Pose Invariant Embedding for Deep Person Re-identification

author:Liang Zheng, Yujia Huang, Huchuan Lu, Yi Yang

摘要

行人未对准主要是由探测器错误和姿势变化引起的，对于健壮的人员重新识别（re-ID）系统而言，这是一个关键问题。为了解决这个问题，本文提出一种姿势不变嵌入方法（pose invariant embedding,PIE）方法作为行人描述符。首先为了对齐行人，提出一种PoseBox结构，然后为了减少PoseBox造成的姿态估计误差和信息损失，文章设计了一种PoseBox fusion（PBF）的CNN架构将原始图片、PoseBox、体态估计作为输入。最终的PIE描述符定义为PBF的全连接层。

简介

有很多影响ReID准确率的因素，例如检测/跟踪错误，光照变化，姿势，视点等。一个很大的影响因素就是行人未对准，主要可以归结为两点：（1）行人本来就有许多姿态，这到此bounding box不准确。如Fig1（2）检测错误

Figure 1. Examples of misalignment correction by PoseBox. Row 1: original bounding boxes with detection errors/occlusions. Every consecutive two boxes correspond to a same person. Row 2: corresponding PoseBoxes. We observe that misalignment can be corrected to some extent.

当行人对齐没做好的时候，ReID准确率会降低，例如在ReID中最常见的是水平切条处理，该方法在轻微的垂直未对准的情况下工作。在Fig1的第二行，人的头可能被匹配到未对准图片的背景上去，所以水平切条可能效果还更差。

作者之前的工作中[8,7]在同组考虑了未对准的问题，均使用了图片结构(pictorial structure,PS)，其中使用了和PoseBox相似的结构，还有和结合归一化body parts的想法类似，作者的工作落脚点为使用基于CNN的姿态估计和局部身体链接，并且PoseBox的组件与PS有所不同，大规模评估证明了这一点。另一个工作是匹配程序，尽管[8，7]并未讨论现实世界数据集中普遍存在的姿势估计误差，但我们表明，这些误差使刚性特征学习/匹配仅与PoseBox产生的结果相比于原始图像差，而三个流PoseBox融合网络可有效缓解此问题。

针对以上问题和有限的方法，本文提出一种姿势不变嵌入方法（PIE）作为鲁棒的描述符。一共有两步：（1）为每个行人外包矩形构造PoseBox，PoseBox描绘了具有标准直立姿势的行人。在姿势估计器的帮助下精心设计[34]，PoseBox的目的是产生对齐良好的行人图像，以便所学习的功能可以在剧烈的姿势变化下找到同一个人。作者使用标准的CNN架构[37、41、44]进行了单独训练，PoseBox产生非常不错的re-ID准确性。（2）为了减缓信息丢失和姿态估计错误的影响(Fig2)，PoseBoxc fusion(PBF)CNN模型被构建出来，三个输入分别是PoseBox、原始图片、姿态估计置信度。PBF在原始图像和PoseBox之间实现了全局优化的权衡。PIE被定义为PBF的全连接激活。

Figure 2. Information loss and pose estimation errors that occur during PoseBox construction. Row 1: important pedestrian details (highlighted in red bounding boxes) may be missing in the PoseBox. Row 2: pose estimation errors deteriorate the quality of PoseBoxes. For each image pair, the original image and its PoseBox are on the left and right, respectively.

本文有三个贡献：
（1）提出了PoseBox，它与以前的工作具有相似的性质[8]。 它可以很好地匹配行人，并在单独使用时产生令人满意的re-ID性能（次要）

（2）姿势不变嵌入（PIE）被提议作为PoseBox Fusion（PBF）网络的一部分。 PBF将原始图像，PoseBox和姿势估计错误融合在一起，从而在姿势估计失败时提供一种回退机制（主要）

（3）使用PIE在Market-1501，CUHK03，VIPeR数据集上有很高的准确率。

相关工作

姿态估计

在DeepPose后，姿态估计从传统方法转向深度徐热西，当前的模型采取复杂尺度特征和学习机制来结合这些特好曾。通过调整一元分数和成对比较来注入人体关节之间的空间关系也是有效的。本文采取了卷积姿态机（convolution pose machines, CPM)来做姿态估计。

ReID的深度学习

由于基于深度学习的模型有很好的效果，近两年基本支配了ReID社区。在两个早期工作中[20,39]，使用将两个图像作为输入的暹罗模型(siamese model)。在以后的工作中，以各种方式改进了该模型，例如注入更复杂的空间约束[1，6]，使用LSTM [32]建模身体部位的顺序属性，以及为不同的图像对挖掘判别匹配部位[31]。 siamese model只使用了简单的id标签。之前很多工作接受了分类模型。[41]中把视频帧作为每个人的训练样本，[37]中，有效的神经元被发明来每个训练域和新的dropout策略被提及。[36]中，手工提取的低层特征和FC+softmax整合到一起。本文的网络和[36]的比较相似，姿态估计分数与两个FC层整合到一起。

ReID中的姿态

尽管在许多之前的工作中有提到姿态变化是ReID的影响因素，只有很少的文献中能找到他们之间联系的描述。Farenzena等[12]提出检测不同身体部位的对称轴并提取姿势变化后的特征。在[35]中，HOG检测器提供了上身定向的粗略估计，然后将上身渲染为关节3D模型的纹理。Bak等[3]进一步将每个人分为三种姿势类型：正面，背面和侧面。两项工作[3，9]均根据不同的测试姿势对应用视点特定的距离度量。最接近PoseBox的作品是[8，7]，它构造了绘画结构（PS），与PoseBox的概念相似。 他们使用传统的姿态估计器和手工制作的描述符，这些描述符在很大程度上不如CNN。 本文的工作采用了一整套更强大的技术，并设计了更有效的CNN结构，以大规模数据集上具有竞争力的re-ID准确性为证。

提出的方法

PoseBox构造

PoseBox的构造分两步：

姿态估计

本文接受了CPM，简而言之，CPM是顺序卷积体系结构，可强制执行中间监督以防止梯度消失。

身体部位发现和仿射投影

从检测点根据检测到的关节，可以描绘出10个身体部位（Fig3)

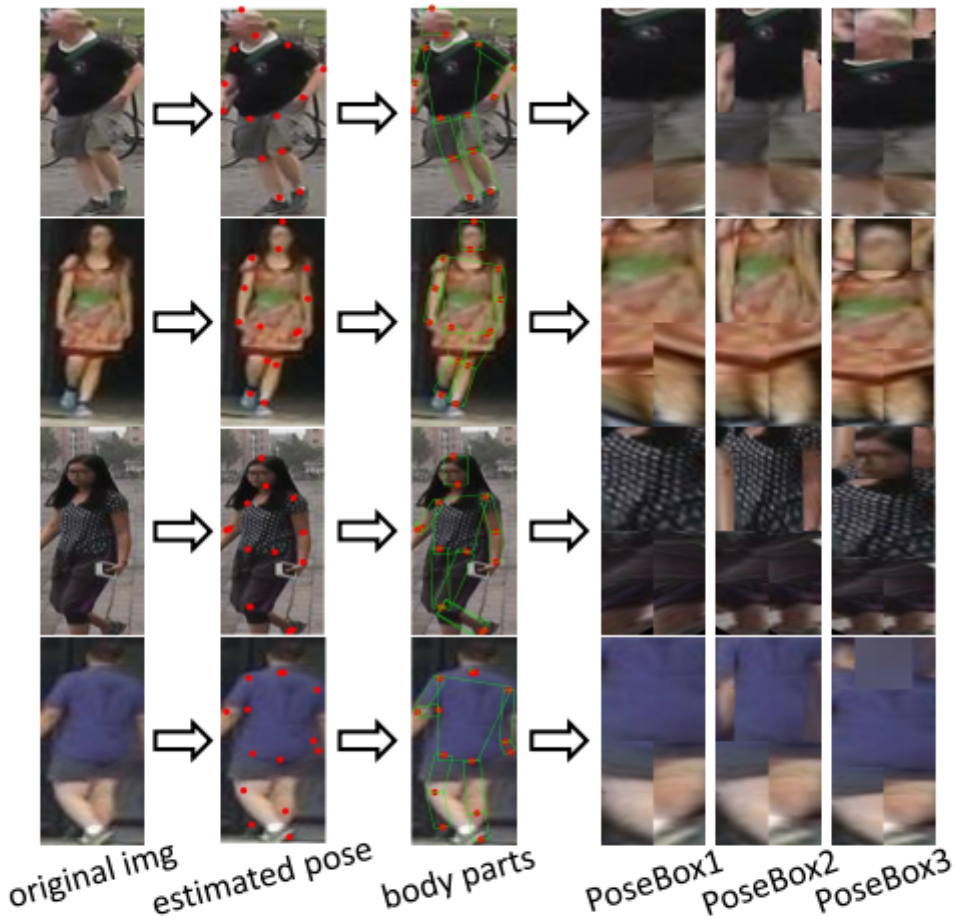

Figure 3. PoseBox construction. Given an input image, the pedestrian pose is estimated by CPM [34]. Ten body parts can then be discovered through the body joints. Three types of PoseBoxes are built from the body parts. PoseBox1: torso + legs; PoseBox2: PoseBox1 + arms; PoseBox3: PoseBox2 + head.

三种PoseBox

PoseBox1由躯干和两条腿组成。 腿由上腿和下腿组成。 PoseBox1包括两个最重要的身体部位，并且是其他两种PoseBox类型的基准

PoseBox2基于PoseBox1，本文进一步添加了左臂和右臂。 臂包括上臂和下臂子模块。 在本文的实验中，由于手臂带来的丰富信息，本文证明PoseBox2优于PoseBox1

PoseBox3在PoseBox 2的基础上，本文将头的box放在躯干的box顶部。在[8]中表明，包含头部带来了边际性能的提高。在本文的案例中，本文发现PoseBox3稍逊于PoseBox2，这可能是由于频繁的头颈部估计错误。

备注

PoseBox的优点是双重的。 首先，可以校正姿势变化。 其次，可以大大消除背景噪音。 PoseBox也有两个方面的限制。 首先，姿势估计错误经常发生，导致关节的检测不精确。 其次，PoseBox是手动设计的，因此就信息丢失或重新ID准确性而言，不能保证它是最佳的。

基准

本文在原始行人图片和PoseBox上采用了两个baseline，AlexNet和Residual-50。

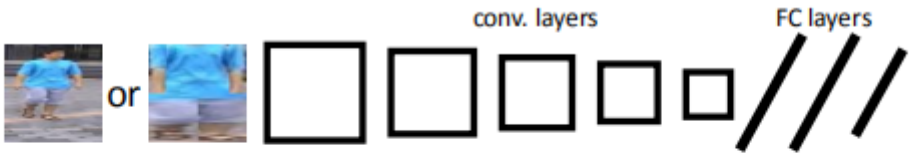

Figure 4. The baseline identification CNN model used in this paper. The AlexNet [19] or ResNet-50 [15] with softmax loss is used. The FC activations are extracted for Euclidean-distance testing.

Baseline1：原始图片（resize到224*224）作为CNN的输入

Baseline2：PoseBox（resize到224*224）作为CNN的输入，每次只用一种PoseBox

PBF网路

动机

针对第一个问题（姿态估计错误）：通过置信度得分来大致预测姿势估计的质量

针对第二个问题（图像部分缺失）：通过重新引入原始图像来挽救丢失的视觉提示，从而使区分性细节被深度网络捕获

	head	neck	Rshoulder	Relbow	Rwrist	Lshoulder	Lelbow	Lwrist	Rhip	Rknee	Rankle	Lhip	Lknee	Lankle
(a)	.05	.01	.06	.13	.07	.03	.04	.03	.16	.38	.51	.12	.55	.60
(b)	.03	.01	.05	.05	.07	.04	.06	.08	.04	.01	.02	.06	.02	.01
(c)	.75	.79	.64	.57	.45	.74	.43	.50	.49	.76	.52	.56	.72	.13
(d)	.15	.34	.25	.71	.62	.58	.72	.55	.09	.04	.01	.11	.01	.01

Figure 5. Examples of pose estimation errors and the confidence scores. **Upper:** four pedestrian bounding boxes named with (a), (b), (c), and (d), and their pose estimation results. **Lower:** pose estimation confidence scores of the four images. A confidence vector consists of 14 numbers corresponding to the 14 body joints. We highlight the correctly detected joints in green.

网络

三流PoseBox Fusion网络，两个图片输入（原始图片、PoseBox）喂给CNN，因为两种图片的不同，两个CNN流的参数不共享。FC6和FC7两层都和这些conv层相连。参个FC7层串接到一起与FC8相连。三个Softmax loss的和就是最终的loss。

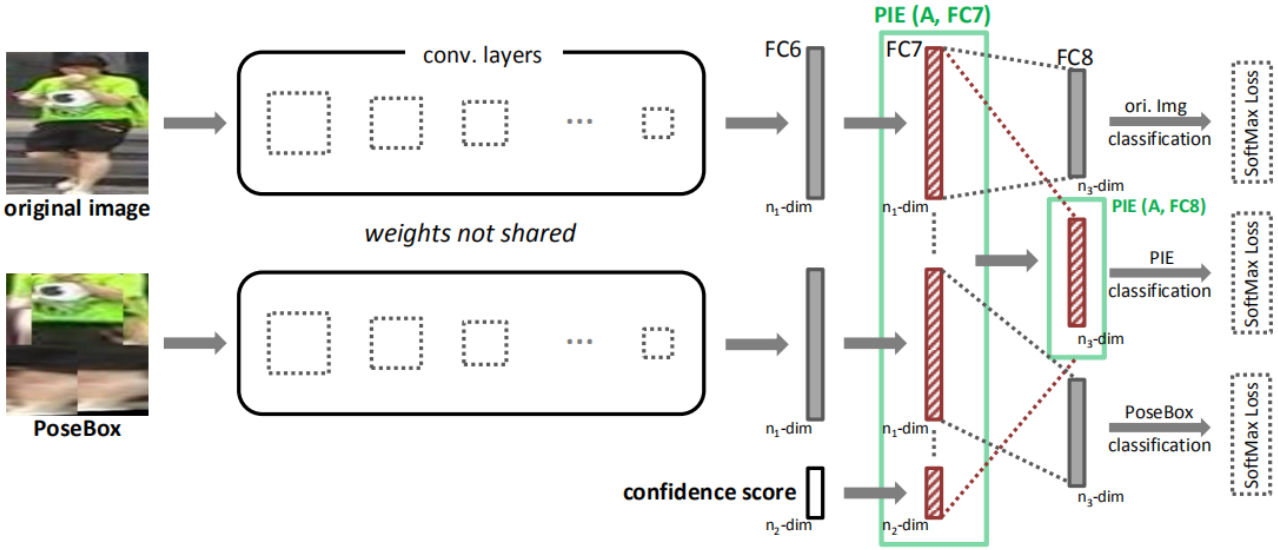

Figure 6. Illustration of the PoseBox Fusion (PBF) network using AlexNet. The network inputs, *i.e.*, the original image, its PoseBox, and the pose estimation confidence, are highlighted in **bold**. The former two undergo convolutional layers before the fully connections (FC). The confidence vector undergoes one FC layer, before the three FC7 layers are concatenated and fully connected to FC8. SoftMax loss is used. Two alternatives of the PIE descriptor are highlighted by green boxes. For AlexNet and Market-1501, PIE(A, FC7) is 8,206-dim, and PIE(A, FC8) is 751-dim; For ResNet-50, there would be no FC6, and PIE(R, Pool5) is 4,110-dim, and PIE(R, FC) is 751-dim.

训练的时候，针对三种输入都喂进PBF，三个loss的和一起反向传播给卷积层。

在测试阶段，PIE作为描述符，使用欧拉距离来判断

PBF有三个优点：（1）置信度向量是PoseBox是否可靠的指标。 这提高了PBF作为静态嵌入网络的学习能力，因此可以在PoseBox和原始图像之间找到全局权衡（2）原始图像不仅在姿态估计失败时启用回退机制，而且还重新训练了在PoseBox构造过程中可能丢失的行人细节，但对识别身份很有用（3）PoseBox为原始图像提供了重要的补充线索。 使用正确预测的关节，行人匹配可以通过对齐良好的图像更加准确。 因此可以减少检测误差和姿势变化的影响。

实验

数据集

文章使用VIPeR,CUHK03,Market-1501

实验步骤

实验直接使用了现成的卷积位姿机（CPM），该机是使用在MPII人类位姿数据集上训练的多级CNN模型训练的使用AlexNet时，$n_1 = 4.96, n_2 = 14, n_3 = 751$，使用ResNet-50时，PBF中取消FC6层，FC7层用Pool5代替，$n_1 = 2048, n_3 = 751$。、

评估

基准

Table 1. Comparison of the proposed method with various baselines. PoseBox2 is employed here. Baseline1: training using the original image. Baselin2: training use the PoseBox. PIE: proposed pose invariant embedding. A: AlexNet. R: ResNet-50.

Methods	dim	*Market-1501*					*CUHK03*				*Market-1501* → *VIPeR*			
		1	5	10	20	mAP	1	5	10	20	1	5	10	20
Baseline1 (A, FC7)	4,096	55.49	76.28	83.55	88.98	32.36	57.15	83.50	90.85	95.70	17.44	31.84	41.04	51.36
Baseline1 (A, FC8)	751	53.65	75.48	82.93	88.51	31.60	58.80	85.80	91.90	96.25	17.15	32.06	41.68	51.55
Baseline1 (R, Pool5)	2,048	73.02	87.44	91.24	94.70	47.62	51.60	79.60	87.70	95.00	23.42	42.31	51.96	63.80
Baseline1 (R, FC)	751	70.58	84.95	90.02	93.53	45.84	54.80	84.20	91.70	97.60	15.85	28.80	37.41	47.85
Baseline2 (A, FC7)	4,096	52.22	71.53	78.95	85.04	28.95	39.90	71.40	82.30	90.00	17.28	32.59	42.25	55.09
Baseline2 (A, FC8)	751	51.10	72.24	79.48	85.60	29.91	42.30	75.05	84.35	92.00	16.04	33.45	42.66	54.97
Baseline2 (R, Pool5)	2,048	64.49	79.48	85.07	88.95	38.16	36.90	68.40	78.70	86.70	21.11	37.18	45.89	54.34
Baseline2 (R, FC)	751	62.20	78.36	83.76	88.84	37.91	41.70	72.70	84.20	92.50	15.57	26.68	33.54	41.71
PIE (A, FC7)	8,206	64.61	82.07	87.83	91.75	38.95	59.80	85.35	91.85	95.85	21.77	38.04	46.61	56.61
PIE (A, FC8)	751	65.68	82.51	87.89	91.63	41.12	**62.40**	88.00	93.70	96.50	18.10	31.20	38.92	49.40
PIE (R, Pool5)	4,108	**78.65**	**90.26**	**93.59**	**95.69**	**53.87**	57.10	84.60	91.40	96.20	**27.44**	**43.01**	**50.82**	**60.22**
PIE (R, FC)	751	75.12	88.27	92.28	94.77	51.57	61.50	**89.30**	**94.50**	**97.60**	23.80	37.88	47.31	56.55

PIE的有效性

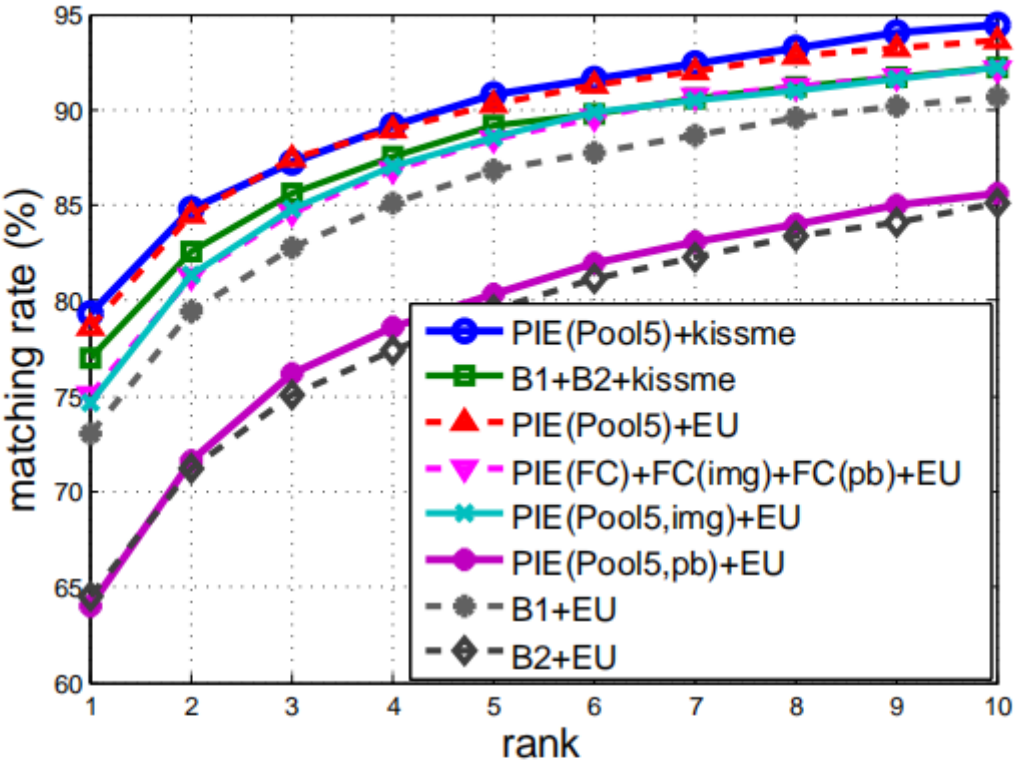

Figure 7. Comparison with various feature combinations on the Market-1501 dataset. ResNet-50 [15] is used. Kissme [18] is used for distance metric learning. "EU": Euclidean distance. "PIE(Pool5,img)" and "PIE(Pool5,pb)" denote the 2,048-dim sub-vectors of the full 4,108-dim PIE(Pool5) vector, corresponding to the image and PoseBox streams of PBF, respectively. "FC(img)" and "FC(pb)" are the 751-dim FC vectors of the image and Pose-Box streams of PBF, respectively. "B1" and "B2" represent base-line 1 and 2, respectively, using the 2,048-dim Pool5 features.

三种PoseBox的对比

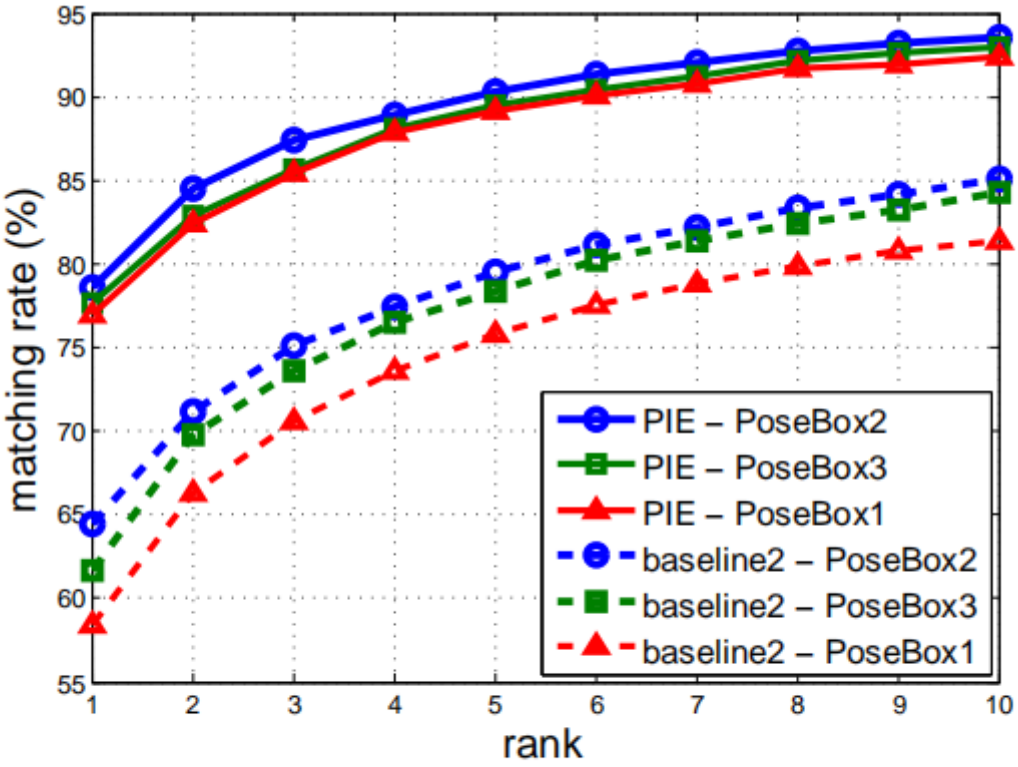

Figure 8. Re-ID accuracy of the three types of PoseBoxes. Results of both the baseline and PIE are presented on the Market-1501 dataset.

Albation实验

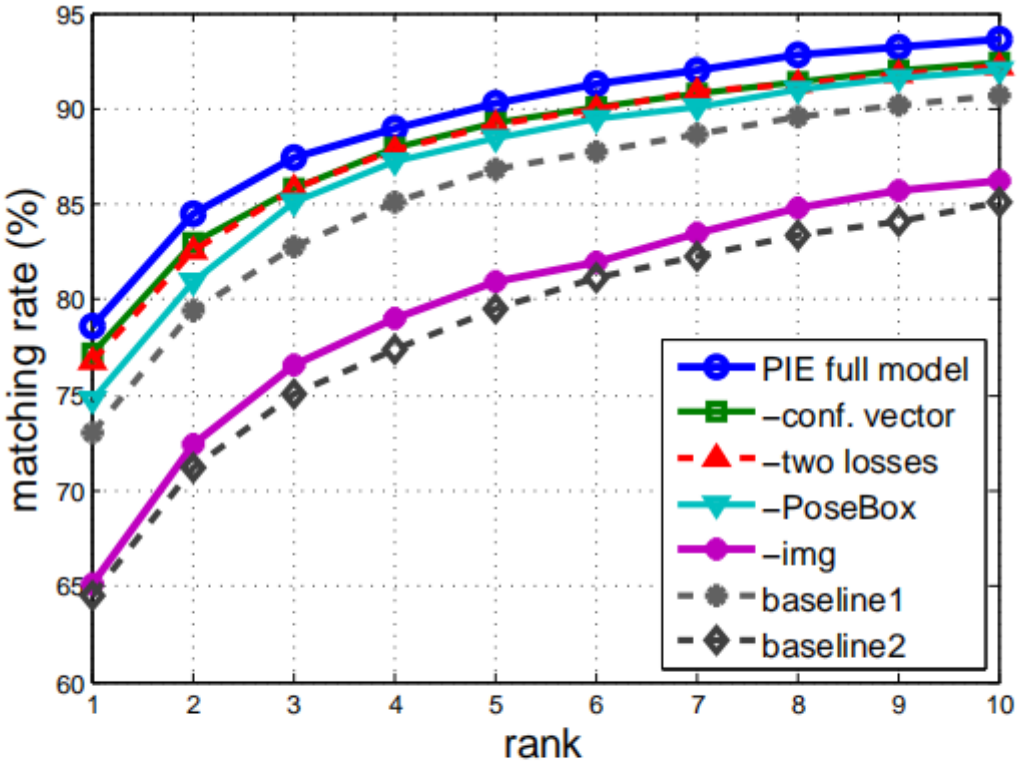

Figure 9. Ablation studies on Market-1501. From the "full" model, we remove one component at a time. The removed components include PoseBox, original image, the confidence vector, and the two losses of the PoseBox and the original image.

与SOTA的对比

Table 2. Comparison with state of the art on Market-1501.

Methods	rank-1	rank-5	rank-10	rank-20	mAP
BoW+Kissme [42]	44.42	63.90	72.18	78.95	20.76
WARCA [17]	45.16	68.12	76	84	-
Temp. Adapt. [24]	47.92	-	-	-	22.31
SCSP [4]	51.90	-	-	-	26.35
Null Space [40]	55.43	-	-	-	29.87
LSTM Siamese [32]	61.6	-	-	-	35.3
Gated Siamese [31]	65.88	-	-	-	39.55
PIE (Alex)	65.68	82.51	87.89	91.63	41.12
PIE (Res50)	*78.65*	*90.26*	*93.59*	*95.69*	*53.87*
+ Kissme	**79.33**	**90.76**	**94.41**	**96.52**	**55.95**

Table 3. Comparison with state of the art on CUHK03 (detected).

Methods	rank-1	rank-5	rank-10	rank-20	mAP
BoW+HS [42]	24.30	-	-	-	-
Improved CNN [1]	44.96	76.01	83.47	93.15	-
XQDA [21]	46.25	78.90	88.55	94.25	-
SI-CI [33]	52.2	74.3	92.3	-	-
Null Space [40]	54.70	84.75	94.80	95.20	-
LSTM Siamese [32]	57.3	80.1	88.3	-	46.3
MLAPG [22]	57.96	87.09	94.74	98.00	-
Gated Siamese [31]	61.8	80.9	88.3	-	51.25
PIE (Alex)	62.60	87.05	92.50	96.30	67.91
PIE (Res50)	*61.50*	*89.30*	*94.50*	*97.60*	*67.21*
+ Kissme	**67.10**	**92.20**	**96.60**	**98.10**	**71.32**

Table 4. Comparison with state of the art on VIPeR. The top 6 rows are unsupervised; the bottom 10 rows use supervision.

Methods	rank-1	rank-5	rank-10	rank-20
GOG [25]	21.14	40.34	53.29	**67.21**
Enhanced Deep [36]	15.47	34.53	43.99	55.41
SDALF [23]	19.87	38.89	49.37	65.73
gBiCov [23]	17.01	33.67	46.84	58.72
BOW [42]	21.74	-	-	60.85
PIE	**27.44**	**43.01**	**50.82**	60.22
XQDA [21]	40.00	67.40	80.51	91.08
MLAPG [22]	40.73	-	82.34	92.37
WARCA [17]	40.22	68.16	80.70	91.14
Null Space [40]	42.28	71.46	82.94	92.06
SI-CI [33]	35.8	67.4	83.5	-
SCSP [4]	53.54	82.59	91.49	96.65
Mirror [5]	42.97	75.82	87.28	94.84
Enhanced [36] + Mirror [5]	34.87	66.68	79.30	90.38
LSTM Siamese [32]	42.4	68.7	79.4	-
Gated Siamese [31]	37.8	66.9	77.4	-
PIE+Mirror [5]+MFA[38]	43.29	69.40	80.41	89.94
Fusion+MFA	**54.49**	**84.43**	**92.18**	**96.87**

Figure 10. Sample re-ID results on the Market-1501 dataset. For each query placed on the left, the three rows correspond to the rank lists of baseline 1, baseline 2, and PIE, respectively. Green bounding boxes denote correctly retrieved images.

总结

本文明确地解决了人员重新识别中的行人错位问题。提出姿势不变嵌入（PIE）作为行人描述符。 首先用卷积姿势机[34]检测到的16个关节构造PoseBox。 PoseBox可帮助纠正由摄像机视线，人的动作和检测器错误引起的姿势变化，并实现对齐的脚蹬Trian匹配。因此，通过PoseBox融合（PBF）网络学习了PIE，在该网络中，原始图像与PoseBox和姿势估计置信度融合在一起。PBF减少了在PoseBox构建过程中姿势估计错误和细节损失的影响。 证明PoseBox可以产生相当高的准确性，而与最新技术相比，PIE可以产生有竞争力的准确性。