entropy

MDPI

Article

ICIRD: Information-Principled Deep Clustering for Invariant, Redundancy-Reduced and Discriminative Cluster Distributions

Aiyu Zheng [1,2,*], Robert M. X. Wu [3], Yupeng Wang [1] and Yanting He [1]

[1] School of Electronic Information and Engineering, Taiyuan University of Science and Technology, Taiyuan 030024, China; yupengwang@tyust.edu.cn (Y.W.); yantinghe@tyust.edu.cn (Y.H.)
[2] School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
[3] Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney 2007, Australia; mingxuan.wu@uts.edu.au
* Correspondence: zheng_aiyu@tyust.edu.cn

Abstract

Deep clustering aims to discover meaningful data groups by jointly learning representations and cluster probability distributions. Yet existing methods rarely consider the underlying information characteristics of these distributions, causing ambiguity and redundancy in cluster assignments, particularly when different augmented views are used. To address this issue, this paper proposes a novel information-principled deep clustering framework for learning invariant, redundancy-reduced, and discriminative cluster probability distributions, termed ICIRD. Specifically, ICIRD is built upon three complementary modules for cluster probability distributions: (i) conditional entropy minimization, which increases assignment certainty and discriminability; (ii) inter-cluster mutual information minimization, which reduces redundancy among cluster distributions and sharpens separability; and (iii) cross-view mutual information maximization, which enforces semantic consistency across augmented views. Additionally, a contrastive representation mechanism is incorporated to provide stable and reliable feature inputs for the cluster probability distributions. Together, these components enable ICIRD to jointly optimize both representations and cluster probability distributions in an information-regularized manner. Extensive experiments on five image benchmark datasets demonstrate that ICIRD outperforms most existing deep clustering methods, particularly on fine-grained datasets such as CIFAR-100 and ImageNet-Dogs.

Keywords: deep clustering; contrastive learning; discriminative learning; information-principled deep clustering; discriminative distribution sharpness; multi-view inter-cluster distribution redundancy reduction

Academic Editor: Friedhelm Schwenker

Received: 4 November 2025
Revised: 21 November 2025
Accepted: 25 November 2025
Published: 26 November 2025

Citation: Zheng, A.; Wu, R.M.X.; Wang, Y.; He, Y. ICIRD: Information-Principled Deep Clustering for Invariant, Redundancy-Reduced and Discriminative Cluster Distributions. *Entropy* **2025**, *27*, 1200. https://doi.org/10.3390/e27121200

1. Introduction

Deep clustering refers to a family of methods that jointly learn feature representations and cluster assignments through neural networks to produce semantically meaningful and discriminative clusters [1]. This paradigm has shown broad potential in visual, graph, and biological data, where clustering enables structural pattern discovery and semantic organization [2,3]. In numerous advanced deep clustering algorithms, the cluster probability distribution plays a central role as the functional carrier of the clustering objective [4]. Its differentiable probabilistic form characterizes an interpretable and measurable correspondence between samples and clusters, while in neural network models, it is generated from

learned feature representations, thereby building a relational bridge between representation learning and clustering optimization. With the incorporation of data augmentation and self-supervised learning into deep clustering frameworks, the performance of deep clustering has achieved unprecedented improvement [5].

However, existing methods still fail to sufficiently model the information characteristics of cluster probability distributions, especially under cross-view scenarios. Consequently, during practical training, these distributions are often disturbed by representation perturbations, augmentation discrepancies, and inter-class redundancy, leading to uncertain predictions, insufficient separability, and inconsistency across augmentations. This phenomenon is evident in representative early methods: DEC [6] iteratively refines embeddings and cluster assignments via a self-training target distribution, which can propagate pseudo-label errors and is sensitive to clustering hyperparameters; JULE [7] couples agglomerative clustering with a CNN in a recurrent loop and is vulnerable to noise and early merge decisions; DeepCluster [8] uses offline k-means to generate hard pseudo-labels for end-to-end training, but assignments can be unstable across iterations and sensitive to augmentation choices; IIC [9] maximizes mutual information of paired augmentations to enforce consistency, but performance hinges on augmentation design and captured correlations. Accordingly, stabilizing the modeling of the cluster probability distribution is a key challenge for improving robustness and performance in deep clustering.

Building on the advances of contrastive learning [10–12], recent works have aimed to alleviate the aforementioned limitations by strengthening the discriminability and consistency of cross-view cluster distributions in joint representation–clustering frameworks. Contrastive Clustering (CC) [13] employs instance-level and cluster-level contrast to reduce augmentation-induced prediction bias. DCRN [14] reduces redundancy by minimizing dual correlations at the feature and cluster levels. DeepCluE [15] integrates multi-layer ensemble clustering with contrastive representations to enhance clustering robustness and accuracy. DHCL [16] enforces hierarchical contrastive constraints across global and local views, enhancing both intra-cluster compactness and cross-view consistency. CCGCC [17] further models the information characteristics of cluster distributions through cross-cluster and graph-level contrastive objectives.

Nevertheless, most recent methods constrain the cluster probability distribution from only a single or partial perspective, lacking a systematic and comprehensive modeling paradigm. On the other hand, approaches based on cross-view cluster distribution contrast often overlook the cluster alignment problem. When the cluster assignments of the same instance are inconsistent across different views, such contrastive operations may instead undermine the stability and accuracy of the model [5]. Therefore, there remains considerable room for exploration in the field of deep clustering.

To address the aforementioned challenges, this paper proposes an information-principled deep clustering framework for learning invariant, redundancy-reduced, and discriminative cluster probability distributions, termed ICIRD. Specifically, ICIRD models clustering as maximizing discriminative mutual information under a constrained information channel, thereby achieving a balance between information preservation and redundancy suppression. First, conditional entropy is minimized to enhance the certainty and discriminability of cluster assignments. Second, inter-cluster mutual information is minimized within each view to suppress statistical redundancy and further improve cluster separability. Finally, cross-view mutual information is maximized to reinforce the semantic invariance and stability of cluster assignments. In addition, a contrastive representation mechanism with both discriminative and information-bottleneck properties is introduced to provide stable and reliable feature inputs for the cluster probability distributions. ICIRD realizes a unified information-principled learning paradigm that bridges representation and

clustering learning, enabling the model to learn sharp, independent, and consistent cluster distributions, thereby improving the stability and robustness of the clustering results. The major contributions of this work are summarized as follows:

- A unified information-principled deep clustering framework, termed ICIRD, is proposed. It systematically imposes information constraints on the cluster probability distributions under data augmentation scenarios, covering discriminability, redundancy reduction, and cross-view invariance.
- Three complementary information-principled modules are designed for the cluster probability distribution. The DDS (Section 4.2) module minimizes conditional entropy to enhance assignment certainty and sharpness. The MIDR (Section 4.3) module suppresses redundancy of cluster distributions by minimizing inter-cluster mutual information within each view. The CIDC (Section 4.4) module maximizes cross-view mutual information to preserve the semantic structural stability of cluster assignments.
- Extensive experiments conducted on multiple benchmark datasets demonstrate that ICIRD achieves superior clustering performance compared with state-of-the-art methods. In addition, an analysis of cross-view IDR further shows that the view alignment strategy can partially alleviate the structural distortion caused by cross-view IDR.

The remainder of this paper is structured as follows. Section 2 reviews related studies about ICIRD. Section 3 introduces the preliminaries, including the fundamental information-theoretic concepts and problem definitions of ICIRD. Section 4 presents the proposed ICIRD framework and its information-principled objective formulation. Section 5 provides comprehensive experimental validation of ICIRD across multiple benchmarks. Section 6 illustrates the applications and scalability of ICIRD. Section 7 summarizes the main contributions and conclusions of this work.

2. Related Works

2.1. Deep Clustering

The core idea of deep clustering is to use deep neural networks to learn clustering-oriented representations and jointly optimize them with clustering objectives, overcoming the limitations of traditional methods on high-dimensional, nonlinear data [1,5]. Early approaches fall into three paradigms: embedding-based, generative-based and pseudo-labeling-based schemes [2,3]. DEC [6] first integrated deep embeddings with self-training to iteratively refine soft assignments. DAC [18] reformulated clustering as pairwise classification with adaptive pseudo-labels. DeepCluster [8] alternated between k-means pseudo-labeling and model updating for large-scale training.

These methods still face two core limitations: (i) Alternating optimization often introduces mismatches between feature and cluster spaces, where pseudo-label noise causes semantic drift and fuzzy boundaries. (ii) The absence of explicit modeling at the cluster-probability level limits the controllability of soft assignments. To overcome these issues, IIC [9] maximizes mutual information between augmented views of the same instance to prevent collapse and enforce probability-level consistency. ProtoCon [19] applies prototypical consistency and online pseudo-label refinement to mitigate noise and stabilize decision boundaries. Despite these advances, a unified theoretical framework that jointly ensures sharpness, independence, and cross-view consistency of probabilistic predictions remains elusive.

2.2. Mutual Information for Deep Clustering

Mutual Information (MI) measures the shared information between random variables and serves as a key objective in unsupervised and self-supervised learning. In deep models, direct MI computation is intractable, so it is typically approximated by differentiable bounds

optimized with neural discriminators or contrastive objectives. Representative methods include NWJ [20], MINE [21] and InfoNCE [10]. Besides, Deep InfoMax [22] extends MI by maximizing local–global MI to learn discriminative representations. Together, these approaches establish a unified framework that transforms MI into trainable objectives for representation learning and clustering.

In clustering, Mutual Information (MI) serves as a natural and interpretable optimization target. IMSAT [23] maximizes MI between inputs and discrete outputs with augmentation consistency; DCCM [24] extends MI to triplet interactions for better cluster discrimination; and DDC [25] maximizes MI between inputs and cluster labels with a semantic generator for interpretable clustering. MIMC [26] generalizes MI to multi-view settings, jointly optimizing completeness, compactness, and diversity while disentangling common and private information. DIB [27] estimates mutual information through kernel Gram matrices to realize compact and consistent representations for multi-view clustering. SDCIB [28] integrates contrastive and information-bottleneck objectives across feature and cluster levels to enhance accuracy and robustness in multi-modal clustering.

Overall, MI in deep clustering has progressed from representation-level to distribution-level modeling and from single-view to cross-view formulations, aligning with the separability and consistency of soft assignments. Yet a smooth transition from self-supervised to clustering-oriented representations remains unrealized, motivating the introduction of clustering-oriented information alignment for joint optimization between the encoder and clustering head—the central focus of this study.

2.3. Deep Contrastive Clustering

Contrastive deep clustering extends contrastive learning by constructing positive and negative pairs through data augmentations or multi-view generation, learning discriminative and cluster-friendly representations without labels [11,12]. Core designs include the following: (i) combining instance-level and cluster-level contrast to constrain both sample and assignment spaces, and (ii) maximizing cross-view mutual information to ensure distributional consistency under augmentations. These mechanisms are often integrated with InfoNCE loss, memory queues, and multi-view augmentation for scalable and stable training [12,29,30], while debiasing strategies further mitigate negative-sampling and long-tail effects [31].

Methodologically, CC [13] applies contrastive objectives at both instance and cluster levels and introduces column-space contrast on the assignment matrix to mitigate semantic drift and representation collapse. Prototype-based approaches further stabilize learning: SwAV [29] replaces explicit negatives with swapped assignments for robust large-scale pretraining, and PCL [30] jointly discovers prototypes and performs instance contrast to enhance semantic aggregation. To refine pseudo-labels, SCAN [32] enforces neighbor consistency, SPICE [33] filters labels via semantic confidence, and ProtoCon [19] integrates prototype consistency for online clustering. For structured data, GCC combines graph contrast with cluster-level optimization [34], DCRN reduces sample-level and feature-level redundancy [14]. SACC [35] aligns samples across augmentations using adaptive cluster-level contrast to enhance semantic consistency, while IcicleGCN [36] incorporates graph convolutional reasoning into instance–cluster contrast to exploit relational dependencies in structured data. CoHiClust [37] learns hierarchical cluster structures through contrastive representation learning and a differentiable tree-based head.

While contrastive deep clustering effectively improves representation discriminability and assignment consistency by coupling instance-level and cluster-level objectives, two common gaps persist: (i) Constraints in the feature domain and in the probability-output (clustering-head) domain are often optimized in a loosely coupled manner, lacking

a unified treatment of cluster independenceand distributional stability. (ii) Pseudo-label noise and strong augmentations can amplify instability in closed-loop training, impeding early-stage convergence. Motivated by these issues, this paper proposes an information-principled framework that directly enforces sharpness, inter-cluster independence, and cross-augmentation consistency at the clustering head, while leveraging contrastive learning to stabilize the representation space, thereby achieving coordinated optimization across the probability and feature domains.

3. Preliminaries

3.1. Information-Theoretic Preliminaries

Let the input random variable X be transformed via a parameterized mapping into an intermediate representation Z, which is then projected through a decision mapping to yield the output random variable Y. This setup induces a two-layer Markov chain:

$$X \to Z \to Y. \tag{1}$$

According to the *Data Processing Inequality* (DPI) [38], the flow of information satisfies:

$$I(X;Y) \leq I(X;Z), \tag{2}$$

indicating that the discriminative information contained in the output cannot exceed that of the learned representation.

In the context of discriminative clustering [4], Y is interpreted as a latent random variable representing the cluster assignment of each sample, and its distribution reflects the underlying cluster structure. The natural measure of discriminative dependence between X and Y is the mutual information, which can be expressed as:

$$I(X;Y) = H(Y) - H(Y \mid X), \tag{3}$$

where $H(Y)$ denotes the entropy of the marginal cluster distributions, and $H(Y \mid X)$ reflects the conditional entropy associated with each prediction.

The two-layer Markov chain suggests that the representation Z should retain the information in Z that is most relevant for predicting Y, while filtering out irrelevant variability. In parallel, a discriminative perspective emphasizes maximizing $I(X;Y)$ appropriately balancing $H(Y)$ and $H(Y \mid X)$. By jointly considering the discriminative clustering and the Information Bottleneck (IB) principle [39], ICIRD preserves relevant information while reducing redundancy, yielding cluster predictions that are both confident and robust [22].

3.2. Problem Definition

The problem addressed by ICIRD is defined as follows: given an unlabeled dataset $X = \{x_i\}_{i=1}^N$ with a predefined number of clusters K, the model consists of a parameterized encoder $f_\theta : X \to \mathbb{R}^d$, a clustering head $g_\phi : \mathbb{R}^d \to \Delta^{K-1}$ (where Δ^{K-1} denotes the K-class probability simplex), and a contrastive head $c_\rho : \mathbb{R}^d \to \mathbb{R}^m$ for representation-level contrastive learning. Denote the conditional probability output of the model for X is $P_{\theta,\phi}(Y \mid X) = g_\phi(f_\theta(X))$, where $p_{\theta,\phi}(y_k \mid x_i) = [g_\phi(f_\theta(x_i))]_k$ and the predicted cluster assignment is $\hat{Y} = \arg\max_k P_{\theta,\phi}(Y|X)$, where $k \in K$. Under a random augmentation set $\mathcal{T}$, the model is trained to learn parameters (θ, ϕ, ρ) such that the predicted assignments $\{\hat{y}_i\}_{i=1}^N$ align with the latent semantic labels $\{y_i\}_{i=1}^N$ up to a permutation of cluster indices.

4. Proposed Method

The ICIRD framework is designed to maximize discriminative information under a constrained channel, preserving cluster-relevant content while compressing redundant inputs. The DDS module sharpens conditional cluster distributions to produce confident predictions under diverse augmentations, thereby reducing decision uncertainty. The MIDR module minimizes inter-dimensional redundancy within the probability space to decouple categories and improve separability. The CIDC module enforces consistent predictions for identical instances across augmentations, preventing semantic drift and structural fragmentation. The CRA backbone strengthens local semantics and instance discrimination, providing a stable foundation for distribution-level optimization. Together, these four modules form an information-preserving and redundancy-suppressing loop that enables ICIRD to learn sharp, independent, and consistent cluster probabilities, which are directly used as the final clustering outputs (see Figure 1).

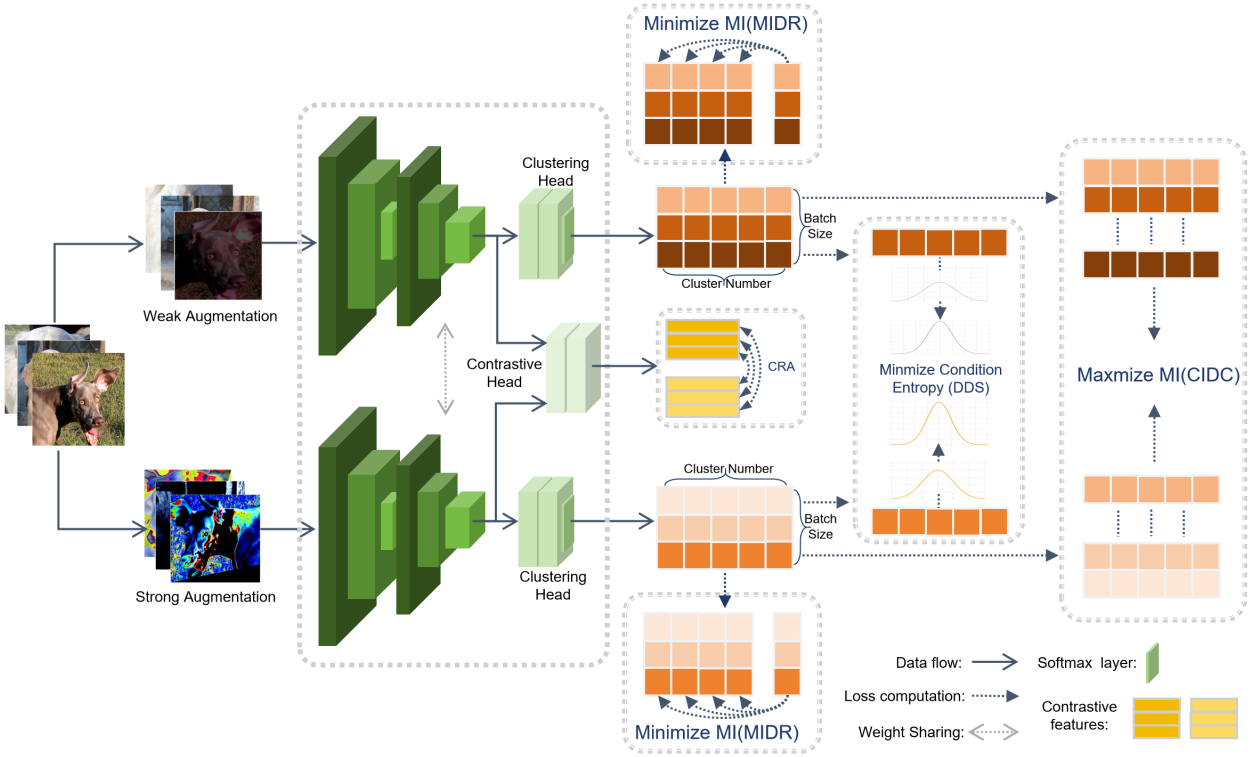


Figure 1. The framework of the proposed ICIRD. The model receives dual-view augmented data, which are processed by a shared encoder to extract representations. These representations are then fed into the clustering head and contrastive head, respectively. In the clustering branch, the DDS, MIDR, and CIDC modules jointly constrain the discriminability, consistency, and redundancy of the predicted distributions. In the contrastive branch, the CRA module aligns cross-view representations at the instance level. The four losses are jointly optimized to obtain discriminative and distributionally balanced clustering results.

4.1. Methodological Origins and Contributions

To improve the transparency of methodological origins and to clearly distinguish inherited components from novel contributions, we provide a unified explanation of the four modules that compose the ICIRD framework, namely CRA, CIDC, DDS, and MIDR. While some modules are directly based on established methods, others introduce structural modifications or new formulations tailored for deep clustering. Together, these components enable a principled, information-theoretic optimization of cluster probability distributions.

The CRA module follows the standard formulation of contrastive learning methods such as SimCLR [12] and InfoNCE [10]. It is adopted without modification in this work. Its primary purpose is to provide stable, invariant, and information-bottleneck–oriented feature representations, thereby supplying reliable inputs for subsequent cluster distribution learning rather than introducing a new contrastive objective.

The CIDC module draws directly from the cross-view mutual information maximization principle used in Invariant Information Clustering (IIC) [9]. By enforcing consistent cluster assignment distributions across augmented views of the same instance, CIDC ensures semantic stability under data augmentation. Its form remains consistent with the original IIC formulation, serving as a complementary constraint alongside the discriminative and redundancy-oriented components of ICIRD.

Building upon IMSAT's discriminative regularization [23], the DDS module introduces a key structural modification. Whereas IMSAT minimizes conditional entropy while maximizing marginal entropy, ICIRD removes the marginal entropy term and retains only conditional entropy minimization. This adjustment is motivated by the fact that CIDC and MIDR already enforce distributional structures related to marginal entropy, making its explicit inclusion redundant. By focusing solely on conditional entropy, DDS provides a clearer and more direct sharpening signal for cluster probability distributions, better aligning with the objectives of deep clustering. This modification constitutes a structural contribution specific to the clustering scenario.

Extending beyond the principles of IIC and CIDC, the MIDR module reconstructs an inter-cluster redundancy measure based on statistical mutual information and minimizes it to suppress redundant dependencies between different cluster assignments. Unlike IIC, which primarily emphasizes cross-view invariance, MIDR explicitly addresses the separability and independence of clusters by reducing inter-cluster statistical overlap. This formulation has not been previously established in deep clustering and therefore represents a novel redundancy-oriented information constraint.

Overall, the main contribution of ICIRD lies not in proposing four entirely new modules, but in formulating a unified, information-theoretic optimization framework for cluster probability distributions. DDS enhances distribution discriminability, MIDR suppresses inter-cluster redundancy, CIDC enforces cross-view invariance, and CRA provides stable representation support. These components collectively yield a structured and complementary optimization of both feature representations and cluster distributions, forming the core innovation of the ICIRD framework.

4.2. Discriminative Distribution Sharpness Module

The discriminative mutual information satisfies Equation (3). In fully unsupervised clustering, directly controlling the marginal entropy $H(Y)$ on mini-batches often incurs distributional bias and training instability [40]. In contrast, ICIRD adopts the strategy of minimizing the conditional entropy $H(Y \mid X)$ offers a more stable and direct route: it encourages sharper and more confident conditional predictions, thereby—under a controlled $H(Y)$—equivalently increasing $I(X;Y)$ via its deterministic component [23,39].

For random variables X and Y, the conditional entropy is:

$$H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x) \log p(y \mid x). \tag{4}$$

Since the true $P(X)$ is unknown, it is approximated by the empirical distribution $\hat{p}(X)$ at the mini-batch or dataset level. With a parameterized encoder and clustering head

producing discriminative cluster probabilities $p_{\theta,\phi}(y \mid x)$, and batch size B, a Monte-Carlo estimator the conditional entropy:

$$\hat{H}(Y \mid X) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K} p_{\theta,\phi}(y_k \mid x_i)\log p_{\theta,\phi}(y_k \mid x_i). \tag{5}$$

Under two augmented views a, b, the variance is reduced by averaging the two conditional entropies and defining the discriminative distribution sharpness (DDS) loss as:

$$\mathcal{L}_{\text{DDS}} = \tfrac{1}{2}\Big[\hat{H}(Y^a \mid X^a) + \hat{H}(Y^b \mid X^b)\Big]. \tag{6}$$

This loss penalizes flat predictive distributions and compresses conditional uncertainty. Meanwhile, the marginal entropy term $H(Y)$ in the mutual information formulation is handled through the following mechanisms: (i) Multi-view Inter-cluster Distribution Redundancy Reduction (MIDR) module suppresses inter-cluster statistical coupling and effectively increases the usable marginal entropy $H(Y)$, thus forming—together with DDS—a complete, structured maximization of $I(X;Y)$. (ii) Cross-view Instance Distribution Consistency (CIDC) module stabilizes $p_{\theta,\phi}(y \mid x)$ against view perturbations. Combined, these modules mitigate the degeneration and collapse risks inherent to pure entropy minimization and shape sharp, separable, and augmentation-consistent cluster predictions under minimal assumptions.

4.3. Multi-View Inter-Cluster Distribution Redundancy Reduction Module

Inspired by IIC [9] and data augmentation, a mutual information-based multi-view inter-cluster distribution redundancy reduction (MIDR) constraint is proposed to minimize mutual information between cluster distributions. This method directly acts on the marginal distributions over clusters, reducing redundant dependencies among cluster probabilities from a probabilistic perspective, thus enhancing the separability and independence of predictive distributions. Compared with conventional deep clustering algorithms that impose constraints on the representation space, MIDR optimizes the cluster assignment layer directly, achieving more effective performance improvement. Moreover, data augmentation helps mitigate the marginal bias introduced by mini-batch sampling, providing stable support for improved discrimination and generalization.

As previously described, a neural network produces a probability matrix $P_{\theta,\phi}(Y \mid X) \in \mathbb{R}^{B\times K}$ (abbreviated as P), where B denotes the batch size and K the number of categories. In clustering, the goal is to make the dimensions of this matrix (i.e., its column vectors) as independent as possible to ensure discriminative cluster probabilities. Therefore, the mutual information between category dimensions is minimized to suppress redundancy:

$$min_{i\neq j} I(u_i, u_j), \tag{7}$$

where u_i and u_j are two distinct probability columns of P. According to the KL-divergence definition of mutual information,

$$I(X;Y) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right], \tag{8}$$

the joint distribution $P(u_i, u_j)$ and the marginal distributions $P(u_i)$ and $P(u_j)$ are required. Furthermore, since the correlation between dimensions is positively related to their joint distribution, the joint distribution through inter-dimensional correlations is approximated. The correlation matrix D is obtained after column-wise normalization of P as $D = (P(Y|X))^{\top} P(Y|X)$. The matrix $D \in \mathbb{R}^{K\times K}$ expresses the full inter-dimensional

relationships, and once normalized as a probabilistic matrix, it can be used to compute the mutual information.

Data augmentation is then introduced, where X^a and X^b are two augmented versions of the unlabeled dataset X. The parameterized neural network with ϕ and θ produces corresponding probability matrices P^a and P^b. Each row vector in P^a or P^b represents a sample's conditional probability over K categories, while each column vector estimates the marginal distribution over clusters.

Based on the previous correlation formulation, D^a and D^b are obtained for each view, where D^a_{ij} and D^b_{ij} denote the redundancy between dimensions i and j. Let $\text{sum}(D^a)$ and $\text{sum}(D^b)$ denote the total sums of each matrix; they are normalized as $D^a = \frac{D^a}{\text{sum}(D^a)}$ and $D^b = \frac{D^b}{\text{sum}(D^b)}$. For any matrix D, the sum of row i gives the marginal $P(u_i) = \sum_{j=1}^{K} D_{ij}$, the sum of column j gives $P(u_j) = \sum_{i=1}^{K} D_{ij}$, and the element D_{ij} represents the joint $P(u_i, u_j)$. According to the KL-based definition in Equation (8), the inter-cluster distribution redundancy reduction loss (IDR) is given by:

$$\mathcal{L}_{\text{IDR}} = \frac{2}{K(K-1)} \sum_{i<j} P(u_i, u_j) \log \frac{P(u_i, u_j)}{P(u_i)\, P(u_j)}. \tag{9}$$

Because D is symmetric, only the upper triangular part is computed to avoid self-correlation and redundancy. In practice, unstable marginals may lead to trivial solutions, so smoothing terms can be added to stabilize training. Extending this formulation to the multi-view case, the MIDR loss becomes:

$$\mathcal{L}_{\text{MIDR}} = \mathcal{L}^a_{\text{IDR}} + \mathcal{L}^b_{\text{IDR}}. \tag{10}$$

Redundancy reduction is applied only within each view rather than across views, since augmented views may suffer from label permutation inconsistency. In the early stages of training, the cluster assignments of different views for the same instance are still likely to be inconsistent. Although cross-view consistency losses such as CIDC can partially mitigate this issue, enforcing cross-view independence early in training leads to instability before the label spaces align. We will discuss this issue in detail in Section 5.6. By minimizing inter-cluster mutual information in a multi-view setting, the MIDR loss effectively suppresses inter-class redundancy while maintaining semantic consistency, achieving joint optimization of class independence and view consistency. Compared with the single-view IDR, MIDR realizes higher robustness and output stability under strong augmentations.

4.4. Cross-View Instance Distribution Consistency Module

A mutual information-based cross-view instance distribution consistency (CIDC) constraint is introduced that is directly applied to the discriminative cluster probabilities, aiming to achieve an unbiased estimation for clustering tasks.

For a data set X and its two augmented versions X^a and X^b, after being mapped by parameterized neural networks ϕ and θ, the corresponding discriminative cluster probability matrices are $P(Y^a \mid X^a)$ and $P(Y^b \mid X^b)$. We focus on their row vectors q^a_i and q^b_i ($i \in N$), each representing the conditional probability distribution of sample i over K categories under the respective views. According to the consistency constraint, the cluster probabilities q^a_i and q^b_i of the same instance across different views should be similar, which in information theory corresponds to having a large mutual information. Therefore, the mutual information objective between these two vectors can be defined as:

$$\max I(q^a_i, q^b_i). \tag{11}$$

The joint probability matrix between the two views can be computed as $q_i^a \cdot (q_i^b)^\top$. Consequently, the averaged joint probability over all paired instances is expressed as:

$$Q = \frac{1}{B} \sum_{i=1}^{B} q_i^a \cdot (q_i^b)^\top, \tag{12}$$

where $Q \in \mathbb{R}^{K \times K}$ denotes the cross-view joint probability distribution matrix. Each element Q_{ij} represents the joint probability between category i in one view and category j in the other. The sum of the i-th row corresponds to the marginal probability Q_i, and the sum of the j-th column corresponds to Q_j. Considering the symmetry of mutual information, i.e., $I(q_i^a, q_i^b) = I(q_i^b, q_i^a)$, a symmetrized form is adopted in practice: $Q = \frac{1}{2}(Q + Q^\top)$ Substituting the above formulation into the KL-divergence expression of mutual information, the loss function is given by:

$$\mathcal{L}_{CIDC} = -\sum_{i=1}^{K} \sum_{j=1}^{K} Q_{ij} \log \frac{Q_{ij}}{Q_i Q_j} \tag{13}$$

The CIDC loss maximizes the mutual information between cluster probabilities of different augmented views, thereby implicitly enhancing the marginal entropy $H(Y)$ and promoting balanced cluster assignments. By directly aligning semantic distributions at the probabilistic output layer, CIDC achieves a balance between stability and discriminability in clustering and unsupervised classification tasks. Specifically, during the maximization process, according to the mutual information identity $I(Y^a; Y^b) = H(Y^a) - H(Y^a \mid Y^b) = H(Y^b) - H(Y^b \mid Y^a)$, the marginal entropy is simultaneously maximized. This loss effectively compensates for the deficiency of DDS loss in constraining marginal entropy.

4.5. Contrastive Representation Anchoring Module

To enhance the discriminability and robustness of the representation space, a contrastive constraint $\mathcal{L}_{\text{CRA}}$ [13,30] is introduced. Given a batch of samples $\{x_i\}_{i=1}^B$, for each sample, two augmented views are generated x_i^a, x_i^b, where a and b denote two augmentation policies. In practice, a "weak" augmentation (e.g., random crop, light flip, mild color jitter) and a "strong" augmentation (e.g., heavy color distortion, extreme cropping, large rotation) are used. Through the encoder $f_\theta(\cdot)$ and the contrastive projection head $c_\rho(\cdot)$, the following representations are obtained: $h_i^v = \text{norm}\big(c_\rho(f_\phi(x_i^v))\big), \quad v \in \{a, b\}$, where $\text{norm}(\cdot)$ denotes ℓ_2-normalization to stabilize the similarity measure.

Let $\text{sim}(u, v) = u^\top v$ denote cosine similarity (which coincides with the inner product after ℓ_2 normalization). Let $\tau > 0$ be a temperature hyperparameter. For each i, the positive sample pair is (h_i^a, h_i^b). The negative samples are taken from the other instances in the same batch, $\{h_j^b\}_{j \neq i}$ (and may be extended via a cross-batch memory bank). The one-direction contrastive objective is:

$$\ell_i^{a \to b} = -\log \frac{\exp\big(\text{sim}(h_i^a, h_i^b)/\tau\big)}{\sum_{j=1}^{N} \exp\big(\text{sim}(h_i^a, h_j^b)/\tau\big)}. \tag{14}$$

To avoid directional bias, the objective is symmetrized, and the CRA loss is expressed as:

$$\mathcal{L}_{\text{CRA}} = \frac{1}{2N} \sum_{i=1}^{N} \Big(\ell_i^{a \to b} + \ell_i^{b \to a}\Big). \tag{15}$$

CRA provides robust, view-invariant embeddings that reduce noise in discriminative cluster probabilities, improving the reliability of DDS, MIDR, and CIDC estimations. Be-

sides, it indirectly strengthens entropy-based discrimination and cross-view alignment, leading to more stable and separable cluster formation.

4.6. Theoretical Interpretation of Objectives

To clarify the theoretical foundation of the information-driven design, this section demonstrates how the three major modules, DDS, MIDR, and CIDC, jointly realize a structured maximization of discriminative mutual information in Equation (3). Specifically, DDS explicitly reduces the conditional entropy, while MIDR and CIDC implicitly increase the marginal entropy, together achieving a balanced optimization of the mutual information components.

The CIDC module maximizes the cross-view mutual information:

$$I(Y^a; Y^b) = H(Y^a) + H(Y^b) - H(Y^a, Y^b), \tag{16}$$

where $H(Y^a) \approx H(Y^b) \approx H(Y)$. Maximizing $I(Y^a; Y^b)$ effectively encourages larger $H(Y)$ while suppressing collapse, thereby enhancing balanced cluster utilization and distributional stability across augmented views.

The proposed MIDR objective enhances the marginal entropy $H(Y)$ by reducing inter-cluster redundancy. Let $P_{\theta,\phi}(Y \mid X) \in \mathbb{R}^{B \times K}$ denote the conditional cluster probability matrix and π_k the marginal class probability over cluster k. When cluster assignments are sharp (near one-hot), the dimensions become mutually exclusive. In this limit, minimizing inter-cluster mutual information drives the marginal distribution π toward uniformity, where $H(Y)$ reaches its maximum. Hence, under sharp assignments, MIDR is equivalent to maximizing $H(Y)$.

In the general soft-assignment case, each column of $P_{\theta,\phi}(Y \mid X)$ represents a probabilistic cluster dimension, and correlations among these columns imply statistical redundancy between cluster assignments. Minimizing such inter-cluster correlations through MIDR reduces over-dependence among dimensions, encouraging a more uniform marginal distribution $\pi = \{\pi_k\}_{k=1}^{K}$, where $\pi_k = \sum_j \tilde{D}_{kj}$. As the distribution flattens, the quadratic term $\sum_k \pi_k^2$ decreases, leading to an increase in the Rényi-2 entropy:

$$H_2(\pi) = -\log \sum_{k=1}^{K} \pi_k^2,$$

which provides a lower bound on the Shannon entropy $H(Y)$. Therefore, by decorrelating cluster dimensions and balancing marginal probabilities, MIDR effectively increases (or lower-bounds) $H(Y)$, yielding more independent and evenly distributed cluster assignments.

From the Information Bottleneck (IB) perspective in Equations (1) and (2), the three modules jointly realize the trade-off between *information sufficiency* and *compression*. MIDR and CIDC jointly realize the compression aspect of the IB principle by eliminating nuisance factors in clustering. Specifically, MIDR minimizes inter-dimensional correlations within the cluster probability space, thereby reducing statistical redundancy and encouraging more independent and balanced cluster representations. CIDC, on the other hand, maximizes cross-view invariance to minimize the dependence of cluster assignments on the augmentation variable T, effectively filtering out augmentation-induced noise and enforcing semantic consistency across views. Within this framework, DDS plays a complementary role: by minimizing the conditional entropy $H(Y \mid X)$, it sharpens predictions and enhances discriminative sufficiency, but does not directly contribute to the compression of $I(X; Y)$.

4.7. Information-Principled Objective Formulation

Without loss of generality, the overall objective loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{CRA}} + \lambda_1 \mathcal{L}_{\text{MIDR}} + \lambda_2 \mathcal{L}_{\text{CIDC}} + \lambda_3 \mathcal{L}_{\text{DDS}} \tag{17}$$

where $\mathcal{L}_{\text{CRA}}$ is the contrastive representation anchoring loss with Equation (15), $\mathcal{L}_{\text{MIDR}}$ is the multi-view inter-cluster distribution redundancy reduction loss with Equation (10), $\mathcal{L}_{\text{CIDC}}$ is the cross-view instance distribution consistency loss with Equation (13) and $\mathcal{L}_{\text{DDS}}$ is the discriminative distribution sharpness loss with Equation (6). λ_1, λ_2 and λ_3 are the balance hyper-parameters of each loss.

From a theoretical standpoint, the composite objective of ICIRD establishes a cooperative optimization mechanism that integrates representation compactness, distribution independence, and cross-view invariance within a unified information-principled framework. This design maximizes discriminative mutual information between inputs and predictions while suppressing redundant dependencies, thereby yielding compact yet expressive clustering representations. The overall training process is summarized in Algorithm 1.

Algorithm 1: ICIRD

Input : Dataset X; training epochs E; batch size B; temperature τ; cluster number K; hyper-parameters $\lambda_1, \lambda_2, \lambda_3$; strong augmentation $\mathcal{T}$, weak augmentation T, neural network g_ϕ, f_θ, c_ρ

Output: Clustering result $\{\hat{y}\}_{i=1}^{N}$

for $epoch = 1$ **to** E **do**
  sample a batch $\{x_i\}_{i=1}^{b}$ from X;
  obtain two augmentations X^a, X^b according to $X^a = \mathcal{T}(X)$ and $X^b = T(X)$;
  compute the representations Z^a and Z^b by $Z^a = f_\theta(X^a)$, $Z^b = f_\theta\left(X^b\right)$;
  compute the discriminative probabilities P^a, P^b through clustering head $P^a = g_\phi(Z^a)$ and $P^b = g_\phi(Z^b)$;
  compute the contrastive outputs H^a, H^b through contrastive head $H^a = c_\rho(Z^a)$ and $H^b = c_\rho(Z^b)$;
  compute DDS loss $\mathcal{L}_{\text{DDS}}$ through Equation (6);
  compute MIDR loss $\mathcal{L}_{\text{MIDR}}$ through Equation (10);
  compute CIDC loss $\mathcal{L}_{\text{CIDC}}$ through Equation (13);
  compute CRA loss $\mathcal{L}_{\text{CRA}}$ through Equation (15);
  update ϕ, θ, ρ through gradient descent to minimize in Equation (17);
end
Feed the samples $\{x_i\}_{i=1}^{N}$ into the network g_ϕ, f_θ to obtain the soft assignments $p_{\theta,\phi}(y_k|x_i)$;
Obtain the labels according to argmax: $\hat{y}_i = \arg\max_k p_{\theta,\phi}(y_k|x_i)$;
return $\{\hat{y}\}_{i=1}^{N}$;

5. Experiments

5.1. Experiment Setting

5.1.1. Datasets

Five well-known benchmark datasets are employed to comprehensively evaluate the effectiveness of the proposed method. CIFAR-10 [41] contains 10 balanced categories of natural images. CIFAR-100 [41] extends CIFAR-10 to 100 fine-grained classes grouped into 20 superclasses, providing a more challenging test of representational capacity and generalization. STL-10 [42] includes 10 categories with higher-resolution images. ImageNet-

10 [43] is a subset of the ImageNet dataset comprising 10 semantically distinct categories. ImageNet-Dogs [43] consists of 15 dog breeds and serves as a fine-grained benchmark. Table 1 provides detailed information about these datasets.

Table 1. The detailed information of datasets used in experiments.

Dataset	Class	Images Number	Image Size
CIFAR-10	10	60,000	$32 \times 32 \times 3$
CIFAR-100	20/100	60,000	$32 \times 32 \times 3$
STL-10	10	13,000	$96 \times 96 \times 3$
ImageNet-10	10	13,000	$224 \times 224 \times 3$
ImageNet-Dogs	15	19,500	$224 \times 224 \times 3$

5.1.2. Evaluation Metrics

Three widely used metrics are adopted to evaluate the clustering results, including Normalized Mutual Information (NMI) [44], Clustering Accuracy (ACC) [6], and Adjusted Rand Index (ARI) [45]. Note that higher values of the three evaluation metrics indicate better clustering performances.

5.1.3. Implementation Details

Unless otherwise specified, the neural network backbone is based on ResNet-34, which is randomly initialized before training to ensure a fair comparison with other deep clustering algorithms. The original 7×7 stride-2 convolution of ResNet is replaced with a 3×3 stride-1 convolution, and the max-pooling layer is removed to avoid early downsampling. This adjustment preserves fine-grained spatial information and makes the network better suited for small-resolution datasets such as CIFAR. For ImageNet-10 and ImageNet-Dogs, we resize the images to 224×224, while for Stanford-Dogs, we resize them to 96×96. Other datasets use their original image resolutions. To accommodate datasets with different image resolutions, the input layer of the model is appropriately modified. The feature dimension of the backbone's penultimate layer is set to 512 to preserve sufficient representational information. A two-layer MLP with ReLU activation functions is employed as the projection head for contrastive representation learning, producing a unified output dimension of 128. The temperature parameter τ in the contrastive loss is set to 0.5. Another two-layer MLP with Softmax activation is used as the clustering head, whose output dimension corresponds to the number of classes in each dataset, as summarized in Table 1. The model is optimized using the Adam optimizer with a learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}. The batch size is set to 256, and training is performed for 1500 epochs to ensure convergence. The balance parameters are set to $\lambda_1 = 0.5$, $\lambda_2 = 1$, and $\lambda_3 = 0.5$. Except the analysis of representation dimension conducted on an A100, all experiments are conducted on a workstation equipped with an NVIDIA RTX 3090Ti GPU, a 12th Gen Intel Core i9 CPU, and 64 GB RAM. For data augmentation, the weak augmentation strategy follows SimCLR [12], including RandomResizedCrop, ColorJitter, Grayscale, HorizontalFlip, and GaussianBlur. The strong augmentation strategy follows Strongly Augmented Contrastive Clustering [46], including AutoContrast, Brightness, Color, Contrast, Equalize, Identity, Posterize, Rotate, Sharpness, ShearX/Y, Solarize, and TranslateX/Y.

5.1.4. Comparison Methods

ICIRD was compared with other competing clustering methods, including: VAE [47], JULE [7], DCGAN [48], DEC [6], DAC [18], ID [49], DCCM [24], PICA [50], DRC [51], IDFD [52], CC [13], DCDC [53], DCSC [54], SACC [35], DeepCluE [15], IcicleGCN [36],

DCHL [16], CoHiClust [37], CCGCC [17]. Table 2 illustrates the clustering metric results on five image data sets, where the highest and second-highest values are tagged in red and blue, respectively.

Table 2. Clustering performance by different competing clustering algorithms on five data sets.

Datasets	**CIFAR-10**			**CIFAR-100**			**STL-10**			**ImageNet-10**			**ImageNet-Dogs**		
Metric	**ACC**	**NMI**	**ARI**	**ACC**	**NMI**	**ARI**	**ACC**	**NMI**	**ARI**	**ACC**	**NMI**	**ARI**	**ACC**	**NMI**	**ARI**
VAE [47]	0.291	0.245	0.167	0.152	0.108	0.040	0.282	0.200	0.146	0.334	0.193	0.168	0.179	0.107	0.079
JULE [7]	0.272	0.192	0.138	0.137	0.103	0.033	0.277	0.182	0.164	0.300	0.175	0.138	0.138	0.054	0.028
DCGAN [48]	0.315	0.265	0.176	0.151	0.120	0.045	0.298	0.210	0.139	0.346	0.225	0.157	0.174	0.121	0.078
DEC [6]	0.301	0.257	0.161	0.185	0.136	0.050	0.359	0.276	0.186	0.381	0.282	0.203	0.195	0.122	0.079
DAC [18]	0.522	0.396	0.306	0.238	0.185	0.088	0.470	0.366	0.257	0.527	0.394	0.302	0.275	0.219	0.111
ID [49]	0.440	0.309	0.221	0.267	0.221	0.108	0.514	0.362	0.285	0.632	0.478	0.420	0.365	0.248	0.172
DCCM [24]	0.623	0.496	0.408	0.327	0.285	0.173	0.482	0.376	0.262	0.710	0.608	0.555	0.383	0.321	0.182
PICA [50]	0.696	0.591	0.512	0.337	0.310	0.171	0.713	0.611	0.531	0.870	0.802	0.761	0.352	0.352	0.201
DRC [51]	0.727	0.621	0.547	0.367	0.356	0.208	0.747	0.644	0.569	0.884	0.830	0.798	0.389	0.384	0.233
IDFD [52]	0.815	0.711	0.663	0.425	0.426	0.264	0.756	0.643	0.575	0.954	0.898	0.901	0.591	0.546	0.413
CC [13]	0.790	0.705	0.637	0.429	0.431	0.266	0.850	0.764	0.726	0.893	0.859	0.822	0.429	0.445	0.274
DCDC [53]	0.699	0.585	0.506	0.349	0.310	0.179	0.734	0.621	0.547	0.879	0.817	0.787	0.365	0.360	0.207
DCSC [54]	0.798	0.704	0.644	0.469	0.452	0.293	0.865	**0.792**	**0.749**	0.904	0.867	0.838	0.443	0.462	0.299
SACC [35]	0.851	0.765	0.724	0.443	0.448	0.282	0.759	0.691	0.626	0.905	0.877	0.843	0.437	0.455	0.285
DeepCluE [15]	0.764	0.727	0.646	0.457	0.472	0.288	-	-	-	0.924	0.882	0.856	0.416	0.448	0.273
IcicleGCN [36]	0.807	0.729	0.660	0.461	0.459	0.311	-	-	-	**0.955**	0.904	0.905	0.415	0.456	0.279
DHCL [16]	0.801	0.710	0.654	0.446	0.432	0.275	0.821	0.726	0.680	-	-	-	0.511	0.495	0.359
CoHiClust [37]	0.839	0.779	0.731	0.437	0.467	0.229	0.613	0.584	0.474	0.953	**0.907**	0.899	0.355	0.411	0.232
CCGCC [17]	0.864	0.778	0.742	0.482	0.486	0.316	0.779	0.698	0.645	0.904	0.859	0.833	0.579	0.568	0.449
ICIRD	**0.877**	**0.801**	**0.750**	**0.504**	**0.496**	**0.333**	**0.887**	0.774	0.737	0.945	0.893	**0.906**	**0.601**	**0.571**	**0.464**

The highest and second highest values are tagged in **bold** and underline, respectively.

5.2. Quantitative Analysis of Clustering Results

To evaluate the clustering quality of ICIRD, its performance is presented in comparison with several baseline methods on five commonly used benchmark datasets in Table 2. As shown in the results, ICIRD consistently outperforms most existing methods across all datasets, with particularly outstanding results on CIFAR-100 and ImageNet-Dogs. Both datasets contain a large number of classes with subtle inter-class differences, indicating that ICIRD possesses stronger fine-grained discriminative ability and a superior capacity for modeling complex data distributions. Moreover, ICIRD achieves either the best or second-best performance on the remaining datasets, demonstrating its strong generalization capability and ability to perform well on both coarse-grained and fine-grained clustering tasks.

5.3. Clustering on Fine-Grained Dataset

To further verify its applicability in more challenging fine-grained scenarios, additional experiments were conducted on the Stanford-Dogs dataset. Stanford-Dogs [55]: Derived from ImageNet, this dataset contains 120 dog breeds with balanced samples and 96×96 images. It features complex backgrounds and fine-grained inter-class variations, posing a challenging benchmark for fine-grained recognition and representation learning.

Several competitive algorithms from Table 2 were selected as comparison methods and reproduced following their publicly recommended configurations. The results, reported in Table 3, clearly show that ICIRD achieves the best performance on this dataset, further confirming its superior discriminability and robustness under highly similar and complex data distributions. This superior performance can be attributed to the redundancy suppression mechanism and discriminative constraint introduced in the clustering head, which jointly promote feature independence in the representation space. In addition, by emphasizing consistency constraints during both representation learning and clustering,

ICIRD effectively encourages intra-class compactness, thereby enhancing the separability and stability of the overall clustering structure.

Table 3. Clustering performance on Stanford-Dogs dataset.

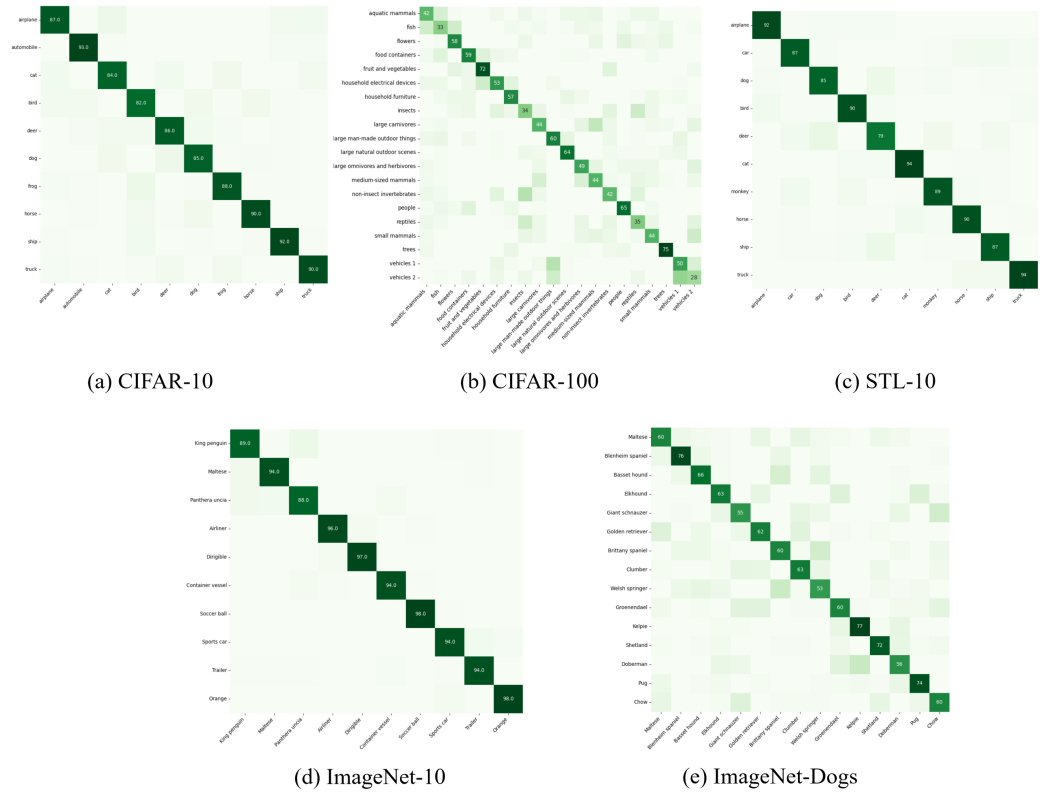
Dataset	ACC	NMI	ARI
DCSC	0.097	0.259	0.034
IDFD	0.132	0.331	<u>0.062</u>
CCGCC	<u>0.140</u>	<u>0.341</u>	0.059
IcicleGCN	0.082	0.235	0.035
DHCL	0.105	0.316	0.064
ICIRD	**0.176**	**0.366**	**0.101**

The highest and second highest values are tagged in **bold** and <u>underline</u>, respectively.

5.4. Qualitative Analysis of Clustering Results

5.4.1. Confusion Matrices

To further evaluate the class-wise discriminative capability of the proposed model, the confusion matrices on five benchmark datasets are visualized based on the ACC values in Table 2, as shown in Figure 2. These matrices provide an intuitive understanding of the intra-class consistency and inter-class separability achieved by our method across different levels of granularity and visual complexity. Overall, the confusion matrices collectively demonstrate that our model maintains strong discriminative ability under coarse-grained settings and exhibits reasonable generalization to fine-grained categories, although subtle appearance variations among closely related classes remain challenging. This analysis validates the robustness and generality of our clustering framework across diverse visual domains.



(a) CIFAR-10 (b) CIFAR-100 (c) STL-10

(d) ImageNet-10 (e) ImageNet-Dogs

Figure 2. The confusion matrices of ICIRD on five datasets, where the x-axes are the ground-truth labels and the y-axes are the predicted labels. The clearer the diagonal structure in the confusion matrix, the better it is represented.

5.4.2. Case Studies

In this section, several representative cases from the clustering results are analyzed as illustrated in Figure 3. Specifically, the cases are divided into three categories: (i) cases correctly assigned to their corresponding categories (CAC); (ii) cases incorrectly assigned to other categories (IAO); and (iii) cases incorrectly assigned to the current category (IAC).

Taking the Stanford-Dogs dataset as an example, the model performs well on high-accuracy categories such as the African hunting dog, where distinctive coat patterns and ear shapes clearly differentiate them from other breeds. However, under uncommon viewing angles or in juvenile images, these samples are often misclassified into visually similar categories such as Whippet or Greyhound. For low-accuracy categories like Border Terrier, the model frequently confuses them with breeds such as Norfolk Terrier, Cairn Terrier, and Miniature Schnauzer, all sharing similar facial textures and drooping ears. In the CIFAR-100 dataset, the limited image resolution causes the model to rely primarily on color, texture, and background cues. Consequently, even well-recognized categories such as Flowers are sometimes confused with Fruits and Vegetables or Trees due to color overlap, whereas objects with clean backgrounds and distinct silhouettes tend to be classified more accurately. For the ImageNet-Dogs dataset, misclassifications mainly occur in images containing multiple objects, occlusions, or ambiguous perspectives—such as rear views or extreme close-ups—where discriminative features become less perceptible.

(a) CIFAR-100: right-flowers, middle-aquatic mammals, left-vehicles2

(b) ImageNet-Dogs: right-blenheim spaniel, middle-giant schnauzer, left-Welsh springer spaniel

(c) StanFord-Dogs: right-Border terrier, middle-Walker Hound, left-Aferican Hunting Dog

Figure 3. The case studies of ICIRD on CIFAR-100, ImageNet-Dogs and Stanford-Dogs datasets. For each dataset, three classes are visualized, and their samples are highlighted with red, green, and blue bounding boxes corresponding to the left, middle, and right groups in each subfigure. Each row then shows a different case criterion (CAC, IAO, IAC), as indicated on the left side of the subfigure.

Overall, ICIRD can accurately distinguish subtle differences across fine-grained datasets, while its misclassifications are mainly concentrated in visually ambiguous or difficult cases.

5.5. Ablation Studies

5.5.1. Effectiveness Analysis

In this section, the proposed framework's losses are evaluated through ablation experiments. Specifically, the following ablation settings are designed to clarify the contribution of each constraint: (i) without the CRA loss; (ii) without the DDS loss; (iii) without the MIDR loss; (iv) without the CIDC loss; and (v) the complete ICIRD model. To comprehensively validate the effectiveness of each constraint, experiments are conducted on CIFAR-10, CIFAR-100, STL-10, and ImageNet-Dogs, which together cover datasets of varying resolution, class number, and granularity. The results are shown in Figure 4. The vertical axis of all subplots is fixed to the range [0, 1] for consistent cross-dataset comparison. It can be observed that removing any loss function leads to a degradation of performance across all datasets, confirming the overall effectiveness of our framework. The importance of each loss is analyzed individually as follows:

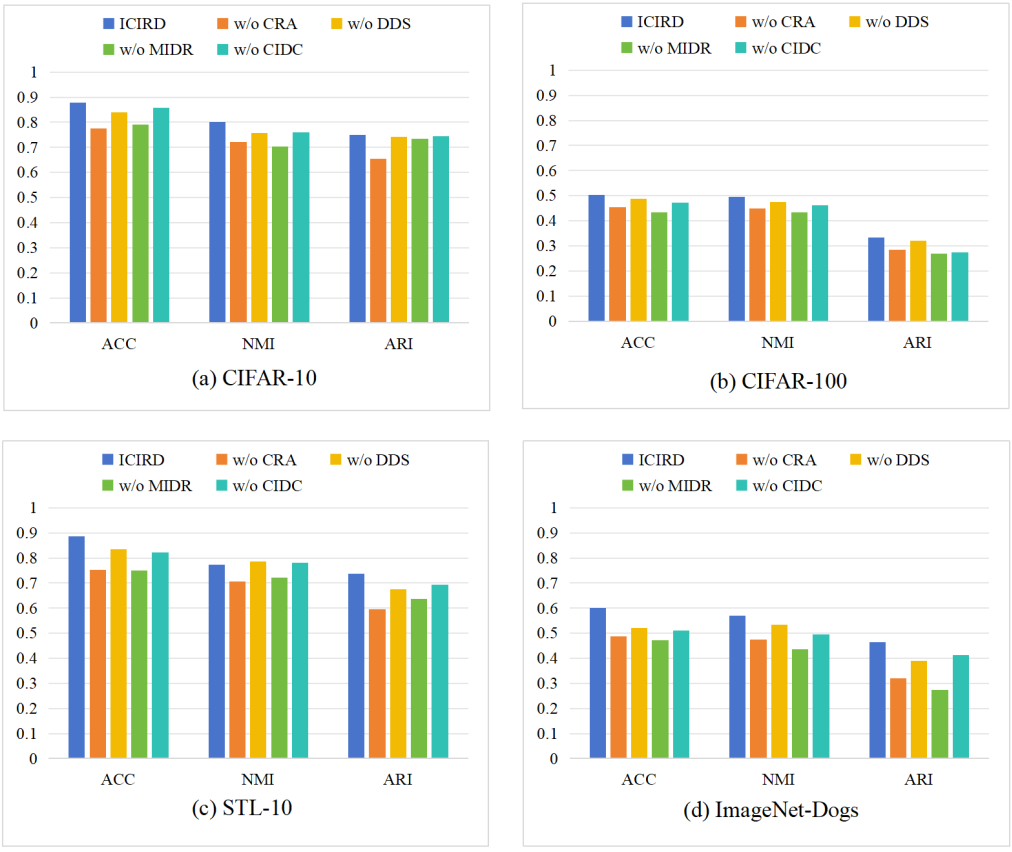


Figure 4. Effectiveness ablation results of the loss functions on four datasets with ACC, NMI, ARI.

(i) When the CRA loss is removed, the performance significantly drops on all datasets, though the model remains stable. This indicates that basic representation learning is necessary for stability, while the remaining three losses still preserve partial discriminative capacity. (ii) When the DDS loss is removed, the decrease is relatively small, implying that it acts as an auxiliary regularization. This aligns with its role in providing stable cluster probabilities for MIDR and CIDC. Even without DDS, the CRA loss can still transfer stability from representation learning to the clustering head. (iii) When the MIDR loss is

removed, performance drops substantially across all datasets, showing that this loss has the greatest impact on clustering quality. In this case, only CIDC and DDS impose weak clustering guidance, which is insufficient to maintain strong discriminative structure, yet they still preserve stable performance. (iv) When the CIDC loss is removed, the decline is minor, likely due to partial functional overlap between CIDC and CRA.

Cross-dataset variations are then analyzed to further interpret the ablation results: (i) For the two major losses, CRA and MIDR, we observe that in Figure 4a,c, "w/o CRA" performs worse than "w/o MIDR," while in Figure 4b,d, the opposite holds. This suggests that MIDR plays a more significant role in fine-grained datasets. (ii) For the two auxiliary losses, DDS and CIDC, Figure 4b,d, show that "w/o DDS" outperforms "w/o CIDC," whereas Figure 4a shows the reverse. This indicates that CIDC benefits fine-grained datasets, whereas DDS is more effective for low-resolution images.

Overall, all losses in the ICIRD framework contribute positively to the final performance. The four modules complement one another, ensuring robust clustering across datasets with different resolutions and granularity levels.

5.5.2. Convergence Analysis

Furthermore, the convergence curves of ACC across four ablation settings on four datasets were investigated. As illustrated in Figure 5, the ACC values were recorded every 50 epochs to construct line charts. The results demonstrate that ICIRD is consistently outperformed by no other setting throughout the entire training process, confirming the effectiveness and stability of the proposed framework. In addition, it was observed that the CIDC and DDS losses significantly accelerate model convergence, and on certain datasets, their impact on convergence speed and stability even surpasses that of the MIDR loss. These findings further indicate that each loss component within the framework is crucial for enhancing the overall performance of the model.

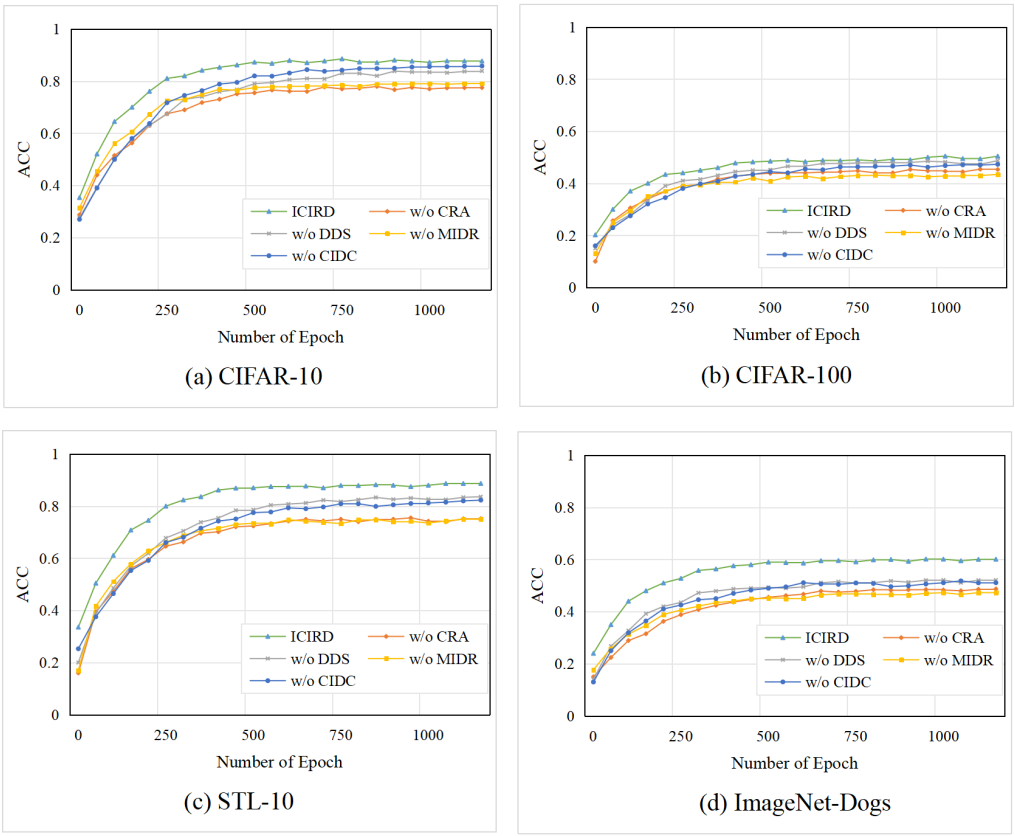


Figure 5. Convergence ablation results of the loss functions on four datasets with ACC.

5.6. Discussion of Cross-View IDR

In this section, the feasibility of cross-view IDR (CIDR) is explored and evaluated. The cross-view IDR loss is designed based on the discriminative outputs of two views, where P^a and P^b are multiplied to obtain D^{ab}, and $\mathcal{L}_{\text{IDR}}^{ab}$ is subsequently computed according to Equation (9). By further incorporating the MIDR loss introduced in Section 4.3, we obtain a unified multi-view and cross-view IDR objective, termed MCIDR, which is formulated as follows:

$$\mathcal{L}_{\text{MCIDR}} = \mathcal{L}_{\text{IDR}}^{a} + \mathcal{L}_{\text{IDR}}^{b} + \frac{1}{2}(\mathcal{L}_{\text{IDR}}^{ab} + \mathcal{L}_{\text{IDR}}^{ba}). \tag{18}$$

where the CIDR loss is $\frac{1}{2}(\mathcal{L}_{\text{IDR}}^{ab} + \mathcal{L}_{\text{IDR}}^{ba})$, and the symmetric cross-view terms are numerically averaged. Ideally, the set of negative samples is expected to be expanded for redundancy reduction through the cross-view mechanism, whereby inter-view associations are established and model performance is further enhanced.

However, the MIDR loss was originally designed as an intra-view constraint because aligning cluster labels across different views in deep learning is inherently challenging. Misaligned cluster probabilities are difficult to compare directly, especially during the early training stages. Nevertheless, the following can be reasonably inferred: (i) In the mid-to-late training stages, the cluster probabilities tend to be stabilized, particularly for samples located near cluster centers. (ii) The CIDC loss is implicitly used to promote distributional alignment across views through consistency constraints. (iii) Noise interference is suppressed by the DDS loss through its sharpening effect. Therefore, the adoption of the CIDR loss can be considered feasible under certain conditions.

In summary, two major factors influence CIDR: (i) the timing of its introduction during training and (ii) the confidence of cluster probabilities. The former is intuitive—CIDR should be introduced in the mid-to-late training stage—and can be controlled via the View Consistency (VC) ratio, while the latter is regulated by the confidence threshold ζ. The view consistency ratio is first defined as follows:

$$\text{VC} = \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}\big[\arg\max_k p_i^a(k) = \arg\max_k p_i^b(k)\big]. \tag{19}$$

Here, p_i^a and p_i^b denote the predicted probability distributions of sample i under views a and b, respectively. This metric represents the proportion of consistent predictions within a batch and reflects the stability of cluster probabilities. The CIDR loss is introduced only when this ratio exceeds the threshold.

Next, the confidence threshold ζ is defined. Taking a single view as an example, for the filtered probability matrix $P' \in \mathbb{R}^{B' \times K}$, $B' < B$, each cluster probability distribution p_i in P' is normalized using a power function: $\tilde{p}_i = \frac{p_i^\alpha}{\sum_j^{B'} p_j^\alpha}$, to further mitigate adverse effects. Finally, the processed matrices $\tilde{P}^a$ and $\tilde{P}^b$ from the two views are used to compute the CIDR loss following the method described in Section 4.3 and Equation (18).

Next, we conduct experiments to validate the proposed CIDR loss and alignment strategy. Specifically, we compare three settings: ICIRD with MIDR (denoted as $ICIRD_{MIDR}$), ICIRD with MCIDR (denoted as $ICIRD_{MCIDR}$), and ICIRD with MCIDR but without the alignment strategy (denoted as $ICIRD_{MCIDR-}$). In the experiments, VC was set to 0.7 and ζ to 0.8 (α to 2), and then conducted on the CIFAR-10, STL-10, and Stanford-Dogs datasets. The clustering results were reported in Table 4. As shown in the table, $ICIRD_{MCIDR-}$ performs significantly worse than $ICIRD_{MCIDR}$, demonstrating that performing cross-view IDR without the alignment strategy adversely affects model performance. On the other hand, only marginal performance improvements were yielded by the inclusion of the CIDR loss in $ICIRD_{MCIDR}$ on the CIFAR-10 and STL-10 datasets. On

STL-10, $ICIRD_{MCIDR}$ was observed to perform slightly worse than $ICIRD_{MIDR}$ in terms of NMI and ARI. However, on the Stanford-Dogs dataset, $ICIRD_{MCIDR}$ was found to significantly outperform $ICIRD_{MIDR}$. This indicates that cross-view redundancy reduction can be beneficial under stable cluster probability conditions, and this scheme is beneficial for discrimination on fine-grained datasets. How to obtain more stable cluster probabilities remains worth investigating.

Table 4. Performance comparison with $ICIRD_{MIDR}$, $ICIRD_{MCIDR}$, and $ICIRD_{MCIDR-}$ on three datasets.

Datasets	**CIFAR-10**			**STL-10**			**Stanford-Dogs**		
Meric	**ACC**	**NMI**	**ARI**	**ACC**	**NMI**	**ARI**	**ACC**	**NMI**	**ARI**
$ICIRD_{MIDR}$	0.877	0.801	0.750	0.887	**0.774**	**0.737**	0.176	0.366	0.101
$ICIRD_{MCIDR}$	**0.886**	**0.815**	**0.753**	**0.894**	0.765	0.721	**0.194**	**0.391**	**0.109**
$ICIRD_{MCIDR-}$	0.852	0.768	0.725	0.864	0.747	0.708	0.139	0.317	0.083

The highest values are tagged in **bold**.

5.7. Analysis of Hyper-Parameters

5.7.1. Analysis of Representation Dimension

The choice of representation dimensionality directly influences the optimization stability and information regularization of each module. To analyze this effect, we evaluate MCIDR under varying representation dimensionalities in $\{128, 256, 512, 768, 1024\}$ across three datasets (CIFAR-10, STL-10, and ImageNet-Dogs), with the results shown in Figure 6. As observed, larger representation dimensions improve MCIDR's performance on high-resolution or fine-grained datasets. However, excessively large dimensions cause performance degradation. This degradation can be attributed to two factors: (i) Higher dimensionality introduces greater spatial redundancy, making redundancy reduction more difficult. (ii) Excessive dimensionality may exceed the representational capacity of ResNet-34.

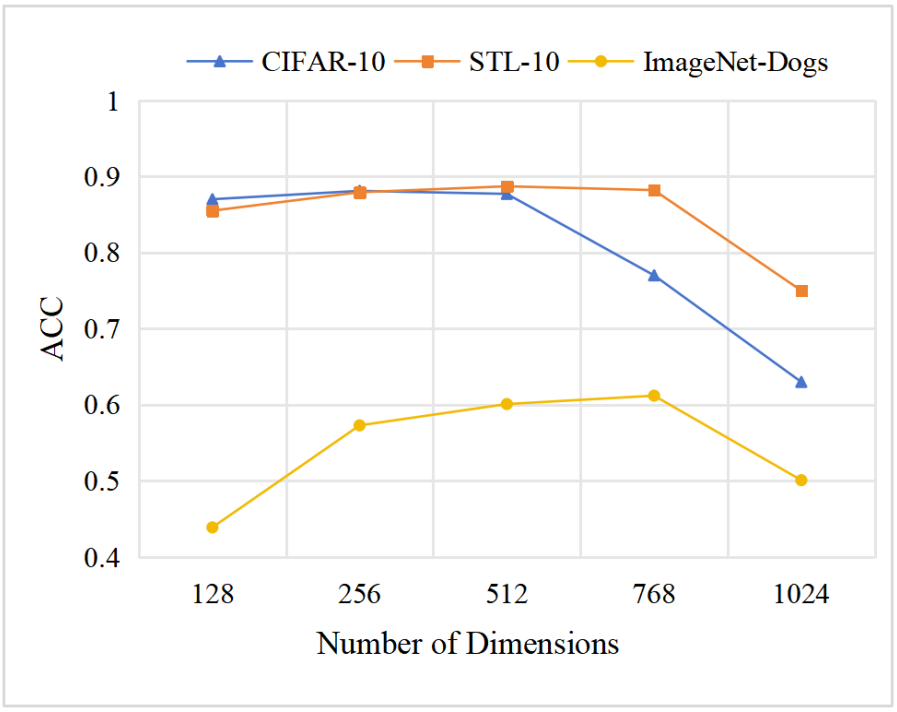


Figure 6. Influence of different representation dimension on three datasets.

5.7.2. Analysis of Cluster Number

In previous experiments, the number of clusters K was predefined. However, in real-world scenarios, such prior knowledge is often unavailable. To evaluate the sensitivity of MCIDR to semantic granularity and the robustness of its distributional constraints, we varied the value of K with the value in $\{5, 10, 20, 30, 40, 50\}$ on the CIFAR-10 and CIFAR-100 datasets and reported the corresponding NMI results, as shown in Figure 7. The results indicate that when K deviates from the true number of categories in CIFAR-10, the NMI drops sharply, suggesting that MCIDR tends to learn cluster structures consistent with the intrinsic category structure of the data. In contrast, for CIFAR-100, the NMI steadily increases as K grows. This aligns with the dataset's hierarchical organization, where 100 subclasses are grouped under 20 superclasses. This experiment demonstrates that MCIDR effectively captures underlying semantic structures.

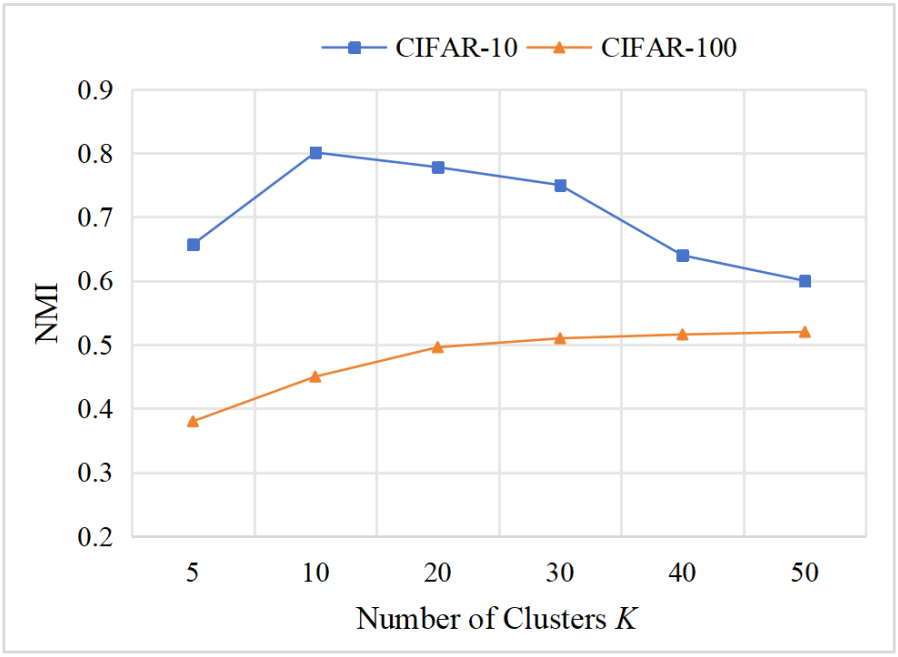
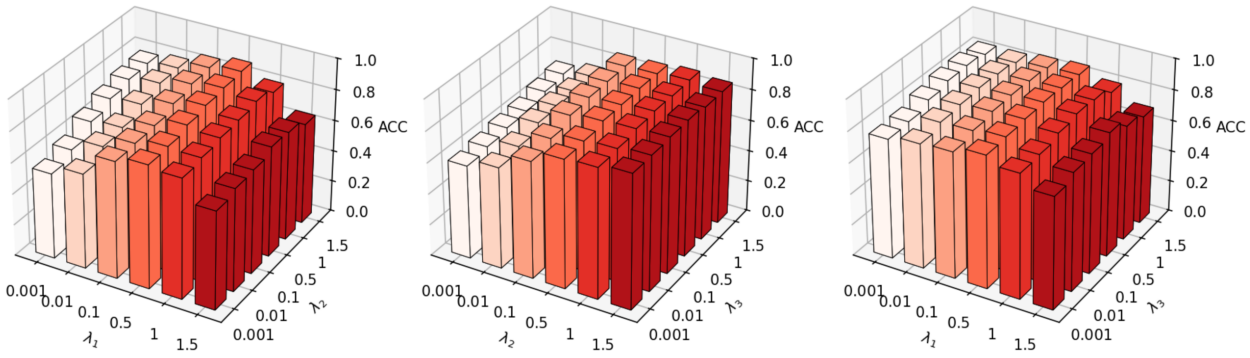


Figure 7. Influence of predefined cluster number on CIFAR-10/100 datasets.

5.7.3. Analysis of Balance Parameters

To further investigate the relationships among the three balancing parameters λ_1, λ_2, and λ_3 in the total loss of MCIDR and their influence on experimental results, we construct the visualization shown in Figure 8 based on the CIFAR-10 dataset. Each parameter is sequentially set to the values in $\{0.001, 0.01, 0.1, 0.5, 1, 1.5\}$. (i) In Figure 8a, an increase in λ_1 gradually improves model performance; however, when $\lambda_1 > 1$, the performance begins to degrade. This may be attributed to unstable cluster probabilities during the early stages of joint training, where an excessively large λ_1 can impair model stability. (ii) In contrast, a higher λ_2 generally enhances model performance except when $\lambda_1 > 1$. This indicates that the CIDC loss positively influences MIDR and effectively improves overall performance. Furthermore, it suggests that the MIDR loss has a more significant impact on the framework than the CIDC loss. (iii) Figure 8b further reveals a clear positive correlation between CIDC and DDS: when one of these parameters remains fixed, model performance increases with the other until convergence plateaus. (iv) According to Figure 8c, the relationship between λ_1 and λ_3 is similar to that between λ_1 and λ_2, suggesting implicit synergy between the two losses. Moreover, the framework is more sensitive to an excessively large λ_1 than to a

smaller one. In summary, the loss terms in MCIDR exhibit complementary and mutually reinforcing effects.



(a) (b) (c)

Figure 8. (**a**–**c**) Three-dimensional bar chart analysis of balance parameters $\lambda_1, \lambda_2, \lambda_3$ on CIFAR-10 datasets.

6. Illustrative Applications and Scalability of ICIRD

In this section, we first discuss the scalability of ICIRD across different data modalities and then discuss its practical application in various downstream tasks.

6.1. Extending ICIRD to Non-Image Modalities

In this study, we primarily conduct experiments on image datasets and adopt ResNet as the backbone feature extractor. When extending ICIRD to other modalities such as textual data or structured tabular data, the main challenges arise from the selection of an appropriate backbone architecture and the design of data augmentation (i.e., multi-view construction) strategies. It is worth noting that, aside from the contrastive learning component, which depends on modality-specific augmentations, the other three information-theoretic modules of ICIRD (DDS, MIDR, and CIDC) remain structurally unchanged across different data types. Their sensitivity to modality is minimal, and thus only minor adaptations based on input feature formats and view-construction strategies are required.

For textual datasets, widely used Transformer-based encoders such as BERT [56], DeBERTa-v3 [57], or more recent variants such as the model studied by Guedes and da Silva [58] can be employed. Multi-view textual augmentations are typically constructed through dropout masking, span masking, synonym substitution, or sentence-level perturbation; these augmentation strategies have been demonstrated to be effective in unsupervised text clustering, as shown in SimCSE [59] and CoCLR-text [60].

For structured behavioral or attribute-based tabular data, Transformer architectures specifically designed for tabular modeling, such as TabTransformer [61] and FT-Transformer [62], can be adopted. In unsupervised settings, multi-view construction for tabular data often relies on feature dropout, feature-subset sampling, or noise injection, similar to the masking-based pseudo-view strategy introduced in TabularSSL [63]. ICIRD can operate directly on such backbone–augmentation pipelines and produce stable and consistent cluster distributions across samples.

6.2. Practical Application of ICIRD in Downstream Tasks

To clarify how ICIRD functions in practical scenarios, we next discuss its behavior across representative downstream tasks and the adaptations required in each setting. In image recognition tasks, deep clustering methods such as ICIRD primarily contribute in

two core directions: unsupervised clustering and novel category discovery. For unsupervised clustering, one typically assumes a reasonable estimate of the true number of categories in order to set the dimensionality of the clustering head. The model is then trained using the standard ICIRD configuration to learn cluster probability distributions, which subsequently serve as pseudo-labels for downstream image recognition. In novel category discovery scenarios, a single supervised warm-up on the labeled subset is first performed—an established practice in prior work, including UNO [64] and GCD [65]. After this warm-up stage, ICIRD's information-theoretic optimization is applied to the unlabeled subset, enabling the model to construct a stable cluster structure under a fixed number of unknown categories K and to produce reliable predictions for the novel classes.

In other domains, such as medical diagnostics and customer segmentation, ICIRD can be applied in essentially the same manner as in image recognition. The four information-theoretic modules of ICIRD typically require no structural modification. Only "outer-layer" adaptations are needed—such as replacing the backbone, redesigning the view-construction strategy, or adjusting loss weights—to accommodate the characteristics of each modality.

When a task simultaneously requires high-quality feature embeddings and reliable clustering outputs, certain modifications to ICIRD become necessary. One feasible approach is to shift the MIDR component from the cluster-distribution space to the representation space, thereby computing inter-cluster mutual-information redundancy directly on feature embeddings to enhance their linear separability. In addition, the DDS module can be strengthened—for instance, by introducing a marginal-entropy term—to compensate for the absence of MIDR at the probability-distribution level, forming a more complete discriminative mutual-information objective that stabilizes the clustering assignments. Under such a configuration, however, ICIRD effectively transitions into a representation-oriented deep clustering framework, and thus, the entire model requires comprehensive re-evaluation and re-adjustment to ensure consistency and effectiveness.

In scenarios where cross-modal data inevitably arises, ICIRD still does not require structural modifications to its four information-theoretic modules. Instead, the additional modality can simply be incorporated as an extra form of data augmentation and integrated into the multi-view framework. This design is consistent with recent advances in cross-modal contrastive learning, such as the CLIP-style [66] modality alignment used in MAC [67], which provides a practical reference for extending ICIRD to multimodal settings.

7. Conclusions

This paper proposes ICIRD, a deep clustering framework that formulates clustering as a process of optimizing the informational structure of cluster probability distributions. ICIRD jointly regularizes feature representations and cluster assignments through information preservation, redundancy reduction, and cross-view consistency, enabling the model to learn sharp, independent, and invariant cluster distributions. The framework establishes a controllable balance between information retention and compression, ensuring stable and bounded optimization. Extensive experiments on multiple benchmarks validate that ICIRD achieves superior clustering accuracy, robustness compared with existing methods. In addition, ICIRD investigates the role of view alignment in improving clustering consistency and robustness across augmented views. Overall, ICIRD provides a unified and principled paradigm for information-theoretic deep clustering, offering both theoretical insight and practical guidance for future research in representation learning and unsupervised semantic modeling.

Author Contributions: Conceptualization, A.Z.; methodology, A.Z.; software, A.Z.; validation, R.M.X.W.; writing, A.Z.; supervision, Y.H.; investigation, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: The work is supported by the National Natural Science Foundation of China (Grant Nos. 12473105, 12473106), Projects of Science and Technology Cooperation and Exchange of Shanxi Province (Grant Nos. 202204041101037, 202204041101033), The central government guides local funds for science and technology development (YDZJSX2024D049), the science research grant from the China Manned Space Project with No. CMS-CSST-2021-B03.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in CIFAR-10 (https://www.cs.toronto.edu/~kriz/cifar.html, (accessed on 3 November 2025)), STL-10 (https://cs.stanford.edu/~acoates/stl10/, (accessed on 3 November 2025)), ImageNet (https://image-net.org/download-images.php, (accessed on 3 November 2025)), Stanford-Dogs (http://vision.stanford.edu/aditya86/ImageNetDogs/main.html, (accessed on 3 November 2025)).

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

Abbreviations

The following abbreviations are used in this manuscript:

ICIRD	Information-principled Deep Clustering for Invariant, Redundancy-reduced and Discriminative Cluster Distributions
DDS	Discriminative Distribution Sharpness
MIDR	Multiview Inter-cluster Distribution Redundancy Reduction
CIDC	Cross-view Instance Distribution Consistency
CRA	Contrastive Representation Anchoring
MI	Mutual Information
DPI	Data Processing Inequality
IDR	Inter-cluster Distribution Redundancy
KL	Kullback–Leibler Divergence

References

1. Ren, Y.; Pu, J.; Yang, Z.; Xu, J.; Li, G.; Pu, X.; Yu, P.S.; He, L. Deep Clustering: A Comprehensive Survey. *arXiv* **2022**, arXiv:2212.07473. [CrossRef]
2. 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]
3. Aljalbout, E.; Golkov, V.; Siddiqui, Y.; Strobel, M.; Cremers, D. Clustering with Deep Learning: Taxonomy and New Methods. *arXiv* **2018**, arXiv:1801.07648. [CrossRef]
4. Ohl, L.; Mattei, P.A.; Precioso, F. A tutorial on discriminative clustering and mutual information. *arXiv* **2025**, arXiv:2505.04484. [CrossRef]
5. Zhou, S.; Xu, H.; Zheng, Z.; Chen, J.; Li, Z.; Bu, J.; Wu, J.; Wang, X.; Zhu, W.; Ester, M. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions. *arXiv* **2024**, arXiv:2206.07579. [CrossRef]
6. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. In Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 20–22 June 2016.
7. 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.
8. Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
9. Ji, X.; Henriques, J.F.; Vedaldi, A. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9865–9874. [CrossRef]
10. Parulekar, A.; Collins, L.; Shanmugam, K.; Mokhtari, A.; Shakkottai, S. InfoNCE Loss Provably Learns Cluster-Preserving Representations. In Proceedings of the Thirty Sixth Conference on Learning Theory, Bangalore, India, 12–15 July 2023; Proceedings of Machine Learning Research; Volume 195, pp. 1914–1961.

11. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
12. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020.
13. Li, Y.; Wang, L.; Wang, Y.; Liu, T.; Zhang, L. Contrastive Clustering. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtually, 2–9 February 2021.
14. Liu, Y.; Tu, W.; Zhou, S.; Liu, X.; Song, L.; Yang, X.; Zhu, E. Deep Graph Clustering via Dual Correlation Reduction. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Online, 22 February–1 March 2022.
15. Huang, D.; Chen, D.H.; Chen, X.; Wang, C.D.; Lai, J.H. Deepclue: Enhanced deep clustering via multi-layer ensembles in neural networks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2024**, *8*, 1582–1594. [CrossRef]
16. Huang, D.; Deng, X.; Chen, D.H.; Wen, Z.; Sun, W.; Wang, C.D.; Lai, J.H. Deep clustering with hybrid-grained contrastive and discriminative learning. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 9472–9483. [CrossRef]
17. Zhao, Y.; Bai, L. Contrastive clustering with a graph consistency constraint. *Pattern Recognit.* **2024**, *146*, 110032. [CrossRef]
18. 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.
19. Nassar, I.; Karlinsky, L.; Feris, R.; Tay, Y.; Padhy, S.; Noy, A.; Zhang, L.; Elhoseiny, M.; Tsai, Y.H.; Nevatia, R. ProtoCon: Pseudo-Label Refinement via Online Clustering and Prototypical Consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
20. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861. [CrossRef]
21. 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, Stockholm, Sweden, 10–15 July 2018.
22. 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. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
23. Hu, W.; Miyato, T.; Tokui, S.; Matsumoto, E.; Sugiyama, M. Learning Discrete Representations via Information Maximizing Self-Augmented Training. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 1558–1567.
24. 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 October–2 November 2019; pp. 8150–8159.
25. Zhang, H.; Liu, S.; Wang, C.; Yu, X. Deep Descriptive Clustering: Unifying Representation, Interpretability, and Discriminability. In Proceedings of the International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 19–27 August 2021.
26. Zhang, L.; Li, H.; Chen, Q.; Zhang, W. Mutual Information-Driven Multi-View Clustering. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023.
27. Yan, X.; Jin, Z.; Han, F.; Ye, Y. Differentiable information bottleneck for deterministic multi-view clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 27435–27444.
28. Lou, Z.; Zhang, K.; Wu, Y.; Hu, S. Super Deep Contrastive Information Bottleneck for Multi-modal Clustering. In Proceedings of the Forty-Second International Conference on Machine Learning, Vancouver, BC, Canada, 13–19 July 2025.
29. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020.
30. Li, J.; Zhou, P.; Xiong, C.; Hoi, S.C.H. Prototypical Contrastive Learning of Unsupervised Representations. In Proceedings of the Ninth International Conference on Learning Representations: ICLR 2021, Vienna, Austria, 4–8 May 2021.
31. Chuang, C.; Robinson, J.; Lin, Y.; Torralba, A.; Jegelka, S. Debiased Contrastive Learning. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Online, 6–12 December 2020.
32. Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; Van Gool, L. Learning to Classify Images without Labels. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
33. Niu, C.; Shan, H.; Wang, G. SPICE: Semantic Pseudo-Labeling for Image Clustering. *IEEE Trans. Image Process.* **2022**, *31*, 7172–7185. [CrossRef]
34. Zhong, H.; Wu, J.; Chen, C.; Huang, J.; Deng, M.; Nie, L.; Lin, Z.; Hua, X.-S. Graph Contrastive Clustering. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021.
35. Deng, X.; Huang, D.; Chen, D.H.; Wang, C.D.; Lai, J.H. Strongly augmented contrastive clustering. *Pattern Recognit.* **2023**, *139*, 109470. [CrossRef]

36. Xu, Y.; Huang, D.; Wang, C.D.; Lai, J.H. Deep image clustering with contrastive learning and multi-scale graph convolutional networks. *Pattern Recognit.* **2024**, *146*, 110065. [CrossRef]
37. Znalezniak, M.; Rola, P.; Kaszuba, P.; Tabor, J.; Śmieja, M. Contrastive hierarchical clustering. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Turin, Italy, 18–22 September 2023; pp. 627–643.
38. Gray, R.M. *Entropy and Information Theory*; Springer Science & Business Media: New York, NY, USA, 2011.
39. Tishby, N.; Pereira, F.C.; Bialek, W. The Information Bottleneck Method. *arXiv* **2000**, arXiv:physics/0004057. [PubMed]
40. Grandvalet, Y.; Bengio, Y. Semi-supervised Learning by Entropy Minimization. In Proceedings of the 18th International Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 13–18 December 2004.
41. Krizhevsky, A.; Hinton, G. Learning Multiple Layers of Features from Tiny Images. 2009. Available online: https://bibbase.org/network/publication/krizhevsky-hinton-learningmultiplelayersoffeaturesfromtinyimages-2009 (accessed on 3 November 2025)
42. Coates, A.; Ng, A.; Lee, H. An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; JMLR Workshop and Conference Proceedings; pp. 215–223.
43. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
44. Strehl, A.; Ghosh, J. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.
45. Hubert, L.; Arabie, P. Comparing Partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
46. Wang, X.; Qi, G.J. Contrastive learning with stronger augmentations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5549–5560. [CrossRef]
47. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
48. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
49. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3733–3742.
50. Huang, J.; Gong, S.; Zhu, X. Deep semantic clustering by partition confidence maximisation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8849–8858.
51. Zhong, H.; Chen, C.; Jin, Z.; Hua, X.S. Deep robust clustering by contrastive learning. *arXiv* **2020**, arXiv:2008.03030. [CrossRef]
52. Tao, Y.; Takagi, K.; Nakata, K. Clustering-friendly representation learning via instance discrimination and feature decorrelation. *arXiv* **2021**, arXiv:2106.00131. [CrossRef]
53. Dang, Z.; Deng, C.; Yang, X.; Huang, H. Doubly contrastive deep clustering. *arXiv* **2021**, arXiv:2103.05484. [CrossRef]
54. Zhang, F.; Li, L.; Hua, Q.; Dong, C.R.; Lim, B.H. Improved deep clustering model based on semantic consistency for image clustering. *Knowl.-Based Syst.* **2022**, *253*, 109507. [CrossRef]
55. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Fei-Fei, L. Novel Dataset for Fine-Grained Image Categorization. In Proceedings of the First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011.
56. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL, Minneapolis, MN, USA, 2 June–7 June 2019.
57. He, P.; Liu, X.; Gao, J.; Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In Proceedings of the ICLR, Virtual Event, Austria, 3–7 May 2021.
58. Guedes, G.B.; da Silva, A.A.E. Classification and Clustering of Sentence-Level Embeddings of Scientific Articles Generated by Contrastive Learning. *arXiv* **2024**, arXiv:2404.00224. [CrossRef]
59. Gao, T.; Yao, X.; Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the EMNLP, Virtual Event/Punta Cana, Dominican Republic, 7–11 November 2021.
60. Yan, Y.; Zhang, Y.; Lin, X.; Li, X. CoCLR-Text: Contrastive Cross-View Learning for Text Clustering. In Proceedings of the Findings of ACL, Dublin, Ireland, 22–27 May 2022.
61. Huang, X.; Khetan, A.; Cvitkovic, M.; Bansal, V. TabTransformer: Tabular Data Modeling Using Contextual Embeddings. In Proceedings of the NeurIPS, Virtual, 6–12 December 2020.
62. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting Deep Learning Models for Tabular Data. In Proceedings of the NeurIPS, Virtual, 6–14 December 2021.
63. Ucar, T.; Artelt, A.; Hammer, B. Self-Supervised Learning for Tabular Data via Masked Feature Reconstruction. In Proceedings of the ICML, Honolulu, HI, USA, 23–29 July 2023.
64. Fini, E.; Lathuilière, S.; Sangineto, E.; Zhong, Z.; Nabi, M.; Sebe, N.; Ricci, E. A Unified Objective for Novel Class Discovery. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021.

65. Vaze, S.; De Melo, N.C.; Bojanowski, P.; Joulin, A.; Douze, M. Generalized Category Discovery. In Proceedings of the NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.
66. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the ICML, Virtual, 18–24 July 2021.
67. Qiu, L.; Zhang, Q.; Chen, X.; Cai, S. Multi-Level Cross-Modal Alignment for Image Clustering. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vancouver, BC, Canada, 26–27 February 2024; Volume 38, pp. 14695–14703. [CrossRef]