IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 14, 2019, accepted July 3, 2019, date of publication July 8, 2019, date of current version July 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927170

Joint Attention Mechanism for Person Re-Identification

SHANSHAN JIAO, JIABAO WANG, GUYU HU, ZHISONG PAN, LIN DU, AND JIN ZHANG
Command and Control Engineering College, PLA Army Engineering University, Nanjing 210007, China

Corresponding author: Zhisong Pan (hotpzs@hotmail.com)

ABSTRACT Although person re-identification (ReID) has drawn increasing research attention due to its potential to address the problem of analysis and processing of massive monitoring data, it is very challenging to learn discriminative information when the people in the images are occluded, in large pose variations or from different perspectives. To address this problem, we propose a novel joint attention person ReID (JA-ReID) architecture. The idea is to learn two complementary feature representations by combining a soft pixel-level attention mechanism and a hard region-level attention mechanism. The soft pixel-level attention mechanism learns a discriminative embedding for the fine-grained information by exploring the salient parts in the feature maps. The hard region-level attention mechanism conducts uniform partitions on the convolutional feature maps for learning local features. We have achieved competitive results in three popular benchmarks, including Market1501, DukeMTMC-reID, and CUHK03. The experimental results verify the adaptability of the joint attention mechanism to non-rigid deformation of the human body, which can effectively improve the accuracy of ReID.

INDEX TERMS Computer vision, attention, person re-identification, saliency.

I. INTRODUCTION

Person re-identification (ReID) aims to tell whether a person can be found in other non-overlapping surveillance camera views by matching person images [1]. The key to ReID problem is how to discover the identity discriminative information from difficult samples. Difficult samples can be divided into two categories: one is due to environmental changes, (i.e. view variations of cameras, illumination, occlusion); the other is related to pedestrian attributes, (i.e. pose changes, intra-person appearance variation, inter-person appearance similarity), as shown in Fig.1. Non-rigid deformation is a common problem in both types of difficult samples. Non-rigid deformation means that the relative distances between parts of the human body are changed. In real-world applications, non-rigid deformation severely affects the performance of person ReID.

To address the problem, the recent ReID methods focus on how to extract more discriminative feature representations. Some works cut the feature maps into strips [49] or grids [5] to extract regional features. However, these deep methods implicitly assume the accuracy of bounding boxes by simply adopting existing deep architectures with high complexity in model design and lack of interpretability. Some other works focus on prior knowledge to localize the discriminative components, such as poses or body landmarks [36], [51], [52]. However, the performance of these methods is highly dependent on the robustness of pose or landmark estimation models. Erroneous estimation of pose or landmark may greatly

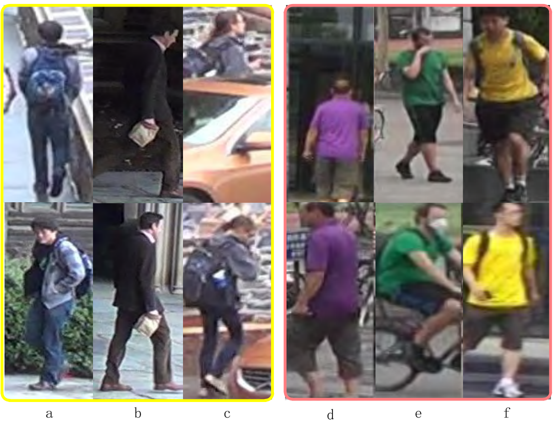

FIGURE 1. Difficult samples in CUHK03 and Market1501. The yellow box on the left shows difficult samples due to environmental changes: a. view variations of cameras, b. illumination, c. occlusion. The pink box on the right shows difficult samples related to pedestrian attributes: d. pose changes, e. intra-person appearance variation, f. inter-person appearance similarity. Non-rigid deformation exists in both categories.

The associate editor coordinating the review of this manuscript and approving it for publication was Shuhan Shen.

influence the identification result. Therefore, we want to design an end to end architecture, which not only can effectively solve the misalignment because of inaccuracy bounding boxes, but also does not need additional auxiliary labels.

We propose a joint attention person ReID architecture (JA-ReID), which combines the soft pixel-level attention and hard region-level attention. The *soft pixel-level attention* which is based on saliency can dynamically select the useful features. The soft pixel-level attention mechanism can automatically localize the most activated part in the feature maps by aggregating all the pixels cross-channels into one feature map and getting the largest connected component. The advantage of this method is that it can remove background noise and less distinctive parts of the image without learning parameters, which benefits the ReID problem. The *hard region-level attention* which divides the convolutional feature rigidly into several parts can select features according to fixed scales. We introduce the hard region-level attention to get the coarse-grained feature representations. This complementary attention mechanism can make the extracted feature representations more discriminative.

Moreover, we adopt two effective strategies to improve the accuracy of ReID. First, we reduce the dimension of each part separately after dividing the feature into parts. The experimental results show that the method performs better than reducing the dimension of the whole feature directly. Second, we adopt an improved pooling strategy. Average pooling perceives the information of the whole image including the background. Max pooling is concerned with the activated part. We concatenate both of them in the width direction. It performs better than using any pooling strategy alone in the experiments.

The main contributions of this paper are as follows: (1) We propose a soft pixel-level attention mechanism, which can get fine-grained information of the image. (2) A novel Joint Attention person re-identification architecture (JA-ReID) is proposed by combining the soft pixel-level attention and hard region-level attention, which can maximize the correlated complementary information. (3) We have achieved competitive results on three large datasets including Market1501, DukeMTMC-reID, CUHK03. The experimental results verify the adaptability of the joint attention mechanism to non-rigid deformation of the human body, which can effectively improve the accuracy of ReID.

The rest of this paper is organized as follows. Sec. II introduces the related work on person re-identification (ReID) task. The details of the proposed JA-ReID method is presented in Sec. III. In Sec. IV, we have done sufficient experiments on three datasets and further discussed the effectiveness of the proposed method. Sec. V gives a brief summary and discussion of our work.

II. RELATED WORK

Early person ReID methods are based on hand-crafted features, which have two components, image description and distance metrics. Hand-crafted features always use color [2], texture features [3] and SIFT [4]. With the great breakthrough of deep learning in the field of computer vision, person ReID problem has gradually entered the era of deep learning. In this section, we review several deep learning person ReID methods from three perspectives.

A. DEEP LEARNING GLOBAL FEATURES

Most of the early work based on deep learning use global features. These methods regard person ReID as a matching or classification problem. The matching approaches use a siamese network with image pairs or triplets [25], [46], [47], [49], [50], [53] as input. The disadvantage of these methods is that the query has to pair or triplet with each gallery image before being sent into the network, which is a time-consuming process in large datasets. The classification approaches [48] can make full use of ReID labels, unlike the siamese network. In fact, the siamese network only needs to consider pairwise (or triplet) labels. On larger datasets, such as PRW and MARS [54], [55], the classification approaches achieve good performance without careful training sample selection. However, the classification approaches assume that person images are well aligned, which rely heavily on the accuracy of bounding boxes. In fact, in many real-world scenarios, bounding boxes are not perfect.

B. DEEP LEARNING LOCAL FEATURES

To overcome the limitation of global features, many approaches focus on the local features which have more regional information. This type of approach can be divided into three categories. The first type is that the partition is cropped into pre-defined strips. Sun *et al.* [6] proposes a Part based Convolutional Baseline (PCB) to learn discriminative partition features. Yi *et al.* [49] divides an image into three strips, each of which makes an SCNN, and concatenates part features. The second type is dividing the image into grid patches. Ahmed *et al.* [50]not only matches the grids in the same position but also in the neighborhood grid of another image. The third type is that the methods based on some prior knowledge such as human pose estimation or landmarks [51], [52]. Su *et al.* [36] embeds pose information into the architecture and generates a modified image. However, the performance of these methods is highly dependent on the robustness of pose or landmark estimation models. Whether the image is cropped based on patches or human posture key points, because each part is rigid segmentation, they may suffer some outliers, which make the inconsistency in each partition.

C. DEEP MULTI-VIEW LEARNING

Some researchers consider person ReID as a cross-view classification problem, that recognizes persons from different cameras. Learning a common feature space from multi-view spaces becomes an effective approach to solve the problem. Cao *et al.* [58] proposes a unified solution for subspace learning methods using the Rayleigh quotient, which is extensible for multi-views. You *et al.* [59] presents a

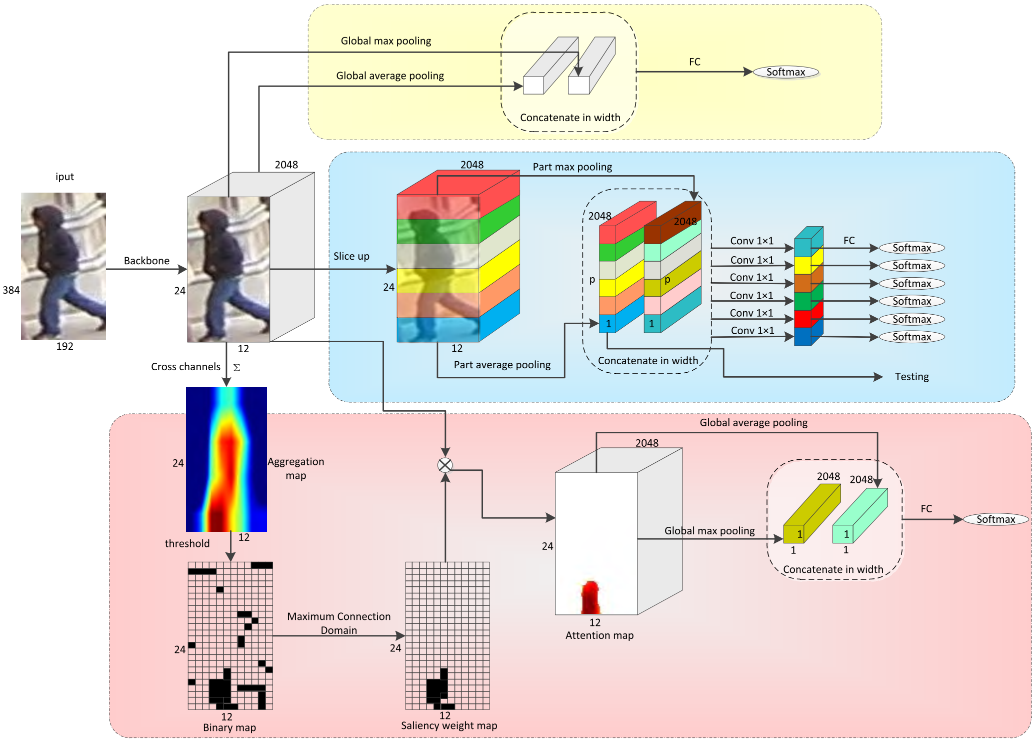

FIGURE 2. An overview of our architecture JA-ReID. There are three branches from top to bottom. Top: Global feature branch. Middle: Hard region-level attention branch (Sec.III.C). Bottom: Soft pixel-level attention branch (Sec.III.B).

MvCCDA algorithm for cross-view classification to handle view discrepancy, discriminability and nonlinearity in a joint manner. Zhu *et al.* [60] proposes a multi-view PHDL approach, which aims to learn a type-specific discriminative representation space from the visual appearance features of images and this type of video features.

III. METHODOLOGY

A. JA-ReID OVERVIEW

Our target is to train a deep feature representation architecture, which can reduce the interference caused by non-rigid deformation. The proposed JA-ReID architecture can learn the feature representation from three scales, the global scale, the regional scale (hard region-level attention) and fine-grained scale (soft pixel-level attention) as shown in Fig. 2. The hard region-level attention may be disturbed by some useless information (i.e. background noise) in each partition, which makes each regional features not completely indicate the feature representation of the human body. The soft pixel-level attention can locate the most discriminative part of the feature maps, which can reduce the impact of hard region-level attention errors. The complementary attention mechanism is designed to identify the best visual patterns for ReID problem, which simulates the dorsal and ventral attention mechanism of human brain [9]. In addition, we adopt two effective strategies to improve the accuracy of Rank1 rate and mAP. First, after dividing the feature into p parts, we reduce the dimensions of each part separately. Second, we concatenate the average pooling and max pooling features in the width direction for training.

Apart from the pre-trained classification model on ImageNet [15], our attention mechanism does not require additional pre-trained or labels. It can take various models as the backbone, such as VGG [41], Resnet [8] and Google Inception [42] . Because of its good performance in classification, we choose ResNet50 [8] as the CNN backbone. The structure after the original global average pooling (GAP) layer is removed (included GAP layer).

Global feature branch consists of a global average pooling (GAP) layer and a global max pooling (GMP) layer. We concatenate the two feature maps in the width direction. Then the concatenated feature maps are put into the fully-connected (FC) layer and classified by Softmax loss. Some methods think that without considering global features, only focusing on local features can achieve good results such as PCB [6]. But we think that global features contain some useful information. Models can learn the relationship between the background and the person under different camera views, for example, human contour information. Local feature methods segment the background and lose the complete information of the image.

Hard region-level attention branch segments the feature maps horizontally into p parts. Each part feature maps are put into a part average pooling layer and a part max pooling layer respectively. After pooling, the two part feature maps are concatenated in the width direction, put into the fully-connected (FC) layer and classified by Softmax loss.

Soft pixel-level attention branch is more concerned with salient information, focusing on the most discriminative part. In some cases where the bounding boxes detected are inaccurate, the regional features acquired from the hard region-level attention branch do not match the corresponding part. The soft pixel-level attention mechanism can effectively lessen the impact of misalignment. We get a saliency weight map from soft pixel-level attention branch. With a tensor multiplication, it can select the most representative features.

Remarks PCB [6] is the closest competitor, which also leverages partial-based learning for person Re-ID. However, it has three major drawbacks. 1)The performance of PCB depends on the precise bounding boxes otherwise the pre-fined partition cannot be aligned very well. In practical application scenarios, the current detection models cannot be sufficient to do that. 2) It benefits from a post-processing approach called RPP, which makes the optimized model cannot be trained in an end-to-end manner. 3) Global information is an important clue to identification and recognition, which is completely ignored in PCB. Global features are often robust to subtle view changes and internal changes. Our method JA-ReID has improved on all three shortcomings.

B. SOFT PIXEL-LEVEL ATTENTION LEARNING

After *resnet_5c* in ResNet50, the input image I can be represented by an 3-dimensional tensor T, which is a sparse and distributed representation [7], [10]. The task of soft pixel-level attention is to produce a saliency weight map W, which is the same size as T. With a tensor multiplication from W and T, it can remove unimportant information such as backgrounds and select the most representative features.

We first sum the pixels of T through the channel direction to get a 2-dimensional feature representation called the aggregation map S. The $h \times w \times c$ tensor T becomes an $h \times w$ tensor S, formulated as

$$S = \sum_{n=1}^{c} f_n \tag{1}$$

where h, w and c denote the number of pixels in height, width and channel dimensions respectively. f_n denotes the $c - th$ feature map in T. We have two considerations for this summation. First, from the perspective of cross-channels feature fusion, the activations of different channels are sparse for the same component in an image, and the useful information should be activated on most channels. Therefore we sum all the pixels in the same spatial position across channels, displaying the most activated positions which contain the most important recognition information. Second, in terms of spatial distribution of features, a single channel contains at most some weak semantic information, most of which is noise, and cross-channel feature fusion can highlight the part that actually contains semantic information.

The higher the activation response of a specific location (i, j), the greater the possibility that the corresponding region becomes a main part of the person. In order to distinguish the more activated part from S, we calculate the mean value $\overline{s}$ of all the positions in S as the threshold, where the activation response is higher than $\overline{s}$ indicates the more discriminative component of the image, formulated as

$$M_{i,j} = \begin{cases} 1, & if\ S_{i,j} > \overline{s} \\ 0, & otherwise. \end{cases} \tag{2}$$

Then we get a binary mask M with the same size as S. Comparing the binary mask M with the original image I, we find that in addition to the central parts of M being labeled 1, at the edge of M some positions are also labeled 1. That means there are still some small background noise parts on the mask M, which are mostly distributed at the edge of the image, smaller than the person's main component. Therefore, in order to further narrow down the effective attention range, we choose the largest connected component, removing the noise parts, and get the final useful feature positions, which is the saliency weight map W. The process is shown in Algorithm 1, which aims to get the largest connected component in the binary mask.

Algorithm 1 The Largest Connected Component Based on Flood-Fill Algorithm

```
Input: a binary mask $M$
Output: a saliency weight map $W$
while there is an unlabelled pixel do
    $L = 1$;
    select an unlabelled pixel and label it $L$;
    number of $L$ is $L_{num} = 1$;
    if there is an unlabelled pixel connected with $L$ then
        Label it $L$;
        $L_{num} = L_{num} + 1$;
    end if
    $L = L + 1$;
    search for the next unlabelled pixel and label it $L$;
end while
select the largest $L_{num}$ and label its components 1;
label other components 0;
return a saliency weight map $W$
```

We can compute efficiently the attention map A from feature maps T and a saliency weight map W with a tensor multiplication, formulated as

$$A = W \times T \tag{3}$$

which means the most distinguish salient information is reserved, all the other parts are 0.

C. HARD REGION-LEVEL ATTENTION LEARNING

The soft pixel-level attention focuses on fine-grained features and the hard region-level attention pays more attention to the regional features of the human body. According to the distribution of the human body, we divide the tensor T into p parts horizontally. For the selection of hyper-parameter p, we draw on the method of dividing pedestrians of PCB [6]

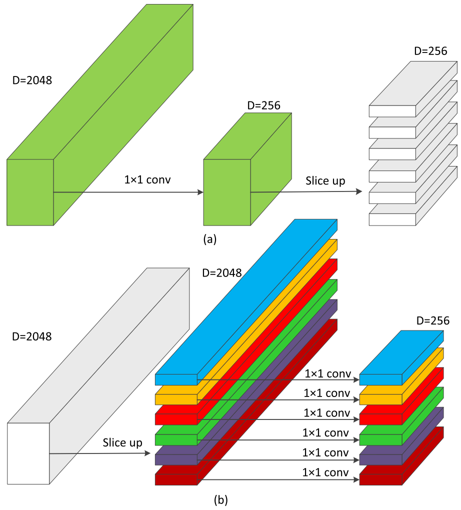

FIGURE 3. **Two methods of reducing the dimensions. (a) The whole feature maps are reduced dimensions by a 1 × 1 convolution layer. (b) The feature maps are first segmented into different parts, and then each part uses an independent parameter 1 × 1 convolution layer to reduce the dimensions.**

$p = 6$, which has been demonstrated the best. Then we downsample each piece of tensor by part average pooling and part max pooling. We concatenate the two features in the width direction and use p 1×1 convolution layer to reduce the dimension in order to reduce the computation. After that, each part feature maps are put into a classifier, which consists of a fully-connected (FC) layer and the following Softmax loss.

Remarks: For how to segment and reduce the dimension of T, we consider two methods. The first method has fewer parameters. The dimension of T is reduced through a 1×1 convolution layer, and then it is divided into p parts, as shown in Fig. 3.(a). The other method is to divide T into p parts firstly and then reduce the dimension of each component by p independent parameter 1×1 convolution layer, as shown in Fig. 3.(b). Our architecture uses the second set, which does not share the parameter of 1×1 convolution layer. We will provide extreme ablation experiments in the following section (Sec.IV.D.2) to verify the effectiveness of our settings.

D. POOLING STRATEGY

The average pooling is widely used in many classification problems because it forces a corresponding relationship between the feature representations and the category. This average operation gives all parts of the image the same weight and equal treatment. But when the background is similar to the pedestrian, average pooling may cause a low response to the component of the pedestrian and ignore it. On the other hand, the max pooling extracts the most discriminative information but lacks global relevance to the whole image.

These two pooling strategies have their own advantages. In order to combine these two complementary features to get a more representative feature, we concatenate the two features obtained by average pooling and max pooling in the width direction for training. This concatenated feature is put into the FC layer and computed Softmax loss.

TABLE 1. **Details of datasets.**

dataset	camers	train ID	test ID	images
Market1501	6	751	750	32668
DukeMTMC-reID	8	702	702	36411
CUHK03	2	767	700	14097

IV. EXPERIMENTS

A. DATASETS AND PROTOCOL

We evaluate the proposed JA-ReID architecture on three large benchmarks Market1501 [18], DukeMTMC-reID [11], CUHK03 [12]. Table 1 shows the structure of the three datasets. Market1501 contains 2793 distractors. DukeMTMC-reID has 8 cameras, whose backgrounds are more complicated than Market1501. There are also distractors in DukeMTMC-reID, so it is the most challenging dataset at present. CUHK03 offers two types of annotations: human labeled and detected by DPM [43]. Our experiments are based on the detected label images.

We use the cumulative matching characteristic (CMC) and mean Average Precision (mAP) metrics to measure the performance. We set all the experiments with the single query evaluation. Moreover, in order to verify the effectiveness of the approach, we do not use data augmentation, reranking [13] or random erasing [16], which considerably improve mAP.

B. IMPLEMENTATION DETAILS

We implement our JA-ReID in the Pytorch [14] framework with a NVIDIA 1080*ti* GPU. The backbone ResNet50 [8] is pre-trained on ImageNet [15]. The training images are resized to 384 × 192. We set batch size to 32 and train the model for 50 epochs with base learning rate initialized at 0.0003. We use AMSGrad [17] with two moment terms $\beta_1 = 0.9$, $\beta_2 = 0.999$. All datasets share the same experiment setting as above.

In the testing stage, we concatenate 6 regional features after part average pooling as the final feature maps.

C. COMPARISON WITH THE STATE OF THE ART METHODS

We compare JA-ReID with the state of the art on three datasets. **Note**, we do not use any data argumentation methods, such as scaling, rotation, color distortion, neither model pre-trained. Most deep learning ReID methods benefit greatly from these operations, which require not only heavier computation but also time-consuming. As a closest competitor, which also leverages partial-based learning for person Re-ID, we reproduce the PCB [6] method in order to better evaluate the performance of the two approaches under the same experimental environment. Maybe there are some differences between our implementation and the PCB paper's, which are not detailed. The reproduced results are different from those on the paper, especially on CUHK03 dataset. In fact, if our reproduced results of PCB are better, our method can also make a further performance improvement. In this paper, we compare the performance with the reproduced results

TABLE 2. Comparison of JA-ReID with the state-of-the-art on Market1501. PCB* refers to the reproduced experimental results of PCB. Other results are obtained from the papers.

Methods	Rank1	mAP
XQDA [27]	43.8	22.2
BoW+kissme [18]	44.4	20.8
WARCA [19]	45.2	-
KLFDA [20]	46.5	-
SCS [28]	51.9	26.3
DNS [29]	61.0	35.6
CRAFT [30]	68.7	42.3
CAN [31]	60.3	35.9
G-SCNN [32]	65.8	39.5
SOMAnet [21]	73.9	47.9
SVDNet [22]	82.3	62.1
PAN [23]	82.8	63.4
Transfer [24]	83.7	65.5
Triplet Loss [25]	84.9	69.1
DML [26]	87.7	68.8
MultiRegion [37]	66.4	41.2
HydarPlus [38]	76.9	-
MSCAN [39]	80.3	57.5
PAR [40]	81.0	63.4
PDC [36]	84.1	63.4
JLML [35]	85.1	65.5
PartLoss [61]	88.2	69.3
MultiScale [44]	88.9	73.1
HA-CNN [45]	**91.2**	75.7
PCB*	89.9	75.5
Camera style adaptation [56]	88.12	68.72
MLFN [57]	90.0	74.3
JA-ReID	90.4	**76.1**

TABLE 3. Comparison of JA-ReID with the state-of-the-art on DukeMTMC-reID. PCB* refers to the reproduced experimental results of PCB. Other results are obtained from the papers.

Methods	Rank1	mAP
BoW+kissme [18]	25.1	12.2
LOMO+XQDA [27]	30.8	17.0
SVDNet-CaffeNet [22]	67.6	45.8
PAN [23]	71.6	51.5
JLML [35]	73.3	56.4
TriNet+Era [16]	73.0	56.6
SVDNet-ResNet50 [22]	76.7	56.8
Pose-transfer [33]	78.5	56.9
AOS [34]	79.2	62.1
HA-CNN [45]	80.5	63.8
PCB*	79.5	65.3
Camera style adaptation [56]	75.27	53.48
MLFN [57]	**81.0**	62.8
JA-ReID	80.9	**65.7**

TABLE 4. Comparison of JA-ReID with the state-of-the-art on CUHK03. PCB* refers to the reproduced experimental results of PCB. Other results are obtained from the papers.

Methods	Rank1	mAP
BoW+kissme [18]	6.4	6.4
LOMO+XQDA [27]	12.8	11.5
SVDNet-CaffeNet [22]	27.7	24.9
PAN [23]	36.3	34.0
SVDNet-ResNet50 [22]	41.5	37.3
Pose-transfer [33]	41.6	38.7
HA-CNN [45]	41.7	38.6
AOS [34]	47.1	43.3
TriNet+Era [16]	55.5	50.7
PCB*	56.1	53.4
MLFN [57]	52.8	47.8
JA-ReID	**58.0**	**56.5**

of PCB. All the other results of methods are obtained from the published papers.

Comparisons on Market1501 are detailed in Table 2. The comparison methods are divided into three categories, namely, hand-crafted methods, deep learning methods with global feature and deep learning methods with local features. The method we proposed, JA-ReID, has a greater improvement on Rank1 and mAP than the previous methods including PDC [36] which needs auxiliary component labels to deliberately align components. Compared with HA-CNN [45], which does not use any augmentation, our method has a slight gap in Rank1 (−0.8%), but it has an improvement (+0.4%) in mAP. Compared with the camera style adaptation [56] and MLFN [57] of 2018, JA-ReID achieves higher Rank1 accuracy and mAP. Compared with PCB, our method has an improvement in Rank1 (+0.5%) and mAP (+0.3%).

Person images from DukeMTMC-reID have more variations in illumination and backgrounds because of wider camera views and more complex scene layout compared with Market1501. Our JA-ReID approach is again superior to the most recent methods, as shown in Table 3. Although our method performs a little worse than MLFN [57] in Rank1 (-0.1%), it has a great improvement in mAP (+2.9%). JA-ReID achieves higher Rank1 accuracy and mAP, compared with the camera style adaptation [56]. Compared with PCB, our method performs better (+1.4%) in Rank1 and (+0.4%)mAP.

Our model JA-ReID achieves a large margin (+2.5% in Rank1 and +5.8% in mAP) over TriNet+Era [16] on the detected set of CUHK03 which is enhanced by extra data augmentation as shown in Table 4. Compared with HA-CNN, JA-ReID not only achieves a great improvement on Rank1 (+16.3%) but also makes a bigger improvement on the mAP (+17.9%). JA-ReID achieves higher Rank1 (+5.2%) and mAP (+8.7%) than MLFN [57]. Compared with PCB, JA-ReID performs better in both Rank1 (+1.9%) and mAP (+3.1%). Especially, JA-ReID gets a higher mAP (+2.3%) than the mAP of PCB (+54.2%) in the paper.

We visualize some results of Market1501 in Fig. 4. Whether the people in the images are occluded, in large pose variations or from different perspectives, our method can extract discriminative feature representations. The experimental results verify the adaptability of the proposed JA-ReID architecture to non-rigid deformation of the human body, which can effectively improve the accuracy of ReID.

D. ABLATION STUDY

1) EFFECT OF SOFT PIXEL-LEVEL ATTENTION

We set experiments about the effect of soft pixel-level attention on Market1501, DukeMTMC-reID and CUHK03 with the same training set. First, we only adopt the global feature branch for training and get the global features for testing. Moreover, we train the model combined global feature

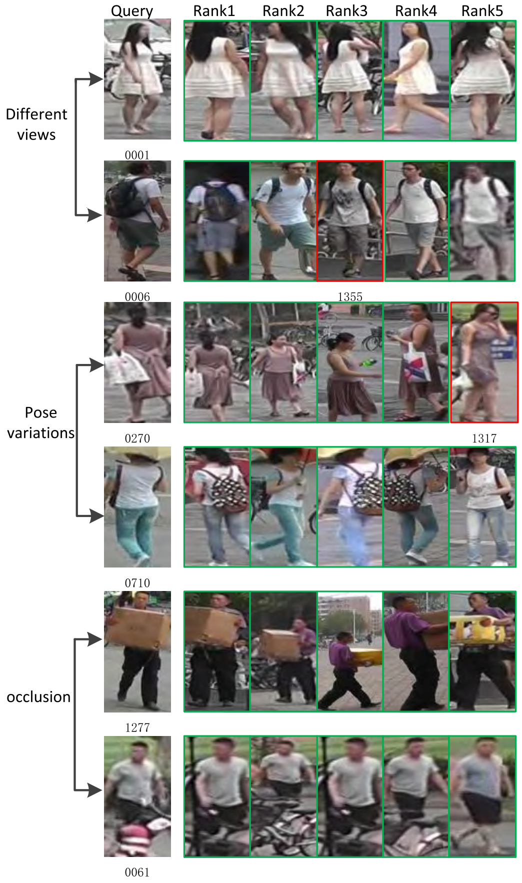

FIGURE 4. Visualization of some results of Market1501. We choose three different types of samples: Differential views, pose variations and occlusion. Green bounding boxes mean the correct matching. Red bounding boxes mean the wrong matching.

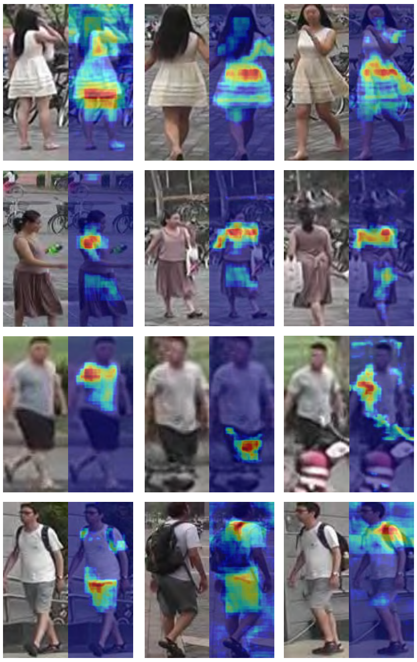

FIGURE 5. Heatmaps of features got from soft pixel-level attention branch.

TABLE 5. Comparison of different dimension reduction strategies in hard region-level attention branch.

Method	Market1501		DukeMTMC-reID		CUHK03	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
share 1 × 1conv	89.1	73.6	79.8	64.9	50.1	48.4
independent 1 × 1 conv	**90.4**	**76.1**	**80.9**	**65.7**	**58**	**56.5**

branch with soft pixel-level attention branch, and get the global features for testing. In addition, we train the model combined global feature branch with hard region-level attention branch, and get the regional feature maps concatenated for testing. It can be seen from Fig. 6 that the with the soft pixel-level attention branch, the global feature branch achieves an increase in Rank1 (+3%) and mAP (+5.5%) on Market1501, in Rank1 (+1.3%) and mAP (+4.2%) on DukeMTMC-reID, in Rank1 (+3.8%) and mAP (+5.5%) on CUHK03. With the soft pixel-level attention branch, JA-ReID achieves an improvement in Rank1 (+0.4%) and mAP (+2.4%) on Market1501, in Rank1 (+1.1%) and mAP (+0.9%) on DukeMTMC-reID, in Rank1 (+0.9%) and mAP (+2.6%) on CUHK03 than global feature branch just combined with hard region-level attention branch. In order to better illustrate the role of soft pixel-level attention branch, we visualize the feature maps got from soft pixel-level attention branch as shown in Fig. 5. For example, the third person is occluded in the third image. The feature maps focus on the unoccluded upper body of the pedestrian. When the people in the images are in large pose variations or from different perspectives (such as the first, second, fourth person), our soft pixel-level attention can locate the most discriminative part.

2) COMPARISON OF TWO DIMENSION REDUCTION METHODS

In hard region-level attention branch, we use 1×1 convolution layer to reduce the dimension, which aims to reduce computational complexity. We first use one 1 × 1 convolution layer for the whole feature maps, which has the advantage of fewer parameters. In addition, we try the independent parameter 1 × 1 convolution layer for each part. We set experiments on Market1501, DukeMTMC-reID and CUHK03. As shown in Table 5, setting an independent parameter 1 × 1 convolution layer for each part can make an improvement in Rank1 (+1.3%) and mAP (+2.5%) on Market1501, in Rank1 (+1.1%) and mAP (+0.8%) on DukeMTMC-reID, in Rank1 (+7.9%) and mAP (+8.1%) on CUHK03. This shows that it is better to reduce the dimension of each component separately with an independent parameter 1 × 1 convolution layer.

3) EFFECT OF DIFFERENT POOLING STRATEGIES

Before putting the features into classifiers, we consider three pooling strategies, using GAP (global average pooling), GMP (global max pooling) and concatenated features for training. As can be seen from Table 6, the Rank1 and mAP of max pooling strategy are a bit higher than the ones of average pooling strategy. That is because max pooling strategy focuses

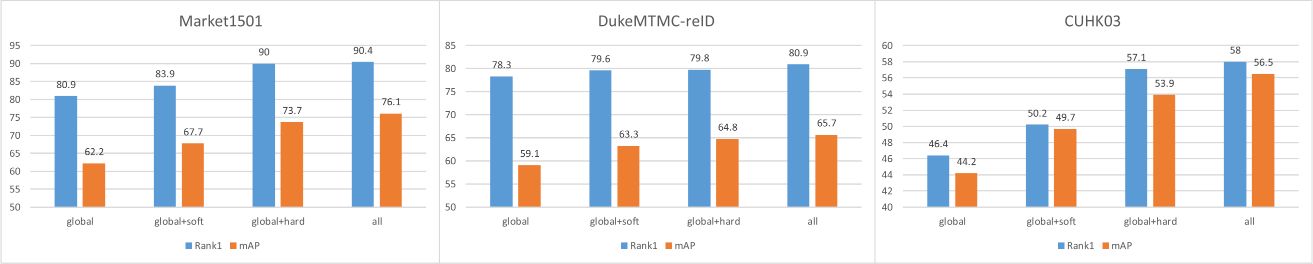

FIGURE 6. Results of different combined branches on Market1501,DukeMTMC-reID and CUHK03. We show that soft pixel-level attention branch can improve the performance over the global branch. In addition, soft pixel-level attention branch yields consistent improvement over the combination of global branch and hard region-level attention branch.

TABLE 6. Results of different pooling strategies in JA-ReID.

Method	Market1501		DukeMTMC-reID		CUHK03	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
GAP	88.7	72.5	78.8	63.2	52.4	50.3
GMP	89.0	73.9	80.4	65.1	57.5	55.9
GAP& GMP	**90.4**	**76.1**	**80.9**	**65.7**	**58**	**56.5**

more on discriminative parts of the feature maps. Average pooling strategy gives the same weight to all parts of the feature maps, which is easily influenced by the background. Each of these two strategies has unique advantages. We concatenate the two features obtained by average pooling and max pooling in the width direction for training. Table 6 shows that this strategy achieves the best results.

V. CONCLUSION

In this work, we propose a novel Joint-Attention ReID (JA-ReID) approach, which can effectively solve the current challenge of non-rigid deformation of the human body in ReID task. Unlike most methods just using one constraint, JA-ReID combines two types of attention: soft pixel-level attention and hard region-level attention. The advantage of the joint attention mechanism is that it can focus on feature representations of fine-grained and coarse-grained scale, which can improve the robustness of the model and make the features extracted more discriminative. The proposed soft pixel-level attention mechanism can locate the most discriminative part of the feature maps without additional auxiliary labels and training. Moreover, we adopt two effective strategies to improve the accuracy of ReID. First, we reduce the dimension of each part separately after dividing the feature into parts. Second, we adopt an improved pooling strategy. The experimental results show that our approach JA-ReID has achieved very competitive results on three datasets. Ablation study well demonstrates the effectiveness of our JA-ReID approach. The experimental results verify the adaptability of the joint attention mechanism to non-rigid deformation of the human body, which can effectively improve the accuracy of ReID.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: https://arxiv.org/abs/1610.02984

[2] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 262–275.

[3] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2666–2672.

[4] Z. Rui, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3586–3593.

[5] Y. Zhang, X. Li, L. Zhao, and Z. Zhang, "Semantics-aware deep correspondence structure learning for robust person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3545–3551.

[6] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2017, pp. 480–496.

[7] G. E. Hinton, "Learning distributed representations of concepts," in *Proc. 8th Conf. Cogn. Sci. Soc.*, 1989, pp. 1–12.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 770–778.

[9] S. Vossel, J. J. Geng, and G. R. Fink, "Dorsal and ventral attention systems: Distinct neural circuits but collaborative roles," *Neuroscientist*, vol. 20, no. 2, pp. 150–159, 2013.

[10] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[11] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3754–3762.

[12] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. CVPR*, Jun. 2014, pp. 152–159.

[13] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. CVPR*, Jul. 2017, pp. 1318–1327.

[14] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in Pytorch," in *Proc. Neural Inf. Process. Syst.*, 2017, pp. 1–4.

[15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[16] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: https://arxiv.org/abs/1708.04896

[17] J. Sashank and R. S. K. Kumar, "On the convergence of adam and beyond," in *Proc. ICLR*, 2018, pp. 1–23.

[18] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, Dec. 2015, pp. 1116–1124.

[19] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *Proc. ECCV*, 2016, pp. 875–890.

[20] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets," 2016, *arXiv:1605.09653*. [Online]. Available: https://arxiv.org/abs/1605.09653

[21] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis, "Looking beyond appearances: Synthetic training data for deep CNNs in re-identification," 2017, *arXiv:1701.03153*. [Online]. Available: https://arxiv.org/abs/1701.03153

[22] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. ICCV*, Oct. 2017, pp. 3800–3808.

[23] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," 2017, *arXiv:1707.00408*. [Online]. Available: https://arxiv.org/abs/1707.00408

[24] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," 2016, *arXiv:1611.05244*. [Online]. Available: https://arxiv.org/abs/1611.05244

[25] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: https://arxiv.org/abs/1703.07737

[26] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," 2017, *arXiv:1705.00384*. [Online]. Available: https://arxiv.org/abs/1705.00384

[27] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. CVPR*, Jun. 2015, pp. 2197–2206.

[28] D. Chen, Z. Yuan, B. Chen, and N. Zheng, "Similarity learning with spatial constraints for person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1268–1277.

[29] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. CVPR*, Jun. 2016, pp. 1239–1248.

[30] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 392–408, Feb. 2017.

[31] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, May 2017.

[32] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. ECCV*, 2016, pp. 135–153.

[33] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 4099–4108.

[34] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. CVPR*, Jun. 2018, pp. 5098–5107.

[35] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," in *Proc. IJCAI*, 2017, pp. 1–10.

[36] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. ICCV*, Oct. 2017, pp. 3960–3969.

[37] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," 2015, *arXiv:1512.05300*. [Online]. Available: https://arxiv.org/abs/1512.05300

[38] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang, "Hydraplus-net: Attentive deep features for pedestrian analysis," in *Proc. ICCV*, Oct. 2017, pp. 350–359.

[39] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. CVPR*, Jul. 2017, pp. 384–393.

[40] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. ICCV*, Oct. 2017, pp. 3219–3228.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/arXiv:1409.1556

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE CVPR*, Jun. 2016, pp. 2818–2826.

[43] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. CVPR*, Jun. 2008, pp. 1–8.

[44] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. ICCVW*, Oct. 2017, pp. 2590–2600.

[45] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," 2018, *arXiv:1802.08122*. [Online]. Available: https://arxiv.org/abs/1802.08122

[46] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," 2016, *arXiv:1604.02426*. [Online]. Available: https://arxiv.org/abs/1604.02426

[47] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[48] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1249–1258.

[49] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *Proc. 22nd Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 34–39.

[50] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

[51] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," 2017, *arXiv:1701.07732*. [Online]. Available: https://arxiv.org/abs/1701.07732

[52] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. ACM Multimedia*, 2017, pp. 420–428.

[53] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognit.*, vol. 48, no. 10, pp. 2993–3003, 2015.

[54] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "MARS: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.

[55] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," 2016, *arXiv:1604.02531*. [Online]. Available: https://arxiv.org/abs/1604.02531

[56] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 5157–5166.

[57] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2109–2118.

[58] G. Cao, A. Iosifidis, K. Chen, and M. Gabbouj, "Generalized multi-view embedding for visual recognition and cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 48, no. 9, pp. 2542–2555, Sep. 2018.

[59] X. You, J. Xu, W. Yuan, X.-Y. Jing, D. Tao, and T. Zhang, "Multi-view common component discriminant analysis for cross-view classification," *Pattern Recognit.*, vol. 92, pp. 37–51, Aug. 2019.

[60] X. Zhu, X.-Y. Jing, X. You, W. Zuo, S. Shan, and W.-S. Zheng, "Image to video person re-identification by learning heterogeneous dictionary pair with feature projection matrix," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 3, pp. 717–732, Mar. 2018.

[61] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "Deep representation learning with part loss for person re-identification," 2017, *arXiv:1707.00798*. [Online]. Available: https://arxiv.org/abs/1707.00798

SHANSHAN JIAO received the B.S. and M.S. degrees in safety engineering from the University of Science and Technology Beijing, Beijing, in 2011 and 2013, respectively. She is currently pursuing the Ph.D. degree in computer science and technology with PLA Army Engineering University, Nanjing, China. Her research interests include computer vision and machine learning.

JIABAO WANG received the Ph.D. degree in computational intelligence from the PLA University of Science and Technology, Nanjing, China, in 2013. He is currently an Assistant Professor with PLA Army Engineering University, Nanjing. His current research interests include computer vision and machine learning.

GUYU HU received the B.S. degree in radio communication from Zhejiang University, Hangzhou, China, in 1983, and the M.S. degree in computer application technology and the Ph.D. degree in communications and information systems from the Nanjing Institute of Communications, Nanjing, China, in 1989 and 1992, respectively. Since 1990, he has been involved in the research on network management. Since 1997, he has been a Full Professor with the PLA Army Engineering University, China. Since 1998, his research interests include intelligent of network management, mainly on failure-finding from data with pattern recognition, machine learning, and neural networks.

ZHISONG PAN received the Ph.D. degree in computational intelligence from the PLA University of Science and Technology, Nanjing, China, in 2013. He is currently a Professor with PLA Army Engineering University, Nanjing. His current research interests include computer vision and machine learning.

LIN DU received the B.S. degree in computer science and technology from Northeastern University, Shenyang, China, in 2014, and the M.S. degree in communication and information system from the PLA University of Science and Technology, Nanjing, China, in 2017. He is currently pursuing the Ph.D. degree with PLA Army Engineering University, Nanjing. His research interests include computer vision and machine learning.

JIN ZHANG received the B.S. and M.S. degrees in nuclear power engineering from the PLA Naval University of Engineering, Wuhan, in 2007 and 2009, respectively. He is currently pursuing the Ph.D. degree in computer science and technology with PLA Army Engineering University, Nanjing, China. His research interests include computer vision and machine learning.