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].

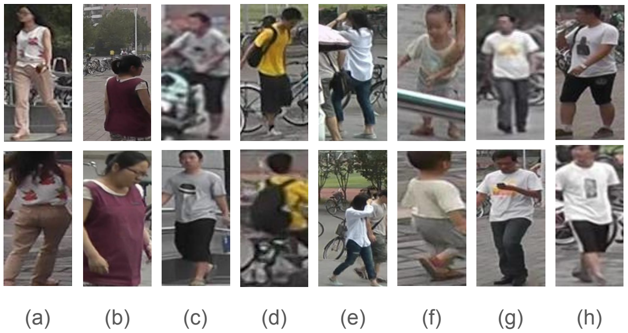

Figure 1. Challenges in ReID: (a-b) inaccurate detection, (c-d) pose misalignments, (e-f) occlusions, (g-h) very similar appearance.

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

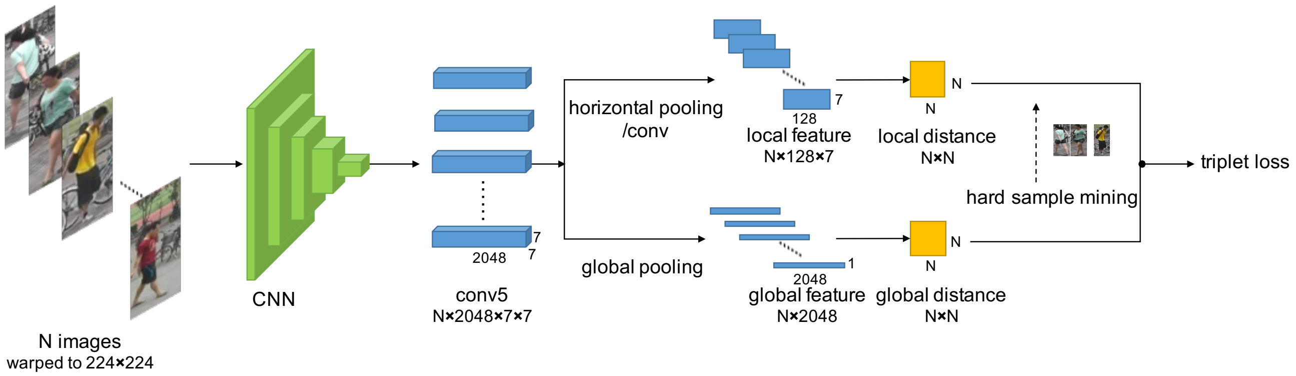

Figure 2. The framework of AlignedReID. Both the global branch and the local branch share the same convolution network to extract the feature map. The global feature is extracted by applying global pooling directly on the feature map. For the local branch, one 1×1 convolution layer is applied after horizontal 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.

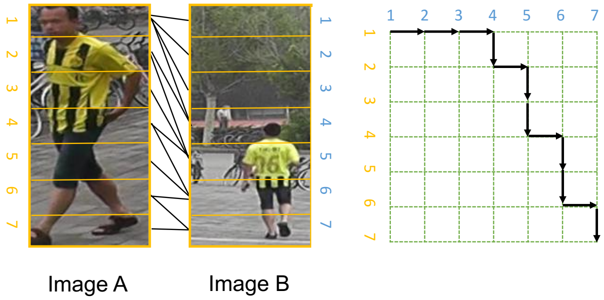

Figure 3. Example of AlignedReID local distance computed by finding the shortest path. The black arrows show the shortest path in the corresponding distance matrix on the right. The black lines show the corresponding alignment between the two images on the left.

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

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

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:

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

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

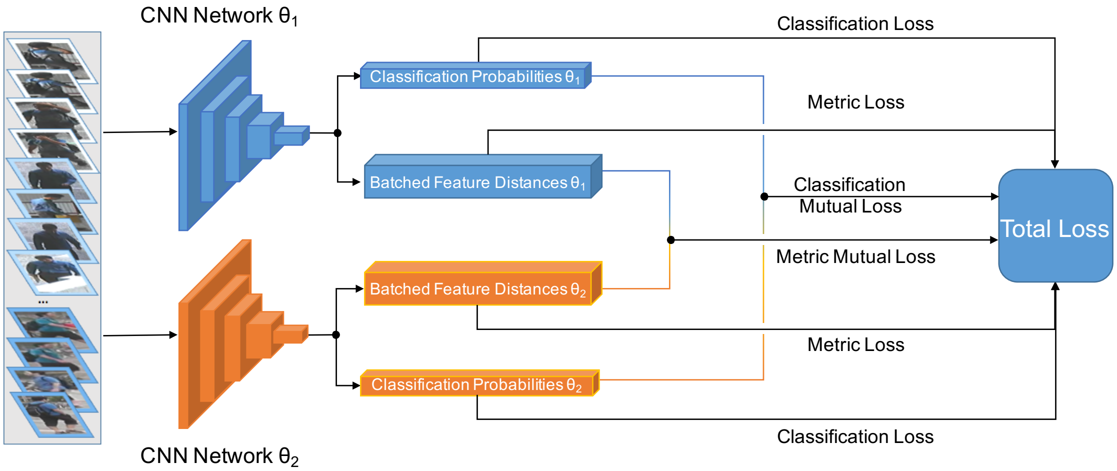
CNN Network θ1
CNN Network θ2
Classification Probabilities θ1
Batched Feature Distances θ1
Batched Feature Distances θ2
Classification Probabilities θ2
Classification Loss
Metric Loss
Classification Mutual Loss
Metric Mutual Loss
Metric Loss
Classification Loss
Total Loss
...

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.

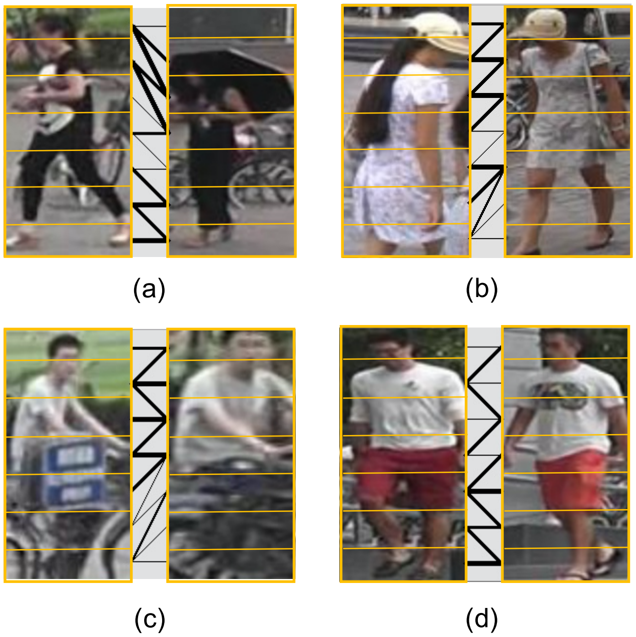

Figure 5. The black lines show the alignments of local parts between two persons: the thicker the line is, the greater it contributes to the shortest path. Persons have the same identities in (a-c), while persons have different identities in (d).

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

		Market1501			MARS			CUHK-SYSU			CUHK03		
Base model	Methods	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.

		Market1501			MARS			CUHK-SYSU			CUHK03		
Loss	Base model	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**

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.

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**

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

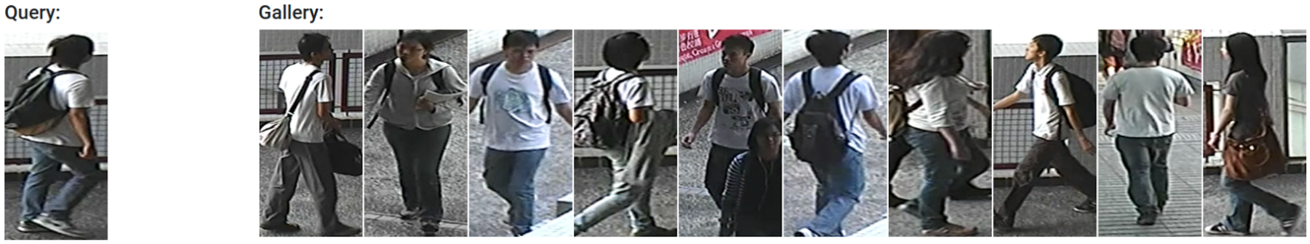

Figure 6. Interface of our human performance evaluation system for CUHK03. The left side shows a query image and the right side shows 10 images sampled using our deep model.

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-

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**

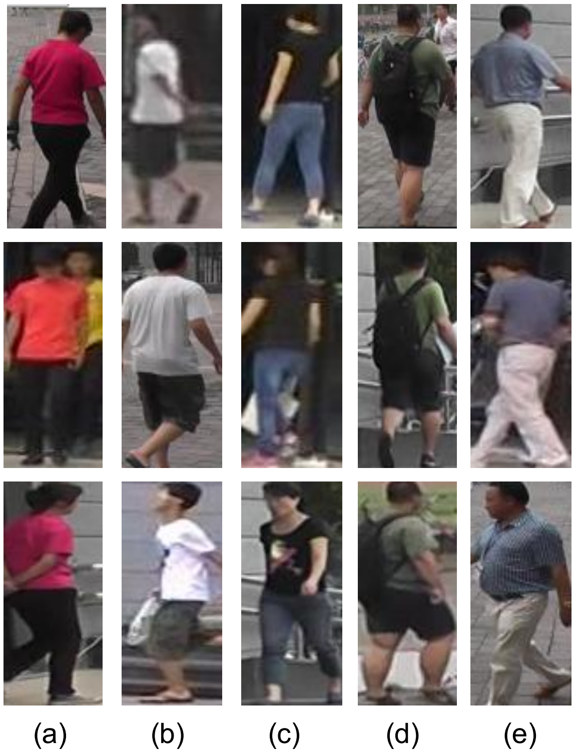

Figure 7. Top: query image. Middle: the result picked by an annotator. Bottom: top1 result by our method.

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

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

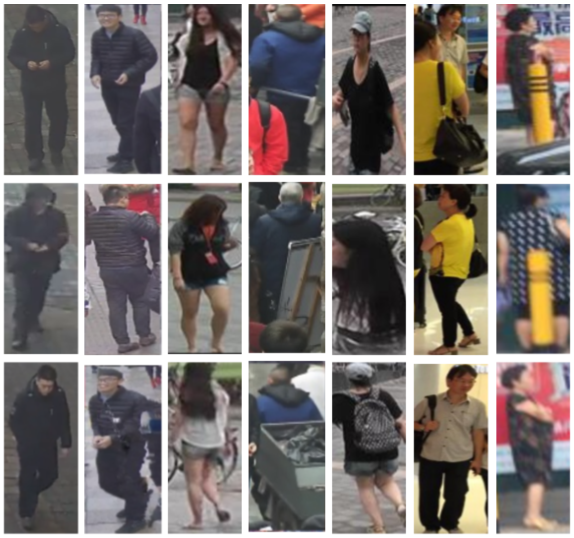

Figure 8. Top: query image. Middle: top1 result by our method. Bottom: ground truth.

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.