Occluded Person Re-Identification Method Based on Multiscale Features and Human Feature Reconstruction

YANGYANG LI\textsuperscript{1}, (Senior Member, IEEE), ZICHEN YANG\textsuperscript{1}, YANQIAO CHEN\textsuperscript{2}, DANQING YANG\textsuperscript{1}, RUJIAO LIU\textsuperscript{1}, AND LICHENG JIAO\textsuperscript{3}, (Fellow, IEEE)

\textsuperscript{1}Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Joint International Research Laboratory of Intelligent Perception and Computation, International Research Center for Intelligent Perception and Computation, Collaborative Innovation Center of Quantum Information of Shaanxi Province, School of Artificial Intelligence, Xidian University, Xi’an 710071, China

\textsuperscript{2}54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China

Corresponding author: Yanqiao Chen (chenyanqiao2016@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772399 and Grant 62101517, in part by the Key Research and Development Program in Shaanxi Province of China under Grant 2019ZDLGY09-05, and in part by the 111 Project.

\section*{ABSTRACT}
Occluded Re-ID task is proposed mainly because people are often occluded by various obstacles in the real world, which greatly affects the accuracy of model matching. In view of the challenge of occluded Re-ID, the main work of this paper is as follows: (i) Aiming at the incompleteness of human body under occlusion, an occluded Re-ID method based on multi-scale features is proposed. A partial human body locator is constructed by using the target detection algorithm to automatically recognize and cut partial human body in this method. Then this method designs a horizontal pyramid pooling strategy to extract multi-scale features and enhance the robustness of the model under the occlusion problem. Experiments show that, this method has better matching accuracy in the occluded Re-ID task. (ii) Aiming at the problem that it is difficult to align the local features between different people images under occlusion, an occluded Re-ID method based on human feature reconstruction is proposed. This method is an unaligned method, which uses sparse representation to reconstruct human body features. Difficult sample triplet loss function was improved by using human feature reconstruction distance and the proportion of similar parts to matching correlation was increased. Experiments show that this method can effectively improve the occlusion resistance of the model.

\section*{INDEX TERMS}
Person re-identification, multi-scale features, human feature reconstruction.

\section*{I. INTRODUCTION}
There are a variety of sub-tasks in the field of computer vision, and the person Re-ID [1], [2] (Person Re-identification) task is one of its basic tasks. Generally, previous person Re-ID methods [3], [4], [5], [6] used the entire person body to design matching models. However, in the actual situation, peoples are often occluded by various obstacles, so the matching model based on complete peoples is not suitable for such scenes.

The deficiencies of existing methods in solving the problem of occluded person Re-ID mainly come from the following aspects [9], [10]:

(1) The incompleteness of the human body under occlusion. The early representation-based learning and metric learning methods mainly focused on human body matching, but could not extract effective person features well under occlusion. Some person Re-ID methods rely on manual clipping to obtain partial human body images that are not covered. Manual clipping is time-consuming and laborious, and also introduces human bias to the clipping results.

(2) It is difficult to align local features between different person images under occlusion. Although the person Re-ID
method based on local features has better generalization ability for occlusion, it often takes a lot of time to design and train local feature alignment, and the model effect depends heavily on the degree of alignment.

In view of the above difficulties and challenges, this paper proposes corresponding solutions, mainly including the following aspects:

1. A multi-scale features method is proposed. Aiming at the incompleteness of human body under occlusion, this method uses target detection algorithm to construct a partial human body locator, which can automatically identify and clip parts of the human body, and designs a horizontal pyramid pooling strategy, solves the problem of multi-scale input and simultaneously integrates multi-scale person features at the same time, to extract more effective the person characteristics, enhance the robustness of the model. This method can extract a more robust person Re-ID model without manual clipping.

2. A method for reconstruction of human body features is proposed. Aiming at the problem that it is difficult to align the local features between different people images under occlusion, this method proposed the human feature reconstruction distance to improve the difficult sample triplet loss function. Sparse representation method was used to reconstruct human body features, and unaligned method was used to increase the proportion of similar parts to match correlation, which improved the anti-occlusion performance of the model. This method does not need to rely on strict local feature alignment, and uses allocation-free method to increase the correlation of matching.

II. RELATED WORK

In the early days, person Re-ID was widely referred to as multi-camera tracking. In 1997, Huang and Rassel [23] proposed a Bayesian formula to predict the posterior probability of objects appearing in cameras based on features found in images obtained by other camera devices. Person Re-ID was formally proposed in 2005, originally based on a simple question, “Have I seen this person before?”. In this paper, Zajedel [24] et al. defined a dynamic Bayesian network to encode the probabilistic relationship between tags and features from tracks for person identification. Subsequently, in mid-2006, person Re-ID was separated from multi-camera tracking [25] and began as an independent computer vision task. Until 2014, deep learning was successfully applied to person Re-ID task for the first time [3], [26] and has become a popular direction now.

Early convolutional neural network [27] was applied to person Re-ID tasks, and most methods mainly considered the design of complete human body matching model. Person Re-ID studies based on metric learning and representation learning [3], [4], [5], [6], [7] were mainly concerned in this period, but these methods did not consider the noise problem in real scenes. The performance of the shaded person Re-ID task was mediocre, and the effective features of people could not be extracted well. Li et al. [8] formulate a novel Harmonious Attention CNN (HACNN) model for joint learning of soft pixel attention and hard regional attention along with simultaneous optimisation of feature representations, dedicated to optimise person Re-ID in uncontrolled (misaligned) images.

Local features-based methods [10], [11], [12], [13], [14], [15], [16] usually generate local features of various body parts by using human body posture estimation to generate corresponding key points or roughly horizontal division, so as to make them more robust to person Re-ID problems containing a lot of noise information in the real world. Miao et al. [10] propose a pose guided feature alignment (PGFA) method to match the local patches of probe and gallery images based on the human semantic key-points. Sun et al. [11] designed a PCB (Part-based Convolutional Baseline) method, which uniformly divides the human body into multiple horizontal parts to learn partial human body features [12]. Similarly, Zhao et al. [13] decomposed human body into discriminant regions for human body matching, calculated the representation of these regions accordingly, and aggregated the similarity calculated between a pair of query images and corresponding regions of gallery images into the overall matching score. On the other hand, Su et al. [14] propose a PDC (Poor-driven Deep Convolutional Model), which improves the learned attitude information and integrates the global and local features of the human body for model matching. Zhao et al. [16] proposed a new type of convolutional neural network based on multi-stage feature decomposition and tree structure feature fusion guided by human body regions, called Spindle Net. Although local feature matching is considered in these models, they ignore that the difficulty of occluded person Re-ID in the real environment lies in incomplete human body information and spatial imbalance. Therefore, such methods often rely heavily on the local alignment of images or the strict alignment of human key points. When the images in the query set are occluded, the results of methods based on local features are poor.

Partial person Re-ID is a method to match the local image with the overall image of the library. The reason why this problem is raised is that in the complex real world environment, the camera often cannot capture the complete human body. Therefore, some scholars manually crop the occluded human body image and retain partial human body for model matching training. Zheng et al. [17] proposed a local Matching model named AMC (Ambiguity-sensitive Matching Classifier) and introduced an SWM (Sliding Window Matching), which can provide complementary spatial layout information. However, AMC and SWM are quite expensive to calculate because most features are double-counted and do not undergo further acceleration. He et al. [18] proposed an unaligned matching model named DSR (Deep Spatial Feature Reconstruction), which is faster than SWM method. He et al. further utilize a dictionary learning based Spatial Feature Reconstruction (SFR) to match different sized feature maps for the Partial Re-ID problem. Fan et al. [19] use a spatial-channel parallelism network (SCPNet) to encode
part features to specific channels and fuse the holistic and part features to get discriminative features. Wang et al. [20] presents a framework utilizing key-points estimation to learn high-order relation information (HOReID) for discriminative features and human-topology information for robust alignment. Jia et al. [21] presents Matching on Sets (MoS) that positions occluded person re-ID as a set matching task without requiring spatial alignment. Tan et al. [22] presents a novel person re-identification model named Multi-Head Self-Attention Network (MHSA-Net), to prune unimportant information and capture key local information from person images. Although partial person Re-ID to some extent solves the problem of partial human body matching in the case of occlusion, such methods rely heavily on manual cropping of the occluded image to match a partial human body. Manual cutting is often time-consuming and laborious, and it will introduce artificial bias to the cutting results. Different from these works, this paper considers how to automatically crop the occluded image to generate partial human body.

III. PROPOSED APPROACH
A. MULTI-SCALE FEATURES METHOD
1) PARTIAL HUMAN BODY LOCATOR
In this paper, partial human body locator is proposed to be used in the occluded person Re-ID task. R-CNN [29], [30], [31], [32] series algorithms Faster R-CNN [31] are used to locate partial human bodies and automatically cut shielding objects, so that the model only focuses on the visible partial human bodies and is used for subsequent model matching training. As shown in Fig. 1, a schematic diagram of partial human body locator is given:

![FIGURE 1. Schematic diagram of partial human body locator.](image1)

2) HORIZONTAL PYRAMID POOLING STRATEGY
In the process of using partial human body image and complete human body image matching, there will be some problems, such as unfixed input, loss of human body information and distortion of position information in multiscale images. As shown in Fig. 2, vertical division of human body images in traditional pyramid pooling will destroy the features of human body parts, while horizontal division can better retain the integrity of local features.

Therefore, considering the particularity of human body images, horizontal feature blocks of 1 × 1, 2 × 1 and 3 × 1 are used in this paper to divide feature maps and obtain horizontal feature blocks of different scales. Then input to the fully connected layer to train its loss function. The specific partitioning process is shown in Fig. 3:

In the process of horizontal pyramid pooling as shown in Fig. 3, due to the particularity of feature map blocks in person Re-ID task, the size of feature map after pooling may be inconsistent with the expected feature map by using the pooling kernel and step calculation method in spatial pyramid pooling [33] under general conditions. In this regard, the influence of filling layer needs to be considered. If the feature block is $n_h \times n_w$ (1 × 1, 2 × 1, 3 × 1 in this paper), the size of the nucleus, step size and filling layer of spatial pyramid pooling can be calculated as follows [33]:

$$
\begin{align*}
K_h &= \left\lfloor \frac{h_{in}}{n_h} \right\rfloor K_w = \left\lfloor \frac{w_{in}}{n_w} \right\rfloor \\
S_h &= \left\lfloor \frac{h_{in}}{n_h} \right\rfloor S_w = \left\lfloor \frac{w_{in}}{n_w} \right\rfloor \\
p_h &= \frac{k_h \times n_h - h_{in} + 1}{2} p_w = \frac{k_w \times n_w - w_{in} + 1}{2}
\end{align*}
$$

where, $K_h$, $S_h$, $p_h$ represent the height of the pooled core, the step length and filling amount of the height direction respectively; $K_w$, $S_w$, $p_w$ represent the width of the pooled core, the step length and filling amount of the width direction respectively. $K_h$ and $S_h$, $K_w$ and $S_w$ are calculated using the same formula, $h_{in}$ and $w_{in}$ are the size of the feature map. According to the number of different horizontal pyramid pools, the required pool core, step size and fill amount can be calculated.

B. HUMAN FEATURE RECONSTRUCTION
1) HUMAN FEATURE RECONSTRUCTION DISTANCE
As shown in Fig. 4, the image that input to model for matching training is actually a partial human body image and a complete human body image. It can be seen that in the actual model training process, partial human body image is only locally similar to the complete image in essence, but due to the multi-scale problem, it is difficult to carry out the
alignment and matching of local features. Therefore, if we simply calculate the similarity measure distance between the features of two human body images in the loss function, a large amount of edge background information noise will be introduced.

![Schematic diagram of matching partial human body and complete human body.](image)

**FIGURE 4.** Schematic diagram of matching partial human body and complete human body.

In order to reduce the impact of irrelevant edge noise information on person matching, increase the proportion of similar parts to matching correlation, and improve the occlusion resistance of the model, the idea of image sparse representation [34] is adopted in this paper. The idea of this method is to find the sparse representation coefficients of partial human body samples to be tested. In this way, the complete human body image can be reconstructed for some human body samples by sparse representation coefficient. After the reconstructed image features, the samples to be tested can be restored as much as possible, and then the spatial feature error between them can be calculated to obtain the reconstruction distance of human body features. As shown in Fig. 5, the specific realization process of human feature reconstruction distance is introduced:

![Schematic diagram of calculation process of human feature reconstruction distance.](image)

**FIGURE 5.** Schematic diagram of calculation process of human feature reconstruction distance.

Suppose that given a pair of images, one is a partial human body image \(I\) and the other is a complete human body image \(J\). Wherein, \(x\) and \(y\) respectively represent the feature images of partial human body image \(I\) and complete human body image \(J\) extracted by ResNet-50 [28], and their formulas are as follows:

\[
\begin{align*}
    x &= \text{conv}(I, \theta) = w_x \times h_x \times d \\
    y &= \text{conv}(J, \theta) = w_y \times h_y \times d
\end{align*}
\]  
(2)

where, \(X\) and \(Y\) are the vectorized tensors \(w_x \times h_x \times d\) and \(w_y \times h_y \times d\). Respectively, \(w_x\) and \(h_y\) are the width and height of the feature graph, and \(d\) is the number of channels.

For sparse representation of images, in order to better describe its local features and reduce the amount of subsequent computation, dimension reduction of feature graphs \(X\) and \(Y\) obtained by convolution is required first to facilitate subsequent linear representation. In this regard, the feature graph \(x\) and \(y\) are divided into \(N\) and \(M\) blocks respectively, which are expressed as follows:

\[
\begin{align*}
    X &= [x_n]_{n=1}^N \in \mathbb{R}^{d \times N} \\
    Y &= [y_m]_{m=1}^M \in \mathbb{R}^{d \times M}
\end{align*}
\]  
(3)

where, \(N = w_x \times h_x; M = w_y \times h_y\); \(x_n\) and \(y_m\) represent feature blocks of size \(1 \times 1 \times d\).

Since partial human body image \(I\) and complete human body image \(J\) are similar in some areas, the feature matrix \(X\) of the former can theoretically be approximately linearly represented by the feature matrix \(Y\) of the latter, that is, there exists a linear representation coefficient matrix \(W\) of \(Y\) with respect to \(X\), so that:

\[
X \approx WY \quad W \in \mathbb{R}^{N \times M}
\]  
(4)

According to the above deduction, the sparse representation model of \(X\) to \(Y\) can be obtained as follows:

\[
\min_W ||X - WY||_2^2 + \beta ||W||_1
\]  
(5)

where, formula (4) is equivalent to formula (5), \(\beta\) is its regularization parameter, \(\|\cdot\|_1\) represents the 1-norm of its matrix, and \(\|\cdot\|_2\) represents the 2-norm of its matrix.

The above equivalent formula (4) and formula (5) are solved by the least square method, the linear representation coefficient matrix \(W\) of \(Y\) with respect to \(X\) is as follows:

\[
W = (Y^T Y + \beta \cdot I)^{-1} Y^T X
\]  
(6)

Therefore, the feature matrix \(X\) of partial human body image \(I\) reconstructed from the feature matrix \(Y\) of the complete human body image \(J\) can be expressed as follows:

\[
\hat{X} = WY = (Y^T Y + \beta \cdot I)^{-1} Y^T XY
\]  
(7)

According to the feature matrix \(\hat{X}\) after reconstruction of the feature matrix \(X\) of partial human body image \(I\) and the feature matrix \(Y\) of the complete human body image \(J\), the human feature reconstruction distance between partial human body image \(I\) and the complete human body image \(J\) can be calculated by the following formula:

\[
d = \frac{1}{N} ||X - \hat{X}||_2^2
\]  
(8)

2) IMPROVED DIFFICULT SAMPLE TRIPLET LOSS FUNCTION

Difficult sample triplet loss function is to select one of the most difficult positive and one of the most difficult negative samples for each picture in the batch and form a triplet with them. Triplet loss is to calculate the similarity of two images in the embedded feature space, that is, the distance in the feature space.
In this paper, sparse representation method is used to reconstruct human body image features. The similarity measure between reconstructed partial human body and complete human body image is obtained, which is defined as reconstruction distance of human body features. This distance is used to improve the Euclidean distance in the triplet loss function of difficult samples, and human body feature similarity is calculated in the reconstructed feature space. The calculation formula is as follows [35]:

\[ L_{th} = \frac{1}{P \times K} \sum_{a \in \text{batch}} (\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha)_+ \tag{9} \]

where, \( P \times K \) means that each person has \( K \) different pictures among \( P \) randomly selected people. In general, \( P \) is related to the batch size in model training, and \( K \) is generally set to 4 in literature on person Re-ID algorithm. \( a \) represents any picture in the batch; \( p \) represents a positive sample of same identities with \( a \); \( n \) denotes negative samples of different identities with \( a \); \( (Z)_+ \) stands for \( \max(Z,0) \); \( \alpha \) is the threshold, which can be set based on actual experience. \( d_{a,p} \) represents the distance between the samples to be tested and positive and negative samples, which is the similarity measure between samples.

Based on the above theory, this paper uses the improved human feature reconstruction distance, which is expressed as follows:

\[ d_{a,s} = \frac{1}{N} ||X_a - \hat{X}_s||_2^2 \tag{10} \]

**C. ALGORITHM FRAMEWORK**

Existing methods often rely on manual clipping to obtain partial human body images when people are occluded. In addition, existing methods also face difficulties in processing partial human body matching. Such as different input image sizes, incomplete human body information and distorted spatial location information due to clipping deformation. Based on the above issues, firstly, a partial human body locator is constructed by target detection method, positioning and automatic cutting the obscured human body images, keep body parts. A horizontal pyramid pooling strategy was proposed, and a feature extraction method combining global and local features was designed to obtain more descriptive features to enhance the robustness of the model and reduce the influence of occlusion. Secondly, the sparse representation method is used to construct the human feature reconstruction distance, and then the distance is used to replace the common Euclidean distance to calculate the triplet loss of difficult samples in the model. The improved human feature reconstruction distance can increase the proportion of matching correlation between partial human body and complete human body image, reduce the interference of noise information, and improve the occlusion resistance of the model. The main algorithm framework of this chapter is shown in Fig. 6:

The whole process of the algorithm is mainly divided into: The first step adopts the Faster R-CNN algorithm to construct a partial human body locator, which automatically locates and cuts the occluded human body image to generate a partial human body. In the second step, because the backbone network commonly used by most algorithms in person Re-ID is ResNet-50, for better comparison, we also use ResNet-50 as the backbone network. ResNet-50 is used as the backbone network to extract human body image features, and the lower sampling step in conv5_1 was set to 1 to retain more human body local features and details. The third step is the horizontal pyramid pooling of human body feature images, which divides human body features into different horizontal feature blocks and integrates multi-scale human body features. In the fourth step, the improved difficult sample triplet loss function and cross entropy loss function are used to train the extracted human body features respectively. In this step the similarity measure distance used for difficult sample triplet loss function is the human feature reconstruction distance.

**IV. EXPERIMENTS**

In order to verify the effectiveness of the algorithm in this paper on the occluded person Re-ID, the datasets adopted in this paper are Occluded-DukeMTMC, Partial-REID [17] and Partial-iLIDS [34]. The Occluded-DukeMTMC dataset is derived from the occluded images in the DukeMTMC-REID dataset, and the duplicate images are removed. It contains 15618 training samples, 17661 images in the gallery set and 2210 images in the occluded query set. Where, Occluded-DukeMTMC is the training dataset, The training of the model was only performed on the Occluded-DukeMTMC dataset, and the performance tests were performed on the Partial-REID, Partial-iLIDS and the Occluded-DukeMTMC dataset. In the experiment, CMC and mAP were used as evaluation indexes to evaluate performance. The software and hardware experimental environment settings in the experiment are shown in Table 1. Python is used as the programming language of this experiment, and the open Y. Li et al.: Occluded Person Re-identification Method Based on Multi-scale Features and Human Feature Reconstruction 6 VOLUME XX, 2017 source library Pytorch is used to build the convolutional neural network model in this paper.

The batch size adopted in this experiment was set as 32, epochs as 120, optimizer as ADAM, weight attenuation as 5e-4, momentum as 0.9, loss function as cross entropy loss and difficult sample triplet loss, margin as 1.2, and learning
rate as $2 \times 10^{-4}$. In partial human body locator, the Faster R-CNN algorithm is used to locate some human bodies, and the parameter settings are consistent with the Faster R-CNN algorithm. The Faster R-CNN do not use pre-train weights.

In order to verify the effectiveness of the multi-scale features method and the human feature reconstruction method, several excellent person Re-ID algorithms were selected to conduct comparative experiments with the proposed method. Part of the experimental results of the proposed method on the Occluded-DukeMTMC dataset are presented below. As shown in Fig. 7, the results in the red box are the results of incorrect matching, and the rest are the results of correct matching.

As can be seen from the results in Table 2, the last column of the table shows the training time. The accuracy of the proposed method based on multi-scale features and the method based on human feature reconstruction is significantly better than that of most methods. At the same time, compared with the recent algorithms, the training time is obviously lower than them when the accuracy is close. The algorithm model has stronger robustness and generalization ability in the shielding person Re-ID task. Compared with PCB of baseline algorithm in person Re-ID task, the performance of the proposed method in CMC and mAP is optimal. Can be seen through the contrast experiment, the proposed human feature reconstruction distance is very effective in improving the difficult sample triplet loss function which based on the multi-scale features. The new model further improves the matching accuracy of the model in the person Re-ID task, and makes the model have higher anti-occlusion performance.

As can be seen from the results in Table 3, the method proposed in this paper also performs well in some datasets of person Re-ID. In comparison with other partial person Re-ID methods, the accuracy index of the proposed method on the Partial-REID and Partial-iLIDS datasets exceeds that
of most methods. At the same time, compared with the algorithms in recent years, the proposed algorithm can also achieve a relatively good result in a shorter training time. In addition, this paper compares the performance difference between the PCB method of baseline algorithm, which is widely used in comparative experiments, and the algorithm in this paper on Partial-REID and Partial-iLIDS datasets. It can be seen that the generalization performance of the proposed method on partial person Re-ID datasets is also better than the baseline algorithm, and it is more suitable for the person Re-ID task with incomplete human body under occlusion. This further verifies the effectiveness of the proposed method. Finally, the performance difference between the multi-scale features based method with human feature reconstruction and the multi-scale features based method without human feature reconstruction is compared. It can be seen that the performance of the improved new method exceeds that of the original method on the above two datasets, and the model accuracy is improved again. It is verified that the reconstructed distance of human features is very effective to improve the occlusion resistance of the model.

Improve the quality of the detection frame when using the target detection algorithm to automatically locate part of the human body. Although the performance of the proposed method in occlusion is better than most existing methods, its matching accuracy still has a lot of room for improvement. The bottleneck of its accuracy improvement is that this paper improves the partial person Re-ID algorithm based on manual clipping, and constructs a partial human body locator to make the model automatically focus on the visible human body. But at the same time, this method also has a certain impact on the cutting results. In the case of serious occlusion, it is prone to false detection and missed detection, which hinders the further improvement of the accuracy of the model. Therefore, how to design an algorithm to make the model pay more attention to the visible parts of the human body and eliminate the interference of the occluded area is a major problem faced at present.

Finally, this paper presents the change of the loss function of this model in the training process, as shown in Fig. 8. It can be seen that the model proposed in this paper has good convergence in the training process.

V. CONCLUSION
This paper studies the occluded person Re-ID tasks facing some problems. Including the incompleteness of human body under occlusion leads to the problem that the effective human body features cannot be well extracted and the problem that it is difficult to align the local features between different human body images under occlusion. The effect of the model depends heavily on the degree of alignment. This paper proposes a occluded person Re-ID based on multi-scale features, and the difficult sample triplet loss function was improved by using human feature reconstruction distance and the proportion of similar parts to matching correlation was increased. Experiments show that the proposed method based on multi-scale features is significantly more accurate than other methods. The algorithm model has stronger robustness and generalization ability in the occluded person Re-ID task. The multi-scale features based method with human feature reconstruction is added to improve the accuracy of the model again, which further verifies that the human feature reconstruction distance is very effective to improve the anti-occlusion of the model.

ACKNOWLEDGMENT
The authors would like to express their sincere thanks to the editors and the anonymous reviewers for their valuable comments and contributions.

REFERENCES
[1] S. Gong and T. Xiang, “Person re-identification,” in Visual Analysis of Behaviour. London, U.K.: Springer, 2011, pp. 301–313.
[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[3] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Deep metric learning for person re-identification,” in Proc. 22nd Int. Conf. Pattern Recognit., Aug. 2014, pp. 34–39.

[4] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: A deep quadruplet network for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 403–412.

[5] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, “Multi-scale deep learning architectures for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5399–5408.

[6] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3908–3916.

[7] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, “Person re-identification by local maximal occurrence representation and metric learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 2197–2206.

[8] W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Feb. 2018, pp. 2285–2294.

[9] J. Zhao, Z. Chen, J. Lai, and G. Wang, “Occluded person re-identification,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2018, pp. 1–6.

[10] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, “Pose-guided feature alignment for occluded person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 542–551.

[11] 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, pp. 480–496.

[12] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, “A Siamese long short-term memory architecture for human re-identification,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 135–153.

[13] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3219–3228.

[14] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, “Pose-driven deep convolutional model for person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3960–3969.

[15] L. Zhao, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 402–419.

[16] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, “SpindleNet: Person re-identification with human body region guided feature decomposition and fusion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1077–1085.

[17] W.-S. Zheng, X. Li, Y. Zhuang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4678–4686.

[18] L. He, J. Liang, H. Li, and Z. Sun, “Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7073–7082.

[19] X. Fan, H. Luo, X. Zhang, L. He, C. Zhang, and W. Jiang, “SCPNet: Spatial-channel parallelism network for joint holistic and partial person re-identification,” in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2018, pp. 19–34.

[20] G. Wang, S. Yang, H. Liu, Z. Wang, Y. Yang, S. Wang, G. Yu, E. Zhou, and J. Sun, “High-order information matters: Learning relation and topology for occluded person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6449–6458.

[21] M. Jia, X. Cheng, Y. Zhai, S. Lu, S. Ma, Y. Tian, and J. Zhang, “Matching on sets: Conquer occluded person re-identification without alignment,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 1673–1681.

[22] H. Tan, X. Liu, B. Yin, and X. Li, “MHSA-Net: Multichannel self-attention network for occluded person re-identification.” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 21, 2022, doi: 10.1109/TNNLS.2022.3144163.

[23] T. Huang and S. Russell, “Object identification in a Bayesian context,” in Proc. IJCAI, vol. 97, 1997, pp. 1276–1282.

[24] W. Zajdel, Z. Zivkovic, and B. J. A. Krose, “Keeping track of humans: Have I seen this person before?” in Proc. IEEE Int. Conf. Robot. Autom., Apr. 2005, pp. 2081–2086.
YANQIAO CHEN was born in 1991. He received the Ph.D. degree in pattern recognition and intelligent system from Xidian University, Xi’an, China, in 2019. He is currently working as an Engineer with 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang, China. His research interests include computer vision and telemeter and telecontrol of UAV.

DANQING YANG received the B.S. degree in mathematics and applied mathematics from Xidian University, Xi’an, China, in 2019, and the master’s degree from the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China, International Research Center of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, in 2022. Her current research interests include deep learning and person re-identification.

RUIJIAO LIU received the B.S. degree in computer science and technology from Hainan Normal University, Hainan, China, in 2015, and the M.S. degree in computer technology from the Xi’an University of Technology, Xi’an, China, in 2018. She is currently pursuing the Ph.D. degree with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education of China. Her research interests include deep learning and synthetic aperture radar image processing and interpretation.

LICHENG JIAO (Fellow, IEEE) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees in electronic engineering from Xi’an Jiaotong University, Xi’an, China, in 1984 and 1990, respectively. From 1990 to 1991, he was a Postdoctoral Fellow with the National Key Laboratory for Radar Signal Processing, Xidian University, Xi’an. Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University. He is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University. He is in charge of about 40 important scientific research projects, and published more than 20 monographs and 100 papers in international journals and conferences. His research interests include image processing, natural computation, machine learning, and intelligent information processing. He is a member of the IEEE Xi’an Section Execution Committee, the Chairperson of awards and recognition committee, the Vice Board Chairperson of the Chinese Association of Artificial Intelligence, the Councilor of the Chinese Institute of Electronics, and the Committee Member of the Chinese Committee of Neural Networks. He is an Expert of Academic Degrees Committee of the State Council.