UNIVERSIDADE FEDERAL FLUMINENSE

JOSE MIGUEL HUAMAN CRUZ

Benchmarking person re-identification approaches and training datasets for practical real-world implementations

NITERÓI

2022

JOSE MIGUEL HUAMAN CRUZ

Benchmarking person re-identification approaches and training datasets for practical real-world implementations

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Orientador:
ESTEBAN WALTER GONZALES CLUA

Coorientador:
JORIS MICHEL GÉRARD DANIEL GUERIN

NITERÓI

2022

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

H874b Huaman Cruz, Jose Miguel
Benchmarking person re-identification approaches and training datasets for practical real-world implementations / Jose Miguel Huaman Cruz. - 2022.
73 f.

Orientador: Esteban Walter Gonzalez Clua.
Coorientador: Joris Michel Gérard Daniel Guerin.
Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2022.

1. Person re-identification. 2. Benchmark study. 3. Practical deployment. 4. Live Re-ID. 5. Produção intelectual. I. Clua, Esteban Walter Gonzalez, orientador. II. Guerin, Joris Michel Gérard Daniel, coorientador. III. Universidade Federal Fluminense. Instituto de Computação. IV. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

JOSE MIGUEL HUAMAN CRUZ

Benchmarking person re-identification approaches and training datasets for practical real-world implementations

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Aprovada em Decembro de 2022.

BANCA EXAMINADORA

Prof. Esteban Walter Gonzales Clua, PhD. – Orientador, UFF

Prof. Joris Michel Gérard Daniel Guerin, PhD. – Co-orientador, UFF

Prof. Leandro Augusto Frata Fernandes, PhD. – UFF

Prof. Luiz Eduardo Soares Oliveira, PhD. – UFP

Niterói

2022

To my parents Lily and Jose, for all your support during these years of study. To my sisters Gloria and Mercedes for all the happiest moments during these years. My grandparents Gloria and Fausto, saw all my achievements and always protected me from the sky. This achievement is for you, mommy Lily, you are the most important person in my life.

Agradecimentos

I am very grateful for the opportunity that Instituto de Computação-UFF gave me to do my master's degree. For the support I always received from all the members of this beautiful place. For me it was always a dream to study outside of my city Arequipa and now I am making that dream come true.

I would like to thank my advisor and why not consider him as father, thank you Professor Esteban Clua. I still remember the meeting we had in 2018 when I hadn't finished my studies. But there you encouraged us to postulate to the master's program and since I arrived in the beautiful city of Niteroi you were always available to talk and support us in whatever was necessary.

I would also like to thank my co-advisor, Professor Joris Guerin, whom I consider a friend. Thank you for the time you always gave me to answer questions, review the results of the experiments and motivate us to move forward.

Finally, but not less importantly I would like to say Thank You to my family, my friends and all person who during these years encouraged me to move forward and never give up. I believe without your constant greeting and motivation this was impossible. You'll be in my heart forever.

Resumo

A Reidentificação de Pessoas (Re-ID) está recebendo muita atenção recentemente. Grandes conjuntos de dados contendo imagens rotuladas de vários indivíduos foram liberados, permitindo aos pesquisadores desenvolver e testar muitas abordagens bem-sucedidas. No entanto, quando esses modelos de Re-ID são implantados em uma nova cidade ou ambiente, a tarefa de pesquisar pessoas em uma rede de câmeras de segurança provavelmente enfrentará uma importante mudança de domínio, resultando em desempenho reduzido. De fato, enquanto a maioria dos conjuntos de dados públicos foram coletados em uma área geográfica limitada, as imagens de uma nova cidade apresentam características diferentes (por exemplo, etnia e estilo de roupa das pessoas, clima, arquitetura etc.).

Além disso, os quadros inteiros dos fluxos de vídeo devem ser convertidos em imagens recortadas de pessoas usando modelos de detecção de pedestres, que se comportam de forma diferente dos anotadores humanos que criaram o conjunto de dados usado para treinamento. Para entender melhor a extensão desse problema, este trabalho apresenta uma metodologia completa para avaliar as abordagens de Re-ID e conjuntos de dados de treinamento com relação à sua adequação para implantação não supervisionada para operações ao vivo. Esse método é usado para comparar quatro abordagens de Re-ID e três conjuntos de dados, fornecendo insights e diretrizes que podem ajudar a projetar melhores pipelines de Re-ID no futuro.

Palavras-chave:reidentificação de pessoa, estudo de benchmark, implantação prática

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 for 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 ethnicity 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 work introduces a complete methodology to evaluate Re-ID approaches and training datasets with respect to their suitability for unsupervised deployment for live operations. This method is used to benchmark four Re-ID approaches and three datasets, providing insight and guidelines that can help design better Re-ID pipelines in the future.

Keywords: person re-identification, benchmark study, practical deployment

List of Figures

List of Tables

Acronyms

AGW Attention Generalized Network Weighted

BB Bounding Boxes

BN Batch Normalization

BNNeck Batch Normalization Neck

BoT Bags of tricks

DPM Deformable Part Model

GPU Graphics Processing Unit

MGN Multi Granularity Network

Re-ID Re-Identification

SBS Strong Base Line

Contents

Chapter 1

Introduction

1.1 Context and Motivation

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 work deals with person Re-Identification (Re-ID), a computer vision problem that intends to find 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; AGGARWAL; JAIN, 2021), among others.

A lot of ReID methods are being developed and different datasets are being used for training these methods. Some methods obtain over 90% Rank-1 accuracy (YE et al., 2021) but when these are used in real-world scenarios their performance decrease. This happens in live scenarios where the query and the gallery are generated by person detection models in combination with person trackers. In this case, it is hard to have good-quality images because it is possible to have occlusion, missing body parts, or changes in illumination. Even the video stream can have some noise that makes the correct detection and cropping difficult. The goal of this dissertation is to propose a way to benchmark different methods and different training datasets to see how well they perform in practical Re-ID settings.

1.2 Definition of the Problem

In the literature, the problem of Re-ID is studied under different settings depending on the application context (The different Re-ID paradigms are presented in detail in Chapter 2 - 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).

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 deployment in a city with different characteristics than the training dataset. Hence, Sumari et al. (2020) recently introduced a setting (called *live Re-ID*) considering specifically the constraints related to implementing Re-ID for use during live operations. In this previous work, we showed that training a successful Re-ID model with respect to standard Re-ID metrics does not guarantee good performance when evaluated in a specific live Re-ID context.

The first contribution of this dissertation is to better formalize the definition and constraints associated with the live Re-ID setting. We also extend the live Re-ID evaluation metrics proposed in (SUMARI et al., 2020) in order to facilitate interpretation.

1.3 Objectives

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 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 work is to answer the following questions:

1. Which characteristics of a standard Re-ID dataset (diversity, size) are most important to train standard Re-ID models for the live Re-ID setting?
2. Which standard Re-ID approaches can be successfully deployed for practical implementations in 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?

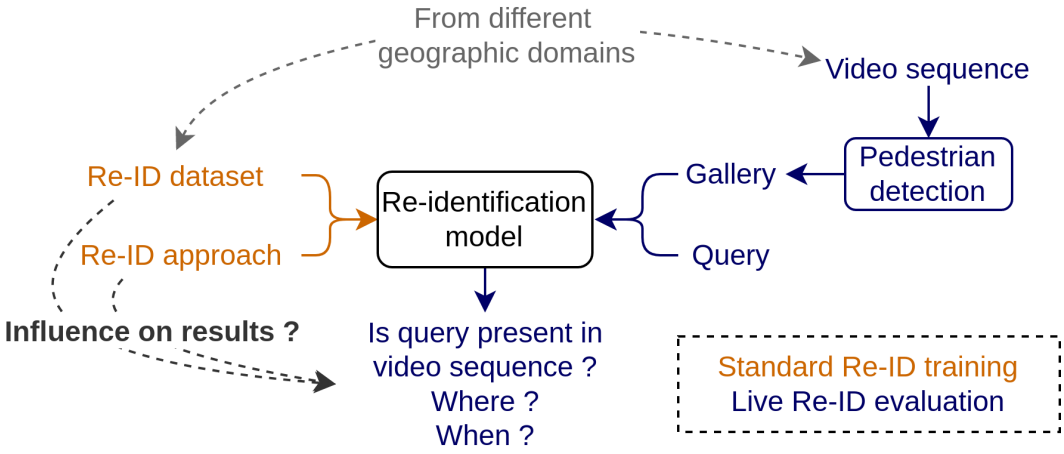


Figure 1: **Conceptual overview**. Visualization of the objectives of our benchmark study. This work aims at evaluating how different standard Re-ID approaches and training datasets behave for practical deployment in new environments, i.e., the live Re-ID setting.

We present a study using three standard Re-ID datasets and four recent standard Re-ID approaches to answer these questions. For each approach-dataset pair, the Re-ID model obtained is evaluated against the other two datasets and against a fourth live Re-ID dataset. We also try to combine training datasets to investigate how dataset size and diversity influence the generalization of the obtained standard Re-ID model. A conceptual overview of the objectives of our study is represented in Figure 1.

In this work, 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. On the one hand, the unsupervised domain adaptation problem consists in leveraging unlabeled data from the target domain to improve the performance of the standard Re-ID model (ZHAO et al., 2020; MEKHAZNI et al., 2020). There are other methods from the transfer learning field (ZHAO et al., 2020) that have been applied to fine-tune standard Re-ID models for new contexts where a small amount of labeled data is available (CHEN, H. et al., 2018).

Such domain adaptation approaches are not tested in this work. Still, 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. On another note, Xiao et al. (2017) have shown that considering bounding box extraction 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 for some configurations. Likewise, we believe that the results from our study can be 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.

1.4 Dissertation organization

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

Chapter 2

Related work

In this chapter, we will review different Re-Identification settings that are currently state of the art. Also, we present clear definitions of previous benchmark studies about Re-ID.

2.1 Person re-identification settings

The security for governmental and private organizations across the world is very important, especially in public areas (WANG, 2013). It requires financial investment and significant effort to provide it. A very common solution is the installation of security cameras at strategic points, but these cameras need to be constantly monitored by security agents or the stream must be recorded for future analysis. Most of the time, these videos are saved in raw format and the quantity of videos saved from the cameras quickly becomes huge. Suppose then, that a security agent, a police officer, or a camera owner wants to look for a person in the recorded videos. Performing this task can take a lot of time due to the difficulty to compare all people with the one you are looking for. The idea of person re-identification is to automate this process.

The field of Re-ID was first formalized by (GHEISSARI; SEBASTIAN; HARTLEY, 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.

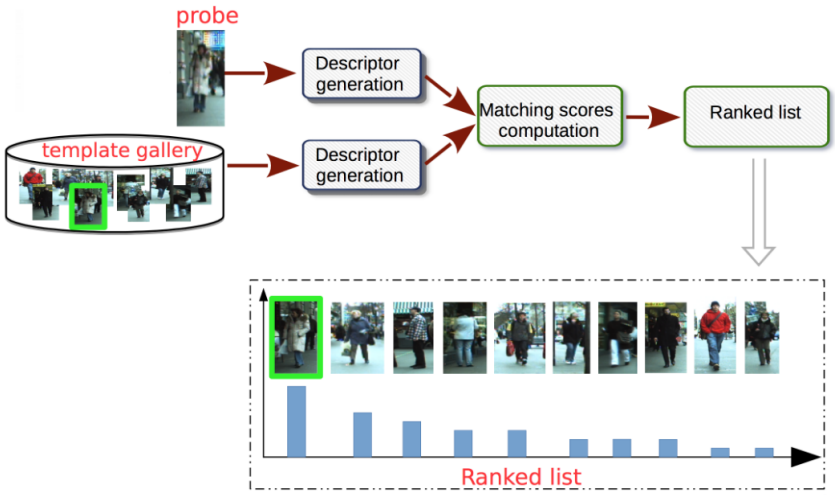


Figure 2: Typical pipeline for the standard Re-ID setting (Source: (LAVI et al., 2020)).

2.1.1 Popular settings

Different Re-Identification settings were introduced to address different problems. Here we present the most relevant settings with respect to the live Re-ID problem.

2.1.1.1 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. The first proposed solutions used hand-crafted features to process images or videos. For this reason, the number of images per query and gallery was limited (ZHENG; YANG; HAUPTMANN, 2016). It is sometimes called closed-set Re-ID as it assumes that the query person has at least one representative in the gallery. Those papers presented a pipeline where the *query image* is called a *probe*. Researchers also used different methods to get descriptors from both query and gallery after applying a matched scores computation and finally getting a ranked list. In Figure 2, we can see the well-cropped images composing the query and gallery, as inputs for a standard Re-ID approach. Also, the different steps involved in a standard Re-ID pipeline are presented: feature extraction for both inputs, matching scores computation, and finally a ranked list with the most similar at the beginning.

According to the statistics presented in (PAPERS WITH CODE. . . , 2021), this is the most studied Re-ID setting, in terms of the number of papers, datasets and benchmarks published. Some standard Re-ID datasets and successful methods are used for this study and presented in Chapter 3. For a more complete overview of standard Re-ID approaches,

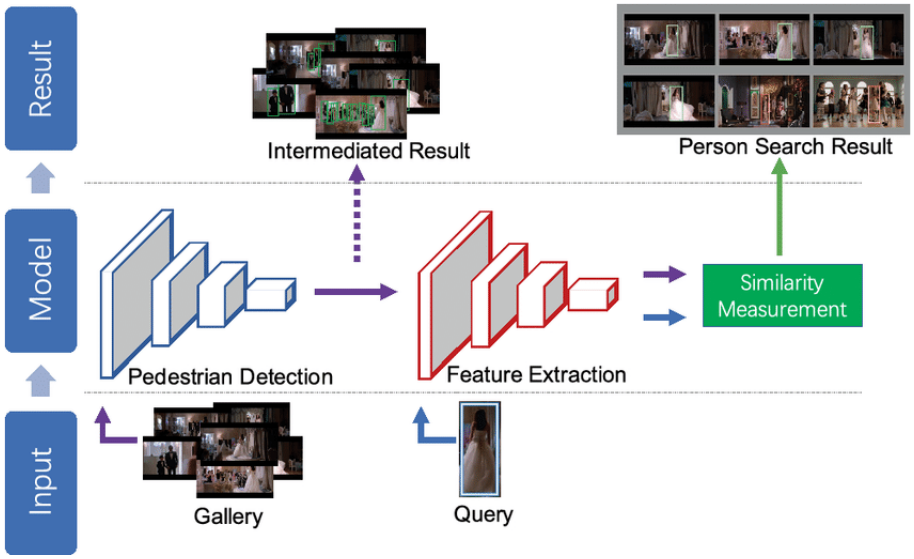


Figure 3: Typical Person Search pipeline (Source (CHEN, Z. et al., 2021)).

we refer the reader to the following surveys (LAVI et al., 2020; YE et al., 2021).

2.1.1.2 Person search

The *person search* setting was introduced in (XU et al., 2014). It consists in replacing the gallery items with 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. In this setting, there is a combination of two models one for pedestrian detection and another for person re-identification. A standard person search pipeline is achieved by combining these models.

Passing a full frame to generate automatically the gallery is the beginning of a pipeline that supports video and images as input when stopping to crop BB by hand but there is a high dependency on how fast and good the BB is made, being also possible to find occlusion between them. Actual computers with GPUs help to improve the speed of these models, and sometimes, it is possible to achieve real-time processing. In Figure 3 we can observe a common pipeline for person search. The final result differs from *Standard Re-ID* because the BBs are placed in every gallery image where the query was found. A survey about person search approaches was proposed in (ISLAM, 2020).

2.1.1.3 Open-set Re-ID

The *open-set Re-ID* setting was first defined in (LIAO et al., 2014). It differs from standard Re-ID in a way 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. This setting is one of the biggest challenges because sometimes the query isn't present in the gallery, so when this happens in the rank list, many similar people have similar clothes. There is a verification step to be sure if the person is on the list to solve this issue. The reader can refer to the survey in (LENG; YE; TIAN, 2019) for an overview of recent open-set Re-ID approaches.

2.1.1.4 Video-based Re-ID

The *video-based Re-ID* setting was first studied in (WANG, T. 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. Ye et al. (2021) proposed a complete review of video-based Re-ID.

2.1.1.5 Other Re-ID settings

For completeness, we mention the existence of other Re-ID variants in the literature, namely unsupervised Re-ID (YANG; QI; JIA, 2021), semi-supervised Re-ID (MOSKVYAK et al., 2021), human-in-the-loop Re-ID (WANG, H. et al., 2016), or federated Re-ID (ZHUANG et al., 2020). However, their specificity 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 the *live Re-ID* setting, which is inspired by our previous work (SUMARI et al., 2020). It takes into account all relevant aspects for deploying Re-ID models in practical real-world applications. An overview of the live Re-ID workflow can be seen in Figure 4.

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. Using these live Re-ID characteristics, it was recently shown that using

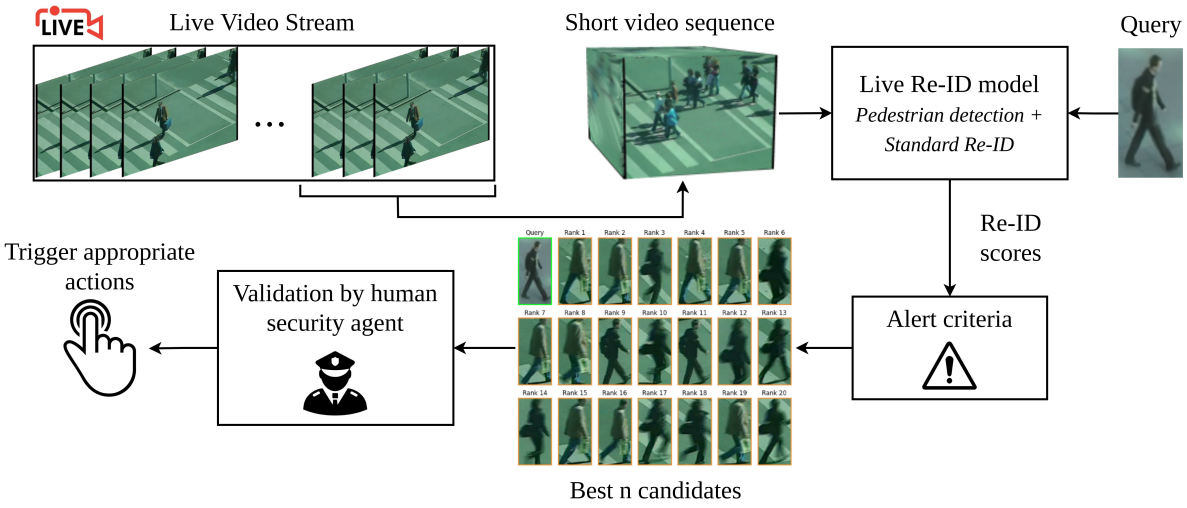


Figure 4: **The live Re-ID setting**. When deploying person re-identification models in practical applications, the galleries are composed of whole scene video sequences. When the Re-ID system raises an alert, the data are verified by a security agent to decide whether actions should be triggered.

tracking and anomaly detection to reduce the size of the generated 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. This highlights the importance of studying *cross domain* transfer of Re-ID, Luo, Jiang, et al. (2020) was the first that discussed and highlighted it.

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 triggers appropriate actions. We would like to clarify that classic Re-ID had a high Rank-1 accuracy is very important because we are performing evaluation over well build datasets and most of the time we are training and evaluating over the same dataset. But when we are evaluating in Live Re-Id setting had a high Rank-1 accuracy depending on the quality of the processed videos and also we won't train again the Re-ID approach with these new images. 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, two evaluation metrics representing both dimensions of the problem were introduced in (SUMARI et al., 2020) (see Chapter 3, Section 3.3.3). The experiments conducted in this dissertation 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 presents a 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.

Gou et al. (2018) conducted a large-scale benchmark experiment 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. They define two classifications for datasets:

1. Academic Re-ID datasets: that are composed of well-cropped images extracted by hand, such that the *probe (query)* always has at least one match image in the gallery.
2. Real-world end-to-end datasets: composed of probes and galleries that were generated automatically by person detection, tracking algorithms, and cropping. This dataset had no guarantee that for every query image, there is a match inside the gallery because the data is unlabeled.

They also built a new dataset to represent constraints relevant to real-world implementations, such as pedestrian detection errors and illumination variations, among others.

However, they do not consider cross-domain performance to be able to deploy Re-ID algorithms in new cities without having to develop context-dependent specialized datasets. Another drawback is that all evaluations are conducted in the closed-set setting, which is a major limitation regarding future deployments. To address these limitations, all important requirements for future deployments of Re-ID algorithms are taken into account in our live Re-ID evaluation experiment. In addition, Zheng, Bie, et al. (2016) proposed a smaller systematic evaluation of video-based Re-ID approaches.

Zheng, Zhang, et al. (2017) conducted another extensive set of experiments 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, Lingxiao He et al. (2020b) proposed the first benchmark regarding the cross-domain transfer of Re-ID approaches. 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 for assessing 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 work 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.

Chapter 3

Benchmark methodology

The objective of this work 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. In particular, we aim to assess the quality of different Re-ID approaches within the context of practical implementation of live Re-ID (SUMARI et al., 2020). In addition, another goal is to understand if training these approaches on different publicly available Re-ID datasets (LI; ZHAO; XIAO, et al., 2014; RISTANI et al., 2016; ZHENG; SHEN, et al., 2015; HIRZER et al., 2011) will lead to different results. This chapter 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. Figure 5 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.

3.1.1 Standard Re-ID datasets

This section presents the standard Re-ID datasets used in this study. Table 1 summarizes relevant statistics.

(a) **Market-1501**.

(b) **DukeMTMC**.

(c) **CUHK03**.

(d) **Original PRID-2011**.

(e) **m-PRID**. Bounding Boxes (BB) extracted from PRID-2011 videos using YOLO-V3 for pedestrian detection. Blue indicates good images for standard Re-ID, while red BB are likely to generate re-identification errors.

Figure 5: **Benchmarking datasets**. Example images from the datasets used in our experimental study.

3.1.1.1 Market-1501

Market-1501 was released in (ZHENG; SHEN, et al., 2015). The authors found different studies where the number of cameras and the cropped number of images taken for pedestrians was very limited. Also, in most of these datasets, the pedestrians were aligned to hand-drawn boxes and when pedestrian detectors are applied over them, the results weren't so good because pedestrians weren't aligned. In some cases, there were missing body parts. To address these issues, they built Market-1501.

Market-1501 was collected at a supermarket in Tsinghua University, Beijing, China. The training set is composed of 12,936 images of 751 different identities, and the testing set contains 3368 query images and 15,931 gallery images of 750 identities. 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. As illustrated by Figure 5a, cropped images appear to present a high level of details about the people represented (i.e., images are taken from a close perspective or videos are high resolutions). Lighting conditions in this dataset are also good to distinguish specific features.

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

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

3.1.1.2 DukeMTMC

DukeMTMC (Multi Target, Multi Camera) was released in (RISTANI et al., 2016). It was collected at the Duke University campus, Durham, North Carolina, USA. This data set contains more than 2,700 identities extracted from 8×85 minutes of $1080p$ videos recorded at 60 frames per second.

All the cameras were static and deployed over the campus when pedestrian traffic was heavy. The training set is composed of 16,522 images of 702 identities, and the testing set contains 2,228 queries and 17,661 gallery images of 702 other identities. In addition, 408 distractor identities are included in the test gallery.

Another difference with single-camera benchmarks was the persistent tracking across different cameras, where there is a group of 891 persons that walk only for one camera and challenge the tracker to determine if there are false or true positives. Analyzing and processing all the videos saved took in some cases up to six days in one computer to analyze background masks and seven days to generate all person detection on a cluster of 192 cores using Deformable Part Model (DPM). The Bounding Boxes (BB) in DukeMTMC are semi-automatically generated but all annotations of identities are conducted by hand. As shown in Figure 5b, lighting conditions are good but the resolution of the BB is relatively low.

3.1.1.3 CUHK03

CUHK03 was introduced in (LI; ZHAO; XIAO, et al., 2014), using video footage collected at the campus of the Chinese University of Hong Kong. They generated images cropped

by hand and also used some state-of-the-art techniques to crop and save BB automatically.

The authors found some problems that were rarely reported for other datasets such as misalignment, body parts missing, and occlusions. Also, they used more than two cameras to generate the BBs, and finally, they recorded the videos over months so there are identities that are the same across the time the videos were recorded. The time and climate conditions changed the illumination and shadows over the videos recorded.

Each identity was associated with 4.8 images on average. In our work, we used the manually labeled version of the BB. Cropped images are high resolution but illumination is dark, which reduces image quality (Figure 5c). The BBs were cropped from six surveillance camera videos, where there are 1360 IDs, and it includes 13164 cropped images. There were two previous versions of this dataset named CUHK01 (LI; ZHAO; WANG, 2013) and CUHK02 (LI; WANG, 2013). We use the newest version of it: CUHK03. Inside this dataset, there are two folders for cropped images hand-labeled and another detected automatically. As explained above, we use a hand-labeled folder in our experiments.

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., 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, the 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 gathers information about each person it contains (identifier, frames where it appears, bounding box coordinates). For evaluation, a total of 73 queries are considered.

To better grasp the influence of the pedestrian detection model, we also evaluate our different models on the original PRID-2011 dataset for standard Re-ID. Figure 5d shows cropped images of poor resolution, taken from relatively high camera angles compared to other datasets. This way, we can see if the performance decrease on the live Re-ID setting

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

is due to the domain shift of PRID or to the pedestrian detector inaccuracies (Figure 5e).

3.2 Re-ID evaluated approaches

This work studies the performance of four successful standard Re-ID approaches. To complement previous benchmark studies (Chapter 2, Section 2.2). Only very recent approaches are selected for this work. Our experimental results should help us understand which neural network architecture design choices are most important for domain adaptation and generalization to the live Re-ID setting.

3.2.1 Bag of Tricks (BoT)

The *Bag of Tricks* approach proposed in (LUO; GU, et al., 2019) resulted from the observation that previous works were expanded on poor state-of-the-art approaches, only two in twenty-three of them surpassed 90% rank-1 accuracy on Market1501 Dataset. Most improvements for these approaches come from neural network training tricks rather than Re-ID approaches themselves. They found that some them were unfairly compared because the improvements were in the training stage rather than the method.

Also, the industry prefers working pipelines but sometimes effective models come after concatenating many local features in the inference stage. Researchers use additional information to have more discriminative features, but these in-production scenarios may have extra work for the model and it'll take more time to process all this additional information. 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, K. et al., 2015). In particular:

1. Initialize ResNet-50 backbone with weights pre-trained 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 where for every 4 persons there are 16 images,
4. Images are resized to 256 X 128 and pad images with 10 pixels with zero values,
5. The images are flipped horizontally with 0.5 probability,
6. Both the model output features and prediction logits are used respectively to compute triplet loss and cross-entropy loss are used, and

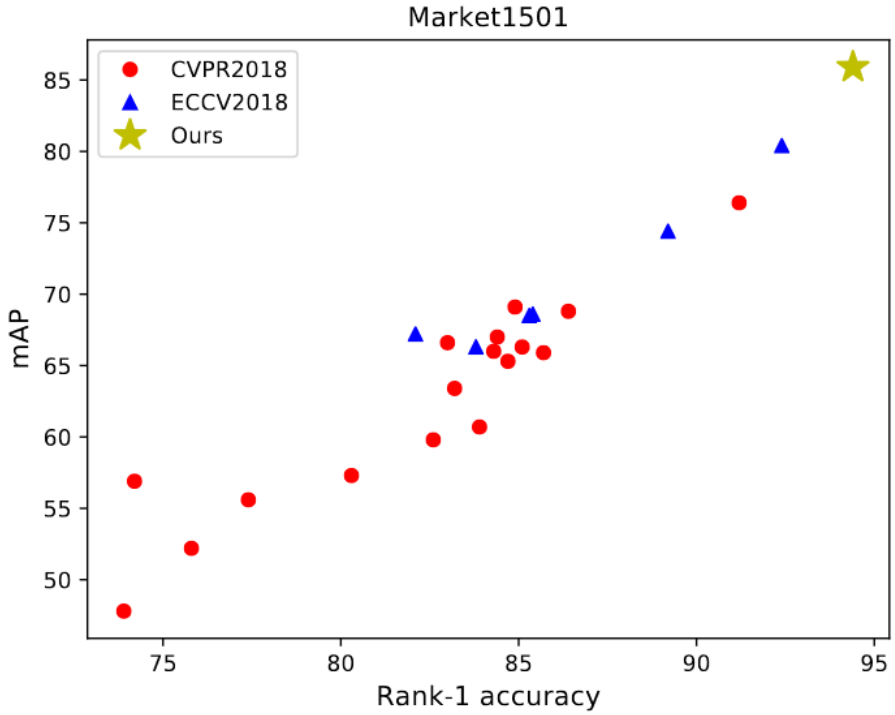


Figure 6: Performance of strong baseline, compared with other approaches. (Source (LUO; GU, et al., 2019)).

7. Adam is used for optimization. (initial learning rate: 3.5×10^{-4} and 120 epochs).

In Figure 6 we can see that the architecture proposed by the authors achieves great results over the different approaches evaluated.

3.2.2 Strong Baseline and Batch Normalization Neck (SBS)

The approach named *Strong Baseline and Batch Normalization Neck* is a extended version of (LUO; JIANG, et al., 2020). They proposed a baseline over ResNet-50 that had similarities with other state-of-the-art approaches. The global aim of this baseline is to add different strategies to improve the training without changing the model architecture. Here we'll present the different strategies proposed:

1. **Warmup Learning Rate:** applied a warmup strategy due that the learning rate had a great impact on the Re-ID model through a standard baseline that was trained with a constant and large learning rate.

2. **Random Erasing Augmentation:** When analyzing a video there, occlusion is a very recurrent problem that is inherent to people's behavior when changing the

Figure 7: Examples of random erasing augmentation. The first row shows five original images and the processed images are presented in the second row. (Source (LUO; GU, et al., 2019)).

walking direction without advising other people or when there may be objects in front of them. To address this problem and improve the generalization of Re-ID models (ZHONG et al., 2017) proposed a new data augmentation technique named Random Erasing Augmentation (REA). They introduced the quantity p_e, which is the probability to apply REA in the image. When REA is applied to an image, a rectangular area is selected randomly to be erased in the image. In Figure 7 there is an example of different images after REA was applied to them.

3. **Label Smoothing:** In most Re-ID architectures, the last layer is a fully-connected layer with a hidden size equal to the number of persons N. The person ID determines the category of the classification, so the authors call this the loss function. When there aren't images of the test set in the training set, person Re-ID can be regarded as a one-shot learning task and generate overfitting. Label Smoothing (LS) proposed in (SZEGEDY et al., 2015) is used to reduce overfitting.

4. **Last Stride:** To increase the size of the feature map, (SUN et al., 2018) removed the last spatial down-sampling from the backbone network. The authors named this as the last stride, where they change the last stride from 2 to 1, so they can get a feature map with a higher spatial size (16×8). Finally, they show that such a higher spatial resolution brings significant improvement.

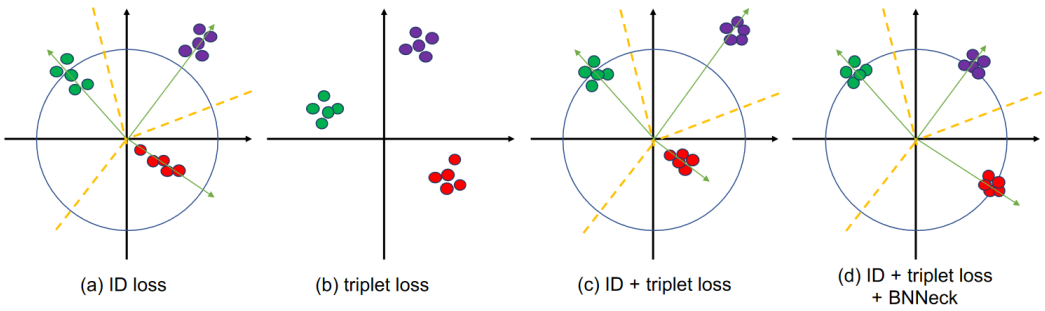


Figure 8: Two-dimensional visualization of sample distribution in the embedding space supervised by (a) ID Loss, (b) Triplet Loss, (c) ID + triplet loss and (d) ID + triplet loss + BNNeck. Points of different colors represent embedding features from different classes. The yellow dotted lines stand for the supposed classification hyperplanes. (Source (LUO; GU, et al., 2019)).

5. **Batch Normalization Neck:** Most of the Re-ID models combine ID loss and triplet loss together for training. However, when this combination is analyzed, there are two problems: first ID loss constructs several hyperplanes to separate the embedding space into different subspaces; second, triplet loss enhances the intra-class compactness and interclass separability in the Euclidean space. Combining both losses generate a possible phenomenon that one loss may be reduced, while the other loss is oscillating or even increases. Hence, the authors designed a structure named Batch Normalization Neck (BNNeck) which adds a batch normalization (BN) layer after the feature extraction layer (and before the classifier FC layer). In Figure 8 there is a representation in two-dimensional visualization of sample distribution in the embedding supervised space. For example, Figure 8(a) shows how the ID loss constructs hyperplanes to separate the embedding space into different subspaces, in Figure 8(b) we can see the intra-class compactness and inter-class separability produced by triplet loss. Figure 8(c) shows the commonly used combination of ID loss and triplet loss. The addition of BN Neck leads to easier convergence for the triplet loss and the ID loss and increases the inter-class separability.

6. **Center Loss:** The center loss, which simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers, makes up for the drawbacks of the triplet loss. Figure 9 shows the final baseline with the selected training tricks proposed in (LUO; GU, et al., 2019).

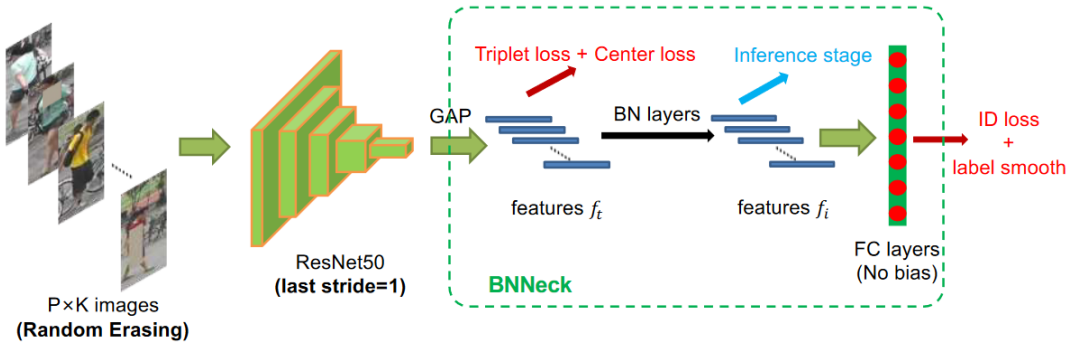


Figure 9: Strong Baseline and Batch Normalization Neck (SBS) architecture proposed. (Source (LUO; GU, et al., 2019)).

3.2.3 Attention Generalized mean pooling with Weighted triplet loss (AGW)

After an extensive analysis, Ye et al. (2021) summarizes that most of the Re-ID models had achieved a rank-1 performance better than humans. The most common dataset used in the state-of-the-art methods is Market1501. Its performance can also be improved using re-ranking or metric fusion. Most of the state-of-the-art methods developed recently adopt the features aggregation paradigm, combining the part-level and full human body features. After analyzing attention schemes in some methods, attention captures the relationship between different convolutional channels and different body parts/regions, which is important for discriminative Re-ID model learning. A different combination of loss is also a good strategy to improve the Re-ID learning stage.

After this analysis, *Attention Generalized mean pooling with Weighted triplet loss* was developed in (YE et al., 2021). They present a complete review of the latest approaches and datasets, also propose a new metric evaluation based on an Inverse Negative Penalty(INP) and a new method, called Attention Generalized mean pooling with Weighted triplet loss(AGW). It was also designed on top of BoT (LUO; GU, et al., 2019) using ResNet-50 as a backbone for their implementation. They also added three major improved components described hereafter.

1. **Non-local Attention Block:** After the authors did experiments with nineteen methods over four datasets, they found that part/global and attention feature learning had an influence on the final Re-ID discriminative process. They adopted a powerful non-local attention block (WANG, X. et al., 2018) to obtain the weighted sum of the features at all positions, represented by:

$$Z_i = W_z * \phi(X_i) + X_i \tag{3.1}$$

Figure 10: Sample of visual MINP. (Source (YE et al., 2021)).

Where W_z is a weight matrix to be learned, $\phi(.)$ is a non-local operation, and X_i is a residual learning strategy.

2. **Generalized-mean (GeM) pooling:** Ye et al. (2021) proposed a learnable pooling layer replacing max and average pooling to capture the domain-specific discriminative features and adopted a layer named generalized-mean (GeM) pooling (RADENOVIĆ; TOLIAS; CHUM, 2018), which learns in the back-propagation process. Finally, the use of weighted regularization triplet loss inherits the advantages of relative distance optimization between positive and negative pairs without introducing additional parameters to the architecture.

3. **mINP: A New Evaluation Metric for Re-ID:** This paper (YE et al., 2021) also introduced a new Re-ID evaluation metric. The closest images to the query should have the lowest rank within the Re-ID model outputs. Also, there should be at least one correct match in the final output list. When there is more than one correct match to the query in the list, the last match position can help to compare the results from two different Re-ID models. For example, Figure 10 present two rank list where green boxes are correct matches, red ones are wrong matches and there are only three correct matches. Evaluating Cumulative Matching Characteristics (CMC), both rank list gets 1 because in Rank-1 they had a correct match with the target. When the Average Precision is calculated, list 1 obtains AP=0.77 while list 2 obtains AP=0.70. The proposed mINP metric suggests that list 2 is actually better than list 1 because the last true positive has a lower rank (see subsubsection 3.3.1.3)

4. Finally, in Figure 11 we can see the final architecture proposed for AGW, with the ResNet-50 Backbone, where there are Non-local Attention layers, GeM layer, and WRT layer.

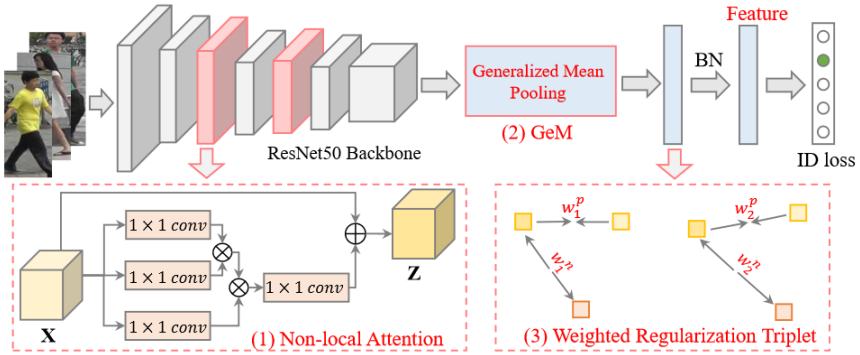


Figure 11: Final Attention Generalized mean pooling with Weighted triplet loss (AGW) proposed architecture. (Source (YE et al., 2021)).

3.2.4 Multiple Granularities Network (MGN)

The MGN is an intuitive idea for extracting features from a person inside an image that represents it. It could be made by using all body parts to have a more complete representation, but it is not necessarily the best option. Summarizing this information to have a unique representation may cause some unusual or hard discriminative features that could be lost in this process. Having this point in mind, (WANG, G. et al., 2018) presents the *Multiple Granularities Network* to represent local features as necessary to find significant body parts, which is an interesting approach to improve Re-ID accuracy. Locating these body parts contains a small percentage of information in relation to the complete body also at the same time noise around these regions is filtered by local operations. These part-based methods according to their part location can be divided into three:

1. Based on empirical knowledge about the human body.
2. Using region proposal methods to locate partial regions.
3. Enhancing features by middle-level attention on salient body parts.

Unfortunately, some limitations reduce the effectiveness of these methods for example occlusion or pose variations. Some methods focus only on specific parts and most of the methods aren't end-to-end learning processes.

Guanshuo Wang et al. (2018) proposed the combination of global and local information but with different granularity. They define the Global branch as containing only one whole partition and when they strip it into different numbers it generates a set of body part images. The granularity depends on how many strips of the image are done; features of

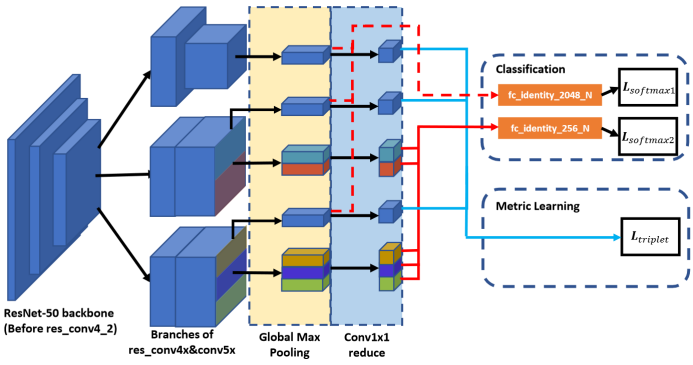


Figure 12: Architecture of Multiple Granularities Network (MGN). (Source (WANG, G. et al., 2018)).

local parts can concentrate more on finer discriminative information when the number of partitions increases.

The design of MGN was based on the idea of a multi-branch network architecture divided into one global and two local branches based on ResNet-50 backbone. Comparing with other techniques this architecture had better results than other part-based methods.

3.2.4.1 Network Architecture

The backbone of the MGN network is a ResNet-50 which helps to achieve competitive performances in some Re-ID models. After the *res_conv4_1* block, the authors propose to divide it into three independent branches:

1. Global Branch: They propose this branch to learn the global feature representations without any partition information and also they employ down-sampling with a stride-2 convolution layer following a global max-pooling operation on the corresponding output feature map.

2. Part-N Branch: They had the same idea of the Global branch but instead of applying a down-sampling they uniformly split into several stripes horizontally. N is the number of splits for this architecture and can vary from 2 to 3.

In Figure 12, we can see the proposed architecture by (WANG, G. et al., 2018). After the *res_conv4_1* residual block, the ResNet-50 backbone is split into three branches: Global Branch, Part-2 Branch, and Part-3 Branch. The reduced features are concatenated together as the final feature representation of a pedestrian image.

The improvements presented by authors about loss employed a combination of Softmax loss for classification and tripĺet loss for metric learning.

3.3 Proposed experiments

To compare the Re-ID datasets and approaches presented above, several experiments are proposed and 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. In figure13 we can see an example of this proposed experiment where we train on *CUHK03* and evaluate on the same dataset our four approaches used in this study. 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:

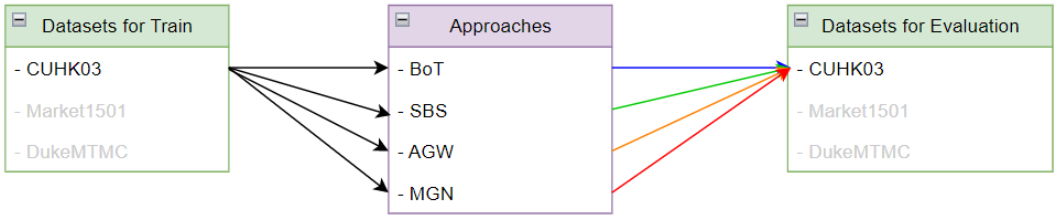


Figure 13: Example of Single dataset experiment.

3.3.1.1 Rank-n

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.

3.3.1.2 mAP

The computation of *mean average precision* for Re-ID takes into account the predicted ranks of all existing matches (ZHENG; SHEN, et al., 2015). 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. mAP is defined in equation 3.2:

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{3.2}$$

Where n is the number of queries in the set and AP_i is the average precision for a given query,

3.3.1.3 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. mINP is defined in equation3.3:

$$mINP = \frac{1}{n}\sum_{i} \frac{|G_i|}{R_i^{hard}} \tag{3.3}$$

Where R_i^{hard} indicates the rank position of the hardest match, G_i represents the total number of correct matches for query i.

These three metrics represent different skills of the evaluated Re-ID model. Computing them might help understand which of these skills are 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

After analyzing the classic Re-ID, we evaluate the different combinations of using one or more datasets for training and evaluating on another not used.

3.3.2.1 Simple cross-dataset evaluation

The simple cross-dataset experiment from (HE, L. et al., 2020b) 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). In figure14 we can see an example of this proposed experiment where we train on *Market1501* and

evaluate on the others dataset(DukeMTMC, CUHK03 and PRID) the four approaches used in this study.

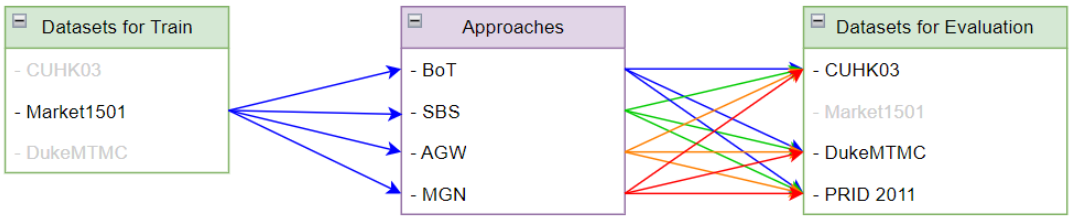


Figure 14: Example of Simple cross-dataset experiment.

As the datasets were built in different geographic areas, these results can give first insights into 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 a 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 of dataset-approach pairs.

3.3.2.2 Combine cross-dataset evaluation

For the cross-datasets experiments, we also try to combine training datasets to see if it improves test performance. Here we'll present some combinations proposed:

1. In the $\text{COMBINED}_{\text{all}}$ experiments, 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 evaluating if adding data from other sources can help improve standard Re-ID in the traditional supervised setting.

2. In the $\text{COMBINED}_{\text{others}}$ experiments, 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, in figure15 we can see an example of this proposed experiment where we train on $\text{COMBINED}_{\text{all}}$ and evaluate on all datasets (Market1501, DukeMTMC, CUHK03 and PRID) the four approaches used in this study. Also in figure16 we observe a combination of $\text{COMBINED}_{\text{others}}$ for train and evaluate.

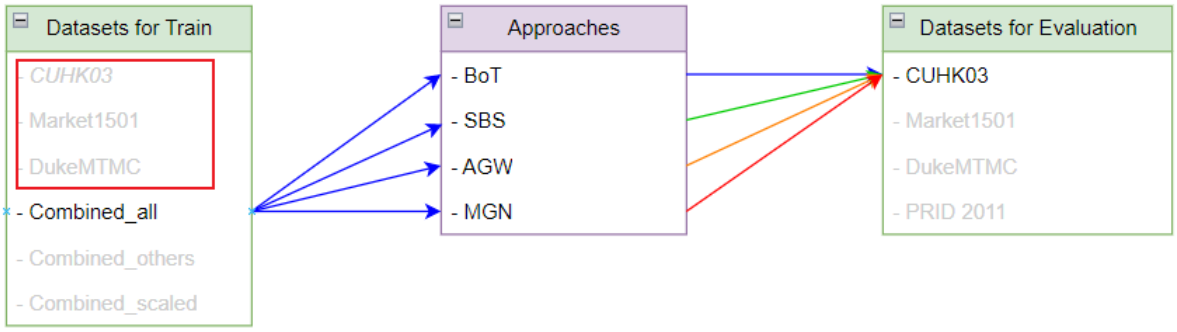


Figure 15: Example of $\text{COMBINED}_{\text{all}}$ experiment, in this sample the datasets in red box are the ones combined for $\text{COMBINED}_{\text{all}}$.

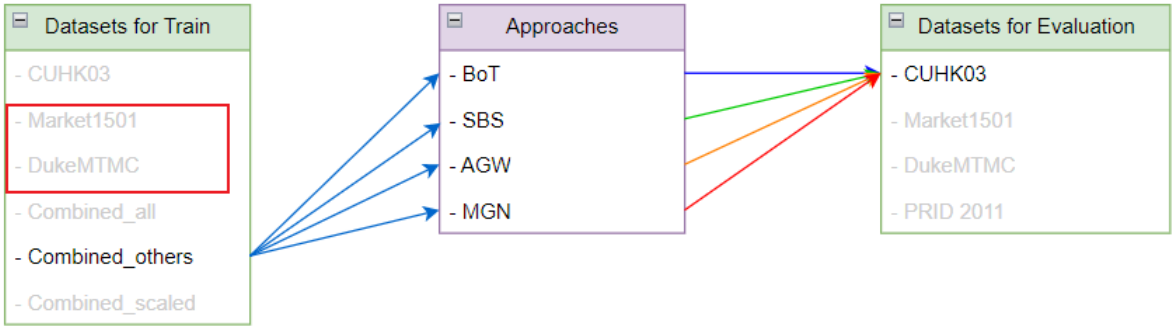


Figure 16: Example of $\text{COMBINED}_{\text{others}}$ experiment, for this case the datasets in red box are the ones combined for $\text{COMBINED}_{\text{others}}$.

3.3.2.3 Scaled combine cross-dataset evaluation

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. For example, when combining datasets A and B, respectively of size N_A and N_B, we only take fractions N_A^* and N_B^* of each datasets such that $N_A^* + N_B^* = \max(N_A, N_B)$ and $N_A^* = N_B^*$.

Comparing $\text{COMBINED}_{\text{scaled}}$, with $\text{COMBINED}_{\text{others}}$ allows us to compare the influence of dataset size and diversity in the generalization power of a dataset. In figure17 we can see an example of this proposed experiment.

The different sizes of combined datasets used in this work:

- When evaluating on CUHK03, the $\text{COMBINED}_{\text{scaled}}$ dataset is composed of 8261 images from DukeMTMC and 8261 from Market-1501.
- When evaluating on DukeMTMC it contains 6468 images from both CUHK03 and Market-1501.

- When evaluating on Market-1501 it contains 9754 images from DukeMTMC and 7368 images from CUHK03.
- Finally, when evaluating on PRID-2011, $\text{COMBINED}_{\text{scaled}}$ is composed of 5507 images from CUHK03, 5508 from DukeMTMC and 5507 from Market-1501.

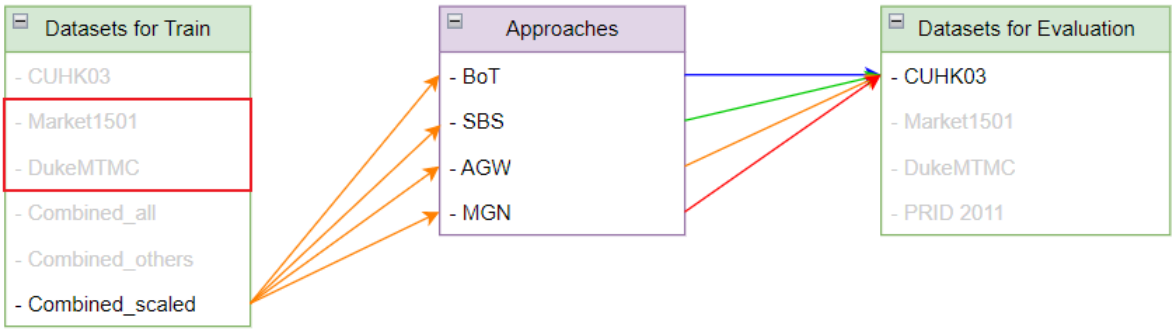


Figure 17: Example of Scaled combine cross-dataset experiment.

For evaluations on PRID-2011, which is not among the training datasets, $\text{COMBINED}_{\text{all}}$ and $\text{COMBINED}_{\text{others}}$ are identical and referred to as COMBINED.

3.3.3 Live Re-ID evaluation

This experiment aims to see if the best approaches and datasets from previous experiments are also the best from the perspective of practical implementation in new cities. In figure18 we can see an example of this proposed experiment where we train on *Market1501* and evaluate on *mPRID* dataset the four approaches used in this study.

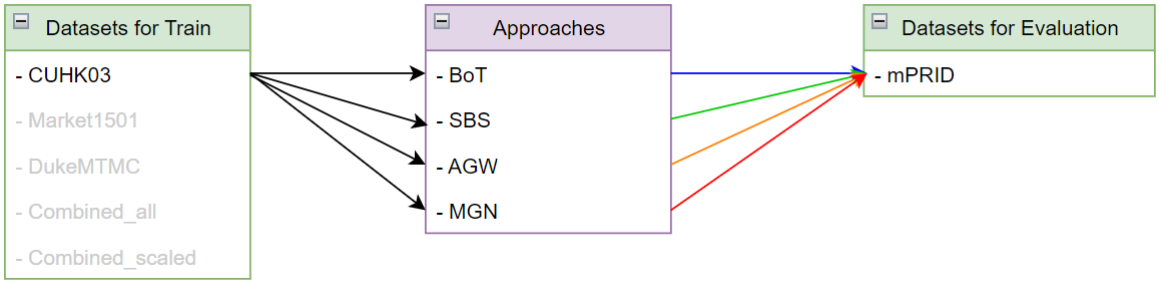


Figure 18: Example of Live Re-ID experiment. We only select five datasets from all combinations for train and evaluate on *mPRID* .

Combined datasets experiments are also conducted for the live Re-ID setting. However, as PRID-2011 is not one of the training datasets used in our experiments, as follows:

- $\text{COMBINED}_{\text{all}}$ and $\text{COMBINED}_{\text{others}}$ are actually equivalent here and simply referred to as COMBINED.
- They are also compared against $\text{COMBINED}_{\text{scaled}}$ results to study the impact of dataset size and diversity.
- The $\text{COMBINED}_{\text{scaled}}$ training dataset for live Re-ID experiments on m-PRID are composed of 5507 images from CUHK03, 5508 from DukeMTMC and 5507 from Market-1501.

Finally, we introduce the metrics used to evaluate the Live Re-ID setting and propose two more metrics to better understand them.

3.3.3.1 Live Re-ID metrics for evaluation

Each standard Re-ID approach and dataset pair is 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, Bounding Boxes (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 (MARTIéN 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 BBs. 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 the 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.

To compare the different models, we use the live Re-ID metrics introduced in (SUMARI et al., 2020):

1. *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.
2. *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, which can be problematic when many cameras need to be monitored simultaneously.

Sumari et al. (2020) defined FR equation 3.4 and TVR equation 3.5 as follows:

$$FR = \frac{TC}{TC + TMC + FS} \tag{3.4}$$

$$TVR = \frac{TC}{TC + TMC + FC} \tag{3.5}$$

Where TC is True Call, TMC is True Missed Call, FC is False Call and FS is False Silence.

3.3.3.2 Proposed metrics for evaluation

In this work 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.

1. 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.

2. The second unified metric consists in computing a weighted harmonic mean of FR and TVR, similar to the F-score computation for precision and recall. We call the resulting metric F_γ, which is defined as follows:

$$F_\gamma = (1 + \gamma^2).\frac{\text{FR.TVR}}{(\gamma^2.\text{FR}) + \text{TVR}}. \tag{3.6}$$

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; PAULA CANUTO; GONCALVES, 2020), consisting in evaluating a model by its performance at the optimal configuration.

The result is called *optimal* F_γ (F_γ^*), 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 obtained by quick experiments in the practical implementation

context. An F_γ^* score of 1 means that there exists a Re-ID threshold β such that it always finds the query when it is in the video sequence, but never raises alerts when it is not.

3.3.4 FastReID

The FastReID(HE, L. et al., 2020a) toolbox has become one of the open-source projects of Jingdong Artificial Intelligence Research (JD AI Research). It is a research/open-source project for academia and industry.

We used this framework to build the approaches and also to train and evaluate over the datasets. Finally, we adapt it to use in our Live Re-ID setting and perform the evaluation proposed in this study.

Chapter 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: A.1 contains all the results from cross-dataset evaluation, A.2 contains all the results from combined cross-dataset evaluation, A.3 contains all the results from scaled cross-dataset evaluation, A.4 contains the missing metrics and the TVR vs FR curves for live Re-ID evaluations. Overall, the curated results presented in the core chapter are representative of the complete results and are sufficient to draw our conclusions.

4.1 Single dataset results

The results for single dataset evaluation are reported in Table 2. The results obtained were rank-1 and mAP are around 70% for the worst approach on the most difficult dataset. We used the architecture without modifications and compared our results with each paper where we obtain the same results.

Comparing the different metrics shows that MGN obtains better results in every dataset. Using BoT backbone obtains the worst results in comparison to SBS and that's what we expected because the backbone is the same but the training tricks are only applied to SBS. All the results agree with the results reported on each respective paper. Finally, the results obtained show that the different approaches generalize very differently to new contexts.

Table 2: **Single dataset evaluations**. Results were 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

4.2 Simple Cross-dataset results

The simple cross-dataset results are presented in Table 3. We evaluate every approach, using classic metrics such as Rank-n, mAP, and mINP, over four datasets: one for training and three for testing but in this experiment, we didn't use the PRID dataset for training in any combination.

When we are analyzing different standard approaches over classic datasets makes a lot of sense to compare them focusing on Rank-1 metric because we want to find the best Re-ID model also these approaches were developed to have the best results in the datasets. But when we begin to shift domain training in dataset A and evaluate in dataset B, the Rank-n results of the approach will decrease, that is why we decided for a future analysis focus on Rank-10. Finally, the complete results show that the ranking of approaches is stable under different values of n. But having a high Rank-10 for live Re-ID is more important than lower ranks, we explained it also in Section 2.1.2.

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 3). We observe one more time the importance of choosing a good training dataset.

As we explained, before using PRID(HIRZER et al., 2011) only for evaluation, as we can see in Table 3 the best approach is MGN and there is an influence in the dataset used

for training.

Table 3: **Cross-dataset evaluations**. The results are 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	0.08	**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

4.3 Combine cross-dataset results

We conducted a complete pair cross-dataset experiments to see if it improves test performance. In this experiment, we select two of the three datasets for training and evaluation in the other one and also over PRID-2011 this is $\text{COMBINED}_{\text{others}}$. For example, when we are performing evaluation on CUHK03 and PRID we train on Market-1501, DukeMTMC.

Also, we propose $\text{COMBINED}_{\text{all}}$ experiments, 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 evaluating if adding data from other sources can help improve standard Re-ID in the traditional supervised setting. We present in Table 3 the different combinations explained before. Here we found a clue about the best option for

training datasets, as we observe the results after training with DukeMTMC and CUHK03 and evaluate over PRID the approach MGN gets similar results as $\text{COMBINED}_{\text{all}}$ presented in Table 3.

Finally, as we can see in Table 3, combining datasets improves the R10 and mAP in all approaches in the evaluation.

4.4 Scaled combined cross-dataset results

We also try to scale the size of the combined datasets used in Section 4.3. These results are presented in Table 3. Here we observe that training on $\text{COMBINED}_{\text{SCALED-all}}$ dataset and evaluating in PRID-2011 had better results in Rank-10 over 46% and mAP over 36% than the same dataset without scaling as shown in Table 3. There is a similar behavior when we train on DukeMTMC & CUHK03 $_{\text{SCALED}}$ dataset and evaluate over Market-1501, where we had a similar score in Rank-10 over 92% than the same dataset without scaling as shown in Table 3.

As an initial analysis of these results, we think that when we perform live Re-ID experiments the best results will be $\text{COMBINED}_{\text{all}}$ for training and with the MGN approach. Both results lead to over 45% in Rank-10. We also observe that MGN evaluated on Market-1501 had the best result in Rank-10 with 93%. It is possible that the best evaluation results on Market-1501 are due to the quality of images provided by this dataset.

4.5 Live Re-ID results

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.

Table 4: **Live Re-ID evaluation**. Results were 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		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**

Chapter 5

Discussion

In this chapter, we propose to discuss the results achieved in our proposal, highlighting several key insights regarding training standard Re-ID models for cross-domain live Re-ID.

5.1 Impact of the training dataset

Proper selection of the training dataset clearly influences the results obtained in a different evaluation domain. However, there is no clear winner between Market-1501 and DukeMTMC to know which individual dataset should be used for any context. In addition, the cross-dataset results do not allow us to choose the best individual 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). 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.

5.1.1 Can data from a different domain improve results in the standard Re-ID scenario ?

To answer this question, we compare results from Table 2 and the $\text{COMBINED}_{\text{all}}$ rows in Table 3. Overall, for both Rank-10 and mAP, the results for $\text{COMBINED}_{\text{all}}$ appear slightly better than the results obtained when learning only on the training set of the evaluated dataset. To illustrate this, we computed the mean and standard deviation across all evaluation datasets and approaches. When using only the training set we obtain the following results: $R10 = 0.962 \pm 0.026$ and $mAP = 0.798 \pm 0.068$. When

combining the three available datasets for training we have: $R10 = 0.964 \pm 0.022$ and $mAP = 0.805 \pm 0.067$.

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 $\text{COMBINED}_{\text{all}}$ are statistically significant. The p-values obtained are 0.2750 for R10 and 0.2313 for mAP, suggesting that the Null Hypothesis cannot be rejected, i.e., *we cannot conclude that using more data from a different domain is beneficial to the standard Re-ID training process.*

5.1.2 Between dataset size and diversity, which is most important for cross-domain generalization ?

The first question we want to answer is whether combining datasets from different domains can help cross-domain generalization. To evaluate this, we can compare the results for $\text{COMBINED}_{\text{others}}$ (COMBINED for PRID-2011) against the results from the best individual dataset in Table 3. The mean and standard deviation across all evaluation datasets and approaches are $R10 = 0.509 \pm 0.242$ and $mAP = 0.224 \pm 0.099$ for the best individual dataset, and $R10 = 0.580 \pm 0.241$ and $mAP = 0.297 \pm 0.124$ when combining all the available training datasets (except the one corresponding to the evaluated test set). The Paired Sample T-Test gives p-values of 0.0001 for both R10 and mAP, which is extremely statistically significant. In other words, our experiments confirm that *combining several training datasets from different domains allows us to train Re-ID models that generalize better to new unknown domains.*

We then want to know if simply increasing the diversity in the training dataset without increasing its size also helps for cross-domain generalization. To evaluate this, we can compare the results for $\text{COMBINED}_{\text{scaled}}$ against the results from the best individual dataset in Table 3. As a reminder, $\text{COMBINED}_{\text{scaled}}$ consists in building a training dataset by taking data from all available training sets (except the one corresponding to the evaluated test set) in such a way that the total number of training data does not exceed that size of the largest individual training set. The mean and standard deviation across all evaluation datasets and approaches are $R10 = 0.509 \pm 0.242$ and $mAP = 0.224 \pm 0.099$ for the best individual dataset, and $R10 = 0.555 \pm 0.245$ and $mAP = 0.279 \pm 0.125$ for $\text{COMBINED}_{\text{scaled}}$. The Paired Sample T-Test gives p-values of 0.0001 for R10 and 0.0005 for mAP, which is statistically significant. In other words, 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 unknown domains.

In view of the two encouraging results presented above, we now want to know whether the size of the training dataset is actually helping cross-domain generalization or if adding diversity is actually sufficient. To evaluate this, we can compare the results for $\text{COMBINED}_{\text{others}}$ against the results for $\text{COMBINED}_{\text{scaled}}$ in Table 3. The mean and standard deviation across all evaluation datasets and approaches are $R10 = 0.580 \pm 0.241$ and $mAP = 0.297 \pm 0.124$ for $\text{COMBINED}_{\text{others}}$, and $R10 = 0.555 \pm 0.245$ and $mAP = 0.279 \pm 0.125$ for $\text{COMBINED}_{\text{scaled}}$. The Paired Sample T-Test gives p-values of 0.0020 for R10 and 0.0075 for mAP, which is statistically significant. In other words, our experiments confirm that *adding more data from domains that are already present in the training set helps generalization to new unknown domains.*

5.2 Live Re-ID results

The live ReID results on m-PRID (Table 4) confirm the conclusions drawn from the cross-dataset experiments. In particular, the $\text{COMBINED}_{\text{scaled}}$ results appear better than the results with a single training set, suggesting the importance of training data diversity for practical live Re-ID implementation in a new context. The COMBINED results are themselves better than $\text{COMBINED}_{\text{scaled}}$, which suggests that one should use all the available data to train a good standard Re-ID model for live Re-ID implementation. Finally, we emphasize the good results obtained by training MGN on the COMBINED training dataset. These results are very encouraging after the pessimistic results reported in (SUMARI et al., 2020) for live Re-ID.

5.3 Impact of the standard Re-ID approaches

All the approaches tested in this study performed 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 the different body parts. In view of our results, this property appears to be desirable for generalization to the live Re-ID setting.

Besides MGN, the SBS approach also appears to present much better generalization than its other two 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, as described in section 3.2.

Finally, we acknowledge that training a classic Re-ID approach with a high quantity of images from different datasets prepares the approach to have better results when it's part of a Live Re-ID setting. Our final results in section 5.2 present an insight that we can improve our original live Re-ID setting. In this study we didn't develop an more complex end-to-end architectures for live Re-ID but we define an initial guide to select and train a Re-ID approach. We'll continue working on the idea of improve our initial trained approach and improve it.

Chapter 6

Conclusion

6.1 Overview

This dissertation 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 in this setting. We also conduct simple cross-dataset experiments to see if they can be used to predict which datasets and approaches will generalize better to the live Re-ID setting.

The main conclusions from this study can be summarized as follows:

1. Although very pessimistic results were reported in (SUMARI et al., 2020), our experiments showed that it is possible to obtain much better live Re-ID pipelines by properly choosing the standard Re-ID model and combining publicly available training datasets.

2. Proper choice of the standard re-ID approach and training dataset can influence greatly the results when transferring the model to the cross-domain 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 improving 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.

Although we only studied the straightforward transfer strategy without fine-tuning, we believe the results presented here can serve as a good starting point to develop better live Re-ID models in the future.

We are providing a repository in GitHub where there are the different configurations used in this study using FastReID also we are adding the Live Re-ID pipeline with the modifications made to adapt the toolbox in the pipeline and perform the evaluation. Link: https://github.com/josemiki/person_reid_FF_PRID_2022

6.2 Future work

The outputs of this study suggest several interesting future research directions. First, it would be very valuable to build new live Re-ID datasets, 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 work 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.

Another interesting research direction would be to see if existing unsupervised cross-dataset adaptation methods could help the generalization of standard Re-ID models for the live Re-ID setting. 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.

Bibliography

ALTUNAY, Damla Gül et al. Intelligent surveillance system for abandoned luggage. In: IEEE. 2018 26th Signal Processing and Communications Applications Conference (SIU). [S.l.: s.n.], 2018. P. 1–4.

BEDAGKAR-GALA, Apurva; SHAH, Shishir K. A survey of approaches and trends in person re-identification. **Image and vision computing**, Elsevier, v. 32, n. 4, p. 270–286, 2014.

CHEN, Haoran et al. Deep transfer learning for person re-identification. In: IEEE. 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM). [S.l.: s.n.], 2018. P. 1–5.

CHEN, Zhicheng et al. FLAG: feature learning with additional guidance for person search. **The Visual Computer**, v. 37, Apr. 2021. DOI: 10.1007/s00371-020-01880-y.

DEB, Debayan; AGGARWAL, Divyansh; JAIN, Anil K. Identifying Missing Children: Face Age-Progression via Deep Feature Aging. In: IEEE. 2020 25th International Conference on Pattern Recognition (ICPR). [S.l.: s.n.], 2021. P. 10540–10547.

FELZENSZWALB, Pedro F et al. Object detection with discriminatively trained part-based models. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 32, n. 9, p. 1627–1645, 2009.

GHEISSARI, Niloofar; SEBASTIAN, Thomas B; HARTLEY, Richard. Person reidentification using spatiotemporal appearance. In: IEEE. 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). [S.l.: s.n.], 2006. v. 2, p. 1528–1535.

GOU, Mengran et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 41, n. 3, p. 523–536, 2018.

GUÉRIN, Joris; PAULA CANUTO, Anne Magaly de;
GONCALVES, Luiz Marcos Garcia. Robust Detection of Objects under Periodic Motion with Gaussian Process Filtering. In: IEEE. 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA). [S.l.: s.n.], 2020. P. 685–692.

HE, Kaiming et al. Deep Residual Learning for Image Recognition. **arXiv:1512.03385 [cs]**, Dec. 2015. arXiv: 1512.03385. Available from: <http://arxiv.org/abs/1512.03385>. Visited on: 4 Oct. 2021.

HE, Lingxiao et al. FastReID: A Pytorch Toolbox for General Instance Re-identification. **arXiv preprint arXiv:2006.02631**, 2020.

_____. Fastreid: A pytorch toolbox for general instance re-identification. **arXiv preprint arXiv:2006.02631**, 2020.

HIRZER, Martin et al. Person re-identification by descriptive and discriminative classification. In: SPRINGER. SCANDINAVIAN conference on Image analysis. [S.l.: s.n.], 2011. P. 91–102.

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

LAVI, Bahram et al. Survey on Reliable Deep Learning-Based Person Re-Identification Models: Are We There Yet? **arXiv preprint arXiv:2005.00355**, 2020.

LENG, Qingming; YE, Mang; TIAN, Qi. A survey of open-world person re-identification. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 30, n. 4, p. 1092–1108, 2019.

LI, Wei; WANG, Xiaogang. Locally Aligned Feature Transforms across Views. en. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, June 2013. P. 3594–3601. ISBN 978-0-7695-4989-7. DOI: 10.1109/CVPR.2013.461. Available from: <http://ieeexplore.ieee.org/document/6619305/>. Visited on: 20 Nov. 2022.

LI, Wei; ZHAO, Rui; WANG, Xiaogang. Human Reidentification with Transferred Metric Learning. en. In_____. **Computer Vision – ACCV 2012**. Berlin, Heidelberg: Springer, 2013. (Lecture Notes in Computer Science), p. 31–44. ISBN 978-3-642-37331-2. DOI: 10.1007/978-3-642-37331-2_3.

LI, Wei; ZHAO, Rui; XIAO, Tong, et al. Deepreid: Deep filter pairing neural network for person re-identification. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2014. P. 152–159.

LIAO, Shengcai et al. Open-set person re-identification. **arXiv preprint arXiv:1408.0872**, 2014.

LIN, Tsung-Yi et al. **Microsoft COCO: Common Objects in Context**. [S.l.: s.n.], 2015. arXiv: 1405.0312 [cs.CV].

LUO, Hao; GU, Youzhi, et al. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. en. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach, CA, USA: IEEE, June 2019. P. 1487–1495. ISBN 978-1-72812-506-0. DOI: 10.1109/CVPRW.2019.00190. Available from: <https://ieeexplore.ieee.org/document/9025455/>. Visited on: 15 Feb. 2021.

LUO, Hao; JIANG, Wei, et al. A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification. **IEEE Transactions on Multimedia**, v. 22, n. 10, p. 2597–2609, Oct. 2020. arXiv: 1906.08332. ISSN 1520-9210, 1941-0077. DOI: 10.1109/TMM.2019.2958756. Available from: <http://arxiv.org/abs/1906.08332>. Visited on: 15 Apr. 2021.

MACHACA, Luigy et al. TrADe Re-ID–Live Person Re-Identification using Tracking and Anomaly Detection. **arXiv preprint arXiv:2209.06452**, 2022.

MARTIéN ABADI et al. **TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems**. [S.l.: s.n.], 2015. Software available from tensorflow.org. Available from: <https://www.tensorflow.org/>.

MEKHAZNI, Djebril et al. Unsupervised domain adaptation in the dissimilarity space for person re-identification. In: SPRINGER. EUROPEAN Conference on Computer Vision. [S.l.: s.n.], 2020. P. 159–174.

MOON, Hyeonjoon; PHILLIPS, P Jonathon. Computational and performance aspects of PCA-based face-recognition algorithms. **Perception**, SAGE Publications Sage UK: London, England, v. 30, n. 3, p. 303–321, 2001.

MOSKVYAK, Olga et al. Going Deeper into Semi-supervised Person Re-identification. **arXiv preprint arXiv:2107.11566**, 2021.

PAPERS WITH CODE, person re-identification. [S.l.: s.n.], 2021. https://paperswithcode.com/task/person-re-identification. Accessed: 2021-09-28.

RADENOVIĆ, Filip; TOLIAS, Giorgos; CHUM, Ondřej. **Fine-tuning CNN Image Retrieval with No Human Annotation**. [S.l.]: arXiv, July 2018. arXiv:1711.02512 [cs]. Available from: <http://arxiv.org/abs/1711.02512>. Visited on: 18 Nov. 2022.

REDMON, Joseph; FARHADI, Ali. **YOLOv3: An Incremental Improvement**. [S.l.: s.n.], 2018. arXiv: 1804.02767 [cs.CV].

RISTANI, Ergys et al. Performance measures and a data set for multi-target, multi-camera tracking. In: SPRINGER. EUROPEAN conference on computer vision. [S.l.: s.n.], 2016. P. 17–35.

SUMARI, Felix O et al. Towards practical implementations of person re-identification from full video frames. **Pattern Recognition Letters**, Elsevier, v. 138, p. 513–519, 2020.

SUN, Yifan et al. **Beyond Part Models: Person Retrieval with Refined Part Pooling (and a Strong Convolutional Baseline)**. [S.l.]: arXiv, Jan. 2018. arXiv:1711.09349 [cs]. DOI: 10.48550/arXiv.1711.09349. Available from: <http://arxiv.org/abs/1711.09349>. Visited on: 18 Nov. 2022.

SZEGEDY, Christian et al. Rethinking the Inception Architecture for Computer Vision. **arXiv:1512.00567 [cs]**, Dec. 2015. arXiv: 1512.00567. Available from: <http://arxiv.org/abs/1512.00567>. Visited on: 4 Oct. 2021.

WANG, Guanshuo et al. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. **Proceedings of the 26th ACM international conference on Multimedia**, p. 274–282, Oct. 2018. arXiv: 1804.01438 version: 1. DOI: 10.1145/3240508.3240552. Available from: <http://arxiv.org/abs/1804.01438>. Visited on: 13 Apr. 2021.

WANG, Hanxiao et al. Human-in-the-loop person re-identification. In: SPRINGER. EUROPEAN conference on computer vision. [S.l.: s.n.], 2016. P. 405–422.

WANG, Taiqing et al. Person re-identification by video ranking. In: SPRINGER. EUROPEAN conference on computer vision. [S.l.: s.n.], 2014. P. 688–703.

WANG, Xiaogang. Intelligent multi-camera video surveillance: A review. en. **Pattern Recognition Letters**, v. 34, n. 1, p. 3–19, Jan. 2013. ISSN 01678655. DOI: 10.1016/j.patrec.2012.07.005. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S016786551200219X>. Visited on: 20 Nov. 2022.

WANG, Xiaolong et al. **Non-local Neural Networks**. [S.l.]: arXiv, Apr. 2018. arXiv:1711.07971 [cs]. Available from: <http://arxiv.org/abs/1711.07971>. Visited on: 20 Nov. 2022.

XIAO, Tong et al. Joint detection and identification feature learning for person search. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2017. P. 3415–3424.

XU, Yuanlu et al. Person search in a scene by jointly modeling people commonness and person uniqueness. In: PROCEEDINGS of the 22nd ACM international conference on Multimedia. [S.l.: s.n.], 2014. P. 937–940.

YANG, ChangShui; QI, Feng; JIA, Huizhu. Survey on Unsupervised Techniques for Person Re-Identification. In: IEEE. 2021 2nd International Conference on Computing and Data Science (CDS). [S.l.: s.n.], 2021. P. 161–164.

YE, Mang et al. Deep learning for person re-identification: A survey and outlook. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, 2021.

ZHAO, Fang et al. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In: SPRINGER. EUROPEAN Conference on Computer Vision. [S.l.: s.n.], 2020. P. 526–544.

ZHENG, Liang; BIE, Zhi, et al. Mars: A video benchmark for large-scale person re-identification. In: SPRINGER. EUROPEAN Conference on Computer Vision. [S.l.: s.n.], 2016. P. 868–884.

ZHENG, Liang; SHEN, Liyue, et al. Scalable person re-identification: A benchmark. In: PROCEEDINGS of the IEEE international conference on computer vision. [S.l.: s.n.], 2015. P. 1116–1124.

ZHENG, Liang; YANG, Yi; HAUPTMANN, Alexander G. **Person Re-identification: Past, Present and Future**. [S.l.]: arXiv, Oct. 2016. arXiv:1610.02984 [cs]. Available from: <http://arxiv.org/abs/1610.02984>. Visited on: 20 Nov. 2022.

ZHENG, Liang; ZHANG, Hengheng, et al. Person re-identification in the wild. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.: s.n.], 2017. P. 1367–1376.

ZHONG, Zhun et al. Random Erasing Data Augmentation. **arXiv:1708.04896 [cs]**, Nov. 2017. arXiv: 1708.04896. Available from: <http://arxiv.org/abs/1708.04896>. Visited on: 4 Oct. 2021.

ZHUANG, Weiming et al. Performance optimization of federated person re-identification via benchmark analysis. In: PROCEEDINGS of the 28th ACM International Conference on Multimedia. [S.l.: s.n.], 2020. P. 955–963.

APPENDIX A – Complete results from our experiments

This appendix presents all the results from our cross-dataset Re-ID experiments.

A.1 Simple Cross-dataset results

Four standard Re-ID approaches are trained on three different standard Re-ID datasets, and evaluated on a different dataset. We are using standard Re-ID metrics as Rank-n, mAP and mINP. The complete results from these experiments are reported in Table 5.

A.2 Combine cross-dataset results

For this experiment, we're combining two datasets for train and using two different to evaluate four standard Re-ID approaches. There is one special combination that's $\text{COMBINED}_{\text{all}}$ where we combine three datasets and use only one for evaluation. The complete results from these experiments are reported in Table 6.

A.3 Combine scaled cross-dataset results

The idea used here is similar to A.2 but instead of using all the dataset size we reduce it. The complete results from these experiments are reported in Table 7.

A.4 Live Re-ID results

The complete results from these experiments are reported in Table 8.

Table 5: Complete results from our cross-dataset experiments using only one dataset for train.

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
	PRID	AGW	0.07	0.14	0.18	0.11	0.11
		MGN	0.21	0.31	0.35	0.26	0.26
		SBS	0.14	0.25	0.29	0.20	0.20
		BoT	0.06	0.10	0.13	0.09	0.09
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
	PRID	AGW	0.08	0.15	0.20	0.12	0.12
		MGN	0.23	0.36	0.42	0.30	0.30
		SBS	0.12	0.22	0.26	0.17	0.17
		BoT	0.03	0.11	0.16	0.07	0.07
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
	PRID	AGW	0.14	0.22	0.26	0.19	0.19
		MGN	0.22	0.35	0.40	0.28	0.28
		SBS	0.15	0.24	0.30	0.20	0.20
		BoT	0.08	0.18	0.23	0.13	0.13

Table 6: Complete results from our cross-dataset experiments using two or more datasets for training.

Train	TEST	Approach	Rank-1	Rank-5	Rank-10	mAP	mINP
Market-1501 & DukeMTMC	CUHK03	AGW	0.13	0.25	0.32	0.14	0.08
		MGN	0.28	0.45	0.55	0.27	0.17
		SBS	0.26	0.42	0.52	0.24	0.14
		BoT	0.12	0.21	0.28	0.11	0.06
	PRID	AGW	0.15	0.25	0.28	0.20	0.20
		MGN	0.27	0.37	0.42	0.32	0.32
		SBS	0.16	0.28	0.33	0.22	0.22
		BoT	0.15	0.24	0.29	0.20	0.20
CUHK03 & Market-1501	DukeMTMC	AGW	0.44	0.59	0.65	0.29	0.05
		MGN	0.63	0.77	0.81	0.44	0.08
		SBS	0.62	0.74	0.79	0.41	0.07
		BoT	0.35	0.48	0.55	0.21	0.03
	PRID	AGW	0.14	0.22	0.27	0.19	0.19
		MGN	0.26	0.37	0.42	0.32	0.32
		SBS	0.20	0.29	0.34	0.24	0.24
		BoT	0.11	0.20	0.26	0.16	0.16
DukeMTMC & CUHK03	Market-1501	AGW	0.65	0.78	0.83	0.38	0.07
		MGN	0.78	0.89	0.93	0.52	0.13
		SBS	0.74	0.87	0.91	0.47	0.10
		BoT	0.61	0.75	0.80	0.34	0.06
	PRID	AGW	0.09	0.19	0.24	0.14	0.14
		MGN	0.28	0.40	0.45	0.34	0.34
		SBS	0.18	0.27	0.32	0.23	0.23
		BoT	0.06	0.13	0.16	0.09	0.09
$\text{COMBINED}_{\text{all}}$	CUHK03	AGW	0.73	0.89	0.94	0.71	0.61
		MGN	0.83	0.93	0.96	0.82	0.74
		SBS	0.77	0.90	0.94	0.76	0.66
		BoT	0.69	0.86	0.92	0.68	0.58
	DukeMTMC	AGW	0.88	0.95	0.96	0.79	0.44
		MGN	0.91	0.96	0.97	0.82	0.49
		SBS	0.88	0.95	0.96	0.78	0.43
		BoT	0.87	0.94	0.96	0.77	0.41
	Market-1501	AGW	0.95	0.98	0.99	0.88	0.65
		MGN	0.96	0.99	0.99	0.91	0.70
		SBS	0.95	0.98	0.99	0.88	0.64
		BoT	0.94	0.98	0.99	0.86	0.60
	PRID	AGW	0.14	0.26	0.32	0.20	0.20
		MGN	0.29	0.42	0.45	0.35	0.35
		SBS	0.17	0.27	0.33	0.23	0.22
		BoT	0.11	0.18	0.24	0.15	0.15

Table 7: Complete results from our scaled cross-dataset experiments using two or more datasets for train.

Train	TEST	Approach	Rank-1	Rank-5	Rank-10	mAP	mINP
Market-1501 & DukeMTMC$_{\text{SCALED}}$	CUHK03	AGW	0.13	0.23	0.31	0.13	0.07
		MGN	0.24	0.40	0.52	0.23	0.14
		SBS	0.21	0.37	0.46	0.20	0.12
		BoT	0.10	0.19	0.23	0.09	0.05
	PRID	AGW	0.12	0.21	0.24	0.17	0.17
		MGN	0.27	0.38	0.44	0.32	0.32
		SBS	0.16	0.27	0.31	0.21	0.21
		BoT	0.13	0.20	0.26	0.17	0.17
CUHK03 & Market-1501$_{\text{SCALED}}$	DukeMTMC	AGW	0.41	0.56	0.62	0.26	0.04
		MGN	0.60	0.73	0.78	0.40	0.07
		SBS	0.55	0.70	0.75	0.35	0.06
		BoT	0.30	0.45	0.51	0.18	0.02
	PRID	AGW	0.10	0.19	0.24	0.15	0.15
		MGN	0.27	0.37	0.41	0.32	0.32
		SBS	0.19	0.29	0.33	0.24	0.24
		BoT	0.08	0.15	0.18	0.12	0.12
DukeMTMC & CUHK03$_{\text{SCALED}}$	Market-1501	AGW	0.65	0.78	0.83	0.38	0.08
		MGN	0.77	0.88	0.92	0.52	0.13
		SBS	0.74	0.85	0.89	0.46	0.09
		BoT	0.58	0.72	0.78	0.32	0.05
	PRID	AGW	0.11	0.18	0.22	0.15	0.15
		MGN	0.26	0.37	0.43	0.31	0.31
		SBS	0.19	0.29	0.33	0.24	0.24
		BoT	0.09	0.15	0.18	0.12	0.12
COMBINED$_{\text{SCALED-all}}$	CUHK03	AGW	0.68	0.84	0.88	0.67	0.56
		MGN	0.75	0.87	0.91	0.72	0.62
		SBS	0.73	0.87	0.90	0.72	0.61
		BoT	0.65	0.82	0.88	0.61	0.52
	DukeMTMC	AGW	0.82	0.90	0.93	0.67	0.28
		MGN	0.84	0.91	0.94	0.70	0.30
		SBS	0.81	0.89	0.93	0.66	0.27
		BoT	0.77	0.88	0.92	0.63	0.25
	Market-1501	AGW	0.91	0.97	0.98	0.79	0.46
		MGN	0.94	0.98	0.99	0.83	0.52
		SBS	0.93	0.97	0.98	0.81	0.49
		BoT	0.89	0.96	0.98	0.75	0.41
	PRID	AGW	0.13	0.22	0.24	0.18	0.18
		MGN	0.30	0.42	0.46	0.36	0.36
		SBS	0.20	0.32	0.36	0.26	0.26
		BoT	0.10	0.19	0.22	0.15	0.15

Table 8: 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
$\text{COMBINED}_{\text{all}}$	AGW	0.49	0.49	0.56	0.71
	BoT	0.30	0.37	0.45	0.63
	SBS	0.71	0.71	0.71	0.79
	MGN	0.80	0.77	0.81	0.86
$\text{COMBINED}_{\text{SCALED-all}}$	AGW	0.39	0.45	0.49	0.63
	BoT	0.31	0.39	0.44	0.58
	SBS	0.72	0.73	0.68	0.77
	MGN	0.75	0.80	0.77	0.84

APPENDIX B – Graphs results for live Re-ID experiments

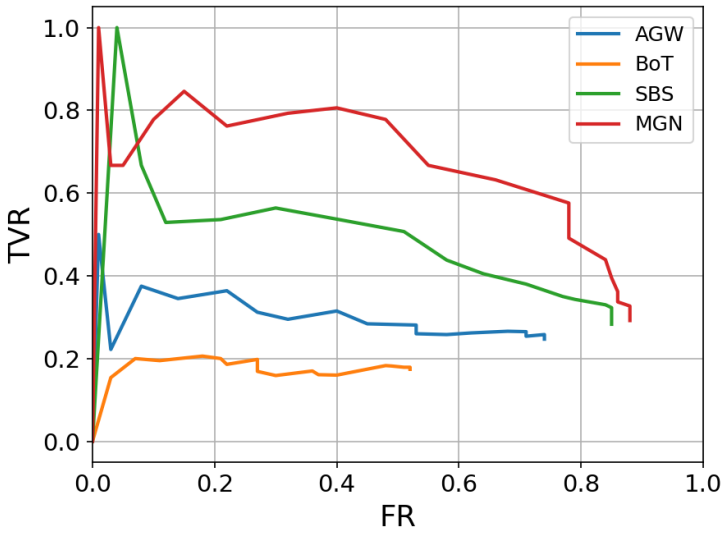
This section presents all the results from our live Re-ID experiments. But using graphs showing Finding Rate(FR) vs True Validation Rate(TVR) curves corresponding to these experiments.

We present in Figure 19, the 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.

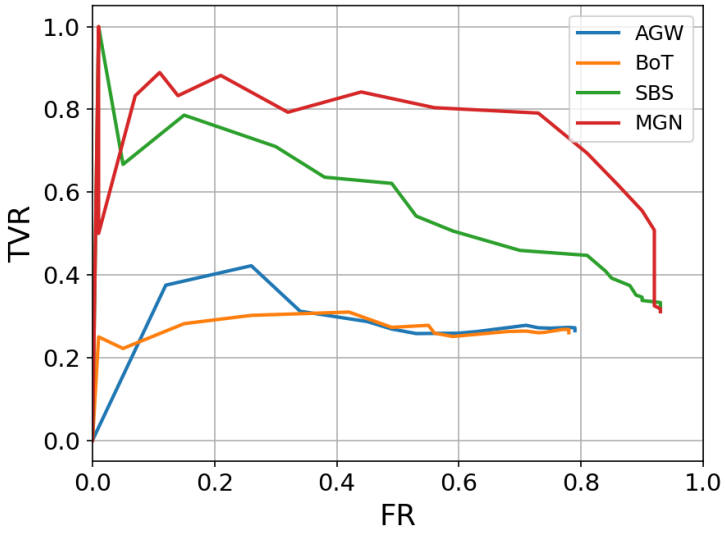
- Approach AGW that had attention blocks in its architecture, had average results as we can see in Figures 19b and 19c and in some points similar behavior as BoT.
- Other observations extracted from Figure 19a that have an initial backbone as Resnet-50 and use some strategies for improving training such as SBS, perform an improvement, especially for a hard dataset such as CUHK03.
- We also observe that the MGN approach generalizes better than the other three. But if you use $\text{COMBINED}_{\text{all}}$ for training, as shown in Figure 19d, it is possible to improve the TVR to over 80%.

Finally, we also present in Figure 20, the influence of the training dataset over different Re-ID approaches. Evaluation is conducted on the *modified PRID-2011* dataset for live Re-ID too.

- In Figure 20b, $\text{COMBINED}_{\text{all}}$ dataset for training over BoT had the worst result, this behavior is very similar for other datasets too. Maybe because the initial baseline is not enough.
- For the AGW approach, in Figure 20a, training with Market-1501 and $\text{COMBINED}_{\text{all}}$ at the beginning had high TVR but it decreases very quickly.

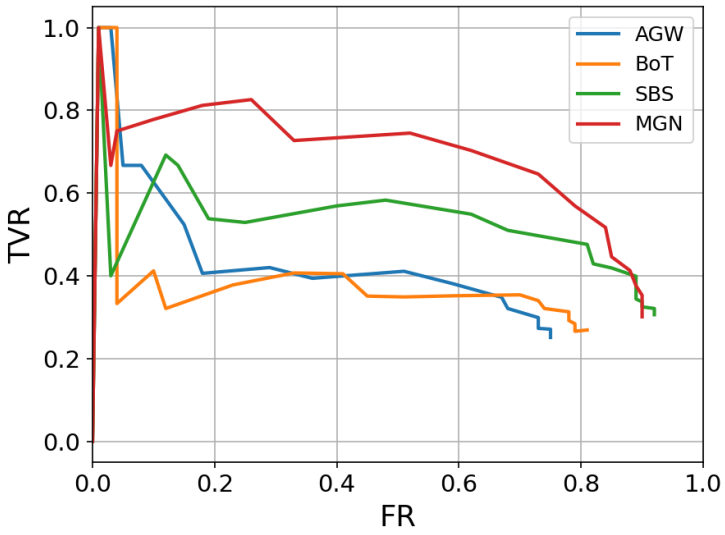


(a) Training on CUHK03.

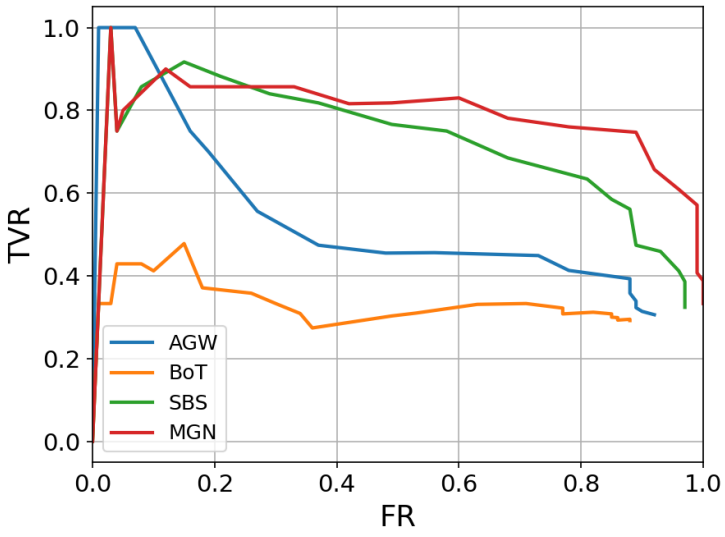


(b) Training on DukeMTMC.

- One more time in Figure 20d we observe that this approach has a good performance using $\text{COMBINED}_{\text{all}}$ for training.

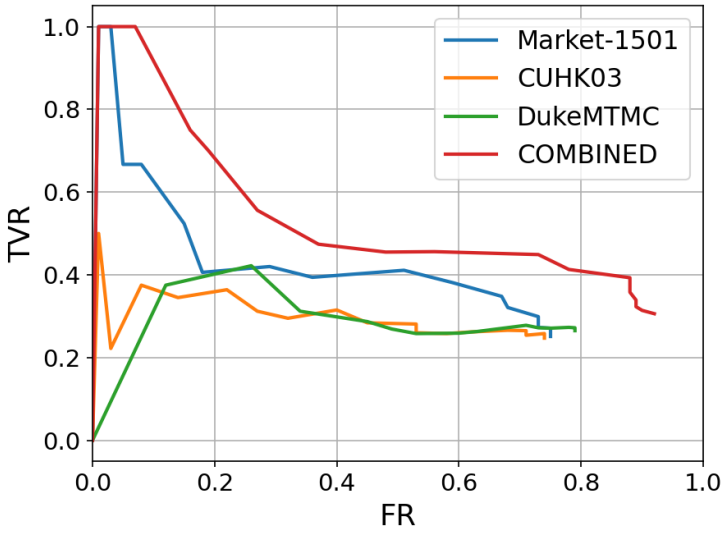


(c) Training on Market-1501.

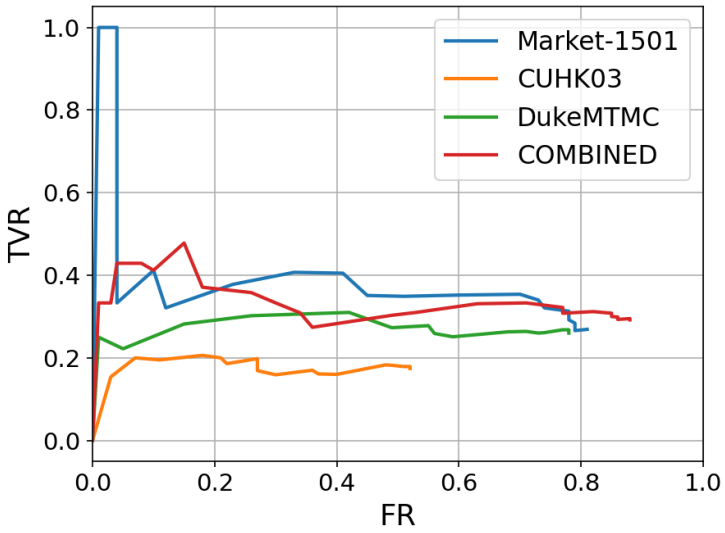


(d) Training Combining$_{\text{all}}$ Datasets.

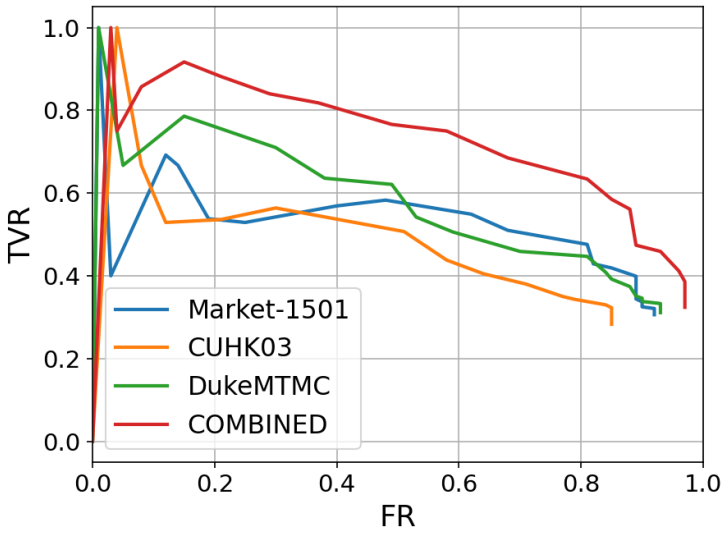
Figure 19: **Influence of the standard Re-ID approach**. TVR vs FR curves of different standard Re-ID approaches for different training datasets. Evaluation is conducted on the *modified PRID-2011* dataset for live Re-ID.


Market-1501
CUHK03
DukeMTMC
COMBINED
TVR
FR
0.0
0.2
0.4
0.6
0.8
1.0

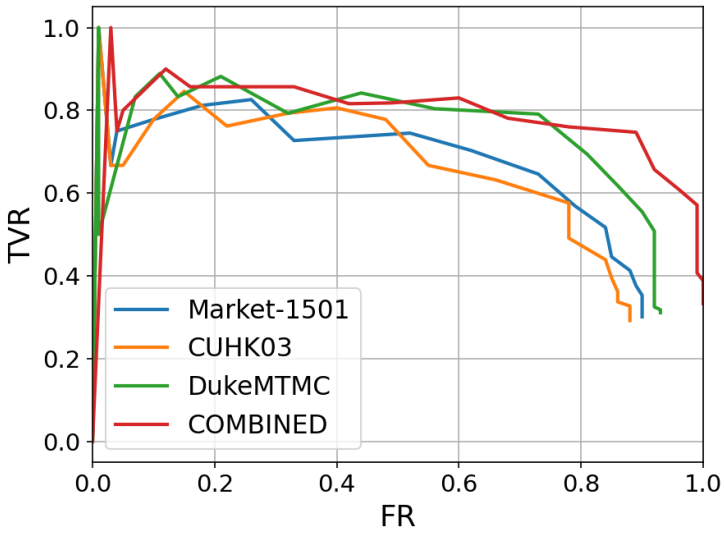
(a) AGW.


Market-1501
CUHK03
DukeMTMC
COMBINED
TVR
FR
0.0
0.2
0.4
0.6
0.8
1.0

(b) BoT.



(c) SBS.



(d) MGN.

Figure 20: **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.