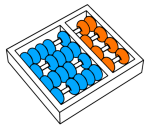
Universidade Estadual de Campinas
Instituto de Computação



Edgar Rodolfo Quispe Condori

Improved Person Re-Identification Through Semantic Parsing and Saliency Maps

Re-Identificação de Pessoas Refinada com Mapas de Segmentação Semântica e Saliências

CAMPINAS
2019

Edgar Rodolfo Quispe Condori

Improved Person Re-Identification Through Semantic Parsing and Saliency Maps

Re-Identificação de Pessoas Refinada com Mapas de Segmentação Semântica e Saliências

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. Hélio Pedrini

Este exemplar corresponde à versão final da Dissertação defendida por Edgar Rodolfo Quispe Condori e orientada pelo Prof. Dr. Hélio Pedrini.

CAMPINAS
2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Q48i

Quispe Condori, Edgar Rodolfo, 1994-
Improved person re-identification through semantic parsing and saliency maps / Edgar Rodolfo Quispe Condori. – Campinas, SP : [s.n.], 2019.

Orientador: Helio Pedrini.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Redes neurais convolucionais. 2. Visão por computador. 3. Reconhecimento de padrões. 4. Aprendizado de máquina. 5. Redes neurais (Computação). I. Pedrini, Helio, 1963-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Re-identificação de pessoas refinada com mapas de segmentação semântica e saliências
Palavras-chave em inglês:
Convolutional neural networks
Computer vision
Pattern recognition
Machine learning
Neural networks (Computer science)
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Hélio Pedrini
Otavio Augusto Bizetto Penatti
Ricardo da Silva Torres
Data de defesa: 14-03-2019
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-1661-3720
- Currículo Lattes do autor: http://lattes.cnpq.br/6931027168436055

Universidade Estadual de Campinas
Instituto de Computação

Edgar Rodolfo Quispe Condori

Improved Person Re-Identification Through Semantic Parsing and Saliency Maps

Re-Identificação de Pessoas Refinada com Mapas de Segmentação Semântica e Saliências

Banca Examinadora:

- Prof. Dr. Hélio Pedrini
  Instituto de Computação - UNICAMP
- Prof. Dr. Otavio Augusto Bizetto Penatti
  Samsung Research Institute Brazil
- Prof. Dr. Ricardo da Silva Torres
  Instituto de Computação - UNICAMP

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 14 de março de 2019

Acknowledgements

To my eternal hero, my mother Corina, because she taught me to work hard for my dreams and never give up. Many thanks!

I am grateful to my family, Jose, Hanan, Adolfo, Brisaida, Barbara and Gregorio, for their moral and emotional support. I am also thankful to my other family members for their good wishes.

Special gratitude to my colleagues from the Laboratory of Visual Informatics (LIV) in the Institute of Computing (IC). Special mentions to Anderson, Eldrey, George, and Helena. I will always consider them as my friends. Thanks for all that funny time learning about new things in research and life. A special mention to my Peruvian friends, because they helped me to feel that I am still in my country, especially to Darwin, Jose Luis and Jadisha. I have been lucky enough to have them as really close friends.

To my advisor Helio Pedrini for all his support, I learned a lot during the Master's degree and he was always open to listen to me. Thank you so much! Also, a special mention to my other professors as well: Adin Ramirez, Lehilton Chaves, Guido Costa, Julio Cesar dos Reis, Tomasz Kowaltowski, because they showed me new ways to learn and teach.

Finally, I would like to thank CAPES for the financial support and all the people who work in the Institute of Computing: Shirley, Regina, Denise, Wilson and all others who make our life easier in this wonderful workplace. Thanks to everyone who has had an impact on my life for the past two years, it was an amazing time.

Thank you all!
Rodolfo Quispe

Resumo

Reidentificação de pessoas é o processo de recuperar todas as instâncias de uma mesma pessoa presente em vídeos ou imagens capturadas por câmeras diferentes com uma visão não sobreposta. Esta é uma tarefa desafiadora devido a fatores como oclusões, diferentes pontos de vista, condições de iluminação, fundo complexo, configurações de câmera e deformação não rígida de corpos humanos. Além disso, este é considerado um tópico de pesquisa bastante ativo na indústria e nas universidades, devido às suas aplicações em vários campos, como vigilância, ciência forense, robótica e multimídia. Neste trabalho, uma nova abordagem, denominada Re-Identificação de Pessoas Baseada em Saliências e Segmentação Semântica (SSP-ReID) é apresentada e implementada para aproveitar as capacidades de ambas as características (mapas de saliência e de segmentação semântica) para guiar uma rede neural convolucional (CNN) de modo a aprender representações complementares que melhorem os resultados sobre as redes originais. O princípio de combinar várias características baseia-se em cenários específicos, em que uma resposta é superior à outra, favorecendo assim a fusão delas para aumentar a efetividade do método. Devido à sua definição, nossa abordagem pode ser facilmente aplicada a uma ampla variedade de redes e, em contraste com outros métodos competidores, nosso processo de treinamento segue protocolos simples e conhecidos. Uma extensa avaliação de nossa abordagem é apresentada com o uso de cinco módulos e três bases de dados. Experimentos de observação são realizados a fim de obter uma melhor compreensão do desempenho de cada módulo e os resultados experimentais demonstram a eficácia de nossa abordagem para reidentificação de pessoas. Além disso, a abordagem proposta foi combinada com técnicas de re-ranqueamento e os resultados obtidos foram comparados com abordagens da literatura, alcançando resultados competitivos.

Abstract

Person Re-Identification is the process of retrieving all instances of the same person from videos or images taken from different cameras with a non-overlapping view. This is a challenging task due to occlusions, different viewpoints, illumination conditions, background clutters, camera settings and non-rigid deformation of human bodies. Moreover, it is considered a very active research topic in the industry and academia due to its applications in various fields, such as surveillance, forensics, robotics, multimedia. In this work, we present and implement a novel framework, named Saliency-Semantic Parsing Re-Identification (SSP-ReID), for taking advantage of the capabilities of both clues (saliency and semantic parsing maps) to guide a backbone convolutional neural network (CNN) to learn complementary representations that improve the results over the original backbones. The insight of fusing multiple clues is based on specific scenarios in which one response is better than other, thus favoring the combination of them to increase effectiveness. Due to its definition, our framework can be easily applied to a wide variety of networks and, in contrast to other competitive methods, our training process follows simple and standard protocols. We present an extensive evaluation of our approach through five backbones and three benchmarks. Ablation experiments are performed in order to gain a better understanding of the performance of each module and experimental results demonstrate the effectiveness of our person re-identification framework. In addition, we combine our framework with re-ranking techniques and compare the results against state-of-the-art approaches, achieving competitive results.

List of Figures

List of Tables

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
CIFAR	Canadian Institute For Advanced Research
CMC	Cumulated Matching Characteristics
CNN	Convolutional Neural Network
CROSSE	Cros Entropy Loss
CUDA	Compute Unified Device Architecture
DPM	Deformable Part Model
FCNN	Fully Convolutional Neural Network
GAP	Global Average Pooling
GPU	Graphics Processing Unit
HA-CNN	Harmonious Attention Network
IC	Institute of Computing
JJPNet	Joint Human Parsing and Pose Estimation Network
LIP	Look into Person
LIV	Laboratory of Visual Informatics
LSR	Label Smoothing Regularizer
mAP	mean Average Precision
MLFN	Multi-Level Factorization Net
Re-Id	Person Re-Identification
ResNet	Residual Network
RR	Re-Ranking
S-ReID	Saliency Re-Identification
SP-ReID	Semantic Parsing Re-Identification
TRIP	Triplet Loss
SSP-ReID	Saliency-Semantic Parsing Re-Identification
UNICAMP	University of Campinas

Contents

Chapter 1

Introduction

This chapter describes the problem under investigation in this dissertation, as well as its motivations, challenges, objectives, contributions, research questions, and text organization.

1.1 Problem and Motivation

Security is a serious concern in society today. In this sense, technology is an important instrument to address this issue by creating intelligent methods that enable fast checking to block possible problematic circumstances. From this perspective, computer vision and machine learning fields play a major role because most security systems are camera-based.

A reliable capability for a camera-based security system is to be able to determine if two people from different camera views are the same person. This task is known as Person Re-Identification (Re-ID) and has been the focus of large research groups in academia and industry [20, 26, 47, 49, 51, 53, 60].

It is important to notice that Re-ID can be misinterpreted with person tracking. In fact, Re-ID arises from tracking, however, there are differences between them: tracking considers small variations in position, point of view, appearance, biometric profile. On the other hand, Re-ID aims to match instances of people during a time delay and/or a change in viewpoint, so that changes in position and point of view are more drastic.

Formally, Re-ID is defined as the task of matching all instances of the same person across multiple cameras with non-overlapping views [12]. There are two terms that are widely used in this problem: *probe* and *gallery*. Re-ID aims to compare a person of interest (*probe*), against a set of candidates (*gallery*). In addition, there are other important constraints to consider: (i) the problem is defined in the scenario with low-resolution cameras (e.g., security cameras at airports or universities). For this reason, methods such as face recognition are not applicable, thus current literature is based on the appearance of people, for instance, color of clothes, type of objects, among others; (ii) people are represented by their bounding boxes (Figure 1.1) and not the complete scene; (iii) Re-ID considers that there is not a long period of time (e.g., more than a few hours) between

the acquisition of the probe and the gallery; this is very important because appearance is the main clue used in state-of-the-art approaches [20, 26, 47, 49, 51, 53, 60].

Due to complex data variations caused by occlusions, different viewpoints, illumination conditions, background clutters, camera settings, and non-rigid deformation of human bodies (Figure 1.1), Re-ID remains an open problem. Moreover, in order to analyze Re-ID in a "real world" scenario, some existing works consider it as an identification task, which means that the probe has at least one instance in the gallery. On the other hand, this may not always occur in a real-world configuration.

Some authors suggest the use of a similarity function (for example, between probe and gallery) and set a threshold in this function to solve this specific issue, however, due to complex variations, it is common to have a low acceptance similarity value [12]. Liao et al. [30] consider that real-world Re-ID is divided into two steps: detection and recognition. Detection determines whether a probe is present in the gallery, whereas recognition assigns an ID (identity) to the accepted probe. This entails a new challenge because it is desirable that the gallery be dynamic, in a way that allows adding new individuals as already "seen", without a high computational cost. Therefore, there is a long way until Re-ID can be considered fully solved.

Re-ID can be addressed both in images (image Re-ID) and video sequences (video Re-ID). Image-based approaches use the term single-shot to refer the case where only two images of each person are available, one in the gallery and the other as the probe, and multi-shot when more than two images per person are available. Most of the video-based approaches consider Re-ID as a multi-shot scenario, however, the use of spatiotemporal information is a challenging topic to be explored [73].

Figure 1.1: Examples of input data for Re-ID, first and second image have the same identity, third and fourth image also have the same identity. Occlusions, different viewpoints, illumination conditions, background clutters are challenging issues. Adapted from Market1501 Dataset [69].

In this work, we use Deep Learning concepts to develop a method capable of achieving competitive results in image Re-ID. We consider these concepts due to outstanding results achieved in the Re-ID task [20, 26, 47, 49, 51]. More specifically, we propose the use of multiple clues to guide the process of learning features that really represent people. Based on this idea, the following hypothesis is proposed: "Deep architectures applied to Person Re-Identification can be enhanced by the use of multiple clues".

Since we aim to evaluate deep architectures in the context of Re-ID, we consider the most popular networks in the computer vision field and combine them with saliency/semantic parsing maps using a training process named fine-tuning. We consider the weights of the networks pre-trained by Krizhevsky et al. [22] and then train the network with new loss functions and feature combination.

1.2 Objectives

The main objective of this work is to develop a new approach to the Re-ID problem based on convolutional neural networks (CNNs) guided by multiple clues with competitive performance. For this objective, specific goals are considered:

- Exploration and evaluation of saliency and semantic segmentation applied to the Re-ID task in order to enhance results.
- Investigation of various state-of-the-art CNNs created for image classification for Re-ID.
- Evaluation of different loss functions in the context of Re-ID.
- Conduction of extensive experiments on various datasets to validate all modules of the proposed network.

1.3 Contributions

The main contributions of our work for image Re-ID are:

- A novel framework using saliency and semantic parsing for image Re-ID is proposed and evaluated. To the best of our knowledge, this is the first work that combines these two clues for Re-ID.
- Extensive experiments on three datasets and five backbones show the ability of our method to improve results and demonstrate that it can be used with many other backbones due to its definition.
- Different from other competitive methods, our framework takes full advantage of pre-trained models and requires a minimum number of fine-tuning epochs to reach competitive results. Moreover, our training process does not need to combine multiple Re-ID benchmarks. Our approach is able to achieve competitive results in all datasets evaluated: Market1501 (second/third best), CUHK (first/second best), DukeMTMC-ReID (second/third best).

1.4 Text Organization

This text is organized into five chapters. In Chapter 1, we presented an introduction to the problem, as well as main objectives and contributions of the research work. In Chapter 2, we describe a review of relevant concepts and works related to the Re-ID problem. The proposed method is explained in details in Chapter 3. Experiments and results are shown in Chapter 4. Finally, concluding remarks and directions for future work are summarized in Chapter 5.

Chapter 2

Background

This chapter is divided into two sections. Initially, relevant concepts related to our work are introduced. Then, a brief review of Person Re-Identification methods available in the literature is presented.

2.1 Related Concepts

This section examines important topics to help readers understand our research work, such as Person Re-Identification, Deep Learning, Saliency Detection and Semantic Parsing.

2.1.1 Person Re-Identification

Person Re-Identification [12, 23, 43, 58], commonly known as Re-ID, is defined as the task of matching all instances of a same person across multiple cameras with non-overlapping views. This definition is simple, however, can be misleading as the reader may consider multiple outlines. In this subsection, we initially show the evolution of the problem and then we break down the constraints in which this work is developed.

The origins of Re-ID are related to the task of tracking people, however, as pointed by Vezzani et al. [58] there are differences between them: tracking considers small variations in position, point of view, appearance, and biometric profile. On the other hand, Re-ID aims to match instances of people during a time delay and/or a change in viewpoint, so that changes in position and point of view are more drastic.

Before describing the Re-ID evolution, we define two important concepts: probe and gallery. If we consider Re-ID as a retrieval task, the probe is analogous to the query (person of interest) and the gallery to the space of search. Probe and gallery can be composed of images (image Re-ID) or videos (video Re-ID).

The evolution of Re-ID problem is associated with the outline in data and label format. From its early years and up to 2017, the problem considered as input bounding boxes that contain mainly one person, whereas labels only contain the ID. Input images, usually in RGB format, were captured from security cameras at universities [25, 46, 70],

supermarkets [69], airports [21], underground stations [39] among other public places.

Due to the task complexity, labels about the attributes of people (e.g., gender, hair length, sleeve, height) have been started to be released in most popular Re-ID datasets [32]. Moreover, other variants of the problem have been proposed; some interesting ones include (i) a scenario in which probe has RGB images and gallery has infrared images [61], (ii) query is based on portrait images and gallery consists of movie scenes [19], (iii) Re-ID for groups of people [27] and (iv) input considers the complete scene instead of only bounding boxes [4].

In this work, we consider the following constraints:

- we address the Re-ID problem using images, this is, probe and gallery consist of images in the RGB format;
- input images correspond to bounding boxes of people and labels contain only their ID;
- our method uses saliency and semantic parsing for enhancing purposes, however, the datasets used for evaluation do not need a further annotation for these tasks;
- input images were captured from security cameras; for this reason, methods based on face recognition are not feasible to use. Then, the proposed method is based on the appearance of people.

2.1.2 Deep Learning

Deep Learning is a subfield of Machine Learning that focuses on models based on Neural Networks, but with large amounts of layers. Each layer learns to represent different levels of representation, hence the term "deep". Although this topic has become very active in recent years, its origins date back to the 1940s under terms such as *cybernetics* [11], however, at that time, it did not become a strong research field due to certain issues, such as computational power and automatic mechanisms for training these types of models.

Many members of the research community consider the year of 2012 as the breaking point for Deep Learning. At that time, the ImagiNet [48] challenge was released, which is a classification task that contains 1000 classes of objects (for instance, cats, dogs, cars) and approximately 1 million images.

Krizhevsky et al. [22] proposed a Convolutional Neural Network (CNN) that achieved the best results in the competition by a large margin. As a result, the research community has made great efforts to apply deep models to other tasks. Currently, Deep Learning is the state of the art in various problems, achieving outstanding results in tasks where the classic "feature-engineering" has had problems to generalize.

On the other hand, this type of models requires large amounts of data. This is a serious issue in areas/problems where the cost of labeling or collecting data is really expensive. In addition, Deep Learning detractors criticize the "architecture engineering" used to create

new models, mainly because it is very difficult to have a true understanding of the meaning of deep model internal parameters and also due to the high computational power used in the trial-and-error process of the network design. However, novel approaches [75] suggest that future deep models will be designed by other representations using reinforcement learning.

Two properties of deep models that are central to our research work are the feature transferability and different levels of representation. Feature transferability allows us to reuse trained networks in related tasks; in the Person Re-Identification problem, a considerable amount of works takes advantage of the strong feature representation learned in ImagiNet [48]. More importantly, the ability of deep models to abstract different levels of representation allows our work to combine different clues to improve results.

In this work, we adopted well-established architectures [6,13,18,54] proposed for image classification tasks to extract features. We will denominate them as backbones.

2.1.3 Saliency Detection

Saliency detection is a task that focuses on identifying the fixation points on which a human viewer would focus at first glance. It has applications in various vision domains, such as image segmentation, object detection, video summarization, compression, just to name a few of them [59].

Two main steps of visual processing are used for human beings to find saliency points: first, a parallel, fast, but simple pre-attentive process; then, a serial, slow, but complex attention process. In the pre-active step, certain features such as orientation, edges or pixel intensities usually attract viewer's attention, which are considered as candidates of an object [17].

Similarly to many computer vision problems, the initial approaches to saliency detection were driven to employ local low-level features – such as intensity, color, orientation and texture – or global features based on finding regions in the image, which implies unique frequencies in the Fourier domain [9]. Then, deep learning methods have become popular in this task.

Zhang et al. [66] indicated that saliency detection methods can be roughly categorized into two groups: bottom-up methods and top-down methods. Bottom-up methods can be categorized into global and local. Global approaches compute saliency scores based on the uniqueness of each image element in the entire image. On the other hand, local methods consider the difference between neighborhoods of image patches, regions, and pixels. Top-down methods usually employ prior knowledge and are usually task-dependent used clues include faces, humans, text, and animals.

In the scenario of Re-ID, there are various cases where salient objects can guide the process of matching people, for instance, consider a person using a red backpack in a white background, then the backpack can be used to differentiate and further re-identify people. In this work, we use saliency on one of the streams of our framework in order

to force the network to focus on salient parts of the image, this is done by masking the feature maps. We used the method proposed by Li et al. [28] to compute the saliency maps from the people bounding boxes. It is a multi-task saliency model based on Fully Convolutional Neural Networks (FCNN), whose approach aims to build a model trained to combine the intrinsic correlations between saliency and semantic image segmentation. Then, a Laplacian regularized nonlinear regression scheme is used to refine the saliency maps. We decided to use this method because of its capability and multi-task nature. Some qualitative results reported by Li et al. [28] are shown in Figure 2.1.

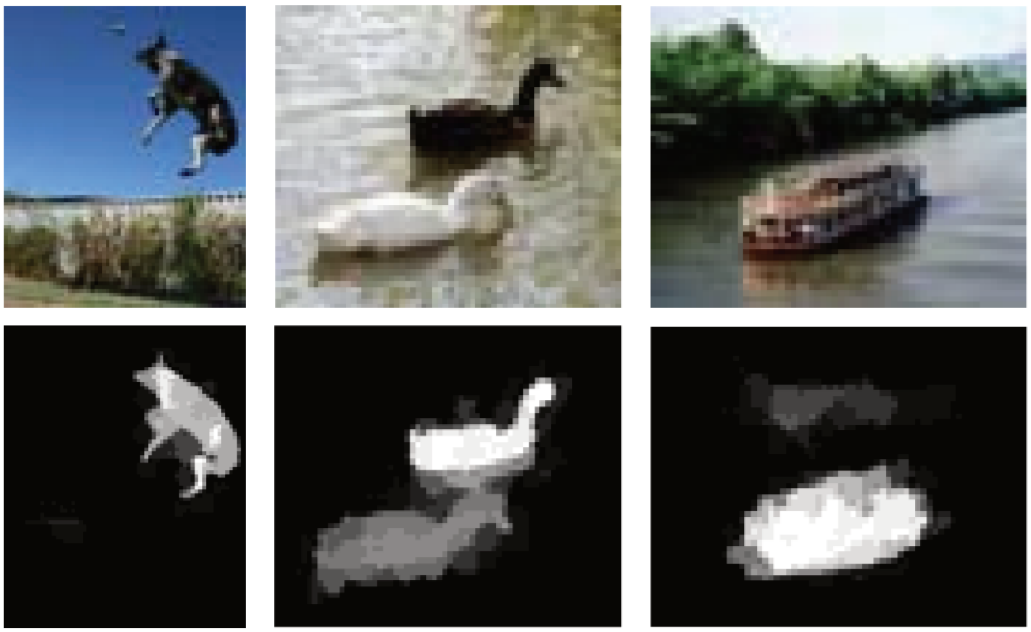


Figure 2.1: Examples of saliency results. The first row shows some input images, whereas the second row shows the corresponding saliency maps. Adapted from [28].

2.1.4 Semantic Parsing

Also known as human semantic segmentation, semantic parsing aims to segment images containing humans into regions with fine-grained meaning, which has applications in Re-ID and human behavior analysis [10]. In its general form, semantic parsing has applications in several other domains, such as image montage, object colorization, stereo scene parsing, indoor navigation, autonomous driving, remote sensing, and medical segmentation [37].

One of the main problems in semantic parsing is to define what makes an object or image part meaningful and this can be ambiguous. This is a complicated process even for humans, since different people have different perception of object grouping (Figure 2.2).

Current literature in semantic parsing can be divided into unsupervised methods, weakly/semi-supervised methods and fully-supervised methods [74]. Unsupervised methods group local pixels based on features such as color or texture and have no explicit object model, where some of the general ideas include graph-based [8], clustering [62],

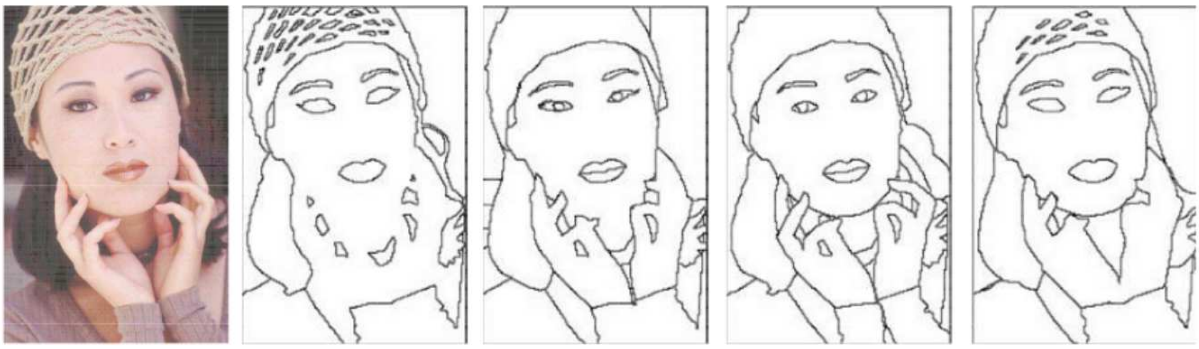


Figure 2.2: Example of an image and hand labels semantic segmentation annotated by four different people. It can be seen that the definition of meaningful areas can be ambiguous. Adapted from [74].

and superpixel approaches [57]. For semi-supervised methods, there are interactive and co-segmentation approaches. Interactive ideas usually consider an initial constraint and apply a repetitive refinement process after the user adds new constraints to each iteration [42]. On the other hand, co-segmentation extracts common objects from a set of images. Fully-supervised methods [1] are the most successful methods, but they are usually limited to specific scenarios (for instance, person semantic parsing) because of the cost of labeled data [74].

In this work, we use the approach proposed by Gong et al. [10] to compute the semantic parsing maps from the images. Gong et al. [10] proposed a large-scale dataset for person semantic parsing, named Look into Person (LIP), composed of representative samples with strong articulation changes, partial occlusions, truncation at image borders, various clothing appearances, diverse viewpoints, and background clutters. We decided to use their method because it uses human pose approximated from parsing annotations to guide human parsing. In addition, it is capable to behave correctly in the complicated scenarios of LIP. This is important since LIP presents similar problems to Re-ID datasets (Figure 2.3).

2.2 Related Work

In this section, we present a review of the most reliable works, based on results quality, in the context of Re-ID problem. Early works focused on handcrafted features, such as color and texture, however, due to extreme viewpoint and illumination variations, these types of characteristics are not sufficiently discriminative [2, 5, 15, 16, 24, 33, 34, 35, 38, 40, 41, 64]. Thus, this literature review is heavily focused on works using deep learning techniques since they present the best results in the current approaches.

Chang et al. [3] proposed Multi-Level Factorization Net (MLFN) that encodes features at multiple semantic-levels. MLFN is composed of various stacked blocks with the same architecture and block selection modules that learn to interpret content of input images.

Figure 2.3: Examples of eight different people from semantic parsing dataset Look into People [10]. It presents various challenges that are common in the Re-ID task, thus, semantic parsing methods trained in this dataset can be applied of-the-shelf in the context of Re-ID.

The insight behind the selection blocks is to control and specialize the features that each block is learning. Moreover, MLFN is tested in the task of object categorization using CIFAR-100 dataset achieving superior results than networks, such as ResNet [13]. Authors of MLFN present an insightful analysis of latent factors learned by MLFN that suggest that shallow blocks learn features such as color and texture, whereas deeper blocks learn clothing style and gender. Our proposed framework employs well-established backbones, so we do not need to do the architecture engineering work done for MLFN blocks. Actually, our proposed framework can use the MLFN as a backbone due to its general definition.

Zhao et al. [65] used GoogLeNet [55] to extract features. Then, a multi-branch architecture uses these features to detect discriminative regions and create a part-aligned representation. From this idea, they were able to overcome misalignment and pose changes. Differently from this work, Su et al. [52] extracted person parts directly from the input images through a pose estimator trained independently. Then, they extracted features from complete images and parts. In the case of local clues, their architecture considers affine transformations. Finally, because pose estimation may be affected by pose changes or occlusions, they combined parts and global features using a weighted sub-net. Our proposed framework lies more in the line of Su et al. [52] than Zhao et al. [65] because multiple clues are extracted from raw images and then introduced in the network for results improvement.

Li et al. [26] proposed Harmonious Attention Network (HA-CNN), which focuses on learning fine-grained relevant pixels and coarse latent regions at the same time. HA-CNN is based on Inception [55] blocks in a multi-branch structure for global and local representation. They further introduced a method for combining these representations in a harmonious way, this is, a combination of hard regional, soft spacial, and channel attention. Compared to the proposed framework, we use a similar idea using attention determined by semantic parsing and saliency. An advantage of HA-CNN is that the attention is computed inside the network, not requiring external clues. On the other hand, our proposed framework is more scalable as other clues or types of attention can be easily added.

Kalayeh et al. [20] demonstrated that the use of semantic parsing can boost up results in Re-ID. They proposed to use an Inception-based network [55] that computes semantic maps and generates features for global representation of the input. Then, the feature map before the last average pooling is multiplied by the parsing maps to create a local representation. Our proposed framework further includes saliency and, differently from Kalayeh et al. [20], our training process does not require a collection of ten Re-ID datasets and uses less than two hundred epochs.

Zhao et al. [67] pioneered the concept of saliency in the context of Re-ID. Their approach is based on a patch matching-based method. Each image patch has an associated saliency that is computed in an unsupervised fashion, then the matching is computed inside the patch-neighborhood using hand-crafted features. A matching between patches with too different saliency brings a penalty to the model. Thus, the model is fitted to minimize the total cost of patch matching. This work follows more classical machine learning pipelines, so there is a large difference between it and our proposal. In our work, we use saliency to guide network for better feature learning.

Liu et al. [36] proposed an attentive-based method, named HydraPlus-Net. Although the authors do not use the concept of saliency, the idea is related because they guide their network to focus more on specific regions of the image. HydraPlus-Net is designed to use Inception [55] blocks for its construction. This network design follows a standard combination of local and global representation, however, in addition, the authors take into account different scales. This is important because difficult cases for Re-ID include differences in small details such as shoes or hairstyle.

Similar to HydraPlus-Net, Zhou et al. [72] proposed to learn saliency maps and Re-ID at the same time. They introduced a weighted version of bilinear coding [31] to encode higher-order channel-wise interactions. This idea is interesting since previous works generally considered global average pooling (GAP) to combine intermediate multi-branch features, as pointed by the authors. GAP generates only a linear combination of features that may not be good enough to represent all the complex variations in the data.

Qian et al. [44] proposed a network that learns saliency from their Re-ID pipeline. They accurately pointed out that features at different scales are not a well-solved problem for Re-ID. Their proposal, named MuDeep Net, is a network capable of learning

features at different scales and creating saliency masks to emphasize channels with highly discriminative features.

Differently from Liu et al. [36], Zhou et al. [72] and Qian et al. [44], we use an off-the-shelf Fully Convolutional Neural Network (FCNN) in order to get saliency from raw input. This is helpful since other multiple clues (e.g., semantic parsing) can also be introduced and combined for enhancing results. Moreover, this is more intuitive because computing saliency and Re-ID features at the same time is generally based on sigmoid layers without specific loss to optimize the saliency map generation. Thus, the saliency maps generated with these methods do not have a clear meaning when overlapping over input images.

To compensate for viewpoint changes, Sarfraz et al. [49] proposed to create a pose-discriminative embedding. They trained a network that learns specialized features depending on the pose of the input: front, back or side. They also used body joint keypoints to guide the CNN's attention. In addition, they proposed a new re-ranking (RR) technique, named Expanded Cross Neighborhood. Results were improved based on the distance between gallery and probe features.

Zhong et al. [71] proposed an RR method based on k-reciprocal nearest neighbors and Jaccard distance. It is worth mentioning that RR methods are unsupervised and do not need any human interaction. In our work, we use RR to improve the final results, so that the comparison with the state of the art is conducted with and without RR for fair validation.

Chapter 3

Saliency and Semantic Parsing Guided Re-ID

In this chapter, we present our novel methodology for Re-ID in images [45]. It is named Saliency-Semantic Parsing Re-Identification (SSP-ReID), for taking advantage of the capabilities of both clues, saliency, and semantic parsing maps, to guide a backbone convolutional neural network (CNN) to learn complementary representations that improve the results over the original backbones.

The insight of fusing multiple clues is based on specific scenarios, where one response is better than another, thus favoring the combination of them to increase performance. Due to its definition, our framework can be easily applied to a wide variety of networks and, in contrast to other competitive methods, our training process follows simple and standard protocols. Our approach is able to achieve competitive results in all evaluated datasets: Market1501 (second/third best), CUHK (first/second best) and DukeMTMC-ReID (second/third best).

3.1 Problem Formulation

For the scenario of image Re-ID, we consider it as a retrieval process, that is, given an image of a query person x_p with ID y_p and a gallery of m people $X = \{x_1, x_2, \ldots, x_m\}$ with IDs $Y = \{y_1, y_2, \ldots, y_m\}$, then Re-ID aims to retrieve all $x_i, (1 \leq i \leq m)$ such that $y_i = y_p$.

Suppose that a model $M(\theta)$ with learned parameters θ is capable of representing x_p and people in X with feature maps f_p and $F = \{f_1, f_2, \ldots, f_m\}$, respectively. Thus, we can use Euclidean distance to compare f_p against each element of F and construct a ranked list based on the similarity of the feature maps. Depending on the application and context in which Re-ID is used, this ranked list may be cut off in the top 1, 5 or more. The resulting list L (also referred to as ranked list) is employed to represent only people that have identity equal to y_p[1].

[1] L may not be totally correct, since $M(\theta)$ may not be perfect.

In this approach, we create a model M' that uses M as backbone, such that the list L' generated by M' is superior than L. The qualitative definition of *superior* here is based on the mean Average Precision (mAP) and cumulated matching characteristics (CMC). Both metrics are explained in the experiments chapter. Due to the definition of $M(\theta)$, our framework can be applied to many different backbones.

3.2 Person Re-ID Framework

Based on the fact that a challenging issue for the re-identification task is caused by dramatic pose/viewpoint changes, we propose to combine global representation with saliency and semantic parsing masks. As shown in our experiments, these two types of masks generate complementary feature maps that improve results over the original CNN backbones. Saliency is important for Re-ID because, in specific scenarios (Figure 3.1) where people have certain items, it can guide the re-identification process.

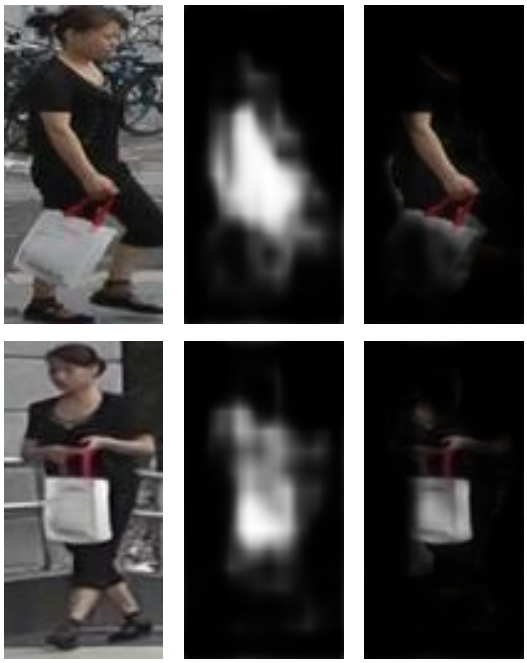


Figure 3.1: Examples of saliency detection for the same person with two different views (from left to right): original image, saliency map created using proposed method by Li et al. [28], and result of overlap saliency map over the original image. The focus of the saliency is on the arm and white bag. Our framework uses this information to guide the feature learning process.

However, saliency is not a complete solution to the problem because it focuses on some areas of the image and may be affected by occlusions. Thus, we use semantic parsing to encode every part of the person and overcome misalignment in the bounding box detection and occlusions (Figure 3.2).

We propose the Saliency-Semantic Parsing (SSP-ReID) framework, as shown in Figure 3.3, which is composed of two streams. Both of them have the same backbone architec-

Figure 3.2: Examples of parsing with five semantic regions of the same person with two different views. We use these maps to overcome misalignment and occlusions.

ture, however, without sharing weights. One of the streams (named S-ReID subnetwork) focuses on getting global-saliency features, whereas the other (named SP-ReID subnetwork) focuses on getting global-semantic-parsing features. The output of our framework is a feature map that is used to compare query and gallery images.

Given the input image, we feed forward it to two CNN backbones, with same architecture but without shared weights, and apply average pooling to create global representations of the person. Moreover, we compute the saliency and semantic parsing maps from the input using off-the-shelf deep methods [10,28]. For the semantic parsing, we follow the work of Kalayeh et al. [20] and consider 5 semantic areas: head, upper body, lower body, shoes and complete body. Then, we take the feature map generated by the feed forward pass from an intermediate layer[2] of the CNN backbones and join it with saliency map in one stream, and semantic parsing maps in the other. As a result, we have S feature vector from the saliency stream and SP feature vector from the semantic parsing stream. Finally, we append the global representation, S and SP feature vectors to create SSP feature vector that encodes reliable information of the input image. Later, SSP vector will be used to compare images for the Re-ID task.

In order to include saliency/semantic parsing maps in our pipeline, the following process is executed: given the output of an intermediate layer, also denominated tensor, τ with height h, width w and c channels, $\tau \in \mathbb{R}^{h \times w \times c}$, and a saliency/semantic parsing map ω with height h' and width w', $\omega \in \mathbb{R}^{h' \times w'}$, in order to join intermediate feature tensor and saliency/semantic parsing information, we initially apply a bilinear interpolation

[2]We use the term *intermediate layer* to refer to the output of a layer that is located near the middle of the CNN. We expect the size of the output of this layer be 1/4 of the input (height and width).

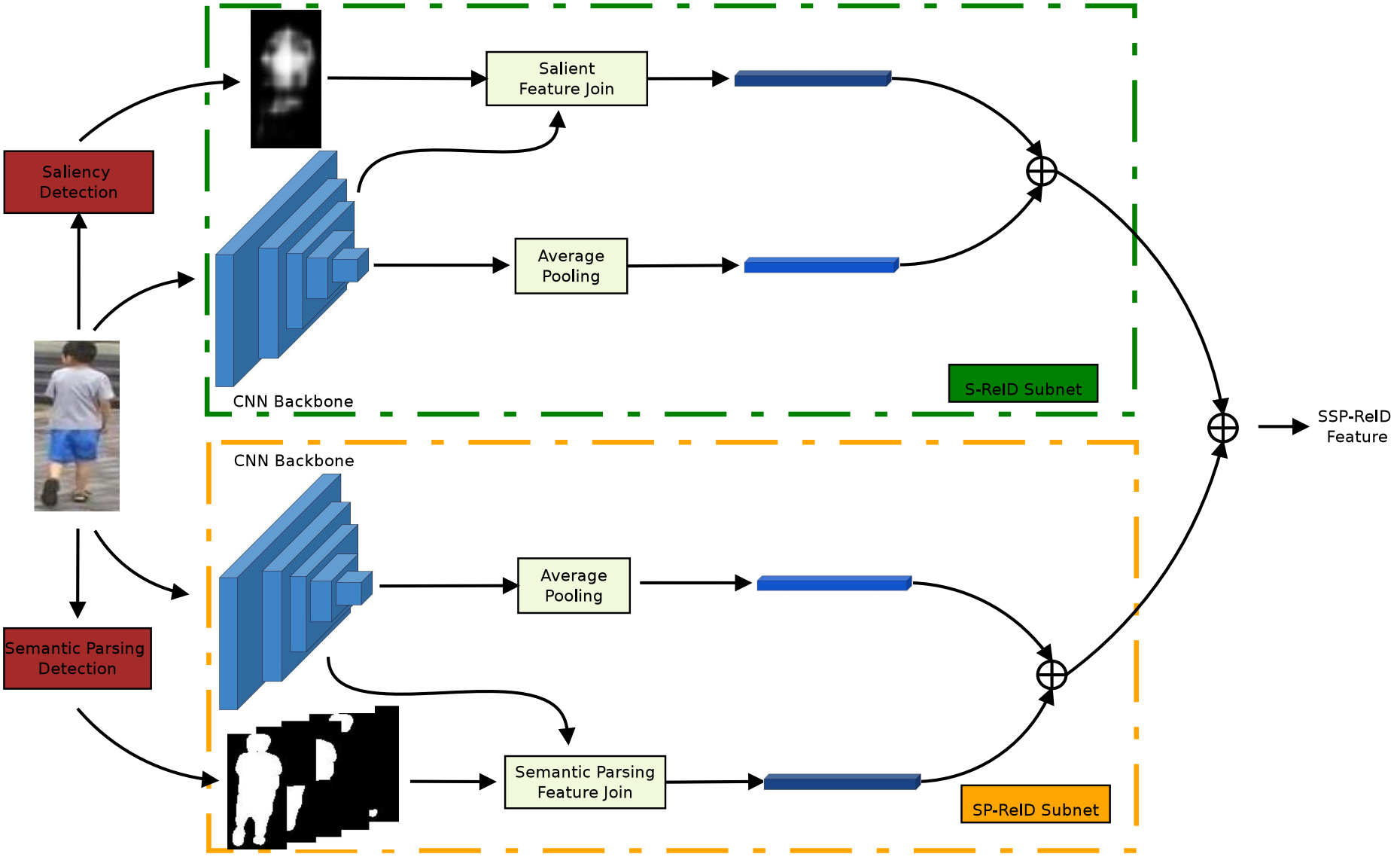


Figure 3.3: SSP-ReID is a framework based on semantic parsing (SP-ReID subnet) and saliency (S-ReID subnet) to learn individual-similar performance representations. At the same time both representations are complementary because the union leads to an increase in performance.

over the tensor to transform $\tau \in \mathbb{R}^{h' \times w' \times c}$. Then, we apply element-wise product between every channel of the tensor and the map. Finally, we use average pooling to obtain the feature vector v. For the saliency feature join, the output feature is inside $\mathbb{R}^c$, whereas for semantic parsing feature join is inside $\mathbb{R}^{5c}$ due to the 5 semantic regions considered.

We use an intermediate layer since it is widely known that CNNs encode more abstract and higher semantic-level features as they go deeper (for instance, the relationship of head location between the input image and the very deep feature map may not be clear because the receptive fields grow). Thus, it is more intuitive to combine raw saliency/semantic parsing maps with an intermediate layer as it does not have too abstract information and, at the same time, it encodes rich information.

In order to train our network, we consider cross-entropy loss function with label smoothing regularizer (LSR) [56] and triplet loss with hard positive-negative mining [14]. Cross-entropy with LSR is defined as:

$$H(q', p) = -\sum_{k=1}^{K} \log p(k) q'(k) = (1 - \epsilon) H(q, p) + \epsilon H(u, p) \tag{3.1}$$

where K is the size of training batch, ϵ is a regularizer value, $p(k)$ is the output of the model, q is the ground-truth distribution, u is the uniform distribution and q' is defined as:

$$q'(k) = (1 - \epsilon) q(k) + \frac{\epsilon}{K} \tag{3.2}$$

LSR is a change in the ground-truth labels distribution, which aims to make the model more adaptable by adding prior distribution over the labels. We consider this loss over general cross entropy in order to avoid the largest logit from becoming much larger than all others, this prevents overfit.

Triplet loss with hard positive-negative mining is defined as:

$$T(X) = \sum_{i=1}^{P} \sum_{a=1}^{N} [m + \max_{p=1 \ldots N} D(f(x_a^i), f(x_p^i)) - \min_{\substack{p=1 \ldots N \\ n=1 \ldots N \\ i \neq j}} D(f(x_a^i), f(x_n^j))]_+ \tag{3.3}$$

where X is a training batch, with P people and N images per person, $f(\ .\ )$ is the output feature map of the network, x_j^i is the j-th image of the i-th person, $D(\ .\ ,\ .\ )$ is a distance function (e.g., Euclidean), $[\sigma]_+$ denotes $\max(\sigma\ ,\ 0)$ and m is hyperparameter named margin. Basically, this loss finds the pair of images of the same person with maximum distance and the pair of images of different people with minimum distance and guides the model to make the difference between these two at least equal to the margin m.

SP-ReID and S-ReID subnetworks are trained separately and, depending on the backbone, we use the sum of both losses or only cross-entropy with LSR. To train our network, we add a multi-class classification layer to the end of the subnetwork. Figure 3.4 illustrates the network architecture for the training step, as well as its relation to the loss

functions.

3.3 Computational Resources

The proposed method was implemented in Python programming language due to the large availability of well-documented libraries for data manipulation, numerical computation, visualization, image processing, and deep learning, such as SciPy, NumPy, Matplotlib, and Pytorch.

The experiments were performed on computers available in the Laboratory of Visual Informatics (LIV) of the Institute of Computing (IC) at University of Campinas (UNICAMP). The machines are equipped with Linux operating system, Intel i7-3770, 3.50 GHz processor, 32GB of RAM and an NVidia GeForce GTX 1080ti video card, with 2880 CUDA cores, 11178 MB DDR5 standard memory and 7.0 Gbps clock.

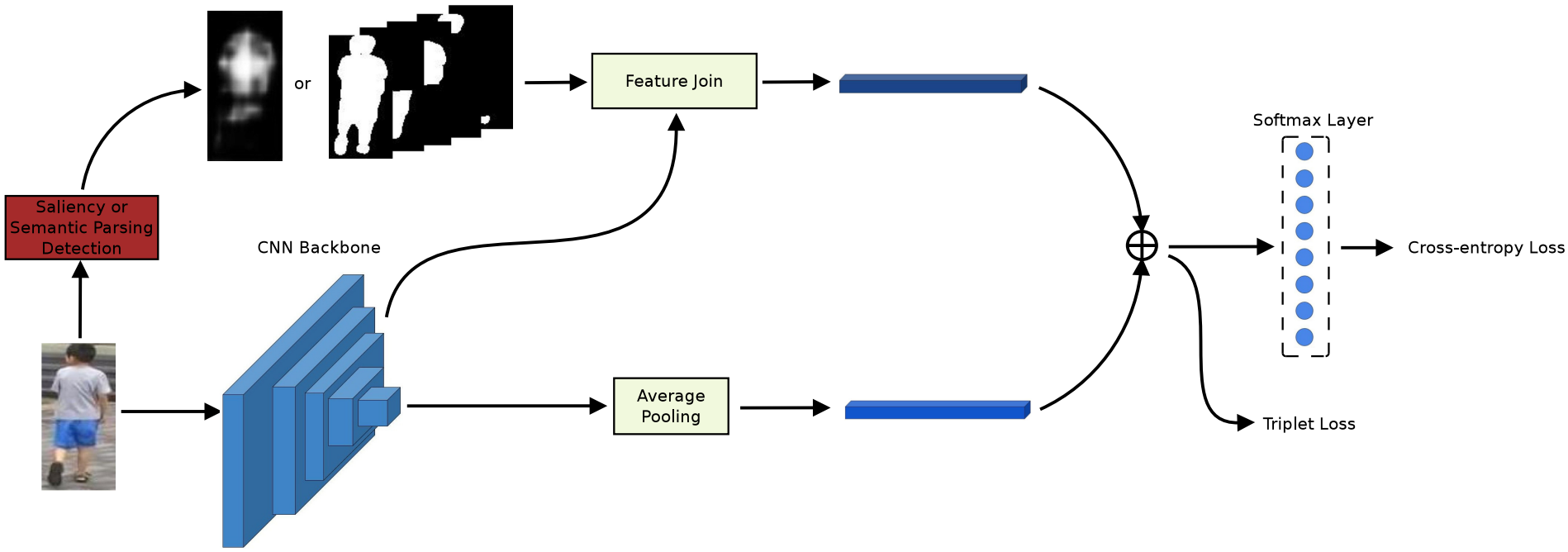


Figure 3.4: Training setup for S-ReID and SP-ReID subnetworks. When training out framework, we consider triplet and cross-entropy loss functions. For the triplet loss, we take the feature vector before softmax layer and use it to compare images based on the Euclidean distance. The triplet loss may be ignored depending on the CNN backbone.

Chapter 4

Experiments

In this chapter, we present the validation protocols and three datasets used in the experiments. Then implementation details for exploring the backbones are explained. Finally, extensive experiments conducted to validate every network module of the proposed framework are presented.

4.1 Validation Protocols

Quantitative results for every dataset are based on mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC). The mAP considers the order in which the gallery is sorted for a given query, defined as:

$$\text{mAP} = \frac{\text{AP}}{\#queries} \tag{4.1}$$

where Average Precision (AP) is defined as:

$$\text{AP} = \frac{\sum_{k=1}^{n} \text{P}(k) \times rel(k)}{\#relevant\ items} \tag{4.2}$$

where n is the number of retrieved items, $rel(k)$ is equal to 1 if the k-th item is relevant to the query and 0 otherwise, and $\text{P}(k)$ is defined as:

$$\text{P}(k) = \frac{\sum_{i=1}^{k} rel(i)}{k} \tag{4.3}$$

The CMC represents the probability that a correct match with the query identity will appear in variable-sized ranked list:

$$\text{CMC}(r) = \frac{in(r)}{\#queries} \tag{4.4}$$

where $in(r)$ is the number of queries that have a relevant element within the first r items in the ranked list. We consider $r = 1, 5, 10, 20$ and refer to them as rank-1, rank-5, rank-10 and rank-20, respectively.

4.2 Image Re-ID Datasets

We evaluate our framework on three widely used datasets. A summary of them is shown in Table 4.1, which reports the number of people, bounding boxes and cameras present in each benchmark setup.

Table 4.1: Comparative summary of Image Re-ID datasets used in our experiments.

Dataset	# **People**	# **BBox**	# **Cameras**
Market1501 [69]	1501	32668	6
CUHK03 [25]	1467	14096	6
DukeMTMC-ReID [70]	1812	36411	8

The Market1501 dataset [69] was created through pedestrian detection using Deformable Part Model (DPM) in order to simulate a real-world scenario. Data was captured from the supermarket in front of the Tsinghua University using 6 cameras, including 5 high-resolution and 1 low-resolution. For validation, we use the fixed splits provided with the dataset. There are 19732 images for testing and 12936 for training.

The CUHK03 dataset [25] has an average of 4.8 images per view and was created in the Chinese University of Hong Kong. Misalignment, occlusions and missing body parts are very common. For validation, we use the new validation protocol [71] with partition of 767/700, which means that images of 767 people are used for training and images of 700 people are used for testing. Moreover, we evaluate detected (CUHK03 (D)) and labeled (CUHK03 (L)) versions of the dataset, for CUHK03 (D) the bounding boxes were automatically detected using the pedestrian detection method of Felzenszwalb et al. [7] and for CUHK03 (L) the bounding boxes were manually labeled.

The DukeMTMC-ReID dataset [70] is a subset of the DukeMTMC dataset [46] for image-based re-identification with hand-drawn bounding boxes. Original dataset has 75 minutes of high resolution video from 8 cameras inside the campus of Duke University and has an average of 20 images per ID. Bounding boxes of different sizes with outdoor scenes as background are their main characteristics. This dataset was designed to be compatible with Market1501, that is, validation has fixed training and testing sets proposed in the original protocol, in a total of 408 distractor identities, 16522 training images of 702 identities, 2228 query images of the other 702 identities and 17661 gallery images.

4.3 Implementation Details

The saliency detection is performed via the off-the-shelf FCNN proposed by Li et al. [28], whereas the semantic parsing detection is computed through the Joint Human Parsing and Pose Estimation Network (JJPNet) [10] trained in the Look into Person (LIP) dataset [29].

We evaluate our framework using 5 different backbones. Table 4.2 summarizes the loss function and intermediate layer used for each backbone. The initial weights of all backbones are ImagiNet [48] pretrained models.

We decide to use the intermediate layers because their outputs are almost in the middle of the backbones, thus they encode enough abstract information (e.g., they are deep). At the same time, when applying bilinear interpolation to combine them with saliency/semantic parsing maps, it is not necessary to interpolate too many values: relation between saliency/semantic parsing maps and layer outputs is about 1/4. Later, we will show that choosing an intermediate layer may not have a huge impact in the final results as CNN backbones are capable of readjusting their weights according to the layer choice.

Table 4.2: Loss function and intermediate layer used for each backbone. CROSSE stands for Cross Entropy Loss with LSR [56], whereas TRIP stands for Triplet Loss with hard positive-negative mining [14]. Using TRIP in Inception-V4 [54] and Xception [6] raises exploding gradient.

Backbone	**Loss Function**	**Intermediate Layer**
ResNet50[1]	TRIP + CROSSE	Res4F
Densenet[2]	TRIP + CROSSE	second composite function
Resnet50-M[3]	TRIP + CROSSE	Res4F
Inception-V4[4]	CROSSE	last Inception-B block
Xception[5]	CROSSE	Middle Flow

We adjust the size of the input to 254×128 pixels and saliency/semantic parsing maps to 128×64 pixels. Adam optimizer is used with training batch of 32, initial learning rate of 0.0003, weight decay of 0.0005, and a learning rate decay factor of 0.1 every 60 epochs.

We also fix the number of training epochs to 180 for every backbone. In the LSR implementation, we set $\epsilon = 0.1$ and for triplet loss $m = 0.3$.

Finally, we use the re-ranking (RR) method proposed by Zhong et al. [71] to boost up results. Comparison is made considering the state-of-the-art approaches with and without RR [6].

[6]Code and trained models are available at `https://github.com/RQuispeC/saliency-semantic-parsing-reid`

4.4 Results

In this section, we present extensive experiments to validate each module of the framework on three datasets. We also compare our results against available methods of the literature.

4.4.1 Ablation Study of Saliency and Semantic Parsing Impact

In this section, we evaluate and compare different aspects of our Image Re-ID framework: backbone (e.g., ResNet [13]), saliency subnet (S-ReID), semantic parsing subnet (SP-ReID) and the complete framework (SSP-ReID). Backbone results are generated using the features of the CNN after average pooling without using any other clue.

Results are summarized in Table 4.3. Overall, ResNet50-M + SSP-ReID produced the best results for all datasets, whereas there are marginal differences in using ResNet50 [13] and DenseNet [18] as backbones. On the other hand, Inception-V4 [54] and Xception [6] yielded the worse performance.

It is worth observing that DenseNet [18] achieves interesting results despite of its lower number of parameters. In addition, our framework raises consistently an improvement over all backbones, which suggests that our framework can be used as an enhancing method for image Re-ID. Moreover, due to its definition, it can be applied to other clues (e.g., person pose) and other tasks (e.g., combining relevant clues for Re-ID).

In the light of all datasets, S-ReID achieved improvements up to 2.5% for mAP and up to 4.4% for rank-1 over individual backbones, however, in general the improvements were marginal. There are also cases where results were marginally worse.

This same scenario is repeated for SP-ReID, but if we consider the complete framework (combination of S-ReID and SP-ReID), we consistently obtained better results, with improvements up to 7.4% (mAP) and 7.2% (rank-1) in Market1501, 6.8%(mAP) and 7.7%(rank-1) in CUHK03 (D), 4.9%(mAP) and 4.5%(rank-1) in CUHK03 (L), 6.7%(mAP) and 8.5% (rank-1) for DukeMTMC-reID.

Results of CMC for ResNet50-M for all datasets are shown in Figure 4.1. It can be seen that using SSP-ReID enhances model capability for every rank value. For the Market1501 [69], the improvement is lower since its backbone performance values are already really high (that is, above 0.9). For the other cases, the gain is higher. This same effect takes place as rank value grows.

These results suggest that S-ReID and SP-ReID are learning representations with similar performance (e.g., close qualitative results), but with complementary information, which improves the model capacity when they are combined.

It can be observed that the method improvement is inversely proportional to the capacity of the backbone. For better backbones (for instance, ResNet50-M [63]), the improvements are smaller when compared to the lower performance backbones (for instance, Xception [6]). This is related to the complexity of datasets: as we start achieving high results in mAP or rank-1, we need much more higher discriminative models that can deal

Table 4.3: Results of framework in Re-ID. ResNet + S-Reid stands for Saliency subnet using ResNet as backbone. Analogously, SP-ReID refers to Semantic Parsing subnet, whereas SSP-ReID refers to the complete framework. We highlight in red color the cases in which the subnetwork/framework is worse than the original backbone, whereas cases with better results than backbone are highlighted in blue color. ResNet [13] means that we used ResNet and append an average pooling to generate feature vectors. Analogously for other backbones.

	Market1501	**CUHK03 (D)**	**CUHK03 (L)**	**DukeMTMC-reID**
Method	**mAP / rank-1(%)**	**mAP / rank-1(%)**	**mAP / rank-1(%)**	**mAP / rank-1(%)**
ResNet [13]	72.9 / 88.1	52.9 / 55.6	56.7 / 58.8	62.1 / 77.7
ResNet + S-ReID	73.0 / 87.6	53.4 /56.0	54.4 /55.9	63.1 / 78.9
ResNet + SP-ReID	72.4 / 87.8	53.5 /56.2	55.5 / 57.3	62.7 / 78.0
ResNet + SSP-ReID	75.9 / 89.3	57.1 / 59.4	58.9 / 60.6	66.1 / 80.1
ResNet-M [63]	77.5 / 91.2	56.3 / 58.7	58.9 / 61.1	63.5 / 78.8
ResNet-M + S-ReID	77.6 / 91.2	56.7 / 59.4	59.7 / 62.1	65.2 / 80.6
ResNet-M + SP-ReID	76.6 / 90.9	57.3 / 59.9	59.7 / 61.4	64.9 / 79.6
ResNet-M + SSP-ReID	80.1 / 92.5	60.5 / 63.1	63.3 / 65.6	68.6 / 81.8
DenseNet [18]	72.0 / 89.3	42.2 / 44.1	45.6 / 47.4	62.5 / 79.7
DenseNet + S-ReID	72.3 / 89.7	43.1 / 44.9	44.3 / 46.7	62.6 / 80.3
DenseNet + SP-ReID	72.9 / 89.6	43.3 / 44.6	44.3 / 44.9	62.9 / 79.8
DenseNet + SSP-ReID	76.7 / 90.9	48.1 / 48.1	49.5 / 49.1	67.1 / 82.2
Inception-V4 [54]	64.0 / 81.9	38.7 / 38.7	40.7 / 42.4	49.6 / 71.9
Inception-V4 + S-ReID	62.8 / 81.4	41.2 / 43.1	42.5 / 42.4	49.1 / 70.6
Inception-V4 + SP-ReID	62.1 / 80.6	34.7 / 35.6	36.1 / 37.4	49.0 / 70.6
Inception-V4 + SSP-ReID	67.7 / 85.4	45.5 / 46.4	45.5 / 45.2	55.0 / 75.5
Xception [6]	50.1 / 69.9	26.0 / 26.3	25.2 / 25.2	36.1 / 55.4
Xception + S-ReID	49.8 / 68.2	23.7 / 23.4	24.2 / 24.1	33.6 / 52.9
Xception + SP-ReID	47.5 / 70.9	20.7 / 22.9	20.8 / 21.6	34.6 / 56.6
Xception + SSP-ReID	57.5 / 77.1	29.4 / 29.9	30.0 / 29.6	42.8 / 63.9

with more specific and often sparse cases.

4.4.2 Are Saliency and Semantic Parsing Meaningful for Re-ID?

Furthermore, we evaluate that combining saliency/semantic parsing maps with intermediate layers, in fact, guides the network for learning meaningful latent characteristics. To do so, we compare it against networks versions (named "+ intermediate") that, instead of using saliency/semantic parsing maps to force better intermediate layers learning, use the raw intermediate layer after average pooling activations. In this case, the "intermediate" version has only one stream.

The results shown in Table 4.4 demonstrate that using saliency and semantic parsing maps guides the network to learn more reliable features that truly represent people.

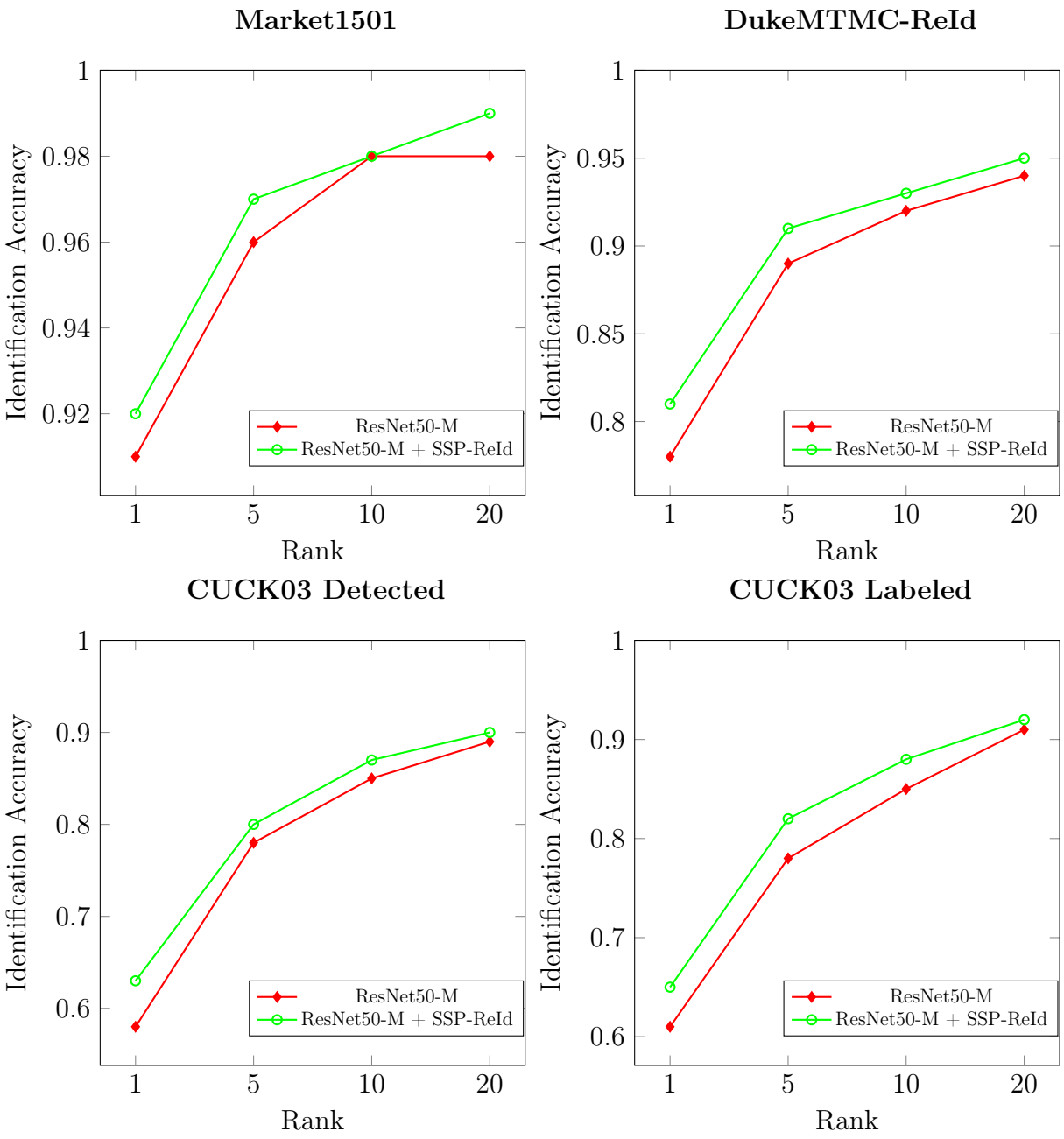


Figure 4.1: CMC curves for ResNet50-M.

4.4.3 Choosing an Intermediate Layer

As can be seen in Table 4.3, the best backbone for SSP-ReId is the ResNet50-M [63]. Thus, we validate how the choice of intermediate layer influences the results for Resnet50-M. Four intermediate layers are evaluated for fusion with saliency/semantic parsing maps: CONV1, RES2C, RES3D, RES4F. The position of these layers varies from very shallow, CONV1 is the first convolutional layer at the beginning of the network, to very deep, RES4F is the last residual layer at the end of the network. Results are shown in Table 4.5.

It can be seen that using RES4F is slightly better than other layers, but the improvement is not really significant (standard deviation between 0.216 and 1.108). Thus, it is

Table 4.4: Comparison of our framework ("+ SSP-ReID") against networks using intermediate layer, but without saliency/semantic parsing map (" + intermediate").

	Market1501	**CUHK03 (D)**	**CUHK03 (L)**	**DukeMTMC-reID**
Method	**mAP / rank-1(%)**	**mAP / rank-1(%)**	**mAP / rank-1(%)**	**mAP / rank-1(%)**
ResNet + intermediate	72.5 / 87.5	52.7 / 54.6	56.0 / 58.6	63.3 / 78.0
ResNet + SSP-ReID	**75.9 / 89.3**	**57.1 / 59.4**	**58.9 / 60.6**	**66.1 / 80.1**
ResNet-M + intermediate	77.2 / 90.1	57.0 / 60.1	58.9 / 60.6	65.3 / 81.2
ResNet-M + SSP-ReID	**80.1 / 92.5**	**60.5 / 63.1**	**63.3 / 65.6**	**68.6 / 81.8**
DenseNet + intermediate	72.8 / 89.6	42.0 / 43.5	43.2 / 45.1	62.6 / 79.3
DenseNet + SSP-ReID	**76.7 / 90.9**	**48.1 / 48.1**	**49.5 / 49.1**	**67.1 / 82.2**
Inception-V4+intermediate	65.3 / 83.7	33.0 / 33.4	31.7 / 31.9	49.1 / 71.1
Inception-V4+SSP-ReID	**67.7 / 85.4**	**45.5 / 46.4**	**45.5 / 45.2**	**55.0 / 75.5**
Xception + intermediate	50.3 / 69.8	21.9 / 22.4	24.1 / 23.8	38.7 / 59.8
Xception + SSP-ReID	**57.5 / 77.1**	**29.4 / 29.9**	**30.0 / 29.6**	**42.8 / 63.9**

Table 4.5: Comparison of intermediate layers choice for Resnet50-M.

	Market1501	**CUHK03 (D)**	**CUHK03 (L)**	**DukeMTMC-reID**
Method	**mAP / rank-1(%)**	**mAP / rank-1(%)**	**mAP / rank-1(%)**	**mAP / rank-1(%)**
ResNet-M - CONV1	80.2 / 92.0	60.0 / 62.3	62.8 / 64.9	68.7 / 82.7
ResNet-M - RES2C	80.3 / **92.6**	**60.5** / 62.2	62.5 / 63.0	68.8 / **83.8**
ResNet-M - RES3D	**80.6** / 92.5	60.2 / 61.6	62.6 / 64.2	**68.9** / 82.8
ResNet-M - RES4F	80.1 / 92.5	**60.5 / 63.1**	**63.3 / 65.6**	68.6 / 81.8
Average	80.25 / 92.5	60.35 / 62.25	62.7 / 64.55	68.75 / 82.75
Standard Deviation	0.216 / 0.270	0.244 / 0.616	0.355 / 1.108	0.129 / 0.818

considered that intermediate layer choice does not have a large impact because, at each configuration, the network is capable of learning new latent representations that handle layer choice.

4.4.4 Qualitative Results

For qualitative results, we analyze various examples. Figure 4.2 shows rank-2 results for DukeMTMC-ReID [70] in which SSP-ReID is capable of taking correct predictions of both, SP-ReID and S-ReID, to reach correct rank-2 predictions.

Other similar examples of the capability of SSP-ReiD are shown in Figure 4.3. In this case, the proposed framework is capable of ranking correct matches of SP-ReID and S-ReID in the top-3 results.

Figure 4.4 has an example that allows us to understand better how SSP-ReID is combining SP-ReID and S-ReID. In this case, SP-ReID and S-ReID have same correct second results, so SSP-ReID considers it as a good match and rank it as its first prediction. If we consider following predictions of SP-ReID and S-ReID, we can notice that next same prediction is the fifth for S-ReID and the third for SP-ReID. Thus, SSP-ReID considers

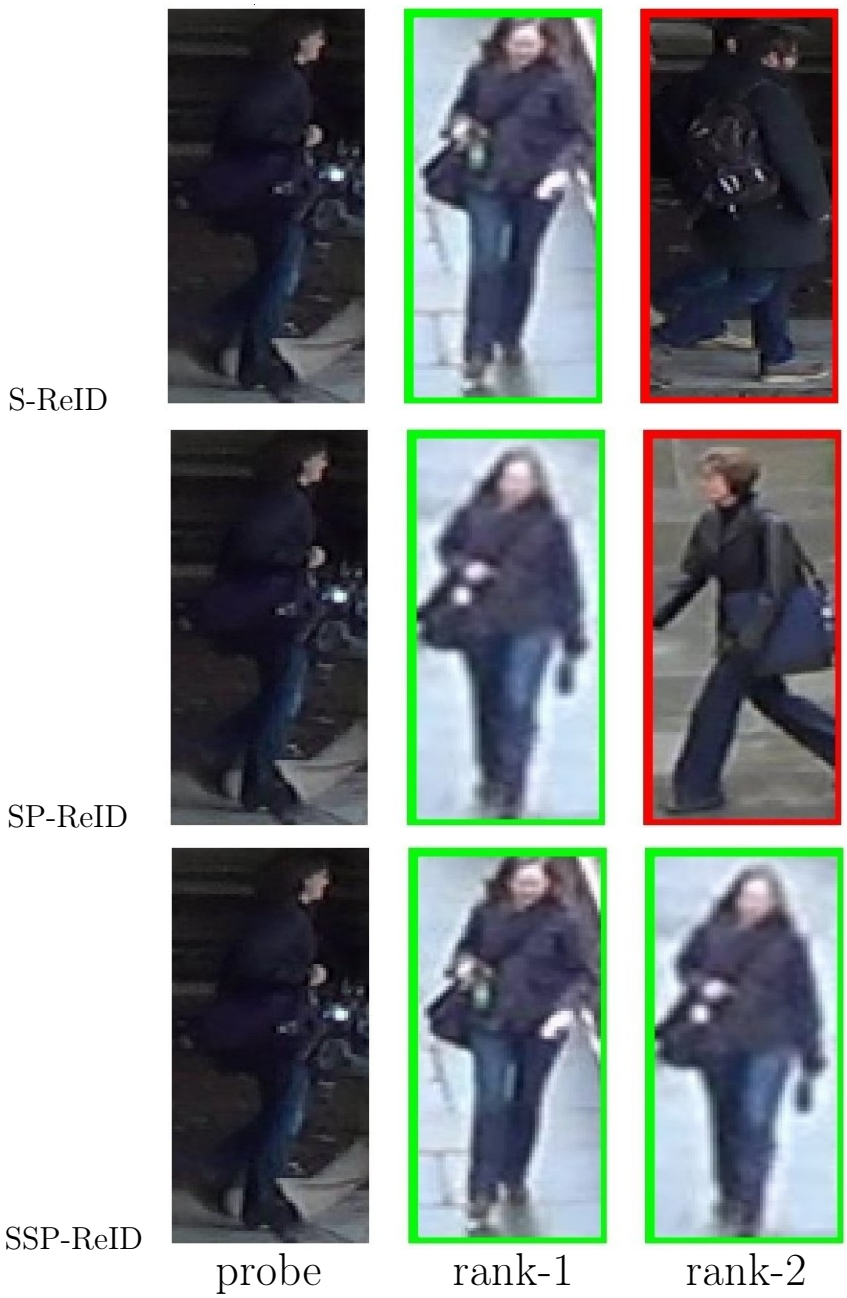


Figure 4.2: For the same query, S-ReID and SP-ReID achieved different correct rank-1 results and wrong rank-2 results. On the other hand, SSP-ReID can combine the best characteristics of both subnets and reach correct rank-2. From dataset Market1501 [69].

it as second prediction regardless if it is not correct.

A similar pattern can be observed in the previous examples (Figure 4.3), where SSP-ReID considers images that look similar in both SP-ReID and S-ReID: a woman with white t-shirt, black backpack in a bicycle and back-view. Then, because they are well ranked in SP-ReID and S-ReID, they end up being well ranked in SSP-ReID as well. This suggests that SSP-ReID combines results of SP-ReID considering similar appearance of candidates and their rank in SP-ReID and S-ReID.

Finally, we show a really difficult example with heavy occlusions. Figure 4.5 shows the

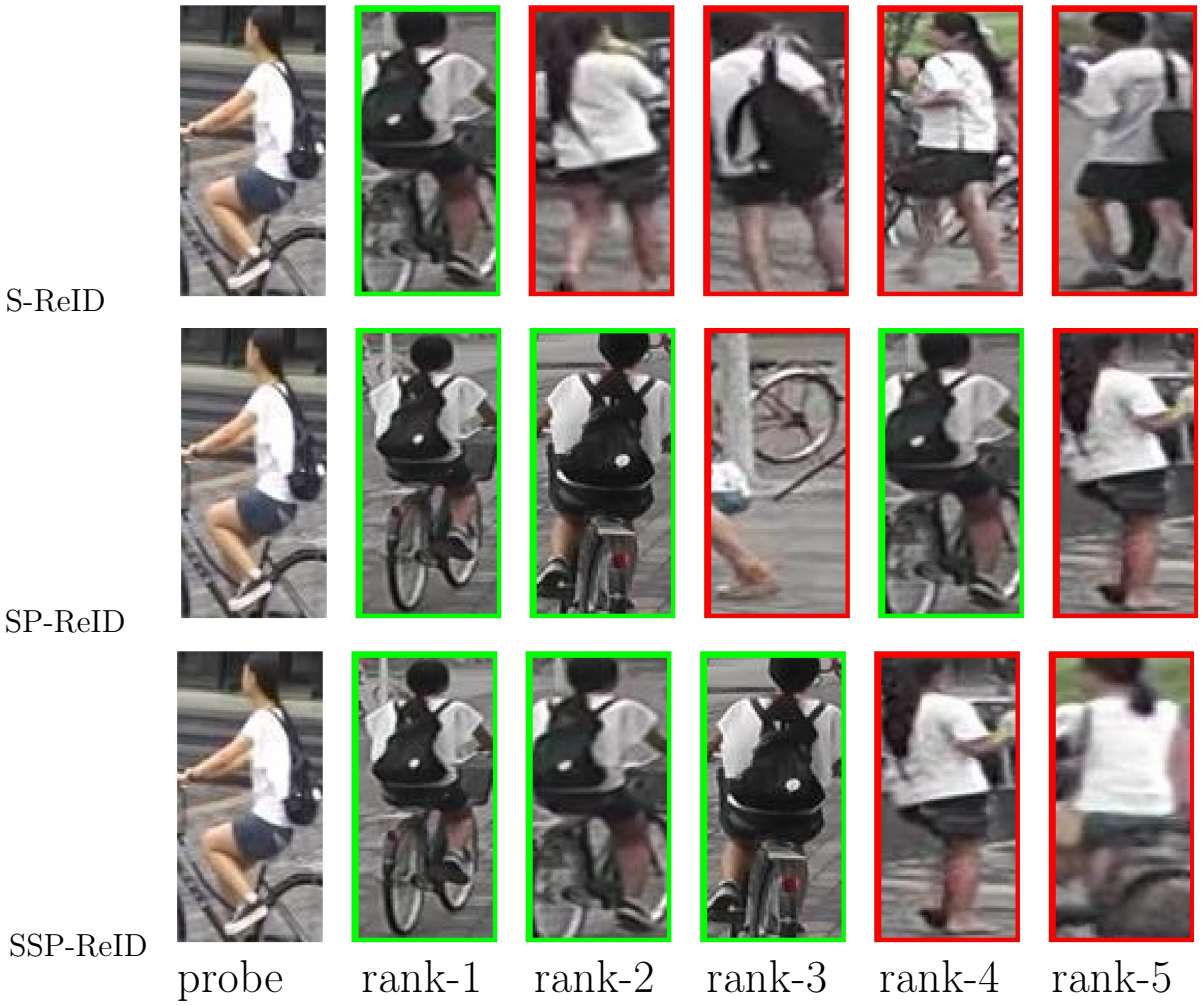


Figure 4.3: In this example from dataset Market1501 [69], SSP-ReID considers images that contains a woman with white T-shirt, black backpack in a bicycle and back-view and because they are well ranked in S-ReID and SP-ReID they are considered in the top-3 results for SSP-ReID.

same pattern as previous examples: candidates with similar appearance in both SP-ReID and S-ReID are ranked top in SSP-ReID regardless their are correct or not. Thus, as correct third match of S-ReID does not have a similar match in SP-ReID causes that this image produces a worse rank in SSP-ReId. This is a good example to illustrate how difficult is the task of ReID since it is really hard, even for humans, to say reliably the correct answer.

4.4.5 Performance Comparison for Image Re-ID

We evaluate our method with various state-of-the-art approaches available in the literature. Performance comparison for each dataset is analyzed independently into two groups: with (if reported) and without re-ranking (RR). Results for Market1501, CUHK03 and DukeMTMC-reID datasets are presented in Tables 4.6, 4.7 and 4.8.

For the Market1501 dataset, our method is the second best according to CMC metric

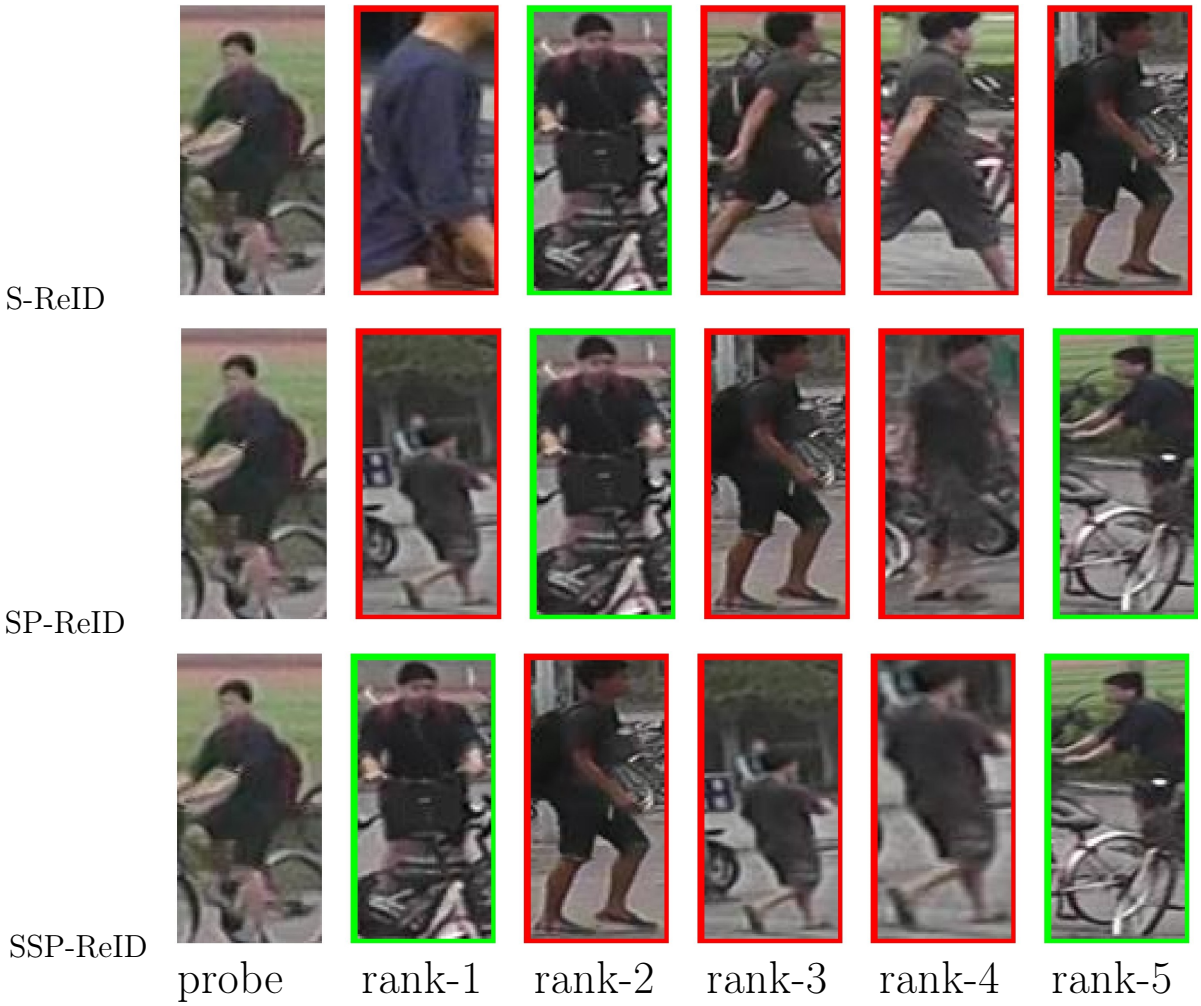


Figure 4.4: Example of proposed framework applied to images from Market1501 dataset [69]. We can observe that SSP-ReID considers similarity of candidates and their rank in S-ReID and SP-ReID to generate its results. This is expected since SSP-ReID combines features of both subnetworks and they are trained independently.

with and without RR. It achieves the third best position without RR and the second best with RR according to mAP.

For the CUHK03(D) dataset without RR, our method is the best with mAP, second best for rank-1 (with a small difference of 0.2) and best for rank-5/rank-10. For CUHK03(D) dataset with RR, our method achieves the best results with improvements of 3.4% and 1.8% according to mAP and rank-1, respectively. Similarly, for the CUHK03(L) dataset, our method is the best with mAP, the second best with rank-1 (with a difference of 0.5) and the best with rank-5/rank-10. For CUHK03(D) with RR, our method achieves the best results with improvements of 2.8% and 0.8% with mAP and rank-1, respectively.

For the DukeMTMC-ReID dataset without RR, our method achieves a competitive third best position with CMC and mAP metrics, whereas it is the second best for DukeMTMC-ReID with RR.

Our approach is able to achieve competitive results in all datasets evaluated: Market1501 (second/third best), CUHK (first/second best), DukeMTMC-ReID (second/third

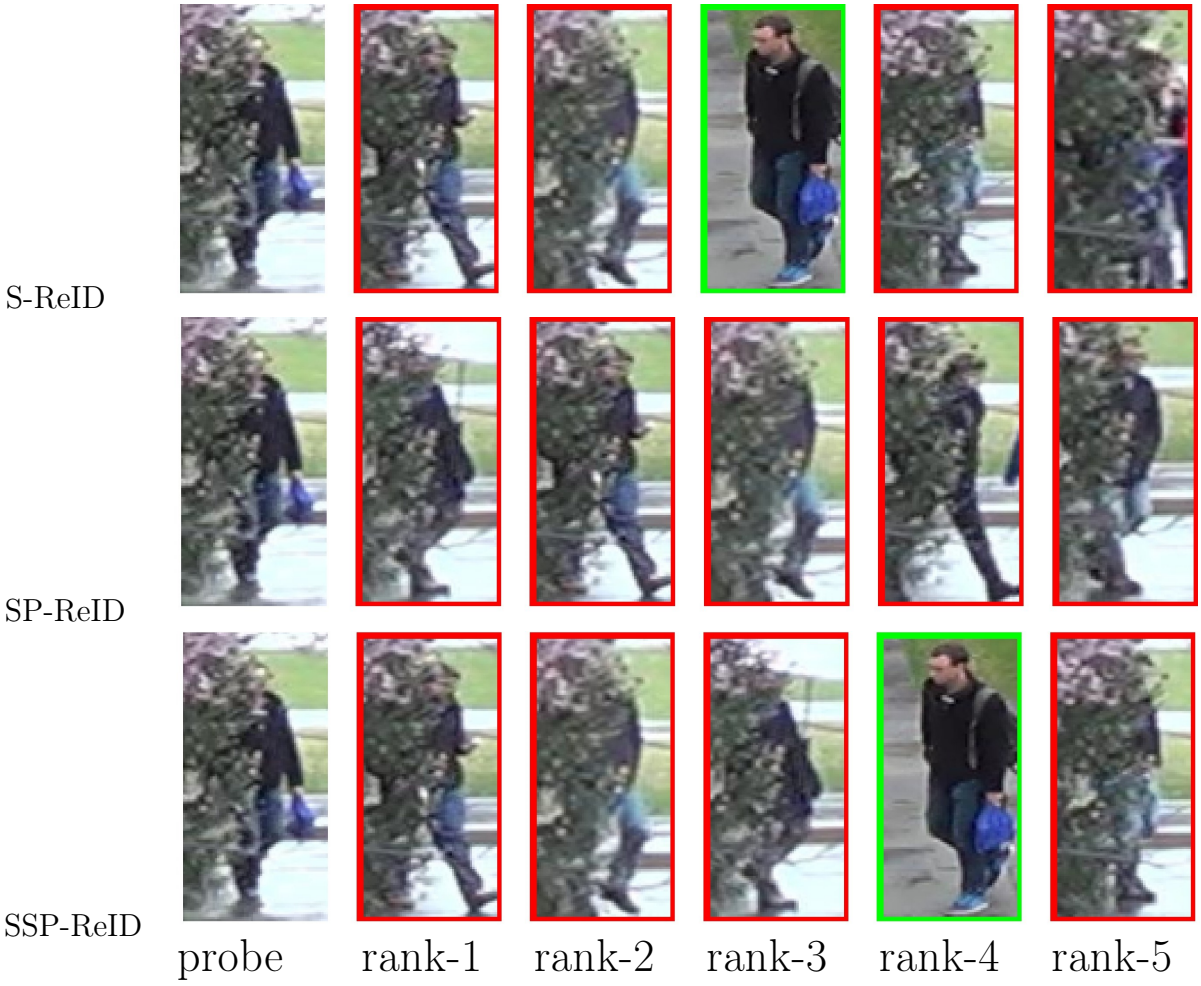


Figure 4.5: Example of a difficult case with heavy occlusions from CUHK03 dataset [25], which even for humans it is really hard to tell whether there is a correct answer, demonstrating that ReID is a challenging task.

best). SPreID [20] is the method with the best performance in two of four validation setups. However, unlike that approach, our framework does not need to be trained with 10 datasets and several thousands of epochs, we only need 180 epochs in each dataset. In addition, our framework is easier to implement compared to other methods [3, 26, 47, 60] because each module (e.g., backbone, saliency/semantic parsing detector) can be changed to a better one without breaking the pipeline, and its idea of combining multiple clues to improve the backbone performance can be applied to networks specifically proposed for Re-ID and other tasks (e.g., using relevant clues for that specific task). One the other hand, the proposed method duplicates the number of parameters. This is a drawback if the size of the network is important for the context in which Re-ID is applied.

Table 4.6: Comparison with state-of-the-art approaches for the Market1501 dataset. RR stands for re-ranking. The first, second and third highest results are highlighted in blue, red and green colors, respectively.

Market1501				
Method	**mAP (%)**	**rank-1 (%)**	**rank-5 (%)**	**rank-10 (%)**
DaRe(De)+RE [60]	76.0	89.0	–	–
SVDNet [53]	62.1	82.3	92.3	95.2
ATWL [47]	75.6	89.4	–	–
HA-CNN [26]	75.7	91.2	–	–
DuATM [51]	76.6	91.4	97.0	–
SPreID [20]	83.3	93.6	97.5	98.4
PSE [49]	84.0	90.3	–	–
ResNet-M + SSP-ReID (ours)	80.1	92.5	97.1	98.2
SPreID+RR [20]	90.9	94.6	96.8	97.6
DaRe(De)+RE+RR [60]	86.7	90.9	–	–
ResNet-M + SSP-ReID + RR (ours)	90.8	93.7	96.4	97.2

Table 4.7: Comparison with state-of-the-art approaches for the CUHK03 dataset. RR stands for re-ranking. The first, second and third highest results are highlighted in blue, red and green colors, respectively.

CUHK03 (D)				
Method	**mAP (%)**	**rank-1 (%)**	**rank-5 (%)**	**rank-10 (%)**
SVDNet [53]	37.8	40.9	–	–
HA-CNN [26]	41.0	44.4	–	–
MLFN [3]	47.8	52.8	–	–
DaRe(De)+RE [60]	59.0	63.3	–	–
ResNet-M + SSP-ReID (ours)	60.5	63.1	80.9	87.2
DaRe(De) + RE + RR [60]	71.6	70.6	–	–
ResNet-M + SSP-ReID + RR (ours)	75.0	72.4	81.6	87.4

CUHK03 (L)				
Method	**mAP (%)**	**rank-1 (%)**	**rank-5 (%)**	**rank-10 (%)**
SVDNet [53]	37.2	41.5	–	–
HA-CNN [26]	38.6	41.7	–	–
MLFN [3]	49.2	54.7	–	–
DaRe(De)+RE [60]	61.6	66.1	–	–
ResNet-M + SSP-ReID (ours)	63.3	65.6	82.6	88.6
DaRe(De) + RE + RR [60]	74.7	73.8	–	–
ResNet-M + SSP-ReID + RR (ours)	77.5	74.6	83.1	89.5

Table 4.8: Comparison with state-of-the-art approaches for the DukeMTMC-reID dataset. RR stands for re-ranking. The first, second and third highest results are highlighted in blue, red and green colors, respectively.

DukeMTMC-reID				
Method	**mAP (%)**	**rank-1 (%)**	**rank-5 (%)**	**rank-10 (%)**
DaRe(De)+RE [60]	64.5	80.2	–	–
SVDNet [53]	56.8	76.7	86.4	89.9
MLFN [3]	62.8	81.0	–	–
ATWL [47]	63.4	79.8	–	–
HA-CNN [26]	63.8	80.5	–	–
DuATM [51]	64.5	81.8	90.1	–
SPreID [20]	73.3	85.9	92.9	94.5
PSE [49]	79.8	85.2	–	–
ResNet-M + SSP-ReID (ours)	68.6	81.8	91.6	93.8
DaRe(De) + RE + RR [60]	80.0	84.4	–	–
SPreID+RR [20]	84.9	88.9	93.2	94.7
ResNet-M + SSP-ReID + RR (ours)	83.7	86.4	91.7	94.4

Chapter 5

Conclusions and Future Work

Solutions to the Re-ID problem have evolved rapidly due their relevance in security systems. The first attempts using deep learning were based on the global representation of input images, then local representations were considered. Early attempts to use local information usually divide feature tensors in horizontal stripes and combine them using techniques such as global average pooling or bilinear coding [50, 68].

This work presented a novel framework for the Re-ID problem, named SSP-ReID, which aims to use local information, but in a smarter way. It is based on saliency and semantic parsing information for guiding the learning process. SSP-ReID is composed of two subnetworks, a saliency-guided subnet that aims to focus on learning in specific parts of the image and a semantic parsing-guided subnet for dealing with misalignments, occlusions, and other challenging issues for the person re-identification task.

To validate the SSP-ReID enhancement power, we tested more than five different backbones. Ablation experiments show that the representation learned from the saliency-guided and semantic parsing-guided subnetworks has similar performance to that of the individual backbones, however, both combined boost up performance, indicating that the learned representation is complementary.

In addition, we show that layer choice for combination with saliency/semantic parsing maps does not have a significant effect on the final performance, since the backbones are able to readjust their weights.

Due to its formal definition, SSP-ReID can be applied in several networks and can combine various other clues (for instance, human pose and human orientation), which is a relevant advantage of SSP-ReID, since it is expected that the use of more clues will handle more sparse and difficult cases of Re-ID.

An extensive evaluation of the proposed Image Re-ID framework was conducted on three challenging datasets. Comparison with state-of-the-art approaches shows that our method was able to achieve competitive results in all evaluated datasets: Market1501 (second/third best), CUHK (first/second best), and DukeMTMC-ReID (second/third best). These results are promising because the backbones used were originally proposed for image classification task and the used training process follows simple and standard protocols.

Based on the conducted experiments, the hypothesis that deep architectures applied to Person Re-Identification can be enhanced by the use of multiple clues for the problem of Re-ID has been validated in our work.

In general, the main problem in current literature of Re-ID is to correctly encode the appearance of people. Therefore, directions for future work could focus on this issue. The evaluation of other clues and networks specifically developed for Re-ID as backbones should also be considered. The proposed SSP-ReID doubles the number of parameters, so evaluating methods that focus on reducing this number is critical to a more realistic scenario in which the framework is deployed in embedded systems with low computational power. An initial idea would be to share the weights of the two streams and train the end-to-end network. In addition, applying SSP-ReID in semi-supervised and unsupervised scenarios would also be an interesting investigation since the open-world version of the problem still needs extensive research work.

We expect that this work will inspire the research community to use multiple clues as enhancement technique on other classification tasks. In action recognition, for instance, clues such as optical flow and semantic parsing could be relevant. Thus, a network that included these clues in two streams would be a first attempt.

Bibliography

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017.

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Person Re-Identification Using Haar-based and DCD-based Signature. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 1–8, 2010.

[3] X. Chang, T. M. Hospedales, and T. Xiang. Multi-Level Factorisation Net for Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1–10, 2018.

[4] D. Chen, S. Zhang, W. Ouyang, J. Yang, and Y. Tai. Person Search via A Mask-Guided Two-Stream CNN Model. In *European Conference on Computer Vision*, pages 734–750, 2018.

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom Pictorial Structures for Re-Identification. In *British Machine Vision Conference*, volume 1, pages 68.1–68.11, 2011.

[6] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

[9] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-Aware Saliency Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, 2012.

[10] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin. Look Into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6757–6765, July 2017.

[11] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep Learning*, volume 1. MIT press Cambridge, 2016.

[12] A. Hauptmann, Y. Yang, and L. Zheng. Person Re-Identification: Past, Present and Future. *arXiv preprint arXiv:1610.02984*, 2016.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] A. Hermans, L. Beyer, and B. Leibe. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017.

[15] M. Hirzer, C. Beleznai, M. Koestinger, P. M. Roth, and H. Bischof. Dense Appearance Modeling and Efficient Learning of Camera Transitions for Person Re-Identification. In *IEEE International Conference on Image Processing*, pages 1617–1620, Sept. 2012.

[16] M. Hirzer, P. Roth, and H. Bischof. Person Re-Identification by Efficient Impostor-Based Metric Learning. In *IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, pages 203–208, Sept. 2012.

[17] X. Hou and L. Zhang. Saliency Detection: A Spectral Residual Approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[18] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 4700–4708, 2017.

[19] Q. Huang, W. Liu, and D. Lin. Person Search in Videos with One Portrait Through Visual and Temporal Links. In *European Conference on Computer Vision*, pages 1–17, 2018.

[20] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah. Human Semantic Parsing for Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1062–1071, 2018.

[21] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. *arXiv preprint arXiv:1605.09653*, 2016.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[23] V. J. R. Layza. Uso de Técnicas de Recuperação de Imagens para o Problema de Reidentificação de Pessoas. Master's thesis, Institute of Computing, University of Campinas, Mar. 2015.

[24] Q. Leng, R. Hu, C. Liang, Y. Wang, and J. Chen. Person Re-Identification with Content and Context Re-Ranking. *Multimedia Tools and Applications*, 74(17):6989–7014, Apr. 2014.

[25] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep Filter Pairing Neural Network for Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014.

[26] W. Li, X. Zhu, and S. Gong. Harmonious Attention Network for Person Re-Identification. In *IEEE International Conference on Computer Vision*, volume 1, pages 2285–2294, 2018.

[27] X. Li, A. Wu, and W.-S. Zheng. Adversarial Open-World Person Re-Identification. In *European Conference on Computer Vision*, pages 280–296, 2018.

[28] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang. Deepsaliency: Multi-Task Deep Neural Network Model for Salient Object Detection. *IEEE Transactions on Image Processing*, 25(8):3919–3930, 2016.

[29] X. Liang, K. Gong, X. Shen, and L. Lin. Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):871–885, Apr. 2019.

[30] S. Liao, Z. Mo, J. Zhu, and S. Z. Hu, Yangand Li. Open-Set Person Re-Identification. *arXiv preprint arXiv:1408.0872*, 2014.

[31] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear CNN Models for Fine-Grained Visual Recognition. In *IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[32] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving Person Re-Identification by Attribute and Identity Learning. *arXiv preprint arXiv:1703.07220*, 2017.

[33] C. Liu, S. Gong, and C. C. Loy. On-the-Fly Feature Importance Mining for Person Re-Identification. *Pattern Recognition*, 47(4):1602–1615, Apr. 2014.

[34] C. Liu, C. Loy, S. Gong, and G. Wang. POP: Person Re-Identification Post-Rank Optimisation. In *IEEE International Conference on Computer Vision*, pages 441–448, Dec. 2013.

[35] L. Liu, X. Lu, Y. Yuan, and X. Li. Person Re-Identification by Bidirectional Projection. In *ACM International Conference on Internet Multimedia Computing and Service*, pages 1–5, 2014.

[36] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis. *IEEE International Conference on Computer Vision*, pages 350–359, 2017.

[37] Z. Liu, J. Zhu, J. Bu, and C. Chen. A Survey of Human Pose Estimation: The Body Parts Parsing based Methods. *Journal of Visual Communication and Image Representation*, 32:10–19, 2015.

[38] C. C. Loy, C. Liu, and S. Gong. Person Re-Identification by Manifold Ranking. In *20th IEEE International Conference on Image Processing*, pages 3567–3571, Sept. 2013.

[39] C. C. Loy, T. Xiang, and S. Gong. Multi-Camera Activity Correlation Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995, 2009.

[40] L. Ma, X. Yang, and D. Tao. Person Re-Identification over Camera Networks Using Multi-Task Distance Metric Learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, Aug. 2014.

[41] N. Martinel and C. Micheloni. Re-Identify People in Wide Area Camera Network. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 31–36, June 2012.

[42] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille. Weakly-and Semi-Supervised Learning of a Deep Convolutional Network for Semantic Image Segmentation. In *IEEE International Conference on Computer Vision*, pages 1742–1750, 2015.

[43] G. Prado, H. Pedrini, and W. Schwartz. A Verify-Correct Approach to Person Re-Identification Based on Partial Least Squares Signatures. In *8th International Conference on Biometrics*, page 222–228, May 2015.

[44] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multi-Scale Deep Learning Architectures for Person Re-Identification. *arXiv preprint arXiv:1709.05165*, 2017.

[45] R. Quispe and H. Pedrini. Improved Person Re-Identification Based on Saliency and Semantic Parsing with Deep Neural Network Models. *Image and Vision Computing*, 2019 (submitted).

[46] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, pages 17–35, Oct. 2016.

[47] E. Ristani and C. Tomasi. Features for Multi-Target Multi-Camera Tracking and Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2018.

[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[49] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A Pose-Sensitive Embedding for Person Re-Identification With Expanded Cross Neighborhood Re-Ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[50] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person Re-Identification with Correspondence Structure Learning. In *IEEE International Conference on Computer Vision*, pages 3200–3208, 2015.

[51] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang. Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–11, 2018.

[52] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-Driven Deep Convolutional Model for Person Re-Identification. In *IEEE International Conference on Computer Vision*, pages 3980–3989, 2017.

[53] Y. Sun, L. Zheng, W. Deng, and S. Wang. SVDNet for Pedestrian Retrieval. *IEEE International Conference on Computer Vision*, pages 3800–3808, 2017.

[54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, volume 4, page 12, 2017.

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[57] Z. Tian, L. Liu, Z. Zhang, and B. Fei. Superpixel-based Segmentation for 3D Prostate MR Images. *IEEE Transactions on Medical Imaging*, 35(3):791–801, 2016.

[58] R. Vezzani, D. Baltieri, and R. Cucchiara. People Reidentification in Surveillance and Forensics: A Survey. *ACM Computing Surveys*, 46(2):29, 2013.

[59] W. Wang, J. Shen, and L. Shao. Deep Learning For Video Saliency Detection. *arXiv preprint arXiv:1702.00871*, 2017.

[60] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger. Resource Aware Person Re-Identification across Multiple Resolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8042–8051, 2018.

[61] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai. RGB-Infrared Cross-Modality Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5380–5389, 2017.

[62] S. Yin, Y. Qian, and M. Gong. Unsupervised Hierarchical Image Segmentation through Fuzzy Entropy Maximization. *Pattern Recognition*, 68:245–259, 2017.

[63] Q. Yu, X. Chang, Y.-Z. Song, T. Xiang, and T. M. Hospedales. The Devil is in the Middle: Exploiting Mid-Level Representations for Cross-Domain Instance Matching. *arXiv preprint arXiv:1711.08106*, 2017.

[64] Z. Zhang and V. Saligrama. Person Re-Identification via Structured Prediction. *arXiv preprint arXiv:1406.4444*, 2014.

[65] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-Learned Part-Aligned Representations for Person Re-Identification. In *IEEE International Conference on Computer Vision*, pages 3219–3228, 2017.

[66] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency Detection by Multi-Context Deep Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015.

[67] R. Zhao, W. Ouyang, and X. Wang. Person Re-Identification by Salience Matching. In *IEEE International Conference on Computer Vision*, pages 2528–2535, 2013.

[68] R. Zhao, W. Ouyang, and X. Wang. Unsupervised Salience Learning for Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013.

[69] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable Person Re-Identification: A Benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[70] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled Samples Generated by GAN Improve the Person Re-Identification Baseline in vitro. In *IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[71] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-Ranking Person Re-Identification with k-Reciprocal Encoding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3652–3661, 2017.

[72] Q. Zhou, H. Fan, H. Su, H. Yang, S. Zheng, and H. Ling. Weighted Bilinear Coding over Salient Body Parts for Person Re-Identification. *arXiv preprint arXiv:1803.08580*, 2018.

[73] Z. Zhou, W. Huang, Yanand Wang, L. Wang, and T. Tan. See the Forest for the Trees: Joint Spatial and Temporal Recurrent Neural Networks for Video-Based Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6776–6785, July 2017.

[74] H. Zhu, F. Meng, J. Cai, and S. Lu. Beyond Pixels: A Comprehensive Survey from Bottom-up to Semantic Image Segmentation and Cosegmentation. *Journal of Visual Communication and Image Representation*, 34:12–27, 2016.

[75] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning Transferable Architectures for Scalable Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8697–8710, June 2018.