

# Learning Clustering-Friendly Representations via Partial Information Discrimination and Cross-Level Interaction

Hai-Xin Zhang<sup>a</sup>, Dong Huang<sup>a,b</sup>, Hua-Bao Ling<sup>c</sup>, Weijun Sun<sup>d</sup>, Zihao Wen<sup>a</sup>

<sup>a</sup>College of Mathematics and Informatics, South China Agricultural University, China

<sup>b</sup>Key Laboratory of Smart Agricultural Technology in Tropical South China, Ministry of Agriculture and Rural Affairs, China

<sup>c</sup>School of Computer Science and Engineering, Sun Yat-sen University, China

<sup>d</sup>School of Automation, Guangdong University of Technology, China

## Abstract

Despite significant advances in the deep clustering research, there remain three critical limitations to most of the existing approaches. First, they often derive the clustering result by associating some distribution-based loss to specific network layers, neglecting the potential benefits of leveraging the contrastive sample-wise relationships. Second, they frequently focus on representation learning at the full-image scale, overlooking the discriminative information latent in partial image regions. Third, although some prior studies perform the learning process at multiple levels, they mostly lack the ability to exploit the interaction between different learning levels. To overcome these limitations, this paper presents a novel deep image clustering approach via Partial Information discrimination and Cross-level Interaction (PICI). Specifically, we utilize a Transformer encoder as the backbone, coupled with two types of augmentations to formulate two parallel views. The augmented samples, integrated with masked patches, are processed through the Transformer encoder to produce the class tokens. Subsequently, three partial information learning modules are jointly enforced, namely, the partial information self-discrimination (PISD) module for masked image reconstruction, the partial information contrastive discrimination (PICD) module for the simultaneous instance- and cluster-level contrastive learning, and the cross-level interaction (CLI) module to ensure the consistency across different learning levels. Through this unified formulation, our PICI approach for the first time, to our knowledge, bridges the gap between the masked image modeling and the deep contrastive clustering, offering a novel pathway for enhanced representation learning and clustering. Extensive experimental results across six image datasets demonstrate the superiority of our PICI approach over the state-of-the-art. In particular, our approach achieves an ACC of 0.772 (0.634) on the RSOD (UC-Merced) dataset, which exhibits

---

Email addresses: reganzhx@stu.scau.edu.cn (Hai-Xin Zhang),  
huangdonghere@gmail.com (Dong Huang), linghb5@mail2.sysu.edu.cn (Hua-Bao Ling),  
gdutswj@gdut.edu.cn (Weijun Sun), zihao.wen@hotmail.com (Zihao Wen)

an improvement of 29.7% (24.8%) over the best baseline. The source code is available at <https://github.com/Regan-Zhang/PICI>.

*Keywords:* Data clustering, Deep clustering, Image clustering, Masked image modeling, Contrastive learning.

## 1. Introduction

Data clustering is a fundamental yet challenging technique in machine learning and data mining. Traditional clustering methods usually assume that some hand-crafted features are given and seek to partition the dataset based on these given features, which, however, lack the feature representation learning ability and may yield suboptimal clustering results, especially for some high-dimensional complex data such as images and videos. With the joint ability of representation and clustering, the deep learning-based clustering methods, also known as the deep clustering methods [1, 2, 3, 4, 5, 6, 7], have captured significant attention in recent years.

Previous deep clustering methods [1, 2, 3, 4, 5] typically employ some deep neural network to learn feature representations and then obtain the clustering result by optimizing some cluster distribution-based clustering loss. Despite the considerable progress that has been achieved, these methods mostly suffer from three critical limitations. First, they often rely on some distribution-based loss (e.g., the Kullback-Leibler (KL) divergence-based clustering loss) to learn the clustering result, which neglect the contrastive information among the sample-wise relationships. Second, most of them tend to perform feature learning at the full-image scale, which overlook the opportunities to discover more discriminative semantics from partial (or masked) regions. Third, although the instance-level modeling and the cluster-level modeling are two key components in many deep clustering works, yet it is surprising that few of them have leveraged the cross-level interaction to adaptively and mutually enhance the multi-level learning.

Recently some efforts have been carried out to partially address one or two of the above three limitations. To utilize the contrastive information among the sample-wise relationships, some contrastive learning-based deep clustering methods have been developed, such as Contrastive Clustering (CC) [8], Instance Discrimination and Feature Decorrelation (IDFD) [9], and Prototypical Contrastive Learning (PCL) [10], which aim to explore the instance-wise (i.e., sample-wise) relationships through the contrastive learning paradigm. However, these contrastive deep clustering methods still rely on representation learning at the full-image scale, while lacking the partial information discrimination ability. On the other hand, they usually perform the instance-level contrastive learning and the cluster-level contrastive learning via two separate projectors, respectively, without considering taking advantage of the interaction between them. To train the network with partial information discrimination, Masked Image Modeling (MIM) [11] arises as a new trend for self-supervised learning. As a representative model in MIM, Masked Auto-Encoder (MAE) utilizes a specific

task that reconstructs the masked images to improve the representation learning ability via Vision Transformer (ViT) [12], which shows that the reconstruction and discrimination of partial image information may significantly benefit the representation learning. Yet the MAE is typically devised for the representation learning task only, and still lacks the deep clustering ability as well as the cross-level contrastive learning ability. Despite these impressive efforts, it remains an open problem how to overcome the above three limitations simultaneously and furthermore formulate the contrastive relationships, the cross-level interaction, and the partial information discrimination in a unified deep representation learning and clustering framework.

To jointly address the above three limitations, this paper presents a novel deep clustering framework based on Partial Information discrimination and Cross-level Interaction (PICI) (as shown in Fig. 1). Different from previous deep clustering methods that mostly utilize the convolutional neural network (CNN), we take advantage of a Transformer encoder as the backbone so as to better capture the global relevance information via the self-attention mechanism. Particularly, two types of augmentations are first performed on an input image to generate two augmented samples. Each of the augmented samples is split into a sequence of patches, where the patches are then randomly masked in order to incur the partial information loss that plays an important role in learning discriminant and semantic information from images. Thereafter, the two parallel views of augmented samples with masked patches are fed to the encoder (with position embedding) to produce the class tokens, denoted as [CLS]. Subsequently, three types of learning modules are jointly formulated, including two modules for partial information discrimination, namely, the partial information self-discrimination (PISD) module and the partial information contrastive discrimination (PICD) module, and one module for cross-level learning, namely, the cross-level interaction (CLI) module.

To be more specific, in PISD, a Transformer decoder is utilized to reconstruct the masked patches of the augmented sample, which seeks to learn the semantic information of the image by recovering the missing local regions. In PICD, the class tokens [CLS] of each augmented view are fed to two multi-layer perceptron (MLP)-based projectors, namely, the instance-MLP and the cluster-MLP, through which the instance-level contrastive learning and the cluster-level contrastive learning can be enforced, respectively. To bridge the gap between the two levels of contrastive learning, the cross-level adaptive learning is further enabled in the CLI module. From the feature representations learned in the instance-MLP, the clustering assignments (i.e., pseudo labels) are generated by self-labeling, which are then connected to the clustering assignments learned in the cluster-MLP by minimizing a cross-entropy loss between them. Thereby, the consistency between the two levels of contrastive learning is imposed for the joint learning. Extensive experiments are conducted on six challenging image datasets, which demonstrate the effectiveness of our PICI approach.

For clarity, the main contributions of this work are summarized below.

- This paper for the first time, to the best of our knowledge, bridges the gap

between masked image modeling and deep contrastive clustering. Specifically, to enjoy the advantages of masked image modeling, we jointly enforce two types of partial information discrimination learning, including the PISD with masked image reconstruction and the PICD with masked contrastive learning, for enhanced representation learning and clustering.

- Different from previous contrastive clustering methods that often neglect the connections between the instance-level and the cluster-level, we design a cross-level interaction mechanism to adaptively guide these two levels of contrastive learning in the dual label spaces with their consistency constrained and optimized.
- We propose a novel deep contrastive clustering approach termed PICI. Extensive experiments have been conducted on a variety of benchmark datasets, which confirm the superiority of our approach over the state-of-the-art deep clustering approaches.

The remainder of this paper is organized as follows. The related works on self-supervised learning and deep learning are reviewed in Section 2. We described the proposed PICI approach in Section 3. The experimental results on multiple real-world image datasets are reported in Section 4. Finally, the conclusion of this paper is presented in Section 5.

## 2. Related Work

In this section, we review the related works on self-supervised learning and deep clustering in Sections 2.1 and 2.2, respectively.

### 2.1. Unsupervised learning and Self-Supervised Learning

Unsupervised learning is a machine learning paradigm where the algorithm learns patterns and structures from unlabeled data without explicit guidance or supervision from labeled examples. It encompasses various techniques, such as clustering, dimensionality reduction, and generative modeling, that enable the extraction of valuable information from unannotated data. In recent years, self-supervised learning, as a subset of unsupervised learning, leverages the inherent structure and content within unlabeled data to create surrogate supervisory signals, enabling the model to learn more discriminative and comprehensive representations [13]. These learned representations can then be transferred to downstream tasks, such as classification, object detection, and semantic segmentation. Indeed, self-supervised learning has emerged as a new paradigm of unsupervised learning.

Recent developments in contrastive learning have greatly advanced the research in self-supervised learning. In a nutshell, the core idea of contrastive learning is to first construct positive and negative pairs, and then to pull the positive pairs close while pushing the negatives far away in the embedding subspace. Previous contrastive learning methods [14, 15, 16] often require a relatively large number of negative samples for instance discrimination, which in

fact treat each sample as a category and may lead to extra computational costs and increased memory storage. For example, the Instance Recognition (IR) method [14] utilizes a discrete memory bank to store the features of each sample. Meanwhile, the SimCLR method [15] requires a large batch size (e.g., 4096) while the MoCo method [16] employs a memory queue to temporarily save the representations produced by a momentum encoder with an exponential moving average.

Some recent studies suggest that the negative pairs are not essential for achieving instance discrimination in contrastive learning [17, 18, 19]. For example, both BYOL [17] and SimSiam [18] adopt an online predictor to avoid collapsing solutions while eliminating the dependence on negative pairs. Alternatively, the SwAV method [19] swaps the predictions where it predicts the code of a view from the representation of another view, which incorporates the clustering into a siamese network. More recently, the Masked Image Modeling (MIM) [11, 20, 21, 22] has emerged as a new trend in self-supervised learning, which designs a prediction task that reconstructs masked images to enhance the representation learning capability of the ViT [12]. Following the emergence of the MAE [11], which is a representative model in MIM, the Context Auto-Encoder (CAE) [22] further incorporates a latent contextual regressor with the alignment constraint, while the Masked Siamese Network (MSN) [20] and the Contrastive Masked Auto-Encoder (CMAE) [21] incorporate the contrastive learning paradigm into the MIM framework. Meanwhile, some researchers have adopted ViT and multi-head attention mechanisms as replacements for CNN. For example, UKSSL [23] aims to efficiently learn the underlying knowledge from unlabeled data with the help of ViT and achieves remarkable performance. Our main motivation for adopting ViT as the backbone is also based on its promising performance in various domains [24, 25].

## 2.2. Deep Clustering

In recent years, the deep clustering methods [1, 2, 8, 9, 7, 4, 26, 27, 28, 29] have attracted increasing attention and made significant progress with the support of deep neural networks. Some early works in deep clustering, such as Deep Embedding Clustering (DEC) [1], Improved Deep Embedding Clustering (IDEC) [2], and Adaptive Self-Paced Deep Clustering with Data Augmentation (ASPC-DA) [4], often utilize the reconstruction loss to pre-train the network and further achieve the clustering result either by associating some clustering loss with a specific layer in the network or by simply performing the  $K$ -means clustering on the learned representation. Alternatively, the Deep Clustering and Visualization (DCV) method [26] integrates the clustering task with data visualization to preserve the geometric structure. The DeepCluster method [28] utilizes the self-labeling to enhance the clustering performance by transforming the unsupervised image clustering problem into a supervised one guided by the pseudo-labels.

More recently, the contrastive learning has demonstrated its promising potential in deep clustering. To incorporate the contrastive learning paradigm

into the deep clustering, data augmentation is required to construct the sample pairs, where the samples augmented from the same instance are treated as positive pairs, while the other pairs are treated as negative pairs. Specifically, the Instance Discrimination and Feature Decorrelation (IDFD) method [9] aims to learn similarities by the instance-level contrastive learning while reducing the correlations within features in the meantime. Furthermore, some recent methods [8, 27, 29] seek to conduct the contrastive learning at both the instance-level and the cluster-level for simultaneous representation learning and clustering, among which a representative method is the Contrastive Clustering (CC) method [8]. Different from the CC method which utilizes two identical and weight-sharing networks for the two augmented views, respectively, Deng et al. [27] presented a Heterogeneous Tri-stream Clustering Network (HTCN), which extends the two-stream contrastive learning network into three streams of heterogeneous networks, including two online networks and a target network, for learning clustering-friendly representations for the image clustering task.

### 3. Proposed Framework

In this paper, we present a novel unsupervised deep image clustering model termed PICI, which can be trained in an unsupervised manner. The training of our PICI model mainly involves three learning modules, namely, the PISD module for the self-discrimination learning of the masked images, the PICD module with two levels of contrastive learning, and the CLI module for imposing the cross-level consistency.

Specifically, given an input image, we first perform two types of augmentations on the image and thus obtain two augmented samples. These two augmentations formulate two parallel views of the backbone network (i.e., a Transformer encoder). Each of the two augmented samples is split into a sequence of patches, which are randomly masked and then fed to the backbone to produce the class tokens [CLS]. Thereafter, three learning modules are jointly incorporated. In the PISD module, a decoder is leveraged to recover the original masked images. In the PICD module, the [CLS] tokens are utilized to achieve the instance- and the cluster-level contrastive learning. Furthermore, the CLI module is exploited to enforce the consistency between the dual levels of contrastive learning via the instance-MLP and the cluster-MLP, respectively, where the set of pseudo labels in the instance-level space are generated by self-labeling and thereby the cross-level consistency is imposed through the maximum match cluster labels between the pseudo labels and the cluster assignment from the cluster-level space. Finally, with the overall network trained, the cluster assignments in the cluster-MLP can be obtained as the final clustering result.

#### 3.1. Backbone with Parallel Views

In the proposed PICI model, we construct two parallel views through data augmentations, which will be fed to the Transformer-based encoder and further leveraged for the contrastive learning process. Specifically, we adopt two

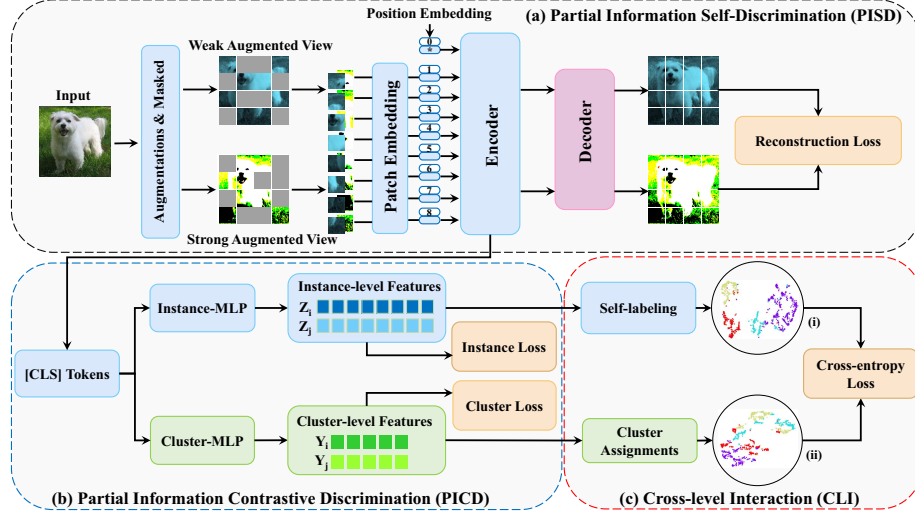


Figure 1: An overview of our PICI framework, which encompasses three partial information learning modules, namely, (a) the PISD module, which enforces the partial information self-discrimination upon the masked images via a Transformer auto-encoder, (b) the PICD module, which takes the class tokens [CLS] as input and performs two levels of contrastive learning, and (c) the CLI module, which enables the mutual interaction between the instance- and cluster-level subspaces by constraining their cross-level consistency.

*different* transformation families for the two views, where the first view is associated with a conventional weak augmentation (without severe distortions) and the second view is associated with a stronger augmentation (with more severe distortions) so as learn more discriminative features from the image. Formally, let the weak augmentation be denoted as  $T_w$  and the strong augmentation be denoted as  $T_s$ . Given an input image, say,  $x_i$ , a weak augmentation and a strong one are performed on the image, respectively, leading to its two corresponding views, denoted as  $x_i^a = T_w(x_i)$  and  $x_i^b = T_s(x_i)$ .

It is noteworthy that most previous contrastive learning works tend to adopt some data augmentations without severe distortions, which are called weak augmentations. In some recent studies, it has been proven that mixing weak and strong augmentations may lead to better contrastive learning performance [7, 30]. In this work, we adopt the weak transformations in the weakly augmented view and the strong transformations in the strongly augmented view. Given the parallel views, a weight-sharing backbone is used to extract features  $h$  (i.e., [CLS]) from the augmented samples. Note that the augmented samples in the two parallel views will be split into sequences of patches and then randomly masked. Thereafter, we adopt the ViT [12] as the backbone, where the self-attention mechanism of the Transformer [31] can bring in the information of global dependencies for enhanced representation learning.

### 3.2. Partial Information Discrimination

Different from conventional deep clustering methods that mostly work at the full-image scale, in this paper, we seek to enforce two types of partial information discrimination, corresponding to the PISD module and the PICD module, respectively, for representation learning via the masked images. Specifically, the PISD module is utilized to train the network by minimizing the discrimination between the original image and the image recovered from the masked image, which will be described in Section 3.2.1. Meanwhile, the PICD module is incorporated to perform two levels of contrastive learning, aiming to minimize the discrimination between the positive pairs while maximizing that between the negative pairs at the instance-level and the cluster-level, respectively, which will be described in Section 3.2.2.

#### 3.2.1. Partial Information Self-Discrimination (PISD)

Given an augmented image, we split it into a sequence of regular non-overlapping patches following the ViT [12]. Then a subset of patches are randomly selected as visible patches and the rest of them as masked (or invisible) patches. Following the MAE [11], we select the random patches to be masked w.r.t. a uniform distribution. It is argued that the masked random sampling with a certain masking ratio (e.g., 50%) may largely eliminate the redundancy of image data [11] and the sparse (masked) input is conducive to improving the ability of representation learning for images.

Formally, let the Transformer encoder be denoted as  $F(\cdot)$ , which operates only on the unmasked patches. These unmasked (or visible) patches are projected with position embeddings, which are employed to retain the positional information for the later recovery. Concretely, we adopt the standard learnable 1-D position embeddings and utilize the resultant sequence of embedding vectors as the input to the encoder. Note that the masked patches are removed and only the unmasked tokens are used, which only takes partial information into consideration and also alleviates the time and memory consumption of the model training process.

In the PISD module, to learn the discriminative representations by recovering the missing patches, we incorporate a decoder  $D(\cdot)$  through another series of Transformer blocks for image reconstruction. Specifically, to reconstruct the image from the embeddings produced by the encoder  $F(\cdot)$ , we add positional embeddings to all tokens so that the masked tokens will be located in their corresponding positions in the image. With the encoder and the decoder forming an auto-encoder, it is employed to reconstruct the input image by predicting the pixel values for all masked patches. Without loss of generality, we calculate the Mean Squared Error (MSE) between the original and reconstructed images at the pixel level as the reconstruction loss. Then we have the overall reconstruction loss of the PISD module as follows:

$$\mathcal{L}_{PISD} = \frac{1}{2}(\ell_{re}^a + \ell_{re}^b). \quad (1)$$



where  $\ell_{re}^a$  and  $\ell_{re}^b$  denote the MSE losses for the first and second views, respectively.

### 3.2.2. Partial Information Contrastive Discrimination (PICD)

In the PICD module, to simultaneously enforce the instance- and cluster-level contrastive learning, two independent MLP projectors are utilized to project the representations (i.e., [CLS]) extracted by the backbone to the instance- and cluster-level subspaces, respectively, which will further be associated with the instance loss and the cluster loss for contrastive representation learning. Before delving into the details of these two contrastive losses, we first present the overall loss of the PICD module, that is

$$\mathcal{L}_{PICD} = \mathcal{L}_{ins} + \mathcal{L}_{clu}. \quad (2)$$

where  $\mathcal{L}_{ins}$  is the instance contrastive loss associated with the instance-MLP, and  $\mathcal{L}_{clu}$  is the cluster contrastive loss associated with the cluster-MLP.

*Instance-level Contrastiveness.* The instance-level contrastive learning essentially works by maximizing the similarities between the positive sample pairs while minimizing that of the negative ones. In contrastive learning, it is a fundamental issue to define the positive and negative pairs. In recent years, many methods for constructing positive and negative sample pairs have been developed. For example, one can define the pairs of within-class samples to be positive and the between-class pairs to be negative. However, without prior knowledge, whether two arbitrary samples belong to the same class is hard to determine. In this paper, following the standard protocol in contrastive learning[8, 15], we regard the pair of samples augmented from the same sample as positive and other pairs as negative.

Given a mini-batch of  $N$  samples, our PICD performs weak and strong data augmentations on each sample and then  $2 \cdot N$  augmented samples, denoted as  $\{x_1^a, \dots, x_N^a, x_1^b, \dots, x_N^b\}$ , are produced. For a specific sample  $x_i$ , the pair of  $\{x_i^a, x_i^b\}$  is treated as its exclusive positive pair while the other  $2 \cdot (N - 1)$  pairs as its negative pairs. Note that directly conducting contrastive learning on the feature representation  $h$  may induce information loss [15]. Hence we stack a two-layer nonlinear MLP  $G_I(\cdot)$ , i.e., the instance-MLP, to map the feature representations  $h_i^a$  and  $h_i^b$  onto a low-dimensional subspace, denoted as  $Z_i^a = G_I(h_i^a)$  and  $Z_i^b = G_I(h_i^b)$ , respectively.

Let the instance representations set for the first view be denoted as  $I^a = \{Z_1^a, Z_2^a, \dots, Z_N^a\}$  and that for the second view as  $I^b$ . Then the pair-wise similarity between two feature vectors is measured by the cosine similarity, that is

$$sim(Z_i^{k_1}, Z_j^{k_2}) = \frac{(Z_i^{k_1})^\top (Z_j^{k_2})}{\|Z_i^{k_1}\| \cdot \|Z_j^{k_2}\|}, \quad (3)$$

with  $k_1, k_2 \in \{a, b\}$  and  $i, j \in [1, N]$ . It is noteworthy that if  $Z_i$  and  $Z_j$  are normalized to unit norm, the cosine similarity in Eq. (3) can be simplified into

a dot-product form, that is

$$\text{sim}(Z_i^{k_1}, Z_j^{k_2}) = (Z_i^{k_1})^\top (Z_j^{k_2}). \quad (4)$$

Further, we utilize the InfoNCE loss [32] for instance-level contrastive learning. Thus, we have the contrastive loss for a given sample  $x_i^a$  as

$$\ell_i^a = -\log \frac{\exp(\text{sim}(Z_i^a, Z_i^b)/\tau_I)}{\sum_{k \in a, b} \sum_{j=1}^N \exp(\text{sim}(Z_i^a, Z_j^k)/\tau_I)}, \quad (5)$$

where  $\tau_I$  is the instance temperature parameter that adjusts the degree of attraction and repulsion between samples.

Finally, the instance contrastive loss for a mini-batch of  $N$  input images can be represented as

$$\mathcal{L}_{ins} = \frac{1}{2 \cdot N} \sum_{i=1}^N (\ell_i^a + \ell_i^b). \quad (6)$$

where  $\ell_i^a$  is the instance contrastive loss for a given sample  $x_i$  for the first view and  $\ell_i^b$  for the second view.

*Cluster-level Contrastiveness.* The idea of “label as representation” in online clustering implies that, when a data sample is projected into a space, where the number of dimensions corresponds to the number of clusters, the  $i$ -th element of its feature vector can be regarded as its likelihood of being part of the  $i$ -th cluster. Consequently, this feature vector can be seen as the data sample’s soft label [33].

Similar to the instance-MLP  $G_I(\cdot)$  in the instance-level contrastive learning, we employ another two-layer MLP  $G_C(\cdot)$ , i.e., the cluster-MLP, with an extra softmax layer to project the representations  $h_i^a$  and  $h_i^b$  into an  $M$ -dimensional space. Formally, let  $C^a = \{Y_1^a, Y_2^a, \dots, Y_N^a\} \in \mathbb{R}^{N \times M}$  represent the cluster assignment probabilities for a mini-batch of  $N$  samples in the first view, and let  $C^b = \{Y_1^b, Y_2^b, \dots, Y_N^b\}$  represent those in the second view, where  $M$  is the number of clusters. Consequently, the  $i$ -th column of  $C^a$  can be interpreted as a representation of the  $i$ -th cluster, and all columns should be distinct from one another. Let  $\{Y_i^a, Y_i^b\}$  represent a positive cluster pair while leaving the other  $2 \cdot M - 2$  pairs as negative pairs. Here, the cosine similarity is employed to measure the similarity between cluster pairs.

$$\text{sim}(Y_i^{k_1}, Y_j^{k_2}) = \frac{(Y_i^{k_1})^\top (Y_j^{k_2})}{\|Y_i^{k_1}\| \cdot \|Y_j^{k_2}\|}, \quad (7)$$

with  $k_1, k_2 \in \{a, b\}$  and  $i, j \in [1, M]$ . Here, the cluster-level contrastive loss is defined to differentiate cluster  $Y_i^a$  from all other clusters except its counterpart  $Y_i^b$ . Thus, the contrastive loss for cluster  $Y_i^a$  can be calculated as

$$\tilde{\ell}_i^a = -\log \frac{\exp(\text{sim}(Y_i^a, Y_i^b)/\tau_C)}{\sum_{k \in \{a, b\}} \sum_{j=1}^M \exp(\text{sim}(Y_i^a, Y_j^k)/\tau_C)} \quad (8)$$

where  $\tau_C$  is the cluster temperature parameter. Directly minimizing the above contrastive loss might lead to a trivial solution that assigns most samples to a single cluster or a few clusters. To circumvent this issue, an entropy term is introduced to constrain the cluster assignment probabilities, that is

$$\mathcal{H}(C) = - \sum_{i=1}^M [P(Y_i^a) \log P(Y_i^a) + P(Y_i^b) \log P(Y_i^b)] \quad (9)$$

$$P(Y_i^k) = \frac{\sum_{j=1}^N C_{ji}^k}{\|Y_i^k\|_1}, k \in \{a, b\}, i \in [1, M] \quad (10)$$

where  $C_{ji}^k$  denotes the probability of sample  $j$  being assigned to cluster  $i$ . Therefore, the overall cluster contrastive loss function can be defined as follows:

$$\mathcal{L}_{clu} = \frac{1}{2 \cdot M} \sum_{i=1}^M (\tilde{\ell}_i^a + \tilde{\ell}_i^b) - \mathcal{H}(C) \quad (11)$$

### 3.3. Cross-level Interaction (CLI)

In the PICD module, as described in Section 3.2.2, the model simultaneously learns the instance-wise features and the cluster assignments  $\{C^k\}, k \in \{a, b\}$  through two levels of contrastive learning. These two levels of contrastive learning effectively capture the instance- and cluster-level contrastive information, respectively. To strengthen the connection between these two learning levels, in this section, we further incorporate the CLI module to achieve the joint learning with cross-level consistency adaptively enforced.

However, in practice [8, 34], the instance-level subspace is semantic-richer compared with the cluster-level subspace. Based on this observation, we propose Cross-level Interaction (CLI) to build a bridge between instance- and cluster-level subspaces. Let  $\{C^k\}$  denote anchors and align them with the clusters among  $I^k$ . In this way, the cluster information contained in the instance-level subspace is leveraged to improve the clustering effect of the semantic labels.

Specifically, we employ the  $K$ -means clustering to produce the pseudo-labels of all samples in the instance-level subspace. For the  $k$ -th view with  $k \in \{a, b\}$ , let  $\{u_m^k\}_{m=1}^M \in \mathbb{R}^I$  denote the  $M$  cluster centroids. Thus, we have

$$\min_{u_1^a, u_2^a, \dots, u_M^a} \sum_{i=1}^N \sum_{j=1}^M \|Z_i^a - u_j^a\|_2^2. \quad (12)$$

Then the pseudo-labels of all samples  $p^k \in \mathbb{R}^N$  (for  $k \in \{a, b\}$ ) are obtained as

$$p^k = \arg \min_j \|Z_i^k - u_j^k\|_2^2. \quad (13)$$

Let  $q^k \in \mathbb{R}^N$  (for  $k \in \{a, b\}$ ) denote the cluster labels generated from the cluster-level subspace, which can be represented as

$$q_i^k = \arg \max_j q_{ij}^k. \quad (14)$$

Further, to enable the interaction between the instance-level and the cluster-level, we treat  $q^k$  as the anchors to align with  $p^k$  according to the maximum matching criterion, that is

$$\begin{aligned} & \min_{W^k} V^k W^k, \\ & s.t. \sum_{i=1} w_{ij}^k = 1, \sum_{j=1} w_{ij}^k = 1, \\ & w_{ij}^k \in \{0, 1\}, i, j = 1, 2, 3, \dots, M, \end{aligned} \quad (15)$$

where  $W^k \in \{0, 1\}^{M \times M}$  is a boolean matching matrix (to be learned) and  $V^k \in \mathbb{R}^{M \times M}$  denotes the cost matrix with  $V^k = \max_{i,j} \tilde{v}_{ij}^k - \tilde{V}^k$  and  $\tilde{v}_{ij}^k = \sum_{n=1}^N \mathbb{1}[q_n^k = i] \cdot \mathbb{1}[p_n^k = j]$ . Here,  $\mathbb{1}[\cdot]$  is the indicator function. The modified cluster assignments  $\tilde{p}_i^k \in \{0, 1\}^M$  for the  $i$ -th sample is defined as a one-hot vector. The  $m$ -th element of  $\tilde{p}_i^k$  is 1 if there exists  $s \in \{1, 2, \dots, M\}$  such that  $\mathbb{1}[w_{ms}^k = 1] \cdot \mathbb{1}[p_i^k = s] = 1$ . Formally, we adopt the cross-entropy loss to align the distribution of the pseudo-labels with the cluster assignments, that is

$$\mathcal{L}_{CLI} = -\frac{1}{2} \cdot \sum_{k \in \{a, b\}} \tilde{P}^k \log C^k, \quad (16)$$

where  $\tilde{P}^k = \{\tilde{p}_1^k, \tilde{p}_2^k, \dots, \tilde{p}_N^k\} \in \mathbb{R}^{N \times M}$ . In this way, the semantic-rich instance features are leveraged to guide the cluster-level learning. This, in turn, optimizes the backbone network via back-propagation, which subsequently benefits the instance-level learning. Thereby, the overall representation learning of the proposed model is enhanced through the adaptive interaction between the two levels of contrastive learning in the CLI module.

### 3.4. Training Strategy

In this work, the network training is performed in three stages. In the first stage, to enhance the stability and representation learning capability of the proposed model, we first pre-train the masked auto-encoder with the reconstruction loss as defined in Eq. (1). Then, the reconstruction-based learning in PISD and the two levels of contrastive learning in PICD are jointly enforced in the second stage. Finally, the contrastive learning in PICD and the cross-level learning in CLI are further conducted in the boosting stage. For clarity, the overall process of the proposed PICI approach is summarized in Algorithm 1.

## 4. Experiments

In this section, we conduct extensive experiments to benchmark the proposed PICI approach against a variety of non-deep and deep clustering approaches on six real-world image datasets. In addition, we present qualitative analyses and ablation experiments to provide a more comprehensive and clear perspective on our proposed approach.

---

**Algorithm 1:** Algorithm for PICI

---

**Input:** Dataset  $\mathcal{X}$ ; Pre-training epochs  $E_1$ ; Training epochs  $E_2$ ; Boosting epochs  $E_3$ ; Batch size  $N$ ; Masked random sampling  $\mathbb{M}$ ; Temperature parameters  $\tau_I$  and  $\tau_C$ ; Cluster number  $M$ ; Network structure of  $\mathcal{T}$ ,  $F$ ,  $D$ ,  $G_I$ , and  $G_C$ .

**Output:** Cluster assignments.

```
// Pre-training
1 for epoch = 1 to  $E_1$  do
2   Sample a mini-batch  $\{x_i\}_{i=1}^N$  from  $\mathcal{X}$ 
3   Sample two augmentations  $T_w, T_s \sim \mathcal{T}$ 
4    $\tilde{h}_i^a = F(\mathbb{M}(T_w(x_i)))$ ,  $\tilde{h}_i^b = F(\mathbb{M}(T_s(x_i)))$ ,  $\tilde{x}_i^a = D(\tilde{h}_i^a)$ ,  $\tilde{x}_i^b = D(\tilde{h}_i^b)$ 
5   Compute  $\mathcal{L}_{PISD}$  through Eq. (1)
6   Update  $F, D$  to minimize  $\mathcal{L}_{PISD}$ 
7 end
// Training
8 for epoch = 1 to  $E_2$  do
9   Conduct dual contrastive learning
10  Compute  $\mathcal{L}_{PICD}$  through Eq. (2)
11  Update  $F, D, G_I, G_C$  to minimize  $\mathcal{L}_{PISD} + \mathcal{L}_{PICD}$ 
12 end
// Boosting
13 for epoch = 1 to  $E_3$  do
14  Match cluster labels between instance- and cluster-level spaces by solving Eq. (15).
15  Compute  $\mathcal{L}_{CLI}$  through Eq. (16)
16  Update  $F, G_I, G_C$  to minimize  $\mathcal{L}_{PICD} + \mathcal{L}_{CLI}$ 
17 end
// Test
18 for  $x$  in  $\mathcal{X}$  do
19  Extract representations by  $h = F(x)$  Calculate the cluster assignment by  $c = \arg \max G_C(h)$ 
20 end
```

---

#### 4.1. Implementation Details

In this work, we adopt two distinct families of data augmentations to generate the weak and strong augmentations, respectively. Specifically, the strong augmentation  $T_s$  is randomly drawn from the augmentation family that includes ResizedCrop, ColorJitter, Grayscale, HorizontalFlip, and GaussianBlur, while the weak augmentation  $T_w$  merely resizes and normalizes the images. A lightweight version of ViT, namely the ViT-Small [41], is utilized as the backbone, with a dimension size of 384 and 6 blocks of encoder layers. In contrast, the corresponding decoder consists of 8 Transformer blocks and 16 multi-heads with a dimension of 512. The combined encoder and decoder can be regarded

Table 1: The parameter settings of PICI in our experiments.

Name	#Value
epochs	1050
image size	$224 \times 224 \times 3$
batch size	96
learning rate	0.0001
instance temperature	0.5
cluster temperature	0.1
mask ratio	0.5
patch size	16
layers of encoder	6
layers of decoder	8
multi-heads of encoder	12
multi-heads of decoder	16
projection dimension of encoder	384
projection dimension of decoder	512

Table 2: The image datasets used in our experiments.

Dataset	#Samples	#Classes	#Image Size
RSOD [35]	976	4	—
UC-Merced [36]	2,100	21	$256 \times 256$
SIRI-WHU [37]	2,400	12	$200 \times 200$
AID [38]	10,000	30	$600 \times 600$
D0 [39]	4,508	40	$200 \times 200$
Chaoyang [40]	6,160	4	$512 \times 512$

as a vanilla MAE [11]. Each input image is resized to dimensions of  $224 \times 224$ .

In PICD, two types of projectors are incorporated, including the instance-MLP with an output dimension of 128 and the cluster-MLP with the output dimension setting to the target cluster number. The temperature parameters  $\tau_I$  and  $\tau_C$  are set to 0.5 and 1.0, respectively. To optimize the network, we use the Adam optimizer with an initial learning rate  $1e^{-4}$ . Note that the masked random sampling varies between the training phase and the testing phase. During training, the masking ratio is set to 50% across all datasets. During testing, the masked random sampling is omitted. The model is pre-trained for 200 epochs and trained for 800 epochs, followed by 50 boosting epochs for all image datasets. The batch size is set to 96. All experiments are executed on a single NVIDIA RTX 3090 GPU, using the Ubuntu 18.04 platform with CUDA 11.0 and PyTorch 1.7.0. For clarity, our experimental parameter settings are summarized in Table 1.

#### 4.2. Datasets and Evaluation Metrics

In our experiments, six real-world image datasets are used for evaluation, which are described as follows:

- **RSOD** [35] is an open remote sensing dataset, which consists of a total of 976 images and 4 classes, namely, the aircraft, the playground, the overpass, and the oil tank.
- **UC-Merced** [36] is a land use image dataset with 21 classes. Each class includes 100 images. The pixel size of each image is  $256 \times 256$ .
- **SIRI-WHU** [37] is a 12-class remote sensing image dataset, which is constructed by the RS-IDEA Group in Wuhan University (SIRI-WHU). Each class in the dataset includes 200 images with a size of  $200 \times 200$ .
- **AID** [38] is a large-scale remote sensing dataset, which includes 30 aerial scene types and a total of 10,000 images.
- **D0** [39] is an image dataset with 40 common pest species, which consists of a total of 4,508 images.
- **Chaoyang** [40] comprises 6,160 images from 4 classes of colon slides, collected from Chaoyang Hospital.

For clarity, we summarize the statistics of the six image datasets in Table 2. Note that the images in the RSOD dataset are not uniform in size. Following the standard evaluation protocol for the image clustering task, we adopt three widely-used evaluation metrics in our experiments, including the Normalized Mutual Information (NMI) [42], the accuracy (ACC) [43], and the Adjusted Rand Index (ARI) [44]. The value range for NMI and ACC is  $[0, 1]$ , whereas ARI varies in the range of  $[-1, 1]$ . It’s worth noting that higher values of these metrics indicate better clustering results.

#### 4.3. Results and Analysis

In this section, we evaluate the proposed PICI method against several state-of-the-art clustering methods on six real-world image datasets. These methods can be divided into two categories, i.e., the non-deep clustering methods and the deep clustering methods. The non-deep clustering methods include  $K$ -means [45], Spectral Clustering (SC) [46], Agglomerative Clustering (AC) [47], Non-negative Matrix Factorization (NMF) [48], Principle Component Analysis (PCA) [49], Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [50] and GMM [51]. The deep clustering methods include Deep Embedding Clustering (DEC) [1], Improved Deep Embedding Clustering (IDEC) [2], Adaptive Self-Paced Deep Clustering with Data Augmentation (ASPC-DA) [4], Instance Discrimination and Feature Decorrelation (IDFD) [9], Contrastive Clustering (CC) [8], Deep Clustering and Visualization (DCV) [26] and Heterogeneous Tri-stream Clustering Network (HTCN) [27]. For NMF and PCA, the clustering results are obtained by performing  $K$ -means on the extracted

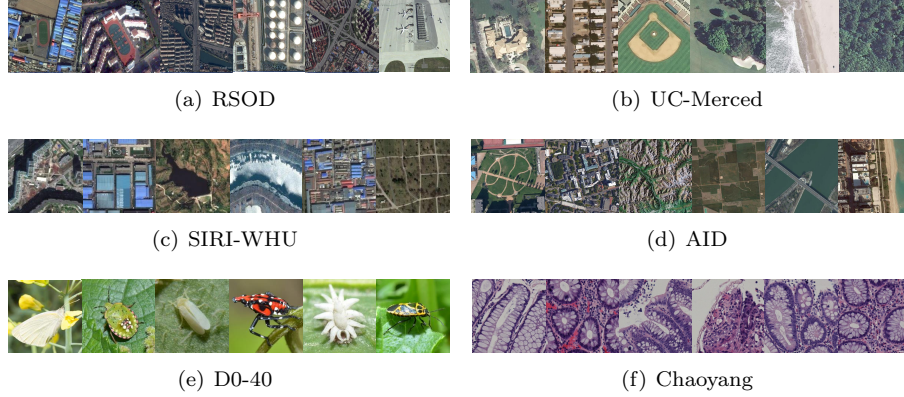


Figure 2: Some examples of the six image datasets used for evaluation, including four remote sensing datasets [35, 36, 37, 38], a crop pest dataset [39], and a medical dataset [40].

Table 3: The **NMI** scores of different clustering methods on the six datasets.

Dataset	RSOD	UC-Merced	SIRI-WHU	AID	D0-40	Chaoyang
<i>K</i> -means [45]	0.162	0.204	0.145	0.209	0.299	0.024
SC [46]	0.146	0.211	0.161	0.189	0.305	0.022
AC [47]	0.168	0.214	0.166	0.204	0.319	0.026
NMF [48]	0.176	0.202	0.245	0.193	0.255	0.018
PCA [49]	0.163	0.206	0.164	0.216	0.308	0.024
BIRCH [50]	0.148	0.225	0.162	0.204	0.315	0.026
GMM [51]	0.160	0.198	0.160	0.205	0.289	0.024
DEC [1]	0.296	0.120	0.183	0.217	0.328	0.001
IDEC [2]	0.209	0.119	0.178	0.207	0.309	0.001
ASPC-DA [4]	0.054	0.137	0.103	0.060	0.153	0.026
IDFD [9]	0.391	0.572	0.540	0.696	0.663	0.309
CC [8]	0.457	0.609	0.603	0.752	0.693	0.365
DCV [26]	0.178	0.102	0.128	0.127	0.139	0.024
HTCN [27]	0.557	0.596	0.534	0.797	0.721	0.276
PICI (Ours)	<b>0.583</b>	<b>0.681</b>	<b>0.658</b>	<b>0.800</b>	<b>0.731</b>	<b>0.382</b>

features. For the other algorithms, the model settings will be set as suggested by their corresponding papers.

The experimental results w.r.t. NMI, ACC, and ARI of different non-deep and deep clustering methods are reported in Tables 3, 4, and 5, respectively. It is obvious that the deep clustering methods perform much more effectively than the non-deep ones on the image datasets, probably due to the robust representation learning ability of deep neural networks. According to the results shown in Tables 3 and 5, our proposed PICI method outperforms the baseline



Table 4: The **ACC** scores of different clustering methods on the six datasets.

Dataset	RSOD	UC-Merced	SIRI-WHU	AID	D0-40	Chaoyang
<i>K</i> -means [45]	0.388	0.200	0.229	0.163	0.204	0.320
SC [46]	0.425	0.183	0.210	0.123	0.195	0.312
AC [47]	0.371	0.188	0.222	0.151	0.209	0.329
NMF [48]	0.420	0.208	0.275	0.161	0.187	0.305
PCA [49]	0.388	0.198	0.227	0.173	0.220	0.320
BIRCH [50]	0.396	0.202	0.222	0.147	0.205	0.329
GMM [51]	0.382	0.193	0.239	0.169	0.189	0.318
DEC [1]	0.534	0.147	0.257	0.185	0.232	0.421
IDEC [2]	0.458	0.141	0.255	0.192	0.213	0.424
ASPC-DA [4]	0.464	0.073	0.183	0.079	0.107	0.325
IDFD [9]	0.595	0.456	0.545	0.628	0.507	0.512
CC [8]	0.538	0.480	0.604	0.622	0.511	0.575
DCV [26]	0.418	0.121	0.195	0.100	0.095	0.321
HTCN [27]	0.584	0.508	0.496	0.709	<b>0.576</b>	0.547
PICI (Ours)	<b>0.772</b>	<b>0.634</b>	<b>0.672</b>	<b>0.748</b>	0.568	<b>0.595</b>

Table 5: The **ARI** scores of different clustering methods on the six datasets.

Dataset	RSOD	UC-Merced	SIRI-WHU	AID	D0-40	Chaoyang
<i>K</i> -means [45]	0.075	0.065	0.053	0.051	0.080	0.017
SC [46]	0.096	0.038	0.041	0.029	0.039	0.005
AC [47]	0.071	0.057	0.057	0.048	0.080	0.010
NMF [48]	0.052	0.089	0.118	0.056	0.068	0.002
PCA [49]	0.075	0.064	0.063	0.054	0.088	0.017
BIRCH [50]	0.068	0.066	0.049	0.046	0.080	0.010
GMM [51]	0.069	0.062	0.062	0.053	0.074	0.016
DEC [1]	0.325	0.053	0.083	0.075	0.105	0.006
IDEC [2]	0.144	0.042	0.079	0.073	0.093	–
ASPC-DA [4]	0.005	0.002	0.035	0.014	0.021	0.005
IDFD [9]	0.362	0.354	0.389	0.547	0.439	0.259
CC [8]	0.371	0.356	0.450	0.550	0.423	0.343
DCV [26]	0.144	0.020	0.043	0.025	0.017	0.017
HTCN [27]	0.465	0.359	0.336	0.646	0.470	0.262
PICI (Ours)	<b>0.510</b>	<b>0.492</b>	<b>0.518</b>	<b>0.663</b>	<b>0.500</b>	<b>0.382</b>

methods on all the six benchmark datasets. Especially, PICI surpasses the most competitive baseline (i.e., IDFD) by 0.177 on RSOD in terms of ACC. On the UC-Merced dataset, our model results in an NMI of 0.681, an ACC of 0.634, and an ARI of 0.492, which exhibit a significant margin of 11.8%, 24.8%, and 36.7%, respectively, over the second best scores. On the other datasets, similar

Table 6: The influence of different ViT architectures.

Dataset	Model	Dimension	#Layers	#Heads	NMI	ACC	ARI
RSOD	ViT-Tiny	192	4	12	0.523	0.750	0.480
	ViT-Small	384	6	12	<b>0.583</b>	<b>0.772</b>	<b>0.510</b>
	ViT-Base	768	12	12	—	—	—
Chaoyang	ViT-Tiny	192	4	12	0.335	0.572	0.330
	ViT-Small	384	6	12	<b>0.382</b>	<b>0.595</b>	<b>0.382</b>
	ViT-Base	768	12	12	0.357	0.579	0.346

advantages of PICI can also be seen in comparison with the other non-deep and deep clustering methods.

#### 4.4. Ablation Study

In this section, we experimentally analyze the influence of different components in PICI. Specifically, the influence of the ViT architecture in PISD, the influence of the PISD and CLI modules, and the influence of the two contrastive projectors in PICD will be tested in Sections.

##### 4.4.1. Influence of ViT Architectures

To assess the impact of the backbone, we evaluate several ViT architectures of different scales: ViT-Tiny, ViT-Small, and ViT-Base [41]. Table 6 details the different ViT network architectures and their corresponding clustering results. For this ablation study, we default the mask ratio to 0.5 and maintain the training strategy described in Section 3.4. For smaller datasets, such as RSOD, we observe that the ViT-Base runs with masked random sampling may be unstable. Nevertheless, the ViT-Small architecture consistently outperforms both ViT-Tiny and ViT-Base on the RSOD and Chaoyang datasets. Empirically, when using ViT-Small as the backbone, our proposed PICI approach is adept at learning robust representations while striking a balance between effectiveness and efficiency.

##### 4.4.2. Influence of PISD and CLI

In our PICI model, we incorporate an image reconstruction task that employs ViT as the backbone combined with masked random sampling in the PISD module, and seek to cultivate more reliable and discriminative representations by enabling mutual interaction between instance- and cluster-level spaces in the CLI module. To assess the impact of PISD and CLI, we exclude each in turn to ascertain their individual significance. When PISD is excluded, we substitute the CNN (e.g., ResNet34) with ViT as the backbone, which only consists of ResNet and PICD and can be regarded as a vanilla CC [8]. Evidently, the performance achieved by only employing PISD surpasses that of vanilla CC, implying that ViT offers superior representation learning capability compared to CNN. Furthermore, integrating both PISD and CLI yields more remarkable

Table 7: The influence of the PISD and CLI modules.

Dataset	PISD	CLI	NMI	ACC	ARI
RSOD	$\times$	$\times$	0.489	0.550	0.409
	$\checkmark$	$\times$	0.527	0.564	0.434
	$\checkmark$	$\checkmark$	<b>0.583</b>	<b>0.772</b>	<b>0.510</b>
Chaoyang	$\times$	$\times$	0.211	0.490	0.188
	$\checkmark$	$\times$	0.278	0.533	0.263
	$\checkmark$	$\checkmark$	<b>0.382</b>	<b>0.595</b>	<b>0.382</b>

Table 8: The influence of the two contrastive projectors in the PICD module.

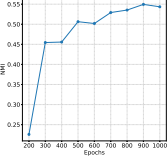
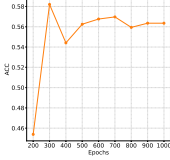
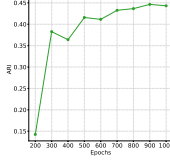
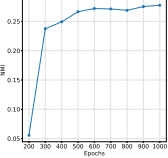
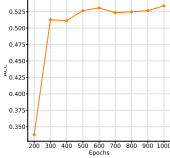
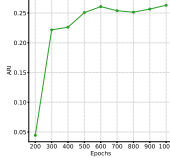
Dataset	Instance-MLP	Cluster-MLP	NMI	ACC	ARI
RSOD	$\checkmark$	$\times$	0.342	0.544	0.159
	$\times$	$\checkmark$	0.293	0.524	0.244
	$\checkmark$	$\checkmark$	<b>0.441</b>	<b>0.546</b>	<b>0.367</b>
Chaoyang	$\checkmark$	$\times$	0.307	0.431	0.201
	$\times$	$\checkmark$	0.124	0.432	0.108
	$\checkmark$	$\checkmark$	<b>0.351</b>	<b>0.568</b>	<b>0.331</b>

clustering results than using PISD only, which confirms the contribution of both the PISD and CLI modules.

#### 4.4.3. Influence of Contrastive Projectors

In the PICD module, we enforce contrastive learning with two contrastive projectors, namely, the instance-MLP for the instance contrastive loss and the cluster-MLP for the cluster contrastive loss. To evaluate the efficacy of these two projectors (corresponding to two contrastive losses, respectively), we perform ablation studies by excluding one projector at a time and then training the model from scratch. When the cluster-MLP is excluded, we simply apply the  $K$ -means method to the representations produced by the instance-MLP to obtain the final cluster assignments. As depicted in Table 8, jointly employing both instance-MLP and cluster-MLP consistently yields superior clustering performance (w.r.t. NMI, ACC, and ARI) compared to using just one projector. Specifically, on the RSOD dataset, only using the instance-MLP leads to an NMI of 0.342, while only employing the cluster-MLP yields a lower NMI of 0.293. In terms of the NMI on the Chaoyang dataset, the performance of the cluster-MLP is much better than that of the instance-MLP. To summarize, we have two observations from the experimental results. First, the instance-MLP can usually learn more discriminative representations than the cluster-MLP. Second, the joint incorporation of these two contrastive projectors can provide more robust clustering performance than using one projector only.

Table 9: The performance curves of our PICI method (as the number of epochs increases) on the RSOD and Chaoyang datasets.

Dataset	NMI	ACC	ARI
RSOD			
Chaoyang			

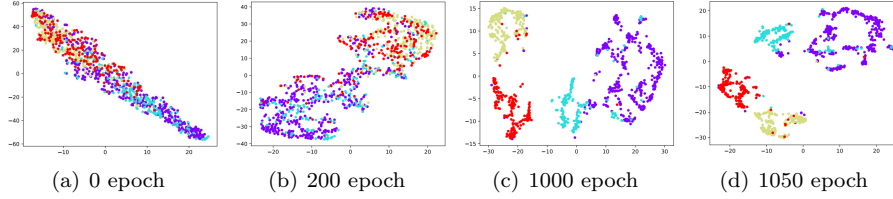


Figure 3: The t-SNE visualization of PICI on the RSOD dataset.

#### 4.5. Convergence Analysis

In this section, we assess the convergence of the proposed PICI method with an increasing number of epochs. Specifically, we report the clustering scores (w.r.t. NMI, ACC, and ARI) on the RSOD and Chaoyang datasets every 100 epochs, as illustrated in Table 9. The NMI, ACC, and ARI scores exhibit a sharp increase during the initial 200 epochs on the benchmark datasets. Following this, the scores of the proposed PICI method consistently improve with more epochs and eventually stabilize.

Furthermore, to visualize the convergence of PICI, we apply t-SNE [52] on the representations learned by the instance-MLP. Each instance-level feature representation is marked with a specific color according to cluster assignment predicted by the cluster-MLP. As shown in Fig. 3, at the beginning, the distribution of the feature representations are chaotic and mostly mixed. As the training process goes on, the distribution of feature representations becomes more distinguishable and balanced and gradually reaches stability with strengthened intra-cluster compactness and well-separated clusters.

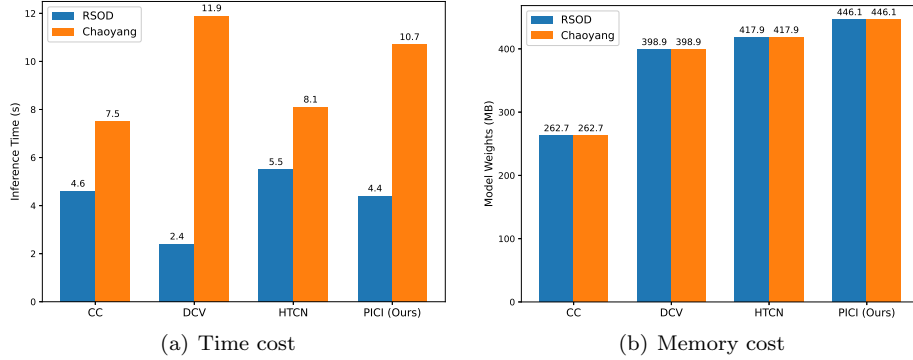


Figure 4: The inference time and memory costs of PICI on the RSOD and Chaoyang datasets in comparison with three baseline methods.

#### 4.6. Computational Analysis

In this section, we analyze the computational cost of our PICI method. Specifically, we compare the inference time and memory costs of our PICI method against three state-of-the-art deep clustering methods, namely, CC [8], DCV [26], and HTCEN [27]. As shown in Fig. 4, in terms of inference time, our proposed PICI method achieves competitive efficiency on the RSOD and Chaoyang datasets. Although PICI is slightly slower than CC and HTCEN on the Chaoyang dataset, it is worth noting that PICI achieves consistently better clustering performance than these baseline methods without a significant increase in model weight, which helps compensate for the efficiency issue.

### 5. Conclusion and Future Work

In this paper, we present a novel deep image clustering approach termed PICI, which enforces the partial information discrimination and cross-level interaction in a joint learning framework. In particular, we utilize a Transformer encoder as the backbone, through which the masked image modeling (MIM) with two parallel augmented views is formulated. After deriving the class tokens from the masked images by the Transformer encoder, three partial information learning modules are further incorporated, including the PISD module for training the auto-encoder via masked image reconstruction, the PICD module for employing two levels of contrastive learning, and the CLI module for mutual interaction between the instance- and cluster-level subspaces. Extensive experiments have been conducted on six real-world image datasets, which demonstrate the superior clustering performance of the proposed PICI approach over the state-of-the-art deep clustering approaches. Especially, on the RSOD (or UC-Merced) dataset, PICI achieves an ACC of 0.772 (or 0.634), which exhibits an improvement of 29.7% (or 24.8%) over the best baseline.

To the best of our knowledge, this paper for the first time bridges the gap between deep contrastive clustering and MIM. While MIM is typically designed for image data, which aims to enhance the image representation learning by

reconstructing masked images, our approach based on MIM is thereby inherently suitable for images, but may not be directly feasible for some other types of data. In the future work, we plan to extend our proposed framework from image data to more data types, such as text data and time series data, and possibly from general-scale datasets to large-scale, noisy, or even incomplete datasets.

In total, our proposed PICI framework offers significant contributions from both theoretical research and practical application perspectives. On one hand, as an unsupervised clustering framework, our proposed PICI can be easily transferred to other downstream tasks such as image classification, unsupervised anomaly detection and semantic segmentation. On the other hand, PICI notably exhibits exceptional clustering performance on remote sensing data, suggesting its potential value in practical applications such as remote sensing image classification and detection.

## Acknowledgments

This work was supported by the NSFC (61976097), the Natural Science Foundation of Guangdong Province (2021A1515012203), and the Science and Technology Program of Guangzhou, China (202201010314).

## References

- [1] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: Proc. of International Conference on Machine Learning (ICML), 2016.
- [2] X. Guo, L. Gao, X. Liu, J. Yin, Improved deep embedded clustering with local structure preservation., in: Proc. of International Joint Conference on Artificial Intelligence (IJCAI), 2017.
- [3] X. Guo, E. Zhu, X. Liu, J. Yin, Deep embedded clustering with data augmentation, in: Proc. of Asian Conference on Machine Learning (ACML), PMLR, 2018, pp. 550–565.
- [4] X. Guo, X. Liu, E. Zhu, X. Zhu, M. Li, X. Xu, J. Yin, Adaptive self-paced deep clustering with data augmentation, IEEE Transactions on Knowledge and Data Engineering 32 (2020) 1680–1693.
- [5] Z. Kang, X. Lu, Y. Lu, C. Peng, W. Chen, Z. Xu, Structure learning with similarity preserving, Neural Networks 129 (2020) 138–148.
- [6] M. Zhao, W. Yang, F. Nie, Deep graph reconstruction for multi-view clustering, Neural Networks 168 (2023) 560–568.
- [7] X. Deng, D. Huang, D.-H. Chen, C.-D. Wang, J.-H. Lai, Strongly augmented contrastive clustering, Pattern Recognition 139 (2023) 109470.

- [8] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, X. Peng, Contrastive clustering, in: Proc. of AAAI Conference on Artificial Intelligence (AAAI), 2021.
- [9] Y. Tao, K. Takagi, K. Nakata, Clustering-friendly representation learning via instance discrimination and feature decorrelation, arXiv preprint arXiv:2106.00131 (2021).
- [10] J. Li, P. Zhou, C. Xiong, S. C. Hoi, Prototypical contrastive learning of unsupervised representations, arXiv preprint arXiv:2005.04966 (2020).
- [11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16000–16009.
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: Proc. of International Conference on Learning Representations (ICLR), 2021.
- [13] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE transactions on pattern analysis and machine intelligence 43 (2020) 4037–4058.
- [14] Z. Wu, Y. Xiong, S. X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3733–3742.
- [15] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proc. of International Conference on Machine Learning (ICML), 2020.
- [16] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
- [17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, in: Advanced in Neural Information Processing Systems (NeurIPS), 2020.
- [18] X. Chen, K. He, Exploring simple siamese representation learning, in: Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15750–15758.
- [19] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, in: Advanced in Neural Information Processing Systems (NeurIPS), 2020.

- [20] M. Assran, M. Caron, I. Misra, P. Bojanowski, F. Bordes, P. Vincent, A. Joulin, M. Rabbat, N. Ballas, Masked siamese networks for label-efficient learning, in: Proc. of European Conference on Computer Vision (ECCV), Springer, 2022, pp. 456–473.
- [21] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, J. Feng, Contrastive masked autoencoders are stronger vision learners, arXiv preprint arXiv:2207.13532 (2022).
- [22] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, J. Wang, Context autoencoder for self-supervised representation learning, arXiv preprint arXiv:2202.03026 (2022).
- [23] Z. Ren, X. Kong, Y. Zhang, S. Wang, UKSSL: Underlying knowledge based semi-supervised learning for medical image classification, IEEE Open Journal of Engineering in Medicine and Biology (2023).
- [24] Z. Ren, S. Wang, Y. Zhang, Weakly supervised machine learning, CAAI Transactions on Intelligence Technology 8 (2023) 549–580.
- [25] Y. Zhang, L. Deng, H. Zhu, W. Wang, Z. Ren, Q. Zhou, S. Lu, S. Sun, Z. Zhu, J. M. Gorriz, et al., Deep learning in food category recognition, Information Fusion 98 (2023) 101859.
- [26] L. Wu, L. Yuan, G. Zhao, H. Lin, S. Z. Li, Deep clustering and visualization for end-to-end high-dimensional data analysis, IEEE Transactions on Neural Networks and Learning Systems 34 (2023) 8543–8554.
- [27] X. Deng, D. Huang, C.-D. Wang, Heterogeneous tri-stream clustering network, Neural Processing Letters 55 (2023) 6533–6546.
- [28] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep clustering for unsupervised learning of visual features, in: Proc. of European conference on computer vision (ECCV), 2018, pp. 132–149.
- [29] H. Zhong, J. Wu, C. Chen, J. Huang, M. Deng, L. Nie, Z. Lin, X.-S. Hua, Graph contrastive clustering, in: Proc. of IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9224–9233.
- [30] X. Wang, G.-J. Qi, Contrastive learning with stronger augmentations, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 5549–5560.
- [31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advanced in Neural Information Processing Systems (NeurIPS), 2017.
- [32] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).



- [33] Y. Li, M. Yang, D. Peng, T. Li, J. Huang, X. Peng, Twin contrastive learning for online clustering, *International Journal of Computer Vision* 130 (2022) 2205–2221.
- [34] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, L. He, Multi-level feature learning for contrastive multi-view clustering, in: *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16051–16060.
- [35] Y. Long, Y. Gong, Z. Xiao, Q. Liu, Accurate object localization in remote sensing images based on convolutional neural networks, *IEEE Transactions on Geoscience and Remote Sensing* 55 (2017) 2486–2498.
- [36] Y. Yang, S. Newsam, Bag-of-visual-words and spatial extensions for land-use classification, in: *Proc. of SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.
- [37] B. Zhao, Y. Zhong, G.-S. Xia, L. Zhang, Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery, *IEEE Transactions on Geoscience and Remote Sensing* 54 (2015) 2108–2123.
- [38] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, X. Lu, Aid: A benchmark data set for performance evaluation of aerial scene classification, *IEEE Transactions on Geoscience and Remote Sensing* 55 (2017) 3965–3981.
- [39] C. Xie, R. Wang, J. Zhang, P. Chen, W. Dong, R. Li, T. Chen, H. Chen, Multi-level learning features for automatic classification of field crop pests, *Computers and Electronics in Agriculture* 152 (2018) 233–241.
- [40] C. Zhu, W. Chen, T. Peng, Y. Wang, M. Jin, Hard sample aware noise robust learning for histopathology image classification, *IEEE Transactions on Medical Imaging* 41 (2021) 881–894.
- [41] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: *Proc. of International Conference on Machine Learning (ICML)*, 2021.
- [42] S.-G. Fang, D. Huang, X.-S. Cai, C.-D. Wang, C. He, Y. Tang, Efficient multi-view clustering via unified and discrete bipartite graph learning, *IEEE Transactions on Neural Networks and Learning Systems* (2023).
- [43] D. Huang, C.-D. Wang, J.-S. Wu, J.-H. Lai, C.-K. Kwok, Ultra-scalable spectral clustering and ensemble clustering, *IEEE Transactions on Knowledge and Data Engineering* 32 (2020) 1212–1226.
- [44] D. Huang, C.-D. Wang, J.-H. Lai, Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity, *IEEE Transactions on Knowledge and Data Engineering* 35 (2023) 11388–11402.

- [45] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proc. of Mathematical Statistics and Probability, 1967.
- [46] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: Advanced in Neural Information Processing Systems (NeurIPS), 2005.
- [47] K. C. Gowda, G. Krishna, Agglomerative clustering using the concept of mutual nearest neighbourhood, Pattern Recognition 10 (1978) 105–12.
- [48] D. Cai, X. He, X. Wang, H. Bao, J. Han, Locality preserving nonnegative matrix factorization, in: Proc. of International Joint Conference on Artificial Intelligence (IJCAI), 2009.
- [49] A. Martinez, A. Kak, Pca versus lda, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 228–233.
- [50] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: An efficient data clustering method for very large databases, in: Proc. of SIGMOD International Conference on Management of Data, 1996.
- [51] C. Fraley, A. E. Raftery, Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST, Journal of Classification 20 (2003) 263–286.
- [52] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.