Domain adaptation for person re-identification on new unlabeled data

Tiago de C. G. Pereira[a], Teofilo E. de Campos[b]
Departmento de Ciência da Computação, Universidade de Brasília - UnB, Brasília-DF, Brazil
tiagodecarvalhopereira@gmail.com, t.decampos@oxfordalumni.org

Keywords: Domain Adaptation, Person Re-identification, Deep Learning

Abstract: In the world where big data reigns and there is plenty of hardware prepared to gather a huge amount of non structured data, data acquisition is no longer a problem. Surveillance cameras are ubiquitous and they capture huge numbers of people walking across different scenes. However, extracting value from this data is challenging, specially for tasks that involve human images, such as face recognition and person re-identification. Annotation of this kind of data is a challenging and expensive task. In this work we propose a domain adaptation workflow to allow CNNs that were trained from one domain to be applied to another domain without the need for new annotation of the target data. Our results show that domain adaptation techniques really improve the performance of the CNN when applied in the target domain.

1 INTRODUCTION

The purpose of person re-identification is to match images of persons in non-overlapping cameras views. It can be helpful in some important applications as intelligent video surveillance (Wang, 2013), action recognition (Wei Niu et al., 2004) and person retrieval (Sun et al., 2017).

For problems related to identifying people in images, the first method of choice is usually based on face recognition. This is because such algorithms have already matched the human capacity, as we can see in Taigman et al.'s work (Taigman et al., 2014), where a 97.35% accuracy was achieved in the LFW dataset (Huang et al., 2008) while the human accuracy on the same data is 97.53%. However, face recognition algorithms have little value on surveillance images because the subjects are usually far away from the cameras, so there is not enough resolution in the area of the face. Furthermore, the surveillance viewpoint is usually such that a high amount of (self-)occlusion happens, to the point that the faces are not visible at all. For these reasons, person re-identification algorithms usually take the whole body into account. The typical workflow to train a person re-identification system follows this steps:

1. Use a CCTV system to gather non structured data;
2. Filter this data using a person detector and tracker;
3. Annotate person bounding boxes;
4. Train a metric learning CNN in the annotated data;
5. Deploy the trained CNN to match people that appear in different cameras.

The biggest problem with this workflow is step 3, because CNNs need a huge amount of data to be properly trained and the process of annotating all this data is very expensive (in terms of time and manpower). We therefore propose to replace this step by an unsupervised domain adaptation technique. According to Pan and Yang (Pan and Yang, 2010), domain adaptation is a type of transfer learning where only source domain data is labeled and both domains have the same task.

In our technique, we use a public dataset as our source domain and the non structured data from the CCTV as our target domain. In our source domain all the annotation and image filtering have already been done, then we use unsupervised image-image translation to create an intermediate dataset. This dataset has the labels of the source domain, but the appearance of people is similar to those in the target domain. Next, we proceed to the metric learning step using that intermediate dataset. As the intermediate dataset is similar to the target domain, we expect that the CNN trained in it will perform well in the target domain.

In addition, we use this learned metric to annotate the target domain using a clustering algorithm. That way, we have pseudo labels available for the target

[a] https://orcid.org/0000-0002-9200-9795
[b] https://orcid.org/0000-0001-6172-0229

domain, then we fine tune our CNN in these pseudo labels and learn specific characteristics from the target domain. As the training is performed with the actual target domain images we expect to increase even more our performance, even though the pseudo label generate a noisy label space for the target domain.

In our experiments, we evaluated the CNN performance in the target domain (direct transfer), we evaluated the same CNN trained with the dataset adapted by a cycleGAN to the target distribution and we also evaluated the same CNN trained in the target domain using our pseudo label method. Our method surpasses the baseline accuracy in all test cases. All of that is achieved by replacing step 3 by our technique and will be explained in more details in sections 3 and 4.

In addition, we observed that the highly unbalanced nature of the person re-identification problem means that training batches may be heavily biased towards negative samples. To deal with that, we use a batch scheduler algorithm that allows to train a CNN with a triplet loss in cases where the data is noisy.

Next section discusses related work. Section 3 presents our method and Section 4 presents experiments and results. This paper concludes in Section 5.

2 RELATED WORK

The state-of-art on person re-identification follows a pattern of using either attention-based neural networks (Liu et al., 2017), factorization neural networks (Chang et al., 2018) or body parts detection (Zhao et al., 2017). The common point in these works is trying to disregard the background information, so they can give the proper weight for the image areas where the person is visible. These methods achieve great results, but have a high complexity, as they are based on combinations of several elements. However, different datasets have different characteristics and certain combination of methods may not work across all datasets. In this paper, our focus is on the exploitation of domain adaptation for this application. To design more controlled experiments, we use a relatively simple end-to-end system based on the ResNet-50 (He et al., 2016) as a backbone.

Typically, the person re-identification challenge is approached as a metric learning task (Zhao et al., 2017; Deng et al., 2018). But it can also be approached as a classification task where each person from the dataset is a class (Liu et al., 2017; Chang et al., 2018). The problem of the classification-based approach is that the space of labels is fixed and has a large cardinality. Such methods are rarely applicable in practice, unless the set of identities of people who transit through a set of environments is always the same. Our target application is public spaces, therefore it is not possible to restrict the set of labels. Therefore we approach this as a metric learning challenge. Further to being applicable to public spaces, the task of comparing samples is the same across different domains. This enables the application of unsupervised domain adaptation methods to adapt the marginal distribution of the data.

Recently, some works presented domain adaptations techniques for person re-identification. (Zhao et al., 2017) created a new dataset to evaluate the generalization capacity of his model. Their CNN was evaluated in it without further training. (Zhong et al., 2018) used a cycleGAN to approximate the camera views in a dataset trying to learn a camera latent space metric. (Xiao et al., 2016) trained his CNN with a super dataset created concatenating multiple datasets. They proposed a domain guided dropout to further specialize their CNN for each dataset. In this work, we consider that the target domains have no labeled data, then we cannot use the approaches of (Zhong et al., 2018) or (Xiao et al., 2016). The approach of (Zhao et al., 2017) can be called direct transfer, because it just evaluates a CNN on a target domain. We shall demonstrate that our method outperforms direct transfer.

3 PROPOSED METHOD

Our technique is based on training a CNN to learn a metric, so we can ensure that distinct domains will have the same task. Therefore, we train a ResNet-50 (Section 3.1) with the triplet loss (Section 3.2) to learn the desired metric in an Euclidean vector space. The core of the domain adaptation method is based in a cycleGAN that will perform an image-image translation to approximate source and target domains (Section 3.3). Then, we use the CNN trained in the intermediate dataset to extract the features of the target domain images and use a clustering algorithm to generate pseudo-labels for the target domain (Section 3.4).

3.1 Baseline CNN

As said in Section 2, the state-of-art in person re-identification use techniques that exploit information from CNNs at multiple levels, bringing multiple semantic levels to the final features. Those semantic levels may carry specific person attributes like gender, clothing, textures and clothing, which are important for matching people across views.

We choose ResNet-50 architecture in our work because we believe that residuals blocks help to propagate information from multiple semantic levels when they are relevant for the output. Although the residual blocks may not perform as well as a specific architecture, the main goal of our work is to propose a domain adaptation workflow.

To have an initial boost (Donahue et al., 2014), we start with a ResNet-50 CNN pre-trained on ImageNet (Deng et al., 2009). We then transfer learn it to the problem of person re-identification using a public dataset. This is done by replacing the last fully connect layer by a new fully connected layer with 128 features which are used as an embedding for metric learning. We use Adam optimizer and the triplet loss.

3.2 Triplet Loss and Batching Strategies

A siamese-like loss is ideal when trying to learn a metric because it allows one to perform an end-to-end learning from a dataset to an embedding space. The siamese loss receives as input a pair of feature vectors and tries to approximate them if they are from the same person or set them apart if they are from different people. This generates an embedding space where feature vectors from the same person tend to lie near each other.

The triplet loss is an upgrade from the siamese loss which instead of using a pair of samples as input, it uses an anchor, a positive sample and a negative sample. Therefore, the triplet loss approximates feature vectors from the same person while it also separates features of different people, according to equation 1 (defined for each anchor sample $\mathbf{x}_a$). This way, one can expect better samples separation in the embedding space.

$$\mathcal{L}(\mathbf{x}_a) = \max\left(0\ ,\ m + D\Big(\mathbf{f}_a, \mathbf{f}_p\Big) - D\Big(\mathbf{f}_a, \mathbf{f}_n\Big)\right), \quad (1)$$

where m is a margin so the loss does not go to zero, $\mathbf{f}$ is the CNN output, i.e., a lower dimensional embedding of image $\mathbf{x}$; (sub indexes a, p and n mean anchor, positive and negative, respectively) and $D(\cdot)$ can be any distance measurement algorithm, in our case is the Euclidean distance defined by

$$D(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^{d} (u_i - v_i)^2}. \quad (2)$$

A question that arises from the triplet loss use is "how to choose the positive/negative examples?" (Hermans et al., 2017) investigated this problem and came to a conclusion that the best learning is achieved when using the hardest positive/negative samples during training. This approach was coined *batch hard* and it works as follows: for each anchor sample $\mathbf{x}_a$ from the batch, the choice of positive sample $\mathbf{x}_p$ is chosen as the one that maximizes $D(\mathbf{f}_a, \mathbf{f}_p)$ and the negative sample $\mathbf{x}_n$ is chosen as the one that minimizes $D(\mathbf{f}_a, \mathbf{f}_n)$. Using this strategy, equation 1 can be rewritten as

$$\mathcal{L}_{BH}(\mathbf{x}_a) = \max\left(0\ ,\ m + \max_p D\Big(\mathbf{f}_a, \mathbf{f}_p\Big) - \min_n D\Big(\mathbf{f}_a, \mathbf{f}_n\Big)\right), \quad (3)$$

where positive and negative samples are chosen within each batch and the losses across all anchors in a batch are averaged out.

Figure 1 illustrates how samples are chosen for a batch. All the rectangles at the top represent samples from a person and the rectangles at the bottom represent sample of another person. The triplet will choose each rectangle as anchor at a time, calculate the loss for it and in the final sum all the losses. From the green rectangle as an anchor, the numbered arrows indicate the distance $D(\cdot)$ from it to the samples, where Pos_i, i $= 1, 2, 3$, are possible positive samples and Neg_j, j $= 1, 2, 3, 4$, are the possible negative samples. In a batch hard approach, Pos_2 is selected as positive sample, Neg_3 as negative sample and $\mathcal{L}_{BH} = m + 0.361 - 0.490$.

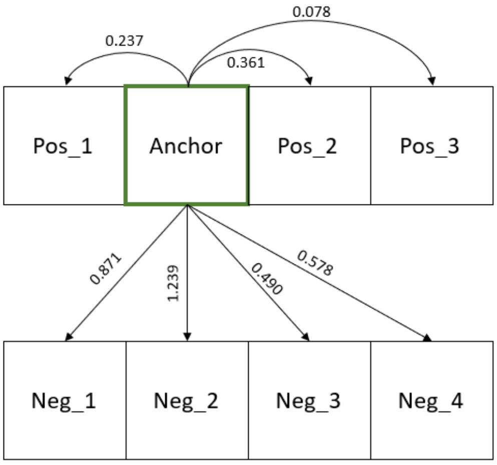

Figure 1: Example of a batch hard triplet selection.

(Hermans et al., 2017) proved the batch hard effectiveness, but choosing the hardest samples at each batch increases the training complexity. Furthermore, we work with an intermediate dataset that can be noisy, meaning that the separation between positive and negative samples may be less trivial, which increases the training cost even more. The consequence is that the training process may never converge with this strategy. When using the triplet loss, a non converging training process can be identified if the loss is stuck at the margin (m), because that means $D(\mathbf{f}_a, \mathbf{f}_p) = D(\mathbf{f}_a, \mathbf{f}_n)$, meaning that all the features are converging to vectors of 0s.

While training with the triplet loss, the goal is to make $D(\mathbf{f}_a, \mathbf{f}_p) < D(\mathbf{f}_a, \mathbf{f}_n)$. However, if the batch is big, the number of negative examples is way bigger than the number of positive examples, particularly in the case of person re-identification. It is therefore possible to have a negative sample that is nearer to the anchor than the hardest positive sample. This way the loss will always be greater than the margin ($\mathcal{L}_{BH} > m$), then the optimizer learns that outputting vectors of 0s will reduce the loss to the margin, i.e., ($\mathcal{L}_{BH} = m$).

Our solution was to use a batch scheduler algorithm to decrease the number of negative samples and lower the training complexity. This way we ease the training convergence, and once the training is converging we slowly increase the batch size (and therefore its complexity, having an impact in the loss). This enables us to learn step by step and converge the training even with a noisy dataset. Our batch scheduler algorithm is shown in Algorithm 1.

Algorithm 1 Batch Scheduler

```
$batch\_size = 8$
$m = 0.5$    // $m$ is the loss margin of Eq. 1
for $i = 0$ to $num\_epochs$ do
    $loss = train(i, batch\_size)$
    if $loss < 0.8 \times m$ then
        $batch\_size = batch\_size + 8$
    end if
end for
```

3.3 Image-Image Translation for Domain Adaptation

To give some background, the definitions and notations used in this paper are based on (Csurka, 2017) and (Pan and Yang, 2010). A domain $\mathcal{D}$ is composed of a d dimensional feature space $\mathcal{X} \subset \mathbb{R}^d$ with a marginal probability distribution $P(\mathbf{X})$ and a task $\mathcal{T}$ defined by a label space $\mathcal{Y}$ and the conditional probability distribution $P(\mathbf{Y}|\mathbf{X})$, where $\mathbf{X}$ and $\mathbf{Y}$ are sets of random variables (which usually are multivariate). Given a particular sample set $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \in \mathcal{X}$, with corresponding labels $\mathbf{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_n\} \in \mathcal{Y}$, $P(\mathbf{Y}|\mathbf{X})$ in general can be learned in a supervised manner from these feature-label pairs $\{\mathbf{x}_i, \mathbf{y}_i\}$.

For simplicity, let us assume that there are two domains: a source domain $\mathcal{D}^s = \{\mathcal{X}^s, P(\mathbf{X}^s)\}$ with $\mathcal{T}^s = \{\mathcal{Y}^s, P(\mathbf{Y}^s|\mathbf{X}^s)\}$ and a target domain $\mathcal{D}^t = \{\mathcal{X}^t, P(\mathbf{X}^t)\}$ with $\mathcal{T}^t = \{\mathcal{Y}^t, P(\mathbf{Y}^t|\mathbf{X}^t)\}$. Those domains are different $\mathcal{D}^s \neq \mathcal{D}^t$, because $P(\mathbf{X}^s) \neq P(\mathbf{X}^t)$ due to domain shift. Also, we do not have the target domain labels $\mathbf{Y}^t$, so we do not have the feature-label pairs $\{\mathbf{x}_i, y_i\}$ to learn $P(\mathbf{Y}|\mathbf{X}^t)$ in a supervised manner.

The person re-identification task $\mathcal{T}$ consists in learning a projection from $\mathbf{x} \in \mathcal{X}$ to a feature $\mathbf{f}$ in a Euclidean space where $\mathbf{f}$ is closer to other vectors if they originated from the same person, more distant to vectors from other people. The set of labels can be thought of as the space of all possible person identities in the world, which impractical. Alternatively, the person re-ID problem can be seen as a binary problem that takes two samples as input, indicating whether or not they come from the same person. Therefore, each person re-ID dataset (or indeed each camera surveillance environment) can be seen as a different domain, however the task is always the same, i.e., telling if two images contain the same person or not. Domain adaptation are transductive transfer learning methods where it is assumed $\mathcal{T}^s = \mathcal{T}^t$, according to Csurka (Csurka, 2017). Therefore, we can use domain adaptation to exploit the related information from $\{\mathcal{D}^s, \mathcal{T}^s\}$ to learn $P(\mathbf{Y}^t|\mathbf{X}^t)$.

In our method, we have images from source domain $\mathbf{X}^s$ and target domain $\mathbf{X}^t$, but we do not have the labels from target domain $\mathcal{Y}^t$. So, we approximate data from images of a known source domain to images of a target domain generating an intermediate dataset.

We use, as source domain, a public dataset which has ground truth annotation of positive/negative examples for each anchor. An unsupervised domain adaptation method can be used to generate an intermediate dataset $\mathcal{D}^i$ that leverages the source domain annotation $\mathcal{Y}^s$ and is similar to the target domain. For that, we follow an approach based on Generative Adversarial Networks – GANs (Goodfellow et al., 2014). More specifically, we use the cycleGAN method proposed by (Zhu et al., 2017) and applied to person re-identification by (Deng et al., 2018).

The idea is to use images from the source domain ($\mathbf{X}^s$) as input and train a GAN to generate outputs which are similar to the images from the target domain ($\mathbf{X}^t$). However, once we have no paired images between domains the problem has a high complexity. Zhu et al. proposed to train two generators G and F where $G : \mathcal{X}^s \to \mathcal{X}^t$ is a mapping from the source domain to the target and $F : \mathcal{X}^t \to \mathcal{X}^s$ is a mapping from the target domain to the source. Also, a cyclic component is added to the loss:

$$\begin{aligned}\mathcal{L}(G,F,D_{X^s},D_{X^t}) = \mathcal{L}_{GAN}(G,D_{X^t},\mathbf{X}^s,\mathbf{X}^t)+& \\ \mathcal{L}_{GAN}(F,D_{X^s},\mathbf{X}^t,\mathbf{X}^s)+& \qquad (4)\\ \lambda\mathcal{L}_{cyc}(G,F),&\end{aligned}$$

where both $\mathcal{L}_{GAN}$ components are the basic GAN loss proposed by Goodfellow et al. and the $\mathcal{L}_{cyc}$ is the cyclic component added by Zhu et al., wich is given by:

$$\begin{aligned}\mathcal{L}_{cyc}(G,F) = E_{\mathbf{X}^s\sim p_{data}(X^s)}\left[\|F(G(\mathbf{X}^s))-\mathbf{X}^s\|_1\right]+& \\ E_{\mathbf{X}^t\sim p_{data}(X^t)}\left[\|G(F(\mathbf{X}^t))-\mathbf{X}^t\|_1\right]& \quad (5)\end{aligned}$$

the cyclic component is there to do an identity match between source domain images $\mathbf{X}^s$ and their double transformed pairing images $F(G(\mathbf{X}^s))$, and vice-versa. By minimizing this cyclic loss we expect to have transformations that can map both domains.

Therefore, we use the generator $G: X^s \rightarrow X^t$ in all images of our source domain to generate an intermediate dataset. That is, we create a dataset that leverages from the labeled data of the source domain and have similar characteristics to the target domain. This way we can expect that a training in this intermediate dataset will perform well in the target domain.

3.4 Pseudo Labels for Re-Identification

In Section 3.2, we used the triplet loss to learn a distance metric in an Euclidean vector space. In Section 3.3, we showed that both source and target domains have the same label space $\mathcal{Y}$. We also presented a method to train our CNN in an intermediate dataset that leverages from the labeled data of the source domain and have similar characteristics to the target domain. The CNN therefore should already present a reasonable performance in target domain.

We use the CNN to extract all features $\mathbf{f}_i^t$ from target domain images $\mathbf{X}^t$ and these features belong to an Euclidean vector space. Then, we used a clustering algorithm to group these features, using the obtained group identifications as target domain with pseudo-labels $\mathbf{Y}^t$. In addition, we fine tune the CNN using the feature-label pairs $\{\mathbf{x}_i, y_i\}$ with the real images from target domain and the pseudo-labels generated by the clustering algorithm.

Even though the pseudo labels generated may contain a lot of errors, this next training step uses the real images from target domain $\mathbf{X}^t$. Therefore, the CNN is be able to learn more robust features for the target domain, because it learns the exact characteristics of the target domain.

We choose the k-means (Hartigan and Wong, 1979) clustering algorithm to group the features in the Euclidean vector space. The value of k was chosen as a proportion of the size of each target dataset. Table 1 indicates the values used in this paper (the datasets are discussed later). However, the naive assignment of samples to clusters is a flawed strategy to annotate the data, because a simple look at the data may cluster viewpoints rather than people. In other words, features from different people taken from the same camera view are often more similar to each other than features from the same person from different camera views.

Table 1: The chosen k for each dataset when using k-means algorithm.

Dataset	k
CUHK03	2000
Market1501	1600
Viper	632

Our solution is to use k-means algorithm to generate k clusters for each camera view, then use a nearest neighbor algorithm to group these clusters across the camera views. This way, we guarantee that every person from our pseudo-labels space have images from each camera. That results in a noisy annotation, because that assumption is not a true in the real label space of the dataset. However, using this approach we ease the CNN task of learning features robust for multiple cameras views and achieve better results in validation.

4 EXPERIMENTAL RESULTS

In our work, we produced results using three well known person re-identification datasets, they are the CUHK03 (Li et al., 2014), the Market1501 (Zheng et al., 2015) and the Viper (Gray et al., 2007). For all the experiments, we did not use any label information in the target domain, except to evaluate the results.

Our work produces two kind of results that must be analyzed to understand the method effectiveness. These results are the generation of a intermediate dataset (discussed in section 4.1) and the CNN evaluation in the target domain after the complete workflow was done using pseudo-labels (discussed in section 4.2).

4.1 Intermediate Dataset

As said in section 3.3 our method tries to approximate the source domain to the target domain. This is done training a cycleGAN between both domains and using the generator to create an intermediate dataset that shifts the source domain samples so that they become more similar to the target domain data. The idea is to generate images that preserve the person morphology, but are visually adapted to the target domain. While there is no guarantee that a GAN preserves person morphology, the cyclic loss contributes towards this goal, as it has an identity match component.

Figure 2 presents examples of transformation results between all domains. It is interesting to note that the person morphology have been well preserved and the changes have been more in the colors, texture and background. That means we could produce a great approximation of how a person would look like in the view of another dataset.

The CUHK03 dataset was created using surveillance cameras from a university in Hong Kong with an elevated viewpoint, so normally the background of their images consists in a granular floor. While the Market1501 dataset was created with cameras in a park, so the images usually have grass in the background of their views. Viper is the oldest dataset used in this work, it was published in 2007 and is composed of low resolution outdoor images.

These characteristics of the datasets make it easy to understand the effects seen in Figure 2. When using CUHK03 as the target domain, the transformed images tend to have a granular background to approximate the floor texture in CUHK03 images. When using Market1501 as target domain, images from CUHK03 had a background transformation from the granular floor to grass, and images from Viper had just a color transformation, because both datasets are from outdoor images. When using Viper as target domain, images from Market1501 had a color transformation and images from CUHK03 had a texture background transformation and a brightness enhancement.

4.2 Domain Adaptation Results

4.2.1 Image-Image Translation Method

After successfully generating an intermediate dataset that approximates both domains we used that intermediate dataset to fine-tune the CNN trained in the source domain. We evaluated all the results in the target domain using the CMC score with rank-1, rank-5 and rank-10.

The cycleGAN method was compared with the direct transfer method, where the direct transfer method consists in evaluating in the target domain a CNN trained in the source domain without further training. The direct transfer method therefore shows how different are both domains and is used as a baseline.

As one can see in Table 2 the cycleGAN method presents huge rank-1 improvements when using CUHK03 as target domain (26% improvement for Viper as source domain and 14.9% improvement for Market1501 as source domain). This happens because the CUHK03 images have granular background texture as a strong characteristic that was easily learned by our cycleGAN.

A great rank-1 improvement was also obtained for Market1501 as target domain and CUHK03 as source domain, where the cycleGAN method achieved a 9% improvement compared with the baseline. Furthermore, for Market1501 as target and Viper as source domain our method achieved 1% improvement, meaning that the color transformation helped to approximate these domains, but this was not as significant as texture changes that occurred when working with CUHK03 images.

For Viper as a target domain the cycleGAN method achieved 1.5% rank-1 improvement using CUHK03 as source domain and 1.9% rank-5 improvement for Market1501 as source domain. Again, this means that texture transformations are more significant than color transformations. Although those are not our best results, they are very significant because as Viper is an old dataset it has a lot less images than the others (only 1264 images), so learning to create the intermediate dataset in a unsupervised manner without much data is extremely hard.

4.2.2 Pseudo-Labels Method

Section 4.2.1 proved the effectiveness of domain adaptation and that the cycleGAN successfully shifted images to the target domain appearance, carrying their source label with them. Also, it was clear that texture transformations are more significant than color transformations.

Although the cycleGAN did a great job shifting images between domains, when using the pseudo-labels method we achieved even better results. This is because the training is now performed with the actual target domain images and estimated pseudo-labels. So, there is no longer the problem of images in which the person morphology was not preserved. The target dataset characteristics are better represented. Figure 3 illustrates the dataset created using pseudo-labels – as one can see the estimated labels are not perfect, but the grouped images show a strong color similarity.

As one can see in Table 2, our method showed great improvements in all test cases. Even when using the Viper dataset as target domain our method could improve the cycleGAN results in 2% or more. For the Market1501 dataset the rank-1 improvement was around 2% also and for the CUHK03 our method achieved improvements of 4% in rank-1 accuracy.

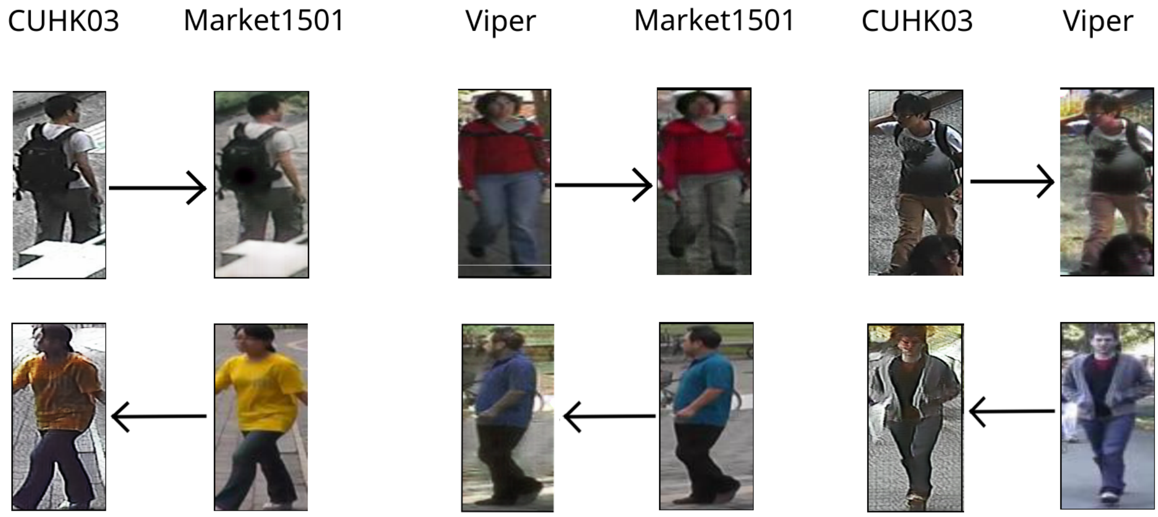

Figure 2: Examples of the cycleGAN transformations between domains.

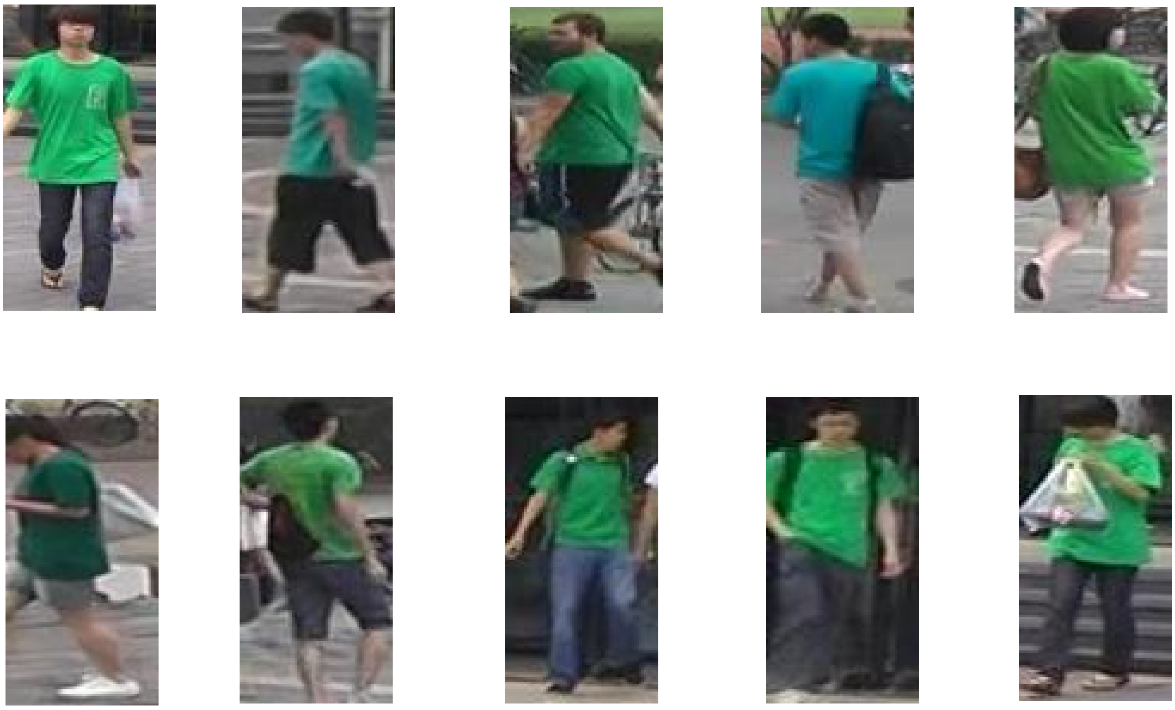

Figure 3: Images from a final cluster when using the pseudo-labels method. The cluster were achieved using Viper as source dataset and Market1501 as target dataset.

It is important to notice that the pseudo-labels have a stronger positive impact on smaller target datasets. This is because small datasets require fewer clusters to annotate the data. This was very significant for the great results presented for Viper dataset.

In summary our method is significantly better than direct transfer without adaptation. It is important to emphasize that our method does not make use of any label from the target domain, completely removing the burden of annotating new data when the application domain changes.

5 CONCLUSIONS

In person re-identification, each type of environment (e.g. airport, shopping center, university campus, etc.) has its own typical appearance, so a system that is trained in one environment does not perform very well in another environment. This observation was confirmed by our cross-dataset (direct transfer) experiments, indicating that each dataset can be treated as a domain. Therefore, we showed that a domain adaptation method based on cycleGAN can be applied to transform the marginal distribution of samples from a source dataset to a target dataset. This enables us to retrain a triplet CNN on adapted samples so that their performance is improved on the target dataset without using a single labeled sample from the target set. Furthermore, we showed that using this CNN and a clustering algorithm to generate pseudo-labels and retrain the triplet CNN leads to a significant boost in the performance on target dataset. This opens doors for the deployment of person re-ID software to real applications, as it completely removes the burden of annotating new data.

Further to proposing a domain adaptation technique for this problem, we also presented the use of a batch scheduler which increases the batch size as training starts to converge.

For future works, we believe it would be interesting to try our technique with other datasets, using more robust CNN architectures as backbone and with different clustering algorithms. But it is proved that this technique brings great contribution to the field of person re-identification.

ACKNOWLEDGEMENTS

The authors would like to thank FAPDF (fap.df.gov.br) and CNPq grant PQ 314154/2018-3 (cnpq.br).

REFERENCES

Chang, X., Hospedales, T. M., and Xiang, T. (2018). Multi-level factorisation net for person re-identification. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR)*.

Csurka, G. (2017). A comprehensive survey on domain adaptation for visual applications. In Csurka, G., editor, *Domain Adaptation in Computer Vision Applications*, pages 1–35. Springer International Publishing, Cham.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., and Jiao, J. (2018). 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)*.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, Bejing, China. PMLR.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems (NIPS) 27*, pages 2672–2680. Curran Associates, Inc.

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro*.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hermans, A., Beyer, L., and Leibe, B. (2017). In defense of the triplet loss for person re-identification. Technical Report arXiv:1703.07737, Cornell University Library. http://arxiv.org/abs/1703.07737.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR)*, pages 152–159.

Table 2: Results with all domains combinations.

			Accuracy		
Target Domain	**Source Domain**	**Method**	**Rank-1**	**Rank-5**	**Rank-10**
Market1501	**Viper**	Direct Transfer	5.7%	15.5%	22.2%
		CycleGAN	6.7%	17.0%	23.7%
		Ours	**8.6**%	**20.5**%	**28.4**%
	CUHK03	Direct Transfer	26.8%	45.9%	55.1%
		CycleGAN	35.8%	56.5%	65.7%
		Ours	**37.3**%	**60.4**%	**70.4**%
CUHK03	**Viper**	Direct Transfer	5.9%	18.1%	29.0%
		CycleGAN	31.9%	64.4%	77.5%
		Ours	**36.1**%	**69.2**%	**81.3**%
	Market1501	Direct Transfer	19.9%	49.4%	63.2%
		CycleGAN	34.8%	66.7%	79.1%
		Ours	**38.2**%	**69.7**%	**81.6**%
Viper	**CUHK03**	Direct Transfer	10.1%	22.5%	29.0%
		CycleGAN	11.6%	25.5%	34.7%
		Ours	**13.6**%	**33.9**%	**46.0**%
	Market1501	Direct Transfer	12.5%	25.0%	33.1%
		CycleGAN	9.8%	26.9%	36.4%
		Ours	**13.9**%	**29.0**%	**40.7**%

Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., and Wang, X. (2017). Hydraplus-net: Attentive deep features for pedestrian analysis. In *The IEEE International Conference on Computer Vision (ICCV)*.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Sun, Y., Zheng, L., Deng, W., and Wang, S. (2017). Svdnet for pedestrian retrieval. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3820–3828.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, X. (2013). Intelligent multi-camera video surveillance: A review. *Pattern Recognition Letters*, 34(1):3 – 19. Extracting Semantics from Multi-Spectrum Video.

Wei Niu, Jiao Long, Dan Han, and Yuan-Fang Wang (2004). Human activity detection and recognition for video surveillance. In *IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, volume 1, pages 719–722 Vol.1.

Xiao, T., Li, H., Ouyang, W., and Wang, X. (2016). Learning deep feature representations with domain guided dropout for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., and Tang, X. (2017). Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proc 30th IEEE Conf on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26*.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*.

Zhong, Z., Zheng, L., Zheng, Z., Li, S., and Yang, Y. (2018). Camera style adaptation for person re-identification. In *Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*.