Journal of New Media
DOI:10.32604/jnm.2020.014278
Review

Review of Image-Based Person Re-Identification in Deep Learning

Junchuan Yang*

School of Computer & Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
*Corresponding Author: Junchuan Yang. Email: 20181221019@nuist.edu.cn
Received: 10 September 2020; Accepted: 11 September 2020

Abstract: Person Re-identification (re-ID) is a hot research topic in the field of computer vision now, which can be regarded as a sub-problem of image retrieval. The goal of person re-ID is to give a monitoring pedestrian image and retrieve other images of the pedestrian across the device. At present, person re-ID is mainly divided into two categories. One is the traditional methods, which relies heavily on manual features. The other is to use deep learning technology to solve. Because traditional methods mainly rely on manual feature, they cannot adapt well to a complex environment with a large amount of data. In recent years, with the development of deep learning technology, a large number of person re-ID methods based on deep learning have been proposed, which greatly improves the accuracy of person re-ID.

Keywords: Person re-identification; deep learning; video surveillance system

1 Introduction

Public security has always been an important part of social security, among which pedestrian video surveillance technology as an important means has been widely concerned in recent years. Due to the continuous improvement of the quality of video surveillance equipment and the continuous decrease of the cost, video surveillance has covered every corner of the city's public places. With the continuous application and construction of surveillance video, it plays an important role in the practice of deterring crimes and maintaining the stability of social security. In addition, pedestrian video is also used in criminal investigation and plays a great role. Because of the large number of cameras, it is difficult to detect the video manually. Moreover, the background of each public place is complex, such as a series of problems with light, shielding, different angles, etc., all of which make person re-ID a great challenge. Therefore, it has become a research hotspot in recent years to solve the problems in the field of video surveillance by using computer vision as the leading method.

Pedestrians weighed more than recognition across the earliest can be traced back to the camera target tracking (MTMC tracking). In 2005 years, the paper [1] is mainly discussed in across the camera system, when the target after the traveler lost in a camera view how the trajectory in the perspective of the other camera again associated problems, the article puts forward how to extract the pedestrian feature of creative and person re-ID of other core issues such as how to carry out similarity metric. Therefore, person re-ID is also extracted from MTMC tracking problem by researchers as an independent research subject, Dr. Zheng Liang proposed a person re-ID system [2], which can be regarded as a combination of pedestrian detection and person re-ID. With the development of deep learning technology, pedestrian detection technology has gradually become mature. At present, most of the data sets directly take the detected pedestrian images as training sets and test sets.

At present, person re-ID technology has been widely used in many scenes. In surveillance video, high-quality face pictures cannot be obtained due to different camera angles and resolutions. Therefore, in the case of the failure of face recognition technology, person re-ID technology becomes a very important

substitute technology.

Person re-ID is widely considered as a subproblem of image retrieval. It means that under the condition of given specific pedestrian image or video, videos or images taken by cameras from different locations at different time periods come to the target of the person in the line.

Prior to the emergence of deep learning technology, early person re-ID research mainly focused on manual design of better visual features and how to learn better similarity metric. However, with the extensive development of deep learning in recent years, deep learning technology has been widely applied in person re-ID task. Different from traditional methods, deep learning method can adaptively learn the feature of pedestrians in images, and also learn a similarity metric with good effect. Of course, with the development of deep learning today, the method applied to person re-ID has also experienced a process from simple to complex. In the early stage of deep learning, researchers mainly focus on learning the global feature of the whole picture, which can be divided into representational learning and metric- learning methods. However, when global features reach a peak, researchers find that local features have a great promotion effect on person re-ID, so person re-ID begins to shift to the study of local features. In recent years, due to the gradual maturity of Generative Adversarial Networks (GAN), some person re-ID methods based on GAN also arose gradually, and GAN generated satisfactory results in the expansion of data sets. At present, most person re-ID still belongs to the category of supervised learning, but researchers have also carried out extensive research on transfer learning, semi-supervised learning and unsupervised learning methods [3].

2 Approaches Based on Feature

Feature-based learning methods [4–9] are widely used in various computer vision tasks, such as video retrieval, target detection, object recognition and image retrieval. In person re-ID, feature expression can also be used to solve this problem. Though the person re-ID other ultimate goal is to match two images of similarity, but based on the feature of learning and not directly when training the network considering the similarity between images, and put the heavy person re-ID task as a classification problem or validation issues, the main feature of this approach is that the last layer of full connection network output is not the final use of the image feature of vector, but after a SoftMax activation function to calculate the characterization study loss, before a full connection layer is usually feature vector layer. Briefly speaking, the problem of person re-ID classification means that the ID or attribute of pedestrian is used as the training label to train the model. Only one picture is needed to be input at a time. However, the problem of person re-ID verification is to input two pictures at a time and let the network learn whether these two pictures belong to the same pedestrian.

The loss usually used in classification network is ID loss [4]. If each pedestrian is regarded as a category of classification problem, and the ID of the pedestrian is used as a label of training data to train the CNN network, the loss of this network is called ID loss. Assume that the training set has an N_{id} picture of M pedestrians, input the image i into the network f, and the last layer of the network outputs the vector $W_c^{kl} \in R$, for the ID of the image. Therefore, ID loss is defined as:

$$L^{id} = \frac{1}{N_{id}} \sum_i \sum_{k,\ l} s\left(\left(W_c^{kl} \right)^T x_i(l,k) \right) \tag{1}$$

$$L^{id} = \frac{1}{N_{id}} \sum_i \sum_{k,\ l} -log \frac{\left(W_c^{kl} \right)^T x_i(l,k)}{\sum_j \left(W_j^{kl} \right)^T x_i(l,k)} \tag{2}$$

where N_{id} is the number of images to use, C is the identity of the input image I, S is the SoftMax function, and W_c^{kl} is the output of the full connection layer.

With the in-depth study of the problem, it is not enough to learn a model with strong generalization ability only by using pedestrian ID information. Therefore, researchers began to use additional tagged pedestrian attribute information, such as human gender, hair, clothing and other attributes. By introducing pedestrian attribute labels to calculate the attribute loss, the trained network could not only predict

pedestrian ID but also predict various pedestrian attributes, which greatly increased the generalization ability of the network. As shown in the figure below, Fig. 1 is the method in CVPR2019 [8], which proposes a new attribute attention network (AA-NET). The paper uses ResNet-50 as the baseline network to generate the feature map X, and then the network inputs the feature map forward into three structures, namely GFN, PFN and AFN, the outputs of the three structures are combined. Like ordinary person re-ID network, GFN uses SoftMax to calculates ID loss. The PFN network, by calculating each part of the body, forms 6 ROI regions, pools each region and classifies each region is responsible for obtaining key human body attribute information in the network, and the AFN network contains two tasks: one is to generate AAM, the other is to classify human body attributes. First blocks the feature map and extracts different features from each part. Such as the upper features capture the hair, hats and other features.

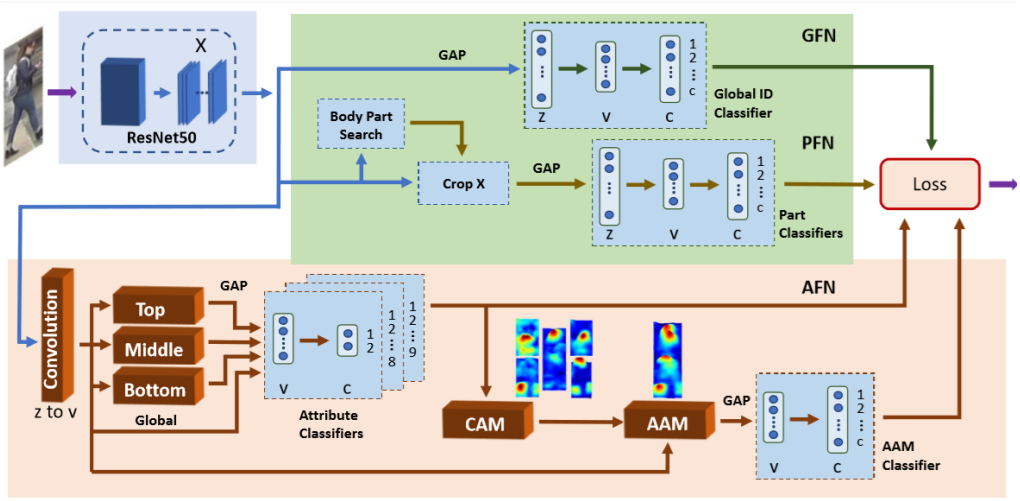

Figure 1: The pipeline of AA-Net

Verification network is another common representation learning method for person re-ID task. Different from classification networks, verification networks require input of two images at a time, which pass through a Shared CNN network and then merge the two feature vectors output by the network into an FC layer, so as to predict whether these two images belong to the same pedestrian. As shown in Fig. 2 [9] below, since the verification network is essentially a single-output dichotomy network, it is usually very inefficient to train the network only by using the verification loss, so the verification loss is often combined with the ID loss for network training.

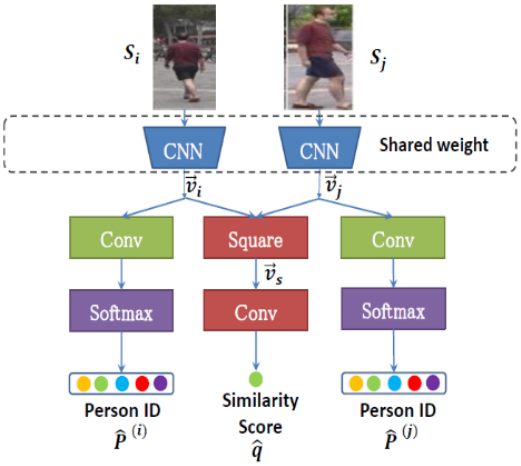

Figure 2: Verification network

3 Approaches Based on Metric-Learning

The main purpose of the method based on metric learning is that the smaller the distance between two samples of the same category, the greater the distance between samples of different categories, the better. Metric learning hopes to learn the similarity between two pedestrian images through the network. In other words, by minimizing the network's metric loss, an optimal mapping can be found to minimize the distance between the same pedestrian image from different perspectives and increase the distance between different pedestrian images. This mapping is the networks that we need to train.

The most commonly used method to metric learning loss is to contrast loss [10–13] and triplet loss [14–16]. If there are two input images *I1* and *I2*, we can obtain their eigenvectors *F1* and *F2* form the networks. And then we need to define a distance metric function, which is not unique, but any function that can describe the similarity or the difference of the eigenvectors in the eigenspace. However, in most networks, we need the metric function to be as continuously differentiable as possible. In general, we use the Euclidean distance or cosine distance of the normalized feature as the metric function.

Contrast loss [10] is used for training twin networks. Similar to the authentication network, the input to the twin network is a pair of images. These two pictures can be the same pedestrian or different pedestrian. Each pair of training pictures has a label Y, in which y = 1 means that the two pictures belong to the same pedestrian, and y = 0 means that they belong to different pedestrians.

Triplet loss [14–16] is a widely used metric of learning loss. Currently, triplet loss is widely used in a large number of studies.

Unlike the contrast loss above, the triple loss requires three input images, including a pair of positive samples and a pair of negative samples, namely three images: the original image, a positive sample image and a negative sample image.

Given three samples I,I_p,I_n, where I,I_p are samples with the same ID, and I,I_n are samples with different ID [14,15]. The goal of triplet loss is to learn a new eigenspace in which the sample pairs $I,I_p$are much smaller than the sample pairs I,I_n

Thus, the triple loss is defined as:

$$L^{tp} = \frac{1}{N_{tp}} \sum_{I,I_p,I_n} \left[d\left(f(I), f\left(I_p\right)\right) - d\left(f(I), f(I_n)\right) + \delta \right]_+ \tag{3}$$

where, δ is a super parameter, representing the Margin, to control the distance difference between samples' is the reachable triplet; And $[\,]_+ = max(\cdot, 0)$ [16].

The fundamental purpose of metric learning is to shorten the positive sample distance and lengthen the negative sample distance in the characteristic space. It can be regarded as clustering in the feature space of samples. The closer of positive samples reduces the distance within the class, while the farther of negative samples increases the distance between the classes. In the final convergence, the sample presents the effect of clustering in the feature space.

The main difference between metric learning and representation learning is that metric learning does not require the addition of an additional full connection layer for classification at the end of the network, so it is not sensitive to the number of IDs in the training set and can be applied to the network training of super-large data sets.

4 Approaches Based on Local-Feature

From the perspective of image features, the method of person re-ID can also be divided into global feature method and local feature method [17–23]. Global feature method is to let the network extract a feature from the whole image, which does not consider all local information.

However, with the complexity of pedestrian data, a single global feature cannot meet the performance requirements, so the extraction of more complex local features has become a hot research issue. Local features are those that manually or automatically allow the network to focus on key local areas and then

extract the local features of those areas.

At present, most person re-ID networks are studied by integrating global and local features [22].

As shown in Fig. 3, a two-channel network is proposed to extract global appearance features and local features respectively, and finally the final features are obtained by combining them. The appearance feature branch is a normal re-ID network, while the local feature branch is a network initialized by open-pose, which is used to extract a feature map containing pedestrian attitude information. Then, bilinear pooling is carried out by integrating the feature map of the two parts to obtain the final features.

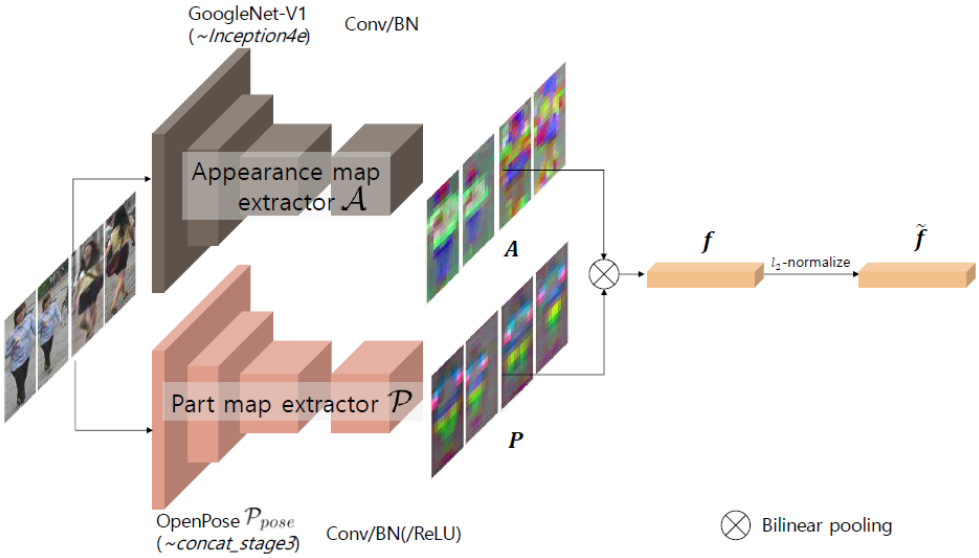

Figure 3: The pipeline of one part-based network named PABR

Similarly, good results can be obtained by simply using local features [23]. As shown in Fig. 4, VPM locates visible regions on a given image and learns their own region-level feature against these visible regions. Through such a perception method, VPM compares the image similarity between two images by focusing on the Shared area of the two images.

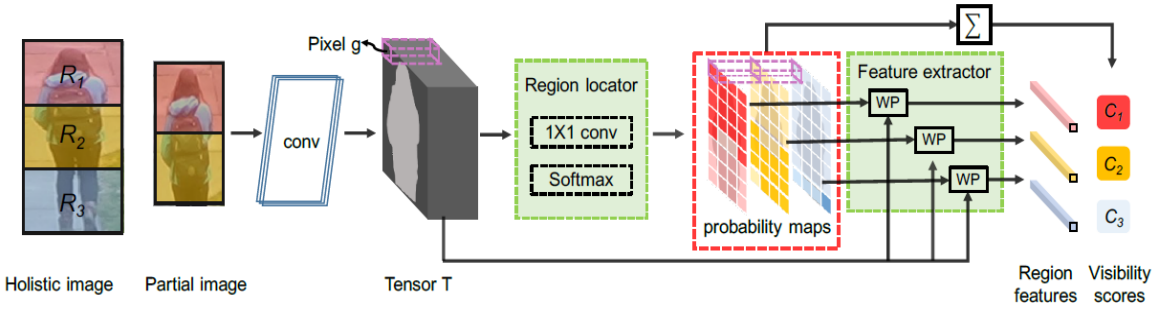

Figure 4: The pipeline of one part-based network named VPM

First, the image is divided into a complete image and a partial image. The image is divided into P = M × N blocks (in the figure above, it is divided into 3 × 1 blocks). Then, the segmented images are put into the convolutional network for feature extraction, and the corresponding feature map is obtained in a certain layer, that is, the tensor T in the figure above. Then we defined pixel g as the row vector and reduced the dimension by 1 × 1 conv. SoftMax function was used to classify g and predict its probability value. From the output of the SoftMax function above, we can get probability maps, which. The feature extractor uses the probability of g to weight g, so as to get three new weighted tensors T, and then global average pooling is performed for the three weighted tensors T, and C is the visibility score of each feature after partitioning. Throughout the network training, ID loss (cross-entropy loss) and triple loss are used.

Local feature of pedestrians has gradually been proved to be an effective feature, which can solve the problem of pedestrian diversification to a certain extent. Therefore, the fusion of global and local features gradually becomes the mainstream [23].

5 Approaches Based on Posed-Feature

With the development of person re-ID network, researchers found that in the case of multiple cameras, the walking posture of pedestrians can be used as a unique attribute to enhance the robustness of person re-ID network [24].

As shown in Fig. 5, the paper focuses on posture and Attention, and presents Attention-Aware Compositional Network (AACN) [24]. AACN network mainly includes two independent links PPA (Pose-guided Part Attention) and AFC (Attention-Aware Feature Composition).

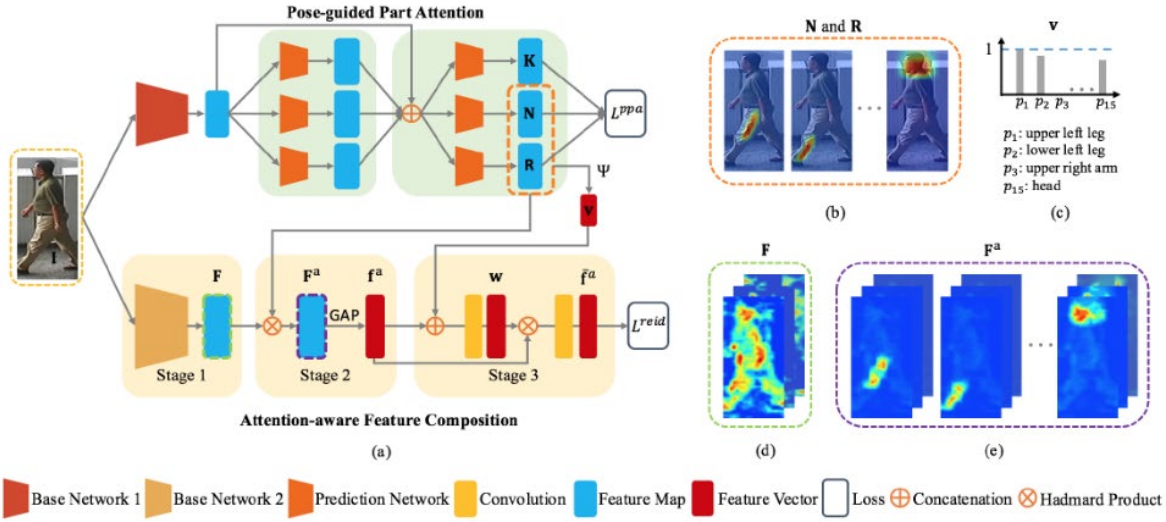

Figure 5: The pipeline of one pose-based network named AACN

AFC fuses the information output from PPA and the original global feature map information to obtain the final feature. The main role of PPA is to extract the person's attitude information and part information, and remove the background interference. And the PPA will also visually score each part. In the Fig. 5, a is the infrastructure of AACN network, b is attention visualization, c is visibility score of PPA, d and e is activation map of the network. PPA consists of three predefined posture information, one is the most original key information, one is the regular rectangular box area, and one is the joint area. Three branches are derived for the three parts information to calculate the loss.PPA can predict the location of three different types of parts on the image, and also gives visibility scores for predefined parts, which are then sent to the next network consists of three stages. The first stage is the baseline network, which is used to train a global feature network and extract the global features of images. In the second stage, the global feature map and PPA's part Attention are combined to get the attention aware feature. The third stage uses PPA to predict the visibility score of each part to get a weight vector, which is then connected with the results of the second stage to train for the final recognition feature.

6 Approaches Based on GAN

GAN has developed rapidly in recent years and is widely used in the generation of pictures. As the deep learning method relies on a large amount of training data, the current person re-ID data set is generally small. Therefore, the use of GAN to solve person re-ID task has gradually become a research hotspot [25–33].

As shown in Fig. 6 below, the paper [28] mainly proposed a new network called SPGAN. Compared with other GAN methods, SPGAN's advantage mainly lies in that this method combines the generation of

images with recognition, so as to solve the domain bias problem in person re-ID. SPGAN can transfer a recognition model trained in the source domain to a new target domain without any additional annotation, thus improving the performance of recognition in the target domain.

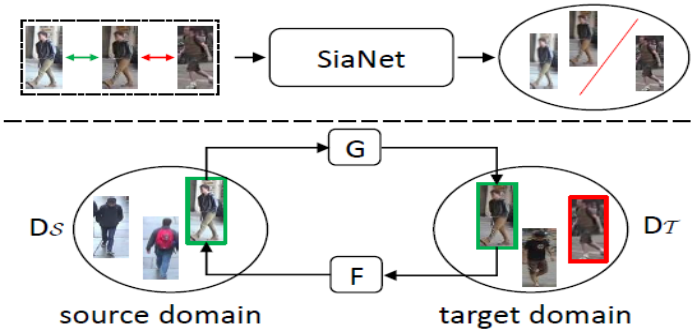

Figure 6: The pipeline of one GAN-based network named SPGAN

First, SPGAN will input the positive sample pair into Siamese network, including the original image of the source domain and the generated image of the target domain, and the negative sample is the original image of the target domain. Since the positive sample comes from other data sets, it must not appear in the target domain, that is, the natural negative sample. By comparing the losses, the positive and negative samples can be pushed aside, so that the recognition model can have better performance in the target domain, instead of the unsupervised method based on extra false label commonly used now.

In addition to the method of simply generating images, researchers have found that using attitude information can better generate images that are helpful for person re-ID [29–33]. As shown in Fig. 7, the paper [33] uses attitude information to generate images. Some data sets have little attitude change, while some data sets have a lot of attitude change, such as MRAS. Taking advantage of this feature, the paper uses GAN to create rich attitude pictures in the data set of pedestrians with little attitude change and other attitude information to train a more robust person re-ID model.

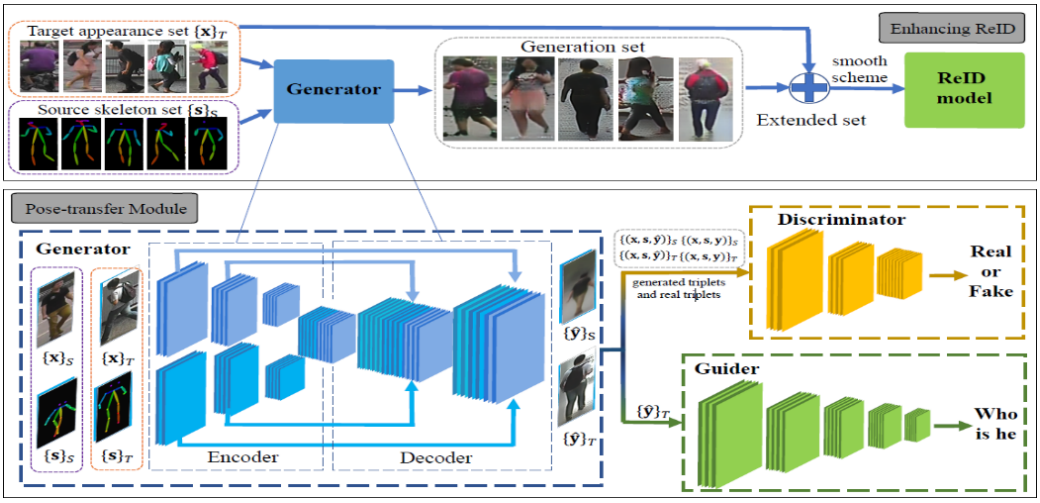

Figure 7: The pipeline of one GAN-based network named PTPR

7 Other Approaches in Person Re-ID

In addition to the above related methods of person re-ID, other methods of person re-ID also need to be introduced.

Person re-ID combined with multi-stage network [34–37]. As shown in Fig. 8, a multi-level factors network (MLFN) is proposed in this paper [37]. In the paper, it is believed that the feature of pedestrians

can be decomposed into several fine-grained feature, although the visual appearance of a person under different cameras may change dramatically due to illumination, background, camera vision and body posture. However, the judgment of people's real identity depends more on the invariable visual features or factors of visual representation. A person's appearance can be described by multiple semantic level factors, and it is very important for the discriminant factors that try to be invariant for cross-attempt matching. Each row shows two attempts by the same person. MLFN took this feature into full consideration by designing a multi-level component-decomposition network that allows the network to learn the various potentially different components of pedestrian's layer by layer. Finally, integrate the components learned from different levels at high-level.

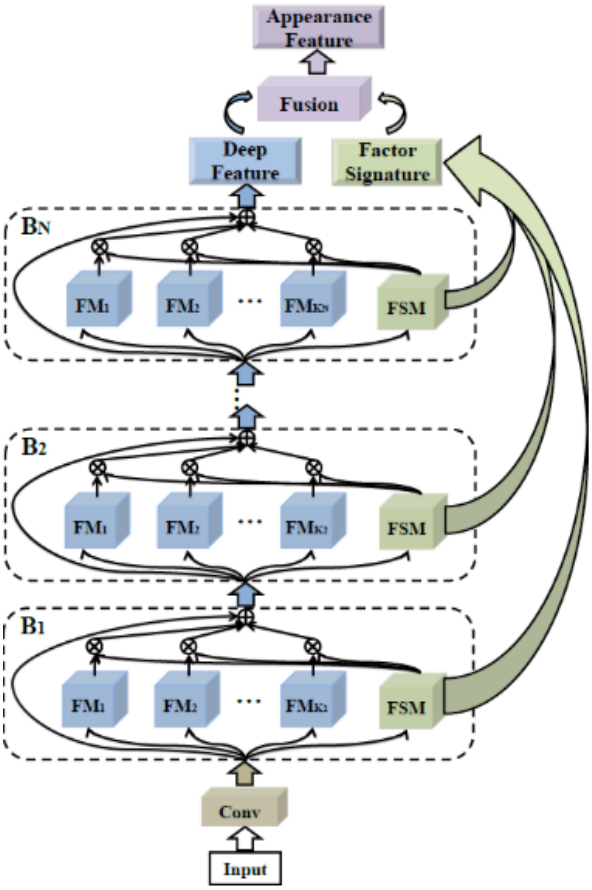

Figure 8: The pipeline of MLFN

For different images, their recognition difficulty is different. In some cases, just color statistics will suffice, while in other cases, high level features and low-level features may need to be calculated. At present, the popular person re-ID methods use the deep convolutional network to conduct a classification operation at the end of the deep network. But doing so may limit the accuracy of some samples or impose unnecessary time costs on some simple samples. In this paper [38], CNN network is segmented into blocks, and feature maps with different resolutions are obtained after each block convolution. The paper uses the residual network as the baseline and divides it into four stages, each of which represents the feature of different resolutions and different levels of abstraction. In a convolutional network layer, the shallow network is mainly concerned with the local detail features of the image. With the deepening of the network, the network layer is more and more concerned with the global abstract features. After global average pooling, the feature map output by each convolutional block (stage) will get a feature vector, and then all features will be merged after the network. The fusion method is that each feature is assigned a weight, which is obtained by using network training.

8 Datasets

Now, some of the datasets for image-based person re-Dd have been published. The commonly used datasets and their main information are shown in Tab. 1.

VIPeR [39] contains two cameras, each of which takes a picture of each person. It also provides an Angle of view for each image. Although it has been tested by many researchers, it remains one of the most challenging data sets. QMUL-iLIDS [40] is based on iLIDS-MCTS, a multi-camera CCTV system collects data sets during airport rush hours. Almost every identity has four images from two non-overlapping cameras. This data set has a scene of severe occlusion and attitude change. GRID [41] was captured by eight disconnected cameras in a busy subway station. Each identity has two images from different views, and there are more images in the Gallery than in the Probe and the image quality of this dataset is poor.

There were 385 tracks from Camera A and 749 tracks from Camera B in The PRID2011 [42] dataset, among which only 200 people appeared in both cameras. There is also a single-lens version of this dataset that contains randomly selected snapshots. Some tracks are not well synchronized, which means people can "jump" between successive frames.

The CUHK01 [43] dataset contains two images for each identity from each camera. The data set has a pair of disjoint cameras and the image quality is good. CUHK02 [44] is an extended data set from CUHK01.In addition to the CUHK01 camera pair, there are four camera pair Settings.

CUHK03 [45] includes 14,097 images from 1,467 identities observed from 2 different cameras. There are two ways to obtain the annotations: manually labeled and DPM detected bounding boxes. For each camera, each person selects one image as the probe and we choose the rest images to construct the gallery set. The labelled dataset contains 767 identities, 7,368 training, 5,328 gallery and 1,400 query images while the detected set includes 767 identities, 7,365 training, 5,332 gallery and 1,400 query images.

Table 1: All datasets are faced with some practical challenges: Disappearing some parts of person due to occlusions, changes in light and viewpoint, or bounding box errors because of the object detectors

Dataset	Re#time	Identities#	Cameras#	Images#	Label method
VIPeR	2007	632	2	1264	Hand
QMUL-iLIDs	2009	119	2	476	Hand
GRID	2009	1025	8	1275	Hand
PRID2011	2011	934	2	24541	Hand
CUHK01	2012	971	2	3884	Hand
CUHK02	2013	1816	10	7264	Hand
CUHK03	2014	1467	10	13164	Hand/DPM
Market-1501	2015	1501	6	32217	Hand/DPM
MARS	2016	1261	6	1191003	DPM
DukeMTMC-reID	2017	1812	8	36441	Hand
Airport	2017	9651	6	39902	ACF
MSMT17	2018	4101	15	126441	Faster RCNN

The Market-1501 [46] dataset includes 1,501 different identities of 32,668 images observed from six cameras with overlapping and one camera is low-resolution, five cameras are high-resolution. Following the same setting in PCB, 751 IDs with 12,936 images are allocated for training and the rest 750 IDs with 19,732 gallery images and 3,368 query images building the testing set. Later in the ICCV 2015 release version, 500K distractors are integrated to make this dataset really large scale.

The MARS [47] (Motion analysis and re-identification set) dataset is an extended version of the Market1501 dataset. This was the first large-scale video-based human re-ID data set. Because all bounding boxes and tracks are automatically generated, it contains interference, and each identity may have multiple tracks. Pre-computed depth features can also be used on the site.

The DukeMTMC-reID [48,49] consists of 1,404 identities, 2,228 queries, 17,661 gallery images, and 16,522 training images captured from 8 high-resolution cameras. The training set is randomly selected from 702 identities and the rest 702 pedestrians are utilized for testing. In addition, 408 additional dis-related identities are regarded as distractors.

MSMT17 [50] is a new person re-ID dataset, which includes 4,101 pedestrians and 126,441 bounding boxes. Different from other datasets, MSMI17 is randomly divided according to the ratio of training and testing 1:3. The training set includes 32,621 images with 1041 identities, while the testing set includes 93,820 images with 3060 identities.

9 Conclusion

From the analysis of domestic and foreign research status, it can be concluded that the introduction of convolutional neural network greatly improves the accuracy of person re-ID, but at the same time, it also creates new problems. First of all, due to the feature of the convolutional neural network itself, the deep neural network always ignores some important features of the shallow layer when abstracting features, which are very necessary for person re-ID. Secondly, with the local feature has been applied gradually, and at the same time, the importance of the global feature was repeatedly ignored, so the study should find a middle point, to solve the application of the relationship between the global features and local feature.

Funding Statement: The author(s) received no specific funding for this study

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study

References

[1] W. Zajdel, Z. Zivkovic and B. J. A. Krose, "Keeping track of humans: Have I seen this person before?" in *Proc. ICRA*, 2006.

[2] L. Zheng, Y. Yang and A. G. Hauptmann, "Person re-identification: Past, present and future", arXiv preprint arXiv:1610.02984, 2016.

[3] Y. Sun, L. Zheng, Y. Yang, Q. Tian and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)" in *Proc. ECCV*, 2018.

[4] M. Y. Geng, Y. W. Wang, Y. Xiang and Y. H. Tian, "Deep transfer learning for person re-identification," arXiv preprint arXiv:1611.05244, 2016.

[5] Y. T. Lin, L. Zheng, Z. D. Zheng, Y. Wu and Y. Yang, "Improving person re-identification by attribute and identity learning," arXiv preprint arXiv:1703.07220, 2017.

[6] Y. Zhang, T. Xiang, T. M. Hospedales and H. C. Lu, "Deep mutual learning," arXiv:1706.00384, 2017.

[7] T. Matsukawa and E. Suzuki, "Person re-identification using CNN features learned from combination of attributes" in *Proc. ICPR*, pp. 2428−2433, 2016.

[8] C. P. Tay, S. Roy and K. H. Yap, "AANet: Attribute attention network for person re-identifications" in *Proc. CVPR*, pp. 7134–7143, 2019.

[9] J. Lv, W. Chen and Q. Li, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns" in *Proc. CVPR*, pp.7948-7956, 2018.

[10] R. R. Varior, M. Haloi and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *Proc. ECCV*, pp. 791−808, 2016.

[11] R. R. Varior, B. Shuai, J. W. Lu, D. Xu and G. Wang, "A Siamese long short-term memory architecture for human reidentification," in *Proc. ECCV*, pp. 135−153, 2016.

[12] Y. C. Wang, Z. Z. Chen, F. Wu and G. Wang, "Person reidentification with cascaded pairwise convolutions," in *Proc. CVPR*. pp. 1470−1478, 2018.

[13] D. Cheng, T. H. Gong, S. P. Zhou, J. J. Wang and N. N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. CVPR*, pp. 1335−1344, 2016.

[14] A. Hermans, L. Beyer and B. Leibe, "In defense of the triplet loss for person re-identification," arXiv preprint arXiv:1703.07737, 2017.

[15] H. Liu, J. S. Feng, M. B. Qi, J. G. Jiang and S. C. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, no. 27, pp. 3492−3506, 2017.

[16] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," arXiv:1803.10859, 2018.

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

[18] Q. Q. Xiao, K. L. Cao, H. N. Chen, F. Y. Peng and C. Zhang, "Cross domain knowledge transfer for person re-identification," arXiv preprint arXiv:1611.06026, 2016.

[19] X. Zhang, H. Luo, X. Fan, W. L. Xiang, Y. X. Sun *et al.,* "Aligned Reid: surpassing human-level performance in person re-identification," arXiv preprint arXiv:1711.08184, 2017.

[20] H. Y. Zhao, M. Q. Tian, S. Y. Sun, J. Shao, J. J. Yan *et al.,* "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. CVPR*, pp. 907−915, 2017.

[21] L. Zheng, T. Huang, H. C. Lu and Y. Yang, "Pose invariant embedding for deep person re-identification," arXiv preprint arXiv:1701.07732, 2017.

[22] S. Yumin and J. D. Wang, "Supplementary material: Part-aligned bilinear representations for person re-identification," in *Proc. ECCV*, 2018

[23] Y. F. Sun, Q. Xu and Y. Li, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proc. CVPR*, pp.393–402, 2019.

[24] J. Xu, R. Zhao and F. Zhu, "Attention-aware compositional network for person reidentification," in *Proc. CVPR*, pp. 2119-2128, 2018.

[25] Y. J. Zhu, T. Park, P. Isola and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc*. *CVPR*, Italy: IEEE, pp. 2242−2251, 2017.

[26] Z. L. Yi, H. Zhang, P. Tan and M. L. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. ICCV*, Venice, Italy: IEEE, pp. 2868−2876, 2017.

[27] T. Kim, M. Cha, H. Kim, J. K. Lee and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," arXiv:1703.05192, 2017.

[28] W. J. Deng, L. Zheng, Q. X. Ye, G. L. Kang, Y. Yang *et al.,* "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification," arXiv:1711.07027, 2018.

[29] Y. Huang, J. S. Xu, Q. Wu, Z. D. Zheng, Z. X. Zhang *et al.,* "Multi-pseudo regularized label for generated data in person re-identification," arXiv preprint arXiv:1801.06742, 2018.

[30] X. L. Qian, Y. W. Fu, T. Xiang, W. X. Wang, J. Qiu *et al.,* "Pose-normalized image generation for person reidentification," arXiv preprint arXiv:1712.02225, 2018.

[31] Z. Zhong, L. Zheng, Z. D. Zheng, S. Z. Li and Y. Yang, "Camera style adaptation for person re-identification," arXiv:1711.10295, 2018.

[32] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever *et al.,* "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," arXiv:1606.03657, 2016.

[33] L. Ma, Q. Sun and S. Georgoulis, "Disentangled person image generation," in *Proc. CVPR*, pp. 99–108, 2018.

[34] L. An, Z. Qin and X. Chen, "Multi-level common space learning for person re-identification" *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1, 2017.

[35] Z. Wang, R. Hu and Y. Yu, "Multi-level fusion for person re-identification with incomplete marks." in *Proc. ACMMM*, ACM, 2015.

[36] S. C. Shi, C. C. Guo and J. H. Lai, "Person re-identification with multi-level adaptive correspondence models," *Neurocomputing*, no. 168, pp. 550–559, 2015.

[37] X. Chang, T. M. Hospedales and T. Xiang, "Multi-level factorization net for person re-identification," in *Proc. CVPR*, pp. 2109–2118, 2018.

[38] Y. Wang, L. Wang and Y. You, "Resource aware person re-identification across multiple resolutions," in *Proc. CVPR*, IEEE, 2018.

[39] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008.

[40] W. S. Zheng, S. G. Gong and T. Xiang, "Associating groups of people," in *Proc. BMVC*, 2009.

[41] C. C. Loy, C. Liu and S. Gong, "Person re-identification by manifold ranking." in *Proc. ICIP*, pp. 3567–3571, 2013.

[42] M. Hirzer, C. Beleznai, P. M. Roth and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Proc. SCIA*, pp. 91–102, 2011.

[43] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proc. ECCV*, 2008.

[44] W. Li, R. Zhao and X. Wang, "Human Reidentification with Transferred Metric Learning," in *Proc. ACCV*, 2012.

[45] W. Li and X. Wang, "Locally Aligned Feature Transforms across Views," in *Proc. CVPR*, 2013.

[46] W. Li, R. Zhao, T. Xiao and X. Wang, "Deep reid: Deep filter pairing neural network for person re-identification. "in *Proc.CVPR*, pp. 152–159, 2014.

[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. ICCV*, pp. 1116–1124, 2015.

[48] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su *et al.,* "Mars: A video benchmark for large-scale person re-identification," in *Proc. ECCV*, pp. 868–884, 2016.

[49] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps *et al.,* "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Transactions on Pattern Analysis and Machine Intelligenc*e, 2018.

[50] L. Wei, S. Zhang, W. Gao and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. CVPR*, 2018.