

Self-similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-identification

Yang Fu¹, Yunchao Wei^{1,2}, Guanshuo Wang³, Yuqian Zhou¹
 Honghui Shi^{4,1,5}, Thomas S. Huang¹

¹University of Illinois at Urbana-Champaign, ²ReLER, University of Technology Sydney,

³Shanghai Jiao Tong University, ⁴IBM Research, ⁵University of Oregon

Abstract

Domain adaptation in person re-identification (re-ID) has always been a challenging task. In this work, we explore how to harness the similar natural characteristics existing in the samples from the target domain for learning to conduct person re-ID in an unsupervised manner. Concretely, we propose a Self-similarity Grouping (SSG) approach, which exploits the potential similarity (from the global body to local parts) of unlabeled samples to build multiple clusters from different views automatically. These independent clusters are then assigned with labels, which serve as the pseudo identities to supervise the training process. We repeatedly and alternatively conduct such a **grouping** and **training** process until the model is stable. Despite the apparent simplify, our SSG outperforms the state-of-the-arts by more than 4.6% (DukeMTMC→Market1501) and 4.4% (Market1501→DukeMTMC) in mAP, respectively. Upon our SSG, we further introduce a clustering-guided semi-supervised approach named SSG⁺⁺ to conduct the one-shot domain adaption in an open set setting (i.e. the number of independent identities from the target domain is unknown). Without spending much effort on labeling, our SSG⁺⁺ can further promote the mAP upon SSG by 10.7% and 6.9%, respectively. Our Code is available at: <https://github.com/OasisYang/SSG>.

1. Introduction

Person re-identification (re-ID) aims at matching images of a person in one camera with the images of this person from other different cameras. Because of its essential 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 [20, 36, 38], it is reported that deep re-ID models trained on the source domain

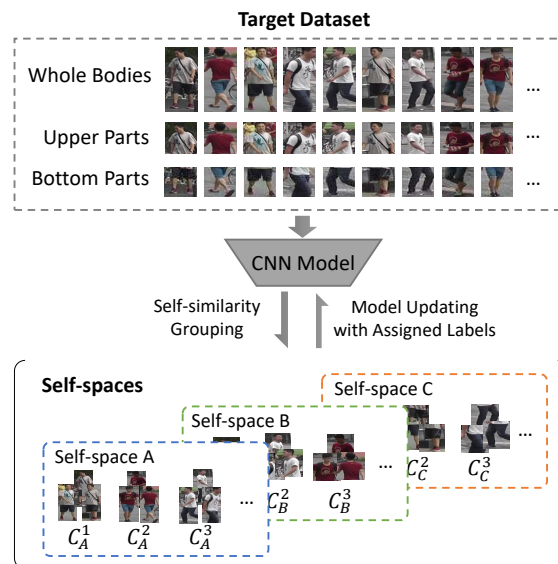


Figure 1. Illustration of proposed Self-similarity Grouping (SSG). We group target images by three cues, whole bodies, upper parts and bottom parts, independently and assign labels according to the corresponding group. The global information of the whole body and the local information of body parts can help us learn a better representation of individuals.

may have a significant performance drop on the target domain [8, 10] due to the data-bias existing between source and target datasets. Since it is costly and unfeasible to annotate all images in target dataset, one of the most popular solutions for such issue is unsupervised domain adaptation (UDA).

Currently, the common UDA has been studied extensively in image classification, object detection, face recognition and semantic segmentation. However, the traditional UDA approaches [4, 5, 27] always have an assumption that the source and target domain share the same set of classes, which does not hold for the person re-ID problem. Notably, in the person re-ID, different datasets have different identities (i.e. classes). Recently, several unsupervised domain

adaptation approaches for person re-ID have been proposed and achieve some promising improvements. Some works aim to translate the appearance of images from the source domain to the target domain based on the generative adversarial networks [8, 40] by preserving the annotation information of the source domain. Besides, the disparities of cameras is another critical factor influencing re-ID performance, and HHL [49] is proposed to address intra-domain image variations caused by different camera configurations. However, the performances of these UDA approaches are still far behind their fully-supervised counterparts. The main reason is that most previous works focus on increasing the training samples or comparing the similarity or dissimilarity between the source dataset and the target dataset but ignoring the similar natural characteristics existing in the training samples from the target domain.

In order to address the problem above and discover the similarities among person images in target dataset, we proposed unsupervised Self-similarity Grouping (SSG) to mine the potential similarities from global to local manner. The critical idea of SSG arises from some recent re-ID works based on part matching, where different parts contain different discriminative information of a person. Fig 1 illustrates our proposed SSG approach. In particular, we extract features of all persons in target dataset and group them by three different cues, whole bodies (A), upper parts (B) and lower parts (C) independently. Then, we can obtain three sets of groups: C_A, C_B, C_C . By assigning a pseudo label to each group, we can pair every person with different pseudo labels. For instance, given a person x_i , it should be assigned by three pseudo labels, C_A^i, C_B^i and C_C^i . As a result, we can establish a new dataset with pseudo labels, which can be treated as the normally labeled dataset. Since individuals with the same pseudo label should share lots of similarities, we iteratively mine these by finetuning the pre-trained model with the established dataset.

Upon our SSG, we further present a semi-supervised solution based clustering-guided annotation to approach the performance of the fully-supervised counterpart and efficiently achieve the adaption from the source domain to the target one. It is no more natural for us to think of the simplest semi-supervised solution, *i.e.* one shot learning. In particular, the traditional one shot learning is based on the setting that only one sample from each category is labeled. However, the traditional one shot setting is not suitable for person re-ID case. Unlike image recognition problem, which is usually based on a closed set assumption, person re-ID problem is actually an open set problem. In other words, we cannot know in advance how many identities are included in a given unlabeled target dataset. Thus, the superior characteristics from traditional one shot setting cannot be directly applied to the re-ID case. To tackle the above-mentioned issue, we innovatively provide

a clustering-guided semi-supervised solution. The proposed semi-supervised training strategy is based on the clustering-guided annotations, which samples a single image from each clustering. By doing this, we can significantly avoid choosing the same identity as two different ones. Hence, it allows us to leverage traditional one shot learning method and achieve similar performance.

We summarize our **contributions** as follows:

- We propose the Self-similarity Grouping (SSG) method, which is a simple yet effective unsupervised domain adaptation (UDA) framework for person re-ID, in order to recover the performance of its fully supervised counterpart.
- We introduce a similarity-guided semi-supervised training strategy for person re-ID and integrate it into the UDA framework, so that we can train unsupervised branch and semi-supervised branch jointly and effectively boost the process of domain adaption.
- We conduct extensive experiments and ablation study on several popular benchmarks including Market1501 [45], DukeMTMC-ReID [31, 46] and MSMT [40], to demonstrate the effectiveness of proposed SSG and semi-supervised solution.

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 the 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 a 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.

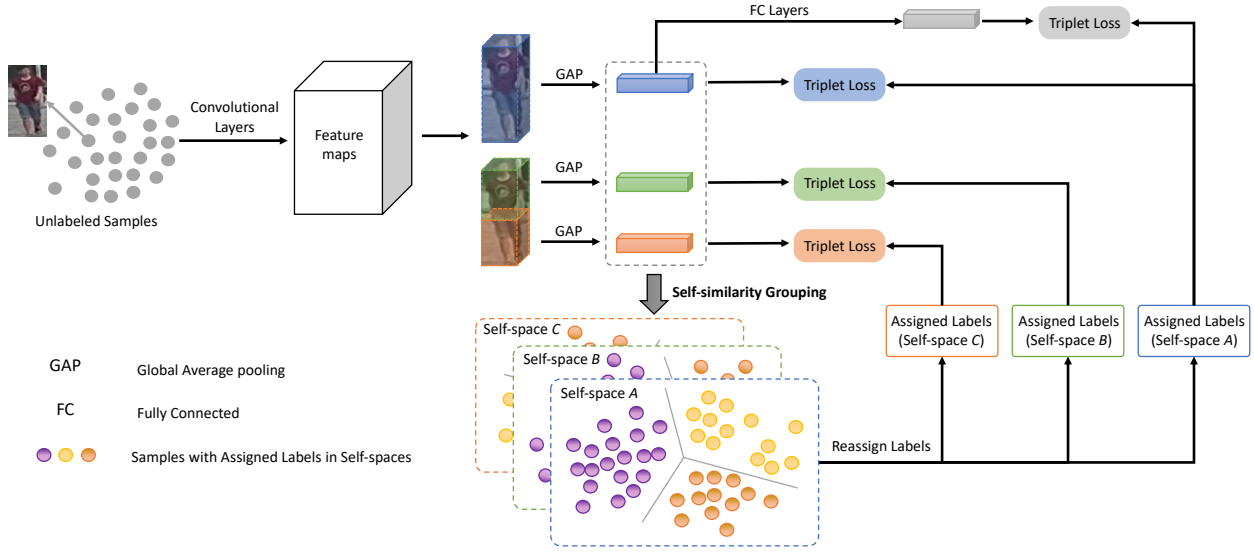


Figure 2. Overview of the proposed SSG approach. The CNN model is ResNet50 and pre-trained on source dataset. For each iteration, after feature extraction, we (1) split the feature maps into an upper part and a lower part and employ GAP on the whole, upper and lower feature maps. (2) Then, we group person images with different feature representations (blue, green and orange) and assign each different self-pseudo labels by grouping results (A, B, C). (3) Next, we update the CNN model by minimizing the triplet loss with each pseudo label. (4) During testing, we concatenate three feature representations together as the final representation of query person.

Unsupervised re-ID. Some methods based on hand-craft features [2, 18, 24] can be directly applied for unsupervised person re-ID. However, these methods always have a poor performance on large-scale dataset because they ignore the distribution of samples in the dataset. Benefit 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 the source domain to the target domain by proposed similarity preserving generative adversarial network (SPGAN). And the translated images are utilized for training 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 the 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 progress, the performance is still unsatisfactory compared with the fully supervised approaches.

Semi-supervised re-ID. Semi-supervised 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. However, the previous one shot re-ID works actually does not make sense. As described in the previous section, re-ID problem is an open set problem which means we cannot know how many identities in that dataset, so we cannot achieve the one/few shot setting under this situation. Based on the above analysis, in this paper, we aim to address person re-ID domain adaptation by Self-similarity grouping approach and further improve the performance by clustering-guided semi-supervised training.

3. Proposed Method

Problem Definition For unsupervised domain adaptation in person re-ID, we have a labeled source dataset $\{X_S, Y_S\}$, which contains N_s person images and each image x_s^i has a corresponding label y_s^i , where $y_s^i \in \{1, 2, \dots, P_s\}$. P_s is the number of identities in the source dataset. Also, we have another target dataset $\{X_t\}$, which consists of N_t person images. Note that the identity of each image x_t in target dataset $\{X_t\}$ is unknown. The goal of UDA person re-ID is to learn great discriminative embeddings of target dataset by only using the supervised information of source dataset.

3.1. Fully Supervised Pre-training

Many existing UDA approaches are based on a model pre-trained on source dataset, and we follow the similar set-

ting in [14, 47, 49, 51] to obtain the pre-trained model. In particular, we first utilize ResNet50 [19] pre-trained on ImageNet [7] as backbone network. The last fully connected (FC) layer is discarded and two additional FC layers are added. The first one has 2048 dimensions named as “FC-2048”. The output of second FC layer is P_s dimensional, where P_s is the number of identity in source dataset, named as “FC-#ID”. Given each labeled image x_s in the 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, the cross-entropy loss is employed with “FC-#ID” by casting the training process as a classification problem and hard-batch triplet loss is employed with “FC-2048” by treating the training process as a verification problem. We name this model as **baseline** in this paper. The **baseline** model achieves good performance with fully labeled data, but always fails when adopted to a new dataset.

3.2. Unsupervised Self-similarity Grouping

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 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 severely 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 the re-ID task, we propose Self-similarity Grouping (SSG) approach. The overview of proposed SSG approach is shown in Fig 2. The motivation of SSG is that we aim to encourage the model to discover the similarities existing in target dataset by self-similarity grouping. Then each unlabeled person is assigned with a pseudo label based on grouping results, which can be further used to reconstruct the target dataset and fine-tune the **baseline** model. Inspired by recent re-ID work [15, 38], we compare the similarities between two persons not only by global information obtaining from the whole body but also by more fine and local information getting from upper and lower parts of a person. By combining the global and local information, we can obtain a more robust and discriminative representation of a person, which is more informative for self-similarity grouping.

To formulate the proposed SSG algorithm, we first feed each unlabeled images x_t^i in target dataset into the **baseline** model trained by configurations described in Sec 3.1 for feature extraction, denote as $F_t^i \in \mathbb{R}^{H \times W \times C}$ (blue one in Fig 2). Then, we split F_t^i into two parts horizontally and each part contains the information of upper body or lower body and denoted as $F_{t,up}^i \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$ (green one) and $F_{t,low}^i \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$ (orange one). Next, we employ the Global Average Pooling (GAP) operation on whole feature

map and two sliced feature maps, i.e. F_t^i , $F_{t,up}^i$ and $F_{t,low}^i$, to obtain three feature vectors f_t^i , $f_{t,up}^i$ and $f_{t,low}^i$. We repeat above steps on every unlabeled images to generate three sets of feature vectors.

$$\begin{cases} f_t = \{f_t^1, \dots, f_t^k, \dots, f_t^{N_t}\} \\ f_{t,up} = \{f_{t,up}^1, \dots, f_{t,up}^k, \dots, f_{t,up}^{N_t}\} \\ f_{t,low} = \{f_{t,low}^1, \dots, f_{t,low}^k, \dots, f_{t,low}^{N_t}\} \end{cases} \quad (1)$$

Based on these feature vectors sets, we utilize unsupervised clustering algorithm [9] on each set to obtain a series of groups, leading to every person image can be assigned with a pseudo label according to the group it belongs to, named as self-label. As shown in Fig 2, we group images according to the three kinds of feature vectors, thus we can get three self-labels for each image x_t^i , denoted as y_t^i , $y_{t,up}^i$ and $y_{t,low}^i$. As a result, we can establish a new target dataset where each image has three self labels based on grouping result of three feature vectors, described as following.

$$X_T = \{x_t^i : (y_t^i, y_{t,up}^i, y_{t,low}^i); 1 \leq i \leq N_t\} \quad (2)$$

In addition to the feature set in equation (1), following the setting of the **baseline** model, we also employ one FC layer after the f_t^i to get a global embedding vector $f_{t,e}^i$, which is 2048-dims and shares the same self label with f_t^i . Note that this FC layer will also be updated during training.

Finally, we employ the self-labels as the supervised information to fine-tune the pre-trained model for cross dataset adaptation using triplet loss $L_{triplet}$, which will be elaborated in Section 3.4. Specifically, given an image, each feature vector and its corresponding self-label are used as two inputs for $L_{triplet}$. The full objective function of SSG is formulated as following,

$$L_{ssg} = L_{triplet}(f_t, y_t) + L_{triplet}(f_{t,up}, y_{t,up}) + L_{triplet}(f_{t,low}, y_{t,low}) + L_{triplet}(f_{t,e}, y_t) \quad (3)$$

During training, we follow the above steps to fine-tune the **baseline model** and employ the SSG for each iteration. During testing, we concatenate f_t^i , $f_{t,up}^i$, $f_{t,low}^i$ together as the final representation of each image x_t^i .

3.3. Clustering-guided Semi-Supervised Training

Although unsupervised domain adaptation for person re-ID has been studied extensively [8, 40, 39, 49], there is still more than 25% and 15% performance drop in mAP and Rank-1 accuracy comparing to their fully supervised counterparts. In order to narrow the enormous performance gap, we further extend our SSG to a semi-supervised setting.

Previous semi-supervised person re-ID works are mostly under an one shot setting, which is actually not practical in

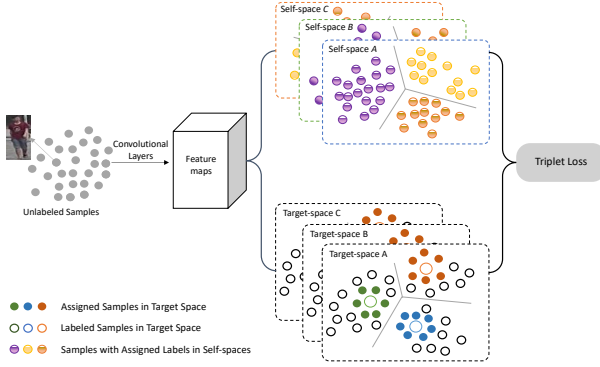


Figure 3. Overview of proposed clustering-guided semi-supervised training strategy combined with SSG. Given a target dataset, we annotate a few images based on the clustering results. And from these labeled images, we employ the step-wise learning approach to exploit the dataset gradually and obtain a more robust model.

the re-ID case since it is hard to know the number of independent identities in advance for a new dataset (as discussed in Section 2). Hence, we introduce a new semi-supervised training strategy based on clustering-guided annotation, which is more practical and useful for real-world applications. As shown in Fig 3, we employ the unsupervised clustering algorithm on f_t to generate N_g groups. Then, we randomly sample a single image from each group to form a very small sub-dataset X_g with N_g images. Next, we label this small sub-dataset manually and perform labels assignment based on this annotation. Specifically, we extract features of all images in sub-dataset X_g and follow the same operation described in Section 3.2 to obtain three feature vector sets f_g , $f_{g.up}$ and $f_{g.low}$ and treat each of them as an identity dictionary. Given an unlabeled image x_t^i , we find the most similar images from X_g by different cues, whole bodies, upper parts and lower parts, and assign x_t^i with corresponding labels y_{tg}^i , $y_{tg.up}^i$ and $y_{tg.low}^i$, which can be formulated as following.

$$\begin{cases} y_{tg}^i = \arg \min_{k=1:N_g} \text{dist}\{f_t^i, f_g^k\} \\ y_{tg.up}^i = \arg \min_{k=1:N_g} \text{dist}\{f_{t.up}^i, f_{g.up}^k\} \\ y_{tg.low}^i = \arg \min_{k=1:N_g} \text{dist}\{f_{t.low}^i, f_{g.low}^k\} \end{cases} \quad (4)$$

Note that we employ the k-reciprocal encoding [47], a variation of Jaccard distance between nearest neighbor sets, as the distance metric for similarity measurement. Since each image in X_g is from different groups, it's less possible that two different images share the same identity, which allows us to adopt some one shot learning approaches and further improve the performance. In particular, we follow the step-wise learning approach proposed in [41] to exploit the whole dataset during training stage progressively.

Furthermore, since SSG and clustering-guided semi-supervised training strategy share the same feature space, we can design a simple yet effective way to train the whole framework jointly and end-to-end, as shown in Fig 3. And the superiority of joint training strategy will be illustrated in ablation study.

3.4. Loss Function

Fully Supervised Training. 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:

$$L_{triplet} = \sum_{i=1}^P \sum_{a=1}^K \left[\alpha + \overbrace{\max_{p=1 \dots K} \|x_a^{(i)} - x_p^{(i)}\|_2}^{\text{hardest positive}} - \underbrace{\min_{\substack{n=1 \dots K \\ j=1 \dots P \\ j \neq i}} \|x_a^{(i)} - x_p^{(j)}\|_2}_{\text{hardest negative}} \right]_+ \quad (5)$$

where $x_a^{(i)}$, $x_p^{(i)}$, $x_n^{(i)}$ are features extracted from the anchor, positive and negative samples respectively, and α 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:

$$L_{softmax} = - \sum_{i=1}^P \sum_{a=1}^K \log \frac{e^{W_{y_{i,a}}^T x_{a,i}}}{\sum_{k=1}^C e^{W_k^T x_{a,i}}} \quad (6)$$

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:

$$L_{baseline} = L_{softmax} + L_{triplet} \quad (7)$$

Unsupervised and Semi-Supervised Training Unsupervised and semi-supervised training share the same loss function, and we just leverage the hard-batch triplet loss for metric learning. Also, each loss function has four components, whole body, upper body, lower body and global embedding, which can be formulated as follows:

$$\begin{aligned} L_{ssg} &= L_{triplet}(f_t, y_t) + L_{triplet}(f_{t.up}, y_{t.up}) \\ &\quad + L_{triplet}(f_{t.low}, y_{t.low}) + L_{triplet}(f_{t.e}, y_t) \\ L_{semi} &= L_{triplet}(f_t, y_{tg}) + L_{triplet}(f_{t.up}, y_{tg.up}) \\ &\quad + L_{triplet}(f_{t.low}, y_{tg.low}) + L_{triplet}(f_{t.e}, y_{tg}) \end{aligned} \quad (8)$$

Hence, the objective function used in jointly training strategy is

$$L_{jointly} = L_{ssg} + L_{semi}. \quad (9)$$

4. Experiments

In this section, we evaluate the proposed method on three large-scale re-ID datasets, *i.e.* Market1501 [45], DukeMTMC-ReID [31, 46] and MSMT17 [40].

4.1. Datasets and Evaluation Protocol

Market1501 [45] contains 32,668 images of 1,501 labeled persons from six camera views. Specifically, 12,936 person images of 751 identities detected by DPM [12] are used for training. For testing, in total 19,732 person images of 750 identities plus some distractors form the gallery set, and 3,368 manually cropped person regions from 750 identities form the query set.

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

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 include 12 outdoor and 3 indoor ones. Faster RCNN [30] is utilized for pedestrian bounding box detection. To the authors' best knowledge, the MSMT17 is the most challenging re-ID dataset with large-scale images and multiple 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. For Market-1501 and DukeMTMC-ReID, we use the evaluation packages provided by [45] and [46], respectively. Moreover, for simplicity, all results reported in this paper are under the single-query setting, and no post-processing like re-ranking [47] is applied.

4.2. Implementation Details

Baseline training As described in Section 3.1, we first train a baseline model on the source dataset by following the training strategy described in [51]. Specifically, we keep the size of input images and resize them to 256×128 . For data augmentation, we employ random cropping, 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 3×10^{-4} and decays to 3×10^{-5} after 100 epochs.

Unsupervised and Semi-supervised training. For unsupervised branch and semi-supervised branch, we follow

the same data augmentation strategy and triplet loss setting. And we decrease the initial learning rate from 3×10^{-4} to 6×10^{-5} and change training epoch from 150 to 70. For fairness, we randomly select a single image from each clustering to annotate and preserve them for all ablation study. Besides, the whole framework is trained for several iterations until the model is stable.

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-the-art unsupervised method. The performance of supervised **baseline** method and the direct transfer method are specified in Table 1. When comparing two methods, we can clearly find that there is a 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 30.5% and 16.1% when tested on DukeMTMC-reID, where the performance gap is more than 55%. Also, a similar drop can be observed when DukeMTMC-reID is used as the training set and tested on Market1501. Nevertheless, many recently proposed unsupervised domain adaptation methods cannot well address this great performance gap. For example, when transferring from DukeMTMC-reID to Market1501, to our best knowledge, the best UDA approach [8] only achieved 53.7% and 75.8% on mAP and rank-1 accuracy, which is lower than the fully supervised method by about 20%.

Effectiveness of Self-similarity Grouping. We perform several ablation studies to prove the effectiveness of proposed SSG as listed in Table 1. Specifically, with SSG, we improve the performance by 21.2% and 27.1% in mAP and rank-1 accuracy when the model is transferred from DukeMTMC-reID to Market1501. Similarly, when the model is trained on Market1501 and tested on DukeMTMC-reID, the performance gain becomes +37.9% and +42.9% in rank-1 accuracy and mAP, respectively. Moreover, compared with state-of-the-art UDA method, SSG can improve the performance by more than 5% on Market1501 and more than 3% on DukeMTMC-reID. This is because proposed SSG mines the potential similarity from global to fine manner, which can learn a more robust and discriminative model compared with [8]. Thus, SSG is a simple yet effective unsupervised method for domain adaptation in person re-ID.

In addition, we compare the performance of SSG with different number of sliced horizontal feature spatial parts. Table 2 suggests that SSG achieves the best results when we only split the feature maps into two parts: the up-

Methods	DukeMTMC-Re-ID \rightarrow Market1501				Market1501 \rightarrow DukeMTMC-ReID			
	mAP	R1	R5	R10	mAP	R1	R5	R10
Baseline(Upper Bound)	80.8	92.5	97.5	98.4	70.5	82.6	92.3	94.4
Direct Transfer	26.6	54.6	71.1	77.1	16.1	30.5	45.0	51.8
UDA(paper) [33]	53.7	75.8	89.5	93.2	49.0	68.4	80.1	83.5
Baseline + UDA	53.0	74.7	86.9	90.3	50.5	69.3	80.2	83.1
Baseline + SSG	58.3	80.0	90.0	92.4	53.4	73.0	80.6	83.2
Baseline + SSG ⁺	62.5	81.4	91.6	93.8	56.7	74.2	83.5	86.7
Baseline + SSG ⁺⁺	68.7	86.2	94.6	96.5	60.3	76.0	85.8	89.3

Table 1. Comparison of various methods on the target domains. When tested on DukeMTMC-reID, Market-1501 is used as the source and vice versa. “Baseline” denotes using the full identity labels on the corresponding target dataset(See Section 3.1). “Direct Transfer” means directly applying the source-trained model on the target domain. “UDA” stands for the state-of-art unsupervised domain adaptation approach. “Baseline+xxx” means using “xxx” domain adaptation method upon baseline model. “SSG” means Self-similarity Grouping in Section 3.2. “SSG⁺” and “SSG⁺⁺” stand for proposed SSG enhanced by clustering-guided semi-supervised training w/o and w/ joint training strategy described in Section 3.3.

Methods	mAP	R1	R5	R10
UDA	49.0	68.4	80.1	83.5
SSG(2 parts)	53.4	73.0	80.6	83.2
SSG(3 parts)	50.4	71.3	79.1	82.0
SSG(4 parts)	49.1	69.1	78.4	81.6

Table 2. The performance of proposed SSG with different number of splitted parts when trained DukeMTMC-reID dataset and tested on Market1501 dataset.

Methods	DukeMTMC-reID \rightarrow Market1501		
	mAP	R1	R10
SSG*	66.8	84.5	95.3
SSG ⁺⁺	68.7	86.2	96.5
Methods	Market1501 \rightarrow DukeMTMC-reID		
	mAP	R1	R10
SSG*	57.6	74.4	88.1
SSG ⁺⁺	60.3	76.0	89.3

Table 3. Comparison of proposed cluster-guided annotation with random sampling annotation on Market1501 dataset and DukeMTMC-reID dataset. SSG* is semi-supervised training with random sampling annotation

per and lower body. It infers that the upper and lower body contain the most discriminative information for re-ID. More fragments of features may break this information and yield worse similarity mining and matching.

Effectiveness of clustering-guided semi-supervised training. Table 1 shows the performance of the proposed cluster-guided semi-supervised method (SSG⁺). Compared to the unsupervised one (SSG), SSG⁺ outperforms SSG by 4.2% and 1.4% in terms of mAP and rank-1 accuracy when tested on Market1501. In addition, it outperforms the direct transfer method by 35.9% in mAP and 26.8% in rank-1 accuracy. These demonstrate the effectiveness of the clustering-guided semi-supervised training.

Effectiveness of joint training strategy Instead of fine-tuning the model obtained by SSG for SSG⁺, we conduct experiments to jointly train the model parameters using both SSG and SSG⁺ losses, and we denote this jointly trained model as SSG⁺⁺. In Table 1, when transferred from DukeMTMC-reID to Market1501, SSG⁺⁺ outperforms SSG⁺ by 6.2% and 4.8% in terms of mAP and rank-1

accuracy. The performance of testing on DukeMTMC-reID also boosts by 3.6% and 1.8% in mAP and rank-1 accuracy. It indicates that joint training strategy has its superiority on both two datasets, and training also becomes more efficient.

Effectiveness of clustering-guided annotation Intuitively, compared to random sampling from the target domain, clustering-guided annotation (*i.e.* sampling from unsupervised clustered groups and annotating) will increase the identity diversity in the sample set, and enhance the learned feature representation ability with limited supervised information. To validate this intuition, we compare clustering-guided and random sampling annotation on the jointly learned SSG⁺⁺. For fairness, we randomly sampled the same number of images from the whole dataset for the latter. In Table 3, SSG⁺⁺ surpasses the random sampling one by 2.7% and 1.6% in mAP and rank-1 accuracy when testing on DukeMTMC-reID and the similar improvement on performance can be found when testing on Market1501 as well. It verifies that the proposed clustering-guided annotation is better than the random one.

In conclusion, SSG⁺⁺ with clustering-guided annotation yields the best performance on both Market1501 and DukeMTMC-reID dataset. For instance, we achieve 42.1% and 31.6% improvements in mAP and rank-1 accuracy when testing on Market1501 compared with direct adaptation.

4.4. Comparison with State-of-arts

In this section, we compare the proposed method with state-of-the-art unsupervised learning methods on Market1501, DukeMTMC-reID and MSMT17 in Table 4, Table 5 and Table 6 respectively. SSG outperforms existing approaches with dominantly advantage. In particular, our model outperforms the best published method ARN [23] by almost 20% on mAP when testing on Market1501 and DukeMTMC-reID dataset. Moreover, it also surpasses the unpublished UDAP [8] and MAR(CVPR2019) [44] by a large margin.

Methods	mAP	R1	R5	R10
LOMO [24]	8.0	27.2	41.6	49.1
Bow [45]	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 [33]	53.7	75.8	89.5	93.2
MAR [44]	40.0	67.7	81.9	-
ENC [50]	43.0	75.1	87.6	91.6
SSG	58.3	80.0	90.0	92.4
SSG++	68.7	86.2	94.6	96.5

Table 4. Comparison of proposed SSG approach with state-of-arts unsupervised domain adaptive person re-ID methods on Market1501 dataset.

Results on Market1501 On Market-1501, we compare our results with two hand-crafted features, *i.e.* Bag-of-Words (BoW) [45] and local maximal occurrence (LOMO) [24], three unsupervised methods, including UMDL [29], PUL [10] and CAMEL [43], and six unsupervised domain adaptation methods, including PTGAN [40], SPGAN [8], TJ-AIDL [39], ARN [23], UDAP [33] and MAR [44]. The two hand-crafted features are directly applied to the test dataset without any training process, but it is obvious that both features fail to obtain competitive results. While training on target set, unsupervised methods always obtain higher results than hand-crafted features. Comparing with unsupervised domain adaptation methods, our proposed SSG is superior. Only in unsupervised setting, we achieve rank-1 accuracy = 80.0% and mAP = 58.3%, which outperforms the best unsupervised method [33] by 4.6% and 4.2%. Furthermore, with clustering-guided semi-supervised training strategy, we further improve the performance by 10% on mAP and 6% on rank-1 accuracy.

Results on DukeMTMC-reID The similar improvement can also be observed when we test it on DukeMTMC-reID dataset. Specifically, we achieve mAP = 53.4% and rank-1 accuracy = 73.0% by unsupervised SSG and obtain mAP = 60.3% and rank-1 accuracy = 76.0% under semi-supervised setting. Compared with best unsupervised method, our result is 4.4%/11.3% and 4.6%/7.6% higher on mAP and rank-1 accuracy. Therefore, the superiority of the proposed SSG adaptation approach for person re-ID can be concluded. In addition, we improve the performance by a large margin and recover about 90% performance of the fully supervised method by clustering-guided semi-supervised training strategy.

Results on MSMT17 In addition, we further evaluate the proposed SSG approach on MSMT17 dataset, which is the largest and most challenging re-ID dataset. We achieve mAP = 13.6% and rank-1 accuracy = 32.6% when

Methods	mAP	R1	R5	R10
LOMO [24]	4.8	12.3	21.3	26.6
Bow [45]	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 [33]	49.0	68.4	80.1	83.5
MAR [44]	48.0	67.1	79.8	-
ENC [50]	40.4	63.3	75.8	80.4
SSG	53.4	73.0	80.6	83.2
SSG++	60.3	76.0	85.8	89.3

Table 5. Comparison of proposed SSG approach with state-of-arts unsupervised domain adaptive person re-ID methods on DukeMTMC dataset.

Methods	DukeMTMC-reID → MSMT17		
	mAP	R1	R10
PTGAN [40]	3.3	11.8	27.4
SSG	13.3	32.2	51.2
SSG++	18.3	41.6	62.2

Methods	Market1501 → MSMT17		
	mAP	R1	R10
PTGAN [40]	2.9	10.2	24.4
SSG	13.2	31.6	49.6
SSG++	16.6	37.6	57.2

Table 6. Comparison of proposed SSG approach with state-of-arts unsupervised domain adaptive person re-ID methods on MSMT17 dataset.

trained DukeMTMC-reID, which improves state-of-the-arts by 10.0% and 30.4%. Also, similar improvement can be observed while trained on Market1501 as well, which further verifies the effectiveness of our proposed method.

5. Conclusion

In this work, we proposed Self-similarity Grouping (SSG), which can mine the potential similarity existing in target dataset automatically by different appearance cues (from global to local) in an unsupervised manner, to tackle the challenging domain adaption in person re-ID. Furthermore, we introduce a clustering-guided semi-supervised approach upon proposed SSG to adopt traditional one shot learning method to person re-ID, which is an open set problem. Extensive experimental results demonstrate that the performance of our approach outperforms the state-of-the-arts by a large margin.

Acknowledgements: This work is part supported by IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) and ARC DECRA DE190101315.

References

- [1] Slawomir Bak and Peter Carr. One-shot metric learning for person re-identification. In *IEEE CVPR*, pages 1571–1580, 2017. 3
- [2] Loris Bazzani, Marco Cristani, and Vittorio Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 2013. 3
- [3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *IEEE CVPR*, 2017. 2
- [4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *IEEE CVPR*, 2018. 1
- [5] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *IEEE CVPR*, 2018. 1
- [6] Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. Best practices for fine-tuning visual classifiers to new domains. In *ECCV*, pages 435–442, 2016. 2
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009. 4
- [8] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In *IEEE CVPR*, 2018. 1, 2, 3, 4, 6, 7, 8
- [9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, 1996. 4
- [10] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2018. 1, 8
- [11] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE TPAMI*, 2006. 3
- [12] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 2010. 6
- [13] Dario Figueira, Loris Bazzani, Ha Quang Minh, Marco Cristani, Alexandre Bernardino, and Vittorio Murino. Semi-supervised multi-feature learning for person re-identification. In *AVSS*, pages 111–116, 2013. 3
- [14] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, 2019. 4
- [15] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019. 4
- [16] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *IEEE ICML*, 2015. 2
- [17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 2016. 2
- [18] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 3
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 4
- [20] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 4, 5
- [21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 2
- [22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [23] Yu-Jhe Li, Fu-En Yang, Yen-Cheng Liu, Yu-Ying Yeh, Xiaofei Du, and Yu-Chiang Frank Wang. Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification. In *IEEE CVPRW*, 2018. 7, 8
- [24] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE CVPR*, pages 2197–2206, 2015. 3, 8
- [25] Ming-Yu Liu and Oncl Tuzel. Coupled generative adversarial networks. In *NIPS*, pages 469–477, 2016. 2
- [26] Xiao Liu, Mingli Song, Dacheng Tao, Xingchen Zhou, Chun Chen, and Jiajun Bu. Semi-supervised coupled dictionary learning for person re-identification. In *IEEE CVPR*, pages 3550–3557, 2014. 3
- [27] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *IEEE ICCV*, volume 2, page 3, 2017. 1, 2
- [28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6
- [29] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *IEEE CVPR*, pages 1306–1315, 2016. 3, 8
- [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 6
- [31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 2, 6
- [32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE CVPR*, 2015. 5
- [33] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsuper-

- vised domain adaptive re-identification: Theory and practice. *arXiv preprint arXiv:1807.11334*, 2018. 7, 8
- [34] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. 2
- [35] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, pages 443–450, 2016. 2
- [36] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. *arXiv preprint arXiv:1711.09349*, 2017. 1
- [37] Eric Tzeng, Coline Devin, Judy Hoffman, Chelsea Finn, Pieter Abbeel, Sergey Levine, Kate Saenko, and Trevor Darrell. Adapting deep visuomotor representations with weak pairwise constraints. *arXiv preprint arXiv:1511.07111*, 2015. 2
- [38] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Multimedia*, 2019. 1, 4
- [39] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *IEEE CVPR*, 2018. 3, 4, 8
- [40] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE CVPR*, 2018. 2, 3, 4, 6, 8
- [41] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *IEEE CVPR*, 2018. 3, 5
- [42] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 2
- [43] Hong-Xing Yu, Ancong Wu, and Wei-Shi Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *IEEE ICCV*, 2017. 8
- [44] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. In *IEEE CVPR*, 2019. 7, 8
- [45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE ICCV*, pages 1116–1124, 2015. 2, 6, 8
- [46] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv preprint arXiv:1701.07717*, 2017. 2, 6
- [47] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *IEEE CVPR*, pages 3652–3661, 2017. 4, 5, 6
- [48] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 6
- [49] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, pages 172–188, 2018. 2, 3, 4, 8
- [50] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *IEEE CVPR*, 2019. 8
- [51] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *IEEE CVPR*, pages 5157–5166, 2018. 4, 6