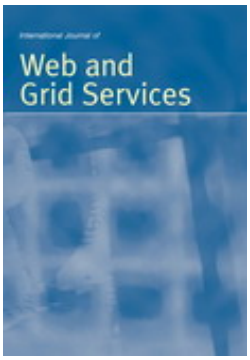


International Journal of Web and Grid Services

ISSN online: 1741-1114 - ISSN print: 1741-1106
https://www.inderscience.com/ijwgs

Attention-based mechanism and feature fusion network for person re-identification

Mingshou An, Yunchuan He, Hye-Youn Lim, Dae-Seong Kang

DOI: 10.1504/IJWGS.2024.10062990

Article History:

Received:	04 September 2023
Last revised:	19 November 2023
Accepted:	13 December 2023
Published online:	25 March 2024

Attention-based mechanism and feature fusion network for person re-identification

Mingshou An and Yunchuan He

School of Computer Science and Engineering,
Xi'an Technological University,
Xi'an, China
Email: anmingshou@xatu.edu.cn
Email: heyunchuan@163.com

Hye-Youn Lim and Dae-Seong Kang*

Department of Electronics Engineering,
Dong-A University,
Busan, South Korea
Email: hylim@dau.ac.kr
Email: dskang@dau.ac.kr
*Corresponding author

Abstract: For the problem that person features cannot be sufficiently extracted in person re-identification, a person re-identification model based on attention mechanism is proposed. Firstly, person features are extracted using a hybrid network combining transformer's core multi-headed self-attentive module with the convolutional neural network ResNet50-IBN-a. Secondly, an efficient channel attention mechanism ECANet is embedded to make the model of this paper more focused on the key information in the person foreground. Finally, fusing the mid-level and high-level features in the model can avoid some discriminative features loss. The experimental results show that the provide model achieves 94.8% rank-1 and 84.5% mAP on the Market-1501 dataset; achieves 84.9% rank-1 and 65.9% mAP on the DukeMTMC-reID dataset; and achieves 40.3% rank-1 and 33.3% mAP on the Occluded-Duke MTMC dataset. Our proposed model compares well with some of the existing person re-identification models on these datasets mentioned above.

Keywords: attention mechanism; person re-identification; feature fusion; convolutional neural network; occluded person detection.

Reference to this paper should be made as follows: An, M., He, Y., Lim, H-Y. and Kang, D-S. (2024) 'Attention-based mechanism and feature fusion network for person re-identification', *Int. J. Web and Grid Services*, Vol. 20, No. 1, pp.74–92.

Biographical notes: Mingshou An received his PhD in Electronics Engineering at Dong-A University. He is a Lecturer at the School of Computer Science and Engineering, Xi'an Technological University. His research interests include issues related to artificial intelligence, pattern recognition, deep learning, and industrial measurement. He is author of a great deal of research studies published at national and international journals, conference proceedings.

Yunchuan He received his Master's in Computer Science and Engineering at Xi'an Technological University. His research interests include issues related to artificial intelligence, pattern recognition, deep learning, and industrial measurement. He is author of a great deal of research studies published at national and international journals, conference proceedings.

Hye-Youn Lim received her PhD in Electronics Engineering at Dong-A University. She is a Lecturer at the Electronics Engineering, Dong-A University. Her research interests include issues related to artificial intelligence, pattern recognition, and deep learning. She is author of a great deal of research studies published at national and international journals, conference proceedings.

Dae-Seong Kang received his PhD in Electrical Engineering at Texas A&M University. He is a Professor at the Electronics Engineering, Dong-A University. His research interests include issues related to artificial intelligence, pattern recognition, and deep learning. He is author of a great deal of research studies published at national and international journals, conference proceedings.

This paper is a revised and expanded version of a paper entitled 'Fusion self-attention feature clustering mechanism network for person ReID' presented at The 13th International Conference on Frontier Computing (FC 2023), Tokyo, Japan, 10–14 July 2023.

1 Introduction

Currently many researchers study on person re-identification (Zajdel et al., 2005), but there are still some obstacles that need to be addressed in practical applications. They include significant changes in light intensity under different cameras, changes in image resolution caused by camera shake, and the possibility that persons under the camera may be obstructed by vehicles, signs, road signs, umbrellas, and other obstacles, making it impossible to obtain a complete person image. The overall flowchart for person re-identification is shown in Figure 1.

Person re-identification is mainly divided into two steps: feature extraction and metric learning. The first step is to extract discriminative features from query images through manual or deep learning methods; the second step is to calculate the similarity of extracted features between query images in the previous step and the gallery images in the database. The traditional methods mainly use traditional machine learning to manually extract low-level visual features, such as colour histograms (RGB) (Koestinger et al., 2012), texture features (Gabor) (Zheng et al., 2012), local features (SIFT) (Lowe, 1999), etc. Metric learning refers to calculating the distance or similarity between features in the feature space to make the intra-class distance smaller and the inter-class distance larger. The common metric learning methods include Marxian distance (Martinel et al., 2015), explicitly weighted metric learning (Li et al., 2013), and locally adaptive decision functions (Xu et al., 2018), etc. The above-mentioned low-level visual feature extraction algorithms are difficult to extract discriminative features when faced with person image samples (lighting, complex backgrounds, etc.) with large variations in style. The rapid development of deep learning algorithms and convolutional neural

networks and the improvement of computing hardware, especially the arithmetic power of graphics processing units (GPU), have promoted the application of deep learning methods for image-based person re-identification. Unlike traditional methods, deep learning-based person re-identification methods integrate the image feature extraction module and the metric learning comparison similarity module into one model, which greatly increases the efficiency.

Figure 1 Flowchart of person re-identification (see online version for colours)

Most current research works related to person re-identification combine global and local branches trained together to extract features from person images. However, local features often require additional models such as human pose estimation (Martinel et al., 2015) or human semantic masks (Li et al., 2013) to locate persons in the images. The additional models not only increase the complexity of the model, but also the inaccurate localisation will directly affect the later work. Therefore, this paper obtained effective results by only applying the global branch, including multiple attention mechanisms and feature fusion methods to extract features.

2 Related works

2.1 Person re-identification

The global feature approach extracts the representational information of pedestrians without any spatial information; however, in non-controllable environments such as illumination and occlusion, person re-identification methods that rely only on global features cannot resolve large intra-class differences and usually lead to a significant degradation of retrieval performance and can no longer accurately identify pedestrians. Qian et al. (2017) developed a multi-scale deep feature representation model to capture discriminative cues at different scales. SVDNet (Li et al., 2014) uses only convolutional neural networks to learn the overall characteristics of pedestrians. In order to solve above

problems, some methods based on local feature extraction are usually applied to the person re-identification. The features of each group of channels in SCPNet (Fan et al., 2019) provide the pedestrians features of a spatial region of the pedestrian's body and use spatial-channel correlation to supervise the network to learn a robust feature. PCB-U-RPP (Sun et al., 2019b) obtains the tensor through the baseline network, then divides the parts equally, after which a classifier is learned for each part separately. Sun et al. (2018) divided the image into pre-defined equal panels to extract the corresponding panel features, and later to capture the relationship between multiple body parts, Varior et al. (2016) applied a long and short term memory architecture, and Hou et al. (2019) predict the occluded body parts from the information of unoccluded body parts in the current pedestrian image.

There are also methods that combine pose estimation to locate key points of the human skeleton and human semantic parsing to help pedestrians better extract features. Su et al. (2017) dividing the human body into parts by estimating the key points of human posture, and Zhang et al. (2019) use human semantic partial alignment to improve the robustness of the model to complex backgrounds and attenuate the effect of background on extracting pedestrian features. Some researchers have proposed that combining global features and local features can improve the performance of the model. Although the multi-branch network structure can mine richer pedestrian semantic information, it also makes the network training more difficult and slows down the convergence speed of the network. Suh et al. (2018) proposes a bilinear representation combining global feature and local feature representation to further extract pedestrian refined features. Wang et al. (2018) used a three-branch structure to extract multiple granularities of pedestrian semantic information by splitting the features into blocks vertically, with one branch used to extract global features and the remaining two branches used for local feature representation to enhance the pedestrian feature representation by fusing global information with discriminative multi-grain local information.

2.2 Attention mechanism

The attention mechanism is used to extract representation learning to solve image misalignment problems because of its property of enhancing important features and suppressing irrelevant features. Yang et al. (2019) proposed a combination of spatial attention and channel attention is proposed to learn to capture features that distinguish between the overall pedestrian image and the pedestrian part image. In addition, an interactive attention module was designed to enable the network to learn optimal weights adaptively. Li et al. (2018) found that the existing methods are insufficient for soft attention, so they combined hard and soft attention mechanisms to learn important features at the region level and pixel level to solve the problem of large disparity between different graph phases of the same pedestrian, and also proposed a cross-attention interaction learning mechanism to learn global features and local features efficiently and jointly.

2.3 Occluded person re-identification

Since pedestrian occlusion under the camera is common, the pedestrian information that can be captured from the image is limited, and some researchers directly use the person

re-identification model to deal with the occlusion problem, which will be less effective using the occluded dataset. Therefore, the performance of the approach that considers the pedestrian as a whole is not high. Other researchers simply segment the image into local facets, obtain relevant features by sliding windows, and then aggregate the local facets to form the final features. Zheng et al. (2015a) applied sliding window matching (SWM) by constructing a sliding window of the same size as the search image, dividing the image into smaller local matches and using it to match the most similar regions in each gallery image, and achieving alignment between local image blocks by a sparse classifier based on a fuzzy sensitive matching classifier (AMC); however, the above methods are not computationally efficient due to repeated extraction of He et al. (2018a) proposed deep spatial feature reconstruction (DSR), which does not require image block alignment and is applicable to pedestrian images of arbitrary size due to the use of full convolutional networks. Sun et al. (2019a) proposed the visibility-aware component model (VPM), which applies a region localiser that learns which part of the pedestrian image is visible or not by self-supervised learning, and focuses on components that share visible regions when comparing two images, significantly improving the accuracy of person re-identification.

2.4 Metric learning

In metric-based learning methods, metric learning refers to learning a similarity function that pulls features from the same identity samples closer together and pushes different identity features away from each other. Through the continuous development of deep learning, metric learning is now mainly divided into identity loss, verification loss and triplet loss, and many improved methods are improved based on these three losses. Classification loss, also known as id loss (Xiao et al., 2016) uses the pedestrian's id as a training label to train the model, and only one image is input at a time; verification loss (Chen et al., 2017) inputs two pedestrian images and lets the model learn whether these two images belong to the same pedestrian, which is equivalent to a binary classification problem; triplet loss (Hermans et al., 2017) uses a triplet input unit, and each triplet unit contains three image samples: anchor point, a positive sample (with the same features as the anchor point) and a negative sample (with different features from the anchor point). Usually, combining multiple losses can effectively improve the effectiveness of the model, so the metric learning in this paper combines classification loss and ternary group loss.

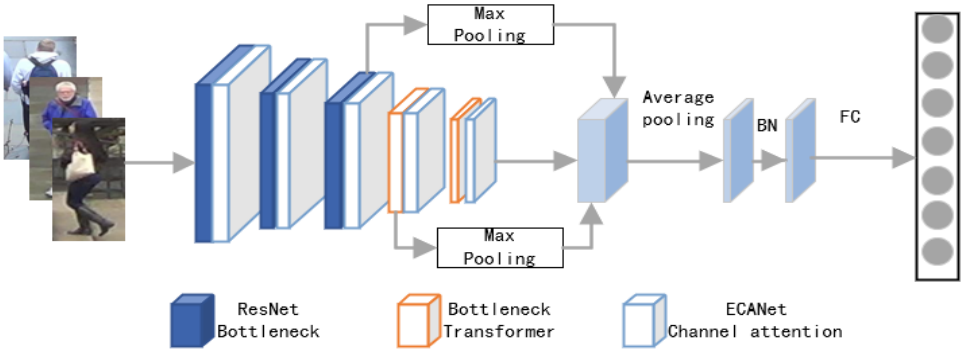
3 Proposed method

The general framework of person re-identification based on attention mechanism proposed in this paper is shown in Figure 2.

There are mainly baseline networks, multi-head self-attention (MHSA) module (Srinivas et al., 2021), efficient channel attention (ECA) module (Wang et al., 2020a), and feature fusion module. In the first step pedestrian images is extracted features through the ResNet50-IBN-a (Pan et al., 2018) baseline network, where the 3 × 3 convolutional modules in the Conv4_x and Conv5_x residual blocks are replaced with multi-head attention modules, while channel attention modules are accessed after each residual block from Conv1_x to Conv5_x. The feature fusion module fuses the features

of the last three residual blocks, where the output features of the Conv3_x and Conv4_x residual blocks are downsampled by maximum pooling so that they can be better fused with the output features of the Conv5_x residual block. The second step combines multiple losses such as triplet loss and cross-entropy loss for loss optimisation.

Figure 2 Base framework of our method (see online version for colours)



3.1 Baseline network

The ResNet50-IBN-a network is a variant of the ResNet50 network with better results in image classification problems and a smaller number of parameters and computation. The unification of instance normalisation (IN) and batch normalisation (BN) is achieved. In short, IN learns features that are shape invariant without affecting the deep texture differences, while BN learns the differences between features. So IN is suitable for the shallow layers of the network, while BN is suitable for the deep layers of the network. Therefore, ResNet50-IBN-a adds IN to the first three groups Conv2_x to Conv4_x of ResNet50, and Conv5_x only uses BN. The residual block of ResNet50 is shown as Figure 3(a), and the residual block of ResNet50-IBN-a is shown as Figure 3(b).

3.2 MHSA module

Since deep learning entered the computer field of vision, it still relies mainly on convolutional neural networks to extract features. However, vision transformer (Dosovitskiy et al., 2020) is able to achieve results comparable to the best convolutional neural networks by introducing a pure transformer model. The disadvantage of the pure transformer model is that the size of the input image is fixed and the computational effort increases exponentially. Therefore, Srinivas et al. (2021) proposed bottleneck transformers for visual recognition, i.e., embedding the blocks of transformer directly into the blocks of convolutional neural network, constituting a hybrid model of convolutional neural network and transformer.

In this paper, the MHSA module in transformer is applied to person re-identification, which can improve the performance of person re-identification by applying MHSA to replace the 3 × 3 convolution in the residual blocks of the baseline model Conv4_x and Conv5_x for person re-identification, as shown in Figure 4(a), represents a bottleneck block in the Conv block, and Figure 4(b) represents the bottleneck block in the transformer. Figure represents the multi-headed self-attentive module.

Figure 3 Baseline network of our method, (a) ResNet50 (b) ResNet50-IBN-a (see online version for colours)

Figure 4 Bottleneck block before and after change, (a) ResNet bottleneck (b) bottleneck transformer (see online version for colours)

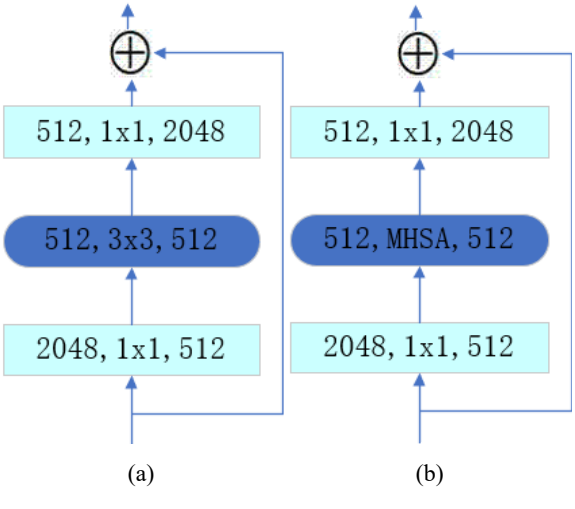


Figure 5 MHSA (see online version for colours)

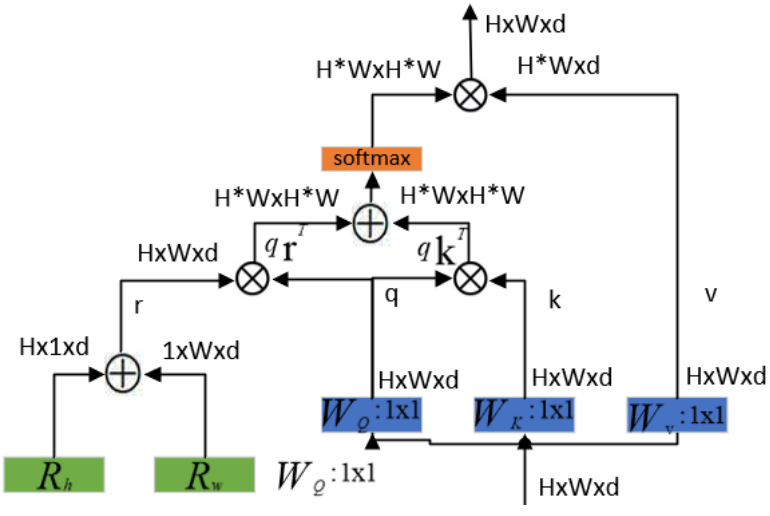


Table 1 Network structure based on MHSA

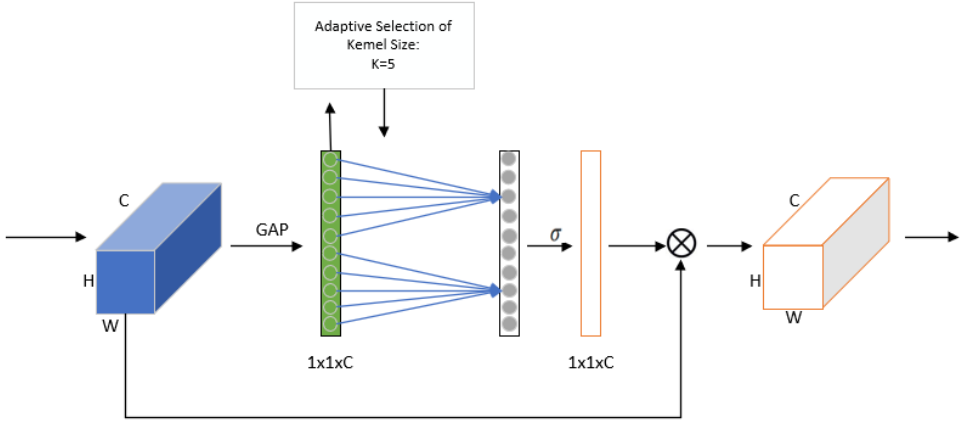
Layer name	*ResNet50-IBN*	*ResNet50-IBN-MHSA*
Conv1	7 × 7, 64, stride2	
	3 × 3, max pool, stride2	
Conv2_x	1×1, 64 3×3, 64 1×1, 256 ×3	1×1, 64 3×3, 64 1×1, 256 ×3
Conv3_x	1×1, 128 3×3, 128 1×1, 512 ×4	1×1, 128 3×3, 128 1×1, 512 ×4
Conv4_x	1×1, 256 3×3, 256 1×1, 1,024 ×6	1×1, 256 3×3, 256 1×1, 1,024 ×6
Conv5_x	1×1, 512 3×3, 512 1×1, 2,048 ×3	1×1, 512 *MHSA*, 512 1×1, 2,048 ×3

Table 1 shows the design concept of the residual network structure composed of ResNet50-IBN and an improved residual module based on multi-head attention mechanism (ResNet50-IBN-MHSA). Each bracket in the table represents a residual block.

3.3 Channel attention module

The human eye observes images to obtain discriminative details and suppress irrelevant information through attentional mechanisms. This is a means for human beings to quickly filter out high-value information from a large amount of information with limited attention resources, and it is a survival mechanism formed by human beings during long-term evolution, which greatly improves the efficiency and accuracy of processing visual information.

Figure 6 ECA module (see online version for colours)



In order to make the model pay more attention to the pedestrian foreground, a channel attention mechanism is introduced in the paper, but the traditional channel attention mechanism such as SENet uses a new structural unit called by SE block that negatively affects the channel attention prediction, and it is inefficient and unnecessary to obtain the dependencies, so (Wang et al., 2020b) proposed the ECA module, shown in Figure 6, which involves only a few parameters but has significant effect gain, proposes a avoids dimensionality reduction and effectively captures cross-channel interaction strategies. Avoiding downscaling and moderate cross-channel interactions can not only have an impact on the channel attention learning effect, but also significantly reduce the complexity of the model while maintaining the accuracy. Change the linear function of a mapping between the kernel size k and channel dimension C of a one-dimensional convolution to a nonlinear function $\phi(k) = r * k - b$, as function (1).

$$C = \phi(K) = 2^{r*k-b} \tag{1}$$

So, the kernel size k and channel dimension C of a given one-dimensional convolution can be adaptively determined, as shown in the following function (2).

$$k = \psi(C)\left|\frac{\log_2(C)}{r} + \frac{b}{r}\right|od, \tag{2}$$

where $|T|od$ represents the nearest odd number of T. In this article, we will set r and b to 2 and 1, respectively. Obviously, by mapping ψ, and by using nonlinear mapping, high-dimensional channels have longer distance interactions, while low-dimensional channels have shorter distance interactions.

Table 2 Network structure based on ECA

Layer name	*ResNet50-IBN-MHSA*	*ResNet50-IBN-MHSA_ECA*
Conv1	7 × 7, 64, stride2	
	3 × 3, max pool, stride2	
Conv2_x	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$
		ECA
Conv3_x	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$
		ECA
Conv4_x	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1{,}024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1{,}024 \end{bmatrix}\times6$
		ECA
Conv5_x	$\begin{bmatrix} 1\times1, & 512 \\ MHSA, & 512 \\ 1\times1, & 2{,}048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, & 512 \\ MHSA, & 512 \\ 1\times1, & 2{,}048 \end{bmatrix}\times3$
		ECA

3.4 Feature fusion module

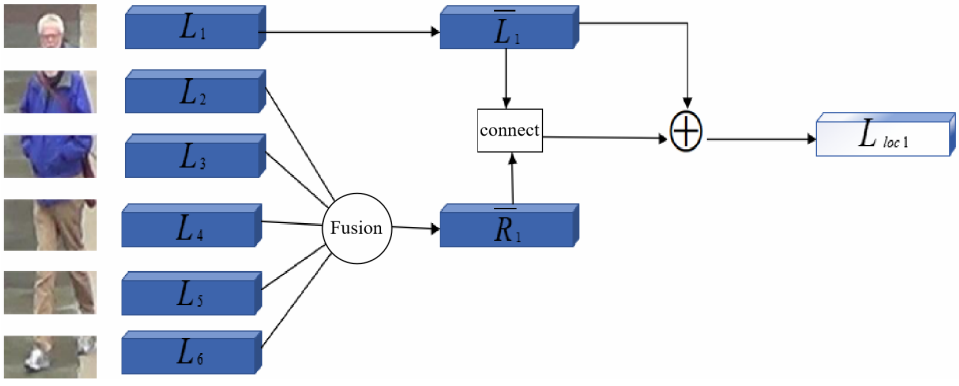
Convolutional neural networks extract features one by one based on residual blocks, and use the output feature of the last residual block as a discriminator but the final extracted features are all high-level human features, and the identification of intermediate features is ignored, which has a certain impact on the identification rate of subsequent character recognition. Therefore, fusing intermediate and high-level features can make up for the shortcomings between them.

When extracting pedestrian features, the output features of Conv3_x are 32 × 16 × 512, the output features of Conv4_x are 16 × 8 × 1,024, and the output features of Conv5_x are 8 × 4 × 2,048. In order to make the feature maps output by Conv3_x to Conv5_x uniform and achieve the fusion of features from multiple scales, the feature maps output by Conv3_x and Conv4_x in Figure 1 are downsampled using the maximum pooling operation.

Although this can make pedestrian images structured and robust to complex scenes. But when facing occlusion problems, local features only cover a small part of the image. Most importantly, the correlation between local features was not considered. That is to say, the local features of each part are relatively independent and have no feature information associated with other body parts. This leads to similarity calculation between pedestrians with similar attributes in the non-occluded area of pedestrians. To address the above issues, this article utilises the relationships between different body regions to

express pedestrian features. Specifically, by introducing a feature association module, the relationship between different regions of the body is utilised. Each local feature can contain information about the corresponding region itself and other body regions.

Figure 7 Local feature relationship extraction module (see online version for colours)



The specific relationship between local features is as follows: the pedestrian feature map extracted by the human pose estimation network is divided into six local features, each with a size of $1 \times 1 \times C$. Subsequently, an average pooling operation is performed on all local features. In the feature aggregation module, except for the specified local features, the aggregation operation from the remaining five local features is shown in function (3):

$$R_i = \frac{1}{5}\sum L_j (i \in 1, \ldots, 6,\ j \in 1, \ldots, 6,\ i \neq j), \tag{3}$$

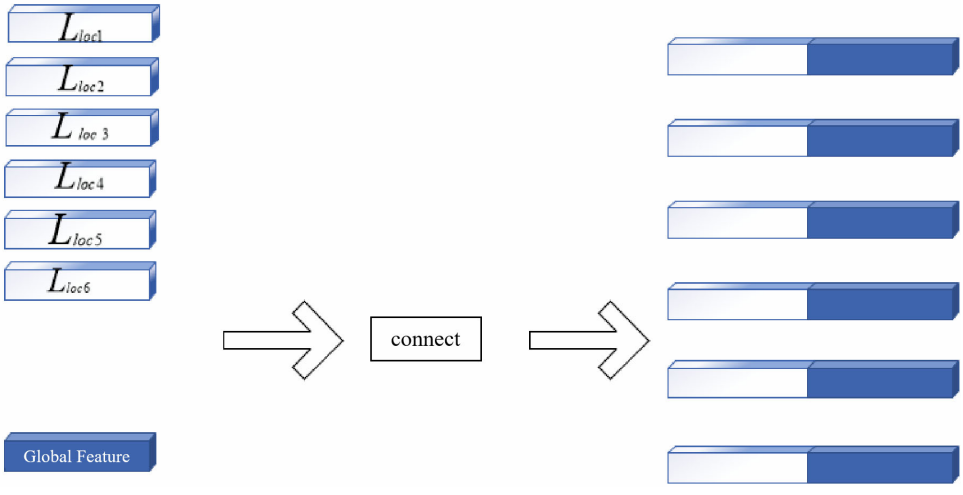
where R_i is the local feature set and L_i is the local feature point. In order to obtain a sum feature map of size $1 \times 1 \times C$, a convolutional layer of 1×1 is added after each and. The next step is to connect these two features ($\overline{L}_i$ and $\overline{R}_i$) through an association relationship, and output a local relationship feature L_{loci} for each L_i. This connection operation as follows:

$$L_{loci} = \overline{L}_i + SKIP\left(Connect\left(\overline{L}_i, \overline{R}_i\right)\right), \tag{4}$$

where *SKIP* is a module composed of 1×1 convolutional layer, ReLU layer, and BN layer, using Connect to represent the connections between features. The association module in this article can better utilise the local features of pedestrians, making them more discriminative and robust against occlusion. An example of extracting local feature relationships is shown in Figure 7, and extracting other local feature relationships is also similar. The association between global features and local features is as follows: whether the pedestrian involved in the local feature association does not include the entire part information of the pedestrian. Therefore, by further integrating global features with local features that are associated, the pedestrian features obtained not only include the relationship between certain features of the pedestrian and the remaining features of other parts. At the same time, it also includes local features and global features of the overall pedestrian. The specific process is as follows: perform a 1×1 convolution operation on

both the extracted global pedestrian features and the local features obtained through association relationships, and perform a flat pooling operation, as shown in Figure 8.

Figure 8 Global and local feature association module (see online version for colours)



3.5 Loss function

In this experiment, the joint training of cross-entropy loss function and triplet loss function is used. The cross-entropy loss function is a common classification loss, which describes the distance between two samples, when the smaller the cross-entropy indicates the closer the two are to each other, as follows:

$$L_{softmax} = -\sum_{i=1}^{k} \log \frac{\exp\left(W_{y_i}^T x_i + b_{y_i}\right)}{\sum_{j=1}^{n} \exp\left(W_j^T x_i + b_j\right)} \tag{5}$$

In function (3), k is the size of the batch, n is the total number of person identities, y_i is the identity of the person image, x_i is the feature vector of the i^{th} sample in the class, W is the weight, and b is the bias.

Triplet loss adopts a triplet input unit, with each triplet unit containing three image samples: reference image sample a, positive sample p with the same features as the reference image, and negative sample n with different features from the reference image. Three input image samples form a set of positive and negative sample sets, where images a and p are a pair of positive sample groups, and images a and n are a pair of negative sample groups.

The purpose is to make the distance between positive sample groups closer, and the distance between incorrect sample groups farther, thereby improving the efficiency of person re-identification networks. Namely, the purpose is to bring a certain sample x_a closer to its positive sample x_p relative to the negative sample x_n. The function is as follows:

$$L_{tri} = \max\left(\left\|x_a - x_p\right\| - \left\|x_a - x_n\right\| + margin, 0\right) \tag{6}$$

The final total loss function is expressed as:

$$L_{total} = L_{softmax} + L_{tri} \tag{7}$$

4 Experiments

4.1 Datasets

To evaluate the validity of the experimental model, it was evaluated on top of three publicly available datasets, including two full-body pedestrian datasets Market-1501 (Zheng et al., 2015b) and DukeMTMC-reID (Ristani et al., 2016), and one occluded pedestrian dataset Occluded-DukeMTMC (Miao et al., 2019).

Market-1501 was acquired within Tsinghua University in 2015 with images from six different resolution cameras, with a training set consisting of 751 individuals containing 12,936 images and a test set consisting of 750 individuals containing 19,732 images.

DukeMTMC-reID was acquired at Duke University in 2017 with images from eight different resolution cameras, a training set consisting of 702 individuals containing 16,522 images, and a test set consisting of 702 individuals containing 19,889 images.

Occluded-DukeMTMC is an occluded dataset extracted from the DukeMTMC-reID dataset. Among them, this training set consisting of 702 individuals containing 15,618 images and a test set consisting of 1,110 individuals containing 17,661 images and 2,210 images are available in the gallery and query set, respectively.

4.2 Experimental environment and parameters

The hardware platform used for the experiments is RTX3090 with 24GB GPU and 48GB CPU. The software platform is Pytorch 1.10.0, python version 3.8 and CUDA version 11.3.

The experimental baseline model is ResNet50-IBN-a pre-trained on ImageNet, the learning rate is 0.0001, and the input image size is 256 × 128, the optimiser is chosen as SGD, the input image batch is 64, and the number of network iterations is 200.

4.3 Evaluation indicators

In this paper, we use the two most commonly used evaluation metrics for person re-identification, rank-n accuracy and mean average precision (mAP). rank-n reflects the probability that the top n images with matching values among the person images to be selected are the persons to be queried, and mAP integrates accuracy and recall, which can reflect the degree to which the query images are at the top of the image The mAP reflects the degree to which all correct images in the library are at the top of the retrieved list.

4.4 Experimental results

Due to prove the effectiveness of the model in this paper, the model in this paper was tested on mainstream person re-identification datasets such as Market-1501 and

DukeMTMC-reID, and the experiments were done to compare with some familiar methods in recent years. map values on Market-1501 dataset and DukeMTMC-reID dataset were 82.4% and 65.9%, and the values of rank-1 are 94.8% and 84.5%, respectively, both of which have achieved more satisfactory results.

In Table 3, the CGEA model using graph neural network and cross-plot embedding alignment layer to jointly learn each person key point region and embed topological information achieves good results on both Market-1501 dataset and DukeMTMC-reID dataset, but our model has low computational complexity and a simple structure. In the Market-1501 dataset, our model improves 7.2% and 3.6%, respectively, compared to the SCPNet model rank-1 and mAP. In the DukeMTMC-ReID dataset, the algorithm of this paper improves 3.3% and 4.2%, respectively, compared with the SCPNet model rank-1 and mAP. The results show that the model in our paper can extract person features relatively well.

Table 3 Comparison of results of different methods on Market-1501 and DukeMTMC-reID datasets %

Methods	*Market-1501*		*DukeMTMC-reID*	
	mAP (%)	*Rank-1 (%)*	*mAP (%)*	*Rank-1 (%)*
SVDNet (Li et al., 2014)	62.1	82.3	56.8	76.7
DLPAR (Zhao et al., 2017)	63.4	81	-	-
SCPNet (Fan et al., 2019)	75.2	91.2	62.6	80.3
CASN+IDE (Zheng et al., 2019)	78	92	67	84.5
HACNN (Li et al., 2018)	75.7	91.2	63.8	80.5
AlignedReID++ (Luo et al., 2019)	79.1	91.8	69.7	82.1
HOReID (Wang et al., 2020a)	*84.9*	94.2	75.6	86.9
PCB-U+RPP (Sun et al., 2018)	81.0	93.1	70.7	84.3
Our method	82.4	*94.8*	*75.8*	80.5

Table 4 Comparison of the results of different methods on the Occluded-DukeMTMC dataset %

Methods	*Occluded-DukeMTMC*			
	mAP (%)	*Rank-1 (%)*	*Rank-5 (%)*	*Rank-10 (%)*
LOMO+XQDA (Liao et al., 2015)	5.0	8.1	17.0	22.2
DIM (Yu et al., 2017)	14.4	21.5	36.1	42.8
DLPAR (Zhao et al., 2017)	20.2	28.8	44.6	51.0
Random erasing (Zhong et al., 2020)	30.0	40.5	59.6	66.8
HACNN (Li et al., 2018)	26.0	34.4	51.9	59.4
Adver Occluded (Huang et al., 2018)	32.2	*44.5*	-	-
DSR (He et al., 2018a)	30.4	40.8	58.2	65.2
SFR (He et al., 2018b)	32.0	42.3	*60.3*	67.3
Our method	*33.3*	40.3	59.2	*67.7*

The model in this paper was also experimented on a larger occluded person re-identification Occluded-DukeMTMC against some mainstream methods, and the

experimental results are shown in Table 2 with map values, rank-1, rank-5 and rank-10 of 33.3%, 40.3%, 59.2% and 66.7%, respectively.

In Table 4, the HACNN model is relatively early to use attention mechanism to solve the person re-identification related problems, which mainly learns global features and local features jointly, soft attention and hard attention, and allows interactive learning between soft and hard attention. In contrast, the structure of the model in this paper is relatively simple, without redundant parameters. The DSR model needs to manually crop the occlusion region to deal with the occlusion problem, which adds considerable workload to the model. This paper can directly input person images for feature extraction, thus avoiding the step of manual crop, and the final experimental results achieved are also well.

4.5 Ablation experiments

Due to explain the validity of each module added, ablation experiments were done for this purpose, and all the ablation experiments in this paper were tested with the Market-1501 dataset as an example. The data obtained from the ablation experiment in Table 5 shows that the use of multi head attention mechanism to improve network structure, combined with ECA mechanism and feature fusion for person re-identification, gradually increases both rank-1 and mAP values on the basis of the baseline network. This is attributed to the joint channel attention, self attention mechanism, and feature fusion being able to extract more significant robust features of person without losing key features that can distinguish different persons.

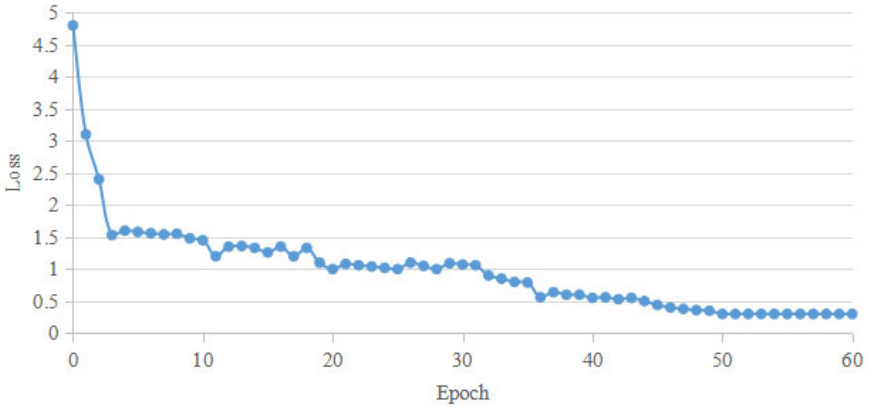
Table 5 Experimental results under different block

Methods	*mAP (%)*	*Rank-1 (%)*
Baseline	71.7	82.4
Baseline + MHSA	76.3	89.6
Baseline + MHSA + ECA	80.5	92.1
Baseline + MHSA + ECA + feature fusion	82.4	94.8

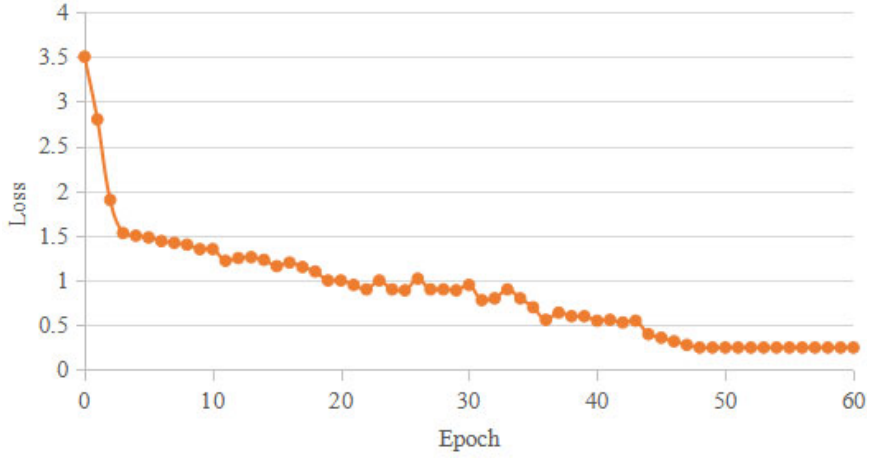
4.6 Loss function results

Figure 9 shows the loss curve results during the training period of the Market-1501 dataset using our method. Firstly, by observing the Loss curves in the two graphs, it can be clearly observed that the overall effect is first rapidly decreasing and then tending to a relatively stable state. This fully conforms to the state of neural network training in deep learning. Figure 9(a) is the loss curve validated by the pedestrian recognition results of the network with channel attention mechanism module added to the baseline network. The change of the curve starts with a rapid decrease and a small fluctuation at the third epoch, and then stabilises at the 50th epoch; Figure 9(b) is a loss curve based on feature fusion networks with different residual blocks. The curve changes rapidly first. At the third epoch, it slows down and begins to fluctuate slightly. Starting to stabilise at the 48th epoch; In order to prevent overfitting of the network model, the network loss value continuously decreases after multiple iterations, and ultimately reaches a stable state.

Figure 9 Loss curve during training (a) training results based on attention mechanism (b) training results based on feature fusion (see online version for colours)



(a)



(b)

5 Conclusions

In this paper, we provide a person re-identification model based on the attention mechanism. Firstly, we consider that the visual transformer has better results in the field of image processing compared with the traditional convolutional neural network, but using pure Transformer will add a large number of parameters and lead to a significant increase in computation, thus we use the core MHSA mechanism in transformer and convolutional neural network. Secondly, a simple and effective channel attention mechanism is added to focus the model of this paper more on the important parts of the person foreground. Finally, fusing the mid-level and high-level features in the model avoids the loss of some distinguishing features. Experimental results on three major datasets, Market-1501, DukeMTMC-reID, and Occluded-DukeMTMC, show that the

performance of the proposed method is improved, and the performance metrics exceed many existing person re-identification models.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (No. RS-2023-00247045); Xi'an Science and Technology Planning Project under Grant No. 22GXFW0046.

References

Chen, W., Chen, X., Zhang, J. and Huang, K. (2017) 'A multi-task deep network for person re-identification', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31, No. 1.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. and Houlsby, N. (2020) *An Image is Worth 16 x 16 Words: Transformers for Image Recognition at Scale*, arXiv preprint arXiv:2010.11929.

Fan, X., Luo, H., Zhang, X., He, L., Zhang, C. and Jiang, W. (2019) 'SCPNET: spatial-channel parallelism network for joint holistic and partial person re-identification', in *Computer Vision – ACCV 2018: 14th Asian Conference on Computer Vision*, Perth, Australia, 2–6 December 2018, No. Part 2/14, pp.19–34.

He, L., Liang, J., Li, H. and Sun, Z. (2018a) 'Deep spatial feature reconstruction for partial person re-identification: alignment-free approach', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.7073–7082.

He, L., Sun, Z., Zhu, Y. and Wang, Y. (2018b) *Recognizing Partial Biometric Patterns*, arXiv preprint arXiv:1810.07399.

Hermans, A., Beyer, L. and Leibe, B. (2017) *In Defense of the Triplet Loss for Person Re-identification*, arXiv preprint arXiv:1703.07737.

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S. and Chen, X. (2019) 'VRSTC: occlusion-free video person re-identification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7183–7192.

Huang, H., Li, D., Zhang, Z., Chen, X. and Huang, K. (2018) 'Adversarially occluded samples for person re-identification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5098–5107.

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M. and Bischof, H. (2012) 'Large scale metric learning from equivalence constraints', in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp.2288–2295.

Li, W., Zhao, R., Xiao, T. and Wang, X. (2014) 'Deepreid: deep filter pairing neural network for person re-identification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.152–159.

Li, W., Zhu, X. and Gong, S. (2018) 'Harmonious attention network for person re-identification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2285–2294.

Li, Z., Chang, S., Liang, F., Huang, T.S., Cao, L. and Smith, J.R. (2013) 'Learning locally-adaptive decision functions for person verification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.3610–3617.

Liao, S., Hu, Y., Zhu, X. and Li, S.Z. (2015) 'Person re-identification by local maximal occurrence representation and metric learning', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2197–2206.

Lowe, D.G. (1999) 'Object recognition from local scale-invariant features', in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Vol. 2, pp.1150–1157.

Luo, H., Jiang, W., Zhang, X., Fan, X., Qian, J. and Zhang, C. (2019) 'Alignedreid++: dynamically matching local information for person re-identification', *Pattern Recognition*, October, Vol. 94, pp.53–61.

Martinel, N., Micheloni, C. and Foresti, G.L. (2015) 'Saliency weighted features for person re-identification', in *Computer Vision – ECCV 2014 Workshops, Proceedings*, Zurich, Switzerland, 6–7 and 12 September 2014, No. Part 3/13, pp.191–208.

Miao, J., Wu, Y., Liu, P., Ding, Y. and Yang, Y. (2019) 'Pose-guided feature alignment for occluded person re-identification', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.542–551.

Pan, X., Luo, P., Shi, J. and Tang, X. (2018) 'Two at once: enhancing learning and generalization capacities via IBN-Net', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.464–479.

Qian, X., Fu, Y., Jiang, Y.G., Xiang, T. and Xue, X. (2017) 'Multi-scale deep learning architectures for person re-identification', in *Proceedings of the IEEE International Conference on Computer Vision*, pp.5399–5408.

Ristani, E., Solera, F., Zou, R., Cucchiara, R. and Tomasi, C. (2016) 'Performance measures and a data set for multi-target, multi-camera tracking', in *European Conference on Computer Vision*, Springer International Publishing, Cha, pp.17–35.

Srinivas, A., Lin, T.Y., Parmar, N., Shlens, J., Abbeel, P. and Vaswani, A. (2021) 'Bottleneck transformers for visual recognition', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.16519–16529.

Su, C., Li, J., Zhang, S., Xing, J., Gao, W. and Tian, Q. (2017) 'Pose-driven deep convolutional model for person re-identification', in *Proceedings of the IEEE International Conference on Computer Vision*, pp.3960–3969.

Suh, Y., Wang, J., Tang, S., Mei, T. and Lee, K.M. (2018) 'Part-aligned bilinear representations for person re-identification', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.402–419.

Sun, Y., Xu, Q., Li, Y., Zhang, C., Li, Y., Wang, S. and Sun, J. (2019a) 'Perceive where to focus: learning visibility-aware part-level features for partial person re-identification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.393–402.

Sun, Y., Zheng, L., Li, Y., Yang, Y., Tian, Q. and Wang, S. (2019b) 'Learning part-based convolutional features for person re-identification', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 3, pp.902–917.

Sun, Y., Zheng, L., Yang, Y., Tian, Q. and Wang, S. (2018) 'Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.480–496.

Varior, R.R., Shuai, B., Lu, J., Xu, D. and Wang, G. (2016) 'A Siamese long short-term memory architecture for human re-identification', in *Computer Vision – ECCV 2016: 14th European Conference, Proceedings*, Amsterdam, The Netherlands, 11–14 October, No. Part 7/14, pp.135–153.

Wang, G., Yuan, Y., Chen, X., Li, J. and Zhou, X. (2018) 'Learning discriminative features with multiple granularities for person re-identification', in *Proceedings of the 26th ACM International Conference on Multimedia*, pp.274–282.

Wang, G.A., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S. and Sun, J. (2020a) 'High-order information matters: learning relation and topology for occluded person re-identification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.6449–6458.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q. (2020b) 'ECA-Net: efficient channel attention for deep convolutional neural networks', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11534–11542.

Xiao, T., Li, H., Ouyang, W. and Wang, X. (2016) 'Learning deep feature representations with domain guided dropout for person re-identification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1249–1258.

Xu, J., Zhao, R., Zhu, F., Wang, H. and Ouyang, W. (2018) 'Attention-aware compositional network for person re-identification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2119–2128.

Yang, F., Yan, K., Lu, S., Jia, H., Xie, X. and Gao, W. (2019) 'Attention driven person re-identification', *Pattern Recognition*, February, Vol. 86, pp.143–155.

Yu, Q., Chang, X., Song, Y.Z., Xiang, T. and Hospedales, T.M. (2017) *The Devil is in the Middle: Exploiting Mid-Level Representations for Cross-Domain Instance Matching*, arXiv preprint arXiv:1711.08106.

Zajdel, W., Zivkovic, Z. and Krose, B.J.A. (2005) 'Keeping track of humans: have I seen this person before', in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, IEEE, Piscataway, pp.2081–2086.

Zhang, Z., Lan, C., Zeng, W. and Chen, Z. (2019) 'Densely semantically aligned person re-identification', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.667–676.

Zhao, L., Li, X., Zhuang, Y. and Wang, J. (2017) 'Deeply-learned part-aligned representations for person re-identification', in *Proceedings of the IEEE International Conference on Computer Vision*, pp.3219–3228.

Zheng, M., Karanam, S., Wu, Z. and Radke, R.J. (2019) 'Re-identification with consistent attentive Siamese networks', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5735–5744.

Zheng, W.S., Gong, S. and Xiang, T. (2012) 'Reidentification by relative distance comparison', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 3, pp.653–668.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q. (2015a) 'Scalable person re-identification: a benchmark', in *Proceedings of the IEEE International Conference on Computer Vision*, pp.1116–1124.

Zheng, W.S., Li, X., Xiang, T., Liao, S., Lai, J. and Gong, S. (2015b) 'Partial person re-identification', in *Proceedings of the IEEE International Conference on Computer Vision*, pp.4678–4686.

Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y. (2020) 'Random erasing data augmentation', in *Proceedings of the AAAI Conference on Artificial Intelligence*, April, Vol. 34, No. 7, pp.13001–13008.