GeCC: Generalized Contrastive Clustering with Domain Shifts Modeling

Yujie Chen[1], **Wenhui Wu** [1, 2 *], **Le Ou-Yang** [1, 3 *], **Ran Wang** [4], **Debby D. Wang** [5]

[1] College of Electronics and Information Engineering, Shenzhen University, Shenzhen, 518060, China
[2] Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, 518060, China
[3] Faculty of Engineering, Shenzhen MSU-BIT University, Shenzhen, 518116, China
[4] College of Mathematics and Statistics, Shenzhen University, Shenzhen, 518060, China
[5] School of Science and Technology, Hong Kong Metropolitan University, HongKong, China
2110436012@email.szu.edu.cn, {wuwenhui, leouyang, wangran}@szu.edu.cn, dwang@hkmu.edu.hk

Abstract

Contrastive clustering performs clustering and data representation in a unified model, where instance- and cluster-level constrastive learning are conducted simultaneously. However, commonly-used data augmentation methods make contrastive mechanism effect but may cause representation learning getting stuck in domain-specific information, which further deteriorates clustering performance and limits generalization ability. To this end, we propose a new framework, named Generalized Contrastive Clustering with domain shifts modeling (GeCC), which can integrate diverse domain knowledge to improve the clustering performance. Specifically, we first design a cluster-guided domain shifts modeling module to synthesize a reference view with diverse domain information. Then, we introduce instance representation and cluster assignment contrastive modules with well-designed attention weights to guide the representation learning and clustering. In this way, our method can maximize the extraction of cluster-related information and avoid overfitting domain-specific features. Experimental results on four benchmark datasets demonstrate that our proposed method consistently outperforms other state-of-the-art methods.

Code — https://github.com/mia-7/GeCC

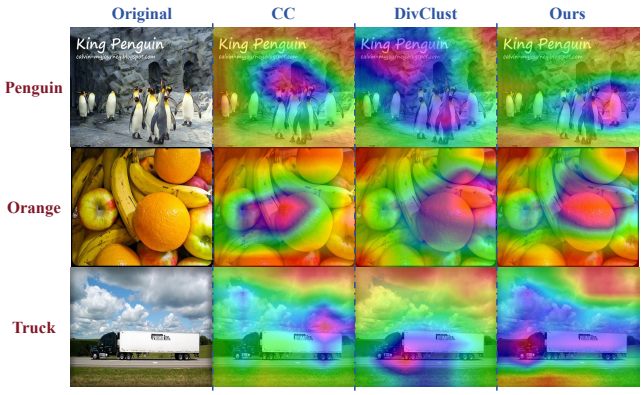


Figure 1: The class activation map (CAM) visualization of clustering prediction. Results from vanilla contrastive clustering methods (i.e., CC and DivClust) get stuck in domain-specific feature such as the background, impeding effective clustering. Nevertheless, the visualization for the proposed method focuses more on the objects, demonstrating its ability in extracting cluster-related features.

Introduction

Over the past decade, deep learning has achieved excellent performance based on large amounts of annotated training samples. However, acquiring supervisory information can be time-consuming and laborious in certain practical scenarios. Clustering, as one of the most fundamental unsupervised learning methods, has garnered significant attention. It involves grouping data into distinct clusters without the need for any labels. Leveraging the capacity of neural networks to model intricate data, deep clustering (Ren et al. 2022; Wu et al. 2023) learns informative representations, facilitating subsequent downstream tasks such as multi-view clustering (Xue et al. 2021; Liu et al. 2022) and cell clustering (Mrabah et al. 2023; He et al. 2023).

Extensive research has concentrated on integrating representation learning and clustering within an end-to-end framework. For example, several methods (Yang et al., 2016; Chang et al., 2017; Caron et al., 2018; Asano et al., 2019) utilize additional clustering algorithms to derive pseudo-labels, which are then employed as supervision to guide network training. However, these two-step approaches may accumulate errors during the alternating process between clustering and representation learning, potentially leading to suboptimal performance. Consequently, some methods (Ji et al., 2019; Huang et al., 2020; Peng et al., 2023; Niu et al., 2022) have been devised to simultaneously learn representations and cluster assignments.

Recently, contrastive learning (Chen et al. 2020; Grill et al. 2020) has demonstrated promising performance in unsupervised scenarios. Data augmentation is usually utilized to create multiple views. Treating different views of a sample as positive pairs and maximizing their similarities, and considering different samples as negative pairs and minimizing their similarities, contrastive learning can distill representations with discriminative content. Building upon this contrastive mechanism, various clustering methods have been proposed (Li et al. 2021b; Liu et al. 2023b; Huang et al.

*Corresponding authors.

2024; Metaxas, Tzimiropoulos, and Patras 2023; Chen et al. 2024). Specifically, contrastive clustering (Li et al. 2021b) trains the network with the normalized temperature-scaled cross entropy loss (Chen et al. 2020) on both the representations and cluster assignments, achieving superior clustering results.

In general, due to the limited variation of augmented operations, the domain information underlying positive pairs may be partially identical, which will be included in the data embedding learned via contrastive learning (Lin et al. 2021; Liu et al. 2023a). Therefore, these vanilla contrastive clustering methods usually get stuck in the domain-specific information. As shown in Figure 1, features extracted by vanilla contrastive clustering methods (Li et al. 2021b; Metaxas, Tzimiropoulos, and Patras 2023) include a large proportion of domain knowledge, which overtakes the cluster-related information and misleads the grouping results.

To alleviate this issue, we propose a Generalized Contrastive Clustering (GeCC) framework, equipped with the cluster-guided domain shifts modeling module, instance representation and cluster assignment contrastive modules. More specifically, our motivation is to restraint the network from over-fitting the domain features that inherent in the predefined augmentation, and maximize the extraction of cluster-related information. The cluster-guided domain shifts modeling module generates a reference view from the feature statistics, which is computed with cluster assignments to incorporate cluster-related information. To utilize samples from different domains more effectively, both of the instance representation contrastive learning and cluster assignment contrastive learning are enhanced with the well-designed attention weights. We measure the relative domain (or distribution) difference between reference and augmented views and add the domain (or distribution) attention weights to the instance-level (or cluster-level) contrastive loss. During the training, the instance projector continuously learns from features with diverse domains, while the clustering projector is optimized to align the label distributions and mine the cluster-related information. Through the cooperation and interaction among these modules, GeCC is able to consistently improve the clustering performance and generalization ability, excluding the undesirable impact induced by data augmentation.

To summarize, the contributions of our work are as follows:

1. A generalized contrastive clustering method is proposed, which incorporates a cluster-guided domain shifts modeling module, instance representation and cluster assignment contrastive modules to enhance the extraction of cluster-related information under data augmentation.
2. A cluster-guided domain shifts modeling module, which models the uncertainty in feature statistics with the guidance of clustering information, is introduced to generate reference view with variant domains. Additionally, well-designed attention weights are computed to guide the training process effectively.
3. Contrastive learning is performed on data pairs with attention weights. During the training process, the instance projector learns from diverse domains while the cluster projector excavates the cluster-related information. In this way, GeCC is disentangled from the augmentation-specific feature.
4. Extensive experiments on four benchmark datasets are performed to demonstrate the superiority of our proposed method. Elaborate ablation experiments demonstrate the effectiveness of each module in our method.

Related Work

Deep Clustering

With the powerful ability of neural networks in extracting feature representation, various deep clustering models have been proposed. Existing models can be roughly categorized into two groups, i.e., alternate learning and joint learning. In alternate learning, representation learning and clustering are performed interchangeably. For example, several methods (Xie, Girshick, and Farhadi 2016; Chang et al. 2017; Caron et al. 2018; Asano, Rupprecht, and Vedaldi 2019) construct pseudo-labels by grouping representation and update the model according to pseudo-labels. However, these methods perform clustering and representation learning separately, potentially constraining their performance. In contrast, joint learning simultaneously conducts clustering and representation learning. For example, some methods (Ji, Henriques, and Vedaldi 2019; Yang et al. 2020; Zhao et al. 2020; Huang, Gong, and Zhu 2020) impose different constraints on the cluster assignments and train the clustering models in an end-to-end way. By jointly learning cluster assignments and representations, these approaches can achieve superior clustering performance.

Contrastive Clustering

Contrastive learning (Chen et al. 2020; Grill et al. 2020) has achieved good performance in unsupervised representation learning. The basic idea of contrastive learning is to put attractive (or repulsive) forces on positive pairs (or negative pairs), with the goal of distilling the common content. However, most existing contrastive learning methods (Wang and Qi 2022; Yeh et al. 2022) focus on learning the representation as a pretext task before the downstream tasks. The separation between representation learning and clustering may lead to locally optimal solution. Contrastive clustering (Li et al. 2021b; Metaxas, Tzimiropoulos, and Patras 2023; Chen, Wu, and Ou-Yang 2023) learns discriminative representations and performs clustering simultaneously by conducting cluster-level contrastive learning on another independent subspace, which achieves better performance than other deep clustering methods. However, because augmentation is predefined, existing contrastive clustering methods inevitably extract domain-specific features, which may bring obstacles to the clustering process.

Domain Shifts with Uncertainty

Previous works (Huang and Belongie 2017; Li et al. 2021a) demonstrate that the domain characteristics of data can be captured by feature statistics, i.e., mean and standard deviation of the learned features. Domain Shifts with Uncertainty

(DSU) (Li et al. 2022) treats feature statistics as stochastic variables and reconstructs samples with synthesized feature statistics, imitating the shifts of diverse domains.

To be specific, for encoded features $P \in \mathbb{R}^{B\times U\times H\times W}$ in a mini-batch of size B, where H and W denote the height and width of the feature maps and U is the channel number, the channel-wise mean of features $\mu \in \mathbb{R}^{B\times U}$ and standard deviation $\sigma \in \mathbb{R}^{B\times U}$ can be computed across spatial dimensions independently for each channel:

$$\mu = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} p_{b,u,h,w},$$
$$\sigma^2 = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(p_{b,u,h,w}-\mu)^2. \tag{1}$$

Then, DSU proposes a non-parametric method to compute the variance of the feature mean $\Sigma_\mu \in \mathbb{R}^U$ and that of feature standard deviation $\Sigma_\sigma \in \mathbb{R}^U$, respectively,

$$\Sigma_\mu^2 = \frac{1}{B}\sum_{b=1}^{B}(\mu_{b,:}-\frac{1}{B}\sum_{i=1}^{B}\mu_{i,:})\odot(\mu_{b,:}-\frac{1}{B}\sum_{j=1}^{B}\mu_{j,:})$$
$$\Sigma_\sigma^2 = \frac{1}{B}\sum_{b=1}^{B}(\sigma_{b,:}-\frac{1}{B}\sum_{i=1}^{B}\sigma_{i,:})\odot(\sigma_{b,:}-\frac{1}{B}\sum_{j=1}^{B}\sigma_{j,:}), \tag{2}$$

where $\odot$ is the Hadamard product. The distribution of the new feature statistics can be acquired by the re-parameterization trick, whose mean β and standard deviation γ are formulated as follows:

$$\beta_{i,:} = \mu_{i,:} + \epsilon_\mu \Sigma_\mu, \quad \epsilon_\mu \sim \mathcal{N}(\mathbf{0},\mathbf{1}),$$
$$\gamma_{i,:} = \sigma_{i,:} + \epsilon_\sigma \Sigma_\sigma, \quad \epsilon_\sigma \sim \mathcal{N}(\mathbf{0},\mathbf{1}), \tag{3}$$

where ϵ_μ and ϵ_σ both follow the standard Gaussian distribution. Therefore, the generated feature is

$$\mathrm{DSU}(P_{:,:,i,j}) = \gamma \odot \left(\frac{P_{:,:,i,j}-\mu}{\sigma}\right) + \beta. \tag{4}$$

In this way, DSU allows the trained model to access images from more diverse domains, improving generalization performance.

Generalized Contrastive Clustering

As shown in Figure 2, the proposed Generalized Contrastive Clustering (GeCC) is designed with a Cluster-guided Domain Shifts Modeling module (CDSM), instance representation and cluster assignment contrastive modules. The network architecture of GeCC consists of a weight-sharing encoder $f_E(\cdot)$ and two projectors, i.e., the instance-level projector $f_I(\cdot)$ and cluster-level projector $f_C(\cdot)$. Given a mini-batch dataset $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_B\}$ with B unlabeled images from K different categories, two augmented versions X^1 and X^2 can be obtained by stochastic augmentation. The encoder extracts embedding $P^1 = f_E(X^1)$ and $P^2 = f_E(X^2)$, where one of two branches is integrated with CDSM to generate a reference embedding P^d. Then, $f_I(\cdot)$ and $f_C(\cdot)$ map these embedding to latent representation $Z^m \in \mathbb{R}^{B\times M}$ and cluster assignment $C^m \in \mathbb{R}^{K\times B}$, respectively, where $m \in \{1, 2, d\}$.

Algorithm 1: Details of CDSM

Input: Encoded embedding $P^1 \in \mathbb{R}^{B\times U\times H\times W}$, cluster representation $C^1 \in \mathbb{R}^{K\times B}$;
Output: Reference embedding $P^d \in \mathbb{R}^{B\times U\times H\times W}$;
Obtain pseudo label $\hat{y} = argmax_k [C^1]_k$, group embedding into $\{\hat{P}_1^1, ...\hat{P}_1^K\}$
while $1 \le i \le K$ **do**
  Compute μ^k, θ^k by substituting $\hat{P}_i^k$ into Eq. (1);
  Compute $\Sigma\mu^k \in \mathbb{R}^U$, $\Sigma\sigma^k \in \mathbb{R}^U$ by Eq. (5);
  Compute synthetic feature statistics β^k and γ^k for k-th cluster according to Eq. (3);
  Obtain distribution $CDSM(\hat{P}_i^k)$ by substituting β^k and γ^k into Eq. (4);
  Generate embedding P_d^i by sampling from distribution $CDSM(\hat{P}_i^k)$;
  $i = i + 1$;
end while
$P^d = \{P_d^1, ...P_d^K\}$;

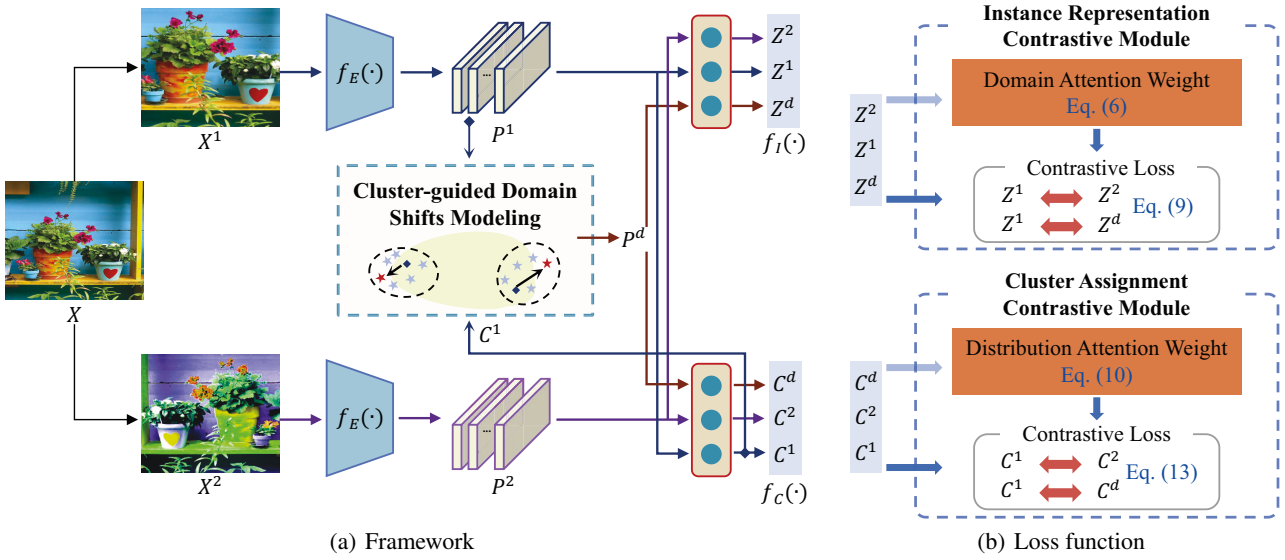
Cluster-guided Domain Shifts Modeling

As mentioned above, the vanilla contrastive clustering may get stuck in the domain-specific information induced by data augmentation. To mitigate such impact, we intend to use DSU to improve the generalization of clustering network with accessing more diverse domains. However, DSU synthesizes new instances based on feature statistics, which neglects the cluster-related information during the modeling and cannot be adapted to contrastive clustering directly. Therefore, CDSM is formulated to make DSU more suitable for clustering task.

Specifically, for newly generated samples, the clustering information is inserted to reduce the semantic difference of samples in the same cluster. Instead of treating all samples with the same standard deviation, we compute the feature statistics for each cluster. According to pseudo label $\hat{\boldsymbol{y}}$, the instances are grouped into K subsets $\{\hat{P}^1, ...\hat{P}^K\}$, where the pseudo label for b-th sample is derived from $\hat{y}_b = \underset{k}{argmax}[\boldsymbol{c}_b^1]_k$, i.e., the index of the largest element in cluster indicator vector $\boldsymbol{c}_b^1$. Then, the mean μ^k and standard deviation θ^k for each cluster are obtained by substituting feature matrix $\hat{P}^k$ into Eq. (1). Then, the variance of the feature mean $\Sigma_{\mu^k} \in \mathbb{R}^U$ and that of feature standard deviation $\Sigma_{\sigma^k} \in \mathbb{R}^U$ can be computed by

$$\Sigma_{\mu^k}^2 = \frac{1}{b_k}\sum_{b=1}^{b_k}(\mu_{b,:}^k-\frac{1}{b_k}\sum_{i=1}^{b_k}\mu_{i,:}^k)\odot(\mu_{b,:}^k-\frac{1}{b_k}\sum_{j=1}^{b_k}\mu_{j,:}^k)$$
$$\Sigma_{\sigma^k}^2 = \frac{1}{b_k}\sum_{b=1}^{b_k}(\sigma_{b,:}^k-\frac{1}{b_k}\sum_{i=1}^{b_k}\sigma_{i,:}^k)\odot(\sigma_{b,:}^k-\frac{1}{b_k}\sum_{j=1}^{b_k}\sigma_{j,:}^k), \tag{5}$$

in which b_k equals to the number of samples in $\hat{P}^k$. Then, we can obtain the cluster-guided statistics by Eq. (3). The generated feature for samples in k-th cluster can be acquired by substituting cluster-guided statistics into the Eq. (4).



(a) Framework (b) Loss function

Figure 2: The flowchart of Generalized Contrastive Clustering with domain shifts modeling (GeCC). (a) Framework: GeCC is equipped with a weight-sharing encoder $f_E(\cdot)$ to extract data embedding from augmented views. Then, instance-level projector $f_I(\cdot)$ and cluster-level projector $f_C(\cdot)$ consisting of two fully connected layers are used to obtain instance representations and cluster assignments respectively. Furthermore, a cluster-guided domain shifts modeling (CSDM) module is built upon one of the augmented view, synthesizing a reference embedding with more variant domains. (b) Loss function: Based on the newly-generated reference and augmented views, GeCC computes the well-designed attention and distribution weights. Instance representation and cluster assignment contrastive modules incorporate the attention and distribution weights respectively, formulating the objective function.

Consequently, in addition to P^1 and P^2, we integrate the branch that learns P^1 with CDSM to obtain the reference embedding $P^d = CDSM(P^1, C^1)$. In this way, P^d is equipped with different domain knowledge while retains cluster-related information. The specific flowchart is provided in Algorithm 1. Because the quality of clustering matters for the subsequent contrastive learning module, even the overall model. Given the lower accuracy of the pseudo label during the initial stage, CDSM is performed at a probability $t = (epoch//100 + 1) \times 0.1$ after each convolution block of the encoder ResNet, in which $epoch$ is the current training iteration and $//$ denotes the modulo operation. Based on the probability setting, probability will be enlarged in the latter training stage. Then more accurate clustering information would contribute to the formation of a semantic reference view, ultimately facilitating the clustering process.

Instance Representation Contrastive Module

Generally, because the variation of predefined augmentation is limited, the augmented views share partially identical information covering their domain knowledge. When the contrastive learning is performed to extract consistent content, the learned representation inevitably involves the augmentation-specific features, which will deteriorate the clustering performance. To increase the proportion of cluster-related information and reduce the undesirable effects of data augmentation, the network gets access to three views of embedding P^1, P^2 and P^d and pays more attention to learn from more diverse domains.

Therefore, we define the domain attention weight $W_z \in \mathbb{R}^{B\times B}$ for instance representation to measure the relative domain difference between newly-generated reference and augmented view, with element

$$w_{ij}^z = 1 + \frac{\hat{D}\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^d\right)}{\hat{D}\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2\right) + \hat{D}\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^d\right)}, \tag{6}$$

in which $\hat{D}(\cdot,\cdot)$ denotes the normalized Euclidean distance. The normalization operation can be expressed as

$$\hat{D}\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2\right) = \frac{D\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2\right)}{max_{k=1}^{B} D\left(\boldsymbol{z}_i^1, \boldsymbol{z}_k^2\right)}. \tag{7}$$

When the domain difference for i-th sample with j-th reference sample is larger than that with j-th augmented sample, i.e., $\hat{D}(\boldsymbol{z}_i^1, \boldsymbol{z}_j^d) > \hat{D}(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2)$, w_{ij}^z will be large.

Then, the instance contrastive loss for a given sample $\boldsymbol{x}_i$ can be computed in the form of

$$l_{i,12}^z = \sum_{m=\{1,2\}} log \frac{w_{ii}^z e^{s\left(\boldsymbol{z}_i^1, \boldsymbol{z}_i^2\right)/\tau_i}}{\sum_{j=1}^{B} w_{ij}^z \left[e^{s\left(\boldsymbol{z}_i^m, \boldsymbol{z}_j^m\right)/\tau_i} + e^{s\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2\right)/\tau_i}\right]},$$

$$l_{i,1d}^z = \sum_{m=\{1,d\}} log \frac{w_{ii}^z e^{s\left(\boldsymbol{z}_i^1, \boldsymbol{z}_i^d\right)/\tau_i}}{\sum_{j=1}^{B} w_{ij}^z \left[e^{s\left(\boldsymbol{z}_i^m, \boldsymbol{z}_j^m\right)/\tau_i} + e^{s\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^d\right)/\tau_i}\right]},$$

$$l_i^z = l_{i,12}^z + l_{i,1d}^z, \tag{8}$$

where τ_i denotes the instance-level temperature coefficient and $s(\cdot,\cdot)$ measures the cosine similarity. The overall contrastive loss for instance representation is computed as

$$L_{ins} = -\frac{1}{2B}\sum_{i=1}^{B} l_i^z. \tag{9}$$

When minimizing Eq. (9), $\boldsymbol{z}_i^1$ will stay away from similar domain information undelying $\boldsymbol{z}_j^2$ and approach new domain knowledge from $\boldsymbol{z}_j^d$. Consequently, the network will involve knowledge from more diverse domains and disentangle from domain-specific information.

Delving into the underlying mechanism, the Euclidean distance for a negative pair $\{\boldsymbol{z}_i^1, \boldsymbol{z}_j^2\}$ equals to $\|2 - 2s\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2\right)\|_F$ because the representation is normalized to the unit sphere. When w_{ij}^z is large, $s\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^2\right)$ is greater than $s\left(\boldsymbol{z}_i^1, \boldsymbol{z}_j^d\right)$. After contrastive learning is performed to minimize the similarity of negative pairs, z_i^1 will be pushed away from z_j^2 compared to z_j^d. On the contrary, when w_{ij}^z is small, i.e., the domain difference between z_i^1 and z_j^d is more subtle than that between z_i^1 and z_j^2, z_i^1 will be kept away from z_j^d. In summary, the instance projector takes the newly-generated data as a reference and dynamically approaches more variant domains. The proof is detailed in the Supplementary Material.

Cluster Assignment Contrastive Module

Since the data embedding from the three views shares the same category information, cluster-level contrastive loss is posed on pairs from cluster assignments, i.e., $\{C^1, C^2\}$ and $\{C^1, C^d\}$, to align the label distributions. Namely, for each cluster, the sample distributions under different views will be aligned while distributions of different clusters will be distinct. When the assignment of i-th cluster blends with that of j-th cluster under augmentation, similarity for the pair of cluster assignments should be diminished. Likewise, we utilize the discrepancy between cluster assignments from different views and introduce the distribution attention weight $W_c \in \mathbb{R}^{K\times K}$ whose element is

$$w_{ij}^c = 1 + \frac{\hat{D}\left(\boldsymbol{c}_i^1, \boldsymbol{c}_j^d\right)}{\hat{D}\left(\boldsymbol{c}_i^1, \boldsymbol{c}_j^2\right) + \hat{D}\left(\boldsymbol{c}_i^1, \boldsymbol{c}_j^d\right)}. \tag{10}$$

And then the cluster assignment contrastive loss for i-th cluster can be formulated as

$$l_{i,12}^c = \sum_{m=\{1,2\}} log\frac{w_{ii}^c e^{s\left(\boldsymbol{c}_i^1, \boldsymbol{c}_i^2\right)/\tau_c}}{\sum_{j=1}^{K} w_{ij}^c \left[e^{s\left(\boldsymbol{c}_i^m, \boldsymbol{c}_j^m\right)/\tau_c} + e^{s\left(\boldsymbol{c}_i^1, \boldsymbol{c}_j^2\right)/\tau_c}\right]},$$

$$l_{i,1d}^c = \sum_{m=\{1,d\}} log\frac{w_{ii}^c e^{s\left(\boldsymbol{c}_i^1, \boldsymbol{c}_i^d\right)/\tau_c}}{\sum_{j=1}^{K} w_{ij}^c \left[e^{s\left(\boldsymbol{c}_i^m, \boldsymbol{c}_j^m\right)/\tau_c} + e^{s\left(\boldsymbol{c}_i^1, \boldsymbol{c}_j^d\right)/\tau_c}\right]},$$

$$l_i^c = l_{i,12}^c + l_{i,1d}^c, \tag{11}$$

where τ_c is the temperature coefficient for cluster assignment.

Algorithm 2: Training Procedures of GeCC

Input: Dataset X, training epochs E, batch size B, cluster number K, probability t, temperature parameter τ_i and τ_c;
Output: Cluster assignments;

// Training
while epoch $< E$ **do**
  Sample a mini-batch $\{\boldsymbol{x}_i\}_{i=1}^B$ from X;
  Obtain two augmentations $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_i^2$;
  Compute instance and cluster representations by
  $\boldsymbol{p}_i^1 = f_E(\boldsymbol{x}_i^1)$, $\boldsymbol{p}_i^2 = f_E(\boldsymbol{x}_i^2)$
  $\boldsymbol{z}_i^1 = f_I(\boldsymbol{p}_i^1)$, $\boldsymbol{z}_i^2 = f_I(\boldsymbol{p}_i^2)$
  $\boldsymbol{c}_i^1 = f_C(\boldsymbol{p}_i^1)$, $\boldsymbol{c}_i^2 = f_C(\boldsymbol{p}_i^2)$;
  Generate reference embedding $\boldsymbol{p}^d$ by performing cluster-guided domain shifts modeling at probability $t = 0.1 \times (\text{epoch}//100 + 1)$;
  Output instance and cluster representations by $\boldsymbol{z}_i^d = f_I(\boldsymbol{p}_i^d)$, $\boldsymbol{c}_i^d = f_C(\boldsymbol{p}_i^d)$;
  Compute attention weights W_z and W_c;
  Compute loss for instance representations by Eq. (9);
  Compute loss for cluster assignments by Eq. (13);
  Update network through minimizing Eq. (14);
  epoch = epoch + 1 ;
end while
// Testing
Extract embedding $\boldsymbol{p}_i = f_E(\boldsymbol{x}_i)$;
Output cluster assignment $\boldsymbol{c}_i = f_C(\boldsymbol{p}_i)$;
Acquire predicted label $y_i = \underset{k}{argmax}[\boldsymbol{c}_i]_k$;

When w_{ij}^c is large, the variation between (c_i^1, c_j^2) is less than that of (c_i^1, c_j^d), which means that samples assigned to the i-th category have more overlap with j-th category under augmentation. The negative pair (c_i^1, c_j^2) with large distribution attention will be punished more heavily, such that the distributions for different clusters will be discriminative. Vice verse, if w_{ij}^c is small, the overlap for (c_i^1, c_j^d) will be reduced.

To avoid the trivial solution that most instances are assigned to the same cluster, the regularization term is composed on three cluster assignment matrices, i.e.,

$$\begin{aligned} R(C) = \sum_{i=1}^{K} \big[& H(\boldsymbol{c}_i^1) logH\left(\boldsymbol{c}_i^1\right) + H(\boldsymbol{c}_i^d) logH\left(\boldsymbol{c}_i^d\right) \\ & + H(\boldsymbol{c}_i^2) logH\left(\boldsymbol{c}_i^2\right)\big], \end{aligned} \tag{12}$$

in which $H(\boldsymbol{c}_i^m) = \frac{\sum_{b=1}^{B} c_{ib}^m}{\|C^m\|_1}, m \in \{1, 2, d\}$. The overall cluster contrastive loss is computed as

$$L_{cls} = -\frac{1}{2K}\sum_{i=1}^{K} l_i^c + R(C). \tag{13}$$

Objective Function

The optimization of our model is a one-stage process shown in Algorithm 2. The overall objective function consists of the

Datasets	Size (n)	Classes (c)	Split
CIFAR-10	60,000	10	Train+Test
CIFAR-100	60,000	20	Train+Test
ImageNet-10	13,000	10	Train
ImageNet-Dogs	19,500	15	Train

Table 1: Details of the datasets.

instance representation and cluster assignment contrastive losses, i.e.,

$$L = \lambda_1 L_{ins} + \lambda_2 L_{cls}. \tag{14}$$

In this paper, λ_1 and λ_2 are set to 1 on all experimental datasets.

Experiments

In this section, the performance of the proposed method GeCC is evaluated by comparing with state-of-the-art methods on four benchmark datasets.

Datasets

We evaluate the proposed method on four image datasets, whose details are shown in Table 1. CIFAR-10 consits of 10 categories and CIFAR-100 takes 20 super-classes as the label. ImageNet-10 and ImageNet-Dogs contains 10 randomly selected subjects and 15 types of dogs, respectively.

Implementation Details

We adopt the same backbone encoder ResNet-34 for all datasets except ImageNet-10, which shows better performance with ResNet-18. The encoder extracts an embedding of size 512 for each image. The instance and cluster projector both consist of two fully connected layers. The row dimension of the instance representation is set to 128 and the column dimension of the cluster assignment equals to the number of categories. We apply Adam optimizer with learning rate of 0.0003 to simultaneously optimize the backbone encoder and two projectors. The size of mini-batch is set as 256. We train the network for $E = 1000$ epochs. The temperature for instance and cluster representation contrastive are fixed to 0.5 and 1.0 on all datasets, respectively.

As for the predefined augmentation, we first resize all input images to the size of 224×224, then perform flip, color jitter, grayscale and GaussianBlur in sequence. For small image datasets including CIFAR-10 and CIFAR-100, we leave out the GaussianBlur augmentation.

Three widely-used metrics, including Normalized Mutual Information (NMI), Accuracy (ACC) and Adjusted Rand Index (ARI), are utilized to assess the clustering performance. Values closer to 1 indicate higher performance.

Experimental Results

We compare the proposed method with 19 representative clustering approaches, including 4 traditional clustering methods: k-means (MacQueen et al. 1967), SC (Shi and Malik 2000), AC (Gowda and Krishna 1978), NMF (Cai et al. 2009), and 10 deep clustering networks: AE (Bengio et al. 2006), DAE (Vincent et al. 2010), GAE (Radford, Metz, and Chintala 2015), DECNN (Zeiler et al. 2010), VAE (Kingma and Welling 2013), JULE (Yang, Parikh, and Batra 2016), DEC (Xie, Girshick, and Farhadi 2016), DAC (Chang et al. 2017), DCCM (Wu et al. 2019), IIC (Ji, Henriques, and Vedaldi 2019), and 5 state-of-the-art contrastive clustering methods: PICA (Huang, Gong, and Zhu 2020), DRC (Zhong et al. 2020), CC (Li et al. 2021b), DeepCluE (Huang et al. 2024), DivClust (Metaxas, Tzimiropoulos, and Patras 2023).

The clustering performance is firstly evaluated on four benchmark datasets. According to the results shown in Table 2, we can draw the following conclusions.

- It can be seen from the results that the deep learning-based methods clearly outperform the traditional clustering methods, mainly due to the powerful representation learning capabilities of neural networks. For instance, the worst-performing baseline among deep learning-based methods achieves 27.20% clustering accuracy in CIFAR-10, while traditional methods score below 25%.
- Among the deep learning-based methods, we can see that contrastive learning-based methods excel in clustering performance. Take ImageNet-10 for example, all contrastive methods achieve accuracy much higher than 80%, while other deep clustering methods achieve accuracy below 72%. It demonstrates that the contrastive mechanism significantly contributes to learning clustering-friendly representation.
- Overall, the proposed method significantly outperforms other compared methods on most datasets in terms of three evaluation metrics. In particular, for ImageNet-Dogs, compared to the second best method DivClust, our method achieves performance improvements of up to 3.35%, 7.38% and 4.53% in terms of NMI, ACC and ARI, respectively. DeepCluE based on CC performs ensemble clustering from the output of multiple layers in the network, achieving the best results on CIFAR-100. However, our proposed method still achieves comparable performance without the help of additional clustering models.

The above results can well demonstrate the effectiveness and superiority of our proposed method.

Ablation Experiments

In this section, ablation studies are carried out to demonstrate the importance and effectiveness of the cluster-guided domain shifts modeling module and well-designed attention weights. The experimental results are shown in Table 3.

The necessity of introducing cluster-guided domain shifts is first assessed. '*w* Aug' and '*w* DSU' denote that replacing the cluster-guided domain shifts with random augmentation and DSU, respectively. As we can see, generating reference view using domain shifts marginally outperforms random augmentation. Comparing '*w* DSU' with 'CC', we can see that the clustering performance degrades slightly for small image collections including CIFAR-10 and CIFAR-100. This may be due to the DSU confusing semantic information without considering cluster assignment. Moreover,

Datasets	CIFAR-10			CIFAR-100			ImageNet-10			ImageNet-Dogs		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
K-means	0.0870	0.2290	0.0490	0.0840	0.1300	0.0280	0.1190	0.2410	0.0570	0.0550	0.1050	0.0200
SC	0.1030	0.2470	0.0850	0.0900	0.1360	0.0220	0.1510	0.2740	0.0760	0.0380	0.1110	0.0130
AC	0.1050	0.2280	0.0650	0.0980	0.1380	0.0340	0.1380	0.2420	0.0670	0.0370	0.1390	0.0210
NMF	0.0810	0.1900	0.0340	0.0790	0.1180	0.0260	0.1320	0.2300	0.0650	0.0440	0.1180	0.0160
AE	0.2390	0.3140	0.1690	0.1000	0.1650	0.0480	0.2100	0.3170	0.1520	0.1040	0.1850	0.0730
DAE	0.2510	0.2970	0.1630	0.1110	0.1510	0.0460	0.2060	0.3040	0.1380	0.1040	0.1900	0.0780
GAN	0.2650	0.3150	0.1760	0.1200	0.1510	0.0450	0.2250	0.3460	0.1570	0.1210	0.1740	0.0780
DeCNN	0.2400	0.2820	0.1740	0.0920	0.1330	0.0380	0.1860	0.3130	0.1420	0.0980	0.1750	0.0730
VAE	0.2450	0.2910	0.1670	0.1080	0.1520	0.0400	0.1930	0.3340	0.1680	0.1070	0.1790	0.0790
JULE	0.1920	0.2720	0.1380	0.1030	0.1370	0.0330	0.1750	0.3000	0.1380	0.0540	0.1380	0.0280
DEC	0.2570	0.3010	0.1610	0.1360	0.1850	0.0500	0.2820	0.3810	0.2030	0.1220	0.1950	0.0790
DAC	0.3960	0.5220	0.3060	0.1850	0.2380	0.0880	0.3940	0.5270	0.3020	0.2190	0.2750	0.1110
DCCM	0.4960	0.6230	0.4080	0.2850	0.3270	0.1730	0.6080	0.7100	0.5550	0.3210	0.3830	0.1820
IIC	-	0.6170	-	-	0.2570	-	-	-	-	-	-	-
PICA	0.5910	0.6960	0.5120	0.3100	0.3370	0.1710	0.8020	0.8700	0.7610	0.3520	0.3520	0.2010
DRC	0.6210	0.7270	0.5470	0.3560	0.3670	0.2080	0.8300	0.8840	0.7980	0.3840	0.3890	0.2330
CC	0.7050	0.7900	0.6370	0.4310	0.4290	0.2660	0.8590	0.8930	0.8220	0.4450	0.4290	0.2740
DeepCluE	0.7270	0.7640	0.6460	**0.4720**	**0.4570**	**0.2880**	0.8820	0.9240	0.8560	0.4480	0.4160	0.273
DivClust	0.7240	0.8190	0.6810	0.4400	0.4370	0.2830	0.8910	0.9360	0.8780	0.5160	0.5290	0.3760
Ours	**0.7559**	**0.8440**	**0.7108**	0.4419	0.4518	0.2878	**0.9092**	**0.9615**	**0.9167**	**0.5495**	**0.6028**	**0.4213**

Table 2: Clustering performance of various methods on four datasets. The best results are in boldfaced and the suboptimal results are underlined.

Datasets	CIFAR-10			CIFAR-100			ImageNet-10			ImageNet-Dogs		
Metrics	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI
CC	0.7050	0.7900	0.6370	0.4310	0.4290	0.2660	0.8590	0.8930	0.8220	0.4450	0.4290	0.2740
w Aug	0.6858	0.7787	0.6197	0.4349	0.4200	0.2789	0.8475	0.8940	0.8168	0.4600	0.4650	0.3034
w DSU	0.6943	0.7893	0.6734	0.4230	0.4105	0.2564	0.8521	0.9057	0.8260	0.5158	0.5453	0.3800
w/o W	0.6255	0.7058	0.5348	0.4330	0.4371	0.2809	0.8543	0.8944	0.8213	0.4369	0.4385	0.2789
w W_c	0.5831	0.6530	0.4752	0.3962	0.3911	0.2359	0.8111	0.8672	0.7737	0.3461	0.3675	0.2039
w W_z	0.6141	0.6710	0.5114	0.4374	0.4427	0.2768	0.8422	0.8872	0.8090	0.4292	0.4455	0.2733
Ours	**0.7559**	**0.8440**	**0.7108**	**0.4419**	**0.4518**	**0.2878**	**0.9092**	**0.9615**	**0.9167**	**0.5495**	**0.6028**	**0.4213**

Table 3: The clustering performance for ablation experiments on four datasets.

the performance gain for ImageNet datasets is limited. Compared with the proposed method, it can be seen that the incorporation of clustering information significantly boosts the performance. The results demonstrate that the cluster-guided domain shifts modeling module effectively preserves the key category information while injecting more variational features within the generated reference.

We also conduct ablation studies on the importance of attention weights. '*w/o W*' denotes performing contrastive learning without weighting strategy. '*w* W_c' and '*w* W_z' denote the incorporation of distribution and domain attention weights derived from cluster assignments and instance representations, respectively. It can be seen that without the weighting strategy, the performance improvement on most datasets is limited. In addition, the results deteriorate significantly when only the cluster assignment contrastive module utilizes the distribution attention weights. By comparison, the proposed method achieves state-of-the-art performance by leveraging the attention weights in both instance and cluster representation learning, demonstrating that the proposed method can better integrate the cluster-related information from the available views.

Conclusion

This paper proposed a novel Generalized Contrastive Clustering with domain shifts modeling (GeCC) method that maximizes the extraction of cluster-related information under augmentation. GeCC integrates a cluster-guided domain shift modeling module to synthesize reference views with variant domains. This module facilitates the computation of well-designed attention weights, which are then employed to guide the contrastive learning process. By optimizing the contrastive loss over instance representations and cluster assignments, the model disentangles itself from the domain-specific information inherent in augmentation and mines the cluster-related information across different views. Extensive experimental results on four benchmark datasets demonstrate the dominant performance of our proposed method.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62376162, 62173235, 62473266, 62176160 and U22B2035, in part by Guangdong Basic and Applied Basic Research Foundation under Grants 2024A1515010205, 2024B1515020059 and 2022A1515010146, in part by Shenzhen Science and Technology Program under Grants RCYX20221008092922051 and JCYJ20230808105802006, and in part by the (Key) Project of Department of Education of Guangdong Province under Grant 2022ZDZX1022.

References

Asano, Y. M.; Rupprecht, C.; and Vedaldi, A. 2019. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*.

Bengio, Y.; Lamblin, P.; Popovici, D.; and Larochelle, H. 2006. Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, 19.

Cai, D.; He, X.; Wang, X.; Bao, H.; and Han, J. 2009. Locality preserving nonnegative matrix factorization. In *Twenty-first International Joint Conference on Artificial Intelligence*.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision*, 132–149.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, 5879–5887.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

Chen, Y.; Wu, W.; Ou-Yang, L.; Wang, R.; and Kwong, S. 2024. GRESS: Grouping Belief-Based Deep Contrastive Subspace Clustering. *IEEE Transactions on Cybernetics*.

Chen, Y.-J.; Wu, W.-H.; and Ou-Yang, L. 2023. Contrastive Subspace Clustering with Dissimilarity Regularization. In *2023 International Conference on Machine Learning and Cybernetics (ICMLC)*, 128–133. IEEE.

Gowda, K. C.; and Krishna, G. 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10(2): 105–112.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33: 21271–21284.

He, Y.; Chen, X.; Tu, N. H.; and Luo, J. 2023. Deep Multi-Constraint Soft Clustering Analysis for single-cell RNA-seq data via zero-inflated autoencoder embedding. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Huang, D.; Chen, D.-H.; Chen, X.; Wang, C.-D.; and Lai, J.-H. 2024. Deepclue: Enhanced deep clustering via multi-layer ensembles in neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Huang, J.; Gong, S.; and Zhu, X. 2020. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8849–8858.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.

Ji, X.; Henriques, J. F.; and Vedaldi, A. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Vomputer Vision*, 9865–9874.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, B.; Wu, F.; Lim, S.-N.; Belongie, S.; and Weinberger, K. Q. 2021a. On feature normalization and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12383–12392.

Li, X.; Dai, Y.; Ge, Y.; Liu, J.; Shan, Y.; and DUAN, L. 2022. Uncertainty Modeling for Out-of-Distribution Generalization. In *International Conference on Learning Representations*.

Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021b. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 8547–8555.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11174–11183.

Liu, Y.; Wang, Y.; Chen, Y.; Dai, W.; Li, C.; Zou, J.; and Xiong, H. 2023a. Promoting semantic connectivity: Dual nearest neighbors contrastive learning for unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3510–3519.

Liu, Y.; Yang, X.; Zhou, S.; Liu, X.; Wang, S.; Liang, K.; Tu, W.; and Li, L. 2023b. Simple Contrastive Graph Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.

Liu, Z.; Li, Y.; Yao, L.; Wang, X.; and Nie, F. 2022. Agglomerative Neural Networks for Multiview Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7): 2842–2852.

MacQueen, J.; et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 281–297. Oakland, CA, USA.

Metaxas, I. M.; Tzimiropoulos, G.; and Patras, I. 2023. DivClust: Controlling Diversity in Deep Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3418–3428.

Mrabah, N.; Amar, M. M.; Bouguessa, M.; and Diallo, A. B. 2023. Toward convex manifolds: a geometric perspective for deep graph clustering of single-cell RNA-seq data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4855–4863.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Ren, Y.; Pu, J.; Yang, Z.; Xu, J.; Li, G.; Pu, X.; Yu, P. S.; and He, L. 2022. Deep clustering: A comprehensive survey. *arXiv preprint arXiv:2210.04142*.

Shi, J.; and Malik, J. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 888–905.

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

Wang, X.; and Qi, G.-J. 2022. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5549–5560.

Wu, J.; Long, K.; Wang, F.; Qian, C.; Li, C.; Lin, Z.; and Zha, H. 2019. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8150–8159.

Wu, L.; Yuan, L.; Zhao, G.; Lin, H.; and Li, S. Z. 2023. Deep Clustering and Visualization for End-to-End High-Dimensional Data Analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8543–8554.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, 478–487. PMLR.

Xue, Z.; Du, J.; Zheng, C.; Song, J.; Ren, W.; and Liang, M. 2021. Clustering-Induced Adaptive Structure Enhancing Network for Incomplete Multi-View Data. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 3235–3241.

Yang, J.; Parikh, D.; and Batra, D. 2016. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5147–5156.

Yang, X.; Deng, C.; Wei, K.; Yan, J.; and Liu, W. 2020. Adversarial learning for robust deep clustering. *Advances in Neural Information Processing Systems*, 33: 9098–9108.

Yeh, C.-H.; Hong, C.-Y.; Hsu, Y.-C.; Liu, T.-L.; Chen, Y.; and LeCun, Y. 2022. Decoupled contrastive learning. In *European Conference on Computer Vision*, 668–684. Springer.

Zeiler, M. D.; Krishnan, D.; Taylor, G. W.; and Fergus, R. 2010. Deconvolutional networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2528–2535. IEEE.

Zhao, J.; Lu, D.; Ma, K.; Zhang, Y.; and Zheng, Y. 2020. Deep image clustering with category-style representation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 54–70. Springer.

Zhong, H.; Chen, C.; Jin, Z.; and Hua, X.-S. 2020. Deep robust clustering by contrastive learning. *arXiv preprint arXiv:2008.03030*.