Learn By Guessing: A Pseudo-Label Approach For Person Re-Identification

Tiago de C. G. Pereira*, Teofilo E. de Campos†
Departmento de Ciência da Computação
Universidade de Brasília - UnB
Brasília-DF, Brazil
Email: pereira.tiago@aluno.unb.br, t.decampos@oxfordalumni.org†*

Abstract—**Due to application limitations there is a gap between academic research and industry in the field of computer vision. Normally, the best academic results are in well controlled environments and this is not different for person re-identification. This brings an urge to develop more generalized systems or robust domain adaptation techniques to allow person re-identification applications in real world. In this work, we present a set of tools to enhance a model performance in the target domain without the need to annotate any new data. We show the efficiency of our techniques in the Visual Domain Adaptation challenge released by ECCV 2020, where we increase the baseline mAP performance from 21.97% to 59.50%.**

1. Introduction

Person re-identification (reID) is an image retrieval task, where the objective is to match images from the same person in different non-overlapping cameras views. It is an indispensable task for intelligent video surveillance [1] and smart cities [2].

At 2019 CVPR, Luo et al. [3] presented a person reID baseline achieving 94.5% rank-1 accuracy in Market1501 [4] and 86.4% rank-1 accuracy in DukeMTMC-reID [5]. Their work evidence the high standard for supervised person reID.

Despite the high performance obtained by supervised person reID methods, real-world tasks demand algorithms that can perform in a variety of domains (datasets). We attempt to narrow this industry-academy gap using domain adaptation and enhance a baseline algorithm to perform well in other domains.

Domain adaptation for person reID is a trending topic at the moment. In 2020 a Visual Domain Adaptation (VisDA) challenge was launched, the objective was to test person reID methods ability to transfer knowledge from a source domain to novel target domains. They released a source dataset (personX [6]), a SPGAN [7] translated dataset and an unlabeled target training dataset.

Our main contributions with this work are:

- We present an unsupervised domain adaptation framework that achieved competitive results in VisDA 2020 challenge (top 20 in the validation dataset);
- We compare two different clustering algorithms for pseudo-labels generation;
- We propose a feature normalization method to decrease the camera bias in target domain;
- We present a set of post processing tools to enhance our final model results.

2. Related works

One of the approaches for domain adaptation relies on pseudo-labels generation at target domain. For classification tasks, pseudo-labels generation is direct, once you assume the algorithm is correct and label the input using its prediction. Typically, person reID is approached as a metric learning task and the model prediction is not a label, so we use clustering algorithms and define each cluster as a pseudo-label (or person ID).

K-means [8] is a classical clustering algorithm and has been used several times for person reID pseudo-labels generation. It was used by Hehe et al. [9] with one variation, they disregard samples far from centroids in attempt to avoid outliers. Also, they proposed an unsupervised progressive learning method, where they repeated the process of generating pseudo-labels and fine-tuning their CNN until it does not converges anymore.

In [10] we also used k-means for pseudo-labels generation. But we realized that the clusters were mainly formed by images from a single view. This could harm the learning process, once it is important for the CNN to learn camera-camera translation. Therefore, we proposed to use k-means for each camera and then use K nearest neighbors (KNN) to merge clusters from different cameras.

Zeng et al. [11] believe that k-means can not handle the outliers, because these points drag the centroids far from interesting regions. Then, they proposed a minimal spanning tree based clustering method where each image would be its own cluster, which is then merged with the nearest cluster at each iteration. In addition, they used a PK sampling technique, where their pseudo labeled dataset would consider only clusters with at least K samples.

3. Methodology

3.1. Datasets

VisDA challenge organizers released three training datasets and one validation dataset, their statistics are presented in table 1. PersonX dataset is used as source domain to train a baseline model. PersonX SPGAN is a translated dataset that take advantage from PersonX labels and looks similar to target domain. The target domain consisted in 13198 unlabeled images for training and 3977 labeled images for validation.

TABLE 1. Datasets Statistics for VisDA 2020 challenge

Dataset	Nº Images	Labeled
PersonX	20280	✓
PersonX SPGAN	20280	✓
Target Training	13198	✗
Target Validation	3977	✓

3.2. Baseline

For our baseline we used ResNet50-IBN [12] as backbone and trained on the PersonX dataset with an approach similar to Luo et al.'s [3].

We initialized the ResNet50-IBN with weights pretrained on ImageNet and changed the last fully connected layer to output an N dimensional feature, where N is the number of identities in the dataset.

We used PK sampling to create the training batchs, where we choose 16 identities (P) and 4 images (K) from each identity. Then, our batch size is $4 \times 16 = 64$.

Our data augmentation procedure included resizing the image into 256×128 pixels (original size was 128×64 pixels), padding each dimension with 10 zero valued pixels and randomly cropping it with size 256×128. Also, we flipped the image horizontally and random erased [13] it with a 0.5 probability each.

During training, our model outputs a features vector f with 2048 dimensions and an ID prediction logits p. We used the features vector in a batch hard triplet loss($\mathcal{L}_{tri}$) [14] with 0.3 margin and in center loss ($\mathcal{L}_{cent}$) [15]. The perdiction logits were used in a cross entropy label smooth loss($\mathcal{L}_{ID}$) [16]. Our loss function was given by Equation 1.

$$\mathcal{L} = \mathcal{L}_{tri} + \mathcal{L}_{ID} + 0.005\mathcal{L}_{cent} \tag{1}$$

Finally, we used Adam optimizer for 40 epochs and a learning rate schedule to avoid overfitting. Our learning rate started in 10^{-5} and linearly increased to 10^{-4} in the first 10 epochs, then we multiplied the learning rate by a 0.1 factor in epochs 15 and 25. After training in personX dataset we fine tuned our model in personX SPGAN following the same training configurations.

3.3. Pseudo labels

After training our baseline method, we used clustering algorithms to generate pseudo-labels in target training data. Then, we fine tuned our model in this dataset and evaluated it in target validation data. We did these recursively (progressive learning) while the Mean Average Precision (mAP) in target validation dataset kept increasing.

3.3.1. K-means.

3.3.1. K-means. The use of triplet loss for training ensures that the features vector f is part of an Euclidean vector space. Then, we can use k-means to group features from different images and generate pseudo-labels.

As we stated in [10], when working in a new domain there is a camera domain shift, so features from the same camera tend to be nearer than features from different cameras. It is fundamental to overcome it because having images from different views is crucial at the learning stage.

Therefore, we used our strategy [10] to generate identities with examples from different views. Their strategy was to use k-means for each camera view and then use KNN to group clusters from different cameras. In our case, we choose $k = 500$ and have 5 different views, so we applied k-means for each camera and generated $5 \times 500 = 2500$ clusters. Then, we grouped the inter camera clusters to generate the final 500 clusters with images from all the 5 cameras views.

3.3.2. Minimal spanning tree clustering. For the minimal spanning tree clustering, we followed the method presented by Zeng et al. [11]. Their method set each image as a cluster and calculate all inter cluster distances using UPGMA (unweighted pairgroup method with arithmetic means) [17].

Then, they merge a percentage (mp) of the nearest clusters (we used $mp = 7\%$) and repeat this process for s steps (we used $s = 14$). Finally, they use PK sampling (select K images from P identities to form the training batch) to generate the pseudo labeled dataset, so only clusters with at least 4 images are inserted in the dataset.

3.4. Post processing

We used some post processing techniques to enhance our results on the target domain. The three techniques that we used were model ensemble, re-rank and camera normalization.

The model ensemble technique was used to take advantage from both k-means and minimal spanning tree clustering. For this technique, we did a weighted sum with the distance matrix generated by each method and used a grid search to find the best weight. Then, our final distance matrix $\mathcal{M}$ was given by Eq. 2.

$$\mathcal{M} = \alpha\mathcal{M}_{k-means} + (1-\alpha)\mathcal{M}_{hierarchical} \tag{2}$$

As said before, the features space suffer from a camera domain shift, then Zhuang et al. [18] proposed to adapt batch normalization layers, so it normalize features from each specific camera view. With that in mind, we normalized the features from the target validation for each camera before calculating the distance matrix $\mathcal{M}$. Also, we used the re-ranking approach proposed by Zhong et al. [19].

4. Experiments

TABLE 2. RESULTS (IN %) FOR ALL THE IMPLEMENTED METHODS. RR MEANS RE-RANKING AND CN IS CAMERA NORMALIZATION.

		CMC Scores		
Method	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**
Baseline	21.97	40.32	62.33	69.50
Baseline + SPGAN	26.49	46.95	61.80	69.50
Minimal Spanning Tree	35.28	57.82	76.66	81.96
K-Means	37.05	59.95	74.80	81.17
Ensemble	39.48	60.74	76.92	84.88
Ensemble + RR	53.23	64.99	78.51	83.02
Ensemble + RR + CN	**59.50**	**70.29**	**81.43**	**85.68**

The weakness of the baseline result highlights the domain shift between source and target datasets. Our baseline model, trained on the PerxonX dataset, achieved a mAP of 21.97% in the target domain as shown in Table 2.

The PersonX SPGAN used a CycleGAN to translate images from source domain to target domain. The translation reduces the shift between source and target domains, while taking advantage from source domain labels. When fine tuning our baseline at PersonX SPGAN dataset, we increased the mAP in 4.52% and the CMC rank-1 in 6.63%. This is an expected result once the images are more similar to target domain.

Although PersonX SPGAN translates source images to appear target images, training with actual target domain images is even better. Then, we used k-means and minimal spanning tree to generate pseudo-labels on target training images and fine-tuned our model in this pseudo labeled dataset. As one can see in Table 2, the minimal spanning tree clustering technique increased the mAP in 8.79% and the k-means increased it in more than 10%, when comparing to our baseline SPGAN model.

For both pseudo labels methods we used the concept of Progressive Learning, in Figure 1 we show how these methods progressed through each step.

The minimal spanning tree based clustering is designed to disregard outliers, but it was too conservative. It continues converging for 8 steps and even in the last iteration only 4852 from the 13198 images were considered in the pseudo labels. Using just 36.76% of the available target domain data was the reason this method achieved worse results than k-means, maybe a tweak in the parameters of the method could result in better results and faster converging.

The k-means method used don not deal with outliers, so it already uses all images available in the first step. Using all the images is a strong factor for the method quick increase in performance, but it also does not enable progressive learning.

Each clustering technique has its own characteristics, so are the pseudo labels generated by them. Then, it is expected that models trained learn different ways to classify the data. With that in mind, we ensemble the k-means and the minimal spanning tree models and achieved an even better performance. The result presented was achieved with $\alpha = 0.76$.

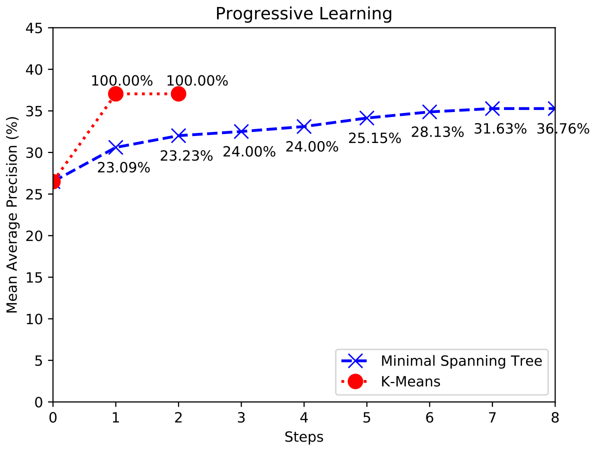

Figure 1. The mAP achieved at each step while using K-Means and Minimal Spanning Tree pseudo-labels methods. The numbers in the graph are the percentage of total available images each method have selected.

As person re-identification is a image retrieval task, re-ranking is critical to improve its accuracy. In re-rank process the similarity relationship from similar samples are captured and k-reciprocal features are considered to match an image, then a more robust matching system is created. The result in Table 2 indicates how essential is the re-ranking with an mAP increase of 13.75% compared with the ensemble method alone.

The biggest challenge in person re-identification is how to encode the person information while leaving camera information out. By normalizing the features by camera we reduce the camera information in the encoding and enhance the model performance. The 6.27% increase in mAP and 5.30% increase in CMC Rank-1 proves that reducing the camera information in the encoding is key to achieve great results in person re-identification.

5. Conclusion

In this work, we show how different person reID datasets can be and how it impact the model performance. Also, we presented some tools to gradually enhance the model performance in the new domain without the burden of annotating it.

Camera domain shift is proven to be a major obstacle to a robust person re-identification model. Therefore, it is interesting to see how camera-based feature normalization enhances the performance on new domains.

Pseudo-labels generation was the base of our domain adaptation techniques and two totally different clustering algorithms were tested. Both methods significantly increased our results, even more when ensemble. So we can assume each one had its own characteristics and complete each other. This is very useful, because this kind of approach may be used in real world to keep improving models performance and adapting itself to new situations with minimal human intervention.

References

[1] Wei Niu, Jiao Long, Dan Han, and Yuan-Fang Wang, "Human activity detection and recognition for video surveillance," in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 1, June 2004, pp. 719–722.

[2] S. Zhang and H. Yu, "Person re-identification by multi-camera networks for internet of things in smart cities," *IEEE Access*, vol. 6, pp. 76 111–76 117, 2018.

[3] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

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

[5] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[7] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[9] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *ACM Transactions on Multimedia Computing, Communications, and Applications TOMM*, vol. 14, no. 4, pp. 83:1–83:18, 2018.

[10] T. Pereira and T. E. de Campos, "Domain adaptation for person re-identification on new unlabeled data," in 15^{th} *International Conference on Computer Vision Theory and Applications - part of VISIGRAPP*, vol. 4: VISAPP, February 27-29 2020, pp. 695–703.

[11] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[12] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[13] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *CoRR*, vol. abs/1708.04896, 2017. [Online]. Available: http://arxiv.org/abs/1708.04896

[14] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," Cornell University Library, Tech. Rep. arXiv:1703.07737, 2017, http://arxiv.org/abs/1703.07737.

[15] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 499–515.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] R. Sokal, C. Michener, and U. of Kansas, *A Statistical Method for Evaluating Systematic Relationships*, ser. University of Kansas science bulletin. University of Kansas, 1958.

[18] Z. Zhuang, L. Wei, L. Xie, T. Zhang, H. Zhang, H. Wu, H. Ai, and Q. Tian, "Rethinking the distribution gap of person re-identification with camera-based batch normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[19] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.