BENCHMARKING PERSON RE-IDENTIFICATION TRAINING DATASETS AND APPROACHES FOR PRACTICAL REAL-WORLD IMPLEMENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Person Re-Identification (Re-ID) is receiving a lot of attention recently. Large datasets containing labeled images of various individuals have been released, allowing researchers to develop and test many successful approaches. However, when such Re-ID models are deployed in a new city or environment, the task of searching people within a network of security cameras is likely to face an important domain shift, thus resulting in decreased performance. Indeed, while most public datasets were collected in a limited geographic area, images from a new city present different features (e.g., people's ethnicities and clothing style, weather, architecture, etc.). In addition, the whole frames of the video streams must be converted into cropped images of people using pedestrian detection models, which behave differently from the human annotators who created the dataset used for training. To better understand the extent of this issue, this paper introduces a complete methodology to evaluate Re-ID approaches and training datasets with respect to their suitability for deployment for live operations. This method is used to benchmark four Re-ID approaches and three datasets, providing insight and guidelines that can help designing better Re-ID pipelines in the future.

1 INTRODUCTION

As many cameras are being deployed in public places (e.g., airports, malls, parks), real-time monitoring of the video streams by security agents becomes impractical. Automated video processing appears as a promising solution to analyze the whole network in real time and select only relevant sequences for verification by human operators. This paper deals with person Re-Identification (Re-ID), a computer vision problem that aims at finding an individual in a network of non-overlapping cameras (Bedagkar-Gala & Shah, 2014). It has diverse potential security applications, such as suspect searching (Liao et al., 2014), identifying owners of abandoned luggage (Altunay et al., 2018), or recovering missing children (Deb et al., 2021), among others.

In the literature, the problem of Re-ID is studied under different settings depending on the application context (see Section 2.1). On the one hand, the most studied Re-ID paradigm, which we refer to as *standard Re-ID*, tries to find images representing the query person within a gallery of pre-cropped images of persons, containing at least one correct match (Lavi et al., 2020). On the other hand, Sumari et al. (2020) recently introduced a setting considering specifically the constraints related to implementing Re-ID for use during live operations. We call it *live Re-ID*, and a first contribution of this work is to formalize the definition and constraints associated to this setting. We also extend the live Re-ID evaluation metrics proposed by Sumari et al. (2020) in order to facilitate interpretation.

Standard Re-ID is not the best suited paradigm for practical implementations, as it does not consider the influence of potential domain shift due to pedestrian detection errors or to deployment in a city with different characteristics than the training dataset. Indeed, in their experiments, Sumari et al. (2020) showed that training a successful Re-ID model with respect to standard Re-ID metrics does not guarantee good performance when evaluated on specific live Re-ID metrics. Nevertheless, most publicly available large scale datasets for Re-ID focus on the standard Re-ID setting, and many successful approaches have been developed for this specific purpose. For this reason, we believe

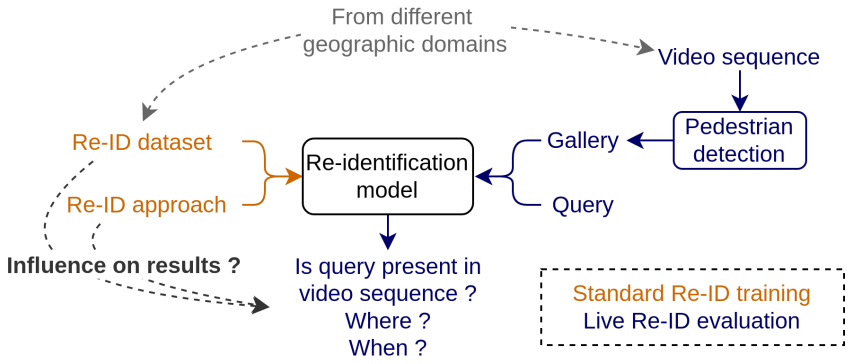


Figure 1: **Conceptual overview**. Visual representation of the objectives of our benchmark study. This work aims at evaluating how different standard Re-ID approaches and training datasets behave for practical deployment in new environments, i.e., live Re-ID setting.

that it is essential to study if these datasets and approaches can be used to implement and deploy practical applications in different contexts. More specifically, the objective of this paper is to answer the following questions:

1. Which standard Re-ID approaches can be successfully deployed for practical implementations in the live Re-ID setting?
2. Which standard Re-ID dataset is better suited to train standard Re-ID models for the live Re-ID setting?
3. Do different Re-ID approaches have different optimal datasets for deployment?
4. Can we use a simple cross-dataset evaluation methodology to assess the deployability of a given approach-dataset pair?

To answer these questions, we conducted a study using three standard Re-ID datasets and four recent standard Re-ID approaches. For each approach-dataset pair, the Re-ID model obtained was evaluated against the other two datasets and against a fourth dataset configured for the live Re-ID setting. A conceptual overview of the objectives of our study is represented in Figure 1. Xiao et al. (2017) showed that considering pedestrian detection and Re-ID separately is not as good as end-to-end approaches for person search, i.e., galleries of whole scene images. However, our results show that this two step approach can perform well on the live Re-ID setting. In addition, we believe that the results from our study can be very useful to pre-train successful initial live Re-ID models and to guide the development of more complex end-to-end architectures for live Re-ID.

This paper is organized as follows: Section 2 discusses the relevant related literature. The methodology for the proposed benchmark experiments is detailed in Section 3. The results are presented in Section 4 and discussed in Section 5. Finally, Section 6 presents our conclusions and potential future work.

2 Related work

A complete literature review about Re-ID approaches is not the purpose of this paper. Instead, we present clear definitions of the different existing Re-ID settings in Section 2.1. Previous benchmark studies about Re-ID are also discussed in Section 2.2.

2.1 Person re-identification settings

The field of Re-ID was first formalized by Gheissari et al. (2006), it consists in retrieving instances of a given individual, called the *query person*, within a complex set of multimedia content called the *gallery*. The different settings presented here are defined by how they represent the query person, the format of the gallery items, the constraints on the gallery content, the boundaries of the Re-ID system, and the constraints imposed on the evaluation methodology.

2.1.1 POPULAR SETTINGS

Standard Re-ID In the *standard Re-ID* setting, both the *query image* (representing the query person), and all items in the gallery are well-cropped images representing the entire body of a person. It is sometimes called closed-set Re-ID as it assumes that the query person has at least one representative in the gallery. According to the statistics in Papers with Code (PwC), it is the most studied Re-ID setting by a large margin, in terms of number of papers, datasets and benchmarks published. Some standard Re-ID datasets and successful methods are used for our benchmark study and presented in Section 3. For a more complete overview of standard Re-ID approaches, we refer the reader to the surveys by Lavi et al. (2020) and Ye et al. (2021).

Person search The *person search* setting was introduced by Xu et al. (2014). It consists in replacing the gallery items by whole scene images (Xiao et al., 2017). In other words, a person search model must return not only the index of the gallery image where the query is present, but also its location in terms of Bounding Box (BB) coordinates. A survey about person search approaches was proposed by Islam (2020).

Open-set Re-ID The *open-set Re-ID* setting was first defined by Liao et al. (2014). It differs from standard Re-Id in that there is no guarantee that the query person is represented in the gallery, i.e., an open-set Re-ID model should be able to answer whether the gallery contains the query. The reader can refer to the survey by Leng et al. (2019) for an overview of recent open-set Re-ID approaches.

Video-based Re-ID The *video-based Re-ID* setting was first studied by Wang et al. (2014). In this setting, all images (query and gallery) are replaced by image sequences extracted from consecutive frames of a video. Sequences are composed of well-cropped entire body images representing the same person. A complete review of video-based Re-ID was proposed by Ye et al. (2021).

Others For completeness, we mention the existence of other Re-ID variants in the literature, namely unsupervised Re-ID (Yang et al., 2021), semi-supervised Re-ID (Moskvyak et al., 2021), human-in-the-loop Re-ID (Wang et al., 2016), or federated Re-ID (Zhuang et al., 2020). However, their specificities lie in how Re-ID models are trained while the other settings above focus on constraints at inference time. For this reason, these Re-ID paradigms are not presented further here.

2.1.2 LIVE RE-ID SETTING

In this section, we clearly define and formalize the *live Re-ID* setting, which is inspired by the work of Sumari et al. (2020). It takes into account all relevant aspects for deploying Re-ID models in practical real-world applications.

When looking for a query person during live operations, whole scene videos need to be processed in near real time, hence the galleries for live Re-ID are composed of the consecutive **whole scene frames** from **short video sequences**. The live Re-ID context is also highly **open-set** as the probability to have the query in a short video sequence from a given camera is low. Hence, this setting combines elements from several of the Re-ID settings mentioned above.

Another key characteristic of live Re-ID is that the training context is different from the deployment context. Indeed, building new specialized datasets for deployment in every shopping mall or small city is unrealistic from the perspective of future advances in the field. This highlights the importance of studying **cross domain** transfer of Re-ID, which was first discussed and highlighted by Luo et al. (2020).

Finally, this setting also takes into account that Re-ID model predictions need to be **processed by a human security agent**, who takes the final decision and trigger appropriate actions. This way, very high rank-1 accuracy is not mandatory for live Re-ID, as the operator can find the query in later ranks. On the other hand, false alarm rates must be kept low to avoid overloading the human operators, who have limited processing capacity. To evaluate these two objectives, Sumari et al. (2020) introduced two evaluation metrics representing both dimensions of the problem (see Section 3.3.3). The experiments conducted in this paper aim at studying the transferability of standard Re-ID approaches and datasets for deployment in the live Re-ID setting.

2.2 PERSON RE-IDENTIFICATION BENCHMARKS

Most recent research dealing with Re-ID present comparative evaluation of different approaches. While listing all these papers is out of the scope of this work, this section presents several benchmark studies considering different Re-ID settings or specific aspects of the Re-ID pipeline.

A large scale benchmark experiment was conducted by Gou et al. (2018) to compare various approaches for standard Re-ID and video-based Re-ID. By evaluating more than 30 approaches on 16 public datasets, they produced the largest Re-ID benchmark to date. In addition, they built a new dataset to represent several constraints relevant for real-world implementations, such as pedestrian detection errors and illumination variations, among others. However, they do not consider cross domain performance and all evaluations are conducted in the closed-set setting, which are major limitations regarding future deployments. In addition, a smaller systematic evaluation of video-based Re-ID approaches was proposed by Zheng et al. (2016).

Another extensive set of experiments was conducted by Zheng et al. (2017) to evaluate different pedestrian detection models on a two-step person search pipeline. They demonstrated that the best performing models on standard object detection metrics are not necessarily the best suited for Re-ID from whole scene frames. In addition, a first benchmark regarding cross-domain transfer of Re-ID approaches was proposed by He et al. (2020). Their experiments consisted in training an approach on one standard Re-ID dataset and evaluating on another. Finally, on another note, Zhuang et al. (2020) compared different approaches for federated Re-ID, i.e. learning Re-ID across decentralized clients to preserve privacy.

The studies presented above have brought valuable insights to the Re-ID community. However, none of them allows to assess the performance of a Re-ID model against all the challenges involved during deployment in a new environment for practical use in security applications. This paper contributes to bridging this gap by conducting experiments within the live Re-ID setting, which was designed to take into account all these challenges. In particular, we consider the influence of different standard Re-ID approaches and training datasets on live Re-ID results.

3 BENCHMARK METHODOLOGY

The objective of this paper is to study if different standard Re-ID approaches and training datasets can be used to build efficient live Re-ID pipelines, ready for practical deployment. This section presents the different components of the proposed benchmarking evaluation, i.e., the datasets and approaches compared, metrics used, and experiments conducted.

3.1 DATASETS

In our experiments, we used three public datasets to train and evaluate standard Re-ID models, and a live Re-ID dataset to evaluate the trained Re-ID models within the context of live operations.

3.1.1 STANDARD RE-ID DATASETS

This section presents briefly the standard Re-ID datasets used in this study. Table 1 summarizes relevant statistics regarding the datasets. **CUHK03** was built by Li et al. (2014), using video footage collected at the campus of the Chinese University of Hong Kong. In our work, we used the manually labeled version of the BB. **DukeMTMC** was released by Ristani et al. (2016). It was collected at the Duke University campus, Durham, North Carolina, USA. The BB in DukeMTMC are hand drawn. **Market-1501** was released by Zheng et al. (2015b). It was collected at a supermarket in Tsinghua University, Beijing, China. The cropped images are detected automatically using a Deformable Part Model (DPM) (Felzenszwalb et al., 2009), which outputs are filtered manually to keep only good BB representing human bodies. This automated way of extracting BB is closer to realistic settings, which might improve live Re-ID results for models trained on Market-1501.

3.1.2 LIVE RE-ID DATASET

To evaluate the different standard Re-ID models for the live Re-ID setting, we used the same dataset as Sumari et al. (2020), which we call **m-PRID**. It is a modified version of PRID-2011 (Hirzer et al.,

Table 1: **Standard Re-ID training datasets**. Characteristics of the standard Re-ID training datasets evaluated in this benchmark paper.

Dataset	# Cameras	Split	Input type	# IDs	# Images
CUHK03	2	Train	–	767	7368
		Test	Query	700	1400
			Gallery	700	5328
DukeMTMC	8	Train	–	702	16522
		Test	Query	702	2228
			Gallery	1110	17661
Market-1501	6	Train	–	751	12936
		Test	Query	750	3368
			Gallery	751	15913

2011), built from the raw video footage and the original annotations that were used to create the official curated version of PRID-2011[1]. The videos were collected from two non-overlapping cameras (A and B), located in Graz, Austria. This way, compared to the training datasets above, evaluation on m-PRID represents a geographic domain shift. In total PRID-2011 contains 385 different identities for camera A and 749 for camera B, of which 200 identities appear in both cameras. The m-PRID dataset is composed of several two minutes videos (30 from A and 33 from B). For each short video sample, a ground truth file gather information about each person it contains (identifier, frames where it appears, bounding box coordinates). For evaluation, a total of 73 queries are considered.

3.2 RE-ID APPROACHES EVALUATED

This work studies the performance of four successful standard Re-ID approaches. To complement previous benchmark studies (Section 2.2), only very recent approaches are selected for this work.

BoT The *Bag of Tricks* approach proposed by Luo et al. (2019) resulted from the observation that most improvements for Re-ID baselines come from neural network training tricks rather than Re-ID approaches themselves. As a result, they came up with a simple recipe to successfully train standard Re-Id models on top of a ResNet-50 backbone (He et al., 2015). In particular: 1. the model is pretrained on ImageNet, 2. the dimension of the fully connected layer is set to the number of training identities, 3. the batch size is set to 64, 4. images are resized to 256 X 128, 5. both triplet loss and cross entropy loss are used, and 6. Adam is adopted for optimization.

SBS The approach named *Strong Baseline and Batch Normalization Neck* also comes from Luo et al. (2020). It extended BoT by adding the following tricks: 1. a warmup strategy (Fan et al., 2019) is applied to bootstrap the network, 2. random erasing augmentation (Zhong et al., 2017) is used to account for potential occlusion of the person, 3. label smoothing (Szegedy et al., 2015) is used to reduce overfitting, 4. the last stride of the ResNet-50 backbone is set to 1 to increase spatial resolution, 5. batch normalization layers are added, and 6. a new center loss is introduced to account for the clustering effect of tracking.

AGW *Attention Generalized mean pooling with Weighted triplet loss* was developed by Ye et al. (2021). It was also designed on top of BoT with three major improved components: 1. a powerful non-local attention block is developed to mix the part and global attention features, 2. a learnable pooling layer replaces max and average pooling to better capture the domain-specific discriminative features, 3. the use of weighted regularization triplet loss inherits the advantages of relative distance optimization between positive and negative pairs without introducing additional parameters.

MGN The *Multiple Granularities Network* was designed by Wang et al. (2018), as a new feature learning strategy. It combines local and global information in different image granularities. The

[1]We kindly thank the authors of the original PRID-2011 paper for their responsiveness and cooperation.

feature maps are split in N horizontal stripes to learn local feature representations. Then, the ResNet-50 backbone is divided into a multi-branch network after the fourth residual stage.The global branch learns global features using down-sampling with a stride-2 convolution layer and representations with 256 dimension features. The part-N branch has a similar architecture without down-sampling.

3.3 PROPOSED EXPERIMENTS

To compare the Re-ID datasets and approaches presented above, several experiments are conducted.

3.3.1 SINGLE DATASET EVALUATION

We first evaluate each approach and dataset pair individually. The standard Re-ID approach is simply fitted to the training split of the dataset, and evaluated on the testing split. The quality of the Re-ID model's predictions on the testing set is assessed using standard Re-ID metrics, coming from the field of information retrieval:

Rank-n was first discussed for Re-ID by Moon & Phillips (2001). It represents the proportion of queries for which at least one correct match was predicted within the n highest ranked gallery images. In practice, we report results for $n \in \{1, 5, 10\}$. This metric represents the Re-ID model's ability to retrieve the easiest match.

mAP The computation of *mean average precision* for Re-ID takes into account the predicted ranks of all existing matches (Zheng et al., 2015a). To have a perfect mAP, all the gallery images corresponding to the query need to be ranked in the first places. It represents an average performance of the model across all existing instances of the query.

mINP The *mean inverse negative penalty* was introduced recently by Ye et al. (2021). It reflects the position of the worst ranked match from the gallery. In other words, it reflects the capacity of a Re-ID model to find all instances of the query in the gallery.

These three metrics represent different skills of the evaluated Re-ID model. Computing them can help understand which of these skill is important regarding generalization to new contexts and to more complex real-world scenarios, i.e., live Re-ID in different cities.

3.3.2 CROSS-DATASET EVALUATION

The simple cross-dataset experiment from He et al. (2020) is also conducted. It consists in training an approach on one of the three standard Re-ID datasets, and evaluating it on the other two. The same metrics are used (rank-n, mAP and mINP). As the datasets were built in different geographic areas, these results can give first insights about domain generalization of the different training datasets and approaches. Conducting such cross-dataset evaluation is also much easier than evaluating the system in the live Re-ID setting. Hence, another objective of this experiment is to discover if simple cross-dataset evaluation can be used as a proxy to quickly test new datasets and approaches for live implementations. In other words, we want to know if there is correlation between cross-dataset results and live Re-ID results of dataset-approach pairs.

3.3.3 LIVE RE-ID EVALUATION

Finally, each standard Re-ID approach and dataset pairs are evaluated in the live Re-ID setting using the m-PRID dataset. We apply the evaluation methodology from Sumari et al. (2020). For each short video sequence, BB of pedestrians are extracted using a YOLO-V3 object detector (Redmon & Farhadi, 2018), trained on COCO (Lin et al., 2015) and available in TensorFlow (Abadi et al., 2015). The score threshold used to decide which predicted BB to keep is set to 0.5. Then, the trained standard Re-ID approaches are applied to the gallery composed of these BB. Following the notations of Sumari et al. (2020), the length of video sequences evaluated τ is set to 1000 frames and the number of candidates shown to the monitoring agent η is set to 20. These values generated best results by a large margin in their experiments. For the threshold β on Re-ID scores used to generate alerts, we test all values between 0 and 1 with a step size of 0.02.

Table 2: **Single dataset evaluations**. Results obtained by training and evaluating Re-ID approaches with the train and test splits of the same dataset. For each dataset, the best Re-ID approach is in bold.

Dataset	Approach	Rank-1	Rank-5	Rank-10	mAP	mINP
CUHK03	AGW	0.73	0.88	0.92	0.72	0.63
	MGN	**0.78**	**0.91**	**0.95**	**0.76**	**0.66**
	SBS	0.74	0.89	0.93	0.73	0.62
	BoT	0.69	0.86	0.92	0.67	0.55
DukeMTMC	AGW	0.89	0.95	**0.97**	0.80	0.46
	MGN	**0.91**	**0.96**	0.97	**0.82**	**0.47**
	SBS	0.89	0.95	0.96	0.79	0.44
	BoT	0.87	0.94	0.96	0.77	0.41
Market-1501	AGW	0.95	0.99	0.99	0.88	0.66
	MGN	**0.96**	**0.99**	**0.99**	**0.89**	**0.66**
	SBS	0.95	0.98	0.99	0.88	0.66
	BoT	0.94	0.98	0.99	0.86	0.61

To compare the different models, we use the live Re-ID metrics introduced by Sumari et al. (2020). On the one hand, the Finding Rate (**FR**) represents the proportion of videos where the query was present, such that an alert was shown to the monitoring agent and where the query was among the selected candidates. A low FR means that the query was missed frequently. On the other, the True Validation Rate (**TVR**) represents the proportion of alert shown to the monitoring agent in which the query was present among the candidates. A low TVR means that the agent was frequently disturbed for no reason, which can be problematic when many cameras need to be monitored simultaneously.

In this paper we also define two new metrics to represent the performance of a live Re-ID approach with a single number, to facilitate comparisons and interpretation. The first one is based on the observation that the meanings of FR and TVR are respectively very close to the meanings of recall and precision. This way, similarly to object detection evaluation, we can plot *TVR vs FR* curves and compute the mean Average Precision (**mAP**) as the area under the curve. The second unified metric consists in computing a weighted harmonic mean of FR and TVR, similarly to the F-score computation for precision and recall. We call the resulting metric F_γ, which is defined as follows:

$$F_\gamma = (1+\gamma^2).\frac{\text{FR.TVR}}{(\gamma^2.\text{FR}) + \text{TVR}}. \tag{1}$$

In practice, we compute F_γ for $\gamma \in \{0.5, 1, 2\}$. In $F_{0.5}$, we consider that having a high TVR is two times more important than a high FR. In F_2, we consider FR two times more important than TVR, and in F_1 FR and TVR contribute equally to the results. However, for each value of the threshold β, there is a different corresponding value of F_γ. To solve this issue, we use the same approach as Guérin et al. (2020), consisting in evaluating a model by its performance at the optimal configuration. The result is called optimal F_γ ($\mathbf{F}^*_\gamma$), and corresponds to the highest F_γ across values of β. The value of β corresponding to F^*_γ can be viewed as the operating point of the Re-ID model, which can be obtain by quick experiments in the practical implementation context. An F^*_γ score of 1 means that there exist a Re-ID threshold β such that it always find the query when it is in the video sequence, but never raises alerts when it is not.

The objective of this experiments is to see if the best approaches and datasets from previous experiments are also the best ones from the perspective of practical implementation in new cities.

4 RESULTS

In order to improve clarity, only a condensed version of the results is presented here. The complete results can be found in the appendix: Section A.1 contains all the results from cross-dataset evaluation, Section A.2 contains the missing metrics and the TVR vs FR curves for live Re-ID evaluations. Overall, the curated results presented in the core paper are representative of the complete results and are sufficient to draw our conclusions.

Table 3: **Cross-dataset evaluations**. Results obtained by training Re-ID approaches on one dataset and evaluating on another. For each evaluation dataset: the best Re-ID approach for a given dataset is in bold; the best training dataset for a given approach is in blue. R10 means Rank-10.

Evaluation dataset	Training dataset	AGW		MGN		SBS		BoT	
		R10	mAP	R10	mAP	R10	mAP	R10	mAP
CUHK03	Market-1501	0.21	0.80	**0.47**	**0.22**	0.40	0.18	0.15	0.04
	DukeMTMC	0.18	0.06	0.34	**0.14**	**0.35**	0.13	0.15	0.05
DukeMTMC	Market-1501	0.58	0.22	**0.77**	**0.39**	0.74	0.34	0.49	0.15
	CUHK03	0.50	0.17	**0.70**	**0.31**	0.60	0.21	0.36	0.10
Market-1501	DukeMTMC	0.75	0.26	**0.87**	**0.37**	0.82	0.31	0.71	0.22
	CUHK03	0.73	0.29	**0.86**	**0.39**	0.80	0.34	0.66	0.22

Table 4: **Live Re-ID evaluation**. Results obtained by training Re-ID approaches on one standard Re-ID dataset and evaluating on m-PRID for the live Re-ID setting. For each training dataset, the best approach is in bold and for each approach, the best dataset is in blue.

Approach	CUHK03		DukeMTMC		Market-1501	
	F_1^*	mAP	F_1^*	mAP	F_1^*	mAP
AGW	0.39	0.23	0.40	0.25	0.46	0.33
BoT	0.27	0.10	0.40	0.22	0.47	0.32
SBS	0.51	0.43	0.58	0.54	0.60	0.50
MGN	**0.66**	**0.60**	**0.76**	**0.72**	**0.69**	**0.63**

The results for single dataset evaluation are reported in Table 2. The results obtained are good: rank-1 and mAP are around 70% for the worst approach on the most difficult dataset. They are also relatively homogeneous: for each dataset-metric pairs, all four methods perform similarly (less than ten points difference).

The cross-dataset evaluation results are presented in Table 3. We only report Rank-10 for two reasons. First, the complete results show that the ranking of approaches is stable under different values of n. Second, having a high Rank-10 is more important than lower ranks for live Re-ID, as explained in Section 2.1.2. The results obtained show that the different approaches generalize very differently to new contexts. For instance training MGN on Market-1501 leads to 47% rank-10 accuracy on CUHK03, while the same experiment using BoT only reaches 15%. For comparison, when training was conducted on CUHK03 itself, only a 3 point difference was observed between the two approaches (Table 2). The choice of the training dataset is also more important. For example, when training a MGN model for CUHK03, Market-1501 is 13% better than DukeMTMC.

Finally, the live Re-ID evaluation results are presented in Table 4. They also illustrate that it is crucial to properly select the training dataset and approach for such task transfer. Overall, MGN appears to generalize much better for use in a live Re-ID setting. For training, Market-1501 appear to work best for most approaches except MGN. The best combination is MGN trained on DukeMTMC, reaching a mAP of 0.72 and an optimal F1 of 0.76, which are very good results for a simple pipeline considering pedestrian detection and Re-ID separately.

5 Discussion

All the approaches tested in this study perform well in the single dataset scenario. However, when it comes to generalization for use during live operations in a different context, MGN has a clear advantage against the other three techniques. This conclusion could already be intuited from the cross-dataset experiments, which suggests a simple yet powerful approach to test future standard Re-ID approaches before live deployment. MGN is the only approach involving a specific image

splitting, forcing the network to focus on different body part. In view of our results, this property appears to be desirable for generalization to the live Re-ID setting.

Proper selection of the training dataset also influences the results obtained in a different evaluation domain. However, there is no clear winner between Market-1501 and DukeMTMC to know which network should be used for any context. In addition, the cross-dataset results do not allow to choose the best dataset for training models for the live Re-ID setting. Indeed, Table 3 suggested that the best dataset for MGN should be Market-1501, whereas it is outperformed by DukeMTMC for live Re-ID (Table 4).

The conclusions from this study can be summarized as follows:

1. It is possible to build a successful live Re-ID pipeline by concatenating a good pedestrian detector with a standard Re-ID model.
2. Proper choice of the standard re-ID approach and training dataset is paramount to obtain satisfying results when transferring the model to the live Re-ID setting.
3. Simple cross-dataset evaluation can be used to quickly assess the generalization performance of future standard Re-ID techniques for live Re-ID.
4. To properly assess the performance of new training datasets, the complete evaluation methodology proposed in this paper should be conducted.

6 Conclusion

6.1 Overview

This paper presents a comprehensive evaluation methodology to benchmark different standard Re-ID approaches and training datasets with respect to their ability to be deployed in practical applications from a different context. To do so, we first formalized the new live Re-ID setting, and define new unified evaluation metrics to facilitate interpretation. The performance of different standard Re-ID models is evaluated on this setting. We also conduct simple cross-dataset experiments to see if it can be used to predict which datasets and approaches will generalize better to the live Re-ID setting.

The results obtained suggest that it is possible to build a good live Re-ID pipeline by simply concatenating a pedestrian detection model and a standard Re-ID model. However, the proper choice of a good Re-ID approach and the appropriate corresponding training dataset is paramount to obtain satisfying results. The simple cross-dataset experiments presented in this paper allow to quickly evaluate the transferability of a given approach, but it fails to assess the best training dataset. To evaluate other training datasets for practical Re-ID implementations, one need to go through the steps of evaluating the learned models on the live Re-ID setting. The proposed benchmarking methodology can be used straightforwardly to evaluate new datasets and new standard Re-ID approaches in the future.

6.2 Future work

The outputs of this study suggest several interesting future research directions. First, it would be very valuable to build a new live Re-ID dataset, allowing not only to confirm the results obtained in this study, but also to see if good live Re-ID performance is consistent across different scenarios. Then, this benchmark experiment can be extended to account for different pedestrian detection models, another important component of the live Re-ID pipeline. In particular, it would be interesting to study if specific Re-ID approaches combine better with specific object detection models. The evaluation methodology proposed in this paper could be used to answer this question. Another valuable contribution would be to create a ready-to-use website implementing the proposed benchmarking methodology for researchers to test their new approaches easily. In view of the results obtained, it could also be interesting to try to combine public datasets to train Re-ID models that are better suited for live Re-ID. Finally, it would be interesting to study how the good design choices identified in this study can be leveraged to develop successful end-to-end approaches for live Re-ID.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Damla Gül Altunay, Naciye Karademir, Okan Topçu, and Cem Direkoğlu. Intelligent surveillance system for abandoned luggage. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4. IEEE, 2018.

Apurva Bedagkar-Gala and Shishir K Shah. A survey of approaches and trends in person re-identification. *Image and vision computing*, 32(4):270–286, 2014.

Debayan Deb, Divyansh Aggarwal, and Anil K Jain. Identifying missing children: Face age-progression via deep feature aging. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10540–10547. IEEE, 2021.

Xing Fan, Wei Jiang, Hao Luo, and Mengjuan Fei. Spherereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation*, 60:51–58, Apr 2019. ISSN 1047-3203. doi: 10.1016/j.jvcir.2019.01.010. URL `http://dx.doi.org/10.1016/j.jvcir.2019.01.010`.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

Niloofar Gheissari, Thomas B Sebastian, and Richard Hartley. Person reidentification using spatiotemporal appearance. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pp. 1528–1535. IEEE, 2006.

Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):523–536, 2018.

Joris Guérin, Anne Magaly de Paula Canuto, and Luiz Marcos Garcia Goncalves. Robust detection of objects under periodic motion with gaussian process filtering. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 685–692. IEEE, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. URL `http://arxiv.org/abs/1512.03385`. arXiv: 1512.03385.

Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020.

Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pp. 91–102. Springer, 2011.

Khawar Islam. Person search: New paradigm of person re-identification: A survey and outlook of recent works. *Image and Vision Computing*, 101:103970, 2020.

Bahram Lavi, Ihsan Ullah, Mehdi Fatan, and Anderson Rocha. Survey on reliable deep learning-based person re-identification models: Are we there yet? *arXiv preprint arXiv:2005.00355*, 2020.

Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4):1092–1108, 2019.

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*, pp. 152–159, 2014.

Shengcai Liao, Zhipeng Mo, Jianqing Zhu, Yang Hu, and Stan Z Li. Open-set person re-identification. *arXiv preprint arXiv:1408.0872*, 2014.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1487–1495, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72812-506-0. doi: 10.1109/CVPRW.2019.00190. URL https://ieeexplore.ieee.org/document/9025455/.

Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609, October 2020. ISSN 1520-9210, 1941-0077. doi: 10.1109/TMM.2019.2958756. URL http://arxiv.org/abs/1906.08332. arXiv: 1906.08332.

Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001.

Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Going deeper into semi-supervised person re-identification. *arXiv preprint arXiv:2107.11566*, 2021.

PwC. Papers with Code, person re-identification. https://paperswithcode.com/task/person-re-identification, 2021. Accessed: 2021-09-28.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

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*, pp. 17–35. Springer, 2016.

Felix O Sumari, Luigy Machaca, Jose Huaman, Esteban WG Clua, and Joris Guérin. Towards practical implementations of person re-identification from full video frames. *Pattern Recognition Letters*, 138:513–519, 2020.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *arXiv:1512.00567 [cs]*, December 2015. URL http://arxiv.org/abs/1512.00567. arXiv: 1512.00567.

Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. *Proceedings of the 26th ACM international conference on Multimedia*, pp. 274–282, October 2018. doi: 10.1145/3240508.3240552. URL http://arxiv.org/abs/1804.01438. arXiv: 1804.01438 version: 1.

Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang. Human-in-the-loop person re-identification. In *European conference on computer vision*, pp. 405–422. Springer, 2016.

Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European conference on computer vision*, pp. 688–703. Springer, 2014.

Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3415–3424, 2017.

Yuanlu Xu, Bingpeng Ma, Rui Huang, and Liang Lin. Person search in a scene by jointly modeling people commonness and person uniqueness. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 937–940, 2014.

ChangShui Yang, Feng Qi, and Huizhu Jia. Survey on unsupervised techniques for person re-identification. In *2021 2nd International Conference on Computing and Data Science (CDS)*, pp. 161–164. IEEE, 2021.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015a.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pp. 1116–1124, 2015b.

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 *European Conference on Computer Vision*, pp. 868–884. Springer, 2016.

Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1367–1376, 2017.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random Erasing Data Augmentation. *arXiv:1708.04896 [cs]*, November 2017. URL http://arxiv.org/abs/1708.04896. arXiv: 1708.04896.

Weiming Zhuang, Yonggang Wen, Xuesen Zhang, Xin Gan, Daiying Yin, Dongzhan Zhou, Shuai Zhang, and Shuai Yi. Performance optimization of federated person re-identification via benchmark analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 955–963, 2020.

A APPENDIX

A.1 COMPLETE RESULTS FOR CROSS-DATASET EXPERIMENTS

This section presents all the results from our cross-dataset Re-ID experiments. Four standard Re-ID approaches are trained on three different standard Re-ID datasets, and evaluated on a different dataset. The complete results from these experiments are reported in Table 5.

Table 5: Complete results from our cross-dataset experiments.

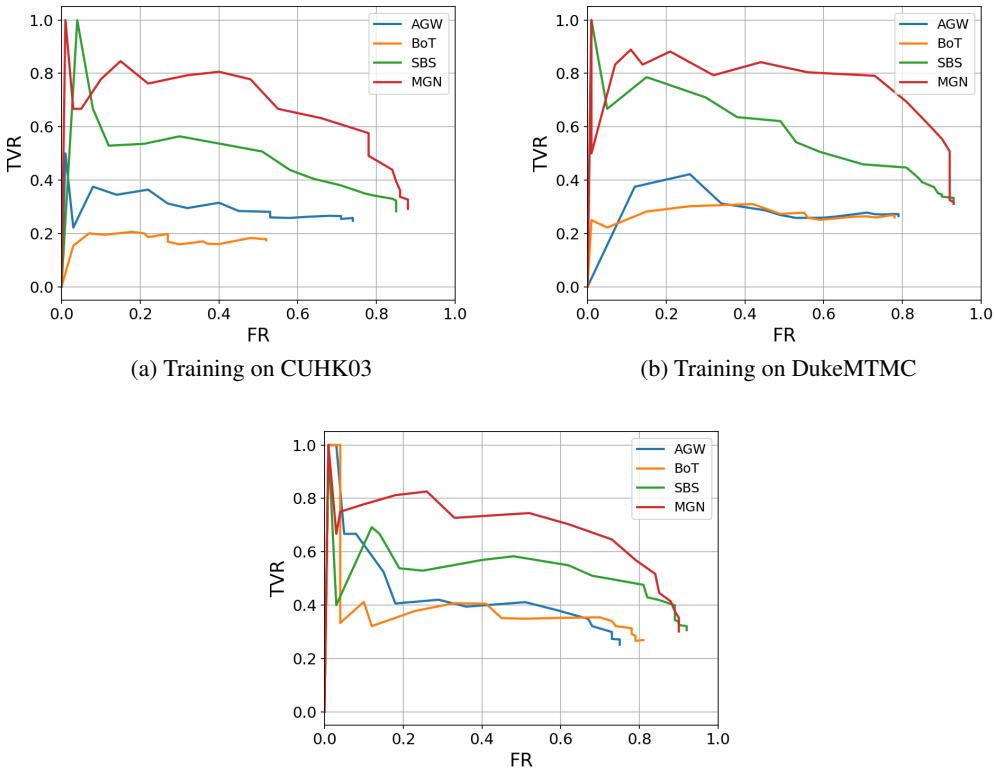
Train	TEST	Approach	Rank-1	Rank-5	Rank-10	mAP	mINP
CUHK03	Market-1501	AGW	0.54	0.67	0.73	0.29	0.05
		MGN	0.66	0.81	0.86	0.39	0.08
		SBS	0.60	0.75	0.80	0.34	0.07
		BoT	0.46	0.61	0.66	0.22	0.03
	DukeMTMC	AGW	0.29	0.44	0.50	0.17	0.02
		MGN	0.50	0.65	0.70	0.31	0.04
		SBS	0.39	0.54	0.60	0.21	0.02
		BoT	0.19	0.30	0.36	0.10	0.01
DukeMTMC	Market-1501	AGW	0.53	0.68	0.75	0.26	0.03
		MGN	0.67	0.82	0.87	0.37	0.06
		SBS	0.61	0.77	0.82	0.31	0.03
		BoT	0.49	0.65	0.71	0.22	0.02
	CUHK03	AGW	0.06	0.13	0.18	0.06	0.03
		MGN	0.14	0.26	0.34	0.14	0.07
		SBS	0.13	0.27	0.35	0.13	0.06
		BoT	0.05	0.10	0.15	0.05	0.03
Market-1501	DukeMTMC	AGW	0.37	0.52	0.58	0.22	0.03
		MGN	0.58	0.73	0.77	0.39	0.06
		SBS	0.54	0.68	0.74	0.34	0.05
		BoT	0.28	0.43	0.49	0.15	0.02
	CUHK03	AGW	0.08	0.15	0.21	0.08	0.04
		MGN	0.22	0.38	0.47	0.22	0.13
		SBS	0.19	0.31	0.40	0.18	0.11
		BoT	0.04	0.11	0.15	0.04	0.02

A.2 COMPLETE RESULTS FOR LIVE RE-ID EXPERIMENTS

This section presents all the results from our live Re-ID experiments. Four standard Re-ID approaches are trained on three different standard Re-ID datasets, and combined with a pedestrian detector for evaluation at the live Re-ID task, using the m-PRID dataset. The complete results from these experiments are reported in Table 6. The TVR vs FR curves for corresponding to these experiments are shown in Figures 2 and 3.

Table 6: Complete results from our live Re-ID experiments.

Training set	Approach	mAP	$F^*_{0.5}$	F^*_1	F^*_2
CUHK03	AGW	0.23	0.33	0.39	0.54
	BoT	0.10	0.21	0.27	0.38
	SBS	0.43	0.51	0.51	0.64
	MGN	0.60	0.69	0.66	0.73
DukeMTMC	AGW	0.25	0.38	0.40	0.57
	BoT	0.22	0.33	0.40	0.56
	SBS	0.54	0.59	0.58	0.70
	MGN	0.72	0.78	0.76	0.80
Market-1501	AGW	0.33	0.43	0.46	0.57
	BoT	0.32	0.41	0.47	0.60
	SBS	0.50	0.56	0.60	0.71
	MGN	0.63	0.69	0.69	0.75

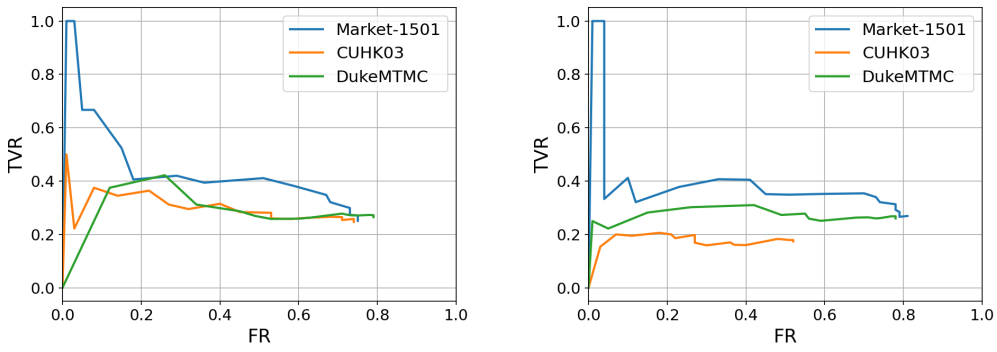


(a) Training on CUHK03

(b) Training on DukeMTMC

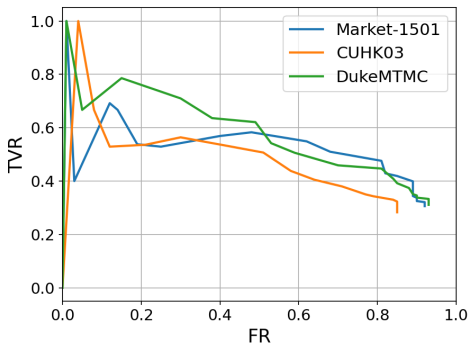
(c) Training on Market-1501

Figure 2: **Influence of the standard Re-ID approach**. TVR vs FR curves of different standard Re-ID approaches for different training datasets. Evaluation is conducted on the *modified PRID-2011* dataset for live Re-ID.

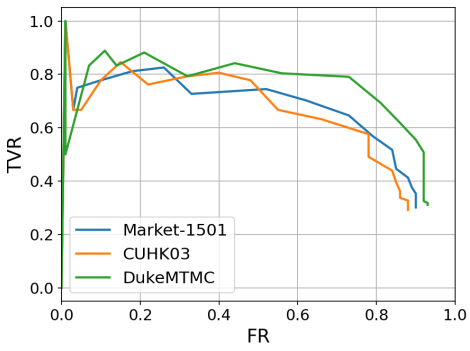


(a) AGW

(b) BoT



(c) SBS



(d) MGN

Figure 3: **Influence of the training dataset**. TVR vs FR curves using different standard Re-ID datasets for training different Re-ID approaches. Evaluation is conducted on the *modified PRID-2011* dataset for live Re-ID.