Discriminative Clue Alignment Network for Both Image- and Video-Based Person Re-Identification

Panwen Hu[1,2]
panwenhu@link.cuhk.edu.cn
Xinyu Zhou[1]
xinyuzhou@link.cuhk.edu.cn
Rui Huang *[1,2]
ruihuang@cuhk.edu.cn

[1] The Chinese University of Hong Kong, Shenzhen, Guangdong, China
[2] Shenzhen Institute of Artificial Intelligence and Robotics for Society, Guangdong, China

Abstract

The body misalignment problem resulted from various factors, e.g., scale and pose variation, occlusion, etc., has always been a great challenge in person re-identification. It is intuitive to model individual body parts and match them through pose estimation or human parsing, which, however, requires additional annotations and training, and may fail with occlusion. Some recent studies employ self-attention mechanisms to discover the Discriminative Clues (DCs) on the body. But, unlike the body parts, the DCs are not naturally aligned properly. To this end, we propose a Discriminative Clue Alignment Network (DCANet), along with a discriminant constraint, to automatically identify various DCs and then align them into a fixed pattern, without explicitly analyzing body parts. Moreover, such an alignment scheme makes the temporal aggregation of features from video frames extremely simple, because the DCs are effectively aligned across frames. Therefore our method can be easily applied to video-based person re-identification as well. Experiments on several popular public benchmarks show that DCANet can achieve state-of-the-art performance on both image- and video-based re-identification tasks.

1 Introduction

Person Re-IDentification (ReID) has drawn significant attention due to its wide applications in tracking people across cameras, searching people in a large gallery, or grouping personal images, etc. The fundamental problem of ReID is to compare a pair of images or videos, each depicting a person's body appearance, possibly taken by different cameras at different times, and determine whether they belong to the same person. In the past decade, many methods [23, 27] have been proposed to improve ReID by addressing various difficulties. With the rapid development of deep learning, training a deep neural network to generate the appearance representations in an end-to-end fashion has dominated the recent ReID studies. Although a large variety of network architectures along with various metric learning loss functions have been proposed, and great progress has been made, the new methods still suffer from the same problems.

The body misalignment problem resulted from various factors, e.g., scale and pose variation, occlusion, etc., has always been one of the most difficult challenges. For example,

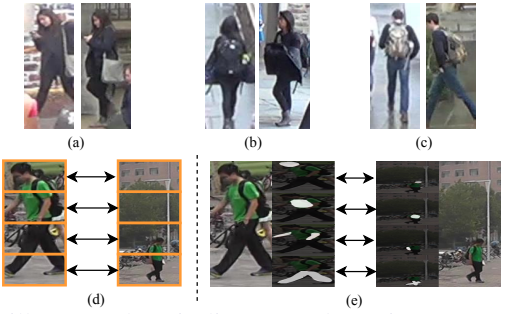

Figure 1: The first row illustrates the misalignments due to inaccurate detection (a), body deformation (b), and occlusion (c). In the second row, the conventional stripe-based method (d) cannot ensure the body part clues of two images are aligned, hence we propose to automatically extract the hidden DCs (highlighted with white masks) i.e., head, body, and arrange their features into an ordered pattern (e).

the person may appear in different locations and scales in the bounding box given by the detection algorithms (Fig.1(a)); pose variation and non-rigid body deformation may lead to completely different configuration of the body parts (Fig.1(b)); the occlusion of body parts may introduce irrelevant context (Fig.1(c)), etc. Therefore, directly training a feature model without considering the spatial structures can easily cause mismatch of the feature vectors.

Many ReID methods have tried to address this problem. Some approaches divided the whole image into multiple grid cells [15] or horizontal stripes [16, 19] to represent the body parts, which oversimplify the body part configuration and the matching procedure, as shown in Fig.1(d). Others [17, 35] employ the self-attention mechanisms to discover the Discriminative Clues (DCs), e.g., the context of body parts, the attributes, some inexplicable cues, on the body and strengthen them in the final features. These DCs are more sophisticated than the cells or stripes, but, unlike the body parts, the DCs are not naturally aligned properly. Some recent studies attempted to generate the part-aligned or DC-aligned features under the guidance of pose estimation [34] or human parsing results [5], which, however, require additional annotations and heavily rely on the accuracy of the pose estimator or the human parsing model.

Some researchers started to move on to video-based ReID [10, 25, 30, 39], because videos naturally contain more poses of the same person and therefore cover larger pose variation than images do. Occlusion may also appear only in part of the videos, hence may become a smaller problem. In video-based ReID, the features of the person are extracted from an image sequence/video to aggregate more information along the time axis to generate more robust global representations. However, most of the existing video-based ReID methods mainly focus on exploiting the temporal relations among the video frames, e.g., [7, 11, 32], or the spatio-temporal clues, e.g., [8, 29]. Due to the nature of such methods, they cannot be applied to image-based ReID and have limited applications.

We argue that even if we do not have sophisticated human parsing models to extract the body parts, or videos to cover larger pose variation, we can still align the DCs for image-based ReID. Moreover, if we align the DCs in images effectively, the frame-wise features extracted by our model can be easily aggregated (e.g., through simple temporal averaging)

*Rui Huang is the correspondence author. This work is supported in part by funding from Shenzhen Institute of Artificial Intelligence and Robotics for Society, and Shenzhen NSF JCYJ20190813170601651.

to achieve promising performance on video-based ReID as well. To this end we propose a Discriminative Clue Alignment Network (DCANet) to automatically identify various DCs and then align them into a fixed feature pattern. Specifically, we treat the feature vector of an image as a fixed pattern and propose a Discriminative Clue Alignment Module (DCAM) to extract multiple semantic DCs that are filled into the corresponding parts of the feature pattern. The main idea is illustrated as Fig.1(e). Thus the semantic DCs are inherently aligned for the calculation of the feature distance, without the dynamic DC matching process [17, 28]. Besides, we further introduce a Discriminant Loss (DL) to ensure that different DCs are discovered, and the parts of the feature pattern are distinguishable. More details will be discussed in Sec.3.

In summary, the main contributions of this work are three-fold. (1) As traditional DC-based methods cannot ensure that DCs are aligned in the feature space, we propose DCAM, which adopts the multi-attention mechanism, to extract and align DC features into a fixed pattern. (2) We further design a discriminant loss to diversify the focuses of multiple attention blocks in DCAM, so that each part of the final feature vectors has its own semantics. (3) We show that the proposed method can be easily and effectively extended to video-based ReID. In Sec.4, both quantitative and qualitative results are presented to prove its effectiveness in both image- and video-based ReID.

2 Related Works

Image-based ReID methods for misalignment. Different categories of methods have been proposed to tackle the misalignment problem. Some studies attempted to split the images or feature map into small patches [15] or stripes [16, 19], thereby the local features are extracted from these patches or stripes using the hand-crafted descriptors [16] or neural networks[19]. However, the semantics of the stripes in the query and gallery images are not aligned explicitly. Other methods [5, 8, 13, 34] applied a pose estimator or human parsing model to accurately extract the features of body parts. With the generated body part mask, these methods were able to compute the part-aligned representations. However, additional accurate parsing models are required to ensure high ReID performance for these methods. Some studies proposed to dynamically calculate the shortest distance of the local features as a supervision signal to train the model [17], or dynamically match the local features when calculating the global distance [28], yet the exhaustive calculation results in high computational burden. The attention-based approaches [20, 35] used the attention modules to suppress the identity irrelevant context while strengthening the DCs in the final representations, however the DCs are not aligned in nature.

Video-based ReID methods. As videos provide richer spatial appearance and temporal cues, video-based ReID has become more and more prevalent. Apparently, the misalignment problem is less severe in this task, since one can aggregate more pose information into a single feature vector. For example, Zhang et al. [30] proposed to temporally align the human gaits by extracting the walking cycles from the videos. Many recent methods devote a large amount of effort on highlighting the DCs with the attention mechanisms [10, 33]. The works like [25, 26] employed the recurrent neural network to extract the features from the sequential frames, and some other approaches [7, 11] learned a temporal attention module to adaptively assign different importance scores to the video frames when aggregating the features. However, these models obviously cannot be directly trained on image-based datasets and applied to image-based Re-ID. Our work, instead, is competitive in both image- and video-based ReID.

3 Approach

In this section, we present our simple yet effective DCANet. As shown in Fig.2, it mainly consists of a backbone network with the DCAMs inserted into different stages. In line with previous works, we employ ResNet50 as the backbone. Our DCANet takes as input a single image I or a video (frame sequence) $V = \{I_1, I_2, \ldots, I_T\}$ to extract the frame-wise features f or $\{f_1, f_2, \ldots, f_T\}$. The final video representation is aggregated by the average pooling, $f = \frac{1}{T}\sum_{t=1:T} f_t$. For conciseness, we will describe the feature extraction for a single frame in Sec.3.1. Moreover, to ensure that different DCs can be identified by the DCAM, unlike previous methods [3, 8, 29] that utilize self-attention mechanism, we also propose a supervision constraint, or discriminant loss, as explained in Sec.3.2.

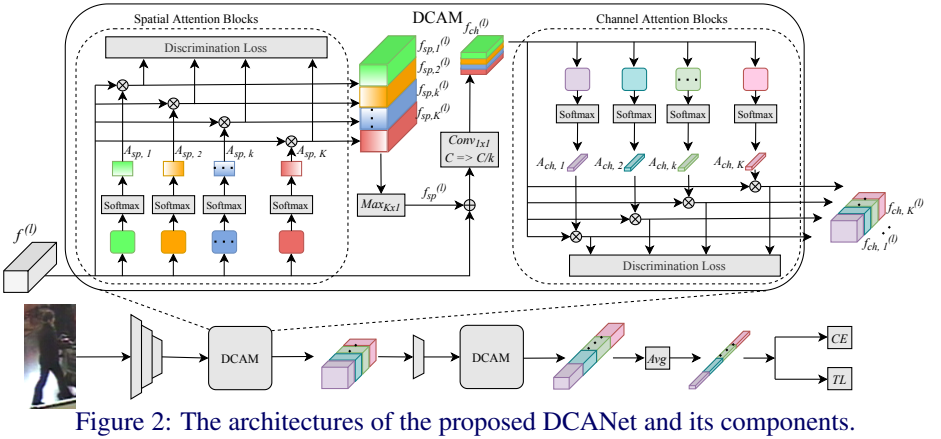

Figure 2: The architectures of the proposed DCANet and its components.

3.1 Aligned feature extraction

As shown in Fig.2, the input image is first passed to the shallow layers of the backbone to generate the semantic features $f^{(l)} \in \mathbf{R}^{C\times H\times W}$, where l denotes the l-th stage of the backbone. These will then be passed to DCAM, which strengthens the DCs and aligns them thereby. Different from [10, 35], which compute the attentions of DCs merely along spatial dimension, DCAM adopts dual multi-attention module to identify various DCs and align them along both spatial and channel dimensions. Specifically, in the spatial module of DCAM, K spatial attention blocks followed by the $Softmax$ operation are introduced to compute K attention distributions $\{A_{sp,k} | A_{sp,k} \in \mathbf{R}^{H\times W}; k \in 1,\ldots,K\}$, where the subscript sp, k indicates the spatial attention from the k-th block. Each attention block focuses on its own DCs spatially with two 1×1 convolutional layers and one ReLU activation in between. Thus the feature $f^{(l)}_{sp,k} \in \mathbf{R}^{C\times H\times W}$ with the DCs highlighted by spatial attention $A_{sp,k}$ is represented as:

$$f^{(l)}_{sp,k}(c,x,y) = f^{(l)}(c,x,y) \times A_{sp,k}(x,y) \tag{1}$$

Unlike previous method [10] which applies maximum pooling on $\{f^{(l)}_{sp,k} | f^{(l)}_{sp,k} \in \mathbf{R}^{C\times H\times W}; k = 1,\ldots,K\}$ along the attention axis, we stack them vertically, followed by the operation of maximum pooling $Maxpool_{K\times 1}$ with the kernel size of $K\times 1$ to preserve the salient infor-

mation. The final spatially aligned feature $f_{sp}^{(l)} \in \mathbf{R}^{C\times H\times W}$ is therefore computed as:

$$f_{sp}^{(l)} = Maxpool_{K\times 1}([f_{sp,1}^{(l)}|f_{sp,2}^{(l)},\cdots,|f_{sp,K}^{(l)}]) \tag{2}$$

where $|$ denotes the concatenation operation along the vertical axis. As a result, each attention block has its own focused DCs, and the position of each DC is fixed, e.g., the DCs found by the first block are presented in the first part of $f_{sp}^{(l)}$, and the second part contains some other fixed DCs found by the second block, etc. Thereby the semantics of the features are aligned.

Similar to most of previous approaches, e.g., the studies [16, 19, 20], we also preserve the channel dimension in the final image representation. Hence, to align the features directly for distance calculation, we further propose a channel multi-attention structure in DCAM, which functions as the spatial attention to find the DCs and then align them along the channel axis. Specifically, K channel attention blocks take as input $\hat{f}_{sp}^{(l)} = f_{sp}^{(l)} \oplus f^{(l)}$ under the skip connection structure, where $\oplus$ is the element-wise summation. These will be used to generate K channel attention score vectors. $\hat{f}_{sp}^{(l)}$ is first passed through a 1×1 channel reduction convolutional layer to generate $f_{ch}^{(l)} \in \mathbf{R}^{C/K\times H\times W}$, then a maximum pooling $Maxpool$ and an average pooling $Avgpool$ operations followed by K two-layer fully connected blocks FC_k are applied to compute K channel attention vectors $\{A_{ch,k}|A_{ch,k} \in \mathbf{R}^{C/K}; k = 1,\ldots,K\}$, so $A_{ch,k}$ is expressed as:

$$A_{ch,k} = Softmax(FC_k(Maxpool(f_{ch}^{(l)})) \oplus FC_k(Avgpool(f_{ch}^{(l)}))) \tag{3}$$

Similar to spatially aligned features, channel weighted features $\{f_{ch,k}^{(l)}|f_{ch,k}^{(l)} \in \mathbf{R}^{C/K\times H\times W}; k = 1,\ldots,K\}$ are derived as:

$$f_{ch,k}^{(l)}(c,x,y) = A_{ch,k}(c) \times f_{ch}^{(l)}(c,x,y). \tag{4}$$

We concatenate those K channel weighted features $f_{ch,k}^{(l)}$ along the channel dimension to obtain the channel aligned feature $\hat{f}_{ch}^{(l)} \in \mathbf{R}^{C\times H\times W}$. The output of the overall network, i.e., the final image representation is attained by applying a spatial average pooling operation on $\hat{f}_{ch}^{(l)}$. In this representation, DCs found by the channel attention blocks are already aligned.

3.2 Objective functions

Discriminant loss It is noticeable that the DCANet aligns the features under the assumption that different attention blocks can identify different DCs. Therefore, instead of using the self-attention mechanism without any constraints, we propose to impose a DL on the dual multi-attention modules to diversify their focuses. Ideally, different $f_{sp,k}^{(l)}$ or $f_{ch,k}^{(l)}$ contain the features of different DCs, which are discriminative. To achieve this goal, we apply two cross entropy objective functions on the training of the spatial and channel multi-attention modules, respectively. Specifically, the spatial dimensions of each $f_{sp,k}^{(l)}$ and $f_{ch,k}^{(l)}$ are reduced to 1 first by an average pooling operation. Then we use two fully connected layers as two discriminators to classify $f_{sp,k}^{(l)}$ and $f_{ch,k}^{(l)}$ into K classes, meaning that each $f_{sp,k}^{(l)}$ or $f_{ch,k}^{(l)}$ is from one of the K attention blocks. Consequently, K attention blocks are capable of identifying different and distinguishable DCs.

For spatial attention, let $P(j|f_{sp,k}^{(l)})$ denotes the probability that $f_{sp,k}^{(l)}$ is from the j-th block, so the loss function is written as:

$$L_{sp} = \sum_{k=1}^{K}\sum_{j=1}^{K} -y_j \log(P(j|f_{sp,k}^{(l)})) \tag{5}$$

where y_j equals to 1 if $j = k$, otherwise it is 0. We also design a loss function L_{ch} with the same idea on the channel attention module. Hence the discriminant loss L_{dis} is the sum of L_{sp} and L_{ch}.

Other ReID Losses In addition to the proposed DL used to constrain the attention modules, we further introduce the cross entropy loss for person identity classification and the triplet loss in the training process, following previous works[17, 16].

4 Experiments

4.1 Datasets and metrics

Datasets Unlike previous studies that adopt only Image-based ReID (IReID) datasets [5, 12, 31, 35] or Video-based ReID (VReID) datasets [6, 29] as benchmarks, we evaluate our DCANet on both to verify its effectiveness and the generalization ability.

IReID: *CUHK03*, *DukeMTMC-ReID*, and *MSMT17* datasets are adopted as the image-based benchmarks. *CUHK03* dataset contains 14,097 images of 1,467 pedestrians, and we split it into training (767 identities) and testing (other 700 identities) sets following the more challenging protocol [38]. The manually labeled bounding boxes are used in the evaluation. The *DukeMTMC-ReID* includes 36,411 labeled images from 1,404 identities, where 16,522 images of 702 identities are selected as the training set, and the rest are split into query and gallery sets as the testing set. *MSMT17* is the most challenging and large-scale dataset, which consists of 126,441 bounding boxes from 4,101 pedestrians captured by 15 cameras.

VReID: The VReID datasets we adopt include *MARS*, *DukeMTMC-VideoReID*, and *iLIDS-VID*. *MARS* dataset is one of the largest public datasets, consisting of 20,715 sequences including 3,248 distracting tracklets, and it contains 1,261 pedestrians captured by at least 2 cameras, out of 6 cameras. *DukeMTMC-VideoReID* is another large-scale dataset which is composed of 4,832 tracklets from 1,812 pedestrians captured in outdoor scenes. For these two datasets, we follow the original protocols to split the training and testing sets, ensuring that no overlaping identities exist. *iLIDS-VID* is a small but typical dataset which consists of 600 videos from 300 identities. We randomly split the probe/gallery identities to construct the training and testing sets following the protocol of [22].

Metrics To quantitatively evaluate our approach, we adopt the widely-used evaluation protocols [1]. The matching process performs the similarity calculation between query and gallery images first, and then sorts the gallery images according to the similarities. The performances are evaluated by the Cumulative Matching Characteristic (CMC) curve, which is an estimate of the expectation of finding the correct match in the top n matches. The mean Average Precision (mAP) scores are also reported.

4.2 Implementation details

The experiments are implemented on Pytorch platform with one Nvidia GeForce RTX 2080Ti GPU. There are 8 identities in each batch, and 4 samples of each identities are included. The images or video frames are resized to 256×128, following by a random horizontal flip operation for data augmentation. We adopt the Adam optimizer with weight decay 0.0005 to update the parameters. The learning rate is initialized to 3×10^{-4}, and decreased by $\times 0.1$ every 60 epochs, and there are 130 epochs in total. The weight for the DL is set to 0.5 for all datasets. When training the VReID model, we randomly sample 4 frames with a stride of 8

frames from each tracklet to extract the frame-wise features, and the average feature along the temporal axis is used to represent the video clip. In the testing process, the final video representation is the averaged feature of all frames.

4.3 Ablation study

For notation convenience, we use DI, DV, and **BS** to represent the *DukeMTMC-ReID*, *DukeMTMC-VideoReID*, and the baseline model (ResNet50) in the rest of this paper, respectively. The bold-type and the underlined numbers indicate the best and second best results in each table, respectively.

Impact of the positions to place the DCAM As mentioned before, the proposed DCAM can be inserted behind any stages of the backbone, hence we study influence of the positions with $K = 8$ in this part. Table 1 compares the results of inserting the DCAMs into different stages, where $stage_l$ denotes the DCAM is inserted behind the stage l. It can be seen that adding the DCAMs into any stages of the backbone can improve the performance by a large margin, compared to **BS**. Moreover, the improvements by $+stage_3$ are greater than that of $+stage_2$ and $+stage_4$ among the results of single stage insertion. It is likely that the low-level features in $+stage_2$ are insufficient to provide precise semantic information, thus the attention blocks can not identify the DCs precisely. In contrast, the spatial dimension of the features in $stage_4$ is small, so the spatial cues found are limited.

Methods	CUHK03		DI		MSMT17	
	R1	mAP	R1	mAP	R1	mAP
BS	73.7	69.8	85.9	71.8	73.8	47.2
$+stage_2$	78.4	75.8	88.7	78.1	79.2	**56.7**
$+stage_3$	**81.9**	**77.6**	**89.2**	**78.3**	**79.3**	56.3
$+stage_4$	79.3	75.7	87.9	76.1	76.9	52.7
$+stage_{2,3}$	79.7	76.7	89.0	78.1	**79.9**	57.5
$+stage_{3,4}$	79.4	76.8	89.3	78.1	79.6	56.2
$+stage_{2,3,4}$	**81.4**	**78.1**	**89.9**	**78.7**	**79.9**	**57.6**

Table 1: Impacts of inserting DCAMs into different stages.

Compared to the single stage insertion, the multi-stage insertion achieves better overall performance, especially at $stage_{2,3,4}$. This is due to the fact that multi-grained features are exploited as the network inference proceeds, and the coarse-to-fine DCs are aligned during this process, thus the final representations are more discriminative. However, this is not always consistent. For example, the rank-1 accuracy of $+stage_{2,3,4}$ is slightly lower than that of $+stage_3$ on *CUHK03* dataset. This might be because the number of attention blocks K is identical for all the DCAMs in different stages, whereas the feature maps have different sizes and channel dimensions. Hence there is a possibility that varying K according to the stages of DCAMs, or even the types of the attention modules (spatial or channel), may lead to greater improvements. We will use $+stage_3$ as a basis in the following analysis.

Impact of the attention block number To investigate the influence of K, we experiment on the IReID datasets with the structure of $stage_3$, while varying K from 2 to 16. From Table 2, we can observe that as K is increased from 2 to 8, the performances are getting better, since the network is able to discover more DCs. Furthermore, DL enforces multiple attention blocks to focus on different DCs, and the attention maps are independent, hence the attention magnitudes on the DCs will be stronger as K increases. When K is sufficiently large, e.g., $K = 8$, the network will reach the maximum performance. However, an overlarge K may hurt the performance, since the sizes of feature maps in deep backbone layers are

small, the information exploited by each attention block become more and more limited as K increases.

K	CUHK03		DI		MSMT17	
	R1	mAP	R1	mAP	R1	mAP
2	78.8	75.6	88.8	78.0	78.7	55.8
4	79.5	76.1	89.0	78.1	79.0	56.2
8	**81.9**	**77.6**	**89.2**	**78.3**	**79.3**	**56.3**
16	80.4	77.0	89.1	77.8	79.2	56.2

Table 2: The results using different number K of attention blocks.

Methods	iLIDS-VID	DV		MARS	
	R1	R1	mAP	R1	mAP
AlignReID	72.0	83.5	80.7	77.9	64.5
DuATM	-	-	-	81.2	67.7
MGN	82.7	90.2	88.9	85.7	77.5
BS	82.7	90.6	89.7	80.8	74.0
BS + DCAM	88.7	95.6	95.1	89.0	83.9
DCANet (**Ours**)	**90.7**	**96.6**	**95.9**	**89.6**	**84.5**

Table 3: The results of image-based methods and DCANet on VReID benchmarks

Generalization performance of DCANet To verify the generalization ability of the DCANet and its components, we implement other two popular IReID methods, a dynamically matching-based method AlignReID [12] and a part-based method MGN [21], on the VReID benchmarks with the same testing manner as that of DCANet. Table 3 lists the performances of these methods, along with the reported results of an attention-based method DuATM [14]. It can be observed that our DCANet surpasses the traditional image-based methods on all three VReID datasets. Although our DCANet is proposed to align frame-wise features as AlignReID and MGN, it is also suitable for VReID task with the average temporal feature integration. Moreover, the improvements by introducing DCAM alone and both of DCAM and DL further verify the effectiveness of the proposed DCAM and DL.

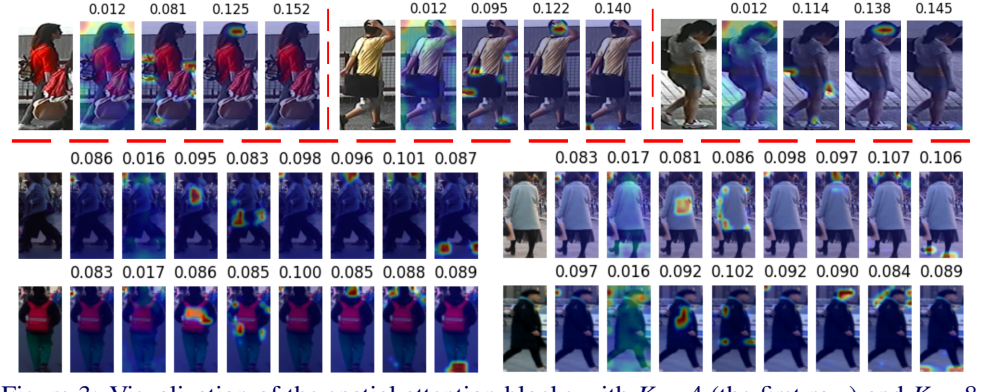

Figure 3: Visualization of the spatial attention blocks with $K = 4$ (the first row) and $K = 8$ (the last two rows). In each sample, the leftmost image is the input image, and the left K images present the attention maps of K blocks, each block has its own focused clues.

Visualization analysis The attention maps from $K(= 4, 8)$ spatial attention blocks are visualized as Fig.3, and the number above each image represents the maximum attention score. It can be seen that different blocks can focus on different cues, and each block has its own interests. For example, in the case of $K = 4$, the second block tends to discover the accessories, e.g., bags, the third one focuses on the human head, etc. In the case of $K = 8$, the third, the

sixth, and the eighth blocks pay attention to the body, head, and feet, respectively. In our DCANet, all these found cues are filled into an ordered pattern to attain the well-aligned features. Furthermore, the last three blocks prone to capture the identity-relevant cues, therefore their maximum attention scores are higher than that of the first block, which produces the dispersed attentions on the background.

4.4 Comparisons with related approaches

Comparisons on the IReID datasets Table 4 summarizes the comparisons with the IReID state of the arts, where AlignReID [12] is a **D**ynamic alignment method, PCB [16], MGN[21], Pyramid[37], HPM [4] belong to the **P**art-based methods, DSA-reID[18], SAN[8], P^2-Net [5], PGFA[13] are classified into the **H**uman pose or part parsing-based (**H**-based) methods, and DuATM [14], Mancs[20] are the **A**ttention-based (**A**-based) methods. Our DCANet achieves the highest performances on all datasets among the **A**-based methods. Unlike Mancs which learns the salient information from only the channel dimension, our DCANet employs the dual multi-attention strategy to identify different DCs from both spatial and channel dimensions, and then aligns them into a fixed pattern. Thus the features from our method are more representative. Furthermore, our method also surpasses other two kinds of methods on almost all the metrics except for the mAP on the DI dataset, even though the **H**-based methods require additional pose or part annotations.

Methods	CUHK03		DI		MSMT17	
	R1	mAP	R1	mAP	R1	mAP
AlignReID(**D**)	61.5	59.6	82.1	69.7	69.8	43.7
PCB(**P**)	63.7	57.5	83.3	69.2	-	-
MGN(**P**)	68.0	67.4	83.3	69.2	-	-
HPM (**P**)	63.9	57.5	86.6	74.3	-	-
Pyramid(**P**)	78.9	<u>76.9</u>	<u>89.0</u>	**79.0**	-	-
P^2-Net(**H**)	78.3	73.6	86.5	73.1	-	-
PGFA(**H**)	-	-	81.9	65.3	-	-
DSA-reID(**H**)	78.9	75.2	86.2	74.3	-	-
SAN (**H**)	<u>80.1</u>	76.4	87.9	75.5	<u>79.2</u>	<u>55.7</u>
DuATM(**A**)	-	-	81.8	64.6	-	-
Mancs (**A**)	69.0	63.9	84.9	71.8	-	-
DCANet(**A**)	**81.9**	**77.6**	**89.2**	<u>78.3</u>	**79.3**	**56.3**

Table 4: Comparison with state-of-the-arts on IReID datasets.

Comparisons on the VReID datasets In Table 5, we further compare our method with the VReID state of the arts, which include the **T**emporal-sequence based (**T**-based)methods, e.g., Snippet[2], STA[3], SCAN [29], GLTR[9], VRSTC[6], MG-RAFA[32], TCLNet [7], and the **I**mage set based (**I**-based) methods, e.g., EUG[24], AttDriven [36], on the VReID datasets. The results show that our DCANet outperforms all the **I**-based methods on all datasets consistently. Compared with the **T**-based methods, our method can still achieve the best result on *iLIDS-VID*, and the second best results on the DV dataset. Whereas the accuracies of our DCANet on the DV and MARS datasets are lower than that of TCLNet, because our method dose not exploit the temporal cues among the frames. Nevertheless, the competitive performances achieved can still verify the effectiveness of our approach, and the importance of the feature alignment in both IReID and VReID tasks.

Qualitative results To qualitatively verify the superiority of our DCANet, we visualize the top-5 retrieval results of the stripe-based method, MGN [21], human parsing-based method, P^2-Net [5] and our approach in Fig.4. The images in the first column of each sub-figure are the query images, and the last five columns are the top-5 similar images to each query image.

Methods	iLIDS-VID	DV		MARS	
	R1	R1	mAP	R1	mAP
Snippet(**T**)	85.4	-	-	86.3	76.1
STA(**T**)	85.3	96.2	94.9	86.3	80.8
SCAN(**T**)	88.0	-	-	87.2	77.2
GLTR(**T**)	86.0	96.3	93.7	87.0	78.5
VRSTC(**T**)	83.4	95.0	93.5	88.5	82.3
MG-RAFA(**T**)	88.6	-	-	88.8	**85.9**
TCLNet(**T**)	86.6	**96.9**	**96.2**	**89.8**	85.1
EUG(**I**)	-	83.6	78.3	80.8	67.4
AttDriven(**I**)	86.3	-	-	87.0	78.2
DCANet(**I**)	**90.7**	96.6	95.9	89.6	84.5

Table 5: Comparison with state-of-the-arts on VReID datasets.

The number above each image indicates the identity of the person. It can be seen that the top-1 results of MGN and P^2-Net are all failure cases, due to the occlusion (the first row for MGN, the second row for P^2-Net), the similar appearance (the second row for MGN, the first row for P^2-Net). In contrast, our DCANet successfully retrieves the correct images for all these challenging query images, due to its ability to discover and align DCs.

Figure 4: The retrieving results of different methods. The first two rows show the comparisons between the proposed DCANet and a stripe-based method, MGN. The last two rows compare DCANet with P^2-Net, which utilizes additional human parsing results.

5 Conclusions

Body misalignment is a great challenge in both image- and video-based ReID tasks. Although many attempts have been made to tackle the problem, these methods have limitations and usually cannot work well on both tasks. In this paper, we propose a flexible network, DCANet, to extract and align the frame-wise discriminative features with a dual multi-attention strategy. Extensive experiments, including the ablation studies and the comparisons with the state-of-the-art methods, on both IReID and VReID benchmarks, have been carried out to demonstrate the effectiveness of our approach.

References

[1] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3908–3916, 2015.

[2] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018.

[3] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8287–8294, 2019.

[4] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8295–8302, 2019.

[5] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3642–3651, 2019.

[6] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019.

[7] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *European Conference on Computer Vision*, pages 388–405. Springer, 2020.

[8] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11173–11180, 2020.

[9] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3958–3967, 2019.

[10] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

[11] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8786–8793, 2019.

[12] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 94:53–61, 2019.

[13] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 542–551, 2019.

[14] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5363–5372, 2018.

[15] Arulkumar Subramaniam, Moitreya Chatterjee, and Anurag Mittal. Deep neural networks with inexact matching for person re-identification. In *Advances in Neural Information Processing Systems*, pages 2667–2675, 2016.

[16] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018.

[17] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7134–7143, 2019.

[18] Chaoping Tu, Yin Zhao, and Longjun Cai. Esa-reid: Entropy-based semantic feature alignment for person re-id. *arXiv preprint arXiv:2007.04644*, 2020.

[19] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang. A siamese long short-term memory architecture for human re-identification. In *European Conference on Computer Vision*, pages 135–153. Springer, 2016.

[20] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, 2018.

[21] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 274–282, 2018.

[22] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer, 2014.

[23] Di Wu, Si-Jia Zheng, Xiao-Ping Zhang, Chang-An Yuan, Fei Cheng, Yang Zhao, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, and De-Shuang Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 337: 354–371, 2019.

[24] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.

[25] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4733–4742, 2017.

[26] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016.

[27] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[28] Fufu Yu, Xinyang Jiang, Yifei Gong, Shizhen Zhao, Xiaowei Guo, Wei-Shi Zheng, Feng Zheng, and Xing Sun. Devil's in the details: Aligning visual clues for conditional embedding in person re-identification. *arXiv e-prints*, pages arXiv–2009, 2020.

[29] Ruimao Zhang, Jingyu Li, Hongbin Sun, Yuying Ge, Ping Luo, Xiaogang Wang, and Liang Lin. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing*, 28(10):4870–4882, 2019.

[30] Wei Zhang, Bingpeng Ma, Kan Liu, and Rui Huang. Video-based pedestrian re-identification by adaptive spatio-temporal appearance model. *IEEE Transactions on Image Processing*, 26(4):2042–2054, 2017.

[31] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019.

[32] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10407–10416, 2020.

[33] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3186–3195, 2020.

[34] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1077–1085, 2017.

[35] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3219–3228, 2017.

[36] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian-sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4913–4922, 2019.

[37] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8514–8522, 2019.

[38] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017.

[39] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017.