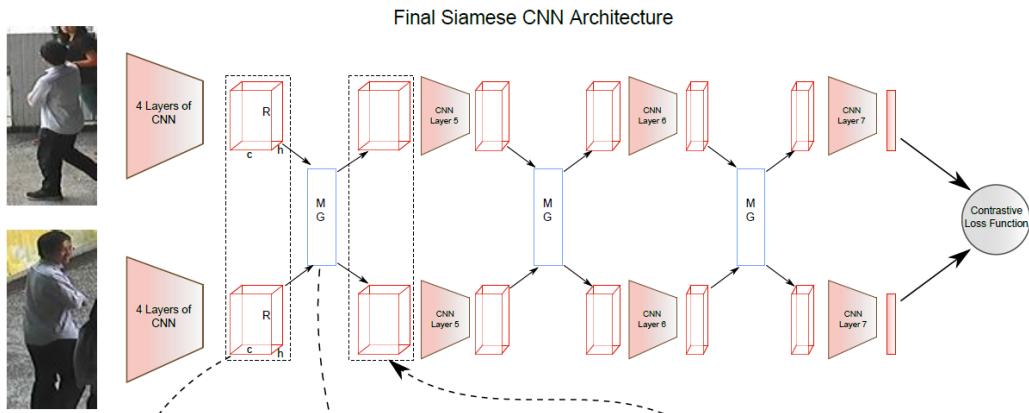


图 2-4 孪生网络结构示意图

(2) 三元组损失(Triplet loss)

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:

$$L_t = (d_{a,p} - d_{a,n} + \alpha)_+ \quad (2-5)$$

如图2-5所示, 三元组可以拉近正样本对之间的距离, 推开负样本对之间的距离, 最后使得相同ID的行人图片在特征空间里形成聚类, 达到行人重识别的目的。

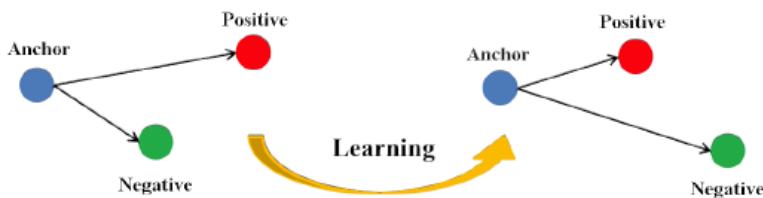


Figure 4. Triplet loss.

图 2-5 三元组损失^[18]

论文^[16]认为公式(4-1)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):

$$L_it = d_{a,p} + (d_{a,p} - d_{a,n} + \alpha)_+ \quad (2-6)$$

公式(2-6)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。

(3) 四元组损失(Quadruplet loss)

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(Anchor) a ，正样本图片(Positive) p ，负样本图片1(Negative1) n_1 和负样本图片2(Negative2) n_2 。其中 n_1 和 n_2 是两张不同行人ID的图片，其结构如图2-6所示。则，四元组损失表示为：

$$L_q = (d_{a,p} - d_{a,n_1} + \alpha)_+ + (d_{a,p} - d_{n_1,n_2} + \beta)_+ \quad (2-7)$$

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。相比于三元组损失只考虑正负样本间的相对距离，四元组添加的第二项不共享ID，所以考虑的是正负样本间的绝对距离。因此，四元组损失通常能让模型学习到更好的表征。

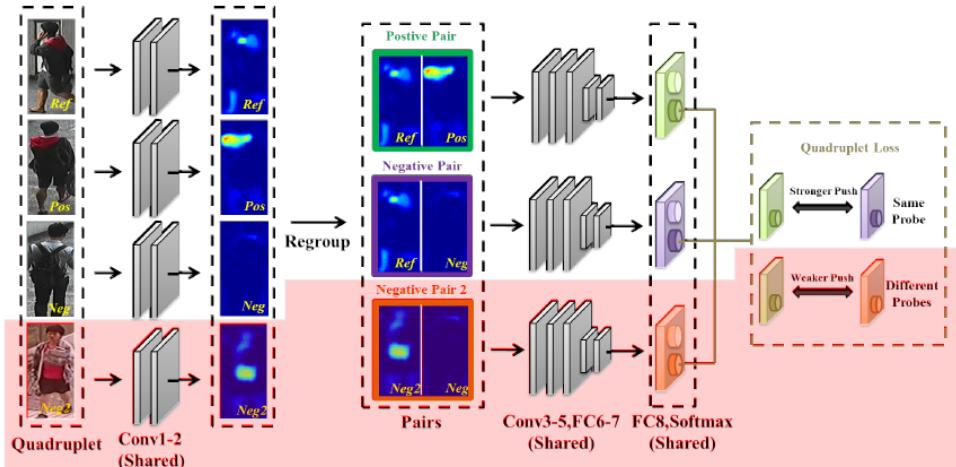


Figure 3. The framework of the proposed quadruplet deep network. The red shadow region indicates elements of the new constraint.

图 2-6 四元组损失网络结构图

(4) 难样本采样三元组损失(Triplet loss with hard sample mining)

难样采样三元组损失（本文之后用TriHard损失表示）是三元组损失的改进版。传统的三元组随机从训练数据中抽样三张图片，这样的做法虽然比较简单，但是抽样出来的大部分都是简单易区分的样本对。如果大量训练的样本对都是简单的样本对，那么这是不利于网络学习到更好的表征。大量论文发现用更难的样本去训练网络能够提高网络的泛化能力，而采样难样本对的方法很多。论文^[19]提出了一种基于训练批量(Batch)的在线难样本采样方法——TriHard损失。

TriHard损失的核心思想是：对于每一个训练batch，随机挑选 P 个ID的行人，每个行人随机挑选 K 张不同的图片，即一个batch含有 $P \times K$ 张图片。之后对于batch中的每一张图片 a ，我们可以挑选一个最难的正样本和一个最难的负样本和 a 组成一个三元组。

首先我们定义和 a 为相同ID的图片集为 A ，剩下不同ID的图片图片集为 B ，则**TriHard**损失表示为：

$$L_{th} = \frac{1}{P \times K} \sum_{a \in batch} (\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha)_+ \quad (2-8)$$

其中 α 是人为设定的阈值参数。**TriHard**损失会计算 a 和batch中的每一张图片在特征空间的欧式距离，然后选出与 a 距离最远（最不像）的正样本 p 和距离最近（最像）的负样本 n 来计算三元组损失。通常**TriHard**损失效果比传统的三元组损失要好。

2.1.5 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。而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姿态矫正等等。

图片切块是一种很常见的提取局部特征方式^[20,21]。如图2-7所示，图片被垂直等分为若干份，因为垂直切割更符合我们对人体识别的直观感受，所以行人重识别领域很少用到水平切割。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。但是这种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。

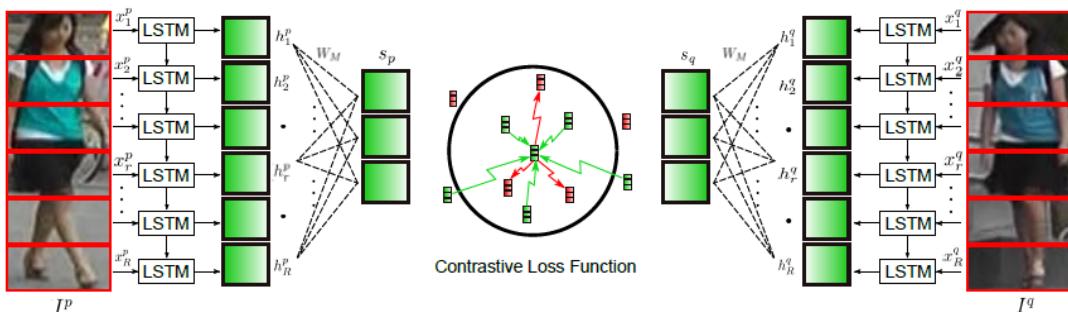


图 2-7 利用图片切块提取局部特征示例

为了解决图像不对齐情况下手动图像切片失效的问题，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(Skeleton)模型。论文^[22]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。如图2-8所示，一个行人通常被分为14个关键点，这14个关键点把人体结果分为若干个区域。为了提取不同尺度上的局部特征，作者设定了三个不同的PoseBox组合。之后这三个PoseBox矫正后的图片和原始未矫正的图片一起送到网络里去提取特征，这个特征包含了全局信息和局部信息。特别提出，如果这个仿射变换可以在进入网络之前的预处理中进行，也可以在输入到网络后进行。如果是后者的话需要对仿射变换做一个改进，因为传统的放射变化是不可导的。为了使得网络可以训练，需要引入可导的近似放射变化，在本文中不赘述相关知识。

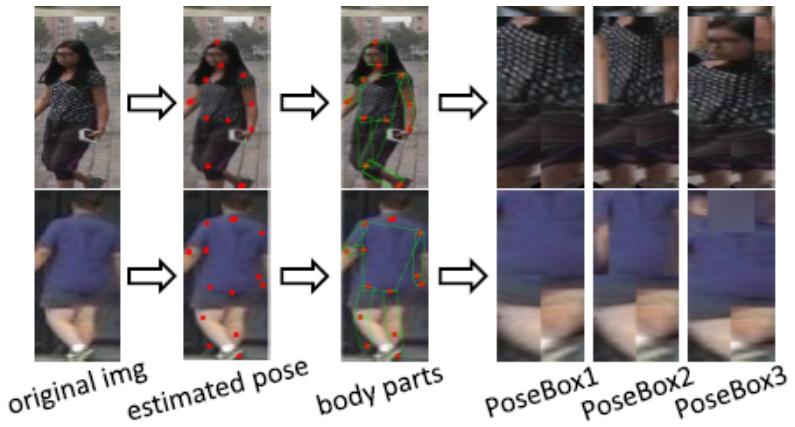


图 2-8 姿态对齐示意图

CVPR2017的工作Spindle Net^[23]也利用了14个人体关键点来提取局部特征。和论文^[22]不同的是，Spindle Net并没有用仿射变换来对齐局部图像区域，而是直接利用这些关键点来抠出感兴趣区域(Region of interest, ROI)。Spindle Net网络如图2-9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。网络中所有提取特征的CNN（橙色表示）参数都是共享的，这个CNN分成了线性的三个子网络FEN-C1、FEN-C2、FEN-C3。对于输入的一张行人图片，有一个预训练好的骨架关键点提取CNN（蓝色表示）来获得14个人体关键点，从而得到7个ROI区域，其中包括三个大区域（头、上身、下身）和四个四肢小区域。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。

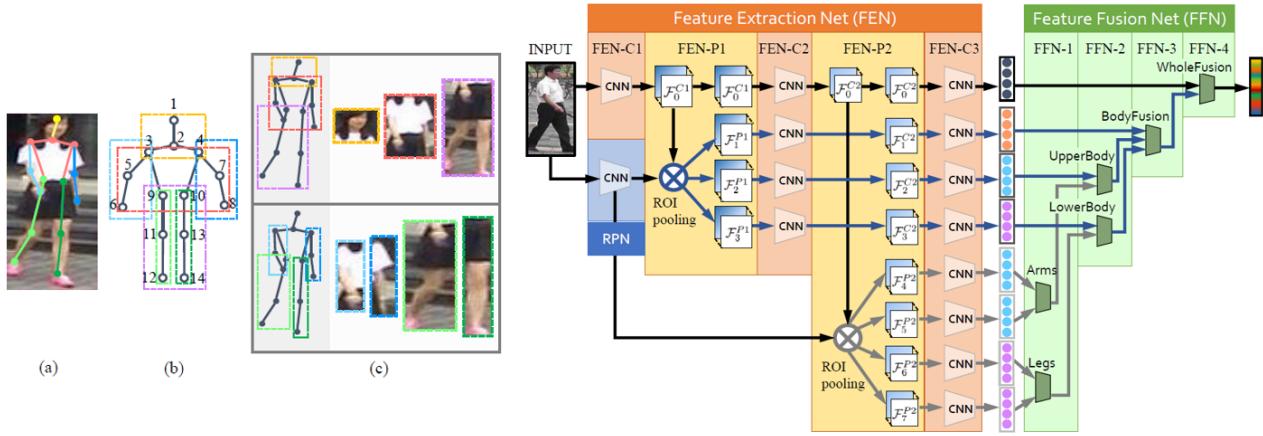


图 2-9 Spindle Net 结构示意图

论文^[24]提出了一种全局-局部对齐特征描述子(Global-Local-Alignment Descriptor, GLAD)，来解决行人姿态变化的问题。与Spindle Net类似，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。

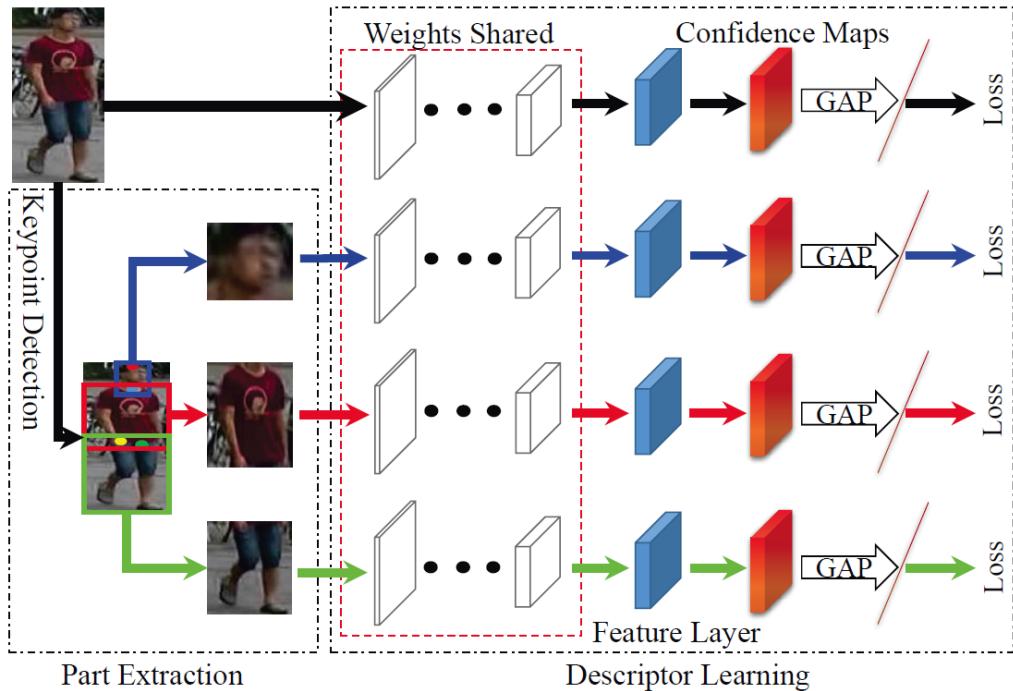


图 2-10 GLAD 结构示意图

2.1.6 基于视频序列的方法

以上介绍的方法都是基于单帧图像的方法，通常单帧图像的信息是有限的，因此有很多工作集中在利用视频序列来进行行人重识别方法的研究^[25-31]。基于视频序列的方法最主要的不同点就是这类方法不仅考虑了图像的内容信息，还考虑了帧与帧之间的运动信息等。

基于单帧图像的方法主要思想是利用CNN来提取图像的空间特征，而基于视频序列的方法主要思想是利用CNN来提取空间特征的同时利用递归循环网络(Recurrent neural networks, RNN)来提取时序特征。图2-11是非常典型的思路，网络输入为图像序列。每张图像都经过一个共享的CNN提取出图像空间内容特征，之后这些特征向量被输入到一个RNN网络去提取最终的特征。最终的特征融合了单帧图像的内容特征和帧与帧之间的运动特征。而这个特征用于代替前面单帧方法的图像特征来训练网络。

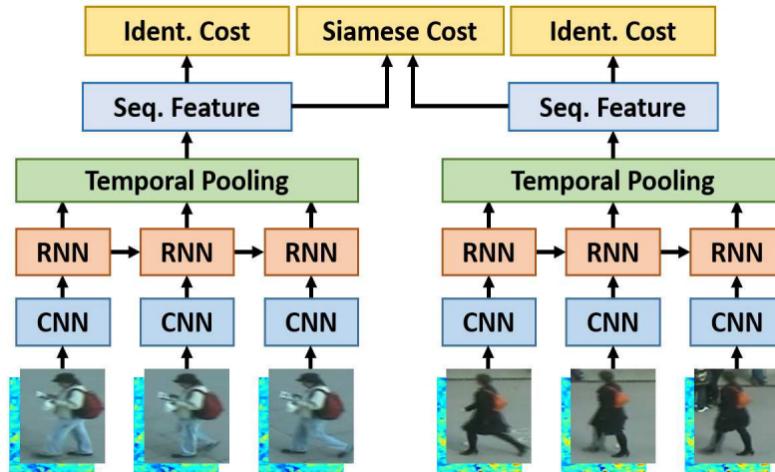


图 2-11 基于视频序列的行人重识别网络结构示意图

视频序列类的代表方法之一是累计运动背景网络(Accumulative motion context network, AMOC)^[31]。AMOC输入的包括原始的图像序列和提取的光流序列。通常提取光流信息需要用到传统的光流提取算法，但是这些算法计算耗时，并且无法与深度学习网络兼容。为了能够得到一个自动提取光流的网络，作者首先训练了一个运动信息网络(Motion network, Moti Nets)。这个运动网络输入为原始的图像序列，标签为传统方法提取的光流序列。如图2-12所示，原始的图像序列显示在第一排，提取的光流序列显示在第二排。网络有三个光流预测的输出，分别为Pred1, Pred2, Pred3，这三个输出能够预测三个不同尺度的光流图。最后网络融合了三个尺度上的光流预测输出来得到最终光流图，预测的光流序列在第三排显示。通过最小化预测光流图和提取光流图的误差，网络能够提取出较准确的运动特征。

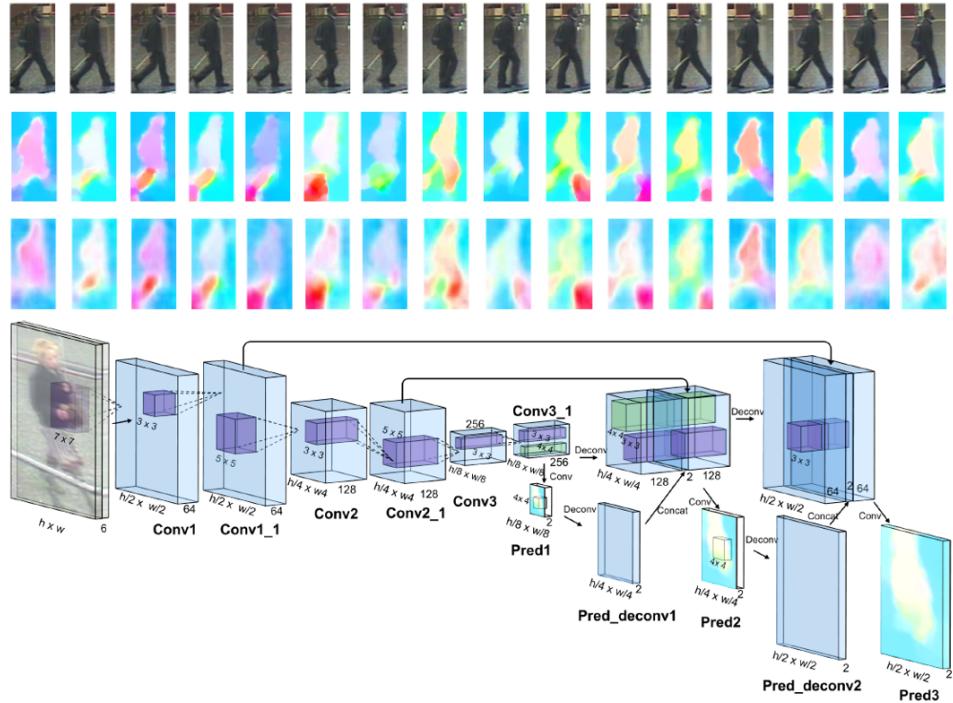


图 2-12 运动网络结构示意图

AMOC的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2-13所示。AMOC拥有空间信息网络(Spatial network, Spat Nets)和运动信息网络两个子网络。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。

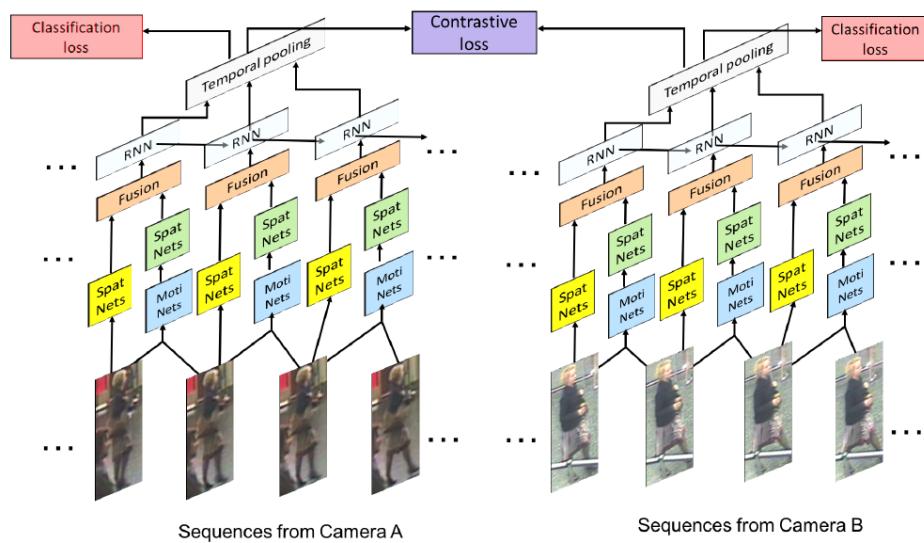


图 2-13 AMOC结构示意图

2.2 跨摄像头多目标跟踪

2.2.1 相关数据集

(1) DukeMTMC DukeMTMC^[32]是杜克大学五位博士研究生花费一年多时间标注的一个跨摄像头多目标跟踪的数据集，是目前最好最新的MTMC数据集。该数据集总共包括8个固定摄像头拍摄的85分钟视频数据，视频为60fps的1080p分辨率图像。总共人工标注了2,000,000帧图像数据，其中超过2,000名行人，比目前所有的MTMC数据集加起来还多。所有的跟踪序列加起来时长超过了30多个小时。针对单个摄像头，单帧图像包含行人数最少0人，最多54人。总共有4,159次轨迹的切换以及50个轨迹线的盲点，此外还有1,800自我遮挡发生。有两对相机(2-8,3-5)的视野有微量的重合，所以这个既可以用来研究有重合的跨摄像头跟踪，也可以用来研究无重合的跨摄像头跟踪。每个摄像头前5分钟的视频用来作为训练集或者验证集，剩下的80分钟用来作为测试集。总共有891个行人只在一个摄像头中出现过，跟踪器很容易产生FP，这对于跟踪器也是一个非常大的考验。DukeMTMC数据集的数据示例如图2-14所示。

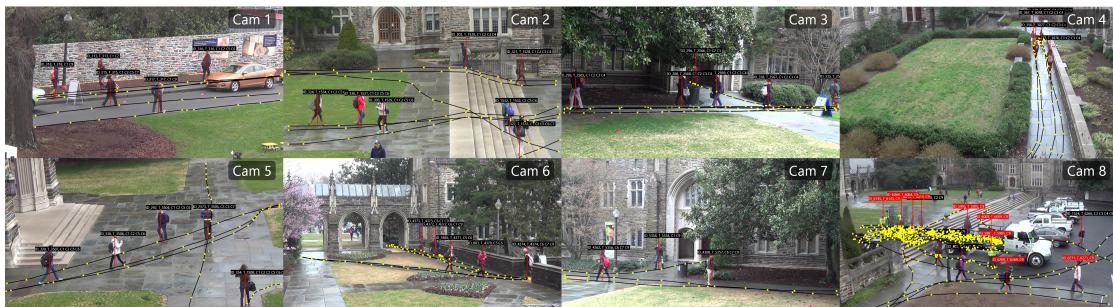


图 2-14 DukeMTMC数据集示意图

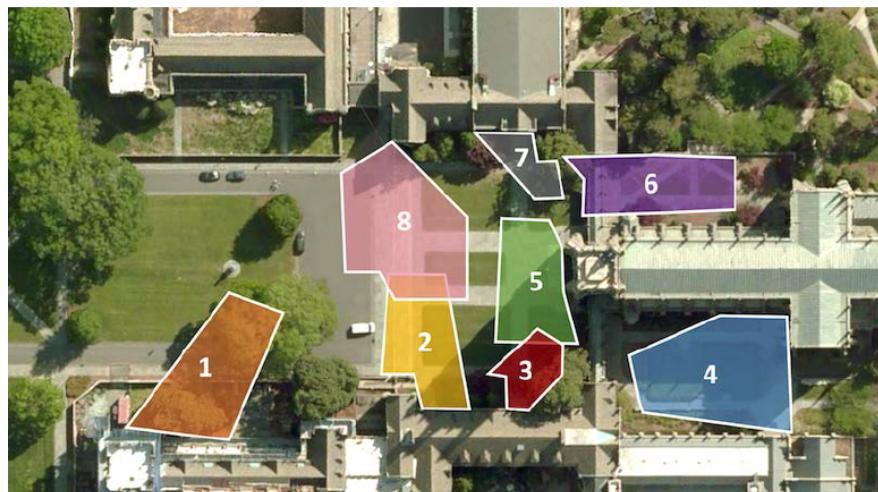


图 2-15 DukeMTMC数据集示意图

当然除了标注行人的ID和跟踪序列，考虑到时空信息关联的问题，DukeMTMC数据集还给定了8个摄像头的相机标定数据，既提供相机内参也提供相机外参数数据。这些相机标定数据使得研究人员可以得到跟踪目标在世界坐标系下的时空坐标，这个信息对于滤波相关的方法是非常有利的。利用相机标定数据重建的8个相机的视野在杜克大学校园地图上的投影分布如图2-15。

(2) MOT16

MOT16^[33]是一个单摄像头的多目标跟踪数据集，总共包括14个视频数据，其中7个训练数据，剩下7个为测试数据。这些视频来自于7个不同的场景，每个场景的视频都是随机切成完成不重合的两部分，分别作为训练和测试数据。和DukeMTMC不同的是，DukeMTMC只关注行人目标，而MOT数据集主要关注行人目标，其次还关注、汽车、自行车、摩托车等12种常见目标。由于视频的来源各自不同，所以视频的分辨率和帧率等属性也各自不同，具体属性数据如图2-16所示。可以看出，视频既有固定视角拍摄的，也有手持移动拍摄，总体来讲是一个非常多多样性、非常具有挑战性的多目标跟踪数据集。

Training sequences											
Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions	Source	
MOT16-02	30	1920x1080	600 (00:20)	49	17,833	29.7	static	medium	cloudy	new	
MOT16-04	30	1920x1080	1,050 (00:35)	80	47,557	45.3	static	high	night	new	
MOT16-05	14	640x480	837 (01:00)	124	6,818	8.1	moving	medium	sunny	[13]	
MOT16-09	30	1920x1080	525 (00:18)	25	5,257	10.0	static	low	indoor	new	
MOT16-10	30	1920x1080	654 (00:22)	54	12,318	18.8	moving	medium	night	new	
MOT16-11	30	1920x1080	900 (00:30)	67	9,174	10.2	moving	medium	indoor	new	
MOT16-13	25	1920x1080	750 (00:30)	68	11,450	15.3	moving	high	sunny	new	
Total training			5,316 (03:35)	512	110,407	20.8					

Testing sequences											
Name	FPS	Resolution	Length	Tracks	Boxes	Density	Camera	Viewpoint	Conditions	Source	
MOT16-01	30	1920x1080	450 (00:15)	23	6,395	14.2	static	medium	cloudy	new	
MOT16-03	30	1920x1080	1,500 (00:50)	148	104,556	69.7	static	high	night	new	
MOT16-06	14	640x480	1,194 (01:25)	217	11,538	9.7	moving	medium	sunny	[13]	
MOT16-07	30	1920x1080	500 (00:17)	55	16,322	32.6	moving	medium	shadow	new	
MOT16-08	30	1920x1080	625 (00:21)	63	16,737	26.8	static	medium	sunny	new	
MOT16-12	30	1920x1080	900 (00:30)	94	8,295	9.2	moving	medium	indoor	new	
MOT16-14	25	1920x1080	750 (00:30)	230	18,483	24.6	moving	high	sunny	new	
Total testing			5,919 (04:08)	830	182,326	30.8					

图 2-16 MOT16数据集各视频属性

MOT16数据集标注是十分精确了，每个目标的检测框都对齐的足够精细了，几乎没有漏框目标的任何一个像素也没有额外消耗多余的像素，也就是基本框住了目标的轮廓边界。最后MOT16数据集总共标注了215,166个检测框，平均每帧有19.15个检测框，即每帧平均有19.15个跟踪目标。其中有一大半是普通的行人，因此可以用于本课题的研究。MOT16数据集的示例如图2-17所示，总共有7个场景的14个视频序列，上面的为训练集，下面的为测试集。



图 2-17 MOT16数据集各视频属性

(3) PETS16

PETS16^[34]是一个多目标跟踪的数据集，因为和本课题研究的应用场景没有非常契合，因此简单介绍一下。该数据集总共包括14个视频数据，其中6个为训练集，8个为测试集。视频分为480p、512p、960p和1280p四种分辨率，帧率为25fps或者30fps两种规格。训练集总共标注了7,051个检测框，测试集标注8,025个检测框。该数据集主要提供道路上行走的行人和水面上高速移动的船只的跟踪序列标注，拍摄的视角基本都是低空的摄像头，比如从卡车副驾驶座拍摄的。数据示例如图2-18所示。



图 2-18 MOT16数据集各视频属性

2.2.2 准确度评估准则

跨摄像头多目标跟踪准确度的评价准则有很多，主要包括单摄像头的多目标跟踪和跨摄像头两部分，本小节将逐个介绍^[32,35,36]。

- TP：真正(True Positive, TP)是指被模型预测为正的正样本，可以称为判断为正的正确率。
- TN：真负(True Negative, TN)是指被模型预测为负的负样本，可以称为判断为负的正确率。
- FP：假正(False Positive, FP)是指被模型预测为正的负样本，可以称为误报率。
- FN：假负(FALSE Negative, FN)是指被模型预测为负的正样本，可以称为漏报率。
- Accuracy：准确度是指被分类器判定正确的比重，公示表示为：

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (2-9)$$

- Precision：精确度是指被分类器判定的正例中真正的正例样本的比重，公示表示为：

$$P = \frac{TP}{TP + FP} \quad (2-10)$$

- Recall：召回率是指被分类器正确判定的正例占总的正例的比重，公示表示为：

$$R = \frac{TP}{TP + FN} \quad (2-11)$$

- MOTA：多目标跟踪准确度(Multiple Object Tracking Accuracy, MOTA)是衡量单摄像头多目标跟踪准确度的一个指标，公示表示为：

$$MOTA = 1 - \frac{FN + FP + \Phi}{T} \quad (2-12)$$

其中 FN 是指所有帧的假负数之和，即假设 fn_t 为第 t 帧的假负数，则 $FN = \sum_t fn_t$ ，同理 $FP = \sum_t fp_t$ 。 T 是指所有帧真正目标数的总和，即假设第 t 帧有 g_t 个目标，则 $T = \sum_t g_t$ 。 Φ 是指所有帧目标发生跳变数(Fragmentation)， ϕ_t 为第 t 帧的目标跳变数，则 $\Phi = \sum_t \phi_t$ 。换而言之，这三项依次表示缺失率、误判率和误配率。

- **MOTP:** 多目标跟踪精确度(Multiple Object Tracking Precision, MOTP)是衡量单摄像头多目标跟踪位置误差的一个指标，公式表示为：

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad (2-13)$$

其中 c_t 表示第 t 帧的匹配个数，对每对匹配计算匹配误差 d_t^i ，表示第 t 帧下目标 O_i 与其配对假设位置之间的距离。

- **MT:** 多数跟踪数(Mostly tracked)是指跟踪部分大于80%的跟踪轨迹数，数值越大越好。
- **ML:** 多数丢失数(Mostly lost)是指丢失部分大于80%的跟踪轨迹数，数值越小越好。
- **IDS:** ID转变数(ID switches)是指跟踪轨迹中行人ID瞬间转换的次数，通常能反应跟踪的稳定性，数值越小越好。
- **IDP:** 识别精确度(Identification Precision)是指每个行人框中行人ID识别的精确度。公式为：

$$IDP = \frac{IDTP}{IDTP + IDFP} \quad (2-14)$$

其中 $IDTP$ 和 $IDFP$ 分别是真正ID数和假正ID数。

- **IDR:** 识别召回率(Identification Recall)是指每个行人框中行人ID识别的召回率。公式为：

$$IDP = \frac{IDTP}{IDTP + IDFN} \quad (2-15)$$

其中 $IDFN$ 是假负ID数。

- **IDF1:** 识别F值(Identification F-Score)是指每个行人框中行人ID识别的F值。公式为：

$$IDP = \frac{2IDTP}{2IDTP + IDFP + IDFN} \quad (2-16)$$

- **MCTA:** 跨摄像头跟踪准确度(Multi-camera Tracking Accuracy)是衡量多个摄像头下跟踪的准确度，是目前少有的专门用来衡量多摄像头跟踪性能的评价指标。公式为：

$$MCTA = \underbrace{\frac{2PR}{P+R}}_{F1} \underbrace{(1 - \frac{M^w}{T^w})}_{\text{within camera}} \underbrace{(1 - \frac{M^h}{T^h})}_{\text{handover}} \quad (2-17)$$

其中 P, R 为前文介绍的精准度和召回率， M^w 是单相机内行人ID的错误匹配数， T^w 是单相机内（标注）正确检测数， M^h 是跨相机行人ID的错误匹配数， T^h 是指跨相机（标注的）正确检测数（即某个目标从某个相机中消失而下次再出现在另外一个相机的情况）。MCTA的范围是 $[0, 1]$ 。

2.2.3 表观模型

表观模型(Appearance model)既包括目标的视觉特征表达，也包括目标间相似性、相异性的度量。视觉表达肯定是基于图像特征的，在深度学习出现之前，学者们通常手动提取一些传统特征，由于在本课题研究中主要使用行人重识别的深度学习特征，所以在这里只是简单提及一下这些传统特征。传统的图像特征包括：

- Point feature, 比如Harris角点、SIFT角点、SURF角点等等
- Color/intensity features, 比如最简单的模板、颜色直方图等
- Optical flow, 光流特征蕴含了时域信息
- Gradient/pixel-comparison features, 基于梯度的特征，典型的如HOG特征
- Region covariance matrix features, 该特征对于光照和尺度变换相对鲁棒
- Depth, 即深度信息，对于视频这种3D数据作用还是蛮大的

这些图像特征在传统的多目标跟踪中有非常广泛的应用，但是随着深度学习和行人重识别的发展，ReID特征逐渐成为一种优秀的表观模型^[37,38]。

论文^[37]结合ReID特征和其他一些数据关联(Data association, DA)，给定两个检测框的图片 I_1 和 I_2 ，它们的距离（和相似度呈反比）公示表示如下：

$$d(I_1, I_2) = \frac{d_{pos}(I_1, I_2)}{N_{pos}} \frac{d_{app}(I_1, I_2)}{N_{app}} \quad (2-18)$$

其中 d_{app} 是两张图片的ReID特征的距离，比如最常用的欧式距离。 N_{pos} 和 N_{app} 是归一化的参数，使得 d_{pos} 和 d_{app} 能够处于相同的数量级。 d_{pos} 是一些传统数据关联方法，这里可以被任何相关的方法所取代。

论文^[38]则更加直接，通过训练一个孪生网络来判断两个detection框是否是需要匹配。如图2-19所示，检测器检测出若干个detections，之后通过一个已经训练好的行人重识别孪生网络，把相同的小段轨迹(Tracklet)关联起来，之后通过一个线性规划(Linear Programming)方法得到最终的跟踪轨迹(Trajectory)。这种方法比较简单，只需要一个检测器和ReID模型便可以实现MTMC跟踪问题，是业界非常常用的一种方法。但是这种方法的缺点也很明显，非常依赖检测器和ReID模型的性能。在本课题中我们不关注检测器的研究，即是在有一个很好的检测器的前提下开展MTMC工作，那么这种方法最后的性能完全由ReID模型的影响。

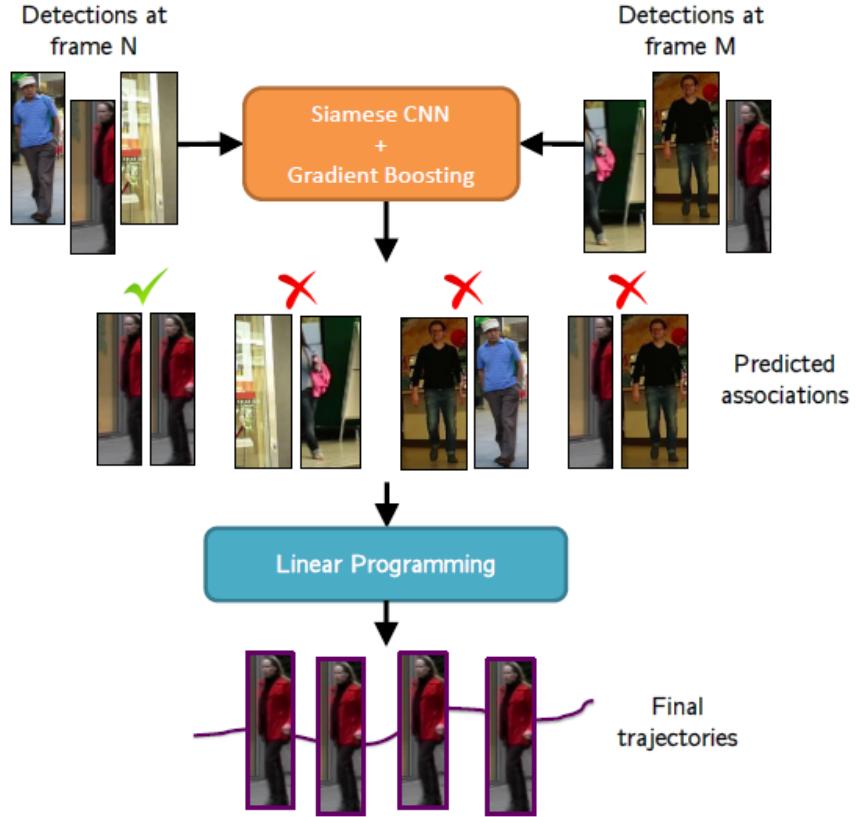


图 2-19 基于行人重识别的孪生网络实现多目标跟踪^[38]

2.2.4 基于相关滤波的多目标跟踪

相关滤波是一类非常常见的目标跟踪方法，贝叶斯滤波(Bayesian filter)、卡尔曼滤波(Kalman filter)、粒子滤波(Particle filter)等都在目标跟踪问题上有所应用。然而在跨摄像头多目标跟踪问题上，本来相关工作就不多，和行人重识别结合起来的工作就更加少之又少。这里介绍一篇结合ReID特征和贝叶斯滤波的MTMC论文^[37]。

给定一个ReID模型 f_θ ，图片 I_p 可以得到一个嵌入特征 $e_p = f_\theta(I_p)$ 。对于完整的某一帧图像 I ，可以得到一个 $D_I(e_p) = (||e_{i,j} - e_p||)_{i,j}$ ， $e_{i,j}$ 代表图像 I 该位置中心的图像切片与目标图片的ReID特征之间的距离，距离越小代表越相似。最后可以得到一个嵌入特征距离图(Embedding distance map) D_I 。之后这个 D_I 可以转化为一个贝叶斯滤波器的观测，观测模型表示为：

$$P(z_t | X_t, z_{1:t-1}) = \text{softmax}(D_I(f(X_t, z_{1:t-1}))) \quad (2-19)$$

其中 $e_p = f(X_t, z_{1:t-1})$ 可以通过很多方式更新，在论文中就简单用第一次出现的代替并不进行更新，即 $e_p = f_\theta(z_1)$ 。当然为了适应这个新的观测器模型，我们需要对传统的最优贝

叶斯滤波进行一个重构，重构利用概率的贝叶斯规则，最终表达如下：

$$P(X_t | z_t, z_{1:t-1}) \propto \overbrace{P(z_t | X_t, z_{1:t-1})}^{\text{new measurement}} \overbrace{P(X_t | z_{1:t-1})}^{\text{belief propagation}} \quad (2-20)$$

公示被分解为两个部分，前一部分是当前最新的观测结果，后一部分可以用前一时刻的状态进行状态估计。当我们有了观测模型之后，下一步就是用贝叶斯滤波器进行状态估计。公示(2-20)的后面一项可以利用全概率模型和马尔科夫规则进行进一步的分解：

$$P(X_t | z_{1:t-1}) = \int \overbrace{P(X_t | x_{t-1})}^{\text{dynamics model}} P(x_{t-1} | z_{1:t-1}) dx_{t-1} \quad (2-21)$$

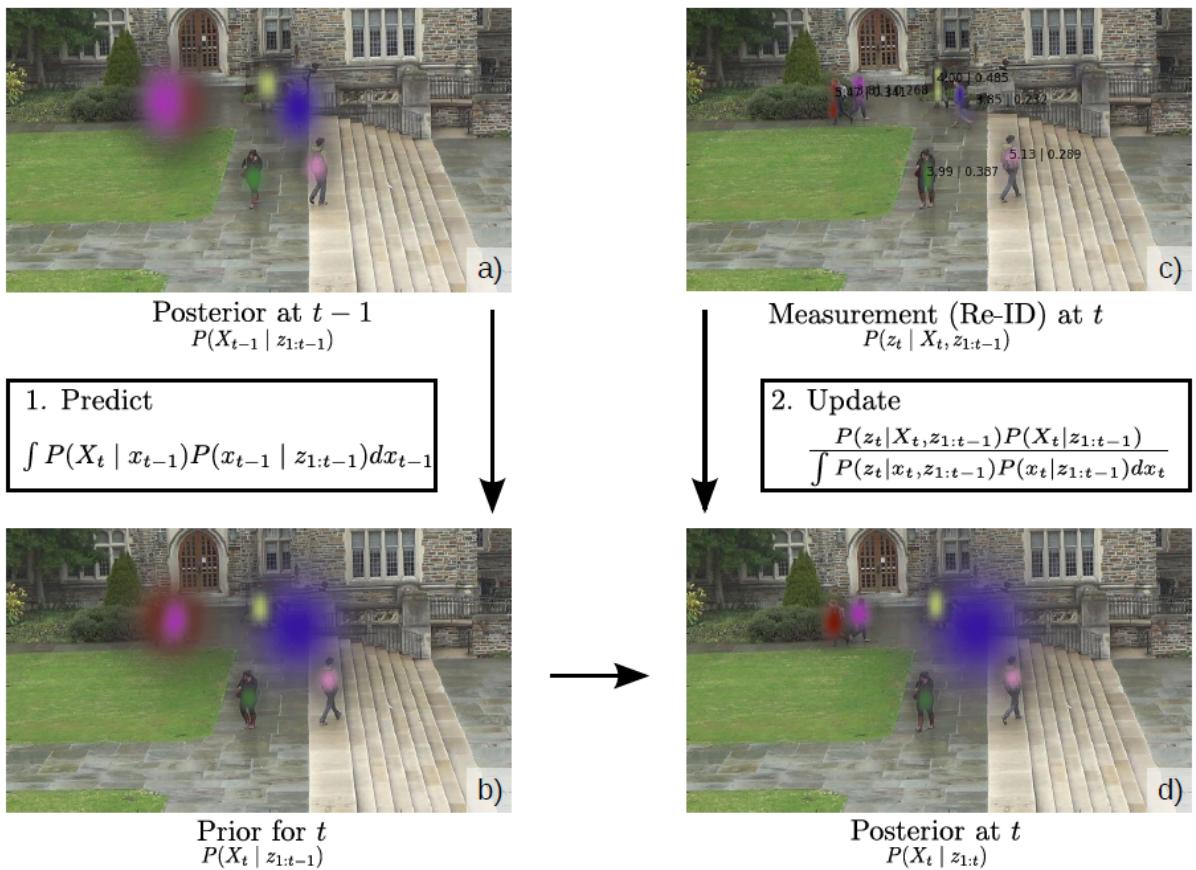


图 2-20 基于行人重识别和贝叶斯滤波的多目标跟踪^[37]

2.2.5 基于代价函数的多目标跟踪

2.2.6 跨摄像头匹配

3 研究内容与技术路线

3.1 研究内容

该课题主要包含两个子课题，一个是行人重识别，一个是跨摄像头多目标跟踪。

对于行人重识别问题，主要研究思路为从单帧图像向跟踪序列发展，从高质量图像向实际干扰很多的图像发展。对于单帧图像的行人重识别方法，核心目标在于在目前已有的公开行人ReID数据集上取得State-of-the-arts，甚至超越人类准确度的结果。为了达到这个目标，需要研究以下内容：

1. 得到更好的图像特征表达。无论是基于表征学习还是度量学习的方法，其本质目标就在于如何让网络学习到一个更加鲁邦的特征，来提高行人重识别的准确度，所以设计一个好的学习模型是最核心的研究内容。
2. 考虑全局特征和局部特征。目前的大量研究工作表明，只考虑全局特征对于行人重识别来说并不足以取得令人惊喜的结果。因为全局特征可以解决大部分的情况，但是在一些外观相似的情况下，只能依靠局部的细节来区分。因此联合全局特征和局部特征是进一步提高行人重识别准确度的重要手段。
3. 多模型的联合与集成。单模型所能取得的性能终究是有限，多个模型集成是提高准确度的一种通用手段，所以在单模型遇到瓶颈的情况下，多模型的研究可以是另外一条继续提升的方法。
4. 室内行人重识别。目前所有的ReID数据集都是远距离拍摄的视频数据，这种数据的特点是图片中行人结构比较完整，基本都是全身照。但是在室内场景下由于摄像头视角的原因，存在着很多半身以及遮挡的情况，如何做好半身以及遮挡图片与全身图片的匹配也是一个值得研究的内容。
5. 基于序列的行人重识别。基于跟踪序列的ReID模型对于MTMC任务是有直接的贡献，因为MTMC的本质核心就是把几个跟踪序列给匹配关联起来。

对于跨摄像头多目标跟踪问题，主要研究思路为以**ReID**作为最主要的表象特征，辅之以相关滤波、时空关联等数据关联技术；借助已有的单摄像头单目标的跟踪技术，加入**ReID**特征改进为多目标跟踪技术。最终能够得到一个可以实用的跨摄像头多目标跟踪系统，并在DukeMTMC等数据集上取得State-of-the-arts的结果。为了达到这个目标，需要研究一下内容：

1. 多目标的匹配关联。单目标跟踪只需要关注当前跟踪的目标，不会出现大范围的目标突变。然而多目标需要关注所有的跟踪目标，首先最直接的一点就是如何把不同时刻的多个目标关联起来。
2. 多目标错匹配的找回。单目标跟踪一般不会出现目标匹配混乱的问题，最多也就是当前目标在一些情况下丢失了。但是多目标跟踪可能由于多个目标的位置重叠等原因，导致目标丢失甚至匹配互换，如何研究一个合理的机制在目标错匹配的情况下再重新联结回来是一个值得研究的内容。
3. 时空数据关联。只用**ReID**技术当然足以完成多目标跟踪任务，但是并不能保证**ReID**技术已经足以保证多目标跟踪的实际需求。在一些**ReID**失效的情况下，利用相关滤波以及时空关联技术，是进一步提升跟踪器性能的思路。
4. 跨摄像头的目标匹配。跨摄像头的目标匹配是**MTMC**和其他跟踪任务最不一样的地方，这是在目标一定丢失并且可能不再出现的先验条件下重新找回目标的任务，因此是一个非常有挑战的任务，也是**MTMC**任务必须解决的一个问题。
5. 室内跨摄像头多目标跟踪数据集。目前**MTMC**相关的数据集很少，仅存的几个数据集也是校园采集的室外场景。随着无人超市、无人商场等应用需求的提升，一个室内**MTMC**的数据集能够非常好的促进相关技术的研究与落地。

3.2 技术路线

3.2.1 基于深度学习的行人重识别技术

深度学习已经取代传统的视觉方法成为行人重识别领域的必用方法。训练一个行人重识别的深度学习模型包括数据处理、网络模型设计、损失函数设计、模型训练、特征提取以及图像检索等环节。数据处理包括数据预处理和数据增广，数据预处理是指把图片的尺寸和内容等归一化到一个统一的尺寸，数据增广是为了防止过拟合而把一张图像增广到多

张图像，典型的技术包括翻转、模糊、平移、裁剪以及光照变化等等。之后是为了满足特定任务设计一个提取特征的卷积网络，可以采用一些预训练好的卷积网络，例如ResNet、GoogleNet等，也可以根据需求自定义网络，比如为了提取不同尺寸规模的特征可以融合不同层次的feature maps。设计合适的损失函数有利于网络收敛到一个更合理的最优点，表征学习和度量学习的思路不同使得其损失函数的设计思路不同，但最终的目的都是为了训练出一个泛化能力很好的卷积网络。前面的工作完成了就可以训练网络，深度学习都是基于梯度下降算法训练模型，因此需要设计好合适的学习率和优化器，使得网络能够更快更好的收敛。训练好了一个网络模型之后，我们需要利用这个模型来完成行人重识别任务，总共包括两个环节，而特征提取是第一个环节，利用已经训练的网络模型，对于每一张图像我们都能提取至少一个特征，这个特征包含了图像的全局或者局部内容特征。特征提取之后是检索环节，对于每一张图像模型都能提取一个特征，之后利用这些特征计算两幅图像之间的相似性，可以用最简单的欧式距离来计算，也可以用一些更加鲁邦的检索方法来实现相似度排序。

3.2.2 基于行人重识别特征的跨摄像头多目标跟踪技术

4 现有成果与研究计划

4.1 现有成果

本章节介绍了现在已经取得的一些科研成果，并且根据已有的成果初步制定了一下未来的研究计划。

4.1.1 边界样本挖掘损失

边界样本挖掘损失(Margin sample mining loss, MSML)是一种引入难样本采样思想的度量学习方法。度量学习的目标是学习一个函数 $g(x) : \mathbb{R}^F \rightarrow \mathbb{R}^D$ ，使得 \mathbb{R}^F 空间上语义相似度反映在 \mathbb{R}^D 空间的距离上。通常我们需要定义一个距离度量函数 $D(x, y) : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ 来表示嵌入空间(Embedding space)的距离，而这个距离也用来重识别行人图片。

在国内外研究现状里面介绍的三元组损失、四元组损失和TriHard损失都是典型度量学习方法。给定一个三元组 $\{a, p, n\}$ ，三元组损失表示为：

$$L_t = (d_{a,p} - d_{a,n} + \alpha)_+ \quad (4-1)$$

三元组损失只考虑了正负样本对之间的相对距离。为了引入正负样本对之间的绝对距离，四元组损失加入一张负样本组成了四元组 $\{a, p, n_1, n_2\}$ ，而四元组损失也定义为：

$$L_q = (d_{a,p} - d_{a,n_1} + \alpha)_+ + (d_{a,p} - d_{n_1,n_2} + \beta)_+ \quad (4-2)$$

假如我们忽视参数 α 和 β 的影响，我们可以用一种更加通用的形式表示四元组损失：

$$L_{q'} = (d_{a,p} - d_{m,n} + \alpha)_+ \quad (4-3)$$

其中 m 和 n 是一对负样本对， m 和 a 既可以是一对正样本对也可以是一对负样本对。但是直接使用(4-3)并不能取得很好的结果，因为随着数据量的上升，可能四元组组合数量急剧上升。绝大部分样本对都是比较简单的，这限制了模型的性能。为了解决这个问题，我们采用了TriHard损失^[19]使用的难样本采样思想。TriHard损失是在一个batch里面计算三元组

损失。对于batch中的每一张图片 a , 我们可以挑选一个最难的正样本和一个最难的负样本和 a 组成一个三元组。我们定义和 a 为相同ID的图片集为 A , 剩下不同ID的图片图片集为 B , 则TriHard损失表示为:

$$L_{th} = \frac{1}{P \times K} \sum_{a \in batch} (\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha)_+ \quad (4-4)$$

而TriHard损失同样只考虑了正负样本对之间的相对距离, 而没有考虑它们之间的绝对距离。于是我们把这种难样本采样的思想引入到(4-3), 可以得到:

$$L_{msml} = (\max_{a,p} d_{a,p} - \min_{m,n} d_{m,n} + \alpha)_+ \quad (4-5)$$

其中 a, p, m, n 均是batch中的图片, a, p 是batch中最不像的正样本对, m, n 是batch 中最像的负样本对, a, m 皆可以是正样本对也可以是负样本对。概括而言TriHard损失是针对batch中的每一张图片都挑选了一个三元组, 而MSML损失只挑选出最难的一个正样本对和最难的一个负样本对计算损失。所以MSML是比TriHard更难的一种难样本采样, 此外 $\max_{a,p} d_{a,p}$ 可以看作是正样本对距离的上界, $\min_{m,n} d_{m,n}$ 可以看作是负样本对的下界。MSML是为了把正负样本对的边界给推开, 因此命名为边界样本挖掘损失。MSML只用了两对样本对计算损失, 看上去浪费了很多训练数据。但是这两对样本对是根据整个batch的结果挑选出来了, 所以batch中的其他图片也间接影响了最终的损失。并且随着训练周期的增加, 几乎所有的数据都会参与损失的计算。总的概括, MSML是同时兼顾相对距离和绝对距离并引入了难样本采样思想的度量学习方法。

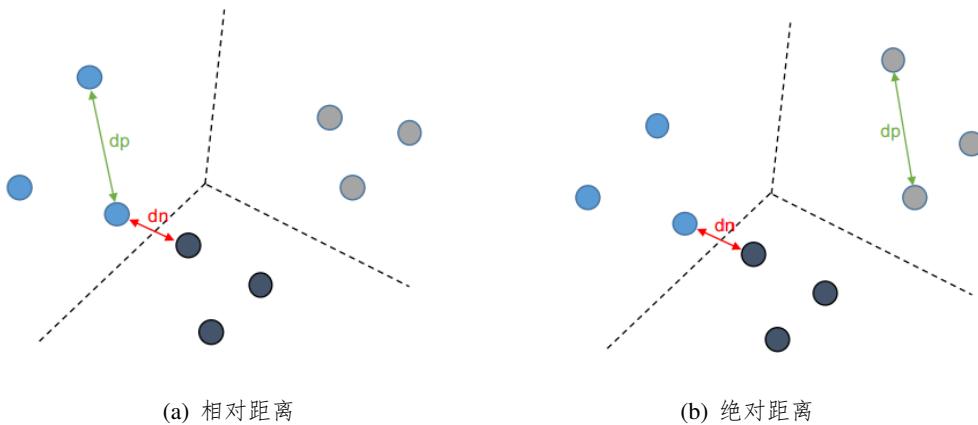


图 4-1 MSML的两种情况

为了对比MSML和传统度量学习方法的结果, 我们在Market1501、MARS、CUHK-SYSU和CUHK03公开数据集上进行了对比实验。为了减少实验数量, 我们用所有的数据训练一个模型, 并在几个数据集上分别进行测试。其中CUHK03只测评了CMC, 而剩下的数

表 4-1 MSML和其他度量学习方法的结果对比

Base model	Methods	Market1501			MARS			CUHK-SYSU			CUHK03		
		mAP	r = 1	r = 5	mAP	r = 1	r = 5	mAP	r = 1	r = 5	r=1	r=5	r = 10
ShuffleNet	Cls	38.4	64.7	83.1	43.6	61.9	78.3	76.7	80.0	90.7	54.3	74.3	80.0
	Tri	60.3	79.6	93.1	59.2	74.0	87.0	88.6	90.2	96.6	78.7	94.8	97.9
	Quad	57.9	77.9	91.9	58.1	73.3	86.8	87.2	89.2	96.5	78.6	94.5	97.3
	TriHard	65.2	82.3	93.6	69.7	81.0	91.7	84.6	86.5	95.1	81.3	95.4	97.8
	MSML	69.7	85.3	94.4	72.0	82.6	93.5	87.9	90.0	96.3	85.7	97.2	98.8
Resnet50	Cls	41.3	65.8	83.5	43.3	59.3	75.2	70.7	75.0	88.1	51.2	72.6	81.8
	Tri	54.8	75.9	89.6	62.1	76.1	89.6	82.6	85.1	94.1	73.0	92.0	96.0
	Quad	61.1	80.0	91.8	62.1	74.9	88.9	85.6	87.8	95.7	79.1	95.3	97.9
	TriHard	68.0	83.8	93.1	71.3	82.5	92.1	82.4	85.1	94.7	79.5	95.0	98.0
	MSML	69.6	85.2	93.7	72.0	83.0	92.6	87.2	89.3	96.4	84.0	96.7	98.2
Inception-v2	Cls	40.7	66.3	84.1	45.0	62.6	77.9	74.2	78.2	89.7	50.5	68.8	77.4
	Tri	57.9	78.3	91.8	55.5	70.7	85.2	87.7	89.7	96.6	76.9	93.7	97.2
	Quad	66.2	83.9	93.6	65.3	77.8	89.9	88.3	90.2	96.6	81.9	96.1	98.3
	TriHard	73.2	86.8	95.4	74.3	84.1	93.5	83.5	86.1	95.2	85.5	97.2	98.7
	MSML	73.4	87.7	95.2	74.6	84.2	95.1	88.4	90.4	96.8	86.3	97.5	98.7
Resnet50-X	Cls	46.5	70.8	87.0	48.0	63.8	80.2	74.2	78.2	89.7	57.2	77.7	85.6
	Tri	69.2	86.2	94.7	68.2	79.5	91.7	89.6	91.4	97.0	82.0	96.3	98.4
	Quad	64.8	83.3	93.8	63.6	77.7	89.4	87.3	89.6	96.2	80.7	94.9	97.9
	TriHard	71.6	86.9	94.7	69.9	82.5	92.4	86.4	88.8	96.3	82.8	96.1	98.1
	MSML	76.7	88.9	95.6	72.0	83.4	93.3	89.6	90.9	97.4	87.5	97.7	98.9

据集同时测评了CMC和mAP。为了证明算法有效性，我们利用了不同的base model多次进行实验，包括Resnet50^[39]、ShuffleNet^[40]、Inception-v2^[41]、Resnet50-Xception等主流网络。Resnet50-X是指用Xception单元^[42]替代了Resnet50中所有的 3×3 卷积层，而一个Xception单元包括一个 3×3 的深度分离卷积层(Depthwise separable/Channel-wise convolutional layer)和一个 1×1 的卷积层。实验结果如表4-1所示，我们结合了度量学习损失和分类损失。Cls代表分类损失，Tri代表三元组损失，Quad代表四元组损失，TriHard代表TriHard损失。黑色粗体标注了最好结果，可以看出绝大多数的黑色粗体都出现在MSML的实验中。当然从实验结果也可以看出，不同的度量学习方法可能适用于不同的数据集，例如三元组损失在CUHK-SYSU上表现很好。具体实验结果可在表中查阅。

为了进一步分析MSML考虑正负样本对绝对距离带来的性能优势，我们挑选几组正负样本对并计算了它们之间的距离。最后的结果显示在图4-2中，其中蓝色的框表示正样本对，红色的框表示负样本对。框下面的黑色数字表示两幅图片在特征空间中欧式距离。如图4.2(a)，TriHard损失由于不考虑正负样本对之间的绝对距离，所以理论上存在正样本对

的距离大于负样本对的情况，实际的结果也验证了这一点。而MSML考虑了绝对距离，并且MSML的目的是为了推开正样本对和负样本对的边界，如图4.2(b)所示，正样本对和负样本对在特征空间拥有非常清晰的分界线。而MSML这种分界面清晰的特性对于一些应用是非常重要的，例如跨摄像头的多目标跟踪问题等。

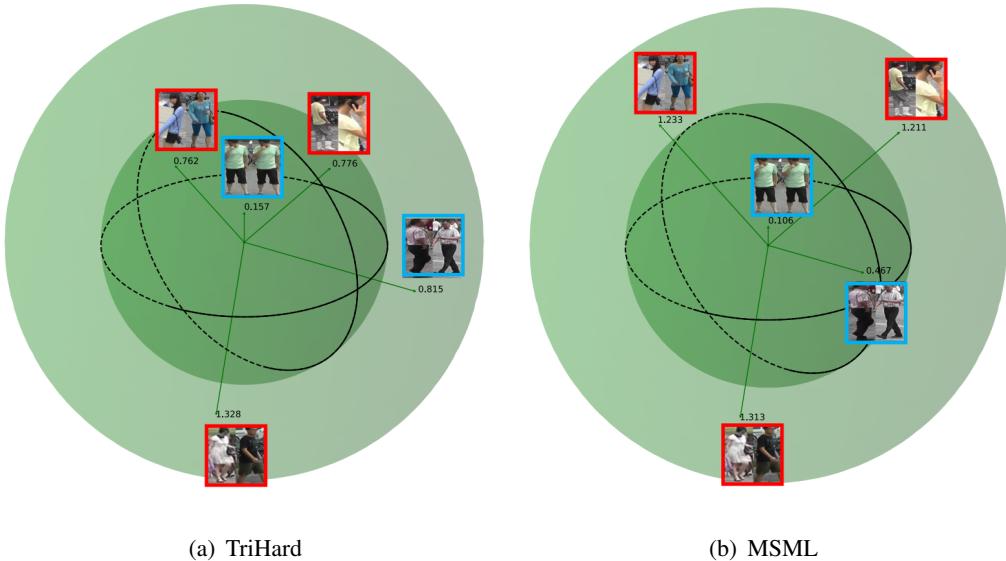


图 4-2 TriHard损失和MSML正负样本对距离分布示意图

4.1.2 最短路径距离

最短路径距离(Shortest path distance, SP distance)是一种适用于自动对齐模型(Auto-alignment model, AAM)的距离度量方法。

在大多数应用中，这个距离度量选择最简单的欧式距离，但是欧式距离不考虑局部空间信息。如果只考虑全局信息，姿态、角度等变化会严重影响识别准确度。为了弥补欧式距离忽略局部空间信息的缺点，很多方法选择手动提取局部特征。基于局部特征的方法已经在前面章节介绍过了，所以这里不再赘述。在笔者所知的范围内，目前所有的局部特征对齐方法都需要一个额外的骨架关键点或者姿态估计的模型。而训练一个可以达到实用程度的模型需要收集足够多的训练数据，这个代价是非常大的。

为了解决以上问题，本文提出基于SP距离的自动对齐模型(Auto-Alignment Model, AAM)在不需要额外信息的情况下自动对齐局部特征。对于每一张图片我们使用CNN来提取最后一层的特征图(Feature map)，传统方法通常使用全局平均池化(Global average pooling, GAP)来得到全局特征。使用GAP就使得最后得到的特征失去了全部的空间信息，为此我们选择水平平均池化(Horizontal average pooling, HAP)。HAP先将特征图在垂直方向

分成相等的若干份，然后在每一行上取平均值来作为当前切片的特征值。经过HAP，我们可以得到一个 $H \times C$ 的特征，其中 H 表示特征图被分割成 H 均分， C 表示特征图的通道数。之所以选择垂直方向的切分是因为这更符合我们对人体认知的感官直觉。据我们所知，人体结构是非常固定的，从上到下分别是头、胸脯、腰、大腿、小腿等等。在监控图片中，通常不存在水平翻转的图片。也就是说，如果按照从上往下看的顺序头部是最先出现的部位。在真实的场景下，因为遮挡和相机视角的变化，其中一张图片的头部也许不能对齐另外一张图片的头部。此外，模型提取的行人边框也有可能不准确。所以，我们提出利用最短路径理论来自动对齐两张图片中相同的行人部位。

给定两张图片的局部特征 $F = \{f_1, \dots, f_H\}$ and $G = \{g_1, \dots, g_H\}$ ，首先我们计算每个元素欧式距离，然后用以下公式归一化到 $[0, 1]$:

$$d_{i,j} = \frac{e^{\|f_i - g_j\|_2} - 1}{e^{\|f_i - g_j\|_2} + 1} \quad i, j \in [1, 2, 3, \dots, H] \quad (4-6)$$

其中 $d_{i,j}$ 表示图片1第*i*部分和图片2第*j*部分的归一化后的距离，之后我们便可以得到一个 $H \times H$ 的距离矩阵 D 。考虑到人体结构是一个连续不变结构，即各部分的相关顺序是不变的，所以我们从上到下比较各部分的相似度。两张图片的SP距离可以定义为矩阵从 $(1, 1)$ 到 (H, H) 的最短路径，这个可以通过动态规划来求解：

$$S_{i,j} = \begin{cases} d_{i,j} & i = 1, j = 1 \\ S_{i-1,j} + d_{i,j} & i \neq 1, j = 1 \\ S_{i,j-1} + d_{i,j} & i = 1, j \neq 1 \\ \min(S_{i-1,j}, S_{i,j-1}) + d_{i,j} & i \neq 1, j \neq 1 \end{cases} \quad (4-7)$$

其中 $S_{i,j}$ 表示当前从 $(1, 1)$ 到 (i, j) 的最短路径，而 $S_{H,H}$ 是两张图片的最终最短路径距离。由于并不存在复杂的计算子，所以该方法的反向传播可以通过大部分框架都自带的自动求导实现。

为了更加直观的解释最短路径距离，我们给出了一个示例。如图4-3所示，其中黑色实线连接了左右两幅图像中最相似的两部分。在本论文中我们采用残差网络(Resnet)^[39]作为基础模型(Base model)，所以最后的feature map的尺寸为 7×7 ，因此图片被分割成了7部分。图片中的数字表示区域ID，右边的黑色箭头是最后求解出来的最短路径。首先，我们连接 f_1 和 g_1 ，因为它们是路径的开端。之后我们比较 $d_{1,2}$ 和 $d_{2,1}$ 发现 $d_{2,1} > d_{1,2}$ ，所以我们认为图片A的第1区域和图片B的第2区域更加相似。所以我们连接图片A的第1区域和图片B的第2区域，在右边的图片反映为从 $(1, 1)$ 走到 $(1, 2)$ 。之后我们比较 $d_{2,2}$ 和 $d_{1,3}$ ，因为 $d_{2,2} > d_{1,3}$ ，

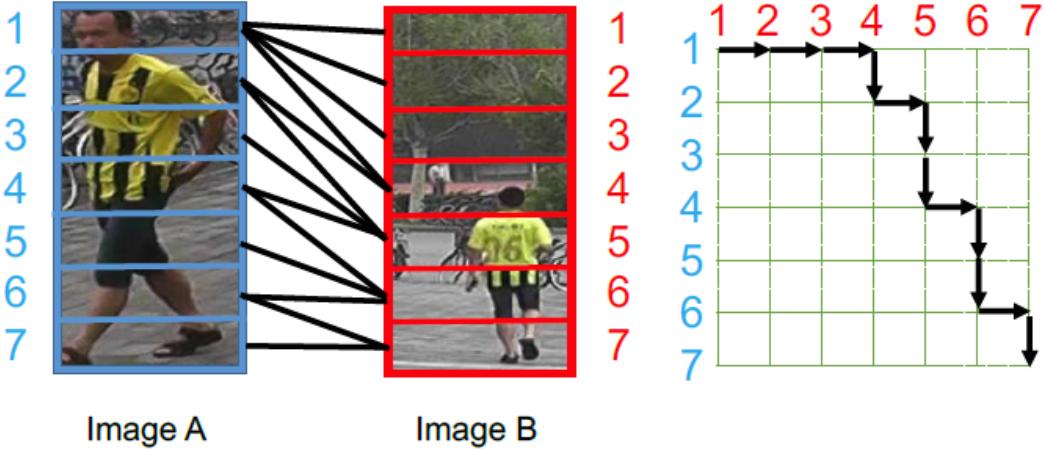


图 4-3 基于SP距离的AAM模型算法示例

所以连接图片A的第1区域和图片B的第3区域，在右边的图片反映为从(1, 2)走到(1, 3)。按照这种规律寻找下去，最终的最短路径如图4-3右半所示。

我们的SP距离可以应用于各种度量学习损失，在本论文中我们使用前文介绍的TriHard损失^[19]。对于每一个训练batch包含 $N = P \times K$ 张图片，即P个ID的行人，每个行人随机挑选K张不同的图片。给定 I_1 和 I_2 两张图片，我们定义 $S_{local}(I_1, I_2)$ 为它们的SP距离，定义 $D_{global}(I_1, I_2)$ 为它们的全局L2距离。如图4-4，我们结合了L2 TriHard损失和SP TriHard损失，最终损失函数 L_t 为：

$$\begin{aligned}
 L_t &= \frac{1}{N} \sum_{a \in batch} \left[\left(S_{local}(a, p) - S_{local}(a, n) + \alpha \right)_+ + \left(D_{global}(a, p) - D_{global}(a, n) + \beta \right)_+ \right] \\
 p &= \arg \max_{p \in A} D_{global}(a, p) \\
 n &= \arg \min_{n \in B} D_{global}(a, n)
 \end{aligned} \tag{4-8}$$

值得说明的是，我们选择用L2距离来进行难样本采样(Hard sample mining, HSM)，HSM的结果用于同时计算L2 TriHard损失和SP TriHard损失。最终采用的网络结构示意图如图4-4所示，网络每个训练batch会输入N张图片。每张图片经过CNN的几个卷积层得到feature maps，feature maps之后分为两个分支。其中一个分支和传统的方法一样经过GAP得到全局L2特征，并计算得到一个 $N \times N$ 的L2距离矩阵，该矩阵用于进行HSM。另外一个分支用HAP得到局部特征并采用一个卷积层CONV降低特征的通道数，最终得到的feature maps用于计算SP距离矩阵。最终的损失函数融合了两个分支的损失，共享HSM结果是为了保证两个分支计算损失时用的是相同的采样图片。之所以选择L2距离来进行HSM主要基于两点考虑。一个是SP距离计算比L2距离计算耗时，在实现上可

以只计算HSM样本的SP距离（虽然本文实现上并非如此）。另一个是我们发现利用SP进行HSM并没有显著地带来性能提升。

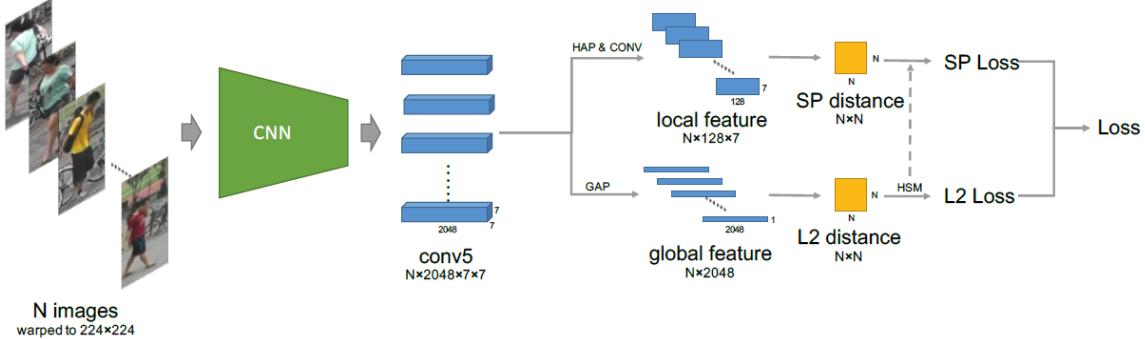


图 4-4 基于TriHard损失的AAM网络结构示意图

在本论文中，我们采用Resnet50和Resnet50-Xception(Resnet50-X)分别进行了对别实验。对于同一个base model，我们采用传统的L2距离和本文提出的结合L2和SP距离的AAM模型进行了对比。我们采用了Market1501、MARS、CUHK-SYSU和CUHK03四个公开数据集来训练一个模型，并分别在这四个数据集上进行了测试。除了CUHK03数据集只计算了CMC以外，另外三个数据集都评测了CMC和mAP。结果表明，AAM模型能够显著地提高行人重识别的准确度，Resnet50和Resnet50-X都得到了类似的结果，具体数值可以看表4-2。

表 4-2 SP距离和L2距离度量结果对比

Base model	Methods	Market1501			MARS			CUHK-SYSU			CUHK03		
		mAP	r = 1	r = 5	mAP	r = 1	r = 5	mAP	r = 1	r = 5	r=1	r=5	r = 10
Resnet50	TriHard+L2	71.2	86.5	94.2	72.1	83.2	92.5	86.0	88.4	95.7	82.7	95.7	98.1
	TriHard+SP	79.0	91.3	95.8	78.8	86.7	94.7	91.0	93.1	97.4	88.8	97.4	98.6
Resnet50-X	TriHard+L2	71.6	86.9	94.7	69.9	82.5	92.4	86.4	88.8	96.3	82.8	96.1	98.1
	TriHard+SP	79.4	91.0	96.3	78.3	86.1	95.0	91.5	93.4	97.6	88.2	97.0	98.5

4.1.3 基于度量学习的互学习方法

本小节介绍了一种基于度量学习的互学习方法(Mutual learning)。深度互学习(Deep mutual learning, DML)是论文^[43]首次提出，是指利用两个或者多个网络同时训练互相学习，来提升单个网络的性能。该论文利用KL散度(Kullback-Leibler divergence)衡量两个分类器的性能差异。如图4-5,假设 p_1 和 p_2 是两个分类器的预测，定义 p_1 到 p_2 的KL距离为：

$$D_{KL}(p_2||p_1) = \sum_{i=1}^N \sum_{m=1}^M p_2^m(x_i) \log \frac{p_2^m(x_i)}{p_1^m(x_i)} \quad (4-9)$$

DML目标是为了使两个分类器的预测性能足够接近，即KL距离足够小，于是在两个分类器的损失函数上加上各自的KL距离。最终两个分类器的损失函数写作：

$$\begin{aligned} L_{\theta_1} &= L_{C_1} + D_{KL}(p_2 || p_1) \\ L_{\theta_2} &= L_{C_2} + D_{KL}(p_1 || p_2) \end{aligned} \quad (4-10)$$

其中 L_{C_1} 和 L_{C_2} 分别是两个网络的分类损失。

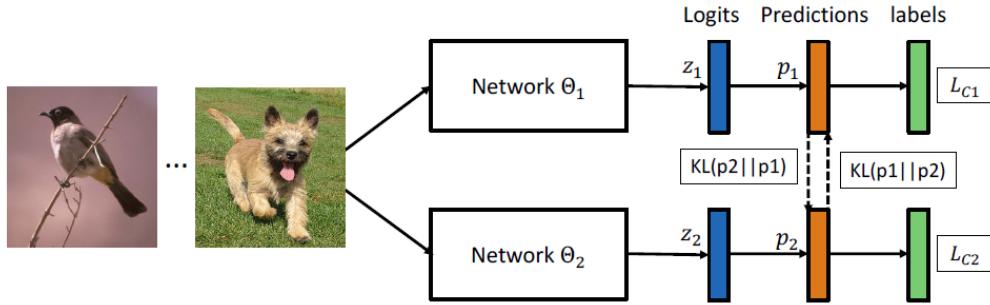


图 4-5 基于分类损失的深度互学习结构

DML是一个通用的基于分类损失的互学习模型。在行人ReID领域，DarkRank^[44]首次采样这种思想。DarkRank的核心思想是，首先训练一个性能非常好的大网络，这个网络成为Teacher网络。之后用这个Teacher网络和一个小网络一起进行互学习，这个小网络称为Student网络。通过这种互学习方法，Student网络能够从Teacher网络身上学习到知识，最终达到和Teacher网络差不多的性能。于是，小网络Student网络可以达到它自学无法达到的性能。但是由于KL散度只适合衡量两个分类器的性能差异，所以不适合应用到度量学习相关的方法中。然而在行人重识别以及人脸识别等目前主流应用中，度量学习是比表征学习更加有效的方法。因此，我们提出一种基于度量学习的互学习方法(Mutual learning based on metric learning)。

首先，给定一个包含 N 张图片的batch，每个网络都提取它们的特征，之后我们可以计算任意两张图片之间的距离并得到一个 $N \times N$ 的batch距离矩阵(batch distance matrix, BDM)。假设我们有网络 θ_1 和 θ_2 ，它们的BDM分别是 M^{θ_1} 和 M^{θ_2} ，则损失函数可以简单地定义为：

$$L_{M_1} = \frac{1}{N^2} \sum_i^N \sum_j^N (M_{ij}^{\theta_1} - M_{ij}^{\theta_2})^2 \quad (4-11)$$

之后我们可以计算 L_{M_1} 关于 $M_{ij}^{\theta_1}$ 的一阶偏导：

$$\frac{\partial L_{M_1}}{\partial M_{ij}^{\theta_1}} = \frac{2}{N^2} (M_{ij}^{\theta_1} - M_{ij}^{\theta_2}) \quad (4-12)$$

考虑到一些优化算法用到了二阶偏导，于是我们计算它的二阶偏导：

$$\frac{\partial^2 L_{M_1}}{\partial M_{ij}^{\theta_1} \partial M_{ij}^{\theta_2}} = -\frac{2}{N^2} \quad (4-13)$$

从中可以看出，在二阶偏导上网络 θ_1 和网络 θ_2 是不正交的。收到论文^[43]，我们希望两个网络更新参数的时候尽可能独立的。于是我们将公式4-11简单地扩展为：

$$L_{M_2} = \frac{1}{N^2} \sum_i^N \sum_j^N \left([ZG(M_{ij}^{\theta_1}) - M_{ij}^{\theta_2}]^2 + [M_{ij}^{\theta_1} - ZG(M_{ij}^{\theta_2})]^2 \right) \quad (4-14)$$

其中 $ZG(\cdot)$ 代表零梯度函数(Zero gradient function)。当计算梯度时，零梯度函数把变量当常数对待。类似的我们计算一阶偏导和二阶偏导：

$$\frac{\partial L_{M_2}}{\partial M_{ij}^{\theta_1}} = \frac{2}{N^2} (M_{ij}^{\theta_1} - ZG(M_{ij}^{\theta_2})) = \frac{\partial L_{M_1}}{\partial M_{ij}^{\theta_1}} \quad (4-15)$$

$$\frac{\partial^2 L_{M_2}}{\partial M_{ij}^{\theta_1} \partial M_{ij}^{\theta_2}} = 0 \quad (4-16)$$

一阶导数依然保持不变，但是二阶偏导等于0，也就是说在二阶偏导上两个网络是完全正交的。当我们优化 L_{M_2} 时，子网络们可以互相从对方学习知识。我们让两个网络独立地从对方学习知识比他们“一起作弊”要好。DarkRank是让一个Student网络从Teacher网络里学习知识，也就是说Student网络的性能上界就是Teacher网络在，这也要求Teacher网络需要有很好的性能。和DarkRank不同，我们的方法是让几个网络互相学习，并没有限制网络能够达到的性能上界，也不需要一个性能比较好的网络来带领其他网络。最终，我们的互学习方法结构如图4-3所示，整个损失函数包括分类损失、基于SP的度量损失 L_t 、基于分类的互学习KL损失和本文提出的基于度量学习的互学习损失 L_{M_2} 。

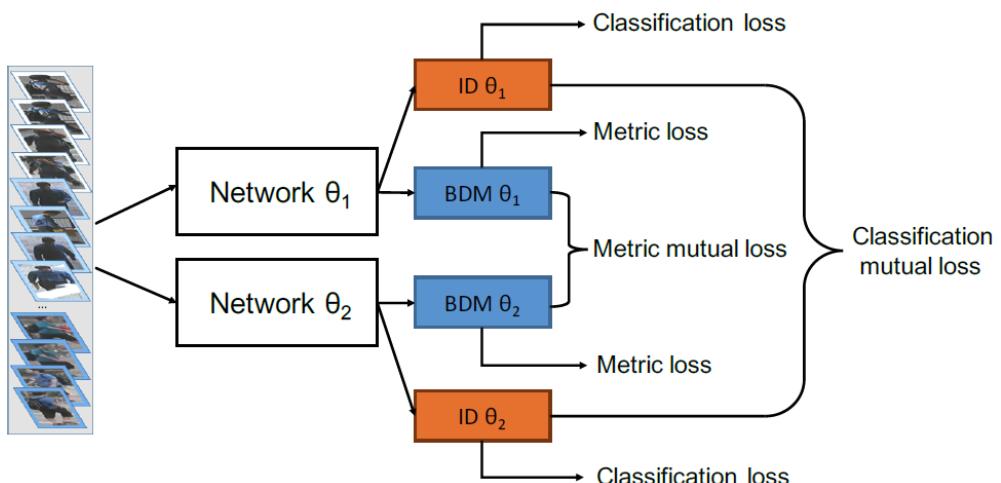


图 4-6 基于度量学习的互学习方法结构示意图

表 4-3 互学习实验结果对比

Loss	Base model	Market1501			MARS			CUHK-SYSU			CUHK03		
		mAP	r = 1	r=5	mAP	r = 1	r=5	mAP	r = 1	r=5	r = 1	r = 5	r = 10
L2+/-	Resnet50	71.2	86.5	94.2	72.1	83.2	92.5	86.0	88.4	95.7	82.7	95.7	98.1
	Resnet50-X	71.6	86.9	94.7	69.9	82.5	92.4	86.4	88.8	96.3	82.8	96.1	98.1
L2+KL	Resnet50	77.3	90.5	96.5	74.2	84.9	94.8	89.6	91.7	96.8	86.5	96.7	98.4
	Resnet50-X	77.1	90.6	96.4	74.4	84.9	93.7	89.6	92.1	96.8	86.8	96.7	98.2
L2+Ours	Resnet50	77.6	90.9	96.6	75.0	85.1	94.8	91.3	93.4	98.5	87.5	97.5	98.8
	Resnet50-X	78.3	90.9	96.6	75.8	85.7	94.9	91.7	93.7	97.7	88.2	97.6	98.8
SP+/-	Resnet50	79.0	91.3	95.8	78.8	86.7	94.7	91.0	93.1	97.4	88.8	97.4	98.6
	Resnet50-X	79.4	91.0	96.3	78.3	86.1	95.0	91.5	93.4	97.6	88.2	97.0	98.5
SP+KL	Resnet50	79.3	91.1	97.1	75.3	84.1	93.6	92.1	94.1	97.9	90.6	98.4	99.2
	Resnet50-X	79.1	91.0	96.3	76.3	85.5	94.8	91.5	93.3	97.5	88.4	97.8	99.0
SP+Ours	Resnet50	82.2	92.4	97.1	79.1	86.8	95.2	93.7	95.3	98.5	91.9	98.7	99.4
	Resnet50-X	82.3	92.6	97.2	78.5	87.3	95.3	93.2	94.6	98.4	91.1	98.6	99.3
	Ensemble	83.9	93.3	97.3	80.2	87.7	95.6	93.7	95.2	98.5	92.5	98.7	99.4

在本论文中，我们使用Market1501、MARS、CUHK-SYSU和CUHK03公开数据集进行了实验，实验采取Resnet50 和Resnet50-X作为互学习的两个网络。为了证明有效性，我们使用L2距离和SP距离分别进行了对比实验，最终实验结果如表4-3所示。其中字符”/”代表没有使用互学习，分别单独训练的基准实验。KL为论文^[43]提出的基于分类损失的互学习方法，而ours是指本文提出的互学习方法。Ensemble是指我们连接了两个模型提取的特征来提升性能。对于L2和SP距离，实验表明我们的方法比只用KL损失的互学习方法和不用互学习方法要取得更好的结果。并且实验在另一个侧面进一步证明了SP距离比L2距离能够取得更好的结果。

4.1.4 行人重识别的人类准确度评估

为了测试我们的行人重识别方法的性能，我们设计了一个人类准确度评估系统(Human performance evaluation system)来测评人类在行人重识别问题上的表现。该系统使用了前文介绍的Market1501、CUHK03和MARS数据集。

首先，我们先用我们训练好的模型提取所有数据集测试集中的图像特征，之后计算query集中每一张图片和gallery集中每一张图片的欧拉距离。对于query中的每一张probe图片，我们挑选一张距离最近（最像/最简单）的正样本和九张距离最近（最像/最难的）的负样本。为了方便测试，我们开发了一个评估系统，系统截图如图4-7所示。待测者要从这10张图片中选择出哪张和probe图片是相同的ID。

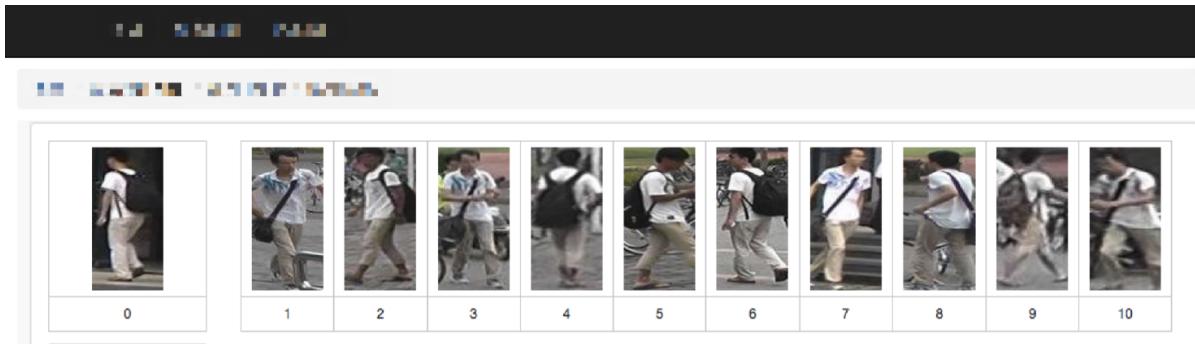


图 4-7 行人重识别的人类准确度评估界面

总共有10位志愿者参加了我们的测试，其中包括我们课题的研究人员、专业的标注全职员工和一般的普通市民。根据我们系统的设计原理，我们只能统计CMC曲而不能计算mAP准确度。最终为了我们统计10位测试者的在Market1501、CUHK03和MARS三个数据集上rank-1准确度，并且选择了其中成绩最好的结果作为人类准确度(Human performance)。结果如表所示4-4，CUHK03因为gallery集数据量比较少，所以最后结果最高达到了95.7%的rank-1准确度。在Market1501和MARS数据集上，人类分别取得了91.8%和90.2%的rank-1准确度。

表 4-4 人类准确度评估结果.

	Market1501	CUHK03	MARS
Rank-1	91.8	95.7	90.2

4.1.5 与现有方法结果对比

在本小节我们将会对比我们的方法与一些现有方法的结果。此外，我们还进行了一场行人重识别的人机大战，对人类准确度和计算机的准确度。结果表明，我们的方法在所使用的Market1501、MARS、CUHK-SYSU以及CUHK03这四个数据集上，都大大超过了现有方法的准确度，并且首次超越了人类准确度。

结果对比如表4-5 ~ 4-8所示。我们用论文题目首字母表示该论文方法的结果，缺损的结果用短横线表示，* 表示Arxiv上未发表的论文。Our-single是指我们最好的单模型取得的结果。Our-ensemble是指我们融合了互学习训练的两个网络的特征取得的结果。RK表示我们用论文^[45]提出的重排序(Re-ranking)来代替L2距离检索图片。黑色粗体表示现有方法的最好结果，蓝色粗体表示人类的准确度，绿色粗体表示我们最好的单模型结果，红色表示我们最好的多模型集成的结果。

表 4-5 Market1501结果对比

Methods	mAP	r=1
Temporal ^[46]	22.3	47.9
Learning ^[47]	35.7	61.0
Gated ^[13]	39.6	65.9
Person ^[48]	45.5	71.8
Re-ranking ^[45]	63.6	77.1
Pose ^[22]	56.0	79.3
Scalable ^[49]	68.8	82.2
Improving ^[9]	64.7	84.3
In ^[19]	69.1	84.9
In (RK) ^[19]	81.1	86.7
Spindle ^[23]	-	76.9
Deep ^{[43]*}	68.8	87.7
DarkRank ^{[44]*}	74.3	89.8
Human Performance	-	91.8
Our-single	82.3	92.6
Our-single (RK)	91.2	94.0
Our-ensemble	83.9	93.3
Our-ensemble (RK)	92.0	94.7

表 4-6 CUHK03结果对比

Methods	r=1	r=5	r=10
Person ^[50]	44.6	-	-
Learning ^[47]	62.6	90.0	94.8
Gated ^[13]	61.8	-	-
A ^[51]	57.3	80.1	88.3
Re-ranking ^[45]	64.0	-	-
In ^[19]	75.5	95.2	99.2
Joint ^[52]	77.5	-	-
Deep ^{[8]*}	84.1	-	-
Looking ^{[53]*}	72.4	95.2	95.8
Unlabeled ^{[54]*}	84.6	97.6	98.9
A ^{[55]*}	83.4	97.1	98.7
Spindle ^[23]	88.5	97.8	98.6
DarkRank ^{[44]*}	89.7	98.4	99.2
Human Performance	95.7		
Our-single	91.9	98.7	99.4
Our-single (RK)	96.1	99.5	99.6
Our-ensemble	92.5	98.7	99.4
Our-ensemble (RK)	97.0	99.6	99.7

表 4-7 MARS结果对比

Methods	mAP	r=1
Re-ranking ^[45]	68.5	73.9
Multi ^{[56]*}	-	68.2
MARS ^[2]	49.3	68.3
In ^[19]	67.7	79.8
In (RK) ^[19]	77.4	81.2
Quality ^{[57]*}	51.7	73.7
See ^[58]	50.7	70.6
Human Performance	-	90.2
Our-single	79.1	86.8
Our-single (RK)	85.6	87.5
Our-ensemble	80.2	87.7
Our-ensemble(RK)	86.5	87.8

表 4-8 CUHK-SYSU结果对比

Methods	mAP	r=1
End ^[4]	55.7	62.7
Neural ^{[59]*}	77.9	81.2
Deep ^{[60]*}	74.0	76.7
Our-single	93.7	95.3
Our-ensemble	93.7	95.2

在Market1501数据集上，目前现有最好的方法取得了89.8%的rank-1准确度和81.1%的mAP，而我们的单模型取得92.6%的rank-1准确度和82.3%的mAP，远远超过了现有的方法。经过RK之后，这个结果涨到了94.0%的rank-1准确度和91.2%的mAP。通过集成互学习的两个模型，我们取得了93.3%的rank-1准确度和83.8%的mAP，通过RK这个结果提升到94.7%的rank-1准确度和92.0%的mAP。我们测试的人类准确度是91.8%rank-1准确度，超越了现在所有的方法，但是我们的单模型就已经超越人类准确度。

在CUHK03数据集上，目前现有最好的方法取得了89.7%的rank-1准确度、98.4%的rank-5准确度和99.2%rank-10准确度，而我们的单模型取得91.7%的rank-1准确度、98.7%的rank-5准确度和99.4%的rank-10准确度，超过了现有的方法。经过RK之后，rank-1涨到了96.1%。通过集成互学习的两个模型，我们取得了92.5%的rank-1准确度、99.5%的rank-5准确度和99.6%rank-10准确度，通过RK这个结果提升到97.0%的rank-1准确度、99.6%的rank-5准确度和99.7%rank-10准确度。由于CUHK03的gallery集比较小，所以人类准确度也比较高，达到了95.7%rank-1准确度，远远超越了现在所有的方法和我们的模型。但是通过RK进行重排序之后，我们的单模型最终依然超越了人类的准确度。

在MARS数据集上，目前现有最好的方法取得了81.2%的rank-1准确度和77.4%的mAP，而我们的单模型取得86.8%的rank-1准确度和79.1%的mAP，远远超过了现有的方法。经过RK之后，这个结果涨到了87.5%的rank-1准确度和85.6%的mAP。通过集成互学习的两个模型，我们取得了87.7%的rank-1准确度和80.2%的mAP，通过RK这个结果提升到87.8%的rank-1准确度和86.5%的mAP。我们测试的人类准确度是90.2%rank-1准确度，超越了现在所有的方法以及我们的所有模型。但是MARS提供图像的序列特征，而我们的方法是基于单帧图像，我们的方法依然有潜能超越这个结果。

在CUHK-SYSU数据集上，我们并没有测评人类准确度，并且关注这个数据集的论文也不多，因此我们也没有使用RK等技术来提高准确度。目前现有最好的方法取得了81.2%的rank-1准确度和77.9%的mAP。我们的单模型取得95.3%的rank-1准确度和93.7%的mAP。经过RK之后，这个结果涨到了94.0%的rank-1准确度和91.2%的mAP。通过集成互学习的两个模型，我们取得了93.6%的rank-1准确度和95.2%的mAP。集成模型并没有带来准确度的提升，说明我们的模型已经达到了一个比较难提升的地步，结果也表明我们的结果远远超过了现在所有的方法。

综上概括，我们基于SP距离的互学习方法在行人重识别问题上取得了很好的结果，超越了所有现存的方法，并且首次超越了人类的准确度。并且我们的方法很容易浮现，而且也是一个比较通用的方法，在图像检索相关的问题上都可以取得应用。

4.2 成果作品

本文作者已经撰写学术论文三篇和发明专利两篇。其中学术论文有一篇一作期刊论文已经录用，一篇共同一作论文在投计算机视觉顶会CVPR2018，一篇二作论文arxiv已经挂出待投。发明专利其中一篇处于审核状态，一篇处于编修状态。

- (1) H Luo, Z Luo, C Xu, W Jiang, et al. Optical plasma boundary reconstruction based on least square for EAST Tokamak[J]. *Frontiers of Information Technology & Electronic Engineering*, In Press.
- (2) H Luo¹, X Zhang¹, X Fan, W Jiang, C Zhang, et al. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification. Contribute to CVPR, 2018.
- (3) Q. Xiao, H. Luo, and C. Zhang. Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification. *arXiv preprint arXiv:1710.00478*, 2017.
- (4) 罗浩, 张弛. 一种联合五元组和分类损失的行人重识别方法。
- (5) 罗浩, 张弛. 一种基于边界样本挖掘的行人重识别方法。

4.3 研究计划

¹Equation contribution

参考文献

- [1] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, Qi Tian. Scalable person re-identification: A benchmark[C]//Computer Vision, IEEE International Conference. 2015.
- [2] Springer. MARS: A Video Benchmark for Large-Scale Person Re-identification[J], 2016, 2016.
- [3] Wei Li, Rui Zhao, Tong Xiao, Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification[J]. 2014:152–159.
- [4] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, Xiaogang Wang. End-to-end deep learning for person search[J]. arXiv preprint arXiv:1604.01850, 2016.
- [5] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking. 2016.
- [6] Doug Gray, Shane Brennan, Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking[J]. 2007.
- [7] Martin Hirzer, Csaba Beleznai, Peter M. Roth, Horst Bischof. Person re-identification by descriptive and discriminative classification[C]//Scandinavian Conference on Image Analysis. 2011:91–102.
- [8] Mengyue Geng, Yaowei Wang, Tao Xiang, Yonghong Tian. Deep transfer learning for person re-identification[J]. arXiv preprint arXiv:1611.05244, 2016.
- [9] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Yi Yang. Improving person re-identification by attribute and identity learning[J]. arXiv preprint arXiv:1703.07220, 2017.
- [10] Liang Zheng, Yi Yang, Alexander G Hauptmann. Person re-identification: Past, present and future[J]. arXiv preprint arXiv:1610.02984, 2016.
- [11] Tetsu Matsukawa, Einoshin Suzuki. Person re-identification using cnn features learned from combination of attributes[C]//Pattern Recognition (ICPR), 2016 23rd International Conference on. IEEE, 2016:2428–2433.
- [12] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks[C]//International Conference on Neural Information Processing Systems. 2012:1097–1105.
- [13] Rahul Rama Varior, Mrinal Haloi, Gang Wang. Gated siamese convolutional neural network architecture for human re-identification[C]//European Conference on Computer Vision. Springer, 2016:791–808.
- [14] Florian Schroff, Dmitry Kalenichenko, James Philbin. Facenet: A unified embedding for face recogni-

- tion and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:815–823.
- [15] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, Shuicheng Yan. End-to-end comparative attention networks for person re-identification[J]. IEEE Transactions on Image Processing, 2017.
- [16] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1335–1344.
- [17] Weihua Chen, Xiaotang Chen, Jianguo Zhang, Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification[J]. arXiv preprint arXiv:1704.01719, 2017.
- [18] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles[C]//Computer Vision and Pattern Recognition. 2016:2167–2175.
- [19] Alexander Hermans, Lucas Beyer, Bastian Leibe. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.
- [20] Qiqi Xiao, Kelei Cao, Haonan Chen, Fangyue Peng, Chi Zhang. Cross domain knowledge transfer for person re-identification[J]. arXiv preprint arXiv:1611.06026, 2016.
- [21] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, Gang Wang. A siamese long short-term memory architecture for human re-identification[C]//European Conference on Computer Vision. Springer, 2016:135–153.
- [22] Liang Zheng, Yujia Huang, Huchuan Lu, Yi Yang. Pose invariant embedding for deep person re-identification[J]. arXiv preprint arXiv:1701.07732, 2017.
- [23] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, Xiaou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]. CVPR, 2017.
- [24] Longhui Wei, Shiliang Zhang, Hantao Yao, Wen Gao, Qi Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval[J]. arXiv preprint arXiv:1709.04329, 2017.
- [25] Taiqing Wang, Shaogang Gong, Xiatian Zhu, Shengjin Wang. Person re-identification by discriminative selection in video ranking[J]. IEEE transactions on pattern analysis and machine intelligence, 2016. 38(12):2501–2514.
- [26] Dongyu Zhang, Wenxi Wu, Hui Cheng, Ruimao Zhang, Zhenjiang Dong, Zhaoquan Cai. Image-to-video person re-identification with temporally memorized similarity learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017.
- [27] Jinjie You, Ancong Wu, Xiang Li, Wei-Shi Zheng. Top-push video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1345–1353.

- [28] Xiaolong Ma, Xiatian Zhu, Shaogang Gong, Xudong Xie, Jianming Hu, Kin-Man Lam, Yisheng Zhong. Person re-identification by unsupervised video matching[J]. Pattern Recognition, 2017. 65:197–210.
- [29] Niall McLaughlin, Jesus Martinez del Rincon, Paul Miller. Recurrent convolutional network for video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1325–1334.
- [30] Rui Zhao, Wanli Oyang, Xiaogang Wang. Person re-identification by saliency learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2017. 39(2):356–370.
- [31] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, Jiashi Feng. Video-based person re-identification with accumulative motion context[J]. arXiv preprint arXiv:1701.00193, 2017.
- [32] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking[J]. 2016:17–35.
- [33] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, Konrad Schindler. Mot16: A benchmark for multi-object tracking[J]. 2016.
- [34] Luis Patino, Tom Cane, Alain Vallee, James Ferryman. Pets 2016: Dataset and challenge[C]//Computer Vision and Pattern Recognition Workshops. 2016:1240–1247.
- [35] B. Keni, S. Rainer. Evaluating multiple object tracking performance: the clear mot metrics[J]. Eurasip Journal on Image & Video Processing, 2008. 2008(1):246309.
- [36] Anton Milan, Konrad Schindler, Stefan Roth. Challenges of ground truth evaluation of multi-target tracking[C]//Computer Vision and Pattern Recognition Workshops. 2013:735–742.
- [37] Lucas Beyer, Stefan Breuers, Vitaly Kurin, Bastian Leibe, Lucas Beyer, Stefan Breuers, Vitaly Kurin, Bastian Leibe. Towards a principled integration of multi-camera re-identification and tracking through optimal bayes filters[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017:1444–1453.
- [38] Laura Leal-Taixé, Cristian Canton-Ferrer, Konrad Schindler. Learning by tracking: Siamese cnn for robust target association[J]. 2016:418–425.
- [39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:770–778.
- [40] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices[J]. arXiv preprint arXiv:1707.01083, 2017.
- [41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:1–9.
- [42] François Chollet. Xception: Deep learning with depthwise separable convolutions[J]. arXiv preprint arX-

- iv:1610.02357, 2016.
- [43] Ying Zhang, Tao Xiang, Timothy M Hospedales, Huchuan Lu. Deep mutual learning[J]. arXiv preprint arXiv:1706.00384, 2017.
- [44] Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer[J]. arXiv preprint arXiv:1707.01220, 2017.
- [45] Zhun Zhong, Liang Zheng, Donglin Cao, Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding[J]. arXiv preprint arXiv:1701.08398, 2017.
- [46] Niki Martinel, Abir Das, Christian Micheloni, Amit K Roy-Chowdhury. Temporal model adaptation for person re-identification[C]//European Conference on Computer Vision. Springer, 2016:858–877.
- [47] Li Zhang, Tao Xiang, Shaogang Gong. Learning a discriminative null space for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1239–1248.
- [48] Ying-Cong Chen, Xiatian Zhu, Wei-Shi Zheng, Jian-Huang Lai. Person re-identification by camera correlation aware feature augmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.
- [49] Song Bai, Xiang Bai, Qi Tian. Scalable person re-identification on supervised smoothed manifold[J]. arXiv preprint arXiv:1703.08359, 2017.
- [50] Shengcai Liao, Yang Hu, Xiangyu Zhu, Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:2197–2206.
- [51] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, Gang Wang. A siamese long short-term memory architecture for human re-identification[C]//European Conference on Computer Vision. 2016:135–153.
- [52] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, Xiaogang Wang. Joint detection and identification feature learning for person search[C]//Proc. CVPR. 2017.
- [53] Igor Barros Barbosa, Marco Cristani, Barbara Caputo, Aleksander Rognhaugen, Theoharis Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification[J]. arXiv preprint arXiv:1701.03153, 2017.
- [54] Zhenzhong Zheng, Liang Zheng, Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[J]. arXiv preprint arXiv:1701.07717, 2017.
- [55] Zhenzhong Zheng, Liang Zheng, Yi Yang. A discriminatively learned cnn embedding for person re-identification[J]. arXiv preprint arXiv:1611.05666, 2016.
- [56] Yonatan Tariku Tesfaye, Eyasu Zemene, Andrea Prati, Marcello Pelillo, Mubarak Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets[J]. arXiv preprint arXiv:1706.06196, 2017.

- [57] Yu Liu, Junjie Yan, Wanli Ouyang. Quality aware network for set to set recognition[J]. arXiv preprint arXiv:1704.03373, 2017.
- [58] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification[J].
- [59] Hao Liu, Jiashi Feng, Zequn Jie, Karlekar Jayashree, Bo Zhao, Meibin Qi, Jianguo Jiang, Shuicheng Yan. Neural person search machines[J]. arXiv preprint arXiv:1707.06777, 2017.
- [60] Arne Schumann, Shaogang Gong, Tobias Schuchert. Deep learning prototype domains for person re-identification[J]. arXiv preprint arXiv:1610.05047, 2016.