SliceNet: Mask Guided Efficient Feature Augmentation for Attention-Aware Person Re-Identification

Zhipu Liu[0000−0002−4014−1233] and Lei Zhang*[0000−0002−5305−8543]

School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China
{zpliu,leizhang}@cqu.edu.cn

Abstract. Person re-identification (re-ID) is a challenging task since the same person captured by different cameras can appear very differently, due to the uncontrolled factors such as occlusion, illumination, viewpoint and pose variation etc. Attention-based person re-ID methods have been extensively studied to focus on discriminative regions of the last convolutional layer, which, however, ignore the low-level fine-grained information. In this paper, we propose a novel SliceNet with efficient feature augmentation modules for open-world person re-identification. Specifically, with the philosophy of divide and conquer, we divide the baseline network into three sub-networks from low, middle and high levels, which are called slice networks, followed by a Self-Alignment Attention Module respectively to learn multi-level discriminative parts. In contrast with existing works that uniformly partition the images into multiple patches, our attention module aims to learn self-alignment masks for discovering and exploiting the align-attention regions. Further, SliceNet is combined with the attention free baseline network to characterize global features. Extensive experiments on the benchmark datasets including Market-1501, CUHK03, and DukeMTMC-reID show that our proposed SliceNet achieves favorable performance compared with the state-of-the art methods.

Keywords: Person re-identification · SliceNet · Self-Alignment Attention.

1 Introduction

The goal of person re-identification(re-ID) is to identify the query images from a large gallery across multiple cameras with non-overlapping views. It has attracted lots of research interest because of its valuable applications in video surveillance, such as multi-camera tracking, multi-camera object detection [28], and pedestrian retrieval [18,14]. In recent years, a large number of deep learning methods are proposed for person re-ID, and achieve big breakthrough. However, it is still an unsolved task due to the drastic appearance changes caused by various viewpoint, resolution, poses, background noise as well as occlusions.

Early re-ID methods [20] learn the global representation from whole-body images, but lose discriminative information lying around body parts. To capture the local discriminative information, a uniform-partition method was proposed to learn local representations from some predefined horizontal partition strips. However, person images

collected by automatic detectors often suffer from misalignment and the re-ID accuracy can be compromised. So horizontal stripes may be less effective when severe misalignment happens. To address this issue, pose estimation based methods [29,5] were proposed to detect key points of body such as head, foots etc. However, severe occlusions can affect the accuracy of these methods. Also, most of these methods need extra training for pose estimation, and the complexity and difficulty of feature learning are casted.

Visual attention has the ability to guide the learning toward informative and discriminative regions, and can be exploited to capture fine-grained saliency regions. That is, it can help to discover the most discriminative regions by producing attention maps containing more personal information. Consequently, a number of attention based person re-ID methods [30,22] were proposed, but most existing works only pay attention to the high-level semantic features but ignore the low-level fine-grained information such as clothing color. The conventional strategy to use the low-level information is intuitively concatenate the convolutional feature maps into the fully-connected layer. However, it ignores the training inequality of each level and may lead to unreliable features. Therefore, to achieve more reliable person re-ID, we propose to exploit the rich fine-grained features from a network, by taking into account both low-level features (e.g., shape, color etc.) and high-level semantic details (e.g., identity).

In order to augment the high-level semantic by exploiting the low-level features and make them compatible, with the philosophy of divide and conquer, we focus on the parallel learning from low-level to high-level. Additionally, we propose a self-alignment attention module in each level to enable the person feature learning. Specifically, in this paper, we propose a novel SliceNet which aims to address the open-world person re-ID from three perspectives: 1) we propose to learn each level in parallel by cutting the network into multiple sub-networks instead of the conventional layer-wise feature concatenation, such that the training is balanced level-to-level and enable reliable person re-ID; 2) we propose the mind of self-alignment attention and aim to find out the discriminative align-attention local saliency regions by learning channel and spatial attention masks from their own feature maps; 3) we leverage the mind that global whole-body features has high semantic discrimination, the high-level information is used for feature augmentation. The main contributions of this paper are three-fold:

(1) With the philosophy of divide and conquer, we propose a SliceNet comprising of three sub-networks with different depth and one global network, which aims to learn the informative features from low-level to high-level in parallel. The three sub-networks contain two, three and four stages of ResNet-50 [11], respectively. The global network is the complete ResNet-50 for whole-body feature learning.

(2) We propose a novel Self-Alignment Attention Module in each subnetwork to capture discriminative local align-attention saliency regions for addressing open-world person re-ID challenges.

(3) Exhaustive experiments on three large datasets including Market-1501, CUHK03, and DukeMTMC show that the proposed SliceNet outperform a number of state-of-the arts in person re-identification.

2 Related Work

With the development of deep learning and the availability of large datasets, deep learning based approaches have dominated the re-ID research community due to their significant superiority in discriminative feature representation. Existing deep learning based re-ID methods can be divided into two categories: 1) learning global whole-body features [12] and 2) learning local information (e.g., part, pose estimation, attention etc.) [29,5,30,22].

Specifically, for the former, Lin et al [17] proposed a simple but effective convolutional neural network, which integrates an identity classification loss and a number of attribute classification losses. Hermans et al [12] proposed a variant of triplet loss to focus on the hard positive and hard negative examples. Chen et al [2] designed a quadruplet loss to supervise the training of their model. Zhong et al [37] proposed a camera style transfer model to address the issue of image style variations caused by different cameras. Although these methods achieved improved performance, the local discriminative regions caused by background noise, pose and occlusion were ignored.

For the latter, recently, a number of parts based deep learning methods have been proposed to capture richer and finer local visual cues. These newly proposed approaches can be broadly classified into three subcategories according to the parts learning scheme, including uniformly partition methods [4,16,31], pose estimation based methods [29,5] and attention based methods [30]. The uniformly partition methods can learn local information, but can not address the issue when the same person in two images are misaligned. Although the pose estimation based methods can address the issue of misalignment, they may be compromised when body parts are occluded or missing during pedestrian detection. Attention based methods is an another subcategory that have been widely explored in various tasks, including image classification [9], object recognition [13], image captioning as well as person re-ID [30,22]. Zhao et al [30] proposed a part-align human representation, which detects the discriminative human body regions beneficial to person matching. Sun et al [22] proposed a Part-based Convolutional Baseline(PCB) network and a refined part pooling(RPP) method to learn discriminative part-information features for person retrieval. Si et al [21] proposed a Dual Attention Matching network (DuATM) to learn context-aware feature sequences and perform attentive sequence comparison from a dual attention mechanism including intra-sequence and inter-sequence attention strategies.

Our attention module is inspired by the part-align method [30], which learns discriminative regions by computing the corresponding regions of a pair of probe and gallery images. Nevertheless, our work differs significantly from part-align methods in three aspects. First, the principle of mask learning is different. the mask in [30] is learned only in channel-wise while ignoring the spatial context-aware information. Our attention mind aims to learn both channel attention and spatial attention masks by a self-alignment attention module, such that more spatial context-aware information can be exploited to guide the attention maps toward informative and discriminative regions. In addition, the same person captured by different cameras can appear in various pose, as a result, the two images do not contain part-align regions, therefore we adopt a soft cross-entropy loss for slight relax on the labels. Finally, with the philosophy of divide

and conquer, we propose to parallel learn the local information level by level for reliable re-ID instead of intuitively concatenate the convolutional feature maps.

3 Methodology

In this section, we present the details of the proposed SliceNet, as illustrated in Figure 1, which includes three subnetworks embedded with a self-alignment attention module for local information learning and one global network for identity semantic feature learning.

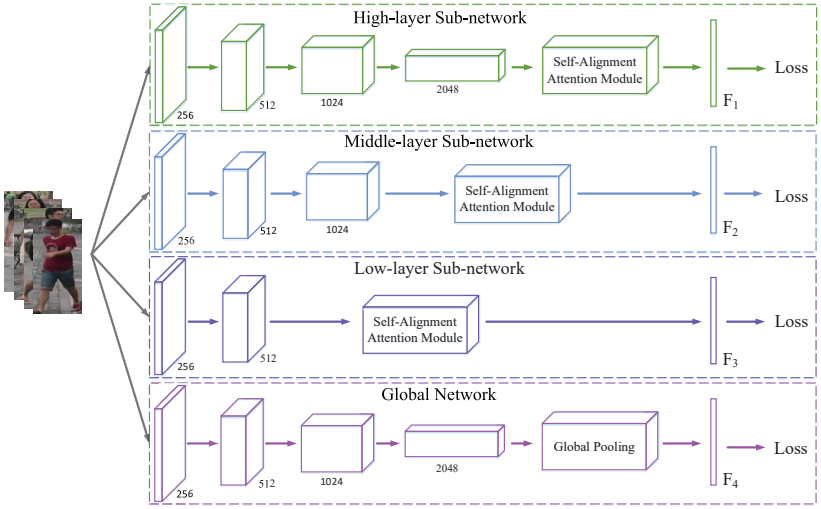

Fig. 1. The architecture of proposed SliceNet. For parallel learning each level equally, with the philosophy of divide and conquer, we divide the network into three subnetworks: high-level subnetwork, middle-level subnetwork and low-level subnetwork, and combined with the attention-free baseline Global Network. For exploiting the local information of subnetworks, a Self-Alignment Attention Module is embedded. Note that all networks are separately learned. During testing, the distance of each feature between probe image and gallery is summed as the final similarity metric.

3.1 Architecture of SliceNet

Previous attention based person re-ID approaches only focus on learning the last convolutional layer local feature, while discarding many local details contained in low-level layers. The low-level information, e.g., color and texture of attention regions, are also important clues for person re-ID. To address this problem, by cutting the baseline network level by level, a SliceNet is proposed as illustrated in Figure 1. The baseline

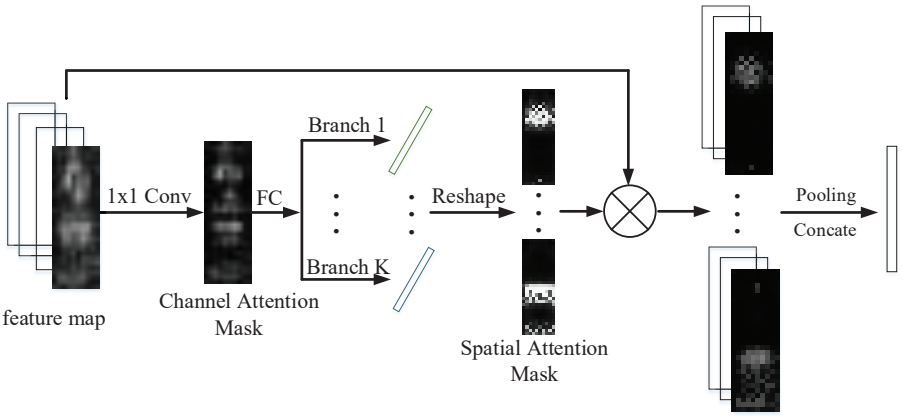

Fig. 2. The framework of the proposed self-alignment attention module. The feature map, extracted from the SliceNet, are followed by a 1×1 convolutional layer to get a channel attention mask. Then, it is reshaped to a 1-dimensional vector which is feeded into the K branches for estimate K spatial attention masks. The masks are applied to the original feature map through element-wise product operation, followed by global pooling and feature concatenation operation.

of our network is ResNet-50 [11], that contains four stages. Each stage (block) comprises of multiple convolutional layers. At the end of each stage, the feature is spatially down-sampled and fed into the next layer.

SliceNet consists of three sub-networks, which contains four, three and two stages of ResNet-50. The last global average pooling layer is replaced by the proposed Self-Alignment Attention Module. The first subnetwork contains all stages of ResNet-50 and is named high-level subnetwork, and the second and third subnetwork are middle-level subnetwork and low-level subnetwork, as shown in Figure 1. The feature map extracted from the three subnetworks are followed by the proposed self-alignment attention module. The last network is an attention-free baseline named Global Network, followed by a global average pooling layer without dimensionality reduction operation.

During training phase, each subnetwork is trained equally, independently and efficiently, which is the philosophy of divide and conquer, such that the mutual negative impacts between different layers can be eliminated and their respective local attention saliency regions can be effectively explored. During testing phases, with the feature augmentation, the distance of each feature-pair between probe image and gallery is summed together with different weights as the final similarity metric.

3.2 Self-Alignment Attention Module

The Self-Alignment Attention Module, as illustrated in Figure 3, first learns a channel attention mask which is fed into the following K branches, and K spatial attention masks can be obtained.

The input of the attention module is the feature map extracted from each subnetwork in SliceNet. First, a 1×1 convolutional layer is utilized to learn a 2-dimensional channel

attention mask. Based on the channel attention mask, to detect multiple discriminative regions, we design multiple branches to learn multiple spatial attention masks. Note that all branches share the same architecture but are deployed with different parameters. Take one branch for example, the channel attention mask is reshaped to a 1-dimensional vector, followed by fully connected layers to learn the weights of local regions, and then reshaped to a 2-dimensional spatial attention mask with same size of original feature map. In this way, a series of self-alignment guided spatial attention masks are learned. These attention masks are applied to the original feature map through element-wise product operation, followed by global average pooling and fully-connected layers, then we concatenate all these local features as the final attention feature.

Let a 3-dimensional tensor $\mathbb{T}$ denote the feature maps extracted from the baseline network and use (x, y, c) to represent the feature map size. (x, y) is the spatial location and c represents the cth channel. A 1×1 convolutional layer is applied to learn a 2-dimensional channel attention mask M_c, and reshaped to a vector V_c, which is the input of the subsequent multiple branches. Each branch contains two fully-connected layers:

$$A_k = F_k(V_c) \tag{1}$$

where F_k represents the fully connected layers of the kth branch, A_k denotes spatial attention vector, followed by a reshape operator for transforming into a 2-dimensional spatial attention mask M_k with the same size of M_c:

$$M_k = R_k(A_k) \tag{2}$$

where R_k represents the reshape operator of the kth branch. Then, the spatial attention mask M_k represents the saliency weights for local regions, which is applied on the feature maps $\mathbb{T}$:

$$T_k(x, y, c) = T(x, y, c) \odot M_k(x, y) \tag{3}$$

where $\odot$ represents element-wise product. Then, by adding a global average pooling layer after T_k, i.e., $f_k = AvePooling(T_k)$, a feature vector f_k is obtained. Further, for reducing the dimension of f_k, two fully-connected layers are used to transform the feature f_k to a 128-dimensional feature f'_k in each branch. Finally, we concatenate all the local features to obtain the final attention feature of local subnetworks in Figure 1:

$$F = [f'_1\ f'_2\ ...\ f'_k] \tag{4}$$

where F is the final feature of all attention regions.

3.3 Loss function

To improve the learning ability of SliceNet, the loss functions we use to train our network is the combination of triplet loss and soft cross-entropy loss.

Triplet loss has been widely used in re-ID, which aims to learn features such that the distance between positive samples decreases and the distance between negative samples increases. Given a batch of images X, consisting of P individuals and K images per person,

and the triplet loss can be represented as follows:

$$\mathcal{L}_{triplet} = \sum_{p=1}^{P} \sum_{k=1}^{K} \left[d_{pos}^{p,k} - d_{neg}^{p,k} + m \right]_+ \tag{5}$$

where $d_{pos}^{p,k} = \max_{a=1,\cdots,K} D(\phi(x_p^k), \phi(x_p^a))$ and $d_{neg}^{p,k} = \min_{q \neq p} D(\phi(x_p^k), \phi(x_q^b))$ represent the distance of the hard positives and hard negative, respectively. $D(.,.)$ represents the L2 distance between two features and $\phi(x_p^k)$ represents the feature of image k with respect to person p. m is a margin that controls the distance between positives and negatives.

A network with only triplet loss considered can easily lead to over-matching when the same person appear various poses captured in different cameras. To alleviate this problem, we add an extra soft cross-entropy loss, which is deployed after the linear layer activated by softmax probability function. The cross-entropy loss is

$$\mathcal{L}_{softmax} = -\sum_{n=1}^{N} \sum_{i=1}^{C} y_{n,i} \log p_{n,i} \tag{6}$$

where y_n denotes the one-hot encoded label vector of image x_n. Compared with traditional softmax loss, we use $y_{n,i} = 0.7$ for x_n and $0.3/(n-1)$ for other bits. C is the number of classes. N is the batch size. The probability $p_{n,i}$ is computed by softmax function, shown as

$$p_{n,i} = \frac{\exp(W_i f_{n,i})}{\sum_{i=1}^{C} \exp(W_i f_{n,i})} \tag{7}$$

where W_i is the weights of linear layer.

The total loss of our SliceNet is the combination of triplet loss and softmax loss:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{triplet} + \lambda_2 \mathcal{L}_{softmax} \tag{8}$$

where λ_1 and λ_2 are the trade-off parameters. In our experiments, both are set to 1.

3.4 Implementation Details

Details of the Backbone Network. Inspired by the success of deep object detection methods [19,6], the granularity of feature can be enriched by removing the last spatial down-sampling operation in the backbone network. This was introduced to person re-ID in [22] and we follow this setting in the backbone of ResNet-50. In addition, the feature map extracted from the low-level subnetwork that only contains two stages, has a very large scale, which is difficult to learn local attention regions, so we add the spatial down-sampling operation at the end of first stage in the subnetwork. The input images are resized to 480×160 for training and testing all subnetworks in our SliceNet. We set the batch size to 128 with $P = 32$ and $K = 4$ to train our model. Our model loads the weights of backbone network ResNet-50 pre-trained on ImageNet [7]. Note that the global whole-body feature extracted from the last Global Network is a 2048-dimensional vector without further dimensionality reduction.

Details of the Self-alignment Attention Module. In this module, the output of backbone network is a 3-dimensional feature map, which is followed by a 1×1 convolutional layer and computes a channel attention mask. Then, the channel attention mask is reshaped to a vector and then feed into the *K* branches to further learn *K* spatial attention masks. In this paper, we set *K* to 10. In each branch, the spatial attention mask is learned by two fully-connected layers. The first fully-connected layer has 800 neural nodes followed by batch normalization, ReLU function and dropout operation with 0.5. The second layer has 300 units and is reshaped into a 2-dimensional spatial mask 30×10 with the same size as the feature map extracted from subnetworks. After element-wise product between spatial attention mask and the feature map, two fully-connected layers are followed for dimension reduction. The first layer has 1024 units followed by batch normalization operation and ReLU function, and the second layer has 128 units.

Network Training. We use Adam optimizer [8] to train our network, and update the learning rate as follows:

$$lr(t) = \begin{cases} lr_0 & , t \leq t_0 \\ l_0 0.001^{\frac{t-t_0}{t_1-t_0}} & , t_0 \leq t \leq t_1 \end{cases} \tag{9}$$

where lr_0 is the initial learning rate, set as $lr_0 = 3e-4$. $t_0 = 300$ and $t_1 = 600$. The margin m of the triplet loss is set to 0.3.

4 Experiments

4.1 Datasets and Evaluation Metrics

To evaluate the performance of our proposed method, we conduct exhaustive experiments on three large datasets: Market-1501 [32], CUHK03 [24] and DukeMTMC [34].

Market1501 were captured from 6 different cameras and contains 12,936 training images of 751 identities and 19,732 testing images of 750 identities. The pedestrians are automatically detected by DPM-detector [10]. During test, it contains single-query and multiple-query models. The single-query model only contains 1 query image of a person and has 3368 query images. The multiple-query model use the avg- or max-pooling features of multiple images.

CUHK03 contains 13,164 images of 1467 persons captured from 6 cameras and each person is captured by 2 cameras. The bounding boxes of pedestrians contain both manually labelled and DPM detected, and we adopt the latter in this paper. The original training/testing protocol is to randomly select 100 identities for testing and the remaining ones for training. 20 random train/test splits [24] are considered, but time-consuming for deep learning.

DukeMTMC-reID is a subset of DukeMTMC captured from 8 high-resolution cameras and detected by manually labelled. It contains 16,522 training images, 17,661 gallery images and 2,228 queries from total 1404 identities.

Evaluation Protocol. In our experiments, we employ the standard cumulative matching characteristics (CMC) accuracy (Rank-1) and mean average precision(mAP) [32] on all datasets to evaluate the performance of different re-ID methods. On Market-1501 dataset, we select the single query model. To simplify the evaluation procedure on CUHK03 dataset, we adopt the new training/testing protocol proposed by [35].

4.2 Comparison with State-of-the Arts

Methods	Rank-1	mAP
part-aligned [30]	81.00	63.40
APR [17]	84.29	64.67
TriNet [12]	84.92	69.14
DaRe (R) [26]	86.40	69.30
PL-Net [27]	88.20	69.30
HA-CNN [25]	91.20	75.70
DuATM [21]	91.42	76.62
SPReID [15]	93.68	83.36
PCB+RPP [22]	93.80	81.60
SliceNet (Ours)	**95.43**	**86.86**
TriNet (RR) [12]	86.67	81.07
DaRe (R, RR) [26]	88.30	82.00
SPReID (RR) [15]	94.63	90.96
SliceNet (Ours, RR)	**96.35**	**94.44**

Table 1. Rank-1 and mAP comparison (%) of SliceNet with other state-of-the arts on Market-1501. 'RR' represents the re-ranking operation proposed by [35].

We compare our proposed method with state-of-the arts on three widely used datasets: Market-1501, CUHK03 and DukeMTMC-reID.

Comparison on Market-1501. Table 1 shows the results of our proposed method and other state-of-the-art methods on Market-1501. As shown in Table 1, our proposed method achieves rank-1 accuracy of 95.43 and mAP of 86.86, which shows competitive performance compared with all of them. By comparing with the very recent PCB+RPP [22], our method outperforms it by 1.63% in rank-1 accuracy and 5.20% in mAP. The attention based methods part-aligned [30], SPReID [17], DuATM [21] and PCB+RPP [22] achieve superior accuracy with 81.00, 91.20, 91.42 and 93.80 of Rank-1. By comparing to attention based methods, the rank-1 accuracy can be improved by 14.43, 4.23, 4.01 and 1.63, respectively, by using our SliceNet.

Comparison on CUHK03. The results on CUHK03 (Detected) is summarized in Table 2, from which we observe that our method achieves Rank-1 of 69.71 and mAP of 66.81. Our model outperforms all the compared methods by a large increment. It exceeds the state-of-the art PCB+RPP [22] by 6.0 in Rank-1 and 9.3 in mAP.

Comparison on DukeMTMC-reID. On DukeMTMC-reID datasets, our proposed method achieves the best Rank-1 accuracy of 88.7 and mAP of 76.1, as is shown in Table 2. PCB+RPP [22] achieves Rank-1 of 83.3 and mAp of 69.2, which surpasses all other compared methods of table 2, but our methods outperforms it by 5.4 in Rank-1 and 6.9 in mAP. Therefore, the effectiveness of the proposed SliceNet is verified and the importance of self-alignment attention mask learning for multi-level fine-grained local saliency region features is obvious.

Methods	CUHK03		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
PAN [33]	36.3	34.0	71.6	51.5
DPFL [3]	40.7	37.0	79.2	60.6
SVDNet [23]	41.5	37.3	76.7	56.8
HA-CNN [25]	41.7	38.6	80.5	63.8
MLFN [1]	52.8	47.8	81.0	62.8
TriNet+Era [36]	55.5	50.7	73.0	56.6
PCB+RPP [22]	63.7	57.5	83.3	69.2
Ours	**69.7**	**66.8**	**88.7**	**76.1**

Table 2. Rank-1 and mAP Comparison (%) with state-of-the arts on DukeMTMC-reID and CUHK03 (Detected) with the same setting as [35].

4.3 Discussions

Local Details Visualization in Different Layers We know that the representations in low-level focus on learning color and texture information, but features extracted from deeper of network tend to be more abstract. As a result, the feature learned from different layers contain different semantic information, by focusing on different local details. The visualization of the learned attention saliency regions on different layers are shown in Figure **??**, in which the attention maps learned from middle-level subnetwork and high-level subnetwork are described. The first column is the original feature map extracted from baseline, and the following ten columns show ten attention maps learned from by using our SliceNet. We can see that the attention map in the two subnetworks are different, and the attention regions in middle-level subnetwork are bigger than that of high-level subnetwork. Additionally, the attention map learned from high-level subnetwork are more focused and the attention maps learned from middle-level subnetwork have obvious noises. The reason is that the semantic information learned from low-level is weaker than high-level subnetwork.

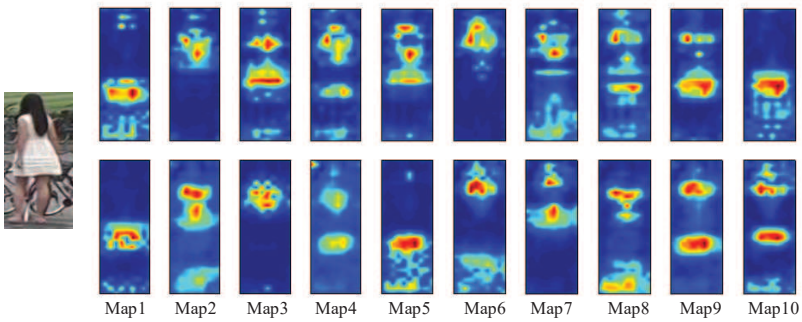

Fig. 3. Examples of visualization results learned from self-alignment attention module. The first row is the result extracted from middle-level subnetwork and the second row is from high-level subnetwork.

5 Conclusion

In this paper, we propose a novel trainable architecture named SliceNet, followed by self-alignment attention module to discover and exploit the multi-level discriminative local attention saliency regions. The framework is with the philosophy of divide and conquer for parallel learning of each level equally, independently, and efficiently. Compared with most existing attention based methods that only focus on learning local features of the last convolutional layer, the SliceNet not only learns high-, middle- and low- level local saliency features but also combines the global whole-body information, which can achieve more reliable person re-ID tasks. Exhaustive experiments on three widely used datasets in person re-ID show that our proposed method has superior performance over the state-of-the arts.

References

1. Chang, X., Hospedales, T.M., Tao, X.: Multi-level factorisation net for person re-identification. In CVPR (2018) 10
2. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In CVPR (2017) 3
3. Chen, Y., Zhu, X., Gong, S.: Person re-identification by deep learning multi-scale representations. In: IEEE International Conference on Computer Vision Workshop (2017) 10
4. 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: Computer Vision & Pattern Recognition (2016) 3
5. Chi, S., Li, J., Zhang, S., Xing, J., Wen, G., Qi, T.: Pose-driven deep convolutional model for person re-identification. In ICCV (2017) 2, 3
6. Dai, J., Yi, L., He, K., Jian, S.: R-fcn: Object detection via region-based fully convolutional networks. In: In CVPR (2016) 7
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision & Pattern Recognition (2009) 7
8. Diederik Kingma, J.B.: Adam: A method for stochastic optimization. In ICLR (2015) 8
9. Fei, W., Jiang, M., Chen, Q., Yang, S., Cheng, L., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In CVPR (2017) 3
10. Felzenszwalb, P.F., Mcallester, D.A., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. in cvpr. In: IEEE Conference on Computer Vision & Pattern Recognition (2008) 8
11. He K, Zhang X, R.S.e.a.: Deep residual learning for image recognition. In CVPR (2016) 2, 5
12. Hermans A, Beyer L, L.B.: In defense of the triplet loss for person re-identification. In CVPR (2017) 3, 9
13. Jimmy Ba, Volodymyr Mnih, K.K.: Multiple object recognition with visual attention. In CVPR (2014) 3
14. Jing, X., Rui, Z., Feng, Z., Wang, H., Ouyang, W.: Attention-aware compositional network for person re-identification. In CVPR (2018) 1
15. Kalayeh, M.M., Basaran, E., Gokmen, M., Kamasak, M.E., Shah, M.: Human semantic parsing for person re-identification. In CVPR (2018) 9
16. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. In CVPR (2017) 3

17. Lin, Y., Liang, Z., Zheng, Z., Yu, W., Yi, Y.: Improving person re-identification by attribute and identity learning. In CVPR (2017) 3, 9
18. 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 A Publication of the IEEE Signal Processing Society **26**(7), 3492–3506 (2017) 1
19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision (2016) 7
20. Prosser, B., Zheng, W.S., Gong, S., Tao, X.: Person re-identification by support vector ranking. In: British Machine Vision Conference (2010) 1
21. Si, J., Zhang, H., Li, C.G., Kuen, J., Kong, X., Kot, A.C., Gang, W.: Dual attention matching network for context-aware feature sequence based person re-identification. In CVPR (2018) 3, 9
22. Sun, Y., Liang, Z., Yi, Y., Qi, T., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: European Conference on Computer Vision (2018) 2, 3, 7, 9, 10
23. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: IEEE International Conference on Computer Vision (2017) 10
24. Wei, L., Rui, Z., Tong, X., Wang, X.G.: Deepreid: Deep filter pairing neural network for person re-identification. In: Computer Vision & Pattern Recognition (2014) 8
25. Wei, L., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In CVPR (2018) 9, 10
26. Yan, W., Wang, L., You, Y., Xu, Z., Weinberger, K.Q.: Resource aware person re-identification across multiple resolutions. In CVPR (2018) 9
27. Yao, H., Zhang, S., Zhang, Y., Li, J., Qi, T.: Deep representation learning with part loss for person re-identification. IEEE Transactions on Image Processing **PP**(99), 1–1 (2017) 9
28. Zhang, S., Wen, L., Xiao, B., Zhen, L., Li, S.Z.: Single-shot refinement neural network for object detection. In CVPR (2017) 1
29. Zhao, H., Tian, M., Sun, S., Jing, S., Yan, J., Shuai, Y., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: IEEE Conference on Computer Vision & Pattern Recognition (2017) 2, 3
30. Zhao L, Li X, W.J.e.a.: Deeply-learned part-aligned representations for person re-identification. In ICCV (2017) 2, 3, 9
31. Zheng, F., Sun, X., Jiang, X., Guo, X., Yu, Z., Huang, F.: A coarse-to-fine pyramidal model for person re-identification via multi-loss dynamic training. In CVPR (2019) 3
32. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: IEEE International Conference on Computer Vision (2015) 8
33. Zheng, Z., Liang, Z., Yi, Y.: Pedestrian alignment network for large-scale person re-identification. In CVPR (2017) 10
34. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: IEEE International Conference on Computer Vision (2017) 8
35. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. In: IEEE Conference on Computer Vision & Pattern Recognition (2017) 8, 9, 10
36. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In CVPR (2017) 10
37. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In CVPR (2017) 3