

分类号

密级

UDC

编号

# 中国科学院 博士后研究报告

## 视频行人重识别优化研究

蒋小可

合作导师 乔宇研究员、闫俊杰博士

工作完成日期 2019 年 6 月— 2021 年 6 月

报告提交日期 2021 年 6 月

2021 年 6 月

Typeset by L<sup>A</sup>T<sub>E</sub>X 2 <sub>$\varepsilon$</sub>  at June 3, 2021

With package **PostDocRep** v0.1e of CT<sub>E</sub>X.ORG

# 视频行人重识别优化研究

## A Study on the Optimization of Video Person Re-Identification

博士后姓名

蒋小可

合作导师

乔宇研究员、闫俊杰博士

流动站（一级学科）名称

计算机科学与技术

专业（二级学科）名称

计算机科学与技术

研究工作起始时间

2019 年 6 月

研究工作期满时间

2021 年 6 月

单位名称

中国科学院深圳先进技术研究院

报告提交日期

2021 年 6 月



## 摘要

行人重识别 (Person Re-Identification) 技术逐渐深入到日常生活中，如商场智能分析，智能视频监控等领域，它可以帮助机器人凭背影识别顾客并提供相应服务，也可以在游乐园寻找走失的儿童。相比人脸识别，行人重识别也有独特优势，技术上也更有挑战。人体目标明显，一般视频系统拍到的人体比人脸数量要多一个数量级以上，它可以实现更精确的轨迹还原，甚至在识别不了人脸的情况下进行身份认证。但由于人体不是刚体，而且行人重识别的图像一般是在行人移动过程中捕获，而不是在用户配合的情况下拍摄，这造成从图片提取人体特征更加困难。加之不同的人体姿态，拍照光照、视角、遮挡等干扰因素，人体特征一般不如人脸特征有判别能力，容易导致行人重识别的错误。在海量数据的系统中，错误被放大和累积，最后极大影响了整个行人重识别系统的准确性和可靠性。

本文致力于通过时空信息来优化行人重识别系统的准确性。本文中的时空信息涵盖两个方面：第一个使用视频作为输入，实际应用中的绝大多数摄像头都是视频摄像头，记录下了现场一段时间内的时空信息；另一方面，在具体应用中，系统会收集到很多人的行为记录，以地铁场景为例，乘客的进站出站信息可以通过票据记录获得准确的信息，这些记录提供了长时间内行人的时空信息（空间信息是粗略的），提供了数学建模的原始数据。

视频提供现场时空信息，相比于依赖单图的行人重识别，视频多帧的信息可以起到互相补充参照的作用：

- 可变信息互补：不同角度的人体不同的视觉信息互补，拼凑一个更完整的人体信息，从而获得更好的人体特征。
- 不变信息去噪：利用不同帧之间变化的信息，可以有利于把人体与背景部分区分开来，从而减少噪声干扰。

我们通过注意力 (Attention) 机制，达到上面两点效果，提取出更加有判断力的人体特征。

对长期行人时空信息的利用则以贝叶斯概率作为主要手段，根据行人的长时问内历史行为信息，提前计算出行人在当前时间、地点出现的概率，经过离散、平滑等处理后，再与视觉特征相似度一起计算联合概率，从而获得更加准确的结果。在我们的模型中，我们把认为查询和目标当做是两个时空中有序的事件，基于时空序列增加约束。

上述两种方法分别代表了两种不同的优化思路：1) 首先考虑到人体特征判别

## 摘要

---

力弱且容易受到干扰，则从视频可以提供更多信息，把多个视频帧输入到深度网络中，从现场的时空信息中提取出更有判别力的特征。为了实现这个目标，我们提出了一个从全图搜索，寻找关键点局部特征的注意力机制，已经对应的无监督、半监督训练方法。2) 另一个研究是根据行人的长时间的时空行为模式建模，计算出现在当下时空的概率，来提升准确性。以视频作为输入和以行人的长期记录作为建模原始数据，在很多场合是可以全部成立或者部分成立的。我们从建立了比现有研究更加准确和更具有一般性的模型，并且给出了离散化、近似和求解和应用的方法。最后，这两个优化思路即可以独立使用，又可以联合使用。

**关键词：**行人重识别；时空信息，注意力机制

## Abstract

The applications of person Re-Identification has been growing in recent years, e.g. business intelligent analysis in shopping malls, intelligent video surveillance and other fields. It helps robots to identify customers from their backs and provide corresponding services, and they can also find lost children in kid parks. Compared with face recognition, person re-identification also has unique advantages and is technically more challenging. The human target is obvious, and the number of human body captured by the video system is more than an order of magnitude more than the number of faces. It can achieve more accurate trajectory restoration, and even perform identity authentication when the face cannot be recognized. However, because the human body is not a rigid body, and the image of person re-identification is generally captured during the movement of the pedestrian, rather than being shot with the cooperation of the user, this makes it more difficult to extract the human body features from the picture. Coupled with different human postures, photo-lighting, viewing angle, occlusion and other interference factors, human body features are generally not as capable of discriminating as human facial features, which can easily lead to errors in pedestrian re-identification. In a system with massive data, errors are amplified and accumulated, which ultimately greatly affects the accuracy and reliability of the entire pedestrian re-identification system. This article is dedicated to optimizing the accuracy of the pedestrian re-identification system through temporal and spatial information. The spatio-temporal information in this article includes two aspects: the first one uses video as input. Most cameras in practical applications are video cameras, which record the spatio-temporal information of the scene for a period of time; on the other hand, in a long running system, it will collect the behavior records of many people during a long period. Taking the subway scene as an example, the passenger's entry and exit information can obtain accurate information through ticket records. These records provide pedestrians' spatio-temporal information over a long period of time, provides the original data for mathematical modeling. Video provides on-site spatio-temporal information. Compared with person re-identification that relies on a single image, the information of multiple frames of video can serve as a complementary reference to each other:

- Variable information for complementation: The different visual information of the human body from different angles complement each other to piece together a more

---

## Abstract

---

- complete human body information, so as to obtain better human body features.
- Unchanging information for denoising: Using information that changes between different frames can help distinguish the human body from the background part, thereby reducing noise interference.

We use the Attention mechanism to achieve the above two effects and extract more discriminative human features. The use of long-period pedestrian spatio-temporal information uses Bayesian probability as the main method. According to the historical behavior information of pedestrians over a long period of time, the probability of pedestrians appearing at the current time and location is calculated in advance. Then we can calculate the joint probability together with the visual feature similarity to obtain more accurate results. In our model, we regard the query and the target as two ordered events in time and space, and add constraints based on the time and space sequence. The above two methods represent two different optimization aspects: 1) First, considering that the human body features are weak in discriminative power and susceptible to interference, more information can be provided from the video, and multiple video frames are input into the deep network. More discriminative features are extracted from the spatio-temporal information of the scene. In order to achieve this goal, we have proposed an attention mechanism that searches from the entire image to find the local features of key points, and has a corresponding unsupervised and semi-supervised training method. 2) Another study is to model the long-term temporal and spatial behavior of pedestrians and calculate the probability of appearing in the current time and space to improve accuracy. Taking video as input and long-term records of pedestrians as modeling raw data can be established in whole or in part in many cases. We have established a model that is more accurate and more general than the existing research, and given the methods of discretization, approximation, solution and application. Finally, these two optimization ideas can be used independently or in combination.

**Keywords:** Person ReID; Spatial-temporal; Attention

## 目 录

附表清单.....	J
插图清单.....	K
<b>第 1 章 引言 .....</b>	<b>1</b>
1.1 课题研究的背景和意义 .....	1
1.2 国内外研究现状 .....	3
1.2.1 特征表示与损失函数 .....	3
1.2.2 视频行人重识别 .....	4
1.2.3 注意力机制 .....	5
1.2.4 3D 卷积 .....	5
1.2.5 基于时空概率建模 .....	6
1.2.6 行人重识别系统设计 .....	6
1.3 本文研究内容和方法 .....	7
1.4 文章结构 .....	7
<b>第 2 章 相关技术原理 .....</b>	<b>8</b>
2.1 行人重识别概述与视频行人重识别 .....	8
2.1.1 基于表征学习的行人重识别 .....	8
2.2 距离度量 .....	8
2.2.1 基于局部特征的方法 .....	10
2.2.2 视频行人重识别 .....	11
2.3 三维卷积 .....	11
2.4 注意力机制 .....	11
<b>第 3 章 基于注意力的视频行人重识别优化方法 .....</b>	<b>15</b>
3.1 本章引言 .....	15
3.2 方案设计 .....	17
3.2.1 自分离网络 .....	18
3.2.2 三维卷积网络块 .....	21
3.2.3 训练和损失函数 .....	21

## 目 录

---

3.3 实验结果 .....	22
3.3.1 实验设置 .....	22
3.3.2 SSN+3D 分析 .....	23
3.3.3 与现有方法比较 .....	24
3.3.4 SSN 分析 .....	27
3.4 本章总结 .....	33
<b>第 4 章 基于时空行为模式的行人重识别优化 .....</b>	<b>34</b>
4.1 本章引言 .....	34
4.2 进站模型 .....	36
4.3 出站模型 .....	37
4.4 联合建模 .....	37
4.4.1 离散和平滑 .....	37
4.4.2 标准化 .....	38
4.4.3 联合指标 .....	38
4.5 优化和求解 .....	38
4.6 实验 .....	39
4.6.1 实验配置 .....	40
4.6.2 实验细节 .....	41
4.6.3 实验结论 .....	43
4.7 贡献 .....	44
<b>参考文献 .....</b>	<b>46</b>
<b>致 谢 .....</b>	<b>51</b>
<b>个人简历 .....</b>	<b>53</b>
基本情况 .....	53
教育状况 .....	53
工作经历 .....	53
研究兴趣 .....	53
联系方式 .....	53
<b>发表文章目录 .....</b>	<b>54</b>

## 附表清单

表 3.1 不同学习策略的影响。半监督学习效果最好，无监督学习效果其次。 ..	23
表 3.2 两轮分类是否共享权重的影响。共享权重效果比不共享高出近 20%，突出了共享权重的重要作用，也说明整个设计中两轮分类的重要性 .....	24
表 3.3 iLIDS-VID 数据集比较.....	25
表 3.4 MARS 数据集比较 .....	26
表 3.5 DukeMTMC-Video 行人重识别数据集比较.....	26
表 4.1 概率优化方法参数表 .....	39

## 插图清单

图 1.1 SSN3D 整体架构设计, 它的输入为若干张视频帧, 输出位一个高维向量表示的人体特征 .....	2
图 1.2 基于视觉、时空行为两路特征的行人重识别, 图引自 <sup>[5]</sup> .....	6
图 1.3 行人重识别系统系统设计: 首先从海量视频中检测行人, 提取特征并建立索引, 从而建立起特征库 (gallery) 并向外提供查询接口; 查询时输入目标行人的视频帧, 提取特征, 然后从索引中进行搜索并返回结果。现代 AI 系统一般部署在云平台, 硬件资源如 CPU, GPU, 内存, 存储全部被虚拟化, 以供平云台软件 (如 Kubernetes) 调度, 从而确保系统的可靠性、可扩展性。	7
图 2.1 各种损失函数在尽可能公平情况下进行对比的结果。该表引用自 <sup>[15]</sup> ...	10
图 2.2 图片切割提取体征典型方案 PCB。该图引用自 <sup>[19]</sup> .....	10
图 2.3 Scale Dot Product Attention (左) 和对应的多 head attention (右)。该图引用自 <sup>[2]</sup> .....	12
图 2.4 Non-local Neural Network 设计。该图引用自 <sup>[3]</sup> .....	13
图 2.5 DANET 设计。该图引用自 <sup>[25]</sup> .....	13
图 2.6 DANET 设计。该图引用自 <sup>[25]</sup> .....	14
图 3.1 SSN3D 整体架构设计, 它的输入为若干张视频帧, 输出位一个高维向量表示的人体特征 .....	16
图 3.2 SSN 的结构。它以图像作为输入, 并输出 $N$ 个空间特征。注意分类器将 $X$ 的每个像素分类两次 .....	19
图 3.3 The top-1 and mAP on (a)iLIDS-VID, (b)MARS(b), and (c)DukeMTMC-Video .....	25
图 3.4 Strong Learning Ability. The x-axis is the training iterations, and the y-axis is a total loss .....	27
图 3.5 Performance of Unsupervised Learning. Our SSN model is able to spot similar parts between different paintings when appreciating our Amber Abstract dataset. Type 1 captures the pattern with a light blue protuberance on the left; type 2 spots the feature of an orange diamond above a blue stripe; and each type 3 contains a small deep blue block .....	28
图 3.6 更多 Amber Abstract 上的可视化结果 .....	29

图 3.7 半监督学习在 Pedestrian 128 数据集上的结果。从图中可见人体关键点寻找基本准确。	30
图 3.8 无监督学习在 Pedestrain 128 数据集上的结果。它会有一些瑕疵，比如说一些非人体部分学到，一些人体部分不稳定	32
图 4.1 行人的时空行为记录具备序列性，且满足马尔科夫性质	34
图 4.2 在不给定中间状态情况下，任意两个节点都具备马尔科夫性质	36
图 4.3 统计用户当月乘坐地铁次数，双次数比单数次多很多，往返均乘坐地铁的情况占多数	40
图 4.4 用户当月乘车次数的占比	41
图 4.5 不同频次乘客的乘车次数在当月总乘车次数的占比	41
图 4.6 异常图片中戴口罩乘车比例	41
图 4.7 108 张异常数据中，有两位用户分别出现 21 次和 20 次，剩余的数据用户错误次数均在 5 次以下	41
图 4.8 加入时空分数前	42
图 4.9 加入时空分数后	42
图 4.10 时间维度平滑前	43
图 4.11 时间维度平滑后	43
图 4.12 视觉分数	43
图 4.13 联合分数	43
图 4.14 有效负样本的原始视觉分数	44
图 4.15 有效负样本的联合分数	44
图 4.16 出站模型联合分数	44

## 第1章 引言

### 1.1 课题研究的背景和意义

随着深度学习(Deep Learning)技术取得突破性进展，人工智能(Artificial Intelligence)目前正处于一个蓬勃发展中，各种不同形式的人工智能应用逐渐深入日常生活，行人重识别(Person Re-Identification)是其中代表性应用之一。行人重识别指的是在所有广泛覆盖的(无重叠)摄像头视角下出现的所有行人中识别出目标人物，该技术广泛应用与商业智能分析(Business Intelligence)，智能视频监控等领域，比如在游乐园寻找走失儿童、机器人可以凭背影准确识别顾客并提供相应服务。

行人重识别与人工智能的另一种代表性技术——人脸识别，有类似的技术目标——识别出人的身份，但二者一个依赖人体，一个依赖人脸。这形成了行人重识别有独特的优势和挑战。相比于人脸，人体目标更大，更容易被摄像头拍到，所以在各种监控摄像头中，拍到人体数量往往比人脸高一个数量级，可以更多的捕捉到人的信息，更完整的还原人的行动轨迹。行人重识别的图像一般是在行人移动过程中捕获，而不是在用户配合的情况下拍摄，所以拍摄到的画面因运动而导致模型。另一方面人体不是刚体，并且不同视角、不同姿态、不同光照下的人体形态有较大的变化，导致人体特征不如人脸特征那样具有判别力，更容易导致错误；更糟糕的是，在大规模海量数据场景下，海量的人体特征，巨大的人的ID底库，小概率的错误识别被放大并不断累积，最后极大影响了整个系统的准确性和可靠性。

为了提升行人重识别的准确性，我们有不同的思路。最直接、最核心的提升是提取更加准确的人体特征；除此之外，还可以利用系统长时间运行积累的信息，从这些累积信息中做数据分析和挖掘，来辅助人体重识别。本文主要利用现场和长期的行人时空信息来优化行人重识别。

现场的时空信息指的是视频。为了提取更准确的人体特征，视频行人重识别领域(Video Re-Identification)领域受到关注。顾名思义，视频行人重识别，AI系统接受视频输入(同一个人的连续多帧图片)，然后从这一段视频不同角度、姿态、位置、光线的人体中提取到人体的特征，因为这种不变的特征与角度、姿态、位姿、光线等无关，所以更加具有识别能力。提取到的特征必须具备下列属性：

- 可度量：特征必须具备度量属性，即同一个人从不同图片里提取出来的特征距离小，不同的人的特征距离大。只有具备可度量的属性，特征才可以“区分”开来。可以把这样的特征视为可度量高维空间里的点，点与点之间有距离，它们可以进行聚类。

- 不变性：特征必须与角度、光照、姿势等无关，同一个人在不同的光照、角度和姿势下必须具备类似的特征，不因环境因素而变化，而是真正反映人体本身的“不变”的特征。

在实际情况中，特征一般都已高维向量表示，特征之间的距离以余弦距离度量，也就是说距离在  $[-1, 1]$  之间。越相似的特征距离越接近 1。

视频保存着行人出现时的时间和空间的投影（图片）信息，相比于依赖一张图片的行人重识别，视频时空信息更加丰富，这些视频帧之间可以彼此补充，达到两个目的：

- 可变信息互补：不同角度的人体不同的视觉信息互补，拼凑一个更完整的人体信息，从而获得更好的人体特征。
- 不变信息去噪：利用不同帧之间变化的信息，可以有利于把人体与背景部分区分开来，从而减少噪声干扰。

但从视频多帧中提取到不变的人体特征，也面临着一系列的技术问题，其中最典型问题排除角度、姿态、位置和遮挡物的干扰。本研究通过自注意力 (self-attention) 的方式对不同图像中（时域）人体的不同部位（空域）进行对齐，再通过三维卷积对不同部位分别提取特征，最后综合形成一个充分利用视频中时空信息，具有强不变性的人体特征，其整体设计如图1.1所示。

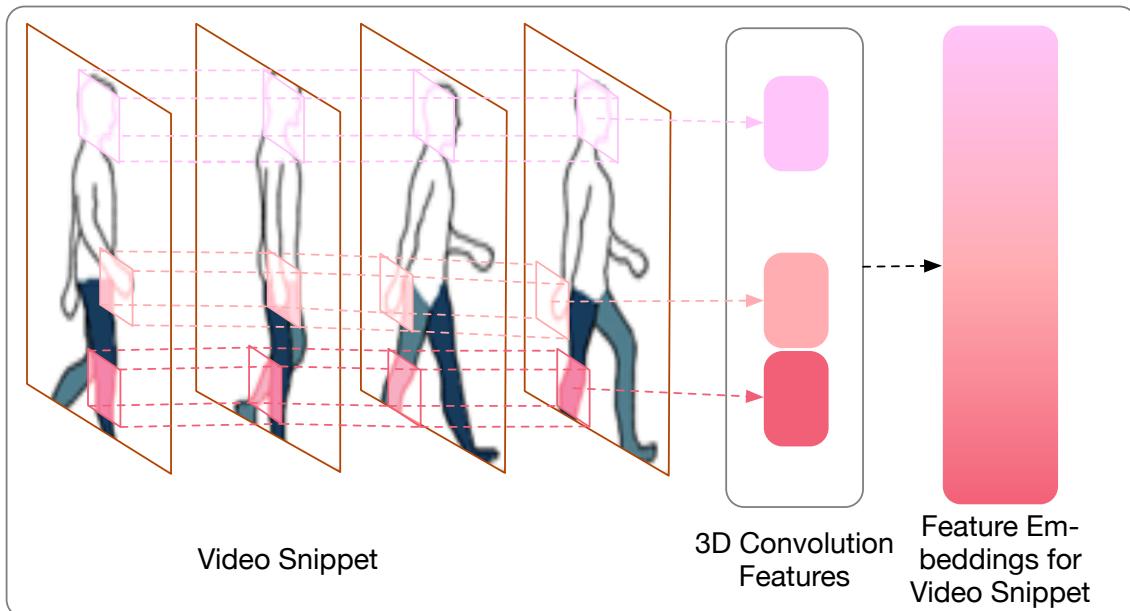


图 1.1 SSN3D 整体架构设计，它的输入为若干张视频帧，输出位一个高维向量表示的人体特征

长期的时空信息，是指 AI 系统在长期运行过程中积累的行人行为数据。在一般的场景下（例如安防），通过对特征进行聚类，或者对已归类数据的整理，我们

粗略地收集到了不同行人在较长期范围内的运动轨迹通过这些轨迹可以建立模型：

$$h' = \arg \max_h J(V_h, P_h)。 \quad (1.1)$$

在以上定义中， $h$  为行人的 ID， $J$  是具有给定视觉相似性评分 ( $V$ ) 和乘客时空概率评分 ( $P$ ) 的联合度量函数。上面式子的目标，把从原本最大化视觉分  $V_h$  的目标转向最大化视觉和行为联合概率。通过对真实数据的统计显示，真实乘客的平均时空概率分比其视觉最近邻乘客的平均时空行为概率大 300 倍。这构成了使用行人长期时空行为概率的来提升行人重识别准确度的基础假设。在实际系统中，我们要考虑离散概率和  $J$  的平滑特性，具体函数的设计还需要符合实际情况，不能简单讲  $V, P$  相乘，这样会导致明显荒谬的结果，即使  $V$  大而  $P$  为零，简单的  $V \cdot P$  也会导致零概率。

上述两种方法分别代表了两种不同的优化思路：首先考虑到人体特征判别力弱，则从视频可以提供更多信息，把多个视频帧输入到深度网络中，从现场的时空信息中提取出更有判别力的特征；然后在根据行人的长时间的是空信息建模，计算出出现在当下时空的概率，来提升准确性。以视频作为输入和以行人的长期记录作为建模原始数据，在很多场合是可以全部成立或者部分成立的。并且，这两个优化思路即可以独立使用，又可以联合使用。

## 1.2 国内外研究现状

最近一波人工智能的热潮主要得益于于深度学习为代表的技术取得突破，因此本文主要关注深度学习相关的视频行人重识别技术。

行人重识别虽然面临很大挑战，但是其数据采集便利，应用场景广阔，具有很强的安全和商业价值，在研究界也收到很多关注。在 1996 年就有了相关研究；2006 年，这一研究领域再次重回研究社区<sup>[1]</sup>；此后，相关工作和成果便不断涌现，各种数据集也不断发布。

行人重识别与人脸识别技术一样，都包含两个部分：第一部分是提取特征，第二部分是对比查找。其中特征提取是核心，也是本文关注的部分。

### 1.2.1 特征表示与损失函数

特征表示的方法，核心是提取行人特征的不变性。传统视觉上采用的方法有颜色直方图，Gabor 特征、颜色名称等。现在通过深度学习得到的特征，其表达能力已经完全超越了这些传统方法获得特征。这些特征都具备上文定义的可度量、不变的特性，但在深度学习领域，对如何获得计算特征的模型有不同的思路。一般

来讲，都会设计一个网络结果并定义一个损失函数，然后使用带标签的数据来寻求这个网络。当把两张属于同一个人的照片输入到一个特征提取的行人重识别网络中，损失函数下降；相反，如果将两张不属于同一个人的输入到网络，则损失函数上升。这种通过损失函数的上升和下降的方法来引导网络参数迭代，达到同一个人特征距离近，不同人特征距离远的效果。

在上述方法的基础上，研究者提出三元组损失函数（及其各种变种）训练的模型，函数输入为三位组，以其中一张照片作为基准，再取一张与基准照同一个人的照片，和一张与基准照不同人的照片，三张照片构成一个正样本对，一个负样本对，输入到网络中去进行训练。考虑到随机选取时，正负样本往往是容易判断的样本，这将导致网络泛化性能差，因此研究者基于训练批次在线难样本抽样的方法（online hard example mining, OHEM）提出一种 TriHard 损失函数。从同一批次的图片中，选取最难判别的正样本（同一个人最不相似的照片），和最难判别的负样本（不同人最相似的照片），实验结果显示，TriHard 损失函数优于传统的一般三元组损失函数。

### 1.2.2 视频行人重识别

以视频多帧作为输入数据，而不是单张图片，来做行人重识别的方式，被称作视频行人重识别（Video Person Re-Identification）。由于行人重识别的数据集本身比较模糊，传统的方法如光流，HOG3D、步态提取等图像运动信息很难再取得本质提升，一些深度学习的方法被应用到这个领域。视频行人重识别多了一个时间维度的信息，为了处理时间维度的信息，研究者们将 RNN/LSTM 等适合序列数据处理的技术应用于此。具体思路上，对于时间维度信息的使用则有不同的考虑，例如帧之间的运动信息，多帧特征融合，帧质量判断等，总体思路是通过多帧信息去掉噪声，提取更正确特征。

从信息上说，多帧之间的信息可以互相补充，做如下两种方式的信息提取：

- 可变信息互补，例如不同角度的人体可以拼凑一个更好的人体特征。
- 不变信息提取：利用不同帧之间变化的信息，可以有利于把人体与背景部分区分开来，从而减少噪声干扰。

为了实现上面两点，最关键的技术是对齐，即将人体各个部分做对齐。只有做好了对齐，就能区分背景和人体，区分人体不同部位，才能提取出不变的特征和排除杂质的干扰。

### 1.2.3 注意力机制

注意力机制 (attention model) 原本是应用于自然语言处理领域的技术，但最近开始广泛应用于计算机视觉领域，并取得了突破性的成绩，是深度学习技术中最值得关注和深入了解的技术之一。

注意力机制是人大脑处理信息的机制之一，当人看图片的时候，会先快速扫描全局图像，获得感兴趣的区域（比如人的脸部，文章的标题、首句和加粗部分），也就是注意力焦点，然后再仔细观察观察焦点区域，排除无关区域的干扰，获得更多、更精确的信息。这是人类利用有限智力资源，从大量信息中心快速提取有价值信息的方法，极大的提高了视觉信息处理的效率和准确性。

深度学习中的注意力机制思路和目标与大脑注意力机制如出一辙。Google 提出 transformer 论文<sup>[2]</sup>中重提了自注意力 (Self-Attention) 机制，宣称”Attention is All You Need”，并取得了巨大成功。该文章提出以 (Key, Query, Value) 三元组来捕捉长距离依赖，其中最重要的模块是缩放点积注意力 (scaled dot-product attention)，Key 和 Query 通过点乘的方式获得注意力权重，以此权重与 Value 做点乘可得到最终的输出。

Transformer 与 Attention 的机制引发了一系列后续研究。Non-Local Neural Network<sup>[3]</sup>继承了 (Key, Query, Value) 三元组的建模思路，但具体设计上则采用 3 个  $1 \times 1 \times 1$  卷积实现了一个 non-local 的模块。DANET 并且把注意力机制分别用在 Spatial 和 Channel 上，然后进行特征融合，逻辑清晰，并取得了很好的效果。

自注意力机制往往能带来较好的效果，但是因为要为每一像素都捕捉全局的上下文信息，导致注意力模块有较大的计算开销和显存开销，例如在 DANET 中，空间注意力的 attetnion map 尺寸为  $(H \times W) \times (H \times W)$ ，其中  $H$  和  $W$  是输入特征的高和宽。

### 1.2.4 3D 卷积

在计算机视觉领域经常使用的是二维卷积 (2D Dimensional Convolution)，这是因为图像本身在空间尺度上是二维的（高和宽）<sup>①</sup>。但是有一些数据本身除了空间两个维度之外，还有时间维度，典型如视频数据。这种三维数据比二维数据多了一个维度，为了更好的处理这类数据，研究者提出了三维卷积 (3D Dimensional Convolution)。三维卷积的卷积核本身是三维的<sup>②</sup>，并且其滑动也是在三维空间里进行。早在 2013 年，<sup>[4]</sup> 将三维卷积用于视频分类，取得了良好的效果。

<sup>①</sup> RGB 图片数据是三维的，但是卷积是在二维上滑动

<sup>②</sup> RGB 帧组成的视频，则形成四维数据

### 1.2.5 基于时空概率建模

[5]是利用时空概率建模的研究的代表作。该文章的思路清晰。如图1.2所示，其设计分为两路 (two streams) 处理，图上部分为是从图片中提取视觉特征，下部分是统计从一个摄像头到另一个摄像头的时间统计。两路数据做一个联合统计，得到最后结果。此文的思路跟本文相似，但是本人模型更加精确，考虑了不同时刻，

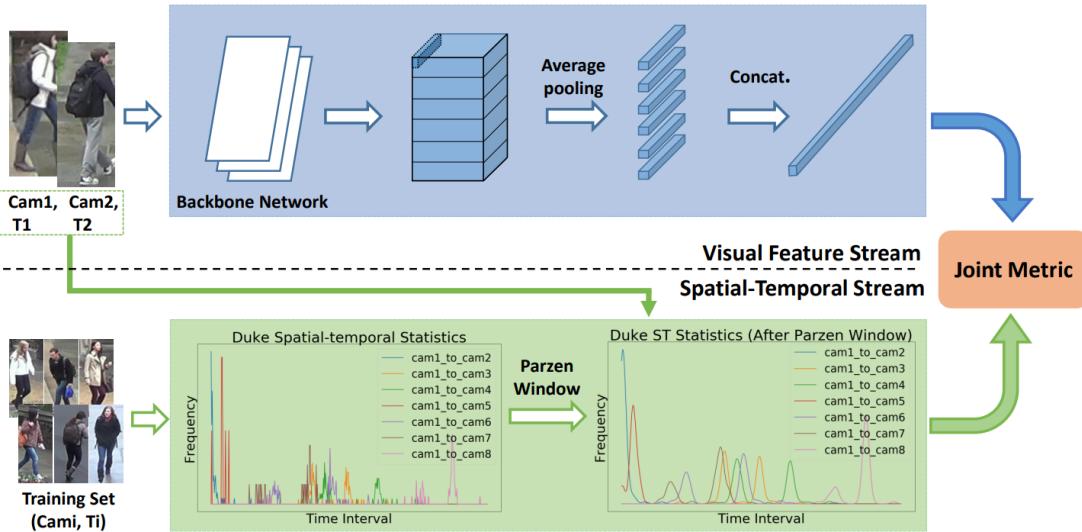


图 1.2 基于视觉、时空行为两路特征的行人重识别，图引自<sup>[5]</sup>

到达和离开等更细致的情况，模型更加贴近现实情况。

### 1.2.6 行人重识别系统设计

在计算机系统中视频处理往往需要比较大的资源。视频首先是摄像头通过透射投影对某一个时刻的采集视野范围内的信息，把采集到的信息以 RGB 颜色的形式存储下来，常见的分辨率为 1080p (即高 1080x 宽 1920 个像素)。在时间维度，一般以 24 帧/秒、30 帧/秒的速度进行采集。摄像头采集到的数据后，经过视频编码 (常见 H.264, VP9 等)，一般通过 rtsp 这样的流传输协议传输到服务器上，进行保存或者直接处理。在处理视频数据之前，先对视频进行解码，还原出每一帧的信息。

本文视频行人重识别工作也是假设一段视频被解码形成若干个视频帧，以此作为我们系统的输入。考虑到视频处理中较大的资源 (计算、内存、存储) 开销，加上 AI 系统本身就需要很多的计算资源，现代行人重识别系统往往都比较复杂。一个典型的设计如图1.3所示。底层的硬件资源往往先虚拟化成为一个资源池，提供给云平台进行统一的管理和调度。云平台软件如 kubernetes 提供了很多的管理和调度功能，比如自动重启，自动扩容，智能升级等，从而保障系统的可靠性

和可靠性。

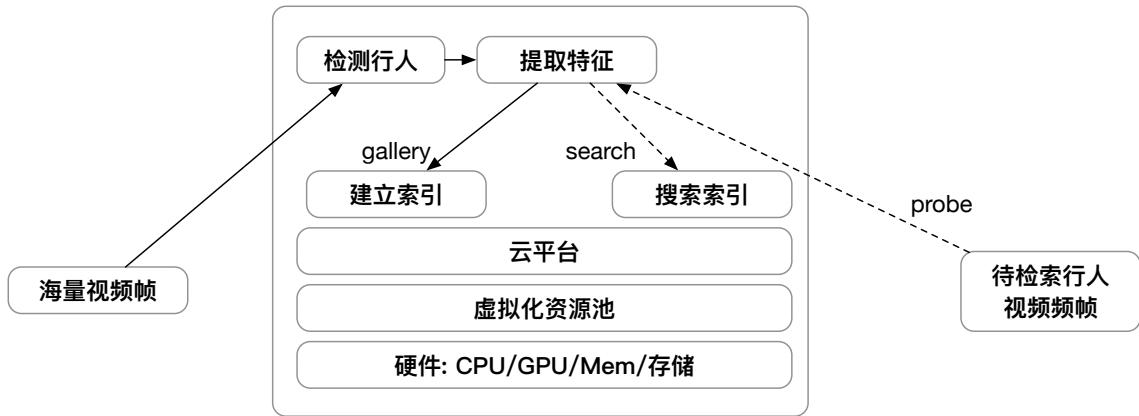


图 1.3 行人重识别系统系统设计：首先从海量视频中检测行人，提取特征并建立索引，从而建立起特征库 (gallery) 并向外提供查询接口；查询时输入目标行人的视频帧，提取特征，然后从索引中进行搜索并返回结果。现代 AI 系统一般部署在云平台，硬件资源如 CPU, GPU, 内存，存储全部被虚拟化，以供平云台软件（如 Kubernetes）调度，从而确保系统的可靠性、可扩展性。

### 1.3 本文研究内容和方法

为了获得更具有判别力的行人特征，帮助基于行人重识别的应用更好的落地，本文针对视频行人重识别中的对齐问题进行优化。具体方法是采用自注意力的方式，通过自注意力模块将视频里的多帧图片（时间维度）中的不同人体部位（空间）进行对齐，然后每个部位采用一个三维卷积分别提取特征，在此基础上综合提取一个基于视频的行人特征。为了实现这个目标，我们提出了一个从全图搜索，寻找关键点局部特征的注意力机制，以及对应的无监督、半监督训练方法。

另一个研究是，根据行人的长时间的时空行为模式建模，计算出出现在当下时空的概率，来提升准确性。以视频作为输入和以行人的长期记录作为建模原始数据，在很多场合是可以全部成立或者部分成立的。我们从建立了比现有研究更加准确和更具有一般性的模型，并且给出了离散化、近似和求解和应用的方法。。最后，这两个优化思路即可以独立使用，又可以联合使用。

### 1.4 文章结构

本问前文介绍了行人重识别的应用场景和挑战，本文的研究思路，简要介绍了相关研究。后面三章安排如下。第三章做相关工作做了一个更仔细的梳理，涵盖行人重识别里的表征学习、度量学习和基于局部特征的方法，视频行人重识别的最近的进展，三维卷积和注意力机制。第三章介绍基于注意力的行人重识别，第四章介绍基于时空行为模型的优化方法。

## 第2章 相关技术原理

### 2.1 行人重识别概述与视频行人重识别

#### 2.1.1 基于表征学习的行人重识别

基于表征学习 (Representation Learning) 行人重识别<sup>[6-9]</sup>，主要方法是利用卷积神经网络提取出表征特征 (Representation)。训练时，行人重识别问题会被当做是分类问题 (Classification) 或验证问题 (Verification)。这一点在训练时用的 loss 上体现明显：分类问题使用行人的 ID (有时还有属性) 作为标签来训练模型；验证问题则是输入一对行人图片，让网络判断是否属于同一个人。这两种 loss 可以一起使用，<sup>[6]</sup> 同时使用了两种 loss。经过足够数据的训练之后，每次输入一张测试图片，网络就会自动输出一个特征 (FC 层之前的输出)，这个特征可以用于行人重识别任务。

有研究者认为，仅依赖上述两种与 ID 有关的 loss 并不足以得到一个具有强泛化能力的模型，通过额外引入其它信息，如性别、头发、衣着属性可以更加准确。在引入属性标签后，模型同时要预测人的 ID，还要预测行人各项属性，这既大大增加了模型的返回能力，也提供了新的功能，运行通过属性对行人进行搜索和查找。表征学习方法鲁棒，训练稳定，是行人重识别领域的基础，不足之处是容易在数据集上过拟合。

### 2.2 距离度量

度量学习 (Metric Learning) 是一种广泛应用于图像检索领域的技术。度量学习思路是让网络学习到一对图片的相似度，到达网络能实现同一个人的不同图片 (正样本对) 间的相似度大于不同行人图片 (负样本对)。具体训练中设计损失函数让正样本距离尽可能小，负样本距离尽可能大。常见损失函数有对比损失 (Contrastive Loss)<sup>[10]</sup>，三元组损失<sup>[10-12]</sup>，四元组损失 (QuadLoss)<sup>[13]</sup>，难样本采样三元组损失<sup>[14]</sup>，边缘挖掘损失<sup>[15]</sup>等。下文对这些距离度量方法逐一介绍。

对比损失函数定义如下：

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

上面式子中， $(z)_+$  表示  $\max(z, 0)$ ， $\alpha$  根据实际需求调整的常数。当输入为正样本时， $y = 1$ ；否则  $y = 0$ 。为了最小化损失函数，当输入为正样本时， $d_{I_a, I_b}$  会变小，否

则会变大直到超过  $\alpha$ 。

三元组损失 (Triplet Loss) 目前使用非常广泛，它的思路是三张图片作为输入，其中一张作为基准 (anchor) 图片  $a$ ，与基准图片同一个人的正样本图片  $p$ ，和负样本图片  $n$ 。三元组损失表示为：

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

训练过程中为了最小化三元组 loss，模型会拉近正样本间的距离，增大负样本之间的距离。

<sup>[12]</sup> 认为 Triplet loss 没有考虑正样本之间的绝对距离，为此提出改进三元组损失：

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

QuadLoss<sup>[13]</sup> 在三元组损失基础上增加一个负样本图片，提出四元组损失：

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

其中  $n1, n2$  是两个负样本， $\alpha, \beta$  是手动设置的常数，前一项为强推动常数，后一项为若推动。该 loss 第二项不共享 ID，考虑的是正负样本间的绝对距离，实验显示，四元组损失通常学到更好的特征。

难样本采样三元组损失<sup>[14]</sup>的设计思想是，从训练的 batch 中挑选处最难区分的正样本和负样本进行训练，从而让网络具备更好的泛化性能。假设一个 batch 中有  $P$  个行人，每个行人随机挑选了  $K$  张照片，对于 batch 中的每一张照片，都可以为它选择一个最难得的正样本和一个最难的负样本，组成三元组。定义与选中的基准图片  $a$  同一个人的照片集合为  $A$ ，剩下的图片构成集合  $B$ ，则损失表示为：

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

边界挖掘损失 (MSML)<sup>[15]</sup> 也采用了难样本采样的思路。基于四元组损失函数，假设我们忽视  $\alpha, \beta$  的影响，可以更紧凑的方式表达四元组损失：

$$L_q = (d_{a,p} - d_{m,n} + \alpha)_+$$

上面公式中， $(m, n)$  是一个负样本对， $(a, p)$  是 batch 中最难的正样本对， $(m, n)$  是 batch 中最难的负样本对， $(a, m)$  既可以是正样本对，也可以是负样本对。也就是说，MSML 损失挑选处了最难的正样本和负样本，公式化表达如下：

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

文章<sup>[15]</sup> 对各种损失函数进行了尽可能的评测，评测记过如下表2.1。

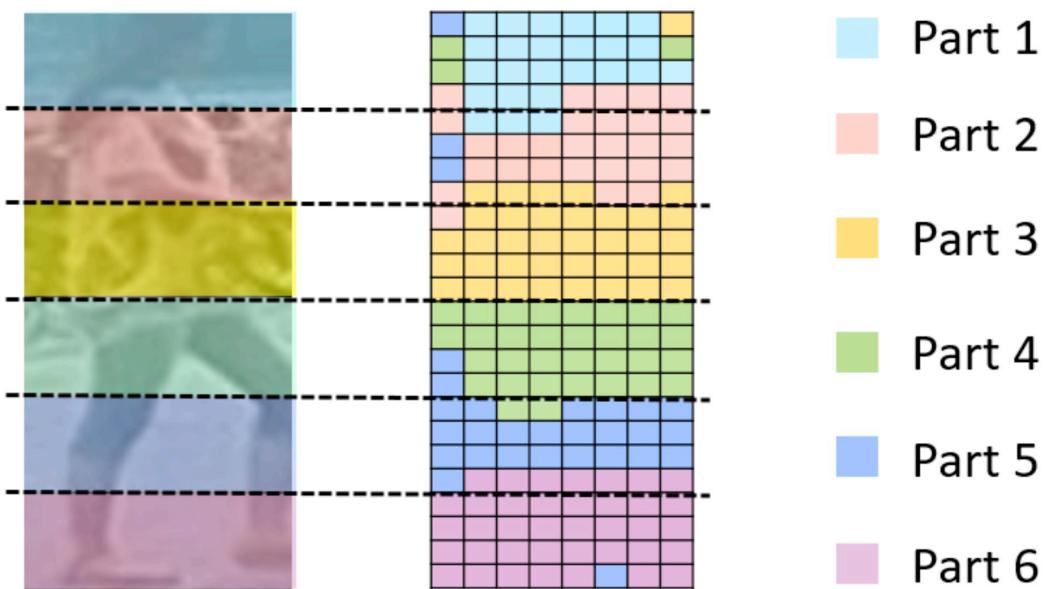
Base model	Methods	Market1501			MARS			CUHK-SYSU			CUHK03		
		mAP	r = 1	r = 5	mAP	r = 1	r = 5	mAP	r = 1	r = 5	r = 1	r = 5	r = 10
Resnet50	Cls	41.3	65.8	83.5	43.3	59.3	75.2	70.7	75.0	88.1	51.2	72.6	81.8
	Tri	54.8	75.9	89.6	62.1	76.1	89.6	82.6	85.1	94.1	73.0	92.0	96.0
	Quad	61.1	80.0	91.8	62.1	74.9	88.9	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	<b>69.6</b>	<b>85.2</b>	<b>93.7</b>	<b>72.0</b>	<b>83.0</b>	<b>92.6</b>	<b>87.2</b>	<b>89.3</b>	<b>96.4</b>	<b>84.0</b>	<b>96.7</b>	<b>98.2</b>
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	<b>95.4</b>	74.3	84.1	93.5	83.5	86.1	95.2	85.5	97.2	98.7
	MSML	<b>73.4</b>	<b>87.7</b>	95.2	<b>74.6</b>	<b>84.2</b>	<b>95.1</b>	<b>88.4</b>	<b>90.4</b>	<b>96.8</b>	<b>86.3</b>	<b>97.5</b>	<b>98.7</b>
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	<b>89.6</b>	<b>91.4</b>	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	<b>76.7</b>	<b>88.9</b>	<b>95.6</b>	<b>72.0</b>	<b>83.4</b>	<b>93.3</b>	<b>89.6</b>	90.9	<b>97.4</b>	<b>87.5</b>	<b>97.7</b>	<b>98.9</b>

图 2.1 各种损失函数在尽可能公平情况下进行对比的结果。该表引用自<sup>[15]</sup>

### 2.2.1 基于局部特征的方法

早期研究往往关注全局特征，即用整图得到一个人体特征，再进行检索。后来研究者发现，局部特征能取得更好的结果。常用局部特征的方法有图像切块<sup>[16]</sup>，骨架关键点定位和姿态矫正<sup>[17-18]</sup>。

图像切换是最常见的一种局部特征提取方式。<sup>[16]</sup> 把图片垂直分成若干等分，分隔号的图像块按照顺序喂送到 LSTM，最后的特征融合了所有局部的特征。典型方案是 PCB 极其改进版本<sup>[19]</sup>，如图2.2所示。这种方法缺点非常明显，要求图像是对齐的，如果没有对齐，比如头和上身在同一个切分里，将带来噪声。

图 2.2 图片切割提取体征典型方案 PCB。该图引用自<sup>[19]</sup>

为了解决上述对齐问题,<sup>[17]</sup>先用姿态估计的方法估计出行人身体的关键点,再使用仿射变换将关键点对齐。该文章利用了14个关键点。文章<sup>[18]</sup>也利用了14个关键点,不同的是,该文章不是讲关键点对齐,而是直接从关键点抠出感兴趣的区域。

### 2.2.2 视频行人重识别

视频行人重识别最近研究颇多,如MG-RAFA<sup>[20]</sup>, MGH<sup>[21]</sup>, ST-GCN<sup>[22]</sup>等。In particular, we investigated MG-RAFA, MGH, and ST-GCN. MG-RAFA通过有监督学习来训练像素级别的权重来构造注意力图(Attention Map)。平均池化(Average Pooling)是该方法依赖的一个重要方法。MGH和ST-GCN都通过图卷积的方式来建模帧内、帧间的不同部分的关系。一种类似与PCB的机制用于把特征分成若干个区域,这些不同区域构成了图卷积中的图节点。

AP3D<sup>[23]</sup>和我们的工作有相似之处。两者都包含了对齐和特征聚合的模块。但是具体的设计,二者有较大差别。AP3D的对其模块使用像素间语义相似度,而本文则采用的是简化的自注意力机制、交叉熵损失函数和两轮的分类。

在特征聚合方面,AP3D用一个特别的分支从多帧全局特征中捕捉时间维度的抖动。本文则采用三维卷积来捕捉时间维度的信息。

SpaAttn<sup>[24]</sup>与本文工作一样,也采用了注意力机制,但是注意力机制的设计也不同。SpaAttn是用一个基于KL散度。

## 2.3 三维卷积

三维卷积被设计出来用于处理除了空间两维之外还有第三个维度(如时间)的数据,典型如视频数据。三维卷积的卷积核本身是三维的,并且其滑动也是在三维空间里进行。与二维卷积(编码了2D域中目标的空间关系)类似,3D卷积可以描述3D空间中目标的空间关系。对某些应用(比如生物医学影像中的3D分割/重构)而言,这样的3D关系很重要,比如在CT和MRI中,血管之类的目标会在3D空间中蜿蜒曲折。早在2013年,<sup>[4]</sup>将三维卷积用于视频呢分类,取得了良好的效果。由于三维卷积是普通卷积非常自然的拓展,概念容易理解,本文不再赘述其基本原理和设计。

## 2.4 注意力机制

文章“Attention is All You Need”<sup>[2]</sup>提出了以自注意力取代循环神经网络,产生了重大的影响。文章设计了缩放点积注意力模块,如图2.3所示。

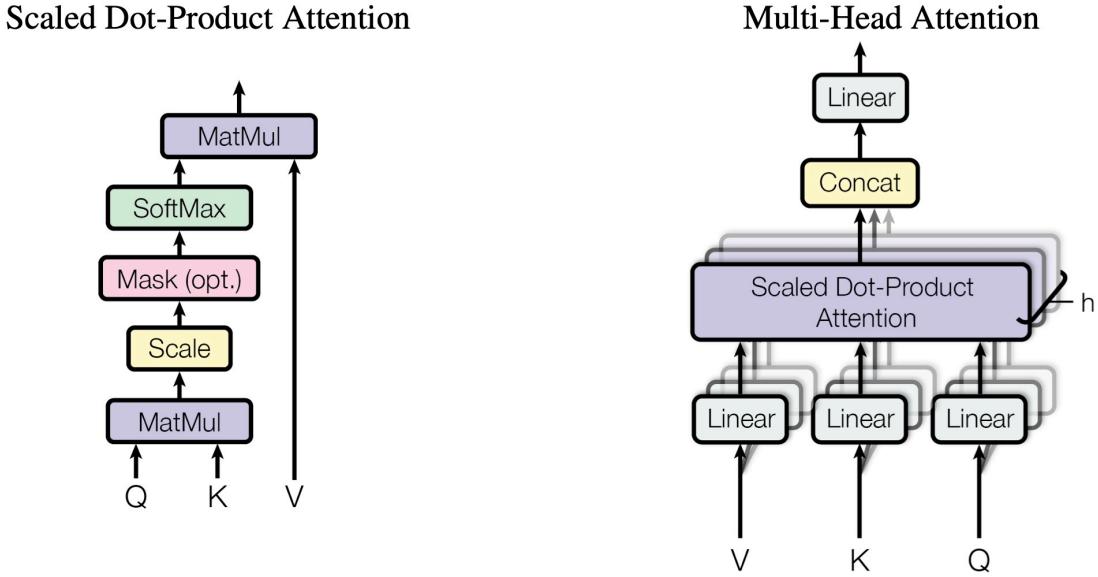


图 2.3 Scale Dot Product Attention（左）和对应的多 head attention（右）。该图引用自<sup>[2]</sup>

上图右侧是缩放点积注意力模块的多 head 版本，也就是说实现了多个 attention map（类似与多通道）。

如图所示，该文章提出以 (Key, Query, Value) 三元组来捕捉长距离依赖，其中最重要的模块是缩放点积注意力 (scaled dot-product attention)，Key 和 Query 通过点乘的方式获得注意力权重，以此权重与 Value 做点乘可得到最终的输出。

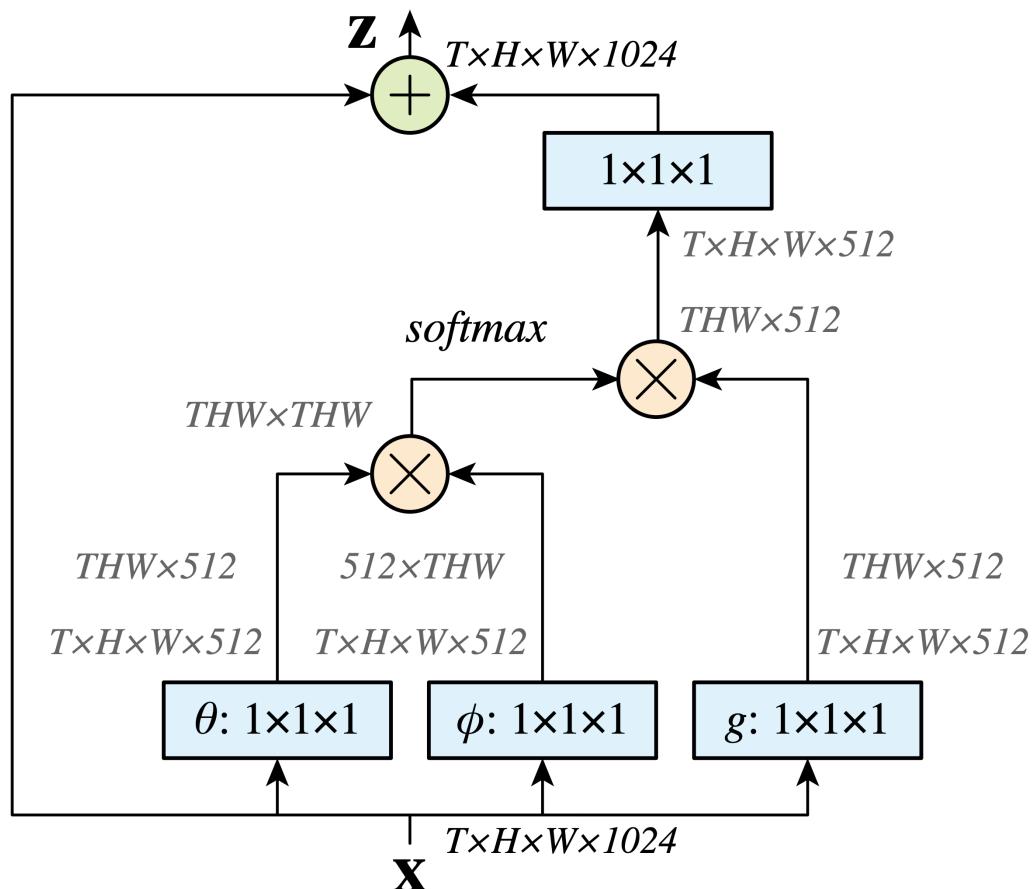
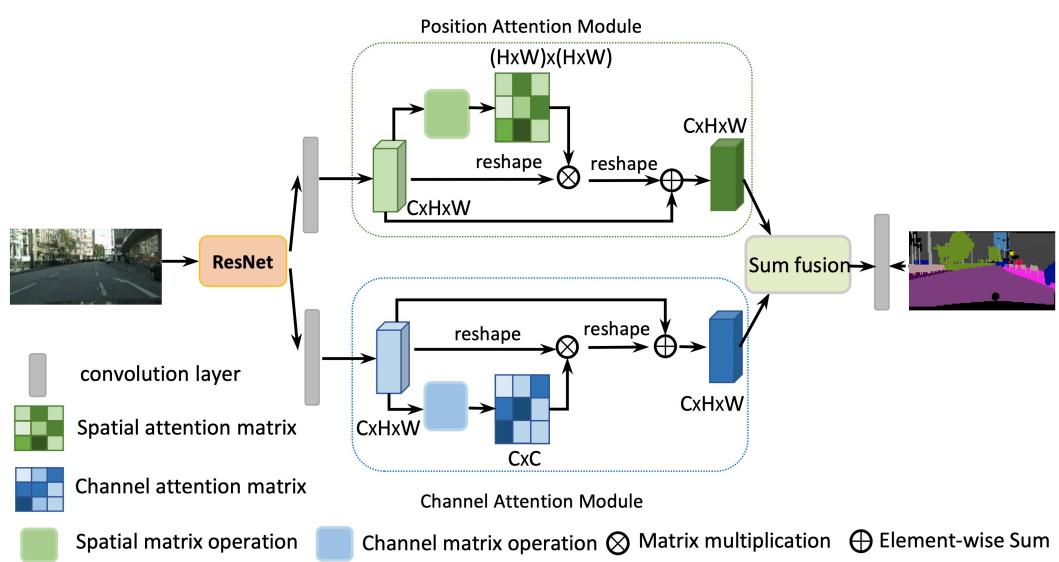
Non-local Neural Network<sup>[3]</sup> 继承了 (Key, Query, Value) 三元组的设计。但具体设计上则采用 3 个  $1 \times 1 \times 1$  卷积实现了一个 non-local 的模块，如图2.4。

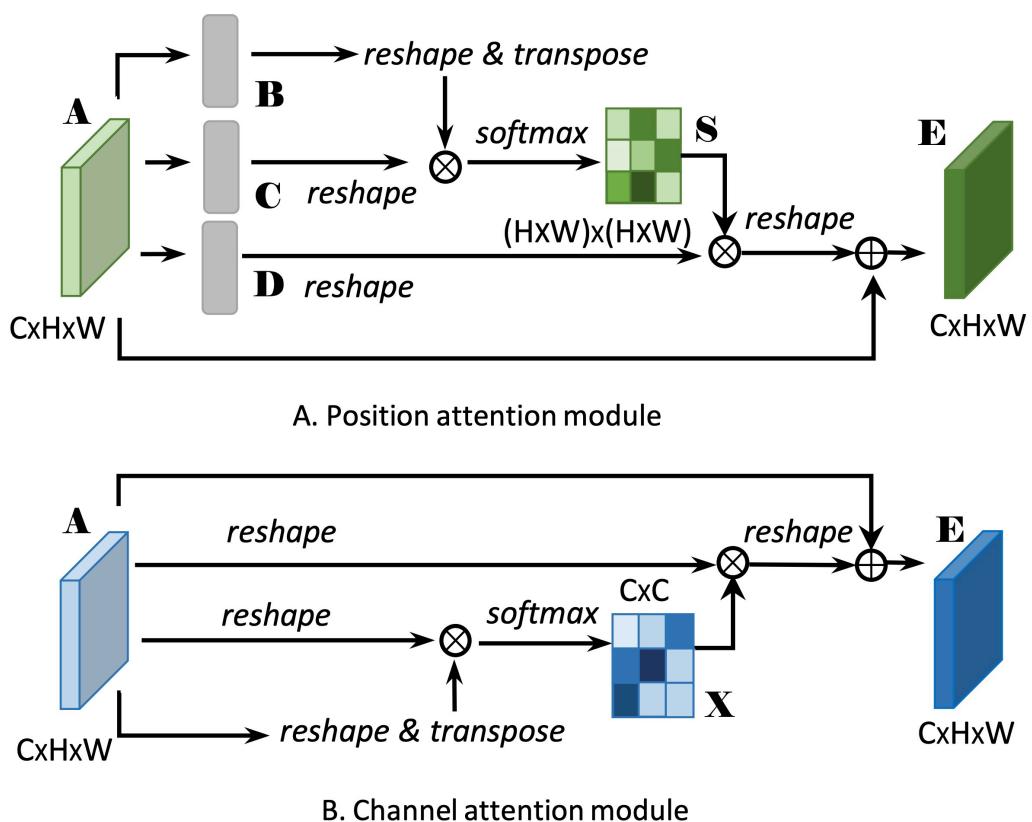
$$y_i = softmax(\theta(x_i)^T \phi(x_j)) \cdot g(x_j) = \frac{1}{\sum_{\forall j} e^{\theta(x_i)^T \phi(x_j)} e^{\theta(x_i)^T \phi(x_j)}} W_g x_j$$

该设计中最关键的步骤是乘法:  $\theta \cdot \phi$ ，这一步的输入是  $[h * w, c/2]$  的矩阵，输出是  $[h \cdot w, h \cdot w]$  的注意力图。

DANET<sup>[25]</sup> 也继承了三元组的设计，更进一步，把注意分别用在空间 (Spatial) 和通道 (Channel) 两个方面，然后进行特征融合，如图2.5所示。

在 DANET 设计中，Spatial 部分的 attetion 模块设计如下图。也是通过卷积实现 (Q,K,V) 三元组，如图2.6所示。

图 2.4 Non-local Neural Network 设计。该图引用自<sup>[3]</sup>图 2.5 DANET 设计。该图引用自<sup>[25]</sup>

图 2.6 DANET 设计。该图引用自<sup>[25]</sup>

## 第3章 基于注意力的视频行人重识别优化方法

### 3.1 本章引言

视频行人重识别以巨大的应用前景和研究挑战近年来吸引了很多研究者的关注<sup>[26-28]</sup>，它的目标是从视频中的空间和时间不同维度信息中提取去具有不变形、可度量的人体特征，从而实现更准确行人重识别。在具体应用时，一段目标人物的视频被用做探针 (probe)，用这个探针从大量的视频集 (gallery) 搜索出同一个人的视频。为了从视频中提取和聚合有效的信息，LSTM<sup>[29-30]</sup>，3D 卷积<sup>[31-33]</sup>和基于 non-local operation<sup>[3]</sup>的方案被广泛探索。其中基于三维卷积的方案以其处理三维数据特殊的设计，在这个任务上取得了很好的效果<sup>[32]</sup>。但是目前这些方案，都需要克服对齐的问题，即当从连续视频帧的不同空间位置里提取特征时，需要对这些图片中的不同的人体位置、相机视角进行对齐。我们检视了多个视频行人重识别的数据集，发现数据集本身并不完美，不完美体现在两个方面：1) 由于检测算法并不完全准确，在多帧里同一个人的人体框 (bounding box) 并不准确，存在遗漏或者错误的情况；2) 在多人情况下，同一个人的框在前后帧结果不一致。在这些数据集上使用姿态估计的模型，除了上述不准确的情况之外，连续多帧之间的关键点也出现不一致的情况。基于这样的原因，一些方法<sup>[34-35]</sup>，包括 non-local attention 被提出用来解决这种对齐错误的问题。

本章的目标是在三维卷积中优化其对齐问题，提升视频行人重识别准确性。如图3.1所示，其设计包含两个深度学习网络模块。第一叫自分离网络 (Self-Separated Network, SSN)，这个模块利用注意力机制，寻找人体上的不同部位。每一个视频帧都会进行这样的操作，也就是说，不同视频帧里同一个人的不同部位会被分别寻找出来，从而实现了对齐。通过这个模块，视频里每一个人体部位，都用一个四维张量来描述。另一个模块是基于三维卷积的特征聚合模块。每一个描述人体部位的四维张量都被作为一个三维卷积的输入，利用三维卷积在空间 (二维) 和时间三个维度的滑动，提取该人体部位的特征。每个部位提取出来的特征再被放入一个融合层 (Fusion Layer)，统一输出一个表征整个人体的特征向量。因为这个设计包含了 SSN 和三维卷积 (3D Convolution) 两部分，本章的设计被命名为 **SSN3D**。

SSN 也运用了注意力模型。并且本章设计了一种两轮分类法来利用 SSN，两轮分类法要求通过 SSN 注意力机制“注意”到的像素在前后两次分类中类别必须是一致的。这种一致性提供了一种可以用于训练的信号，利用这种信号再配合交叉熵损失函数，我们可以通过无监督或者半监督的方法来把人体的不同部分区分

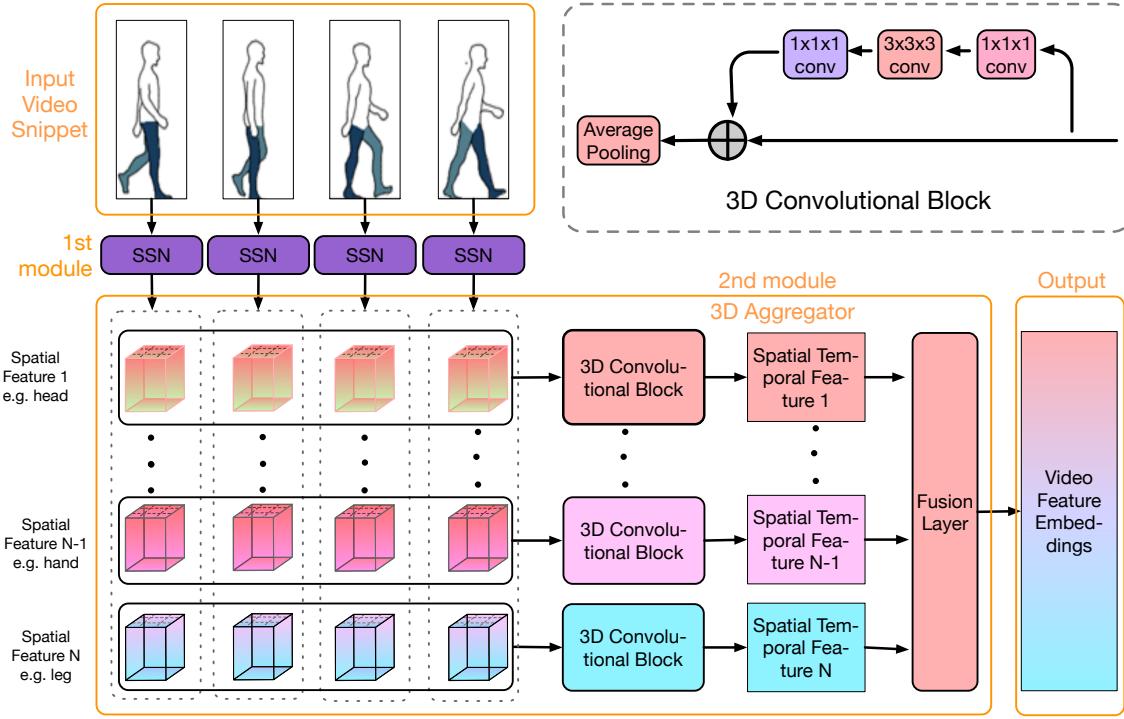


图 3.1 SSN3D 整体架构设计，它的输入为若干张视频帧，输出位一个高维向量表示的人体特征

出来。这也是我们叫它自分离网络的原因。但是通过无监督学习的方式学到的人体特征，具备较强的判别力，但因为缺乏人先验知识的引导只能依靠盲目的搜索，不能确保学到的人体部分/特征是人类可理解的，也不能确保每次学到的关键部位都是一样的。因此我们同时设计了一种有监督的学习，在有监督学习中把人的关键点输入网络作为瞄点，引导网络去学习我们希望学到的部分。但是，现有的姿态检测、关键点检测的模型<sup>[36-37]</sup>需要消耗大量的时间，并且在应用在视频的时候出现了不一致的情况。为了同时利用这两种策略的优势又规避它们的缺点，我们使用了一些高质量的标注数据来引导 SSN 进行训练，让网络学到我们的“目标”部位后，然后再用无标签的数据继续训练。通过这种半监督学习的策略，我们选择出来的人体特征不但具备可识别性，同样是人类可理解并且稳定的。实验结果显示这种训练方式效果超过了无监督模型，也超过了通过 OpenPose<sup>[37]</sup>来标注关键点的有监督模型。

SSN + 3D 这个组合非常适合聚合时间维度的信息。不同于传统的通过池化方法来融合一帧里的特征<sup>[27,38]</sup>，本章通过 SSN 把每张图里的关键部位提取出来，然后三维卷积在空间和时间三个维度滑动能够在对齐的部位信息进行特征提取，充分利用了三维卷积在处理视频数据的优势。相比于现有方法，SSN+3D 的方法很好地清除因为灯光、姿态、遮挡等导致的在时间和空间维度上的噪声，实现了更细致的人体部位对齐和特征聚合，让多帧之间互相补充，从变化信息中区分出人体，

并将多角度人体信息集成更好的特征。

总结起来，本章的研究方法有三个贡献：

- 本章提出了 SSN+3D 来处理视频行人重识别中关键部位对齐和信息提取的方法，它能够很好的处理空间和时间维度上的噪声。我们的方法在 iLIDS-VID 和 DukeMTMC-Video 数据集上都超过了当前最佳水平 (SOTA)，在 MARS 数据集上也与当前水平相仿。
- SSN 注意力机制与两轮分类的机制，可以通过像素级分类一致性来支持无监督的训练，这种方法可以用于各种不同的注意力机制。
- 不同部分分别用 3D 卷积的设计，相比于此前基于池化的信息聚合方法，我们基于三维卷积的方法粒度更细，取得更好的效果。

## 3.2 方案设计

图3.1展示了特征提取模型的总体框架，该图由两个模块，即 SSN 和三维卷积特征聚合器组成。第一个模块 SSN 是一个深度卷积网络，它在每个帧中选出人体不同的部分，其设计如图3.2中所示。SSN 从每个图像中找出  $N$  个人的不同部分。来自视频片段的相同类别人体部位 (例如头部) 对齐以形成四维张量。第二个模块是三维卷积。它将来自多个帧的属于同一部分的特征组合在一起，然后将结果输入到三维卷积，三维卷积沿时间维度提取这些部分的特征。三维卷积模块将  $N$  个四维张量输入到  $N$  三维卷积以进行特征聚合。整个系统以端到端的方式进行训练。

通过使用独立策略训练的注意力机制，SSN 有助于对齐那些空间变化的特征，以帮助三维卷积网络迅速提取特征。在第3.2.1节中，我们介绍了 SSN 的设计和实现方案。在第3.2.2节，我们将深入了解三维卷积网络体系结构。

在本章中，我们对数学符号做一些约定。我们使用粗体大写符号 (例如  $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ ) 表示张量；大写符号 (例如  $A, B, C$ ) 代表二维矩阵；粗体小写符号 (例如  $\mathbf{a}, \mathbf{b}, \mathbf{c}$ ) 代表矢量；小写符号 (例如  $a, b, c$ ) 代表一个标量。但出于习惯的原因，有一些例外，即  $H, W, C, N, K$  和  $T$ ，它们都是标量。 $H, W$  和  $C$  表示图像特征图的高度，重量和通道数。 $N$  表示第3.2.1节中提到的空间特征 (即选择的人体部分) 的数量。 $K$  表示由第3.2.1节中定义的注意力分类器生成的向量的长度。 $T$  是指样本图片的数量 (在第3.2.2节中使用)。

### 3.2.1 自分离网络

如图3.2中所示的 SSN，将图像作为输入并输出大小为  $3 \times 3 \times C$  的  $N$  个空间特征 (在本章的设计中， $N = 9, C = 1024$ )。我们使用 ResNet50<sup>[39]</sup>作为卷积层的 backbone，其输出具有大小  $H \times W \times C$  的特征。为了计算注意力图，我们在最后一个共享的卷积层上的卷积特征图上滑动了一个小窗口。滑动窗口中的特征点被输入到注意力分类器 (第一轮分类) 中，该分类器由卷积层和紧随其后的可将特征点映射到较低维度的全连接层组成。值得注意的是，分类器的末尾没有 softmax 层，因为我们将在空间上将这个 softmax 应用到注意力图上。我们使用  $3 \times 3$  滑动窗口。在对注意力图进行 softmax 操作之后，每个注意力图中所有元素的总和将为 1。此注意力图的每个组件都描述了其在特征图中对应滑动窗口的权重。在基于此计算滑动窗口特征的加权总和后，我们可以输出  $N$  个空间特征。

对于 SSN 训练，所产生的空间特征将通过 softmax 层 (第二轮分类) 被输入到同享权重的同一注意力分类器中。第二轮分类的结果为无监督训练创建标签，该结果应与第一轮分类保持一致。在这个过程里，即使我们不知道这些部分代表什么，我们使用交叉熵损失来确保每个标签都描述了不同的部分。我们的实验表明，损失将收敛在一个很小的值上，这表明该模型确实学习了不同的部分。此外，我们还可以使用半监督学习方法来引导目标成为我们想要的部分。在以下小节中，我们将详细解释图3.2。

#### 3.2.1.1 网络体系结构

我们介绍 SSN 的组件，如下所述。

**注意力分类器：**图3.2中展示的注意力分类器将卷积特征映射到  $N$  维向量。我们的注意力分类器有一个卷积层，它使用  $3 \times 3$  内核和  $K$  输出通道。这个卷积层将特征点映射到带有  $K$  元素的向量 (在我们的示例中， $K = 512$ )。在 ReLU 层之后，然后将向量输入到  $K \times N$  全连接层。我们定义

$$\mathbf{a}_{h,w} = \text{AttentionClassifier}(\mathbf{X}_{h,w}) \quad (3.1)$$

其中  $\mathbf{a}_{h,w} \in \mathbb{R}^N$  是输出，而  $\mathbf{X}_{h,w} \in \mathbb{R}^{3 \times 3 \times C}$  是在位置  $(h, w)$  处从特征图中滑动样本的特征，*AttentionClassifier* 为本章定义的注意力分类器。

在第一轮分类中，我们将注意力分类器应用于输入特征点的像素。Softmax 不应用在这个结果上，这意味着此轮将保留属于不同标签的像素的可能性。

**Attention Map:** 我们有从特征图生成的  $N$  个注意力图，其中第  $i$  个表示为  $A^i \in$

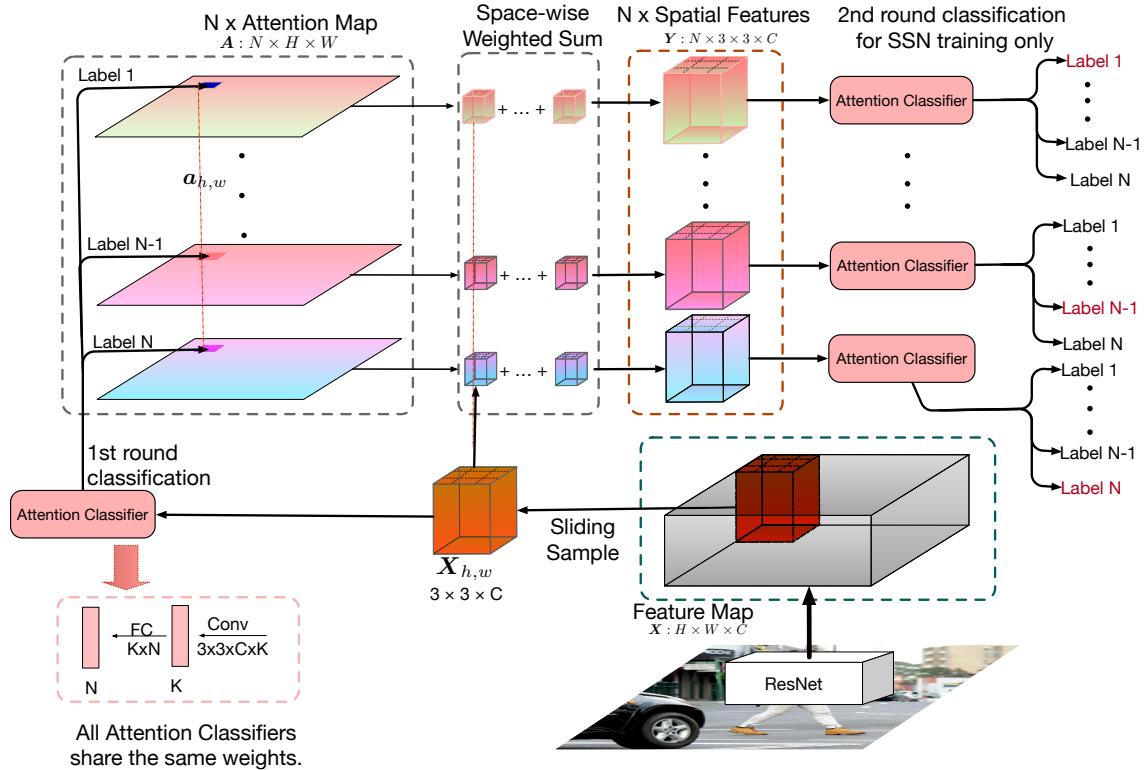


图 3.2 SSN 的结构。它以图像作为输入，并输出  $N$  个空间特征。注意分类器将  $X$  的每个像素分类两次

$\mathbb{R}^{H \times W}$ 。每个注意力图是由先前的注意力分类器计算的，即

$$A_{h,w}^i = a_{h,w}^i, \quad (3.2)$$

其中  $a_{h,w}^i$  是  $a_{h,w}$  的第  $i$  个元素。

在上一节中，我们已经解释了在第一轮分类中没有将 Softmax 应用于注意力图，并且当使用注意力图来计算具有滑动窗口特征的相应权重时，我们会执行此操作。注意图  $\hat{A}^i$  的最终输出元素的计算公式为

$$\hat{A}_{h,w}^i = \frac{e^{A_{h,w}^i}}{\sum_{h=1}^H \sum_{w=1}^W e^{A_{h,w}^i}}. \quad (3.3)$$

**空间加权总和：**现在，我们输出了  $N$  个注意图。如前所述，每个注意力图都为图像的特征提供了一组权重。我们用

$$Y^i = \sum_{h=1}^H \sum_{w=1}^W \hat{A}_{h,w}^i \cdot X_{h,w} \quad (3.4)$$

表示第  $i$  个用于无监督学习的加权后的特征。

### 3.2.1.2 第二轮分类训练

现在我们有了  $N$  个空间特征，接下来要关注的是如何训练我们的模型。在这里，我们将相同的注意力分类器应用于这  $N$  个空间特征，第二轮分类的结果应与第一轮一致。例如，在另一个关注分类器之后，通过标签 1 的注意力图生成的空间特征仍应属于标签 1。交叉熵损失用于训练模型。交叉熵损失的性质保证了每个标签都描述了不同的信息，因为在训练时，它将最大化一个滑窗属于一个类的可能性的同时抑制另外其属于另一个的可能性。结果，属于同一标签的帧之间的空间特征表达了同质信息，而不同的标签之间完全是不同的信息，实际上，这是我们之前展示的对齐操作。通常，我们将第一个特征点输出的目标设置为标签 1，将第二个特征点输出的目标设置为 2，依此类推。如果损失收敛到很小的值，则意味着该模型确实已经学到了一些不变的信息。因此，我们将损失定义为

$$\mathcal{L}_{SSN} = - \sum_{i=1}^N \log\left(\frac{e^{p_{i,i}}}{\sum_{j=1}^N e^{p_{i,j}}}\right). \quad (3.5)$$

其中  $p_{i,j}$  是  $i$  部分属于类别  $j$  的预测概率。我们希望分类器在时间和空间维度上具有一致性，这意味着由我们的注意力分类器计算的第  $i$  个注意力图生成的特征又可以归类为  $i$  相同关注度分类器的标签。这就是为什么我们在所有注意力分类器中均分权重，并将此训练方案称为独立训练。请注意，第二轮分类仅用于训练。

通过两轮分类和交叉熵损失，可以在无监督策略下训练 SSN。但是这种无监督的训练或多或少是“盲目的”。注意分类器由两次分类的一致信号来引导，而不管每一个具体的类别代表什么，无论是头还是手。本质上，它是利用局部像素值的相似性 (RGB 特征或高维深度特征)。在我们的情况下，网络可能会找出像素值相似的人体部位。但是，如果不指定要选择的部分，则具有相似像素值的部分可能会相互干扰。结果，无监督学习会找出处于不同位置但彼此相似的部分。在行人重识别<sup>[40-41]</sup>的上下文中，使用像素相似度已被广泛接受，因为像素值是最有区别的特征之一。我们实验证明了这种两轮分类无监督训练方案的功能如此强大，以至于它甚至可以从高斯分布生成的样本中选择不同的模式，但不擅长处理相似的人体部分。因此，我们修改了方法，以启用半监督训练以获得更好的性能。根本原因在于交叉熵本身依赖于局部像素值 (RGB 特征或高维特征)，而没有考虑其位置。在行人重识别<sup>[40-41]</sup>中，依赖像素相似性已被广泛应用。在我们的实验中，无监督策略下的结果也非常好。这就是为什么 SSN 很难区分左腿和右腿或膝盖，因为一个人的那些部分具有接近的像素值。对于无监督和半监督策略，通过标记数据，提供了像素值及其位置信息，从而提供了更强的约束力，并准确找到稳定的部分。

### 3.2.1.3 半监督培训

在上文中，通过两轮分类，我们能够在无监督策略下训练 SSN。但是，我们不知道每个空间特征代表什么，标签 1 还是标签 2 是人的手还是头。在这里，我们提出了半监督策略，以引导目标成为我们想要的部分。这个想法很简单，我们同时向网络提供标注和未标注的数据。标记的数据起着“锚定”的作用，这是一个显著的引导信号，可以指示注意力分类器选择我们想要的  $N$  个部分。我们会提供一些标记数据来引导，然后在无人监督的策略中对其进行培训。详细地说，我们将绕过 SSN 的某些功能混合到 SSN 产生的那些功能中。那就是用下面的公式替换公式3.4中的  $Y^i$

$$\mathbf{Y}^i = \mathbf{X}_{h_i, w_i} \circ \quad (3.6)$$

其中  $(h_i, w_i)$  是我们第  $i$  个目标的手动标记的关键点。因此，标记的数据和未标记的数据可以一起输入到网络。请注意， $\mathbf{Y}$  的大小为  $N \times 3 \times 3 \times C$ 。

在半监督训练中，使用此锚点，注意力分类器能够在接下来的无人监督阶段利用锚点的优势。这样，半监督 SSN 可以找出在一个帧内不同且在多个帧内稳定的人体部分。

### 3.2.2 三维卷积网络块

由于我们已通过 SSN 对齐了空间特征，因此三维卷积现在可以有效地提取时间信息。图3.1已证明，在将每个帧输入到 SSN 之后，已经生成了大小为  $3 \times 3 \times C$  的  $N$  个空间特征。然后，我们将在时间维度上属于同一标签的那些空间特征连接起来，以产生  $T \times 3 \times 3 \times C$  形状的特征(四维张量)(在本文中， $T = 4$ )。这些特征将被输入到三维卷积中以进行特征提取。请注意，由于三维卷积是不同的部分，因此它们不共享权重。在三维卷积网络将所有  $N$  个空间特征映射到一维向量后，我们将它们全部串联起来，并应用卷积网络为视频生成最终的特征。

### 3.2.3 训练和损失函数

#### 3.2.3.1 使用硬样本挖掘进行训练

与<sup>[42]</sup>使用的常规做法不同，后者将分类损失和三元损失相结合以进行时空表示学习。我们将以人的身份作为类别标签的分类损失丢弃，并用我们的 SSN 中的交叉熵损失代替。

难样本挖掘三元组<sup>[14]</sup>是人员行人重识别任务的一种常见做法，它被描述为

$$\mathcal{L}_{tri} = \sum_{i=1}^B [m + \max_{f_p \in S_i^+} \frac{\|f_i - f_p\|_2}{\sqrt{d}} - \min_{f_n \in S_i^-} \frac{\|f_i - f_n\|_2}{\sqrt{d}}]_+. \quad (3.7)$$

其中  $m$  是预定义的边距， $d$  是输出特征的维度， $f_i$  是第  $i$  个样本的视频特征，而  $[z]_+ = \max(0, z)$ 。 $S_i^+$  和  $S_i^-$  分别是第  $i$  个样本的正样本集和负样本集。

最终目标函数  $\mathcal{L}$  被公式化为 SSN 损失和三元组损失的加权和，即

$$\mathcal{L} = L_{tri} + \lambda \cdot L_{SSN}. \quad (3.8)$$

我们给 SSN 损失赋予系数  $\lambda$  的原因是，它收敛速度非常快，实验显示，需要把它控制在较小范围内，才可以更好的训练。

### 3.3 实验结果

在本节中，我们将首先在视频行人重识别任务上评估 SSN + 3D，然后在合成数据和真实数据上评估我们的核心模块 SSN。然后，我们将其与三维卷积网络相结合以完成人员行人重识别任务。在本节中，我们将首先在综合数据和真实数据上评估我们的 SSN。然后，我们将其与三维卷积网络相结合以完成人员行人重识别任务。

#### 3.3.1 实验设置

##### 3.3.1.1 数据集

我们在三个视频人物行人重识别数据集中评估了我们的方法。特别是 *iLIDS-VID* 包含 600 个视频序列，其中两个摄像机捕获了 300 个不同的人。视频序列的长度从 23 到 192 帧不等。*MARS* 具有 1,261 个人体 ID，并具有从 6 个摄像机捕获的 20,000 多个视频序列。边界框由 DPM 检测器<sup>[43]</sup>和 GMMCP 跟踪器<sup>[44]</sup>产生。*DukeMTMC-Video* 行人重识别是跟踪 *DukeMTMC-Video*<sup>[45]</sup> 基准的子集。行人图像每秒从视频中裁剪 12 帧，以生成视频小片段。

##### 3.3.1.2 评价指标

我们采用 Mean Average Precision (mAP)<sup>[46]</sup> 和 Cumulative Matching Characteristics<sup>[47]</sup> 作为实验中的衡量指标

### 3.3.1.3 实现细节

*CNN backbone*: 我们使用通过 ImageNet 预训练的 ResNet50 用于 CCN backbone 网络，来提取图片特征。请注意，每个 ResNet 模型共有 5 层，我们仅使用其中的前 4 层，它们会产生 1024 个通道的特征图。

**监督标签**: 对于每个数据集中的监督标签，我们使用 OpenPose<sup>[37]</sup>标记其头部，身体，胯部，左右手肘，膝盖和脚的位置。训练时，如果选择了一个视频片段，则还将选择其相应的带标签的图像。

**最终特征表示**: 3D CNN 的体系结构如图3.1所示。融合层将 9 个串联的特征点强制转换为 1024 维向量。

**训练和评估**: 在训练阶段，对于每个视频小轨迹，我们以 8 步的步幅随机采样 4 帧以形成视频剪辑。每批包含 8 个人，每个人具有 4 个视频片段。将所有输入图像调整为  $256 \times 128$  像素。使用权重衰减为 0.0005 的 Adam<sup>[48]</sup>来更新参数。该网络总共训练了 150 个周期，初始学习率为  $3 \times 10^{-4}$ 。50 个周期后，学习率降低，衰减率为 0.1。在测试阶段，每个视频小轨迹被分为多个 32 帧视频片段。然后，我们提取每个视频片段的特征表示，并使用它们的平均值来表示它们。

## 3.3.2 SSN+3D 分析

### 3.3.2.1 不同的学习策略

Learning Strategy	iLIDS-VID		MARS		DukeMTMC-Video	
	top-1	mAP	top-1	mAP	top-1	mAP
Supervised Learning	73.4	75.8	69.8	61.1	86.3	79.6
Unsupervised Learning	83.1	84.2	82.4	67.5	89.9	86.2
Semi-Supervised Learning	<b>88.9</b>	<b>89.2</b>	<b>90.1</b>	<b>86.2</b>	<b>96.8</b>	<b>96.3</b>

表 3.1 不同学习策略的影响。半监督学习效果最好，无监督学习效果其次。

有监督的学习时，我们无需使用注意力图来计算每个帧的空间特征，而是将标记的数据直接输入到以下三维卷积网络以进行特征提取。但是，无人监督的训练是让模型找到没有任何标签的独特人体部位。半监督学习是从每个视频轨迹中选择两个带有高质量标签的图像。训练方法在上一节中进行了描述，即当我们训练 SSN 时，来自标记图像的空间特征将直接输入到注意力分类器和三维卷积网络中。其余图像将以无监督的方式进行训练。

如表3.1所示，训练模型的最好方法是提供一些高质量的标签，然后让模型学习其余信息。这种半监督学习的方法指示我们的注意力分类器注意我们所需的身体部位。由于 OpenPose 模型可能在某些图像上得到错误的结果，我们认为这是监

督学习效果不如其他图像的原因。无监督学习的表现比有监督的好，但是由于缺乏引导，它仍然落后于半监督学习。我们的模型发现的不同部分可能有交集，而关键点的某些部分可能被忽略了。我们在 iLIDS-VID 和 MARS 上设置  $\lambda = 0.1$ ，在 DukeMTMC-Video 上设置  $\lambda = 0.05$ 。

如果无特殊说明，下文中的实验都采用半监督学习策略。

### 3.3.2.2 权重共享

Attention Classifiers	iLIDS-VID		MARS		DukeMTMC-Video	
	top-1	mAP	top-1	mAP	top-1	mAP
NonSharing Weights	69.4	73.2	72.3	70.9	84.9	71.2
Sharing Weights	<b>88.9</b>	<b>89.2</b>	<b>90.1</b>	<b>86.2</b>	<b>96.8</b>	<b>96.3</b>

表3.2 两轮分类是否共享权重的影响。共享权重效果比不共享高出近 20%，突出了共享权重的重要作用，也说明整个设计中两轮分类的重要性

在我们的设计中，单个框架的空间注意力分类器和多个框架的时间注意力分类器具有相同的权重。但是，我们可以放宽此约束，并探索它将为我们带来的结果。如果不共享注意力权重，则每个空间特征都是我们无法控制的方式，是图像不同部分的特定组合，我们也不知道每个空间特征真正代表了什么。表3.2已向我们证明了在空间和时间分类器之间进行这种权重共享的好处。如果不分配权重，则在训练过程中使用  $T$  个独立 SSN 损失。在运行实验时， $\lambda$  在 iLIDS-VID 和 MARS 上设置为 0.1，在 DukeMTMC-Video 上设置为 0.05。有权重共享和没有权重共享之间的较大余量暗示了 SSN 的影响。

### 3.3.2.3 $\lambda$ 参数选择

$\lambda$  是用于平衡 SSN 丢失的相对影响的参数。我们分别分析了  $\lambda$  对 iLIDS-VID，MARS 和 DukeMTMC-Video 数据集上的结果。我们观察到，当我们设置  $\lambda$  为相对较小时，例如，我们的方法可以获得最佳性能。0.1 或 0.05。这是因为 SSN 强大的学习能力可能会导致快速收敛，如果没有足够的数据表示，这将是有害的。我们在图3.3中展示了我们的结果。

### 3.3.3 与现有方法比较

表3.3, 3.4和3.5报告了我们的方法和其他最新方法在 iLIDS-VID, MARS 上的性能，和 DukeMTMC-Video 结果比较。这些表中的方法包含方方面面，从深度模

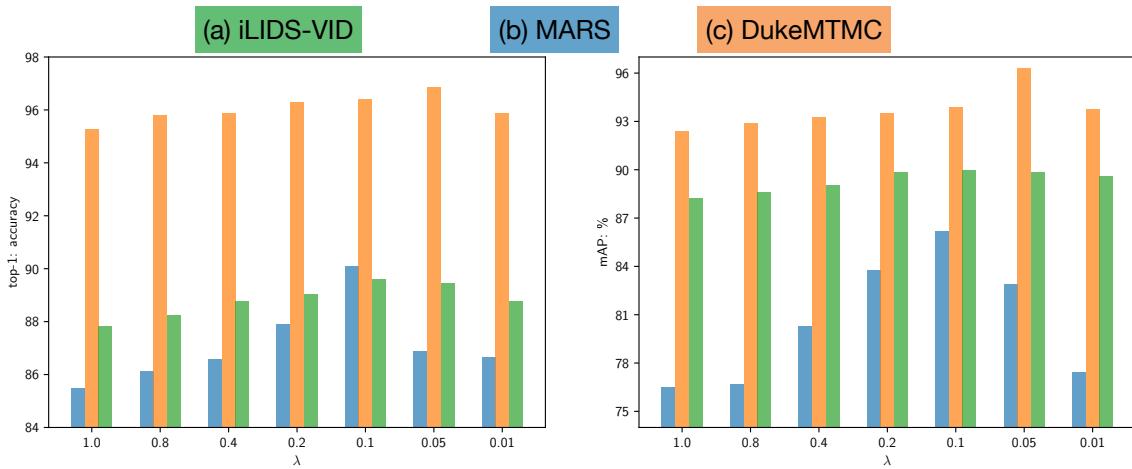


图 3.3 The top-1 and mAP on (a)iLIDS-VID, (b)MARS(b), and (c)DukeMTMC-Video

Methods	top-1	top-5	top-10
LFDA <sup>[49]</sup>	32.9	68.5	82.2
KISSME <sup>[50]</sup>	36.5	67.8	78.8
LADF <sup>[51]</sup>	39.0	76.8	89.0
STF3D <sup>[52]</sup>	44.3	71.7	83.7
TDL <sup>[53]</sup>	56.3	87.6	95.6
MARS <sup>[27]</sup>	53.0	81.4	-
SeeForest <sup>[54]</sup>	55.2	86.5	91.0
CNN+RNN <sup>[55]</sup>	58.0	84.0	91.0
Seq-Decision <sup>[56]</sup>	60.2	84.7	91.7
ASTPN <sup>[57]</sup>	62.0	86.0	94.0
QAN <sup>[58]</sup>	68.0	86.8	95.4
RQEN <sup>[59]</sup>	77.1	93.2	97.7
STAN <sup>[24]</sup>	80.2	-	-
Snippet <sup>[60]</sup>	79.8	91.8	-
Snippet+OF <sup>[60]</sup>	85.4	96.7	<b>98.8</b>
VRSTC <sup>[28]</sup>	83.4	95.5	97.7
AP3D <sup>[23]</sup>	86.7	-	-
<b>SSN3D</b>	<b>88.9</b>	<b>97.3</b>	<b>98.8</b>

表 3.3 iLIDS-VID 数据集比较

Methods	top-1	top-5	top-10	mAP
Mars <sup>[27]</sup>	68.3	82.6	89.4	49.3
SeeForest <sup>[54]</sup>	70.6	90.0	97.6	50.7
Seq-Decision <sup>[56]</sup>	71.2	85.7	91.8	-
Latent Parts <sup>[61]</sup>	71.8	86.6	93.0	56.1
QAN <sup>[58]</sup>	73.7	84.9	91.6	51.7
K-reciprocal <sup>[62]</sup>	73.9	-	-	68.5
RQEN <sup>[59]</sup>	77.8	88.8	94.3	71.7
TriNet <sup>[14]</sup>	79.8	91.3	-	67.7
EUG <sup>[63]</sup>	80.8	92.1	96.1	67.4
STAN <sup>[24]</sup>	82.3	-	-	65.8
Snippet <sup>[60]</sup>	81.2	92.1	-	69.4
Snippet+OF <sup>[60]</sup>	86.3	94.7	<b>98.2</b>	76.1
VRSTC <sup>[28]</sup>	88.5	96.5	97.4	82.3
AP3D <sup>[23]</sup>	<b>90.1</b>	-	-	85.1
SSN3D	<b>90.1</b>	<b>96.6</b>	98.0	<b>86.2</b>

表 3.4 MARS 数据集比较

Methods	top-1	top-5	top-10	mAP
EUG <sup>[63]</sup>	83.6	94.6	97.6	78.3
VRSTC <sup>[28]</sup>	95.0	<b>99.1</b>	<b>99.4</b>	93.5
AP3D <sup>[23]</sup>	96.3	-	-	95.6
SSN3D	<b>96.8</b>	98.6	<b>99.4</b>	<b>96.3</b>

表 3.5 DukeMTMC-Video 行人重识别数据集比较

型到传统模型皆有涵盖。在 iLIDS-VID 上，SSN3D 在 top-1, top-5 和 top-10 CMC 方面胜过其他公司。在 DukeMTMC-Video 上，SSN3D 在 top-1, top-10 和 mAP 方面的表现优于其他同类产品。在 MARS 上，SSN3D 还可以通过最先进的方法获得可比的分数。我们认为这种改进主要来自我们的 SSN + 3D 组合设计。SSN 和两轮分类是我们最重视的。因此，除了视频人行人重识别，我们希望 SSN 可以实际地应用于其他任务。

### 3.3.4 SSN 分析

在本小节中，我们将研究我们的核心设计 SSN。为了证明收敛能力，并证明我们的模型在无监督和半监督学习的训练下的结果，我们为每个任务准备以下三个数据集。考虑到数据集的不同性质，我们使用不同的模型设置来确保可视化结果。我们使用权重衰减为  $5 \times 10^{-4}$  的 Adam 优化器<sup>[48]</sup>来更新参数。请注意，我们仅评估不带三维卷积的 SSN。

#### 3.3.4.1 随机生成数据测试

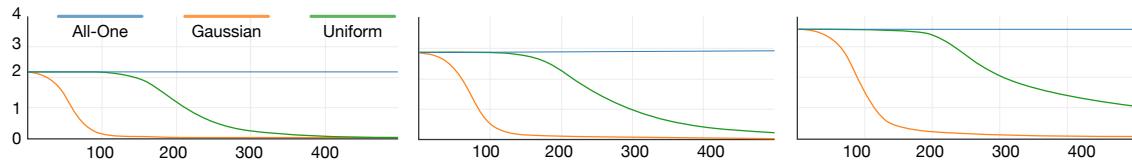


图 3.4 Strong Learning Ability. The x-axis is the training iterations, and the y-axis is a total loss

为了证明我们模型的强大学习能力，我们使用了由不同的随机分布（包括高斯分布和均匀分布）生成的数据。随机分布生成大小为  $128 \times 16 \times 16$  的随机样本，并且我们的滑动窗口大小设置为  $3 \times 3$ 。我们不使用任何卷积 backbone 进行特征提取，而是将样本直接输入到注意力分类器中。我们含 128 个样本的 batch 来训练模型。

如图3.4所示，我们为此数据集运行了 9 个实验。显而易见，在随机数据上运行的模型学习曲线使我们相信了模型的强大学习能力。即使随机生成数据，loss 也可以非常快地收敛到较小的值，尤其是当标签  $N$  的数量较小且数据更具动态性时（从正态分布生成的数据比从均匀分布生成的数据动态得多）。由于不可能从同类信息中发现不同的模式，因此，全一特征图的学习曲线不收敛是合理的。值得注意的另一件事是，当标签数量增加时，我们的模型似乎很难学习。我们认为这是因为很难在固定大小的特征图中发现更多不同的部分。总而言之，此随机数据实验揭示了我们模型的非凡学习能力。

### 3.3.4.2 Amber Abstract 抽象数据集测试

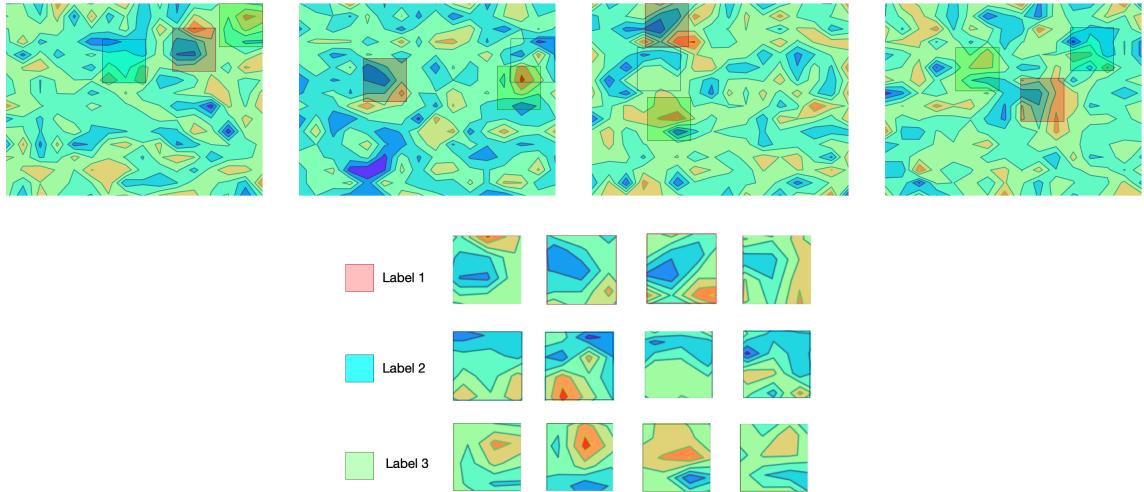


图 3.5 Performance of Unsupervised Learning. Our SSN model is able to spot similar parts between different paintings when appreciating our Amber Abstract dataset. Type 1 captures the pattern with a light blue protuberance on the left; type 2 spots the feature of an orange diamond above a blue stripe; and each type 3 contains a small deep blue block

为了探索通过无监督学习训练的模型的性能，我们将此数据集用于人体部位对齐任务。由于没有足够的现成数据集可用于此类挑战，因此我们以数字抽象表现主义的形式来生成此数据集。受到琥珀的形状和颜色的启发，我们以相同的风格生成了 128 幅不同的画。在创作这些作品时，我们无意在不同作品之间绘制相似的部分，因此共享区域没有“正确答案”，我们使用此数据集为读者提供直观直观的感觉，以了解我们作品的有效性通过无监督学习训练的 SSN，我们将此数据集称为 **Amber Abstract**。我们将原始绘画的大小调整为  $480 \times 640$  到  $120 \times 160$  的大小，我们没有使用任何卷积 backbone 进行特征提取。我们将绘画直接输入到注意力分类器中，该分类器将每个  $3 \times 19 \times 19$  滑动窗口映射到  $1 \times 9$  向量。batch 大小为 4。

我们在 **Amber Abstract** 数据集上得到的一些结果如图3.5所示。为了让读者更直接地了解我们的模型产生了什么，我们让模型找到对齐的部分，这些部分是来自不同随机创建的相似部分。在训练过程中没有标记的数据，并且我们的模型可以很容易地发现这些人体部位，如在不同颜色的边框中所证明的那样，而无需额外的代价。图??展示了更多的结果

### 3.3.4.3 Pedestrian 128 数据集测试

通过从互联网上收集的 128 个行人边界框，以  $60 \times 120$  的形式展示了半监督学习结果。我们手动标记了其中 16 个（并非全部）的身体关键点，包括头部，手，

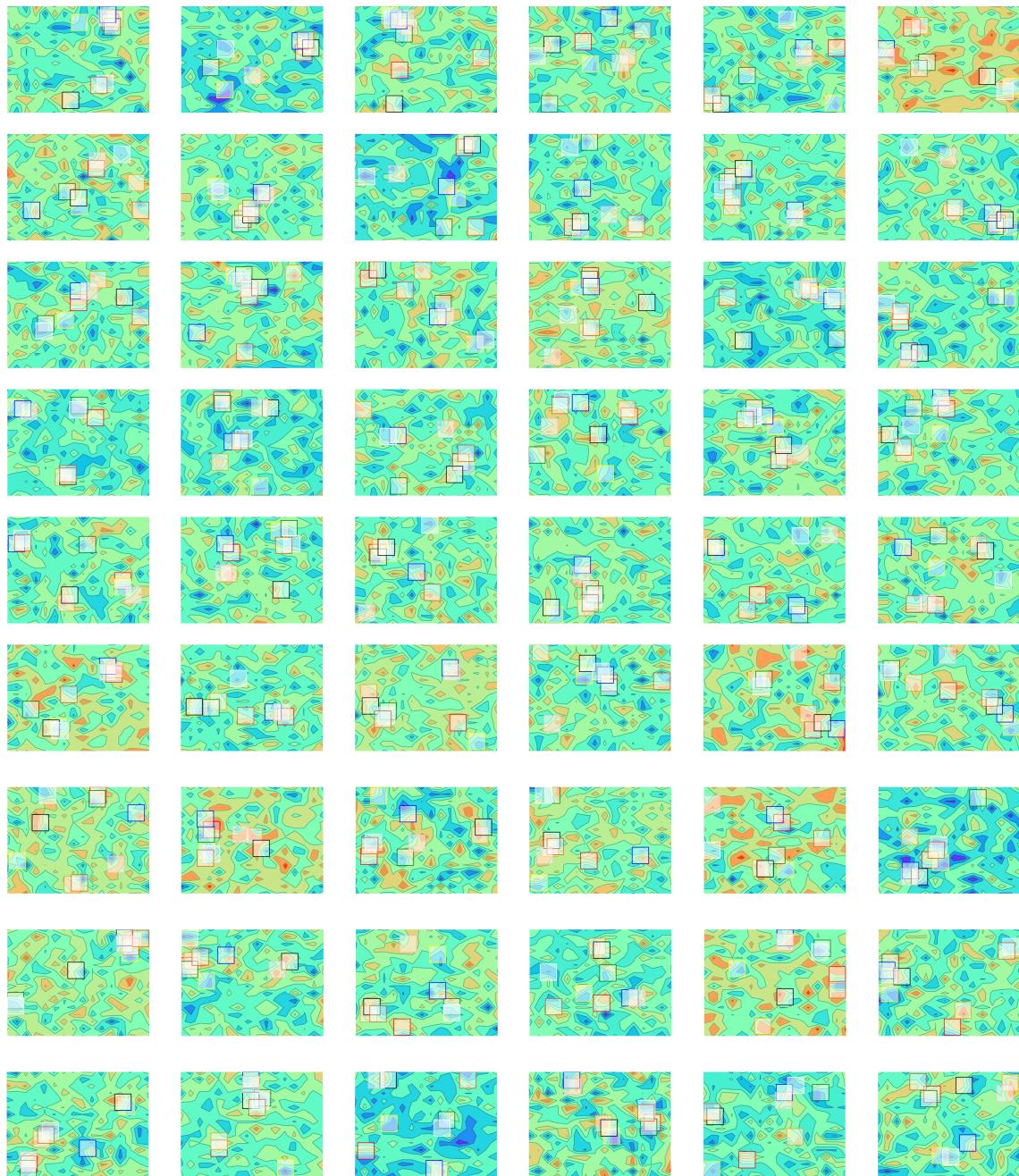


图 3.6 更多 Amber Abstract 上的可视化结果

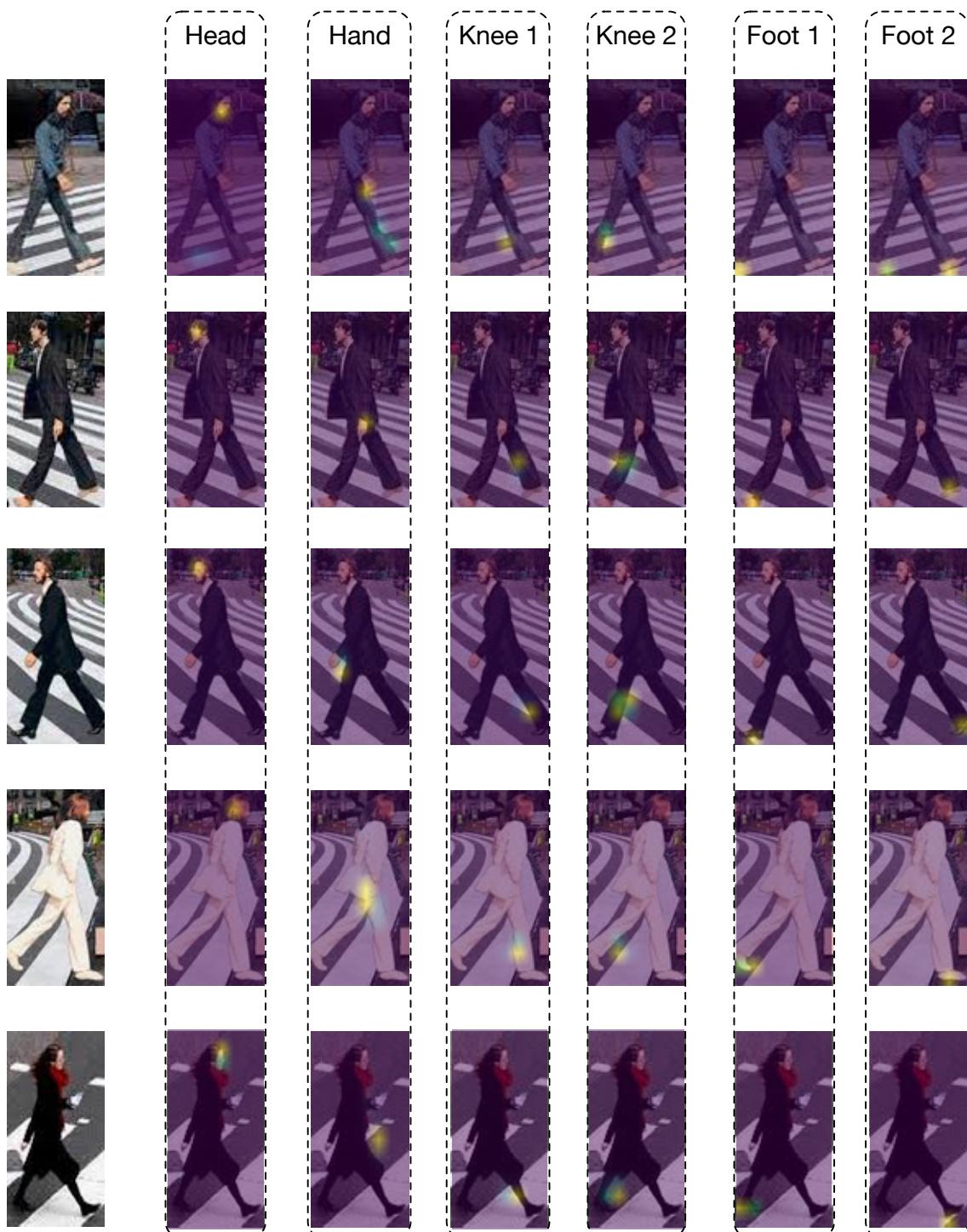


图 3.7 半监督学习在 Pedestrian 128 数据集上的结果。从图中可见人体关键点寻找基本准确。

膝盖和脚。我们使用 ResNet34 的 3 块为每个图像提取  $128 \times 8 \times 15$  特征图。滑动窗口的大小为  $3 \times 3$ 。batch 大小为 32。

我们随机选择要标记的 16 张图像，这意味着它们的空间特征不是通过空间加权总和来计算的，而是手动设置为图像的特定部分（例如，头，手或脚）。在训练过程中，我们使用标记的图像来引导 SSN 训练。之后，我们将继续使用未标记的数据训练 SSN。在图3.7中，我们展示了一些不同标签的注意图以及它们的原始图像。为了确保我们的模型可以发现我们想要的部分，我们使用了半监督学习策略。我们随机选择 16 张图像，并标记其身体关键点，包括头部，身体，手，膝盖和脚。在训练模型时，我们不会使用注意力图来生成3.2中提到的空间特征，而是将这些带有标签的部分及其相应的标签直接提供给我们的网络。考虑到我们的模型在人体关键点检测方面的出色表现，这种半监督学习模式将进一步应用于我们的人体行人重识别任务。

我们还会使用无人监督的策略来训练 SSN。如我们所料，无监督的 SSN 选择不同的部分，但是有些部分是人类无法理解的，如图3.8所示。而且，所选人体部位不如半监督策略中的稳定或准确。

我们发现两个问题：1) 选择了一些没有意义的部分，例如马路；2) 对于某些人体部分，某些选定的部分不稳定。

例如，SSN 可以仅区分左腿和右腿，并且当所选部分是膝盖时，根本原因是交叉熵本身依赖于像素值（RGB 特征或高维特征），而没有考虑其位置。在行人重识别<sup>[40-41]</sup>中，依赖像素相似性已被广泛接受。由于像素值是最重要的属性，因此效果很好。因此，我们在无监督策略下的结果也非常好。但是，如果两个部分相似，则无法分辨出差异。这就是为什么 SSN 很难区分左腿和右腿或膝盖，因为一个人的那些部分具有接近的像素值。对于无监督和半监督策略，通过标记数据，可以提供像素值及其位置信息，从而提供了更强的约束力，并准确地找到稳定的人体部位。

在这一部分中，我们从整个设计中取出 SSN，并通过三个数据集验证其稳健性。实验展示出其寻找不同部分的强大学习能力，在无监督策略中进行训练时的缺陷以及在半监督策略中的较好结果。我们的模型甚至可以收敛于随机生成的数据，这一事实证明了其强大的学习能力。此外，在 Amber Abstract 和 Pedestrian 128 上令人印象深刻的性能使我们对模型在无监督学习和半监督学习中的有效性有了直观的认识。

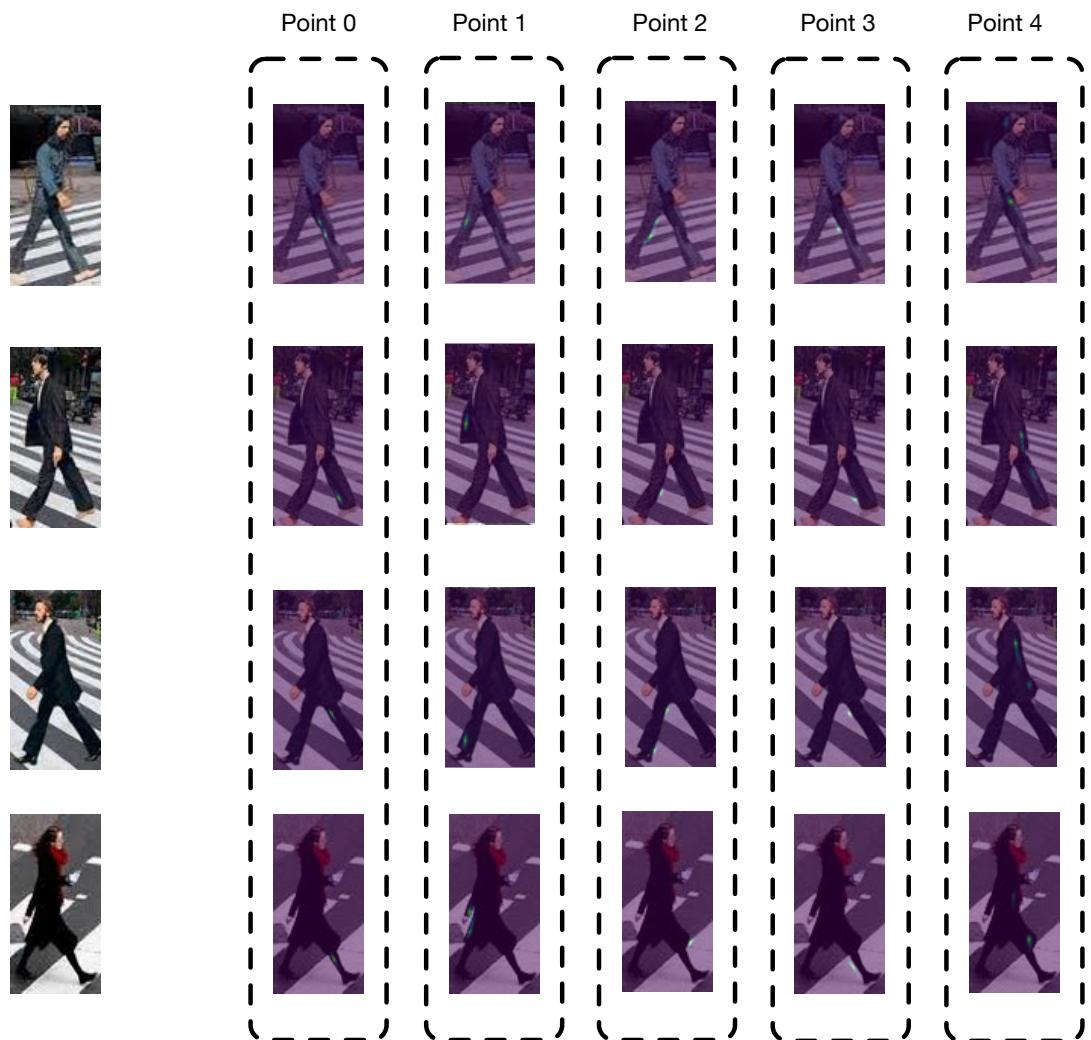


图 3.8 无监督学习在 Pedestrain 128 数据集上的结果。它会有一些瑕疵，比如说一些非人体部分学到，一些人体部分不稳定

### 3.4 本章总结

在本章中，我们提出了一种新颖的设计，称为 SSN3D。SSN 和两轮分类提供了一种以不同策略（当然还有不同表现）训练注意力网络的一般方法。SSN 可以从全图搜索，寻找关键点局部特征，可以进行对无监督、半监督训练。SSN 的主要优点是保证时空信息的可区分性和稳定性。基于三维卷积的聚合机制具有出色的能力来处理不同部分的时间变化。SSN 将为后续三维卷积网络提出一组对齐的空间特征，以进行视频特征提取。我们在 SSN 上的实验已向我们证明了其非凡的学习能力。

在实际应用中，整个系统在视频人行人重识别任务上取得了令人印象深刻的结果。在将来的工作中，我们期望在其他任务上看到更多此类对齐模型。此外，我们计划探索更灵活的半监督学习策略，例如带有较少或部分标记的数据。

## 第4章 基于时空行为模式的行人重识别优化

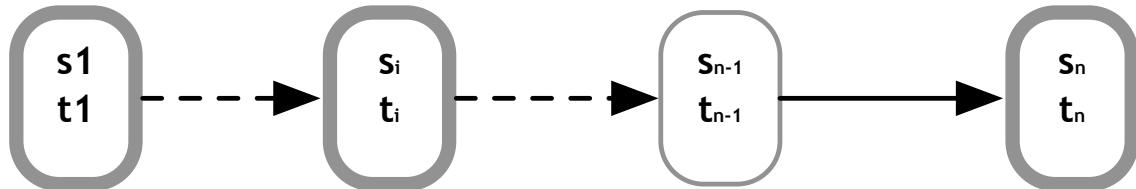
### 4.1 本章引言

在上一章，我们利用现场视频中的时空信息，提取到更有判别力的人体特征。本章我们利用行人长期的时空行为数据，建模时空模型，计算当下某行人出现在此时此地的概率。将这个概率与视觉相似度进行联合建模，从而得到更加准确的结果。

在行人重识别场景下，如文章<sup>[5]</sup>所描述，给出 probe 和 gallery 中 candidate 的时空约束即可。但是实际运行的系统中，如视频安防，商业 BI 中，往往都有历史数据积累，往往以天为单位处理数据，历史上大部分人都有历史记录，利用这些数据可以建立更加准确的模型；并且真实的数据一个人往往有若干个轨迹点，而不是单独一个。也就是说真实场景是实时增量聚类，而行人重识别是这个问题的子集。

在本章中我们贴近真实情况做一些假设：

- 系统中存有历史数据，该历史数据记录了大多数行人的行为数据；
- 人的行为数据是有序的时空状态，并且局部马尔科夫性质，即当前状态仅与前一个状态有关，如图4.1



$\{s_i\}$ 为空间位置序列， $\{t_i\}$ 为时间序列

图 4.1 行人的时空行为记录具备序列性，且满足马尔科夫性质

需要说明，本章方法需要与视觉相似度进行联合建模，人脸识别和人体重识别都能提供视觉相似度。因此本方法同样可以应用与人脸识别领域。人脸验证已在某些领域广泛使用，例如海关出入境，飞机登机，手机支付和地铁乘客入口，取得了非常好的效果。但是，在长时间的 COVID-2019 流行期间，公共卫生政策强烈建议在公共场所戴口罩，这给面部验证带来了新的挑战：蒙面人脸验证。由于遮盖住了鼻子、下巴、嘴等很多关键的特征，人脸识别的准确性下降了，我们通过本章描述的来减少人脸识别的错误，其技术原理和模型完全一样。我们从真实运行的

人脸识别的系统收集的数据进行分析，发现其中所有的错误记录里有 91% 的错误记录人脸带有口罩。

不失一般性，我们将人体或者人脸的特征相似度统称为视觉相似度。为了更加简洁和便于理解，我们将以地铁场景作为例子来描述我们的方案。地铁场景非常具有代表性，首先有进站和出站，对应着一般场景下人出现时间的连续性。其次，地铁场景下空间封闭，摄像头点位较多，我们能捕捉到更好的。为了衡量这个方法的优劣，因为缺少直接的时空数据，我们采集了一组地铁场景下行人刷脸的真实数据集。固然这不是行人重识别的数据，但是其数据本身准确性高，规模较大，而且场景真实，是具有代表性的。实际上，人体图片往往不是在人自由行走，而没有主动配合的情况下拍摄的，人体特征质量判别能力比人脸特征弱，基于贝叶斯概率的方法帮助更大。

相比于现有的利用时空信息的工作，我们的工作有一下贡献：

- 增加了模型推导，近似和求解部分，完善了其数学理论基础。
- 增加了时间和行人的维度，我们认为在不同时间段，每个人出现在不同地点的概率均是不同；文章<sup>[5]</sup>的结果是，是在我们的模型忽略不同时刻、不同行人的区别之后近似的结果。
- 通过建立了更一般的模型，增加了事件连续性的约束，我们的模型可以适用于更一般的场景，例如具有事件关联的人脸比对，人脸、人体聚类等。

乘客的行为可以根据其历史进展出站数据进行建模，进而结合时空行为概率和视觉分数来定义联合分数。则本章寻找最可能的乘客 ID 问题形式化定义为

$$h' = \arg \max_h = J(V(f, G_h), P(h, s, t)). \quad (4.1)$$

在以上定义中， $h$  表示乘客的 ID (human ID)， $J$  是具有给定视觉相似性评分 ( $V$ ) 和乘客时空行为概率评分 ( $P$ ) 的联合度量函数。 $f$  是输入的视觉特征， $G$  待搜索的基准特征底库， $G_h$  是  $G$  中 ID 为  $h$  的乘客的基准视觉特征 (底库特征)， $s$  (station) 是车站 ID(可以分类为进/出车站)， $t$  是进出站的时刻。真实数据显示，真实乘客的平均时空行为概率比其视觉最近邻乘客的平均时空行为概率大 300 左右。

$V(f, G_h)$  是视觉相似度，这一点已经在计算机视觉领域研究了多年，上一章计算出来的视觉特征就具备度量特性。在本章中，我们重点介绍  $P(h, s, t)$  和  $J(V, P)$ 。一般问题里，只计算出现的概率，但是来和往的概率不同，尤其是像这类时间序列的情况。进站和出站的时空行为概率相差很大。对于车站，我们可以给出给定乘客，车站和时间的概率。对于站外情况，其进出事件之间的时间间隔也限制了概率。

为了把上模的模型应用在真实系统中，还需要考虑更多的因素，包括离散参

数（时刻和时间开销）、处理没有乘车记录的乘客等。 $J(V, P)$  的定义应考虑罕见情况。即使  $J$  大而  $V$  为零，简单的  $V \cdot P$  也会导致零概率。

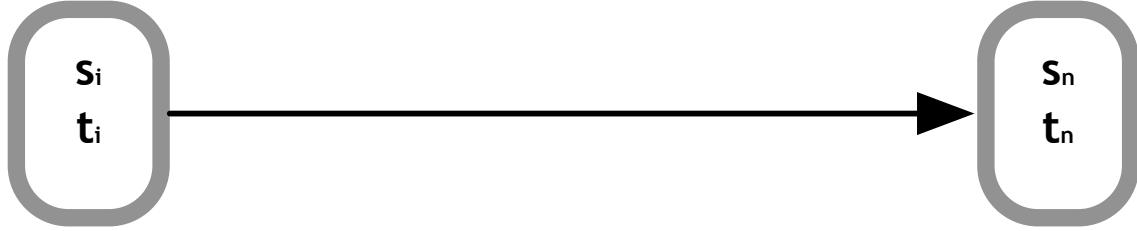


图 4.2 在不给定中间状态情况下，任意两个节点都具备马尔科夫性质

在具体应用中，我们首先对马尔科夫性质做出推导，在不给定中间状态下，任意两个节点都具备马尔科夫性质，如图所示4.2。以马尔科夫性质可知， $p(j|i), p(n|j)$  是确定值，则以公式4.2可知，在不给定中间状态  $j$  的情况下，状态  $n$  也只与此前已知的任意一个状态  $i$  有关。

$$p(n|i) = \sum_{j \in J} P(j|i) \cdot p(n|j) \quad (4.2)$$

## 4.2 进站模型

在进站场景里，我们使用  $i$  (in) 来替代  $s$  (station)，表示车站。公式4.1中的  $P(h, i, t)$  表示在站点  $i$  和时刻  $t$  中给定人类  $h$  的概率，这与我们的优化目标一致，即找到一个人类使关节得分最大化。因此，将  $P(h, s)$  定义为

$$P(h, i, t) = p_t(h|i) = \frac{p_t(h, i)}{p_t(i)} \quad (4.3)$$

上面的模型看起来很简单，但遗憾的是，在实践中如何这样处理，会出现荒谬的结果：最大的问题是对于那些以前从未在  $i$  站出现过的人，我们不应该将其设为 0。在这里，我们提供了一种平滑方法。假设有  $M$  名乘客。人类  $h_j$  的记录为  $x_{h_j}, j \in [1, M]$  记录，我们得到

$$p_t(h_j|i) = \frac{x_{h_j} + 1}{\sum_{j=1}^M x_{h_j} + M + 1} \quad (4.4)$$

我们使用  $h_0$  指代从未出现在车站的人，并且

$$p_t(h_0|i) = \frac{1}{\sum_{j=1}^M x_{h_j} + M + 1} \quad (4.5)$$

我们可以看到  $\sum_{j=0}^{j=M} p_t(h_j|i) = 1$ 。

### 4.3 出站模型

出站的情况下，与站内相比，我们有更多信息。我们可以得到相应人员的进站位置以及出入口之间的时间间隔。我们不能忽略时间  $t$ ，因为在这种情况下时间间隔变成一个约束。我们用  $o$  (out) 表示  $s$  (station)。因此，对于外站，我们将概率建模为

$$P(h, o, t) = p_t(h|o) \cdot p_t(\Delta t^h|i^h, o, h)。 \quad (4.6)$$

这里  $i^h$  是乘客  $h$  进站的车站 ID， $\Delta t^h$  是乘客  $h$  从  $i^h$  进站，从  $o$  出站之间的时间间隔。

实际上，在公式4.6中统计  $p_t(\Delta t^h|i^h, o, h)$  并不简单。假设在我们的方案中有 10 个车站，10 个耗时离散时长（将连续的时长简化成 10 个离散的时长），20 个时间片（将连续的时间简化成 20 个离散的时间片）。那么， $(i, o, t, \Delta t)$  的组合为  $10 \times 10 \times 10 \times 20$ 。为了测量  $p_t(\Delta t^h|i^h, o, h)$ 。为此，我们使用整个乘客记录进行估算，即

$$p_t(\Delta t^h|i^h, o, h) \approx p_t(\Delta t^h|i^h, o) \approx p(\Delta t^h|i^h, o)。 \quad (4.7)$$

值得注意的是，公式4.7这种近似忽略了人类的速度差异，把老年人与年轻人在车站之间的耗时等价了起来。当然如果数据积累足够，我们完全可以不做近似，构造准确的模型；但当数据量还不够的时候，可以先做近似。如果数据量较少可以做进一步近似，公式4.8忽略了不同时段两站之间的延迟，把早晚高峰期和空闲时段的延迟统一一起来看。

$$p_t(\Delta t^h|i^h, o) \approx p(\Delta t^h|i^h, o)。 \quad (4.8)$$

通过两步近似，我们获得站外的可能性如公式4.9，其中  $p(\Delta t^h|i^h, o)$  也正是论文<sup>[5]</sup>给出的时空行为概率结果。

$$P(h, o, t) = p_t(h|o) \cdot p(\Delta t^h|i^h, o)。 \quad (4.9)$$

### 4.4 联合建模

#### 4.4.1 离散和平滑

对于  $p(h|i)$ ，需要沿时间维度进行平滑。在这里，我们使用基于高斯的平滑<sup>[5]</sup>。对于  $p(h|i)$ ，假定时间  $t$  分为  $R$  个区间：

$$p_t(h|i) = \frac{1}{Z} \sum_r p_t(h|i) \cdot K(r-t), r \in [1, R], \quad (4.10)$$

这里的  $K(\cdot)$  是高斯核，即

$$K(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{\sigma^2}}, \quad (4.11)$$

$Z$  是：

$$Z = \sum_t p_t(h|i), t \in [1, R]. \quad (4.12)$$

对于  $p(\Delta t^h | i^h, o)$  的处理，与上面类似；对于  $P(\Delta t^h | i^h, o, t)$ ，由于存在两个时间维，我们需要一个相应的二维高斯平滑核，该核可以由两个一维高斯向量。假设时间  $t$  被拆分为  $R_1$  个时间片，成本时间  $\Delta t^h$  被拆分为  $R_2$  个时间长度。

$$P(\Delta t^h | i^h, o, t) = \frac{1}{Z} \sum P(\Delta t^h | i^h, o, t) \cdot K(r_1 - t, R_2 - \Delta t^h), r_1 \in [1, R_1], R_2 \in [1, R_2]. \quad (4.13)$$

#### 4.4.2 标准化

平滑后，我们可以通过简单地计算以下内容将分数归一化为  $[0,1]$  范围：

$$f(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)}. \quad (4.14)$$

#### 4.4.3 联合指标

我们将视觉分和时空行为概率分用于计算联合概率，得到

$$f(x; \lambda, \gamma) = \frac{1}{1 + \lambda e^{-\gamma x}}. \quad (4.15)$$

其中  $\lambda$  和  $\gamma$  是常数。 $\lambda$  是平滑因子，而  $\gamma$  是收缩因子。

对于进站情况， $J$  为

$$J(V, P) = f(v, \lambda_1, \gamma_1) \cdot f(p, \lambda_2, \gamma_2). \quad (4.16)$$

对于出站模型， $J$  为

$$J(V, P) = f(v, \lambda_1, \gamma_1) \cdot f(p, \lambda_2, \gamma_2) \cdot f(p, \lambda_3, \gamma_3). \quad (4.17)$$

上面的方法也用于<sup>[5]</sup>中，其中定义了联合度量函数。

### 4.5 优化和求解

参数包括：

变量名	含义
$N$	被选择并应用概率模型的候选数
$R_1$ for $p_t(h s)$	时间段数。定时点也要分割箱
$L_1$ for $p_t(h s)$	Parzen 窗口长度，它表示可以计算的邻域时间段总数，以进行平滑处理
$\sigma_1$ for $p_t(h s)$	Parzen 窗口平滑系数
$R_2$ for $P(\Delta t i, o, t)$	时间间隔的数量
$R_3$ for $P(\Delta t i, o, t)$	成本时间间隔区间的数量
$\Delta T_{max}$ for $P(\Delta t i, o, t)$	的最长间隔，如果时间成本长于此，则忽略它
$L_2$ for $p_t(h s)$	Parzen 窗口的时间长度
$L_3$ for $p_t(h s)$	耗时维度的 Parzen 窗口长度
$\sigma_2$ for $p_t(\Delta t i, o)$	Parzen 窗口平滑系数
$\lambda_1, \lambda_2, \lambda_3, \gamma_1, \gamma_2, \gamma_3$	联合概率的系数

表 4.1 概率优化方法参数表

给定许多地铁历史记录，鉴于以下优化目标，我们可以推断出最佳参数。假设一个查询特征  $q$ ，我们发现其对应人员的地面真实特征  $G_h$ （假设其 ID 为  $h$ ），以及与其最相似的邻居特征  $G_n$ （假设其 ID 为  $n$ ）。 $G_h$  和  $G_n$  是库中最接近查询  $q$  的前 2 个相似功能。我们将优化目标定义为

$$L_{joint} = J(V(q, G_h), P(h, s, t)) - J(V(q, G_n), P(n, s, t)) \quad (4.18)$$

为了摆脱过度拟合，我们添加了一个  $L_2$  参数归一化损失，最终损失为

$$L_{total} = L_{model} + L_2(\sigma_1, \sigma_2, \lambda_1, \lambda_2, \lambda_3, \gamma_1, \gamma_2, \gamma_3)。 \quad (4.19)$$

我们的目标是找到一组使损失函数最小的参数，即

$$\arg \max_{\sigma_1, \sigma_2, \lambda_1, \lambda_2, \lambda_3, \gamma_1, \gamma_2, \gamma_3} (L_{total})。 \quad (4.20)$$

我们收集了一个旅客记录数据集，我们使用 BFGS 来求解方程 4.20。

## 4.6 实验

在这部分，我们首先评估了利用用户历史乘车信息协助人体/人脸验证的可行性，接着我们分别评估了我们的核心模块，进站和出站两个模型在真实乘车数据上的结果。

### 4.6.1 实验配置

#### 4.6.1.1 数据集

我们从真实的人脸验证系统收集了时长跨度一个月的某地铁线路作为数据集。原始数据共 166842 条，去除三分钟内同乘客重复刷脸的记录 164894 条。每条记录包含用户 ID，进出站类型，站点号，进站时间，刷脸得分。一天中的时间分布从 05:00-23:30，共包含 8611 位乘客。

如图 4.3、图4.4和图4.5所示，乘坐公共交通的历史记录中，多数乘车记录是由具有明显乘车习惯的用户产生的。图4.4统计了用户当月乘车次数的占比，对该数据集进行统计分析，具有明显乘车习惯的用户数量占据多数，65% 用户月乘坐地铁 5 次以上，49% 用户乘坐 10 次以上，31% 用户乘坐 20 次以上，最多有乘坐 155 次的。图4.5统计了不同频次乘客的乘车次数在当月总乘车次数的占比。尽管只有 26% 的用户乘坐 25 次以上，但这部分用户贡献了 71% 的乘车记录，而乘车记录少于 5 次的用户虽然人数占总用户的 35%，乘车记录量只占总乘车记录的 5%。基于此，我们可以通过利用历史的用户的乘车习惯校正单次人脸识别的结果。

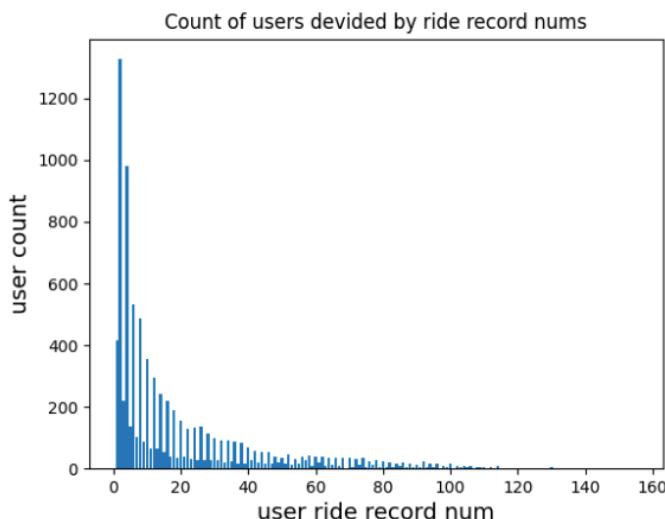


图 4.3 统计用户当月乘坐地铁次数，双数次比单数次多很多，往返均乘坐地铁的情况占多数

#### 4.6.1.2 异常数据分析

存在 108 条比对人脸特征分数存在异常的数据（存在其他人脸底图比第一次判定为 gt 的人脸底图分数更高），需要通过人工验证标注出正确的人脸。

如图4.7显示，108 张异常 query 图片中 98 张为戴口罩乘车，10 张为不戴口罩乘车，绝大部分异常数据是戴口罩时产生的，戴口罩带来的精度损失较为明显。

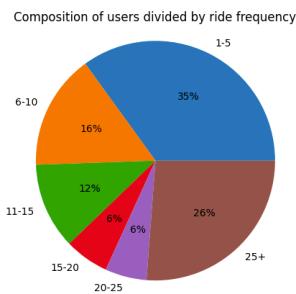


图 4.4 用户当月乘车次数的占比

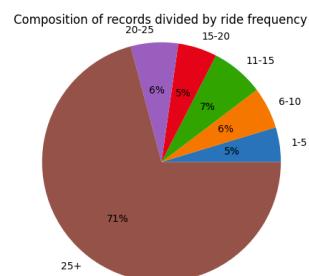


图 4.5 不同频次乘客的乘车次数在当月总乘车次数的占比

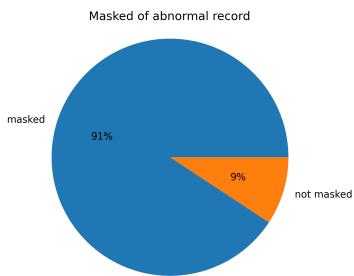


图 4.6 异常图片中戴口罩乘车比例

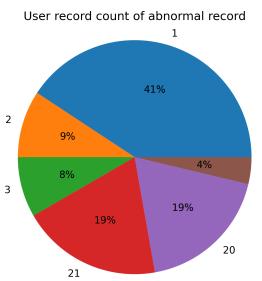


图 4.7 108 张异常数据中，有两位用户分别出现 21 次和 20 次，剩余的数据用户错误次数均在 5 次以下

## 4.6.2 实验细节

### 4.6.2.1 进站模型

按照用户 ID、站点 ID、时间频次格式处理历史记录，按照每三十分钟为一个间隔，一天中 05:00-23:30 将乘车时间分为 37 个时间区间。如用户拥有大于 k（当前设置 k=5）条乘车记录，则为该用户计算用户的历史时空分布。

将所有用户历史乘车数据逐条输入统计，构造三个维度的模型，三个维度分别是用户 ID、站点 ID、进站时间区间。用用户在当前站乘车记录除以所有用户在当前站总乘车记录代表该用户的当前乘车概率。

对该概率分布在时间维度和对空间维度做分布做高斯平滑，针对历史乘车次数  $\leq k$  的乘客，则用普遍这个时间段在这个站上车的概率代替他们的乘车概率，即历史所有乘客在该站点该时间段出现的次数除以历史所有乘客在该时间段乘车的次数。

存储所有用户的概率统计值，并在特征搜索阶段根据当前时间所处时间段，读取用户时空得分，与视觉分数计算获得融合分数。

#### 4.6.2.2 出站模型

按照用户进站站点 ID、出站站点 ID、进站时间、出站时间格式进行数据建模，需要将用户乘车记录配对，去掉单独的乘车记录。

将所有用户历史乘车数据逐条输入统计，构造三个维度的模型，站点 i，站点 j，从站点 i 到站点 j 花的时间长度（当前以 10 分钟为一时间区间）。用花了当前长度的时间从 i 站到 j 站的记录数除以从 i 站到 j 站的所有记录数代表该用户乘车概率。

由于时间区域有较强的相关性，则应该对相邻空间作出概率上的平滑。如果花了某时间长度从站点 i 到站点 j 乘车次数多，相邻时间区域内也可以平滑得到一个比较大的概率，反之则时间越远概率越低。存储所有概率统计值，在特征搜索阶段根据用户进站站点、出站站点、进出站间隔时间读取该用户的时空得分，与视觉分数计算获得融合分数。

#### 4.6.2.3 联合分数

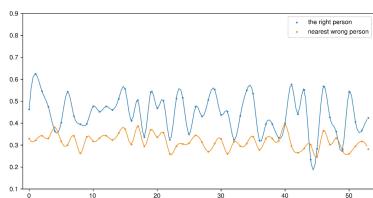


图 4.8 加入时空分数前

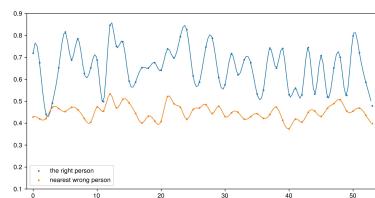


图 4.9 加入时空分数后

针对数据中单个用户数据分析联合分数带来的收益。将某用户的乘车记录汇总，得到的 54 条乘车记录可视化生成以上两张散点图，其中横坐标为乘车记录 id(0 53)，纵坐标为置信度 (0 1)。

图4.8为视觉特征输出概率，蓝色点为该记录人脸图片和该用户的底库图片的视觉相似度，黄色点为该记录人脸图片和除该用户以外的所有用户的底图中最像的一张图片的视觉相似度。在  $x=[4, 44]$  时，黄色点高于蓝色点，则存在除该真实用户以外的其他用户的视觉相似度更高，因而判断错误。

图4.9为视觉特征加入时空分布输出的联合概率：蓝色点为融合视觉和时空分布后的该用户和底库图片间的相似度，黄色点为融合视觉和时空分布后的该记录和除正确标签以外的所有记录中置信度最高的一条记录的相似度。所有蓝色点均高于黄色点，则所有记录均判断为正确标签，该用户的乘车记录没有误判。

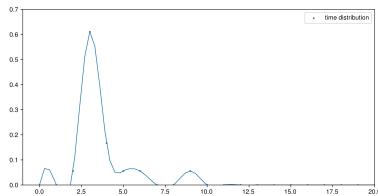


图 4.10 时间维度平滑前

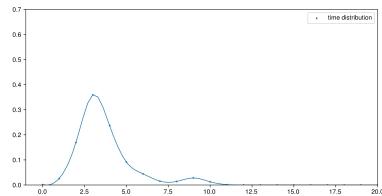


图 4.11 时间维度平滑后

#### 4.6.2.4 高斯平滑

选择单个站点 i 进, j 站点出的分布进行汇总, 横坐标为时间间隔, 每个单位的间隔为 10min, 生成图4.10。将统计分布值归一化后, 做平滑得到图4.11。经过高斯平滑后的时间概率分布更加稳定。

### 4.6.3 实验结论

#### 4.6.3.1 进站模型

总体上看, 对整体数据计算时空概率, 真实标签用户的时空分数是视觉上最接近搜索图的用户的时空分数的 2.79 倍, 证明时空分数可以作为视觉分数的辅助判断。联合分数由于加入了时空分数联合计算, 可以将真实标签用户与视觉分数最相近的用户的得分差距进一步拉开。

联合分数中出现两条与原始视觉分数判断出的结果不符的记录, 经人工研判, 其中一条数据正确标签为联合分数更高的记录, 而非原始仅凭视觉分数判断出的标签。人工标注数据中存在 44 条有效负样本, 其中通过联合分数可以减少其中的 16% 误判。

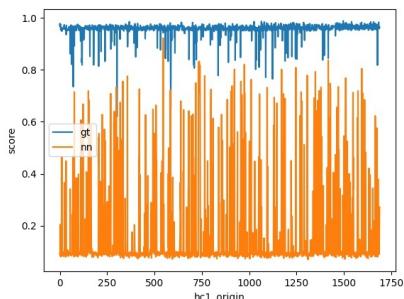


图 4.12 视觉分数

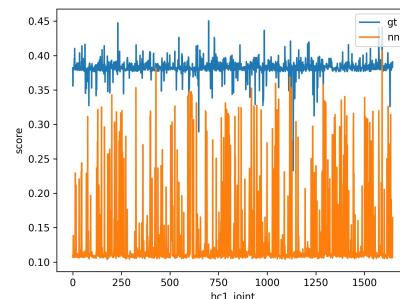


图 4.13 联合分数

#### 4.6.3.2 出站模型

对两小时内有进站记录的历史出站数据计算时空概率, 真实标签用户的时空分数是除真实用户外最相似用户的 2.7 倍, 证明时空分数可以作为视觉分数的辅助

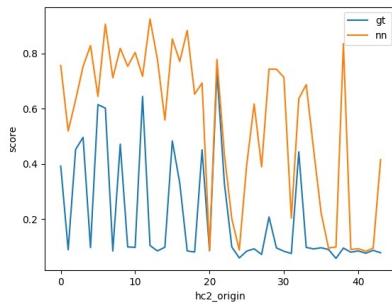


图 4.14 有效负样本的原始视觉分数

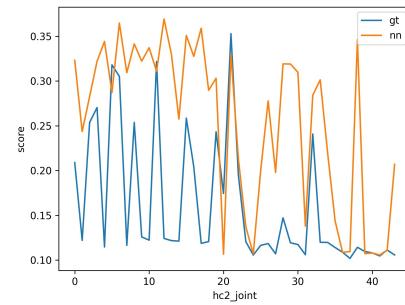


图 4.15 有效负样本的联合分数

判断。将出站模型统计分数加入联合分数，可以将真实标签用户与除真实用户外最相似用户的得分差距进一步拉开。

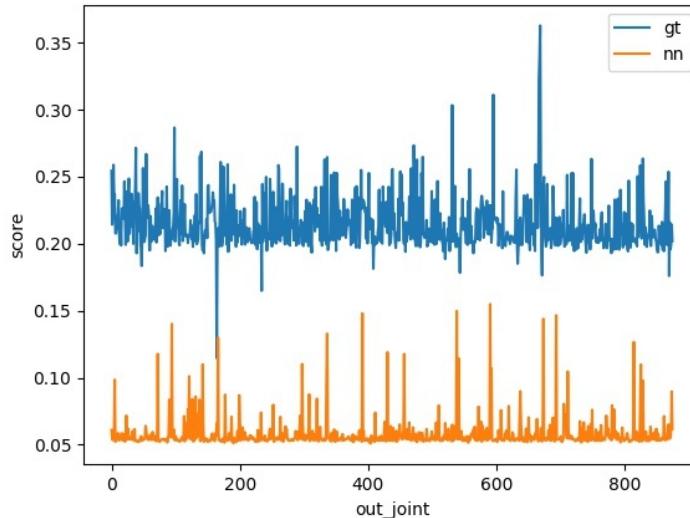


图 4.16 出站模型联合分数

## 4.7 贡献

本章的贡献包括三个方面：

- 增加了模型推导，近似和求解部分，完善了其数学理论基础。
- 增加了时间和行人的维度，我们认为在不同时间段，每个人出现在不同地点的概率均是不同；文章<sup>[5]</sup>的结果是，是在我们的模型忽略不同时刻、不同行人的区别之后近似的结果。
- 通过建立了更一般性的模型，增加了事件连续性的约束，我们的模型可以适用于更一般性的场景，例如具有事件关联的人脸比对，人脸、人体聚类等。

在应用层面，本章工作被工程化和产品化，产品层面已经应用部署。但在理论层面，目前的工作还不完整。

- 只考虑了前后进出两个事件，这是典型的马尔科夫过程的假设，但在真正的应用中，人的轨迹可能是多个时间地点共同构成，本文建模时忽略了这一点。
- 由于缺乏数据，本章的实验是通过地铁人脸比对的数据完成。这样处理的益处是在基本原理相同情况下做验证，不利之处是，人脸特征与人体特判别能力有较大差异，在人脸数据上取得的效果不能直接映射到人体数据上。所以本章现有实验只能验证原理和特性。

## 参考文献

- [1] Han J, Bhanu B. Individual recognition using gait energy image[J]. IEEE transactions on pattern analysis and machine intelligence, 2005, 28(2): 316-322.
- [2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
- [3] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.
- [4] Ji S, Xu W, Yang M, et al. 3d convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 221-231.
- [5] Wang G, Lai J, Huang P, et al. Spatial-temporal person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 33. 2019: 8933-8940.
- [6] Chen H, Wang Y, Shi Y, et al. Deep transfer learning for person re-identification[C]//2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM). IEEE, 2018: 1-5.
- [7] Lin Y, Zheng L, Zheng Z, et al. Improving person re-identification by attribute and identity learning[J]. Pattern Recognition, 2019, 95: 151-161.
- [8] Zheng L, Yang Y, Hauptmann A G. Person re-identification: Past, present and future[J]. arXiv preprint arXiv:1610.02984, 2016.
- [9] Matsukawa T, Suzuki E. Person re-identification using cnn features learned from combination of attributes[C]//2016 23rd international conference on pattern recognition (ICPR). IEEE, 2016: 2428-2433.
- [10] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[C]//European conference on computer vision. Springer, 2016: 791-808.
- [11] Liu H, Feng J, Qi M, et al. End-to-end comparative attention networks for person re-identification[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3492-3506.
- [12] Cheng D, Gong Y, Zhou S, et al. 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.
- [13] Chen W, Chen X, Zhang J, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 403-412.
- [14] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint arXiv:1703.07737, 2017.
- [15] Xiao Q, Luo H, Zhang C. Margin sample mining loss: A deep learning based method for person re-identification[J]. arXiv preprint arXiv:1710.00478, 2017.
- [16] Varior R R, Shuai B, Lu J, et al. A siamese long short-term memory architecture for human re-identification[C]//European conference on computer vision. Springer, 2016: 135-153.

- 
- [17] Zheng L, Huang Y, Lu H, et al. Pose-invariant embedding for deep person re-identification[J]. *IEEE Transactions on Image Processing*, 2019, 28(9): 4500-4509.
  - [18] Zhao H, Tian M, Sun S, et al. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1077-1085.
  - [19] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 480-496.
  - [20] Zhang Z, Lan C, Zeng W, et al. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10407-10416.
  - [21] Yan Y, Qin J, Chen J, et al. Learning multi-granular hypergraphs for video-based person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2899-2908.
  - [22] Yang J, Zheng W S, Yang Q, et al. Spatial-temporal graph convolutional network for video-based person re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3289-3299.
  - [23] Gu X, Chang H, Ma B, et al. Appearance-preserving 3d convolution for video-based person re-identification[C]//ECCV. 2020.
  - [24] Li S, Bak S, Carr P, et al. Diversity regularized spatiotemporal attention for video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 369-378.
  - [25] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.
  - [26] Wang T, Gong S, Zhu X, et al. Person re-identification by video ranking[C]//European conference on computer vision. Springer, 2014: 688-703.
  - [27] Zheng L, Bie Z, Sun Y, et al. Mars: A video benchmark for large-scale person re-identification [C]//European Conference on Computer Vision. Springer, 2016: 868-884.
  - [28] Hou R, Ma B, Chang H, et al. Vrstc: Occlusion-free video person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7183-7192.
  - [29] Hochreiter S, Schmidhuber J. Long short-term memory[J]. *Neural computation*, 1997, 9(8): 1735-1780.
  - [30] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4694-4702.
  - [31] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

- 
- [32] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset [C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.
  - [33] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks [C]//proceedings of the IEEE International Conference on Computer Vision. 2017: 5533-5541.
  - [34] Buades A, Coll B, Morel J M. A non-local algorithm for image denoising[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05): volume 2. IEEE, 2005: 60-65.
  - [35] Liao X, He L, Yang Z, et al. Video-based person re-identification via 3d convolutional networks and non-local attention[C]//Asian Conference on Computer Vision. Springer, 2018: 620-634.
  - [36] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
  - [37] Cao Z, Hidalgo G, Simon T, et al. Openpose: realtime multi-person 2d pose estimation using part affinity fields[J]. arXiv preprint arXiv:1812.08008, 2018.
  - [38] Liu H, Jie Z, Jayashree K, et al. Video-based person re-identification with accumulative motion context[J]. IEEE transactions on circuits and systems for video technology, 2017, 28(10): 2788-2802.
  - [39] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
  - [40] Shen Y, Xiao T, Li H, et al. End-to-end deep kronecker-product matching for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 6886-6895.
  - [41] He L, Wang Y, Liu W, et al. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 8450-8459.
  - [42] Wang G, Yuan Y, Chen X, et al. Learning discriminative features with multiple granularities for person re-identification[C]//Proceedings of the 26th ACM international conference on Multimedia. 2018: 274-282.
  - [43] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 32(9): 1627-1645.
  - [44] Dehghan A, Modiri Assari S, Shah M. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4091-4099.
  - [45] Ristani E, Solera F, Zou R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//European Conference on Computer Vision. Springer, 2016: 17-35.
  - [46] Zheng L, Shen L, Tian L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1116-1124.
  - [47] Bolle R M, Connell J H, Pankanti S, et al. The relation between the roc curve and the cmc [C]//Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05). IEEE, 2005: 15-20.

- 
- [48] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
  - [49] Pedagadi S, Orwell J, Velastin S, et al. Local fisher discriminant analysis for pedestrian re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3318-3325.
  - [50] Koestinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 2288-2295.
  - [51] Li Z, Chang S, Liang F, et al. Learning locally-adaptive decision functions for person verification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3610-3617.
  - [52] Liu K, Ma B, Zhang W, et al. A spatio-temporal appearance representation for video-based pedestrian re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 3810-3818.
  - [53] You J, Wu A, Li X, et al. Top-push video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1345-1353.
  - [54] Zhou Z, Huang Y, Wang W, et al. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4747-4756.
  - [55] McLaughlin N, Martinez del Rincon J, Miller P. Recurrent convolutional network for video-based person re-identification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1325-1334.
  - [56] Zhang J, Wang N, Zhang L. Multi-shot pedestrian re-identification via sequential decision making[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6781-6789.
  - [57] Xu S, Cheng Y, Gu K, et al. Jointly attentive spatial-temporal pooling networks for video-based person re-identification[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4733-4742.
  - [58] Liu Y, Yan J, Ouyang W. Quality aware network for set to set recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5790-5799.
  - [59] Song G, Leng B, Liu Y, et al. Region-based quality estimation network for large-scale person re-identification[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
  - [60] Chen D, Li H, Xiao T, et al. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1169-1178.
  - [61] Li D, Chen X, Zhang Z, et al. Learning deep context-aware features over body and latent parts for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 384-393.
  - [62] Zhong Z, Zheng L, Cao D, et al. Re-ranking person re-identification with k-reciprocal encoding [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1318-1327.

- [63] Wu Y, Lin Y, Dong X, et al. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5177-5186.

## 致 谢

“我走过很远的路，吃过很多苦，才将这份博士学位论文送到你的面前”，一位中科院博士在他的博士学位论文里写道。求知的道路往往不是平原走马，同是偏远农村学子，我的成长没有经历那么多不幸，但在求知这条道路上一样走过了很多路。

我第一次接触计算机的时候，就发现中国的网络不好，又贵又慢，而且连通性差；所以从本科的开始我就对网络情有独钟，想着怎样能网络变的更快。本科毕业论文写的是 IPv6 的部署规模测量，后来博士课题做的一种更革新的网络体系结构，命名数据网 (Named Data Networking, NDN) 部署中的挑战。我发现在网络领域，研究与革新中间隔着大江大河，但直到博士毕业，IPv6 和 NDN 都没有大规模应用起来。广域网——以我个人有限的理解——本质上是共识与合作，然而作为世界上最大的人工系统，想要取得的共识太难；更多的是资本与利益在博弈：本质上，这不是一个计算机领域的问题。

后来有幸参与了计算机视觉相关的项目，我推开了一个新的世界。这个世界里计算机有了眼睛和大脑，猫狗可以分类，人脸可以度量，电脑可以“代替”奥巴马做视频演讲。这是一个五光十色的世界，计算机通过像素去理解人类世界，有时候还创造出像素来欺骗人类的眼睛。这还是一个百花齐放，以点破面的领域，它允许我们就做一个应用点，把一个点做好就能帮助一些人。

但是进入这个领域，我需要补数学，补传统的计算视觉基础，补深度学习。本来就不宽裕的时间，变的更加紧张。连出差的时候，我都带着一本厚厚的《凸优化》，看不懂就强迫自己硬看，一遍看不懂就多看几遍。睡觉之前想着一个公式推导，越想越兴奋，推导想明白了，漫漫长夜却再也睡不着了。特别累的时候，也跟自己说，世界上让人快乐的事情除了读书和求知之外，还有很多；可是等有余闲的时候，发现让我快乐的依然是读书，而不是电视、游戏或者别的什么。

在深圳这样一个繁华大都市，在人生最有好奇心的几年里，我一个人度过很多冷清的周末和孤单的夜晚，终于站在了这里。为求知者喝彩，求知的灵魂就像是旷野里的大大小小的火炬，他们燃烧起来，或明或暗或强或弱，一起驱散黑暗照亮迷途；为那迷人的夜色和星空喝彩，那些人类无法触及，智力不足描述的神秘之地，好奇心将在那里的星辰大海种满玫瑰。

最后也是最重要的，衷心感谢我的合作导师乔宇老师、闫俊杰老师的指导和鼓励。两位老师在选题、研究、论文撰写、职业发展等方面及时指导，让我突破迷

## 致 谢

---

障。我此生的可能遗憾之一是，无法向两位老师一样成为追求真理的学者，那我就努力做一个终生学习者；如果能应用所学，做一点让世界变的更美好的事情也算不忘初心，殊途同归了。

## 个人简历

### 基本情况

蒋小可，男，湖北黄冈人，1987年2月20日出生，未婚。

### 教育状况

- 2006.9 - 2010.7，清华大学软件学院，获得学士学位，专业：计算机软件
- 2010.9 - 2016.6，清华大学计算机系，获得博士学位，以当年计算机系优秀博士生毕业，专业：计算机科学与技术

### 工作经历

2016.5 - 2018.10，深圳市看到科技高级算法工程师，主要负责 VR 全景视频的数据传输和播放。获得看到科技 2017 年内优秀员工。

### 研究兴趣

当前主要精力投入在计算机视觉，尤其是工业视觉，如表面缺陷检测相关研究方向。对计算机网络、未来互联网网络体系结构亦保持一定关注。

### 联系方式

- 通讯地址：广东省深圳市南山区南贸中心3栋405
- 邮编：518000
- E-mail: shock.jiang@gmail.com

## 发表文章目录

### 论文和专利

#### 学术论文：

- [1] Xiaoke Jiang, Qu Qiao, Junjie Yan, Qichen Li, Wanrong Zheng, Dapeng Chen, SSN3D: Self-Separated Network to Align Parts for 3D Convolution in Video Person Re-Identification, AAAI2021

#### 专利：

- [2] 蒋小可、丁思杰、鲍纪奎、季聪，一种基于人脸识别与人体关联的地铁逃票监测系统，申请号：202011529962.8，申请日：2020.12.22
- [3] 郑莞蓉、蒋小可、鲍纪奎、李启琛、季聪，基于历史客流大数据的轨道交通人脸识别乘车解决方案，申请号：202011132611.3，申请日：2020.10.21
- [4] 孙哲、郑莞蓉、蒋小可、姚兴华、季聪，一种隔栏递包行为检测解决方案，申请号：202110129505.8，申请日：2021.1.29
- [5] 蒋小可、邱达明、马睿、马志友，一种全景视频动态切流后的画面质量评估方法及装置，公开号：CN108833976B，申请日：2018.6.27，授权日：2020.1.24
- [6] 蒋小可、马睿、马志友，全景视频的画面质量显示方法及装置，公开号：CN108810513B，申请日：2018.6.27，授权公告日：2020.3.13
- [7] 蒋小可、王伦、蒋捷，全景画面生成方法及装置，公开号：CN10727474B，申请日：2017.6.30，授权日：2019.6.25