Self-Critical Attention Learning for Person Re-Identification

Deep learning research project Report

Represented by:

1- Bilal Mohamed Elhlwany-20398551
2- Ahmed Mahmoud-20398552
3- Yara Hassan-20398563

Under supervision of:
Prof: Hazem Abass
Eng: Asif Mahfuz

1. What is the problem statement of the paper?

The problem statement of the paper is to propose a -self-critical attention learning (SCAL)- method for person re-identification (ReID).

Person re-identification is a challenging problem that aims to match and return a specified person from a large-scale gallery set collected by multiple non-overlapping camera views deployed at different locations. But this is still a challenging problem due to the difficulty of visual features matching with the illumination changes, pose variations, occlusions, and cluttered backgrounds. Also, the paper addresses the limitations of existing methods that train the attention mechanism in a weakly-supervised manner and ignore the attention confidence level.

2. What are the objectives of the paper and do you think the authors managed to achieve these goals? Explain

1) One of the most important objectives of the paper is introducing a self-critical attention learning (SCAL) method and show its ability to improve the accuracy, robustness and the performance of human pattern recognition of person REID systems.
2) Demonstrate not only the effectiveness of the SCAL method on several benchmark datasets but also compare its performance to other state-of-the-art methods including but not limited to SVDNet, CamStyle, Pose-transfer, DaRe.

In my opinion, the authors successfully achieved their goals since:

1) The Self-Critical Attention Learning (SCAL) method proposed in the paper has shown to be highly effective and outperforms many other deep learning methods for person re-identification on several benchmark datasets including Market-1501, DukeMTMC-reID, CUHK03 by a large margin.
2) The paper presents an effective approach for person re-identification that has the potential to advance the state-of-the-art in the field of person REID.
3) It confirms the effectiveness of the attention evaluator and the self-critical supervisory signal.
4) I also think the paper could address some potential limitations or challenges of their method, such as the computational cost, the generalization ability, and the robustness to noise or occlusion.

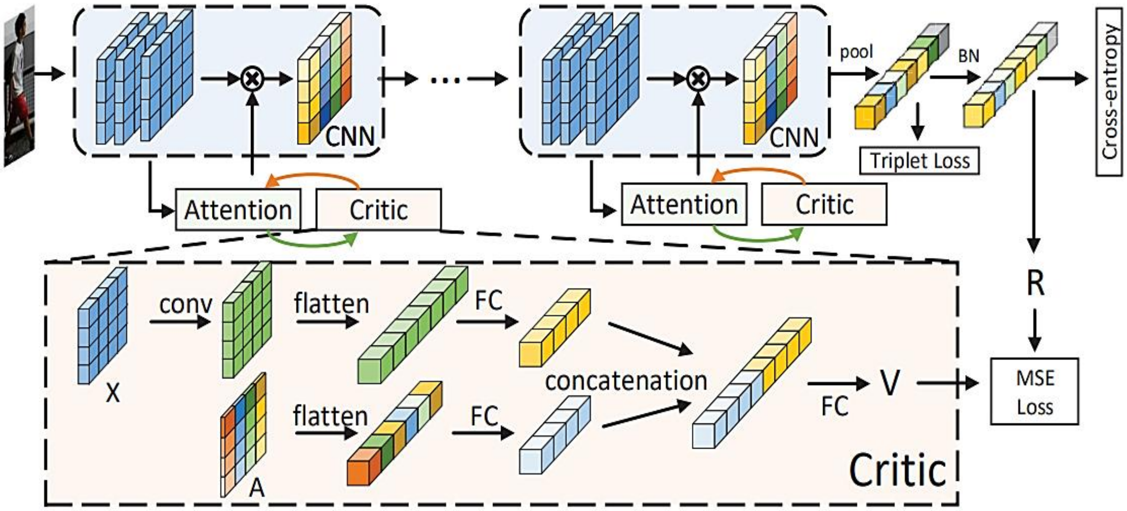
Overall, the paper offers a new and effective approach to identity rediscovery that addresses some of the key challenges in the field.

3. What is the DL method used in this paper?

- **The self-critical attention learning (SCAL) method**

The method helps in improving person re-identification performance, it’s based on **deep learning especially (**It is based on **attention mechanism)** and is used to learn a similarity metric between pairs of images. How it works?

- It improves the learning process, permitting to fully exploit the effectiveness of attention. Instead of the weakly-supervised manner.
- The attention model evaluates itself and guide the optimization with evaluation performance.
- It consists of a convolutional backbone network, an attention agent, and a critic module. The backbone consists of a series of convolutional blocks, where we encode the attention maps on top of each block. The critic as an important component of the attention model, takes the feature maps X and the attention map A as input, and outputs the critic value V as the additional supervisory signal of attention learning.



[1]

Why SCAL Method is based on attention mechanism what is the benefits from it?

- Since most effective evaluation indicators are usually non-differentiable, the gain of attention model over the basic network, self-critical attention model is optimized by reinforcement learning algorithms. Specifically, the state is the input image, and the agent is the attention model that predicts the attention maps based on the current state. The critic takes the state and attention as input and evaluates the quality of the attention model.

The steps of implementation of SCAL method in the paper:

- The authors of the paper employed the ResNet-50 as the basic backbone network for The SCAL method.
- And they initialized them with the ImageNet pre-trained parameters in order to preserve the resolution of the image.
- They applied a convolution layer with stride = 1 and stacked five attention models on the ResNet-50 network which are placed on top of the first convolution layer of the network and the output layer of each residual block.
- They employed three data augmentation methods, including random cropping, horizontal flipping, and erasing.
- They evaluate the SCAL on three large-scale datasets including Market-1501, DukeMTMC-ReID and CUHK03.
- For optimization the authors use triplet loss and classification loss. And the triplet loss function aims to preserve the rank relationship among a triplet of samples with a large margin, which increases the inter-class distance and reduces the intra-class one

4. What are the other state-of-the-art methods that can be applied to the same problem?

According to the mentioned in the paper the authors claim that SCAL outperforms previous methods on several benchmark datasets, such as Market1501, DukeMTMC-reID, and CUHK03.

But the date in which the paper Added to IEEE Xplore: 27 February 2020 [4]

However, there are also other recent methods that have achieved competitive or even better results on person Re-ID such as:

1. A paper published in June 2021.The paper presents a novel locally aware network and blockwise fine-tuning techniques for person re-identification. The proposed method is called LA-Transformer with blockwise fine-tuning and it achieves rank-1 accuracy of 94.8% with standard deviation of 0.5% on the Market-1501 dataset and rank-1 accuracy of 89.6% with standard deviation of 0.7% on the CUHK03 dataset respectively, outperforming SCAL and other state-of-the-art methods. [2]
2. A paper published in January 2022 proposes a novel gait recognition method, called RealGait, that extracts silhouettes from an existing video person re-identification challenge and uses them as biometric identifiers.
   - Both papers use the same datasets (Market-1501 and DukeMTMC-reID) for evaluation, which makes the comparison fair and consistent.
   - Both papers achieve competitive or better results than other methods on both datasets, which demonstrates their effectiveness and robustness for person re-identification.
   - The SCAL paper outperforms the paper this one on Market-1501 (94.5% vs 93.6%), but slightly underperforms on DukeMTMC-reID (86.4% vs 88.7%). This suggests that there is no clear winner between these two methods, and their performance may depend on the characteristics of the datasets or the evaluation protocols. [3]

To summarize, these two papers propose different methods for person re-identification, and they achieve similar results on two common datasets. However, there is no conclusive evidence to claim that one method is superior to the other, and more experiments and analysis are needed to compare them in depth.

5. Would you apply any of the other methods other than the DL method used in this paper? Explain your answer?

Yes, I would apply other methods for person re-identification. Deep learning-based person Re-ID methods have emerged and some of these methods include Deep Learning for Person Re-Identification and some of these methods are not based on deep learning:

Other methods for person re-identification that are not based on deep learning:

"Person Re-Identification by Local Maximal Occurrence Representation and Metric Learning": This method uses a local maximal occurrence (LOMO) feature descriptor to represent the appearance of a person and learns a metric to measure the similarity between two LOMO features. This method has been shown to be effective in handling pose variations and occlusions. (Related work)

"Person Re-Identification by Matching Compositional Parts via Joint Convolutional Forests": This method uses a joint convolutional forest (JCF) to learn the appearance of different parts of a person's body and then matches these parts across different cameras. This method has been shown to be effective in handling changes in viewpoint and illumination.

LMNN (Large Margin Nearest Neighbor) is a metric learning algorithm that has been used for person re-identification. It learns a Mahalanobis distance metric that maximizes the distance between samples from different classes while keeping samples from the same class close together. This method has been shown to be effective in handling small training sets and noisy data. (Related work)

kernel-based metric learning methods have been used for person re-identification. These methods learn a distance metric between samples by mapping them into a high-dimensional feature space using a kernel function. This method has been shown to be effective in handling complex variations in human appearances from different camera views. (Related work)

Other methods for person re-identification that based on deep learning:

Generative adversarial networks (GANs) have been used for person re-identification to generate new samples and improve the performance of person re-identification1. GAN-based person re-ID adopts adversarial learning where the generator and discriminator can interact with each other in the process of adversarial learning, which are conducive to improving the recognition performance of person re-ID.

Part-aligned is a method for person re-identification that uses part-based feature representations to improve the performance of person re-identification. The method divides the image into several parts and learns a feature representation for each part. The method has been shown to achieve state-of-the-art performance on several benchmark datasets. (In the paper)

6. What datasets have been used in this paper? Do you think the result is generalizable for any datasets?

According to the mentioned in the paper, the datasets that have been used for person Re-ID are Market-1501, DukeMTMC- ReID, and CUHK03. These datasets are widely used benchmarks for person Re-ID that contain images of pedestrians captured by multiple cameras with different viewpoints and backgrounds. The authors trained their model with the data in the Market-1051 dataset and tested it with the samples in the DukeMTMC-ReID dataset, and vice versa.

The paper reports the performance of SCAL on these datasets using metrics such as rank1 accuracy and mean average precision (mAP). However, the generalizability of the results to other datasets may depend on several factors:

- The size and diversity of the datasets
- The quality and resolution of the images
- The degree of occlusion and illumination variation.
- The similarity or difference between the training and testing domains

The paper does not provide any analysis or discussion on these aspects, nor does it compare SCAL with other methods on more challenging or realistic datasets, such as MSMT17 or PRW. Therefore, it is hard to say how well SCAL would perform on any datasets without further experiments and evaluations.

Note:

PRW is a large-scale dataset for end-to-end pedestrian detection and person recognition in raw video frames.

MSMT17 is a multi-scene multi-time person re-identification dataset. The dataset consists of 180 hours of videos, captured by 12 outdoor cameras, 3 indoor cameras, and during 12 time slots. The videos cover a long period of time and present complex lighting variations, and it contains a large number of annotated identities, i.e., 4,101 identities and 126,441 bounding boxes.

7. Discuss the results presented in the paper. Compare the results with other state-of-the art methods used to solve this problem.

- **In the Table 2, Table 3, and Table 4 all consist of three parts:**
  a) The top groups **b)** The bottom group **c)** Their approach method
- **In the top groups**, the authors compared their approach against the state-of-the-art methods (13 conventional deep learning methods) on the Market-1501, DukeMTMC-ReID, and CUHK03 datasets.
- **In the bottom group**, summarizes the performance of deep learning methods with attention model (9 attention-based methods).
- It's obvious from the tables that the proposed SCAL methods on both spatial and channel domain achieve superior performance over all comparing methods substantially on the three benchmarks.
- **The authors discuss the results on three datasets individually:**

For Market-1051 dataset (Table 2)

The channel-based SCAL with ResNet-50 achieved state-of-the-art results of mAP/Rank-1 = 89.3%/95.8%, outperforming SPReID by +5.9% on mAP and +2.1% on Rank-1. (Top group)

Although the attention-based methods have achieved great performance recently, the proposed attention model with self-critical outperforms them by a large margin, 7% on mAP and 2.7% on Rank-1. (Bottom group)

For DukeMTMC-ReID dataset (Table 3)

They outperformed the second-best method SPReID substantially by 6.3% and 3.0% respectively on the mAP score and Rank-1 accuracy.

For CUHK03 dataset (Table 4)

For both labeled and detected settings, the proposed SCAL achieved the improvement by a large margin (12.1% on mAP and 10.3% on Rank-1 in the labeled version; 10.5% on mAP and 9.5% on Rank-1 in the detected version) over the best alternative DaRe method with the same ResNet-50 base-model.

Note:

mAP stands for mean Average Precision, which is a commonly used performance metric in object detection and person re-identification tasks.

Rank-1 refers to the accuracy of the model in correctly identifying the most similar image to a given query image.

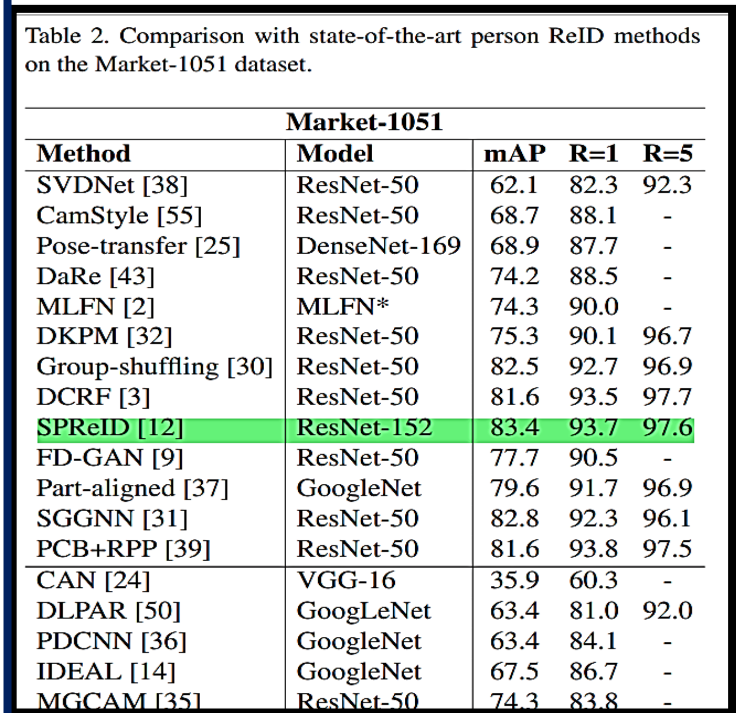


Table 2. Comparison with state-of-the-art person ReID methods on the Market-1051 dataset.

Market-1051				
Method	**Model**	**mAP**	**R=1**	**R=5**
SVDNet [38]	ResNet-50	62.1	82.3	92.3
CamStyle [55]	ResNet-50	68.7	88.1	-
Pose-transfer [25]	DenseNet-169	68.9	87.7	-
DaRe [43]	ResNet-50	74.2	88.5	-
MLFN [2]	MLFN*	74.3	90.0	-
DKPM [32]	ResNet-50	75.3	90.1	96.7
Group-shuffling [30]	ResNet-50	82.5	92.7	96.9
DCRF [3]	ResNet-50	81.6	93.5	97.7
SPReID [12]	ResNet-152	83.4	93.7	97.6
FD-GAN [9]	ResNet-50	77.7	90.5	-
Part-aligned [37]	GoogleNet	79.6	91.7	96.9
SGGNN [31]	ResNet-50	82.8	92.3	96.1
PCB+RPP [39]	ResNet-50	81.6	93.8	97.5
CAN [24]	VGG-16	35.9	60.3	-
DLPAR [50]	GoogLeNet	63.4	81.0	92.0
PDCNN [36]	GoogleNet	63.4	84.1	-
IDEAL [14]	GoogleNet	67.5	86.7	-
MGCAM [35]	ResNet-50	74.3	83.8	-

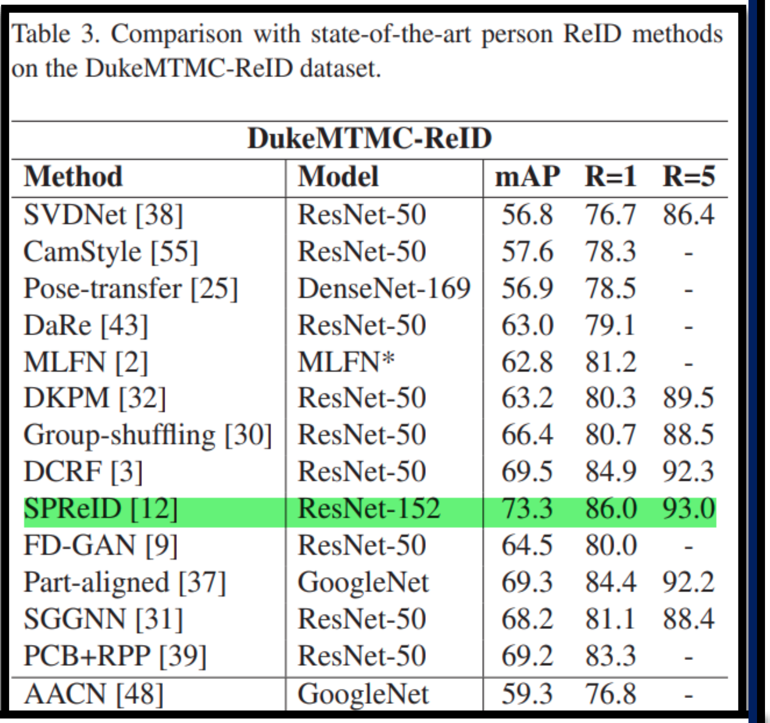


Table 3. Comparison with state-of-the-art person ReID methods on the DukeMTMC-ReID dataset.

DukeMTMC-ReID				
Method	**Model**	**mAP**	**R=1**	**R=5**
SVDNet [38]	ResNet-50	56.8	76.7	86.4
CamStyle [55]	ResNet-50	57.6	78.3	-
Pose-transfer [25]	DenseNet-169	56.9	78.5	-
DaRe [43]	ResNet-50	63.0	79.1	-
MLFN [2]	MLFN*	62.8	81.2	-
DKPM [32]	ResNet-50	63.2	80.3	89.5
Group-shuffling [30]	ResNet-50	66.4	80.7	88.5
DCRF [3]	ResNet-50	69.5	84.9	92.3
SPReID [12]	ResNet-152	73.3	86.0	93.0
FD-GAN [9]	ResNet-50	64.5	80.0	-
Part-aligned [37]	GoogleNet	69.3	84.4	92.2
SGGNN [31]	ResNet-50	68.2	81.1	88.4
PCB+RPP [39]	ResNet-50	69.2	83.3	-
AACN [48]	GoogleNet	59.3	76.8	-

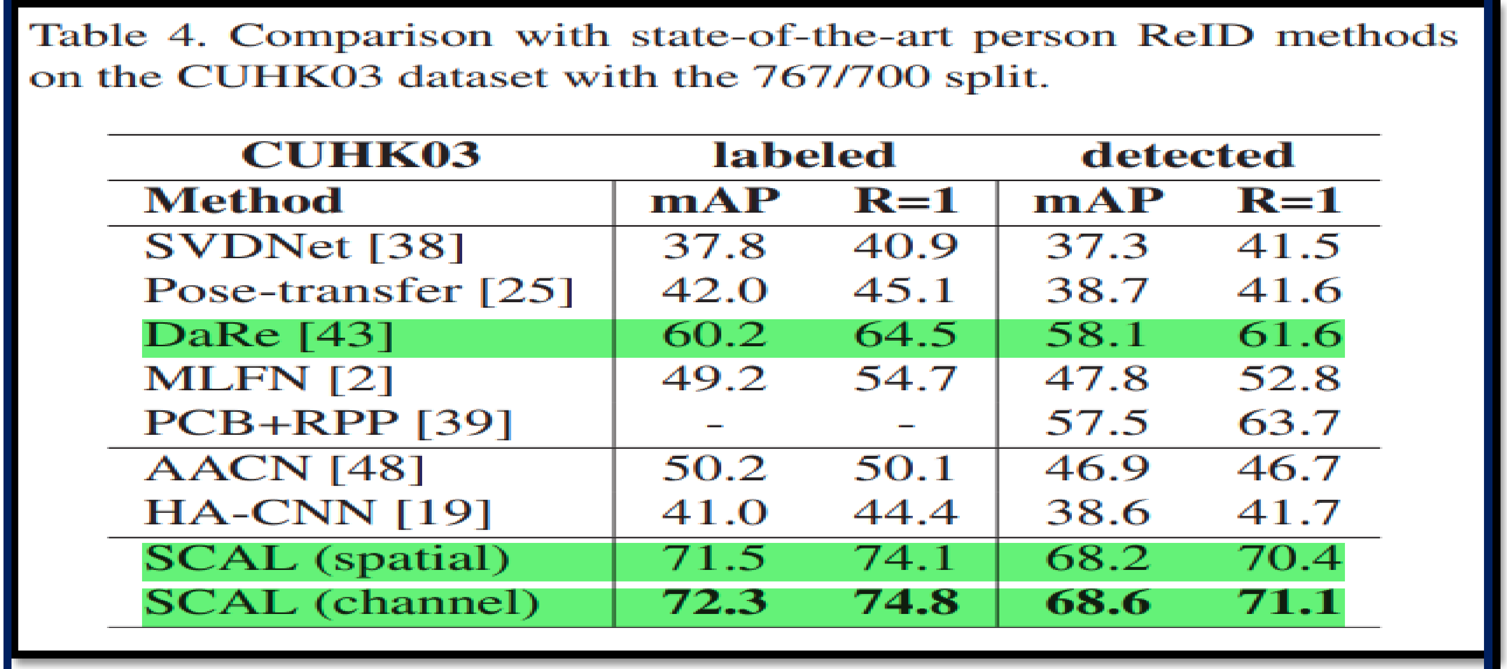


Table 4. Comparison with state-of-the-art person ReID methods on the CUHK03 dataset with the 767/700 split.

CUHK03	labeled		detected	
Method	**mAP**	**R=1**	**mAP**	**R=1**
SVDNet [38]	37.8	40.9	37.3	41.5
Pose-transfer [25]	42.0	45.1	38.7	41.6
DaRe [43]	60.2	64.5	58.1	61.6
MLFN [2]	49.2	54.7	47.8	52.8
PCB+RPP [39]	-	-	57.5	63.7
AACN [48]	50.2	50.1	46.9	46.7
HA-CNN [19]	41.0	44.4	38.6	41.7
SCAL (spatial)	71.5	74.1	68.2	70.4
SCAL (channel)	**72.3**	**74.8**	**68.6**	**71.1**

8. What would you like to criticize about the paper? Could you suggest any improvements?

AS a team We think the paper is well-written and presents an interesting approach to person re-identification also presents an effective method for person Re-ID. However, we also have some criticisms and suggestions for improvement such as:

- The paper could have been improved by providing more detailed explanations of the attention mechanism and how it works in practice.
- The paper could have included more experiments to test the effectiveness of the deep learning based SCAL method.
- It was mentioned in the paper that the self-critical attention learning method outperforms existing state-of-the-art methods by a large margin in accuracy but the paper has not any visualization to these accuracy comparisons or even a link to a code that show the implementation of the visualization of these comparisons to validate them. So, the paper could have included these visualization of the accuracy (mAP score and Rank-1 accuracy).
- The paper does not provide any qualitative results or visualizations on the critic module or the attention quality estimation of the SCAL Method. It would be useful to see how the critic module evaluates the attention maps and how it guides the learning process of SCAL.
- To evaluate the generalizability of SCAL, the authors or the programmers who works on this paper could conduct experiments on more diverse and challenging datasets, such as MSMT17 or PRW which are known to be challenging for person re-identification tasks and have larger scale, more cameras, more occlusions, and more background clutter. These datasets and would provide a good benchmark for evaluating the performance of the SCAL method.

References:

- https://linchunze.github.io/papers/ICCV19_self-critical_attention.pdf [1]
- https://arxiv.org/pdf/2106.03720.pdf[2]
- 2201.04806v1.pdf (arxiv.org)[3]
- Self-Critical Attention Learning for Person Re-Identification | IEEE Conference Publication | IEEE Xplore[4]