This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2873599, IEEE Transactions on Circuits and Systems for Video Technology

Pedestrian Alignment Network for Large-scale Person Re-identification

Zhedong Zheng, Liang Zheng, Yi Yang

***Abstract*—Person re-identification (re-ID) is mostly viewed as an image retrieval problem. This task aims to search a query person in a large image pool. In practice, person re-ID usually adopts automatic detectors to obtain cropped pedestrian images. However, this process suffers from two types of detector errors: excessive background and part missing. Both errors deteriorate the quality of pedestrian alignment and may compromise pedestrian matching due to the position and scale variances. To address the misalignment problem, we propose that alignment be learned from an identification procedure. We introduce the pedestrian alignment network (PAN) which allows discriminative embedding learning pedestrian alignment without extra annotations. We observe that when the convolutional neural network (CNN) learns to discriminate between different identities, the learned feature maps usually exhibit strong activations on the human body rather than the background. The proposed network thus takes advantage of this attention mechanism to adaptively locate and align pedestrians within a bounding box. Visual examples show that pedestrians are better aligned with PAN. Experiments on three large-scale re-ID datasets confirm that PAN improves the discriminative ability of the feature embeddings and yields competitive accuracy with the state-of-the-art methods.**

***Index Terms*—Person Re-identification, Person Search, Person Alignment, Image Retrieval, Deep Learning**

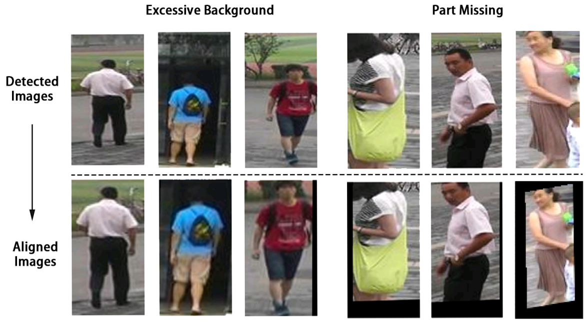


Fig. 1: Sample images influenced by detector errors (the first row) which are aligned by the proposed method (the second row). Two types of errors are shown: excessive background and part missing. We show that the pedestrian alignment network (PAN) corrects the misalignment problem by 1) removing extra background or 2) padding zeros. In implementation, we pad zeros to the feature maps. For visualization, we pad zeros to original images. The aligned images thus benefit the subsequent matching step.

I. Introduction

Person re-identification (person re-ID) aims at spotting the target person in different cameras, and is mostly viewed as an image retrieval problem, *i.e.,* searching for the query person in a large image pool. The recent progress mainly consists in the discriminatively learned embeddings using the convolutional neural network (CNN) on large-scale datasets. The learned embeddings extracted from the fine-tuned CNNs are shown to outperform the hand-crafted features [1]–[4].

Among many influencing factors, misalignment is a critical one in person re-ID. This problem arises due to the usage of pedestrian detectors. In realistic settings, the hand-drawn bounding boxes, existing in some previous datasets such as VIPER [5], CUHK01 [6] and CUHK02 [7], are infeasible to acquire when millions of bounding boxes are to be generated. So recent large-scale benchmarks such as CUHK03 [8], Market1501 [9] and MARS [10] adopt the Deformable Part Model (DPM) [11] to automatically detect pedestrians. This pipeline largely saves the amount of labeling effort and is closer to realistic settings. However, when detectors are used, detection errors are inevitable, which may lead to two common noisy factors: excessive background and part missing. For the former, background may take up a large proportion of a detected image. For the latter, a detected image may contain only part of the human body (see Fig. 1).

Pedestrian alignment and re-identification are two connected problems. When we have identity labels of pedestrian bounding boxes, we might be able to find optimal affine transformation that contains the most informative visual cues to discriminate between different identities. With affine transformation, pedestrians can be better aligned. Furthermore, with superior alignment, more discriminative features can be learned, and the pedestrian matching accuracy can in turn be improved.

Motivated by the above-mentioned aspects, we propose to incorporate pedestrian alignment into a re-identification architecture, yielding the pedestrian alignment network (PAN). Given a detected image, this network simultaneously learns to re-localize the person and categorize the person into predefined classes. Therefore, PAN takes advantage of the complementary nature of pedestrian alignment and person re-ID.

In a nutshell, the training process of PAN is composed of the following components: 1) a network to predict the identity of an input image, 2) an affine transformation to be estimated which re-localizes the input image, and 3) another network to predict the identity of the re-localized image. For components 1) and 3), we use two convolutional branches called the

Zhedong Zheng, and Yi Yang are with Centre for Artificial Intelligence, University of Technology Sydney, NSW 2007, Australia. (e-mail: zdzheng12@gmail.com, yee.i.yang@gmail.com) Liang Zheng is with the Research School of Computer Science, Australian National University. (e-mail:liangzheng06@gmail.com) We thank the support of Data to Decisions Cooperative Research Centre (www.d2dcrc.com.au), Google Faculty Research Award and NVIDIA Corporation with the donation of TITAN X (Pascal) GPU.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2018.2873599, IEEE Transactions on Circuits and Systems for Video Technology

base branch and alignment branch, to respectively predict the identity of the original image and the aligned image. Internally, they share the low-level features and during testing are concatenated at the FC layer to generate the pedestrian descriptor. In component 2), the affine parameters are estimated using the feature maps from the high-level convolutional layer of the base branch. The affine transformation is later applied on the lower-level feature maps of the base branch. In this step, we deploy a differentiable localization network: spatial transformer network (STN) [12]. With STN, we can 1) crop the detected images which may contain too much background or 2) pad zeros to the borders of feature maps with missing parts. As a result, we reduce the impact of scale and location variances caused by misdetection and thus make pedestrian matching more precise.

Note that our method addresses the misalignment problem caused by detection errors, while the commonly used patch matching strategy aims to discover matched local structures in well-aligned images. For methods that use patch matching, it is assumed that the matched local structures locate in the same horizontal stripe [8], [13]–[17] or square neighborhood [18]. Therefore, these algorithms are robust to some small spatial variance, *e.g.,* position and scale. However, when misdetection happens, due to the limitation of search scope, this type of methods may fail to discover matched structures, and raise the risk of part mismatching. Therefore, regarding the problem to be solved, the proposed method is significantly different from this line of works [8], [13]–[18]. We speculate that our method is a good complementary step for those using part matching.

Our contributions are summarized as follows:

- We propose the pedestrian alignment network (PAN), which simultaneously aligns pedestrians within images and learns pedestrian descriptors. It addresses the misdetection problem and person re-ID together, and improves the person re-ID accuracy;
- We conduct extensive experiments to validate the performance of the proposed network, and achieve competitive accuracy compared to the state-of-the-art methods on three large-scale person re-ID datasets (Market-1501 [9], CUHK03-NP [19] and DukeMTMC-reID [20]).

II. Related work

A. Hand-crafted Systems for re-ID

Person re-ID needs to find the robust and discriminative features among different cameras. Several pioneering approaches have explored the person re-ID by extracting local hand-crafted features such as LBP [21], Gabor [22] and LOMO [16]. In a series of works by [14], [15], the 32-dim LAB color histogram and the 128-dim SIFT descriptor are extracted from each 10×10 patches. In [9], [23], Zheng *et al.* use color name descriptor for each local patch and aggregate them into a global vector through the Bag-of-Words model. Yang *et al.* use Gaussian of Gaussian feature [24] to conduct semi-supervised learning [25]. Differently, Cheng *et al.* [26] localize the parts first and calculate color histograms for part-to-part correspondences. This line of works is beneficial from the local invariance in different viewpoints.

Besides the robust feature, metric learning is nontrivial for person re-ID. Kostinger *et al.* [27] propose “KISSME” based on Mahalanobis distance and formulate the pair comparison as a log-likelihood ratio test. Further, Liao *et al.* [16] extend the Bayesian face and KISSME to learn a discriminant subspace with a metric. Aside from the methods using Mahalanobis distance, Prosser *et al.* apply a set of weak RankSVMs to assemble a strong ranker [22]. Wang *et al.* transform the feature description from characteristic vector to discrepancy matrix [28] and apply the cross-view consistency [29] .

B. Deeply learned Models for re-ID

CNN-based deep learning models have been popular since [30] won ILSVRC’12 by a large margin. Yi *et al.* [13] split a pedestrian image into three horizontal parts and respectively trained three part-CNNs to extract features. Similarly, Cheng *et al.* [17] split the convolutional map into four parts and fuse the part features with the global feature. Li *et al.* [8] add a new layer that multiplies the activation of two images in different horizontal stripes. They use this layer to allow patch matching in CNN explicitly. Later, Ahmed *et al.* [18] improve the performance by proposing a new part-matching layer that compares the activation of two images in neighboring pixels. Lin *et al.* [31], [32] propose a boosting-based approach to learn a correspondence structure, which indicates the patch-wise matching probabilities between images from a target camera pair. Besides, Varior *et al.* [33], [34] combine CNN with some gate functions, similar to long-short-term memory (LSTM [35]) in spirit, which aims to focus on the similar parts of input image pairs adaptively. But it is limited by the computational inefficiency because the input should be in pairs. Similarly, Liu *et al.* [36] propose a soft attention-based model to focus on parts and combine CNN with LSTM components selectively; its limitation also consists of the computation inefficiency. Recently, He *et al.* [37] propose a fully convolutional network to conduct the feature reconstruction. They also calculate the pair-wise similarity. If the number of candidate images is large, it may take a long time to calculate the similarity of each pair.

Moreover, a convolutional network has the high discriminative ability by itself without explicit patch-matching. For person re-ID, Zheng *et al.* [38] directly use a conventional fine-tuning approach on Market-1501 [9] and their performance outperform other recent results. Wang *et al.* [39] propose an adaptive margin loss to solve the data imbalance. Wu *et al.* [40] combine the CNN embedding with hand-crafted features. Wu *et al.* [41] and Qian *et al.* [42] deepen the network and use filters of smaller size. Lin *et al.* [43] use person attributes as auxiliary tasks to learn more information. Zheng *et al.* [44] propose combining the identification model with the verification model and improve the fine-tuned CNN performance. Ding *et al.* [45] and Hermans *et al.* [46] use triplet samples for training the network which considers the images from the same people and the different people at the same time. Recent work by Zheng *et al.* combined original training dataset with GAN-generated images and regularized the model [47].

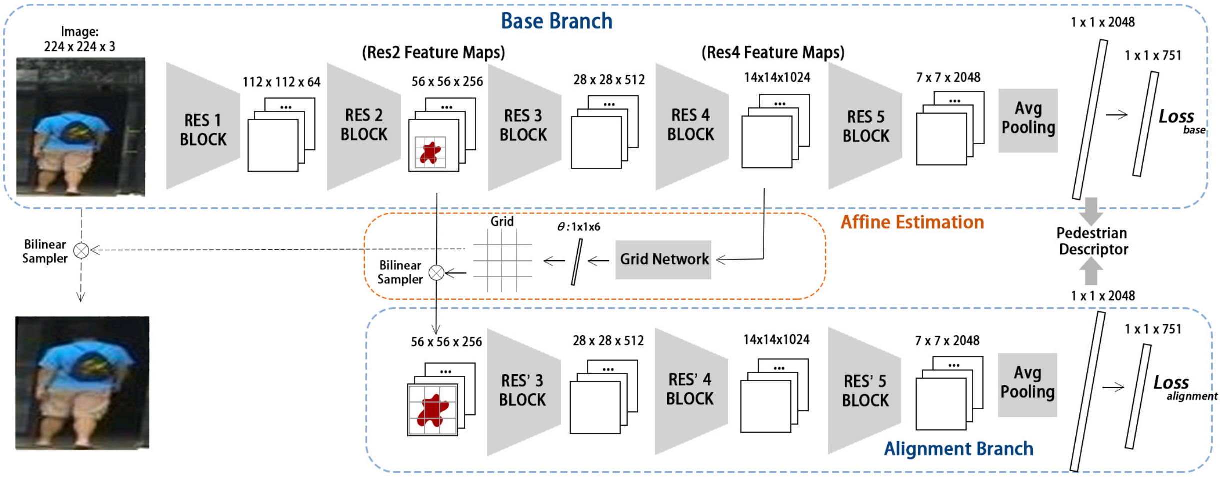


Fig. 2: Architecture of the pedestrian alignment network (PAN). It consists of two identification networks and an affine estimation network. We denote the Res-Block in the base branch as Res$i$, and denote the Res-Block in the alignment branch as Res'i. The base branch predicts the identities from the original image. We use the high-level feature maps of the base branch (Res4 Feature Maps) to predict the grid. Then the grid is applied to the low-level feature maps (Res2 Feature Maps) to re-localize the pedestrian (red star). The alignment stream then receives the aligned feature maps to identify the person again. Note that we do not perform alignment on the original images (dotted arrow) as previously done in [12] but directly on the feature maps. In the training phase, the model minimizes two identification losses. In the test phase, we concatenate the two $1 \times 1 \times 2048$ FC embeddings to form a 4096-dim pedestrian descriptor for retrieval.

C. *Objective Alignment*

Face alignment (here refer to the rectification of face misdetection) has been widely studied. Huang *et al.* [48] propose an unsupervised method called funneled image to align faces according to the distribution of other images and improve this method with convolutional RBM descriptor later [49]. However, it is not trained in an end-to-end manner, and thus following tasks *i.e.,* face recognition take limited benefits from the alignment. On the other hand, several works introduce attention models for task-driven object localization. Jadeburg *et al.* [12] deploy the spatial transformer network (STN) to fine-grained bird recognition and house number recognition. Johnson *et al.* [50] combine faster-RCNN [51], RNN and STN to address the localization and description in image caption. Aside from using STN, Liu *et al.* use reinforcement learning to detect parts and assemble a strong model for fine-grained recognition [52].

In person re-ID, there are several works using body patch matching [53], [54]. The work that inspires us the most is "PoseBox" proposed by [53]. The PoseBox is a strengthened version of the Pictorial Structures proposed in [26]. PoseBox is similar to our work in that 1) both works aim to solve the misalignment problem, and that 2) the networks have two convolutional streams. Nevertheless, our work differs significantly from PoseBox in two aspects. First, PoseBox employs the convolutional pose machines (CPM) to generate body parts for alignment in advance, while this work learns pedestrian alignment in an end-to-end manner without extra steps. Second, PoseBox can tackle the problem of excessive background but may be less effective when some parts are missing, because CPM fails to detect body joints when the body part is absent. However, our method automatically provides solutions to both problems, *i.e.,* excessive background and part missing.

III. Pedestrian Alignment Network

A. *Overview of PAN*

Our goal is to design an architecture that jointly aligns the images and identifies the person. The primary challenge is to develop a model that supports end-to-end training and benefits from the two inter-connected tasks. The proposed architecture draws on two convolutional branches and one affine estimation branch to simultaneously address these design constraints. Fig. 2 briefly illustrates our model. To better illustrate our method, we use the ResNet-50 model [55] as the base model which is applied on the Market-1501 dataset [9]. Each Res$i, i = 1, 2, 3, 4, 5$ block in Fig. 2 denotes several convolutional layers with batch normalization, ReLU, and optionally max pooling. We denote the Res-Block in the base branch as Res$i$, and denote the Res-Block in the alignment branch as Res'i. After each block, the feature maps are down-sampled to be half of the size of the feature maps in the previous block.

B. *Base and Alignment Branches*

There are two main convolutional branches exist in our model, called the base branch and the alignment branch. Both branches are classification networks that predict the identity of the training images. Given an originally detected image, the base branch not only learns to distinguish its identity from the others but also encodes the appearance of the detected image and provides the clues for the spatial localization (see Fig. 3). The alignment branch shares a similar convolutional network but processes the aligned feature maps produced by the affine estimation branch.

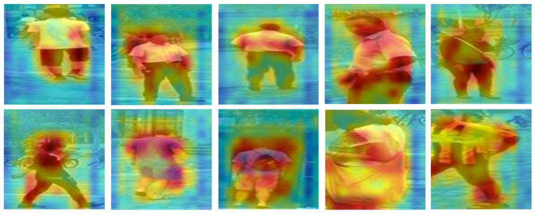


Fig. 3: We visualize the Res4 Feature Maps. High responses are mostly concentrated on the pedestrian body.

In the base branch, we train the ResNet-50 model [55], which consists of five down-sampling blocks and one global average pooling. We deploy the model pre-trained on ImageNet [56] and remove the final fully-connected (FC) layer. There are $K = 751$ identities in the Market-1501 training set, so we add a FC layer to map the CNN embedding of size $1 \times 1 \times 2048$ to 751 unnormalized probabilities. The alignment branch, on the other hand, is comprised of three ResBlocks and one average pooling layer. We also add a FC layer to predict the multi-class probabilities. The two branches do not share weight. We use W_1 and W_2 to denote the parameters of the two convolutional branches, respectively.

More formally, given an input image x, we use $p(k|x)$ to denote the probability that the image x belongs to the class $k \in \{1...K\}$. Specifically, $p(k|x) = \frac{exp(z_k)}{\sum_{k=1}^{K} exp(z_i)}$. Here z_i is the outputted probability from the CNN model. For the two branches, the cross-entropy losses are formulated as:

$$l_{base}(W_1, x, y) = -\sum_{k=1}^{K} (log(p(k|x))q(k|x)), \quad (1)$$

$$l_{align}(W_2, x_a, y) = -\sum_{k=1}^{K} (log(p(k|x_a))q(k|x_a)), \quad (2)$$

where x_a denotes the aligned input. It can be derived from the original input $x_a = T(x)$. Given the label y, ground-truth distribution $q(y|x) = 1$ and $q(k|x) = 0$ for all $k \neq y$. If we discard zero terms in Eq. 1 and Eq. 2, losses are equivalent to:

$$l_{base}(W_1, x, y) = -log(p(y|x)), \quad (3)$$

$$l_{align}(W_2, x_a, y) = -log(p(y|x_a)). \quad (4)$$

At each iteration, we wish to minimize the total entropy, which equals to maximizing the possibility of the correct prediction.

C. Affine Estimation Branch

To address the problems of excessive background and part missing, the key idea is to predict the position of the pedestrian and do the corresponding spatial transformer. When excessive background exists, a cropping strategy should be used; under part missing, we need to pad zeros to the corresponding feature map borders. Both strategies need to find the parameters for the affine transformation. In this paper, this function is implemented by the affine estimation branch.

The affine estimation branch receives two input tensors of activations $14 \times 14 \times 1024$ and $56 \times 56 \times 256$ from the base branch. We name the two tensors the Res2 Feature Maps and the Res4 Feature Maps, respectively. The Res2 Feature Maps contain shallow feature maps of the original image and reflects the local pattern information. On the other hand, since the Res4 Feature Maps are closer to the classification layer, it encodes the attention on the pedestrian and semantic cues for aiding identification. The affine estimation branch contains one bilinear sampler and one small network called Grid Network. The Grid Network contains one ResBlock and one average pooling layer. We pass the Res4 Feature Maps through Grid Network to regress a set of 6-dimensional transformer parameters. The learned transforming parameters θ are used to produce the image grid. The affine transformation process can be written as below,

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}, \quad (5)$$

where (x_i^t, y_i^t) are the target coordinates on the output feature map, and (x_i^s, y_i^s) are the source coordinates on the input feature maps (Res2 Feature Maps). $\theta_{11}, \theta_{12}, \theta_{21}$ and θ_{22} deal with the scale and rotation transformation, while θ_{13} and θ_{23} deal with the offset. In this paper, we set the coordinates as: (-1,-1) refer to the pixel on the top left of the image, while (1,1) refer to the bottom right pixel. We use a bilinear sampler to make up the missing pixels, and we assign zeros to the pixels located out of the original range. So we obtain an injective function from the original feature map V to the aligned output U. More formally, we can formulate the equation:

$$U_{(m,n)}^c = \sum_{x^s}^{H} \sum_{y^s}^{W} V_{(x^s,y^s)}^c max(0, 1 - |x^t - m|) max(0, 1 - |y^t - n|). \quad (6)$$

$U_{(m,n)}^c$ is the output feature map at location (m, n) in channel c, and $V_{(x^s,y^s)}^c$ is the input feature map at location (x_s, y_s) in channel c. If (x_t, y_t) is close to (m, n), we add the pixel at (x_s, y_s) according to the bilinear sampling.

In this work, we do not perform pedestrian alignment on the original image; instead, we choose to achieve an equivalent function on the shallow feature maps. By using the feature maps, we reduce the running time and parameters of the model. This explains why we apply re-localization grid on the feature maps. The bilinear sampler receives the grid and the feature maps to produce the aligned output x_a. We provide some visualization examples in Fig. 3. Res4 Feature maps are shown. Through ID supervision, we are able to re-localize the pedestrian and correct misdetections to some extent.

D. Pedestrian Descriptor

Given the fine-tuned PAN model and an input image x_i, the pedestrian descriptor is the weighted fusion of the FC features of the base branch and the alignment branch. That is, we are able to capture the pedestrian characteristic from the original image and the aligned image. In the Section IV-C, the experiment shows that the two features are complementary to each other and improve the person re-ID performance.

Batchsize	Dropout rate	Rank@1	mAP
16	0.75	**80.17**	**59.14**
32	0.75	76.13	55.22
64	0.75	73.04	51.63
16	0.75	**80.17**	**59.14**
16	0.5	78.15	56.28
16	0.25	78.09	56.25

TABLE I: Sensitivity of the baseline method towards batchsize and dropout rate on Market-1501.

In this paper, we adopt a straightforward late fusion strategy, *i.e.,* $f_i = g(f_i^1, f_i^2)$. Here f_i^1 and f_i^2 are the FC descriptors from two types of images, respectively. We reshape the tensor after the final average pooling to a 1-dim vector as the pedestrian descriptor of either branch. The pedestrian descriptor is then represented as:

$$f_i = \left[\alpha|f_i^1|^{\mathrm{T}}, (1-\alpha)|f_i^2|^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{7}$$

The $|\cdot|$ operator denotes an l^2-normalization step. We concatenate the aligned descriptor with the original descriptor, both after l^2-normalization. α is the weight for the two descriptors. If not specified, we simply use $\alpha = 0.5$ in our experiments.

E. Re-ranking for re-ID

We can obtain the rank list $N(q,n) = [x_1, x_2, ...x_n]$ by sorting the Euclidean distance of gallery images to the query. Distance is calculated as $D_{i,j} = (f_i - f_j)^2$, where f_i and f_j are l^2-normalized features of image i and j, respectively. We then perform re-ranking to obtain better retrieval results. Several post-processing methods can be applied in person re-ID [19], [57]–[62]. Specifically, we adopt the re-ranking method [19] to further improve performance. In [19], the query is augmented by other unlabeled queries and highly ranked unlabeled images. However, these unlabeled images may also suffer from the mis-alignment problem, which compromises the feature fusion process. In this work, our method does not change the re-ranking process. Instead, our method extract features that reflect the aligned quality of these unlabeled images. With better features, re-ranking is more effective.

IV. Experiments

A. Datasets

Market-1501 is a large-scale person re-ID dataset, which contains 19,732 gallery images, 3,368 query images and 12,936 training images collected from six cameras. There are 751 identities in the training set and 750 identities in the testing set without overlapping. Every identity in the training set has 17.2 photos on average. All images are automatically detected by the DPM detector [11]. The misalignment problem is common, and the dataset is closer to the realistic settings.

CUHK03-NP contains 14,097 images of 1,467 identities [8]. We follow the new training/testing protocol proposed in [19] to evaluate the re-ID performance, which is more challenging. First, the new protocol has a larger candidate pool which has 5,332 images of 700 identities. In comparison, the original protocol only has 100 images of 100 identities. Second, the new protocol has a smaller training set (767 identities) while the original protocol has 1467 identities. In the "detected" set, all the bounding boxes are produced by DPM; in the "labeled" set, the images are all hand-drawn. In this paper, we evaluate our method on "detected" and "labeled" sets. If not specified, "CUHK03-NP" denotes the detected set.

DukeMTMC-reID is a subset of DukeMTMC [20] and contains 36,411 images of 1,812 identities shot by 8 cameras. The pedestrian images are cropped manually. The dataset consists 16,522 training images of 702 identities, 2,228 query images of the other 702 identities and 17,661 gallery images. We follow the evaluation protocol in [47].

B. Implementation Details

ConvNet. We first fine-tune the base branch on the person re-ID datasets. Then, the base branch is fixed while we fine-tune the whole network. In fine-tuning the base branch, the learning rate decreases from 10^{-3} to 10^{-4} after 30 epochs. We stop training at the 40th epoch. When we train the whole model, the learning rate also decreases from 10^{-3} to 10^{-4} after 30 epochs. Our implementation is based on the Matconvnet [73] package. The input images are uniformly resized to the size of 224×224. We perform simple data augmentation such as cropping and horizontal flipping following [44]. **STN.** For the affine estimation branch, the network may fall into a local minimum in early iterations. To stabilize training, a small learning rate is useful. We, therefore, use a learning rate of 1×10^{-5} for the final convolutional layer in the affine estimation branch. In addition, we set the all $\theta = 0$ except that $\theta_{11}, \theta_{22} = 0.8$ and thus, the alignment branch starts training from looking at the center part of the Res2 Feature Maps.

C. Evaluation

Evaluation of the ResNet baseline. We implement the baseline according to the routine proposed in [1], with the details specified in Section IV-B. We report our baseline results in Table II. The Rank@1 accuracy is 80.17%, 30.50%, 31.14% and 65.22% on Market1501, CUHK03-NP (detected), CUHK03-NP (labeled) and DukeMTMC-reID respectively. The baseline model is on par with the network in [1], [44]. In our recent implementation, we use a smaller batch size of 16 and a dropout rate of 0.75. We obtain a higher baseline Rank@1 accuracy 80.17% on Market-1501 than 73.69% in the [1], [44]. A baseline ablation study is in Table I. The reason for improvement might be due to dataset scale. When using a large batch size, CNN training might be prone to the overfitting problem. So large dropout rate and small batchsize helps. A similar observation has been reported in [75]. For a fair comparison, we present the results of our methods built on this new baseline. Note that this baseline result itself is higher than many previous works [34], [44], [53], [68].

Base branch vs. alignment branch To investigate how alignment helps to learn discriminative pedestrian representations, we evaluate the Pool5 descriptors of the base branch and the alignment branch, respectively.

Methods	dim	*Market-1501*				*CUHK03-NP (detected)*				*CUHK03-NP (labeled)*				*DukeMTMC-reID*			
		1	5	20	mAP	1	5	20	mAP	1	5	20	mAP	1	5	20	mAP
Base	2,048	80.17	91.69	96.59	59.14	30.50	51.07	71.64	29.04	31.14	52.00	74.21	29.80	65.22	79.13	87.75	44.99
Alignment	2,048	79.01	90.86	96.14	58.27	34.14	54.50	72.71	31.71	35.29	53.64	72.43	32.90	68.36	81.37	88.64	47.14
PAN	4,096	82.81	93.53	97.06	63.35	36.29	55.50	75.07	34.00	36.86	56.86	75.14	35.03	71.59	83.89	90.62	51.51

TABLE II: Comparison of different methods on Market-1501, CUHK03-NP (detected), CUHK03-NP (labeled) and DukeMTMC-reID. Rank@1, 5, 20 accuracy (%) and mAP (%) are shown. Note that the base branch is the same as the classification baseline [1]. We observe consistent improvement of our method over the individual branches on the three datasets.

Methods	Backbone	Single Query		Multi. Query	
		Rank@1	mAP	Rank@1	mAP
BoW+kissme [9]	-	44.42	20.76	-	-
MST CNN [63]	*	45.1	-	55.4	-
FisherNet [64]	*	48.15	29.94	-	-
SL [65]	-	51.90	26.35	-	-
S-LSTM [33]	*	-	-	61.6	35.3
DNS [66]	-	55.43	29.87	71.56	46.03
Gated Reid [34]	*	65.88	39.55	76.04	48.45
CADL [67]	*	73.84	47.11	80.85	55.58
SOMAnet [68]	*	73.87	47.89	81.29	56.98
PIE [53]	Res50	78.65	53.87	-	-
Verif.-Identif. [44]	Res50	79.51	59.87	85.84	70.33
DCF [69]	*	80.31	57.53	86.79	66.70
DPR [70]	GoogLe	81.0	63.4	-	-
SSM [71]	Res50	82.21	68.80	88.18	76.18
SVDNet [72]	Res50	82.3	62.1	-	-
GAN+re-rank [47]	Res50	83.97	66.07	88.42	76.10
Basel.	Res50	80.17	59.14	87.41	72.05
Ours	Res50	82.81	63.35	88.18	71.72
Ours+re-rank	Res50	85.78	76.56	89.79	83.79
Ours (GAN)	Res50	86.67	69.33	90.88	76.32
Ours (GAN)+re-rank	Res50	**88.57**	**81.53**	**91.45**	**87.44**

TABLE III: Rank@1 accuracy (%) and mAP (%) on Market-1501. - the respective papers use hand-crafted feature, * the respective papers use their own specific network.

Methods	Back bone	Detected		Labeled	
		Rank@1	mAP	Rank@1	mAP
BOW+XQDA [9]	-	6.36	6.39	7.93	7.29
IDE [1]	Res50	21.3	19.7	22.2	21.0
IDE +DaF [74]	Res50	26.4	30.0	27.5	31.5
IDE+XQDA [19]	Res50	31.1	28.2	32.0	29.6
IDE+XQDA+re-rank [19]	Res50	34.7	37.4	38.1	40.3
Basel.	Res50	30.5	29.0	31.1	29.8
Ours	Res50	36.3	34.0	36.9	35.0
Ours+re-rank	Res50	**41.9**	**43.8**	**43.9**	**45.8**

TABLE IV: Rank@1 accuracy (%) and mAP (%) on CUHK03-NP. We evaluate the proposed method on the "detected" and "labeled" subsets according to the new protocol in [19]. - the respective papers use hand-crafted feature.

Methods	Backbone	Rank@1	mAP
BoW+kissme [9]	-	25.13	12.17
LOMO+XQDA [16]	-	30.75	17.04
Gan [47]	Res50	67.68	47.13
Verif.-Identif. [44]	Res50	68.9	49.3
APR [43]	Res50	70.69	51.88
Basel. [47]	Res50	65.22	44.99
Ours	Res50	71.59	51.51
Ours+re-rank	Res50	**75.94**	**66.74**

TABLE V: Rank@1 accuracy (%) and mAP (%) on DukeMTMC-reID. We follow the evaluation protocol in [47].- the respective papers use hand-crafted feature.

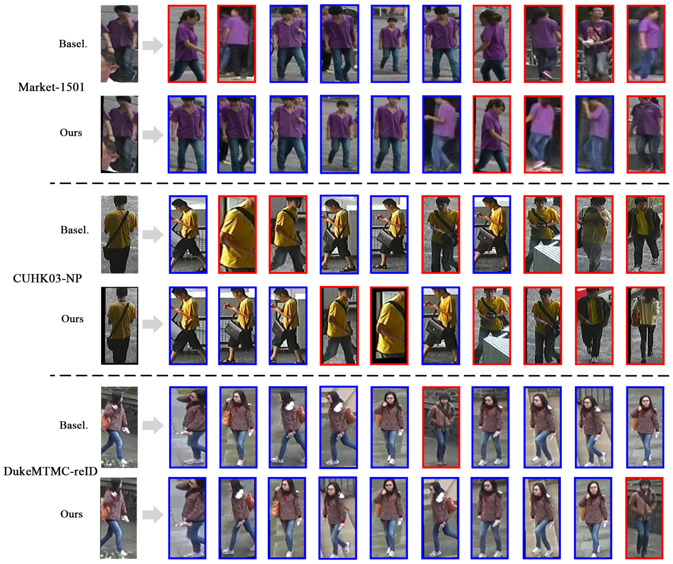


Fig. 4: Sample retrieval results on three datasets. The images in the first column are queries. The retrieved images are sorted according to the similarity score from left to right. The first row shows the result of baseline [1], and the second row denotes the results of PAN. Correct and false matches are in the blue and red rectangles, respectively.

First, as shown in Table II, the alignment branch yields higher performance *i.e.,* 3.64% / 4.15% on the two dataset settings (CUHK03-NP detected/labeled) and 3.14% on DukeMTMC-reID, and achieves a very similar result with the base branch on Market-1501. We speculate that Market-1501 contains more intensive detection errors than the other three datasets and thus, the effect of alignment is limited.

Second, though the CUHK03-NP (labeled) dataset and the

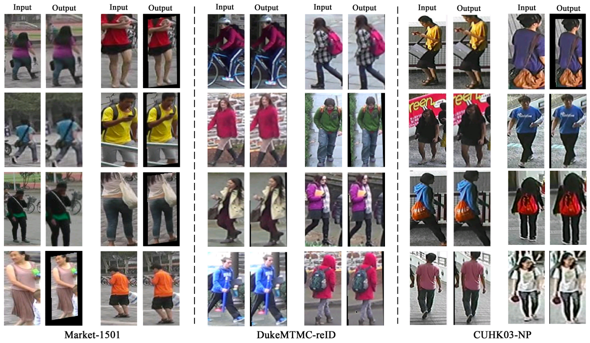


Fig. 5: Examples of pedestrian images before and after alignment on three datasets. Pairs of input images and aligned images are shown.

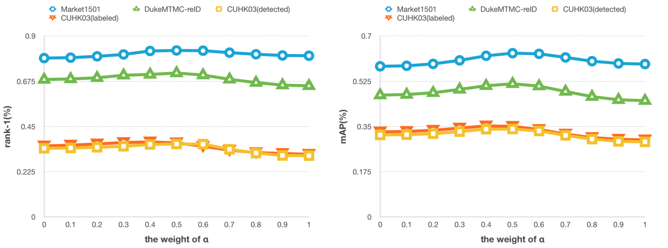


Fig. 6: Sensitivity of person re-ID accuracy to parameter α. Rank@1 accuracy(%) and mAP(%) are shown.

DukeMTMC-reID dataset are manually annotated, the alignment branch still improves the performance of the base branch. This observation demonstrates that the manual annotations may not be good enough for the machine to learn a good descriptor. In this scenario, alignment is non-trivial and makes the pedestrian representation more discriminative.

The complementary of the two branches. As mentioned, the two descriptors capture the different pedestrian characteristic from the original image and the aligned image. The results are summarized in Table II. We observe a constant improvement on the three datasets when we concatenate the two descriptors. The fused descriptor improves 2.64%, 2.15%, 1.63% and 3.23% on Market-1501, CUHK03-NP(detected), CUHK03-NP(labeled) and DukeMTMC-reID, respectively. The two branches are complementary and thus, contain more meaningful information than a separate branch. Aside from the improvement of the accuracy, this simple fusion is efficient sine it does not introduce additional computation.

Parameter sensitivity. We evaluate the sensitivity of the person re-ID accuracy to the parameter α. As shown in Fig. 6, we report the Rank@1 accuracy and mAP when tuning the α from 0 to 1. The change of Rank@1 accuracy and mAP are relatively small corresponding to the α. Our reported result simply use $\alpha = 0.5$. $\alpha = 0.5$ may not be the best choice for a particular dataset. But if we do not foreknow the distribution of the dataset, it is a simple and straightforward choice.

Comparison with the state-of-the-art methods. We compare our method with the state-of-the-art methods on Market-1501, CUHK03-NP and DukeMTMC-reID in Table III, Table IV and Table V, respectively. On Market-1501, we achieve Rank@1 accuracy = 85.78%, mAP = 76.56% after re-ranking, which is the best result compared to the published paper. Our model is also adaptive to previous models. One of the previous best results is based on the model regularized by GAN [47]. We combine the model trained on GAN generated images and thus, achieve the state-of-the-art result Rank@1 accuracy = 88.57%, mAP = 81.53% on Market-1501. On CUHK03-NP, we arrive at a competitive result Rank@1 accuracy = 36.3%, mAP=34.0% on the detected dataset and Rank@1 accuracy = 36.9%, mAP = 35.0% on the labeled dataset. On DukeMTMC-reID, we also observe a state-of-the-art result Rank@1 accuracy = 75.94% and mAP = 66.74% after re-ranking. Despite the visual disparities among the three datasets, *i.e.,* scene variance and detection bias, we show that our method consistently improves the re-ID performance.

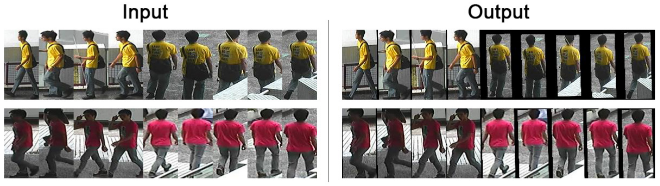


Fig. 7: Alignment results on CUHK03 (labeled). The original images, which are manually annotated, still contain the position/scale bias.

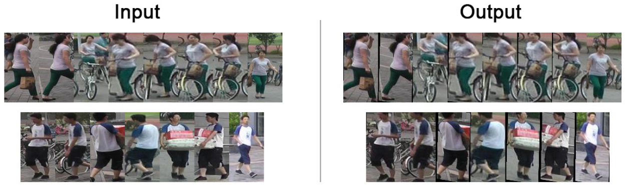


Fig. 8: Alignment results on occluded pedestrians. Our alignment is robust to the such level of occlusion, and still predicts reasonable alignment.

Visualization of the alignment. We further visualize the aligned images in Fig. 5, Fig. 7 and Fig. 8. As aforementioned, the proposed network does not process the alignment on the original image. To visualize the aligned images, we extract the predicted affine parameters and then apply the affine transformation on the originally detected images manually. **We observe that the network does not perform perfect alignment as the human, but it more or less reduces the scale and position variance, which is critical for the network to learn the representations.** As shown in Fig. 8, our method is robust to some occlusion, and still predicts reasonable alignment. So the proposed network improves the performance of the person re-ID.

V. Conclusion

Pedestrian alignment and re-identification are two inner-connected problems, which inspires us to develop an attention-based system. In this work, we propose the pedestrian alignment network (PAN), which simultaneously aligns the pedestrians within bounding boxes and learns the pedestrian descriptors. Experiments on three datasets indicate that our method is competitive with the state-of-the-art methods.

References

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

[2] L. Zheng, S. Wang, J. Wang, and Q. Tian, "Accurate image search with multi-scale contextual evidences," *International Journal of Computer Vision*, vol. 120, no. 1, pp. 1–13, 2016.

[3] Z. Ma, X. Chang, Y. Yang, N. Sebe, and A. G. Hauptmann, "The many shades of negativity," *IEEE Trans Multimedia*, vol. 19, no. 7, pp. 1558–1568, 2017.

[4] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016.

[5] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *PETS*, vol. 3, no. 5. Citeseer, 2007.

[6] W. Li, R. Zhao, and X. Wang, "Human reidentification with transferred metric learning," in *ACCV*, 2012.

[7] W. Li and X. Wang, "Locally aligned feature transforms across views," in *CVPR*, 2013.

[8] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014.

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015.

[10] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.

[11] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *CVPR*, 2008.

[12] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *NIPS*, 2015.

[13] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014.

[14] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *ICCV*, 2013.

[15] R. Zhao, W. Ouyang, and Wang, "Learning mid-level filters for person re-identification," in *CVPR*, 2014.

[16] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015.

[17] 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 *CVPR*, 2016.

[18] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015.

[19] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017.

[20] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *ECCVW*, 2016.

[21] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *CVPR*, 2012.

[22] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in *BMVC*, 2010.

[23] L. Zheng, S. Wang, and Q. Tian, "Coupled binary embedding for large-scale image retrieval," *IEEE transactions on image processing*, vol. 23, no. 8, pp. 3368–3380, 2014.

[24] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016, pp. 1363–1372.

[25] X. Yang, M. Wang, R. Hong, Q. Tian, and Y. Rui, "Enhancing person re-identification in a self-trained subspace," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2017.

[26] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *BMVC*, 2011.

[27] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012.

[28] Z. Wang, R. Hu, C. Chen, Y. Yu, J. Jiang, C. Liang, and S. Satoh, "Person reidentification via discrepancy matrix and matrix metric," *IEEE transactions on cybernetics*, 2017.

[29] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, "Zero-shot person re-identification via cross-view consistency," *TMM*, 2016.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[31] W. Lin, Y. Shen, J. Yan, M. Xu, J. Wu, J. Wang, and K. Lu, "Learning correspondence structures for person re-identification," *TIP*, 2017.

[32] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," in *ICCV*, 2015.

[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 *ECCV*, 2016.

[34] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *ECCV*, 2016.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[36] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *TIP*, 2016.

[37] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *CVPR*, 2018.

[38] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian, "Person re-identification in the wild," *arXiv:1604.02531*, 2016.

[39] J. Wang, Z. Wang, C. Gao, N. Sang, and R. Huang, "Deeplist: Learning deep features with adaptive listwise constraint for person reidentification." *TCSVT*, 2017.

[40] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *WACV*, 2016.

[41] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv:1601.07255*, 2016.

[42] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," *CVPR*, 2017.

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

[44] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, p. 13, 2017.

[45] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

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

[47] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," *ICCV*, 2017.

[48] G. B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *ICCV*, 2007.

[49] G. Huang, M. Mattar, H. Lee, and E. G. Learned-Miller, "Learning to align from scratch," in *NIPS*, 2012.

[50] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *CVPR*, 2016.

[51] R. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[52] X. Liu, T. Xia, J. Wang, Y. Yang, F. Zhou, and Y. Lin, "Fully convolutional attention networks for fine-grained recognition," *arXiv:1611.05244*, 2016.

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

[54] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[57] M. Ye, C. Liang, Z. Wang, Q. Leng, and J. Chen, "Ranking optimization for person re-identification via similarity and dissimilarity," in *ACM Multimedia*, 2015.

[58] Y. Lin, Z. Zheng, H. Zhang, C. Gao, and Y. Yang, "Bayesian query expansion for multi-camera person re-identification," *Pattern Recognition Letters*, 2018.

[59] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *CVPR*, 2011.

[60] S. Bai and X. Bai, "Sparse contextual activation for efficient visual re-ranking," *TIP*, 2016.

[61] S. Bai, X. Bai, Q. Tian, and L. J. Latecki, "Regularized diffusion process on bidirectional context for object retrieval," *TPAMI*, 2018.

[62] S. Bai, Z. Zhou, J. Wang, X. Bai, L. J. Latecki, and Q. Tian, "Ensemble diffusion for retrieval," in *ICCV*, 2017.

[63] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei, "Multi-scale triplet cnn for person re-identification," in *ACM Multimedia*, 2016.

[64] L. Wu, C. Shen, and A. v. d. Hengel, "Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification," *Pattern Recognition*, 2016.

[65] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *CVPR*, 2016.

[66] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," *CVPR*, 2016.

[67] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *CVPR*, 2017.

[68] 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:1701.03153*, 2017.

[69] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *CVPR*, 2017.

[70] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *ICCV*, 2017.

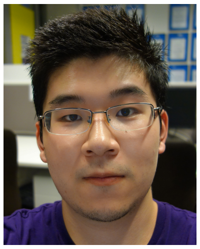
[71] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *CVPR*, 2017.

[72] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," *ICCV*, 2017.

[73] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," in *ACM Multimedia*, 2015.

[74] R. Yu, Z. Zhou, S. Bai, and X. Bai, "Divide and fuse: A re-ranking approach for person re-identification," in *BMVC*, 2017.

[75] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *ICLR*, 2017.



Zhedong Zheng received the B.S. degree in Fudan University, China, in 2016. He is currently a Ph.D. student in University of Technology Sydney, Australia. His research interests include image retrieval and person re-identification.

Liang Zheng is a Lecturer and a Computer Science Futures Fellow in the Research School of Computer Science, Australian National University. He received the Ph.D degree in Electronic Engineering from Tsinghua University, China, in 2015, and the B.E. degree in Life Science from Tsinghua University, China, in 2010. He was a postdoc researcher in the Center for Artificial Intelligence, University of Technology Sydney, Australia. His research interests include image retrieval, and person reidentification.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, China, in 2010. He is currently a professor with University of Technology Sydney, Australia. He was a Post-Doctoral Research with the School of Computer Science, Carnegie Mellon University, USA. His current research interest includes machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis and video semantics understanding.