Attention-Aligned Network for Person Re-Identification

Sicheng Lian*, Weitao Jiang*, Haifeng Hu*
*School of Electronics and Information Technology, Sun Yat-sen University, 510006

Abstract—Currently, attention mechanism receives enormous interest and has been extensively employed in the fields of Person Re-Identification (RE-ID), as it gains superior performance in learning discriminative feature representations. However, most off-the-shelf attention methods are still vulnerable to cross-view inconsistency problem. Besides, they merely exploit imprecise channel attention information and coarse-grained spatial attention of homogeneous scales, being insufficient to capture subtle differences among highly-similar individuals. To this end, we propose a novel Attention-Aligned Network (AANet) to address the aforementioned problems, in which a novel Omnibearing Foreground-aware Attention (OFA) module, Attention Alignment Mechanism (AAM) and an improved triplet loss with hard mining are proposed to learn foreground attentive features for RE-ID. Specifically, AANet firstly leverages OFA module to exploit heterogeneous-scale spatial attention and foreground-aware channel attention information. Then AANet further reduces the impact of background clutter and learns camera-invariant and background-invariant representations by virtue of AAM. Last but not least, an improved triplet loss with hard mining is also introduced to enhance the feature learning capability, which can jointly minimize the intra-class distance and maximize the inter-class distance in each triplet unit. Extensive experiments are carried out to demonstrate that the proposed method outperforms most current methods on three main RE-ID benchmarks.

Index Terms—Person Re-identification, Attention-Aligned Network, Omnibearing Foreground-aware Attention

1 Introduction

Person Re-Identification (RE-ID), a crucial component in video surveillance system, aims to correctly match pedestrian images captured by non-overlapping cameras. Currently, it plays a prominent role in extensive surveillance applications, encompassing multi-camera activity analysis [1], multi-camera tracking [2] and crowd counting [3].

Despite many significant advances in recent years, RE-ID still remains unsolved due to three major problems which are clearly presented in Fig.1. One of the main issues degrading the robustness of feature representation can be mainly attributed to disjoint camera views. Given that pedestrian images are captured arbitrarily in complicated and diversified circumstances, one person's appearance tends to change dramatically on account of misalignment as well as significant variances in person poses, occlusion and illumination conditions. Besides, similar background among different pedestrians can cause considerable difficulties during the attempts to identify the person of interest. Moreover, it's also challenging for RE-ID when the distinctions which can be utilized to distinguish between pedestrians are subtle. As shown in Fig.1, global visual cues such as clothing color and body shape are insufficient for correct identification whilst discriminative local visual cues (e.g. hairstyle and attachment) need to be exploited.

Many tentative efforts have been devoted to coping with above three aforementioned problems. Notably, the importance of attention mechanism is now more pronounced in the field of RE-ID [1] [2] [3] [4]. The main contribution of attention mechanism is biasing the distribution of existing resources towards the most informative part. Specifically, representations strengthened by the attention mechanism can better represent the images since more distinguishable distinctions are provided for pedestrians recognition. How

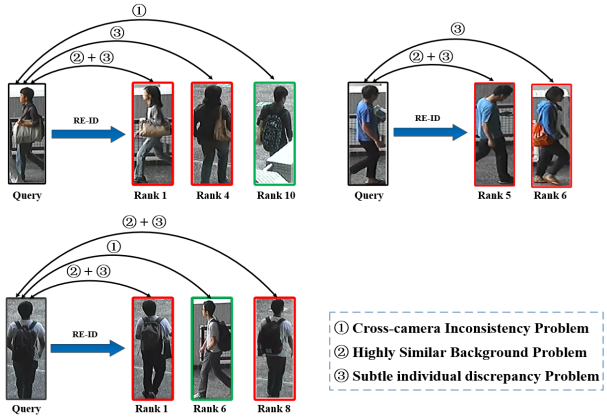


Fig. 1. Some typical examples vividly demonstrate three main problems which hinder real-world use of RE-ID. Based on our baseline model, gallery pictures are ranked by the distance to the query picture from left to right. The green/red bounding box indicates that the specific gallery image has the same/different identities as the query picture.

to employ attention mechanism and exploit attention information is now at the research frontier. Yet, the attention information obtained from existing attention models is still coarse-grained, thereby being insufficient to capture subtle discrepancies among individuals and failing to achieve accurate matching. More technically, most existing attention methods [2] [3] [5] tend to treat all pixels in each feature map equally and employ homogeneous-scale spatial attention, therefore being confined to mining coarse and simple information. What's more, these traditional attention models are also incapable of coping with cross-camera inconsistency which is a pressing problem and demands further study.

To fill the research gap and tackle all three challenges concurrently, in this paper, we propose a novel Attention-Aligned Network (AANet) which can not only learn discriminative feature representation and reduce adverse impacts of background by exploiting more comprehensive and precise attention information, but also realize cross-view consistency by performing attention alignment. We argue that it is better to incorporate the discriminative feature learning and attention alignment in an end-to-end network, since they are complementary with high compatibility and can benefit from each other during the training phase. Our contributions can be summarized and highlighted as follows:

(i) With the purpose of learning more discriminative features, we improve both channel and spatial attention mechanisms and then propose a novel Omnibearing Foreground-aware Attention (OFA) module which cascades Foreground-aware Channel Attention (FCA) and Multiscale Spatial Attention (MSA) in a coherent way. In this way, the OFA module produces the desired result in collecting comprehensive attention information of feature maps, explicitly preventing the loss of effective information.

(ii) We propose a novel Attention Alignment Mechanism (AAM) to explicitly guide attention map of last convolutional layer, ensuring spatial consistency of attentive regions among the images of the same identity. AAM amounts to supervise the end-to-end learning of consistent attentive regions of the same person. Specifically, it can be divided into three parts: (a) acquiring attention map, (b) selecting salient foreground areas, and (c) learning attention consistency via a new Attention Cosine Distance Loss (ACDL).

(iii) An improved triplet loss with hard mining is introduced to jointly minimize the intra-class distance and maximize the inter-class distance. With the purpose of remedying the drawback of neglecting to minimize the intra-class distance, we improve the triplet loss with hard mining by means of adding an extra loss item. By virtue of this extra constraint, the intra-class distance is strictly constrained, which is beneficial to realize the full potential of the original TriHard loss.

(iv) In order to confirm the superiority of AANet, we extensively conduct experiments on three mainstream RE-ID benchmark datasets. As the results indicate, AANet reaches rank-1 accuracy of 82.4% and 77.3% on CUHK03-NP (Labeled) and CUHK03-NP (Detected) [6] [7], 96.1% on Market-1501 [8] and 89.7% on DukeMTMC-reID [9] [10], which surpasses a broad range of state-of-the-art RE-ID models.

The rest of this work is organized as follows. Sec.2 illustrates some related works about Deep Learning-Based RE-ID models and Attention Mechanism. In Sec.3, the network structure of AANet and the proposed methods of Omnibearing Foreground-aware Attention module, Attention Alignment Mechanism and Classification Module will be elaborated respectively. Sec.4 gives the details of training strategies, describes our experiments, discusses the proposed method and shows the experiment results. Finally, the whole work is concluded in Sec.5.

2 Related Work

2.1 Deep Learning-Based RE-ID models.

Person Re-identification (RE-ID), a challenging task in computer vision, seeks to correctly match cross-view persons. Ideally, RE-ID models are expected to learn discriminative feature representations invariant and robust to external interference. Early works in the field of RE-ID mostly leverage hand-crafted feature representations [6] and learning latent spaces [49]. At present, deep learning methods dominate this community and demonstrate convincing superiority by delivering state-of-the-art results. The state-of-the-art RE-ID methods can be mainly divided into two groups accordingly, which are feature learning and metric learning methods.

The feature learning methods aim to learn more representative pedestrian features by virtue of various designed feature extractors. Chen *et al.* [11] develop a CNN-based appearance model to jointly learn scale-specific features and explicitly exploit information of various scales. Sun *et al.* [12] construct Part-based Convolutional Baseline (PCB) network to employ part-level features of pedestrian images. They also propose a refined part pooling method which is particularly designed to re-assign outliers in the parts. Zhang *et al.* [13] propose a unique dynamic programming to search the shortest path between two sets of features and achieve remarkable performance. Recently, attention mechanism is also proposed and employed in numerous RE-ID models, which will be elaborated in Sec. 2.2.

With similar purpose of obtaining discriminative feature representations, metric learning methods choose a different way to achieve the goal. They attempt to optimize inter-group and intra-group distance of feature representations, learning effective metrics to measure the similarity among images. Ding *et al.* [14] utilize triplet samples to train CNN-based models and attempt to decrease intra-group feature distances and increase inter-group feature distances. Furthermore, Chen *et al.* [15] further revise the triplet loss and propose a deep quadruplet network. In addition, other types of methods have also been introduced for RE-ID, such as transfer learning methods [16], dictionary learning methods [17], unsupervised learning [18] , etc.

2.2 Attention Mechanism.

Visual Attention is the process in which human brains always preferentially select the most salient regions to cope with. It selects salient regions intensively [19], ignores other uninformative areas [20] [21] [3], and then strengthens representations at those areas.

Currently, many researchers have made tentative efforts towards applying attention mechanism into deep neural network. In [22], a weakly supervised part selection method is proposed with spatial constraints, which is free of part annotations and bounding boxes in both training and testing phases. Peng *et al.* [23] firstly localize the objects of images with the saliency map, and then select discriminative parts of objects, and finally jointly employ both of them to boost the multi-view and multi-scale feature learning. Thereafter, Hu *et al.* [21] propose a novel architectural unit named squeeze-and-excitation block which are able to recalibrate channel-wise feature responses adaptively by modeling interdependencies between channels. Worthy of mention is

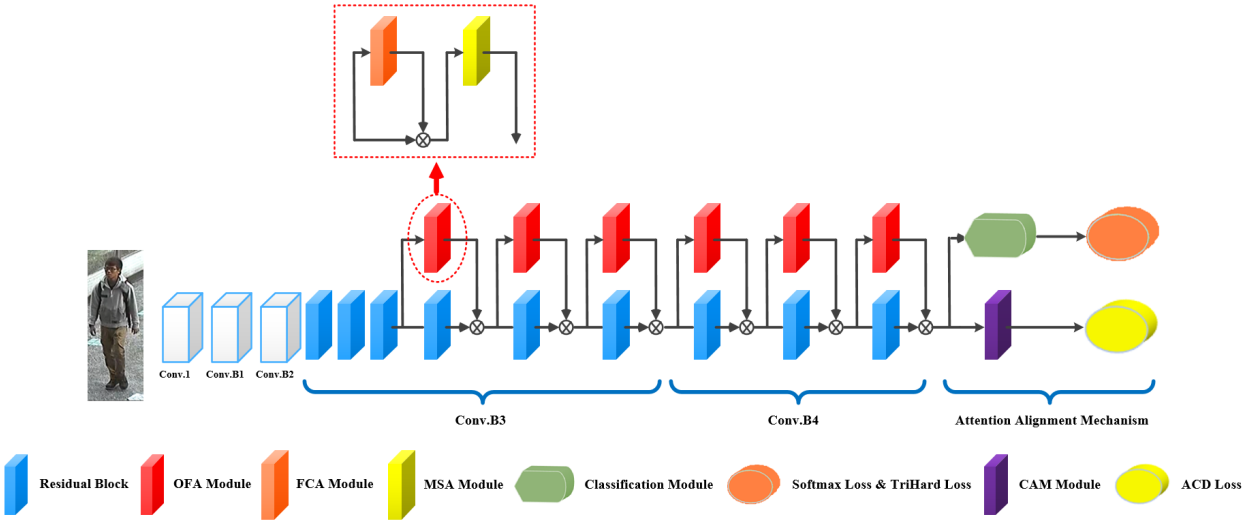


Fig. 2. Overview of the proposed AANet. Note that Conv.1, Conv.B1, Conv.B2, Conv.B3 and Conv.B4 in AANet all come from ResNet-50. The schematic diagram of OFA module is also highlighted in this figure, which consists of FCA module and MSA module.

that attention mechanism is not only used in CNNs [3] [4] [1] [2] [5], but also often applied in Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) [24] to process sequential decision problems [25].

In general, the attention fall into two types: spatial attention [1] [2] and channel attention [5] [21]. More specifically, spatial attention is proposed to identify the salient regions and allocate attention preferentially towards more informative regions. Channel attention is utilized to consider intrinsic interaction among different channels. But it's actually not an optimal choice to simply employ spatial attention or channel attention in attention network, since the former rarely considers intrinsic interaction among different channels whilst the latter fails to allocate attention preferentially towards more useful regions. To this end, [3] [4] [26] successfully combine spatial attention and channel attention together and propose soft attention models, bringing remarkable improvements in performance. In [3], Li *et al.* have come up with a jointly learning attention selection, and further maximize the complementary information of different levels of visual attention subject to RE-ID discriminative learning constraints. Chen *et al.* [4] construct a novel Mixed High-Order Attention Network (MHN) to utilize the complex and high-order statistics attention information.

3 Proposed Approach

In this section, as presented in Fig. 2, we propose an Attention-Aligned Network (AANet) for RE-ID which can be further divided into four parts to elaborate. In the first place, a coherent framework of AANet is presented in Sec. 3.1. After which, the formulations of Omnibearing Foreground-aware Attention (OFA) module and Attention Alignment Mechanism (AAM) are given in Sec. 3.2 and Sec. 3.3 respectively. At last, the classification module which contains loss function is proposed in Sec. 3.4.

3.1 Network Structure

The framework of our proposed AANet is shown in Fig. 2. Clearly, AANet consists of four parts, including Backbone network, OFA module, attention alignment mechanism and classification module. As aforementioned, AANet aims to learn discriminative features and cope with cross-camera inconsistency (e.g. occlusions, and background clutter), thereby two requirements need to be fulfilled in the network design.

Firstly, the proposed network is expected to adopt a reliable, flexible backbone network and design a powerful attention module, so as to extract discriminative features at the output layer. Without loss of generality, we follow most of the previous works and choose the ResNet-50 [10] as backbone network, where the global average pooling layer and the Fully Connected (FC) layer are removed. The convolutional layers of ResNet-50 network can be explicitly divided into five parts: conv.1 (i.e. the 1st layer), conv.B1, conv.B2, conv.B3 and conv.B4. Among them, the last four parts contain 3, 4, 6, 3 residual blocks respectively, each of which is composed of three convolutional layers. To exploit beneficial attention information and enhance discriminative representations, we meticulously propose an Omnibearing Foreground-aware Attention (OFA) module and embed it in the ResNet-50. As presented in Fig.2, the last six residual blocks and OFA modules are placed in a parallel way, therefore biasing the attention of model towards salient foreground regions in heterogeneous spatial scales.

Secondly, an efficient alignment module should be designed to vertically align attention maps among various images of the same pedestrian. Motivated by this idea, we propose a powerful Attention Alignment Mechanism (AAM) to achieve cross-view consistency. It firstly obtains Class Activation Mapping (CAM) [27] from last convolutional layer, then utilizes dual soft threshold to select salient foreground regions and reduce background clutters, and finally performs attention alignment via Cosine distance loss. Main innovations of AANet, including OFA module, AAM and improved TriHard loss, are elaborated in this section.

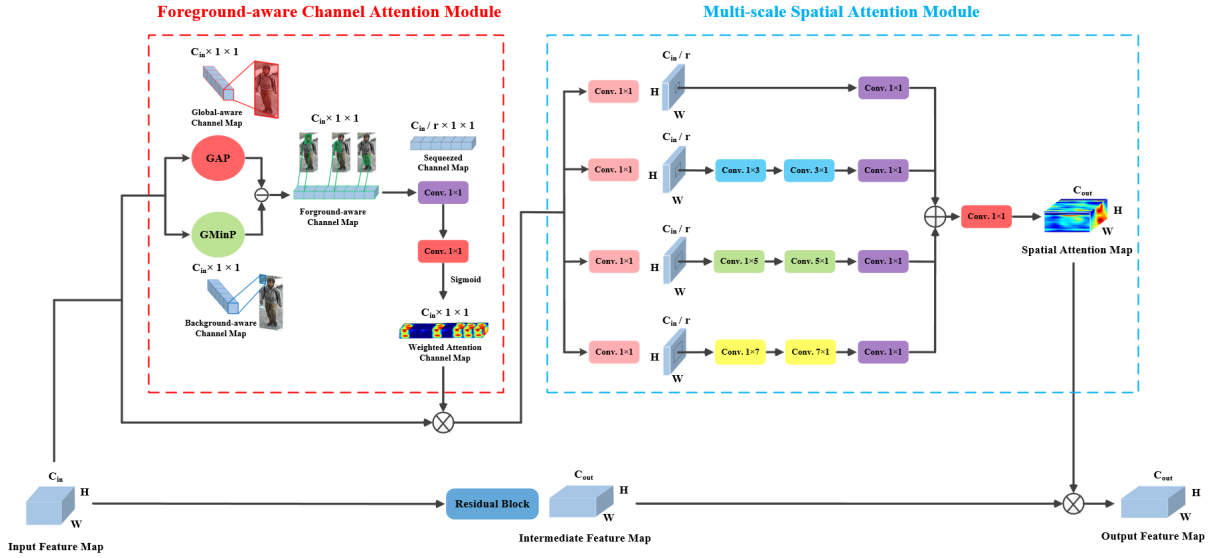


Fig. 3. Detailed schematic diagram of Omnibearing Foreground-aware Attention (OFA) module. This figure shows the exact position and structure details of OFA module when integrated in backbone network. Moreover, the structure of Foreground-aware Channel Attention (FCA) and Multiscale Spatial Attention (MSA) modules are also presented.

3.2 Omnibearing Foreground-aware Attention Module

Existing works [3] [28] [4] show that there is a close relationship between spatial attention and channel attention. It will certainly sacrifice feature map information and compromise the module performance if we merely employ spatial attention or channel attention to the pedestrian image. The soft attention, the parallel connection of spatial and channel attention, has been proved to be effective due to the advantage of fully exploiting both spatial and channel information. However, most existing soft attention approaches merely extract two kinds of attention separately with two independent branches and their mechanism are still coarse-grained and incomplete, therefore being insufficient to provide discriminative feature representations.

To this end, we dedicate to improving attention mechanism so as to facilitate model with learning more features from informative foreground regions. As shown in Fig. 3, Omnibearing Foreground-aware Attention (OFA) module is meticulously proposed to extract more comprehensive and precise attention maps, which connects **Foreground-aware Channel Attention (FCA)** module and **Multiscale Spatial Attention (MSA)** module in series. In each OFA module, multiscale spatial attention maps are obtained on the basis of feature map strengthened by foreground-aware channel attention. This unique serial connection combines the merit of two kinds of attention information, hence gaining an important performance boost.

Formally, we define the input to an OFA module as a 3-D tensor $\boldsymbol{F}_{in} \in \mathbb{R}^{c_{in} \times h \times w}$ where c_{in}, h, and w indicate the number of channel, height and width, respectively. The channel and spatial attention maps obtained from FCA and MSA are denoted by $\boldsymbol{C} \in \mathbb{R}^{c_{in} \times 1 \times 1}$ and $\boldsymbol{S} \in \mathbb{R}^{c_{out} \times h \times w}$, respectively. Then, the saliency weighted attention map learned by OFA module can be written as:

$$\boldsymbol{A} = \boldsymbol{F}_{in} \times \boldsymbol{C} \times \boldsymbol{S} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{c_{out} \times h \times w}$, and $\times$ means tensor multiplication where broadcasted operation is included to make tensors have compatible shapes for arithmetic operations.

Moreover, motivated by the prior work [29] which constructs perception branch and attention branch in the design of network, we intend to model "where" pathway (i.e. attention module) and "what"pathway (i.e. residual module) simultaneously by means of the parallel connection. Specifically, we place OFA modules in parallel with residual blocks retrieved from ResNet-50, whilst traditional RE-ID attention network tends to concatenate them in series. The final output feature strengthened by OFA module can be considered as:

$$\boldsymbol{F}_{out} = f(\boldsymbol{F}_{in}, \{W\}) \otimes \boldsymbol{A} \tag{2}$$

where $\boldsymbol{F}_{out} \in \mathbb{R}^{c_{out} \times h \times w}$ is the output feature map, the function $f(\boldsymbol{F}_{in}, \{W\})$ represents the convolution mapping of a residual block which contains multiple convolutional layers, and $\otimes$ denotes the Hadamard product.

3.2.1 *Foreground-aware Channel Attention*

Traditional channel attention performs a squeeze operation via a simple Global Average Pooling (GAP) , with feature information distributed across the spatial space being equally weighted. Clearly, it's not the optimal scheme to represent the importance of the specific channel. Through GAP layer, all pixels in each feature map have equal influence, which indicates that the obtained channel attention maps can be easily distracted by background clutter and thus fails to provide precise information for inter-channel dependency modelling.

To address this problem, we propose Foreground-aware Channel Attention (FCA) module to guide the attention towards foreground regions which are more helpful for feature representation. In fact, only few studies have formally assessed Global Minimum Pooling (GMinP) operation. In [30], min-pooling is utilized to provide interpretation of

outlierness. Li *et al.* [31] claim that the attention mechanism with GMinP improves the detection performance on small objects. As presented in Fig. 3, a hybrid pooling approach of GAP and GMinP is introduced to aggregate spatial information and alleviate the impact of background clutter.

Here, we denote the channel signatures obtained from GAP and GMinP by $\boldsymbol{C_{avg}}$ and $\boldsymbol{C_{min}}$ respectively. Intuitively, in each feature map, GAP operation treats all pixels equally whilst GMinP operation selects the minimum pixel value. With the training process of neural network, the pixels from background regions in attention map are trending to small value. For this reason, during the training phase, GMinP operation tends to concentrate on the inconspicuous regions which are likely to be background areas. Motivated by this idea, we compute a contrastive channel signature by subtracting $\boldsymbol{C_{min}}$ from $\boldsymbol{C_{avg}}$, which inherits the advantages of GAP and GMinP.

Specifically, it can convey the per-channel filter response from the salient human body regions and demonstrate the robustness to background clutters. In this way, we naturally obtain a foreground-aware channel signature which effectively assigns the body parts with larger weights and the background/occlusion parts with smaller weights, as vividly shown in Fig. 8.

Thereafter, we perform squeeze-expansion operation to employ channel attention by two global convolutional layers of $\{c/r, 1\times1, 1\}$ and $\{c, 1\times1, 1\}$ where the first parameter stands for the number of output channels, second one is the kernel size, and the last one is the stride. In this way, the foreground feature information distributed across the spatial space can be precisely weighted and dynamically compressed into channel attention maps.

3.2.2 Multiscale Spatial Attention

Multiscale method has been extensively adopted in many works [32] [33]. However, existing multiscale methods are mostly applied to learn multiscale feature. To our knowledge, it is unusual to embed multiscale mechanism into spatial attention module in the field of RE-ID. To fill this gap, we propose a new Multiscale Spatial Attention (MSA) module which is capable of exploiting more comprehensive spatial attention with various receptive fields. In our design, we spare no efforts to decrease computational cost and the size of module parameter, making up for the deficiency of multiscale methods.

As shown in Fig. 3, four branches with receptive fields of 1×1 , 3×3 , 5×5 and 7×7 are designed to mine multi-scale spatial attention information. The first step of traditional spatial attention module is often to take a cross-channel average pooling and flatten feature map from $h\times w\times c$ to $h\times w\times 1$, which neglects the discrepancy of spatial attention map among channels. Instead, In each branch of MSA module, the original feature map of $h\times w\times c$ is firstly convoluted by a convolution layer with a kernel in size of 1×1 which compresses the channels to c/r (r is experimentally set to 16). In this way, more inter-channel spatial information is retained.

Then, following the successful practices of [34] which argues that any $n\times n$ convolution can be equivalently replaced by a $1\times n$ convolution followed by a $n\times1$ convolution, we employ this convolution factorization in three branches to reduce computation load and parameters size, e.g. 7×7 convolution filter is replaced by 1×7 and 7×1 convolution layers in series. Afterward, another convolutional layer of $\{c, 1\times1, 1\}$ is applied to extend the feature map in channel dimension. Furthermore, features of heterogeneous spatial scales in four branches are aggregated, followed by a 1×1 convolution where the attention map can be up-sampled to $h\times w\times c'$ and match the channel size of output feature from residual module in backbone network. At last, a sigmoid function is used to normalize the spatial attention map in the range of $[0, 1]$. It's noteworthy that appropriate padding is applied in each convolution layer so as to guarantee the invariant of spatial dimensions of feature maps.

To sum up, OFA module manages to harmonize all attention information from different layers by integrating FCA information and MSA information effectively.

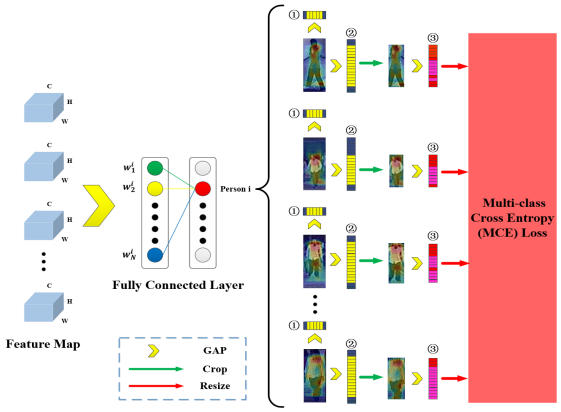


Fig. 4. Illustrations of the Attention Alignment Mechanism. Note that GAP is the abbreviation of Global Average Pooling and (1)(2)(3) represent horizontal, vertical intermediate vectors and final attention vector respectively, which will be elaborated in Sec.3.3.

3.3 Attention Alignment Mechanism

As aforementioned, both backbone network and OFA module can't ensure attention consistency or effectively reduce adverse impact of cross-view variations. To this end, we propose a novel Attention Alignment Mechanism (AAM) which aims to provide explicit guidance towards attention consistency.

In the first place, we localize the attentive regions based on Class Activation Mapping (CAM) [27]. As a powerful and reasonably interpretable technique, CAM is introduced to generate attention maps which represent the salient regions and accurately explains the feature learned by the model. With the help of CAM, "visual attention explanations" for deep-learning-based models is provided which is an intuitive insight and valuable understanding into model mechanism.

Here, we intend to elaborate the first stage of generating CAM as follows. The first stage consists of a Global Average Pooling (GAP) layer and a Fully Connected (FC) layer. The spatial average of the feature map of each unit at the last convolutional layer is firstly obtained from GAP layer. Note that the unit denotes the convolutional filter which is also known as convolutional kernel. More technically, the activation of unit k in the last convolutional layer at spatial

location (x, y) is presented as $\mathbf{A_k}(\mathbf{x}, \mathbf{y})$. In AANet, the last convolutional layer before CAM module is of size 1×1, stride 1, with 2048 units (filters).

Thus, for every unit k, it can be easily obtained that the activation result after GAP layer is $\mathbf{G}_k = \sum_{x,y} \mathbf{A_k}(\mathbf{x}, \mathbf{y})$. Furthermore, for any arbitrary person i, the output of FC layer $\mathbf{F_i}$ can be written as:

$$\begin{aligned} \mathbf{F}_i &= \sum_k w_k^i \mathbf{G_k} \\ &= \sum_k w_k^i \sum_{x,y} \mathbf{A_k}(x, y) \\ &= \sum_{x,y} \sum_k w_k^i \mathbf{A_k}(x, y) \end{aligned} \tag{3}$$

where w_k^i is the weight corresponding to person i for unit k, the value of which determines the importance of $\mathbf{G_k}$ for person i. We define $\mathbf{M}_i$ as the CAM for person i, where each spatial element is given by

$$\mathbf{M_i}(x, y) = \sum_k w_k^i \mathbf{A_k}(x, y) \tag{4}$$

Intuitively, it can be learned that the CAM is actually a weighted linear sum of the presence of these visual patterns at different spatial locations. Later, the output of the class activation map is further normalized as below:

$$\mathbf{M_i^{norm}} = \frac{\mathbf{M}_i - \min(\mathbf{M}_i)}{\max(\mathbf{M}_i) - \min(\mathbf{M}_i)} \tag{5}$$

Secondly, with the purpose of remedying adverse impact of background clutters, we utilize dual soft threshold to crop attention map with respect to height and width, thereby acquiring attention map of human body regions. Given an attention map $\mathbf{M^{norm}}$, we initially utilize the average-pooling operation to compute the response across vertical and horizontal row of pixels, and then obtain two intermediate vectors, $\mathbf{V^{vertical}}$ and $\mathbf{V^{horizontal}}$. Then, two soft thresholds, T_w and T_h, are applied in these two vectors respectively to determine the clipping region of original attention map and produce a cropped attention map $\mathbf{M^{cropped}}$. Afterward, another average-pooling operation is applied across horizontal row of pixels of $\mathbf{M^{cropped}}$, providing us with the final attention vector which is denoted by $\hat{\mathbf{V}}$. Furthermore, all attention vectors are resized to have the same dimensions via interpolation.

In the end, we propose a novel Attention Cosine Distance Loss (ACDL) to learn attention consistency. Cosine similarity is a metric of similarity between two non-zero vectors whilst cosine distance seeks to express the vector dissimilarity. We argue that, compared with Euclidean distance, the cosine distance is advantageous and more suitable for achieving consistency of attention vectors, since Euclidean distance accounts for magnitude while cosine distance does not. For instance, the distance between vector of $(10, 30, 80)$ and $(1, 3, 8)$ is large when calculating with Euclidean distance whereas the result computed by cosine distance is 0. In this case, both two vectors have the same direction and thus can be considered as two consistent vectors. Consequently, when cosine distance is introduced as loss function, the attention maps of same identity among various cameras are learned to have similar distribution. Generally, for arbitrary two vectors $\mathbf{V_1}, \mathbf{V_2}$, we define cosine similarity as:

$$S(\mathbf{V_1}, \mathbf{V_2}) = \frac{\mathbf{V_1}^T \mathbf{V_2}}{||\mathbf{V_1}|| \cdot ||\mathbf{V_2}||} \tag{6}$$

where S stands for cosine similarity and $||\cdot||$ is L2-norm.

Furthermore, for every targeted person i with randomly sampling K images, it can be obtained that

$$\mathcal{L}_{\text{ACD}} = \frac{1}{\binom{K}{2}} \sum_{i=1}^{N} \sum_{a,b}^{K} 1 - S(\hat{\mathbf{V}}_\mathbf{i}^\mathbf{a}, \hat{\mathbf{V}}_\mathbf{i}^\mathbf{b}) \tag{7}$$

where $1 \leq a, b \leq K$, $\binom{K}{2}$ means the number of combinations of K items taking 2 items at a time.

3.4 Classification Module

Inspired by [12], the original feature map $\mathbf{F}$ is further split into 6 local feature maps $(\mathbf{F}_1, \ldots, \mathbf{F}_6)$ in the branch of local features extraction. As presented in Fig.5, in addition to local feature maps, we also obtain holistic feature $\mathbf{F}_7$. Each of feature maps is further fed into a classifier, respectively. Each identity classifier is implemented with a FC layer and a sequential softmax layer.

In terms of the first six branches for extracting local features, during training, given a set of images $\{\mathbf{I_n}\}_{n=1}^{N} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N\}$ encompassing N images, each identity classifier predicts the identity of the input image and is supervised by Softmax Loss:

$$\mathcal{L}_{\text{Softmax}} = -\frac{1}{N} \left(\sum_{z=1}^{6} \sum_{i=1}^{N} \log \frac{e^{y_i^{(z)}}}{\sum_{k=1}^{N_c} e^{y_k^{(z)}}} \right) \tag{8}$$

where z denotes the index of extracted features in six local feature branches, N_c is the number of person labels in a training batch, and $y_i^{(z)}$ represents the prediction of targeted person i from the identity classifier in z-th branch for input image $\mathbf{I}_n$.

As for the global feature branch, we improve the Triplet loss with Hard mining (TriHard) by placing extra constraints on intra-class distance, therefore facilitating model with learning more discriminative features. The original TriHard loss merely attaches importance to increasing the distance between positive groups and negative groups, Yet, it fails to decrease intra-class distance in positive groups effectively and we are dedicated to addressing this problem by making a modification to TriHard Loss function. Worthy of mention is that, during the inference stage for RE-ID, this global feature branch is no longer needed.

Formally, given an anchor point $\mathbf{F_i^a}$, the projection of a positive point $\mathbf{F_i^p}$ of same identity i is pulled closer to the anchor's projection than that of a negative point $\mathbf{F_j^p}$ belonging to person j. Eventually all points of the same class will form a single cluster with the model training. During the training stage, P classes (person identities) are sampled randomly and K images of each identity are then randomly selected as well. Hardest positive and negative samples are

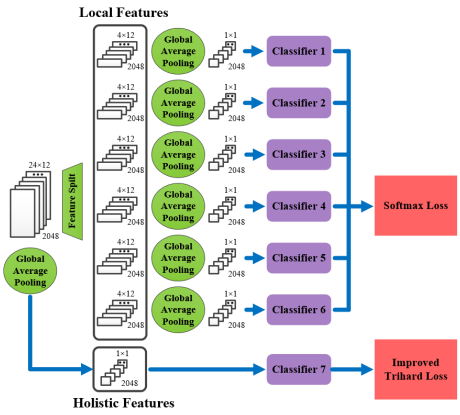


Fig. 5. Structure of Classification Module. It mainly contains two branches which are specifically proposed to obtain local and holistic features.

selected in the improved TriHard Loss within the a batch of $P \times K$ images, which can be written as:

$$\mathcal{L}_{\text{TriHard}} = \frac{1}{PK} \overbrace{\sum_{i=1}^{P} \sum_{a=1}^{K}}^{\text{all anchors}} \overbrace{\max_{p=1\ldots K} D\left(\mathbf{F_i^a}, \mathbf{F_i^p}\right)}^{\text{An Extra constraint on intra-class distance}} + \Big[m + \overbrace{\max_{p=1\ldots K} D\left(\mathbf{F_i^a}, \mathbf{F_i^p}\right) - \min_{\substack{j=1\ldots P \\ n=1\ldots K \\ j\neq i}} D\left(\mathbf{F_i^a}, \mathbf{F_j^n}\right)}^{\text{Pushing apart positive and negative groups}}\Big]_+ \tag{9}$$

where m is margin, $D(\cdot)$ means the Euclidean distance between two features and $[a]_+$ represents $\max(a, 0)$.

From Formula (9), one can observe that the improved TriHard Loss is able to pull positive groups together and push negative groups apart. More importantly, an extra constraint on intra-class distance hopefully realizes the full potential of the previous TriHard Loss, hence further promoting the model's performance.

By combining all the above losses, the final objective function for end-to-end training can be written as minimizing the loss function below:

$$\mathcal{L} = \mathcal{L}_{\text{Softmax}} + \alpha \cdot \mathcal{L}_{\text{TriHard}} + \beta \cdot \mathcal{L}_{\text{ACD}} \tag{10}$$

where α and β are constant parameters. By adjusting the value of parameter α and β to optimal value, the final objective function enables model to obtain the effective information from intra group, inter group and identity label.

On the whole, these three loss items complement each other indispensably with different emphasis. Softmax Loss focuses on the classification of identities and the improved TriHard Loss places more emphasis on optimizing intra and inter-group distance, facilitating model with learning more discriminative features. ACD Loss is an effective loss item to supervise attention consistency among the images of same identity under disjoint cameras. Consequently, the final objective function fully exploits the advantages of both three losses by integrating them effectively, and the desired results are produced naturally.

TABLE 1
The detailed and comprehensive illustrations of Market-1501, DukeMTMC-reID, CUHK03-NP (Labeled) and CUHK03-NP (Detected), datasets, where Duke means DukeMTMC-reID, CUHK03 (L) means CUHK03-NP (Labeled) and CUHK03 (D) means CUHK03-NP (Detected)

Datasets	Training		Testing			
	IDs	Images	Query		Gallery	
			IDs	Images	IDs	Images
Market-1501	751	12,936	750	3,368	750	19,732
Duke	702	16,522	702	2,228	1,110	17,661
CUHK03 (L)	767	7,368	700	1,400	700	5,328
CUHK03 (D)	767	7,365	700	1,400	700	5,332

4 EXPERIMENTS

4.1 Datasets and Settings

In order to verify the superiority of AANet, we tested our proposed model on three mainstream RE-ID datasets which encompass CUHK03-NP [6] [7], Market-1501 [8] and DukeMTMC-reID [9] [10]. Among these datasets, pedestrian images are all captured by non-overlapping surveillance cameras and each image is denoted by a specific label. Comprehensive comparisons on datasets are presented in Table 1. Note that this paper reports the rank-1 accuracy and mean average precision (mAP) for all approaches.

4.2 Implementation Details

Baseline network. For fair comparison, BaseLine (BL) network adopts the same backbone network, i.e. ResNet-50. As opposed to AANet, in BL network, the OFA module is removed and Attention Alignment Mechanism (AAM) is not adopted. Besides, the global feature branch is also excluded and only Softmax Loss is utilized as objective function.

System Settings We implemented AANet model with the Pytorch deep learning framework, including torch 1.0.1, cudnn 6.0.21 and cuda 8.0.61.2. The hardware of the Server contains 12G TITAN XP GPU, 64G memory, Intel Core i7-7820X CPU @ 3.60GHz×16. The Operating System is ubuntu 14.04.

Training. During training, we modify the number of neurons in the fully-connected layer as 767, 751, 702 neurons for Market-1501 [8], DukeMTMC-reID [9] [10] and CUHK03 [6] [7] accordingly. The original input pictures are all resized to 384×192. We manage to adopt an augment strategy for training data by normalization, horizontal flip and random erasing [13]. Based on the results of adequate experiments, we tend to give equal weights to every loss item and the parameters for the final objective function are set to $\alpha = \beta = 1$ accordingly. Considering the trade-off between the computational cost and predictive accuracy, we assign $P = 10$ in TriHard Loss, and $K = 3$ is set in both TriHard Loss and ACD Loss to train our proposed model. The batch size is 30. Our proposed AANet is trained in end-to-end fashion and we use the weights of ResNet-50 pre-trained on ImageNet for fine-tuning. Stochastic Gradient Descent (SGD) is used to optimize the network with initial learning rate of 0.01 and a momentum of 0.9. After 40 epochs, the learning rate of the convolution layers decay from 0.01 to 0.001. The total training epochs are 80, 100, 120 for

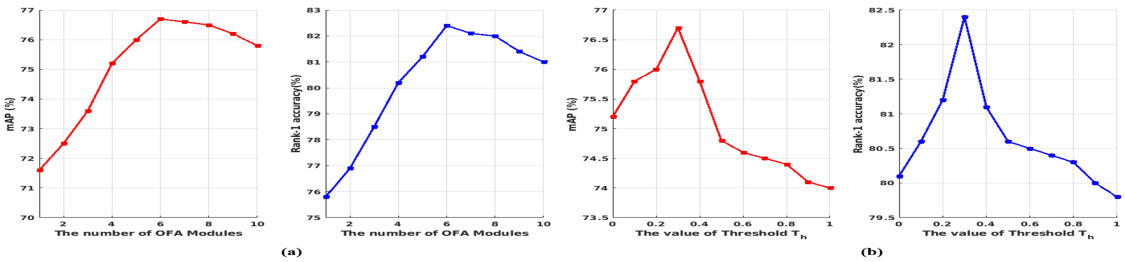


Fig. 6. Parameters analysis. (a) The impact of the number of OFA modules. (b) The impact of the threshold T_h, given $T_w = 0.3$ beforehand.

CUHK03, Market-1501, DukeMTMC-reID respectively. With a NVIDIA TITAN XP GPU and Pytorch platform, training a AANet and BL model on Market-1501 consumes about 260 and 200 minutes, respectively.

Testing. Our proposed AANet classifies by a 2048×6-dimension fused local features extracted from classification module. The cosine distance is applied to calculate the similarity score between two pictures. Our results are all acquired under the single-query mode and compared with reported results without the strategy of re-ranking [7]. As for inference efficiency, we report the testing time per image and total testing time in Table 2. As the results show, AANet is applicable and practical to be implemented in reality scenarios with efficiency.

TABLE 2
Testing time of AANet on the Market-1501, DukeMTMC-ReID, CUHK03-NP (Labeled) and CUHK03-NP (Detected).

Dataset	Total time (s)	Time per image (ms)
Market-1501	252	10.9
DukeMTMC-reID	210	10.6
CUHK03-NP (Labeled)	81.3	12.1
CUHK03-NP (Detected)	81.8	12.2

4.3 Ablation study

In order to demonstrate how much each component contributes to the final performance, various experiments are meticulously designed and conducted on all datasets extensively, as summarized in Table 3.

From the first six rows, it can be easily learned that every component in isolation brings performance gain. Among them, OFA module is arguably the biggest contributor. The improvements brought by OFA module are particularly evident on CUHK03-NP (Labeled). With the performance gains of 14.4% and 15.7% for mAP and rank-1 accuracy, the results convincingly confirm the effectiveness of OFA module. Besides, the second and third contributors are AAM and the improved TriHard loss function respectively. With respect to the mAP and rank-1 accuracy, AAM brings the improvement of 5.8% and 3.3% whilst the improved TriHard loss function gives a further performance boost of 2.5% and 1.9% on CUHK03-NP (Labeled) .

Based on the powerful attention module (i.e. OFA module), we provide further analysis of other two important parts. The seventh row demonstrates the effect of the improved TriHard loss function, which gives the performance gains of 1.2% and 0.9% for mAP and rank-1 accuracy respectively on CUHK03-NP (Labeled), which is quite noticeable. Furthermore, as the results of the eighth row indicates, OFA+AAM gives a further performance boost, suggesting such combination is capable of exploiting the complementary benefits. Based on the comprehensive attention information exploited by OFA module, AAM further provides an explicit guidance towards achieving attention consistency against misalignment. From above results and analysis, we observe that there is a progressive improvement in recognition performance from BL to AANet, and using all components performs best.

4.4 Qualitative Analysis

To provide more analysis of OFA modules, we further carry out some comparative experiments for qualitative evaluation. For a fair comparison, in all comparative experiments, other components in AANet and implementation setting for training/testing still remain unchanged.

In the first place, we evaluate AANet with variant placement of OFA modules. From Conv.B1 to Conv.B4, the OFA modules are placed in parallel with these four different parts of ResNet-50 separately. As shown in Table 4, one can easily observe that when OFA modules are placed in the low-level feature extraction layers (i.e. Conv.B1), the modified AANet has little improvement over the baseline, even slightly worse. Conversely, a significant performance boost is acquired when OFA modules are placed in parallel with high-level feature layers. Then we intend to place OFA module densely, taking similar placing way as DenseNet [35]. Specifically, OFA modules are connected in dense-fashion with each other. For each OFA module, the attention maps of all preceding OFA modules are added as inputs, and its own produced attention maps are fed into all subsequent OFA modules.

Yet, when OFA layers are placed densely (denoted by AANet-DenseOFA), it will inevitably induce high GPU memory consumption because of the reuse of feature. Given that the value of batchsize is limited by available GPU memory, the model performance decrease because of the reduction of batchsize. Then we amount to ensure sufficient GPU memory and assign the same value of batchsize as AANet, which is denoted as AANet-DenseOFA-2 in Table 4. The performance has been improved slightly at the cost of extra huge GPU memory, which is not satisfactory.

Moreover, we carry out the experiment of concatenating OFA module and residual block in series connection, which is denoted as AANet-2. As the results indicate, AANet outperforms AANet-2, which validates the advantage of our

TABLE 3
Quantitative ablation study for revealing the impact of each imponent of AANet on the final performance. BL : BaseLine model; OFA : Omnibearing Foreground-aware Attention; FCA : Foreground-aware Channel Attention; MSA : Multiscale Spatial Attention; AAM: Attention Alignment Mechanism; TriHard: Improved Triplet loss with Hard mining.

BL	OFA		AAM	Improved TriHard	Market-1501		Duke		CUHK03(L)		CUHK03(D)	
	FCA	MSA			r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
✓					92.6	79.9	83.0	71.4	63.1	59.4	59.6	54.2
✓				✓	93.1	80.5	83.7	72.0	65.7	61.5	60.3	54.9
✓			✓		94.3	81.8	84.4	72.5	69.0	62.9	65.5	60.2
✓	✓				94.8	82.9	86.0	74.8	73.1	68.1	67.1	61.0
✓		✓			94.7	83.2	86.6	75.2	75.4	69.1	70.8	65.1
✓	✓	✓			95.1	86.7	87.8	76.6	78.9	74.0	73.1	66.7
✓	✓	✓		✓	95.4	87.0	88.3	77.0	79.8	75.2	74.0	67.8
✓	✓	✓	✓		95.8	87.3	89.1	79.2	81.1	76.1	76.2	69.4
✓	✓	✓	✓	✓	**96.1**	**87.5**	**90.2**	**79.5**	**82.4**	**76.7**	**77.2**	**70.5**

method and convincingly proves that the parallel connection produces the better effect to the performance. Besides, we attempt to freeze the weights of the backbone model which is pre-trained on ImageNet whilst other components of AANet are trained on RE-ID dataset. For convenience, we denote the AANet with such setting as AANet-Freeze. As shown in Table 4, it turns out that AANet-Freeze is prone to have very poor performance, demonstrating the significance of backbone model.

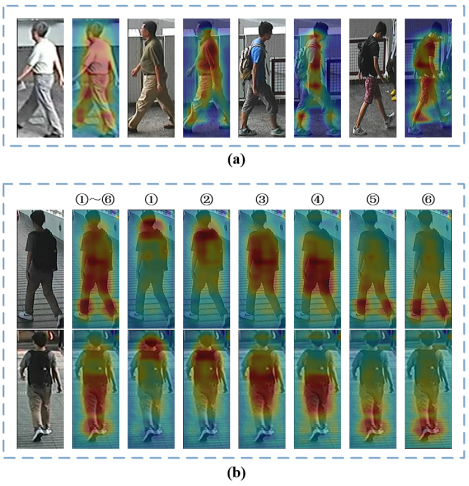


Fig. 7. Visualizations of the attention maps in our proposed AANet.

TABLE 4
Comparison of rank and mAP for Qualitative Analysis of OFA modules on CUHK03(Labeled).

Method	r=1	r=5	r=10	mAP
AANet (without OFA modules)	70.1	84.0	90.4	64.9
AANet-Conv.B1	70.3	84.5	91.0	64.6
AANet-Conv.B2	71.7	87.1	91.4	68.2
AANet-Conv.B3	80.3	90.9	94.1	74.8
AANet-Conv.B4	**80.7**	**91.4**	**94.9**	**75.2**
AANet-DenseOFA	74.1	86.6	91.0	68.7
AANet-DenseOFA-2	**82.7**	**92.3**	**95.9**	**77.0**
AANet-Freeze	47.1	72.4	81.4	43.6
AANet-2	79.6	91.0	94.7	74.4
AANet	**82.4**	**91.9**	**95.3**	**76.7**

4.5 Parameters Analysis

4.5.1 The Number of OFA Modules

To provide a quantitative analysis of the number of OFA modules, we conduct some comparative experiments on CUHK03 (Labeled).

From Fig.6(a), it can be learned that as the number of OFA modules increases, the performance of model gradually improves at first. However, the accuracy does not always increase. As the result indicates, the optimal number of OFA modules is 6, which is actually a reasonable result. If we utilize few OFA modules, feature representations extracted from model is highly likely to be compromised as the beneficial attention information is unexploited. In another case, when we stack too many OFA modules in model, the performance is also compromised with a slight decrease. And it tends to bring at least two main drawbacks but not limited to: (1) It will inevitably introduce considerable computational overheads and make AANet complex and hard to tune. (2) Since the value of batchsize is limited by available GPU memory, the performance can be easily affected with the reduction of batchsize. Suppose we do assume the GPU memory is sufficient, it will only bring little improvement when OFA modules are intergraded into front stages of the backbone network.

The manipulation of the number of OFA modules should take into consideration the trade-off between the computational cost and efficacy. To further demonstrate this trade-off relationship, we provide a quantitative investigation of the optimal number of OFA modules, which is presented in Table 5. As the number of OFA modules increases from 6 to 9, it brings small improvements of 0.1% in rank-1 and 0.2% in mAP, and meanwhile demands an extra 3.6M memory complexity and 0.7G computational cost. Similarly, one can easily learn that when OFA modules are placed in parallel with all convolutional blocks, the contribution of the increased OFA modules is also trivial. Compared with other schemes, the configuration of six OFA modules is more cost-effective as it brings a relatively small memory-

TABLE 5
Comparisons of model size, computation complexity and performance with varying quantity of OFA modules on CUHK03(Labeled). PN: Parameter Number. FLOPs: the number of FLoating-point OPerations;

Method	The number of OFA modules	PN (M)	FLOPs (G)	r=1	mAP
AANet (without OFA modules)	0	29.8	9.2	70.1	64.9
AANet-Conv.B4	3	37.2	10.9	80.7	75.2
AANet	6	39.6	11.4	82.4	76.7
AANet-Conv.B4+B3	9	43.2	12.1	82.5	76.9
AANet-Conv.B4+B3+B2	13	43.9	13.0	82.5	77.2
AANet-Conv.B4+B3+B2+B1	16	44.1	13.6	82.6	77.3

TABLE 6
Comparison of total training and testing time with different number of OFA modules on CUHK03(Labeled). Note that total training epochs on this dataset are 80.

The number of OFA modules	Total Training time (s)	Total Testing time (s)
1	8428	79.1
2	9153	79.3
3	9779	79.6
4	10458	79.9
5	11091	80.7
6	11553	81.3

TABLE 7
Comparison of rank accuracy and mAP with RR for different values of α and β on CUHK03(Labeled).

α	β	r=1	r=5	r=10	mAP
0.2	1.0	81.3	90.0	94.7	76.2
0.4	1.0	81.5	90.5	95.2	76.3
0.6	1.0	82.0	91.1	95.3	76.4
0.8	1.0	82.3	91.7	95.2	76.6
1.0	**1.0**	**82.4**	**91.9**	**95.3**	**76.7**
1.0	0.2	80.0	91.3	94.9	75.4
1.0	0.4	80.6	91.4	95.4	75.8
1.0	0.6	81.0	91.4	95.5	76.2
1.0	0.8	82.3	91.6	95.7	76.5
1.0	**1.0**	**82.4**	**91.9**	**95.3**	**76.7**

computational cost and a considerable performance gain.

Moreover, we also report the training and testing time on CUHK03(Labeled) as well, which is presented in Table 6. With the increase of the number of OFA modules, the time for training and testing increases as well.

4.5.2 The value of dual soft threshold T_w and T_h

As aforementioned in Sec.3.3, these two hyper-parameters are of vital importance to AAM, as they determine the clipping region of original attention map given the value of vertical and horizontal row of pixels in attention maps.

To obtain the optimal values, different values of T_w and T_h are tested extensively. In fact, we firstly determine the value of T_w based on the intuitive display of attention maps and is assigned to the optimal value of 0.3. Then we amount to find the optimal value of T_h by conducting comparative experiments under the same experiment setting. Specifically, we take values of T_h in an interval of 0.1 within the range from 0 to 1. As presented in Fig.6(b), the trend of the change is vividly revealed and we find that the model performance reaches at the peak when $T_h = 0.3$.

With optimal value of T_w and T_h, it can ensure the cropped attention map contains the least information from background clutters and facilitate model with mining more attention information from informative human regions.

4.5.3 Effect of α and β in Loss Function

Here, we conduct experiments with different value of α and β in formula (10) to evaluate their effects on model performance and then analyze the results.

To obtain the optimal values, different values of α and β are tested extensively. Note that we take values of α and β in an interval of 0.2 within the range from 0 to 1. After conducting a large number of experiments, we find that the model performance reaches at the peak when α and β are at specific values. Here, for a brief and clear presentation, we choose the same interval for two parameters in Table 7 respectively, which effectively reveals the trend of the change. From Table 7, we observe that when $\alpha = 1.0$ and $\beta = 1.0$, the model gains the best performance. With optimal value of α and β, the joint loss function is capable of making best use of all three loss items, hence facilitating model with learning more discriminative information from intra group, inter group and identity label.

4.6 Visualizations of Attention Mechanism

Robustness to Background Clutters. The attention maps acquired from our proposed AANet are shown in Fig.7(a). Clearly, with the proposed OFA module, we obtain more discriminative attention maps which attaches great importance to human body parts. The visualizations in Fig.7 fully demonstrate that features learned by AANet are robust to background clusters.

Besides, as illustrated in Sec.3.4, the output feature map $\mathbf{F}$ is further split into 6 local feature maps in classification module. Correspondingly, we further evaluate the attention maps obtained from local feature maps, which are clearly presented in Fig.7(b). The results indicate that local features not only explicitly attend to different human body regions from top to bottom, but they also show robustness to the background clutters.

Effectiveness of Attention Alignment Mechanism. Some representative attention maps of input images are shown in Fig. 8, in which the images represent the same person under different disjoint camera views. The results indicate that the obtained attention maps are robust to misalignment as well as background clutters, compellingly validating the effectiveness of our proposed OFA module and AAM. Moreover, the effectiveness of AAM also reflect

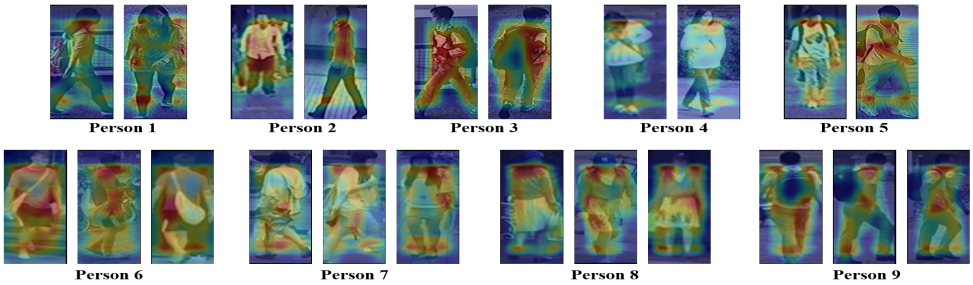


Fig. 8. Demonstrating the efficacy of the proposed OFA module and AAM. The original images in first row (i.e. from Person 1 to Person5) are retrieved from CUHK03-NP, and in the second row (i.e. from Person 6 to Person 9) are retrieved from Market-1501.

TABLE 8
Rank and mAP comparison of AANet and BL on the CUHK03-NP (Labeled), CUHK03-NP (Detected), Market-1501, and DukeMTMC-ReID. Note that BL stands for BaseLine. Note that all experiments are run for 5 times with different random seeds.

Method	Dataset							
	Market-1501		DukeMTMC-reID		CUHK03-NP (Labeled)		CUHK03-NP (Detected)	
	r=1	mAP	r=1	mAP	r=1	mAP	r=1	mAP
BL	92.6±0.4	79.9±0.2	83.0±0.3	71.4±0.2	63.1±0.3	59.4±0.4	59.6±0.3	54.2±0.5
AANet	**96.1**±0.3	**88.6**±0.2	**90.2**±0.1	**79.5**±0.1	**82.4**±0.3	**76.7**±0.2	**77.2**±0.3	**70.5**±0.2

on the consistence of vertical space distribution among the attention maps of same identity. In brief, AAM provides a practicable scheme for addressing cross-camera inconsistency problem, and has proven to be beneficial to learn view-invariant and background-invariant features for Re-ID.

4.7 Performance Comparison

4.7.1 Comparison with BaseLine Model

We firstly compare the rank-1 and mAP between AANet and BaseLine (BL) on three mainstream RE-ID datasets, which is presented in Table 8. Compared to the BaseLine, our network gets significantly promoted in both rank-1 and mAP accuracy on all three datasets by a large margin. The increment of performance has also been presented in Table 8. The comparative results compellingly validates the superiority of our proposed approaches.

TABLE 9
Comparison of rank and mAP for various Attention methods on CUHK03(Labeled).

Method	r=1	r=5	r=10	mAP
BL+CA	72.3	85.2	91.7	67.1
BL+SA	71.0	84.5	90.9	65.6
BL+SOA	74.2	86.6	91.4	68.2
BL+HA	74.0	86.5	91.2	68.0
BL+3DA	77.6	89.9	93.8	73.1
BL+OFA	**78.9**	**90.5**	**94.6**	**74.0**
AANet-CA	75.6	87.9	92.8	68.9
AANet-SA	74.4	86.1	92.2	68.4
AANet-SOA	77.8	89.8	93.9	72.9
AANet-HA	76.7	88.6	93.0	72.2
AANet-3DA	81.0	90.2	94.6	75.4
AANet	**82.4**	**91.9**	**95.3**	**76.7**

4.7.2 Comparison with Existing Attention Modules

In order to validate the effectiveness of OFA module, we carefully design and conduct two sets of comparative experiments. In the first set, the BaseLine (BL) network is integrated with traditional attention modules, including Channel Attention (CA), Spatial Attention (SA) and SOft Attention (SOA). Furthermore, we also compare our OFA modules with other off-the-shelf attention modules containing Harmonious Attention (HA) module [3] and 3DA module [36], as presented in the first half of Table 9. Intuitively, OFA module gains better performance than any other comparative attention module, with a lead of 1.3% in rank-1 and 0.9% in mAP over the second one. In the second set, for fair comparison, we merely replace the OFA module with other attention module and maintain other components and experimental settings unchanged. As the results in Table 9 indicate, AANet still outperform any other comparative model. Compared with the AANet-3DA, AANet gains the accuracy improvement of 1.4% in rank-1 and 1.3% in mAP, which further confirms the superiority of OFA module.

In addition, based on Table 4 and Table 9, we observe that BL+FCA outperforms BL+CA, which proves the superiority of FCA module and also gives an indirect evidence to prove the effectiveness of GMinP operation.

TABLE 10
Comparison of rank and mAP for Improved TriHard Loss and other ranking loss on CUHK03(Labeled)

Method	r=1	r=5	r=10	mAP
Original TriHard Loss	81.9	91.5	94.9	76.1
Quadruplet Loss	81.6	91.2	94.5	75.8
Angular Loss	82.1	**92.0**	**95.2**	76.4
Improved TriHard Loss	**82.4**	91.9	**95.2**	**76.7**

4.7.3 Comparison with other ranking loss

To compare the effects between the improved Trihard loss and other ranking loss functions, we utilize them to train AANet respectively, along with softmax loss. Experimental

TABLE 11
Comparison results of rank and mAP on Market-1501. '-' indicates not reported. Numbers in bold indicate the best performance and underscored ones are the second best.

Method	r=1	r=5	r=10	mAP
PABR [37] (ECCV-2018)	88.8	95.6	97.3	73.9
MLFN [38] (CVPR-2018)	90.0	-	-	74.3
HA-CNN [3] (CVPR-2018)	91.2	-	-	75.7
KPM [39] (CVPR-2018)	90.1	96.7	97.9	75.3
PCB+RPP [12] (ECCV-2018)	93.8	97.5	98.5	81.6
Mancs [28] (ECCV-2018)	93.1	-	-	82.3
SPReID [40] (CVPR-2018)	93.7	97.6	98.4	83.4
DNN-CRF [41] (CVPR-2018)	93.5	97.7	-	81.6
DATRL-ReID [42] (TCSVT-2019)	94.4	98.1	98.8	81.5
MHN-6 (PCB) [4] (ICCV-2019)	95.1	98.1	98.9	85.0
3DTANet [36] (TCSVT-2020)	95.3	98.1	98.8	86.9
RN [43] (AAAI-2020)	95.2	-	-	**88.9**
SCSN [44] (CVPR-2020)	95.7	-	-	88.5
RGA-SC [45] (CVPR-2020)	**96.1**	-	-	88.4
AANet	**96.1**	**98.7**	**99.2**	88.6

TABLE 12
Comparison results of rank and mAP on DukeMTMC-reID.

Method	r=1	r=5	r=10	mAP
DaRe(De) [46] (CVPR-2018)	74.5	-	-	56.3
HA-CNN [3] (CVPR-2018)	80.5	-	-	63.8
MLFN [38] (CVPR-2018)	81.0	-	-	62.8
KPM [39] (CVPR-2018)	80.3	89.5	91.9	63.2
PABR [37] (ECCV-2018)	82.1	90.2	92.7	64.2
PCB+RPP [12] (ECCV-2018)	83.3	90.5	92.5	69.2
DNN-CRF [41] (CVPR-2018)	84.9	92.3	-	69.5
Mancs [28] (ECCV-2018)	84.9	-	-	71.8
SPReID [40] (CVPR-2018)	86.0	93.0	94.5	73.3
DATRL-ReID [42] (TCSVT-2019)	86.3	93.1	95.1	72.9
MHN-6 (PCB) [4] (ICCV-2018)	89.1	94.6	96.2	77.2
3DTANet [36] (TCSVT-2020)	89.9	94.4	95.6	78.4
RN [43] (AAAI-2020)	89.7	-	-	78.6
SCSN [44](CVPR-2020)	**91.0**	-	-	79.0
AANet	90.2	**94.7**	**96.3**	**79.5**

results are presented in Table 10 from which we obviously learn that the improved Trihard loss is superior to the original one. Besides, we notice that model with quadruplet loss [15] has poorer performance compared with original TriHard loss. When compared with the angular loss [47], our improved TriHard loss still obtains better experimental results. In conclusion, the improved Trihard loss function places extra constraints on the intra-class distance and gains better performance.

4.7.4 Comparison with State-of-the-art

Evaluations on Market-1501. We firstly evaluate AANet on Market-1501, which is one of the largest RE-ID datasets. As shown in Table 11, AANet achieves competitive performance compared with existing state-of-the-art methods. Specifically, We report the accuracy of 96.1% in rank-1 and 88.6% in mAP on Market-1501, which validates the superiority of AANet and further verifies the importance of the combination of attention mechanism and attention alignment in performing RE-ID.

Evaluations on DukeMTMC-reID. AANet is also evaluated on DukeMTMC-reID. Similar to Market-1501, it also has a large amount of pedestrian images captured in campus, but meanwhile the images in this dataset has more occlusion and complex background. One can obviously learn from Table 12 that AANet outperforms most compared methods on DukeMTMC-reID, especially leading by a competitive margin of 0.5% in mAP. It further suggests that the proposed OFA manages to exploit heterogeneous-scale spatial attention and foreground-aware channel attention, therefore extracting discriminative features for RE-ID. More importantly, it also confirms that AAM can effectively guide the attention of same identity in various cameras towards alignment, hence facilitating model with learning view-invariant and background-invariant features.

TABLE 13
Comparison results of rank and mAP on CUHK03 (Labeled).

Method	CUHK03(Labeled)			
	r=1	r=5	r=10	mAP
HA-CNN [3] (CVPR-2018)	44.4	-	-	41.0
MLFN [38] (CVPR-2018)	54.7	-	-	49.2
DaRe(De) [46] (CVPR-2018)	56.4	-	-	52.2
DaRe(De)+RE [46] (CVPR-2018)	66.1	-	-	61.6
Mancs [28] (ECCV-2018)	69.0	-	-	63.9
DATRL-ReID [42] (TCSVT-2019)	69.9	85.8	91.9	64.7
MHN-6 (PCB) [4] (ICCV-2019)	77.2	-	-	72.4
RN [43] (AAAI-2020)	77.9	-	-	75.6
3DTANet [36] (TCSVT-2020)	80.2	91.8	95.1	75.2
RGA-SC [45](CVPR-2020)	81.1	-	-	77.4
SCSN [44](CVPR-2020)	**86.8**	-	-	**84.0**
AANet	82.4	**91.9**	**95.3**	76.7

TABLE 14
Comparison results of rank and mAP on CUHK03 (Detected).

Method	CUHK03(Detected)			
	r=1	r=5	r=10	mAP
HA-CNN [3] (CVPR-2018)	41.7	-	-	38.6
MLFN [38] (CVPR-2018)	52.8	-	-	47.8
DaRe(De) [46] (CVPR-2018)	54.3	-	-	50.1
DaRe(De)+RE [46] (CVPR-2018)	63.3	-	-	59.0
PCB+RPP [12] (ECCV-2018)	63.7	80.6	86.9	57.5
Mancs [28] (ECCV-2018)	65.5	-	-	60.5
DATRL-ReID [42] (TCSVT-2019)	66.4	83.1	89.6	60.6
MHN-6 (PCB) [4] (ICCV-2019)	71.7	-	-	65.4
RN [43] (AAAI-2020)	74.4	-	-	69.6
3DTANet [36] (TCSVT-2020)	75.2	88.6	92.3	68.9
SCSN [44](CVPR-2020)	**86.8**	-	-	**84.0**
AANet	77.2	**89.2**	**92.8**	70.5

Evaluations on CUHK03. On the CUHK03 dataset, we compare our proposed AANet with the state-of-the-art methods under both manually labeled and automatically detected settings in Table 13 and Table 14 respectively. Experimental results show that AANet can achieve competitive performance on both two settings with the 82.4% and 76.7% accuracy in rank-1 and mAP on CUHK03 (Labeled) as well as 77.2% and 70.5% on CUHK03 (Detected).

Discussion. To verify the superiority of our method, we firstly compare AANet with the other existing attention models, including Mancs [28], HA-CNN [3], MHN-6 [4], 3DTANet [36] and SCSN [44], on three RE-ID datasets. According to the results from Table 11 to Table 14, we observe that our proposed AANet achieves competitive performance on all datasets and occupies a relatively leading position on Market-1501 and DukeMTMC-reID.

As a matter of fact, AANet is distinctive from existing attention networks which can be mainly concluded in following two aspects: (1) **Improvement of attention mechanism.** Our proposed OFA module is superior to other attention modules and the comparative results are presented in Sec.4.7.2. With exploiting heterogeneous-scale spatial attention and more precise channel attention, OFA module can explicitly guide model to learn more discriminative features. (2) **Attention Alignment Mechanism.** All of these aforementioned attention modules fail to cope with the cross-view inconsistency problem, with performance being affected by large viewpoint variations. Instead, as vividly shown in Fig.8, our adopted AAM can perform attention alignment, which is crucial for learning view-invariant and background-invariant features.

Here, we specifically provide a comparative analysis between AANet and SCSN [44]. SCSN is dedicated to solving the problems of how to extract discriminative features and how to integrate these features whilst AANet places more emphasis on the pressing problem of large viewpoint variations (i.e. cross-camera inconsistency). Notice that latest attention model, SCSN, gains relatively high performance on three RE-ID datasets. By virtue of a cascaded suppression strategy, SCSN attaches great importance to mining diverse potential useful features that be masked by the other salient features stage-by-stage. However, it also needs to be pointed out that the non-local multistage feature fusion, the crucial component of SCSN, increases memory requirements and creates high computational overheads. More importantly, attention alignment among cross-view images of same identity is neglected in the design of attention modules.

Besides, we also report that AANet outperforms other compared methods, including the methods of calculating group consistent similarity [41], addressing misalignment problems [12] [39] [37] [48], adopting deep supervision [46], integrating human semantic parsing [40] and other innovative methods [38] [42]. The competitive results further support the claim that AANet is entirely feasible and effective, with a great potential and broad prospects of development.

5 Conclusion

This paper proposes an Attention-Aligned Network (AANet) for Person Re-Identification. AANet aims to address the challenges of the Re-ID, including cross-camera inconsistency problem and highly similar background problem, and subtle individual discrepancy problem. We firstly leverage a new Omnibearing Foreground-aware Attention (OFA) module to learn background-invariant and discriminative feature representations. Then, we adopt a novel Attention Alignment Mechanism to directly solve misalignment problem as well as cope with significant viewvariations. Furthermore, we improve the triplet loss with hard mining to further optimize intra and inter group distance, facilitating model with learning more discriminative features. Our proposed AANet outperforms most of the state-of-the-art methods on three large RE-ID datasets. Extensive comparative experiments and ablation analysis of our framework fully validate the effectiveness of AANet as well as every individual component.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62076262, 61673402, 61273270, 60802069), the Natural Science Foundation of Guangdong Province (2017A030311029).

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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*. IEEE, 2017, pp. 3652–3661.

[8] 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.

[9] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 17–35.

[10] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3754–3762.

[11] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2590–2600.

[12] 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*, 2018, pp. 480–496.

[13] 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*, pp. 1–10, 2017.

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

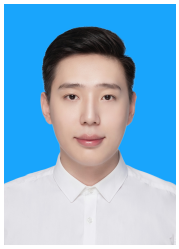
[15] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp. 1320–1329.

[16] H. Chen, Y. Wang, Y. Shi, K. Yan, M. Geng, Y. Tian, and T. Xiang, "Deep transfer learning for person re-identification," in *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, Sep. 2018, pp. 1–5.

[17] S. Li, M. Shao, and Y. Fu, "Person re-identification by cross-view multi-level dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2963–2977, Dec 2018.

[18] M. Li, X. Zhu, and S. Gong, "Unsupervised tracklet person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[19] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.
[20] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 483–499.
[21] 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.
[22] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in *Proceedings of the Association for the Advance of Artificial Intelligence*, 2017, pp. 4075–4081.
[23] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1487–1500, 2018.
[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[25] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual qa," in *Advances in Neural Information Processing Systems*, 2016, pp. 361–369.
[26] Y. Huang, S. Lian, H. Hu, D. Chen, and T. Su, "Multiscale omnibearing attention networks for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–14, 2020.
[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2016, pp. 2921–2929.
[28] 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.
[29] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," *Computer Vision and Pattern Recognition*, pp. 10 705–10 714, 2019.
[30] J. Kauffmann, K.-R. Mller, and G. Montavon, "Towards explaining anomalies: A deep taylor decomposition of one-class models," *Pattern Recognition*, vol. 101, p. 107198, 2020.
[31] Q. Li, N. Guo, X. Ye, D. Fan, and Z. Tang, "Pixel-semantic revising of position: One-stage object detector with shared encoder-decoder," in *The 27th International Conference on Neural Information Processing (ICONIP2020)*, 2020, pp. 1–13.
[32] S. Gao and X. Zhuang, "Multi-scale deep neural networks for real image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1–10.
[33] R. Liao, Z. Zhao, R. Urtasun, and R. S. Zemel, "Lanczosnet: Multi-scale deep graph convolutional networks," pp. 1–18, 2019.
[34] 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.
[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.
[36] Y. Huang, S. Lian, S. Zhang, H. Hu, D. Chen, and T. Su, "Three-dimension transmissible attention network for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–14, 2020.
[37] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–419.
[38] 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.
[39] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6886–6895.
[40] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.
[41] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8649–8658.
[42] Y. Huang, Y. Huang, H. Hu, D. Chen, and T. Su, "Deeply associative two-stage representations learning based on labels interval extension loss and group loss for person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–14, 2019.
[43] H. Park and B. Ham, "Relation network for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
[44] X. Chen, C. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, and Y. Yang, "Salience-guided cascaded suppression network for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[45] 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 (CVPR)*, June 2020.
[46] 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.
[47] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2612–2620.
[48] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5735–5744.



Sicheng Lian received the B.E. degree from Sun Yat-sen University in 2019. He is currently pursuing the M.E. degrees in School of Electronics and Information Technology at Sun Yat-Sen University. His research interests encompass person re-identification, object detection, and image processing.

Weitao Jiang received the B.E. degree in electronic information science and technology from Dalian Maritime University, Dalian, China, in 2019. He is currently pursuing the M.E. degree in information and communication engineering with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China. His current research interests include computer vision and image captioning.

Haifeng Hu received the Ph.D. degree from Sun Yat-sen University in 2004, and now he is a professor of School of Electronics and Information Engineering at Sun Yat-sen University. His research interests are in computer vision, pattern recognition, image processing and neural computation. He has published about 140 papers since 2000.