UNIVERSIDADE FEDERAL FLUMINENSE

FELIX OLIVER SUMARI HUAYTA

Towards Practical Implementations of Person Re-Identification from Full Video Frames

NITERÓI 2021

UNIVERSIDADE FEDERAL FLUMINENSE

FELIX OLIVER SUMARI HUAYTA

Towards Practical Implementations of Person Re-Identification from Full Video Frames

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 Gonzalez Clua

Coorientador: Joris Michel Gérard Daniel Guerin

> NITERÓI 2021

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

H874t	Huayta, Felix Oliver Sumari Towards Practical Implementations of Person Re- Identification from Full Video Frames / Felix Oliver Sumari Huayta ; Esteban Walter Gonzalez Clua, orientador ; Joris Michel Gérard Daniel Guerin, coorientador. Niterói, 2021. 83 f.
	Dissertação (mestrado)-Universidade Federal Fluminense, Niterói, 2021.
	DOI: http://dx.doi.org/10.22409/PGC.2021.m.06537662702
	 Re-identificação de pessoas. Sistemas Segurança. Video Analisis. 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 -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

FELIX OLIVER SUMARI HUAYTA

TOWARDS PRACTICAL IMPLEMENTATIONS OF PERSON RE-IDENTIFICATION FROM FULL VIDEO FRAMES

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 março de 2021.

BANCA EXAMINADORA

Prof. Esteban Walter Gonzalez Clua, PhD. - Orientador,

UFF

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

Prof. Aura Conci, PhD. – UFF

Prof. João Paulo Papa, PhD. – UNESP

Niterói 2021

Á minha família toda, em especial para minha Mãe Bernardina e meu pai Félix pelo apoio e dedicação, á meus irmãos pelos sorrisos e á meus avôs por acreditar em mim.

Agradecimentos

Este trabalho foi realizado no Instituto de Informática da UFF como parte do programa de pós-graduação para obtenção do título de mestre em ciência da computação.

Em primeiro lugar, gostaria de agradecer a Esteban Clua, professor da UFF, que me deu a oportunidade de ser seu orientado e o apoio para poder estudar no PGC. Da mesma forma, a Joris Guerin, por também ser meu orientador e aconselhar-me na minha pesquisa ao longo do meu período de estudos.

Em segundo lugar, gostaria de agradecer a todos que fizeram parte dessa parte da minha vida, como meus amigos da UFF e meus amigos de Arequipa que apoiaram a mim e a todos os meus professores da UFF pelos conselhos.

Agradeço aos meus pais, Felix e Berna, assim como aos meus irmãos Erick e Nereyda, bem como a todos os membros da minha família; tios, tias, primos, primos, que sempre me apoiaram e incentivaram a continuar no caminho que decidi. Por fim, aos meus avós que estão no céu, principalmente à minha avó Zélia, que sem ela eu não seria quem sou.

Resumo

Com a grande adoção da automação para a segurança das cidades, a reidentificação de pessoas (Re-ID) vem sendo amplamente estudada recentemente. Nesta dissertação, argumentamos que a forma atual de estudar a reidentificação de pessoas, ou seja, tentando reidentificar uma pessoa dentro de imagens já detectadas e pré-recortadas de pessoas, não é suficiente para implementar aplicações práticas de segurança, onde as entradas para o sistema são os quadros completos das transmissões de vídeo. Para apoiar esta afirmação, apresentamos a configuração Full Frame Person Re-ID (FF-PRID) e definimos métricas específicas para avaliar as implementações de FF-PRID. Para melhorar a robustez, também formalizamos a estrutura de colaboração híbrida homem-máquina, que é inerente a qualquer aplicativo de segurança Re-ID. Para demostrar a importância de considerar a configuração FF-PRID, construímos um experimento que mostra-nos que combinar uma boa técnica de detecção de pessoas com um bom método Re-ID não produz necessariamente bons resultados para a aplicação final. Isso sublinha uma falha da formulação atual na avaliação da qualidade de um modelo Re-ID e justifica o uso de diferentes métricas. Esperamos que este trabalho motive a comunidade de pesquisa a considerar o problema completo, a fim de desenvolver algoritmos que sejam mais adequados a cenários do mundo real.

Palavras-chave: Aplicação de Segurança, Re-identificação de Pessoas, Detecção de Pedestres.

Abstract

With the major adoption of automation for cities security, person re-identification (Re-ID) has been extensively studied. In this dissertation, we argue that the current way of studying person re-identification, i.e. by trying to re-identify a person within already detected and pre-cropped images of people, is not sufficient to implement practical security applications, where the inputs to the system are the full frames of the video streams. To support this claim, we introduce the Full Frame Person Re-ID setting (FF-PRID) and define specific metrics to evaluate FF-PRID implementations. To improve robustness, we also formalize the hybrid human-machine collaboration framework, which is inherent to any Re-ID security applications. To demonstrate the importance of considering the FF-PRID setting, we build an experiment showing that combining a good people detection network with a good Re-ID model does not necessarily produce good results for the final application. This underlines a failure of the current formulation in assessing the quality of a Re-ID model and justifies the use of different metrics. We hope that this work will motivate the research community to consider the full problem in order to develop algorithms that are better suited to real-world scenarios.

Keywords: Security application, Person re-identification, Pedestrian detection.

List of Figures

1.1	Illustration of the Classic Person Re-ID (C-PRID) setting (Source: author).	2
3.1	Multi-camera surveillance network illustration of Re-ID (Source: Gala [3]).	12
3.2	Images of the same person taken from different cameras to illustrate the appearance changes. The top row images were captured on the same day, bottom row images were captured on different days (Source: Gala [3]).	13
3.3	The pipeline for a practical person Re-ID system, including five main steps: 1) Raw Data Collection, 2) Bounding Box Generation, 3) Training Data Annotation, 4) Model Training and 5) Person Retrieval	13
3.4	Different feature representation (Source: author).	16
3.5	Same person in different cameras represented using a multi-shot version that contains multiple images(called tracklet in other works [51]) per person. (Source: https://www.tugraz.at/institute/icg/research/team-bis/lrs/downloads/prid11)	schof/ 18
3.6	Three kinds of widely used loss functions in the literature. (Source: Mang Ye [135]).	20
3.7	Person search(End-to-End Re-ID) is about finding a query person (yel- low rectangle) within a gallery image (the target green rectangle).(Source: author)	25
3.8	YOLO flow for object detection. (Source: YOLO [84])	28
3.9	Architecture of YOLO V3. (Source: https://plos.figshare.com/articles figure/YOLOv3_architecture_/8322632/1)	/ 30
3.10	Architecture of Siam-IDL Re-ID. (Source: Ahmed [17])	31
4.1	Full Frame Person Re-ID (FF-PRID) setting. (Source: author)	35
4.2	Hybrid Human-Machine Framework and proposed Pipeline for Full Frame Person Re-Identification.(Source: author)	37

5.1	Example of a building with two different static surveillance cameras called	
	A and B. (Source: https://www.tugraz.at/institute/icg/research/	
	<pre>team-bischof/lrs/downloads/prid11/)</pre>	40
6.1	Different possible mistakes for cropping. (Source: author)	44
6.2	CMC curve on CUHK-03 validation set and on view A of PRID-2011, using a SiamIDL model trained on CUHK-03 training set. (Source: author)	45
6.3	Example images from the CUHK-03 and the PRID-2011 datasets. (Source: CUHK[54] and PRID2011[36])	46
6.4	Graphic with $\tau = 10$. (Source: author).	47
6.5	Graphic with $\tau = 100$. (Source: author)	48
6.6	Graphic with $\tau = 1000$. (Source: author).	49
6.7	Example of the interface for alert validation with $\eta = 6$. (Source: author).	50

List of Tables

3.1	Close-world vs. Open-world Person Re-ID	14
3.2	Details about commonly used datasets for closed-world person Re-ID. "both"	
	means that it contains both hand-cropped and detected bounding boxes.	
	"C&M" means both CMC and mAP are evaluated	23
6.1	Evaluation of the YOLO-v3 model for pedestrian detection on the raw	
	video B from the PRID-2011 dataset. For the Original Bounding Boxes	
	(OBB) rows, metrics were computed using only the bounding boxes avail-	
	able from the original dataset as ground truth. For the OBB $+$ Manually	
	added Bounding Boxes (MBB) rows, the bounding boxes added using the	
	LabelImg tool were also considered	44

Glossary

Re-ID	:	Re-Identification;
PRID	:	Person Re-Identification;
FF-PRID	:	Full Frame Re-Identification;
FOVs	:	Fields of views;
SIR	:	Single Image Representation;
OD	:	Object Detection;
YOLO	:	You Only Look One;
Siam-IDL	:	Siamese Improve Deep Learning
CUHK	:	Chinese University of Hong Kong
CNN	:	Convolutional Neural Network
R-CNN	:	Region based Convolutional Neural Network
SSD	:	Single Shot Multi-box Detector
PCB	:	Part-based Convolutional Baseline
MLFN	:	Multi-Level Factorization Net
DGD	:	Domain Guided Dropout
GAN	:	Generative Adversarial Net-work
CMC	:	Cumulative Matching Characteristics
mAP	:	mean Average Precision
NPSM	:	Neural Person Search Machine
\mathbf{FR}	:	Finding Rate
TVR	:	True Validate Rate
TC	:	True Call
TMC	:	True Missed Call
\mathbf{FS}	:	False Silence
\mathbf{FC}	:	False Call
TS	:	True Silence
IOU	:	Intersection Over Union
OBB	:	Original Bounding Boxes
MBB	:	Manually added Bounding Boxes

CIR	:	Cross-Image Representation
IDE	:	ID-Discriminative Embedding
SVM	:	Support Vector Machine
LAAM	:	Locality-Aware Appearance Metric
DGM	:	Dynamic Graph Matching
TAUDL	:	Tracklet Association Unsupervised Deep Learning
DSR	:	Deep Spatial feature Reconstruction
DNet	:	Distribution Net
APN	:	Adversarial Person Net
LDM	:	Logistic Distance Metric
LMNN	:	Largest Margin Nearest Neighbor
PD	:	Pedestrian Detection

Contents

1	Intr	oduction	1
	1.1	Context and Motivation	1
	1.2	Definition of the Problem	3
	1.3	Justification	4
	1.4	Objectives	4
		1.4.1 Main objective	4
		1.4.2 Secondary objectives	4
		1.4.3 Contributions	4
	1.5	Thesis organization	5
2	$\operatorname{Lit}\epsilon$	rature Review	6
	2.1	Person re-identification	6
	2.2	Object detection	8
3	$\mathrm{Th}\epsilon$	oretical referential	10
	3.1	Computer Vision	10
		3.1.1 Definition	10
		3.1.2 Applications	10
		3.1.3 Research task	11
		3.1.3.1 Recognition	11
		3.1.3.2 Motion Analysis	11
	3.2	Person Re-identification	12

3.3	Closed	Closed-World Person Re-Identification		15
	3.3.1	3.3.1 Feature Representation Learning		16
		3.3.1.1	Global Feature Representation Learning	16
		3.3.1.2	Local Feature Representation Learning	17
		3.3.1.3	Auxiliary Feature Representation Learning	17
		3.3.1.4	Video Feature Representation Learning	18
	3.3.2	Deep M	etric Learning	19
		3.3.2.1	Loss Function Design	19
	3.3.3	Datasets	s and Evaluation Metrics	21
3.4	Open-	World Pe	erson Re-Identification	22
	3.4.1	Heterog	eneous Re-ID	24
		3.4.1.1	Depth-based Re-ID	24
		3.4.1.2	Text-to-Image Re-ID	24
		3.4.1.3	Visible-Infrared Re-ID	24
		3.4.1.4	Cross-Resolution Re-ID	25
	3.4.2	End-to-]	End Re-ID (Person Search)	25
	3.4.3	Semi-su	pervised and Unsupervised Re-ID	26
	3.4.4	Noise-R	obust Re-ID	26
	3.4.5	Open-se	t Re-ID and Beyond	27
3.5	Objec	t Detecto	r: You Only Look One (YOLOv3)	28
	3.5.1	Training	g Details	29
	3.5.2	Test De	tails	29
	3.5.3	Architec	eture	30
3.6	Persor	n Re-ident	tifier: Improved Architecture (Siam-IDL)	31
	3.6.1	Tied Co	nvolution	32
		3.6.1.1	Cross-Input Neighborhood Differences	32

4	Proposed Methodology		34
	4.1	Full Frame Person Re-Identification	34
	4.2	A Human-Machine Hybrid Framework for FF-PRID	35
		4.2.1 Framework	36
		4.2.2 Validation measures	36
5	Exp	perimental Setup: Dataset and FF-PRID pipeline details	39
	5.1	Dataset used for validation	39
	5.2	Overview of the Full Frame Re-ID pipeline	41
	5.3	Object Detection	41
	5.4	Classic Person Re-ID	42
6	Res	ults	43
	6.1	Evaluation of the Object Detection model	43
	6.2	Evaluation of the Person Re-ID model	45
	6.3	Evaluation of the full pipeline for FF-PRID	47
		6.3.1 Influence of the FF-PRID parameters	48
		6.3.2 Qualitative Evaluation	49
		6.3.3 Further considerations	50
	6.4	Some Observations	51
7	Con	nclusions	53
	7.1	Conclusion	53
	7.2	Future works	54
R	efere	nces	55

Chapter 1

Introduction

1.1 Context and Motivation

In recent years, many security cameras were deployed in public places such as streets, malls or airports. Today, most of these video streams are monitored in real-time by security agents, which is expensive and rather inefficient as the amount of videos to analyze is tremendous. In contrast, automated video analysis [29] can process large amounts of videos simultaneously but is more prone to errors for complex tasks such as person reidentification [135]. In addition, even for automated video analysis systems, the final decision often rests with a human security agent, who triggers the appropriate actions. Hence, in practice it seems good to adopt hybrid approaches, where artificial intelligence models can screen the whole network in real time and select only relevant sequences for the monitoring agents.

There are many automated video analysis jobs that are part of an inquiry area called Computer Vision, where there are investigations such as Pedestrian Tracking, Detection of anomalous actions, and Person Re-Identification. This work addresses the Person Re-Identification (Re-ID) task, which has various definitions in different research areas such as metaphysics [80], psychology [87], and logic [12]. Person Re-Identification (Re-ID) problem aims at searching a given person (query) in a network of non-overlapping cameras and raising an alert when this person appears in one of the video streams. It seeks to reproduce and enhance the human ability to recognize people in different scenarios, e.g. wearing different clothes, in a different pose, different illumination conditions, etc.

The current formulation to address Re-ID is based on large databases of images representing human beings in a real-world environment [148, 157, 54, 26, 36]. These images are usually extracted using pedestrian detection models [4] and filtered manually to meet certain standards: each image should contain the entire body of exactly one person, centered and occupying most of the image(examples are shown on Fig. 1.1). From these datasets, a given image is selected as the query and the others constitute the search gallery. Then, the objective is to look for the query person within the gallery [135].



Figure 1.1: Illustration of the Classic Person Re-ID (C-PRID) setting (Source: author).

This approach is illustrated in Fig.1.1 where the output of a C-PRID model is an ordered list with the most similar person on top. Sometimes, individual images are replaced by sequences of successive cropped images and the problem is called video-based Re-ID [79, 52]. From now on, the Re-ID setting considering pre-cropped images of persons as input is referred to as Classic Person Re-Identification (C-PRID). Recent successful methods to address C-PRID are mostly based on deep learning [17, 109, 156, 129, 74].

In practical, tasks of person re-ID system in video surveillance can be divided into three sub-modules[150]; (1)person detection ,(2)person tracking and (3)person retrieval. In general, the first two steps are investigated independently, so C-PRID works are focused on the last module in state of the art. Therefore, our motivation is to discuss the three modules as one task and solve the practical application problems.

1.2 Definition of the Problem

Current research by focusing only on the third module for the recovery of a person limits its performance somehow, causing that in practical scenarios, the application of Re-ID presents failures. However, when thinking about addressing the three modules as a whole, it is a complex problem that as result we call a practical person re-identification system. Here are some problems encountered if we consider C-PRID for resolve our problem:

- C-PRID methods do not consider the full frames of the video stream as input. Therefore, they rely on pre-processing using manual human trimming or detection methods.
- 2. Re-ID is to find if a person is present in a camera network in the real-world practical use case. In this sense, the C-PRID formulation does not satisfy the need to previously extract and crop the people's images to recover the person from being found.
- 3. In this way, the C-PRID can be leveraged to build a new configuration that would be part of a complete task to solve the practical Re-ID problem but is not sufficient on its own.

In this work, our hypothesis is a system that is deployed in a network of non-overpositioned cameras, where the system receives several stream video inputs. Therefore, the system will automatically process the videos generating images of people. The system user will decide which person(query) he wants to search in the camera network. Finally, the system will find the query using re-identification algorithms, alerting the user of the system. Therefore, a new configuration is needed that considers what has been previously reported to create and investigate complete methods in practical terms.

Besides security applications [89, 107, 59, 7], C-PRID is a useful building block for other practical applications such as 3D Multi Object-Tracking [91] or executing visual tasks for drones [105]. This work focuses on the practical security application, which consists of identifying in a network of non-overlapping cameras, a specific person being followed by human surveillance activity. To improve the clarity of this work, from now on this practical application is referred to as Full Frame Person Re-Identification (FF-PRID).

1.3 Justification

This research introduces a new setting of the person Re-ID problem, called FF-PRID, which is better suited to implement and evaluate security applications. However, the inquiry does not claim to solve the FF-PRID setting but rather to demonstrate that the current way of approaching Re-ID is not suited for practical scenarios. By proposing an alternative framework and evaluation method, we hope that this work will motivate the community to consider the FF-PRID setting in order to develop algorithms that are better adapted for real-world scenarios.

1.4 Objectives

1.4.1 Main objective

The main objective is to implement and evaluate real-world security applications based on our Re-ID problem setting, called FF-PRID.

1.4.2 Secondary objectives

- Formalize the natural collaboration between an automated Re-ID system and the monitoring agents like a hybrid framework to address the FF-PRID problem.
- Formulate complementary metrics to assess the quality of any FF-PRID pipeline.
- Reformulate a dataset of C-PRID that satisfying our new setting.
- Elaborate experiments are conducted to demonstrate the importance of considering the FF-PRID problem in its entirety.

1.4.3 Contributions

Our results were published in Pattern Recognition Letters [99], where the main contributions of our work is the proposal of a new pipeline of the person Re-ID problem, called FF-PRID, which is better suited to implement and evaluate real-world security applications. By formalizing the natural collaboration occurring between an automated Re-ID system and the human monitoring agents, a hybrid and robust framework to address the FF-PRID problem is proposed, as well as two complementary metrics to assess the quality of any FF-PRID pipeline. Then, experiments are conducted to demonstrate the importance of considering the FF-PRID problem in its entirety. The most natural pipeline for FF-PRID is implemented within the proposed framework, which consists in using a pedestrian detection model and a C-PRID model sequentially, with both models performing well on standard datasets for their respective tasks. This FF-PRID pipeline is then tested on a modified version of the PRID-2011 dataset [36], using the metrics introduced in our paper[99]. Our experimental results demonstrate that this combination struggles to produce good results for the FF-PRID problem, despite the apparent success of its two independent components. This shows the importance of considering the person Re-ID problem in its Full Frame setting, using adapted metrics. This research does not claim to have solved the FF-PRID setting but rather to demonstrate that the current way of approaching Re-ID is not suited for practical scenarios. By proposing an alternative framework and evaluation method, we hope that this work will motivate the community to consider the FF-PRID setting, in order to develop algorithms that are better adapted for real-world scenarios.

1.5 Thesis organization

This research is organized as follows: in Chapter 2, a literature review about person Re-ID and object detection is presented. The Chapter 3 explain many topics for understanding person Re-ID progressively. In Chapter 4, our reformulation is introduced together with the proposed metric to evaluate a model following this framework. Chapter 5 presents the dataset and practical implementation of the FF-PRID pipeline used for our experimental validation. These experiments consist in demonstrating the use of our new metrics to evaluate a pretrained Re-ID model, coupled with a pretrained object detection model. The results obtained are reported and discussed in Chapter 6. Finally, Chapter 7 shows some conclusions and possible future work.

Chapter 2

Literature Review

In this chapter, we mention the relevant works related to our research. First, it should be emphasized that a new re-identification paradigm is proposed in this work. Therefore, there is not much literature. However, we relate practical approaches and new re-id approaches to ours. On the other hand, we also mention important works related to object detection, which is an important part of our hypothesis.

In this work, the literature search was done using Science Direct and Scopus databases. Then, to obtain knowledge of Re-id we rely on articles and surveys related to Deep Learning Re-ID, Real Systems Re-ID, Practical Re-ID and Video Re-ID. As a result of reading surveys, many articles related to new paradigms were investigated.

2.1 Person re-identification

In the past decade, the task of person Re-ID has been widely studied. The most commonly seen Re-ID pipeline consists of two inputs, a query image representing a well-cropped person to be re-identified, and a search gallery containing various images of different people. The goal of a Re-ID model is then to select the image in the gallery that represents the same person as the query [3]. Some variants of this task have also been proposed, for example, the open-world Re-ID problem extends the previous definition by allowing the case where the gallery does not contain the query image, thus adding a level of complexity [45]. Another alternative definition of person Re-ID is called video Re-ID, or multi-shot Re-ID, and it consists in using a sequence of images of the cropped persons instead of a single image [68, 40, 79, 52]. Some approaches at the intersection of single frame and multi-shot Re-ID have also been proposed, addressing the problem of cross-modality of the input [127]. We note that this paper does not deal with the problem of face recognition,

which is useful in many scenarios but not suitable for re-identification cases where the cameras are far, without very high resolution, or when people are backward.

In the early days of person Re-ID, a typical Re-ID system had two components: capturing a unique feature representation of each pre-cropped person and then comparing two descriptors to infer either a match or a non-match case. Early research about Re-ID relied on extracting low-level features from images, such as color and texture [5, 2]. However, these approaches are only valid over short periods of time. Indeed, on different days, the person might change his clothes, which would generate many errors. To solve this problem, early papers treated it as a retrieval or recognition task. In other words, Re-ID models are computing a similarity score for each image in the gallery and the highest score is selected as the re-identification of the query.

Recently, many large Re-ID datasets have been released. Relevant examples of these datasets are Market-1501 [148], CUHK3 [54], DUKE [157], Viper [26] and PRID [36]. Each one of them has a different configuration, quantity and quality of images. These datasets are composed of cropped images representing entire bodies of people on a real-world environment. In the rare cases when other people appear in the images, they are far behind the main character, in the background. Every image inside such a dataset is labeled with an ID that uniquely represents a given person. In most cases, there are at least 5 to 8 images of each person in the dataset. These dataset also contain a separated sub-set for testing. Some of these datasets, such as PRID, contain sequences of such cropped images and can be used for video Re-ID [79, 127].

More recent approaches leverage Deep Learning to accomplish person Re-ID. The rapid progress in this field, together with the emergence of these large datasets, made it possible to train deep convolutional neural networks to solve the Re-ID task. A complete literature review of Deep Learning methods to solve the classic Re-ID problem is out of the scope of this paper, however, we mention some different techniques that we judged relevant. One example of research using deep neural networks to solve Re-ID was proposed in [17], where six layers are used to extract features, apply cross-input neighborhood differences, patch summary features and finally a softmax function to yield the final estimate of whether the input images are of the same person or not. They only trained and tested over CUHK3 and Viper datasets. In [109], another architecture is proposed, applying *Parameter Free Spatial Attention* layer after feature extraction to focus more on the features extracted over a person's body. This method obtained very promising results on Market-1501, DUKE and CUHK3. In [156], a new approach is proposed, using

appearance and structure space to complement the discriminative module that shares the appearance encoder with the generative module. By switching the appearance or structure codes, the generative module is able to generate high-quality cross-id composed images, which are online fed back to the appearance encoder and used to improve the discriminative module. This approach is tested over Market-1501, Duke and MSMT [118]. Another approach considers features from different layers of a trained CNN to learn partlevel attention on different local regions [129], obtaining promising results on Market-1501, DUKE, and CUHK03. Finally, building on the success of ensemble methods, a voting algorithms was used to choose between the outputs of various trained models in [74].

In the last couple of years, new considerations to deal with practical challenges for implementation of Re-ID have started to appear. In this way, a new metric to measure the cost of finding all the correct matches was introduced in [135]. The open-world setting is starting to gain importance because of its higher real-world relevance [45]. Finally, gaitbased Re-ID is a recent field that aims to identify people by their gait in unconstrained scenarios, typical of surveillance video systems. This is better in long-term scenarios than appearance-based Re-ID [77].

The research presented in this thesis is a continuation of these works as it also attempts to deal with real-world constraints in the practical implementation of Re-ID. In our case, we claim that it is necessary to consider full video frames as input instead of pre-cropped images in order to build solutions that can be evaluated and implemented and in practical scenarios.

2.2 Object detection

Object Detection (OD) has been one of the most studied problems in computer vision in the last decade. Its objective is to find object instances from several predefined categories in real-world images. In this regard, Deep Learning approaches have been developed as a robust strategy for determining feature representations directly from data. For a complete overview of the literature about OD, we refer the reader to the two following surveys [66, 146]. In short, OD methods can be divided into two main families of approaches: region proposal based and regression/classification based.

The region proposal based framework presents a two-step process similar to the attentional mechanism of the human brain. It first gives a coarse scan of the scenario and then focus on the different regions of interest. Conversely, a regression/classification based method is a one-step process, mapping directly from image pixels to bounding box coordinates and class probabilities, adopting a unified framework to reduce significantly the time complexity.

Most object detectors based on regions of interest are based on the following process. First, it builds a region proposal framework to generate a large number of potential bounding boxes. Second, a Convolutional Neural Network (CNN) is used to extract feature characteristics to classify each proposal among the different categories. As a result, the time spent handling the different components becomes the bottleneck in realtime application. Some of the most representative methods of this family of techniques include R-CNN [23], Fast R-CNN [24], Faster R-CNN [85] and Mask R-CNN [31].

Some of the most representative architectures of regression/classification based methods include Single Shot Multi-box Detector(SSD) [67] and You Only Look Once (YOLOv1, YOLO-v2, YOLO-v3) [82, 83, 84]. The YOLO methods use all high-level features to predict confidence scores for each category and generate bounding box, with an execution close to real-time. In short, the basic idea behind YOLO is to divide the image in cells, so that each cell is responsible for predicting the object in its center. Each cell predicts bounding boxes that have their respective confidence scores concerning the detected class.

Chapter 3

Theoretical referential

3.1 Computer Vision

3.1.1 Definition

Computer vision is a research area responsible for analyzing and processing images and videos to understand this information through the computer. In this way, it simulates the human visual system using one or more digital cameras for tries to perceive and understand an image or sequence of images to act differently according to a particular situation. Indeed, it develops different theories and algorithms to understand visual information automatically. Thus, it seeks to apply its ideas to build computer vision systems and facilitate our activities.

3.1.2 Applications

This research field has different applications in real life, like in the industrial area where is used to inspect a production line. Industrial robots control processes, assisting humans in detecting objects for managing employees. Computer vision was able to learn 3D shapes with advances in deep learning has made it possible to develop systems that reconstruct 3D objects from depth maps. Many investigations have been applied to real systems such as:

- Automatic inventory in factories.
- Spice identification system.
- Industrial robots.

- Video surveillance.
- Human-computer interaction.
- Analysis of medical images.
- Autonomous robot navigation.
- Image indexing.

3.1.3 Research task

3.1.3.1 Recognition

Recognition is a common problem for Computer Vision, which recognizes characteristics to determine if the image or video contains any activity.

- Object Recognition: Tries to recognize one or more specified or learned object classes.
- Identification: Recognize a part of the specific image, such as the face or fingerprint.
- Detection: From scanned images, it seeks to detect a specific condition. For example, detection of damaged cells or tissues.
- Content-based image retrieval: Retrieves images with specific content in a large set of images.
- Position estimation: Orientation of a specific object concerning the camera.
- Optical Character Recognition: Identify characters in images that contain text.

3.1.3.2 Motion Analysis

Motion estimation is processed from a sequence of images to analyze different events in the images.

- Egomotion: Determine the rigid 3D movement of the camera from images.
- Tracking: From an initial object, estimate its movement.
- Optical Flow: Determine each point of the image, how the camera moves about the scene.

3.2 Person Re-identification



Figure 3.1: Multi-camera surveillance network illustration of Re-ID (Source: Gala [3]).

Person Re-identification (Re-ID) is defined as a process of developing a correspondence between images of a person taken from non-overlapping fields-of-views(FOVs) [3]. It is used to determine whether situations captured by different cameras belong to the same person. In other words, assign a stable ID to several instances of the person. Fig.3.1 gives an example of a surveillance area monitored by many cameras with non-overlapping FOVs. It shows the top view of a building floor plan and the cameras' relative placement concerning the building. Colored points describe different people, and numbers beside the points are the IDs assigned to the people. The dotted lines with arrows represent how certain people move within the camera network.

Re-ID is used to establish a correspondence between separate tracks to accomplish tracking across many cameras. Thus, single-camera tracking and Re-ID across cameras allow for the reconstruction of a person's trajectory across the larger scene.



Figure 3.2: Images of the same person taken from different cameras to illustrate the appearance changes. The top row images were captured on the same day, bottom row images were captured on different days (Source: Gala [3]).

In general, person Re-ID is challenging to automate for several reasons, which we will discuss later. Still, Re-ID's main challenge comes from the variation in a person's appearance across different views. Fig.3.2 presents images of a person taken by different cameras on the same and different days, highlighting the variations in appearance. The top row demonstrates the changes in the appearance of a person across different cameras. It is also exciting to note that the appearance changes significantly inside the same camera view as well.



Figure 3.3: The pipeline for a practical person Re-ID system, including five main steps: 1) Raw Data Collection, 2) Bounding Box Generation, 3) Training Data Annotation, 4) Model Training and 5) Person Retrieval.

On the other hand, building a person Re-ID system for a practical scenario requires five main steps (as shown in Fig. 3.3, Source: Mang Ye [135]):

- 1. Raw Data Collection: Obtaining raw video data from surveillance cameras is the principal requirement of practical scenarios. These cameras are usually located in different places [112]. This raw data includes a large amount of complex and noisy background issues.
- 2. Bounding Box Generation: Selecting the bounding boxes which contain the person images from the raw video data. Frequently, it is impossible to crop all the person images in real applications manually. The bounding boxes are usually obtained by the person detection approaches [28], or person tracking algorithms [16], [86].
- 3. Training Data Annotation: Annotating the cross camera labels. It is usually essential for discriminative Re-ID model learning due to the large cross-camera variations.
- 4. Model Training: Training a discriminative and robust Re-ID model with the previous annotated person images/videos. This step is the nucleus for developing a Re-ID system, and it is also the most extensively studied paradigm in the literature.
- 5. Person Retrieval: The testing stage conducts the person retrieval. Given a person like our target (query) and a gallery set, we extract the feature representations using the Re-ID model learned in the previous step. A retrieved ranking list is obtained by ordering the calculated query-to-gallery similarity.

Table 3.1: Close-world vs. Open-world Person Re-ID

Closed-world (Section 3.3)	Open-world (Section 3.4)
Single-modality Data	Heterogeneous Data
Bounding Boxes Generation	Raw Images/Videos
Sufficient Annotated Data	Unavailable/Limited Labels
Correct Annotation	Noisy Annotation
Query Exists in Gallery	Open-set

According to the five steps mentioned before, [135] classifies existing Re-ID methods into two main trends: closed-world and open-world settings, as shown in Table 3.1. A comparison is shown in the following five aspects:

 Single-modality vs. Heterogeneous Data: Respect to Step 1 all the persons are represented by images/videos captured by single-modality visible cameras in the closed-world setting [149], [147], [27], [158], [55], [119]. In contrast, in practical open-world applications, we might also require to process heterogeneous data, like infrared images [122], [78], sketches [56], depth images [121], or even text descriptions [53].

- 2. Bounding Box Generation vs. Raw Images/Videos : In Step 2, the closed-world person Re-ID usually makes the training and testing based on the generated bounding boxes manually. In contrast, some practical open-world applications require end-to-end person search from the raw images or videos [151], [126].
- 3. Sufficient Annotated Data vs. Unavailable/Limited Labels: For Step 3, the closedworld person Re-ID usually implies that we have enough annotated training data for supervised Re-ID model training. However, we might not have enough annotated data [65] or even without any label information [144].
- 4. Correct Annotation vs. Noisy Annotation: For Step 4, existing closed-world person Re-ID systems usually assume that all the annotations are correct. However, annotation noise is typically unavoidable due to annotation error or imperfect detection/tracking results [101]).
- 5. Query Exists in Gallery vs. Open-set: In the person retrieval stage (Step 5), most existing closed-world person Re-ID works believe that the query must occur in the gallery set by calculating the CMC [113] and mAP [149]. However, in several scenarios, the query person may not appear in the gallery set [110], [163]. This carries us to the open-set person Re-ID.

We will discuss these settings in the next sections, besides a detailed review of stateof-art data-sets and metrics.

3.3 Closed-World Person Re-Identification

This section provides a summary for closed-world person Re-ID. As presented in Section 2.1, this setting ordinarily has the following premises: 1) person appearances are captured by single-modality visible cameras, either by image or video; 2) The persons are represented by bounding boxes, where most of the bounding box area belongs the same identity; 3) The training has enough annotated training data for supervised discriminative Re-ID model learning; 4) The annotations are generally correct; 5) The query person must appear in the gallery set.



Figure 3.4: Different feature representation (Source: author).

3.3.1 Feature Representation Learning

Here, we discuss the feature learning approaches in closed-world person Re-ID. There are four main categories (as shown in Fig. 3.4).

3.3.1.1 Global Feature Representation Learning

Global feature representation learning extracts a global feature vector for each person image, as shown in Fig. 3.4a. Considering that deep neural networks are originally applied in image classification [93], [32], global feature learning is the initial choice when integrating advanced deep learning techniques into the person Re-ID field. Then, to capture the information in global feature learning, a joint learning framework consisting of a singleimage representation (SIR) and cross-image representation (CIR) is developed in [108], trained with triplet loss using specific sub-networks. The widely-used ID-discriminative Embedding (IDE) model [151] builds the training process as a multi-class classification problem by handling each identity as a distinct class.

3.3.1.2 Local Feature Representation Learning

It learns part/region aggregated features such as shown in Fig 3.4b, making it robust against misalignment [103], [106]. The body parts are automatically generated by 1) human parsing/pose estimation or 2) roughly horizontal division.

For 1) automatic body part detection, the standard solution is to join the full-body representation and local part features [98], [143]. Correctly, the multi-channel aggregation [11], multi-scale context-aware convolutions [47], multi-stage feature decomposition [142], and bilinear-pooling [98] are designed to enhance the local feature learning.

For 2) horizontal-divided region features, multiple part level classifiers are learned in Part-based Convolutional Baseline (PCB) [103], which presently serves as a vital part feature learning baseline in the current state-of-the-art [95], [100], [160].

3.3.1.3 Auxiliary Feature Representation Learning

Auxiliary feature representation learning usually requires additional annotated information (e.g., semantic attributes [96] like Fig 3.4c shows) or generated/augmented training samples to augment the feature representation [41], [158].

Semantic Attributes. Collective identity and attribute learning baseline are addressed in [61]. [96] Propose a deep attribute learning framework by incorporating the predicted semantic attribute information, enhancing the feature representation's generalization and robustness in a semi-supervised learning mode.

Viewpoint Information. The viewpoint information is also leveraged to improve the feature representation learning [8], [63]. Multi-Level Factorization Net (MLFN) [8] also proposes to learn the identity-discriminative and view-invariant feature representations at multiple semantic levels. [63] extract a combination of view-generic and view-specific learning.

Domain Information. A Domain Guided Dropout (DGD) algorithm [125] is designed to adaptively mine the domain-sharable and domain-specific neurons for multidomain deep feature representation learning. Using each camera as a distinct domain, [60] proposes a multi-camera consistent matching constraint to take a globally optimal representation in a deep learning framework.

GAN Generation. Here we discuss the use of GAN(Generative Adversarial Network) generated images as the auxiliary information. [158] Start the beginning attempt

to apply the GAN technique for person Re-ID. It improves the supervised feature representation learning with the created person images. Pose constraints are incorporated in [65] to improve the generated person images' quality, developing the person images with new pose variants. A pose-normalized image generation approach is designed in [81], enhancing the robustness against pose variations.

Data Augmentation. For Re-ID, custom operations are random resize, cropping, and horizontal flip [72]. Besides, adversarially occluded samples [41] are generated to augment the variation of training data. A similar random erasing strategy is proposed in [159], adding random noise to the input images. In [1] generate the virtual humans are rendered under different illumination conditions. These methods enhance the supervision with the augmented samples, increasing the generalizability of the testing set.

3.3.1.4 Video Feature Representation Learning



Figure 3.5: Same person in different cameras represented using a multi-shot version that contains multiple images(called tracklet in other works [51]) per person. (Source: https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/prid11)

Video-based or Multi-shot Re-ID(3.4d) is becoming a popular topic [37], where each person is represented by a video sequence (as shown in Fig. 3.5). Due to the rich appearance and temporal information, it has gained increasing interest in the Re-ID community. This also produces new challenges in video feature representation learning. The main challenge is to capture the temporal information accurately. A Recurrent neural network architecture was designed for video-based person Re-ID in [75], which simultaneously optimizes the final recurrent layer for temporal information propagation and the temporal pooling layer. Semantic attributes are also utilized in [145] for video Re-ID with feature disentangling and frame re-weighting. Combined aggregating the frame-level feature and spatio-temporal appearance information is crucial for video representation learning [162], [128], [97].

A different major challenge is the unavoidable outlier tracking frames within the videos. A diversity regularization [51] is employed to work multiple discriminative body parts in each video sequence. An affine hull is adopted to handle the outlier frames within the video sequence [132]. An exciting work [38] utilizes multiple video frames to auto-complete occluded regions. These works demonstrate that handling noisy frames can significantly improve video representation learning.

It is also challenging to handle the diverging lengths of video sequences; in [10], authors divide the long video sequences into multiple short pieces, aggregating the topranked pieces to learn a compact embedding. A clip-level learning strategy [22] exploits both spatial and temporal dimensional attention cues to produce a robust clip-level representation. Both the short and long-term relations [48] are integrated into a self-attention scheme.

3.3.2 Deep Metric Learning

3.3.2.1 Loss Function Design

Here, we only focus on the loss functions designed for deep learning [138]. There are three widely studied loss functions with their variants in the literature for person Re-ID, including identity loss, verification loss, and triplet loss. An illustration of three-loss functions is shown in Fig. 3.6.

Identity Loss. It discusses the training process of person Re-ID as an image classification task [151], *i.e.*, each identity is a distinct class. In the testing phase, the pooling layer's output or embedding layer is adopted as the feature extractor. Given an input image x_i with label y_i , the predicted probability of x_i being recognized as class y_i is encoded with a softmax function, represented by $p(y_i|x_i)$. The identity loss is then computed by the cross-entropy



Figure 3.6: Three kinds of widely used loss functions in the literature. (Source: Mang Ye [135]).

$$\mathcal{L}_{id} = -\frac{1}{n} \sum_{i=1}^{n} \log(p(y_i|x_i)),$$
(3.1)

where n represents the number of training samples within each batch. The identity loss has been widely used in existing methods [41], [158], [102], [152], [71]. Frequently, it is easy to train and automatically mine the hard samples during the training process, as demonstrated in [137]. Several works have also investigated the softmax variants [120], such as the sphere loss in [21] and AM softmax in [71].

Verification Loss. It optimizes the pairwise correlation, either with a contrastive loss [106], [120] or binary verification loss [55], [123]. The contrastive loss improves the relative pairwise distance comparison, expressed by

$$\mathcal{L}_{con} = (1 - \delta_{ij}) \{ max(0, \rho - d_{ij}) \}^2 + \delta_{ij} d_{ij}^2, \qquad (3.2)$$

where d_{ij} represents the Euclidean distance between the embedding features of two input samples x_i and x_j . δ_{ij} is a binary label indicator ($\delta_{ij} = 1$ when x_i and x_j belong
to the equivalent identity, and $\delta_{ij} = 0$, otherwise). ρ is a margin parameter. There are several variants, e.g., the pairwise comparison with ranking SVM in [108]. Binary verification [55], [123] discriminates the positive and negative of an input image pair. Generally, a differential feature f_{ij} is obtained by $f_{ij} = (f_j - f_j)^2$ [123], where f_i and f_j are the embedding features of two samples x_i and x_j . The verification network classifies the differential feature into positive or negative. We use $p(\delta_{ij}|f_{ij})$ to represent the probability of an input pair (x_i and x_j) being recognized as $\delta_{ij}(0 \text{ or } 1)$. The verification loss with cross-entropy is

$$\mathcal{L}_{veri}(i,j) = -\delta_{ij}\log(p(\delta_{ij}|f_{ij})) - (1 - \delta_{ij})\log(1 - p(\delta_{ij}|f_{ij})).$$
(3.3)

The verification is often combined with the identity loss to enhance the performance [94][106].

Triplet loss. It uses the Re-ID model training process as a retrieval ranking task. The basic idea is that the positive pair's distance should be smaller than the negative pair by a pre-defined margin [35] (as shown in Fig. 3.6c). Typically, a triplet includes one anchor sample x_i , one positive sample x_j with the same identity, and one negative sample x_k from a different identity. The triplet loss with a margin parameter is represented by

$$\mathcal{L}_{tri}(i, j, k) = max(\rho + d_{ij} - d_{ik}, 0), \qquad (3.4)$$

where d() estimates the Euclidean distance between two samples. The large proportion of easy triplets will dominate the training process if we directly optimize the above loss function, resulting in limited discriminability. To mitigate this issue, various informative triplet mining methods have been designed [116], [94], [35], [98]. The basic idea is to select the informative triplets [35], [92]. Specifically, moderate positive mining with a weight restriction is introduced in [92], which directly optimizes the feature difference. [35] demonstrate that the online hardest positive and negative mining within each training batch is beneficial for discriminative Re-ID model learning. Some methods also studied the point-to-set similarity strategy for informative triplet mining [161], [139]. This improves robustness against the outlier samples with a soft hard-mining scheme.

3.3.3 Datasets and Evaluation Metrics

In this work, we review the most common datasets for the closed-world setting, including 11 image datasets (VIPeR [27], iLIDS [153], GRID [70], PRID2011 [37], CUHK0103 [55], Market-1501 [149], DukeMTMC [158], Airport [25] and MSMT17 [119]) and 7 video datasets (PRID-2011 [37], iLIDS-VID [111], MARS [147], Duke-Video [124], Duke-Tracklet [49], LPW [50] and LS-VID [48]). The statistics of these datasets are shown in Table 2. In this work, we only focus on the general large-scale datasets for deep learning approaches. We can make several observations in terms of the dataset collection over recent years:

- 1. The dataset scale (both #image and #ID) has increased quickly. Commonly, the deep learning approach can benefit from more training samples. This also increases the annotation difficulty needed in closed-world person Re-ID.
- 2. The camera number is also significantly increased over the years to approximate the large-scale camera network in practical scenarios.
- 3. The bounding boxes generation is usually done automatically detected/tracked, rather than manually cropped. This mimics the real-world scenario with track-ing/detection errors.

Evaluation Metrics. To evaluate a Re-ID system, Cumulative Matching Characteristics (CMC) [113] and mean Average Precision (mAP) [149] are two widely used measurements. CMC- k (a.k.a, Rank- k matching accuracy) [113] represents the probability that a correct match appears in the top- k ranked retrieved results. CMC is accurate when only one ground truth exists for each query since it only considers the first match in the evaluation process. On the other hand, mean Average Precision (mAP) [149] measures the average retrieval performance with multiple ground truths. It is originally widely used in image retrieval.

3.4 Open-World Person Re-Identification

This section reviews open-world person Re-ID as discussed in Section 2.1, including heterogeneous Re-ID by matching person images across heterogeneous modalities (Sub-section 3.4.1), end-to-end Re-ID from the raw images/videos (Sub-section 3.4.2), semi/unsupervised learning with limited/unavailable annotated labels (Sub-section 3.4.3), robust Re-ID model learning with noisy annotations (Sub-section 3.4.4) and open-set person Re-ID when the correct match does not occur in the gallery (Sub-section 3.4.5).

	Image Datasets							
Dataset	Time	$\#\mathrm{ID}$	#image	$\#\mathrm{cam}$	Label	Res.	Eval.	
VIPeR	2007	632	1,264	2	hand	fixed	CMC	
iLIDS	2009	119	476	2	hand	vary	CMC	
GRID	2009	250	1,275	8	hand	vary	CMC	
PRID2011	2011	200	$1,\!134$	2	hand	fixed	CMC	
CUHK01	2012	971	3,884	2	hand	fixed	CMC	
CUHK02	2013	1,816	7,264	10	hand	fixed	CMC	
CUHK03	2014	$1,\!467$	13,164	2	both	vary	CMC	
Market-1501	2015	1,501	32,668	6	both	fixed	C&M	
DukeMTMC	2017	$1,\!404$	36,411	8	both	fixed	C&M	
Airport	2018	$9,\!651$	39,902	6	auto	fixed	C&M	
MSMT17	2018	4,101	$126,\!441$	15	auto	vary	C&M	
	Video datasets							
Dataset	Time	$\#\mathrm{ID}$	$\# \mathrm{image}$	$\# \mathrm{cam}$	Label	Res.	Eval.	
PRID2011	2011	200	400(40k)	2	hand	fixed	CMC	
iLIDS-VID	2014	300	600(44k)	2	hand	vary	CMC	
MARS	2016	1261	20,715(1M)	6	auto	fixed	C&M	
Duke-Video	2018	1,812	4,832(-)	8	auto	fixed	C&M	
Duke-tracklet	2018	1,788	$12,\!647(-)$	8	auto	fixed	C&M	
LPW	2018	2,731	7,694(590K)	4	auto	fixed	C&M	
LS-VID	2019	3,772	14,943(3M)	15	auto	fixed	C&M	

Table 3.2: Details about commonly used datasets for closed-world person Re-ID. "both" means that it contains both hand-cropped and detected bounding boxes. "C&M" means both CMC and mAP are evaluated.

3.4.1 Heterogeneous Re-ID

3.4.1.1 Depth-based Re-ID

Depth images capture the body shape and skeleton information. This gives Re-ID the possibility to work under illumination/clothes changing environments, which is also essential for personalized human interaction applications.

A recurrent attention-based model is proposed in [30] to address depth-based person Re-ID. In a reinforcement learning framework, they joint the convolutional and recurrent neural networks to identify small, discriminative local regions of the human body. [42] Leverage the large RGB datasets to design a split-rate RGB-to-Depth transfer method, connecting the gap between the depth images and the RGB images.

3.4.1.2 Text-to-Image Re-ID

Text-to-image Re-ID engages matching between a text description and RGB images [53]. It is imperative when a query person's visual image cannot be obtained, and only a text description can be alternatively provided. A gated neural attention model [53] with a recurrent neural network acquires the shared features between the text description and the person images. This allows the end-to-end training for text to image pedestrian retrieval. Cheng et al. [9] propose a global discriminative image-language association learning method, capturing the identity discriminative information and local reconstructive image-language association under a reconstruction process.

3.4.1.3 Visible-Infrared Re-ID

Visible-Infrared Re-ID examines the cross-modality matching between the daytime visible and night-time infrared images. It is vital in low-lighting conditions, where the images can only be captured by infrared cameras [122], [78], [?].

Wu et al. [122] start the first attempt to address this problem by proposing a deep zero-padding framework [122] to learn shareable modality features adaptively. A twostream network is introduced in [131], [136] to model the modality-sharable and -specific information, addressing the intra-modality and cross-modality variations concurrently.

3.4.1.4 Cross-Resolution Re-ID

Cross-Resolution Re-ID handles the matching between low-resolution and high-resolution images, addressing the large resolution changes [58], [116]. A cascaded SR-GAN [117] produces the high-resolution person images in a cascaded manner, incorporating the identity data.

3.4.2 End-to-End Re-ID (Person Search)

End-to-end Re-ID eases the reliance on an additional step for bounding boxes generation. It includes the person's Re-ID from raw images or videos and multi-camera tracking. Re-ID in Raw Images/Videos This job demands that the model jointly perform person detection and re-identification in a single structure (as shown in Fig. 3.7) [151], [126]. It is stimulating due to the different focuses of two significant components.



Figure 3.7: Person search(End-to-End Re-ID) is about finding a query person (yellow rectangle) within a gallery image (the target green rectangle).(Source: author)

Zheng et al. [151] present a two-stage framework and systematically evaluate person detection's benefits and limitations for the later stage person Re-ID. Xiao et al. [126] design an end-to-end person search system using a single convolutional neural network for joint person detection and re-identification. A Neural Person Search Machine (NPSM) [64] is developed to recursively refine the searching area and locate the target person by fully exploiting the query's contextual information and the detected candidate region.

Multi-camera Tracking. End-to-end person Re-ID is also strictly related to multiperson, multi-camera tracking [86]. A graph-based formulation to link person hypotheses is proposed for multi-person tracking [104]. The holistic features of the full human body and body pose layout are combined as the representation for each person. Ristani et al. [86] learn the correlation between the multi-target multi-camera tracking and person Re-ID by hard-identity mining and adaptive weighted triplet learning. Lately, a localityaware appearance metric (LAAM) [39] with both intra and inter-camera relation modeling is proposed.

3.4.3 Semi-supervised and Unsupervised Re-ID

Unsupervised Re-ID mainly studies invariant components, *i.e.* dictionary [18], metric [69], or saliency [144], which guides to limited discriminability or scalability. For deeply unsupervised techniques, cross-camera label estimation is one of the popular approaches [134], [20]. Dynamic graph matching (DGM) [133] expresses the label estimation as a bipartite graph matching problem. To further enhance the performance, global camera network constraints [115] are exploited for consistent matching.

For end-to-end unsupervised Re-ID, an iterative clustering and Re-ID model learning is presented in [20]. Likewise, the relations among samples are utilized in a hierarchical clustering framework [141]. Soft multi-label learning [209] mines the soft label information from a reference set for unsupervised learning. A Tracklet Association Unsupervised Deep Learning (TAUDL) framework [49] combined conducts the withincamera tracklet association and models the cross-camera tracklet correlation.

Semi-supervised Re-ID. With limited label information, a one-shot metric learning method is offered in [213], incorporating a deep texture representation and a color metric. A stepwise one-shot learning method (EUG) is introduced in [124] for video-based Re-ID, regularly selecting candidates from unlabeled tracklets to enrich the labeled tracklet set. A multiple instance attention learning framework [114] uses the video-level labels for representation learning, alleviating the dependence on full annotation.

3.4.4 Noise-Robust Re-ID

Re-ID usually suffers from unavoidable noise due to data collection and annotation difficulty. We review noise-robust Re-ID from three aspects:

Partial Re-ID. This addresses the Re-ID problem with substantial occlusions, *i.e.*, only part of the human body is visible [155]. A fully convolutional network [33] is chosen to generate fix-sized spatial feature maps for the incomplete person images. Deep Spa-

tial feature Reconstruction (DSR) is further incorporated to avoid precise alignment by exploiting the reconstructing failure.

Re-ID with Sample Noise. This refers to the person images or the video sequence containing outlying regions/frames, either produced by poor detection/inaccurate tracking results. To handle the outlying areas or background clutter within the person image, pose estimation cues [142], [88], or attention cues [94], [144], [44] are exploited. The basic idea is to suppress the contribution of the noisy regions in the final holistic representation. For video sequences, set-level feature learning [132] or frame-level re-weighting [10] are the ordinarily used approaches to reduce noisy frames' impact.

Re-ID with Label Noise. Label noise is usually inevitable due to annotation errors. Zheng et al. adopt a label smoothing technique to avoid label overfitting issues [158]. A Distribution Net (DNet) that models the feature uncertainty is proposed in [140] for robust Re-ID model learning against label noise, reducing the impact of samples with high feature uncertainty.

3.4.5 Open-set Re-ID and Beyond

Open-set Re-ID is usually expressed as a person verification problem, *i.e.*, discriminating whether two-person images belong to the same identity [110], [163]. The verification usually needs a learned condition τ , *i.e.*, $sim(query, gallery) > \tau$. Early researches design handcrafted systems [154], [110], [163]. For deep learning methods, an Adversarial Person-Net (APN) is proposed in [57], which jointly learns a GAN module and the Re-ID feature extractor. This GAN's basic idea is to generate realistic target-like images (imposters) and enforce the feature extractor is robust to the generated image attack.

Group Re-ID. It aims at associating the persons in groups rather than individuals [153]. Early research focuses on group representation extraction with sparse dictionary learning [62] or covariance descriptor aggregation [6]. The group similarity is also practiced in the end-to-end person search [130] and the individual re-identification [76], [90] to increase the accuracy. Nevertheless, group Re-ID is still challenging since the group variation is longer complicated than the individuals.

Dynamic Multi-Camera Network. The dynamic updated multi-camera network is another challenging issue [15], [73], [14], [13], which needs model adaptation for new cameras or probes. A human-in-the-loop incremental learning method is presented in [73] to update the Re-ID model, adjusting the representation for different probe galleries. Early research also applies active learning [14] for continuous Re-ID in a multi-camera network. A constant adaptation method based on sparse non-redundant representative selection is introduced in [15]. Furthermore, how to apply the deep learning technique for the dynamic multi-camera network is still less investigated.

3.5 Object Detector: You Only Look One (YOLOv3)



Figure 3.8: YOLO flow for object detection. (Source: YOLO [84]).

Object Detection (OD) has been one of the most studied problems in computer vision in the last decade. Its objective is to find object instances from several predefined categories in real-world images. In this regard, Deep Learning approaches have been developed as a robust strategy for determining feature representations directly from data.

In [84] propose a new framework called YOLO, which makes use of all highest level feature maps to predict multiple categories and bounding boxes. YOLO's basic idea is to divide the image from entering $S \times S$ cells, so that each cell is responsible for predicting the object centered in each cell(as shown in Fig. 3.8). Each cell predicts *B* bounding boxes that have their respective confidence scores concerning the detected class [66, 146]

YOLO formally defines the confidence score as, $Pr(Object) * IOU_{pred}^{truth}$, which indicates the probability that objects exist, ($Pr(Object) \ge 0$), and displays the confidence score for its prediction, (IOU_{pred}^{truth}) . At the same time, without depending on the number of bounding boxes, C conditional class probabilities, $(\Pr(\text{Class}_i|\text{Object}))$, should also be predicted for each cell [84].

3.5.1 Training Details

During training, YOLO optimizes the loss function,

$$\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathscr{W}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\ + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathscr{W}_{ij}^{obj} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \\ + \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathscr{W}_{ij}^{obj} \left(C_i - \hat{C}_i \right)^2 \\ + \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathscr{W}_{ij}^{noobj} \left(C_i - \hat{C}_i \right)^2 \\ + \sum_{i=0}^{S^2} \mathscr{W}_i^{obj} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2 \end{cases}$$
(3.5)

, where each cell is represented with i, and the coordinates of the midpoint of each cell is represented by (xi, yi), (wi, hi) are the width and height normalized in relation to the image size, respectively, Ci represents the confidence score, \mathbb{K}_i^{obj} represents the existence of objects and \mathbb{K}_{ij}^{obj} indicates the prediction was driven, by the j - th bounding box production. Also, it may be noted that, when an object is present in the cell, the loss function will penalize the classification of errors. Similarly, when the predictor is responsible for the bounding box of the ground-truth, the bounding box coordinate errors will be penalized [146, 84].

3.5.2 Test Details

For the tests, there is a specific confidence score for each bounding box that is multiplied by the prediction confidence score with the conditional class probability, as shown in the following equation :

$$Pr(Object) * IOU_{pred}^{truth} * Pr(Class_i|Object)$$

$$= Pr(Class_i) * IOU_{pred}^{truth}$$
(3.6)

,where the real probability of class objects in the bounding box and the correspondence between the predicted bounding box and the object bounding box are both taken into consideration [146, 84].

3.5.3 Architecture



Figure 3.9: Architecture of YOLO V3. (Source: https://plos.figshare.com/articles/figure/YOLOv3_architecture_/8322632/1).

This network(as shown in Fig. 3.9) uses 53 convolutional layers with 3x3 kernels in the beginning and 1x1 in the end. The model used was trained on the VOC dataset [19], containing 80 classes. Darknet-53 operates at a level close to state-of-the-art object detectors, but is faster because it uses less floating-point operations.

YOLO divides the input image in a 13 by 13 cell grid. Each of these cells is responsible for predicting 5 Bounding Boxes as well as their associated confidence scores. Each Bounding Box describes a rectangle that encloses an object. For each Bounding Box, YOLO generates a confidence score that tells us how safe this bounding box contains an object [84]. 3.6 Person Re-identifier: Improved Architecture (Siam-IDL)



Figure 3.10: Architecture of Siam-IDL Re-ID. (Source: Ahmed [17]).

The method used to perform classic person Re-ID in this work is called An Improved Deep Learning Architecture for Person Re-Identification (SiamIDL) [17]. Because, we think that is an exciting work that has a considerable influence in literature, where they obtain better results than the other works presented until that day. They propose a deep neural network architecture(as shown Fig. 3.10) as follows: two layers of tied convolution with max pooling, cross-input neighborhood differences, patch summary features, across-patch features, higher-order relationships, and finally a softmax function to yield the final estimate of whether the input images are of the same person or not. They achieved 54.74% vs. 20.65% over the CUHK-03 dataset more than double from the previous methods. This work was compared against KISSME, eSDC, SDALF, ITML, logistic distance metric learning (LDM), largest margin nearest neighbor (LMNN), metric learning to rank (RANK), and directly using Euclidean distance to compare features.

3.6.1 Tied Convolution

In the deep learning literature, convolutional features have proven to provide representations that are useful for various classification tasks. The first two layers of the network are convolution layers, which compute higher-order features on each input image separately. For the features to be comparable across the two images in later layers, our first two layers perform tied convolution, in which weights across the two views, to ensure that both views use the same filters to compute features. As shown in Figure 2, in the first convolution layer, we pass input pairs of RGB images of size $60 \times 160 \times 3$ through 20 learned filters of size $5 \times 5 \times 3$. The resulting feature maps are passed through a max-pooling kernel that halves the width and height of features. These features are passed through another tied convolution layer that uses 25 learned filters of size $5 \times 5 \times 20$, followed by a max-pooling layer that again decreases the width and height of the feature map by a factor of 2. At the end of these two feature computation layers, each input image is represented by 25 feature maps of size 12×37 .

3.6.1.1 Cross-Input Neighborhood Differences

The two tied convolution layers give a set of 25 feature maps for each input image, from which it can learn relationships between the two views. Let f_i and g_i , respectively, represent the *i*th feature map $(1 \le i \le 25)$ from the first and second views. A crossinput neighborhood differences layer computes differences in feature values across the two views around a neighborhood of each feature location, producing a set of 25 neighborhood difference maps K_i .

Since $f_i, g_i \in \mathbb{R}^{12 \times 37}$, $K_i \in \mathbb{R}^{12 \times 37}$, where 5×5 is the size of the square neighborhood. Each K_i is a 12×37 grid of 5×5 blocks, in which the block indexed by (x, y) is denoted $K_i(x, y) \in \mathbb{R}^{12 \times 37}$, where x, y are integers $(1 \le x \le 12 \text{ and } 1 \le y \le 37)$. More precisely.

$$K_i(x, y) = f_i(x, y) \mathbb{1}(5, 5) - \mathcal{N}[g_i(x, y)]$$

where

1 (5,5) $\in \mathbb{R}^{5 \times 5}$ is a 5 × 5 matrix of 1s,

N $[g_i(x, y)] \in \mathbb{R}^{5 \times 5}$ is the 5 × 5 neighborhood of g_i centered at (x, y).

(3.7)

In other words, the 5 × 5 matrix $K_i(x, y)$ is the difference of two 5 × 5 matrices, in the first of which every component is a copy of the scalar $f_i(x, y)$, and the second of which is the 5 × 5 neighborhood of g_i centered at (x, y). The impulse behind taking differences in a neighborhood is to add robustness to positional differences in corresponding features of the two input images. Following the operation in (1) is asymmetric, it also considers the neighborhood difference map K'_i , which is defined just like K_i in (1) except that the roles of f_i and g_i are reversed. This yields 50 neighborhood difference maps, $\{K_i\}_{i=1}^{25}$ and $\{K'_i\}_{i=1}^{25}$, each of which has size $12 \times 37 \times 5 \times 5$. They pass these neighborhood difference maps through a rectified linear unit (ReLu).

Chapter 4

Proposed Methodology

In this section, we introduce a new pipeline of the person Re-ID problem, called FF-PRID [99], which is better suited to implement and evaluate real-world security applications. Also, by formalizing the natural collaboration between an automated Re-ID system and the human monitoring agents, a hybrid and robust framework to address the FF-PRID problem is proposed, and two complementary metrics to assess the quality of any FF-PRID pipeline.

4.1 Full Frame Person Re-Identification

The C-PRID formulation is a useful building block to implement security application of Re-ID, i.e. to identify a person sought by the authorities in a network of security cameras. However it is not sufficient, as for such a practical implementation, the entire image of the video frames must be used as input, instead of carefully selected pre-cropped images of persons. From now on, this application-oriented Re-ID setting is referred to as Full Frame Person Re-ID (FF-PRID).

In short, in the FF-PRID setting, a successful model must analyze full frames to determine if the query is present in the stream, and if it is, when and where it appeared. The FF-PRID setting is illustrated in Fig. 4.1.

One can argue that the C-PRID problem can be easily derived from the FF-PRID setting by applying a pedestrian detection (PD) model [4] on the raw video stream, which is often done in practice. Indeed, some object detection models have demonstrated strong results for detecting human beings over the last few years [85, 84]. However, we argue that not considering the problem as a whole presents several issues:



Figure 4.1: Full Frame Person Re-ID (FF-PRID) setting. (Source: author)

- 1. The bounding boxes extracted by PD models may differ from the images in the reference datasets used for C-PRID training and evaluation, which have been filtered manually to only select clean images. This domain shift between the galleries used for training and the data encountered at inference time can decrease the quality of the model at run time, and thus induces a strong bias for model evaluation.
- 2. Even if both a good pedestrian detection model and a good Re-ID model are used, their small prediction errors might add up to produce poor overall results for the final application.
- 3. Not considering FF-PRID as an independent problem might dissuade the community from trying different approaches for the full application. Indeed, the vast availability of C-PRID datasets might take researchers away from trying other promising approaches such as end-to-end methods or video based methods, which have been shown to work for other computer vision problems [46, 34].
- 4. When developing a practical application, it is crucial to evaluate the quality of the entire pipeline before deploying it in production. To the best of our knowledge, frameworks and metrics to evaluate FF-PRID are missing in the literature.

4.2 A Human-Machine Hybrid Framework for FF-PRID

The classic formulation of person Re-ID consists in comparing a query image with all the images of a search gallery to output a set of similarity scores representing the Re-ID predictions. Conversely, this work considers the Full Frame Re-ID setting, which is better suited to implement and evaluate practical security applications. In this field, we introduce a hybrid framework, using human-machine collaboration to address the FF-PRID problem and we propose two new evaluation metrics to assess the quality of a FF-PRID model on a given dataset.

4.2.1 Framework

In the FF-PRID setting, the inputs to the system are a query image and a raw video from a security camera. Studying this setting is important as the conversion from a camera feed to a C-PRID search gallery is not straightforward and needs to be evaluated to design reliable applications. Ideally, from a query image and a raw video feed, a FF-PRID model should find whether or not the query appears in each frame. This way, the system can raise an alert as soon as the searched person is encountered in any camera. But in practice, the FF-PRID task is complex and highly prone to errors. Because of the criticality of the task in many scenarios, the outputs of the model must be cross-checked by a human operator before triggering any action involving security agents.

Thus, we propose an alternative hybrid framework, which requires validation by a human operator after automatic predictions are made by an artificial intelligence model, to address this problem and evaluate it. The proposed pipeline goes as follows: First, the live video stream is cut into short video segments of τ frames. Then, each of these segments are processed by a pedestrian detection model to extract bounding boxes of all the persons present in the video and create a traditional search gallery. The query and the gallery are then processed by a classic Re-ID model and, if the highest similarity score in the gallery is higher than a given threshold β , the η members of the gallery with highest similarity scores are shown to the monitoring agent, who decides if the predictions are correct triggers actions when necessary. The proposed pipeline is illustrated in Fig. 4.2a. The threshold for raising an alert β , the number of images shown to the agent η and the length of the video segments τ are user defined parameters that influence the final results. We note that the ideal scenario described above can be obtained with this framework if $\tau = 1$, $\eta = 1$, the FF-PRID works perfectly and β is tuned appropriately.

4.2.2 Validation measures

In the case of a perfect FF-PRID model, the operator validation is required in all the cases where the query is present in the τ frames of video sequence and not in any other case. Hence, there are two ways for a model to fail: by missing the query when it is present



(b) Proposed FF-PRID model.

Figure 4.2: Hybrid Human-Machine Framework and proposed Pipeline for Full Frame Person Re-Identification.(Source: author)

in the video segment or by calling the operator when the query is not present. Thus, to evaluate the quality of a model, we define two important indicators that we call *Finding Rate* (FR) and *True Validation Rate* (TVR). They respectively represent the number of sequences in which the query was found when it appeared and the number of times that the query was present when the operator was solicited.

To define these two validation measures formally, some other variables must be introduced first. These variables are influenced by the variables to evaluate the classification task as True Positive (TP), False Negative(FN), True Negative(TN) and False Positive(FP). For a given {query, video} pair, we define:

- A True Call (TC), when the query is present in the video, the highest similarity score is greater than the threshold β and the query is in the top η best candidates. It corresponds to a successful case of re-identification by the system.
- A True Missed Call (TMC), when the query is present in the video, the highest similarity score is greater than β and the query is not in the top η best candidates. It

is the case where the query is present, the system is asking for confirmation but does not provide the correct images to the operator and the query is missed anyways.

- A False Silence (FS), when the query is present in the video, but the highest similarity score is smaller than β . It is the case where the query is missed but the operator is not disturbed.
- A False Call (FC), when the query is not in the video but the highest similarity score is greater than β . It corresponds to the case where the operator is disturbed for nothing.
- A True Silence (TS), when the query is not in the video and the highest similarity score is smaller than β . It is the case where the query is not present and nothing happens.

Then, the FR and TVR can be defined as follows:

$$FR = \frac{TC}{TC + TMC + FS},\tag{4.1}$$

$$TVR = \frac{TC}{TC + TMC + FC}.$$
(4.2)

FR and TVR are comprised between 0 and 1. Hence, FR = 1 means that whenever the query was present in the video, it was successfully identified by the system (model + operator). Likewise, TVR = 1 means that the operator was never called for nothing, i.e. all the time the model asked for verification, the query was actually present in the proposed cropped images. In contrast, FR < 1 means that in some sequences the query was present but it was missed, and TVR < 1 means that in some situations the model asked for operator validation when the query was not present in the suggestions.

Chapter 5

Experimental Setup: Dataset and FF-PRID pipeline details

5.1 Dataset used for validation

To test the proposed framework and metrics, we use a modified version of the PRID-2011 dataset [36], considering raw full frame videos as input instead of the pre-cropped images of the original dataset.

The original PRID-2011 dataset is composed of images extracted from multiple person trajectories recorded from two different static surveillance cameras, named A and B (as shown in Fig. 5.1). Images from these cameras contain a view point change and a stark difference in illumination and background. Since images are extracted from trajectories, several successive poses per person are available in each camera view, with some people appearing in both views. After filtering out manually some heavily occluded persons, corrupted images induced by tracking and annotation errors, the official PRID-2011 dataset contains 385 persons in camera view A and 749 in camera view B. The persons with the first 200 labels appear in both views.

PRID-2011 was created to test classic person Re-ID approaches, as well as video-based Re-ID [79]. To conduct our experiments, we obtained the raw videos and annotations that were used to create the PRID-2011 dataset¹. From now on, the two raw full frame videos will be called view A (1:01:53 hours) and view B (1:06:39 hours). Both views were cut into sub-videos of 2 minutes, to serve as input to the FF-PRID framework (Fig. 4.2a). This way, view A contains 30 videos and view B, 33. For each video, only a few persons

 $^{^1\}mathrm{We}$ kindly thank the authors of the original PRID-2011 paper for their responsiveness and cooperation.



Figure 5.1: Example of a building with two different static surveillance cameras called A and B. (Source: https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/prid11/).

appear simultaneously, e.g., the first video of view B only contains the persons labeled 16 and 204 (the same labels as in the original dataset are used).

For each 2 minute video sample, a ground truth file is generated². For each person that appears in the video, it contains: the person identifier Id; the first frame where it appears fr; the number of times that it appears in the following frames s; the coordinates of the bounding box where it first appears (ulx, uly, brx, bry); the timestamp where it appears sec (calculated from fr); and the number of the sub-sequences where it appears sub. The number of sub-sequences is computed from τ , a user defined parameter introduced in Section 4.2.1.

²Our scripts for processing the raw videos and generating the ground truth files, as well as the implementation of our baseline pipeline, are openly available at: https://github.com/fsumari/FF-PRID-2020.

5.2 Overview of the Full Frame Re-ID pipeline

Fig. 4.2 illustrates the proposed FF-PRID approach. In Fig. 4.2a, we show the application level Re-ID scheme. The original video is split into shorter sequences and passed to a FF-PRID model. When the model returns a high confidence score that the query is present in the sequence, an alert is raised and a group of persons' images are presented to a monitoring agent for human validation. In Fig. 4.2b, the FF-PRID model is shown in details. The video is fed to an object detection model in order to detect pedestrians and generate clippings for the search gallery. After this step, the image of the query person is searched in the gallery by means of a classic Re-ID model, which outputs a list of images similar to the query, ordered from most to least similar. Both the pedestrian detection model and the classic Re-ID model were implemented using TensorFlow 1.14.0 and were executed on a NVIDIA P5000 GPU. We present the implementation of these models in the following subsections.

5.3 Object Detection

For this work, we use the You Only Look Once (YOLO-v3) [84] approach for pedestrian detection. In short, YOLO methods belong to the family of regression/classification based approaches, mapping directly from image pixels to bounding box coordinates and class probabilities to reduce significantly the time complexity. A detailed explanation of YOLO is out of the scope of this sub-section, and for a complete overview of the recent literature about Object Detection (OD), we refer the reader to the two following surveys [66, 146].

In practice, we use the Darknet-53 architecture and pretrained weights proposed in tensorflow. This network uses 53 convolutional layers with 3x3 kernels in the beginning and 1x1 in the end. The model used was trained on the VOC dataset [19], containing 80 classes. Darknet-53 operates at a level close to state-of-the-art object detectors, but is faster because it uses less floating-point operations. The YOLO-v3 model was prepared with a threshold of 0.5 for both Intersection over union (IOU) and the loss function. During our evaluation (Section 6.1), the score threshold to keep a bounding box, as well as the IOU threshold were both set to 0.5 as well. To generate the search galleries, we only use the output corresponding to the person class from the object detector.

5.4 Classic Person Re-ID

The method used to perform classic person Re-ID in this paper is the same as proposed by Ejaz2015, called *An Improved Deep Learning Architecture for Person Re-Identification*. From now on, we refer to this method as SiamIDL. This method used the following deep neural network architecture: two layers of tied convolution with max pooling, cross-input neighborhood differences, patch summary features, across-patch features, higher-order relationships, and finally, a softmax function to yield the final estimate of whether the input images are of the same person or not (architecture is shown in Fig 3.10). They achieved 54.74% vs. 20.65% over the CUHK-03 dataset, more than double from the previous methods. This work was compared against KISSME, eSDC, SDALF, ITML, logistic distance metric learning (LDM), largest margin nearest neighbor (LMNN), metric learning to rank (RANK), and directly using Euclidean distance to compare features.

We select this work because the authors provide a base code repository to perform training and validation. Also, this works fits perfectly in our broad pipeline because it receives two images as input, and the output is a score of similarity between 0 and 1.

The first step was to validate this method's results over the same dataset where it was evaluated(CUHK-03[54]). The dataset provides a folder train where there are 7239 images. The organization is as follows: there are 742 Ids, every Id has between eight and ten hand-labeled images extracted from different frames, these have different sizes, and some have missing body parts. Also, we have another folder for validation where there are 938 images; these images don't appear in the train folder, and Ids are different from the train folder. There are 99 Ids, and every Id has between eight and ten images with different sizes and view positions.

The second step was to evaluate this method over the PRID2011[36] (introduced in 5.1) dataset, our principal data for FF-PRID. For implementation, we used the authors' source code and trained the network using the training set of the CUHK-03 dataset [54]. We use the same parameters as in the original paper: batch_size=50, max_steps=210 000, and learning_rate=0.01. The Cumulative Matching Characteristics (CMC) are computed on both the validation folder of CUHK-03 (938 images) and the original PRID dataset to evaluate the model. Results are presented in Section 6.2. We save final weights to use them to compute over the validation folder. We didn't re-train the model for this step.

Chapter 6

Results

To demonstrate the importance of considering the FF-PRID pipeline as a whole, and thus corroborate the usefulness of the proposed metrics, the evaluation conducted in this paper is three-fold. First, the Object Detection model is evaluated independently on a raw video from the PRID-2011 dataset. Then, the classic Re-ID model is tested on both the CUHK-03 validation set and on the official PRID-2011 dataset. Finally, we evaluate the full FF-PRID pipeline using our metrics on the modified PRID-2011 dataset.

6.1 Evaluation of the Object Detection model

The PRID-2011 dataset was initially created to evaluate classic Re-ID models. Hence, occluded persons, persons with less than five confidence frames, as well as distorted images caused by tracking and annotation errors were removed from the list of bounding boxes (see Figure 6.3b). To achieve a correct evaluation of YOLO-v3 on the PRID-2011 videos, it is necessary to manually add the bounding boxes of these people who were ignored during dataset creation. To do this, the *LabelIMG* tool was used and we added a total of 37.772 bounding boxes for the labels of video B. The results obtained for pedestrian detection with YOLO-v3 on the PRID-2011 videos are presented in Table 6.1. These results correspond to the model that was used to generate the search gallery for the classic Re-ID model (see 4.2b).

When analyzing visually the output produced by YOLO-v3, the results on PRID-2011 video B seem almost perfect. In this way, the difference in the results between OBB and OBB+MBB can be interpreted as the number of entire human bodies which where manually filtered by the annotators of the original dataset (e.g. partially overlapping persons).

Table 6.1: Evaluation of the YOLO-v3 model for pedestrian detection on the raw video B from the PRID-2011 dataset. For the Original Bounding Boxes (OBB) rows, metrics were computed using only the bounding boxes available from the original dataset as ground truth. For the OBB + Manually added Bounding Boxes (MBB) rows, the bounding boxes added using the LabelImg tool were also considered.

	Precision	Recall	F1-score	\mathbf{mAP}
OBB	0.462	0.866	0.603	45.53%
OBB + MBB	0.761	0.824	0.791	69.50%

On the other hand, the remaining errors for the OBB+MBB case mostly correspond to incomplete body parts, such as legs, arms or torso, which we did not include in our ground truth bounding boxes (see Fig. 6.1). An object detector, such as YOLO-v3, is trained to find the particular characteristics of the object of interest in an image and thus generates bounding boxes for the cases mentioned above. These cases constitute an important discrepancy between the domain on which the classic Re-ID model was trained and the images generated by the OD model. Such domain shift in the inputs of the C-PRID model can be a major source of errors for the full FF-PRID pipeline.



Figure 6.1: Different possible mistakes for cropping. (Source: author).

6.2 Evaluation of the Person Re-ID model

To evaluate the SiamIDL model used in our pipeline, we compute the CMC curves for both the validation set of CUHK-03 and PRID-2011. These results can be seen on Fig. 6.2. The evaluation on CUHK-03 is used to validate the training of our model by comparing our results with the ones obtained in the original paper. The blue curve obtained on Fig. 6.2 is very similar to the experimental results obtained in [17]. The red curve on Fig. 6.2 shows the results of a test performed on the first 200 Ids from view A of PRID-2011. We can see that the results obtained were good, with more than 48% on Rank 1 and more than 95% on Rank 20. We note that no additional training was conducted on the PRID-2011 dataset and only the weights trained on CUHK-03 are used in this validation. This last experiment corresponds to the practical scenario of deploying Re-ID in new environments (e.g. new city, new shopping center), where it would be impractical to create a new custom training dataset for every new implementation.



Figure 6.2: CMC curve on CUHK-03 validation set and on view A of PRID-2011, using a SiamIDL model trained on CUHK-03 training set. (Source: author).

The fact that a network trained on CUHK-03 can generalize to data from another dataset shows that the proposed Re-ID model is able to learn cross-domain Re-ID. Indeed, the kind of images used for training are very different than the images encountered at inference time (see Fig. 6.3). This property is interesting as the domains encountered for every new implementation vary a lot depending on the quality of the cameras, the



distance to the people and the illumination, among other factors.

(b) PRID 2011

Figure 6.3: Example images from the CUHK-03 and the PRID-2011 datasets. (Source: CUHK[54] and PRID2011[36])

6.3 Evaluation of the full pipeline for FF-PRID

For evaluation, we selected 10 sets of two minutes videos from each camera view. For each short video sequence, approximately 4 query images were selected. In total, our evaluation consists of 20 videos and 73 queries (36 for view A and 37 for view B). Each query appears in its associated video at least in one frame, but does not necessarily appear in each sub-videos after splitting into shorter sequences (see Fig. 4.1). To evaluate the influence of the different parameters of the FF-PRID pipeline, i.e. the number of frames for video splitting τ , the threshold for alert generation β and the number of candidates shown to the monitoring agent η , we use different values for each parameter. Thus, we test $\tau \in \{10, 100, 1000\}, \eta \in \{1, 10, 20\}$ and the threshold β is computed for various values in the interval [0.5, 0.98]. The Figures. 6.4, 6.5 and 6.6 shows the Finding Rate (FR) and True Validation Rate (TVR) curves for different values of τ , β and η .



Figure 6.4: Graphic with $\tau = 10$. (Source: author).



Figure 6.5: Graphic with $\tau = 100$. (Source: author).

6.3.1 Influence of the FF-PRID parameters

As we can see in these graphs, for all values of τ and η , the FR curves decrease when β increases. This behavior can be explained by the fact that a larger β means that the model will raise less alerts and is more likely to miss the query. However, with $\tau = 1000$, the decreasing effect is less noticeable. This is because when considering larger galleries, the model has more chances of finding a similar image and having at least one high confidence prediction. In contrast, the three TVR curves demonstrate the opposite behavior and are increasing with β . This also makes sense as increasing β correspond to reducing the accepted confidence range and thus calling the agent with less frequency. However, except for the case $\tau = 1000$, we note that the values of the different TVR are all very low, meaning that the human monitoring agent would be called in many unnecessary cases.

Furthermore, as expected, $\eta = 10$ and $\eta = 20$ performed much better than $\eta = 1$ for all configurations of τ and β . Indeed, the C-PRID models are not perfect and training Re-ID models with very high top 1 accuracy is hard. In contrast, decreasing η , reduces



Figure 6.6: Graphic with $\tau = 1000$. (Source: author).

the amount of work for the monitoring agent as it needs to control less image samples.

Finally, the FR curves present better results for $\tau = 100$ than for the two other tested values. This is because the raw video is split into sub-videos which are neither too short nor too long. This way, the query appears on the video for a sufficient amount of time to be recognize and there are not too many distractors to confuse the network.

6.3.2 Qualitative Evaluation

In Fig. 6.7, an example of the propositions shown to the monitoring agent is presented. This is the interface that we used for testing the approach and computing the final metrics scores. The text box present in the interface should be filled with the best ranked image representing the query, and 0 if the query is not present in the proposals. For example, in Fig. 6.7 the operator should enter 1.

When we carried out the evaluation, we observed that way too many alert calls occurred. Also, when a sub-video of τ frames contains too many persons, a variety of



Figure 6.7: Example of the interface for alert validation with $\eta = 6$. (Source: author).

cropped images are presented, not always representing the query. We also noted that the imperfect bounding boxes produced by YOLO-v3 (see Fig. 6.1) had a strong negative influence on the results.

6.3.3 Further considerations

The results obtained for the FF-PRID problem suggest that careful selection of the tunable parameters (τ , β and η) is paramount. Indeed, with proper selection we can reach an FR of almost 80% with a TVR of 26%. Although the score that we managed to reach for the Finding Rate are satisfactory, we acknowledge that the TVR is still too low for the method to be used practically, as the operator would be called too many times if dealing with several cameras at the same time. These mixed results emphasize the importance of considering the FF-PRID problem as a whole and suggest that changing the paradigm for person Re-ID might be the best way to obtain applicable solution for tomorrow's cities.

As already mentioned, the discrepancy between the training domain of the classic Re-ID model and the domain generated by OD is a possible reason for the results obtained. Another possible reason for the low TVR is that SiamIDL is a closed-world method, i.e. it supposes that the query is always present in the search gallery. Hence, the highest ranked prediction tend to have very high confidence scores and to raise alerts very frequently. In the validation example studied here, sometimes the query is not present, thus defining an open set scenario [45].

6.4 Some Observations

We have been able to observe that our approach to solving the FF-PRID problem has meager results. The objective of our work is not to develop an optimal model. However, we analyze the limitations with the following observations:

- An important observation is that we do not have good results for only considering YOLO as a person detector. The ideal would be to consider other options such as SSD, RCNN, etc.
- We are not getting by using a tracking algorithm to aid in processing before reidentification. Therefore, when our approach shows the results to the user, many repeat people are shown.
- In the same way as using only YOLO, in the case of using a C-PRID model, we only consider SiamIDL because it is a classic technique, video-based Re-id or other Re-id approaches are not being considered.
- Another possible reason for the low results is that we use two separate models, the output of YOLO being the input of SiamIDL. However, they do not work as a unified network, this causes the process to be slower, and the characteristics obtained by YOLO are wasted.
- Concerning the characteristics of the images of people, there is a problem with the C-PRID methods, which is to confuse people with others, either because of their way of dressing or because of elements outside of people.

Outside of the proposed model, we believe that it is essential to tune the β , τ , and η parameters of our Hybrid Framework. On the other hand, our model's leading cause has terrible results in the framework because it generates many false alerts. This induces a decrease in TVR, harming FR. False alerts are caused because our Re-id model is a Close world Re-ID algorithm, which means that it has only been trained in situations where the query always exists in the search gallery. Therefore, it cannot differentiate itself from other people, so the model generates false alerts, believing that it found a similar person. All those SiamIDL problems cause it to generate false alerts when the target is not found in our gallery.

Chapter 7

Conclusions

7.1 Conclusion

In the last couple of years, new considerations to deal with practical challenges for implementation of Re-ID have started to appear. In this way, a new metric to measure the cost of finding all the correct matches was introduced in [135]. The open-world setting is starting to gain importance because of its higher real-world relevance [45]. Finally, gaitbased Re-ID is a recent field that aims to identify people by their gait in unconstrained scenarios, typical of surveillance video systems. This is better in long-term scenarios than appearance-based Re-ID [77]. The research presented in dissertation is a continuation of these works as it also attempts to deal with real-world constraints in the practical implementation of Re-ID. In our case, we claim that it is necessary to consider full video frames as input instead of pre-cropped images in order to build solutions that can be evaluated and implemented and in practical scenarios.

In this work we claim that the classic approach for person Re-ID is not sufficient to develop practical implementations of Re-ID for security application, which requires to process the full frames of the cameras stream (FF-PRID) instead of pre-cropped clean images of people. To support this claim, we build a two steps FF-PRID pipeline. First, persons bounding boxes are extracted from the input video using a state of the art object detection model (YOLO-v3) to generate a search gallery. Then, the query is searched in the gallery using a good Re-ID model (SiamIDL). A framework embedding these two sub-modules is presented, including a human monitoring agent in the loop in order to strengthen the results. We present two new metrics in order to evaluate the proposed FF-PRID pipeline. The metrics are used to evaluate how many times the query is found when it is present in the video (Finding Rate) and how many times the query is present when the agent is solicited (True Validation Rate). These framework and metrics are, to the best of our knowledge, the first proposed approaches to evaluate a FF-PRID model, looking for persons directly in the entire video frames.

Our experimental results were conducted on a modified version of the PRID-2011 dataset. We demonstrated that both the OD model and the classic Re-ID model managed to perform well on the new dataset without additional training. However, the final results for FF-PRID, evaluated using FR and TVR, were not sufficient to deploy FF-PRID in production. Although choosing the right parameters in our framework enabled us to reach a good FR score (> 80%), we were not able to obtain a TVR much better than 25%, which means that most of the time the operator calls were unnecessary. Some possible explanations for these results were discussed as well as possible improvements. However, these mixed results emphasize the importance of considering Re-ID in the FF-PRID setting if we want to develop methods that can be used in practical scenarios. We believe that many improvements could be achieved if the community starts investigating Re-ID solutions for the Full Frame setting instead of focusing only on the classic precropped image-based setting.

7.2 Future works

After demonstrating the importance of considering the FF-PRID setting, the next step is to improve the proposed pipeline and get closer to solving FF-PRID. A possible direction to achieve this is to consider video-based classic Re-ID methods [79, 52]. Another natural option is to consider the open-world person Re-ID setting instead of closed-world [45]. We also plan to train more specific pedestrian detection techniques, focusing on recognizing only full-bodies. Finally, another potential improvement to address the problem would consist in building a large dataset of annotated videos, which could be used for training an end-to-end model for the whole FF-PRID application. This approach sound promising in regards with the success of end-to-end approaches in solving complex tasks lately [46, 43].

References

- BAK, S.; CARR, P.; LALONDE, J.-F. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference* on Computer Vision (ECCV) (2018), pp. 189–205.
- [2] BAZZANI, L.; CRISTANI, M.; PERINA, A.; FARENZENA, M.; MURINO, V. Multiple-shot person re-identification by HPE signature. *Proceedings - International Conference on Pattern Recognition*, August (2010), 1413–1416.
- [3] BEDAGKAR-GALA, A.; SHAH, S. K. A survey of approaches and trends in person re-identification. *Image and Vision Computing 32*, 4 (2014), 270–286.
- [4] BRUNETTI, A.; BUONGIORNO, D.; TROTTA, G. F.; BEVILACQUA, V. Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. *Neurocomputing 300* (2018), 17–33.
- [5] CAI, Y.; PIETIKÄINEN, M. Person Re-identification based on global color context. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6468 LNCS, PART1 (2011), 205– 215.
- [6] CAI, Y.; TAKALA, V.; PIETIKAINEN, M. Matching groups of people by covariance descriptor. In 2010 20th International Conference on Pattern Recognition (2010), IEEE, pp. 2744–2747.
- [7] CAMPS, O.; GOU, M.; HEBBLE, T.; KARANAM, S.; LEHMANN, O.; LI, Y.; RADKE, R. J.; WU, Z.; XIONG, F. From the lab to the real world: Re-identification in an airport camera network. *IEEE transactions on circuits and systems for video* technology 27, 3 (2016), 540–553.
- [8] CHANG, X.; HOSPEDALES, T. M.; XIANG, T. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition (2018), pp. 2109–2118.
- [9] CHEN, D.; LI, H.; LIU, X.; SHEN, Y.; SHAO, J.; YUAN, Z.; WANG, X. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 54–70.
- [10] CHEN, D.; LI, H.; XIAO, T.; YI, S.; WANG, X. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 1169–1178.

- [11] CHENG, D.; GONG, Y.; ZHOU, S.; WANG, J.; ZHENG, N. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings* of the iEEE conference on computer vision and pattern recognition (2016), pp. 1335– 1344.
- [12] COCCHIARELLA, N. Sortals, natural kinds and re-identification. Logique et analyse 20, 80 (1977), 439–474.
- [13] DAS, A.; CHAKRABORTY, A.; ROY-CHOWDHURY, A. K. Consistent reidentification in a camera network. In *European conference on computer vision* (2014), Springer, pp. 330–345.
- [14] DAS, A.; PANDA, R.; ROY-CHOWDHURY, A. Active image pair selection for continuous person re-identification. In 2015 IEEE International Conference on Image Processing (ICIP) (2015), IEEE, pp. 4263–4267.
- [15] DAS, A.; PANDA, R.; ROY-CHOWDHURY, A. K. Continuous adaptation of multicamera person identification models through sparse non-redundant representative selection. *Computer Vision and Image Understanding* 156 (2017), 66–78.
- [16] DOLLÁR, P.; WOJEK, C.; SCHIELE, B.; PERONA, P. Pedestrian detection: A benchmark. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (2009), IEEE, pp. 304–311.
- [17] EJAZ, A.; JONES, M.; MARKS, T. K. An Improved Deep Learning Architecture for Person Re-Identification. *Cvpr* (2015), 3908–3916.
- [18] ELYOR, K.; TAO, X.; ZHENYONG, F.; SHAOGANG, G. Person re-identification by unsupervised 11 graph learning. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands (2016), pp. 8–16.
- [19] EVERINGHAM, M.; VAN GOOL, L.; WILLIAMS, C. K.; WINN, J.; ZISSERMAN, A. The pascal visual object classes (voc) challenge. *International journal of computer* vision 88, 2 (2010), 303–338.
- [20] FAN, H.; ZHENG, L.; YAN, C.; YANG, Y. Unsupervised person re-identification: Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14, 4 (2018), 1–18.
- [21] FAN, X.; JIANG, W.; LUO, H.; FEI, M. Spherereid: Deep hypersphere manifold embedding for person re-identification. *Journal of Visual Communication and Image Representation 60* (2019), 51–58.
- [22] FU, Y.; WANG, X.; WEI, Y.; HUANG, T. Sta: Spatial-temporal attention for largescale video-based person re-identification. In *Proceedings of the AAAI conference* on artificial intelligence (2019), vol. 33, pp. 8287–8294.
- [23] GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014), pp. 580–587.
- [24] GIRSHICK, R. B. Fast R-CNN. CoRR abs/1504.08083 (2015).
- [25] GOU, M.; WU, Z.; RATES-BORRAS, A.; CAMPS, O.; RADKE, R. J., ET AL. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE transactions on pattern analysis and machine intelligence 41*, 3 (2018), 523–536.
- [26] GRAY, D.; TAO, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision* (2008), Springer, pp. 262–275.
- [27] GRAY, D.; TAO, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision* (2008), Springer, pp. 262–275.
- [28] HAHNEL, M.; KLUNDER, D.; KRAISS, K.-F. Color and texture features for person recognition. In 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541) (2004), vol. 1, IEEE, pp. 647–652.
- [29] HAMPAPUR, A.; BROWN, L.; CONNELL, J.; PANKANTI, S.; SENIOR, A.; TIAN, Y. Smart surveillance: applications, technologies and implications. In Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint (2003), vol. 2, IEEE, pp. 1133–1138.
- [30] HAQUE, A.; ALAHI, A.; FEI-FEI, L. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition (2016), pp. 1229–1238.
- [31] HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. B. Mask R-CNN. CoRR abs/1703.06870 (2017).
- [32] HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (2016), pp. 770–778.
- [33] HE, L.; LIANG, J.; LI, H.; SUN, Z. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7073–7082.
- [34] HE, R.; TAN, T.; DAVIS, L.; SUN, Z. Learning structured ordinal measures for video based face recognition. *Pattern Recognition* 75 (2018), 4–14.
- [35] HERMANS, A.; BEYER, L.; LEIBE, B. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017).
- [36] HIRZER, M.; BELEZNAI, C.; ROTH, P. M.; BISCHOF, H. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Im*age analysis (2011), Springer, pp. 91–102.
- [37] HIRZER, M.; BELEZNAI, C.; ROTH, P. M.; BISCHOF, H. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Im*age analysis (2011), Springer, pp. 91–102.

- [38] HOU, R.; MA, B.; CHANG, H.; GU, X.; SHAN, S.; CHEN, X. Vrstc: Occlusionfree video person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 7183–7192.
- [39] HOU, Y.; ZHENG, L.; WANG, Z.; WANG, S. Locality aware appearance metric for multi-target multi-camera tracking. arXiv preprint arXiv:1911.12037 (2019).
- [40] HU, T.-Y.; HAUPTMANN, A. G. Multi-shot person re-identification through set distance with visual distributional representation. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval* (2019), pp. 262–270.
- [41] HUANG, H.; LI, D.; ZHANG, Z.; CHEN, X.; HUANG, K. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 5098–5107.
- [42] KARIANAKIS, N.; LIU, Z.; CHEN, Y.; SOATTO, S. Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In *Proceedings of* the European Conference on Computer Vision (ECCV) (2018), pp. 715–733.
- [43] KENDALL, A.; MARTIROSYAN, H.; DASGUPTA, S.; HENRY, P.; KENNEDY, R.; BACHRACH, A.; BRY, A. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer* Vision (2017), pp. 66–75.
- [44] LAN, X.; WANG, H.; GONG, S.; ZHU, X. Deep reinforcement learning attention selection for person re-identification. arXiv preprint arXiv:1707.02785 (2017).
- [45] LENG, Q.; YE, M.; TIAN, Q. A survey of open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* (2019).
- [46] LEVINE, S.; FINN, C.; DARRELL, T.; ABBEEL, P. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research* 17, 1 (2016), 1334– 1373.
- [47] LI, D.; CHEN, X.; ZHANG, Z.; HUANG, K. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE* conference on computer vision and pattern recognition (2017), pp. 384–393.
- [48] LI, J.; WANG, J.; TIAN, Q.; GAO, W.; ZHANG, S. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 3958–3967.
- [49] LI, M.; ZHU, X.; GONG, S. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision* (ECCV) (2018), pp. 737–753.
- [50] LI, M.; ZHU, X.; GONG, S. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision* (ECCV) (2018), pp. 737–753.
- [51] LI, S.; BAK, S.; CARR, P.; WANG, X. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference* on Computer Vision and Pattern Recognition (2018), pp. 369–378.

- [52] LI, S.; BAK, S.; CARR, P.; WANG, X. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference* on Computer Vision and Pattern Recognition (2018), pp. 369–378.
- [53] LI, S.; XIAO, T.; LI, H.; ZHOU, B.; YUE, D.; WANG, X. Person search with natural language description. In *Proceedings of the IEEE Conference on Computer* Vision and Pattern Recognition (2017), pp. 1970–1979.
- [54] LI, W.; ZHAO, R.; XIAO, T.; WANG, X. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer* vision and pattern recognition (2014), pp. 152–159.
- [55] LI, W.; ZHAO, R.; XIAO, T.; WANG, X. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer* vision and pattern recognition (2014), pp. 152–159.
- [56] LI, W.-H.; ZHONG, Z.; ZHENG, W.-S. One-pass person re-identification by sketch online discriminant analysis. *Pattern Recognition 93* (2019), 237–250.
- [57] LI, X.; WU, A.; ZHENG, W.-S. Adversarial open-world person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV) (2018), pp. 280–296.
- [58] LI, X.; ZHENG, W.-S.; WANG, X.; XIANG, T.; GONG, S. Multi-scale learning for low-resolution person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3765–3773.
- [59] LI, Y.; WU, Z.; KARANAM, S.; RADKE, R. J. Real-world re-identification in an airport camera network. In *Proceedings of the International Conference on Distributed Smart Cameras* (2014), pp. 1–6.
- [60] LIN, J.; REN, L.; LU, J.; FENG, J.; ZHOU, J. Consistent-aware deep learning for person re-identification in a camera network. In *Proceedings of the IEEE conference* on computer vision and pattern recognition (2017), pp. 5771–5780.
- [61] LIN, Y.; ZHENG, L.; ZHENG, Z.; WU, Y.; HU, Z.; YAN, C.; YANG, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognition 95* (2019), 151–161.
- [62] LISANTI, G.; MARTINEL, N.; DEL BIMBO, A.; LUCA FORESTI, G. Group reidentification via unsupervised transfer of sparse features encoding. In *Proceedings* of the IEEE International Conference on Computer Vision (2017), pp. 2449–2458.
- [63] LIU, F.; ZHANG, L. View confusion feature learning for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (2019), pp. 6639–6648.
- [64] LIU, H.; FENG, J.; JIE, Z.; JAYASHREE, K.; ZHAO, B.; QI, M.; JIANG, J.; YAN, S. Neural person search machines. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 493–501.

- [65] LIU, J.; NI, B.; YAN, Y.; ZHOU, P.; CHENG, S.; HU, J. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4099–4108.
- [66] LIU, L.; OUYANG, W.; WANG, X.; FIEGUTH, P.; CHEN, J.; LIU, X.; PIETIKÄI-NEN, M. Deep learning for generic object detection: A survey. *International journal* of computer vision 128, 2 (2020), 261–318.
- [67] LIU, W.; ANGUELOV, D.; ERHAN, D.; SZEGEDY, C.; REED, S.; FU, C.-Y.; BERG, A. C. Ssd: Single shot multibox detector. In *European conference on computer vision* (2016), Springer, pp. 21–37.
- [68] LIU, X.; BI, S.; MA, X.; WANG, J. Multi-instance convolutional neural network for multi-shot person re-identification. *Neurocomputing* 337 (2019), 303–314.
- [69] LIU, Z.; WANG, D.; LU, H. Stepwise metric promotion for unsupervised video person re-identification. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2429–2438.
- [70] LOY, C. C.; LIU, C.; GONG, S. Person re-identification by manifold ranking. In 2013 IEEE International Conference on Image Processing (2013), IEEE, pp. 3567– 3571.
- [71] LUO, C.; CHEN, Y.; WANG, N.; ZHANG, Z. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE/CVF International Conference* on Computer Vision (2019), pp. 4976–4985.
- [72] LUO, H.; JIANG, W.; GU, Y.; LIU, F.; LIAO, X.; LAI, S.; GU, J. A strong baseline and batch normneuralization neck for deep person reidentification. arXiv preprint arXiv:1906.08332 (2019).
- [73] MARTINEL, N.; DAS, A.; MICHELONI, C.; ROY-CHOWDHURY, A. K. Temporal model adaptation for person re-identification. In *European conference on computer* vision (2016), Springer, pp. 858–877.
- [74] MARTINEL, N.; MICHELONI, C.; FORESTI, G. L. A pool of multiple person reidentification experts. *Pattern Recognition Letters* 71 (2016), 23–30.
- [75] MCLAUGHLIN, N.; DEL RINCON, J. M.; MILLER, P. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference* on computer vision and pattern recognition (2016), pp. 1325–1334.
- [76] MUNJAL, B.; AMIN, S.; TOMBARI, F.; GALASSO, F. Query-guided end-to-end person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 811–820.
- [77] NAMBIAR, A.; BERNARDINO, A.; NASCIMENTO, J. C. Gait-based person reidentification: A survey. ACM Computing Surveys (CSUR) 52, 2 (2019), 1–34.
- [78] NGUYEN, D. T.; HONG, H. G.; KIM, K. W.; PARK, K. R. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17, 3 (2017), 605.

- [79] OUYANG, D.; ZHANG, Y.; SHAO, J. Video-based person re-identification via spatio-temporal attentional and two-stream fusion convolutional networks. *Pattern Recognition Letters* 117 (2019), 153–160.
- [80] PLANTINGA, A. Things and persons. The Review of Metaphysics (1961), 493–519.
- [81] QIAN, X.; FU, Y.; XIANG, T.; WANG, W.; QIU, J.; WU, Y.; JIANG, Y.-G.; XUE, X. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 650–667.
- [82] REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer* vision and pattern recognition (2016), pp. 779–788.
- [83] REDMON, J.; FARHADI, A. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 7263–7271.
- [84] REDMON, J.; FARHADI, A. Yolov3: An incremental improvement. arXiv (2018).
- [85] REN, S.; HE, K.; GIRSHICK, R.; SUN, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (2015), pp. 91–99.
- [86] RISTANI, E.; TOMASI, C. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 6036–6046.
- [87] RORTY, A. O. The transformations of persons. *Philosophy* 48, 185 (1973), 261–275.
- [88] SARFRAZ, M. S.; SCHUMANN, A.; EBERLE, A.; STIEFELHAGEN, R. A posesensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 420–429.
- [89] SATTA, R.; PALA, F.; FUMERA, G.; ROLI, F. Real-time appearance-based person re-identification over multiple kinecttm cameras. In VISAPP (2) (2013), pp. 407– 410.
- [90] SHEN, Y.; LI, H.; YI, S.; CHEN, D.; WANG, X. Person re-identification with deep similarity-guided graph neural network. In *Proceedings of the European conference* on computer vision (ECCV) (2018), pp. 486–504.
- [91] SHENOI, A.; PATEL, M.; GWAK, J.; GOEBEL, P.; SADEGHIAN, A.; REZATOFIGHI, H.; MARTIN-MARTIN, R.; SAVARESE, S. Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset. arXiv preprint arXiv:2002.08397 (2020).
- [92] SHI, H.; YANG, Y.; ZHU, X.; LIAO, S.; LEI, Z.; ZHENG, W.; LI, S. Z. Embedding deep metric for person re-identification: A study against large variations. In *European conference on computer vision* (2016), Springer, pp. 732–748.
- [93] SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

- [94] SONG, C.; HUANG, Y.; OUYANG, W.; WANG, L. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 1179–1188.
- [95] SONG, J.; YANG, Y.; SONG, Y.-Z.; XIANG, T.; HOSPEDALES, T. M. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings* of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 719–728.
- [96] SU, C.; ZHANG, S.; XING, J.; GAO, W.; TIAN, Q. Deep attributes driven multicamera person re-identification. In *European conference on computer vision* (2016), Springer, pp. 475–491.
- [97] SUBRAMANIAM, A.; NAMBIAR, A.; MITTAL, A. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 562–572.
- [98] SUH, Y.; WANG, J.; TANG, S.; MEI, T.; LEE, K. M. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference* on Computer Vision (ECCV) (2018), pp. 402–419.
- [99] SUMARI, F. O.; MACHACA, L.; HUAMAN, J.; CLUA, E. W.; GUÉRIN, J. Towards practical implementations of person re-identification from full video frames. *Pattern Recognition Letters* 138 (2020), 513–519.
- [100] SUN, X.; ZHENG, L. Dissecting person re-identification from the viewpoint of viewpoint. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 608–617.
- [101] SUN, Y.; XU, Q.; LI, Y.; ZHANG, C.; LI, Y.; WANG, S.; SUN, J. Perceive where to focus: Learning visibility-aware part-level features for partial person reidentification. In *Proceedings of the IEEE/CVF Conference on Computer Vision* and Pattern Recognition (2019), pp. 393–402.
- [102] SUN, Y.; ZHENG, L.; DENG, W.; WANG, S. Svdnet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision (2017), pp. 3800–3808.
- [103] SUN, Y.; ZHENG, L.; YANG, Y.; TIAN, Q.; WANG, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European conference on computer vision (ECCV) (2018), pp. 480– 496.
- [104] TANG, S.; ANDRILUKA, M.; ANDRES, B.; SCHIELE, B. Multiple people tracking by lifted multicut and person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 3539–3548.
- [105] TOGOOTOGTOKH, E.; MICHELONI, C.; FORESTI, G. L.; MARTINEL, N. An efficient uav-based artificial intelligence framework for real-time visual tasks. arXiv preprint arXiv:2004.06154 (2020).

- [106] VARIOR, R. R.; SHUAI, B.; LU, J.; XU, D.; WANG, G. A siamese long shortterm memory architecture for human re-identification. In *European conference on computer vision* (2016), Springer, pp. 135–153.
- [107] WANG, C.-Y.; CHEN, P.-Y.; CHEN, M.-C.; HSIEH, J.-W.; LIAO, H.-Y. M. Real-time video-based person re-identification surveillance with light-weight deep convolutional networks. In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (2019), IEEE, pp. 1–8.
- [108] WANG, F.; ZUO, W.; LIN, L.; ZHANG, D.; ZHANG, L. Joint learning of singleimage and cross-image representations for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 1288–1296.
- [109] WANG, H.; FAN, Y.; WANG, Z.; JIAO, L.; SCHIELE, B. Parameter-Free Spatial Attention Network for Person Re-Identification.
- [110] WANG, H.; ZHU, X.; XIANG, T.; GONG, S. Towards unsupervised open-set person re-identification. In 2016 IEEE International Conference on Image Processing (ICIP) (2016), IEEE, pp. 769–773.
- [111] WANG, T.; GONG, S.; ZHU, X.; WANG, S. Person re-identification by video ranking. In *European conference on computer vision* (2014), Springer, pp. 688–703.
- [112] WANG, X. Intelligent multi-camera video surveillance: A review. Pattern recognition letters 34, 1 (2013), 3–19.
- [113] WANG, X.; DORETTO, G.; SEBASTIAN, T.; RITTSCHER, J.; TU, P. Shape and appearance context modeling. In 2007 ieee 11th international conference on computer vision (2007), Ieee, pp. 1–8.
- [114] WANG, X.; LIU, M.; RAYCHAUDHURI, D. S.; PAUL, S.; WANG, Y.; ROY-CHOWDHURY, A. K. Learning person re-identification models from videos with weak supervision. *IEEE Transactions on Image Processing 30* (2021), 3017–3028.
- [115] WANG, X.; PANDA, R.; LIU, M.; WANG, Y.; ROY-CHOWDHURY, A. K. Exploiting global camera network constraints for unsupervised video person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).
- [116] WANG, Y.; WANG, L.; YOU, Y.; ZOU, X.; CHEN, V.; LI, S.; HUANG, G.; HARIHARAN, B.; WEINBERGER, K. Q. Resource aware person re-identification across multiple resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8042–8051.
- [117] WANG, Z.; YE, M.; YANG, F.; BAI, X.; SATOH, S. Cascaded sr-gan for scaleadaptive low resolution person re-identification. In *IJCAI* (2018), vol. 1, p. 4.
- [118] WEI, L.; ZHANG, S.; GAO, W.; TIAN, Q. Person transfer gan to bridge domain gap for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

- [119] WEI, L.; ZHANG, S.; GAO, W.; TIAN, Q. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer* vision and pattern recognition (2018), pp. 79–88.
- [120] WOJKE, N.; BEWLEY, A. Deep cosine metric learning for person re-identification. In 2018 IEEE winter conference on applications of computer vision (WACV) (2018), IEEE, pp. 748–756.
- [121] WU, A.; ZHENG, W.-S.; LAI, J.-H. Robust depth-based person re-identification. IEEE Transactions on Image Processing 26, 6 (2017), 2588–2603.
- [122] WU, A.; ZHENG, W.-S.; YU, H.-X.; GONG, S.; LAI, J. Rgb-infrared crossmodality person re-identification. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 5380–5389.
- [123] WU, Y.; LIN, Y.; DONG, X.; YAN, Y.; OUYANG, W.; YANG, Y. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), pp. 5177–5186.
- [124] WU, Y.; LIN, Y.; DONG, X.; YAN, Y.; OUYANG, W.; YANG, Y. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), pp. 5177–5186.
- [125] XIAO, T.; LI, H.; OUYANG, W.; WANG, X. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE* conference on computer vision and pattern recognition (2016), pp. 1249–1258.
- [126] XIAO, T.; LI, S.; WANG, B.; LIN, L.; WANG, X. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 3415–3424.
- [127] XIE, Z.; LI, L.; ZHONG, X.; ZHONG, L.; XIANG, J. Image-to-video person reidentification with cross-modal embeddings. *Pattern Recognition Letters* 133 (2020), 70–76.
- [128] XU, S.; CHENG, Y.; GU, K.; YANG, Y.; CHANG, S.; ZHOU, P. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In Proceedings of the IEEE international conference on computer vision (2017), pp. 4733–4742.
- [129] YAN, Y.; NI, B.; LIU, J.; YANG, X. Multi-level attention model for person reidentification. *Pattern Recognition Letters* 127 (2019), 156–164.
- [130] YAN, Y.; ZHANG, Q.; NI, B.; ZHANG, W.; XU, M.; YANG, X. Learning context graph for person search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 2158–2167.
- [131] YE, M.; LAN, X.; WANG, Z.; YUEN, P. C. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security* 15 (2019), 407–419.

- [132] YE, M.; LAN, X.; YUEN, P. C. Robust anchor embedding for unsupervised video person re-identification in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 170–186.
- [133] YE, M.; LI, J.; MA, A. J.; ZHENG, L.; YUEN, P. C. Dynamic graph co-matching for unsupervised video-based person re-identification. *IEEE Transactions on Image Processing 28*, 6 (2019), 2976–2990.
- [134] YE, M.; MA, A. J.; ZHENG, L.; LI, J.; YUEN, P. C. Dynamic label graph matching for unsupervised video re-identification. In *Proceedings of the IEEE international* conference on computer vision (2017), pp. 5142–5150.
- [135] YE, M.; SHEN, J.; LIN, G.; XIANG, T.; SHAO, L.; HOI, S. C. Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193 (2020).
- [136] YE, M.; WANG, Z.; LAN, X.; YUEN, P. C. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI* (2018), vol. 1, p. 2.
- [137] YE, M.; ZHANG, X.; YUEN, P. C.; CHANG, S.-F. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 6210–6219.
- [138] YI, D.; LEI, Z.; LIAO, S.; LI, S. Z. Deep metric learning for person reidentification. In 2014 22nd International Conference on Pattern Recognition (2014), IEEE, pp. 34–39.
- [139] YU, R.; DOU, Z.; BAI, S.; ZHANG, Z.; XU, Y.; BAI, X. Hard-aware point-to-set deep metric for person re-identification. In *Proceedings of the European conference* on computer vision (ECCV) (2018), pp. 188–204.
- [140] YU, T.; LI, D.; YANG, Y.; HOSPEDALES, T. M.; XIANG, T. Robust person re-identification by modelling feature uncertainty. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 552–561.
- [141] ZENG, K.; NING, M.; WANG, Y.; GUO, Y. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference* on Computer Vision and Pattern Recognition (2020), pp. 13657–13665.
- [142] ZHAO, H.; TIAN, M.; SUN, S.; SHAO, J.; YAN, J.; YI, S.; WANG, X.; TANG, X. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer* vision and pattern recognition (2017), pp. 1077–1085.
- [143] ZHAO, L.; LI, X.; ZHUANG, Y.; WANG, J. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international* conference on computer vision (2017), pp. 3219–3228.
- [144] ZHAO, R.; OUYANG, W.; WANG, X. Unsupervised salience learning for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition (2013), pp. 3586–3593.

- [145] ZHAO, Y.; SHEN, X.; JIN, Z.; LU, H.; HUA, X.-S. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings* of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 4913–4922.
- [146] ZHAO, Z.-Q.; ZHENG, P.; XU, S.-T.; WU, X. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems 30*, 11 (2019), 3212–3232.
- [147] ZHENG, L.; BIE, Z.; SUN, Y.; WANG, J.; SU, C.; WANG, S.; TIAN, Q. Mars: A video benchmark for large-scale person re-identification. In *European Conference* on Computer Vision (2016), Springer, pp. 868–884.
- [148] ZHENG, L.; SHEN, L.; TIAN, L.; WANG, S.; WANG, J.; TIAN, Q. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference* on computer vision (2015), pp. 1116–1124.
- [149] ZHENG, L.; SHEN, L.; TIAN, L.; WANG, S.; WANG, J.; TIAN, Q. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference* on computer vision (2015), pp. 1116–1124.
- [150] ZHENG, L.; YANG, Y.; HAUPTMANN, A. G. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016).
- [151] ZHENG, L.; ZHANG, H.; SUN, S.; CHANDRAKER, M.; YANG, Y.; TIAN, Q. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 1367–1376.
- [152] ZHENG, M.; KARANAM, S.; WU, Z.; RADKE, R. J. Re-identification with consistent attentive siamese networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 5735–5744.
- [153] ZHENG, W.-S.; GONG, S.; XIANG, T. Associating groups of people. In BMVC (2009), vol. 2, pp. 1–11.
- [154] ZHENG, W.-S.; GONG, S.; XIANG, T. Towards open-world person re-identification by one-shot group-based verification. *IEEE transactions on pattern analysis and machine intelligence 38*, 3 (2015), 591–606.
- [155] ZHENG, W.-S.; LI, X.; XIANG, T.; LIAO, S.; LAI, J.; GONG, S. Partial person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (2015), pp. 4678–4686.
- [156] ZHENG, Z.; YANG, X.; YU, Z.; ZHENG, L.; YANG, Y.; KAUTZ, J. Joint Discriminative and Generative Learning for Person Re-identification.
- [157] ZHENG, Z.; ZHENG, L.; YANG, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3754–3762.
- [158] ZHENG, Z.; ZHENG, L.; YANG, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3754–3762.

- [159] ZHONG, Z.; ZHENG, L.; KANG, G.; LI, S.; YANG, Y. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 13001–13008.
- [160] ZHONG, Z.; ZHENG, L.; LUO, Z.; LI, S.; YANG, Y. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of* the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019), pp. 598–607.
- [161] ZHOU, S.; WANG, J.; WANG, J.; GONG, Y.; ZHENG, N. Point to set similarity based deep feature learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 3741–3750.
- [162] ZHOU, Z.; HUANG, Y.; WANG, W.; WANG, L.; TAN, T. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4747–4756.
- [163] ZHU, X.; WU, B.; HUANG, D.; ZHENG, W.-S. Fast open-world person reidentification. *IEEE Transactions on Image Processing* 27, 5 (2017), 2286–2300.