Generalizing A Person Retrieval Model Hetero- and Homogeneously

Zhun Zhong[1,2], Liang Zheng[2,3], Shaozi Li[1](✉),Yi Yang[2]

[1]Cognitive Science Department, Xiamen University, China
[2] Centre for Artificial Intelligence, University of Technology Sydney, Australia
[3] Research School of Computer Science, Australian National University, Australia
{zhunzhong007,liangzheng06,yee.i.yang}@gmail.com szlig@xmu.edu.cn

Abstract. Person re-identification (re-ID) poses unique challenges for unsupervised domain adaptation (UDA) in that classes in the source and target sets (domains) are entirely different and that image variations are largely caused by cameras. Given a labeled source training set and an unlabeled target training set, we aim to improve the generalization ability of re-ID models on the target testing set. To this end, we introduce a Hetero-Homogeneous Learning (HHL) method. Our method enforces two properties simultaneously: 1) camera invariance, learned via positive pairs formed by unlabeled target images and their camera style transferred counterparts; 2) domain connectedness, by regarding source / target images as negative matching pairs to the target / source images. The first property is implemented by *homogeneous* learning because training pairs are collected from the same domain. The second property is achieved by *heterogeneous* learning because we sample training pairs from both the source and target domains. On Market-1501, DukeMTMC-reID and CUHK03, we show that the two properties contribute indispensably and that very competitive re-ID UDA accuracy is achieved. Code is available at: https://github.com/zhunzhong07/HHL

Keywords: Person re-identification, Unsupervised domain adaptation

1 Introduction

Given a query, person re-identification (re-ID) aims to retrieve the same person from a database collected by different cameras from the query. Despite the dramatic performance improvement obtained by the convolutional neural network (CNN), it is reported that deep re-ID models trained on the source domain may have a large performance drop on the target domain [7,10]. The main reason is that the data distribution of the source domain is usually different from the target domain. In this paper, we consider the setting of unsupervised domain adaptation (UDA), in which during training we are provided with labeled source training images and unlabeled target training images. Performance is evaluated on the target testing database.

Unsupervised domain adaptation [37,26,16], which has been studied extensively in image classification, object detection and semantic segmentation, faces new challenges in the context of person re-ID. On the one hand, the source and target domains in person re-ID have entirely different classes (person identities), while in generic UDA, the source and target share the same set of classes. On the other hand, a critical factor that leads to domain variance in person re-ID can be clearly identified, *i.e.,* the disparities of cameras. Even in the unlabeled target domain, camera information, *i.e.,* the camera by which an image is captured, is known. However, it remains unknown in the UDA community how to effectively leverage the camera information for person re-ID.

In this paper, our design is motivated in two aspects, closely associated with the new challenges mentioned above. First, a critical part of our motivation arises from the *intra-domain* image variations caused by different camera configurations. This perspective is largely overlooked in recent methods addressing the UDA problem in person re-ID. These recent works either concentrate on content-preserving source-target translation models [39,7] or employ both the attribute and identity labels to learn a transferable model [38]. To our knowledge, these methods only consider the overall *inter-domain* differences, but do not explicitly consider the *intra-domain* image style variations caused by different camera configurations. In fact, the *intra-domain* camera style difference is a critical influencing factor for person re-ID, because during testing, the query and its ground truth matches are captured by different cameras. Without considering the fine-grained *intra-domain* image variations, a transfer learning model trained on the source set will probably only capture the overall data bias between the two domains and have problems when encountering the large *intra-domain* image variations in target domain testing set.

Second, we consider the prior that the source and target sets have entirely different classes / identities, so a source image and a target image naturally form a negative training pair. A similar idea has been explored by Deng *et al.* [7]. However, the two papers differ in the purpose of using this prior. In [7], Deng *et al.* use the negative pairs to improve the image-image translation model, so that the generated images will largely preserve their identity label, a desirable property for UDA. In comparison, we directly use these negative pairs to learn person embeddings within a triplet loss formulation.

With the two considerations, we propose a new unsupervised domain adaptation method, named **H**etero-**H**omogeneous **L**earning (HHL), for the person re-ID task. HHL is constructed without target supervision, *i.e.,* we do not require laborious manual annotations such as identities in the target set. In fact, the construction of HHL requires a source set (identity labels given), a target set (without identity labels), and the camera information for each image in the target set. Here, we emphasize that the camera ID for each target image can be obtained along with the raw videos: it suffices to simply record the ID of the camera capturing the videos. Therefore, we call the construction of HHL "without target supervision", or in the most strict way "with extremely weak target supervision".

In our method, HHL underpins constraints at two properties. First, we constrain to learn person embeddings which are robust to camera variances in the target domain. To achieve this *camera invariance* property in an unsupervised fashion, positive training pairs are generated by image-image translation, viewing each camera as an individual style. Second, in order to endow *domain connectedness* to the system, we learn the underlying structures between source and target domains using negative training pairs sampled from the source and target sets, respectively. In this paper, imposing the camera invariance property is a *homogeneous learning* process because training images are from the same domain. Imposing the domain connectedness property implies a *heterogeneous learning* procedure because the training samples are from two domains. The two properties produce a positive pair homogeneously and a negative pair heterogeneously, which, bridged by an anchor image to be fed into a triplet loss training.

To summarize, this paper is featured in the three aspects. First, a Hetero-Homogeneous Learning (HHL) scheme is introduced. Through a triplet loss, it brings about camera invariance and domain connectedness to the system, which are essential properties towards an effective UDA approach in person re-ID. Second, HHL is a new method for training sample construction in UDA. It is robust to parameter changes. The insights and indispensability of camera invariance and domain connectedness are validated through experimental studies. Third, we report new state-of-the-art UDA accuracy on the Market-1501, CUHK03 and DukeMTMC-reID datasets.

2 Related Work

Unsupervised domain adaptation. Our work is closely related to unsupervised domain adaptation (UDA) where the target domain is unlabeled. Most of the previous methods try to align the source to the target domain by reducing the divergence of feature distributions [37,26,11,35,36,41]. These methods are motivated by the theory stating that the error for the target domain is bounded by the difference between domains [2]. CORAL [35] aligns the mean and covariance of source domain and target domain distributions and achieves promising results in various visual recognition tasks. Further, deep CORAL [36] extends the approach by incorporating the CORAL loss into deep model. There exist many methods which aim at providing pseudo-labels to unlabeled samples. Several methods utilize similarity of features to give pseudo-labels to unlabeled target samples [31,33]. In [33], an approach is presented to estimate the labels of unlabeled samples by using the k-nearest neighbors. Then, the predicted labels are leveraged to learn the optimal deep feature. Alternatively, many methods try to predict labels to unlabeled samples by leveraging the predictions of a classifier and retraining the classifier with both labeled samples and pseudo-labeled samples, which is called co-training [50]. The underlying assumption of these methods is that high-confidence prediction is a mostly correct class for an unlabeled sample. In [4], the idea of co-training is applied to domain adaptation by gradually adding the target samples of high-confidence predictions to the training set. Saito *et al.* [32] propose to generate pseudo-labels for target domain samples through three classifiers asymmetrically and train the final classifier with predicted labels.

Recently, Many Generative Adversarial Networks (GAN) [12] based domain adaptation approaches focus on learning a generator network that transforms samples in the pixel space from one domain to another [24,3,16]. CyCADA [16] adapts representations at both the pixel-level and feature-level via pixel cycle-consistency and semantic losses, it achieves high performance on both digit recognition and semantic segmentation. Most of existing unsupervised domain adaptation methods assume that class labels are the same across domains, while the person identities (classes) of different re-ID datasets are totally different. Hence, the approaches mentioned above fail to be utilized directly for the problem of unsupervised domain adaptation in person re-ID.

Unsupervised person re-ID. Hand-craft features can be directly applied for unsupervised person re-ID, for example, ELF [13], LOMO [23], and SDALF [1], which aim to design or learn robust feature for person re-ID. These methods often ignore the distribution of samples in the dataset and fail to perform well on large-scale dataset. Benefit from the remarkable success of deep learning [21,14,9,8,27], recent works [10,25,40] attempt to predict pseudo-labels to unlabeled samples based on the deep learning framework. Fan *et al.* [10] propose an unsupervised re-ID approach for iteratively applying k-means clustering to assign labels to unlabeled samples and fine-tuning the deep re-ID model on the target domain. Liu *et al.* [25] estimate labels with k-reciprocal nearest neighbors [29] and iteratively learn features for unsupervised video re-ID. Wu *et al.* [40] propose a progressive sampling method to gradually predict reliable pseudo labels and update deep model for one-shot video-based re-ID.

Few works [7,38,39,28,17] have studied on unsupervised domain adaptation for re-ID. Peng *et al.* [28] propose to learn a discriminative representation for target domain based on asymmetric multi-task dictionary learning. Deng *et al.* [7] learn a similarity preserving generative adversarial network based on CycleGAN [49] to translate images from source domain to target domain. The translated images are utilized to train re-ID model in a supervised way. In [38], a transferable model is proposed to jointly learn attribute-semantic and identity discriminative feature representation for target domain. These approaches aim at reducing the gap between source domain and target domain on either the image-level space [7,39] or feature-level space [38,28,17], while overlook the image style variations caused by different cameras in target domain. In this work, we explicitly consider the intra-domain image variations caused by cameras to learn discriminative re-ID model for target domain.

3 Proposed Method

Problem definition. For unsupervised domain adaptation in person re-ID, we have a labeled source set $\{X_s, Y_s\}$ consisting of N_s person images. Each image x_s corresponds to a label y_s, where $y_s \in \{1, 2, ..., M_s\}$, and M_s is the number of identities. We also have N_t unlabeled target images from unlabeled target set

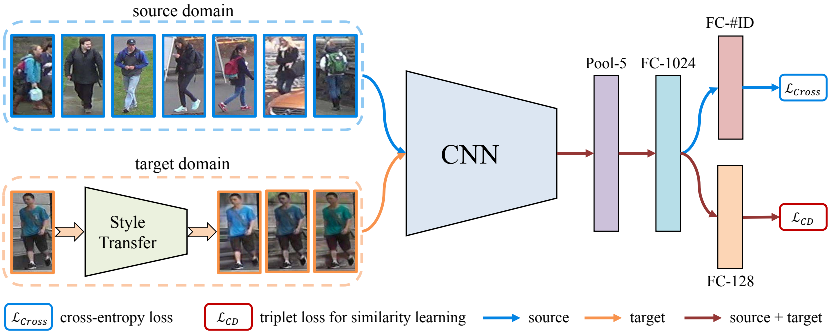

Fig. 1. The framework of the proposed approach. It consists of two loss functions: 1) cross-entropy loss for classification, which learned by labeled source samples; 2) triplet loss for similarity learning, which imposes camera invariance and domain connectedness to the model and learned through labeled source samples, unlabeled target samples and cameras style transferred samples.

$\{X_t\}$. The identity of each target image x_t in $\{X_t\}$ is unknown. The goal of this paper is to leverage both labeled source training images and unlabeled target training images to learn discriminative embeddings for target testing set.

3.1 Baseline Configuration

We use ResNet-50 [14] as backbone and follow the training strategy in [48] which fine-tunes on the ImageNet [6] pre-trained model. We discard the last 1,000-dim fully connected (FC) layer and add two FC layers. The output of the first FC layer is 1,024-dim named as "FC-1024", followed by batch normalization [18], ReLU and Dropout [34]. The output of the second FC layer, named as "FC-#ID" is M_s-dim, where M_s is the number of identities (classes) in the labeled training set.

Given the labeled training images, an effective strategy is to learn the ID-discriminative embedding (IDE) [44] for person re-ID. The cross-entropy loss is employed by casting the training process as a classification problem. The cross-entropy loss is written as,

$$\mathcal{L}_{Cross} = -\frac{1}{n_s}\sum_{i=1}^{n_s} \log p_i(y), \qquad (1)$$

where n_s is the number of labeled training images in a batch, $p_i(y)$ is the predicted probability of the input belonging to ground-truth class y. We name this model as **baseline** throughout this paper.

The IDE-based methods [44,46,47] achieve good performance on fully labeled datasets, but often fail to generalize to a new target set. Next, we will describe the Hetero-Homogeneous Learning (HHL) approach to improve the transferability of the baseline.

3.2 Network Architecture

The network used in this paper is shown in Fig. 1. It has two branches. The first branch is the same with the **baseline**, which is an identification task. The second branch is different from the first branch in two aspects: 1) a 128-dim FC layer named "FC-128" is used instead of the "FC-#ID" layer; 2) a triplet loss is used instead of the cross-entropy loss. Therefore, our network has two loss functions, a cross-entropy loss for classification and a triplet loss for similarity learning. For similarity learning, we employ the triplet loss used in [15], which is formulated as,

$$\mathcal{L}_T(X) = \sum_{x_a, x_p, x_n} [m + D_{x_a,x_p} - D_{x_a,x_n}], \ \forall \ x_a, x_p, x_n \in X, \tag{2}$$

where X represents images in a training batch, x_a is an anchor point. x_p is a hardest (farthest) sample in the same class with x_a, and x_n is a hardest (closest) sample of a different class to x_a. m is a margin parameter and $D(\cdot)$ is the Euclidean distance between two images in the embedding space. We use the output of FC-128 as the embedding feature and set m to 0.3. Note that during re-ID testing, we use the output of Pool-5 (2,048-dim) as person descriptor.

3.3 Camera Invariance Learning

The variation of image style caused by cameras is a critical influencing factor during person re-ID testing procedure. To achieve the camera invariance property in target domain, we impose the camera invariance constraint by learning with both unlabeled target images and their counterparts containing the same person but with different camera styles.

In order to generate new target images that more or less preserve the person identity and reflect the style of another camera, we employ the CamStyle approach [48] to learn camera style transfer model in the target set. Different from [48] which uses CycleGAN [49] for image-image translation, we build CamStyle based on StarGAN [5]. This is because StarGAN allows us to train multi-camera image-image translation with a single model, while CycleGAN needs to train a translation model for each pair of cameras. Suppose we have C cameras in the target set. We first train a StarGAN model which enables image-image translation between every camera pair. With the learned StarGAN model, for a real image $x_{t,j}$ collected by camera j ($j \in 1, 2, ..., C$) in the target set, we generate C fake (camera style transferred) images $x_{t^*,1}$, $x_{t^*,2}$,..., $x_{t^*,C}$ which more or less contain the same person with $x_{t,j}$ but whose styles are similar to camera $1, 2, ..., C$, respectively. Note that the C images include the one transferred to the style of camera j, that is, the style of the real image $x_{t,j}$. Examples of real images and fake images generated by CamStyle [48] are shown in Fig. 2.

To learn camera invariant person embeddings for the target set, we view $x_{t,j}$ and its corresponding fake images $x_{t^*,1}$, $x_{t^*,2}$,..., $x_{t^*,C}$ as belonging to the same class. We view all the other images as belonging to a different class with $x_{t,j}$. For simplicity, we omit the subscript of camera. Specifically, we compute

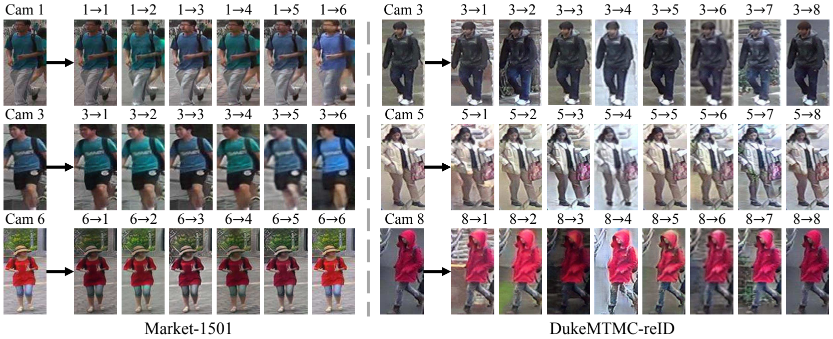

Fig. 2. Examples of camera style transfer on Market-1501 and DukeMTMC-reID. An image collected by a certain camera is transferred to the style of other cameras. In this process, the identity information is preserved to some extent. The real image and its corresponding fake images are assumed to belong to the same class during training.

a triplet loss through the unlabeled target domain samples $\{x_t^i\}_{i=1}^{n_t}$ and their corresponding camera transferred samples $\{x_{t^*}^i\}_{i=1}^{n_t^*}$. The loss function of camera invariance learning can be written as,

$$\mathcal{L}_C = \mathcal{L}_T(\{x_t^i\}_{i=1}^{n_t} \cup \{x_{t^*}^i\}_{i=1}^{n_t^*}), \tag{3}$$

where n_t is the number of real target images in a training batch, and n_t^* is the number of camera style transferred samples. In our experiment, we generate C fake images for each real target image, *i.e.* $n_t^*/n_t = C$, where C is the number of cameras. In a training batch, x_t^i is randomly selected from the target set, and we assume that x_t^1, x_t^2, ..., $x_t^{n_t}$ as belonging to different classes. Technically speaking, this assumption is incorrect, because each target training class has several images, and it may well be the case that two images of the same class are selected into the training batch. That being said, we will show in Section 3.6 and Fig. 3 that our assumption does not affect the performance noticeably.

3.4 Domain Connectedness Learning

In person re-ID, different domains have completely different classes / identities, so a source image and a target image naturally form a negative training pair. With this prior, we propose to endow domain connectedness to the system by regarding source / target images as negative matching pairs to the target / source images. Given an anchor image from the source, we use source domain labels to construct a positive pair. We then choose a target domain image to form a negative pair with the anchor. Formally, given the labeled source domain samples $\{x_s^i\}_{i=1}^{n_s}$ and the unlabeled target domain samples $\{x_t^i\}_{i=1}^{n_t}$, the loss function of domain connectedness learning can be defined as,

$$\mathcal{L}_D = \mathcal{L}_T(\{x_s^i\}_{i=1}^{n_s} \cup \{x_t^i\}_{i=1}^{n_t}), \tag{4}$$

where n_s is the number of source images, and n_t is the number of target images. In this loss function, since the identities of target images do not overlap with the identities in source domain, each source image and each target image form a negative pair. Therefore, the relationship between the source and target samples is considered, so that the communication and the underlying structures between two domains can be achieved to some extent.

3.5 Hetero-Homogeneous Learning

In this paper, we argue that camera invariance and domain connectedness are complementary properties towards an effective UDA system for person re-ID. To this end, we propose to jointly learn camera invariance and domain connectedness using a single loss in a training batch. Specifically, a training batch contains labeled source images $\{x_s^i\}_{i=1}^{n_s}$, unlabeled real target images $\{x_t^i\}_{i=1}^{n_t}$, and their corresponding fake images $\{x_{t^*}^i\}_{i=1}^{n_t^*}$. The triplet loss function of camera invariance learning and domain connectedness learning can be written as,

$$\mathcal{L}_{CD} = \mathcal{L}_T(\{x_s^i\}_{i=1}^{n_s} \cup \{x_t^i\}_{i=1}^{n_t} \cup \{x_{t^*}^i\}_{i=1}^{n_t^*}). \tag{5}$$

In this loss function, we enforce two properties simultaneously: 1) camera invariance, learned through real target images and its corresponding fake images; 2) domain connectedness, mapping the source and target samples into shared feature space by regarding source / target samples (including their camera style transferred samples) as negative matching pairs to the target / source samples.

Finally, the overall loss function (Fig. 1) in a training batch is expressed as,

$$\mathcal{L}_{HHL} = \mathcal{L}_{Cross} + \beta \mathcal{L}_{CD}, \tag{6}$$

where β is the weight of the joint camera invariance and domain connectedness loss. We name this learning method "Hetero-Homogeneous learning (HHL)" because of the heterogeneous sample selection scheme of domain connectedness learning, and because of the homogeneous sample selection scheme of camera invariance learning. Also, we note that the cross-entropy loss is indispensable in Eq. 6, which provides a basic discriminative ability *learned on the source only.* Without the cross-entropy loss, the system will be harmed significantly.

3.6 Discussion

Why use camera style transfer? In Table 1, we compare the distance between images that undergoes different data augmentation method, *i.e.* random cropping, random flipping and camera style transfer. It is clearly that, the re-ID model trained on source set is robust to random cropping and random flipping on target set, but is sensitive to image variations caused by cameras. Therefore, the change of image style caused by different cameras on target set is a key influencing factor that should be explicitly considered in person re-ID UDA.

How to sample training images from target domain? We compare three sampling strategies, 1) random sampling, we randomly sample n_t target

Table 1. The average distance between two images that undergo different data augmentation techniques. We use the baseline re-ID model (Section 3.1) trained on the source set to extract image descriptors (Pool-5, 2,048-dim) on the target set.

Source	Target	Random Crop	Random Flip	CamStyle Transfer
Duke	Market-1501	0.049	0.034	**0.485**
Market-1501	Duke	0.059	0.044	**0.614**

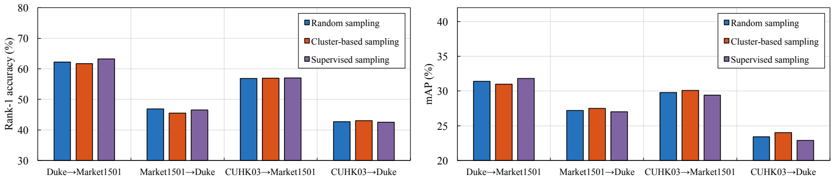

Fig. 3. Comparison of different sampling strategies on the target set, including random sampling, cluster-based sampling and supervised sampling. Rank-1 accuracy and mAP are reported. We set $\beta = 0.5$, $n_t = 16$. We find that different sampling methods achieve very similar results. So for simplicity, we use random sampling throughout the paper.

images in each mini-batch and assign non-overlap randomly identity for each image, *i.e.* each image has a different identity in a mini-batch; 2) cluster-based sampling, at begin of each training epoch, we apply k-means to cluster target images into n_t clusters based on currently learned re-ID model, and sample one image from each cluster to compose training data of target domain in a mini-batch. The cluster-based sampling strategy could effectively avoid to sample the same identity in a mini-batch; 3) supervised sampling, assume that we are provided with labeled target set, we randomly select n_t images in a supervised way ensuring that each target image comes from a different identity. The comparison of different sampling strategies is shown in Fig. 3. It is clearly that random sampling yields quite approximate results with the other two strategies. It is because of the probability of images to be the same identities is very low when sampling few images from target set including a large number of images and identities. Therefore, we use random sampling in this paper.

4 Experiment

4.1 Datasets

We evaluate our method on three re-ID datasets which are considered as large-scale in the community, *i.e.*, Market-1501 [43], DukeMTMC-reID [45,30], and CUHK03 [22]. **Market-1501** [43] contains 32,668 labeled images of 1,501 identities collected from 6 cameras. For evaluation, 12,936 images from 751 identities are used for training, and 19,732 images from 750 identities plus some distractors form the gallery / database. Moreover, 3,368 hand-drawn bounding boxes

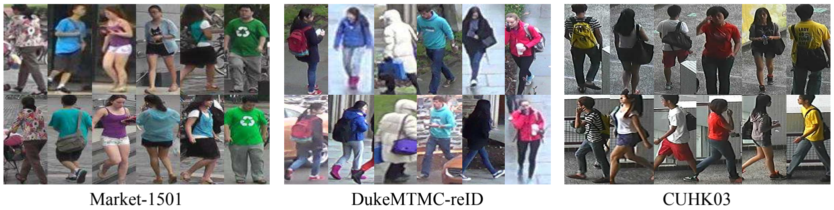

Fig. 4. Example images of the Market-1501, DukeMTMC-reID and CUHK03 datasets. Images in each column represent the same identity / class collected from different cameras. We observe that the image style of the three datasets is very different and that within each dataset, the image style of different cameras is different as well.

from 750 identities are used as queries to retrieve the corresponding person images in the database. We use the single-query evaluation in our experiment. **DukeMTMC-reID** [45] has 8 cameras and 36,411 labeled images belonging to 1,404 identities. Similar to the division of Market-1501, the dataset contains 16,522 training images from 702 identities, 2,228 query images from another 702 identities and 17,661 gallery images. **CUHK03** [22] contains 14,096 images of 1,467 identities. Each identity is captured from two cameras. The dataset has two train / test settings: using labeled bounding boxes and using DPM detected bounding boxes. We use the detected setting because it is more challenging and closer to practical scenarios. Note that images in CUHK03 do not have camera labels, so we cannot perform *camera invariance learning*. Therefore, we only use CUHK03 as the *source domain* instead of the *target domain*. We use the conventional rank-n accuracy and mean average precision (mAP) for evaluation on all datasets. Example persons of different re-ID datasets are shown in Fig. 4.

4.2 Experiment Settings

Camera style transfer model. Given a target set collected by C cameras, we use StarGAN [5] to train an image-image translation model to transfer images between every camera pair. We follow the same architecture as [5]. Specifically, the generator contains 2 convolutional layer, 6 residual blocks and 2 transposed convolution layers, while the discriminator is the same as PatchGANs [19]. The input images are resized to 128 × 64. In training, we use the Adam optimizer [20] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Two data augmentation methods, random flipping and random cropping, are employed. The learning rate is 0.0001 for both generator and discriminator at the first 100 epochs and linearly decays to zero in the remaining 100 epochs. In camera style transfer, for each image in the target set, we generate C style-transferred images (including the one transferred to the camera style of the original real image). These C fake images are regarded as containing the same person with original real image.

Re-ID model training. To train the re-ID model, we employ the training strategy in [48]. Specifically, we keep the aspect ratio of input images and resize them to 256×128. For data augmentation, random cropping and random flipping

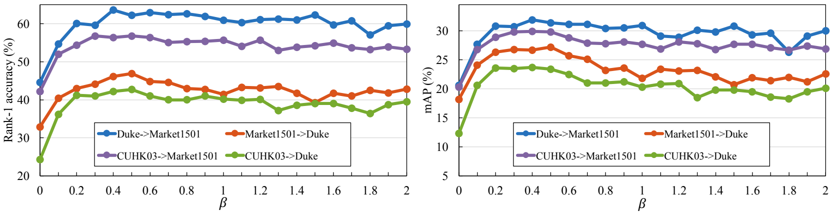

Fig. 5. Sensitivity to parameter β (weight of the triplet loss) in Eq. 6. We fix $n_t = 16$.

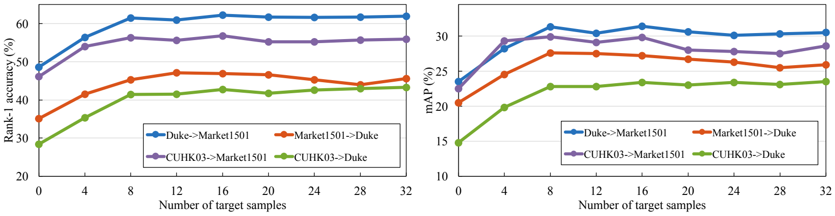

Fig. 6. Sensitivity to the number of real target images n_t in a batch. β is fixed to 0.5.

are applied. Dropout probability is set to 0.5. Learning rate is initialized to 0.1 for the classification layer and to 0.01 for the rest of the layers. Learning rate is divided by 10 after 40 epochs. We set the mini-batch size of source images to 128 and 64 for IDE and triplet loss, respectively. The model is trained with the SGD optimizer in a total of 60 epochs. In testing, we extract the output of the 2,048-dim Pool-5 layer as the image descriptor and use the Euclidean distance to compute the similarity between the query and database images.

4.3 Important Parameters

We evaluate two important parameters, *i.e.* the weight of the triplet loss β and the number of real target images n_t in a batch. When evaluating one parameter, we fix the other one. Results are shown in Fig. 5 and Fig. 6, respectively.

Weight of the triplet loss. When $\beta = 0$, our method reduces to the baseline (with cross-entropy loss only, Section 3.1). It is clearly shown that, our approach significantly improves the baseline at all values. The rank-1 accuracy and mAP improve with the increase of β and achieve the best results when β is between 0.4 to 0.8.

Number of the real target images in a training batch. When $n_t = 0$, only source images are used for training the re-ID model with IDE and triplet loss, so our method reduces to "baseline+$\mathcal{L}_T$". From Fig. 5, we observe that when increasing the number of real target images and their corresponding camera style transferred samples in a training batch, our method consistently outperforms "baseline+$\mathcal{L}_T$". Performance becomes stable after $n_t = 16$.

Based on the above analysis, our method is robust to parameters changes. In the following experiment, we set $\beta = 0.5$ and $n_t = 16$.

Table 2. Methods comparison using Duke / Market as source, and using Market / Duke as target. S: labeled source set, T: labeled target set, T^u: unlabeled target set.

Methods	Train set	Duke → Market-1501					Market-1501 → Duke				
		R-1	R-5	R-10	R-20	mAP	R-1	R-5	R-10	R-20	mAP
Basel.	T	83.8	93.3	95.6	97.1	66.3	72.3	84.1	88.1	90.9	53.5
Basel.	S	44.6	62.5	69.6	76.5	20.6	32.9	49.5	54.8	61.7	16.9
Basel.+$\mathcal{L}_T$	S	48.6	66.4	73.3	78.9	23.5	35.1	50.7	57.6	64.0	20.5
Basel.+$\mathcal{L}_D$	S+T^u	49.8	67.8	74.5	80.5	23.8	36.8	52.3	59.1	64.9	21.1
Basel.+$\mathcal{L}_C$	S+T^u	60.6	77.1	83.0	87.6	28.5	42.5	56.8	62.9	67.9	22.1
Basel.+$\mathcal{L}_{CD}$	S+T^u	**62.2**	**78.8**	**84.0**	**88.3**	**31.4**	**46.9**	**61.0**	**66.7**	**71.9**	**27.2**

4.4 Evaluation

Baseline accuracy. We present results of the baselines (see Section 3.1) in Table 2 and Table 3. When trained and tested both on the target set, high accuracy can be observed. However, performance drops significantly when the model is trained on the source set and directly deployed on the target set. For example, the baseline model trained and tested on Market-1501 yields a rank-1 accuracy of 83.8%, but drops to 44.6% when trained on DukeMTMC-reID and tested on Market-1501. The reason is the data distribution bias among datasets.

Effectiveness of domain connectedness learning over baseline. Because the loss function of domain connectedness learning in Eq. 4 includes both source labeled samples and unlabeled target samples, we first add triplet loss with source samples into baseline (Basel.+$\mathcal{L}_T$). As shown in Table 2 and Table 3, the performance of "Basel.+$\mathcal{L}_T$" is consistently improved in all settings. Specially, the rank-1 accuracy of "Basel.+$\mathcal{L}_T$" is increased from 42.2% to 46.1% when using CUHK03 as the source set and tested on Market-1501. Then, we inject domain connectedness learning into "Basel.+$\mathcal{L}_T$" by adding unlabeled target samples into triplet loss. Comparison to "Basel.+$\mathcal{L}_T$", when tested on Market-1501, "Basel.+$\mathcal{L}_D$" leads to +1.2% and +2.8% improvement in rank-1 accuracy when using Duke and CUHK03 as the source set, respectively.

Effectiveness of camera invariance learning over baseline. We verify the effectiveness of camera invariance learning over baseline in Table 2 and Table 3. It is clear that, "Basel.+$\mathcal{L}_C$" significantly outperforms the baseline in all settings. For example, when tested on Market-1501, "Basel.+$\mathcal{L}_C$" gives rank-1 accuracy of 60.6% when using Duke as source set. This is +16% higher than the baseline in rank-1 accuracy. Similar improvement is observed when tested on DukeMTMC-reID. The consistent improvement indicates that camera invariance learning is critical for improving the discriminate ability in target domain.

Benefit of Hetero-Homogeneous learning. We study the benefit of hetero-homogeneous learning in Table 2 and Table 3. The "Basel.+$\mathcal{L}_{CD}$" achieves higher performance than the model trained independently with camera invariance learning (Basel.+$\mathcal{L}_C$) or domain connectedness learning (Basel.+$\mathcal{L}_D$). For example, when Market-1501 is the target set, the "Basel.+$\mathcal{L}_{CD}$" obtains rank-1

Table 3. Comparison of various methods on unsupervised domain adaptation from CUHK03 to Market-1501 and DukeMTMC-reID (Duke).

Methods	Train set	CUHK03 → Market-1501					CUHK03 → Duke				
		R-1	R-5	R-10	R-20	mAP	R-1	R-5	R-10	R-20	mAP
Basel.	T	83.8	93.3	95.6	97.1	66.3	72.3	84.1	88.1	90.9	53.5
Basel.	S	42.2	59.1	66.1	73.8	20.3	24.3	38.2	45.0	51.9	12.3
Basel.+$\mathcal{L}_T$	S	46.1	63.8	71.1	78.1	22.5	28.4	43.4	49.6	55.9	14.8
Basel.+$\mathcal{L}_D$	S+T^u	48.9	66.7	74.6	79.6	23.3	29.2	44.5	50.7	57.5	15.7
Basel.+$\mathcal{L}_C$	S+T^u	53.6	71.0	77.6	82.7	25.6	40.9	55.9	60.9	66.2	20.8
Basel.+$\mathcal{L}_{CD}$	S+T^u	**56.8**	**74.7**	**81.4**	**86.3**	**29.8**	**42.7**	**57.5**	**64.2**	**69.1**	**23.4**

Table 4. Unsupervised person re-ID performance comparison with state-of-the-art methods.

Methods	Duke → Market-1501				Market-1501 → Duke			
	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
LOMO [23]	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
Bow [43]	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
UMDL [28]	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
PTGAN [39]	38.6	-	66.1	-	27.4	-	50.7	-
PUL [10]	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
SPGAN [7]	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
CAMEL [42]	54.5	-	-	26.3	-	-	-	-
SPGAN+LMP [7]	57.7	75.8	82.4	26.7	46.4	62.3	**68.0**	26.2
TJ-AIDL [38]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
HHL	**62.2**	**78.8**	**84.0**	**31.4**	**46.9**	**61.0**	66.7	**27.2**

accuracy in 56.8% by using CUHK03 as source set, surpassing the "Basel.+$\mathcal{L}_D$" and "Basel.+$\mathcal{L}_C$" by +7.9% and +3.2%, respectively. Similar improvement is observed in other settings, indicating that camera invariance and domain connectedness are indispensable to improve the transferability of the re-ID model in UDA.

4.5 Comparison with the state-of-the-art methods

We compare our method with the state-of-the-art unsupervised learning methods. Table 4 presents the comparison when Market-1501 / Duke is the source set and Duke / Market-1501 is the target. We compare with two hand-crafted features, *i.e.* BoW [43] and LOMO [23], three unsupervised methods, including CAMEL [42], PUL [10], and UMDL [28], and three unsupervised domain adaptation approaches, including PTGAN [39], SPGAN [7] and TJ-AIDL [38]. The two hand-crafted features are directly applied on target testing set without training. Both features fail obtain competitive results. With training on target set, unsupervised methods obtain higher results than hand-crafted features. For

Table 5. Unsupervised person re-ID performance comparison with state-of-the-art methods when trained on CUHK03.

Methods	CUHK03 → Market-1501					CUHK03 → Duke				
	R-1	R-5	R-10	R-20	mAP	R-1	R-5	R-10	R-20	mAP
PTGAN [39]	31.5	-	60.2	-	-	17.6	-	38.5	-	-
PUL [10]	41.9	57.3	64.3	70.5	18.0	23.0	34.0	39.5	44.2	12.0
SPGAN [7]	42.3	-	-	-	19.0	-	-	-	-	-
HHL	**56.8**	**74.7**	**81.4**	**86.3**	**29.8**	**42.7**	**57.5**	**64.2**	**69.1**	**23.4**

example, CAMEL [42] achieves 54.4% rank-1 accuracy when using DukeMTMC-reID as source set and tested on Market-1501 (multi-query setting). Comparing with unsupervised domain adaptation methods, our method is superior. Specifically, when tested on Market-1501, our results are higher than all the competing methods, achieving **rank-1 accuracy = 62.2% and mAP = 31.4%**. For example, comparing with the recently published TJ-AIDL method [38], our results are higher by +4.0% in rank-1 accuracy and +4.9% in mAP. When tested on DukeMTMC-reID, our method achieves **rank-1 accuracy = 46.9% and mAP = 27.2%**, higher than previous methods as well. So this paper sets a new state of the art on Duke → Market-1501 and yields competitive results on Market-1501 → Duke.

Table 5 presents comparisons of methods using CUHK03 as the source set. Our method outperforms the state-of-the-art methods by a large margin. Specifically, HHL yields an mAP of 29.8% when Market-1501 is the target set. This is higher than SPGAN [7] (19.0%) by +10.8%.

5 Conclusion

In this paper, we present Hetero-Homogeneous Learning (HHL), a new unsupervised domain adaptation approach for person re-identification (re-ID). Taking advantage of the unique challenges of UDA approaches in the context of person re-ID, we propose to learn camera invariance and domain connectedness simultaneously to obtain more generalized person embeddings on the target domain. Experiment conducted on Market-1501, DukeMTMC-reID and CUHK03 confirms that our approach achieves very competitive performance compared with the state of the art.

Acknowledgements This work is supported by the National Nature Science Foundation of China (No. 61572409, No. U1705286 & No. 61571188), Fujian Province 2011Collaborative Innovation Center of TCM Health Management, Collaborative Innovation Center of Chinese Oolong Tea Industry-Collaborative Innovation Center (2011) of Fujian Province, Fund for Integration of Cloud Computing and Big Data, Innovation of Science and Education, the Data to Decisions CRC (D2D CRC) and the Cooperative Research Centres Programme. Zhun Zhong thanks Wenjing Li for encouragement.

References

1. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification. CVIU (2013)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning (2010)
3. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: CVPR (2017)
4. Chen, M., Weinberger, K.Q., Blitzer, J.: Co-training for domain adaptation. In: Advances in neural information processing systems. pp. 2456–2464 (2011)
5. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR (2018)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
7. Deng, W., Zheng, L., Kang, G., Yang, Y., Ye, Q., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In: CVPR (2018)
8. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: CVPR (2018)
9. Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y.: Supervision-by-Registration: An unsupervised approach to improve the precision of facial landmark detectors. In: CVPR (2018)
10. Fan, H., Zheng, L., Yang, Y.: Unsupervised person re-identification: Clustering and fine-tuning. arXiv preprint arXiv:1705.10444 (2017)
11. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. JMLR (2016)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
13. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: ECCV (2008)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person reidentification. arXiv preprint arXiv:1703.07737 (2017)
16. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. arXiv preprint arXiv:1711.03213 (2017)
17. Hu, J., Lu, J., Tan, Y.P.: Deep transfer metric learning. In: CVPR (2015)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
19. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
20. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
22. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR (2014)

23. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR (2015)
24. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: NIPS (2016)
25. Liu, Z., Wang, D., Lu, H.: Stepwise metric promotion for unsupervised video person re-identification. In: ICCV (2017)
26. Long, M., Cao, Y., Wang, J., Jordan, M.: Learning transferable features with deep adaptation networks. In: ICML (2015)
27. Luo, Y., Zheng, Z., Zheng, L., Tao, G., Junqing, Y., Yang, Y.: Macro-micro adversarial network for human parsing. In: ECCV (2018)
28. Peng, P., Xiang, T., Wang, Y., Pontil, M., Gong, S., Huang, T., Tian, Y.: Unsupervised cross-dataset transfer learning for person re-identification. In: CVPR (2016)
29. Qin, D., Gammeter, S., Bossard, L., Quack, T., Van Gool, L.: Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In: CVPR (2011)
30. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCVW (2016)
31. Rohrbach, M., Ebert, S., Schiele, B.: Transfer learning in a transductive setting. In: NIPS (2013)
32. Saito, K., Ushiku, Y., Harada, T.: Asymmetric tri-training for unsupervised domain adaptation. arXiv preprint arXiv:1702.08400 (2017)
33. Sener, O., Song, H.O., Saxena, A., Savarese, S.: Learning transferrable representations for unsupervised domain adaptation. In: NIPS (2016)
34. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR (2014)
35. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: AAAI (2016)
36. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV Worksshops (2016)
37. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)
38. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR (2018)
39. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR (2018)
40. Wu, Y., Lin, Y., Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In: CVPR (2018)
41. Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W.: Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: CVPR (2017)
42. Yu, H., Wu, A., Zheng, W.S.: Cross-view asymmetric metric learning for unsupervised person re-identification. In: ICCV (2017)
43. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV (2015)
44. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
45. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: ICCV (2017)
46. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: CVPR (2017)

47. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
48. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: CVPR (2018)
49. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
50. Zhu, X.: Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison (2005)