Self-Critical Attention Learning for Person Re-Identification

Guangyi Chen[1,2,3], Chunze Lin[1,2,3], Liangliang Ren[1,2,3], Jiwen Lu[1,2,3,*], Jie Zhou[1,2,3]
[1]Department of Automation, Tsinghua University, China
[2]State Key Lab of Intelligent Technologies and Systems, China
[3]Beijing National Research Center for Information Science and Technology, China
{chen-gy16,lcz16,renll16}@mails.tsinghua.edu.cn; {lujiwen,jzhou}@tsinghua.edu.cn

Abstract

In this paper, we propose a self-critical attention learning method for person re-identification. Unlike most existing methods which train the attention mechanism in a weakly-supervised manner and ignore the attention confidence level, we learn the attention with a critic which measures the attention quality and provides a powerful supervisory signal to guide the learning process. Moreover, the critic model facilitates the interpretation of the effectiveness of the attention mechanism during the learning process, by estimating the quality of the attention maps. Specifically, we jointly train our attention agent and critic in a reinforcement learning manner, where the agent produces the visual attention while the critic analyzes the gain from the attention and guides the agent to maximize this gain. We design spatial- and channel-wise attention models with our critic module and evaluate them on three popular benchmarks including Market-1501, DukeMTMC-ReID, and CUHK03. The experimental results demonstrate the superiority of our method, which outperforms the state-of-the-art methods by a large margin of $5.9\%/2.1\%$, $6.3\%/3.0\%$, *and* $10.5\%/9.5\%$ *on mAP/Rank-1, respectively.*

1. Introduction

Person re-identification (ReID) aims to identify an individual across multiple non-overlapping camera views deployed at different locations, considering a large set of candidates. It plays an important role in various video surveillance applications such as suspect tracking and missing elderly or children retrieval, and has attracted much attention over the past few years [20, 46, 53, 38, 39, 19].

Despite the recent progress, ReID is still a challenging problem due to the difficulty of visual features matching with the illumination changes, pose variations, occlusions, and cluttered backgrounds. Recently, several attention-based deep learning models for ReID have been proposed to address these issues [48, 19, 41]. With the attention mechanism, the model learns to focus on discriminative features of the pedestrians and reduce the negative effects of different variations and background interference. However, the attention mechanism is generally trained in a weakly-supervised manner without a powerful supervisory signal to guide the attention module during the training process. As shown in Figure 1, the gradients from this weak supervisory signal might be vanishing in the back propagation process [15]. The attention maps learned in such manner are not always "transparent" in their meaning, and lack discrimination ability and robustness. The redundant and misleading attention maps are hardly corrected without direct and appropriate supervisory signal. Moreover, the quality of the attention during training process can only be evaluated qualitatively by the human end-users, examining the attention map one by one, which is labor-intensive and inefficient.

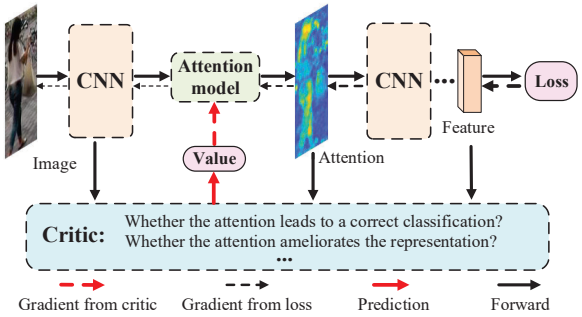

Figure 1. The insight of the self-critical attention learning. Convolutional attention learning is guided by the weak supervisory signal from the loss function. However, this learning manner is not always "transparent" and suffers from the problem of "vanishing" gradients. Differently, our SCAL method exploits a critic module to evaluate the quality of the attention model and provide the strong supervisory information by the predicted critic value.

To overcome the above issues, in this paper, we propose a self-critical attention learning (SCAL) method for person ReID. We simultaneously train an attention agent and a critic module to provide the self-critic and self-correctness ca-

*Corresponding author

pability for the attention model. Specifically, the attention agent produces the visual attention maps to focus the model on discriminative features. The critic module examines the attention and measures the gain on the performance. Based on its observation, the critic provides a direct supervisory signal to the attention agent in order to maximize the gain. We illustrate the self-critical attention learning flowchart in Figure 1. When the attention is incorrectly allocated, the critic provides the feedback to the attention agent, so that it can figure out the mistakes and adapts itself, which alleviates the "vanishing" gradients and "transparent" learning of the weakly-supervised manner. Beyond powerful supervision, the outputs of the critic permit to quantify the quality of attention, which significantly facilitates the interpretation of attention learning process. To train our critic module, we exploit several intuitive evaluation criteria such as the effect of the attention on the final classification results and the relative gain compared to original features without attention. As these criteria are usually non-differentiable, the conventional back-propagation is hardly directly used for learning. This motivates us to formulate our self-critical attention learning process in a reinforcement learning framework, where the state is the input person image, the action is the generated attention. In this framework, the critic receives the state and action to evaluate the quality of attention and is optimized by minimizing the difference between the predicted critic value and actual evaluation criteria. Using our self-critical learning process to train the spatial- and channel-wise attention models substantially outperforms the other state-of-the-art methods on three popular benchmarks including Market-1501, DukeMTMC-ReID, and CUHK03.

2. Related Work

Person Re-identification: Person ReID systems roughly consist of two major components: *representation learning* and *metric learning*. Some conventional methods primarily employ handcrafted features such as color and texture histograms. Liao *et al.* [20] propose a Local Maximal Occurrence (LOMO) method to handle viewpoint changes by maximizing the horizontal occurrence of local features. Matsukawa *et al.* [27] propose a hierarchical Gaussian feature, which models the color and texture cues of each region by multiple Gaussian distributions. Metric learning also has been widely applied for person ReID. LMNN [45] attempts to ensure that for each person its neighbors always belong to the same class while examples from different classes are separated by a large margin. To learn the nonlinear relation of persons, the kernel-based metric learning methods are proposed [47, 21]. Recently, deep learning based person ReID approaches have achieved great success [18, 1, 33, 23, 39] through simultaneously learning the person representation and similarity within one network. Some methods [51, 46] usually learn the representation feature via training a deep classification network. In addition, some works employ deep metric learning method for person ReID such as: pair-wise contrastive loss [7], triplet ranking loss [56] and quadruplet loss [5]. To avoid the effect of the background clutters and pose variations, several body-structural or part-based methods [39, 49, 16, 12, 6, 36] are proposed. These methods leverage the prior human-body information or learning-based pose information to locate salient parts and learn structural representation.

Attention Model: Recently, attention models [28, 42, 22] have gained great success in various fields, such as natural language processing (NLP), image understanding, and video analysis. It is also efficient and effective for person ReID to handle the matching misalignment challenge and enhance the feature representation [24, 19, 35, 17, 14, 50, 48, 34, 19, 11, 42, 13]. For example, Liu *et al.* [24] and Lan *et al.* [14] directly learn attention regions to locate the salient image regions. Xu *et al.* [48] and Zhao *et al.* [50] introduce a body part detector to consider the body structure in the attention model. Some works [26, 34, 17, 4] employ the attention model on the frame or feature sequences to select key parts of sequences. In addition, channel-based attention methods [19, 11, 48] are proposed to refine feature representations. However, the training process of these attention methods is only sustained by a weak supervision signal and the effect of the attention model is invisible for the overall model. We propose therefore the self-critical attention learning method to address these issues. In particular, we develop a critic module to evaluate the quality of the attention model, which provides powerful supervision signal for attention learning and quantificationally measures the effectiveness of the attention model.

3. Approach

In this section, we first present our self-critical attention learning method and then employ it on both spatial- and channel-wise attention models. Finally, we explain the optimization procedure and implementation details.

3.1. Self-critical Attention Learning

The attention module is an important component for person ReID systems to guide the network to find the most discriminative features of an individual. Most attention modules are usually trained in a weakly-supervised manner with the final objective, for example, the supervision from the triple loss or classification loss in the person ReID task. However, as the supervision is not specifically designed for the attention module, it may lead to the sub-optimal benefit of attention. To overcome this issue, we propose the self-critical attention module to improve the learning process, permitting to fully exploit the effectiveness of attention. Instead of the weakly-supervised manner, we let the attention

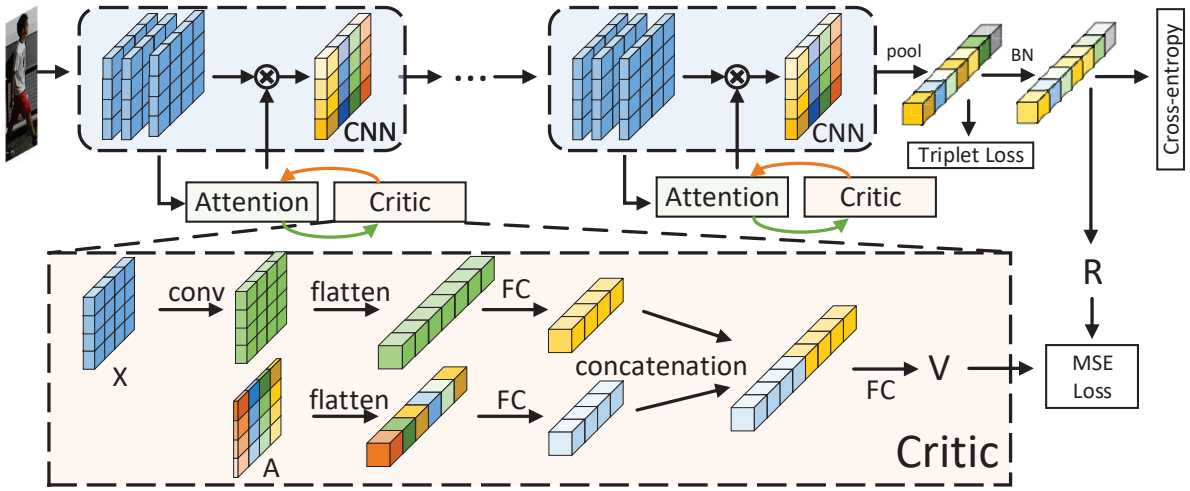

Figure 2. Illustration of the self-critical attention learning method. It is mainly composed of a convolutional backbone network, an attention agent, and a critic module. The backbone consists of a series of convolutional blocks, where we encode the attention maps on top of each block. The critic as 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.

model evaluate itself and guide the optimization with evaluation performance. At each training step, a critic inside the attention module will examine the visual attention map and then transmit a supervisory signal to the attention. With this self-critical supervision, the attention can efficiently figure out whether it is correctly learned and adapt itself.

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

In each step, given the input image I as the state, we first extract the feature maps by the basic network $\mathcal{F}$, which is formulated as

$$X = \mathcal{F}(I|\psi), \tag{1}$$

where ψ denotes the parameters of the basic network. Then the attention agent $\mathcal{A}$ with parameters θ predicts the attention maps A based on these feature maps X.

$$A = \mathcal{A}(X|\theta). \tag{2}$$

To evaluate the attention model and guide the agent to predict more accurate attention, we design a critic module, which is formulated as :

$$V = \mathcal{C}(X, A|\phi), \tag{3}$$

where, V is the predicted evaluation value and ϕ defines the parameters of the critic network. As our critic module is general for different attention agents, we focus on the description of the critic in this section and let the details of the attention agent architecture in the next section. The architecture of the proposed critic module is illustrated in Figure 2. Specifically, it consists of two branches: the state branch employs a convolution layer followed by a fully-connected (FC) layer to extract the state information; while the attention branch applies a single FC layer. Then, the state and attention branches are concatenated and fed in a value-predicted FC layer to output the critic value.

To guide the critic network to predict the actual value of the attention model, we design a reward signal R which reflects our task objective. Specifically, the reward in our experiments includes two parts, the first is the classification criterion R_c denoting whether the attention maps lead to a correct classification, and the second is the amelioration part R_a indicating whether the attention model brings the positive effects. The detail definitions of the classification reward are as follows:

$$R_c = \begin{cases} 1 & y_i^c = y_i^p \\ 0 & y_i^c \neq y_i^p \end{cases}, \tag{4}$$

where y_i^p denotes the prediction label by the attention-based features about person i and the y_i^c is the ground-truth classification label. While the amelioration reward R_a is formulated as:

$$R_a = \begin{cases} 1 & p^k(A_i, X_i) > p^k(X_i) \\ 0 & p^k(A_i, X_i) \leq p^k(X_i) \end{cases}, \tag{5}$$

where p^k indicates the predicted probability of the true classification. The final reward of the attention model is denoted as $R = R_c + R_a$.

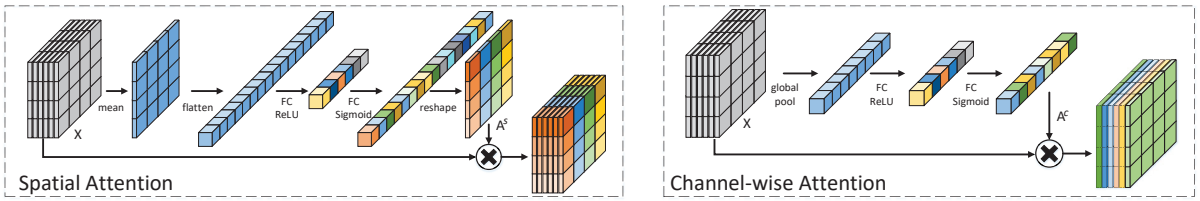

Figure 3. The architectures about spatial and channel-wise attentions. In (a), the spatial attention agent learns a $R^{H\times W}$ attention map to locate the spatial salient region. While in (b), the channel-wise one generates a R^C channel-wise attention vector for feature re-weighting.

3.2. Attention Agent

In this work, we exploit two types of attention model as our attention agent: channel-wise attention and spatial attention.

Spatial Attention: The spatial attention aims to guide the network to focus on most salient regions of the given image. Instead of exploiting all spatial features equally, we discard the irrelevant information and highlight important areas. The proposed spatial attention agent consists of two FC layers, a ReLU layer, and a Sigmod layer. Given the feature maps $X \in \mathbb{R}^{C\times H\times W}$ from the convolutional block, where C is the channels and $H \times W$ denotes the spatial size, the spatial attention agent produces the spatial attention map A^s as:

$$A^s = \sigma\left(W_2^s \max(0, W_1^s \bar{X})\right), \tag{6}$$

where $W_1^s \in \mathbb{R}^{\frac{H\times W}{r}\times(H\times W)}$ and $W_2^s \in \mathbb{R}^{(H\times W)\times\frac{H\times W}{r}}$ correspond to the parameters of two FC layers of the attention agent, respectively. $\bar{X}$ denotes the average across the channel domain of the feature maps, following by a flattening operation. To limit model complexity and improve generalization, we employ a bottleneck structure for our attention agent, where the first FC layer reduces the input dimension C by a ratio r, while the second FC layer restores the dimension. The outputs are then reshaped and expanded to match the shape of the feature maps. Once obtaining the attention maps, we encode the attention information into the feature maps via element-wise production to get the spatially guided feature maps $G = X * A^s$. For more clarity, a detailed architecture of the attention agent is illustrated in Figure 3 (a).

Channel-wise Attention: The different channels of feature maps have specific activation for specific objects. The channel-wise attention aims to enhance the representational ability for various samples by modeling the interdependencies between the convolutional channels. The channel-wise attention agent exploits the "Squeeze-and-Excitation" (SE) block to re-weight the channels of feature maps, by selecting more informative ones and suppressing less useful ones. Specifically, it is composed of a global average pooling layer and two consecutive fully-connected layers. A detail architecture of the attention agent is illustrated in Figure 3 (b). Given the feature maps X, the attention agent produces the channel-wise attention A^c as

$$A^c = \sigma\left(W_2^c \max(0, W_1^c X_{pool})\right), \tag{7}$$

where W_1^c and W_2^c are the parameters of bottleneck FC layers which are similar with spatial ones, and X_{pool} denotes the average pooling on the spatial domain of feature maps X. Differently, the channel-wise attention A^c are applied on the original feature maps via channel-wise multiplication.

Stacked Attention Model: Since it is not trivial to retrieve the most salient features at a single step, we propose to stack multiple attention models at different convolution stages of the backbone network. The model can gradually filter out noises and concentrate on the regions that are highly specific to the identity. The architecture of the stacked attention model is illustrated in Figure 2. Taking the Resnet [10] as an example of the backbone network, we add an attention model on top of each residual block. With the stacked attention structure, the network is progressively guided to focus on the significant features.

3.3. Optimization

The parameters of our network consist of three parts: the backbone network ψ, the attention agent θ, and the critic module ϕ. We design two loss functions to train the backbone network $\mathcal{F}_\psi$ and the attention model $\mathcal{A}_\theta$, including triplet loss and classification loss. 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. It is formulated as:

$$J_{tri}(\psi,\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[||f_i - f_i^+||_2^2 - ||f_i - f_i^-||_2^2 + m\right]_+, \tag{8}$$

where $[\cdot]_+$ indicates the max function $\max(0,\cdot)$, and f_i, f_i^+, f_i^- respectively denote as features of the anchor, positive and negative sample in a triplet. m is a margin to enhance the discriminative ability of learned features. The classification loss focuses on the correctness of predicted identity, which is defined with the cross-entropy:

$$J_{cls}(\psi,\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_i^k \log(p_i^k), \tag{9}$$

Algorithm 1: Self-critical attention learning

Input: Training image data: $\mathbf{I} = \{I\}$, maximal iterative number T, smoothing parameter ϵ, margin m.

Output: The parameters of backbone network ψ, attention model θ ,critic module ϕ

Initialize ψ,θ, and ϕ;
for $t = 1, 2, \ldots, T$ **do**
  Randomly select a batch of images $I_{i=1:N}$ from $\mathbf{I}$;
  Obtain feature map X_i with (1)
  Generate attention A_i with (2)
  Predict critic value V_i with (3)
  Update $\psi \leftarrow \frac{\partial}{\partial \psi}(J_{cls} + J_{tri})$
  Update $\theta \leftarrow \frac{\partial}{\partial \theta}(J_{cls} + J_{tri} + J_{cri})$
  Update $\phi \leftarrow \frac{\partial}{\partial \phi} J_{mse}$
end for
return ψ,θ, and ϕ

where y_i^k is the ground truth identity of ith person on the kth class and p_i^k indicates the predicted probability. In addition, to regularize the model for better generalization ability, we employ the label smooth regularization [40] in our classification loss function. Specifically, we take a uniform distribution $\mu(k) = 1/K$ as the regularization term and reformulate (9) loss as:

$$J_{cls}(\psi, \theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\log(p_i^k)\big((1-\epsilon)y_i^k + \frac{\epsilon}{K}\big), \tag{10}$$

where the $\epsilon \in (0, 1)$ is a smoothing parameter. Since the classification loss is sensitive to the scales of features, we add a batch-norm (BN) layer before classification loss to normalize the scales, as shown in Fig 2. For attention model $\mathcal{A}_\theta$, we introduce an additional powerful supervisory signal predicted by the critic module, which is defined as the critic loss:

$$J_{cri}(\theta) = -\mathcal{V}_\phi^{A_\theta}(X, A). \tag{11}$$

With this critic-based objective function, we update the attention agent to obtain the higher critic value. Finally, we employ the Mean Square Error (MSE) to optimize the critic network $\mathcal{C}_\phi$ by minimizing the gap between the estimated critic value and the real reward. The MSE loss is written as:

$$J_{mse}(\phi) = (\mathcal{V}_\phi^{A_\theta}(X, A) - R)^2. \tag{12}$$

It is worth noting that when we optimize the critic network, the attention network is frozen and vice versa. To explain the optimization more clearly, we provide Algorithm 1 to detail the learning process of SCAL.

3.4. Implementation Details

We employed the ResNet-50 [42] as the basic backbone network for our SCAL method in the experiments, and initialized them with the ImageNet pre-trained parameters. In order to preserve the resolution of the image, we applied a convolution layer with $stride = 1$, instead of original $stride = 2$ convolution layer in the last block of ResNet-50. We 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. During training, we employed three data augmentation methods, including random cropping, horizontal flipping, and erasing. Each mini-batch consists of randomly selected P identities and randomly sampled K images for each identity from the training set to cooperate the requirement of triplet loss. Here we set $P = 24$ and $K = 4$ to train our proposed model. Each input image is resized as 384×192 for exploiting fine-grained information. The margin parameters of triplet loss and the label smoothing regularization rate were set as 0.3 and 0.1 respectively. The weighting coefficients about loss functions $\{J_{cls}, J_{tri}, J_{cri}, J_{mse}\}$ were set as $\{1.0, 1.0, 0.3, 1.0\}$ respectively in the all experiments. We trained our model for 160 epochs in total by the Adam optimizer. The initial learning rate was 0.0004 and was divided by 10 every 40 epochs. The weight decay factor for L2 regularization was set to 0.001. During evaluation, we extracted the features with original images and the horizontally flipped ones and averaged them as the final features. We employed the cosine distance as the metric to measure the similarity of two features. All experiments were implemented with PyTorch 1.0 on 2 Nvidia GTX 1080Ti GPUs. It took about 3 hours with data-parallel acceleration to train the models on the Market-1051 dataset. The above parameter settings were applicable for all three datasets in our experiments.

Table 1. The basic statistics of all datasets in the experiments.

Datasets	**Market-1501**	**DukeMTMC-ReID**	**CUHK03**
Identities	1501	1402	1467
Images	32668	36411	14097
Cameras	6	8	10
Train IDS	751	702	767
Test IDS	750	702	700
Test Setting	SS	SQ	SQ
Labeling	Hand/DPM	Hand	Hand/DPM

4. Experiments

We evaluated our method on three public person ReID benchmarks. In the experiments, we compared the proposed method with other state-of-the-art approaches and conducted ablation studies to analyze our attention model. In addition, we conducted the transfer testing on the cross-dataset to investigate the generalization ability of the SCAL model.

4.1. Experimental Settings

We conducted the experiments on three large-scale datasets including Market-1501 [52], DukeMTMC-ReID [29] and CUHK03 [18]. The detailed statistics and

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	-
AACN [48]	GoogleNet	66.9	85.9	-
DuATM [34]	DenseNet-121	76.6	91.4	97.1
HA-CNN [19]	HA-CNN*	75.7	91.2	-
Mancs [41]	ResNet-50	82.3	93.1	-
SCAL (spatial)	ResNet-50	88.9	95.4	98.5
SCAL (channel)	ResNet-50	**89.3**	**95.8**	**98.7**

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	-
DuATM [34]	DenseNet-121	64.6	81.8	90.2
HA-CNN [19]	HA-CNN*	63.8	80.5	-
Mancs [41]	ResNet-50	71.8	84.9	-
SCAL (spatial)	ResNet-50	**79.6**	**89.0**	95.1
SCAL (channel)	ResNet-50	79.1	88.9	**95.2**

evaluation protocols of all datasets are summarized in Table 1. All the three datasets are collected in a natural real-world scene which is close to the practical application. As shown in Table 1, we followed the standard person ReID experimental setups in [19]. Specifically, we adopted single-query evaluation mode on the Market-1501 dataset in our experiments. For CUHK03 dataset, we applied the CUHK03-NP splits in [54], which selected 767 identities for training and the other 700 ones for testing. For all the datasets, we applied the cumulative matching characteristic (CMC) curve and mean Average Precision (mAP) as the evaluation metric. CMC curves record the true matching within the top n ranks, while mAP considers precision and recall to evaluate the overall performance of methods. To preserve the simplicity and efficiency of the model, we evaluate our method **without post-processings** which are orthogonal to our method and could be integrated in a straightforward manner, such as various re-ranking schemes and metric learning [54, 20].

4.2. Comparison with the State-of-the-Art Methods

In the top groups of Table 2, Table 3, and Table 4, we respectively compared our approach against the state-of-the-art methods on the Market-1501, DukeMTMC-ReID, and CUHK03 datasets. While the bottom group summarizes the performance of deep learning methods with attention model. We observe that the proposed SCAL methods on both spatial and channel domain achieve superior performance over all comparing methods substantially on the three benchmarks. It confirms the effectiveness of the attention evaluator and the self-critical supervisory signal.

For Market-1051 dataset, we selected the single query mode in our experiment and compared with other methods without re-ranking. As shown in Table 2, we evaluated the SCAL method against 13 conventional deep learning methods and 9 attention-based methods. SPReID [12] integrates human semantic parsing in the ReID problem and achieved the best-published result. Our 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. 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. This suggests the importance of the proposed critic module in the attention learning process.

DukeMTMC-ReID is a more challenging person ReID benchmark than Market-1501, due to the more intra-class variations under the wider camera views and more complex background. The performance of the proposed method and other state-of-the-art approaches are summarized in the Table 3. We outperformed the second best method SPReID [12] substantially by 6.3% and 3.0% respectively on the

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**

mAP score and Rank-1 accuracy, which suggests the proposed attention model is an effective manner for the salient location with the cluttered background.

We conducted experiments on both versions of person boxes of the CUHK03 benchmark: manually labeled and auto-detected with a pedestrian detector. We chose the 767/700 identity split rather than 1367/100 since the former is more realistic and challenging. How to learn a robust deep feature representation with limited samples is a common problem of the person ReID systems in the real world. We reported the results of all previous results for both versions in 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 [43] method with the same ResNet-50 base-model.

4.3. Ablation Study

To investigate the contribution of individual components in the SCAL method, we conducted comprehensive ablation evaluations on the Market-1051 dataset in the single query mode. Table 5 shows the comparison results in different settings related to components of SCAL. We separately analyzed each component as follows:

Effect of self-critical module: We compared our SCAL methods with two original attention models, including stacked spatial attention and channel-wise attention. As shown in Table 5, the SCAL methods achieve a significant performance improvement for both spatial attention and channel attention. The consistent improvement over two different basic models demonstrates that the proposed self-critical module is applicable for any attention module.

Spatial attention vs Channel-wise attention: In the experiments, we designed two basic attention model to investigate the generality of the proposed self-critical module. On the Market-1051 and CUHK03 dataset, the channel-based attention usually obtains better performance than the spatial-based one. While on the DukeMTMC-ReID dataset, the spatial-based attention model is superior. We argue that it reflects the images in the DukeMTMC-ReID dataset have lager intra-class spatial variance due to the wider camera views and more cluttered background.

Table 5. Ablation studies of the SCAL method on the Market-1051 dataset with ResNet-50 baseline. Analysis shows the influences of different components and design choices on Rank-1 and mAP (%).

Component	**Design Choice**							
Cross Entropy	✓	✓	✓	✓	✓	✓	✓	✓
Horizontal Flip		✓	✓	✓	✓	✓	✓	✓
Triplet Loss			✓	✓	✓	✓	✓	✓
Label Smooth				✓	✓	✓	✓	✓
Spatial Att					✓		✓	
Channel Att						✓		✓
Self-critical							✓	✓
Rank-1	92.4	92.6	93.5	94.1	94.9	94.9	95.4	95.8
mAP	82.1	82.2	84.1	85.5	87.6	88.1	88.9	89.3

Table 6. Cross-domain evaluation about Market-1051 and DukeMTMC-ReID datasets. M→D indicates that the model is trained on the Market-1501 dataset and tested on the DukeMTMC-ReID dataset, and vice versa.

	M → D		**D → M**	
Method	**mAP**	**R=1**	**mAP**	**R=1**
PTGAN [44]	-	27.4	-	38.6
SPGAN [8]	22.3	41.1	22.8	51.5
Baseline	13.1	25.9	18.8	38.4
SCAN(spatial)	17	30.4	23.1	49
SCAN(channel)	16.4	28.6	23.8	51.7
SCAN(channel)+ SPGAN	**28.4**	**48.4**	**30.4**	**61.0**

Loss functions: We employed the cross-entropy loss as the basic objective functions to optimize our models and additionally improved the performance by introducing the triplet loss as auxiliary rank-based supervisory signal and applying label smoothing regularization (LSR) [40]. As shown in Table 5, the triplet loss obtains the +1.9%/0.9 improvement about mAP/Rank-1 by preserving the rank relationship among a triplet to encourage the intra-class compactness. While LSR further promotes the performance with 1.4% on the mAP score and 0.6% on the Rank-1 accuracy by avoiding the over-fitting.

Horizontal flip: During inference, we average the features from the original and horizontal flipped images, which is a simple trick to reduce the viewpoint variance. As shown in Table 5, it provides about 0.2% gain.

4.4. Cross-Domain Evaluation

In real surveillance systems, it requires intensive human labor to label an overwhelming amount of data. An important evaluation metric about the robustness of ReID system is the generalization ability for unseen persons and scenes.

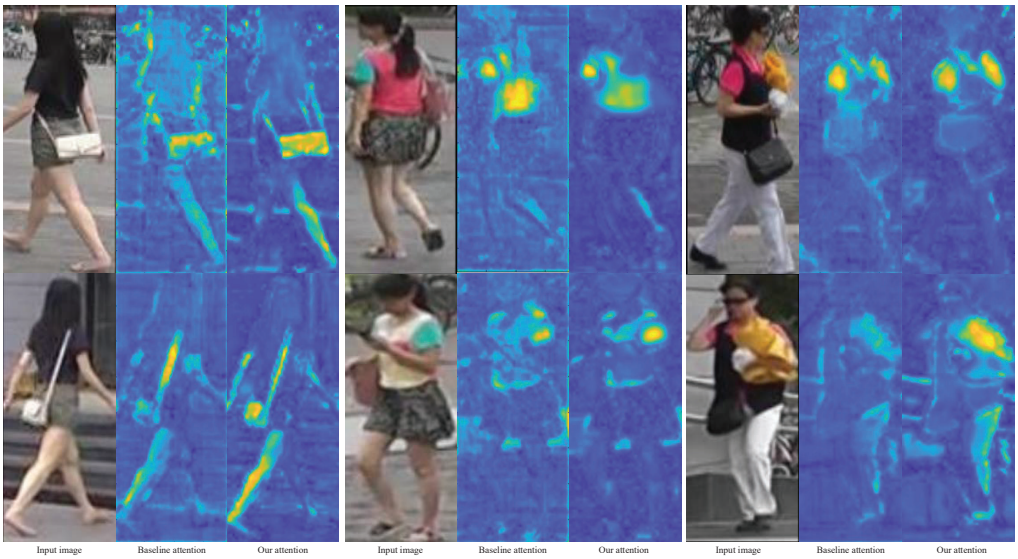

Figure 4. Visualizations of the attention maps. We illustrate three pairs of images, where each pair corresponds to a same individual from the query and gallery sets, respectively. For each sample, from left to right, we show the input image, baseline attention map, and our attention map. We can see that the baseline attention fails to locate the same salient part of people, while our method successes. The comparison results clearly show the effectiveness of our strong supervision for learning more accurate attention. Best viewed in color.

Therefore, we conducted a cross-domain evaluation to investigate the transferability of our SCAL model. Specifically, we trained the model with the data in the Market-1051 dataset and tested it with the samples in the DukeMTMC-ReID dataset, and vice versa. We applied the ResNet-50 network and trained it with cross-entropy+triplet+LSR loss functions as the baseline. As shown in Table 6, both spatial- and channel-wise SCAL methods outperform the baseline by a large margin, which demonstrates the generalization ability of the SCAL methods. Compared with the state-of-the-art transfer learning methods PTGAN [44] and SPGAN [8], we still achieve competitive performance. It is worth pointing out that the person images of the test domain are visible for PTGAN and SPGAN methods in the training process. While in our experiments, both **images and labels** in the test-domain are unseen to evaluate the generalization ability of the proposed attention model. In addition, with the same setting as SPGAN [8], by transferring the style of source-domain into target-domain but replacing the feature extraction part of SPGAN by our SCAL model, we further improve the performance.

4.5. Qualitative Analysis

In order to validate the effectiveness of our self-critical attention learning method, we qualitatively examined the attention maps and the associated critic value. Some examples of visualization are illustrated in Figure 4. Specifically, we chose two images of the same individual from query and gallery sets, respectively. We expected to observe that the attention helps to focus on the same discriminative parts of the person. We can see that the salient features of the same target are highlighted, such as the bag, T-shirt. These qualitative results demonstrate the effectiveness of our SCAL model which guides the network to focus on highly relevant regions. Besides, we also compare our attention maps with the baseline attention map. As shwon in Figure 4 whose middle column is baseline attention map and the right one is our attention map, the proposed critic model provides strong supervision for learning more accurate attention.

5. Conclusion

In this paper, we have proposed a simple yet effective self-critical attention model for person re-identification. Instead of the weak supervision, we learn the attention with a critic which examines the gain from the attention over the backbone network and provides a strong supervisory signal based on its observation. Moreover, the critic can measure the quality of the attention maps which significantly facilitates the interpretation of attention for human end-users. Extensive experimental results show that the proposed self-critical attention learning method outperforms existing state-of-the-art methods by a large margin, which validate the effectiveness of our approach.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, Grant 61672306, and Grant 61572271.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015.

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, pages 2109–2118, 2018.

[3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, June 2018.

[4] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Spatial-temporal attention-aware learning for video-based person re-identification. *TIP*, 28(9):4192–4205, 2019.

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017.

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *CVPR*, pages 1335–1344, 2016.

[7] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, pages 1983–1991, 2017.

[8] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *CVPR*, pages 994–1003, 2018.

[9] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NIPS*, pages 1230–1241, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[12] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, pages 1062–1071, 2018.

[13] Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, and Stefano Soatto. Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In *ECCV*, pages 715–733, 2018.

[14] Xu Lan, Hanxiao Wang, Shaogang Gong, and Xiatian Zhu. Deep reinforcement learning attention selection for person re-identification. *BMVC*, pages 4–7, 2017.

[15] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.

[16] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017.

[17] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018.

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014.

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, page 2, 2018.

[20] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015.

[21] Shengcai Liao and Stan Z Li. Efficient psd constrained asymmetric metric learning for person re-identification. In *ICCV*, pages 3685–3693, 2015.

[22] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. Graininess-aware deep feature learning for pedestrian detection. In *ECCV*, pages 732–747, 2018.

[23] Ji Lin, Liangliang Ren, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *CVPR*, 2017.

[24] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 2017.

[25] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, pages 4099–4108, 2018.

[26] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017.

[27] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016.

[28] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NIPS*, pages 2204–2212, 2014.

[29] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016.

[30] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *CVPR*, June 2018.

[31] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with deep similarity-guided graph neural network. In *ECCV*, September 2018.

[32] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *CVPR*, pages 6886–6895, 2018.

[33] Hailin Shi, Yang Yang, Xiangyu Zhu, Shengcai Liao, Zhen Lei, Weishi Zheng, and Stan Z Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, pages 732–748, 2016.

[34] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*, pages 5363–5372, 2018.

[35] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, pages 1179–1188, 2018.

[36] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *ICCV*, 2017.

[37] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, September 2018.

[38] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *ICCV*, pages 3820–3828, 2017.

[39] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, pages 480–496, 2018.

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[41] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*, September 2018.

[42] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017.

[43] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *CVPR*, pages 8042–8051, 2018.

[44] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[45] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10(Feb):207–244, 2009.

[46] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, pages 1249–1258, 2016.

[47] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16, 2014.

[48] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, pages 2119–2128, 2018.

[49] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017.

[50] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, pages 3219–3228, 2017.

[51] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016.

[52] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[53] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, Qi Tian, et al. Person re-identification in the wild. In *CVPR*, volume 1, page 2, 2017.

[54] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017.

[55] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *CVPR*, pages 5157–5166, 2018.

[56] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, July 2017.