End to End Person Re-Identification Based on Attention Mechanism

Yang Li¹,²,*, Huahu Xu¹,³, and Minjie Bian¹,³

¹ School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China
² School of Information Technology, Shanghai Jianqiao University, Shanghai 201306, China
³ Shanghai Shangda Hairun Information System Co., Ltd, Shanghai 200444, China

*E-mail: liyang162@shu.edu.cn

Abstract. This paper treats person re-identification (re-id) as a sequential model, guide person re-id with person detection, combines recurrent neural network (RNN) with attention mechanism, and proposed an end to end person re-id method for surveillance scenarios. The feature of target person is firstly extracted using ResNet, and then the feature is added to the long short-term memory (LSTM) network to guide the attention model for the region of interest in the surveillance image. Finally, combined with the information observed from the image multiple times, the most similar candidate person in the image is deduced, and the feature distance is calculated and ranked for person re-id. This paper strengthens the relationship between person detection and person re-id, and reduces the error between models. Because the number of candidate person matched with target person is reduced, this method can process person re-id task with less calculation and time. This paper also verified the effectiveness of the proposed method by experiments comparing a variety of person detection and re-id methods on several person re-id datasets.

1. Introduction
Person re-identification (re-id) aims at matching the target person from a series of person images through the appearance of persons. It can be used in many fields, such as video surveillance, suspect tracking, multi-camera person tracking, and person activity analysis. Due to the complex changes of pose, viewpoint, light, occlusion, resolution, background clutter and so on, person re-id is challenging and has attracted wide attention in recent years, and has become a rapidly developing research field [1].

Current person re-id algorithms mainly study how to match the target person image with the candidate person images which has been cropped and extracted from the complete scene image. However, in practical application scenarios, the candidate person images are usually not cropped from the complete scene images, so it is necessary to identify the target person from the complete scene images. Although person detection and person re-id are closely integrated tasks in many scenarios, but they are usually studied as two separate tasks, and the method of combining the two is not yet popular. In order to identify specific target person in surveillance scenarios, a natural method is to detect all candidate persons from the images, then match the search targets one by one, and finally select the matching targets.

In this paper, a new method of person detection and re-id based on attention mechanism is
proposed, which extracts the feature of the target person as the initial memory information and pass to the person detection and re-id network to guide the process of retrieve the target person in the image. Compared with the previous methods, this paper combines the detection and re-id process more closely as an end to end solution. The introduction of attention based memory information makes the detection process more purposeful and reduces the number of re-id calculations, thus reducing the amount of calculation in the whole process.

2. Related Work

**Pedestrian detection.** Hand-crafted features are widely used in early years, the HOG descriptor [2] is based on local image differences, and successfully represents the special head-shoulder shape of pedestrians. A deformable part model (DPM) [3] is proposed to handle deformations and still uses HOG as basic features. More recently, the integral channel feature (ICF) detectors [4,5,6] become popular, as they achieve remarkable improvements while running fast. In recent years, convolutional neural networks (CNN) are used in pedestrian detection and make good progress [7,8,9,10]. Some works use the R-CNN architecture, which relies on ICF for proposal generation [7,8]. Aiming for an end-to-end procedure, Faster R-CNN[11] is adopted and it achieves top results by applying proper adaptations [12,13].

**Person re-id.** Early person re-id mainly use manually designing discriminative features [14,15], using salient regions [16] and learning distance metrics [17,18]. Most of these CNN-based methods can be divided into two categories. The first category uses the siamese model with image pairs [19,20,21,22] or triplets [23,24] as inputs. The main idea of these works is to minimize the feature distance between the same persons and maximize the feature distance between different persons. The second category regards person re-id task as a classification problem [25,26,27]. The main drawback of classification models is that they require more training data. Xiao et al. [25] propose to combine multiple datasets for training and improve feature learning by domain guided dropout. Zheng et al. [26,27] point out that classification models are able to reach higher accuracy than siamese model, even without careful sample choosing. Recently, attention mechanism [21,28,29,30] has been adopted to learn better features for person re-id.

**Detection and re-id.** Xu et al. [31] jointly model pedestrian commonness and uniqueness, and calculate the similarity between query and each sliding window in a rude manner. Zheng et al. [27] studied the effect of detection on re-id by combining detection with re-id using different methods. Other works propose to solve the problem in an end-to-end fashion by employing the Faster R-CNN detector [11] for pedestrian detection and share the base network between detection and re-id [1]. In [32], feature discriminative power is increased by introducing center loss during training. Liu et al. [33] improve the localization policy of Faster R-CNN by recursively shrink the search region from the whole image till achieving precise location of the target person. Xiao et al. [1] proposed a new pedestrian retrieval framework, in which detection and re-id were processed in a unified CNN. They proposed a new loss function Online Instance Matching (OIM) to train the network. In addition, they also collected and labelled a large-scale pedestrian image data set for person re-id tasks.

In this paper, the problem is treated as a sequential model. Using the information of detection to guide re-id, the recurrent neural network (RNN) and attention mechanism are combined to carry out person detection and re-id, which can reduce the computational load needed to be processed in a single cycle.

3. Methodology

3.1. Framework of Attention Based Algorithm

Traditional target detection algorithms usually deal with the whole image's visual information indiscriminately. But human beings usually will not and can't process all the information of an image in detail, just as when we are viewing a painting, although we can observe the whole painting, our attention is usually focused on a small part of the painting at a certain time period, that is, our attention
to the whole picture information is not balanced. Based on this idea, this paper uses the attention mechanism [34] to carry out person detection and re-id, which can focus on the selection of interested parts from the whole image for processing. Compared with traditional methods, this method can reduce the computational load in the whole detection and re-id process.

In this paper, the problem of person detection and re-id is regarded as a sequential model problem, which is processed by combining attention model with RNN and using a memory unit containing target person information. Suppose that we can only observe an image through an observation agent, and the observation agent can only observe a part of the image at each time point. After inputting the target person to be identified, the feature is extracted as the initial memory information of the target person. The initial memory is used to help the attention model select the next observation region. The final search result can be obtained by combining the information obtained from multiple observations. Just as in reading an article, we understand the current text according to the text information we have read before, rather than dividing the information independently, that is, people's thinking is continuous. We regard the whole person detection and re-id process as a continuous process, and the RNN is widely used to deal with sequence problems. This paper uses Long Short-Term Memory (LSTM) [35] network for the task.

This section introduces the overall framework of the algorithm. The whole person detection and re-id process consists of two key parts: memory unit and neural net-work. ResNet50 [36] Part1 is used to extract the features of the target person as memory information and passed to the person re-id network. The person detection and re-id network used in this paper is a LSTM recurrent neural network using attention model. At each step of the recurrent process, only a part of the whole image can be observed. The network is guided by memory information for the observation region of the whole image. This algorithm transforms the traditional two-step process into a simple process, which reduces the number of times of calculating and ranking the feature distance of recognition, and thus reduces the amount of calculation and processing time. The block diagram of the algorithm is shown in Figure 1.

![Figure 1. Block diagram of end to end attention based person re-identification algorithm.](image)
Figure 1 shows the flow of the whole algorithm. After input the image of a target person, the feature extraction is carried out through the ResNet50 Part1, and the feature expression Q of the target person is obtained and passed to the detection and re-id network using attention model. The detection and re-id network in this paper is based on Convolutional LSTM (ConvLSTM) [37]. The network can only observe the image through an observation agent with limited observation range, where \( l_i \) represents the coordinates of the observation position of step \( i \). The image and the current observation coordinate \( l_i \) constitute the observation region of step \( i \) in the cycle. The feature of current observation region is also extracted by the ResNet50 Part1 and passed to the ConvLSTM unit, the ConvLSTM unit combined with historical state \( h_{i-1} \) to generate the next state \( h_{i} \). \( h_{i} \) determines the next observation coordinate \( l_{i+1} \) through a positioning function \( f_1 \). The final feature vector \( h_{N} \) is obtained after a certain number of cycles. After feature extraction through ResNet50 part2, \( h_{N} \) and the target person feature \( Q \) are passed to the re-id network for cosine feature distance calculation and ranking for person re-identification.

3.2. Initial Memory Unit

In the traditional process for person re-id based on surveillance scene, all candidate pedestrian images are firstly detected and cropped from surveillance images by pedestrian detection, and then all the candidate images are sent to the re-id network for feature distance calculation and ranking with the comparison with the target person.

In such a two-stage process, the feature of the target person that need to be identified is not fully utilized in the pedestrian detection stage, and is used only when compared with the candidate images in the re-id stage. Therefore, all pedestrians detected from the surveillance images will be used as candidate pedestrians to participate in feature distance calculation and ranking. If the detection process can be guided by the features of the target person to be retrieved, the number of person re-id processes can be reduced if the object which is similar to the target person features can be found in the image at the detection stage. Firstly, we extract the features of the target person, which can be called initial memory. We put the initial memory into the detection and re-id network, and use the initial memory to guide the next region of attention of the observation agent. In order to achieve such a process, we design an initial memory unit and use the ResNet50 to extract the feature expression of the target person. In each step of the observation process of the recurrent neural network, the feature expression is passed to the LSTM unit to help the detection and re-id network select the next region to be observed.

In this paper, two networks are used. The RNN with convolutional LSTM unit is used as the core network for detection and re-id. The ResNet50 is used to extract person image features. The ResNet-50 is divided into two parts: lower layers are used to extract the features for pedestrian detection, while the higher layers are used to extract the features for person re-id.

3.3. Recurrent Attention Model

There are two ways to process visual information, bottom-up data-driven process and top-down task-driven process. At present, most of the visual attention models are based on bottom-up, namely data-driven visual attention model. Bottom-up data-driven models usually use the underlying visual features of an image, such as colour, brightness and direction, to generate saliency maps of multi-dimensional features of an image.

Although data-driven visual attention is widely used at present, when it is applied to object detection, the effect may not be satisfactory due to the interference of the underlying visual features of the environment itself. So top-down task-driven visual model is more effective. Task-driven visual attention model is a model with prior knowledge, more targeted. Task-driven visual attention model pre-establishes visual expectations, separates the desired objects from the original image, and thus completes the extraction of regions of interest in the image. Task-driven models generally pay attention to different objects in three aspects: scenario priori information, task requirements and object features. For example, in a complex scene to find a car, and the car has four wheels, so after adding
wheel features in the model, the model can quickly eliminate other interference information in the
detect process, and improve the efficiency of the detection. The processing method in this paper is
similar to task-driven visual attention model. Adding feature information of target person to detection
and re-identification tasks is similar to generating visual expectations.

In order to add the feature information of the target person to the process, we can regard the person
detection and re-identification task as a series of decision-making processes of observation agents with
specific tasks and interaction with visual environment. The observation agent can only observe a part
of the whole image at each time. The observation agent can extract information from a smaller region
and decide how to control its perception range, such as changing the location of the observation region.
Observation agents can also influence the real state of the environment by performing some actions.
Since the observation agent can only observe some parts of the image at one time, it needs to learn to
integrate information from different times to determine what actions to perform and where to observe
at the next time. After each decision, the observation agent receives a reward value. Its goal is to
obtain the maximum reward value. When the model is applied to the task of person detection and re-id,
the action is detect and re-identify the target person, and the reward signal reflects whether the action
is correct or not.

RNN is used in this paper to deal with above recurrent attention model, and LSTM is used to
improve the traditional RNN to solve the problem of gradient disappearance and gradient explosion. In
order to make more effective use of the information of the target person, the feature of the target
person are added to the modified ConvLSTM unit as an independent input information. ConvLSTM is
a processing structure specially designed for spatiotemporal sequence. The main disadvantage of
traditional LSTM is that it uses full connection when encoding input to internal state and state
transition, without considering the processing of spatial information. ConvLSTM uses convolution to
make the transition between different states, which can make better use of spatial location information.

In this paper, the feature information of target person is added to ConvLSTM. On the basis of
ConvLSTM unit, the algorithm adds the feature of target person extracted by Resnet-50 part1 to the
input gate, forgetting gate, output gate and input information, which is used to guide the internal state
of ConvLSTM to change toward the region relevant to target person, and discard the information
irrelevant to target person. The feature information of the target person is independent of the
processing flow of the recurrent steps, that is, the feature of the target person is a globally invariant
value in the whole process of detection and re-id. In this paper, the feature information of the target
person is added in the observation process, which makes the whole process of person detection and re-id
more purposeful.

3.4. Training Strategy
The purpose of training is to learn parameters to maximize the reward value of the observation agent
when interacting with the environment, which is a reinforcement learning strategy. The observation
agent acquires the sampling of the interaction sequence through the current strategy and then changes
the parameters of the agent, which increases the probability of selecting the action with high
cumulative reward value and decreases the probability of obtaining low reward action.

In some situations, we know which actions are right. For example, in an object detection task, the
observation agent's action in the final step must output the predictive label of the object. The object
label is known in the training image dataset. After learning the optimization strategy, the observation
agent can output the correct label of the training image at the end of the observation sequence. This
method is a kind of supervised learning. The object detection task can be accomplished by maximizing
the conditional probability of predicting the correct label of the training image.

In the person detection and re-id method described in this paper, since the environment we observe
is a static image frame, we only need to transfer the hₜ state to the re-id network at the last layer of the
observation state for re-id matching. By using the above reinforcement learning strategy, that is,
feedback the corresponding reward value through the correctness of the re-id result output by the final
re-id network, to train the whole detection and re-id network.
3.5. Efficiency Analysis of the Algorithms

Compared with the method of directly connecting person detection and person re-identification in series, this algorithm can effectively reduce the amount of calculation in the whole process.

Person re-identification requires the calculation of feature extraction, feature distance and ranking between the target person image and the images of candidate persons. Therefore, when the candidate set is large, the time of re-id is long. In the traditional method of direct serialize person detection and person re-id, because the information of the target person is not fully used, all the persons detected by the detector in the surveillance video image will participate in the process of person re-id, so the calculation of the whole process is relatively large. In the real world, a person will only appear in a place at a time, so in a surveillance image, the target person will appear at most once. This method uses the features of the target person to guide the process of person detection. In a surveillance image, only one candidate person who is most similar to the target person is inferred for person re-id, thus reducing the number of persons participating in the process of person re-id, and reducing the amount of calculation and time required for the whole process.

For example, in a re-id task in which the number of images in a surveillance scene is n, the average number of persons in each image is assumed to be m. In the process of traditional two-stage person detection and re-id methods, the persons appearing in each surveillance image are detected firstly, and then, as the comparison candidates of person re-id, the feature distance between the candidates and the target persons is calculated and sorted, it needs to do such calculation with m*n candidate persons. The method in this paper only infers one candidate person in each surveillance image, that is, it only needs to calculate and rank the feature distance with n candidate persons after detection, which reduces the time of the whole process.

4. Experiments

In this section, the settings and datasets used in our experiments are firstly introduced, and then the algorithm is tested on CUHK-SYSU[1] and PRW[27] datasets, and compared with traditional methods.

4.1. Experimental Setting and Datasets

Our system is implemented with Pytorch. The ResNet50 is used for target person feature extraction, and is initialized with ImageNet pre-trained model. Conv1 to conv4_3 are used as ResNet50 part1 to extract features at detection stage, conv4_4 to conv5_3 are used as ResNet50 part2 to extract features for person re-id. In each step of the observation process of the RNN, the extracted feature is passed to the LSTM unit to help the detection and re-id network select the next region to be observed. The network is trained using SGD with a batch size of 16. The initial learning rate is 0.001, decayed by a factor of 0.1 at 60K and 80K iterations respectively and kept unchanged until the model converges at 100K iterations.

CUHK-SYSU. CUHK-SYSU is a large-scale person search database consisted of street/urban scene images captured by a hand-held camera or selected from movie snapshots. It contains 18,184 images and 96,143 pedestrian bounding boxes. There are a total of 8,432 labelled identities, and the rest of the pedestrians are served as negative samples for identification.

PRW. The PRW dataset is extracted from video frames captured with six cameras in a university campus. There are 11,816 frames annotated with 34,304 bounding boxes. Among all the pedestrians, 932 identities are tagged and the rest of them are marked as unknown persons similar to CUHK-SYSU.

4.2. Performance Comparison on CUHK-SYSU and PRW

This section compares our method with the traditional person detection methods combined with person re-id methods on CUHK-SYS and PRW datasets. In traditional methods, DPM [3], ACF [38] and Faster R-CNN [11] are selected as detectors to generate candidate regions, and DSIFT [14], LOMO [15] and IDE [27] are used as person re-id methods. In addition, it is also compared with OIM [1]. Table 1 shows the performance comparison between our algorithm and above algorithms on the CUHK-SYSU and PRW datasets.
From Table 1, it can be seen that the accuracy of our algorithm is better than most other methods. Overall, the accuracy performance is consistent between CUHK-SYSU and PRW. The accuracy of each algorithm on PRW is lower than that on CUHK-SYSU due to the poor image quality of PRW compared with CUHK-SYSU. The accuracy of our algorithm on CUHK-SYSU and PRW is slightly lower than that of OIM, but our method reduces the calculation of the whole process, at the expense of a small amount of re-id accuracy, in order to achieve faster re-id speed.

| Table 1. Accuracy comparison of various methods on CUHK-SYSU and PRW. |
|------------------|------------------|---------------|------------------|---------------|
| Method           | CUHK-SYSU        |               | PRW             |               |
|                  | mAP(%) | Rank-1(%) | mAP(%) | Rank-1(%) |
| DPM+DSIFT        | 46.2   | 42.3     | 13.2   | 13.5       |
| DPM+LOMO         | 59.1   | 67.2     | 15.4   | 49.3       |
| DPM+IDE          | 61.2   | 67.9     | 22.6   | 50.1       |
| ACF+DSIFT        | 34.1   | 39.8     | 10.7   | 26.5       |
| ACF+LOMO         | 57.4   | 65.3     | 12.5   | 32.7       |
| ACF+IDE          | 58.3   | 65.4     | 19.5   | 45.2       |
| Faster R-CNN+DSIFT | 49.6  | 55.3     | 21.8   | 49.7       |
| Faster R-CNN+LOMO | 69.5  | 76.2     | 23.1   | 51.1       |
| Faster R-CNN+IDE | 69.3   | 76.5     | 23.6   | 51.3       |
| OIM              | 76.1   | 79.2     | 27.6   | 54.1       |
| Ours             | 75.9   | 78.9     | 27.5   | 53.9       |

4.3. Runtime Comparison

By introducing the feature information of the target person into the process of person detection and re-id, our algorithm reduces the number of candidate person images participating in the process of calculating and ranking the feature distance, thus reducing the amount of calculation and running time of the whole process. Table 2 shows the comparison of running time between OIM and our method, which directly combines person detection and re-id on CUHK-SYSU.

| Table 2. Runtime comparison on CUHK-SYS |
|-----------------------------------------|
| Number of test images | OIM     | Ours     |
|-----------------------|---------|----------|
| 100                   | 193.2s  | 165.4s   |

From Table 2, it can be seen that the processing time of our method is improved by about 14.4% compared with OIM. Each image in CUHK-SYSU average contains about 4 to 5 pedestrians. According to the analysis in section 3.5, person re-id is a 1:N matching process. N represents the number of pedestrians in candidate set. For methods like OIM, the more pedestrians in a surveillance image, the more pedestrians will be generated in the candidate set. However, for your method, the number of pedestrians generated in the candidate set is independent of the number of pedestrians in the image, and only one candidate will be generated in each image. Therefore, the more pedestrians in each image, the speed improvement is bigger of our method compared with other algorithms. Of course, as the number of pedestrians in each image increases, the interference factors increase, and the re-id accuracy decreases slightly.

5. Conclusion

This paper designs an end to end person re-id method based on attention mechanism for surveillance image. The method regards person re-id in surveillance scene as a sequential model problem. Using LSTM network, the original two-stage process of person detection and person re-id is transformed into a unified process of detection and re-id. In this paper, we introduce an initial memory unit containing features extracted using ResNet50 for target person and combine it with person detection and re-id network to detect and re-id the target person. Compared with the two-stage process of person detection
before re-id, the method in this paper is more purposeful in the whole process. In this paper, the attention model is combined with the initial memory unit to guide the observation region of the network, which reduces the number of candidate pedestrians for person re-id, thus reducing the computational load and improve the re-id efficiency.

References
[1] Xiao, T., Li, S., Wang, B., Lin, L., Wang, X.: Joint detection and identification feature learning for person search. In: CVPR. (2017).
[2] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. (2005).
[3] Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. PAMI 32(9) 1627–1645 (2010).
[4] Dollar, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC. (2009).
[5] Zhang, S., Bauckhage, C., Cremers, A.B.: Informed Haar-like features improve pedestrian detection. In: CVPR. (2014).
[6] Zhang, S., Benenson, R., Schiele, B.: Filtered channel features for pedestrian detection. In: CVPR. (2015).
[7] Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: How far are we from solving pedestrian detection? In: CVPR. (2016).
[8] Zhang, S., Benenson, R., Omran, M., Hosang, J., Schiele, B.: Towards reaching human performance in pedestrian detection. IEEE Trans. PAMI 40(4) 973-98 (2018).
[9] Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: ICCV. (2013).
[10] Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR. (2012).
[11] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. PAMI 39(6) 1137-1149 (2017).
[12] Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: CVPR. (2017).
[13] Zhang, S., Yang, J., Schiele, B.: Occluded pedestrian detection through guided attention in CNNs. In: CVPR. (2018).
[14] Zhao, R., Ouyang, W., Wang, X.: Unsupervised salience learning for person re-identification. In: CVPR. (2013).
[15] Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: CVPR. (2015).
[16] Zhao, R., Ouyang, W., Wang, X.: Person re-identification by saliency learning. IEEE Trans. PAMI 39(2) 356–370 (2017).
[17] Paisitkriangkrai, S., Shen, C., Van Den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: CVPR. (2015).
[18] Zhang, L., Xiang, T., Gong, S.: Learning a discriminative null space for person re-identification. In: CVPR. (2016).
[19] Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR. (2015).
[20] Varior, R.R., Shuai, B., Lu, J., Xu, D., Wang, G.: A siamese long short-term memory architecture for human re-identification. In: ECCV. (2016).
[21] Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S.: End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing 26(7) (2017).
[22] Xu, J., Zhao, R., Zhu, F., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In: CVPR. (2018).
[23] Ding, S., Lin, L., Wang, G., Chao, H.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognition 48(10) (2015).
[24] Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In: CVPR. (2016).
[25] Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR. (2016).

[26] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: MARS: A video bench-mark for large-scale person re-identification. In: ECCV. (2016).

[27] Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In: CVPR. (2017).

[28] Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X.: Hydraplus-net: Attentive deep features for pedestrian analysis. In: ICCV. (2017).

[29] Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. arXiv preprint arXiv:1802.08122 (2018).

[30] Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR. (2017).

[31] Xu, Y., Ma, B., Huang, R., Lin, L.: Person search in a scene by jointly modeling people commonness and person uniqueness. In: ACM’MM. (2014).

[32] Xiao, J., Xie, Y., Tillo, T., Huang, K., Wei, Y., Feng, J.: Ian: The individual aggregation network for person search. arXiv preprint arXiv:1705.05552 (2017).

[33] Liu, H., Feng, J., Jie, Z., Jayashree, K., Zhao, B., Qi, M., Jiang, J., Yan, S.: Neural person search machines. In: ICCV. (2017).

[34] Wang F, Jiang M, Qian C, et al.: Residual attention network for image classification. In: CVPR. (2017).

[35] Hochreiter S, Schmidhuber J.: Long short-term memory. Neural Computation 9(8) 1735-1780 (2013).

[36] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016).

[37] Shi X, Chen Z, Wang H, et al.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: CVPR. (2015).

[38] Dollár P, Appel R, Belongie S, et al.: Fast feature pyramids for object detection. IEEE Trans. PAMI 36(8) 1532-1545 (2014).