中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

面向行人重识别的多视角机器学习模型与算法研究

作者姓名：张志忠

指导教师：张文生 研究员

中国科学院自动化研究所

学位类别：工学博士

学科专业：模式识别与智能系统

培养单位：中国科学院自动化研究所

2020 年 6 月

Research on Multi-View Machine Learning Algorithms and Models for Person Re-identification

A dissertation submitted to the

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Pattern Recognition and Intelligent System

By

Zhizhong Zhang

Supervisor: Professor Zhang Wensheng

Institute of Automation, Chinese Academy of Sciences

June, 2020

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：张志忠

日　　期：2020年6月3日

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：张志忠　　导师签名：张文生

日　　期：2020年6月3日　　日　　期：2020年6月3日

摘 要

经济的快速发展带来了不同区域、不同城市间人员的大规模流动，也给公共安全带来了巨大的挑战。特别是，随着安防监控系统的普及，如何对海量的监控数据进行理解与分析，正逐渐成为智能化安防的核心。在这种背景下，行人重识别任务近年来受到了广泛的关注和应用。在给定行人图像的情况下，行人重识别方法能够快速检索出行人的跨视角图像，从而解决大规模监控网络下的行人识别与检索问题，在人物追踪、商场寻人以及反恐安全等方面有着重要应用前景，对于打造智慧城市、提升安防处置能力也有着巨大的科研和应用价值。然而，受限于视频监控探头的安装高度及密度，以及光照变化、行人姿态变化、遮挡、监控数据分辨率低等因素的影响，多视角场景下的目标锁定与查找仍然十分困难，这衍生出一个重要的机器学习问题，即如何对多视角数据进行有效利用，以解决目标对象、数据之间量化关系模糊的难题。

本文聚焦于多视角机器学习模型与算法在行人重识别中的应用，研究如何利用多视角数据中蕴含的一致和差异信息，构建适用于安防场景的相似性度量模型。论文从多视角特征融合、多视角非对称度量和多视角深度损失函数的角度，开展多源信息的关联研究，尝试解决真实安防场景下的行人查找与匹配问题。本文的创新性研究成果主要有:

• 提出了一种多线性多视角特征融合算法 (Multi-linear multi-view feature fusion, MMF)。针对多视角特征中互补信息难以捕获的问题，根据特征的固有特性，提出相似性作用矩阵，挖掘和传播多种特征之间的一致互补信息；通过样本依赖与视角依赖假设，探讨多线性结构与多视角数据中蕴含的一致信息的关系，提出多线性多视角融合算法，实现索引层级的特征融合，在降低内存开销的同时显著提升匹配精度；针对优化目标，提出一种高效的迭代优化求解算法，该求解算法具有较低的计算复杂度和理论收敛性保证。在 Market1501 行人重识别数据集和 Holidays、UKbench 等图像检索数据集上的实验表明，多视角特征融合算法能够有效提升原始特征的判别性，同时降低在线匹配的计算和内存开销。

• 提出了一种张量多视角非对称度量学习模型 (Tensor multi-task learning, t-TML)。针对行人重识别中，由视角差异导致的数据分布不一致问题，提出张量

多视角非对称度量学习框架，通过视角间和视角内的关联结构，学习非对称度量，对齐不同视角下的数据分布；提出无监督张量多视角度量学习模型，在不利用样本标签的情况下，能有效提升跨视角匹配精度，并运用多特征张量，灵活地融合多种视觉特征，有效地挖掘不同特征之间的互补信息。在 ViPeR、CUHK01、CUHK03、Market1501 等行人重识别公开数据集上进行了实验验证，结果表明所提方法的识别性能显著优于相关对比方法，所提出的无监督多视角模型、多特征张量模型，能够进一步提升行人重识别识别准确率。

• 提出了一种多视角深度对齐度量学习模型 (Wasserstein Triplet Loss, W-Triplet)。针对行人重识别中，目标存在偏差，不同视角下样本出现错位的情形，提出基于推土机距离的三元组损失函数，将原有的跨视角对齐问题转化为最优运输问题，通过对齐局部特征上的空间概率分布，运用正则化的推土机距离，解决样本错位问题；提出一种新的注意力机制，学习目标感兴趣区域，生成区域重要性离散概率，对最优运输问题提供监督指导；提出多分支深度网络模型，实现了全局和局部信息的融合，提升了识别准确率。在 CUHK03、Market1501、DukeMtMC-Reid、MSMT17 等多个行人重识别公开数据集上的实验表明，基于推土机距离的三元组损失能够帮助模型学习到目标的感兴趣区域，并依靠感兴趣区域，对齐和消除跨视角下的样本偏差，有效提升深度网络性能。

关键词：行人重识别，多视角机器学习，特征融合，度量学习，深度卷积神经网络

Abstract

The rapid economic development has brought a large-scale flow of people from cities to cites, and has also led to a huge challenge for public security. In particular, with the popularity of surveillance monitoring systems, how to understand and analyze these monitoring data has gradually become a core issue of intelligent security. In such context, person re-identification has aroused extensive attention in recent years. Given a pedestrian image, person re-identification technology can retrieve its cross-temporal and cross-scene images, and thereby solving the problem of pedestrian identification and retrieval under large-scale monitoring network, which has shown great potential for people tracking, shopping mall searching and anti-terrorism security. It also has tremendous scientific research and application value for building a smart city and enhancing security treatment capability. However, limited by the height and density of the monitoring camera, different illuminations, human poses, as well as occlusion and the low resolution of monitoring data, it is very difficult to match individuals across views. This raises an important machine learning problem, *i.e.*, how to effectively use multi-view data to solve the problem of estimating the quantitative relationship of data and objects.

This paper focuses on the multi-view machine learning models and algorithms with the application of person re-identification. It studies how to use the consistency and difference information contained in multi-view data to build a similarity measurement model which is suitable for security scenarios. From the perspective of multi-view feature fusion, multi-view asymmetric measurement and multi-view deep loss function, the thesis conducts research on multi-source information association and attempts to solve the problem of person searching and matching in real security scenarios. The main contributions of this paper are summarized as below:

- This paper proposes a new multi-linear multi-view feature fusion model (Multi-linear multi-view feature fusion, MMF). To mine complementary information from multi-view features, the model learns the functional matrix according to the properties of

features, and propagates similarities among multiple features. On this basis, the sample-dependence and view-dependence assumptions are used to capture the consistent information to achieve feature fusion on index level, which explores the relationship between multi-linear structure and consistent information contained in various feature representations. It reduces the memory cost while significantly improving matching accuracy. In addition, the proposed method offers an efficient solution algorithm, which has lower computational complexity and theoretical convergence guarantee. Experiments on person re-identification *e.g.*, Market1501 and image retrieval tasks *e.g.*, UKbench and Holidays illustrate that the multi-view feature fusion algorithm can effectively improve the discrimination of the original features, while reducing the computation and memory overhead in the on-line stage.

• This paper proposes a new tensor multi-view asymmetric metric learning model (Tensor multi-task learning, t-TML). To reduce the distribution discrepancy caused by the multi-view data, t-TML introduces a tensor multi-view framework by taking advantage of the correlations captured not only across different views but also within the view itself. It learns the asymmetric metric and enables the model to align the data distribution. On this basis, unsupervised tensor multi-view learning is proposed, which improves the identification accuracy without any supervision. The proposed model can also be easily incorporated with multiple visual features and explore their complementary information. Extensive evaluations on ViPeR, CUHK01, CUHK03, Market1501 Re-ID benchmark datasets confirm the effectiveness of the proposed t-TML model.

• This paper proposes a new multi-view alignment deep metric method (Wasserstein Triplet Loss, W-Triplet). To solve the mis-alignment issue, W-Triplet presents a new triplet loss based on Earth Mover Distance. By aligning the probability distribution in support of the local features, it transforms the cross-view alignment problem into an optimal transportation problem and utilizes a regularized Earth Mover Distance to mitigate the mis-alignment issue. Besides, a new attention mechanism is proposed, which can learn the object of interest, generate a discrete probability for local features, and provide supervision for the optimal transportation problems. A multi-branch deep network is also utilized to fuse global and local information, which improves recog-

nition accuracy. Experiments on CUHK03, Market1501, DukeMtMC-Reid, MSMT17 public Re-ID datasets show that the Earth Mover Distance based triplet loss can help the model distinguish the object of interests, such that it is able to align and eliminate the sample bias according to the salient areas. It also shows that the proposed method can effectively promote the performance of deep network.

Keywords: Person Re-identification, Multi-View Machine Learning, Feature Fusion, Metric Learning, Deep Convolutional Neural Network

目 录

图形列表

表格列表

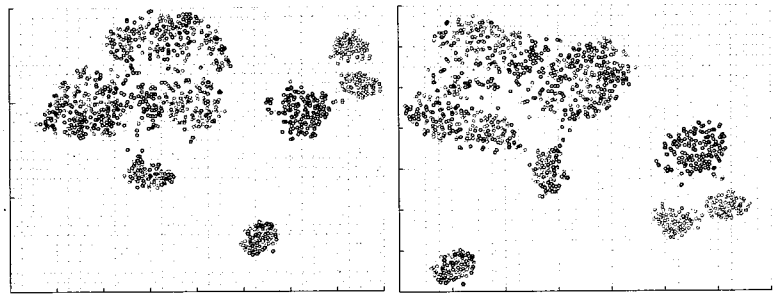
第 1 章 绪论

1.1 研究背景与意义

经济的快速发展带来了不同区域、不同城市间大规模人员的频繁流动，也给公共安全带来了巨大的挑战。在“十三五”期间，国家制定了以信息技术为支撑，打造平安城市、智慧城市的战略目标，给智能安防系统提出了新的业务需求。特别是，随着安防监控系统的普及，面对海量的监控数据，如何对其进行内容理解与分析，正逐渐成为智能化安防的核心问题。据行业研究机构 IHSMarkit 公司[1]估计，中国用于公共和私人的监控摄像头共有 1.76 亿个，庞大的监控网络，无时无刻不在监控着街道、商场、住宅等公共或私人场所，另一方面，面对海量的监控数据，如何自动地发现和锁定关键目标，识别出关键罪犯，也给智能化算法带来了前所未有的挑战。在这种背景下，基于计算机视觉的模式识别技术和机器学习算法得到了广泛的研究和应用。监控视频中的动作识别 [113]、行为分析 [99] 和行人重识别技术 [47]，正成为计算机视觉的热点问题，在安防领域有着广阔的应用前景。

行人作为监控场景中的重点研究对象，是智能安防体系中的重要一环。对行人的查找、识别和行为分析，对于公共安全态势感知和风险预警有着重要作用。作为行人分析的基础，行人重识别 (Person re-identification, Re-ID) 任务近年来受到了学术界和工业界的广泛关注。行人重识别是指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。在给定一个监控行人图像的情况下，行人重识别技术能够检索出该行人的跨场景图像，从而解决大规模监控网络下的行人识别与检索问题，对于打造平安城市、智慧城市，提升安防处置能力有着巨大的科研和应用价值。“重识别”的概念最早源自多目标行人跟踪，用于指代出现在当前摄像机下并重新返回的目标 [1]，随后，Gheissari 等人 [2] 正式提出行人重识别概念，旨在从大规模监控网络中检索出目标行人图像。有别于人脸识别 [131] 等生物识别技术，行人重识别任务不需要假定监控数据中的行人的姿态、分辨率等前提，在人物追踪、商场寻人以及反恐安全等方面有着重要应用前景。

[1]https://ihsmarkit.com



(a) t-SNE 投影结果　　(b) PCA 投影结果

图 1.1 行人重识别数据特征投影结果。

Figure 1.1 The visual results of dimension reduction with person re-identification feature.

对于行人重识别任务而言，其识别过程可以粗略地划分为两个步骤。首先，用一个固定长度的特征向量对行人图像进行特征描述，早期研究者们通过设计光照、旋转不变的手工特征，提取行人图像的纹理、颜色以及梯度信息，以应对极端的场景、视角变化；随后，研究人员对特征向量进行度量计算，借助度量学习 [79; 89],、排序学习 [72]、子空间学习 [61] 等方法，设计出合适的距离度量，来衡量行人图像之间的相似性，最终通过相似性的排序，检索得到目标行人图像。随着深度学习时代的来临，通过端到端的学习方式，特征和度量更为紧密的结合在了一起，依托于不同的网络结构 [54]、训练策略 [65] 和损失函数 [83]，伴随着训练数据规模的不断扩大，深度学习方法极大地提升了行人重识别任务的准确率，取得了里程碑式的成功。

然而，受限于视频监控探头的安装高度及密度，以及光照变化、行人姿态变化、遮挡、数据分辨率低等因素的影响，跨视角的目标锁定与查找仍然十分困难，这衍生出一个重要的机器学习问题，即如何对多视角数据进行有效利用，以解决目标对象、数据之间量化关系模糊的难题。所谓多视角信息是指，对于同一个样本语义对象，从不同层面、视角观察得到的数据，或从多个源头得到的不同维度的特征数据。多视角数据常常展现出特征高维、异构，语义一致但描述方式不同等特点。多视角学习的例子在信息时代随处可见。随着数据采集设备的发展、新采集手段以及新的数据特征提取方式的出现，针对同一描述对象，可以获得大量的产生于不同的数据源或特征子集的数据。以行人重识别任务为例，多视角数据既可以是提取到的纹理、颜色以及卷积神经网络 (Convolutional neural network, CNN) 等多种类型的具备不同表达能力行人图像特征 [86]，也可以是由不同视角

下摄像机采集得到的姿态、视角不同的行人图像 [71; 87]。如图1.1所示，不同颜色的数据点代表了不同视角下行人图像特征的二维投影，其中，每一个蓝色数据点都对应于一个绿色的数据点，二者来自同一个行人的不同视角图像。可见，虽然多视角数据是对同一个样本的数据描述，但不同视角下的数据分布仍然存在明显差异，致使目标对象之间的量化关系难以刻画。

多视角数据由于其特有的多态性、多描述性及多源性，使得传统的机器学习方法效果不佳。仍以行人重识别任务为例，早期的行人重识别方法 [74; 61] 并不考虑场景信息，而对所有视角下的数据，建立统一的度量模型。当面对视角、光照、人体姿态等急剧变化的情况时，跨视角的行人重识别精度往往不高，难以在实际场景中运用。为了解决多视角数据的学习问题，宾夕法尼亚大学 Yarowsky 等人提出了多视角学习 (Multi-view learning) 的概念 [3]。它的定义较为宽泛，只要学习任务所给定的经验数据由多个视角来表示，都称为多视角学习。多视角学习的一种朴素 (naive) 的方法是直接将所有视角特征强行拼接成一个长向量 (特征)，该方法符合将一个复杂问题归结为一个可解决的简单问题的思想，然而其效果往往不如人意。事实上，这种方法完全忽视了各个视角特征间的相互关联，不能够充分利用视角间的关联进行学习，其实质仍然是传统的单视角机器学习算法。

多视角学习的核心在于如何对多源信息进行有效利用。在行人重识别任务背景下，不同的特征表达构成了多源的多视角信息，这类多源信息是对同一对象的不同侧面描述。如何捕获不同视角表达下的一致互补信息，并借助互补信息提升识别准确率，是这类多视角机器学习问题的关键。另一方面，不同场景、视角下捕获的行人图像也构成了另一种多源的多视角信息，这类多源信息是从多个源头获取得到的数据，如何刻画与对齐不同视角下的差异信息，给出适合的跨视角度量，是这类多视角机器学习问题的核心。基于此，本论文以安防监控数据为应用场景，以行人重识别为具体任务，从多视角数据中蕴含的一致信息和差异信息入手，通过机器学习方法的研究，从多视角特征融合、多视角度量学习的角度，开展目标对象间的相似性度量的研究，实现多源信息有效关联，提升目标对象、数据之间匹配识别精度。

1.2 国内外研究现状

1.2.1 图像检索和行人重识别算法研究现状

行人重识别任务近年来受到了学术界和工业界的广泛关注，该任务可以被视作一种特殊的图像检索任务 (Content-based image retrieval, CBIR)，本小节围绕图像检索和行人重识别两大应用中的相关技术，进行简要评述。

图像检索是指根据图像内容，以图像语义特征为线索，从图像数据库中检索出具有相似特性图像的技术，在场景搜索、目标搜索、行人搜索、安防主题搜索等方面，有广阔应用前景。早期的图像检索方法通常运用图像的统计特性，如颜色直方图、灰度直方图等，将图像表示成单一的全局特征 [10; 23; 7]，然后根据全局特征的欧式距离，进行排序检索，受图像光照旋转的影响，这类方法虽然实现较为简单，但检索精度不高。在此之后，为了加强特征的判别能力，以尺度不变特征变换 (Scale-invariant feature transform, SIFT)[4] 为代表的具备旋转和尺度不变性，且对光照强度不敏感的局部特征逐渐引入到图像检索系统中，在这段时期内，汉明编码 [30]，负证据 [31] 和软指派 [16] 等词袋模型被相继提出，这些方法通过降低量化损失 (Quantization loss)，显著提升了词袋模型的表示能力，取得了巨大成功 [30; 29; 12]。与此同时，随着深度学习方法的兴起，其在特征表示方面的优势受到了学者们的广泛关注，一些工作尝试从预训练好的深度模型中提取特征。Babenko 等人 [15] 发现从全连接层 (Fully-connected layer) 提取到的特征对目标具有高阶语义描述性，以此为基础提取到的特征具有很强的语义信息。此外，一些研究工作运用卷积滤波器 (Filter) 的激活值 (Activation) 作为卷积特征 (Convolution feature)，由于滤波器的感受野有限，这些卷积特征具有类似局部特征的性质，并且也保留了深度特征的高阶语义特性 [32]。与此同时，一些研究人员开始尝试运用特征融合方法提升匹配的精度 [13; 8; 14]。其中 Chen 等人 [44] 提出了多索引融合的方法，通过提取不同特征表达中共有的领域结构，解决多特征融合问题。Zhou 等人提出了协同索引嵌入方法 [46]，利用交替索引更新的方案来融合多种特征。通过把特征加入到另一个相关特征上，强化邻域结构，提高检索精度。

将局部特征索引为词袋模型，以及将全局特征索引为紧致的哈希编码是索引方法中的两条重要分支。对于哈希索引来说，以随机多视角哈希 [39]、谱哈希 [40] 和非线性稀疏哈希 [50] 为代表的数据相关哈希方法，通过学习过程来产生

较短的哈希编码。相较于数据无关的哈希方法，这类方法更为便捷和高效。对于倒排索引来说，将其应用于图像检索的开创性工作来自于 [17]，它对一个视频片段中出现的目标进行了匹配。在此基础上，大量的工作集中于帮助倒排索引索引更多的细节信息。Zhou 等人 [33] 运用空间编码的方式，将局部特征的位置信息索引到倒排表中。Zhang 等人 [35] 将局部描述子和语义线索同时引入到倒排索引中。Chen 等人 [45] 利用方向和位置信息，学习得到一个背景敏感的码本。总体而言，图像检索任务侧重于判别性的特征设计和高效的索引结构，旨在解决大规模的图像匹配问题。

与之不同，行人重识别任务更注重判别性的学习。行人重识别是指通过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。在给定一个监控行人图像的情况下，行人重识别技术能够快速检索出行人的跨视角图像，从而解决跨摄像头、跨场景、跨时间下的行人识别与检索问题。随着人工智能技术的飞速发展，行人重识别技术正在公共安防领域大量应用，成为公安部门视频案件分析的又一重要武器。早期行人重识别方法的研究集中于特征设计 [64; 61] 和度量学习 [52; 84; 89; 74] 两大方面。研究者们通过设计光照、视角不变的手工特征，来应对极端的场景、视角变化，在这段时期内，包括行人轮廓 [2]、颜色 [64]、局部特征 [167] 在内的手工特征被相继提出。其中，利用行人的颜色信息 [64] 是行人重识别方法的重要手段之一，然而，背景信息、以及光照视角变化的影响，使得包括颜色特征在内的手工特征行人重识别准确率不高。为此，研究人员开始运用特征融合方式，提升特征的表示能力。SSM[86] 将 GOG 特征 [64] 和 LOMO 特征 [61] 拼接成一个长向量进行特征融合，DMVFL[96] 运用手工特征和深度学习之间的相互协作学习，有效提升匹配准确率。

除特征设计外，相似性度量的学习也广泛应用于行人重识别领域，以此来解决跨视角对象、数据之间量化关系模糊的问题。从广义上而言，这些方法借助度量学习 [79; 89],、排序学习 [72]、子空间学习 [61] 和深度学习 [60]，来提升跨视角匹配的准确度。这些方法可以进一步划分为视角特定 [87] 与视角无关的距离度量方法 [52; 61]。其中，代表性的视角无关方法包括 Metric Ensembles [72]、SCSP[55]、Null Space[52]、XQDA [61]、KISSME [74] 和 MFA [88]。举例来说，KISSME 从概率预测的角度，通过两个高斯分布拟合匹配与不匹配样本对，通过等价限制条件替代原始的标签监督，实现了大规模的度量学习。在此基础上，

XQDA 通过投影子空间的方式，最大化类间距离和类内距离比，取得了不错的识别准确率。总而言之，这类方法由于受到视角差异的影响，匹配精度有限。但是它们的计算复杂度较低，适用于大规模的行人重识别问题。与视角无关方法相对应，视角特定模型要么为每对视角，学习特定的匹配模型，要么为每个视角单独训练投影矩阵。前者的代表方法为 MtMCML[67]，它通过设计多个马氏距离度量，来与摄像网络相关联。而基于投影的方法包括 [87; 75; 71]。其中，Su 等人 [87] 提出了一个多任务学习框架来同时训练视角特定的分类器。Chen 等人 [75] 泛化马氏对称距离到非对称距离，克服场景变化带来的影响。总体而言，早期的行人重识别方法依靠手工设计的特征与度量学习，取得了一定的成功，但泛化性较差，难以在实际场景中进行运用。

2014 年以来，随着行人重识别数据库规模的不断扩大 [47]，基于深度学习的方法开始占据主导地位。与传统方法不同，深度学习方法通过端到端的训练方式，能够同时实现特征的提取和度量的学习，取得了里程碑式成功。在这个框架下，SVDnet[68]、MTDnet[58]、CAN[66] 以及其他卷积神经网络模型 [60] 被相继提出。这些方法在网络结构 [54]、训练策略 [65]、损失函数 [83] 等方面都有显著的区别，极大地提升了行人重识别准确率。总体而言，这些方法通常基于分类网络模型 [97]，使用分类损失 (Cross-entropy loss)[146; 47; 124; 82] 进行特征学习。在此基础上，三元组损失通过采样的方式，采集锚点、正样本点和负样本点，并迫使正样本对之间距离小于负样本对之间距离，提升特征的判别性，逐渐成为行人重识别方法的学习范式。Quadxruplet loss[150] 将三元组关系推广到四元组关系，以此来扩大类间距离、缩小类内差异。此外，许多深度学习模型 [144; 132; 124; 139] 还尝试运用局部特征来加强行人表示。PCB[124] 将网络的中间结果划分为不同的部位，每个部位利用交叉熵损失进行训练，显著提升了识别匹配的准确率。Yang 等人 [144] 提出了一种水平金字塔匹配方法，充分利用了不同池化操作和不同尺度提取到的局部信息，加强了特征的判别性。近些年，由于检测不准确、遮挡引起的行人错位问题，在行人重识别领域也得到了广泛的研究。PAN[159] 增加了一个基于注意力机制的仿射变换估计分支，来自适应地定位和对齐行人。DSA-reid[161] 通过密集的语义对齐解决局部特征错误匹配的问题。AAN[127] 通过学习逐块偏移和逐像素偏移，自动将行人图像从粗到细对齐。姿态估计类的方法 (Pose estimation)[142; 140] 通过姿态估计模块，减少姿态变化

对行人图像影响。Aligned Re-ID[132] 考虑局部特征的对齐问题，将原有的局部特征之间的匹配问题转化为最短路问题，从而提供更好的样本距离估计。虽然上述深度学习方法取得了里程碑式的成功，但是仍然缺乏多视角场景下的建模与分析，致使模型难以应对跨视角的样本偏差。

1.2.2 多视角特征融合学习研究现状

多视角学习起源于宾夕法尼亚大学 Yarowsky 等人的工作 [3]，一经发表就迅速引起了研究者们的关注，其核心思想是如何为多视角数据设计出能够利用视角间关联的学习算法。作为多视角学习的的重要组成部分，多视角特征融合算法受到了学者们的广泛研究，其目的是将多视角特征集成到一个单一的紧致的表示中，以此来提升特征的表示能力。代表性的方法主要包括以下两种：基于图的模型和基于神经网络的模型。从生成模型（基于图的模型）的角度来看，多视角特征融合学习问题可以归结为试图学习一组潜在随机变量的问题，这些变量代表了观测到的多视角数据分布。典型的方法包括多视角稀疏编码 [102]、多视角潜在空间马尔可夫网络 [108; 110] 和多视角深度玻尔兹曼机 [109]。多视角稀疏编码 (Multi-view sparse coding)[102; 103] 通过一组线性映射将共享的潜在表示与多视角数据关联起来，这种潜在表示被称为字典 (Dictionary)。这类方法能够选择合适的基，从而与多视角输入产生高度的相关性。特别是当不同视角的特征互补时，多视角稀疏表示将显著地提升性能。例如 Liu 等人 [103] 提出了将多视角稀疏编码和 Hessian 正则化相结合的多视角 Hessian 判别稀疏编码 (mHDSC)。mHDSC 利用基于 Hessian 正则化的数据局部几何特性，充分考虑了多视角数据的互补信息来提高学习性能。多视角稀疏编码已经成功应用于多个领域，包括人体姿态估计 [102]、图像分类 [104]、web 数据挖掘 [105] 以及跨媒体检索 [107; 106]。在无监督多视角特征融合的背景下，Zhang 等人 [8] 尝试在排序阶段融合不同的特征，构造了一种不同视角下的融合图结构，并通过在图上执行链接预测或计算最大权重子图，提升了图像检索的性能。Zheng 等人 [14] 提出了一种耦合的多索引结构，进行特征融合。Zheng 等人 [13] 提出了一种针对查询的得分层级的特征融合方法，通过估计额外数据集上特征分布，消除长尾分布带来的影响。

多视角卷积神经网络考虑在多个视角数据下学习卷积表示（特征），例如 3D 目标识别 [111]、视频动作识别 [113] 和跨视角的行人重识别 [112]，它试图从不同的角度综合有用的信息，以便为后续的下游任务提供更全面的表示。Su 等人

[111] 引入一种多视角 CNN 模型，它将一个目标的多个二维视图中的信息，通过视角池化集成到一个紧凑的表示中，再经由另一个网络得到目标的预测。其中，视角池化层是一个跨视角的逐元素最大化操作层，这种多视角机制的作用类似于“数据增强”，即在训练期间添加数据的变化，以学习翻转、平移和旋转等不变性。近些年，多视角特征融合网络也广泛应用于行人重识别领域，Ahmed 等人 [112] 引入了一个多视角特征融合层，该层计算输入的邻域差异，以捕获两个输入图像之间的局部关系。Wang 等人 [114] 将多视角网络融合表示问题作为一种跨图像表示学习问题，并提出了一种联合学习框架，将单图像表示和跨图像表示结合起来进行行人重识别。Wang 等人 [139] 探索局部特征和全局特征融合的多视角网络，该网络利用多分支网络结构，加强了行人图像的特征表示能力，有效提升行人重识别的性能。可见，多视角特征融合方法已成为下游任务中的常用手段，能够进一步提升模型性能。

1.2.3 多视角度量学习研究现状

多视角度量学习 (Multi-view metric learning) 是多视角学习中的重要分支，其目的是对齐多视角数据下的偏差，建立合适的度量模型，已广泛应用于图像分类 [100]、行为识别 [99]、3D 目标检索 [101] 和行人重识别 [98] 等领域。早期的研究人员通过构造不同视角下的投影模型，在投影空间获取度量模型。其中，典型相关性分析 (Canonical correlation analysis, CCA)[168] 是其中的代表方法，通过假设所有视角数据都产生于共同的潜在子空间，CCA 将多视角高维数据映射到低维子空间，从而有效的刻画多个视角之间的线性和非线性关联。Luo 等人 [169] 提出了基于张量的典型相关分析模型 (Tensor CCA)，该模型依据张量协方差理论，突破了传统 CCA 的限制可以有效处理任意多个视角数据。偏最小二乘法 (Partial least square, PLS)[115] 是一类用于建模观测变量之间关系的广泛方法，PLS 的基本假设是观测数据是由一个投影矩阵驱动产生的，PLS 通过最大化多视角投影矩阵的协方差，来创建正交投影，得到多视角度量。Li 等人 [116] 提出了 PLS 的最小二乘形式，称为跨模态因子分析 (CFA)。CFA 的目标是通过最小化投影后的多视角距离，找到最优正交投影，其优化问题的解与 PLS 一致。在此基础上，一些工作通过引入非线性变化和额外的监督信息，给出了 CFA 模型的相关扩展 [117; 118]。在行人重识别领域，Yu 等人 [71] 提出了无监督的多视角度量学习方法，通过投影矩阵建立多视角公共空间，并针对视角间和视角内进行了判别性学

习，成功实现了无监督行人重识别。Hu 等人 [98] 提出了一种共享、私有的多视角度量方法 (MvML)，MvML 不仅学习每个视角下的距离度量以保留其私有属性，还学习统一的共享度量以保留公共属性，有效提升了行人重识别准确率。

多视角度量学习的研究还集中于跨模态搜索任务上，Bai 等人 [119] 提出了有监督语义索引 (SSI) 模型，该模型定义了一系列非线性变化，这些非线性变化通过判别性学习，用于将多模态输入对映射为排序得分。实际上，这种非线性变化是一种马氏距离，可以通过排序间隔损失进行优化。此外，为了利用基于核分类器的在线学习方法优势，Grangie 等人 [120] 提出了一种判别性的跨模态排序模型，称为被动-主动图像检索模型 (PAMIR)，该模型不仅采用与最终检索性能相关的学习准则，同时也考虑不同的核分类器。为了解决跨模态度量问题，Frome 等人 [121] 提出了一种深度语义嵌入模型 (DeViSE)，该模型通过跨模态映射将两个深度神经网络连接起来。受到 DeViSE 设计的启发，Norouzi 等人 [122] 提出了一种将图像映射到连续语义嵌入空间的语义嵌入模型 (ConSE)。与 DeViSE 不同，ConSE 模型保持了卷积网络以及 softmax 层的完整性。此外，Fang 等人 [123] 提出一个深度多模态相似性度量模型，该模型学习两个神经网络，将图像和文本片段映射到一个通用的空间进行矢量表示。Dai 等人 [166] 提出了一种跨模态三元组损失，用于推近跨模态同类样本，推远跨模态不同类样本，有效提升了可见光-红外行人重识别算法的识别准确率。

1.3 研究内容与主要贡献

目标对象、数据之间的量化关系模糊，是当前安全态势感知与预警的主要困难，究其根本在于难以对多源信息进行有效利用。为此，本文从多视角特征融合、多视角度量学习的角度，开展目标对象间的相似性度量研究，实现多源信息的有效关联，提升行人重识别精度。

对本文研究的主要难点总结如下：

• 在大规模图像检索和行人重识别任务中。检索模型需要在保证检索精度的同时，还要尽量降低在线的检索时间和内存开销。虽然运用多种特征，提升检索精度，已经成为图像检索类任务的范式，但一方面，不同特征的维度、特性以及建立索引的方式，使得特征层级的特征融合十分困难，另一方面，多特征融合也不可避免的带来了额外的计算和存储开销。因此，如何设计计算开销较小的多视

角特征融合算法是需要解决的首要问题。此外，多视角特征侧重不一，具有不同的表示能力，如何捕获和利用其中的互补信息，提升大规模图像检索和行人重识别任务的精度，也是亟待解决的难题。

• 在行人重识别任务中，由于数据获取的视角差异，使得行人图像处于不同的视角、不同的光照和不同行人姿态下，致使多视角数据分布不一致的情形发生，给跨视角的识别和匹配带来了巨大困难。此外，当视角急剧变化时，匹配的训练样本难以获取，训练数据有限，如何利用视角内部以及视角间的关联关系，提升模型的泛化性，是行人重识别度量学习的一大重要问题。

• 典型的行人重识别任务是针对图像数据而言的，从不同视角下的摄像机拍摄获取的监控数据，经由行人检测器或人为标注得到行人图像，由于检测的不准确，不同视角下的行人姿态变化，遮挡等因素的影响，致使跨视角下的样本偏差，不同视角下样本出现错位的情形。此外，从摄像机捕捉到的监控数据通常分辨率较低，难以捕获行人面部、体态信息，也给重识别任务带来了额外的困难。

论文主要贡献

本论文的主要贡献可以归纳为以下几个方面：

• 针对图像检索和行人重识别任务中，难以捕获多种特征之间互补信息的问题，提出了多线性多视角特征融合模型 (MMF)。主要创新点如下：1）根据不同特征的固有性质，学习相似性作用矩阵，传播多种特征之间的互补信息。2）提出多视角特征融合模型，通过样本依赖与视角依赖假设，捕获多种特征之间蕴含的一致信息，实现索引层级的特征融合，在降低内存开销的同时显著提升匹配精度。3）提出一种高效的求解算法，具有较低的计算复杂度和理论收敛性保证。在行人重识别和目标匹配上的实验表明，多视角特征融合算法能够有效提升原始特征的判别性，同时降低在线匹配的计算和内存开销。

• 针对行人重识别中，由于数据获取的视角差异，导致的多视角数据分布不一致问题，提出了张量多视角非对称度量学习方法 (t-MTL)。主要创新点如下：1）提出张量多视角非对称度量学习框架，通过视角间和视角内的关联结构，学习非对称度量对齐不同视角下的数据分布；2）针对所提方法目标函数的非光滑特性，设计了一种高效求解算法，在保证收敛性的同时快速求解优化问题；3）提出无监督张量多视角度量学习模型，可以作为一种新的无监督基准方法，提升多视角数据的判别性；4）提出多特征张量，能够灵活地融合多种视觉特征，有效

地挖掘不同特征之间的互补信息。在行人重识别的四个公共数据上进行了实验验证，实验结果表明所提方法的识别性能显著优于对比方法，所提的无监督多视角模型、多特征张量多视角模型，能够进一步提升行人重识别特征判别能力。

• 针对行人重识别中，目标存在偏差，不同视角下样本出现错位的情形，提出了多视角深度对齐度量学习方法 (W-Triplet)。主要创新点如下：1）提出基于推土机距离的三元组损失，将原有的对齐问题转化为最优运输问题，通过对齐局部特征上的概率分布，解决样本出现错位的情形；2）提出新的注意力机制，能够学习目标感兴趣区域，生成区域重要性离散概率，对最优运输问题提供监督指导；3）提出多分支深度网络，融合全局和局部信息，提升识别准确率。在行人重识别四个公开数据集上的实验表明，所提方法显著优于对比方法，通过自对比实验表明，基于推土机距离的三元组损失能够帮助模型学习到目标的感兴趣区域，并依靠感兴趣区域对齐和消除跨视角带来的样本偏差，有效提升深度网络性能。

1.4 课题来源与论文结构

论文的课题来源

本论文研究围绕多视角机器学习理论与方法，研究内容来源于中国通用技术研究院、公安部第三研究所等安防实际业务需求。

本论文研究主要得到以下课题的支持：

• 国家自然科学基金重点项目 (U1636220)，面向大数据的国际特定主题事件推演与风险预警研究；

• 国家重点研发计划 (2017YFC0803703)，警务大数据智能技术应用研究；

• 国家自然科学基金面上项目 (61772524)，大数据多视图子空间非监督机器学习理论与方法；

论文的结构安排

本论文的章节关系如图1.2所示。第 1 章首先介绍了安防场景下行人重识别任务的背景与意义，阐述了其中存在的多视角机器学习问题，并对涉及到的行人重识别方法、多视角特征融合方法、多视角度量学习方法进行简要回顾。在此基础上，凝练出三个行人重识别任务中存在的多视角机器学习关键难点，最后总结了本文的主要贡献。其余各章节的内容安排如下：

第 2 章为多线性多视角特征融合模型：首先，对安防场景下的大规模图像检

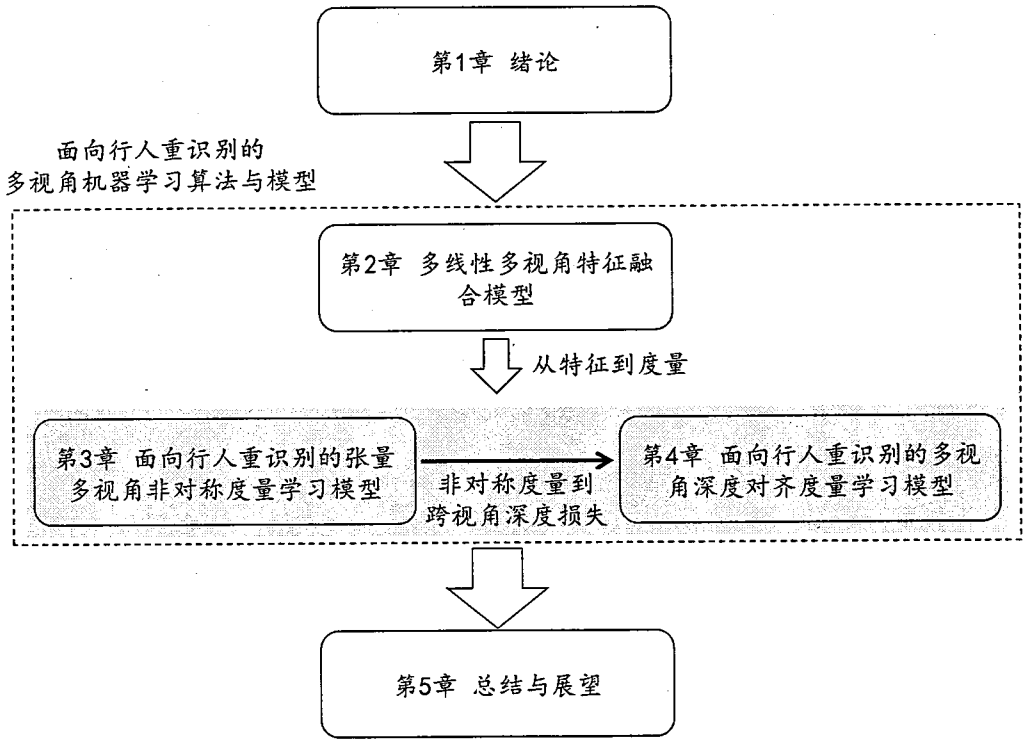


图 1.2 全文章节关系图。

Figure 1.2 The chapter diagram of this dissertation.

索和行人重识别任务进行简要介绍，明确了多特征融合作为多视角学习的重要问题之一，能够有效提升检索准确率，阐述本方法的研究动机在于捕获多视角数据中的一致和互补信息。然后介绍了本方法的相关工作，包括图像检索和行人重识别的特征表示方法、大规模图像检索索引结构和多特征融合方法。接着，详细介绍了多线性多视角特征融合模型，包括模型框架，模型的形式化表达，模型的优化方法和在线匹配流程。最后，在图像检索和行人重识别应用上进行了实验，实验结果表明所提方法不仅匹配精度显著优于其他对比方法，还降低内存开销；通过将所提方法进行分解自比较，证实了各创新点的有效性与必要性。

第 3 章为面向行人重识别的张量多视角非对称度量学习模型：首先，对安防场景中，行人重识别任务的跨视角特性进行了简要介绍，明确了不同视角下数据存在分布不一致的情形，阐述本方法的研究动机在于利用非对称度量，通过视角之间和视角内部的关联结构，对齐多视角数据。然后介绍了本方法的相关工作，包括行人重识别特征设计、行人重识别度量学习和行人重识别无监督学习。接着，详细介绍了张量多视角非对称度量学习模型，包括张量多视角学习框架、优

化方法、无监督非对称度量学习和多特征融合多视角度量学习。最后，在行人重识别的四个公共数据上进行了实验验证，实验结果表明所提方法的识别性能显著优于对比方法。

第 4 章为面向行人重识别的多视角深度对齐度量学习模型：首先，对于行人重识别中目标存在偏差，不同视角下样本出现错位的情形进行了介绍，阐述本方法的研究动机在于利用注意力机制引导模型，消除多视角样本的偏差。然后介绍了本方法的相关工作，包括推土机距离、行人重识别深度度量学习、行人重识别对齐学习以及卷积神经网络注意力机制。接着，详细介绍了提出的多视角深度对齐度量学习模型，包括基于推土机的跨视角三元组损失、注意力引导的对齐概率学习和网络结构。最后，在行人重识别的四个公共数据上进行了实验验证，实验结果表明所提方法能够学习到行人的感兴趣区域，并依靠感兴趣区域对齐和消除跨视角带来的样本偏差，有效提升深度网络性能。

第 5 章为对本论文工作的总结，介绍论文提出方法的实际应用情况，梳理论文的研究内容与主要贡献，同时展望未来工作的研究方向。

第 2 章　多线性多视角特征融合的图像检索模型

2.1　引言

大规模图像检索任务是指从大量的图像数据中找寻相关目标图像的任务，典型的应用包括目标检索 [29]、行人重识别 [47] 和场景搜索 [30] 等。由于该技术能够快速识别和匹配关键目标图像，其在目标搜索、行人搜索、安防主题搜索等安防领域有着极为重要的应用前景。图像检索往往包含两个关键步骤：首先，用一个固定长度的特征向量对图像进行特征描述，典型的特征向量包括词带模型 [17]、局部聚集描述子 (VLAD)[7] 和深度特征 (22; 6)；然后，对特征向量进行度量计算，该度量反映了二者的相似性，通过对相似性的排序，寻找到匹配目标图像。其中，特征表示是图像检索的重中之重，好的特征具有很强的判别性，能够有效区分不同目标，例如 SIFT 特征 [4] 对局部纹理有较好的表示能力，而 CNN 特征 [19; 20] 反映了目标的高阶语义信息。这些特征虽然都能够实现相关目标图像的查找，但得到的结果却不尽相同。因此，从特征的角度而言，不同的特征表示方法构成了同一目标的多视角表达，融合多视角特征的互补信息，提升匹配的精度已经成为了图像检索类任务的研究热点 [13; 8; 9]。然而，不同特征的维度、特性以及索引建立的方式，使得特征层级的融合变得困难，例如融合基于局部特征的词袋模型 (Bag of words, BOW)[15; 18] 和基于全局特征的哈希模型 (Hash)[10; 23; 57]。

一种简单且有效的替代方案是在索引 (Index) 层级进行多视角特征融合 [44; 46]，通过迭代更新索引，隐式地实现特征融合。索引结构通常被视为一种特殊的数据库管理机制，好的索引结构能够避免穷举搜索，极大地提升匹配系统的效率。倒排索引就是这样一种具有代表性的方法 [17]，这种方法首先将局部描述子 [5] 通过最近邻搜索被量化到附近的“单词”（Visual word, 即聚类中心）上，然后通过词频-逆文档频权重 (TF-IDF)，将每一副图像索引为稀疏向量，在进行匹配时，由于只计算稀疏向量的非零元元素的内积，该索引能够有效的降低计算量，从而处理大规模的图像匹配问题，受到了学术界和工业界的广泛关注和应用。

为了充分借助倒排索引结构，多索引融合算法 [44; 46] 在此基础上，通过索引层级的特征融合来捕获不同特征之间的互补信息，在保证匹配系统高效的同

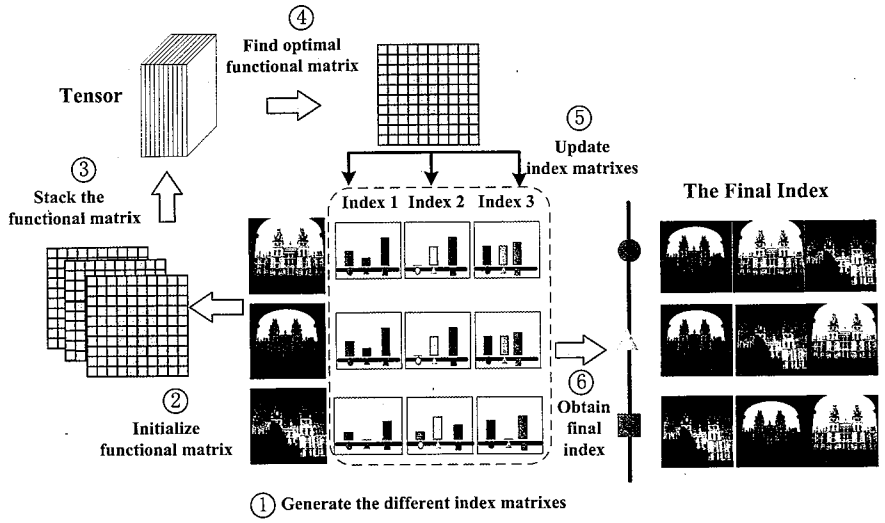


图 2.1 多线性多视角特征融合算法流程图。

Figure 2.1 The flowchart of multi-linear multi-view feature fusion.

时提升了匹配的精度。具体而言，多索引融合算法首先在一个特征索引空间下，通过计算欧式距离，寻找到待匹配库中的近邻关系。如果满足近邻关系的两幅图像恰好互为 k 近邻，则二者大概率是真正相关的，此时将这种近邻关系嵌入到另一种特征索引空间中。通过不停地交替迭代，使得二者之间的近邻关系不断进行迁移，进而提升特征的判别性。由于多索引融合算法不需要将多视角特征同时存储和计算，提供了有借鉴意义多视角特征融合的解决思路。

然而，现有的多索引融合算法往往忽略了特征与样本之间的高阶互补关系。此外，这类算法仅通过近邻关系的相互嵌入，也丢失了两幅图像之间的距离关系。更为重要的是，该类方法 [46] 不能处理多个视角同时更新的问题，严重制约了算法的使用场景。为此，本章提出一种多线性多视角特征融合算法 (MMF)。图 2.1详细地展示了算法流程框架。具体而言，多线性多视角特征融合算法首先根据不同的视角表示方法，设计不同的索引结构。其次，针对每一个视角索引，引入作用矩阵，将相似的样本推进，不相似的样本推远。为了得到理想的作用矩阵，算法将这些作用矩阵堆叠成高阶数组（张量），并在所建立的统一张量空间优化所有作用矩阵，利用张量低秩约束 [27]，比较每一个样本和每一种视角特征，捕获样本与视角之间的多线性关系。值得注意的是，MMF 充分考虑了索引的稀疏结构，在匹配测试阶段，并未带来额外的内存和计算开销；此外，多种视角之间的互补信息可以通过作用矩阵更为全面的进行传递。虽然优化多线性多视角特征融合算法需要较大的计算和内存开销，但 MMF 只需离线训练一次，并

可以通过分块的方式进一步降低计算量。

总结本章的主要贡献如下：

• 提出了一种新的多视角特征融合机制，实现索引层级的特征融合，其中多种视角特征之间的高阶互补信息可以有效地被张量低秩正则项挖掘；

• 提出了一种针对 MMF 目标函数的高效优化算法，并具有理论收敛性的保证；

• 在多个国际公开的数据集上进行了算法有效性验证，所提出的方法在匹配精度上和内存开销上优于相关方法，把行人重识别作为一种特殊的图像检索任务，所提算法能极大提升识别的准确率。

本章接下来内容的组织结构如下：第2.2节介绍相关的工作，符号和预备知识；第2.4节首先回顾多索引融合算法，给出多线性多视角特征融合算法的具体形式，并提出优化算法以及分析它的收敛性；第2.5节介绍实验研究相关的内容；最后，在第2.6节对本章工作进行总结并提出本章工作需要进行的后续研究。

2.2 相关工作

大多数图像检索及行人重识别算法可以被粗略的划分为两类：特征表示和索引结构，本章还与特征融合方法密切相关，本节将对其中代表性的算法进行简单回顾。

2.2.1 图像检索及行人重识别的特征表示方法

特征表示是图像检索的重要组成部分，关于特征表示的讨论一直是模式识别领域经久不衰的话题。早期的检索方法将图像表示成单一的全局统计特征 [10; 23; 7]，例如颜色直方图、形状上下文等。然而，这些简单的特征虽然维度较低，但判别性有限，仅仅适用于规模较小的图像检索数据。为了给出更具判别力的特征表示，以 SIFT 特征 [4] 为代表的具备旋转和尺度不变性的局部特征逐渐引入到图像检索系统中，配合词袋模型 [17]，取得了巨大成功 [30; 29; 12]。而近些年，许多工作致力于设计计算开销更小的全局特征，尤其是随着深度学习在特征表示方面的优势突显，一些工作尝试从预训练好的深度模型中提取特征。Babenko 等人 [15] 发现，经过 ImageNet 预训练过的卷积网络，提取到的特征具有很强的判别性。而一些研究工作开始运用卷积滤波器的激活值作为卷积特征，由于滤波器的感受野有限，这些卷积特征具有类似局部特征的性质，并且也保留了深度特

征的高阶语义特性 [32]，往往产生比局部特征更好的匹配结果 [22]。

行人图像的表示方法与之类似，早期研究人员利用行人的颜色信息 [64] 和梯度信息 [61]，进行行人重识别的研究。然而，手工特征 [88] 借助人类经验，使得特征表示能力不强，难以在真实场景中使用。随着深度学习时代的来临，基于高层语义特征的行人重识别方法占据了主流 [47]，此后，基于局部特征的表示 [139] 进一步提升了行人重识别精度。总体而言，不同的特征表示方法具有不同优势，可以产生不同的检索结果。

2.2.2 大规模图像检索索引结构

将局部特征索引为词袋模型，以及将全局特征索引为紧致的哈希编码是索引方法中的两条重要分支。对于哈希索引来说，数据无关的哈希编码可以产生较高的碰撞概率 [37; 38].，但往往依赖于较长的哈希位数 (Hash bits) 和较多的哈希表 (Hash tables) 个数。与之不同，以随机多视角哈希 [39]、谱哈希 [40] 和非线性稀疏哈希 [50] 为代表的数据相关的哈希方法，通过学习过程来产生较短的哈希编码，相较于数据无关的哈希编码，这种索引结构更为便捷和高效。虽然哈希方法能够提供较为准确的匹配结果，但这种索引结构是一种有损的索引方法（丢失信息），与此相反，作为无损的代表性方法之一，倒排索引结构展示出了处理大规模匹配的能力，逐渐成为图像检索系统中的主流方法 [30]。

对于倒排索引来说，将其应用于图像检索的开创性工作来自于 Zisserman 等人 [17]，它对一个视频片段中出现的目标进行了匹配。在此之后，大量的工作集中于帮助倒排索引存储更多的细节信息。Babenko 等人 [36] 提出了多索引倒排结构来减少量化损失。Zhou 等人 [33] 运用空间编码的方式，将局部特征的位置信息索引到倒排表中，取得了不错的效果。Zhang 等人 [35] 将局部描述子和语义线索同时引入到倒排索引中。Chen 等人 [45] 利用方向和位置信息，学习得到一个背景敏感的码本，极大地提升了性能。近些年，Mohedano 等人 [34] 利用倒排表的方式将深度特征进行索引，取得了不错的检索精度。

2.2.3 基于多特征融合的图像检索方法

为了有效利用不同特征之间的互补信息，一些工作已经开始尝试运用特征融合方法提升匹配的精度 [13; 8; 14]。其中 Zhang 等人 [8] 尝试在排序阶段融合不同的特征，通过在图上执行链接预测，提升了匹配性能。Zheng 等人 [13] 提出

了一种得分层级的特征融合方法。Zheng 等人 [14] 提出了一种耦合的多索引结构进行特征融合。在行人重识别领域，PCB[124] 融合了针对不同身体部位的局部特征，DMVFL[96] 运用手工特征和深度学习之间的相互协作学习，有效提升匹配准确率。然而，大部分的方法独立地处理每一种特征表示，忽略了不同特征之间的互补信息。此外，查询操作必须对多个索引执行多次，效率低下。

为了克服这些缺点，一些研究者开始尝试在索引层级进行特征融合。在这些方法中，一个共同的假设是：在一种特征表现形式下，两个样本如果相似，则假定它们是相关样本。通过在其他特征空间中将它们推得更近，可以大大提高搜索精度。在这一原则的指导下，Zhou 等人提出了协同索引嵌入方法 [46]，利用交替索引更新的方案来融合多种特征。通过把特征加入到另一个相关特征上，强化邻域结构，提高检索精度。Chen 等人 [44] 将该模型推广到多索引融合问题。但这两种方法都忽略了原始特征空间的距离信息。此外，还忽视了视角与样本之间的高阶特性。为了减轻这些影响。Xie 等人 [25] 在自表示矩阵上施加基于 t-SVD 的范数进行子空间聚类，为多索引融合方法提供了一种更为有效的解决思路。

2.3 符号系统和预备知识

2.3.1 符号系统

本章运用小写粗体字母 $\mathbf{x}$ 表示向量，用大写粗体字母 $\mathbf{X}$ 表示矩阵，以及用小写字母 x_{ij} 表示矩阵的元素。符号 $\|\mathbf{X}\|_F := (\sum_{i,j} |x_{ij}|^2)^{\frac{1}{2}}$，$\|\mathbf{X}\|_{2,1} := \sum_j (\sum_i x_{ij}^2)^{\frac{1}{2}}$ 和 $\|\mathbf{X}\|_1 := \sum_{i,j} |x_{ij}|$ 分别表示 Frobenius 范数，$\ell_{2,1}$ 范数和 ℓ_1 范数。核范数 $\|\mathbf{X}\|_*$ 的定义为 $\|\mathbf{X}\|_* := \sum_i \sigma_i(\mathbf{X})$，其中 $\sigma_i(\mathbf{X})$ 表示矩阵 $\mathbf{X}$ 第 i 大的特征值。令 $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ 表示一个三阶张量。2D 部分 $\boldsymbol{\mathcal{X}}(i,:,:)$, $\boldsymbol{\mathcal{X}}(:,i,:)$ 和 $\boldsymbol{\mathcal{X}}(:,:,i)$（此处沿用 Matlab 符号帮助理解）分别表示为张量 $\boldsymbol{\mathcal{X}}$ 的第 i 个水平切片 (Horizontal slices)，侧向切片 (Lateral slices) 以及正向切片 (Frontal slices)。与此类似，1D 部分 $\boldsymbol{\mathcal{X}}(i,j,:)$, $\boldsymbol{\mathcal{X}}(i,:,j)$ 和 $\boldsymbol{\mathcal{X}}(:,i,j)$ 分别表示第 i 个模式-1、模式-2 和模式-3 纤维 (Fiber)。特别地，用 $\boldsymbol{\mathcal{X}}^{(k)}$ 表示第 k 个正向切片 $\boldsymbol{\mathcal{X}}(:,:,k)$，$\boldsymbol{\mathcal{X}}_f$ 表示张量 $\boldsymbol{\mathcal{X}}$ 沿第三个维度做快速傅立叶变换后得到的张量。

2.3.2 张量预备知识

为了帮助理解基于 t-SVD 的张量核范数 (TNN-norm)[26; 43; 27; 42]，下面将给出一些张量运算的定义。

定义 2.1 (张量乘法). 令 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ 和 $\mathcal{Y} \in \mathbb{R}^{n_2 \times n_4 \times n_3}$ 为两个三阶张量，则两个张量的乘积 (t-product)$\mathcal{M} = \mathcal{X} * \mathcal{Y}$ 定义为：

$$\begin{bmatrix} \mathcal{M}^{(1)} \\ \mathcal{M}^{(2)} \\ \vdots \\ \mathcal{M}^{(n_3)} \end{bmatrix} = \begin{bmatrix} \mathcal{X}^{(1)} & \mathcal{X}^{(n_3)} & \cdots & \mathcal{X}^{(2)} \\ \mathcal{X}^{(2)} & \mathcal{X}^{(1)} & \cdots & \mathcal{X}^{(3)} \\ \vdots & \ddots & \ddots & \vdots \\ \mathcal{X}^{(n_3)} & \mathcal{X}^{(n_3-1)} & \cdots & \mathcal{X}^{(1)} \end{bmatrix} \cdot \begin{bmatrix} \mathcal{Y}^{(1)} \\ \mathcal{Y}^{(2)} \\ \vdots \\ \mathcal{Y}^{(n_3)} \end{bmatrix}. \tag{2.1}$$

其中 $\mathcal{M}$ 是一个维度为 $n_1 \times n_4 \times n_3$ 的张量，$\cdot$ 表示标准的矩阵乘法。

定义 2.2 (张量转置). 假设 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$，则它的转置 $\mathcal{X}^{\mathrm{T}}$ 是一个维度为 $n_2 \times n_1 \times n_3$ 的张量，其中转置操作是将原张量 $\mathcal{X}$ 的每一个正向切片先转置，然后按第 2 到第 n_3 个切片的逆序排列。

定义 2.3 (张量正交). 一个张量 $\mathcal{Q} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ 是正交的，那么

$$\mathcal{Q}^{\mathrm{T}} * \mathcal{Q} = \mathcal{Q} * \mathcal{Q}^{\mathrm{T}} = \mathcal{I}, \tag{2.2}$$

其中 $\mathcal{I} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ 是单位张量，它的第一个正向切片是单位矩阵，其余的正向切片是零矩阵。

根据这些定义，易于发现张量乘法可以转换为在傅立叶域的正向切片的矩阵乘法。数学上，公式 (2.1) 等价于：

$$\mathcal{M}_f^{(k)} = \mathcal{X}_f^{(k)} \mathcal{Y}_f^{(k)}, \ k = 1, \ldots, n_3. \tag{2.3}$$

基于此，可以得到一个重要的理论性质，即张量 SVD 分解 (t-SVD)[26]，它与矩阵形式十分类似：

定理 2.1 (t-SVD). 令 $\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ 为一个实值张量. 那么 $\mathcal{X}$ 可以被分解为：

$$\mathcal{X} = \mathcal{U} * \mathcal{S} * \mathcal{V}^{\mathrm{T}}, \tag{2.4}$$

其中 $\mathcal{U} \in \mathbb{R}^{n_1 \times n_1 \times n_3}$ 和 $\mathcal{V} \in \mathbb{R}^{n_2 \times n_2 \times n_3}$ 为正交张量. 且 $\mathcal{S} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ 的正交切片为对角矩阵。

定理 2.1表明，与矩阵 SVD 分解类似，任意一个实值张量都可以被写作为三个张量乘积的形式。此外，公式 (2.4) 在傅立叶域可以被写作为：

$$\begin{bmatrix} \mathcal{X}_f^{(1)} & & \\ & \ddots & \\ & & \mathcal{X}_f^{(n_3)} \end{bmatrix} = \begin{bmatrix} \mathcal{U}_f^{(1)} & & \\ & \ddots & \\ & & \mathcal{U}_f^{(n_3)} \end{bmatrix} \cdot \begin{bmatrix} \mathcal{S}_f^{(1)} & & \\ & \ddots & \\ & & \mathcal{S}_f^{(n_3)} \end{bmatrix} \cdot \begin{bmatrix} \mathcal{V}_f^{(1)} & & \\ & \ddots & \\ & & \mathcal{V}_f^{(n_3)} \end{bmatrix}^{\mathrm{T}}, \tag{2.5}$$

其中 $\mathcal{X}_f^{(i)} = \mathcal{U}_f^{(i)} \mathcal{S}_f^{(i)} (\mathcal{V}_f^{(i)})^{\mathrm{T}}, i = 1, \ldots, n_3$ 为标准的矩阵 SVD 分解。因此，基于 t-SVD 张量核范数可以以如下形式给出：

$$\|\mathcal{X}\|_{\circledast} := \sum_{i=1}^{\min(n_1, n_2)} \sum_{k=1}^{n_3} |\mathcal{S}_f(i, i, k)|. \tag{2.6}$$

由于在傅立叶域，对角块矩阵可以转换为原域的循环矩阵形式。因此基于 t-SVD 张量核范数也可以被写作：

$$\|\mathcal{X}\|_{\circledast} = \| \begin{bmatrix} \mathcal{X}^{(1)} & \mathcal{X}^{(n_3)} & \ldots & \mathcal{X}^{(2)} \\ \mathcal{X}^{(2)} & \mathcal{X}^{(1)} & \ldots & \mathcal{X}^{(n_3)} \\ \vdots & \ddots & \ddots & \vdots \\ \mathcal{X}^{(n_3)} & \mathcal{X}^{(n_3-1)} & \ldots & \mathcal{X}^{(1)} \end{bmatrix} \|_* \tag{2.7}$$

不同于现有的张量核范数，基于 t-SVD 张量核范数通过比较每一个正交切片的每一行和每一列来得到张量的秩，具有清晰的物理含义。

2.4 多线性多视角特征融合模型

2.4.1 多索引融合模型

多索引融合算法是一种能够在索引层级隐式地实现特征融合的技术，它可以只保留一种特征索引，从而完成快速的检索任务。具体而言，假设有 V 种不同视角的特征索引，表示为 $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_V \in \mathbb{R}^{d_v \times N}$，其中每一列为一个特征向量（例如词带模型向量）。$d_v$ 是第 v 种特征的维度，N 是待匹配库中样本的数目。

现有的多索引融合算法主要通过近邻结构，在不同视角的特征索引间传递

相似性。例如在 [46] 中，多索引融合可以形式上表示为：

$$\begin{aligned}\widetilde{\mathbf{X}}_1 &= \mathbf{X}_1 + \alpha \cdot g(\mathbf{X}_1) \odot \mathbf{X}_1 \cdot \Phi_2, \\ \widetilde{\mathbf{X}}_2 &= \mathbf{X}_2 + \beta \cdot g(\mathbf{X}_2) \odot \mathbf{X}_2 \cdot \Phi_1,\end{aligned} \tag{2.8}$$

其中 α 和 β 为常系数，$g(\cdot)$ 是一个示零函数，当其变量为零时取 1，其余为 0。$\odot$ 表示为逐元素相乘。Φ_1 和 Φ_2 表示为：

$$\begin{cases}\Phi_m(k,i) = 1, \text{ 如果 } k \neq i, \mathbf{x}_k \in \Re_m(\mathbf{x}_i) \\ \Phi_m(k,i) = 0, \text{ 其余情况}\end{cases}, \tag{2.9}$$

其中，$\Re_m(\mathbf{x}_i)$, $m = 1, 2$ 表示样本 i 在特征空间 m 上的近邻集。公式 (2.8) 和公式 (2.9) 假定如果在一个特征空间发现两个样本互为近邻，那么在另一个特征空间上，需要把一个样本的特征加入到另一个样本上，进而推近相似样本的距离。然而，[46] 只能融合两种类型的特征索引，即便 Chen 等人 [44] 将其拓展到了多索引情形，但也忽略了视角与样本之间的高阶关系，致使较低的检索精度。

2.4.2 多线性多视角特征融合模型

与之不同，为了保留各个视角表达下的自身特性，所提出的多线性多视角特征融合模型针对每一个视角，学习一个特定于视角的作用矩阵，以此来更新原有的索引矩阵。形式上，MMF 的更新方式如下式所示：

$$\mathbf{X}_v^* = \mathbf{X}_v(\mathbf{I} + \mathbf{Z}_v), \ v = 1, 2, \ldots, V, \tag{2.10}$$

其中，$\mathbf{I}$ 是单位矩阵，$\mathbf{Z}_v \in \mathbb{R}^{N \times N}$ 表示引入的作用矩阵，它的值 $z^v(i,j)$ 大于 0 表明样本 i 和样本 j 的距离需要被推近。这种融合过程可以被视为一种“离线”的查询扩展 (Query-expansion)。在给定作用矩阵的情况下，不同视角下测量到的相似性度量可以传播到不同的特征表示中，这样索引矩阵就可以得到更具判别性的加强。接下来的小节将介绍如何学习得到作用矩阵 $\mathbf{Z}_v$。

2.4.3 作用矩阵

为了得到理想的作用矩阵，MMF 引入两个基本假设，第一个假设是指将待匹配库视为一个整体空间，相关的样本组成了特殊的子空间结构。该假设是基于相关样本的特征向量之间彼此相似得出的，这种关系类似于子空间结构，这种假设被称为样本相关假设。除此之外，另一种假设是指不同视角下的距离相关假

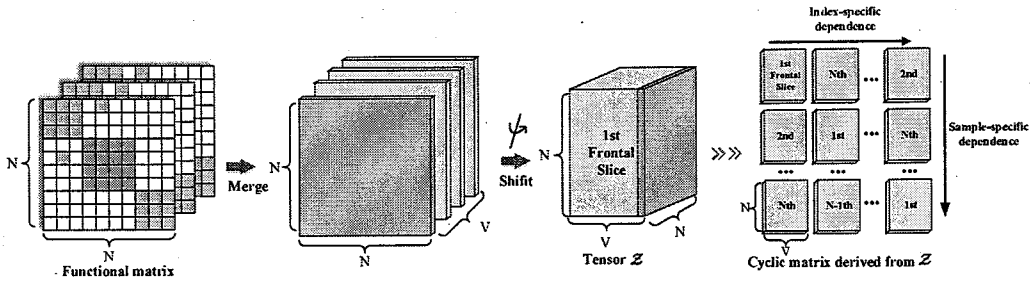


图 2.2 张量 $\mathcal{Z}$ 的建立方法及由它产生的循环矩阵。

Figure 2.2 The construction of tensor $\mathcal{Z}$ and its derived cyclic matrix.

设。具体而言，在不同的视角下，两个样本之间的相似性距离仍然高度一致。也就是说，由于不同的特征是对同一个样本的不同侧面表示，即便有时匹配结果差异很大，但在大多数特征空间中，相关目标的距离仍然很近，这种假设被称为视角相关假设。因此，对于多索引融合问题而言，作用矩阵的目的是通过两个基本假设，学习到不同特征空间的一致信息。

根据上述分析，MMF 提出一种多线性优化算法同时刻画两种相关性。特别地，在自表示模型基础上 [24]，多线性多视角特征融合模型形式如下：

$$\min_{\mathbf{Z}_v,\mathbf{E}} \lambda\|\mathbf{E}\|_{2,1} + \|\mathcal{Z}\|_{\circledast} + \sigma\sum_{v}\|\mathbf{Z}_v\|_1 \tag{2.11}$$

$$\text{s.t.} \quad \mathbf{X}_v = \mathbf{X}_v\mathbf{Z}_v + \mathbf{E}_v, v = 1, 2, \ldots, V,$$

其中 σ 和 λ 是两个常数，分别控制重建误差和作用矩阵的稀疏性。$\mathbf{X}_v \in \mathbb{R}^{d_v \times N}$ 表示第 v 个索引矩阵，$\mathbf{Z}_v \in \mathbb{R}^{N\times N}$ 表示第 v 个作用矩阵。$\mathcal{Z} = \Phi(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_V) \in \mathbb{R}^{N\times V\times N}$ 是一个通过堆叠不同作用矩阵并翻转得到的 3 阶张量，$\mathbf{E} = [\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_V]$ 表示残差矩阵，$\|\cdot\|_{\circledast}$ 表示基于张量 SVD 的核范数。图 2.2展示了张量的具体构建过程以及对应的循环矩阵形式。

直观而言，公式 (2.11) 中的 $\|\mathbf{E}\|_{2,1}$ 试图降低重建误差，其目的是尽量少地更新索引矩阵，以此来保持特征的原始表达。张量核范数 $\|\mathcal{Z}\|_{\circledast}$ 用于刻画两种相关性假设，从公式 (2.7) 中可以发现，通过比较作用矩阵中的每一行（样本相关）和每一列（视角相关），两个基本假设可以同时满足。对于样本相关性假设而言，算法假定每一个作用矩阵 Z_v 具有低秩结构，从而保证相关样本之间的子空间结构；对于视角相关性假设而言，算法假定不同作用矩阵之间具有高度的相关性，从而保证不同特征表达的一致性。稀疏项 $\|\mathbf{Z}_v\|_1$ 有两个目的，一是试图尽量少的

更新样本特征，二是保证原有稀疏的索引结构。通过该优化目标，相关的样本信息可以被嵌入到新的特征中去，进而推近相关样本，推远不相关样本。

2.4.4 优化方法

该优化目标可以通过增广拉格朗日乘子(Augmented Lagrange Multiplier)[28]来进行有效地求解。通过引入额外的辅助张量 $\mathcal{G}$ 和辅助矩阵 $\mathbf{M}_v$, $v = 1, 2, \ldots, V$，该优化问题可以被转换为无约束优化问题：

$$\begin{aligned}
&\mathcal{L}(\mathbf{Z}_1, \ldots, \mathbf{Z}_V; \mathbf{E}_1, \ldots, \mathbf{E}_V; \mathbf{M}_1, \ldots, \mathbf{M}_V; \mathcal{G}) \\
&= \sum_v (\sigma ||\mathbf{M}_v||_1 + \langle \mathbf{Y}_v, \mathbf{X}_v - \mathbf{X}_v\mathbf{Z}_v - \mathbf{E}_v \rangle \\
&+ \frac{\mu}{2} ||\mathbf{X}_v - \mathbf{X}_v\mathbf{Z}_v - \mathbf{E}_v||_F^2 + \langle \mathbf{N}_v, \mathbf{Z}_v - \mathbf{M}_v \rangle \\
&+ \frac{\xi}{2} ||\mathbf{Z}_v - \mathbf{M}_v||_F^2) + \lambda ||\mathbf{E}||_{2,1} + ||\mathcal{G}||_{\circledast} \\
&+ \langle \mathcal{W}, \mathcal{Z} - \mathcal{G} \rangle + \frac{\rho}{2} ||\mathcal{Z} - \mathcal{G}||_F^2,
\end{aligned} \tag{2.12}$$

其中矩阵 $\mathbf{N}_v$,$\mathbf{Y}_v$ 和张量 $\mathcal{W}$ 是拉格朗日乘子。μ, ξ 和 ρ 是惩罚参数。对 $\mathbf{E}_u^v$、$\mathbf{M}_u^v$、$\mathbf{Z}_u^v$ 和 $\mathcal{G}$ 进行精确地联合求解看起来十分困难，因此本小节提出一种交替的优化算法，将无约束问题交替划分为以下四个步骤。

子问题 $\mathbf{Z}_v$：当 $\mathcal{G}, \mathbf{E}, \mathbf{M}$ 固定时，原问题转化为：

$$\begin{aligned}
&\min_{\mathbf{Z}_v} \langle \mathbf{Y}_v, \mathbf{X}_v - \mathbf{X}_v\mathbf{Z}_v - \mathbf{E}_v \rangle + \frac{\mu}{2} ||\mathbf{X}_v - \mathbf{X}_v\mathbf{Z}_v \\
&- \mathbf{E}_v||_F^2 + \langle \mathbf{N}_v, \mathbf{Z}_v - \mathbf{M}_v \rangle + \frac{\xi}{2} ||\mathbf{Z}_v - \mathbf{M}_v||_F^2 \\
&+ \langle \mathbf{W}_v, \mathbf{Z}_v - \mathbf{G}_v \rangle + \frac{\rho}{2} ||\mathbf{Z}_v - \mathbf{G}_v||_F^2.
\end{aligned} \tag{2.13}$$

由于闭式解的存在，该优化问题易于求解，通过将导数置为零，即可获得最优解 $\mathbf{Z}_v^*$：

$$\begin{aligned}
\mathbf{Z}_v^* =& (\mathbf{X}_v{}^T\mathbf{Y}_v + \mu\mathbf{X}_v{}^T\mathbf{X}_v - \mu\mathbf{X}_v{}^T\mathbf{E}_v - \mathbf{W}_v - \mathbf{N}_v \\
&+ \rho\mathbf{G}_v + \xi\mathbf{M}_v)/(\rho + \xi)(\mathbf{I} + \frac{\mu}{\rho + \xi}\mathbf{X}_v{}^T\mathbf{X}_v)^{-1}.
\end{aligned} \tag{2.14}$$

子问题 $\mathbf{M}_v$：当 $\mathcal{G}, \mathbf{E}$ 和 $\mathbf{Z}$ 固定时，求解公式 (2.12) 等价于最小化下述优化问题：

$$\min_{\mathbf{M}_v} \sigma ||\mathbf{M}_v||_1 + \frac{\xi}{2} ||\mathbf{M}_v - (\mathbf{Z}_v + \frac{1}{\xi}\mathbf{N}_v)||_F^2. \tag{2.15}$$

该问题可以运用软阈值 (Soft Thresholding) 方法求解：

$$\begin{aligned}\mathbf{M}_v^*(i,j) =& sign(\mathbf{Z}_v(i,j)+\frac{1}{\xi}\mathbf{N}_v(i,j))\cdot \\ & max(|\mathbf{Z}_v(i,j)+\frac{1}{\xi}\mathbf{N}_v(i,j)|-\frac{\sigma}{\xi},0).\end{aligned} \tag{2.16}$$

子问题 $\mathbf{E}_v$：对于给定的 $\mathcal{G}$、$\mathbf{Z}$ 和 $\mathbf{M}$，原问题转换为：

$$\begin{aligned}\mathbf{E}^* =& \text{argmin}_{\mathbf{E}}\lambda\|\mathbf{E}\|_{2,1}+\frac{\mu}{2}\|\mathbf{X}_v-\mathbf{X}_v\mathbf{Z}_v-\mathbf{E}_v\|_F^2 \\ & +\sum_v\langle\mathbf{Y}_v,\mathbf{X}_v-\mathbf{X}_v\mathbf{Z}_v-\mathbf{E}_v\rangle \\ =& \text{argmin}_{\mathbf{E}}\lambda\|\mathbf{E}\|_{2,1}+\frac{1}{2}\|\mathbf{E}-\mathbf{D}\|_F^2,\end{aligned} \tag{2.17}$$

其中 $\mathbf{D}$ 是由将矩阵 $(\mathbf{X}_v-\mathbf{X}_v\mathbf{Z}_v+\frac{1}{\mu}\mathbf{Y}_v), v=1,2,\ldots,V$ 竖直拼接得到。这个子问题可由 [24] 提出的方法进行求解。

子问题 $\mathcal{G}$：最终，当 $\mathbf{E}, \mathbf{Z}, \mathbf{M}$ 固定时，原问题转换为下述子问题求解张量 $\mathcal{G}$：

$$\mathcal{G}^* = \text{argmin}_{\mathcal{G}}\|\mathcal{G}\|_{\circledast}+\frac{\rho}{2}\|\mathcal{G}-(\mathcal{Z}+\frac{1}{\rho}\mathcal{W})\|. \tag{2.18}$$

该问题在傅立叶域转化为：

$$\mathcal{G}_f^* = \text{argmin}_{\mathcal{G}_f}\sum_{j=1}^{N}\tau'\|\mathcal{G}_f^{(j)}\|_*+\frac{\rho}{2}\|\mathcal{G}_f^{(j)}-(\mathcal{Z}+\frac{1}{\rho}\mathcal{W})_f^{(j)}\|_F^2. \tag{2.19}$$

该张量优化问题 (2.19) 可以在傅立叶域中划分为 N 独立的矩阵核范数问题。[25] 中提出的方法可用于解决此子问题。

除此之外，拉格朗日乘子也需要按如下方式更新：

$$\begin{aligned}\mathbf{Y}_v^* &= \mathbf{Y}_v+\mu(\mathbf{X}_v-\mathbf{X}_v\mathbf{Z}_v-\mathbf{E}_v), \\ \mathcal{W}^* &= \mathcal{W}+\rho(\mathcal{Z}-\mathcal{G}), \\ \mathbf{N}_v^* &= \mathbf{N}_v+\xi(\mathbf{Z}_v-\mathbf{N}_v).\end{aligned} \tag{2.20}$$

上述步骤不断重复，直至满足收敛条件。虽然从理论上证明算法的收敛性并不容易，但该优化方法满足由 [24] 提出的两个收敛充分条件，并且在实际优化过程中，该算法收敛良好。此外，作用矩阵上的微小数值可能不会影响检索的精度，但是可能会将噪声引入到新的索引中。因此本算法还将低于某个阈值 θ_1 的数值过滤为零：

$$sparse(\mathbf{Z}) = \begin{cases} 0, & |z_{ij}|<\theta_1 \\ z_{ij}, & |z_{ij}|\geq\theta_1 \end{cases} \tag{2.21}$$

算法 1 MMF 算法

输入：索引矩阵 $\mathbf{X}_v,\ v = 1, 2, \ldots, V$

输入参数：超参数 $\lambda > 0, \sigma > 0$ 、稀疏系数 θ_1, θ_2 以及迭代次数 $iters$

for $iter = 1$ to $iters$ **do**

  初始化：$\mathbf{Z}_v = \mathbf{E}_v = \mathbf{Y}_v = \mathbf{M}_v = \mathbf{N}_v = \mathbf{0}$、$\mathcal{G} = \mathcal{W} = \mathbf{0}$

  初始化：$\mu = \rho = \xi = 10^{-5}$、$\eta = 2$、$\mu_{\max} = \rho_{\max} = \xi_{\max} = 10^{10}$、$\varepsilon = 10^{-7}$

  while 未收敛 **do**

    更新 $\mathbf{Z}_v,\ v = 1, 2, \ldots, V$ 通过公式 (2.14)

    更新 $\mathbf{E}$ 通过求解子问题 (2.17)

    更新 $\mathbf{M}_v,\ v = 1, 2, \ldots, V$ 通过公式 (2.16)

    堆叠得到张量 $\mathcal{Z} = \Phi(\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_V)$

    更新 $\mathcal{G}$ 通过子问题 (2.19)

    更新拉格朗日乘子 $\mathcal{W}, \mathrm{Y}_v, \mathrm{N}_v,\ v = 1, 2, \cdots, V$ 通过公式 (2.20)

    更新参数 $\mu = \min(\eta\mu, \mu_{\max}),\ \rho = \min(\eta\rho, \rho_{\max}),\ \xi = \min(\eta\xi, \xi_{\max})$

    拆解得到矩阵 $(\mathbf{Z}_1, \ldots, \mathbf{Z}_V) = \Phi^{-1}(\mathcal{Z})$

    拆解得到矩阵 $(\mathbf{G}_1, \ldots, \mathbf{G}_V) = \Phi^{-1}(\mathcal{G})$

    检查收敛条件：$||\mathbf{X}_v - \mathbf{X}_v\mathbf{Z}_v - \mathbf{E}_v||_\infty < \varepsilon$

    检查收敛条件：$||\mathbf{Z}_v - \mathbf{G}_v||_\infty < \varepsilon$

    检查收敛条件：$||\mathbf{Z}_v - \mathbf{M}_v||_\infty < \varepsilon$

  end while

  $\mathbf{Z}_v^* = sparse(\mathbf{Z}_v),\ v = 1, 2, \ldots, V$

  $\mathbf{X}_i = \mathbf{X}_i \sum_{v=1}^{V}(\mathbf{Z}_v^* + \mathbf{Z}_v^{*T}),\ i = 1, 2, \ldots, V$

  $\lambda = 10 \cdot \lambda,$

  $\sigma = 10 \cdot \sigma$

end for

$\mathbf{X}_v^* = sparse(\mathbf{X}_v),\ v = 1, 2, \ldots, V$

输出：融合后的索引矩阵 $\mathbf{X}_v^*,\ v = 1, 2, \ldots, V$

2.4.5 索引更新和在线匹配

在融合过程中，直接使用公式 (2.10) 更新索引可能是一个次优选择，这是由于 Z_v 不能保留不同索引之间的相似关系。因此，为了平衡不同作用矩阵，本算法按照如下方式更新索引：

$$\mathbf{X}_v^* = \mathbf{X}_v(\mathbf{I} + \frac{1}{V}\sum_v(\mathbf{Z}_v + \mathbf{Z}_v^T)), v = 1, 2, \ldots V. \tag{2.22}$$

算法反复融合索引 T 次，直至获得最佳的检索精度。在每次迭代更新后，算法还对每个新索引执行归一化操作，以便于下次更新。此外，在下一次优化公式 (2.11) 获取作用矩阵时，算法还将超参数 λ 和 σ 的取值扩大十倍，来保证原始的特征表达。当索引完成后，只保留一个索引作为最终索引进行在线匹配。为了利用倒排索引的稀疏结构来降低内存和计算开销，算法还将最终索引上低于阈值 θ_2 的元素设置为零，这样即保证了匹配的精度，又极大地提升了匹配的效率。

在在线的匹配阶段，给定查询样本 q，系统首先提取原始特征表达 $\mathbf{x}_u^v(q)$（与保留的索引特征类型一致）。然后，算法通过倒排索引计算查询样本与每个数据库样本之间的相似度，其中数据库样本是由融合后的索引进行表示存储的。值得注意的是，算法充分利用特征索引的稀疏性，使得在线检索的计算复杂度大大降低。最后，MMF 通过对相似性度量进行降序排序，返回匹配结果。算法1总结了整个融合过程。

2.5 实验

在本章的实验中，将对所提方法的有效性进行全面评估。本节针对两个图像检索中的典型应用，图像检索和行人重识别任务，进行实验。其中，行人重识别作为一个特殊的图像检索任务，仅对算法的匹配精度进行评测，而针对目标检索任务，同时对算法的检索精度和内存开销进行评测。本节主要对基准方法和多个有代表性的图像检索算法进行对比，以此来说明算法的有效性。所有实验都是在一台上工作站上用 Matlab 实现的，该工作站配备了 Intel Xeon E5-2630@2.30GHz CPU、128GB RAM 和 TITANX GPU（12GB 缓存）。

2.5.1 实验设置

实验在三个公共的基准数据集，UKBench[29]、Holidays[30] 和 Market-1501[47] 上进行评估，其中 UKBench 和 Holidays 用于图像检索任务，Market-1501 用于行

人重识别任务。UKBench 数据集包含 10200 个样本，该数据集中的所有样本作为查询 (Query)，其余作为待匹配库 (Gallery) 的样本，每个目标都有 4 个相关目标，通过 N-S 得分 (N-S score) 来评估算法的匹配精度，N-S 得分是针对每一个查询样本，计算 Gallery 中排序前 4 个样本中相关样本的个数，最后取均值。Holidays 数据集包含从个人照片中获取的 1491 张图像，其中选择 500 张图像进行查询。采用平均精度 (mAP) 来评价检索精度，假设检索返回个数为 N，相关样本个数为 P，整个数据集中真实相关样本个数为 A，则查全率为 P/A，查准率为 P/N，mAP 表示不同查询下，查全率和查准率所组成的曲线的平均曲下面积。此外，对于 UKBench 和 Holidays 数据集，在线匹配内存开销 (OQMC) 用于评估匹配时的内存开销，OQMC 表示当前索引结构存储一张图片的内存开销。Market-1501 数据集是在清华大学的一家超市前收集的。总的来说，这个数据集包含 32668 个带注释的行人检测框，共 1501 个行人。其中 12936 幅图像用于训练，其余 19732 幅图像用于测试。采用累积匹配曲线 (CMC) 和平均精度 (mAP) 进行评价，其中 CMC 表示前 K 个检索结果中出现相关样本的的概率，通常用 Rank-1、Rank-5、Rank-10 accuracy 表示 $K = 1, 5, 10$ 的情形。如果返回的第一个图像为相关样本，则 Rank-1 accuracy=1，否则为 0。值得注意的是，算法只使用测试集的 Gallery 来执行多索引融合过程。

2.5.2 实施细节

本小节将介绍算法的具体实施细节，包括特征的表示方法和索引的构建方法。在 UKBench 和 Holidays 数据集上，算法分别提取三种类型的特征来构建索引，即局部 SIFT 特征、卷积神经网络全连接特征和卷积神经网络卷积特征。具体而言，对于 SIFT 特征，算法首先提取 SIFT 特征描述子 [5; 4]，并使用根 SIFT(Root SIFT) 来归一化每个描述子 [12]。为了构建倒排索引，算法首先预先训练好码本 (Codebook)，为了避免量化过程中的损失，算法将每个描述子分配给最近的三个视觉单词 (Visual word)[16; 11]。遵循传统的词袋模型 [17]，算法以 TF-IDF 方式将每个样本表示为 20K 稀疏向量，这种索引被称作 SIFT index。对于卷积神经网络全连接特征，算法首先将每个图像的大小调整为 224×224，然后将其通过深度卷积网络 (AlexNet)[19]，该网络是在 ImageNet 上由 Caffe[21] 预训练完成。最终提取全连接层 (FC6) 的输出作为图像的特征表示，从索引的角度来看，特征向量的每一维都同样可以被视作一个视觉单词，这种索引被称作

FC index。对于卷积神经网络卷积特征，算法使用 VggNet[20] 作为卷积特征提取器，该网络也是在 ImageNet 上预先训练完成的。为了得到卷积特征，算法提取了 conv54 层的特征图 (Featuremap)，其大小为 $14 \times 14 \times 512$。然后将激活响应 $\mathbf{f}_h(m) \in \mathcal{R}^{512}, m = 1, 2, \ldots, 196$ 作为算法的特征向量。类似于标准词袋模型的量化策略，首先用这些得到的特征，预训练长度为 10K 码本，然后通过 TF-IDF 加权的方式，将每个 $\mathbf{f}_h(m)$ 量化为三个最近的视觉单词，这种索引被称作 HC index。对于 Market-1501 数据集，算法遵循 [48; 47] 提出的基准方法，提取三种不同类型的特征。

2.5.3 图像检索实验结果

图像检索的结果在表2.1上展示，其中 MMF-表示多线性多视角特征融合后的结果，* 表示基准方法。与融合前的特征索引相对比，经过 MMF 融合后的结果在 UKBench 和 Holidays 数据集上都取得了巨大提升。在 UKBench 数据集上，SIFT index、FC index 和 HC index 的 N-S 得分分别为 3.94、3.92 和 3.87，与基准方法相比，分别有 30.8%、14.6% 和 18.0% 的提升。对于 Holidays 数据集，算法将 SIFT index 的 mAP 从 31.8% 提升至 84.8%，FC index 的 mAP 从 70.4% 提升至 93.6%，HC index 的 mAP 从 74.3% 提升至 94.1%。实验结果表明，所提方法能够有效地捕获 SIFT 特征和 CNN 特征之间的互补信息，从而大幅度提升原有特征表达的检索精度。此外，所提方法具有很强的鲁棒性，在面对较弱的视觉特征时，例如 Holidays 数据集上的 SIFT index，MMF 也能很好的提升各个索引的性能。另外一方面，不同的索引特征对最终的融合结果有很大的影响，例如高维 MMF-SIFT index 优于 UKbench 上的其他索引，而 MMF-HC index 则在 Holidays 数据集上表现优异。产生这种现象的原因是这两个数据集场景差别较大，Holidays 数据集包含了许多复杂的地标场景，UKbench 数据集则包含了相对简单的物体场景，这使得在 UKbench 数据集上融合前的 SIFT index 匹配精度要远优于其在 Holidays 数据集上的表现。

此外，MMF 算法还显著优于其他特征融合方法，例如查询融合 (QaLF)[13]，排序融合 (QSF)[8] 和耦合索引 (c-MI)[14]，这些对比方法结果直接摘自原始论文。结果表明，融合后的特征索引不仅取得了更高的匹配精度，内存开销还更少。尽管多索引融合方法 MFSMP[44] 也不需要过多的内存开销，但其匹配精度远低于 MMF。同时，为了取得准确的匹配结果，CIE[46] 需要良好的基准方法进

表 2.1 MMF 在 UKBench 和 Holidays 数据集上检索精度和内存开销的比较。

Table 2.1 Comparison of retrieval accuracy and memory cost on UKBench and Holidays.

方法	UKBench(NS-score)	Holiday(mAP)	OQMC
SIFT Index*	3.03	31.8	21.5KB
FC Index*	3.42	70.4	5.1KB
HC Index*	3.28	74.3	1.5KB
c-MI [14]	3.85	85.8	13.5KB
QSF [8]	3.77	84.6	20KB
QaLF [13]	3.84	88.0	62KB
CIE [46]	3.86	89.2	4KB
MFSMP [44]	3.78	78.8	0.38KB
CoInd [35]	3.60	80.9	24KB
MMF-SIFT	**3.94**	84.8	10.1KB
MMF-FC	3.92	93.6	2.8KB
MMF-HC	3.87	**94.1**	1.2KB

行融合，例如需要 N-S score 为 3.53 的深度索引和 N-S score 为 3.33 的 SIFT 特征索引，这也间接证明了 MMF 融合策略的有效性。

在在线匹配阶段，多线性多视角特征融合算法的主要存储开销在于存储 MMF 索引文件。假设每张图片 SIFT 描述子的个数为 2000，索引矩阵中的每个非零元素需要占据 8 个字节来存储倒排索引的权重和图像 ID。在应用稀疏操作之后，融合后的索引文件需要的内存开销甚至少于原始索引文件，同时融合后的索引还保证了不错的检索精度。在线匹配的计算复杂度也得益于索引文件的稀疏性，大大降低了匹配响应时间。

2.5.4 行人重识别实验结果

遵从 [48; 47] 中的方法，算法在 market-1501 上提取了三种图像特征，即词袋特征、ResNet50 特征和 AlexNet 特征（CaffeNet 特征）。如表4.1所示，所提出的方法在单次查询模式 (Single-query) 和多次查询模式 (Multi-query) 下，检索精度都远超基准方法。尽管 CaffeNet 特征和词袋特征得到了更多的提升，但 MMF-ResNet50 仍然取得了最佳的 Rank-1 accuracy 和 mAP，同时其初始特征也是三者中最具判别力的一个。此外，mAP 的提升远高于 Rank-1 accuracy，不难理解，多索引融合模型实质上是一种离线的查询扩展 (query expansion) 或重排序技术 (reranking)，这类方法不能从根本上提升视觉特征的表示能力，但可以有效地把

表 2.2 MMF 在 Market-1501 数据集上检索精度的比较。

Table 2.2 Comparison of retrieval accuracy on Mraket-1501.

Market-1501	单次查询		多次查询	
方法	Rank-1	mAP	Rank-1	mAP
BOW*[47]	35.84	14.75	44.36	19.41
CaffeNet*[48]	49.36	32.10	66.63	41.25
ResNet50*[48]	74.02	49.36	81.26	59.10
NULL[52]	61.02	35.68	71.56	46.03
Reranking[51]	**77.11**	**63.63**	-	-
LBA[49]	73.87	47.89	81.29	56.98
Gate-SCNN[54]	65.88	39.55	76.04	48.45
S-LSTM[53]	-	-	65.6	35.31
SCSP[55]	51.9	26.35	-	-
CIE[46]	73.77	57.55	79.39	65.24
SSDAL[56]	39.40	19.60	49.00	25.80
MMF-BOW	55.73	37.64	63.81	44.37
MMF-CaffeNet	69.63	53.93	76.93	61.79
MMF-ResNet50	**77.11**	62.39	**82.51**	**69.58**

Gallery 中相关的样本聚拢在一起。而重排序方法 (Reranking)[51] 也得到了类似的结果。另一方面，如果能够设计出更具辨别力的特征表达，MMF 就能进一步提高特征的判别能力。一些代表性的匹配结果在图2.3中展示，其中黑色边框表示干扰图像或来自同一摄像头下的图像，红色边框表示真正匹配的样本，没有边框的样本代表错误的匹配。检索结果表明，不同特征之间的互补信息可以通过 MMF 进行有效传递，例如运用 CaffeNet 特征检索得到的前两名的图像，它们与查询图像一样，都来自同一行人，MMF 可以把这种相似关系向其他的特征表示方法（ResNet50 特征）传递。

如表 1 所示，所提出的方法与相关的对比方法取得了相近（甚至更好的）结果，这些对比方法包括门孪生卷积神经网络 (Gate-SCNN) [54]、判别无空间 (NULL) [52]、空间约束度量 (SCSP)[55]、长短时记忆孪生网络 (S-LSTM) [53]、合成训练数据 (LBA) [49]、深度特征网络 (SSDAL) [56] 和重排序 (Reranking)[51]。由于多索引融合方法 (CIE)[46] 没有提供行人重识别任务的实验结果，为了公平比较，本小节复现了该论文的算法，并在 Market-1501 上数据集上运行，与原论文相比，复现的 CIE 方法类似 MMF 提取了三种不同特征，并融合其中的任意两

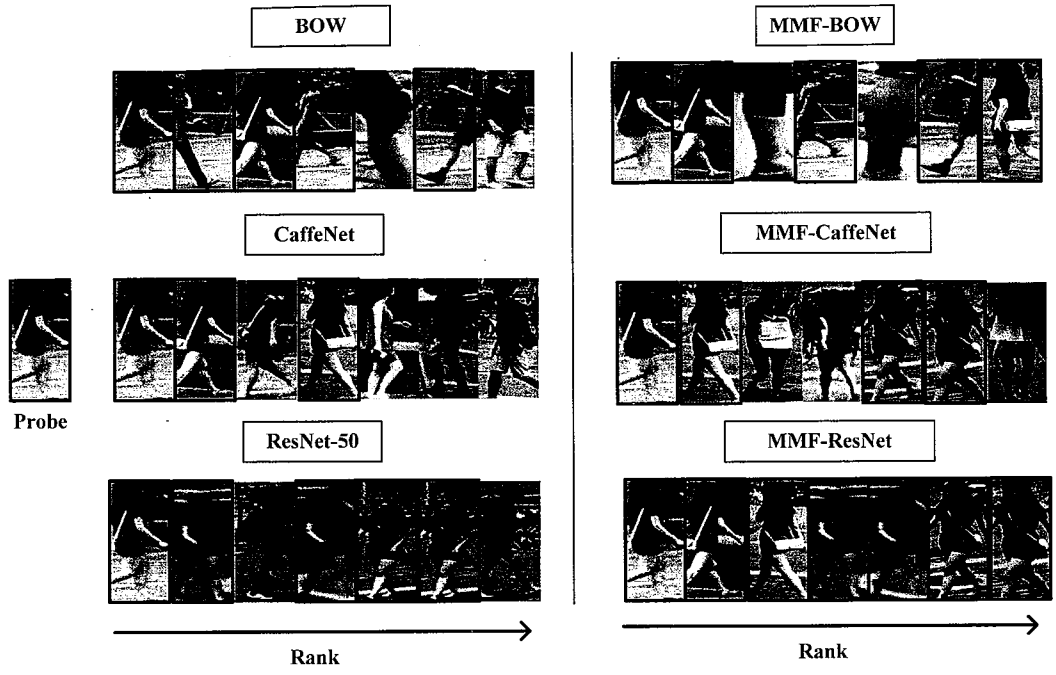


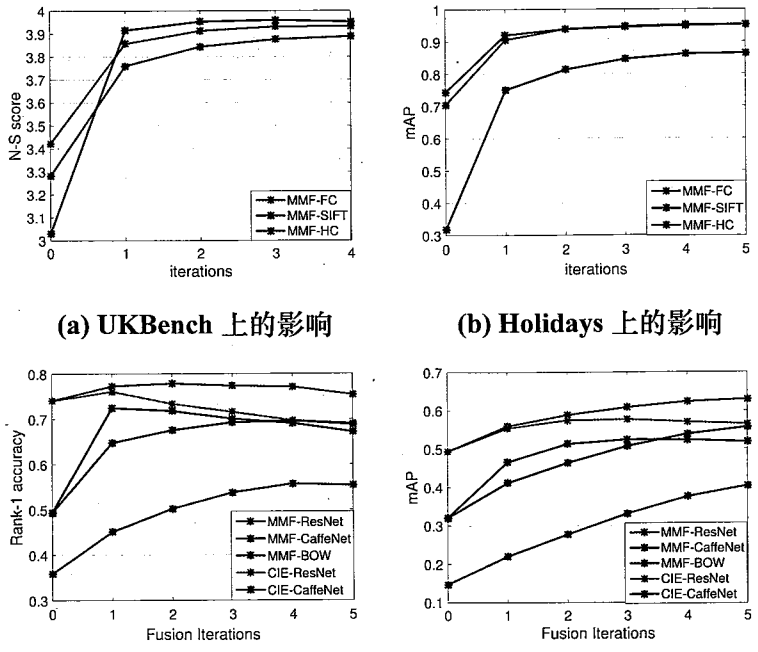
图 2.3 Market-1501 数据集上多线性多视角特征融合算法代表性检索结果。

Figure 2.3 Representative retrieval results of multi-linear multi-index fusion on the Market-1501 dataset.

个。实验表明，CIE 融合 ResNet50 和 CaffeNet 特征的结果取得了最佳性能。然而，由 MMF 得到的融合结果的 Rank-1 accuracy 和 mAP 相较于该方法都有明显的提升，验证了本算法的有效性。

2.5.5 参数分析

本小节将讨论不同参数对算法性能的影响。五个关键参数很大程度上影响着检索系统的性能，它们包括公式 (2.11) 中的 λ 和 σ，作用矩阵的稀疏阈值 θ_1，最终索引的稀疏阈值 θ_2，以及融合迭代数 T。本小节首先评估融合迭代次数 T 对检索精度的影响，如图3.8所示，随着融合迭代次数的增加，UKbench 和 Holidays 上的检索精度先迅速达到峰值，然后保持稳定。在 UKBench 上，算法迭代融合 3 次，直至 MMF-SIFT 索引获得最佳的检索精度，此时 MMF-HC 索引仍有上升的空间。在 Holidays 数据集上，当所有索引的检索精度保持稳定时，算法融合了 4 次。对于 Market-1501 数据集，MMF-ResNet50 的 Rank-1 accuracy 在 $T = 2$ 时达到最高，之后略有下降，该现象在 CIE 上同样出现，但各索引的 mAP 不断提升。最终算法选择 $T = 4$ 用于 Market-1501，以获得相对稳定的性能。虽然 CIE 方法在选取 $\alpha = \beta = 0.4$, $p = q = 9$, $m = 20$（通过网格搜索得到）时取得了同样



(a) UKBench 上的影响 **(b) Holidays 上的影响**

(c) Market-1501 上 CMC 影响 **(d) Market-1501 上 mAP 影响**

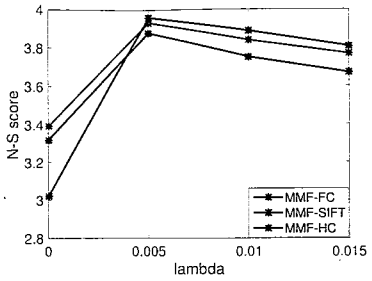
图 2.4 融合迭代次数 T 在不同数据集上对匹配精度的影响。

Figure 2.4 Influence of iteration number on retrieval accuracy.

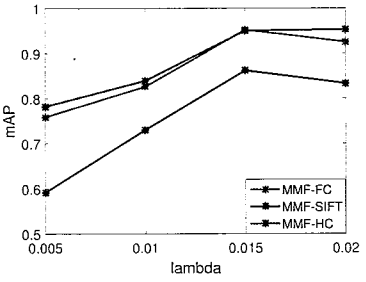
不错的结果，但是它的 Rank-1 accuracy 会在几个迭代以后迅速下降。这个现象会在章节2.5.6中详细讨论。

本小节接下来使用重建误差参数 λ 和稀疏参数 σ 的不同取值，来评估二者对性能的影响。如图2.5所示，虽然 λ 和 σ 对性能的起着决定性作用，但是大多数的结果仍然优于基准方法。当固定 $\sigma = 0.001$，λ 从 0 开始增加时，检索的精度首先攀升到峰值点，然后在两个数据集上缓慢下降。造成这个现象的原因可能是由于 λ 越大，对索引的改变就越少。当 λ 减少到 0 时，作用矩阵退化为单位矩阵。与之类似，当固定 $\lambda = 0.0005$，稀疏参数 σ 从 0 开始增加时，检索的精度也是先上升再缓慢下降。当 σ 增加到一定程度时，由于过多的稀疏约束，作用矩阵的所有值都被抑制，这导致检索精度，例如 MMF-SIFT 索引的急剧下降。对于 Market-1501 数据集而言，算法设置 $\lambda = 0.010$ 和 $\sigma = 0.001$ 以获得最佳性能。经验上，λ 通常取 0.005 到 0.015 之间，而取 $\sigma = 0.001$ 适用于大多数情况。

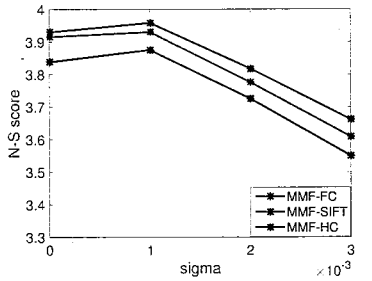
图2.6展示了阈值 θ_2 对最终索引的稀疏性和检索精度的影响。从图中可以很轻易地得到，融合迭代次数越大，索引结构的稀疏性就越低，同样地，这种现象也出现在多索引融合算法 [44; 46] 中。当 θ_2 增大时，两个数据集的索引稀疏度急



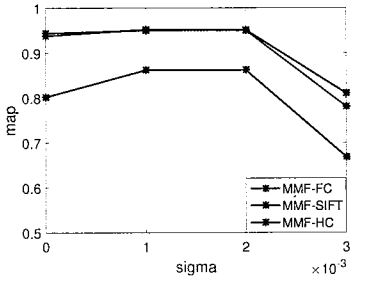
(a) λ 在 UKBench 上的影响



(b) λ 在 Holidays 上的影响



(c) σ 在 UKBench 上的影响



(d) σ 在 Holidays 上的影响

图 2.5 参数 λ 和 σ 在不同数据集上对匹配精度的影响。

Figure 2.5 Influence of λ and σ on retrieval accuracy.

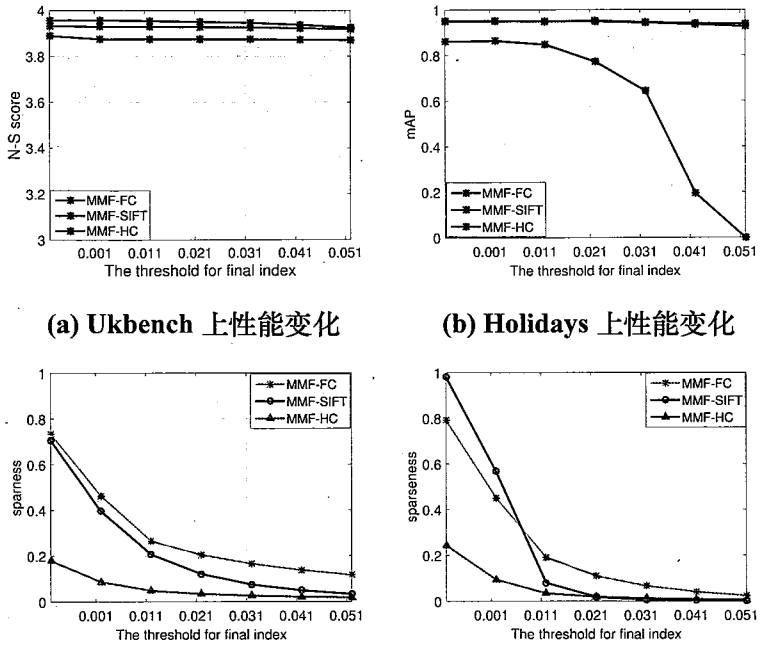
剧下降，而检索精度保持稳定。Holiday 上的情况与 UKbench 上略有不同，尤其是对于 MMF-SIFT 索引而言。这是由于 UKbench 上，算法选用了更多的融合迭代次数，以及该数据集上具有更多的 SIFT 描述子个数导致的。阈值 θ_1 对函数矩阵的影响很小，当它从 0.005 增加到 0.02 时，可以稍微提高性能。详细的对索引稀疏度分析将在章节2.5.6中介绍。

此外，按照算法1 所述的参数设置，优化目标可以快速收敛到最优值。图. 2.7 展示了优化算法的收敛情况，其中三条曲线分别记录了每一步的误差，如下所示：

$$\begin{aligned} \mathrm{Err}_1 &= ||\mathbf{X}_v - \mathbf{X}_v\mathbf{Z}_v - \mathbf{E}_v||_\infty \\ \mathrm{Err}_2 &= ||\mathbf{Z}_v - \mathbf{G}_v||_\infty \\ \mathrm{Err}_3 &= ||\mathbf{Z}_v - \mathbf{M}_v||_\infty. \end{aligned} \tag{2.23}$$

2.5.6 讨论与分析

为了充分地解释和理解多线性多索引融合算法的特性，本小节进行了进一步的分析和实验。



(a) Ukbench 上性能变化 **(b) Holidays 上性能变化**

(c) Ukbench 上索引稀疏度变化 **(d) Holidays 上索引稀疏度变化**

图 2.6 θ_2 对性能和索引稀疏度的影响。

Figure 2.6 Influence of threshold θ_2 on the sparseness and retrieval accuracy.

扩展性分析：尽管多线性多视角特征融合算法复杂度较高，但正如之前所述，整个融合过程仅离线训练一次。在算法1的整个迭代过程中，只要选择合适的参数，矩阵求逆的运算仅需计算一次。算法的主要计算瓶颈在于求解子问题 $\mathcal{Z}$，但求解该问题等价于计算 $\frac{N-1}{2}$ 个矩阵的 SVD 分解，其中每个矩阵的维数是 $N \times V$。这种特殊的结构可以很容易被并行化，并将在今后的工作中进行进一步研究。总而言之，计算快速傅立叶变换及其逆运算需要 $\mathcal{O}(2N^2V\log(N))$，同时需要 $\mathcal{O}(N^2V^2)$ 计算矩阵 SVD。对于子问题 $\mathbf{E}$ 和 $\mathbf{M}_v$，每次迭代需要 $\mathcal{O}(N^2V)$ 求解这两个问题。由于 $\log(N) \gg V$，多线性多视角特征融合算法的复杂度为：

$$\mathcal{O}(TK(2N^2V\log(N))), \tag{2.24}$$

其中 K 表示优化过程中的迭代次数。实际上，T 通常取 $3-4$，K 取 $30-50$。表2.3给出了 MMF 算法及相对应的基准方法的运行时间，其中 #M 表示原始索引所需的运行时间，#A 表示由多线性多视角特征融合算法带来的额外时间开销。* 表示使用了多核并行检索。由于 MMF 在离线训练过程中，实质上是求解一个基于图的优化问题时，因此对于规模较大的数据集，离线上的训练时间增加较多。但在在线测试阶段，由于融合索引的稀疏性，多线性多索引融合算法带来的额外时间

表 2.3 MMF 运行时间分析。

Table 2.3 Quantitative analysis for timing of MMF.

MMF	离线时间		在线时间	
数据集	#M	#A	#M	#A
Holidays (HC-index)	-	+0.08h	37.54s	+0.51s.
UKbench (HC-index)	-	+3.15h	588.3*s	-87.6*s
Market-1501 (ResNet-50)	-	+7.85h	56.39*s	-4.19*s

开销在 Holiday 上占比很小。对于 UKbench 和 Market-1501 数据集，在线匹配的时间甚至降低了。

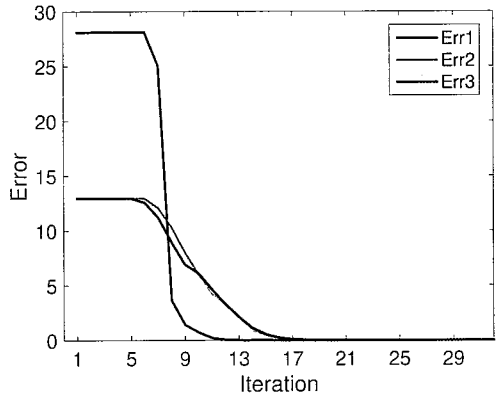


图 2.7 Holiday 数据集上 MMF 的收敛曲线。

Figure 2.7 Convergence curves of MMF on Holiday dataset.

与此同时，如图2.8所示，作用矩阵上的数值表示两个样本的相关性的强弱，图中的作用矩阵具有很清晰的分块结构，这一方面说明了样本依赖假设的合理性。另一方面，分块结构表明，整个数据集可以被进一步切分成不同的样本组，多索引融合算法可以在各个样本组上分别运行，从而进一步减少计算和内存开销，而不降低系统的性能，这将在今后的工作中进一步研究。

鲁棒性：算法还提取了 GIST 特征作为 UKbench 上的第 4 个特征索引，由 GIST 特征作为索引得到的 N-S 分数为 1.89。当算法反复融合迭代三次后，它的 N-S 得分提高至 3.24，对于其他特征索引，N-S 分数只降低了 0.01，得到了与不加入该特征进行融合的相似结果，具体结果如图2.9所示。这一现象表明了多线性多视角特征融合算法在面对较差特征时的鲁棒性。

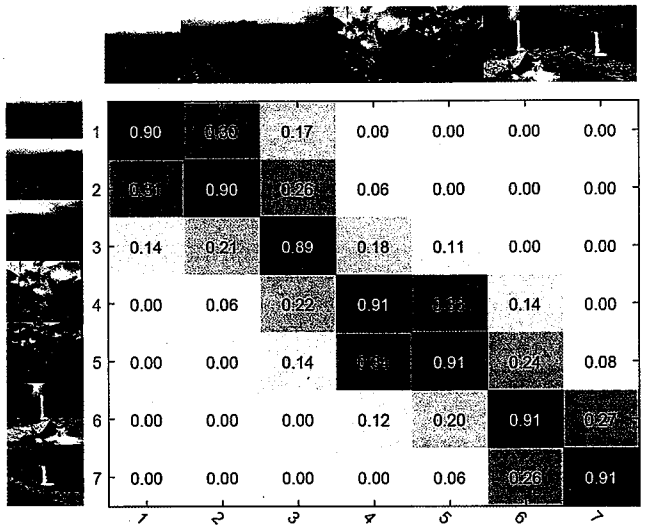


图 2.8 Holiday 数据集上 MMF 学习得到的作用矩阵 Z。

Figure 2.8 Representative functional matrix Z of MMF learned on the Holidays datasets.

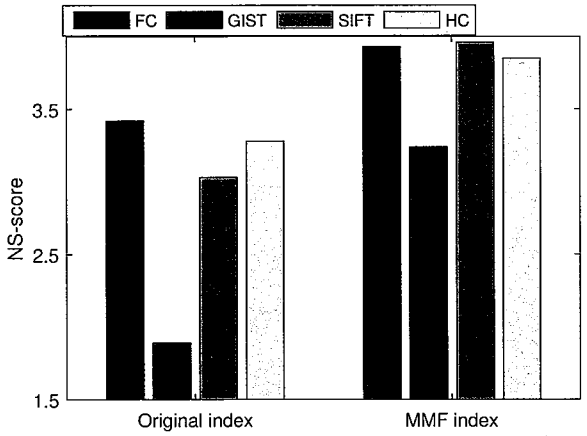


图 2.9 UKbench 数据集上 MMF 融合前后性能对比。

Figure 2.9 Comparison between the original index and the proposed MMF inedx in terms of accuracy on UKbench dataset.

作用矩阵 Z 的分析：为了评估作用矩阵的稀疏性在模型中的影响，本小节去掉了作用矩阵上的稀疏操作而进行实验，在 Market-1501 上，它的 Rank-1 accuracy 的下降了 5%。结果表明，作用矩阵的稀疏性不仅对多索引融合框架可扩展性有着重要影响，同时也对检索精度有着极大的促进作用。在 Market-1501 上的实验表明，学习得到的作用矩阵中约有 97% 的元素为零，这揭示了数据库中的子空间结构。此外，在每次迭代过程中，只有少数相关图像被更新，这也能尽可能少地破坏原始的特征索引。

噪声：图.2.8表明，由于作用矩阵 Z 上一些较小的值，所提出的方法将引入

额外的噪声。但这可能是多索引融合方法的一个共性问题。如图3.8所示，对于CIE 和 MMF 方法，Rank-1 accuracy 均是先上升再下降，其中 Rank-1 accuracy 是一个噪声敏感的评价指标，其下降表明多索引融合算法确实引入了噪声。为了减轻这种影响，CIE 只选择最为可能相关的近邻图像进行更新，而多线性多索引算法希望借助作用矩阵的稀疏性来抑制噪声。直觉上，CIE 将更好地解决噪声问题，但这可能会严重影响融合后的算法性能，因为 CIE 选择固定数量的近邻样本进行更新，而这种情形对于实际问题而言往往是不适用的（例如 Market-1501）。

局限性和未来的工作：虽然 MMF 取得了不错的匹配准确度，但是在这项工作中仍然存在一些局限性，在未来的工作中将进一步研究。首先在没有真实标签情况下，特别是当遇到判别性较差的特征索引（例如 GIST 索引）时，MMF 难以选择最终的索引。其次，在迭代过程中，特征索引的稀疏性不能得到直接的保证。稀疏操作会很大程度上破坏原始特征的表示，迫使相关图像趋于一致。对于这些局限性，可以利用先验知识和较大的 θ 来解决，但仍需进一步的研究。

2.6 本章小结

本章提出的多线性多视角特征融合方法继承了多索引融合技术的核心思想，能够实现索引层级多视角特征融合，从而不需要存储和计算额外的特征表达。与传统的特征融合方法相比，所提模型利用索引相关和视角相关假设，通过统一的张量空间，寻找到最优的作用矩阵，以此来更新索引矩阵和传播不同视角特征的相似关系，进而模型能够有效捕获不同特征表达之间的互补信息，提升图像检索的精度。在三个公开数据集上的实验表明，所提方法在检索精度上明显优于基准方法和其它对比方法，具有优异的泛化性能。此外，在线检索阶段的内存和计算开销也同样低于对比方法。通过自对比实验，验证了多线性多视角特征融合模型的鲁棒性和有效性。未来的研究将包括以下几个方面：1）t-SVD 的并行计算；2）最终索引选择方法；3）分组并行的可拓展化的多索引融合方法。

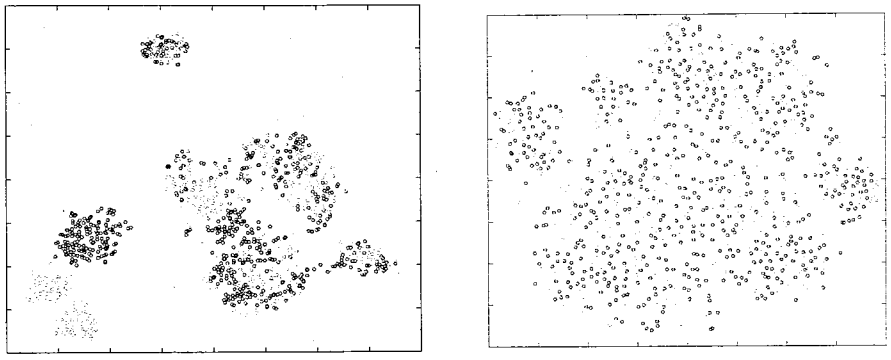
第 3 章　面向行人重识别的张量多视角非对称度量学习

3.1　引言

行人重识别任务 (Re-ID)[85; 58; 47; 74] 由于其在行人跟踪 [69]、行为分析 [59] 等智能安防应用中的巨大潜力，近年来受到学术界和工业界的广泛关注。典型的 Re-ID 任务是针对图像数据而言，从不同视角下拍摄获取的监控数据，经行人检测器 [94] 或人为标注得到行人图像，再通过以图搜图的方式，从大量的监控数据中检索和识别出目标人物图像。在这种背景下，用于行人重识别的行人图像，往往处于不同的视角、不同的光照和不同人体姿态的情形下，有较大的类内差异，较小的类间差距。此外，从监控摄像机捕捉到的监控数据通常分辨率较低，也给重识别带来了额外的困难。

为了应对这些问题，多年来研究人员尝试了多种方法 [77; 64; 54] ，其中，判别性的度量学习取得了巨大成功 [74; 61]。这类方法通常在训练数据上建立一个视角无关或视角相关的投影模型，通过投影矩阵学习判别性的度量。具体而言，视角无关模型对所有视角下的数据建立统一的映射模型，然后在投影空间上学习区分不同行人的度量 [52; 72]，而视角相关模型 [87; 67] 则显式地捕获视角之间的差异。如图3.1所示，蓝色和绿色的点分别代表不同的视角下样本的二维投影。子图 (a) 表示原始的特征空间，子图 (b) 表示本章所提算法的投影结果。可见，由于视角无关模型无法很好地对齐跨视角的特征分布（请参见图3.1获取更多详细信息），因此视角相关模型近年来更受推崇。

对于视角相关模型而言，一个主要问题在于训练样本的不足，特别是在视角急剧变化时，匹配的训练样本更加难以获取。因此，现有的方法通常采用矩阵正则化方法 [71; 87; 75; 67] 学习非对称度量，以此来描述视角差异，通过视角间的关联结构，提升模型的泛化性。该类方法的一个关键假设是，所有视角下的投影矩阵都是通过某种结构进行关联的，例如，低秩正则化约束 [87]、Bregman 差异约束 [75][71]。然而，这类视角相关方法仍然存在以下缺点，使得重识别精度不高：1）只关注成对跨视角投影矩阵之间的关联关系；2）在多视角学习框架下，关联关系只在视角之间共享，忽略视角内部的模式挖掘；3）模型的复杂度随着监控网络规模的增大而增大。



(a) 原始空间结果　　(b) 所提方法投影结果

图 3.1 t-SNE 可视化特征分布对比图。

Figure 3.1 The visual comparison of feature distributions alignment by using t-SNE.

为了解决上述问题，本章提出了一种新的张量多视角非对称度量学习模型(t-MTL)。t-MTL 将每个视角下的行人重识别任务视为一个子任务[1]，通过一对多(one-vs-all) 的多分类学习框架，学习得到多个特定于视角的分类器。其中，一个分类器用于区分一个特定的行人，而某个视角下的所有分类器都被堆叠到特定于视角的投影矩阵中。为了提升模型的泛化能力，对于不同任务，t-MTL 假定同一行人在不同视角下的特征是由一个潜在的/公共的子空间生成的；而在特定的任务内部，t-MTL 假定同一视角下的不同身份行人应该共享某种相似的模式。通过一种新的低秩张量约束，t-MTL 可以捕捉到投影矩阵之间的高阶关联关系，从而使不同视角下的特征分布，在投影的公共空间中被很好地对齐（详见图.3.1）。该张量核范数的循环代数不仅可以捕获沿第三维，即不同任务之间相关信息，而且可以获取不同列，即任务内部的关联结构。这表明高阶关联不仅存在于不同的视角之间，而且存在于特定视角（特定任务）下的分类器内部。图3.2 展示了所提算法的流程。图中假定有三个不同的视角（学习任务），这些用于区分特定行人的线性分类器构成了某个视角下的投影矩阵（如灰色矩阵，蓝色矩阵和紫色矩阵）。通过将这些投影矩阵堆叠成张量结构，并对其施加正则化，所有视角下的线性分类器被同时优化，最后通过交替迭代的方式求解得到最优投影矩阵。在在线测试阶段，在给定查询样本及其视角类别的情况下，利用相应的投影矩阵，将查询样本和待匹配样本的原始特征投影到由分类器输出的公共空间。t-MTL 最终在公共空间计算余弦距离来进行排序，完成重识别。

[1]本章此后不区分视角和任务，二者等同。

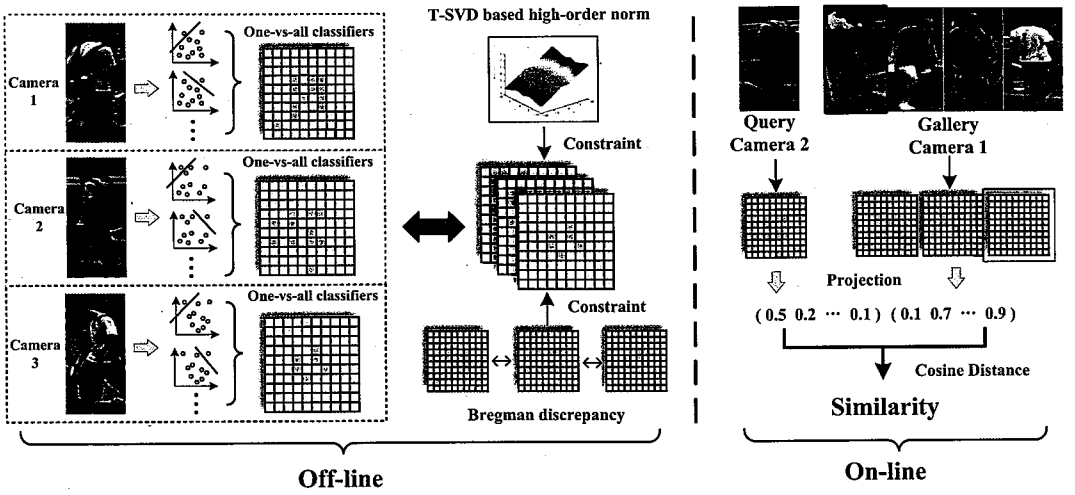


图 3.2 张量多视角学习框架图。

Figure 3.2 The framework of tensor multi-view learning.

本章的主要贡献概括如下：

• 本章提出了一种新的多视角度量学习模型，在有监督和无监督两种模式下，利用一个新的张量正则化约束来有效处理行人重识别问题，该模型不仅可以捕获不同任务之间的关联，还可以挖掘任务内部多分类器的相关信息。

• 本章提出了一种高效的优化算法，用于求解所提出的多视角度量学习模型，该优化算法具有较低的计算复杂度和理论收敛性保证。

• 本章提出的无监督多视角度量学习模型可以作为一种新的无监督基准方法，提升行人重识别特征判别能力。

• 多视角学习模型能够灵活地融合多种视觉特征。无论如何构造多视角张量，通过高阶约束，都可以有效地挖掘不同特征之间的一致性和互补性。

• 本章在几个基准数据集上对所提方法进行了评估，验证了方法的有效性。

本章接下来内容的组织结构如下：第3.2节介绍相关的工作；第3.3节首先详细地给出所提模型的动机，然后形式化的介绍模型表达，并给出求解该模型的优化算法，最后对所提方法进行了扩展；第3.4节介绍实验研究相关的内容，还进行了分析和讨论；最后，在第3.5节对本章工作进行总结并提出本章工作需要进行的后续研究。

3.2 相关工作

大多数行人重识别方法属于特征设计和有监督/无监督度量学习的范畴。下面简要回顾其中一些代表性方法优点和局限性。

3.2.1 行人重识别特征设计

特征设计类的方法 [61; 64] 往往希望通过设计鲁棒的特征来对抗极端的场景/视角/光照变换，为了实现这一目标，研究者们设计了多种手工特征，包括 LBP 描述子 [88]、颜色 [64]、基于局部特征的词袋模型 [47] 等，这些手工特征借助人类的先验知识，往往不需要相应场景信息，从而适用于大多数情况。Liao 等人 [61] 利用 Retinex 算法来克服场景变换的影响，Matsukawa 等人 [64] 采用分层的高斯分布来保留判别性信息。近年来，深度学习的兴起，给特征设计带来了新的机遇和挑战，对于充足的训练样本，研究者们提出了 IDE 网络 [47]、深度多任务学习网络 [58]、CAN[66] 以及其他的深度神经网络 [60]，通过提取网络的高层语义特征，获取判别性的行人图像表达。这些方法区别于网络结构 [54]、训练策略 [65] 和损失函数 [83]，产生了不同特征的提取方式。另一方面，很多深度学习方法从迁移的角度，尝试设计领域无关的特征，例如 JSTL[82]、Deep Transfer 模型 [73] 和 HIPHOP[62] 等方法。与此同时，大量的工作试图结合多种特征来提升匹配的准确性，例如 MHJLw[70] 利用共享和个体信息设计深度网络。DMVFL[96] 运用手工特征和深度学习之间的相互协作学习，有效提升行人重识别准确率。

3.2.2 行人重识别度量学习

除特征设计外，距离学习仍然是目标对象间相似性排序不可或缺的一环。这些方法借助度量学习 [79; 89]、排序学习 [72]、子空间学习 [61] 和深度学习 [60] 提供可靠的相似性度量。这些方法可以进一步划分为视角相关 [87] 与视角无关的距离度量方法 [52; 61]，代表性的视角无关方法包括 Metric Ensembles [72]、SCSP[55]、Null Space[52]、XQDA [61]、KISSME [74] 和 MFA [88] 等，通常而言，视角无关方法不考虑场景信息而建立一个统一的映射模型。这类方法由于受到视角差异的影响，重识别精度往往有限。但这些方法的计算复杂度通常较低，对于大规模的行人重识别问题具有较强的扩展性。

与之相反，视角相关模型要么为每对视角之间学习匹配模型，要么对每个视

角单独训练投影矩阵。前者的代表方法是 MtMCML[67]，它设计多个马氏距离度量来与监控网络相关联。然而，该方法的复杂度随着监控网络规模的增大而增大，在实际应用中受限颇多。而基于投影的度量学习框架近年来受到了更多的关注。Su 等人 [87] 提出了一个多任务学习框架来同时训练视角特定的分类器。Chen 等人 [75] 泛化马氏距离到非对称距离，克服场景变化带来的影响。这类方法都遵循一个基本假设，即不同视角下的投影矩阵虽然不同却相关，从而保证不同视角的信息通过投影矩阵进行传递，有效提升了模型的泛化性。

3.2.3 行人重识别无监督度量学习

不借助训练标签，而训练一个相似性度量模型具有更广泛的实际应用价值。为了实现这一目标，研究人员相继提出了大量的迁移学习算法 [63; 73; 82]，这类方法借助额外数据库获取的知识并进行迁移。在这一框架下，Peng 等人 [63] 借助多任务字典学习提出了一种跨数据库的迁移模型，Geng 等人 [73] 通过一个精心设计的深度结构和一个损失相关的 dropout 策略，尝试解决数据稀疏性问题。Xiao 等人 [82] 设计了一个领域相关的 dropout 策略，将不同域的数据在同一个深度框架下进行训练，取得了不错的效果。

除迁移学习外，很多方法尝试运用未标记的数据进行训练。Fan 等人 [65] 提出了一种贪婪的训练策略，不断地进行聚类和 fine-tuning，实现了无监督的度量学习。Kodirov 等人 [80] 将 L1 图拉普拉斯引入到字典学习中，同时学习目标的特征表达和判别性信息。Lisanti 等人 [81] 同样探索临近样本的附加信息来解决无监督行人重识别问题。Yu 等人 [71] 也借助 [75] 的思想，通过聚类过程提出一种非对称的无监督度量学习模型。

3.3 张量多视角非对称度量学习框架

本节将正式介绍所提出的张量多视角非对称度量学习框架。具体而言，假设 $\mathbb{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^N$ 为数量 N 的样本集合，其中 d 是特征向量的维数。典型的 Re-ID 度量学习目标是学习得到马氏距离 (Mahalanobis Distance)，通过马氏距离来匹配多个视角下的行人目标。形式上，给定任意两个样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$，它们之间

的距离定义为：

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \mathbf{M} (\boldsymbol{x}_i - \boldsymbol{x}_j)} \\ = ||\mathbf{U}^T \boldsymbol{x}_i - \mathbf{U}^T \boldsymbol{x}_j||_2, \tag{3.1}$$

其中 $\mathbf{M} = \mathbf{U}\mathbf{U}^T$。通过计算查询行人图像 (Query) 和待匹配行人图像特征向量之间的距离，对待匹配库 (Gallery) 图像进行排序，最终完成识别匹配。然而，由于对称模型（公式 (3.1)）对所有视角样本进行统一的映射，因此当遇到视角急剧变化的情况时，对称模型的性能往往不太理想。

因此，考虑一个更一般的情形，即假设有 $V \geq 2$ 个视角采集图像，它们具有显著的视角差异。令 $\mathbf{X}^{(v)} = [\boldsymbol{x}_1^{(v)}, \boldsymbol{x}_2^{(v)}, ..., \boldsymbol{x}_{n_v}^{(v)}] \in \mathbb{R}^{d \times n_v}$ 表示第 v 个视角下，行人图像提取到的特征矩阵，其中 n_v 表示该视角下的样本数目。值得注意的是，所提算法并不假设不同视角下的训练样本数目相等，这种设定适用于大多数 Re-ID 应用。假设存在有 V 个视角特定的投影矩阵 $\mathbf{U}^{(v)}, v = 1, 2, \ldots, V$，这些投影将每个原始特征向量转换到一个潜在公共空间，在这个公共空间下，每副行人图像都可以用投影特征来表示。因此，给定任意两个样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$，它们之间的距离被重新定义为：

$$d(\boldsymbol{x}_i^{(v_1)}, \boldsymbol{x}_j^{(v_2)}) = ||\mathbf{U}^{(v_1)T} \boldsymbol{x}_i^{(v_1)} - \mathbf{U}^{(v_2)T} \boldsymbol{x}_j^{(v_2)}||_2. \tag{3.2}$$

通过这种方式，不同的投影矩阵将在不同视角下特征分布进行对齐，建模视角之间的差异，从而将对称度量（公式 (3.1)）推广到非对称度量（公式 (3.2)）。因此，在潜在公共空间中计算的距离更适合于行人匹配，提升性能。

为了学习得到适合的投影矩阵，本章将行人重识别任务视为一种一对多的多分类学习模型。形式上，给定 V 个分类器集合，其中每一个分类器用于区分一个特定的行人。令 $\mathbb{U} = \{\mathbf{U} \in \mathbb{R}^{d \times C}\}_{i=1}^{V}$ 表示这些分类器集合，其中 $\mathbf{U}^{(v)}$ 表示第 v 个视角下的投影矩阵，它的列，即 $\mathbf{U}_i^{(v)}$，表示第 v 个视角下用于区分第 i 个行人的分类器，C 表示训练的类别数目。分类器、投影矩阵和任务之间的关系展示在图3.3。其中每一个投影矩阵表示一个视角相关的学习任务，矩阵的每一列表示一个特定于行人的分类器，任务内和任务间的关联关系都集成在所构建的张量结构中。

因此，给定训练标签 $\mathbf{Y}^{(v)} \in (0, 1)^{C \times n_v}$，其中 $\mathbf{Y}^{(v)}$ 的第 c 个维度用于区分它

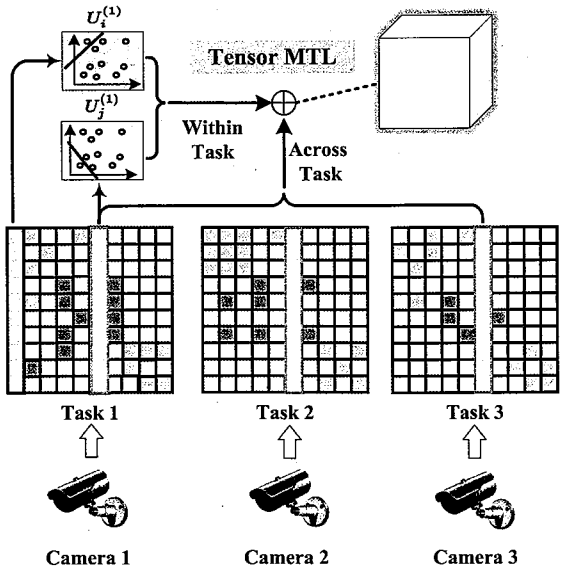


图 3.3 张量多视角学习框架下的分类器、投影矩阵和任务之间的关系。

Figure 3.3 The relationship among classifier, projection matrix and task under the tensor multi-task learning framework.

是否属于第 c 个类别，一个通用分类模型由下式给出：

$$\mathbf{U}^{(v)*} = \text{argmin}\mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)}) + \lambda\mathfrak{R}(\mathbf{U}^{(v)}), \tag{3.3}$$

其中 $\mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)})$ 表示分类项，$\mathfrak{R}(\mathbf{U}^{(v)})$ 表示分类器的正则项。λ 是一个超参数用于平衡分类和正则项。

然而，由于训练样本的有限，学习到的分类器容易出现过拟合现象。因此，正则化项 $\mathfrak{R}(\mathbf{U}^{(v)})$ 将在学习过程中起到关键作用。此外，从迁移学习的角度来看，一个视角下获取的知识如果被其他视角下的分类器加以利用，可以进一步提高分类器的识别和泛化能力。这促使本章采用基于张量正则化的多任务（MTL）框架来共同训练分类器。值得注意的是，典型的多任务学习模型 [87] 只允许任务间的知识共享（即知识只允许任务间而不在任务内部传递），这使得从一个视角下学习到的视角知识不能被同个视角下其他分类器学习得到，也就是说，以往的研究主要集中在成对任务之间的非对称度量学习上，而忽略了视角和类别之间的高阶信息。在下面，本章将首先介绍所提出的的有监督/无监督张量多任务学习（t-MTL）模型，该模型能够捕获任务之间和特定任务内的高阶关联关系，从而有效提升识别精度；然后再介绍相应的优化算法及其相应的多特征扩展方法。

3.3.1 有监督张量多视角度量学习模型

在多任务学习中，一个关键的假设是指，不同任务下的分类器是通过一定的结构相互关联，从而知识可以在任务之间互相传递。为此，本章提出了一种新的张量结构用于行人重识别。形式上，模型定义如下：

$$\min_{\mathbf{U}^{(v)}} \sum_{v} \mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)}) + \alpha \|\mathcal{U}\|_{\circledast} + \beta \sum_{i \neq j} \|\mathbf{U}^{(i)} - \mathbf{U}^{(j)}\|_F^2 \tag{3.4}$$

其中，α 和 β 表示超参数。$\mathcal{U} = \Phi(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(V)}) \in \mathbb{R}^{d \times C \times V}$ 是一个张量结构，该张量通过将投影矩阵 $\mathrm{U}^{(v)}$ 沿着第三方向堆叠得到，换言之，张量的 $\mathcal{U}$ 的第 v 正向切片 (frontal slices) 表示第 v 个视角下的投影矩阵。$\|\cdot\|_{\circledast}$ 表示张量低秩约束 [27; 26]，其定义如下所示：

$$\begin{aligned} \|\mathcal{U}\|_{\circledast} &= \|\mathrm{bcirc}(\mathcal{U})\|_* \\ &= \| \begin{bmatrix} \mathbf{U}^{(1)} & \mathbf{U}^{(\mathrm{V})} & \cdots & \mathbf{U}^{(2)} \\ \mathbf{U}^{(2)} & \mathbf{U}^{(1)} & \cdots & \mathbf{U}^{(3)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}^{\mathrm{V}} & \mathbf{U}^{(\mathrm{V}-1)} & \cdots & \mathbf{U}^{(1)} \end{bmatrix} \|_*. \end{aligned} \tag{3.5}$$

该低秩假设使得模型通过比较每个投影矩阵中的每一列（任务内），以及沿第三维度中的每一个正向切片（任务间）来捕获高阶关联关系。具体而言，通过比较正向切片（即 $\mathbf{U}^{(i)}$）的每一列，来捕获不同视角下，对应于同一行人分类器间的关联关系；通过比较正向切片的每一行，来捕获同一视角下，对应于不同行人分类器间的相似模式。因此，通过该假设，模型具有更佳的识别性能以及更强的泛化能力。

除此之外，分类项 $\mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)})$ 可以是任意光滑的凸函数，用于衡量真实标签和预测之间的差异。不失一般性，分类项由均方误差给出：

$$\mathfrak{L}(\mathbf{U}^{(v)}, \mathbf{X}^{(v)}, \mathbf{Y}^{(v)}) = \|\mathbf{U}^{(v)T}\mathbf{X}^{(v)} - \mathbf{Y}^{(v)}\|_F^2. \tag{3.6}$$

通过这种方式，$\mathbf{U}^v$ 矩阵的每一列对应于训练集中的每一个类别，因此，$\mathbf{U}^v$ 将特征向量转换为标签空间而替代原有的特征空间。尽管训练集和测试集中包含的类别是不完全重叠，但是采用这种方式主要有以下四方面的优势：1）不需要额外的映射，可以直接计算变换空间中的分类损失；2）不需要假定测试集中

的类别数，模型可以将特征空间直接映射到二值训练标签空间，由于测试集中未知的类别可以用其他训练类别来近似表示，在标签空间的样本表示同样也具有判别性；3）由于标签空间具有明确的物理意义，不需要再额外区分视角内和视角间的情况 [75]，这样可以显著减少计算量；4）在多任务分类框架下，通过张量结构可以更深入地探索任务之间和任务内部的一致信息。模型还引入了 Bregman 差异 $\sum_{i\neq j}\|\mathbf{U}^{(i)}-\mathbf{U}^{(j)}\|_F^2$ 来保证不同视角下投影矩阵的一致性，进而所提出的多任务学习模型可以在对称模型和非对称模型之间灵活切换。

3.3.2 优化过程

该优化问题（公式（3.4））似乎难以解决，不仅因为 $\mathcal{U}$ 上的张量低秩范数，而且还由于 Bregman 差异约束。为了便于求解，本小节首先构造分块矩阵，以更紧凑的形式重写公式（3.4）：

$$\tilde{\mathbf{U}}=[\mathbf{U}^{(1)};\mathbf{U}^{(2)};\ldots;\mathbf{U}^{(V)}], \tag{3.7}$$

$$\tilde{\mathbf{Y}}=[\mathbf{Y}^{(1)},\mathbf{Y}^{(2)},\ldots,\mathbf{Y}^{(V)}]. \tag{3.8}$$

因此分类项，即公式 (3.6) 可以被重新定义为：

$$\sum_{v}||\mathbf{U}^{(v)T}\mathbf{X}^{(v)}-\mathbf{Y}^{(v)}||_F^2=||\tilde{\mathbf{U}}^T\tilde{\mathbf{X}}-\tilde{\mathbf{Y}}||_F^2 \tag{3.9}$$

其中，$\tilde{\mathbf{X}}$ 是由特征矩阵拼接而成：

$$\tilde{\mathbf{X}}=\begin{bmatrix}\mathbf{X}^{(1)} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \mathbf{X}^{(2)} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}^{(V)}\end{bmatrix}. \tag{3.10}$$

此外，本章还定义了块矩阵 $\mathbf{M}$：

$$\mathbf{M}=\begin{bmatrix}(V-1)\mathbf{I} & -\mathbf{I} & \cdots & -\mathbf{I}\\ -\mathbf{I} & (V-1)\mathbf{I} & \cdots & -\mathbf{I}\\ \vdots & \vdots & \ddots & \vdots\\ -\mathbf{I} & -\mathbf{I} & \cdots & (V-1)\mathbf{I}\end{bmatrix}, \tag{3.11}$$

其中 $\mathbf{I}\in\mathbb{R}^{d\times d}$ 表示单位矩阵。根据这些定义，Bregman 差异约束可以被写作：

$$\sum_{i\neq j}\|\mathbf{U}^{(i)}-\mathbf{U}^{(j)}\|_F^2=tr(\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}}). \tag{3.12}$$

因此，模型可以由分块矩阵重新定义为：

$$\min_{\tilde{\mathbf{U}}} ||\tilde{\mathbf{U}}^T\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}||_F^2 + \alpha||\mathcal{U}||_{\circledast} + \beta tr(\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}}). \tag{3.13}$$

上述优化问题可以运用增广拉格朗日乘子 (ALM)[28] 求解。但采用交替方向最小化策略，需要使目标函数分离。故引入辅助张量变量 $\mathcal{G}$，将优化问题转化为以下无约束问题：

$$\begin{aligned}\mathcal{L}(\tilde{\mathbf{U}};\mathcal{G}) =& ||\tilde{\mathbf{U}}^T\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}||_F^2 + \alpha||\mathcal{G}||_{\circledast} + \beta tr(\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}}) \\ &+ \langle\mathcal{W},\mathcal{U} - \mathcal{G}\rangle + \frac{\rho}{2}||\mathcal{U} - \mathcal{G}||_F^2\end{aligned} \tag{3.14}$$

其中张量 $\mathcal{W}$ 表示拉格朗日乘子，ρ 表示惩罚因子。上述无约束问题可以划分为两个独立子步骤，采用交替求解。

$\tilde{\mathbf{U}}$-子问题: 当张量 $\mathcal{G}$ 固定时，则原问题转换为：

$$\begin{aligned}\tilde{\mathbf{U}}^* =& \mathbf{argmin}_{\tilde{\mathbf{U}}}||\tilde{\mathbf{U}}^T\tilde{\mathbf{X}} - \tilde{\mathbf{Y}}||_F^2 + \beta tr(\tilde{\mathbf{U}}^T\mathbf{M}\tilde{\mathbf{U}}) \\ &+ < \tilde{\mathbf{W}}, \tilde{\mathbf{U}} - \tilde{\mathbf{G}} > + \frac{\rho}{2}||\tilde{\mathbf{U}} - \tilde{\mathbf{G}}||_F^2\end{aligned} \tag{3.15}$$

其中，$\tilde{\mathbf{W}} = [\mathbf{W}^{(1)};\mathbf{W}^{(2)};\ldots;\mathbf{W}^{(V)}]$ 和 $\tilde{\mathbf{G}} = [\mathbf{G}^{(1)};\mathbf{G}^{(2)};\ldots;\mathbf{G}^{(V)}]$ 是块矩阵，$\mathbf{G}^{(v)} = \Phi_v^{-1}(\mathcal{G})$ 和 $\mathbf{W}^{(v)} = \Phi_v^{-1}(\mathcal{W})$，而 Φ_v^{-1} 是 Φ 的逆操作，表示切取张量的第 v 个正向切片。由于闭式解的存在，只需将公式（3.15）的导数置为 0，即可得到子问题的最优解：

$$\tilde{\mathbf{U}}^{T*} = (\tilde{\mathbf{Y}}\tilde{\mathbf{X}}^T + \rho\tilde{\mathbf{G}}^T - \tilde{\mathbf{W}}^T)(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T + \beta\mathbf{M} + \rho\tilde{\mathbf{I}})^{-1} \tag{3.16}$$

其中，$\tilde{\mathbf{I}} \in \mathbb{R}^{Vd\times Vd}$ 表示单位阵。

$\mathcal{G}$-子问题： 当 $\tilde{\mathbf{U}}$ 固定时，求解公式（3.14）等价于最小化下述问题：

$$\mathcal{G}^* = \mathrm{argmin}_{\mathcal{G}}\alpha(||\mathcal{G}||_{\circledast} + \frac{\rho}{2\alpha}||\mathcal{G} - (\mathcal{U} + \frac{1}{\rho}\mathcal{W})||_F^2). \tag{3.17}$$

上式又可以写作：

$$\min_{\mathcal{G}} \tau||\mathcal{G}||_{\circledast} + \frac{1}{2}||\mathcal{G} - \mathcal{F}||_F^2 \tag{3.18}$$

其中 $\tau = \frac{\alpha}{\rho}$，$\mathcal{F} = (\mathcal{U} + \frac{1}{\rho}\mathcal{W})$。该问题的最优解可以由下面定理给出。

定理 3.1. *对于任意 $\tau > 0$ 以及 $\mathcal{G},\mathcal{F} \in \mathbb{R}^{n_1\times n_2\times n_3}$, 给定优化问题：*

$$\min_{\mathcal{G}} \tau||\mathcal{G}||_{\circledast} + \frac{1}{2}||\mathcal{G} - \mathcal{F}||_F^2 \tag{3.19}$$

则该优化问题的全局最优解可以由张量管收缩算符得到：

$$\mathcal{G} = \mathcal{C}_{n_3\tau}(\mathcal{F}) = \mathcal{U} * \mathcal{C}_{n_3\tau}(\mathcal{S}) * \mathcal{V}^{\mathrm{T}}, \tag{3.20}$$

其中 $\mathcal{F} = \sum_{i=1}^{\min(n_1,n_2)} \mathcal{U}(:,i,:) * \mathcal{S}(i,i,:) * \mathcal{V}(:,i,:)^{\mathrm{T}}$，$\mathcal{C}_{n_3\tau}(\mathcal{S}) = \mathcal{S} * \mathcal{J}$，而 $\mathcal{J}$ 是一个 $n_1 \times n_2 \times n_3$ f-对角张量，它正向切片在傅立叶域的对角元素为 $\mathcal{J}_f(i,i,j) = (1 - \frac{n_3\tau}{\mathcal{S}_f^{(j)}(i,i)})_+$.

算法 2 张量核范数优化算法

1: **输入**：张量 $\mathcal{F} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ 和常数 $\tau > 0$

2: **输出**：张量 $\mathcal{G}$

3: $\mathcal{F}_f = \mathrm{fft}(\mathcal{F}, [\,], 3)$, $\tau' = n_3\tau$

4: **for** $i \leftarrow 1 : n_3$ **do**

5: $\quad [\mathcal{U}_f^{(j)}, \mathcal{S}_f^{(j)}, \mathcal{V}_f^{(j)}] = \mathrm{SVD}(\mathcal{F}_f^{(j)})$

6: $\quad \mathcal{J}_f^{(j)} = \mathrm{diag}\{(1 - \frac{\tau'}{\mathcal{S}_f^{(j)}(i,i)})_+\}, \quad i = 1, \ldots, \min(n_1, n_2)$

7: $\quad \mathcal{S}_{f,\tau'}^{(j)} = \mathcal{S}_f^{(j)} \mathcal{J}_J^{(j)}$

8: $\quad \mathcal{G}_f^{(j)} = \mathcal{U}_f^{(j)} \mathcal{S}_{f,\tau'}^{(j)} \mathcal{V}_f^{(j)^{\mathrm{T}}}$

9: **end for**

10: $\mathcal{G} = \mathrm{ifft}(\mathcal{G}_f, [\,], 3)$

证明. 在傅立叶域，原优化问题 (3.13) 可以写作：

$$\min_{\mathcal{G}_f} \tau \| \begin{bmatrix} \mathcal{G}_f^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathcal{G}_f^{(2)} & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathcal{G}_f^{(V)} \end{bmatrix} \|_* + \frac{1}{2V} \|\mathcal{G}_f - \mathcal{F}_f\|_F^2 \tag{3.21}$$

其中 $\mathcal{G}_f$ 表示 $\mathcal{G}$ 沿第三维进行快速傅立叶变换得到的张量，而 $\mathcal{G}_f^{(i)}$ 表示 $\mathcal{G}$ 的第 i 个正向切片[2]. 由于块对角矩阵结构，该张量优化问题可以被分解为在傅立叶域的 V 个独立的矩阵子问题：

$$\mathcal{G}_f^{(i)*} = \mathrm{argmin}_{\mathcal{G}_f^{(i)}} \tau' \|\mathcal{G}_f^{(i)}\|_* + \frac{1}{2} \|\mathcal{G}_f^{(i)} - \mathcal{F}_f^{(i)}\|_F^2, \tag{3.22}$$

[2]Matlab 符号为 $\mathcal{G}_f = \mathrm{fft}(\mathcal{G}, [], 3)$ 和 $\mathcal{G}_f^{(i)} = \mathcal{G}_f(:,:,i)$

其中 $\tau' = V\tau$ 且 $i = 1, 2, \ldots, V$。上述优化问题实质上为基于 F 范数的矩阵低秩优化问题。该问题可以通过在 $\mathcal{F}_f^{(i)}$ 上执行 SVD 分解，然后对其奇异值进行软阈值操作求解：

$$\mathcal{G}_f^{(i)*} = \mathcal{M}_f^{(i)} \mathcal{S}_{f,\tau'}^{(i)} \mathcal{N}_f^{(i)} \tag{3.23}$$

其中 $\mathcal{G}_f^{(i)} = \mathcal{M}_f^{(i)} \mathcal{S}_f^{(i)} \mathcal{N}_f^{(i)}$ 是标准的矩阵 SVD 分解，且 $\mathcal{S}_{f,\tau'}^{(i)}$ 表示 $diag\{(\mathcal{S}_f^{(i)}(i,i) - \tau')_+\}$。最后通过快速傅立叶逆运算，可以获得公式（3.22）的全局最优解：

$$\mathcal{G} = \text{ifft}(\mathcal{G}_f, [], 3) \tag{3.24}$$

□

算法 3 张量多多视角非对称度量学习优化算法

1: **输入**：特征向量 $\mathbf{X}_v$, $v = 1, 2, \ldots, V$，标签 $\mathbf{Y}_v$, $v = 1, 2, \ldots, V, \alpha > 0, \beta > 0$
2: **输出**：分类器 $\mathbf{U}_v$, $v = 1, 2, \ldots, V$
3: **初始化**：$\mathbf{U}_v = \mathbf{0}; \mathcal{G} = \mathcal{W} = \mathbf{0}; \rho = 10^{-5}, \rho_{\max} = 10^{10}$
4: **构造**：$\tilde{\mathbf{U}}$, $\tilde{\mathbf{X}}$, $\tilde{\mathbf{Y}}$, $\mathbf{M}$
5: **while** 未满足收敛条件 **do**
6: 更新 $\tilde{\mathbf{U}}$ 通过公式 (3.16)
7: 拼接 $\tilde{\mathbf{U}}$ 得到 $\mathcal{U}$
8: 更新 $\mathcal{G}$ 通过算法 2
9: 更新拉格朗日乘子 $\mathcal{W}$ 通过公式 (3.25)
10: 拼接 $\mathcal{G}$ 和 $\mathcal{W}$ 得到 $\tilde{\mathbf{G}}$ 和 $\tilde{\mathbf{W}}$ 通过
11: 更新参数 ρ: $\rho = \min(\eta\rho, \rho_{\max})$
12: **end while**
13: 拆分 $\tilde{\mathbf{U}}$ 得到 $\mathbf{U}_v$, $v = 1, 2, \ldots, V$
14: **返回**分类器 $\mathbf{U}_v$, $v = 1, 2, \ldots, V$

算法 2总结了求解张量低秩优化问题的流程。此外，拉格朗日乘子 $\mathcal{W}$ 也需按如下方式更新：

$$\mathcal{W}^* = \mathcal{W} + \rho(\mathcal{U} - \mathcal{G}) \tag{3.25}$$

上述两个步骤不断交替优化，直至满足收敛条件。整个算法流程总结在算法 3。与此同时，根据 [91]，有以下定理保证了优化算法的收敛性。

定理 3.2. *由算法 3每一步产生的序列 $(\mathcal{G}, \tilde{\mathbf{U}})$，可以收敛到一个聚点。不仅如此，该聚点是优化问题（3.13）的全局最优解。*

实际上，所提方法在实际应用中表现良好且收敛迅速，本章将所有数据集上的最大迭代次数固定为 30，以保证运行效率。

3.3.3 无监督张量多视角度量学习

前面的小节介绍了如何运用 t-MTL 模型解决有监督的行人重识别问题。在该框架下，模型可以学习到合适的投影矩阵，将不同视角下的特征映射到统一的潜在空间，从而提升多个视角下匹配和识别个体的准确率。然而，在实际应用中，并不总是能够保证有足够的行人标签进行训练。因此一个直观的替代方案是充分利用易于获取的未标记数据进行训练。但是，在无监督的环境下，由于缺乏标签数据来指导模型区分相似个体，训练模型变得更为困难。为此，基于 [75]，本小节提出一种新的无监督 t-MTL 模型，该模型利用伪标签替代样本的真实标签进行训练。具体而言，无监督 t-MTL 模型首先通过聚类的方式（例如 k-means clustering），为每个训练样本生成一个伪标签。然后，类似于有监督 t-MTL 方法，最小化以下目标函数以获得投影矩阵：

$$\min_{\mathbf{U}^{(v)}} \sum_{v} ||\mathbf{U}^{(v)T}\mathbf{X}^{(v)} - \mathbf{P}||_F^2 + \alpha\|\mathcal{U}\|_{\circledast} + \beta \sum_{i \neq j} \|\mathbf{U}^{(i)} - \mathbf{U}^{(j)}\|_F^2 \tag{3.26}$$

其中 $\mathbf{P} \in \mathbb{R}^{C \times K}$ 表示训练样本的伪标签，K 表示聚类数目。尽管伪标签相对真实标签来说并不准确，但是由于张量结构，无监督 t-MTL 模型可以利用任务内/任务间共享的知识来提高泛化能力。同时，与 [71] 不同，无监督 t-MTL 模型仅需运行一次聚类过程，从而提升算法效率。

3.3.4 多特征融合张量多视角学习

现有的行人重识别方法通常使用多种视觉特征，通过串联特征向量的方式来提升性能。然而，这种级联操作往往忽略了不同特征之间的互补信息，进而影响性能。

由于提出的张量结构，t-MTL 可以进一步挖掘多种特征之间的共享信息。通过叠加相应的特征投影矩阵，t-MTL 可以很容易与多种特征学习结合。假设每个训练样本都有 L 种特征表达，令 $\mathbf{U}_{(l)}^{(v)}$ 表示在第 v 个视角下的第 l 种特征投影矩阵，$\{\mathbf{U}_{(l)}^{(v)}\}_{l=1}^{L}$ 可以隐式地挖掘隐藏在不同视觉表示中的共享信息。在实际中，多

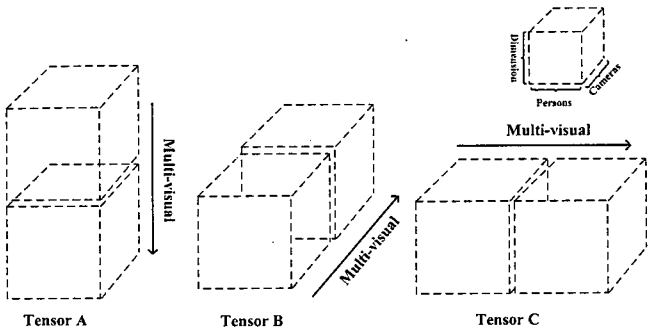


图 3.4 多特征投影矩阵构造的张量结构。

Figure 3.4 The tensor structure constructed by stacking the multi-feature classifiers.

特征融合 t-MTL 模型考虑了三种方法来构造多任务张量，如图所示3.3，分别表示为$\mathcal{U}_A$，$\mathcal{U}_B$ 和$\mathcal{U}_C$。实际上，由于下面的定理，这三种构造方法是等价的。

定理 3.3. *高阶张量范数在张量 A，张量 B 和张量 C 上等价。*

证明. 根据高阶张量范数的定义，以及：

$$||\begin{bmatrix} \mathrm{A} & \mathrm{B} \end{bmatrix}||_* = ||\begin{bmatrix} \mathrm{A} \\ \mathrm{B} \end{bmatrix}||_* \tag{3.27}$$

有：

$$||\mathcal{U}_B||_{\circledast} = ||\begin{bmatrix} \mathbf{U}_{(1)}^{(1)} & \mathbf{U}_{(V)}^{(L)} & \cdots & \mathbf{U}_{(2)}^{(1)} \\ \mathbf{U}_{(2)}^{(1)} & \mathbf{U}_{(1)}^{(1)} & \cdots & \mathbf{U}_{(V)}^{(L)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}_{(V)}^{L} & \mathbf{U}_{(V)}^{(L-1)} & \cdots & \mathbf{U}_{(1)}^{(1)} \end{bmatrix}||_* \tag{3.28}$$

$$= ||\begin{bmatrix} \mathbf{U}_{(1)}^{(1)} & \mathbf{U}_{(V)}^{(1)} & \cdots & \mathbf{U}_{(2)}^{(1)} \\ \mathbf{U}_{(1)}^{(2)} & \mathbf{U}_{(V)}^{(2)} & \cdots & \mathbf{U}_{(2)}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{U}_{V}^{L} & \mathbf{U}_{(V)}^{(L-1)} & \cdots & \mathbf{U}_{(1)}^{(1)} \end{bmatrix}||_* = ||\mathcal{U}_A||_{\circledast} \tag{3.29}$$

根据上述等式，易知 $||\mathcal{U}_B||_{\circledast} = ||\mathcal{U}_C||_{\circledast}$，定理得证。 □

最终本章采用张量 A 的方式实现多特征融合 t-MTL 模型（见图.3.3）。与直接拼接特征向量不同，所提方法利用循环代数来比较每一种特征的分类器以获取一致性。因此，互补信息被隐式地嵌入到投影矩阵中，从而获得更好的性能。综上所述，无论如何构造多任务张量，通过高阶正则化约束，都能有效地挖掘不同特征之间的一致性和互补性。

3.3.5 在线识别测试

在在线测试阶段，给定视角为 $v_q \in [1,\ldots,V]$ 的查询样本 q，首先提取查询样本的多种特征 $\mathbf{x}(q)=[\mathbf{x}_{(1)}(q),\ldots,\mathbf{x}_{(L)}(q)]$。值得注意的是，对于 Re-ID 任务，训练集中包含的类别可以测试集中的类别不同。也就是说，一旦训练获得了 v_q 视角下的投影矩阵 $\mathbf{U}_{(l)}^{(v_q)}, l=1,\ldots,L$，可以直接利用投影矩阵，将提取的特征转换到潜在公共空间，不管它的标签是否出现在在训练集中，即：

$$\boldsymbol{x}'(q)=[\mathbf{U}_{(1)}^{(v_q)T}\boldsymbol{x}_{(1)}(q),\mathbf{U}_{(2)}^{(v_q)T}\boldsymbol{x}_{(2)}(q),\ldots,\mathbf{U}_{(L)}^{(v_q)T}\boldsymbol{x}_{(L)}(q)]. \tag{3.30}$$

投影后的特征被进一步归一化处理。对于待查询库中的样本，模型也将其原始特征向量转化到所需的公共空间。根据非对称模型（公式（3.2）），利用余弦距离来度量这两个特征之间的相似性。最后，对待查询库中的样本进行排序，完成重识别任务。

3.4 实验结果和分析

在本章的实验中，将对所提方法的有效性进行全面评估。所有实验都是在一台上工作站上用 Matlab 实现的，该工作站配备了 Intel Xeon E5-2630@2.30GHz CPU、128GB RAM 和 TITANX GPU（12GB 缓存）。

为了清楚地说明实验方法，本节首先介绍实验设置，包括实验采用的数据集、特征表示方法和评价指标。然后本节给出 t-MTL 方法在有监督和无监督情形下的主要结果，其中实验分别比较了基准方法和其他具有代表性的相关方法。同时，本节还对 t-MTL 方法的相关变种进行了实验，通过集成多种特征来验证所提方法的有效性。最后，本节还分析了 t-MTL 方法的参数敏感性、收敛性、计算复杂度等特点，以及分析了无监督 t-MTL 的一些特性。

数据集：VIPeR[76] 数据集显示了两个视角之间的光照变化和姿态变化。该数据集共有 632 对样本，这些样本被随机分成两组，一组用于训练，另一组用于测试。测试时，将查询集中的每个样本与待匹配库中的所有样本进行匹配。CUHK01[77] 数据集是在校园环境中用两个视角捕获的。这个数据集包含 971 个人，每个人在每个视角下均有有两幅图像。用其中的 485 人进行训练，486 人进行测试。该数据集提供了两种评估方式：单次查询和多次查询模式。CHUK03[78] 包含 13164 张从 6 个监控摄像头拍摄的 1360 名行人的图像。除了手工裁剪的图

像外，还提供了行人检测器检测的样本。此外，模型还在 Market-1501[47] 数据集上进行了评测。

有监督模型的特征表示方法：为了获取图像的特征表示，本节使用了两种特征对图像进行描述，它们分别是 LOMO 特征 [61] 和 GOG 特征 [64])。此外，实验还同时使用 LOMO 和 GOG 特征来评估多特征下 t-MTL 算法的性能。对于 Market-1501 数据集，本节遵循由 [48] 提出的识别模型，训练两个 CNN 网络，即 CaffeNet[19] 和 ResNet50[47]，作为特征提取器。

无监督模型的特征表示方法：实验使用了 [82] 提出的基于深度学习的 JSTL 特征。JSTL 通过由卷积层、感知模块和全连接层组成的 GoogleNet 来实现 [95]，它最终生成一个 256-维的特征。对于 Market-1501 而言，实验采用原始的训练策略来提取该数据集上的特征。值得注意，原始 JSTL 的训练集中包含 VIPeR，违反了无监督设置。因此，本节训练了一个新的 JSTL 模型来提取特征，其训练过程并不使用 VIPeR 中的训练数据。

评价指标：实验使用两种常用的评价指标来评估性能，即累积匹配曲线 (CMC) 和平均精度 (mAP)。在有监督模式下，实验按照 [87; 75; 61; 64; 70] 的方式采用 CMC 对 VIPeR、CUHK01 和 CHUK03 进行评估，而对于 Market-1501[47]，实验则同时采用 CMC 和 mAP 进行性能评估。在无监督的模式下，实验遵循 [71] 的方式，对 VIPeR 数据集采用 Rank-1 accuracy 评估，而对于 Market-1501，则分别采用 Rank-1 accuracy 和 mAP。此外，在 VIPeR 和 CUHK01 中，实验取运行 10 次的均值作为最终的结果，而对于 CUHK03 数据集，报告的指标则是运行 20 次的均值结果。

参数设置：只有两个参数 α 和 β 需要被调节。有关这些参数的更多详细信息将在节3.4.2.1中详细讨论。其他对比方法的参数设置，均在原始文献建议的范围内进行优化，以显示最佳结果。

3.4.1 实验结果

3.4.1.1 有监督 t-MTL 实验结果

VIPeR. 对于 VIPeR 数据集，实验首先将 t-MTL 方法与基准模型进行比较，如表3.1所示。对于 LOMO、GOG 和多个特征，t-MTL 方法分别取得了 44.7%、50.6% 和 56.1% 的 Rank-1 accuracy。值得注意的是，t-MTL 方法通过利用多种特征的互补信息，较现有的非对称度量方法有显著提升，如 CVDCA[75] 和 MTL-

表 3.1 有监督 t-MTL 方法在 VIPeR 数据集上的识别精度。

Table 3.1 Performances of supervised t-MTL on VIPeR.

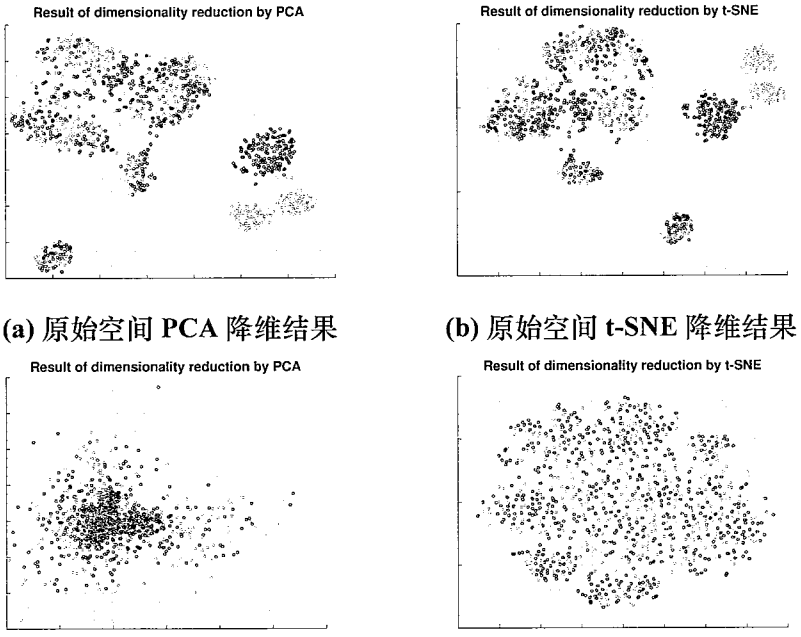
方法	特征	Rank-1	Rank-5	Rank-10	Rank-20
MTL-LORAE[87]	LBP+Attribute	42.3	72.2	81.6	89.6
CVDCA[75]	LOMO	43.7	74.1	84.8	91.9
XQDA[61]	LOMO	40.0	-	80.5	91.1
t-MTL($\alpha = 0$)	LOMO	31.2	64.5	79.1	89.6
t-MTL($\beta = 0$)	LOMO	32.6	58.0	69.7	80.2
t-MTL	LOMO	44.7	74.1	84.7	91.8
CVDCA[75]	GOG	50.4	78.8	88.0	94.5
XQDA[61]	GOG	49.7	-	88.6	94.5
t-MTL	GOG	50.6	78.4	87.3	93.3
CVDCA[75]	LOMO+GOG	49.5	78.6	87.7	94.1
XQDA[61]	LOMO+GOG	53.3	-	90.9	95.7
t-MTL (L1 loss)	LOMO+GOG	53.4	80.5	88.7	94.6
t-MTL (Concatenate)	LOMO+GOG	55.8	82.1	90.3	95.5
t-MTL (Tensor)	LOMO+GOG	56.1	82.1	90.3	95.5

LORAE[87]。此外，与对称度量学习方法 XQDA 相比，t-MTL 方法也在较大程度上提高了 Rank-1 acccuracy，对于不同的特征，分别提升了 4.7%、0.9% 和 2.8%。但是对于 Rank-5，Rank-10，Rank-20 的准确度，t-MTL 方法与 XQDA 方法性能类似。原因可能是 t-MTL 方法是基于分类的模型，而分类模型侧重于于区分个体，而不擅长提高排序性能。

为了验证 t-MTL 方法的有效性，本小节还对所提方法的变体，即移除张量低秩范数或 Bregman 差异约束进行了实验。对于 LOMO 特征，通过设置 $\alpha = 0$ 或 $\beta = 0$，分别移除张量约束或 Bregman 差异，t-MTL 方法在 Rank-1 accuracy 上均下降了约 10%（表中第四和第五行），这体现了正则项在多任务学习过程中的作用。与此同时，实验还对多个特征的不同张量构造方式进行了评估，例如张量 A 和张量 B，实验表明（表的最后两行），二者获得了几乎相同的 CMC 曲线。值得注意的是，这两种构造方式都显著优于 XQDA 方法，而 t-MTL 对于单独特征的识别准确率却与 XQDA 类似，这表明 t-MTL 方法能够捕获 LOMO 特征和 GOG 特征之间的互补信息，并进一步提升识别准确率。此外，本节对于分类项，还采用 L1 损失 $||\mathbf{U}^{(v)T}\mathbf{X}^{(v)} - \mathbf{Y}^{(v)}||_1$ 替代 L2 损失 $||\mathbf{U}^{(v)T}\mathbf{X}^{(v)} - \mathbf{Y}^{(v)}||_2^2$ 进行了实验，结果显示在表3.1中，可以发现，L2 损失较 L1 损失表现稍好，原因可能是一个

平滑的分类项有助于识别。

由于大量的算法在 VIPeR 数据集上进行了测试，因此，实验只比较了近五年或与 t-MTL 方法关系紧密的算法。如表3.2所示，所提方法与相关方法（包括域转移方法 [58]、多任务学习方法 [87]）取得了可比（甚至更好）的结果，并且也优于多特征融合方法 [70]。值得注意的是，SSM[86] 作为一种后处理方法，也取得了不错的识别准确率。可以预见，SSM 和 t-MTL 将相互受益，并将进一步提升性能。同时，DictRW[84] 的 Rank-10 和 Rank-20 accuracy 也超过了 t-MTL 方法。这是由于 DictRW 在模型中嵌入了三元组关系约束，进而提升 CMC 指标。此外，DictRW 采用了深度特征进行训练。相比之下，t-MTL 只采用手工设计的特征，但识别准确率更高，这验证了 t-MTL 方法的有效性。



(a) 原始空间 PCA 降维结果 **(b) 原始空间 t-SNE 降维结果**

(c) t-MTL 的 PCA 降维结果 **(d) t-MTL 的 t-SNE 降维结果**

图 3.5 ViPeR 数据集上的特征降维结果（LOMO 特征）。

Figure 3.5 Result of dimensionality reduction for ViPeR dataset (LOMO feature).

此外，对于 ViPeR 数据集，实验还通过 PCA 和 t-SNE 对特征进行了降维，可视化了原始特征空间和潜在公共空间中特征分布结果。如图所示，从两个视角获得的两个特征分布（蓝点和绿点）在原始特征空间中的重叠率相对较低，这表明在原始特征空间中，不同视角下的特征分布截然不同，这导致了识别准确率不高。但是，在经过 t-MTL 投影之后（参见图3.5），投影后的特征分布在潜在公共空间中被很好地对齐，并且数据点比以前更加分离，从而加大类间距离，提升识

表 3.2 有监督 t-MTL 方法在 VIPeR 数据集上与相关方法的对比

Table 3.2 Performances of supervised t-MTL on Viper with state-of-the art methods

方法	参考文献	Rank-1	Rank-5	Rank-10
MTL-LORAE[87]	ICCV2015	42.3	81.6	89.6
XQDA[61]	CVPR2015	40.0	80.5	91.1
MetricEnsemple[72]	CVPR2015	45.9	88.9	95.8
SSDAL[56]	ECCV2016	43.5	81.5	89.0
NULL[52]	CVPR2016	51.2	90.5	95.9
GOG[64]	CVPR2016	49.7	88.6	94.5
KCVDCA[75]	TCSVT2017	43.3	83.5	92.2
MHJLw[70]	TNNLS2017	45.4	84.0	92.5
SSM[86]	CVPR2017	53.7	91.5	96.1
MTDnet[58]	AAAI2017	47.5	82.6	-
DictRW[84]	IJCAI2017	55.7	91.5	96.7
t-MTL (Tensor)	-	56.1	90.3	95.5

别准确率。

Market-1501：对于 Market-1501 数据集，由于该数据集规模较大，有充足的训练样本，故基于深度学习的方法 [47] 取得了不错的性能。如表3.3所示，可以发现，通过挖掘任务间/任务内的相关性，t-MTL 方法对于训练好的深度特征，可以进一步提升性能。具体来说，在单次查询 (Single Query) 模式下，经过 t-MTL 映射得到的特征显著优于基准模型，对于 CaffeNet，t-MTL 方法对于 Rank-1 accuracy 和 mAP 分别取得了 2.82% 和 2.70% 的提升；对于 ResNet-50，t-MTL 方法则获得了 2.71% 和 1.98% 的提升。在多次查询 (Multiple Query) 模式下，t-MTL 也得到了类似的结果。与典型的度量学习方法相比，即 KISSME[74] 和 XQDA[61]，t-MTL 对于 Rank-1 accuracy 有显著提升，而对 mAP 的提升并不明显，这与之前的实验结果相类似。然而，当融合两种特性时，t-MTL 的结果略低于原始的 ResNet-50 特征识别准确率。造成这种异常现象原因可能是 CaffeNet 特征和 ResNet-50 特征属于同质特征，在性能上也存在巨大差异，因此不能挖掘互补信息提升性能。值得一提的是，SVDnet[68] 也采用了张量 SVD 投影，显著地提升了重识别性能，但其本质与 t-MTL 方法不同，SVDnet 对于网络的最后一个线性层，利用 SVD 分解产生正交层投影特征。与 SVDnet 相比，t-MTL 的目标是寻找一个低秩张量，以获得更好的泛化能力，其中 t-SVD 用于优化求解，两者可以相互使用。重排序

方法 [51] 与 t-MTL 模型也紧密相关，但这类方法侧重于将待查询库中的样本之间的关系嵌入到学习的度量中，而 t-MTL 旨在通过探索任务之间/任务内部的共享信息来学习合适的度量。

表 3.3 有监督 t-MTL 方法在 Market-1501 数据集上的识别精度。

Table 3.3 Performances of supervised t-MTL.

Market-1501	单次查询		多次查询	
方法	Rank-1	mAP	Rank-1	mAP
CaffeNet[47]	59.53	32.85	66.63	41.25
CaffeNet+ CVDCA[75]	59.80	35.69	-	-
CaffNet+XQDA[61]	62.00	37.55	70.28	46.78
CaffNet+KISSME[74]	61.02	37.72	69.86	45.34
CaffNet+t-MTL	62.35	35.60	71.38	44.68
ResNet50[47]	75.62	50.68	81.26	59.10
ResNet-50+CVDCA[75]	74.82	50.21	-	-
ResNet50+XQDA[61]	76.01	52.98	81.12	61.09
ResNet50+KISSME[74]	77.52	53.88	82.16	61.54
ResNet50+t-MTL	78.33	52.66	84.14	61.73

CUHK01：CUHK01 数据集上的主要结果展示在表 3.4上。t-MTL 的结果与在 VIPeR 上的结果类似，在单次查询 (single-shot) 和多次查询 (multi-shot) 模式下，对于 LOMO 特征，t-MTL 优于度量学习方法 XQDA [61]；对于 GOG 特征，t-MTL 表现与 XQDA 类似。但当融合两种特征时，在单次查询模式下，t-MTL 相较于 XQDA，取得了 6.5% 的 Rank-1 accuracy 提升，这一结果优于大多数方法在多次查询模式下的结果；在多次查询模式下，t-MTL 相较于 XQDA 也取得了 3.4% 的 Rank-1 accuracy 提升，提升相较于单次查询下略低，这是由于 t-MTL 在 GOG 特征下的性能较差。与最新的相关方法 [72; 75; 58; 52; 70] 相比（表3.4中的第一块），对于单种特征，t-MTL 取得了可比的重识别准确率，但当融合多种特征，t-MTL 显著优于这些方法。

CUHK03：由于 CUHK03 数据集的规模较大，实验首先采用主成分分析将 GOG 和 LOMO 特征的维数降到 1000 维。表3.5展示了 CUHK03 数据集上的主要结果，对 7 个有代表性的 Re-ID 方法 [78; 54; 53; 52; 86; 64; 61] 与 t-MTL 方法进行了比较。其中 XQDA[61] 取得了非常出色的结果，对于 GOG 特征，XQDA 显著优于 t-MTL 方法。在此基础上，SSM[86]，通过平滑所学习的度量，在所有

表 3.4 有监督 t-MTL 方法在 CUHK01 数据集上的识别精度。

Table 3.4 Performances of supervised t-MTL on CUHK01

CUHK-1	单次查询			多次查询		
方法	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
MetricEnsemple[72]	53.4	76.4	84.4	-	-	-
KCVDCA[75]	47.8	74.2	83.4	-	-	-
CVDCA+LOMO+GOG[75]	73.0	89.1	93.9	-	-	-
MTDnet[58]	-	-	-	78.5	96.5	97.5
Null+LOMO[52]	-	-	-	65.0	85.0	89.9
MHJLw[70]	-	-	-	64.5	-	91.1
XQDA+LOMO[61]	48.7	73.0	81.3	63.2	83.9	90.0
XQDA+GOG[64]	57.8	79.1	86.2	67.3	86.9	91.8
XQDA+LOMO+GOG	68.5	87.3	92.4	76.9	91.5	95.4
t-MTL+LOMO	50.1	73.8	81.3	64.4	84.9	90.3
t-MTL+GOG	58.0	78.6	85.1	66.0	85.1	90.1
t-MTL+LOMO+GOG	74.1	89.1	95.9	80.3	92.4	95.1

表 3.5 有监督 t-MTL 方法在 CUHK03 数据集上的识别精度。

Table 3.5 Performances of supervised t-MTL on CUHK03

方法	Rank-1	Rank-5	Rank-10	Rank-20
DeepReID[78]	19.9	49.0	64.3	-
S-LSTM[53]	57.3	80.1	88.3	
Null[52]	54.7	84.8	94.8	95.2
S-CNN[54]	61.8	80.9	88.3	-
SSM[86]	72.7	92.4	96.1	-
XQDA+LOMO[61]	46.3	78.9	88.6	94.3
XQDA+GOG[64]	64.0	88.6	94.2	97.6
CVDCA+LOMO+GOG[75]	59.6	86.6	93.9	97.3
XQDA+LOMO+GOG	68.1	90.2	95.0	98.0
t-MTL+LOMO	50.5	78.5	86.3	92.2
t-MTL+GOG	59.3	84.5	91.5	96.2
t-MTL+LOMO+GOG	66.5	88.3	93.3	97.0

的对比方法中取得了最优性能。然而，对于 LOMO 特征，t-MTL 略优于 XQDA，通过融合两种特征，t-MTL 取得了 66.5% 的 Rank-1 accuracy，与 XQDA 十分接近。值得注意的是，CVDCA[75] 在相对较小的数据集上表现良好，但对于规模较大的数据集，其性能变差。此外，由于训练样本充足，大多数深度学习方法，

如 S-LSTM[53]、S-CNN[54]，往往取得不错的识别准确率。

3.4.1.2 无监督 t-MTL 实验结果

本小节展示了无监督 t-MTL 方法在 VIPeR 和 Market-1501 数据集上的结果。**ViPeR**：在表 3.6中，实验采用了三种类型特征（即 LOMO，GOG 和 JSTL 特征）来对无监督 t-MTL 方法在 VIPeR 数据集上进行全面评估。首先，实验采用单一特征对所提方法进行性能评估。对于 LOMO，GOG 和 JSTL 特征，t-MTL 分别取得了 21.8%，25.3% 和 30.3% 的 Rank-1 accuracy，远超使用相同特征的欧几里得距离。此外，当 GOG 和 LOMO 融合时，t-MTL 方法的识别准确率进一步提升，取得了 28.6% 的 Rank-1 accuracy，在相同实验设置下比 CAMEL[71] 高出 2.1%，对于深度学习特征 JSTL，t-MTL 也略高于 CAMEL。

表 3.6 无监督 t-MTL 方法在 VIPeR 数据集上的识别精度。

Table 3.6 Performances of unsupervised t-MTL. Measured by rank-1 accuracies for VIPeR.

方法	特征	Rank-1	方法	特征	Rank-1
t-MTL	JSTL	31.8	l_2	JSTL	30.0
t-MTL	LOMO	21.8	l_2	LOMO	16.5
t-MTL	GOG	25.3	l_2	GOG	15.4
CAMEL[71]	LOMO	26.5	CAMEL[71]	JSTL	30.6
t-MTL	LOMO+GOG	28.6	t-MTL	JSTL	31.8

Market-1501：对于 Market-1501，如表3.7所示，在单次查询模式下，t-MTL 的 Rank-1 accuracy 达到 51.57%，mAP 达到 22.71%，在多次查询模式下，Rank-1 accuracy 和 mAP 则分别达到了 59.44% 和 30.75%。其中 t-MTL 方法仅使用训练样本，而不使用训练标签，与采用原始特征的欧几里德距离 [82] 相比，Rank-1 accuracy 分别提升了 8.61% 和 6.51%。与 CAMEL(71) 相比，在采用相同实验设置的情况下，Rank-1 accuracy 在多次查询模式下也提升了 4.9%，这归功于基于张量的多任务正则化具有良好的泛化能力，可以利用不同视角下的共享信息。此外，与聚类方法 PUL [65] 相比，t-MTL 利用了 Markte-1501 的未标记训练数据和 JSTL 特征提取器，基准方法略强于 PUL。图3.6显示了由无监督 t-MTL 模型产生的一些代表性识别结果，其中由 (47) 定义的废图像 (Junk imgae)，表示那些干扰项或与查询来自同一视角的图像。从中可以发现，因为仅使用聚类结果来训练模型，相似的外表行人更容易被识别出来。

表 3.7 无监督 t-MTL 方法在 Market-1501 数据集上的识别精度。

Table 3.7 Performances of unsupervised t-MTL for Market-1501

Market-1501	单次查询		多次查询	
方法	Rank-1	mAP	Rank-1	mAP
JSTL[82]	43.0	19.2	52.9	25.7
PUL[65]	45.5	20.5	-	-
CAMEL[71]	-	-	54.5	-
t-MTL	51.6	22.7	59.4	30.8

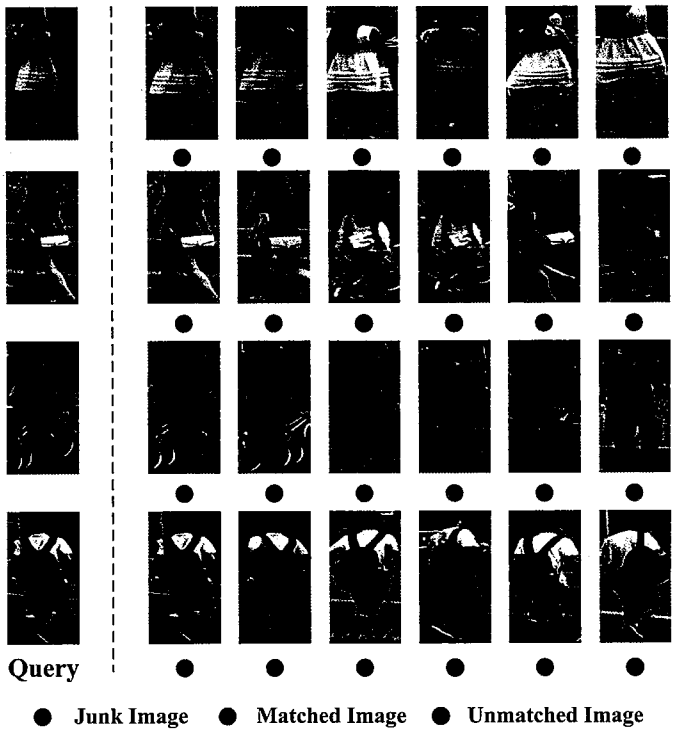


图 3.6 t-MTL 模型在 Market-1501 上得到的代表性识别结果。

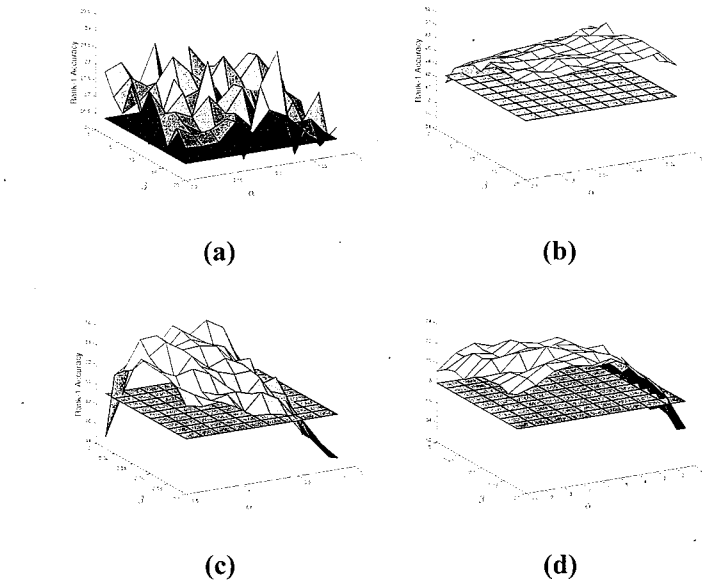
Figure 3.6 Representative RE-ID results on Market-1501 produced by unsupervised t-MTL.

3.4.2 模型分析

本小节将对 t-MTL 方法进行进一步的分析和实验，以便更好地了解 t-MTL 的相关特性。

3.4.2.1 敏感性分析

实验首先对参数进行敏感性分析。α 和 β 两个关键参数在 t-MTL 方法中起着重要作用。但是无论二者如何取值，大多数结果仍然优于基准方法。值得注意的是，所有结果都是通过随机切分数据集获得的，以避免过拟合现象发生。对于有监督 t-MTL 方法，实验通过使用 α 和 β 不同的取值来评估参数对算法性能的



(a) (b)

(c) (d)

图 3.7 参数 α 和 β 对性能的影响。图3.7a 和3.7b表示 ViPeR 和 Markte-1501 上无监督 t-MTL 的结果，图3.7c 和3.7d 表示 ViPeR 和 CUHK01 上有监督 t-MTL 的结果。

Figure 3.7 Influence of parameter α and β in terms of Rank-1 accuracy. Fig. 3.7a and 3.7b show unsupervised results for ViPeR and Markte-1501. Fig. 3.7c and 3.7d show supervised results for ViPeR and CUHK01.

影响。如图3.7c所示，水平面表示基准模型 [64] 的性能。在 ViPeR 数据集上，当 α 和 β 增加时，t-MTL 的 Rank-1 accuracy 先上升，然后保持相对稳定。但当它们超过某个特定阈值时，性能开始下降。通过将 α 设置在 $[0.5, 1]$ 之间以及将 β 设置在 $[0.01, 0.05]$ 之间，可以获得最佳性能。图3.7d展示了 t-MTL 在 CUHK01 数据集上的参数灵敏度分析，其结果与在 ViPeR 数据集上的结果类似，图中的平面仍然表示基准模型 [64] 的性能。随着训练数据规模增加，分类器堆叠的矩阵增大，较大的 α 在该数据集上更为适合。

对于无监督 t-MTL 方法，由于数据集的随机切分，t-MTL 模型在 ViPeR 数据集上随参数变化波动较大。当 $\alpha = 0.08$ 和 $\beta = 18$ 时，t-MTL 取得最佳性能。值得注意的是，如图3.7a所示，ViPeR 数据集上 t-MTL 模型没有明显的性能差异，受参数的影响较小，此外，大多数结果明显优于基准面 [71]。Market-1501 的主要结果如图3.7b所示。当 α 增加时，Rank-1 accuracy 先上升后下降。β 与 α 的结果类似。

3.4.2.2 无监督 t-MTL 方法分析

为了全面了解无监督 t-MTL 方法，本小节进一步讨论了聚类类别数目对于最终识别性能的影响。如表3.8所示，总体而言，当聚类类别数增加时，无监督 t-MTL 的识别准确率也会缓慢提升。

但值得注意的是，在 VIPeR 和 Market-1501 数据集上，当聚类类别数增加到真实类别数目时，识别准确率保持稳定，但继续增加聚类类别数，Rank-1 accuracy 依然缓慢增长。直觉上而言，当聚类类别数与真实别类别数吻合时，无监督算法能取得最佳性能。但是最近一些论文表明，在无监督的学习环境中，个体识别对于无监督学习是非常有效的，它们认为一个典型学习模型可以从数据本身中学习到判别性，而并不需要语义注释。因此，这些方法将每个样本视为一个独立的类别，并尝试训练一个模型来将区分每一个样本。从这个角度出发，在弱监督下，聚类类别数目的增加将有利于个体判别性的学习，同时模型也会不被远离聚类中心的难样本所破坏。因此，在 VIPeR 数据集上，取训练样本数作为聚类类别数目，t-MTL 达到最佳性能，这与 [93; 92] 实验结论类似。同时，所提的两个正则化项作为先验知识，也有助于模型获得更好的泛化能力。CAMEL[71] 也得到了类似的实验结果，因此，更多的聚类类别数将对无监督 t-MTL 方法有所帮助。

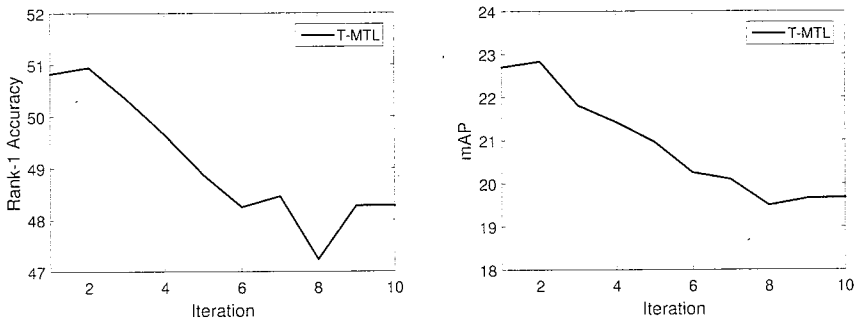
表 3.8 聚类类别数目对无监督 t-MTL 方法在 Market-1501 和 VIPeR 上的性能影响。

Table 3.8 Performances of unsupervised t-MTL when number of clusters K varies on Market-1501 and VIPeR.

聚类数目	500	800	1000	1500	2000
Market-1501	49.0	50.8	50.8	51.2	51.6
聚类数目	200	300	400	500	632
VIPeR	15.0	19.5	22.3	24.9	28.6

此外，在 CAMEL 中，为了提高性能，需要迭代逐步生成伪标签。但是在本章的实验中，发现这个过程对于无监督 t-MTL 模型来说并无帮助。如图所示3.8，Rank-1 accuracy 在最初的几次迭代中达到峰值，然后缓慢下降，最后保持稳定。在 ViPeR 数据集上的实验同样如此，随着迭代次数的增加，性能先保持稳定，然后不论聚类类别数目，性能都略有下降。为了深入分析这种异常现象，本节在 Market-1501 上进行了额外的实验。实验从某一类别中的样本，选取前 25 个训练

样本，观察它们在每次迭代中伪标签的变化情况。结果表明，在第一次迭代之后，更多的样本被划分在同一类别中，这表明在投影空间中，特征变得更具区分性。随着迭代次数的增加，这些样本的伪标签的变化很小，甚至出现了错误的监督，并且每次迭代后，具有相同伪标签的样本仍然以很高的概率分配到同一个聚类中心。这表明，聚类得到的弱监督信息不能进一步提升模型的性能。



(a) (b)

图 3.8 迭代次数对 t-MTL 性能的影响。

Figure 3.8 Influence of iteration number on Rank-1 accuracy and mAP for Markte-1501.

本小节还针对不同的聚类算法进行了实验，评估聚类算法对无监督 t-MTL 方法性能的影响。实验选取了两种聚类方法，k-means 和 t-SVD-MSC[25]。其中 t-SVD-MSC 在多个数据集上取得了不俗的聚类性能，因此选取它作为对比方法。由于 t-SVD-MSC 是一种多视图聚类方法，因此实验以 GOG 和 LOMO 作为输入，结果显示在表3.9中。在相同的参数设置下，结果表明对于不同的聚类类别数目 K，二者的性能差别不大，聚类算法对于无监督 t-MTL 方法的性能并没有起到关键作用。

表 3.9 聚类算法对无监督 t-MTL 方法在 VIPeR 上的性能影响。

Table 3.9 Performances of unsupervised t-MTL with different clustering methods on VIPeR.

聚类数目	200	300	400	500	632
k-means	15.0	19.5	22.3	24.9	28.6
聚类数目	200	300	400	500	632
T-SVD-MSC	16.9	19.7	20.3	23.2	28.6

3.4.2.3 算法复杂度和收敛性分析

尽管 t-MTL 的优化复杂度较高，但整个过程仅在离线训练时执行一次。算法的主要计算瓶颈在于求解子问题 $\mathcal{G}$，但求解该问题等价于计算 $\frac{V-1}{2}$ 个矩阵 SVD 分解，其中每个矩阵的维度为 $d \times C$。总之，计算快速傅立叶变换及其逆运算需要 $\mathcal{O}(2N^2V\log(N))$，同时需要 $\mathcal{O}(N^2V^2)$ 计算矩阵 SVD。

因此，假设有 V 个视角，$\mathcal{G}$ 的计算复杂度为 $\mathcal{O}(\min(VC^2d, Vd^2C))$，由于 $min(d, C) \gg \log(V)$，t-MTL 算法的复杂度为：

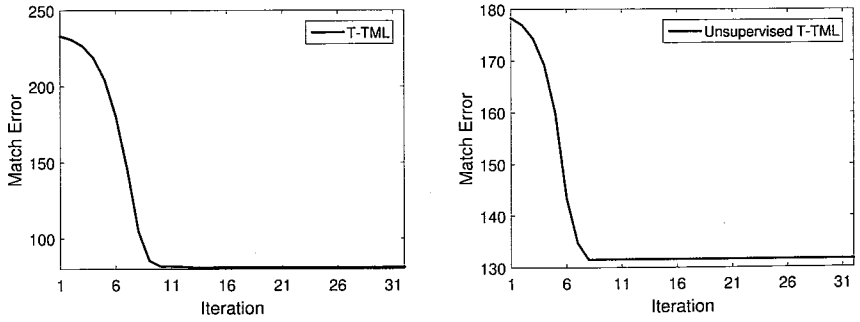
$$\mathcal{O}(\min(KVC^2d, KVd^2C))), \tag{3.31}$$

其中 K 表示优化过程中的迭代次数。实际上，K 通常位于 30 ~ 50 之间，经验上，所有实验设置 $K = 30$。此外，在实际应用中，类别的数量似乎在模型的可扩展性方面起着关键作用，然而，t-MTL 能够应对几万类别数据的情形，满足现阶段的行人重识别任务需求。表3.10展示了 t-MTL 方法的训练时间，对于 Market-1501 数据集，也仅需要 1200 秒就能够完成训练，远低于深度学习模型。

表 3.10 t-MTL 训练时间分析。

Table 3.10 Quantitative analysis of execution time for our supervised algorithm.

数据集	ViPeR	CUHK01	CUHK03	Market-1501
运行时间	7.2s	101.3s	52.4s	1267.3s



(a) **(b)**

图 3.9 ViPeR 数据集上 t-MTL 的收敛曲线。

Figure 3.9 Convergence Curve on ViPeR Dataset.

此外，算法在每次迭代产生的序列，使目标函数单调递减，子问题得到了精

确求解。因此，有：

$$||\mathcal{U}_{t+1} - \mathcal{U}_t||_F \to 0. \tag{3.32}$$

而 t-MTL 的收敛性可以由以下准则表示：

$$\text{Match Error} = ||\mathcal{U} - \mathcal{G}||_\infty. \tag{3.33}$$

实际上，t-MTL 方法在实际中收敛得很快，如图3.9所示，图中曲线记录了每个迭代步骤中的误差值。

3.5 本章小结

本章提出了一种新的张量多视角度量模型来克服视角差异，在多视角分类框架下，通过张量低秩和 Bregman 差异约束，探索视角之间与视角内部的关联关系，对齐由视角偏差导致数据分布差异，提升行人重识别准确率。此外，本章还对 t-MTL 算法进行了无监督扩展以及构建多特征张量提升算法性能。在四个公开数据集上的实验结果表明，t-MTL 在有监督和无监督两种模式下，能够有效对齐多视角特征分布，提升特征的判别性。然而，t-MTL 模型还存在一些弊端需进一步改进。主要包括参数敏感性问题和扩展性问题，由于数据集的规模和差异，t-MTL 模型需要针对每个数据集进行参数微调，而且，在无监督的情况下，t-MTL 无法通过交叉验证的方式获取参数。因此，下一步的研究工作需要集中于自适应的参数获取方法。此外，如何通过端到端的训练模式训练非对称度量，以应对模型扩展性的问题，也是亟待解决的问题。

第 4 章　面向行人重识别的多视角深度对齐度量学习

4.1　引言

随着人工智能技术和深度学习方法的发展，近些年行人重识别方法取得了里程碑式的成功，识别准确率不断攀升，借助大规模深度卷积神经网络 (Convolutional Neural Network)[68; 78; 47] 的强大判别能力，基于深度学习的行人重识别方法提供了强有力的行人特征表达，取得了不俗的成绩。

其中，损失函数 [131; 149] 作为深度网络的学习目标，是深度 Re-ID 方法的重要组成部分。初期的 Re-ID 方法要么使用分类损失（例如交叉熵损失）[47; 146]，将不同类别的样本分离，要么使用深度度量损失（例如三元组损失）[83]，在分离不同类别样本的同时，聚拢同类样本。其中大多数方法 [83; 149] 是为全局特征表示而设计的。然而，近年来研究者们发现基于全局表示的方法 [132; 142]，在跨视角场景下，经常会遇到所谓的“错位”现象 (Mis-alignment)（图4.1中行人图像）。“错位”现象通常是由姿势/视角变化、不准确的行人检测、遮挡引起的，当面对这种情形时，提取到的全局特征难以对行人部位、区域进行有效对齐，致使模型的重识别准确率不高。相反，近期的 Re-ID 方法开始尝试使用行人图像的局部表示，即局部特征来提取细粒度信息。常见的做法是利用姿态估计或基于纵向的平均分割，生成对应于头部、腿部和其他身体部分的局部特征，以此来加强行人的特征表示，提升 Re-ID 性能。

然而，大多数基于局部特征的 Re-ID 方法 [124; 139] 直接使用全局损失来独立地训练每个局部部分，从而忽略了不同视角下，行人局部特征之间的关系（例如错位情形），这会导致跨视角样本之间的距离估计不准确，这对于深度度量方法有着严重影响。为了解决这一问题，Aligned Re-ID[132] 提出运用基于动态规划的对齐距离来替代原有的欧式距离，通过计算两组跨视角局部特征之间的最短路径，给出样本之间的对齐距离。但值得注意的是，Aligned Re-ID 实质是通过贪心的算法，求解最短路径问题，该方法需要计算起始局部区域之间的距离，因此仍然会受到错位问题的影响。此外，Aligned Re-ID 对每一个局部特征赋予相同的权重，而不强调行人的显著性区域，提升有限。

为了解决上述问题，本章提出一种基于注意力机制的跨视角对齐方法。其

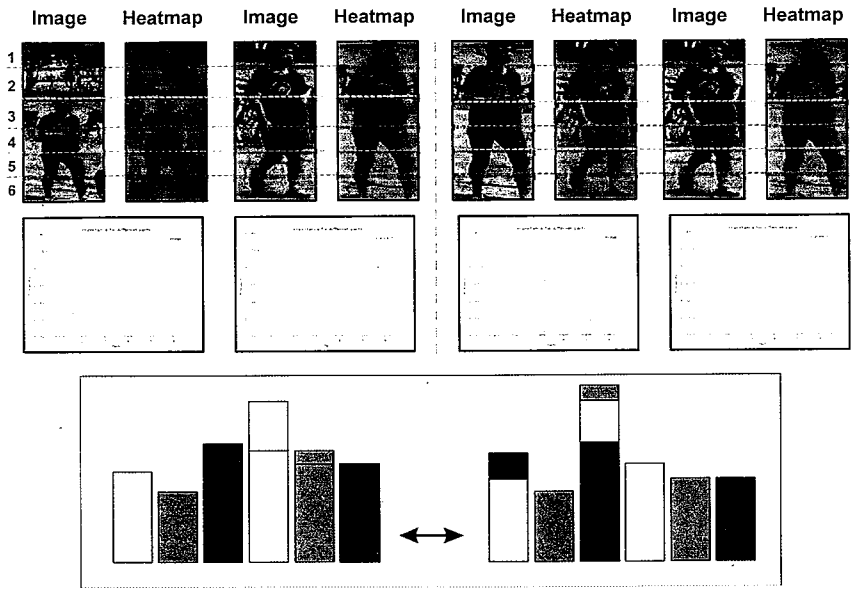


图 4.1 行人图像错位情形示意图。

Figure 4.1 An example of mis-alignment issue.

中，注意力机制 [135; 158]，可以在给定图像级别标注的情况下，生成图像的感兴趣区域。通过注意力的学习，模型能够快速区分行人的目标区域与无关部分，对齐和强调这些感兴趣区域为错位问题提供了潜在的解决方案。具体而言，如图4.1所示，图像的热图是由所提出的注意力方法生成的，从蓝色到红色的热图颜色代表了注意力由小到大的取值，反映了各个区域对识别的重要性。中间的直方图是基于纵向的平均分割得到平均注意力，可以发现直方图的对齐为错位问题提供了潜在的解决方案。基于这些观察结果，本章提出了一种新的三元组损失 (Wasserstein triplet loss, W-Triplet)，在试图解决错位问题的同时，更加关注行人显著区域。这种损失基于 Wasserstein 距离，又名推土机距离，它测量移动两组数据点来对齐两个分布所需的功。

在该框架下，W-Triplet 损失首先将网络的中间结果纵向分割成局部特征图 (Featuremap)，再经由平均池化生成局部特征，受 [151] 的启发，W-Triplet 学习得到不同区域的注意力，生成对应于局部特征的概率分布。与现有的注意力方法不同 [66; 152]，所提出的注意力模块将感受野为整幅图像的全局特征，与感受野为图像子集的局部特征进行相容性打分（即内积），得到模型的注意力。该相容性得分实质上反映了特征图中局部特征的重要性，也就是特征图上每一点的注意力。通过池化各个区域的注意力，再通过一个 sofmax 层，生成所谓的注意力概

率，从而将原来的跨视角对齐学习问题转化为跨视角分布对齐问题。

随后，W-Triplet 损失给出一个基于最优运输问题的优化目标，也被称为 Wasserstein 距离 [143]。通过求解该运输问题，模型可以得到一个最优输运计划，该运输计划作为权值，进一步加权得到跨视角局部特征的的距离。Wasserstein 距离实质上是寻找一个最优的运输方案，来搬运局部特征以对齐相应的注意力概率。该运输计划，即运输矩阵的元素表示概率值从一个点搬运到另一个点的数量。通过该最优运输问题，Wasserstein 距离直接导出了一种新的样本距离测量方式，并以三元组损失的形式用于模型训练。结果表明，Wasserstein 距离能够以端到端的方式在大多数行人重识别系统中实现，由此得到的 W-Triplet 损失能够有效缓解错位情况，显著提高模型的性能。

本章的主要贡献概括如下：

• 本章提出了一种新的多视角深度对齐损失。该损失通过求解最优运输问题，导出了一种新的跨视角样本距离估计方式，并以三元组损失的形式用于模型训练。该损失能够有效缓解错位情况，显著提高模型的性能。

• 本章运用一种新的空间注意力机制区分图像中显著区域与无关区域。该注意力机制作为注意力概率，能够引导 Wasserstein 距离学习，得到对齐的跨视角样本距离。

• 本章提出了一种新的多分支网络，能够有效融合全局和局部特征。

• 所提方法能够以端到端的方式进行训练，并易于在大多数行人重识别方法中实现。本章在几个基准数据集和基准模型上对所提方法进行了测试，验证了方法的有效性。

本章接下来内容的组织结构如下：第4.2节介绍相关的工作以及背景知识。第4.3节介绍深度损失的形式化表达，然后给出注意力模块、网络结构和模型解释。第4.4节介绍实验研究相关的内容，还进行了分析和讨论。最后，在第4.5节对本章工作进行总结并提出后续研究。

4.2 相关工作

4.2.1 行人重识别深度度量学习

深度度量学习 [131; 149]，作为深度网络的学习目标，是行人重识别模型的重要组成部分之一。早期的深度 Re-ID 方法通常采用分类网络 [97]，使用交叉

熵损失 (Cross-entropy loss)[146; 47; 124; 82] 进行训练，试图训练模型能够正确识别每一个行人个体。然而，交叉熵损失只能分离不同类别的数据点，而无法聚拢同一类别的样本。为了克服这一问题，人脸识别领域率先引入了三元组损失 (Triplet loss)[131]，并逐渐成为一种学习范式 [58]，三元组损失通过采样的方式，采集锚点 (Anchor)、正样本点和负样本点，通过假定正样本对之间距离小于负样本对之间距离，提升特征的判别性。在此基础上，Hermans 等人 [83] 提出了难样本挖掘 (Hard sample mining) 方法，对于每一个锚点，只优化批样本中最难的正负样本，缓解三元组损失带来的训练不稳定和模型收敛性问题。Quadxruplet loss[150] 将三元组关系推广到四元组关系，以此来扩大类间距离、缩小类内差异。Aligned Re-ID[132] 考虑局部特征的对齐问题，将原有的对齐距离学习问题转化为最短路问题，从而提供更好的跨视角样本距离估计。

4.2.2 行人重识别局部特征对齐学习

许多深度学习方法 [144; 132; 124; 139] 已经开始尝试运用局部特征来加强行人的特征表达。其通常做法是将特征图沿竖直方向上切分成几个局部区域，每一个区域对应于一个局部特征。PCB[124] 是其中的典型代表，PCB 将网络的中间结果划分为不同的部分，每个部分利用交叉熵损失进行训练，显著提升了识别匹配的准确率。Yang 等人 [138] 提出了一种水平金字塔匹配方法，该方法充分利用了不同池化操作和不同尺度提取到的局部信息，加强了局部特征的判别性。MGN[139] 利用了多分支表示的优点，融合全局和局部信息，取得了很好的性能。SPReID[130] 增加了一个额外的行人语义分析分支，用于生成局部概率，该概率进一步用作行人部位的损失权重，优化模型。

行人错位问题在行人重识别领域也得到了广泛的研究。PAN[159] 增加了一个基于注意力机制的仿射变换估计分支，来自适应地定位和对齐行人。DSA-reid[161] 通过密集的语义对齐，解决局部特征错误匹配的问题。AAN[127] 通过学习逐块偏移和逐像素偏移，自动地将行人图像从粗到细对齐。姿态估计 (Pose estimation)[142; 140] 通过姿态估计模块，减少姿态变化的影响，但这类方法需要额外的行人关键点、姿态标注，限制了其实际应用价值。

4.2.3 卷积神经网络注意力机制

注意力机制 [135; 158]，在不需要额外标注的情况下，以其强大的区分感兴趣区域和无关区域的能力，近年来得到学者们广泛关注，并成功地应用于行人重识别领域。其中，CAN[66] 是代表性的 Re-ID 注意力方法，CAN 通过对比行人图像之间的差异，提取行人图像的感兴趣区域，DuATM[158] 与之类似，通过双通道的匹配网络，提取序列行人图像注意力区域。与二者不同，Selfattention[165] 通过多卷积通路自对比方式，强化目标区域，用于图像生成。[129] 设计了一种前景注意力神经网络，通过编解码框架增强前景的正面影响，同时削弱背景的负面影响。Spatial Attention[126] 将特征图上的每一个点，沿通道维度运用平均池化和 softmax，产生空间注意力热图。Saumya 等人 [151] 提出了一种注意力学习机制，该方法利用网络的中间表征和高层特征的内积来产生空间注意力。经过注意力机制加强的网络中间层直接连接分类损失。Li 等人 [152] 提出了一种软像素注意和硬区域注意联合学习的特征表示模型。

4.2.4 推土机距离及预备知识

计算机视觉和机器学习中的许多问题，都可以落入推近两个概率分布的范畴 [148; 147]。对于该问题，一种直观的解决方法是定义一种距离来衡量两个分布之间的差异，并通过最小化分布之间的距离来迫使分布接近。推土机距离（Earth Mover Distance，也称为 Wasserstein 距离）就是这样一种距离，它测量从一个分布搬运数据点到另一个分布上所做的最少功。形式上，假设有两个离散分布为 $\mathbf{r}$ 和 $\mathbf{c} \in \mathbb{R}^n$，搬运数据点的运输代价定义为 $\mathbf{M} \in \mathbb{R}^{n \times n}$，其中 n 是离散变量的数目。则运输计划，即运输矩阵 $\mathbf{T} \in \mathbb{R}^{n \times n}$ 表示从分布 $\mathbf{r}$ 上搬运到分布 $\mathbf{c}$ 上数据点的数量。通过求解该最优传输矩阵，可以得到推土机距离，即最优运输问题的目标函数：

$$\begin{aligned} \mathcal{D}_{\mathbf{M}}(\mathbf{r}, \mathbf{c}) &= \min_{\mathbf{T}} \langle \mathbf{T}, \mathbf{M} \rangle \\ \text{s.t.} \quad & \mathbf{T}\mathbf{1} = \mathbf{r} \\ & \mathbf{T}^T\mathbf{1} = \mathbf{c}, \end{aligned} \tag{4.1}$$

其中 $\langle \cdot, \cdot \rangle$ 表示 Frobenius 内积。

推土机距离基于最优运输理论，提供了一种强有力的直方图匹配度量 [143]。其中有大量的工作关于该最优运输问题的求解，但事实上，只需要微小的改动，

算法 4 近似最优运输问题求解算法

1: **输入**：运输成本 $\mathbf{M} \in \mathbb{R}^{n \times n}$，离散分布 $\mathbf{r}$ 和 $\mathbf{c} \in \mathbb{R}^n$，常数 λ。
2: **输出**：最优运输矩阵 $\mathbf{T}^*$
3: 随机初始化 $\mathbf{u}$ 和 $\mathbf{v} \in \mathbb{R}^n$;
4: $\mathbf{K} = \exp(\frac{1}{\lambda}\mathbf{M})$
5: **for** $i \leftarrow 1$: max iteration **do**
6: $\quad \mathbf{u}^{i+1} = \mathbf{r}./(\mathbf{K}\mathbf{v}^i)$;
7: $\quad \mathbf{v}^{i+1} = \mathbf{c}./(\mathbf{K}^T\mathbf{u}^i)$;
8: **end for**
9: $\mathbf{T}^* = diag(\mathbf{u})\mathbf{K}diag(\mathbf{v})$;
10: 返回最优运输矩阵 $\mathbf{T}^*$。

就可以使原问题以迭代和可微的方式求解。特别的，通过加入 $\mathbf{T}$ 的信息熵正则项，可以得到最优运输问题的近似问题 [153]:

$$\mathcal{D}_{\mathbf{M}}(\mathbf{r}, \mathbf{c}) = \min_{\mathbf{T}} \langle \mathbf{T}, \mathbf{M} \rangle + \lambda \sum_{ij} \mathbf{T}_{ij} \log \mathbf{T}_{ij} \tag{4.2}$$

$$\text{s.t.} \quad \mathbf{T}\mathbf{1} = \mathbf{r}$$

$$\mathbf{T}^T\mathbf{1} = \mathbf{c},$$

其中 λ 是一个控制近似程度的超参数，并固定为 0.1。该正则化问题是严格凸的，并易于在现有的深度学习框架下实现。具体而言，通过引入两个辅助向量 $\mathbf{u}$ 和 $\mathbf{v} \in \mathbb{R}^n$，并对二者进行交替求解，可以得到运输问题 (4.2) 的解。形式上，在第 i 次迭代，辅助向量按如下方式更新:

$$\mathbf{u}^{i+1} = \mathbf{r}./(\mathbf{K}\mathbf{v}^i) \tag{4.3}$$

$$\mathbf{v}^{i+1} = \mathbf{c}./(\mathbf{K}^T\mathbf{u}^i),$$

其中，$\mathbf{K}$ 表示由 $\mathbf{M}$ 计算的核矩阵，而 ./ 表示点除。当迭代满足终止条件或超过迭代最大次数时，最优运输问题的解可以写作 $\mathbf{T}^* = diag(\mathbf{u})\mathbf{K}diag(\mathbf{v})$ 的形式。整个求解过程总结在算法4中。

4.3 基于推土机距离的多视角三元组损失

本节将详细阐述深度多视角对齐度量学习的形式化描述，以进一步捕捉局部特征之间的关系，此外，在现有的行人重识别模型中，将其实现为三元组损失形式，以一个可训练的深度模型呈现。

4.3.1 推土机三元组损失

深度神经网络通常以批样本 (Batch samples) 作为输入，生成相应的特征表示或嵌入 (Embeddings)$f(\mathbf{x}_i,\theta):\mathcal{X}\rightarrow\mathbb{R}^d$，其中 θ 表示网络的参数，$\mathbf{x}_i\in\mathcal{X}$ 表示第 i 个样本。为了使这种特征表达更具判别性，模型通常采用深度度量损失进行训练，例如配对损失 (Pairwise loss)、三元组损失等。数学上，三元组损失以如下形式给出：

$$\mathcal{L}(\mathcal{X},\mathbf{y};\theta)=\sum_{i,p,n}\max(0,-(\mathcal{D}(f(\mathbf{x}_i;\theta),f(\mathbf{x}_p;\theta)) - \mathcal{D}(f(\mathbf{x}_i;\theta),f(\mathbf{x}_n;\theta))+\epsilon), \tag{4.4}$$

其中，$\mathbf{x}_i$ 表示从数据中随机选取的锚点，$\mathbf{x}_p$ 表示正样本点，通常与锚点来自同一类别，$\mathbf{x}_n$ 表示负样本点，通常与锚点来自不同类别，ϵ 表示间隔 (Margin)。此外，在公式 (4.4) 中，$\mathcal{D}(\cdot,\cdot)$ 表示距离函数，用于衡量跨视角对象之间的远近程度，一般情况下，$\mathcal{D}(\cdot,\cdot)$ 由欧几里德距离给出：

$$\mathcal{D}(f(\mathbf{x}_i;\theta),f(\mathbf{x}_j;\theta))=||f(\mathbf{x}_i;\theta)-f(\mathbf{x}_j;\theta)||_2. \tag{4.5}$$

直观上而言，三元组损失能够聚拢同类样本，同时推远不同类样本，从而提升特征的表示能力。

然而，在跨视角行人图像匹配背景下，直接使用欧几里德距离可能会导致错位问题。为此，本节提出一种新的距离函数，来对齐跨视角样本偏差。具体而言，令 $\mathbf{f}_i=\{\mathbf{f}_{i1},\mathbf{f}_{i2},\ldots,\mathbf{f}_{im}\}$ 表示行人图像的局部特征，类似于 [124]，该局部特征是将网络特征图沿竖直方向等分，再经由平局池化生成的，其中 m 表示局部特征生成的数目[1]。除此之外，考虑在这些局部特征上的离散概率分布 $\mathbf{p}_i=(p_i(1),p_i(2)\ldots,p_i(m))$，其中每一个离散概率值对应于一个单独的局部特征。这种离散分布本质上是下节介绍的注意力机制，其中较高的取值表示图像显著性的区域，较低的值表示无关或背景区域。

[1]具体的生成细节请参见4.3.3。

基于上述定义，一个最优运输问题可以被形式化的给出，其目的是寻找最优运输矩阵，对齐局部特征上的分布，即：

$$\mathcal{D}_{\mathbf{M}}(\mathbf{p}_i, \mathbf{p}_j) = \min_{\mathbf{T}} \langle \mathbf{T}, \mathbf{M} \rangle + \lambda \sum_{k,l} \mathbf{T}_{kl} \log \mathbf{T}_{kl} \tag{4.6}$$

$$\text{s.t.} \quad \mathbf{T}\mathbf{1} = \mathbf{p}_i \quad \text{and} \quad \mathbf{T}^T\mathbf{1} = \mathbf{p}_j,$$

其中，$\mathbf{M}_{kl} = ||\mathbf{f}_{ik} - \mathbf{f}_{jl}||_2^2$ 表示运输代价，即局部特征之间的距离。该最优运输问题直接来自于公式 (4.2)，并可以通过算法4进行求解。学习到的最优运输矩阵 $\mathbf{T}^*$ 能够将两个局部特征上的概率以最少的功对齐，其值越大，表示移动的数据点的数量就越多。该最优运输矩阵被进一步用作校正原始距离的权值，得到两组跨视角局部特征之间的距离。这种加权的距离也被称为 Wasserstein 距离：

$$\begin{aligned} \mathcal{D}(f(\mathbf{x}_i;\theta), f(\mathbf{x}_j;\theta)) &= \mathcal{D}(\mathbf{f}_i, \mathbf{f}_j, \mathbf{p}_i, \mathbf{p}_j; \theta) \\ &= \frac{\sum_{kl} \mathbf{T}^*_{kl} ||\mathbf{f}_{ik} - \mathbf{f}_{jl}||_2^2}{\sum_{kl} \mathbf{T}^*_{kl}}. \end{aligned} \tag{4.7}$$

在此基础上，通过将 Wasserstein 距离代替原始的欧几里德距离，本节提出一种新的 Wasseratein 三元组损失 (W-Triplet)。通过该损失，原始的跨视角对齐学习问题就转化为跨视角分布对齐问题（公式（4.6））。

与此同时，出于批样本中三元组规模和训练稳定性的考量，W-Triplet 遵循 [132; 83] 提出的方法，设计了基于全局特征的难样本挖掘分支。具体而言，W-Triplet 将批样本中的每一个样本作为锚点，同时针对每一个锚点，只选取最难的正、负样本进行训练，即与锚点距离（全局特征欧几里得距离）最远的正样本以及与锚点距离最近的负样本。此外，虽然 W-Triplet 获得了可微的 $\mathbf{T}^*$，但是实验结果表明回传运输矩阵的梯度不仅会降低性能，还很耗时。所以算法只将 $||\mathbf{f}_{ik} - \mathbf{f}_{jl}||_2$ 的梯度回传。综上所述，基于 Wasseratein 三元组损失的训练损失如下：

$$\mathcal{L}_{\text{total}}(\mathcal{X}, \mathbf{y}; \theta) = \mathcal{L}_{\text{W}}(\mathcal{X}, \mathbf{y}; \theta) + \mu \mathcal{L}_{\text{C}}(\mathcal{X}, \mathbf{y}; \theta), \tag{4.8}$$

其中，μ 是一个超参数，$\mathcal{L}_{\text{W}}(\mathcal{X}, \mathbf{y}; \theta)$ 表示 Wasseratein 三元组损失，$\mathcal{L}_{\text{C}}(\mathcal{X}, \mathbf{y}; \theta)$ 表示交叉熵损失。事实上，如果 W-Triplet 能够设计出在局部特征上合理的概率分布，那么 Wasseratein 三元组损失将通过给显著性区域赋予更大的权重 $\mathbf{T}$ 来增强特征表达，并通过只计算匹配的局部特征之间的距离来克服错位问题。接下来的小节将介绍如何学习概率分布来指导 Wasseratein 三元组损失。

4.3.2 注意力引导的对齐概率学习

局部特征上的概率分布是为了引导模型强调并对齐显著区域，从而解决跨视角错位问题。因此 W-Triplet 采用注意力机制，在仅用图像级别的标注的情况下，试图区分行人感兴趣区域和不相关或误导的背景区域，并以此来产生对应于局部特征的概率分布。具体而言，令 $\mathbf{f}_s = \{\mathbf{f}_{s1}, \mathbf{f}_{s2}, \ldots, \mathbf{f}_{sk} \in \mathbb{R}^{d_s}\}$ 表示从卷积层 $s \in \{1, 2, \ldots, S\}$ 提取到的局部特征集合，其中 d_s 表示通道数。每一个 $\mathbf{f}_{si}$ 表示位于总共 k 个空间位置的特征图上第 i 个位置的激活响应。此外，令 $\mathbf{g} \in \mathbb{R}^{d_g}$ 表示维度为 d_g 的全局特征。在行人重识别背景下，模型经常会采用一个嵌入模块来增强特征的判别性，该嵌入模块输出的特征一般作为测试特征，具有很强的行人图像表示能力。因此，本章选用该特征作为全局特征。

在给定全局特征和局部特征的情况下，W-Triplet 以内积形式的兼容性函数 $C(\mathbf{f}^i, \mathbf{g})$ 来生成特征图上每个点的注意力：

$$C(\mathbf{f}_{si}, \mathbf{g}) = \langle \mathbf{u}, \mathbf{f}_{si} + \mathbf{g}' \rangle, \tag{4.9}$$

其中，$C(\mathbf{f}_{si}, \mathbf{g})$ 以两个维度相等的特征作为输入，输出标量来指示空间位置 i 的重要性。$\mathbf{g}'$ 表示经由线性变换后的全局特征，由于全局特征和局部特征的维度可能存在不一致情形，W-Triplet 将全局特征 $\mathbf{g}$ 送入一个线性层 (Linear layer) 产生 $\mathbf{g}'$。最后，相容性得分再经由一个 softmax 层输出注意力：

$$a_i^s = \frac{\exp(c_i^s)}{\sum_j \exp(c_j^s)}, \tag{4.10}$$

其中，$c_i^s = C(\mathbf{f}_{si}, \mathbf{g})$。通过这种方式，模型最终得到了与特征图空间尺寸一样的注意力 $\mathbf{a}^s$，为了得到最终的离散概率，W-Triplet 将 $\mathbf{a}^s$ 按局部特征的生成方式，池化为一个一维张量，并将其送入另一个 softmax 层得到注意力引导的对齐概率 $\mathbf{r}$ 和 $\mathbf{c}$ [2]。

实际上，注意力引导的对齐概率是由网络全局特征与中间结果之间内积产生的直方图。它代表了高层语义特征和局部特征之间的关联程度，从而显示了不同区域对于识别的重要性。此外，模型假定每个行人图像的注意区域应该是行人的各个身体部位，因此，注意力引导的对齐概率能够有效指导最优运输矩阵的学习，并帮助最优运输矩阵找到匹配和显著的局部区域，从而有效解决错位问题。

[2]虽然本章称得到的注意力得分为“离散概率”，但它不符合概率的严格定义，即有限结果概率（或发生频率）。

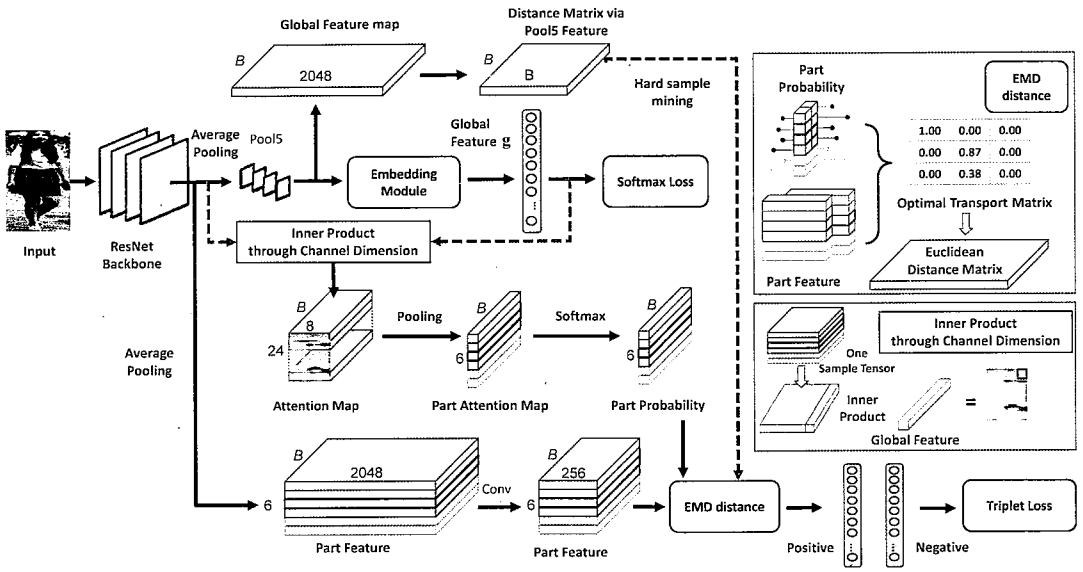


图 4.2 网络结构示意图。

Figure 4.2 The network architecture of the proposed method.

4.3.3 网络结构

W-Triplet 的网络架构如图4.2所示。该结构遵循识别模型 [124]，并进行了微调，这些变化主要集中在网络的末端。与典型的 Re-ID 模型一样，W-Triplet 利用一个骨架网络提取特征，骨架网络可以是为分类设计，去除分类器的任意网络，本章利用 ResNet-50 作为骨架网络。在骨架网络之后，除原识别分支外，模型对提出的 Wasserstein 三元组损失增加了三个并行分支。其中，原有的识别分支由嵌入模块和分类器组成，其中分类器是针对交叉熵损失设计的，由于嵌入模块是一个极富技巧性的设计，因此 W-Triplet 直接遵循 [124; 157] 的设计。

第一个增加的分支用于全局特征的难样本挖掘，有助于 Wasserstein 三元组损失的收敛以及提升模型训练的稳定性。该分支提取全局特征并计算批样本中任意两幅图像的欧几里得距离。通过该距离，模型针对每一个锚点，选取最难的正负样本，然后沿图中标记为红色虚线，将这些样本的索引送至 Wasserstein 三元组损失中进行难样本挖掘。

第二个额外的分支是所提出的注意力模块，在图中标记为蓝色虚线。该注意力模块有两个输入，一个是从最后一个线性层提取的高层全局特征，另一个是从 Pool5 层提取到的网络特征图。由于该网络特征图和全局特征的维度并不一致，注意力模块还将高层特征送至线性层进行维度变换。然后，注意力模块沿着通道维度对两个输入进行内积操作，得到注意力热图。最后，如图的右侧所

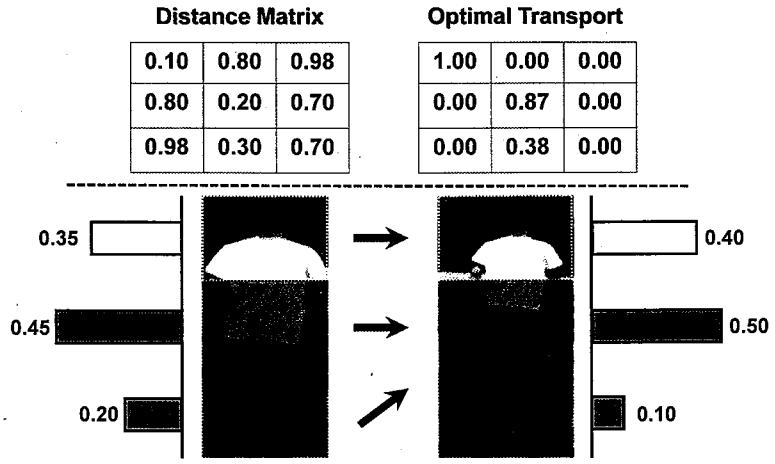


图 4.3 最优运输问题解决错位问题的示意图。

Figure 4.3 The diagram of how optimal transportation problem solve the mis-alignment issue.

示，注意力模块将注意力热图送至一个平均池化层和一个 softmax 层来生成对应于每个局部特征的对齐概率。

第三个附加分支用于生成基于行人部位的局部特征。它由一个平均池化层组成，其作用是将网络 Pool5 层特征图分成独立的几个部分（图中为 6 个）。然后将得到的局部特征送至一个卷积层，以降低局部特征的维数。最终，将网络得到的局部特征、对齐概率以及需要训练的难样本索引，送至 Wasserstein 三元组损失模块，进行最优运输问题的求解，该过程也在图的右侧进行了说明。通过该网络结构，模型以端到端的训练方式求解最优运输问题，并最终得到 Wasserstein 三元组损失。

4.3.4 Wasserstein 距离如何解决错位问题

本小节给出了一个 Wasserstein 距离解决跨视角错位问题的例子。如图4.3所示，直接使用欧几里德距离会导致严重的错位问题，从而导致错误的跨视角距离估计。为此，模型将每个行人图像（即网络特征图）沿竖直方向平均切分成三等份，则运输的代价为距离矩阵 $||\mathbf{f}_{ki} - \mathbf{f}_{lj}||_2$（即局部特征距离，图中的 Distance Matrix）。通过最优运输问题的求解，可以获得最优传输矩阵（图中的 Optimal Transport）。该运输计划告诉模型，对齐两个直方图（即注意力概率 $\mathbf{r}$ 和 $\mathbf{c}$，图中红色、蓝色和黄色长条）所做的最少的功应该是将左侧图像的第二部分和第三部运输到第右侧图像的第二部分，其中搬运的数量应该为最优近似运输矩阵的取值。因此，通过最优运输矩阵校正后的样本距离，并不计算任何未对准区域之间

的距离，同时，还通过赋予较大的权值来强化头部和身体等显著性的区域，加强特征的表达。虽然图4.3是一个极端的情形，但实验表明，推土机距离可以显著提升特征的判别性。

4.4 实验

在本章的实验中，将对所提方法的有效性进行全面评估。所有实验都是在一台上工作站上基于PyTorch实现的，该工作站配备了Intel Xeon E5-2630@2.30GHz CPU、128GB RAM 和 2 块 TITANX GPU（12GB 缓存）。为了清楚地说明实验方法，本节首先介绍实验设置，然后，章节4.4.3给出了实验的主要结果，包括和基准方法以及最新的相关方法的结果比较，同时，章节4.4.4进行了消融实验，验证所提方法各成分的有效性。最后，本节还分析了 W-Triplet 方法的参数敏感性、收敛性，以及给出模型的可视化结果。

4.4.1 实验数据集与设置

实验在四个公开数据集上进行了验证，即 CUHK03 [78]、DukeMTMC-Reid [133]、MSMT17 [136] 和 Market-1501 [47]。CHUK03 包含从六个监控摄像机中捕获的 1,360 名行人的 13,164 张图像。除了手工裁剪的图像外，还提供了由行人检测器检测到的行人图像。由于原始的评估方式对于深度学习类的方法非常耗时，因此实验遵循 [134] 提出的数据集划分方法进行评测。DukeMTMC-Reid 是最具挑战性的行人重识别数据集之一。它包括来自 8 个摄像头的 36411 张包含 1404 个行人个体的图像，其中 16522 张用于训练，2228 张用于查询和 17661 张作为待匹配库。MSMT17 标注了 4,441 个行人的 126,441 张图像。训练集包含 1,621 个行人个体的 32,621 个行人图像，而测试集包含 3,060 个个体的 93,820 幅图像。从测试集中，随机选择 11659 幅图像作为查询图像，而剩下的 82161 幅图像作为图库图像。实验使用两种常用的评价指标来评估性能，即累积匹配曲线（CMC）和平均精度（mAP），其中实验只报告单次查询模式下结果。值得注意的是，实验不采任何重排序（例如 [51]）技巧，为了公平地比较，对比方法也不使用这类方法。

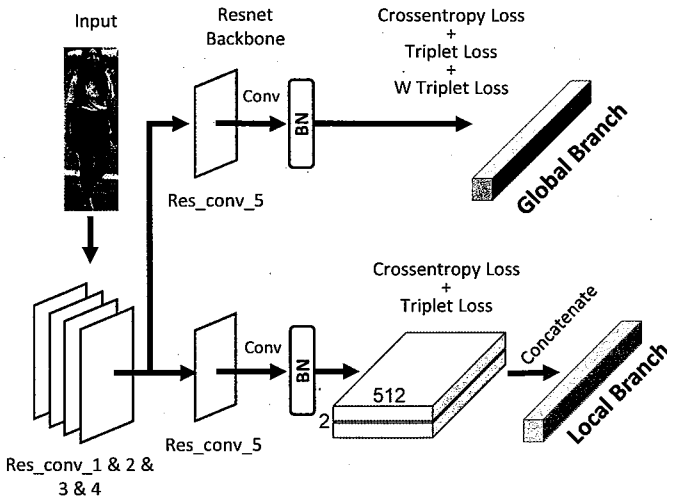


图 4.4 多分支网络结构图。

Figure 4.4 The multi-branch network architecture of the proposed method.

4.4.2 实施细节

实验使用交叉熵损失和 Wasserstein 三元组损失，以及原始的三元组损失来训练模型，章节4.4.4将详细的分析这一现象。此外，实验还在三种基准模型上进行了测试，包括 Strong baseline [157]，多分支模型和 PCB[124]。

Strong baseline & 多分支模型：为了与相关方法进行比较，实验遵循 [157] 的网络结构及其各种训练技巧（Warm up[162]、Label smooth[95]）训练网络，实验利用 ImageNet 上经过预训练的 ResNet-50(97) 作为骨架网络，在最后的池化层后加入 Batch Normalization 层 [137]，以提升特征判别能力。为了显示最佳性能，实验还设计了多分支网络（类似 MGN (139)）来提升识别准确率。如图4.4所示，多分支网络在 Strong baseline（全局分支）的基础上添加了一个局部分支，局部分支相较于全局分支，增加了局部池化层，以得到局部特征。在每个局部特征上，多分支网络运用三元组损失和交叉熵损失进行训练，而 Wasserstein 三元组损失仅在全局分支上使用。所有图像大小调整为 312×156，实验将公式（4.8）中的 μ 设置为 3，公式（4.4）中的 ϵ 设置为 0.3。实验使用 Adam 优化方法，将初始学习速率设为 3.5×10^{-4}，并在第 40、70 和 120 个训练周期 (epoch)，将原有的学习率减少为原来的 0.1，共训练 150 个训练周期。

PCB：为了验证方法的有效性，实验还在 IDE，PCB 和 PCB + RPP[124] 进行了测试。其中，IDE 和 PCB 模型的主要区别在于图4.2中所示的嵌入模块。IDE 中的嵌入模块为全局分类器，而 PCB 中的嵌入模块为基于局部特征的分类模块。

RPP 利用改进的局部池化来替代原始的平均池化，略微提高性能。请参考 [124] 了解更多详细信息。在基于 PCB 的模型中，图像的大小调整为 384×128。实验将公式（4.8）中的 μ 设置为 0.5，公式（4.4）中的 ϵ 设置为 0.3。实验使用 SGD 优化方法，将初始学习速率设为 0.1，并每隔 40 个训练周期 (epoch)，将原有的学习率减少为原来的 0.1，共训练 100 个训练周期。

对于所有基准模型，实验都使用随机水平翻转和裁剪进行数据增强。实验选取 64 作为批样本的大小，其中，一个批样本中随机选取 16 个不同的行人个体进行训练，每个行人个体包含 4 幅不同的图像。在测试阶段，对于 Strong baseline 和 PCB 方法，实验提取 pool5 之后的特征进行测试，而对于多分支模型，拼接局部特征和全局特征进行测试。

4.4.3 实验结果

本节展示了 Wasserstein 三元组损失的实验结果，并与相关方法进行比较。由于大量方法在这些基准数据集上进行了测试，因此，实验只比较了近三年或与所提方法紧密相关的算法，包括深度度量学习算法 (Aligned Re-ID [132], Triplet loss [83], Deep mutual learning[141] 和 RandomWalk [128])，局部特征模型 (PCB [124]) 以及基于注意力机制的方法 (DuATM [158], MLFN [155]), MGN [139], DSA-reID[161])。

Market-1501 上的实验结果。如表4.1所示，* 表示复现结果，W-Triplet 与最新的对比方法相比，取得了高度可比（甚至更好）的识别准确率。特别是，在多分支模型上，所提方法的 Rank-1 accury 为 95.5%，mAP 为 88.2%。值得注意的是，模型是与交叉熵损失和三元组损失共同训练的，这表明 W-Triplet 损失与这些损失相辅相成，并能够进一步改善模型的性能。在 Strong baseline 上，W-Triplet 也取得了一定的性能提升，这表明 Wasserstein 三元组损失对于不同的基准模型，都能够提升其性能，验证了方法的有效性。

对于深度度量学习而言，RandomWalk 利用数据库图像之间的关联关系进行模型优化；而 Aligned Re-ID 诉诸于对齐三元组损失和深度相互学习进行模型训练，这些方法取得了不俗的识别准确率，然而，W-Triplet 在使用相同的 ResNet-50 骨架网络情况下，显著优于这些度量学习方法。PCB 是基于局部特征的代表性模型之一，PCB 直接使用分类损失来训练每个局部特征，加强了行人的特征表达。但是，所提出的多分支模型，在融合局部和全局信息后，取得了更好的重识别精

表 4.1 W-Triplet 在 Market-1501 数据集上的实验结果。

Table 4.1 Performance of W-Triplet compared with state-of-the-art models on Market-1501.

方法	Rank-1	Rank-5	Rank-10	mAP
SVDNet[68]	82.3	92.3	95.2	62.1
TripletLoss[83]	84.9	94.2	-	69.1
AlignedReID[132]	91.8	96.3	-	79.3
MLFN[155]	90.0	-	-	74.3
HA-CNN[152]	91.2	-	-	75.7
DuATM[158]	91.4	-	-	76.6
Deep-Person[156]	92.3	-	-	79.5
RandomWalk[128]	92.7	96.9	98.1	82.5
Part + Aligned[164]	91.7	-	-	71.6
PCB[124]	92.3	97.2	98.2	77.4
PCB+RPP[124]	93.8	97.5	98.5	81.6
MGN[139]	95.9	-	-	87.4
DSA-reID[161]	95.7	-	-	87.6
Strong baseline [157]	94.1	-	-	85.6
Joint [125]	94.8	-	-	86.0
Strong baseline*	94.0	98.3	99.0	85.4
+ Wasserstein Triplet loss	94.5	98.3	99.0	85.8
Multi-branch	95.0	98.5	99.0	87.7
+ Wasserstein Triplet loss	**95.5**	**98.6**	**99.2**	**88.2**

度，关于 PCB 模型上的 W-Triplet 实验，将在4.4.4部分中详细介绍。值得注意的是，基于切分的局部模型（例如 DSA-reID [161] 和 MGN [139]），同样利用了多分支结构，提供更具判别性的行人特征，进而取得了优异的重识别性能。在此基础上，W-Triplet 能进一步提升这类模型的性能，这表明多分支方法与 Wasserstein Triplet 损失可以相互受益。

DukeMTMC-Reid 和 CUHK03 上的结果：DukeMTMC-Reid 的结果显示在表4.2中。与 Market-1501 相比，实验结果非常相似。由于多分支模型继承了 Strong baseline[157] 和 MGN [139] 的优点，因此该基准模型性能显著优于这两种对比方法。同时，在利用 Wasserstein 三元组损失后，多分支模型的 mAP 的提升了 0.6%，Rank-1 accuracy 提升了 0.3%。识别精度的提高可以归因于基于注意力的 Wasserstein 三元组损失，该损失在局部特征对齐过程中，对每个局部特征赋予不同的权重。最终，所提出的方法取得了 89.1% 的 Rank-1 accruacy，以及 79.8% 的

表 4.2 W-Triplet 在 DukeMTMC-Reid 和 CUHK03 数据集上的实验结果。

Table 4.2 Performance of W-Triplet on DukeMTMC-Reid and CUHK03.

W-Triplet	DukeMTMC-Reid		CUHK03	
方法	Rank-1	mAP	Rank-1	mAP
SVDNet[68]	76.7	56.8	41.5	37.3
MLFN[155]	81.0	62.8	54.7	49.2
HA-CNN [152]	80.5	63.8	41.7	38.6
DuATM[158]	81.8	64.6	-	-
Deep-Person[156]	80.9	64.8	-	-
RandomWalk [128]	80.7	66.4	-	-
Part + Aligned[164]	84.4	69.4	-	-
PCB[124]	81.9	65.3	61.3	54.2
PCB+RPP [124]	83.3	69.2	63.7	57.5
MGN[139]	88.7	78.4	68.0	67.4
DSA-reID[161]	86.2	74.3	78.9	75.2
Strong baseline*	86.8	76.1	57.1	55.6
+ Wasserstein Triplet loss	87.5	76.4	58.7	56.5
Multi-branch	88.8	79.2	70,9	69.2
+ Wasserstein Triplet loss	**89.1**	**79.8**	**72.5**	**69.8**

mAP，提供了非常有竞争力的重识别性能。W-Triplet 相较于对比方法（例如深度度量方法 (128)，对齐的方法 [164]，和基于注意力的方法 [155])，也取得了巨大提升。

表4.2还展示了 CUHK03 数据集上的实验结果。由于实验遵循 [134] 提出的新评估方法，只有少数方法在这种评估方式下进行了测试。因此，实验选取了基于识别的方法 [68]，基于注意力的方法 [158] 和一些其他代表性方法 (124; 152; 139; 161) 进行对比。如表中所示，Strong baseline 取得了为 57.1% 的 Rank-1 accuracy 和 55.6% 的 mAP，在此基础上，Wasserstein 三元组损失仍然可以提升该模型的性能。特别地，W-Triplet 取得了 1.6% 的 Rank-1 accuracy 提升。此外，DSA-reID [161] 的性能显著优于其他方法。而所提出的多分支模型也实现了不错的识别精度。

MSMT17 上的结果：MSMT17 是行人重识别领域中最大的数据集之一。该数据集利用部署在校园中的 15 个摄像机网络来捕获行人图像。表4.3总结了 MSMT17 数据集的实验结果。由于该数据集的场景和背景复杂，现有方法（例

表 4.3 W-Triplet 在 MSMT17 数据集上的实验结果。

Table 4.3 Performance of W-Triplet on MSMT17.

方法	Rank-1	Rank-5	Rank-10	mAP
GoogleNet[160]	47.6	65.0	71.8	29.7
PDC[142]	58.0	73.6	79.4	23.0
GLAD [163]	61.4	76.8	79.4	23.0
PCB [124]	68.2	81.2	85.5	40.4
Joint [125]	77.2	87.4	90.5	52.3
Strong baseline*	75.0	85.9	89.5	51.5
+ Wasserstein Triplet loss	76.1	86.7	89.9	51.6
Multi-branch	80.7	89.6	92.1	59.4
+ Wasserstein Triplet loss	**81.4**	**89.8**	**92.4**	**59.7**

如 [160]，[142]，[163] 和 [124]）的识别精度有限。而所提出的基准模型由于运用多种训练技巧，以及融合了局部和全局信息，相较于这些方法取得了巨大提升。多分支模型取得了 80.7% 的 Rank-1 accuracy 和 59.4% 的 mAP，而 W-Triplet 可以进一步提升基准模型的性能，对于 Strong baseline 而言，W-Triplet 的 Rank-1 accuracy 较相应的基准模型提升了 1.1%。在该复杂数据上的提升表明，提出的 Wasserstein 三元组损失通过计算匹配的区域，缓解了错位问题的负面影响，能有效提升深度模型性能。

4.4.4 消融实验

本节进行了多组消融实验以验证所提方法的有效性，包括不同训练技巧对性能的影响、W-Triplet 与基准模型的性能比较、不同三元组损失对性能的影响、不同注意力机制对性能的影响。其中实验对对比方法进行了优化，以得到公平的对比结果。

不同训练技巧对性能的影响：表4.4展示了不同训练技巧对最终识别性能的影响，实验测试了 Warm up，Label smooth 和 Random erasing 等常用的 Re-ID 训练技巧，与此同时，实验还测试了将 ResNet-50 最后一个卷积模块中的滑动距离设置为 1 的结果 (Last stride=1)，并使用了 Batch Normalization 层作为嵌入模块 (BNneck)。除 Random erasing 外，几乎所有的训练技巧均可显著提升性能。Random erasing 旨在使用随机数值将原始图片中的像素进行替换，一定程度上会提升模型的泛化性。但是对于 W-Triplet 方法，会显著降低其识别精度，这是

表 4.4 不同训练技巧对 W-Triplet 性能的影响。

Table 4.4 Performance of the proposed method with various training tricks.

W-Triplet	DukeMTMC-Reid		Market-1501	
方法	Rank-1	mAP	Rank-1	mAP
Baseline	79.7	63.7	88.0	73.3
+warm up	79.9	65.8	87.9	75.0
+label smooth	81.5	66.7	90.3	77.1
+last stride=1	82.6	68.0	90.4	78.4
+BNneck	83.8	70.1	93.2	81.5
-Triplet	82.8	67.7	91.7	79.2
+Wasserstein Triplet loss	**85.8**	**72.7**	**94.5**	**83.2**
+random erasing	83.3	71.7	92.4	81.4
+Triplet loss	**87.5**	**76.4**	**94.5**	**85.8**

由于 W-Triplet 强调跨视角对齐匹配问题，而 Random erasing 中的随机数值将会对学习过程造成不利影响。在这种情况下，在使用了除 Random erasing 外的所有训练技巧后，Strong baseline 在 Market-1501 数据集达到了 93.2% 的 Rank-1 accuracy 和 81.5% 的 mAP，在 DukeMTMC-Reid 数据集上，达到了 83.8% 的 Rank-1 accuracy 和 70.1% 的 mAP。在相同的实验设置下，仅将原始的三元组替换为提出的 Wasserstein 三元组损失后，W-Triplet 在 Market-1501 数据集达到了 94.5% 的 Rank-1 accuracy 和 83.2% 的 mAP，在 DukeMTMC-Reid 数据集达到了 85.8% 的 Rank-1 accuracy 和 72.6% 的 mAP，显著提升识别精度。

然而，当实验同时使用传统三元组损失，Wasserstein 三元组损失以及 Random erasing 后，W-Triplet 的性能会略微有所提升。对比原始的 Strong baseline，所提方法在两个基准数据集上表现出了优越的性能。最终，W-Triplet 在 Market-1501 数据集达到了 94.5% 的 Rank-1 accuracy 和 85.8% 的 mAP，在 DukeMTMC-Reid 数据集达到了 87.5% 的 Rank-1 accuracy 和 76.4% 的 mAP。这验证了所提出的 Wasserstein 三元组损失的有效性。

W-Triplet 与基准模型的性能比较：实验对比了 W-Triplet 与对应基准模型的性能比较，结果如表4.5所示，其中 W 表示 Wasserstein 三元组损失。* 表示运用传统三元组损失共同训练。实验选取了 IDE、PCB、Strong baseline 和多分支模型作为基准模型，其中 IDE 与 PCB 遵循 [124] 提出的方法，使用交叉熵损失进行训练，而其余方法则同时使用交叉熵损失和 triplet 损失进行训练，所有的实验

结果都是开源代码运行结果，虽然部分结果与原始论文公布结果略有出入，但不影响实验结论。

表 4.5 W-Triplet 与基准模型的性能比较。

Table 4.5 Performance of the proposed method compared with various baselines.

W-Triplet	DukeMTMC-Reid		Market-1501	
方法	Rank-1	mAP	Rank-1	mAP
IDE	83.3	63.8	75.6	57.9
IDE + W	89.6	70.2	78.9	59.9
PCB	92.1	77.0	83.9	69.9
PCB + W	92.8	77.2	84.5	69.4
S*	94.0	85.4	86.8	76.1
S* + W	94.5	85.8	87.5	76.4
M*	95.0	87.7	88.8	79.2
M* + W	95.5	88.2	89.1	79.8

在使用相同网络架构的基础上，W-Triplet 可以有效地提升基准模型的识别准确率。对于不同基准方法，使用 Wasserstein 三元组损失均可实现 0.5% 左右的 Rank-1 accuracy 和 mAP 提升。对于 IDE 方法，W-Triplet 取得了较大提升，这是由于在该实验中，基准模型并未使用 triplet 进行训练。值得注意的是，W-Triplet 对于较强的基准方法仍然有效，以多分支模型为例，W-Triplet 在 DukeMTMC-Reid 和 Market-1501 数据集上，Rank-1 accuracy 实现了 0.5% 与 0.3% 的性能提升，mAP 实现了 0.5% 与 0.6% 的性能提升。在以 PCB 模型为基础的实验中，W-Triplet 性能的提升相对较为波动，但也能够显著提升 Rank-1 accuracy。

不同三元组损失对性能的影响。为验证本文提出模型的有效性，本节也设计了多组实验以对比不同三元组损失对性能的影响。两种三元组损失（即难样本挖掘三元组损失 [83] 记作 TripletLoss 和 Aligned ReID[132] 记作 AlignedDistance）用于基准模型的训练中，除损失形式外网络架构保持一致。实验选取了 IDE、PCB 和 PCB+RPP 三种模型作为基准模型，其实验结果如表4.6所示。在这三种基准模型基础上 Wasserstein 三元组损失在 Market-1501 数据集上的 Rank-1 accuracy 分别为 89.6%、92.8% 和 93.4%，在 DukeMTMC-Reid 数据集上分别为 78.9%、84.5% 和 83.6%，在 CUHK03 数据集上分别为 38.0%、54.9% 和 60.6%。

注意到，除 PCB+RPP 基准模型外，W-Triplet 的实验结果均达到预期效果。在

表 4.6 不同三元组损失对 W-Triplet 性能的影响。

Table 4.6 Performances of the baseline method with various triplet loss.

W-Triplet	Market-1501			DukeMTMC-Reid			CUHK03		
模型	r-1	r-5	mAP	r-1	r-5	mAP	r-1	r-5	mAP
IDE[124]	83.3	94.4	63.8	75.6	87.4	57.9	39.0	58.0	35.8
IDE+TripletLoss[83]	87.0	95.4	71.7	78.1	88.8	61.4	47.6	69.0	45.5
IDE+AlignedDistance[132]	89.8	96.1	72.8	76.5	87.8	59.3	40.9	60.9	36.9
IDE+W	89.6	96.4	70.2	78.9	89.6	59.9	38.0	57.8	33.6
PCB[124]	92.1	97.2	77.0	83.9	92.1	69.9	59.1	77.4	56.1
PCB+TripletLoss[83]	91.9	97.1	76.4	83.0	91.6	69.4	49.8	70.6	47.3
PCB+AlignedDistance[132]	91.7	96.7	76.1	84.2	91.8	70.8	52.4	72.9	50.2
PCB+W	92.8	97.2	77.2	**84.5**	**91.5**	**69.4**	54.9	73.5	51.2
PCB+RPP[124]	92.9	97.5	81.5	84.3	92.2	71.7	59.4	77.0	56.6
PCB+RPP+TripletLoss[83]	93.1	97.4	81.2	83.7	91.5	70.6	52.9	72.4	51.4
PCB+ RPP+AlignedDistance[132]	93.0	97.4	79.6	84.2	91.9	71.8	55.6	74.2	53.4
PCB+W+RPP	**93.4**	**97.5**	**80.6**	83.6	91.4	69.8	**60.6**	**78.6**	**57.1**

PCB+RPP 上结果不稳定的原因在于其采用了两步分离训练的方式。此外，RPP 采用了一种软池化 (Soft-pooling) 的策略，致使 W-Triplet 在 PCB+RPP 的实现方式上异于 PCB 和 IDE 模型。在 PCB 模型中，实验先得到 Pool5 特征图的空间尺寸为 24×8，随后再生成注意力概率，最后再对其进行平均池化操作；而在 PCB+RPP 模型中，实验为了与局部特征相对应，直接在 RPP 的软池化之后，与局部特征生成注意力概率。因此，W-Triplet 在 PCB+RPP 模型上的实验结果波动较大，未达到预期。

同时，对于 IDE 模型，Aligned Re-ID 的结果显著优于三元组损失 [83]，这表明基于对齐的三元组损失确实能提升模型的性能。此外，分类损失结合度量损失可有效提升识别准确率。但是 Aligned Re-ID 对于 PCB 与 PCB+RPP 等较强的基准方法，成效并不明显，甚至会降低性能。造成这种状况的原因可能是 Aligned Re-ID 计算了不相关局部特征之间的距离。相反，W-Triplet 对于几乎所有的基准方法都可以有效提升性能，但对于 PCB 与 PCB+RPP 等较强的基准方法上，提升有限。

CUHK03 的数据集规模相对较小，因此，使用三元组损失训练的模型相比于仅用分类损失的 PCB 原始模型，性能有所降低。不仅如此，IDE 模型在使用

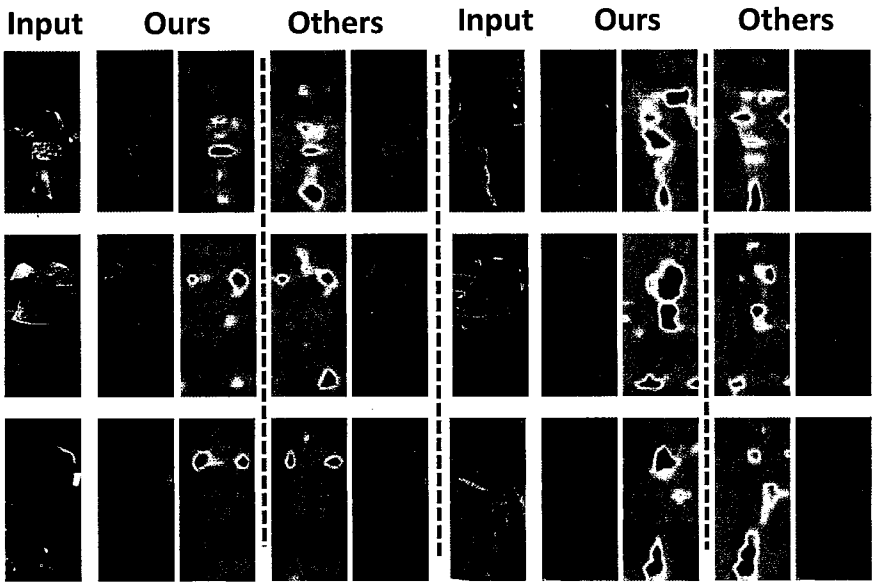


图 4.5 不同注意力方法得到的注意力热图。

Figure 4.5 Attention visualization comparison between ours and other methods.

表 4.7 不同注意力机制对 W-Triplet 性能的影响。

Table 4.7 Performance of the proposed method compared with other attention module.

方法	Rank-1	mAP
Strong Baseline	93.2	81.5
Strong Baseline+W(Parameter-free)	93.4	82.6
Strong Baseline+W(Ours)	**94.5**	**83.2**

三元组损失进行训练时，无法收敛，因此在每次迭代中，实验首先使用随机采样的方式，得到批样本数据，然后利用分类损失预训练模型。而这种训练策略并未在 Wasserstein 三元组损失和 Aligned 三元组损失的训练过程中使用，而这可能会提升模型的识别准确率，因此三元组损失 [83] 对于 IDE 模型，取得了最优性能。值得注意的是，提出的 IDE+W 方法在 CUHK03 数据集上，未达到预期效果，这与 Aligned Re-id 情况类似，这可能是由于训练样本不足导致，此外，SGD 优化方法在此数据集上表现较差，使用 Adam 优化器的基准模型表现稳定。

不同注意力机制对 W-Triplet 性能的影响：对于 Strong baseline，实验还使用了不同的空间注意力方法来实现 Wasserstein 三元组损失，并在 Market-1501 上进行了定性与定量的实验。为了比较不同注意力机制对 W-Triplet 性能的影响，实验使用了一种无参数的注意力方法 [126]。该方法对输入的特征图，沿着通道维度进行平均池化操作，然后将得到的特征图输入到 softmax 层中，产生最终的空

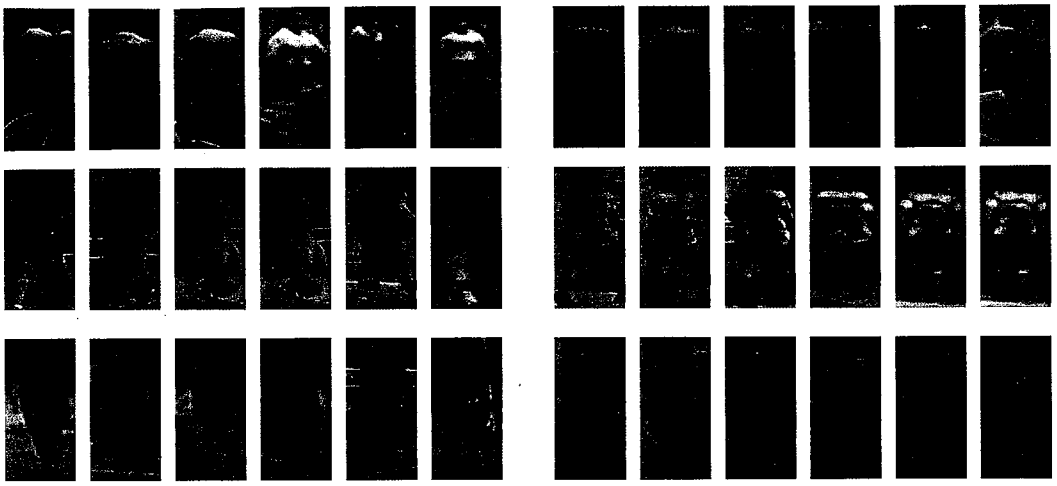


图 4.6 Market-1501 上 PCB 模型产生的注意力热图。

Figure 4.6 The attention map produced by PCB on Market-1501 dataset.

间注意力。由于这种注意力方法与提出方法类似，因此实验选择了该方法进行对比，实验结果如表4.7所示，使用由无参数的注意力模块得到的 W-Triplet，超过了使用传统三元组损失训练的基准模型。而在使用所提出的注意力模块后，模型的性能可进一步得到提升，验证了所提出注意力机制的优越性。同时，实验还对相应的注意力结果进行了可视化展示。如图4.5所示，可以发现，所提出的注意力方法相较于 [126]，更偏向于关注感兴趣的目标区域，而非背景区域，人体轮廓更为清晰。因此，所提注意力方法热图展现出了更有物理意义的指导作用。

4.4.5 可视化结果

实验对网络输出的注意力进行了可视化，该注意力热图由 PCB 模型中的 Pool5 层输出得到，结果如图4.6所示。可以发现，学习得到的注意力热图偏向于关注图像的感兴趣区域，并同时对背景区域进行有效抑制（例如图4.6中的第一行所示）。此外，注意力热图还可以对行人图像的不同部位进行区分，并将其中对识别有益的部分进行突出强调（例如图4.6中第二行左侧与右侧图像），由此生成的注意力概率（对注意力图进行全局池化得到）同样也保持了相同的特性。由此可见，W-Triplet 得到的注意力区域更为可靠，可针对性地解决错位问题（例如图4.6中第三行）。

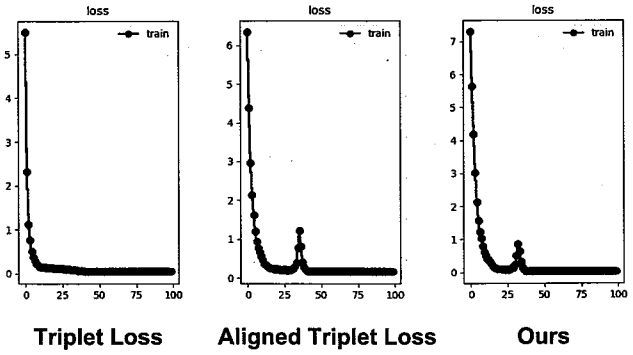


图 4.7 CUHK03 数据集上的 IDE 模型收敛曲线。

Figure 4.7 The convergence curves drawn for IDE baseline on CUHK03 dataset.

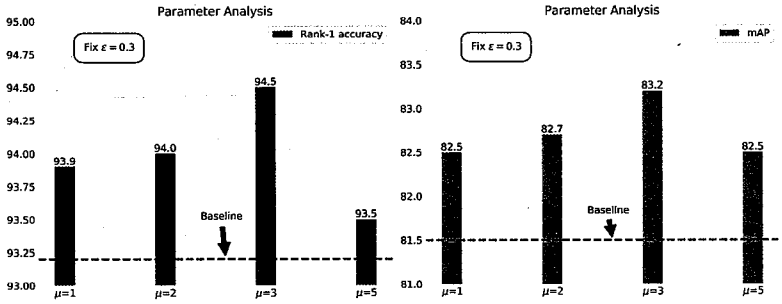


图 4.8 Market1501 数据集上的参数敏感性分析。

Figure 4.8 Sensitivity analysis of parameters on Market1501 dataset.

4.4.6 收敛性与敏感性分析

模型收敛曲线如图4.7所示。对于在 CUHK03 数据集上训练得到的 IDE 模型而言，实验将训练过程中不同三元组损失变化进行了记录，其中，Aligned Re-ID 与 Wasserstein 三元组损失结果较为类似，这两种方法的训练损失均在第 30 个训练周期，表现出一定程度的震荡，其原因在于学习率的减小。

本节还对模型超参数的敏感性进行实验与分析。所有公布的实验结果均经过重复运行所得，以避免过拟合情况发生。两个关键参数（间隔参数 ϵ 与权衡参数 μ）在模型中具有重要作用。在 Market-1501 数据集上的实验结果表明 (未运用 random erasing)，这两个关键参数在很大的范围内能够保持性能的稳定，而且优于基准方法。结果如图4.8所示，其中红色虚线表示由三元组损失 [83] 训练得到的基准模型性能。对于 Strong basline 而言，在各个数据集上，W-Triplet 均在 $\epsilon = 0.3$ 以及 $\mu = 3$ 时达到最优结果。

4.5 本章小结

本章提出了一种深度多视角对齐度量损失，旨在为跨视角局部表示提供合适的度量目标。为了更好地估计局部特征之间的距离，所提方法将原始的跨视角错位对齐问题转换为最优运输问题。通过求解该运输问题，模型可以对齐由注意力机制产生的注意力离散概率，而无需任何额外的监督。通过这种方式，学习得到的最优运输矩阵能够对匹配的局部区域赋予更大的权值，得到加权的对齐局部特征距离，从而加强行人感兴趣区域以及抑制不相关局部部位。实验结果表明，Wasserstein 三元组损失能够有效区分图像中的目标区域，并能显著提升模型性能。

然而，所提方法中还存在一些弊端需进一步改进。例如如何与现有 Re-ID 中的训练技巧相结合，提升模型泛化性，进一步强化深度模型性能。此外，Wasserstein 三元组损失在规模相对较小的数据集上训练并不稳定，如何解决小规模数据训练问题，也是亟待解决的问题。

第 5 章　总结与展望

在“十三五”期间，国家制定了以信息技术为支撑，打造平安城市、智慧城市的战略目标，给智能安防系统提出了新的业务需求。在这种背景下，基于计算机视觉的模式识别技术和机器学习算法得到了广泛的研究和应用。监控视频中的行人重识别技术正成为计算机视觉的热点问题，在安防领域有着广阔的应用前景。本论文从特征和度量的角度，开展面向行人重识别的多视角机机器学习算法与模型研究，设计了多视角特征融合模型、多视角度量学习方法，有效提升了行人重识别的识别精度。研究成果既可以服务于监控场景中目标的相似性学习，又能更具一般性地推广到多视角数据建模和数据挖掘应用。

本论文的研究在“面向大数据的国际特定主题事件推演与风险预警研究”、“警务大数据智能技术应用研究”和“大数据多视图子空间非监督机器学习理论与方法”等项目的支持下开展，研究内容来源于中国通用技术研究院、公安部第三研究所等安防实际业务需求。其中，本文第2章提出的多线性多视角特征融合算法已成功应用于开源情报分析系统中，通过图像搜索和行人搜索的方式，挖掘开源数据中的情报信息。本文第3章和第4章中提出的张量多视角非对称度量学习模型和多视角深度对齐度量学习方法，已成功服务于公安三所开发的智能警务系统中，实现了行人再识别模块。

5.1　主要研究内容与贡献

总结本论文的主要工作与贡献如下：

1. 本文提出了一种新的多线性多视角特征融合算法 (MMF)，MMF 根据不同的特征表示方法，设计不同视角下的特征索引。然后，针对每一个特征索引，引入视角特定的作用矩阵，从而将相似的样本推进，不相似的样本推远。为了得到理想的作用矩阵，算法将作用矩阵堆叠成高阶数组，并在建立的高维数组空间中统一优化所有的作用矩阵，利用高阶低秩约束，比较每一个样本和每一种视角特征，捕获样本与视角之间的多线性关系。针对 MMF 的目标函数，算法给出了优化方法，该优化方法具有收敛速度快且有理论保证等特点。在行人重识别和图像检索上的实验表明，多视角特征融合算法能够有效提升原始特征的判别性，

同时降低在线匹配的计算和内存开销。

2. 本文提出了一种新的张量多视角非对称度量学习模型 (t-MTL)。t-MTL 将每个视角下的行人重识别任务视为一个子任务，通过一对多的多分类学习框架，学习得到多个视角特定的分类器。其中，一个分类器用于区分一个特定的行人，而某个视角下的所有分类器都被堆叠到特定于任务的投影矩阵中。为了获得更好的泛化能力，所提模型通过视角间和视角内的关联结构，学习非对称度量，对齐不同视角下的数据分布。在此基础上，提出了无监督张量多视角度量学习模型，多特征张量多视角模型，分别在不借助行人标记和利用多特征互补信息的情况下，提升识别准确率。在行人重识别的四个公共数据上的实验验证了 t-MTL 的有效性。

3. 本文提出了一种新的多视角深度对齐度量学习方法 (W-Triplet)。W-Triplet 损失首先将网络的中间结果纵向分割成局部特征图，再经由平均池化生成局部特征，并通过注意力机制，在给定图像标注的情况下，学习得到不同区域的注意力热图，生成对应于局部特征的概率分布。在此基础上，W-Triplet 引入最优运输问题，将原有的跨视角样本对齐学习问题转化为跨视角分布对齐问题，给出一种新的样本距离测量方式，并以三元组损失的形式用于模型训练，从而解决跨视角样本错位问题。W-Triplet 还提出多分支深度网络，融合全局和局部信息，提升识别准确率。在行人重识别的四个公开数据集上实验表明，基于推土机距离的三元组损失能够帮助模型学习到目标的感兴趣区域，并依靠感兴趣区域对齐和消除跨视角带来的样本偏差，有效提升深度网络性能。

5.2 进一步研究展望

本文从特征融合和度量对齐的角度，针对行人重识别任务，设计了一系列多视角机器学习算法，在行人重识别基准数据集上取得了一定的提升，但是由于时间和精力所限，除了各章小结中提到的若干后续研究外，仍然有许多问题需要进一步完善和深入探索：

- **深度多视角学习算法**。本文在第4章中介绍了多视角度量模型，该模型设计了一种新的三元组损失来解决跨视角样本错位问题，但模型本身不涉及网络结构的研究。实际上，从特征设计的角度而言，针对局部特征和全局特征的多视角深度网络模型，更有利于行人图像的表示。因此，如何针对全局和局部特征，

设计多视角深度网络，值得进一步研究。

• **半监督多视角机器算法**。现有行人重识别方法的成功，一方面归功于深度学习，尤其是大型卷积神经网络的成功应用，另一方面也是建立在特定监控网络下，足够数据标注的基础之上。当模型迁移到新的监控场景时，效果往往不尽如人意，严重制约了其在实际安防场景中的应用。对于该问题，一种潜在解决方案是利用未标记的场景数据来提升识别的准确率。从迁移的角度来看，不同场景下的数据构成了多视角信息，该半监督多视角机器学习实质上是解决一个视角下有标注，另一个视角下无标注的机器学习问题，如何建立标记数据与未标记数据的关联关系，探索未标记数据的内部结构，是一个有广泛应用潜力的研究课题。

• **跨模态多视角机器学习算法**。跨模态问题是行人重识别的重要问题，也是多视角机器学习重要的研究领域之一。现有的行人重识别模型都是针对可见光图像设计的，当遇到夜间光照不足的情况时，这类方法难以发挥作用。而当前监控系统中，大部分监控探头支持红外图像的获取，因此产生了可见光-红外图像的跨模态行人重识别问题。有别于现有行人重识别的多视角问题，跨模态行人重识别的视角（模态）差异更大，属于不同源信息，如何刻画这种多视角数据的关系，也是亟待解决的热点问题。