Improved Person Re-Identification Based on Saliency and Semantic Parsing with Deep Neural Network Models

Rodolfo Quispe[a], Helio Pedrini[a]

[a] *Institute of Computing, University of Campinas, Campinas, SP, Brazil, 13083-852*

Abstract

Given a video or an image of a person acquired from a camera, person re-identification is the process of retrieving all instances of the same person from videos or images taken from a different camera with non-overlapping view. This task has applications in various fields, such as surveillance, forensics, robotics, multimedia. In this paper, we present 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 improves the results over the original backbones. The insight of fusing multiple clues is based on specific scenarios in which 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. We present extensive evaluation of our approach through five backbones and three benchmarks. Experimental results demonstrate the effectiveness of our person re-identification framework. In addition, we combine our framework with re-ranking techniques to achieve state-of-the-art results on three benchmarks.

Keywords: person re-identification, deep learning, multi-clue guided learning, human semantic parsing, saliency detection, convolutional neural networks

1. Introduction

Person re-identification (Re-ID) is a very challenging problem that aims to find all entities that have the same identification (ID) across cameras with respect to a gallery of individuals for a given probe (query). The probe and gallery are recorded from different camera views.

Some challenges associated with the Re-ID problem include occlusions, complex background, illumination conditions. However, the most difficult scenario is the occurrence of extreme changes in pose/viewpoint.

Re-ID is typically defined in a setting where no high resolution images are available (for example, security cameras installed in universities and airports). Since methods based on face recognition are not effective to be applied, current approaches are based on the appearance of people. More recently, the use of Convolutional Neural Networks (CNN) has become popular in this task.

All of the previously mentioned problems and constraints make Re-ID a difficult task, even for humans. Consider a scenario in which two different people are wearing similar clothing, with only a few difference in the colors of their belts and shoes. These details can be key clues to distinguishing people. In a security camera, small detail, such as the color of shoes or belt, may not be sufficiently clear so that they may be ignored by a human operator, where the two people would be considered as the same person. This realistic situation occurs in several Re-ID datasets, which present low resolution images, scaling changes, or misalignment in bounding boxes.

It is worth mentioning that the Re-ID task considers as input the bounding boxes around the people in the scene, which can be a sequence of images or videos. In this work, we focus on images, however, our framework can be easily extended to the video person Re-ID.

Since pose/viewpoint change are crucial issues for Re-ID, several approaches are based on separate horizontal-stripe images and compare people based on them, but such a method is not a complete solution. As pointed out by Kalayeh et al. [1], semantic parsing is a natural improvement for horizontal stripes because it

Email address: helio@ic.unicamp.br (Helio Pedrini)

provides labels at the pixel level, so we decided to use this insight in our framework. In addition, we realized that not every part of people is equally informative; in some cases, a backpack with a bright color or other salient objects may be a clue to the Re-ID. Thus, we designed a unified framework that unities semantic parsing and saliency to improve performance. The idea of combining multiple clues is natural to this problem, because each subnet stream of the framework can learn to solve different scenarios.

The main contributions of this paper are summarized as follows. Initially, we introduce a novel framework using saliency and semantic parsing. 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 suggest that it can be used with many other backbones because of its definition. Different from other competitive methods, our framework takes full advantage of pretrained models and require 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 framework, combined with re-ranking techniques, achieved state-of-the-art results on the three most widely used and challenging Re-ID datasets. We compared our work with the most competitive approaches available in the literature, yielding improvements of up to 4.1% in mAP and 1.8% rank-1.

The remainder of this paper is organized as follows. Section 2 briefly reviews saliency detection, semantic parsing detection, as well as methods that use these concepts in the context of person re-identification. Section 3 defines the re-identification problem and models that can be used as backbones of our framework, then it describes our method. Section 4 offers implementation details, validation protocols, evaluation and comparison with the state-of-art. Section 5 concludes the paper with some final remarks and directions for future work.

2. Background

This section reviews some relevant concepts and works associated with the research topic investigated in this work. Techniques for salient object detection, human semantic parsing, and person re-identification are described.

2.1. Salient Object Detection

Saliency detection is a task that aims to identify the fixation points that a human viewer would focus at the first glance. It has applications in various vision tasks, such as image segmentation, object detection, video summarization, compression, just to mention a few of them [2].

Early approaches to saliency detection were driven through 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 [3]. In the last years, deep models have become the mainstream solution due to the CNNs capacity for representing multi-scale and multi-level features. Some current approaches include Multi-Layer Perceptrons (MLP) and Fully Convolutional Neural Network (FCNN) [4].

The first work that used the concept of saliency in the context of person re-identification was proposed by Zhao et al. [5]. 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 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. Differently to this work, we do not use any patch matching-based approach and use deep features to encode person characteristics.

Liu et al. [6] 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. Differently from this work, our approach first computes the salient object map from the input image and then uses this map to weigh an intermediate layer of the CNN backbone. Another difference is that our training pipeline does not include the saliency detection step as part of its process. Finally, our framework is designed to be capable to use different type of backbones (ResNet [7], DenseNet [8], among others), whereas HydraPlus-Net is designed to use Inception [9] blocks for its construction.

Similar to HydraPlus-Net, Zhou et al. [10] proposed to learn saliency maps and Re-ID at the same time. They introduced a weighted version of bilinear coding [11] to encode higher-order channel-wise interactions. The main difference from our framework is that saliency maps are computed through the raw image in our pipeline, whereas the work by Zhou et al. [10] uses the output of the GoogLeNet [9] as input to their saliency Part-Net.

Qian et al. [12] 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. In our framework, we guide the network to learn from saliency and semantic parsing maps, without using multi-scale information.

2.2. *Human Semantic Parsing*

Human semantic parsing aims to segment human image into regions with fine-grained meaning, which has applications in Re-ID and human behavior analysis [13]. In its general form, semantic parsing has applications in several other domains, such as image montage, object colorization, stereo scene parsing, and medical segmentation [14].

Kalayeh et al. [1] demonstrated that the use of semantic parsing can boost up results in Re-ID. They proposed to use an Inception-based network [9] 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 framework presents similarities with their method in the sense that we also employ human semantic parsing in Re-ID, but there are some key differences: the first one is that we use intermediate layers instead of the last one since we consider that very deep layer representation encodes too abstract information and it is not intuitive to combine it in a meaningful way with semantic and saliency maps. Second, we introduce saliency in our framework, as our experiments indicated, saliency and parsing maps contain complementary information able to enhance results. In fact, when integrated with re-ranking techniques, we achieved state-of-the-art performance. Finally, our training process does not require to combine various benchmarks for creating a huge training dataset.

2.3. *Person Re-Identification*

Person re-identification (Re-ID) is defined as the task of matching all instances of the same person across multiple cameras, that is, comparing a person of interest, named probe, against a gallery of candidates previously captured. Re-ID has applications related to surveillance of public areas/events for preventing dangerous events (such as terrorism and murder). For this reason, it has received attention from the computer vision community and has been widely studied over the last years.

Early works focused on handcrafted features such as color and texture, however, due to extreme viewpoint and illumination changes, these types of characteristics are not sufficiently discriminative. Currently, Deep Learning has established a new paradigm in the Re-ID problem.

Chang et al. [15] 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. The insight behind the selection blocks is to control and specialize the features that each block is learning.

Zhao et al. [16] uses GoogLeNet [9] 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. [17] 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.

Li et al. [18] 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 [9] blocks in a multi-branch structure for global and local representation. They further introduced a method for combining these representations in a harmonious way.

To compensate for viewpoint changes, Sarfraz et al. [19] 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 in order to guide the CNNs attention. Moreover, they proposed a new re-ranking technique, named Expanded Cross Neighborhood. Results were improved based on distance between gallery and probe features. Zhong et al. [20] proposed a re-ranking method based on k-reciprocal nearest neighbors and Jaccard distance. It is worth mentioning that re-ranking methods are unsupervised and do not need any human interaction.

3. Proposed Method

In this section, we describe the person re-identification problem more formally and present our framework.

3.1. *Problem Formulation*

We consider Re-ID as a retrieval process, that is, given 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 recover 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 as ranked list) is employed to represent only people that have identity equal to y_p[1].

In this work, we create a model M' that uses M as backbone, such that the list L' generated by M' is *better* than L. The quantitative definition of *better* is based on the mean Average Precision (mAP) and cumulated matching characteristics (CMC). Both metrics are explained in the experiment section. 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 1) where people have certain items that can guide the re-identification 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 2).

We propose the Saliency-Semantic Parsing (SSP-ReID) framework, as shown in Figure 3, which is composed of two streams. Both of them have the same backbone architecture, 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.

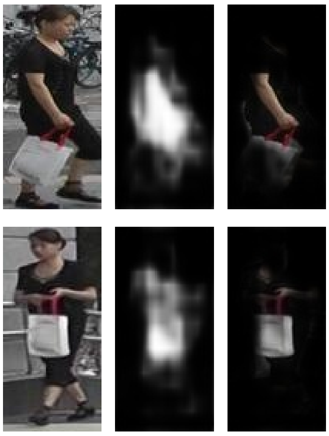

Figure 1: Examples of saliency detection for the same person (from left to right): original image, saliency map, and result of overlap saliency map over the original image. The focus of the saliency is on the arm and white bag. Our framework uses this information to guide the feature learning process.

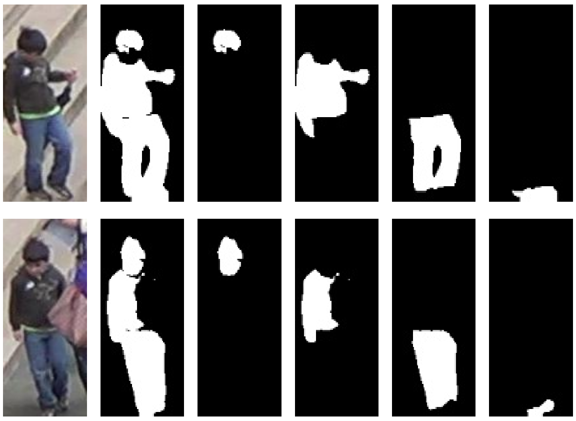

Figure 2: Examples of parsing with five semantic regions of the same person with two different views. We use these maps to overcome misalignment and occlusions.

Given the input image, we compute the saliency and semantic parsing maps using off-the-shelf deep methods [21, 13]. For the semantic parsing, we consider 5 semantic areas [2]. Then, we use a CNN to create a global representation of the person. Moreover, we take the feature map from an intermediate layer and join it with saliency map in one stream, and semantic parsing maps in the other. We decide to 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

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

[2] Head, upper body, lower body, shoes and complete body.

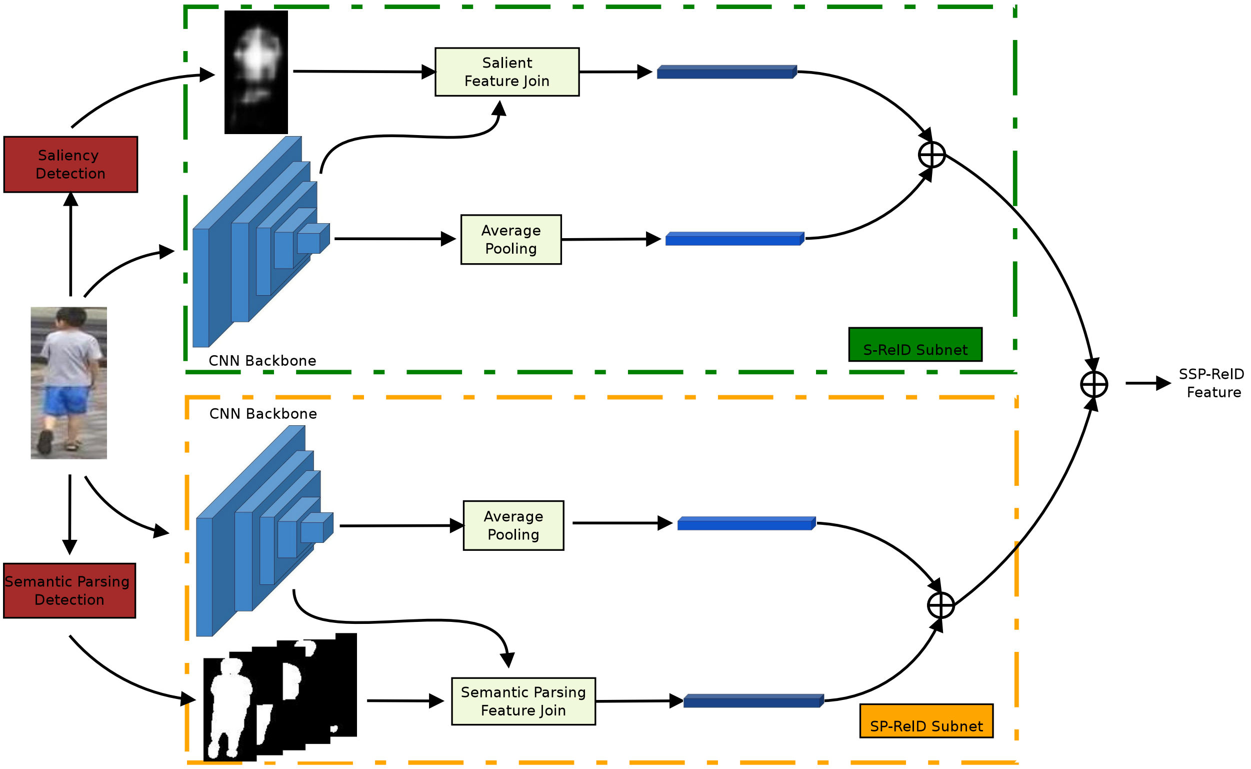

Figure 3: SSP-ReID is a framework based on semantic parsing (SP-ReID subnet) and saliency (S-ReID subnet) to learn individual-similar performance representations. At the same time both representations are complementary because the union leads to an increase in performance.

may not be clear at first glance). 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.

Given an intermediate tensor $\tau \in \mathbb{R}^{h \times w \times c}$ and a saliency/semantic parsing map $\omega \in \mathbb{R}^{h' \times w'}$, in order to join intermediate feature tensor and saliency/semantic parsing information, we initially apply a bilinear interpolation 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.

In order to train our network, we consider cross-entropy loss function with label smoothing regularizer (LSR) [22] and triplet loss with hard positive-negative mining [23].

Cross-entropy with LSR is defined as:

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

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 :

$$q'(k) = (1-\epsilon) q(k) + \frac{\epsilon}{K} \tag{2}$$

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:

$$T(X) = \sum_{i=1}^{P} \sum_{a=1}^{N} [m + \max_{p=1...N} D(f(x_a^i), f(x_p^i)) \\ - \min_{\substack{p=1...N \\ n=1...N \\ i \neq j}} D(f(x_a^i), f(x_n^j))]_+ \qquad (3)$$

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 4 illustrates the network architecture for the training step, as well as its relation to the loss functions.

4. Experimental Results

In this section, the parameter setting and datasets used in our experiments will be first introduced. Next, we will show the results obtained with our proposed method and compare them with state-of-the-art approaches.

4.1. Implementation Details

The saliency detection is performed via the off-the-shelf FCNN proposed by Li et al. [21], whereas the semantic parsing detection is computed through the Joint Human Parsing and Pose Estimation Network (JJPNet) [13] trained in the Look into Person (LIP) dataset [24].

We evaluate our framework using 5 different backbones. Table 1 summarizes the loss function used for each backbone. The initial weights of all backbones are Imaginet [25] pretrained models. In the case of the intermediate layer to be combined with saliency and semantic parsing maps: (i) we use the output of layer Res5C for ResNet50 [7] and ResNet50-M [26]; (ii) we use the output last Inception-B block for Inception-V4 [27]; (iii) we use the output of Middle Flow layer for Xception [28]; and (iv) we use the output of the second composite function for DenseNet121 [8].

Table 1: Loss function used for each backbone. CROSSE stands for Cross Entropy with LSR [22], whereas TRIP stands for Triplet Loss with hard positive-negative mining [23]. Using TRIP in Inception-V4 [27] and Xception [28] raises exploding gradient.

Backbone	Loss function
ResNet50 [7]	TRIP + CROSSE
Densenet121 [8]	TRIP + CROSSE
Resnet50-M [26]	TRIP + CROSSE
Inception-V4 [27]	CROSSE
Xception [28]	CROSSE

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 triplet loss as $m = 0.3$. Finally, we use the re-ranking method proposed by Zhong et al. [20] to compare our results with state-of-the-art approaches [3].

4.2. Datasets and Validation Protocols

We evaluate our framework on three widely used datasets. A summary of them is shown in Table 2, which reports the number of people, bounding boxes and cameras present in each benchmark setup.

Table 2: Comparative summary of Re-ID datasets used in our experiments.

Dataset	# People	# BBox	# Cameras
Market1501 [29]	1501	32668	6
CUHK03 [30]	1467	14096	6
DukeMTMC-ReID [31]	1812	36411	8

The DukeMTMC-ReID dataset [31] is a subset of the DukeMTMC dataset [32] for image-based re-identification with hand-drawn bounding boxes. Bounding boxes of different sizes with outdoor scenes as background are available. For validation, we use the fixed training and testing sets proposed in the original protocol of the dataset.

The Market1501 [29] was created through Deformable Part Model (DPM) in order to simulate a real-world scenario. For validation, we use the fixed training and testing sets provided with the dataset.

[3] Code and models will be available upon acceptance.

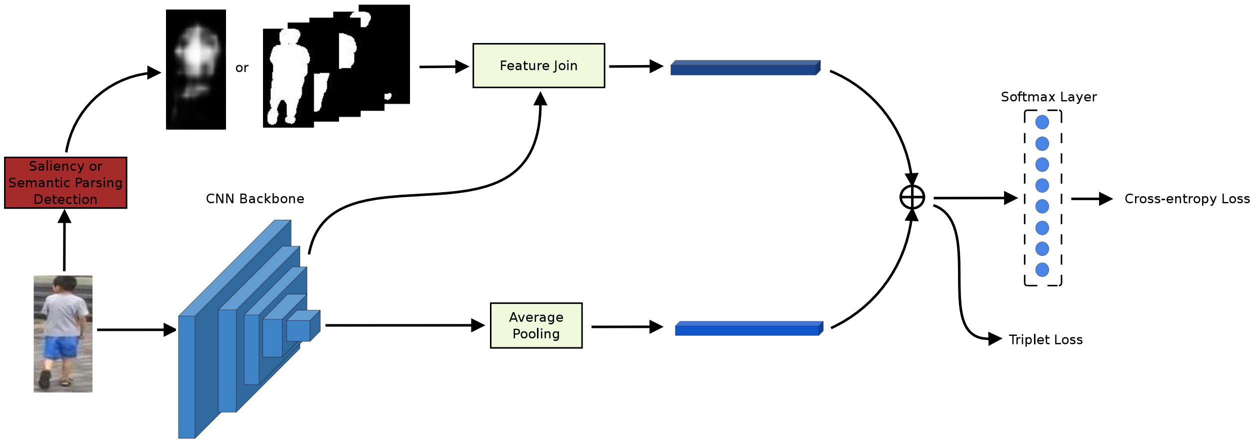

Figure 4: Training setup for S-ReID and SP-ReID subnetworks. When training out framework, we consider triplet and cross-entropy loss functions. For the triplet loss, we take the feature vector before softmax layer and use it to compare images based on the Euclidean distance. The triplet loss may be ignored depending on the CNN backbone.

The CUHK03 [30] has an average of 4.8 images per view. Misalignment, occlusions and missing body parts are quite common. For validation, we use the new validation protocol [20] with partition of 767/700. Moreover, we evaluate detected (CUHK03 (D)) and labeled (CUHK03 (L)) versions of the dataset.

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

$$mAP = \frac{AP}{\#queries} \quad (4)$$

where AP is defined as

$$AP = \frac{\sum_{k=1}^{n} P(k) \cdot rel(k)}{\#relevant\ items} \quad (5)$$

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

$$P(k) = \frac{\sum_{i=1}^{k} rel(i)}{k} \quad (6)$$

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

$$CMC(r) = \frac{in(r)}{\#queries} \quad (7)$$

where $in(r)$ is the number of queries that have a relevant element within the first r items in the ranked list. We set $r = 1$ and refer to it as rank-1.

4.3. Performance in Re-ID

In this section, we evaluate and compare different aspects of our framework: backbone (e.g., ResNet [7]), saliency subnet (S-ReID), semantic parsing subnet (SP-ReID) and the complete framework (SSP-ReID). Results are summarized in Table 3. Overall, ResNet50-M + SSP-ReID produced the best results for all datasets, whereas there are marginal differences in using ResNet50 [7] and DenseNet [8] as backbones. On the other hand, Inception-V4 [27] and Xception [28] yielded the worse performance. Note that DenseNet [8] gets interesting results despite of its lower number of parameters. In addition, our framework raises consistently an improvement over all backbones, this suggest that our framework can be used as a enhancing method in future works.

In the light of all datasets, S-ReID achieved improvements up to 1.3% 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. This suggests that

Table 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.

	Market1501		**CUHK03 (D)**		**CUHK03 (L)**		**DukeMTMC-reID**	
Method	**mAP(%)**	**rank-1(%)**	**mAP(%)**	**rank-1(%)**	**mAP(%)**	**rank-1(%)**	**mAP(%)**	**rank-1(%)**
ResNet [7]	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 [26]	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 [8]	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 [27]	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 [28]	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

S-ReID and SP-ReID are learning similar performance representation, but with complementary information, which improves the model capacity when they are combined. Figure 5 shows an example of rank-2 results for DukeMTMC-ReID [31].

It can be observed that the method improvement is inversely proportional to the capacity of the backbone. For better backbones (for instance, ResNet50-M [26]), the improvements are smaller when compared to the lower performance backbones (for instance, Xception [28]). 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 with more specific and often sparse cases.

4.4. Performance Comparison

We evaluate our method with various competitive approaches available in the literature. A performance comparison is presented in Table 4, where we applied re-ranking (RR) to boost our final results.

Our method was able to achieve state-of-the-art in all datasets. Overall, performance improvement is higher for mAP than rank-1, because mAP considers a greater number of elements in its definition. Thus, it is more sensitive to any change in the ranking list. SPreID [1] is the method with the closest performance, however, unlike such approach, our framework does not have to be trained with 10 datasets, we only need 180 epochs in each dataset. In addition, our framework is easier to implement compared to other methods [15, 18, 33, 35].

5. Conclusions and Future Work

In this work, we presented SSP-ReID, a framework based on saliency and semantic parsing information, which achieved state-of-the-art results on three challenging datasets for the re-identification task. 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.

Table 4: Comparison with the state-of-art, in **bold** the best results, RR stands for re-ranking

	Market1501		**CUHK03 (D)**		**CUHK03 (L)**		**DukeMTMC-reID**	
Method	**mAP(%)**	**rank-1(%)**	**mAP(%)**	**rank-1(%)**	**mAP(%)**	**rank-1(%)**	**mAP(%)**	**rank-1(%)**
SPreID [1]	83.3	93.6	—	—	—	—	73.3	85.9
DaRe(De)+RE+RR [33]	86.7	90.9	71.6	70.6	74.7	73.8	80.0	84.4
DuATM [34]	76.6	91.4	—	—	—	—	64.5	81.8
HA-CNN [18]	75.7	91.2	41.0	44.4	38.6	41.7	63.8	80.5
ATWL [35]	75.6	89.4	—	—	—	—	63.4	79.8
PSE [19]	84.0	90.3	—	—	—	—	79.8	85.2
MLFN [15]	74.3	90.0	47.8	52.8	49.2	54.7	62.8	81.0
SVDNet [36]	62.1	82.3	37.8	40.9	37.2	41.5	56.8	76.7
ResNet + SSP-ReID	75.9	89.3	57.1	59.4	58.9	60.6	66.1	80.1
ResNet + SSP-ReID + RR	88.2	91.5	71.1	67.6	72.4	68.4	81.4	84.8
ResNet-M + SSP-ReID	80.1	92.5	60.5	63.1	63.3	65.6	68.6	81.8
ResNet-M + SSP-ReID + RR	**90.8**	**93.7**	**75.0**	**72.4**	**77.5**	**74.6**	**83.7**	**86.4**
DenseNet + SSP-ReID	76.7	90.9	48.1	48.1	49.5	49.1	67.1	82.2
DenseNet + SSP-ReID + RR	89.9	93.3	63.1	58.4	64.7	59.9	83.3	86.2

We conducted extensive evaluation of our framework on five different backbones and three datasets. 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 the results, indicating that the learned representation is complementary.

Our framework can be easily adapted to multiple CNN backbones, further improving performance over original networks. We expect that the use of multiple clues (for instance, human pose and multi-scale strategies) inspires other re-identification works.

6. Acknowledgments

The authors are thankful to FAPESP (grant #2014/12236-1) and CNPq (grant #305169/2015-7) for their financial support.

References

[1] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, M. Shah, Human Semantic Parsing for Person Re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1062–1071.

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

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

[4] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, J. Li, Salient Object Detection: A Survey, arXiv preprint arXiv:1411.5878.

[5] R. Zhao, W. Ouyang, X. Wang, Person Re-identification by Salience Matching, in: IEEE International Conference on Computer Vision, 2013, pp. 2528–2535.

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

[7] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[8] G. Huang, Z. Liu, K. Q. Weinberger, L. van der Maaten, Densely Connected Convolutional Networks, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2017, p. 3.

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

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

[11] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear CNN Models for Fine-Grained Visual Recognition, in: IEEE International Conference on Computer Vision, 2015, pp. 1449–1457.

[12] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, X. Xue, Multi-scale Deep Learning Architectures for Person Re-identification, arXiv preprint arXiv:1709.05165.

[13] K. Gong, X. Liang, D. Zhang, X. Shen, 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, 2017.

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

[15] X. Chang, T. M. Hospedales, T. Xiang, Multi-Level Factorisation Net for Person Re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, 2018, p. 2.

[16] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-Learned Part-Aligned Representations for Person Re-identification, in: IEEE

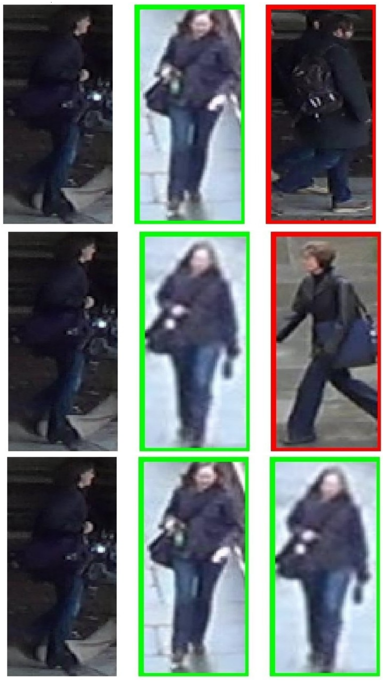

Figure 5: Qualitative example of performance of framework modules. The first column represents the query, the second and third columns represent the results of rank-1 and rank-2, respectively (the correct results are in green, otherwise they are in red). The first, second and third rows represent the output of S-ReID, SP-ReID and SSP-ReID, respectively. 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.

International Conference on Computer Vision, 2017.

[17] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven Deep Convolutional Model for Person Re-identification, in: IEEE International Conference on Computer Vision, IEEE, 2017, pp. 3980–3989.

[18] W. Li, X. Zhu, S. Gong, Harmonious Attention Network for Person Re-identification, in: IEEE International Conference on Computer Vision, Vol. 1, 2018, p. 2.

[19] M. Saquib Sarfraz, A. Schumann, A. Eberle, 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, 2018, pp. 420–429.

[20] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking Person Re-identification with *k*-reciprocal Encoding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[21] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, J. Wang, Deepsaliency: Multi-task deep neural network model for salient object detection, IEEE Transactions on Image Processing 25 (8) (2016) 3919–3930.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception architecture for computer vision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.

[23] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737.

[24] X. Liang, K. Gong, X. Shen, L. Lin, Look into Person: Joint Body Parsing & Pose Estimation Network and a New Benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[25] 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) (2015) 211–252.

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

[27] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning, in: Thirty-First AAAI Conference on Artificial Intelligence, Vol. 4, 2017, p. 12.

[28] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable Person Re-identification: A Benchmark, in: IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[30] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep Filter Pairing Neural Network for Person Re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.

[31] Z. Zheng, L. Zheng, Y. Yang, Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro, in: IEEE International Conference on Computer Vision, 2017.

[32] E. Ristani, F. Solera, R. Zou, R. Cucchiara, 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, 2016.

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

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

[35] E. Ristani, C. Tomasi, Features for Multi-Target Multi-Camera Tracking and Re-Identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[36] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for Pedestrian Retrievala, arXiv preprint.