Deep Learning Prototype Domains for Person Re-Identification

Arne Schumann
Fraunhofer IOSB, Karlsruhe, Germany
arne.schumann@iosb.fraunhofer.de

Shaogang Gong
Queen Mary University of London, UK
s.gong@qmul.ac.uk

Tobias Schuchert
Fraunhofer IOSB, Karlsruhe, Germany
tobias.schuchert@iosb.fraunhofer.de

Abstract

Person re-identification (re-id) is the task of matching multiple occurrences of the same person from different cameras, poses, lighting conditions, and a multitude of other factors which alter the visual appearance. Typically, this is achieved by learning either optimal features or matching metrics which are adapted to specific pairs of camera views dictated by the pairwise labelled training datasets. In this work, we formulate a deep learning based novel approach to automatic prototype-domain discovery for domain perceptive (adaptive) person re-id (rather than camera pair specific learning) for any camera views scalable to new unseen scenes without training data. We learn a separate re-id model for each of the discovered prototype-domains and during model deployment, use the person probe image to select automatically the model of the closest prototype-domain. Our approach requires neither supervised nor unsupervised domain adaptation learning, i.e. no data available from the target domains. We evaluate extensively our model under realistic re-id conditions using automatically detected bounding boxes with low-resolution and partial occlusion. We show that our approach outperforms most of the state-of-the-art supervised and unsupervised methods on the latest CUHK-SYSU and PRW benchmarks.

1. Introduction

The task of re-identifying the same person across different cameras has attracted much interest in recent years. Person re-identification is at its core a cross-domain recognition problem. Datasets are usually recorded in a camera network setting with a fixed set of cameras and viewing angles. Consequently, most approaches interpret each camera as a separate visual domain and focus on developing features or metrics that can robustly recognize a person within such *camera-view-perspective* domains. In this work, we
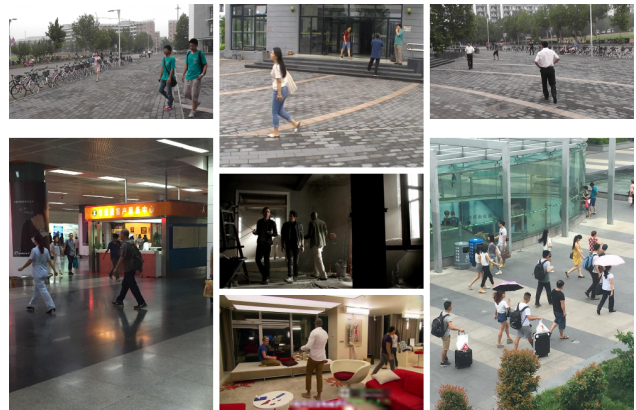

Figure 1. We use the latest PRW [37] (top row) and CUHK-SYSU [33] (bottom row) datasets as target (test) domains for evaluation, unavailable to model learning. Both datasets provide many camera views and unsegmented video frames which requires auto-person-detection for a more realistic person re-id evaluation.

consider other *camera-view-independent* factors, such as pose, illumination, occlusions, and background influence the visual appearance of a person, and we wish to explore them as visual domains in constructing *camera-view independent re-id* models for better scalability to unknown camera views.

In this work, we propose a two-stage approach to automatically discover visual domains in large amounts of diverse data and use them to learn feature embeddings for person re-identification (see Figure 1). In the first stage, we pool data from a large amount of re-identification datasets, independent from the test domains, to capture a large degree of visual variation in the training data. We then explore clustering based on feature learning in convolutional neural networks (CNNs) to automatically discover dominant (prototype) visual domains. In the second stage, we again apply CNNs to learn feature embeddings in each of the prototype

domains in order to support domain perceptive (sensitive) person re-id during testing with automatic domain selection. We learn one embedding per domain. This allows the model to learn specific details about each individual prototype domain while ignoring others. For example, an embedding learned for a domain which predominantly contains people of dark-dress does not need to encode information relevant to distinguishing a person dressed in light blue colors from a person dressed in white clothes. By doing so, the domain perceptive embedding focuses on learning subtle discriminative characteristics among similar visual appearances. On testing, a probe image is first matched to its most likely domain. Then, the feature embedding learned on that domain is used to perform re-identification. Note, this approach is purely *inductive*. It does not require any training data (labelled or unlabelled) from the target (test) domains, and the model is designed to scale to any new target domains. Our approach is particularly well suited to scenarios in which no fixed set of camera views is available (i.e. no fixed domain borders are specified). We thus evaluate it on the latest CUHK-SYSU and PRW datasets, which contain images from diverse sources of mobile cameras, movies and fixed view cameras, with multitude of view angles, backgrounds, resolutions and poses. Our approach yields the state-of-the-art accuracy on CUHK-SYSU and is competitive on PRW. This is *without* using target domain data in our model training whilst *all* other methods compared in the evaluation exploit target domain data in their model learning.

Our contributions are: **(1)** We formulate a novel approach to automatic discovery of prototype-domains, characterising person visual appearance with domain perceptive awareness. **(2)** We develop a deep learning model for domain perceptive (DLDP) selection and re-id matching in a single automatic process without any supervised nor unsupervised domain transfer learning. **(3)** We show the significant advantage of our model by outperforming the state-of-the-art on the CUHK-SYSU benchmark [33] with up to 5.6% at Rank-1 re-id, and being competitive on the PRW benchmark [37] of 45.4% Rank-1 re-id compared to the 47.7% state-of-the-art, notwithstanding that the latter benefited from model learning on target domain data.

2. Related Work

Most re-id approaches can be grouped into two categories: feature based approaches and metric based approaches. The former type aims to develop a robust feature representation. The latter approach focuses on optimizing a distance metric that, given any feature, yields small distances for matching person images and large distances of images of different people. In recent years, deep learning methods have gained significant advantages on image classification, and have been applied to person re-id. Many deep learning approaches focus on feature learning. Yi *et*

al. [35] split person images into three regions and learn separate feature maps which are combined into a final feature through a fully connected layer. Cosine distance is used to perform re-id. Ding *et al*. [7] apply a triplet loss in order to train a feature whose Euclidean distance of a matching image pair is smaller than that of a pair of images of different people. Xiao *et al*. [32] propose to use dataset specific dropout to learn features over multiple smaller datasets simultaneously. Cheng *et al*. [4] propose an improved triplet loss function which emphasises small distances for similar image pairs. We *et al*. [31] combine hand-crafted features with CNN features for re-id metric learning. Other deep learning approaches focus on studying network layers specifically designed for person re-id. Li *et al*. [18] describe a filter pairing network to model translation, occlusion and background clutter in its architecture. Ahmend *et al*. [1] introduce a neighborhood matching layer for improving robustness to translation and pose change. This layer is also applied by Wu *et al*. [30] to train an end-to-end re-id net which directly outputs a (dis-) similarity decision without relying on a separate distance function. Xiao *et al*. [33] propose an approach which combines person detection and re-id into a single CNN for simultaneous person detection and computing re-id feature for each detection.

A few studies have addressed cross-domain re-id by using target domain data for supervised [16, 23, 29] or unsupervised [22, 25] domain adaptation. Others have evaluated their models on datasets without any adaptation to the target domain [12, 24, 35]. To our knowledge, the proposed model in this work, for the first time, does not rely on domain adaptation using target domain data whilst learning domain perceptive re-id for unknown target domains.

3. Methodology

The central objective of our approach is to learn a domain adaptive re-id model (domain perceptive) which is scalable to new and unseen data without requiring any manual labelling for model training on the new target domains. We propose a two-stage approach to achieve this: (1) In the first stage, characteristic and dominant (prototype) domains are automatically discovered in large amounts of diverse data (Section 3.1.2); (2) In the second stage, this information is used to train a number of domain specific embeddings by deep learning for person re-id (Section 3.2.2). An overview of our approach is given in Figure 2.

3.1. Automatic Domain Discovery

3.1.1 Divergent Data Sampling

A key requirement for a meaningful domain discovery is *divergent data sampling* which aims to provide a large range of realistic visual variation. In order to achieve such a high degree of variation, we pool a number of publicly available
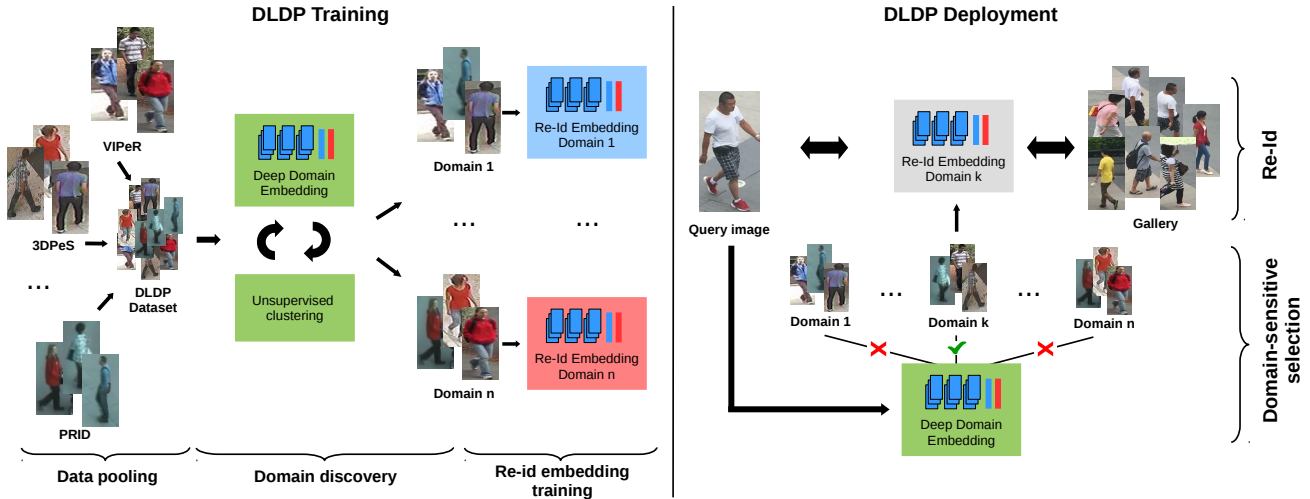
DLDP Training

DLDP Deployment

Figure 2. Overview of DLDP. During training we discover domains in a large pooled dataset. For each domain a domain-specific re-id model is trained. At deployment, the query image is used to identify the closest matching domain and use the corresponding domain-specific model to rank the gallery.

	Persons	Cameras	M-BBoxes	A-BBoxes
HDA [9]	85	13	850	-
GRID [21]	250	8	500	-
3DPeS [2]	200	8	1,012	-
CAVIAR4REID [5]	72	2	1,221	-
i-LIDS [38]	119	2	476	-
PRID [11]	200	2	400	-
VIPeR [10]	632	2	1,264	-
SARC3D [3]	50	1	200	-
CUHK2 [17]	1,816	10	7,264	-
CUHK3 [18]	1,360	6	14,096	14,097
Total	4,786	54	27,283	14,097

Table 1. Ten sources for the DLDP re-id domain discovery dataset. The data consists of manually labelled bounding boxes (M-BBoxes) as well as person detections (A-BBoxes).

person re-identification datasets into a new, large dataset for domain discovery, called DLDP domain discovery dataset[1]. We combine 10 datasets which together contain images of 4,786 different persons with a total of 41,380 bounding boxes. Of these bounding boxes 27,283 are manually annotated and 14,097 are obtained by a person detector. Table 1 shows the sources used to construct the DLDP domain discovery dataset. We resize all bounding boxes to a uniform size of 160×80 pixels. This DLDP divergent data sampling allows us to discover domains which cover a large and diverse spectrum of possible variation in person visual appearance. We show in our experiments its suitability for generalisation to person re-id in new/unseen camera domains.

---
[1]The DLDP dataset will be made publically available.

3.1.2 Prototype-Domain Discovery

We wish to explore deep learning based clustering to discover dominant (prototype) visual domains from the multi-source pooled DLDP dataset. In particular, we exploit the concept of unsupervised deep embedding space learning proposed in [34], but importantly, adopted to utilise the available person id labels from the re-id datasets. Our *supervised* deep learning clustering model alternates between (1) training a CNN to learn a feature embedding from the re-id image datasets and (2) applying conventional k-means clustering in the embedding space to find clusters. To initialize the weights of our feature embedding CNN, we train the model using the person ID labels available in the data. Our model architecture is given in Table 2 (Section 3.2.2 for more details). We set the last, fully connected layer of the network to 4,786 dimensions and train using person ID labels in a one-hot encoding and a softmax loss for person ID classification. The multi-source dataset ensures that the influence of any particular dataset's bias on the initial feature embedding is reduced. Moreover, we apply data augmentation by cropping and flipping the images, and also ensure unbiased sampling by selecting images from different data sources in DLDP with equal frequency. Image cropping is performed by resizing an image to 30×10 pixels larger than the net requires and randomly cropping it down to the correct size. Through data augmentation, the hypothetical data pool size is increased by a factor of 600. We ensure unbiased sampling by selecting images from different data sources in DLDP with equal frequency. This results in more data augmentation on the smaller data sources (among the ten sources in the DLDP dataset) so to prevent the CNN

from overfitting to the larger data sources. For the domain discovery part (k-means clustering) the person ID softmax loss layer is replaced by a softmax loss which corresponds to the number of clusters, set to eight in our current model[2]. Thus, after a supervised initialization, the domain discovery continues in an unsupervised manner.

The initialization from a person re-identification net is crucial to the success of our prototype domain discovery. The re-identification training ensures that the initial model does *not* react strongly to the dataset biases present in our feature pool. *This prevents the clustering from simply discovering trivial dataset boundaries as prototype domain boundaries* and instead, lets the model focuses more on the content of each person bounding box.

3.1.3 Training Strategy

For training of our deep clustering model we use a low initial learning rate of 0.001. This ensures that the cluster embedding does not deviate too quickly from its re-id label constrained initialization. Given the initial embedding, we perform 25 runs of k-means clustering in the embedding space and select the best result for the next refinement of the embedding (*i.e.* step 2 in Section 3.1.2). This ensures stability of the iterative training process. The refinement (fine-tuning) of the embedding CNN is then performed for a further 10,000 training iterations (*i.e.* step 1 in Section 3.1.2). We divide the learning rate of the embedding by 10 every two iterations of the discovery process. This iterative process is repeated until less than 1% of images change their cluster assignments. Some examples of learned prototype domains (*i.e.* clusters in the embedding feature space) with their corresponding images are shown in Figure 3.

3.2. Deep Learning Domain Perceptive Re-Id Model

The second stage of our DLDP re-id model consists of learning a domain-sensitive re-id model for each prototype domain. That is, we train one feature embedding *with all person ID labels* for each of the discovered clusters in the feature embedding space resulted from the first stage. To that end, we start by training a common generic baseline re-id model on all available data without considering the domains. The individual domain models are then trained by fine-tuning this baseline model. The same baseline model is also used as initialization for the domain discovery approach described in Section 3.1.2.

3.2.1 Baseline Generic Model

As a baseline approach we train a model of the architecture given in Table 2 (Section 3.2.2 for more details) on all

---

[2]We choose 8 clusters as a tradeoff between the number of domains available to our model and the computational effort involved in training our domain-specific embeddings. Future work can further optimise it.

Figure 3. Example domains discovered by our approach using the proposed initialization with a re-id net (top 3 rows, supervised initialization) and initialization by weights learned through autoencoding (AE) (bottom 3 rows, unsupervised initialization). The re-id initialization leads to more semantically meaningful domain (*e.g.* light-colored, yellow and blue clothing). The AE initialization is strongly influenced by dataset bias and learns domains corresponding to datasets (*e.g.* CAVIAR4REID, 3DPeS, PRID).

available training data to learn a generic feature embedding without domain specific adaptation. We train the baseline model for 60,000 iterations. The initial learning rate is set to 0.1 and divided by 10 after every 20,000 iterations. We use the output of the 512 dimensional layer (fc feat in Table 2) just before the loss as our feature embedding for person re-id. The resulting features are compared using cosine distance.

3.2.2 Domain Embeddings

In order to learn feature embedding focused on each of the domains we need to first create suitable domain-specific training data. For any person ID in a given domain we thus select all of that person's images and add them to the training data for the domain. This data sampling method allows the domain models to specialize and focus particularly on the visual cues relevant to persons from their domain while not having to also learn how to distinguish persons from different domains.

The architecture (Table 2) we use to learn the domain-specific feature embedding is motivated by a number of recent studies. It consists of an inital set of four convolutional layers with filter sizes of 3×3. This configuration of multiple layers with small filter sizes was shown to perform well for image classification in the VGG nets [26].

name	patch size, stride	output dim	# filters
input		$3 \times 160 \times 64$	
conv 1-4	$3 \times 3, 1$	$32 \times 160 \times 64$	
pool	$2 \times 2, 2$	$32 \times 80 \times 32$	
inception 1a		$256 \times 80 \times 32$	64
inception 1b	stride 2	$384 \times 40 \times 16$	64
inception 2a		$512 \times 40 \times 16$	128
inception 2b	stride 2	$768 \times 20 \times 8$	128
inception 3a		$1024 \times 20 \times 8$	128
inception 3b	stride 2	$1536 \times 10 \times 4$	128
fc feat		512	
fc loss		#person ids	

Table 2. DLDP model architecture for prototype domain discovery.

We further adopt insights from [27] and [28] to add multiple (four) inception layers to our network. We modify the original inception architecture by replacing the 5×5 layer with two 3×3 layers, reducing the grid size and expanding filter banks as suggested in [28]. We apply batch normalization [13] after each layer and use a softmax loss based on the person ID labels for training. Our feature embeddings are of size 512.

3.2.3 Training Strategy

We begin training by disregarding the identified domain borders and combining all available person IDs into one softmax layer. We train this net for an initial 60,000 iterations with a learning rate on 0.1 which is divided by 10 every 20,000 iterations. After this, we continue to train individually for each domain relying only the corresponding data pool. The dimension of the softmax layers is adapted accordingly. For each domain we continue training for 30,000 iterations at an inital learning rate of 0.001. Our input images are resized to a size of 210×70. Data augmentation is then performed by randomly flipping images and randomly cropping them to a final input size of 180×60. Similar to [31] we apply hard negative mining by selecting misclassified training images and fine-tuning each net for a further 10,000 iterations at a reduced learning rate of 0.00001.

3.2.4 Automatic Domain Selection

After model training and during model deployment, a probe person image is first matched to its most likely domain by the deep clustering model (Section 3.1). The corresponding domain embedding (domain specific re-id model) is then used to rank the gallery images by computing the corresponding 512 dimensional embedding and using cosine distance for matching the probe image. Note, the camera view and the target domain of the probe image is new, *i.e.* unseen and independent from any of the multi-sources used to construct the DLDP dataset.

4. Experiments

Datasets: We evaluate our model on two publicly available large re-id datasets: CUHK-SYSU [33] and PRW [37], both of which are independent/unseen from the ten multi-source data pool used to construct our DLDP domain discovery training dataset. Both datasets contain a large number of viewing angles. CUHK-SYSU consists of pedestrian images collected by handheld cameras as well as scenes from movies and the PRW dataset was collected with six cameras on a campus environment. The datasets contain 8432 and 932 person ids and 99,809 and 34,304 bounding boxes, respectively. Both datasets provide full images to enable automatically detected person bounding boxes to be evaluated in person re-id, subject to occlusion, bbox misalignment, and large changes in resolution/low-resolution. Some example images of both datasets are depicted in Figure 1. These characteristics of the two datasets allow us to investigate the generalization capability of our approach, its ability to handle large amounts of varying views and to evaluate its performance against automatically detected person bboxes for more realistic evaluation.

Evaluation protocol: A central objective of our approach is *not* to require any training data on the target domain for the re-id task. To that end, in the experiments we only used the test part of both datasets. The CUHK-SYSU dataset contains a fixed set of 2,900 query persons and gallery sets of multiple sizes (at most 6,978 images). The PRW dataset contains a fixed query set of 2,057 bounding boxes and a gallery size of 6,112 test images. Note that in both datasets each gallery image contains multiple persons and an automatic person detector may generate additional false posistive bounding boxes. We follow the exact evaluation protocols specified in [33] and [37] respectively, and used the provided evaluation code where applicable. Also note, both datasets contain many persons without id in the galleries, *i.e.* the re-id tasks in these datasets are potentially *open-set* given the unknown distractors in the target population. To give a direct comparison to the reported results in [33] and [37], we also adopt mean Averaged Precision (mAP) and Rank-1 accuracy as evaluation metrics.

Comparison with the state-of-the-art: To demonstrate the effectiveness of our approach, we compared our model directly to the state-of-the-art reported in [33] and [37], using both manually labelled person bounding boxes (ground truth) and automatically detected bounding boxes. Results on the CUHK-SYSU dataset for gallery sizes of 100 images

		mAP	Rank-1
GT	Euclidean [33]	41.1	45.9
	KISSME [14]	56.2	61.9
	BoW [36]	62.5	67.2
	IDNet [33]	66.5	71.1
	Baseline Model	68.4	70.3
	DLDP	**74.0**	**76.7**
Detections	Person Search [33]	55.7	62.7
	Person Search rerun	55.79	62.17
	DLDP (SSD VOC300)	49.53	57.48
	DLDP (SSD VOC500)	**57.76**	**64.59**

Table 3. DLDP re-id performance comparison against both supervised (KISSME, IDNet, Person Search) and unsupervised (Euclidean, BoW) methods on the CUHK-SYSU dataset.

		mAP	Rank-1
DPM Inria	IDE [37]	13.7	38.0
	IDE$_{det}$ [37]	**18.8**	**47.7**
	BoW + XQDA [37]	12.1	36.2
	Baseline Model	12.9	36.5
	DLDP	15.9	45.4
SSD	BoW + XQDA (SSD VOC300)	6.8	26.6
	DLDP (SSD VOC300)	10.1	35.3
	DLDP (SSD VOC500)	**11.8**	**37.8**

Table 4. DLDP re-id performance on the PRW dataset in comparison to state-of-the-art. All results are obtained by considering 5 bounding boxes per image. Note that all approaches except ours were trained (supervised) on the PRW dataset.

are given in Table 3. Our baseline generic re-id model (Section 3.2.1) given manually labelled person bounding boxes (ground truth) as input outperforms not only [33] using conventional image features but also the deep IDNet model which has the advantage of being trained on the CUHK-SYSU dataset itself at Rank-1 by 1.9%. The reason is likely a combination of our deeper 10 layer network architecture, the use of inception layers and batch normalization. For our domain adaptive model given manually labelled person bounding boxes, our model outperforms [33] by 7.5% and 5.6% in mAP and Rank-1 respectively, a further improvement of 6% in both mAP and Rank-1 over our generic baseline model. This suggests that the DLDP model learning for prototype-domain adaptive re-id is more effective than the "blind" generic model.

For automatic detection generated person bounding boxes, we adopt the SSD VOC500 person detector [20]. For re-id given these automatic detections, our prototype-domain adaptive model outperforms the state-of-the-art person search deep model [33] by 2.06% and 1.89% on mAP and Rank-1 respectively, despite a very critical difference

that the person search deep model [33] was trained jointly for person detection and re-identification *using part of the CUHK-SYSU dataset, i.e.* both their detector and their re-id matching model were trained and fine-tuned on the target domain. In contrast, our DLDP model did not benefit from training detectors in the target domain, nor fine-tuning re-id model on the target domain data.

For the evaluation on the PRW benchmark, we compared DLDP to a baseline using BoW features and XQDA metric learing [19] and two deep feature embeddings IDE and IDE$_{det}$ from [37] which are based on the AlexNet [15] architecture, trained on ImageNet and fine-tuned for re-id on PRW. For person detection, we used both the DPM person detector [8] trained on the INRIA dataset [6] provided by [37] and the SSD detectors for a fair comparison. Our results are shown in Table 4. All reported results were obtained by considering five bounding boxes per gallery image which is the value at which the methods reported in [37] perform best. It is evident that the SSD detectors decrease re-id performance for all models as the SSD detectors seem to perform poorly on the PRW dataset. Regardless, our model outperforms both the BOW+XQDA baseline and the deep IDE feature embedding reported in [37] when identical DPM person detector was used, by 2.2% and 7.4% in mAP and Rank-1, respectively (for the more accurate DPM detections). The improved deep IDE$_{det}$ embedding of [37] is trained by fine-tuning the AlexNet first for person/background classification followed by further fine-tuning for re-id. It outperforms DLDP by 2.9% and 2.3% in mAP and Rank-1 accuracy. However, our performance remains competitive and has its unique advantage over IDE$_{det}$ embedding. This is because that IDE$_{det}$ was not only *trained directly on PRW* but also *specifically adapted on the PRW* for sensitivity to false positive person detections. DLDP benefited from none of that.

In summary, our DLDP model (given the output from comparable/identical person detectors) outperforms most of the state-of-the-art person re-identification methods on both the CUHK-SYSU and the PRW benchmark datasets. It is even more significant that our results are obtained without any labelled data training on the target test domains whilst all other methods require training data from the target domains. Qualitative examples on both datasets are shown in Figure 8, including two failure cases in the last row. Note that the incorrect results for all queries have a color composition or clothing configuration that is reasonably similar to the query image. In particular, DLPD understandably ranks near-identically looking people (PRW, row 2) very high. In the failure case on PRW our model appears to focus on the structural pattern created by the bikes in combination with white-dressed persons.

Effects from gallery size increase: The CUHK-SYSU dataset offers multiple gallery sets of varying sizes. This

		50	100	500	1000	2000	4000	all (6978)
Person Search [33]	mAP	58.72	55.79	47.38	43.41	39.14	35.70	32.69
DLDP (SSD VOC500)		**60.8**	**57.7**	**49.2**	**45.2**	**41.4**	**38.1**	**35.2**
Person Search [33]	Rank-1	64.83	62.17	53.76	49.86	45.21	41.86	38.69
DLDP (SSD VOC500)		**67.6**	**64.6**	**57.0**	**52.9**	**49.2**	**46.1**	**43.1**

Table 5. Comparison of DLDP to [33] for different gallery sizes on CUHK-SYSU. Results of [33] were obtained using the provided code.

	Deep+Kissme [33]		ACF+BOW [33]		Person Search [33]		Person Search [33] rerun		DLDP (SSD VOC500)	
	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1
Whole	39.1	44.9	42.4	48.4	55.7	62.7	55.79	62.17	**57.7**	**64.6**
Occlusion	18.2	17.7	29.1	32.1	**39.7**	**43.3**	34.97	37.43	38.9	39.0
LowRes	43.8	42.8	44.9	53.8	35.7	42.1	35.21	40.00	**41.9**	**49.0**

Table 6. Comparisons on the CUHK-SYSU occlusion and low resolution tests.

allows us to evaluate the influence of larger numbers of detractors on re-identification accuracy in a more realistic open-set setting. Table 5 shows results of our DLDP model in comparison to those achieved by the end-to-end person search deep network model [33], where results were obtained by running the code provided by the authors. Our obtained results closely match those reported in [33] (Figure 7 b)). Overall, our DLDP model consistently ourperforms the end-to-end person search deep model by a constant 2% in mAP regardless gallery size; 3% in Rank-1 for low gallery sizes of 50 images (correspoding to 256 bounding boxes) and up to 5.4% in Rank-1 for the largest possible gallery of all 6978 images (36984 bounding boxes). This suggests that the DLDP model is less sensitive to increase in gallery size, even *without* benefiting from learning on target domains.

Effects from occlusion and low-resolution: Finally, we evaluated the effects of occlusion and low-resolution probe images. The CUHK-SYSU dataset provides two probe subsets for this purpose, which were created by sampling heavily occluded probe images and the 10% probe images with the lowest resolutions, respectively. Gallery sizes for this evaluation are fixed at 100 images. We report results using the SSD VOC500 person detection in Table 6 and compared to the end-to-end person search deep network model. Consistent with the observation made by [33], an occluded probe image causes more difficulty for re-id than that of low-resolution imagery. For low-resolution, our DLDP model suffers only a 15% loss in mAP and Rank-1, as compared to a 20% decrease for the end-to-end person search deep model. On occlusion, the reported results on the end-to-end person search model are less affected (reduced by 16.0% mAP and 19.4% Rank-1) than our DLDP model whose performance is reduced by 18.8% in mAP and 25.6% in Rank-1. However, using the author provided code, we could not re-create the same results as reported in [33] for the occlusion test. Instead, we obtained the result on the end-to-end person search model for occlusion test 5% lower

than reported, almost identical to that of DLDP.

To gain more insight, we further report the re-id results from our DLDP model on the occlusion and low-resolution tests but this time using manually labelled person bounding boxes (Table 7). The overall results are much improved by relying on ground truth detections. This may partly be due to that ground truth labelled bboxes resemble more closely to data the model was trained on, therefore the negative impact of low-resolution query images is less severe. This suggests that the resolution gap between probe and gallery can be handled well by DLDP provided person bbox detection is reasonably accurate without significant misalignment.

5. Conclusion

In this work, we presented a novel approach to domain sensitive person re-identification by deep learning *without* the need for training data from the target (test) domains. The new model DLDP automatically discovers prototype domains from independent diverse datasets and learns specific feature embedding for each of the discovered domains. In model deployment, each query image is used to select for the most suitable feature embedding with its corresponding domain best fitting the query before the ranking match against the gallery candidates. Our approach has a singificant *unique* advantage, over *all* existing models, of not requiring any target domain data for model learning. Our extensive comparative evaluation on two latest benchmark datasets demonstrate clearly that the proposed DLDP model outperforms the state-of-the-art or is competitive, notwithstanding that all other models benefit from having been trained on the target domain data. It is also evident that the proposed new DLDP model copes well with real-world re-id conditions when automatic person detection, occlusion, low resolution and very large gallery sizes (*i.e.* open world) are unavoidable in model deployment. Future work includes investigating in more detail the impact of the num-

	mAP	Rank-1	Rank-5	Rank-10	Rank-20
DLDP GT Whole	74.0	76.7	86.4	89.7	92.9
DLDP GT Occlusion	56.0	54.0	69.5	76.5	83.4
DLDP GT LowRes	72.0	74.1	86.9	89.7	93.1

Table 7. Results of DLDP on the occlusion and low resolution test sets using ground-truth detections.

Query   **Top rank 8 re-id matches**   **Query**   **Top rank 8 re-id matches**

a)   b)

Figure 4. (a) The top 8 re-id matches by the DLDP model on the CUHK-SYSU test data for a set of five randomly chosen queries from the 100 image gallery setting, and (b) five randomly chosen queries on the PRW test data. Note, rank-2 and rank-3 in the "yellow T-shirt" example in (b) are false matches even though they look very similar. The bottom examples from both (a) and (b) show failure cases when the model failed to find a match in the top 8 ranks.

ber of domains on the accuracy of our approach as well as ways of coupling the domain discovery and learning of domain embeddings more directly in an end-to-end approach.

6. Acknowledgements

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the joint ACM workshop on Human gesture and behavior understanding*. ACM, 2011.

[3] D. Baltieri, R. Vezzani, and R. Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identification. In *International Conference on Image Analysis and Processing*, 2011.

[4] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2011.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), 2005.

[7] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48, 2015.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 2010.

[9] D. Figueira, M. Taiana, A. Nambiar, J. Nascimento, and A. Bernardino. The hda+ data set for research on fully automated re-identification systems. In *Proceedings of the European Conference on Computer Vision*, 2014.

[10] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.

[11] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Proceedings of the Scandinavian conference on Image analysis*, 2011.

[12] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Z. Li. Cross dataset person re-identification. In *Asian Conference on Computer Vision*, 2014.

[13] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[14] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE, 2012.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., 2012.

[16] R. Layne, T. M. Hospedales, and S. Gong. Domain transfer for person re-identification. In *Proceedings of the 4th ACM/IEEE international workshop on Analysis and retrieval of tracked events and motion in imagery stream*, 2013.

[17] W. Li and X. Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[19] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015.

[21] C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90, 2010.

[22] A. J. Ma, J. Li, P. C. Yuen, and P. Li. Cross-domain person reidentification using domain adaptation ranking svms. *IEEE Transactions on Image Processing*, 24, 2015.

[23] L. Ma, X. Yang, and D. Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE Transactions on Image Processing*, 23, 2014.

[24] N. McLaughlin, J. M. Del Rincon, and P. Miller. Data-augmentation for reducing dataset bias in person re-identification. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2015.

[25] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. 2016.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv:1512.00567*, 2015.

[29] X. Wang, W.-S. Zheng, X. Li, and J. Zhang. Cross-scenario transfer person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2015.

[30] L. Wu, C. Shen, and A. van den Hengel. Personnet: Person re-identification with deep convolutional neural networks. *arXiv preprint arXiv:1601.07255*, 2016.

[31] 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. 2016.

[32] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[33] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, 2016.

[34] J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning (ICML)*, 2016.

[35] D. Yi, Z. Lei, S. Liao, S. Z. Li, et al. Deep metric learning for person re-identification. In *Proceedings of the International Conference on Patter Recognition (ICPR)*, 2014.

[36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

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

[38] W.-S. Zheng, S. Gong, and X. Tao. Associating groups of people. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2009.

Appendix

Evaluation on Market-1501

We additionally evaluted DLDP on the Market-1501 [36] dataset. This dataset does not provide full images but was created using the DPM detector instead of manual annotations. Results of DLDP in comparison to state-of-the-art approaches are given in Table 8. DLDP outperforms many recent approaches and performs en-par with the approach of [1] (-0.12% mAP and -0.44 Rank-1). DLPD is clearly outperformed by [9] and [7] which both achieve a significant improvement over the previous state-of-the-art and beat DLPD by up to 13.32% in mAP and 14.42% in Rank-1. All approaches given in Table 8 make full use of the Market training dataset with the noteable exception of [5] which uses only 13.58% of the training data to adapt model pretrained on other data. DLPD clearly outperforms this result without using any of the training data.

A qualitative impression of the results of DLPD on the Market-1501 dataset is depicted in Figure 8. Again, it can be observed that most of the incorrect results are quite reasonable and share one or more salient features with the query person (e.g. the black backpack in the second row or either a bag-strap or gray shirt in the fourth row). The last row shows a query for which only one out of five true matches is returned among the top 15 results. We consider this a near failure case. A success implies that the model finds *all* of the existing true matches.

	mAP	Rank-1
Gated S-CNN [7]	**39.55**	**65.88**
DNS [9]	35.68	61.02
SCSP [1]	26.35	51.90
DLDP	26.23	51.46
Multiregion Bilinear DML [6]	26.11	45.58
End-to-end CAN [3]	24.43	48.24
TMA LOMO [5]	22.31	47.92
WARCA-L [2]	-	45.16
MST-CNN [4]	-	45.1

Table 8. DLDP's performance in context of many recent state-of-the-art approaches for the single-query setting on the Market-1501 dataset.

Ground-Truth Detections on PRW

In Table 9 we show the performance of DLDP and our baseline model (see Section 3.2.1 in the main paper) on the PRW dataset. We compare to the BoW+XQDA baseline which is provided with the evaluation code of [37]. Compared to the CUHK-SYSU dataset (compare Table 3, main paper) the improvement in accuracy achieved by relying on ground-truth is much less pronounced for all approaches.

This is likely due to the fact that the ground-truth on PRW was obtained in part using the DPM-Inria person detector, thus giving that detector an unusually high localization accuracy on the dataset. This also explains the comparatively weak performance of the SSD detectors observed in Table 4 of the main paper. DLDP is able to maintain its advantage over the BoW+XQDA baseline on ground-truth and outperforms it by 1.8% mAP and 8.5% Rank-1.

	mAP	Rank-1
BoW + XQDA	16.7	38.8
Baseline Model	16.1	43.2
DLDP	**18.5**	**47.3**

Table 9. DLDP re-id performance on the PRW dataset using ground-truth annotations instead of automatic detections.

Full CMC-Curves

In Figures 5, 6 and 7 we give the full CMC curves for our main experiments on CUHK-SYSU, PRW, and Market-1501, respectively. All curves were generated using the evaluation code provided with the datasets and are compared to the strongest baseline approaches for which code was provided.

In Figure 5 we show DLDP's average accuracy over the first 50 ranks on the CUHK-SYSU dataset. DLDP shows a consistent improvement of more than 5% over the baseline model and narrowly but consistently outperforms the deep PersonSearch approach which integrates detection and re-id into one CNN.
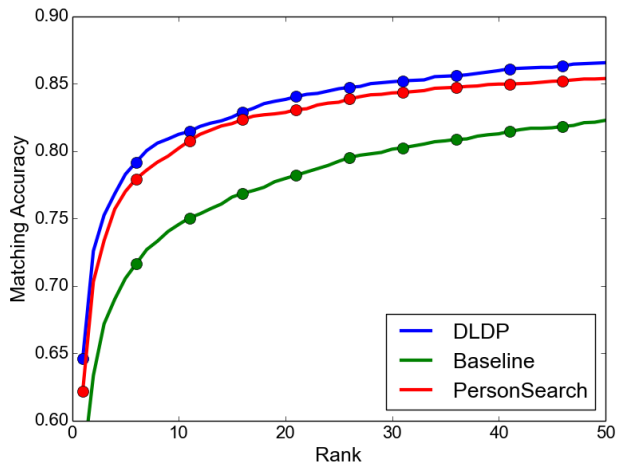

Figure 5. CMC curve of DLDP on CUHK-SYSU compared to our baseline model and the PersonSearch approach presented in [33]. Results are obtained using the default gallery size of 100 images and the SSD-VOD500 detector for DLDP and the baseline model.

Figure 6 shows the first 50 ranks on the PRW dataset.

Our baseline model performs en-par with the BoW+XQDA baseline. Both appraoches are again consistently outperformed by more than 5% by DLDP.
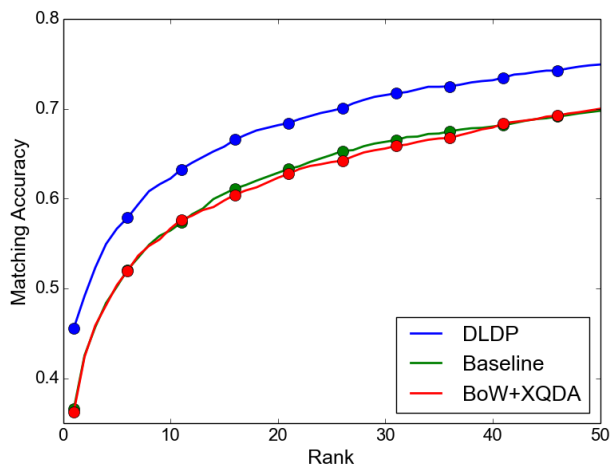

Figure 6. CMC curve of DLDP on PRW compared to our baseline model and the BoW+XQDA baseline provided with the evaluation code. Results were obtained by considering the top 5 detections in each image. All approaches are evaluated on the provided detections of the DPM-Inria detector.

The average accuracy over the first 50 ranks on the Market-1501 dataset is depicted in Figure 7. The difference between DLDP, our baseline and the BoW+KISSME baseline is less significant on this dataset. However, in the important segment of ranks 1-20 DLDP has a clear advantage. For ranks above 38 the BoW+KISSME approach actually outperforms both our baseline model and DLDP narrowly.
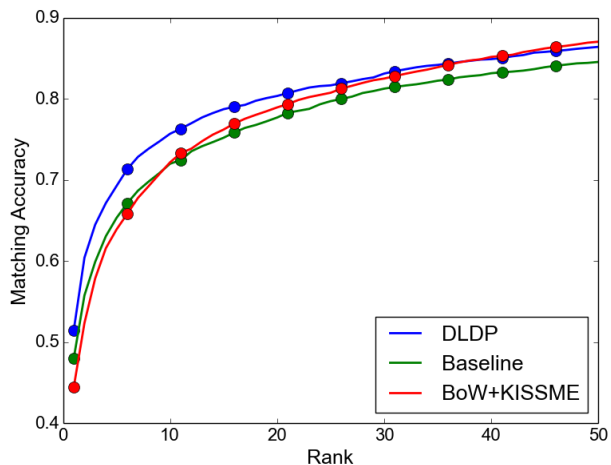

Figure 7. CMC curve of DLDP on Market-1501 compared to our baseline model and the BoW+KISSME baseline provided with the evaluation code. Results are for the single-query setting.

Appendix-Literature

[1] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] C. Jose and F. Fleuret. Scalable metric learning via weighted approximate rank component analysis. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[3] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *arXiv preprint arXiv:1606.04404*, 2016.

[4] 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 *Proceedings of the 2016 ACM Multimedia Conference (ACMMM)*, 2016.

[5] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury. Temporal model adaptation for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[6] E. Ustinova, Y. Ganin, and V. Lempitsky. Multiregion bilinear convolutional neural networks for person re-identification. *arXiv preprint arXiv:1512.05300*, 2015.

[7] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[8] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, 2016.

[9] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

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

Query   **Top rank 15 re-id matches**

Figure 8. The top 15 re-id matches by the DLDP model on the Market-1501 dataset. Correct matches are framed red. Matches labelled as "junk" in the dataset and not considered in the evaluation protocol are framed blue. The last row shows a failure case where only one (out of five) correct images could be found in the top 15 results.