Dynamic Matching of Local Features for Re-Identification of Pedestrians

Seokhyun Ahn* and Nam Ik Cho*
* Department of ECE, INMC, Seoul National University, Seoul, Republic of Korea
E-mail: fervent@ispl.snu.ac.kr, nicho@snu.ac.kr

***Abstract*—This paper presents a person re-identification (re-id) method based on a convolutional neural network (CNN). There are many CNNs for the re-id problems, where few of them considered the image properties in real-world situations that the images could be deformed due to the perspective view of surveillance cameras and that pedestrian detectors do not give perfect bounding box. In this paper, we address the problem of perspective view and incomplete bounding box by proposing a new network architecture and metric learning method. Specifically, we compensate for the vertical and horizontal misalignment due to the incomplete bounding boxes of the pedestrian detector, and also horizontal squeezing that was not considered in the existing algorithms. For this, we partition the bounding box of pedestrian detection results into M horizontal region and N vertical regions. Then, we apply a dynamic matching technique in both horizontal and vertical directions to compensate for the effects of unfit bounding boxes and squeezed appearance due to the perspective views. The partitioning and dynamically matched distance are also considered for the metric training of CNN. We compare our method with state-of-the-art ones and validate the improved performance.**

***Index Terms*—Person Re-Identification, Feature alignment, Hard sample mining, Triplet loss**

I. Introduction

Person re-identification (re-id) is an important task for various purposes, especially in the field of surveillance [27], [28], [29], [30], [31], [1]. The role of re-id is to compare pedestrians in multiple cameras and to determine whether they are the same person or not. Like in many computer vision tasks, deep-learning features [50], [46] are shown to provide better performance for this problem than the engineered ones [51], [34], [52], [53], [54], [55], [56]. Hence, most of the recent re-id researches also use the features extracted from convolutional neural networks (CNN [60]). Specifically, the deep learning-based person re-id algorithms [32], [9] extract feature maps by forwarding images of pedestrians to a CNN, which is usually trained by a metric learning method. Hence, the CNN learns features that make their distance small for similar persons, and large for dissimilar ones. In the test stage, the CNN identifies pedestrians by calculating their feature distances. Although most deep learning-based re-id algorithms have the above procedure in common, the performance varies on what information is used in training and how metric learning is applied.

Early studies calculated the classification loss [33], [42], [43], [46], [47], [48], [49] and deep metric loss by matching feature maps from the entire bounding box given by pedestrian detectors [20], without considering the incompleteness of bounding box detection [9], [8]. Hence, they had some challenges such as unmatched bounding box for the target, pose misalignments due to camera perspectives, occlusions by other objects, etc. Therefore, many different ways [34], [27], [35], [38], [11], [14], [59], [58], [57], [62], [63], [64], [65], [66], [67], [68], [61], [2] to overcome these problems have been studied, and compensating for these problems is the mainstream of person re-id researches. For example, AlignedReID++ [2] is one of the re-id deep networks that apply alignment in detected bounding box images. It divides the feature map of the last convolution layer of Resnet50 [39] into several horizontal stripes and dynamically matched local information to find the shortest local distance in the local branch. As a result, it achieved better performance on several re-id datasets than the methods using global features. But this method focused on only vertical misalignment and overlooked horizontal misalignment and distortion that may occur during the pedestrian detection process.

In this paper, we propose a method that can dynamically align not only the horizontal but also the vertical direction of features, which was not considered in previous methods. It may seem that this is an incremental extension compared to the methods that considered the vertical direction distortions only, but it needs a new branch of the network that correctly extracts the horizontal features, and careful re-training of the CNN with a new loss function that balances the horizontal and vertical loss very well. In addition, we also evaluate the performance of varying numbers of partitions in each direction. For this, we add a new branch to the CNN of [2] as the baseline, which extracts the features into the horizontal direction as well. Then the branch is merged to the one that extracts features into the vertical direction, with appropriate loss functions considering the number of stripes in each direction. Experiments on three widely used datasets, CUHK03 [9], Market1501 [7], and DukeMTMCReID [8] show that our method brings some gains over the baseline, especially for the CUHK03 that contains more challenging images.

The rest of this paper is organized as follows. In Section II, we briefly review the related works. Section III introduces the proposed network and training process in detail. The experimental results, comparisons, and analyses are presented in Section IV. Conclusion comes in Section V.

Fig. 1: Examples of horizontally misaligned bounding box of pedestrian detection.

II. Related Work

A. Feature alignment in Person Re-ID

The recent widely adopted CNN-based object detectors usually give the detection result as a bounding box that fits the object [20], [44], [45], [22], [21], [23], [24], [25], [26]. But the detectors are not perfect such that there can be a wide area of background in the bounding box or conversely, part of the object is out of the bounding box. We found that most of the re-id errors are due to this kind of bounding box misfit by observing the results on re-id datasets such as CUHK03[9] and DukeMTMCReID[8]. Specifically, the re-id usually fails when the bounding box of pedestrian detectors is shifted out of the target. The re-id also fails when the view of two objects are largely different, such that the aspect ratio of two objects appear differently. Hence, a number of studies have been conducted to overcome the imperfection of bounding box of pedestrian detectors and differences of perspective views. For example, [18], [35], [36], [37], [38] used multiple shots of a person to enhance body part alignment. [19] made use of the periodicity property of pedestrians and divided the walking cycle into several segments, which are described by temporally aligned pooling. More recently, [9], [2] divided feature maps into horizontal local areas and corrected vertical misalignment. They achieved competitive performances, but they overlooked the horizontal misalignment that can occur as often as the vertical ones (Fig. 1). To alleviate this problem, we propose a method that divides feature maps into M horizontal and N vertical local parts, and balances and matches them by a dynamic matching method.

B. Triplet loss in Person Re-ID

Triplet loss was first used for person re-id in [5], [15], [16]. This metric encourages the distances between feature vectors of inter-class pairs to be less than those of intra-class pairs with a specific margin. Triplet loss is also frequently used with hard sample mining [40], [15], [41]. However, in a system that adopts complex metric learning, it is hard to apply the triplet loss to all instances in a batch. Therefore, it is often recommended to apply it to only a few mined hard positive/negative samples. By doing this, the loss can concentrate on the key hard samples and bypass the enormous quantity of computation. In this paper, we follow this trend and adopt triplet loss for training the overall structures using mined hard samples based on global distance.

III. Proposed Method

In this section, we illustrate the structure of proposed network, which is summarized as Fig. 2, and also explain the training method.

A. The overall framework

During the training, the goal of our network is to classify and match the pedestrians using entire features at one branch, while correcting misalignments both horizontally and vertically at the other two branches. Each image in the training sets is resized to be the same ($32M \times 32N$). Feature map of size $M \times N \times 2048$ is extracted at the last convolution layer of the backbone network. At the global branch, global average pooling is applied to obtain global feature vector f_g. Here, the classification loss of L_{cls} is calculated after following the fully connected layer and a softmax layer. At the same time, distance of the global feature vector pairs in the batch are calculated. Next, hard sample mining is performed based on this global distance. At the two local branches, horizontal and vertical max poolings are performed, which yield two local feature vectors f_{lh} ($M \times 1 \times 128$) and f_{lv} ($1 \times N \times 128$) following the channel reduction. After that, dynamic local matching of this information is performed to calculate local distances. Finally, triplet losses (L^g_{tri}, L^l_{tri}) are defined from mined hard samples, and the total loss for training the network is defined as

$$L = L_{cls} + L^g_{tri} + L^l_{tri}. \tag{1}$$

In the following paragraphs, we explain the losses step by step.

B. Classification Loss L_{cls}

Person re-id is to identify K different ID labels for each instance in a given training or gallery (test) set, in terms of classification. Hence, we define the classification loss L_{cls} as a kind of cross-entropy calculated as:

$$L_{cls} = -\sum_{i=0}^{B-1}\sum_{k=0}^{K-1} \hat{p}(i,k) \log p(i,k) \tag{2}$$

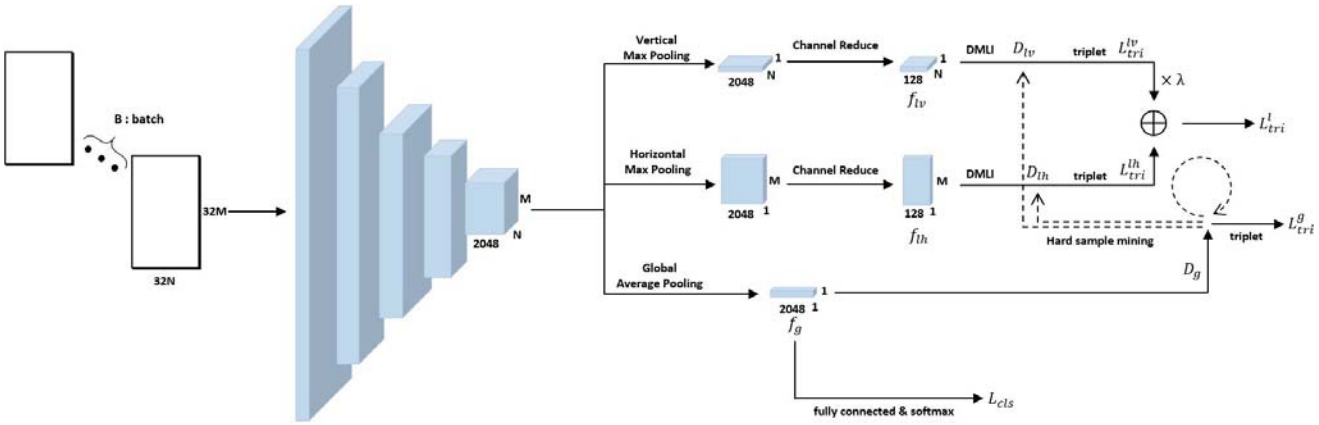

Fig. 2: The overall structure of the proposed network. Resnet50 is employed as a backbone network to extract features of input images. In the following three branches, the global branch applies global average pooling to raw features for classifying each sample, calculating global distances, and performing hard sample mining. The other two local branches adopt horizontal and vertical max pooling to the raw features and obtain local distances of hard samples dynamically.

$$\hat{p}(i,k) = \begin{cases} 0, & k \neq y_i \\ 1, & k = y_i \end{cases}, \quad p(i,k) = \frac{\exp(z_k)}{\sum_{l=0}^{K-1} \exp(z_l)},$$

where B is the batch size, K is the number of ID labels in a training set, y_i is the ground-truth label of the sample i, $\hat{p}(i,k)$ denotes one hot ground-truth label, z_k is the k-th output of $2048 \times K$ sized fully connected layer, and $p(i,k)$ is the final softmax output.

C. Hard Sample Mining & Global Triplet Loss L_{tri}^{g}

Next, hard sample mining is performed to obtain the triplet loss. Before mining, we obtain $B \times B$ global distance matrix D_g over all pairs in a batch. The global distance $D_g(i,j)$ between two global feature vectors f_g^i and f_g^j of instances i and j is defined as:

$$D_g(i,j) = ||f_g^i - f_g^j||_2. \tag{3}$$

After that, one hard positive index p_i and one hard negative index n_i are mined using this global distance from each sample i as shown in Fig. 3, which are expressed as:

$$p_i = \underset{j \in \{j | y_i = y_j\}}{\operatorname{argmax}} D_g(i,j), \quad n_i = \underset{j \in \{j | y_i \neq y_j\}}{\operatorname{argmin}} D_g(i,j) \tag{4}$$

After gaining hard sample indices for all samples in a batch, the global triplet loss is calculated as:

$$L_{tri}^{g} = \frac{1}{B} \sum_{i=0}^{B-1} max(D_g(i,p_i) - D_g(i,n_i) + m, 0) \tag{5}$$

where m is the preset margin value. If the difference between $D_g(i,n_i)$ and $D_g(i,p_i)$ is smaller than m, (5) is activated, which is otherwise 0. This means if the gap between hard positive and negative is less than the margin m, then we push them away from each other to distinguish between positive and negative cases clearly.

D. Local Triplet Loss L_{tri}^{l}

At local branches, only local distances of hard samples are calculated to evade too much computational cost. As stated previously, horizontal and vertical local feature vectors f_{lh} and f_{lv} are divided into M horizontal and N vertical regions. Before matching, the local partial distances are calculated over all $M \times M$ pairs of horizontal parts and $N \times N$ pairs of vertical parts in sample i and j. The local part distance between the p-th region of i and the q-th region of j, is defined as:

$$d_x^{i,j}(p,q) = \frac{\exp(||f_{lx}^i(p) - f_{lx}^j(q)||_2) - 1}{\exp(||f_{lx}^i(p) - f_{lx}^j(q)||_2) + 1} \tag{6}$$

$$p, q \in 0, 1, \ \dots \ , T-1, \quad i, j \in 0, 1, \ \dots \ , B-1,$$

$$x \in \{h, v\}, \quad T \in \{M, N\}$$

and then these distances are allocated to all grid points one-by-one on $M \times M$ or $N \times N$ size lattice Fig. 7. Again, at the same point, $S_x^{i,j}(p,q)$ is calculated dynamically as Fig. 5:

$$S_x^{i,j}(p,q) = \begin{cases} d_x^{i,j}(p,q) & p = 0, \ q = 0 \\ S_x^{i,j}(p-1,q) + d_x^{i,j}(p,q) & p > 0, \ q = 0 \\ S_x^{i,j}(p,q-1) + d_x^{i,j}(p,q) & p = 0, \ q > 0 \\ \min(S_x^{i,j}(p-1,q), S_x^{i,j}(p,q-1)) & \\ \quad + d_x^{i,j}(p,q) & p > 0, \ q > 0 \end{cases} \tag{7}$$

$$i, j \in 0, 1, \ \dots \ , B-1, \quad x \in \{h, v\}$$

and the final local distance is the last element of the above process, denoted as :

$$D_{lh}(i,j) = S_h^{i,j}(M-1, M-1), D_{lv}(i,j) = S_v^{i,j}(N-1, N-1) \tag{8}$$

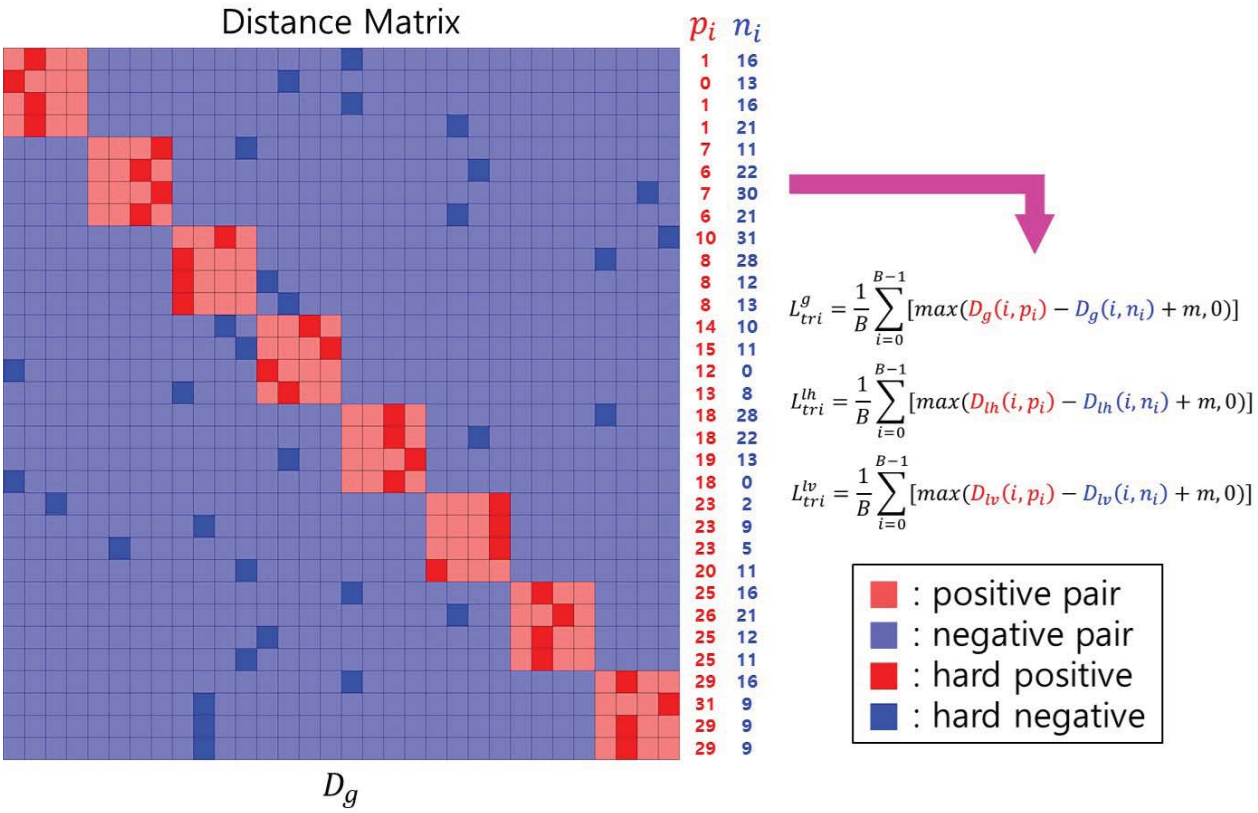

Fig. 3: Hard Sample Mining on global distance matrix D_g. Indices of hard samples are obtained by applying (4) row-wise on this matrix. After that, the final three triplet losses are calculated from these indices.

Finally, horizontal local triplet loss L_{tri}^{lh} and vertical local triplet loss L_{tri}^{lh} are defined as:

$$L_{tri}^{lh} = \frac{1}{B}\sum_{i=0}^{B-1} max(D_{lh}(i,p_i) - D_{lh}(i,n_i) + m, 0) \quad (9)$$

$$L_{tri}^{lv} = \frac{1}{B}\sum_{i=0}^{B-1} max(D_{lv}(i,p_i) - D_{lv}(i,n_i) + m, 0) \quad (10)$$

which have the same formation as (5). The total local triplet loss used for the training is the sum of (9) and (10):

$$L_{tri}^{l} = \frac{1}{2}(L_{tri}^{lh} + \lambda L_{tri}^{lv}) \quad (11)$$

where λ is the balancing parameter between two local triplet losses.

E. Implementation Details

We implement the proposed network in the Pytorch deep learning framework. We use Resnet50 as the backbone network for feature extraction using one NVIDIA TITAN Xp GPU with the batch size $B = 32$. The feature map fed to the following three branches is extracted at the last convolution layer of the backbone network. Since Resnet50 downsamples input by a factor of 32, we set the size of network input to 256×128 for gaining the desired size of the feature map. During training, we train network for 300 epochs, and set margin value $m = 0.3$ for all 3 triplet losses (5), (9), and (10). At the vertical branch, a total of $2N - 1$ values are summed, and $2M - 1$ at the horizontal branch. Hence, when we obtain the local triplet loss L_{tri}^{l}, if we use a direct sum of (9) and (10), vertical alignment's influence is far lesser than that of horizontal counterpart. Therefore, we set $\lambda = \frac{2M-1}{2N-1}$ of (11) to balance the influences of two local branches equally. By default, we set $M = 8$, $N = 4$, and $\lambda = \frac{15}{7}$. In sec IV, the above settings are followed unless specified differently. In ablation studies, we modify these default settings and measure the effect of each hyperparameter.

IV. Experiments

A. Datasets

For evaluating the performances, experiments are conducted on three benchmark data sets, CUHK03 [9], Market1501 [7], and DukeMTMCReID [8], which are widely used in the person re-id field. These benchmark datasets are divided into three subsets: train set, query set, and gallery (test) set. In the

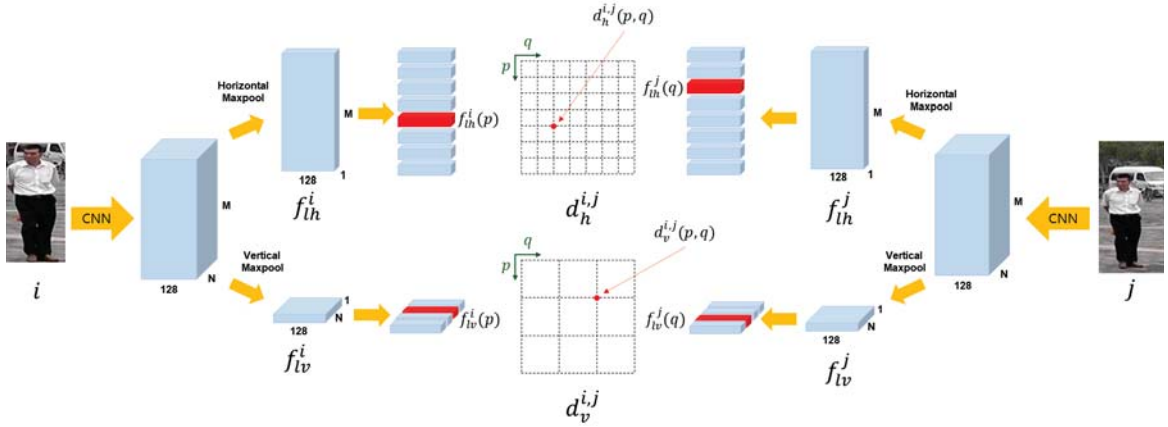

Fig. 4: In two local branches, $M \times N \times 128$ size feature maps of a pair are horizontally and vertically max-pooled each. Next, the max-pooled feature maps are separated M horizontal part and N vertical part feature vectors. Finally, distances (6) among all $M \times M$ pairs of horizontal part feature vectors and $N \times N$ pairs of vertical part feature vectors of two images are allocated on corresponding position of lattice graph one-by-one. For example, a distance value between p-th part vector of i and q-th part vector of j is allocated at (p, q) on the grid.

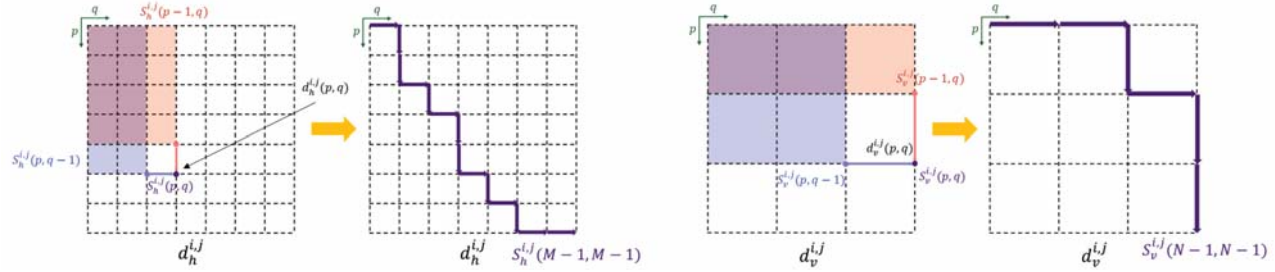

Fig. 5: Dynamic process for calculating the final local distance (7). From the lattices of Fig. 4, the final local distance is obtained by finding minimum sum recursively among all paths from left-top to right-bottom.

training process, only the train set is used, and in the test process, only the query set and the gallery set are used, and the performance is measured by finding and ranking the same person in the gallery set by using the query set as a probe. The IDs of pedestrians in the query set and gallery set are the same, and there is no same person as the pedestrians in the train set used for learning.

More precisely, the CUHK03 dataset consists of a train set consisting of 767 pedestrian IDs, 7365 images, 700 pedestrian IDs, a query set consisting of 1400 images, 700 pedestrian IDs as a query set, and a gallery set consisting of 5332 images. The CUHK03 dataset is generally known to have harder cases so that the re-id algorithms usually show lower performance for this set than others such as Market1501 and DukeMTMCReID.

The Market1501 consists of 1501 pedestrian IDs and 32217 images taken from six camera views. Among them, the train set consists of 751 pedestrian IDs and 12936 images, the query set consists of 750 pedestrian IDs, 3368 images, and the gallery set consists of the same 750 pedestrian IDs and 19732 images. In this paper, we adopt a single query mode.

The DukeMTMCReID is a reconstruction of the pedestrians detected in the eight videos for pedestrian recognition. Together with Market1501 and CUHK03, the DukeMTMCReID is the most widely used benchmark dataset in this field. The train set consists of 702 pedestrian IDs and 16522 images. The query set contains 702 person IDs and 2228 images that do not exist in the train set. The test set is not included in the query set. Similar but different 408 distractor IDs are added to disturb the trained model, consisting of 1110 pedestrian IDs and 17661 images.

B. Visualization

First, we tested the qualitative effectiveness of our proposed method with that of AlignedReID++[2], which is an up-to-date algorithm correcting vertical misalignment, by measuring unaligned and aligned distance between 2 positive pairs (one of the pairs is well aligned, the other is horizontally misaligned) in Market1501. We calculate the unaligned distance by matching each parts of two images one by one in order, averaging distances of diagonal parts on the grids. On the other hands, we calculate aligned distance by eq. 7 and average of the number of points on the path, 2M-1 or 2N-

	AlignedReID++	Proposed
Unaligned Distance	0.5295	0.6316
Aligned Distance	0.5338	0.5787

	AlignedReID++	Proposed
Unaligned Distance	0.6106	0.4980
Aligned Distance	0.6227	0.4669

Fig. 6: Visualization examples of two positive pairs in Market1501 dataset. Since our goal is smaller intra-class distance even though one or both of positive pair is horizontally misaligned, it is desirable that aligned distance is far smaller than unaligned distance in order to prevent false negative. As seen in the tables, the proposed method compensates the horizontal misalignment better than AlignedReID++.

TABLE I: Comparison with state-of-the-art methods on CUHK03, Market1501, and DukeMTMCReID. We set test distance as local distance. For a fair comparison, we show the results of using only DMLI local distance on Market1501, DukeMTMCReID in case of AlignedReID++.

Method	CUHK03		Market1501		DukeMTMCReID	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
FPNN [9]	-	19.9	-	-	-	-
PAN [3]	34.0	36.3	63.4	82.8	51.5	71.6
SVD [13]	37.3	41.5	62.1	82.3	56.8	76.7
FMN [17]	39.2	42.6	67.1	86.0	56.9	74.5
PCE&ECN [14]	27.3	30.2	69.0	87.0	62.0	79.8
HA-CNN [4]	41.0	44.4	75.7	91.2	63.8	80.5
PIE [11]	41.2	45.9	69.3	87.3	64.1	80.8
MLFN [12]	47.8	52.8	74.3	90.0	62.8	81.0
AlignedReID++ [2]	59.7	60.9	77.4	91.1	67.7	81.0
Proposed	**60.2**	**61.9**	**78.9**	**91.7**	**67.8**	**81.1**

1. Here, we allocated L2-distances $||f^i_{lx}(p) - f^j_{lx}(q)||_2$ on the grid points instead of (6). The final distance is average of horizontal local distance and vertical local distance. As seen in the Fig. 6, in AlignedReID++, unaligned distance is larger than aligned distance. Here, we can see that the AlignedReID++ doesn't compensate horizontal distortion well, considering that aligned distance is smaller than unaligned distance. Meanwhile, our proposed method corrects it better than AlignedReID++, evidenced by the opposite result.

C. Comparision with state-of-the-art

We compare the performance of the proposed method with those of state-of-the-art person re-id methods that consider the misalignment. Table I illustrates the experimental results on CUHK03, Market1501, and DukeMTMCReID. We adopt the mean Average Precision (mAP) and Top-1 rank score metric to measure and compare the performances of each method. Here, we set the test distance as dynamically matched local distance. For a fair comparison, we decide not to use any post-processing, such as re-ranking [6], MultiQueryFusion [7], Label smoothing [10], etc. We compare our method with several representative methods such as SVD [13], HA-CNN [4], MLFN [12], and AlignedReID++ [2]. From the table, we can see that our method achieves competitive results on these three major benchmark datasets.

D. Ablation studies

To analyze the effect of balancing factor in (11), we set $\lambda = 1$ (direct sum without balance) and compare it to the original setting. As shown in Table II, scaling L^{lv}_{tri} by a factor of $\lambda = \frac{2M-1}{2N-1}$ outperforms the case of setting $\lambda = 1$ by a margin of 1.5%p (Top-1)/1.5%p(mAP), 0.5%p(Top-1)/0.6%p(mAP) on CUHK03 and Market1501 respectively. On the other hand, the latter outperforms the former by a margin of 0.1%p(mAP)/0.3%p(Top-1) on DukeMTMCReID, but the margin is far lesser than CUHK03 and Market1501 cases. From this, we conclude that balancing the influences of horizontal and vertical alignment brings better results than simply adding two local triplet losses (9) and (10).

Next, we change the margin m to 0.6 or 1.2, and compare them with default setting $m = 0.3$. On the CUHK03, setting $m = 1.2$ yields slightly better result than $m = 0.3$, by 0.3%p in Top-1. The former outperforms the latter by a margin of 0.5%p(mAP)/1.5%p(Top-1) on DukeMTMCReID. On the

TABLE II: Comparison of several parameter settings. We change balancing factor λ to 1 or margin m to 0.6 or 1.2 and compare each performance to that of the original setting.

Setting	CUHK03		Market1501		DukeMTMCReID	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
$\lambda = 1$	58.7	60.4	78.4	91.1	67.9	81.4
$m = 0.6$	59.7	60.5	78.0	90.5	68.2	81.1
$m = 1.2$	**60.2**	**62.2**	78.2	90.5	**68.3**	**82.6**
default($m = 0.3, \lambda = \frac{15}{7}$)	**60.2**	61.9	**78.9**	**91.7**	67.8	81.1

TABLE III: Comparison on various feature map sizes. Here, we do not change any part of the backbone network but input image size. Since the input image is downsampled by a factor of 32, we set network input size $32M \times 32N$ so that we can obtain the desired $M \times N$ size of the feature map. For simplicity, we only tested on $M = 2N$ cases.

Size of feature map	CUHK03		Market1501		DukeMTMCReID	
	mAP	Top-1	mAP	Top-1	mAP	Top-1
$M = 4, N = 2$	46.8	47.1	68.5	85.2	58.3	75.0
$M = 6, N = 3$	56	58	75.9	89.7	65.7	79.8
default($M = 8, N = 4$)	60.2	**61.9**	**78.9**	**91.7**	67.8	**81.1**
$M = 10, N = 5$	**60.4**	61.4	78.1	90.8	**68.0**	80.6
$M = 16, N = 8$	58.4	57.7	74.7	90.1	64.8	79.7

other hand, the default case outperforms the modified cases by a margin of 0.7%p(mAP)/1.2%p(Top-1) on Market1501. The other setting, $m = 0.6$ does not give better results than the above two cases. Of course, setting $m = 1.2$ seems slightly better than $m = 0.3$, considering the sum of the margins. Yet, it is hard to say that the former is definitely superior to the latter. Meanwhile, we set $m = 0.3$ when comparing the method with others, for a fair comparison.

Additionally, we change the feature map size variously and compare the performances of each case. We control the input size to obtain the desired size of the feature map indirectly without any change to the backbone network. Input images in the dataset are resized to $32M \times 32N$ before being fed to the network so that the resized input is downsampled by a factor of 32 to $M \times N$. For simplicity, we only tested on $M = 2N$ cases, considering the general aspect ratio of pedestrians. In these cases, the balancing factor λ depends on the feature map size. As shown in Table III, setting $(M, N) = (8, 4)$ and $(M, N) = (10, 5)$ outperform the other cases by large margins. Since we prefer Top-1 score and less computational complexity to mAP, we choose $M = 8$ and $N = 4$ as a default setting.

V. Conclusion

In this paper, we have proposed a person re-id method, which works robustly against vertical and horizontal misalignment of bounding boxes that contain the targeting pedestrians. By developing a dynamic horizontal and vertical feature matching algorithm, with appropriate loss functions for the metric learning of local features, the proposed method is shown to perform better than the state-of-the-art methods in several widely used benchmarks. We expect that the proposed method would contribute to the construction of robust person re-id systems, in that there are many different camera views and obstacles that distort and occlude pedestrians in practical situations.

Acknowledgment

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2016-0-00288) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation).

References

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

[2] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, Chi Zhang (2019) AlignedReID ++ : Dynamically matching local information for person re-identification. Pattern Recognition, 94, 53-61

[3] Zheng, Z., Zheng, L., Yang, Y. (2018) Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 29(10), 3037-3045

[4] Wei Li, Xiatian Zhu, Shaogang Gong (2018) Harmonious Attention Network for Person Re-Identification. IEEE conference on computer vision and pattern recognition, 2285-2294

[5] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, Nanning Zheng (2016) Person Re-Identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function. IEEE International Conference on computer vision and pattern recognition, 1335-1344

[6] Z. Zhong, L. Zheng, D. Cao, S. Li (2017) Re-ranking person re-identification with k-reciprocal encoding. IEEE International Conference on Computer Vision and Pattern Recognition, 3652–3661

[7] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian (2015) Scalable person re-identification: A benchmark. IEEE International Conference on Computer Vision, 1116-1124

[8] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C. (2016) Performance measures and a data set for multi-target, multi-camera tracking. European Conference on Computer Vision, 17-35

[9] W. Li, R. Zhao, T. Xiao, X. Wang (2014) Deepreid: Deep filter pairing neural network for person re-identification. IEEE International Conference on Computer Vision and Pattern Recognition, 152–159

[10] C. Szegedy , V. Vanhoucke , S. Ioffe , J. Shlens , Z. Wojna (2016) Rethinking the inception architecture for computer vision. IEEE International Conference on Computer Vision and Pattern recognition, 2818–2826

[11] Zheng, L., Huang, Y., Lu, H., and Yang, Y. (2019) Pose-invariant embedding for deep person re-identification. IEEE Transactions on Image Processing, 28(9), 4500-4509
[12] X. Chang , T.M. Hospedales , T. Xiang (2018) Multi-level factorisation net for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 2109-2118
[13] Y. Sun , L. Zheng , W. Deng , S. Wang (2017) Svdnet for pedestrian retrieval. IEEE International Conference on Computer Vision, 3800–3808
[14] Saquib Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R. (2018) A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. IEEE Conference on Computer Vision and Pattern Recognition, 420-429
[15] Alexander Hermans, Lucas Beyer, and Bastian Leibe (2017) In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737
[16] Wentong Liao, Michael Ying Yang, Ni Zhan, and Bodo Rosenhahn (2017) Triplet-based deep similarity learning for person re-identification. IEEE International Conference on Computer Vision Workshop, 385–393
[17] G. Ding, S. Khan, Z. Tang, and F. Porikli (2017) Let features decide for themselves: Feature mask network for person re-identification. arXiv:1711.07155
[18] S. Karanam, Y. Li, and R. J. Radke (2015) Sparse re-id: Block sparsity for person re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 33–40
[19] C. Gao, J. Wang, L. Liu, J.-G. Yu, and N. Sang (2016) Temporally aligned pooling representation for video-based person re-identification. IEEE International Conference on Image Processing, 4284–4288
[20] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009) Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9), 1627-1645
[21] Redmon, J., and Farhadi, A. (2017) YOLO9000: better, faster, stronger. IEEE conference on Computer Vision and Pattern Recognition, 7263-7271
[22] Ren, S., He, K., Girshick, R., and Sun, J. (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Conference on Neural Information Processing Systems, 91-99
[23] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C. (2016) Ssd: Single shot multibox detector. European Conference on Computer Vision, 21-37
[24] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017) Feature pyramid networks for object detection. IEEE conference on Computer Vision and Pattern Recognition, 2117-2125
[25] Dai, J., Li, Y., He, K., and Sun, J. (2016) R-fcn: Object detection via region-based fully convolutional networks. Neural Information Processing Systems, 379-387
[26] Lin, T. Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017) Focal loss for dense object detection. IEEE International Conference on Computer Vision, 2980-2988
[27] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux (2008) Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. in ACM/IEEE International Conference on Distributed Smart Cameras, 1–6
[28] X. Wang (2013) Intelligent multi-camera video surveillance: A review. Pattern recognition letters, 34(1), 3–19
[29] W. Zajdel, Z. Zivkovic, and B. Krose (2005) "Keeping track of humans: Have i seen this person before?". IEEE International Conference on Robotics and Automation, 2081–2086
[30] D. Gray, S. Brennan, and H. Tao (2007) Evaluating appearance models for recognition, reacquisition, and tracking. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, 3(5), 1-7
[31] D. Baltieri, R. Vezzani, and R. Cucchiara (2011) 3dpes: 3d people dataset for surveillance and forensics. The 2011 joint ACM workshop on Human gesture and behavior understanding, 59–64
[32] Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014) Deep metric learning for person re-identification. IEEE 22nd International Conference on Pattern Recognition, 34-39
[33] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems, 1097-1105
[34] Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010) Person re-identification by symmetry-driven accumulation of local features. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (pp. 2360-2367). IEEE.
[35] Ma, B., Su, Y., and Jurie, F. (2012) Local descriptors encoded by fisher vectors for person re-identification. European Conference on Computer Vision, 413-422
[36] Ma, B., Su, Y., Jurie, F. (2012) Bicov: a novel image representation for person re-identification and face verification
[37] Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011) Person re-identification by descriptive and discriminative classification. Scandinavian Conference on Image Analysis, 91-102
[38] Li, W., and Wang, X. (2013) Locally aligned feature transforms across views. IEEE Conference on Computer Vision and Pattern Recognition, 3594-3601
[39] He, K., Zhang, X., Ren, S., and Sun, J. (2016) Deep residual learning for image recognition. IEEE conference on computer vision and pattern recognition, 770-778
[40] E. Ristani, and C. Tomasi (2018) Features for multi-target multi-camera tracking and re-identification. IEEE Conference on Computer Vision and Pattern Recognition, 6036–6046
[41] Q. Xiao, H. Luo, and C. Zhang (2017) Margin sample mining loss: a deep learning based method for person re-identification. arXiv: 1710.00478
[42] Luo, H., Gu, Y., Liao, X., Lai, S., and Jiang, W. (2019) Bag of tricks and a strong baseline for deep person re-identification. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 0-0
[43] Weinberger, K. Q., Blitzer, J., and Saul, L. K. (2006) Distance metric learning for large margin nearest neighbor classification. Neural Information Processing Systems, 1473-1480
[44] R. Girshick, J. Donahue, T. Darrell, and J. Malik (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, 580–587
[45] Girshick, and R. (2015) Fast r-cnn. IEEE International Conference on Computer Vision, 1440-1448
[46] Xiao, T., Li, H., Ouyang, W., and Wang, X. (2016) Learning deep feature representations with domain guided dropout for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1249-1258
[47] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016) Mars: A video benchmark for large-scale person re-identification. European Conference on Computer Vision, 868-884
[48] Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., and Yang, X. (2016) Person re-identification via recurrent feature aggregation. European Conference on Computer Vision, 701-716
[49] Xie, L., Wang, J., Wei, Z., Wang, M., and Tian, Q. (2016) Disturblabel: Regularizing cnn on the loss layer. IEEE Conference on Computer Vision and Pattern Recognition, 4753-4762
[50] Varior, R. R., Haloi, M., and Wang, G. (2016) Gated siamese convolutional neural network architecture for human re-identification. European Conference on Computer Vision, 791-808
[51] Gheissari, N., Sebastian, T. B., and Hartley, R. (2006) Person reidentification using spatiotemporal appearance. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2, 1528-1535 IEEE.
[52] Gray, D., and Tao, H. (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. European conference on computer vision, 262-275
[53] Prosser, B. J., Zheng, W. S., Gong, S., Xiang, T., and Mary, Q. (2010) Person re-identification by support vector ranking. BMVC, 2(5), 6
[54] Zheng, W. S., Gong, S., and Xiang, T. (2012) Reidentification by relative distance comparison. IEEE transactions on pattern analysis and machine intelligence, 35(3), 653-668
[55] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3567–3574
[56] A. Mignon and F. Jurie (2012) Pcca: A new approach for distance learning from sparse pairwise constraints. IEEE Conference on Computer Vision and Pattern Recognition, 2666–2672
[57] Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., ... and Tang, X. (2017) Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1077-1085
[58] Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q. (2017) Glad: Global-local-alignment descriptor for pedestrian retrieval. The 25th ACM international conference on Multimedia, 420-428

[59] C. Su , J. Li , S. Zhang , J. Xing , W. Gao , Q. Tian (2017) Pose-driven deep convolutional model for person re-identification. IEEE International Conference on Computer Vision, 3980–3989
[60] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324
[61] X. Zhang, X. Luo, W. Xiang, Y. Sun, Q. Xiao, Q. Jiang, C. Zhang, J. Sun (2017) Alignedreid: Surpassing human-level performance in person re-identification. arXiv: 1711.08184
[62] Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S. (2018) Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). European Conference on Computer Vision, 480-496
[63] L. He , J. Liang , H. Li , Z. Sun (2018) Deep spatial feature reconstruction for partial person re-identification: alignment-free approach. IEEE Conference on Computer Vision and Pattern Recognition, 7073–7082
[64] Zhang, Z., Si, T., and Liu, S. (2018) Integration convolutional neural network for person re-identification in camera networks. IEEE Access, 6, 36887-36896
[65] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, W. Jiang, (2018) Scpnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. arXiv:1810.06996
[66] Varior, R. R., Shuai, B., Lu, J., Xu, D., and Wang, G. (2016) A siamese long short-term memory architecture for human re-identification. European Conference on Computer Vision, 135-153
[67] Xiao, Q., Cao, K., Chen, H., Peng, F., and Zhang, C. (2016) Cross domain knowledge transfer for person re-identification. arXiv:1611.06026.
[68] Yang, F., Yan, K., Lu, S., Jia, H., Xie, X., and Gao, W. (2019) Attention driven person re-identification. Pattern Recognition, 86, 143-155