Fused-Grain Feature Learning for Unsupervised Person Re-identification

Hua HAN*, Li HUANG, Yujin ZHANG, Jiamin TANG

Shanghai University of Engineering Science, School of Electronic and Electrical Engineering, Shanghai, 201620, China
2070967@mail.dhu.edu.cn (**Corresponding author*), huanglili622@126.com, yjzhang@sues.edu.cn, 151066723@qq.com

Abstract: Most supervised learning methods are currently used to solve the task of person re-identification (Re-ID) and yield excellent results. But these methods usually need manual annotation of training data. Especially for large data sets, they need too high cost of manual annotation and the data are difficult to obtain for fully pairwise labeling. So unsupervised learning becomes a necessarily trend for person Re-ID. This paper is trying to solve the problem by unsupervised learning method. Moreover, global features focus on spatial integrity of person features, and local ones help to highlight discriminative features of different patches. Therefore, fused-grained unsupervised (FGU) learning framework of global and local branches' feature learning is proposed to solve Re-ID task. Specifically, for the local branch, one extracts patches from a feature map which learned on a PatchNet network of images, and learns their fine-grained features to pull close the similar patches and push away the dissimilar ones. For the global branch, one maximizes the diversity between classes by repelled loss and similarity within classes through attracted loss, then similarity and diversity in the unlabeled data sets are used as information for unsupervised cluster merging and learning their coarse-grained features. The two branches are used to jointly achieve the effect of increasing inter-class differences and intra-class similarity. A large number of experiments verify the superiority of the proposed method for unsupervised person re-identification.

Keywords: Unsupervised Learning, Fused-Grained Feature Learning, Cluster Merging, Spatial Integrity.

1. Introduction

Person re-identification is derived from multiple camera tracking which needs to determine whether different images of people captured in non-overlapping fields of view belong to the same person. The problem of correlative identification of person target in surveillant network of non-overlapping area is collectively referred to as person re-identification problem by the researchers. Person Re-ID is based on the use of computer vision, machine learning, pattern recognition, image processing, and others. It is an intersection of these frontier fields and can be widely used in areas such as intelligent video surveillance, security, and criminal investigations. The appearance of a person with both rigid and flexible characteristics would be susceptible to be impacted by clothing and posture, changes in viewing angle, light and shade in the complex environment, which make the task face huge technical challenges. However, in recent years, person re-identification technologies have already received a wide attention in academia and industry, and many excellent methods have been proposed (Gong et al., 2014), which make it become one of the research hotspots in the field of computer vision.

The study of re-identification dates back to 2003, when Porikli (2003) used a correlation coefficient matrix to build a non-parametric model between camera pairs, which can obtain the color distribution changes of the target between different cameras, and achieve cross-view target matching. Gheissari et al. (2006) were the first to introduce the concept of person re-identification, and proposed the usage of color and Salient Edge Histogram for Person Re-ID.

Person Re-ID consists of two core parts:

1. Feature extraction and representation. Based on the appearance of a person, a feature representation vector with robust and strong differentiation is extracted to effectively represent the person.

2. Measures of similarity. Similarity comparisons between feature vectors are used to determine the similarity of persons. The development of person Re-ID techniques can be divided into two stages, depending on feature extraction and representation: 1) Stage of artificial design feature before 2012; 2) Stage of deep feature learning after 2012 (Li et al., 2018).

Features are the basis of person re-identification, and feature quality directly impacts the final recognition performance. By adopting a reasonable similarity metric method, matching rate can be further improved. Many existing methods of person re-identification are attempting to establish a robust feature representation (Cheng et al., 2011,

https://doi.org/10.24846/v31i2y202204

Gray & Tao, 2008), and learn coarse-grained features from a global perspective. Cheng et al. (2011) exploited an image structure, which took into consideration part-based color information and color shifts for human re-identification. By using AdaBoost, Gray & Tao (2008) proposed a method to select good features from a set of color and texture features. Ma et al. (2012) converted a local descriptor to Fisher Vector to generate a global representation of an image. Farenzena, et al. (2010) presented a symmetry-driven accumulation of local features (SDALF) method with symmetry and asymmetry to address viewpoint variability. The metric system is a very important element in person Re-ID, and many methods have been applied in computer vision and proved to be effective (Subotic et al., 2020, Yousuf Uddin et al., 2021, Rădulescu & Rădulescu, 2020, Han et al., 2020, 2021b, Ma et al., 2020), for example, Keep It Simple and Straightforward Metric learning (KISSME) (Kostinger et al., 2012), Locally-Adaptive Decision Function (LADF) (Li et al., 2013) and Cross-View Quadratic Discriminant Analysis (XQDA) (Liao et al., 2015). These algorithms have shown excellent results in face recognition and person Re-ID. By considering the log-likelihood ratio test of two Gaussian distributions, KISSME can obtain a simplified and very efficient solution. LADF is a joint model of distance metric and local adaptive threshold rules, which aims to create a unified quadratic classifier. XQDA simultaneously learns identification subspace and distance metrics, and it also can perform dimensionality-reduction and select the best dimensions.

Since Krizhevsky et al. (2012) won ILSVRC'12 classification contest, deep learning based on convolutional neural networks (CNN) is widely used in the field of computer vision (Zhang et al., 2021; Han et al., 2021a; Wang et al., 2020). Li et al. (2014) were the first to apply deep learning to pedestrian re-identification, and proposed a CNN-based filter pairing neural network (FPNN). Various person re-identification methods based on deep learning have been proposed, and they usually perform much better than conventional methods.

1.1 Motivation

The study investigates the challenges of deep learning systems in the field of unsupervised learning. This paper mainly studies the following two problems.

1. In unsupervised learning, triplet loss function is widely used due to its ability to close distance between sample and its positive one. However, it just learns the relative distance between samples, and considers only the differences between classes, does not learn absolute distance, and ignores similarity within classes.

2. Since different pictures of the same person in different cameras would have relatively large differences in appearance, the differences will be amplified if the maximum distance criterion is used for cluster merging, which would result in a failure of merging pictures of the same person from different cameras.

1.2 Main Contributions

This paper proposes a fused-grained unsupervised learning framework (FGU) for person Re-ID. Patch-based discriminative feature learning loss (PEDAL) (Yang et al., 2019) and unlabeled datasets are used to guide and learn fine-grained features with discriminative properties to close the similar patch features and push away the dissimilar ones. At the same time, each image in unsupervised datasets has no identity label, so it is assigned to its own clustering center at the beginning. Samples are pushed farther apart during learning process for expanding the diversity between each training sample. However, images with same identity still have similar visual features, so they will still be closer in feature space. The convolution model realizes parameter updating by maximizing the difference between cluster centers. In order to increase feature similarity of a same identity, the features in a same cluster are gathered to the center to maximize similarity within cluster. A combined loss based on repelled and attracted feature learning (RAFL) is used to guide coarse-grained feature learning of unlabeled dataset.

The main contributions of this work are as follows:

1. For the first problem, this paper innovatively proposes a method to increase inter-class differences through guidance of repelled loss and increase intra-class similarity through guidance of attracted loss. The method is well suited for feature learning and parameter updating.

2. For the second problem, this paper uses minimum distance criterion to cluster people

with the same identity who differ widely across cameras, and then gradually merge multiple clusters.

This paper is organized as follows. Section 2 presents the proposed method, Section 3 shows the experiment and analysis, Section 4 treats the main results, and Section 5 concludes the paper.

2. The Proposed Method

2.1 Overview of the FGU

In this unsupervised person re-identification work, a framework based on fused-grained feature learning is developed to obtain discriminative features from both global and local channels. As shown in Figure 1, *U* patches are firstly obtained for each feature map. *U* CNNs are used to extract their fine-grained features separately for *U* patches, and *U* losses are obtained, the mean of which is regarded as FEDAL Loss. Next, different cluster centers are assigned for each image. In Figure 1, each circle represents an image, and the same color stands for the similar identity. The similar samples can be gradually merged with an identity by a cluster algorithm. By combining repelled and attracted loss, RAFL loss is obtained to pull the similar images together and push away the dissimilar ones.

In terms of global branch, parameters are firstly updated by allowing the model to distinguish different people, so that the differences between these features can be expanded. By using a likely cross-entropy exclusion loss function, the convolutional model is optimized and the variance between different people is increased. Then, an attracted loss function that takes advantage of similarity of identities and treats pictures of similar features as the same person is used to reduce variability among clusters. They are then aggregated to form a cluster, and the convolutional model will update the parameters by maximizing the differences between cluster centers. The coarse-grained feature loss function (RAFL) is obtained by combining repelled and attracted loss functions. Finally, in order to minimize the variation within clusters, all the features that are in the same cluster are grouped towards its center, which increases similarity of features with the same identity. Using the structured information in feature space, data are clustered and merged via minimum distance criterion.

As for the local branch, in order to provide discriminative guided learning for local features in an unlabeled dataset, a fine-grained feature learning loss (PEDAL) based on patch-based discriminative feature is used to pull similar partial blocks closer together and push dissimilar partial ones further apart.

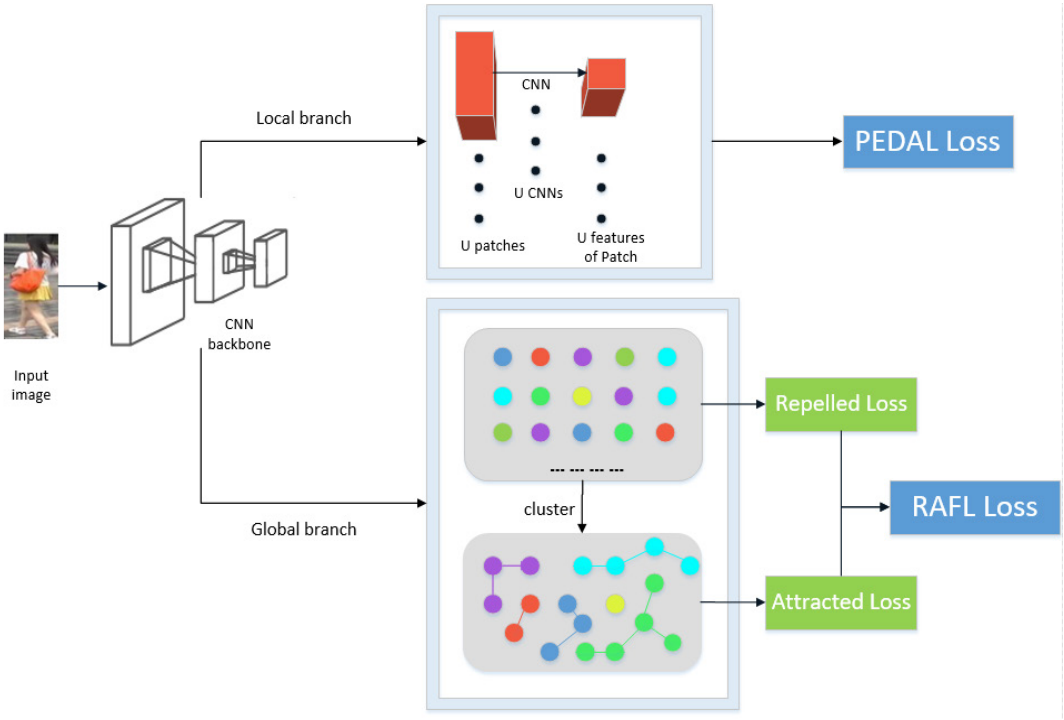

Figure 1. An illustration of the proposed FGU

2.2 Loss Functions

2.2.1 Coarse-grained Feature Loss of Global Branch

1. *Repelled loss*

None of the images of the known unsupervised datasets has an identity label, so each image is assigned to its own clustering center at the beginning, *i.e.* $\{\hat{y}_i = i | 1 \le i \le N\}$ ($\hat{y}_i$ is a dynastical index with variational numbers of cluster for x_i , x_i is the *i*-th image, *N* is the number of images, *i* is the index of *N*).The advantage of this approach is that the network can learn to recognize each training sample cluster rather than each person, and can maximize diversity between training samples. Then, as data parameters is updated, similar images are merged into the same clustering center step by step as a representation of an identity.

The possibility that image x belongs to the cluster is as follows:

$$p(c \mid x_i, \mathbf{V}) = \frac{\exp(\mathbf{V}_c^T v_i / \tau)}{\sum_{j=1}^{C}, \exp(\mathbf{V}_j^T v_i / \tau)} \quad (1)$$

where *C* is the number of clusters in the current state. In the starting state, *C*=*N* (*N* represents the number of training samples), that is, the number of clusters is equal to the number of images, and the number of clusters *C* gradually decreases as similar images are merged in the following stages. $v = \frac{\phi(\theta; x_i)}{\|\phi(\theta; x_i)\|}$ indicates the l_2 paradigm of data x_i in the feature space, and $\|v_i\| = 1$, $\mathbf{V} \in R^{C \times n_\phi}$ is a query list which saved each feature of cluster. V_j is the feature of column j in $\mathbf{V}$. τ is a scalar parameter. It is introduced to control the softness of the probability on class, like the peaks of the probability distribution produced by the softmax function. Since the range of similarity would be fixed between [-1,1] if cosine were used, so it is then necessary to introduce a scalar factor to control it. In subsequent experiments, it will be set to 0.1, according to Xiao et al. (2016).

In the foreground, the cosine similarity between data and all the other data is calculated by $\mathbf{V}^T \cdot v_i$. The backward processing updates the $\hat{y}_i - th$ column of the table with $\mathbf{V}_{\hat{y}_i} \leftarrow (\mathbf{V}_{y_i} + v_i)/2$. The original clustered features are summed and averaged with the new data features. The convolutional model can then be optimized using a likely cross-entropy loss function as a repelled loss. The variation between different people is enlarged through equation (2), since cross-entropy itself is a loss function that distinguishes categories.

$$L_r = -\log(p(c \mid x, \mathbf{V})) \quad (2)$$

By minimizing the repelled loss of equation (2), it can calculate and maximize the cosine distance between each image feature v_i and each central feature $\mathbf{V}_{j \ne \hat{y}_i}$ of cluster. It can also calculate and minimize the cosine distance between each image feature v_i and the corresponding central feature $\mathbf{V}_{j \ne \hat{y}_i}$ of the cluster. Therefore, the similarity and diversity can be balanced in order to pull close the similar images and push away dissimilar ones.

In the optimization process, $\mathbf{V}_j$ contains the features of all images in the $j - th$ cluster, and can be considered as the "center" of cluster. In every training stage, due to the high time complexity of directly calculating cluster centers, redundant calculations are avoided by querying table $\mathbf{V}$. So there is no need to extract features from all the training data from one time to the next.

2. *Attracted loss*

In addition to considering the repelled loss function, the present works intends to distinguish between different clustering centers while reducing the intra-class differences, so the loss function of attraction is proposed as follows:

$$L_a = \frac{1}{2} \sum_{i=1}^{m} \| v_i - c_{y_i} \|_2^2 \quad (3)$$

where $c_{y_i} \in R^d$ represents the $y_i - th$ feature embedding of clustering centers. At each iteration, calculate the mean value of all features belonging to cluster after merging step, and use it as cluster center feature. In each small batch, parameters are updated via equation (4). Repelled and attracted loss functions are jointly used as signals to train convolutional models for unsupervised feature learning.

The coarse-grain loss function (RAFL) formula for the global branch at this point is as follows:

$$\begin{aligned} L_g &= L_r + \beta L_\alpha \\ &= -\sum_{i=1}^{m} \frac{\exp(\mathbf{V}_c^T v_i / \tau)}{\sum_{j=1}^{C}, \exp(\mathbf{V}_j^T v_i / \tau)} + \frac{\beta}{2} \sum_{i=1}^{m} \| v_i - c_{y_i} \|_2^2 \end{aligned} \quad (4)$$

where β is a parameter to balance the two losses and m represents the total number of images.

This loss function is used for feature learning and parameters updating for the model.

3. *Strategy for cluster merging*

A similar measure is available, a numerical-based cluster criterion is also needed, which can classify similar samples into the same classes and dissimilar samples into different classes. A key to cluster merging is the calculation of the distance between clusters formed in each iteration, and between clusters and samples. Different distance functions will get different calculation results. The main distance calculation criteria are as follows: Minimum distance, Maximum distance, Median method, Centroid method, Average linkage, *etc*. However, in the current case, using a maximum distance criterion would not only result in a failure to merge the same identity under different cameras, but also in a failure to merge images of different identities from the same camera. When using intermediate distance criterion, there is an over-reliance on camera information and a lack of extraction of information about person themselves. These would result in ignoring the differences of samples in a same cluster. Therefore, it was decided to use a minimum distance criterion to calculate the dissimilarity value $D(A,B)$ between cluster A and B.

At the beginning of the process, training samples tend to be pushed away from each other in feature learning space, and every image is assigned as its own independent cluster center. But the images of the same identity still have similar visual features, so they should be relatively close in feature space. In this way, the similarity of the same identity image is exploited. In the process of exploring similarity, structured information from feature space was used to merge datum into clusters. At the beginning, the shortest distance between each image was used as dissimilarity. Depending on similarity between clusters, multiple cluster pairs will be merged.

The minimum distance criterion is a measure of dissimilarity that takes the shortest distance between all the images of two clusters. By using this criterion, the result will be the following:

As long as there are really similar image pairs in two clusters, the two clusters should be merged, no matter how much the other images look unlike. The reason for this is that the images of people will be more similar under the same camera, and the images of the same person will have greater differences under different cameras, whereas cross-camera images have more useful information. Because such images can provide more information for difficult samples, it is still difficult for direct cluster to bring such images together. So, by using minimum distance criterion it is possible to exploit similarity of people in the same camera to cluster people who are more different in across-cameras, which can guarantee the accuracy of image merge. The formula of merging is as follows:

$$D_{distance}(A,B) = \min_{x_a \in A, x_b \in B} d(x_a, x_b) \tag{5}$$

where $d(x_a, x_b)$ denotes Euclidean distance of images in feature space, i.e., $d(x_a, x_b) = \|v_a - v_b\|$. At each merge step, it is intended to reduce the number of n clustering centers. $n = N \times \gamma$ is defined, where $\gamma \in (0,1)$ denotes the speed of merge. Each time the clusters are merged, the n clusters with the smallest distances are merged. The number of clusters at the very beginning is initialized as $C = N$, which means that each training sample is a cluster. After t times of cluster merging, the number of clusters is reduced to $C = N - t \times n$.

2.2.2 Fine-Grained Features Loss of Local Branch

In order to further improve the performance of deep learning in person Re-ID, the main focus of many studies has been on the enhancement of local features. Some studies have aligned and matched the global images to improve metric learning performance while others have taken an area-based approach to localize body parts, using space to enhance attention regarding local features. There are also studies that use coarse slicing approaches to divide and express local features of human body. In short, how to combine the information of local features well to improve the matching rate, and to further make network performance greater are the tasks of local branch to be achieved.

In this work, a patch-based discriminative feature loss function (PEDAL) is used in an unsupervised framework to pull similar features close and push the dissimilar ones away, and to learn patch features in unlabeled datasets. The formula is as follows:

$$L_s^u = -\log \frac{\sum_{\mathbf{w}_j^u \in k_i^u} e^{-\frac{s}{2}\|x_i^u - \mathbf{w}_j^u\|_2^2}}{\sum_{j=1, j\neq i}^{N} e^{-\frac{s}{2}\|x_i^u - \mathbf{w}_j^u\|_2^2}} \tag{6}$$

where $W^u = \{\mathbf{W}_j^u\}_{j=1}^N$ is used to store patch features in batches (Wu et al., 2018, Xiao et

al., 2017). N is the number of training images. x_i^u denotes the feature of the $u-th$ patch of the $i-th$ image in a batch. k_i^u is a collection of k nearest patches of x_i^u which calculate the pairwise distance of W^u via each x_i^u. s is a scaling parameter. Because the features of similar images of people are directly pulled together, it may be possible to combine the features of people with different identities but visually similar, which would ignore the identity information and lead to a lower matching rate. So, by dividing a person's image into parts, different patches of the same image would contain different information about that person, and would bring out potential concealed information.

2.2.3 Overall Loss Function

Based on coarse-grained loss function (RAFL) and fine-grained loss function (PEDAL) used in the above unlabeled dataset framework, the final total loss function formed of each image in the batch can be expressed as:

$$L = \lambda \frac{1}{U} \sum_{u=1}^{U} L_s^u + L_g \tag{7}$$

where U denotes the numbers of patches of each image and λ is a parameter to control the weight.

3. Experimental Analysis

This section will demonstrate the effectiveness and progress of the proposed method through some experimental results. Firstly, the datasets used in the experiments will be presented, then evaluation metrics for the results of the whole task will be introduced, afterwards the details of the algorithm will be presented, and finally, the method will be compared with the current state-of-the-art to validate its effectiveness.

3.1 Dataset

Two of the more commonly used large-scale datasets for person re-identification are employed, as show in Table 1: Market-1501, DukeMTMC-reID. Market-1501 has 1501 identities captured by 6 cameras from a total of 32 668 detected images. Each person was captured by at least two cameras and there may be more than one image in one camera. The experiments made based on this dataset include a training set which consists of 751 identities and contains 12 936 images, and a test set which consists of 750 identities and contains 19 732 images. Query images were randomly selected from 750 identities in the test set, so that one person has up to 6 queries, and a total of 3 368 images. DukeMTMC-reID is a person Re-ID version of DukeMTMC dataset. It contains 1404 identities captured by 8 cameras from a total of 36 411 images. The experiments made based on this dataset include a training set which consists of 702 identities and contains 16 522 images, and a test set which consists of 702 identities and contains 17 661 images. For each camera, query images were randomly selected from 702 identities in the test set. The two datasets have a lot of changes in viewpoint variation, occlusion, illumination, pose, etc.

3.2 Evaluation Criteria

Cumulative matching characteristic (CMC) curves and mean average precision (mAP) were used to evaluate the performance of the proposed method. For each query image, its average precision (AP) is determined by its accuracy-recall curve, and the mean precision (mAP) is obtained by calculating the mean precision of all the query images which reflects the recall rate. The cumulative matching characteristic curve (CMC) is calculated from the score of Rank-l, Rank-5, Rank-10 which reflect the retrieve precision.

3.3 Implementation Details

ResNet-50 was used as the backbone network of the convolutional network for all experiments, and pre-trained weights of ImageNet were used to initialize the model. The last fully connected layer was removed and the stride of the last residual block was set to 1. For local branches, feature map was divided into U horizontal stripe blocks with

Table 1. Dataset description

Dataset	Cams	Identities	Images	Train	Test
				images	images
Market-1501	6	1501	32668	12936	19732
DukeMTMC-reID	8	1404	36411	16522	19889

the same size. In FEDAL loss function, k is set to 15, that is, for a patch x_i^u in an image, a set k_i^u was obtained from the top 15 nearest patches of x_i^u which were selected by calculating the pairwise distance. The scale number s varied with different data sets. It was set to 15 in Market-1501 and to 5 in DukeMTMC-reID. For global branch, β was set to 0.5, meanwhile τ was set to 0.1 in equation (4), and the clustering speed $\gamma \in (0,1)$ was set to 0.05. λ was set to 0.8 in equation (7). The training epoch was 60, and the batch size was 32. Stochastic descent gradient method with a momentum of 0.9 was used to train model. The learning rate was set to 0.0001 at the initial initialization, and decreased by 0.1 every 50 epochs.

4. Main Results

4.1 Ablation Study

1. *The effectiveness of attracted loss in global branch*

A comparison experiment was conducted by using the model with and without attracted loss in the global branch, and the results are shown in Table 2. The performance of attracted loss achieves an improvement on both Market-1501 and DukeMTMC-reID datasets. Specifically, the addition of attracted loss results in a 3.2% improvement in rank-1 accuracy on Market-1501 dataset and a 3.5% improvement on DukeMTMC-reID dataset. There will be a large intra-class variation in learning feature embeddings in the feature space if there is no attracted loss in global branch. Conversely, adding an attracted loss will gather cluster centers together. Learning each clustering center at the same time and setting appropriate value of β will significantly improve the discriminative power of depth features.

2. *The effectiveness of the cluster merge criterion*

Minimum distance criterion was selected from several cluster merging criteria, by comparing the results on Market-1501, was shown in Table 3.

The minimum distance criterion has the highest rank-1 accuracy, followed by the center distance criterion, with a slight decrease, while the maximum distance criterion has the lowest accuracy. Since images in dataset are from different cameras, different images of the same person captured by different cameras will have relatively large differences in appearance characteristics. So, the usage of maximum distance criterion will amplify the differences, which would result in the inability to merge images of the same person from different cameras.

3. *The effectiveness of FEDAL in local branch*

Here the effectiveness of FEDAL function is verified and the experimental results can be seen in Table 4. The performance results with FEDAL function are better than those obtained without it on both datasets. This happens mainly because the loss function provides an effective guide. For

Table 2. The result of the model whether or not has the attracted loss on unlabeled datasets

methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
no attracted loss	70.2	41.0	68.7	48.8
attracted loss	73.4	43.2	72.2	53.6

Table 3. The result of three common cluster merge criterions on Market-1501 dataset

Criterion	Rank-1	Rank-5	Rank-10	mAP
Maximum	69.3	41.4	70.6	51.9
Centroid	72.8	42.7	70.9	53.1
Minimum	73.4	43.2	72.2	53.6

Table 4. The result of FEDAL in local branch on unlabeled datasets

methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
no FEDAL loss	70.5	40.7	67.3	48.8
FEDAL loss	73.4	43.2	72.2	53.6

each person image, features located in different image regions have different information. By using different network branches, patch features can be learned in different parts of images on unlabeled datasets. Thus, the features that may be overlooked can be extracted more accurately, and eventually a more discriminating model can be obtained for different local features of a person. Furthermore, there is a high likelihood that similar areas will exist in similar images. The similarity between regional patches is much greater than that between whole images. But images with similar patches are not necessarily similar samples, so using FEDAL in local branch can also reduce the error rate of determining whether two images belong to the same person.

4.2 Further Analysis

1. *Analysis of parameter k in PEDAL*

In local branches, PEDAL is used, where the parameter k is used to determine the number of similar patches of an image patch. This threshold is employed to determine and distinguish whether it is a positive or a negative sample patch. As shown in Figure 2 and Table 5, if k is too small, many samples with a same identity will be lost. Otherwise, if k is too large, many samples with different identities will be pushed closer together. Both of the two situations will lead to large deviation in the results.

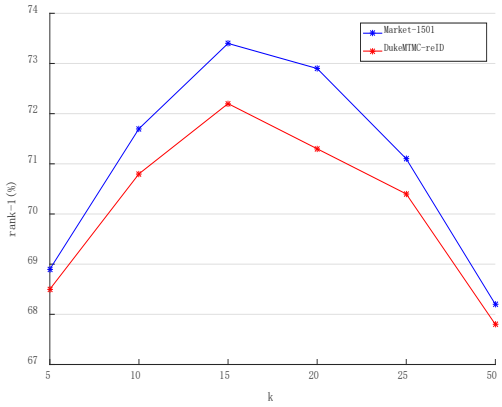

Figure 2. Analysis of the parameter k in PEDAL

When $k=[10, 20]$, the performance of model is in a smooth and better stage. When k=15, experiment achieves best result.

2. *Analysis of parameter β of RAFL*

On global branch, the joint of repelled and attracted loss functions is used, because that repelled loss can amplify the variation between samples of different identities and attracted loss can push similar identities closer. As shown in Figure 3 and Table 6, when $\beta = 0.5$, the results are optimal. This leads to the conclusion that repelled loss provides more energy than attracted loss.

3. *Analysis of cluster merging*

For cluster merging strategy, as previously shown in Table 3, the performance was improved by

Table 5. Analysis of the parameter k in PEDAL

	k	5	10	15	20	25	50
Rank-1	Market-1501	68.9	71.7	73.4	72.9	71.1	68.2
	DukeMTMC-reID	68.5	70.8	72.2	71.3	70.4	67.8

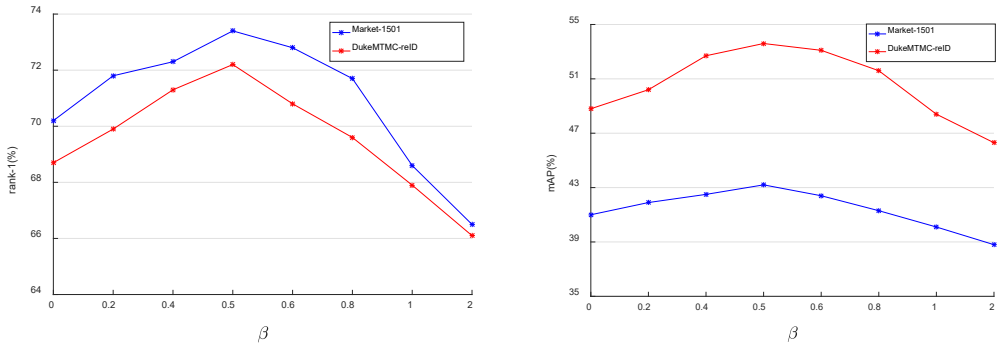

Figure 3. Analysis of the parameter β of RAFL

Table 6. Analysis of the parameter β of *RAFL*

	β	0	0.2	0.4	0.5	0.6	0.8	1	2
Market-1501	Rank-1	70.2	71.8	72.3	73.4	72.8	71.7	68.6	66.5
	mAP	41.0	41.9	42.5	43.2	42.4	41.3	40.1	38.8
DukeMTMC-reID	Rank-1	68.7	69.9	71.3	72.2	70.8	69.6	67.9	66.1
	mAP	48.8	50.2	52.7	53.6	53.1	51.6	48.4	46.3

cluster merging on the Market-1501 dataset which indicated that, as the number of the remaining clusters is gradually reduced with one merge at a time, rank-1 and mAP values are also improving very much. The number of clusters also slowly decreases with merging from single-sample-single-cluster at the beginning. Thus, it can be observed that both reduction of clusters and improvement of results are continuous and progressive. From diverse and similar images, one learns to obtain discriminative feature representations.

4. *Analysis of weight λ of final total loss function*

The effect of parameter λ in total loss is shown in Figure 4 and Table 7. Through the combination of global RAFL and local PEDAL, better results can be achieved, where PEDAL learns discriminative fine-grained features and RAFL guides discriminative coarse-grained features. The rank-1 and mAP steadily increase with weight λ. λ was set to 0.8, because that RAFL contributed a little more to the results.

4.3 Comparison with State-of -the-art

The proposed method was compared with the state-of-the-art when testing it on Market-1501 and DukeMTMC-reID datasets, including: (1) Models based on hand-crafted feature representation; (2) Models based on deep learning feature representation, such as: a) fake label learning; b) unsupervised domain adaptation. Table 8 shows the results of these comparisons.

1. *Comparison with models based on hand-crafted feature representation*

When the proposed method was compared to the hand-crafted feature methods, for example Local Maximal Occurrence (LOMO) (Liao et al., 2015), Unsupervised Multitasking Dictionary Learning Bag-of-Words (Bow)) (Zheng et al., 2015), UMDL (Peng et al., 2016), it proved to be significantly better than them. This happens mainly because the research on hand-crafted features started early in the research of person Re-ID. Taking into account that most of the early studies are based on idea-based design, and that there were not many learning methods that could be referred to, it was no possible to learn excellent discriminative features.

Table 7. Analysis of the weight λ of the final total loss of function

	λ	0	0.3	0.6	0.7	0.8	0.9	1	2
Market-1501	Rank-1	70.5	71.7	72.3	72.9	73.4	73.1	72.6	71.8
	mAP	40.7	41.9	43.1	43.8	43.2	43.0	42.4	41.3
DukeMTMC-reID	Rank-1	67.3	69.1	70.9	71.5	72.2	71.2	70.4	68.7
	mAP	48.8	49.9	52.0	53.1	53.6	52.4	51.7	50.3

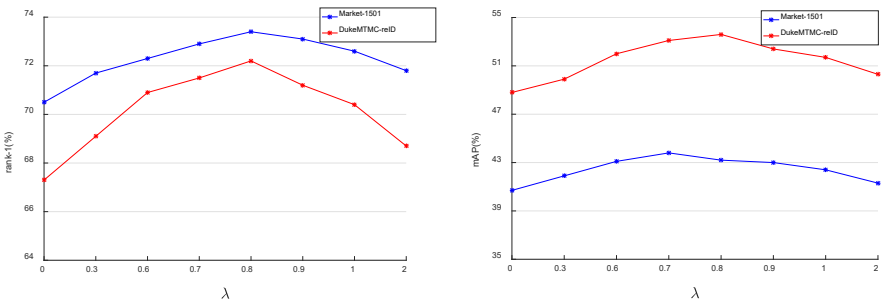

Figure 4. Analysis of the weight λ of the final total loss function

Table 8. Performance (%) comparison to the state-of-the-art on Market-1501 and DukeMTMC-reID

datasets		Market-1501				DukeMTMC-reID			
models	Methods	Rank-1	Rank-5	Rank-10	mAP	q	Rank-5	Rank-10	mAP
hand-crafted feature	LOMO (2015)	27.2	41.6	49.1	8.0	12.3	21.3	26.6	4.8
	Bow (2015)	35.8	52.4	60.3	14.8	17.1	28.8	34.9	8.3
	UMDL (2016)	34.5	52.6	59.6	12.4	18.5	31.4	37.6	7.3
fake label learning	CAMEL (2015)	54.5	-	-	26.3	40.3	57.6	-	26.3
	PUL (2018)	45.5	60.7	66.7	20.5	30.0	43.4	48.5	16.4
	DECAMEL (2020)	60.2	76.0	81.1	32.4	-	-	-	-
unsupervised domain adaptation	PTGAN (2018)	38.4	-	66.1	-	27.4	-	50.7	-
	TJ-AIDL (2018)	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
	SPGAN (2018)	51.5	70.1	76.8	22.8	41.1	56.6	63.0	22.3
	MMFA (2018)	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7
	CamStyle (2018)	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
	HHL (2018)	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
	MAR (2019)	67.7	81.9	87.3	40.0	67.1	79.8	84.2	48.0
	NSSA (2020)	-	-	-	-	65.5	77.9	81.3	45.5
The proposed method	**FGU**	**73.46**	**86.40**	**90.08**	**43.20**	**72.17**	**81.91**	**85.46**	**53.60**

2. *Comparison with models based on deep learning feature representation*

a. fake label learning

The proposed method obviously outperforms fake label learning ones based on unsupervised models, such as progressive unsupervised learning (PUL) (Fan et al., 2018), Cross-view Asymmetric Metric Learning (CAMEL) (Yu et al., 2017), and Deep Clustering-based Asymmetric Metric Learning (DECAMEL) (Yu et al., 2020). This happens because these methods assigned directly the fake labels by comparing visual features, and they ignored the potential discriminative information, which resulted in unsatisfactory performance results.

b. unsupervised domain adaptation

When the proposed method was compared to unsupervised domain adaptation-based methods, such as, Person Transfer GAN (PTGAN) (Wei et al., 2018), (TJ-AIDL) (Wang et al., 2018), Similarity Preserving Generative Adversarial Network (SPGAN)(Deng et al., 2018), Multi-task Mid-level Feature Alignment (MMFA) (Lin et al., 2018), CamStyle (Zhong et al., 2019), Hetero-Homogeneous Learning (HHL) (Zhong et al., 2018), Multilabel Reference learning (MAR) (Yun et al., 2019), Neighbor Similarity and Soft-label Adaptation (NSSA) (Zhao & Lu, 2020), the results of these methods also proved to be slightly worse than those obtained by the proposed method. One of the point reasons is that the models of most methods take into consideration only the discriminative feature information in source domain, while ignoring the effective discriminative potential information in unlabeled target domain. In addition, the discriminative feature information in source domain will vary greatly with the change of dataset, so the effectiveness and diversity of its own in target dataset are reduced. Moreover, since the similarity between image patches must be larger than that of images, the result of local branch in the proposed method which, is based on image patch to learn features, is better than the result of the methods which are based on images.

5. Conclusion

Person Re-ID has become an important branch in the field of computer vision and pattern recognition in recent years, and it gives a great boost to a range of applications of intelligent video surveillance, such as cross-camera target tracking and cross-camera behavior analysis. In this paper, an unsupervised learning framework based on global and local features is proposed to solve the re-identification task. Specifically, for local branch, the work focuses on extracting patches from feature map by learning on the PatchNet network of images, and by learning fine-grained patch features at different locations for images within unlabeled datasets. Sometimes it can solve the annoyance caused by occlusion, which demonstrates the effectiveness of local feature learning in unsupervised Re-ID. For global branch, the similarity and diversity of unlabeled datasets are used as information to learn its coarse-grained features. The two loss functions of attraction and repulsion are used to continuously increase intra-classes similarity and inter-classes diversity. The similarity between features is pulled

close through minimum distance criterion which is carried out in unsupervised cluster merging. Finally, a series of experiments were designed to verify the effectiveness of each part in the whole method. Experimental results proved that the proposed method is significantly effective for solving the task that guides the learning of inter-class diversity and intra-class similarity, as well as the discriminatively fused-grained feature learning for unsupervised person re-identification. With the implementation of person re-identification technology, some new problems have emerged, such as: cross-modal learning. In the future, the focus will be on the extraction of cross-modal features.

Acknowledgements

The research reported in this paper was supported in part by the National Nature Science Foundation of China (Nos. 62103257, 61305014), the Natural Science Foundation of Shanghai, China (No. 22ZR1426200), "Chen Guang" project supported, in turn, by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (No. 13CG60).

REFERENCES

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L. & Murino, V. (2011). Custom pictorial structures for re-identification. In *Proceedings of the British Machine Vision Conference (BMVC), Dundee, Scotland* (pp. 1-68).

Deng, W. J., Zheng, L., Ye, Q. X., Kang, G. L., Yang, Y. & Jiao, J. B. (2018). Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA (pp. 994-1003).

Fan, H., Zheng, L. & Yan, C. (2018). Unsupervised person re-identification: Clustering and fine-tuning, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14*(4), 83.

Farenzena, M., Bazzani, L., Perina, A., Murino, V. & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA (pp. 2360-2367).

Gheissari, N., Sebastian, T. B. & Hartley R. (2006). Person reidentification using spatiotemporal appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* New York, USA (pp. 1528-1535).

Gong, S. G., Cristani, M., Yan, S. C. & Loy, C. C. (eds.), (2014). *Person Re-Identification,* 301-313. Springer.

Gray, D. & Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision (ECCV), Marseille, France* (pp. 262-275).

Han, H., Ma, W., Zhou, M., Guo, Q. & Abdullah, A. (2021a). A Novel Semi-supervised Learning Approach to Pedestrian Reidentification, *IEEE Internet of Things Journal, 8*(4), 3042-3052.

Han, H., Zhou, M. C., Shang, X. W., Cao, W. & Abusorrah, A. (2021b). KISS+ for Rapid and Accurate Pedestrian Re-identification, *IEEE Transactions on Intelligent Transportation Systems*, *22*(1), 394-403.

Han, H., Zhou, M. C. & Zhang, Y. J. (2020). Can Virtual Samples Solve Small Sample Size Problem of KISSME in Pedestrian Re-identification of Smart Transportation?, *IEEE Transactions on Intelligent Transportation Systems*, *21*(9), 3766-3776.

Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P. M. & Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, OI, USA (pp. 2288-2295).

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, USA (pp. 1097-1105). Curran Associates Inc.

Li, W., Zhao, R., Xiao, T. & Wang, X. G. (2014). Deep Re-ID: deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA (pp. 152-159).

Li, Y. J., Zhuo, L., Zhang, J., Li, J. F. & Zhang, H. (2018). A survey of person re-identification, *Acta Automatica Sinica, 44*(9), 1554−1568.

Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L. & Smith, J. R. (2013). Learning locally-adaptive decision functions for person verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA (pp. 3610-3617).

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

Lin, S., Li, H. L., Li, C. T. & Kot, A. C. (2018). *Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification.* Available at: <arXiv:1807.01440>, last accessed: May 19, 2021.

Ma, B., Su, Y. & Jurie, F. (2012). Local descriptors encoded by fisher vectors for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Firenze, Italy (pp. 413-422).

Ma, W. J., Han, H., Kong, Y. & Zhang, Y. J. (2020). A New Date-Balanced Method Based on Adaptive Asymmetric and Diversity Regularization in Person Re-identification, *International Journal of Pattern Recognition and Artificial Intelligence*, *34*(9), 2056004.1-20.

Peng, P. X., Xiang, T., Wang, Y., Gong, S. G., Huang, T. J. & Tian, Y. H. (2016). Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA (pp. 1306-1315).

Porikli, F. (2003). Inter-camera color calibration by correlation model function. In *IEEE International Conference on Image Processing (Cat.No.03CH37429)*, (pp. II-133). DOI: 10.1109/ICIP.2003.1246634

Rădulescu, C. & Rădulescu, M. (2020). A Group Decision Approach for Supplier Selection Problem Based on a Multi-criteria Model, *Studies in Informatics and Control, 29*(1), 35-44. DOI: 10.24846/v29i1y202004

Subotic, M., Manastijevic, A. & Kupusinac, A. (2020). Parallelized Multiple Swarm Artificial Bee Colony (PMSABC) Algorithm for Constrained Optimization Problems, *Studies in Informatics and Control, 29*(1), 77-86. DOI: 10.24846/v29i1y202008

Wang, C. H., Han, H., Shang, X. W. & Zhao, X. L. (2020). A New Deep Learning Method Based on Unsupervised Domain Adaptation and Re-ranking in Person Re-identification, *International Journal of Pattern Recognition and Artificial Intelligence, 34*(13), 2052011.1-20.

Wang, J. Y., Zhu, X. T., Gong, S. G. & Li, W. (2018). Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA (pp. 2275-2284).

Wei, L. H., Zhang, S. L., Gao, W. & Tian, Q. (2018). Person transfer GAN to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA (*pp. 7*9-88).

Wu, Z. R., Xiong, Y. J., Yu, S. & Lin, D. H. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA (pp. 3733-3742).

Xiao, T., Li, H. & Ouyang, W. (2016). Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA (pp. 1249-1258). DOI: 10.1109/CVPR.2016.140

Xiao, T., Li, S., Wang, B. C., Lin, L. & Wang, X. G. (2017). Joint detection and identification feature learning for person search, Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA (pp. 3376-3385).

Yang, Q., Yu, H. X., Wu, A. & Zheng, W. S. (2019). Patch-based Discriminative Feature Learning for Unsupervised Person Re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, LA, USA (pp. 3633-3642). DOI: 10.1109/CVPR.2019.00375

Yousuf Uddin, M., Awad Abdeljaber, H. & Ahamed Ahanger, T. (2021). Development of a Hybrid Algorithm for efficient Task Scheduling in Cloud Computing environment using Artificial Intelligence, *International Journal of Computers Communications & Control, 16*(5), 4087.

Yu, H. X., Wu, A. C. & Zheng, W. S. (2017). Cross-view asymmetric metric learning for unsupervised person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy (pp. 994-1002).

Yu, H. X., Wu, A. C. & Zheng, W. S. (2020). Unsupervised person re-identification by deep asymmetric metric embedding, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(4), 956-973.

Yun, H. X., Zheng, W. S., Wu, A., Guo, X, Gong, S. & Lai, J-H. (2019). Unsupervised Person Re-identification by Soft Multilabel Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, LA, USA (pp. 2143-2152).

Zhang, L., Han H., Zhou, M., Al-Turki, Y. & Abusorrah, A. (2021). An Improved Discriminative Model Prediction Approach to Real-time Tracking of Objects with Camera as Sensors, *IEEE Sensors Journal, 21*(15), 17308-17317.

Zhao, Y. & Lu, H. (2020). Neighbor Similarity and Soft-label Adaptation for Unsupervised Cross-dataset Person Re-identification, *Neurocomputing*, *388*, 246-254.

Zheng, L., Shen, L. Y., Tian, L., Wang, S. G., Wang, J. D. & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile (pp. 1116-1124).

Zhong, Z., Zheng, L., Li, S. Z. & Yang, Y. (2018). Generalizing a person retrieval model hetero- and homogeneously. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 11217 LNCS (pp. 176-192).

Zhong, Z., Zheng, L., Zheng, Z. H., Li, S. Z. & Yang, Y. (2019). Camstyle: A novel data augmentation method for person re-identification, *IEEE Transactions on Image Processing*, *28*(3), 1176-1190.