# Anchor-ReID: A Test Time Adaptation for Person Re-identification

Mohammed Almansoori, Mustansar Fiaz, and Hisham Cholakkal(✉)

Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{mohammed.almansoori,mustansar.fiaz,
hisham.cholakkal}@mbzuai.ac.ae

**Abstract.** Person re-identification (ReID) is a challenging computer vision problem where the objective is to retrieve a person of interest from a gallery of images. Conventional person ReID methods struggle to generalize across different domains, leading to inferior cross-domain performance. To address this, domain generalization (DG) person re-id methods are proposed. The majority of DG studies focus on designing more robust models that are trained on source datasets in an offline setting with explicit components or training mechanisms for domain generalization. Moreover, these frameworks do not take into consideration that in the real world the environment is in continual change resulting in the degradation of the discriminative feature extraction. To tackle these problems, we propose a new problem formulation of Test-time adaption for ReID (TTA ReID). Our TTA ReID takes a pre-trained model that was not designed for domain generalization from source domain, and aims in inference to adapt it to the current target domain in an unsupervised setting. To this end, we propose *Anchor ReID*, a framework designed to adapt a pre-trained model without altering its network architecture and make it robust to domain shift. Comprehensive experiments on CUHK03, Duke-mtmc, and Market 1501 datasets demonstrate the benefits of the proposed approach. The proposed Anchor ReID framework can improve a pre-trained model (that was not designed for domain generalization) and achieves an absolute gain of mAP 10 in the CUHK03 dataset.

## 1 Introduction

Person ReID seeks to match a pedestrian's visual representation to a gallery of images taken by a camera network. In general, Person ReID is a difficult problem to solve due to numerous challenges such as (i) requirement of re-identifying (matching) a person across different camera sensors, (ii) intra class variations like changes in clothings, view variations, scale changes and deformations (iii) background clutter and distracting objects. Despite these challenges, deep learning based models [1–5] made considerable progress in recent years achieving state-of-the-art performance in a supervised setting by extracting high-quality discriminative features. The majority of recent studies focus on designing novel loss functions and proposing better feature-extraction strategies for person Reid. Additionally, alternative strategies such as the effectiveness of batch normalization [1] and centroid loss [2] are also investigated in the literature.


R. Gade et al. (Eds.): SCIA 2023, LNCS 13886, pp. 599–612, 2023.
https://doi.org/10.1007/978-3-031-31438-4_39

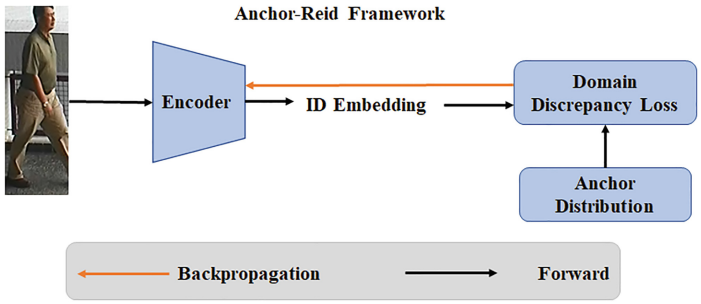



**Fig. 1.** The objective of a Test time adaptation (TTA) model is to update the encoder parameter $\theta$ through the testing phase stage, resulting in updating the weights $f_\theta \rightarrow f_{\theta+1}$, in each step. The proposed method Anchor uses feature distribution alignment to minimize the divergence of the domain shift.

Although recent supervised deep learning based approaches achieve impressive performance in scenarios where the source and target datasets are from the same domain, they struggle to generalize across different domains. However, in practical re-identification applications, it is highly probable that we may have to deploy the re-identification model on target domains that are different to the source domain used to train the model. For example, various aspects such as the camera sensors, camera viewpoints, background scenarios, illuminations, clothing style of people, etc. may differ between the source and target domains. In recent years, progress in unsupervised domain adaptation (UDA) ReID [6–8] achieved great performance that avoided the extensive labor work for labeling target domain data. Utilizing the source domain labeled data with the target domain unlabeled data allowed for better discriminative feature embedding and improvement in results. Pseudo-labeling for UDA led to great progress in learning better discriminate features in the target domain with methods [7] that exploit the dynamic memory aspect of training and pseudo-label the target domain. These solutions require more resources and training time for a model to adapt better to a new domain. Even, with methods that utilize a more adversarial approach of adapting to new domains [9–11] they require access to the source data and are expensive in terms of resources needed for training.

Domain Generalization (DG) in person Reid goes beyond the simple premise in UDA, by designing a more robust model to domain shift and does not access the target domain data. The common strategy here is to train a model on single/multiple datasets and test it on other target dataset/s to measure the generalizability to unseen domains. A combination of batch normalization (BN) and instance normalization (IN) [12–14] improved the model statistical normalization to be invariant to domain shift. Another direction is to match the convolution feature maps [3,15] between the gallery and the query images, resulting in a more robust generalizable framework.

The majority of DG studies focus on designing more robust models that are trained on source datasets in an offline setting with explicit components or training mechanisms for domain generalization. Similarly, UDA methods [7–9,16,17] utilize labeled source domain data as well as unlabelled target domain data and train their model in an offline learning setting. Popular DG [18–20] and UDA approach disregards the fact that the

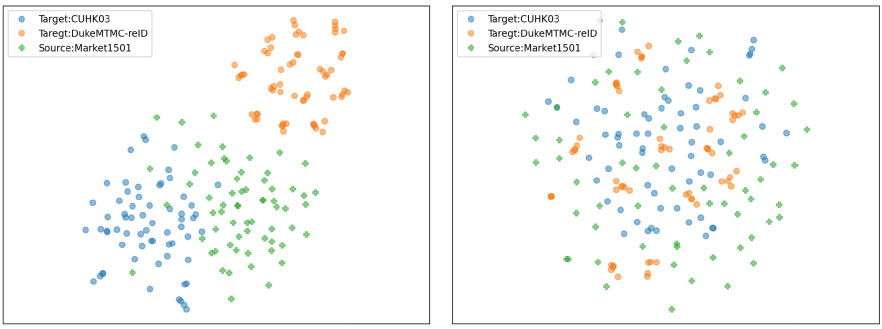
real-world environment is in continual change and trains their model in an offline setting, without any method to update the model parameters when encountering data from unseen domains. A major setback to a more practical model for real-world application is the adaptability of the model to changes to the environment and deployment in a domain that is different in setting than the source domain. Thus, we propose a new problem formalization designed for more practical usage in real-world applications, named as Test-time adaption (TTA) for re-identification. Our TTA problem setting takes a pre-trained model (that is not designed for domain generalization) and adapts it to the given target domain during inference, while not having access to the pre-trained source dataset.

In this work, we propose a framework for TTA named Anchor Reid, that utilizes a parameter $\theta$ updating method similar to [18] with a novel anchor distribution strategy to summarize the source domain as seen in Fig. 1. The proposed Anchor ReID utilizes KL-divergence to measure the discrepancy between the two data sources [21]. Our framework can take any pre-trained re-id model from the source domain and it does not require additional components or explicit training strategy on the source data for domain generalization. Therefore, the proposed method is designed to be compatible with any fully supervised person ReID method. In summary, the contributions of this paper are as follows:

- Introducing a new problem formulation of TTA in Person Reid that focuses on cross-domain evaluation with a unique set of challenges.
- We design a new framework that takes a pre-trained Reid model and built a summarization distribution from the source domain to effectively adapt to the target domain.
- We show that by using a pre-trained model trained on Market1501 [22], Anchor ReID achieved an absolute gain of over 10 mAP on CUHK03 [23] and Duke-mtmc [24] benchmarks when compared to the baseline [1] model.

## 2 Related Work

**Domain Generalizability:** The objective for DG ReID is to learn a more robust model that is invariant to domain shift to perform better to unseen target domains. When compared to standard ReID the aim there is for a more discriminative visual representation for each instance. For DG [14] proposed to use instance normalization (IN) to eliminate style variations, and to disentangle relevant features from the irrelevant features for a more robust model. Normalization has a major role on the model robustness which is explored in [13] and [25] by replacing batch normalization or adding instance normalization to bottleneck. These approaches require a lot of trial and error to get a consistent and robust solution. MetaBIN [12] proposed the combination of BN and IN together with using meta-learning few-shot reinforcement learning to boost the generalization capability. Image augmentations are used in [26] to make a model generalizable to unseen data. Instead of using embedding ReID features for matching another approach is to extract a feature map and maintain the spatial information of the image. [3,15] Extract feature maps and then pass them into a kernel that calculates each pair's similarity score in a network. Furthermore, utilizing a better sampling [3] of data through a graph improves the model's accuracy.




(a) Baseline[1] distributions (b) Anchor ReID distributions

**Fig. 2.** We visualize distribution for each domain by taking 64 samples and projecting them into 2D feature space using TSNE. On the right: each domain distribution is occupying a region of the 2D space, resulting in the degradation of the discriminative feature extraction. In domain adaptation, the main objective is for the encoder to adapt to new domains. Thus, in the right image the Anchor ReID blurs boundaries between domains, which enforces the encoder to push the target domains means $\mu$ into the Source domain.

**Test-Time Adaptation:** When it comes to UDA the methods require access to the source domain and the target domain. However, in practice, TTA does not require access to the source domain for the adaptation at inference. In recent years, fine-tuning a source-trained model is a more cost-effective approach in terms of resources and training time. An approach to minimize the uncertainty of a model is to minimize the entropy of the prediction by updating the model normalization parameters. Test entropy minimization (TENT) [18] calculates the cross-entropy of the prediction and updates the batch normalization layers to effectively adapt to the new domain. By only updating normalization parameters $\theta_{norm}$ the model reduces the risk of diverging. [19] utilize self-supervising learning with SimClR to train the model, and then generate a summarization for the source domain. In inference, TTT++ uses the distribution of the source domain to minimize the distance of the target domain. Similar to Tent, [16] focuses on updating only the classification layer alone, which makes the model a propagation-free solution for TTA. This is done by generating pseudo labels of the previous inferences to update the classifiers.

To the best of our knowledge, there is no work that has been published for a TTA re-identification solution, which makes our work the first in the field.

## 3 Problem Formulation

It is important to describe in detail the problem formalization of TTA ReID with the constraints. The work here is based on the work done in TTA for classification, but the work here adds some constraints for privacy since ReID is a more sensitive security application. The proposed method differs from the other ReID settings as in Table 1. TTA ReID doesn't access Source domain images for updating the model in inference

**Table 1.** Summary of the different ReID domain settings and how our proposed method differs from previous methods. The training Data describes how a model needs to use data in either the training for updating the model in the training or testing stages. Since Anchor ReID takes Pre-trained weights it does already trained, but it is only updating the model $\theta$ in inference, unlike other settings.

| | Training Data | | Learning Phases | |
|---|---|---|---|---|
| ReID Settings | Training Set | Testing Set | Training Stage | Testing Stage |
| Supervised ReID | Source | Source | Yes | No |
| UDA | Source and Target | Target | Yes | No |
| DG ReID | Source | Target | Yes (Target-aware) | No |
| Proposed (Anchor ReID) | Source(Pretrained) | Target | No (Pretrained) | Yes (Unsupervised) |

similar to the other domains. The main objective is to make a model adapt in inference, unlike other domains focusing on better feature discrimination. A pre-trained model $f_\theta(x)$ that is trained on $(X^s, Y^s)$ is used for inference on a stream of data in $(X^t)$. In addition, $(x_i^t)$ is only accessible once through inference. The goal is to update the model $\theta$ for it to adapt better to the target domain. The target set is split into multiple streams with specific size $X^t = (X_1^t, X_2^t, X_3^t, ....X_n^t)$, and the streams are sent sequentially to the model. Each instance of the stream is only processed once by the model to mimic the real world. The model process the output of the model $f_{\theta i-1}(X_i^t)$ and the model update in each iteration $\theta_i \rightarrow \theta_{i+1}$.

## 4 Method

Our main objective is to minimize the disparity between the source and target setting for the ReID model in a test-time adaptation setting, in which target-set data are divided into a stream of data and are only accessible ones. Anchor-ReID is built on top of multiple modules that together make it possible to adapt an off-the-shelf ReID model. In Sect. 4.1 the model utilizes a sampling strategy to select the samples that summarize the source domain. In 4.2 describes the policy of selecting parameters for updating and optimization. In this case, the BN is the only parameter that is being optimized. Finally, we design a way to measure the discrepancy between the different datasets.

### 4.1 Sampling the Anchor Memory

Taking inspiration from works in unsupervised Domain adaptations [3,9,17], Anchor ReID takes the alignment learning approach and tries to minimize the discrepancy between the source and target datasets. In the offline stage Fig. 3, we construct a subset from the source data by randomly selecting images from each ID class. Assuming that $f_\theta(x)$ is robust on the training set, randomly selecting samples for each ID will have a minimal effect since each ID would have an almost identical instance embedding. The $f_\theta(x)$ would generate the $A \in R^{C \times d}$ where C is the number of ID classes, and A represents the Anchor Memory. The next step is to minimize the number of samples in the memory by selecting P samples. With the given distribution, a pairwise distance

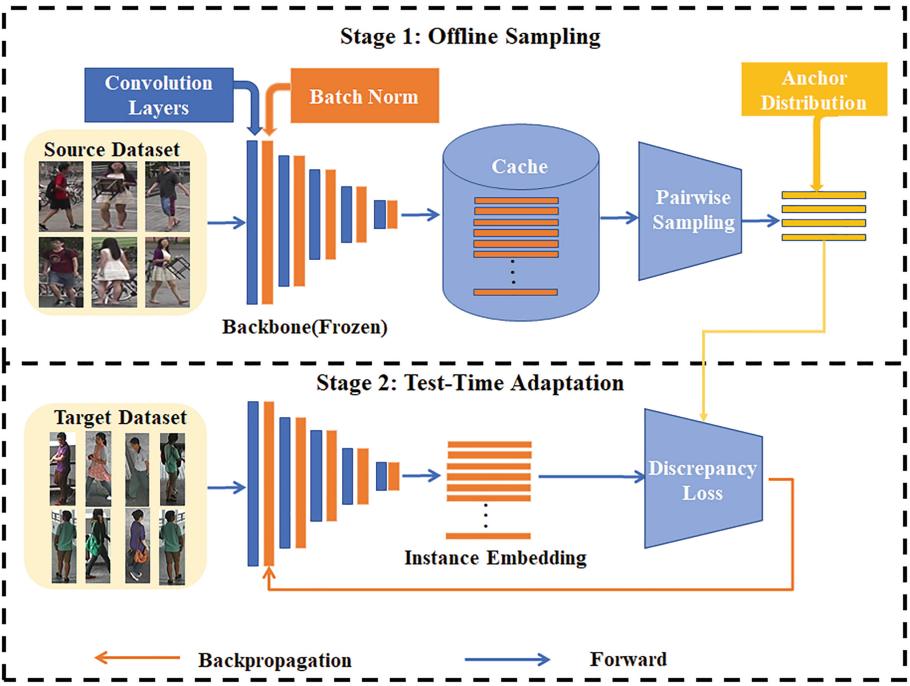



**Fig. 3.** The proposed Anchor ReID framework comprises of two stages: offline sampling and a test time adaptation (TTA). (i) At first, in the offline sampling stage we take an off-the-shelf person ReID model and extract a single embedding for each ID class in the source dataset $X^s$. These embeddings are stored inside a cache among which P samples are further selected as anchors, following an anchor distribution, that summarizes the source data. (ii) Next, in the test-time adaptation stage, we obtain the embeddings for the test data which are utilized to compute the disparity between the distributions of test data and the anchor distribution. Finally, the normalization layers in our model are updated to minimize the disparity between the anchor and target distributions.

is calculated for each class ID $dist \in R^{C \times C}$. Next, the most centered ID instance in the distribution is selected, and then the nearest $P-1$ samples are selected. Selecting samples that are near each other is essential since it reduces the variance in distribution. The Anchor memory is acting as a summarization of the source domain and enables the framework to eliminate the need to accessing $X^s$ during TTA.

### 4.2 Parameter Selection and Update

ReID models encode the feature representations of the pedestrian into a single embedding vector. While the model has a high level of visual discrimination features, it makes the process of optimizing the model parameter $\theta$ more difficult. Optimizing the $\theta$ of the whole model may cause it to diverge significantly by damaging the robustness of the visual extraction. To minimize the sensitivity of optimizing the model $\theta$, we select the Normalization layers in order to adapt the model to the target domain. Similar to

[18] the model normalization $\theta$ is updated through inference. Through updating the normalization running mean and running variance the model would shift them to be more in line with the target domain, thus reducing diverged risk of the model optimal $\theta$. In sense, ReID models are designed to be highly discriminative of different IDs, by extracting features embedding that is highly similar to the same ID class and repealing embeddings of different IDs. In Anchor ReID case the BN on the final layer [1] is a more important aspect since it affects the similarity measurement as can be seen in Fig. 2.

Next, a challenging aspect is to measure the uncertainty of the predicted output $Z_i^t$, since in TENT [18] and other TTA frameworks they use cross-entropy to measure the loss. Given that $X^t$ is the only data available, an adversarial feature alignment UDA method is required for a robust adaptation in test time. The strategy is to impose the feature distribution of the target domain to be near the training domain. For feature alignment, Maximum Mean Discrepancies [21] (MMD) is wildly used for adversarial learning in unsupervised Domain adaptation, by measuring the distance between two different means of distributions acting as a discriminator loss function. The Anchor memory $A \in R^{P \times d}$ is used as an anchor constraining the feature distribution of $X_i^t$ to stay near the source distribution, allowing to minimize the latent feature discrepancy as seen in Fig. 2. The features are reproduced using the kernel Hilbert $\mathcal{H}_k$ on $\mathbb{P}$ distribution to generate mean embedding of the $\mu_k(\mathbb{P})$. Given Anchor $A$ and Target $X^t$, MMD distance between reads $MMD^2(A, Z^t) = ||\mu_A - \mu_Z||^2_{\mathcal{H}_k}$. In the case of the proposed method the model doesn't have access to the full distribution thus we estimate it as follows using Gaussian kernel:

$$MMD^2 = l(A, Z^t) = \frac{1}{m(m-1)} \sum_i \sum_{j \neq i} k(z_i, z_j) - 2\frac{1}{m.m} \sum_i \sum_j k(z_i, a_j) + \frac{1}{m(m-1)} \sum_i \sum_{i \neq j} k(a_i, a_j) \tag{1}$$

## 5 Results

### 5.1 Implementation Details

Our experiments are adapted from the official Pytorch code of [1]. It was selected to mimic the experimenting environment of TTA papers since it uses Resnet50 as a backbone and the availability of multiple public pre-trained weights from different datasets and configuration settings. TTA main premise is to take off a shelf model and adapt it in inference. Thus, by stating these design decisions we are showing that we are taking pre-trained weights and just adapted in inference. The batch size is set to 128 with a similar number of samples for $P = 128$. Adam optimizer is used for Anchor ReID with a learning rate of 0.0003 and beta between $(0.9 - 0.99)$.

**Table 2.** Comparison on cross-domain evaluation for person re-identification using CMC and mAP evaluation metrics. The [1] weights are taken as baseline to measure (%) change with Anchor ReID. The Anchor ReID (%) is the mean and standard deviation from running the experiment 10 times, due to shuffling the Test set.

| | | Target:CUHK03 | | Target:Duke-mtmc | |
|---|---|---|---|---|---|
| Source: Market-1501 | | mAP | r = 1 | mAP | r = 1 |
| Supervised ReID | Baseline [1] | 4.72 | 4.14 | 15.2 | 27.0 |
| DG ReID | MuDeep [27] | 10.3 | 9.1 | 27.6 | 47.5 |
| | QACONV [15] | 9.9 | 8.6 | 33.6 | 54.4 |
| | METABIN [12] | – | – | 33.1 | 55.2 |
| | QACONV-GS [3] | 19.1 | 18.1 | – | – |
| TTA ReID | Anchor ReID on [1] | 14.7 ± 0.7 | 13.5 ± 1 | 25.5 ± 0.5 | 41.1 ± 0.8 |
| | | Target:CUHK03 | | Target:Market1501 | |
| Source: Duke-mtmc | | mAP | r = 1 | mAP | r = 1 |
| Supervised ReID | Baseline [1] | 5.12 | 4.85 | 21.4 | 47.4 |
| DG ReID | MuDeep [27] | – | – | – | – |
| | QACONV [15] | 7.9 | 6.8 | 31.6 | 62.8 |
| | METABIN [12] | – | – | 35.9 | 69.2 |
| | QACONV-GS [3] | 19.1 | 18.1 | – | – |
| TTA ReID | Anchor ReID on [1] | 8.8 ± 0.6 | 8.5 ± 0.6 | 24.1 ± 0.5 | 52.4 ± 1.3 |

### 5.2 Datasets

The datasets were selected w.r.t the public pre-trained weights of [1]. CUHK03 [23], Market-1501 [22], and Duke-mtmc [24] are the datasets for experiments. The pretrained weights are trained on Market-1501 and Duke-mtmc. CUHK03 dataset is composed of 1,360 pedestrians from 13,164 images. The Market-1501 dataset includes 1,501 IDs which are split into 12,936 images for training and 19,732 images for testing. Finally, the Duke-etc dataset contains 16,522 training images of 702 identities and 702 identities with 19,889 images for testing.

**Evaluation Protocol.** TTA standard evaluation metric consists of minimizing the error of the model prediction. However, for Anchor ReID we followed the ReID community evaluation metric of Cumulative Matching characteristics (CMC) curve and mean Average Precision (mAP). Similar to domain generalization the objective is to measure cross-domain performance.

### 5.3 Results

Table 2 compares Anchor ReID (%) utilizing the baseline [1] public weights. For comparison with other state-of-the-art models, we indirectly compared Anchor ReID with DG models, due to similar evaluation metrics and show the challenges TTA ReID needs to overcome. Overall, Anchor ReID shows an impressive improvement in all case scenarios with +10 mAP in CUHK03 when trained on Market1501 and +3.7 mAP when

the source is Duke-mtmc. For example, in Market1501 → Duke-mtmc and on Duke-mtmc → Market1501 cases, our proposed method improves the baseline significantly averaging between +4 to +10 mAP. However, in these cases, the proposed method is still behind DG models. Interestingly, in Market1501/Duke-mtmc → CUHK03 the (%) of Anchor ReID is overcoming the majority of DG models and showing a performance gain of +10 mAP. In general, the Anchor ReID is capable of improving on the baseline in all cases, but when compared with Domain Generalization models, it still lacking in some scenarios. QACONV-GS [3] shows a better performance in CHUK-03 in all cases. However, in consideration, the Anchor ReID doesn't add or change anything of the original model (Resnet50) when compared with DG. Both TTA and DG over have different constraints and challenges. A limiting factor for Anchor ReID is that the model $f_\theta$ is updating through testing resulting in better instances embedding at the end. The first few iterations pull down the mAP and top-1 scores. The numbers reported in Table 2 are the mean and standard deviation when Anchor ReID is tested with a randomized stream of data in testing. Randomization of the data stream shows a better improvement in overall accuracy when adapting to the target sets. This is the result of how the dataset data are formatted. For CUHK03, Market1501, Duke-mtmc the data are ordered by ID class, thus when we measure the divergence of the distributions, this results in a sub-optimal understanding of the real distribution of the target set. For comparison, Table 4 shows the mean of 14.7 mAP when shuffling and 10 mAP when not shuffling. This is explored more in detail in Table 4.

### 5.4 Ablation Study

We performed analytic experiments to understand the limitations and effectiveness of Anchor ReID in different settings. Table 3 shows how data streaming and sampling have an effect on the model performance. A limiting factor of the study is the randomization aspects of the testing set order, which results in a different score each time. To limit the effect of the randomization all the reported numbers are the results of measuring the mean and standard deviation of the experiments. Each experiment is done 5 times for a more consistent testing environment. In Table 3, the results variate for each testing set, but there is a vast difference in Market1501 to Duke-mtmc when the anchor samples are selected randomly. This results in a decrease of $-1.7$ mAP and $-3.4$ top-1 accuracy. This is explained since the sampling the Anchor distribution from the pairwise sampling method the distribution of the mean would be closest to the actual mean of the source data, while the randomizing would affect by shifting the mean in unpredictable directions.

A limiting factor of the Anchor ReID is how many samples are needed to improve the mAP and CMC scores in cross-domain testing. Table 4 show the effect of increasing the number of samples in two sets no shuffling and shuffling testing set. The Pattern in both scenarios is that the more samples you introduce in the Anchor ReID the more it improves the overall score. The Anchor ReID is a summarization of the $X^s$ distribution and the more samples you introduce the more accurately the disparency of the domains is calculated. For the majority of the testing scenarios, the model would start to adapt better as long, as it has at least 64 samples in Anchor. Furthermore, when not shuffling the data the model would degrade significantly as long the model has few samples to

**Table 3.** Sampling methods (%) on Anchor selection. For each sampling method, we tested them in two scenarios of shuffling and not shuffling the testing set.

| Random Sampling | | | | |
|---|---|---|---|---|
| | Market1501 → Duke-mtmc | | Duke-mtmc → Market1501 | |
| Method | mAP | r = 1 | mAP | r = 1 |
| No Shuffling | 22.2 | 36.3 | 24.6 | 50.4 |
| Shuffling | 22.7 ± 1.3 | 36 ± 2.3 | 24.3 ± 0.3 | 52.9 ± 1 |
| Pairwise Sampling | | | | |
| | mAP | r = 1 | mAP | r = 1 |
| No Shuffling | 22.8 | 37.7 | 24.2 | 49.6 |
| Shuffling | 25.5 ± 0.5 | 41.1 ± 0.8 | 24.1 ± 1.3 | 52.4 ± 1.3 |

**Table 4.** The performance (%) of Anchor ReID with the change of the number of samples used in testing in a Cross-domain setting. The proposed model performs better with smaller batch sizes when data are shuffled

| | | Market1501 → CUHK03 | | Market1501 → Duke-mtmc | |
|---|---|---|---|---|---|
| Sample size | | mAP | r = 1 | mAP | r = 1 |
| Shuffling | P = 16 | 9.2 ± 0.8 | 9.9 ± 1.2 | 3.6 ± 0.7 | 11.74 ± 1.68 |
| | P = 32 | 13.5 ± 0.6 | 13.4 ± 0.9 | 14.8 ± 1.2 | 30.8 ± 1.7 |
| | P = 64 | 14.3 ± 0.4 | 13.6 ± 0.5 | 22.1 ± 1 | 38.3 ± 1.2 |
| | P = 128 | 14.7 ± 0.7 | 13.5 ± 1 | 25.5 ± 0.5 | 41.1 ± 0.8 |
| No Shuffling | P = 16 | 0.9 | 1 | 0.6 | 1.8 |
| | P = 32 | 2.9 | 2.8 | 3.5 | 10 |
| | P = 64 | 7.1 | 7.2 | 22.8 | 37.7 |
| | P = 128 | 10 | 9.4 | 22.8 | 37.7 |
| | | Duke-mtmc → CUHK03 | | Duke-mtmc → Market1501 | |
| Sample size | | mAP | r = 1 | mAP | r = 1 |
| Shuffling | P = 16 | 7.38 ± 0.4 | 7.54 ± 0.4 | 21.8 ± 0.3 | 50.78 ± 0.9 |
| | P = 32 | 8.52± 0.6 | 8.6 ± 0.8 | 23.44 ± 0.3 | 52.26 ± 0.7 |
| | P = 64 | 8.42 ± 0.2 | 8.2 ± 0.5 | 23.78 ± 0.37 | 52.82 ± 1 |
| | P = 128 | 8.8 ± 0.6 | 8.5 ± 0.6 | 24.11 ± 0.57 | 52.4 ± 1.3 |
| No Shuffling | B = 16 | 2 | 2.3 | 21.1 | 46.2 |
| | B = 32 | 3.5 | 3.2 | 23.3 | 48.1 |
| | B = 64 | 5.3 | 5.1 | 24.1 | 49.8 |
| | B = 128 | 6.54 | 6.42 | 24.2 | 49.6 |

measure the divergence. The datasets sort image order by ID class, thus when using the feature alignment method of the latent embedding the $\mu$ is affected by sample $X_i^t$. Resulting in less than optimal (%).

### 5.5 Further Evaluations

Anchor ReID is affected by the number of samples available in inference. To simulate more data Table 5 shows that we took the training set from the target domain and used only the $X^t$ to update the model. We divided the testing into two settings. The first

**Table 5.** Performance (%) Comparison on how increasing the number of samples affects TTA. The Ablation is done by first TTA the model on the Training set of the target Domain with label data. Frozen means that after going through the training set the model is frozen in inference. In Full the model goes through the whole set without freezing for evaluation. On each category, the experiment is done by Shuffling the data and vice versa.

| | | Market1501 → CHUK03 | | Market1501 → Duke-mtmc | |
|---|---|---|---|---|---|
| Setting | Shuffling | mAP | r = 1 | mAP | r = 1 |
| Frozen | ✓ | 16.5 ± 0.7 | 16.3 ± 0.35 | 24.6 ± 0.6 | 39.9 ± 1.7 |
| Frozen | ✗ | 9.1 | 8.7 | 25.8 | 41.3 |
| Full | ✓ | 16.6 ± 0.2 | 16.6 ± 0.7 | 25.4 ± 0.7 | 41.2 ± 1.3 |
| Full | ✗ | 9.3 | 9.6 | 25.7 | 41.8 |

is to update the model on the training set and then freeze it back to the evaluation mode in inference (Frozen). Comparing the Suffled results in 2 and here there is a huge improvement in the CUHK03 results by +2.8 mAP and a slight decrease in Duke-mtmc by −1.4 mAP. The increase in CUHK03 could be attributed to the huge difference in the domains which shows that the first few predictions drag back the model's overall accuracy. In the second scenario, the model is not frozen in the inference stage and it continues adapting the new updated model. The model improves slightly in both target domains, but the most interesting insight is that the standard deviation is decreasing in the CMC and top-1 recall. This shows that over time Anchor ReID going to stabilize. This shows that in practicality Anchor ReID is more likely to perform better in real-world applications.

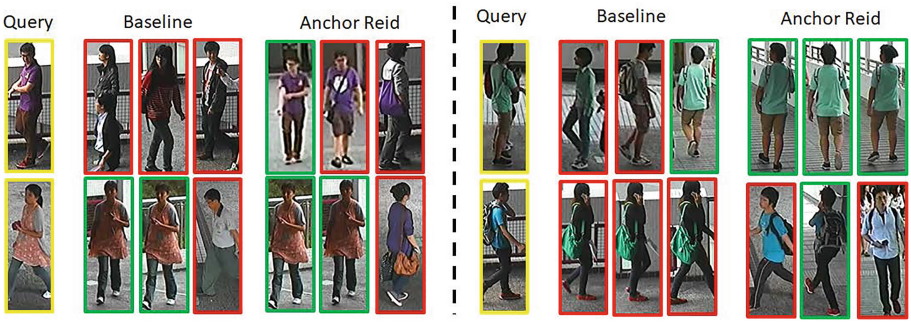



**Fig. 4.** Qualitative Results of the baseline and the proposed method. Yellow represents the Query image, Red for the wrong output, and Green for the correct output. (Color figure online)

Figure 4 we show visualize results of baseline [1] and the proposed method on CUHK03 dataset and returning the top-3 results. As can be observed, the Anchor ReID is able to retrieve images of pedestrians with similar visual cues. The disturbing observation is that the baseline is retrieving images of people who clearly don't have similar

features to the queue image. The baseline seems to confuse the background features with the query and instead returns images from the same viewpoint with individuals with a drastically different set of clothing. In cupper-right and bottom-left the baseline doesn't differentiate between colors and instead returns images of people with the same pose position and direction of the queue. While Anchor ReID is able to handle better color cues and return results from different viewpoints.

**Table 6.** Performance of proposed method on continual testing setting through multiple datasets. The testing goes through Target A and B. The number after $\pm$ is the standard deviation after 5 experiments.

| Market1501 → CHUK03 → Duke-mtmc | | | |
|---|---|---|---|
| CUHK03 | | Duke-mtmc | |
| mAP | r = 1 | mAP | r = 1 |
| 13.9 ± 0.6 | 13.5 ± 0.8 | 23.3 ± 1.3 | 38.9 ± 1.6 |
| Market1501 → Duke-mtmc → CHUK03 | | | |
| Duke-mtmc | | CHUK03 | |
| mAP | r = 1 | mAP | r = 1 |
| 23.5 ± 1 | 40.8 ± 1.7 | 14 ± 1.3 | 13.7 ± 1.1 |
| Duke-mtmc → CHUK03 → Market1501 | | | |
| CHUK03 | | Market1501 | |
| mAP | r = 1 | mAP | r = 1 |
| 8.5 ± 0.1 | 8 ± 0.4 | 24.5 ± 0.2 | 52.7 ± 1 |
| Duke-mtmc → Market1501 → CHUK03 | | | |
| Market1501 | | CHUK03 | |
| mAP | r = 1 | mAP | r = 1 |
| 24 ± 0.2 | 52.4 ± 0.5 | 9.1 ± 0.6 | 9 ± 1.1 |

To provide a more comprehensive study of the proposed method a continual test-time adaptation setting was conducted as shown in Table 6. The experiment illustrates a practical setting in which, the data stream source may change instantly depending on the task, simulating a real-world scenario. Conducting 4 different settings the proposed method maintained performance.

## 6 Conclusion

In this paper, we conduct an in-depth analysis of the potential of test-time adaptation for the person re-identification setting, and from our knowledge and we are the first to introduce this setting. To tackle the limited research on TTA for re-identification we proposed the novel method of Anchor ReID that is evaluated on cross-domain testing. The proposed method incorporates off-the-shelf models. To effectively adapt these models

we introduce a sampling method to build an Anchor distribution that summarizes the distribution of the source domain. The summarization helps to reduce the divergence of the target domain from the source domain. The proposed method was efficient in adapting to a new domain and showed promising improvements when compared to Domain Generalization models.

## References


1. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification (2019)
2. Wieczorek, M., Rychalska, B., Dabrowski, J.: On the unreasonable effectiveness of centroids in image retrieval (2021)
3. Yang, S., Kang, B., Lee, Y.: Sampling agnostic feature representation for long-term person re-identification. IEEE Trans. Image Process. **31**, 6412–6423 (2022)
4. Wang, H., Shen, J., Liu, Y., Gao, Y., Gavves, E.: NFormer: robust person re-identification with neighbor transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7297–7307, June 2022
5. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: TransReID: transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15013–15022, October 2021
6. Ge, Y., Chen, D., Li, H.: Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification (2020)
7. Ge, Y., Zhu, F., Chen, D., Zhao, R., Li, H.: Self-paced contrastive learning with hybrid memory for domain adaptive object re-id (2020)
8. Hu, Z., Zhu, C., He, G.: Hard-sample guided hybrid contrast learning for unsupervised person re-identification. arXiv preprint arXiv:2109.12333 (2021)
9. Dai, Y., Liu, J., Sun, Y., Tong, Z., Zhang, C., Duan, L.-Y.: IDM: an intermediate domain module for domain adaptive person re-id (2021)
10. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: exemplar memory for domain adaptive person re-identification (2019)
11. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Learning to adapt invariance in memory for person re-identification (2019)
12. Choi, S., Kim, T., Jeong, M., Park, H., Kim, C.: Meta batch-instance normalization for generalizable person re-identification (2020)
13. Jia, J., Ruan, Q., Hospedales, T.M.: Frustratingly easy person re-identification: generalizing person re-id in practice (2019)
14. Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification (2020)
15. Liao, S., Shao, L.: Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 456–474. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_27
16. Iwasawa, Y., Matsuo, Y.: Test-time classifier adjustment module for model-agnostic domain generalization. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 2427–2440. Curran Associates Inc. (2021)
17. Zheng, K., Lan, C., Zeng, W., Zhang, Z., Zha, Z.: Exploiting sample uncertainty for domain adaptive person re-identification. CoRR, vol. abs/2012.08733 (2020)

18. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization (2020)
19. Liu, Y., Kothari, P., van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: TTT++: when does self-supervised test-time training fail or thrive? In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 21808–21820. Curran Associates Inc. (2021)
20. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation (2022)
21. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A Kernel two-sample test. J. Mach. Learn. Res. **13**(25), 723–773 (2012)
22. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: a benchmark. In: IEEE International Conference on Computer Vision (2015)
23. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: CVPR (2014)
24. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 17–35. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_2
25. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: enhancing learning and generalization capacities via IBN-Net (2018)
26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision (2015)
27. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV (2019)