Benchmarking Person Re-Identification Datasets and Approaches for Practical Real-World Implementations

Jose Huaman[1], Felix O. Sumari H.[1], Luigy Machaca[1], Esteban Clua[1] and Joris Guérin[2]
[1]*Instituto de Computação, Universidade Federal Fluminense, Niteroi-RJ, Brazil*
[2]*Espace-Dev, Univ. Montpellier, IRD, Montpellier, France*

Keywords: Person Re-Identification, Practical Deployment, Benchmark Study.

Abstract: Person Re-Identification (Re-ID) is receiving a lot of attention. Large datasets containing labeled images of various individuals have been released, and successful approaches were developed. However, when Re-ID models are deployed in new cities or environments, they face an important domain shift (ethnicity, clothing, weather, architecture, etc.), resulting in decreased performance. 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 built the training dataset. 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 unsupervised deployment for live operations. We benchmark four Re-ID approaches on three datasets, providing insight and guidelines that can help to design better Re-ID pipelines.

1 INTRODUCTION

Person Re-Identification (Re-ID) is a computer vision problem aiming to find an individual in a network of cameras. It has diverse potential 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). In the literature, the problem of Re-ID is studied under different settings. On the one hand, the most studied Re-ID paradigm, which we call *standard Re-ID*, tries to find images representing the query within a gallery of pre-cropped images of persons, containing at least one correct match (Lavi et al., 2020). On the other hand, we recently introduced a setting considering all the constraints to implement Re-ID for live operations, which we call *live Re-ID* (Sumari et al., 2020). The first contribution of this paper is to better formalize the live Re-ID definition and to extend the evaluation metrics to facilitate interpretation.

Standard Re-ID is not the best-suited paradigm for practical implementations, as it does not consider the influence of domain shift due to pedestrian detection errors or deployment in a new city. Indeed, in our previous experiments (Sumari et al., 2020), we showed that training a successful standard Re-ID model does not guarantee good performance when evaluated in a live Re-ID context. Nevertheless, most publicly available large-scale datasets focus on the standard Re-ID setting, and many successful approaches have been developed for this task. Hence, we believe that it is essential to study if these datasets and approaches can be used to implement and deploy practical applications in different contexts.

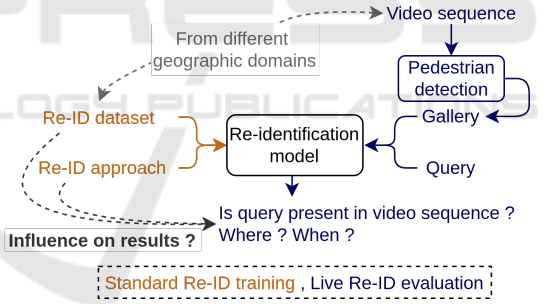

Figure 1: Objectives of our benchmark study.

More specifically, the objective of this paper is to answer these questions:

1. Which characteristics of a standard Re-ID dataset are most important for live Re-ID deployment?
2. Which standard Re-ID approaches can be successfully deployed in the live Re-ID setting?
3. Do different Re-ID approaches have different optimal datasets for deployment?

Huaman, J., Sumari H., F., Machaca, L., Clua, E. and Guérin, J.
Benchmarking Person Re-Identification Datasets and Approaches for Practical Real-World Implementations.
DOI: 10.5220/0011633800003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 495-502
ISBN: 978-989-758-634-7; ISSN: 2184-4321

4. Can we use cross-dataset evaluation to assess the deployability of a given approach-dataset pair?

To answer these questions, we present a study using three standard Re-ID datasets and four recent approaches. For each approach-dataset pair, the Re-ID model obtained is evaluated against the other two datasets and against another one configured for live Re-ID. We also combine training datasets to investigate how dataset size and diversity influence the generalization of the standard Re-ID model (Figure 1).

In this paper, we consider the evaluation of Re-ID models without additional training on images from the target domain. More sophisticated approaches have been proposed for domain adaptation of standard Re-ID models, e.g., unsupervised domain adaptation (Zhao et al., 2020). Such approaches are not tested in this work, but we believe standard Re-ID models performing well without target domain training (our experiments) are likely to be good initialization for more sophisticated fine-tuning approaches.

2 RELATED WORK

Here, we define the different Re-ID settings and discuss existing benchmark studies about Re-ID.

2.1 Person Re-Identification Settings

The field of Re-ID consists in retrieving instances of a given individual, called the *query*, within a complex set of multimedia content called the *gallery*. Different settings are defined by how they represent the query and the gallery items, the constraints on the gallery content, and the boundaries of the Re-ID system.

Standard Re-ID. Both the query and all gallery items are well-cropped images representing entire human bodies. It is sometimes called closed-set Re-ID as it assumes that the query has at least one representative in the gallery. It is the most studied Re-ID setting in terms of the number of papers, datasets, and benchmarks published (Papers with Code, 2021). For an overview of standard Re-ID approaches, see (Ye et al., 2021).

Person Search. Gallery items are replaced by whole scene images (Xiao et al., 2017), i.e., a person search model must return the gallery image where the query is present and its location in the image. A survey about person search was proposed in (Islam, 2020).

Open-Set Re-ID. In this setting, there is no guarantee that the query is present in the gallery. A survey about open-set Re-ID model was proposed in (Leng et al., 2019).

Video-Based Re-ID. Images (query and gallery) are replaced by image sequences extracted from consecutive video frames. Sequences are composed of well-cropped entire body images representing the same person. A survey about video-based Re-ID was proposed in (Ye et al., 2021).

2.2 Live Re-ID Setting

The *live Re-ID* setting (Sumari et al., 2020) takes into account all relevant aspects for deploying Re-ID in practical real-world applications (Figure 2).

When searching a query 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 *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. We recently showed that reducing the size of the gallery improves live Re-ID results (Machaca et al., 2022).

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.Finally, live Re-ID also takes into account that predictions need to be *processed by a human agent*, who triggers 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 human operators.

2.3 Person Re-Identification Benchmarks

The largest Re-ID benchmark to date was proposed in (Gou et al., 2018). They evaluated 30 approaches on 16 public datasets for standard and video-based Re-ID. They also built a new dataset to represent real-world constraints, e.g., pedestrian detection errors and illumination variations. 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.

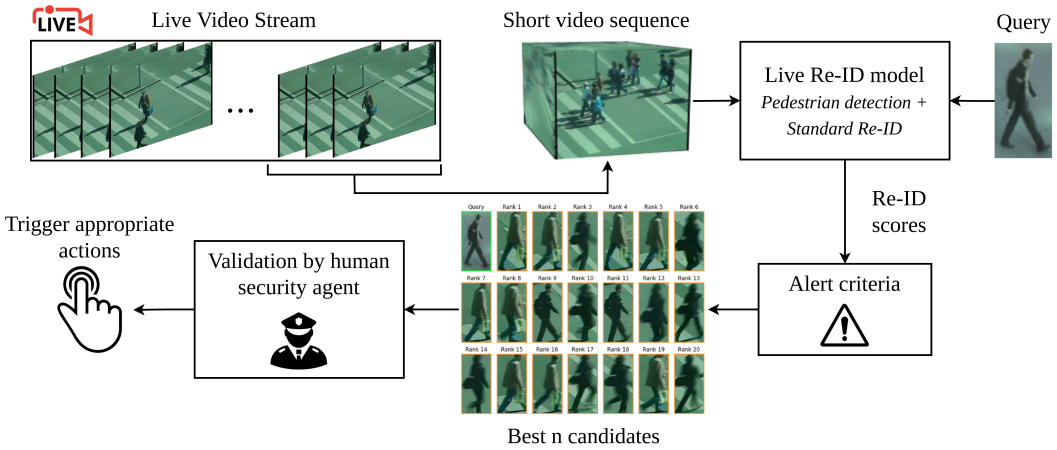

Figure 2: The live Re-ID setting. When deploying Re-ID models in practice, the galleries are composed of whole scene video sequences. When an alert is raised, the data are verified by a security agent to decide whether actions should be triggered.

A smaller benchmark for video-based Re-ID was proposed in (Zheng et al., 2016).

Extensive experiments were conducted to compare pedestrian detection models on a two-step person search pipeline (Zheng et al., 2017). It showed that the best models at object detection are not necessarily the best suited for person search. In addition, the first benchmark regarding the cross-domain transfer of Re-ID approaches was proposed in (He et al., 2020). Their experiments consisted in training an approach on one standard Re-ID dataset and evaluating on another.

However, none of these studies allows for assessing the performance of a Re-ID model against all the challenges involved during live deployment in a new environment. Our paper contributes to bridging this gap by conducting experiments within the live Re-ID setting. In particular, we consider the influence of different standard Re-ID approaches and training datasets on live Re-ID results.

3 BENCHMARK METHODOLOGY

This section presents the different components of the proposed benchmarking evaluation.

3.1 Datasets

In our experiments, we used three public datasets to train standard Re-ID models and a live Re-ID dataset. Figure 3 shows example images from the datasets, where we can see that they represent people from different geographic regions, under different resolutions, lighting conditions, and camera angles.

Table 1: Characteristics of the standard Re-ID datasets.

Dataset	Input type	# IDs	# Images
CUHK03	Train	767	7368
	Test (Query)	700	1400
	Test (Gallery)	700	5328
DukeMTMC	Train	702	16522
	Test (Query)	702	2228
	Test (Gallery)	1110	17661
Market-1501	Train	751	12936
	Test (Query)	750	3368
	Test (Gallery)	751	15913

Standard Re-ID Datasets. Table 1 summarizes relevant statistics about the standard Re-ID datasets used. *Market-1501* was collected in Beijing, China (Zheng et al., 2015). The cropped images are detected automatically using a Deformable Part Model and are filtered manually to keep only good BBs representing humans. Cropped images appear to have high resolution and good lighting conditions (Figure 3a). *DukeMTMC* was collected in Durham, North Carolina, USA (Ristani et al., 2016). The BB in DukeMTMC are hand drawn, and lighting conditions are good but the resolution of the BBs is relatively low (Figure 3b). *CUHK03* was built using video footage collected in Hong Kong (Li et al., 2014). In our work, we used the manually labeled version of the BBs. Cropped images are high resolution but illumination is dark, which reduces image quality (Figure 3c).

Live Re-ID Dataset. For the live Re-ID setting, we used the same dataset as our previous work (Sumari et al., 2020), which we call *m-PRID*. It is a modified version of PRID-2011 (Hirzer et al., 2011), built from the raw video footage and the original annotations used to build the official PRID-2011. The videos were collected from two cameras (A and B), in Graz,

(a) Market-1501. (b) DukeMTMC. (c) CUHK03.

(d) Original PRID-2011. (e) m-PRID. Extracted with YOLO-V3.

Figure 3: Benchmarking datasets. Example images from the datasets used in our experimental study.

Austria. This way, compared to the training datasets above, the evaluation on m-PRID represents a geographic domain shift. In total, PRID-2011 contains 385 different identities for A and 749 for B, of which 200 identities appear in both cameras. The m-PRID dataset is composed of two minutes videos (30 from A and 33 from B). For each short video sample, a ground truth file gathers information about each person it contains (frames where it appears, BB coordinates). For evaluation, 73 queries are considered.

To better grasp the influence of pedestrian detection, we also evaluate our models on the original PRID-2011 dataset. Figure 3d shows cropped images of poor resolution, taken from relatively high camera angle compared to other datasets. This way, we can see if the performance decrease on the live Re-ID setting is due to the domain shift of PRID or to the pedestrian detector inaccuracies (Figure 3e).

3.2 Re-ID Approaches Evaluated

Four recent standard Re-ID approaches are tested.

Bag of Tricks (BoT). This approach resulted from the observation that most Re-ID improvements come from neural network training tricks rather than Re-ID approaches themselves (Luo et al., 2019). As a result, they came up with a simple recipe to successfully train standard Re-ID models.

Strong Baseline and Batch Normalization Neck (SBS). This approach (Luo et al., 2020) extended BoT by adding more tricks, such as a warm-up strategy and random erasing augmentation

Attention Generalized Mean Pooling with Weighted Triplet Loss (AGW). This technique (Ye et al., 2021) was also designed on top of BoT with three new components: a non-local attention block, a learnable pooling layer, and the use of weighted regularization triplet loss.

Multiple Granularity Network (MGN). This approach (Wang et al., 2018) combines local and global information in different image granularity.

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/dataset pair individually. The standard Re-ID approach is fitted to the training split of the dataset and evaluated on the test split. The performance of the Re-ID model on the testing set is assessed using standard Re-ID metrics:

Rank-n. 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 $n \in \{1, 5, 10\}$. It represents the model's ability to retrieve the easiest match.

mAP. The *mean average precision* for Re-ID takes into account the ranks of all existing matches. It is the average performance across all instances of the query.

mINP. The *mean inverse negative penalty* reflects the position of the worst ranked match from the gallery (Ye et al., 2021). It represents the capacity of a model to find all instances of the query.

These three metrics represent different skills of a Re-ID model. Computing them might 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

A simple cross-dataset experiment is also conducted. We train an approach on one of the three standard Re-ID datasets and evaluate it on the other two. The same metrics are used. As the datasets were built in different geographic areas, it can give first insights into domain generalization of the different training datasets and approaches. Conducting such cross-dataset evaluation is 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 a correlation between cross-dataset results and live Re-ID results.

For the cross-datasets experiments, we also try to combine training datasets to see if it improves test performance. For $\text{COMBINED}_{\text{all}}$, training is conducted on all training sets available (Market-1501, DukeMTMC, and CUHK03), including the one corresponding to the test set of interest. This allows us to evaluate if adding data from other sources can help to improve standard Re-ID. For $\text{COMBINED}_{\text{others}}$, the training set corresponding to the test dataset is excluded. For example, when evaluating on CUHK03, the standard Re-ID models are trained on Market-1501 and DukeMTMC. Finally, the $\text{COMBINED}_{\text{scaled}}$ setting is similar to $\text{COMBINED}_{\text{others}}$, but we ensure that the total number of training data is equal to the number of data in the largest dataset, e.g., when evaluating on CUHK03, $\text{COMBINED}_{\text{scaled}}$ is composed of 8261 images from DukeMTMC and 8261 from Market-1501. Comparing $\text{COMBINED}_{\text{scaled}}$ with $\text{COMBINED}_{\text{others}}$ allows us to evaluate how size and diversity affect the generalization power of a dataset. As PRID-2011 is not among the training datasets, $\text{COMBINED}_{\text{all}}$ and $\text{COMBINED}_{\text{others}}$ are identical and called COMBINED.

3.3.3 Live Re-ID Evaluation

Finally, each standard Re-ID approach and dataset pair is evaluated in the live Re-ID setting using m-PRID. We apply the evaluation methodology from (Sumari et al., 2020). For each short video sequence, BBs of pedestrians are extracted using YOLO-V3 (Redmon and Farhadi, 2018), trained on COCO, and available in TensorFlow. The score threshold used to decide which predicted BBs to keep is set to 0.5. Then, the trained standard Re-ID approaches are applied to the gallery composed of these BBs. Also, 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 the best results by a large margin in previous experiments (Sumari et al., 2020). For the β threshold on Re-ID scores, used to generate alerts, we test values between 0 and 1 with a step size of 0.02.

To compare the different models, we use the live Re-ID metrics from (Sumari et al., 2020). The *Finding Rate* (FR) represents the proportion of videos where the query was present, such that an alert was shown to the agent and where the query was among the selected candidates. A low FR means that the query was missed frequently. The *True Validation Rate* (TVR) represents the proportion of alerts 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.

Table 2: Single dataset evaluations. For each dataset, the best Re-ID approach is in bold.

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

To facilitate comparisons and interpretation, we also propose two new live Re-ID metrics. The first one is based on the observation that the meanings of FR and TVR are respectively very close to recall and precision. This way, we can plot *TVR vs FR* curves and compute the *mean Average Precision* (mAP) as the area under the curve. The second metric is inspired by the F_1-score:

$$F_1 = 2.\frac{\text{FR.TVR}}{\text{FR}+\text{TVR}}. \quad (1)$$

However, for each value of β, there is a different corresponding value of F_1. To solve this issue, we use the same approach as in (Guérin et al., 2020) consisting in evaluating a model by its performance at the optimal configuration. The result is called F_1^* and corresponds to the highest F_1 across values of β. The value of β corresponding to F_1^* can be viewed as the operating point of the Re-ID model, which can be obtained by quick experiments in the practical implementation context. An F_1^* score of 1 means that there exists a β such that it always finds the query when it is present, but never raises alerts when it is not.

Combined datasets experiments are also conducted for live Re-ID. However, as PRID-2011 is not one of the training datasets used in our experiments, $\text{COMBINED}_{\text{all}}$ and $\text{COMBINED}_{\text{others}}$ are actually equivalent and simply called COMBINED. We aim to know if the best approaches and datasets from previous experiments are also the best for live Re-ID.

4 RESULTS AND DISCUSSION

In order to improve clarity, only a condensed version of the results is presented. Complete results: github.com/josemiki/benchmarking_person_Re_ID.

The results for single dataset evaluation are reported in Table 2, and the cross-dataset evaluation re-

Table 3: Cross-dataset evaluations. 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	0.08	**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
	$\text{COMBINED}_{\text{all}}$	0.94	0.71	**0.96**	**0.82**	0.94	0.76	0.92	0.68
	$\text{COMBINED}_{\text{others}}$	0.32	0.14	**0.55**	**0.27**	0.52	0.24	0.28	0.11
	$\text{COMBINED}_{\text{scaled}}$	0.31	0.13	**0.52**	**0.23**	0.46	0.20	0.23	0.09
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
	$\text{COMBINED}_{\text{all}}$	0.96	0.79	**0.97**	**0.82**	0.96	0.78	0.96	0.77
	$\text{COMBINED}_{\text{others}}$	0.65	0.29	**0.81**	**0.44**	0.79	0.41	0.55	0.21
	$\text{COMBINED}_{\text{scaled}}$	0.62	0.26	**0.78**	**0.40**	0.75	0.35	0.51	0.18
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
	$\text{COMBINED}_{\text{all}}$	0.99	0.88	**0.99**	**0.91**	0.99	0.88	0.99	0.86
	$\text{COMBINED}_{\text{others}}$	0.83	0.38	**0.93**	**0.52**	0.91	0.47	0.80	0.34
	$\text{COMBINED}_{\text{scaled}}$	0.83	0.38	**0.92**	**0.52**	0.89	0.46	0.78	0.32
PRID-2011	CUHK03	0.18	0.11	**0.35**	**0.26**	0.29	0.20	0.13	0.09
	DukeMTMC	0.20	0.12	**0.42**	**0.30**	0.26	0.17	0.16	0.07
	Market-1501	0.26	0.19	**0.40**	**0.28**	0.30	0.20	0.23	0.13
	COMBINED	0.32	0.20	**0.45**	**0.35**	0.33	0.23	0.24	0.15
	$\text{COMBINED}_{\text{scaled}}$	0.24	0.18	**0.46**	**0.36**	0.36	0.26	0.22	0.15

sults in Table 3. We only report Rank-10 for two reasons: the complete results show that the ranking of approaches is stable under different n, and Rank-10 is more important for live Re-ID (Section 2.1).

Single dataset results are good: Rank-10 (resp. mAP) is above 90% (resp. 70%) for the worst approach on the most difficult dataset. They are also relatively homogeneous: for each dataset-metric pair, all four methods perform similarly (less than 10% difference). However, the tested approaches generalize differently to new contexts. For example, 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% difference was observed between the approaches. The choice of the training dataset is also important, e.g., when training MGN 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 using a single dataset is MGN trained on DukeMTMC, reaching a mAP of 0.72 and an optimal F1 of 0.76.

4.1 Impact of the Training Dataset

Proper selection of the training dataset influences the results in a different evaluation domain. However, there is no clear winner between Market-1501 and DukeMTMC to know which dataset should be used for any context. In addition, the cross-dataset results do not help to choose the best dataset for training live Re-ID models. Indeed, Table 3 suggests that Market-1501 should be used for MGN, whereas it is outperformed by DukeMTMC for live Re-ID (Table 4). In the remaining of this section, we discuss the results obtained on the combined datasets settings to gain new insights regarding building standard Re-ID datasets for efficient training of live Re-ID models.

4.1.1 Can Data from a Different Domain Improve Results in the Standard Re-ID Scenario?

To answer, we compare results from Table 2 and the $\text{COMBINED}_{\text{all}}$ rows (Table 3). The results obtained for $\text{COMBINED}_{\text{all}}$ seem slightly better than the results obtained when learning only on the training set of the evaluated dataset (Rank-10 and mAP). To confirm this intuition, we conduct a Paired Sample T-Test to determine whether the mean difference between the results obtained using the single in-domain training set and the $\text{COMBINED}_{\text{all}}$ is statistically significant. The p-values obtained are 0.2750 for R10 and 0.2313 for mAP, suggesting that *we cannot conclude that using more data from a different domain is beneficial to the standard Re-ID training process*.

Table 4: Live Re-ID evaluation. For each dataset, the best approach is in bold. For each approach, the best dataset is in blue.

Approach	CUHK03		DukeMTMC		Market-1501		COMBINED		$\text{COMBINED}_{\text{scaled}}$	
	F_1^*	mAP	F_1^*	mAP	F_1^*	mAP	F_1^*	mAP	F_1^*	mAP
AGW	0.39	0.23	0.40	0.25	0.46	0.33	0.56	0.49	0.49	0.39
BoT	0.27	0.10	0.40	0.22	0.47	0.32	0.45	0.30	0.44	0.31
SBS	0.51	0.43	0.58	0.54	0.60	0.50	0.71	0.71	0.68	0.72
MGN	**0.66**	**0.60**	**0.76**	**0.72**	**0.69**	**0.63**	**0.81**	**0.80**	**0.77**	**0.75**

4.1.2 Comparison of Dataset Size and Diversity for Cross-Domain Generalization

We want to know whether combining datasets from different domains help cross-domain generalization. Hence, we compare the results for $\text{COMBINED}_{\text{others}}$ (COMBINED for PRID-2011) against the results from the best individual dataset in Table 3. The Paired Sample T-Test gives p-values of 0.0001 for both R10 and mAP. Hence, our experiments confirm that *combining several training datasets from different domains allows to train Re-ID models that generalize better to new unknown domains*.

We then want to know if increasing the diversity in the training dataset without increasing its size also helps for cross-domain generalization (i.e., $\text{COMBINED}_{\text{scaled}}$ vs. best individual dataset). The Paired T-Test gives p-values of 0.0001 for R10 and 0.0005 for mAP. Hence, our experiments confirm that *increasing diversity in the training dataset, even without increasing its size, allows us to train Re-ID models that generalize better to new domains*.

Finally, we also want to know whether the size of the training dataset is actually helping cross-domain generalization or if diversity is sufficient. To evaluate this, we compare $\text{COMBINED}_{\text{others}}$ against $\text{COMBINED}_{\text{scaled}}$. The Paired Sample T-Test gives p-values of 0.0020 for R10 and 0.0075 for mAP. Hence, our experiments confirm that *adding more data from domains that are already present in the training set helps generalization to unknown domains*.

4.1.3 Live Re-ID Results

The results on m-PRID (Table 4) confirm the conclusions from the cross-dataset experiments. The $\text{COMBINED}_{\text{scaled}}$ results are better than the results with a single dataset, i.e., training data diversity is important for live Re-ID. The COMBINED results are themselves better than $\text{COMBINED}_{\text{scaled}}$, suggesting that one should use all the available data to train a good model for live Re-ID. Finally, we emphasize the good results obtained by training MGN on the COMBINED training dataset. These results are very encouraging after the pessimistic live Re-ID results reported in (Sumari et al., 2020).

4.2 Impact of the Approach

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

Besides MGN, SBS approach also appears to present much better generalization than its competitors (Table 3 and 4). Hence, a promising research direction for live Re-ID research would be to design a new standard Re-ID architecture combining features from MGN and SBS.

5 CONCLUSION

This paper presents a comprehensive benchmark of standard Re-ID approaches and training datasets with respect to their ability to be deployed in practical applications (live Re-ID setting). We also conduct cross-dataset experiments to see if they can be used to predict which approaches will generalize better to live Re-ID. The main conclusions from this study are:

1. It is possible to design good live Re-ID pipelines by properly choosing the standard Re-ID model and combining publicly available training dataset.
2. Proper choice of the standard re-ID approach and dataset influences greatly the results when transferring the model to the live Re-ID setting.
3. Increasing training dataset diversity helps generalization to the cross-domain live Re-ID setting.
4. Increasing training dataset size allows to improve cross-domain generalization even further.
5. Simple cross-dataset evaluation can be used to quickly assess the generalization performance of future standard Re-ID techniques for live Re-ID.

For future works, it would be valuable to build new live Re-ID datasets, to confirm our results and

to see if good live Re-ID performance is consistent across different scenarios. Our benchmark could also be extended to account for pedestrian detectors, as it would be interesting to study which Re-ID approach combines better with which OD models. Finally, it would also be interesting to see if existing unsupervised cross-dataset adaptation methods could help generalization to the live Re-ID setting.

REFERENCES

Altunay, D. G., Karademir, N., Topçu, O., and Direkoğlu, C. (2018). Intelligent surveillance system for abandoned luggage. In *26th Signal Processing and Communications Applications Conf. (SIU)*, pages 1–4. IEEE.

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

Gou, M., Wu, Z., Rates-Borras, A., Camps, O., Radke, R. J., et al. (2018). 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.

Guérin, J., de Paula Canuto, A. M., and Goncalves, L. M. G. (2020). Robust detection of objects under periodic motion with gaussian process filtering. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 685–692. IEEE.

He, L., Liao, X., Liu, W., Liu, X., Cheng, P., and Mei, T. (2020). Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*.

Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer.

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

Lavi, B., Ullah, I., Fatan, M., and Rocha, A. (2020). Survey on reliable deep learning-based person re-identification models: Are we there yet? *arXiv preprint arXiv:2005.00355*.

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

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159.

Liao, S., Mo, Z., Zhu, J., Hu, Y., and Li, S. Z. (2014). Open-set person re-identification. *arXiv preprint arXiv:1408.0872*.

Luo, H., Gu, Y., Liao, X., Lai, S., and Jiang, W. (2019). 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)*, pages 1487–1495, Long Beach, CA, USA. IEEE.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., and Gu, J. (2020). A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification. *IEEE Transactions on Multimedia*, 22(10):2597–2609. arXiv: 1906.08332.

Machaca, L., Huaman, J., Clua, E., Guerin, J., et al. (2022). TrADe Re-ID–live person re-identification using tracking and anomaly detection. *21st IEEE International Conference on Machine Learning and Applications (to appear)*.

Papers with Code (2021). person Re-ID. paperswithcode.com/task/person-re-identification.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer.

Sumari, F. O., Machaca, L., Huaman, J., Clua, E. W., and Guérin, J. (2020). Towards practical implementations of person re-identification from full video frames. *Pattern Recognition Letters*, 138:513–519.

Wang, G., Yuan, Y., Chen, X., Li, J., and Zhou, X. (2018). Learning Discriminative Features with Multiple Granularities for Person Re-Identification. *Proceedings of the 26th ACM international conference on Multimedia*, pages 274–282. arXiv: 1804.01438 version: 1.

Xiao, T., Li, S., Wang, B., Lin, L., and Wang, X. (2017). Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3415–3424.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., and Hoi, S. C. (2021). Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, F., Liao, S., Xie, G.-S., Zhao, J., Zhang, K., and Shao, L. (2020). Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *European Conference on Computer Vision*, pages 526–544. Springer.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124.

Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., and Tian, Q. (2017). Person re-identification in the wild. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376.