Unsupervised Domain Adaptation for Person Re-Identification with Few and Unlabeled Target Data

George Galanakis[1,2](✉), Xenophon Zabulis[2], and Antonis A. Argyros[1,2]

[1] Computer Science Department, University of Crete, Rethymno, Greece
{ggalan,argyros}@csd.uoc.gr
[2] Institute of Computer Science, FORTH, Heraklion, Greece
{ggalan,zabulis,argyros}@ics.forth.gr

Abstract. Existing, fully supervised methods for person re-identification (ReID) require annotated data acquired in the target domain in which the method is expected to operate. This includes the IDs as well as images of persons in that domain. This is an obstacle in the deployment of ReID methods in novel settings. For solving this problem, semi-supervised or even unsupervised ReID methods have been proposed. Still, due to their assumptions and operational requirements, such methods are not easily deployable and/or prove less performant to novel domains/settings, especially those related to small person galleries. In this paper, we propose a novel approach for person ReID that alleviates these problems. This is achieved by proposing a completely unsupervised method for fine tuning the ReID performance of models learned in prior, auxiliary domains, to new, completely different ones. The proposed model adaptation is achieved based on only few and unlabeled target persons' data. Extensive experiments investigate several aspects of the proposed method in an ablative study. Moreover, we show that the proposed method is able to improve considerably the performance of state-of-the-art ReID methods in state-of-the-art datasets.

Keywords: Person re-identification · Unsupervised domain adaptation · Agglomerative clustering

1 Introduction

During the recent years, person re-identification (ReID) has received a lot of attention in the computer vision research community [30]. This is especially due to the increased interest in surveillance applications related to security, crime prevention and crowd analytics. The goal of person ReID, is to match people across non-overlapping camera views at different times. In an effort towards more accurate person identification, modern solutions propose learning discriminative, appearance-based features with increased robustness against illumination and pose variations, but also tolerant to missing information, such as occlusions.

G. Bebis et al. (Eds.): ISVC 2020, LNCS 12510, pp. 357–373, 2020.
https://doi.org/10.1007/978-3-030-64559-5_28

Table 1. Comparison of ReID methods according to their type and level of supervision. Rows: requirements of the methods due to their supervision type and the resulting suitability in real world settings (see text for details).

Supervision level → Properties ↓	Fully-sup.	Unsup.	Semi-sup.	UFT-reID
Requires auxiliary non-target dataset	-	optional	optional	✓
Uses target training dataset	✓	✓	✓	-
Requires ID annotations in target dataset	✓	-	partial	-
Uses views from all cameras in target dataset	✓	✓	✓	optional
Easy to deploy in new settings	×	✓	×	✓
Suitable for small galleries	×	×	×	✓

These features are learned in a supervised, semi-supervised or unsupervised manner, based on several public surveillance data.

Table 1 summarizes the requirements and properties of the person ReID methods with respect to their type of supervision. *Supervised learning* is the most prominent methodology, as it incorporates a lot of information from the target domain in which these methods will need to operate. Such information includes images depicting persons from multiple cameras, paired with the corresponding person ids. The outcome of the learning process is a model capable of extracting features to represent persons. The hope is that the training set contains enough variability, therefore it is expressive and generalizes well. The model is afterwards evaluated on images of unseen persons, while the cameras and other conditions (e.g., illumination conditions) remain the same. As it is often demonstrated [13], supervised models do not generalize well to new domains. This means that even if two datasets are obtained in visually similar conditions training in one of them (source) and directly using the model on the other (target) dataset, results in very significant ReID performance degradation.

To overcome the expense of labeling requirements in the target domain, *semi-supervised* transfer learning techniques have been developed. These techniques incorporate labeled data from an auxiliary (source, non-target) domain and partially labeled data from the new (target) domain. Other techniques, also referred to as *unsupervised* domain adaptation, require no labelled data from the new domain. Moreover, some recent unsupervised techniques such as the one proposed in [18] depend only on unlabeled auxiliary data from the target domain. The common ground of these techniques is their dependence on the availability of a substantially large set of auxiliary data from the new, target domain.

In real world applications, enough amounts of data (even unlabeled) from the target domain may be few, hard to obtain, or even unavailable. This holds especially for places where, within a specific time-frame, passers-by are in the dozens rather than in the hundreds. This situation makes most of the existing ReID methods inadequate, due to limited input for learning. For this reason, we argue that effective, real-world solutions must require no same-domain auxiliary data.

In that direction, we propose UFT-ReID, a novel ReID method that is unsupervised with respect to the target domain and operates directly on unlabeled, test target data, without requiring a training target dataset. Moreover, contrary to existing approaches that use training input from all available cameras of the target domain, the proposed ReID method is demonstrated in situations were part of the viewpoints are not available. Due to its loose supervision requirements, UFT-ReID can be applicable even in small person galleries. Indeed, extensive experiments with UFT-ReID prove that the proposed approach is very effective even in such constrained settings.

2 Related Work

Modern person ReID approaches learn robust person representations by incorporating appropriate *CNN architectures* and objective functions that result from different *training and supervision objectives*. Below, we discuss these two components separately as usually, they are independent to each other.

2.1 CNN Architectures

Initial ReID approaches, including [1,11,16] borrowed or got inspired by CNN models that were designed specifically for object classification. This trend is followed by current works [10,15,18], too. Recent efforts in CNN architecture design incorporate methods which take into account that person images are constrained, i.e., they only contain standing persons rather than generic object classes. Towards this direction, some methods propose rough segmentation of a person's body into parts [26,27,33,39]. Other methods directly integrate pose estimation [23,25,32] or dense part correspondence [40] for extracting part-aware features. In contrast to competing works, [21,38,45] do not adopt an ordinary backbone network, but propose their own. Finally, some recent works proposed modified architectures that better generalize to new domains [14,44].

2.2 Training Objectives

CNN models are trained by minimizing some properly designed loss objective function. In order to express and quantify the loss, this function may (or may not) utilize the ground truth person IDs. In this light, approaches may be classified to supervised, semi-supervised or unsupervised.

Supervised Methods: The majority of proposed works present supervised methods, while most common losses are classification and metric loss. Classification loss is inspired by object classification and is usually implemented as cross-entropy loss. In this case, each person is regarded as a different object class. Approaches which adopt the classification loss include [9,17,23,45]. More interestingly [34] measured separate classification loss for each body part. Metric loss is inspired by the distance metric learning framework. Its objective is

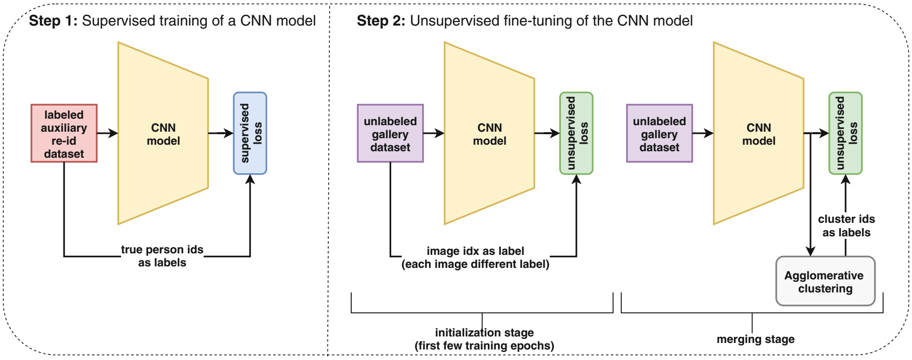

Step 1: Supervised training of a CNN model
labeled auxiliary re-id dataset
CNN model
supervised loss
true person ids as labels
Step 2: Unsupervised fine-tuning of the CNN model
unlabeled gallery dataset
CNN model
unsupervised loss
image idx as label (each image different label)
unlabeled gallery dataset
CNN model
unsupervised loss
cluster ids as labels
Agglomerative clustering
initialization stage (first few training epochs)
merging stage

Fig. 1. Overview of the two-step approach followed by the proposed UFT-ReID method.

twofold; to simultaneously minimize intra-class and maximize inter-class distance. Variances of this loss, also referred to as triplet loss, are adopted by the methods in [4,11,22,24,37]. Some works proposed novel loss functions [8,38]. Finally, combinations of losses are incorporated in [2,21,38,40].

Unsupervised Methods: Methods that do not use any supervision with respect to target domain fall into two subcategories. First, domain adaptation methods, aim at adapting well-performing models that are trained on auxiliary data, to a different target domain. Recent methods which follow this approach are [6,7,33,35,36]. Completely unsupervised methods, including [5,18] do not include supervision at any stage of training. Some of these works propose generating pseudo-labels [7,18,35] or pseudo-positive samples [33] that can be utilized as previously within classification or metric losses, respectively. The notion of such pseudo labels has been effectively explored in other visual tasks such as [3,20], too. Other methods propose more appropriate unsupervised objectives [28,35].

Semi-supervised Methods: Semi-supervised approaches require that only a part of target data is labeled. These works borrow and combine methodologies from both supervised and unsupervised settings. Some recent works in this domain are [19,29].

Our Approach and Contribution: The proposed UFT-ReID method corresponds to domain adaptation techniques, while it is achieved by unsupervised fine-tuning on the target domain. In contrast to other works, UFT-ReID operates directly on target persons' data. This is a novel formulation that is more relevant to real-world situations, where the requirement for same-domain auxiliary data is hard or even impossible-to-fulfill. We demonstrate that this setting is feasible and that our approach is very effective, especially for small person galleries. Our work adopts the pseudo-label generation approach and corresponding loss function from [18], however it deviates from them as follows. Firstly and most importantly, our method and corresponding evaluation scheme do not require same-domain auxiliary person data. Such data are extensively utilized by [18] to facilitate learning in a completely unsupervised way. Instead, we pre-train on auxiliary domain dataset (domain adaptation). We argue that for real deployment one should always perform supervised pre-training on existing re-id datasets, since this results to already capable baselines [45]. Finally, our architectures are specially designed for the ReID task, in contrast to the general ResNet utilized in [18].

3 The UFT-ReID Method

The workflow of UFT-ReID (Fig. 1) consists of two steps: (1) supervised model training and (2) unsupervised model fine-tuning. In both cases, model refers to the same deep neural network. Supervised training is accomplished using an auxiliary domain data, while fine-tuning is performed directly on the target domain and especially on target persons; those of ReID interest.

3.1 Supervised Model Training

Given a labeled person dataset from a particular domain, our goal is to learn a mapping from the original image space, to a feature space in which, images of the same person are close, while images of different persons are more distant. This mapping can be effectively structured by typical supervised learning approaches, using conventional or problem-specific CNN architectures, loss functions and training methodologies. The outcome of learning is a feature extraction model, robust for unseen persons in this particular domain, though limited for other domains. For this step we utilized and experimented with two modern architectures as described in Sect. 4.2.

3.2 Unsupervised Model Fine-Tuning

Unsupervised training techniques are naturally more suitable for real-world applications where we are not aware of the identity of each person in a gallery and live manual annotation is highly undesirable, as it is costly or even impossible. Instead of training from scratch, we propose to refine the model of the previous (supervised model training) step, solely in an unsupervised fashion. This model has been trained on a separate domain; different cameras, lighting conditions etc, but on the same task (person ReID). It is expected that a tuning to the new domain should be sufficient to adapt the model to the new persons' appearances, without mitigating its generalizability.

Model fine-tuning is a well known technique for transfer learning. Given a pre-trained CNN, the common practice is (a) to modify its output layer to contain the new classes and (b) train the rest of the model according to some initial, relatively small learning rate. Note that, depending on the model architecture, it may be useful to define different learning rates for groups of layers. This is explained by the structure of CNNs, where first convolutional layers represent primitive information, which is a common base across different datasets and even tasks.

Traditional fine-tuning depends on availability of class labels (supervision). As previously posed, supervised methods require resource-intensive identity annotation, therefore they are not easily applicable to new domains. In this work we are interested in the unsupervised setting which is more suitable for our task. Lack of person ids guided us to seek for an unsupervised solution with respect to target domain. In this light, we are inspired by a recent unsupervised framework [18], also employed for person ReID, however for a different setting. More specifically, our approach relies on the use of pseudo-labels. A common method for proposing pseudo-labels during learning is clustering. This approach is followed for example in [7,18,35]. However, in [18] clustering is naturally integrated with training. A bottom-up clustering approach is proposed, for jointly optimizing the CNN and the relationship between individual samples (images).

In UFT-ReID we utilize this framework as follows. At first, we initialize our CNN with weights pre-trained from the previous, supervised step. We also prepare the model for the unsupervised task, by keeping only relevant layers and assign unique labels to each training sample. This is because initially, all training samples are considered as independent clusters. At each training stage, cluster numbers are utilized as pseudo-labels for optimizing the CNN. However, the number of clusters is not constant, but gradually lowers as clusters merge. In more detail, the training scheme operates as follows:

1. **Initialization stage:** All training samples are regarded as unique clusters. The CNN is trained for e_i epochs, with respect to minimization of the repelled loss.
2. **Merging stage:** The current state of the CNN is utilized for features extraction. Afterwards, according to a merging criterion, m clusters are merged and CNN is trained for another e_m epochs.
3. **Stopping criterion:** Training stops when the number of clusters due to several merges reaches m.

The repelled loss is defined as the negative log probability (cross-entropy) that a sample belongs to the correct cluster. For a single sample x_i it corresponds to:

$$\mathcal{L} = -log(p(\hat{y_i}|x_i, \mathbf{V})), \tag{1}$$

where p is defined as:

$$p(c|x, \mathbf{V}) = \frac{exp(\boldsymbol{V}_c^{\mathrm{T}} \boldsymbol{v}/\tau)}{\sum_{j=1}^{C} exp(\boldsymbol{V}_j^{\mathrm{T}} \boldsymbol{v}/\tau)}. \tag{2}$$

In the above equations, x corresponds to the input samples within the batch, $\boldsymbol{v}$ are the $L2$-normalized features of these samples as extracted by the CNN at the current state, C is the current number of clusters, $\mathbf{V}$ is a lookup table which maintains the features of the centroid of each cluster, and τ is a temperature parameter which controls the softness of the probability distribution [12]. The contribution of repelled loss is that it computes probabilities based on feature

Table 2. Choices of learning rate (lr) parameter for step 2 (fine-tuning). For each case, lr is 10x smaller for all layers up to and including top base layer.

Model	**#IDs**	lr	**Top base layer**	**Model**	**#IDs**	lr	**Top base layer**
PCB	10-33	1e-3	ResNet-50.layer3	OSNet	10-83	1.5e-4	OSNet.conv4
	45-282	1e-4			113-282	1.5e-5	
	383-702	1e-5			383 - 702	1.5e-6	

similarity and simultaneously trades off intra-cluster similarity and inter-cluster diversity, over the whole training set.

Cluster merging is based on the minimum distance criterion. This criterion takes the shortest Euclidean distance between samples in two clusters as the dissimilarity measure. In order to ensure that all clusters contain approximately the same number of samples, a diversity regularization term is introduced. This boosts merging smaller clusters. The overall dissimilarity merging score is computed as a sum of the minimum distance criterion and the regularization term, whose impact is controlled by a parameter λ. For more details about bottom-up clustering, we refer the reader to [18].

In UFT-ReID we utilize a pre-trained CNN on an auxiliary source domain and we fine-tune it using the above framework. Fine-tuning is performed in fewer data, containing only target persons. With respect to the CNN itself, we choose and experiment with modern architectures (PCB [26] and OSNet [45], see Sect. 4.2) specifically designed for the ReID task. These architectures take into account that the image belongs to a person, therefore they are able to exploit contextual information. Furthermore, the experimentation with different architectures essentially reveals that our approach is capable of applying successfully transfer learning via unsupervised fine-tuning that is irrelevant to the backbone architecture.

4 Experiments and Discussion

Our experiments where conducted on a PC equipped with Intel Core i7 CPU, 16GB RAM and an NVIDIA GTX1080 GPU. We re-implemented [18] as extension to Torchreid framework [43], based on the original reference code and appropriate modifications for supporting additional features related to our training and evaluation strategies.

4.1 Datasets

To evaluate UFT-ReID we employ two recent datasets, Market1501 [41] and DukeMTMC-reid [42] which comprise of multiple persons and multiple views per person. Both datasets are utilized either as auxiliary/source or as target, in different experiments. Let $M \rightarrow D$ denote fine-tuning on DukeMTMC-reid,

Table 3. $M \rightarrow D$ fine-tuning results using the proposed UFT-ReID method for different target gallery sizes.

Architecture		10	13	18	24	33	45	61	83	113	153	208	282	382	518	702
PCB	Baseline	70.9	66.3	61.9	59.8	56.0	52.7	49.8	46.5	43.8	41.1	38.3	35.6	32.7	30.2	27.6
	With UFT-ReID	78.6	74.3	73.5	70.8	68.1	64.7	62.9	60.5	57.6	55.2	51.7	47.5	40.5	38.6	35.6
	% Benefit	7.6	8.0	11.6	11.0	12.1	12.0	13.1	14.0	13.8	14.1	13.4	11.9	7.9	8.4	8.0
OSNet	Baseline	77.8	73.6	71.9	68.5	64.5	61.2	58.1	55.4	52.8	50.4	47.6	45.0	42.0	39.1	36.0
	With UFT-ReID	75.1	75.3	76.0	79.4	78.5	76.2	72.3	66.5	64.9	62.8	60.3	56.4	48.9	46.9	43.9
	% Benefit	−2.8	1.7	4.0	11.0	14.0	15.1	14.1	11.1	12.1	12.4	12.7	11.3	7.0	7.9	7.9

Table 4. Left: $D \rightarrow M$ fine-tuning results using the proposed UFT-ReID method. Fewer target gallery sizes have been tested because of the size of the target dataset (Market1501). Right: $M \rightarrow D_r$ UFT-ReID results.

Architecture		20	30	50
	Baseline	79.2	76.4	71.7
PCB	With UFT-ReID	83.8	82.3	75.3
	% Benefit	4.6	5.9	3.7
	Baseline	81.5	78.7	75.3
OSNet	With UFT-ReID	85.9	83.1	79.8
	% Benefit	4.4	4.4	4.5

Architecture		33	45	61
	Baseline	63.2	58.9	55.1
PCB	With UFT-ReID	65.4	64.4	57.1
	% Benefit	2.1	5.5	2.0
	Baseline	71.8	67.8	66.0
OSNet	With UFT-ReID	77.0	74.9	71.6
	% Benefit	5.2	7.1	5.6

based on a model which is pre-trained on Market1501, and $D \rightarrow M$ the opposite. We point out that $M \rightarrow D$ is a more difficult task because Market1501 contains fewer images observed from less viewpoints, i.e., it is less general. As a consequence, the pre-trained model is less expressive.

In our work we experiment with person galleries of varying size. In order to simulate such galleries, we generate random subsets for each target dataset. Let P be the number of persons contained in the test part of the target dataset and k the number of persons in the subset. In this work we focus on galleries containing $k \geq 10$ persons. In order to simulate diverse scenarios, we choose 15 values of k spaced evenly on a logarithmic scale in the range $[10, P]$. We randomly select k person ids and we repeat 20 times. In total, 300 random galleries of varying ids and sizes are generated. We do such randomization once and prior to all our experiments and store the galleries for further utilization. We further refer to a gallery sized x as G_x, e.g. G_{30} denotes gallery containing 30 persons.

It should be noted that in both Market1501 and DukeMTMC-reid datasets, the original gallery and query subsets contain person images obtained from the same cameras. However, this comes in contrast to a more realistic, cross-camera evaluation setting. Some person ReID methods [31,43] address this issue, but only during evaluation, where images from same cameras are discarded during pairwise matching. In contrast, in a separate experiment we purposely utilized images from only two cameras from DukeMTMC-reid's gallery during fine-tuning, while leaving the rest six cameras for query. We refer to this reduced

Table 5. Left: comparison with [33], standard eval. protocol. Right: comparison with [33], our eval. protocol and gallery type G_{33}.

Method	**% rank-1**
PatchnetUn	65.7
PatchnetUn w. Pedal + Ipfl ([33])	72.0
OSNet	68.1
OSNet w. repelled (UFT-ReID)	69.6

Method	**% rank-1**
PatchnetUn	88.33
PatchnetUn w. Pedal + Ipfl ([33])	**89.95**
OSNet	89.08
OSNet w. repelled (UFT-ReID)	89.50
OSNet w. repelled + Pedal + Ipfl (UFT-ReID)	**90.47**

dataset as D_r. This setting is more challenging, as refinement is based on images captured by part of the cameras which are different to those used to capture the test images.

4.2 Settings

CNN Architectures: We utilize two recently proposed person ReID architectures named PCB [26] and OSNet [45]. PCB mainly aims at learning discriminative part-informed features, without the need for exact part/pose estimation. It is based on ResNet-50 architecture and augmented with additional layers. These layers are parallel for each body subdivision, while the final descriptor is the concatenation of the feature vectors from the separate layers. OSNet is a novel architecture, capable for multi-scale feature learning at each level of the architecture. The core of this architecture is an omni-scale residual block which allows the propagation of smaller scale features to higher layers. For both architectures we used the implementations provided by [43]. We trained models on source datasets using the default training parameters (Fig. 1, step 1).

Fine-Tuning Parameters: In the second step, we utilize the trained CNN model, apart from its output layer (classifier). Batch size is chosen to be small (16). Experiments with larger values are provided in Sect. 4.4. Learning rate is also chosen to be small so that the original model does not alter much. During preliminary experiments we found out that it is more beneficial to lower the learning rate as the number of persons grows. Thus, we choose variable learning rates, depending on the number of persons in the gallery. Furthermore, we choose smaller learning rates for the base layers of the network. Detailed learning rate settings are given in Table 2. The number of training epochs is not predefined, but dependent on the total number of samples and clustering algorithm parameters, as explained below.

Unsupervised Algorithm Parameters: As explained in Sect. 3.2, during the first e_i fine-tuning epochs, the number of clusters is equal to the number of gallery samples. Afterwards, and every e_m epochs, the number of clusters is reduced by m, due to merges. This is an iterative process which stops when the total number of clusters reaches m. In our experiments we set $e_i = 20$

Table 6. Comparison with OSNet variants [44,45] using our eval. protocol. Gallery type: G_{33}.

Architecture	Baseline	+UFT-reid	% Benefit
OSNet (original)	64.5	78.5	14.0
OSNet-IBN [45]	74.5	77.6	3.1
OSNet-AIN [44]	76.5	**80.0**	3.5

Table 7. Average fine-tuning training duration in minutes (OSNet architecture).

Auxiliary source → Target	33	45	61
$M \rightarrow D$	07:46	10:55	15:03
$M \rightarrow D_r$	06:01	08.16	11:17

and $e_m = 2$. Nevertheless, we realized that by increasing e_m we obtain a clear performance gain, at the cost of more time-consuming training. m is set to the number of gallery ids. Finally, we experimentally confirmed that the optimal value for the diversity regularization parameter λ is 0.05, as suggested in [18]. Section 4.4 presents additional experiments using various options for e_m and λ.

4.3 Experimental Results

Tables 3, 4 demonstrate the effectiveness of our method in various settings. In all experiments we report the rank-1 accuracy, obtained by the CMC curve. More specifically, Table 3 shows the impact of UFT-ReID fine-tuning for a wide range of gallery persons. The particular experiment is conducted using a CNN model pre-trained on Market1501, while galleries are sampled from DukeMTMC-reid ($M \rightarrow D$). It is shown that fine-tuning using UFT-ReID achieves considerably better rank-1 accuracy, increasing the baseline up to 14% and 15.1% for PCB and OSNet architectures, respectively. For both architectures, higher accuracy is obtained in the case of mid-sized galleries. This may indicate that a mid-sized gallery is a good balance for appearance diversity. Too small diversity encountered in smaller galleries is not enough for generalized learning. On the other hand, too large diversity, encountered in larger galleries may encompass persons that are similar to each other, negatively affecting the overall accuracy.

In Table 4 (left) we present similar results for fine-tuning models pre-trained on DukeMTMC-reid using some random Market1501 galleries ($D \rightarrow M$). In this case, the benefit is smaller on average. This is expected, because base models are trained on a larger dataset, (DukeMTMC-reid), therefore better equipped against a smaller dataset.

Finally, Table 4 (right) shows experimental results using the reduced DukeMTMC-reid gallery dataset which, as explained in Sect. 4.1, contains images from cameras that were not used for the unsupervised fine tuning step ($M \rightarrow$

D_r). Even in this case, UFT-ReID increases the accuracy of the baseline methods, although as expected, the benefit is on average smaller compared to the $M \rightarrow D$ experiment.

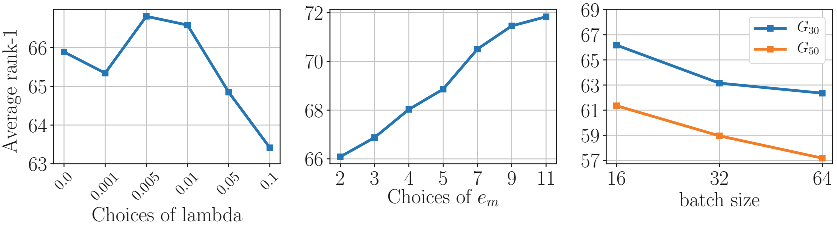
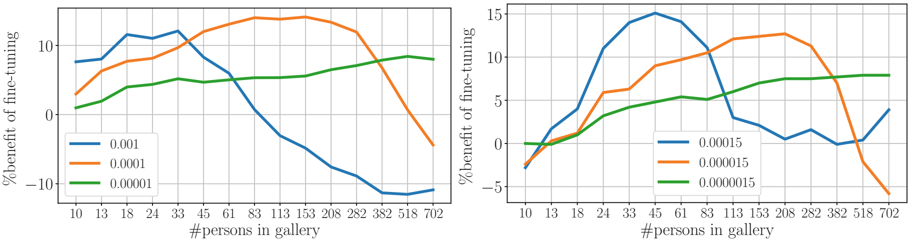


Fig. 2. Left: effect of diversity regularization parameter λ on average rank-1 accuracy. Middle: effect of merging epochs e_m on accuracy. Right: accuracy for variable batch size, G_{30} and G_{50}.



(a) PCB

(b) OSNet

Fig. 3. Average % benefit of fine-tuning with respect to rank-1 accuracy. The experiment includes variable gallery sizes and variable learning rates using PCB and OSNet architectures.

Comparison to State-of-the-Art: We compare our approach to that of [33] which considers MSMT17 as source, while supervised training is conducted in a combined train + test dataset. MSMT17 is the largest ReID dataset containing 4101 persons captured from up to 15 views, much larger than Market1501 and DukeMTMC-reid. For fairness, we also utilized MSMT17 as the source dataset. To perform the aforementioned comparison, we conducted two types of experiments. First, we evaluated our work using the standard protocol they also use. Corresponding results are presented in Table 5 (left). Our method is able to obtain some benefit with respect to the baseline model, however smaller than the one of [33]. We stress that this experiment regards a very large dataset (MSMT17) as source. This captures a wide variety of appearance characteristics, letting the supervised training to generalize better. Subsequently, impact of fine-tuning on the target domain is smaller. In a real-world scenario however, such suitable source dataset will not always be available. We also note that the particular experiment incorporates imagery of auxiliary persons from the same domain (DukeMTMC-reid, train); a hard requirement which limits deployment opportunities. Under these circumstances we consider that this experiment does not address real-world demands. Second, in contrast to standard evaluation, we employ our novel experimental protocol which considers multiple random subsets of two datasets in order to simulate real-world situations, *without* utilizing any same-domain auxiliary dataset. The results of the comparison with [33] using our experimental protocol are shown in Table 5 (right)). In this case, both methods achieve a small benefit, while the largest accuracy is obtained by UFT-ReID with a combined loss approach (third row). An explanation for this can be based on the evidence provided in Tables 3 and 5 (right) for the case of G_{33}. The accuracy of the Market1501-trained base model is much lower than the MSMT17-trained base model for the same (OSNet) architecture, i.e. 64.5% vs. 89.08%. Thus, the MSMT17 dataset results in an already capable baseline model, therefore fine-tuning is left with less room for substantial improvement (Table 5 (right)).

Comparison to OSNet Variants: In all previous experiments, OSNet refers to the originally proposed architecture, precisely denoted as "osnet_x1_0". This version was proposed for same-domain supervised person ReID. More recently, the authors of OSNet released two subsequent versions of their architecture [44,45], specialized to generalize to new domains. These new architectures, denoted as OSNet-IBN, and OSNet-AIN, are modifications to the original, however they are able to address the substantial domain shift, resulting to better baselines. Table 6 presents the benefit that UFT-ReID is able to achieve starting from these new baselines, in an experiment with a G_{33} gallery. The original OSNet architecture achieves a fairly low average rank-1 accuracy (64.5%), which is increased by 14% through UFT-ReID-based refinement. The OSNet-IBN variant sets a much better baseline performance (74.5%) than the original OSNet, which is again further improved by UFT-ReID by 3.1%. Interestingly, UFT-ReID refinement of the original OSNet, results in better accuracy (78.5%) compared to either the baseline or the refined OSNet-IBN variants. OSNET-AIN starts with an even better baseline (76.5%). Still, UFT-ReID improves it further by 3.5%, achieving the best result among the three variants.

Training Execution Time: Table 7 shows average training execution times of UFT-ReID in a couple of experimental settings involving the OSNet architecture. Duration varies, depending on the amount of training images contained in the gallery. We notice that such durations are acceptable for some non time-critical applications , such as for crowd analytics or cross-camera person tracking in smart spaces.

We also compare the execution time of UFT-ReID to that of [33]. For a fair comparison, we measured parts of the training process that are relevant to each method, i.e. initializations, computations of losses and clustering. A G_{33} experiment has shown that the execution time of [33] is 2.5× the execution time of UFT-ReID. This is attributed mainly to the heavy computations required within the Ipfl loss.

4.4 Ablation Study

We now show how different parameters affect the performance of our framework. In most cases we chose small galleries, which is the main focus of our work. In all following experiments we used the Market1501 as the source dataset and the DukeMTMC-reid as the target dataset.

Internal Parameters: At first we experimented with two internal parameters of the bottom-up clustering algorithm; the diversity regularization λ and the number of merging epochs e_m. Experimentation on different values of λ is also conducted by [18]. Our motivation for repeating the experiment, is both due to the different formulation of the problem, as well as the utilization of fewer data. More specifically, the chance that two persons share similar appearance is smaller in the case of small-sized galleries. Experimentation was conducted on 20 random galleries containing 30 persons.

Figure 2 (left) demonstrates the effect of altering λ while keeping all other parameters fixed. We confirmed that the optimal value for λ is 0.005.

Next, we experimented with the effect of increasing e_m. Our motivation is to let the CNN stabilize between two cluster merging operations where the state pseudo-labels remain the same. Our experiments included the following options for e_m: $\{2, 3, 4, 5, 7, 11\}$.

Figure 2 (middle) shows the average rank-1 accuracy for these options. It turns out that by increasing e_m, we obtain better accuracy. However, this is at the cost of a more time-consuming fine-tuning, as training time increases considerably. More specifically, in our experiments the training time when using $e_m = 11$ was about three times more, compared to using $e_m = 2$. In all other experiments, for achieving reasonable training duration, we kept $e_m = 2$.

Batch Size: We conducted experiments using random 30- and 50-person galleries. Figure 2 (right) depicts a negative trend on the average rank-1 accuracy when using batch sizes larger than 16. Therefore, we chose $b = 16$ for subsequent experiments.

Data Augmentation: Typical CNN optimization requires lots of training images in an effort to generalize to diverse scenarios. To compensate for lack of such images, various online or offline data augmentation methods have been proposed.

In [18] images are randomly cropped and horizontally flipped, in an online fashion. We further experimented with randomly altering the color properties of images, including brightness, contrast, hue and saturation. The motivation is that our method is evaluated on smaller galleries, therefore lack of appearance diversity is expected as opposed to the case of a complete dataset. We conducted an experiment of random 30-person galleries to investigate the impact of augmentation based on color jittering. Using such augmentation, rank-1 accuracy increased from 65.7% to 68.1% in the particular experiment. Its impact was quite small, however it was used in subsequent experiments.

Learning Rate: As discussed in Sect. 3.2, lr is a crucial parameter for fine-tuning, as it balances the adaptation to the new domain and the preservation of the already learned knowledge. Preliminary experiments with small galleries (20–50 persons) shown that the initial learning rate should be fixed around 10% of the final value of a stepped learning rate reduce approach during supervised training. Our experiments in the full-scale experiment and dataset ranges (Sect. 4.1) confirmed that such choice for learning rate is satisfactory for galleries with few persons. For larger gallery sizes performance degraded significantly, and even worsens the original trained model. For this case we experimented with lower lr.

Figure 3 demonstrates the benefit of fine-tuning for variable gallery sizes and three options of lr. The outcome of the experiment is that larger gallery sizes require smaller learning rates. As shown in Fig. 3, this happens regardless of the choice of architectures (i.e., PCB or OSNet). We interpret this result as follows. The original model is trained on a large variety of person appearances. By fine-tuning we want to extend the model to new persons (i.e. appearances) from the target domain. In traditional supervised training, the choice of learning rate controls how large of a step to take in the direction of the negative gradient of the loss function. The original model correctly represents such appearance diversity, therefore smaller learning rates are required for not moving too far from the initial solution. On the other hand, in the case of smaller gallery sizes we want to mitigate the expressiveness of the model. This is achieved by adapting to fewer data while using larger learning rates.

5 Summary and Discussion

We presented UFT-ReID, a novel method for person re-identification. UFT-ReID performs fine-tuning and adaptation of a model already learned on an auxiliary, source dataset to a new, target one. It does so, with no requirement for training data on the target domain. Thus, it is compatible with real-world applications that require easy deployment of ReID methods in novel settings.

We also presented a new evaluation protocol, that is more suitable for real-world demands. Several experiments were conducted, demonstrating that UFT-ReID is able to adjust models and improve their accuracy, bringing them above state-of-the-art performance. Additionally, a number of experiments explored the parameter space of UFT-ReID, providing evidence on proper parameter settings and relevant justifications.

Ongoing work considers the extension of UFT-ReID by incorporating other unsupervised losses, training methodologies including early stop, as well as experimentation with novel optimizers. Another important topic of ongoing research considers the improvement of the cluster merging criterion described in Sect. 3.2. The selection of this criterion is crucial because it relates to the pseudo-labels generated during the fine-tuning. Errors due to false cluster merges, eventually propagate as errors within the objective function. Indicatively, in [5], another merging criterion is proposed, exploiting feature affinities within and between clusters. Preliminary experiments with a G_{33} gallery using the PCB architecture shows that the use of this criterion increases further the rank-1 accuracy of UFT-ReID by 6.6%.

Acknowledgments. This work is partially funded by the H2020 project CONNEXIONs (GA 786731) and the FORTH-ICS internal RTD Programme "Ambient Intelligence and Smart Environments".

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR (2015)
2. Bai, X., Yang, M., Huang, T., Dou, Z., Yu, R., Xu, Y.: Deep-person: learning discriminative deep features for person re-identification. Pattern Recogn. **98**, 107036 (2020)
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: ECCV (2018)
4. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: CVPR (2017)
5. Ding, G., Khan, S.H., Tang, Z.: Dispersion based clustering for unsupervised person re-identification. In: BMVC, p. 264 (2019)
6. Ding, Y., Fan, H., Xu, M., Yang, Y.: Adaptive exploration for unsupervised person re-identification. ACM TOMM **16**(1), 3:1–3:19 (2020)
7. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification: clustering and fine-tuning. ACM TOMM **14**(4), 83 (2018)
8. Fan, X., Jiang, W., Luo, H., Fei, M.: Spherereid: deep hypersphere manifold embedding for person re-identification. Vis. Commun. Image Represent. **60**, 51–58 (2019)
9. Fu, Y., et al.: Horizontal pyramid matching for person re-identification. In: AAAI, vol. 33 (2019)
10. Guo, Y., Cheung, N.M.: Efficient and deep person re-identification using multi-level similarity. In: CVPR (2018)
11. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
12. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
13. Hsu, H.K., et al.: Progressive domain adaptation for object detection (2019)
14. Jia, J., Ruan, Q., Hospedales, T.M.: Frustratingly easy person re-identification: generalizing person re-id in practice. arXiv preprint arXiv:1905.03422 (2019)
15. Kalayeh, M.M., Basaran, E., Gökmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In: CVPR (2018)
16. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReiD: deep filter pairing neural network for person re-identification. In: CVPR (2014)
17. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: CVPR (2018)
18. Lin, Y., Dong, X., Zheng, L., Yan, Y., Yang, Y.: A bottom-up clustering approach to unsupervised person re-identification. In: AAAI (2019)
19. Liu, W., Chang, X., Chen, L., Yang, Y.: Semi-supervised Bayesian attribute learning for person re-identification. In: AAAI (2018)

20. Lucic, M., Tschannen, M., Ritter, M., Zhai, X., Bachem, O., Gelly, S.: High-fidelity image generation with fewer labels. arXiv preprint arXiv:1903.02271 (2019)
21. Quan, R., Dong, X., Wu, Y., Zhu, L., Yang, Y.: Auto-ReID: searching for a part-aware convnet for person re-identification. arXiv preprint arXiv:1903.09776 (2019)
22. Ristani, E., Tomasi, C.: Features for multi-target multi-camera tracking and re-identification. In: CVPR (2018)
23. Saquib Sarfraz, M., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: CVPR (2018)
24. Shen, C., et al.: Sharp attention network via adaptive sampling for person re-identification. IEEE CAS (2018)
25. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.: Pose-driven deep convolutional model for person re-identification. In: ICCV (2017)
26. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV (2018)
27. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM Multimedia Conference. ACM (2018)
28. Wang, J., Zhu, X., Gong, S., Li, W.: Transferable joint attribute-identity deep learning for unsupervised person re-identification. In: CVPR (2018)
29. Wu, A., Zheng, W.S., Lai, J.H.: Distilled camera-aware self training for semi-supervised person re-identification. IEEE Access **7**, 156752–156763 (2019)
30. Wu, D., et al.: Deep learning-based methods for person re-identification: a comprehensive review. Neurocomputing **337**, 354–371 (2019)
31. Xiao, T.: Open-ReID framework (2016). https://github.com/Cysu/open-reid
32. Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: CVPR (2018)
33. Yang, Q., Yu, H.X., Wu, A., Zheng, W.S.: Patch-based discriminative feature learning for unsupervised person re-identification. In: CVPR (2019)
34. Yao, H., Zhang, S., Hong, R., Zhang, Y., Xu, C., Tian, Q.: Deep representation learning with part loss for person re-identification. IEEE Trans. Image Process. **28**(6), 2860–2871 (2019)
35. Yu, H.X., Wu, A., Zheng, W.S.: Unsupervised person re-identification by deep asymmetric metric embedding. TPAMI (2019)
36. Yu, H.X., Zheng, W.S., Wu, A., Guo, X., Gong, S., Lai, J.H.: Unsupervised person re-identification by soft multilabel learning. In: CVPR (2019)
37. Yu, R., Dou, Z., Bai, S., Zhang, Z., Xu, Y., Bai, X.: Hard-aware point-to-set deep metric for person re-identification. In: ECCV (2018)
38. Yu, T., Li, D., Yang, Y., Hospedales, T.M., Xiang, T.: Robust person re-identification by modelling feature uncertainty. In: ICCV (2019)
39. Zhang, X., et al.: AlignedReiD: surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184 (2017)
40. Zhang, Z., Lan, C., Zeng, W., Chen, Z.: Densely semantically aligned person re-identification. In: CVPR, June 2019
41. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: ICCV (2015)
42. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In: ICCV (2017)
43. Zhou, K., Xiang, T.: Torchreid: a library for deep learning person re-identification in pytorch. arXiv preprint arXiv:1910.10093 (2019)

44. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Learning generalisable omni-scale representations for person re-identification. arXiv preprint arXiv:1910.06827 (2019)
45. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV (2019)