Unsupervised Visual Feature Learning Based on Similarity Guidance

Xiaoqiang Chen[a], Zhihao Jin[a], Qicong Wang[a,*], Wenming Yang[b], Qingmin Liao[b], Hongying Meng[c]

[a]*Department of Computer Science and Technology, Xiamen University, Xiamen, 361000, China*
[b]*Shenzhen International Graduate School/Department of Electronic Engineering, Tsinghua University,*
[c]*Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge UB8 3PH, UK*

Abstract

The availability of a large amount of image data and the impracticality of annotating each sample, coupled with various changes in the target class, such as lighting, posture, etc., make the performance of feature learning disappointing on unlabeled datasets. Lack of attention to hard sample pairs in network modeling and one-sided consideration of similarity measurement in the process of merging have exacerbated the huge performance gap between supervised and unsupervised feature expression. In order to alleviate these problems, we propose an unsupervised network that gradually optimizes feature expression under the guidance of similarity. It employs the deep network to train high-dimensional features and small-scale merge to generate high-quality labels to alternately execute the two steps. Feature learning is guided by gradually generating high-quality labels, thereby narrowing the huge gap between unsupervised learning and supervised learning. The proposed method has been evaluated on both general datasets and the datasets for person re-identification (person re-ID) with superior performance in comparison with existing state-of-the-art methods.

Keywords: Unsupervised learning, similarity measurement, feature generation, image retrieval.

*Corresponding author
Email address: qcwang@xmu.edu.cn (Qicong Wang)

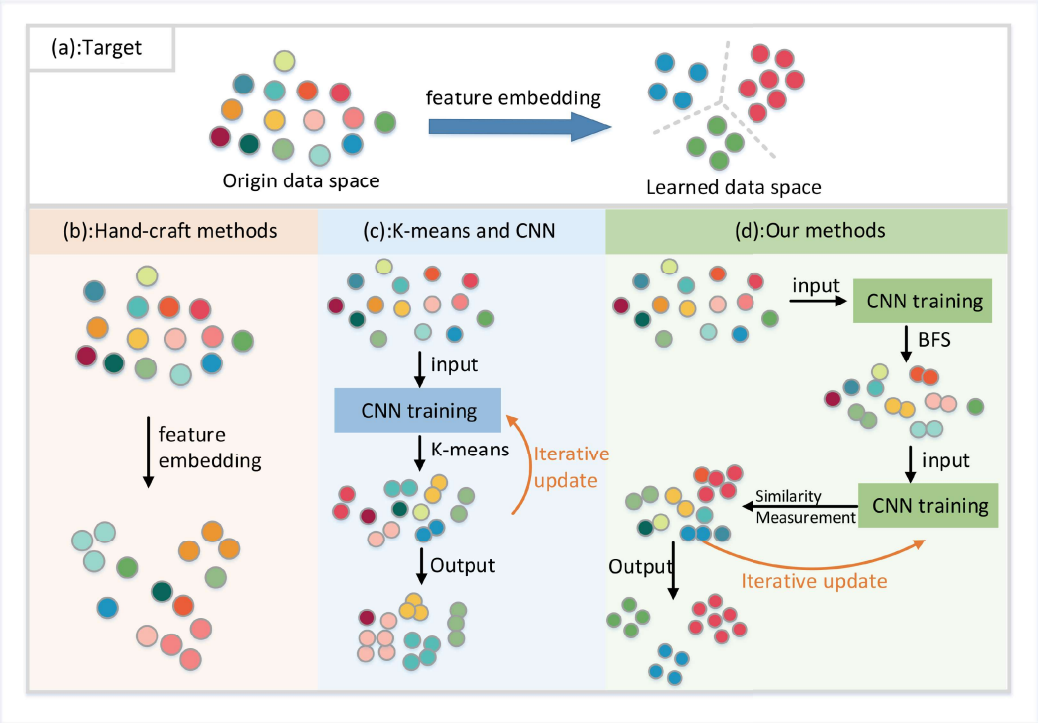

Figure 1: **Principle comparison of methods. First, (a) explains the goal pursued by the model. Furthermore, (b) represents a hand-craft method. (c) is the current method of combining clustering and neural network, which uses the k-means algorithm [1] and CNN training repeatedly to optimize the feature extraction network. Finally, (d) represents our proposed method, which refines and taps the potential of the mergeing process and training process. Through the comparison of these ideas(b,c,d), it is highlighted that our method pays more attention to the sample similarity relationship in the training process.**

1. Introduction

With the widespread application of image classification and image retrieval, accompanied by the rise of large-scale unlabeled image data, how to use unlabeled data-driven methods to learn the feature expression of images, and how to use high-dimensional vectors to represent the discriminative features of images become hot research topics. However, the complexity of

the content of the image, the lack of discernibility of the image with the characteristics of hand-designed, and the guidance of learning without true labels exacerbate the difficulty of the task. Traditional methods are mainly manual features [2, 3] where the feature expression of images relies on prior knowledge. With the development of neural networks, the discrimination and robustness of image features under supervisory signal has been greatly improved, which significantly improves the accuracy of image classification tasks and image retrieval tasks. However, in real life, there are more unlabeled image data, and it is impractical to make precise annotations for each sample. Some previous works combined deep networks and clustering algorithms together to narrow the huge gap in performance between supervised and unsupervised features, such as using k-means clustering tags for reverse training of the network [1, 4] and employing minimum distance as the merge standard to gradually generate pseudo-label to optimize the network [5]. In recent work, the attention mechanism is further mined and applied for more discriminative feature learning [6, 7]. However, most methods did not fully exploit the relations or constraints among samples, which carry crucial information for feature representation.

In order to reduce the gap with supervised methods, we propose a dynamic framework based on multiple similarities to combine samples between different classes and gradually optimize feature representation. On the one hand, in the feature network, the multi-similarly loss is used to guide the learning of features to narrow the sample pairs of the same label and increase the difference under different labels. On the other hand, in the first merging process, we use a clustering algorithm based on limited threshold and breadth-first search to classify similar labels with higher confidence, and generate reliable labels. Then we use the distance between the samples based on different similarity considerations to sort as the basis for the merge of classes.

Specifically, we use the differences between the various images to initialize the network. Due to the same network architecture, images with larger differences during training are more likely to be far away from each other. This provides the basis for our preliminary clustering. Furthermore, using the breadth-first search algorithm, the adjacent samples in the feature space are divided into multiple sets and given different pseudo-labels. We want to aggregate highly similar samples to form multiple high-reliability sets, instead of dividing all samples correctly at once. Although the number of classes will be greatly increased and the original class will be divided into multiple classes, the samples are aggregated by a strong confidence and the generated

pseudo-label has a higher credibility. The Euclidean distance between samples, the consideration of the surrounding environment information where the two samples are located, and the optional sample equilibrium restriction constitute our distance measurement, which is called the individual-group similarity. This distance is used as a criterion for whether to merge classes during the merging process. It takes into account not only the spatial distance between sample pairs, but also the relationship between the sets in the feature space where the samples are located. Based on the assumption that the distribution of various classes in the dataset tends to the mean, the distance measurement should be added according to the specific dataset.

We propose multi-similarly loss to train the feature extraction network. It can narrow the distance between samples of the same class and make samples of different classes far away from each other. As shown in Figure 1, compared to previous works, we divide the network training into multiple stages to facilitate the use of more accurate similarity-based merging measurement in the process of merging classes. The proposed method makes the gradually generated pseudo-label more credible, and it can optimize the entire feature network.

Our contributions are as follows:

1) We propose an unsupervised feature learning network based on context similarity and gradual optimization of feature expression. Our proposed method can generate high-quality pesudo-label, based on which we use iterative merging methods to train the network. The network then can gradually express the similarity distribution contained in various classes in the dataset.

2) In order to effectively use the rich similarity relationship between samples when combining, we propose the individual - group similarity between samples as a measurement method. The proposed distance can consider the similarity relationship between both individual samples and sample groups. In addition, the constraints of sample balance when measuring sample spatial distance are also considered.

3) Since there are dissimilar features between the same class and similar features between different classes, it will cause the interference problem of the similarity measurement of different classes in the training process. Based on the principle of similar attraction, we propose a multi-similarly loss function to alleviate this problem.

2. Related Work

2.1. Unsupervised Feature Learning

Unsupervised feature learning has received extensive research and attention in many tasks, such as image recognition, image classification and image retrieval tasks [8]. The previous works [2, 3] mainly used manual methods to generate features and used them in subsequent tasks. However, the hand-designed features are not normally discriminatory enough because they are limited by people's prior knowledge. Chen et al. [9] integrated extreme learning machine withunsupervised feature selection for clustering. Wang et al. [10] proposed a new network structure for both representation learning and Gaussian Mixture Model-based representation modeling. Jiao et al. [11] combined graph-based clustering and high-level semantic features into an unsupervised segmentation method. Du et al. [12] proposed an unsupervised deep network to map images to hierarchical representations without any label information. Ding et al. [13] proposed to carry out the feature selection process in the learned latent representation space. Dosovitskiy et al. [14] used data augmentation to generate proxy labels to guide the network learning. Bautista et al. [15] used the similarity classification in the sample to ease the huge imbalance between a positive sample and many negative samples and to solve the unreliable relationship between most samples. Wu et al. [16] proposed a non-parametric softmax classifier, and used noise contrast estimation to solve the computational difficulties caused by a large number of samples. Different from these methods, our framework not only considers the diversity of samples, but also makes use of the similarity between them. Compared with these unsupervised feature learning methods, our framework has better performance in image classification and image retrieval tasks.

2.2. Self-supervised Learning

At present, a popular form of unsupervised learning is called "self-supervised learning" [17], which uses a priori operation on the original data to obtain "pseudo-label" to replace artificially annotated labels to guide the network to learn features. For example, some researchers [18, 19] explored how to use spatial environment and video spatiotemporal information as free and rich supervision signals to train rich visual representations, and in [20] the rotation of the image served as a supervised signal to guide the learning of the unsupervised network. Recently, Jing et al. [17] classified and summarized the research works related to self-supervision. It is worth noting that

the current contrastive learning as a form of self-supervision has attracted the attention of many researchers. Oord et al. [21] predicted the data representation in the latent space by using a powerful autoregressive model based on the comparison prediction coding and the proposed infoNCE loss function. In [22] , it is assumed that the effective data identification pair [21] could be further applied and improved by making the variability in the natural signal more predictable. The difference between these methods and our work is that these methods rely on prior knowledge or keen "intuition" and require professional knowledge to carefully design pseudo-label so that they may generate useful features for auxiliary tasks. Our proposed method does not require these addtional conditions and can still perform well on general tasks.

2.3. Data Representation and Similarity Measurement

The success of machine learning algorithms usually depends on the data representation. We assume that this is because different feature representations can represent different hidden explanatory factors behind the data. Although domain knowledge can be used to help design representations, learning can also be used. In recent years, more powerful representation learning algorithms have been put forward for visual tasks [23, 24, 25]. In general, data representation is the core determinant factor on the performance of most machine learning algorithms on a particular application. Furthermore, the storage and availability of large amounts of data remain barriers, and the cost of annotating in human and material resources is also very high. How to express a large amount of original data into robust high-dimensional feature representations without supervision has become an important research topic. In particular, clustering method generating pseudo-label and feature learning training mutually promote each other as a kind of mainstream. The similarity measurement plays a key role in the clustering process. An important factor that affects whether two sample groups can be combined into the same class is the distance measurement between samples. The effect of the combination directly affects the quality of the pseudo-label. Currently, the measurement standards for sample pairs are often unitary, such as direct use of Euclidean distance or Cosine distance to measure. In [5], the minimum distance between the sample pairs and the average distance based on the class level are compared in experiments, and it is concluded that the use of the minimum distance as the standard for the combination measurement has better performance improvement. It uses a single metric on sample pairs to determine its

image retrieval performance. However, we believe that only using the spatial distance between samples for measurement has limitations, and the groups to which the samples belong (the same label or similar samples) have a certain reference value for their similarity measurement. Zhang et al. [26] proposed an entropy-based distance metric that quantifies the distance between categories by exploiting the information provided by different attributes that correlate with the target one. Zhao et al.[27] introduced a distance metric which incorporates inner-domain neighbor similarity. In [28, 29], the sample pairs were measured taking into account the surrounding information, and the original ordering in the image retrieval is rearranged. Fan et al. [30] proposed a dual-level progressive similar instance selection method to build relationship for each instance with its neighbors. In our unsupervised feature learning method, merging pseudo-label is crucial, and the similarity measurement of sample pairs determines the quality of pseudo-label. By considering the spatial information and neighboring information of the sample pair, we propose a measurement of the similarity of the individual-group distance and use the gradual combination to improve the network.

3. Method

We will describe the overall framework and show its algorithm flow in section A. The design of similarity measurement and loss functions will be introduced in sections B and C, respectively. It is worth noting that in Table 1, we define some important notations.

3.1. Overall Framework

In unsupervised learning tasks, we can use only N unlabeled images $X = \{x_1, x_2, x_3, ..., x_N\}$. Our goal is to train a neural network model capable of extracting discriminative features $F = \{f_1, f_2, f_3, ..., f_N\}$ (Resnet-50 [31] is used by default in this paper), that is, $f_i = function(z, x_i)$, where z is the learnable parameter in the model. The key to unsupervised learning is whether it can capture effective low-level appearance information and high-level semantic information while suppressing interference from unknown changes, such as pose, lighting, background, etc. Due to intra-class variation, inter-class similarity and the uncertainty of interference factors, the task is more difficult. We divide the optimization model parameters into three stages, including initial training, re-training with pseudo label and iterative

Table 1: **Notations and Definitions**

Notations	**Definitions**
N	the total number of samples in the dataset
X	the samples
x_i	the i-th sample
F	the sample feature
f_i	the i-th sample feature
L	the pseudo-label
l_i	the label of the i-th sample
υ	the feature extraction network
ξ	the classifier
θ_1 or θ_2	the parameter corresponding to f_i
ζ	the clustering algorithm
eps	the threshold in clustering algorithm
$d_{1,2,3}$ and α, β, γ	different similarity measurements and corresponding weight parameters
D	the individual-group similarity
a, b	the feature representation from different samples
m	the dimension of feature representation
g	the gallery images
κ or ω	the different loss functions
ι	the multi-similarly Loss

merging process. Figure 2 and Algorithm 1 show the proposed method more intuitively.

Initial training. In this stage, we classify each image in the training set into an independent class, so that the network can maximize the difference between each image. Set the training set images as $X = \{x_1, x_2, x_3, ..., x_N\}$, and map them to N classes, namely $L = \{l_1, l_2, l_3, ..., l_N\}$, which corresponds to each image in the image set. This will cause each image in the training process to tend to different classes. In fact, if images with similar features tend to be in the same network, they tend to be close to each other, that is, features generated from similar images are also similar. In this process, we

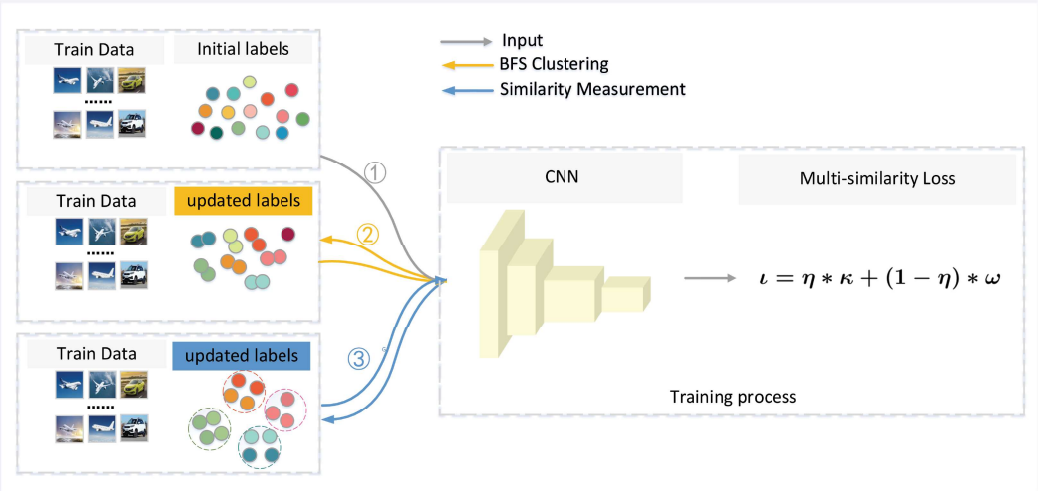

Figure 2: **The training process of the proposed overall architecture. Network update and merging are carried out interactively (Numbers 1,2,3 in the picture), aiming to generate distinguishing features. The detailed multi-stage process is described in section 3. It is worth noting that the yellow arrow represents the clustering algorithm based on breadth-first search.**

need to optimize the conventional neural network:

$$\min_{\theta_1,\theta_2} \frac{1}{N} \sum_{i=1}^{N} Loss(\xi(\theta_2; \upsilon(\theta_1; x_i)), l_i)) \tag{1}$$

where $\xi(\theta_2; \upsilon(\theta_1; x_i))$ is classifier. θ_1 are the parameters of the feature extraction network. θ_2 are the parameters of the classifier. $\upsilon(\theta_1; x_i)$ is the image features. Generally speaking, *Loss* is the cross-entropy softmax loss, but in the case of unsupervised, the lack of real labels causes the classifier to fail to be optimized effectively, so we use the Repelled loss without classifier parameters to train the network. This loss is based on the non-parametric characteristics of non-parametric loss [16] . The multi-stage and multi-similarly loss function specifically proposed in the framework will be described in detail in section C. We use mini-batch stochastic gradient descent and backpropagation to calculate the gradient to minimize this cost function. This stage is called the model initialization training stage, which can be seamlessly migrated to other unlabeled datasets for model pre-training tasks.

Algorithm 1 A class merging framework based on similarity guidance

Require: Unlabeled data $X = \{x_1, x_2, ..., x_N\}$, merge percent $p \in (0, 1)$, CNN model (υ).

Ensure: Best CNN model (υ^*) and accuracy (A^*).

Initialize υ, cluster label (L) and number of cluster (C).
Train υ with L, number of merging images $num = N * p$.
for step 1:1/p **do**
    Calculate distance between sample pairs using individual-group similarity $\rho = D(\upsilon(X))$.
    if $step = 1$ **then**
        Generate new labels L_{new} through clustering algorithm based on breadth-first search.
    else
        Merge classes based on similarity (L_{new}) and current number of classes $C = C - num$.
    end if
    Update label L with L_{new}.
    Initialize the lookup table V with the multi-similarly loss, which temporarily stores the features of each class.
    Re-train υ with L_{new}.
    Evaluate on the validation set and obtain accuracy A.
    if $A > A^*$ **then**
        $A^* = A$.
        $\upsilon^* = \upsilon$.
    end if
end for
return υ^* and A^*.

Re-training with pseudo label. After the initial training of the first stage, we need better labels to guide the training of the feature network in reverse. A subset of the sample set with high similarity in the feature space is required to form the first generation of pseudo-label. We choose a breadth propagation algorithm based on high-confidence sample pairs to perform the combination. A high-confidence threshold can ensure that the generated tags have a high degree of similarity. Although this will result in many more classes than expected, we will alleviate this problem in the third phase of the iterative algorithm. Choosing a clustering algorithm based on

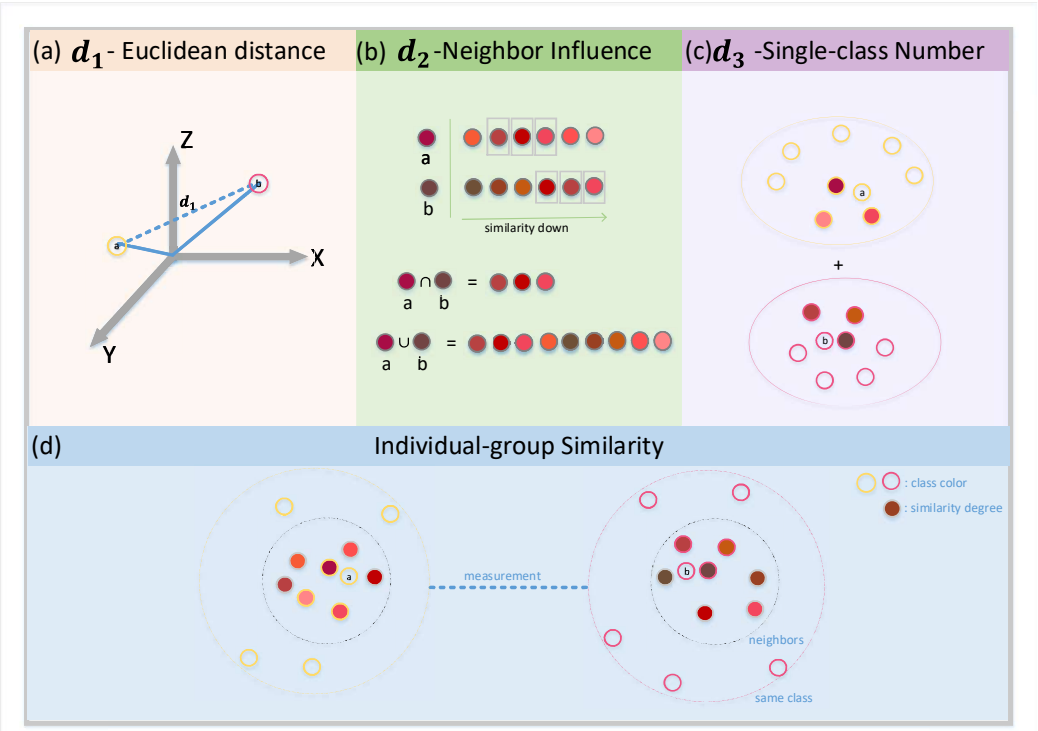

Figure 3: **The proposed individual-collective similarity is in turn determined by the spatial distance (a), the influence of neighbors (b), and the number constraints of the same kind (c) in the picture. It is worth noting that in the Figure (b), the ratio of the intersection and union of the neighbors of a and b is used to determine the degree of influence of their neighbors.**

breadth-first search can take into account the similarity between each pair of samples, so as to form a high-confidence set of these highly similar samples, and mark them as the same class:

$$L = \zeta(eps, \upsilon(\theta_1; X)) \qquad eps \in \{0.1, 0.2, 0.3\} \tag{2}$$

The *eps* is selected from 0.1 to 0.3 according to the actual dataset, that is, the similarity distance between the sample pairs is smaller than the *eps* to have the same pseudo-label. In this way, the high-confidence sample pairs can be merged into one class, which provides a good basic environment for the following iterative algorithms. For ζ, we use the DBSCAN [32] algorithm as the unification algorithm for our breadth-first search. We classify it as a

small-scale cluster algorithm and call this stage pseudo-label initialization.

Iterative merging process. When we have the preliminary pseudo-label, we will train the network again. The network will further improve the feature network due to these label information, so as to obtain more distinctive features. When the preliminary potential of pseudo-label has been tapped, we will carry out an iterative merging process. Every time classes are merged, it will be accompanied by a network update. The purpose is to fully explore the supervisory role of pseudo-label, and further enable the similarities of the same kind to be reflected and the differences of different kinds to be expanded. In the merge process, we select the sample pairs of different classes with the smallest joint distance to merge. This joint distance is called individual-group distance and will be described in detail in section B. It is worth mentioning that not every merger is beneficial to the update of the network. The main reason is the influence of the wrong merger behavior in the merge process. The difference of the loss function at each stage will be described in detail in section C.

3.2. Individual-group Similarity

The similarity measurement of sample pairs is very important in the unsupervised feature extraction method of deep learning. It affects the effect of clustering, thereby further affecting the quality of pseudo-label, and ultimately affecting the update trend of the feature network. We consider that only using the Euclidean distance between two samples to measure the similarity between samples has limitation, because it only considers the spatial information between the sample pairs, and ignores the surrounding information. On the basis of Euclidean distance, we additionally adopt the nearest neighbor algorithm and the idea that each sample's neighbors should include each other. And inspired by the local to global idea, we propose a joint distance that needs to consider the feature of the sample itself and multiple surrounding information. Our default setting is (20, 10, 5). In other words, each layer has higher and higher requirements for data similarity. We optionally add sample balance control to make the total number of samples between each class relatively balanced.

$$d_1 = (\sum_{i=1}^{m}(a_i - b_i)^p)^{\frac{1}{p}} \tag{3}$$

where $p = 2$, $a = (a_1, a_2, ..., a_m)$ and $b = (b_1, b_2, ..., b_m)$ represents a single sample feature, and m represents the dimension of the feature. d_1 can be

any conventional metric based on two sample features, here we use Euclidean distance.

$$d_2 = 1 - |\frac{\psi(x,k) \cap \psi(g,k)}{\psi(x,k) \cup \psi(g,k)}| \tag{4}$$

where $\psi(x,k)$ is the number of the sample and its neighbors each containing each other under the premise of knn algorithm. Among them, we expand the number of neighbor samples in order to weaken the influence of the artificial factors of k. The main idea is to compare the similarity between the neighbor samples (or multiple neighbors) of the sample and the neighbor samples of the sample.

On the other hand, considering the efficiency of the algorithm and the weight of the surrounding samples, it is simplified as:

$$d_2 = 1 - \frac{\sum_{j=1}^{N} min(\varphi(x,g_j), \varphi(g_i,g_j))}{\sum_{j=1}^{N} max(\varphi(x,g_j), \varphi(g_i,g_j)} \tag{5}$$

where

$$\varphi(x,g_i) = \begin{cases} e^{-d_1(X,g_i)} & g_i \in \psi(x,k) \\ 0 & otherwise \end{cases} \tag{6}$$

the distance is mainly a measurement of the similarity of the surrounding conditions of the sample.

$$d_3(a,b) = |Q| + |T| \tag{7}$$

where $|Q|, |T|$ are the number of classes of the a and b samples respectively. The distance is mainly to control the number of samples of the same kind to be in equilibrium.

$$D = \alpha * d_1 + \beta * d_2 + \gamma * d_3 \tag{8}$$

where α, β, γ are the hyperparameters, we set them to 0.5, 0.5, 0.03. Among them, d_3 has limitations on the dataset, and we selectively add it in specific experiments. Figure 3 visualizes the principles of each similarity.

3.3. Multi-similarly Loss

In the first stage, since each sample is treated as a separate class, there is no sample pair of the same kind. We only use the formula to update the network:

$$\kappa = \sum_{i=1}^{N} -log(\frac{e^{V_{ic}^T f_i/r}}{\sum_{j=1}^{C} e^{V_j^T f_i/r}}) \tag{9}$$

where V_{ic} represents the class center to which the i-th sample belongs. C represents the number of classes and r is the hyperparameter.

When using this loss to update the network, it is emphasized that the features under similar tags can be further narrowed. Although this can strengthen the similarity of the same class, hard samples farther from the class center between the same class require additional training to narrow the distance and some between the different classes need to be far away, which is not clearly reflected. Therefore, we use triplet loss function to alleviate this problem. So that the similarity between dissimilar species is suppressed while the similarity of hard samples in the same class can be fully expressed.

$$\begin{aligned} \omega = \sum_{i=1}^{N} max(D(f(x_i^z), f(x_i^p)) \\ -D(f(x_i^z), f(x_i^n)) + margin, 0) \end{aligned} \tag{10}$$

The $margin$ is the hyperparameter of the difference between the distance from the sample x_i^z to the positive sample x_i^p and the distance from the sample to the negative sample x_i^n. In the other words, the $margin$ is the minimum distance between positive and negative pairs in the same sample.

The final loss function is:

$$\iota = \eta * \kappa + (1 - \eta) * \omega \tag{11}$$

where η is a hyperparameter, which determines the proportion of each part's loss. ι considers that the features of the samples of the same type are close to the average features of the class, while taking into account the training between pairs of samples with obvious differences in the same classes and those with smaller differences in the different classes.

4. Experimental Results

In this section, the datasets used are introduced firstly, including the general datasets and the datasets for person re-identification (person re-ID). Furthermore, the evaluation criteria of metrics are introduced, including the criteria for finding one-to-one correspondence between true labels and pseudo-label based on the Hungarian algorithm in unsupervised clustering [33] and the commonly used performance evaluation mAP and Rank-k ($k = 1, 5, 10$) [5]. Finally, we introduce the experimental implementation details and specific experiments including parameter comparison experiments,

ablation experiments, and experiments comparing with the most advanced methods.

4.1. Dataset

4.1.1. General Dataset

STL10 [34]. An ImageNet-adjusted dataset for developing unsupervised feature learning, deep learning, and self-learning learning algorithms, which contains 500/800 train/test samples from 10 classes and 500/800 train/test samples for each class, as well as auxiliary unknown classes of 100,000 unlabeled images. A dataset of 96×96 color images. That is, there are a total of 10 classes, and each class has 1300 examples.

CIFAR10/100 [35]. A natural image dataset containing $50,000/10,000$ train/test images from 10 (/100) object classes.

MNIST [36]. A handwritten digit dataset containing 60,000/10,000 train/test images of 10 digit classes. The MNIST dataset consists of 70,000 handwritten digits and the size is 28×28 pixels. The numbers are centered and the dimensions are standardized.

4.1.2. Person Re-identification Dataset

Market [37]. A large-scale dataset of pedestrian pictures captured by 6 cameras on a university campus. Each pedestrian is captured by at least 2 cameras, and there may be multiple images in one camera. It contains 12,936 images from 751 identities for training, 3368 pictures of pedestrians from 750 identities for query and 19,732 images for testing.

Duke [38, 39]. A large-scale person re-identification dataset derived from the DukeMTMC dataset. It contains 16,522 images of 702 identities used for training, 2,228 images of other 702 identities used for query, and 17,661 gallery images.

4.2. Metric

4.2.1. Unsupervised Standard Metric

Evaluation is based on accuracy, that is, the correct number of samples divided by the total number of samples. For the true label and pseudo-label correspondence problem, we use the k-means method based on the number of classes equal to the true label to divide the learned features and find the best one-to-one permutation mapping according to the standard evaluation scheme. This is not to use true labels for learning, but to correspond to pseudo-label. Thus, the true label situation corresponding to the final

pseudo-label is obtained, and then the accuracy [40] is obtained. We use both the training set and the test set for model learning, which is the same as the experimental setting of Huang et al. [33] .

4.2.2. Person Re-ID Metric

For image retrieval tasks, the images are divided into train set, test set and gallery set. The test set is used to find and sort the same person under different cameras in the gallery, and get its mAP and Rank-1, Rank-5, Rank-10 measurements.

4.3. Experimental Details

In this paper, the Resnet-50 network structure is selected. The parameters trained based on the imagenet dataset are used as pre-training parameters. As the input data of the network, the image width and height are set to 128 and 256 respectively. The training batches are 4 iterations except for the 20 iterations in the first round, because the first round of training changes to network parameters fluctuates greatly. We call the completion of the first round of training as model initialization. The default batch parameter is 64. $instances = 4$ means that 4 samples of each class will be randomly selected for training during the training process. If the number of samples is less than 4, they will be randomly copied from the samples of this class. Merge $percent = 0.05$ controls the number of classes merged each time. For example, there are currently 1000 classes, the next time the number of classes will be reduced by $1000 * 0.05 = 50$ classes. It is worth noting that in the training process of the STL10 dataset, the imagenet dataset pre-training model is not used, instead the CIFAR10 dataset is used for pre-training. Because the STL10 dataset is a subset of the ImageNet dataset.

4.4. Parameter Comparison Experiment

From the data in Table 2, it can be concluded that the best hyperparameter scheme is $proportion = 0.1$, $margin = 0.2$, and $eps = 0.1$. Among them, mAP is the main criterion and Rank-1 is the secondary criterion. In order to more intuitively express the influence of hyperparameters on the model during the training process, we show the specific changes of mAP and Rank-1 in different stages of different values of *eps* in Figures 4 and 5. Generally speaking, within a certain range, the change of hyperparameters has little effect on the model, that is, the difference between the main evaluation indicators mAP and Rank-1 in the comparative experiment is about 1%. It

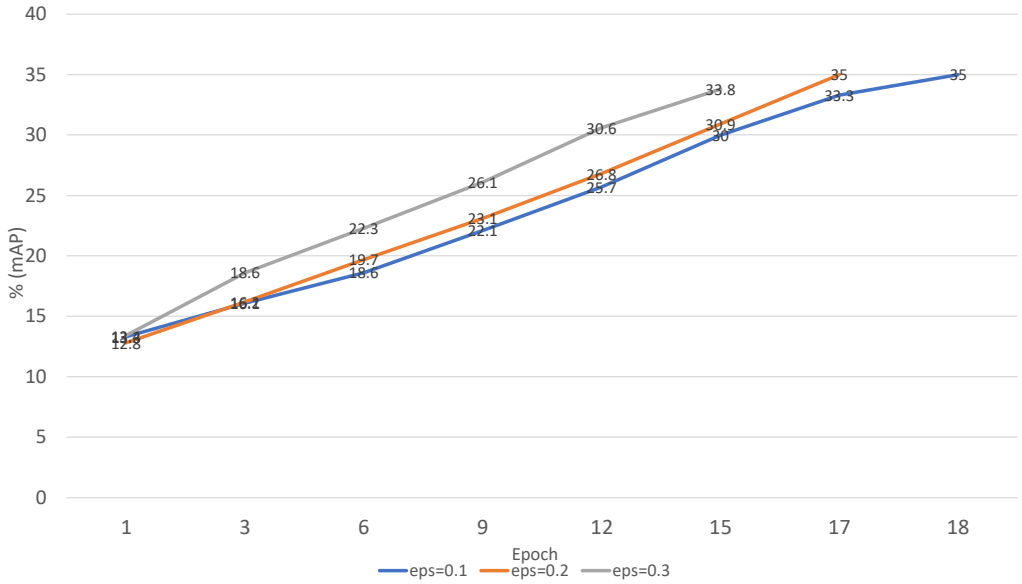

Figure 4: **The mAP comparison of different** *eps* **parameters on the Market dataset**

shows that our model is less dependent on hyperparameters. Details of each hyperparameter are analyzed as follows.

4.4.1. Factors Affecting Joint Losses

On the one hand, the *margin* adjustment determines the distance between the sample with the farthest distance from the same class and the sample with the closest distance to a different class. From the table, we can determine that $margin = 0.2$ is the most ideal in the Market dataset, but it is not much different from a margin of 0.3 and does not have an advantage in the $Rank - 1$ criterion, which indicates that *margin* changes in a small range have little effect on the model. On the other hand, it can be observed that the adjustment of *proportion* determines the influence of the proportion of the triplet loss function in the joint function on the result, and that the effect is best when $proportion = 0.1$.

4.4.2. Determination of The Value in The First Cluster

In the selection process of the *eps*'s value, we tend to choose a smaller value. But it cannot be close to 0, which will cause all samples to remain

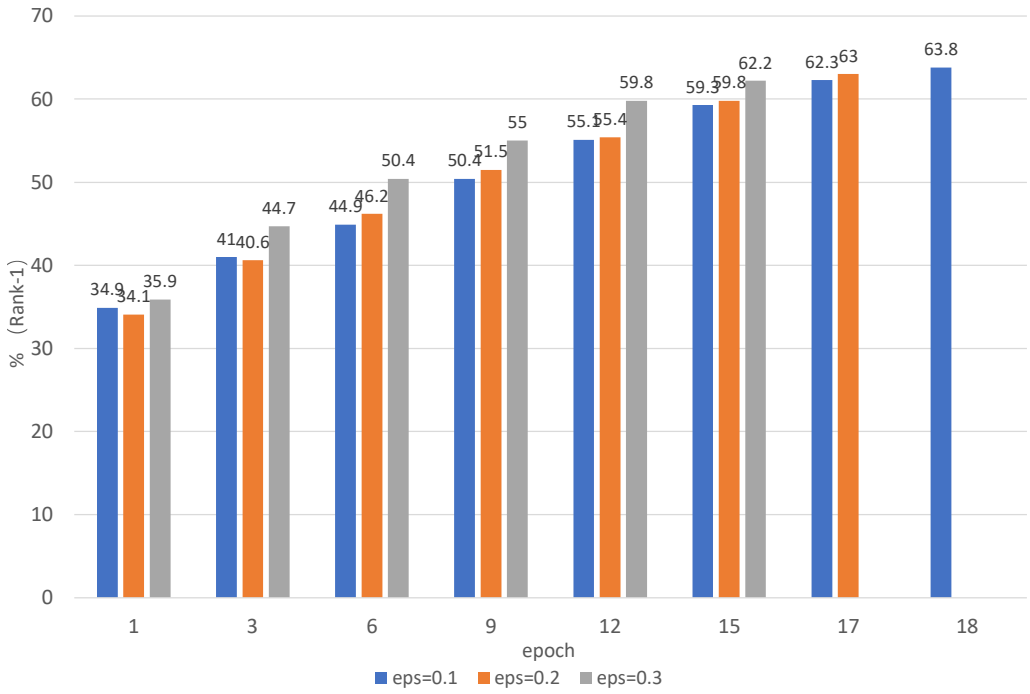

Figure 5: **The Rank-1 comparison of different** *eps* **parameters on the Market dataset**

in one class alone or only a few samples are allocated together. So it can be found from the table that 0.1 is the best choice. In Figures 4 and 5, we can find that as *eps* increases, the number of iterations will decrease. This is because in the first clustering, there are too many samples and labels are assigned in advance which will increase the probability of subsequent merge errors. The gradual decline in the accuracy of the model can also prove the correctness of our views.

4.5. Ablation Experiment

We first conduct several experiments to analyze the effects of using joint loss on the model. From the comparison between conventional model and that uses joint loss in Table 3, we can get through the evaluation criteria that the use of joint loss can make the features generated by the model more discriminative. It shows that only using a single strategy to map images of the same label to the feature space is limited. We believe that the feature difference between images may be relatively large even if they are under the

Table 2: **The influence of the change of main parameters on the model.**

parameter	mAP	Rank-1	Rank-5	Rank-10
proportion=0.1	35.0	63.8	76.9	82.1
proportion=0.2	34.4	63.2	77.4	82.7
proportion=0.3	33.9	62.4	76.3	82.4
margin=0.1	34.3	63.0	76.7	82.2
margin=0.2	35.0	63.8	76.9	82.1
margin=0.3	34.5	63.9	77.7	83.3
eps=0.1	35.0	63.8	76.9	82.1
eps=0.2	35.0	63.0	77.0	83.2
eps=0.3	33.8	62.2	76.4	81.7

same label, and the difference between the features under different labels may not be far away as expected either. Therefore, our improvement to the loss function can effectively encourage the principle that samples of the same class are close to each other and samples of different classes are far away from each other during the training process. Figure 6 intuitively shows the effect of our proposed method through ablation experiment.

We also study whether the use of additional clustering algorithms in the first stage has effects on the performance of the model. The clustering method is used to assign pseudo-label to all samples at once, that is, to rigidly assign pseudo-label to samples without training. It will make some pseudo-label immeasurable and inferior, resulting in subsequent feature learning networks based on pseudo-label supervision signals not able to be optimized in the expected direction. For example, the k-means algorithm is directly used to generate pseudo-label and then guide the network, trying to use all pseudo-label generated at once to complete the optimization of the network, while ignoring the sustainable learning nature of the network [1]. This paper does not seek to assign high-quality pseudo-label to all samples at once, but uses sample pairs based on a high-confidence similarity measurement as the initial merging standard, and uses local high-quality pseudo-label to train the network first before the merge process. It can gradually increase the quality of pseudo-label, gradually optimize the network, and further highlight the similarity relationship between samples. In the first stage, the clustering and division of samples with high similarity can be used to generate a small

Table 3: **Ablation experiments and Comparison results on Market. "basic" means only using the first stage of pre-training. "conventional" represents the conventional merging and training. "+loss" represents adding the loss function we proposed on the basis of "basic". "+clustering" represents adding a clustering algorithm based on "+loss". "+distance" represents replacing the euclidean distance in "+clustering" with the distance function we proposed.**

model	mAP	Rank-1	Rank-5	Rank-10
basic	13.3	34.9	52.8	60.9
conventional	28.0	59.8	70.8	76.0
+loss	32.0	62.5	75.5	80.6
+clustering	33.4	62.8	76.0	81.2
+distance(ours)	35.0	63.8	76.9	82.1

number of high-quality labels in the first round of merging, which provides a basis for subsequent merging. From the comparison between model with only loss function and that with both loss function and the traditional clustering algorithm, it can be clearly seen that there is considerable improvement in the quality of the pseudo-label, and the follow-up training maintains a good trend, which confirms our perspective. This reflects that the quality of the pseudo-label at the initial stage is crucial for the subsequent training of the merge and feature network. It also shows that using the control of the thresh-old value can classify the samples with high similarity into one class and train the model, which can improve the performance to a certain extent. But it will cause the problem of too many classes, and we will alleviate this problem in subsequent iterations.

We test the effect that the adoption of individual-group has on the model. In Table 3, we can intuitively see the difference between the model using clustering based on Euclidean distance and the model using joint distance, which highlights the positive impact on the overall quality of pseudo-label generation. It shows that it is defective to only use the Euclidean distance between samples as the standard of union. This is because the Euclidean distance can only consider the spatial distance between the two samples, and ignore the different label groups in which the two samples are located. It did not consider the difference of the two label groups or the surrounding environment of their samples and fail to take the similarity of their "class" level into consideration. Therefore, we added the influence based on the

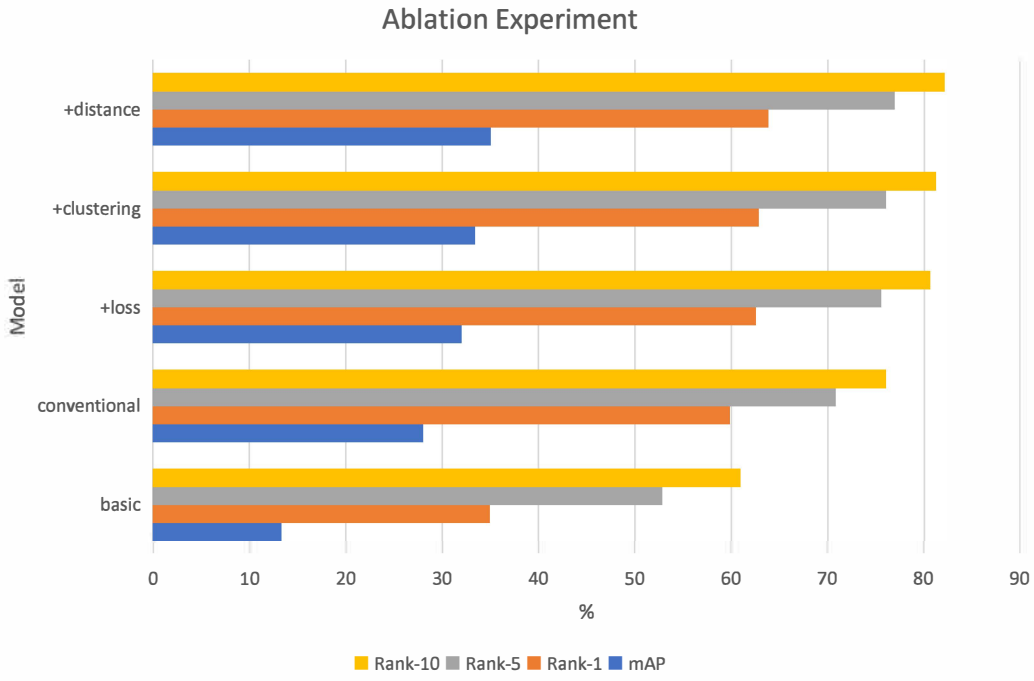

Figure 6: **Ablation experiments and Comparison results on Market.**

surrounding environment of the sample, and added the control of sample balance to form the joint distance. It is worth noting that the sample balance idea performs well in the person re-identification datasetes, but it needs to be adjusted in other general datasets.

Finally, we try to find out how each similarity affects the individual-group distance. In Table 4, we conduct ablation experiments on the proposed similarity distance from single similarity to joint similarity. From experimenting with only Euclidean distance, to adding surrounding information or adding dataset constraints, the overall return is positive. At the end, the individual-group distance we proposed reaches the best state. It is worth mentioning that in the comparison between adding surrounding information and adding dataset constraints, although we can see that the effect of adding dataset constraints is more prominent, the factor that we cannot ignore is that the ablation experiment was carried out on the person re-identification dataset of Market, and the distribution of the dataset itself tends to be balanced, so the effect will be more advantageous.

Table 4: **Ablation experiments and Comparison results about the Individual-group Similarity. "d_1": Euclidean distance. "d_2": surrounding information. "d_3": dataset constraints.**

model	mAP	Rank-1	Rank-5	Rank-10
d_1	30.1	60.3	74.7	79.9
$d_1 + d_2$	32.8	62.0	76.2	81.4
$d_1 + d_3$	33.4	62.8	76.0	81.2
$d_1 + d_2 + d_3$	35.0	63.8	76.9	82.1

Table 5: **Comparision results on image clustering of unsupervised learning methods. "+": Used k-means.**

Methods	MNIST	STL10	CIFAR10	CIFAR100
JULE[41]	96.4	27.7	27.2	13.7
DEC[40]	84.3	35.9	30.1	18.5
DAC[42]	97.8	47.0	52.2	23.8
ADC[43]	**99.2**	53.0	32.5	16.0
IIC[44]	98.4	**59.8**	57.6	25.5
Random CNN+	48.1	20.1	18.6	10.3
Triplets+[45]	52.5	24.4	20.5	9.9
AE+[46]	81.2	30.3	31.4	16.5
Sparse AE+[47]	82.7	32.0	29.7	15.7
Denoising AE+ [48]	83.2	30.2	29.7	15.1
Variational Bayes AE+[49]	83.2	28.2	29.1	15.2
SWWAE+[50]	82.5	27.0	28.4	14.7
DCGAN+[51]	82.8	29.8	31.5	15.1
DeepCluster+[4]	65.6	33.4	37.4	18.9
PAD[33]	98.2	46.5	62.6	28.8
Ours	95.0	58.7	**63.5**	**39.2**

4.6. Comparisons with State-of-the-art Methods

4.6.1. General Dataset

In Table 5, we compare the proposed method with two different types of methods. The first representation method is based on clustering. The second

Table 6: **Comparision with the latest method on 2 datasets based on person re-identification, i.e., the Market dataset and the Duke dataset. The "Label" list indicates whether to use labels and their format. "Transfer" means using the information of another dataset with full annotations. "OneEx" represents an example annotation, where everyone in the dataset is annotated with a labeled example. "*" means that the results are reproduced by us.**

Methods	Label	Year	Market-1501				DukeMTMC-reID			
			mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
BOW [52]	None	2015	14.8	35.8	52.4	60.3	8.3	17.1	28.8	34.9
OIM* [53]	None	2018	14.0	38.0	58.0	66.3	11.3	24.5	38.8	46.0
UMDL [54]	Transfer	2016	12.4	34.5	52.6	59.6	7.3	18.5	31.4	37.6
PUL[1]	Transfer	2018	20.1	44.7	59.1	65.6	16.4	30.4	46.4	50.7
EUG*[55]	OneEx	2018	22.5	49.8	66.4	72.7	24.5	45.2	59.2	63.4
SPGAN[56]	Transfer	2018	26.7	58.1	76.0	**82.7**	26.4	46.9	62.6	68.5
TJ-AIDL[57]	Transfer	2018	26.5	58.2	-	-	23.0	44.3	-	-
UPRSSL[58](w/o part and CCE)	None	2020	29.8	58.7	70.4	76.3	17.4	31.6	48.3	53.4
BUC*[5]	None	2019	30.0	61.7	73.1	77.7	22.1	40.4	52.5	58.2
ours	None	2021	**35.0**	**63.8**	**76.9**	82.1	**31.8**	**53.2**	**65.2**	**70.1**

method focuses on general expression learning. It is worth noting that for the second method, we use the k-means algorithm for clustering to facilitate comparison under the same standard. The experimental data comes from IIC [44]. Although the focus of the method is different, the goal of generating more discriminative feature representations under unsupervised conditions is the same. Some analysis results are drawn from the observation of the above two methods and our proposed method:

On the one hand, the first method tends to generate better clustering results. This is because they use the known number of real classes to jointly learn feature representation and clustering during end-to-end model training, that is, consistency between training and testing goals. Among them, IIC achieved the best results. On the other hand, without clustering as the target, the second group of methods are relatively poor in modeling the structure of the data group. Among them, PAD uses the affinity between the sample pairs to combine to achieve the best performance in this group.

Finally, our model is still very competitive with all the advanced methods mentioned above, and achieved the best results on CIFAR10/100. We believe that the main reason is that the proposed model considers the relationship between the data in the process of focusing on the small-scale combined sample and the neural network feedback, which makes the generated features more unique. Furthermore, our model adopts progressive combination for end-to-end learning without the known number of true classes.

4.6.2. Person Re-identification Dataset

Table 6 shows the comparison between our method and the most advanced methods on the image retrieval dataset. We use the person re-identification task based on the Market and Duke datasets as our experimental evaluation criteria for image retrieval. And in these two datasets our proposed method reaches current optimal level. For BOW, OIM* and other experimental data are derived from BUC*. On the Market dataset, our method reaches mAP=35.0%, Rank-1=63.8%. It is better than the most advanced unsupervised methods by 5% and 2% respectively, such as BUC* and UPRSSL. Similarly, improvements of 9% and 12% were obtained on the duke dataset. It is worth mentioning that for UPRSSL, we chose a model without the prior knowledge of camera id and block. Because we don't use these techniques to optimize.

We also compare our experimental results with the last two popular methods. The first is to use cross-domain learning methods between different datasets under the same task conditions, such as UMDL, PUL, SPGAN and TJ-AIDL. Furthermore, we implement experiments on one-shot methods which use a very small number of labeled samples and a large number of unlabeled samples, such as EUG*. The numerical comparison between mAP and Rank-k proves that the proposed model has greater advantages than these methods. At the same time, it highlights that our method can achieve better feature representation without supervision.

5. Conclusion

In this paper, we propose an unsupervised visual feature network based on similarity guidance, aiming to solve the problem of image feature generation under the unsupervised manner. The essence is to use a variety of similarities to enrich the similarity measurement between samples, and then use multi-stage network training to jointly optimize the pseudo-label quality and the neural network. It achieves state of the art on two person re-identification datasets, and competitive performance on four general classification datasets.

The focus of future work lies in three aspects. Firstly, we will focus on exploring and using more similarity measurement for feature learning of high-dimensional features. Secondly, how to apply our method to massive data is also one of our future research directions. Thirdly, we consider to apply the idea of adversarial learning in self-supervision manner to our proposed framework to obtain a better pre-training model in the early stage.

Acknowledgement

This work was supported in part by Shenzhen Science and Technology Projects under Grant JCYJ20200109143035495.

References

[1] H. Fan, L. Zheng, C. Yan, Y. Yang, Unsupervised person re-identification: Clustering and fine-tuning, ACM Transactions on Multimedia Computing, Communications, and Applications 14 (4) (2018) 1–18.

[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1, IEEE, 2005, pp. 886–893.

[3] T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, in: Proceedings of the European Conference on Computer Vision, Springer, 2004, pp. 469–481.

[4] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep clustering for unsupervised learning of visual features, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 132–149.

[5] Y. Lin, X. Dong, L. Zheng, Y. Yan, Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8738–8745.

[6] Huang, Yangru and Peng, Peixi and Jin, Yi and Li, Yidong and Xing, Junliang, Domain adaptive attention learning for unsupervised person re-identification, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11069–11076.

[7] Z. Ji, X. Zou, X. Lin, X. Liu, T. Huang, S. Wu, An attention-driven two-stage clustering method for unsupervised person re-identification, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16, Springer, 2020, pp. 20–36.

[8] B. Peng, J. Lei, H. Fu, C. Zhang, T.-S. Chua, X. Li, Unsupervised video action clustering via motion-scene interaction constraint, IEEE Transactions on Circuits and Systems for Video Technology 30 (1) (2018) 131-144.

[9] J. Chen, Y. Zeng, Y. Li, G.-B. Huang, Unsupervised feature selection based extreme learning machine for clustering, Neurocomputing 386 (2020) 198–207.

[10] J. Wang, J. Jiang, Unsupervised deep clustering via adaptive gmm modeling and optimization, Neurocomputing 433 (2021) 199–211.

[11] X. Jiao, Y. Chen, R. Dong, An unsupervised image segmentation method combining graph clustering and high-level feature representation, Neurocomputing 409 (2020) 83–92.

[12] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, D. Tao, Stacked convolutional denoising auto-encoders for feature representation, IEEE Transactions on Cybernetics 47 (4) (2016) 1017–1027.

[13] D. Ding, X. Yang, F. Xia, T. Ma, H. Liu, C. Tang, Unsupervised feature selection via adaptive hypergraph regularized latent representation learning, Neurocomputing 378 (2020) 79–97.

[14] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, T. Brox, Discriminative unsupervised feature learning with convolutional neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 766–774.

[15] M. A. Bautista, A. Sanakoyeu, E. Tikhoncheva, B. Ommer, Cliquecnn: Deep unsupervised exemplar learning, in: Advances in Neural Information Processing Systems, 2016, pp. 3846–3854.

[16] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.

[17] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1–1doi:10.1109/TPAMI.2020.2992393.

[18] C. Doersch, A. Gupta, A. A. Efros, Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1422–1430.

[19] D. Luo, C. Liu, Y. Zhou, D. Yang, C. Ma, Q. Ye, W. Wang, Video cloze procedure for self-supervised spatio-temporal learning, arXiv preprint arXiv:2001.00294 (2020).

[20] S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, arXiv preprint arXiv:1803.07728 (2018).

[21] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[22] O. J. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. Es-lami, A. v. d. Oord, Data-efficient image recognition with contrastive predictive coding, arXiv preprint arXiv:1905.09272 (2019).

[23] W. Zhang, Q. J. Wu, Y. Yang, T. Akilan, H. Zhang, A width-growth model with subnetwork nodes and refinement structure for representation learning and image classification, IEEE Transactions on Industrial Informatics 17 (3) (2020) 1562–1572.

[24] W. Zhang, Q. J. Wu, Y. Yang, T. Akilan, M. Li, Hkpm: A hierarchical key-area perception model for hfswr maritime surveillance, IEEE Transactions on Geoscience and Remote Sensing (2021).

[25] W. Zhang, Q. J. Wu, Y. Yang, T. Akilan, Multimodel feature reinforcement framework using moore-penrose inverse for big data analysis, IEEE Transactions on Neural Networks and Learning Systems (2020).

[26] Y. Zhang, Y.-M. Cheung, A new distance metric exploiting heterogeneous interattribute relationship for ordinal-and-nominal-attribute

data clustering, IEEE Transactions on Cybernetics (2020) 1–14doi:10.1109/TCYB.2020.2983073.

[27] Y. Zhao, H. Lu, Neighbor similarity and soft-label adaptation for unsupervised cross-dataset person re-identification, Neurocomputing 388 (2020) 246–254.

[28] M. Saquib Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, 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, 2018, pp. 420–429.

[29] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.

[30] H. Fan, P. Liu, M. Xu, Y. Yang, Unsupervised visual representation learning via dual-level progressive similar instance selection, IEEE Transactions on Cybernetics (2021) 1–11doi:10.1109/TCYB.2021.3054978.

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[32] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, Vol. 96, 1996, pp. 226–231.

[33] J. Huang, Q. Dong, S. Gong, X. Zhu, Unsupervised deep learning via affinity diffusion., in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 11029–11036.

[34] A. Coates, A. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 215–223.

[35] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).

[36] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[37] L. Zheng, L. Shen, L. Tian, S. Wang, J. Bu, Q. Tian, Person re-identification meets image search, arXiv preprint arXiv:1502.02171 (2015).

[38] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 3754–3762.

[39] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 17–35.

[40] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: International Conference on Machine Learning, 2016, pp. 478–487.

[41] J. Yang, D. Parikh, D. Batra, Joint unsupervised learning of deep representations and image clusters, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5147–5156.

[42] J. Chang, L. Wang, G. Meng, S. Xiang, C. Pan, Deep adaptive image clustering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5879–5887.

[43] P. Haeusser, J. Plapp, V. Golkov, E. Aljalbout, D. Cremers, Associative deep clustering: Training a classification network with no labels, in: German Conference on Pattern Recognition, Springer, 2018, pp. 18–32.

[44] X. Ji, J. F. Henriques, A. Vedaldi, Invariant information clustering for unsupervised image classification and segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9865–9874.

[45] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, Advances in Neural Information Processing Systems 16 (2003) 41–48.

[46] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, Advances in Neural Information Processing Systems 19 (2006) 153–160.

[47] A. Ng, et al., Sparse autoencoder, CS294A Lecture Notes 72 (2011) (2011) 1–19.

[48] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, L. Bottou, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion., Journal of Machine Learning Research 11 (12) (2010).

[49] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[50] J. Zhao, M. Mathieu, R. Goroshin, Y. Lecun, Stacked what-where auto-encoders, arXiv preprint arXiv:1506.02351 (2015).

[51] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434 (2015).

[52] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[53] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3415–3424.

[54] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, Y. Tian, Unsupervised cross-dataset transfer learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1306–1315.

[55] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5177–5186.

[56] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, J. Jiao, 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, 2018, pp. 994–1003.

[57] J. Wang, X. Zhu, S. Gong, W. Li, Transferable joint attribute-identity deep learning for unsupervised person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2275–2284.

[58] Y. Lin, L. Xie, Y. Wu, C. Yan, Q. Tian, Unsupervised person re-identification via softened similarity learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3390–3399.