applied sciences

MDPI

Article

Deep Image Clustering Based on Label Similarity and Maximizing Mutual Information across Views

Feng Peng [1] and Kai Li [1,2,*]

[1] School of Cyber Security and Computer, Hebei University, Baoding 071000, China
[2] Hebei Machine Vision Engineering Research Center, Baoding 071000, China
* Correspondence: likai@hbu.edu.cn

Abstract: Most existing deep image clustering methods use only class-level representations for clustering. However, the class-level representation alone is not sufficient to describe the differences between images belonging to the same cluster. This may lead to high intra-class representation differences, which will harm the clustering performance. To address this problem, this paper proposes a clustering model named Deep Image Clustering based on Label Similarity and Maximizing Mutual Information Across Views (DCSM). DCSM consists of a backbone network, class-level and instance-level mapping block. The class-level mapping block learns discriminative class-level features by selecting similar (dissimilar) pairs of samples. The proposed extended mutual information is to maximize the mutual information between features extracted from views that were obtained by using data augmentation on the same image and as a constraint on the instance-level mapping block. This forces the instance-level mapping block to capture high-level features that affect multiple views of the same image, thus reducing intra-class differences. Four representative datasets are selected for our experiments, and the results show that the proposed model is superior to the current advanced image clustering models.

Keywords: image clustering; extended mutual information; unsupervised learning

Citation: Peng, F.; Li, K. Deep Image Clustering Based on Label Similarity and Maximizing Mutual Information across Views. *Appl. Sci.* **2023**, *13*, 674. https://doi.org/10.3390/app13010674

Academic Editor: Huibing Wang

Received: 14 November 2022
Revised: 27 December 2022
Accepted: 28 December 2022
Published: 3 January 2023

1. Introduction

Clustering is one of the basic tasks in machine learning, computer vision and other fields. It aims to assign unlabeled samples to different clusters so that samples in the same cluster are similar to each other while samples in different clusters are as dissimilar as possible [1]. With the development of deep learning, the need for large-scale labeled datasets is a major obstacle to the application of deep learning in many scenarios. However, manually labeling these datasets is costly and time-consuming. To make better use of unlabeled data, unsupervised clustering has gradually attracted widespread attention. The main research directions of clustering at this stage can be divided into image, text [2], multi-view clustering [3,4] and so on. Among these data, the simplest and most intuitive feeling is image data. Traditional clustering methods, such as k-means [5], Gaussian mixture model [6], density peak clustering [7] and other methods are limited when they meet high-dimensional data, such as images. The reason is the failure of their similarity measures. With the help of fuzzy set theory and its extensions, such as intuitionistic fuzzy sets [8] and others, many novel similarity measures have emerged and made some progress compared to conventional similarity measures. However, a more intuitive solution attracts attention. These methods [9,10] map high-dimensional data into a low-dimensional representation space and perform clustering in the low-dimensional space. These low-dimensional representations are not informative enough, resulting in no significant improvement in the clustering performance. Therefore, image clustering is still a challenging task.

Deep learning has been gradually combined with computer vision. Due to the powerful representation extraction ability of deep neural networks, it has achieved success

in the fields of classification [11,12], object recognition [13–15] and image retrieval [16]. Many scholars turn their attention to deep clustering [17]. Deep clustering refers to the process of automatically learning the data representations and guiding clustering through neural networks. Widely used network architecture includes Convolutional Neural Networks [12,18], Autoencoders [19] and Variational Autoencoders [20]. Following the success of Generative Adversarial Networks [21], many works [22–25] introduced the idea of GANs into representation learning and achieved good results. Combining neural networks with traditional clustering algorithms is an effective solution, which consists of two stages. In the first stage, the neural network captures the low-dimensional representations of the samples. In the second stage, the learned low-dimensional representations are served as the input for traditional clustering methods. Some progress has been made compared to traditional methods.

Recently, many unsupervised deep learning methods have been proposed to learn clustering. By introducing data augmentation, most of the latest methods look into deep clustering from the perspective that the original image and its transformation should share similar semantic clustering assignments [26–28]. These methods have achieved some progress on complex natural image datasets. However, the above methods have some limitations: (1) Since generative networks based on autoencoders require pixel-level reconstruction. It is easy to capture redundant image information, such as background, color and other trivial information; (2) These methods only learn class-level representation. However, the instance-level representation could be quite different even if they are assigned to the same cluster since the softmax function is only sensitive to the maximum value. In other words, class-level representation alone is not enough to describe the differences between images belonging to the same cluster.

In order to solve the above limitations, this paper proposes a new image clustering model, DCSM. Based on the class-level mapping block (C-block), DCSM adds an instance-level mapping block (IR-block). The two blocks share the same backbone network, and the backbone network acts as an information transmission channel between the two blocks. First, the C-block learns discriminative class-level representation by selecting pairwise similar (dissimilar) samples. Second, errors will occur when constructing similar labels of sample pairs, so select high-confidence cluster-standard samples to correct network parameters because samples with high-confidence predictions are often already assigned to the proper cluster. Third, the proposed extended MI maximizes the MI between the global-local and local-local features of multiple views, forcing the IR block to capture the essential information. This can narrow the intra-class diversities in the representation feature space. Fourth, by introducing the theory of entropy maximization, the probability distribution of clusters is pushed closer to a uniform distribution to solve the trivial solution problem.

In summary, the major contributions of this paper are:

1. An end-to-end clustering network model is proposed to jointly learn representations and cluster assignments while preventing the network from falling into trivial solutions.
2. An extended MI method is proposed. It calculates the MI between features from multiple views of the same image, forcing the instance-level feature to capture invariant essential information and thus reducing the intra-class diversities in the representation feature space.

The rest of this paper is organized as follows. In Section 2, some related works on deep clustering and mutual information are outlined. Section 3 introduces the proposed method. Section 4 discusses the experimental results of DCSM, and Section 5 concludes our proposed method.

2. Related Work

2.1. Deep Clustering

In various deep architectures, one simple yet effective neural network is the autoencoder (AE). Therefore, many methods applied AE to extract features for clustering. For example, Yang et al. [29] applied AE to learn a K-means-friendly embedding representa-

tion. Fard et al. [30] proposed a Deep K-means approach using AE. Lv et al. [31] proposed a pseudo-label-based deep subspace clustering method. Diallo et al. [32] employed contractive autoencoder and self-augmentation techniques to develop deep clustering models suitable for document datasets. Ren et al. [33] proposed a deep density clustering framework by combining density clustering and AE. Since the convolutional neural network has demonstrated promising performance in image processing tasks, CNN has also been used in deep clustering. Yang et al. [34] used an RNN to implement aggregated clustering on the representations output by CNN and trained both networks in an end-to-end fashion. Sun et al. [35] combined active learning and deep clustering models to manually annotate uncertain and informative sample pairs and used these manually annotated samples to guide the training of CNN.

Different from the above methods using AE or CNN, another unsupervised deep architecture is generative models, such as GAN and VAE. Mukherjee et al. [24] designed the ClusterGAN-2 model to perform clustering in the latent space by optimizing GANs and cluster-specific losses. Wang et al. [25] utilized generated and raw data to train a feature extractor to learn efficient representations. Ntelemis et al. [36] introduced the Sobel operation before the discriminator to enhance the separability of learning features. Xu et al. [1] proposed a variational encoder deep clustering method based on MI maximization and derived a new generalization evidence lower bound.

It is also a good idea to directly design a specific clustering loss according to the desired clustering assumption. Xie et al. [37] use only one encoder when clustering, updating the encoder part of the AE by minimizing the KL divergence loss. Chang et al. [26] transformed the clustering problem into a binary discrimination problem by selecting pairs of similar and dissimilar samples to train the network. Wu et al. [27] mine and exploit associations between unlabeled data from three aspects to study category information, and introduce MI to further help learn more discriminative features. Xu et al. [28] learned cluster assignments by maximizing the mutual information of data to probabilistic features. Gansbeke et al. [38] generated pseudo-labels by mining nearest neighbors in the latent embedding space.

2.2. Mutual Information

Information theory has become an important tool for training deep neural networks. Mutual information quantifies the amount of information obtained by observing one random variable about another random variable. MINE [39] is an MI estimation technique for continuous variables that is scalable in the sample size, trained via backpropagation and is strongly consistent. DIM [40] adopts a simple alternative based on Jensen–Shannon divergence, which is an unbiased estimator that is insensitive to the number of negative samples. More importantly, DIM can also use the local structure of the input to improve the learned embedding representation ability and improve the quality of downstream tasks. RIM [41] introduces a regularization term that penalizes conditional models with complex decision boundaries to produce a reasonable clustering solution. Inspired by RIM, IMSAT [42] utilizes data augmentation to impose invariance on discrete representations by maximizing the mutual information between the data and its representations. IIC [28] constrains the output variables to be discrete variables and performs a simple and accurate mutual information calculation for the discrete variables.

3. Deep Image Clustering under Similarity and Mutual Information

The purpose of this paper is to cluster a set of N images $X = \{x_i\}_{i=1}^{N}$ into K classes by training a neural network without using any annotations. The neural network can be conceptually divided into two parts: a feature module that maps images to instance-level representations, $z_i = f_r(x_i; w_r)$, and a clustering module that maps images to class-level representations, $l_i = f_l(r_{i,j}; w_l)$, where w_r and w_l represent the trainable parameters of IR-block and C-block combined with the backbone network, respectively. This paper uses the output of the C-block to measure the similarity between samples driving the separation between clusters. The IR-block is constrained by the proposed extended mutual information, which

captures the presence of a key object in the image to reduce the differences in intra-class feature representations. In addition, two tasks are designed to ensure the accuracy and stability of our model. They are fine-tuning the network using cluster-standard samples and maximizing the entropy loss. In the following subsections, this paper introduces each part in detail.

3.1. Clustering Module

The relationship of pairwise images is binary, i.e., similar or dissimilar. Based on such observations, the clustering task can be smoothly recast into a binary pairwise classification model [26]. It is natural to use the loss for the classification task instead of the clustering objective function:

$$minL_{DT}(w_l) = \sum_{i=1}^{n}\sum_{j=1}^{n} L(r_{i,j}, g(x_i, x_j; w_l)) \tag{1}$$

where $L_{DT}(r_{i,j}, g(x_i, x_j; w_l))$ represents the cross-entropy loss between the ground-truth similarity $r_{i,j}$ and the estimated similarity $g(x_i, x_j; w_l)$, and w_l indicates the learnable parameter of the clustering module.

In Equation (1), the ground-truth similarity $r_{i,j}$ of samples x_i and x_j is unknown, and this problem needs to be solved. Previous methods used a high threshold to select pairwise similar (dissimilar) samples to assign $r_{i,j}$, where $r_{i,j} = 1$ indicates that samples x_i and x_j belong to the same cluster and $r_{i,j} = 0$ otherwise [26,27]. Due to the existence of the threshold, the similarity relationship between pairwise samples does not satisfy transitivity, i.e., $r_{i,j} = 1 = r_{i,k} \nRightarrow r_{j,k} = 1$.

This can lead to unstable network training. Different from the existing methods, this paper uses K-means as a flexible threshold selection strategy to construct $r_{i,j}$:

$$r_{i,j} = \begin{cases} 1, & \mathrm{c_i} = \mathrm{c}_j \; or \; i = j, \\ 0, & otherwise. \end{cases} \tag{2}$$

$r_{i,j} = 1$ indicates that x_i and x_j belong to the same cluster c or the same sample, $i = j$, and $r_{i,j} = 0$ otherwise. When pairwise samples belong to the same cluster, their similarity relationship must be transitive. The estimated similarity $g(x_i, x_j; w_l)$ can be formulated as:

$$\begin{aligned} &g(x_i, x_j, w_l) = f_l(x_i; w_l) \cdot f_l(x_j; w_l) = l_i \cdot l_j \\ &s.t. \forall i ||l_i||_2 = 1, \, 0 \le l_{ih} \le 1, \, h = 1, \ldots, k. \end{aligned} \tag{3}$$

where f_l is the clustering module, and the operator "$\cdot$" represents the dot product between two label features. If the optimal value of Equation (1) is attained, the learned label features are k-diverse one-hot vectors ideally [26]. Images can be directly clustered based on the learned class-level representations.

3.2. Fine-Tuning through Cluster-Standard Samples

Pseudo-similar labels $r_{i,j}$ are obtained in Section 3.1 to learn discriminative features. However, errors in the construction of $r_{i,j}$ are inevitable. These false $r_{i,j}$ lead to predictions for which the network is less certain. It has been observed in experiments that samples with high-confidence predictions tend to be assigned to the proper cluster. It can be regarded as cluster-standard samples to fine-tune the network's errors caused by incorrect $r_{i,j}$. Cluster-standard samples are selected by setting a large threshold:

$$W_i = \begin{cases} 1, & p_{ij} \ge threshold, \\ 0, & otherwise. \end{cases} \tag{4}$$

If the largest element p_{ij} in the label feature l_i is greater than the threshold, it can be regarded as a cluster-standard sample, $W_i = 1$; otherwise, it is not, $W_i = 0$. Then,

the clustering module can be fine-tuned in the backward pass by minimizing the following cross-entropy loss:

$$L_{FT} = -\sum_{i=1}^{N} w_i(s_i \log(l_i) + (1 - s_i)\log(1 - l_i)) \tag{5}$$

where s_i is the one-hot vector corresponding to label feature l_i.

3.3. Extended Mutual Information

The discriminative class-level features have been obtained in the above section. In this section, by exploring the correlation between the features from multiple views of the same image, e.g., original and transformed images, the essential information in the image is learned, and the intra-class diversities in the instance-level representation feature space are also reduced. Maximizing MI between the complete input and the encoder output is not sufficient to learn useful representations [40]. The average MI between the global representation and the local features of the input plays a stronger role in improving the quality of representation than the global MI. Inspired by DIM, this paper only uses local MI and considers that KL divergence has no upper bound theoretically, so it is replaced by JS divergence (JSD) similar to f-gan [43]. The JSD version of MI is defined as:

$$MI(f_{map}, z) = \max E_J[-sp(-T(f_{map}, Z))] - E_M[sp(T(f_{map}, Z))] \tag{6}$$

where f_{map} corresponds to the local features (feature map), Z corresponds to the global feature (instance-level representation), sp is the softplus function and T is a discriminator trained to distinguish whether f_{map} and Z are sampled from the joint distribution J or from the marginal distribution M. In this paper, only f_{map} and Z belong to the same cluster and follow J; otherwise, they follow M. A pair of samples that satisfy the joint distribution is called a pair of positive samples; otherwise, a pair of negative samples.

In order to force the instance-level feature z to capture the essential information, our model extends DIM by introducing data augmentation techniques. Let x and $\tilde{x}$ be the original and transformed samples, respectively, $\{C_{i,j}(x)|i = 1, 2, \ldots, h; j = 1, 2, \ldots, w\}$ and $\{\tilde{C}_{i,j}(\tilde{x})|i = 1, 2, \ldots, h; j = 1, 2, \ldots, w\}$ correspond to their feature map. The negative samples of the feature map are all represented by the symbol $C'_{i,j}$. By concatenating $C_{i,j}$ and z to get a larger feature map as the input of the discriminator, the estimation of the global-local MI loss can be obtained by minimizing the following loss function:

$$\begin{aligned} &L_{LMI}(\tilde{z}, C_{i,j}) \\ &= -\max I(\tilde{z}, C_{i,j}) = \min -(E_J[-sp(-T(\tilde{z}, C_{i,j}))] - E_M[sp(T(\tilde{z}, C'_{i,j}))]) \end{aligned} \tag{7}$$

$$\begin{aligned} &L_{LMIT}(\tilde{z}, \tilde{C}_{i,j}) \\ &= -\max I(\tilde{z}, \tilde{C}_{i,j}) = \min -(E_J[-sp(-T(\tilde{z}, \tilde{C}_{i,j}))] - E_M[sp(T(\tilde{z}, C'_{i,j}))]) \end{aligned} \tag{8}$$

Meanwhile, not only the global-local MI is calculated. The original images and the augmented images' intermediate layer features should also be related, and their MI should be as large as possible. The estimation of local-local MI loss can be obtained by minimizing the following loss function:

$$\begin{aligned} &L_{LMIC}(\tilde{C}_{i,j}, C_{i,j}) \\ &= -\max I(\tilde{C}_{i,j}, C_{i,j}) = \min -(E_J[-sp(-T(\tilde{C}_{i,j}, C_{i,j}))] - E_M[sp(T(\tilde{C}_{i,j}, C'_{i,j}))]) \end{aligned} \tag{9}$$

By maximizing the MI between features extracted from multiple views with the same information, it implies that the features need to capture high-level information with decisive

influence, such as the existence of an object, so the extension of this paper is reasonable. Since DIM selects negative samples randomly, it is unreasonable to minimize the MI of pairwise samples when they belong to the same cluster. Therefore, in a similar way to DCCM, the positive and negative pairs of the same sample are selected according to the similarity of the sample pairs in this batch. Similar pairwise samples are regarded as positive pairs; otherwise, they are negative pairs. Maximizing MI between positive pairs promotes intra-class aggregation. Minimizing MI between negative pairs promotes inter-class separation. Based on the above-mentioned analysis and exploration, the final loss function is summarized as:

$$L_{MI} = \alpha L_{LMI} + \beta L_{LMIT} + \delta L_{LMIC} \tag{10}$$

where α, β and δ are constants to balance the contributions of different terms.

3.4. Maximize Entropy Loss

In deep clustering, it is easy to fall into a local optimal solution that assigns most samples into a minority of clusters. Inspired by GATCluster [44], this problem can be solved by introducing the technique of maximizing entropy, which can be formulated as:

$$\begin{aligned} L_E &= -H(q) = \sum_{i=1}^{K} q_i \log(q_i) \\ s.t.\ q_k &= \frac{1}{m} \sum_{i=1}^{m} l_{ik},\ k = 1, \ldots, K \end{aligned} \tag{11}$$

where l_{ik} indicates the k-th element of l_i, q_k represents the frequency of the k-th cluster and H is entropy.

Then the overall objective function of DCSM can be formulated as:

$$L = \varepsilon L_{DT} + L_{FT} + L_{MI} + \lambda L_E \tag{12}$$

where ε and λ are constants to balance the contributions of different terms. Figure 1 shows the framework of the proposed DCSM model.

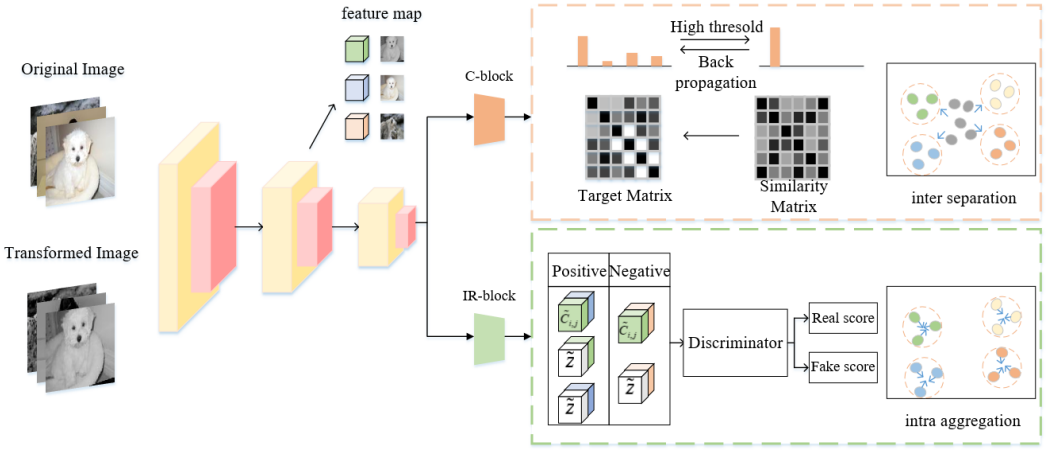


Figure 1. Model Overview. The original image and its transformed image are regarded as a pair of inputs to learn class-level (C-block) and instance-level (IR-block) representations using a shared deep neural network. The similarity matrix is generated from the class-level feature similarity relationship of the transformed images, and the target matrix is generated using K-means in the original image class-level feature space to guide the similarity matrix. Cluster standard samples are selected according to the threshold to modify the network parameters. The discriminator network estimates the mutual information of local-global and local-local representations, respectively.

4. Experiments

In this section, the effectiveness of DCSM is demonstrated on four image datasets through experiments. Furthermore, five ablation experiments are performed to systematically and comprehensively analyze the developed DCSM model. All experiments are implemented on a PC machine installed in a 64-bit operating system with two Nvidia GeForce GTX 1080 Ti GPUs with 8-GB video memory.

4.1. Datasets

This paper selects four representative image datasets for clustering, including CIFAR10, MNIST, STL-10 and Imagenet10. As a common setting [29,35], for each dataset, the training and test sets are jointly used in our experiments. For clarity, Table 1 reports the details of each dataset, such as the number of categories, the number of samples and so on.

Table 1. Statistics of the Different Datasets.

Datasets	Training Sets	Test Sets	Categories	Image Size
CIFAR10	50,000	10,000	10	$32 \times 32 \times 3$
MNIST	60,000	10,000	10	$28 \times 28 \times 1$
STL-10	13,000	N/A	10	$96 \times 96 \times 3$
ImageNet-10	13,000	N/A	10	$128 \times 128 \times 3$

4.2. Evaluation Metrics

For a complete comparison, this paper employs three clustering metrics to evaluate the clustering performance, i.e., Normalized Mutual Information (NMI), Adjusted Rand Index (ARI) and clustering Accuracy (ACC). In addition, these metrics range in [0,1], and higher scores indicate better clustering results.

4.3. Experimental Settings

The backbone network architecture used in our model is VGG, and ReLU activation is used on all hidden layers. The reason why we use ReLU activation is that it overcomes the vanishing gradient problem, allowing models to learn faster and perform better. Both the class-level mapping block and the instance-level mapping block are fully connected networks, and their output layer dimensions are set to 10 and 30, respectively. For the local MI objective, the global representation is concatenated with the feature map at each location. A 1×1 convolutional discriminator is then used to score the pair. The network parameters are optimized using Adam [45], and the learning rate is set to 0.001. The batch size of STL10, ImageNet-10, Cifar10 and MNIST is all 32. The number of training repetitions for each batch of MNIST is 4, and the for rest of the datasets, it is 8. The small perturbations used in the experiments include rotation, shift, color adjustment and so on. The setting of the remaining three hyperparameters is given in Table 2.

Table 2. Parameter Settings of the Different Datasets.

Datasets	α	β	δ
CIFAR10	0.1	0.5	0.2
MNIST	0.2	1	0.1
STL-10	0.1	1	0.1
ImageNet-10	0.3	1	0.3

4.4. Compared Methods

In the experiment, both traditional and deep learning-based methods are compared, including K-means, SC, AE, DAE, GAN, DECNN, VAE, DEC [37], JULE [34], DAC [26], DCCM [27], IIC [28], ICGAM [36] and EDCN [25]. DCSM-oc is a version of our model that uses only class-level map blocks.

4.5. Results and Analysis

In Table 3, the quantitative clustering results of the compared methods on four image datasets are orderly reported. From Table 3, our DCSM method consistently achieves superior performance on different datasets, which empirically signifies that DCSM is in a position to cluster data effectively. From further analysis, several tendencies can be observed in Table 3.

First, the performance of the clustering methods based on deep learning is generally superior to the traditional methods (e.g., K-means, SC). It shows that representation learning is more important than clustering techniques for unsupervised learning.

Second, the methods (e.g., DEC, DAC) of joint representation learning and cluster assignment outperform traditional methods and the representation-based clustering methods. It indicates that clustering and representation learning can promote each other and achieve better performance consequently.

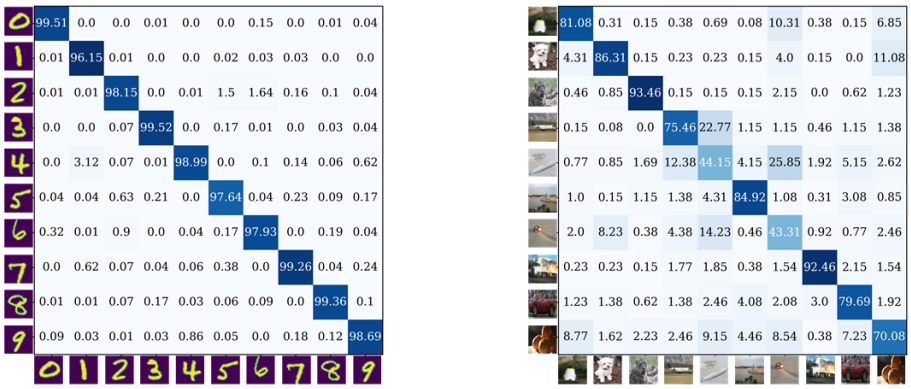
Third, the performance of DCCM and IIC using the data augmentation technique is better than other algorithms. It means that the introduction of the data augmentation technique in unsupervised clustering can help the model to be optimized.

Besides of these common conclusions, our DCSM method outperforms methods that only use class-level representation (e.g., DCSM-oc, DCCM). It means that the clustering performance is related to both class-level representation and instance-level representation.

Figure 2 visualizes the confusion matrix on the MNIST and Imagenet10 datasets. Figure 3 visualizes the class-level representations of the DCSM on the ImageNet10 datasets using t-SNE at the different training stages.

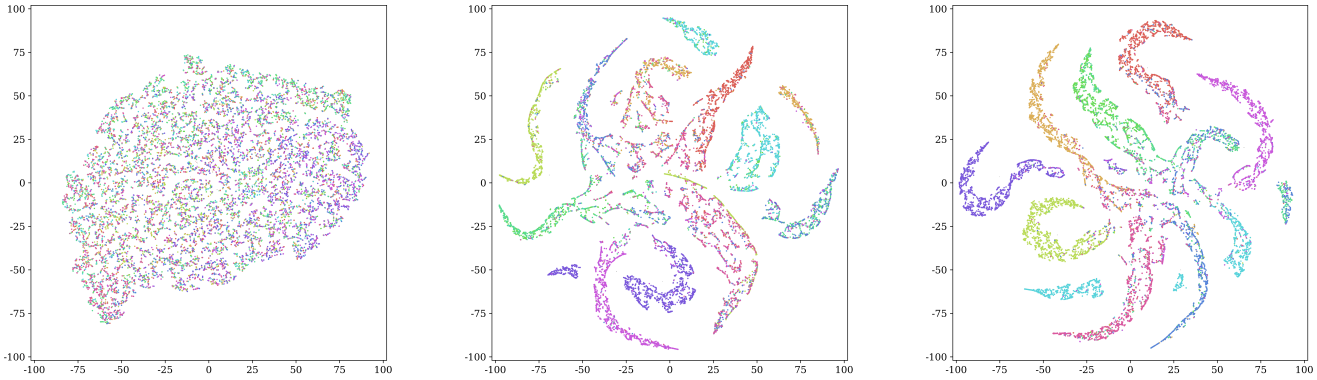
Table 3. Clustering performance of different methods on four datasets. The best results are highlighted in bold.

	MNIST			**CIFAR-10**			**STL-10**			**ImageNet-10**		
	NMI	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**
K-means	0.501	0.572	0.365	0.087	0.229	0.049	0.125	0.192	0.061	0.119	0.241	0.057
SC	0.662	0.695	0.521	0.103	0.247	0.085	0.098	0.159	0.048	0.151	0.274	0.076
AE	0.725	0.812	0.613	0.239	0.314	0.169	0.250	0.303	0.161	0.210	0.317	0.152
DAE	0.756	0.831	0.647	0.251	0.297	0.163	0.224	0.302	0.152	0.206	0.304	0.138
GAN	0.763	0.736	0.827	0.265	0.315	0.176	0.210	0.298	0.139	0.225	0.346	0.157
DeCNN	0.757	0.817	0.669	0.240	0.282	0.174	0.227	0.299	0.162	0.186	0.313	0.142
VAE	0.876	0.945	0.849	0.245	0.291	0.167	0.200	0.282	0.146	0.193	0.334	0.168
DEC	0.771	0.843	0.741	0.257	0.301	0.161	0.276	0.359	0.186	0.282	0.381	0.203
JULE	0.913	0.964	0.927	0.192	0.272	0.138	0.182	0.277	0.164	0.175	0.300	0.138
DAC	0.935	0.977	0.948	0.396	0.522	0.306	0.366	0.470	0.257	0.394	0.527	0.302
IIC	-	0.992	-	-	0.617	-	-	0.610	-	-	-	-
DCCM	0.951	0.982	0.954	0.496	0.623	0.408	0.376	0.482	0.262	0.608	0.710	0.555
ICGAM	-	0.990	-	-	**0.700**	-	-	0.587	-	-	-	-
EDCN	0.962	0.985	-	**0.603**	0.544	-	0.357	0.482	-	-	-	-
DCSM-oc	0.950	0.981	0.959	0.451	0.589	0.368	0.431	0.576	0.342	0.561	0.701	0.523
DCSM	**0.967**	**0.994**	**0.973**	0.505	0.636	0.413	**0.483**	**0.625**	**0.410**	**0.623**	**0.769**	**0.581**



(**a**) (**b**)

Figure 2. Confusion matrixes of the cluster results. (**a**,**b**) represent the clustering results on MNIST and Imagenet10 datasets, respectively.



(**a**) (**b**) (**c**)

Figure 3. Visualization of the class−level representations for different training stages of the proposed DCSM on the Imagenet10 datasets. (**a**) Initial stage of DCSM, (**b**) Middle stage of DCSM and (**c**) Final stage of DCSM.

4.6. Ablation Study

In this section, to further verify the performance of our model, five additional ablation experiments are performed.

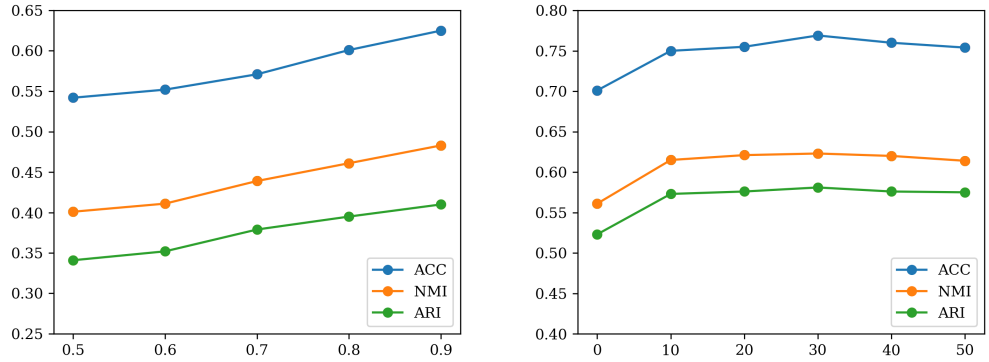
The influence of different strategies to construct $r_{i,j}$. The $r_{i,j}$ constructed by the high threshold in DCCM and constructed by the K-means in this paper are used to guide the training of the two identical networks on Stl-10 datasets. The results are shown in Table 4. It can be seen that our strategy has enough advantages, indicating that K-means is a better alternative solution than the threshold.

Table 4. The influence of different strategies on constructing $r_{i,j}$.

Methods	ACC	NMI	ARI
OUR	0.529	0.402	0.311
DCCM	0.391	0.297	0.213

Impact of threshold. The clustering performance of STL-10 datasets under different thresholds in Equation (2) is displayed in Figure 4a. For STL-10, the clustering performance increases gradually as the threshold increases. It indicates that the correct rate of the selected cluster-standard samples is also gradually increasing.

The dimensional impact of z. As shown in Figure 4b, varying the dimension of z from 10 to 50 does not affect the clustering performance much. However, when the dimension of z becomes 0, which means that the network only learns class-level representations, the performance drops a lot. The reason is that class-level representation alone is not enough to describe the differences between images belonging to the same cluster. A good clustering model should assign data to clusters to keep inter-group similarity low while maintaining high intra-group similarity.



(**a**) (**b**)

Figure 4. Clustering performance under different thresholds and z. (**a**) Impact of threshold. (**b**) The dimensional impact of z.

The sensitivity of the model to mutual information. For convenience, the mutual information loss L_{MI} is divided into two parts: the global-local MI $L_{ZC} = L_{LMI} + L_{LMIT}$ and the local-local MI L_{LMIC}. To comprehensively examine the effect of global-local and local-local MI on the clustering performance of the model, L_{MI} was changed to L_{ZC} and L_{LMIC}, respectively. The experimental results are shown in Table 5. From the results, the two parts of MI have similar performances in improving the clustering performance. It also proves our proposed extended mutual information is reasonable.

Table 5. The sensitivity of the model to mutual information.

	STL-10			**ImageNet-10**		
	NMI	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**
L_{ZC}	0.464	0.605	0.387	0.589	0.736	0.542
L_{LMIC}	0.455	0.596	0.376	0.583	0.727	0.564
$L_{ZC} + L_{LMIC}$	0.483	0.625	0.410	0.623	0.769	0.581

The impact of each part. In the case of the combination of each part, the clustering performance is shown in Table 6. Due to the lack of L_E loss, the network will have unstable results, so the maximum entropy loss is not removed in this part. It can be found that it is very effective to use cluster-standard samples to correct the network parameters.

Table 6. The impact of each part.

	STL-10			**ImageNet-10**		
	NMI	**ACC**	**ARI**	**NMI**	**ACC**	**ARI**
L_{DT}	0.402	0.529	0.311	0.501	0.638	0.469
$L_{DT} + L_{FT}$	0.431	0.576	0.342	0.561	0.701	0.523
$L_{DT} + L_{FT} + L_{MI}$	0.483	0.625	0.410	0.623	0.769	0.581

5. Conclusions

This paper proposes an end-to-end clustering network model, DCSM, which aims to simultaneously reduce intra-class differences and improve inter-class separation. The proposed extended mutual information method achieves the purpose of reducing intra-class differences by further exploring the relationships between the original image features and transformed image features. Based on the pseudo-similar labels of sample pairs, DCSM encourages similar samples to move closer to each other and dissimilar samples to move away from each other to promote inter-class separation. The experiments demonstrate the advantages of our approach on four benchmark datasets. Future work will explore how to achieve more accurate pseudo-similar labels of sample pairs. This may include the use of contrastive learning or soft assignment methods to generate the pseudo-similar labels of sample pairs.

Author Contributions: Conceptualization, F.P. and K.L.; methodology, F.P.; software, F.P.; validation, F.P. and K.L.; formal analysis, F.P.; investigation, F.P.; resources, F.P.; data curation, F.P.; writing—original draft preparation, F.P.; writing—review and editing, F.P. and K.L.; visualization, F.P.; supervision, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: Natural Science Foundation of Hebei Province (No. F2022201009); Hebei University High-level Scientific Research Foundation for the Introduction of Talent (No. 521100221029); Postgraduate's Innovation Fund Project of Hebei University (No. HBU2022ss022).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets are publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, C.; Dai, Y.; Lin, R.; Shi, W. Deep clustering by maximizing mutual information in variational auto-encoder. *Knowl.-Based Syst.* **2020**, *205*, 106260. [CrossRef]
2. Tang, X.; Dong, C.; Zhang, W. Contrastive Author-aware Text Clustering. *Pattern Recognit.* **2022**, *130*, 108787. [CrossRef]
3. Jiang, G.; Peng, J.; Wang, H.; Ze, M.; Xian, F. Tensorial Multi-view Clustering via Low-rank Constrained High-order Graph Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *30*, 5307–5318. [CrossRef]
4. Wang, H.; Jiang, G.; Peng, J.; Ruo, D.; Xian, F. Towards Adaptive Consensus Graph: Multi-view Clustering via Graph Collaboration. *IEEE Trans. Multimed.* **2022**, *10*, 1–13. [CrossRef]
5. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Los Angeles, CA, USA, 21 June–18 July 1967; Volume 1, pp. 281–297.
6. Von Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **2007**, *17*, 395–416. [CrossRef]
7. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496. [CrossRef]
8. Ghareeb, A.; Rida, S.Z. Image quality measures based on intuitionistic fuzzy similarity and inclusion measures. *J. Intell. Fuzzy Syst.* **2018**, *34*, 4057–4065. [CrossRef]
9. Xiao, Z.; Shi, Z.; Yong, L.; Ji, Z.; Li, Y.; Yue, F. Low-rank sparse subspace for spectral clustering. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 1532–1543.
10. Wang, H.; Wang, Y.; Zhang, Z.; Xian, F.; Li, Z.; Ming, X. Kernelized multiview subspace analysis by self-weighted learning. *IEEE Trans. Multimed.* **2020**, *23*, 3828–3840. [CrossRef]
11. Ding, X.; Zhang, X.; Ma, N.; Jun, H.; Gui, D.; Jian, S. Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13733–13742.
12. Kaur, P.; Harnal, S.; Tiwari, R.; Alharithi, F.S.; Almulihi, A.H.; Noya, I.D.; Goyal, N. A hybrid convolutional neural network model for diagnosis of COVID-19 using chest X-ray images. *Int. J. Environ. Res. Public Health* **2021**, *18*, 12191. [CrossRef]
13. Hui, W.; Jin, P.; Guang, J.; Feng, X.; Xian, F. Discriminative feature and dictionary learning with part-aware model for vehicle re-identification. *Neurocomputing* **2021**, *438*, 55–62.
14. Wang, H.; Peng, J.; Zhao, Y.; Fu, X. Multi-path deep cnns for fine-grained car recognition. *IEEE Trans. Veh. Technol.* **2020**, *69*, 10484–10493. [CrossRef]

15. Wang, H.; Peng, J.; Chen, D.; Jiang, G.; Zhao, T.; Fu, X. Attribute-guided feature learning network for vehicle reidentification. *IEEE Multimed.* **2020**, *27*, 112–121. [CrossRef]
16. Chugh, H.; Gupta, S.; Garg, M.; Gupta, D.; Mohamed, H.G.; Noya, I.D.; Singh, A.; Goyal, N. An Image Retrieval Framework Design Analysis Using Saliency Structure and Color Difference Histogram. *Sustainability* **2022**, *14*, 10357. [CrossRef]
17. Min, E.; Guo, X.; Liu, Q.; Zhang, G.; Cui, J.; Long, J. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access* **2018**, *6*, 39501–39514. [CrossRef]
18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
19. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
20. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
21. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde, D.; Ozair, S. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
22. Yu, Y.; Zhou, W. Mixture of GANs for clustering. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 3047–3053.
23. Ghasedi, K.; Wang, X.; Deng, C.; Huang, H. Balanced self-paced learning for generative adversarial clustering network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 4391–4400.
24. Mukherjee, S.; Asnani, H.; Lin, E.; Kannan, S. Clustergan: Latent space clustering in generative adversarial networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 22 February–1 March 2019; pp. 4610–4617.
25. Cao, W.; Zhang, Z.; Liu, C.; Li, R. Unsupervised discriminative feature learning via finding a clustering-friendly embedding space. *Pattern Recognit.* **2022**, *129*, 108768. [CrossRef]
26. Chang, J.; Wang, L.; Meng, G.; Xiang, S.; Pan, C. Deep adaptive image clustering. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5879–5887.
27. Wu, J.; Long, K.; Wang, F.; Qian, C.; Li, C.; Lin, Z.; Zha, H. Deep comprehensive correlation mining for image clustering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 8150–8159.
28. Xu, J.; Henriques, J.F.; Vedaldi, A. Invariant information clustering for unsupervised image classification and segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 9865–9874.
29. Yang, B.; Fu, X.; Sidiropoulos, N.D.; Hong, M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 3861–3870.
30. Fard, M.M.; Thonet, T.; Gaussier, E. Deep k-means: Jointly clustering with k-means and learning representations. *Pattern Recognit. Lett.* **2020**, *138*, 185–192. [CrossRef]
31. Lv, J.; Kang, Z.; Lu, X.; Xu, Z. Pseudo-supervised deep subspace clustering. *IEEE Trans. Image Process.* **2021**, *30*, 5252–5263. [CrossRef] [PubMed]
32. Diallo, B.; Hu, J.; Li, T.; Khan, G.; Liang, X.; Zhao, Y. Deep embedding clustering based on contractive autoencoder. *Neurocomputing* **2021**, *433*, 96–107. [CrossRef]
33. Ren, Y.; Wang, N.; Li, M.; Xu, Z. Deep density-based image clustering. *Knowl.-Based Syst.* **2020**, *197*, 105841. [CrossRef]
34. Yang, J.; Parikh, D.; Batra, D. Joint unsupervised learning of deep representations and image clusters. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5147–5156.
35. Sen, B.; Zhou, P.; Du, L.; Li, X. Active deep image clustering. *Knowl.-Based Syst.* **2022**, *252*, 109346. [CrossRef]
36. Ntelemis, F.; Jin, Y.; Thomas, S.A. Image clustering using an augmented generative adversarial network and information maximization. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 7461–7474. [CrossRef]
37. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised deep embedding for clustering analysis. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 478–487.
38. Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; Van Gool, L. Scan: Learning to classify images without labels. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 268–285.
39. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual information neural estimation. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 531–540.
40. Hjelm, R.D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; Bengio, Y. Learning deep representations by mutual information estimation and maximization. *arXiv* **2018**, arXiv:1808.06670.
41. Krause, A.; Perona, P.; Gomes, R. Discriminative clustering by regularized information maximization. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 531–540.
42. Hu, W.; Miyato, T.; Tokui, S.; Matsumoto, E.; Sugiyama, M. Learning discrete representations via information maximizing self-augmented training. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1558–1567.
43. Nowozin, S.; Cseke, B.; Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 271–279.

44. Niu, C.; Zhang, J.; Wang, G.; Liang, J. Gatcluster: Self-supervised gaussian-attention network for image clustering. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 735–751.
45. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–15.