Hierarchical Cross-View Alignment for Multi-View Clustering via Decoupled Information Distillation

Taichun Zhou[1], Siwei Wang[2*], Zhibin Dong[1], Jiaqi Jin[1], Ke Liang[1], Baili Xiao[1], Miaomiao Li[3], Xinwang Liu[1*], En Zhu[1*]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, China, 410073
[2]Intelligent Game and Decision Lab, Academy of Military Sciences, Beijing, China, 100850
[3]School of Computer, Changsha College, Changsha, 410022, China
taichunzhou@163.com

Abstract

Multi-view clustering aims to learn robust and comprehensive clustering structures by exploiting shared and complementary information from diverse data sources. However, the inherent heterogeneity across views poses substantial challenges to effective collaboration and information integration. Although recent studies introduce distillation-based mechanisms to reduce heterogeneity and enhance cross-view consistency, they often rely on manually designed transfer paths or fixed fusion weights, limiting their ability to model complex and dynamic view relationships. To address this limitation, we propose HOARD, a novel framework for Hierarchical crOss-view Alignment for multi-view clusteRing via Decoupled information distillation. HOARD decouples multi-view representations into shared and specific components and performs hierarchical alignment across them. Specifically, we design a granular-ball contrastive alignment to enforce semantic consistency among shared features, and develop a prototype-based collaborative transmission strategy to align specific features while preserving their view-dependent structures. In addition, an information distillation unit adaptively models cross-view knowledge transfer in both feature spaces, and an attention-based fusion module integrates shared and specific representations. Extensive experiments on multiple benchmark datasets demonstrate that HOARD substantially improves alignment quality and clustering performance, achieving state-of-the-art results.

Introduction

In real-world data collection scenarios, an object is often described through multiple perceptual views or information channels, forming a multi-view data (Wen et al. 2023; Ren et al. 2025; Liang et al. 2025; Hu et al. 2024; Wan et al. 2024; Feng et al. 2025). For instance, an image can be characterized using complementary descriptors such as color histograms, edge shapes, and spatial layouts. A biomedical sample may contain various views including gene expression profiles, DNA methylation levels, and clinical annotations. These diverse views collectively provide a rich and complementary semantic characterization of the object. Effectively mining the consistency and latent correlations across views has thus become a central objective in multi-view clustering (Hu et al. 2025a; Du et al. 2025; Wang et al. 2024b).

In recent years, deep multi-view clustering has witnessed remarkable progress by harnessing the nonlinear modeling capacity of neural networks (Chen et al. 2023a; Hu et al. 2025b; Xu et al. 2025; Yu et al. 2025; Dong et al. 2025b). These methods excel at capturing cross-view interactions and uncovering complex relationships, delivering impressive clustering performance across a variety of practical applications. However, view heterogeneity—caused by differences in sensing modalities, representation granularity, and sampling mechanisms—remains a critical barrier to unified modeling (Gan et al. 2025; Wang et al. 2024a). In particular, the relative importance of a sample may vary drastically across views in real-world settings, leading to the entanglement of shared semantics and view-specific structures, which exacerbates the modeling difficulty (Lu et al. 2024).

Early fusion approaches (Wang et al. 2019, 2021) often simplify multi-view modeling by directly concatenating or projecting all views into a unified latent space. However, such indiscriminate fusion tends to entangle view-specific characteristics with shared semantics, resulting in semantic interference and degrading the ability to capture structural consistency across views. To mitigate this, some studies (Wang et al. 2024c; Ke et al. 2023) have introduced knowledge distillation mechanisms to facilitate inter-view collaboration. Yet, most existing methods rely on predefined distillation paths or static weighting schemes, which lack adaptability to diverse view combinations and sample-level variations. Moreover, pervasive distributional shifts across views further limit the effectiveness of unified distillation strategies.

To address these challenges, we propose a novel hierarchical cross-view alignement framework via view- decoupling, termed HOARD. Our framework begins with representation decoupling, decomposing each view into a shared (homogeneous) component and a specific (heterogeneous) component, explicitly tackling semantic consistency and view heterogeneity, respectively. To enhance cross-view consistency, we design a granular-ball contrastive alignment mechanism that constructs structured contrastive objectives to enforce semantic alignment and structural integrity across shared representations. To alleviate feature shifts caused by hetero-

*Corresponding authors.

geneity, we introduce a prototype-level collaborative transport strategy that leverages optimal transport to build fine-grained cross-view prototype alignments, guided by distributional discrepancy measures.

Furthermore, to enable adaptive and efficient collaboration across decoupled spaces, we propose a information distillation unit that models inter-view knowledge transfer paths and distillation weights via a dynamic graph structure, allowing the network to learn cross-view information flow in a data-driven manner. Finally, we incorporate an attention-based fusion module to integrate the shared and specific representations, enhancing high-order semantic modeling while reducing inconsistencies across views. Our contributions are summarized as follows:

- We propose a novel framework, HOARD, which performs hierarchical and collaborative modeling of view-specific and shared semantics via representation decoupling and dual-stream alignment, significantly enhancing cross-view integration.
- We design a dynamic information distillation unit that learns adaptive knowledge transfer paths and distillation weights within decoupled spaces, supporting flexible and effective cross-view information flow.
- Extensive experiments on multiple benchmark datasets demonstrate that our method achieves superior semantic consistency and clustering performance while preserving view-specific characteristics, consistently outperforming state-of-the-art baselines.

Related Work

Deep Multi-view Representation Learning

Deep multi-view representation learning aims to extract complementary information from multiple views to derive unified or collaborative feature representations (Wang et al. 2023; Cui et al. 2024a; Zhou et al. 2025; Cui et al. 2024b; Dong et al. 2025a), thereby enhancing the performance of downstream tasks such as clustering and classification. Early works primarily relied on canonical correlation analysisand its extensions, such as DCCA (Wang et al. 2015) and DGCCA (Xu et al. 2021), which project multi-view data into a shared latent space using deep neural networks. More recent studies have adopted autoencoder-based or graph neural network frameworks to reconstruct each view while capturing inter-view correlations. For instance, Multi-VAE (Xu et al. 2021) introduces a variational autoencoder framework with disentangled view-common and view-specific latent variables, which facilitates the discovery of discrete clustering structures across views. Similarly, TTAE (Wang et al. 2025b) leverages a tensor transformer autoencoder to model information exchange between views while applying a self-expression mechanism to enhance subspace clustering. However, these methods often assume strict alignment or overlook inter-view heterogeneity, reducing their clustering performance.

Disentangled Representation Learning

Disentangled representation learning seeks to separate semantically meaningful and independent factors from data. In multi-view learning, this involves decomposing view-invariant shared representations from view-feature ones. Recent works (Bao 2021; Wu et al. 2024; Xu et al. 2023) enhance disentanglement via structural mechanisms. MRDD (Ke et al. 2024) introduces distilled disentangling with masked cross-view prediction, filtering redundant consistency signals to improve both shared and private representations. TGM-MVC (Wang et al. 2024a) constructs a graph across views and preserves inter-view differences via a minimum spanning tree. DIVIDE (Lu et al. 2024) employs high-order random walks to identify informative pairs, mitigating false positives while preserving semantic coherence. However, many methods rely on implicit disentanglement and lack hierarchical modeling. To address this, we propose an explicit hierarchical disentanglement mechanism with contrastive alignment and graph-based distillation, enhancing both representation quality and cross-view consistency.

Methodology

In this section, we propose a deep multi-view clustering framework named HOARD, which consists of four key modules: multi-view Feature Disentanglement; Shared Feature Alignment Enhancement; Cross-View Discrepancy Mitigation; Attention-based Fusion Mechanism. The overall framework is illustrated in Figure 1.

Preliminary Work

Let the multi-view dataset be denoted as $\mathcal{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(V)}\}$, where $\mathbf{X}^{(v)} \in \mathbb{R}^{N \times d_v}$ represents the feature matrix of the v-th view, with N samples and input dimension d_v. Each view captures distinct and complementary aspects of the same set of objects. Each sample $\mathbf{x}_i^{(v)}$ is decomposed into a shared representation $\mathbf{z}_i^{(v)}$ and a specific representation $\mathbf{s}_i^{(v)}$ via a set of parameterized nonlinear mappings. This design disentangles shared and view-specific information while maintaining their independence and input reconstructability. Specifically, we employ view-specific encoders $\mathcal{E}^{(v)}(\cdot;\psi^{(v)})$ and decoders $\mathcal{D}^{(v)}(\cdot;\varphi^{(v)})$, along with shared modules $\mathcal{E}(\cdot;\psi)$ and $\mathcal{D}(\cdot;\varphi)$ that are common across all views except for a few linear layers at the input and output ends. The final clustering representation is constructed by concatenating $\mathbf{z}_i^{(v)}$ and $\mathbf{s}_i^{(v)}$.

Multi-view Feature Disentanglement

To exploit consistency and heterogeneity across views, we first decouple features into shared and specific representations via a shared and a view-specific module, optimized by reconstruction loss $\mathcal{L}_{rec}$ and orthogonality loss $\mathcal{L}_{ort}$. The disentanglement is performed as follows:

$$\mathbf{z}_i^{(v)} = \mathcal{E}^{(v)}(\mathbf{x}_i^{(v)};\psi^{(v)}), \quad \mathbf{s}_i^{(v)} = \mathcal{E}(\mathbf{x}_i^{(v)};\psi), \qquad (1)$$

for $i = 1, 2, \cdots, N$ and $v = 1, 2, \cdots, V$. The reconstruction loss based on the disentangled representation is formulated as follows:

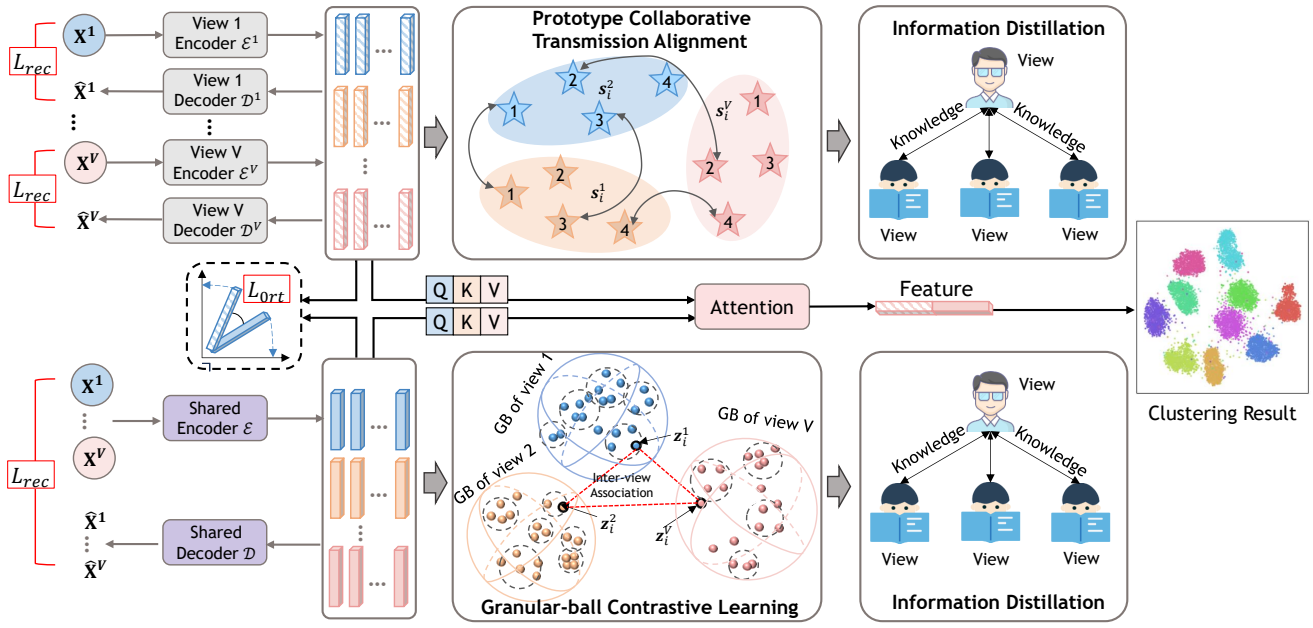


Figure 1: The framework of HOARD consists of three parts: Decouple, Align, and Fuse. In the first stage, each view is encoded into shared and specific features via a dual-branch reconstruction module. The second stage performs hierarchical alignment, where shared representations are aligned via granular-ball contrastive learning, and specific representations are aligned via prototype collaborative transmission alignment strategy. In the final stage, graph-based information distillation and attention fusion are applied to enhance representation quality and generate clustering results.

$$\mathcal{L}_{rec} = \sum_{v=1}^{V}\sum_{i=1}^{N}\Big(\Big\|\mathbf{x}_i^{(v)} - \mathcal{D}^{(v)}(\mathcal{E}^{(v)}(\mathbf{x}_i^{(v)};\psi^{(v)});\varphi^{(v)})\Big\|_2^2 + \Big\|\mathbf{x}_i^{(v)} - \mathcal{D}(\mathcal{E}(\mathbf{x}_i^{(v)};\psi);\varphi)\Big\|_2^2\Big), \tag{2}$$

where the reconstructed losses for shared and specific representations from all views are aggregated. Furthermore, to ensure that the shared representation $\mathbf{z}_i^{(v)}$ and the specific representation $\mathbf{s}_i^{(v)}$ are statistically independent, we introduce an orthogonality loss $\mathcal{L}_{ort}$.

$$\mathcal{L}_{ort} = \sum_{v=1}^{V}\sum_{i=1}^{N}\pi\left(\mathbf{z}_i^{(v)}, \mathbf{s}_i^{(v)}\right), \tag{3}$$

where $\pi(\cdot,\cdot)$ denotes the cosine similarity between the instance features. Therefore, the loss in the decoupling process is:

$$\mathcal{L}_{dec} = \mathcal{L}_{rec} + \mathcal{L}_{ort}. \tag{4}$$

Shared Feature Alignment Enhancement

To enhance feature homogeneity and facilitate effective knowledge transfer, we first align shared representations$\{\mathbf{z}_i^{(v)}\}_{i=1}^N$ from each view $v \in \{1,\ldots,V\}$ via granule-based contrastive learning. This promotes semantic consistency and structural clarity, providing a unified space that enables the distillation graph to model inter-view relations more accurately and propagate reliable cross-view knowledge.

Granular-ball Construction for Shared Representations. To capture consistency and structure, we employ granular-ball contrastive learning (Su et al. 2025; Xia et al. 2021) to associate granular regions within and across views. For the v-th view, we partition the shared representations $\mathbf{z}_i^{(v)}$ into g_v granular regions by applying k-means clustering:

$$g_v = \max\left(\left\lfloor \frac{N}{\theta} \right\rfloor, 1\right), \tag{5}$$

where θ is a granularity control parameter. This yields a set of cluster centers $\mathbf{C}^{(v)} = \{\mathbf{c}_1^{(v)},\ldots,\mathbf{c}_{g_v}^{(v)}\}$ and corresponding sample indices $I_i^{(v)} = \{j \mid \mathbf{z}_j^{(v)} \in b_i^{(v)}\}$ for each granular ball $b_i^{(v)}$.

To reflect local density, we assign a radius to each granular ball as:

$$r_i^{(v)} = \frac{1}{|b_i^{(v)}|}\sum_{\mathbf{z}_j^{(v)} \in b_i^{(v)}}\left\|\mathbf{z}_j^{(v)} - \mathbf{c}_i^{(v)}\right\|_2. \tag{6}$$

Cross-view Association of Granular Balls. To establish semantic correspondences across views, we define a binary association matrix $\mathbf{E}^{(m,n)} \in \{0,1\}^{g_m \times g_n}$ of view m and view n:

$$e_{ij}^{(m,n)} = \begin{cases} 1, & \text{if } \frac{|I_i^{(m)} \cap I_j^{(n)}|}{\min(|b_i^{(m)}|,|b_j^{(n)}|)} \geq \delta, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

where δ is the threshold controlling the overlap, which is set to 0.1 by default.

Contrastive Learning on Granular Centers. We apply contrastive learning on the center representations to ensure alignment. First, we define the combined center matrix:

$$\mathbf{C}^{(m,n)} = [\mathbf{C}^{(m)}; \mathbf{C}^{(n)}],$$

and construct the positive mask matrix:

$$\mathbf{H}^{(m,n)} = \begin{bmatrix} \mathbf{A}^{(m)} & \mathbf{E}^{(m,n)} \\ (\mathbf{E}^{(m,n)})^\top & \mathbf{A}^{(n)} \end{bmatrix}, \tag{8}$$

where $\mathbf{A}^{(v)}$ denotes intra-view adjacency based on neighborhood overlap, when $\left\|\mathbf{c}_i^{(v)} - \mathbf{c}_j^{(v)}\right\|_2 < r_i^{(v)} + r_j^{(v)}$, $a_{ij}^{(v)} = 1$, otherwise 0.

The granular-ball contrastive loss is then defined as:

$$l_g^{(m,n)} = \frac{1}{|B|} \sum_{i \in B} \sum_{j \in \Delta_i^+} \frac{\exp(\cos(\mathbf{c}_i, \mathbf{c}_j))}{\sum_{k \in \Delta_i^-} \exp(\cos(\mathbf{c}_i, \mathbf{c}_k))}, \tag{9}$$

where the postive sample set $\Delta_i^+ = \{j \mid \mathbf{h}_{ij}^{(m,n)} = 1, \forall j\}$ and negative sample set $\Delta_i^- = \{k \mid \mathbf{h}_{ik}^{(m,n)} = 0, \forall k\}$. B denotes the total number of granular balls, and $\cos(\cdot)$ is used to measure the similarity between two vectors.

Finally, we obtain the total alignment loss by averaging across all view pairs:

$$\mathcal{L}_G = \frac{2}{V(V-1)} \sum_{m<n} l_g^{(m,n)}. \tag{10}$$

This process encourages the semantic consistency and structural alignment of shared representations, thereby facilitating more reliable cross-view knowledge distillation.

Information Distillation Unit. To enable adaptive knowledge transfer, we introduce an Information Distillation Unit (IDU) that models inter-view relations as a directed graph and learns to distill predictive signals with weighted edges. Each view is treated as a node v_i, with $w_{i \to j}$ denoting the distillation weight from view i to j. The distillation signal $\eta_{i \to j}$ captures the discrepancy between their logits, and the aggregated loss for target view v_j is defined as:

$$\mu_{:j} = \sum_{v_i \in \mathcal{N}(v_j)} w_{i \to j} \cdot \eta_{i \to j}, \tag{11}$$

where $\eta_{i \mapsto j} = \|\mathbf{z}_i - \mathbf{z}_j\|_2$, and $\mathcal{N}(v_j)$ denotes the set of source views contributing to v_j.

To compute adaptive and dynamic weights $w_{i \to j}$, we embed both the logits and hidden representations of each view into a shared distillation space. The injection weight is then computed as:

$$w_{i \to j} = h\left([f(\mathbf{z}_i, \sigma_1), f(\mathbf{z}_j, \sigma_1), \mathbf{z}_i, \mathbf{z}_j]; \sigma_2\right), \tag{12}$$

where $f(\cdot, \sigma_1)$ is a learnable function that projects inputs to the logit space, and $h(\cdot, \sigma_2)$ is a fully connected layer that outputs pairwise edge weights given concatenated features. The final edge matrix $\mathbf{W}$ is constructed by evaluating $w_{i \to j}$ for all pairs and normalizing through a softmax operation to ensure probabilistic interpretation.

Finally, we define the overall distillation loss as:

$$\mathcal{L}_{dis} = \|\mathbf{W} \odot \mathbf{E}\|_1 + \|\bar{\mathbf{w}} - \mathbf{a}\|_2^2, \tag{13}$$

where $\mathbf{E}$ stores the pairwise distillation errors $\eta_{i \to j}$, and $\odot$ denotes element-wise multiplication. The second term regularizes the average weights $\bar{\mathbf{w}}$ towards a uniform prior $\mathbf{a}$ via $\|\bar{\mathbf{w}} - \mathbf{a}\|_2^2$. This design enables the network to emphasize reliable modalities and suppress noisy ones. Through iterative optimization, IDU adaptively models cross-view dependencies and facilitates both shared and specific knowledge distillation.

Ultimately, the overall loss for shared feature enhancement is the sum of the Granular-ball contrastive loss and the information distillation loss:

$$\mathcal{L}_{sha} = \mathcal{L}_G + \mathcal{L}_{dis}. \tag{14}$$

Cross-View Discrepancy Mitigation

In multi-view tasks, view-specific features carry unique semantics, but discrepancies across views may cause semantic shifts and hinder collaboration. To address this, we propose a prototype collaborative transmission alignment strategy to reduce cross-view discrepancies while preserving specificity, followed by information distillation to enhance shared representations.

Prototype Collaborative Transmission Alignment

Differentiable Prototype Extraction. For specific features $\mathbf{s}_i^{(v)}$, we first extract a set of prototypes for each view via a differentiable K-means process. Specifically, we denote the prototype set for view v as $\mathbf{P}^{(v)}$, obtained by:

$$\mathbf{P}^{(v)} = \left[\mathbf{p}_1^{(v)}, \ldots, \mathbf{p}_Q^{(v)}\right]. \tag{15}$$

To stabilize the prototype evolution across training iterations, we introduce a momentum-based prototype smoothing mechanism. For each view-specific prototype $\mathbf{P}^{(v)}$, its update incorporates the historical estimate ${\mathbf{P}_{his}}^{(v)}$ as:

$$\tilde{\mathbf{P}}^{(v)} \leftarrow \alpha \cdot \mathbf{P}^{(v)} + (1 - \alpha) \cdot \mathbf{P}_{his}^{(v)}, \tag{16}$$

where $\alpha \in (0, 1)$. To maintain gradient flow via soft reassignment, we first compute the soft-assignment scores

$$f_{iq}^{(v)} = -\left\|\mathbf{s}_i^{(v)} - \tilde{\mathbf{p}}_q^{(v)}\right\|_2^2, \tag{17}$$

transform them into probabilities

$$\gamma_{iq}^{(v)} = \frac{\exp\left(f_{iq}^{(v)}\right)}{\sum_{q=1}^{Q} \exp\left(f_{iq}^{(v)}\right)}, \tag{18}$$

and re-estimate the centroids in a fully differentiable manner:

$$\hat{\mathbf{p}}_q^{(v)} = \frac{\sum_{i=1}^{N} \gamma_{iq}^{(v)} \mathbf{s}_i^{(v)}}{\sum_{i=1}^{N} \gamma_{iq}^{(v)}}, \hat{\mathbf{P}}^{(v)} = \left[\hat{\mathbf{p}}_1^{(v)}, \ldots, \hat{\mathbf{p}}_Q^{(v)}\right]. \tag{19}$$

Prototype-level Optimal Transport. We quantify the discrepancy between their prototype sets by the cost matrix:

$$\mathbf{o}_{kl}^{(m,n)} = \left\|\hat{\mathbf{p}}_k^{(m)} - \hat{\mathbf{p}}_l^{(n)}\right\|_2^2, \tag{20}$$

where $\mathbf{o}_{kl}^{(m,n)}$ represents the pairwise alignment cost between views m and n.

Given $\mathbf{O}^{(m,n)}$, we compute the Sinkhorn transport plan $\mathbf{T}^{(m,n)}$ under entropy regularization:

$$\mathbf{T}^{(m,n)} = \arg\min_{\mathbf{T}\in\Pi(\mathbf{a},\mathbf{b})} \langle \mathbf{T}, \mathbf{O}^{(m,n)} \rangle - H(\mathbf{T}). \tag{21}$$

where $\mathbf{T}$ denotes a candidate transport plan, $H(\mathbf{T}) = -\sum_{k,l} \mathbf{T}_{kl} \log \mathbf{T}_{kl}$, and $\Pi(\mathbf{a},\mathbf{b}) = \{\mathbf{T} \in \mathbb{R}_+^{Q\times Q} \mid \mathbf{T}\mathbf{1} = \mathbf{a},\ \mathbf{T}^\top \mathbf{1} = \mathbf{b}\}$ with uniform marginals $a_k = b_k = \frac{1}{Q}$.

Collaborative Transmission Loss. The final alignment loss aggregates the OT cost and the prototype-level MMD:

$$l_c^{(m,n)} = \left\langle \mathbf{T}^{(m,n)}, \mathbf{O}^{(m,n)} \right\rangle + \mathrm{MMD}\big(\hat{\mathbf{P}}^{(n)}, \bar{\mathbf{P}}^{(m\to n)}\big), \tag{22}$$

Here, the MMD term is computed using an RBF kernel with automatic bandwidth selection, encouraging semantic consistency under distributional alignment.Transporting prototypes from view m to n gives $\bar{\mathbf{P}}^{(m\to n)} = \mathbf{T}^{(m,n)}\,\hat{\mathbf{P}}^{(m)}$.

And the overall loss averages across all view pairs

$$\mathcal{L}_C = \frac{2}{V(V-1)} \sum_{2\le m<n\le V} l_c^{(m,n)}. \tag{23}$$

Information Distillation Unit. After the prototype collaborative transmission alignment of the specific representations, we apply an Information Distillation Unit—identical in structure to the one used for the shared representations to transfer knowledge across views. The resulting distillation loss is then given by:

$$\mathcal{L}_{dis} = \|\mathbf{W} \odot \mathbf{E}\|_1 + \|\bar{\mathbf{w}} - \mathbf{a}\|_2^2. \tag{24}$$

Ultimately, the overall loss for specific feature part is the sum of the collaborative transmission loss and the information distillation loss:

$$\mathcal{L}_{spe} = \mathcal{L}_C + \mathcal{L}_{dis}. \tag{25}$$

Attention-based Fusion Mechanism

To integrate multi-view information, we employ an attention-based fusion mechanism that separately processes the distilled shared representations $\mathbf{z}_i^{(v)}$ and specific representations $\mathbf{s}_i^{(v)}$. For both, we adopt identical self-attention networks to capture high-order semantics and derive the unified features $\mathbf{Z}$ and $\mathbf{S}$.

For a given view v, the attention mechanism begins by projecting the input feature $\mathbf{X}^{(v)} \in \mathbb{R}^{N\times d}$ into query, key, and value matrices via:

$$\mathbf{Q}^{(v)} = \mathbf{X}^{(v)}\mathbf{W}_Q^{(v)}, \mathbf{K}^{(v)} = \mathbf{X}^{(v)}\mathbf{W}_K^{(v)}, \mathbf{V}^{(v)} = \mathbf{X}^{(v)}\mathbf{W}_V^{(v)}, \tag{26}$$

where $\mathbf{W}_Q^{(v)}, \mathbf{W}_K^{(v)}, \mathbf{W}_V^{(v)} \in \mathbb{R}^{d\times d_k}$ are learnable projection matrices, and d_k is the dimensionality of the transformed space. The attention score matrix is computed via scaled dot-product attention:

$$\text{Attention}(\mathbf{Q}^{(v)}, \mathbf{K}^{(v)}) = \text{softmax}\left(\frac{\mathbf{Q}^{(v)}\mathbf{K}^{(v)\top}}{\sqrt{d_k}}\right), \tag{27}$$

which measures the similarity between queries and keys. The resulting normalized scores are used to weight the value matrix to produce attention-enhanced features $\mathbf{A}^{(v)} = \text{Attention}(\mathbf{Q}^{(v)}, \mathbf{K}^{(v)})\mathbf{V}^{(v)}$,where $\mathbf{A}^{(v)} \in \mathbb{R}^{N\times d_k}$ is the attention-weighted representation for view v.

After obtaining the attention-enhanced features $\mathbf{A}^{(v)}$ for all views, we average them to produce the global shared feature $\mathbf{Z}$ and global specific feature $\mathbf{S}$:

$$\mathbf{Z} = \sum_{v=1}^{V} \varepsilon_{sha}^{(v)} \mathbf{A}_{sha}^{(v)}, \quad \mathbf{S} = \sum_{v=1}^{V} \varepsilon_{spe}^{(v)} \mathbf{A}_{spe}^{(v)}, \tag{28}$$

where $\mathbf{A}_{sha}^{(v)}$ and $\mathbf{A}_{spe}^{(v)}$ denote the attention-enhanced representations for shared and specific features of view v, respectively. $\varepsilon_{sha}^{(v)}$ and $\varepsilon_{spe}^{(v)}$ are weights of attention.

Finally, we feed the concatenation of $\mathbf{Z}$ and $\mathbf{S}$ into a multilayer perceptron (MLP) to produce the final unified representation:

$$\mathbf{H} = \text{MLP}([\mathbf{Z}\|\mathbf{S}]), \tag{29}$$

where $[\cdot\|\cdot]$ denotes concatenation and $\mathbf{H}$ is the final fused representation that incorporates both global shared semantics and view-specific nuances.

Objective Function and Optimization

Given the preceding definitions, the HOARD framework aims to minimize a composite loss comprising three components: $\mathcal{L}_{dec}$, $\mathcal{L}_{sha}$, $\mathcal{L}_{spe}$, corresponding to decouple, contrastive, and fusion consistency losses, respectively. The complete objective is defined as:

$$\mathcal{L} = \mathcal{L}_{dec} + \lambda_1 \mathcal{L}_{sha} + \lambda_2 \mathcal{L}_{spe}, \tag{30}$$

where λ_1 and λ_2 are trade-off hyperparameters that balance different optimization components.

Dataset	Samples	Views	Dimensionality
Synthetic3d	600	3	[3,3,3]
CUB	600	2	[1024,300]
UCI	2000	3	[216,76,64]
MFeat	2000	2	[76,240]
CiteSeer	3312	4	[3703,3312,3312,3312]
Hdigit	10000	2	[784,256]
NoisyMNIST	70000	2	[784,784]

Table 1: Information of seven datasets.

Experiments

Experimental Settings

Datasets and Experimental Settings. To thoroughly assess the effectiveness of the proposed model, we perform evaluations on seven widely used multi-view datasets: Synthetic3d, CUB, UCI, MFeat, CiteSeer, Hdigit and NoisyMNIST. The characteristics of these datasets are presented in Table 1. Our model uses two hyperparameters λ_1 and λ_2, both ranging between $[0.1, 1000]$, and a learning rate of 0.0001. We implemented the model in PyTorch 2.0.1 with an NVIDIA GeForce RTX 3090 GPU and 64GB RAM, utilizing the Adam optimizer with default settings.

Dataset	MFLVC	CVCL	DealMVC	GCFAgg	ACCMVC	DMAC	DMVC-CE	HOARD
				ACC(%)				
Synthetic3d	96.33	61.00	97.00	66.50	97.50	66.00	70.10	**97.67**
CUB	39.50	72.33	50.33	60.67	74.83	71.33	23.10	**79.83**
CiteSeer	27.20	35.00	29.14	21.26	43.32	21.62	19.34	**46.98**
MFeat	72.45	80.15	81.70	45.80	78.18	82.75	29.19	**93.30**
Hdigit	99.14	99.26	89.38	37.12	98.19	86.45	36.65	**99.32**
UCI	91.20	90.50	84.40	63.75	90.50	82.80	23.79	**93.65**
NoisyMNIST	69.64	97.75	96.65	34.26	96.24	OM	17.42	**97.94**
				NMI(%)				
Synthetic3d	85.68	55.39	87.43	41.91	88.88	27.50	49.42	**89.42**
CUB	50.55	67.13	61.92	58.30	67.67	72.73	15.19	**75.44**
CiteSeer	8.69	14.49	11.34	0.24	18.14	1.05	0.33	**22.94**
MFeat	74.65	76.89	76.99	48.50	76.07	75.79	23.15	**87.47**
Hdigit	97.43	97.78	94.04	45.79	94.92	75.59	34.70	**97.83**
UCI	85.25	83.52	82.28	59.33	83.66	74.67	15.00	**88.06**
NoisyMNIST	83.02	93.87	94.81	48.30	90.71	OM	7.96	**95.36**
				ARI(%)				
Synthetic3d	89.38	43.94	91.27	39.66	92.66	25.70	42.45	**93.11**
CUB	26.24	53.60	46.09	45.02	56.08	60.91	7.72	**65.06**
CiteSeer	3.76	8.78	6.29	0.10	17.04	0.73	0.08	**19.00**
MFeat	63.94	68.82	70.49	27.99	66.30	68.33	12.49	**85.85**
Hdigit	98.10	**98.36**	87.85	22.09	98.19	72.87	21.82	98.32
UCI	82.18	80.66	77.42	43.89	80.37	66.82	7.50	**87.22**
NoisyMNIST	54.59	95.14	93.24	23.35	92.01	OM	3.91	**95.16**

Table 2: Clustering algorithms are compared on benchmark datasets using ACC, NMI, and ARI. The best performance is highlighted in **bold**, while the second-best is shown with underline. "OM" denotes out-of-memory failure.

Compared Methods and Evaluation Metrics. To evaluate our method, we compare it with seven state-of-the-art baselines: MFLVC (Xu et al. 2022), CVCL (Chen et al. 2023b), DealMVC (Yang et al. 2023), GCFAgg (Yan et al. 2023), ACCMVC (Yan et al. 2024), DMAC (Wang et al. 2025a), and DMVC-CE (Zhang et al. 2025). We assess clustering performance in the unsupervised setting using three standard metrics: clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI).

Comparison with State-of-the-art Baselines

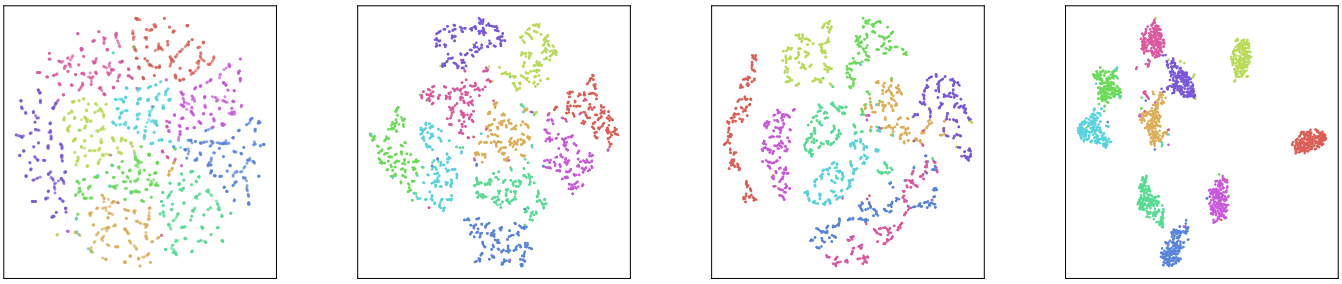
We evaluate HOARD on seven benchmark multi-view datasets against seven state-of-the-art methods using three standard metrics, as shown in Table 2. Overall, HOARD consistently achieves either the best or second-best performance across most datasets. In terms of ACC, HOARD ranks first on all datasets—Synthetic3d (97.67%), CUB(79.83%), CiteSeer(46.98%), MFeat (93.30%), Hdigit (99.32%), UCI (93.65%), and NoisyMNIST (97.94%)—outperforming all baselines by a notable margin. Particularly on challenging datasets like NoisyMnist, where several methods suffer from out-of-memory (OOM) failures, HOARD maintains robust and stable performance, highlighting its scalability and noise resilience. For NMI and ARI, it also ranks among the top performers, demonstrating its ability to preserve semantic consistency and capture latent cluster structures. While some methods (e.g., DealMVC, ACCMVC) perform well on specific datasets, their overall stability degrades on high-dimensional or heterogeneous data. In contrast, the hierarchical cross-view alignment mechanism, information distillation module, and structure-preserving strategy introduced in HOARD work collaboratively to build a highly consistent and discriminative clustering framework, delivering more robust and superior performance across multi-view datasets with diverse types and multi-scale features.

Ablation Studies

Datasets	$\mathcal{L}_{dec}$	$\mathcal{L}_{sha}$	$\mathcal{L}_{spe}$	ACC	NMI	ARI
CUB	✓	✓	✓	**79.83**	**75.44**	**65.06**
	✓		✓	41.09	16.87	11.75
	✓	✓		71.67	70.95	58.07
	✓			65.17	62.63	47.69
MFeat	✓	✓	✓	**93.30**	**87.47**	**85.85**
	✓		✓	46.17	59.13	37.20
	✓	✓		76.60	73.69	65.84
	✓			77.10	72.58	64.78
Hdigit	✓	✓	✓	**99.32**	**97.83**	**98.32**
	✓		✓	64.76	49.82	41.21
	✓	✓		97.89	95.48	95.30
	✓			70.59	62.12	53.94

Table 3: Ablation study on three datasets with different loss situation.

- **Loss Component Analysis.** We evaluate the impact of each loss by removing the shared alignment loss ($\mathcal{L}_{sha}$) or the specific alignment loss ($\mathcal{L}_{spe}$) on three representative datasets. As shown in Table 3, removing any loss term leads to noticeable performance drops, with $\mathcal{L}_{sha}$ having the largest impact. The full model performs best

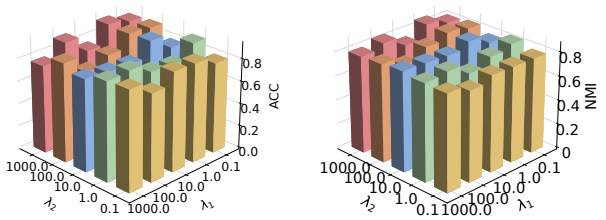


(a)Raw(ACC=73.70%) (b)Shared(ACC=76.15%) (c)Specific(ACC=76.95%) (d)Final(ACC=93.30%)

Figure 2: t-SNE visualization of clustering results for different feature spaces on the MFeat dataset.

Datasets	GB	OT	ACC	NMI	ARI
CUB	✓	✓	**79.83**	**75.44**	**65.06**
	✓		66.00	74.83	63.24
		✓	61.17	62.89	45.47
			55.17	56.51	37.76
MFeat	✓	✓	**93.30**	**87.47**	**85.85**
	✓		86.45	77.44	73.12
		✓	72.15	66.77	56.40
			73.35	72.11	63.22
Hdigit	✓	✓	**99.32**	**97.83**	**98.32**
	✓		97.85	95.39	95.19
		✓	68.79	57.64	49.97
			73.39	70.27	63.22

Table 4: Ablation study on three datasets with different module combinations.



(a) ACC on MFeat (b) NMI on MFeat

Figure 3: Parameter sensitivity analysis on MFeat.

overall, demonstrating the necessity of jointly optimizing decoupling, shared alignment, and specific modeling.

- **Module Contribution Evaluation.** We further ablate the Granular-Ball contrastive module (GB) and the Prototype-based OT module (OT). Results in Table 4 show that removing either module reduces clustering performance. Eliminating GB notably degrades results on CUB and MFeat, while removing OT diminishes semantic discriminability. The full model combining both modules achieves the highest performance, indicating their complementary effects.

Sensitivity Analysis and Visaulization

We perform sensitivity analysis on the hyperparameters λ_1 and λ_2 within the range $\{0.1, 1, 10, 100, 1000\}$, as shown in Figure 3. The model maintains stable clustering performance on the MFeat dataset, with ACC and NMI consistently above 0.8, demonstrating low sensitivity to changes in hyperparameters. Figure 2 illustrates the t-SNE visualizations of features at different stages on the MFeat dataset. The original features (a) exhibit poor separability, with significant overlap between clusters. The shared features (b) capture local discrimination, while the specific features (c) improve global alignment. The final fused features (d) form compact and well-separated clusters, validating the effectiveness of integrating shared and specific representations.

Conclusion

In this paper, we propose HOARD, a novel hierarchical alignment framework for unsupervised multi-view clustering. By decoupling shared and specific representations, HOARD enables explicit modeling of inter-view consistency and diversity. The introduction of granular-ball contrastive alignment ensures fine-grained alignment of homogeneous semantics, while a prototype collaborative transmission alignment strategy mitigates view-specific discrepancies at the distributional level. Furthermore, the adaptive distillation unit dynamically captures inter-view knowledge flow based on representation similarity, enhancing cross-view collaboration. Extensive experiments on multiple benchmarks demonstrate that HOARD achieves state-of-the-art performance.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China under Grant 2022ZD0209103; in part by the National Natural Science Foundation of China under Project 62325604, Project 62276271, Project 62406329, Project 62476280, Project 62506371, and Project 62441618; and in part by the National Natural Science Foundation of China Joint Found under Grant U24A20323.

References

Bao, F. 2021. Disentangled Variational Information Bottleneck for Multiview Representation Learning. In *Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part II*, 91–102. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-93048-6.

Chen, J.; Mao, H.; Woo, W. L.; and Peng, X. 2023a. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, 16752–16761.

Chen, J.; Mao, H.; Woo, W. L.; and Peng, X. 2023b. Deep Multiview Clustering by Contrasting Cluster Assignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16752–16761.

Cui, J.; Li, Y.; Huang, H.; and Wen, J. 2024a. Dual Contrast-Driven Deep Multi-View Clustering. *IEEE Transactions on Image Processing*, 33: 4753–4764.

Cui, J.; Li, Y.; Huang, H.; and Wen, J. 2024b. Dual Contrast-Driven Deep Multi-View Clustering. *IEEE Transactions on Image Processing*, 33: 4753–4764.

Dong, Z.; Hu, D.; Jin, J.; Wang, S.; Liu, X.; and Zhu, E. 2025a. Selective Cross-View Topology for Deep Incomplete Multi-View Clustering. *IEEE Transactions on Image Processing*, 34: 4792–4805.

Dong, Z.; Liu, M.; Wang, S.; Liang, K.; Zhang, Y.; Liu, S.; Jin, J.; Liu, X.; and Zhu, E. 2025b. Enhanced then Progressive Fusion with View Graph for Multi-View Clustering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15518–15527.

Du, S.; Fang, Z.; Tan, Y.; Wang, C.; Wang, S.; and Guo, W. 2025. OpenViewer: Openness-Aware Multi-View Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15): 16389–16397.

Feng, Y.; Liang, W.; Wan, X.; Liu, J.; Li, M.; and Liu, X. 2025. Incremental Multi-View Clustering: Exploring Stream-View Correlations to Learn Consistency and Diversity. *IEEE Transactions on Knowledge and Data Engineering*.

Gan, Y.; You, Y.; Huang, J.; Xiang, S.; Tang, C.; Hu, W.; and An, S. 2025. Multi-View Clustering via Multi-Stage Fusion. *IEEE Transactions on Multimedia*, 27: 4571–4583.

Hu, D.; Dong, Z.; Liang, K.; Yu, H.; Wang, S.; and Liu, X. 2024. High-order Topology for Deep Single-cell Multi-view Fuzzy Clustering. *IEEE Transactions on Fuzzy Systems*.

Hu, S.; Tian, B.; Liu, W.; and Ye, Y. 2025a. Self-supervised Trusted Contrastive Multi-view Clustering with Uncertainty Refined. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16): 17305–17313.

Hu, S.; Zhang, C.; Zou, G.; Lou, Z.; and Ye, Y. 2025b. Deep Multiview Clustering by Pseudo-Label Guided Contrastive Learning and Dual Correlation Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2): 3646–3658.

Ke, G.; Wang, B.; Wang, X.; and He, S. 2024. Rethinking multi-view representation learning via distilled disentangling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26774–26783.

Ke, G.; Yu, Y.; Chao, G.; Wang, X.; Xu, C.; and He, S. 2023. Disentangling multi-view representations beyond inductive bias. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2582–2590.

Liang, K.; Meng, L.; Li, H.; Wang, J.; Lan, L.; Li, M.; Liu, X.; and Wang, H. 2025. From Concrete to Abstract: Multi-view Clustering on Relational Knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Lu, Y.; Lin, Y.; Yang, M.; Peng, D.; Hu, P.; and Peng, X. 2024. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 14193–14201.

Ren, Y.; Pu, J.; Yang, Z.; Xu, J.; Li, G.; Pu, X.; Yu, P. S.; and He, L. 2025. Deep Clustering: A Comprehensive Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4): 5858–5878.

Su, P.; Huang, S.; Ma, W.; Xiong, D.; and Lv, J. 2025. Multi-view Granular-ball Contrastive Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20637–20645.

Wan, X.; Xiao, B.; Liu, X.; Liu, J.; Liang, W.; and Zhu, E. 2024. Fast Continual Multi-View Clustering With Incomplete Views. *IEEE Transactions on Image Processing*, 33: 2995–3008.

Wang, B.; Zeng, C.; Chen, M.; and Li, X. 2025a. Towards Learnable Anchor for Deep Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 21044–21052.

Wang, F.; Jin, J.; Dong, Z.; Yang, X.; Feng, Y.; Liu, X.; Zhu, X.; Wang, S.; Liu, T.; and Zhu, E. 2024a. View Gap Matters: Cross-view Topology and Information Decoupling for Multi-view Clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 8431–8440. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Wang, F.; Jin, J.; Hu, J.; Liu, S.; Yang, X.; Wang, S.; Liu, X.; and Zhu, E. 2024b. Evaluate then cooperate: shapley-based view cooperation enhancement for multi-view clustering. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Wang, J.; Feng, S.; Lyu, G.; and Gu, Z. 2023. Triple-granularity contrastive learning for deep multi-view subspace clustering. In *Proceedings of the 31st ACM international conference on multimedia*, 2994–3002.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2021. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30: 1771–1783.

Wang, Q.; Zhang, Z.; Feng, W.; Tao, Z.; and Gao, Q. 2025b. Contrastive Multi-view Subspace Clustering via Tensor Transformers Autoencoder. In *Proceedings of the*

AAAI Conference on Artificial Intelligence, volume 39, 21207–21215.

Wang, S.; Liu, X.; Zhu, E.; Tang, C.; Liu, J.; Hu, J.; Xia, J.; and Yin, J. 2019. Multi-view clustering via late fusion alignment maximization. In *IJCAI*, 3778–3784.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International conference on machine learning*, 1083–1092. PMLR.

Wang, X.; Wang, Y.; Ke, G.; Wang, Y.; and Hong, X. 2024c. Knowledge distillation-driven semi-supervised multi-view classification. *Information Fusion*, 103: 102098.

Wen, J.; Zhang, Z.; Fei, L.; Zhang, B.; Xu, Y.; Zhang, Z.; and Li, J. 2023. A Survey on Incomplete Multiview Clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2): 1136–1149.

Wu, S.; Zheng, Y.; Ren, Y.; He, J.; Pu, X.; Huang, S.; Hao, Z.; and He, L. 2024. Self-Weighted Contrastive Fusion for Deep Multi-View Clustering. *IEEE Transactions on Multimedia*, 26: 9150–9162.

Xia, S.; Zheng, S.; Wang, G.; Gao, X.; and Wang, B. 2021. Granular ball sampling for noisy label classification or imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4): 2144–2155.

Xu, H.; Wang, Q.; Wang, B.; and Gao, Q. 2025. Deep Fair Multi-View Clustering with Attention KAN. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5061–5070.

Xu, J.; Ren, Y.; Shi, X.; Shen, H. T.; and Zhu, X. 2023. UNTIE: Clustering analysis with disentanglement in multi-view information fusion. *Information Fusion*, 100: 101937.

Xu, J.; Ren, Y.; Tang, H.; Pu, X.; Zhu, X.; Zeng, M.; and He, L. 2021. 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*, 9234–9243.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-Level Feature Learning for Contrastive Multi-View Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16051–16060.

Yan, W.; Zhang, Y.; Lv, C.; Tang, C.; Yue, G.; Liao, L.; and Lin, W. 2023. GCFAgg: Global and Cross-View Feature Aggregation for Multi-View Clustering. 19863–19872.

Yan, W.; Zhang, Y.; Tang, C.; Zhou, W.; and Lin, W. 2024. Anchor-sharing and cluster-wise contrastive network for multiview representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2): 3797–3807.

Yang, X.; Jiaqi, J.; Wang, S.; Liang, K.; Liu, Y.; Wen, Y.; Liu, S.; Zhou, S.; Liu, X.; and Zhu, E. 2023. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*, 337–346.

Yu, S.; Liu, S.; Wang, S.; Tang, C.; Luo, Z.; Liu, X.; and Zhu, E. 2025. Sparse Low-Rank Multi-View Subspace Clustering With Consensus Anchors and Unified Bipartite Graph. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 1438–1452.

Zhang, Y.; Cai, J.; Wu, Z.; Wang, P.; and Ng, S.-K. 2025. Mixture of Experts as Representation Learner for Deep Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22704–22713.

Zhou, T.; Dong, Z.; Wang, S.; Liang, K.; Li, M.; Liu, X.; Zhu, E.; and Dong, X. 2025. DPFMVC: Dynamic Progressive Fusion for Multi-view Clustering. MM '25, 1102–1111. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.