Semantics-Aware Deep Correspondence Structure Learning for Robust Person Re-identification

Yaqing Zhang, Xi Li*, Liming Zhao, Zhongfei Zhang
Zhejiang University, Hangzhou, China
{yaqing,xilizju,zhaoliming,zhongfei}@zju.edu.cn

Abstract

In this paper, we propose an end-to-end deep correspondence structure learning (DCSL) approach to address the cross-camera person-matching problem in the person re-identification task. The proposed DCSL approach captures the intrinsic structural information on persons by learning a semantics-aware image representation based on convolutional neural networks, which adaptively learns discriminative features for person identification. Furthermore, the proposed DCSL approach seeks to adaptively learn a hierarchical data-driven feature matching function which outputs the matching correspondence results between the learned semantics-aware image representations for a person pair. Finally, we set up a unified end-to-end deep learning scheme to jointly optimize the processes of semantics-aware image representation learning and cross-person correspondence structure learning, leading to more reliable and robust person re-identification results in complicated scenarios. Experimental results on several benchmark datasets demonstrate the effectiveness of our approach against the state-of-the-art approaches.

1 Introduction

Person re-identification (Re-ID) is a task of matching persons captured from non-overlapping camera views, which has important applications in video surveillance systems including tracking in the camera network, human retrieval and threat detection.

In computer vision, the task of person Re-ID focuses on constructing a robust feature space for describing the person in the image, so as to minimize the intra-personal variance while maximizing the inter-personal margin. Many existing efforts employ low-level appearance feature representations, including color, texture, shape, to describe the appearance of each person. However, there exist a variety of appearance which lead to instability of these features. One of the major challenging problems is the misalignment caused by viewpoint changes. For example, in Figure 1, the yellow dashed bounding boxes of the same person captured by two camera views appear to be quite different since they respectively show the front and back views. Such variations can enlarge the intra-personal difference which the traditional feature representations fail to cope with.

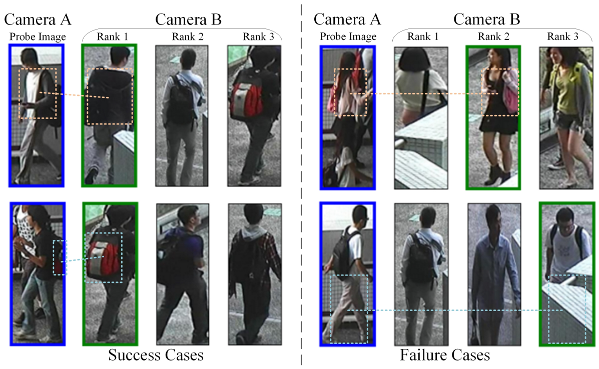


Figure 1: Examples of person Re-ID. Given the probe image of a person in camera A marked by a blue window, the task is to find the same person in the gallery set of camera B. The figure shows top-3 of our predicted results in CUHK03 dataset, with the groundtruth images marked by the green bounding boxes. The dashed bounding boxes illustrate several difficulties in person Re-ID.

Inspired by the perception of humans, the images of persons can be composed of a set of latent semantic units (e.g., head, front and back upper body, belongings), and the spatial configurations of each person are intrinsic even if the viewpoint changes. To identify the correct person given the probe image, we first want to recognize these semantic components according to the inherent structure representations and correspond them to those of the images from other camera views, respectively. Based on such intuition, we are able to analyse the person by a semantic-based perspective, which is robust to misalignment and variations.

With such structure representations, the remaining task of person re-identification is to match these components of the image pair across camera views. Many methods focus on building the correspondence structure in pursuit of establishing spatial correlations between images, which can be achieved by one-to-one matching strategy or a pre-defined

*Corresponding author

matching function [Li *et al.*, 2014; Ahmed *et al.*, 2015; Shen *et al.*, 2015]. However, such matching strategies may either be of low computing efficiency or have limitations by ignoring the spatial structures. From our point of view, matching is supposed to be a data-driven approach that proper matching function will automatically select the critical components and their feature representations of each image and compare them respectively while ignoring the misaligned or missing components (e.g., the front and back of body in Figure 1).

Thus the robust person Re-ID mainly focuses on the following two aspects: (1) Find the semantic components to better describe the appearance and the spatial structure of each person. (2) Match the representations of the image pair using a learned function to adaptively discover the correspondence between the images. Based on the aforementioned observations and motivations, we propose a novel and unified approach called **Deep Correspondence Structure Learning (DCSL)** as shown in Figure 2 to deal with the challenges in person Re-ID. In our approach, we try to capture the intrinsic structural information of persons by learning a semantics-aware image representation, and then we match each component of the structure using a learned hierarchical matching strategy. We unify our approach into an end-to-end deep learning framework to learn feature representation without any heuristic motivations. In the experiments, we employ mixture models for cross-domain person Re-ID learning to further enhance the robustness of the person Re-ID system for cross-dataset evaluation.

To summarize, we have several contributions as follows.

- We propose an end-to-end deep learning framework to deal with the challenges in person Re-ID task. The proposed approach enables the feature representation of an image to keep stable under lots of variations. Furthermore, it also seeks to adaptively learn a data-driven feature matching function with efficient hierarchical matching strategy. The learned image and correspondence representations are robust to be adapted to other camera settings or scenarios.
- The proposed DCSL approach employs a unified framework to learn the semantics-aware image representation and adaptive correspondence representation jointly by a data-driven approach. The framework enables both representations to be regularized from the person Re-ID task to find the most discriminative representations.

2 Related Work

In the proposed work, we mainly aim at establishing semantics-aware correspondence between images captured from non-overlapping cameras. In the literature, previous efforts of person Re-ID mainly focus on the following two points: feature extraction for image representation and image matching based on the extracted features.

Early papers pay attention to effective feature representations for better modelling the person in the image. People employ hand-crafted features for detecting determinative information including HSV color histogram [Farenzena *et al.*, 2010], SIFT [Zhao *et al.*, 2013], LBP features [Li and Wang, 2013] and the combination of them. Unlike hand-crafted features, the methods based on deep learning [Li *et al.*, 2014; Zhao *et al.*, 2014a] learn feature representation directly from tasks which have shown significant improvement compared with traditional features .

With the extracted features, many existing efforts are devoted to handling the overall appearance variations between the given images [Paisitkriangkrai *et al.*, 2015; Hirzer *et al.*, 2012]. In order to learn the spatial relationship between images from different views, people utilize location-related matching strategies to find higher-level correlation between the images. Among these efforts, patch-based matching methods [Li *et al.*, 2014; Shen *et al.*, 2015; Ahmed *et al.*, 2015] decompose images into patches and perform patch-wise matching strategy to find the spatial relationship. To reduce the computing cost due to the comparisons among the patch compositions, some saliency based methods are proposed to guide spatial matching methods [Zhao *et al.*, 2014b]. Both patch-based and saliency-based methods capture the appearance correlations such as color or texture. Compositional approaches [Xu *et al.*, 2013] first localize the body parts and search for part-to-part correspondence between reference samples and observations, which have promising results on the challenging datasets. These methods depend on the performance of body parser, which are limited in some specific scenes.

3 Our Approach

Before describing this work, we first introduce a number of notations used hereinafter. For notational convenience, we let an uppercase boldfaced letter denote a feature map blob, and a lowercase boldfaced letter denote a feature vector. Specifically, for a given feature map blob $\mathbf{X}$, $\mathbf{x}_{ij}$ is associated with the feature vector at the (i, j)-th location of $\mathbf{X}$, while x_{ijk} stands for the k-th element of $\mathbf{x}_{ij}$. Namely, $\mathbf{X}$ is generated from a 2D spatial stack of $\mathbf{x}_{ij}$ along the two spatial dimensions of $\mathbf{X}$. To distinguish feature maps or vectors from different views, we employ the superscript letters for notational representations (e.g., $\mathbf{X}^A$ and $\mathbf{X}^B$ respectively for camera views A and B). Moreover, a general function is written in the form of $f(\mathbf{X}_1, \mathbf{X}_2, ...; \boldsymbol{\theta})$ with $\{\mathbf{X}_i\}$ and $\boldsymbol{\theta}$ respectively being the input variables and the parameters to be learned.

3.1 Problem Formulation

In the beginning, we collect the person Re-ID dataset $\{(\mathbf{X}_n, l_n), n = 1, ..., N\}$, where $\mathbf{X}_n = (\mathbf{I}^A, \mathbf{I}^B)_n$ denotes the paired images of the n-th sample and $l_n \in \{0, 1\}$ denotes its label, and here $l_n = 1$ indicates that the pair belongs to the same person. In our approach, the task of person Re-ID is to learn a classifier $s(\mathbf{X}_n; \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$ to distinguish whether the image pair $\mathbf{X}_n$ represents the same person or not by measuring the similarities between them.

In order to perform the robust similarity measurement, the proposed approach involves two factors: (1) image representation of each image by a semantics-aware feature extraction approach, (2) image matching based on the representations of the image pair to adaptively discover the correspondence of them and measure the similarity guided by the correspondence.

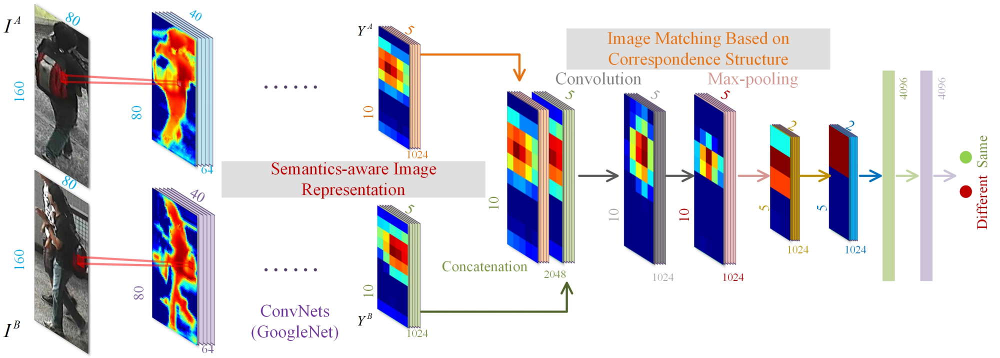


Figure 2: Architecture of Deep Correspondence Structure Learning (DCSL). The input of the network is the paired images of size 160 × 80 × 3. The network consists of two kernel procedures: (1) semantics-aware image representation to represent the semantic components of each image by modified GoogLeNet, (2) hierarchical image matching using multi-layer CNNs based on the representations of shape 10 × 5 × 1024 of each image. Finally we will compute the final decision of whether the image pair belongs to the same person or not by Softmax activations.

3.2 Semantics-aware Image Representation

In person Re-ID, image representation is confronted with a number of challenges, such as occlusion and viewpoint variations. As discussed previously, there exists inherent semantic structure in an image, and such structure is composed of intrinsic latent components such as upper body and head, which are robust to the variations of views and background. Based on the semantic representations of the image, we will be able to capture the correlations of the image pair.

To effectively encode these components in an adaptive manner, we adopt convolutional neural networks (CNNs) to model and abstract the multi-scale semantic information for a hierarchical representation parameterized by $\boldsymbol{\theta}_1$ to represent the image:

$$\mathbf{Y} = f_{\text{CNN}}(\mathbf{I}; \boldsymbol{\theta}_1) \tag{1}$$

As observed in [Zeiler and Fergus, 2014], these representations contain both low-level visual information and high-level semantic components of the image. Without loss of generality, we decompose the representations to $\mathbf{Y} = \{\mathbf{V}, \mathbf{C}\}$, where $\mathbf{V}$ denotes the low-level representations and $\mathbf{C}$ stands for the representation of semantic components. The representation of each component is associated with a subset of feature maps in $\mathbf{C}$, and here we assign $\mathbf{E}$ for representing one specific component.

3.3 Image Matching via Correspondence Structure Learning

Given paired images from different camera views, we extract their corresponding feature representations by Eq. (1), written as $\mathbf{Y}^A = \{\mathbf{V}^A, \mathbf{C}^A\}$ and $\mathbf{Y}^B = \{\mathbf{V}^B, \mathbf{C}^B\}$. With $\mathbf{Y}^A$ and $\mathbf{Y}^B$, the problem of image matching is converted to the task of how to correspond the components by their representation maps and to perform the similarity measurement based on the correspondence representation. First we will discuss the correspondence structure learning of the specific component representations $\mathbf{E}^A$ and $\mathbf{E}^B$.

Correspondence structure learning based on $\mathbf{E}^A$ and $\mathbf{E}^B$. We first define the correspondence structure based on the semantic components of the person to cope with the misalignment in person Re-ID.

In our approach the correspondence structure $\mathbf{E}^A \rightarrow \mathbf{E}^B$ encodes the spatial correspondence distributions between the specific component of the image pair.

Many methods adopt a discrete distribution as the correspondence structure [Shen *et al.*, 2015], which is a set of pair-wise matching probabilities: $\{\mathbf{P}(\mathbf{e}^A_{ij}, \mathbf{E}^B)\}$, where $\mathbf{P}(\mathbf{e}^A_{ij}, \mathbf{E}^B)$ describes the correspondence distribution in $\mathbf{E}^B$ for the feature vector $\mathbf{e}^A_{ij}$ in $\mathbf{E}^A$. The distribution $\mathbf{P}(\mathbf{e}^A_{ij}, \mathbf{E}^B)$ can be obtained through different strategies. For example, (1) compute an optimal assignment matrix from $\mathbf{e}^A_{ij}$ to every vector $\{\mathbf{e}^B_{st}\}$ in $\mathbf{E}^B$ by dense matching; (2) conduct local search in a small neighbourhood of the given location. For (1), one-to-one location-specific matching strategy searches in the whole space of possible solutions, resulting in low computational efficiency. Moreover, it may induce additional matching noise since it lacks the capability of modelling the spatial neighbourhood matching consistency of each image; for (2), local spatial neighbourhood matching is often trapped in local correspondence optimums as a result of heuristically predetermined searching ranges.

To deal with these difficulties, we construct the correspondence structure from a different perspective, which aims to adaptively capture the correspondence relationships between the multi-scale semantic component representations in a hierarchical matching fashion. As shown in Figure 3, the correspondence structure is discovered through a pyramid-like matching strategy. We generate the correspondence maps between the feature maps of each image. The value of each location (i, j) indicates the correspondence probability at that location. To deal with misalignment, we further down-sample the feature maps by max-pooling so as to preserve the most discriminative information of the component and align it in a larger region.

In this case, we denote the set of multi-level feature maps as $\{\mathbf{E}^{A,t}\}$ and $\{\mathbf{E}^{B,t}\}$ generated from $\mathbf{E}^A$ and $\mathbf{E}^B$ by max-

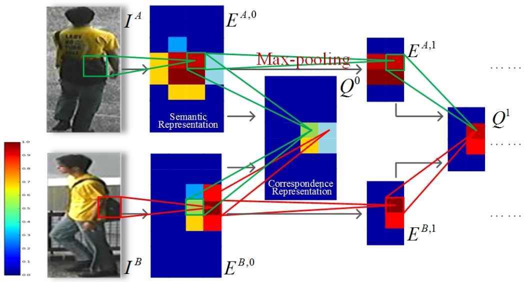


Figure 3: Illustration of correspondence structure learning by pyramid matching. The representation of component "shoulder bag" is represented by $\{E^{A,t}\}$ and $\{E^{B,t}\}$ respectively. The red and green bounding boxes indicate that the bag appears at a different location and misaligns in the first correspondence map $\mathbf{Q}^0$. We solve the problem by max-pooling for choosing the most relevant pairs.

pooling using the stride of 2:

$$e^t_{ijk} = \begin{cases} e_{ijk}, & t = 0 \\ \max\limits_{0\le u,v<k_p} e^{t-1}_{2i+u,2j+v,k}, & t = 1,2,3,... \end{cases} \tag{2}$$

where k_p denotes the kernel size in the pooling stage, and t is the level of max-pooling.

The final correspondence representation can be defined as a set of hierarchical correspondence maps $\{\mathbf{Q}^t\}$, where $\mathbf{Q}^t$ denotes the correspondence distribution at level t. Inspired by [Han *et al.*, 2015], we calculate it by learning the interactions between all the feature vectors $\mathbf{e}^{A,t}_{ij}$ and $\mathbf{e}^{B,t}_{ij}$ using the learned weights $\boldsymbol{\theta}^t_2$:

$$\mathbf{Q}^t = f_{\text{CNN}}([\mathbf{E}^{A,t}, \mathbf{E}^{B,t}]; \boldsymbol{\theta}^t_2) \tag{3}$$

where $[\mathbf{E}^{A,t}, \mathbf{E}^{B,t}]$ is the concatenation of the two feature maps, and the set $\{\mathbf{Q}\}^t$ will be augmented by introducing more semantic-level components from $\mathbf{C}^A$ and $\mathbf{C}^B$.

To summarize, the correspondence structure learning by pooling strategy has two advantages in learning context similarities: 1) it effectively improves the cross-view matching robustness by using a hierarchical multi-stage matching strategy; 2) it makes full use of multi-scale spatial matching consistency by multi-level max pooling. The above two parts are modelled in a totally data-driven learning scheme, resulting in the flexibility in practice.

Similarity measurement based on $\mathbf{V}^A$ and $\mathbf{V}^B$. Before matching two images by their correspondence representations, we have to first measure the similarity given the low-level feature vectors of the selected pairs. We first construct the hierarchical feature representations $\{\mathbf{V}^{A,t}\}$ and $\{\mathbf{V}^{B,t}\}$, which are built from $\mathbf{V}^{A,t}$ and $\mathbf{V}^{B,t}$ in the same way as that of $\{\mathbf{E}^{A,t}\}$ and $\{\mathbf{E}^{B,t}\}$. Then the similarity representations $\{\mathbf{S}^t\}$ upon the feature representations learn the interactions between the vectors by the weights $\boldsymbol{\theta}^t_3$:

$$\mathbf{S}^t = f_{\text{CNN}}([\mathbf{V}^{A,t}, \mathbf{V}^{B,t}]; \boldsymbol{\theta}^t_3) \tag{4}$$

Correlation of the images based on $\{\mathbf{Q}^t\}$ and $\{\mathbf{S}^t\}$. Correlation of the paired images is characterized by evaluating the similarities of the mutually corresponding semantic components after correspondence structure learning. It is formulated as a multi-level matching function with $\boldsymbol{\theta}_4$:

$$s = f_{\text{CNN}}\left([\{\mathbf{S}^t\}, \{\mathbf{Q}^t\}]; \boldsymbol{\theta}_4\right) \tag{5}$$

3.4 Unified Correspondence Structure Learning

In principle, image representations and correspondence structure learning are two correlated and complementary problems. Namely, better semantic representations contribute to guiding correspondence network to find more meaningful structures. Meanwhile, better correspondence structures regularize feature representation network to learn more effective features for matching.

Thus we develop a unified end-to-end data-driven framework, where the feature, correspondence and metric representations are learned jointly and adaptively in a supervised setting using the structure illustrated in Figure 2. We train the network by minimizing the cross-entropy error:

$$E_\theta = -\frac{1}{N}\sum_{n=1}^{N}\left[l_n \log \hat{l}_n + (1-l_n)\log(1-\hat{l}_n)\right] \tag{6}$$

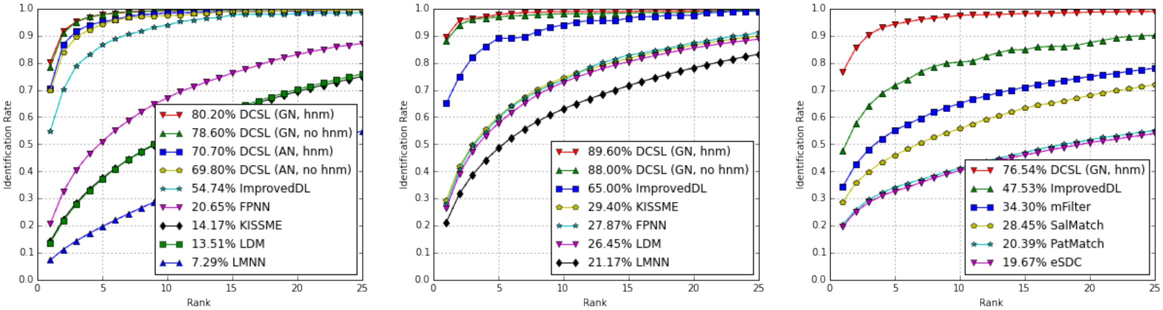
over a training set of N pairs using stochastic gradient descent. l_n is the 1/0 label for the input pair $\mathbf{X}_n$ while $\hat{l}_n$ and $(1-\hat{l}_n)$ is the Softmax activations computed on the two output nodes of the similarity function $s(\mathbf{X}_n; \boldsymbol{\theta})$, namely $s_0(\mathbf{X}_n; \boldsymbol{\theta})$ and $s_1(\mathbf{X}_n; \boldsymbol{\theta})$:

$$\hat{l}_n = \frac{\exp(s_1(\mathbf{X}_n; \boldsymbol{\theta}))}{\exp(s_0(\mathbf{X}_n; \boldsymbol{\theta})) + \exp(s_1(\mathbf{X}_n; \boldsymbol{\theta}))} \tag{7}$$

The proposed approach is finally simplified to the unified framework by synthesizing Eq. (1) for semantics-aware image representation and Eqs. (3)-(5) for correspondence-based image matching, which is defined as the deep correspondence network (DCN) with $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \{\boldsymbol{\theta}^t_2\}, \{\boldsymbol{\theta}^t_3\}, \boldsymbol{\theta}_4\}$:

$$\begin{aligned} & s(\mathbf{X}_n; \boldsymbol{\theta}) \\ =& f_{\text{CNN}}\left([\{\mathbf{S}^t\}, \{\mathbf{Q}^t\}]; \boldsymbol{\theta}_4\right) \\ =& f_{\text{CNN}}\left([\mathbf{Y}^A, \mathbf{Y}^B]; \boldsymbol{\theta}_4, \left\{\boldsymbol{\theta}^t_3\right\}, \left\{\boldsymbol{\theta}^t_2\right\}\right) \\ =& f_{\text{CNN}}\left([f_{\text{CNN}}(\mathbf{I}^A; \boldsymbol{\theta}_1), f_{\text{CNN}}(\mathbf{I}^B; \boldsymbol{\theta}_1)]; \boldsymbol{\theta}_4, \left\{\boldsymbol{\theta}^t_3\right\}, \left\{\boldsymbol{\theta}^t_2\right\}\right). \end{aligned} \tag{8}$$

In summary, the deep correspondence structure learning is performed to capture the semantics-aware components automatically from the person Re-ID task. We employ convolutional neural networks for robust semantic representations. The concatenation of feature maps captures correspondence between images from different camera views and also enables part-wise matching by the correspondence structures, together with the efficient hierarchical alignment strategy by max-pooling operations. The models of the two tasks are learned jointly by combining the appearance and semantic representation with the correspondence structures to find the optimal representation of the input image pair to be matched.



(a) CUHK03 labelled (100 test IDs)

(b) CUHK01 (100 test IDs)

(c) CUHK01 (486 test IDs)

Figure 4: CMC curves of intra dataset experiments on CUHK03 labelled and CUHK01 dataset. Rank-1 identification rates are shown in the figure beside the method legend. In the legends of DCSL, "GN" means that we use GoogLeNet for feature extraction of each image while "AN" stands for AlexNet. We also evaluate the effect of hard negative mining ("hnm") by training with or without hard negative mining.

4 Experimental Results

We have implemented our architecture using Caffe [Jia *et al.*, 2014], which is a widely used deep learning framework in recent years. We use an NVIDIA TITAN X GPU for training the networks. It takes around 30 hours for training one DCSL model and 1.7ms for computing the similarity given a pair of images.

In this section, we compare our approach with the state-of-the-art approaches on several datasets. To avoid accidental results, experiments are conducted on the datasets with 10 random training and testing splits. We evaluate all the approaches with Cumulative Matching Characteristics (CMC) curves by single-shot results. The CMC curve characterize a ranking result for every image in the gallery given the probe image. We first report results of our proposed model on three challenging datasets with the same evaluation criteria. Finally, we report cross domain transfer learning on the trained networks with cross dataset evaluation.

4.1 Dataset

We evaluate our method on three public datasets, namely CUHK03(labelled) [Li *et al.*, 2014], CUHK01 [Li *et al.*, 2012], VIPeR [Gray *et al.*, 2007]. Table 1 shows the description of each dataset and our experimental settings with the training and testing splits.

Table 1: Datasets and settings in our experiments.

Dataset	**CUHK03**	**CUHK01**	**VIPeR**
# identities	1360	971	632
# images	13,164	3,884	1,264
# cam./ ID	2	2	2
# train IDs	1,160	871;485	316
# test IDs	100	100;486	316

Among the three datasets, CUHK03 is a relatively large dataset, which can cover a number of variations to construct the spatial configurations and also contain several challenging situations including illumination changes, cross-view deformations, occlusions, etc.

In addition, we also conduct cross-dataset experiments to evaluate the performance on CUHK01 and VIPeR using the trained model on CUHK03 and Market-1501 [Zheng *et al.*, 2015].

4.2 Training the Network

In this section, we use pre-trained AlexNet [Krizhevsky *et al.*, 2012] and GoogLeNet [Szegedy *et al.*, 2015] models as our deep architecture for extracting the feature maps as shown in Figure 2. The parameters of these parts are initialized with the model trained on ImageNet-1K [Russakovsky *et al.*, 2014]. In our experiments, we use "norm2" maps of AlexNet and "inception_5b/output" maps of GoogLeNet to represent semantics-aware information of each image. To improve the ability of spatial representations, we remove "pool4/3x3_s2" layer of GoogLeNet. As a result, given the input shape of $160 \times 80 \times 3$, the shapes of feature maps which we obtained are $19 \times 9 \times 256$ in AlexNet and $10 \times 5 \times 1024$ in GoogLeNet. These feature maps are directly fed to the concatenation layer to learn the correspondence of the image pair.

In our method, we use stochastic gradient descent for updating the weights of the network. Training pairs are divided into mini-batches. For training efficiency, we set the base learning rate as $\eta^{(0)} = 0.01$ and use a polynomial decay for learning rate to train around 100,000 batches of size 100. We use a momentum of $\mu = 0.9$ and weight decay $\lambda = 0.0002$.

Data Augmentation. To generate more training pairs, we also apply the data augmentation strategy for training the neural network. We randomly perform affine transformation over the shape of the original images and thus generate 5 augmented images per image for training.

Hard Negative Mining. There still exist some scenarios that are hard to distinguish, which require more negative samples for training. To increase the efficiency of model training, we use the trained model to classify all the negative pairs and choose the top ranked ones for retraining our network.

Mixture of the DCSL networks. To deal with the dataset biases when transferring the trained model to real-world applications, We employ a mixture of DCSL networks to represent the domains of different distributions inspired by [Ge

Table 2: Top recognition rate (%) of the various methods over CUHK03 dataset with 100 test IDs, CUHK01 dataset with 100 test IDs, CUHK01 dataset with 486 test IDs with rank = 1, 5, 10. We use the benchmarks provided by [Ahmed *et al.*, 2015] and some results on CUHK01 with 486 test IDs and VIPeR are not reported.

Methods	CUHK03 (100 test IDs)			CUHK01 (100 test IDs)			CUHK01 (486 test IDs)			VIPeR (316 test IDs)		
	r = 1	r = 5	r = 10	r = 1	r = 5	r = 10	r = 1	r = 5	r = 10	r = 1	r = 5	r = 10
DCSL	**80.20**	**97.73**	**99.17**	**89.60**	**97.80**	**98.90**	**76.54**	**94.24**	**97.49**	**44.62**	**73.42**	**82.59**
ImprovedDL	54.74	86.50	93.88	65.00	89.00	94.00	47.53	71.60	80.25	34.81	63.61	75.63
KISSME	14.17	37.46	52.20	29.40	60.18	74.44	-	-	-	19.60	48.00	62.20
FPNN	20.65	50.94	67.01	27.87	59.64	73.53	-	-	-	-	-	-
LMNN	7.29	19.64	30.74	21.17	48.51	62.98	13.45	31.33	42.25	-	-	-

et al., 2015]. First we denote $\beta_k(\mathbf{X}_n;\boldsymbol{\theta}_e)$ as the probability of the image pair $\mathbf{X}_n$ belonging to the k-th component which can be calculated using parameters $\boldsymbol{\theta}_e$. The output score of the image pair is finally composed by all the K domain components using the weights $\{\beta_k\}_{k=1}^{K}$: $s(\mathbf{X}_n;\boldsymbol{\theta}) = \sum_{k=1}^{K} \beta_k s_{(k)}(\mathbf{X}_n;\boldsymbol{\theta}_{(k)}))$ where $s_{(k)}(\mathbf{X}_n;\boldsymbol{\theta}_{(k)})$ is the output score of the k-th DCSL model .

4.3 Experimental Results

We compare our deep correspondence structure learning (DCSL) method with several methods in recent years, including KISSME [Koestinger *et al.*, 2012], FPNN [Li *et al.*, 2014], ImprovedDL [Ahmed *et al.*, 2015], LDM [Guillaumin *et al.*, 2009], eSDC [Zhao *et al.*, 2013], LMNN [Weinberger *et al.*, 2005], mFilter [Zhao *et al.*, 2014a], SalMatch and PatMatch [Zhao *et al.*, 2013]. Figure 4 and Table 2 illustrates the recognition rate of these methods. Figure 1 shows a few qualitative results.

Comparisons on CUHK03 and CUHK01. From Figure 4 and Table 2, we see that DCSL outperforms the state-of-the-art methods by more than 20% on the performance of rank-1 accuracy (80.20% vs. 54.74%, 89.60% vs. 65.00%, 76.54% vs. 47.53%). Deep learning based methods outperform the traditional methods by its outstanding ability of data-driven feature learning. Note that the rank-5 recognition rate of DCSL reaches around 95%, meaning that the trained model has high probability of finding the correct person from other cameras given more than 100 candidate images.

Comparisons on VIPeR. The last column of Table 2 illustrates that DCSL is able to compete with the state-of-the-art methods even though the training samples are limited for deep neural networks. We have a relative better performance than deep learning based method ImprovedDL (44.62% vs. 34.81%), which employs the pre-trained model on the CUHK03 and CUHK01 dataset.

Effect of different feature extraction models. We both try AlexNet (AN) and GoogLeNet (GN) models for feature extraction with the same correspondence networks. Figure 4(a) illustrates that the performance of GoogLeNet exceeds that of AlexNet due to its outstanding ability to abstract the image features.

Effect of hard negative mining. Hard negative mining also has positive impact on the performance of our DCSL method by mining more training pairs for further tuning the model. We see the absolute gain over 1% compared with the same model without hard negative mining, as shown in Figure 4(a)(b).

Table 3: Top recognition rate (%) of the cross dataset experiments.

Test Set	Training Set	Model	r = 1	r = 5	r = 10
CUHK01	CUHK03	1-DCSL	64.22	86.69	93.31
	Market	1-DCSL	52.00	71.00	78.00
	CUHK03+Market	1-DCSL	69.14	85.71	91.57
	CUHK03+Market	m-DCSL	**72.48**	**88.64**	**94.31**
VIPeR	CUHK03	1-DCSL	14.24	34.18	45.25
	Market	1-DCSL	17.82	35.28	44.46
	CUHK03+Market	1-DCSL	21.89	39.24	49.68
	CUHK03+Market	m-DCSL	**23.40**	**42.11**	**55.80**

Results of cross-dataset evaluations. We both use the single DCSL model (1-DCSL) and a mixture of DCSL models (m-DCSL) for the cross-dataset evaluations. Table 3 illustrates that our proposed DCSL model is robust to be generalized to other target domains. With the mixture models of DCSL, we are able to deploy our proposed method to the real-world applications. Note that the rank-1 recognition rate of CUHK01 using the model trained by CUHK03 and Market-1501 can reach the score of 69.14%, which outperforms most of the state-of-the-art methods that conduct training on the CUHK01.

5 Conclusion

In this paper, we propose and evaluate a unified approach for person Re-ID that jointly learns a deep convolutional neural network for representing the semantics-aware information of the image as well as a network for robust image matching based on correspondence structure learning. Our model outperforms the state-of-the-art approaches by a large margin and has a considerable computing efficiency. This work demonstrates that unified deep neural networks can be effective for general person Re-ID and can be easily generalized through the mixture of the knowledge from different domains.

Acknowledgments

This work was supported in part by the National Basic Research Program of China under Grant 2012CB316400 and Grant 2015CB352300, National Natural Science Foundation of China under Grant U1509206 and Grant 61472353, Zhejiang Provincial Engineering Center on Media Data Cloud

Processing and Analysis, Fundamental Research Funds for the Central Universities, and equipment donation by Nvidia.

References

[Ahmed *et al.*, 2015] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015.

[Farenzena *et al.*, 2010] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367, 2010.

[Ge *et al.*, 2015] ZongYuan Ge, Alex Bewley, Christopher McCool, Ben Upcroft, Peter Corke, and Conrad Sanderson. Fine-grained classification via mixture of deep convolutional neural networks. *arXiv:1511.09209*, 2015.

[Gray *et al.*, 2007] Douglas Gray, Shane Brennan, and Hai Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3. Citeseer, 2007.

[Guillaumin *et al.*, 2009] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, pages 498–505, 2009.

[Han *et al.*, 2015] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR*, pages 3279–3286, 2015.

[Hirzer *et al.*, 2012] Martin Hirzer, Peter M Roth, Martin Köstinger, and Horst Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793. 2012.

[Jia *et al.*, 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678, 2014.

[Koestinger *et al.*, 2012] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, pages 2288–2295, 2012.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Li and Wang, 2013] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601, 2013.

[Li *et al.*, 2012] Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *ACCV*, pages 31–44, 2012.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[Paisitkriangkrai *et al.*, 2015] Sakrapee Paisitkriangkrai, Chunhua Shen, and Anton van den Hengel. Learning to rank in person re-identification with metric ensembles. In *CVPR*, pages 1846–1855, 2015.

[Russakovsky *et al.*, 2014] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, pages 1–42, 2014.

[Shen *et al.*, 2015] Yang Shen, Weiyao Lin, Junchi Yan, Mingliang Xu, Jianxin Wu, and Jingdong Wang. Person re-identification with correspondence structure learning. In *ICCV*, pages 3200–3208, 2015.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[Weinberger *et al.*, 2005] Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2005.

[Xu *et al.*, 2013] Yuanlu Xu, Liang Lin, Wei-Shi Zheng, and Xiaobai Liu. Human re-identification by matching compositional template with cluster sampling. In *ICCV*, pages 3152–3159, 2013.

[Zeiler and Fergus, 2014] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

[Zhao *et al.*, 2013] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013.

[Zhao *et al.*, 2014a] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Learning mid-level filters for person re-identification. In *CVPR*, pages 144–151, 2014.

[Zhao *et al.*, 2014b] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by saliency learning. *arXiv:1412.1908*, 2014.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.