

Strongly Augmented Contrastive Clustering

Xiaozhi Deng, Dong Huang, *Member, IEEE*, Ding-Hua Chen
Chang-Dong Wang, *Member, IEEE*, Jian-Huang Lai, *Senior Member, IEEE*,

Abstract—Deep clustering has attracted increasing attention in recent years due to its capability of joint representation learning and clustering via deep neural networks. In its latest developments, the contrastive learning has emerged as an effective technique to substantially enhance the deep clustering performance. However, the existing contrastive learning based deep clustering algorithms mostly focus on some carefully-designed augmentations (often with limited transformations to preserve the structure), referred to as weak augmentations, but cannot go beyond the weak augmentations to explore the more opportunities in stronger augmentations (with more aggressive transformations or even severe distortions). In this paper, we present an end-to-end deep clustering approach termed Strongly Augmented Contrastive Clustering (SACC), which extends the conventional two-augmentation-view paradigm to multiple views and jointly leverages strong and weak augmentations for strengthened deep clustering. Particularly, we utilize a backbone network with triply-shared weights, where a strongly augmented view and two weakly augmented views are incorporated. Based on the representations produced by the backbone, the weak-weak view pair and the strong-weak view pairs are simultaneously exploited for the instance-level contrastive learning (via an instance projector) and the cluster-level contrastive learning (via a cluster projector), which, together with the backbone, can be jointly optimized in a purely unsupervised manner. Experimental results on five challenging image datasets have shown the superiority of our SACC approach over the state-of-the-art. The code is available at <https://github.com/dengxiaozhi/SACC>.

Index Terms—Data clustering, Deep clustering, Image clustering, Contrastive learning, Deep neural network

I. INTRODUCTION

DATA clustering is a fundamental task in unsupervised learning, which aims to group a set of data samples into different unlabeled clusters. The traditional clustering algorithms, such as K -means [1], agglomerative clustering (AC) [2], and spectral clustering (SC) [3], typically rely on the hand-crafted data features, which lack the representation learning ability and may lead to poor clustering performance when dealing with some complex high-dimensional data, such as images and videos, where the proper features are not easy to be manually extracted.

This project was supported by the NSFC (61976097, 61876193 & 62076258) and the Natural Science Foundation of Guangdong Province (2021A1515012203). (*Corresponding author: Dong Huang*)

X. Deng, D. Huang, and D.-H. Chen are with the College of Mathematics and Informatics, South China Agricultural University, Guangzhou, China. E-mail: dengxiaozhi45@gmail.com, huangdonghere@gmail.com, dhchen@stu.scau.edu.cn.

C.-D. Wang and J.-H. Lai are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, and also with Guangdong Key Laboratory of Information Security Technology, Guangzhou, China, and also with Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China.

E-mail: changdongwang@hotmail.com, stsljh@mail.sysu.edu.cn.

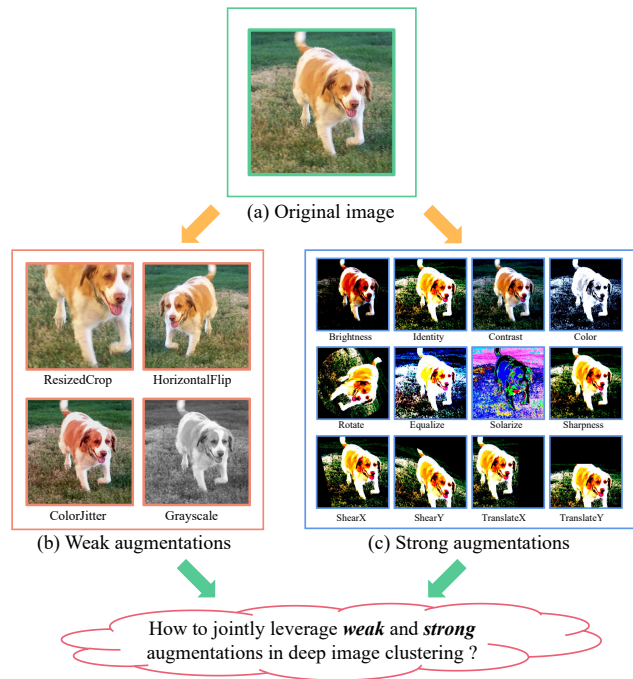


Fig. 1. The visualization of weak and strong augmentations. (a) Original image. (b) Weak augmentations. (c) Strong augmentations.

With the rapid development of deep learning, the deep neural network has recently been adopted to learn proper representations for the clustering task on complex high-dimensional data. In the past few years, many clustering algorithms based on deep neural networks (referred to as deep clustering algorithms) have been devised [4]–[8]. As one of the earliest deep clustering works, Xie et al. [9] proposed a deep embedded clustering (DEC) method to simultaneously perform feature representations learning and clustering in a deep neural network, where the distribution of soft labels and an auxiliary target distribution are constrained via a Kullback-Leibler (KL) divergence based loss. Guo et al. [5] developed an improved deep embedded clustering (IDEC) method by learning the feature representation and the cluster assignment with local structure preservation. Caron et al. [6] iteratively clustered the learned features via K -means and update the weights of the deep neural network by using the cluster assignment as soft labels. Ji et al. [7] designed a deep clustering method termed invariant information clustering (IIC), which seeks to maximize the mutual information between the original image and the augmented one for more robust representation learning and clustering.

Despite the considerable progress, these methods [4]–[7]

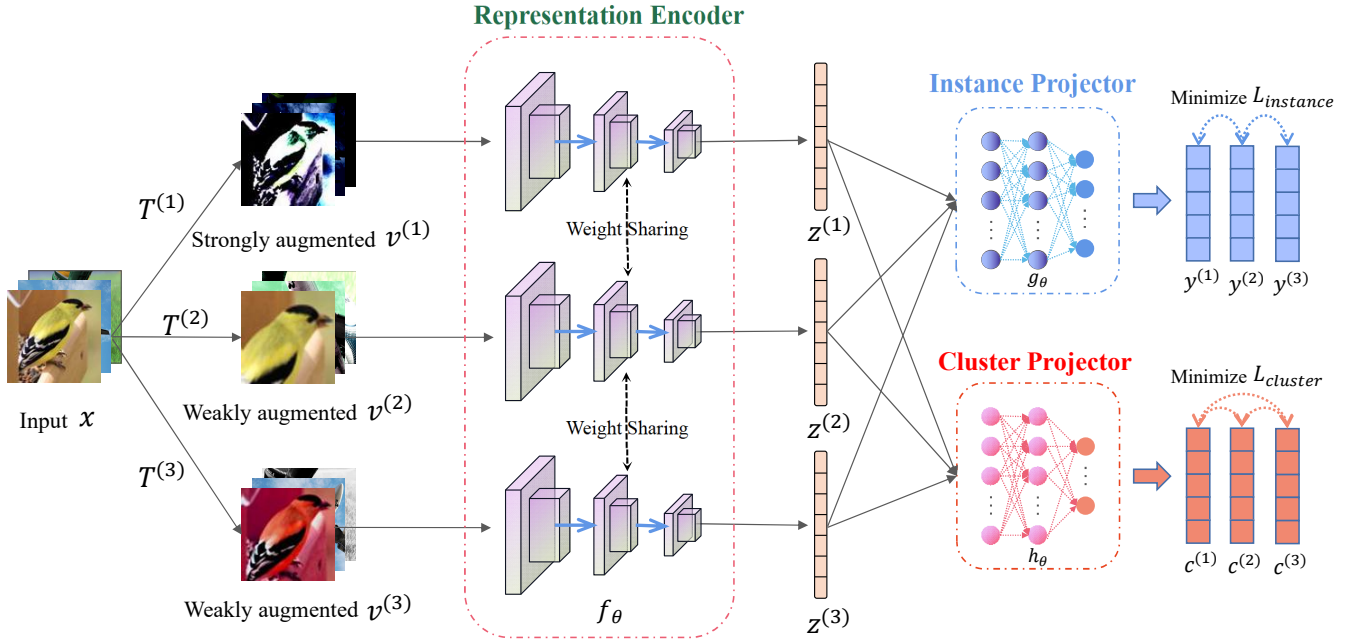


Fig. 2. Illustration of the proposed SACC framework. By simultaneously incorporating strong and weak augmentations, a backbone network with triply-shared weights is used to learn the representations of three augmentation views, which are then fed to two types of projectors for instance-level and cluster-level contrastive learning, respectively.

often perform the feature representation learning and clustering by considering the overall distributions (such as the distribution of soft labels or some other target distributions), which overlook the the sample-wise relationships and their contrastiveness. Recently, the contrastive learning has emerged as an effective technique for enhancing the deep clustering performance, which typically generates positive sample pairs and negative sample pairs via data augmentations, and aims to maximize the agreement between positive pairs and minimize the agreement between negative pairs. For example, van Gansbeke et al. [10] proposed a two-stage deep clustering method termed semantic clustering by adaptive nearest neighbors (SCAN), which first utilizes the contrastive learning to learn the discriminant features for finding the K nearest neighbors, and then trains the network via a loss function that aims to pull each sample and its K nearest neighbors closer. Dang et al. [11] extended the SCAN method by matching both local and global nearest neighbors. Li et al. [12] devised a one-stage method termed contrastive clustering (CC), which jointly exploits instance-level and cluster-level contrastive learning in an end-to-end manner.

Though these contrastive learning based deep clustering methods [10]–[12] have shown substantial improvements on some complex image datasets, yet there are still two limitations to most of them. First, the previous deep clustering methods with contrastive learning tend to utilize weak augmentations (with limited transformations to preserve the structure) for original images, but mostly overlook the opportunities in stronger augmentations which may even be coupled with more aggressive transformations or distortions (as illustrated

in Fig. 1). Second, they mostly design the network with two augmentation views (typically drawn from the same family of augmentations), but cannot go beyond two augmentation views to explore multiple views of augmentations (especially with varying degrees of transformations or distortions). More recently, Wang and Qi [13] shown that the incorporation of stronger augmentations can enhance the feature representation learned by contrastive learning, which, however, is designed for the general-purpose contrastive learning but lacks the ability to achieve the representation learning and clustering simultaneously. It remains an open problem how to simultaneously leverage the strong and weak augmentations while extending the conventional two-view network architecture to explore opportunities in multiple augmentation views in a unified deep clustering framework.

In light of this, this paper presents a novel end-to-end deep clustering approach termed strongly augmented contrastive clustering (SACC), which is able to jointly learn the feature representation and the cluster assignments with the strong and weak augmentations simultaneously leveraged in a network of multiple augmentation views (as illustrated in Fig. 2). In particular, our SACC approach utilizes a backbone network with triply-shared weights to produce the feature embeddings of a strongly augmented view and two weakly augmented views, upon which the weak-weak view pair and the strong-weak view pairs can be constructed for the instance-level and cluster-level contrastive learning. With the strong and weak augmentations as well as the instance-level and cluster-level contrastive learning jointly leveraged, the network training can thus be performed in a purely unsupervised manner and

the deep clustering result is therefore obtained. Extensive experiments are conducted on five challenging image datasets, which demonstrate the superiority of our SACC approach over the state-of-the-art deep clustering approaches.

For clarity, the key contributions of this work are summarized as follows.

- This paper for the first time, to the best of our knowledge, jointly leverages strong and weak augmentations for the task of unsupervised image clustering.
- A novel end-to-end deep clustering approach termed SACC is proposed, which utilizes three augmentations views for simultaneous instance-level and cluster-level contrastive learning.
- Extensive experimental results have confirmed that our SACC approach outperforms the state-of-the-art deep clustering approaches on several challenging image datasets.

The remainder of this paper is organized as follows. Section II reviews the related works on deep clustering. Section III describes the proposed SACC framework. Section IV reports the experimental results. Finally, we conclude this paper in Section V.

II. RELATED WORK

Deep learning has proved to be an advantageous technique for unsupervised clustering of very complex data. Many deep clustering methods have been designed [4]–[8], [14]–[20], whose difference can often be reflected by their network losses, such as the reconstruction loss of autoencoder (AE), the variational loss of variational autoencoder (VAE) [21], the loss of generative adversarial network (GAN) [22], and some specific clustering losses [23], [24].

The AE-based deep clustering methods generally optimize the networks by both the reconstruction loss and some clustering loss. The reconstruction loss measures the disagreement between the original input and the reconstruction. Yang et al. [14] presented the deep clustering network (DCN) method with the dimensionality reduction with the K -means clustering jointly modeled. Ji et al. [15] proposed an AE-based deep clustering method with a self-expressive layer for deep subspace clustering. Dizaji et al. [16] developed the deep embedded regularized clustering (DEPICT) method based on AE embedding and relative entropy minimization.

The VAE-based deep clustering methods utilize the VAE to regularize the network training to avoid over-fitting by enforcing the latent space to follow some predefined distribution. Jiang et al. [17] presented a variational deep embedding (VaDE) method that optimizes the VAE by maximizing the evidence lower bound. Dilokthanakul et al. [18] proposed a Gaussian mixture variational autoencoder (GMVAE) method by incorporating a variational Bayes in its optimization objective.

The GAN-based deep clustering methods seek to train the network with a min-max adversarial game. Springenberg [19] proposed a categorical generative adversarial network (CatGAN) method that jointly exploits GAN and regularized information maximization (RIM) to train the network. Chen

et al. [25] developed an information maximizing generative adversarial network (InfoGAN) method that aims to extract interpretable and disentangled features for deep clustering.

Different from the above three categories that usually combine the clustering loss with some network losses (such as the losses of AE, VAE, and GAN), another category of deep clustering methods aim to train the network with only the clustering loss. For example, Yang et al. [23] leveraged a convolutional neural network to learn representation feature and image clusters by a weighted triplet loss. Xie et al. [9] devised a deep embedding clustering (DEC) method that jointly optimizes deep embedding and clustering with a KL divergence based loss between the distribution of soft labels and an auxiliary target distribution. Guo et al. [24] developed an adaptive self-paced deep clustering with data augmentation (ASPC-DA) method that incorporates data augmentation and self-paced learning into deep clustering.

Recently, the contrastive learning has become a popular topic [26], and several attempts have been made to utilize the contrastive loss to improve the deep clustering performance [10]–[12]. Typically, van Gansbeke et al. [10] proposed a two-stage deep clustering method which adopts the contrastive learning as a pretext task to learn discriminant features and then exploits the K nearest neighbors (via the learned features) in the second-stage network training. Dang et al. [11] presented a nearest neighbor matching (NNM) method by considering not only the global nearest neighbors but also the local nearest neighbors. Li et al. [12] performed contrastive learning at both instance-level and cluster-level and obtained the clustering result via a cluster projector.

III. PROPOSED FRAMEWORK

In this section, we describe the proposed SACC framework. Specifically, an overview of the framework is given in Section III-A. The weak and strong augmentations are introduced in Section III-B. The design of the network architecture is provided in Section III-C. Finally, the implementation details are presented in Section III-D.

A. Framework Overview

The overall framework of SACC is illustrated in Fig. 2. In SACC, we utilize a backbone network with triply-shared weights, where the representations of three augmentation views (including one strong augmentation view and two weak augmentation views) are learned. Specifically, given a mini-batch of N image samples, we perform one type of strong augmentation and two types of weak augmentations on each input image, denoted as x_i , which lead to $3 \cdot N$ augmented samples denoted as $\{v_1^1, \dots, v_N^1, v_1^2, \dots, v_N^2, v_1^3, \dots, v_N^3\}$, with N strongly augmented samples and $2 \cdot N$ weakly augmented samples. The backbone network f_θ transforms each augmented sample v_i^j to z_i^j , with $i \in [1, N]$ and $j \in \{1, 2, 3\}$, which will then be fed to the instance projector and the cluster projector. With the instance projector g_θ transforming z_i^j to y_i^j and the cluster predictor h_θ transforming z_i^j to c_i^j , two types of feature matrices are built (for this mini-batch of samples) via the two projectors, respectively. Thereafter, unsupervised

network training can be performed by simultaneous optimizing the instance-level contrastive loss (in the row space of the feature matrix in the instance projector) and the cluster-level contrastive loss (in the column space of the feature matrix in the cluster projector).

B. Augmentations: From Weak to Strong

The contrastive learning has shown its promising ability in unsupervised representation learning [26], and has been utilized in some recent deep clustering methods [10]–[12]. In previous deep clustering methods with contrastive learning, some weak augmentations (with limited transformations to preserve the image structure) are generally exploited to form the positive pairs. However, few of them have gone beyond the weak augmentations to utilize some stronger augmentations (with more aggressive transformations or even severe distortions).

In this paper, we have shown that the joint use of strong and weak augmentations can substantially strengthen the contrastive learning ability for representation learning and clustering in the deep clustering framework. Specifically, in terms of the weak augmentations, we adopt a family of four often-used augmentations, namely, ResizedCrop, HorizontalFlip, ColorJitter, and Grayscale, to generate weakly augmented samples. For the two weak augmentation views in SACC, two augmentations are randomly chosen from the family of weak augmentations for each input image. Besides the weak augmentations, we adopt a family of fourteen stronger transformations [13], including AutoContrast, Brightness, Color, Contrast, Equalize, Identity, Posterize, Rotate, Sharpness, ShearX/Y, Solarize and TranslateX/Y. Since the strong augmentations transform the original image more aggressively, it can provide some additional clues that do not exist in the weak augmentations for learning distinctive representations. By jointly modeling strong and weak augmentations, the proposed SACC framework is able to obtain more representative patterns and semantic information of images for learning clustering-friendly representations.

C. Network Architecture

The network architecture of SACC consists of three modules, namely, the backbone network, the instance projector, and the cluster projector. An instance-level contrastive loss and a cluster-level contrastive loss are utilized in the instance projector and the cluster projector, respectively, which are jointly trained with both the weak and strong augmentation views. In the following, we will describe the three modules as well as the overall objective function in detail.

1) *Backbone Network with Triply-Shared Weights:* In the proposed framework, we utilize a backbone network with triply-shared weights (as shown in Fig. 2). Specifically, three augmentation views, including a strong augmentation view and two weak augmentation views share the backbone network f_θ , through which three views of representations can be learned, denoted as $z_i^1 = f(v_i^1)$, $z_i^2 = f(v_i^2)$, and $z_i^3 = f(v_i^3)$. Then three feature representations are fed to each of the two projectors for the later instance-level and cluster-level

contrastive learning. Note that we can adopt different network structures as the backbone. In this work, we adopt the widely-used ResNet-34 [27] as our backbone network.

2) *Instance Projector with Weak-Strong Augmentations:* In the instance projector, a two-layer nonlinear multilayer perceptron (MLP), denoted as $g(\cdot)$, is used to transform z_i^j to a lower-dimensional space, that is, $y_i^j = g(z_i^j)$, where y_i^j is interpreted as the instance representation, with $i \in [1, N]$ and $j \in \{1, 2, 3\}$.

As there are three augmentation views, including two weak augmentation views and a strong augmentation view, we use a weak-weak pair and a strong-strong pair to form the positive pairs and the negative pairs. Specifically, for each input image, its two weakly augmented samples form a positive pair, and its strongly augmented sample and its first weakly augmented sample form another positive pair. In the meantime, negative pairs are formed between the augmented samples from different input images.

With the positive pairs and negative pairs defined, the instance-level contrastive loss is utilized to maximize the agreement of positive pairs while increasing the distance of negative pairs. To measure the similarity of instance pairs, the cosine similarity can be used, that is

$$s(u, v) = \frac{u^\top v}{\|u\| \|v\|}, \quad (1)$$

where u and v denote two feature vectors. To optimize the agreement of contrastive pairs constructed from two augmentations, say, augmentation a and augmentation b , the contrastive loss for an augmented sample v_i^a is defined as

$$\ell_i^a = -\log \frac{\exp(s(y_i^a, y_i^b)/\tau_g)}{\sum_{j=1}^N [\exp(s(y_i^a, y_j^a)/\tau_g) + \exp(s(y_i^a, y_j^b)/\tau_g)]}, \quad (2)$$

with $i, j \in [1, N]$ and $a, b \in \{1, 2, 3\}$. The parameter τ_g is the temperature parameter. In order to identify all positive pairs from two augmentations (say, a and b), the instance-level contrastive loss is calculated over every augmented examples, that is

$$\mathcal{L}_{instance(a,b)} = \frac{1}{2N} \sum_{i=1}^N (\ell_i^a + \ell_i^b) \quad (3)$$

In the instance projector, we construct two augmented view pairs for each original input. One pair consists of a strongly augmented view and a weakly augmented view, while the other one consists of two weakly augmented views. Thus the contrastive loss for the instance projector is defined as

$$\mathcal{L}_{instance} = \mathcal{L}_{instance(1,2)} + \mathcal{L}_{instance(2,3)} \quad (4)$$

3) *Cluster Projector with Weak-Strong Augmentations:* The cluster projector is a two-layer nonlinear MLP with a softmax layer, denoted as $h(\cdot)$. The dimension of the output layer of the cluster projector, denoted as M , is equal to the number of classes (or the desired number of clusters). The output representation (for each sample) in the cluster projector, computed by $\tilde{c}_i^j = h(z_i^j)$, can be treated as the probabilities of this sample belonging to different classes. Thus, \tilde{c}_i^j can serve as a soft label for the augmented sample.

For each of the three augmentation views, a feature matrix with N rows and M columns can be obtained for a mini-batch of N samples, where c_m^j denotes the m -th column of the feature matrix, with $m \in [1, M]$ and $j \in \{1, 2, 3\}$. That is, c_m^j can be regarded as the distribution of the N samples in the m -th cluster of the augmentation j . We treat the same cluster from two different augmentation views as a positive cluster pair, and the other cluster pairs as the negative cluster pairs. Then, for a cluster c_m^j , the cluster-level contrastive loss can be defined as

$$\hat{\ell}_m^a = -\log \frac{\exp(s(c_m^a, c_m^b)/\tau_h)}{\sum_{n=1}^M [\exp(s(c_m^a, c_n^a)/\tau_h) + \exp(s(c_m^a, c_n^b)/\tau_h)]} \quad (5)$$

with $m, n \in [1, M]$ and $a, b \in \{1, 2, 3\}$. The parameter τ_h is the temperature parameter. After traversing all clusters, the cluster-level contrastive loss can further be represented as

$$\mathcal{L}_{cluster(a,b)} = \frac{1}{2M} \sum_{m=1}^M (\hat{\ell}_m^a + \hat{\ell}_m^b) - H(Y), \quad (6)$$

$$H(Y) = -\sum_{m=1}^M [P(c_m^a) \log P(c_m^a) + P(c_m^b) \log P(c_m^b)], \quad (7)$$

where $H(Y)$ is the entropy of the cluster assignment probabilities with $P(c_m^k) = \sum_{n=1}^N Y_{nm}^k / \|Y\|_1$, for $k \in \{a, b\}$ within a mini-batch under each data augmentation. This term is incorporated to avoid the trivial solution that most samples are assigned to the same cluster.

For every original image, we utilize three augmentation view pairs in the cluster projector, i.e., every two augmentation views form a view pair. Then the contrastive loss in the cluster projector is defined as

$$\mathcal{L}_{cluster} = \mathcal{L}_{cluster(1,2)} + \mathcal{L}_{cluster(1,3)} + \mathcal{L}_{cluster(2,3)} \quad (8)$$

4) *Overall Objective*: The optimization of the backbone network, the instance projector, and the cluster projector is jointly performed in an end-to-end manner. The overall objective function is composed of the instance-level contrastive loss and the cluster-level contrastive loss, that is

$$\mathcal{L} = \mathcal{L}_{cluster} + \mathcal{L}_{instance}. \quad (9)$$

Thereby, the unsupervised network training of our SACC approach can be conducted with both instance-level and cluster-level contrastive learning upon three weak/strong augmentation views.

D. Implementation Details

In the proposed framework, all original images of different sizes are resized to the size of 224×224 . We use the ResNet-34 as the backbone network which is designed for input images with 224×224 pixels. As for the instance projector, its output dimensionality set to 128 in order to hold sufficient information after the transformation. As for the cluster projector, the output dimensionality is set to the desired number of clusters, where the output feature vector can be treated as the soft label. The temperature parameters of the instance projector and the cluster projector are fixed to 0.5 and 1, respectively. In the

TABLE I
DESCRIPTION OF THE BENCHMARK DATASETS.

Dataset	#Images	#Classes
CIFAR-10	60,000	10
CIFAR-100	60,000	20
STL-10	13,000	10
ImageNet-10	13,000	10
ImageNet-Dogs	19,500	15

training process, we adopt Adam optimizer with a learning rate of 0.0003 to simultaneously optimize the backbone network and the two projectors. The batch size for training is set to 200, and the number of training epochs is set to 1000.

IV. EXPERIMENTS

In this section, we conduct extensive experiments on five image datasets to evaluate the clustering performance of our SACC algorithm against eighteen traditional and deep clustering algorithms.

A. Datasets and Evaluation Metrics

In our experiments, five challenging image datasets are used, namely, CIFAR-10 [39], CIFAR-100 [39], STL-10 [40], ImageNet-10 [33], and ImageNet-Dogs [33]. Similar to the previous deep clustering works [12], [35], [37], we jointly utilize the training and testing samples of each dataset, due to the unsupervised nature of the clustering task. Specifically, the five benchmark datasets are introduced below.

- **CIFAR-10** is a natural image dataset which consists of 60,000 images from 10 object classes.
- **CIFAR-100** has the same size and the same number of samples as the CIFAR-10 dataset, but it contains 20 super-classes, which can be further divided into 100 classes. Following the previous works, we use the 20 super-classes as the ground-truth when evaluating the clustering performance.
- **STL-10** is an ImageNet-sourced dataset which collects 13,000 color images with the size of 96×96 from 10 classes.
- **ImageNet-10** is a subset of ImageNet with 10 classes, each of which consists of 1,300 samples with varying image sizes.
- **ImageNet-Dogs** is constructed in a similar way to ImageNet-10, but it selects a total of 19,500 dog images of 15 breeds from the ImageNet dataset.

For clarity, we visualize some samples from the five image datasets in Fig. 3, and provide the statistics of them in Table I.

To evaluate the clustering results of different clustering algorithms, we adopt three widely-used evaluation metrics [12], [35], [41], namely, clustering accuracy (ACC), normalized mutual information (NMI) and adjusted rand index (ARI). Note that greater values of the three metrics indicate better clustering results.

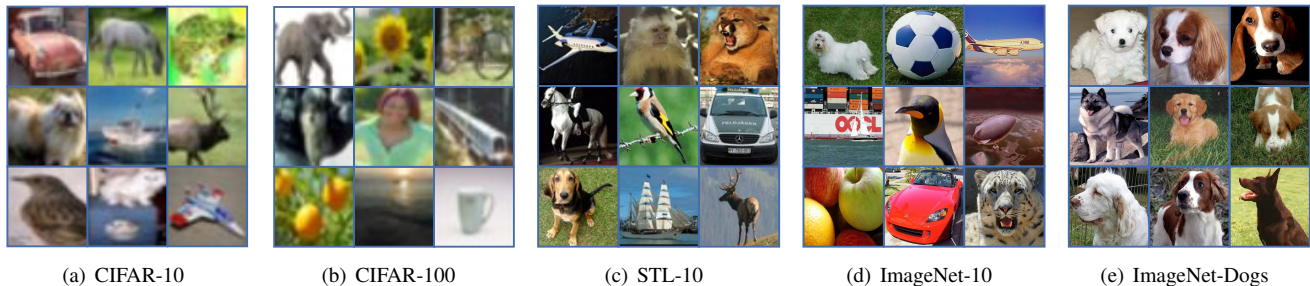


Fig. 3. Visualization of some image samples from the five benchmark datasets.

TABLE II
THE CLUSTERING PERFORMANCE (W.R.T. NMI(%)) BY DIFFERENT CLUSTERING ALGORITHMS ON THE FIVE IMAGE DATASETS. THE BEST SCORE IN EACH COLUMN IS IN **BOLD**.

Dataset	CIFAR-10	CIFAR-100	STL-10	ImageNet-10	ImageNet-Dogs
K-means [1]	8.7	8.4	12.5	11.9	5.5
SC [3]	10.3	9.0	9.8	15.1	3.8
AC [2]	10.5	9.8	23.9	13.8	3.7
NMF [28]	8.1	7.9	9.6	13.2	4.4
AE [29]	23.9	10.0	25.0	21.0	10.4
DAE [30]	25.1	11.1	22.4	20.6	10.4
DCGAN [31]	26.5	12.0	21.0	22.5	12.1
DeCNN [32]	24.0	9.2	22.7	18.6	9.8
VAE [21]	24.5	10.8	20.0	19.3	10.7
JULE [23]	19.2	10.3	18.2	17.5	5.4
DEC [9]	25.7	13.6	27.6	28.2	12.2
DAC [33]	39.6	18.5	36.6	39.4	21.9
DDC [34]	42.4	-	37.1	43.3	-
DCCM [35]	49.6	28.5	37.6	60.8	32.1
IIC [7]	51.1	22.5	49.6	-	-
GATCluster [36]	49.6	28.5	44.6	59.4	28.1
PICA [37]	59.1	31.0	61.1	80.2	35.2
DRC [38]	62.1	35.6	64.4	83.0	38.4
CC [12]	68.1	42.4	67.4	86.2	40.1
SACC(our)	76.5	44.8	69.1	87.7	45.5

B. Baseline Methods

In the experiments, the proposed SACC method is compared with both traditional clustering methods and deep clustering methods. Specifically, eighteen baseline clustering methods are compared, including four traditional clustering methods, namely, K -means [1], agglomerative clustering (AC) [2], spectral clustering (SC) [3], and nonnegative matrix factorization (NMF) [28], and fourteen deep clustering methods, including autoencoders (AE) [29], denoising autoencoder (DAE) [30], deep convolutional generative adversarial networks (DCGAN) [31], deconvolutional networks (DeCNN) [32], variational auto-encoder (VAE) [21], joint unsupervised learning (JULE) [23], deep embedded clustering (DEC) [9], deep adaptive clustering (DAC) [33], deep discriminative clustering (DDC) [34], deep comprehensive correlation mining (DCCM) [35], invariant information clustering (IIC) [7], Gaussian attention network for image clustering (GATCluster) [36], partition

confidence maximization (PICA) [37], deep robust clustering (DRC) [38] and contrastive clustering (CC) [12]. For the CC method, the NMI, ACC, and ARI scores are reproduced by running the authors' code [12], while the scores of the other baseline methods are taken from the corresponding papers.

C. Results and Analysis

In this section, we compare the proposed SACC method against both traditional and deep clustering methods on the five benchmark datasets.

The NMI, ACC, and ARI scores of different clustering methods are reported in Tables II, III, and IV, respectively. From Table II, we can observe that the deep clustering methods can significantly outperform the traditional (non-deep) clustering methods, due to the representation learning ability of the deep neural networks. In terms of the proposed method, SACC achieves the best NMI score on the five benchmark

TABLE III

THE CLUSTERING PERFORMANCE (W.R.T. ACC(%)) BY DIFFERENT CLUSTERING ALGORITHMS ON THE FIVE IMAGE DATASETS. THE BEST SCORE IN EACH COLUMN IS IN **BOLD**.

Dataset	CIFAR-10	CIFAR-100	STL-10	ImageNet-10	ImageNet-Dogs
K-means [1]	22.9	13.0	19.2	24.1	10.5
SC [3]	24.7	13.6	15.9	27.4	11.1
AC [2]	22.8	13.8	33.2	24.2	13.9
NMF [28]	19.0	11.8	18.0	23.0	11.8
AE [29]	31.4	16.5	30.3	31.7	18.5
DAE [30]	29.7	15.1	30.2	30.4	19.0
DCGAN [31]	31.5	15.3	29.8	34.6	17.4
DeCNN [32]	28.2	13.3	29.9	31.3	17.5
VAE [21]	29.1	15.2	28.2	33.4	17.9
JULE [23]	27.2	13.7	27.7	30.0	13.8
DEC [9]	30.1	18.5	35.9	38.1	19.5
DAC [33]	52.2	23.8	47.0	52.7	27.5
DDC [34]	52.4	-	48.9	57.7	-
DCCM [35]	62.3	32.7	48.2	71.0	38.3
IIC [7]	61.7	25.7	59.6	-	-
GATCluster [36]	62.3	32.7	58.3	73.9	32.2
PICA [37]	69.6	33.7	71.3	87.0	35.2
DRC [38]	72.7	36.7	74.7	88.4	38.9
CC [12]	76.6	42.6	74.7	89.5	34.2
SACC(our)	85.1	44.3	75.9	90.5	43.7

TABLE IV

THE CLUSTERING PERFORMANCE (W.R.T. ARI(%)) BY DIFFERENT CLUSTERING ALGORITHMS ON THE FIVE IMAGE DATASETS. THE BEST SCORE IN EACH COLUMN IS IN **BOLD**.

Dataset	CIFAR-10	CIFAR-100	STL-10	ImageNet-10	ImageNet-Dogs
K-means [1]	4.9	2.8	6.1	5.7	2.0
SC [3]	8.5	2.2	4.8	7.6	1.3
AC [2]	6.5	3.4	14.0	6.7	2.1
NMF [28]	3.4	2.6	4.6	6.5	1.6
AE [29]	16.9	4.8	16.1	15.2	7.3
DAE [30]	16.3	4.6	15.2	13.8	7.8
DCGAN [31]	17.6	4.5	13.9	15.7	7.8
DeCNN [32]	17.4	3.8	16.2	14.2	7.3
VAE [21]	16.7	4.0	14.6	16.8	7.9
JULE [23]	13.8	3.3	16.4	13.8	2.8
DEC [9]	16.1	5.0	18.6	20.3	7.9
DAC [33]	30.6	8.8	25.7	30.2	11.1
DDC [34]	32.9	-	26.7	34.5	-
DCCM [35]	40.8	17.3	26.2	55.5	18.2
IIC [7]	41.1	11.7	39.7	-	-
GATCluster [36]	40.8	17.3	36.3	55.2	16.3
PICA [37]	51.2	17.1	53.1	76.1	20.1
DRC [38]	54.7	20.8	56.9	79.8	23.3
CC [12]	60.6	26.7	60.6	82.5	22.5
SACC(our)	72.4	28.2	62.6	84.3	28.5

datasets. Especially, on the CIFAR-10 and ImageNet-Dogs datasets, SACC achieves NMI(%) scores of 76.5 and 45.5, respectively, where the second best NMI(%) scores are only 68.1 and 40.1, respectively. On the other three datasets, SACC

also obtains better NMI scores than the other deep clustering methods. Similar advantages can be observed in Tables III and IV, where our SACC method also achieves the best ACC and ARI scores on all the five benchmark datasets.

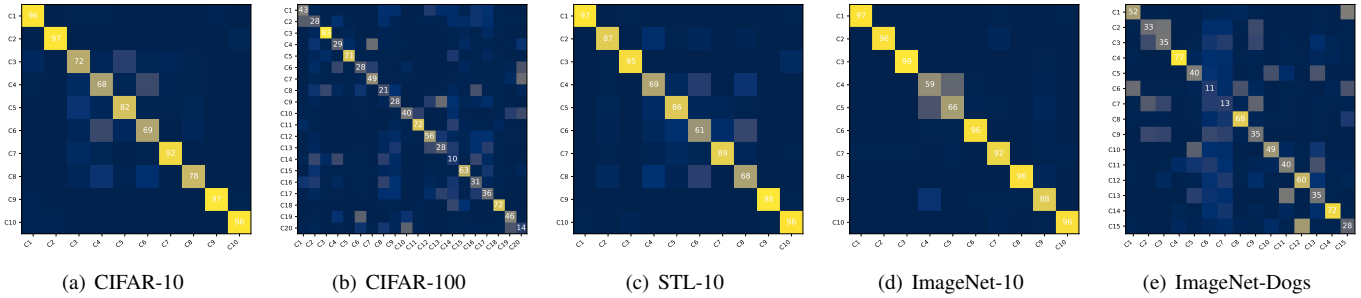


Fig. 4. The confusion matrices on the five image datasets, where the rows are the ground-truth labels and the columns are the predicted labels by SACC.

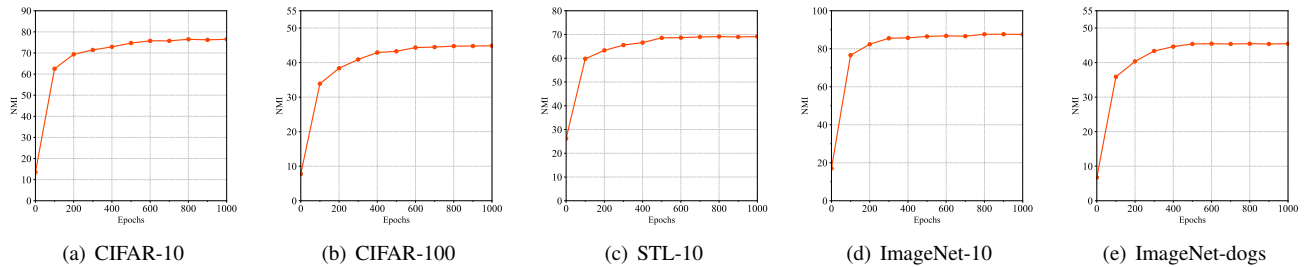


Fig. 5. The NMI(%) performance of SACC as the number of epochs increases.

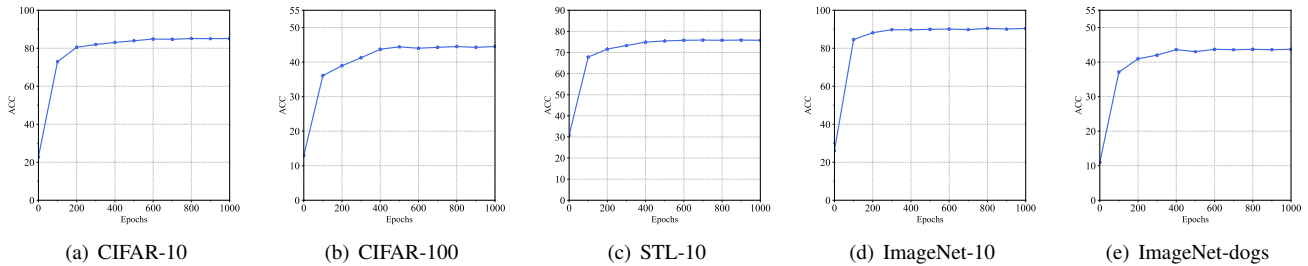


Fig. 6. The ACC(%) performance of SACC as the number of epochs increases.

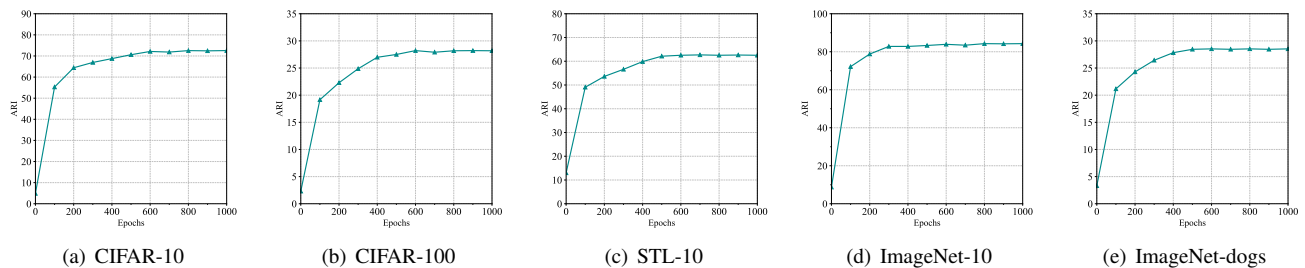


Fig. 7. The ARI(%) performance of SACC as the number of epochs increases.

Besides the quantitative evaluation, we further provide visual analysis on the clustering results of our SACC method. Specifically, Fig. 4 illustrates the confusion matrices between the true and predicted labels on the five image datasets. As shown in Fig. 4, clear block-diagonal structures can be observed in the confusion matrices for the CIFAR-10, STL-10, and ImageNet-10 datasets. Even for the more challenging

datasets of CIFAR-100 and ImageNet-Dogs, we can still observe the block-diagonal structures, though they are not as clear as the other three datasets. Notably, even for the challenging datasets like CIFAR-100 and ImageNet-Dogs, our SACC method can still yield better clustering performance than the state-of-the-art deep clustering methods (as shown in Tables II, III, and IV).

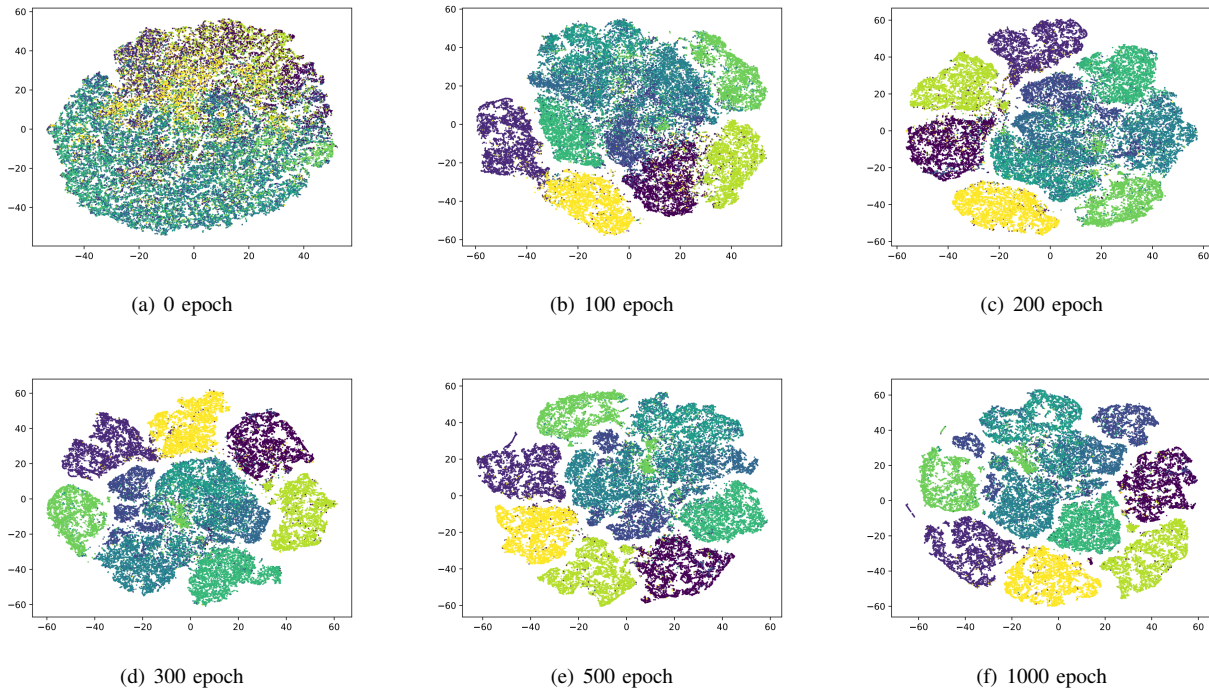


Fig. 8. The t-SNE visualization of SACC on the CIFAR-10 dataset.

TABLE V
THE NMI(%), ACC(%), AND ARI(%) BY SACC USING DIFFERENT COMBINATIONS OF WEAK AND STRONG AUGMENTATIONS.

Augmentations	NMI	ACC	ARI
Weak+Weak	68.1	76.6	60.6
Weak+Strong	72.6	82.6	67.9
Weak+Weak+Strong	76.5	85.1	72.4

TABLE VI
THE NMI(%), ACC(%), AND ARI(%) BY SACC USING ONE OR TWO OF ITS CONTRASTIVE PROJECTORS.

Used Projectors	NMI	ACC	ARI
With Only Cluster Projector	67.5	73.4	58.8
With Only Instance Projector	69.2	75.5	61.5
With Both Projectors	76.5	85.1	72.4

D. Ablation Study

In this section, we conduct the ablation study on the CIFAR-10 dataset to test the influence of the strong and strong augmentations and that of the two contrastive projectors in our SACC method.

1) *Influence of Strong and Weak Augmentations:* In our SACC framework, three augmentation views are utilized, including a strongly augmented views and two weakly augmented views (as shown in Fig. 2). In the section, we test

the influence of the three augmentation views. Note that at least two augmentation views should be preserved to make contrastive learning feasible. As shown in Table V, using a weak augmentation and a strong augmentation leads to better clustering performance than using two weak augmentations. Further, the proposed SACC method using all three augmentations significantly outperforms that variant of using two augmentations (weak+weak or weak+strong), which demonstrate the substantial benefits brought in by our network architecture with three augmentation views.

2) *Influence of Two Contrastive Projectors:* In SACC, we utilize two projectors, namely, the instance projector and the cluster projector, for the instance-level and cluster-level contrastive learning, respectively. In the section, we test the influence of these two projectors by evaluating the performance of SACC with one of the two projectors removed. Note that in SACC, the clustering result is obtained in the cluster projector. For the variant with the cluster projector removed, we obtain the clustering result by performing K -means on the feature representation learned by the instance projector. As shown in Table VI, using both projectors can lead to significantly better clustering performance (w.r.t. NMI, ACC, and ARI) than using only one of the two projectors, which confirm the advantage of the joint instance-level and cluster-level contrastive learning in our SACC method.

E. Convergence Analysis

In this section, we evaluate the convergence of the proposed SACC method. Specifically, the NMI, ACC, and ARI scores of the proposed method are recorded for each 100 epochs, which

are then plotted in Fig. 5, 6 and 7. As shown Fig. 5, 6 and 7, the NMI, ACC, and ARI scores of SACC consistently increase as the number of epochs grows. On most of the datasets, the proposed SACC method can reach high-quality clustering results when the number of epochs is greater than 500. In this paper, the number of epochs is set to 1,000 for all datasets.

Further, we visualize convergence of SACC by performing t-SNE [42] on the learned feature representations the learned feature representation on the CIFAR-10 dataset, which is shown in Fig. 8, where different colors denote different ground-truth labels. As can be observed in Fig. 8, at the beginning, the data samples are mostly mixed. After training 100 epochs, many samples in the same class have been grouped closer. As the learning process proceeds, the distribution of the data samples reach stability after about 500 epochs, where the separability of the samples in most classes becomes relatively clear.

V. CONCLUSION

In this paper, we propose a novel deep clustering approach termed SACC. Different from the previous contrastive learning based deep clustering approaches which typically use some weak augmentations with limited transformations, our SACC approach is able to jointly leverage strong and weak augmentations for enhancing the simultaneous contrastive representation learning and clustering. In particular, we utilize a backbone network with triply-shared weights for three augmentation views, including a strongly augmented view and two weakly augmented views. Three views of representations (for weakly and strongly augmented samples) can be obtained from the backbone network, which are then fed to two types of projectors, namely, the instance projector and the cluster projector, so as to enable the instance-level contrastive learning and the cluster-level contrastive learning, respectively. Further, the unsupervised training is performed to simultaneously optimize the backbone and the two projectors in an end-to-end manner, where the final clustering result can therefore be achieved in the cluster projector. Extensive experiments are conducted on five benchmark image datasets, which have confirmed the superior clustering performance of the proposed SACC approach over the state-of-the-art deep clustering approaches.

REFERENCES

- [1] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, 1967, pp. 281–297.
- [2] K. C. Gowda and G. Krishna, "Agglomerative clustering using the concept of mutual nearest neighbourhood," *Pattern Recognition*, vol. 10, no. 2, pp. 105–112, 1978.
- [3] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," *Advances in Neural Information Processing Systems*, vol. 17, 2004.
- [4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 31–35.
- [5] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *International Joint Conference on Artificial Intelligence*, 2017, pp. 1753–1759.
- [6] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *European Conference on Computer Vision*, 2018, pp. 132–149.
- [7] X. Ji, J. F. Henriques, and A. Vedaldi, "Invariant information clustering for unsupervised image classification and segmentation," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9865–9874.
- [8] T. Wang, W. W. Y. Ng, J. Li, Q. Wu, S. Zhang, C. Nugent, and C. Shewell, "A deep clustering via automatic feature embedded learning for human activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 210–223, 2022.
- [9] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International Conference on Machine Learning*, 2016, pp. 478–487.
- [10] W. van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "SCAN: Learning to classify images without labels," in *European Conference on Computer Vision*, 2020, pp. 268–285.
- [11] Z. Dang, C. Deng, X. Yang, K. Wei, and H. Huang, "Nearest neighbor matching for deep clustering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 693–13 702.
- [12] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *AAAI Conference on Artificial Intelligence*, 2021.
- [13] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," *arXiv preprint arXiv:2104.07713*, 2021.
- [14] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong, "Towards k-means-friendly spaces: Simultaneous deep learning and clustering," in *International Conference on Machine Learning*, 2017, pp. 3861–3870.
- [15] P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid, "Deep subspace clustering networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [16] K. G. Dizaji, A. Herandi, C. Deng, W. Cai, and H. Huang, "Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization," in *IEEE International Conference on Computer Vision*, 2017, pp. 5736–5745.
- [17] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, "Variational deep embedding: An unsupervised and generative approach to clustering," *arXiv preprint arXiv:1611.05148*, 2016.
- [18] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan, "Deep unsupervised clustering with gaussian mixture variational autoencoders," *arXiv preprint arXiv:1611.02648*, 2016.
- [19] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.
- [20] X. Dong, L. Liu, L. Zhu, Z. Cheng, and H. Zhang, "Unsupervised deep k-means hashing for efficient image retrieval and clustering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3266–3277, 2021.
- [21] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.
- [23] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5147–5156.
- [24] X. Guo, X. Liu, E. Zhu, X. Zhu, M. Li, X. Xu, and J. Yin, "Adaptive self-paced deep clustering with data augmentation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 9, pp. 1680–1693, 2019.
- [25] X. Chen, Y. Duan, R. Houthoofd, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, 2020, pp. 1597–1607.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [28] D. Cai, X. He, X. Wang, H. Bao, and J. Han, "Locality preserving nonnegative matrix factorization," in *International Joint Conference on Artificial Intelligence*, 2009.
- [29] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in Neural Information Processing Systems*, vol. 19, 2006.
- [30] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. 12, 2010.

- [31] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.
- [32] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2528–2535.
- [33] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep adaptive image clustering," in *IEEE International Conference on Computer Vision*, 2017, pp. 5879–5887.
- [34] J. Chang, Y. Guo, L. Wang, G. Meng, S. Xiang, and C. Pan, "Deep discriminative clustering analysis," *arXiv preprint arXiv:1905.01681*, 2019.
- [35] J. Wu, K. Long, F. Wang, C. Qian, C. Li, Z. Lin, and H. Zha, "Deep comprehensive correlation mining for image clustering," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8150–8159.
- [36] C. Niu, J. Zhang, G. Wang, and J. Liang, "GATcluster: Self-supervised gaussian-attention network for image clustering," in *European Conference on Computer Vision*, 2020, pp. 735–751.
- [37] J. Huang, S. Gong, and X. Zhu, "Deep semantic clustering by partition confidence maximisation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8849–8858.
- [38] H. Zhong, C. Chen, Z. Jin, and X.-S. Hua, "Deep robust clustering by contrastive learning," *arXiv preprint arXiv:2008.03030*, 2020.
- [39] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [40] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 215–223.
- [41] D. Huang, C.-D. Wang, J.-S. Wu, J.-H. Lai, and C.-K. Kwok, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 6, pp. 1212–1226, 2020.
- [42] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.



Ding-Hua Chen received the B.S. degree in computer science and technology in 2019 from South China Agricultural University, Guangzhou, China, where he is currently pursuing the master degree in computer science with the College of Mathematics and Informatics. His current research interests include deep learning and clustering analysis.



Chang-Dong Wang received the B.S. degree in applied mathematics in 2008, the M.Sc. degree in computer science in 2010, and the Ph.D. degree in computer science in 2013, all from Sun Yat-sen University, Guangzhou, China. He was a visiting student at the University of Illinois at Chicago from January 2012 to November 2012. He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. His current research interests include machine learning and data mining. He has published more than 100 scientific papers in international journals and conferences such as IEEE TPAMI, IEEE TKDE, IEEE TNNLS, IEEE TSMC-C, ACM TKDD, Pattern Recognition, SIGKDD, ICDM and SDM. His ICDM 2010 paper won the Honorable Mention for Best Research Paper Award. He was awarded 2015 Chinese Association for Artificial Intelligence (CAAI) Outstanding Dissertation.



Xiaozhi Deng received the B.S. degree in 2020 from South China Agricultural University, Guangzhou, China, where he is currently pursuing the master degree in computer science with the College of Mathematics and Informatics. His research interests include clustering analysis and deep learning.



Dong Huang received the B.S. degree in computer science in 2009 from South China University of Technology, Guangzhou, China. He received the M.Sc. degree in computer science in 2011 and the Ph.D. degree in computer science in 2015, both from Sun Yat-sen University, Guangzhou, China. He joined South China Agricultural University in 2015, where he is currently an Associate Professor with the College of Mathematics and Informatics. From July 2017 to July 2018, he was a visiting fellow with the School of Computer Science and Engineering,

Nanyang Technological University, Singapore. His research interests include data mining and machine learning. He has published more than 50 papers in international journals and conferences, such as IEEE TKDE, IEEE TCYB, IEEE TSMC-S, ACM TKDD, SIGKDD, AAAI, and ICDM. He was the recipient of the 2020 ACM Guangzhou Rising Star Award.



Jian-Huang Lai received the M.Sc. degree in applied mathematics in 1989 and the Ph.D. degree in mathematics in 1999 from Sun Yat-sen University, China. He joined Sun Yat-sen University in 1989 as an Assistant Professor, where he is currently a Professor with the School of Data and Computer Science. His current research interests include the areas of digital image processing, pattern recognition, multimedia communication, wavelet and its applications. He has published more than 200 scientific papers in the international journals and conferences on image processing and pattern recognition, such as IEEE TPAMI, IEEE TKDE, IEEE TNN, IEEE TIP, IEEE TSMC-B, Pattern Recognition, ICCV, CVPR, IJCAI, ICDM and SDM. Prof. Lai serves as a Standing Member of the Image and Graphics Association of China, and also serves as a Standing Director of the Image and Graphics Association of Guangdong.