CONSENSUS-GUIDED DEEP MULTI-VIEW CLUSTERING WITH DUAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering aims to leverage complementary information from different data sources to improve clustering performance. However, many existing methods emphasize cross-view consensus in deep representations while underexploiting structural cues retained in shallower features, leaving complementary information underused and making the learned clusters less discriminative. Moreover, they often explore clustering patterns through centroid assignment or alignment, insufficiently modeling inter-sample relationships. Same-cluster samples may remain dispersed while different-cluster samples lie close together, leaving clusters dispersed with unclear boundaries. To address these challenges, we propose **CO**nsensus-guided deep multi-view clustering with dual **RE**presentation learning (CORE), a framework that combines multi-hierarchical representation learning, semantic invariance, and hard-samples contrastive alignment. Specifically, CORE employs deep and shallow branches to extract abstract semantics and preserve structural cues, respectively, allowing the model to use complementary information when learning discriminative representations. It then constructs a global clustering target from concatenated deep representations to guide semantic invariance across views. Our analysis derives an explicit bound showing when a discriminative view's distance advantage outweighs conflicts to determine the consensus alignment. Finally, CORE uses consensus-derived pseudo-labels to jointly mine hard positive and negative samples. Contrastive learning then increases the similarity of the selected hard positives and decreases that of the hard negatives, improving separation between clusters. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art methods.

1 INTRODUCTION

The proliferation of data acquisition technologies has made multi-view data ubiquitous. Each view provides a distinct yet partial glimpse into the same set of entities, promising a more complete and robust understanding than any single source alone. A fundamental task in analyzing such data is multi-view clustering, which aims to partition data points into coherent groups by synthesizing information from all available views without manual labels (Jiang et al., 2025; Liu et al., 2026). The promise is clear: by fusing complementary evidence, algorithms can overcome the limitations and biases inherent in any single view, leading to discoveries that are more stable and generalizable (Zhu et al., 2025; Wang et al., 2025b; Yin et al., 2025).

Recent advances in deep learning have significantly promoted multi-view clustering (MVC), leading to deep MVC methods with improved representation and clustering performance (Gao et al., 2024; Chen et al., 2025a). Most existing approaches employ autoencoders to learn view-specific representations, align them in a shared latent space to ensure cross-view consistency, and perform clustering through various learning objectives (Ma et al., 2026; Dong et al., 2025). To better address inter-view discrepancies, techniques such as contrastive learning and label distribution alignment have been introduced. For example, MFLVC (Xu et al., 2022) constructs contrastive consistency objectives for high-level features and semantic labels. Similarly, MCMC (Bian et al., 2025) performs contrastive learning at the cluster, instance, and prototype levels to improve intra-cluster compactness, inter-cluster separability, and assignment consistency. Overall, deep MVC methods mainly improve clustering by learning discriminative embeddings while exploiting cross-view consistency and complementary information.

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.

Existing multi-view clustering methods often emphasize consensus across views among deep representations, leaving structural cues available in shallower features underused. The resulting embeddings may capture shared semantics without fully exploiting the complementary information. Moreover, clustering patterns explored through centroid assignments or alignment may provide limited guidance on relationships between individual samples: samples assigned to the same cluster may remain far apart, while those assigned to different clusters may remain close, leaving clusters dispersed with unclear boundaries. These challenges motivate a fundamental research question: *How can multi-view clustering combine semantic and structural information while refining both within-cluster and between-cluster relationships?*

To address these limitations, we propose **Co**nsensus-Guided Deep Multi-View Clustering with Dual **Re**presentation Learning (CORE), a framework that integrates semantic abstraction with structural detail to capture complementary information across views. Specifically, CORE employs a deep branch to extract abstract semantic features and a shallow branch to retain structural and contour information encoded in each view. Cross-view semantic invariance is then driven by a consensus derived from the deep representations. Our analysis identifies an explicit bound showing when a discriminative view's distance advantage outweighs conflicting views and cluster-frequency correction to determine the consensus alignment. Furthermore, we fuse the deep and shallow representations and use pseudo-labels to guide hard positive and negative samples mining in the fused embedding space. The resulting contrastive objective promotes within-cluster cohesion and between-cluster separation, connecting complementary representation learning with consensus-guided discriminative refinement. The key contributions of our work are threefold:

- We propose a dual-representation framework that combines semantic abstraction with structural information to capture complementary cues. A global clustering target derived from concatenated deep representations guides cross-view semantic invariance.
- We introduce HaSE, a hardness-aware structure exploration mine hard positive and negative samples using consensus-derived pseudo-labels. By refining fused deep and shallow representations, HaSE promotes within-cluster cohesion and between-cluster separation.
- Extensive experiments on multiple benchmark datasets demonstrate that CORE achieves state-of-the-art performance, outperforming nine competitive baselines.

2 Related Works

2.1 Multi-view Clustering

Conventional multi-view clustering methods can be broadly categorized into subspace-based (Shi et al., 2024; Lan et al., 2022), matrix factorization-based (Zhao et al., 2021; Yang et al., 2021), and graph-based methods (Wu et al., 2020; Zhong & Pun, 2021). However, their limited representation ability and high computational complexity often restrict performance on complex real-world data. Benefiting from deep learning, deep multi-view clustering (DMVC) (Fei et al., 2025; Xu et al., 2026) has made substantial progress by exploiting complementary information and structural cues across views. ADMC (Zhao et al., 2024) learns view-specific features through autoencoders, adaptively fuses them, and uses actively acquired labels to guide optimization. EASEMVC (Xiao et al., 2025) employs Optimal Transport for adaptive view selection and dual-selection contrastive learning to mitigate low-quality views. EMLFCL (Yi et al., 2025) constructs an anchor view and aligns each view with it at both feature and clustering levels. Nevertheless, these methods mainly emphasize cross-view consistency while insufficiently preserving valuable structural and contour information, limiting the discriminability of the learned representations.

2.2 Contrastive Learning

Contrastive learning (Su et al., 2025; Yan et al., 2023) has recently emerged as an effective representation learning paradigm. In multi-view settings, different views of the same sample are treated as positive pairs, while views from different samples form negative pairs. The widely used InfoNCE loss improves representations by increasing agreement between positive pairs and separating negative ones. CVCL (Chen et al., 2023) employs cluster-level contrastive learning to align cross-view assignments and discover consistent semantic information. To mitigate false negatives, DCMVC

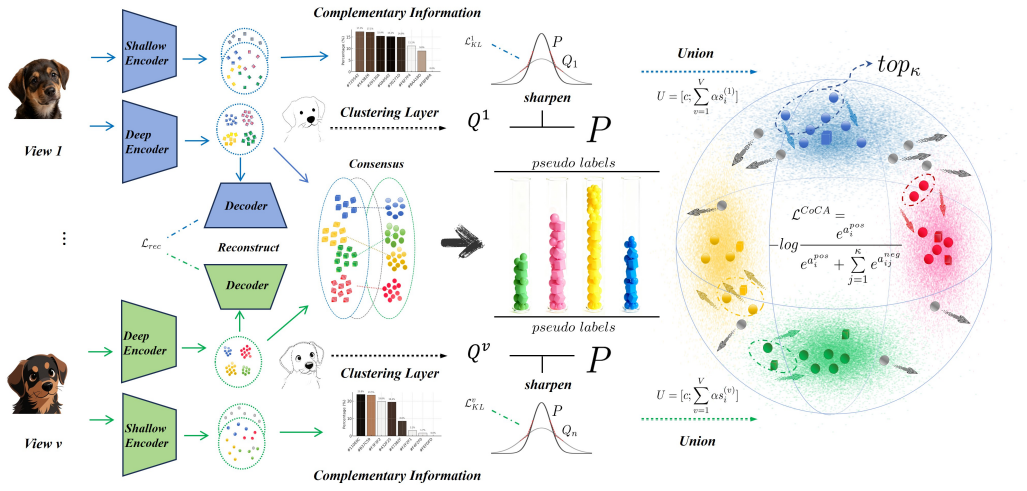


Figure 1: Overview of our method.

(Cui et al., 2024) incorporates nearest-neighbor structural guidance into instance-level contrastive learning and enhances inter-cluster separation at the cluster level. PGCL (Hu et al., 2025c) further uses pseudo-labels to progressively remove false negative pairs. MCMVC (Geng et al., 2025) introduces view-label-guided complementary learning and performs contrastive learning from both feature and view-label perspectives. Nevertheless, these methods generally assume that fused consensus representations are reliable, overlooking prototype shifts caused by view diversity. Optimization based on such biased representations and prototypes may therefore reduce clustering reliability.

3 METHOD

This section presents CORE, whose architecture is illustrated in Figure 1. 3.1 describes multi-hierarchical representation learning. 3.2 explains how consensus guides semantic invariance across views. 3.3 details feature fusion and hard samples contrastive learning.

3.1 MULTI-HIERARCHICAL REPRESENTATION EXTRACTION

The Multi-hierarchical representation extraction (MRE) captures semantic and structural information at different levels of abstraction to support discriminative clustering. To accomplish this, MRE construct deep and shallow encoding–decoding branches within each view, each trained through reconstruction to retain informative content from the input. The deep branch learns abstract semantic features, whereas the shallow branch helps preserve structural and contour cues.

For each view v, let $\mathbf{X}^{(v)} \in \mathbb{R}^{N\times d_v}$ contain N samples with d_v input features. A deep encoder $\mathbf{E}_d^v(\theta_d^v)$ and a shallow encoder $\mathbf{E}_s^v(\theta_s^v)$ produce

$$C_i^{(v)} = \mathbf{E}_d^v(\mathbf{X}^v), \quad S_i^{(v)} = \mathbf{E}_s^v(\mathbf{X}^v) \tag{1}$$

where θ_d and θ_s are their learnable parameters. The deep encoder is a three-layer MLP that supports a higher-capacity transformation and provides the representations used to construct the global clustering target, whereas the shallow encoder uses a single linear layer to provides a lower-capacity encoding path to preserve structural information. Following processing through both encoders, we obtain semantic encoded by deep encoder $C_i^{(v)} \in \mathbb{R}^{N\times d}$ and structural information encoded by shallow encoder $S_i^{(v)} \in \mathbb{R}^{N\times d}$, which undergo reconstruction via corresponding decoders $\mathbf{D}_d^v(C_i^{(v)}; \phi_d^v)$ and $\mathbf{D}_s^v(S_i^{(v)}; \phi_s^v)$, where ϕ denotes the learnable parameters of the decoders. MRE optimize the feature extraction process by minimizing the reconstruction loss between the reconstructed and original data, formalized as:

$$\mathcal{L}_{\text{rec}} = \sum_{v=1}^{V} \left\| \mathbf{X}^v - \mathbf{D}_d^v\big(\mathbf{E}_d^v(\mathbf{X}^v)\big) \right\|_2^2 + \left\| \mathbf{X}^v - \mathbf{D}_s^v\big(\mathbf{E}_s^v(\mathbf{X}^v)\big) \right\|_2^2. \tag{2}$$

3.2 SEMANTIC-INVARIANT CONSENSUS LEARNING

The semantic-invariant consensus learning(SCL) integrates clustering target from deep representations across views to guide semantic learning. To achieve this, SCL constructs soft assignments from the deep features across views and their concatenation, both using a Student's t kernel, then refines the assignments into a shared target distribution. By minimizing the Kullback–Leibler divergence from this target to each view-specific assignment distribution, SCL promotes cross-view semantic invariance while preserving separate representation spaces across views.

Soft assignments For v-th view, $C_i^{(v)}$ of sample i is embedded into a space parameterized by cluster centroids $\{\mu_j^{(v)}\}_{j=1}^K$. A soft assignment is computed using a t-distribution kernel:

$$q_{ij}^{(v)} = \frac{\left(1 + \|C_i^{(v)} - \mu_j^{(v)}\|^2/\beta\right)^{-\frac{\beta+1}{2}}}{\sum_{k=1}^K \left(1 + \|C_i^{(v)} - \mu_k^{(v)}\|^2/\beta\right)^{-\frac{\beta+1}{2}}}, \tag{3}$$

where $q_{ij}^{(v)}$ denotes the probability of assigning $C_i^{(v)}$ to cluster j, and β is a degrees-of-freedom parameter. Similarly, we concatenate the deep representations across views as $C_i = \mathrm{Concat}(C_i^{(1)}, \ldots, C_i^{(V)})$ and obtain global centroids m_j via K-means. Applying the same Student's t-distribution kernel yields the global assignments q_{ij}.

Setting and notation Partition each global centroid into its view-specific blocks: $m_j = \mathrm{Concat}\left(m_j^{(1)}, \ldots, m_j^{(V)}\right)$. Here, $m_j^{(v)}$ denotes v-th block of the global centroid. Then we define $d_{ij}^{(v)} = \|C_i^{(v)} - m_j^{(v)}\|_2^2$, $D_{ij} = \sum_{v=1}^V d_{ij}^{(v)}$, $f_j = \sum_{i=1}^N q_{ij}$. For finite representations and centroids, these assignments are strictly positive.

Target distribution To stabilize training and highlight high-confidence assignments, we refine q_{ij} into a target distribution π_{ij} by

$$\pi_{ij} = \frac{q_{ij}^2/\sum_i q_{ij}}{\sum_{k=1}^K \left(q_{ik}^2/\sum_i q_{ik}\right)} = \frac{q_{ij}^2/f_j}{\sum_{k=1}^K q_{ik}^2/f_k}. \tag{4}$$

This reweighting sharpens confident predictions while downweighting ambiguous ones, producing a more discriminative and consistent target for clustering. We assume probabilities are clipped to $[\varepsilon, 1]$ so all logs are finite.

SCL loss The clustering-driven objective, designed to achieve semantic-invariant consensus, is formulated as the Kullback–Leibler divergence between the target distribution Π and the current soft assignment $Q^{(v)}$:

$$\mathcal{L}_{\mathrm{Scl}}^{(v)} = \sum_{i=1}^N \sum_{j=1}^K \pi_{ij} \log \frac{\pi_{ij}}{q_{ij}^{(v)}}. \tag{5}$$

The overall SCL loss is obtained by averaging across all views.

Lemma 1 (Log-odds under target refinement)**.** *For any sample i and clusters j, k,*

$$\log \frac{\pi_{ij}}{\pi_{ik}} = 2 \log \frac{q_{ij}}{q_{ik}} - \log \frac{f_j}{f_k}. \tag{6}$$

Thus, refinement increases the log-odds in favor of j exactly when $q_{ij}/q_{ik} > f_j/f_k$; equal cluster frequencies double the original log-odds.

Proof. In π_{ij}/π_{ik}, the sample-wise normalizer cancels, leaving $(q_{ij}/q_{ik})^2(f_k/f_j)$. □

Lemma 2 (View contributions to the consensus)**.** *For any clusters j, k, the shared target satisfies*

$$\log \frac{\pi_{ij}}{\pi_{ik}} = (\beta + 1) \log \left(1 + \frac{\sum_v (d_{ik}^{(v)} - d_{ij}^{(v)})}{\beta + D_{ij}}\right) - \log \frac{f_j}{f_k}. \tag{7}$$

Proof. Concatenation gives $D_{ij} = \sum_v d_{ij}^{(v)}$. The Student's t kernel gives $\log(q_{ij}/q_{ik}) = \frac{\beta+1}{2}\log[(\beta + D_{ik})/(\beta + D_{ij})]$. Substitute this identity into Lemma 1 proves the result. □

Theorem 1 (Conditional discriminative-view dominance)**.** *Fix a sample i and a candidate cluster $j^\star$. Suppose a view $v^\star$ provides a positive distance margin against every competing cluster:*

$$\delta_i := \min_{k \neq j^\star} \left(d_{ik}^{(v^\star)} - d_{ij^\star}^{(v^\star)} \right) > 0. \tag{8}$$

Define the maximum opposing contribution from the remaining views and the cluster-frequency correction as

$$\Gamma_i = \max_{k \neq j^\star} \sum_{v \neq v^\star} \left(d_{ij^\star}^{(v)} - d_{ik}^{(v)} \right), \quad B_i = \max_{k \neq j^\star} \log \frac{f_{j^\star}}{f_k}. \tag{9}$$

If $\delta_i > \Gamma_i$, then

$$\log \pi_{ij^\star} - \max_{k \neq j^\star} \log \pi_{ik} \geq (\beta + 1) \log \left(1 + \frac{\delta_i - \Gamma_i}{\beta + D_{ij^\star}} \right) - B_i. \tag{10}$$

In particular, $j^\star$ is the unique consensus winner whenever the right-hand side is positive.

Proof. For every $k \neq j^\star$, $D_{ik} - D_{ij^\star} = d_{ik}^{(v^\star)} - d_{ij^\star}^{(v^\star)} + \sum_{v \neq v^\star} \left(d_{ik}^{(v)} - d_{ij^\star}^{(v)} \right) \geq \delta_i - \Gamma_i$. By Lemma 2 and monotonicity of the logarithm,

$$\log \frac{\pi_{ij^\star}}{\pi_{ik}} = (\beta + 1) \log \left(1 + \frac{D_{ik} - D_{ij^\star}}{\beta + D_{ij^\star}} \right) - \underbrace{\log \frac{f_{j^\star}}{f_k}}_{\leq B_i} \tag{11}$$

Taking the minimum over competing clusters proves (10). Under (11), all these log-ratios are strictly positive, so $\arg\max_j \pi_{ij} = j^\star$ uniquely. □

Corollary 1 (Consensus-label alignment)**.** *Suppose the global target has a unique winning cluster $j^\star$, with probability margin $\Delta_i = \pi_{ij^\star} - \max_{k \neq j^\star} \pi_{ik} > 0$. For $\varepsilon_i^{(v)} = \mathrm{KL}(\pi_i \| q_i^{(v)})$,*

$$q_{ij^\star}^{(v)} - \max_{k \neq j^\star} q_{ik}^{(v)} \geq \Delta_i - \sqrt{2\varepsilon_i^{(v)}}, \tag{12}$$

v-th assignment has the same unique winning cluster as the global target when $\varepsilon_i^{(v)} < \Delta_i^2/2$.

Proof. Using natural logarithms, Pinsker's inequality gives $\|\pi_i - q_i^{(v)}\|_1 \leq \sqrt{2\varepsilon_i^{(v)}}$. For every $k \neq j^\star$,

$$q_{ij^\star}^{(v)} - q_{ik}^{(v)} \geq \pi_{ij^\star} - \pi_{ik} - |q_{ij^\star}^{(v)} - \pi_{ij^\star}| - |q_{ik}^{(v)} - \pi_{ik}| \geq \Delta_i - \underbrace{\|q_i^{(v)} - \pi_i\|_1}_{\leq \sqrt{2\varepsilon_i^{(v)}}}. \tag{13}$$

The claimed label alignment follows when the final lower bound is strictly positive. □

Discussion. Lemma 1 shows how squared assignments and cluster frequencies shape the global target, while Lemma 2 links its preferences to view-wise contributions. Theorem 1 identifies when one view's distance advantage determines the consensus despite conflicting evidence. Together with Corollary 1, these results characterize conditional consensus formation and cross-view alignment.

3.3 HARDNESS-AWARE STRUCTURE EXPLORATION

The Hardness-aware structure exploration(HaSE) uses consensus-derived pseudo-labels to identify challenging sample relationships and refine the discriminative structure of the embedding space. To accomplish this, HaSE aggregates the deep and shallow representations across views and then select the least similar positive and the most similar negatives for each valid anchor, jointly addressing within-cluster dispersion and between-cluster ambiguity through a contrastive objective.

Consensus-driven pseudo labels and anchors For sample i and view v, let $q_i^{(v)} \in \Delta^{K-1}$ be the per-view soft assignment (Sec. 3.3). We aggregate views by

$$\bar{q}_i \;=\; \frac{1}{V}\sum_{v=1}^{V} q_i^{(v)} \;\in\; \Delta^{K-1}, \tag{14}$$

and define the pseudo label and its confidence as

$$\hat{y}_i \;=\; \arg\max_j \bar{q}_{ij}, \qquad \kappa_i \;=\; \max_j \bar{q}_{ij}. \tag{15}$$

With a threshold $\tau \in (0,1)$, the anchor set is $\mathcal{A} \;=\; \{\, i \;:\; \kappa_i \geq \tau \,\}$.

Fused embedding and cosine similarity We form a fused representation by concatenating deep feature $C_i^{(v)}$ and shallow feature $S_i^{(v)}$, followed by ℓ_2 normalization:

$$\begin{aligned} u_i \;&=\; \mathrm{norm}\Big(\big[\,\bar{C}_i\,;\,\textstyle\sum_{v=1}^{V} a_v\, S_i^{(v)}\,\big]\Big), \\ \bar{C}_i \;&=\; \tfrac{1}{V}\displaystyle\sum_{v=1}^{V} C_i^{(v)}, \quad a \;=\; \mathrm{softmax}(\alpha w) \in \mathbb{R}^V. \end{aligned} \tag{16}$$

Stack a mini-batch as $U = [u_1, \ldots, u_B]^\top$. The cosine similarity matrix (dot product after normalization) is

$$\mathbf{S} \;=\; UU^\top, \qquad S_{im} = [\mathbf{S}]_{im} \in [-1,1], \qquad S_{ii} = -\infty. \tag{17}$$

For scale stabilization, min–max normalize off-diagonal entries to $[0,1]$:

$$\tilde{\mathbf{S}} = \frac{\mathbf{S} - s_{\min}\mathbf{1}\mathbf{1}^\top}{s_{\max} - s_{\min}} \begin{pmatrix} s_{\min} = \min\limits_{i \neq m} S_{im}, \\ s_{\max} = \max\limits_{i \neq m} S_{im}, \\ [\tilde{\mathbf{S}}]_{ii} = 0 \end{pmatrix}. \tag{18}$$

Class masks and hard-example mining Define class-consistency masks :

$$\mathrm{same}_{im} \;=\; \mathbb{I}[\hat{y}_i = \hat{y}_m], \qquad \mathrm{neg}_{im} \;=\; \mathbb{I}[\hat{y}_i \neq \hat{y}_m]. \tag{19}$$

For each anchor $i \in \mathcal{A}$, select the *hardest positive* and the *top-k hardest negatives* row-wise:

$$\begin{pmatrix} s_{\mathrm{pos}}(i) = \min\{\, S_{im} : \mathrm{same}_{im} = 1 \,\}, \\ \{s_{\mathrm{neg},j}(i)\}_{j=1}^{k'} = \text{Top-}k\{\, S_{im} : \mathrm{neg}_{im} = 1 \,\}, \;\; k' \leq k \end{pmatrix}. \tag{20}$$

If no positive exists for i, the anchor is skipped. This selection focuses learning on samples that share a pseudo-label but remain dissimilar, and on samples with different pseudo-labels that remain geometrically close.

HaSE objective(temperature-scaled contrastive loss) Given a temperature $\tau_\ell > 0$, the *per-anchor Hardness-aware Structure Exploration objective* is

$$\ell_i^{\mathrm{HaSE}} = -\log \frac{\exp\big(s_{\mathrm{pos}}(i)/\tau_\ell\big)}{\exp\big(s_{\mathrm{pos}}(i)/\tau_\ell\big) + \sum_{j=1}^{k_i} \exp\big(s_{\mathrm{neg},j}(i)/\tau_\ell\big)}. \tag{21}$$

The objective encourages the selected positive to become more similar to the anchor than the selected negatives, thereby aligning the geometry of the fused representations with the clustering relationships inferred from the consensus.

An anchor is valid only when it admits both a positive and a negative candidate. Accordingly, we define

$$\mathcal{I}_{\mathrm{valid}} = \left\{ i \in \mathcal{A} \;\middle|\; \begin{matrix} \exists\, m \neq i :\; \hat{y}_m = \hat{y}_i, \\ \exists\, n \neq i :\; \hat{y}_n \neq \hat{y}_i \end{matrix} \right\}. \tag{22}$$

The *batch* objective is then

$$\mathcal{L}_{\mathrm{HaSE}} = \frac{1}{|\mathcal{I}_{\mathrm{valid}}|} \sum_{i \in \mathcal{I}_{\mathrm{valid}}} \ell_i^{\mathrm{HaSE}}. \tag{23}$$

When $\mathcal{I}_{\mathrm{valid}}$ is empty, we set $\mathcal{L}_{\mathrm{HaSE}} = 0$. Overall, The total loss of CORE is as follows:

$$L \;=\; L_{\mathrm{rec}} + \lambda_{\mathrm{Scl}} L_{\mathrm{Scl}} + \lambda_{\mathrm{HaSE}} L_{\mathrm{HaSE}}, \tag{24}$$

where λ_{Scl} and λ_{HaSE} are trade-off parameters.

4 EXPERIMENT

4.1 EXPERIMENT SETTINGS

Benchmark datasets: This experiment utilizes six benchmark datasets: MSRC (Winn & Jojic, 2005), ORL (Samaria & Harter, 1994), Caltech-4V (Fei-Fei et al., 2004), Caltech-5V (Fei-Fei et al., 2004), LandUse-21 (Yang & Newsam, 2010) and ALOI-100(Geusebroek et al., 2005). Together, these datasets span different sample sizes, class counts, and feature dimensions, enabling evaluation across varied multi-view configurations. The details of the six multi-view datasets is shown in Table 1.

Table 1: Details of the datasets.

Dataset	Samples	Classes	Views	Dimensions
MSRC	210	7	5	24/576/512/256/254
ORL	400	40	4	256/256/256/256
Caltech-4V	1400	7	4	40/254/928/512
Caltech-5V	1400	7	5	40/254/928/512/1984
LandUse-21	2100	21	2	59/40
ALOI-100	10800	100	4	77/13/64/125

Comparison methods: Nine methods are selected as baselines, i.e., AEVC (CVPR,24) (Liu et al., 2024), MVCAN (CVPR,24) (Xu et al., 2024), PMIMC (TIP,25) (Yuan et al., 2025), ALPC (AAAI,25) (Chen et al., 2025b), DMAC (AAAI,25) (Wang et al., 2025a), STCMC (AAAI,25) (Hu et al., 2025b), MSDIB (AAAI,25) (Hu et al., 2025a), BONE (AAAI,26) (Xu et al., 2026), BMvC (TPAMI,26) (Li et al., 2026). For the fair comparison, nine methods are implemented in strict accordance with the configurations detailed in their original publications, and the grid search experiments of trade-off parameters are conducted to elicit their best performance.

Evaluation metrics: Three common evaluation metrics, i.e., accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI), are selected to validate the cutting-edge performance of CORE, and a higher value on the three metrics signifies superior performance.

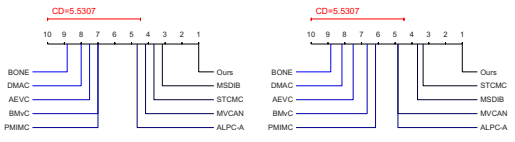
Table 2: Clustering results (Mean ± Std %) on six benchmark datasets. The best results are highlighted in **bold**, and the second-best results are underlined.

	MSRC			ALOI			Caltech-4V		
Method	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
AEVC	70.87±2.48	64.96±1.39	54.48±2.52	66.01±2.58	80.92±2.44	53.44±2.61	64.71±2.54	53.13±2.42	46.00±2.57
MVCAN	87.14±1.21	79.23±1.25	75.21±1.18	71.50±2.47	84.62±2.36	60.12±2.51	76.29±1.33	67.05±2.41	61.95±2.38
PMIMC	74.29±2.52	69.49±2.44	59.76±2.55	75.72±2.41	90.35±1.18	69.74±1.37	80.43±1.29	67.30±2.35	62.52±2.42
ALPC-A	83.81±2.38	72.64±2.41	67.40±2.49	83.86±1.22	90.79±1.15	76.86±1.24	75.07±2.45	64.35±2.51	60.51±2.47
DMAC	68.57±2.61	59.31±2.55	47.71±2.63	77.18±2.44	86.98±1.31	67.43±2.48	70.43±2.55	54.72±2.62	50.58±2.59
STCMC	79.05±2.42	80.40±1.33	71.65±2.41	82.52±1.28	89.52±1.19	75.23±1.31	84.93±1.21	77.26±1.17	73.25±1.25
MSDIB	85.71±2.65	85.14±2.71	82.36±2.68	83.03±1.25	88.39±1.22	73.57±1.29	78.93±2.38	69.84±2.44	63.17±2.41
BONE	85.71±9.68	84.95±8.04	84.59±9.93	69.96±1.85	84.07±1.10	60.71±2.98	69.34±5.71	62.68±3.33	53.77±6.26
BMvC	88.64±4.32	86.51±3.19	84.10±4.92	84.50±0.61	89.02±0.35	73.34±0.88	84.63±3.00	75.14±1.98	72.43±1.95
Ours	**94.76±1.65**	**90.17±1.54**	**88.23±1.77**	**88.76±1.86**	**93.31±1.53**	**83.56±1.92**	**86.07±0.71**	**77.39±0.62**	**74.70±0.84**

	Caltech-5V			ORL			LandUse		
Method	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
AEVC	83.71±2.41	70.37±2.48	68.42±2.45	64.72±2.42	81.31±1.29	53.63±2.48	23.90±2.62	24.21±2.68	10.48±2.71
MVCAN	87.93±1.25	79.25±1.29	76.02±1.27	67.25±2.38	81.10±2.31	53.61±2.41	25.19±2.45	31.86±1.38	11.27±2.55
PMIMC	82.57±2.47	71.34±2.52	67.08±2.51	56.25±2.58	76.21±2.49	42.87±2.64	22.52±2.68	28.05±2.65	9.14±2.74
ALPC-A	86.29±2.35	74.82±2.44	74.52±2.39	64.25±2.45	80.10±2.36	49.69±2.52	24.33±2.51	31.17±2.44	11.93±2.59
DMAC	73.57±2.62	59.79±2.68	54.03±2.72	58.50±2.51	76.50±2.44	43.62±2.59	23.81±2.58	29.30±2.55	10.63±2.62
STCMC	90.14±1.19	82.91±1.21	80.66±1.24	68.55±1.31	81.24±1.28	64.09±1.35	24.24±2.53	30.08±2.48	12.76±1.41
MSDIB	88.15±1.24	80.72±1.28	77.25±1.26	65.42±2.41	77.34±2.45	50.08±2.49	28.76±2.61	35.44±2.52	14.69±2.45
BONE	78.49±2.51	70.05±2.49	64.84±3.25	57.20±1.03	76.17±0.33	43.24±1.07	25.13±0.50	30.86±0.72	10.41±0.47
BMvC	90.07±2.28	80.02±2.72	77.07±3.69	76.22±1.85	87.61±1.09	68.21±2.21	28.96±0.69	33.76±0.79	14.98±0.53
Ours	**95.00±1.58**	**90.82±1.51**	**89.00±1.69**	**82.75±2.72**	**90.76±2.54**	**75.68±2.81**	**30.86±0.94**	**37.95±0.68**	**15.49±0.97**

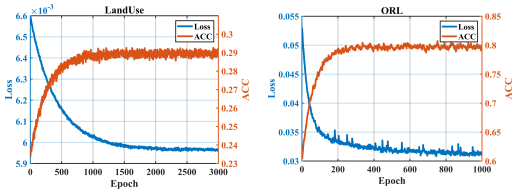
4.2 CLUSTERING PERFORMANCE

Table 2 records numerical clustering results of CORE and comparison methods on six dataset. The experimental results demonstrate that: (1) our method achieves the best clustering performance among all compared methods across the six datasets, consistently ranking first in ACC, NMI, and ARI. On the challenging LandUse dataset, it achieves an NMI of 37.95%, outperforming the runner-up by 2.51 percentage points. (2) Our method achieves particularly pronounced gains on Caltech-5V and ORL. On Caltech-5V, it obtains 95.00% ACC and 90.82% NMI, exceeding the strongest com-



(a) ACC (b) NMI

Figure 2: Nemenyi tests across six datasets.



(a) LandUse-21 (b) ORL

Figure 3: Convergence on two datasets.

peting results by 4.86 and 7.91 percentage points, respectively. On ORL, it improves ACC and ARI by 6.53 and 7.47 percentage points, respectively, over the runner-up. These observations support the effectiveness of the proposed framework: dual-representation learning separates semantic information from complementary information to facilitate cross-view consensus, while HaSE dynamically emphasizes hard sample pairs to enhance discriminative learning.

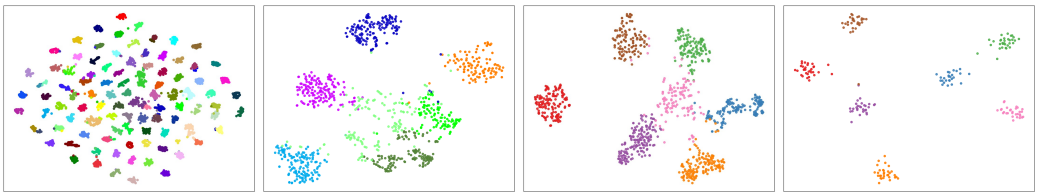
Meanwhile, post-hoc Nemenyi tests are executed according to the average ACC, NMI and ARI on six datastets to verify the statistical superiority of CORE. As illustrated in Figure 2, the results demonstrate that CORE significantly outperforms all nine comparison methods at a significance level of $p = 0.05$.

4.3 The Ablation Study

To assess the contribution of each component, we construct three ablation variants: (i) CORE-S removes the shallow branch, its reconstruction loss, and its contribution to feature fusion, using only deep representations; (ii) CORE-C removes $\mathcal{L}_{\text{Scl}}$; and (iii) CORE-H removes $\mathcal{L}_{\text{HaSE}}$. Table 3 show that removing any single component leads to performance degradation, confirming their individual necessity. Specifically, the variant "CORE-C" achieves lower performance than the full model on all datasets, indicating that "CORE-C" plays a critical role in enhancing cluster discrimination. Similarly, "CORE-S" and "CORE-H" also underperform the full model, with the largest drops observed in NMI and ARI, suggesting these losses contribute significantly to semantic alignment. Our method consistently achieves the highest scores across all metrics and datasets, demonstrating that the synergy of all components is essential for achieving optimal clustering performance. This ablation analysis validates the design choices and underscores the complementary benefits of each component in improving representation learning.

Table 3: Ablation results (%) on three benchmark datasets. The best results are shown in **bold**.

	MSRC			ALOI			Caltech-5V		
Variant	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
CORE-S	92.33	89.21	87.66	86.74	92.41	83.24	93.77	88.62	87.06
CORE-C	91.43	84.05	81.08	85.44	92.30	80.06	90.43	82.15	80.65
CORE-H	93.81	87.22	85.96	75.98	85.53	66.91	93.71	87.75	86.11
CORE	**94.76**	**90.17**	**88.23**	**88.76**	**93.31**	**83.56**	**95.00**	**90.82**	**89.00**

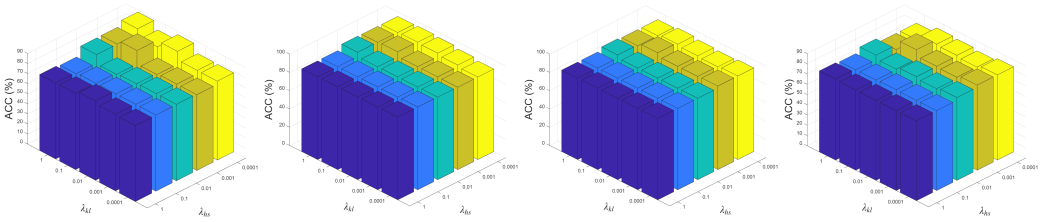


(a) ALOI (b) Caltech-4V (c) Caltech-5V (d) MSRC

Figure 4: The t-SNE visualization results.

4.4 The Comprehensive Study

Convergence: To investigate the convergence behavior of CORE, we report the objective loss and clustering performance versus the number of iterations on the LandUse and ORL datasets in Fig-



(a) Caltech-4V (b) Caltech-5V (c) MSRC (d) ORL

Figure 5: The analysis results of the trade-off parameters λ_{Scl} and λ_{HaSE}.

ure 3. As the optimization proceeds, the loss decreases steadily and gradually approaches a stable value, while the corresponding clustering metric increases rapidly before reaching a plateau. Specifically, the algorithm converges after approximately 1,500 iterations on LandUse and 200 iterations on ORL. The minor fluctuations observed during the later stage do not affect the overall convergence trend. These results empirically demonstrate that the proposed alternating optimization algorithm exhibits stable convergence behavior on different datasets.

Visualization: Figure 4 presents t-SNE visualizations of the clustering results on four datasets, with each point colored by its assigned cluster label. ALOI shows clear group boundaries and distinct, localized clusters, indicating good separability among identities. For Caltech-4V/5V, the visualizations reveal highly structured and well-separated clusters with minimal overlap, suggesting that the proposed method effectively preserves semantic consistency across views. MSRC also exhibits relatively well-defined clusters, despite slight intermixing in some regions, consistent with its multi-label nature. Overall, these visualizations corroborate the quantitative results: datasets with higher clustering performance, such as Caltech-4V/5V, form more compact and separated clusters, validating the method's ability to enhance discriminative structures in the embedding space.

Parameter analysis: Figure 5 illustrates the parameter sensitivity of the proposed clustering framework on four datasets by showing ACC under different values of λ_{Scl} and λ_{HaSE}. As shown in the 3D bar plots (a)–(d), ACC generally increases when the two parameters vary from small to moderate values, reaching its peak around (λ_{Scl}=1) and ($\lambda_{\mathrm{HaSE}} \in [10^{-4}, 10^{-3}]$). Beyond these ranges, the performance tends to stabilize or slightly decline, indicating that a proper balance between feature alignment and contrastive regularization is important. Moreover, the gradual performance variation suggests that the parameters mainly stabilize optimization rather than causing drastic changes. Similar trends are observed across all datasets, demonstrating the robustness and generalizability of the framework. Overall, the proposed method is not overly sensitive to parameter selection within reasonable ranges, supporting its practical applicability.

5 Conclusion

In this work, we presented CORE, a deep multi-view clustering framework that combines semantic and structural information through dual representation learning. Deep representations form global clustering target guides semantic invariance across views, while shallow representations contribute complementary cues to the fused embedding. Hardness-aware structure exploration (HaSE) uses consensus-derived pseudo-labels to identify hard positive and negative samples, refining both within-cluster cohesion and between-cluster separation. Experiments on six benchmark datasets show improved clustering performance over nine comparison methods. Future work will examine how conflicting views and unreliable pseudo-labels affect its quality.

AI Use Statement

No generative AI tools are used in any stage of this work, including research design, methodology development, experimental implementation, data analysis, and manuscript preparation. All content is independently produced and reviewed by the authors. We take full responsibility for the final content of this work.

Reproducibility statement

To ensure reproducibility, we provide the source code in the Supplementary Material and will make it publicly available upon acceptance of the paper.

References

Jintang Bian, Yixiang Lin, Xiaohua Xie, Chang Dong Wang, Lingxiao Yang, Jian Huang Lai, and Feiping Nie. Multilevel contrastive multiview clustering with dual self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10422–10436, 2025.

Haishun Chen, Cai Xu, Ziyu Guan, Wei Zhao, and Jinlong Liu. Biased incomplete multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15767–15775, 2025a.

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

Yawei Chen, Huibing Wang, Jinjia Peng, and Yang Wang. Anchor learning with potential cluster constraints for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15939–15947, 2025b.

Jinrong Cui, Yuting Li, Han Huang, and Jie Wen. Dual contrast-driven deep multi-view clustering. *IEEE Transactions on Image Processing*, 33:4753–4764, 2024.

Zhibin Dong, Dayu Hu, Jiaqi Jin, Siwei Wang, Xinwang Liu, and En Zhu. Selective cross-view topology for deep incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 2025.

Lunke Fei, Junlin He, Qi Zhu, Shuping Zhao, Jie Wen, and Yong Xu. Deep multi-view contrastive clustering via graph structure awareness. *IEEE Transactions on Image Processing*, 34:3805–3816, 2025.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Computer Vision and Image Understanding*, pp. 178–178, 2004.

Jing Gao, Meng Liu, Peng Li, Jianing Zhang, and Zhikui Chen. Deep multiview adaptive clustering with semantic invariance. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9): 12965–12978, 2024.

Chuanxing Geng, Aiyang Han, and Songcan Chen. Explicit view-labels matter: A multifacet complementarity study of multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2520–2532, 2025.

JM Geusebroek, GJ Burghouts, and AWM Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

Shizhe Hu, Jiahao Fan, Guoliang Zou, and Yangdong Ye. Multi-aspect self-guided deep information bottleneck for multi-modal clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17314–17322, 2025a.

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, 2025b.

Shizhe Hu, Chengkun Zhang, Guoliang Zou, Zhengzheng Lou, and Yangdong Ye. Deep multi-view clustering by pseudo-label guided contrastive learning and dual correlation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3646–3658, 2025c.

Mudi Jiang, Lianyu Hu, Zengyou He, and Zhikui Chen. Interpretable multi-view clustering. *Pattern Recognition*, 162:111418, 2025.

Mengcheng Lan, Min Meng, Jun Yu, and Jigang Wu. Generalized multi-view collaborative subspace clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3561–3574, 2022.

Zhenglai Li, Jun Wang, Chang Tang, Xinzhong Zhu, Wei Zhang, and Xinwang Liu. Balanced multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Meng Liu, Yuzhe Li, Zhikui Chen, and Yilong Lin. Information-driven complementarity and consistency mining for multi-view clustering. *IEEE Signal Processing Letters*, 33:216–220, 2026.

Suyuan Liu, Ke Liang, Zhibin Dong, Siwei Wang, Xihong Yang, Sihang Zhou, En Zhu, and Xinwang Liu. Learn from view correlation: An anchor enhancement strategy for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26151–26161, 2024.

Shubin Ma, Zhikui Chen, Lin Yuanbo Wu, and Liang Zhao. Multi-view aligned clustering via sample-bundled optimization: Anchor graph enhancement and contrastive propagation. *IEEE Transactions on Multimedia*, pp. 1–12, 2026.

Ferdinand Samaria and Andy Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of Second IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 138–142, 1994.

Long Shi, Lei Cao, Jun Wang, and Badong Chen. Enhanced latent multi-view subspace clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):12480–12495, 2024.

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

Bocheng Wang, Chusheng Zeng, Mulin Chen, and Xuelong Li. Towards learnable anchor for deep multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21044–21052, 2025a.

Suixue Wang, Shilin Zhang, Qingchen Zhang, Peng Li, and Weiliang Huo. Master: A multi-granularity invariant structure clustering scheme for multi-view clustering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6415–6423, 2025b.

John M. Winn and Nebojsa Jojic. Locus: Learning object classes with unsupervised segmentation. In *Proceedings of International Conference on Computer Vision*, pp. 756–763, 2005.

Jianlong Wu, Xingxu Xie, Liqiang Nie, Zhouchen Lin, and Hongbin Zha. Unified graph and low-rank tensor learning for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6388–6395, 2020.

Baili Xiao, Zhibin Dong, Ke Liang, Suyuan Liu, Siwei Wang, Tianrui Liu, Xingchen Hu, En Zhu, and Xinwang Liu. Easemvc:efficient dual selection mechanism for deep multi-view clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 20716–20726, 2025.

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

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

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22957–22966, 2024.

Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19863–19872, 2023.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

Zuyuan Yang, Naiyao Liang, Wei Yan, Zhenni Li, and Shengli Xie. Uniform distribution non-negative matrix factorization for multiview clustering. *IEEE Transactions on Cybernetics*, 51(6): 3249–3262, 2021.

Yugen Yi, Ningyi Zhang, Zehui Zhang, Yijian Fu, Lei Chen, and Jianzhong Wang. Emlfcl: An efficient multilevel fusion contrastive learning for multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):14228–14242, 2025.

Jun Yin, Pei Wang, Shiliang Sun, and Zhonglong Zheng. Incomplete multi-view clustering via multi-level contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

Honglin Yuan, Yuan Sun, Fei Zhou, Jing Wen, Shihua Yuan, Xiaojian You, and Zhenwen Ren. Prototype matching learning for incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 34:828–841, 2025.

Helin Zhao, Wei Chen, and Peng Zhou. Active deep multi-view clustering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 5554–5562, 2024.

Wei Zhao, Cai Xu, Ziyu Guan, and Ying Liu. Multiview concept learning via deep matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):814–825, 2021.

Guo Zhong and Chi-Man Pun. Latent low-rank graph learning for multimodal clustering. In *Proceedings of the IEEE International Conference on Data Engineering*, pp. 492–503, 2021.

Zhengzhong Zhu, Pei Zhou, Zeting Li, Kejiang Chen, and Jiangping Zhu. Multi-label text classification with label attention aware and correlation aware contrastive learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 8420–8428, 2025.