Deep-Shallow Occlusion Parallelism Network for Person Re-Identification

Qian Wang\textsuperscript{a}, Meibin Qi\textsuperscript{b}, Kaiyuan Jin\textsuperscript{c} and Jianguo Jiang\textsuperscript{d}

DSP Laboratory, School of Computer and Information, Hefei University of Technology, Hefei, Anhui, 230009, China

\textsuperscript{a}tinaw25@foxmail.com; \textsuperscript{b}qimeibin@163.com; \textsuperscript{c}kimcarrysa@mail.hfut.edu.cn; \textsuperscript{d}jgjiang@hfut.edu.cn

Abstract. In recent years, the progress of person re-identification has advanced significantly, but person in real scenes are often obscured by various objects. This problem has been ignored or solved based on an incomplete assumption by the previous person re-ID methods. This paper proposes a new training mechanism, Deep-Shallow Occlusion Parallelism network, which responds to multi-scale occlusion in a more universal case and alleviate the negative impact of occlusion effectively. Specifically, the proposed consists of shallow occlusion and deep occlusion, to which are applied with shielding simulation to enhance the difficulty of training samples. What’s more, channel-spatial attention is applied to various positions in each branch to concentrate on discriminative features after occlusion. In the end, the weighted fusion of the two branches not only informs the network of the information of changes before and after occlusion, but also complements the deep and shallow information effectively. It makes the network more robust. This method has excellent performance on the three basic re-ID datasets and the largest partial re-ID dataset, with the state-of-the-art reached to or even surpassed.

1. Introduction

The task for Person re-identification (re-ID) is to identify the same person between query images and gallery images in different monitoring ranges. These methods require the features to be intensively robust in changes in pose, illumination and observation angle. Some works [1] improve feature learning by locating body parts, aligning poses, segmenting evenly. They assume that each picture covers the full view of a person and are vulnerable to interference such as background and occlusion.

However, person images captured by cameras are usually incomplete. The main reasons include but not limited to: (1) only partial body are covered by the captured images due to changes in the camera's viewing angle; (2) as shown in Figure 1, person images are frequently kept out by other stationary or dynamic objects. In these cases of occlusion, dispersion of target information is enormously challenging for cross-domain person re-ID probably.

A few works have started with the progress of research on addressing occlusion. Some of them [2, 3,4] are based on the ideal assumption that occlusion only exists in the person image in probe, but not in gallery. Some use the exposed part as a new query class after image cropping to learn the remaining local features. Nevertheless, there is still room for further research. First of all, the above assumptions are not comprehensive in reality, as the images in the gallery may only show part of the person. Secondly, manual cropping is subject to subjective deviation of the operator, leading to unsatisfactory experimental results. Thirdly, some useful information has been lost when all occlusions are removed.
roughly, such as dynamic occlusions like luggages and umbrellas with auxiliary information for identification of the target person.

In this term, this paper proposes a more general method called deep-shallow occlusion parallelism network (DSOP). For the dynamic and static multi-scale occlusion issue, it carries out parallel computing of two branches and conducts weight fusion in the end to effectively integrate the information of low-level and high-level features. With the real case of occlusion simulated, DSOP with coexistence of multi-scale occlusion utilizes an attention mechanism to learn discriminative features before and after occlusion, respectively. Moreover, in the end of DSOP, those features integrated for richer complementary information. It has a necessary and important impact on the theoretical research and practical application of re-ID. The model was been tested on three re-ID benchmark datasets and one partial re-ID dataset, with the SOTA reached to or even surpassed.

Figure 1. Samples of person be occluded by obstacles (e.g. cars, wall, bags, umbrellas and other persons).

For prevent it from overfitting and increase of the difficulty of training, the occlusion masks endow model with considerable robustness. Furthermore, random occlusion in the one batch forces the network to learn other more discriminative details in the remaining areas. The utilization of channel-spatial attention highlights important channels and areas. It considers the visible body area more with attention less paid to the occlusion area, so as to improve the network performance. It is worth noting that person images are always rectangular due to the body shape. So the previous works [1,5] concentrate on horizontal segmenting. For real scenes, as the occluded person in images are probably in a square area, the effect of horizontal segmentation will be enormously derogated. Since the proposed method for random appearance of occlusion mask is applicable for any shape, which can be applied extensively.

The summary of the main content of this method and contributions is as follows:

1. Occlusion mask mapping is set in both deep and shallow branches, which makes DSOP more suitable for persons in actual scenarios and more robust.

2. The attention mechanism is applied to the deep and shallow occlusion branches. With more attention paid to person and other useful features in the occluded images, the negative impact of occlusion is reduced greatly.

3. Experiments demonstrate that DSOP model was been tested on three re-ID benchmark datasets and the largest partial dataset, with the current level reached to or even surpassed. Our method increases the performance to 57.7% (+6.3%) Rank-1 accuracy on the partial datasets Occluded-DukeMTMC [4].

2. The Proposed Method

2.1. Problem and Structure
Person re-ID aims at matching images of the same identity across camera views. However, the feature representation is often corrupted by the visual appearances of occlusion. This paper focuses on the severe challenge brought by occlusion to re-ID.
This paper proposes a more comprehensive network DSOP, which takes into account the occlusion in both probe and gallery. DSOP model with channel-spatial attention mechanism utilizes the coexistence of deep and shallow occlusion branches. It learns discriminative features after occlusion in shallow branch, and learns regional features before occlusion in deep branch respectively. In the end, the weighted fusion of the two branches not only informs the network of the information of changes before and after occlusion, but also complements the deep and shallow information effectively to make the network more robust.

2.2. Backbone and Occlusion

We use the ResNet-50 [6] as the backbone. To deal with dynamic and static multi-scale occlusion, this paper proposes two-way occlusion branches after backbone, as shown in Figure 2. Random occlusion mask (OM) is applied in the two occlusion, which interferes with the original training images in local areas.

In the shallow occlusion branch, after backbone, the OM randomly zeroes some isolated features of the input tensor in this batch. As the receptive field is small, the shallow network learns the details such as the edges and contours that are easily affected by background and occlusion. Therefore, the output is fed into the attention module four more times to make the size of features keep up with that in the deep occlusion branch. After occlusion, the local information are paid more attention to alleviate the negative impact caused by occlusion.

In the deep occlusion branch, the input features are fed into stage 4 of ResNet-50 [6] to ensure the feature size is large enough. But the down-sampling operation is not employed [7]. In the deep layer receptive field is larger, the features learns the semantic information. Through the attention module, the extracted features obtain the discrimination information of person. Then, the OM, which is the same size as the OM of the shallow occlusion branch, is mapped to that deep semantic features. The area where the occlusion mask appears is still random. The deep occlusion branch is less affected by details. In the end, we choose to weighted fusion to the deep and shallow occlusion branch. It not only enriches feature expressions and complements useful information, but also prevents overfitting.

2.3. Channel Spatial Attention Modules

Inspired by researches [7,8,9], this paper combines the advantages of spatial attention and channel attention [10]. It is denoted with C-S attention module. The shallow occlusion network passes through C-S attention four times after occlusion mask mapping to get the larger features. Similarly, the deep occlusion network pass through C-S attention module before occlusion mask mapping.

In Figure 2, the input feature map $X \in \mathbb{C} \times \mathbb{H} \times \mathbb{W}$ is fed into the channel attention module to obtain an attention vector with a length of $C$. Multiply these $C$ values by the channels of the input feature map.
X to get the output feature map $X'$. Along the channel the enhanced feature map $X'$ is used to calculate the spatial attention. Then, the output is refined attention tensor $X'' \in (H, W)$. The formula is as follows:

$$X' = X \otimes P_x(X)$$

(1)

$$X'' = X \otimes P_m(X)$$

(2)

where $\otimes$ denotes element-wise multiplication. $P_x$ and $P_m$ are inferred as channel attention map and a spatial attention map.

Channel attention assigns different weights to each channel in the unit of feature map. It uses global average-pooled and global max-pooled features simultaneously. The results of two poolings are input into the same fully convolutional network (FCN) to obtain the C-channel attention vector. The channel attention vector and feature map $X$ are multiplied channel by channel to output the final output $P_x$. In short, the channel attention is computed as:

$$P_m(X) = \sigma(\text{FCN}(\text{GAP}(X)) + \text{FCN}(\text{GMP}(X)))$$

$$= \sigma(W_1(W_0(X_{\text{avg}}^c)) + W_1(W_0(X_{\text{max}}^c)))$$

(3)

where $\sigma$ denotes the sigmoid function, and $X_{\text{avg}}^c$ and $X_{\text{max}}^c$ denote average-pooled features and max-pooled features respectively. Note that the FCN weights, $W_0$ and $W_1$, are shared for both inputs and the ReLU activation function is followed by $W_0$.

Spatial attention is based on each pixel of the feature map to assign the weight. The input $X'$ is applied by max-pooling and average-pooling. The convolution operation with the filter size of $7 \times 7$ is used. The calculation is as follows:

$$P_m(X) = \sigma(\text{CNN}[\text{AvgPool}(X); \text{MaxPool}(X)])$$

$$= \sigma(\text{CNN}[X_{\text{avg}}'; X_{\text{max}}'])$$

(4)

where we aggregate channel information of a feature map by using two pooling operations, generating two 2D maps $X_{\text{avg}}'$ and $X_{\text{max}}'$. Each denotes average-pooled features and max-pooled features across the channel. The convolutional layer outputs spatial attention with more concentrated information.

2.4. Loss Function

Both shallow and the deep occlusion branch use soft margin batch-hard triplet loss [11], and the softmax loss. Firstly, we calculate the loss of the two branches, and then the two types of loss. The sum is calculated at the end.

3. Experiment

3.1. Datasets and Implementation Details

Experiments demonstrate that DSOP model on three re-ID benchmark datasets and the largest partial dataset, including Market-1501 [12], CUHK03-Detect [13], DukeMTMC-reID [14] and Ooccluded-DukeMTMC [4]. The new protocol of CUHK03-Detect selects challenging query images for evaluation, and it is more challenging dataset among the basic three. Notice that Occluded DukeMTMC is more practical since both probe and gallery images have occlusions and does not need a manually cropping.

The Adam optimizer [1] is used with the base learning rate initialized to 1e-4. The input images are re-sized to $384 \times 128$. DSOP is trained using 2 RTX 2080Ti GPUs with a batch size of 108. Each identity contains 4 instance images in a batch. The whole training procedure has 800 epochs and takes approximately 4 hours. We adopt mean average precision (mAP) and CMC Rank-1 accuracies as evaluation metrics. All the experiments are conducted in a single-query setting without re-ranking.

3.2. Comparison with State-of-the-Art

The statistical comparison between our DSOP in CUHK03-Detect, DukeMTMC-reID and Market-1501 and the latest methods is shown in Table 1. Remarkably, our method achieves the largest improvement over previous methods on CUHK03-Detect dataset. DSOP Network achieves 76.9% Rank-1 accuracy on CUHK03-Detect dataset, which is 10.1% higher than the second. It is proved that DSOP is robust
and more applicable. For the Market-1501, our model achieves similar performance to MGN [5]. However, it is worth to point out that MGN benefits from a much larger and more complex network which generates 8 feature vectors with 8 branches supervised by 11 loss functions. In comparison, DSOP involves only a simple network with two branches.

Table 1. DSOP achieves advanced performance on all datasets and far exceeds most methods.

| Method                      | CUHK03-Detect | Market-1501 | DukeMTMC-reID |
|-----------------------------|--------------|-------------|---------------|
|                             | Rank-1 | mAP  | Rank-1 | mAP  | Rank-1 | mAP  |
| MLFN[15]                    | 52.8   | 47.8 | 90.0   | 74.3 | 81.0   | 62.8 |
| HA-CNN[16]                  | 41.7   | 38.6 | 91.2   | 75.5 | 80.5   | 63.8 |
| PCB+RPP[1]                  | 62.8   | 56.7 | 93.8   | 81.6 | 83.3   | 69.2 |
| HPM[17]                     | 63.9   | 57.5 | 94.2   | 82.7 | 86.6   | 74.3 |
| CAMA[18]                    | 66.6   | 64.2 | 94.7   | 84.5 | 85.8   | 72.9 |
| MGN[5]                      | 66.8   | 66.0 | 95.7   | 86.9 | 88.7   | 78.4 |
| IANet[19]                   | -      | -    | 94.4   | 83.1 | 87.1   | 73.4 |
| DSR[3]                      | 61.7   | 56.8 | 91.2   | 75.6 | 82.4   | 68.7 |
| SFR[2]                      | 63.8   | 58.9 | 93     | 81   | 84.8   | 71.2 |
| VPM[20]                     | -      | -    | 93.0   | 80.8 | 83.6   | 72.6 |
| Ours                        | 76.9   | 73.2 | 95.4   | 85.9 | 88.2   | 77.0 |

3.3. Choice of Occlusion Branches
The motivation of two occlusion parallelism is to learn richer appearance and discriminative features. We experiment it on three datasets to analyse the effectiveness of occlusion on each branch. First situation, the structure of network with only deep occlusion branch (the first line of the Table 2). The second, the network with only shallow occlusion branch (the second line of the Table 2). In addition, we also design a simple global branch based on ResNet-50 [6], and combined it with the above two branches and compared the results (the third and fourth line of the Table 2).

In Table 2, only one occlusion branch, the networks learns the occluded image, and the results are not ideal. The third and fourth line of the Table 2 show that even with the holistic person features as supervision, the structure of the occlusion + global only basically reach the current level of re-ID. DSOP complements the deep and shallow information effectively to make the network more robust.

Table 2. The performance of DSOP on three public datasets under different structures.

|                     | CUHK03-Detect | Market-1501 | DukeMTMC-reID |
|---------------------|--------------|-------------|---------------|
|                     | Rank-1 | mAP  | Rank-1 | mAP  | Rank-1 | mAP  |
| Deep Occlusion      | 71.7   | 65.3 | 93.5   | 81.1 | 85.8   | 69.3 |
| Shallow Occlusion   | 71.4   | 63.8 | 93.9   | 81.7 | 83.0   | 67.6 |
| Deep Occlusion(+global) | 75.9  | 71.2 | 94.5   | 84.3 | 87.8   | 73.0 |
| Shallow Occlusion(+global) | 75.1  | 71.2 | 94.1   | 82.8 | 84.4   | 70.6 |
| DSOP (Deep+Shallow Occlusion) | 76.9  | 73.2 | 95.3   | 85.9 | 88.2   | 77.0 |

3.4. Size of Occlusion Masks
The size of OM directly affects the appearance features of person extracted by the occlusion branches. In this section, we change the size of OM on Market-1501. The width of OM and the proportion of input person feature map are fixed to 1:1. The ratio(r) of height OM to input feature map is changed to 10%, 30%, 50%, 70% and 90%. From Figure 3, when the height ratio(r) is 30%, the best performance is achieved. The size of OM selected in DSOP is the most suitable setting for re-ID.
3.5. Experiment on Occlusion Dataset

In Table 3, the proposed method and the results of previous work on the Occluded-DukeMTMC [4] dataset are shown. The first group of methods is mainly designed for the holistic person re-ID. The methods in the second group are designed for the problem of partial re-ID. The results show that DSOP achieves 57.7% Rank-1 accuracy and 45.3% mAP, which surpasses all the previous methods. Compared with PGFA [4], the accuracy of Rank-1 and mAP of DSOP increased by 6.3% and 8.0% respectively. DSOP effectively alleviates the negative impact of occlusion on person re-ID.

3.6. Benefit of C-S Attention Module

In the deep and shallow occlusion branch, DSOP uses a joint channel-spatial attention mechanism [10]. In Table 4, the results show that the utilization of C-S Attention effectively improves the performance of network on Market-1501 and the largest occlusion dataset Occluded-DukeMTMC. Using attention in DSOP not only learns the changes of person features before and after occlusion, but also more focuses on the remaining person features after occlusion.

Table 3. As shown in Table 3, the DSOPNet proposed in this paper achieves the state-of-the-art effect on the occlusion dataset Occluded-DukeMTMC.

| Method              | Rank-1 | mAP  |
|---------------------|--------|------|
| Random Erasing [21] | 40.5   | 30.0 |
| HA-CNN [16]         | 34.4   | 26.0 |
| Adver Occluded [22] | 44.5   | 32.2 |
| PCB[1]              | 42.6   | 33.7 |
| DSR[3]              | 40.8   | 30.4 |
| SFR[2]              | 42.3   | 32.0 |
| PGFA[4]             | 51.4   | 37.3 |
| DSOP                | 57.7   | 45.3 |

Table 4. The comparison experiment on Market-1501 and the occlusion dataset Occluded-DukeMTMC shows that the performance of DSOP with C-S Attention is better.

| DSOP | Occluded-DukeMTMC | Market-1501 |
|------|-------------------|-------------|
|      | Rank-1 | mAP  | Rank-1 | mAP  |
| non-Attention | 50.5   | 36.5 | 93.8   | 84.3 |
| C-S Attention  | 57.7   | 45.3 | 95.4   | 85.9 |

4. Conclusion

This paper proposes a new training method, Deep-Shallow Occlusion Parallelism network, which effectively responds to multi-scale occlusion in a universal case and alleviate the negative impact of occlusion. This method jointly uses C-S attention to improve the network. In DSOP net, two different occlusion branches are utilized. The deep occlusion learns the local information, and pays more attention to under the larger receptive field. The shallow occlusion learns more detailed feature. A large number of experiments on the re-ID dataset show that the DSOP significantly improves the accuracy of person re-ID.
Acknowledgments
This work is supported by the National Natural Science Foundation of China Grant (61771180, 61876056). Qian Wang is the corresponding author of this paper.

References
[1] Sun Y, Zheng L, Yang Y, et al. Beyond Part Models: Person Retrieval with Refined Part Pooling[J]. 2017.
[2] He L, Sun Z, Zhu Y, et al. Recognizing Partial Biometric Patterns[J]. 2018.
[3] He L, Liang J, Li H, et al. Deep Spatial Feature Reconstruction for Partial Person Re-identification: Alignment-Free Approach[J]. 2018
[4] Miao, Jiaxu, et al. "Pose-Guided Feature Alignment for Occluded Person Re-Identification." Proceedings of the IEEE International Conference on Computer Vision. 2019.
[5] Wang G, Yuan Y, Chen X, et al. Learning Discriminative Features with Multiple Granularities for Person Re-Identification[C]// 2018 ACM Multimedia Conference. ACM, 2018.
[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. InCVPR, 2016.
[7] Dai Z, Chen M, Gu X, et al. Batch Drop Block Network for Person Re-identification and Beyond[J]. 2018.
[8] Liu H, Feng J, Qi M, et al. End-to-End Comparative Attention Networks for Person Re-identification[J]. IEEE Transactions on Image Processing, 2017:1-1
[9] Wu J, Jiang J, Qi M, et al. Independent metric learning with aligned multi-part features for video-based person re-identification[J]. Multimedia Tools and Applications, 2019.
[10] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[J]. 2018
[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. Indefense of the triplet loss for person re-identification. arXiv:1703.07737, 2017
[12] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. InICCV,2015.
[13] Zhong Z, Zheng L, Cao D, et al. Re-ranking Person Re-identification with k-reciprocal Encoding[J]. 2017.
[14] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline invitro. InICCV, 2017
[15] Chang, Xiaobin, Timothy M. Hospedales, and Tao Xiang. "Multi-level factorisation net for person re-identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[16] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. InCVPR, 2018
[17] Fu Y, Wei Y, Zhou Y, et al. Horizontal Pyramid Matching for Person Re-identification[J]. 2018.
[18] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. Towards rich feature discovery with class activation maps augmentation for person re-identification. InCVPR, 2019.
[19] Hou R, Ma B, Chang H, et al. Interaction-and-Aggregation Network for Person Re-identification[J]. 2019.
[20] Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., & Wang, S., et al. (2019). Perceive where to focus: learning visibility-aware part-level features for partial person re-identification.
[21] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. arXiv preprint arXiv:1708.04896, 2017
[22] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In CVPR, 2018.