浙江大学

本科生毕业设计

文献综述和开题报告

姓名与学号 王兴路 3140102282

指导教师 李英明

年级与专业 2014 级 信息工程

所在学院 信息与电子工程学院

一、题目：深度行人再识别学习

二、指导教师对文献综述和开题报告的具体内容要求：

指导教师（签名）_____

年　　月　　日

目 录

文献综述

一、 背景介绍

行人再识别（Person Re-identification，简称 ReID），也称行人重识别 [1]，如图 1，是利用计算机视觉技术，在图像或者视频集合（gallery）中找到与询问图片（query）相似行人的任务。理论上来说，视频监控中的行人再识别系统应该被分解为三个子模块：行人检测、行人跟踪、行人检索。前两个模块是计算机视觉中已经存在的任务，因此大多研究者所指的行人再识别问题着力解决行人检索问题 [2]。本文也不例外，着力解决光照、姿态、视角急剧变化下的行人再识别/检索问题。

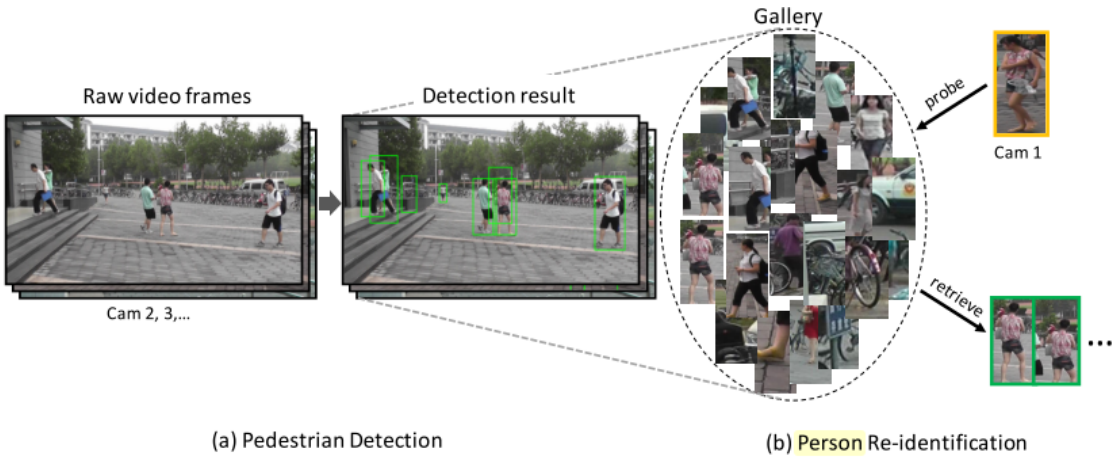

图 1　典型行人再识别系统的流程，整个行人再识别系统包括：视频行人检测、行人跟踪、行人检索。目前的工作主要集中在行人检索问题上。因此大多研究者所指的行人再识别问提指的是其中的子模块——行人检索问题。

行人再识别中存在许多挑战。由于行人的非刚性运动、检测器的误差、摄像头的视角变化，同一行人的不同图片往往存在严重的空间失配（Spatial Misalignment）；行人没有可靠的生物特征，只能从属性、语义层面的特征加以区分；未标定的摄像机参数、巨大的时空跨度，这些都进一步增加了再识别的难度；同时现有的数据集规模相对较小，不存在 ImageNet 或者 MegaFace 这样的大规模、可以泛化迁移（Transfer）到任意子领域（domain）的数据集。这导致数据集间存在较大偏差（domain bias/domain shift），从一个数据集到另一个数据集，模型的性能通常都会下降。

二、 国内外研究现状

由于行人再识别的应用和研究的价值，他在计算机视觉领域受到了越来越多的关注和巨大的发展。近几年，再识别领域投稿数量增加，性能也呈指数增长，2015 年好的模型 cmc-1（Cumulative Matching Characteristic）只有 65%。17 年在 80% – 85%。但是在 17 年短短 11 月，Arxiv 上公布的文章已经达到了 90% – 96%。18 年行人再识别领域的 cmc-1 已经能够在大型数据集上超越人类的表现。目前的研究者大多从几个固定的方面深入研究。从损失函数分类上看，行人再识别主要使用分类损失、度量损失，分别对应表征学习和度量学习。这是再识别的基本方法，在此基础上，国内外的研究者往往从特征学习、网络结构的设计（如引入结构先验——行人由几个部分构成、具有相对空间位置关系）或者提出新的问题（视频再识别、生成数据、大规模检索、弱监督学习）等方面着手。本综述也将从这些方面展开。

1. 表征学习与度量学习

表征学习 (Representation learning)：通过转化为分类 (Classification/ Identification) 问题或者验证 (Verification) 问题，进行监督性学习训练，取全连接层的特征作为很好的表征。之所以说是很好的表征，是因为输入图片原本线性不可分，但是最后一层线性分类层的特征变得线性可分，从而使得后续任务变得简单。w 无论是分类还是验证问题，在转化之后，会使用 softmax 作为激活函数。在有的论文中，作者认为光靠行人的 ID 信息不足以学习出一个泛化能力足够强的模型，于是额外标注了行人图片的属性特征，例如性别、头发、衣着等属性。通过引入行人属性标签，模型不但要准确地预测出行人 ID，还要预测出各项正确的行人属性。

度量学习 (Metric learning)：度量学习广泛应用于图像检索。不同于表征学习，度量学习通过设计距离度量函数，直接学习特征。优化目标直接就是同一行人不同图片特征之间的相似度更小，不同行人的更大。损失函数使得相同行人图片（正样本对）的距离尽可能小，不同行人图片（负样本对）的距离尽可能大。常用的度量学习损失方法有对比损失 (Contrastive loss) [3]、三元组损失 (Triplet loss)、四元组损失 (Quadruplet loss)、难样本采样三元组损失 (Triplet hard loss with batch hard mining, TriHard loss)[4]、边界挖掘损失 (Margin sample mining loss, MSML)

两者最终学到的特征，都是语义上紧凑的表示，能够根据 ID 聚成不同的类别，有

利于后续任务的进行。但是区别在于表征学习是通过定义其他有关联的分类任务间接学到表征的。

	Type	Market-1501 SQ			Market-1501 MQ			MARS		
		mAP	rank-1	rank-5	mAP	rank-1	rank-5	mAP	rank-1	rank-5
TriNet	E	**69.14**	**84.92**	**94.21**	**76.42**	**90.53**	**96.29**	**67.70**	**79.80**	**91.36**
LuNet	E	60.71	81.38	92.34	69.07	87.11	95.16	60.48	75.56	89.70
IDE (R) + ML ours	I	58.06	78.50	91.18	67.48	85.45	94.12	57.42	72.42	86.21
LOMO + Null Space [36]	E	29.87	55.43	-	46.03	71.56	-	-	-	-
DTL† [8]	E	41.5	63.3	-	49.7	72.4	-	-	-	-
DTL† [8]	I+V	65.5	83.7	-	73.8	89.6	-	-	-	-
ResNet 50 (I+V)† [41]	I+V	59.87	79.51	90.91	70.33	85.84	94.54	-	-	-
Gated siamese CNN [31]	V	39.55	65.88	-	48.45	76.04	-	-	-	-
CNN + DCGAN† [42]	I	56.23	78.06	-	68.52	85.12	-	-	-	-
IDE (R) + ML [43]	I	49.05	73.60	-	-	-	-	55.12	70.51	-
IDE (C) + ML [38]	I	-	-	-	-	-	-	47.6	65.3	82.0
CNN + Video† [37]	I	-	-	-	-	-	-	-	55.5	70.2
TriNet (Re-ranked)	E	**81.07**	**86.67**	**93.38**	**87.18**	**91.75**	**95.78**	**77.43**	**81.21**	**90.76**
LuNet (Re-ranked)	E	75.62	84.59	91.89	82.61	89.31	94.48	73.68	78.48	88.74
IDE (R) + ML ours (Re-ra.)	I	71.38	81.62	89.88	79.78	86.79	92.96	69.50	74.39	85.86
IDE (R) + ML (Re-ra.) [43]	I	63.63	77.11	-	-	-	-	68.45	73.94	-

图 2　表示学习、度量学习的方法比较，其中，TriHard 方法获得了最好的效果。在不使用 Rerank 后处理的情况下，得到了 84.92 的 rank-1 指标，在 rerank 和使用 multi-shot 查询的情况下，获得了 91.75 的 rank-1 指标。

表征学习实现简单，当每个 ID 都有充足的训练图片时效果很有竞争力。如图 2，实践证明，在行人再识别、人脸验证这一类问题，度量学习只要训练得当，就能够获得更快的收敛速度与泛化能力。因此，本节将重点叙述表征学习的方法。

对比损失用于训练孪生网络 (Siamese network)，网络含有两条分支，三元损失用于训练三条分支的网络。但是由于各条分支之间参数共享，所以可以使用一条分支高效实现。首先介绍对比损失，假设输入图像 a 和 b，提取到的特征之间的距离定义为

$$d_{a,b} = \| f_{I_a} - f_{I_b} \|^2 \tag{1.1}$$

当两张图属于同一行人时，标签 y=1，不同行人时，标签 y=0，则对比损失为

$$L_c = y d_{a,b}^2 + (1-y)(\alpha - d_{a,b})_+^2 \tag{1.2}$$

三元损失的三个输入分别为锚定图片 (Anchor) a，正样本图片 (Positive) p 和负样本图片 (Negative) n。图片 a 和图片 p 为一对正样本对，图片 a 和图片 n 为一对负样本对。三元组损失表示为：

$$L_t = (d_{a,p} - d_{a,n} + \alpha)_+ \tag{1.3}$$

传统的三元组随机从训练数据中抽样三张图片，由于随机抽取的往往是无用信息，导致训练时间长、收敛慢，而且容易训练崩塌。因而长期以来，研究者普遍认为使用度量损失往往比表示学习中的分类验证损失效果差。[5] 提出了一种基于训练批量 (Batch) 的在线难样本采样方法——TriHard Loss，比所有 17 年所有的最新方法性能和收敛速度都好了一大截。对于每一个训练 batch，随机挑选 P 个 ID 的行人，每个行人随机挑选 K 张不同的图片，即一个 batch 含有 $P \times K$ 张图片。之后对于 batch 中的每一张图片 a ，我们可以挑选一个最难的正样本和一个最难的负样本和 a 组成一个三元组。

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

$$L_t = \frac{1}{PK} \sum_{a \in batch} \left(\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha \right)_+ \tag{1.4}$$

2. 融合局部特征

在行人再识别中，一个很热门的研究方向就是提取更好的局部特征 [6, 7, 8, 9]，18 年性能很高的几篇论文基本上都使用了局部特征 [7, 8]。其中对齐再识别 [10] 使用最短路径自动匹配局部特征，算法复杂度较高，通过辅助全局特征的训练获得了超过人类的表现，详细可以参考文献翻译。

上下文特征 [11] 使用 STN 结合先验定位可变性的行人部分，从原始图片中预测定位参数，从而便于后续的网络将注意力集中于具有潜在语义的身体部分。考虑到视频监控环境下行人的姿态先验——行人通常直立于地面，作者使用了 4 个自由度的仿射变换矩阵，建模了尺度、平移方面的可变性变换。

$$\begin{pmatrix} x_i^{in} \\ y_i^{in} \end{pmatrix} = \begin{bmatrix} s_x & 0 & t_x \\ 0 & s_y & t_y \end{bmatrix} \begin{pmatrix} x_i^{out} \\ y_i^{out} \\ 1 \end{pmatrix} \tag{1.5}$$

由于具有了可变性的特性，学到的身体部分能够减轻视角和背景带来的类内差异。但是由于随机生成的反射矩阵参数会产生巨大形变，作者不得不增加正则项将其限制于先验的附近。最终将全局特征与身体各部分的特征融合，结合分类损失训练。

3. GAN 生成样本

如图 3所示，ReID 有一个非常大的问题就是数据获取困难，截止 CVPR18 deadline 截稿之前，最大的 ReID 数据集只有 1k ID，36k 图片。视频数据集的图片数量当然达到上万，但是冗余较高，有效的图片仍然少于几千张。因此，使用 gan 生成图片提升泛化能力显得非常重要。

Dataset	*MSMT17*	*Duke* [40]	*Market* [38]	*CUHK03* [20]	*CUHK01* [19]	*VIPeR* [8]	*PRID* [10]	*CAVIAR* [3]
BBoxes	**126,441**	36,411	32,668	28,192	3,884	1,264	1,134	610
Identities	**4,101**	1,812	1,501	1,467	971	632	934	72
Cameras	**15**	8	6	2	10	2	2	2
Detector	**Faster RCNN**	hand	DPM	DPM, hand	hand	hand	hand	hand
Scene	**outdoor, indoor**	outdoor	outdoor	indoor	indoor	outdoor	outdoor	indoor

图 3　ReID 常用数据集统计信息，可以看见，ReID 领域公开数据集规模通常较小

试管实验一文 [12] 是第一篇用 GAN 做 ReID 的文章，由于生成的图像质量较差，没有明确的 id 信息，论文使用的是标签平滑的方法，将 label vector 每一个元素的值取为各个类别的均匀分布。生成的图像作为训练数据加入到训练之中，相当于在训练过程中引入了适当的噪声，避免了过拟合提升了泛化能力。

在此基础上，姿态归一化一文 [13] 提出了改进，一方面能够相同 ID 的行人，另一方面克服了行人姿态不同带来的类内差异。

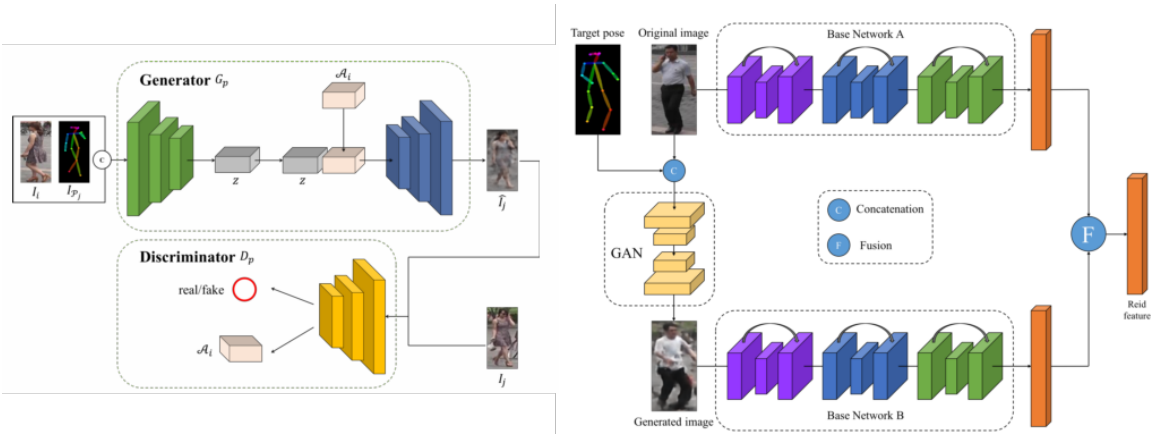

图 4　姿态归一化一文 [13] 所使用的网络结构。左图：最终网络结构——原始图片提取特征与姿态归一化后图片提取的特征融合，用于再识别。右图：生成姿态归一化图片的 GAN 网络结构，融合姿态、ID、属性等信息，生成较高质量的图片。

论文首先使用在 coco 上预训练的 pose estimation 网络提取骨骼关键点，由于数据集之间存在偏差，往往存在很高的误检漏检。将 pose 信息和原图共同作为条件信

息，输入 pixel2pixel 网络，并使用属性和 ID 作为额外的监督信息，将原本无监督的 GAN 生成问题转化为半监督学习，生成了质量较好的行人图片。作者选取了 8 个视角，通过 GAN 将 single query 转化为了 multi query，通过 max-pool 形式，将不同视角的行人图片提取的特征与 baseline 网络提取的特征融合。

最终在 Market1501 上再次获得了超过人类表现的性能，获得了 95.52% 的 rank-1 指标，在 CUHK03 上也获得了不错的性能。在 CUHK03 上性能没有超过人类的原因在于，属性标签只在 Marke1501 有标注，泛化到 CUHK03 上反而会由于数据集的偏差降低性能。同时该方法在半监督学习的设置下也获得了很好的效果，不使用目标数据集调优，在 VIPeR 小数据集上直接得到了 68.67% 的 rank-1 指标。

图 5　在 Market1501 和 DukeMTMC-reID 数据集上生成图片的效果。

三、 其他方向

目前视觉领域逐渐开始从图片转向了视频，对于再识别而言，两者最大的区别在于询问视频序列具有多张图片。对此，常用的有两种方案：多张 query 图片多次查询，聚合排序结果；聚合查询图片的特征，然后只进行一次查询。两种方法相比，后者在大规模检索问题中更具实用性和扩展性。因此再识别领域往往借鉴视频动作识别中的想法将多张图片的特征聚合。比如捕捉特征随时间的演化模式、通过在 CNN 中嵌入三维卷积直接得到视频级别的特征描述。

累计运动背景网络 (Accumulative motion context network, AMOC)[14] 使用视频动作识别中效果最好的双流网络提取空间特征和光流（运动）特征，融合后输入到一

个 RNN 来提取时序特征。其中，由于光流运动信息的提取非常耗时，作者首先训练了一个类似 U-Net 结构的运动信息网络，输入原始图像序列，前馈网络预测不同尺度的光流图像拼接为最终的预测光流。通过 AMOC 网络，每个图像序列都能被提取出一个融合了内容信息、运动信息的特征。网络采用了分类损失和对比损失来训练模型。融合了运动信息的序列图像特征能够提高行人重识别的准确度。

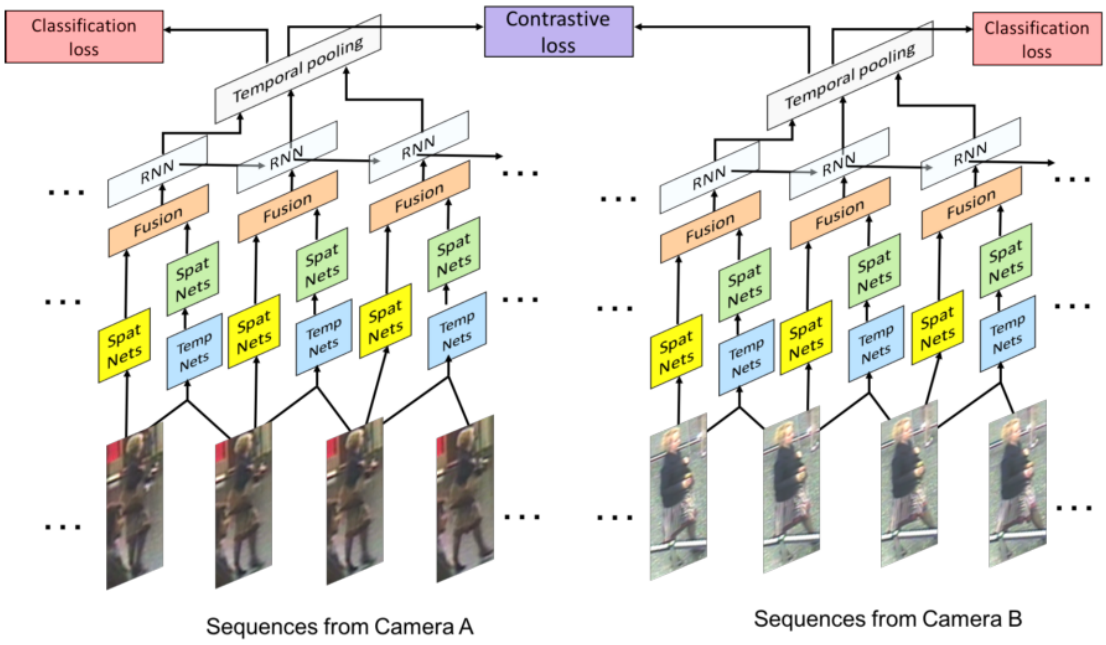

图 6　累计运动背景网络结构

另一方面，行人再识别领域的 gallery 集合中图片数量不断扩大，从传统数据集的 100k 增长到了 500k，开始逐渐迈向大规模检索的现实场景。随着 gallery 集合的增大，会带来两方面的挑战：性能的下降和速度的下降。

性能方面，由于迷惑项的增加，mAP 降幅最大可达 7%。同时为了达到实时的速度，在性能与速度的权衡后，我们往往采用近似最近邻搜索，牺牲一定的精度，导致性能进一步下降。

速度方面，我们非常需要紧凑的特征表达，将一张图片映射到 2048 维的语义空间和 128 维的语义空间，在距离矩阵的计算速度上会有巨大差别。从 500k 到 10M 的 gallery 集合，耗时会增加 60.7s！于是我们有必要从图像检索领域借鉴想法。但是与图像检索有所区别的是，在训练阶段再识别问题被标注有 ID 信息，可以作为分类问题；而在测试阶段，再识别问题完全变成了检索问题。除了算法上存在难题，数据集上也还没有达到图像检索的规模。于是近期，悉尼大学的学者将会提出一个具有更多迷惑项、强调检索实时性的行人再识别数据集。

参考文献

[1] Zheng L, Yang Y, Hauptmann A G. Person re-identification: Past, present and future[J]. arXiv preprint arXiv:1610.02984, 2016.

[2] Zheng, Liang and Zhang, Hengheng and Sun, Shaoyan and Chandraker, Manmohan and Tian Q. Person Re-identification in the Wild[J/OL]. arXiv Prepr., 2017. http://dx.doi.org/10.1109/CVPR.2017.357.

[3] Varior R R, Haloi M, Wang G. Gated siamese convolutional neural network architecture for human re-identification[C] // European Conference on Computer Vision. 2016 : 791 – 808.

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

[5] Liu Y, Yan J, Ouyang W. Quality aware network for set to set recognition[C] // Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.. 2017 : 5790 – 5799.

[6] Zhong Z, Zheng L, Cao D, et al. Re-ranking Person Re-identification with k-reciprocal Encoding[J/OL]. CoRR, 2017, abs/1701.08398. http://arxiv.org/abs/1701.08398.

[7] Liu X, Zhao H, Tian M, et al. Hydraplus-net: Attentive deep features for pedestrian analysis[J]. arXiv preprint arXiv:1709.09930, 2017.

[8] Zhao H, Tian M, Sun S, et al. Spindle net: Person re-identification with human body region guided feature decomposition and fusion[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 : 1077 – 1085.

[9] Wei L, Zhang S, Yao H, et al. GLAD: Global-Local-Alignment Descriptor for Pedestrian Retrieval[J/OL]. CoRR, 2017, abs/1709.04329. http://arxiv.org/abs/1709.04329.

[10] Zhang X, Luo H, Fan X, et al. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification[J/OL]. CoRR, 2017, abs/1711.08184. http://arxiv.org/abs/1711.08184.

[11] Li D, Chen X, Zhang Z, et al. Learning Deep Context-aware Features over Body and Latent Parts for Person Re-identification[J/OL]. CoRR, 2017, abs/1710.06555. http://arxiv.org/abs/1710.06555.

[12] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[J]. arXiv preprint arXiv:1701.07717, 2017, 3.

[13] Qian X, Fu Y, Wang W, et al. Pose-Normalized Image Generation for Person Re-identification[J]. arXiv preprint arXiv:1712.02225, 2017.

[14] Liu H, Jie Z, Jayashree K, et al. Video-based person re-identification with accumulative motion context[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017.

开题报告

一、 研究开发的背景、意义与目的

1. 背景介绍

大多数研究者所指的行人再识别目标是跨摄像头识别同一行人 [4]。由于姿态、遮挡和背景干扰的存在，再识别任务急需一种强大的特征表示，具有较小的类内间距和较大的类间间距。对此，我们可以从全局特征、局部特征的融合，或者采用全新的训练方式，提升特征的鲁棒性和表达能力。

2. 本研究的意义和目的

从多个角度研究行人再识别，通过广泛阅读行人再识别方面的文献和动手实验，了解行人再识别中真正有用的关键技术。同时也阅读相关领域的论文，行人再识别领域一个很明显的趋势就是不停地从其他领域借鉴新问题和新方法，从人脸验证到图像检索再到信息检索（数据挖掘）。目前人脸验证中的开放环境验证（Open-set Verification）、各种 loss 函数、使用 GAN 生成多视角的样本，图像检索中的迷惑图片（Distractor）、大规模信息检索中速度与精度的权衡、可缩减的特征表示 [2]、rerank 后处理 [3] 等思路都已经引入行人再识别。信息检索、web 搜索中还有一些概念可能还没有引入，比如专家（Expert/Local Model）、社区（Community）。最后还可以进行的方向包括：结合认知科学中对人类视觉皮层结构的研究成果，提出可变性（Deformable）、动态（Dynamic、Running Time）的模块，提取鲁棒的特征，利用结合 IRGAN 利用对抗网络挖掘最具竞争力的样本，利用强化学习选择最富有信息的样本，使用记忆网络摒弃所有 point2point 的方法将 set2set 的问题 [10] 真正地解决。我们也将从行人再识别的各个方面着手，先达到 State of the Art 的效果，再尝试提出自己的想法。

二、 主要研究开发内容

1. 主要研究内容

主要研究行人再识别中的起作用的关键技术。包括再识别中的基础技术：siamese network、match network、histogram 特征的提取，再识别中起作用的技术：Triplet loss、online hard negative mining、rerank 后处理，以及最后在 IRGAN、NTM 方面继续尝试。

2. 技术路线

从基本方法开始，首先熟悉再识别数据集和任务，然后尝试复现达到最新论文的效果，最后基于之前积累的经验，选择几个最有可能成功的方向，提出想法、进行试验从而能有自己的创新。

3. 可行性分析

分为三个阶段，从熟悉再识别任务，到复现最新方法，再到提出自己的方法，前面两个阶段比较基础，最后一个阶段具有挑战性。针对发表论文的目标而言，这样的技术路线是否可行？从研究的问题上来说，行人再识别作为新兴的研究方向投稿论文增加，本身课题具有可行性；从使用的方法上来说，我们可能会采用 IRGAN、RL、GAN、NTM 等方法，或者提出一些基础通用的模型，具有可行性。

针对学习最新方法的目标而言，这样的技术路线是否可行？一方面，我们直接从开源代码、文档、最新论文着手，可以在最短时间内掌握一个最前沿的方向。另一方面，虽然我们没有时间系统阅读 Deep Learning、Reinforcement Learning: An Introduction、Element Of Statistics Learning 或者学习一些公开课，系统地了解深度学习领域的知识，但是我们会取长补短，仔细选择和仔细阅读相关部分论文，方便撰写论文。具有可行性。

三、 进度安排及预期目标

1. 进度安排

(1) 熟悉典型的行人再识别数据集

行人再识别中大型的数据集包括 CUHK03、Market1501，每个数据集都有自己强调的创新之处，比如 CUHK03 强调图片数量和行人数目是 15 年最多的数据集，Market1501 在此基础上强调使用 DPM 检测器，图片质量具有自然场景下应有的噪声与挑战。比较不同的数据集主要可以从行人 ID 数据、图片数据、训练测试数量、gallery 集合大小、数据质量方面衡量。这一阶段，重点掌握各个数据集的共性，准备好各个数据集调用的同一接口，从而方便之后的研究对比实验。

表 1 常用数据集的统计特性

cuhk03.combine	#ids	#images	market1501	#ids	#images
train	1267	12183	train	651	11281
val	100	948	val	100	1655
trainval	1367	13131	trainval	751	12936
query	100	965	query	750	16483
gallery	100	965	gallery	751	19281

当然目前行人再识别领域数据集频出，各种新的任务也不断出现。传统的 CUHK01、VIPeR 通常只在弱监督、迁移学习中会被研究和使用。视频 ReID 任务常常使用 PRID 2011，iLIDS-VID 和 MARS。为了衡量每帧图片质量，Sensetime 新推出了 Labeled Pedestrian in the Wild (LPW) 数据集，其中包含 7,694 个 tracklets，超过 590,000 个图片。为了衡量现实场景下行人搜索任务的性能，悉尼大学的 Zheng Liang 提出了 PRW 数据集。为了研究迁移学习，比现有数据集规模再大 3.5 倍的 MSMT17 即将开源，该数据集在时空跨度、行人多样性、背景干扰方面的挑战更大。如果有余力，也可以整理一下这方面的数据接口，有助于将来进行 Ablation Study。

(2) 广泛阅读文献，熟悉典型的行人再识别方法

使用 CNN 做行人再识别的典型方法包括：改进的行人再识别（ImprovedReID）、基于深度语义特征的行人再识别（DCSL）、基于行人局部特征的行人再识别（Part-reid）。从这些方面的论文着手，了解以前和现在的研究者对行人再识别常用方法

存在怎么样的理解、分类甚至偏见。比如悉尼大学的 Zheng Liang 观察到当每个 ID 的训练样本达到 10 张以上时，基于 identification 方法（IDE，identity discriminative embedding）的模型能很轻松地实现高准确率，而基于 Siamese 模型的 Verification 方法，每次只能看见两个样本判断相似与否，难以完全利用所有 ID 的标注信息。因而该组的学者后续提出的方法的 baseline 都是 IDE 模型。这种方法简单有效，但是该组学者没有探索通过挖掘富有信息的样本尽可能利用所有 ID 标注信息，提升模型泛化能力的方法。因而，我们需要广泛阅读世界各地研究者对再识别的见解，获得全方位的了解。

(3) 通过实验熟悉行人再识别方法，复现最新工作

TriHard、Part-reid 等方法都有开源代码，我们会基于这些工作进行实验，复现最新工作的效果。第一步，通过实验，理解这些方法为什么能起作用，存在什么缺点。现有模型往往在较小随机选择的训练集子集上达到近乎完美的效果（cmc-1 为 99%），但是在测试集上泛化能力不佳（cmc-1 为 80% 85%）。但是事实上，如果使用整个训练集合而非子集合来测试模型的能力，我们发现 mAP 会下降 3.4%。这说明：一方面，训练集合存在噪声——标注错误，异常数据（Outlier）；另一方面，训练集中还存在许多富有信息的样本，在随机选择的过程中被忽略。TriHard 的出发点就在于在一个 batch 里在线地挖掘富有信息的样本。但是这一工作其实具有一定的局限性，比如是否每一个 anchor 样本都需要寻找对应的难样本、寻找几个。当一个随机选择的 batch 无法提供更多的信息时，是否需要通过其他方法来寻找最富有信息量的样本。

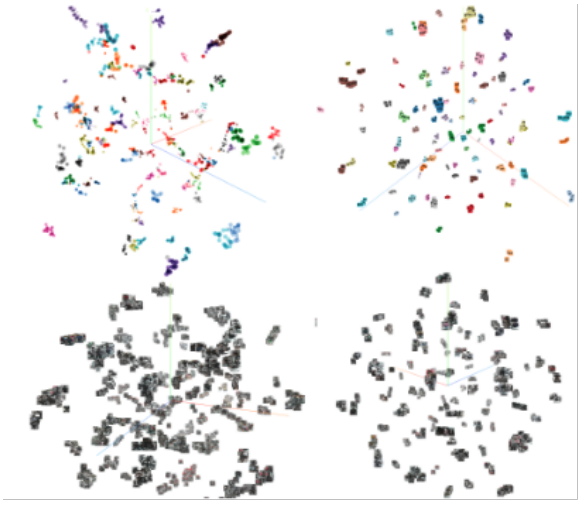

图 1　测试集（左图）与训练集合的子集（右图）特征降维可视化。

我们也会尝试将在线难样本挖掘用于含有匹配网络的双孪模型。含有匹配网络

的双孪模型增大对齐能力的同时也增大了计算量。在这种模型上实现在线难样本挖掘只需要多进行 128*128 次前向计算（增加 1.6s/iter 左右的时间），选择最富有信息的样本对即可。这类模型的优点在于可学习的匹配函数，摆脱了欧氏距离的缺点。我们知道目前含有匹配网络的双孪模型较好的水平为 80%（在 CUHK03 上，cmc-1 指标），我们有信心使用在线难样本挖掘将他提升到 85.43%。而用了在线难样本挖掘的三元损失模型（TriHard）只有 85%。因此如果进一步在损失函数、模型结构上扩展，我们也许有机会达到更高的水平。如果按着这一研究方向进行，我们尝试引入一些非线性模型的想法。

在局部特征与全局特征的融合方面，我们会尝试使用 Part-reid 方法。使用时注意观察 TriHard 中提到的注意事项会有怎样的影响。比如激活函数的使用，ReLU，L2 Normalize，Sigmoid 函数究竟该不该用、用在哪里。小心一些不合理的设置，比如 BatchNorm 之前和之后同时使用 ReLU 函数，当模型变得复杂时，很容易出现这些简单的错误。这时候不能归咎于方法不起作用，而应该通过单元测试、输出中间变量可视化等方式找到错误。同时这也有助于我们了解模型的特点和优缺点。

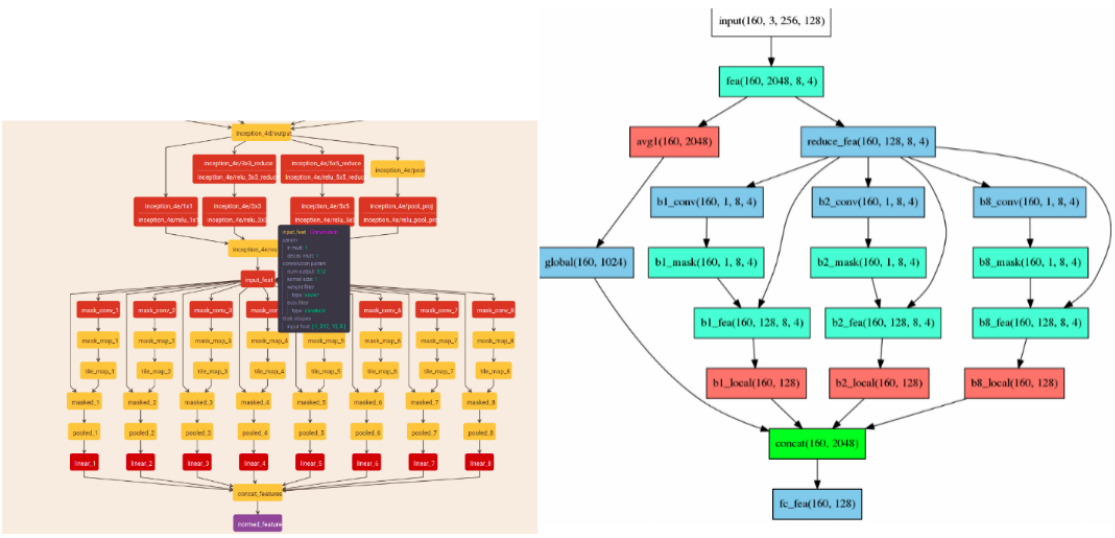

图 2　Part-reid 的网络结构，左图：论文中的使用的结构。右图：我们计划实现的版本

(4)　提出我们的方法

我们将在观察到的现象的基础上，提出我们的方法。虽然想法很重要，决定了一篇文章是否有影响力，但是目前很难下定论，什么样的想法能够起作用。按照最初课题的方向，我们可以从 GAN、Video 的角度着手，比如使用 IRGAN 选择富有信息的样本，GAN 的训练比较困难会遇到训练崩溃的情况，对此一定要耐心调试、多吸取前

人的经验。针对 Video 数据，一方面可以提取时序特征，作为行人的描述子；另一方面，可以将问题转化为 multi-shot 的问题，考虑每帧图片的质量的基础上进行特征融合。同时也可以在其他方面着手，比如目前的 reid 研究都强行将 set2set 的问题 [1] 转化为 point2point，无论训练还是测试都只比较两张图片的相似度。这才让 rerank 有机可乘。可以尝试结合 NTM 将 rerank 利用上下文信息的步骤端到端地结合到网络中，从而在一定意义上实现自动的 single-shot 到 multi-shot 的转换。同时我们也会在选择样本、特征提取、基础卷积模块的设计上着手，选择最有效果的方向深入研究。

2. 预期目标

1. 熟悉典型的行人再识别数据集：编写统一的调用接口。
2. 广泛阅读文献，熟悉典型的行人再识别方法：全面了解行人再识别问题和解决方法。
3. 通过实验熟悉行人再识别方法，复现最新工作：达到前沿方法的效果。
4. 提出我们的方法：通过对模型的了解，尝试不同方法，提出新的想法。

四、 参考文献

[1] Y. Liu, J.Yan, andW. Ouyang. Quality aware network for set to set recognition. arXiv preprint arXiv:1704.03373, 2017

[2] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. 2017.

[3] Zhong, Z., Zheng, L., Cao, D., & Li, S. (2017). Re-ranking Person Re-identification with k-reciprocal Encoding. http://arxiv.org/abs/1701.08398

[4] Zheng, L., Yang, Y., & Hauptmann, A. G. (2016). Person Re-identification: Past, Present and Future. http://arxiv.org/abs/1610.02984

[5] Wang, J., Wang, B., Zhang, P., & Zhang, D. (2017). IRGAN : A Minimax Game for Unifying Generative and Discriminative Information Retrieval Models.

外文翻译

文献原文：

Zhang X, Luo H, Fan X, et al. AlignedReID: Surpassing Human-Level Performance in Person Re-Identification, 2017.

对齐再识别：超越人类的表现

摘要

在这篇论文中，我们提出了一种全新的方法，叫做对齐再识别。最终部署时只使用全局特征，全局特征的鉴别能力是通过与局部特征的共同学习获得的。在共同学习的过程中，局部特征通过计算最短路径对齐和匹配不同人体区域，不需要额外的监督信息。在学习完成之后，我们只使用全局特征计算图像的相似度，进行检索。我们的方法在 Market1501 数据集上获得了 94.0% 的 rank-1 准确率，在 CUHK03 上获得了 96.1% 的 rank-1 准确率，均比目前最好的方法提升了很多。我们也评估了人类所能达到的极限，发现我们的方法首次在两个公开数据集上超过了人类表现！

一、 介绍

行人再识别，是计算机视觉中的一个子任务，目标是在不同的时空鉴别出目标行人。他应用广泛，从跨摄像头跟踪，到销售商店客流量分析，均可发挥巨大作用。同很多视觉任务一样，再识别的难点也在于视角，光照的变化和遮挡的存在。传统的方法通常在底层特征上着力。但自从深度学习复兴之后，卷积神经网络成为了这一难题的主流方法，通过端到端的特征学习，以及丰富的度量学习损失函数设计（比如对比损失，三元损失，四元损失和难样本挖掘三元损失），这一方法达到了前所未有的准确率。很多卷积神经网络的方法仅仅考虑学习全局特征，而忽略了行人的空间结构。这些方法主要有以下缺点：

1. 不准确的行人检测会影响特征学习。

2. 行人被遮挡时，全局特征会引入无关的背景信息。
3. 视角变化和非刚性物体的形变使度量学习更加困难。
4. 外观相似但并非同一人的情况大量存在，这时候需要突出局部特征才能鉴别出这类行人。

为了解决上述问题，近期的一些工作在划分人体区域学习局部特征上着手，但是仍然无法解决检测不准确、姿态变化、遮挡等难题。也有研究者从姿态估计着手，但是这种方法需要额外的监督信息，同时在姿态估计的步骤中往往更容易发生错误。在本文中，我们提出了一种全新的对齐再识别的方法，虽然也是学习全局特征。但是在学习的过程中使用局部特征对齐协同学习，不需要额外的监督信息或者显式的姿态估计。在学习阶段，我们有两条支路分别学习局部特征和全局特征。在局部特征支路，我们引入了最短路径损失。在部署使用阶段，我们只使用全局特征。因为我们发现只使用全局特征已经和多种特征融合的效果同样好！这从另一方面说明，在协同训练的过程中，局部特征帮助全局特征变得更有鉴别能力。同时简洁的全局特征使得我们的方法便于部署到大尺度行人再识别系统中。

二、 方法

1. 对齐再识别

对于每张输入图片，我们使用 Resnet50 提取特征，得到 2048X7X7 的特征图，然后使用全局池化得到 2048 维的全局特征。对于局部特征我们首先使用使用垂直方向的池化得到 2048X7 的特征图，然后使用动态匹配局部区域，得到最小距离，作为局部特征之间的距离度量。这样的特征提取方式隐含着人在图片中通常直立的假设。给定两幅图片的局部特征，在我们的实验中 H 为 7，即行人从头到脚被分为 7 个部分。我们首先计算两两之间的距离矩阵：

$$d_{i,j} = \frac{e^{\|f_i - g_j\|_2} - 1}{e^{\|f_i - g_j\|_2} + 1} \qquad i, j \in 1, 2, 3, \cdots, H \tag{3.1}$$

采用这样形式的变换的原因是：改距离可以归一化到 [0,1] 之间。然后定义局部特征时间的距离为距离矩阵 D 中，从 (1,1) 到 (H,H) 之间的最短路径。改最短路可以由动态规划计算得到。如图 1 所示，图像 A 和图像 B 是同一行人的不同视角的图片。身体不同部分的对齐，比如图 A 的第一部分和图 B 的第 4 部分，在最短路径的第一

个转折中体现出来。同时，也存在一些不对应部分的对齐，比如图 A 的第一部分和图 B 的第一部分。我们认为，不对应部分的对齐是有助于维护垂直对齐的顺序的。不对应部分有着更大的 L2 距离，梯度更接近于 0，英尺这些对齐在最短路中贡献可以忽略。因此最短路的总距离主要由对应部分的对齐决定。

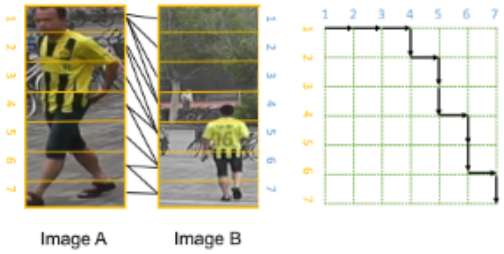

图 1　图像 A 和图像 B 是同一行人的不同视角的图片。身体不同部分的对齐，比如图 A 的第一部分和图 B 的第 4 部分，在最短路径的第一个转折中体现出来。

全局和局部特征共同定义了图像在学习阶段的距离。我们选择 TriHard 损失作为度量损失函数。对于每一个样本，我们会根据全局特征之间的距离选取最有竞争力的样本对，获得 3 元组。而三元组的损失则是根据全局特征距离和局部特征距离共同计算的。这样做的原因出于两点考虑：1. 全局距离的计算更加快；2. 我们观察到使用全局距离和局部距离作为难样本挖掘的手段，没有很大的差别。

2. Mutual Learning 应用于度量学习

我们将 mutual learning 应用于对齐再识别中，来进一步提升效果。在知识蒸馏一文中，作者使用小的学生网络学习大的教师网络中的知识。他的工作用于分类任务中，与他的工作不同，我们采用矩阵级别的均方误差作为新提出的相互学习损失，用于度量学习的任务中。

总体的损失由度量损失、分类损失、分类相互学习损失和度量相互学习损失构成。其中，度量损失是由局部和全局特征距离共同决定的，但是度量相互损失仅由全局特征距离决定。下面介绍度量相互损失的实现：给定一个 batch 的图片，可以使用全局特征计算出距离矩阵。使用 ResNet50 和 Resnet50-Xception 作为基础骨架模型，得到两个不同的学生模型。使用距离矩阵的均方误差作为损失函数，使得两个学生相互学习。在实践中，我们发现使用停止梯度的操作可以加快收敛。即停止 Resnet50 学生网络的梯度更新，将其作为常数，使得 Resnet50-Xception 学生逼近 Resnet50，以及用相同的方式让 Resnet50 逼近 Resnet50-Xception。

三、 实验

1. 数据集

Market1501 数据集包含 32668 张图像，1501 个行人，6 个摄像视角。训练集中有 750 个行人，测试集中有 751 个行人。与提出数据集的作者相同，我们也将 mAP 作为我们的指标的一种。

CUHK03 包含 13164 张图片，1360 个行人，他同时提供了 DPM 检测器检测的行人和手工标注的行人。

MARS 数据集是 Market1501 的扩展版本，所有的行人都是自动检测的。因此可能包含一些虚警增加难度，同时每个 ID 可能包含多余 1 个候选答案。他包含 20478 个 tracklet 和 1261 个行人，6 个拍摄视角。

CUHK-SYSU 是一个大尺度行人搜索的数据集，包含 18184 张图片，每张图中含有多个行人，共计 99809 个行人检测框，8432 个行人。

注意我们使用全部数据集的样例训练单个模型，对于 MARS、CUHK-SYSU 和 Market1501 数据集我们使用官方的训练和评估协议。但是在 CUHK03 上，由于我们使用了全部 benchmark 的数据集训练了一个模型，和官方的测试方式（随机划分数据集为训练集和测试集，测试包含 100 个行人，重复 20 次）有些不同。因此，我们仅仅随机划分数据集一次，训练集与其他数据集的训练集一同作为整个模型的训练集，将测试集作为 CUHK03 上衡量 CUHK03 数据集性能的测试集。由于我们划分的测试集含有 200 个行人，可以认为这个评估协议比原始的协议更具有挑战性。

2. 实现细节

我们使用 Resnet50 和 Resnet50-Xception 制品为基础模型。输入图片缩放到 224X224。数据增强使用随机水平翻转和随机截取。TriHard 损失的 margin 设为 0.3，mini-batch 设为 128，在一个 batch 中确保每个 id 的图片有 4 张。每个 epoch 包含 2000 个 mini-batch。我们采用 adam 优化器，初始学习率设为 0.001，在第 80 和第 160 个 epoch 时，衰减 0.1 训练至收敛。对于对偶学习，相互分类损失的权重设为 0.01，相互度量损失的权重设为 0.001，使用 adam 优化器。初始学习率设为 0.0003，分别在 60 和 120epochs 时衰减到 0.0001 和 0.00001。

3. 对齐再识别的优点

我们首先定性分析对齐的效果。在图 2 的（a）中，由于右边行人的检测不准确，两张原始图片原本无法对齐。对齐再识别算法将左边第一部分和右边的上上个部分匹配在一起，实现了对齐。在（d）中两个行人外观相似，但属于不同 id，右边行人衣服上的标志找在左边找不到相似的部分，因此最短路中对应部分的路径也变得更长。

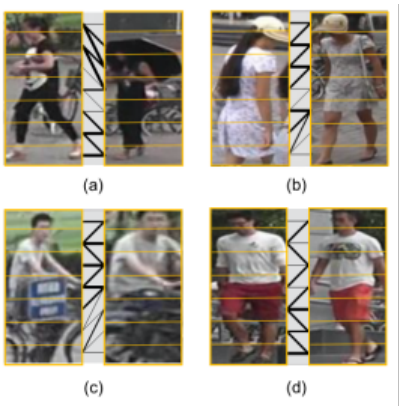

图 2　黑色线条表示两个新人局部的对齐：加粗表示对最短路贡献大。在（a-c）中行人有着相同的 id，而在（d）中行人有着不同的 id。

然后我们将对齐再识别与基准模型对比。基准模型没有使用局部特征支路。两个结果使用相同的网络和训练参数设置。我们观察到对齐再识别方法能提升 3.5% 到 6.0% 的 rank-1 指标和 5.0% 到 8.4% 的 mAP 指标。我们发现将局部特征和全局特征放在一起，rank-1 指标只进一步提升了 0.3% 到 0.5%，但该方法会变得更加耗时，因此我们推荐只是用全局特征。

4. 与其他最新方法比较

在 Market1501 数据集上，使用 rerank 的 GLAD 达到了 89.9% 的 rank-1 指标和 81.1% 的 mAP。而我们的对齐再识别在没有使用 rerank 的情况下就达到了 92.6% 的 rank-1 指标和 82.3% 的 mAP。如果使用 rerank，则进一步提升到 94% 和 91.2%

在 CUHK03 数据集上，在不使用 rerank 的情况下 HydraPlus-Net 获得了 91.8% 的 rank1，而我们的对齐再识别为 91.9%。再次强调，我们的测试候选集为 200 张，我们的测试协议更加困难。同时在使用 rerank 的情况下，我们的方法获得了 96.1% 的 rank-1。

Methods	mAP	r=1
Temporal [23]	22.3	47.9
Learning [47]	35.7	61.0
Gated [32]	39.6	65.9
Person [5]	45.5	71.8
Re-ranking [57]	63.6	77.1
Pose [52]	56.0	79.3
Scalable [1]	68.8	82.2
Improving [16]	64.7	84.3
In [13]	69.1	84.9
In (RK)[13]	**81.1**	86.7
Spindle[50]	-	76.9
Deep[49]*	68.8	87.7
DarkRank[4]*	74.3	89.8
GLAD[37]*	73.9	**89.9**
HydraPlus-Net[20]*	-	76.9
AlignedReID	82.3	92.6
AlignedReID (RK)	**91.2**	**94.0**

Methods	r=1	r=5	r=10
Person [15]	44.6	-	-
Learning [47]	62.6	90.0	94.8
Gated [32]	61.8	-	-
A [34]	57.3	80.1	88.3
Re-ranking [57]	64.0	-	-
In [13]	75.5	95.2	99.2
Joint [42]	77.5	-	-
Deep [10]*	84.1	-	-
Looking [2]*	72.4	95.2	95.8
Unlabeled [56]	84.6	97.6	98.9
A [55]*	83.4	97.1	98.7
Spindle[50]	88.5	97.8	98.6
DarkRank[4]*	89.7	**98.4**	**99.2**
GLAD[37]*	85.0	97.9	99.1
HydraPlus-Net[20]*	**91.8**	**98.4**	99.1
AlignedReID	91.9	98.7	99.4
AlignedReID (RK)	**96.1**	**99.5**	**99.6**

图 3　左图：在 Market1501 数据集上的性能比较，右图：在 CUHK03 数据集上的性能比较

5. 总结

在这篇文章中，我们阐明了隐式局部特征对齐对提升全局特征学习的巨大作用。这一惊奇的结果给予了我们重要启发：1). 端到端的学习需要结构先验，否则就是盲学。2). 尽管我们的方法在 Market1501 和 cuhk03 数据集上超过了人类的表现，但是机器想要在更广泛的领域战胜人类仍然有很长的路要走，比如在验证集中我们发现了一些低级的错误。

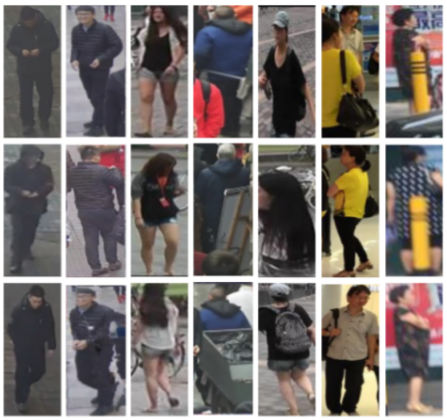

图 4　在验证集中发现的一些低级错误

AlignedReID: Surpassing Human-Level Performance in Person Re-Identification

Xuan Zhang[1]*, Hao Luo[1,2]*‡, Xing Fan[1,2]‡, Weilai Xiang[1], Yixiao Sun[1], Qiqi Xiao[1],
Wei Jiang[2], Chi Zhang[1], Jian Sun[1]
[1]Megvii, Inc. (Face++) [2]Zhejiang University
{zhangxuan,xiangweilai,sunyixiao,xqq,zhangchi,sunjian}@megvii.com
{haoluocsc,xfanplus,jiangwei_zju}@zju.edu.cn

Abstract

In this paper, we propose a novel method called AlignedReID that extracts a global feature which is jointly learned with local features. Global feature learning benefits greatly from local feature learning, which performs an alignment/matching by calculating the shortest path between two sets of local features, without requiring extra supervision. After the joint learning, we only keep the global feature to compute the similarities between images. Our method achieves rank-1 accuracy of 94.0% *on Market1501 and* 96.1% *on CUHK03, outperforming state-of-the-art methods by a large margin. We also evaluate human-level performance and demonstrate that our method is the first to surpass human-level performance on Market1501 and CUHK03, two widely used Person ReID datasets.*

1. Introduction

Person re-identification (ReID), identifying a person of interest at other time or place, is a challenging task in computer vision. Its applications range from tracking people across cameras to searching for them in a large gallery, from grouping photos in a photo album to visitor analysis in a retail store. Like many visual recognition problems, variations in pose, viewpoints illumination, and occlusion make this problem non-trivial.

Traditional approaches have focused on low-level features such as colors, shapes, and local descriptors [9, 11]. With the renaissance of deep learning, the convolutional neural network (CNN) has dominated this field [24, 32, 6, 54, 16, 24], by learning features in an end-to-end fashion through various metric learning losses such as contrastive loss [32], triplet loss [18], improved triplet loss [6], quadruplet loss [3], and triplet hard loss [13].

Figure 1. Challenges in ReID: (a-b) inaccurate detection, (c-d) pose misalignments, (e-f) occlusions, (g-h) very similar appearance.

Many CNN-based approaches learn a global feature, without considering the spatial structure of the person. This has a few major drawbacks: 1) inaccurate person detection boxes might impact feature learning, e.g., Figure 1 (a-b); 2) the pose change or non-rigid body deformation makes the metric learning difficult, e.g., Figure 1 (c-d); 3) occluded parts of the human body might introduce irrelevant context into the learned feature, e.g., Figure 1 (e-f); 4) it is non-trivial to emphasis local differences in a global feature, especially when we have to distinguish two people with very similar appearances, e.g., Figure 1 (g-h). To explicitly overcome these drawbacks, recent studies have paid attention to part-based, local feature learning. Some works [33, 38, 43] divide the whole body into a few fixed parts, without considering the alignment between parts. However, it still suffers from inaccurate detection box, pose variation, and occlusion. Other works use pose estimation result for the alignment [52, 37, 50], which requires additional supervision and a pose estimation step (which is often error-prone).

In this paper, we propose a new approach, called AlignedReID, which still learns a global feature, but performs an automatic part alignment during the learning, without requiring extra supervision or explicit pose estimation. In the

*Equal contribution
‡The work was done when Hao and Xing were interns at MegVii, Inc. (Face++)

learning stage, we have two branches for learning a global feature and local features jointly. In the local branch, we align local parts by introducing a shortest path loss. In the inference stage, we discard the local branch and only extract the global feature. We find that only applying the global feature is almost as good as combining global and local features. In other words, the global feature itself, with the aid of local features learning, can greatly address the drawbacks we mentioned above, in our new joint learning framework. In addition, the form of global feature keeps our approach attractive for the deployment of a large ReID system, without costly local features matching.

We also adopt a mutual learning approach [49] in the metric learning setting, to allow two models to learn better representations from each other. Combining AlignedReID and mutual learning, our system outperforms state-of-the-art systems on Market1501, CUHK03, MARS, and CUHK-SYSU by a large margin. To understand how well human perform in the ReID task, we measure the best human performance of ten professional annotators on Market1501 and CUHK03. We find that our system with re-ranking [57] has a higher level of accuracy than the human. To the best of our knowledge, this is the first report in which machine performance exceeds human performance on the ReID task.

2. Related Work

Metric Learning. Deep metric learning methods transform raw images into embedding features, then compute the feature distances as their similarities. Usually, two images of the same person are defined as a positive pair, whereas two images of different persons are a negative pair. Triplet loss [18] is motivated by the margin enforced between positive and negative pairs. Selecting suitable samples for the training model through hard mining has been shown to be effective [13, 3, 39]. Combining softmax loss with metric learning loss to speed up the convergence is also a popular method [10].

Feature Alignments. Many works learn a global feature to represent an image of a person, ignoring the spatial local information of images. Some works consider local information by dividing images into several parts without an alignment [33, 38, 43], but these methods suffer from inaccurate detection boxes, occlusion and pose misalignment.

Recently, aligning local features by pose estimation has become a popular approach. For instance, pose invariant embedding (PIE) aligns pedestrians to a standard pose to reduce the impact of pose [52] variation. A Global-Local-Alignment Descriptor (GLAD) [37] does not directly align pedestrians, but rather detects key pose points and extracts local features from corresponding regions. SpindleNet [50] uses a region proposed network (RPN) to generate several body regions, gradually combining the response maps from adjacent body regions at different stages. These methods require extra pose annotation and have to deal with the errors introduced by pose estimation.

Mutual Learning. [49] presents a deep mutual learning strategy where an ensemble of students learn collaboratively and teach each other throughout the training process. DarkRank [4] introduces a new type of knowledge-cross sample similarity for model compression and acceleration, achieving state-of-the-art performance. These methods use mutual learning in classification. In this work, we study mutual learning in the metric learning setting.

Re-Ranking. After obtaining the image features, most current works choose the L2 Euclidean distance to compute a similarity score for a ranking or retrieval task. [35, 57, 1] perform an additional re-ranking to improve ReID accuracy. In particular, [57] proposes a re-ranking method with k-reciprocal encoding, which combines the original distance and Jaccard distance.

3. Our Approach

In this section, we present our AlignedReID framework, as shown in Figure 1.

3.1. AlignedReID

In AlignedReID, we generate a single global feature as the final output of the input image, and use the L2 distance as the similarity metric. However, the global feature is learned *jointly* with local features in the learning stage.

For each image, we use a CNN, such as Resnet50 [12], to extract a feature map, which is the output of the last convolution layer ($C \times H \times W$, where C is the channel number and $H \times W$ is the spatial size, e.g., $2048 \times 7 \times 7$ in Figure 1). A global feature (a C-d vector) is extracted by directly applying global pooling on the feature map. For the local features, a horizontal pooling, which is a global pooling in the horizontal direction, is first applied to extract a local feature for each row, and a 1×1 convolution is then applied to reduce the channel number from C to c. In this way, each local feature (a c-d vector) represents a horizontal part of the image for a person. As a result, a person image is represented by a global feature and H local features.

The distance of two person images is the summation of their global and local distances. The global distance is simply the L2 distance of the global features. For the local distance, we dynamically match the local parts from top to bottom to find the alignment of local features with the minimum total distance. This is based on a simple assumption that, for two images of the same person, the local feature from one body part of the first image is more similar to the semantically corresponding body part of the other image.

Given the local features of two images, $F = \{f_1, \cdots, f_H\}$ and $G = \{g_1, \cdots, g_H\}$, we first normalize

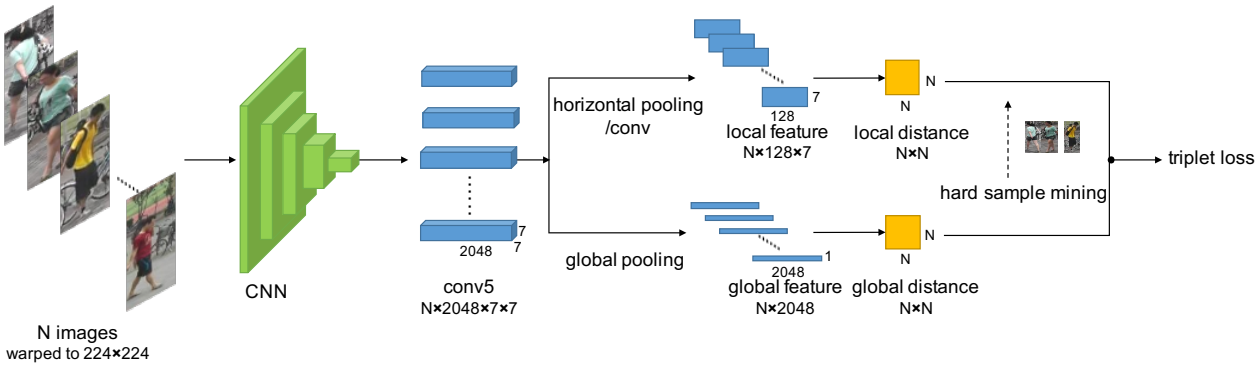

Figure 2. The framework of AlignedReID. Both the global branch and the local branch share the same convolution network to extract the feature map. The global feature is extracted by applying global pooling directly on the feature map. For the local branch, one 1×1 convolution layer is applied after horizontal pooling, which is a global pooling with a horizontal orientation. Triplet hard loss is applied, which selects triplet samples by hard sample mining according to global distances.

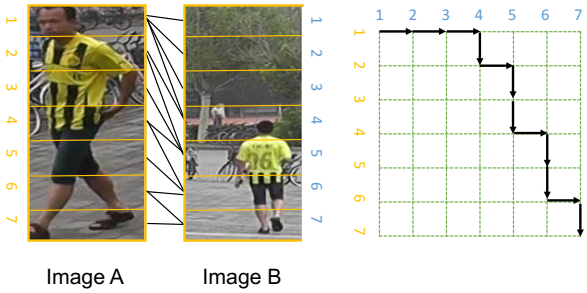

Figure 3. Example of AlignedReID local distance computed by finding the shortest path. The black arrows show the shortest path in the corresponding distance matrix on the right. The black lines show the corresponding alignment between the two images on the left.

the distance to $[0, 1)$ by an element-wise transformation:

$$d_{i,j} = \frac{e^{||f_i - g_j||_2} - 1}{e^{||f_i - g_j||_2} + 1} \quad i, j \in 1, 2, 3..., H, \tag{1}$$

where $d_{i,j}$ is the distance between the i-th vertical part of the first image and the j-th vertical part of the second image. A distance matrix D is formed based on these distances, where its (i, j)-element is $d_{i,j}$. We define the local distance between the two images as the total distance of the shortest path from $(1, 1)$ to (H, H) in the matrix D. The distance can be calculated through dynamic programming as follows:

$$S_{i,j} = \begin{cases} d_{i,j} & i = 1, j = 1 \\ S_{i-1,j} + d_{i,j} & i \neq 1, j = 1 \\ S_{i,j-1} + d_{i,j} & i = 1, j \neq 1 \\ min(S_{i-1,j}, S_{i,j-1}) + d_{i,j} & i \neq 1, j \neq 1, \end{cases} \tag{2}$$

where $S_{i,j}$ is the total distance of the shortest path when walking from $(1, 1)$ to (i, j) in the distance matrix D, and $S_{H,H}$ is the total distance of the final shortest path (i.e., the local distance) between the two images.

As shown in Fig. 3, images A and B are samples of the same person. The alignment between the corresponding body parts, such as part 1 in image A, and part 4 in image B, are included in the shortest path. Meanwhile, there are alignments between non-corresponding parts, such as part 1 in image A, and part 1 in image B, still included in the shortest path. These non-corresponding alignments are necessary to maintain the order of vertical alignment, as well as make the corresponding alignments possible. The non-corresponding alignment has a large L2 distance, and its gradient is close to zero in Eq.1. Hence, the contribution of such alignments in the shortest path is small. The total distance of the shortest path, i.e., the local distance between two images, is mostly determined by the corresponding alignments.

The global and local distance together define the similarity between two images in the learning stage, and we chose TriHard loss proposed by [13] as the metric learning loss. For each sample, according to the global distances, the most dissimilar one with the same identity and the most similar one with a different identity is chosen, to obtain a triplet. For the triplet, the loss is computed based on both the global distance and the local distance with different margins. The reason for using the global distance to mine hard samples is due to two considerations. First, the calculation of the global distance is much faster than that of the local distance. Second, we observe that there is no significant difference in mining hard samples using both distances.

Note that in the inference stage, we only use the global features to compute the similarity of two person images. We make this choice mainly because we unexpectedly observed that the global feature itself is also almost as good as the combined features. This somehow counter-intuitive phenomenon might be caused by two factors: 1) the feature

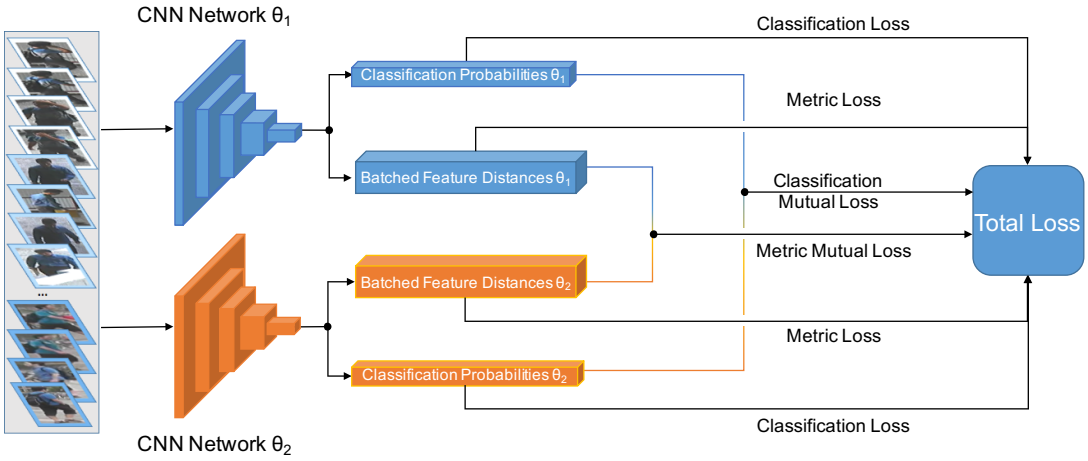

Figure 4. Framework of the mutual learning approach. Two networks with parameters θ_1 and θ_2 are trained together. Each network has two branches: a classification branch and a metric learning branch. The classification branches are trained with classification losses, and learn each other through classification mutual loss. The metric learning branches are trained with metric losses, which include both global distance and local distance. Meanwhile, the metric learning branches learn each other by metric mutual loss.

map jointly learned is better than learning the global feature only, because we have exploited the structure prior of the person image in the learning stage; 2) with the aid of local feature matching, the global feature can pay more attention to the body of the person, rather than over fitting the background.

3.2. Mutual Learning for Metric Learning

We apply mutual learning to train models for AlignedReID, which can further improve performance. A distillation-based model usually transfers knowledge from a pre-trained large teacher network to a smaller student network, such as [4]. In this paper, we train a set of student models simultaneously, transferring knowledge between each other, such as [49]. Differing from [49], which only adopts the Kullback-Leibler (KL) distance between classification probabilities, we propose a new mutual learning loss for metric learning.

The framework of our mutual learning approach is shown in Fig. 4. The overall loss function includes the metric loss, the metric mutual loss, the classification loss and the classification mutual loss. The metric loss is decided by both the global distances and the local distances, while the metric mutual loss is decided only by the global distances. The classification mutual loss is the KL divergence for classification as in [49].

Given a batch of N images, each network extracts their global features and calculates the global distance between each other as an $N \times N$ batch distance matrix, where $M_{ij}^{\theta_1}$ and $M_{ij}^{\theta_2}$ denote the (i,j)-th element in the matrices separately. The mutual learning loss is defined as

$$L_M = \frac{1}{N^2}\sum_{i}^{N}\sum_{j}^{N}\Big([ZG(M_{ij}^{\theta_1}) - M_{ij}^{\theta_2}]^2 + [M_{ij}^{\theta_1} - ZG(M_{ij}^{\theta_2})]^2\Big), \tag{3}$$

where $ZG(\cdot)$ represents the zero gradient function, which treats the variable as constant when calculating gradients, stopping the backpropagation in the learning stage.

By applying the zero gradient function, the second-order gradients is

$$\frac{\partial^2 L_M}{\partial M_{ij}^{\theta_1}\partial M_{ij}^{\theta_2}} = 0. \tag{4}$$

We found that it speeds up the convergence and improves the accuracy compared to a mutual loss without the zero gradient function.

4. Experiments

In this section, we present our results on four most widely used ReID datasets: Market1501 [53], CUHK03 [14], MARS [30], and CUHK-SYSU [41].

4.1. Datasets

Market1501 contains 32,668 images of 1,501 labeled persons of six camera views. There are 751 identities in the training set and 750 identities in the testing set. In the original study on this proposed dataset, the author also uses mAP as the evaluation criteria to test the algorithms.

CUHK03 contains 13,164 images of 1,360 identities. It provides bounding boxes detected from deformable part models (DPMs) and manual labeling.

MARS (Motion Analysis and Re-identification Set) dataset is an extended version of the Market1501 dataset. Because all bounding boxes and tracklets are generated automatically, it contains distractors, and each identity may have more than one tracklet. In total, MARS has 20,478 tracklets of 1,261 identities of six camera views.

CUHK-SYSU is a large-scale benchmark for a person search, containing 18,184 images (99,809 bounding boxes) and 8,432 identities. The training set contains 11,206 images of 5,532 query persons, whereas the test set contains 6,978 images of 2,900 persons.

Note that we only train a single model using training samples from all four datasets, as in [40, 50]. We follow the official training and evaluation protocols on Market1501, MARS, and CUHK-SYSU, and mainly report the mAP and rank-1 accuracy. For CUHK03, because we train one single model for all benchmarks, it is slightly different from the standard procedure in [14], which splits the dataset randomly 20 times, and the gallery for testing has 100 identities each time. We only randomly split the dataset once for training and testing, and the gallery includes 200 identities. It means our task might be more difficult than the standard procedure. Similarly, we evaluate our method with rank-1, -5, and -10 accuracy on CUHK03.

4.2. Implementation Details

We use Resnet50 and Resnet50-Xception (Resnet-X) pre-trained on ImageNet [28] as the base models. Resnet50-Xception replaces the 3×3 filter kernel through the Xception cell [7], which contains one 3×3 channel-wise convolution layer and one 1×1 spatial convolution layer. Each image is resized into 224×224 pixels. The data augmentation includes random horizontal flipping and cropping. The margins of TriHard loss for both the global and local distances is set to 0.3, and the mini-batch size is set to 128, in which each identity has 4 images. Each epoch includes 2000 mini-batches. We use an Adam optimizer with an initial learning rate of 10^{-3}, and shrink this learning rate by a factor of 0.1 at 80 and 160 epochs until achieving convergence.

For mutual learning, the weight of classification mutual loss (KL) is set to 0.01, and the weight of metric mutual loss is set to 0.001. The optimizer uses Adam with an initial learning rate of 3×10^{-4}, which is reduced to 10^{-4} and 10^{-5} at 60 epochs and 120 epochs until convergence is achieved.

Re-ranking is an effective technique for boosting the performance of ReID [57]. We follow the method and details in [57]. In all of our experiments, we combined metric learning loss with classification (identification) loss.

Figure 5. The black lines show the alignments of local parts between two persons: the thicker the line is, the greater it contributes to the shortest path. Persons have the same identities in (a-c), while persons have different identities in (d).

4.3. Advantage of AlignedReID

In this section, we analyze the advantage of our AlignedReID model.

We first show some typical results of the alignment in Fig 5. In FIg 5(a), the detection box of the right person is inaccurate, which results in a serious misalignment of heads. AlignedReID matches the first part of the left image with the first three parts of the right image in the shortest path. Fig 5(b) presents another difficult situation where human body structure is defective. The left image does not contain the parts below the knee. In the alignment, the skirt side of the right image are associated with the skirt parts of the left one, while the leg parts of the right image provide small contribution to the shortest path. Fig 5(c) shows an example of occlusion, where the lower part of the persons are invisible. The alignment shows that the occlude parts contribute small in the shortest path, hence the other parts are paid more attention in the learning stage. Fig 5(d) show two different persons with similar appearances. The shirt logo of the right person has no similar part in the left person, which results in a large shortest path distance (local distance) between these two images.

We then compare our *AlignedReID* with a *Baseline* without local feature branch. Two results are obtained by using the same network and the same training setting. The results are shown in Table 1. They show that AlignedReID boots $3.5\% \sim 6.0\%$ rank-1 accuracy and $5.0\% \sim 8.4\%$ mAP on

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

Table 1. Experiment results of AlignedReID. We combine metric learning loss with classification loss in our experiments.

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

Table 2. Results of mutual learning. MC stands for experiments with classification mutual loss. MM stands for experiments with both classification mutual loss and metric mutual loss.

all datasets. The local feature branch helps the network focus on useful image regions and discriminates similar person images with subtle differences.

We find that if we apply the local distance together with the global distance in the inference stage, rank-1 accuracy further improves approximately $0.3\% \sim 0.5\%$. However, it is time consuming and not practical when searching in a large gallery. Hence, we recommend using the global feature only.

4.4. Analysis of Mutual Learning

In the mutual learning experiment, we simultaneously train two AlignedReID models. One model is based on Resnet50, and the other is based on Resnet50-Xception. We compare their performances for three cases: with both metric mutual loss and classification mutual loss, with only classification mutual loss, and with no mutual loss. We also conduct a similar mutual learning experiment as a baseline, where the global features are trained without local features. The results are shown in Table 2.

Both experiments show that the metric mutual learning method can further improve performance. With the baseline mutual learning experiment, the classification mutual loss significantly improves performance on all datasets. However, with the AlignedReID mutual learning experiment, because the models without mutual learning perform well enough, the classification mutual loss cannot further improve performance. Particularly for MARS, it may even lead to a reduction of approximately $2.0\% \sim 3.5\%$ in rank-1 accuracy and mAP. However, metric mutual loss consistently helps the model achieve better performance for both the baseline and AlignedReID.

4.5. Comparison with Other Methods

In this subsection, we compare the results of AlignedReID with state-of-the-art methods, in Table 3 $\sim$ 6. In the tables, AlignedReID represents our method with mutual learning, and AlignedReID (RK) is our method with both mutual learning and re-ranking [57] with k-reciprocal encoding.

On Market1501, GLAD [37] achieves an 89.9% rank-1 accuracy and [13] obtains 81.1% for mAP owing to the use of re-ranking. Our AlignedReID achieves a 92.6% rank-1 accuracy and a 82.3% mAP, exceeding both of them. With the help of re-ranking, rank-1 accuracy and mAP are further improved to 94.0% and 91.2% in our AlignedReID (RK), outperforming the best of previous works by 4.1% and 10.1% separately.

On CUHK03, without re-ranking, HydraPlus-Net [20] achieves 91.8% rank-1 accuracy and our AlignedReID yields 91.9%. Note that our test gallery size is two times large as that used in [20]. Furthermore, our AlignedReID (RK) obtains a 96.1% rank-1 accuracy, exceeding state-of-the-art by 4.3%.

We also show our results for MARS, which is based on tracklets. However, we ignore the sequence information provided in MARS. For each tracklet, its feature is calculated by simply averaging features of all its bounding boxes. In this way, AlignedReID with/without re-ranking obtains 87.5% and 86.8% rank-1 accuracy, which is better than all other state-of-the-art methods by a large margin.

There have not been many studies reported on CUHK-SYSU. With this dataset, AlignedReID achieves 93.7% mAP and 95.3% rank-1 accuracy, which is much higher than any published results.

Table 3. Comparison on **Market1501** in single query mode

Methods	mAP	r=1
Temporal [23]	22.3	47.9
Learning [47]	35.7	61.0
Gated [32]	39.6	65.9
Person [5]	45.5	71.8
Re-ranking [57]	63.6	77.1
Pose [52]	56.0	79.3
Scalable [1]	68.8	82.2
Improving [16]	64.7	84.3
In [13]	69.1	84.9
In (RK)[13]	**81.1**	86.7
Spindle[50]	-	76.9
Deep[49]*	68.8	87.7
DarkRank[4]*	74.3	89.8
GLAD[37]*	73.9	**89.9**
HydraPlus-Net[20]*	-	76.9
AlignedReID	82.3	92.6
AlignedReID (RK)	**91.2**	**94.0**

Table 4. Comparison on **CUHK03** labeled dataset

Methods	r=1	r=5	r=10
Person [15]	44.6	-	-
Learning [47]	62.6	90.0	94.8
Gated [32]	61.8	-	-
A [34]	57.3	80.1	88.3
Re-ranking [57]	64.0	-	-
In [13]	75.5	95.2	99.2
Joint [42]	77.5	-	-
Deep [10]*	84.1	-	-
Looking [2]*	72.4	95.2	95.8
Unlabeled [56]	84.6	97.6	98.9
A [55]*	83.4	97.1	98.7
Spindle[50]	88.5	97.8	98.6
DarkRank[4]*	89.7	**98.4**	**99.2**
GLAD[37]*	85.0	97.9	99.1
HydraPlus-Net[20]*	**91.8**	**98.4**	99.1
AlignedReID	91.9	98.7	99.4
AlignedReID (RK)	**96.1**	**99.5**	**99.6**

Table 5. Comparison on **MARS** in single query mode

Methods	mAP	r=1
Re-ranking [57]	68.5	73.9
Learning [48]*	-	55.5
Multi [31]*	-	68.2
MARS [30]	49.3	68.3
In [13]	67.7	79.8
In (RK)[13]	**77.4**	**81.2**
Quality [21]*	51.7	73.7
See [58]	50.7	70.6
AlignedReID	79.1	86.8
AlignedReID (RK)	**85.6**	**87.5**

Table 6. Comparison with existing methods on **CUHK-SYSU**

Methods	mAP	r=1
End[41]	55.7	62.7
Deep [29]*	74.0	76.7
Neural [17]	**77.9**	**81.2**
AlignedReID	**93.7**	**95.3**

5. Human Performance in Person ReID

Given the significant improvement of our approach, we are curious to find the quality of human performance. Thus, we conduct human performance evaluations on Market1501 and CUHK03.

To make the study feasible, for each query image, the annotator does not have to find the same person from the entire gallery set. We ask him or her to pick the answer from a much smaller set of selected images.

In CUHK03, for each query image, there is only one image for the identical person in the gallery set. The annotator looks for the identical person among 10 images selected: our ReID model first generates the top10 results in the gallery set for the query image; if the "ground truth" is not among the top10 results, we replace the 10th result with the ground truth.

For Market1501, there may be more than one ground truth in the gallery set. The annotator needs to pick one from 50 images selected as follows: our ReID model generated the top50 results in the gallery set for the query image; if any ground truth is not among them, it would be used to replace one non-ground truth result with the lowest rank. In this way, we make sure that all ground truths are in the 50 selected images.

The interface of the human performance evaluation system is presented in Fig 6. The images are randomly shuffled before being displayed to the annotator. The evaluation

Figure 6. Interface of our human performance evaluation system for CUHK03. The left side shows a query image and the right side shows 10 images sampled using our deep model.

website is available now [†].

Ten professional annotators participate in the evaluation. Because only one candidate is chosen, we are unable to obtain the mAP of human beings as a standard evaluation. The rank-1 accuracies are computed for each annotator on all datasets. The best accuracy is then used as the human performance, which is shown in Table 7.

On Market1501, human beings achieve a 93.5% rank-1 accuracy, which is better than all state-of-the-art methods. The rank-1 accuracy in our AlignedReID (RK) reaches 94.0% rank-1, exceeding the human performance. On CUHK03, the human performance reaches a 95.7% rank-1 accuracy, which is much higher than any known state-of-the-art methods. Our AlignedReID (RK) obtains a 96.1% rank-1 accuracy, surpassing the human performance.

Figure 7 shows some examples, where an annotator selected a wrong answer, while the top1 result provided by our method is correct. There are several reasons why the annotator makes mistakes. First, the annotator usually summarizes some attributes, such as gender, age, and etc., to decide whether the images contain the same person. However, the summarized attributes might be incorrect. For example, the person in the query image of (a) seems to be a man, but actually a woman. In (b), the bag appeared in the ground truth image is occluded in the query image. Second, color bias exists between cameras, and it could make the same person looks differently in the query and ground truth images such as in (c). Last, different camera angles and human poses might mislead the judgement of body shapes as shown in (d-e).

6. Conclusion

In this paper, we have demonstrated that an implicit alignment of local features can substantially improve global feature learning. This surprising result gives us an important insight: the end-to-end learning with structure prior is more powerful than a "blind" end-to-end learning.

Although we show that our methods outperform humans in the Market1501 and CUHK03 datasets, it is still early to claim that machines beat humans in general. Figure 8 presents a few "big" mistakes which seldom confuses hu-

[†]Market1501:http://reid-challenge.megvii.com
CUHK03:http://reid-challenge.megvii.com/cuhk03

Table 7. Results of human performance evaluation. We show the accuracies of the five annotators who did best in the evaluation. We also show our AlignedReID results with re-ranking.

	Market1501	CUHK03
Annotator Rank 1	**93.5**	**95.7**
Annotator Rank 2	91.1	91.9
Annotator Rank 3	90.6	91.2
Annotator Rank 4	90.0	91.1
Annotator Rank 5	88.3	90.0
AlignedReID (RK)	**94.0**	**96.1**

Figure 7. Top: query image. Middle: the result picked by an annotator. Bottom: top1 result by our method.

mans. This indicates that the machine still has a lot of room for improvement.

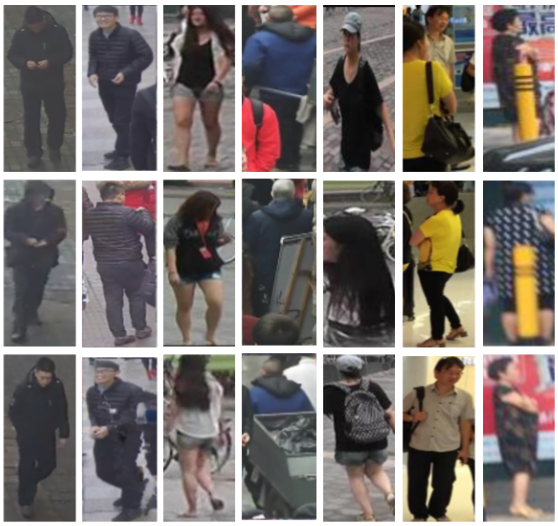

Figure 8. Top: query image. Middle: top1 result by our method. Bottom: ground truth.

Acknowledgement

The authors gratefully appreciate Xiangyu Zhang for the use of the pre-trained Resnet50 and Resnet50-Xception on ImageNet dataset.

We also appreciate Jianan Wu for developing the human performance evaluation system, and Sipeng Zhang etc. for participating in the human performance evaluation.

References

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. *arXiv preprint arXiv:1703.08359*, 2017.

[2] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. *arXiv preprint arXiv:1701.03153*, 2017.

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. *arXiv preprint arXiv:1704.01719*, 2017.

[4] Y. Chen, N. Wang, and Z. Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *arXiv preprint arXiv:1707.01220*, 2017.

[5] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016.

[7] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*, 2016.

[8] H. Fan, L. Zheng, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. *arXiv preprint arXiv:1705.10444*, 2017.

[9] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE, 2010.

[10] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv preprint arXiv:1611.05244*, 2016.

[11] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *Distributed Smart Cameras, 2008. ICDSC 2008. Second ACM/IEEE International Conference on*, pages 1–6. IEEE, 2008.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[14] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. pages 152–159, 2014.

[15] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[16] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv preprint arXiv:1703.07220*, 2017.

[17] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan. Neural person search machines. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[18] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017.

[19] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *arXiv preprint arXiv:1701.00193*, 2017.

[20] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. 2017.

[21] Y. Liu, J. Yan, and W. Ouyang. Quality aware network for set to set recognition. *arXiv preprint arXiv:1704.03373*, 2017.

[22] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong. Person re-identification by unsupervised video matching. *Pattern Recognition*, 65:197–210, 2017.

[23] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-

identification. In *European Conference on Computer Vision*, pages 858–877. Springer, 2016.

[24] T. Matsukawa and E. Suzuki. Person re-identification using cnn features learned from combination of attributes. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 2428–2433. IEEE, 2016.

[25] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2016.

[26] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1315, 2016.

[27] F. Radenović, G. Tolias, and O. Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*, pages 3–20. Springer, 2016.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[29] A. Schumann, S. Gong, and T. Schuchert. Deep learning prototype domains for person re-identification. *arXiv preprint arXiv:1610.05047*, 2016.

[30] Springer. *MARS: A Video Benchmark for Large-Scale Person Re-identification*, 2016.

[31] Y. T. Tesfaye, E. Zemene, A. Prati, M. Pelillo, and M. Shah. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv preprint arXiv:1706.06196*, 2017.

[32] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016.

[33] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016.

[34] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153, 2016.

[35] J. Wang, S. Zhou, J. Wang, and Q. Hou. Deep ranking model by large adaptive margin learning for person re-identification. *arXiv preprint arXiv:1707.00409*, 2017.

[36] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2501–2514, 2016.

[37] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian. Glad: Global-local-alignment descriptor for pedestrian retrieval. *arXiv preprint arXiv:1709.04329*, 2017.

[38] Q. Xiao, K. Cao, H. Chen, F. Peng, and C. Zhang. Cross domain knowledge transfer for person re-identification. *arXiv preprint arXiv:1611.06026*, 2016.

[39] Q. Xiao, H. Luo, and C. Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv preprint arXiv:1710.00478*, 2017.

[40] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016.

[41] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016.

[42] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proc. CVPR*, 2017.

[43] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv preprint arXiv:1707.00798*, 2017.

[44] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1353, 2016.

[45] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai. Image-to-video person re-identification with temporally memorized similarity learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[46] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[47] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1239–1248, 2016.

[48] W. Zhang, S. Hu, and K. Liu. Learning compact appearance representation for video-based person re-identification. *arXiv preprint arXiv:1702.06294*, 2017.

[49] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. *arXiv preprint arXiv:1706.00384*, 2017.

[50] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. CVPR, 2017.

[51] R. Zhao, W. Oyang, and X. Wang. Person re-identification by saliency learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):356–370, 2017.

[52] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017.

[53] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference*, 2015.

[54] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[55] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *arXiv preprint arXiv:1611.05666*, 2016.

[56] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[57] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *arXiv preprint arXiv:1701.08398*, 2017.

[58] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification.

毕业设计文献综述和开题报告的考核

一、对文献综述、外文翻译和开题报告评语及成绩评定：

评语：

成绩：

开题报告答辩小组负责人 (签名) _____