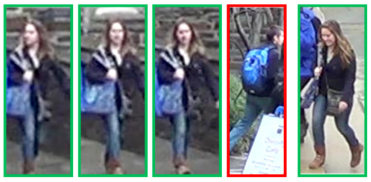
Person re-identification based on attention mechanism and adaptive weighting



Re-identificación de personas basada en mecanismo de atención y en ponderación adaptativa

Yangping Wang[1,3*], **Li Li**[1], **Jingyu Yang**[1,2] **and Jianwu Dang**[1,2]

[1] School of Electronic and Information Engineering,Lanzhou Jiaotong University, Anning West Road 88, Lanzhou, 730070, Gansu, China, w_yp73@163.com
[2] Gansu Provincial Engineering Research Center for Artificial Intelligence and Graphic & Image Processing, Anning West Road 88,Lanzhou, 730070, Gansu, China
[3] Gansu Provincial Key Lab of System Dynamics and Reliability of Rail Transport Equipment, Anning West Road 88, Lanzhou 730070, Gansu,China

DOI: https://doi.org/10.6036/9981 | Received: 30/oct/2020 • Reviewing: 30/oct/2020 • Accepted: 04/jan/2021

To cite this article: YANGPING, Wang; LI, Li; JINGYU, Yang; JIANWU, Dang. PERSON RE-IDENTIFICATION BASED ON ATTENTION MECHANISM AND ADAPTIVE WEIGHTING. *DYNA*, March 2021, vol. 96, no. 2, p. 186-193. DOI: https://doi.org/10.6036/9981

RESUMEN

- Debido a factores como el cambio de postura, la condición de la iluminación, el desorden del fondo y la oclusión, la re-identificación de la persona (re-ID) basada en los fotogramas de vídeo es una tarea difícil. Para utilizar la información de relevancia a nivel de píxel y la información discriminatoria del cuerpo local de la imagen y mejorar la precisión de la re-identificación en el caso de cambios complejos de postura y diferencias de puntos de vista, se propuso en este estudio una red de re-identificación de personas basada en el mecanismo de atención y el peso adaptativo. Sobre la base de la detección de puntos clave humanos, se integró un mecanismo de atención para examinar la información discriminatoria en diversas partes del cuerpo humano. En la red se adoptó el método de ponderación adaptativa, que proporciona a las características locales extraídas diferentes pesos según la información discriminatoria de las distintas partes del cuerpo humano. La exactitud de la reidentificación del modelo de la red se verificó mediante experimentos. Los resultados demuestran que el modelo de red propuesto puede extraer con precisión las características de las regiones discriminatorias de diversas partes del cuerpo humano integrando el mecanismo de atención y el peso de la región adaptativa, mejorando así el rendimiento de la reidentificación de la persona. Nuestro método se compara con los actuales modelos de red de identificación de personas ampliamente utilizados como la AACN y la HACN. En el conjunto de datos de Market-1501, los valores de Rank-1 y mAP se mejoran en un 4,79% y 2,78% así como en un 8% y 3,52%, respectivamente, y en el conjunto de datos de DukeMTMC-reID, en un 4,92% y 3,26% así como en un 5,17% y 3,17%, respectivamente. En comparación con el anterior modelo de red GLAD, los valores de Rango-1 y mAP en dos conjuntos de datos experimentales se incrementan en más de un 2%. El método propuesto proporciona un buen enfoque para optimizar el descriptor de peatones para la identificación de personas en entornos complejos.
- **Palabras clave:** Reidentificación de personas, Peso adaptativo, Mecanismo de atención, Red neuronal convolucional

ABSTRACT

Owing to factors such as pose change, illumination condition, background clutter, and occlusion, person re-identification (re-ID) based on video frames is a challenging task. To utilize pixel-level saliency information and discriminative local body information of the image and improve re-ID accuracy in the case of complex pose change and viewpoint difference, a person re-ID network based on attention mechanism and adaptive weight was proposed in this study. Based on the detection of human key points, an attention mechanism was integrated to screen the discriminative information in various parts of the human body. The adaptive weighting method was adopted in the network, providing the extracted local features different weights according to the discriminative information of different human parts. The re-ID accuracy of the network model was verified by experiments. Results demonstrate that the proposed network model can accurately extract the features of discriminative regions in various parts of the human body by integrating the attention mechanism and adaptive region weight, thereby improving the performance of person re-ID. Our method is compared with current widely used person re-ID network models as AACN and HAC. On the Market-1501 dataset, the Rank-1 and mAP values are improved by 4.79% and 2.78% as well as 8% and 3.52%, respectively, and on the DukeMTMC-reID dataset, by 4.92% and 3.26% as well as 5.17% and 3.17%, respectively. Compared with the previous GLAD network model, Rank-1 and mAP values on two experimental datasets are increased by more than 2%. The proposed method provides a good approach to optimize the descriptor of pedestrians for person re-ID in complex environments.

Keywords: Person re-identification, Adaptive weight, Attention mechanism, Convolutional neural network.

1. INTRODUCTION

With the wide application of surveillance video, person re-identification (re-ID) technology plays an important role in intelligent security by using it to determine whether a specific pedestrian is present in massive video data and then to track the target. Given one query image of one specific person, person re-ID technology is expected to provide all the samples with the same identification (ID) from a large number of video images captured by multiple non-overlapping cameras at different times. Owing to various factors such as hardware configuration, camera viewpoint, and pedestrian posture of surveillance video in different scenarios, images with the same ID captured by different cameras or the same camera at different times will be very different, introducing considerable challenges to person re-ID. Person re-ID comprises two main research contents: (1) extracting robust and discriminative features from pedestrian images and (2) designing efficient metric methods to address the intra-class and inter-class problem. To address the huge differences in images with the same ID caused by viewpoint, posture, and illumination, many robust handcrafted pedestrian descriptors have been proposed, such as color histogram [1-3], local binary pattern [4-5], and Gaussian descriptor [6]. Some representative metric learning methods have also been proposed [7-8]. However, for person re-ID, which is of great practical significance and is affected by many factors, the handcrafted pedestrian descriptors have poor generalization performance and cannot completely represent pedestrian characteristics. In recent years, deep learning has achieved considerable success in various computer vision and pattern recognition tasks, and it has been gradually applied to the field of person re-ID to solve this problem [9-13]. Deep learning is used to solve the problem of person re-ID, which integrates the two processes of feature extraction and metric learning. Deep learning can automatically extract more discriminative features and map the features to a better metric space. Scholars have designed many network models [14-18] on how to use neural networks to extract comprehensive and highly discriminative person features, playing an important role in the development of neural networks for person re-ID. However, these network models have some shortcomings. Among them, the network based on the global features tends to ignore the local discriminative information of the body parts, and it is less discriminative to pedestrians with a similar appearance. The network model based on the partial features of pedestrians cannot easily calculate the weight parameters between each part manually, resulting in the poor performance of person re-ID.

Based on the above analysis, the attention mechanism is embedded into the network model on the basis of using human key points to detect human body parts. And an adaptive weighting module is designed to learn the proportion of different body parts. The discrimination of feature descriptor is enhanced by combining the significant information of the global feature with the information of three weighted part features. The re-identification accuracy of the whole network model can be improved.

2. STATE OF THE ART

At present, scholars have conducted many studies on person re-ID based on deep learning networks, and they have proposed many effective network models. Li [19] pioneered in applying deep learning to person re-ID and proposed a filter pairing neural network (FPNN) model based on a convolutional neural network (CNN). The first layer of FPNN was the convolution layer with a max pooling operation, and then the block matching layer was added to match the filter response across the visual field. FPNN could reduce the influence of factors such as misalignment, occlusion, and background clutter on re-ID performance under a unified framework. By using the CNN-based Siamese network as the baseline, Ahmed et al. [20] proposed an enhanced deep learning framework to learn the relationship between features across the visual field by taking advantage of multi-input nearest neighbor difference and brief patch features. McLaughlin [21] introduced a recurrent neural network based on the Siamese network to fuse the temporal characteristics and deep features of video to improve re-ID accuracy. By adding a self-connected hidden layer across time points, the temporal features were introduced into pedestrian features to aggregate deep features, reduce the feature dimension, and enhance the robustness of features. Varior et al. [22] adopted long short-term memory (LSTM) on the basis of the Siamese network to process patches in sequence to memorize the spatial information of images, thereby enhancing the discriminability of deep features. Liu [23] integrated LSTM into the triplet network to imitate the human visual system through an end-to-end contrast attention model. Yan [24] proposed a recursive feature aggregation network that used LSTM to record the change information of pedestrian body parts with time. The approach used low-level handcrafted features, rather than CNN deep features, as input time nodes to avoid overfitting caused by training CNN on small datasets. LSTM could memorize and propagate the features with excellent performance and ignored poor features to establish the global features of the video. Aiming at the disadvantage of poor generalization of traditional triplet loss, Khatun [25] proposed a deep four-stream convolutional neural network for person re-ID. This method used four input images, two of which having the same identity and the other two having different identities. The network used double identification and verification losses in a single framework to minimize the intra-class distance and maximize the inter-class distance. Erbeti [26] presented an end-to-end integrated person re-ID method to solve the problem of overfitting in discriminative models. On a small dataset, this method could effectively deal with the overfitting problem and had high computational efficiency. To a certain extent, the above methods have achieved good results on small datasets made by overfitting. However, these methods require considerable time to train for re-ID under complex scene changes. Huang [27] designed an enhanced aggregated channel feature. Aiming at the need for real-time application, person detection and re-ID were implemented in the same network model. The algorithm met the real-time requirements, but the robustness of person detection and feature extraction in complex scenes with changeable postures was poor. Zheng [16] used the verification and identification model to jointly guide the network learning, aiming at learning more discriminative features. Wu [28] constructed the Spindle Net network integrating facial features. After detecting the key points of the human body, facial and other body parts could be divided into the specified size through the region of interest pooling. All features were learned by the network, and then the pedestrian descriptor was obtained by fusing the partial features. However, when fusing multiple partial features, the operation of simply slicing and obtaining the maximum activation value of each region did not fully use the information of extracted local features. Franco [29] proposed a convolutional covariance feature (CCF) based on the covariance descriptor. CCF is calculated by the adaptive and trainable features in the coarse to fine transfer learning strategy, with the invariable property of knowledge and noise, and CCF achieves a good effect on person re-ID. Barbosa [30] designed a convolutional neural network based on the inception architecture. The network could capture the structural attributes of the shape of the human body such as height, obesity, and gender. The network was not limited by the information of human appearance. Sun [17] et al. proposed a part-based convolutional baseline and used refined part pooling (RPP) to classify pixel positions in the later stage to achieve the effect of “soft segmentation.” However, the single division size could not effectively divide the local information of all images. In the test stage, the local features were only in conjunction with the global feature as the final description of pedestrians. To address the problem of pedestrian misalignment in the dataset, Wei [31] established a network called the global–local-alignment descriptor (GLAD). Based on the key points of the pedestrian body, the head, upper body, and lower body of the pedestrian were first calculated. Three local features were concatenated with a global feature in a fixed proportion as the final pedestrian descriptor. The GLAD network used the key point detection network to accurately divide the pedestrian into three parts according to their positions, extract the global features, and finally concatenate them to form the overall pedestrian descriptor. When the whole pedestrian picture was input, compared with the "hard segmentation" of the previous deep re-ID network, GLAD could accurately learn the local feature and easily adapt to the complex and changeable monitoring scene. However, because the head, upper body, and lower body contain considerable differences in the amount of discriminative information, local and global features concatenated with the fixed ratio could not effectively utilize the extracted pedestrian features in the matching process. The max pooling operation made the final pedestrian descriptor lose some information and reduce the overall discriminability[32].

In summary, due to the complexity of current monitoring video scenes, the various viewpoints, and the inaccurate pedestrian detection algorithm, the existing network models tend to ignore the pixel-level saliency information, and the fixed weighting of different pedestrian parts is insufficiently reasonable. Aiming at these problems, we propose a person re-ID network based on attention mechanism and adaptive weight. Our network uses the key point detection network to divide pedestrian parts. Combined with the attention mechanism, the model extracts fine-grained features of local and global pedestrian characteristics and uses adaptive weight to fuse multiple partial features as the final pedestrian descriptor, which is then classified through the softmax layer. The effectiveness of the proposed algorithm is verified on the person re-ID datasets Market-1501 and DukeMTMC-reID. Results demonstrate that the proposed algorithm is superior to traditional algorithms and other similar networks.

The remainder of this study is organized as follows. The pedestrian preprocess and the structure of our person re-ID network model based on attention mechanism and adaptive weight are introduced in detail in Section 3. Implementation details and experimental results are presented in Section 4. Several concluding remarks are drawn in Section 5.

3. METHODOLOGY

3.1. GLAD MODEL ARCHITECTURE

Owing to the considerable differences in the posture of the human body captured by different cameras, pedestrian images suffer from various misalignments, seriously affecting the performance of person re-ID. To address this problem, the GLAD network calculates three body parts, namely, the pedestrian's head, upper body, and lower body, according to the coordinates of the four body key points. Then, the entire pedestrian image and the three parts are input to learn the global and local features, respectively. The GoogLeNet network with shared weights is used as the backbone network, and two convolutional layers are used instead of the fully connected layers as the final classifier. The former convolutional layer performs the reduction of feature dimension, and the latter convolutional layer generates C feature maps, where C corresponds to the number of output categories. Then, a C-dimensional vector of the predicted category is obtained by global average pooling, and each branch calculates the loss separately to guide the network learning together. The overall structure of the network is shown in Figure 1.

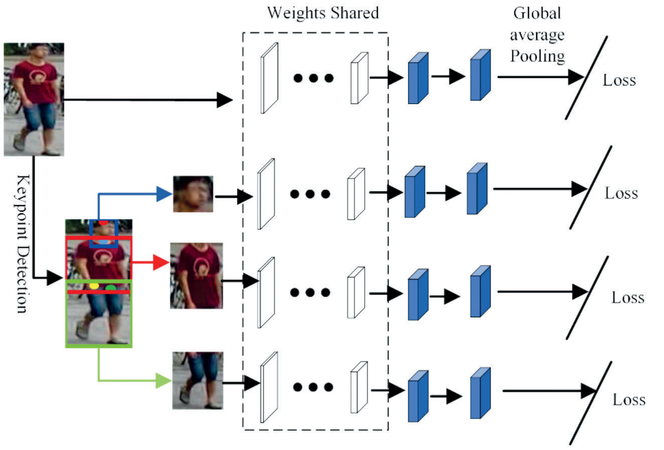


Fig. 1. The structure of GLAD

3.2. SEGMENTATION OF PEDESTRIAN PARTS

Prior to extracting the features of the input image of the GLAD network, the Deeper Cut [33] method is used to detect the four main key points of the pedestrian's body: the upper head, the neck, the left hip, and the right hip, as shown in Figure 2 (see section: supplementary material).

The head can be positioned by two key points: the upper head and the neck. Supposing a person image with a size of HxW, the coordinates of the upper head and the neck are (x_1, y_1) and (x_2, y_2), the head region B^h is calculated with Eq. (1).

$$\begin{cases} B^h = \left[\left(|x_c - w/2|_+ , |y_1 - \alpha|_+ \right), \left(x_c - w/2, y_2 + \alpha \right) \right] \\ w = y_2 - y_1 + 2 \cdot \alpha \\ x_c = \left(x_1 + x_2 \right) / 2 \end{cases} \quad (1)$$

where α is a parameter that controls the overlapping region between neighboring body parts. For a 512*256 image, the parameter α is experimentally set as 15. $|*|_+$ indicates a non-negative constrain, and its value is equal to 0, when * is negative.

Supposing that the coordinates of the left hip and the right hip are respectively (x_3, y_3) and (x_4, y_4), the upper body region B^{ub} and the lower body region B^{lb} is calculated with Eq. (2).

$$\begin{cases} B^{ub} = \left[\left(0, |y_2 - 2 \cdot \alpha|_+ \right), \left(W - 1, y_c + 2 \cdot \alpha \right) \right] \\ B^{lb} = \left[\left(0, |y_2 - 2 \cdot \alpha|_+ \right), \left(W - 1, H - 1 \right) \right] \\ y_c = \left(y_3 + y_4 \right) / 2 \end{cases} \quad (2)$$

The first picture in Figure 2 (see section: supplementary material) is the key point detection of a pedestrian image with normal posture completely unobstructed, and the second and third images are failure cases for key point detection caused by occlusion or the low resolution. However, the four key points, namely, upper-head, neck, right-hip, and left-hip, can be robustly detected even in those failure cases. Therefore, three regions of the body could be reliably sliced with the four key points in complex situations.

3.3. GLOBAL DESCRIPTOR

For an input pedestrian image with a size of 512*256, the GLAD network fixes the sliced pedestrian part to a specified size, scales it to 96*96 for the head region and 224*256 each for the upper body and lower body regions. Then, each region and the entire image are separately end-to-end trained for classification. The entire network is designed to use global and local input regions to correctly classify pedestrians and to calculate the classification loss for each region separately. Generally, the network is trained to focus on each part of the body and learn the representation of each region.

In the test phase, after extracting the feature representations of the entire body and each body part through the network, the global average pooling (GAP) is used to obtain the output features. The head, upper body, and lower body features are weighted with weights of 0.2, 0.4, and 0.4, respectively, and they are then concatenated with the global feature as the final pedestrian descriptor, as shown in Eq. (3):

$$f^{GLAD} = \left[f^{G}; f^{h}; f^{ub}; f^{lb}\right] \tag{3}$$

where f^{GLAD} denotes the final pedestrian descriptor with a 4096-dimensional vector. f^{G} represents the global feature f^{h}, f^{ub} and f^{lb} are the local features of the head, upper body, and lower body, respectively, all of which are 1,024-dimensional vectors.

3.4 GLAD NETWORK MODEL COMBINED WITH ATTENTION

The attention mechanism imitates the selective attention of human vision to different regions when making judgments. Human receives all the images that need to be judged through their eyes and perceive some unique parts, which requires additional energy to obtain some representative details of the target. Human then increases the weight of the representative information and decreases the influence of useless information to make a correct judgment. Person re-ID is also a complex visual classification task, so some studies focus on person re-ID based on attention mechanism. Insafutdinov [33] designed a network combined with hard and soft attention. Hard attention learns to find the discriminative region, and soft attention selects fine-grained features to improve the performance of person re-ID. Zhao [34] used the pose information to learn the attention mask as the local feature and then fused the global feature and the local feature to obtain the final pedestrian descriptor. The attention-embedded GLAD network described in this study is different from this attention mechanism. The current attention mechanism is to learn local and global features based on hard attention perception of discriminative regions or learning attention masks, thereby improving person re-ID accuracy. First, our method uses a key point detection network to detect local pedestrian regions and avoid the interference of excessive background noise. Then, the channel-wise attention is added to enhance the relationship between channels in feature extraction to focus on the channel features with a large amount of information in the same region. Another reason for the addition is to suppress the unimportant channel features thereby achieving an attention mechanism similar to human vision. Thus, this method can improve the discrimination of the features extracted by the network[35].

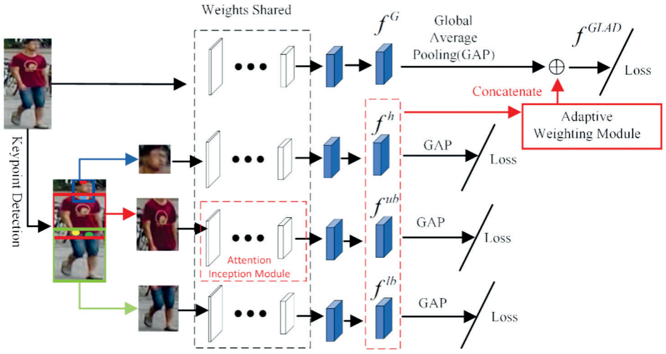


Fig. 3. The framework of the proposed network

Figure 3 shows the general framework of the proposed network. The input consists of four parts: the entire image, the head, the upper body, and the lower body. The GoogLeNet with shared weights is used as the basic network. After embedding the attention module, the fine-grained feature of each part is extracted. The three parts and global features are weighted by the adaptive weighting module to obtain the final pedestrian descriptor. Each part is studied as a separate classification task.

3.5 ATTENTION MODULE

For the GoogLeNet network, using the multi-branch structure based on the Inception module, the convolution operation will fuse multi-scale spatial information. However, the fusion process is not a focus, and the addition operation is directly used to mix the feature relationship between the channels mixed with the spatial relationship learned by the convolution kernel. The attention Inception network module proposed in this study is to separate it from this mixture so that the model directly learns the relationship between channels. The module structure is shown in Figure 4.

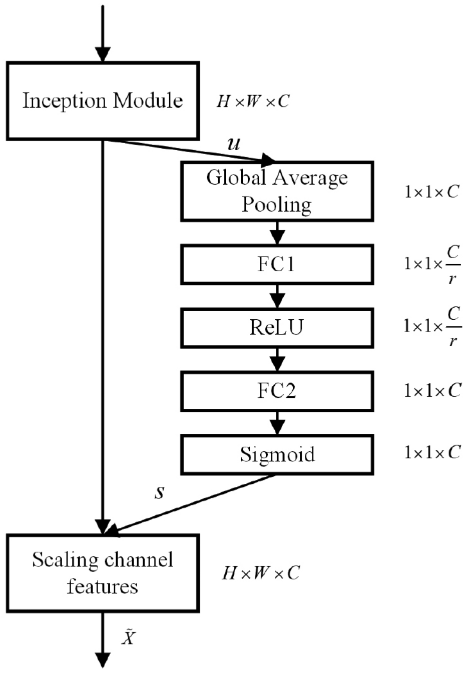


Fig. 4. Inception module with embedded attention

The entire attention module is embedded between the output of the previous Inception module and the input of the next Inception module. Assuming that the output of the Inception module u is a feature matrix of $HxWxC$ dimension. We perform a squeeze operation on u, implement GAP along the spatial dimensions HxW, and use a real number to represent the channel feature of each dimension, which has a global receptive field, as shown in Eq.(4).

$$z_c = \mathrm{F}_{sq}\left(u_c\right) = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H} u_c\left(i,j\right) \tag{4}$$

Then, the squeezed output is subjected to an excitation operation. First, a fully connected layer operation with parameters W_1 and reduction ratio r (set to 16 here) is used to perform a dimensionality-reduction mapping on the squeezed output. The reduction of the channel dimension means the decrease of calculated parameters. After a ReLU layer, nonlinear mapping is performed, and then a fully connected layer with a parameter of W_2 is used for dimension increase. Then, the vector of $1x1xC$ dimension is output, and the sigmoid function is used as the weight s between channels.

$$s = F_{ex}\left(z, \mathrm{W}\right) = \sigma\left(W_2 \delta(W_1 z)\right) \tag{5}$$

where δ represents the ReLU operation and σ represents the sigmoid function.

When the input is weighted, the channel-wise multiplication is implemented between the input feature u and the weight s, which is shown in Eq. (6):

$$\%o_c = s_c \cdot u_c \tag{6}$$

Through the sigmoid function, the elements of the vector s belong to (0,1). If the weighted value is directly used as the input of the next layer of the network, the value of the learned feature will become gradually smaller as the depth of our network increases, making the model difficult to train. Therefore, the weighted feature and the original feature are added as the input of the next layer, ensuring the update of the parameters of the neural network and also playing a vital role in highlighting important information and suppressing unnecessary information.

3.6 ADAPTIVE WEIGHTING MODULE

Given that the amount of information that can distinguish pedestrians contained in the head, upper body, and lower body is not equal, the direct concatenation will inhibit the contribution of some region features with abundant information to the re-ID accuracy. To fully use the feature information of each part, we propose an adaptive weighting module based on channel attention, as shown in Figure 5. The red arrow in Figure 3 points the position of the module embedded in the overall network.

In Figure 5, the features of the head, upper body, and lower body are marked as f^h, f^{ub}, and f^{lb} respectively, then

$$F = \left[f^h; f^{ub}; f^{lb}\right] \tag{7}$$

We consider the feature of each part as a channel and squeeze it to obtain a vector z with a global receptive field

$$z_{\mathrm{k}} = \mathrm{GAP}_{\mathrm{H}}\left(\mathrm{GAP}_{\mathrm{W}}\left(f\right)\right) \tag{8}$$

where $f \in (f^h, f^{ub}, f^{lb})$, $k \in (1,2,3)$ represents the number of pedestrian body parts, and GAP_H and GAP_W represents the height-wise and width-wise pooling, respectively.

The excitation operation is shown in Eq. (9).

$$s_k = \sigma\left(FC_2\left(\delta\left(FC_1\left(Z\right)\right)\right)\right) \tag{9}$$

where Z represents the concatenated vector, δ represents the ReLU operation, σ represents the sigmoid function, and FC represents the fully connected operation.

The weighted features are calculated as follows:

$$\widetilde{\boldsymbol{F}} = [s_1 \cdot f^h{}_; s_2 \cdot f^{ub}{}_; s_3 \cdot f^{lb}] \tag{10}$$

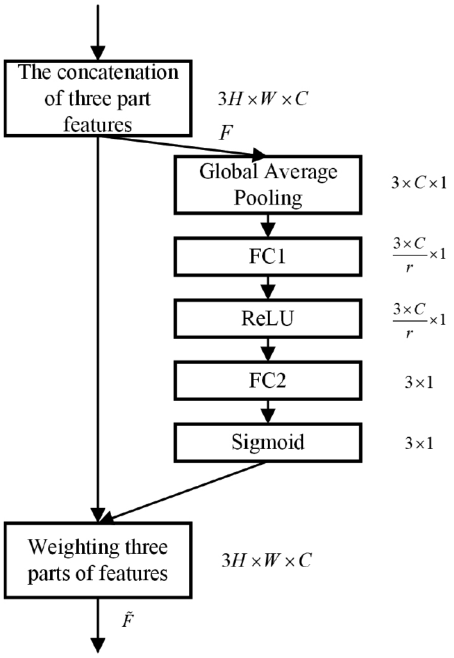


Fig. 5. Adaptive weighting module

3.7 LOSS FUNCTIONS

The network proposed in this study uses two convolutional layers as classifiers instead of fully connected layers. The first convolutional layer is used for the dimensionality-reduction. The second convolutional layer directly outputs features with N channels, corresponding to the total pedestrian number N of the training set. Then GAP is used to obtain a final N-dimensional classification vector. The experiment uses the cross-entropy loss function, which evaluates the loss through the prediction of the classification vector and the ground truth of the pedestrian.

$$\mathrm{Loss} = \sum_{\mathrm{i}=1}^{N} \text{-}\mathrm{y}^{(\mathrm{i})} \log(\hat{y}^{(i)}) \tag{11}$$

where N represents the total number of pedestrians in the dataset, $y^{(i)}$ is the ground truth of the sample, and $\hat{y}^{(i)}$ is the predicted result of the sample by the network model.

Given that the multiple branches of this network have different contents, the loss should also be weighted. According to the learned weights of different parts, the loss is weighted the same, and the total loss function is shown in Eq. (12).

$$\mathrm{L} = Loss_G + s_1 \cdot Loss_h + s_2 \cdot Loss_{ub} + s_3 \cdot Loss_{lb} \tag{12}$$

where $Loss_G$ represents the overall feature loss, s represents the weight of different parts, and $Loss_h$, $Loss_{ub}$, and $Loss_{lb}$ represent the loss value of the three part branches.

We train the entire network by minimizing the joint loss function. In the testing stage, the images are input into the network. Then, the weighted concatenation of the global feature and three part features is used as the final pedestrian descriptor. The Euclidean distance between the query image and all the images in the gallery is calculated as the similarity. On this basis, the result sequence is obtained by sorting.

4. RESULT ANALYSIS AND DISCUSSION

We mainly evaluate the proposed method on two widely used person re-ID dataset Market-1501 and DukeMTMC-reID.

In the Market-1501 dataset, each pedestrian appears under at least two cameras, and occurs between different cameras. The pedestrian bounding box is automatically marked by DPM. Among them, 12,936 images of 751 identities were used for training, and 19,732 images of 750 identities were used for testing. More than bounding boxes of 3,000 pedestrians in the query image are drawn manually, whereas the bounding boxes of pedestrians in the test set are drawn by pedestrian detection algorithm.

The DukeMTMC-reID dataset is a subset of a multi-target pedestrian tracking dataset applied to person re-ID task. All the pedestrians cross over under different cameras, sampling an image per 120 frames. The training set is a random sample of 702 identities, including 17,661 images, and 408 identities only appearing under one camera are added to the training set as interference information. The remaining 702 identities are used as the test set, which contains 16,522 images. In the test set, a query image is selected from each ID in each camera, with a total of 2,228 images.

In this study, the performance of person re-ID algorithm is evaluated by two evaluation protocols: (1) cumulative match characteristic curve (CMC) and (2) mean average precision (mAP). CMC regards person re-ID as a sorting problem. The probability that the image ranked in the first place hit the target is represented by Rank-1, which is obtained by averaging through experiments many times. mAP is the average value of average precision (AP), which is the evaluation standard when person re-ID is regarded as an image retrieval task. The formulas of AP and mAP are as follows:

$$\mathrm{AP}=\frac{\sum_{k}^{n}\left(P(k)\times B(k)\right)}{N} \tag{13}$$

where n is the total number of gallery images returned by the network, N is the number of images hitting the query target in the returned images, P(k) is the retrieval accuracy in the returned image number k, and B(k) is the indicator function. when the k-th returned gallery image hits the query target, the value of it is 1; otherwise, it is 0.

$$\mathrm{mAP}=\frac{\sum_{q=1}^{Q}\mathrm{AP}(q)}{Q} \tag{14}$$

where Q is the number of queries.

The experimental platform in this study is the Ubuntu 16.04 operating system. All experiments are conducted on a computer equipped with a Quadro p4000 GPU, Intel i7 CPU, and 32 GB memory. The GoogLeNet network, which was pre-trained on the ImageNet dataset, is used as the baseline. After embedding the attention and adaptive weighting modules, the network is trained on a person re-ID dataset. We set the maximum number of training epochs to 70 and the batch size of image pairs to 32. The initial learning rate is set to 0.0015, and the learning rate is updated to one-fifth of the original 10 epochs.

In the training process, the entire dataset images need to be extracted by the pre-trained DeeperCut network for key points. On this basis, the input is divided into three parts. The training of each sample is divided into two steps. First, the three parts of pedestrians are used for classification training to obtain the classification loss and update the network parameter. The three parts share the weight of the feature extraction network. Given the different input sizes, the structure after the feature extraction layer will learn independent weights. In the second step, the whole pedestrian sample is input, and the adaptive weight module is used to adaptively weight the three partial features extracted in the first step. The weighted partial features are concatenated with the global feature for classification to update the parameters of the network.

Tables I and II, compare the results of the traditional algorithm (LOMO+XQDA、BoW+Kissme) and other classical deep learning algorithms which achieve the state of art performance at every research stage on Market-1501 and DukeMTMC-reID dataset, respectively. Among them, AACN, HAC algorithm and the proposed algorithm all use attention mechanism. The method proposed here can improve results on both datasets.

On the Market-1501 dataset, our algorithm is approximately 45% higher than the traditional person re-ID algorithm LOMO+XQDA and BoW+Kissme on Rank-1. Our algorithm is approximately 6% higher than the TriNet network on Rank-1. AACN, a deep learning network based on the attention mechanism, is approximately 5% and 8% lower than our algorithm on Rank-1 and mAP respectively. Our method outperforms the HAC network by approximately 2% on Rank-1 and mAP. Compared with the GLAD network before modification, the Rank-1 and mAP in this study are improved by approximately 1%.

On the DukeMTMC-reID dataset, the proposed algorithm improves by approximately 14% compared with a generative adversarial network (GAN) on Rank-1 and approximately 5% compared with SVDNet. Compared with the attention network, the mAP and Rank-1 of our network are higher than AACN by approximately 5%, while the mAP and Rank-1 of HAC attention network are approximately 2% lower than our study, respectively. For the GLAD network before modification, compared with the Rank-1 and mAP value, the proposed method increases both by approximately 1%.

Methods	Rank-1(%)	mAP(%)
LOMO+XQDA[1]	43.8	22.2
BoW+Kissme[38]	44.4	20.8
Spindle[34]	76.9	-
GAN[41]	78.12	56.25
SVDNet[40]	82.3	62.1
TriNet[39]	84.9	69.1
AACN[37]	85.9	66.87
HAC[36]	89.00	71.25
GLAD[31]	89.9	73.9
Ours	91.78	74.87

Table I Comparison of Market-1501 results

Methods	Rank-1(%)	mAP(%)
BoW+Kissme[38]	25.1	12.2
LOMO+XQDA[1]	30.75	17.04
GAN[41]	67.68	47.13
SVDNet[40]	76.7	56.8
AACN[37]	76.84	58.25
HAC[36]	78.50	60.25
GLAD[31]	80.0	62.2
Ours	81.76	63.42

Table II Comparison of DukeMTMC-reID results

Table III shows the person re-ID performance using different pedestrian descriptors on the Market-1501 dataset, where A represents the attention module proposed in this study. If only three parts of the GLAD network are weighted as the final pedestrian descriptor, the Rank-1 and mAP accuracy of manual weighting will both be reduced by approximately 1% compared with the value of adaptive weighted. Compared with the baseline, the GoogLeNet embedding the attention has been improved by approximately 4% and 2% on Rank-1 and mAP, respectively, indicating that the attention module can extract more discriminative information and improve the re-ID performance. Based on the attention module, manual and adaptive weighting were respectively used to generate pedestrian descriptors. The performance of the former was approximately 0.5% lower than that of the latter, indicating that compared with other methods, adaptive weighting could better measure the importance of each part of the body. In summary, compared with other methods, the method proposed in this study can extract more discriminative information combined with the attention mechanism and can better weigh the importance of each part with an adaptive weighting module to improve the performance of the network model.

Descriptors	Rank-1(%)	mAP(%)
Manual weighting (W)	85.5	62.8
Adaptive weighting ($\tilde{W}$)	**86.2**	**63.4**
GoogLeNet	79.1	58.2
GoogLeNet+A	**83.2**	**60.0**
GoogLeNet+A+W	**91.16**	**72.73**
GoogLeNet+A+$\tilde{W}$	**91.78**	**74.87**

Table III Performance comparison of different descriptors in Market-1501 dataset

Figures 6(a) and 6(b) are examples of Rank-10 results of the proposed algorithm on the Market-1501 and DukeMTMC-reID datasets, respectively. The leftmost is the query target. From the left to the right, the results are arranged in order of decreasing similarity with the query image, in which the green box represents the pedestrian with the same ID and the red box represents the pedestrian with a different ID. The figures clearly illustrate that the proposed algorithm can correctly identify the front and the misaligned images of the target (such as Rank-4 in Fig. 6(b)). The target pedestrian can still be correctly identified when the pedestrian is flank and back and no other pedestrian dressed is similarly, such as Rank-8 and Rank-10 in Fig. 6(b). However, the side or back image with similar clothes is easily predicted incorrectly. For example, Rank-6 in Fig. 6(a) is partially occluded, but the colors of the clothes and backpack are very similar. When the front and side images appear at the same time, the front image has a higher degree of similarity. As shown in Fig. 6(b), the similarity of Rank-9 is higher than that of Rank-10.

A comparison between the proposed algorithm and the GLAD network is shown in Table IV (see section: supplementary material). In terms of the number of parameters, the time needed to identify a person and the time needed to extract all pedestrian features in the test set. The three indexes all increase to different degrees. According to Tables I and II, compared with the GLAD network, the proposed network model has a slight increase in the three indexes, but the re-ID accuracy has been significantly improved.

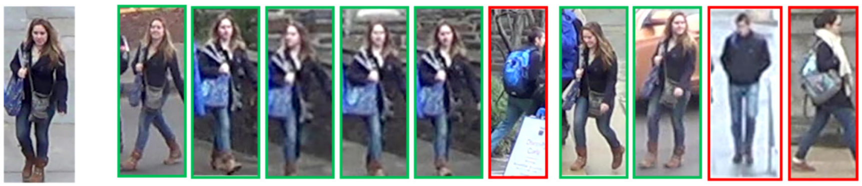


Fig. 6(a) An example of person re-identification Rank-10 in DukeMTMC-reID

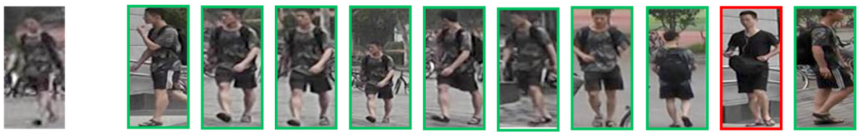


Fig. 6(b) An example of person re-identification Rank-10 in Market-1501 dataset

5. CONCLUSION

To extract additional discriminative characteristics of pedestrians in the complex monitoring environment to improve the person re-ID performance, we embeded the attention module on the basis of the GLAD network and designed an adaptive weighting module for different body parts to concatenated feature. The performance was verified on two major datasets Market-1501 and DukeMTMC-reID. The following conclusions could be drawn.

(1) The attention mechanism is integrated with the key point detection, which can effectively extract the pixel-level saliency information to obtain discriminative pedestrian features under complex changes.

(2) By introducing the adaptive weight design into the network, the weight is assigned according to the information of different body regions to utilize regional features and improve re-ID accuracy.

(3) From the experimental results, under the same dataset, the re-ID accuracy of the proposed network model is significantly better than traditional algorithms and other networks that also use attention mechanisms.

The proposed network model based on the attention mechanism and adaptive weight compensates for the lack of pixel-level saliency information of the global feature in the GLAD network. This model also solves the problem that manual weighting is not appropriate for each pedestrian block feature. When compared with the GLAD network before the improvement, the proposed network model has a slight increase in the number of parameters of the network and the run time required to identify individuals. However, the re-ID accuracy has been significantly improved with complete pedestrian features from the feature extraction aspect. The study has a good reference for person re-ID in surveillance videos.

In reality, there is a big gap in lighting, pedestrian clothing and camera style in surveillance video. The existing person re-identification data sets are far from meeting the requirements of person re-identification in actual video monitoring system. In order to further explore the above problems, it will be considered to expand the dataset with a GAN in subsequent studies to increase the suitability of the training data for actual situations. Subsequent studies will attempt to combine data enhancement with person re-identification to train end-to-end, so as to reduce the impact of image generation error on re-identification accuracy.

6. APPRECIATION

This study was funded by the National Natural Science Foundation of China under Grant No.62067006, the Gansu Provincial Science and Technology Project under Grant No. 18JR3RA104, the Industrial Support Program for Colleges and Universities in Gansu Province under Grant No. 2020C-19, the Science and technology project of State Administration of market supervision and Administration under Grant No.2019MK150, the Lanzhou Science and Technology Project under Grant No.2019-4-49, the Key Laboratory Project of Plateau Traffic Information Engineering and Control in Gansu Province under Grant No.20181103,Tianjin University-Lanzhou Jiaotong University Independent Innovation Fund Cooperation Project.

REFERENCES

[1] Dong H, Lu P, Zhong S, et al. "Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning". Neurocomputing. September 2018. Vol. 307. p.25-37.DOI: https://doi.org/10.1016/j.neucom.2018.04.013

[2] Kviatkovsky I, Adam A, Rivlin E. "Color invariants for person reidentification". IEEE Transactions on pattern analysis and machine intelligence. November 2012. Vol.35-7. p.1622-1634. DOI: https://doi.org/10.1109/TPAMI.2012.246

[3] Ma B, Su Y, Jurie F. "Covariance descriptor based on bio-inspired features for person re-identification and face verification". Image and Vision Computing. June–July 2014. Vol.32-6. p.379-390. DOI: https://doi.org/10.1016/j.imavis.2014.04.002

[4] Ojala T, Pietikainen M, Maenpaa T. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns". IEEE Transactions on pattern analysis and machine intelligence. August 2002. Vol.24-7. p.971-987. DOI: https://doi.org/10.1109/TPAMI.2002.1017623

[5] Singh C, Walia E, Kaur K P. "Color texture description with novel local binary patterns for effective image retrieval". Pattern recognition. April 2018. Vol.76. p.50-68. DOI: https://doi.org/10.1109/TPAMI.2002.1017623

[6] Matsukawa T, Okabe T, Suzuki E, et al. "Hierarchical gaussian descriptor for person re-identification". In: Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, USA: IEEE, 2016, pp.1363-1372. DOI: https://doi.org/10.1109/CVPR.2016.152

[7] Xiong M, Chen D, Chen J, et al. "Person re-identification with multiple similarity probabilities using deep metric learning for efficient smart security applications". Journal of Parallel and Distributed Computing. October 2019. Vol.132. p.230-241. DOI: https://doi.org/10.1016/j.jpdc.2017.11.009

[8] Wu W, Tao D, Li H, et al. "Deep Features for Person Re-Identification on Metric learning". Pattern Recognition. May 2020. DOI: https://doi.org/10.1016/j.patcog.2020.107424

[9] Apurva B, Shishir KS. "A survey of approaches and trends in person re-identification". Image and Vision Computing. April 2014. Vol.32-4. p.270-286. DOI: https://doi.org/10.1016/j.imavis.2014.02.001

[10] Hnin TC, Sinchai K, Krongthong W. "Sleep apnea detection using deep learning". Tehnicki Glasnik-Technical Journal. December 2019. Vol. 13-4. p.261-266.DOI: https://doi.org/10.31803/tg-20191104191722

[11] Song WR, Zhao QQ, Chen C, et al. "Survey on pedestrian re-identification research". CAAI transactions on intelligent systems. December 2017. Vol.12-6. p.770-780. DOI: https://doi.org/10.11992/tis.201706084

[12] Li YJ, Zhuo L, Zhang J, et al. "A Survey of Person Re-identification". Acta Automatica Sinica. September 2018. Vol.44-9. p.1554-1568. DOI: https://doi.org/ 10.16383/j.aas.2018.c170505

[13] Luo H,Jiang W,Fan X,et al. "A Survey on Deep Learning Based Person Re-identification". Acta Automatica Sinica. November 2019. Vol.45-11. p.2032-2049. DOI: https://doi.org/10.16383/j.aas.c180154

[14] Ding SY, Lin L, Wang G, et al. "Deep feature earning with relative distance comparison for person re-identification". Pattern Recognition. October 2015. Vol.48-10. p.2993-3003. DOI: https://doi.org/10.1016/j.patcog.2015.04.005

[15] Chen WH, Chen XT, Zhang JG, et al. "A multi-task deep network for person re-identification". In: Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, USA: AAAI Press, 2017.p.3988-3994

[16] Zheng ZD, Zheng L, Yang Y. "A discriminatively learned CNN embedding for person reidentification". ACM Transactions on Multimedia Computing, Communications, and Applications. December 2017. Vol.14-1. p.13. DOI: https://doi.org/10.1145/3159171

[17] Sun YF, Zheng L, Yang Y, et al. "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)". In: Proceedings of the European Conference on Computer Vision, Berlin, Germany: Springer, 2018, pp.480-496. DOI: https://doi.org/10.1007/978-3-030-01225-0_30

[18] Saquib SM, Schumann A, Eberle A, et al. "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA: IEEE, 2018, pp.420-429. DOI: https://doi.org/10.1109/CVPR.2018.00051

[19] Li W, Zhao R, Xiao T, Wang XG. "DeepReID: deep filter pairing neural network for person re-identification". In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, USA: IEEE, 2014, pp.152–159. DOI: https://doi.org/10.1109/CVPR.2014.27

[20] Ahmed E, Jones M, Marks T K. "An improved deep learning architecture for person re-identification". In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA: IEEE, 2015, pp.3908–3916. DOI: https://doi.org/10.1109/CVPR.2015.7299016

[21] McLaughlin N, Martinez Del Rincon J, Miller P. "Re-current convolutional network for video-based person re-identification". In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA: IEEE, 2016, pp.1325–1334. DOI: https://doi.org/10.1109/CVPR.2016.148

[22] Varior RR, Shuai B, Lu J, et al. "A siamese long short-term memory architecture for human re-identification". In: Proceedings of the European Conference on Computer Vision, Amsterdam, USA: IEEE, 2016, pp.135-153. DOI: https://doi.org/10.1007/978-3-319-46478-7_9

[23] Liu H, Feng J, Qi M, et al. "End-to-end comparative attention networks for person re-identification". IEEE Transactions on Image Processing. July 2017. Vol.26-7. p.3492-3506. DOI: https://doi.org/10.1109/TIP.2017.2700762

[24] Yan YC, Ni BB, Song ZC, et al. "Person re-identification via recurrent feature aggregation". In: Proceedings of the 14th European Conference on Computer Vision, Amsterdam, USA: Springer, 2016, pp.701–716. DOI: https://doi.org/10.1007/978-3-319-46466-4_42

[25] Khatun A, Denman S, Sridharan S, et al. "Joint identification-verification for person re-identification: A four stream deep learning approach with improved quartet loss function". Computer Vision and Image Understanding. August 2020. Vol.197-198. p.102989. DOI: https://doi.org/10.1016/j.cviu.2020.102989

[26] Erbeti A, Yusuf SA. "End-to-End Training of CNN Ensembles for Person Re-Identification". Pattern Recognition. August 2020. Vol.104 p.107319. DOI: https://doi.org/10.1016/j.patcog.2020.107319

[27] Huang X, Xu J,Guo G,et al. "Real-Time Pedestrian Reidentification Based on Enhanced Aggregated Channel Features". Laser & Optoelectronics Progress. January 2017. Vol.54-9. p.119-127. DOI: https://doi.org/10.3788/LOP54.091001

[28] Wu D,Fang M,Fu F. "Person Re-Identification Net of Spindle Net Fusing Facial Feature". Journal of Northwestern Polytechnical University. October 2019. Vol.37-5. p.1070-1076. DOI: https://doi.org/10.1051/jnwpu/20193751070

[29] Franco A, Oliveira L. "Convolutional covariance features: Conception, integration and performance in person re-identification". Pattern Recognition. January 2017. Vol.61. p.593-609. DOI: https://doi.org/10.1016/j.patcog.2016.07.013

[30] Barbosa IB, Cristani M, Caputo B, et al. "Looking Beyond Appearances: Synthetic Training Data for Deep CNNs in Re-identification". Computer Vision and Image Understanding. February 2018. Vol.167. p.50-62. DOI: https://doi.org/10.1016/j.cviu.2017.12.002

[31] Wei L, Zhang S, Yao H, et al. "GLAD: Global–local-alignment descriptor for scalable person re-identification". IEEE Transactions on Multimedia. April 2019. Vol.21-4. p.986-999. DOI: https://doi.org/10.1109/TMM.2018.2870522

[32] Li Y, Cao X, Geng X. "A novel intelligent assessment method for SCADA information security risk based on causality analysis". Cluster Computing. October 2019.Vol.22, p.5491–5503.DOI: https://doi.org/10.1007/s10586-017-1315-4

[33] Insafutdinov E, Pishchulin L, Andres B, et al. "Deepercut: A deeper, stronger, and faster multi-person pose estimation model". In: European Conference on Computer Vision, Amsterdam, USA: Springer, 2016, pp.34-50. DOI: https://doi.org/10.1007/978-3-319-46466-4_3

[34] Zhao HY, Tian MQ, Sun SY, et al. "Spindle net: Person re-identification with human body region guided feature decomposition and fusion". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA: IEEE, 2017, pp.1077-1085. DOI: https://doi.org/10.1109/CVPR.2017.103

[35] Yang L, Liu J. "TuningMalconv: Malware Detection with Not Just Raw Bytes". IEEE Access. August 2020. Vol.8. p.140915-140922. DOI: https://doi.org/10.1109/ACCESS.2020.3014245

[36] Li W, Zhu X, Gong S. "Harmonious attention network for person re-identification". In: Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, USA: IEEE, 2018, pp.2285-2294. DOI: https://doi.org/10.1109/CVPR.2018.00243

[37] Xu J, Zhao R, Zhu F, et al. "Attention-aware compositional network for person re-identification". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA: IEEE, 2018, pp.2119-2128. DOI: https://doi.org/10.1109/CVPR.2018.00226

[38] Zheng L, Shen L, Tian L, et al. "Scalable person re-identification: A benchmark". In: Proceedings of the IEEE international conference on computer vision, Santiago, Chile: IEEE, 2015, pp.1116-1124. DOI: https://doi.org/10.1109/ICCV.2015.133

[39] Chen W, Chen X, Zhang J, et al. "Beyond triplet loss: a deep quadruplet network for person re-identification". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, USA: IEEE, 2017, pp.403-412. DOI: https://doi.org/10.1109/CVPR.2017.145

[40] Sun Y, Zheng L, Deng W, et al. "Svdnet for pedestrian retrieval". In: Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy: IEEE, 2017, pp.3800-3808. DOI: https://doi.org/10.1109/ICCV.2017.410

[41] Zhong Z, Zheng L, Zheng Z, et al. "Camera style adaptation for person re-identification". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA: IEEE, 2018, pp.5157-5166. DOI: https://doi.org/10.1109/CVPR.2018.00541

SUPPLEMENTARY MATERIAL

https://www.revistadyna.com/documentos/pdfs/_adic/9981-1.pdf