Alma Mater Studiorum · University of Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica

On the Adaptability of Person Re-Identification Models

Supervisor:
Prof. Andrea Asperti

Co-Supervisor:
Dr. Salvatore Fiorilla

Candidate:
Leonardo Naldi

Session II
2023/2024

Abstract

Person Re-Identification is a computer vision task that aims at retrieving a target person across multiple, non overlapping cameras. It is a central component of any intelligent surveillance system, and has seen a steady increase in research efforts. Despite the problem's considerable difficulty, state-of-the-art models have achieved impressive results over all benchmark datasets, which has motivated us to investigate the applicability of such models in a real-world scenario. To this end, we present a brief overview of the most important aspects of person re-identification, as well as some experimental results regarding the domain gap problem of person re-identification models, which to this day remains the main challenge to the adoption of such models in a real-world setting.

Contents

Chapter 1

Introduction

The increase in deep neural network performance over the last decade and its consequent surge in popularity has led to a never-before seen amount of interest go towards the implementation of intelligent systems into practical scenarios. One scenario that has seen a constant increase in demand has been that of camera-based surveillance systems, which have steadily grown in size and complexity, making relying on a human operator more and more arduous. Person re-identification, often shortened to person re-id, is a computer vision task that sits at the heart of any such system: its aim is to retrieve a person of interest's identity across multiple, non-overlapping cameras [13].

Even in its most basic form, person re-id has proven itself to be an extremely challenging task, mostly due to the reliance of the models on intrinsically weak visual features (such as clothing, general appearance and objects carried) to perform the re-identification [13]. Additionally person re-id datasets are hard to both gather and annotate.

Nevertheless, research has proceeded steadily in the past years, achieving extremely promising results across all available person re-id datasets (on which state-of-the-art models generally achieve a Rank-1 score of at least 90% [48, 38]), some of which will be discussed in Chapter 2.

These results, together with the aforementioned interest in real-world intelligent systems, have motivated us to believe the applicability of such models is to be investigated. To this end, we directed our efforts towards evaluating the performance of person re-id state-of-the-art models on never-before seen data, to evaluate how such models would perform in a real-world setting. This is done by measuring the model's performance on a dataset different from the one it was trained on, which evaluates the model on a new set of subjects, camera angles, levels of occlusion, light levels, environmental conditions, backgrounds and image qualities, all of which are crucial challenges that will be faced by any person re-id model deployed in a real-world scenario. We used this method to estimate, among all datasets at our disposal, the two most *different* datasets, meaning the two datasets that present the biggest gap in performance when one is used for training and the other is used for testing. We then attempt to bridge this gap by merging the two datasets, retraining the model on the result and measure its performance on the remaining datasets.

This thesis is structured as follows: in this chapter we give a brief overview of person re-identification and the domain gap, we then move on to an overview of some of the most important datasets in Chapter 2 (including every dataset used in our experiments). Chapter 3 contains an outline of the models we used, as well as the most common metrics in person re-id. The results of our experiments relating to the domain gap will be presented in Chapter 4.

1.1 Person Re-Identification

The basic functionality of an intelligent camera-based surveillance system would be to be to detect all pedestrians present in the current feed, track them as they move and retrieve all images of a given person (which is essential in a number of surveillance applications, for example long-term multi-camera tracking [13]) [48], both in the current feed of other cameras and from a data bank of stored images/videos. In computer vision terms these tasks are called as follows:

- **Object Detection** [61], which is the task of detecting and locating objects in a given image or video. Most commonly camera-based surveillance systems will only need to track people.
- **Object Tracking** [46], which is the task of tracking a set of bounding boxes as they move. Each bounding box will be paired with an ID which will have to be kept coherent throughout the tracking.
- **Person Re-Identification** [13, 48, 53, 52], which will implement the person retrieval part of the surveillance system.

Person Re-Identification (person re-id for short) is, as said, the computer vision task that implements person retrieval: in its simplest and most widely studied form, it consists of matching a given a person's image, which will be called a *query*, against a set of cross-camera images, which will be called a *gallery*, as shown in Fig 1.1.

Person Re-Identification presents numerous challenges. Depending on a number of factors (some of which include camera angle and quality, distance from the subject and subject pose, environmental factors and lighting conditions etc.), relying on more robust biometric features, as is done in face recognition, to perform person re-id could be unfeasible [13]. Instead, person re-id models have to rely on more generic appearance based features, such as clothing, which can be weak for associating people. For example, a person recorded by two non-overlapping cameras could take off a coat, put one on, or, most commonly, be wearing very similar clothing to some other person, as shown in Figure 2.4. This can result in cross-camera images representing two different identities being

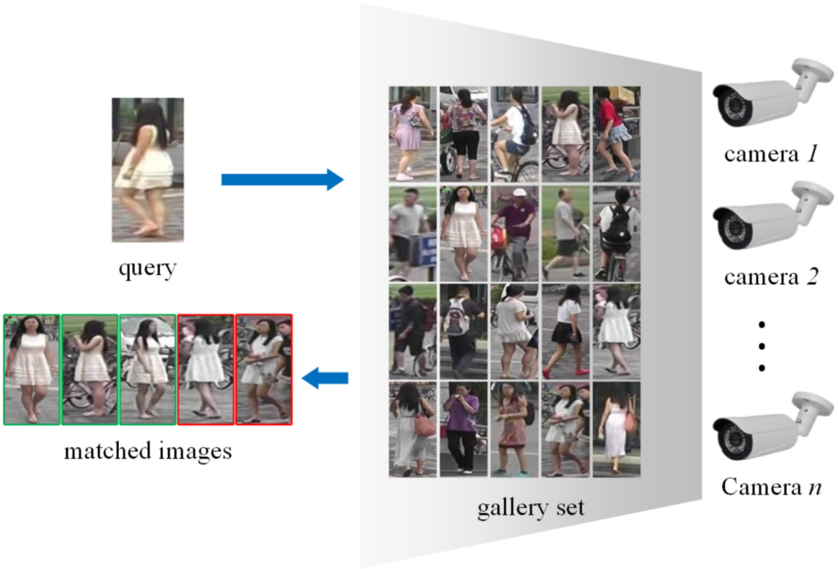


Figure 1.1: [48] The Person Re-Identification task: the query image of a woman is matched against the gallery set produced by n independent cameras. The output is the gallery subset that matches the query identity.

more alike than two cross-camera images representing the same identity, in classification terms this means that intra-class variability can be greater than inter-class variability [13].

The intrinsically weak nature of appearance features is compounded by the difficulties inherent to the multi-camera setting of person re-id. First is the pose variation of the subjects: given that images of the same person will be recorded from different cameras at different times, there will be a great amount of variation in the subject's poses (e.g. the subject could be walking towards one of the cameras and away from the other), in the levels of environmental occlusion (e.g. other pedestrians, cars) and lighting conditions. Secondly, the fact that images will be taken from different cameras further adds complexity to the problem, as the images will also be taken from different angles and, possibly, with varying image qualities and backgrounds. Lastly, the training images which will comprise the datasets are often generated using an object detector, which can sometimes be inaccurate, producing misaligned training images or images that only partially contain the subject, as shown in Figure 1.2 (f). However this is, as will be discussed in Chapter 2, more of an asset than a liability: as said, it is reasonable to assume that an intelligent camera-based surveillance system would employ an object detector and an object tracker, meaning that that person re-id models should be prepared to sometimes face malformed bounding boxes.

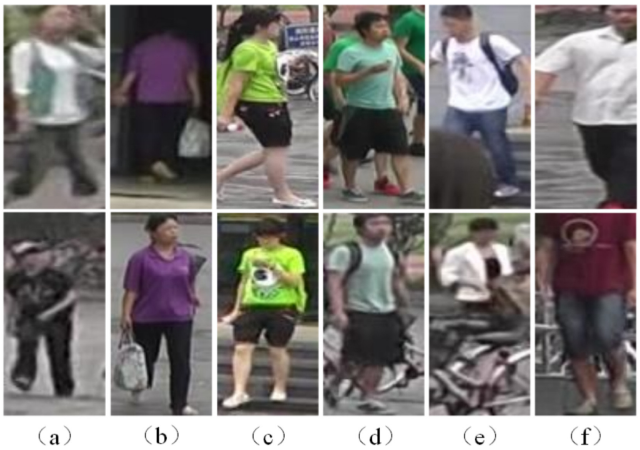


Figure 1.2: [48] Some of the challenges presented by person re-id: (a) low resolution, (b) pose and camera angle changes, (c) different imaging techniques, (d) environmental occlusion, (f) inaccurate bounding box.

Variations

In this thesis we will solely focus on image-to-image Person Re-Identification, where both the query and the gallery are image sets, but variations do exist and are worth mentioning. The most closely related one is Vehicle Re-Identification, in which both the query and the gallery will be images of vehicles, such as cars, trucks and motorbikes, instead of people [3]. Other commonly studied variations include text-to-image person re-id, where the gallery is searched not based on a query image, but based on a textual description that acts as the query [41], or video person re-id, where either the gallery, or the query, or both are comprised of videos instead of images [17].

1.2 The Domain Gap

One of the most important problems plaguing person re-id models is the domain gap. Models, even high performing models, suffer a large performance degradation when tested on a never-before seen dataset [47, 52].

The reason seems to be twofold: as said, person re-id is an extremely complicated task, which in turn causes even small variations in the nature of the data (e.g. lighting changes, different camera angles, different clothing styles, etc.), which will be extremely common in data taken from different cameras and in different environments than those used to create the model's training dataset, have a great impact on model performance. Secondly, because of the costly nature of gathering and annotating person re-id data, most available datasets are, as will be discussed in Chapter 2, rather small in size, making it even more difficult for models to learn generalized features.

The domain gap is currently the biggest challenge to the deployment of person re-id models in a real world setting, since a good performance cannot be guaranteed on never-before seen data.

Chapter 2

Datasets

As with all deep learning tasks, data plays a crucial role in the model's performance. In this chapter, we will present a brief overview of some of the most influential person re-id datasets.

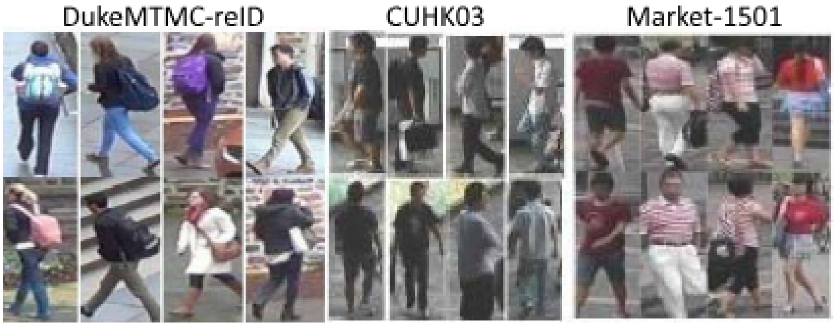


Figure 2.1: [49] Examples of images from three different person re-id datasets: DukeMTMC-reID [36], CUHK03 [23] and Market-1501 [57].

Any effective person re-identification dataset will present challenges similar to the ones present in a real-life scenario. To better capture these challenges most datasets are produced from video footage recorded in a public setting (like a university campus). The frames from the video footage are then passed to an object detector or processed by hand in order to generate the bounding boxes for each pedestrian, some examples from three different datasets can be seen in Figure 2.1. As said in the previous chapter, person re-identification aims at retrieving a person's identity (called a *query*) from an image set (called *gallery*). The datasets for person re-id are, as is often the case with deep learning datasets, split into the training set and the testing set. The testing set will be further split into a query set and a gallery set. In order for this to be feasible, each query identity (that is, each identity that appears at least once in the query set) needs to be present in at least one, but preferably more, gallery images taken from a

Dataset	Year	Identities	Bounding Boxes	Cameras
CUHK03 [23]	2014	1,360	13,164	6
Market-1501 [57]	2015	1,501	32,668	6
DukeMTMC-reID [36]	2016	1,812	36,441	8
MSMT17 [47]	2017	4,101	126,441	15
Airport [19]	2018	9,651	39,902	6
ENTIRe-ID [54]	2024	13,540	4.45M	37
ENTIRe-ID (Testing Set[1])	2024	2,741	13,415	37

Table 2.1: Characteristics of the person re-id datasets that will be discussed in this chapter. Highlighted datasets are the one used for our experiments in Chapter 4.

different camera. The bounding boxes produced from the raw video that satisfy these prerequisite will be selected, the others will either be discarded or added as distractor images. Distractor images, as will be discussed in Section 2.2, help make the dataset closer to a real-world environment.

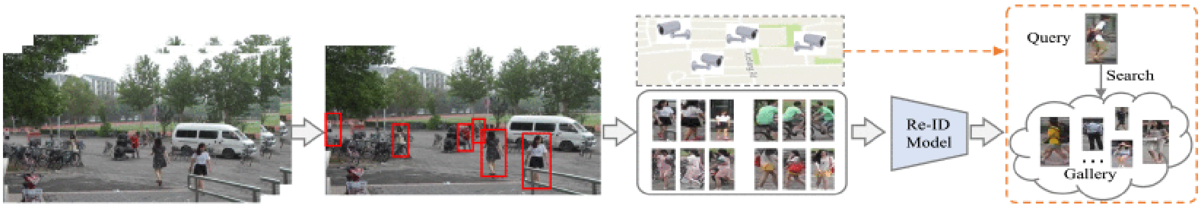


Figure 2.2: [53] The person re-identification data pipeline: first the raw footage is captured by a public camera, the bounding boxes are then made (by hand or with an object detector), annotated, and are finally used by a model.

Each of the selected bounding boxes is then annotated with an ID corresponding to the person it depicts, the camera's ID and, optionally, a timestamp and/or track ID and is finally added to the dataset. Table 2.1 provides a quick summary of the person re-id datasets we will be presenting in this chapter.

2.1 CUHK03

CUHK03 is a dataset released in 2014 by Li et al. [23]. It was developed at the Chinese University of Hong Kong and its main contributions to the field lie in its size and its annotation method.

[1] As of the time of writing, only the testing set has been released.

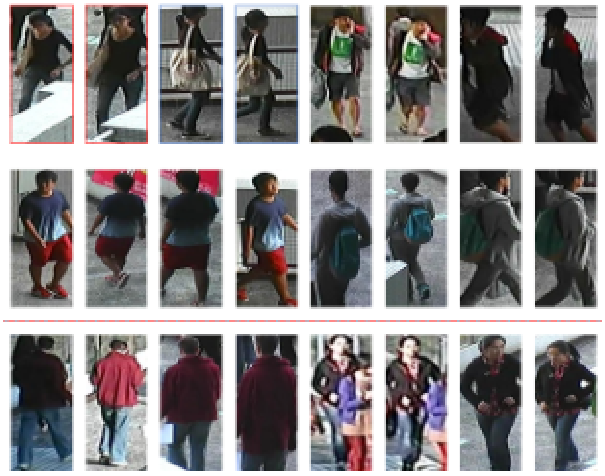


Figure 2.3: [23] Some examples of CUHK03 images. Each pair of adjacent images share the same identity.

Compared to earlier works [39, 58, 27, 7, 16, 14], CUHK03 exhibits a much greater magnitude in terms of data volume with 13,164 bounding boxes and 1,360 identities extracted from footage recorded by 6 cameras, with each identity being captured by two cameras and having an average of 4.8 images per identity in each of the two camera views. For comparison, the VIPeR dataset [14] contains 632 image pairs taken by 2 cameras (each identity has one pair of images), and the CUHK02 dataset [22], released by the same authors in 2013, contains 1,816 identities, with each identity appearing in two cameras, but with only two images per camera view (for a total of 7,264 images). This increase in size was pivotal for allowing the training of deep neural networks, which generally requires large amounts of data to be effective.

The second major contribution of the dataset is the availability of both hand-crafted and automatically detected bounding boxes. The hand-crafted bounding boxes (which are the standard in earlier works) allow researchers to train and test the models under ideal conditions, whereas the automatically detected ones (generated using a state of the art object detector, the Deformable Part Model [11]) bring the dataset closer to what would be seen by a real world automatic surveillance system, by introducing misalignment and missing body parts (as can be seen in Figure 2.3), which can rarely be found in hand-crafted bounding boxes.

Lastly, the footage used to create the dataset was recorded by six different cameras (running on different settings) over the span of multiple months, introducing illumination changes caused from weather variations, sun directions and shadows even within a single camera view, thus adding a further layer of complexity (and realism) to the data.

2.2 Market-1501

Introduced by Zheng et al. [57] in 2015, Market-1501 is a person re-identification dataset consisting of 32, 668 bounding boxes representing, as the name implies, 1, 501 different identities. The dataset was created from footage recorded by six cameras in front of a campus market and introduced three main innovations: the usage of an object detector to craft bounding boxes (following the lead of CUHK03 [23], Section 2.1), having each identity potentially be captured by more than two cameras and supplementing the dataset with a distractor set to further amplify the effects of automatic pedestrian detection.

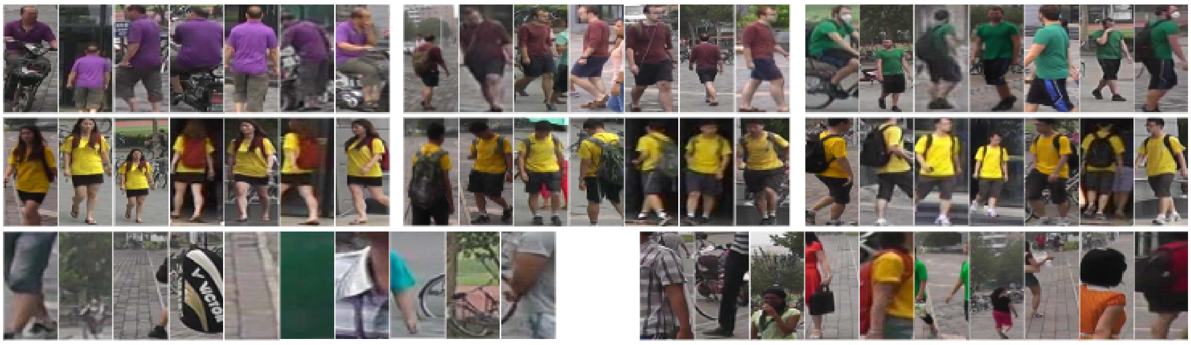


Figure 2.4: [57] Some examples of Market-1501 images: the top row consists of three distinct identities that are easily told apart, the middle row contains three distinct identities that share very similar appearance and the bottom row consists of distractor images (left) as well as junk images (right).

The pedestrian bounding boxes for the gallery set were generated using an object detector (a Deformable Part Model [11], the same one used in the CUHK03 dataset [23]), as opposed to manually hand-crafted bounding boxes. The usage of an object detector introduces misaligned images and false alarms, which, contrary to intuition, help the dataset account for detection errors, which will always be present in an intelligent surveillance systems that employs object detectors to identify pedestrians. Each detected bounding box was categorized using a ground truth hand-crafted bonding box and calculating the ratio between the overlapping areas and the union of the two areas. The bounding box was categorized as *good* if the overlap was at least 50% of the area ratio, *distractor* if the ratio was less than 20% and *junk* otherwise, as shown in Figure 2.4. The inclusion of junk and distractor images is, together with its size, what differentiates Market-1501 from CUHK03 [23], where even the automatically detected bounding boxes are all relatively accurate.

Market was also one of the first datasets where each identity potentially appears in more than two cameras and with multiple images under each camera, as opposed to the usual approach of having each identity only captured by two cameras [23, 14, 22], meaning that both the query and the gallery set may contain images from multiple cameras of the same identity. This allows the model to obtain better discriminative information from the multiple query images and is also more consistent with the practical use case.

Lastly, scale plays a vital role in every deep learning dataset, thus Market was further enhanced with 500K distractor images (Figures 2.4 and 2.5 contain some examples). These images consist of false alarms by the object detector as well as pedestrian bounding boxes that do not reappear in the other cameras.

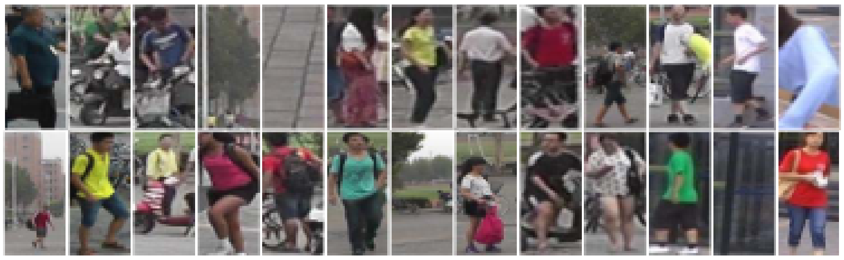


Figure 2.5: [57] Some examples of distractor images from the Market-1501 dataset.

This addition contributes to the dataset's scale, as well as pushing it closer to the real-world scenario, where detection errors and images containing non-reappearing identities will be at least somewhat common.

2.3 DukeMTMC-reID

DukeMTMC-reID (Duke Multi Target Multi Camera re-identification) [36] is a dataset released by Ristani et al. in 2016. It was crafted from footage recorded by 8 high quality cameras placed on Duke University's campus. It is one of the largest dataset consisting of hand-crafted bounding boxes [49] extracted from high quality images. The footage consists of an 85 minute recording taken from the time in between lectures (to guarantee high pedestrian activity) for each camera. The cameras record at 1080p and 60 frames per second. Each person was manually tracked by recording the person's foot point of contact with the ground, as can be seen in Figure 2.6. Finally, for each of the eight cameras, the first five minutes of footage were reserved for training, and the remaining for testing.

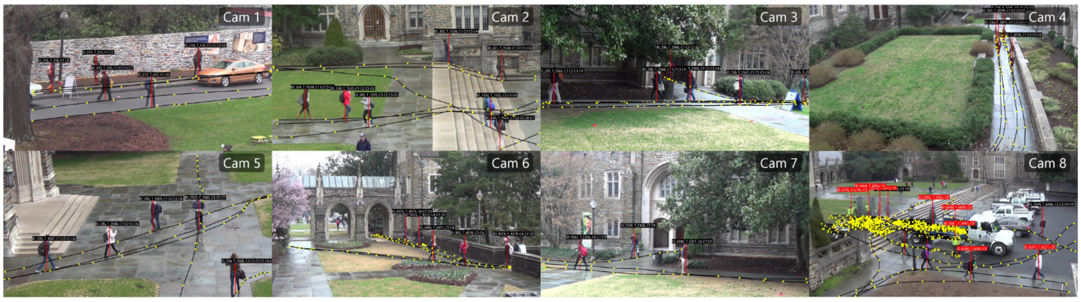


Figure 2.6: [36] A frame with all the manually-annotated trajectories for each camera.

The people depicted within Duke often carry a wide variety of items and occlusion is rather common.

The footage used to create the dataset was all recorded in the same scene (university campus with soft overhead light), which limits the dataset's variety in terms of background and lighting conditions. Additionally, since the bounding boxes are extracted from hand-crafted trajectories, the dataset does not account for object detector or object tracker errors, which will be seen by a real-world person re-id system.

2.4 MSMT17

The MSMT17 (**M**ulti-**S**cene **M**ulti-**T**ime) person re-id dataset [47] was released in 2017 by Wei et al. It remains, at of the time of writing, one of the largest and most challenging supervised person re-identification datasets.

The dataset was extracted from 180 hours of video recorded by 15 cameras set up on a university campus. Contrary to previous works where all cameras recorded an outside scene, the cameras for MSMT17 were split between outside (12 cameras) and inside (3 cameras). Some statistics on the on the distribution of identities amongst cameras and time slots can be found in Figure 2.8. The video was extracted by selecting four days with different weather conditions among a month of recording. For each day and for each camera three hours of recording (taken in the morning, at noon and in the afternoon respectively) were used. The presence of multiple scenes (indoor and outdoor), the high number of cameras, as well as multiple time slots and varied weather conditions serves to introduce complex illumination changes, scene and background changes, in addition to the challenges commonly present in other person re-identification datasets (e.g. occlusion and pose variations), as can be seen in Figure 2.7.

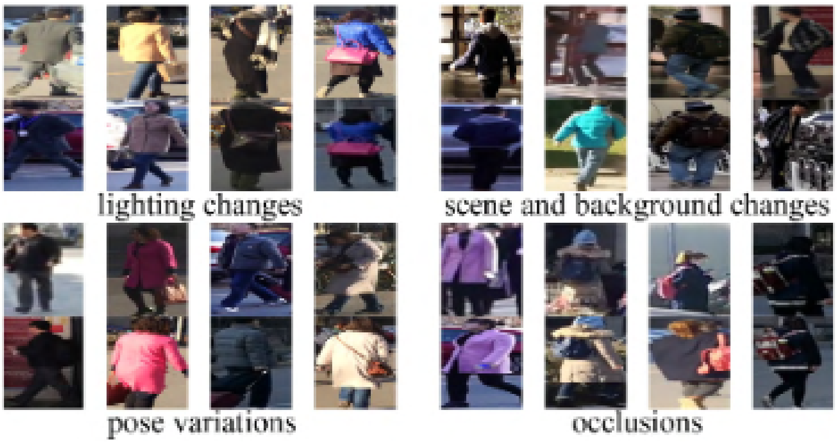
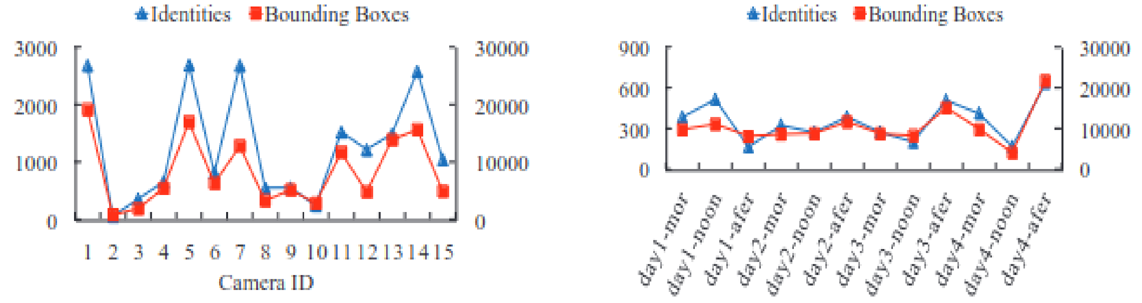


Figure 2.7: [47] Some examples of MSMT17 images and the challenges they present.

The pedestrian bounding boxes were detected automatically using a Faster R-CNN [35], which improves on the DPM [11] object detector that was the standard in earlier works [57, 23] and provides fairly accurate detection without being as time-consuming as manually handcrafting the bounding boxes.

Finally, the datasets contains 126,441 bounding boxes spanning 4,101 identities, which is a significant improvement over earlier datasets, all of which contained less than 2,000 identities [57, 23, 22, 14], as can be seen in Table 2.1.



(a) Identities and bounding boxes on each camera.

(b) Identities and bounding boxes for each time slot.

Figure 2.8: [47] Some statistics on the MSMT17 dataset.

2.5 Airport

Airport [19] is a person re-identification dataset constructed from footage recorded inside a major airport. The footage was recorded over the course of 12 hours in a single day by six cameras placed after the airport's security checkpoints. Each 12 hour video was randomly split into 40 five minute clips which were then used to create the dataset.

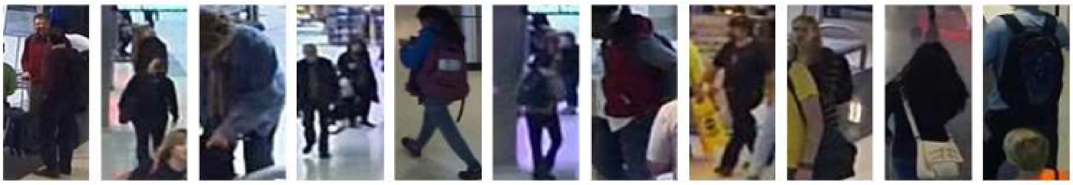


Figure 2.9: [19] Some bounding boxes from the Airport dataset.

Airport's first contribution lies in its camera network and its environment: the cameras used were the airport's actual security cameras, not a camera system setup by a research team in order to collect a person re-identification dataset. The security oriented setup of this camera network introduces new challenges that had not been previously considered. For example, in most earlier works, the cameras were set up parallel to the ground, whereas in an actual security system the cameras would be set much higher, close to the ceiling. Furthermore, most earlier datasets [47, 23] used footage recorded on a university campus, thus limiting the variety of people recorded. In contrast, the inside of an airport contains a much greater variety of people, and much more erratic crowd dynamics (which generally depend on flights schedules). The cameras used in legacy security system generally record on a lower quality than those used in earlier datasets, as can be seen in Figure 2.9.

The second contribution lies in the data annotation process. The dataset was created during the experimental deployment of a real time intelligent surveillance system [6], and thus the bounding boxes were automatically generated using the ACF framework [9] to detect people and a mix of the KLT tracker [28] and the FAST corner features [37] to do the tracking. This naturally introduces false alarms and misaligned images into the dataset. Of the 9,651 identities contained in the dataset, 1,382 reappear in at least two cameras, the remaining unpaired identities were added to the dataset to make it more challenging and realistic.

Disclaimer The ALERT Airport Re-Identification Dataset used in the research related to this publication was generated and provided by ALERT (Awareness and Localization of Explosives-Related Threats), a Department of Homeland Security Center of Excellence (COE). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

2.6 ENTI-ReID

ENTIRe-ID [54] is, to the best of our knowledge, the most recent and most extensive supervised person re-identification dataset, having been released in 2024 and containing more than four million images.

Most person re-identification datasets are, as can be seen in Table 2.1, rather small in size when compared to other benchmarking datasets used in other areas of computer vision (or deep learning in general). This contributes, as will be discussed in Chapter 4, to the presence of a rather sizeable domain gap, which remains to this day the main obstacle to implementing person re-identification into real-world scenarios.

To contrast this, the ENTIRe-ID datasets contains 4.45 million images spanning 13,540 identities. The source footage was taken from 37 publicly available internet cameras located in four different continents. This massively improves the standard for dataset size: the closest ones are the LaST dataset [42], which is comprised of 228,156 images spanning 10,862 identities, but was created from movie footage, and the MARS dataset [56], which is a video extension of the Market-1501 dataset containing more than one million of bounding boxes, but spanning only 1,261 identities captured by 6 cameras.

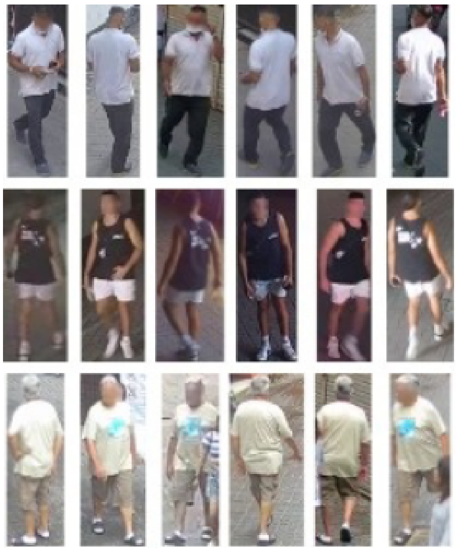


Figure 2.10: [54] Examples of images from the ENTIRe-ID dataset.

The 37 cameras used to capture the source footage were, as said, spread across four different continents. This contributes greatly to diversify the environments captured by each camera, which are influenced by both environmental factors (e.g. rain, fog, snow, sunlight, seasons, etc.) as well as cultural factors, which have a great impact on people's appearance. A larger scale also allows to capture a wider variety of activities, such as riding vehicles (cars, bikes, bicycles, etc.), carrying items and walking pets on leashes.

Given the impracticality of manual labeling for such a large dataset, an object detection model (the YOLOv8 [18] model) was used, and only individual crops that exceeded a threshold in confidence scores and pixel size were considered for the dataset. Frames from each camera sequence were sampled, this was done both to improve efficiency and account for the fact that consecutive frames would display very little variance, and an object tracking algorithm (the ByteTrack [55] algorithm) was used to correctly identify people in consecutive frames. To make the algorithm more robust (accounting for both object detection misfires and FPS drops in the live streams), all frames from the streams were included in the object tracking process, even those containing confidence scores below the threshold. This process results in person sequences for each camera, each containing a maximum of 250 images. The researchers composed the sequences and minimized the errors within by hand.

Lastly, the face of each person contained in the dataset was blurred. This was done both to preserve the subject's privacy as well as to push models away from learning facial features which, depending on factors like pose, clothing and occlusion, may not always be available.

2.7 Data Augmentation

Finally, given the relatively modest size of most person re-identification datasets, data augmentation plays an important role during training, and thus we will discuss here two data augmentation techniques that have become the norm in training person re-identification models.

2.7.1 Horizontal Flip

Horizontal flipping is one of the most basic forms of data augmentation in person re-identification and computer vision in general. It consists of flipping the image horizontally with a given probability. In person re-identification, it can help the model to better generalize pose information.

2.7.2 Random Erasing Augmentation

As discussed above, pedestrian bounding boxes often contain some form of occlusion, which contributes to the complexity of the problem. In order to combat this, researchers have attempted to include a wide variety of occlusion types in the datasets (e.g. vehicles, railings, crowds, object detectors producing misaligned images, etc.). Nevertheless, most person re-id datasets are rather limited in scale, so to further push the model to adapt to partial occlusion, the **R**andom **E**rasing **A**ugmentation, REA for short, was proposed by Zhong et al. [59]. REA works as follows: each image will have a probability of undergoing REA equal to p_e, and those that do undergo REA have a randomly selected rectangle erased with random pixels, as can be seen in Figure 2.11. The effects of REA have been extensively studied by Luo et al. [29]. The study was conducted using a ResNet50 [15] as a backbone and the Market-1501 and the DukeMTMC-reID [36] datasets for training. During training each image was resized to 256x128 pixels, padded with 10 zero value pixels and then randomly cropped to a 256x128 image.

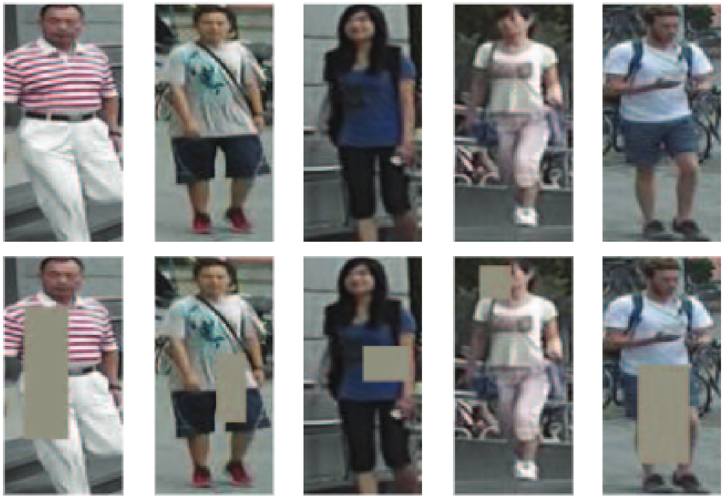


Figure 2.11: [29] Bounding boxes undergoing REA.

The images then underwent horizontal flipping with a 50% chance. Finally, each image was encoded in 32-bit float values in the range $[0, 1]$, and the RGB channels were normalized. The only training trick used in the baseline was a warmup period for the learning rate of 10 epochs. The impact of REA was measured in terms of mAP and CMC rank-1 (see Section 3.1) in two different settings: the *same dataset* setting, where the model was trained and tested on the same dataset, and the *cross-dataset* setting, where the model was trained on one of the two datasets and tested on the other.

Within the same domain, REA was found to improve both mAP and CMC rank-1 by between 1% and 4%, as can be seen in Table 2.2. However, in the cross-dataset setting, applying REA appears to harm the model's performance by a similar margin. This is conceivably due to the model overfitting the training set in order to compensate for the extra occlusion [29].

Settings		**Market**		**DukeMTMC**	
		mAP	rank-1	mAP	rank-1
Same Dataset	Baseline	75.2	88.7	65.1	80.6
	+REA	79.3	91.3	68.3	81.5
Cross-Dataset	Baseline	17.4	39.7	14.1	26.3
	+REA	13.5	32.5	10.2	21.5

Table 2.2: The effects of REA [29]. The cross-dataset results are presented for the testing dataset (using the other for training).

Chapter 3

Models

Modern research into person re-identification relies heavily (one could even say entirely) on deep learning. As said, person re-id consists of finding images in a gallery set containing the same person as a given query image. This is achieved by making the models learn deep discriminative features, which are then used to sort the gallery set according to the distance of each sample from the query. This sorted list is what will be given as output and will, hopefully, contain most, if not all, the gallery samples containing the query. The model's performance will be evaluated in how many *correct* samples are put at the *beginning* of the sorted list (what this means in more mathematical terms will be explored in Section 3.1).

Unfortunately, the field is too vast for us to give a complete overview of all possible models and all the different training techniques that have been attempted, so, in this chapter, we will give a brief overview of the metrics used to evaluate the models, since these stay consistent across all research and will also be used in Chapter 4, and we will focus on the CLIP framework by Radford et al. [34] and its very promising applications in the field of person re-id [21].

3.1 Metrics

We will begin by explaining the two most common metrics used in person re-id, which are called Cumulative Matching Characteristics and Mean Average Precision. They are both metrics that can be used to evaluate retrieval systems, meaning systems whose aim is to retrieve items *similar* to a given query (what an *item* is and what the world *similar* means depends on the task, in person re-id, for example, the items are images depicting a single person and the word similar means an image that shares the same identity as the query image). These functions are essential for evaluating a trained model's performance, they are widely used across all person re-id research and will be used for our experiments in Chapter 4.

3.1.1 Cumulative Matching Characteristics

Cumulative Matching Characteristics Rank-k, often shortened to CMC_k or *Rank-k*, where k will be a positive integer, is, together with mean average precision, one of the two most used metrics to evaluate person re-id models. It represents the probability that a correct sample appears in the highest k scored gallery samples [53, 5, 38].

In mathematical terms, given a dataset $D = (T, Q, G)$ (training set, query set and gallery set respectively), a model M, and a distance function d we have that

$$CMC_k(M, Q, G, d) = \frac{1}{|Q|} \sum_{q \in Q} acc_k(q, G')$$

Where $| \cdot |$ indicates the cardinality of a set and G' will be the output of model M on G sorted according to the distance from the output of model M on the query sample q (in layman's terms, G' will be the gallery set sorted according to how close the model currently thinks each gallery image is to query image q), and where acc_k is the accuracy at k function:

$$acc_k(x, S) = \begin{cases} 1 & \text{if a sample matching } x \text{ is in the top-k samples of } S \\ 0 & \text{otherwise} \end{cases}$$

As said, in the case of person re-id, *”a sample matching x”* means an image containing the same person as x, but the definition can be easily extended for classification, or any other supervised retrieval task.

3.1.2 Mean Average Precision

The CMC_k score gives the probability that a match for a given query image is present the first k samples of a given scored list. It is well defined if, for every query image q, there is only one gallery image g_q that has the same identity as q (which is called a *ground truth* for q), as is the case for older datasets like VIPeR [14]. In more modern datasets however, for each query image multiple ground truths are present in the gallery (e.g. the Market-1501 dataset contains an average of 14.8 ground truths for each query sample [57]). Mean Average Precision [57] is a performance metric often used to evaluate retrieval systems. It is, together with CMC_k, the most used metric to evaluate person re-id models. It is calculated, perhaps unsurprisingly, as the mean of the average precision over all queries. The average precision is, in turn, the average value of the precision of the model over all the queries. We will therefore start with explaining the precision function P_k (sometimes written as $P@k$).

Given a sorted scored list $G' = (g_1, \ldots, g_n)$ of items, each labeled by a ground truth t_i, a query sample q and its ground truth t_q (in the case of person re-id the samples will be images and the ground truths will be the person's identity) then we can define the precision at k as the ratio of true positives within the first k samples of G' and k. A sample g_i is considered a true positive if its ground truth t_i matches t_q. Thus the number of true positives within the first k samples can be written as:

$$TP_k = |\{g_i \in (g_1, \ldots, g_k) : t_i = t_q\}|$$

Where $| \cdot |$ is the cardinality of a set. We can then write the precision at k for query sample q as:

$$P_k = \frac{TP_k}{k}$$

We also need to define the relevance-at-k function:

$$rel_k = \begin{cases} 1 & \text{if } t_k = t_q \\ 0 & \text{else} \end{cases}$$

With $k \in \{1, \ldots, n\}$. Then we can define the average precision for a given query q, its ground truth t_q and the relative scored list as:

$$AP_q = \frac{1}{TP_n} \sum_{i=1}^{n} P_k \cdot rel_k$$

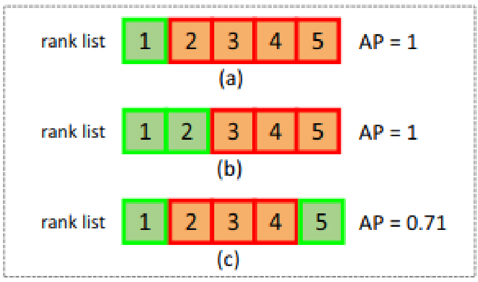


Figure 3.1: [57] An example of Average Precision being different from CMC (which will be 1.0 for all three rows). True positives are in in green and false positives are in red.

Once this is done the mean average precision is simply the mean of the AP function over all the query samples [38].

$$mAP = \frac{1}{|Q|} \sum_{q \in Q} AP_q$$

3.2 CLIP

Deep learning models are usually trained to perform a very specific task, and are generally poor when applied, without any further training, to a new task [34]. In some cases a model will perform poorly even on a different dataset that lies within the same task, this is the case for person re-id [47, 52]. Another example of this comes from image classification, which is the task of pairing a given image with a label, taken from a predetermined set of labels. Image classifiers that are trained and achieve extremely high results on benchmark datasets often perform poorly on more challenging sets of data, for example images that underwent some form of distortion [8], texture alterations [12], objects presented in unfamiliar poses [2], and when presented with real-world images that are not as curated as the ones presented in common benchmarking datasets [4] (e.g. occlusion, varied backgrounds, misalignment and pose variations). Additionally, most classifiers are not usable on a new set of labels (different than the ones present in the training dataset) without further training and, thus, additional data.

In addition to this, the datasets, which are essential to the training, are effort and time consuming to produce (in person re-id in particular, data annotation can take a long time, for example it took a year to manually annotate all the trajectories in the DukeMTMC-reID dataset [36], which is rather small when compared to most modern classification benchmarks).

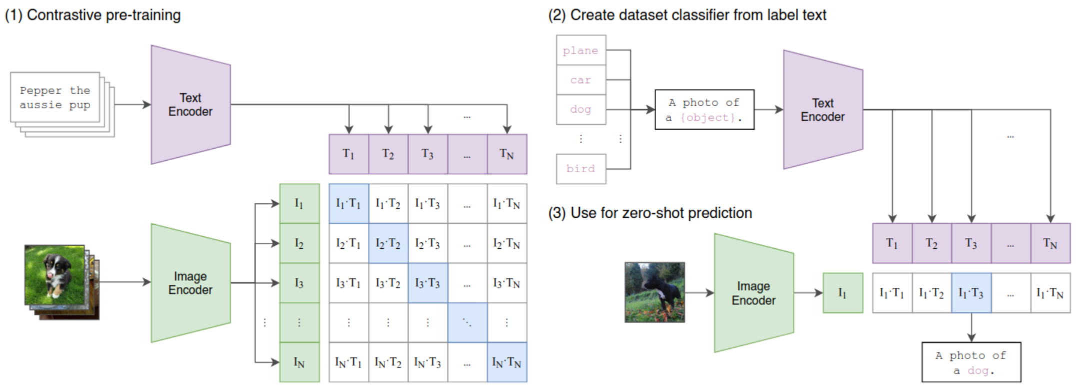

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 I_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ ... $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ ... $I_3{\cdot}T_N$
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ ... $I_N{\cdot}T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
A photo of a dog.

Figure 3.2: [34] (1) The CLIP training process. (2) CLIP being adapted to image classification. This is done by crafting a prompt for each class label; CLIP will then predict which of the prompts goes with the image.

Contrastive Language Image Pre-training (CLIP for short) [34] is a framework from OpenAI which aims at pairing visual features with natural language. This allows the model to learn visual features from natural language, which has a nearly unlimited scope, as opposed to a fixed number of classes. The model learns generalized features and can more easily adapt to never-before seen data or tasks without the need for further training, which is referred to as zero-shot capabilities.

Architecture

CLIP works by training two models, a text encoder and an image encoder, which will map the given text and the given image respectively into a low-dimensional shared embedding space. The text encoder is a Transformer [45], which works on the encodings of each token made with a lower case BPE [40], using a vocabulary size of 49,152. The image encoder consists of either a ResNet [15] or a ViT [10]. Both these models produce a low-dimensional (in this context "*low*" means low when compared to the size of the original input) output vectors, which are then fed to two linear layers which project them into the shared embedding space.

Dataset

As existing datasets were either too small (like MS-COCO [24]) or lack high quality metadata (like YFCC100M [43]) which could be used to create the (image, text) pairs, the dataset used to train CLIP was created ad hoc. An additional, more practical, advantage of using natural language for image supervision is the vast availability of such data on the internet, which allows to gather large amounts of data with minimum human effort required. In order to exploit this, the dataset was constructed from 400 million (image, text) pairs extracted from the internet, using a set of 500,000 textual queries. The queries were constructed starting from words present in at least 100 Wikipedia articles and the resulting dataset has a word-count similar to that of WebText [33, 34].

Training

Training a model to learn an open-ended set of visual concepts through natural language supervision requires high efficiency. This is because modern image classifiers, which train for a limited number of labels, already require a large amount of resources to train. For example, Xie et al. [51] had to train their EfficientNet-L2 on a Cloud TPU v3 Pod containing 2048 cores for six days. In order for CLIP's training to be feasible without employing an inordinate amount of resources, a contrastive objective was used in place of a generative one: instead of predicting the exact words that would go with a given image, during training CLIP is given a batch of N images and N pieces of text, and it will try to predict which of the possible $N \times N$ pairings actually occurred. This has been shown to learn higher quality visual features compared to generative objectives [44], which also require a higher amount of resources.

This is done with a contrastive loss: a loss that will minimize the distance between some feature embeddings, referred to as positive pairs, while maximizing the distance between others, referred to as negative pairs. For CLIP, this will be used to bring the features of an image and of a piece of text closer together if the (image, text) pair appears in the batch, otherwise it will increase the distance between them.

In practice, CLIP learns a multi-modal embedding space, which is shared by both the image and the text features, and it maximizes the cosine similarity of the true N pairs of the batch while minimizing it for the remaining $N^2 - N$ negative pairs. In more mathematical terms, given a batch of size $B \in \mathbb{Z}^+$ composed of $\{img_1, \ldots, img_B\}$ images and $\{text_1, \ldots, text_B\}$ pieces of text, the similarity between image img_i and text $text_i$ is computed as:

$$S(V_i, T_i) = g_V(\mathcal{I}(x_i)) \cdot g_T(\mathcal{T}(text_i)) \tag{3.1}$$

Where [21, 34]:

- $\mathcal{I}(\cdot)$ and $\mathcal{T}(\cdot)$ are the functions computed by the image and text encoders, respectively.

- $g_V(\cdot)$ and $g_T(\cdot)$ are the linear layers that project the given (image and text, respectively) embedding into the shared embedding space.

These similarities are optimized using two contrastive losses, called image-to-text and text-to-image respectively [34, 21]:

$$\mathcal{L}_{i2t}(i) = -\log \frac{\exp\left(S(V_i, T_i)\right)}{\sum_{\substack{a=1 \\ a \neq i}}^{B} \exp\left(S(V_i, T_a)\right)} \tag{3.2}$$

$$\mathcal{L}_{t2i}(i) = -\log \frac{\exp\left(S(V_i, T_i)\right)}{\sum_{\substack{a=1 \\ a \neq i}}^{B} \exp\left(S(V_a, T_i)\right)} \tag{3.3}$$

In these two losses, the numerator uses the similarity of the $image, text$ pair that actually matches, and thus has to be maximized, whereas the denominator uses the similarities between the other, non-matching, pairs, which will be minimized [21].

Zero-Shot Transfer

In the context of CLIP, zero-shot transfer is the generalization to unseen datasets [34], which in turn can be seen as the model's ability to learn a *task* rather than just optimizing for a benchmark. This can be seen as a proxy for zero-data learning, as theorized by Larochelle, Erhan, and Bengio [20].

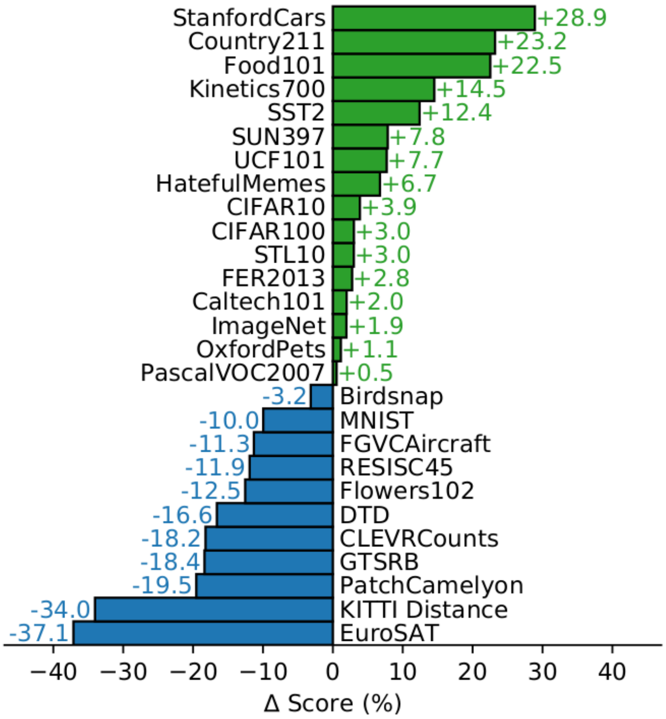


Figure 3.3: [34] Zero-shot CLIP scores compared to the scores of a fully supervised baseline.

CLIP can easily be adapted to image classification by creating a textual prompt for each class, for example ”An image of a {object}”, where the token ”{object}” will be replaced with the name of the class (e.g. cat, dog, car, etc.), the class predicted by CLIP for a given image will be the one corresponding to the textual prompt that CLIP deemed the more likely (e.g. if the prompt ”An image of a dog” was the more likely, then the output will be the class ”dog”, as shown in Figure 3.2 (2)). This can sometimes be hard to do in datasets where classes are simply numerical IDs, without an explicit textual description, as is the case in some image classification datasets (e.g. Flowers-102 [32]).

Zero-shot CLIP has shown competitive results with a fully supervised baseline (a trained ResNet [15]), outperforming it in 16 of the 27 datasets used, as can be seen in Figure 3.3, which is an extremely promising result considering that it is obtained in a Zero-Shot setting, a setting where CLIP had not been trained on any of the data beforehand.

3.3 CLIP-ReID

As stated in the previous section, CLIP works by jointly training an image and a text encoder to learn a multi modal embedding space. This can also be viewed as using the text encoder as a hyper-network to generate the weights for a linear classifier based on a description of the image made using natural language [34]. This, in turn, means that CLIP can learn a much richer set of visual features encoding higher semantics from the textual description. This can be useful for person re-id, where a model's performance relies heavily on the quality of the extracted features.

Furthermore, the CLIP architecture has shown extremely promising results in a zero-shot setting, particularly in zero-shot classification, meaning the classification of never-before seen classes, even surpassing fully-supervised baselines. Person re-id can, in some sense, be regarded as a zero-shot classification task, where the *classes* are simply the people's IDs: the models are trained on a dataset containing a finite number of IDs, but will then be evaluated (and faced in the real world) with people that were not present in the training set, meaning new, unseen IDs.

While exploiting CLIP seems like a promising premise, in practice it presents some challenges. Namely, the person labels used in person re-id are simple numerical values which lack any concrete natural language description. This has first been explored by Li, Sun, and Li [21], whose work we will now explain in more detail.

Architecture and Training

As said, the main challenge to adapting CLIP to person re-id is the lack of concrete class labels, which will just be numerical IDs representing each person captured in the dataset, and thus will be devoid of any concrete textual description. This is overcome by dividing the training in two stages:

- During the first stage, the two encoders are kept frozen (meaning their weights do not get altered), and a set of learnable tokens $[X]_1, \dots, [X]_M$ is optimized by using it to construct a textual prompt "A photo of a $[X]_1 \dots [X]_M$ person", as can be seen in Figure 3.4 (c). This effectively overcomes the need for a concrete text label. The idea of using learnable tokens to create CLIP's textual prompt was first introduced by Zhou et al. [60] (Figure 3.4 (b)), who sought to overcome the limitations of manual prompt engineering.

- During the second stage, the learnable token's weights are frozen and the image encoder's weights are unfrozen, as can be seen in Figure 3.4.

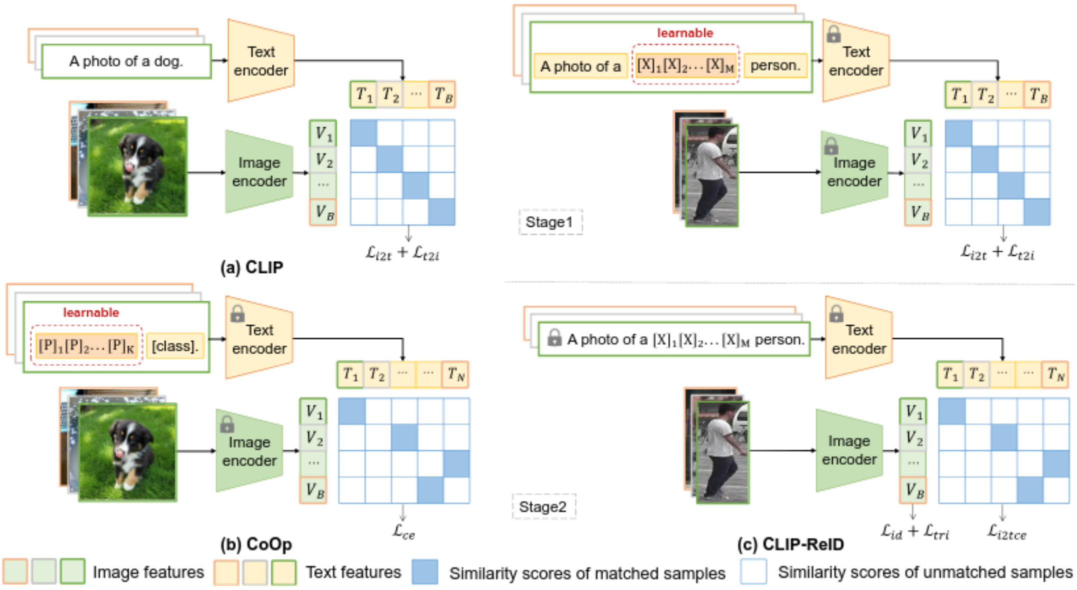


Figure 3.4: [21] (a) The CLIP [34] training process. (b) The CoOp [60] approach which adds learnable tokens to construct the text encoder's prompt, this removes the need for prompt engineering. (c) The CLIP-ReID [21] training process, during the first stage the tokens are learned for each identity in the dataset, during the second stage the image encoder is optimized.

Each image img_i in the training dataset is expected to have a person ID (pid for short) associated with it, y_i. This pid is used to compute the textual prompt (specifically the tokens) associated with image img_i as "An image of a $[X]_1 \dots [X]_M$ person."

First Stage During the first training stage, all parameters in the network, save for the learnable tokens, are kept frozen. The learnable tokens $[X]_1, \dots, [X]_M$, with M being a hyper-parameter ($M = 4$ was empirically determined to yield the best results [21]), are optimized using the same two losses as CLIP (Equations (3.2) and (3.3)), with only some minor modifications. In the similarity definition (Equation (3.1)), $text_i$ is replaced with $text_{y_i}$, since the textual prompt solely depends on the person's identity. Furthermore, unlike CLIP, two different images may share the same pid, and thus share the same textual prompt, so the $\mathcal{L}_{t2i}$ loss was changed to:

$$\mathcal{L}_{t2i}(i) = -\frac{1}{|P(y_i)|} \sum_{p \in P(y_i)} \log \frac{\exp\left(S(V_i, T_{y_i})\right)}{\sum_{\substack{a=1 \\ a \neq i}}^{B} \exp\left(S(V_a, T_{y_i})\right)} \tag{3.4}$$

Where $P(y_i) = \{p \in \{1, \dots, B\} : y_p = y_i\}$ is the set of all positives for y_i present in the batch. Then, just like in CLIP, the final loss was calculated as the sum of these two

Algorithm 1 [21] Pseudo-code for the CLIP-ReID training process.

```
Input: Batch of $x_i$ images and the corresponding $t_{y_i}$ texts, with $i \in \{1, \dots, B\}$. The
    number $M$ of tokens to be used.
Initialize $\mathcal{I}, \mathcal{T}, g_V$ and $g_T$ from the pre-trained CLIP.
Initialize the $X_1, \dots, X_M$ tokens randomly.
while 1st stage do
    $S(V_i, T_{y_i}) = g_V(\mathcal{I}(x_i)) \cdot g_T(\mathcal{T}(t_{y_i}))$
    Optimize $[X]_1, \dots, [X]_M$ by Eq. (3.5)
end while
for all $i \in \{1, \dots, N\}$ do
    $text_{y_i} = g_T(\mathcal{T}(t_{y_i}))$
end for
while 2nd stage do
    $S(V_i, T_{y_i}) = g_V(\mathcal{I}(x_i)) \cdot text_{y_i}$
    Optimize $\mathcal{I}$ by Eq. (3.10)
end while
```

losses [21]:

$$\mathcal{L}_{stage1} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i} \tag{3.5}$$

Second Stage During the second stage only the image encoder $\mathcal{I}(\cdot)$ is optimized, with the rest of the network being kept frozen. The optimization was done in accordance with the work of Luo et al. [29], employing both triplet and ID loss.

ID Loss is the person re-id equivalent of cross-enthropy, which is normally calculated as [29]:

$$\mathcal{L}_{id}(ID) = \sum_{i=1}^{N} -q_i \log p_i \tag{3.6}$$

Where p_i are the prediction logits for class i and q_i is the following constant:

$$q_i = \begin{cases} 1 & \text{if } i = ID \\ 0 & \text{otherwise} \end{cases}$$

However, in person re-id the model will be evaluated on identities that were not present in the training set, and thus it is important to prevent overfitting, especially considering the small size of most datasets, so a different definition of q_i was used, which again was introduced by Luo et al. [29]:

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon & \text{if } i = ID \\ \varepsilon / N & \text{otherwise} \end{cases} \tag{3.7}$$

Where $\varepsilon \in \mathbb{R}^+$ is a small positive constant, which helps prevent overfitting of the training set.

Triplet loss is a contrastive loss calculated as [29, 21]:

$$\mathcal{L}_{tri} = [d_p - d_n + \alpha]_+ \tag{3.8}$$

Where $[x]_+ = max\{x, 0\}$, d_p and d_n are the distances of the *anchor* from the *positive* and the *negative* respectively (meaning the distance of the image being considered from another image depicting the same person, and an image depicting a different person), and α is a margin constant [29].

Additionally, the image-to-text loss from CLIP, Equation (3.2), can still be used, allowing to better exploit the features generated by the text encoder, albeit with some slight modifications [21]:

$$\mathcal{L}_{i2tce}(i) = -\sum_{k=1}^{N} q_k \log \frac{\exp\left(S(V_i, T_{y_k})\right)}{\sum_{\substack{y_a=1 \\ y_a \neq y_k}}^{N} \exp\left(S(V_i, T_{y_a})\right)} \tag{3.9}$$

Where q_k is the same label smoothing mechanism used in ID Loss, Equation (3.7). The final loss used in the second stage is simply defined as [21]:

$$\mathcal{L}_{stage2} = \mathcal{L}_{id} + \mathcal{L}_{tri} + \mathcal{L}_{i2tce} \tag{3.10}$$

Evaluation

The experiments conducted by Li, Sun, and Li [21] were made using the same text encoder provided by CLIP [34] and two different image encoders: a CNN (ResNet-50 [15]) with a global attention pooling layer, and a vision transformer (a ViT-B/16 [10]). Two linear layers were used on the image and text features to match their dimensions. Each of these models was evaluated using the mAP and CMC Rank-1 metrics (Section 3.1) on four different datasets: MSMT17 (Section 2.4 [47]), Market-1501 (Section 2.2 [57]), DukeMTMC-reID (Section 2.3 [36]) and Occluded-Duke [31]. For comparison, each backbone (CNN and ViT) was also trained as a stand-alone model to better estimate the impact of the CLIP-ReID framework.

The results are shown in Tables 3.1 and 3.2. It is apparent that the addition of the CLIP-ReID framework provides a performance boost, which can be further increased by adding Side Information Encoding (SIE) and Overlapping Patches (OLP) in the ViT model [21]. The scores obtained on MSMT17 [47] in particular are, at the time of writing, the highest scores achieved on the dataset on the Papers With Code website [1].

Model	**Method**	**MSMT17** [47]		**Market-1501** [57]	
		mAP (%)	Rank-1 (%)	mAP (%)	Rank-1 (%)
CNN	Baseline	60.7	82.1	88.1	94.7
	CLIP-ReID	63.0	84.4	89.8	95.7
ViT	Baseline	66.1	84.4	86.4	93.3
	CLIP-ReID	73.4	88.7	89.6	95.5
	CLIP-ReID+SIE+OLP	75.8	89.7	90.5	95.4

Table 3.1: [21] CLIP-ReID results on the MSMT17 [47] and Market-1501 [57] datasets.

Model	**Method**	**DukeMTMC-reID** [36]		**Occluded-Duke** [31]	
		mAP (%)	Rank-1 (%)	mAP (%)	Rank-1 (%)
CNN	Baseline	79.3	88.6	47.4	54.2
	CLIP-ReID	80.7	90.0	53.5	61.0
ViT	Baseline	80.0	88.8	53.5	60.8
	CLIP-ReID	82.5	90.0	59.5	67.1
	CLIP-ReID+SIE+OLP	83.1	90.8	60.3	67.2

Table 3.2: [21] CLIP-ReID results on the DukeMTMC-reID [36] and Occluded-Duke [31] datasets.

The framework was also evaluated on two Vehicle Re-Identification (Section 1.1) datasets, VeRi-77 [26] and VehicleID [25].

Chapter 4

Experimental Results

In order to investigate the applicability of person re-id models into the real world, we sought to evaluate their adaptability to new scenarios which could differ widely from the conditions presented into any training dataset. Unsurprisingly, the great amount of variation that person re-id models have to account for (both due to human appearance itself, the way this appearance is recorded and other environmental factors) cannot be fully captured by a single dataset [52], and it has been found that models trained and evaluated on a dataset will perform poorly on others. As we briefly mentioned in Section 1.2, this is referred to as domain gap.

This domain gap is a known challenge in person re-id [47, 52] and other computer vision tasks [34] and is a fundamental obstacle to the deployment of a real-world re-identification system. We therefore endeavored to accurately measure its effects on the performance of the CLIP-ReID models [21], and perform some small experiments attempting to mitigate them. Attempts towards bridging this gap have already been made, usually in the form of domain adaptation [47, 52] or achieving more generalized features, and although some success has been found, this challenge remains far from being solved.

4.1 Methodologies

We choose to focus on the CLIP-ReID (Section 3.3 [21]) framework and models both due to the high scores achieved (particularly on MSMT17 [47], which remains one of the most challenging supervised person re-id datasets) and due to the fact that person re-id can be seen as a zero-shot classification task (meaning the classification of unseen labels, as explained in Section 3.2 and 3.3), in which CLIP has shown extremely promising results [34]. Furthermore, these models include two simple backbones, made following the baseline set by Luo et al. [29], which underlies many other person re-id frameworks and models, and thus we believe that the experimental results obtained on them can be expected to remain consistent across other modern state-of-the-art models.

In order to assess the model's performance on new data, we took the models trained on the Market-1501 [57] and DukeMTMC-reID [36] datasets provided by Li, Sun, and Li [21], and measured their performance (through the mAP and CMC Rank-1 metrics, see Section 3.1) on different datasets. These tests are defined as *Cross-Dataset Tests*, and are performed using the following datasets: CUHK03 [23], Airport [19], the test set of ENTIRe-ID [54], Market-1501 [57] (for the models trained on DukeMTMC-reID) and DukeMTMC-reID [36] (for the models trained on Market-1501). The results of these tests are shown in the following section.

We used these cross-dataset tests to determine, for each model, the dataset it performed the poorest on, and inferred this dataset to be the one most ”distant” from the model's original training dataset. We then merged this dataset with the original training dataset and re-trained the model on this union, in hopes that this could allow the model to learn more generalized features, and thus diminish the performance degradation.

4.2 Cross-Datasets Tests

Model	**Test Set**	**mAP** (%)	**Rank-1** (%)
ViT Baseline	Airport [19]	16.0	18.4
	ENTIRe-ID [54]	29.3	29.2
	DukeMTMC re-ID [36]	44.8	64.3
	CUHK03 [23]	34.9	36.9
ResNet Baseline	Airport [19]	3.8	5.1
	ENTIRe-ID [54]	11.9	11.7
	DukeMTMC-reID [36]	17.3	30.8
	CUHK03 [23]	8.7	8.6
ViT CLIP-ReID	Airport [19]	20.0	22.4
	ENTIRe-ID [54]	38.9	38.4
	DukeMTMC re-ID [36]	50.2	68.9
	CUHK03 [23]	38.6	40.4
ResNet CLIP-ReID	Airport [19]	4.8	6.0
	ENTIRe-ID [54]	13.3	12.8
	DukeMTMC-reID [36]	21.0	36.8
	CUHK03 [23]	8.8	8.9

Table 4.1: Cross-dataset scores for CLIP-ReID models [21] trained on the Market-1501 [57] dataset.

Table 4.1 shows the results of the cross-dataset tests (meaning tests made on a dataset in a zero-shot setting) for all the models trained on the Market-1501 dataset [57]. Unsurprisingly, the models that employ vision transformers [10] (ViT) as their image encoders outperform their ResNet-based [15] counterparts, and the boost in performance provided by the CLIP-ReID framework seems to carry on to the cross-dataset setting. Nevertheless, when comparing these scores to the ones the models obtained on the Market-1501 [57] test set shown in Table 3.1, the drop-off in performance is massive, particularly on ResNet-based models, which experience a drop in mAP on the Airport dataset of 84.3% (ResNet Baseline) and 83.3% (ResNet CLIP-ReID). The highest cross-dataset mAP score was 50.2%, obtained by the ViT CLIP-ReID model on the DukeMTMC-reID dataset, which is still 39.4% lower than what the exact same checkpoint had achieved on the Market-1501 test set, as shown in Table 3.1. The CMC Rank-1 score suffered a similar degradation across the board.

The checkpoints trained on the DukeMTMC-reID [36] obtained similar cross-dataset scores, as shown in Table 4.2, but a bit lower than those obtained by their Market-1501-trained counterparts. This leads us to believe that market is, despite its smaller size, more complex than duke and thus models learn slightly more general features from it.

Model	**Test Set**	**mAP** (%)	**Rank-1** (%)
ViT Baseline	Airport [19]	10.0	12.9
	ENTIRe-ID [54]	33.6	32.9
	Market-1501 [57]	37.4	62.2
	CUHK03 [23]	19.2	20.4
ResNet Baseline	Airport [19]	4.5	6.3
	ENTIRe-ID [54]	19.7	19.4
	Market-1501 [57]	21.5	47.5
	CUHK03 [23]	6.1	5.9
ViT CLIP-ReID	Airport [19]	14.3	17.4
	ENTIRe-ID [54]	40.2	40.3
	Market-1501 [57]	43.4	70.7
	CUHK03 [23]	25.9	27.6
ResNet CLIP-ReID	Airport [19]	5.3	7.2
	ENTIRe-ID [54]	21.2	20.4
	Market-1501 [57]	24.6	52.2
	CUHK03 [23]	5.7	5.0

Table 4.2: Cross-dataset scores for CLIP-ReID models [21] trained on the DukeMTMC-reID [36] dataset.

Across all models and all tests we performed, Airport [19] is the dataset where the lowest scores are obtained. This is probably due to the difference in scene between airport and most other datasets: both Market-1501 and DukeMTMC-reID were created from footage taken in a university campus [57, 36], with the cameras being setup by researchers, whereas airport was created using the surveillance cameras of an actual airport. This leads to a wide difference in camera angles, image quality, activities, clothing etc., making Airport the most challenging dataset for models trained on Market-1501 and DukeMTMC-reID.

4.3 Training on the Union of Datasets

In order to help close the domain gap, we re-trained the best performing model (ViT CLIP-ReId [21]) on the union of Market-1501 [57] (the original training dataset which yielded the best results in a cross-dataset setting) and Airport [19] (the dataset that proved to be the most challenging in a cross-dataset setting) in the hopes that more general re-id features could be learned from the resulting union. We first re-trained the ViT CLIPRe-ID model, using the same training algorithm and settings used by Li, Sun, and Li [21] and repeated the cross-dataset tests with the remaining datasets. The results can be seen in Table 4.3.

Test Set	**mAP** (%)	**ΔmAP** (%)	**Rank-1** (%)	**ΔRank-1** (%)
ENTIRe-ID [54]	46.2	+7.3	45.9	+7.5
DukeMTMC-reID [36]	53.5	+3.3	71.6	+2.7
CUHK03 [23]	41.8	+3.2	42.6	+2.2
Market-1501 [57]	89.0	-0.6	95.0	-0.4
Airport	64.0	+44.0	56.9	+34.5

Table 4.3: Cross-dataset test result for the ViT CLIP-ReID model [21] trained on the union of Market-1501 [57] and Airport [19]. The difference in scores (the Δ columns) are given with respect to the corresponding cross-dataset scores obtained by the same model trained on the Market-1501 dataset alone, as shown in Table 4.1.

The idea of training a model on the union of two or more datasets is not new. For example, Marchwica, Jamieson, and Siva [30] sought to achieve scene-independent re-id by training on a larger amount of data, obtained by merging between two and six datasets. It is worth mentioning that, following the work of Xiao et al. [50], they balanced the union of datasets by keeping ten images per person, ensuring to keep images from all cameras the person appeared in (except for when the person appeared in more than ten cameras, in which case one image per camera is kept). This is done to account for the fact that different datasets might have more or less images per person, which could lead to overfitting one of the datasets [30]. However, when merging only two datasets, this balancing can lead to a great reduction in the number of images used, so we opted to not balance the union.

Additionally, Luo et al. [29] found that Random Erasing Augmentation (Section 2.7.2), which is used in the CLIPRe-ID training process [21], harms model performance in a cross-dataset setting, and therefore we repeated the above training process and testing procedures without it, the results can be seen in Table 4.4.

Test Set	**mAP** (%)	**ΔmAP** (%)	**Rank-1** (%)	**ΔRank-1** (%)
ENTIRe-ID [54]	43.6	+4.7	43.1	+4.7
DukeMTMC-reID [36]	54.3	+4.1	72.3	+3.4
CUHK03 [23]	42.7	+4.1	44.3	+3.9
Market-1501 [57]	87.3	-2.3	94.6	-0.9
Airport [19]	69.4	+49.4	63.1	+40.7

Table 4.4: Cross-dataset test result for the ViT CLIP-ReID model [21] trained on the union of Market-1501 [57] and Airport [19], without using Random Erasing Augmentation [29, 59]. The difference in scores (the Δ columns) are given with respect to the corresponding cross-dataset scores obtained by the same model trained on the Market-1501 dataset alone, as shown in Table 4.1.

As can be seen in Table 4.4, the mAP score over all datasets improved by between 4.1% and 4.7% and, similarly, the CMC Rank-1 improved by between 3.4% and 4.7% for the model trained without REA. Meanwhile, adding Airport to the dataset did not considerably hinder performance on the original Market-1501 dataset, which dropped by 2.3% (mAP) and 0.9% (Rank-1). Naturally, the performance on the Airport [19] dataset was greatly improved. It is worth noting that, in order to exploit the differences between Market-1501 [57] and Airport [19] as much as possible, we created the training set by adding every Airport image annotated with a valid and reappearing person ID to Market's training set, so the scores obtained by these models on Airport should not be given much weight, they were added for the sake of completeness. Finally, as it was expected, the cross-dataset scores receive a slight improvement when removing random erasing augmentation from the training process, except on the ENTIRe-ID [54] dataset, where the model trained with REA obtained, surprisingly enough, higher cross-dataset scores.

Chapter 5

Conclusions

Despite the rapid advancements made in the person re-id field [53, 48, 52], its deployment into the real world with the same degree of accuracy obtained on benchmarks appears to still be far.

We experimentally measured the effects of the domain gap on state-of-the-art models [21] using various benchmarking datasets (including Airport [19], Market-1501 [57] and DukeMTMC-reID [36]). The evidence we found shows that the CLIP-ReID [21] models still suffer a massive performance degradation when faced with with data that presents considerable differences with the training dataset, despite their impressive performance in a more traditional setting. We attempted to bridge this gap by re-training the best performing model on the union of the two most distant dataset in an attempt to mitigate the domain gap without needing to dramatically improve the amount of labeled data required, and achieved an average mAP improvement of 4.3% and an average CMC Rank-1 improvement of 3.8%.

Acknowledgements

I would like to extend my thanks to Professor Andrea Asperti and Dr. Salvatore Fiorilla for supporting me during the writing of this thesis and lending me their precious time and expertise, as well as for the endless patience they have shown me.

I would also like to thank Giulia, who has been tirelessly supporting me for the past ten years, and without whom i would never have been able to graduate, and Veronica, who has been endlessly pestered during the writing of this thesis, yet has been kind enough to keep encouraging me.

Finally, i would like to thank all my friends and my sisters for supporting and believing in me, even when I wouldn't.

References

[1] *"Papers With Code"*. URL: https://paperswithcode.com/ (visited on 09/30/2024).

[2] Michael A. Alcorn et al. "Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4840–4849. DOI: 10.1109/CVPR.2019.00498.

[3] Ali Amiri, Aydin Kaya, and Ali Seydi Keçeli. "A Comprehensive Survey on Deep-Learning-based Vehicle Re-Identification: Models, Data Sets and Challenges". In: *ArXiv* abs/2401.10643 (2024). URL: https://api.semanticscholar.org/CorpusID:267061389.

[4] Andrei Barbu et al. "ObjectNet: a large-scale bias-controlled dataset for pushing the limits of object recognition models". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[5] R.M. Bolle et al. "The relation between the ROC curve and the CMC". In: *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*. 2005, pp. 15–20. DOI: 10.1109/AUTOID.2005.48.

[6] Octavia Camps et al. "From the Lab to the Real World: Re-identification in an Airport Camera Network". In: *IEEE Transactions on Circuits and Systems for Video Technology* 27.3 (2017), pp. 540–553. DOI: 10.1109/TCSVT.2016.2556538.

[7] Dong Seon Cheng et al. "Custom Pictorial Structures for Re-identification". In: *British Machine Vision Conference*. 2011. URL: https://api.semanticscholar.org/CorpusID:6485959.

[8] Samuel F. Dodge and Lina Karam. "A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions". In: *2017 26th International Conference on Computer Communication and Networks (ICCCN)* (2017), pp. 1–7. URL: https://api.semanticscholar.org/CorpusID:40148280.

[9] Piotr Dollár et al. "Fast Feature Pyramids for Object Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.8 (2014), pp. 1532–1545. DOI: 10.1109/TPAMI.2014.2300479.

[10] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021.

[11] Pedro F. Felzenszwalb et al. "Object Detection with Discriminatively Trained Part-Based Models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), pp. 1627–1645. DOI: 10.1109/TPAMI.2009.167.

[12] Robert Geirhos et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *ArXiv* abs/1811.12231 (2018). URL: https://api.semanticscholar.org/CorpusID:54101493.

[13] S. Gong et al. *Person Re-Identification*. Advances in Computer Vision and Pattern Recognition. Springer London, 2014. ISBN: 9781447162964. URL: https://books.google.it/books?id=7iu5BAAAQBAJ.

[14] Douglas Gray, Shane Brennan, and Hai Tao. "Evaluating Appearance Models for Recognition, Reacquisition, and Tracking". In: 2007. URL: https://api.semanticscholar.org/CorpusID:15225312.

[15] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[16] Martin Hirzer et al. "Relaxed Pairwise Learned Metric for Person Re-identification". In: *Computer Vision – ECCV 2012*. Ed. by Andrew Fitzgibbon et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 780–793. ISBN: 978-3-642-33783-3.

[17] Xinyang Jiang et al. "Rethinking Temporal Fusion for Video-Based Person Re-Identification on Semantic and Time Aspect". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (Apr. 2020), pp. 11133–11140. DOI: 10.1609/aaai.v34i07.6770.

[18] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *Ultralytics YOLO*. Version 8.0.0. Jan. 2023. URL: https://github.com/ultralytics/ultralytics.

[19] Srikrishna karanam et al. "A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.3 (2019), pp. 523–536. DOI: 10.1109/TPAMI.2018.2807450.

[20] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. "Zero-data learning of new tasks". In: *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*. AAAI'08. Chicago, Illinois: AAAI Press, 2008, pp. 646–651. ISBN: 9781577353683.

[21] Siyuan Li, Li Sun, and Qingli Li. “CLIP-ReID: Exploiting Vision-Language Model for Image Re-identification without Concrete Text Labels”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.1 (2023), pp. 1405–1413. DOI: 10.1609/aaai.v37i1.25225. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25225.

[22] Wei Li and Xiaogang Wang. “Locally Aligned Feature Transforms across Views”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 3594–3601. DOI: 10.1109/CVPR.2013.461.

[23] Wei Li et al. “DeepReID: Deep Filter Pairing Neural Network for Person Re-identification”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 152–159. DOI: 10.1109/CVPR.2014.27.

[24] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 740–755. ISBN: 978-3-319-10602-1.

[25] Hongye Liu et al. “Deep Relative Distance Learning: Tell the Difference between Similar Vehicles”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2167–2175. DOI: 10.1109/CVPR.2016.238.

[26] Xinchen Liu et al. “A Deep Learning-Based Approach to Progressive Vehicle Re-identification for Urban Surveillance”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 869–884.

[27] Chen Change Loy, Tao Xiang, and Shaogang Gong. “Multi-camera activity correlation analysis”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 1988–1995. DOI: 10.1109/CVPR.2009.5206827.

[28] Bruce D. Lucas and Takeo Kanade. “An iterative image registration technique with an application to stereo vision”. In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI’81. Vancouver, BC, Canada: Morgan Kaufmann Publishers Inc., 1981, pp. 674–679.

[29] Hao Luo et al. “Bag of Tricks and a Strong Baseline for Deep Person Re-Identification”. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019, pp. 1487–1495. DOI: 10.1109/CVPRW.2019.00190.

[30] Paul Marchwica, Michael Jamieson, and Parthipan Siva. “An Evaluation of Deep CNN Baselines for Scene-Independent Person Re-identification”. In: *2018 15th Conference on Computer and Robot Vision (CRV)*. 2018, pp. 297–304. DOI: 10.1109/CRV.2018.00049.

[31] Jiaxu Miao et al. “Pose-Guided Feature Alignment for Occluded Person Re-Identification”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 542–551. DOI: 10.1109/ICCV.2019.00063.

[32] Maria-Elena Nilsback and Andrew Zisserman. "Automated Flower Classification over a Large Number of Classes". In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. 2008, pp. 722–729. DOI: 10.1109/ICVGIP.2008.47.

[33] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[34] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[35] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pp. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

[36] Ergys Ristani et al. "Performance Measures and a Data Set for Multi-target, Multi-camera Tracking". In: *Computer Vision – ECCV 2016 Workshops*. Ed. by Gang Hua and Hervé Jégou. Cham: Springer International Publishing, 2016, pp. 17–35. ISBN: 978-3-319-48881-3.

[37] Edward Rosten, Reid Porter, and Tom Drummond. "Faster and Better: A Machine Learning Approach to Corner Detection". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.1 (2010), pp. 105–119. DOI: 10.1109/TPAMI.2008.275.

[38] Prodip Kumar Sarker, Qingjie Zhao, and Md. Kamal Uddin. "Transformer-Based Person Re-Identification: A Comprehensive Review". In: *IEEE Transactions on Intelligent Vehicles* 9.7 (2024), pp. 5222–5239. DOI: 10.1109/TIV.2024.3350669.

[39] William Robson Schwartz and Larry S. Davis. "Learning Discriminative Appearance-Based Models Using Partial Least Squares". In: *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*. 2009, pp. 322–329. DOI: 10.1109/SIBGRAPI.2009.42.

[40] Rico Sennrich, Barry Haddow, and Alexandra Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

[41] Zhiyin Shao et al. "Learning Granularity-Unified Representations for Text-to-Image Person Re-identification". In: *Proceedings of the 30th ACM International Conference on Multimedia* (2022). URL: https://api.semanticscholar.org/CorpusID:250627620.

[42] Xiujun Shu et al. "Large-Scale Spatio-Temporal Person Re-Identification: Algorithms and Benchmark". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.7 (2022), pp. 4390–4403. DOI: 10.1109/TCSVT.2021.3128214.

[43] Bart Thomee et al. "YFCC100M: the new data in multimedia research". In: *Commun. ACM* 59.2 (Jan. 2016), pp. 64–73. ISSN: 0001-0782. DOI: 10.1145/2812802. URL: https://doi.org/10.1145/2812802.

[44] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive Multiview Coding". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 776–794. ISBN: 978-3-030-58621-8.

[45] Ashish Vaswani et al. "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 6000–6010. ISBN: 9781510860964.

[46] Zhongdao Wang et al. "Towards Real-Time Multi-Object Tracking". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 107–122. ISBN: 978-3-030-58621-8.

[47] Longhui Wei et al. "Person Transfer GAN to Bridge Domain Gap for Person Re-identification". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 79–88. URL: https://api.semanticscholar.org/CorpusID:6258614.

[48] Wenyu Wei et al. "Person re-identification based on deep learning — An overview". In: *Journal of Visual Communication and Image Representation* 82 (2022), p. 103418. ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2021.103418. URL: https://www.sciencedirect.com/science/article/pii/S1047320321002765.

[49] Lin Wu et al. "Deep Coattention-Based Comparator for Relative Representation Learning in Person Re-Identification". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.2 (2021), pp. 722–735. DOI: 10.1109/TNNLS.2020.2979190.

[50] Tong Xiao et al. "Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1249–1258. DOI: 10.1109/CVPR.2016.140.

[51] Qizhe Xie et al. "Self-Training With Noisy Student Improves ImageNet Classification". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 10684–10695. DOI: 10.1109/CVPR42600.2020.01070.

[52] Ankit Yadav and Dinesh Kumar Vishwakarma. "Deep learning algorithms for person re-identification: Sate-of-the-art and research challenges". In: *Multimedia Tools and Applications* 83.8 (Aug. 2023), pp. 22005–22054. DOI: 10.1007/s11042-023-16286-w.

[53] Mang Ye et al. "Deep Learning for Person Re-Identification: A Survey and Outlook". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.6 (2022), pp. 2872–2893. DOI: 10.1109/TPAMI.2021.3054775.

[54] Serdar Yıldız and Ahmet Nezih Kasım. "ENTIRe-ID: An Extensive and Diverse Dataset for Person Re-Identification". In: *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. 2024, pp. 1–5. DOI: 10.1109/FG59268.2024.10581945.

[55] Yifu Zhang et al. "ByteTrack: Multi-object Tracking by Associating Every Detection Box". In: *Computer Vision – ECCV 2022*. Ed. by Shai Avidan et al. Cham: Springer Nature Switzerland, 2022, pp. 1–21. ISBN: 978-3-031-20047-2.

[56] Liang Zheng et al. "MARS: A Video Benchmark for Large-Scale Person Re-Identification". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 868–884. ISBN: 978-3-319-46466-4.

[57] Liang Zheng et al. "Scalable Person Re-identification: A Benchmark". In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1116–1124. DOI: 10.1109/ICCV.2015.133.

[58] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. "Associating Groups of People". In: *British Machine Vision Conference*. 2009. URL: https://api.semanticscholar.org/CorpusID:10150893.

[59] Zhun Zhong et al. "Random Erasing Data Augmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.07 (Apr. 2020), pp. 13001–13008. DOI: 10.1609/aaai.v34i07.7000. URL: https://ojs.aaai.org/index.php/AAAI/article/view/7000.

[60] Kaiyang Zhou et al. "Learning to prompt for vision-language models". In: *International Journal of Computer Vision* 130.9 (July 2022), pp. 2337–2348. DOI: 10.1007/s11263-022-01653-1.

[61] Zhengxia Zou et al. "Object Detection in 20 Years: A Survey". In: *Proceedings of the IEEE* 111.3 (2023), pp. 257–276. DOI: 10.1109/JPROC.2023.3238524.