CLUSTERING AND DYNAMIC SAMPLING BASED UNSUPERVISED DOMAIN ADAPTATION FOR PERSON RE-IDENTIFICATION

Jinlin Wu\textsuperscript{1,2}, Shengcai, Liao\textsuperscript{3}, Zhen Lei\textsuperscript{1,2}, Xiaobo Wang\textsuperscript{4}, Yang Yang\textsuperscript{*1,2}, Stan Z. Li\textsuperscript{1,2}

\textsuperscript{1}CBSR&NLPR, Institute of Automation Chinese Academy of Sciences, Beijing, China, 100190
\textsuperscript{2}University of Chinese Academy of Sciences, Beijing, China, 100049
\textsuperscript{3}Inception Institute of Artificial Intelligence
\textsuperscript{4}JD AI Research, Beijing, China

\{jinlin.wu, zlei, szli, yang.yang\}@nlpr.ia.ac.cn, scliao@ieee.org, wangxiaobo8@jd.com

ABSTRACT

Person Re-Identification (Re-ID) has witnessed great improvements due to the advances of the deep convolutional neural networks (CNN). Despite this, existing methods mainly suffer from the poor generalization ability to unseen scenes because of the different characteristics between different domains. To address this issue, a Clustering and Dynamic Sampling (CDS) method is proposed in this paper, which tries to transfer the useful knowledge of existing labeled source domain to the unlabeled target one. Specifically, to improve the discriminability of CNN model on source domain, we use the commonly shared pedestrian attributes (e.g., gender, hat and clothing color etc.) to enrich the information and resort to the margin-based softmax (e.g., A-Softmax) loss to train the model. For the unlabeled target domain, we iteratively cluster the samples into several centers and dynamically select informative ones from each center to fine-tune the source-domain model. Extensive experiments on DukeMTMC-reID and Market-1501 datasets show that the proposed method greatly improves the state of the arts in unsupervised domain adaptation.

Index Terms— Clustering, Dynamic Sampling, Pedestrian Attributes, A-Softmax

1. INTRODUCTION

Person Re-identification (Re-ID) is fundamental and of great practical value in surveillance video analysis, of which the core is to match the same pedestrian across multiple camera views. Recently, benefiting from the development of deep learning [1], the performance of Re-ID has obtained significant improvements by adopting supervised learning [2, 3, 4]. However, in real applications, it is expensive and time-consuming to obtain sufficient manually labeled data. Moreover, due to the diversity of data acquisition, the domain gap between different datasets is a critical issue. For example, the pedestrians from DukeMTMC-reID [5] usually wear long and thick clothes while those from Market-1501 [6] almost wear short and thin clothes. Besides the foreground gap, the backgrounds between these two datasets are also significantly different (see Fig.1). These factors lead to a sharp performance decrease when directly applying the trained model from the labeled source domain to the unlabeled target one (e.g., Rank-1 rate declines from 76.2\% when trained and evaluated on the Market-1501 to 36.1\% when directly evaluated on the DukeMTMC-reID).

To address the domain gap, a large number of methods have been proposed. One group considers this issue as a style transferring task. The methods in [7, 8] adopt the generative adversarial nets [9] transferring the unlabeled target image style to labeled source image style, trying to eliminate the domain gap between the different datasets. However, the discriminant information might be lost in the transferring process, which leads to limited Rank-1 accuracy. Another group resort to the unsupervised domain transfer meth-
Fig. 2: The framework of the proposed Clustering and Dynamic Sampling (CDS) method: we firstly train our source-domain model with both identity classification and attribute classification. Seven attributes including gender, hat, backpack, bag, handbag, upper-body clothing color and lower body clothing color are employed to enhance the source training of the source-domain model. Then we iteratively cluster the target-domain samples and dynamically select informative samples in target domain to fine-tune the source-domain model.

ods. Approaches in [10, 11, 12, 13, 14] adopt unsupervised approaches to enhance the model’s ability on the target unlabeled dataset. Wang et al. [10] propose an unsupervised transfer method called TJ-AIDL, transferring source domain knowledge to target domain by learning the attribute-semantic and identity-discriminative feature space. Fan et al. [14] propose a progressive unsupervised learning (PUL) approach obtaining pseudo labels to fine-tune the source-domain model. However, its performance is limited to the source-domain model’s poor generalization and the fixed-threshold sampling.

In this paper, we propose a novel clustering and dynamic sampling (CDS) scheme to address the cross-domain unsupervised adaption issue. The framework of our CDS is shown in Fig. 2. Specifically, we firstly train the source-domain model with both identities and pedestrian attributes, under the supervision of the angular softmax loss (A-Softmax [15]). Generally speaking, that identities labels from different pedestrian datasets are non-overlapping, while the attributes [3] (e.g., hat, clothing color, gender, bags and backpack) are commonly shared. In this work, we use attribute classification to improve the discriminability of the baseline source-domain model. We adopt the angular softmax loss (A-Softmax [15]) to supervise the training process. Then, to overcome the drawbacks of the fixed-threshold sampling in clustering centers, we employ a dynamic sampling strategy to select informative training samples on the target domain. To sum up, the contributions of this paper can be summarized as follows:

- We propose a novel Clustering and Dynamic Sampling (CDS) scheme to address unsupervised cross-domain adaptation task in person re-identification.
- We employ the commonly shared attributes and the margin-based A-Softmax loss to improve the discriminability of the source-domain model and adopt a dynamic threshold to select informative target-domain samples for unsupervised adaptation.

- Extensive experiments on the DukeMTMC-reID and Market-1501 datasets show that the proposed method largely improves the state of the arts in unsupervised domain adaptation.

2. RELATED WORK

2.1. Supervised Person Re-Identification

Most of existing Re-ID models are built by learning similarity or representation via cross-camera data [4, 16, 17, 3, 2, 18] in a supervised manner. In early works, Yi et al. [4] introduce part priors into deep convolutional network for similarity learning. In recent works, Hermans et al. [3] propose a multi-task training approach, combining ID classification and attributes classification tasks to improve the Re-ID performance. Fan et al. [2] introduce an angular margin to the softmax function, reducing intra-class variations and increasing inter-class variations simultaneously. These approaches achieve superior performance in a single domain task, but may not be effective for cross-domain person Re-ID.

2.2. Cross-Domain Person Re-Identification

To address the cross-domain issue, researchers shift their attention to the unsupervised methods, including unsupervised clustering [14], image-style transfer [7, 19, 8, 20] and unsupervised domain transfer [11, 10, 21]. To unsupervised clustering, Fan et al. [14] propose a clustering and self-pace training frame named PUL, which adopts the k-means algorithm to obtain pseudo labels and fine-tunes the source-domain model with these pseudo labels. However, in this work, the threshold for controlling training data sampling is
3. PROPOSED METHODS

Notation. Given a labeled source dataset \( X^s \) and an unlabeled target dataset \( X^t \), where \( X^t = \{x^t_1, x^t_2, ..., x^t_N\} \) contains \( N \) images, the source-domain model \( \phi_0 \) is pre-trained on the source dataset, where \( \theta \) denotes the corresponding parameters. Since the target dataset is unlabeled, we consider the latent ground truth labels of \( X^t \) as \( Y^t = \{y^t_1, y^t_2, ..., y^t_N\} \). While the pseudo labels predicted by the k-means algorithm are donated as \( \hat{Y}^t = \{\hat{y}^t_1, \hat{y}^t_2, ..., \hat{y}^t_N\} \).

3.1. Supervised Learning on Source Domain

Attribute Knowledge. Person identities are usually non-overlapping across different Re-ID datasets, while pedestrian attributes are commonly shared. For instance, as shown in Fig. 3, the attributes female, backpack, red are shared on the Market-1501 and DukeMTMC-reID. It seems that the attribute knowledge is easier to transfer than identity knowledge. To this end, besides the identity classification, we additionally use seven common attributes (gender, hat, backpack, bag, handbag, upper-body clothing color, lower body clothing color) to improve the generalization ability of the source-domain model, so as to make a better starting point for the latter fine-tuning.

A-Softmax. For the loss function of the supervised learning, as we expect that the representations learned by the source-domain model are discriminative and are easy to transfer. Therefore, we adopt the angular softmax loss (A-Softmax [15]) to train the source-domain model. Specifically, A-Softmax loss aims to increase the inter-class variations and meanwhile decrease the intra-class variations by adding an angular margin to the original softmax loss. The A-Softmax loss is formulated as follows:

\[
L_{\text{A-Softmax}} = -\frac{1}{N} \sum_{i=1}^{N} \log \sum_{j \neq y^t_i} e^{z_{y^t_i}} + e^{z_{j}}
\]

where \( k \in [0, m - 1] \) and \( m \) is the margin factor as a hyper parameter. For more details, please refer to the work [15].

3.2. Clustering and Dynamic Sampling

Clustering. On the target domain, since the samples are unlabeled, we firstly cluster them into several centers by using the well-known k-means algorithm. The formulation can be summarized as:

\[
\min_{y^t_1, y^t_2, ..., y^t_K} \sum_{k=1}^{K} \sum_{x_i^t \in y^t_k} \| \phi(x_i^t) - c_k \|_2,
\]

where \( K \) is the number of clusters pre-defined empirically, and \( c_i, i \in \{1, 2, ..., K\} \) are the corresponding cluster centers.

Dynamic Sampling. Generally, samples in each cluster may contain noisy pseudo labels or outliers due to the limited ability of the initial source-domain model. Regarding this, we define a binary mask \( v_i \) and use a threshold \( \lambda \) to indicate whether a sample belongs to a certain cluster or not:

\[
v_i = \begin{cases} 1, & \cos(\phi_0(x_i^t), c_{z_i}) > \lambda, \\ 0, & \text{otherwise}, \end{cases}
\]

where \( c_{z_i} \) denotes the corresponding cluster center of the image \( x_i^t \). From the definition, we can see that if the cosine similarity \( \cos(\phi_0(x_i^t), c_{z_i}) \) is larger than a threshold \( \lambda \), the sample \( x_i^t \) and its pseudo label \( y_i^t \) would be temporarily selected. Otherwise, it will be discarded. After that, we use the selected samples to fine-tune the source-domain model by optimizing the following objective:

\[
\min_{v_w} \sum_{i=1}^{N} L(y_i^t, f_w(\phi(x_i^t))),
\]

where \( L \) is the A-Softmax loss function, and \( f_w \) is the classifier of the target domain.

For the threshold \( \lambda \), a large value prefer to choose samples belonging to the same identity under the same camera. Such selected samples are relatively reliable, but they may be useless for cross-camera person retrieval. In contrast, a small value usually involves noisy samples or outliers. As a result, the model may be misled by them during the fine-tuning process. To solve this dilemma, we design a dynamic sampling strategy. Specifically, at the beginning of the training, since the model’s discriminability is weak on the target
domain, it is easily misled by noisy samples or outliers. Thus in this case we set a larger threshold to cautiously select reliable samples, although these samples may belong to the same camera and the same ID. When the model’s discriminability becomes stronger, we can set a lower threshold to select more informative samples including the cross-camera pairs to improve the model’s cross-camera retrieval ability. To this end, we develop a dynamic sampling strategy by using a decreasing rate $\eta$ to iteratively decrease the sampling threshold $\lambda$:

$$\lambda \leftarrow \lambda - \eta(U - L),$$  

(5)

where a large threshold $U$ is served as the upper bound and a small threshold $L$ is used as the lower bound. The threshold $\lambda$ is decreasing from the upper bound to the lower bound, as the model’s discriminative ability is being enhanced gradually. For clarity, the proposed clustering and dynamic sampling (CDS) scheme is summarized in Algorithm 1.

**Algorithm 1 Clustering and Dynamic Sampling (CDS)**

**Input:** Unlabeled dataset $X$; Upper bound $U$ and lower bound $L$; Source domain pre-trained model $\phi_0$; Target domain classifier $f_w$; Relaxing rate $\eta$.

**Output:** Fine-tuned model $\phi_f$.

1. Initial $\lambda_0 \leftarrow U$, $\theta_t \leftarrow \theta_0$.
2. repeat
3. extracting deep feature: $f_i = \phi_0(x_i^t)$ for all $x_i^t$;
4. $L_2$-normalization for deep feature $f_i$;
5. $k$-means clustering;
6. updating pseudo labels $y_i^t$ and cluster center $C$;
7. for $i = 1$ to $N$ do
8. find the nearest cluster center $c_s^t$ of $x_i^t$;
9. $S = f_i \cdot c_s^t$;
10. if $S > \lambda$ then
11. selecting $x_i^t$: $v_i = 1$;
12. else
13. discarding $x_i^t$: $v_i = 0$;
14. end if
15. end for
16. training $< \phi_f, f_w >$ with the selected samples
17. updating $\lambda$: $\lambda \leftarrow \lambda - \eta(U - L)$
18. until ($\lambda \leq l$)

4. EXPERIMENTS

4.1. Datasets

We conduct extensive experiments on DukeMTMC-reID [5], Market-1501 [6] and CUHK03 [16] to evaluate our approach. Data statistics and evaluation settings are shown in Table 1. On both Market-1501 and DukeMTMC-reID, we adopt the standard single-query protocol [6] as the testing protocol. On CUHK03, we follow the testing protocol of [22] to evaluate our method.

4.2. Implementation details

Our approach is implemented with Pytorch framework. The model structure and the training pipeline are shown in 2. The classifier $f_w$ of target domain consists of one linear layer. The backbone is pre-trained with ID classification and attributes classification tasks on the source dataset. Seven pedestrian attributes (i.e., gender, hat, backpack, bag, handbag, upper-body clothing color, lower body clothing color) are used in the source training process. A-Softmax is used as the loss function for ID classification task on both source training and target training. The $m$ of A-Softmax is set to 3 in the experiments. The input images are resized to $288 \times 188$. We use Adam optimizer with the hyper-parameters (\(\epsilon = 10^{-8}, \beta_1 = 0.9, \beta_2 = 0.99, lr = 0.0001\)). For the dynamic sampling, we set the dynamic decreasing rate to $1.5 \times 10^{-3}$. The upper bound $U$ and lower bound $L$ are set to 0.8 and 0.7 respectively. The batch size is set to 16.

4.3. Ablation Experiments

In this section, we analyze the impact of the three main strategies (Attributes Knowledge, A-Softmax, Dynamic Sampling) and cluster number $K$ of CDS.

**Table 1:** Data splits (/#persons/#images) for experiments on the Market-1501 and DukeMTMC-reID datasets.

<table>
<thead>
<tr>
<th>Domain</th>
<th>#All</th>
<th>#Train</th>
<th>#Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market-1501</td>
<td>1501/32688</td>
<td>751/12936</td>
<td>750/19732</td>
</tr>
<tr>
<td>DukeMTMC-reID</td>
<td>1812/36411</td>
<td>702/16522</td>
<td>702/19889</td>
</tr>
<tr>
<td>CUHK03</td>
<td>1467/14097</td>
<td>767/7365</td>
<td>708/6732</td>
</tr>
</tbody>
</table>

**Table 2:** Ablation Experiments. The "Attributes" denotes whether using attributes on the source domain training. For the “Fixed or Dynamic”, “Fixed” denotes using a fixed threshold during the sampling procedure and "Dynamic" means using dynamic sampling (DS) strategy. The "A-Softmax" denotes whether using the A-Softmax as loss function in the target training procedure.

**Strategy** | **Market->Duke** | **Duke->Market** |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>** Attributes**</td>
<td><strong>mAP</strong></td>
<td><strong>Rank-1</strong></td>
</tr>
<tr>
<td><em>yes</em></td>
<td>19.9</td>
<td>37.2</td>
</tr>
<tr>
<td><em>no</em></td>
<td>15.8</td>
<td>31.0</td>
</tr>
<tr>
<td>** Fixed or Dynamic**</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>yes</em></td>
<td>0.6</td>
<td>9.8</td>
</tr>
<tr>
<td><em>no</em></td>
<td>0.8</td>
<td>39.6</td>
</tr>
<tr>
<td>** A-Softmax**</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>yes</em></td>
<td>0.7</td>
<td>35.6</td>
</tr>
<tr>
<td><em>no</em></td>
<td>0.6</td>
<td>29.71</td>
</tr>
</tbody>
</table>

**Table 3:** Ablation Experiments. The "Fixed or Dynamic" denotes whether using attributes on the source domain training. The "A-Softmax" denotes whether using the A-Softmax as loss function in the target training procedure.
**Table 3**: Comparing CDS with other state-of-the-art methods on DukeMTMC-reID and Market-1501 datasets.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP</td>
<td>Rank-1</td>
<td>Rank 5</td>
</tr>
<tr>
<td>LOMO [23]</td>
<td>8.0</td>
<td>27.2</td>
</tr>
<tr>
<td>UMDL [12]</td>
<td>12.4</td>
<td>34.5</td>
</tr>
<tr>
<td>PTGAN [24]</td>
<td>-</td>
<td>38.6</td>
</tr>
<tr>
<td>PUL [14]</td>
<td>20.5</td>
<td>45.5</td>
</tr>
<tr>
<td>CAMEL [25]</td>
<td>26.3</td>
<td>54.5</td>
</tr>
<tr>
<td>SPGAN+LMP [7]</td>
<td>26.7</td>
<td>57.7</td>
</tr>
<tr>
<td>TJ-AIDL [10]</td>
<td>26.5</td>
<td>58.2</td>
</tr>
<tr>
<td>HHL [8]</td>
<td>31.4</td>
<td>62.2</td>
</tr>
<tr>
<td>CamStyle [26]</td>
<td>27.4</td>
<td>58.8</td>
</tr>
<tr>
<td>CamStyle + LMP [27]</td>
<td>30.4</td>
<td>64.7</td>
</tr>
<tr>
<td>ARN [11]</td>
<td>39.4</td>
<td>70.3</td>
</tr>
<tr>
<td>CDS</td>
<td><strong>39.9</strong></td>
<td><strong>71.6</strong></td>
</tr>
</tbody>
</table>

**Fig. 4**: The impacts of the number of cluster centers. The blue curve denotes Rank-1 accuracy and the yellow curve denotes mAP.

**4.4. Comparing with the State-of-the-art Methods**

The experimental results are displayed in Table 3 and Table 4. Comparing with the hand-crafted features LOMO [23], three image-style transform method [24, 7, 8] (i.e., PTGAN, SPGAN+LMP and HHL), two unsupervised domain transfer methods TJ-AIDL and ARN, and two clustering methods PUL and CAMEL, the proposed CDS method outperforms these approaches in Rank-1, Rank-5 and mAP on both the Market-1501, DukeMTMC-reID and CUHK03 datasets. Comparing with the second best method ARN, the proposed CDS method outperforms ARN by 7% in Rank-1, 9.3% in mAP on the DukeMTMC-reID dataset, while achieving 1.3% and 0.5% gains in Rank-1 and mAP, respectively, on DukeMTMC-reID dataset. Comparing with the fixed sampling method PUL, the proposed CDS method outperforms PUL by 29.5% in Rank-1 and 19.4% mAP in the Market-1501, 26.3% in Rank-1 and 37.2% in mAP on the DukeMTMC-reID dataset. Together with the ablation experiments, we conclude that the improvement of the proposed CDS for cross-domain person Re-ID mainly benefits from the attribute guided training, dynamic sampling in clusters, and the A-Softmax loss function.

**5. CONCLUSION**

In this paper, we have proposed a novel Clustering and Dynamic Sampling (CDS) approach to address cross-domain person re-identification task. It uses the pedestrian attributes to improve the discriminability of the baseline model, the A-Softmax loss to train our model, and the dynamic sampling strategy to select informative samples from the clustering results. We have achieved the state-of-the-art results on both the DukeMTMC-reID and Market-1501 datasets under unsupervised domain adaptation. Particularly on the Market-1501,
we exceed the current state-of-the-art method ARN by 7% in Rank-1 accuracy and 9% in mAP.

6. REFERENCES