DISCOVERING OR PRESCRIBING? REFLEXIVE STRUCTURAL SUPERVISION FOR CONTRASTIVE CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning seeks to derive learning signals from unlabeled data itself, allowing meaningful representations to emerge without externally provided semantic supervision. For clustering, however, this principle leads to a fundamental tension. The organization of the data is precisely what the model is expected to discover, yet many contrastive clustering methods construct their supervision under an assumed clustering structure before that organization has been revealed. This tension is particularly relevant in unlabeled data, where multiple levels of organization may coexist and the appropriate granularity is generally unknown a priori. Committing representation learning to a single prescribed view of the data may therefore favor one possible organization while leaving other meaningful structural relationships under-explored. We develop Discovering Intrinsic Structure via Contrastive Clustering (DISC) as a reflexive learning framework, in which structural supervision is progressively inferred from the current representation space and fed back into representation learning. To avoid overlooking complementary structural information under a single prescribed view, DISC derives supervisory signals from discovered multi-granularity structural relationships. Furthermore, to reduce discrepancies between assignment results across granularities, DISC imposes cross-granularity consistency constraints between adjacent levels. Comparisons with recent alternatives demonstrate that DISC achieves strong representation quality and clustering performance without relying on a predefined clustering structure, while the ablation results consistently show the benefit of discovering and exploiting multi-granularity structure throughout learning. These findings suggest a shift in contrastive clustering from learning under a prescribed structure toward learning supervision through progressive structure discovery.

1 INTRODUCTION

Deep clustering aims to discover the underlying structure among samples from unlabeled data and assign similar samples to the same cluster (Xie et al., 2016; Caron et al., 2018). In recent years, contrastive learning has been increasingly adopted in deep clustering to transform clustering structures into self-supervised signals, thereby jointly optimizing feature representations and cluster assignments (Caron et al., 2020; Li et al., 2021b). These methods construct supervision under a predefined clustering structure (Zhu et al., 2026; Li, 2026). Such a structure typically takes the form of a specified number of clusters (Ji et al., 2019), clustering hierarchies (Li et al., 2021a; Guo et al., 2022), or relationships among samples (Zhong et al., 2021; Yang et al., 2023). Such a structure determines how samples are assigned to clusters or related to one another, which can then be transformed into pseudo-labels (Niu et al., 2022) or positive/negative sample relationships (Liu et al., 2023) to provide supervision for representation learning. Concretely, some methods use a clustering head whose output space is determined by the predefined number of clusters, thereby learning cluster assignments at a prescribed granularity (Do et al., 2021); others directly exploit predefined sample relationships or hierarchical structures to construct contrastive supervision.

However, the underlying structure of the data is often unknown and complex, making it difficult to determine an appropriate clustering structure before learning. More importantly, when the predefined structure is inconsistent with the actual organization of the data, it can lead to incorrect sample

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

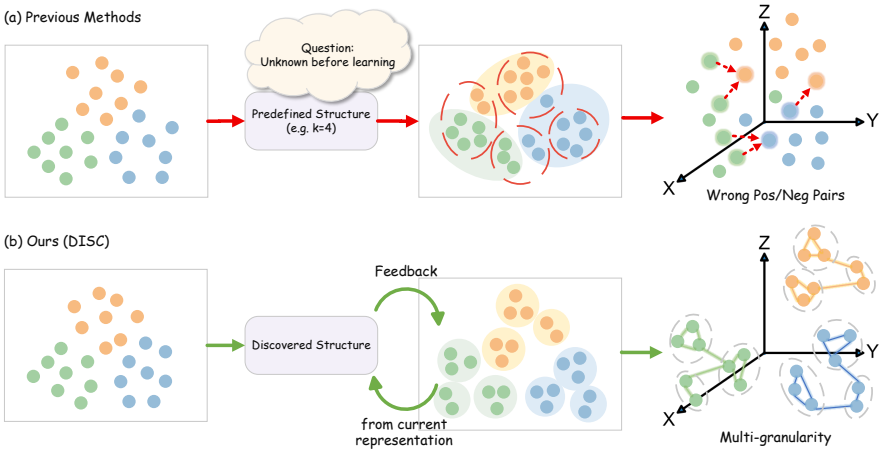

Figure 1: Comparison of structure-guided contrastive clustering and DISC. (a) Previous methods rely on a predefined clustering structure to guide learning. (b) DISC discovers multi-granularity structure from feature representations and iteratively refines the representations.

relationships and supervisory signals, thereby constraining the learned representation. For example, as illustrated in Figure 1(a), an inappropriate predefined structure can result in incorrect cluster assignments, which further induce misleading positive/negative sample pairs and distort the learned representation. This creates a fundamental tension: the organizational structure of the data should be discovered through clustering, yet it is assumed before learning. Is contrastive clustering truly discovering how the data are organized, or merely adapting the representation space to a predefined structure?

Based on this observation, we propose DISC (Discovering Intrinsic Structure via Contrastive Clustering), a data-driven contrastive clustering framework. DISC introduces a teacher-student architecture to iteratively couple structure discovery with representation learning, with the teacher and student responsible for structure discovery and representation optimization, respectively. The teacher discovers multi-granularity structures from the current feature representations, and the student uses the discovered structures to optimize the representations. As illustrated in Figure 1(b), the updated representations are then fed back for subsequent structure discovery, allowing the structural information to evolve with the representation space. Through this iterative process, DISC allows supervisory signals to be continuously derived from the evolving feature representations rather than from a predefined clustering structure. In summary, our main contributions are as follows:

- A fundamental mismatch between structure discovery and supervision construction is revealed in existing contrastive clustering: self-supervised signals are constructed from predefined clustering structures, although these structures are themselves the objects to be discovered.
- DISC is proposed as a contrastive clustering framework that iteratively couples structure discovery with representation optimization, continuously deriving multi-granularity supervisory signals from evolving feature representations.
- DISC establishes a new paradigm for contrastive clustering, shifting the role of clustering structure from a structural prior to an object of discovery during learning, thereby providing a new perspective on self-supervised learning.

2 Related Work

Deep Clustering is a class of unsupervised learning methods that simultaneously optimize representation learning and clustering objectives via deep neural networks, aiming to discover intrinsic group structures in unlabeled data (Zhou et al., 2024). Existing deep clustering methods can be broadly categorized into two paradigms: representation learning based methods and clustering head based

methods (Li et al., 2026). The former focuses on learning discriminative feature representations and performs clustering in the learned representation space using algorithms such as K-means (Yang et al., 2017; Chen et al., 2024), while the latter introduces a trainable clustering head after the feature extraction network, directly producing cluster assignments via pseudo labels (Gansbeke et al., 2020; Qian, 2023a). Despite significant progress, most existing methods rely heavily on human priors, such as pre-specifying the number of clusters K (Ronen et al., 2022; Aghayengejeh et al., 2026) or implicitly assuming balanced data distributions (Zhang et al., 2023; Li & Jia, 2025). Such priors are often unavailable in real world scenarios, severely limiting the generalization and practicality of these methods.

Contrastive Learning is an instance-based unsupervised learning method that learns shared representations across similar instances and discriminates between dissimilar instances, bringing similar instances closer together and pushing dissimilar instances farther apart in the feature space (Wu et al., 2018; He et al., 2020). According to the granularity of contrast, contrastive learning methods can be broadly categorized into instance-level and prototypical contrastive learning (Peng et al., 2022). Instance-level contrastive learning treats each sample as a distinct class to achieve discrimination. However, its reliance on a large number of negative samples may introduce the false negative problem, where semantically similar samples are incorrectly pushed apart (Chuang et al., 2020; Balmaseda et al., 2025). To address this challenge, prototypical contrastive learning emerges, which introduces learnable prototypes to capture semantic centers and performs contrastive learning at the cluster level, aiming to alleviate the false negative issue while learning more globally structured representations (Caron et al., 2020; Li et al., 2021a). Despite these advances, existing prototypical contrastive learning methods still face critical limitations. They either adopt an independent multi-granularity strategy, where different granularities are optimized independently without mutual interaction, or establish hierarchical multi-granularity relationships while still relying on predefined cluster numbers and other priors. Such dependence on human priors can limit the generalization ability of these models in open-world scenarios.

Discussions. Existing contrastive clustering methods typically rely on predefined clustering structures to construct self-supervised signals, constraining the model with specific structural assumptions before the data structure is discovered and limiting the autonomous discovery of sample relationships. Therefore, a more fundamental question arises: in the absence of reliable external priors, how can structural information for constructing self-supervised signals be discovered from the data itself and continuously updated as representation learning progresses? Existing contrastive clustering research has yet to systematically explore this question.

3 Proposed Method

3.1 Motivation

Self-supervised learning aims to move beyond the reliance on human annotations by constructing supervision signals directly from the data to learn meaningful representations (Gui et al., 2024). For deep clustering, a more fundamental question arises: is the model actively discovering the structure of the data, or merely adapting the learned feature representations to a human-prespecified structure? Existing deep clustering methods often still require predefined structural priors, such as specifying the number of clusters K, based on which pseudo-labels, positive/negative sample relations, or cluster assignments are constructed to guide representation learning. In other words, before the data structure is discovered, the model is already required to use a structural prior to construct supervision signals and organize the relationships among samples.

This issue goes beyond whether the specified K is correct. More fundamentally, from the perspective of representation learning, structural priors directly participate in the construction of supervision signals and thereby influence the optimization of the feature representation space (Li et al., 2021b; Liu et al., 2023). To provide a unified description, we represent the sample relations induced by a structural prior as a relation matrix $\mathbf{S}$, where $S_{ij} \in \{0, 1\}$ indicates whether the structural prior considers samples i and j similar. For a fixed relation matrix $\mathbf{S}$ obtained from the current structural discovery step, given sample feature representations $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^{N}$, a structure-guided self-supervised

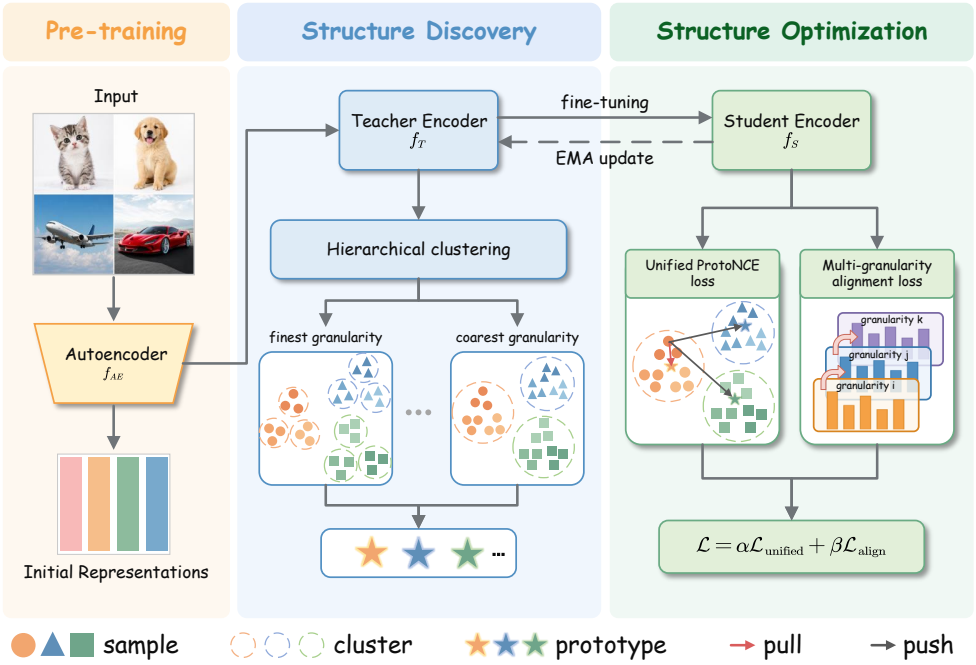

Figure 2: The framework of DISC.

objective can be abstracted as

$$\mathcal{L}(\mathbf{Z};\mathbf{S}) = \sum_{i,j} \left[S_{ij}\ell^{+}(\mathbf{z}_i,\mathbf{z}_j) + (1-S_{ij})\ell^{-}(\mathbf{z}_i,\mathbf{z}_j) \right], \tag{1}$$

where ℓ^{+} and ℓ^{-} encourage sample representations to be pulled together and pushed apart, respectively. Let $\mathcal{L}_i$ denote the anchor-wise contribution associated with $\mathbf{z}_i$. Its gradient can then be written as

$$\nabla_{\mathbf{z}_i}\mathcal{L}_i = \sum_{j} \left[S_{ij}\nabla_{\mathbf{z}_i}\ell^{+} + (1-S_{ij})\nabla_{\mathbf{z}_i}\ell^{-} \right]. \tag{2}$$

This formulation shows that a structural prior is not merely an auxiliary condition outside representation learning. Rather, it directly enters the optimization process through the induced sample relations. For two structural priors $\mathbf{S}$ and $\mathbf{S}'$, the corresponding difference in optimization gradients is

$$\nabla_{\mathbf{z}_i}\mathcal{L}_i(\mathbf{S}) - \nabla_{\mathbf{z}_i}\mathcal{L}_i(\mathbf{S}') = \sum_{j} \left(S_{ij} - S'_{ij}\right)\left(\nabla_{\mathbf{z}_i}\ell^{+} - \nabla_{\mathbf{z}_i}\ell^{-}\right). \tag{3}$$

Therefore, different structural priors can induce different optimization signals for the feature representations. Furthermore, data may exhibit multiple organizations at different granularities, and such structural information is generally unknown before learning. Consequently, a single prescribed structure may capture only part of the underlying organization, while constraining representation learning to a particular view and limiting the exploitation of complementary structural information across different granularities.

This perspective motivates us to reconsider the source of self-supervised signals in contrastive clustering. Existing methods typically rely on a predefined clustering structure to construct supervision for representation optimization. A more natural paradigm is to continuously discover structural information from evolving feature representations and use it to construct supervision for subsequent representation learning. Under this paradigm, structural information is no longer externally prescribed but progressively discovered and exploited throughout the learning process.

Based on this observation, we propose DISC, which progressively discovers multi-granularity hierarchical structures from current feature representations and converts them into self-supervised signals for representation optimization without requiring a predefined clustering structure. By continuously coupling Structure Discovery with Structure Optimization, DISC allows the supervisory signals to evolve with the representation space, forming a contrastive clustering paradigm that discovers and exploits structural information throughout learning rather than prescribing a target organization in advance.

3.2 DISC: Discovering Intrinsic Structure via Contrastive Clustering

DISC derives supervisory signals directly from the evolving feature representations. Rather than fixing the clustering structure before representation learning, it progressively infers structural information from the current representation space and feeds it back into subsequent representation optimization. In this way, structure discovery and representation learning continuously inform each other throughout the learning process.

The network architecture consists of three components: an Autoencoder f_{AE} for initializing the feature space, a Student Encoder f_{S}, and a Teacher Encoder f_{T}. The Student Encoder is optimized via gradient descent, while the parameters of the Teacher Encoder are updated via exponential moving average (EMA) of the Student's parameters, thereby providing stable feature representations. The training process follows a two-stage paradigm consisting of Structure Discovery and Structure Optimization. In the Structure Discovery stage, the Teacher Encoder obtains feature representations and applies the FINCH algorithm (Sarfraz et al., 2019). To bridge the two stages, we incorporate a lightweight contrastive fine-tuning mechanism between the E-step and M-step. It rapidly realigns the pseudo labels with the current hierarchy, reducing structural inconsistency caused by the lag between the Student Encoder and Teacher Encoder. In the Structure Optimization stage, we optimize the feature representations using two complementary losses: the Unified ProtoNCE loss $\mathcal{L}_{\text{unified}}$ and the multi-granularity alignment loss $\mathcal{L}_{\text{align}}$. This iterative process enables the hierarchical structure and feature representations to evolve collaboratively within an EM-style optimization framework, as illustrated in Figure 2.

Unified ProtoNCE loss. After discovering the hierarchical structure, DISC uses the resulting prototypes and augmented views to provide prototype-level and instance-level self-supervised signals for representation learning. In practice, we adopt a similar approach to PCL (Li et al., 2021a) to better balance local smoothness and global semantic compactness. The loss encourages the sample to align with both its assigned prototypes and augmented view, while pushing away non-assigned prototypes and other instances:

$$\mathcal{L}_{\text{unified}} = \frac{1}{2LB} \sum_{l=1}^{L} \sum_{i=1}^{B} \left[\ell_{\text{proto}}^{(l)}(i) + \ell_{\text{inst}}^{(l)}(i) \right], \tag{4}$$

where $\ell_{\text{proto}}^{(l)}(i)$ and $\ell_{\text{inst}}^{(l)}(i)$ denote the prototype-level and instance-level contrastive losses for sample i at the l-th granularity, respectively. Here, L is the number of granularity levels and B is the batch size.

For the prototype-level discrimination, the loss for sample i at the l-th granularity is defined as

$$\ell_{\text{proto}}^{(l)}(i) = -\log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{c}_i^{(l)} / \tau_{\text{proto}}\right)}{\mathcal{D}_i^{(l)}}, \tag{5}$$

where $\mathbf{z}_i$ is the L2-normalized feature of sample x_i, $\mathbf{c}_i^{(l)}$ is the prototype assigned to sample i at the l-th granularity, τ_{proto} is the temperature for prototype contrast, and $\mathcal{D}_i^{(l)}$ denotes the shared normalization term. Similarly, the instance-level contrastive loss is defined as

$$\ell_{\text{inst}}^{(l)}(i) = -\log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_i^{+} / \tau_{\text{inst}}\right)}{\mathcal{D}_i^{(l)}}, \tag{6}$$

where $\mathbf{z}_i^{+}$ is the feature of the augmented view of sample x_i and τ_{inst} is the temperature for instance contrast. The shared normalization term $\mathcal{D}_i^{(l)}$ contains all prototypes at the l-th granularity and the augmented views of all instances in the batch, placing prototype-level and instance-level candidates

under a common normalization. The detailed definitions of the two contrastive terms are provided in Appendix B.

Multi-granularity alignment loss. To enforce consistency across adjacent levels, DISC aligns the child-level and parent-level assignment distributions using the discovered hierarchy. Following the KL-based assignment alignment in DEC (Xie et al., 2016), we extend this principle to the multi-granularity setting. For each level, we obtain the soft assignment vector $\mathbf{q}_i^{(l)}$ by applying softmax to the prototype similarities. The child-level distribution is then projected into the parent-level space using the fixed child-to-parent mapping matrix $\mathbf{G}^{(l)}$:

$$\tilde{\mathbf{q}}_i^{(l+1)} = \mathbf{q}_i^{(l)} \mathbf{G}^{(l)}, \tag{7}$$

where $\mathbf{G}^{(l)}$ maps the prototypes at the l-th level to their corresponding parent prototypes at the $(l+1)$-th level, and $\tilde{\mathbf{q}}_i^{(l+1)}$ denotes the projected child-level distribution in the parent-level space. We then align the projected distribution with the parent-level assignment distribution using KL divergence:

$$\mathcal{L}_{\text{align}} = \frac{1}{L-1} \sum_{l=1}^{L-1} \frac{1}{B} \sum_{i=1}^{B} D_{\text{KL}} \left(\tilde{\mathbf{q}}_i^{(l+1)} \| \mathbf{q}_i^{(l+1)} \right), \tag{8}$$

where L is the number of granularity levels and B is the batch size. The alignment loss encourages the projected child-level assignments to remain consistent with the parent-level structure, promoting a coherent hierarchy across granularities.

Overall procedure and objective. At each EM round, the hierarchical structure discovered from the Teacher representations is fixed during the subsequent optimization step and used to jointly optimize the Student Encoder with the two loss components introduced above. The overall objective is defined as

$$\mathcal{L} = \alpha \mathcal{L}_{\text{unified}} + \beta \mathcal{L}_{\text{align}}, \tag{9}$$

where α and β are trade-off parameters balancing the two loss terms. After optimization, the Teacher Encoder is updated toward the Student Encoder via exponential moving average, and the updated Teacher representations are used for the next round of structure discovery. After the EM iterations converge, FINCH is applied to the final Teacher representations to obtain the multi-granularity clustering results.

4 Experiments

4.1 Experiment Settings

Datasets. We conducted experiments on several benchmark datasets, including HW2 (Zhang et al., 2026), Wiki (Rasiwasia et al., 2010), MNIST (LeCun et al., 1998), USPS (Hull, 1994), CIFAR-10 (Krizhevsky, 2009) and Fashion-MNIST (Xiao et al., 2017). For HW2, Wiki and USPS, we use all available samples. For the remaining datasets, we randomly sample subsets from their standard training sets: 5,000 images from MNIST, 10,000 images from CIFAR-10 and Fashion-MNIST. All datasets are preprocessed following standard protocols to ensure fair comparison with existing methods.

Implementation Details. We adopt a fully connected autoencoder as the feature extraction network, with the encoder architecture configured as $\dim(\mathbf{X})$-500-500-2000-128. All experiments are conducted on NVIDIA GPUs. The training process consists of two stages: autoencoder pre-training and EM alternating optimization. During the pre-training stage, the autoencoder is trained with mean squared error (MSE) loss using the Adam optimizer with a learning rate of $1\text{e}{-3}$ for 200 epochs. After pre-training, a copy of the encoder is maintained as the Teacher Encoder, whose parameters are updated from the Student Encoder via exponential moving average (EMA) with a momentum coefficient of 0.99. During the M-step, the Student Encoder is optimized using Adam with a cosine annealing learning-rate schedule. To improve generalization, Gaussian noise augmentation with a standard deviation of 0.05 is applied during training to construct two augmented views for instance-level contrastive learning. All experiments are performed with fixed random seeds to ensure reproducibility.

4.2 Main Results

Table 1: Clustering performance on benchmark datasets, with the best results highlighted in bold and the second-best results underlined. TLE denotes that the time limit was exceeded.

Datasets	FINCH *CVPR'19*	DeepDPM *CVPR'22*	PBC *PR'24*	Gauging-δ *TPAMI'25*	DPRC-KNN *PR'26*	HypCSE *AAAI'26*	SCMax *AAAI'26*	Ours
				ACC				
HW2	0.3745	0.1000	0.5150	0.1000	0.4600	0.5975	0.6115	**0.8345**
Wiki	0.3569	0.5042	0.3726	0.1619	0.4543	0.5816	0.5967	**0.6054**
MNIST	0.4710	0.3754	0.4024	0.1000	0.4910	0.5978	0.6004	**0.8336**
USPS	0.3030	0.6088	0.2854	TLE	0.4127	0.5934	0.5613	**0.8037**
Cifar10	0.1978	0.3756	0.3012	TLE	0.2012	0.8378	0.7650	**0.9006**
Fashion-MNIST	0.2013	0.5145	0.2290	TLE	0.3781	0.4860	0.4118	**0.6006**
				NMI				
HW2	0.5480	0.0000	0.6008	0.0000	0.4469	0.5721	0.6436	**0.7552**
Wiki	0.3909	0.4984	0.4858	0.0069	0.4389	0.5558	**0.5609**	0.5491
MNIST	0.6411	0.4786	0.6313	0.0000	0.5796	0.6000	0.7021	**0.7977**
USPS	0.4185	0.6521	0.6373	TLE	0.5082	0.5911	0.6891	**0.8383**
Cifar10	0.3678	0.5594	0.4842	TLE	0.1772	0.7299	0.7409	**0.7972**
Fashion-MNIST	0.3112	0.6064	0.4946	TLE	0.4928	0.5505	0.5881	**0.6448**
				PUR				
HW2	0.3745	0.1000	0.7050	0.1000	0.4940	0.6250	0.6115	**0.8345**
Wiki	0.3569	0.5094	**0.6818**	0.1619	0.4655	0.6389	0.6403	0.6420
MNIST	0.4710	0.3754	**0.8568**	0.1000	0.5300	0.6144	0.7658	0.8336
USPS	0.3030	0.6562	**0.9337**	TLE	0.4703	0.6512	0.5613	0.8047
Cifar10	0.1978	0.3756	0.7563	TLE	0.2550	0.8378	0.7999	**0.9006**
Fashion-MNIST	0.2013	0.5145	**0.7376**	TLE	0.4279	0.5336	0.7077	0.6007
				F-score				
HW2	0.2250	0.0182	0.2121	0.0182	0.3630	0.5933	0.5538	**0.8338**
Wiki	0.1556	0.4254	0.1212	0.0349	0.3000	0.5278	0.5538	**0.5606**
MNIST	0.3425	0.2154	0.0904	0.0182	0.4092	0.5880	0.4553	**0.8334**
USPS	0.0927	0.5212	0.0503	TLE	0.3064	0.5860	0.3960	**0.7015**
Cifar10	0.0659	0.2207	0.0506	TLE	0.1594	0.8384	0.7483	**0.9004**
Fashion-MNIST	0.0709	0.3946	0.0308	TLE	0.3004	0.4636	0.2100	**0.5401**

We first evaluate the clustering performance of DISC on multiple benchmark datasets and compare it with representative methods, including FINCH (Sarfraz et al., 2019), DeepDPM (Ronen et al., 2022), PBC (Motallebi, 2024), Gauging-δ (Yao et al., 2025), DPRC-KNN (Xie et al., 2026), HypCSE (Zeng et al., 2026) and SCMax (Zhang et al., 2026). The experimental results are presented in Table 1. Overall, DISC achieves the best or second-best clustering performance on most datasets, indicating its effectiveness across diverse data structures. Notably, DISC does not require a predefined clustering structure throughout the learning process. Instead, it progressively discovers structural information from the current feature representations and further transforms it into self-supervised signals to guide representation learning. Although some competing methods rely on predefined clustering structures or other structural priors, DISC achieves comparable or even superior performance.

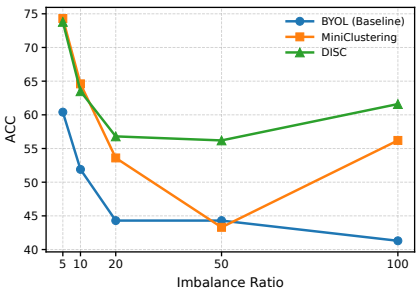

Figure 3: ACC comparison on CIFAR-10 under different imbalance ratios.

To further evaluate the robustness of DISC under distributional imbalance, following (Li et al., 2026), we conduct experiments on CIFAR-10, CIFAR-20 (Krizhevsky, 2009), and STL-10 (Coates et al., 2011) under different imbalance ratios and compare DISC with representative methods, including BYOL (Grill et al., 2020), SCAN (Gansbeke et al., 2020), SDCLR (Jiang et al., 2021), CC (Li et al., 2021b), IDFD (Tao et al., 2021), CoNR (Yu et al., 2023), ProPos (Huang et al., 2023), DMICC (Li et al., 2023b), SeCu (Qian, 2023b), LFSS (Li et al., 2025), ConMix (Li & Jia, 2025) and MiniClustering (Li et al., 2026). Unlike methods specifically designed to handle long-tailed distributions, DISC does not require reweighting strategies based on differences in sample distributions, nor does it adopt other specialized treatments for imbalanced data. The experimental results are presented in Table 2 and Figure 3. DISC maintains stable and competitive performance across different degrees of data imbalance. Even under substantially skewed data distributions, DISC can discover the intrinsic distributional structure of the data without additional human guidance.

Table 2: Clustering performance (%) on CIFAR-10, CIFAR-20, and STL-10 (imbalance ratio = 5).

Datasets	CIFAR-10				CIFAR-20				STL-10			
Metrics	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
SCAN *ECCV'20*	55.6	60.3	54.0	39.3	38.8	39.7	42.6	25.1	48.8	49.8	46.2	34.9
SDCLR *ICML'21*	44.1	50.5	43.4	38.0	40.2	38.7	40.3	23.8	35.8	37.4	34.4	19.3
CC *AAAI'21*	25.2	20.9	18.3	1.47	16.0	12.0	18.0	0.00	41.5	30.0	46.2	22.4
IDFD *ICLR'21*	56.7	63.0	51.8	37.1	31.2	30.7	30.5	16.2	41.6	41.2	39.6	24.3
CoNR *NeurIPS'23*	41.4	43.5	34.9	23.3	23.8	23.4	21.6	12.1	32.4	31.3	29.0	18.2
ProPos *TPAMI'23*	51.4	59.2	52.2	34.1	40.7	39.3	42.6	26.4	37.2	39.3	34.5	21.8
DMICC *AAAI'23*	40.6	42.5	36.9	25.4	25.2	23.1	20.6	7.5	45.9	45.6	45.3	35.5
SeCu *ICCV'23*	56.1	59.0	54.5	40.3	34.8	33.7	35.8	19.9	51.6	50.8	48.2	39.1
LFSS *ICML'25*	58.1	63.1	59.7	43.3	43.5	42.2	46.8	28.6	46.6	43.7	46.0	34.0
ConMix *ICLR'25*	61.6	65.4	59.8	45.6	42.8	41.0	43.9	27.7	48.9	49.7	48.8	35.9
MiniClustering *ICLR'26*	**74.3**	**72.6**	**69.9**	62.1	45.6	42.3	47.5	30.7	54.7	52.7	**54.4**	**42.8**
DISC (Ours)	73.8	71.0	69.8	**65.8**	**45.9**	**43.3**	**47.7**	**31.1**	**56.4**	**59.4**	50.7	40.2

Table 3: Multi-granular structures discovered by DISC on different datasets. Each cell reports the number of clusters K (first row), followed by the Silhouette Coefficient and Davies–Bouldin (DB) index (second row).

Dataset	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
HW2	480 0.1530 / 1.1572	108 0.1407 / 1.2929	30 0.4351 / 0.8064	**11** **0.9041 / 0.3219**	4 0.4647 / 1.3438	– – / –
Wiki	687 0.1381 / 1.1500	158 0.0390 / 1.3279	36 0.1938 / 1.0674	**9** **0.7708 / 0.5347**	4 0.5222 / 1.1114	– – / –
MNIST	1140 0.1008 / 1.2956	266 0.1080 / 1.4357	64 0.1372 / 1.0634	15 **0.7679 / 0.3793**	**7** 0.7124 / 0.7528	2 0.2447 / 1.6348
USPS	1596 -0.0457 / 1.7711	285 -0.0581 / 2.0988	58 -0.0651 / 1.8819	**13** 0.3907 / 1.1138	5 **0.7268 / 0.8551**	– – / –
CIFAR-10	1974 0.0128 / 1.5128	393 -0.0318 / 1.7973	74 0.0109 / 1.4729	**16** **0.6176 / 0.6063**	4 0.4515 / 1.1934	2 0.2553 / 1.4233
Fashion-MNIST	2110 0.0324 / 1.4196	403 -0.0542 / 1.8237	80 -0.0965 / 1.7782	**11** 0.5710 / **0.8789**	4 **0.6716** / 1.0069	– – / –

Overall, DISC demonstrates good adaptability across different datasets and distribution conditions. More importantly, these experimental results indicate that Reflexive Structural Supervision for Contrastive Clustering is a feasible learning paradigm, in which the structural supervision required for representation learning does not necessarily depend on structures prescribed by humans before learning, but can instead be progressively revealed by the data itself during the learning process. Moreover, DISC can discover and produce clustering partitions at multiple granularities, providing a direct reference for understanding and analyzing the intrinsic structure of the data, as shown in Table 3.

4.3 Ablation Study

The ablation results are presented in Table 4. We analyze DISC from the perspective of self-supervised signal construction. First, compared with single-granularity structural supervision using the ground-truth number of clusters K, multi-granularity supervision yields further performance improvements, indicating that a single prescribed granularity is insufficient to capture structural information across different granularities. Furthermore, the variant that independently learns different granularities performs worse than the one that explicitly models hierarchical relationships, indicating that the value of multi-granularity supervision lies not only in covering different granularities, but also in revealing the relationships among different levels of the data distribution.

Further examination shows that discovering the hierarchy only at the beginning of training and then keeping it fixed is still insufficient to reach the performance of the full model. As the feature representations continuously evolve, the fixed structure cannot keep reflecting these changes, limiting its ability to guide subsequent clustering optimization. Moreover, replacing the EMA update of the

Table 4: Ablation study on three representative datasets.

Variants	HW2				Wiki				MNIST			
	ACC	NMI	PUR	F-score	ACC	NMI	PUR	F-score	ACC	NMI	PUR	F-score
Fixed-Single Granularity	0.6725	0.6016	0.6725	0.6755	0.5161	0.5195	0.6015	0.4910	0.6178	0.6067	0.6614	0.6064
Fixed-Independent MG	0.6765	0.5842	0.6765	0.6668	0.5604	0.5259	0.6127	0.5325	0.6556	0.6517	0.6762	0.6543
Fixed-Hierarchical MG	0.7975	0.7356	0.7975	0.7916	0.5513	0.5386	0.6134	0.4794	0.7338	0.7473	0.7592	0.7269
One-shot Structure Discovery	0.6550	0.5747	0.6550	0.6555	0.5523	0.5450	0.6235	0.5127	0.5986	0.5787	0.6430	0.5884
No-Momentum Teacher	0.6670	0.6291	0.6750	0.6645	0.5722	**0.5519**	0.6120	0.5177	0.7500	0.7426	0.7662	0.7413
DISC w/o $\mathcal{L}_{\text{unified}}$	0.3580	0.4502	0.3950	0.2642	0.3566	0.3469	0.3772	0.3137	0.2732	0.1814	0.2860	0.2500
DISC w/o $\mathcal{L}_{\text{align}}$	0.8155	0.7394	0.8155	0.8156	0.5747	0.5467	0.6197	0.5520	0.7804	0.7865	0.8228	0.7587
DISC	**0.8345**	**0.7552**	**0.8345**	**0.8338**	**0.6054**	0.5491	**0.6420**	**0.5606**	**0.8336**	**0.7977**	**0.8336**	**0.8334**

Teacher with direct synchronization of the Student parameters also leads to a performance drop, indicating that structure discovery requires not only continuous updates but also stable structural signals.

Finally, removing either $\mathcal{L}_{\text{unified}}$ or $\mathcal{L}_{\text{align}}$ leads to a performance drop, further indicating that Structure Discovery and Structure Optimization need to be effectively coupled. Specifically, $\mathcal{L}_{\text{unified}}$ uses the currently discovered multi-granularity structure to optimize the representations, while $\mathcal{L}_{\text{align}}$ further enforces structural consistency across different granularities. Overall, these results support the core learning paradigm of DISC: self-supervised signals no longer rely on a predefined clustering structure, but are progressively derived from multi-granularity structural information discovered from the current feature representations and continuously fed back into representation learning, allowing Structure Discovery and representation learning to mutually reinforce each other throughout training. Detailed experimental settings for all variants are provided in Appendix D.

4.4 Convergence and Efficiency Analysis

We further analyze the convergence behavior and computational efficiency of our framework. During training, the objective value generally decreases as the EM rounds proceed and gradually becomes stable after several rounds, indicating that the iterative interaction between structure discovery and representation optimization progressively reaches a stable state. As the feature representations become more stable, the resulting structures and supervisory signals also tend to change less substantially, leading to stable optimization in subsequent EM rounds.

Each EM round mainly involves feature extraction by the Teacher Encoder, hierarchical structure discovery with FINCH, and representation optimization by the Student Encoder. FINCH constructs the hierarchy from first-neighbor relations (Sarfraz et al., 2019), without computing or storing a complete pairwise distance matrix, achieving $O(N \log N)$ time and $O(N)$ space complexity. Moreover, the input size decreases at coarser hierarchy levels, as FINCH operates on cluster-level mean representations. Therefore, despite repeated structure discovery across EM rounds, the overall computational cost remains manageable.

5 Conclusion

In this work, we revisit a fundamental question in contrastive clustering: should the structure of data be prescribed before learning, or discovered through learning itself? Our study shows that predefined structural assumptions can constrain representation learning and may conflict with the goal of discovering structure from unlabeled data. Based on this insight, we propose DISC, which incorporates structure discovery into the contrastive clustering process, allowing structural information to progressively emerge from the data and serve as self-supervised signals. Experimental results show that, without introducing structural priors, contrastive clustering can still learn effective representations and achieve competitive clustering performance. This finding suggests a shift in contrastive clustering from adapting representations to a prescribed organization toward discovering the structural supervision needed for representation learning from the data itself.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research and paper writing, including discussing research ideas and supporting research execution, retrieving and discovering related work, deriving and proving mathematical claims, and assisting with English writing, language polishing, and translation. All AI-assisted content was manually reviewed by the authors. Mathematical derivations and theoretical conclusions were independently verified by the authors, and the research methods, experimental designs, and final research contributions were developed and determined by the authors. The authors take full responsibility for all final content of the paper.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the DISC framework, training procedure, and optimization objectives in Section 3.2, with theoretical analyses and corresponding proofs provided in Appendices B and C. Detailed designs of the ablation experiments are further provided in Appendix D. The source code used in our experiments is provided as supplementary material, including the implementation of DISC and the relevant experimental code, to facilitate the reproduction of our experimental results.

REFERENCES

Nazila Pourhaji Aghayengejeh, M. A. Balafar, Jafar Tanha, and M. Alper Selver. Beyond predefined clusters: A comprehensive review of clustering methods for unknown numbers of clusters. *IEEE Transactions on Knowledge and Data Engineering*, 38(7):4121–4138, 2026. doi: 10.1109/TKDE.2026.3680286.

Vicente Balmaseda, Bokun Wang, Ching-Long Lin, and Tianbao Yang. Discovering global false negatives on the fly for self-supervised contrastive learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2697–2714. PMLR, 2025.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Computer Vision – ECCV 2018*, volume 11218 of *Lecture Notes in Computer Science*, pp. 132–149. Springer, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Guikun Chen, Xia Li, Yi Yang, and Wenguan Wang. Neural clustering based visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5714–5725, 2024.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/63c3ddcc7b23daa1e42dc41f9a44a873-Abstract.html.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 215–223. PMLR, 2011.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x.

Kien Do, Truyen Tran, and Svetha Venkatesh. Clustering by maximizing mutual information across views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9928–9938, 2021.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Computer Vision – ECCV 2020*, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024. doi: 10.1109/TPAMI.2024.3415112.

Yuanfan Guo, Minghao Xu, Jiawen Li, Bingbing Ni, Xuanyu Zhu, Zhenbang Sun, and Yi Xu. Hcsc: Hierarchical contrastive selective coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9706–9715, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. doi: 10.1109/CVPR42600.2020.00975.

Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7509–7524, 2023. doi: 10.1109/TPAMI.2022.3216454.

J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9865–9874, 2019.

Ziyu Jiang, Tianlong Chen, Bobak J. Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4927–4939. PMLR, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40238–40271. Curran Associates, Inc., 2023a. doi: 10.52202/075280-1749. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7e8bb8d17bb1cb24dfe972a2f8ff2500-Paper-Conference.pdf.

Hongyu Li, Lefei Zhang, and Kehua Su. Dual mutual information constraints for discriminative clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8571–8579, 2023b. doi: 10.1609/aaai.v37i7.26032.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021a. URL https://openreview.net/forum?id=KmykpuSrjcq.

Wei Li. Adaptive prototypical contrastive learning for time series clustering. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.

Yunfan Li, Peng Hu, Jerry Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 8547–8555. AAAI Press, 2021b.

Zhixin Li and Yuheng Jia. Conmix: Contrastive mixup at representation level for long-tailed deep clustering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3lH8WT0fhu.

Zhixin Li, Yuheng Jia, Hui Liu, and Junhui Hou. Learning from sample stability for deep clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34904–34919. PMLR, 2025.

Zhixin Li, Yuheng Jia, Guanliang Chen, Hui LIU, and Junhui Hou. Mini-cluster guided long-tailed deep clustering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3JlljaiQwR.

Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Zhen Wang, Ke Liang, Wenxuan Tu, Liang Li, Jingcan Duan, and Cancan Chen. Hard sample aware network for contrastive deep graph clustering. In *Proc. of AAAI*, 2023.

Hassan Motallebi. Efficient and robust clustering based on backbone identification. *Pattern Recognition*, 155:110635, 2024. ISSN 0031-3203. doi: 10.1016/j.patcog.2024.110635.

Chuang Niu, Hongming Shan, and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278, 2022.

Meixin Peng, Xin Juan, and Zhanshan Li. Graph prototypical contrastive learning. *Information Sciences*, 612:816–834, 2022. doi: 10.1016/j.ins.2022.09.013.

Qi Qian. Stable cluster discrimination for deep clustering. In *IEEE/CVF International Conference on Computer Vision (ICCV 2023)*, 2023a.

Qi Qian. Stable cluster discrimination for deep clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16645–16654, 2023b.

Nikhil Rasiwasia, José Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 251–260, 2010. doi: 10.1145/1873951.1873987.

Meitar Ronen, Shahaf E. Finder, and Oren Freifeld. Deepdpm: Deep clustering with an unknown number of clusters. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9851–9860, 2022. doi: 10.1109/CVPR52688.2022.00963.

Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8926–8935, 2019. doi: 10.1109/CVPR.2019.00914.

Yaling Tao, Kentaro Takagi, and Kouta Nakata. Clustering-friendly representation learning via instance discrimination and feature decorrelation. In *International Conference on Learning Representations*, 2021.

C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1): 95–103, 1983. doi: 10.1214/aos/1176346060.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3733–3742, 2018. doi: 10.1109/CVPR.2018.00393.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Jingyu Xie, Caiming Zhong, Yu Tian, Lianyu Hu, and Tiejun Pan. Robust density clustering based on density peak regions and k-nearest neighbors. *Pattern Recognition*, 174:113021, 2026. ISSN 0031-3203. doi: 10.1016/j.patcog.2025.113021.

Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 478–487. PMLR, 2016.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3861–3870. PMLR, 2017.

Xihong Yang, Yue Liu, Sihang Zhou, Siwei Wang, Wenxuan Tu, Qun Zheng, Xinwang Liu, Liming Fang, and En Zhu. Cluster-guided contrastive graph clustering network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10834–10842, 2023.

Jinli Yao, Jie Pan, and Yong Zeng. Gauging-δ: A non-parametric hierarchical clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4897–4907, 2025. doi: 10.1109/TPAMI.2025.3545573.

Chunlin Yu, Ye Shi, and Jingya Wang. Contextually affinitive neighborhood refinery for deep clustering. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5778–5790, 2023. doi: 10.52202/075280-0253.

Guangjie Zeng, Hao Peng, Angsheng Li, Li Sun, Chunyang Liu, Shengze Li, Yicheng Pan, and Philip S. Yu. Hyperbolic continuous structural entropy for hierarchical clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28094–28102, 2026. doi: 10.1609/aaai.v40i33.40035.

Lijun Zhang, Suyuan Liu, Siwei Wang, Shengju Yu, Xueling Zhu, Miaomiao Li, and Xinwang Liu. Parameter-free clustering via self-supervised consensus maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28310–28318, 2026.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023. doi: 10.1109/TPAMI.2023.3268118.

Huasong Zhong, Jianlong Wu, Chong Chen, Jianqiang Huang, Minghua Deng, Liqiang Nie, Zhouchen Lin, and Xian-Sheng Hua. Graph contrastive clustering. In *IEEE ICCV*, 2021.

Sheng Zhou, Hongjia Xu, Zhuonan Zheng, Jiawei Chen, Zhao Li, Jiajun Bu, Jia Wu, Xin Wang, Wenwu Zhu, and Martin Ester. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *ACM Comput. Surv.*, 57(3), 2024. doi: 10.1145/3689036.

Zhengzhong Zhu, Pei Zhou, Lanxi Bai, Li Cheng, Jia Nie, Shiquan Min, and Jiangping Zhu. Reliable clustering number estimation for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30162–30171, 2026.

A Algorithm Procedure

The overall optimization procedure of DISC is summarized in Algorithm 1. We first pretrain an autoencoder to initialize the Student Encoder, which is then copied to initialize the Teacher Encoder. During each EM iteration, the Teacher Encoder performs hierarchical structure discovery based on the current feature representations. The resulting prototypes and hierarchical relationships are then used to construct self-supervised signals for the Student Encoder. The Student Encoder then updates its feature representations under these signals, exploring a representation space that reflects the current organization of the data. Subsequently, the Teacher Encoder is updated via exponential moving average (EMA), providing a relatively stable representation reference for subsequent structure discovery. Through this iterative optimization process, DISC progressively refines its representation of the intrinsic data structure and ultimately reveals the multi-granular hierarchical organization of the data.

Algorithm 1 DISC: Discovering Intrinsic Structure via Contrastive Clustering

Input: Dataset $\mathcal{X}$; loss weights α, β; EM rounds T; inner epochs E.
Output: Final multi-granularity hierarchy $\mathcal{H}^*$.

```
Pretrain AutoEncoder on X and initialize Student Encoder f_S.
Initialize Teacher Encoder: f_T ← f_S.
for t = 1, ..., T do
    Z_T ← f_T(X).
    H_t ← FINCH(Z_T).
    Obtain prototypes q, assignments Y, and parent-child mappings G from H_t.
    Fine-tune f_S using the self-supervised signals from H_t.
    for e = 1, ..., E do
        for each mini-batch B do
            Generate two augmented views B^(1), B^(2).
            Compute L_unified and L_align.
            Update f_S by minimizing L = αL_unified + βL_align.
            Update f_T using EMA.
        end for
    end for
end for
Z* ← f_T(X).
H* ← FINCH(Z*).
```

B Gradient Derivation of the Total Loss

In the main text, we define the total loss as the following weighted combination:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{unified}} + \beta \mathcal{L}_{\text{align}}, \tag{10}$$

where $\mathcal{L}_{\text{unified}}$ is the Unified ProtoNCE loss, $\mathcal{L}_{\text{align}}$ is the multi-granularity alignment loss, α is a dynamic weight that decays during training, and β is a fixed hyperparameter.

Since the total loss is a linear combination of the two loss terms, its gradient with respect to the encoder parameters θ is given by the weighted sum of the individual gradients derived in Appendices B.1 and B.2:

$$\frac{\partial \mathcal{L}}{\partial \theta} = \alpha \frac{\partial \mathcal{L}_{\text{unified}}}{\partial \theta} + \beta \frac{\partial \mathcal{L}_{\text{align}}}{\partial \theta}. \tag{11}$$

B.1 Gradient of Unified ProtoNCE Loss

The Unified ProtoNCE loss $\mathcal{L}_{\text{unified}}$ is defined as:

$$\mathcal{L}_{\text{unified}} = -\frac{1}{L}\sum_{l=1}^{L}\frac{1}{B}\sum_{i=1}^{B}\frac{1}{2}\Bigg[\log\frac{\exp(\mathbf{z}_i \cdot \mathbf{c}_{y_i^{(l)}}^{(l)}/\tau_{\text{proto}})}{\sum_{k=1}^{K_l}\exp(\mathbf{z}_i \cdot \mathbf{c}_k^{(l)}/\tau_{\text{proto}}) + \sum_{j=1}^{B}\exp(\mathbf{z}_i \cdot \mathbf{z}_j^{+}/\tau_{\text{inst}})} + \log\frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_i^{+}/\tau_{\text{inst}})}{\sum_{k=1}^{K_l}\exp(\mathbf{z}_i \cdot \mathbf{c}_k^{(l)}/\tau_{\text{proto}}) + \sum_{j=1}^{B}\exp(\mathbf{z}_i \cdot \mathbf{z}_j^{+}/\tau_{\text{inst}})}\Bigg], \tag{12}$$

where $\mathbf{z}_i$ is the embedding of sample i, $\mathbf{z}_i^+$ is its augmented view, $\mathbf{c}_k^{(l)}$ is the k-th prototype at level l, K_l is the number of prototypes at that level, $y_i^{(l)}$ is the corresponding prototype index, and τ_{proto} and τ_{inst} are the temperatures.

To simplify the derivation, we denote the shared denominator as:

$$Z_i = \sum_{k=1}^{K_l}\exp(\mathbf{z}_i \cdot \mathbf{c}_k^{(l)}/\tau_{\text{proto}}) + \sum_{j=1}^{B}\exp(\mathbf{z}_i \cdot \mathbf{z}_j^{+}/\tau_{\text{inst}}). \tag{13}$$

We then define the prototype-level and instance-level softmax probabilities as:

$$p_{i,k}^{(l)} = \frac{\exp(\mathbf{z}_i \cdot \mathbf{c}_k^{(l)}/\tau_{\text{proto}})}{Z_i}, \tag{14}$$

and

$$q_{i,j} = \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j^+ / \tau_{\text{inst}})}{Z_i}. \tag{15}$$

The loss can thus be rewritten as:

$$\mathcal{L}_{\text{unified}} = -\frac{1}{L}\sum_{l=1}^{L}\frac{1}{2B}\sum_{i=1}^{B}\left(\log p^{(l)}_{i,y_i^{(l)}} + \log q_{i,i^+}\right). \tag{16}$$

Taking the derivative with respect to $\mathbf{z}_i$, we obtain:

$$\frac{\partial \mathcal{L}_{\text{unified}}}{\partial \mathbf{z}_i} = -\frac{1}{LB}\sum_{l=1}^{L}\left[\frac{\mathbf{c}^{(l)}_{y_i^{(l)}}}{2\tau_{\text{proto}}} + \frac{\mathbf{z}_i^+}{2\tau_{\text{inst}}} - \sum_{k=1}^{K_l} p^{(l)}_{i,k}\frac{\mathbf{c}^{(l)}_k}{\tau_{\text{proto}}} - \sum_{j=1}^{B} q^{(l)}_{i,j}\frac{\mathbf{z}_j^+}{\tau_{\text{inst}}}\right]. \tag{17}$$

The gradient with respect to the encoder parameters θ is then obtained via the chain rule:

$$\frac{\partial \mathcal{L}_{\text{unified}}}{\partial \theta} = \sum_{i=1}^{B}\frac{\partial \mathcal{L}_{\text{unified}}}{\partial \mathbf{z}_i}\cdot\frac{\partial \mathbf{z}_i}{\partial \theta}. \tag{18}$$

This gradient reflects the balance between attraction and competition. The attraction comes from prototype-level and instance-level supervision, while the competition arises from the common normalization denominator. Its analytic form is well-defined and can be directly optimized via standard backpropagation, confirming that the Unified ProtoNCE loss is differentiable and amenable to gradient-based optimization.

B.2 Gradient of the Multi-granularity Alignment Loss

The multi-granularity alignment loss $\mathcal{L}_{\text{align}}$ is defined as:

$$\mathcal{L}_{\text{align}} = \frac{1}{L-1}\sum_{l=1}^{L-1}\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K_{l+1}} q^{(l+1)}_{i,k}\log\frac{q^{(l+1)}_{i,k}}{\tilde{q}^{(l)}_{i,k}}, \tag{19}$$

where K_l is the number of prototypes at level l, $q^{(l)}_{i,k}$ is the k-th component of the soft assignment vector of sample i at level l, and $\tilde{q}^{(l)}_{i,k}$ is the projected distribution of the child-level assignments into the parent-level space.

To simplify the derivation, we rewrite the alignment loss as:

$$\mathcal{L}_{\text{align}} = \frac{1}{L-1}\sum_{l=1}^{L-1}\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K_{l+1}} q^{(l+1)}_{i,k}\left(\log q^{(l+1)}_{i,k} - \log \tilde{q}^{(l)}_{i,k}\right). \tag{20}$$

Taking the derivative with respect to $\mathbf{z}_i$, noting that for any level l:

$$\mathbf{q}_i^{(l)} = \text{softmax}\left(\mathbf{z}_i \cdot \mathbf{C}_l^\top / \tau_{\text{proto}}\right), \tag{21}$$

and $\tilde{q}^{(l)}_{i,k}$ is obtained from $q^{(l)}_{i,k}$ via a fixed projection matrix, we apply the chain rule level-wise:

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{align}}}{\partial \mathbf{z}_i} = \frac{1}{L-1}\sum_{l=1}^{L-1}\frac{1}{B}\Bigg[&\sum_{k=1}^{K_{l+1}}\frac{\partial \mathcal{L}_{\text{align}}}{\partial q^{(l+1)}_{i,k}}\cdot\frac{\partial q^{(l+1)}_{i,k}}{\partial \mathbf{z}_i} \\ &+ \sum_{k=1}^{K_{l+1}}\frac{\partial \mathcal{L}_{\text{align}}}{\partial \tilde{q}^{(l)}_{i,k}}\cdot\sum_{k'=1}^{K_l}\mathbf{G}_{k',k}\frac{\partial q^{(l)}_{i,k'}}{\partial \mathbf{z}_i}\Bigg],\end{aligned} \tag{22}$$

where

$$\frac{\partial \mathcal{L}_{\text{align}}}{\partial q^{(l+1)}_{i,k}} = \log\frac{q^{(l+1)}_{i,k}}{\tilde{q}^{(l)}_{i,k}} + 1, \tag{23}$$

and

$$\frac{\partial \mathcal{L}_{\text{align}}}{\partial \tilde{q}_{i,k}^{(l)}} = -\frac{q_{i,k}^{(l+1)}}{\tilde{q}_{i,k}^{(l)}}, \tag{24}$$

with $\mathbf{G}_{k',k}$ being the fixed mapping matrix that projects each child-level prototype k' to its corresponding parent-level prototype k, where both the prototypes $\mathbf{C}_l$ and the mapping matrix $\mathbf{G}$ are treated as fixed with respect to $\mathbf{z}_i$ during the gradient derivation, as they are obtained from the teacher encoder and remain constant during the feature update step.

Taking the derivative of the softmax function with respect to $\mathbf{z}_i$, we have:

$$\frac{\partial q_{i,k}^{(l)}}{\partial \mathbf{z}_i} = \frac{1}{\tau_{\text{proto}}} q_{i,k}^{(l)} \left(\mathbf{C}_{l,k} - \sum_{k'} q_{i,k'}^{(l)} \mathbf{C}_{l,k'} \right). \tag{25}$$

The gradient with respect to the encoder parameters θ is then obtained via the chain rule:

$$\frac{\partial \mathcal{L}_{\text{align}}}{\partial \theta} = \sum_{i=1}^{B} \frac{\partial \mathcal{L}_{\text{align}}}{\partial \mathbf{z}_i} \cdot \frac{\partial \mathbf{z}_i}{\partial \theta}. \tag{26}$$

This gradient naturally decomposes into two contributions: one from the parent-level soft assignments that directly supervise the feature, and one from the child-level assignments projected through the mapping matrix $\mathbf{G}$, which encourages the child-level distributions to align with the parent-level structure. Its analytic form is well-defined and enables standard backpropagation, ensuring that the multi-granularity alignment loss is differentiable and trainable.

C Analysis of the Alternating Optimization

C.1 Convergence Analysis of the Alternating Optimization

Our optimization procedure follows an EM-like alternating structure. The E-step extracts features from the Teacher Encoder and constructs a multi-granularity hierarchy via hierarchical clustering. The M-step first performs contrastive fine-tuning of the Student Encoder using the current clustering labels, followed by gradient-based optimization of the full objective with the hierarchy fixed. After each round, the Teacher Encoder is updated via exponential moving average (EMA).

Let θ_t and $\mathcal{H}_t$ denote the Student encoder parameters and the hierarchical structure at the beginning of outer iteration t, respectively. The objective function is:

$$F_t(\theta, \mathcal{H}) = \alpha_t \mathcal{L}_{\text{unified}}(\theta, \mathcal{H}) + \beta \mathcal{L}_{\text{align}}(\theta, \mathcal{H}), \tag{27}$$

where α_t is a dynamic weight that decays during training and β is a fixed hyperparameter.

Since the hierarchical structure constructed in the E-step remains fixed during the M-step, and the loss function has been shown to be differentiable in Appendix B, we assume that for fixed $\mathcal{H}$, $F_t(\cdot, \mathcal{H})$ has an L_t-Lipschitz continuous gradient and the learning rate satisfies:

$$0 < \eta_t \leq \frac{1}{L_t}. \tag{28}$$

Then the standard descent inequality for L_t-smooth functions (Li et al., 2023a) gives:

$$F_t(\theta_{t+1}, \mathcal{H}) \leq F_t(\theta_t, \mathcal{H}) - \frac{\eta_t}{2} \|\nabla_\theta F_t(\theta_t, \mathcal{H})\|^2. \tag{29}$$

The E-step updates the hierarchical structure from $\mathcal{H}_t$ to $\mathcal{H}_{t+1}$ via hierarchical clustering, without changing θ_t. However, since the clustering procedure does not necessarily minimize F_t exactly, it may introduce a non-negative optimization error in the objective.

In practice, the EMA update of the Teacher encoder and the decay of α_t also introduce small perturbations to the objective. We collectively treat the E-step error, EMA drift, and α-induced variation as a total error term:

$$\Delta_t = \epsilon_t^{\mathrm{E}} + \delta_t^{\mathrm{EMA}} + \delta_t^{\alpha}. \tag{30}$$

In our setting, the effect of these perturbations is bounded. We therefore assume that the total perturbation is summable, i.e.:

$$\sum_{t=0}^{\infty} \Delta_t < \infty. \tag{31}$$

Combining the M-step descent and the E-step error, we obtain the following perturbed descent inequality:

$$F_{t+1} \leq F_t - \frac{\eta_t}{2} \|\nabla_\theta F_t\|^2 + \Delta_t. \tag{32}$$

Since $\mathcal{L}_{\text{unified}} \geq 0$ and $\mathcal{L}_{\text{align}} \geq 0$, and $\alpha_t \geq 0$, $\beta \geq 0$, the total loss is bounded below by zero:

$$F_t \geq 0. \tag{33}$$

Under this assumption, summing the perturbed descent inequality gives:

$$\sum_{t=0}^{\infty} \eta_t \|\nabla_\theta F_t\|^2 < \infty. \tag{34}$$

Assuming a lower-bounded learning rate $\eta_t \geq \eta_{\min} > 0$, it follows that:

$$\lim_{t \to \infty} \|\nabla_\theta F_t\| = 0, \tag{35}$$

which means the optimization converges to a first-order stationary point of the limiting objective. In the idealized case where the perturbations vanish ($\Delta_t = 0$), the objective sequence is monotonically non-increasing and bounded below, and therefore converges.

In summary, our alternating optimization framework is essentially a perturbed descent process that drives the gradient norm to zero, implying that every accumulation point is first-order stationary for the limiting objective. The analysis is inspired by classical convergence results for EM-type procedures (Dempster et al., 1977; Wu, 1983), and is further supported by the loss curves observed in our experiments.

D Ablation Study Details

D.1 Single- vs. Multi-Granularity Supervision

To isolate the effect of supervision granularity, we construct a Fixed-Single Granularity variant, which uses the ground-truth number of clusters K and performs K-Means to obtain a single-granularity clustering structure, replacing the multi-granularity hierarchical structure in the full model while keeping all other training settings unchanged.

The experimental results are presented in Table 4. Even when given the ground-truth number of clusters as prior knowledge, the single-granularity variant performs substantially worse than the full model across all four metrics on the three datasets. This indicates that single-granularity structural supervision is insufficient to fully characterize the intrinsic structure of the data. In contrast, DISC automatically discovers a multi-granularity hierarchical structure from the current feature representations and jointly exploits structural information at different granularities for clustering optimization, thereby overcoming the limitation of single-granularity supervision in oversimplifying the intrinsic structure of the data.

D.2 Fixed-Independent Multi-Granularity Supervision

To further investigate whether hierarchical relationships among different granularities are essential for clustering optimization, we construct a Fixed-Independent MG variant. Specifically, we use the

hierarchy chain output by the full model, $K_1, K_2, \ldots, K_m$, as mutually independent supervision signals. For each K_i, we separately perform K-Means on the same feature representations to obtain the corresponding supervision for guiding feature learning. Unlike the full model, no parent-child relationships are established among different granularities in this variant.

The experimental results show that the Fixed-Independent MG variant performs substantially worse than the full model, although it consistently outperforms the single-granularity variant. This shows that multi-granularity supervision indeed provides richer structural information than single-granularity supervision, but simply combining supervision signals from different granularities is insufficient to fully exploit this information. When structural information at different granularities is obtained independently, the resulting supervision signals may impose inconsistent relational constraints on the feature representations. For example, samples assigned to the same cluster at a fine granularity may be assigned to different clusters at a coarser granularity, causing conflicting optimization directions.

In contrast, DISC establishes parent-child relationships across granularities to maintain structural consistency and propagate information between levels. These results demonstrate that the benefit of multi-granularity supervision comes not only from the richer structural information provided by multiple granularities, but also from organizing this information into a coherent hierarchical structure.

D.3 Fixed-Hierarchical Multi-Granularity Supervision

To further investigate whether a predefined hierarchical structure is sufficient to achieve the benefit of multi-granularity supervision, we construct a Fixed-Hierarchical MG variant. Specifically, we use the multi-granularity hierarchical structure discovered by the full model, $K_1, K_2, \ldots, K_m$, as a fixed structural prior, and recursively perform K-Means according to these cluster numbers to preserve the parent-child relationships across granularities. Unlike DISC, this variant does not discover the hierarchical structure during training, but directly uses the fixed structure to construct multi-granularity supervision.

The experimental results show that the Fixed-Hierarchical MG variant still performs substantially worse than the full model. During training, the predefined hierarchy is used to establish explicit relationships across different granularities. However, as the feature representations are continuously updated, the fixed hierarchy lags behind the evolving representations, which may make the resulting supervision gradually inconsistent with the current feature space. In contrast, DISC continuously rediscovers multi-granularity structural information from updated feature representations and uses it to construct subsequent supervision, thereby maintaining consistency between structure discovery and clustering optimization.

D.4 One-shot Structure Discovery

To investigate the necessity of continuous structure discovery, we construct a One-shot Structure Discovery variant. This variant performs hierarchical clustering only once during training, while keeping all other training settings unchanged from those of the full model. The experimental results show that this variant still performs substantially worse than the full model, indicating that a single structure discovery step is insufficient to adapt to the continuously evolving feature representations. In contrast, DISC continuously updates the structural information, allowing subsequent supervision to adapt to changes in the representation space and thereby better coordinate structure discovery and clustering optimization.

D.5 No-Momentum Teacher

To investigate the role of the Teacher's momentum update in stabilizing structural information, we construct a No-Momentum Teacher variant by setting the momentum coefficient to 0, so that the Teacher no longer incorporates historical Student parameters through momentum updates, while keeping all other training settings unchanged from those of the full model.

The experimental results show that this variant performs worse than the full model, indicating that directly discovering structures from Student features is susceptible to unstable hierarchical clustering

Table 5: Clustering performance (%) on CIFAR-10, CIFAR-20, and STL-10 (imbalance ratio = 10).

Datasets	CIFAR-10				CIFAR-20				STL-10			
Metrics	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI	ACC	CAA	NMI	ARI
SCAN *ECCV'20*	52.0	50.1	59.4	43.8	37.0	32.4	37.8	23.1	47.1	45.4	47.5	34.2
SDCLR *ICML'21*	38.9	44.3	42.5	26.5	37.8	35.9	39.6	22.9	34.6	37.9	32.2	17.3
CC *AAAI'21*	40.6	27.5	43.9	18.8	19.9	14.3	21.9	1.1	43.0	35.3	44.7	25.4
IDFD *ICLR'21*	47.5	54.9	48.4	33.1	28.7	27.2	28.6	15.1	38.6	34.7	36.8	22.8
CoNR *NeurIPS'23*	31.4	44.3	29.2	17.8	20.3	17.9	17.3	8.0	34.8	32.5	30.7	21.0
ProPos *TPAMI'23*	46.1	49.3	52.5	34.2	36.8	33.6	40.1	22.5	35.6	38.2	37.2	23.9
DMICC *AAAI'23*	36.6	39.5	36.8	25.9	24.7	21.9	20.7	10.1	41.3	41.9	38.7	30.3
SeCu *ICCV'23*	51.5	57.0	54.8	39.6	32.0	30.7	33.3	17.6	48.2	49.6	50.0	36.8
LFSS *ICML'25*	56.3	59.7	57.9	43.0	41.1	40.0	46.8	28.4	43.3	42.3	45.4	30.8
ConMix *ICLR'25*	53.3	58.2	57.1	40.8	41.7	39.3	43.6	27.0	47.4	48.7	48.2	33.9
MiniClustering *ICLR'26*	**64.6**	**61.4**	**63.9**	**56.7**	**44.3**	**41.5**	**47.2**	**30.4**	**52.0**	**54.4**	**50.8**	**39.1**
DISC (Ours)	63.4	58.8	59.9	51.3	40.3	34.5	41.8	26.9	50.0	50.9	46.5	37.5

results. DISC addresses this issue by accumulating historical representations through the momentum Teacher and smoothly updating the structural information. This further demonstrates that the Teacher/Student architecture is not an auxiliary engineering design, but an important component for stabilizing dynamic structural supervision.

E Additional Experimental Results

E.1 Robustness under Different Imbalance Ratios

In the main paper, we set the imbalance ratio to 5 on CIFAR-10, CIFAR-20, and STL-10 to evaluate the performance of DISC under distributional imbalance. To further assess the robustness of DISC under more severe distributional skew, we increase the imbalance ratio to 10 and conduct additional experiments using the same datasets and experimental settings. The results are presented in Table 5.