One Shot Domain Adaptation for Person Re-Identification

Yang Fu\textsuperscript{1}, Yunchao Wei\textsuperscript{1}, Guanshuo Wang\textsuperscript{2}
Jiwei Li\textsuperscript{3}, Xi Zhou\textsuperscript{3,2}, Honghui Shi\textsuperscript{4,1}, Thomas Huang\textsuperscript{1}
\textsuperscript{1}IFP Group, Beckman, UIUC, \textsuperscript{2}Shanghai Jiao Tong University, \textsuperscript{3}CloudWalk Technology, \textsuperscript{4}IBM Research

Abstract

How to effectively address the domain adaptation problem is a challenging task for person re-identification (re-ID). In this work, we make the first endeavour to tackle this issue according to one shot learning. Given an annotated source training set and a target training set that only one instance for each category is annotated, we aim to achieve competitive re-ID performance on the testing set of the target domain. To this end, we introduce a similarity-guided strategy to progressively assign pseudo labels to unlabeled instances with different confidence scores, which are in turn leveraged as weights to guide the optimization as training goes on. Collaborating with a simple self-mining operation, we make significant improvement in the domain adaptation tasks of re-ID. In particular, we achieve the mAP of 71.5\% in the adaptation task of DukeMTMC-reID to Market1501 with one shot setting, which outperforms the state-of-arts of unsupervised domain adaptation more than 17.8\%. Under the five shots setting, we achieve competitive accuracy of the fully supervised setting on Market-1501. Code will be made available.

1. Introduction

Person re-identification (re-ID) aims at matching images of a person in one camera with the images of this person from another different cameras. Because of its important applications in security and surveillance, person re-ID has been drawing lots of attention from both academia and industry. Despite the dramatic performance improvement obtained by the convolutional neural network \cite{15,20,36,38}, it is reported that deep re-ID models trained on the source domain may have a large performance drop on the target domain \cite{8,10} due to data bias existing between source and target dataset. Since it is extremely expensive to label all images in target dataset, one of the most popular solutions for that problem is unsupervised domain adaptation (UDA).

The common UDA has been studied extensively in image classification, object detection, face recognition and semantic segmentation. \cite{4,5,27}. However, the traditional UDA approaches always have an assumption that the source and target domain share the same set of classes, which does not hold for person re-ID problem. For person re-ID problem, different datasets have totally different person identities (classes). Recently, several unsupervised domain adaptation approaches for person re-ID have been proposed and achieve some promising improvements. Some works aim to translate images from source domain to target domain based on generative adversarial network \cite{8,40} by preserving the annotation information of source domain. In addition, the disparities of cameras is another critical factor influencing re-ID performance, and HHL \cite{49} is proposed to address intra-domain image variations caused by different camera configurations. However, the performances of these UDA approaches are still unsatisfactory. Specifically, the state-of-art UDA methods are about 20\% to 30\% lower than the corresponding fully supervised baselines, making person re-ID limited in real world scenarios. Moreover, most existing person re-ID UDA methods need to generate lots of images with target domain style as the first step, which is time-consuming, especially, for the dataset with many cameras, and this kind of pre-processing step makes the model cannot be trained end-to-end.

To approach the performance of the fully-supervised
counterpart and efficiently achieve the adaption from the source domain to the target one, we make the first attempt of leveraging one shot learning to tackle this problem. In particular, one shot learning is based on the setting that only one sample from each category is labeled, which does not require much more human effort compared with the UDA and is also cheaper and feasible compared with its fully-supervised counterpart. Under such a setting, we propose a simple, efficient yet effective framework that is composed by two basic components, i.e., self-mining and one shot mining. Concretely, we employ the self-mining to learn the discriminative feature representations according to the appearances of samples, which is conducted in an unsupervised manner. For the one shot mining, we introduce a similarity-guided strategy to assign pseudo labels with confidence scores to all unlabeled samples and progressively update such scores as the training goes on. In this way, all the samples from the target domain can be employed for training from beginning, which can effectively boost the process of domain adaption.

By taking the advantages of both one shot mining and self-mining, we achieve the mAP scores of 71.5% and 55.9% on Market1501 and DukeMTMC-reID, which outperforms the state-of-the-arts more than 17.8% and 6.9% respectively. More importantly, under five shots setting, we recover more than 98% and 95% performance of fully supervised method on Market1501 and DukeMTMC-reID respectively, as shown in Fig 1. The contribution of this work can be summarized as following:

- We propose a simple yet effective one shot domain adaptation framework for person re-ID, which can recover the performance of its fully supervised counterpart with few annotations.
- We introduce a novel similarity-guided strategy for person re-ID one shot mining and integrate it into UDA framework, so that we can train two branches of unsupervised domain adaptation and one shot domain adaptation jointly and effectively boost the process of domain adaption.
- We conduct extensive experiments and ablation study on Market1501 [44] and DukeMTMC-ReID [31, 46] to demonstrate the effectiveness of one shot domain adaptation and each component.

2. Related Work

**Unsupervised domain adaptation.** Our work is closely related to unsupervised domain adaptation (UDA) where no data in target domain are labeled during training. Some works in this community try to address this problem by reducing the discrepancy between source domain and target domain [6, 34, 42]. For example, CORAL [34] learns a linear transformation that aligns the mean and covariance of feature distribution between two domains. And Sun [35] proposes deep CORAL to extend original approach to deep neural networks with a nonlinear transformation. Some other methods aim to learn a transformation to generate samples that are similar to target domains by adversarial learning approach [3, 25, 21]. Recently, some works solve this problem by mapping the source data and target data to a same feature space for the domain-invariant representations [16, 17, 27, 37]. For instance, Ganin et al. [16] propose a gradient reversal layer (GRL) and integrate it into standard deep neural network for minimizing the classification loss while maximizing domain confusion loss. However, most of existing unsupervised domain adaptation methods are based on an assumption that class labels are the same across domains, while the person identities of different re-ID datasets are entirely different. Hence, the approaches mentioned above cannot be utilized directly for person re-ID task.

**Unsupervised re-ID.** Some methods based on hand-craft features [2, 18, 24] can be directly applied for unsupervised person re-ID. However, these method always have a poor performance on large-scale dataset because they ignore the distribution of samples in the dataset. Benet from the success of deep learning, some recent works [8, 29, 39, 40] attempt to address unsupervised domain adaptation based on deep learning framework. Deng et al. [8] aim to translate images from source domain to target domain by proposed similarity preserving generative adversarial network (SPGAN). And the translated images are utilized to train re-ID model in a supervised way. In [39], a Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) is proposed to learn an attribute-semantic and identity discriminative feature representation space for target domain without using additional labeled data in target domain. In [49], Zhong et al. introduce a Hetero-Homogeneous Learning (HHL) method, which aims to improve the generalization ability of re-ID models on the target set by achieving camera invariance and domain connectedness simultaneously. Although these unsupervised domain adaptation approaches achieve promising progresses, the performance is still unsatisfactory compared with fully supervised approach.

**One shot re-ID.** One-shot learning aims at learning a task from one or very few training examples [11] and there are some works of one shot person re-ID [1, 13, 26, 41]. In [1], Bak et al. utilize a metric learning approach for a pair of cameras which can be split into texture and color components for one shot image-based re-ID. Wu et al. [41] propose a progressive sampling method to gradually predict reliable pseudo labels and update deep model for one shot video-based re-ID. To the best of our knowledge, there are no previous works on one shot domain adaptation for per-
son re-ID and existing one shot re-ID methods can hardly apply to the domain adaptation directly. Based on the above analysis, in this paper, we aim to address person re-ID domain adaptation with similarity-guided one shot learning approach.

3. Proposed Method

Problem Definition For one shot domain adaptation in person re-ID, we have a labeled source dataset \( S = \{X_s, Y_s\} \), which contains \( N_s \) person images and each image \( x_s \) has a corresponding label \( y_s \), where \( y_s \in \{1, 2, ..., P_s\} \), \( P_s \) is the number of identities in source dataset. Also, we have another target dataset \( T \), which consists \( N_t \) unlabeled images and can be split into two sub-datasets, \( T = T_L + T_U \), where \( T_L \) just has single labeled person image for each identity and \( T_U \) has large number of remaining unlabeled person images. The goal of this work is to leverage source dataset \( S \), labeled one shot target dataset \( T_L \) and unlabeled target dataset \( T_U \) to learn great discriminative embeddings of target dataset \( T \).

3.1. Fully Supervised Pre-training

Our proposed one shot domain adaptation framework is based on a model pre-trained on source dataset \( S \). In order to obtain the baseline model, we utilize ResNet50 [19] pre-trained on ImageNet [7] as backbone network. The fully connected (FC) layer is named as FC and the number of output channels is changed from 1000 to \( P_s \), where \( P_s \) is the number of identities in \( S \). And the outputs of global average pooling (GAP) and FC are noted as \( f_1 \) and \( f_2 \).

Given each labeled image \( x_s \) in source dataset and its ground truth identify \( y_s \), we train the baseline model with cross-entropy loss and hard-batch triplet loss [20]. Specifically, cross-entropy loss is employed with \( f_2 \) by casting the training process as a classification problem and hard-batch triplet loss is employed with \( f_1 \) by treating the training process as a verification problem. We name this model as baseline throughout this paper.

The baseline model achieves good performance with fully labeled data [14,45,47], but always fails when adopt to a new target dataset. Even some recent works on unsupervised domain adaptation still have a large performance gap compared with baseline model. In the following sections, we will describe proposed one shot domain adaptation method to reduce this gap and approach the baseline performance as close as possible.

3.2. One Shot Domain Adaptation in Person re-ID

The overview of proposed one shot domain adaptation in re-ID is shown in Fig 2. First, we train a baseline model based on configurations described in Sec 3.1. Then, we feed each person image of target dataset \( T \) into baseline model for feature extraction. With these feature vectors, we conduct domain adaptation by self-mining and similarity-guided one shot mining. For self-mining, we encourage the model to discover the similarities existing in the target dataset by iterative clustering. For one shot mining, instead of some step-wised approaches which exploit the unlabeled data in target dataset gradually, we propose a novel similarity-guided one shot mining approach, which can effectively boost the domain adaption process. Furthermore, we propose a training strategy to achieve the self-similarity mining and one shot mining simultaneously for a better re-ID performance. Finally, with the robust model obtained by self-similarity mining and one shot mining, we can assign each unlabeled person image a pseudo label with a high confidence and refine model by the baseline training configuration to progressively achieve better performance.

3.2.1 Domain Adaptation with Self-mining

Although the re-ID performance drops dramatically when directly adopted to another dataset, it is still much better than the performance of directly applying the ResNet50 pre-trained on ImageNet, which is almost zero. From this observation, we have every reason to believe that the model trained on source dataset still learns some useful representations of a person for the re-ID task.

The reason why it performs so badly on target dataset is that the similarities among different person images cannot be discovered correctly. In order to mine these similarities and make use of them for re-ID task, we employ a simple version of self-mining approach inspired by [33]. Specifically, we extract the feature of each person by baseline and then we obtain a feature space for every person in target dataset. Next, we employ a clustering approach [9] on the feature space to generate a series of clusters and every person image in each cluster can be assigned a pseudo label, called self-labels. Finally, we change the number of FC output channels from \( P_s \) to 128 and feed the \( f_1 \) and \( f_2 \) to train the self mining framework iteratively, as shown in Fig 2. As mentioned in [33], this self-mining approach is simple yet effective for unsupervised re-ID domain adaptation, which provides a good starting point for our following one shot mining.

3.2.2 Domain Adaptation with one shot mining

Although unsupervised domain adaptation for person re-ID has been studied extensively [8,40,49], recently, there is still more than 25% and 15% performance drop in mAP and Rank-1 accuracy comparing to fully supervised baseline, as shown in Fig 1. In order to narrow the huge performance gap, we propose to conduct domain adaptation approach in an one (few) shot learning manner.

For one shot setting, we have single labeled image for each identity, note as \( T_L = \{x^1_{tl}, x^2_{tl}, ..., x^P_{tl}\} \), where \( P_t \) is
Figure 2. Overview of the framework. Different colors indicate different categories according to target labels or self-labels. The CNN model is ResNet50 and initially trained on ImageNet. For each iteration, after feature extraction, we (1) assign each unlabeled image with a target-pseudo label by the similarity to all labeled images and obtain a confidence for each target-pseudo label. (2) Then, we assign each images (labeled and unlabeled) a self-pseudo label by clustering algorithm. (3) Next, we update the CNN model by minimizing the triplet loss of self-labeled images and target-labeled images. For each triplet loss, we compute it for twice, with feature vector after GAP and FC layer. Note that target-labeled images employ the same triplet loss with confidence term. For evaluation, we take the feature vector after GAP as the representation of query person image.

the number of identity of target dataset, and large unlabeled dataset $T_U : \{x_{tu}^1, x_{tu}^2, ..., x_{tu}^{Q_U}\}$, where $Q_U = N_t - P_t$. We first feed them into baseline model for feature extraction, then we use the feature vector after the GAP as the representation of a person, note as $F_{TL} : \{f_{t1}, f_{t2}, ..., f_{tP_t}\}$ and $F_{TU} : \{f_{tu1}, f_{tu2}, ..., f_{tuQ_U}\}$. Based on $F_{TL}$ and $F_{TU}$, we compute the similarity between each labeled image and each unlabeled image. Specifically, we employ the k-reciprocal encoding [47], which is a variation of Jaccard distance between nearest neighbors sets, as the distance metric for similarity measure, where more distance means less similarity. We denote the similarity matrix as $D$ with size of $P_t \times Q_t$.

$$D = \begin{bmatrix}
d_{11} & d_{12} & d_{13} & ... & d_{1Q_t} \\
d_{21} & d_{22} & d_{23} & ... & d_{2Q_t} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
d_{P_t1} & d_{P_t2} & d_{P_t3} & ... & d_{P_tQ_t}
\end{bmatrix}. \quad (1)$$

With the similarity matrix $D$, we assign a pseudo target label $y_{ui}$ to each unlabeled image $x_{ui}$ by corresponding the most similar labeled image

$$y_{ui} = \arg \min_{j=1, \ldots, P_t} d_{ji}. \quad (2)$$

However, it is obvious that the assigned pseudo labels may not be correct and they have different confidences on whether they are right or not, so it is not a good way to treat all pseudo labels as the same. In order to address the variance among different pseudo labels, we introduce a confidence term to measure the probability of statement that the assigned pseudo label is correct, which is described as Eqn 3. Recall that the larger distance means less similarity, so the unlabeled image who has the largest distance to a specific labeled image is the most unlikely sample to share the same identity. We denote the maximum distance corresponding to image $x_{ui}$ as $d_{max(i)}$. Furthermore, if the minimum distance used for pseudo label assigning is closer to $d_{max(i)}$, this pseudo label is more likely to be wrong and the confidence term $w_{ui}$ is closer to zero.

$$\begin{cases}
d_{max(i)} = \max_{k=1, \ldots, Q_t} d_{ui,k} \\
w_{ui} = 1 - \frac{d_{ui}}{d_{max(i)}} \quad (3)
\end{cases}$$

Since self-mining and one shot mining share the same feature space, we then design a simple yet effective way to train the whole framework jointly and end-to-end. As shown in Fig 2 after feature extraction on each unlabeled image, $f_1$ is feed into two branch: self-mining and one shot mining. After assigning self pseudo labels and target pseudo labels, the self mining branch uses the hard-batch triplet loss [20] as object function and one shot mining branch employs the same triplet loss with confidence term. Meanwhile, $f_2$ is used for training as well.
3.2.3 Model Refinement

After training the one shot domain adaptation framework, we adapt an additional step of model refinement for achieving better performance. Since the model obtained by one shot domain adaptation framework has already performed very well on the target dataset, the assigned target pseudo labels are more likely to be right finally. Based on this assumption, we repeat the one shot mining branch once for target label assignment and treat them as the ground truth labels, then follows the baseline training model configuration for further model refinement.

3.3. Loss Function

Baseline and Model Refinement. For the baseline model and model refinement, they share the same training configuration. As describe in Sec 3.1 we utilize both the batch-hard triplet loss proposed in [20] and the softmax loss jointly. The triplet loss with hard mining is first proposed in [20] as an improved version of the original semi-hard triplet loss [32]. We randomly sample $P$ identities and $K$ instances for each mini-batch to meet the requirement of the batch-hard triplet loss. Typically, the loss function is formulated as follows:

$$L_{triplet} = \sum_{i=1}^{P} \sum_{a=1}^{K} \left[ \alpha + \max_{p=1\ldots K} \left| \left| w_{x_a} \ast x_a^{(i)} - w_{x_p} x_p^{(i)} \right| \right|_2 ight]$$

(4)

where $x_a^{(i)}$, $x_p^{(i)}$, $x_n^{(i)}$ are features extracted from the anchor, positive and negative samples respectively, and $\alpha$ is the margin hyperparameter. Besides batch-hard triplet loss, we employ softmax cross entropy loss for discriminative learning as well, which can be formulated as follows:

$$L_{softmax} = - \sum_{i=1}^{P} \sum_{a=1}^{K} \log \frac{e^{w^T y_{i,a} x_{a,i}}}{\sum_{k=1}^{C} e^{W^T y_{k,a} x_{a,i}}}$$

(5)

where $y_{i,a}$ is the ground truth identity of the sample $(a, i)$, and $C$ is number of identity. Our loss function for optimization is the combination of softmax loss and batch-hard triplet loss as follows:

$$L_{baseline} = L_{softmax} + L_{triplet}$$

(6)

One Shot Learning Domain Adaptation For the one shot mining and self-mining, we just leverage the hard-batch triplet loss for metric learning and the only difference is that we introduce a confidence term to original hard-batch triplet loss for one shot mining, which can be formulated as follows:

$$L_{oneshot} = \sum_{i=1}^{P} \sum_{a=1}^{K} \left[ \alpha + \max_{p=1\ldots K} \left| \left| w_{x_a} \ast x_a^{(i)} - w_{x_p} x_p^{(i)} \right| \right|_2 ight]$$

$$- \min_{j=1\ldots P, j \neq i} \left| \left| w_{x_a} \ast x_a^{(i)} - w_{x_p} x_p^{(i)} \right| \right|_2 + \right]$$

(7)

Hence, our object function for one shot domain adaption framework is

$$L_{dp} = L_{triplet} + L_{oneshot}.$$  

(8)

4. Experiments

4.1. Datasets and Evaluation Protocol

In this section, we evaluate the proposed method on three re-ID datasets which are considered as large scale in the community, i.e. Market1501 [44] and DukeMTMC-ReID [31, 46] and MSMT17 [40].

Market1501 [44] contains 32,668 images of 1,501 labeled persons of six camera views. Specifically, 12,936 person images from 751 identities detected by DPM [12] are used for training and 19,732 person images from 750 identities plus some distractors form gallery set. In addition, 3,368 hand-drawn bounding boxes from 750 identities are used as query set to retrieve the corresponding person images in the gallery set.

DukeMTMC-ReID [46] is a subset of the DukeMTMC dataset [31]. It contains 1,812 identities captured by 8 cameras. There are 2,228 query images, 16,522 training images and 17,661 gallery images, with 1,404 identities appear in more than two cameras. Also, similar with the Market1501, the rest 408 identities are considered as distractor images.

MSMT17 [40] is the largest re-ID dataset, which contains 126,441 bounding boxes of 4,101 identities taken by 15 cameras during 4 days. These 15 cameras contain 12 outdoor cameras and 3 indoor cameras. And Faster RCNN [30] is utilized for pedestrian bounding box detection. The MSMT17 can be viewed as the most challenging re-ID dataset up to now with so many images under so many cameras.

Evaluation Protocol In our experiment, we use Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP) to evaluate the performance of re-ID. CMC represents the accuracy of the person retrieval, it is accurate when each query only has one ground truth. However, when multiple ground truths exist in the gallery, the goal is to return all right matches to the user. In this case, CMC may not have enough discriminative ability, but the mAP could reflect the recall. For Market-1501 and
DukeMTMC-ReID, we use the evaluation packages provided by [44] and [46], respectively. Moreover, for simplicity, all results reported in this paper are under the single-query setting and do not use the re-ranking proposed in [47] as post-processing.

4.2. Implementation Details

Baseline training As described in Section 3.1, we first train a baseline model on source dataset and follow the training strategy described in [50]. Specifically, we keep the size of input images and resize them to 256 × 128. For data augmentation, we employ random cropping, random flipping and random erasing [48]. To meet the requirement of hard-batch triplet loss, each mini-batch is sampled with randomly selected $P = 16$ identities and randomly sampled $K = 8$ images for each identity from the training set, so that the mini-batch size is 128. And in our experiment, we set the margin parameter to 0.5. During training, we use the Adam [22] with weight decay 0.0005 to optimize the parameters for 150 epochs. The initial learning rate is set to $3e^{-4}$ and decays to $3e^{-5}$ after first 100 epochs.

One shot domain adaptation learning. Since there is at least one image for each identity under each camera, we choose the first image taken by the first camera of each identity as the labeled data. And for fairness, we preserve them for all following experiments. During training, we follow the same settings of data augmentation and triplet loss to train the one shot adaptation framework. The number of FC layer output channels is set to 128. And we change the initial learning rate from $3e^{-4}$ to $6e^{-5}$ and training epoch from 150 to 70. In addition, the whole framework is trained for several iterations until convergence.

Our model is implemented on Pytorch [28] platform and trained with two NVIDIA TITAN X GPUs. All our experiments on different datasets follow the same settings as above.

4.3. Ablation Study

Comparison between supervised learning, direct transfer and state-of-arts unsupervised method. The performance of supervised baseline method and the direct transfer method are specified in Table I. When comparing two methods, we can clearly find that there is large performance drop when directly adopting source-trained model on target dataset. For instance, the baseline model trained on Market1501 tested on Market1501 achieves 92.5% in rank-1 accuracy and 80.8% in mAP, but it drops to 26.9% and 13.7% when tested on DukeMTMC-reID, where the performance gap is more than 60%. And the similar drop can be observed when DukeMTMC-reID is used as training set and tested on Market1501.

Although, lots of recent unsupervised based domain adaptation works proposed to address this performance drop, the performance on target is still unsatisfactory. For example, when training set is DukeMTMC-reID and testing set is Market1501, to our best knowledge, the best UDA approach achieves 53.7% and 75.8% on mAP and rank-1 accuracy, which is lower than the fully supervised method by about 20%, and it is described in Table 2 [8] as well. By our proposed one shot domain adaptation framework, we achieve 71.5% and 87.5% on mAP and rank-1 accuracy respectively when trained on DukeMTMC-reID while tested on Market1501, which is only 9% and 5% lower than baseline model. Moreover, when we extend the one shot framework to five shots setting, the performance on mAP and rank-1 accuracy is 78.8% and 91.5%, which is extremely close to the baseline model.

The effectiveness of self mining. As described in Section 3.2.1, the first step of one shot domain adaptation framework is mining the useful information existing in pre-trained model. As shown in Table I only with self mining, we improve the performance by +19.3% and +24.1 in mAP and rank-1 accuracy when trained on DukeMTMC-reID and tested on Market1501. When trained on Market1501 and tested on DukeMTMC-reID, the performance gain is +26.8% and +42.4% in rank-1 accuracy and mAP, respectively. Through this method, we can learn a more effective and strong baseline model for domain adaptation.

The effectiveness of one shot domain adaptation. We conduct several experiments to verify the influence of the one shot domain adaptation on performance in Table I. First, we only use the one shot labeled data based on model trained by self mining. However, without any data exploration strategies, there are no gains on performance but even a little drop on DukeMTMC-reID. This is because during training, the framework only see one shot rather that the whole target dataset. Then, we employ our proposed similarity-guided one shot domain adaptation approach and it is clear that we achieve better result on both datasets. For instance, compared with baseline, we gain +37% and +29% on mAP and rank-1 accuracy when tested on Market1501. Also, comparing to one shot learning without similarity-guided strategy, we improve mAP and rank-1 accuracy by +7.5% and +5.4%.

In addition, we compare proposed similarity-guided one
The effectiveness of joint training strategy As described in Section 3.2.2, we further train it with self mining together but no on the top of model by trained by self mining. From Table 1 we gain +7.7% and +4.5% in mAP and rank-1 accuracy, respectively, when trained on DukeMTMC-reID and tested on Market1501. When tested on DukeMTMC-reID, the gains are +2.8% and +1.6% in mAP and rank-1 accuracy, respectively. By jointly training strategy, we can not only improve the reID performance on both dataset, but also save the training time.

The effectiveness of model refinement As shown in Fig 4, the accuracy of pseudo label prediction during training iterations on Market1501 and DukeMTMC-reID figureatons, as described in Section 3.2.3. From Table 1 the model refinement improves mAP and rank-1 accuracy by 3.1% and 3.4% on Market1501.

So far all components in one shot domain adaptation framework have been evaluated and validated, we achieve promising performances on both Market1501 and DukeMTMC-reID. For instance, we achieve 46.8% and 36.9% improvements in mAP and rank-1 accuracy when trained on DukeMTMC-reID and tested on Market1501.

Extension of few shot settings Our proposed one shot domain adaptation framework can be easily extended to few shots setting, i.e., three shots, five shots, and we conduct experiment to evaluate the performance under those few shot setting. As shown in Table 1 the performance are improved consistently with increasing of labeled data. Compared with one shot setting, we achieve +4.0% and 7.3% in mAP over Market1501 under three shots and five shots setting, respectively. Also, the gains are 5.8% and 7.9% under three shot and five shots setting when tested on DukeMTMC-reID. Moreover, the performance under five shots setting is extremely close to fully supervised baseline method. Specif-
| Methods       | mAP | R1  | R5  | R10 |
|--------------|-----|-----|-----|-----|
| LOMO [24]    | 8.0 | 27.2| 41.6| 49.1|
| Bow [44]     | 14.8| 35.8| 52.4| 60.3|
| UMDL [29]    | 12.4| 34.5| 52.6| 59.6|
| PTGAN [40]   | -   | 38.6| -   | 66.1|
| PUL [10]     | 20.5| 45.5| 60.7| 66.7|
| SPGAN [8]    | 22.8| 51.5| 70.1| 76.8|
| CAMEL [43]   | 26.3| 54.5| -   | -   |
| SPGAN+LMP [8]| 26.7| 57.7| 75.8| 82.4|
| TJ-AIDL [39] | 26.5| 58.2| 74.8| 81.1|
| HHL [49]     | 31.4| 62.2| 78.8| 84.0|
| ARN [23]     | 39.4| 70.3| 80.4| 86.3|
| UDAP [43]    | 53.7| 75.8| 89.5| 93.2|
| Ours(one)    | 71.5| 87.5| 95.2| 96.8|
| Ours(five)   | 78.8| 91.5| 96.9| 98.1|

Table 2. Comparison of proposed one shot domain adaptation with state-of-arts unsupervised domain adaptive person re-ID methods on Market1501 dataset.

| Methods       | mAP | R1  | R5  | R10 |
|--------------|-----|-----|-----|-----|
| LOMO [24]    | 4.8 | 12.3| 21.3| 26.6|
| Bow [44]     | 8.3 | 17.1| 28.8| 34.9|
| UMDL [29]    | 7.3 | 18.5| 31.4| 37.4|
| PTGAN [40]   | -   | 27.4| -   | 50.7|
| PUL [10]     | 16.4| 30.0| 43.4| 48.5|
| SPGAN [8]    | 22.3| 41.1| 56.6| 63.0|
| CAMEL [43]   | -   | -   | -   | -   |
| SPGAN+LMP [8]| 26.2| 46.4| 62.3| 68.0|
| TJ-AIDL [39] | 23.0| 44.3| 59.6| 65.0|
| HHL [49]     | 27.2| 46.9| 61.0| 66.7|
| ARN [23]     | 33.4| 60.2| 73.9| 79.5|
| UDAP [43]    | 49.0| 68.4| 80.1| 83.5|
| Ours(one)    | 55.9| 72.4| 84.0| 87.7|
| Ours(five)   | 63.8| 78.8| 89.2| 92.2|

Table 3. Comparison of proposed one shot domain adaptation with state-of-arts unsupervised domain adaptive person re-ID methods on DukeMTMC-reID dataset.

4.4. Comparison with State-of-arts

To our best knowledge, there are no previous works on one shot domain adaptation for person re-ID, so we compare the proposed method with the state-of-the-art unsupervised learning methods on Market1501, DukeMTMC-reID and MSMT17 in Table 2, Table 3 and Table 4 respectively.

**Results on Market1501** On Market-1501, we compare our results with two hand-crafted features, *i.e.* Bag-of-Words (BoW) [44] and local maximal occurrence (LOMO) [24], three unsupervised methods, including UMDL [29], PUL [10] and CAMEL [43], and five unsupervised domain adaptation methods, including PTGAN [40], SPGAN [8], TJ-AIDL [39], ARN [23] and UDAP [33]. The two hand-crafted features are directly applied on test dataset without any training process, but it is obvious that both features fail to obtain competitive results. With training on target set, unsupervised methods always obtain higher results than hand-crafted features. Comparing with unsupervised domain adaptation methods, our method is superior. In the one shot setting, we achieve rank-1 accuracy $= 87.5\%$ and mAP $= 71.5\%$, which outperforms the best unsupervised method [33] by $11.7\%$ and $+17.8\%$. In addition, in the five shots setting, we achieve $78.8\%$ and $91.5\%$ in mAP and rank-1 accuracy, which are $25.1\%$ and $+15.7\%$ higher than all other unsupervised domain adaptation methods. The comparisons indicate the competitiveness and effectiveness of the proposed method on Market-1501.

**Results on DukeMTMC-reID** The similar improvement can also be observed when we tested on DukeMTMC-reID dataset. Specifically, we achieve mAP $= 55.9\%$ and rank-1 accuracy $= 72.4\%$ under one shot setting and mAP $= 63.8\%$ and rank-1 accuracy $= 78.8\%$ under five shots setting. Compared with best unsupervised method, our result is $+6.9\%$ and $14.8\%$ higher in mAP. Therefore, the superiority of proposed one shot domain adaptation for person re-ID can be concluded.

**Results on MSMT17** In addition, we further evaluate proposed one shot domain adaptation approach on MSMT17 dataset, which is the largest and most challenging re-ID dataset. Under one shot setting, we achieve mAP $= 23.6\%$ and rank-1 accuracy $= 43.6\%$ when trained DukeMTMC-reID, which is $20.3\%$ and $32.8\%$ higher than the state-of-art. And similar improvement can be observed when trained on Market1501 as well.

5. Conclusion

In this work, we made the first endeavour to tackle the challenging domain adaptation person re-ID by one (few) shot learning. Different from the common practice of adopting the step-wise strategy to iteratively assign pseudo labels to unlabeled samples, we introduced a similarity-guided strategy that enables the network to “see” all the samples of target domain starting from the training, leading to the parameters can be quickly adapted from the source domain to the target domain. Extensive experimental results demonstrated that the performance of our approach outperformed the state-of-the-arts by a large margin.
Acknowledgements. This work is in part supported by IBM-Illinois Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00341. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. Fu is supported by CloudWalk Technology.

References

[1] S. Bak and P. Carr. One-shot metric learning for person re-identification. In IEEE CVPR, pages 1571–1580, 2017.
[2] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding, 2013.
[3] K. Boumaisalos, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In IEEE CVPR, 2017.
[4] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In IEEE CVPR, 2018.
[5] Y. Chen, W. Li, and L. Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In IEEE CVPR, 2018.
[6] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell. Best practices for fine-tuning visual classifiers to new domains. In ECCV, pages 435–442, 2016.
[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE CVPR, pages 248–255, 2009.
[8] W. Deng, L. Zheng, G. Kang, Y. Yang, Q. Ye, and J. Jiao. Image-domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In IEEE CVPR, 2018.
[9] M. Ester, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In IEEE CVPR, 2017.
[10] Y. Fang, L. Zheng, C. Yan, and Y. Yang. Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2018.
[11] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. IEEE TPAMI, 2006.
[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE TPAMI, 2010.
[13] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino. Semi-supervised multi-feature learning for person re-identification. In AVSS, pages 111–116, 2013.
[14] Y. Fu, X. Wang, Y. Wei, and T. Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In AAAI, 2019.
[15] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang. Horizontal pyramid matching for person re-identification. In AAAI, 2019.
[16] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In IEEE ICML, 2015.
[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 2016.
[18] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In ECCV, 2008.
[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In IEEE CVPR, pages 770–778, 2016.
[20] A. Herrmans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.
[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. arXiv preprint, 2017.
[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
[23] Y.-J. Li, F.-E. Yang, Y.-C. Liu, Y.-Y. Yeh, X. Du, and Y.-C. F. Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In IEEE CVPRW, 2018.
[24] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In IEEE CVPR, pages 2197–2206, 2015.
[25] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In NIPS, pages 459–477, 2016.
[26] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu. Semi-supervised coupled dictionary learning for person re-identification. In IEEE CVPR, pages 3550–3557, 2014.
[27] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. Unified deep supervised domain adaptation and generalization. In IEEE ICCV, volume 2, page 3, 2017.
[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In NIPS-W, 2017.
[29] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian. Unsupervised cross-dataset transfer learning for person re-identification. In IEEE CVPR, pages 1306–1315, 2016.
[30] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 2015.
[31] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In ECCV, 2016.
[32] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In IEEE CVPR, 2015.
[33] L. Song, C. Wang, L. Zhang, B. Du, Q. Zhang, C. Huang, and X. Wang. Unsupervised domain adaptive re-identification:
Theory and practice. *arXiv preprint arXiv:1807.11334*, 2018.

[34] B. Sun, J. Feng, and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[35] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016.

[36] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. Beyond part models: Person retrieval with refined part pooling. *arXiv preprint arXiv:1711.09349*, 2017.

[37] E. Tzeng, C. Devin, J. Hoffman, C. Finn, P. Abbeel, S. Levine, K. Saenko, and T. Darrell. Adapting deep visuo-motor representations with weak pairwise constraints. *arXiv preprint arXiv:1511.07111*, 2015.

[38] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2019.

[39] J. Wang, X. Zhu, S. Gong, and W. Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE CVPR*, 2018.

[40] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE CVPR*, 2018.

[41] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *IEEE CVPR*, 2018.

[42] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[43] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE ICCV*, 2017.

[44] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *IEEE ICCV*, pages 1116–1124, 2015.

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

[46] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017.

[47] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *IEEE CVPR*, pages 3652–3661, 2017.

[48] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

[49] Z. Zhong, L. Zheng, S. Li, and Y. Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, pages 172–188, 2018.

[50] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *IEEE CVPR*, pages 5157–5166, 2018.