IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 6, 2020, accepted March 18, 2020, date of publication March 20, 2020, date of current version March 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2982032

Adversarial Erasing Attention for Person Re-Identification in Camera Networks Under Complex Environments

SHUANG LIU[1,2], (Senior Member, IEEE), XIAOLONG HAO[1,2], RONGHUA ZHANG[3], ZHONG ZHANG[1,2], (Senior Member, IEEE), AND TARIQ S. DURRANI[4]

[1]Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China
[2]College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin 300387, China
[3]College of Information Science and Technology, Shihezi University, Shihezi 832003, China
[4]Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XQ, U.K.

Corresponding author: Zhong Zhang (zhong.zhang8848@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61711530240, in part by the Natural Science Foundation of Tianjin under Grant 19JCZDJC31500, in part by the Open Projects Program of National Laboratory of Pattern Recognition under Grant 202000002, and in part by the Tianjin Higher Education Creative Team Funds Program.

ABSTRACT Person re-identification (Re-ID) in camera networks under complex environments has achieved promising performance using deep feature representations. However, most approaches usually ignore to learn features from non-salient parts of pedestrian, which results in an incomplete pedestrian representation. In this paper, we propose a novel person Re-ID method named Adversarial Erasing Attention (AEA) to mine discriminative completed features using an adversarial way. Specifically, the proposed AEA consists of the basic network and the complementary network. On the one hand, original pedestrian images are used to train the basic network in order to extract global and local deep features. On the other hand, to learn features complementary to the basic network, we propose the adversarial erasing operation, that locates non-salient areas with the help of attention map, to generate erased pedestrian images. Then, we utilize them to train the complementary network and adopt the dynamic strategy to match the dynamic status of AEA in the learning process. Hence, the diversity of training samples is enriched and the complementary network could discover new clues when learning deep features. Finally, we combine the features learned from the basic and complementary networks to represent the pedestrian image. Experiments on three databases (Market1501, CUHK03 and DukeMTMC-reID) demonstrate the proposed AEA achieves great performances.

INDEX TERMS Person re-identification, dynamic strategy, adversarial learning.

I. INTRODUCTION

Person re-identification (Re-ID) in camera networks targets to retrieve a specific pedestrian from a large gallery which is captured by multiple visual sensors [1]–[3]. There are many applications (such as multi-camera tracking [4], [5], crowd counting [6], [7] and so on [8], [9]) that require an accurate person Re-ID algorithm. However, it is a challenging task under complex environments because of significant changes in body poses, viewpoints, illuminations, etc.

Recently, many researchers [10]–[14] employ Convolutional Neural Networks (CNNs) to extract pedestrian representations Afterwards, the distance metric is applied to compute the similarity. There are mainly three kinds of representations, i.e., global deep features, local deep features and combination of them. The global deep features focus on salient areas of the whole pedestrian image. As shown in Fig. 1(a), we visualize the convolutional activation map where only some parts of pedestrian are mainly concerned.

Some researchers exploit local deep features to learn the structural information of pedestrian. Specifically, they directly divide pedestrian images or convolutional activation maps into several sub-regions and extract local deep features from sub-regions [15], [16]. As shown in Fig. 1(b), we visualize the convolutional activation map for local deep features, where the network pays attention to different

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

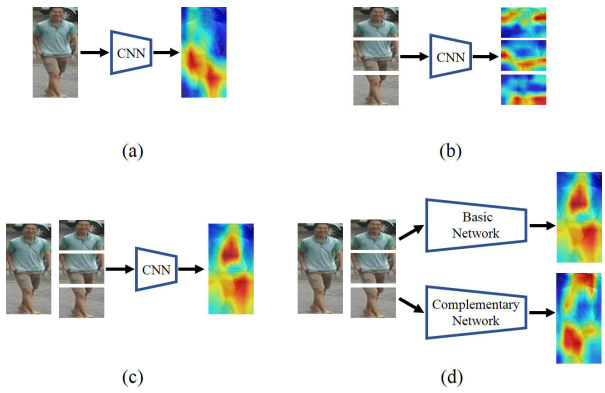

FIGURE 1. Visualization of convolutional activation maps for (a) global deep features, (b) local deep features, (c) global and local deep features, and (d) the proposed AEA. The warmer color denotes higher value, that is, attention regions of CNN.

parts from global deep features. Other partition strategies are proposed to apply external operations, such as pose estimation [17]–[19].

Furthermore, some methods [20]–[22] learn global and local deep features simultaneously to represent pedestrians, which could make full use of their own advantages. As shown in Fig. 1(c), we visualize the convolutional activation map where more parts of pedestrian are concerned by the network. However, some non-salient parts or rare pedestrian information may be easily ignored, which results in an incomplete pedestrian representation.

In this paper, we propose a novel method named Adversarial Erasing Attention (AEA) to learn discriminative completed features using an adversarial manner. For this purpose, we design a deep network including two subnetworks, i.e., basic network and complementary network. The basic network learns global and local deep features for pedestrian images simultaneously. Meanwhile, the complementary network aims at learning complementary information to the basic network. In other words, it is expected to extract features from non-salient areas of pedestrian where the basic network does not focus on. Since attention maps could capture the attention location of deep network, we propose the adversarial erasing operation to locate the non-salient areas using attention maps. Concretely, we erase salient areas based on the attention map of basic network, and therefore the remaining areas are non-salient. Afterwards, we map these erased areas to original pedestrian images, and obtain erased pedestrian images which are utilized to train the complementary network. With such adversarial learning, the complementary network can discover new clues from non-salient areas of pedestrian.

To match the dynamic status of AEA, we propose a dynamic strategy which conducts adversarial erasing operation at each iteration. As a result, it could increase the diversity of training samples with these erased pedestrian images. Finally, we integrate the features learned from the basic and complementary networks to represent pedestrian images. Fig. 1(d) visualizes the attention maps of the proposed AEA where the two networks focus on different parts of pedestrian. Hence, the features extracted from the two subnetworks could obtain discriminative completed representations for pedestrian images. It is worth mentioning that the proposed AEA does not resort to any external operations when learning deep features. In the learning process, we assign the identity label of original image to the erased pedestrian image. To make full use of erased pedestrian images, we utilize the cross-entropy loss and the triplet loss simultaneously.

In a word, the proposed AEA has the following contributions: 1) the complementary network locates non-salient areas of pedestrian in a pixel-wise manner without any external operations, and learns complementary features by an adversarial way; 2) we dynamically erase salient areas of pedestrian at each iteration so as to match the dynamic status of AEA and enrich the diversity of training samples; 3) the proposed AEA surpasses other approaches on three public person Re-ID databases.

II. RELATED WORK

A. PERSON RE-IDENTIFICATION

Recently, person Re-ID approaches mainly utilize CNNs to learn deep features [23]–[27]. Some methods learn global deep features to represent pedestrian images. Qian *et al.* [24] extract the global features of pedestrian images from multiple levels, and integrate them to describe pedestrian images using a weighted strategy. However, global deep features ignore subtle differences between pedestrian images, which results in incomplete pedestrian representation. Hence, some methods are proposed to directly divide pedestrian images or convolutional activation maps into several parts for extracting local deep features. Sun *et al.* [16] uniformly partition convolutional activation maps to extract local deep features. Wang *et al.* [26] obtain deep local features by dividing the convolutional activation map into different strips. Yao *et al.* [14] propose the Part Loss Network (PL-Net) which could minimize the representation learning risk by automatically detecting human body parts and computing the person classification loss.

Furthermore, some approaches resort to external operations, such as pose estimation, to learn local deep features. Zheng *et al.* [28] first employ the pose estimation to generate PoseBoxes and then utilize these PoseBoxes to learn local deep features. Wei *et al.* [29] apply Deeper Cut [30] to locate pedestrian key points and learn robust features from the part of pedestrian. Some approaches combine global deep feature and local deep features to represent pedestrian. Li *et al.* [31] learn global deep features from the whole pedestrian body, and employ Spatial Transformer Networks (STN) to obtain local deep features. Zhang *et al.* [22] jointly learn global feature using the coarse branch and multi-scale local features using different fine branches simultaneously. In this paper, we learn global, local and complementary deep features in a unified framework without any external operations, and obtain discriminative completed features.

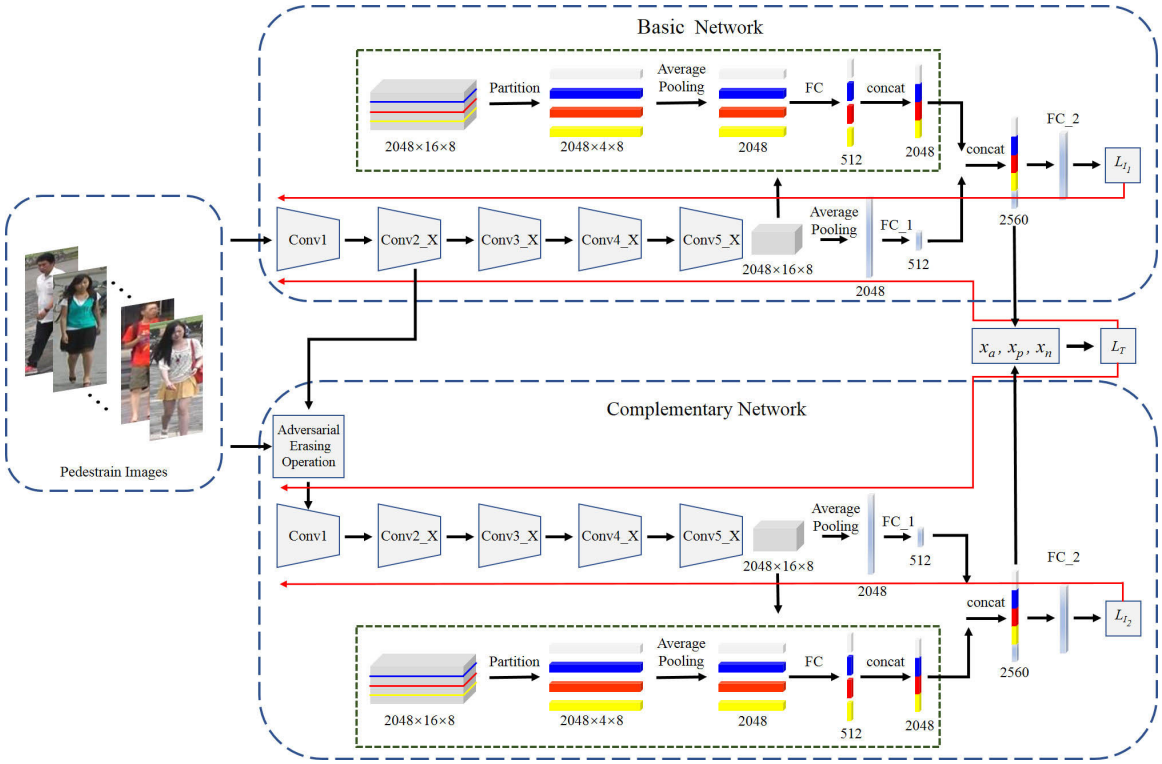

FIGURE 2. The structure of AEA. The black arrow indicates the direction of forward propagation and the solid red line represents the path of back propagation.

B. ATTENTION OF CNN

Attention mechanism is applied in many applications including image caption generation, visual question answering and person Re-ID [32]–[35]. In [32], the spatial attention is employed to relate each word with the corresponding image part, and make the latent attention align different structures for textual-visual matching. Liu *et al.* [33] propose the HydraPlus-Net for pedestrian analysis, which feeds multiple level attention maps into various feature layers so as to learn multiple attentions. Zhou *et al.* [34] introduce the temporal attention model for person Re-ID that automatically selects the discriminative frames. In [35], a multi-stream deep model is proposed to obtain representations from aligned pedestrian images using attention mechanism.

In addition, many approaches focus on how to visualize the attention map of CNN. Zhou *et al.* [36] present the Class Activation Mapping (CAM) to visualize the class-discriminative image regions without any bounding box annotations. Selvaraju *et al.* [37] combine the gradient of attention map and the final convolutional layer to visualize the localization map.

C. ADVERSARIAL LEARNING

Adversarial learning is widely utilized in domain adaptation, image generation and so on. In [38], Generative Adversarial Network (GAN) is proposed to generate new samples, which employs a generative model of learning sample distribution against a discriminative model of distinguishing generated samples from real ones. Tzeng *et al.* [39] present the Adversarial Discriminative Domain Adaptation (ADDA) to implement adversarial learning between the discriminator and the target CNN in order to obtain domain-invariant representations. Recently, adversarial learning has been applied in person Re-ID society. Liu *et al.* [40] propose the Adversarial Binary Coding (ABC) to guide the extraction of binary codes using adversarial learning for efficient person re-identification. Huang *et al.* [41] adversarially generate occluded samples and combine them with training samples to fine-tune the CNN model.

III. METHOD

In this section, we present the proposed AEA including the basic network and complementary network. First, we present the architecture of the basic network, and then describe the complementary network including adversarial erasing operation and dynamic learning. Finally, the losses of AEA are explained in detail.

A. BASIC NETWORK

The performance of AEA heavily depends on the basic network, and therefore we enable the basic network to fuse global and local deep features. The architecture of the basic network is shown in the upper branch of Fig. 2. We modify the ResNet-50 [42] as the backbone network as shown in Table 1. Specifically, we remove the down-sampling of Conv5_X to increase the size of convolutional activation maps, and correspondingly change the filter size of average

TABLE 1. The architecture of the basic network for learning global deep features.

Name	Filters	Stride	Output Size
Conv1	$\left[7 \times 7, 64\right] \times 1$	(2, 2)	$64 \times 128 \times 64$
Max Pooling	3×3	(2, 2)	$64 \times 64 \times 32$
Conv2_X	$\begin{bmatrix} 1 \times 1, & 64 \\ 3 \times 3, & 64 \\ 1 \times 1, & 256 \end{bmatrix} \times 3$	$\begin{bmatrix} (1,1) \\ (1,1) \\ (1,1) \end{bmatrix}$	$256 \times 64 \times 32$
Conv3_X	$\begin{bmatrix} 1 \times 1, & 128 \\ 3 \times 3, & 128 \\ 1 \times 1, & 512 \end{bmatrix} \times 4$	$\begin{bmatrix} (1,1) \\ (2,2),(1,1) \\ (1,1) \end{bmatrix}$	$512 \times 32 \times 16$
Conv4_X	$\begin{bmatrix} 1 \times 1, & 256 \\ 3 \times 3, & 256 \\ 1 \times 1, & 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} (1,1) \\ (2,2),(1,1) \\ (1,1) \end{bmatrix}$	$1024 \times 16 \times 8$
Conv5_X	$\begin{bmatrix} 1 \times 1, & 512 \\ 3 \times 3, & 512 \\ 1 \times 1, & 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} (1,1) \\ (1,1) \\ (1,1) \end{bmatrix}$	$2048 \times 16 \times 8$
Average Pooling	16×8	(1, 1)	$2048 \times 1 \times 1$
FC_1			

pooling to 16×8. In addition, we add a 512-dim fully-connected layer (FC_1) after the average pooling to reduce the feature dimensionality. In order to integrate local deep features, we divide Conv5_X into four horizontal parts, and then follow an average pooling and a 512-dim FC for each horizontal part. Afterwards, we concatenate the four 512-dim features, and further combine with the output of FC_1 to form the pedestrian feature (2,560-dim). Finally, the FC_2 is treated as the classifier to accomplish the classification task. The neuron number of FC_2 is equal to the identity number of person Re-ID dataset, i.e., 751, 702 and 767 for Market1501, DukeMTMC-reID and CUHK03, respectively. Given a pedestrian image with the size of 256×128, the basic network extracts the 2,560-dim vector to represent the pedestrian image.

B. COMPLEMENTARY NETWORK

The basic network learns global and local deep features for pedestrian images, which only focuses on a part of pedestrian. In order to mine new clues from other parts, we design the complementary network to mine the complementary features in an adversarial learning way. The architecture of the complementary network is shown in the lower branch of Fig. 2, and its backbone network is the same as the basic network except for adversarial erasing operation.

The complementary network aims at learning features complementary to the basic network and an intuitive way is to extract features from the parts where the basic network does not concern on, i.e., non-salient areas of pedestrian. Thus, it is key to locating the non-salient areas of pedestrian. Attention is a set of spatial maps, which represents the region of interest

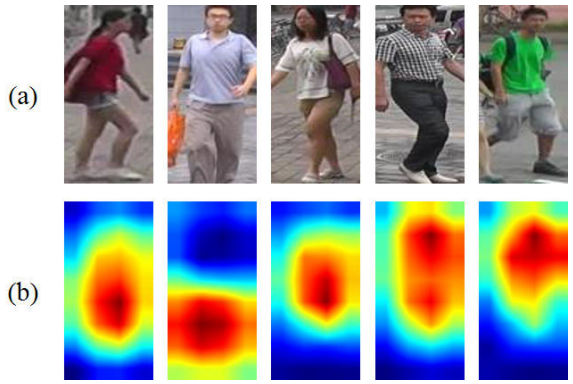

FIGURE 3. (a) Pedestrian images and (b) their attention maps extracted from the basic network. The warmer color denotes higher value, that is, these areas are the focus of basic network.

to the CNN model [43]. We utilize the attention map to locate the non-salient areas of pedestrian. The attention map is defined as:

$$A_l = \sum_c |M_{l,c}|^2 \tag{1}$$

where $M_{l,c}$ expresses the c-th channel of convolutional activation maps in the l-th layer. It should be noted that the operations in Eq. 1 are all element-wise. Fig. 3 shows some pedestrian images and their attention maps extracted from the basic network where we can see that the attention map highlights the salient areas of pedestrian.

To locate the non-salient areas of pedestrian, we conduct the adversarial erasing operation on the attention map and the process is shown in Fig. 4. Specifically, we first feed a pedestrian image with the size of 256×128 into the basic network and extract the convolutional activation maps from Conv2_X which is a $256 \times 64 \times 32$ tensor. Then, we acquire the attention map by using Eq. 1, and its size is 64×32. In order to simplify the mapping operation from the attention map to the original image, we resize the attention map into the same size with the original pedestrian image, i.e., 256×128, using bilinear interpolation. Afterwards, we select the top $R\%$ largest values from the resized attention map as salient areas, and map these salient areas to the corresponding positions of pedestrian image. Finally, we erase these pixels on the pedestrian image to generate the erased pedestrian image. It is worth noting that the erased pedestrian image has the same label as the original one. There are two advantages of erased pedestrian images. Firstly, since the erased pedestrian images are obtained by erasing the salient areas of pedestrian, they are adversarial to the basic network and the generalization ability of AEA could be improved. Secondly, the eased pedestrian images have the same identities as the original ones, but they are different. Hence, the diversity of training images is increased.

We generate the erased pedestrian images based on the attention map, but the attention map is variable during the training process due to the updated parameters of deep model. Fig. 5 shows different statuses of attention maps in the

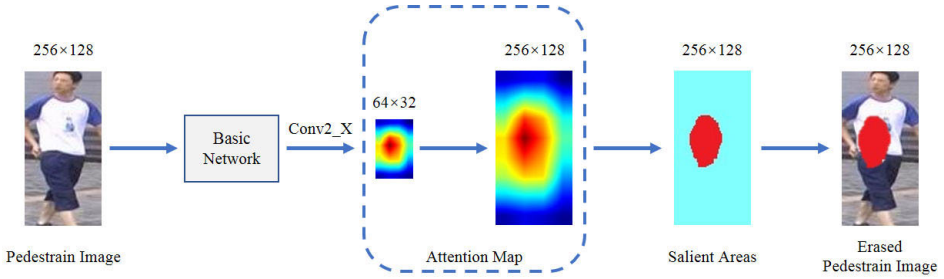

FIGURE 4. The process of adversarial erasing operation.

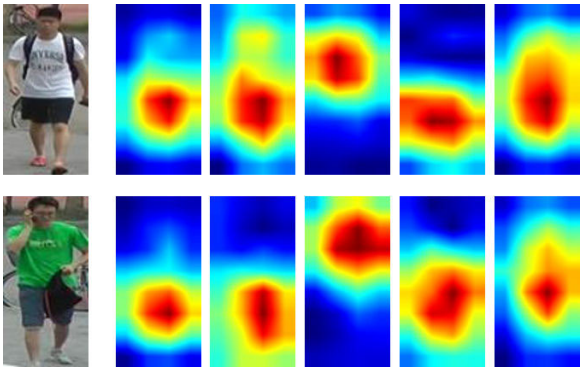

FIGURE 5. The pedestrian images and their different statuses of attention maps, where the first column represents pedestrian images and the remaining columns represent their attention maps at different statuses.

training stage, from which we can see that AEA focuses on different salient areas of pedestrian in different statuses. To match the dynamic status of AEA and generate reasonable erased pedestrian images, we propose a dynamic strategy to erase the salient areas of pedestrian in the attention map. Specifically, we conduct adversarial erasing operation at each iteration, and then generate the erased pedestrian images as the mentioned above.

After obtaining erased pedestrian images using adversarial erasing operation and dynamic strategy, we utilize them to train the complementary network, and meanwhile extract 2,560-dim complementary deep features to represent pedestrians.

C. LOSSES OF AEA

We expect to obtain discriminative completed features of pedestrians, and therefore simultaneously utilize the cross-entropy loss and the triplet loss in the proposed AEA.

Specifically, we apply the cross-entropy loss to train the basic network using original pedestrian images:

$$L_{I_1} = -\sum_{t=1}^{N} q(t)\log(p(t)) \tag{2}$$

where N is the total number of pedestrian identities, $q(t)$ is the label distribution, and $p(t)$ is the predicted probability of pedestrian image belonging to the t-th identity.

The label distribution $q(t)$ is expressed as:

$$q(t) = \begin{cases} 0 & t \neq y \\ 1 & t = y \end{cases} \tag{3}$$

where y represents the ground-truth identity label.

As for the prediction probability $p(t)$, it is formulated as:

$$p(t) = \frac{\exp(v_t)}{\sum_{i=1}^{N} \exp(v_i)} \tag{4}$$

where v_i represents the output of the i-th neuron of FC_2.

Similarly, we employ the entropy loss to train the complementary network using erased pedestrian images, and denote the loss as L_{I_2}. Note that the formulas of L_{I_2} and L_{I_1} are the same.

To take full advantage of erased pedestrian images, the triplet loss is fed by pedestrian images and their erased pedestrian images to learn an effective metric. The triplet loss is formulated as follows:

$$L_T = [\, d_{ap} + m - d_{an}]_+ \tag{5}$$

where the symbol $[Z]_+$ represents the formula $max(Z, 0)$, m is the margin between positive and negative pairs, d_{ap} is the distance between the anchor and positive samples, and d_{an} is the distance between the anchor and negative samples. Specifically, d_{ap} and d_{an} are expressed as:

$$d_{ap} = \max_{p} ||f(x_a) - f(x_p))||_2 \tag{6}$$

$$d_{an} = \min_{n} ||f(x_a) - f(x_n))||_2 \tag{7}$$

where $\{x_a, x_p, x_n\}$ indicate a set of triplets, $f(x_a)$, $f(x_p)$ and $f(x_n)$ represent feature vectors of the anchor sample x_a, the positive sample x_p and the negative sample x_n respectively, and $||.||_2$ expresses L_2 norm, i.e., Euclidean norm. These feature vectors are the concatenation of global and local features, and therefore the dimensions of them are 2,560. Afterwards, these features are all normalized by L_2 norm. Here, the positive sample refers to the pedestrian image with same identity as the anchor sample, while the negative sample is from different identities. It should be noted that triplets contain both original pedestrian images and erased pedestrian images.

The overall objective function of AEA can be written as:

$$L = L_{I_1} + \alpha_1 L_{I_2} + \alpha_2 L_T \tag{8}$$

TABLE 2. Results of the proposed AEA and four baselines on the three databases.

Methods	Market-1501		CUHK03				DukeMTMC-reID	
	rank-1	mAP	Labeled		Detected		rank-1	mAP
			rank-1	mAP	rank-1	mAP		
Baseline 1	92.5	81.5	65.6	62.2	64.1	60.3	83.4	68.9
Baseline 2	90.8	79.8	64.2	60.0	62.9	58.1	81.8	67.2
Baseline 3	93.6	84.7	68.2	64.2	65.1	61.9	85.2	72.0
Baseline 4	94.7	86.6	69.9	66.4	66.2	63.1	86.6	73.6
AEA	**95.9**	**87.9**	**71.2**	**67.6**	**67.8**	**65.5**	**88.4**	**75.5**

where α_1 and α_2 are the parameters which control the relative importance of objective function.

As shown in Fig. 2, we use the solid red line to represent the path of back propagation. During back propagation, the total loss L in which the total loss values of L_{I_1}, L_{I_2} and L_T are utilized to optimize the basic network and the complementary network simultaneously.

In the test stage, we first feed pedestrian images into the trained AEA model, and then extract features from the basic and complementary network, respectively. Afterwards, concatenate them to form final pedestrian feature which is a 5,120-dim feature vector.

IV. EXPERIMENTS

We introduce databases and experimental settings in detail. Afterwards, we proof the advantages of AEA and compare with other approaches, and also analysis key parameters of AEA.

A. DATABASES AND PROTOCOLS

Market-1501 [44] is divided into the training set (12,936 images with 751 identities), the test set (12,936 pedestrian images with 751 identities) and the query set (3,368 pedestrian images).

CUHK03 [45] includes labeled set and detected set, where 7,365 pedestrian images with 767 identities are used for training, 5,332 pedestrian images with 700 identities are utilized for testing and 1,400 pedestrian images are utilized for querying.

DukeMTMC-reID [46] provides 16,522 pedestrian images with 702 identities in training set, 17,661 pedestrian images with 1,110 identities in the test set, and 2,228 pedestrian images in the query set.

B. IMPLEMENTATION DETAILS

We initialize the basic network and complementary network of AEA using ResNet-50 [47]. As for FC_1 in the two sub-networks, we set the parameter of Leaky ReLU to 0.1 and the rate of Dropout to 0.5. In the training stage, we resize all pedestrian images to 256×128. Batch size is set to 32 for the basic network including 8 identities and 4 pedestrian images for each identity. As for the complementary network, every batch is also set to 32 which contains 8 identities and each identity comprises 4 erased pedestrian images.

The margin m of the triplet loss in Eq. 5 is set to 0.3,0.3 and 0.35 for Market-1501, DukeMTMC-reID and CUHK03, respectively. The number of epoch is 100, and learning rate is fixed to 0.01 before 75-th epochs and decreased by 0.1 in the remaining 25 epochs. In order to enable the basic network to predict the salient parts properly, we first train the basic network for 10 epochs, and then we train the basic network and the complementary network simultaneously. In Eq. 8, we experimentally set α_1 and α_2 to 1. We concatenate the features extracted from the basic network and the complementary network as the final representation, thus obtaining a 5,120-dim feature vector. We then normalize features using L_2 norm, and calculate the similarity between them according to the Euclidean distance.

TABLE 3. Results of the proposed AEA and other methods on Market-1501.

Methods	rank-1 (%)	mAP (%)
BoW [44]	34.4	14.1
SSDAL [48]	39.4	19.6
LOMO+XQDA [49]	43.8	22.2
Gated [50]	65.9	39.6
IDE [51]	73.7	51.5
PAR [52]	81.0	63.4
SVDNet [53]	82.3	62.1
LSRO [46]	84.0	66.1
APR [54]	84.3	64.7
TriNet [55]	84.9	69.1
AOS [41]	86.5	70.4
REDA [56]	87.1	71.3
DPFL [57]	88.6	72.6
DML [58]	89.3	70.5
MLFN [59]	90.0	74.3
HA-CNN [60]	91.2	75.7
DuATM [61]	91.4	76.6
PAAN [62]	92.4	77.6
AANet [63]	93.9	83.4
CFCNN [22]	94.0	81.2
TRFD [64]	94.7	84.5
JDGL [65]	94.8	86.0
AEA	**95.9**	**87.9**

C. ADVANTAGES OF AEA

In this subsection, we compare the proposed AEA with four baselines to verify the advantages of AEA.

1) BASELINE 1

It only employs pedestrian images to train the basic network with the cross-entropy loss and the triplet loss. The input of FC_2 is also regarded as the input of the triplet loss. In the test stage, it treats the input of FC_2 (2,560-dim) in the basic network as the pedestrian representations.

2) BASELINE 2

It utilizes original pedestrian images to train the basic network using the cross-entropy loss and generates erased

TABLE 4. Results of the proposed AEA and other methods on CUHK03.

Methods	Labeled Set (%)		Detected Set (%)	
	rank-1	mAP	rank-1	mAP
LOMO+XQDA [49]	14.8	13.6	12.8	11.5
IDE [51]	22.2	21.0	21.3	19.7
DPFL [57]	40.7	37.0	-	-
SVDNet [53]	40.9	37.8	41.5	37.3
HA-CNN [60]	44.4	41.0	41.7	38.6
AOS [41]	-	-	47.1	43.3
TriNet [55]	49.9	46.7	50.5	46.5
MLFN [59]	54.7	49.2	52.8	47.8
REDA [56]	58.1	53.8	55.5	50.7
CFCNN [22]	-	-	64.6	58.4
TRFD [64]	70.1	66.5	66.6	64.2
AEA	**71.2**	**67.6**	**67.8**	**65.5**

pedestrian images using the adversarial erasing operation and the dynamic strategy. Meanwhile, the erased pedestrian images are employed to train the complementary network using the cross-entropy loss and the triplet loss. When testing, the 2,560-dim deep features extracted from the complementary network are treated as the pedestrian representations.

3) BASELINE 3

It adopts the same network structure with the proposed AEA, but erases salient areas based on the original pedestrian image. Specifically, for Baseline 3, we first transform the pedestrian image (RGB) into the gray image, and treat the pixel positions with the top $R\%$ largest values in the gray image as the salient areas. Then, we erase these salient areas at the corresponding positions of the pedestrian image (RGB) to obtain the erased pedestrian image. It should be noted that since the original pedestrian images are invariant, the salient areas and erased pedestrian images of Baseline 3 are immutable in the training process. Hence, for convenience we erase the salient areas only once in the training process. In Baseline 3, the pedestrian image is represented by the 5,120-dim feature vector, which is the same as the proposed AEA.

4) BASELINE 4

Compared with AEA, it does not utilize the dynamic strategy. Specifically, we first train the basic network using the cross-entropy loss and the triplet loss until convergence. Then, we extract the attention maps from Conv2_X of the trained basic network. Afterwards, we use the adversarial erasing operation to obtain erased pedestrian images. Finally, we utilize original pedestrian images as the input of the basic network and meanwhile we utilize erased pedestrian images as the input of the complementary network. Note that we do not adopt the dynamic strategy in Baseline 4, that is, we employ the adversarial erasing operation only once, and therefore the erased pedestrian images are invariant in the training stage. The test stage of Baseline 4 is the same as AEA.

We list the comparison results in Table 2 where the following conclusions can be drawn. Firstly, the proposed AEA outperforms four baselines in all situations. Secondly, the proposed AEA yields higher accuracy than Baseline 1 and Baseline 2 on the three databases. It is because Baseline 1 and Baseline 2 extract deep features from the basic network and the complementary network respectively, while the proposed AEA fuses the two kinds of complementary features in a unified framework. Thirdly, Baseline 1 achieves better results on rank-1 accuracy and mAP than Baseline 2. It is because Baseline 1 learns feature representations from whole pedestrian images, while Baseline 2 extract features from erased pedestrian images where the salient areas are occluded. Fourthly, Baseline 4 surpasses Baseline 3, which demonstrates the attention map contains more semantic information than the original pedestrian image. Hence, the attention map is utilized to locate salient areas in the proposed method. Finally, Baselines 4 is experimentally validated as inferior to the proposed AEA. It is because the proposed AEA adopts the dynamic strategy which could match the dynamic status of AEA while Baseline 4 does not utilize the dynamic strategy.

TABLE 5. Results of the proposed AEA and other methods on DukeMTMC-reID.

Methods	rank-1 (%)	mAP (%)
LOMO+XQDA [49]	30.8	17.0
IDE [51]	65.2	45.0
LSRO [46]	67.7	47.1
APR [54]	70.7	51.9
TriNet [55]	72.4	53.5
SVDNet [53]	76.7	56.8
AOS [41]	79.1	62.1
DPFL [57]	79.2	60.6
REDA [56]	79.3	62.4
HA-CNN [60]	80.5	63.8
MLFN [59]	81.0	62.8
DuATM [61]	81.8	64.6
PAAN [62]	82.6	65.5
CFCNN [22]	85.7	72.4
TRFD [64]	85.8	72.9
JDGL [65]	86.6	74.8
AEA	**88.4**	**75.5**

D. COMPARISON WITH OTHER METHODS

Table 3 shows the results on Market1501 where the proposed AEA obtains the best results (95.9% rank-1, 87.9% mAP). Compared with AANet [63] and CFCNN [22] which utilize global and local deep features to represent the pedestrian, the proposed AEA achieves better results than them, because the proposed AEA not only learns global and local deep

TABLE 6. Some results about REDA, AOS and AEA.

Databases	AEA		REDA		AOS	
	rank-1(%)	mAP(%)	rank-1(%)	mAP(%)	rank-1(%)	mAP(%)
Market-1501	95.9	87.9	93.2	84.3	93.7	84.9
CUHK03 (Labeled Set)	71.2	67.6	67.8	63.6	68.1	64.0
CUHK03 (Detected Set)	67.8	65.5	65.0	61.6	65.8	62.3
DUKEMTMC-REID	88.4	75.5	84.8	71.5	85.9	72.4

features, but also learns completed features using an adversarial manner.

Table 4 shows the results on CUHK03. As for the labeled setting, the proposed AEA obtains 71.2% rank-1 accuracy and 67.6% mAP. As for the detected setting, the proposed AEA obtains 67.8% rank-1 accuracy and 65.5% mAP. These results surpass other methods in Table 4.

Table 5 shows the results on DukeMTMC-reID. The proposed AEA yields 88.4% rank-1 accuracy and 75.5% mAP, which obtains the best results compared with other methods. Since DukeMTMC-reID is an extremely challenging person Re-ID database, the superiority of AEA is verified once again.

The proposed AEA generates erased pedestrian images in the learning stage, and therefore it could be regarded as a kind of data augmentation method. Hence, we also compare AEA with other data augmentation methods, i.e., REDA [56] and AOS [41]. For fair comparison, we utilize the same network structure to implement REDA and AOS. The experimental results are shown in Table 6 where the proposed AEA is superior to REDA and AOS. Compared with REDA [56] that randomly occludes rectangle regions in original pedestrian images, the proposed AEA surpasses REDA by a large margin. It is because the proposed AEA learns erased pedestrian images from the attention map using adversarial learning, whereas REDA does not have the learning process. Furthermore, the proposed AEA obtains better performance than AOS [41] where the erased area is invariant. Additionally, AOS learns the erased area using a separate network, that is the erasing operation and the feature learning are separate. On the contrary, the proposed AEA obtains variant erased areas using dynamic strategy, and learns erased areas and features in a unified framework.

E. PARAMETER ANALYSIS

In this subsection, we evaluate several important parameters which influence the performance of AEA, i.e., the parameter R to regulate the percentage of erased areas, the margin m in Eq. 5 to adjust the distances between positive and negative pairs, and α_1 and α_2 in Eq. 8 to control the relative importance of objective function. In addition, we also discuss the selection of convolutional layer for the attention map.

We first evaluate the influence of the parameter R on the three databases, and the evaluation results are shown in Fig. 6. We can see that when R is too small, the performance decreases because the erased pedestrian image is similar to the original one, and meanwhile when R goes too large, the performance also declines because of erasing too many salient areas. Hence, we set R to 10 for Market-1501 and DukeMTMC-reID, and 20 for CUHK03.

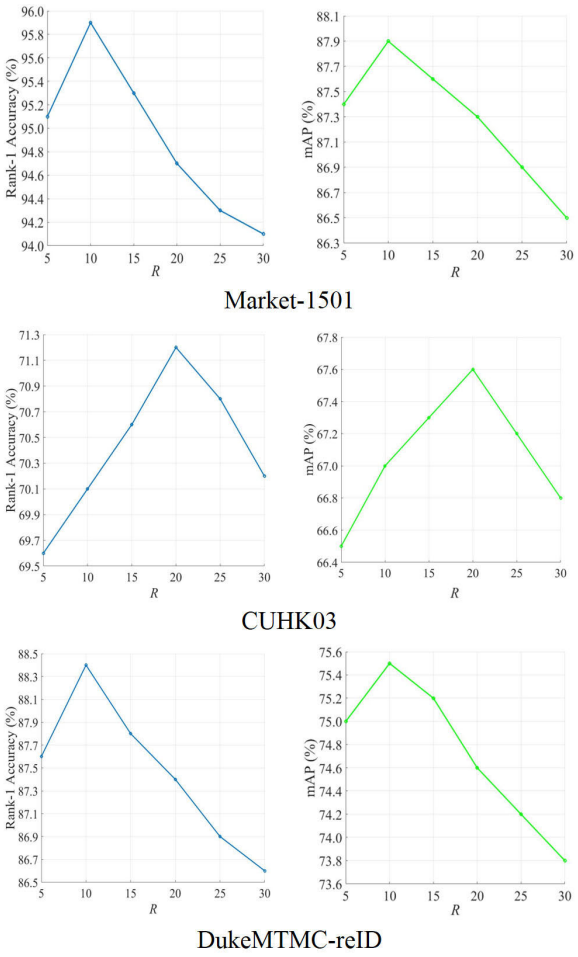

FIGURE 6. Influence of the parameter R in rank-1 accuracy and mAP on the three databases.

TABLE 7. Effect of the parameters α_1 and α_2 in rank-1 accuracy on the Market-1501 database.

rank-1 (%) α_1 \ α_2	0.1	0.5	1	1.5	2
0.1	91.5	92.9	93.4	93.2	93.0
0.5	92.4	94.3	94.5	94.4	93.2
1	93.1	94.9	**95.9**	93.6	94.8
1.5	92.6	94.1	94.6	94.5	92.7
2	91.9	92.3	93.4	93.1	92.5

Then, we test the influence of the margin m, and the results are shown in Fig. 7. We change m from 0.1 to 0.5 with the step of 0.1. When m is equal to 0.3, the rank-1 accuracy and mAP obtain the best results on Market-1501 and DukeMTMC-reID. Similarly, the proposed AEA achieves the best results on CUHK03 when m is set to 0.35.

Afterwards, we discuss the influence of α_1 and α_2 in Eq. 8 on Market-1501, and our experiments have shown

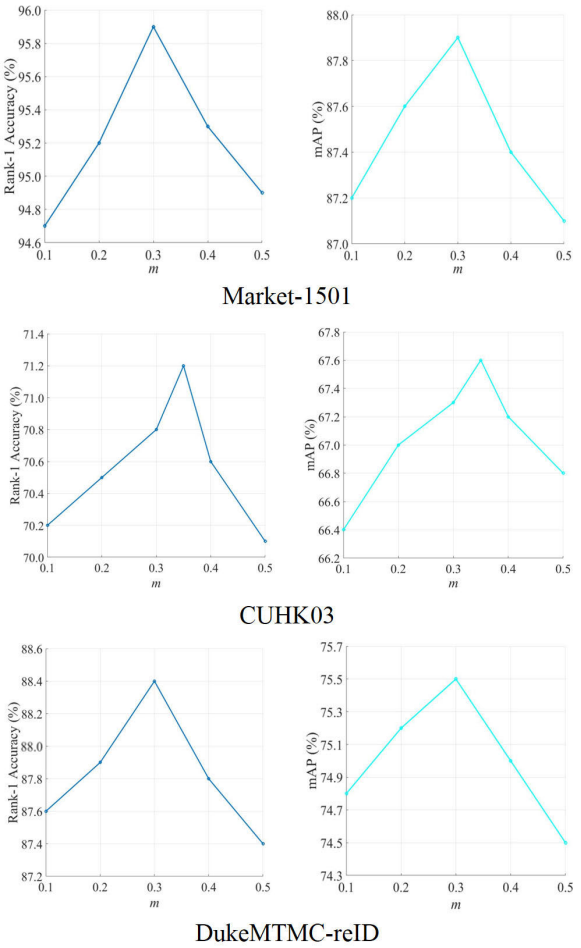

FIGURE 7. Influence of the margin m in rank-1 accuracy and mAP on the three databases.

TABLE 8. Effect of the parameters α_1 and α_2 in mAP on the Market-1501 database.

mAP (%) α_1 \ α_2	0.1	0.5	1	1.5	2
0.1	80.1	81.5	84.6	83.9	83.8
0.5	83.0	84.1	85.3	84.2	84.1
1	83.7	85.2	**87.9**	85.2	84.8
1.5	82.4	82.8	85.6	84.6	84.0
2	81.3	82.2	84.2	83.5	81.7

TABLE 9. Effect of extracting convolutional activation maps from different layers on Market-1501.

	Conv1	Conv2_X	Conv3_X
rank-1 (%)	94.0	**95.9**	94.8
mAP (%)	85.8	**87.9**	86.5

that the conclusions can be generalized to CUHK03 and DukeMTMC-reID as well. We take the values of α_1 and α_2 from a discrete set $\{0.1, 0.5, 1, 1.5, 2\}$. From Table 7 and 8, it is clear that the proposed AEA reaches the best results when $\alpha_1 = \alpha_2 = 1$.

Finally, we also discuss the selection of convolutional activation maps $M_{l,c}$ in Eq. 1 which are used to form the attention map. The results are shown in Table 9 where extracting convolutional activation maps from Conv2_X obtains better performance than others. It should be noted that the experimental results can be generalized to CUHK03 and DukeMTMC-reID.

V. CONCLUSION

In this paper, we have proposed the AEA to learn discriminative completed features for person Re-ID. The AEA extracts global and local deep features from original pedestrian images, and learns complementary features using erased pedestrian images. In order to generate effective erased pedestrian images, we have presented the adversarial erasing operation to locate salient areas on the attention map, and adopted the dynamic strategy to match the dynamic status of AEA. Because of the diversity of training samples, i.e., original and erased pedestrian images, the generalization ability of AEA is improved. The experimental results on three large-scale person Re-ID databases have demonstrated that the proposed AEA achieves better performance than other state-of-the-art methods.

REFERENCES

[1] X. Yang, M. Wang, and D. Tao, "Person re-identification with metric learning using privileged information," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 791–805, Feb. 2018.
[2] S. Tan, F. Zheng, L. Liu, J. Han, and L. Shao, "Dense invariant feature-based support vector ranking for cross-camera person reidentification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 356–363, Feb. 2018.
[3] Z. Zhang, M. Huang, S. Liu, B. Xiao, and T. Durrani, "Fuzzy multi-layer clustering and fuzzy label regularization for unsupervised person re-identification," *IEEE Trans. Fuzzy Syst.*, early access, May 2, 2019, doi: 10.1109/TFUZZ.2019.2914626.
[4] J. Nino-Castaneda, A. Frias-Velazquez, N. Bo Bo, M. Slembrouck, J. Guan, G. Debard, B. Vanrumste, T. Tuytelaars, and W. Philips, "Scalable semi-automatic annotation for multi-camera person tracking," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2259–2274, May 2016.
[5] J. Berclaz, F. Fleuret, and P. Fua, "Multi-camera tracking and atypical motion detection with behavioral maps," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 112–125.
[6] A. B. Chan and N. Vasconcelos, "Bayesian Poisson regression for crowd counting," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 545–551.
[7] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2913–2920.
[8] Y. Zhou, L. Liu, and L. Shao, "Vehicle re-identification by deep hidden multi-view inference," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3275–3287, Jul. 2018.
[9] Z. Zhang, C. Wang, B. Xiao, W. Zhou, S. Liu, and C. Shi, "Cross-view action recognition via a continuous virtual path," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2690–2697.
[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, May 2015, pp. 135–153.
[11] G. G. Ghiasi, T. Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 10727–10737.
[12] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.
[13] W. Huang and L. Mei, "The combination of features extracted from different parts for person re-identification," in *Proc. 2nd Int. Conf. Robot. Automat. Sci. (ICRAS)*, Jun. 2018, pp. 1–5.
[14] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[15] S. Liu, X. Hao, and Z. Zhang, "Pedestrian retrieval via part-based gradation regularization in sensor networks," *IEEE Access*, vol. 6, pp. 38171–38178, 2018.
[16] 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.*, Sep. 2018, pp. 1–17.
[17] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.
[18] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
[19] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 483–499.
[20] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.
[21] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1077–1085.
[22] Z. Zhang, H. Zhang, and S. Liu, "Coarse-fine convolutional neural network for person re-identification in camera sensor networks," *IEEE Access*, vol. 7, pp. 65186–65194, 2019.
[23] 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. (CVPR)*, Jun. 2016, pp. 1249–1258.
[24] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5399–5408.
[25] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for practical person re-identification," in *Proc. Int. Conf. Pattern Recognit.*, Jun. 2014, pp. 34–39.
[26] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.
[27] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, Jun. 2019.
[28] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," 2017, *arXiv:1701.07732*. [Online]. Available: http://arxiv.org/abs/1701.07732
[29] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2017, pp. 420–428.
[30] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 34–50.
[31] 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. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7398–7407.
[32] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1908–1917.
[33] 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," 2017, *arXiv:1709.09930*. [Online]. Available: http://arxiv.org/abs/1709.09930
[34] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6776–6785.
[35] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognit.*, vol. 86, pp. 143–155, Feb. 2019.
[36] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[37] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 618–626.
[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 2672–2680.
[39] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2962–2971.
[40] Z. Liu, J. Qin, A. Li, Y. Wang, and L. Van Gool, "Adversarial binary coding for efficient person re-identification," 2018, *arXiv:1803.10914*. [Online]. Available: http://arxiv.org/abs/1803.10914
[41] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5098–5107.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[43] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*. [Online]. Available: http://arxiv.org/abs/1612.03928
[44] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[45] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[46] 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. (ICCV)*, Oct. 2017, pp. 3754–3762.
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[48] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 475–491.
[49] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.
[50] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 791–808.
[51] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," 2016, *arXiv:1610.02984*. [Online]. Available: http://arxiv.org/abs/1610.02984
[52] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3219–3228.
[53] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3820–3828.
[54] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," 2017, *arXiv:1703.07220*. [Online]. Available: http://arxiv.org/abs/1703.07220
[55] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737
[56] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896
[57] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2590–2600.
[58] A. Dunn and N. Robles, "Polynomial partition asymptotics," 2017, *arXiv:1705.00384*. [Online]. Available: http://arxiv.org/abs/1705.00384
[59] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2109–2118.
[60] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.
[61] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5363–5372.

[62] Y. Zhang, X. Gu, J. Tang, K. Cheng, and S. Tan, "Part-based attribute-aware network for person re-identification," *IEEE Access*, vol. 7, pp. 53585–53595, 2019.
[63] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7134–7143.
[64] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.
[65] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.

SHUANG LIU (Senior Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. She is currently an Associate Professor with Tianjin Normal University, Tianjin, China. She has published over 50 articles in major international journals and conferences. Her research interests include computer vision, remote sensing, and deep learning.

XIAOLONG HAO is currently pursuing the master's degree with Tianjin Normal University, Tianjin, China. His research interests include sensor networks, person re-identification, and deep learning.

RONGHUA ZHANG is currently an Associate Professor with Shihezi University, Xinjiang, China. She has published about 20 articles in major international journals and conferences. Her research interests include computer vision and digital image processing.

ZHONG ZHANG (Senior Member, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently an Associate Professor with Tianjin Normal University, Tianjin, China. He has published about 110 articles in international journals and conferences, such as the IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Pattern Recognition*, the IEEE TRANSACTIONS ON CIRCUITS SYSTEMS VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *Signal Processing* (Elsevier), CVPR, ICPR, and ICIP. His research interests include computer vision, remote sensing, and deep learning.

TARIQ S. DURRANI is currently a Research Professor with the University of Strathclyde, Glasgow, U.K. He has authored 350 publications and supervised 45 Ph.D. students. His research interests include AI, signal processing, and technology management. In 2018, he was elected as a Foreign Member of the U.S. National Academy of Engineering. He is a Fellow of the U.K. Royal Academy of Engineering, the Royal Society of Edinburgh, IET, and the Third World Academy of Sciences.