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

Anonymous CVPR submission

Paper ID ****

Abstract

Person re-identification (ReID) is a challenging task in computer vision. Deep learning with a metric loss has recently become a common framework for ReID. Many works comparing the similarity of two images have only considered the global feature information, which ignores the local details, whereas other works have considered the local features by using extra information, such as a pose estimation. In this paper, we propose a novel method called AlignedReID that dynamically aligns local features by calculating the shortest path, with no need for extra information. Our method achieves rank-1 accuracy of 94.0% on Market1501 and 96.1% on CUHK03, beating other state-of-the-art algorithms by a large margin. Furthermore, we evaluate the human-level performance and demonstrate that our method is also the first method to outperform human-level performance on Person ReID.

1. Introduction

Person re-identification (ReID) is an important and challenging task in computer vision. It has many applications in surveillance video, such as person tracking across multiple cameras, and person searching in a large gallery *etc*. However, certain issues make this task difficult, such as large variations in poses, viewpoints, illumination, background environments, and occlusions. The similarity of appearances among different persons also increases such difficulty. Some traditional ReID approaches have focused on low-level features such as colors, shapes, and local descriptors [9, 11]. With the development of deep learning, the convolutional neural network (CNN) has been commonly used for feature representation [26, 34, 6, 56, 18, 26]. CNN-based methods can present high-level features, and the similarity is trained end-to-end through multiple metric learning losses, such as contrastive loss [34], triplet loss [20], improved triplet loss [6], quadruplet loss [3], and triplet hard loss [13]. In this way, the performance of person ReID has been significantly improved.

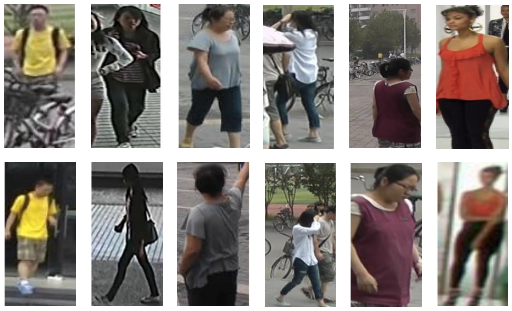

Figure 1. Examples of person ReID datasets. The same identities may be dissimilar because of pose misalignments, various viewpoints, image resolution, occlusions, and illumination *etc*.

Many of the CNN-based features used in person ReID are global features, which ignore the spatial details. In practice, this brings about difficulty in discriminating persons wearing similar clothes, such as shown in Fig 1. To overcome this shortcoming, more and more studies have paid attention to local features. Some divide person images into several parts to extract local features [35, 40, 45], not considering their alignments, whereas others use pose estimation for the alignment [54, 39, 52], which requires an additional model, whose quality largely affects the ReID performance.

In this paper, we propose a novel method, called AlignedReID, which automatically aligns local features without extra information or models. This model is based on the intuition that the corresponing part of the images of the same person should mostly have more similar local features compared to other parts. Moreover, the body of a person is highly structured vertically: the head is always above the chest, whereas the legs are always beneath the abdomen. The alignment between local features of two images should be constrained to be in the same order. For two persons, the minimum sum of the local feature distances among all possible alignments can be used as a proper distance between them.

We also design a mutual learning approach for metric learning to allow two models to learn knowledge from

[†]Equal contribution

each other. The experiments show that our method outperforms other the state-of-the-art methods for Market1501, CUHK03, MARS, and CUHK-SYSU by a large margin. Furthermore, more than ten volunteers were organized to identify the images on Market1501 and CUHK03, and we chose the best result as the human performance. It was shown that our method obtains higher accuracy, and is the first machine able to beat a human with regard to person ReID.

In the following, we overview the main contents of our method and summarize the contributions:

- We propose a novel method for person re-identification, AlignedReID, which dynamically aligns local features of two persons by finding the alignments with the minimum total distance.
- We design a new mutual learning approach for metric learning, which allows two models to learn from each other, thereby improving their performances.
- We collect the human performance on the Market1501 and CUHK03 datasets. The results show our method not only beats other state-of-the-arts methods but is also the first method to outperform a human performance.

The remainder of this paper is organized as follows: related works with additional details are presented in section 2. In section 3, we introduce our SP method and a mutual learning approach. The datasets and experiments are presented in section 4. Finally, some concluding remarks and outlook are presented in section 6.

2. Related Work

2.1. Metric Learning

Before deep learning, most traditional metric learning methods focused on learning a Mahalanobis distance in a Euclidean space. Cross-view Quadratic Discriminant Analysis [17] and Keep It Simple and Straightforward Metric Learning [14] are both previously used classic metric learning methods for person ReID. However, deep metric learning methods usually transform raw images into embedding features, and then compute the similarity scores or feature distances directly in the Euclidean space.

With deep metric learning, two images of the same person are defined as a positive pair, whereas two images of different persons are a negative pair. Triplet loss is motivated by the threshold enforced between positive and negative pairs. In improved triplet loss, a metric learning loss of positive pairs is used to reinforce the clustering of the same person images in the feature space. The positive and negative pairs within a triplet share a common image. A triplet has only two identities. Quadruplet loss adds a new negative pair, and a quadruplet samples four images from three identities. For quadruplet loss, a new loss enforces the distance between positive pairs of one identity and negative pairs of the other two identities. Deep metric learning methods are sensitive to the samples of pairs. Selecting suitable samples for the training model through hard mining has been shown to be effective [13, 3, 41]. A common practice is to pick out dissimilar positive pairs and similar negative pairs according to their similarity scores. Compared with identification or verification loss, metric learning loss for metric learning can lead to a margin between the inter-class distance. However, combining softmax loss with metric learning loss to speed up the convergence is also a popular method.

2.2. Feature Alignments

Aligning local features is a popular approach to person ReID. For instance, pose invariant embedding (PIE) aligns pedestrians to a standard pose to reduce the impact of pose [54] variation and the background. A Global-Local-Alignment Descriptor (GLAD) [39] does not directly align pedestrians, but rather detects key pose points first. It then extracts several regions according to these points, and finally merges the global features with local features. Spindle Net also uses a region proposed network (RPN) to generate several body regions, and it gradually combines the response maps from adjacent body regions at different stages [52]. The above methods all need human pose datasets to train an extra model.

Many research works use a global feature to represent a person image, and compute the distance in the Euclidean space or other embedding spaces to evaluate the similarity between two images. This practice ignores spatial local information of the images. To resolve this disadvantage, some researchers divide images into several parts without an alignment, which suffers from a pose misalignment. Others use a pose or skeleton model to align the parts [54, 39, 52, 35, 40]; however, a significant cost is incurred when training an applicable pose model for person ReID.

2.3. Mutual learning

Mutual learning is at the frontier of deep learning research. [51] presented a deep mutual learning (DML) strategy where an ensemble of students learn collaboratively and teach each other throughout the training process. DML significantly improves the capability of networks. DarkRank [4] introduced a new type of knowledge-cross sample similarity for model compression and acceleration, and obtained a state-of-the-art performance. However, these mutual learning methods are only suitable for a classification problem.

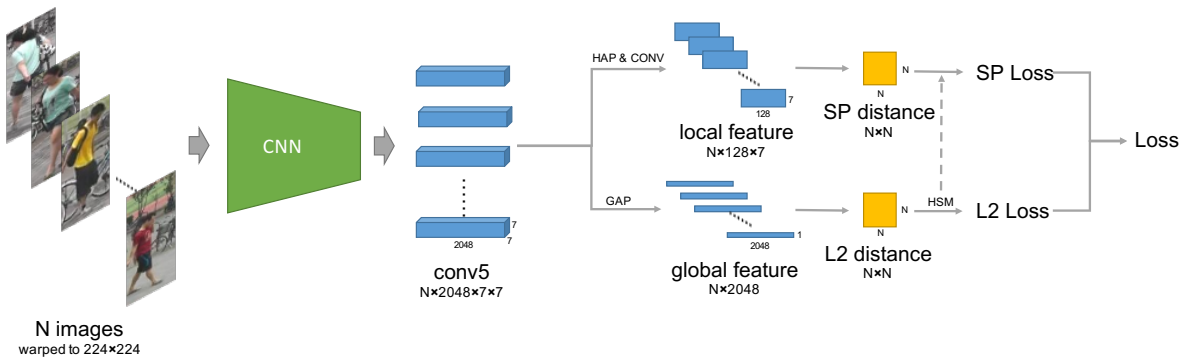

Figure 2. The framework of our AAM method. For a local branch, one 1×1 convolution layer is used to reduce the channels from 2048 to 128 after HAP. Two branches share hard sample mining (HSM) pairs according to the L2 distance. HAP indicates global pooling in a horizontal orientation

2.4. Other Proposed ReID Methods

Some successful unsupervised or transfer learning methods have been recently proposed [8, 29, 28, 48]. One important concern is that bias exists among datasets collected in different environments. Another problem is the lack of labeled data, which can easily cause an overfitting. Despite this, supervised learning methods based on a CNN have been successful for a certain dataset, and a network trained with such a dataset may perform poorly on other datasets. Therefore, one method of transfer learning is to train one task with one dataset, and then fine-tune from the trained model to train another task with another dataset. For example, the model trained on one dataset clusters the other dataset to predict the fake labels that are used to fine-tune the model [8]. In [28], an unsupervised multi-task dictionary learning is proposed to solve the dataset-bias problem.

Natural language description [15] and image data generated by generative adversarial networks (GANs) [58] are respectively regarded as additional information input into the networks. In addition to the image-based learning methods described above, there have also been some video-based person ReID methods developed, which take into account the sequence information, such as the motion or optical flow [38, 47, 46, 24, 27, 53, 21]. RNN architectures and an attention model are also applied when embedding sequence features.

After obtaining the image features, most current works choose the L2 Euclidean distance to compute a similarity score for a ranking or retrieval task. In [37, 59, 1], some re-ranking methods are proposed that clearly improve the ReID accuracy.

3. Our Method

3.1. AlignedReID

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. From the feature map, both the global and local features are extracted. A global feature is extracted by directly applying global pooling to the feature map. For the local features, global pooling is first applied on only the horizontal orientation (HAP) to extract the local feature for each line, and an $1*1$ convolution is then applied to the feature map to reduce the channel number. In this way, each local feature corresponds to a vertical part of the image, or the person. As a result, a person is expressed through a C-channel global feature, and H c-channel local features. The global feature, together with the local features, are then used to compute the similarity of persons.

The similarity of two persons is defined based on their global and local distances. The global distance is easily computed based on the L2 distance of their global features. For the local distance, it is not optimal to compute the L2 distance of each local feature from the corresponding vertical location because there are manifold possible poses, occlusions, camera viewpoint changes, and inaccurate bounding boxes of the human body, as shown in Fig. 1. We intend to compute the L2 distance of each local feature using the same or similar parts of the human body. As is well known, the human body is highly structured in the vertical orientation, and is composed of a head, chest, abdomen, and legs in a fixed topological structure, *e.g.*. This structure is also followed by the local features in the vertical orientation. Therefore, we propose dynamically aligning the local features by matching the local features from top to bottom to find the shortest path. This is based on an intuitive 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 corresponding body part of the other image, when compared to the other parts of the second image.

Given the local features of two images, $F = \{f_1, \cdots, f_H\}$ and $G = \{g_1, \cdots, g_H\}$, the Euclidean distance of each pair is computed, and the following element-

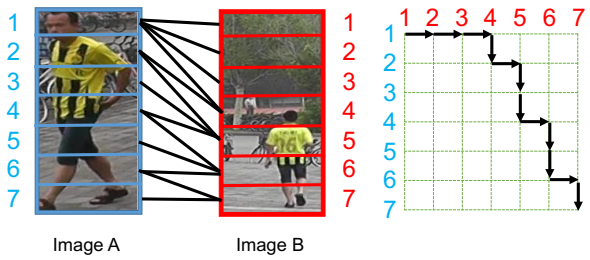

Figure 3. Example of AlignedReID local distance computed by finding the shortest path. The black lines show the shortest path between the two images on the left. The black arrows show the shortest path in the corresponding distance matrix on the right.

wise transformation is applied to normalize it to [0,1), i.e.,

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

Thus, $d_{i,j}$ indicates the distance between the i-th vertical part of the first image and the j-th vertical part of the second image. It is easy to see that, if a pair of local features has a large Euclidean distance, the distance will be close to a constant 1, and the gradient close to 0. A distance matrix D is formed based on these distances, where its (i, j)-element is $d_{i,j}$, as given in Eq.1. The local distance between the two images is defined as the shortest path from $(1, 1)$ to (H, H) in the matrix, which 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 currently the shortest path when walking from $(1, 1)$ to (i, j) in the distance matrix, and $S_{H,H}$ is the shortest path distance of the two images. The back propagation is easy to compute.

As shown in Fig. 3, images A and B are samples of the same person, but the corresponding areas are different. The shortest path is computed following Eq.2. It is shown that the distances between the corresponding body parts, such as the 1st part in the left image, and the 4th part in the right image, are included in the shortest path. Meanwhile, there are distances between non-corresponding parts, such as the 1st part in the left side image, and the 1st part in the right side image, 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. Because the non-corresponding alignment has a large L2 distance, the contribution of such alignments in the shortest path is constant. Furthermore, their gradient is close to 0. Hence, the total distance of the shortest path, i.e., the local distance between two images, is mostly determined by the corresponding alignments.

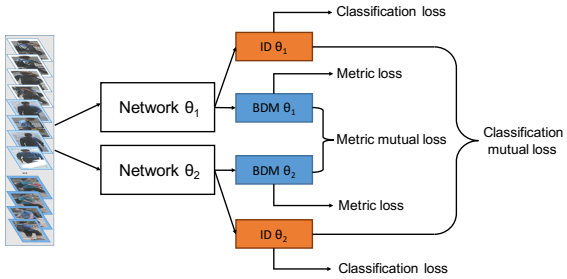

Figure 4. Framework of our mutual learning approach.

The global and local distance together define the similarity between two images, 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 is chosen with the same identity, and the most similar one is chosen with a different identity, 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 depends on two considerations. First, the calculation of the global distance is much faster than that of the local distance. Additionally, we observe that there is no significant difference in mining hard samples using the local distance or both distances.

3.2. Mutual Learning for Metric Learning

We apply mutual learning to train models for AlignedReID, which can further improve the 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 [51]. Differing from [51], which only adopts the Kullback-Leibler (KL) distance between classification probabilities, we propose a new mutual learning loss for metric learning. In this paper, the mutual learning approach trains two networks, simultaneously; however, it is straightforward to extend it to more networks.

Given a batch of N images, each network extracts their ReID features and calculates the distance between each other as an $N \times N$ batch distance matrix. 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 the gradi-

ents. The first- and second-order gradients are then

$$\frac{\partial L_M}{\partial M_{ij}^{\theta_1}} = \frac{2}{N^2}(M_{ij}^{\theta_1} - ZG(M_{ij}^{\theta_2})) = \frac{\partial L_M}{\partial M_{ij}^{\theta_1}} \quad (4)$$

$$\frac{\partial^2 L_M}{\partial M_{ij}^{\theta_1} \partial M_{ij}^{\theta_2}} = 0 \quad (5)$$

When minimizing L_M, both networks can learn knowledge from each other. Differing from [4], which allows better teacher networks to transform knowledge to several student networks, our approach does not limit the upper bound, such as with a teacher network. In AlignedReID, the batch distance matrix for mutual learning loss is computed based on the global distances between images. The reason for this is the same as before: First, the global distance is computed faster. Second, there is no significant difference when using local distances.

The framework of our mutual learning approach is shown in Fig. 4. The overall loss function also includes the classification loss, and KL divergence for classification of mutual loss, as in [51].

4. Experiments

4.1. Datasets

During our experiments, we evaluated our method on public datasets including Market1501 [55], MARS [32], CUHK03 [16], and CUHK-SYSU [43].

Market1501 contains more than 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 used mAP as the evaluation criteria to test the algorithms.

MARS The (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. However, we do not use the sequence information in this paper.

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

CUHK03 contains 14,097 images of 1,467 identities. It provides bounding boxes detected from deformable part models (DPMs) and manually labeling.

Note that we only train one single model using all datasets, as in [42, 52]. We follow the training and evaluation protocol as official papers on Market1501, MARS, and CUHK-SYSU, and mainly report the mAP and rank-1 accuracy in a single query. Because we train one single model for all benchmarks, it is slightly different from the standard procedure in [16], which splits the dataset randomly 20 times on CUHK03. And the gallery for testing only has 100 identities in previous work. However, 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 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 [30] 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 conducted using data augmentaion, including 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 10^{-3}. 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 [59]. In re-ranking experiments, $k1 = 20, k2 = 6$, and $\lambda = 0.3$, which are the same as in [59]. In all of our experiments, we combined metric learning loss with classification (identification) loss. Our models are trained on MegBrain, a self-developed deep learning framework by Megvii, Inc.

4.3. Advantage of AlignedReID

In this section, we compare the proposed AlignedReID method with a similar model that only generates a global feature for ReID. The two models share the same base model, and generate a global feature with the same channel number. The only difference is that AlignedReID also generates local features, which are used to compute the local distance described in Eq.2. The results are shown in Table 1. Resnet50 and Resnet50-Xception achieve a similar performance. The results indicate that AlignedReID improves the $3.5\% \sim 6.0\%$ rank-1 accuracy and $5.0\% \sim 8.4\%$ mAP on all datasets. It helps the network focus on useful image

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

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 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.

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

regions, and discriminates similar regions with slight differences. Finally, we obtain 90.0+% rank-1 accuracy on Market1501 and CUHK-SYSU. On MARS and CUHK03, we reach 86.7% and 88.8% rank-1 accuracy, respectively, for the Resnet50 base model.

4.4. Analysis of Mutual Learning

In the mutual learning experiment, we simultaneously trained two AlignedReID models. One model is based on Resnet50, and the other is based on Resnet50-Xception. We compared their performances for three cases: with both metric mutual loss and classification loss, with only classification mutual loss, and with no mutual loss. We also conducted a similar mutual learning experiment as a baseline, where the local features were removed from the models.

The results are shown in Table 2.

Both experiments show that the metric mutual learning method can further improve the performance. With the baseline mutual learning experiment, the classification of mutual loss can significantly improve the performance on all datasets. However, with the AlignedReID mutual learning experiment, because the models without mutual learning perform well enough, AlignedReID cannot further improve the 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 a better performance for the both baseline and AlignedReID. The experiments also show that the AlignedReID models consistently perform better than the baseline models. Finally, we concatenate the features of the two models trained in one mutual learning experiment, and show the results in the last line of Table 2, which is then used to compare with the human performance in the next section.

5. Human Performance in Person ReID

Experiments on the public datasets show the effectiveness of AlignedReID, which is superior to all known state-of-the-art methods. The question is how does it compare to the performance of a human being. To answer this question, we conducted human performance evaluations and compared our results with the human performance.

5.1. Human Performance Evaluation System

We evaluated the human performance on Market1501 and CUHK03. For each query image, the volunteer did not need to find the identical person from the entire gallery set, but from a much smaller set of selected images.

For CUHK03, for each query image, there is only one image for the identical person in the gallery set. The volunteer looked for the identical person among 10 images selected in the following manner: First, our ReID model generated the top-10 results in the gallery set for the query

Table 3. The results of human performances

	Market1501	CUHK03
Rank-1	93.5	95.7

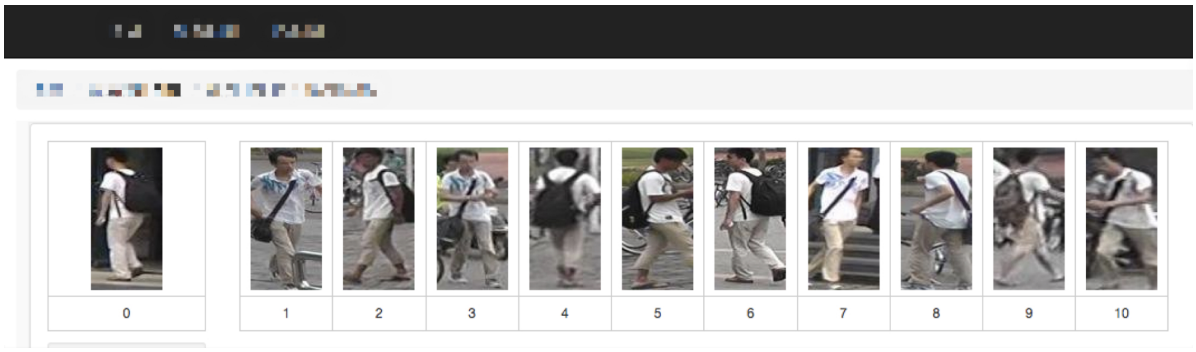

Figure 5. Interface of our human performance evaluation system. The left side shows a query image and the right side shows ten images sampled using our deep model.

image. If the identical person was among these ten results, the top-10 results were selected; otherwise, the top-9 results along with one image of the identical person were selected.

For Market1501, there could be more than one images in the gallery set, representing the same person as the query image. For each query image, the volunteer just needed to choose one image, which was considered to represent the identical person, from 50 images selected in the following manner: First, our ReID model generated the top-50 results in the gallery set for the query image. For each image representing the identical person in the gallery set, if it was not among the results, it would be inserted into the results, meanwhile we removed the image which does not represent the identical person and had the lowest rank in the results. In this way, we make sure that all images with the identical person were in the 50 selected images.

The order of the selected images were then shuffled. The interface of the human performance evaluation system is presented in Fig 5. The evaluation website is also available now: http://reid-challenge.megvii.com for Market1501, and http://reid-challenge.megvii.com/cuhk03 for CUHK03.

Compared to the ReID model, the difficulty of finding an identical person in an entire gallery set was reduced during the evaluation. Including our researchers, professional image labelers, and general citizens, more than ten volunteers participated in the evaluation. Because only one candidate is chosen, we were unable to obtain the mAP of the human beings as a standard evaluation. The rank-1 accuracies were computed for each volunteer on all datasets. The best accuracy was then used as the human performance, which is shown in Table 3.

5.2. Comparison with State-of-the-art Methods and Human Performance

Explanations of Marks. In Tables 4 through $\sim$ 7, * indicates that the respective paper is on ArXiv but has not been published. The black bold marks the highest accuracies of state-of-the-arts methods to the best our knowledge. The blue indicates the best human performance collected during our experiments. The red bold (AlignedReID) shows the results of our best single AlignedReID model. Finally, RK shows the use of re-ranking [59] with k-reciprocal encoding. In this section, we show the results of AlignedReID, the human performance, and the state-of-the-arts methods.

On Market1501, DarkRank[4] obtained an 89.9% rank-1 accuracy and [13] 81.1% for mAP owing to the use of RK. However, human beings achieved 93.5% rank-1 accuracy, which is better than all of the state-of-the-art methods applied. Finally, AlignedReID achieved 92.6% rank-1 accuracy and 82.3% for mAP, and improved to 94.0% and 91.2% after re-ranking, respectively. Thus, AlignedReID after re-ranking outperformed the humans on Market1501.

On CUHK03, in terms of single model without RK, HydraPlus-Net[22] achieve 91.8% rank-1 accuracy and our AlignedReID is 91.9%. Note that, our gallery set is two times as large as that used in [22]. The human performance reached 95.7%, which is much higher than any known state-of-the-art methods. AlignedReID after re-ranking exceeded this, however, and obtained 96.1% rank-1 accuracy, respectively.

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 obtained 87.5% and 86.8% rank-1 accuracy, which is better than all other state-of-the-arts methods.

There have not been many studies on CUHK-SYSU, and we did not evaluate the human performance on this dataset. With this dataset, AlignedReID obtained 93.7% mAP and 95.3% rank-1 accuracy, which is much higher than any known state-of-the-art methods. Additional detailed results can be found in Table 4 $\sim$ 7.

In this paper, we proposed AlignedReID for person re-identification. It dynamically aligns local features of two images by finding a shortest path in the matrix composed using the distances for each pair of the local features. We also proposed a new mutual learning approach for met-

Table 4. Comparison on **Market1501** with single query

Methods	mAP	r=1
Temporal [25]	22.3	47.9
Learning [49]	35.7	61.0
Gated [34]	39.6	65.9
Person [5]	45.5	71.8
Re-ranking [59]	63.6	77.1
Pose [54]	56.0	79.3
Scalable [1]	68.8	82.2
Improving [18]	64.7	84.3
In [13]	69.1	84.9
In (RK)[13]	**81.1**	86.7
Spindle[52]	-	76.9
Deep[51]*	68.8	87.7
DarkRank[4]*	74.3	**89.8**
HydraPlus-Net[22]*	-	76.9
Human Performance	-	**93.5**
AlignedReID	82.3	92.6
AlignedReID (RK)	**91.2**	**94.0**

Table 5. Comparison on **MARS** with single query

Methods	mAP	r=1
Re-ranking [59]	68.5	73.9
Learning [50]*	-	55.5
Multi [33]*	-	68.2
MARS [32]	49.3	68.3
In [13]	67.7	79.8
In (RK)[13]	**77.4**	**81.2**
Quality [23]*	51.7	73.7
See [60]	50.7	70.6
AlignedReID	79.1	86.8
AlignedReID (RK)	**85.6**	**87.5**

ric learning, which further improves the performance of AlignedReID models.

AlignedReID outperforms the other state-of-the-art methods by a large margin. To compare its performance with human beings, ten volunteers participated in our human performance evaluation for person re-identification. The results show that our methods outperform the human performance on Market1501 and CUHK03, which is the first time such an accomplishment has been achieved. The human performance evaluation system will be made open, allowing more human performance results to be collected.

The models were trained on MegBrain, a self-developed deep learning framework by Megvii, Inc. As soon as this framework is made open-source, we will open our models and source code, which will not be difficult owing to its easy implementation.

6. Conclusion

As future work, we will publish a more challenging dataset to promote research on person ReID. Image-based techniques have achieved a good performance when few occlusions occur. We are collecting and labeling a new person ReID dataset, which has many occlusions in a crowded environment.

Table 6. Comparison with existing methods on **CUHK03**

Methods	r=1	r=5	r=10
Person [17]	44.6	-	-
Learning [49]	62.6	90.0	94.8
Gated [34]	61.8	-	-
A [36]	57.3	80.1	88.3
Re-ranking [59]	64.0	-	-
In [13]	75.5	95.2	99.2
Joint [44]	77.5	-	-
Deep [10]*	84.1	-	-
Looking [2]*	72.4	95.2	95.8
Unlabeled [58]*	84.6	97.6	98.9
A [57]*	83.4	97.1	98.7
Spindle[52]	88.5	97.8	98.6
DarkRank[4]*	89.7	**98.4**	**99.2**
HydraPlus-Net[22]*	**91.8**	**98.4**	99.1
Human Performance	**95.7**	-	-
AlignedReID	91.9	98.7	99.4
AlignedReID (RK)	**96.1**	**99.5**	**99.6**

Table 7. Comparison with existing methods on **CUHK-SYSU**

Methods	mAP	r=1
End[43]	55.7	62.7
Neural [19]*	**77.9**	**81.2**
Deep [31]*	74.0	76.7
AlignedReID	**93.7**	**95.3**

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 and co-workers for volunteering to participate 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] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012.

[15] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. *arXiv preprint arXiv:1702.05729*, 2017.

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

[17] 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.

[18] 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.

[19] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan. Neural person search machines. *arXiv preprint arXiv:1707.06777*, 2017.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

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

[52] 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.

[53] 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.

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

[55] 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.

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

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

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

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

[60] 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.