Ranking Models in Unlabeled New Environments

Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, Liang Zheng
Australian National University
{first name.last name}@anu.edu.au

Abstract

Consider a scenario where we are supplied with a number of ready-to-use models trained on a certain source domain and hope to directly apply the most appropriate ones to different target domains based on the models' relative performance. Ideally we should annotate a validation set for model performance assessment on each new target environment, but such annotations are often very expensive. Under this circumstance, we introduce the problem of ranking models in unlabeled new environments. For this problem, we propose to adopt a proxy dataset that 1) is fully labeled and 2) well reflects the true model rankings in a given target environment, and use the performance rankings on the proxy sets as surrogates. We first select labeled datasets as the proxy. Specifically, datasets that are more similar to the unlabeled target domain are found to better preserve the relative performance rankings. Motivated by this, we further propose to search the proxy set by sampling images from various datasets that have similar distributions as the target. We analyze the problem and its solutions on the person re-identification (re-ID) task, for which sufficient datasets are publicly available, and show that a carefully constructed proxy set effectively captures relative performance ranking in new environments. Code is avalible at https://github.com/sxzrt/Proxy-Set.

1. Introduction

In real-world applications, it is not uncommon to see models trained on the source domain (hereafter called source models) directly applied to unlabeled new target environments (hereafter called target domains) at the price of employing some unsupervised domain adaptation (UDA) techniques [15, 32, 19]. Assume that one has access to a pool of source models and can choose appropriate ones. Under this context, it is desirable to obtain the relative performance of different models on the target domain without having to annotate data in the target environment.

To find the appropriate models, we usually evaluate each individual model on a labeled partition (*e.g.*, a validation set) of the target environment and rank them to find the best one (see Fig. 1 **A**). However, annotation is often expensive to obtain, and it becomes prohibitive if we consider data labeling for every new application scenario. As such, an interesting question arises: can we estimate model rankings in new environments in the absence of ground truth labels?

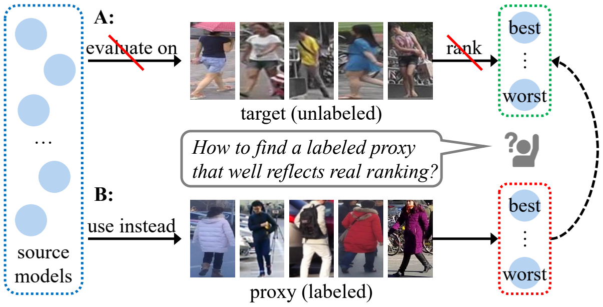


Figure 1: Illustration of the proposed problem and a general solution. Given various models (blue circles) trained on source data (hereon denoted as source models) and an unlabeled target domain, we aim to rank them and find the best one for direct deployment to the target. **A**: Without access to image labels, this objective is unlikely to be achieved using only the target data. **B**: We find a proxy to rank the model performance and use this (red) ranking as a surrogate. Specifically, this proxy should 1) be fully labeled and 2) well reflect the (green) true ranking on the target.

In this work, we aim to find a proxy (or surrogate) to rank the models in answer to the aforementioned question. Specifically, we focus on the person re-identification (re-ID) task, which aims to retrieve persons of the same identity across multiple cameras. For this problem, it is desirable that the proxy can provide similar model rankings, since a target validation set is difficult to acquire in practice. To this end, the proxy should satisfy: 1) have labels for evaluation and 2) well reflect the true model rankings (see Fig. 1 **B**).

For the first requirement (labels), we can either use the target dataset with *pseudo* labels, or other labeled datasets. However, due to the nature of pseudo labels, some of them might not be accurate. Existing works find that the inaccurate pseudo labels greatly influence the model accuracies when used in training [13]. We suspect such inaccurate pseudo labels may even do more harm when used for evaluation. As such, we consider using labels that are *real* and not from the target domain.

For the second requirement (a good reflection of the true ranking), we should consider the target data distribution. If we intuitively use the model rankings on the source domain (assuming a labeled source validation set) for the ranking estimation, we might find them to be very different from the target rankings. This can often be attributed to the distribution difference. For example, one model may outperform another in a certain scenario, but their performances could be dis-similar or even reversed in a different scenario. Therefore, in order to obtain accurate model rankings on the target domain, target data distribution should be considered.

We explore proxy sets that meet these two requirements. **First**, we use existing datasets, where the labels of IDs are available. It could be the source, an arbitrary dataset other than the source or target, or a composite one. This allows us to conveniently compute model accuracies using its labels. **Second**, the proxy is close to the target distribution in terms of two distribution difference measurements: Fréchet Inception Distance (FID) [18] and feature variance gap [12, 23]. This is based on our observation that datasets more similar to the target domain (*i.e*., small FID and small variance gap) are more likely to form better proxies. This observation shares a similar spirit with some key findings in domain adaptation that reduced domain gap can benefit model training. Yet we derive it from a different viewpoint, *i.e*., the quality of a proxy set for performance ranking.

These two measurements are further investigated in a dataset search procedure. An image pool is collected from existing datasets and is partitioned into clusters. Images are sampled from each cluster with a probability proportional to the similarity (FID and variance gap) between the cluster and the target, forming the proxy. Overall, this paper contains the following main points.

- We study a new problem: ranking source model performance on an unlabeled target domain.
- We propose to use a labeled proxy that can give us a good estimation of model ranking. It is constructed via a search process such that the proxy data distribution is close to the target.
- Experiment verifies the efficacy of our method, and importantly, offers us insights into dataset similarities and model evaluation.

2. Related Work

Unsupervised domain adaptation (UDA) is a commonly used strategy to improve source model performance on the target domain where no labeling process is required. This objective can be implemented on the feature level [32], pixel level [63, 10], or based on pseudo labels [13, 61, 42]. While the goal of UDA is to learn an effective model for target scene, we aim to compare the performance of different models that are directly transferred to the target domain.

Predicting model generalization ability. Our work is also related to this area, where model generalization error on unseen images is estimated. Some work predicts the generalization gap using the training set and model parameters [2, 5, 21, 37]. For example, Corneanu *et al*. [5] use the persistent topology measures to predict the performance gap between training and testing errors. There are also works aiming to predict accuracy on unlabeled test samples based on the agreement score among predictions of several classifiers [34, 39, 38, 11]. Platanios *et al*. [38] use a probabilistic soft logic model to predict classifier errors. Recently, Deng *et al*. [9, 8] attempt to estimate classifier accuracy on various unlabeled test sets. Our work differs from the above works. We study a new problem: ranking different models in an unlabeled test domain.

Learning to simulate synthetic data. The objective of this area is to bridge the gap between the synthetic and real-world images by optimizing a set of parameters of a surrogate function that interfaces with a synthesizer [53, 51, 24]. It can be used to make customized data but needs to utilize specific engines and 3D models similar to the target object, which is not often accessible. Some recent works [27, 52] search a dataset from websites or data server for model training. Inspired by them, we attempt to search a proxy set with annotated data to rank models for the target domain.

Learning to rank has been studied in the fields of information retrieval [40, 48, 20], data mining [22, 3] and natural language processing [47, 17]. In general, given a query, the goal is to learn to rank data from the collection and return the top-ranked data. In computer vision, learning to rank is studied in content-based image retrieval [14, 20] and metric learning [16, 3, 31]. These works are concerned with learning metrics so that related samples are mapped to be closer to the query than unrelated ones. While they work on the datum (image) level, our paper deals with model ranking, which is on the model level.

3. Problem and Baseline

3.1. Problem Definition

Let $\{\mathbf{m}_i\}_{i=1}^{M}$ denote a set of M models trained on source domain (we call them source models). $\mathbf{T}$ is a set of unlabeled images collected from the target domain for performance ranking. In order to find the best model for direct application on the target domain, ideally, we should rank the model performances on $\mathbf{T}$ after labeling all images. However, given the high annotation costs, this becomes a less

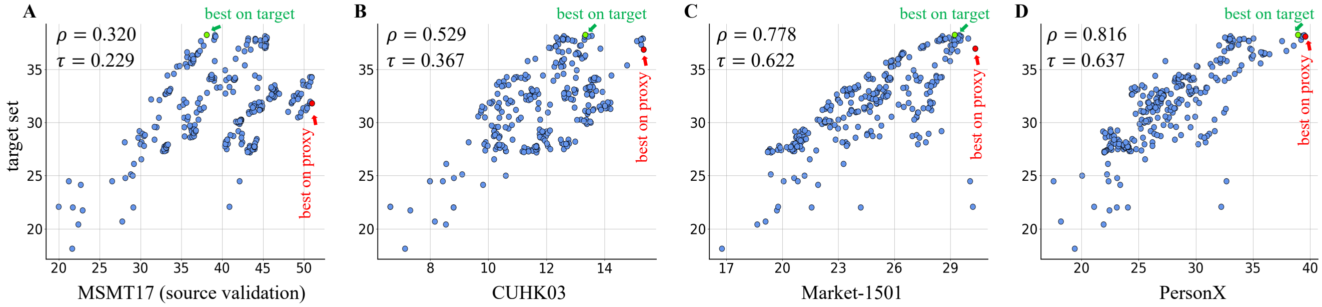


Figure 2: Correlation of model accuracies (mAP, %) on a given target set *vs.* proxy sets. Specifically, we train models with MSMT17 as the source domain and DukeMTMC-reID as the target. Four proxy choices are studied, *i.e.*, from left to right: **A.** source (MSMT17) validation set, **B.** CUHK03, **C.** Market-1501, and **D.** PersonX. For each model (blue circles), we evaluate its mAP scores on the target test set and the proxy set, which are then used to plot a 2-D correlation sub-graph. For each sub-graph, we use Spearman's Rank Correlation (ρ) [43] and Kendall's Rank Correlation (τ) [25] to measure the correlation between the two sets of mAP values. A higher absolute value of ρ (or τ) indicates a stronger correlation. Also shown are the best models on the target (green circle) and the proxy (red circle). We clearly see that source is a relatively poor proxy ($\rho = 0.320$, $\tau = 0.229$), while PersonX ($\rho = 0.816$, $\tau = 0.637$) and Market-1501 ($\rho = 0.778$, $\tau = 0.622$) are much better choices. Aside from the increased correlation (from left to right), we also find that the best models on the target and the proxy are getting closer. It indicates that the best model on the target is more likely to be the best one on the proxy (with a smaller error). All the above correlation coefficients have very high statistical significance due to their p-value < 0.001.

appealing choice. In this paper, we investigate whether it is possible to estimate the model accuracy ranking on the target domain (hereon denoted as the *ground truth* accuracy ranking) without labeling the images in $\mathbf{T}$.

Specifically, given an unlabeled target dataset $\mathbf{T}$ and models $\{\mathbf{m}_i\}_{i=1}^{M}$, we aim to find a labeled proxy set $\mathbf{P}$ whose performance ranking well represents the ground truth accuracy rankings on $\mathbf{T}$. We therefore formulate the goal of this problem as, find $\mathbf{P}$,

$$\text{s.t.} \quad rank\left(\{\mathbf{m}_i\}_{i=1}^{M}, \mathbf{P}\right) \rightarrow rank\left(\{\mathbf{m}_i\}_{i=1}^{M}, \mathbf{T}\right), \tag{1}$$

where $rank(\cdot,\cdot)$ denotes the performance ranking of certain models on a certain dataset.

For each proxy dataset, we use the model accuracies to create a performance ranking, and evaluate the quality of the proxy set as its ranking correlation with the ground truth accuracy ranking on target domain. To quantitatively evaluate the quality of a proxy, we use two rank correlation coefficients: Spearman's Rank Correlation ρ [43], and Kendall's Rank Correlation τ [25]. Both ρ and τ fall into the range of $[-1, 1]$, and a higher absolute value indicates a stronger correlation between rankings, *i.e.*, $rank\left(\{\mathbf{m}_i\}_{i=1}^{M}, \mathbf{P}\right)$ and $rank\left(\{\mathbf{m}_i\}_{i=1}^{M}, \mathbf{T}\right)$ in our problem. Accordingly, a lower absolute value of the correlation scores (with 0 being the lowest) indicates weak (or no) correlation.

3.2. Baseline: Individual Datasets as Proxy

Source validation set as proxy. We first study the relationship between model performance on the source (MSMT17 [50]) validation set (we use the test partition in the absence of validation) and target (DukeMTMC-reID [59, 41]) test set. Specifically, 280 re-ID models trained on MSMT17 are considered, which are shown by blue circles in Fig. 2. We plot these circles according to their accuracies on the proxy (MSMT17) and the target (DukeMTMC-reID). We only report mean average precision (mAP) here and omitted rank-1 precision since both metrics share a very similar trend. The rank correlation coefficients are: $\rho = 0.320$ and $\tau = 0.229$, indicating a weak rank correlation [1, 36] between proxy and target. As an intuitive understanding, the best model according to the proxy (source validation) has mAP 5.5% lower than the best one on the target set. We also witness similar phenomena using different source and target datasets. These results show that the source is a less appealing choice for proxy.

Other datasets as proxy. An annotated dataset from another domain can be a proxy, too. For example, when using MSMT17 and DukeMTMC-reID as source and target, respectively, a third dataset, Market-1501 [56] can serve as a target proxy. There are also other options readily available, such as PersonX [44] and RandPerson [49] (see Fig. 2 **B-D**). When compared with source validation set (MSMT17, Fig. 2 **A**), these datasets consistently achieve higher ranking correlations. For example, when using CUHK03, Market-1501, and PersonX as the proxy, we obtain Spearman's

ρ of 0.529, 0.778, and 0.816, and Kendall's τ of 0.367, 0.622, and 0.637, respectively. These numbers are consistently higher than those calculated from using the source as proxy. Meanwhile, the correlation coefficients suggest that the Market-1501 and PersonX are "moderate to strong" rank correlated with the target test set on model performance [1, 36]. When using different source and target combinations, we also find that these datasets from different domains form better proxies when compared to corresponding source validations. In this case, unless specified, we do not use the source validation as proxy in our further experiments. See Section 4.3 for more discussions.

4. Method: Search a Proxy Set

4.1. Motivation

When using different datasets (other than source and target) as proxy, we find some proxy sets to have higher quality (higher correlations with ground truth ranking) than others. Interested in what causes such proxy quality differences, we further investigate the potential reasons. Inspired by works in domain adaptation [10, 63], we examine the distribution difference between proxy set $\mathbf{P}$ and the target set $\mathbf{T}$. Specifically, we measure the distribution difference via two metrics, Fréchet Inception Distance (FID) [18] and feature variance gap [23]. $\mathrm{FID}(\mathbf{T}, \mathbf{P})$ measures the domain gap between the proxy set $\mathbf{P}$ and target set $\mathbf{T}$. On the other hand, feature variance gap measures how similar two data distributions are in terms of diversity and variation. We compute feature variance gap as the absolute difference between feature variance of $\mathbf{P}$ and $\mathbf{T}$,

$$\mathrm{V}_{\mathrm{gap}}(\mathbf{P}, \mathbf{T}) = |v(\mathbf{P}) - v(\mathbf{T})| \,, \tag{2}$$

where $v(\cdot)$ computes the variance. Notably, to calculate FID and $\mathrm{V}_{\mathrm{gap}}$, we use Inception-V3 [46] pre-trained on ImageNet to extract image features.

Using these two metrics, we further show the relationships between FID, $\mathrm{V}_{\mathrm{gap}}$, and proxy quality (correlations with the ground truth ranking). As shown in Fig. 3 **A**, we can spot an overall trend that smaller FID and $\mathrm{V}_{\mathrm{gap}}$ values often accompany higher proxy quality (ranking correlation coefficients). Moreover, as from Fig. 3 **B** and **C**, there also exist relatively strong correlations between either of the two metrics and the quality of proxy sets.

These experiments show that there might exist a proxy set of even better quality if it is composed of images whose distributions are more similar to the target (in terms of FID and $\mathrm{V}_{\mathrm{gap}}$). Motivated by this observation, we explore how to create a proxy set by searching images in next section.

4.2. The Search Algorithm

Given a data pool $\mathbf{D}$ that includes multiple datasets (other than the source and the target) and an unlabeled target set $\mathbf{T}$, it is our goal to sample data from $\mathbf{D}$ and compose a proxy set $\hat{\mathbf{P}}$ that has small $\mathrm{FID}(\mathbf{T}, \hat{\mathbf{P}})$ and $\mathrm{V}_{\mathrm{gap}}(\mathbf{T}, \hat{\mathbf{P}})$. Based on the findings in Section 4.1, we believe this can lead to a high quality proxy set for target domain. As shown in Fig. 4, we go through the following three steps in our proxy searching approach:

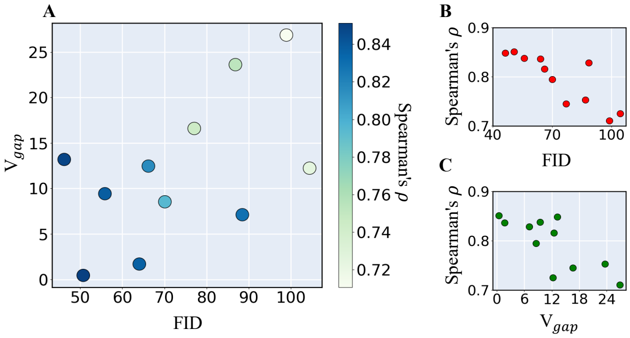


Figure 3: Relationships between FID, variance gap and the proxy set quality (evaluated using ρ). **A**: Influence of FID and variance gap on the proxy set quality. **B**: FID *vs.* proxy set quality. There is a very strong negative correlation (-0.88) with a very high statistical significance (p-value $<$ 0.001) between them. **C**: Variance gap *vs.* proxy set quality. They have a relatively strong negative correlation (-0.65) with a high statistical significance (p-value < 0.05). These three sub-figures verify that both FID and variance gap affect the proxy quality.

First, we cluster the data pool $\mathbf{D}$ into K subsets $\{\mathbf{S}_k\}_{k=1}^{K}$. To this end, we average all image features that belong to the same identity, and use this ID-averaged feature to represent all corresponding images. We then use k-means [29, 35] to cluster the ID-averaged features into K groups, and construct K subsets by including all images of the corresponding IDs that are in that group.

Second, we calculate the $\mathrm{FID}(\mathbf{T}, \mathbf{S}_k)$ and $\mathrm{V}_{\mathrm{gap}}(\mathbf{T}, \mathbf{S}_k)$ between each subset and the target set $\mathbf{T}$.

Lastly, we calculate a sampling score $\{w_k\}_{k=1}^{K}$ for each subset, and then assign a probabilistic weighting for each ID and sample ID form the data pool $\mathbf{D}$ based on the weightings. Specifically, we calculate the sampling score based on $\{\mathrm{FID}(\mathbf{T}, \mathbf{S}_k)\}_{k=1}^{K}$ and $\{\mathrm{V}_{\mathrm{gap}}(\mathbf{T}, \mathbf{S}_k)\}_{k=1}^{K}$. We take the negative of FID and variance gap values when calculating the sampling scores according to the negative correlations between their values and the proxy quality (see Fig. 3). The sampling score is written as,

$$\begin{aligned}\{w_k\}_{k=1}^{K} &= \lambda softmax(\{-\mathrm{FID}(\mathbf{T}, \mathbf{S}_k)\}_{k=1}^{K}) \\ &\quad + (1-\lambda) softmax(\{-\mathrm{V}_{\mathrm{gap}}(\mathbf{T}, \mathbf{S}_k)\}_{k=1}^{K}),\end{aligned} \tag{3}$$

where $softmax(\cdot)$ denotes the softmax function, and $\lambda \in [0, 1]$ is a weighting factor. $\lambda = 0$ or 1 represents only us-

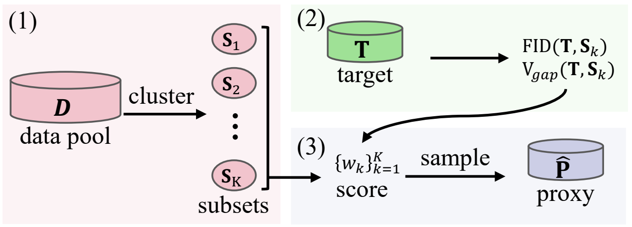


Figure 4: Three steps in our proxy searching process. First, we cluster the data pool into K clusters; second, we compute the distribution differences between clusters and the target; third, we calculate the sampling scores and compose a proxy set accordingly.

ing FID or variance gap to calculate sampling score. Based on the sampling scores of clusters, each ID of each cluster is assigned a probabilistic weighting $\frac{w_k}{|\mathbf{S}_k|}$. Here, $|\mathbf{S}_k|$ is the number of IDs of the cluster $\mathbf{S}_k$. The proxy set is constructed by sampling N examples from the data pool $\mathbf{D}$ at a rate according to probabilistic weightings of IDs.

In addition, if the camera annotation of the target set is available, we can further split the searching process into N steps for N cameras in the target, and then combine the final results as the proxy set $\hat{\mathbf{P}}$. Specifically, we repeat the aforementioned procedure N times (each camera once) to get N proxy sets. Notably, if one identity is sampled multiple times, we keep only one copy of the images of that identities in the final proxy set. We believe such a task-specific design would be helpful as it aligns with the multi-camera matching nature of re-ID problems [57].

4.3. Discussion

Why is the source often a poor proxy? Two reasons would explain the trend in Fig. 2 **A**. First, in our experiment, there is a non-negligible domain gap between the source (*e.g.*, MSMT17) and target (*e.g.*, DukeMTMC-reID). A strong model capable of distinguishing between fine-grained classes on the source may lose such discriminative ability on the target due to their distribution difference. Second, the models may be more or less overfitting the source. It is shown in [26] that when pretrained on ImageNet [7], models that have higher accuracy on ImageNet yields superior accuracy on other classification tasks after fine-tuning. While there seem to be fewer overfitting issues with ImageNet pretrained models, the relatively small source datasets (*e.g.*, MSMT17) in re-ID may cause overfitting, such that a good model on the source may be poor under a different environment.

Distribution difference measurements. This paper computes sampling weights based on both FID and variance gap (V_{gap}). Interestingly, the computation of FID also includes a diversity term between the two distributions, as it uses the covariance matrix. Nonetheless, in the experiment, we find only using either FID or variance gap leads to inferior results than them combined (see Fig. 6), which suggests both of them are indispensable. This suggests that the adopted feature variance gap could really benefit the searching process since it might provide a different angle for diversity difference measurement of data distribution.

Dataset	#ID	#images	#ID in **D**
MSMT17	4,101	126,441	3,060
DukeMTMC-reID	1,812	36,411	702
Market-1501	1,501	32,668	750
CUHK03	1,467	13,164	700
RAiD	43	1,264	43
iLIDS	119	476	119
PKU-Reid	114	1,824	114
PersonX	1,266	227,880	856
RandPerson	8,000	228,655	1,000
UnrealPerson	3,000	120,000	800

Table 1: Data pool composition. Seven real-world datasets and three synthetic datasets are considered. #ID in **D** means the number of identities used in the data pool.

Application scope. The proposed problem and solution allow us, for example, to select the most suitable model for new environments. As shown in Fig. 2 and later experiments, the selection process is fairly reliable. For applications like object recognition, we require that the proxy have the same categories as the target and the source so that the source models can be evaluated. The number of such classification datasets is currently limited (see supplementary material). For applications like person re-identification, we can leverage the abundant datasets available for proxy construction, because it is feasible to evaluate source models on proxy sets with completely different categories. In addition, since the proposed task is new and challenging, we currently focus on models that are directly applied to target data to avoid complicating the problem. As such, we do not consider UDA models [10, 62] that include the target samples in training, but they are worth studying, and we will investigate these models in future works.

5. Experiment

5.1. Experimental Details

Databases. This paper uses a wide range of real-world and synthetic person re-ID datasets. Real-world ones include Market-1501 [56], DukeMTMC-reID [59, 41], MSMT17 [50], CUHK03 [28], RAiD [6], PKU-Reid [33] and iLIDS [58]. Synthetic datasets used are PersonX [44], Randperson [49] and UnrealPerson [54]. Some important details of these datasets are shown in Table 1. From these

Source	Target		Individual Dataset							Other Method				Ours	
			CUHK03	Duke	Market	MSMT17	RandPerson	PersonX	UnrealPerson	Random	Attr. descent [53]	StarGAN [4]	pseudo-label [13]	w/o cam	w/ cam
MSMT17	Duke	ρ	0.529	-	0.778	0.320	0.775	0.816	0.837	0.725	0.756	0.700	0.789	**0.858**	**0.882**
		τ	0.367	-	0.622	0.229	0.602	0.637	0.655	0.537	0.569	0.518	0.625	**0.713**	**0.725**
	Market	ρ	0.180	0.778	-	0.335	0.803	0.874	0.854	0.643	0.638	0.811	0.823	**0.884**	**0.912**
		τ	0.126	0.622	-	0.245	0.616	0.690	0.664	0.507	0.467	0.615	0.648	**0.715**	**0.753**
Market	Duke	ρ	0.374	-	-0.119	0.932	0.905	0.805	0.933	0.713	0.740	0.848	0.899	**0.939**	**0.950**
		τ	0.260	-	-0.048	0.790	0.774	0.626	0.808	0.538	0.551	0.662	0.742	**0.810**	**0.824**
	MSMT17	ρ	0.331	0.932	-0.173	-	0.876	0.727	0.941	0.711	0.790	0.807	0.846	**0.949**	**0.958**
		τ	0.254	0.790	-0.092	-	0.705	0.548	0.817	0.553	0.612	0.624	0.698	**0.822**	**0.829**

Table 2: Comparison of different proxy sets on different source-target configurations. We search proxy sets ("w/ cam" and "w/o cam") under different availability of the target domain camera annotation.

datasets, we can select one as the source and another one as the target. The rest will form the data pool (Section 4.2). When creating the data pool, we only use a portion of identities and their corresponding images. This limits the problem size in our searching process, while preventing dominating the data pool with images in a few datasets. Overall, we consider a total number of 8,144 identities in our data pool.

Models to be ranked. We consider 28 representative baselines and approaches in person re-ID, including ID-discriminative embedding (IDE) [55], part-based convolution baseline (PCB) [45], and record 10 different versions of each model during their training procedure. For hyper-parameters, we follow their original settings (see supplementary material for more details of the models). In total, we have 280 models, *i.e.*, $N = 280$ in $\{\mathbf{m}_i\}_{i=1}^M$. All the models are trained from scratch on the source domain.

Searched proxies. In this work, we choose the hyper-parameters for proxy set searching as $\lambda = 0.6$ for the weighting factor and $K = 20$ for the cluster number. The number of identities of the searched proxy sets is set to 500 (see Section 5.3). For more details on the searched proxy sets, please refer to the supplementary materials. We perform search with one RTX-2080TI GPU and a 16-core AMD Threadripper CPU @ 3.5Ghz.

5.2. Evaluation of the Proposed Method

In Table 2, we compare the quality of our searched proxy with alternative proxy choices, including individual labeled datasets (datasets in Table 1), engine-based synthetic images [53], GAN-based generated images [4, 60], pseudo labels on the target validation [13], and a random sample from all the individual labeled dataset (denoted as "Random" in Table 2). We have the following observations.

Effectiveness of the searched proxy over individual datasets. Our main observation is that the searched proxy is very competitive to individual datasets as proxy. When MSMT17 is used as source and DukeMTMC-reID is used as target, the searched proxy ("w/o cam" in Table 2) achieves very good ranking correlations ($\rho = 0.858$ and $\tau = 0.713$), outperforming both individual datasets and other methods by at least $+0.021$ of ρ and $+0.060$ of τ. Similar results can also be found when Market-1501 is selected as target, where the proposed searching method achieves $\rho = 0.884$ and $\tau = 0.715$, outperforming every alternative by at least $+0.010$ of ρ and $+0.024$ of τ.

Besides, the search method is better than a random combination of individual datasets. As shown in Table 2, "Random" might lag behind some of the better performing individual datasets by up to -0.343 of ρ and -0.300 of τ. Our searching method, on the other hand, constantly outperforms this random combination by at least $+0.133$ of ρ and $+0.176$ of τ, while achieving competitive or even better results to the individual datasets.

Utilizing camera annotations of the target domain yields the best performance of proxy. For example, when MSMT17 and DukeMTMC-reID is used as source and target, repetitively, using camera information in our searching approach further improves the overall proxy quality (ranking correlations) to $\rho = 0.882$ and $\tau = 0.735$. This shows the advantage of a task-centered searching method design, which aligns well with the cross-camera matching the nature of the person re-ID problem.

Study of the composition of proxy set. In Fig. 5, we examine the composition of the searched proxy set. With the MSMT17 dataset as source and the DukeMTMC-reID dataset as target, the searching process ends up using more (63.7%) real-world data compared to synthetic ones, since the real-world data might look more similar to the real-world target of DukeMTMC-reID. Overall, our searched proxy sampled IDs and images look similar to those in the target domain in terms of lighting and colors.

Comparison with generated images and pseudo label methods. Methods designed for generating or synthesizing training data are found less efficient as proxy sets. For example, engine-based synthesis [53] achieves a ρ of 0.756 and τ of 0.569 for rank correlations (MSMT17 as source

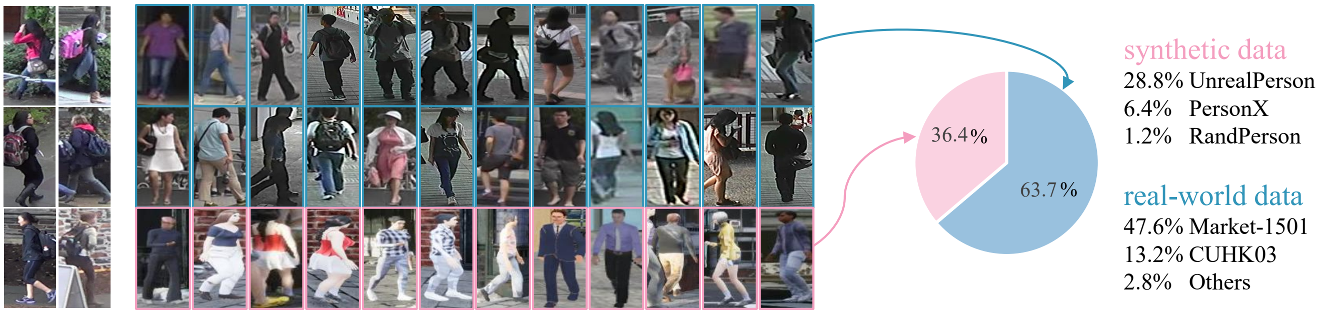


Figure 5: Image samples and composition statistics of the searched proxy (MSMT17 as source and Duke as target). **Left**: unlabeled target; **Middle**: searched proxy; **Right**: composition statistics of the searched proxy. We observe that the searched proxy overall displays similar lighting and color schemes compared with the target.

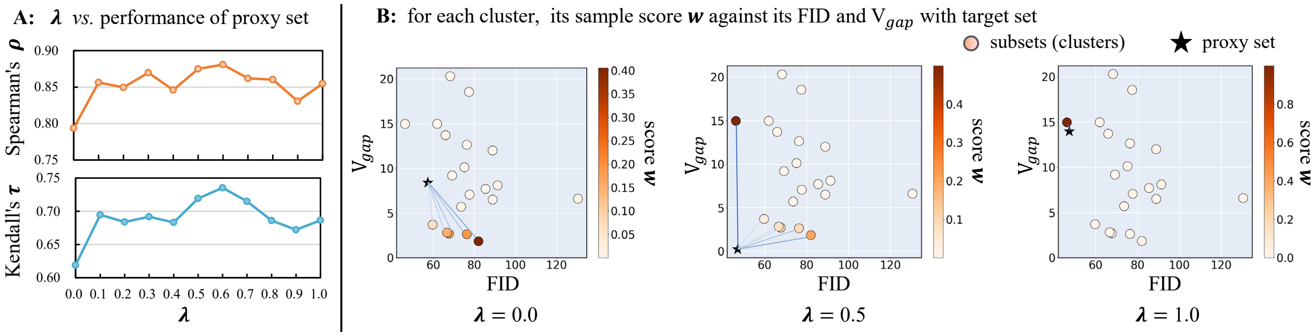


Figure 6: The impact of weighting factor λ in Eq. 3. **A**: searched proxy quality under different λ values. Then overall sample score w_k only considers variance gap when $\lambda = 0$, and only considers FID when $\lambda = 1$. **B**: sampling scores for each cluster and their contribution to the searched proxy set $\star$ under different λ values. Deep colors denote higher sampling scores for the clusters (dots) and higher contributions to the searched proxy (lines). Here, the cluster number K is set to 20. MSMT17 and DukeMTMC-reID are used as source and target, respectively.

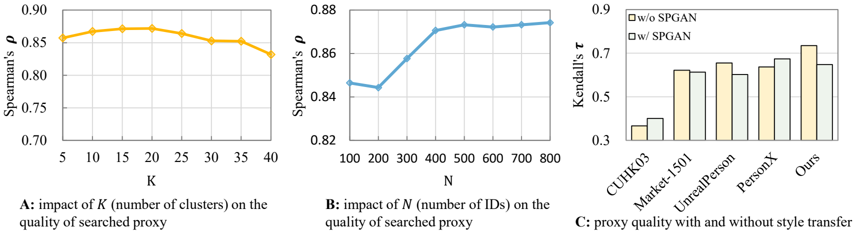
and DukeMTMC-reID as target). This trails behind not only our searched proxy set but also some of the better-performing individual datasets as proxies. As for GAN-based method [60] and pseudo label method [13], both of them create image-label pairs using a network, which might introduce inaccurate labels (network decided image label pairs are less reliable than annotated ones). For this reason, the rank correlations of these methods are also sub-optimal.

Computational cost in searching a proxy. As shown in Fig. 4, there are three steps involved in our proxy searching process. When the MSMT17 is used as source and Market-1501 is used as target, feature extraction and clustering cost about 200 seconds. Then, it takes about 188 seconds to calculate the FIDs and variance gaps. Time for the image sampling process can be neglected. So our algorithm consumes about 400 seconds in total. When camera annotations are available, the searching process has no additional cost in the first step. In fact, feature extraction and clustering results can be reused. The overall searching process takes about 1772 seconds for all 6 cameras in the target set.

5.3. Parameter Analysis

Weighting factor λ for sampling score. λ encodes the trade-off between FID and $\mathrm{V}_{\mathrm{gap}}$ when calculating the sampling score. As shown in Fig. 6 **A**, setting λ to 0.6 (as in our current design) gives the best overall quality of the proxy set (highest Spearman's and Kendall's correlation coefficient). Using only either FID or variance gap (setting λ to 1 or 0) leads to a quality drop of the searched proxy set. Interestingly, only using FID provides slightly better results compared to only using variance gap. One possible reason is that the FID also considers covariance during computation, which might have a slight overlap with the variance gap. In this case, the variance gap is also reduced when only minimizing FID, which might provide a slight edge to the variant that only uses FID over only using variance gap.

For more intuitive understandings, we find that only considering variance gap ($\lambda = 0$) creates a proxy set that has



A: impact of K (number of clusters) on the quality of searched proxy

B: impact of N (number of IDs) on the quality of searched proxy

C: proxy quality with and without style transfer

Figure 7: The impact of **A:** number of clutters K, **B:** number of IDs N and **C:** style transfer on the quality of proxy set. (MSMT17 and DukeMTMC-reID are used as source and target sets, respectively.)

Training Data	R1	R5	mAP
MSMT17	58.0	71.6	36.5
Market-1501	42.1	56.1	23.9
Synthetic data [53]	21.2	39.7	13.5
Pseudo-label [13]	67.5	80.5	50.4
Ours (searched proxy)	47.9	63.2	27.9

Table 3: Performance on DukeMTMC-reID using different training sets. Here, Rk means rank-k accuracy.

an even higher variance gap compared to the clusters that majorly contribute to the proxy (Fig. 6 **B** $\lambda = 0.0$). Only considering FID ($\lambda = 1$) samples samples mainly from only one cluster, and results in a proxy that is very similar in terms of FID (Fig. 6 **B** $\lambda = 1.0$). When jointly considering both FID and variance gap ($\lambda = 0.5$), the resulting proxy has an even lower FID and variance gap compared to the clusters that contribute to it, further indicating the effectiveness of the proposed method (Fig. 6 $\lambda = 0.5$).

Numbers of clusters K and IDs N of Proxy Set. The proposed method clusters the data pool into K groups based on their ID-averaged features and samples N identities to build the proxy set. Here, we further investigate the influence of the cluster number K and the identity number N on the searched proxy quality. As shown in Fig. 7 **A**-**B**, we find that 1) either a too small or too large K can lead to slightly poor proxy quality (here, N is set to 400) and 2) when N gradually becomes larger, the result tends to be stable, so we set the cluster number K to an intermediate value 20, and ID number to 500, to provide relatively good results.

5.4. Further Understandings

Can we improve the proxies by style transfer? Pixel-level alignment [10, 50, 30] is commonly used to reduce the domain gap by transferring the image style of one domain into that of the other. For different proxy sets (individual datasets or searched ones), we employ SPGAN [10] to translate them into the style of the target domain. We present the correlation coefficients in Fig. 7 **C**. Taking the DukeMTMC-reID dataset as target data, we transfer several proxy sets to DukeMTMC-reID style through SPGAN [10] and use the style-transferred proxy sets to ranking models. It is found that SPGAN cannot bring consistent improvements to the model ranking proxies. Despite these mixed results, we note that the best performance is still held by the searched proxy (without style transfer).

Can we train re-ID models on proxy sets for a certain target? In Table 3, we find that directly applying re-ID models (IDE [55]) trained on the searched proxy set does not lead to competitive performance on the target domain, despite the fact that the proxy set is searched for that target specifically. In comparison, pseudo-label [13], a method that underperforms our method in building proxy sets for model ranking, actually achieves the best result in building training sets for domain adaptation models. This suggests that our problem is quite different from training data search, although they might appear similar at first glance.

Effectiveness of MMD in replacing FID. We replace FID with MMD in Eq. 3, which is another way to calculate the distribution difference between two datasets. We use MSMT17 and DukeMTMC-reID as source and target, respectively. We observe that replacing FID with MMD yields -0.0056 Spearman's ρ and -0.0161 Kendall's τ, suggesting that MMD has a similar effect with FID.

6. Conclusion

This paper studies an important and practical problem: when some source models are directly applied to an unseen target domain, can we rank their performance without having to know the ground-truth labels for (a representative subset of) the target domain? We answer this question by using a so-called *target proxy* for un-referenced model evaluation. We first propose a number of baseline approaches, *i.e.*, using the source data as proxy, or using various cross-domain datasets as proxy. We analyze the underlying reasons for the (in)effectiveness of such baselines and identify that the domain gap and diversity gap are two important factors affecting the quality of a proxy. We therefore adopt a search strategy that uses a weighted combination of these two metrics as objective. Experiments on public person re-ID datasets validate our strategy and let us gain rich insights into dataset similarity and model generalization.

Acknowledgement

This work was supported in part by the ARC Discovery Early Career Researcher Award (DE200101283), the ARC Discovery Project (DP210102801). Hongdong Li's research is also funded in part by an Australian Research Council Discovery grant (DP 190102261).

References

[1] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3):91–93, 2018. 3, 4

[2] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296*, 2018. 2

[3] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *CVPR*, pages 1861–1870, 2019. 2

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 6

[5] Ciprian A Corneanu, Sergio Escalera, and Aleix M Martinez. Computing the testing error without a testing set. In *CVPR*, pages 2677–2685, 2020. 2

[6] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In *European conference on computer vision*, pages 330–345, 2014. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. 5

[8] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *ICML*, 2021. 2

[9] Weijian Deng and Liang Zheng. Are labels necessary for classifier accuracy evaluation? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 2

[10] Weijian Deng, Liang Zheng, Qixiang Ye, Yi Yang, and Jianbin Jiao. Similarity-preserving image-image domain adaptation for person re-identification. *arXiv preprint arXiv:1811.10551*, 2018. 2, 4, 5, 8

[11] Pinar Donmez, Guy Lebanon, and Krishnakumar Balasubramanian. Unsupervised supervised learning i: Estimating classification and regression errors without labels. *Journal of Machine Learning Research*, 11(4), 2010. 2

[12] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine-grained classification. In *NIPS*, 2018. 2

[13] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(4):1–18, 2018. 2, 6, 7, 8

[14] Songhe Feng, Zheyun Feng, and Rong Jin. Learning to rank image tags with limited training examples. *IEEE Transactions on Image Processing*, 24(4):1223–1234, 2015. 2

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1

[16] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision*, 2018. 2

[17] Yoav Goldberg. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309, 2017. 2

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 2017. 2, 4

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998, 2018. 1

[20] Yang Hu, Mingjing Li, and Nenghai Yu. Multiple-instance ranking: Learning to rank images for image retrieval. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 2

[21] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. *arXiv preprint arXiv:1810.00113*, 2018. 2

[22] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789, 2017. 2

[23] Dag Jonsson. Some limit theorems for the eigenvalues of a sample covariance matrix. *Journal of Multivariate Analysis*, 12(1):1–38, 1982. 2, 4

[24] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *ICCV*, pages 4551–4560, 2019. 2

[25] Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. 3

[26] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, pages 2661–2671, 2019. 5

[27] Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017. 2

[28] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014. 5

[29] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003. 4

[30] Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *CVPR*, 2019. 8

[31] Tie-Yan Liu. Learning to rank for information retrieval. 2011. 2

[32] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of International Conference on Machine Learning*, pages 97–105, 2015. 1, 2

[33] Liqian Ma, Hong Liu, Liang Hu, Can Wang, and Qianru Sun. Orientation driven bag of appearances for person re-identification. *arXiv preprint arXiv:1605.02464*, 2016. 5

[34] Omid Madani, David Pennock, and Gary Flake. Co-validation: Using model disagreement on unlabeled data to validate classification algorithms. *Advances in neural information processing systems*, 17:873–880, 2004. 2

[35] Mikko I Malinen and Pasi Fränti. Balanced k-means for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 32–41, 2014. 4

[36] Ellen Marshall and Elizabeth Boggis. The statistics tutor's quick guide to commonly used statistical tests. *Statstutor Community Project*, pages 1–57, 2016. 3, 4

[37] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in neural information processing systems*, pages 5947–5956, 2017. 2

[38] Emmanouil Platanios, Hoifung Poon, Tom M Mitchell, and Eric J Horvitz. Estimating accuracy from unlabeled data: A probabilistic logic approach. In *Advances in Neural Information Processing Systems*, pages 4361–4370, 2017. 2

[39] Emmanouil Antonios Platanios, Avinava Dubey, and Tom Mitchell. Estimating accuracy from unlabeled data: A bayesian approach. In *International Conference on Machine Learning*, pages 1416–1425, 2016. 2

[40] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010. 2

[41] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016. 3, 5

[42] Liangchen Song, Yonghao Xu, Lefei Zhang, Bo Du, Qian Zhang, and Xinggang Wang. Learning from synthetic images via active pseudo-labeling. *IEEE Transactions on Image Processing*, 2020. 2

[43] Charles Spearman. The proof and measurement of association between two things. 1961. 3

[44] Xiaoxiao Sun and Liang Zheng. Dissecting person re-identification from the viewpoint of viewpoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3, 5

[45] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision*, pages 480–496, 2018. 6

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 4

[47] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. Learning to rank question answer pairs with holographic dual lstm architecture. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 695–704, 2017. 2

[48] Ji Wan, Pengcheng Wu, Steven CH Hoi, Peilin Zhao, Xingyu Gao, Dayong Wang, Yongdong Zhang, and Jintao Li. Online learning to rank for content-based image retrieval. In *IJCAI*, 2015. 2

[49] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. *arXiv preprint arXiv:2006.12774*, 2020. 3, 5

[50] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3, 5, 8

[51] Zhenfeng Xue, Weijie Mao, and Liang Zheng. Learning to simulate complex scenes. *arXiv preprint arXiv:2006.14611*, 2020. 2

[52] Xi Yan, David Acuna, and Sanja Fidler. Neural data server: A large-scale search engine for transfer learning data. In *CVPR*, pages 3893–3902, 2020. 2

[53] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 6, 8

[54] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. *arXiv preprint arXiv:2012.04268*, 2020. 5

[55] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 868–884, 2016. 6, 8

[56] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 3, 5

[57] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 5

[58] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, 2009. 5

[59] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 3, 5

[60] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *Proceedings of the European Conference on Computer Vision*, 2018. 6, 7

[61] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[62] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Learning to adapt invariance in memory for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2, 4