Human Semantic Parsing for Person Re-identification

Mahdi M. Kalayeh[1*] Emrah Basaran[2*] Muhittin Gökmen[3]
mahdi@eecs.ucf.edu basaranemrah@itu.edu.tr gokmenm@mef.edu.tr
Mustafa E. Kamasak[2] Mubarak Shah[1]
kamasak@itu.edu.tr shah@crcv.ucf.edu

[1]Center for Research in Computer Vision, University of Central Florida
[2]Dept. of Computer Engineering, Istanbul Technical University
[3]Dept. of Computer Engineering, MEF University

Abstract

*Person re-identification is a challenging task mainly due to factors such as background clutter, pose, illumination and camera point of view variations. These elements hinder the process of extracting robust and discriminative representations, hence preventing different identities from being successfully distinguished. To improve the representation learning, usually local features from human body parts are extracted. However, the common practice for such a process has been based on bounding box part detection. In this paper, we propose to adopt human semantic parsing which, due to its pixel-level accuracy and capability of modeling arbitrary contours, is naturally a better alternative. Our proposed SPReID integrates human semantic parsing in person re-identification and not only considerably outperforms its counter baseline, but achieves state-of-the-art performance. We also show that, by employing a simple yet effective training strategy, standard popular deep convolutional architectures such as Inception-V3 and ResNet-152, with no modification, while operating solely on full image, can dramatically outperform current state-of-the-art. Our proposed methods improve state-of-the-art person re-identification on: Market-1501 [48] by ∼**17**% in mAP and ∼**6**% in rank-1, CUHK03 [24] by ∼**4**% in rank-1 and DukeMTMC-reID [50] by ∼**24**% in mAP and ∼**10**% in rank-1.*

1. Introduction

Given a query image, person re-identification is the problem of retrieving all the images of the same identity from a large gallery, where query and gallery images are captured by distinctively different cameras which may or may not have any field-of-view overlap. Hence it can be seen as a cross-camera data association problem.

Person re-identification is a very challenging task. First, when a single person is captured by two different cameras, the illumination conditions, background clutter, occlusion, observable human body parts, and perceived posture of the person can be dramatically different. Second, even within a single camera, the aforementioned conditions can vary through time as the person moves and engages in different actions (*e.g* suddenly taking something out of a bag while walking). Third, gallery itself usually consists of diverse images of a single person from multiple cameras, which given the above factors, generates a huge intra-class variation impeding the generalization of the learned representations. Fourth, compared to problems such as object recognition or detection, images in person re-identification benchmarks are usually of lower resolution, making it difficult to extract distinctive attributes to distinguish one identity from another. Considering the above challenges, an effective person re-identification system is obliged to learn representations that are identity-specific, context-invariant and agnostic with respect to the camera point of view.

In recent past, improving global (image-level) representation by leveraging local (part-level) features extracted from human body parts has been the main theme of person re-identification research. While an image-level representation is prone to background clutter and occlusion, part-level representations are supposed to be more robust. However, part detection in low resolution images has its very own challenges and any error in that stage can propagate to the entire person re-identification system. That is why some research works prefer to simply extract representations from multiple image patches, often horizontal strips, that are loosely associated to human body parts. On the

[*]Authors contributed equally

other hand, almost all of the previous works which involve body parts begin with an often off-the-shelf pose estimation model and infer corresponding bounding boxes from predicted joint locations. The person re-identification systems then process the global and local representations in what can be coarsely seen as multi-branch deep convolutional neural network (CNN) architectures. These models while delivering very good results, usually consist of many sub-models that are trained in multiple stages, tailored specifically for person re-identification problem. By studying recent literature, we raise two major questions, in this paper. **First**, are such *complex* models necessary to improve the performance of person re-identification? **Second**, are the local features best captured using bounding boxes on human body parts?

Addressing the first question, we show that a *simple* model based on Inception-V3 [37] with no bells and whistles, operating solely on full body images and optimized in a straightforward training procedure can outperform current state-of-the art. Unlike recent research works which commonly adopt, binary or triplet losses, we train our model using softmax cross-entropy at two different input resolutions. Using re-ranking as a post processing technique, the improvement margin further increases.

To address the second question, we propose using semantic segmentation, more specifically human semantic parsing, as an alternative to bounding boxes in order to extract local features from human body parts. While bounding boxes are coarse, can include background, and cannot capture deformable nature of human body, semantic segmentation is able to precisely localize arbitrary contours of various body parts even under severe pose variations. We begin by training a human semantic parsing model that learns to segment human body into multiple semantic regions and then use them to exploit local cues for person re-identification. We analyze two variations for integrating human semantic parsing into re-identification and show that they provide complementary representations. The contributions of this paper are as follows:

- Through extensive set of experiments, we show that, our *simple* yet effective training procedure can significantly outperform current state-of-the-art. We verify our observations using two standard deep convolutional architectures, namely Inception-V3 [37] and ResNet-152 [16] on three different benchmarks.

- We propose SPReID, where human semantic parsing is employed to harness local visual cues for person re-identification. To do so, we train our very own semantic segmentation model and show that it not only helps improving person re-identification, but also achieves state-of-the-art performance on human semantic parsing problem, demonstrating the quality of our model.

- We improve state-of-the-art person re-identification performance on: Market-1501 [48] by $\sim$**17**% in mAP and $\sim$**6**% in rank-1, CUHK03 [24] by $\sim$**4**% in rank-1 and DukeMTMC-reID [50] by $\sim$**24**% in mAP and $\sim$**10**% in rank-1.

The remainder of this paper is organized as follows. Section 2 offers a brief overview of the person re-identification literature. We then present our method in Section 3. Experimental results are discussed in Section 4, followed by the implementation details in Section 5. Finally, we conclude the paper in Section 6.

2. Related Work

In recent years, significant progress has been achieved in different computer vision areas, including in person re-identification, thanks to the emergence of deep learning, and in particular deep convolutional neural networks. Challenges due to variations in pose and illumination, occlusion and background clutter in the person re-identification problem have resulted in the research community to focus on two major sub-problems, namely feature representation and similarity or distance metrics. Improvements in feature representation have mainly been achieved by leveraging local cues while in the latter, similarity measures such as contrastive or triplet loss have been studied. Next, we briefly survey the person re-identification literature.

To obtain robust representations, authors in [21, 45] augment a global representation by employing human body parts. Specifically, Li *et al.* [21] learn the body parts roughly as head-shoulder, upper-body and lower-body using a spatial transformer network [19]. Then multiple streams, with shared weights, through a multi-scale CNN structure process these parts and ultimately concatenate them with a global representation. Zhao *et al.* [45] use region proposal network, trained on an auxiliary pose dataset, to detect body parts. Part representations are then gradually combined and finally fused with the global representation. Their proposed model is very complex and is trained through multiple non-trivial stages. While avoiding to explicitly detect human body parts, authors in [52, 11] try to benefit from local cues by extracting multiple patches from image which are loosely associated to human body parts. Such frameworks cannot address the part misalignment properly. Taking a slightly different approach, in [26, 30], authors develop attention-based models where respectively, a Long Short-Term Memory (LSTM) [18] and a gradient-based attention model dynamically focus on distinctive regions in the image. Some works [47, 34] have tried to address the misalignment issue by explicitly integrating pose estimation into person re-identification where off-the-shelf pose estimation models are used to initialize part locations as quadrilaterals which then are aligned via

affine transformation or spatial transformer network [19].

Also, there has been some attempts [33, 35] to improve person re-identification performance using person attributes. These attributes usually contain high-level semantic information that are supposedly invariant to pose, illumination and camera point of view. However, one should note that the exact same conditions make reliable detection of those attributes very challenging.

Several loss functions have been adopted for person re-identification. Some like [38, 39, 24, 1] have used positive and negative image pairs through contrastive and binary (verification) losses to train their neural network models. Others [12, 9, 11] have employed triplet loss which requires a tuple of anchor, positive and negative images where the training objective is to simultaneously pushing the positive image towards the anchor while pulling the negative image away from it. These loss functions are very suitable for person re-identification due to its retrieval nature. However, their effectiveness is highly dependent on how the training pairs/triplets are chosen. Easy to distinguish pairs/triplets do not help the learning since no error signal will be back-propagated while the hard ones can result in the training process to diverge. Unlike the aforementioned approaches, Zhao *et al.* [45], adopt simple multi-class classification loss while [49, 29] use a combination of both classification and verification losses.

In contrast to the above works, we propose to employ human semantic parsing to extract local regions from human body. We argue that semantic segmentation, due to its pixel-level accuracy, is naturally more suitable than bounding box part localization to cope with person re-identification challenges. To the best of our knowledge, we are the first to propose the integration of human semantic parsing into person re-identification.

3. Methodology

In this work, unless specified otherwise, we use Inception-V3 [37] as the CNN backbone for both human semantic parsing and person re-identification models. Therefore, we begin by briefly describing the Inception-V3 [37] architecture. Then, we provide details for our human semantic parsing model and finally explain how to integrate it into our proposed person re-identification framework.

3.1. Inception-V3 Architecture

Inception-V3 [37] is a 48-layers deep convolutional architecture. Since it employs global average pooling instead of fully-connected layer, it can operate on arbitrary input image sizes. While being shallower than different variations of popular ResNet [16], our experiments show that it gives competitive and in cases even better results than ResNet-152 [16], while being dramatically less computationally expensive. We will provide quantitative comparison between different choices of the backbone architecture.

The Inception-V3 [37] has an output stride of 32, where the activation size quickly reduces to $\frac{1}{8}$ of the input image resolution within the first seven layers. Such reduction is achieved by two convolution and one max pooling layer that operate with the stride of 2. The network follows by three blocks of Inception layers separated by two grid reduction modules. Spatial resolution of the activations remains intact within the Inception blocks, while grid reduction modules halve the activation size and increase the number of channels. Then, the output of the last Inception block is aggregated via global average pooling to produce a 2048-D feature vector. For more details on the architecture, readers are encouraged to refer to [37].

3.2. Human Semantic Parsing Model

In order to exploit local cues for person re-identification, we propose to employ human semantic parsing. We argue that semantic segmentation due to its pixel-level accuracy and robustness to pose variation is naturally superior to bounding box part detection.

We use Inception-V3 [37] as the backbone architecture of our human semantic parsing model. However, we make two modifications to adopt it for the semantic segmentation task. The quality of human semantic parsing heavily relies on the final activations to be of sufficient resolution. Hence, we change the stride of the last grid reduction module in the Inception-V3 [37] from 2 to 1 resulting in an output stride of 16 compared to 32 in the original architecture. To cope with the extra computation that consequently is added to the last Inception block, corresponding convolution filters are replaced with the dilated convolution [43]. We then remove the global average pooling and add an atrous spatial pyramid pooling (rates=3,6,9,12)[7] followed by a 1×1 convolution layer as the classifier. This would allow us to perform multi-class classification in pixel-level and is a standard approach, commonly used in semantic segmentation architectures [6, 7].

3.3. Person Re-identification Model

Our person re-identification model, illustrated in Figure 1 consists of a convolutional backbone, a human semantic parsing branch and two aggregation heads. From now on, we refer to it as **SPReID**: Human **S**emantic **P**arsing for Person **Re-id**entification. The person re-identification backbone in SPReID is exactly Inception-V3 [37] with a minor modification of removing global average pooling layer. Hence, it generates a tensor of 2048 channels with the output stride of 32.

Our baseline person re-identification model simply aggregates the output activations of the convolutional backbone using global average pooling. Corresponding aggregation head, shown in Figure 1 generates a 2048-D global

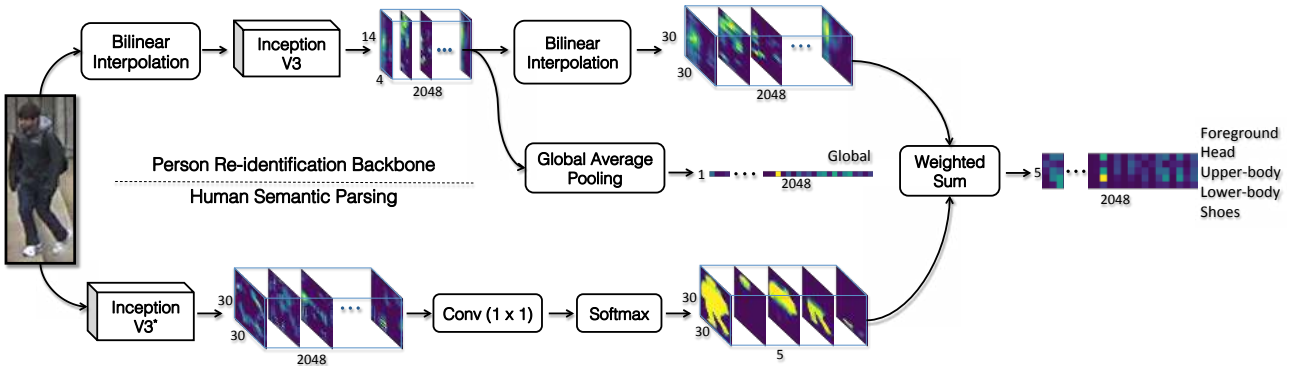

Figure 1: SPReID framework: our proposed person re-identification model first transforms the input RGB image into a tensor of activations via a convolutional backbone while simultaneously generating probability maps associated to different semantic regions of human body using the human semantic parsing branch. Note that the Inception-V3 module in the lower branch is denoted as Inception-V3*. That refers to the modifications which we applied (ref. Section 3.2) to the original Inception-V3 [37] architecture. SPReID then uses the aforementioned probability maps to aggregate the convolutional activations from different semantic regions of human body.

representation. To train the network, we pass it to a multi-class classification (over different identities) objective with softmax cross-entropy loss. To avoid clutter, we are not showing the loss in Figure 1. At the test time, final representations before the classifier layer are used to retrieve correct matches of a given query from the gallery. In Section 4, we show how the performance of the baseline model varies if we change the backbone architecture from Inception-V3 [37] to ResNet-50 [16] and ResNet-152 [16].

To exploit the local visual cues, we use the probability maps associated to five different body regions, namely foreground, head, upper-body, lower-body and shoes. These probability maps are generated by the human semantic parsing model and are ℓ_1-normalized per channel. In SPReID, we pool the output activations of the CNN backbone multiple times, each time using one of the five probability maps. This is in contrast with global average pooling, which is agnostic with respect to where in the spatial domain activations occur. It is not hard to see that exclusively pooling activations within different semantic regions associated to human body parts can be seen as a weighted sum operation where the probability maps are used as weights. From an implementation point of view, this is equal to a matrix multiplication between the output of re-identification backbone and human semantic parsing where their corresponding spatial domain is flattened. Such a procedure results in five 2048-D feature vectors each exclusively representing one human body region. Next, we perform element-wise max operation over representations of head, upper-body, lower-body and shoes and concatenate the outcome with the foreground and previously described global representation from the full image. Our proposed technique is applicable to any convolutional backbone choice and adds minimal computation to the naive global average pooling which serves as our baseline person re-identification model. Note that since the human semantic parsing model usually operates on higher resolution images, the re-identification backbone, as shown in Figure 1 uses bilinear interpolation to initially scale down the input images and then scale up the final activations to match the ones in human semantic parsing branch.

4. Experiments

4.1. Datasets and Evaluation Measures

To evaluate our proposed methods, we use three publicly available large-scale person re-identification benchmarks namely Market-1501 [48], CUHK03 [24] and DukeMTMC-reID [50]. Market-1501 [48] dataset consists of 32,668 images of 1,501 subjects captured by 5 high-resolution and one low-resolution camera. In this dataset, to obtain the person bounding boxes, Deformable Part Model (DPM) [13] is used. Therefore, there are misaligned detected boxes within the dataset. In its standard evaluation protocol, the training set consists of 751 identities and has a total of 12,936 images. In the test set, images of 750 identities which have not appeared in the training are used to create gallery and query sets. These sets respectively contain 19,734 and 3,368 images.

DukeMTMC-reID [50] dataset consists of the person images which are extracted from the DukeMTMC [31] tracking dataset. DukeMTMC contains images taken from 8 high-resolution cameras, and person bounding boxes are hand-annotated. The standard evaluation protocol [50] of DukeMTMC-reID dataset is in the same format as Market-1501. Specifically, 16,522 images of 702 persons are reserved as training set. For gallery and probe, respectively

16,522 and 2,228 images associated to 702 identities that do not appear in the training set are used. In CUHK03 [24] dataset, there are 13,164 images with a total of 1,467 identities. These images were recorded by 6 surveillance cameras and each person is viewed by 2 different cameras. For the experiments conducted on this benchmark, both manually annotated and DPM-detected bounding boxes can be used. The evaluation protocol of CUHK03 is in a different format than the other two datasets. In our experiments, we are following the standard protocol detailed in [24] and reporting the results on the manually annotated images.

In addition to these datasets that were used in evaluation, we utilize 3DPeS [4], CUHK01 [23], CUHK02 [22], PRID [17], PSDB [41], Shinpuhkan [20] and VIPeR [15] datasets to augment our training data. The training splits of these datasets, in addition to Market-1501 [48], CUHK03 [24] and DukeMTMC-reID [50], are aggregated to create a large training set which consists of ∼111,000 images. We evaluate the quality of different person re-identification models using Cumulative Matching Characteristic (CMC) curves and mean average precision (mAP). All the experiments are performed in single query setting.

4.2. Training the Networks

To train our person re-identification models, we aggregate 10 different person re-identification benchmarks, detailed in Section 4.1, which results in a total of ∼111,000 images of ∼17,000 identities. The baseline models solely operate on full image with no use of semantic segmentation. We begin by training them for 200K iterations using input images of size 492×164. Then, we fine-tune each one for an additional 50K iteration but on higher input resolution of 748×246. Fine-tuning is conducted on Market-1501, CUHK03 and DukeMTMC-reID datasets separately. Training of SPReID is done on the aggregation of 10 datasets with the exact same setting as above. The input image resolution in its associated experiments is set to 512×170.

We train the human semantic parsing model on Look into Person (LIP) [14] dataset which consists of ∼30,000 images with 20 semantic labels[1]. The probability of predictions for different regions are then grouped together to create 5 coarse labels[2] in order to parse human body for the person re-identification. Our experiments indicate that the human semantic parsing model is capable of decently localizing various human body parts even under severe pose variation and occlusion. Despite being out of the scope of this work, to demonstrate the quality of our human semantic parsing, we show in Table 1 that, on the validation set of LIP [14], our model outperforms the current state-of-the-art. Figure 2 illustrates how our human semantic parsing model segments example images from DukeMTMC-reID [50] person re-identification benchmark.

[1]Background, Hat, Hair, Glove, Sunglasses, Upper-clothes, Dress, Coat, Socks, Pants, Jumpsuits, Scarf, Skirt, Face, Right-arm, Left-arm, Right-leg, Left-leg, Right-shoe and Left-shoe

[2]Foreground, Head, Upper-body, Lower-body and Shoes

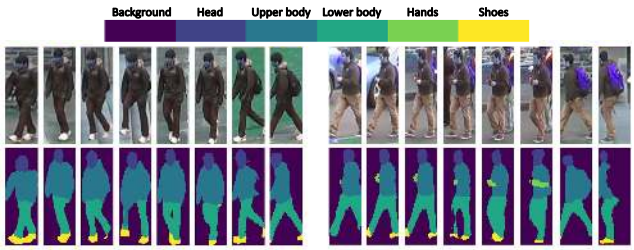

Figure 2: Examples of the segmentation masks generated by our human semantic parsing model on random images from DukeMTMC-reID [50] person re-identification benchmark.

method	overall acc.	mean acc.	mean IoU
SegNet [2]	69.04	24.00	18.17
FCN-8s [28]	76.06	36.75	28.29
DeepLabV2 [6]	82.66	51.64	41.64
Attention [8]	83.43	54.39	42.92
DeepLabV2 + SSL [14]	83.16	52.55	42.44
Attention + SSL [14]	84.36	54.94	44.73
Ours	**85.07**	**60.54**	**48.16**

Table 1: Performance (%) comparison of human semantic parsing on the validation split of LIP [14].

4.3. Person Re-identification Performance

In this section, we begin by analyzing the performance of our baseline person re-identification models. We will show the effect of input image resolution, fine-tuning on large image size, different choices for the re-identification backbone, and finally weight sharing among aggregation heads. We show that the baseline models, thanks to our simple yet well designed training strategy, can outperform the current state-of-the-art with large margin. Then, we quantitatively illustrate the effectiveness of SPReID in harnessing human semantic parsing for person re-identification. We conclude this section by comparison with the state-of-the-art person re-identification on three large-scale benchmarks.

Effect of input image resolution: In Table 2, we show quantitative results from our Inception-V3 baseline model when different input resolutions are used to train the network. Other than that, the rest of settings/parameters are the same for all models. We observe that on all three datasets, training on higher resolution input images yields a better performance measured by either mAP or re-identification rate. Though such gap tends to shrink when we consider rank-10 versus rank-1, as it is expected. Model-S, Model-M

Market-1501				
model	input size	mAP(%)	rank-1	rank-10
Model-S	246×82	64.45	84.06	95.55
Model-M	375×125	72.14	88.18	96.64
Model-L	492×164	73.06	88.87	97.00
Model-L^{ft}	748×246	**76.56**	**90.8**	**97.71**
CUHK03				
model	input size	mAP(%)	rank-1	rank-10
Model-S	246×82	–	81.78	98.12
Model-M	375×125	–	85.66	98.90
Model-L	492×164	–	87.91	98.41
Model-L^{ft}	748×246	–	**88.73**	**98.94**
DukeMTMC-reID				
model	input size	mAP(%)	rank-1	rank-10
Model-S	246×82	53.73	74.87	88.51
Model-M	375×125	59.98	79.85	90.89
Model-L	492×164	59.87	79.08	90.26
Model-L^{ft}	748×246	**63.27**	**80.48**	**91.65**

Table 2: Effect of input image resolution on Inception-V3 baseline model, measured by mAP and re-identification rate. We observe that higher input image resolution and fine-tuning can provide considerable performance gain. Small, medium and large models are respectively indicated as Model-S, Model-M and Model-L.

and Model-L are trained on ∼111K images of ∼17K identities when we merge 10 different person re-identification datasets. Since training on high resolution images is computationally expensive, in order to further push the performance boundaries, we take a trained Model-L and fine-tune it with input images of 748×246 which is ∼1.5 times larger than what Model-L has been originally trained with. Table 2 shows that such a fine-tuning practice, denoted as Model-L^{ft}, yields an average of **4.75**% mAP, and **1.71**% rank-1 score on the top of Model-L. Hence, we confirm the advantages of training person re-identification models using large input image sizes.

Choice of re-identification backbone architecture: Table 3 shows the effect of varying the re-identification backbone architecture in our baseline model. Inception-V3 [37] despite its considerably shallower architecture, provides a very competitive performance with ResNet-152 [16], while significantly outperforming ResNet-50 [16], which is of approximately the same depth. Table 3 also shows that the performance gain achieved by fine-tuning on high resolution images (ref. Table 2) is valid across variety of the architecture choices. In our experiments, we observe

Market-1501			
model	mAP(%)	rank-1	rank-10
Inception-V3	**73.06**	**88.87**	**97.00**
ResNet-50	66.32	85.10	95.75
ResNet-152	72.95	88.33	96.88
Inception-V3^{ft}	76.56	**90.80**	**97.71**
ResNet-50^{ft}	72.97	87.92	96.76
ResNet-152^{ft}	**77.96**	90.71	97.65
CUHK03			
model	mAP(%)	rank-1	rank-10
Inception-V3	–	87.91	98.41
ResNet-50	–	85.88	99.19
ResNet-152	–	**88.01**	**99.27**
Inception-V3^{ft}	–	88.73	98.94
ResNet-50^{ft}	–	89.08	99.15
ResNet-152^{ft}	–	**90.38**	**99.46**
DukeMTMC-reID			
model	mAP(%)	rank-1	rank-10
Inception-V3	59.87	79.08	90.26
ResNet-50	54.77	73.70	88.02
ResNet-152	**62.42**	**79.62**	**90.80**
Inception-V3^{ft}	63.27	80.48	91.65
ResNet-50^{ft}	59.72	77.74	90.84
ResNet-152^{ft}	**67.02**	**83.26**	**92.95**

Table 3: Effect of backbone architecture in our baseline person re-identification model, measured by mAP and re-identification rate.

that ResNet-152 is 3 times more computationally expensive (measured by forward+backward time) than Inception-V3. Hence, given their relatively similar performance, we chose Inception-V3 as our main backbone architecture.

SPReID Performance: Table 4 compares the performance of our proposed SPReID against the Inception-V3 baseline person re-identification. All the models are trained using the settings detailed in Section 4.2. We observe that both with and without foreground variations, respectively denoted as SPReID$^{w/fg}$ and SPReID$^{wo/fg}$, outperform Inception-V3 baseline while their combination (ℓ_2-normalization+concatenation) results in further performance gains. Exploiting human semantic parsing through SPReID improves the baseline re-identification model on: Market-1501 [48] by **6.61**% in mAP and **2.58**% in rank-1, CUHK03 [24] by **3.33**% in rank-1 and DukeMTMC-reID [50] by **8.91**% in mAP and **4.22**% in rank-1. Since the only difference between Inception-V3 baseline and SPReID is

Market-1501			
model	mAP(%)	rank-1	rank-10
Inception-V3	73.06	88.87	97.00
SPReID$^{w/fg}$	78.66	90.97	97.71
SPReID$^{wo/fg}$	78.06	90.74	97.80
SPReID^{combined}	**79.67**	**91.45**	**98.1**
CUHK03			
model	mAP(%)	rank-1	rank-10
Inception-V3	–	87.91	98.41
SPReID$^{w/fg}$	–	89.57	99.19
SPReID$^{wo/fg}$	–	**91.29**	98.93
SPReID^{combined}	–	91.21	**99.2**
DukeMTMC-reID			
model	mAP(%)	rank-1	rank-10
Inception-V3	59.87	79.08	90.26
SPReID$^{w/fg}$	67.20	82.32	92.32
SPReID$^{wo/fg}$	67.11	82.14	92.24
SPReID^{combined}	**68.78**	**83.3**	**92.91**

Table 4: Effect of utilizing human semantic parsing by SPReID to improve Inception-V3 person re-identification baseline, measured by mAP and re-identification rate.

Market-1501			
model	weight sharing	mAP(%)	rank-1
SPReID$^{w/fg}$	N	78.66	90.97
SPReID$^{w/fg}$	Y	77.62	90.88
SPReID$^{w/fg\text{-}ft}$	N	**80.68**	**92.40**
SPReID$^{w/fg\text{-}ft}$	Y	80.54	92.34
CUHK03			
model	weight sharing	mAP(%)	rank-1
SPReID$^{w/fg}$	N	–	89.57
SPReID$^{w/fg}$	Y	–	87.69
SPReID$^{w/fg\text{-}ft}$	N	–	**92.57**
SPReID$^{w/fg\text{-}ft}$	Y	–	89.68
DukeMTMC-reID			
model	weight sharing	mAP(%)	rank-1
SPReID$^{w/fg}$	N	67.20	82.32
SPReID$^{w/fg}$	Y	65.66	81.73
SPReID$^{w/fg\text{-}ft}$	N	**69.79**	**84.02**
SPReID$^{w/fg\text{-}ft}$	Y	69.29	83.80

Table 5: Effect of weight sharing in Inception-V3 backbone between global average pooling, and semantic based pooling branches of SPReID person re-identification architecture.

in how they aggregate the activations of the final convolution layer, we can confirm the advantage of our proposed method in effectively harnessing human semantic parsing to improve person re-identification.

Effect of weight sharing: SPReID model illustrated in Figure 1 has two aggregation heads. One simply performs global average pooling while the other uses probability maps associated to different human body parts as weights to aggregate convolutional activations. Table 5 compares two scenarios based on whether or not the two aggregation heads share the re-identification backbone. We observe that while exclusive backbone achieves slightly better results than weight sharing, with the exception of CUHK03 [24] the margin shrinks after fine-tuning on very high image resolutions. It is worth noting that in both scenarios, SPReID outperforms Inception-V3 baseline (ref. Table 4).

4.4. Comparison with the state-of-the-art

Table 6 shows the performance of our person re-identification models against the current state-of-the-art. For each dataset, the corresponding results are divided into three blocks, first one shows the performance of current state-of-the-art methods. Second block shows the performance of our baseline models with no human semantic parsing cues but trained using our two-stage training procedure. The third block shows the performance of SPReID.

From Table 6, we observe that the baseline person re-identification models when trained using our proposed training procedure outperform the current state-of-the-art. These results are particularly interesting, since the models are less complex and are also trained in a straightforward fashion. When utilizing re-ranking [51], the improvement margin further increases. Therefore, we confirm that a simple model with no bells and whistles is sufficient to achieve state-of-the-art person re-identification performance. Table 6 shows that SPReID can effectively harness local visual cues from human body parts. On all three datasets, SPReID$^{combined\text{-}ft}$ outperforms Inception-V3^{ft} baseline with a large margin. Although, the gap reduces when models are combined with the strong ResNet-152^{ft} baseline. Similar to the previous case, the performance will be further improved by employing re-ranking as post processing.

5. Implementation Details

Person Re-identification: In both training phases, mini-batch size is set to 15, momentum to 0.9 and we use weight decay and gradient clipping with 0.0005 and 2.0 for the respective values. Initial learning rate value is set to 0.01 in the first phase and reduces to 0.001 in the second phase.

Market-1501				
method	mAP(%)	rank-1	rank-5	rank-10
Li *et. al.* [21]	57.5	80.3	–	–
SVDNet [36]	62.1	82.3	92.3	95.2
DPAR [46]	63.4	81.0	92.0	94.7
JLML [25]	65.5	85.1	–	–
Basel.+LSRO [50]	66.1	84.0	–	–
SSM [3]	68.8	82.2	–	–
DaF [44]	72.4	82.3	–	–
Chen *et. al.* [10]	73.1	88.9	–	–
Inception-V3^{ft}	76.56	90.8	96.35	97.71
Inception-V3^{ft*}	82.87	93.14	97.27	98.22
+re-ranking[51]	90.66	94.21	96.76	97.3
SPReID$^{combined\text{-}ft}$	81.34	92.54	97.15	98.1
SPReID$^{combined\text{-}ft*}$	83.36	93.68	**97.57**	**98.4**
+re-ranking[51]	**90.96**	**94.63**	96.82	97.65
CUHK03				
method	mAP(%)	rank-1	rank-5	rank-10
FT-JSTL+DGD [40]	–	75.3	–	–
SSM [3]	–	76.6	94.6	98
Spindle [45]	–	88.5	97.8	98.6
DPAR [46]	–	85.4	97.6	99.4
Chen *et. al.* [10]	82.8	86.7	–	–
HydraPlus [27]	–	91.8	98.4	99.1
Inception-V3^{ft}	–	88.73	97.82	98.94
Inception-V3^{ft*}	–	92.81	98.9	99.35
+re-ranking[51]	–	95.18	99.18	99.6
SPReID$^{combined\text{-}ft}$	–	93.89	98.76	99.51
SPReID$^{combined\text{-}ft*}$	–	94.28	99.04	99.56
+re-ranking[51]	–	**96.22**	**99.34**	**99.7**
DukeMTMC-reID				
method	mAP(%)	rank-1	rank-5	rank-10
Basel.+LSRO [50]	47.1	67.7	–	–
Basel.+OIM [42]	–	68.1	–	–
ACRN [33]	52.0	72.6	84.8	88.9
SVDNet [36]	56.8	76.7	86.4	89.9
Chen *et. al.* [10]	60.6	79.2	–	–
Inception-V3^{ft}	63.27	80.48	88.78	91.65
Inception-V3^{ft*}	72	85.37	92.15	94.21
+re-ranking[51]	84.82	**89.41**	93.18	**94.75**
SPReID$^{combined\text{-}ft}$	70.97	84.43	91.88	93.72
SPReID$^{combined\text{-}ft*}$	73.34	85.95	92.95	94.52
+re-ranking[51]	**84.99**	88.96	**93.27**	**94.75**

Table 6: Comparison with the state-of-the-art. *indicates combination (ℓ_2-normalization+concatenation) with ResNet-152^{ft}.

Throughout training, we decay the learning rate 10 times using exponential shift with the rate of 0.9. We train the models using Nesterov Accelarated Gradient [5] and initialize the weights using pre-trained models on ImageNet [32].

Human Semantic Parsing: We train our human semantic parsing model for 30K iterations where the initial learning rates for the Inception-V3 backbone, atrous spatial pyramid pooling and the 1×1 convolution layer are respectively set to 0.01, 0.1 and 0.1. The rest of the parameters and settings are similar to the ones for person re-identification except the input resolution where 512×512 input images are used.

6. Conclusion

In this paper, we began by raising two major questions. First, whether to achieve state-of-the-art performance, the person re-identification models need to be *complex*. Second, whether bounding boxes on human body parts is the best practice to harness local visual cues. Through this paper, we addressed both of these questions with extensive set of experiments. We showed that, indeed a *simple* deep convolutional architecture when trained properly on large number of high resolution images can outperform the current state-of-the-art. We also demonstrated that, by exploiting human semantic parsing in our proposed SPReID framework, the performance of an state-of-the-art baseline model can be further improved. SPReID applies minimal modifications to the person re-identification backbone and offers a more natural solution for utilizing human body parts. We hope that, this work encourages the research community to invest more in employing human semantic parsing for person re-identification task.

Acknowledgments

This research is based upon work supported in parts by the U. S. Army Research Laboratory and the U. S. Army Research Office (ARO) under contract/grant number W911NF-14-1-0294; and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. D17PC00345. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Emrah Basaran was supported by 2214-A programme of The Scientific and Technological Research Council of Turkey (TÜBİTAK).

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015. 3

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015. 5

[3] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 8

[4] D. Baltieri, R. Vezzani, and R. Cucchiara. 3dpes: 3d people dataset for surveillance and forensics. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 59–64. ACM, 2011. 5

[5] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu. Advances in optimizing recurrent networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8624–8628. IEEE, 2013. 8

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 3, 5

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[8] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2016. 5

[9] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. *arXiv preprint arXiv:1704.01719*, 2017. 3

[10] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2590–2600, 2017. 8

[11] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. 2, 3

[12] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 3

[13] 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. 4

[14] K. Gong, X. Liang, X. Shen, and L. Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. *arXiv preprint arXiv:1703.05446*, 2017. 5

[15] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 1–7. Citeseer, 2007. 5

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 4, 6

[17] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person Re-Identification by Descriptive and Discriminative Classification. In *Proc. Scandinavian Conference on Image Analysis (SCIA)*, 2011. 5

[18] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[19] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 2, 3

[20] Y. Kawanishi, Y. Wu, M. Mukunoki, and M. Minoh. Shinpuhkan2014: A multi-camera pedestrian dataset for tracking people across multiple cameras. In *20th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, volume 5, page 6, 2014. 5

[21] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 384–393, 2017. 2, 8

[22] W. Li and X. Wang. Locally aligned feature transforms across views. In *CVPR*, 2013. 5

[23] W. Li, R. Zhao, and X. Wang. Human reidentification with transferred metric learning. In *ACCV*, 2012. 5

[24] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 1, 2, 3, 4, 5, 6, 7

[25] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017. 8

[26] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017. 2

[27] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2017. 8

[28] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 5

[29] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2016. 3

[30] A. Rahimpour, L. Liu, A. Taalimi, Y. Song, and H. Qi. Person re-identification using visual attention. *arXiv preprint arXiv:1707.07336*, 2017. 2

[31] 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*, 2016. 4

[32] 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. 8

[33] A. Schumann and R. Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1435–1443. IEEE, 2017. 3, 8

[34] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Pose-driven deep convolutional model for person re-identification. *arXiv preprint arXiv:1709.08325*, 2017. 2

[35] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 3

[36] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 8

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 2, 3, 4, 6

[38] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European Conference on Computer Vision*, pages 791–808. Springer, 2016. 3

[39] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016. 3

[40] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1249–1258, 2016. 8

[41] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2016. 5

[42] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proc. CVPR*, 2017. 8

[43] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[44] R. Yu, Z. Zhou, S. Bai, and X. Bai. Divide and fuse: A re-ranking approach for person re-identification. *BMVC*, 2017. 8

[45] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017. 2, 3, 8

[46] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 8

[47] L. Zheng, Y. Huang, H. Lu, and Y. Yang. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*, 2017. 2

[48] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 1, 2, 4, 5, 6

[49] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. *TOMM*, 2017. 3

[50] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2, 4, 5, 6, 8

[51] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. *arXiv preprint arXiv:1701.08398*, 2017. 7, 8

[52] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian. Part-based deep hashing for large-scale person re-identification. *IEEE Transactions on Image Processing*, 2017. 2