Person Re-identification Based on Adaptive Relation Attention Network in Intelligent Monitoring System for the IoB

Journal:	*Transactions on Engineering Management*
Manuscript ID	TEM-21-1362.R2
Manuscript Type:	Special Section: Information Cybersecurity Management in Cloud-Edge Computing using Artificial Intelligence (AI) and Blockchain Technologies
Keywords:	person re-identification, artificial intelligence, attention mechanism, intelligent monitoring, internet of behavior, blockchain
Subject Category:	Digital Technologies and Analytics, Design

Person Re-identification Based on Adaptive Relation Attention Network in Intelligent Monitoring System for the IoB

IEEE Publication Technology Department

***Abstract*—The Internet of Behavior (IoB) which combines Internet of Things (IoT) and Artificial Intelligence (AI) plays an important role in building smart city. As a significant part of data collection and analysis, intelligent monitoring system needs robust algorithms for analyzing data to make corresponding feedback. The development of person re-identification (re-id) has benefited from deep learning methods, especially for the IoB application. However, most existing person re-id methods under intelligent monitoring can't solve some problems in the real world, such as occlusion, background cluster. In this paper, we propose an ARA Network for person re-id based on Adaptive Relation Attention (ARA) mechanism, which can address the above challenges effectively. The ARA module consists of Relation Branch and Adaption Branch. Relation Branch captures the global structural information by mining relation among feature nodes. Adaption Branch generates dynamic weights for attention features Relation Branch produced. Our constructed intelligent IoB system can acquire the behavior status of pedestrian at different times and places, providing corresponding feedback rapidly. Data transmission and storage in our system are built in a decentralized way which is based on blockchain. Our person re-id method outperforms many state-of-the-art methods on the CUHK03, Market-1501 and DukeMTMC-reID, showing excellent robustness.**

***Index Terms*—person re-identification, artificial intelligence, attention mechanism, intelligent monitoring, internet of behavior, blockchain.**

I. Introduction

THE rapid development of AI and IoT techniques facilitated the birth and development of the Internet of Behavior (IoB). As the core carrier of smart city, the IoB system can collect data by the way of IoT and process information in the producing process and conduct data analysis on the AI platform, finally provides intelligent public service (e.g. health, security) according to needs and environment. An ideal IoB system includes wide and various data collection modules and sensor networks, summarizing behavioral patterns on multiple dimensions. In order to achieve this task, there is a series of algorithms in need to analyze data and generate stereo behavioral patterns. In this work, we design a robust construction of intelligent monitoring system for IoB, and mainly focus on person re-identification algorithm based on deep learning applied to the system.

With the rapid growth of population of cities, monitoring system in cities is responsible for important functions, especially for security. It is the basis of building an intelligent and complex IoB system to accurately grasp the identity information of pedestrians in the city and their behavior status at different times and places. Person re-identification (re-id) based on deep learning algorithm is suited for this task well. good person re-identification algorithm can retrieve and match images with the same pedestrian as the query image in a large scale image data set in low resolution and variable lighting conditions. Only on this basis, the IoB system can realize some more advanced and intelligent city service functions. In terms of security, it can help functional departments more efficiently obtain information about people's identities and track. Moreover, during the epidemic of COVID-19, the IoB system based on person re-id also can help government help governments manage staff turnover more intelligently, mitigating the spread of the epidemic.

In order to get more accurate identification results, there are much effort has been invested in studying person re-id algorithm based on deep learning [1], [2], [3], [4], [5], [6], [7]; however, there are still some dominant problems which have to be solved, including background cluster and resource consumption. Recently, plenty of research works [8], [9], [10], [11], [12], [13], [14] resort to visual attention mechanism so as to focus on the informative parts (e.g. characteristic spatial regions) of feature maps and suppress the irrelevant information (e.g. background). The inherent attention module is designed to automatically identify and select the more valuable parts of an pedestrian image, and has a weakly-supervised train manner (i.e., no explicit labelling information is provided to identify which areas are more important). This idea is very helpful to solve the influence of complex urban environment on the recognition results. Specifically, spatial attention mechanism [12], [13], [14] focuses on the mining of spatial domain information which involves directing attention to a location in space, it allows CNN to selectively process visual information of input fearture maps within the visual field. Relatively, channel attention [15], [16] is proposed to adaptively recalibrates channel-wise convolutional responses by explicitly modelling inter-dependencies among channels. And the combination of spatial and channel attention has also achieved satisfactory results applied in person re-Id [13].

In this paper, we proposed Adaptive Relation Attention

Manuscript created October, 2020; This work was developed by the IEEE Publication Technology Department.

Fig. 1. Multiple levels of learned attention in the model. (a) The original image. (b)-(d) The learned attentions of different attention modules. With the stacking of convolutional layers and attention modules, the attention to the input image becomes more focused.

(ARA) module which perform well on mining global relation information to efficiently extract discriminative features for person re-id. And we designed ARA Network based on ARA module which effectively extract valuable information from input data and obtain high-level semantic features pivotal. Our method not only achieves stable and excellent recognition results on several public datasets, but also consumes less computing resources compared with some methods that design complex components or utilize local features, since it only rely on global features. Therefore, it is more suitable for data analysis and fast feedback of IoB system. Fig. 1 shows our learned attention after different attention modules for multiple levels of features on the pedestrian images. With the stacking of convolution layers in the single-flow structure, features are more concentrated on the discriminant regions.

Overall, the contributions of this work are four-fold:

1) We propose a robust construction of intelligent monitoring system for IoB, which combines IoT techniques and deep learning algorithm. This system is a smart integration of hardware and software, aiming at mastering the identity information and behavior track of pedestrians, providing fast feedback and support for more advanced public service functions. The designed smart device uses a modified camera to capture and further process the captured video information, which can be transmitted to remote devices by 5G technology.
2) We propose Adaptive Relation Attention(ARA) module which consists of two branches. Relation Branch mines global structure patterns by utilizing relation information. Adaption Branch generates dynamic attention weights for Relation Branch. By combining these two branches, ARA module can globally learn the relation among feature nodes and adaptively re-weight relation attention during training stage. By using global relational information, ARA module can produce accurate attention features, improving the robustness of the model in complex scenarios.
3) We propose ARA Network based on ARA module for person re-identification task. We integrate some effective tricks to improve the improve the model during training stage, and adopt appropriate data augmentation methods to make the model learn occlusion-tolerant representation. By these approaches, the model has good power on feature extraction and obtain accurate high-level semantic features. Extensive experiments performed on the standard benchmark datasets including CUHK03, Market-1501, DukeMTMC-reID, show that our approach has state-of-the-art performance and good robustness.

The rest of the paper is organized as follows. Section II focuses on the related works. Sections III detail the mentioned modules of our proposed algorithm. Section IV presents detailed experiment setting and experimental results where performance is compared with the state-of-the-art approaches. We conclude the paper and discuss the direction of our future work in Section V.

II. Related Work

A. Internet of Things

The Internet of Behaviors utilizes diverse techniques, such as Iot, artificial intelligence to analyze behavioral patterns. IoT establishes the basic framework and hardware technology foundation for IoB. By the end of 2021, about 28 billion smart devices are connected across the world[17]. With the development of Fifth Generation (5G) , artificial intelligence, edge computing and other technologies, the combination of IoT and these techniques will be the development trend for a long time[18], [19], [20]. By combining computation offloading technology and deep neural network, Chen et al.[19] proposed a feasible method to construct a real-world intelligent application. In [21], Huang et al. proposed a software-defined infrastructure design in a decentralized fashion which is helpful for exploring big data on the Internet adequately.

Inspired by previous works, we propose a robust construction of intelligent monitoring system for IoB. Our framework makes full use of the cloud platform to apply our deep learning algorithm, capturing behavior patterns of pedestrian and providing corresponding feedback.

B. Person Re-identification

In recent years, a lot of previous works in the domain of person re-identification primarily are based on hand-crafted feature. As the research progresses, Convolutional Neural Network (CNN) became the first choice of methods for representation learning over the years, presenting state-of-the-art results in

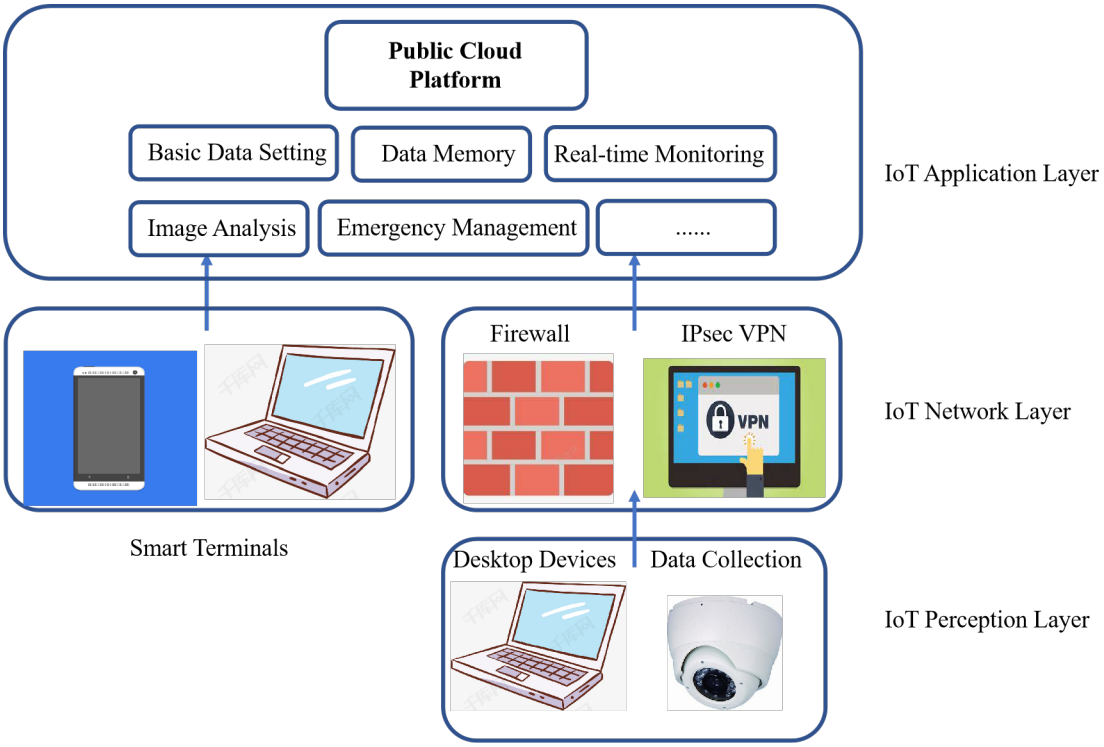

Fig. 2. The overall framework of the proposed system. The IoB system consists of four main parts, including Public Cloud Platform (PCP), smart terminals connected to the platform, a data collection module with desktop devices and smart cameras and an intranet server for safe data transmission from the collection part.

this vision area. In [22], Sarfraz et al. proposed a CNN embedding which incorporates fine-grained and coarse person pose information which perform robust on several benchmarks. Adopting striped-based method, Sun et al. [5] designed Part-based Convolutional Baseline (PCB) and a refined part pooling (RPP) method. In[23], Zheng et al. proposed a siamese network architecture utilizing global feature information to obtain a discriminative CNN embedding and a similarity metric. In [24], Zhong et al. proposed re-ranking method which can improve final results without any human interaction. In [25], Luo et al. design and collect some practical tricks including the BNNeck and center loss, etc. which are often helpful to training model for re-id.

C. Attention Mechanism

Recently, attention mechanisms, inspired by the human sensing process, have been studied extensively in Natural Language Processing [26] and Computer Vision [16]. In person re-id, the person misalignment [4] and background biases [27] discourage learning a robust representation. Visual attention mechanisms aim at emphasizing informative parts for identification, while depreciating irrelevant ones (e.g., background and occluded regions). In order to introduce more contextual information, Woo et al. [28] designed Convolutional Block Attention Module (CBAM) with a large filter size of 7×7 over the spatial features to produce a spatial attention map . For utilizing second order statistical information, Fang et al. [10] proposed Bilinear Attention network with an Attention in Attetion mechanism to build inter-dependency among the second order local and global features. In [29], Zhang et al. proposed RGA module which utilizes global features effectively.

Inspired by previous methods, we propose ARA in this paper. Different from previous works, ARA module can exploit the internal relations of input features and has better robustness under larger intra-class variance.

III. The Main Methods

The focus of this section is on the details of our proposed methods. The intelligent monitoring system for IoB we designed is shown in Fig. 2. The whole structure of our deep learning network is depicted in Fig. 3. The architecture of ARA module is shown in Fig 5.

A. IoB system modeling

As shown in Fig. 2, the framework consists of four main parts, including Public Cloud Platform (PCP), smart terminals connected to the platform, a data collection module with desktop devices and smart cameras and an intranet server for safe data transmission from the collection part. Data transmission and storage should be built in a decentralized way which is based on blockchain[30], [31], [32]. The PCP consists of image analysis, emergency management, real-time monitoring, basic data setting, data memory and some other necessary functions. With the assistance of 5G communication technique[33] and distributed computing, our framework is capable of accomplishing a series of public service functions including public security and some other task that requires searching people.

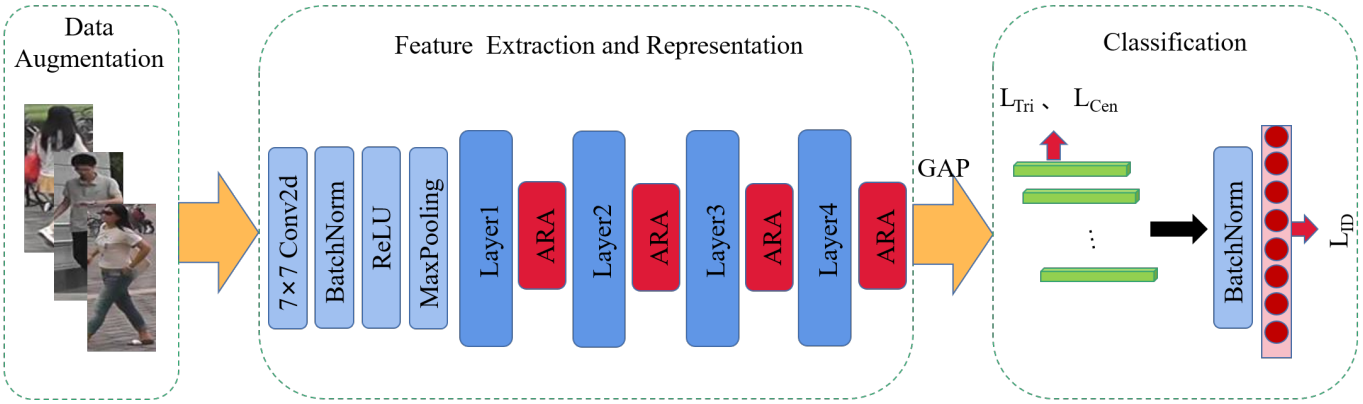

Fig. 3. The overview of total architecture of the model in training stage. The architecture is mainly composed of three parts. The data augmentation part includes a set of image processing methods for re-id. In the feature extraction part, we adopt ResNet-50 as baseline, and Layer1-Layer4 are stacked bottleneck blocks. The classification part combines suitable loss functions which help to improve feature representation power during training.

The data collection module collects the video through the camera, and uploads the collected video to the PCP through the network. The PCP receives the video data and extract the key frame of the video, and then analyzes the key frame image data, which can greatly reduce the computing burden of the server. To make full use of hardware resources and make the web applications high-performance, multi-threading technique is used in both the embedded device side and server side. The embedded equipment of the system contains eight CPU cores, and the hardware resources of the embedded development board can be fully utilized by using multi-thread technique. In order to ensure the stability of I/0, we set up a thread for video transmission over TCP specially. Meanwhile, the main thread and the child threads executing in parallel on different CPUs without affecting video capture, display, and statistics, making the application running better. Moreover, there is an independent thread used for data analysis, and another independent thread is in charge of data storage. For communication between threads, we adopt Qt signal and slot mechanism, which makes each thread executing concurrently and sequentially.

B. Network Architecture

As shown in Fig. 3, during training stage, the network architecture of our proposed method consists of three parts: data augmentation, feature extraction and representation, classification.

The purpose of data augmentation is emphasizing some features which are helpful for the model and enhancing image interpretation and recognition effect. In the data augmentation part, we flip and rotate images randomly which can expand the diversity of training samples according to previous instructional work[34]. Specially, in order to alleviate the frequent occlusion problem in person re-identification task, we adopt Random Erasing Augmentation (REA)[35] method which can effectively improve robustness of the model to occluded images during training.

The feature extraction and representation part is the core of our person re-identification algorithm. An ideal person re-identification model can extract the key information from the input data and finally obtain the discriminant feature representation. Then, in the embedding space of feature representation, the sample with the highest similarity to the query sample is sought as the result of re-identification. Therefore, the quality of the discriminative features generated by the network directly determines the performance of the model. In this part, we adopt a pre-trained classification network as the backbone. By integrating several ARA modules into the backbone, we make the model focus on the key regions and finally obtain the discriminative high-level semantic feature for re-id.

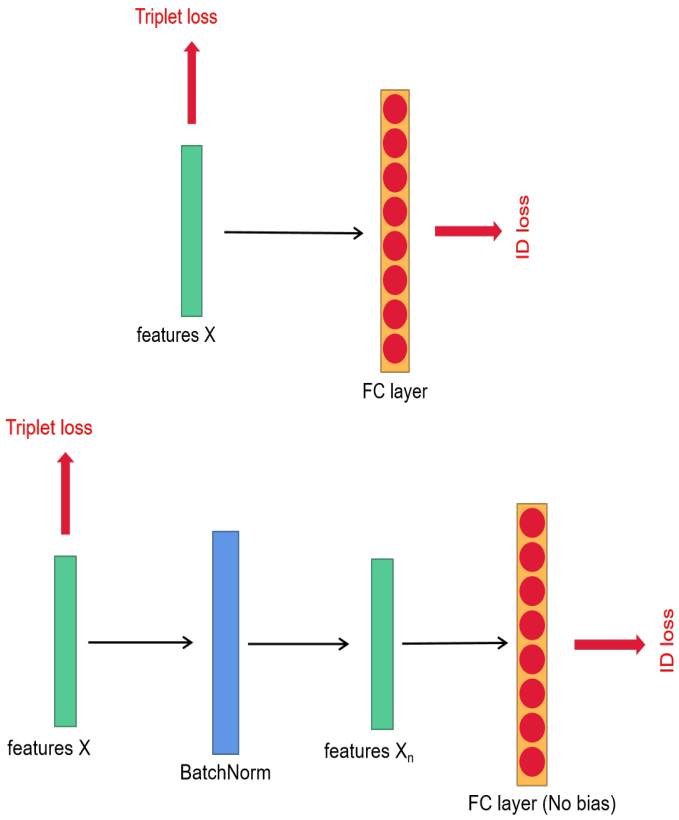

Fig. 4. The BatchNorm layer is the major improvement in the BNNeck structure. It makes loss functions easier to converge in a simple way.

Compared with the traditional classification task, person re-identification is a zero-shot learning task. There is no sample intersection between the training set and the test set, in other words, the pedestrian in the test set will not appear in the training set, which is designed based on the real-world requirements of the monitoring system. But in the training stage, we still train our network in the traditional way of classification task so that the model has good ability of feature representation. In several state-of-the-art methods, the classifier part combining triplet loss and ID loss achieves good results. In the classification part, we use the combination of triplet loss and ID loss as the loss function of the model during the training stage. And on this basis, the BNNeck structure[25] address the inconsistency between ID loss and triplet loss helping both loss functions easier to converge in the training process. The BNNeck structure is shown in Fig. 4. In this work, ID loss is cross entropy loss with label smoothing[36], and triplet loss is computed as:

$$L_{Tri} = [d_p - d_n + \gamma]_+ \tag{1}$$

where d_p and d_n are feature distances of positive pair and negative pair. γ is the margin of triplet loss, and $[z]+$ equals to $max(z, 0)$. According to the experience, we set γ to be 0.3. Usually, the overall optimization target is:

$$L_{=}L_{ID} + L_{Tri} \tag{2}$$

where L_{ID} is cross entropy loss with label smoothing. The disadvantage of triplet loss is that it only considers the gap between d_p and d_n but ignores the absolute values of them. To make up for that, we adopt centor loss[37] which learns a center for features of each class and penalizes the distances between the features and their corresponding class centers in the meantime. The center loss function is formulated as:

$$L_{Cen} = \frac{1}{2}\sum_{j=1}^{B} \|x_j - c_{y_j}\|_2^2 \tag{3}$$

where x_j denotes the feature representing the jth image used to calculate centor loss and y_j is the label of the jth image in a mini-batch. B is the number of batch size. Center loss represents the direction of optimization to increase intra-class compactness. Finally, our overall loss function can be expressed as:

$$L_{=}L_{ID} + \beta_1(L_{Tri} + \beta_2 L_{Cen}) \tag{4}$$

where β_1 and β_2 are two balanced weights. β_1 is set to be around 0.5. And we set β_2 a small number, like 0.001, to make centor loss complementary to triplet loss.

C. Adaptive Relation Attention

In order to capture the global structural information and enhance robustness of attention module in complex scenarios, we propose Adaptive Relation Attention (ARA) module which includes two branches: Attention Branch and Adaption Branch. Attention Branch makes use of the relation among feature nodes to mine global structural patterns. Adaption Branch generates adaptive weights dynamically to enhance robustness of Attention Branch.

In general, for a feature vector set $X = \{x_i \in \mathbb{R}^d, i = 1, \cdots, N\}$ of N correlated features with each of d dimensions, the target of attention mechanism is learning a weight mask denoted by $\boldsymbol{a} = (a_1, \cdots, a_N)$ for the N features to re-weight them according to their importance.

Relation Branch For a feature set $X = \{x_i \in \mathbb{R}^d, i = 1, \cdots, N\}$, the relation between x_i and x_j are denoted as $r_{i,j}$. In this work, the pairwise relations among all the feature nodes are represented by an affinity matrix $R \in \mathbb{R}^{N\times N}$, where the relation between node i and j is $r_{i,j} = R(i, j)$. $r_i = [R(i, :), R(:, i)]$, where $R(i, :)$ denotes the i^{th} row of R and $R(:, i)$ denotes the i^{th} column of R, which represents relation between x_i and all the other feature nodes in some way. As shown in Fig. 5, inside the ARA module, we calculate r_i of each feature node to captures global structural patterns.

Given an intermediate feature tensor $X \in \mathbb{R}^{C\times H\times W}$ of width W, height H, and C channels from a CNN layer, we design Relation Branch which effectively utilize relation information, for learning an attention map of size $H \times W$. On the basis of taking the C-dimensional feature vector at each spatial position as a feature node, we scan the spatial positions and assign their identification number as $1, \cdots, N$. We define the N feature nodes as $x_i \in \mathbb{R}^C$, where $i = 1, \cdots, N$.

As shown in Fig. 6, we define the pairwise relation (i.e. affinity) $r_{i,j}$ from node i to node j as a dot-product affinity in the embedding spaces as:

$$r_{i,j} = f_s(x_i, x_j) = \varphi_s(x_i)^T \cdot \psi_s(x_j) \tag{5}$$

where φ_s and ψ_s are two embedding functions implemented by a 1×1 spatial convolutional layer followed by batch normalization (BN) and ReLU activation,

$$\varphi_s(x_i) = ReLU(W_\varphi x_i) \tag{6}$$

$$\psi_s(x_i) = ReLU(W_\psi x_i) \tag{7}$$

where $W_\varphi \in \mathbb{R}^{\frac{C}{s_1}\times C}$ and $W_\psi \in \mathbb{R}^{\frac{C}{s_1}\times C}$. s_1 is a predefined positive integer that controls the dimension reduction ratio. Note that BN operations are all omitted to simplify the notation. Similarly, we can get the affinity from node j to node i as $r_{j,i} = f_s(x_j, x_i)$. We use the pair $(r_{i,j}, r_{j,i})$ to describe the bi-directional relations between x_i and x_j. Then, we represent the pairwise relations among all the nodes by an affinity matrix $R_s \in \mathbb{R}^{N\times N}$.

From the i^th feature node, we stack its pairwise relations with all the nodes in a certain fixed order, i.e., node identities as $j = 1, 2, \cdots, N$, to obtain a relation vector $r_i = [R(i, :), R(:, i)] \in \mathbb{R}^{2N}$. The set of r_i which defined as $R' = r_i, i = 1, \cdots, N$ can be used as the basis of attention that relatively utilizing relation information applied in the space domain.

Adaption Branch In order to improve robustness of Attention Branch, we design Adaption Branch which dynamically re-weight the relation features making final attention features focusing on discriminative regions accurately. In this branch, we adopt Global Average Pooling (GAP) and Global Max

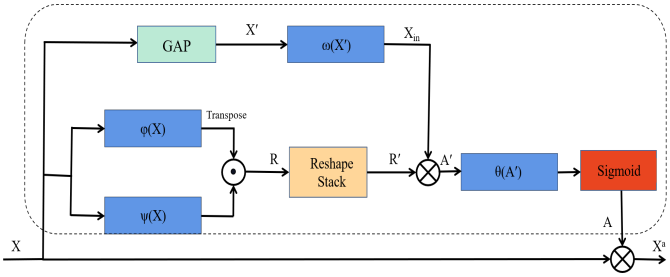

Fig. 5. The succinct structure of Adaptive Relation Attention module. X is a feature set applied attention mechanism on certain scope (space domain or channel domain).$\varphi(\cdot), \psi(\cdot), \omega(\cdot)$ and $\theta(\cdot)$ are embedding functions. $\odot$ indicates matrix multiplication. GAP operates global average pooling. $\bigotimes$ indicates element-wise multiplication.

Pooling (GMP) to aggressively summarize the global scope of input features. A combination of both can capture complementary information, so we use a weighted summation of GMP and GAP by learning their importance for a more robust result. The summation X' can be defined as

$$X' = \alpha G_{GMP} + (1-\alpha) G_{GAP} \tag{8}$$

$$\alpha = Softmax(W_\alpha G) \tag{9}$$

Where $G \in \{GMP, GAP\}$, and matrix W_α is learnable parameter. Then the final output vector of Adaption Branch can be defined as $X_a = \omega(X')$, we define $\omega(X')$ as an embedding function, which map the global feature to the latent space of the same dimension of r_i. The combination of the output of Adaption Branch X_a and the output of Attention Branch r_i, is:

$$A' = X_a \bigotimes R' \tag{10}$$

where $\bigotimes$ indicates element-wise multiplication.

For $A' = a'_i \in \mathbb{R}^{2N}, i = 1, \cdots, N$, the module need to obtain an effective spacial attention map $A \in \mathbb{R}^{H \times W \times 1}$ finally, so we use embedding function $\theta(a'_i)$ to map A' to a tensor with shape of $H \times W \times 1$. We found that the structure with two 1×1 convolutional layers as $\theta(\cdot)$ is better than one with one layer. So we define that:

$$\theta(a'_i) = W'_\theta ReLU(W_\theta a_i) \tag{11}$$

In order to re-weight feature reasonably, the final attention values should be distributed between 0 and 1. In past works, $Sigmoid$ function is often used to achieve this target, and mostly of them achieved good results. So we also adopt this manner and the final produced attention can be defined as:

$$a_i = Sigmoid(W'_\theta ReLU(W_\theta a'_i)) \tag{12}$$

And the spacial attention map $A = a_i \in \mathbb{R}, i = 1, \cdots, N$. Finally, the attention mask emphasize the significant elements of its input feature tensor X by element-wise multiplication as:

$$X^a = X \bigotimes A \tag{13}$$

X^a is the final output of an ARA module passed to the next part of the model.

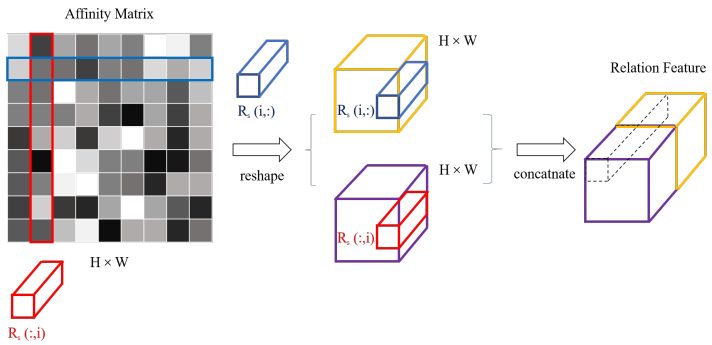

Fig. 6. Diagram of Relation Branch. When computing the attention at a feature position, in order to capture information of global scope, we stack the pairwise relation items, i.e., its affinities with all the feature positions, for learning the attention with convolutional operations.

TABLE I
COMPARISON OF DIFFERENT ADAPTION BRANCH.

	Market1501		DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
Baseline	94.0	83.6	86.1	75.7
w/o	95.1	85.9	87.0	76.8
only GAP	95.7	87.1	87.7	77.4
only GMP	95.6	87.0	87.5	77.2
constant α	95.9	87.9	88.0	78.1
learnable α	96.2	88.6	88.5	78.7

IV. EXPERIMENT

A. Datasets

In this section, we evaluate our model across three benchmark datasets, including **Market-1501**[38], **DukeMTMC-ReID**[39] and **CUHK03**[40]. **Market-1501** is a large person re-ID dataset, consisting of 32668 person images of 1501 identities shooted by 6 different cameras. The training set consists of 12936 training images belonging to 751 identities, whereas the test set consists of 19732 testing images of the remaining 750 identities. All the images are detected using a DPM.

DukeMTMC-ReID consists of 16522 training images of 702 identities, 2228 query images of 702 identities and 17661 gallery images. All the images are oberved under 8 different camera views and are originally collected for video-based person tracking and re-identification. Some samples in DukeMTMC-ReID are shown in Fig. 7.

CUHK03 is collected with 6 cameras and each person captured in this dataset is observed by two disjoint camera views.

TABLE II
EFFECT OF THE NUMBER OF ARA MODULE

	Market1501		DukeMTMC	
	Rank-1	mAP	Rank-1	mAP
Baseline	94.0	83.6	86.1	75.7
ARA × 1	94.6	84.4	86.8	76.4
ARA × 2	95.3	85.3	87.6	77.2
ARA × 3	95.6	86.1	87.9	77.8
ARA × 4	96.2	87.8	88.5	78.7

TABLE III
COMPARISONS TO STATE OF THE ARTS ON MARKET1501, DUKEMTMC-REID AND CUHK03

Method	Market1501		DukeMTMC		CUHK03(Labeled)		CUHK03(Detected)	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Mancs[41]	93.1	82.3	84.9	71.8	69.0	63.9	65.5	60.5
HA-CNN[13]	91.2	75.7	80.5	63.8	44.4	41.0	41.7	38.6
BAT-net[10]	95.1	87.4	87.7	77.3	-	-	73.2	76.2
DuATM[42]	91.4	76.6	81.8	64.6	-	-	-	-
MGCAM[43]	83.8	74.3	-	-	50.1	50.2	46.7	46.9
MHN-6[9]	95.1	85.0	89.1	77.2	77.2	72.4	71.7	65.4
PCB[5]	93.8	81.6	83.3	69.2	63.7	57.5	-	-
PIE[44]	79.3	56.0	-	-	-	-	67.1	71.3
PSE[22]	84.0	90.3	79.8	85.2	-	-	-	-
HPM[45]	94.2	82.7	-	-	63.9	57.5	-	-
OSNet[46]	94.8	84.9	-	-	-	-	72.3	67.8
Ours	96.2	88.6	88.5	78.7	81.5	78.4	80.1	77.3

Fig. 7. Some samples in DukeMTMC-reID dataset. Images are captured by 8 different cameras at random different times. Low resolution images can be taken under many realistic conditions, and suitable for person re-identification algorithm research

It has totally 13164 person images of 1467 identities, of which 767 identities belong to the training set and 700 identities belong to the testing set. There are both DPM-detected and hand-labeled bounding boxes offered by CUHK03, and we do experiment on both two types of the dataset.

We use the standard mean average precision(mAP) and cumulative matching characteristics (CMC) at rank-1 accuracy to evaluate the performance.

B. Implementation Details

Following the common practices in re-id[47], [48], we adopt ResNet-50[34] pre-trained on ImageNet[49] as the backbone network in the feature extraction and representation part. And we integrate our ARA module into ResNet-50 to improve its ability.Within RGA modules, the ratio parameters s1 and s2 are set to be 8.

The training data are cropped to 256 × 128 and augmented by horizontal flipping, rotating and random erasing[35]. The probability of flipping and rotating is set to be 0.5. And the probability of erasing is set to 0.4. These parameters follow the experience of previous works. The batch size is 128. And after experimental trials and adjustments, we adopt the Adam optimizer to train the model for 500 epochs with the learning rate of 8×10^{-4} and the weight decay of 5×10^{4}. All of our experiments are conducted on PyTorch 1.1 with 2 2080Ti GPUs.

C. Ablation Study

Following the common practice, we further perform extra experiments to verify the effectness of our proposed methods on two popular datasets Market1501 and DukeMTMC-ReID.

Table I shows the comparison of different Adaption Branch and the baseline. We observe that Adaption Branch obviously improves the performance of ARA. Ether GAP or GMP is helpful to final result. The combination of GAP and GMP with constant α achieves 1.9% higher Rank-1 accuracy and 4.3% higher mAP than the baseline on Market1501. However when using learnable α, The improvement is even more significant. Compared to using constant α, Adaption Branch with learnable α achieves 0.3% higher Rank-1 accuracy and 0.7% higher mAP on Market1501.

As shown in Table 1, ARA mechanism has a significant improvement on the baseline. For Market1501, the baseline has 94.0% Rank-1 accuracy and 83.6% mAP accuracy. A single ARA module improves by 0.6% and 0.8% on Rank-1 accuracy and mAP accuracy respectively. And the performance of the model improves significantly as more ARA modules were added to the network structure. Finally, the model with four ARA modules applying attention to different levels of feature achieve a large margin over the baseline. According to our experiment, the addition of four ARA modules is the best choice for the baseline in this work, which effectively improves the ability of feature representation and doesn't increase too much computational burden at the same time.

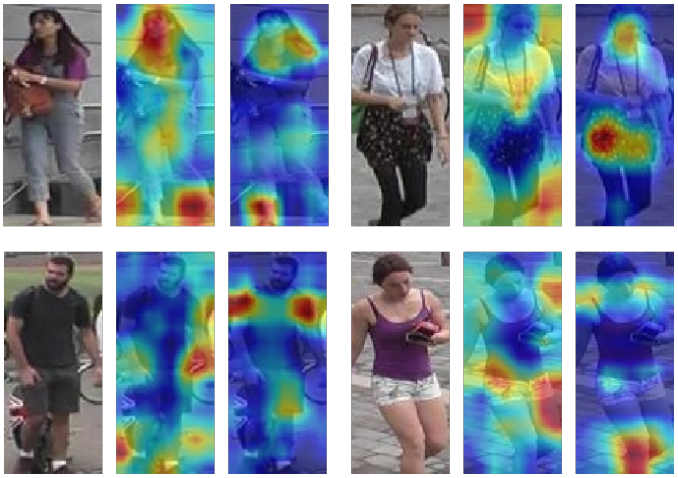

Fig. 8. The three images in each group are, from left to right, the original image, the visualization of CBAM, and the visualization of our model. Colors ranging from blue to red represent the different areas of the image that the model is paying attention to. Blue represents the areas with the least attention, red represents the areas with the most attention.

D. Comparison to State-of-the-Arts

We compare our model with state-of-the-arts methods in Table III. Some previous methods achieved good performance on Market1501, But did not achieve satisfactory accuracy on CUHK03. Compared to them, our method achieved excellent result on Market1501, and it also performs well on CUHK03, showing better robustness. In comparison with all other approaches, our method achieved good performance which outperforms most of the others a large margin. Compared to BAT-net which also adopts attention idea, our method achieves 1.1% higher Rank-1 accuracy and 1.2% higher mAP than BAT-net on Market1501. Overall, compared to state of the arts, our model performs well across three datasets, i.e. Market1501, DukeMTMC and CUHK03, showing good robustness.

E. Visualization

To validate the ability of our method to emphasize key regions, we visualize the final feature representation of baseline and our model respectively on origin images. As shown in Figure 5, we do the visualization results by Grad-CAM[50], where different colors indicate different weight distribution among spacial regions.

As shown in Fig.8, it is obvious that the areas which our model focuses on in images are more centralized and purposeful compared with CBAM. Although CBAM also concentrates on discriminative regions, it still pay much attention to disruptive or useless regions (backgrounds). Our model, which effectively utilize relation information, performs well on identifying important parts over bodies and distinguishing between people and backgrounds. We observe that characteristic parts of body are usually emphasized, e.g. backpack straps and spotted skirt, which play significant roles in identifying people.

Fig. 9 shows the retrieval results, the images are sorted by their similarity to the query images. The images in the red boxes are the incorrectly identifid samples, and the images in the green boxes are the correctly identified samples. Obviously, most of results are correct. For misidentified samples, their clothing and body shape are very similar to the query samples. In this case, a small amount of human work can be done through smart terminals.

Fig. 9. In the re-identification task, the 10 result samples with the highest measure similarity to the query sample are ranked in order. The images in the green boxes are the correctly identified samples, and the images in the red boxes are the incorrectly identified samples

V. Conclusion

In this work, we designed a robust construction of intelligent monitoring system for IoB, which combines IoT techniques and deep learning algorithm. The system is a smart integration of hardware and software, aiming at mastering the identity information and behavior track of pedestrians, providing fast feedback and support for more advanced public service functions. And person re-identification method is the core algorithm of data analysis in our system.

In order to solve problems of occlusion, background cluster and variable illumination in person re-identification, we designed ARA Network based on ARA module which consists of Relation Branch and Adaption Branch. Relation Branch mines global structure patterns by utilizing relation information. Adaption Branch generates dynamic attention weights for Relation Branch which helps to produce better attention feature. By these approaches, our model has good power on feature extraction and obtain accurate high-level semantic features, showing state-of-the-art performance and good robustness on several benchmark datasets including CUHK03, Market-1501, DukeMTMC-reID in the experiment. However, our algorithm relies on labeled datasets, which imposes cost and data privacy limitations. In the future, our follow-up work will focus on unsupervised methods to solve these problems.

VI. Acknowledgements

This work is supported by the National Natural Science Foundation of China (NO. 61702226); the 111 Project (B12018); the Natural Science Foundation of Jiangsu Province (NO. BK20170200); Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Land and Resources (NO. KF-2018-03-065); the Fundamental Research Funds for the Central Universities (NO. JUSRP11854,NO. JUSRP11851).

References

[1] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 384–393.

[2] J. Liu, Z. Yang, T. Zhang, and H. Xiong, "Multi-part compact bilinear cnn for person re-identification," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 2309–2313.

[3] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision*. Springer, 2016, pp. 475–491.

[4] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–419.

[5] 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 *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 480–496.

[6] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8042–8051.

[7] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2109–2118.

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

[9] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 371–381.

[10] P. Fang, J. Zhou, S. K. Roy, L. Petersson, and M. Harandi, "Bilinear attention networks for person retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8030–8039.

[11] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8287–8294.

[12] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 369–378.

[13] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2285–2294.

[14] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2119–2128.

[15] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[17] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[18] B. Lin, Y. Huang, J. Zhang, J. Hu, X. Chen, and J. Li, "Cost-driven offloading for dnn-based applications over cloud, edge, and end devices," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5456–5466, 2019.

[19] X. Chen, M. Li, H. Zhong, Y. Ma, and C.-H. Hsu, "Dnnoff: Offloading dnn-based intelligent iot applications in mobile edge computing," *IEEE Transactions on Industrial Informatics*, 2021.

[20] G. Huang, X. Liu, Y. Ma, X. Lu, Y. Zhang, and Y. Xiong, "Programming situational mobile web applications with cloud-mobile convergence: An internetware-oriented approach," *IEEE Transactions on Services Computing*, vol. 12, no. 1, pp. 6–19, 2016.

[21] G. Huang, C. Luo, K. Wu, Y. Ma, Y. Zhang, and X. Liu, "Software-defined infrastructure for decentralized data lifecycle governance: Principled design and open challenges," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1674–1683.

[22] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 420–429.

[23] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, pp. 1–20, 2017.

[24] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1318–1327.

[25] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[27] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5794–5803.

[28] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[29] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3186–3195.

[30] O. Novo, "Blockchain meets iot: An architecture for scalable access management in iot," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1184–1195, 2018.

[31] Z. Huang, X. Su, Y. Zhang, C. Shi, H. Zhang, and L. Xie, "A decentralized solution for iot data trusted exchange based-on blockchain," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2017, pp. 1180–1184.

[32] S. S. Gill, S. Tuli, M. Xu, I. Singh, K. V. Singh, D. Lindsay, S. Tuli, D. Smirnova, M. Singh, U. Jain *et al.*, "Transformative effects of iot, blockchain and artificial intelligence on cloud computing: Evolution, vision, trends and open challenges," *Internet of Things*, vol. 8, p. 100118, 2019.

[33] S. Li, L. Da Xu, and S. Zhao, "5g internet of things: A survey," *Journal of Industrial Information Integration*, vol. 10, pp. 1–9, 2018.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[38] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[39] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European conference on computer vision*. Springer, 2016, pp. 17–35.

[40] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE*

conference on computer vision and pattern recognition, 2014, pp. 152–159.

[41] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 365–381.

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

[43] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[44] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Transactions on Image Processing*, vol. 28, no. 9, pp. 4500–4509, 2019.

[45] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8295–8302.

[46] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3702–3712.

[47] J. Almazan, B. Gajic, N. Murray, and D. Larlus, "Re-id done right: towards good practices for person re-identification," *arXiv preprint arXiv:1801.05339*, 2018.

[48] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.