Reconstruction Weighting Principal Component Analysis with Fusion Contrastive Learning

Qianqian Wang[12], **Meiling Liu**[1*], **Wei Feng**[3], **Mengping Jiang**[1], **Haiming Xu**[1] and **Quanxue Gao**[1]

[1]School of Telecommunications Engineering, Xidian University, Xi'an, China
[2]Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing, China, 210096.
[3]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
qqwang@xidian.edu.cn, lml1710031211@163.com, weifeng.ft@xjtu.edu.cn, 2047788244@qq.com, kiritobryant8@gmail.com, qxgao@xidian.edu.cn *

Abstract

Principal component analysis (PCA) is a popular unsupervised dimensionality reduction method to extract the principal components of data. However, there are two problems with the existing PCA: (1) Traditional PCA methods treat each sample equally and ignore sample differences. (2) They fail to extract the discriminative features required by recognition tasks. To overcome these problems, we incorporate contrastive learning to develop a novel weighted PCA algorithm. Specifically, our method weights the reconstruction error of individual samples to reduce the influence of outliers. Besides, it integrates contrastive learning into PCA to increase inter-class distances and reduce intra-class distance, which helps to improve PCA's discriminative capability. We further develop an unsupervised strategy to select positive and negative samples, which eliminates pseudo-negative samples guided by clustering labels. Specifically, it employs confidence level to distinguish positive and negative samples. Consequently, our method achieves higher recognition accuracy on benchmark datasets.

1 Introduction

Feature dimensionality reduction is a crucial tool in analyzing complex high-dimensional data [Doersch *et al.*, 2015]. Current feature dimensionality reduction methods are roughly divided into the following types: principal component analysis (PCA) [Abdi and Williams, 2010] and linear discriminant analysis (LDA) [Belhumeur *et al.*, 1997], as well as structure preservation methods such as local preserving projection (LPP) [He and Niyogi, 2003] and neighborhood preserving embedding (NPE) [He *et al.*, 2005]. Among these methods, PCA has gained significant attention as a well-established feature extraction approach and has been widely applied in various scenarios.

*Corresponding Author

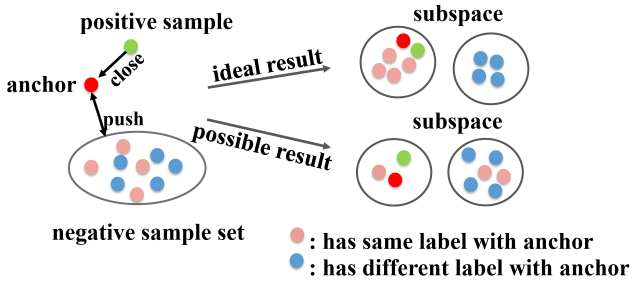


Figure 1: This figure shows the problem of pseudo-negative samples. Samples having the same true label with the anchor are treated as negative data. This may lead to the same class of samples being pushed away from each other. It is suboptimal for learning good visual features.

To overcome the influence of outliers and improve the robustness of PCA, several robust metric PCA has been proposed [Ke and Kanade, 2005; Nie *et al.*, 2014; Ding *et al.*, 2006; Wang *et al.*, 2017b; Wang *et al.*, 2017a]. They replace the squared L2-norm by new distance metric to weaken the influence of noise. Therefore, several studies leverage L1-norm, F-norm, and L21-norm to obtain robust projection vectors [Ding *et al.*, 2006]. For example, Ke and Kanade [2005] proposed L1-PCA that uses L1-norm to measure reconstruction error in the objective function; Kwak [2008] proposed a robust principal component analysis based on L1-norm maximization, which is robust to outliers. However, L1-norm maximization problem is difficult to be solved directly and thus has to be optimized with a greedy strategy. Specifically, the projection direction is sequentially optimized, and the greedy algorithm is easy to fall into the local solution. To mitigate this issue, Wang et al. [2014] developed an optimization algorithm that solves the maximization problem of L1-norm directly with the non-greedy strategy.

However, L1-norm does not satisfy rotational invariance which is an important property in learning methods. To address this problem, Ding et al. [2006] used R1-norm to extract features and proposed a rotational invariant L1-norm PCA (R1-PCA). He et al. [2011] introduced a Half-Quadratic PCA (HQ-PCA), which uses the maximum correntropy criterion to

realize rotational invariant PCA. Zhang et al. [2019] proposed an algorithm named Robust Principal Component Analysis with Adaptive Neighbors (RPCA-AN). In this method, the idea of adaptive weight learning is applied to robust principal component analysis. Nie et al. [2014] proposed an algorithm called Optimal Mean Robust Principal Component Analysis (RPCA-OM). The robust PCA objective function of this method can automatically remove the optimal mean value.

Although the above algorithms effectively suppress the influence of outliers and retain rotational invariance, they do not consider the class structure relationships among samples in the embedding space. However, extracting discriminative features is undoubtedly beneficial to downstream classification tasks [Zhang *et al.*, 2022; Arora *et al.*, 2019]. As a result, contrastive learning has been widely applied to various models and has shown promising results [Ye *et al.*, 2019]. By reducing the distance between positive pairs and increasing the distance between negative pairs, contrastive learning effectively enhances the discriminative capacities of features. Nevertheless, the issue of pseudo-negative samples has consistently influenced its performance [Wang *et al.*, 2020; Wei *et al.*, 2020; Sun *et al.*, 2019], as shown in Figure 1.

Inspired by the capability of contrastive learning to capture sample differences, we apply contrastive learning to the weighted PCA and propose a novel strategy to tackle the "pseudo-negative samples" problem, which helps to enhance the discriminability of the features extracted by PCA. Besides, the connection between PCA and contrastive learning can be established through the projection matrix. Specifically, we first employ the original sample as the anchor point and reconstruct it with the projection matrix generated by PCA, which is treated as a positive sample. For negative samples, we utilize the labels acquired by clustering to remove pseudo-negative samples. Finally, we calculate the confidence level of each pseudo-negative sample and iteratively eliminate those with a high confidence level. Our contributions can be summarized as follows:

- We consider the diversity of the samples by assigning a small weight to the reconstruction error of outliers, which enhances the robustness of PCA and effectively resists the impact of outliers.
- We integrate contrastive learning into weighted PCA, and with the pseudo-labels generated by K-means, we remove the pseudo-negative samples belonging to the same class as the anchors, which improves the discriminative capacity of the features extracted by the PCA algorithm.
- We conduct experiments on six datasets to evaluate the performance of our method, which indicates our method greatly improves the classification accuracy and demonstrates its effectiveness in capturing more information about sample differences.

2 Related Work

In this section, we briefly introduce two related works about our proposed model, i.e., optimal mean principal component analysis and instance-level contrastive learning.

2.1 Optimal Mean Principal Component Analysis

PCA is a well-known unsupervised dimension reduction method that uses a projection matrix to reduce the dimension of data features. To adjust the features on the same scale, PCA needs to center the features with a mean matrix. [Nie *et al.*, 2014] proposed an optimal mean PCA where the mean matrix is optimized during training to mitigate the sensitivity of PCA to noise.

Let $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbf{R}^{d\times N}$ be a data matrix. N is the number of data points. $\mathbf{x}_i \in \mathbf{R}^{d\times 1}$ denotes the ith data point. $\mathbf{W} \in \mathbf{R}^{d\times c}$ is the projection matrix. $\mathbf{m}$ is the mean matrix. The reconstruction error in the form of the optimal mean PCA loss function is as follows:

$$\begin{aligned} &\min_{\mathbf{m},\mathbf{W}} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{m} - \mathbf{W}\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})||_2^2 \\ &s.t. \mathbf{W}^T\mathbf{W} = \mathbf{I} \end{aligned} \tag{1}$$

where $\mathbf{W}\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})$ is the reconstructed data matrix; $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ is the constraint term.

2.2 Instance Level Contrastive Learning

The goal of instance-level contrastive learning is to minimize the distance between positive pairs while maximizing the distance between negative pairs. The contrastive learning cross-entropy loss function is defined as follows:

Definition 1. *For the two descriptions* $\mathbf{A}$ *and* $\mathbf{B}$*, where* $\mathbf{A} \in \mathbf{R}^{N\times d}$*,* $\mathbf{B} \in \mathbf{R}^{N\times d}$*, the number of samples is* N *and the dimension is* d*.* $\mathbf{a}_i$ *and* $\mathbf{b}_i$ *respectively represent the* i*-th line of* $\mathbf{A}$ *and* $\mathbf{B}$*,* i.e.*, the* i*-th sample. Suppose* $s(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}^T}{\|\mathbf{a}\|\|\mathbf{b}\|}$ *represents the cosine similarity of the sample* $\mathbf{a}$ *and the sample* $\mathbf{b}$*, and then the contrastive loss can be defined as below:*

$$L = \frac{1}{2N}(L_A + L_B) \tag{2}$$

where,

$$L_A = \sum_{i=1}^{N} -\log \frac{\exp\left[\frac{s(\mathbf{a}_i,\mathbf{b}_i)}{\tau}\right]}{\sum_{j=1}^{N}\left[\exp\left[\frac{s(\mathbf{a}_i,\mathbf{a}_j)}{\tau}\right] + \exp\left[\frac{s(\mathbf{a}_i,\mathbf{b}_j)}{\tau}\right]\right]} \tag{3}$$

$$L_B = \sum_{i=1}^{N} -\log \frac{\exp\left[\frac{s(\mathbf{b}_i,\mathbf{a}_i)}{\tau}\right]}{\sum_{j=1}^{N}\left[\exp\left[\frac{s(\mathbf{b}_i,\mathbf{b}_j)}{\tau}\right] + \exp\left[\frac{s(\mathbf{b}_i,\mathbf{a}_j)}{\tau}\right]\right]} \tag{4}$$

$\mathbf{a}_i = \mathbf{x}_i$*;* $\mathbf{b}_i = \mathbf{W}\mathbf{W}^T\mathbf{x}_i$*;* $s(\cdot,\cdot)$ *is the cosine similarity function;* $\tau \in (0, 1]$ *is the monitoring temperature.*

3 The Proposed Method

In this section, we provide a detailed discussion of our model. We start with our motivation and objective function. Subsequently, we explain our reconstruction-weighted PCA and instance-level contrastive learning implementation. Furthermore, we introduce the process of using the clustering results to compute the confidence level of each sample and selectively eliminate pseudo-negative samples.

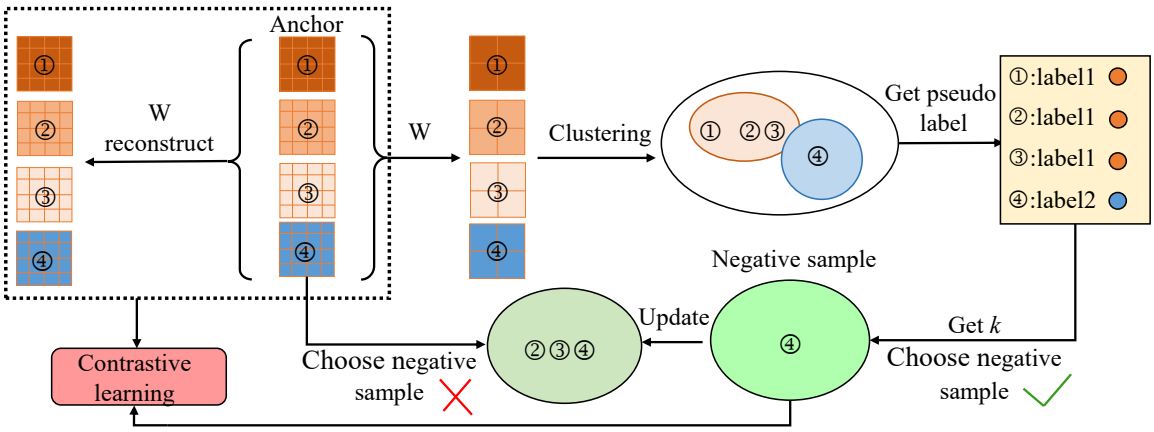


Figure 2: The structure of our proposed PCA method. We use contrastive learning on the real sample and the reconstructed sample obtained using the projection matrix W. Then we exclude the pseudo-negative examples in contrastive learning using the pseudo-labels obtained by clustering in the projection space. Instances ①②③ have the same true label, while ④ has a different label. We use pseudotags to remove pseudo-negative examples ②③ from the instance-level contrast loss, thus improving the classification performance of PCA-extracted features.

3.1 Motivation and Objective Function

To enhance the robustness of PCA, we incorporate a reconstruction error approach. Simultaneously, we combine contrastive learning with PCA to ensure that the projected features retain more class information. Moreover, we use clustering labels to compute each sample's confidence level. Then, we eliminate pseudo-negative samples with high confidence levels. The rate of eliminated samples will increase with the number of iterations.

Specifically, we perform contrastive learning between real and reconstructed samples. The connection between PCA loss and contrastive learning loss is established through the projection matrix $\mathbf{W}$. To keep similar samples from being pushed far away, we use clustering labels to guide contrastive learning to exclude pseudo-negative examples. The block diagram of our model is shown in Figure 2. Our objective function can be written as follows:

$$\min_{\mathbf{r},\mathbf{m},\mathbf{W}} \sum_{i=1}^{N} \frac{1}{r_i} ||\mathbf{x}_i - \mathbf{m} - \mathbf{W}\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})||_2^2 + \lambda L \quad (5)$$

where $\mathbf{W}$ is the projection matrix of PCA; L is the contrastive learning function; λ is the weight of contrastive learning loss; $\mathbf{x}_i$ is the i-th row of $\mathbf{X}$; $\mathbf{m}$ is the mean matrix; r_i is the reconstructed weight, which will be discussed later.

In Eq.(2), $\mathbf{b}_i$ represents the i-th reconstructed sample, and $\mathbf{a}_i$ represents its corresponding original sample. Other original samples and reconstructed samples that are not pseudo-negative samples are treated as negative samples. It can then be considered that $L_A = L_B$, and L in our model can be rewritten as:

$$\begin{aligned} L &= \frac{1}{N} L_B \\ &= \frac{1}{N} \sum_{i=1}^{N} -\log \frac{\exp\left[\frac{s(\mathbf{b}_i,\mathbf{a}_i)}{\tau}\right]}{\sum_{j=1}^{N}\left[\exp\left[\frac{s(\mathbf{b}_i,\mathbf{b}_j)}{\tau}\right] + \exp\left[\frac{s(\mathbf{b}_i,\mathbf{a}_j)}{\tau}\right]\right]} \end{aligned} \quad (6)$$

Therefore, the objective function of our model in Eq.(5) can be rewritten as:

$$\begin{aligned} &\min_{\mathbf{m},\mathbf{r},\mathbf{W}} \sum_{i=1}^{N} \frac{1}{r_i} ||\mathbf{x}_i - \mathbf{m} - \mathbf{W}\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})||_2^2 \\ &+ \lambda \frac{1}{N} \sum_{i=1}^{N} -\log \frac{\exp(\frac{s_1}{\tau})}{\sum_{j=1,j\neq k}^{N}\left[\exp(\frac{s_2}{\tau}) + \exp(\frac{s_3}{\tau})\right]} \\ &s.t. \mathbf{r} \geqslant 0, \mathbf{r}^T\mathbf{1} = 1, \mathbf{W}^T\mathbf{W} = \mathbf{I} \end{aligned} \quad (7)$$

Where $s_1 = s(\mathbf{b}_i, \mathbf{a}_i)$, $s_2 = s(\mathbf{b}_i, \mathbf{b}_j)$; $s_3 = s(\mathbf{b}_i, \mathbf{a}_j)$; k is the index of the pseudo-negative sample, which will be discussed later; $\tau \in (0, 1]$ is the monitoring temperature.

3.2 Weighting the Reconstruction Error of Each Sample

The presence of outliers can significantly impact the projection direction of PCA since their reconstruction errors tend to be much larger compared to normal values. To address this issue, we propose a novel weighted PCA. To be specific, we weight the reconstruction error of each sample instead of weighting the samples directly. This approach allows for a more direct reduction of the reconstruction error of outliers, and the weights are taken in a larger range. The reconstruction weights in our method are as defined follows:

$$\min_{r_i} \sum_{i=1}^{N} \frac{1}{r_i} g(\mathbf{x}_i), s.t. \sum_{i=1}^{N} r_i = 1, 0 \leqslant r_i \leqslant 1 \quad (8)$$

where $g(\mathbf{x}_i)$ is the reconstruction error of $\mathbf{x}_i$, denoted by g_i in the following derivation; r_i is the reconstruction weight.

We use the Lagrange multiplier method to solve the reconstruction weights r_i.

$$\mathcal{L} = \sum_{i=1}^{N} \frac{g_i}{r_i} + \beta(1 - \sum_{i=1}^{N} r_i) + \sum_{i=1}^{N} \gamma_i(-r_i) \quad (9)$$

where β, γ_i are Lagrange multipliers. The KKT conditions are as follows:

$$\begin{cases} -\frac{g_i}{{r_i}^2} - \beta - \gamma_i = 0 \\ \gamma_i r_i = 0 \\ \gamma_i \geqslant 0 \\ \sum_{i=1}^{N} r_i = 1 \end{cases} \tag{10}$$

Then the optimal solution of the weights r_i is below:

$$r_i = \frac{\sqrt{g_i}}{\sum_{i=1}^{N} \sqrt{g_i}} \tag{11}$$

Thus, the reconstruction weighting PCA loss is as follows:

$$\begin{aligned} &\min_{\mathbf{m},\mathbf{r},\mathbf{W}} \sum_{i=1}^{N} \frac{1}{r_i} ||\mathbf{x}_i - \mathbf{m} - \mathbf{W}\mathbf{W}^T(\mathbf{x}_i - \mathbf{m})||_2^2 \\ &s.t. \mathbf{r} \geqslant 0, \mathbf{r}^T\mathbf{1} = 1, \mathbf{W}^T\mathbf{W} = \mathbf{I} \end{aligned} \tag{12}$$

where $\mathbf{r} = [r_1, r_2, ..., r_N]^T$ is the vector of reconstruction weights r_i; $\mathbf{m}$ is the mean matrix; $\mathbf{W}$ is the projection matrix.

3.3 Removing Pseudo-Negative Samples During Contrastive Learning

Considering the low complexity of the K-means algorithm, we first utlize K-means to obtain the pseudo-labels of each projected sample. Next, the confidence level is obtained by calculating the cosine similarity between each sample and the cluster centroid. A high confidence level means that the sample is close to the cluster center, indicating the sample is more likely to be a pseudo-negative example of its corresponding class. The confidence level of $\mathbf{x}_i$ is calculated as follows:

$$con_i = \frac{sim(\mathbf{z}_i, \mathbf{c}_{yi})}{\sum_{j=1, j\neq yi}^{p} sim(\mathbf{z}_i, \mathbf{c}_j)} \tag{13}$$

Where $\mathbf{z}_i = \mathbf{W}^T(\mathbf{x}_i - \mathbf{m})$; $\mathbf{c}_{yi}$ is the cluster center of $\mathbf{z}_i$; p is the class number of $\mathbf{X}$; $\mathbf{c}_j$ is the cluster center except $\mathbf{c}_{yi}$; sim is the cosine similarity function.

Pseudo-negative examples are selected according to the confidence level. First, we rank the confidence levels of each sample in descending order. Then, we select the pseudo-negative example with the highest confidence level and exclude it from the denominator of the contrastive learning objective function. The proportion of choices is constantly increasing with each iteration. The subscript of the selected pseudo-negative example can be expressed as follows:

$$k = \eta * Index(\mathbf{con}) \tag{14}$$

where $\eta = 0 \sim 100\%$ is the rate of selected pseudo-negative samples; $\mathbf{con} = [con_1, ..., con_N] \in [0, 1]$ is the sorted confidence level of all samples; $Index$ is a function that can get the subscript of $\mathbf{z}_i$ corresponding to the confidence level con_i.

Our method prevents similar samples from being projected to greater distances by deleting pseudo-negative examples with the subscript k in contrastive learning. It is worth noting that the acquisition of pseudo-negative examples and the extraction of PCA features are carried out alternately and enhance each other, thereby improving the discriminative quality of the recovered features. The detailed optimization can be found in the **Supplementary material** [1].

Dataset	Quantity	Dimension	Class Number
Extended Yale B	2144	1024	38
CMU PIE	2856	1024	68
AR	3120	2000	120
ORL	400	1024	40
Lung	203	3312	5

Table 1: Parameters of the datasets used in our experiments.

4 Experiment

4.1 Experiment Setting

All our experiments were conducted on the Windows 10 operating system using Python 3.7. The recognition accuracy follows the 1-nearest neighbor approach. To prevent significant differences in clustering outcomes during iterations, we use the K-means and fix random seed as 4 in the training process. Furthermore, for fairness, we run each method for 10,000 epochs and set the algorithm parameters in accordance with the specifications in their papers.

4.2 Datasets Information

Extended Yale B:[Georghiades *et al.*, 2001] The Yale B dataset comprises 5850 facial images from 10 individuals, later expanded and renamed the Extended YaleB dataset. It encompasses 2144 frontal images of 38 subjects captured under diverse lighting setups. Images were resized to 32×32 dimensions and noise blocks with ratios ranging from 0.05 to 0.15 were applied to 14 images per person. From each individual, 32 images were randomly chosen, with 7 noisy images assigned for training, totaling 1216 images. The remaining 1198 images were allocated for testing purposes.

CMU PIE:[Sim *et al.*, 2002] The CMU PIE database was created by Carnegie Mellon University. It contains 4,400 frontal faces of 68 people. These images were taken under 43 kinds of light, 13 kinds of postures, and 4 kinds of expressions. Our experiment adopts the No. 27 posture, which contains 2856 positive images, 42 for each person. We also add random noise to 10 images of each person, and the noise ratio is 0.05 to 0.15. We selected 1428 for training and the rest for testing.

AR: [Martinez and Benavente, 1998] The AR dataset comprises 4,000 color facial images from 56 females and 70 males, featuring diverse lighting conditions, expressions, poses, and occlusions. Biweekly image capture resulted in 26 images per subject. Experimentation involved a subset of 120 subjects, with varying occluded, unobstructed, and subsequent session images. Images were resized to 50x40 pixels,

[1]https://github.com/lml314159/IJCAI2024

Method	CMU PIE	Extended Yale B	AR+All	AR+glass	ORL	Lung
L2-PCA	82.65±0.77	52.25±0.69	59.38±0.72	67.00±0.13	83.50±1.35	92.23±0.53
GreedyL1-PCA	84.52±0.73	54.64±0.64	59.24±1.02	67.50±1.43	85.50±1.53	92.23±1.12
Non-greedyL1-PCA	82.06±0.70	49.95±0.90	59.11±0.53	67.00±0.54	85.50±0.89	92.23±0.75
HQ-PCA	83.28±0.89	53.15±0.77	59.96±0.57	68.00±0.57	86.00±1.04	92.32±1.20
R1-PCA	85.23±0.67	55.57±0.51	60.08±0.51	67.50±0.53	84.50±0.86	93.20±0.30
RPCA-OM	85.18±0.65	55.48±0.58	60.24±0.68	69.00±0.56	83.00±1.56	92.23±0.64
RPCA-AN	85.47±0.60	56.21±0.68	60.13±0.59	68.50±0.35	84.50±1.11	93.20±0.76
TRPCA	85.59±0.64	54.97±0.75	60.21±0.74	68.50±0.76	85.76±0.75	93.32±0.74
EGCFS	85.69±0.54	52.56±0.66	58.87±1.35	66.00±0.85	87.50±0.54	92.23±1.20
PCA-CL(ours)	**86.27±0.64**	**57.60±0.75**	**60.83±0.74**	**70.28±0.76**	**90.50±0.75**	**93.32±0.74**

Table 2: The average recognition accuracy(%) and standard deviation on the six datasets with a fixed dimension of 100.

PCA	RW	CL	k	CMU PIE	Extended Yale B	AR+ALL	AR+glass	ORL	Lung
✓				82.65	52.25	60.13	64.67	85.45	89.32
✓	✓			83.43	53.32	60.32	65.13	86.67	90.42
✓		✓		84.22	55.22	60.54	66.22	87.72	91.43
✓	✓	✓		85.96	56.50	60.77	68.14	88.34	91.67
✓	✓	✓	✓	**86.27**	**57.60**	**60.83**	**70.28**	**90.50**	**93.32**

Table 3: The recognition rate (%) under different settings with a fixed dimension of 100. RW represents the reconstruction error weights, CL represents contrastive learning, and k represents the strategy of removing pseudo-negative examples.

with 1560 images designated for training and the remainder for testing. Occluded images were deemed noisy samples. Two categories emerged: (1) AR+all, using initial session images for training and subsequent session images for testing; (2) AR+glass, employing non-disguised subjects' initial session images along with one glass-wearing image for training, while two additional glass-wearing images were incorporated from both initial and subsequent sessions for testing.

ORL:[Jin *et al.*, 2001] The ORL dataset includes 400 grayscale frontal facial images of 40 subjects, showcasing variations in shooting time, posture (frontal, lateral), expression (open eyes, closed eyes, smiling), lighting conditions, occlusion (presence of sunglasses), and environmental factors. Images were resized uniformly to 32x32 dimensions for experimentation. Half of each subject's images were randomly assigned for training, with the other half reserved for testing. Images with shading or occlusions were classified as noisy samples.

Lung:[Li *et al.*, 2018] The Lung dataset constitutes a biological repository comprising 203 samples categorized into five distinct classes. Our experimental protocol entails randomly partitioning 50% of the images, totaling 539 samples per category, for the training subset, while the residual samples are earmarked for the test subset. This partitioning procedure is iterated tenfold to yield ten independent sets of randomized datasets.

4.3 Comparison Algorithms

We compare our method with L2-PCA [Turk and Pentland, 1991], GreedyL1-PCA [Kwak, 2008], Non-greedyL1-PCA [Wang *et al.*, 2014], HQ-PCA [He *et al.*, 2011], R1-PCA [Ding *et al.*, 2006], RPCA-OM [Nie *et al.*, 2014], RPCA-AN [Zhang and Tong, 2019], TRPCA [Nie *et al.*, 2020], and a feature extraction method based on graph structure: EGCFS [Zhang *et al.*, 2020]. Then we carry out experiments on Extended Yale B [Georghiades *et al.*, 2001], CMU PIE [Sim *et al.*, 2002], AR [Martinez and Benavente, 1998], AR+glass, ORL, and Lung datasets. We will briefly introduce the above comparison methods:

L2-PCA: It uses L2-norm to constrain reconstruction error, which is a traditional PCA algorithm.

Greedy-L1PCA: It uses the L1-norm to measure reconstruction error, which has the advantage of satisfying rotation invariance.

Non-greedy-L1PCA: It proposes a non-greedy algorithm to solve the L1-norm maximization problem, which ensures that the algorithm converges to the global optimal solution.

HQ-PCA: It's based on the maximum entropy criterion and has rotation invariance.

R1-PCA: It uses the R1 norm to replace the traditional F norm, which has rotation invariance, and uses the subspace iteration method to solve the model.

RPCA-OM: It uses the optimal mean to replace the fixed mean so that the mean matrix is optimized during the iteration. This approach effectively reduces the sensitivity of PCA to noise.

RPCA-AN: It introduces the adaptive weight into PCA, effectively eliminates outliers in the training process and enhances the robustness of PCA.

TRPCA: It ignores outliers by truncation and minimizes the reconstruction error in $\ell_{2,1}$-norm.

EGCFS: A graph structure-based feature extraction method.

4.4 Recognition Accuracy

In this section, we compare the recognition accuracy of our method with the mentioned comparison methods. We conduct feature extraction on six datasets with these six methods

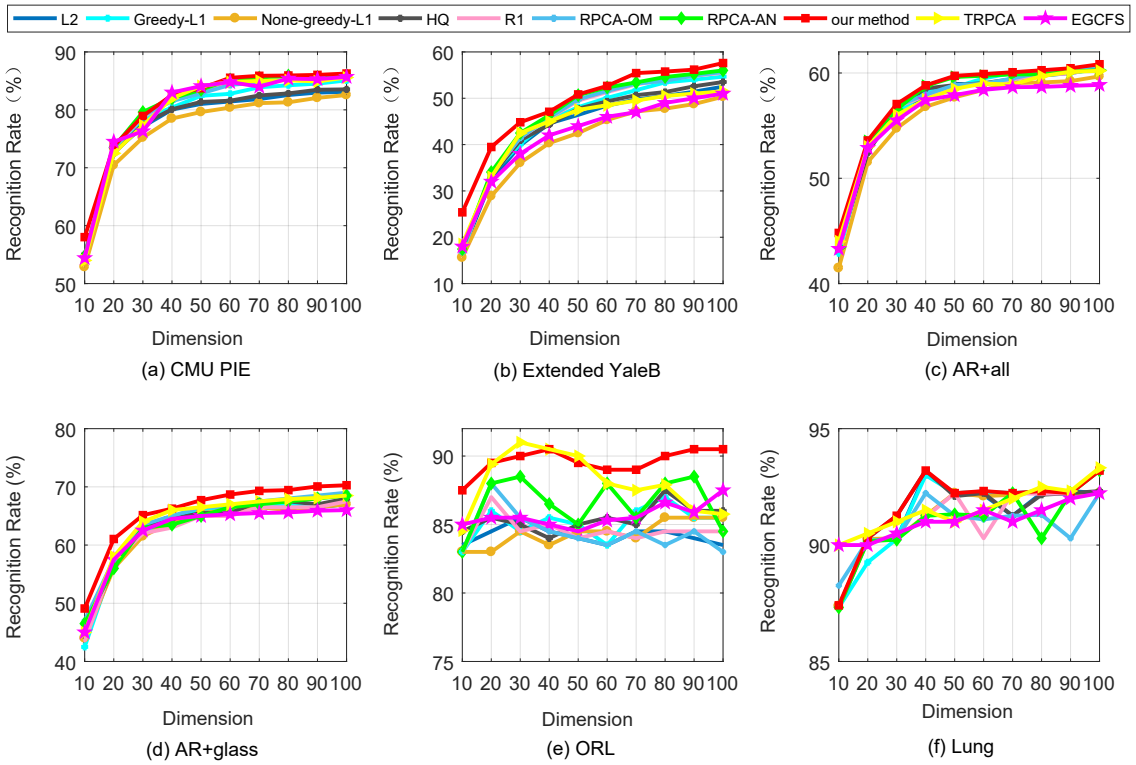
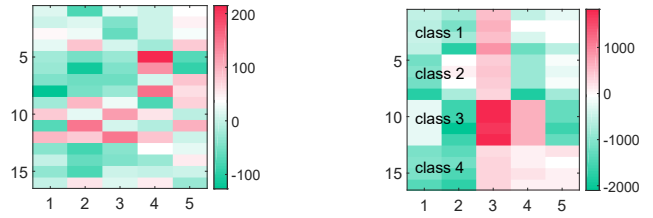


Figure 3: The average recognition curve on the CMU PIE, Extended YaleB, AR+all, AR+glass, ORL, Lung datasets.

and analyze the recognition rate curves across different dimensions, as illustrated in Figure 3. Subsequently, we present the average recognition rates of all methods at a fixed dimension of 100, as summarized in Table 2. The experimental results reveal that our proposed approach initially exhibits commendable performance and surpasses other PCA techniques as dimensionality increases, indicating its robustness against outliers. Furthermore, we conduct ten iterations of experiments and maintain a constant projection dimension of 100 across three datasets. Our method consistently outperforms the compared methods in terms of accuracy, which demonstrates its effectiveness in extracting discriminative features and consequently enhancing classification accuracy. In summary, our method shows its superiority in capturing discriminative features for accuracy improvement.

4.5 Low-dimensional Data Visualization

Figure 4 illustrates that the low-dimensional features obtained by applying the proposed PCA method are more discriminative, which indicates that features within the same category exhibit higher similarity. The figure consists of 16 samples displayed along the vertical axis from the four categories in the AR dataset. Each set of four consecutive samples corresponds to the same category and includes five selected attributes per sample, displayed along the horizontal axis. Figure 4(a) depicts the digitized features before PCA dimensionality reduction, in which different colors represent different digitized features whose values are up to four decimal places. It is evident that the feature values before PCA reduction are ambiguous, posing challenges in distinguishing between classes. Figure 4(b) depicts the visualized features after PCA reduction, which reveals that the feature values of four consecutive samples within the same category exhibit higher similarity, as indicated by color consistency. This observation provides evidence that the PCA method proposed in this paper effectively extracts the discriminative features present in the dataset.



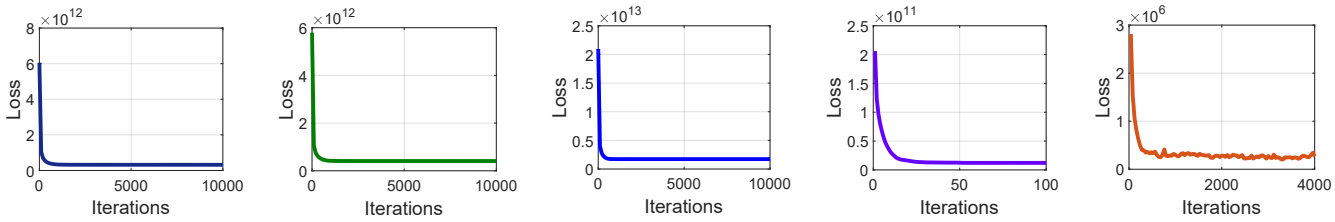
(a) Initial low-dimensional data

(b) Low-dimensional data after 10000 training epoches

Figure 4: A low-dimensional projection of 16 samples from 4 classes in the AR dataset. Each row represents a sample, and each column represents a feature.

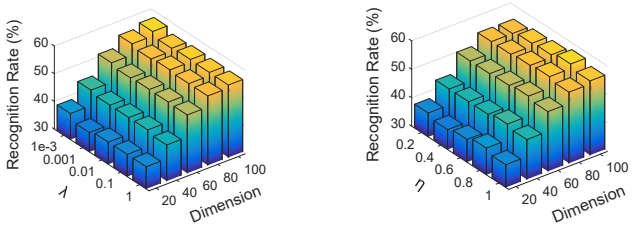
4.6 Ablation Experiment

We conducted the ablation experiments to assess the efficacy of each component of our method. Table 3 presents its recognition outcomes across different configurations. The experimental results indicate that contrastive learning notably



(a) CMU PIE (b) Extended Yale B (c) AR (d) ORL (e) Lung

Figure 5: The convergence curve on CMU PIE, Extended Yale B, AR , ORL and Lung datasets.



(a) Recognition rate for different values of λ in each dimension

(b) Recognition rate for different values of η in each dimension

Figure 6: Parametric analysis on the Extended Yale B dataset

enhances recognition accuracy across six different datasets, demonstrating its capacity to extract discriminative features. Specifically, after the elimination of pseudo-negative instances, the proposed method achieves improvement of 0.31%, 1.1%, 0.06%, 2.14%, 2.16%, and 1.65% on the CMU PIE, Extended YaleB, AR+all, AR+glass, ORL, and Lung datasets, respectively. The efficacy of the pseudo-negative elimination strategy is particularly effective on Yale B, ORL, and Lung datasets, potentially owing to their reduced category diversity and consequent heightened clustering accuracy. Consequently, the removal of pseudo-negative instances yields more precise representations and hence strengthens the efficacy of contrastive learning mechanisms.

4.7 Parametric Analysis

Figure 6 illustrates the effect of two hyperparameters on the Extended Yale B dataset in each dimension. When the hyperparameter η is fixed as 0.8, the recognition rate under each dimension reaches its highest value when the weight λ of contrastive learning is set to around 0.0001. When setting λ with different values, the resulting changes in recognition accuracy are not significant. When fixing λ to 0.0001, it is found that the recognition accuracy under each dimension reaches its highest value when η is set to around 0.8, while the recognition accuracy does not change much when it is set to other values. Therefore, when λ is set to 0.0001 and η is set to 0.8, this method has the highest recognition accuracy.

4.8 Time Complexity Analysis

The time complexity of our method is comparable to the compared methods since they all rely on the SVD decomposition of the projection matrix $\mathbf{W}$. Specifically, when applied to a dataset of dimension $N \times d$ (N is sample number, d is feature dimension), PCA has a complexity of $O(Nd^2)$. The detailed optimization can be seen in the **Supplementary material**[2] .

4.9 Convergence Analysis

We analyze the convergence of our optimization algorithm with an extra experiment on these datasets. Figure 5 depicts the objective value with regard to the iteration number. The objective value decreases as the iteration number increases, and eventually, it converges to a constant value. It indicates that our algorithm successfully minimizes the objective function and achieves good convergence.

5 Conclusion

This work presents a novel weighted PCA method that aims to enhance feature discriminability and mitigate the influence of outliers. The weight assigned to each sample is determined based on its reconstruction error, with higher weights assigned to samples with lower reconstruction errors, which reduces the impact of outliers on the algorithm. It further improves discriminability by leveraging contrastive learning, which removes pseudo-negative samples to enhance the separability between low-dimensional features of different categories. Experiments on six datasets demonstrate the effectiveness and superiority of our method.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant 62176203 and Grant 62102306, the Fundamental Research Funds for the Central Universities, the Natural Science Basic Research Program of Shaanxi Province (Grant 2023-JC-YB-534), Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (LCNBI) and ZJLab, and the Science and technology project of Xi'an (Grant 2022JH-JSYF-0009), Initiative Postdocs Supporting Program (Grant BX20190262), the Fundamental Research Funds for the Central Universities(2242024k30038).

[2]https://github.com/lml314159/IJCAI2024

References

[Abdi and Williams, 2010] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[Arora *et al.*, 2019] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[Belhumeur *et al.*, 1997] Peter N. Belhumeur, Joao P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*, 19(7):711–720, 1997.

[Ding *et al.*, 2006] Chris Ding, Ding Zhou, Xiaofeng He, and Hongyuan Zha. R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *ICML*, pages 281–288, 2006.

[Doersch *et al.*, 2015] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE ICCV*, pages 1422–1430, 2015.

[Georghiades *et al.*, 2001] Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI*, 23(6):643–660, 2001.

[He and Niyogi, 2003] Xiaofei He and Partha Niyogi. Locality preserving projections. *NeurIPS*, 16, 2003.

[He *et al.*, 2005] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang. Neighborhood preserving embedding. In *IEEE ICCV*, volume 2, pages 1208–1213. IEEE, 2005.

[He *et al.*, 2011] Ran He, Bao-Gang Hu, Wei-Shi Zheng, and Xiang-Wei Kong. Robust principal component analysis based on maximum correntropy criterion. *IEEE TIP*, 20(6):1485–1494, 2011.

[Jin *et al.*, 2001] Zhong Jin, Jing-Yu Yang, Zhong-Shan Hu, and Zhen Lou. Face recognition based on the uncorrelated discriminant transformation. *PR*, 34(7):1405–1416, 2001.

[Ke and Kanade, 2005] Qifa Ke and Takeo Kanade. Robust l/sub 1/norm factorization in the presence of outliers and missing data by alternative convex programming. In *IEEE CVPR*, volume 1, pages 739–746. IEEE, 2005.

[Kwak, 2008] Nojun Kwak. Principal component analysis based on l1-norm maximization. *IEEE TPAMI*, 30(9):1672–1680, 2008.

[Li *et al.*, 2018] Xuelong Li, Han Zhang, Rui Zhang, Yun Liu, and Feiping Nie. Generalized uncorrelated regression with adaptive graph for unsupervised feature selection. *IEEE TNNLS*, 30(5):1587–1595, 2018.

[Martinez and Benavente, 1998] Aleix Martinez and Robert Benavente. The ar face database: Cvc technical report, 24. *CVC*, 1998.

[Nie *et al.*, 2014] Feiping Nie, Jianjun Yuan, and Heng Huang. Optimal mean robust principal component analysis. In *ICML*, pages 1062–1070. PMLR, 2014.

[Nie *et al.*, 2020] Feiping Nie, Danyang Wu, Rong Wang, and Xuelong Li. Truncated robust principle component analysis with a general optimization framework. *IEEE TPAMI*, 44(2):1081–1097, 2020.

[Sim *et al.*, 2002] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *IEEE international conference on automatic face gesture recognition*, pages 53–58. IEEE, 2002.

[Sun *et al.*, 2019] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*, 2019.

[Turk and Pentland, 1991] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[Wang *et al.*, 2014] Rong Wang, Feiping Nie, Xiaojun Yang, Feifei Gao, and Minli Yao. Robust 2dpca with non-greedy ℓ_1 -norm maximization for image analysis. *IEEE TC*, 45(5):1108–1112, 2014.

[Wang *et al.*, 2017a] Qianqian Wang, Quanxue Gao, Xinbo Gao, and Feiping Nie. Angle principal component analysis. In *IJCAI*, pages 2936–2942, 2017.

[Wang *et al.*, 2017b] Qianqian Wang, Quanxue Gao, Xinbo Gao, and Feiping Nie. $\ell_{2,p}$-norm based pca for image recognition. *IEEE TIP*, 27(3):1336–1346, 2017.

[Wang *et al.*, 2020] Xudong Wang, Ziwei Liu, and X Yu Stella. Unsupervised feature learning by cross-level discrimination between instances and groups. *arXiv preprint arXiv:2008.03813*, 1(2):4, 2020.

[Wei *et al.*, 2020] Chen Wei, Huiyu Wang, Wei Shen, and Alan Yuille. Co2: Consistent contrast for unsupervised visual representation learning. *arXiv preprint arXiv:2010.02217*, 2020.

[Ye *et al.*, 2019] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *IEEE CVPR*, pages 6210–6219, 2019.

[Zhang and Tong, 2019] Rui Zhang and Hanghang Tong. Robust principal component analysis with adaptive neighbors. *NeurIPS*, 32, 2019.

[Zhang *et al.*, 2020] Rui Zhang, Yunxing Zhang, and Xuelong Li. Unsupervised feature selection via adaptive graph learning and constraint. *IEEE TNNLS*, 33(3):1355–1362, 2020.

[Zhang *et al.*, 2022] Hongjie Zhang, Wenwen Qiang, Jinxin Zhang, Yingyi Chen, and Ling Jing. Unified feature extraction framework based on contrastive learning. *Knowledge-Based Systems*, 258:110028, 2022.