RETHINKING VIEW CONSISTENCY: CONSENSUS-ORIENTED REPRESENTATION LEARNING FOR MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering aims to integrate heterogeneous observations from multiple views to learn a discriminative consensus representation for clustering. Existing methods often place cross-view consistency at the center of representation learning, either by directly aligning different views or by using a global target to constrain individual-view representations. While effective when clustering-relevant information is redundantly distributed across views, such consistency-oriented strategies can become restrictive when discriminative evidence is available only in a subset of views. Enforcing cross-view consistency may weaken view-specific informative structures, while global-to-local constraints may force individual views to reproduce semantics beyond their own observational scope. In this work, we rethink the role of view consistency and propose a consensus-oriented representation learning method for multi-view clustering. Instead of explicitly driving individual views toward cross-view consistency, we preserve their intrinsic relational structures and allow them to contribute differently to a jointly learned consensus. Direct clustering supervision is imposed only on the consensus representation, allowing each view to be optimized according to its contribution to the global clustering structure rather than being forced to match the complete global target. Extensive experiments on multiple benchmark datasets demonstrate that the proposed method effectively preserves view relations and yields more discriminative clustering results. Code will be available upon acceptance.

1 INTRODUCTION

Real-world objects are often described from multiple perspectives, with each view providing a partial observation from a distinct source or feature space Xu et al. (2015); Zhao et al. (2017). Multi-view clustering (MVC) Fang et al. (2023) aims to integrate these views to uncover latent data partitions without labels. Since each view captures only part of the object, clustering-relevant evidence may be observable only from a subset of views Baltrušaitis et al. (2018); Li et al. (2018). Combining multiple views can therefore reveal evidence that cannot be fully observed from any single view. How to combine such evidence into a reliable clustering consensus is a central problem in MVC.

In the absence of labels, explicit semantic supervision for learning a clustering consensus is difficult to obtain. However, the correspondence of the same sample across views provides a natural source of supervision Trosten et al. (2021). Existing deep MVC methods can be broadly grouped into two recurring strategies. The first strategy directly constrains different views and learns consistent structures through cross-view contrast, representation alignment, or clustering-assignment matching Xu et al. (2022b); Chen et al. (2023); Xu et al. (2023); Chen et al. (2026a). The second strategy first combines multiple views into a shared representation, unified assignment, or other global clustering target, which is then used as a consensus proxy to supervise the representation or clustering learning of individual views Xu et al. (2022a); Yan et al. (2023); Liu et al. (2025); Li et al. (2026). Although implemented differently, both strategies promote cross-view consistency: the former through direct constraints between views, and the latter by propagating a global target back to local views. In both cases, the resulting consistent or global structure is used to support the final clustering decision.

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.

However, as shown in Fig. 1, when some clustering-relevant evidence is present only in individual views, both types of consistency-oriented methods may face limitations. First, cross-view consistency constraints may weaken discriminative evidence available only in a subset of views, preventing it from being fully utilized in the final representation. Second, even if a correct global target is obtained by integrating observations from multiple views, directly using this target to supervise all views may require local representations to fit information beyond their own observations.

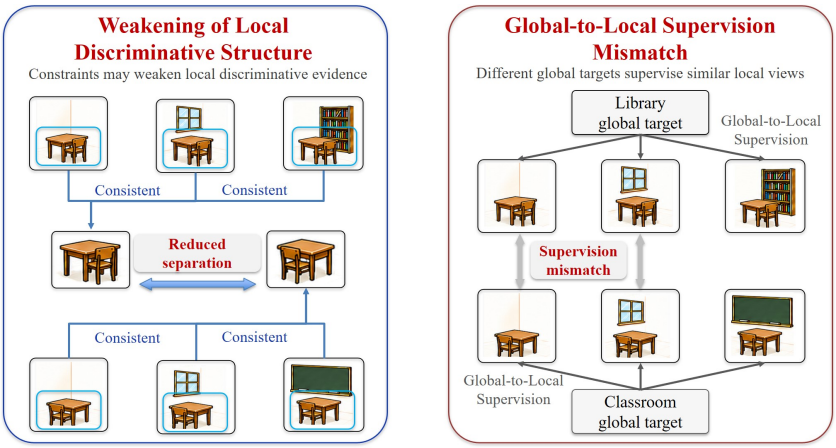


Figure 1: **Left:** The library and classroom appear similar in the first two views, while discriminative evidence is only observable in the third view, such as the bookshelf and blackboard. Strengthening cross-view consistency can weaken discriminative evidence that is available only in a subset of views. **Right:** Even when multi-view information yields a global target that correctly distinguishes the library from the classroom, the first two local views still do not contain such discriminative evidence. Directly using the global target to supervise all local views can force them to fit information beyond their own observations.

Motivated by these observations, we rethink the learning process of view representations and consensus representations in multi-view clustering and propose a consensus-oriented representation learning framework. Specifically, at the view level, instead of explicitly enforcing different views to be consistent, we preserve the local relational structure of each view during representation learning to avoid excessive changes to local structures. At the consensus level, we use self-attention to integrate representations from different views and apply clustering supervision only to the resulting consensus representation. By integrating complementary evidence from different views, the framework can effectively exploit clustering-relevant information that is observable only from a subset of views. By applying clustering supervision only to the consensus representation, it avoids the need for individual views to reproduce the full clustering semantics formed through multi-view integration. Our main contributions are summarized as follows:

- We rethink the relationship between consistency learning and consensus learning, analyze when consistency learning can approximate consensus learning, and reveal its limitations.
- We propose a consensus-oriented representation learning framework for multi-view clustering, which learns representations toward clustering consensus rather than relying on explicit consistency constraints as the primary objective.
- Extensive experiments on benchmark datasets demonstrate the effectiveness of the proposed method, with further empirical analyses and ablation studies validating the roles of the proposed designs.

2 Related Work

Existing consistency-oriented multi-view clustering methods can be broadly divided into two categories. The first directly enforces consistency across views to learn consistent multi-view representations, while the second constructs global targets to further guide view-level learning.

In multi-view clustering without label supervision, the natural correspondence of the same sample across views provides a direct supervisory signal for representation learning. MFLVC Xu et al. (2022b) extracts shared semantics through cross-view contrastive learning, while CVCL Chen et al. (2023) further provides systematic modeling and analysis of consistency among soft cluster assignments. As the field progressed, the focus gradually shifted from simply strengthening consistency to imposing consistency constraints in a more appropriate manner. SEM Xu et al. (2023) adaptively adjusts the contrastive strength according to representation discrepancies between view pairs, thereby alleviating representation degeneration caused by unreliable view pairs; STCMC-UR Hu et al. (2025) estimates uncertainty to select more reliable clustering supervision and improve the reliability of contrastive learning. MGBCC Su et al. (2025) introduces granular balls to extend consistency learning to an intermediate granularity between instances and clusters; RAV Peng et al. (2026) further aligns different views at the level of sample relationships and adjusts semantic alignment strength according to view similarity. MANGO Chen et al. (2026a) instead transfers the consistency constraint to a projected shared structural space, alleviating the restriction imposed by direct alignment on heterogeneous view representations. These methods continuously calibrate cross-view consistency in terms of its strength and the space in which it is imposed, demonstrating that consistent information can provide stable and effective discriminative cues for unsupervised clustering. However, when clustering-relevant evidence is available only in a subset of views, relying excessively on cross-view consistency may weaken such discriminative evidence.

Beyond directly constraining different views, another line of work constructs global representations or clustering targets from jointly available multi-view information and further uses them to guide individual-view learning. SDMVC Xu et al. (2022a) is a representative example: it constructs a unified target distribution from the joint multi-view representation and uses the resulting global discriminative information to guide the learning of each view. Subsequent studies mainly improve this framework along two dimensions: the quality of global information and the way it is fed back to local views. GCFAgg Yan et al. (2023) enhances global representations through cross-sample and cross-view aggregation and constrains view representations using structure-guided contrastive learning; SCMVC Wu et al. (2024) adjusts the contribution of each view during fusion according to the discrepancy between the view representation and the global representation; SparseMVC Liu et al. (2025) further uses the correlation between global and single-view features to adaptively determine the contribution of each view for different samples. Meanwhile, related work has also begun to regulate global feedback according to local predictions and view reliability. SGDMC Huang et al. (2023) reduces the contribution of unreliable samples during training according to the discrepancy between local cluster assignments and global pseudo-labels; AWDC-MVC Wang et al. (2026) employs self-attention to model inter-view relationships and estimate view importance, thereby adjusting the degree of global-to-local alignment at both the feature and decision levels; DRCFMVC Li et al. (2026) combines single-view predictive uncertainty with cross-view conflict to estimate reliability and regulate the contrastive fusion process. These methods continuously improve the quality of global information and its interaction with local views through mechanisms such as view contribution estimation, sample reliability, and alignment strength. However, even when the global consensus correctly captures the clustering semantics, directly using the global target to supervise local views may require local representations to fit information beyond their own observations.

3 METHOD

Consider a multi-view dataset with V views, denoted as $\mathcal{X} = \{\mathbf{X}^v\}_{v=1}^{V}$, where $\mathbf{X}^v = [\mathbf{x}_1^v, \mathbf{x}_2^v, \ldots, \mathbf{x}_n^v]^\top \in \mathbb{R}^{n \times d_v}$. Here, n denotes the number of samples and d_v denotes the feature dimensionality of the v-th view. The goal of multi-view clustering is to partition the n samples into K clusters without label supervision. Since the feature dimensions differ across views in the raw data $\mathcal{X}$, view-specific encoders are widely used in deep multi-view clustering to map these heterogeneous views into a shared representation space:

$$\mathbf{Z}^v = f_v(\mathbf{X}^v; \theta_v) \in \mathbb{R}^{n \times d} \tag{1}$$

where θ_v denotes the parameters of the v-th encoder, $\mathbf{Z}^v = [\mathbf{z}_1^v, \mathbf{z}_2^v, \ldots, \mathbf{z}_n^v]^\top$ and $\mathbf{z}_i^v$ denotes the representation learned from view v.

3.1 RETHINKING CONSISTENCY AND CONSENSUS LEARNING

The two issues illustrated in Fig. 1 motivate a general analysis of the relationship between consistency learning and consensus learning. Their relationship depends on whether clustering-relevant information is redundant across views. For an arbitrary sample pair (i, j), we define the clustering-relevant information unavailable from view v but contained in the remaining views as:

$$\Gamma_{ij}^{v} = I\left(\mathbf{S}_{ij}; \mathbf{z}_i^{-v}, \mathbf{z}_j^{-v} \mid \mathbf{z}_i^{v}, \mathbf{z}_j^{v}\right) \tag{2}$$

where $I(\cdot;\cdot \mid \cdot)$ denotes conditional mutual information, $\mathbf{z}_i^{-v}$ denotes the representations of sample i from all views except v, and $\mathbf{S}_{ij}$ denotes the clustering relation between samples i and j. When $\Gamma_{ij}^{v} \approx 0$ for all views, clustering-relevant information is largely redundant across views. In contrast, $\Gamma_{ij}^{v} > 0$ indicates that view v alone does not fully characterize the multi-view clustering relation.

Existing consistency-oriented methods generally follow two paradigms. The first directly enforces consistency across views, where different views are encouraged to produce consistent representations or clustering structures. At the representation level, this can be expressed as:

$$\mathcal{L}_{\text{consis}} = \frac{1}{nV(V-1)} \sum_{i=1}^{n} \sum_{v=1}^{V} \sum_{u \neq v} D\left(\mathbf{z}_i^{v}, \mathbf{z}_i^{u}\right) \tag{3}$$

This formulation assumes that the structures consistently captured across views are sufficient to characterize the clustering structure. When $\Gamma_{ij}^{v} \approx 0$ for all views, each view already captures most of the clustering-relevant information available from the other views, and consistency learning can therefore provide a reasonable approximation to consensus learning. In contrast, when $\Gamma_{ij}^{v} > 0$ for a subset of views, these views contain incomplete clustering-relevant evidence, suggesting that enforcing representation consistency alone may overlook complementary information.

The second constructs a global target from multiple views and uses it to guide individual views. At the prediction level, it can be generally expressed as:

$$\mathcal{L}_{c \to v} = \frac{1}{nV} \sum_{i=1}^{n} \sum_{v=1}^{V} D\left(\mathbf{q}_i^{v}, \mathbf{q}_i^{c}\right) \tag{4}$$

where $\mathbf{q}_i^{v}$ and $\mathbf{q}_i^{c}$ denote the clustering predictions of sample i from the v-th view and the global target representation, respectively. This strategy implicitly assumes that each view can sufficiently capture the semantics encoded in the global target. For view v, when $\Gamma_{ij}^{v} \approx 0$, the remaining views provide little additional clustering-relevant information beyond that already available from view v, making the global target largely representable by this view. In contrast, when $\Gamma_{ij}^{v} > 0$, view v lacks certain clustering-relevant evidence provided by other views, and the global target may therefore contain information unavailable to view v, resulting in a global-to-local supervision mismatch. Detailed analysis of these two cases is provided in Appendix A.

In contrast to the above consistency-oriented paradigms, consensus learning should integrate the evidence from different views and determine the clustering structure directly at the consensus level:

$$\mathcal{L}_{\text{consen}} = \mathcal{L}_{\text{clu}}\left(\mathbf{C}\right) \tag{5}$$

where $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n]^{\top}$ denotes the consensus representations of all samples, and $\mathbf{c}_i = g(\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^V)$ encodes rich and complementary information from different views. $\mathcal{L}_{\text{clu}}$ evaluates the clustering structure over the consensus representations. Under this formulation, consensus learning does not rely on clustering-relevant information being redundant across views. For a given view v, whether $\Gamma_{ij}^{v} \approx 0$ or $\Gamma_{ij}^{v} > 0$, clustering-relevant evidence from different views can be integrated directly at the consensus level without requiring such evidence to recur in every view. Correspondingly, whether $\Gamma_{ij}^{v} \approx 0$ or $\Gamma_{ij}^{v} > 0$, view v is not directly required to reproduce the global consensus, but contributes its available evidence through the consensus representation $\mathbf{C}$.

This analysis clarifies the relationship between consistency and consensus learning. Consistency learning can provide a reasonable approximation to consensus learning when clustering-relevant information is largely redundant across views. Otherwise, consistency learning may not fully capture the clustering structure supported by complementary evidence. This leads to two principles for consensus-oriented representation learning: (1) preserve each view's local relational structures during representation learning; and (2) apply direct clustering supervision only at the consensus level. Guided by these principles, our framework consists of view-wise relation preservation, sample-wise consensus formation, and consensus-only clustering supervision, as illustrated in Fig. 2.

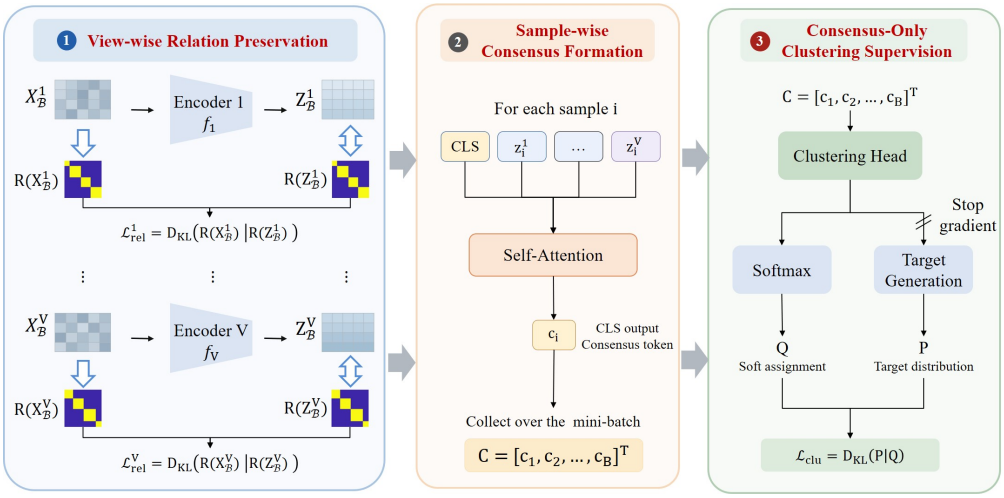


Figure 2: Overview of the proposed consensus-oriented representation learning framework. The framework first preserves view-wise relational structures during encoding and then forms sample-wise consensus representations by integrating multi-view observations. Clustering supervision is subsequently applied only at the consensus level.

3.2 View-Wise Relational Preservation

Since consensus is formed by integrating evidence from multiple views, preserving the local structures captured by each view before integration is important. Therefore, we introduce a view-wise relation preservation objective during representation learning. For a mini-batch $\mathcal{B}$ containing B samples, the v-th view $\mathbf{X}_{\mathcal{B}}^{v}$ is mapped by its view-specific encoder to $\mathbf{Z}_{\mathcal{B}}^{v} = f_v(\mathbf{X}_{\mathcal{B}}^{v}; \theta_v)$, where all views are mapped into the same latent space. To retain the local relations within each view during this transformation, we constrain the relational structures before and after encoding as $R(\mathbf{X}_{\mathcal{B}}^{v}) \approx R(\mathbf{Z}_{\mathcal{B}}^{v})$, where $R(\cdot)$ denotes a normalized relation graph among samples in the mini-batch.

Local neighborhoods provide a concise description of the geometric relations among samples. We therefore use a neighborhood graph to instantiate $R(\cdot)$. Taking $R(\mathbf{X}_{\mathcal{B}}^{v})$ as an example, let $\kappa(\cdot,\cdot)$ denote a pairwise similarity function and $\mathcal{N}_i^v$ the local neighborhood of sample i. The relation between samples i and j for $\mathbf{X}_{\mathcal{B}}^{v}$ is defined as:

$$r_{ij}^{x,v} = \begin{cases} \dfrac{\exp\left(\kappa(\mathbf{x}_i^v, \mathbf{x}_j^v)/\tau_r\right)}{\sum_{k \in \mathcal{N}_i^v} \exp\left(\kappa(\mathbf{x}_i^v, \mathbf{x}_k^v)/\tau_r\right)}, & j \in \mathcal{N}_i^v \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where τ_r is a temperature parameter controlling the concentration of the relation distribution. The encoded relation graph $R(\mathbf{Z}_{\mathcal{B}}^{v})$ is constructed in the same manner from $\mathbf{Z}_{\mathcal{B}}^{v}$. Hence, $R(\mathbf{X}_{\mathcal{B}}^{v})$ and $R(\mathbf{Z}_{\mathcal{B}}^{v})$ describe the local relational states of the same samples before and after encoding, respectively, and can be used to characterize the relation changes introduced by the encoder.

We then use the relation distribution before encoding as a reference and constrain the encoded relation distribution from deviating excessively from it. For the v-th view, the relational preservation loss is defined as:

$$\mathcal{L}_{\text{rel}}^{v} = \frac{1}{B} \sum_{i=1}^{B} D_{\text{KL}}\left(R(\mathbf{X}_{\mathcal{B}}^{v}) \,\|\, R(\mathbf{Z}_{\mathcal{B}}^{v})\right) \tag{7}$$

where $D_{\text{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence, which measures the discrepancy between two probability distributions. Minimizing $\mathcal{L}_{\text{rel}}^{v}$ penalizes excessive deviation of the encoded neighborhood relations from their original reference, thereby limiting local structural distortion during view encoding. Averaging over all views gives:

$$\mathcal{L}_{\text{rel}} = \frac{1}{V} \sum_{v=1}^{V} \mathcal{L}_{\text{rel}}^{v} \tag{8}$$

This constraint provides each view with a local relational reference during representation learning. Unlike cross-view consistency constraints that align representations across views, it only preserves the intra-view relational structure and allows different views to maintain their own observations.

3.3 Sample-wise Consensus Formation

After obtaining view representations, we construct a consensus representation by integrating observations from different views of the same sample. For each sample, we treat its view representations as a set of observations and learn a consensus representation through cross-view interaction. We denote the representation of sample i in the v-th view by $\mathbf{z}_i^v$. The collection $\mathbf{z}_i^1, \ldots, \mathbf{z}_i^V$ therefore describes the sample across all V views. We implement the consensus function $g_\phi(\bullet)$ in Sec. 3.1 with self-attention, enabling adaptive interactions among views based on their observed information. A learnable consensus token $\mathbf{z}_{\text{cls}} \in \mathbb{R}^d$ is introduced to collect cross-view information and produce the consensus representation. The input token sequence is:

$$\mathbf{T}_i = [\mathbf{z}_{\text{cls}}, \mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^V] \in \mathbb{R}^{(V+1)\times d} \tag{9}$$

Self-attention is applied independently to the token sequence of each sample:

$$\widetilde{\mathbf{T}}_i = \text{SA}_\phi\left(\mathbf{T}_i\right) = [\widetilde{\mathbf{z}}_{i,\text{cls}}, \widetilde{\mathbf{z}}_i^1, \ldots, \widetilde{\mathbf{z}}_i^V] \tag{10}$$

where $\widetilde{\mathbf{T}}_i$ denotes the updated token sequence. In this sequence, the resulting view tokens are conditioned on the other observations of the same sample, while the consensus token is updated through interactions with all view tokens. Therefore, the updated representation $\widetilde{\mathbf{z}}_{i,\text{cls}}$ captures the cross-view information of sample i and serves as the sample-wise consensus representation:

$$\mathbf{c}_i = \widetilde{\mathbf{z}}_{i,\text{cls}} \qquad \mathbf{c}_i \in \mathbb{R}^d \tag{11}$$

To obtain clustering supervision, we collect the sample-wise consensus representations within a mini-batch and use them to discover the underlying clustering structure:

$$\mathbf{C}_\mathcal{B} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_B]^\top \in \mathbb{R}^{B\times d} \tag{12}$$

Here, $\mathbf{C}_\mathcal{B}$ contains the consensus representations of the B samples in the mini-batch and is used for clustering supervision in the following subsection.

3.4 Consensus-Only Clustering Supervision

Different from consistency-oriented methods that propagate clustering objectives to individual views, we apply clustering supervision only to the consensus representation in this subsection. Given the consensus representations $\mathbf{C}_\mathcal{B}$ obtained in Sec. 3.3, we further learn discriminative clustering structures from the sample-wise consensus representations within a mini-batch. A clustering head produces the logits $\mathbf{L} = [\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots, \boldsymbol{\ell}_B]^\top \in \mathbb{R}^{B\times K}$, where K denotes the number of clusters. For each sample i, the corresponding clustering logits are given by:

$$\boldsymbol{\ell}_i = g_\psi(\mathbf{c}_i) \tag{13}$$

where $g_\psi(\cdot)$ denotes the clustering head, consisting of an MLP projection followed by an orthogonal prototype-based classifier. Based on $\mathbf{L}$, we construct a clustering prediction distribution and a refined target distribution. The prediction distribution reflects the current clustering assignment, while the target distribution further explores the clustering structure to provide supervision. Therefore, the current clustering prediction is obtained by:

$$\mathbf{Q} = \text{softmax}(\mathbf{L}) \tag{14}$$

The target distribution is further derived from the current prediction by exploring its clustering structure, with stop-gradient used to avoid bidirectional coupling between the two branches:

$$\mathbf{P} = \mathcal{R}\left(\text{sg}(\mathbf{L})\right) \tag{15}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $\mathcal{R}(\cdot)$ denotes the target refinement process, including centering, sharpening, and Sinkhorn-Knopp normalization. The resulting target distribution provides

a structured supervision signal over the mini-batch, encouraging the consensus representations to form discriminative clustering structures. The clustering objective is defined as:

$$\mathcal{L}_{\mathrm{clu}} = D_{\mathrm{KL}}\left(\mathbf{P} \| \mathbf{Q}\right) \tag{16}$$

The target branch is detached during optimization, while gradients are propagated through the prediction branch. As discussed in Sec. 3.1, clustering supervision is therefore applied directly to the consensus representation, and its optimization signal is propagated to the view representations through the consensus formation module. Unlike consistency-oriented strategies that impose clustering targets on individual views, our formulation updates view representations only through their contributions to the consensus representation. Finally, the overall objective is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{clu}} + \lambda \mathcal{L}_{\mathrm{rel}}, \tag{17}$$

where λ is a hyperparameter which controls the contribution of the relational preservation term.

4 Experiments

We evaluate the proposed method on twelve widely used multi-view clustering benchmark datasets, including 100Leaves, Handwritten2V Li et al. (2026), Handwritten Chen et al. (2026a), Mfeat Gu et al. (2026), Wiki Rasiwasia et al. (2010), Animal Sun et al. (2024), AWA-7view Romera-Paredes & Torr (2015), ALOI100 Chen et al. (2026a), STL10 Chen et al. (2026a), YTF10 Chen et al. (2026b), XMedia Peng et al. (2018), and CIFAR100 Wang et al. (2026). We compare our method with eleven representative and recent multi-view clustering methods, including MFLVC Xu et al. (2022b), CVCL Chen et al. (2023), CSOT Zhang et al. (2024), VITAL He et al. (2024), MVCAN Xu et al. (2024), SparseMVC Liu et al. (2025), DDMVC Xu et al. (2025), AWDCMVC Wang et al. (2026), DRCFMVC Li et al. (2026), RAV Peng et al. (2026), and MANGO Chen et al. (2026a). We report Clustering Accuracy (ACC), Normalized Mutual Information (NMI), Purity (PUR), and Adjusted Rand Index (ARI) as evaluation metrics. All experiments are implemented in PyTorch 2.1.0 and conducted on an NVIDIA RTX 3090 GPU. Details of datasets are provided in Appendix B.1.

4.1 Comparison with State-of-the-Art Methods

Table 1 reports the comparison between our method and recent state-of-the-art multi-view clustering methods across multiple datasets. Overall, our method achieves the best or second-best results on the majority of datasets and evaluation metrics, while remaining competitive regardless of the number of views, data scales, and modality compositions. This suggests that the proposed consensus-oriented formulation generalizes well across diverse multi-view settings and can produce discriminative clustering structures.

It is also worth noting that existing consistency-oriented methods remain competitive on some benchmarks, such as Handwritten. This is consistent with our analysis: when clustering-relevant information is highly redundant across views, cross-view consistency learning can provide a reasonable approximation to consensus learning. However, its limitations become more evident when discriminative information is distributed unevenly across views or when the views are highly heterogeneous, such as AWA-7view and XMedia. Our method does not explicitly enforce cross-view consistency. Instead, it preserves view-local relational structures and performs clustering directly on the consensus representation, allowing complementary evidence from different views to be better incorporated. Overall, these results highlight that effective multi-view clustering relies on forming a reliable consensus that can integrate evidence across views.

4.2 Empirical Analysis of Consistency-Oriented Learning

To further examine the distinction between consistency learning and consensus learning, we augment the original model with an explicit cross-view consistency constraint $\mathcal{L}_{\mathrm{consis}}$ and a global-to-local supervision term $\mathcal{L}_{\mathrm{g2l}}$. Their strengths are controlled by β and γ, respectively. We report the corresponding constraint loss, the relational preservation loss $\mathcal{L}_{\mathrm{rel}}$, and ACC in Fig. 3.

As β increases, $\mathcal{L}_{\mathrm{consis}}$ steadily decreases, while $\mathcal{L}_{\mathrm{rel}}$ increases and ACC drops. Thus, stronger consistency constraints indeed make the views more consistent, but at the cost of larger changes to their local relational structures, without yielding a better clustering consensus. This trend is clear on

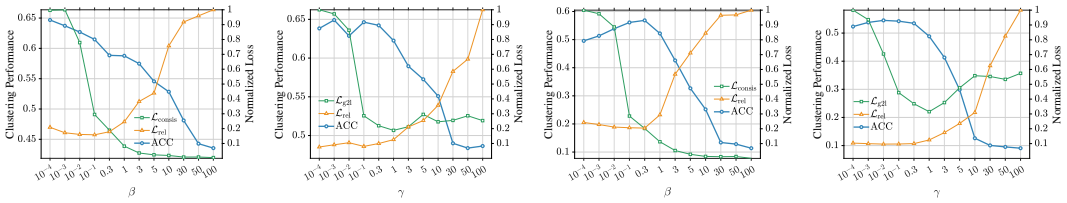
Table 1: Clustering results of different metrics on multi-view datasets

Datasets	Metrics	MFLVC	CVCL	CSOT	VITAL	MVCAN	Sparse MVC	DDMVC	AWDC MVC	DRCF MVC	RAV	MANGO	**Ours**
100Leaves	ACC	0.406	0.484	0.402	0.929	0.883	0.804	0.589	0.586	0.839	0.634	0.877	**0.964**
	NMI	0.716	0.774	0.706	0.978	0.959	0.919	0.831	0.818	0.930	0.833	0.946	**0.979**
	PUR	0.419	0.493	0.424	0.941	0.899	0.742	0.604	0.606	0.858	0.663	0.883	**0.964**
	ARI	0.290	0.415	0.274	0.914	0.848	0.833	0.510	0.484	0.778	0.528	0.817	**0.943**
Handwritten2v	ACC	0.819	0.916	0.752	0.907	0.927	0.858	0.905	0.881	0.915	0.825	0.947	**0.963**
	NMI	0.783	0.853	0.71	0.848	0.875	0.770	0.846	0.843	0.844	0.782	0.895	**0.918**
	PUR	0.819	0.916	0.752	0.907	0.927	0.718	0.905	0.881	0.915	0.825	0.947	**0.963**
	ARI	0.715	0.831	0.621	0.810	0.851	0.858	—	0.794	—	0.722	0.886	**0.919**
Handwritten	ACC	0.866	0.974	0.750	0.879	0.980	0.897	0.978	0.969	0.938	0.888	0.978	**0.986**
	NMI	0.856	0.941	0.745	0.874	0.953	0.834	0.949	0.928	0.875	0.843	0.948	**0.967**
	PUR	0.866	0.974	0.750	0.879	0.980	0.795	0.978	0.969	0.938	0.888	—	**0.987**
	ARI	0.807	—	0.646	0.831	0.955	0.897	0.952	0.931	0.867	0.791	0.952	**0.970**
Mfeat	ACC	0.876	0.946	0.834	0.879	0.979	0.913	0.884	0.924	0.941	0.938	0.975	**0.980**
	NMI	0.870	0.894	0.811	0.876	0.951	0.855	0.898	0.876	0.881	0.898	0.942	**0.952**
	PUR	0.876	0.946	0.834	0.879	0.979	0.827	0.884	0.924	0.941	0.938	0.975	**0.980**
	ARI	0.822	0.887	0.747	0.831	0.953	0.913	0.848	0.852	0.875	0.876	0.944	**0.955**
Wiki	ACC	0.462	0.530	0.488	0.542	0.555	0.590	0.581	0.365	0.604	0.459	0.567	**0.623**
	NMI	0.408	0.416	0.418	0.533	0.550	0.500	0.473	0.287	0.521	0.342	0.459	**0.533**
	PUR	0.497	0.570	0.540	—	0.625	0.423	0.393	0.393	**0.630**	0.510	0.593	0.625
	ARI	0.293	0.354	0.319	0.414	0.436	**0.615**	0.594	0.19	0.441	0.246	0.389	0.432
Animal	ACC	0.257	0.411	0.321	0.559	0.546	0.563	0.356	0.369	0.425	0.356	0.609	**0.648**
	NMI	0.393	0.537	0.479	0.634	0.668	0.664	0.499	0.507	0.540	0.499	0.687	**0.721**
	PUR	0.309	0.448	0.409	0.612	0.588	0.448	0.424	0.446	0.482	0.461	0.640	**0.680**
	ARI	0.171	0.329	0.216	0.439	0.477	0.621	0.246	0.258	0.316	0.252	0.516	**0.560**
AWA-7view	ACC	0.094	0.144	0.093	0.238	0.298	0.297	0.099	0.104	0.125	0.130	0.519	**0.558**
	NMI	0.131	0.221	0.140	0.335	0.447	0.473	0.132	0.160	0.214	0.203	0.580	**0.635**
	PUR	0.104	0.165	0.109	0.273	0.338	0.177	0.11	0.111	0.151	0.156	0.548	**0.595**
	ARI	0.037	0.064	0.026	0.125	0.209	0.373	0.031	0.038	0.054	0.042	0.393	**0.430**
ALOI100	ACC	0.539	0.521	0.845	0.845	0.860	0.822	0.830	0.674	0.856	0.826	0.887	**0.944**
	NMI	0.778	0.784	0.912	0.934	0.920	0.927	0.925	0.804	0.921	0.912	0.910	**0.949**
	PUR	0.554	0.539	0.851	0.868	0.871	0.842	0.838	0.689	0.867	0.830	—	**0.944**
	ARI	0.461	0.501	—	0.808	0.789	—	0.795	0.561	0.802	—	0.803	**0.896**
STL10	ACC	0.333	0.456	0.168	0.813	0.213	0.743	0.349	0.967	0.974	0.405	0.960	**0.979**
	NMI	0.234	0.192	0.021	0.791	0.271	0.734	0.240	0.920	0.934	0.186	0.911	**0.945**
	PUR	0.337	0.456	0.172	0.815	0.215	0.638	0.356	0.967	0.974	0.405	—	**0.979**
	ARI	0.144	0.160	0.011	0.743	0.111	0.748	0.161	0.929	0.943	0.143	0.912	**0.955**
YTF10	ACC	0.642	0.757	0.660	0.643	0.855	0.878	0.807	0.683	0.835	0.660	OM	**0.890**
	NMI	0.724	0.777	0.733	0.736	0.841	0.851	0.854	0.709	0.815	0.733	OM	**0.875**
	PUR	0.723	0.805	0.734	0.740	0.855	0.802	0.860	0.733	0.835	0.728	OM	**0.890**
	ARI	0.574	0.677	0.571	0.548	0.763	**0.878**	0.748	0.599	0.701	0.572	OM	0.809
XMedia	ACC	0.110	0.103	0.066	0.765	0.747	OT	0.109	0.458	0.307	0.314	OM	**0.919**
	NMI	0.304	0.287	0.287	**0.924**	0.873	OT	0.321	0.773	0.587	0.591	OM	0.921
	PUR	0.116	0.112	0.072	0.773	0.778	OT	0.114	0.477	0.338	0.334	OM	**0.919**
	ARI	0.039	0.033	0.034	0.723	0.681	OT	0.039	0.415	0.180	0.216	OM	**0.850**
CIFAR100	ACC	0.808	0.719	0.920	0.957	0.974	0.985	0.907	0.990	0.845	0.901	OM	**0.999**
	NMI	0.950	0.865	0.986	0.991	0.995	0.995	0.984	0.998	0.948	0.956	OM	**0.999**
	PUR	0.808	0.731	0.921	0.968	0.980	0.984	0.909	0.990	0.872	0.916	OM	**0.999**
	ARI	0.795	0.659	0.922	0.959	0.976	0.989	0.912	—	0.804	0.879	OM	**0.999**

* The best results are shown in bold, while the second-best results are indicated with underlining. "–" indicates that the corresponding metric was not reported in the original paper. "OT" denotes that the running time exceeded one week, and "OM" denotes out-of-memory.

both datasets, showing that optimizing consistency is not equivalent to optimizing consensus. As γ increases, $\mathcal{L}_{\mathrm{rel}}$ generally rises while ACC continues to decrease, indicating that overly strong global supervision can increasingly interfere with the relational structure of individual views. On AWA-7view, $\mathcal{L}_{\mathrm{g2l}}$ first decreases and then levels off with a slight rebound. This is because the global target is jointly determined by the view representations. Stronger global-to-local supervision not only changes the local representations but also alters the consensus target that they are required to match. As the supervision becomes stronger, the coupling between local representations and the global consensus increases, and $\mathcal{L}_{\mathrm{g2l}}$ therefore no longer decreases monotonically.

Overall, strengthening either cross-view consistency or global-to-local supervision does not improve clustering performance, but is instead accompanied by larger changes in local relational structure. These observations are consistent with our earlier analysis of consistency and consensus learning.



(a) Animal with $\mathcal{L}_{consis}$ (b) Animal with $\mathcal{L}_{g2l}$ (c) AWA-7view with $\mathcal{L}_{consis}$ (d) AWA-7view with $\mathcal{L}_{g2l}$

Figure 3: Effects of consistency and global-to-local supervision on view relations and clustering performance.

4.3 Ablation Study

To examine the effects of view-wise relational preservation and consensus formation on clustering performance, we conduct an ablation study by evaluating $\mathcal{L}_{\text{rel}}$ under different consensus formation strategies. Specifically, we investigate whether preserving view-local relational structures can facilitate consensus learning, and whether the proposed consensus formation strategy can effectively integrate multi-view evidence. The results are reported in Table 2.

Across different consensus formation strategies, $\mathcal{L}_{\text{rel}}$ shows a consistent positive effect. With relational preservation, Concat.+MLP, Mean Pooling, and Self-Attention either improve or maintain comparable performance, with more pronounced gains observed for Mean Pooling and Self-Attention. Among the consensus formation strategies, Self-Attention shows stronger overall adaptability. Mean Pooling or Concat.+MLP can slightly outperform it on individual datasets or metrics, but the margins are generally small and do not constitute a consistent advantage. In contrast, Self-Attention remains competitive across different data characteristics, with clearer gains when more views are involved and cross-view interactions become more complex.

Table 2: Ablation study of view-wise relational preservation and consensus formation.

$\mathcal{L}_{rel}$	Consensus Formation	Animal				AWA-7view				XMedia			
		ACC	NMI	PUR	ARI	ACC	NMI	PUR	ARI	ACC	NMI	PUR	ARI
✗	Concat.+MLP	0.587	0.710	0.634	0.534	0.313	0.453	0.333	0.217	0.879	0.903	0.879	0.803
✓	Concat.+MLP	0.610	**0.727**	0.671	**0.565**	0.312	0.463	0.330	0.212	0.906	0.918	0.907	0.839
✗	Mean Pooling	0.598	0.716	0.664	0.530	0.457	0.585	0.496	0.382	0.904	0.915	0.906	0.833
✓	Mean Pooling	0.622	0.724	0.665	0.561	0.509	0.628	0.556	0.435	**0.919**	**0.927**	**0.921**	**0.860**
✗	Self-Attention	0.613	0.701	0.637	0.525	0.309	0.452	0.351	0.200	0.869	0.903	0.874	0.801
✓	Self-Attention	**0.648**	0.722	**0.681**	0.560	**0.558**	**0.635**	**0.595**	**0.430**	**0.919**	0.921	0.919	0.850

* The best results are shown in bold, while the second-best results are indicated with underlining.

Overall, relational preservation provides consistent benefits across different consensus formation strategies, while Self-Attention yields a more balanced performance across datasets. We therefore recommend combining $\mathcal{L}_{\text{rel}}$ with Self-Attention for consensus-oriented representation learning.

5 Conclusion

This work revisits the relationship between consistency learning and consensus learning in multi-view clustering. Our analysis shows that consistency can provide an effective approximation to consensus when clustering-relevant information is highly redundant across views, whereas overly strong consistency may suppress view-specific evidence when discriminative information is available only in a subset of views. Based on this observation, we develop a consensus-oriented multi-view representation learning framework that preserves the local relational structure of each view and applies clustering supervision only at the consensus level. Experimental results demonstrate the effectiveness of the proposed framework, while also showing that stronger consistency constraints or global-to-local supervision do not necessarily lead to better results. Overall, multi-view representation learning should be driven by the clustering structure formed at the consensus level rather than by cross-view consistency alone. Future work will explore more effective mechanisms for consensus formation and consensus-level clustering structure discovery.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for implement methods and assist with translation. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpret results. Other required disclosure tasks include Generate synthetic datasets, clean and reformat dataset, support qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for create or edit software code, edit a research paper to improve readability. We have reviewed all AI-assistedwork. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work does not involve human subjects, private data, or sensitive information. All experiments are conducted on publicly available benchmark datasets for research purposes. Therefore, we do not identify any specific ethical concerns requiring further discussion.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work presents a simple yet effective framework for consensus-oriented multi-view clustering, where the main contribution lies in the formulation and learning strategy. The main paper and supplementary materials provide detailed descriptions of the framework, optimization objective, experimental protocols, and hyperparameter settings. All evaluations are performed on publicly available benchmark datasets. Upon acceptance, we will release the source code with complete implementation details to facilitate reproduction and future research.

REFERENCES

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2018.

Feiyang Chen, Ruiqiang Guo, and Zhibin Gu. Multi-scale diffusion-guided graph learning with power-smoothing random walk contrast for multi-view clustering. In *The Fourteenth international conference on learning representations*, 2026a.

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16706–16715. IEEE, 2023.

Xu Chen, Zhiwen Yu, Kaixiang Yang, and CL Philip Chen. A parameter-free multi-view clustering framework with adaptive anchors for large-scale data. *IEEE Transactions on Knowledge and Data Engineering*, 38(9):5594–5606, 2026b.

Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35 (12):12350–12368, 2023.

Wanrong Gu, Xinyu Zhang, Dong Huang, Guangyu Zhang, Haiyan Wang, and Huiling Zhang. One-step anchor-guided multi-view clustering via tensor-consistent graph factorization. *IEEE Transactions on Multimedia*, 2026. doi: 10.1109/TMM.2026.3727168.

Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. Robust variational contrastive learning for partially view-unaligned clustering. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 4167–4176, 2024.

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, pp. 17305–17313, 2025.

Zongmo Huang, Yazhou Ren, Xiaorong Pu, Shudong Huang, Zenglin Xu, and Lifang He. Self-supervised graph attention networks for deep weighted multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 7936–7943, 2023.

Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2018.

Ziqi Li, Junjie Liao, Yonghong Zhang, and Tianyang Xu. Dual-reliable contrastive fusion for multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(6):8329–8341, 2026.

Ruimeng Liu, Xin Zou, Chang Tang, Xiao Zheng, Xingchen Hu, Kun Sun, and Xinwang Liu. Sparsemvc: probing cross-view sparsity variations for multi-view clustering. *Advances in Neural Information Processing Systems*, 38:154594–154622, 2025.

Shuangmei Peng, Zhe Chen, Tianyang Xu, and Xiaojun Wu. Relationship alignment for view-aware multi-view clustering. In *International Conference on Learning Representations*, volume 2026, pp. 147335–147347, 2026.

Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2372–2385, 2018.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG 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.

Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pp. 2152–2161. PMLR, 2015.

Peng Su, Shudong Huang, Weihong Ma, Deng Xiong, and Jiancheng Lv. Multi-view granular-ball contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20637–20645, 2025.

Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):9150–9162, 2024.

Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1255–1265, 2021.

Shuangyang Wang, Lihua Zhou, and Bing Kong. Adaptive weighting-guided dual-level contrastive learning for multi-view clustering. *Knowledge-Based Systems*, pp. 115231, 2026.

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*, 26:9150–9162, 2024.

Chang Xu, Dacheng Tao, and Chao Xu. Multi-view intact space learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2531–2544, 2015.

Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2022a.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16030–16039. IEEE, 2022b.

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.

Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22957–22966. IEEE, 2024.

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

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19863–19872. IEEE, 2023.

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*, 33:4501–4515, 2024.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

A Detailed Analysis of Consistency and Consensus Learning

This section provides a detailed analysis of the two consistency-oriented paradigms discussed in Sec. 3.1.

A.1 Consistency-Oriented Consensus Formation

We first examine how cross-view consistency affects the preservation of clustering-relevant structures. For clarity, consider the representation-level objective in Eq. 3 with squared Euclidean discrepancy. For each sample i, the pairwise consistency objective can be transformed into the following center-based formulation without changing its optimization objective:

$$\mathbf{a}_i^* = \arg\min_{\mathbf{a}_i} \sum_{v=1}^{V} \|\mathbf{z}_i^v - \mathbf{a}_i\|_2^2 \tag{18}$$

whose solution is $\mathbf{a}_i^* = \frac{1}{V}\sum_{v=1}^{V} \mathbf{z}_i^v$. This construction is used only as a canonical abstraction for analyzing the behavior induced by cross-view consistency. Practical consistency objectives need not explicitly construct such a center.

For an arbitrary sample pair (i, j), let $\Delta_{ij}^v = \mathbf{z}_i^v - \mathbf{z}_j^v$ denote their pairwise difference in view v, and let $\Delta_{ij}^a = \mathbf{a}_i^* - \mathbf{a}_j^*$ denote the corresponding difference in the consistency representation. We have:

$$\Delta_{ij}^a = \frac{1}{V}\sum_{v=1}^{V} \Delta_{ij}^v \tag{19}$$

Consider first the case in which a clustering-relevant relation is repeatedly supported across views:

$$\Delta_{ij}^1 \approx \Delta_{ij}^2 \approx \cdots \approx \Delta_{ij}^V \tag{20}$$

Eq. 19 then preserves this pairwise structure in the consistency representation. This canonical case illustrates the regime in which clustering-relevant structures are repeatedly supported across views. From the information-theoretic perspective introduced in Sec. 3.1, the redundancy of clustering-relevant information is characterized by small Γ_{ij}^v, meaning that the remaining views provide little additional information about $\mathbf{S}_{ij}$ beyond view v.

The behavior changes when a discriminative structure is supported by only a subset of views. Consider an illustrative case in which the distinction between samples i and j is primarily captured by the k-th view, with $\Delta_{ij}^k \neq 0$ while $\Delta_{ij}^v \approx 0$ for $v \neq k$. Eq. 19 reduces to:

$$\Delta_{ij}^a \approx \frac{1}{V}\Delta_{ij}^k \tag{21}$$

Thus, the distinction contributed by the k-th view is attenuated because it receives little support from the remaining views. The factor $1/V$ is specific to this canonical averaging case, whereas the underlying effect is more general: cross-view consistency preferentially preserves structures that recur across views and may weaken structures supported by only a subset of views.

In contrast, consensus learning does not require clustering-relevant evidence to be present in every view. With $\mathbf{c}_i = g(\mathbf{z}_i^1, \ldots, \mathbf{z}_i^V)$, evidence from different views contributes directly to the consensus representation. Therefore, consensus learning can incorporate both redundant and complementary evidence across views, corresponding to $\Gamma_{ij}^v \approx 0$ and $\Gamma_{ij}^v > 0$, respectively.

A.2 Consensus Supervision and Global-to-Local Mismatch

We next analyze the second paradigm, where the global target is used to supervise individual views. From the definition in Eq. 2, the conditional mutual information can be rewritten as:

$$\begin{aligned}\Gamma_{ij}^v &= I\left(\mathbf{S}_{ij}; \mathbf{z}_i^{-v}, \mathbf{z}_j^{-v} \mid \mathbf{z}_i^v, \mathbf{z}_j^v\right) \\ &= H\left(\mathbf{S}_{ij} \mid \mathbf{z}_i^v, \mathbf{z}_j^v\right) - H\left(\mathbf{S}_{ij} \mid \mathbf{z}_i^{1:V}, \mathbf{z}_j^{1:V}\right)\end{aligned} \tag{22}$$

where $\mathbf{z}_i^{1:V}$ denotes the representations of sample i from all views, and $H(\cdot|\cdot)$ denotes conditional entropy. When $\Gamma_{ij}^v \approx 0$, we have:

$$H\left(\mathbf{S}_{ij} \mid \mathbf{z}_i^v, \mathbf{z}_j^v\right) \approx H\left(\mathbf{S}_{ij} \mid \mathbf{z}_i^{1:V}, \mathbf{z}_j^{1:V}\right) \tag{23}$$

In this case, the remaining views provide little additional clustering-relevant information about $\mathbf{S}_{ij}$ beyond that already available from view v. Consequently, global-to-local supervision introduces little additional clustering-relevant information beyond the observation of this view.

When $\Gamma_{ij}^v > 0$, we have:

$$H\left(\mathbf{S}_{ij} \mid \mathbf{z}_i^{1:V}, \mathbf{z}_j^{1:V}\right) < H\left(\mathbf{S}_{ij} \mid \mathbf{z}_i^v, \mathbf{z}_j^v\right) \tag{24}$$

The joint multi-view observations therefore contain clustering-relevant information about $\mathbf{S}_{ij}$ that is unavailable from view v alone. Although $\mathbf{S}_{ij}$ is defined at the pairwise level, the collection of such relations determines the clustering structure. Hence, additional evidence about $\mathbf{S}_{ij}$ available from other views can affect the global clustering prediction $\mathbf{q}_i^c$. Directly enforcing $\mathbf{q}_i^v \approx \mathbf{q}_i^c$ may therefore require view v to reproduce a clustering decision that is partly supported by evidence outside its own observation. This gives rise to the global-to-local supervision mismatch described in Sec. 3.1.

Consensus-level clustering avoids imposing such a view-wise matching constraint. Let $\mathbf{c}_i = g(\mathbf{z}_i^1, \ldots, \mathbf{z}_i^V)$ and $\mathcal{L}_{\text{consen}} = \mathcal{L}_{\text{clu}}(\mathbf{C})$. The clustering gradient received by view v is propagated through its contribution to the consensus representation:

$$\frac{\partial \mathcal{L}_{\text{clu}}}{\partial \mathbf{z}_i^v} = \frac{\partial \mathcal{L}_{\text{clu}}}{\partial \mathbf{c}_i}\frac{\partial \mathbf{c}_i}{\partial \mathbf{z}_i^v} \tag{25}$$

Thus, view v participates in clustering through the consensus representation rather than through an independent constraint that forces its prediction to match the global consensus. Whether $\Gamma_{ij}^v \approx 0$ or $\Gamma_{ij}^v > 0$, each view can contribute the evidence available in its own observation, while complementary evidence from other views is integrated at the consensus level.

B Additional Experimental Details

All experiments were conducted on Ubuntu 20.04 using a server equipped with an Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz, 128 GB of RAM, and an NVIDIA RTX 3090 GPU. Upon acceptance, we will release the source code with complete implementation details.

B.1 Dataset Details

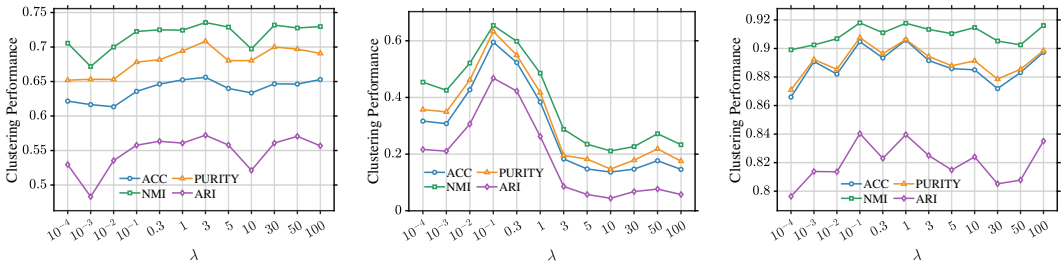
Table 3 summarizes the twelve multi-view benchmark datasets used in our experiments, including the number of samples, clusters, and the dimensionality of each view.

Table 3: Statistics of the multi-view benchmark datasets.

Dataset	Samples	Clusters	$d^{(1)}, d^{(2)}, \cdots, d^{(V)}$
100leaves	1600	100	64,64,64
Handwritten2V	2000	10	240,76
Handwritten	2000	10	240,76,216,47,64,6
Mfeat	2000	10	216,76,64,6,240,47
Wiki	2866	10	128,10
Animal	10158	50	4096,4096
AWA-7view	10158	50	2688,2000,2000,20000,2000,4096,4096
ALOI100	10800	100	77,13,64,125
STL10	13000	10	1024,512,2048
YTF10	38654	10	944,576,512,640
XMedia	40000	200	4096,300
CIFAR100	50000	100	512,2048,1024

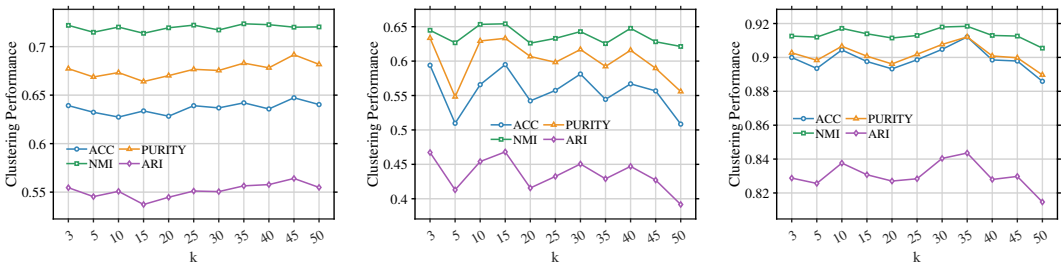
B.2 Sensitivity to Hyperparameters

Our method mainly involves two hyperparameters: the trade-off parameter λ for relational preservation and the neighborhood size k. Specifically, λ controls the weight of relational preservation in the overall objective, while k specifies the number of neighbors used to construct local relations within each view. To study their effects on clustering performance, we conduct parameter sensitivity experiments on Animal, AWA-7view, and XMedia. Specifically, we vary λ over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.3, 1, 3, 5, 10, 30, 50, 100\}$, and vary k from 3 to 50.



(a) Animal (b) AWA7view (c) XMedia

Figure 4: Sensitivity analysis of the trade-off parameter λ



(a) Animal (b) AWA7view (c) Xmedia

Figure 5: Sensitivity analysis of the neighborhood size k

As shown in Fig. 4, the effect of λ varies across different datasets. The performance on Animal and XMedia remains relatively stable over a wide range, while AWA-7view is more sensitive to λ and performs better with smaller values. As λ further increases, the clustering performance does not continue to improve and shows a clear drop for AWA-7view. This suggests that relational preservation

is intended to retain local structures within each view, rather than dominate the overall optimization. An overly large λ may place too much emphasis on local relation preservation and affect the formation of discriminative consensus. Therefore, λ should be adjusted according to the dataset. Based on the results, we recommend searching λ within the range of $0.1 \sim 3$.

For the neighborhood size k, as shown in Fig. 5, the performance is generally stable within a moderate range. Animal is particularly stable with respect to k, while AWA-7view and XMedia also maintain good performance over a reasonable range. This indicates that relational preservation does not rely on a narrowly tuned value of k. However, when k becomes larger, both AWA-7view and XMedia show a certain performance drop. A possible reason is that, as the neighborhood expands, it may gradually extend beyond the actual local structure, making the preserved relations less effective in capturing discriminative local structures. Therefore, we do not recommend using an overly large neighborhood and suggest choosing k within the range of $10 \sim 40$.