Multimedia Tools and Applications (2021) 80:26855–26888
https://doi.org/10.1007/s11042-021-10953-6

# Person re-identification based on metric learning: a survey

Guofeng Zou[1] · Guixia Fu[1] · Xiang Peng[2] · Yue Liu[1] · Mingliang Gao[1] · Zheng Liu[2]

Received: 19 August 2020 / Revised: 5 January 2021 / Accepted: 14 April 2021 /
Published online: 10 May 2021


**Abstract**
Person re-identification is a challenging research issue in computer vision and has a broad application prospect in intelligent security. In recent years, with the emergence of large-scale person datasets and the rapid development of deep learning, many outstanding results have been achieved in person re-identification researches, which mainly involves two critical technologies: feature extraction and distance metric. Among them, feature extraction has been well summarized in the current literature of person re-identification, but there is no systematic analysis of the distance metric method in the current review literature. However, effective and reliable distance metric is crucial to improve the accuracy of person re-identification. Therefore, it is necessary to systematically review and summarize the metric learning methods in person re-identification, so as to provide some references for the researchers of metric learning. In this paper, we make a comprehensive analysis of metric learning methods in the past five years, which can be summarized into three aspects: distance metric method, metric learning algorithm, and re-ranking for the metric results. Then, we compare the performance of some representative metric learning methods and discuss them in-depth. Finally, we make a prospect for the future research direction of metric learning in person re-identification.

**Keywords** Distance metric · Classical metric learning · Deep metric learning · Re-ranking · Person re-identification · Survey

## 1 Introduction

Person re-identification is to establish the corresponding relation between persons in different visual range, which is a typical image retrieval problem [9]. Person re-identification is related

✉ Guofeng Zou
gfzou@sdut.edu.cn

[1] School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo 255049, China

[2] School of Engineering, University of British Columbia, Okanagan, Kelowna, BC V1V 1V7, Canada

to multiple technical fields such as computer vision, pattern recognition, and machine learning. Now, it has been widely concerned in academia and industry, and has become a research hotspot in computer vision [45]. However, due to the complex interference factors such as different camera resolutions, view angle and background changes, illumination variation, occlusion, and person pose changes, person re-identification is facing numerous technical challenges. More importantly, there is still a large gap between current person re-identification technology and practical application.

The flowchart of person re-identification system is shown in Fig. 1. Among which person detection and tracking [46] has become an independent research issue in computer vision, and has gradually matured after years of development. Now, the study of person re-identification mainly focuses on two aspects: (1) person feature extraction and description; (2) person feature distance metric.

In 2003, the research on cross-view target matching is considered to be the origin of re-identification [48], and "Person re-identification" was first proposed in CVPR 2006 [23]. With the progress of computer vision, person re-identification technology has also achieved fruitful results. In 2014, Bedagkar-Gala et al. [6] made the earliest review of person re-identification, which summarized the definition, research status, main challenges, and main datasets of person re-identification. With the deep neural network winning the ImageNet competition [31], image recognition based on deep learning has attracted researchers' attention rapidly. Meanwhile, person re-identification based on deep learning is favored by many researchers. In 2018, Li et al. [71] combed and summarized the development history, research status and typical methods of person re-identification, which involved the analysis of some deep methods. In 2019, Luo et al. [42] made a comprehensive analysis of the person re-identification method based on deep learning and the person datasets applicable to deep learning.

In Ref. [6, 42, 71], person feature extraction and description methods are mainly summarized and analyzed, but there is less systematic analysis of feature distance metric method. In addition, considering the key role of reliable metric in improving person re-identification accuracy, we believe that it is necessary to summarize the metric learning methods in person re-identification in recent years. Therefore, in this work, we focus on the comprehensive analysis and review of metric learning methods in person re-identification.

At present, metric learning in person re-identification usually treats "metric" and "metric learning" as a whole, but there are essential differences between them. Metric refers to the calculation of distance or similarity between features in the embedded space, and the features are obtained by mapping the original samples through the metric matrix. Metric learning usually refers to the process of designing an objective function or a loss function to obtain the metric matrix (mapping relation) by solving the optimization problem. The goal of metric learning is to make the same class objects more compact and more separate between different

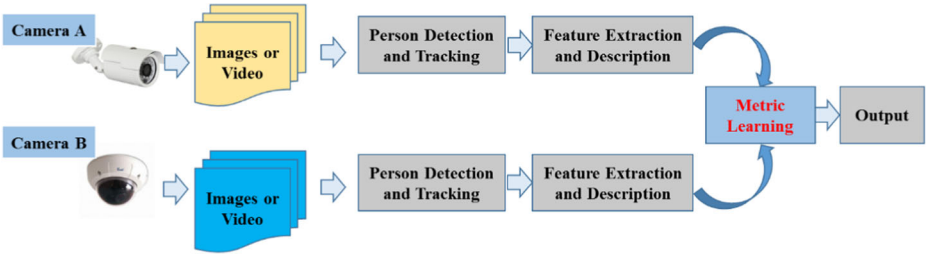


**Fig. 1** Flowchart of person re-identification system

class objects in the embedded space. Therefore, to better comb and analyze the progress of metric learning method in person re-identification, in this paper, we discuss "metric method" and "metric learning algorithm" separately.

In addition, considering that the metric results of person re-identification usually need to be further optimized or re-ranked after distance or similarity calculation, in this paper, we summarize "re-ranking algorithm" as an important part of metric learning method in person re-identification.

In summary, in this paper, we summarize the research progress of person re-identification methods based on metric learning in the past five years. These research results are divided into metric method, metric learning algorithm and re-ranking algorithm. Then, the experimental results of some representative methods are compared and analyzed. Finally, the possible future research trends and hotspot issues of metric learning in person re-identification are discussed. The organizational chart of this paper is shown in Fig. 2.

## 2 Metric methods in person re-identification

### 2.1 Distance metric

#### 2.1.1 Mahalanobis distance metric

The Mahalanobis distance [16] is a classical method for measuring distance or similarity in person re-identification. Assuming that the vector $v = [x_1, x_2, \cdots x_n]$ represents the features of $n$ persons, the Mahalanobis distance between two features $x_i$ and $x_j$ is defined as:

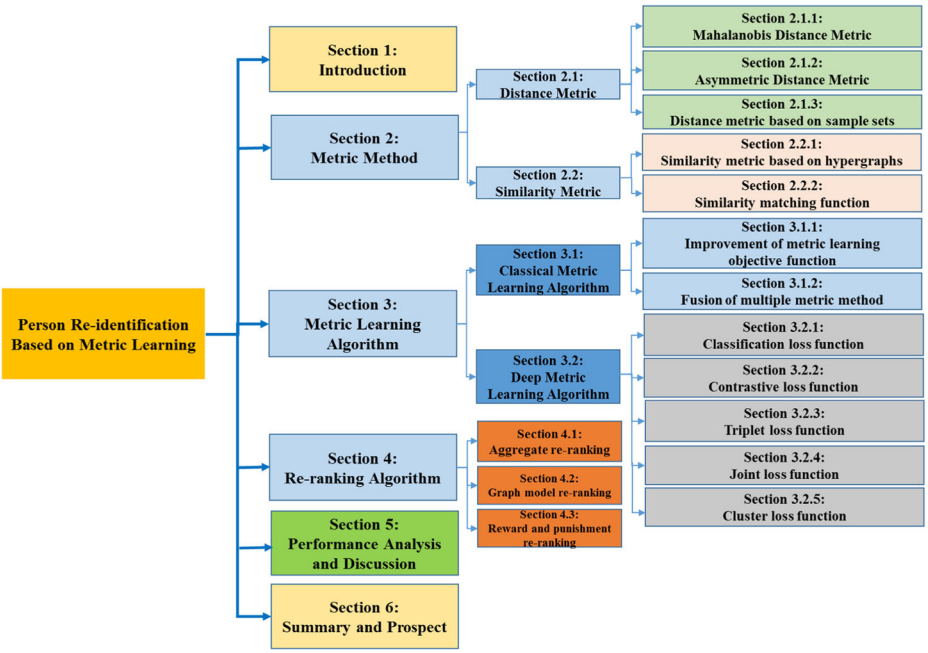


**Fig. 2** The organizational chart of the paper

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)} \tag{1}$$

where the positive semi-definite matrix $M \in R^{d \times d}$ is the Mahalanobis metric matrix to be solved.

The classical Mahalanobis metric is easy to over-fit when the training samples is limited, to overcome the overfitting problem, Qi et al. [49] proposed a regularized independent metric matrix. Unlike the smooth regularization algorithm [58], the proposed method learned the metric matrix separately in four color spaces, and then the regularized independent Mahalanobis metric matrix was obtained by regularization of four metric matrices respectively. Zhou et al. [85] constructed reference datasets based on three different principles and defined a new reference constraint. Based on the constraint, an improved optimized Mahalanobis distance metric matrix was defined. Apart from the classical Mahalanobis metric and improved Mahalanobis metric methods, according to the practical difficulties of person re-identification, some new metric methods or similarity functions are further proposed, the following four sections describe them in detail.

### 2.1.2 Asymmetric distance metric

In person re-identification, the view angle change causes the problem of feature failure and poor metric effect. To solve this problem, Chen et al. [10] proposed the asymmetric distance metric. In this method, the RGB, HSV and YCbCr features are extracted, and then the unmatched features of each person are transformed into a common space. In this space, the cross-view discriminative features are extracted, and based on these features, the cross-view and same-view distance functions are respectively defined:

$$d\left(x_i^p, x_j^q\right) = \left\| U^{pT} x_i^p - U^{qT} x_j^q \right\|_2^2 \tag{2-1}$$

$$d\left(x_i^p, x_j^p\right) = \left\| U^{pT} x_i^p - U^{pT} x_j^p \right\|_2^2 \tag{2-2}$$

where Eq. (2-1) represents the cross-view distance and Eq. (2-2) represents the same-view distance. $U^p$, $U^q$ are the asymmetric metric matrices for camera angle $p$ and camera angle $q$. $x_i^p$ is the $i$-th sample under camera angle $p$. Based on the asymmetric metric matrix, the feature loss and metric problems caused by the deformation of cross-view person images can be effectively overcome.

To overcome the shortcoming of poor scalability of supervised person re-identification, Yu et al. [73] proposed an unsupervised asymmetric metric. Similar to ref. [10], this method transforms person images with different view angle into a common space through an asymmetric mapping, in which the differences of samples from different view angle are alleviated. Feng et al. [21] proposed an unsupervised cross-view metric based on the characteristics of sample distribution, to solve the problem that common metric methods only consider the shared features, but ignore the specific features and cause information loss. This method combines the shared features and inconsistent features of persons through shared mapping and asymmetric specific perspective mapping to improve the effect of cross-view person re-identification.

### 2.1.3 Distance metric based on sample sets

The high-order correlation between samples based on the hypergraph model can improve the metric reliability, but beyond that, comprehensive utilization of the relationship between sample sets can also effectively improve the effect of person re-identification. In 2015, Li et al. [36] defined the difference between pairwise sample sets, and constructed a local metric domain based on this difference, thus forming the sample set distribution of intra-class compact and inter-class separation. Then, the set-to-set matching was realized by using the set-level nearest neighbor modeling method. Tan et al. [57] divided the pairwise image sequences into different groups (sets) and adopted the method of full connection within groups and no connection between different groups to obtain the individual different features of the set. Finally, the set feature was used to train the Adaboost classifier to implement the matching of image sequences from different view angles. Cho et al. [14] used the input image set to estimate the front, rear, left and right postures of persons respectively, so as to realize the reliable correlation of person information under different camera angles. Then, the person re-identification was transformed into the set-to-set matching of persons under fixed pose. This method not only utilizes the local neighborhood relation of the image set, but also introduces the pose constraint to improve the accuracy of multi-view and multi-pose person re-identification.

The set-based metric method extends the paired samples matching to the paired sample sets matching, which effectively uses relevant structure information and context information of the local neighborhood of the sample set, overcomes the variability and sparsity of single person image. This method alleviates the overfitting problem of traditional metric model and helps to enhance the discriminability of metric model, which is helpful to improve the accuracy of person re-identification.

## 2.2 Similarity metric

### 2.2.1 Similarity metric based on hypergraphs

The popular metric methods usually consider the pairwise similarity between the test sample and the target sample, and it is easy to ignore the high-order correlation between the test sample and the target sample. To address this problem, An et al. [2] proposed a person matching scheme based on hypergraph. Through hypergraph learning to mine the pairwise relationship and higher-order relationship between test samples and target samples, the improved similarity score can be obtained and the person re-identification effect is improved. The algorithm flowchart is shown in Fig. 3. Zhao et al. [78] proposed a similarity metric based on multi-hypergraph joint learning. This method extracts the features of gBiCov, HLCNL and LOMO for the input image pair, and constructs three hypergraphs respectively. Then, the multi-hypergraph joint learning algorithm is used to learn the correlation between features for the similarity metric. The algorithm framework is shown in Fig. 4.

Hypergraph-based similarity metric not only considers the pairwise matching of image pairs, but also comprehensively considers the high-order correlation between test data and target dataset. It realizes the sufficient expression of different features and the conveying of more information, it is a flexible and effective similarity metric method for person multi-features.

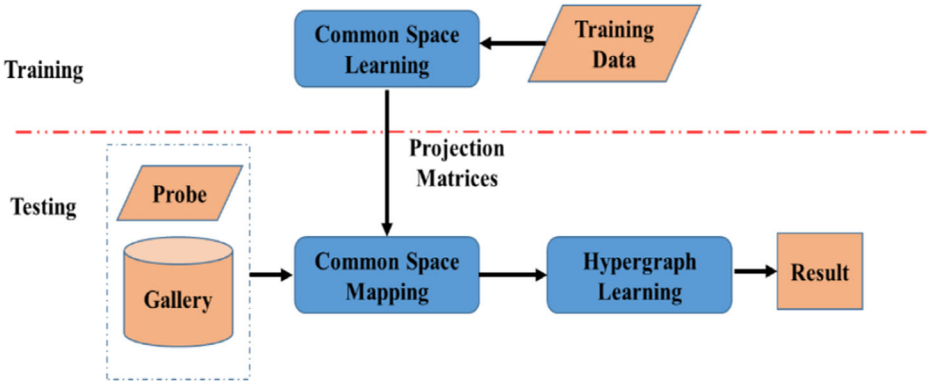


**Fig. 3** Person matching scheme based on hypergraph [2]

### 2.2.2 Similarity matching function

In addition to defining a new distance metric, researchers propose many effective similarity metric functions from the perspective of paired person image matching. In 2016, Du et al. [75] proposed the quadratic similarity function to strengthen the connection between similarity function and person appearance features. The function not only describes the cross-correlation of person image pairs, but also includes the autocorrelation of person image pairs, which can effectively capture the appearance changes of the same person in different scenes. In the joint optimization of deep network and metric learning, the translational invariance leads to the infinite solutions in the low-level feature representation, which complicates the network optimization. To overcome this problem, Chen et al. [7] proposed a weighted inner product similarity function, which can reduce the training difficulty of network parameters and enhance the discriminability of the learned features. Zhu et al. [87] extracted the deep features of person image pairs using the deep convolutional network, and then calculated the absolute

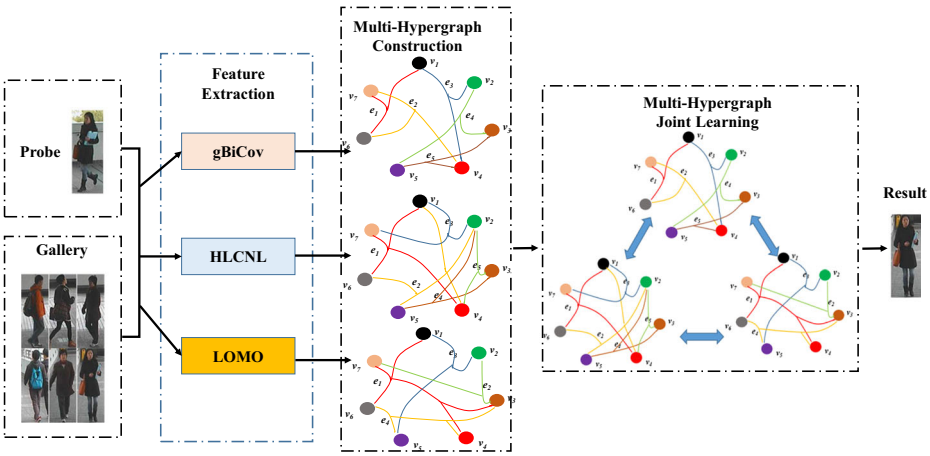


**Fig. 4** Multi-hypergraph joint learning framework for person re-identification [78]

difference and product of person deep features respectively. Finally, the two metrics were fused to obtain a mixed similarity function. Zhao et al. [79] proposed a new joint transfer constraint to learn the similarity function by combining multiple common subspaces, each in charge of a sub-region, which can solve the issue of person re- identification under inconsistent data distributions.

In this section, we summarized and analyzed four aspects of the metric methods: (1) The regularized constraint Mahalanobis metric to overcome the problem of overfitting; (2) The asymmetric metric to solve the difficulty of cross-view metric; (3) Improving person re-identification accuracy based on hypergraph similarity metric and set-to set metric; (4) Defining or designing a new similarity matching function to strengthen the correlation between similarity metric and person features. In order to better compare these new metric methods, some metric and similarity calculation formulas are summarized in Table 1.

The improvement of metric method is an important way to improve the performance of person re-identification, but it is more difficult to propose novel metric. In addition, the learning of metric matrix usually needs to be associated with specific objective function to implement optimization operation, which results in a certain limitation in the generality of the metric method. Therefore, compared with designing new metric method, in the metric learning algorithm, designing a new objective function or loss function is more favored by researchers, and more research results have been obtained.

## 3 Metric learning algorithms

Compared with designing new metric methods, the improvement of metric learning algorithm is more favored by researchers. Considering the influence of deep learning, in this paper, we divide the metric learning algorithms into " classical metric learning algorithm " and "deep metric learning algorithm" to discuss separately.

### 3.1 Classical metric learning algorithm

The classical metric learning algorithm is to define an explicit objective function under certain rule constraints, and then by solving the optimal objective function, the metric matrix with strong discriminability and robustness can be obtained. Currently, the classical metric learning algorithms in person re-identification include: Large Margin Nearest Neighbor (LMNN) distance metric [64], and improved Fast-LMNN [63] and LMNN-R [17] algorithms; Information theoretic metric learning (ITML) [15]; Keep it Simple and Straight Metric (KISSME) learning algorithms [30]; Probabilistic relative distance comparison (PDRC) metric learning [80, 81]; Local Fisher Discrimination Analysis (LFDA) metric learning [47]; Cross-view Quadratic Discriminant Analysis (XQDA) distance metric learning [39]; Metric learning Accelerated Proximal Gradient (MLAPG) [38] etc..

The above classical metric learning algorithms have been widely used in person re-identification, but the learned metric matrix still faces problems such as overfitting, weak classification ability and poor anti-interference ability. Therefore, the classical metric learning algorithms have been improved, which are summarized into two aspects in this paper:(1) improvement of metric learning objective function; (2) multi-metric method fusion.

**Table 1** Some metric and similarity calculation formulas

| Reference | Metric method | Formulas | Explanation |
|---|---|---|---|
| [49] | Regularized Mahalanobis metric | $\widetilde{M} = (1-\lambda)M + \frac{\lambda}{n}tr(M)I$ | $\widetilde{M}$:Metric matrix<br>$\lambda$:Regularization coefficient<br>$I$:Unit matrix<br>$n$:Number of training samples<br>$X$:Original dataset<br>$R$:Reference dataset |
| [85] | Reference constraint Mahalanobis metric | $\widetilde{M} = \left(X^TX + \lambda nI\right)^{-1}X^TRR^TX\left(X^TX + \lambda nI\right)^{-1}$ | |
| [75] | Quadratic similarity function | $f_Q(X,Y) = X^TAX + Y^TAY + X^TBY + Y^TBX + W^TX + W^TY + e$ | $W^TX+W^TY$:First order relation<br>$X^TAX+Y^TAY$:Second order auto-correlation<br>$X^TBY+Y^TBX$:Second order cross-correlation<br>$e$:Bias |
| [7] | Inner product similarity function | $S(h_1,h_2)=w^T(h_1 \odot h_2)$ | $w$:Weight vector<br>$h_1$, $h_2$:Features of image pair |
| [87] | Mixed similarity function | $S_H(X_1,X_2) = w_d^T\lvert X_1-X_2\rvert + w_m^T(X_1.*X_2)$ | $w_d$:Projection matrix of absolute difference<br>$w_m$:Projection matrix of product |

### 3.1.1 Improvement of metric learning objective function

In order to solve the problem of poor robustness of current metric learning algorithms, in 2016, Zhou et al. [83] improved LMNN objective function from the perspective that different samples and different local features contribute differently to recognition. In this method, the sample instance weighting strategy is adopted to make the intra-class samples more compact and inter-class samples more separated, and then the local feature weighting is used to further realize the rewards and punishments for features of different importance. Tan et al. [56] implemented the feature fusion of foreground and background regions by dividing dense areas and extracting features of person images. Then, based on RankSVM framework [28], a ranking model was designed to maximize the interval and minimize fusion feature error. The ranking model was integrated into the objective function of RankSVM to help solve the optimal metric matrix. Wang [61] and Zhou et al. [86] proposed the improved equidistance constrained metric learning (ECML) respectively. The algorithm maps the same class samples to the same vertex of feature simplex and maps different class samples to different vertices, and then the relationship between different class samples is expressed by vertex distance. Based on this assumption, a metric learning objective function that satisfies minimizing intra-class distance and maximizing the minimum interval between different classes is defined. Liu et al. [41] proposed a joint optimization strategy for view angle transformation and similarity metric, and defined a new joint optimization objective function. This method reduces the difference between positive sample pairs, increases the difference between negative sample pairs, and improves the separability of the learned features and the robustness of the metric matrix.

To solve the problem of data imbalance in pair constraint metric learning, Ding et al. [89] proposed a similarity metric learning algorithm based on distance centralization. In the training dataset construction, the eigenvalues of the same class target groups were centralized. Then, the distance between different classes was constructed by using the central eigenvalues, and the intra-class distance still adopted the traditional calculation method. Finally, the metric learning objective function adopted the logarithmic relative distance comparison model. After centralization, the number of intra-class samples is close to the number of inter-class samples, which can effectively alleviate the overfitting problem caused by class imbalance. Dong et al. [18] proposed large margin relative distance learning (LMRDL). This method takes triples as input to extract the color and texture features of different stripes. In the objective function construction, the minimum inter-class distance is used to punish triples, and the logical loss function is used to implement the relative distance comparison, which is helpful to learn more effective metric matrix. The logical loss function is defined as follows:

$$L_M\left(x_i, z_j, z_k\right) = \log\left(1 + \exp\left(\Delta_M\left(x_i, z_j, z_k\right)\right)\right), \left(x_i, z_j, z_k\right) \in T \qquad (3-1)$$

$$\Delta_M\left(x_i, z_j, z_k\right) = d_M^2\left(x_i, z_j\right) - \min_k d_M^2\left(x_i, z_k\right) + \delta \qquad (3-2)$$

where $T$ is the triple sets, $d_M^2(\cdot)$ is the Mahalanobis distance, $\min_k d_M^2(x_i, z_k)$ is the penalty term of minimum inter-class distance, $\delta$ is the minimum interval between positive sample and negative sample.

To overcome the overfitting problem of traditional metric model, He et al. [25] proposed a ring-push metric learning (RPML) algorithm. Unlike the metric learning strategy that punishes too small inter-class distance, this method punishes both those with too small inter-class

distance and those with too large inter-class distance. By using the generalized logic function as the loss, the learning of ring-push metric is transformed into solving the convex optimization problem. The learning objective function of ring-push metric is:

$$L(M) = \frac{1}{N}\sum_{i=1}^{N} l_p(x_i, z_i) + \frac{1}{N(N-1)}\sum_{i\neq j} l_n(x_i, z_j) \tag{4}$$

where $N$ is the number of sample pairs, $l_p(x_i, z_i) = (1/\beta)\log\left\{1 + e^{\beta\left[D_M^2(x_i,z_i)-\mu_1\right]}\right\}$ is loss of similar image pairs, $l_n(x_i, z_j) = (1/\beta)(\log\left\{1 + e^{\beta\left[D_M^2(x_i,z_j)-\mu_3\right]}\right\} + \log\left\{1 + e^{\beta\left[\mu_2 - D_M^2(x_i,z_j)\right]}\right\})$ represents the loss of dissimilar image pairs, $\beta$ is the smooth factor, $D_M^2(\cdot)$ is the Mahalanobis distance, $\mu_1 < \mu_2 < \mu_3$ is three preset thresholds.

In order to compare and analyze the characteristics of these improved methods, Table 2 summarizes some new objective functions that improve the robustness of metric learning.

By adding different constraints to the classical metric learning objective function to generate a new metric learning objective function, which can alleviate data imbalance and class imbalance, and improve the generalization ability and robustness of the learned metric matrix. In addition, these improvements are also conducive to the capture of sample features with stronger identification ability, which improves the accuracy of person re-identification. Although the improvement of the objective function improves the effect of person re-identification, it still cannot overcome the dependence of metric learning algorithm on big training data. At the same time, due to the complexity of the objective function, the convergence speed of the optimization algorithm slows down and the solving efficiency of the optimal metric matrix decreases.

### 3.1.2 Fusion of multiple metric method

The new metric learning objective function mainly improves the performance of a single distance metric matrix, but a single metric matrix is often unable to take into account the metric of different visual features of different persons. Therefore, some multi-metric fusion algorithms are proposed from the perspective of multi-feature subspace metric fusion. Wang et al. [35] proposed the idea of learning different visual metrics based on different candidate datasets. First, based on the visual similarity of the samples, the large-scale training sample sets were divided into different candidate sets and given different weights. Then, several different maximum interval metrics were obtained by the training based on different candidate sets. Syed et.al [54] proposed an adaptive weighted multi-kernel method. First, the color and texture features of the image pairs were extracted, and then these two features were mapped to corresponding subspace by using different kernel functions. Then, Fisher discriminant analysis was used to learn the weighted metric matrix in different kernel spaces, and metric matrix fusion enhances the robustness of the metric to inter-class discrimination. Barman et.al [4] adopted SDALF, SDC_knn, SDC_ocsvm and XQDA metric methods to calculate the distance between image pairs, and then used weighted distance aggregation framework to realize multi-metric fusion and improve the accuracy of person re-identification. Because the metric learning based on feature fusion tends to ignore the difference between different features, and the learned metric matrix cannot accurately express the similarity or difference between samples. To solve this problem, Qi et.al [50] proposed a person re-identification algorithm based on multi-feature subspace and kernel learning. This method maps features to different kernel subspaces through kernel learning, and the similarity function in kernel space is as follows:

**Table 2** Some improved metric learning objective function

| Reference | Improved objective function | Explanation |
|---|---|---|
| [83] | $F(L)_{L^TL=I} = \mu_1\varepsilon_a(L) + \mu_3\|L\|_{2,1} + \mu_2\varepsilon_b(L)$ | $\mu_1$, $\mu_2$, $\mu_3$:Weighting coefficients<br>$\varepsilon_a(L)$:Intra-class distance<br>$\varepsilon_b(L)$:Inter-class distance<br>$\|L\|_{2,1}$:Importance of different features |
| [56] | $(w^*, P^*) = \min_{(w,P)} \|(w^x, Pw^y)\|^2 + C\sum\xi_{ij} \quad s.t.\ f\left(\left(F_{ii}^x, P^T F_{ii}^y\right)\right) > f\left(\left(F_{ij}^x, P^T F_{ij}^y\right)\right) + 1 - \xi_{ij}$ | $P$:Projection matrix<br>$\xi_{ij}$:RankSVM slack variable<br>$C$:Training error interval<br>$\left(F_{ii}^x, F_{ii}^y\right)$:Fusion of similar features<br>$\left(F_{ij}^x, F_{ij}^y\right)$:Fusion of different features |
| [61] | $L(M) = \frac{\gamma}{\|S\|}\sum_{(x_i,z_j)\in S}\left(D_M^2(x_i,z_j)+1\right)^2 + \frac{1-\gamma}{\|D\|}\sum_{(x_i,z_j)\in D}\left(D_M^2(x_i,z_j)-\mu\right)^2 + \frac{\lambda}{2}\|M-I\|_F^2$ | $S$:Dataset of similar samples<br>$D$:Datasets of different class samples<br>$D_M^2(\cdot)$:Mahalanobis distance<br>$\mu$:Equidistance constraint<br>$\gamma\in[0,1]$:Hyper-parameter<br>$\lambda$:Regularization coefficient<br>$I$:Unit matrix<br>$M$:Metric matrix |
| [86] | $F(M) = \frac{\gamma}{\|S\|}\sum_{(x_i,z_j)\in S}\left(\lg\left(1+\exp\left(D_M^2(x_i,z_j)\right)\right)\right) + \frac{1-\gamma}{\|D\|}\sum_{(x_i,z_j)\in D}\lg\left(1+\exp\left(\mu\|M-I\|_F^2 - D_M^2(x_i,z_j)\right)\right) + \frac{2}{\lambda}\times\|M-I\|_F^2$ | |
| [41] | $(M^*, A^*) = \underset{A,M}{\operatorname{argmin}}\frac{1}{\|S\|}\sum_{(x_i,y_j)\in S} d_{M,A}\left(x_i,y_j\right) - \frac{1}{\|D\|}\sum_{(x_i,y_j)\in D} d_{M,A}\left(x_i,y_j\right) + \lambda_f\left\|Y-A^TX\right\|_F^2 + \lambda_M\|M\|_F^2$ | $S$:Dataset of similar samples<br>$D$:Datasets of different class samples<br>$d_{M,A}(\cdot)$:Mahalanobis distance<br>$\lambda_f$:View transformation regularization coefficient<br>$\lambda_M$:Metric matrix regularization coefficient<br>$M$:Metric matrix<br>$A$:View angle transformation matrix |

$$S(k_i, k_j) = (k_i - k_j)^T M (k_i - k_j) \tag{5}$$

where $M$ is the metric matrix in kernel space, $k_i$ is the $i$-th nonlinear feature in kernel space. Different similarity metrics can be learned based on different person features, and then the sum of similarity in different subspaces can be calculated:

$$S = \alpha S_C + (1-\alpha) S_T \tag{6}$$

where $S_C$, $S_T$ represent the similarity functions of different feature spaces, which corresponds to different similarity metric matrices $M_C$, $M_T$.

In the multi-metric fusion method, the metric matrix is trained by using multiple features of person images, which enriches discriminative training data and alleviates the dependence of metric learning methods on big training data. In addition, the metric matrices learned by different features usually have different advantages and metric ability, so the fused metric matrix has stronger identification ability, generalization ability and robustness, which is helpful to implement person image matching in complex situation. Finally, the multi-kernel mapping is used to transform the original features into a high-dimensional nonlinear space, which helps to overcome the non-linear interferences such as illumination changes, pose and view angle changes, and further enhance the robustness of the metric model.

### 3.2 Deep metric learning algorithm

Deep metric learning is to extract features through deep network and construct corresponding loss function based on the deep features. Then, the optimal parameter configuration of the network model is obtained by optimally solving the loss function, so as to realize the essential feature extraction and reliable classification. The loss function is a key factor affecting the deep metric learning effect besides the deep network structure. In addition, most researchers usually equate the design of loss function with the metric learning based on deep network. In this section, we comb and analyze the construction methods of deep loss function in person re-identification, and summarize the deep network loss as: classification loss, contrastive loss, triplet loss, joint loss and cluster loss.

#### 3.2.1 Classification loss function

Softmax function [32] is a classical loss in deep metric learning, which can ensure that the learned deep features are well separated. However, when Softmax loss is directly used for similarity comparison, the intra-class compaction and inter-class separation of deep features cannot be achieved, which affects the parameter learning and re-identification accuracy. To address the problem, Zhu et al. [88] combined center loss and Softmax loss to construct a new loss function, which realized maximum inter-class separation and intra- class compactness. It is conducive to convolutional network learning more discriminative features. Borgia et al. [8] proposed the steering meta center term and the enhancing centers dispersion term, and then combined the two loss items with Softmax loss to form a new loss function. Under the supervision of new loss, the deep separate feature extraction was realized and the inter-class interference was reduced. Feng et al. [20] proposed the concepts of cross-view European constraint and cross-view center loss constraint to solve the challenge of intra-class feature differentiation caused by view angle changes. Combining the two constraints with Softmax

loss respectively, two new loss functions were defined. Based on the cross-view Euclidean constraint, it is beneficial to realize the alignment of the deep features of different views, and the center lossconstraint can narrow the gap between the features of different views.

Softmax loss can only implement the separation of deep features of different class samples, but cannot guarantee the compactness of same class sample features. By introducing the center loss, the inter-class separation is maintained and the intra-class distance is controlled. In addition, to solve some specific problems in person re-identification, by introducing the constraint penalty term into the loss function, the specific interference is alleviated to some extent. In a word, the performance of Softmax loss function is improved effectively through the center loss and special constraint penalty items, which is helpful for deep network parameter optimization and metric performance improvement.

### 3.2.2 Contrastive loss function

Contrastive Loss [24] can effectively express the matching degree of sample pairs, and can better supervise the training of deep feature extraction model, which is widely used in the similarity comparison of person re-identification. Its definition is as follows:

$$L = \frac{1}{2N}\sum_{n=1}^{N} yd^2 + (1-y)\max(margin-d, 0)^2 \tag{7}$$

where $N$ is the number of sample pairs, $d = \|a_n - b_n\|_2$ is the Euclidean distance between the two samples, the label $y = 1$ means that two samples are similar or matched, $y = 0$ means that two samples are mismatched, *margin* is the threshold.

The contrastive loss function can be directly applied to person re-identification, but the matching accuracy is low due to the interference of many complex factors. In this case, the optimal parameter configuration of the deep network cannot be obtained. Therefore, based on the original contrastive loss function, many improved contrastive loss functions are proposed.

In 2018, Wang et al. [62] extracted features based on component deep Siamese network, and constructed an adaptive interval loss function consisting of similarity comparison item and regularization item. This method minimizes the distance of same class samples and maximizes the distance of different samples. Chen et al. [7] defined the weighted inner product distance. Then, the logarithmic loss functions for the positive and negative sample sets were constructed respectively, and these two losses were fused to form a joint contrastive loss. Zhu et al. [87] proposed a weighted mixed similarity metric, based on this metric method, defined a mean logarithmic contrastive loss with regularized term. The proposed method can reasonably allocate the complexity of feature learning and metric learning in the deep network and improve person re-recognition performance. Saquib et al. [51] defined a new extended cross-nearest neighbor distance, which extended the distance metric between sample pairs to the distance metric between cross-neighbor sets of sample pair. Then, a bidirectional contrastive loss function for neighborhood sets was proposed, which implemented the bidirectional distance metric between test sample and training sample.

The contrastive loss function takes the image pair as the input. The optimization of network parameters and similarity metric are implemented by minimizing the distance between same class sample pairs and punishing the distance between different class samples less than the margin threshold. Contrastive loss function plays an important role in person re-identification based on deep metric learning, which usually combines with new distance metric method,

regularized constraint item or person dataset to define new loss form. The improved contrastive loss greatly expands the application scope of contrastive loss function, which can not only reduce the training difficulty of deep network parameters, but also effectively enhance the ability of network to extract discriminative features and improve the re-identification effect.

### 3.2.3 Triplet loss function

In 2015, Schroff et al. [52] proposed the Triplet Loss in face recognition and clustering analysis. The input triplet data consists of three samples, namely an anchor sample randomly selected from the training sample set, a random sample with the same class as the anchor sample and a random sample with class label different from the anchor sample. The goal of triplet loss is to minimize the distance between anchor sample and positive instances, to maximize the distance between anchor sample and negative instances. Finally, the distance between negative sample pairs and the distance between positive sample pairs are kept at a minimum interval. The objective function of the triple loss is defined as follows:

$$L = \sum_{i}^{N} \left[ \left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 - \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2 + \alpha \right] \tag{8}$$

where $\left(x_i^a, x_i^p, x_i^n\right)$ is the triplet, $x_i^a$ is the anchor sample, $x_i^p$ is the positive sample, $x_i^n$ is the negative sample.$N$ represents the number of triplet, $f(\cdot)$ is the network model, $\alpha$ is the interval threshold.

The classical triplet losses usually compare the distance between a single positive and negative sample, ignoring the samples of other classes. Therefore, it can only promote a greater distance between the test sample and the selected negative sample, and cannot guarantee a greater distance between the test sample and other unselected negative samples. For this problem, the improved structural triplet loss is proposed. In 2018, Yang et al. [68] proposed a new learning method for structural deep metric to solve the problems of slow convergence speed, local optimality and insufficient attention paid to difficult positive samples in network training. The algorithm divided training samples into many different small batches. In each small batch training, the positive sample pairs were compared with all the negative sample pairs, and a difficulty weight was adaptively assigned to the positive sample pairs. The weighting strategy made the algorithm focus more on the learning of difficult positive samples, reduced the distance variance of positive/negative sample pairs, and enhanced the generalization ability of the loss function. Yu et al. [72] proposed to construct triplet input by using positive sample pair, negative sample and negative kin sample. Then, the distance between positive sample pair, the distance between negative sample pair, the distance between positive sample and negative sample, the distance between positive sample and negative kin samples, the distance between negative sample and negative kin sample were defined respectively, and a new triplet loss was constructed by integrating regularization term. He et al. [26] proposed an improved lifting structure loss function to supervise deep network learning better feature. This method can solve the problems that the triplet loss cannot make full use of batch information and it is difficult to select negative samples manually. Arindam et al. [53] proposed a batch adaptive triplet loss function. In this strategy, the weights of the hardest samples were adjusted adaptively according to their distances with the anchor. This method well overcomes the influence of scale on person re-identification.

Unlike classical triplet loss, structural triplet loss function makes full use of the structure information between person samples to learn more discriminative features. In the process of parameter updating, the relationship between anchor sample and other negative samples is considered, which promotes the distance between anchor sample and all other classes, and greatly accelerates the convergence speed of the model. The lifting structural loss constructs the most difficult triplet for each positive sample pair dynamically, and at the same time, all negative samples are taken into account. It effectively improves the ability of optimizing network parameters of triplet loss and contributes to the realization of strong classification feature extraction and robust metric.

### 3.2.4 Joint loss function

In Sections 3.2.1-3.2.3, we introduce the classification loss, contrastive loss and triplet loss respectively. The three loss functions show their own characteristics in different recognition tasks. In order to integrate the advantages of different losses and further improve the accuracy of person re-identification, the researchers proposed ideas such as multi-task learning and multi-feature metric. In different learning tasks and feature metrics, different loss functions are usually used to supervise the deep network. Therefore, joint learning of multiple loss functions is widely used.

In 2016, McLaughlin et al. [44] used the Siamese network to extract person features and adopted person verification, person recognition and multi-attribute recognition to form multi-task learning objectives. Then the learning objective losses of different tasks were weighted and fused to form a joint loss. The architecture of the multi-task joint learning network is shown in Fig. 5, which improves the network's comprehensive learning ability through joint learning and multi-task loss fusion. Zhou et al. [84] extracted the features of multiple local parts of person images, and then fused all local features together through the tandem method to

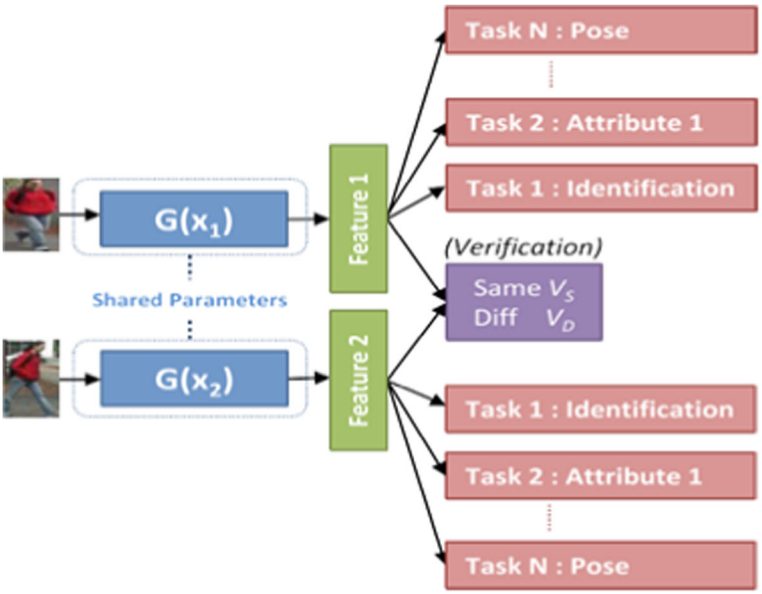


**Fig. 5** Framework of multi-task joint learning [44]

form a feature set. Based on the feature set, classification loss and relative distance loss (triplet loss and contrastive loss) were calculated respectively, and then the three losses were fused to form a joint loss. The distance metric function ensures that the network can obtain robust and separable feature representation. Cheng et al. [13] calculated asymmetric triplet loss and Softmax loss respectively by using triplet features, and then fused the two losses to obtain a new joint objective function. The triplet loss can make the features of same person more compact and the features of different persons separate. The Softmax loss helps the network learn the discriminative features of persons. In 2019, Ling et al. [40] extracted the classification features and attribute features through ResNet-50. In classification feature subnet, person identification and person verification were realized based on Softmax loss and contrastive loss respectively. In attribute feature subnet, Softmax loss and contrastive loss were used to realize attribute recognition and verification. Finally, the four losses were combined to obtain the total loss. The combined loss enhances the network's ability to accurately identify persons. Attribute feature is an effective complement of classification feature, and the combination of person recognition and verification enhances the reliability of decision result.

### 3.2.5 Cluster loss function

With the increase of person samples, the local spatial distribution relation of the samples and the interaction between the sample pair and other sample pairs show unique advantages in the distance metric, which plays an important role in assisting person re-identification. In 2018, Li et al. [37] defined the concept of positive and negative support neighborhood sets containing context information and neighborhood structure for each anchor sample. Based on these neighborhood sets, a new support neighborhood loss function was proposed:

$$L(\theta) = L_{spr}(\theta) + \lambda L_{sqz}(\theta) \tag{9-1}$$

$$L_{spr}(\theta) = -\sum_{i=1}^{N} \log \frac{\sum\limits_{x_p \in P_i} \exp\left(-\sigma D\left(x_i, x_p\right)\right)}{\sum\limits_{x_s \in K_i} \exp\left(-\sigma D\left(x_i, x_s\right)\right)} \tag{9-2}$$

$$L_{sqz}(\theta) = \sum_{i=1}^{N} \left( \max\left( \bigcup_{x_p \in P_i} D\left(x_i, x_p\right) \right) - \min\left( \bigcup_{x_p \in P_i} D\left(x_i, x_p\right) \right) \right) \tag{9-3}$$

where $D(\cdot)$ is the Euclidean distance, $\sigma$is scale factor, $x_i$ is the anchor sample, $P_i$ is the positive neighborhood set, $K_i$ is the union of positive and negative sample neighborhoods. $\bigcup_{x_p \in P_i} D\left(x_i, x_p\right)$ is distance set between the anchor sample and the positive sample neighborhood set. $\lambda$ is used to control the ratio of separation loss $L_{spr}(\theta)$ to aggregation loss $L_{sqz}(\theta)$. The loss function can reliably separate the positive sample neighbor from the negative sample neighbor, and the change in the positive sample neighborhood is minimized.

Yuan et al. [74] proposed a micro clustering loss. The method is to take similar samples as a micro-cluster and then use it as a whole in training. Within each cluster, the maximum distance between each sample is defined as the internal divergence, the minimum distance from the external sample is defined as its external divergence. By limiting the external divergence to be greater than the internal divergence, the method achieves a more compact micro-cluster

structure and improves the generalization ability of the model. The loss function is defined as follows:

$$L_{mini} = \sum_{a=1}^{P} \log\left(1 + \exp\left(\max_{x_i, x_j \in C_a} \left\|f(x_i) - f(x_j)\right\|_2 - \min_{\substack{b=1,2\cdots P \\ b \neq a}} \left\{\min_{\substack{x_k \in C_a \\ x_l \in C_b}} \|f(x_k) - f(x_l)\|_2\right\}\right)\right) \quad (10)$$

where $(x_i, x_j, x_k, x_l)$ represents a tuple, $x_i$, $x_j$, $x_k$ are from the anchor micro-cluster, $x_l$ is from other sample set. In each training batch, P tuples are used to generate a loss penalty.

In order to compare and analyze the loss functions of deep metric learning more clearly, Table 3 summarizes some loss functions of deep metric learning.

The improvement of deep metric learning algorithm can be summarized into three aspects: (1) The improvement of loss function. The main improvement is to introduce a new loss item or regularization term for metric parameters. (2) Multi-task joint learning. It is to obtain the comprehensive decision for person information through different task objectives (classification or verification) to improve the discrimination accuracy. (3) Composite loss of sample cluster. Based on the context information and neighborhood structure relation of the sample cluster, a comprehensive loss function with stronger generalization is defined. Deep metric learning shows great advantages in feature extraction and metric of person re-identification through complex network architecture and optimal loss, which has become the mainstream method to deal with various challenges in person re-identification. However, the effectiveness of deep metric learning depends on the high quality requirements for the training data set. When the training set contains complex noise, abnormal point, class imbalance, small sample and other problems, its effectiveness is often not guaranteed. Therefore, the robustness and adaptability enhancement of deep metric learning algorithm are the key difficulties to be solved urgently.

## 4 Re-ranking algorithm

The researches on metric method and metric learning algorithm have overcome many challenges in person re-identification, and the robustness and generalization of metric matrix have been improved effectively, which provides effective supervision for extracting more discriminative person features. However, the best person re-identification effect cannot be achieved only by the improvement of metric method or metric learning algorithm. As a post-processing method of person similarity metric or person matching, re-ranking provides important ideas for further improving person re-identification accuracy. In this paper, we summarize and analyze re-ranking as an important part of person re-identification based on metric learning. The current re-ranking algorithms can be summarized into three categories: aggregate re-ranking, graph model re-ranking, reward and punishment re-ranking.

### 4.1 Aggregate re-ranking

Aggregate re-ranking means that different initial person sorting results are aggregated by some fusion algorithm to improve the final similarity sorting list. According to the different features used to calculate the initial sorting results, the aggregate re-ranking is divided into two categories: (1) the aggregate re-ranking based on the similarity of structural features; (2) the aggregate re-ranking based on the similarity of multiple features.

**Table 3** Some loss functions of deep metric learning

| Reference | Loss function | Explanation |
|---|---|---|
| [88] | $L = L_s + \lambda L_c = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T x_i + b_j}} + \lambda \sum_{i=1}^{m} \left\|x_i - c_{yi}\right\|_2^2$ | $L_s$:Softmax loss<br>$L_c$:Centralization loss<br>$\lambda$:Proportional coefficient<br>$m$:Scale of training samples<br>$C$:Number of categories<br>$c_{yi}$:The $y_i$th deep feature center |
| [8] | $L = L_{softmax} + \alpha L_{SMC} + \beta L_{ECD} = L_{softmax} + \alpha \frac{1}{2} \sum_{i=1}^{m} \left\| x_i^{(g_i)} - \sum_{j=1}^{s_i} c_{y_i}^{(j)} \right\|_2^2 + \beta \frac{1}{2} \sum_{i=1}^{m} \left[ \sum_{j=1}^{s_i} \left\| x_i^{(g_i)} - c_{y_i}^{(j)} \right\|_2^2 \cdot \sum_{\substack{t=1 \\ t \neq i}}^{m} \sum_{k=1}^{s_t} \frac{1}{\left\| x_i^{(g_i)} - c_{y_i}^{(j)} \right\|_2^2} \right]$ | $L_{softmax}$:Softmax loss<br>$L_{SMC}$:Steering meta center term<br>$L_{ECD}$:Enhancing centers dispel- rsion term<br>$\alpha$, $\beta$:Equilibrium factor<br>$m$:Scale of training samples<br>$s_i$:Number of view angles in class $y_i$<br>$x_i^{(g_i)}$:Input image of camerag$_i$<br>$c_{y_i}^{(j)}$:Sub-center of $y_i$ |
| [7] | $L = \sum_{(i,j) \in T} \left( s_{ij} \lg\left(e^{D_{ij}+t} + 1\right) + \left(1 - s_{ij}\right) \lg\left(e^{t - D_{ij}} + 1\right) \right)$ | $s_{ij} \in \{0, 1\}$:Sample labels<br>$t$:Decision interval<br>$T$:Training sample set<br>$D_{ij}$:Inner product distance |
| [87] | $J(W) = \underset{W}{\operatorname{argmin}} \left\{ \frac{1}{K} \left[ \sum_{k=1}^{K} \log\left(1 + e^{-y_k W^T Z_k}\right) \right] + \frac{1}{2} \alpha \|W\|_2^2 \right\}$ | $\alpha$:Regularization coefficient<br>$K$:Scale of training samples<br>$y_k$:Sample labels<br>$Z_k$:Fusion feature of samples<br>$W$:Network parameter |
| [51] | $L(p, g_i) = \frac{1}{2M} \sum_{j=1}^{M} d\left(pN_j, g_i\right) + d\left(g_i N_j, p\right)$ | $(p, g_i)$:Image pair<br>$pN_j$:The j-th neighbor of nearest neighbor set $N(p, M)$<br>$g_iN_j$:The j-th neighbor of nearest neighbor set $N(g_i, M)$ $M$: Number of neighbors<br>$d(\cdot)$:Distance between two samples |
| [72] | $J = \left\| d\left(h_i^M, u_k^M\right) - d\left(h_j^M, u_k^M\right) \right\|_F^2 - \left\| d\left(h_i^M, u_k^M\right) - d\left(h_i^M, h_j^M\right) \right\|_F^2 + \alpha d\left(h_i^M, h_j^M\right) - \beta d\left(h_i^M, h_k^M\right)$ | $\left\langle h_i^M, h_j^M \right\rangle$:Features of positive sample pair<br>$h_k^M$:Feature of negative sample<br>$u_k^M$:Feature of negative kin sample |

**Table 3** (continued)

| Reference | Loss function | Explanation |
|---|---|---|
| [26] | $L_{struct}=\frac{1}{2\lvert\widehat{P}\rvert}\sum_{(i,k)\in\widehat{P}}\max\left(0,\log\left(\frac{1}{\lvert\widehat{T}_{i,j}\rvert}\left(\sum_{(i,k)\in\widehat{N}}e^{\alpha-D^2_{i,k}}+\sum_{(j,l)\in\widehat{N}}e^{\alpha-D^2_{j,l}}\right)+D^2_{i,j}\right)\right)$ | $\widehat{T}_{i,j}$:Number of negative sample<br>$\widehat{P}$:Positive sample set<br>$\widehat{N}$:Negative sample set<br>$D^2_{i,j}$:Euclidean distance of sample pair $\{i,j\}$<br>$\alpha$:Interval threshold |
| [44] | $L=\sum_{(x_1^i,x_2^i)\in X}\nu\left(x_1^i,x_2^i\lvert y^i;w\right)+\sum_k\alpha_k T\left(G\left(x_1^i\right)\lvert l_1^{i,k};w\right)+\sum_k\alpha_k T\left(G\left(x_2^i\right)\lvert l_2^{i,k};w\right)$ | $T_k(\cdot)$:Loss of task $k$<br>$\nu(\cdot)$:Contrastive loss of verification<br>$\left(x_1^i,x_2^i\right)\in X$:Image pair in set X<br>$l^{i,k}$:True label of one task<br>$\alpha_k$:Weight of different task losses<br>$G(x)$:Image feature<br>$w$:Network parameter |
| [84] | $L=\alpha L_C(X,W,b)+L_S(X,W,b)\ +\beta R(W,b)$ | $L_C(\cdot)$:Classification loss<br>$L_S(\cdot)$:Relative distance loss<br>$R(\cdot)$:Regularization item<br>$\alpha$, $\beta$:Weight coefficient |
| [13] | $\text{argmin}_W L=(1/T)\sum_{i=1}^{T}\Gamma(t_i;W)+\lambda l\left(I_j,y_j;W,\Theta,c\right)$ | $\Gamma(t_i;W)$:Triplet loss<br>$T$:Number of triples<br>$l(I_j,y_j;W,\Theta,c)$:Softmax loss<br>$\lambda$:Proportional coefficient of two losses |
| [40] | $L=(L_{IC}+\beta L_{IV})+\alpha(L_{AC}+\beta L_{AV})$ | $(L_{IC}+\beta L_{IV})$:Identity loss<br>$(L_{AC}+\beta L_{AV})$:Attribute loss<br>$\alpha$:Proportional coefficient of two losses<br>$\beta$:Balance factor of identification and verification loss |

Aggregate re-ranking based on the similarity of structural features. In 2016, Wang et al. [60] derived four image pairs from the original person image: full-scale, semi-scale, upper body region and middle body region. Then, the corresponding convolutional network was used to extract the features for the four image pairs, and the similarity matching score was calculated. Finally, the re-ranking result was calculated through the weighted average of the four scores, the algorithm flowchart is shown in Fig. 6. Zhang et al. [77] proposed a structured matching method for person re-identification. In this method, the training samples are formed into codebooks by small area division and clustering. Then, the codebooks are used to encode the person images into code words to form locally sensitive visual patterns. Finally, the weighted images of different sensitive areas are matched to obtain re-identification results. By capturing local sensitive visual patterns and weighting different sensitive areas, this method highlights the different contributions of structural features. Zhang et al. [76] divided the input image pairs into different strip areas to preserve the person structure information, then calculated the local similarity of each corresponding strip area to form the initial metric vector. Finally, a discriminant subarea aggregation algorithm was adopted to fuse the local similarity score into the final global similarity score. Structural area division and initial similarity fusion can not only effectively strengthen the rationality and reliability of metric results, but also effectively alleviate the impact of person pose change and occlusion.

Aggregate re-ranking based on the similarity of multiple features. In this method, the multiple features of persons are first extracted, then the similarity between corresponding multiple features is calculated as the initial metric result, finally the initial metric results are fused and re-ranked. In 2018, Chen et al. [12] proposed a group constraint similarity learning method based on deep conditional random fields, which extracted multi-scale features of the input image with deep network, and calculated the similarity between the test sample and the target sample in different

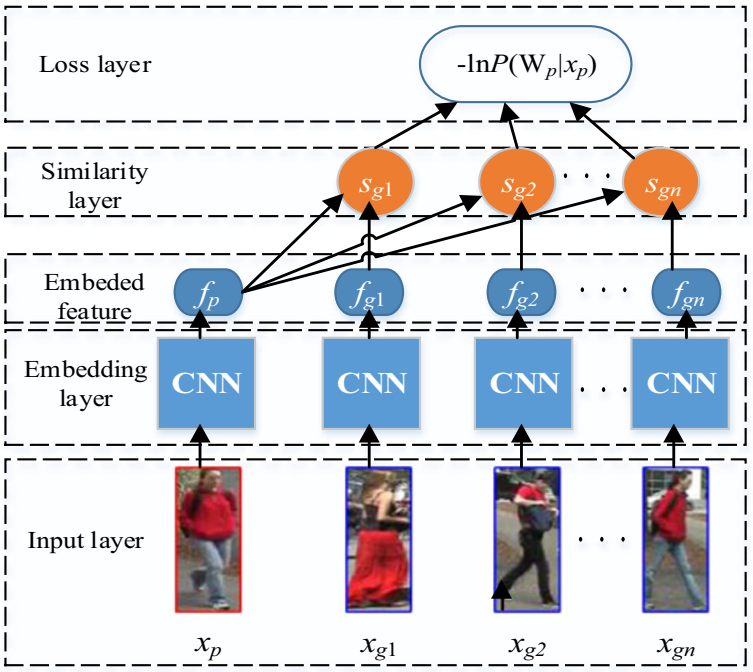


**Fig. 6** Flowchart of multiple similarity weighted fusion [60]

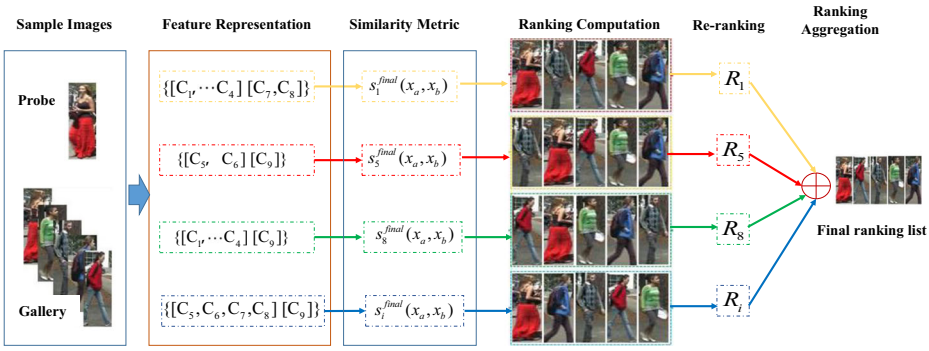


**Fig. 7** Flowchart of similarity matching and re-ranking based on polynomial feature map [29]

scales. Then a continuous conditional random fields algorithm is used to fuse the "group similarity", the optimal group similarity is obtained to realize person re-identification. Junior et al. [29] used deep global network and convolutional network to extract the foreground/background information of person images and adopted them together as person feature descriptions. The person polynomial feature map is constructed based on these feature descriptions, and the similarity between test sample and target sample is calculated by matching the polynomial features, as shown in Fig. 7. Finally, the DCIA algorithm [22] is used to implement the aggregation and re-ranking to improve the accuracy of person re-identification

## 4.2 Graph model re-ranking

The graph model re-ranking maps the initial metric results of person re-identification into a graph model, and then optimizes the graph model to obtain the optimal re-ranking results. In 2016, Xie et al. [66] proposed a multi-metric fusion method based on graph model to solve the problem that the single specific metric can easily lead to overfitting of metric results, as shown in Fig. 8. Firstly, the metric between test sample and gallery samples was calculated in many ways, then the initial metric results were expressed as different graph models. Finally, the complementary re-ranking results were obtained by multi-graph joint ranking. Barman [5] proposed the fractional distance graph for person re-identification. The similarity scores between test sample and gallery samples were calculated, then the gallery sample was expressed as graph model vertex, the similarity scores were expressed as the connection strength of the edges in the graph model. Finally, the greedy algorithm and the ant colony algorithm were used to optimize the graph model. Xie etc. [67] proposed specific sample sorting algorithm based on the structure of hypergraph. A set of initial sorting results were obtained through adaptive metric, and then the sorting results were constructed into a hypergraph structure. Based on the hypergraph, the neighborhood relationship between the test samples and the top 100 target images can be captured. Finally, the neighborhood relationship can be used to re-rank the results to improve person re-identification accuracy.

## 4.3 Reward and punishment re-ranking

Re-ranking based on reward and punishment is to introduce some reward and punishment factors or constraints to modify the initial re-ranking results, so as to improve the rationality and accuracy of the sorting results. In 2015, Leng et al. [34] proposed a bidirectional re-ranking method to correct the initial ranking list based on reverse query results. The pseudocode is shown in Algorithm 1. In

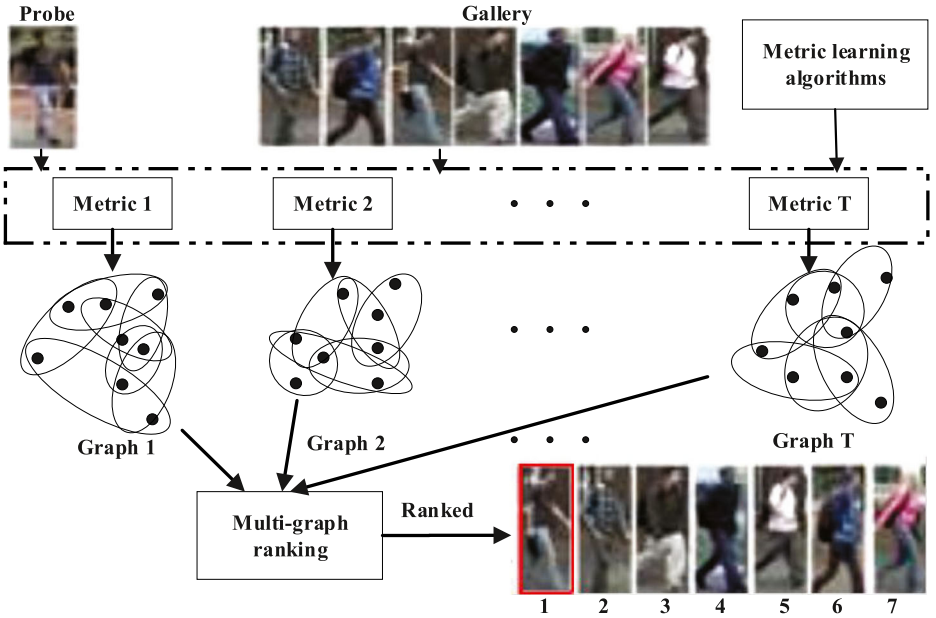


**Fig. 8** Multi-similarity fusion based on graph model [66]

this method, the *k*-nearest neighbor context information similar to the sample features is taken into account by reverse query, which is helpful to improve the person matching accuracy. Chen et al. [11] calculated some similarity scores to form the initial ranking list based on the deep features of person image pairs. Then, a new penalty loss function for mismatched sample sets and test samples.

**Algorithm 1 Bidirectional re-ranking technique**

Algorithm 1 Bidirectional re-ranking technique
**1: Offline**
2: Input: a gallery set $G = \{g_i \mid i = 1, \ldots, n\}$.
3: Querying every gallery person image $g_i$ in the gallery $G \setminus \{g_i\}$ consist of other gallery person images.
4: Output: the ranking list $R_{g_i}(G \setminus \{g_i\})$ for every gallery person image.
5:
**6: Online**
7: Input: probe image $p$ and the ranking list $R_{g_i}(G \setminus \{g_i\})$ for every gallery person image.
8: Inserting the probe into every gallery's ranking list.
9: Compare $p$ with $g_i$ and generate an initial ranking list $R_p(G) = \{g_i^0\}$.
10: **for** $i = 1$ to $n$ **do**
11: Compute the content similarity $S_{cn}(g_i^0)$ between $p$ with $g_i^0$.
12: Compute the context similarity $S_{cx}(g_i^0)$ between $p$ with $g_i^0$.
13: Compute the combination similarity $S_*(g_i^0)$ between $p$ with $g_i^0$.
14: **end for**
15: Use the combination similarities to re-rank the initial ranking list.
16: Output: A new rank of the gallery set is producted for the probe image.

was constructed to further enhance the learning of mismatched samples and minimize the cost of sorting errors Liu et al. [65] proposed a coarse-to-fine ranking iterative algorithm. By adjusting

the value of expected ranking parameters, training data was gradually improved and refined to reduce the penalty for too many sick samples and form the optimal training data subspace. By introducing penalty factors into the training data set, the method optimizes the training data and achieves the optimal results. Ye et al. [70] proposed an optimized sorting framework using both similar and dissimilar cues, as shown in Algorithm 2. This method uses similarity to get a set of ranking results and uses dissimilarity to generate penalty factors. The initial ranking results combine with penalty items to optimize the final ranking list. Yang et al. [69] defined spatiotemporal information and network consistency constraints, as well as camera network topology constraints. Then the two constraints were introduced into the ranking process of similarity metric results, which corrected the image similarity score and improved person re-identification in camera network. Hou et al. [27] constructed an interactive aggregation update module using the global spatial-temporal context information for person feature extraction. Combining classification and ranking, a cross entropy loss was defined, and the attention to hard samples was enhanced in batch training to make the ranking results more reasonable.

As a further optimization step for person re-identification results, re-ranking is an important method to improve recognition accuracy. Through the joint optimization of multiple initial ranking results, the metric results of person multi-local area or person multi-feature are comprehensively utilized to realize complementary information fusion and effectively improve the rationality and reliability of the metric results of person re-identification. By making full use of the advantages of the neighborhood structure of graphs, the re-ranking problem is transformed into the optimization learning of graphs. The re-ranking based on reward and punishment mechanism revises the initial ranking results by reward and punishment factors,

**Table 4** The main challenges and solutions of metric learning methods in person re-identification

| Challenges | Solutions | References |
|---|---|---|
| Poor Robustness | Define new metrics | Asymmetric metric: [10, 21, 73]<br>Hypergraph similarity metric: [2, 78]<br>Set-based metric: [14, 36, 57] |
| | Improved metric learning objective function | [56, 61, 83, 86] |
| | Multi-metric fusion method | [4, 25, 46, 54] |
| | Deep metric learning algorithm | Classification loss: [8, 20, 88]<br>Contrastive loss: [51, 62, 87]<br>Triplet loss: [26, 53, 68, 72]<br>Joint loss: [13, 40, 44, 84]<br>Cluster loss: [37, 74] |
| | Re-ranking algorithm | Aggregate re-ranking: [12, 29, 60, 76, 77]<br>Graph model re-ranking: [5, 66, 67]<br>Reward and punishment re-ranking: [11, 27, 34, 65, 69, 70] |
| Weak generalization ability | Define new metrics | [45, 85], |
| | Improved metric learning objective function | [25] |
| | Re-ranking algorithm | Aggregate re-ranking: [29] |
| Data imbalance | Improved metric learning objective function | [18, 89] |
| | Re-ranking algorithm | Aggregate re-ranking: [60, 76] |
| Difficult to apply across image domains | Cross-domain metric learning algorithm | [43, 59] |
| Limited sample size | Small sample metric learning algorithm | [1, 3, 19, 33, 55, 82] |

which strengthens the factors conducive to accurate ranking list, weakens the possible interferences, and greatly reduces the possibility of error in re-ranking results. However, in recent years, deep metric learning algorithm research has become the mainstream, and the research progress of re-ranking algorithm is slow. Therefore, in future work, effective combination of deep metric learning and re-ranking algorithm should receive some attention.

## 5 Performance analysis and discussion

### 5.1 Relationship between challenges and methods

In above sections, we summarize the research progress of metric learning methods in person re-identification from the perspective of learning theory. Each theoretical algorithm is proposed to solve the corresponding problem. Therefore, in this section, we summarize the metric methods and metric learning algorithms from the perspective of person re-identification challenges, as shown in Table 4. At present, person re-identification based on metric learning faces many problems, among which the more important ones include: poor robustness, overfitting (weak generalization ability), data imbalance, difficult cross-domain application, limited samples, etc.

**Algorithm 2 The SRA Algorithm**

**Input:** A probe image $p$ and a gallery set $G = \{g_i \mid i = 1, \ldots, n\}$

**Output:** A ranking list for the probe image.

**Offline:**

1: Querying every gallery image $g_i$ in the gallery $G$ with two different methods.

2: Achieve the top-*k* galleries of each image $g_i$.

**Online**

1: Probe $p$ in the gallery set $G$.

2: Obtain two original ranking lists the $RL^1(p)$ and $RL^2(p)$ by two different methods for the probe $p$.

3: Get strongly similar galleries set $G_+(p) = N^1_{k+}(p) \cap N^2_{k+}(p)$.

4: **for** $i = 1$ to $|G_+(p)|$ **do**

5: $g_+$ is the $i$-th item in $G_+(p)$.

6: Cross-view based backward requery for $g_+$.

7: Compute weighting coefficient $w(p, g_+)$.

8: Compute new score $Sim^1(p, g_+)$ between $p$ with $g_+$.

9: **end for**

10: Repeat step 5-9 to get $Sim^2(p, g_+)$

11: Late fusion of $Sim^1(p, g_+)$ and $Sim^2(p, g_+)$ by Equation

$$Sim'(p, g_+) = \alpha \cdot Sim^1(p, g_+) + (1 - \alpha) Sim^2(p, g_+)$$

12: Use new scores to re-rank $RL(p)$.

13: The final ranking list $RL'(p)$ is achieved.

Metric learning theory is usually associated with the challenges of person re-identification, among which poor robustness is a common problem caused by many influencing factors

(occlusion, view angle change, pose, illumination, etc.). To solve the problem of poor robustness, the researchers proposed some improved algorithms, which can usually overcome multiple interferences at the same time, rather than being limited to a specific interference. In addition, in the literature that overcomes the problem of poor robustness, there are some literatures that can deal with the problems of weak generalization ability and data imbalance. Considering that the main goal of the proposed methods is to solve the robustness problem, we do not classify these literatures as a solution of weak generalization ability and data imbalance.

The methods to solve the problems of poor robustness, weak generalization ability and data imbalance have been combed and discussed in detail. Therefore, this section focuses on analyzing the difficulties and challenges of metric learning in cross-image domain application and limited sample size.

Difficulty in cross-image domain application of metric learning: Person re-identification based on metric learning usually uses the training samples from two different visual domains to obtain a metric matrix, and then metric the similarity between test sample and gallery samples. However, the real video surveillance is often unable to meet the metric learning requirements under the ideal situation, such as the target person moving through a complex camera network instead of two simple visual domains, persons without label information. Therefore, ref. [43, 59] have done some effective research on the difficulties of metric learning model in cross-image domain or cross-scene application. Cross-domain no-label metric learning implements the transfer of a known person re-identification system in a certain domain to a new domain or scene, which is helpful to solve the problem of person re-identification when the sample label is unknown in the target domain, and strengths the expansibility and practical application value of the person re-identification system.

Difficulty with limited samples of metric learning: Due to the limitation of capture conditions, the available sample quantity is limited, which brings great challenges to person re-identification. Now, the researchers propose corresponding metric learning solutions from two different perspectives: transfer learning and data augmentation. (1) Method based on transfer learning. In 2016, Zheng et al. [82] proposed a local relative distance comparison model to solve the problem that there are few available effective training samples in person re-identification in open environment. First, the relative distance comparison information was learned through the non-target image set. Then, the relative distance comparison was transferred from the non-target image set to the target image set based on the constraint of intra-class change, inter-class change and group separation relative distance comparison. (2) Method based on data augmentation. In 2018, Syed et al. [55] proposed an elastic multi-modal metric method. First, multi-modal information of persons was constructed based on color and texture features, and then the multi-modal information was used to generate false negative sample modes, so as to enrich the negative samples in the training sample set. Then, the metric learning matrix is trained by using the sample set after data augmentation. Dong et al. [19] proposed iterative multi-kernel metric learning, which constructed a pseudo-training sample set using correctly identified test samples to achieve data augmentation of the training samples. Then the original training samples and the pseudo-training samples were used together for the iterative learning of metric matrix to obtain the constantly updated metric matrix, which is helpful to solve the problem of small size samples. T Ali et al. [1] proposed the maximum edge metric learning of null space kernel. First, the maximum edge criterion was used to determine the learning metric on the null space. All training samples with the same category were mapped to a single point, so as to minimize the intra-class divergence. Then the algorithm was extended to the kernel space by nonlinear function and the distance between classes was

separated effectively. Through different transformations, the classification ability of person features is strengthened, and the re-identification accuracy with small sample is improved. Leng et al. [33] proposed a new semi-supervised collaborative metric learning. First, the input sample was decomposed into two pseudo views, and two distance metric matrices were learned through pseudo labels and reference information. Then, the two metric matrices were used to measure the relationship between the training data and the unlabeled data, the two groups of sorting results were obtained. Finally, the two groups of ranking results were used as reference information to supervise the learning of the metric matrix. This method can obtain a practical person re-identification model in the case of insufficient training samples or insufficient sample labels.

### 5.2 Performance analysis and discussion

In order to compare and analyze the performance of the metric learning in person re-identification more intuitively, this section combs the experimental results of some typical algorithms on common person datasets. Because the Ref. [6, 42, 71] have made a detailed introduction to the person re-identification datasets, in this section, we have not explained more about the person datasets. Experimental results of metric methods, classical metric learning algorithm, deep metric learning algorithm and re-ranking algorithm are mainly summarized, as shown in Tables 5, 6, 7 and 8.

**Table 5** Performance comparison and analysis of some metric methods

| Metric method | Reference | Experimental conditions | Dataset | Rank-1 | Rank-5 | Rank-10 | Rank-20 |
|---|---|---|---|---|---|---|---|
| Mahalanobis distance metric | [49] | Software:vs2010+opencv2.4.9<br>Hardware: 2.66GHz Intel core2 Quad CUP Q9400 and 4 GB RAM | VIPeR | 30.0 | 57.3 | 69.1 | 81.9 |
| | | | iLIDS | 51.0 | 78.7 | 85.3 | 91.4 |
| | | | CUHK01 | 22.9 | 44.6 | 54.9 | 65.9 |
| | [85] | – | PRID2011 | 70.9 | 78.7 | 82.7 | 87.3 |
| | | | iLIDS | 42.0 | 52.67 | 60.03 | 66.67 |
| Asymmetric distance metric | [10] | – | VIPeR | 43.29 | 72.66 | 83.51 | 92.18 |
| | | | PRID450S | 57.60 | 82.67 | 89.24 | 93.20 |
| | | | CUHK01 | 47.80 | 74.16 | 83.44 | 89.92 |
| | [73] | – | VIPeR | 30.9 | 51.7 | 61.6 | 72.3 |
| | | | CUHK01 | 57.3 | 80.0 | 86.3 | 91.8 |
| | | | CUHK03 | 31.9 | 59.4 | 70.1 | 80.0 |
| | [21] | – | VIPeR | 38.1 | 56.3 | 63.1 | 70.6 |
| | | | CUHK01 | 56.6 | 78.9 | 86.2 | 92.3 |
| | | | iLIDS | 19.7 | 38.6 | 48.9 | 62.4 |
| Hypergraphs-based Similarity metric | [2] | Software: Matlab<br>Hardware: 2.4 GHz Intel i7 CPU and 8 GB RAM | VIPeR | 34.18 | 66.60 | 79.75 | 90.19 |
| | | | CUHK01 | 35.01 | 58.25 | 69.28 | 80.62 |
| | [78] | Software: —— | VIPeR | 45.35 | 71.49 | 83.99 | 92.53 |
| | | Hardware: 3.4GHz Intel CPU i7 and 12G RAM | CUHK01 | 64.45 | 83.53 | 91.11 | 95.26 |
| | | | GRID | 23.68 | 43.92 | 52.56 | 61.76 |
| Set-based metric method | [36] | – | iLIDS | 65.90 | 87.30 | – | – |
| | | | Caviar 4REID | 32.60 | 65.40 | – | – |
| | [14] | – | PRID2011 | 76.0 | 94.0 | 98.0 | 99.0 |
| | | | iLIDS | 57.3 | 79.3 | 87.3 | 93.3 |
| | | | MARS | 66.3 | 82.2 | – | 89.9 |

**Table 6** Performance comparison and analysis of classical metric learning algorithms

| Classical metric learning | Reference | Experimental conditions | Dataset | Rank-1 | Rank-5 | Rank-10 | Rank-20 |
|---|---|---|---|---|---|---|---|
| Improved objective function | [83] | Software: —— | VIPeR | 36.6 | 67.5 | 80.1 | 90.2 |
| | | Hardware: 2.1GHz Intel E5 CPU | iLIDS | 43.5 | 63.9 | 75.2 | 86.8 |
| | [56] | Software: Matlab | VIPeR | 29.35 | 50.66 | 61.93 | 74.94 |
| | | Hardware: 2.20GHz 16 Intel Xeon CPUs (E5–2660, 8 cores,) and 64 GB RAM. | CUHK01 | 33.46 | 50.88 | 60.97 | 70.97 |
| | [18] | Software: Matlab | VIPeR | 52.18 | 80.54 | 88.92 | 95.60 |
| | | Hardware: 2.60 GHz Intel i7–3720 CPU | GRID | 24.32 | 46.24 | 57.12 | 68.08 |
| | [25] | Software: Matlab | VIPeR | 47.66 | 77.65 | 87.18 | 94.78 |
| | | Hardware: 3.2 GHz 4 Cores CPU and 12 GB RAM | Market1501 | 57.69 | 77.20 | 82.87 | – |
| | | | CUHK01 | 69.42 | 88.79 | 93.97 | 97.10 |
| | [61] | Software: Matlab | VIPeR | 51.46 | 79.56 | 89.08 | 95.85 |
| | | Hardware: 2.8 GHz 16 Cores CPU and 64GB RAM | CUHK01 | 74.22 | 90.51 | 94.42 | 96.91 |
| | | | CUHK03 | 73.21 | 94.48 | 97.93 | – |
| | [86] | – | CUHK03 | 58.72 | 89.07 | 95.28 | – |
| Multi-metric fusion | [50] | Software: vs2010+opencv2.4.9 | VIPeR | 40.7 | 72.37 | 83.95 | 92.08 |
| | | Hardware: Intel Xeon CPU E5506 2.13 GHz and 24GB memory | iLIDS | 38.3 | 66.5 | 79.0 | 88.3 |
| | | | CUHK01 | 36.10 | 62.68 | 72.61 | 81.90 |
| | [54] | – | VIPeR | 36.97 | 69.87 | 80.31 | 90.44 |
| | | | CUHK01 | 31.19 | 57.93 | 70.66 | 81.39 |

**Table 7** Performance comparison and analysis of deep metric learning algorithms

| Loss function | Reference | Experimental conditions | Dataset | Rank-1 | Rank-5 | Rank-10 | Rank-20 |
|---|---|---|---|---|---|---|---|
| Classification loss | [8] | Software: —— | Market-1501 | 80.31 | 91.27 | 94.09 | 96.02 |
| | | Hardware: NVIDIA GeForce GTX Titan X GPU and 3.00GHz Intel Core i7-5960X CPU, 64.0 GB RAM | CUHK03 | 69.55 | 90.96 | 95.07 | 97.54 |
| | [20] | – | VIPeR | 51.90 | 76.60 | 85.40 | 93.40 |
| | | | CUHK01 | 83.50 | 95.20 | 97.30 | 98.80 |
| | | | CUHK03 | 88.60 | 98.20 | 99.20 | 99.70 |
| | | | Market1501 | 88.40 | 94.80 | 96.50 | 98.00 |
| Contrastive loss | [62] | – | RRID2011 | 73.30 | – | 97.50 | 98.30 |
| | | | CUHK01 | 71.90 | 91.80 | 95.80 | 97.20 |
| | | | 2DPeS | 58.30 | 74.00 | – | 88.50 |
| | [7] | – | Market-1501 | 71.37 | 88.51 | 93.17 | – |
| | | | CUHK03 | 78.81 | 96.72 | 99.02 | – |
| | [87] | Software: MATLAB 2016 and Visual Studio 2015 | GRID | 21.20 | – | 54.24 | 65.84 |
| | | Hardware: NVIDIA GPU and Intel Core i7 CPU | VIPeR | 44.87 | – | 86.01 | 93.70 |
| | [51] | – | Market1501 | 84.40 | 93.10 | 95.20 | – |
| | | | DUKE | 71.70 | 83.50 | 87.10 | – |
| Triple loss | [68] | – | Market1501 | 84.26 | 93.59 | 95.99 | – |
| | | | DUKE | 74.50 | 87.66 | 90.98 | – |
| | [26] | Software: —— | VIPeR | 47.3 | 76.6 | 88.1 | – |
| | | Hardware: NVIDIA GTX1080-Ti GPU (11GB RAM) and Intel i7–7700 CPU (8 cores and 3.6GHz) | CUHK03 | 81.9 | 96.7 | 98.7 | – |
| | | | CUHK01 | 70.2 | 90.2 | 95.5 | – |
| Joint loss | [84] | – | CUHK01 | 77.89 | 93.22 | 96.61 | 98.68 |
| | | | CUHK03l | 63.58 | 89.17 | 93.75 | 98.25 |
| | | | PRID2011 | 72.54 | 94.61 | 100.00 | 100.00 |
| | [13] | Software: —— | iLIDS | 59.7 | 81.8 | 90.9 | 96.9 |
| | | Hardware: Tesla K40 GPU with 12GB RAM | PRID2011 | 26.0 | – | 49.0 | 58.0 |
| | | | CUHK03 | 84.7 | 97.6 | 98.9 | 99.6 |
| Cluster loss | [37] | Software: —— | CUHK01 | 79.3 | 94.0 | 97.2 | – |
| | | Hardware: Tesla K40 GPU with 12G memory, 3.6GHz Intel Core CPUs, 48 GB RAM, and an NVIDIA GTX TITAN X GPU | CUHK03 | 90.2 | 98.8 | – | – |

**Table 8** Performance comparison and analysis of re-ranking algorithms

| Re-ranking | Reference | Experimental conditions | Dataset | Rank-1 | Rank-5 | Rank-10 | Rank-20 |
|---|---|---|---|---|---|---|---|
| Aggregate re-ranking | [60] | – | VIPeR | 40.51 | 69.15 | 81.04 | 91.17 |
| | | | CUHK01 | 57.02 | 80.43 | 87.90 | 93.40 |
| | | | CUHK03 | 55.89 | 86.26 | 93.74 | 98.00 |
| | [77] | Software: MATLAB | VIPeR | 35.8 | 69.9 | 80.4 | 89.6 |
| | | Hardware: Xeon E5–2696 v2 CPU and a GTX TITAN GPU | CUHK01 | 52.0 | 71.8 | 79.9 | 85.9 |
| | | | iLIDS | 22.0 | 43.3 | 52.0 | 73.3 |
| | [76] | Software: MATLAB 2013 | VIPeR | 44.02 | – | 85.40 | 92.83 |
| | | Hardware: 2.5 GHz CPU and 8 GB RAM | CUHK01 | 65.03 | – | 91.26 | 95.33 |
| | [29] | Software: MATLAB | VIPeR | 67.21 | 87.78 | 93.39 | 97.82 |
| | | Hardware: 2.30 GHz Intel Core i7 CPU and 8 GB RAM | CUHK01 | 66.91 | 86.95 | 92.12 | 95.7 |
| Graph model re-ranking | [66] | – | VIPeR | 52.59 | 82.50 | 91.14 | 96.84 |
| | [5] | Software: MATLAB | VIPeR | 34.26 | 57.34 | 67.86 | 80.78 |
| | | Hardware: 1.7GHz Intel Core i3 CPU and 8GB RAM | | | | | |
| | [67] | Software: MATLAB | VIPeR | 45.19 | 73.58 | 85.35 | 93.99 |
| | | Hardware: 2.8 GHz Intel i5 dual core CPU | CUHK01 | 68.64 | 88.00 | 92.74 | 95.80 |
| Reward and punishment re-ranking | [11] | – | VIPeR | 52.85 | 81.96 | 90.51 | 95.73 |
| | | | CUHK01 | 57.28 | 81.07 | 88.44 | 93.46 |
| | [65] | Software: MATLAB R2014a | VIPeR | 30.23 | 58.73 | 69.18 | 79.32 |
| | | Hardware: 3.20 GHz Intel Core i5 CPU and 8G RAM | iLIDS | 48.30 | 74.50 | 86.50 | 96.20 |

The common criteria for evaluating the performance of person re-identification algorithms include the Cumulative Match Characteristic (CMC) and Rank-N table. The CMC curve reflects the probability of finding the correct result among the first $k$ matching results. The increasing trend of the CMC curve indicates the better recognition effect. Rank-N table gives the cumulative matching accuracy of key matching points in numerical form, such as Rank-1, Rank-5, Rank-10 and Rank-20. For example, Rank-5 represents the probability that can be correctly matched in the first 5 images. The higher probability value indicates better algorithm performance. Because the Rank-N table is more intuitive than CMC curve, therefore, in this paper, we compared the performance of different algorithms using Rank-N table.

The data in Table 5 show that the overall performance of Mahalanobis metric and improved Mahalanobis metric has a certain gap compared with other methods. The asymmetric distance metric in ref. [10] showed excellent performance in overcoming the influence of perspective and pose variation. Due to the unsupervised restriction in ref. [21, 73], the identification performance was reduced, but the overall performance was still improved compared with the Mahalanobis metric. Hypergraphs-based similarity metric and set-based similarity metric make comprehensive use of context relation and neighborhood relation, which improved the metric results significantly.

The experimental data demonstrate that, compared with the classical metric learning algorithm, the rank-1 accuracy of the deep metric learning is improved greatly. However, the deep metric learning algorithm does not perform better than the classical metric learning algorithm on the person datasets with insufficient training samples such as VIPeR. It indicates that the deep metric learning algorithm is more applicable to the large-scale person dataset. When rank-20 accuracy is compared, the performance difference between classical metric

learning and deep metric learning is very small. The classical metric learning has low dependence on large-scale training sets and low complexity, so it shows more obvious advantages in obtaining candidate result sets of person re-identification in a wide range. However, in person re-identification with high accuracy, the deep metric learning has more advantages due to the diversity of training samples and the strong nonlinear feature extraction ability of deep models.

In addition, compared with the classical metric learning algorithm, the rank-1accuracy after re-ranking is improved to a certain extent. For example, the optimal rank-1 accuracy on VIPeR dataset is 51.46% before re-ranking, and the optimal rank-1 accuracy reaches 52.85% after re-ranking. Comparison of rank-20 accuracy shows that the algorithm performance is improved significantly after re-ranking. It shows that re-ranking is effective for optimizing and correcting the initial metric results.

Finally, from the perspective of time evolution, the proposed algorithm after 2018 has a large performance improvement compared with the algorithm before 2016. This is mainly due to the gradual optimization and improvement of the metric learning algorithm, as well as the researchers' continuous comprehensive understanding of person re-identification. In general, with the continuous enrichment of person datasets, deep metric learning has become the mainstream direction of person re-identification based on metric learning theory. Re-ranking and fusion techniques play a positive role in improving the effect of classical metric learning and deep metric learning.

## 6 Summary and prospect

In this paper, we summarize the research progress of metric learning methods in person re-identification in recent years, and discuss various typical algorithms. Although the metric learning theory in person re-identification has made remarkable progress, different metric learning methods are usually suitable for different person re-identification tasks or application scenarios. Therefore, it is still difficult to find a universally applicable metric or metric learning method. In addition, the current metric learning theory has not been applied maturely in person re-identification of actual video surveillance, and it still faces many challenges. In the future research, there are still many contents that need to be explored and studied:

1) Small sample metric learning. In the real video surveillance, the persons are in a non-cooperative state, so it is very difficult to capture high-quality available person images. Under the small sample dataset, many deep metric learning methods are usually not fully trained, resulting in weak generalization ability and poor robustness of the learned distance metric matrix. In recent years, researchers have paid attention to the small sample person re-identification problem, but more research is to propose solutions from the perspective of sample data augmentation and transfer learning. However, there are few studies on the metric or metric learning methods directly targeting small sample person dataset. Therefore, in future work, in addition to better research on small sample data augmentation methods, an in-depth study should also be implemented from the perspective of metric learning.
2) Metric learning across image domains. Person re-identification under different resolutions, occlusion, cross-scenes, visible-infrared, long time intervals, etc. is often referred to as cross-domain image recognition [31]. Now, there are many studies on cross-domain person metric learning under one certain situation. However, person re-identification under real-world video surveillance is usually a synthesis of multiple cross-domain image

recognition problems. Therefore, how to integrate multiple cross-domain situations to define a more adaptive and robust metric or metric learning method has important theoretical significance and practical value.

3) Semi-supervised and unsupervised metric learning. Many current person re-identification methods, including metric learning, are often based on supervised information. However, the data labeling for obtaining supervised information is costly. Therefore, semi-supervised and unsupervised metric learning are more suitable for person re-identification under real-world video surveillance, but the performance of current unsupervised and semi-supervised metric learning is far inferior to supervised metric learning. Therefore, designing reliable and efficient unsupervised or semi-supervised metric learning algorithms to improve metric learning performance in the absence of supervised information is a key step in improving the performance of person re-identification.
4) New metric definition and re-ranking technology. In addition to the improvement of metric learning algorithm, referring to the current distance metric concept, the definition of a new distance metric plays an essential role in improving person re-identification, but it is more difficult to study. In addition, how to use the intelligent optimization algorithm to perform a deeper comprehensive judgment on the initial metric results is crucial to improve the accuracy of person re-identification, but there have been few relevant studies in the past two years.

**Acknowledgements** This work was funded by the Visiting Project Funds of Shandong University of Technology, the Integration Funds of Shandong University of Technology and Zhangdian District (No.118228), the National Natural Science Foundation of China (No. 61601266, No.61801272), the Natural Science Foundation of Shandong Province of China (No. ZR2015FL029, ZR2016FL14).

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

## References

1. Ali, TMF, Chaudhuri S. (2018) Maximum margin metric learning over discriminative nullspace for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV). : 122–138.
2. An L, Chen X, Yang S (2016) Person re-identification via hypergraph-based matching. Neurocomputing 182:247–254
3. Arindam S, Dibyadip C, Arpan B, et al. (2020) Open-set metric learning for erson re-identification in the wild. In Proceedings of the IEEE International Conference on Image Processing (ICIP), 2356–2360. https://doi.org/10.1109/ICIP40778.2020.9190744.
4. Barman A, Shah SK (2017) Distance aggregation based score fusion for improving person re-identification. In 2017 IEEE International Symposium on Technologies for Homeland Security (HST). IEEE, 2017: 1–8.
5. Barman A, Shah S K (2017) Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems. In Proceedings of the IEEE International Conference on Computer Vision. : 1115–1124.
6. Bedagkar-Gala A, Shah SK (2014) A survey of approaches and trends in person re-identification. Image Vis Comput 32(4):270–286
7. Bing C, Yufei Z, Yunqiang L et al (2018) Shift-variant similarity learning for person re-identification. J Electron Inf Technol 40(10):2381–2387

8. Borgia A, Hua Y, Kodirov E, Robertson NM (2018) Cross-view discriminative feature learning for person re-identification. IEEE Trans Image Process 27(11):5338–5349
9. Cai Y, Pietikäinen M (2010) Person re-identification based on global Y. Cai, M. Pietikäinen, Person re-identification based on global color context, The Tenth International Workshop on Visual Surveillance (in conjunction with ACCV 2010), pp. 205–215.
10. Chen YC, Zheng WS, Lai JH et al (2016) An asymmetric distance model for cross-view feature mapping in person reidentification. IEEE transactions on circuits and systems for video technology 27(8):1661–1675
11. Chen SZ, Guo CC, Lai JH (2016) Deep ranking for person re-identification via joint representation learning. IEEE Trans Image Process 25(5):2353–2367
12. Chen D, Xu D, Li H et al (2018) Group consistent similarity learning via deep crf for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 8649–8658
13. Cheng D, Gong Y, Shi W, Zhang S (2018) Person re-identification by the asymmetric triplet and identification loss function. Multimed Tools Appl 77(3):3533–3550
14. Cho YJ, Yoon KJ (2018) Pamm: pose-aware multi-shot matching for improving person re-identification. IEEE Trans Image Process 27(8):3739–3752
15. Davis J V, Kulis B, Jain P, et al. (2007) Information-theoretic metric learning. In Proceedings of the 24th international conference on Machine learning. ACM, 209–216.
16. De Maesschalck R, Jouan-Rimbaud D, Massart DL (2000) The mahalanobis distance. Chemom Intell Lab Syst 50(1):1–18
17. Dikmen M, Akbas E, Huang TS et al (2010) Pedestrian recognition with a learned metric. In Asian conference on Computer vision. Springer, Berlin, Heidelberg, pp 501–512
18. Dong H, Gong S, Liu C, Ji Y, Zhong S (2017) Large margin relative distance learning for person re-identification. IET Comput Vis 11(6):455–462
19. Dong H, Lu P, Liu C et al (2018) Learning multiple kernel metrics for iterative person re-identification. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 14(3):78–78
20. Feng Z, Lai J, Xie X (2018) Learning view-specific deep networks for person re-identification. IEEE Trans Image Process 27(7):3472–3483
21. Feng Y, Yuan Y, Lu X (2019) Person Reidentification via Unsupervised Cross-View Metric Learning. In: Person Reidentification via unsupervised cross-view metric learning. IEEE transactions on cybernetics
22. García J, Martinel N, Gardel A, Bravo I, Foresti GL, Micheloni C (2017) Discriminant context information analysis for post-ranking person re-identification. IEEE Trans Image Process 26(4):1650–1665
23. Gheissari N, Sebastian T B, Hartley R (2006) Person reidentification using spatiotemporal appearance. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 1528–1535.
24. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06) IEEE 2:1735–1742
25. He B, Yu S (2017) Ring-push metric learning for person reidentification. Journal of Electronic Imaging 26(3):033005
26. He Z, Zhang Z, Jung C (2018) Deep feature embedding learning for person re-identification using lifted structured loss. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 1957–1961.
27. Hou R, Ma B, Chang H, et al. (2020) IAUnet: Global context-aware feature learning for person reidentification. IEEE Transactions on Neural Networks and Learning Systems,, https://doi.org/10.1109/TNNLS.2020.3017939
28. Joachims T (2002) Optimizing search engines using click through data. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM,: 133–142.
29. Junior JCSJ, Baró X, Escalera S (2018) Exploiting feature representations through similarity learning, post-ranking and ranking aggregation for person re-identification. Image Vis Comput 79:76–85
30. Koestinger M, Hirzer M, Wohlhart P, et al. (2012) Large scale metric learning from equivalence constraints. In 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2288–2295.
31. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst:1097–1105
32. Krizhevsky A, Sutskever I, Hinton G E (2012) Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097–1105.
33. Leng Q (2018) Co-metric learning for person re-identification. Advances in Multimedia 2018:1–9
34. Leng Q, Hu R, Liang C, Wang Y, Chen J (2015) Person re-identification with content and context re-ranking. Multimed Tools Appl 74(17):6989–7014
35. Li W, Zhao R, Wang X (2012) Human reidentification with transferred metric learning. In Asian conference on computer vision. Springer, Berlin, Heidelberg, pp 31–44

36. Li W, Wu Y, Mukunoki M, Kuang Y, Minoh M (2015) Locality based discriminative measure for multiple-shot human re-identification. Neurocomputing 167:280–289
37. Li K, Ding Z, Li K, et al. (2018) Support neighbor loss for person re-identification. In 2018 ACM Multimedia Conference on Multimedia Conference. ACM, 2018: 1492–1500.
38. Liao S, Li S Z (2015) Efficient psd constrained asymmetric metric learning for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision. 3685–3693.
39. Liao S, Hu Y, Zhu X et al (2015) Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE conference on computer vision and pattern recognition:2197–2206
40. Ling H, Wang Z, Li P, Shi Y, Chen J, Zou F (2019) Improving person re-identification by multi-task learning. Neurocomputing 347:109–118
41. Liu Z, Lu H, Ruan X, Yang MH (2019) Person Reidentification by joint local distance metric and feature transformation. IEEE transactions on neural networks and learning systems 30(10):2999–3009
42. Luo Hao, Jiang Wei, Fan Xing, et.al. A survey on deep learning based person re-identification. Acta Automat Sin, 2019,45(11): 2032–2049.
43. Ma AJ, Li J, Yuen PC et al (2015) Cross-domain person reidentification using domain adaptation ranking svms. IEEE Trans Image Process 24(5):1599–1613
44. McLaughlin N, del Rincon JM, Miller PC (2016) Person reidentification using deep convnets with multitask learning. IEEE Transactions on Circuits and Systems for Video Technology 27(3):525–539
45. Olszewska JI (2016) Automated face recognition: challenges and solutions. Pattern Recognition Analysis and Applications
46. Pang Y, Cao J, Wang J, Han J (2019) JCS-net: joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. IEEE Transactions on Information Forensics and Security 14(12):3322–3331
47. Pedagadi S, Orwell J, Velastin S et al (2013) Local fisher discriminant analysis for pedestrian re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition:3318–3325
48. Porikli F (2003) Inter-camera color calibration by correlation model function. In: Proceedings of the 2003 International Conference on Image Processing. Barcelona, Spain: IEEE, II-133-6.
49. Qi Meibin, Wang Yunxia, Tan Shengshun, et.al. Person re-identification based on regularization of independent measure matrix. Pattern Recognition and Artificial Intelligence, 2016, 29( 6): 511–518.
50. Qi Mei-Bin, Tan Sheng-Shun, Wang Yun-Xia, et.al. Multi-feature subspace and kernel learning for person re-identification. Acta Automat Sin, 2016, 42(2):299–308.
51. Saquib Sarfraz M, Schumann A, Eberle A et al (2018) 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:420–429
52. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition:815–823
53. Sikdar A, Chowdhury AS (2020) Scale-invariant batch-adaptive residual learning for person re-identification. Pattern Recogn Lett 129:279–286
54. Syed M A, Jiao J (2016) Multi-kernel metric learning for person re-identification. In 2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016: 784–788.
55. Syed MA, Han Z, Li Z et al (2018) Impostor resilient multimodal metric learning for person Reidentification. Advances in Multimedia:2018.1–2018.201811
56. Tan S, Zheng F, Liu L et al (2016) Dense invariant feature-based support vector ranking for cross-camera person Reidentification. IEEE Transactions on Circuits and Systems for Video Technology 28(2):356–363
57. Tan F, Liu W, Huang L, Zhai C, Shi W, Li Y (2017) Person re-identification across multiple non-overlapping cameras by grouping similarity comparison model. Chin J Electron 26(5):905–911
58. Tao DP, Jin LW, Wang YF et al (2013) Person re-identification by regularized smoothing KISS metric learning. IEEE Trans on Circuits and Systems for Video Technology 23(10):1675–1685
59. Wang X, Zheng WS, Li X et al (2015) Cross-scenario transfer person reidentification. IEEE Transactions on Circuits and Systems for Video Technology 26(8):1447–1460
60. Wang J, Wang Z, Gao C et al (2016) Deeplist: learning deep features with adaptive listwise constraint for person reidentification. IEEE Transactions on Circuits and Systems for Video Technology 27(3):513–524
61. Wang J, Wang Z, Liang C, Gao C, Sang N (2018) Equidistance constrained metric learning for person re-identification. Pattern Recogn 74:38–51
62. Wang J, Zhou S, Wang J, Hou Q (2018) Deep ranking model by large adaptive margin learning for person re-identification. Pattern Recogn 74:241–252
63. Weinberger K Q, Saul LK (2008) Fast solvers and efficient implementations for distance metric learning. In Proceedings of the 25th international conference on Machine learning. ACM, 2008: 1160–1167.
64. Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10(Feb):207–244

65. Xiaokai L (2016) Pedestrian re-identification via coarse-to-fine ranking. IET Comput Vis 10(5):368–375
66. Xie Y, Levine MD, Yu H (2016) Person re-identification by graph-based metric fusion. Electron Lett 52(17):1447–1449
67. Xie Y, Yu H, Gong X, Levine MD (2017) Adaptive metric learning and probe-specific reranking for person reidentification. IEEE Signal Processing Letters 24(6):853–857
68. Yang X, Zhou P, Wang M (2018) Person reidentification via structural deep metric learning. IEEE Transactions on Neural Networks and Learning Systems 30(10):2987–2998
69. Yang H, Cheng Z, Chen L (2018) Reranking optimization for person re-identification under temporal-spatial information and common network consistency constraints. Pattern Recogn Lett:1–10
70. Ye M, Liang C, Yu Y, Wang Z, Leng Q, Xiao C, Chen J, Hu R (2016) Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Transactions on Multimedia 18(12):2553–2566
71. You-Jiao L, Li Z, Jing Z et al (2018) A survey of person re-identification. Acta Automat Sin 44(9):1554–1568
72. Yu B, Xu N (2018) Deep triplet-group network by exploiting symmetric and asymmetric information for person reidentification. Journal of Electronic Imaging 27(3):033033
73. Yu H X, Wu A, Zheng W S. (2017) Cross-view asymmetric metric learning for unsupervised person re-identification. In Proceedings of the IEEE International Conference on Computer Vision. 994–1002.
74. Yuan C, Guo J, Feng P, Zhao Z, Luo Y, Xu C, Wang T, Duan K (2019) Learning deep embedding with mini-cluster loss for person re-identification. Multimed Tools Appl 78(15):21145–21166
75. Yu-Ning D, Hai-Zhou A (2016) Learning quadratic similarity function for pedestrian re-identification. Chinese Journal of Computers 39(8):1639–1651
76. Zhang Z, Huang M (2017) Discriminative structural metric learning for person Reidentification in visual internet of things. IEEE Internet Things J 5(5):3361–3368
77. Zhang Z, Saligrama V (2016) Prism: person reidentification via structured matching. IEEE Transactions on Circuits and Systems for Video Technology 27(3):499–512
78. Zhao X, Wang N, Zhang Y, du S, Gao Y, Sun J (2017) Beyond pairwise matching: person reidentification via high-order relevance learning. IEEE transactions on neural networks and learning systems 29(8):3701–3714
79. Zhao C, Wang X, Zuo W et al (2020) Similarity learning with joint transfer constraints for person re-identification. Pattern Recogn 97(107014):1–10
80. Zheng W S, Gong S, Xiang T (2011) Person re-identification by probabilistic relative distance comparison. In 2011 IEEE conference on computer vision and pattern recognition. IEEE, 2011: 649–656.
81. Zheng WS, Gong S, Xiang T (2012) Reidentification by relative distance comparison. IEEE Trans Pattern Anal Mach Intell 35(3):653–668
82. Zheng WS, Gong S, Xiang T (2015) Towards open-world person re-identification by one-shot group-based verification. IEEE Trans Pattern Anal Mach Intell 38(3):591–606
83. Zhou Q, Zheng S, Yang H, et al. (2016) Joint instance and feature importance re-weighting for person re-identification. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 1546–1550.
84. Zhou S, Wang J, Shi R et al (2017) Large margin learning in set-to-set similarity comparison for person reidentification. IEEE Transactions on Multimedia 20(3):593–604
85. Zhou J, Su B, Wu Y (2018) Easy identification from better constraints: multi-shot person re-identification from reference constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition:5373–5381
86. Zhou Z, Liu K, Huang J et al (2019) Improved metric learning algorithm for person re-identification based on equidistance. J Electron Inf Technol 41(2):477–483
87. Zhu J, Zeng H, Liao S et al (2017) Deep hybrid similarity learning for person re-identification. IEEE Transactions on Circuits and Systems for Video Technology 28(11):3183–3193
88. Zhu F, Kong X, Wu Q, Fu H, Li M (2018) A loss combination based deep model for person re-identification. Multimed Tools Appl 77(3):3049–3069
89. Zongyuan D, Wang H, Fuhua C et al (2017) Person re-identification based on distance centralization and projection vectors learning. Journal of Computer Research and Development 54(8):1785–1794