Multiple Granularity Network and Dynamic Label for Domain Adaptive Person Re-identification

Xile Wang\textsuperscript{1,a}, Sihan Zhang\textsuperscript{2,b}, Junyu Song\textsuperscript{2,c} and Miaohui Zhang\textsuperscript{1,2,d,\\*}

\textsuperscript{1} Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, Henan, 475004, China
\textsuperscript{2} School of Artificial Intelligence, Henan University, Kaifeng, Henan, 475004, China

\textsuperscript{a}email: 104753190612@henu.edu.cn, \textsuperscript{b}email: 104754190915@henu.edu.cn, \textsuperscript{c}email: 104754190915@henu.edu.cn
\textsuperscript{d}email: zhmh@henu.edu.cn

\* Corresponding author: \textsuperscript{d}email: zhmh@henu.edu.cn

Abstract. The domain adaptive person re-identification (Re-ID) has become more popular among researchers. Because it can save a lot of resources as it only exploits the source domain knowledge and does not need the complex annotation efforts in target domain. It aims to extend a model trained on a labeled dataset to another dataset which is unlabeled. Many works reduce feature distribution gap between two different datasets to solve the problem. However, these works ignore the problem which is the variations within an unlabeled dataset. In the paper, we propose a domain adaptive person Re-ID framework based on multiple granularity network and dynamic label (MGDL). Specifically, we send the images of two different datasets into multiple granularity network at the same time for joint training to reduce feature distribution gap which is between the two different datasets. The network is trained by two different kinds of pseudo labels, namely, conservative label and radical label. The two kinds of pseudo labels are used to alternating pull and push the feature distribution in the target domain to reduce the variations within an unlabeled dataset. Experiments have shown that the MGDL achieves competitive performance in person Re-ID which is under the cross-domain setting.

1. Introduction
Person Re-ID is committed to match a given person image in an image gallery which is obtained from nonoverlapping cameras. The fully-supervised person Re-ID have obtained a great achievement. However, the performance of the model will decline dramatically when the model trained in a person Re-ID dataset (source domain of labelled data) is directly used in another person Re-ID dataset (target domain of unlabelled data). This problem is caused by the large gap of feature distribution between two different datasets and the variations within an unlabeled dataset. This work considers the problem that the performance degradation of person Re-ID model under the cross-domain setting, and proposes a method to reduce the cross-domain impact on the person Re-ID model.

The domain adaptive person Re-ID aims to learn a Re-ID model from a dataset which is labeled and another dataset which is unlabeled. It can save a lot of resources as it only exploits the source domain knowledge and does not need the complex annotation efforts in target domain. So, it is required for real-world deployments and attractive for researcher. Some studies\cite{1} apply Generative Adversarial Network (GAN) to perform image style transfer, meanwhile maintaining the identity annotations on source domains. TJ-AIDL\cite{2} is proposed a method which uses the attribute semantic
and the discriminative feature to solve the problem in domain adaptive person Re-ID. Yu et al.[3] take a labeled source domain as a vector space, and use the vector space to describe the distribution of pedestrians in target domain. Each pedestrian image in target domain can get the soft multilabel, and uses the pseudo label for clustering.

Compared with traditional methods, we propose MGDL for domain adaptation person Re-ID. We use multiple granularity feature extraction and joint training to reduce feature distribution gap which is between two different datasets. We design a new type of pseudo label, namely, dynamic label which contains conservative label and radical label. During the training process, the dynamic label is used to alternating pull and push the feature distribution to reduce the variations within an unlabeled dataset.

On the whole, the main contribution in this paper is in the following points:
1. We propose a domain adaptive person Re-ID framework based on multiple granularity network and dynamic label (MGDL), which can not only effectively reduce feature distribution gap which is between two different datasets, but also effectively reduce the variations within an unlabeled dataset.
2. We designed experiments on commonly used datasets which are Market1501 [4] and DukeMTMC-ReID[5] to show that our MGDL on the task of domain adaptive person Re-ID is effective.

2. Our Methods
In this section, we will introduce the overall framework of MGDL, which is based on multiple granularity network and dynamic label. After that, we will introduce the design of dynamic label and the generation method of conservative label and radical label in detail.

2.1. The Overall Framework of MGDL
We can see the framework of MGDL in Figure 1. In the MGDL, the backbone of our framework is the ResNet-50[6] which is trained on ImageNet [7]. Specifically, we remove the down-sampling layer in the fourth residual block and design six branches. We use the six branches to extract the features of the upper, middle and lower local blocks from the fine granularity, the upper and lower local blocks from the medium granularity, and the whole pedestrian image from the coarse granularity. After that, we connect these features of multiple granularities and fuse them through a shallow CNN as the final pedestrian features.

To reduce the influence of the gap of feature distribution which is two different datasets, the framework of our method uses the dataset with labelled data and another dataset with unlabelled data for joint training. In this way, the training model not only conforms to the feature distribution on the source domain dataset, but also conforms to the feature distribution on the target domain dataset. Because identities of source images are labeled, we use supervised training for source domain dataset and use the cross-entropy loss to optimize the network. It can be expressed as

$$L_s = -\frac{1}{n_s} \sum_{i=1}^{n_s} \log p(y_i^s|x_i^s) \quad (0)$$

where $n_s$ means the number of source images during small batch training. $x_i^s$ represents the i-th image in dataset which is labeled by people. $y_i^s$ is the identity of i-th image in the dataset with labelled data. $p(y_i^s|x_i^s)$ is the possibility that the $x_i^s$ belongs to the identity $y_i^s$, which is predicted by the classifier.

Due to identities of target images are unknown, we need to optimize the image clustering in the target domain. We use the memory which is proposed in ECN[8] to store the features of all images. The memory $M \in R^{N_t \times d}$ is constantly updated regularly during training:

$$m_t \leftarrow \alpha m_t + (1 - \alpha)f(x_i^s), \quad m_t \leftarrow m_t/\|m_t\|_2 \quad (2)$$

where each slot stores $m_i \in R^d$ the feature of the sample $x_i^s$, $f(x_i^s) \in R^d$ is the $l_2$-normalized feature of $x_i^s$ which is extracted by the model which is obtained in current training. The $\alpha \in [0,1]$ is the updating rate.
2.2. Dynamic Label

In the actual scene, we can easily get the camera-ID of every pedestrian image. For target domain without identity labels, we suppose the images of target domain dataset is taken by $C$ cameras and treat the images taken by the same camera as a separate style domain. Then we will use trained StarGAN[9] to generate $C-1$ images which is belong other style domain.

In this paper, the improved hot-labels is used to design two kinds of pseudo labels for the unlabeled target images. The pseudo label for each image is a vector of $N^f$ dimension, which represents a common $N^f$ classes, where each element is the possibility of the image belonging to this class. $N^f$ represents the total number of pedestrian images in unlabeled dataset which contains the pedestrian images generated by StarGAN. Every image of a person in unlabeled dataset can get two pseudo labels which are conservative label and radical label. We use the two kinds of pseudo labels to alternating pull and push the feature distribution in the target domain to reduce the variations within an unlabeled dataset.

The target image and the images which are generated by StarGAN from the target image must belong to the same person. Target domain images are classified according to this standard to construct pseudo label which is named conservative label. Specifically, assume that the conservative label of image $i$-th is $W^c_i \in R^{1 \times N^c}$, suppose that set $M^c_i$ contains the $i$-th image, the original image of the $i$-th image is generated, and these pedestrian images which are generated by StarGAN from the original image. So, we set the weight of possibility that the $i$-th image belongs to the class $j$ is:

$$w^c_{i,j} = \begin{cases} 1, & j \in M^c_i \\ 0, & \text{others} \end{cases}$$

where $w^c_{i,j}$ is $j$-th element of conservative label $W^c_i$. So the conservative loss function can be expressed as:

$$L_{cc} = -\frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{j} w^c_{i,j} \log p(j|x^f_i).$$

Where $j \in M^c_i$. $n_t$ means the number of images during small batch training. $p(j|x^f_i)$ represents the predicted possibility of the source image $x^f_i$ belongs to $j$-th class, which is obtained by the classifier.

The similarity between the target domain image in the memory module is calculated to obtain top $k$ images which have the highest similarity with the target image. These images are considered to belong to the same class and construct pseudo label which is named radical label. Specifically, assume that the radical label of image $i$-th is $W^r_i \in R^{1 \times N^r}$, suppose that set $M^r_{i,k}$ contains top $k$ images which have the highest similarity with the $i$-th image. So, we set the weight of the possibility that the $i$-th image belongs to the class $j$ is:
\[ w_{i,j}^r = \begin{cases} 1, & i = j \\ \frac{1}{k}, & j \in M_{i,k}^r \\ 0, & \text{others} \end{cases} \]

The \( w_{i,j}^r \) means the \( j \)-th element of the radical label \( W_i^r \). So the radical loss function can be expressed as:

\[ L_{tr} = -\frac{1}{n_t} \sum_{i=1}^{n_t} \sum_j w_{i,j}^r \log p(j|x_i^r) \]

Where \( j \in M_{i,k}^r \). The total loss of training images is denoted as:

\[ L = \beta L_t + (1 - \beta)L_s = \beta (L_{tc} + L_{tr}) + (1 - \beta)L_s \]

where \( \beta \in [0,1] \) control the balance of the target loss and source loss.

3. Results & Discussion
In this section, the experimental details and datasets which are used in this paper will be introduced. Then we made a comparative analysis for the experimental results of MGDL and other works.

3.1. Experimental setting
Our MGDL are evaluated on commonly used datasets which are Market-1501[4] and DukeMTMC-reID[5]. The detailed statistics of the two datasets which are used in this paper are shown in Table 1. The mAP and the CMC are used to evaluate the performance of the methods. Before the model training, we will use StarGAN to obtain the images which are other camera styles from each real target image while remaining the original identity. After that, we will make all pedestrian images to the 256 × 128 and we extend the dataset by random cropping and random erasing.

Table 1. The detailed information of the two datasets: Market-1501 and DukeMTMC-reID

| Dataset          | Market-1501               | DukeMTMC-reID     |
|------------------|---------------------------|-------------------|
| Cameras          | 6                         | 8                 |
| Training set (Image/ID) | 12,936/751       | 16,522/702       |
| Testing set (Image/ID) | Gallery          | 19,732/750       |
|                  | Query                     | 3,368/750         |
|                  |                            | 17,661/1,110      |
|                  |                            | 2,228/702         |

In our model, we remove the last down-sampling layer in the fourth residual block and design six branches in ResNet-50. During the training phase, the batch size of training pedestrian images is set to 128. We set the initial value of the update parameter \( \alpha \) of the memory \( M \) as 0.01, and it changes with the increase of epoch. We set the number of images which are most similar to the target image \( k = 12 \) and weight of losses \( \beta = 0.4 \). The SGD optimizer is used to train 60 epochs. The initial learning rate is 0.1 and it will be shrunken to one tenth of the original every 20 epochs. During the evaluation phase, we will remove the last fully connected layer, and get the features of pedestrian images directly through the previous stage. The metric is calculated according to the cosine distance between two features.

3.2. Experimental results
As shown in Table 2, we compare the results of our MGDL and other methods under the cross-domain setting. Compare to other approaches, our MGDL gets the best results than them on the two datasets. When we test the methods on Market-1501, we will use DukeMTMC-reID as the source domain dataset, our MGDL obtains mAP = 47.4% and rank-1 accuracy = 77.8%. When we test the methods on DukeMTMC-reID, we will use Market-1501 as the source domain dataset, our MGDL obtains mAP = 43.8% and rank-1 accuracy = 67.4%.

Table 2. The Rank-i and mAP of our MGDL and other methods on Market-1501 dataset and DukeMTMC-reID dataset

| Methods        | Market-1501 | DukeMTMC-reID |
|----------------|-------------|---------------|
|                | Rank-1 | Rank-5 | Rank-10 | mAP  | Rank-1 | Rank-5 | Rank-10 | mAP |
| CamStyle[10]   | 58.8   | 78.2   | 84.3    | 27.4 | 48.4   | 62.5   | 68.9    | 25.1 |
To show the superiority of our MGDL intuitively, we show some retrieval results of our MGDL when we test the model which is trained by the two datasets on Market-1501 dataset, including examples of success and failures as shown in Figure 2. The green box means the image belongs to the same identity as the target pedestrian image, and the red box means the image does not belong to the same person as the target pedestrian image.

![Figure 2. Some results are obtained on Market-1501 by MGDL.](image)

### 4. Conclusions

In our work, we design a domain adaptive person Re-ID framework which are based on multiple granularity network and dynamic label (MGDL) to improve performance degradation after cross domain. We start from the two causes of this problem, we use the joint training strategy of two different datasets on the multiple granularity network to reduce the large gap of feature distribution between two different datasets. After that, we design dynamic label to alternating pull and push the feature distribution in the target domain to reduce the variations within an unlabeled dataset. The experimental results demonstrate that MGDL can solve the problem of performance degradation after cross domain very well.

### Acknowledgments

This research was funded by the Foundation of Henan Educational Committee (19A520002) and the National Natural Science Foundation of China (62002100).

### References

[1] Wei L, Zhang S, Gao W and Tian Q 2018 Person Transfer GAN to Bridge Domain Gap for Person Re-identification *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp 79–88

[2] Wang J, Zhu X, Gong S and Li W 2018 Transferable Joint Attribute-Identity Deep Learning for Unsupervised Person Re-identification *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* pp 2275–84

[3] Yu H X, Zheng W S, Wu A, Guo X, Gong S and Lai J H 2019 Unsupervised person re-identification by soft multilabel learning *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*

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

[5] Ristani E, Solera F, Zou R, Cucchiara R and Tomasi C 2016 Performance measures and a data set for multi-target, multi-camera tracking *Lecture Notes in Computer Science (including
6

He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition vol 2016-Decem pp 770–8

Deng J, Dong W, Socher R, Li L-J, Kai Li and Li Fei-Fei 2010 ImageNet: A large-scale hierarchical image database pp 248–55

Zhong Z, Zheng L, Luo Z, Li S and Yang Y 2019 Invariance matters: Exemplar memory for domain adaptive person re-identification Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition vol 2019-June pp 598–607

Choi Y, Choi M, Kim M, Ha J W, Kim S and Choo J 2018 StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition pp 8789–97

Zhong Z, Zheng L, Zheng Z, Li S and Yang Y 2019 CamStyle: A Novel Data Augmentation Method for Person Re-Identification IEEE Trans. Image Process. 28 1176–90

Li Y J, Yang F E, Liu Y C, Yeh Y Y, Du X and Wang Y C F 2018 Adaptation and re-identification network: An unsupervised deep transfer learning approach to person re-identification IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops vol 2018-June pp 285–91

Lin S, Li H, Li C T and Kot A C 2019 Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification Br. Mach. Vis. Conf. 2018, BMVC 2018

Zhong Z, Zheng L, Li S and Yang Y 2018 Generalizing a person retrieval model hetero- and homogeneously Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol 11217 LNCS pp 176–92