Bias Elimination for Domain Adaptive Pedestrain Re-identification

Jianyang Gu[1,2], Hao Luo[2], Weihua Chen[2], Yiqi Jiang[2], Yuqi Zhang[2], Shuting He[1,2], Fan Wang[2], Hao Li[2], and Wei Jiang[1]

[1] Zhejiang University
[2] Alibaba Group
gu_jianyang@zju.edu.cn
{michuan.lh,kugang.cwh}@alibaba-inc.com

Abstract. This paper presents our proposed methods for domain adaptive pedestrian re-identification (Re-ID) task in Visual Domain Adaptation Challenge (VisDA-2020). Considering the large gap between the source domain and target domain, we focused on solving two biases that influenced the performance on domain adaptive pedestrain Re-ID and proposed a two-stage training procedure. At the first stage, a baseline model is trained with images transferred from source domain to target domain and from single camera to multiple camera styles. Then we introduced a domain adaptation framework to train the model on source data and target data simultaneously. Different pseudo label generation strategies are adopted to continuously improve the discriminative ability of the model. Finally, with multiple models ensmebled and additional post processing approaches adopted, our methods achieve 76.56% mAP and 84.25% rank-1 on the test set.

1 Introduction

Pedestrain re-identification (Re-ID) aims to match specific person identities across multiple cameras. As more and more surveillance cameras are being deployed in cities, pedestrain Re-ID can play an indispensable role in modern security systems. In recent years, deep learning methods have made a significant progress on pedestrain Re-ID task [5,6,10]. However, pedestrain Re-ID still faces many challenges, one of which is the large data amount. As the pedestrain Re-ID task is an open-set problem, it is impossible to manually label all the pedestrain images produced by surveillance cameras day by day. Based on the situation, domain adaptation has attracted much attention on the pedestrain Re-ID task [3,12,4,1,8].

Compared with traditional domain adaptation tasks, domain adaptive pedestrain Re-ID is much harder, as the source domain and target domain share no identical classes. Moreover, in VisDA-2020, a synthetic dataset are provided as the labeled source domain [9] and a real-world dataset is adopted as the target domain, where exists a large domain gap. In this work, we analyzed the biases in domain adaptive pedestrain Re-ID task introduced by different datasets and different cameras, and proposed a domain adaptation framework to solve the problem.

The rest of the paper is organized as follows. In Section 2, the proposed methods is introduced. The experimental results are presented in Section 3. And finally Section 4 concludes the paper.

2 Methods

2.1 Data Generation

Domain adaptive pedestrain Re-ID task is faced with two main biases which will introduce disturbance to the discriminative ability of the model.

The first one is the inter-domain gap between different datasets. In the challenge, the source domain contains synthetic pedestrain images while the target domain is consisted of realistic ones. The huge appearance difference between the two domains brings poor performance to directly using the model trained on source domain for testing. To bridge this domain gap, generative adversarial networks (GAN) is commonly adopted to transfer source domain images to target domain. In the challenge, an SPGAN-transferred dataset is provided [3]. Through transferring, realistic texture from thet target domain is added into the labeled source synthetic images. Therefore, conducting supervised learning on the SPGAN-transferred dataset can gain better discriminative ability on the target domain.

The second one is the intra-domain bias introduced by different cameras, which indicate differences on orientation, illumination, occlusion, resolution and many more conditions. In this work, we introduced starGAN to produce images with different camera styles [16,15]. With a large amount of additional images, the model can generalize better among images captured with different cameras. This part of data will be named CamStyle data in the following sections.

2.2 Baseline Model

As the source domain provides labeled images while the target domain doesn't, we train the baseline model in supervised manner on source domain. Both SPGAN-transferred images and CamStyle images are utilized in the training process. Label-smoothed cross entropy loss is adopted for classification and soft-margin triplet loss is adopted for better clustering performance.

2.3 Domain Adpatation

To better eliminate the inter-domain gap in domain adaptation tasks, we introduced a domain adaptation framework to train the model simutaneously on source domain and target domain. We designed a model with backbone and different classifiers for each domain. With this structure, the network can fully utilize the extracted feature to identify classes from each domain and narrow

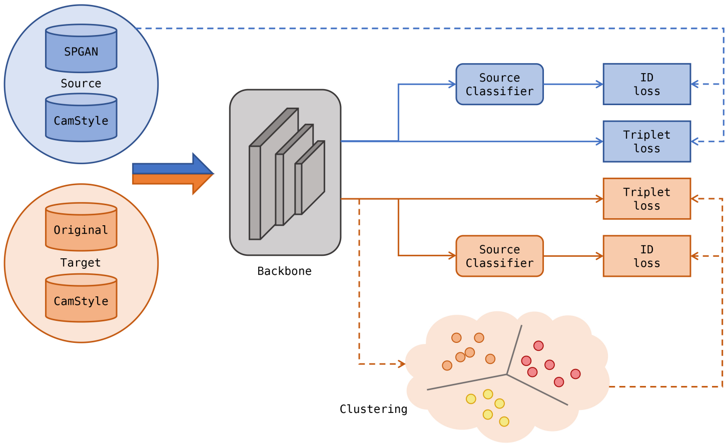

Fig. 1. The pipeline for domain adaptation stage

the feature distribution gap between these two domains. For the source domain, the training procedure is the same as that at baseline model training stage. And for the target domain, an additional clustering operation will be executed to produce pseudo labels. As the exact class number of target domain is unknown, DBSCAN is adopted as the clustering algorithm. The model pipeline is shown in Figure 1, where blue lines indicate source domain data and orange lines indicate target domain data. The dashed lines represent the labels flow. For source domain, the labels come from the dataset, while for target domain, labels come from clustering. The embedded features extracted by backbone network are then passed through corresponding classifiers to get their classification scores.

The domain adaptation training process is separated into two stages with different pseudo label generation strategies. At the first stage, we selected 500 classes with most samples clustered as the target training set and discard the rest. And as the model can better discriminate different identities, the outliers are regarded as classes with few samples. So at the second stage, we added another 200 classes each with one sample into the target training dataset. For the triplet loss calculating, these classes will only contribute to the loss of negative samples. Through the adoption of the two-stage pseudo label generation strategy, the model can continueously improve its performance.

The clustering process is executed every 6 epochs. After the pseudo label generation, the source domain data and target domain data are sampled at a certain rate each to form a mini-batch. And for each mini-batch, original data and CamStyle data are also sampled at a fixed propotion. That means a mini-batch is either composed of source domain data and target domain data, while contains both original data and Camstyle data.

2.4 Post processing

After the features are extracted for testing set, we adopted several post-processing methods to further improve the model performance. The main focus of treatment is on the camera bias, which will largely influence the discriminative ability of the model. Firstly, the mean value of features under the same camera is calculated and subtracted from each feature. Then for each sample, the feature is updated with its closest neighbours. Considering there is no camera label provided in the testing set, we trained an additional camera model to predict the camera label for each image. Besides, inspired by [17], the features extracted by the camera model is utilized to calculate a camera distance matrix, which will be subtracted from the original feature distance matrix at a certain rate. Additionally, we built up a topology map representing the probability of showing up under a certain camera based on the given camera labels in validation set. Images under cameras with larger probability will be assigned larger distance weights. Traditional re-ranking [13] is also adopted to update the distance matrix.

3 Experimental Results

3.1 Implementation Details

The model structure is based on [5,6]. We added an additional linear layer after the feature is extracted to compress the feature dimension to 512. At the baseline model training stage, a 700-class classifier conducts the classification operation for source domain, while at the domain adaptation stage, we designed two classifiers with corresponding dimensions to source domain and target domain. We trained the models on different backbones pretrained on ImageNet [2], among which ResNet50-ibn-a [7], ResNet50-ibn-b [7], ResNet101-ibn-a [7] and HRNetv2-w18 [11] showed better performance when testing on target domain. For data augmentation, we used random horizontal flip, padding and erasing [14].

We utilized SGD optimizer with a original 0.02 learning rate. Warm-up strategy is adopted during the first 10 epochs, and the learning rate is decayed at the $24th$ and $48th$ epochs. The model is trained for 60 epochs in total.

For the standard model training, the images are resized to 384 × 128. We also trained multiple models with different image sizes to compare features in different scales and multiple camera models to better extract the camera features. Finally we fused these models to reach the final score.

3.2 Ablation Study

Effectiveness of Individual Components We compare the evaluation result on validation set to demonstrate the effectiveness of each component in our model structure and the experimental results are summarized in Table 1. In the table, "Direct Transfer" means testing on target domain validation set with model trained on original source domain. It can be seen that directly applying

Table 1. The model performance on validation set with different components

Methods	mAP	Rank-1	Rank-5	Rank-10
Direct transfer	16.2	32.4	50.7	58.6
+ SPGAN	24.7	44.6	62.3	72.4
+ CamStyle	30.7	59.7	77.5	83.3
+ Domain Adpatation	42.2	71.1	83.8	88.9
+ Finetuning	46.3	76.1	86.5	91.0
+ Post Processing	69.4	85.4	91.0	93.1

the model trained on source domain to target domain shows poor performance, with a 16.2% mAP and 32.4% Rank-1. Through the introduction of SPGAN-generated data, part of the domain gap has been narrowed. And by adding extra CamStyle data, the performance is boosted by large margin. It shows that although diminishing the domain gap is effective, the camera bias is also an important issue.

The domain adaptation process further reduces the inter-domain gap, with about 15% growth on mAP and Rank-1. And during the finetuning stage with more samples utilized, there is an additional 2% increase. The post processing methods also play a significant role in the model performance. The experiment is executed on ResNet50-ibn-a backbone, and the stats are similar on the other backbones. Note that the stats are obtained offline, so there might be some differences under different evaluation systems.

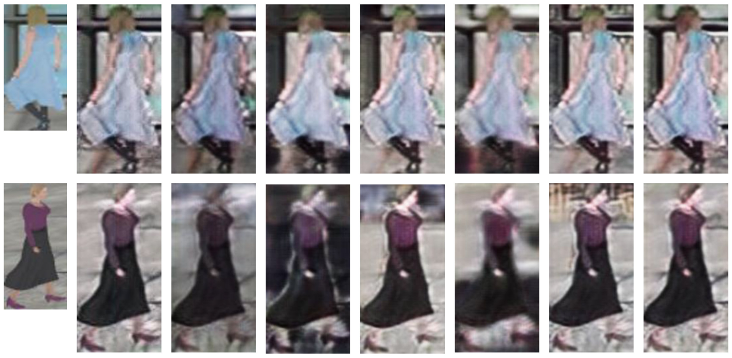

Fig. 2. The generated training data. The first column is from source domain. The second column is from SPGAN-generated data and the other columns are from CamStyle data.

Visualization of Generated Data We selected some of our generated data utilized in training processes to show the influence from data more intuitively. From Figure 2, it can be observed that through the introduction of SPGAN, a large number of texture is added into the images. Add the CamStyle data further improves the diversity of data, where a variety of resolution, illumination conditions are simulated.

Table 2. The performance on testing set with single model and model ensemble

Methods	mAP	Rank-1	Rank-5	Rank-10
Single Model	71.1	79.8	86.8	90.4
Model Ensemble	76.6	84.3	89.6	92.4

Effectiveness of model ensemble With the above generated data and training stages, our best model (ResNet50-ibn-a) can reach about 71.1 mAP and 79.8 Rank-1. We also trained the model with ResNet50-ibn-b, ResNet101-ibn-a and HRNetv2-w18 backbones with different image sizes. Finally we integrated all models and adjusted some post-processing parameters to gain more than 5.5% mAP and 4.7% boost compared to the mean performance of all backbones. Note that our final Rank-1 is the highest among the competitors.

4 Conclusion

We have presented our framework for the domain adaptive pedestrain Re-ID challenge. It mainly focuses on the domain gap and camera bias which would influence the discriminative ability of models in the task. The top performance during the challenge had proved the effectiveness of our proposed methods, which can be further analyzed and better utilized in future works.

References

1. Chen, Y., Zhu, X., Gong, S.: Instance-guided context rendering for cross-domain person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 232–242 (2019)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 994–1003 (2018)

4. Fu, Y., Wei, Y., Wang, G., Zhou, Y., Shi, H., Huang, T.S.: Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6112–6121 (2019)
5. Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Transactions on Multimedia pp. 1–1 (2019). https://doi.org/10.1109/TMM.2019.2958756
6. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
7. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: Enhancing learning and generalization capacities via ibn-net. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 464–479 (2018)
8. Song, L., Wang, C., Zhang, L., Du, B., Zhang, Q., Huang, C., Wang, X.: Unsupervised domain adaptive re-identification: Theory and practice. Pattern Recognition **102**, 107173 (2020)
9. Sun, X., Zheng, L.: Dissecting person re-identification from the viewpoint of viewpoint. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 608–617 (2019)
10. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 480–496 (2018)
11. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence (2020)
12. Zhang, X., Cao, J., Shen, C., You, M.: Self-training with progressive augmentation for unsupervised cross-domain person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8222–8231 (2019)
13. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1318–1327 (2017)
14. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI. pp. 13001–13008 (2020)
15. Zhong, Z., Zheng, L., Li, S., Yang, Y.: Generalizing a person retrieval model hetero- and homogeneously. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 172–188 (2018)
16. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5157–5166 (2018)
17. Zhu, X., Luo, Z., Fu, P., Ji, X.: Voc-reid: Vehicle re-identification based on vehicle-orientation-camera. In: Proc. CVPR Workshops (2020)