

Deep Online Probability Aggregation Clustering

Yuxuan Yan, Na Lu*, and Ruofan Yan

Systems Engineering Institute, Xi'an Jiaotong University
yan1611@stu.xjtu.edu.cn, lvna2009@mail.xjtu.edu.cn,
yanruofan@stu.xjtu.edu.cn

Abstract. Combining machine clustering with deep models has shown remarkable superiority in deep clustering. It modifies the data processing pipeline into two alternating phases: feature clustering and model training. However, such alternating schedules may lead to instability and computational burden issues. To tackle these problems, we propose a centerless clustering algorithm called Probability Aggregation Clustering (PAC), enabling easy deployment in online deep clustering. PAC circumvents the cluster center and aligns the probability space and distribution space by formulating clustering as an optimization problem with a novel objective function. Based on the computation mechanism of the PAC, we propose a general online probability aggregation module to perform stable and flexible feature clustering over mini-batch data and further construct a deep visual clustering framework deep PAC (DPAC). Extensive experiments demonstrate that DPAC remarkably outperforms the state-of-the-art deep clustering methods.¹

Keywords: Deep Clustering · Unsupervised Learning · Fuzzy Clustering

1 Introduction

Clustering analysis [3] is a widely explored domain in the field of unsupervised learning, aiming to group the unlabeled samples into clusters that have common characteristics. Conventional machine clustering is favored by many researchers due to its significant interpretability and stable optimization. In recent years, deep clustering has received more attention due to its powerful representation extraction capability. Previous deep clustering models [8, 24, 55, 56] directly combine deep networks with machine clustering and utilize designed loss functions to guide both representation learning and clustering. For example, Deepcluster [9] and PCL [34] leverage K-means (KM) to cluster unlabeled images. Unfortunately, these offline methods typically require running multiple times of standard KM over the entire dataset, which brings much time and space complexity. Besides, simply grouping data in batches instead of the whole dataset

* Corresponding author

¹ The code is available at <https://github.com/aomandechenai/Deep-Probability-Aggregation-Clustering>

to obtain online clustering causes collapsing and degradation issues. To address these problems, researchers have given two dominant solutions: batch clustering and contrastive clustering.

Batch clustering [20,30,38,57] focuses on modifying the conventional machine clustering algorithms [59] to adapt the data flow of deep models, which has high extensibility. For example, Online Deep Clustering (ODC) [57] decomposes the standard KM process into batch clustering with memory banks and optimizes the clustering and network shoulder-to-shoulder (online) to facilitate stable learning. CoKe [42] proposes the moving average strategy to reassignment pseudo labels and introduces Constrained K-means [7] into training to ensure the minimal size of clusters to avoid collapsing. Most existing batch clustering approaches focus more on center-based machine clustering algorithms, such as KM and fuzzy c-means (FCM) [6], which require specially designed center update rules. Moreover, center-based machine clustering is easily susceptible to the influence of cluster center [4,22]. Random initialization of cluster centers introduces instability to subsequent training. Partitioning based on nearest centers cannot provide fine-grained discrimination hyperplanes for clusters, affecting clustering performance.

Recently, contrastive clustering [36,46,49,60] has achieved significant success in online deep clustering. Contrastive methods perform online clustering by exploring multi-view correlations of data. Formally, instances are augmented into two views using random data augmentation to build contrastive frameworks. The clustering process is then performed by minimizing the designed contrastive loss. PICA [28] proposes cluster-level contrastive loss based on contrastive framework to perform online deep clustering. However, the establishment of contrastive approaches needs a lot of artificial knowledge, including data augmentation, hyperparameter setting, and model architecture. Contrastive models often need thousands of epochs to reach convergence. Besides, they make a balanced assumption for clustering (i.e. each cluster has the same number of samples), which requires additional regular terms to constrain optimization and avoid crash problems (i.e. a few clusters have a majority of instances). The essence of contrastive clustering methods is to leverage the nearest-neighbor relationship of augmented instances in the semantic space to unsupervisedly train the classifier. Such semantic nearest-neighbor learning only uses a portion of data and its corresponding augmented version, failing to capture the global cluster relationship [13] and encode spatial embedding distribution.

In this work, considering the adverse effect of the cluster center, we first introduce a novel objective function quantifying the intra-cluster distances without cluster centers. Furthermore, inspired by fuzzy c-means, a concise optimization program is formulated by incorporating a fuzzy weighting exponent into an objective function. Then we build a centerless machine clustering algorithm called Probability Aggregation Clustering (PAC). In the optimization program of PAC, the probability of one sample belonging to a cluster is aggregated across samples with distance information in an iterative way. Unlike KM which assigns instances by cluster centers, PAC directly outputs probabilities which is more stable and easy to deploy in deep models. Therefore, we extend the PAC to the

online probability aggregation module (OPA), a simple plug-in component for online deep clustering tasks. OPA seamlessly combines the calculation process of PAC with loss computation. It overcomes the disadvantages of both batch and contrastive clustering and implements efficient clustering. Besides, OPA does not impose any constraints on the size of clusters, mitigating the suboptimal solutions introduced by balanced clustering and obtaining more flexible partitioning. It computes clustering codes with the batches of data and updates the network by KL divergence, which leaves out the complicated clustering steps and trains the model in a supervised manner. Based on the above theories, a deep image clustering model Deep PAC (DPAC) is established, which ensures stable learning, global clustering, and superior performance. The major contributions of this work include:

- A novel centerless partition clustering method PAC is proposed to implement clustering by exploring the potential relation between sample distribution and assignment probability.
- An online deep clustering module OPA is exploited based on PAC, which encodes spatial distances into online clustering without incorporating plenty hyper-parameters and components. It leaves out the cluster size constraints to perform flexible partitioning.
- A simple end-to-end unsupervised deep clustering framework DPAC is established for stable and efficient clustering. DPAC achieves significant performance on five challenging image benchmarks compared with the state-of-the-art approaches.

2 Related Work

Deep Clustering: Deep clustering methods [12, 18, 46] combine representation learning with clustering through deep models. ProPos [29] proposes the prototype scattering loss to make full use of K-means pseudo labels. Deepdpm [43] is a density-based approach, which does not require the preset number of class. Different from the above, recent deep clustering methods assume that the output is uniform. SwAV [10] and SeLa [5] adopt a balanced cluster discrimination task via the Sinkhorn-Knopp algorithm. SCAN [50] leverages K-nearest-neighbor information to group samples. Its loss maximizes the agreements of assignments among neighbors, which inevitably need an additional balanced cluster constraint to avoid trivial solutions. SeCu [41] employs a global entropy constraint to relax the balanced constraint to a lower-bound size constraint that limits the minimal size of clusters.

Machine Clustering: Machine clustering [11, 27, 33] tries to decompose the data into a set of disjoint clusters by machine learning algorithms. FCM [6] obtains soft cluster assignment by alternately updating the fuzzy partition matrix and cluster center. Many modified methods [33, 37, 51] aim at improving

the performance and robustness of center-based clustering. In addition, non-parametric methods [21, 23] have received more and more attention in recent years. FINCH [44] performs hierarchical agglomerative clustering based on first-neighbor relations without requiring a specific number of clusters. However, the complex clustering progresses involved in these algorithms hinder their easy deployment in neural networks.

3 Method

The following sections present the theoretical basis of our approach. We first derive a novel objective function and analyze how the proposed objective function relates to existing methods. Second, we present a scalable centerless clustering algorithm PAC. Finally, we extend PAC to a novel online clustering module OPA, and construct a novel online deep clustering model DPAC to learn the semantic knowledge of unlabeled data.

3.1 Objective Function

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ be an N -point dataset, where $\mathbf{x}_i \in \mathbb{R}^{D \times 1}$ is the i -th D -dimensional instance. The clustering algorithm aims to divide $\mathbf{X}$ into K mutually disjoint clusters, where $2 \leq K < N$, $K \in \mathbb{N}$. $\mathbf{P} = [p_{i,k}]_{N \times K}$ is the soft partition matrix, $p_{i,k}$ is the probability of one sample belonging to certain cluster indicating the relationship between sample $\mathbf{x}_i$ and cluster k which satisfies $\mathbf{P} \in \{\Gamma^{N \times K} | \gamma_{i,k} \in [0, 1], \forall i, k; \sum_{k=1}^K \gamma_{i,k} = 1, \forall i; 0 < \sum_{i=1}^N \gamma_{i,k} < N, \forall k\}$. And the cluster prediction of $\mathbf{x}_i$ can be predicted by $\hat{p}_i = \arg \max_k p_{i,k}$, $1 < k \leq K$.

Different from the existing classical center-based methods [6, 53], we utilize the inner product operation of probability vectors instead of cluster center to indicate cluster relations of samples. Formally, we multiply the inner product results with corresponding distance measurements to quantify the global intra-cluster distance of the data. The objective function J_{pac} is defined as:

$$J_{pac} = \sum_{i=1}^N \sum_{j=1}^N \mathbf{p}_i^\top \mathbf{p}_j \|\mathbf{x}_i - \mathbf{x}_j\|^2, \quad (1)$$

where $\mathbf{p}_i = [p_{i,1}, p_{i,2}, \dots, p_{i,K}]^\top$ is the probability vector. $\mathbf{p}_i^\top \mathbf{p}_j \in [0, 1]$ can be regarded as the probability weight for $\|\mathbf{x}_i - \mathbf{x}_j\|^2$. By minimizing Eq. (1), $\mathbf{p}_i^\top \mathbf{p}_j$ can be negatively related to $\|\mathbf{x}_i - \mathbf{x}_j\|^2$, which denotes the probabilities of instances consistent with nearby samples, but not with distant samples.

3.2 Relation to Existing Methods

We provide a new perspective to further understand the proposed objective function. We summarize the difference between our method and Spectral Clustering

(SC) [51] and SCAN [50]. The minimizing problem for J_{pac} can be rewritten as:

$$\min_{\mathbf{P} \in \Gamma^{N \times K}} Tr(\mathbf{P}^\top \tilde{\mathbf{D}}_x \mathbf{P}), \quad (2)$$

where $\tilde{\mathbf{D}}_x$ is the distances matrix, $\tilde{d}_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$. Obviously, $\tilde{d}_{i,j}$ can be replaced by many other distance measurement. We use L_2 distance as the default distance measure in the following experiments. The graph partitioning problem of SC is formulated as:

$$\begin{aligned} \min_{\mathbf{H} \in \mathbb{R}^{N \times K}} Tr(\mathbf{H}^\top \tilde{\mathbf{L}}_x \mathbf{H}), \\ \text{s.t. } \mathbf{H}^\top \mathbf{H} = \mathbf{I}, \end{aligned} \quad (3)$$

where $\tilde{\mathbf{L}}_x$ is the Laplacian matrix of graph. The indicator matrix $\mathbf{H}$ contains arbitrary real values with orthogonality constraint. The semantic clustering loss in SCAN can be reformulated as:

$$\begin{aligned} \max_{\mathbf{P} \in \Gamma^{N \times K}} \sum_{i=1}^N \sum_{j \in \mathcal{N}_i} \log \mathbf{p}_i^\top \mathbf{p}_j - \lambda \mathcal{H}(\mathbf{P}) \\ \Leftrightarrow \max_{\mathbf{P} \in \Gamma^{N \times K}} Tr(\mathbf{P}^\top \tilde{\mathbf{A}}_x \mathbf{P}) - \lambda \mathcal{H}(\mathbf{P}), \end{aligned} \quad (4)$$

where $\mathcal{H}(\mathbf{P}) = \sum_{k=1}^K \frac{\sum_{i=1}^N p_{i,k}}{N} \log \frac{\sum_{i=1}^N p_{i,k}}{N}$, $\mathcal{N}_i$ is the K nearest neighbor set of instance i , $\tilde{\mathbf{A}}_x$ is the adjacent matrix, $\tilde{a}_{i,j} = 1$ when $j \in \mathcal{N}_i$, otherwise $\tilde{a}_{i,j} = 0$. λ is the hyper-parameter. The second term $\mathcal{H}(\mathbf{P})$ in Eq. (4) denotes balanced constrain of cluster. Compared with Eq. (3), Eq. (2) transforms the partitioning problem in Euclidean space into the graph-cut problem. And different from balanced partitioning in Eq. (4), we convert the maximum problem to the minimum problem to efficiently avoid trivial solutions. The intrinsical constraints of probability matrix $\mathbf{P}$ enable J_{pac} directly clustering without using orthogonality and balanced constraints. Therefore, DPAC does not require additional clustering regular terms [35, 46, 50] to avoid collapse and performs more flexible cluster assignment. Moreover, unlike only using neighbors to group, J_{pac} introduces the distance information into optimization to obtain a global clustering.

3.3 Probability Aggregation Clustering

The proposed Eq. (2) is a constrained optimization problem. Inspired by FCM, we incorporate the fuzzy weighting exponent m into the objective function and obtain a scalable machine clustering algorithm based on the Lagrange method. The new objective function with m can be formulated as:

$$\tilde{J}_{pac} = \sum_{i=1}^N \sum_{j=1}^N \varphi(i, j) \tilde{d}_{i,j}, \quad \text{with } \varphi(i, j) = \sum_{k=1}^K p_{i,k}^m p_{j,k}, \quad (5)$$

where $m \in (1, +\infty)$. The corresponding Lagrange function is:

$$\tilde{L}_{pac} = \sum_{i=1}^N \sum_{j \neq i} \varphi(i, j) \tilde{d}_{i,j} + \sum_{i=1}^N \lambda_i (1 - \sum_{k=1}^K p_{i,k}) - \sum_{i=1}^N \sum_{k=1}^K \gamma_{i,k} p_{i,k}, \quad (6)$$

where λ_i and $\gamma_{i,k}$ are the Lagrange multipliers respectively for the sum constraint and the non-negativity constraint on $\mathbf{P}$. The partial derivative of $\tilde{L}_{pac}$ with respect to $p_{i,k}$ should be equal to zero at the minimum as:

$$\frac{\partial \tilde{L}_{pac}}{\partial p_{i,k}} = 2 \sum_{j \neq i} m p_{i,k}^{m-1} p_{j,k} \tilde{d}_{i,j} - \lambda_i - \gamma_{i,k} = 0. \quad (7)$$

And according to the Karush-Kuhn-Tucker conditions we have: $1 - \sum_{k=1}^K p_{i,k} = 0$, $\gamma_{i,k} p_{i,k} = 0$, $\gamma_{i,k} \geq 0$, $\forall i, k$. For soft clustering, endpoints are generally unreachable during optimization. Therefore, we only consider the case when $p_{i,k} \in (0, 1)$, $\gamma_{i,k} = 0$. Let $\alpha = 1/(m-1)$, it can be obtained from Eq. (7) that $p_{i,k} = \lambda_i^\alpha (2m \sum_{j \neq i} p_{j,k} \tilde{d}_{i,j})^{-\alpha}$. Considering the sum constraint, the equation becomes $\lambda_i^\alpha \sum_{k=1}^K (2m \sum_{j \neq i} p_{j,k} \tilde{d}_{i,j})^{-\alpha} = \sum_{k=1}^K p_{i,k} = 1$. By solving λ_i and taking it into Eq. (7), we can finally obtain:

$$p_{i,k} = \frac{s_{i,k}^{-\alpha}}{\sum_{r=1}^K s_{i,r}^{-\alpha}}, \quad \text{with } s_{i,k} = \sum_{j \neq i} p_{j,k} \tilde{d}_{i,j}. \quad (8)$$

Take one element $p_{i,k}$ as a variable and all the rest elements as constant, $\mathbf{P}$ can be iteratively updated with Eq. (8). $s_{i,k}$ aggregates the probabilities and distances to compute a score that $\mathbf{x}_i$ belongs to cluster k . In other words, PAC solves $p_{i,k}$ through all other instances instead of cluster centers. PAC only needs to initialize the $\mathbf{P}$ following approximately uniform distribution, that is $p_{i,k} \approx 1/K$. Therefore, PAC circumvents the delicate cluster center initialization problem caused by disparate data distributions in the feature space [4]. The detailed steps of PAC are summarized in Algorithm 1.

Algorithm 1: PAC Program

```

1 Input: dataset  $\mathbf{X}$ ; weighting exponent  $m$ ; cluster number  $K$ ; initialization  $\mathbf{P}$ .
2 while not convergence do
3   for  $i \leftarrow 1$  to  $N$  do
4     for  $k \leftarrow 1$  to  $K$  do
5        $p_{i,k} \leftarrow$  Eq. (8)
6     end
7   end
8 end
9 Output: Clustering result  $\mathbf{P}$ 

```

3.4 Online Probability Aggregation

A deep neural network $\hat{\mathbf{x}}_i = f(\mathbf{I}_i)$ maps data $\mathbf{I}_i$ to feature vector $\hat{\mathbf{x}}_i$. And a classifier h maps $\mathbf{x}_i$ to K-dimensional class probability $\hat{\mathbf{p}}_i$. We proposed a novel online clustering module OPA, which combines the optimization process of PAC with loss computation to generate pseudo labels step by step. Specifically, B is the size of the mini-batch in the current epoch, OPA has two alternate steps:

Target Computation: Sec. 3.3 demonstrates the optimization program for a single variable, we extend it to the matrix to adopt multivariable. Given the current model $h \circ f$, the clustering score $\mathbf{S} \in \mathbb{R}^{+B \times K}$ is calculated by:

$$\mathbf{S} = \tilde{\mathbf{D}}_{\hat{\mathbf{x}}} \hat{\mathbf{P}}. \quad (9)$$

The target clustering code $\mathbf{Q} \in \Gamma^{B \times K}$ can be obtained by normalizing $\mathbf{S}$, $q_{i,k} = s_{i,k}^{-\alpha} / \sum_{r=1}^K s_{i,r}^{-\alpha}$. We call the operation in Eq. (9) as online probability aggregation. The probability outputs from the classifier are aggregated by matrix multiplication to compute corresponding scores, which not only incorporates historical partitioning knowledge but also encodes distance information.

Self-labeling: Given the current target clustering code $\mathbf{Q}$, the whole model $h \circ f$ is updated by minimizing the following KL divergence:

$$KL(\mathbf{Q} \parallel \hat{\mathbf{P}}) = \sum_{i=1}^N \sum_{k=1}^K q_{i,k} \log \frac{q_{i,k}}{\hat{p}_{i,k}} \quad (10)$$

Different from directly leveraging J_{pac} in Eq. (1) as clustering loss, OPA trains the model in a supervised way instead of solving the clustering problem in Eq. (2) exactly. The pseudo code of OPA is illustrated in Algorithm 2, which only involves a mini-batch matrix multiplication and power, so the computation cost of OPA equals general loss.

Algorithm 2: Pseudo code for OPA in pytorch-style

```

1 Input: distance matrix  $D$ ; probability matrix  $P$ ; weighting exponent  $m$ .
2    $S = \text{torch.matmul}(D.\text{detach}(), P)$  // Aggregate Probability
3    $S = \text{torch.pow}(S, -1/(m-1))$  // Scale Up
4    $Q = S/S.\text{sum}(1).\text{view}(-1,1)$  // Normalize to 1
5 Output:  $(Q * \log Q - Q * \log P).\text{sum}(1).\text{mean}()$  // KL divergence loss
```

3.5 Deep Probability Aggregation Clustering

With the proposed loss function, we construct an online deep clustering framework DPAC, which has two heads: contrastive learning and online clustering. Let

$\hat{\mathbf{I}}_i^1$ and $\hat{\mathbf{I}}_i^2$ denote two-view features of $\hat{\mathbf{I}}_i$ generated by random image augmentation. We reformulate the standard contrastive loss in SimCLR [13] as weight contrastive loss (WCL) to mitigate the semantic distortion caused by negative samples. The weight contrastive loss $\ell(\hat{\mathbf{X}}^1, \hat{\mathbf{X}}^2, \hat{\mathbf{P}})$ is defined as:

$$-\sum_{i=1}^N \log \frac{\exp(\hat{\mathbf{z}}_i^{1\top} \hat{\mathbf{z}}_i^2 / \tau)}{\sum_{j \neq i} \hat{w}_{i,j} \exp(\hat{\mathbf{z}}_i^{1\top} \hat{\mathbf{z}}_j^1 / \tau) + \sum_{j=1}^N \hat{w}_{i,j} \exp(\hat{\mathbf{z}}_i^{1\top} \hat{\mathbf{z}}_j^2 / \tau)}, \quad (11)$$

where τ is the temperature hyper-parameter, $\hat{\mathbf{z}}_i$ is the normalized feature projected by projector g , where $\hat{\mathbf{z}}_i = g(\hat{\mathbf{x}}_i) / \|g(\hat{\mathbf{x}}_i)\|$. $\hat{w}_{i,j} = (1 - \hat{\mathbf{p}}_i^\top \hat{\mathbf{p}}_j)$ is a gate coefficient, which filters the negative samples that belong to same cluster as $\hat{\mathbf{x}}_i$.

In pre-training step, due to the lack of cluster information, $\hat{\mathbf{P}}$ is set to the uniform, $\hat{p}_{i,j} = 1/K, \forall i, j$. And DPAC is pre-trained by the pairwise contrastive loss: $\frac{1}{2}[\ell(\hat{\mathbf{X}}^1, \hat{\mathbf{X}}^2, \hat{\mathbf{P}}) + \ell(\hat{\mathbf{X}}^2, \hat{\mathbf{X}}^1, \hat{\mathbf{P}})]$. Then in clustering step, the whole model is updated by minimizing the sum of contrastive and clustering loss:

$$\min_{\theta_{f,g,h}} \frac{1}{2}[\ell(\hat{\mathbf{X}}^1, \hat{\mathbf{X}}^2, \hat{\mathbf{P}}) + \ell(\hat{\mathbf{X}}^2, \hat{\mathbf{X}}^1, \hat{\mathbf{P}})] + KL(\mathbf{Q} \parallel \hat{\mathbf{P}}^1), \quad (12)$$

$$\min_{\theta_{f,g,h}} \frac{1}{2}[\ell(\hat{\mathbf{X}}^1, \hat{\mathbf{X}}^2, \hat{\mathbf{P}}) + \ell(\hat{\mathbf{X}}^2, \hat{\mathbf{X}}^1, \hat{\mathbf{P}})] + \frac{1}{N} Tr(\hat{\mathbf{P}}^{1\top} \tilde{\mathbf{D}}_{\hat{\mathbf{x}}} \hat{\mathbf{P}}^1), \quad (13)$$

where $\theta_{f,g,h}$ are the parameters of the neural network, classifier, and projector, respectively. Eq. (12) is the deep clustering method based on OPA mentioned in Sec. 3.4. Eq. (13) is the deep clustering method that directly minimizes J_{pac} in Sec. 3.1. The overall training procedure is shown in Algorithm 3. Moreover, for fair comparison in subsequent experiments, we also implement a self-labeling fine-tuning operation as [36, 47] to further improve the clustering performance.

Algorithm 3: Training algorithm for DPAC

```

1 Input: image set  $\mathbf{I}$ ; clustering epochs  $E$ ; batch size  $B$ ; weighting exponent  $m$ .
2 for  $epoch \leftarrow 1$  to  $E$  do
3   Sample a mini-batch  $\{\mathbf{I}_i\}_{i=1}^B$  and conduct augmentations  $\{\mathbf{I}_i^1, \mathbf{I}_i^2\}_{i=1}^B$ ;
4   Get  $\{\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_i^1, \hat{\mathbf{x}}_i^2, \hat{\mathbf{p}}_i, \hat{\mathbf{p}}_i^1\}_{i=1}^B$  through forward propagation;
5   if choose OPA as optimal object then
6     Compute clustering codes  $\{\mathbf{q}_i\}_{i=1}^B$  by Algorithm 2 with  $\{\hat{\mathbf{x}}_i, \hat{\mathbf{p}}_i\}_{i=1}^B$ ;
7     Compute overall loss  $\mathcal{L}$  by Eq. (12) with  $\{\hat{\mathbf{x}}_i^1, \hat{\mathbf{x}}_i^2, \hat{\mathbf{p}}_i, \hat{\mathbf{p}}_i^1, \mathbf{q}_i\}_{i=1}^B$ ;
8   end
9   if choose  $J_{pac}$  as optimal object then
10    Compute overall loss  $\mathcal{L}$  by Eq. (13) with  $\{\hat{\mathbf{x}}_i^1, \hat{\mathbf{x}}_i^2, \hat{\mathbf{p}}_i, \hat{\mathbf{p}}_i^1\}_{i=1}^B$ ;
11  end
12  Update  $\theta_f, \theta_g, \theta_h$  through gradient descent to minimize  $\mathcal{L}$ ;
13 end
14 Output: Deep clustering model  $h \circ f$ 

```

Table 1: Dataset settings for our experiments.

Dataset	Sample	Class	Size	Dataset	Sample	Class	Dimension
CIFAR-10 [32]	60,000	10	32×32	Coil-100 [39]	7,200	100	49,152
CIFAR-100 [32]	60,000	20	32×32	Isolet [16]	7,796	26	617
STL-10 [15]	13,000	10	224×224	Pendigits [2]	10,992	10	16
ImageNet-10 [12]	13,000	10	224×224	MNIST [19]	10,000	10	784
ImageNet-Dogs [12]	19,500	15	224×224				

4 Experiment

Dataset: Four real-world datasets and five widely used natural image datasets are involved to evaluate the clustering ability of PAC and DPAC. The details of the datasets are summarized in the Tab. 1. For CIFAR-100, we used its 20 super-classes rather than 100 classes as the ground truth. For STL-10, its 100,000 unlabeled images are additionally used in the pre-training step of DPAC. ImageNet-10 and ImageNet-Dogs are subsets of ImageNet-1k. Clustering accuracy (ACC), normalized mutual information (NMI), and adjusted random index (ARI) are adopted to compare the clustering results.

4.1 Probability Aggregation Clustering

Hyperparameter and Method Setting The effectiveness of the proposed PAC is verified by comparing it with multiple clustering methods on nine datasets. The m of PAC is set to 1.03 for all datasets. The threshold value of RCC [45] is set to 1. The weighting exponent m of FCM is set to 1.1 for real-world datasets and 1.05 for natural image datasets. We predefine K for all algorithms except FINCH [44]. All algorithms are initialized randomly and run 10 times. The mean and variance of 10 times run are taken as comparison results.

Algorithm Scalability The clustering results of the real-world datasets, which consist of samples with varying numbers, classes, and dimensions, are summarized in Tab. 2. PAM and RCC time out due to the high dimensionality of Coil-100. PAC outperforms all the compared clustering algorithms on Coil-100 and Isolet but is not as effective as RCC on Mnist and Pendigit, which is specially designed for entangled data. The robustness and performance of PAC surpass center-based methods by a large margin. Moreover, we also provide the clustering results on neural network feature data in Tab. 3 to explore the ability of PAC to handle data extracted by neural networks. RCC experience extreme performance degradation on neural network extracted data, so we exclude it from the comparison. PAC also performs well in processing neural network data. The improvement is not significant in CIFAR-100 and ImageNet-Dogs. One possible explanation is that these datasets give subtle differences in object classes, causing the pretrained representations to be indistinguishable.

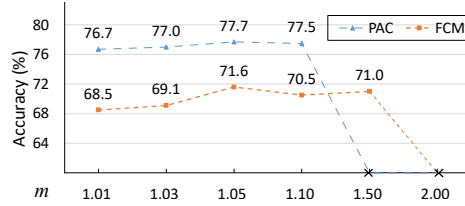
Table 2: Clustering results (Avg $\pm$ Std) and average time (s) of PAC on real-world datasets. The best and second-best results are shown in bold and underlined, respectively. Metric: ACC (%).

Method	Coil-100	Isolet	Pendigits	MNIST	Average Time			
KM [53]	56.4 $\pm$ 1.7	52.7 $\pm$ 4.5	67.0 $\pm$ 4.7	53.0 $\pm$ 3.6	98.1	0.2	0.05	0.07
PAM [33]	N/A	55.5 $\pm$ 0.0	75.6 $\pm$ 2.5	47.2 $\pm$ 1.7	N/A	341.9	141.6	124.0
FCM [6]	61.6 $\pm$ 1.2	<u>55.8</u> $\pm$ 2.3	70.5 $\pm$ 2.1	56.6 $\pm$ 2.6	2001.5	8.6	0.9	0.6
SC [51]	58.2 $\pm$ 0.7	53.5 $\pm$ 2.5	62.4 $\pm$ 4.2	54.6 $\pm$ 2.2	11.7	3.4	5.8	6.2
SPKF [27]	59.7 $\pm$ 1.3	55.2 $\pm$ 2.0	71.4 $\pm$ 4.4	53.9 $\pm$ 2.7	101.6	0.6	0.07	0.2
RCC [45]	N/A	15.3 $\pm$ 0.0	79.6 $\pm$ 0.0	65.7 $\pm$ 0.0	N/A	122.8	6.9	6.9
FINCH [44]	56.4 $\pm$ 0.0	47.5 $\pm$ 0.0	62.7 $\pm$ 0.0	57.9 $\pm$ 0.0	15.1	0.5	0.05	0.05
PAC	65.1 $\pm$ 1.5	61.8 $\pm$ 0.0	<u>78.0</u> $\pm$ 0.0	<u>59.7</u> $\pm$ 3.6	5179.0	249.6	153.6	423.4

Table 3: Clustering results (Avg $\pm$ Std) of PAC on deep features. Metric: ACC (%).

Method	CIFAR-10	CIFAR-100	STL-10	ImageNet-10	ImageNet-Dogs
KM [53]	76.8 $\pm$ 6.8	41.8 $\pm$ 1.7	66.8 $\pm$ 4.3	76.8 $\pm$ 6.8	41.8 $\pm$ 1.7
PAM [33]	77.8 $\pm$ 2.5	41.0 $\pm$ 1.1	64.3 $\pm$ 4.8	79.9 $\pm$ 4.6	52.6 $\pm$ 3.1
FCM [6]	75.9 $\pm$ 2.1	42.3 $\pm$ 0.7	66.6 $\pm$ 4.7	75.9 $\pm$ 2.1	42.3 $\pm$ 0.7
SC [51]	83.5 $\pm$ 0.0	40.0 $\pm$ 1.1	63.8 $\pm$ 2.9	<u>82.9</u> $\pm$ 1.3	47.6 $\pm$ 1.4
SPKF [27]	75.9 $\pm$ 5.7	<u>42.9</u> $\pm$ 1.9	65.8 $\pm$ 5.5	80.6 $\pm$ 7.6	<u>49.1</u> $\pm$ 3.8
FINCH [44]	49.2 $\pm$ 0.0	32.0 $\pm$ 0.0	42.9 $\pm$ 0.0	52.6 $\pm$ 0.0	43.8 $\pm$ 0.0
PAC	87.1 $\pm$ 0.0	43.8 $\pm$ 0.7	74.9 $\pm$ 2.6	95.8 $\pm$ 0.0	47.3 $\pm$ 3.9

Parameter Sensibility Analysis We evaluate the parameter sensitivity of m for both FCM and PAC on Pendigits. Fig. 1 reports the average ACC for different m . It was indicated that in comparison to FCM, PAC has a narrower optimal range of m and smaller results variance, which is not sensitive to parameter m .

**Fig. 1:** The effect of weighting exponent m in PAC and FCM.

Time Complexity Analysis The average calculation time for each algorithm is listed in Tab. 2. The computational complexity of PAC is analyzed in this section. It takes $\mathcal{O}(N)$ time to calculate $\sum_{j \neq i} p_{j,k} \tilde{d}_{i,j}$ in Eq. (8). And PAC updates entire $\mathbf{P}$ by NK iterations. So the time complexity PAC is $\mathcal{O}(N^2K)$, which is the square complexity.

Table 4: Performance comparison of deep clustering methods on five benchmarks. The best and second-best results are shown in bold and underlined, respectively. Metrics: NMI / ACC / ARI (%). Temu* incorporates extra ImageNet-1k data to pretrain the model, so we exclude it in comparison. ₁ denotes online deep clustering methods, while ₂ denotes offline deep clustering methods. Cluster const. denotes cluster size constraint.

Method	Cluster	CIFAR-10			CIFAR-100			STL-10			ImageNet-10			ImageNet-Dogs		
	const.	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
PICA ₁ [28]	✓	59.1	69.6	51.2	31.0	33.7	17.1	61.1	71.3	53.1	80.2	87.0	76.1	35.2	35.2	20.1
PCL ₂ [34]		80.2	87.4	76.6	52.8	52.6	36.3	41.0	71.8	67.0	84.1	90.7	82.2	44.0	41.2	29.9
IDFD ₂ [48]		71.1	81.5	66.3	42.6	42.5	26.4	64.3	75.6	57.5	<u>89.8</u>	<u>95.4</u>	<u>90.1</u>	54.6	59.1	41.3
NNM ₁ [18]	✓	74.8	84.3	70.9	48.4	47.7	31.6	69.4	80.8	65.0	-	-	-	-	-	-
CC ₁ [35]	✓	70.5	79.0	63.7	43.1	42.9	26.6	76.4	85.0	72.6	85.9	89.3	82.2	44.5	42.9	27.4
GCC ₁ [58]	✓	76.4	85.6	72.8	47.2	47.2	30.5	68.4	78.8	63.1	84.2	90.1	82.2	49.0	52.6	36.2
TCC ₁ [46]	✓	79.0	<u>90.6</u>	73.3	47.9	49.1	31.2	73.2	81.4	68.9	84.8	89.7	82.5	55.4	59.5	41.7
SPICE ₁ [40]	✓	73.4	83.8	70.5	44.8	46.8	29.4	<u>81.7</u>	<u>90.8</u>	<u>81.2</u>	82.8	92.1	83.6	<u>57.2</u>	<u>64.6</u>	<u>47.9</u>
SeCu ₁ [41]	✓	79.9	88.5	78.2	<u>51.6</u>	<u>51.6</u>	<u>36.0</u>	70.7	81.4	65.7	-	-	-	-	-	-
Temu ₂ * [1]	✓	82.9	90.0	80.7	59.8	57.8	42.5	93.6	96.7	93.0	-	-	-	-	-	-
DPAC ₁ ^{pac} (Eq. (13))		<u>81.2</u>	89.0	<u>79.1</u>	48.3	50.2	34.4	81.8	89.7	80.0	<u>90.1</u>	<u>96.0</u>	<u>91.1</u>	51.9	53.9	38.9
DPAC ₁ ^{opa} (Eq. (12))		82.7	90.7	81.2	52.9	51.6	36.2	84.5	92.6	84.7	90.8	96.2	91.8	60.2	65.5	50.0
With self-labeling fine-tuning (†):																
SCAN ₂ † [50]	✓	79.7	88.3	77.2	48.6	50.7	33.3	69.8	80.9	64.6	-	-	-	-	-	-
SPICE ₁ † [40]	✓	<u>86.5</u>	92.6	85.2	56.7	53.8	38.7	87.2	93.8	87.0	<u>90.2</u>	<u>95.9</u>	<u>91.2</u>	<u>62.7</u>	<u>67.5</u>	52.6
TCL ₁ † [36]	✓	81.9	88.7	78.0	52.9	53.1	35.7	79.9	86.8	75.7	87.5	89.5	83.7	62.3	64.4	51.6
SeCu ₁ † [41]	✓	86.1	<u>93.0</u>	<u>85.7</u>	55.2	<u>55.1</u>	39.7	73.3	83.6	69.3	-	-	-	-	-	-
DPAC ₁ ^{opa} †		87.0	93.4	86.6	<u>54.2</u>	55.5	<u>39.3</u>	<u>86.3</u>	<u>93.4</u>	<u>86.1</u>	92.5	97.0	93.5	66.7	72.6	59.8

4.2 Deep Probability Aggregation Clustering

Implementation Details ResNet-34 [26] is used as the backbone network in DPAC to ensure a fair comparison. We employed the architecture of SimCLR [14] with an MLP clustering classifier as model architecture. DPAC incorporates the image transformation of SimCLR as one view of augmentation and randomly selects four transformations from Rand Augment [17] as another view of augmentation. We maintain a consistent set of hyperparameters ($m = 1.03, \tau = 0.5$) across all amounts of benchmarks. The model is trained for 1,000 epochs in the pre-training step and 200 epochs in the clustering step. As for self-labeling fine-tuning, we utilize a linear classifier and train the model as [36]. The thresholds are set to 0.95 for each dataset to select sufficient pseudo labels from clustering classifier outputs. Adam [31] with a constant learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} was employed. The batch size is set as 240 and the experiments are implemented on a single NVIDIA 4090 24G GPU.

Comparison with State of the Arts The comparison of DPAC is presented in Tab. 4, where methods with additional cluster size constraints are marked. We have the following observations: (1) DPAC significantly surpasses the performance of SimCLR+PAC in Tab. 3 across all benchmarks. The accuracy of DPAC exceeds PAC by more than 10% on CIFAR-100, STL-10, and ImageNet-Dogs benchmarks, which demonstrates the semantic learning ability of DPAC.

Table 5: Further analysis for DPAC.

(a) Comparison of different contrastive framework on CIFAR-10.		(b) Comparison of AE based clustering methods on MNIST.		(c) Effect of clustering regularization (CR) term on CIFAR-10. Metric: ACC (%).		
Method	ACC	Method	NMI ACC	Method	w/ CR	w/o CR
SimCLR+OPA	89.7	DEC [55]	86.7 88.1	SCAN [50]	85.7	0.1
MoCo+OPA	86.5	IDEC [24]	86.7 88.1	CC [35]	79.2	68.7
DPAC ^{opa}	90.8	EDESC [8]	86.2 91.3	GCC [58]	85.6	68.0
		SSC [52]	95.0 98.2	DPAC ^{Jpac}	88.0	89.0
		AE+OPA	90.3 95.4	DPAC ^{opa}	0.1	90.7

(2) Compared with DPAC^{Jpac}, DPAC^{opa} has better performance. We attribute this to the fact that the self-labeling manner of OPA alleviates the intrinsic bias brought by the objective function of feature clustering. (3) Compared with deep clustering methods with offline K-means, such as IDFD [48] and PCL [34], DPAC has superior performance on all benchmarks due to the stable learning offered by the online manner. (4) Compared with online contrastive clustering methods CC [35], TCC [46], and TCL [36], DPAC incorporates global spatial information to achieve a fine-grained partitioning of cluster boundaries. (5) Compared with balanced clustering methods and minimal cluster size constraint SeCU [36], DPAC omits clustering regular term, is more concise, and outputs more flexible cluster assignments. (6) DPAC^{opa}† demonstrates the remarkable extensibility of our approach, showcasing the potential for integration with diverse deep modules.

Contrastive Framework Analysis We further analyze our DPAC model from different perspectives. We study the effect of proposed contrastive learning. We replace the weighted contrastive loss in Eq. (13) with standard contrastive loss, and denote it as SimCLR+OPA. Besides, we also perform OPA based on MoCo [25]. Conventional contrastive loss treats corresponding augmented samples as positive pairs and others as negative pairs, which ignores the latent semantic structure between negative pairs, leading to the class collision issue [54]. Tab. 5a illustrates our weighted contrastive loss alleviates the cluster collision problem and encodes cluster knowledge into contrastive representation learning.

Pretext Task Analysis We study the effect of different pretext tasks combined with DPAC. The autoencoder (AE) is used as architecture to prove the universality of our module. The clustering results on MNIST are shown in Tab. 5b, which demonstrates that OPA can combine with other self-supervised approaches. Especially, compared with center-based IDEC [55] and SSC [52], our OPA does not require K-means to initialize cluster layer and has higher scalability.

Balanced Constraint Analysis We study the impact of balanced constraints in different deep clustering methods. Most existing online deep clustering meth-

Table 6: Hyperparameter analysis of exponent m in OPA. Metrics:ACC (%).

Weight exponent m	1.01	1.04	1.07	1.1	1.13	1.16	1.2
$\alpha = 1/(m - 1)$	100.0	25.0	14.3	10.0	7.7	6.3	5.0
CIFAR-10	90.8	90.3	90.1	89.3	89.3	89.3	10.0
STL-10	92.4	92.1	92.0	91.8	90.7	10.0	10.0
CIFAR-100	50.2	51.1	51.0	5.0	5.0	5.0	5.0

ods [35, 46, 58] introduce an average entropy as clustering regularization (CR) term to balance the cluster distribution. The clustering regularization experiments are shown in Tab. 5c. SCAN classifies all samples into a single cluster, and CC and GCC descend into a suboptimal solution without (w/o) the CR term. Besides, if the CR term is too large in the total loss, it will affect the clustering performance in these methods. It is noteworthy that DPAC avoids crashes without the CR term. The performance of DPAC^{*J_{pac}*} with (w/) CR term becomes worse. It demonstrates the superiority of unconstrained clustering, that is, no trade-off between trivial solutions and performance. And DPAC^{*opa*} with CR term yields a uniform distribution with no predictive effect. The reason that the constraint of the CR term is too strong so that the classifier cannot accumulate optimization enough information for OPA.

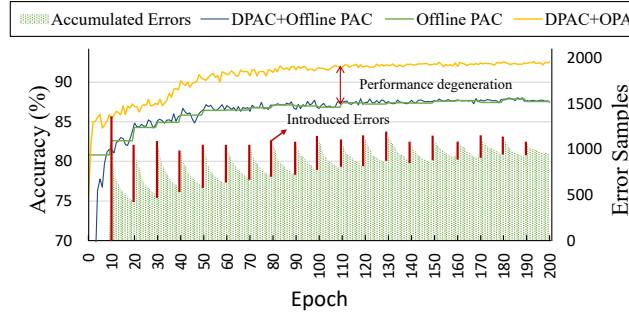
Hyperparameter Analysis As listed in Algorithm 2, weight exponent m is the key hyperparameter for OPA, $\alpha = 1/(m - 1)$ is the power of $\mathbf{s}_{i,k}$ that amplifies clustering score in Eq. (9) to become sharper to obtain distinguishable cluster assignments. The larger m becomes, the smaller the sharpening effect is, so the model tends to uniform assignments, and clustering may fail due to insufficient scaling. The performance of OPA with different m settings is evaluated in Tab. 6. As features become more and more inseparable, the optimal range of m narrows. Therefore, we suggest setting m close to 1 to obtain a universal hyperparameter setting ($m = 1.03$ for all datasets).

Superiority of Online Clustering We perform the offline clustering version of DPAC to facilitate a comparative analysis between online and offline clustering strategies. We adopt KM, FCM, and PAC to compute offline codes of all samples for Eq. (10) every 1, 10, and 200 epochs. The performance and training duration are reported in Tab. 7. It can be observed that the performance of KM and FCM gradually deteriorates as the update frequency decreases, whereas DPAC^{*opa*} exhibits superior performance and lower time complexity.

We recorded accumulated errors during DPAC + offline PAC training progress to analyze the error accumulation issue. Offline PAC was conducted every 10 epochs. As depicted in Fig. 2, errors (network classifies correctly while offline clustering classifies incorrectly) are introduced by offline clustering every 10 epochs and continue to accumulate through the training process. It demonstrates our OPA module effectively mitigates performance degeneration and error accumulation issues to perform stable and efficient clustering.

Table 7: The comparison of online and offline DPAC on STL-10. Metrics: Hour/ACC (%).

Method	Number of Offline Clustering Runs		
	200	20	1
DPAC + offline KM	6.3 / 73.7	3.0 / 72.0	2.0 / 69.3
DPAC + offline FCM	8.5 / 78.7	3.3 / 77.5	2.0 / 68.4
DPAC + offline PAC	52.2 / 83.7	6.7 / 87.5	2.4 / 81.3
DPAC ^{opa}	2.0 / 92.6		

**Fig. 2:** Training process and error accumulation of online and offline DPAC on STL-10.

5 Conclusion

A novel machine clustering method PAC without cluster center was proposed from a very new perspective, which addresses the shortcomings of center-based clustering approaches and is well-suited for integration with deep models. A theoretical model and an elegant iterative optimization solution for PAC have been developed. PAC implements clustering through sample probability aggregation, which makes part samples based calculation possible. Therefore, an online deep clustering framework DPAC has been developed, which has no constraints on cluster size and can perform more flexible clustering. Experiments on several benchmarks verified the effectiveness of our proposal.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. U22B2036).

References

1. Adaloglou, N., Michels, F., Kalisch, H., Kollmann, M.: Exploring the limits of deep image clustering using pretrained models. arXiv preprint arXiv:2303.17896 (2023)
2. Alimoglu, F., Alpaydin, E.: Combining multiple representations and classifiers for pen-based handwritten digit recognition. In: Proceedings of the Fourth International Conference on Document Analysis and Recognition. vol. 2, pp. 637–640. IEEE (1997)
3. Arabie, P., Hubert, J., Soete, D.: Complexity theory: An introduction. Clustering and classification p. 199 (1996)
4. Arthur, D., Vassilvitskii, S., et al.: k-means++: The advantages of careful seeding. In: Soda. vol. 7, pp. 1027–1035 (2007)
5. Asano, Y.M., Rupprecht, C., Vedaldi, A.: Self-labelling via simultaneous clustering and representation learning. arXiv preprint arXiv:1911.05371 (2019)
6. Bezdek, J.C., Ehrlich, R., Full, W.: Fcm: The fuzzy c-means clustering algorithm. Computers & geosciences **10**(2-3), 191–203 (1984)
7. Bradley, P.S., Bennett, K.P., Demiriz, A.: Constrained k-means clustering. Microsoft Research, Redmond **20**(0), 0 (2000)
8. Cai, J., Fan, J., Guo, W., Wang, S., Zhang, Y., Zhang, Z.: Efficient deep embedded subspace clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–10 (2022)
9. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
10. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
11. Celebi, M.E.: Partitional clustering algorithms. Springer (2014)
12. Chang, J., Wang, L., Meng, G., Xiang, S., Pan, C.: Deep adaptive image clustering. In: Proceedings of the IEEE international conference on computer vision. pp. 5879–5887 (2017)
13. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
14. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems **33**, 22243–22255 (2020)
15. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223. JMLR Workshop and Conference Proceedings (2011)
16. Cole, R., Muthusamy, Y., Fanty, M.: The ISOLET spoken letter database. Oregon Graduate Institute of Science and Technology, Department of Computer ... (1990)

17. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
18. Dang, Z., Deng, C., Yang, X., Wei, K., Huang, H.: Nearest neighbor matching for deep clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13693–13702 (2021)
19. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine* **29**(6), 141–142 (2012)
20. Deshmukh, A.A., Regatti, J.R., Manavoglu, E., Dogan, U.: Representation learning for clustering via building consensus. *Machine Learning* **111**(12), 4601–4638 (2022)
21. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: *kdd*. vol. 96, pp. 226–231 (1996)
22. Fränti, P., Sieranoja, S.: How much can k-means be improved by using better initialization and repeats? *Pattern Recognition* **93**, 95–112 (2019)
23. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. *science* **315**(5814), 972–976 (2007)
24. Guo, X., Gao, L., Liu, X., Yin, J.: Improved deep embedded clustering with local structure preservation. In: *Ijcai*. vol. 17, pp. 1753–1759 (2017)
25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
27. Hornik, K., Feinerer, I., Kober, M., Buchta, C.: Spherical k-means clustering. *Journal of statistical software* **50**, 1–22 (2012)
28. Huang, J., Gong, S., Zhu, X.: Deep semantic clustering by partition confidence maximisation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8849–8858 (2020)
29. Huang, Z., Chen, J., Zhang, J., Shan, H.: Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(6), 7509–7524 (2022)
30. Jiao, Y., Xie, N., Gao, Y., Wang, C.C., Sun, Y.: Fine-grained fashion representation learning by online deep clustering. In: *European Conference on Computer Vision*. pp. 19–35. Springer (2022)
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
32. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
33. Van der Laan, M., Pollard, K., Bryan, J.: A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation* **73**(8), 575–584 (2003)
34. Li, J., Zhou, P., Xiong, C., Hoi, S.C.: Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966* (2020)
35. Li, Y., Hu, P., Liu, Z., Peng, D., Zhou, J.T., Peng, X.: Contrastive clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 8547–8555 (2021)
36. Li, Y., Yang, M., Peng, D., Li, T., Huang, J., Peng, X.: Twin contrastive learning for online clustering. *International Journal of Computer Vision* **130**(9), 2205–2221 (2022)

37. Lin, W.C., Ke, S.W., Tsai, C.F.: Cann: An intrusion detection system based on combining cluster centers and nearest neighbors. *Knowledge-based systems* **78**, 13–21 (2015)
38. Nassar, I., Hayat, M., Abbasnejad, E., Rezaatofghi, H., Haffari, G.: Protocon: Pseudo-label refinement via online clustering and prototypical consistency for efficient semi-supervised learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11641–11650 (2023)
39. Nene, S.A., Nayar, S.K., Murase, H., et al.: *Columbia object image library (coil-20)* (1996)
40. Niu, C., Shan, H., Wang, G.: Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing* **31**, 7264–7278 (2022)
41. Qian, Q.: Stable cluster discrimination for deep clustering. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 16645–16654 (2023)
42. Qian, Q., Xu, Y., Hu, J., Li, H., Jin, R.: Unsupervised visual representation learning by online constrained k-means. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16640–16649 (2022)
43. Ronen, M., Finder, S.E., Freifeld, O.: Deepdpm: Deep clustering with an unknown number of clusters. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9861–9870 (2022)
44. Sarfraz, S., Sharma, V., Stiefelhagen, R.: Efficient parameter-free clustering using first neighbor relations. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8934–8943 (2019)
45. Shah, S.A., Koltun, V.: Robust continuous clustering. *Proceedings of the National Academy of Sciences* **114**(37), 9814–9819 (2017)
46. Shen, Y., Shen, Z., Wang, M., Qin, J., Torr, P., Shao, L.: You never cluster alone. *Advances in Neural Information Processing Systems* **34**, 27734–27746 (2021)
47. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* **33**, 596–608 (2020)
48. Tao, Y., Takagi, K., Nakata, K.: Clustering-friendly representation learning via instance discrimination and feature decorrelation. *arXiv preprint arXiv:2106.00131* (2021)
49. Tsai, T.W., Li, C., Zhu, J.: Mice: Mixture of contrastive experts for unsupervised image clustering. In: *International conference on learning representations* (2021)
50. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: Scan: Learning to classify images without labels. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X*. pp. 268–285. Springer (2020)
51. Von Luxburg, U.: A tutorial on spectral clustering. *Statistics and computing* **17**, 395–416 (2007)
52. Wang, H., Lu, N., Luo, H., Liu, Q.: Self-supervised clustering with assistance from off-the-shelf classifier. *Pattern Recognition* **138**, 109350 (2023)
53. Wang, J., Wang, J., Song, J., Xu, X.S., Shen, H.T., Li, S.: Optimized cartesian k-means. *IEEE Transactions on Knowledge and Data Engineering* **27**(1), 180–192 (2014)
54. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: *International Conference on Machine Learning*. pp. 9929–9939. PMLR (2020)

- 55. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International conference on machine learning. pp. 478–487. PMLR (2016)
- 56. Yang, B., Fu, X., Sidiropoulos, N.D., Hong, M.: Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In: international conference on machine learning. pp. 3861–3870. PMLR (2017)
- 57. Zhan, X., Xie, J., Liu, Z., Ong, Y.S., Loy, C.C.: Online deep clustering for unsupervised representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6688–6697 (2020)
- 58. Zhong, H., Wu, J., Chen, C., Huang, J., Deng, M., Nie, L., Lin, Z., Hua, X.S.: Graph contrastive clustering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9224–9233 (2021)
- 59. Zhong, S.: Efficient online spherical k-means clustering. In: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005. vol. 5, pp. 3180–3185. IEEE (2005)
- 60. Znalezniak, M., Rola, P., Kaszuba, P., Tabor, J., Śmieja, M.: Contrastive hierarchical clustering. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 627–643. Springer (2023)