

Online Unsupervised Domain Adaptation for Person Re-identification

Hamza Rami^{1,2}, Matthieu Ospici², Stéphane Lathuilière¹

¹LTCI, Télécom Paris, Institut Polytechnique de Paris.

²Atos.

{hamza.rami, stephane.lathuiliere}@telecom-paris.fr, matthieu.ospici@atos.net

Abstract

Unsupervised domain adaptation for person re-identification (Person Re-ID) is the task of transferring the learned knowledge on the labeled source domain to the unlabeled target domain. Most of the recent papers that address this problem adopt an offline training setting. More precisely, the training of the Re-ID model is done assuming that we have access to the complete training target domain data set. In this paper, we argue that the target domain generally consists of a stream of data in a practical real-world application, where data is continuously increasing from the different network's cameras. The Re-ID solutions are also constrained by confidentiality regulations stating that the collected data can be stored for only a limited period, hence the model can no longer get access to previously seen target images. Therefore, we present a new yet practical online setting for Unsupervised Domain Adaptation for person Re-ID with two main constraints: **Online Adaptation and Privacy Protection**. We then adapt and evaluate the state-of-the-art UDA algorithms on this new online setting using the well-known Market-1501, Duke, and MSMT17 benchmarks.

1. Introduction

Person Re-Identification (Re-ID) aims at recognizing a query of person-of-interest within a gallery of images. Therefore, we can determine whether this person has appeared in another place and/or has been captured by a different camera, or even the same camera at different times. The most typical scenario where Person Re-ID is used is video surveillance. A Person Re-ID model can be used to track people on a network of cameras or to find them given only a picture of them. Re-ID is therefore essential to many important surveillance applications.

Recent advances in deep learning have improved the performance of the general Re-ID. Besides specialized architectures and algorithms, much of the recent progress can be attributed to the availability of large and annotated datasets

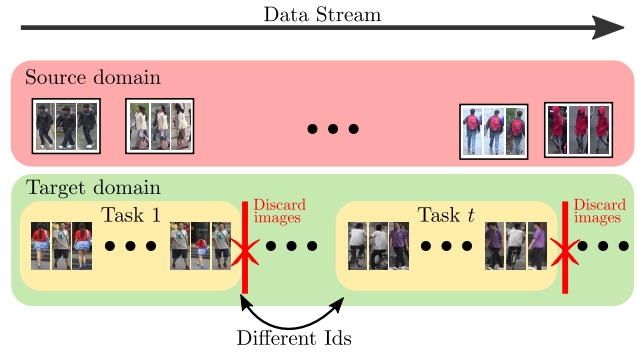


Figure 1. Illustration of the proposed OUDA for Re-ID setting: in Online Unsupervised Domain Adaptation for Person Re-ID, the annotated source dataset is available at any time while the target dataset is divided into annotated tasks. In between each task, the data from the previous task is discarded.

[25, 29, 34]. As obtaining annotated data for Person Re-ID is tedious and costly, pre-trained models are often deployed in new test environments without any adaptation mechanism. However, these models, which work well when the test samples are collected in the same environment as the training data, suffer from important performance degradation when they are deployed in different locations or different camera networks.

To bridge this gap several **Unsupervised Domain Adaptation** (UDA) approaches have been proposed in the literature. We can distinguish two major categories of UDA methods for Person Re-ID: **domain translation-based methods** [5, 29, 36] and **pseudo-label based methods** [7, 8, 10, 22, 27, 32]. Domain translation methods are based on the style transfer from the target domain to the source domain. Methods based on pseudo-labels generate pseudo-labels using clustering algorithms on the target domain and employ cluster index as class labels to perform adaptation [7, 10–12].

Despite their relative efficiency, these UDA methods are all based on the assumption that we have access to a large set of samples from the target domain environment during the training to perform adaptation in an *offline* fashion. In

this work, we argue that this assumption is violated in many real-world scenarios. First, when deploying a Person Re-ID system, we gather images as long as they are recorded under the form of a stream that continuously provides images from different cameras/places. The *offline* process implies that the Re-ID system requires a possibly long data collection phase before deployment. Second, since the Re-ID task evolves person identities, the system is confronted to confidentiality purposes in many countries, forcing the technology provider of such models to discard previously seen images. Therefore, we argue that to match practical scenarios, an unsupervised domain adaptation method for Person Re-ID should respect two main constraints: 1) **Online adaptation**: the target domain data is not accessible all at once, but in a stream fashion where only small batches of images are available at a given instant of time and 2) **Privacy protection**: Images captured by the different cameras can be used to update the Re-ID model and stored for only a limited period of time. To this end, in this paper, we propose and study a practical scenario for Unsupervised Person Re-ID which is the Online Unsupervised Domain Adaptation setting for Person Re-ID (OUDA-Rid). Fig. 1 gives an illustration of the proposed online setting, where we assume that the model has full access to the well-annotated source data set, however, unlike all the previous methods, the target domain dataset is fed to the Re-ID model in an online fashion. Practically, the target domain will be divided into several unlabeled subsets of images, where each subset will be viewed by the Re-ID model only once, hence respecting the two constraints: *online adaptation* and *privacy protection*. Regarding evaluation, we consider an independent and fixed target dataset with identities that do not overlap with any of the training tasks.

Our contributions can be summarized as twofold:

- We propose a new challenging yet practical scenario, the Online Unsupervised Domain Adaptation (OUDA) setting for Person Re-ID to respect two main constraints: Online Adaptation and Privacy Protection of identities.
- We adapt and evaluate three existing frameworks for *offline* UDA to the proposed OUDA setting: the Strong Baseline [7], MMT [11] and SpCL [12]. These methods are evaluated in four different adaptation settings based on three public and widely-used datasets: Market 1501 [34], DukeMTMC-reID [25] and MSMT17 [29]. Our results provide some interesting experimental conclusions regarding the performance and limitations of existing approaches. We hope that this work will stimulate the community to address domain adaptive Re-ID in the OUDA setting.

2. Related Work

Unsupervised domain adaptation (UDA) for person Re-identification has been recently gaining a lot of attention for its practical applications. UDA methods can transfer learned knowledge from an annotated source domain to an unlabeled target domain, thus reducing the cost and discarding the need to have a well-annotated data set. Most of the existing methods and approaches in this area can be divided into two main categories: Domain translation-based and Pseudo label-based methods.

Domain translation-based methods employ style transfer methods to modify the source images to obtain images with the content of the source domain but the appearance of the target. In this way, they obtain images similar to the target with the corresponding label annotations from the source images. These generated images are then used to refine the network parameters. Recent works in this category investigate the integration of generative models [23] [13] [2], as an example, we have [29] which is based on CycleGAN [38] to bridge the domain gap by transferring persons from the source domain to the target. [5] also generates images while preserving the self-similarity of the images before and after the translation and the domain-dissimilarity of the translated source images to the target images. And finally, we can cite the work of Zhong et al. [36] where the proposed framework learns camera-invariant features while enforcing domain connectedness, where two images, one from the source domain and the other one from the target domain, are fed to the network as a negative pair of images.

Pseudo-labeling methods, also called clustering-based methods, employ an iterative process alternating between clustering and finetuning [3, 8, 22, 27, 32]. In its most simple implementation, [7], the cluster indexes obtained in the cluster stage are used as labels to fine-tune the Re-ID network. Despite its simplicity, this simple approach obtains satisfactory results but suffers from limitations that have been addressed in recent works. For instance, Yang Fu et al. proposed a Self-Similarity Grouping (SSG) [10] approach that assigns different pseudo-labels to both global and local features. To mitigate the effects of noisy hard pseudo-labels, Mutual-Mean Teaching (MMT) [11], proposed by Yaxiao et al., adopts a teacher-student framework with two networks that are trained jointly using hard pseudo labels generated by the two networks and soft pseudo labels generated by their Mean Networks, to conduct pseudo-label refinement in the target domain. Moreover, we can mention the work done by Ger et al referred to as SpCL [12] that, unlike previous methods, takes advantage of both labeled source domain images' centroids and un-clustered target instances, stored in a hybrid memory, in addition to the target domain clusters. The memory gives more supervision to the feature extractor during training while minimizing the unified contrastive loss over the three kinds of informa-

tion available in the hybrid memory. Importantly, pseudo-label-based methods achieve better results than translation approaches and maintain up until now the state-of-the-art performances on almost all public datasets [11, 12]. In addition, these approaches avoid the computation overhead of the transfer-based approach that requires image generation. Consequently, our experimental benchmark will focus only on pseudo-labeling approaches.

Even though all the aforementioned methods have shown promising results and great capability to adapt to a new target domain data set, their training process always assumes that they can have access to the entire target domain, which is difficult to hold in a real-world application as previously discussed in Sec. 1.

Lifelong Learning for person Re-Identification. Lifelong Learning, also called Continual Learning or Incremental Learning [16, 17, 28], is a field that aims at mitigating the catastrophic forgetting problem, which means that the model tends to forget previous knowledge acquired during previous tasks when learning new ones. Recently, many approaches have been developed to solve this problem for common vision tasks such as object detection [37], segmentation [1] or even image generation [33]. We can categorize existing methods into three main categories. First, teacher-student frameworks [21, 31], use a teacher module to remind the student network about the knowledge acquired in the past. The second category of methods relies on the regularization of the parameters update when new tasks arrive [20]. Finally, the third category is replay methods that consist in using stored images or an image generation model to feed old-task images along with the current task images into the learning network [30].

Recently, only a few works tackle the problem of lifelong learning in the case of Person Re-ID. [24] propose an Adaptive Knowledge Accumulation (AKA) framework, however, the training process is fully supervised and treats only the domain-incremental scenario. Zhipeng Huang et al. [15] address a scenario similar to ours except that storing images from the previous task is permitted. In our work, we consider that in real-world applications, person images might be subject to confidentiality purposes, and therefore storing images from previous tasks is not permitted.

3. Online Setting for UDA for Person Re-ID (OUDA-Rid)

3.1. Problem Definition

In this section, we describe the proposed online unsupervised domain adaptation setting for person re-identification (OUDA-Rid). We consider that we have access to an annotated source domain data set $D_S = \{(x_i^S, y_i^S) |_{i=1}^{N_S}\}$, where x_i^S and y_i^S denote the i^{th} training sample and its associated person identity label. We consider that we also have access

to a target domain data set D_T where ground truth identity labels are not available. However, differently from the standard UDA setting, we consider that the target data set is accessible as an online stream of data. More precisely, we adopt the **batch-based relaxation** [9] of the online learning scenario. The model will have access to the target domain D_T as a stream of T independent batches $T_t, t \in 1..T$. In analogy with the Continual Learning (CL) setting and to avoid confusion with the *mini-batch* used in Stochastic Gradient Descent (SGD), each target batch will be called task. Each task T_t is composed of N_t images $\{x_i^t, i = 1..N_t\}$ that depict an unknown number of identities. We assume that there is no identity overlap between tasks even if our approach does not strictly require it. This assumption corresponds to the practical scenario where data are collected over several hours or days. Even if the same person can appear again at different times, most detection will correspond to different identities.

Importantly, we consider that at the end of the task, the images of the task T_t cannot be used for the next tasks. This corresponds to a practical scenario where sensitive data can only be stored for a short period of privacy concerns (e.g. camera images from a public area). In addition to the source domain that is accessible at any time, only the parameters of the networks can be kept in memory in between two tasks. Finally, the goal is to deploy the trained model on an unknown target dataset that follows the same distribution as the training target tasks but does not share identities with the training tasks.

In this work, we adapt three frameworks for UDA to our OUDA setting. As detailed in Sec. 2, the UDA methods based on pseudo-labeling dominate most person re-identification benchmarks. Therefore, we focus our work on this paradigm. First, we employ a *Strong baseline* that is a very simple, yet effective, baseline. Then, we consider MMT [11] and SpCL [12], which are two methods that achieve state-of-the-art performance on publicly available datasets. Apart from their performance, what motivates the choice of these two frameworks is that, on the one hand, MMT has attracted a lot of attention lately and it is now considered a reference baseline for the task of UDA for person re-identification. On the other hand, SpCL is included in our benchmark since it illustrates the potential advantage of employing a memory to combine source and target data. Once adapted, the three frameworks will be evaluated and tested under four different configurations to: 1) decide which of the three frameworks is most suited to the OUDA problem 2) measure the drop in performance due to the online constraint 3) study the sensitivity of each model to its hyper-parameters.

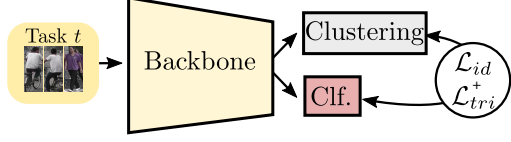


Figure 2. Scheme of *Strong Baseline*: training iterate between clustering and finetuning. The network is trained using a combination of cross-entropy and triplet losses.

3.2. Strong Baseline

The *Strong Baseline* [7] is a simple pseudo labeling pipeline. A feature extractor network F (backbone in Fig. 2) is pre-trained on the source labeled domain data set. After pretraining, the model is then fine-tuned on the target unlabeled data set. The finetuning on target consists of an iterative process where two major steps are alternated until convergence:

1. F is used to extract image features for every target domain image. Then, a standard clustering algorithm (DBSCAN [6] in our experiments) is applied to the extracted target domain features to obtain K clusters. In our case, K is automatically returned by the DBSCAN algorithm. In this way, we assign a **cluster label** to every image.
2. F is then finetuned on the target samples using their cluster labels as pseudo-labels. More precisely, a target domain classifier C with K classes is added to classify the images' features along with their assigned pseudo labels. The network is then trained via the minimization of a combination of an identity loss $\mathcal{L}_{id}^T(\theta)$ and a triplet loss $\mathcal{L}_{tri}^T(\theta)$. Assuming a sample x_i with pseudo-label y_i , the identity loss is given by:

$$\mathcal{L}_{id} = \mathcal{L}_{ce}(C(F(x_i)), y_i), \quad (1)$$

where \mathcal{L}_{ce} denotes the cross-entropy loss. Assuming the hardest positive and hardest negative features in the current mini-batch for the sample x_i , denoted f_i^+ and f_i^- respectively, the triplet can be written:

$$\mathcal{L}_{tri}^T(\theta) = \max[0, \|F(x_i) - f_i^+\| + m - \|F(x_i) - f_i^-\|] \quad (2)$$

where $\|\cdot\|$ denotes the \mathcal{L}^2 -norm and $m = 0.5$ denotes the triplet distance margin.

Adaptation to OUDA.

In our setting, this baseline approach is applied to each task. Instead of using the whole target dataset in the clustering step, we use only the data of the current task. The two steps are applied iteratively for several epochs.

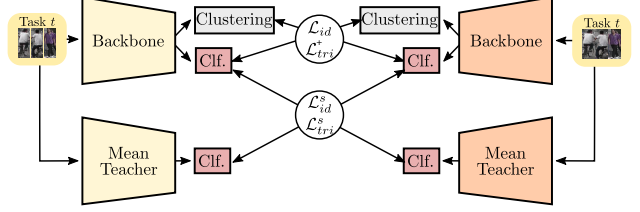


Figure 3. Scheme of *MMT*: two networks are trained thanks to two other momentum encoder networks. The two networks are trained using a combination of cross-entropy and triplet losses.

3.3. MMT

MMT is a recent framework proposed by [11], that integrates the teacher-student framework with two networks that train jointly. The main motivation is to design a framework that uses both hard and soft pseudo labels to learn better features. As shown in Fig.3, MMT extends the *Strong Baseline* in several ways. First, MMT employs two networks F_1 and F_2 instead of a single feature extractor F . To enforce that the networks help each other, the classifier C_1 for the feature extractor F_1 is trained to predict the clustering labels obtained from F_2 and vice-versa. Second, mean teacher networks M_1 and M_2 are introduced. These networks are obtained via estimating the running average on the network parameters of F_1 and F_2 . These networks predict more stable pseudo labels since they combine the knowledge of the networks at previous training iterations. In addition to the losses identity and triplet losses introduced in the *Strong baseline*, the two networks F_1 and F_2 are also optimized with respect to a soft classification loss and a soft triplet loss. Those losses are calculated for each network over the predictions of the other mean network. The losses between F_1 and M_2 are:

$$\mathcal{L}_{sid} = -M_2(x_i) \cdot \log C_1(F_1(x_i)) \quad (3)$$

$$\mathcal{L}_{stri} = -\mathcal{L}_{bce}(\tau_1^F(x_i), \tau_2^M(x_i)), \quad (4)$$

where \mathcal{L}_{bce} denotes the binary cross entropy and τ_1^F and τ_2^M are given by:

$$\tau_1^F(x_i) = \frac{e^{\|F_1(x_i) - F_1(x_i^-)\|}}{e^{\|F_1(x_i) - F_1(x_i^+)\|} + e^{\|F_1(x_i) - F_1(x_i^-)\|}} \quad (5)$$

$$\tau_2^M(x_i) = \frac{e^{\|M_2(x_i) - M_2(x_i^-)\|}}{e^{\|M_2(x_i) - M_2(x_i^+)\|} + e^{\|M_2(x_i) - M_2(x_i^-)\|}} \quad (6)$$

Note that to encourage the two networks to learn different image representations, different random data transformation policies are used for each network pairs (F_1, M_1) and (F_2, M_2) .

Adaptation to OUDA.

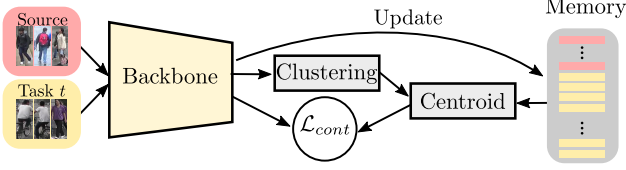


Figure 4. Scheme of *SpCL*: a feature memory is used to perform contrastive learning.

We adapt MMT to the OUDA setting in the following way: at the end of each task, the parameters of the four networks are kept and reused for the next task.

3.4. SpCL

Finally, we consider the *SpCL* method proposed in [12]. This framework (Fig. 4) employs a hybrid memory that stores and continually updates three types of feature vectors: the class centroids for every class of the source domain, cluster centroids for every cluster from the target domain, and the image feature corresponding to the target-domain samples that are not assigned to any cluster and that are considered as outliers. This memory provides supervision to the feature extractor via a contrastive loss over the three types of features in the memory.

Adaptation to OUDA.

We consider two adaptations of the *SpCL* framework. In the first version, referred to *SpCL-SF*, we adopt a source-free SF strategy [19, 26] where we do not use any of the source data when adapting to the target domain. This version is introduced because it allows for a fair comparison with MMT and the *Strong Baseline* that use the source dataset only for pretraining. In our second version (simply referred to as *SpCL*), we use the source dataset during the whole adaptation process in addition to the target data of the current task. In both cases, the memory is emptied, and clustering is performed at the beginning of each task.

4. Experiments

4.1. Datasets

We evaluate the different frameworks on three widely-used real-world person benchmark datasets in Domain Adaptive Person Re-ID:

- Market 1501 [34]: is a large-scale public dataset that contains 1501 identities that are captured by six different cameras. The total number of images is 32,668 for which 12,936 images of 751 identities are used for training and 19,732 images corresponding to the remaining 750 identities are used as a test set. We follow the official testing protocol stating that 3,368 query images should be tested and matched to 19,732 gallery images.

- DukeMTMC-reID [25]: The Duke Multi-Tracking Multi-Camera Re-Identification consists of images extracted from videos captured by 8 different cameras. It contains 16,522 training images corresponding to 702 identities, 2,228 query images of another 702 identities along 17,661 gallery images for testing.

- MSMT17 [29]: The third benchmark is the most challenging dataset since it has a greater diversity in terms of people’s appearances, viewpoints, and scales. It consists of multiple hours of videos captured by 15 different cameras. This dataset is a large-scale dataset consisting of 32,621 images of 1,042 identities as a training set, and 11,659 query images along with 82,161 gallery images corresponding to 3,060 identities as a test set.

4.2. Evaluation Protocol

To evaluate the performance of the different frameworks on our proposed setting, we consider four source-target configurations: Duke→Market, Market→Duke, Market→MSMT17, and Duke→MSMT17. These configurations are widely used in the literature and illustrate domain shifts of diverse difficulty. For each configuration, we randomly and uniformly split the training identities into 5 subsets corresponding to 5 tasks. For the evaluation metrics, we adopt the metrics commonly used in Re-ID [11, 12]: Mean Average Precision (mAP) and Rank-1 [35] accuracies. These metrics are computed on the entire test set of the target domain after each task during the online adaptation process. The proposed testing protocol is chosen to have a global overview of the model’s adaptation capability to the domain shift between source and target, and also to see which framework is the most suited for online adaptation.

In our preliminary experiments, we observed that the number of epochs per task is a key hyper-parameter. Even with a separate validation set, this hyper-parameter could not be chosen by mAP maximization since it requires identity labels and it would break the unsupervised adaptation assumption. On the contrary, using an inappropriate hyper-parameter value would jeopardize the validity of the conclusions of our experiments. Therefore, we used the following procedure: we run the *strong baseline* with four different numbers of epochs ranging from 10 to 40. Then, we observed that training for 20 epochs per task leads to the best performance. Therefore, we use 20 epochs per task for all the methods. Note that we report an ablation study in Sec. 4.6 to measure the sensitivity to this hyper-parameter and we validate that this choice remains satisfactory for the other methods.

	Offline		Direct inference		Online	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseline	75.6	90.9	29.6	62.4	49.4	77.1
MMT	80.9	92.9	29.6	62.4	63.7	87.5
SpCL-SF	76.7	90.3	29.6	62.4	42.9	70.2
SpCL	78.2	90.5	29.6	62.4	47.9	72.9

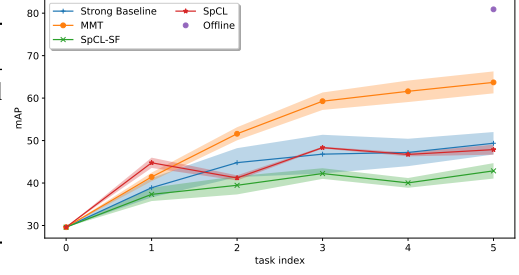


Figure 5. Experimental comparison of the performance of the four methods (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the 5-task OUDA Duke ->Market configuration. We report mAP and Rank-1 accuracy for each method.

	Offline		Direct inference		Online	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseling	60.4	75.9	28.2	50.1	26.8	59.3
MMT	67.7	80.3	28.2	50.1	51.7	72.3
SpCL-SF	68.8	82.9	28.2	50.1	38.7	61.8
SpCL	70.4	83.8	28.2	50.1	42.7	66.0

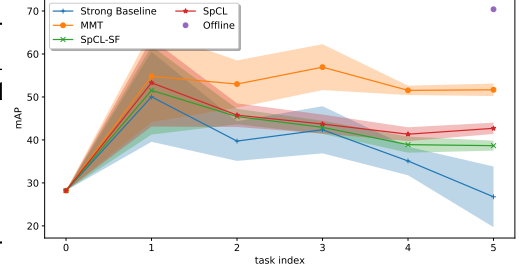


Figure 6. Experimental comparison of the performance of the four methods (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the 5-task OUDA Market ->Duke configuration. We report mAP and Rank-1 accuracy for each method.

	Offline		Direct inference		Online	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseling	9.7	25.8	8.9	28.9	6.1	18.0
MMT	22.9	49.2	8.9	28.9	15.1	31.5
SpCL-SF	26.3	53.4	8.9	28.9	13.1	36.5
SpCL	26.8	53.7	8.9	28.9	14.7	36.7

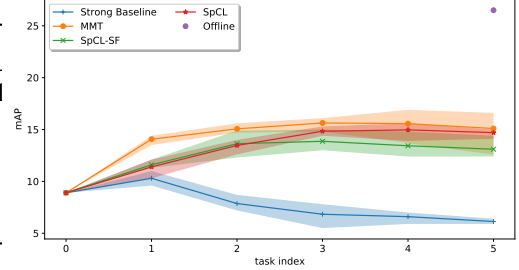


Figure 7. Experimental comparison of the performance of the four methods (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the 5-task OUDA Market ->MSMT configuration. We report mAP and Rank-1 accuracy for each method.

	Offline		Direct inference		Online	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
Strong Baseline	10.9	28.6	11.1	35.2	7.2	19.9
MMT	23.3	50.1	11.1	35.2	17.0	35.0
SpCL-SF	26.3	52.6	11.1	35.2	17.1	43.1
SpCL	26.5	53.1	11.1	35.2	17.8	40.8

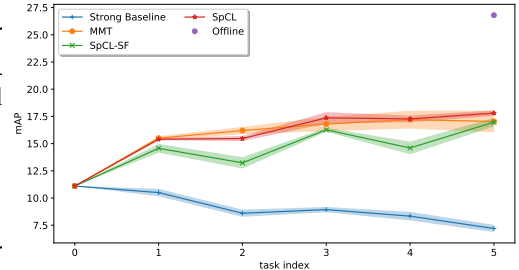


Figure 8. Experimental comparison of the performance of the four methods (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the 5-task OUDA Duke ->MSMT configuration. We report mAP and Rank-1 accuracy for each method.

4.3. Additional Baselines

To better assess the performance of the evaluated approaches, we consider two additional baselines that are not trained following our OUDA setting. First, we report the performance of the model pre-trained on the source and di-

rectly evaluated on the target. This baseline is common to all the frameworks since all the methods use the same pre-trained model and it is referred to as *Direct inference*. The second baseline is specific to each framework. It corresponds to the original method trained in the standard UDA setting and is referred to as *Offline*. It can be interpreted as

an upper bound for the online methods.

4.4. Implementation details

We follow the common practices in the UDA Person Re-ID field by adopting ResNet50 [14] pre-trained on ImageNet [4] as a backbone. For clustering, we use DBSCAN [6] which is frequently used in the pseudo-label based methods as it requires no prior on the number of clusters but only the maximum distance between two samples to consider one in the neighborhood of the other. We employ the maximum distance hyper-parameter set in the original papers of MMT and SpCL. Adam [18] optimizer is adopted with an initialized learning rate equal to $3.5 * 10^{-4}$ and a weight decay of 0.0005 [11, 12]. Finally, all the images are resized to 256 x 128 before being fed into the backbone (backbones for MMT), and the batch size was set to 64 corresponding to 16 different identities with 4 images per ID.

4.5. Results

We report in Figs. 5, 6, 7 and 8 the results of the *Strong Baseline* [7], MMT [11] and SpCL [12] on respectively four OUDA configurations: Duke→Market, Market→Duke, Market→MSMT17 and Duke→MSMT17. For every configuration, we report the final performances of each framework at the end of the adaptation process and plot the evolution of the test performance while the model is adapting to the target domain. Each experiment was repeated 3 times with different batch sampling initializations (*i.e.* seeds). The colored area corresponds to the variance of the performance on the test set at the end of each task, where the points correspond to the mean performances of the different initializations.

First of all, in the four configurations, the results show that the pre-trained ResNet50 on the source domain gives poor performances when directly deploying it into the target domain without any finetuning (*Direct inference*) compared to when it is fine-tuned on the target domain, either in an *Offline* or *Online* setting. This big gap in terms of performance illustrates the problem of domain shift.

Then, when it comes to the 5-tasks Online setting, the conclusions differ between methods and datasets. In the case of the Duke→Market configuration (Fig. 5), we observe that MMT (orange line) performs best among the online methods and reaches 63.7% of mAP. This result is very satisfactory since MMT bridges most of the gap between *Direct inference* and *Offline*. The *strong baseline* obtains lower performance since it plateaus after the second task. However, it surprisingly outperforms the two SpCL variants. Indeed, the performance is not improved significantly after completing the first task. We even observe a small drop when processing the second task for the SpCL variant that uses the source domain images. We also notice that the difference between the two variants of SpCL is mi-

nor illustrating that with a straightforward adaptation of the SpCL method, initially proposed for offline UDA, SpCL does not benefit much from the availability of the source data. On the right-hand side of the Fig. 5, we can see that the *Strong Baseline* shows the highest sensibility to random seeds. Moreover, MMT keeps reaching the best performance independently of the random seed.

Regarding the Market→Duke configuration (see Fig. 6), MMT is again the best performing method even though its gap with respect to the best offline method (purple dot) is larger. This behavior change can be explained by the highest difficulty of this setting as illustrated by the lower score obtained by the offline methods (*e.g.* 70.4% of the map in Market→Duke vs 80.9% of the map in Duke→Market). In this more difficult configuration, the *strong baseline* does not perform well since it achieves the worst performance among all the evaluated methods. The behavior of SpCL is very instructive. At the beginning of training (until the second task), the source-free model performs better but shows degraded performance later in training. This behavior can be explained by a probable divergence of the model that forgets its initial source model and overfits on the target task. On the contrary, the SpCL variant that uses the source data needs more time to handle the domain shift but keeps slowly increasing. Concerning the variance of the performance, we can see that the four frameworks are sensitive to their random seed, especially at the beginning of the adaptation process. However, this variance decreases after a few tasks, showing that the training becomes more stable (after two tasks for most methods) except for the *Strong baseline*, where the variance of the performances becomes even higher on late tasks.

In the Market→MSMT configuration (Fig 7), conclusions drastically change since SpCL has almost the same results as MMT, hence, breaking the big gap between the two methods in performance we observed in previous configurations. This change can be explained by the large training target dataset MSMT. Therefore, every target task contains more images and more identities. This difference is beneficial to both SpCL variants that perform similarly. Regarding, MMT, we see that the performance starts degrading from task 3. Again, it can be explained by the fact that in the case of a large target dataset, MMT can forget the knowledge from the source domain that is not further used during adaptation. Interestingly, the best performance of MMT (end of task 3) is higher than the best performance of SpCL. It illustrates the importance of handling the divergence problem and designing efficient consolidation mechanisms. Finally, we observe that the strong baseline worsens the performance compared to the initial pre-trained model. Regarding the variance of the performances, we can see that MMT, SpCL, and SpCL-SF finally get more or less similar results at the end of the adaptation process. Finally, in the

Duke→MSMT configuration (Fig. 8), the conclusions remain similar to the previous setting. Nevertheless, we can mention that SpCL outperforms MMT in this specific configuration, and observe higher instability on the SpCL-SF method that oscillates in the last tasks.

4.6. Analyses

Model sensitivity: number of training iterations. In this section, we study the effect of the number of epochs on the performance of the four frameworks (*Strong baseline*, MMT, SpCL, and SpCL-SF) in the following configuration: 5-task OUDA Duke→Market. In Fig. 9 we report the performance on the target test set (mAP) of the three frameworks while varying the number of training epochs between 0 and 40 epochs per task. Note that zero epoch corresponds to the *Direct inference* performance of the pretrained model without any training on the target domain. It can be observed that with 20 epochs the *strong baseline* achieves the best performance on the test set. When we increase the number of the training epochs, we see a decrease in the performance on the test set of the three frameworks probably illustrating overfitting issues in the current training task. SpCL, thanks to its memory-based system, that provides supervision from the labeled source domain images to the Re-ID model, needs fewer training epochs per task to converge, compared to the *strong baseline* and MMT. We see that the four aforementioned frameworks are sensitive to the number of training epochs to some extent. These experiments illustrate the difficulty of the OUDA setting where only a few samples are available in each task and where overfitting can appear rapidly.

Impact of the number of tasks. We also conducted further experiments to show the effects of varying the number of tasks on the adaptation performances. In Fig. 10 we report the final performance (mAP) on the target test set of the four methods when considering 1, 3, 5, 8, and 10 tasks. Naturally, when augmenting the number of tasks during the adaptation process, the number of images per task decreases. This affects the fine-tuning of the model, where we can see in Fig. 10 that for all the considered frameworks, the performance drops when considering more challenging online settings by adding more tasks.

We also performed experiments with a number of tasks larger than 10 (typically 15 or 20), however, training did not succeed due to the sampling limitation. More precisely, when the number of tasks increases the number of samples becomes too small to be handled by DBSCAN. In such challenging configurations, only a few clusters are considered, where only a few images per cluster are sampled, hence the sampling of the 16 identities with 4 images per id, which is necessary for the optimization of the triplet loss, becomes impossible. This clustering issue shows the limitation of UDA methods to address our OUDA setting and demon-

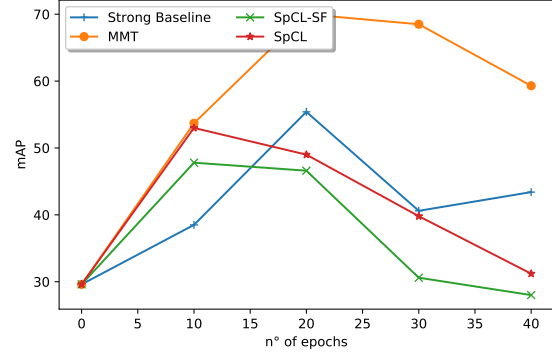


Figure 9. Effect of the number of training epochs per task on the Re-ID performance. At zero, we reported the results from the *direct inference* model.

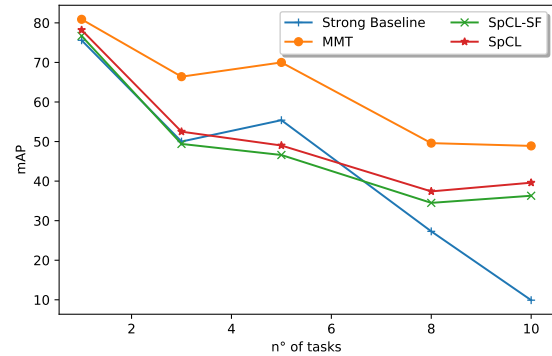


Figure 10. Effect of the number of tasks on the performances of the four frameworks at the end of the adaptation process. We varied the number of epochs from 1 to 10. Note that 1 epoch corresponds to the *offline* setting.

strates the need for new methods tailored for OUDA.

5. Conclusions

In this work, we introduced the Online Domain Adaptation Re-ID problem and presented an empirical benchmark where we adapt and evaluate three state-of-the-art methods previously introduced for the *Offline* UDA setting. Our experiments show that existing methods can achieve satisfactory results in simple online adaptation scenarios but fails to reach the performance achieved in the *Offline* setting. We also show that the best-performing methods depend on the setting. Finally, our experiments highlight the forgetting problem when the source model is not used during adaptation. These conclusions pave the way toward novel approaches for online domain adaptive Re-ID and we hope this work will stimulate the community to address this setting that matches real-world constraints and better protect privacy.

References

- [1] Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. Incremental learning in semantic segmentation from image labels. *arXiv preprint arXiv:2112.01882*, 2021. 3
- [2] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Instance-guided context rendering for cross-domain person re-identification. *ICCV*, 2019. 2
- [3] Guillaume Delorme, Yihong Xu, Stéphane Lathuilière, Radu Horaud, and Xavier Alameda-Pineda. Canu-reid: a conditional adversarial network for unsupervised person re-identification. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4428–4435. IEEE, 2021. 2
- [4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009. 7
- [5] Weijian Deng, Liang Zheng, Guoliang Kang, Yi Yang, Qixiang Ye, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. *CVPR*, 2018. 1, 2
- [6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 1996. 4, 7
- [7] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM TOMM*, 2018. 1, 2, 4, 7
- [8] Hao Feng, Minghao Chen, Jinming Hu, Dong Shen, Haifeng Liu, and Deng Cai. Complementary pseudo labels for unsupervised domain adaptation on person re-identification. *IEEE TIP*, 2021. 1, 2
- [9] Enrico Fini, Stéphane Lathuilière, Enver Sangineto, Moin Nabi, and Elisa Ricci. Online continual learning under extreme memory constraints. *ECCV*, 2020. 3
- [10] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. *ICCV*, 2019. 1, 2
- [11] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *ICLR*, 2020. 1, 2, 3, 4, 5, 7
- [12] Yixiao Ge, Dapeng Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *NeurIPS*, 2020. 1, 2, 3, 5, 7
- [13] Yixiao Ge, Feng Zhu, Rui Zhao, and Hongsheng Li. Structured domain adaptation with online relation regularization for unsupervised person re-id. *arXiv preprint arXiv:2003.06650*, 2020. 2
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 7
- [15] Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, and Zhengjun Zha. Lifelong unsupervised domain adaptive person re-identification with coordinated anti-forgetting and adaptation. *arXiv preprint arXiv:2112.06632*, 2021. 3
- [16] Steven C. Y. Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compact-ing, picking and growing for unforgetting continual learning. *NeurIPS*, 2019. 3
- [17] Prakhhar Kaushik, Alex Gain, Adam Kortylewski, and Alan L. Yuille. Understanding catastrophic forgetting and remembering in continual learning with optimal relevance mapping. *arXiv preprint arXiv:2102.11343*, 2021. 3
- [18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 7
- [19] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *IEEE/CVF CVPR*, 2020. 5
- [20] Jeongtae Lee, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017. 3
- [21] Zhizhong Li and Derek Hoiem. Learning without forgetting. *ECCV*, 2016. 3
- [22] Yutian Lin, Xuanyi Dong, Liang Zheng, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. *AAAI*, 2019. 1, 2
- [23] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. *CVPR*, 2019. 2
- [24] Nan Pu, Wei Chen, Yu Liu, Erwin M. Bakker, and Michael S. Lew. Lifelong person re-identification via adaptive knowledge accumulation. *CVPR*, 2021. 3
- [25] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *ECCV*, 2016. 1, 2, 5
- [26] Cristiano Saltori, Stéphane Lathuilière, Nicu Sebe, Elisa Ricci, and Fabio Galasso. Sf-uda 3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection. In *2020 IEEE 3DV*, 2020. 5
- [27] Liangchen Song, Cheng Wang, Lefei Zhang, Bo Du, Qian Zhang, Chang Huang, and Xinggang Wang. Unsupervised domain adaptive re-identification: Theory and practice. *Pattern Recognition*, 2020. 1, 2
- [28] Cheng-Hao Tu, Cheng-En Wu, and Chu-Song Chen. Extending conditional convolution structures for enhancing multi-tasking continual learning. *APSIPA ASC*, 2020. 3
- [29] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. *CVPR*, 2018. 1, 2, 5
- [30] Chenshen Wu, Luis Herranz, Xialei Liu, Yaxing Wang, Joost van de Weijer, and Bogdan C. Raducanu. Memory replay gans: learning to generate images from new categories without forgetting. *NeurIPS*, 2018. 3
- [31] Fei Ye and Adrian G. Bors. Lifelong teacher-student network learning. *IEEE TPAMI*, 2021. 3
- [32] Mang Ye, Jiawei Li, Andy J. Ma, Liang Zheng, and Pong C. Yuen. Dynamic graph co-matching for unsupervised video-based person re-identification. *IEEE TIP*, 2019. 1, 2
- [33] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong GAN: continual learning for conditional image generation. *ICCV*, 2019. 3

- [34] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. (*ICCV*), 2015. 1, 2, 5
- [35] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. *ICCV*, 2015. 5
- [36] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero- and homogeneously. *ECCV*, 2018. 1, 2
- [37] Wang Zhou, Shiyu Chang, Norma E. Sosa, Hendrik F. Hamann, and David D. Cox. Lifelong object detection. *arXiv preprint arXiv:2009.01129*, 2020. 3
- [38] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV*, 2017. 2