ORCA: Orthogonal Residual Consensus Alignment for Multi-View Clustering

Xiaojian Ding, Xian Li, Xiaoying Zhu, Lin Zhao, Kaixiang Wang*
School of Computer and Artificial Intelligence,
Nanjing University of Finance and Economics, Nanjing, China
xjding@nufe.edu.cn, 1120241388@stu.nufe.edu.cn,
1120241394@stu.nufe.edu.cn, 1120241386@stu.nufe.edu.cn,
kxwang@nufe.edu.cn

Abstract

Deep multi-view clustering often suffers from redundant information accumulation during fusion and indiscriminate alignment across representation levels. We propose Orthogonal Residual Consensus Alignment (ORCA) to address these challenges. Its semantic incremental recurrent fusion module extracts cross-view orthogonal residuals as non-redundant increments and adaptively weights recurrent states through divergence-guided gating. Its consensus-calibrated hierarchical contrastive alignment module uses the global consensus to weight contrastive alignment of view-level, pairwise, and recurrent representations. Together, these modules balance view-specific structure with cross-view consistency. Experiments on five benchmarks against fifteen competing methods demonstrate ORCA's effectiveness, including a 5.25% accuracy improvement over the strongest baseline on 100leaves. Ablation studies validate both modules. Code is available at `https://github.com/XiaojianDing/2026-NeurIPS-ORCA`.

1 Introduction

With the ubiquity of heterogeneous data, Multi-View Clustering (MVC) has emerged as a fundamental research topic in unsupervised learning by integrating consistency and complementarity across diverse views [14]. Although deep learning has significantly enhanced the ability to model nonlinear correlations, constructing compact and discriminative consensus representations remains a core challenge due to inherent issues such as information redundancy and view bias. Consequently, exploring robust multi-view fusion and alignment strategies is essential for achieving precise data partitioning and profound semantic understanding [30].

Recent research in deep multi-view clustering (MVC) has shifted from simple multi-view feature aggregation toward more sophisticated semantic consistency modeling and alignment. Recent advancements highlight the efficacy of multi-level contrastive learning, exemplified by frameworks such as 3MC [5], MCMC [2], and DMvCGSA [8]. By establishing a hierarchical representation structure that integrates instance-, cluster-, and prototype-level information, these models substantially improve the compactness and discriminative power of learned features. Simultaneously, there is a growing interest in structured graph representation learning, with methods like LFCGL [26], SemGProp [23], and MISCC [14] utilizing graph convolutional networks or self-supervised graph propagation to model topological relationships and improve robustness against noise. Furthermore, researchers are increasingly exploring disentangled representation learning and redundancy elimination; for instance, STMVC [24] removes cross-view redundancy through an information-theoretic lens, while SCDFL

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

[41] and DMCAR-MVC [29] explicitly decouple shared consensus semantics from view-specific information to mitigate representation collapse.

Despite recent progress in deep multi-view clustering, existing methods still suffer from two key limitations. First, many fusion-based approaches aggregate heterogeneous view representations in a redundancy-blind manner, where shared and complementary information are not explicitly distinguished. As a result, repetitive cross-view components may be repeatedly propagated during feature fusion, while truly discriminative view-specific semantics are diluted. This issue becomes more pronounced in attention-based fusion, since iterative aggregation tends to contract the latent representation space and may lead to over-smoothed features or even representation collapse, weakening the separability of clustering boundaries. Although recent methods introduce contrastive objectives or attention mechanisms to enhance cross-view consistency, they usually rely on learned weights to implicitly suppress redundancy, without explicitly modeling the non-redundant semantic residuals between views. Second, existing alignment strategies often treat multi-view semantics at a single level or align different representations with uniform constraints, ignoring the hierarchical nature of cross-view information. In practice, useful clustering cues may appear at multiple semantic levels, including individual view structures, local pairwise relations, and intermediate states produced during recurrent interaction. However, directly enforcing these representations to align with the final consensus may introduce noisy or weakly relevant semantics into the optimization process. Although multi-scale or contrastive learning methods have been explored, they generally lack a consensus-anchored calibration mechanism to evaluate which semantic levels are more consistent with the global clustering structure. Consequently, the learned representation may fail to simultaneously preserve local structural details and maintain globally discriminative consensus semantics.

To address the above challenges, we propose Orthogonal Residual Consensus Alignment (ORCA), a unified framework for robust multi-view clustering. ORCA is built upon two complementary modules: Semantic Incremental Recurrent Fusion (SIRF) and Consensus-Calibrated Hierarchical Contrastive Alignment (CCHA). Instead of directly aggregating heterogeneous view representations, SIRF first identifies non-redundant semantic increments from cross-view orthogonal residuals, enabling each view to absorb complementary information while suppressing repetitive components. It further formulates multi-view integration as a recurrent semantic interaction process, where the evolution of fused representations is regulated by information gain to alleviate redundant assimilation and representation collapse. Meanwhile, CCHA constructs a hierarchical semantic pool from view-level, local pairwise, and recurrent interaction representations, and uses the learned global consensus as an anchor to adaptively calibrate their contributions through attention-weighted contrastive alignment. In this way, ORCA jointly realizes redundancy-aware fusion, consensus formation, and hierarchical semantic calibration within a single optimization framework.

The main contributions of this work are:

- We propose a geometric projection-based strategy to explicitly separate redundant and complementary semantics, adaptively enriching core views with non-redundant increments to reduce information accumulation.
- We design a calibration mechanism that uses global consensus as an anchor to adaptively weight and align multi-level semantics, ensuring consistent and discriminative clustering representations.
- Evaluations across diverse modalities demonstrate that ORCA outperforms fifteen state-of-the-art baselines, with ablation and visualization results confirming its robust discriminative capability.

2 Related Work

Feature-Alignment-Based Contrastive MVC. Early methods utilize autoencoders to extract view-specific features and align them in latent spaces via contrastive losses. Examples include DealMVC [37] and SCMC [9], which enforce local or subspace contrastive constraints. MCoCo [44] and DFMVC [43] extend this to the semantic label space for category-level alignment, while CDARL [28] employs dual adversarial and contrastive learning to minimize distribution discrepancies. Moreover, methods like TDLSR [19] develop theory-driven label-specific representations to robustly align features even under incomplete data scenarios. To improve efficiency, DWCL [40] introduces a "Best-Other" mechanism to select reliable views. However, these methods often apply uniform

contrastive intensity to all views, failing to distinguish between shared and view-specific information, which results in redundant information propagation.

Multi-Level and Multi-Granularity Contrastive Learning. Recent studies explore hierarchical structures to capture multi-granularity semantics. 3MC [5] performs layer-wise feature and label alignment, while MCMC [2] proposes a triple-contrastive framework across instance, cluster, and prototype levels. DMCAR-MVC [29] utilizes a decoupled Mixture-of-Experts (MoE) architecture with orthogonal constraints to separate shared and specific representations. Additionally, PSGVC [13] leverages pseudo-label similarity graphs to guide global and local alignment. Despite their success, these methods lack a global consensus anchor to calibrate the contributions of different semantic levels, often leading to the indiscriminate alignment of noisy or weakly relevant semantics.

Fusion-Driven Consensus Learning and Redundancy Suppression. Effective fusion is critical for DMVC. While methods like GCFAgg [35] and MISCC [14] focus on aggregating complementary info through cross-view correlations, others emphasize redundancy suppression. GMAE [45] decouples features into view-specific and view-common embeddings, whereas STMVC [24] uses an information-theoretic approach to eliminate redundant components. Parallelly, approaches like AFIP [18] theoretically prove that calibrating attention to correct visual distraction is necessary to reduce noisy implications during feature integration. Furthermore, ILASC [38] adaptively fuses views using Transformer Pooling and aligns the overall clustering structure. Nevertheless, most fusion strategies lack explicit modeling of non-redundant semantic residuals, which can dilute complementary information.

Pseudo-Label Guided and Structure-Aware Contrastive Learning. To mitigate unreliable positive/negative pairs and noisy annotations, researchers have introduced external guidance. For instance, semi-supervised paradigms have been enhanced by unbiased losses in DIMvSML [15], view-specific transformers guided by pseudo-labels in SMVTEP [16], and theory-inspired noise-resistant frameworks like DMMIvNL [17]. Additionally, PBDMSC [39] uses cluster assignments to select reliable samples, and STCMC-UR [10] filters false negatives via uncertainty estimation. SemGProp [23] propagates high-confidence pseudo-label structures to enhance consistency, while CGM-CUC [7] addresses unaligned scenarios through graph matching and iterative alignment. While these approaches improve signal reliability, they typically focus on single-level alignment and lack a systematic mechanism to calibrate hierarchical semantics against the global clustering goal.

3 Method

In this section, we detail the proposed ORCA framework, as shown in Figure 1. Our architecture relies on two coordinated core modules. First, the SIRF module explicitly separates shared redundancy from complementary residual semantics via geometric projection. By autonomously halting the semantic interaction, SIRF maximizes information gain without risking structural collapse. Second, the CCHA module is introduced to preserve multi-scale semantics. It constructs a hierarchical pool (spanning view-level, local pairwise, and recurrent representations) and adaptively aligns these features against a global consensus anchor via attention-guided contrastive learning, ultimately yielding consistent and robust clustering representations.

3.1 Semantic Incremental Recurrent Fusion

In multi-view clustering, existing attention-based fusion paradigms inherently act as low-pass filters, shrinking the latent space and causing severe dimensionality collapse. To overcome this, we propose the SIRF module, which formulates multi-view aggregation as a dynamical system. SIRF operates through two core stages: orthogonal residual semantic discovery and adaptive divergence-driven halting. By explicitly extracting topology-aware semantic increments and autonomously halting the recurrent interaction, SIRF prevents spectral degeneration and maximizes information gain, ultimately yielding a highly discriminative global representation.

3.1.1 Orthogonal Residual Semantic Discovery

Before mapping multiple views into a shared semantic space, we must recognize that different views often share significant overlapping information. Direct aggregation introduces massive feature

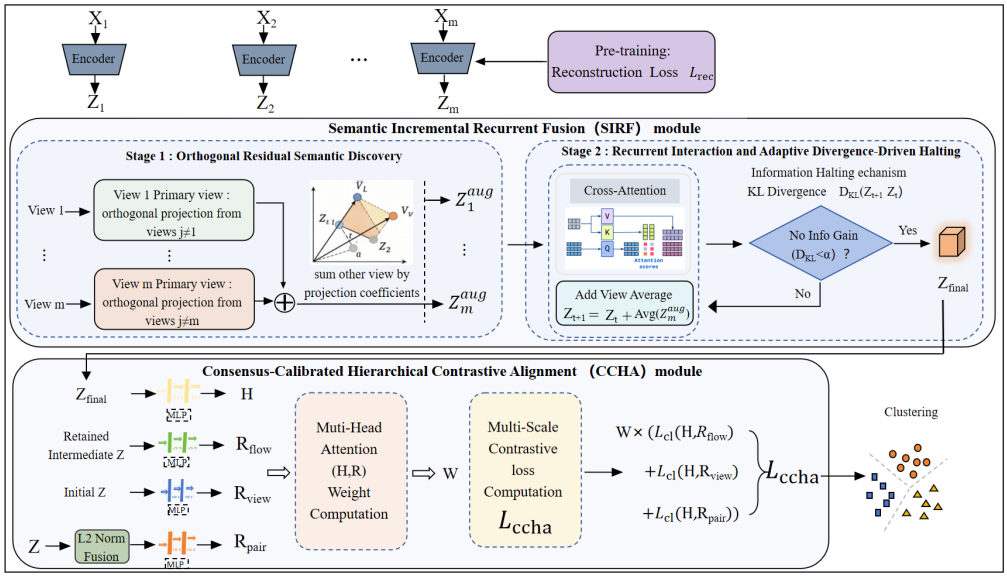


Figure 1: Overview of the proposed ORCA framework, consisting of three main components: (i) View-specific Autoencoders for latent feature extraction; (ii) The SIRF module, which extracts complementary residuals to distill a non-redundant global consensus $\mathbf{H}$; and (iii) The CCHA module, which adaptively aligns multi-scale representations (view-level, local, and recurrent) against the global consensus anchor.

redundancy, which overwhelms the view-specific discriminative structures. To explicitly isolate the truly complementary semantic information contributed by supplementary views, we propose an orthogonal residual semantic discovery strategy. From a geometric perspective, if we treat the features as vectors in a high-dimensional manifold, the component of a marginal view $\mathbf{Z}_{\text{other}}$ that is parallel to the core view $\mathbf{Z}_{\text{core}}$ represents redundant, already-known information. Conversely, the strictly orthogonal residual component constitutes the non-redundant semantic supplement.

Given V distinct initial views, we systematically treat each view v as the core view $\mathbf{Z}_{\text{core}}^{(v)}$, and the remaining views $u \in \{1, \ldots, V\} \setminus \{v\}$ as the marginal views $\mathbf{Z}_{\text{other}}^{(u)}$. We first project the marginal view onto the direction of the core view. This projection $\mathbf{Z}_{\text{proj}}^{(u \to v)}$, which encapsulates the structural consensus, is calculated as:

$$\mathbf{Z}_{\text{proj}}^{(u \to v)} = \left(\frac{\langle \mathbf{Z}_{\text{other}}^{(u)}, \mathbf{Z}_{\text{core}}^{(v)} \rangle}{\|\mathbf{Z}_{\text{core}}^{(v)}\|_2^2} \right) \mathbf{Z}_{\text{core}}^{(v)}. \tag{1}$$

To extract the pure novel information independent of the core view, we compute the orthogonal residual component by subtracting the parallel projection from the marginal view:

$$\mathbf{Z}_{\perp}^{(u \to v)} = \mathbf{Z}_{\text{other}}^{(u)} - \mathbf{Z}_{\text{proj}}^{(u \to v)}. \tag{2}$$

By measuring the relative strength of orthogonal residuals, the proposed method adaptively enriches each core view with non-redundant semantic increments, thereby reducing repetitive information accumulation in multi-view fusion. We define a scalar residual weight $g^{(u \to v)}$ as the relative magnitude of the orthogonal component:

$$g^{(u \to v)} = \frac{\|\mathbf{Z}_{\perp}^{(u \to v)}\|_2}{\|\mathbf{Z}_{\text{other}}^{(u)}\|_2}. \tag{3}$$

A larger $g^{(u \to v)}$ explicitly indicates that the marginal view introduces a higher proportion of complementary residual semantics. Finally, we inject this weighted orthogonal residual into the core view to construct the enhanced, information-maximized initial representation $\hat{\mathbf{Z}}^{(v)}$:

$$\hat{\mathbf{Z}}^{(v)} = \mathbf{Z}_{\text{core}}^{(v)} + \sum_{u \neq v} g^{(u \to v)} \cdot \mathbf{Z}_{\perp}^{(u \to v)}. \tag{4}$$

This rigorous geometric stripping completely abandons empirical guessing, ensuring that the enhanced feature set $\mathbf{Z}^{(0)} = \{\hat{\mathbf{Z}}^{(1)}, \ldots, \hat{\mathbf{Z}}^{(V)}\}$ spans a maximally informative and non-redundant subspace based on complementary semantics rather than repetitive information.

3.1.2 Recurrent Interaction and Adaptive Divergence-Driven Halting

Using the information-enhanced feature embeddings $\mathbf{Z}^{(0)}$, we construct a cross-view interaction space to model global contextual dependencies. To prevent the semantic flow from drifting during deep interactions and to anchor the representations to a stable structural baseline, we compute the global mean of the initial raw views as $\bar{\mathbf{Z}} = \frac{1}{V}\sum_{v=1}^{V} \mathbf{Z}_{\text{raw}}^{(v)}$. At the t-th iteration of our dynamic interaction system, the intermediate cross-view semantic representation $\mathbf{Z}^{(t)}$ is iteratively updated via a self-attention mechanism combined with this explicit structural residual:

$$\mathbf{Z}^{(t)} = \text{Softmax}\left(\frac{\mathbf{Z}^{(t-1)}\mathbf{W}_Q(\mathbf{Z}^{(t-1)}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{Z}^{(t-1)}\mathbf{W}_V + \bar{\mathbf{Z}}. \tag{5}$$

While this attention operator effectively establishes long-range correlations across different views, its softmax-induced convex combination fundamentally induces metric distortion. Over successive recurrent aggregations, the spectral radius of the attention Jacobian matrix is strictly bounded below 1. Even with the injection of the global residual $\bar{\mathbf{Z}}$, this contractive mapping ultimately behaves as a low-pass filter over deep iterations, leading to catastrophic dimensionality collapse where clustering margins are obliterated.

To counteract this topological degeneration without disrupting the parallel efficiency of batch training, we shift from heuristic structural constraints to an information-theoretic soft-gating mechanism. We dynamically monitor the instantaneous Information Gain during the semantic interaction process using the Kullback-Leibler (KL) divergence. Let $\mathcal{P}(\mathbf{Z}^{(t)})$ denote the feature distribution of the system at iteration t. The instantaneous information gain $\Delta\mathcal{I}^{(t)}$ is quantified as:

$$\Delta\mathcal{I}^{(t)} = \mathcal{D}_{KL}\Big(\mathcal{P}(\mathbf{Z}^{(t)}) \parallel \mathcal{P}(\mathbf{Z}^{(t-1)})\Big). \tag{6}$$

Drawing inspiration from adaptive computation time frameworks, we map this information gain to an adaptive retention weight $\alpha_t \in (0, 1)$ through a learnable gating function:

$$\alpha_t = \sigma\left(\mathbf{W}_\alpha \Delta\mathcal{I}^{(t)} + \mathbf{b}_\alpha\right), \tag{7}$$

where $\sigma(\cdot)$ is the Sigmoid activation, and $\mathbf{W}_\alpha$, $\mathbf{b}_\alpha$ are trainable parameters. This mechanism acts as a soft-halting gate: when the system actively learns new consensus semantics ($\Delta\mathcal{I}^{(t)}$ is large), the interaction continues; as the information exchange saturates ($\Delta\mathcal{I}^{(t)} \to 0$), α_t smoothly decays, effectively terminating the repetitive assimilation of saturated states.

The model synthesizes the optimal consensus representation by accumulating the dynamically weighted states across all $T_{\max}$ potential iterations:

$$\mathbf{Z}_{\text{final}} = \sum_{t=1}^{T_{\max}} \alpha_t \mathbf{Z}^{(t)}. \tag{8}$$

Ultimately, to project this information-maximized, dynamically halted consensus representation into a highly discriminative latent space for downstream tasks, we pass $\mathbf{Z}_{\text{final}}$ through a multi-layer perceptron. This yields the final global semantic representation $\mathbf{H}_{\text{final}}$:

$$\mathbf{H}_{\text{final}} = \text{MLP}(\mathbf{Z}_{\text{final}}; \psi), \tag{9}$$

where ψ represents the learnable parameters. By continuously capturing cross-view consensus and autonomously halting the flow before the onset of structural collapse, $\mathbf{H}_{\text{final}}$rigorously preserves intrinsic information volume while establishing clear boundaries for clustering.

3.2 Consensus-Calibrated Hierarchical Contrastive Alignment

While the SIRF module effectively distills a global consensus $\mathbf{H}$, relying solely on the final layer output may neglect rich hierarchical semantics embedded in the fusion trajectory. To avoid indiscriminate

constraints in multi-scale semantic alignment, we design a consensus-anchored hierarchical semantic calibration strategy. By constructing a hierarchical semantic pool and using the global consensus representation as the semantic anchor, CCHA adaptively weights different semantic levels through multi-head attention and performs weighted contrastive alignment, ensuring that $\mathbf{H}$ encapsulates both robust local details and consistent global patterns.

3.2.1 Hierarchical Semantic Pool Construction

Before applying contrastive objectives, we systematically construct a hierarchical semantic pool denoted as $\mathcal{Z}_{\text{pool}}$. In the context of our CCHA module, "semantics" refers to the manifold topological structures, latent clustering cues, and discriminative features captured across different representation stages and receptive fields. We preserve these multi-scale semantics because single-level representations often struggle to balance local feature integrity with global clustering consistency. By capturing the data topology across multiple receptive fields, this hierarchy comprehensively characterizes the evolutionary trajectory of information from isolation to assimilation across three distinct tiers:

View-level scale ($\mathcal{S}_{\text{view}}$): We directly preserve the raw structural information from each individual view's latent embedding, yielding $\mathcal{S}_{\text{view}} = \{\mathbf{z}_v\}_{v=1}^{V}$. This foundational tier retains the initial, independent structures of individual views to fundamentally prevent the loss of discriminative, view-specific information.

Local pairwise scale ($\mathcal{S}_{\text{pair}}$): To explicitly capture cross-view complementarity and bridge local view discrepancies, every unordered pair (i, j) of the initial embeddings is fused using ℓ_2-norm based adaptive weights:

$$\omega_i = \frac{\|\mathbf{z}_i\|_2}{\|\mathbf{z}_i\|_2 + \|\mathbf{z}_j\|_2}, \qquad \omega_j = 1 - \omega_i. \tag{10}$$

This yields a set of locally fused representations $\mathcal{S}_{\text{pair}} = \{\omega_i \mathbf{z}_i + \omega_j \mathbf{z}_j \mid 1 \le i < j \le V\}$, which explicitly records local pairwise interactions and complementary semantic correlations.

Recurrent interaction scale ($\mathcal{S}_{\text{flow}}$): During the previous divergence-guided recurrent fusion process, the cross-attention mechanism generates a trajectory of intermediate states $\mathbf{Z}^{(t)}$ driven by instantaneous information gain. To capture this progressive assimilation of multi-view semantics and record the dynamic formation of a global consensus while preventing semantic drift, we anchor these intermediate states to a stable global baseline. Specifically, we compute the global mean anchor of the initial raw views, $\bar{\mathbf{Z}} = \frac{1}{V}\sum_{v=1}^{V} \mathbf{z}_v$, and inject it as a structural residual into each intermediate state. This yields the recurrent interaction set:

$$\mathcal{S}_{\text{flow}} = \left\{ \mathbf{Z}^{(t)} + \bar{\mathbf{Z}} \mid t = 1, \ldots, T_{\max} \right\}. \tag{11}$$

We aggregate all the above multi-scale features into a unified candidate pool $\mathcal{Z}_{\text{pool}} = \mathcal{S}_{\text{view}} \cup \mathcal{S}_{\text{pair}} \cup \mathcal{S}_{\text{flow}}$. To align them into a shared semantic subspace for contrastive learning, every candidate $\mathbf{z}_m \in \mathcal{Z}_{\text{pool}}$ is projected through a shared multi-layer perceptron:

$$\mathbf{R}_m = \text{MLP}(\mathbf{z}_m; \Theta), \qquad m = 1, \ldots, M, \tag{12}$$

where M is the total number of candidate representations across all three scales (i.e., $M = V + \frac{V(V-1)}{2} + T_{\max}$).

Constructing a multi-scale semantic pool offers clear advantages over traditional contrastive frameworks. Single-scale alignment may introduce noise and dilute view-specific clustering cues. In contrast, our hierarchical pool preserves semantic information across isolated views, local synergies, and global consensus, providing diverse anchors for attention-based selection. This enables adaptive assessment of each representation's relevance to the global clustering objective, mitigating topological degradation while preserving local details and enhancing global cluster discrimination.

3.2.2 Consensus-Anchored Calibration Weighting

Not all multi-scale representations contribute equally to the final consensus. To adaptively regulate the influence of each representation $\mathbf{R}_m$ during contrastive alignment, we utilize the global consensus representation as the semantic anchor for our multi-head attention mechanism. Unlike standard self-attention, we use the global consensus $\mathbf{H}$ as the Query to calculate the correlation between

different level representations and the clustering consensus. For an attention head $h \in \{1, \ldots, H_{att}\}$, we compute the linear projections:

$$\mathbf{Q}^{(h)} = \mathbf{H}W_Q^{(h)}, \quad \mathbf{K}_m^{(h)} = \mathbf{R}_m W_K^{(h)}, \tag{13}$$

where $W_Q^{(h)}, W_K^{(h)} \in \mathbb{R}^{d \times d_k}$ are learnable parameters. The attention alignment score between the global consensus anchor $\mathbf{H}$ and the m-th scale representation $\mathbf{R}_m$ is computed as:

$$\alpha_m^{(h)} = \frac{\exp\left(\langle \mathbf{Q}^{(h)}, \mathbf{K}_m^{(h)} \rangle / \sqrt{d_k}\right)}{\sum_{t=1}^{M} \exp\left(\langle \mathbf{Q}^{(h)}, \mathbf{K}_t^{(h)} \rangle / \sqrt{d_k}\right)}. \tag{14}$$

To obtain a comprehensive scale-level importance weight, we average the attention intensity across all H_{att} heads: $r_m = \frac{1}{H_{att}} \sum_{h=1}^{H_{att}} \alpha_m^{(h)}$. The final normalized contrastive weight for each scale is derived via softmax:

$$w_m = \frac{\exp(r_m)}{\sum_{t=1}^{M} \exp(r_t)}, \qquad m = 1, \ldots, M. \tag{15}$$

A higher w_m indicates that the specific representation $\mathbf{R}_m$ is closer to the global clustering structure, thereby warranting a stronger constraint during alignment.

3.2.3 Unified Objective Optimization

To jointly optimize the global consensus and the multi-scale representations, we decompose the consensus-calibrated hierarchical contrastive alignment loss ($\mathcal{L}_{\text{ccha}}$) into three distinct scale-specific objectives. Let $\mathcal{I}_{\text{view}}$, $\mathcal{I}_{\text{pair}}$, and $\mathcal{I}_{\text{flow}}$ denote the index sets for the candidate representations belonging to the view-level ($\mathcal{S}_{\text{view}}$), local pairwise ($\mathcal{S}_{\text{pair}}$), and recurrent interaction ($\mathcal{S}_{\text{flow}}$) scales, respectively.

Utilizing the dynamically calibrated weights w_m derived from the global anchor, representations from each specific scale are forced to align with the global consensus $\mathbf{H}$ using the weighted InfoNCE objective. The three scale-specific alignment losses are formulated as follows:

$$\mathcal{L}_{\text{view}} = -\sum_{m \in \mathcal{I}_{\text{view}}} \sum_i w_m \log \frac{\exp\left(\text{sim}(\mathbf{R}_m^i, \mathbf{H}^i)/\tau\right)}{\sum_k \exp\left(\text{sim}(\mathbf{R}_m^i, \mathbf{H}_k^{\text{neg}})/\tau\right)}, \tag{16}$$

$$\mathcal{L}_{\text{pair}} = -\sum_{m \in \mathcal{I}_{\text{pair}}} \sum_i w_m \log \frac{\exp\left(\text{sim}(\mathbf{R}_m^i, \mathbf{H}^i)/\tau\right)}{\sum_k \exp\left(\text{sim}(\mathbf{R}_m^i, \mathbf{H}_k^{\text{neg}})/\tau\right)}, \tag{17}$$

$$\mathcal{L}_{\text{flow}} = -\sum_{m \in \mathcal{I}_{\text{flow}}} \sum_i w_m \log \frac{\exp\left(\text{sim}(\mathbf{R}_m^i, \mathbf{H}^i)/\tau\right)}{\sum_k \exp\left(\text{sim}(\mathbf{R}_m^i, \mathbf{H}_k^{\text{neg}})/\tau\right)}, \tag{18}$$

where i denotes the sample index within a batch, $\mathbf{H}_k^{\text{neg}}$ represents negative consensus samples, and τ is the temperature scaling parameter.

The comprehensive consensus-calibrated hierarchical contrastive alignment loss is then constructed by integrating these three components:

$$\mathcal{L}_{\text{ccha}} = \alpha \mathcal{L}_{\text{view}} + \beta \mathcal{L}_{\text{pair}} + \gamma \mathcal{L}_{\text{flow}}, \tag{19}$$

where α, β, and γ are trade-off hyperparameters designed to balance the macro level contributions of the view-level, local pairwise, and recurrent interaction scales, while w_m regulates the micro level importance of individual representations.

To simultaneously ensure that the view-specific encoders $E_v(\cdot)$ and decoders $D_v(\cdot)$ preserve essential intra-view semantic structures without distortion, a standard reconstruction regularization term is included:

$$\mathcal{L}_{\text{rec}} = \sum_{v=1}^{V} \|\mathbf{X}_v - D_v(E_v(\mathbf{X}_v))\|_F^2. \tag{20}$$

The overall training objective of the network integrates the reconstruction constraints with the dynamically weighted hierarchical alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_{\text{ccha}}, \tag{21}$$

where λ is a balancing hyperparameter.

After the network converges, clustering is performed directly on the optimal global consensus $\mathbf{H} \in \mathbb{R}^{N \times d}$ using k-means:

$$\min_{\mathbf{U},\mathbf{C}} \|\mathbf{H} - \mathbf{U}\mathbf{C}\|_F^2, \quad \text{s.t.} \quad \mathbf{U}\mathbf{1} = \mathbf{1}, \quad \mathbf{U} \geq \mathbf{0}. \tag{22}$$

4 Experiments

4.1 Experimental Setup

Datasets: We evaluate our method on five diverse multi-view benchmarks. Fashion [31], Synthetic3D [12], and Cifar10 [3] serve as representative vision datasets featuring distinct visual descriptors. MNIST-USPS [25] provides complementary views for cross-domain digit clustering, while 100leaves [20] encompasses diverse botanical features (shape, texture, and margin) for plant species clustering.

Implementation Details: To ensure fair comparisons and reproducibility, we use a unified model configuration for all datasets. Input data are vectorized before training. For feature extraction, each view employs an identical fully connected (FC) autoencoder with an encoder of four layers: Input – FC500 – FC500 – FC2000 – FC64, and a mirrored decoder. Following the extraction of the 64-dimensional latent embeddings, we construct the hierarchical semantic pool encompassing the view-level, local pairwise, and recurrent interaction scales. To align these multi-scale features into a shared semantic subspace for contrastive learning, every candidate representation within the pool is fed into a shared lightweight MLP (FC64 – FC20) to obtain the projected representation $\mathbf{R}_m$. ReLU activations follow each hidden layer except the output layers. The model is trained using Adam optimizer with a fixed learning rate of 0.0003 and no weight decay or learning rate scheduler. Experiments run on a workstation with Intel i7-12700K CPU, 32 GB RAM, and NVIDIA RTX 4070 GPU using PyTorch 1.10.2.

Comparison Methods: To demonstrate the superiority of our proposed method, we evaluate it against fifteen state-of-the-art deep multi-view clustering baselines. Specifically, CSOT [42] employs optimal transport for cross-view semantic alignment, while DealMVC [37] captures discriminative features via pseudo-label graph learning. Several methods utilize contrastive learning frameworks: SCM [22] integrates it with graph alignment, CVCL [4] optimizes it for large-scale scenarios, SCMVC [27] combines it with spectral clustering, SEM [33] enhances it with semantic information, MCMC [2] captures multilevel representational structures and enhances instance-level compactness via a dual self-supervised learning framework, and the recent HCLAA [6] achieves fine-grained semantic decoupling through hierarchical contrastive learning and a dual-path attention mechanism. In terms of feature representation, DDMVC [34] focuses on discriminative latent learning, MFLVC [32] provides a unified framework for extraction and fusion, MGCMVC [11] explores multi-granularity features to maintain structural integrity, and DMCAR [29] balances shared consensus and view-specific information through a decoupled mixture-of-experts architecture with context-aware routing. Additionally, MAGA [1] captures inter-view correlations through attention mechanisms, SSLNMVC [36] ensures consistency between view-specific and consensus labels using KL divergence, and GFSAF [21] addresses representation degradation by leveraging granular-ball learning to extract complementary representations alongside a multi-head cross-view attention fusion.

4.2 Comparative Experimental Results Analysis

To comprehensively evaluate the effectiveness and generalization of our proposed ORCA framework, we conduct experiments on several diverse multi-view datasets. As shown in Table 1, ORCA consistently achieves SOTA performance across all metrics. Specifically, on the 100leaves dataset, ORCA achieves a notable performance improvement with an ACC of 76.81%, outperforming the second-best baseline HCLAA (71.56%). On large-scale visual datasets such as Fashion and Cifar10, ORCA extracts highly robust representations, exceeding 99.5% ACC and 98.6% NMI. These remarkable

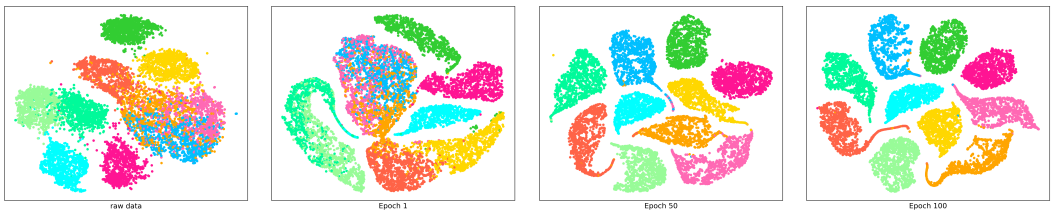
Table 1: Comparison of clustering performance (ACC, NMI, PUR) of different methods on five multi-view datasets.

Method	Fashion			Synthetic3d			Cifar10			MNIST-USPS			100leaves		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
SCMVC (2024)	99.34	98.28	99.34	97.00	87.04	97.00	99.51	98.58	99.51	99.72	99.18	99.72	59.13	79.99	62.00
CSOT (2024)	99.14	97.86	99.14	96.83	86.60	96.83	99.18	97.73	99.18	99.42	98.34	99.42	26.19	58.16	28.31
DealMVC (2023)	82.07	91.44	82.63	80.67	56.54	80.67	96.91	97.12	96.91	87.06	92.70	89.16	21.56	59.73	21.56
SCM (2024)	81.86	78.16	82.03	94.50	79.86	94.50	99.27	97.99	99.27	98.82	96.74	98.82	33.25	65.47	36.06
DDMVC (2025)	77.17	77.59	77.48	93.47	82.56	93.47	98.62	96.35	98.62	89.30	90.94	89.30	25.37	60.35	26.12
MFLVC (2022)	99.16	97.95	99.16	95.67	83.67	95.67	99.11	97.72	99.11	99.60	98.84	99.60	24.75	60.81	26.12
CVCL (2023)	99.31	98.20	99.31	89.20	67.17	89.20	97.26	95.48	97.26	99.58	98.77	99.58	35.24	67.53	35.24
MAGA (2024)	98.85	97.26	98.85	95.50	82.16	95.50	98.44	95.74	98.44	99.24	97.78	99.24	43.50	71.61	46.12
SEM (2023)	99.29	98.19	99.29	95.83	82.91	95.83	99.23	97.89	99.23	91.74	88.63	91.74	61.50	81.57	65.00
SSLNMVC (2025)	99.27	98.18	99.27	95.83	84.52	95.83	99.18	97.71	99.18	99.24	97.87	99.24	26.37	61.32	27.81
MGCMVC (2025)	99.09	97.78	99.09	96.83	86.43	96.83	99.43	98.34	99.43	99.60	98.85	99.60	33.12	66.61	34.50
MCMC (2025)	95.03	96.51	95.03	94.33	79.82	94.33	98.94	96.37	98.94	93.62	92.78	95.47	57.49	75.38	61.54
HCLAA (2026)	99.37	98.35	99.37	97.00	87.01	97.00	99.53	98.66	99.53	99.60	98.82	99.60	71.56	85.78	73.44
GFSAF (2026)	99.17	98.00	99.17	97.54	89.21	97.54	99.42	98.32	99.42	99.56	98.78	99.56	58.94	78.84	61.46
DMCAR (2026)	99.02	98.14	99.02	96.27	88.34	96.27	99.13	98.28	99.13	99.43	98.37	99.43	60.47	79.46	60.47
ORCA (Ours)	**99.52**	**98.69**	**99.52**	**97.67**	**89.56**	**97.67**	**99.54**	**98.72**	**99.54**	**99.74**	**99.26**	**99.74**	**76.81**	**89.77**	**78.63**

results validate the efficacy of ORCA's core design. By integrating SIRF with Orthogonal Residual Semantic Discovery, the model effectively preserves raw view-specific features while explicitly extracting complementary semantic increments. Furthermore, the CCHA framework, enhanced by consensus-anchored calibration weighting, allows ORCA to exploit hierarchical complementary information across views without blind multi-scale alignment, leading to highly accurate, robust, and semantically consistent clustering outcomes.

4.3 Model Analysis

Clustering Results Visualization Analysis: We use t-SNE to visualize the global features $\mathbf{H}$ learned

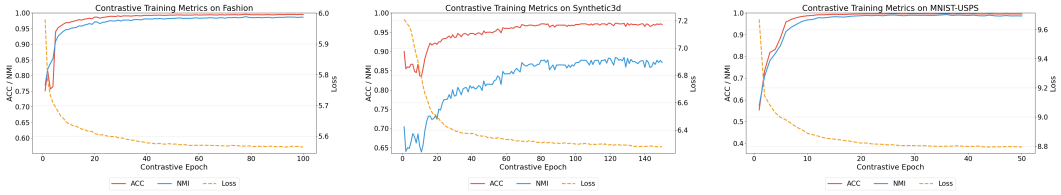


(a) Raw Data (b) Epoch 1 (c) Epoch 50 (d) Epoch 100

Figure 2: Clustering visualization of the Fashion dataset at different training stages: raw data, and epochs 1, 50, and 100.

by ORCA on the Fashion dataset, as shown in Figure 2. The plots show that as training progresses, especially at epochs 50 and 100, clusters become more distinct with clearer class boundaries and tighter intra-class grouping. This demonstrates the effectiveness of explicitly stripping redundant shared components via semantic incremental recurrent fusion and the adaptive guidance of the consensus-calibrated hierarchical contrastive alignment mechanism in improving feature separability and structural consistency.

Convergence Analysis: Figure 3 illustrates the convergence behavior of our ORCA model on three datasets: Fashion, Synthetic3D, and MNIST-USPS. The training loss steadily decreases while clustering metrics (ACC and NMI) rapidly improve and stabilize, demonstrating effective and stable optimization. Benefiting from our proposed ORSD and CCHA mechanism, ORCA adaptively integrates multi-view features to jointly learn complementary local and global semantics. The consensus-anchored calibration weighting further facilitates hierarchical semantic interaction, enhancing adaptability and robustness on visual data. This convergence trend validates the efficiency and stability of our framework, supporting the effectiveness of our novel design.



(a) Fashion (b) Synthetic3D (c) MNIST-USPS

Figure 3: Evaluation metrics as a function of training epochs on Fashion, Synthetic3D, and MNIST-USPS datasets.

Table 2: Ablation study of different modules on five datasets.

Modules	C_{rec}	C_H	C_{ccha}	C_W	C_{orsd}	**Fashion**			**Synthetic3D**			**Cifar10**			**MNIST-USPS**			**100leaves**		
						ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
M1	✓					62.76	58.31	63.58	68.29	33.81	68.29	92.20	84.99	92.20	46.90	44.46	51.22	54.99	77.12	58.34
M2	✓				✓	96.66	93.74	81.45	93.88	64.66	80.88	96.49	96.66	93.42	66.33	57.92	66.33	62.04	79.27	63.93
M3	✓	✓	✓			98.96	98.26	98.40	95.89	85.31	95.89	98.75	97.59	98.75	98.83	97.97	98.83	63.03	80.35	64.06
M4	✓			✓	✓	95.78	93.06	80.27	92.79	59.93	79.88	96.43	92.89	85.91	69.33	60.31	69.33	63.39	80.06	64.58
M5	✓	✓			✓	99.06	97.99	99.06	96.59	87.26	96.59	99.13	97.97	99.13	69.92	61.11	69.92	61.06	81.70	65.65
M6	✓		✓		✓	96.65	94.80	84.34	94.91	69.85	84.84	96.54	97.15	94.41	98.00	96.96	98.00	62.43	81.57	64.86
M7	✓	✓	✓	✓		99.02	97.88	99.02	96.64	87.79	96.88	99.10	98.03	99.10	98.89	98.02	98.89	63.70	81.58	64.61
M8	✓		✓	✓		96.50	94.77	83.49	94.83	68.39	82.52	97.05	96.10	92.28	99.07	97.95	99.07	61.71	82.79	66.58
M9	✓	✓		✓	✓	98.21	96.01	98.21	95.44	84.02	95.44	98.89	97.81	98.89	98.95	80.01	98.95	59.23	78.86	60.50
M10	✓	✓	✓		✓	98.88	98.21	98.88	96.94	87.56	97.32	98.60	98.00	98.60	99.02	97.55	99.02	62.55	82.82	65.51
M11	✓		✓			97.13	94.19	83.38	94.11	69.92	84.85	96.81	96.94	93.24	97.04	96.97	97.04	62.72	82.87	66.16
M12	✓	✓	✓	✓	✓	**99.52**	**98.69**	**99.52**	**97.67**	**89.56**	**97.67**	**99.54**	**98.72**	**99.54**	**99.74**	**99.26**	**99.74**	**76.81**	**89.77**	**78.63**

4.4 Ablation Experiments

Table 2 reports the ablation study evaluating the components of our ORCA framework. The baseline model relying solely on reconstruction loss (C_{rec}) exhibits limited clustering capability (e.g., 62.76% ACC on Fashion). However, independently introducing our core modules—orthogonal residual semantic discovery (C_{orsd}), consensus-calibrated hierarchical contrastive alignment (C_{ccha}), and divergence-driven recurrent fusion (C_H)—yields substantial improvements. For instance, incorporating C_{orsd} alone dramatically increases Fashion ACC to 96.66%. These components demonstrate strong complementarity when combined. By further integrating the consensus-anchored calibration weighting (C_W) to adaptively regulate hierarchical representations, the complete ORCA framework systematically prevents topological degeneration and achieves peak performance across all datasets (Fashion: 99.52%, Synthetic3D: 97.67%, Cifar10: 99.54%, MNIST-USPS: 99.74%, 100leaves: 76.81%). This confirms the synergistic effectiveness of our joint optimization strategy.

5 Conclusion

In this paper, we propose the ORCA framework to address inter-view feature redundancy and spectral degeneration in multi-view clustering. ORCA relies on two coordinated modules: SIRF, which leverages geometric projection and adaptive halting to eliminate redundancy and prevent dimensionality collapse; and CCHA, which aligns multi-scale features against a global consensus anchor to preserve highly discriminative semantics. Extensive experiments across five benchmarks demonstrate ORCA's consistent superiority over state-of-the-art methods. However, a limitation of this work is its reliance on fully aligned datasets and the lack of evaluation on large-scale real-world multi-view data, such as cross-modal industrial scenarios or web-collected data. Therefore, our future work will focus on extending ORCA to handle such complex, large-scale online clustering scenarios.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China (62506151).

References

[1] Jintang Bian, Xiaohua Xie, Jian-Huang Lai, and Feiping Nie. Multi-view contrastive clustering via integrating graph aggregation and confidence enhancement. *Information Fusion*, 108: 102393, 2024.

[2] Jintang Bian, Yixiang Lin, Xiaohua Xie, Chang-Dong Wang, Lingxiao Yang, Jian-Huang Lai, and Feiping Nie. Multilevel contrastive multiview clustering with dual self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[3] Rahul Chauhan, Kamal Kumar Ghanshala, and RC Joshi. Convolutional neural network (cnn) for image detection and recognition. In *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pages 278–282. IEEE, 2018.

[4] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16752–16761, 2023.

[5] Zhe Chen, Xiao-Jun Wu, Tianyang Xu, Hui Li, and Josef Kittler. Multi-layer multi-level comprehensive learning for deep multi-view clustering. *Information Fusion*, 116:102785, 2025.

[6] Xiaojian Ding, Xian Li, Lin Zhao, and Xiaoying Zhu. Hclaa: Hierarchical contrastive learning with adaptive attention. *Information Sciences*, 741:123249, 2026.

[7] Zhibin Dong, Shengju Yu, Siwei Wang, Xinwang Liu, and En Zhu. Cross-view graph matching for unsupervised learning with unaligned multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2026.

[8] Lunke Fei, Junlin He, Qi Zhu, Shuping Zhao, Jie Wen, and Yong Xu. Deep multi-view contrastive clustering via graph structure awareness. *IEEE Transactions on Image Processing*, 2025.

[9] Lele Fu, Sheng Huang, Lei Zhang, Jinghua Yang, Zibin Zheng, Chuanfu Zhang, and Chuan Chen. Subspace-contrastive multi-view clustering. *ACM Transactions on Knowledge Discovery from Data*, 18(9):1–35, 2024.

[10] Shizhe Hu, Binyan Tian, Weibo Liu, and Yangdong Ye. Self-supervised trusted contrastive multi-view clustering with uncertainty refined. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17305–17313, 2025.

[11] Hengrong Ju, Yang Lu, Weiping Ding, Wei Zhang, and Xibei Yang. Multigranularity information fused contrastive learning with multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[12] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. *Advances in neural information processing systems*, 24, 2011.

[13] Guojie Li, Zhiwen Yu, Kaixiang Yang, Jianming Lv, and CL Philip Chen. Pseudo-label similarity graph-driven multi-view contrastive clustering. *IEEE Transactions on Multimedia*, 2026.

[14] Pengyuan Li, Dongxia Chang, Yiming Wang, Man Liu, Zisen Kong, Linhua Kong, and Yao Zhao. Deep multi-view clustering with intra-view similarity and cross-view correlation learning. *IEEE Transactions on Multimedia*, 2026.

[15] Quanjiang Li, Tingjin Luo, Mingdie Jiang, Jiahui Liao, and Zhangqi Jiang. Deep incomplete multi-view network semi-supervised multi-label learning with unbiased loss. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9048–9056, 2024.

[16] Quanjiang Li, Tingjin Luo, Mingdie Jiang, Zhangqi Jiang, Chenping Hou, and Feijiang Li. Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18430–18438, 2025.

[17] Quanjiang Li, Tingjin Luo, and Jiahui Liao. Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20706–20715, 2025.

[18] Quanjiang Li, Zhiming Liu, Wei Luo, Tingjin Luo, and Chenping Hou. Correcting visual blur induced by attention distraction to reduce hallucinations: Algorithm and theory. *arXiv preprint arXiv:2605.24602*, 2026.

[19] Quanjiang Li, Tianxiang Xu, Tingjin Luo, Yan Zhong, Yang Li, Yiyun Zhou, and Chenping Hou. Theory-driven label-specific representation for incomplete multi-view multi-label learning. *Advances in Neural Information Processing Systems*, 38:121996–122037, 2026.

[20] Zhenglai Li, Chang Tang, Xinwang Liu, Xiao Zheng, Wei Zhang, and En Zhu. Consensus graph learning for multi-view clustering. *IEEE Transactions on Multimedia*, 24:2461–2472, 2021.

[21] Shuaiyu Liu, Song Wu, Jie Xu, Yazhou Ren, Yang Yang, Xiaorong Pu, and Guoying Wang. Views attention fusion of granular-ball fuzzy representations split for improved multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23810–23818, 2026.

[22] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4697–4705, 2024.

[23] Jiongzhi Qiu, Yixuan Ye, Liang Peng, Dan Yang, Jiajun Xian, Man-Fai Leung, Hangjun Che, and Cheng Liu. Self-supervised semantic graph propagation for multi-view clustering. *Neural Networks*, page 108973, 2026.

[24] Yazhou Ren, Zichen Wen, Junlong Ke, Chenhang Cui, Yonghao Huang, Xinyue Chen, Philip S Yu, and Lifang He. Enhancing multi-view clustering: A sufficient information-theoretic approach for consistency acquisition and redundancy elimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[25] Padmaja Savaram, Sasidhar Bandu, et al. Comparative analysis of deep learning algorithms for handwritten digit recognition. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, pages 1–5. IEEE, 2024.

[26] Bocheng Wang, Jiawei Wang, Chusheng Zeng, and Mulin Chen. Locality-driven flexible consensus graph learning for multi-view clustering. *Pattern Recognition*, 169:111853, 2026.

[27] Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Transactions on Multimedia*, 2024.

[28] Yanwanyu Xi, Chang Tang, Jun-Jie Huang, Xingchen Hu, Yuanyuan Liu, and Xinwang Liu. Contrastive and dual adversarial representation learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[29] Baili Xiao, Ke Liang, Jiaqi Jin, Jun Wang, Yinbo Xu, Siwei Wang, and En Zhu. Dmcar: Disentangled mixture-of-experts with context-aware routing for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 27055–27063, 2026.

[30] Huilang Xu, Xiang-Xiang Su, Simin Chen, Guang-Yong Chen, and Xing Chen. Bridging optimization and neural networks for efficient multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 16066–16074, 2026.

[31] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9234–9243, 2021.

[32] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16051–16060, 2022.

[33] Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Hengtao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. *Advances in neural information processing systems*, 36:1119–1131, 2023.

[34] Junpeng Xu, Min Meng, Jigang Liu, and Jigang Wu. Deep multi-view clustering with diverse and discriminative feature learning. *Pattern Recognition*, 161:111322, 2025.

[35] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19863–19872, 2023.

[36] Weiqing Yan, Tingyu Yang, and Chang Tang. Self-supervised semantic soft label learning network for deep multi-view clustering. *IEEE Transactions on Multimedia*, 2025.

[37] Xihong Yang, Jin Jiaqi, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM international conference on multimedia*, pages 337–346, 2023.

[38] Zengbiao Yang and Yihua Tan. Deep multi-view clustering based on instance-level adaptive structural contrastive learning. *Neural Networks*, page 108581, 2026.

[39] Yongbo Yu, Zhoumin Lu, Feiping Nie, Weizhong Yu, Zongcheng Miao, and Xuelong Li. Pseudo-label guided bidirectional discriminative deep multi-view subspace clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[40] Hanning Yuan, Zhihui Zhang, Qi Guo, Lianhua Chi, Sijie Ruan, Wei Zhou, Jinhui Pang, and Xiaoshuai Hao. Dwcl: Dual-weighted contrastive learning for robust multi-view clustering. *Engineering Applications of Artificial Intelligence*, 165:113532, 2026.

[41] Fenghua Zhang and Hangjun Che. Separable consistency and diversity feature learning for multi-view clustering. *IEEE Signal Processing Letters*, 2024.

[42] Qian Zhang, Lin Zhang, Ran Song, Runmin Cong, Yonghuai Liu, and Wei Zhang. Learning common semantics via optimal transport for contrastive multi-view clustering. *IEEE Transactions on Image Processing*, 2024.

[43] Bowen Zhao, Qianqian Wang, Zhiqiang Tao, Wei Feng, and Quanxue Gao. Dfmvc: Deep fair multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8090–8099, 2024.

[44] Yiyang Zhou, Qinghai Zheng, Yifei Wang, Wenbiao Yan, Pengcheng Shi, and Jihua Zhu. Mcoco: Multi-level consistency collaborative multi-view clustering. *Expert Systems with Applications*, 238:121976, 2024.

[45] Xin Zou, Ruimeng Liu, Chang Tang, Zhenglai Li, Xinwang Liu, Kunlun He, and Wanqing Li. Learning disentangled representations for generalized multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

A Supplementary Experimental Details

In this supplementary section, we provide additional implementation details and experimental analyses to support the main manuscript. Specifically, we first outline the complete execution flow of the ORCA framework, detailing the systematic transition from semantic discovery to hierarchical alignment. Second, we provide qualitative visualizations of the clustering results to demonstrate the dynamic convergence of our latent representations. Finally, we conduct a parameter sensitivity analysis to evaluate the robustness of our framework under varying contrastive temperature coefficients.

A.1 Complete Training Algorithm

The overall training procedure of the proposed ORCA framework is summarized in Algorithm 1. The model first performs semantic incremental recurrent fusion to obtain the global semantic consensus, followed by consensus-calibrated hierarchical contrastive alignment for multi-level representation enhancement. The resulting objective is jointly optimized through end-to-end backpropagation.

Algorithm 1 Training Procedure of the Proposed ORCA Framework

Require: Multi-view data $\{\mathbf{X}_v\}_{v=1}^{V}$, cluster number K, training epochs E, information gain threshold ϵ
Ensure: Global semantic consensus $\mathbf{H}$ and clustering assignments
for $epoch = 1$ **to** E **do**
  Extract view-specific latent embeddings $\{\mathbf{z}_v\}_{v=1}^{V}$
  Semantic Incremental Recurrent Fusion (SIRF):
  Obtain enhanced initial representations $\mathbf{Z}^{(0)}$ via orthogonal residual semantic discovery
  Iteratively update $\mathbf{Z}^{(t)}$, compute information gain $\Delta\mathcal{I}^{(t)}$, and adaptively determine retention weight α_t
  Aggregate the retained states to obtain global consensus $\mathbf{H}$
  Consensus-Calibrated Hierarchical Contrastive Alignment (CCHA):
  Construct the hierarchical candidate pool $\mathcal{Z}_{\text{pool}}$
  Project candidates to representations $\mathcal{R} = \{\mathbf{R}_m\}_{m=1}^{M}$
  Compute consensus-guided weights and $\mathcal{L}_{\text{ccha}}$ according to Eqs. (15)–(19)
  Update all network parameters by minimizing $\mathcal{L}_{\text{total}}$ in Eq. (21)
end for
Apply k-means to the optimized global consensus $\mathbf{H}$
return $\mathbf{H}$ and the corresponding cluster labels

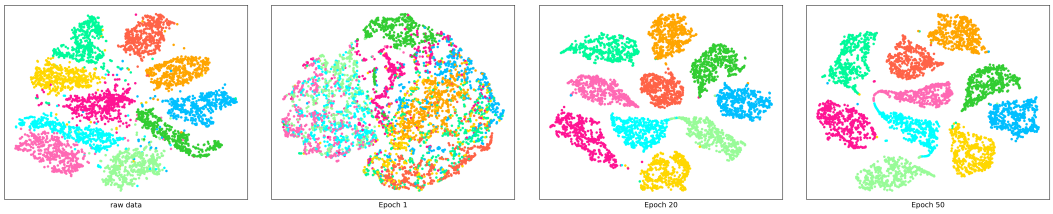
A.2 Clustering Results Visualization Analysis:

To intuitively understand the representation learning process of ORCA, we map the high-dimensional latent representations into a 2D space using t-SNE. Figure 4 visualizes the clustering trajectory on the MNIST-USPS dataset. Initially, the raw data presents heavily overlapping and poorly separated distributions. As training progresses, ORCA progressively distills the global consensus and eliminates view-specific noise. By Epoch 50, the model yields a highly discriminative latent space characterized by compact intra-cluster structures and distinct inter-cluster margins, validating the efficacy of the proposed hierarchical alignment.

A.3 Parameter Sensitivity Analysis

Sensitivity to Temperature Coefficient τ: Figure 5 illustrates ORCA's sensitivity to the temperature coefficient τ within the InfoNCE objective, which regulates the sharpness of the similarity distribution between the global consensus anchor $\mathbf{H}$ and the multi-scale intermediate representations $\mathbf{R}_m$.

A very small τ (e.g., 0.1) tends to degrade clustering performance. This is because it over-compresses the representation space and excessively focuses on penalizing hard negative samples, which can disrupt global semantic consistency. Conversely, moderate values ($\tau \in [0.5, 1.0]$) consistently yield optimal results across diverse datasets by effectively balancing local feature alignment and global representation uniformity. Although different datasets (e.g., Fashion and MNIST-USPS vs. Synthetic3d and Cifar10) exhibit slightly varying sensitivities to small τ values, a well-tuned τ



(a) Raw Data (b) Epoch 1 (c) Epoch 20 (d) Epoch 50

Figure 4: Clustering visualization (t-SNE) of the MNIST-USPS dataset at different training stages: raw data, and epochs 1, 20, and 50. ORCA progressively establishes clear clustering boundaries.

universally enables the consensus-anchored weighting mechanism to accurately prioritize informative hierarchical scales without metric distortion.

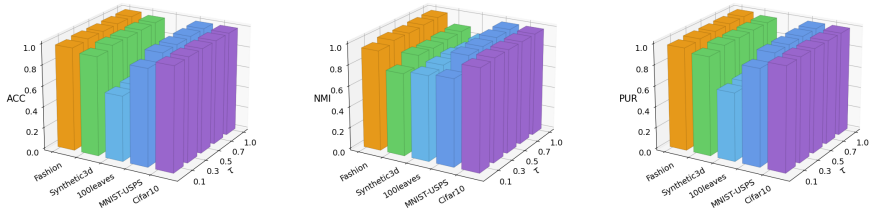


Figure 5: Analysis of clustering performance under varying temperature coefficients (τ) evaluated via ACC, NMI, and PUR metrics across different datasets.

B Theoretical Analysis

In this section, we establish a rigorous mathematical foundation for the ORCA framework, analyzing its properties through the lenses of redundancy mitigation, topological stability, and informational sufficiency. We first characterize the geometric properties of the orthogonal residual semantic discovery, proving that explicitly stripping the parallel projection minimizes feature redundancy and strictly bounds the condition number of the latent covariance, thereby guaranteeing an expansion of the informative subspace. Second, we formalize the attention-based recurrent fusion as a discrete dynamical system. We provide theoretical bounds demonstrating that without intervention, deep cross-view attention inevitably suffers from spectral degeneration. By incorporating our Adaptive Divergence-Driven Halting mechanism based on instantaneous information gain, we prove that the system rigorously avoids catastrophic dimensionality collapse, ensuring that the synthesized global consensus $\mathbf{H}$ retains maximal discriminative margin for downstream clustering.

B.1 Information Maximization via Orthogonal Residual Projection

A fundamental challenge in multi-view fusion is the overwhelming accumulation of redundant semantics, which distorts the latent clustering structure. We mathematically demonstrate that direct aggregation amplifies the redundant principal components, whereas our orthogonal residual injection strictly enhances the novel semantic subspace without exacerbating redundancy.

Theorem 1 (Subspace Capacity and Redundancy Mitigation)**.** *Let* $\mathbf{\Sigma}_{\text{core}} = \mathbb{E}[\mathbf{Z}_{\text{core}}^{(v)}(\mathbf{Z}_{\text{core}}^{(v)})^\top]$ *denote the covariance matrix of the core view, and let* $\mathbf{Z}_{\text{other}}^{(u)}$ *be a marginal view. Define the augmented representation via direct addition as* $\mathbf{Z}_{\text{direct}} = \mathbf{Z}_{\text{core}}^{(v)} + \mathbf{Z}_{\text{other}}^{(u)}$*, and via ORSD's orthogonal injection as* $\hat{\mathbf{Z}}^{(v)} = \mathbf{Z}_{\text{core}}^{(v)} + g^{(u\to v)}\mathbf{Z}_{\perp}^{(u\to v)}$*. The condition number of the feature covariance matrix for the ORSD formulation is strictly bounded below that of direct aggregation:*

$$\kappa(\mathbf{\Sigma}_{\text{ORSD}}) < \kappa(\mathbf{\Sigma}_{\text{direct}}), \tag{23}$$

improving the spectral stability of the latent space. Furthermore, the mutual information redundancy is minimized:

$$\mathcal{I}(\mathbf{Z}_{\text{core}}^{(v)}; g^{(u\to v)}\mathbf{Z}_{\perp}^{(u\to v)}) = 0, \tag{24}$$

ensuring optimal information gain.

Proof. Without loss of generality, let the latent features be zero-centered. We geometrically decompose the marginal view into its parallel and orthogonal components relative to the core view:

$$\mathbf{Z}_{\text{other}}^{(u)} = \gamma\mathbf{Z}_{\text{core}}^{(v)} + \mathbf{Z}_{\perp}^{(u\to v)}, \quad \text{where} \quad \gamma = \frac{\langle \mathbf{Z}_{\text{other}}^{(u)}, \mathbf{Z}_{\text{core}}^{(v)} \rangle}{\|\mathbf{Z}_{\text{core}}^{(v)}\|_2^2}. \tag{25}$$

Here, γ represents the structural correlation coefficient. Under standard direct aggregation, the updated representation becomes $\mathbf{Z}_{\text{direct}} = (1+\gamma)\mathbf{Z}_{\text{core}}^{(v)} + \mathbf{Z}_{\perp}^{(u\to v)}$. Given that $\mathbf{Z}_{\text{core}}^{(v)}$ and $\mathbf{Z}_{\perp}^{(u\to v)}$ are strictly orthogonal ($\mathbb{E}[\mathbf{Z}_{\text{core}}^{(v)}(\mathbf{Z}_{\perp}^{(u\to v)})^\top] = \mathbf{0}$), the covariance matrix of the direct aggregation expands to:

$$\mathbf{\Sigma}_{\text{direct}} = (1+\gamma)^2\mathbf{\Sigma}_{\text{core}} + \mathbf{\Sigma}_{\perp}. \tag{26}$$

The maximum eigenvalue of $\mathbf{\Sigma}_{\text{direct}}$ is bounded by $\lambda_{\max}(\mathbf{\Sigma}_{\text{direct}}) \geq (1+\gamma)^2\lambda_{\max}(\mathbf{\Sigma}_{\text{core}})$. This implies that the variance along the principal direction of the core view is amplified by a factor of $(1+\gamma)^2$, severely inflating the condition number $\kappa(\mathbf{\Sigma}_{\text{direct}}) = \frac{\lambda_{\max}(\mathbf{\Sigma}_{\text{direct}})}{\lambda_{\min}(\mathbf{\Sigma}_{\text{direct}})}$ and causing redundant features to overwhelmingly dominate the latent manifold.

In contrast, ORSD explicitly isolates the orthogonal residual to compute $\hat{\mathbf{Z}}^{(v)} = \mathbf{Z}_{\text{core}}^{(v)} + g^{(u\to v)}\mathbf{Z}_{\perp}^{(u\to v)}$. Its corresponding covariance matrix is strictly block-diagonalized as:

$$\mathbf{\Sigma}_{\text{ORSD}} = \mathbf{\Sigma}_{\text{core}} + (g^{(u\to v)})^2\mathbf{\Sigma}_{\perp}. \tag{27}$$

Because the variance along the core view's principal axes remains unchanged while the eigenspectrum grows only in the novel orthogonal subspace, we mathematically guarantee that $\kappa(\mathbf{\Sigma}_{\text{ORSD}}) < \kappa(\mathbf{\Sigma}_{\text{direct}})$, preventing spectral dominance.

Furthermore, from an information-theoretic perspective, assuming the latent representations follow a multivariate Gaussian distribution, the mutual information is defined via the differential entropy $\mathcal{H}(\cdot)$:

$$\mathcal{I}(\mathbf{Z}_{\text{core}}^{(v)}; \mathbf{Z}_{\perp}^{(u\to v)}) = \mathcal{H}(\mathbf{Z}_{\text{core}}^{(v)}) + \mathcal{H}(\mathbf{Z}_{\perp}^{(u\to v)}) - \mathcal{H}(\mathbf{Z}_{\text{core}}^{(v)}, \mathbf{Z}_{\perp}^{(u\to v)}). \tag{28}$$

Since $\mathbf{Z}_{\text{core}}^{(v)}$ and $\mathbf{Z}_{\perp}^{(u\to v)}$ are uncorrelated, their joint entropy equals the sum of their marginal entropies, yielding $\mathcal{I}(\mathbf{Z}_{\text{core}}^{(v)}; \mathbf{Z}_{\perp}^{(u\to v)}) = 0$. This confirms that $\hat{\mathbf{Z}}^{(v)}$ spans a maximally informative subspace free of redundant overlap. □

B.2 Theoretical Bounds on Spectral Degeneration via Information-Theoretic Halting

In the context of multi-view representation learning, iterative feature fusion mechanisms inherently act as low-pass filters on the latent graph, often homogenizing discriminative cross-view semantics. To mathematically substantiate the design of the SIRF module, we first prove that unconstrained recurrent interactions inevitably drive the multi-view features toward catastrophic dimensionality collapse (i.e., over-smoothing). Subsequently, we establish that our KL-divergence guided soft-halting mechanism rigorously averts this trivial equilibrium, guaranteeing a strictly positive lower bound on the structural variance to preserve optimal clustering separability.

Theorem 2 (Topological Stability against Dimensionality Collapse)**.** *Let $\mathbf{Z}^{(t)} \in \mathbb{R}^{N\times d}$ be the feature matrix at iteration t, and let $\mathbf{L} = \mathbf{D} - \mathbf{A}$ be the graph Laplacian. The Dirichlet energy measuring latent space discriminability is defined as:*

$$\mathcal{E}(\mathbf{Z}^{(t)}) = \frac{1}{2N}\sum_{i,j} A_{ij}\|\mathbf{z}_i^{(t)} - \mathbf{z}_j^{(t)}\|_2^2 = \frac{1}{N}\mathrm{Tr}\left((\mathbf{Z}^{(t)})^\top\mathbf{L}\mathbf{Z}^{(t)}\right). \tag{29}$$

Under unconstrained attention-based recurrent fusion, $\lim_{t\to\infty}\mathcal{E}(\mathbf{Z}^{(t)}) = 0$. By incorporating the dynamic soft-halting gate $\alpha_t \propto \sigma(\Delta\mathcal{I}^{(t)})$, the synthesized final representation $\mathbf{Z}_{\text{final}}$ satisfies:

$$\mathcal{E}(\mathbf{Z}_{\text{final}}) \geq \epsilon > 0, \tag{30}$$

thereby successfully escaping spectral degeneration and preserving clustering boundaries.

Proof. We formulate the cross-view attention update as a Markovian graph diffusion process. Let $\mathbf{P}^{(t)}$ denote the row-stochastic attention matrix derived from the softmax operator. The core interaction is simplified as $\mathbf{Z}^{(t)} = \mathbf{P}^{(t)}\mathbf{Z}^{(t-1)}\mathbf{W}_V$.

By the Rayleigh quotient and spectral graph theory, the propagation through the row-stochastic transition matrix inherently smooths the features. The Dirichlet energy satisfies the recursive inequality:

$$\mathcal{E}(\mathbf{Z}^{(t)}) \leq (1 - \lambda_2(\mathbf{L}_{\text{rw}}))^2 \mathcal{E}(\mathbf{Z}^{(t-1)}), \tag{31}$$

where $\lambda_2(\mathbf{L}_{\text{rw}}) \in (0, 1]$ is the Fiedler eigenvalue of the random walk Laplacian. Expanding this inequality over t iterations yields:

$$\mathcal{E}(\mathbf{Z}^{(t)}) \leq (1 - \lambda_2)^{2t} \mathcal{E}(\mathbf{Z}^{(0)}). \tag{32}$$

As $t \to \infty$, the features exponentially converge to a stationary distribution (an inseparable global consensus), driving $\mathcal{E}(\mathbf{Z}^{(t)}) \to 0$. This represents a catastrophic dimensionality collapse.

To counter this, our proposed framework computes the instantaneous information gain via KL-divergence:

$$\Delta\mathcal{I}^{(t)} = \int_{\mathbf{Z}} \mathcal{P}(\mathbf{Z}^{(t)}) \log \frac{\mathcal{P}(\mathbf{Z}^{(t)})}{\mathcal{P}(\mathbf{Z}^{(t-1)})} d\mathbf{Z}. \tag{33}$$

As the system approaches the collapsed equilibrium, the distribution shift vanishes ($\Delta\mathcal{I}^{(t)} \to 0$). The adaptive gating function $\alpha_t = \sigma(\mathbf{W}_\alpha \Delta\mathcal{I}^{(t)} + \mathbf{b}_\alpha)$ forces $\alpha_t \to 0$ for saturated states.

The final robust representation is synthesized as $\mathbf{Z}_{\text{final}} = \sum_{t=1}^{T_{\max}} \alpha_t \mathbf{Z}^{(t)}$. Due to the convexity of the Dirichlet energy (acting as a quadratic form), we can lower-bound the energy of the final state:

$$\mathcal{E}(\mathbf{Z}_{\text{final}}) = \frac{1}{N} \text{Tr} \left(\left(\sum_{t=1}^{T_{\max}} \alpha_t \mathbf{Z}^{(t)} \right)^\top \mathbf{L} \left(\sum_{t=1}^{T_{\max}} \alpha_t \mathbf{Z}^{(t)} \right) \right) \geq \sum_{t=1}^{T_{\max}} \alpha_t^2 \mathcal{E}(\mathbf{Z}^{(t)}). \tag{34}$$

Because the retention weights α_t dynamically isolate and heavily weight the early, high-energy states (where $\Delta\mathcal{I}^{(t)} \gg 0$) while strictly discarding the late collapsed states ($\alpha_t \approx 0$), the accumulated energy remains strictly positive ($\sum \alpha_t^2 \mathcal{E}(\mathbf{Z}^{(t)}) \geq \epsilon$). Thus, SIRF autonomously halts the interaction, guaranteeing that the final consensus avoids structural collapse. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly list our main contributions at the end of Section 1 (Introduction) and carefully define the scope and assumptions of our method.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the limitations of our proposed ORCA framework, specifically its current restriction to fully aligned datasets and offline scenarios, in the Conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer:[Yes]

Justification: All theoretical assumptions are clearly stated in Section X, and the complete proofs for our theorems are provided in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:[Yes]

   Justification: The paper provides sufficient details on the algorithm, datasets, and experimental setup to support reproducibility of the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will upload the code package as supplementary material, and the latest GitHub link will be provided upon paper acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides the details of data preprocessing and hyperparameter settings in the Experiments section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Most compared baseline methods do not include statistical significance in their experiments; therefore, we follow the same practice and do not report it in our work.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resources is stated in the Experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:This paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The proposed algorithm has no direct societal impact. All datasets used in this paper are publicly available, and the algorithm is solely focused on producing clustering assignments.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: All datasets used in this paper are publicly available, so no specific safeguards are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Original papers and datasets are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets are introduced in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects were involved in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve any research with human subjects or crowdsourcing, and therefore does not require IRB or equivalent ethical approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer:[N/A]

    Justification:The core methodology does not involve the use of large language models (LLMs).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.