DECOUPLED CONDITIONAL ALIGNMENT FOR PARTIALLY VIEW-ALIGNED CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Partially View-aligned Clustering (PVC) aims to handle complex multi-view data where only a small subset of samples exhibits strict cross-view correspondence. However, existing methods often overlook intra-view geometric topology during the feature extraction process. Furthermore, blindly aligning massive unpaired samples risks false-negative repulsion, leading to severe semantic confusion and feature manifold collapse. To address these critical issues, we propose a divide-and-conquer framework termed Decoupled Conditional Alignment (DeCA). DeCA first employs a dual-decoder Graph Autoencoder (GAE) to preserve the native intra-view neighborhood connectivity for extracting structure-preserved latent representations. At the sample level, it constructs stable global semantic anchors via strong cross-view contrast on aligned samples. For unaligned samples, pseudo-label-guided conditional Maximum Mean Discrepancy (MMD) and intra-view neighborhood contrast ensure intra-class sub-distribution matching without cross-view pairings. Finally, a decoupled feature fusion strategy is introduced to mitigate manifold contamination caused by superimposing heterogeneous misaligned features. Extensive experiments on ten benchmark datasets demonstrate DeCA's superiority over various PVC baselines.

1 INTRODUCTION

Unsupervised representation learning heavily relies on Multi-View Clustering (MVC) to distill consensus semantics across heterogeneous modalities (Fang et al., 2023; Qin et al., 2025b). However, the traditional MVC paradigm fundamentally hinges on an idealized premise: a perfect, instance-level bijective mapping across the majority of views (Qin et al., 2025a; Gao et al., 2026). In practical deployments—ranging from asynchronous sensor fusion in autonomous driving to uncoupled multimodal medical imaging—this rigid alignment is frequently compromised, leaving only a sparse subset of instances paired (Yang et al., 2025; Wang et al., 2025). This structural degradation has catalyzed the shift toward Partially View-aligned Clustering (PVC). The primary bottleneck in PVC is bridging the semantic gap for the vast majority of samples lacking explicit cross-view correspondence (Wang et al., 2024; Lin et al., 2026).

The trajectory of PVC research has rapidly evolved from shallow structural formulations to deep representation paradigms. Early efforts primarily focused on graph-theoretic imputations and matrix factorization to deduce missing topologies (Chao et al., 2025; Zhao et al., 2025). However, bottlenecked by limited representational capacity, recent literature is dominated by deep learning architectures (Huang et al., 2020). Prevailing deep PVC frameworks typically cascade autoencoders with cross-view alignment engines, mapping heterogeneous inputs into a shared latent space to progressively calibrate semantic embeddings via end-to-end optimization (Sun et al., 2025b; Qian et al., 2024).

Despite these methodological advances, current deep PVC paradigms still face notable limitations in manifold learning and distribution synchronization. First, a heavy reliance on sparse, paired anchors frequently overlooks the geometric topologies intrinsic to individual views, which can compromise the structural integrity of view-specific embeddings (Zhao et al., 2024a; Cao et al., 2024). Second, directly applying global alignment to unaligned instances is prone to introducing false-negative penalties. This forced matching can affect intra-class compactness and lead to feature contamination during downstream fusion (Deng et al., 2025; Guo et al., 2026).

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.

To circumvent these limitations, we formulate a decoupled architecture, termed Decoupled Conditional Alignment (DeCA) (depicted in Figure 1). Our primary contributions are delineated as follows:

- We introduce a dual-decoder Graph Autoencoder (GAE) that preserves intrinsic intra-view geometric topologies alongside feature attributes, establishing a structurally robust latent space prior to cross-view intervention.
- We devise a sample-level decoupled alignment engine: strictly aligned instances are anchored globally via robust cross-view contrast, whereas unaligned instances undergo a pseudo-label-guided conditional Maximum Mean Discrepancy. This probabilistic matching synchronizes local sub-distributions without requiring explicit cross-view pairings.
- We propose an asymmetric feature fusion protocol to mitigate manifold contamination. It selectively averages aligned representations while maintaining unaligned samples within the template view's coordinate reference.
- Empirical evaluations across ten benchmark datasets validate that DeCA significantly outperforms various existing PVC baselines, demonstrating substantial clustering robustness across varying degrees of view misalignment.

2 Related Work

2.1 Deep Multi-View Clustering

The evolution of deep multi-view clustering (MVC) has been largely driven by the capacity of neural networks to extract non-linear consensus representations (Fang et al., 2025; Bao et al., 2025). Early paradigms relied heavily on autoencoder-based joint reconstruction to forge a shared latent subspace (Wen et al., 2023), which was later augmented by graph neural networks and generative models to delineate sharper decision boundaries (Yu et al., 2025). Currently, cross-view contrastive learning dominates the field, optimizing instance-level discrimination by maximizing the mutual information between paired cross-view samples (Wang et al., 2024; Lin et al., 2026). However, the efficacy of these contrastive frameworks inherently presupposes perfect cross-view correspondence. When deployed in partially view-aligned scenarios, the indiscriminate propagation of supervisory signals from misaligned pairs inevitably introduces semantic noise, severely degrading the learned representations (Hou et al., 2026).

2.2 Partially View-Aligned Clustering

Partially View-Aligned Clustering (PVC) relaxes the rigid bijective assumption of traditional MVC, aiming to infer semantic structures when only a limited fraction of cross-view correspondence is accessible. To bypass the limitations of shallow structural matching, contemporary PVC research has pivoted toward deep representation learning. For instance, He et al. (2024) introduced a variational architecture to disentangle common and view-specific factors under incomplete alignment. To mitigate the impact of faulty pairings, Chen et al. (2026) proposed an iterative positive-pair mining strategy regulated by quantile-adaptive thresholds. Alternatively, Deng et al. (2026) formulated the alignment problem through a game-theoretic lens, utilizing marginal coalition contributions to reduce reliance on pre-aligned anchors. Despite these methodological innovations, existing deep PVC models predominantly focus on instance-level recalibration. Consequently, they often overlook the preservation of intrinsic intra-view geometric manifolds and remain vulnerable to false-negative repulsion during indiscriminate global alignment.

2.3 Partially View-Aligned Clustering

Partially View-Aligned Clustering (PVC) is a generalization of Multi-View Clustering (MVC) under incomplete cross-view correspondence, where only a limited fraction of samples possess known one-to-one correspondence across views. In this context, deep learning has emerged as the mainstream paradigm, with recent works striving to transcend the limitations of simple instance-level matching. For instance, He et al. (2024) employ a variational framework to simultaneously learn

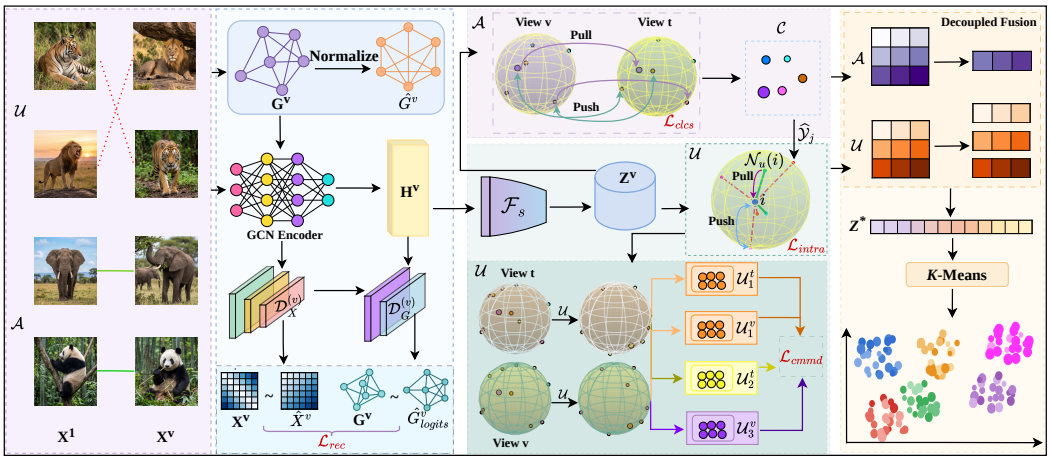

Figure 1: The main framework of DeCA. (1) Dual-decoder GAEs extract structure-preserving embeddings. (2) Cross-view contrast establishes global anchors for aligned instances. (3) Unaligned instances are matched via intra-view neighborhood contrast and conditional Maximum Mean Discrepancy. (4) Decoupled feature fusion mitigates manifold contamination prior to clustering.

common and view-specific information, achieving robust clustering under partially unaligned conditions. Chen et al. (2026) introduce a quantile-adaptive threshold and a uniqueness constraint to iteratively mine reliable cross-view positive pairs while suppressing the negative impacts caused by faulty pairs. From a game-theoretic perspective, Deng et al. (2026) model multi-view samples as game players and quantify the marginal contributions of coalition members; this fundamentally alleviates the reliance on pre-aligned samples and enables robust clustering even under severe view misalignment.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Given a multi-view dataset $\mathcal{X} = \{\mathbf{X}^{(v)} \in \mathbb{R}^{N \times d_v}\}_{v=1}^{V}$ with V views, where N is the total number of samples and d_v is the feature dimension of the v-th view. In the Partially View-aligned Clustering (PVC) task, only a fraction of the samples maintains a strict one-to-one correspondence across different views. Therefore, based on whether the samples are aligned across views, the dataset is partitioned into two mutually exclusive subsets: an aligned subset $\mathcal{A}$ and an unaligned subset $\mathcal{U}$, where the sum of their sample sizes equals the total number of samples N. Ultimately, we partition the N samples into K clusters.

3.2 MOTIVATION

DeCA is specifically tailored for partially view-aligned scenarios, systematically addressing three critical limitations inherent in this setting. First, traditional independent sample reconstruction mechanisms ignore the inherent geometric structures of individual views. To mitigate structural distortion, we construct a dual-decoder Graph Autoencoder (GAE) that leverages neighborhood constraints to strictly preserve native intra-view neighborhood connectivity. Second, to prevent the interference features of unaligned samples from misleading network feature propagation, we design a strong cross-view contrastive learning mechanism exclusively for aligned samples to establish preliminary robust semantic anchors. Third, unaligned samples lack direct cross-view pairings; directly executing contrastive learning on them would introduce massive false negatives, leading to erroneous repulsion. To address this, we propose a local probabilistic conditional alignment module: by mining reliable neighborhood structures within single views and utilizing pseudo-label-guided

conditional Maximum Mean Discrepancy, we achieve the cross-view matching of intra-class sub-distributions.

3.3 TOPOLOGY-AWARE LATENT EMBEDDING

Traditional autoencoders perform sample latent encoding and feature reconstruction in an instance-independent manner, which easily disrupts the inherent intra-view geometric structures during the encoding process. To address this issue, we employ a Graph Autoencoder (GAE)-based structure-preserving latent encoding to maintain the intra-view geometric relationships during encoding, thereby ensuring the structural integrity of the view-specific latent representations.

Specifically, for the v-th view, we first construct a k-nearest neighbor (k-NN) graph from the original features $X^{(v)}$ to capture local structural information. The undirected symmetric adjacency matrix $G^{(v)} \in \mathbb{R}^{N\times N}$ is defined as:

$$G_{ij}^{(v)} = \begin{cases} 1, & \text{if } x_i^{(v)} \in \mathcal{N}_k(x_j^{(v)}) \text{ or } x_j^{(v)} \in \mathcal{N}_k(x_i^{(v)}), \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\mathcal{N}_k(\cdot)$ denotes the k-nearest neighbor set, and $G_{ij}^{(v)}$ represents the connectivity status between sample i and sample j in the adjacency matrix $G^{(v)}$.

To ensure that instances retain their original feature semantics during neighborhood aggregation, we add self-loops and perform symmetric Laplacian normalization. The symmetrically normalized adjacency matrix $\tilde{G}^{(v)}$ is formulated as:

$$\tilde{G}^{(v)} = (D^{(v)} + I)^{-\frac{1}{2}}(G^{(v)} + I)(D^{(v)} + I)^{-\frac{1}{2}}, \tag{2}$$

where $D^{(v)}$ is the degree matrix of $G^{(v)} + I$ with self-loops, and $I \in \mathbb{R}^{N\times N}$ is the identity matrix.

Based on this, we adopt a Graph Convolutional Network (GCN) as the encoder $E^{(v)}(\cdot)$ to aggregate local topology and feature attributes, outputting the data as a structure-preserving latent representation $H^{(v)} \in \mathbb{R}^{N\times d}$:

$$H^{(v)} = E^{(v)}(X^{(v)}, \tilde{G}^{(v)}) = \sigma\Big(\tilde{G}^{(v)}\sigma\big(\tilde{G}^{(v)}X^{(v)}W_1^{(v)}\big)W_2^{(v)}\Big), \tag{3}$$

where $\sigma(\cdot)$ denotes the LeakyReLU activation function, and $W_1^{(v)}$ and $W_2^{(v)}$ are learnable network weights.

To guarantee the encoding capability of the latent space for both feature attributes and topological structures, and to ensure the structural robustness of representation learning, we introduce a dual-decoder paradigm to reconstruct the original data and the graph topology:

$$\hat{X}^{(v)} = D_X^{(v)}(H^{(v)}), \quad G_{logits}^{(v)} = D_G^{(v)}(H^{(v)}), \tag{4}$$

where $D_X^{(v)}$ is the feature decoder, responsible for reconstructing the original view features from the learned latent representation $H^{(v)}$. $D_G^{(v)}$ is the structure decoder, which reconstructs the graph topology by computing the inner product. $\hat{X}^{(v)}$ represents the reconstructed original view feature matrix, and $G_{logits}^{(v)}$ represents the reconstructed graph topology logits matrix.

The reconstruction loss $\mathcal{L}_{rec}$, which jointly minimizes the feature loss and the structure loss, is formulated as:

$$\mathcal{L}_{rec} = \frac{1}{V}\sum_{v=1}^{V}\Big(\|\hat{X}^{(v)} - X^{(v)}\|_F^2 + \beta \cdot \text{BCEWithLogits}\big(G_{logits}^{(v)}, G^{(v)}; w_{pos}^{(v)}\big)\Big), \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $\beta > 0$ is the regularization weight for topology reconstruction, and BCEWithLogits represents the binary cross-entropy loss with logits (Ruby et al., 2020), which stably measures the discrepancy between the reconstructed graph structure $G_{logits}^{(v)}$ and the original graph structure $G^{(v)}$. $w_{pos}^{(v)} = \frac{N^2 - \sum G^{(v)}}{\sum G^{(v)} + \text{1e-8}}$ represents the positive edge weight, which is used to alleviate the severe imbalance between positive and negative edges in the k-NN graph.

3.4 Strong Cross-View Contrastive

The structure-preserving latent representation $H^{(v)}$ is projected into a unified hyperspherical semantic space via a Multi-Layer Perceptron (MLP): $Z^{(v)} = \text{Normalize}(F_s(H^{(v)}))$, satisfying $\|z_i^{(v)}\|_2 = 1$. We select the view with the clearest clustering silhouette as the template view t, and treat the remaining views $v \neq t$ as auxiliary views. To prevent blind cross-view contrast from introducing false-negative noise that disrupts clustering boundaries, instance-level alignment is strictly restricted within the aligned subset $\mathcal{A}$. For aligned samples $i \in \mathcal{A}$ with known one-to-one correspondence, we pull their cross-view representations closer while pushing apart different anchors. The cross-view strong contrastive loss is formulated as:

$$\mathcal{L}_{cscl} = \frac{1}{V-1} \sum_{v \neq t} \left(-\frac{1}{N_a} \sum_{i \in \mathcal{A}} \log \frac{\exp(\text{sim}(z_i^{(t)}, z_i^{(v)})/\tau_s)}{\sum_{j \in \mathcal{A}} \exp(\text{sim}(z_i^{(t)}, z_j^{(v)})/\tau_s)} \right), \tag{6}$$

where $\text{sim}(u, v) = u^\top v$ denotes the cosine similarity, and τ_s is the temperature parameter. This constraint enables aligned samples to form stable semantic anchors.

3.5 Local Pseudo-Label-Guided Conditional Distribution Alignment

Within the unaligned subset $\mathcal{U}$, cross-view instance pairings are invalid, yet the local manifold topology within each individual view remains reliable. To this end, we design a local probabilistic conditional alignment module, comprising intra-view local neighborhood contrast and pseudo-label-guided conditional Maximum Mean Discrepancy.

3.5.1 Intra-View Local Neighborhood Contrastive

For sample $i \in \mathcal{U}$, we mine its k-nearest neighbor set $\mathcal{N}_u^{(v)}(i) \subset \mathcal{U}$ in the semantic space of each view. Within the same view, a sample and its topological neighbors constitute positive pairs, while the remaining non-self samples form negative pairs. By pulling positive pairs closer and pushing negative pairs apart, the intra-view local neighborhood contrastive loss is formulated as:

$$\mathcal{L}_{intra} = \frac{1}{V} \sum_{v=1}^{V} \left(-\frac{1}{N_u} \sum_{i \in \mathcal{U}} \log \frac{\sum_{j \in \mathcal{N}_u^{(v)}(i)} \exp(\text{sim}(z_i^{(v)}, z_j^{(v)})/\tau_l)}{\sum_{k \in \mathcal{U}, k \neq i} \exp(\text{sim}(z_i^{(v)}, z_k^{(v)})/\tau_l)} \right), \tag{7}$$

where $z_i^{(v)}$ is the ℓ_2-normalized feature of the unaligned sample i, and τ_l is the local contrastive temperature parameter. Completely detached from cross-view pairings, this loss reinforces the local compactness and inter-class discriminability of unaligned samples within individual views in a self-supervised manner.

3.6 Pseudo-Label-Guided Conditional Maximum Mean Discrepancy

To avoid class confusion caused by indiscriminate alignment, we strictly restrict cross-view distribution matching to subsets sharing the same semantic category. In each joint training step, we first extract the multi-view fused features of aligned samples to compute the global cluster centers $\mathcal{C} = \{c_k\}_{k=1}^K$. For unaligned samples in each view, discrete pseudo-labels $y_i^{(v)} = \arg\max_k (z_i^{(v)})^\top c_k$ are assigned independently by calculating the cosine distance to the cluster centers; this strategy prevents cross-view feature contamination caused by row index misalignment. For each pseudo-class $c \in \{1, \ldots, C_{ps}\}$, where $C_{ps} = K$, we extract subsets of unaligned samples belonging to class c from the template view and the auxiliary view, respectively:

$$\mathcal{U}_c^{(t)} = \{z_i^{(t)} \mid i \in \mathcal{U}, y_i^{(t)} = c\}, \quad \mathcal{U}_c^{(v)} = \{z_j^{(v)} \mid j \in \mathcal{U}, y_j^{(v)} = c\}. \tag{8}$$

Conditional Maximum Mean Discrepancy imposes Reproducing Kernel Hilbert Space (RKHS) mean distance alignment exclusively between cross-view sub-distributions belonging to the same pseudo-class c. The pseudo-label-guided conditional Maximum Mean Discrepancy loss is formulated as:

$$\mathcal{L}_{cmmd} = \frac{1}{V-1} \sum_{v \neq t} \left(\frac{1}{C_{ps}} \sum_{c=1}^{C_{ps}} \text{MMD}^2(\mathcal{U}_c^{(t)}, \mathcal{U}_c^{(v)}) \right). \tag{9}$$

The multi-kernel Maximum Mean Discrepancy is formulated as:

$$\text{MMD}^2(X_1, X_2) = \frac{1}{|X_1|^2} \sum_{x,x' \in X_1} \kappa(x, x') - \frac{2}{|X_1||X_2|} \sum_{x \in X_1, y \in X_2} \kappa(x, y) + \frac{1}{|X_2|^2} \sum_{y,y' \in X_2} \kappa(y, y'), \tag{10}$$

where the kernel function $\kappa(x, y) = \sum_{m=1}^{M} \exp\left(-\frac{\|x-y\|_2^2}{2\sigma_m^2}\right)$ is a multi-scale Gaussian mixture. This conditional matching constrains the cross-view sub-distributions of the same cluster to align tightly without requiring external supervision.

In summary, the complete loss for the local pseudo-label-guided conditional distribution alignment module is formulated as:

$$\mathcal{L}_{lpca} = \mathcal{L}_{intra} + \mathcal{L}_{cmmd}. \tag{11}$$

3.7 Overall Objective

By integrating Eq 5, Eq 6, and Eq 11, the overall objective function is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_a \mathcal{L}_{cscl} + \lambda_l \mathcal{L}_{lpca}, \tag{12}$$

where $\lambda_a, \lambda_l \geq 0$ are hyperparameters. The network is first pre-trained by optimizing only $\mathcal{L}_{rec}$ to construct robust single-view geometric manifolds; subsequently, full-parameter joint fine-tuning is performed, enabling cross-view alignment and local conditional matching to collaboratively optimize the latent semantic space.

After joint training, a final discriminative feature representation is generated for each sample for K-means clustering. In the partially view-aligned scenario, indiscriminately averaging all view features would erroneously add randomly shuffled features of different categories from auxiliary views to the template view, resulting in severe semantic contamination. Therefore, we establish a decoupled feature fusion strategy:

$$z_i^* = \begin{cases} \frac{1}{V} \sum_{v=1}^{V} z_i^{(v)}, & i \in \mathcal{A}, \\ z_i^{(t)}, & i \in \mathcal{U}, \end{cases} \tag{13}$$

where z_i^* denotes the final discriminative feature vector of sample i after multi-view fusion. That is, aligned samples utilize multi-view consistency averaging, whereas unaligned samples are strictly maintained within the reference coordinate system of the template view.

Finally, the fused features of all samples are concatenated:

$$Z^* = [z_1^*, \ldots, z_N^*]^\top \in \mathbb{R}^{N \times d_z}. \tag{14}$$

Ultimately, Z^* is fed into the K-means algorithm to obtain the final clustering assignments. The complete computational procedure of DeCA is summarized in Algorithm 1 in the Appendix, and the detailed complexity analysis is deferred to Appendix A.2.

4 Experiments

4.1 Datasets Informations

Our datasets consist of the following ten: webkb (Wen et al., 2024), GSE100866 (Deng et al., 2026), UCI3view (Xia et al., 2014), Handwritten (Wen et al., 2024), Scene15 (Fei-Fei & Perona, 2005), Fashion (Xiao et al., 2017), MNIST-USPS (Peng et al., 2019), NoisyMNIST-tune (Wang et al., 2015), ReutersEF (Huang et al., 2019), and Cifar10 (Zhang et al., 2019). Table 1 comprehensively summarizes the specific attributes of the ten multi-view datasets, including the number of samples, views, clusters, and feature dimensions. The datasets span a wide range of scales, from small (1,051 samples) to large (50,000 samples), and vary from two to six dimensions, encompassing a diverse array of view counts and cluster categories.

4.2 COMPARISON METHODS

To comprehensively evaluate the clustering capability of our method on partially aligned data, we compare it against eight PVC methods: DAGF (Jiang et al., 2026), FUMC(Li et al., 2024), PAVuC-ATS(Dong et al., 2026), AFUMC(Sun et al., 2025a), CANDY (Guo et al., 2024), BIN (Deng et al.,

Table 1: Summary statistics of different datasets.

Dataset	Samples	Views	Clusters	Dimensions
webkb	1051	2	2	2949/334
GSE100866	1182	2	6	10/1180
UCI3view	2000	3	10	240/76/6
Handwritten	2000	6	10	216/76/64/6/240/47
Scene15	4485	3	15	1800/1180/1240
MNIST-USPS	5000	2	10	784/256
Fashion	10000	3	10	784/784/784
NoisyMNIST-tune	10000	2	10	784/784
ReutersEF	18758	2	6	100/100
Cifar10	50000	3	10	512/2048/1024

Table 2: Quantitative evaluation of clustering results on ten datasets given an unaligned rate of 0.5. Top and runner-up scores are indicated by boldface and underlines, respectively. Entries marked with "OT" denote algorithms that failed to terminate within a 12-hour limit.

Method	webkb			GSE100866			UCI3view			Handwritten			Scene15		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
DAGF	54.19	2.61	1.69	90.61	84.92	82.01	64.65	64.22	49.33	64.80	64.41	49.55	46.69	42.03	24.97
FUMC	75.86	47.84	56.06	83.84	79.46	72.42	27.40	14.45	8.14	42.20	17.63	13.16	34.05	35.34	18.08
PAVuCATS	66.01	20.94	30.90	56.18	32.38	25.75	50.60	37.46	29.94	49.40	42.09	30.05	41.40	33.14	20.58
AFUMC	37.19	7.10	5.05	25.00	2.18	2.03	71.62	64.60	53.88	64.02	63.42	49.31	46.91	45.28	29.43
CANDY	35.74	3.45	0.49	65.08	74.71	51.34	70.24	67.57	57.15	76.81	74.11	65.54	36.84	27.06	16.66
BIN	48.77	16.31	16.97	**97.59**	**96.09**	95.82	90.60	81.12	80.32	88.72	78.58	76.81	27.10	17.87	9.76
DWInfoNCE	40.89	8.91	3.45	69.83	79.05	57.44	84.67	80.84	76.23	83.28	79.62	74.14	55.96	52.78	36.98
RMCNC	36.95	11.65	-0.95	79.02	87.68	69.79	84.33	80.75	75.70	83.55	80.59	75.01	58.79	53.50	40.33
Ours	**78.82**	**48.99**	**59.07**	95.85	94.99	**98.25**	**95.45**	**90.21**	**90.18**	**94.55**	**88.48**	**88.31**	**62.79**	**56.15**	**44.56**

Method	MNIST-USPS			Fashion			NoisyMNIST-tune			ReutersEF			Cifar10		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
DAGF	42.08	38.20	22.33	47.99	49.34	30.05	42.94	36.07	18.91	28.68	2.42	1.03	OT	OT	OT
FUMC	44.86	35.65	23.59	35.58	19.49	13.54	35.75	28.13	17.47	53.26	30.82	44.03	91.32	81.72	81.96
PAVuCATS	53.34	44.16	30.21	40.25	21.66	16.52	30.98	19.96	12.07	53.79	31.91	37.43	OT	OT	OT
AFUMC	40.68	37.61	23.18	52.54	51.78	34.74	49.66	40.13	29.28	39.20	16.91	33.85	83.94	69.44	68.05
CANDY	34.33	24.55	14.88	53.39	48.71	32.72	40.52	34.45	22.84	28.01	7.37	25.39	69.14	64.08	51.66
BIN	71.21	47.86	46.44	60.25	36.91	33.27	50.59	24.10	21.05	39.91	18.30	36.02	80.25	60.71	61.22
DWInfoNCE	95.17	88.20	89.60	50.83	77.79	69.82	71.58	70.55	59.16	32.87	16.39	40.25	93.77	85.81	86.87
RMCNC	95.66	89.32	90.62	57.71	75.95	72.87	88.19	76.72	75.73	42.39	27.60	**47.04**	93.51	85.12	86.33
Ours	**96.58**	**91.45**	**92.57**	**89.05**	**81.52**	**78.87**	**92.40**	**83.20**	**84.05**	**56.73**	**33.58**	28.82	**95.25**	**88.58**	**89.85**

2026), DWInfoNCE(Chen et al., 2026), and RMCNC (Sun et al., 2024). Among these, the first three are non-deep methods, while the remaining are deep learning methods.

4.3 Implementation Details

All experiments were conducted on an Ubuntu 20.04.3 LTS system equipped with an NVIDIA RTX 3090 GPU, implemented in PyTorch 2.3.1. To ensure fair comparisons, all baseline methods were executed using their official source codes, with hyperparameters optimally tuned according to their original papers. For evaluation, we employ three widely used clustering metrics: clustering Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). All results are reported as percentages (%). For our proposed DeCA model, the experimental settings were as follows: the Graph Autoencoder (GAE) was configured with a 1024-dimensional hidden layer to output a 128-dimensional latent embedding. The first view was designated as the template view. The number of nearest neighbors for the KNN graph, k, was set to 5. The model underwent pre-training for 50 epochs and joint training for 300 epochs, with a batch size of 2000. We utilized the Adam optimizer with a learning rate of 5×10^{-4} and a weight decay of 1×10^{-4}. The temperature parameters τ_s and τ_l were fixed at 0.1 and 0.2, respectively. A grid search was performed for the trade-off hyperparameters λ_a and λ_l within the set $\{0.1, 0.5, 1, 5, 10\}$. We fixed N_a samples to

remain aligned across all views, while the remaining N_u samples were randomly shuffled across different views. The unaligned rate is defined as N_u/N.

4.4 PERFORMANCE COMPARISON

The clustering results of our method and the eight baselines under 50% and 80% unaligned rates are summarized in Tables 2 and 3, respectively. Overall, DeCA achieves optimal performance across the vast majority of datasets. Under the 50% unaligned rate condition, DeCA's Accuracy (ACC) exceeds 95% on multiple datasets, such as MNIST-USPS and Cifar10, significantly outperforming non-deep baseline models. When the unaligned rate escalates to an extreme 80%, deep models like BIN experience a precipitous performance drop due to their over-reliance on sufficient aligned samples. In contrast, DeCA exhibits superior robustness against interference, maintaining an ACC above 95% on the GSE100866 and UCI3view datasets with only marginal performance degradation.

Figure 2 illustrates the performance trends of nine methods, including ours, across three metrics (ACC, NMI, and ARI) under varying unaligned rates (from 0.1 to 0.9) on the UCI3view dataset. As shown in the figure, our method consistently maintains the highest performance level and does not exhibit significant fluctuations as the unaligned rate varies.To further evaluate our model, we provide hyperparameter sensitivity and convergence analyses in Appendix A.3. Additionally, an execution time comparison against the eight baselines is deferred to Appendix A.4.

Table 3: Quantitative evaluation of clustering results on ten datasets given an unaligned rate of 0.8. Top and runner-up scores are indicated by boldface and underlines, respectively. Entries marked with "OT" denote algorithms that failed to terminate within a 12-hour limit.

Method	webkb			GSE100866			UCI3view			Handwritten			Scene15		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
DAGF	53.69	2.63	2.16	90.52	84.85	81.87	64.55	64.12	49.28	65.00	64.20	49.65	46.15	41.90	27.05
FUMC	72.91	39.38	49.78	83.76	78.97	72.75	25.55	14.81	7.97	23.75	4.72	2.49	34.47	34.87	17.65
PAVuCATS	67.98	21.44	33.46	51.61	31.79	23.72	45.60	33.63	25.43	53.95	46.01	35.05	38.51	36.81	20.47
AFUMC	37.19	7.10	5.05	22.50	0.97	0.39	70.90	67.16	56.34	70.10	67.81	57.44	47.00	44.63	29.61
CANDY	35.76	2.32	0.32	67.09	77.92	57.32	55.49	50.80	38.18	64.28	57.74	46.26	32.33	23.67	15.09
BIN	40.39	7.38	6.38	93.31	92.38	88.75	88.10	77.15	75.64	83.05	71.73	66.70	16.45	8.60	3.45
DWInfoNCE	46.40	13.91	10.67	72.05	79.92	58.94	84.55	77.93	73.46	87.11	78.83	75.16	55.05	48.65	35.43
RMCNC	48.11	17.36	13.10	73.10	81.97	62.42	85.32	78.97	74.67	85.87	79.53	75.02	53.07	46.64	33.05
Ours	**76.35**	**44.65**	**57.56**	**96.87**	**92.69**	**95.34**	**95.95**	**90.93**	**91.21**	**92.00**	**84.74**	**83.36**	**61.03**	**53.68**	**42.63**

Method	MNIST-USPS			Fashion			NoisyMNIST-tune			ReutersEF			Cifar10		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
DAGF	44.44	40.30	25.35	48.92	49.77	30.68	41.82	35.63	22.16	27.96	2.57	-0.94	OT	OT	OT
FUMC	40.38	27.50	17.03	33.30	32.68	20.16	36.07	27.76	17.32	53.29	31.28	23.67	89.28	78.90	78.51
PAVuCATS	46.36	40.94	25.08	33.56	22.52	13.54	27.52	18.97	10.60	53.97	32.74	26.15	OT	OT	OT
AFUMC	50.05	42.55	30.17	55.75	53.38	37.10	52.28	44.80	33.59	44.56	22.69	18.76	81.73	66.37	70.24
CANDY	32.18	25.98	14.50	44.75	38.71	23.67	32.28	26.21	15.21	25.60	3.17	1.48	68.93	60.99	50.48
BIN	48.74	25.92	20.79	60.25	36.91	33.27	21.52	3.98	2.48	38.91	13.99	12.45	66.11	41.53	39.86
DWInfoNCE	88.39	75.72	76.16	50.83	**77.79**	69.82	46.24	47.94	24.56	38.31	20.60	7.31	91.92	82.22	83.15
RMCNC	86.27	73.92	72.37	57.71	75.95	**72.87**	51.25	46.67	25.62	37.16	21.61	4.40	92.09	82.50	83.49
Ours	**93.92**	**86.15**	**87.04**	**82.79**	73.74	68.94	**88.03**	**76.23**	**75.69**	**54.75**	**33.27**	**28.86**	**92.65**	**83.56**	**84.56**

4.5 ABLATION RESULTS

Table 4 presents the ablation study results of the DeCA model. The first row indicates replacing the Graph Autoencoder (GAE) with a Multi-Layer Perceptron (MLP), which leads to a performance degradation of the model. "w/o L_{cscl}" denotes the removal of the strong cross-view contrast; this variant experiences a substantial performance drop of 36.11% on the NoisyMNIST-tune dataset, demonstrating that imposing strong cross-view contrast on aligned samples lays a robust structural foundation for subsequent operations. "w/ Global MMD" indicates replacing the conditional alignment with global Maximum Mean Discrepancy, which also leads to a performance decline. "w/o L_{lpca}" denotes the removal of the local pseudo-label-guided conditional distribution alignment, while "w/o L_{cscl} & w/o L_{lpca}" indicates the simultaneous removal of both the strong cross-view contrast and the local pseudo-label-guided conditional alignment. On the Handwritten dataset, these variants suffer precipitous performance drops of 33.3% and 31.6%, respectively. Such significant degradations indicate that removing these essential components leads to the collapse of the entire model framework. Finally, "w/o Decoupled Fusion" denotes the removal of the decoupled feature

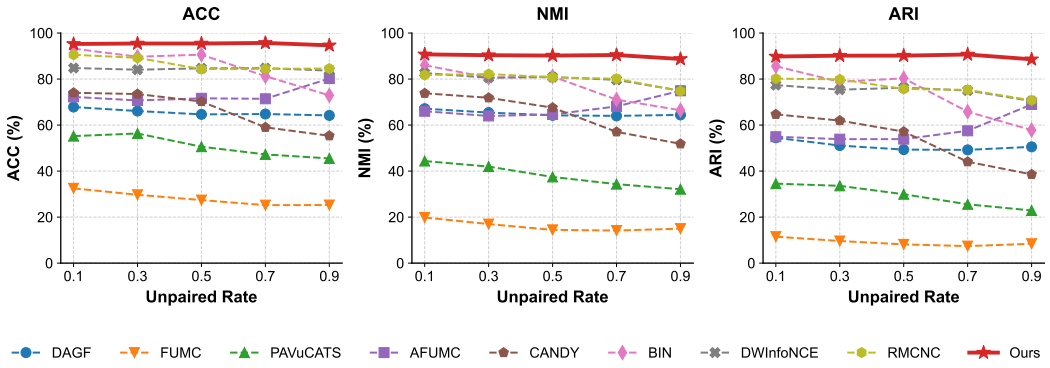

Figure 2: Comparison of clustering performance across nine methods under different unaligned rates on the UCI3view dataset.

fusion strategy, which results in performance drops of 7.4% and 3.78% on the Handwritten and NoisyMNIST-tune datasets, respectively, thereby confirming the effectiveness of this strategy.

Table 4: Ablation results on the Handwritten and NoisyMNIST-tune datasets under an unaligned rate of 0.5. The best results are highlighted in **bold**.

Variant	Handwritten			NoisyMNIST-tune		
	ACC	NMI	ARI	ACC	NMI	ARI
w/ MLP	87.15	76.55	73.81	88.62	77.05	76.89
w/o L_{cscl}	61.50	57.39	41.12	56.29	44.20	33.26
w/ Global MMD	93.40	87.00	85.99	91.99	82.36	83.22
w/o L_{lpca}	61.35	60.10	37.40	44.96	43.13	27.11
w/o L_{cscl} & w/o L_{lpca}	62.95	38.98	34.38	73.22	50.03	49.58
w/o Decoupled Fusion	87.15	76.55	73.81	88.62	77.05	76.89
DeCA	**94.55**	**88.48**	**88.31**	**92.40**	**83.20**	**84.05**

4.6 Conclusion

The proposed Decoupled Conditional Alignment (DeCA) method is specifically tailored for partially view-aligned scenarios. First, to mitigate structural distortion, we construct a dual-decoder Graph Autoencoder that leverages neighborhood constraints to strictly preserve native intra-view neighborhood connectivity. Second, to prevent the interference features of unaligned samples from misleading network feature propagation, we design an efficient cross-view contrastive learning mechanism exclusively for aligned samples to establish preliminary robust semantic anchors. Subsequently, to alleviate the issue where directly applying contrastive learning to unaligned samples introduces massive false negatives and erroneous repulsion, we design a pseudo-label-guided local conditional alignment module. This module achieves cross-view matching of intra-class sub-distributions for unaligned samples. Simultaneously, we introduce a decoupled feature fusion strategy to mitigate the semantic contamination problem. Extensive experiments conducted on ten datasets against eight baseline methods demonstrate that our method achieves state-of-the-art clustering performance.

AI use statement

In this work, we used generative AI tools (Gemini) for text polishing, assisting with code generation, and LaTeX formatting. We have not used generative AI tools to generate core research ideas or theoretical methodologies, nor have we used them to fabricate data. We have reviewed all AI-assisted work. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

To ensure the reproducibility of our method, we have made the source code, pre-processed data, and execution instructions publicly available via an anonymous repository: https://anonymous.4open.science/r/DeCA-9F4F/

References

Shifeng Bao, Zhe Xue, Qi Chen, Shilong Ou, Amin Beheshti, Quan Z Sheng, Anton van den Hengel, and Yuankai Qi. Causalmvc: Causal content-style representation learning for deep multi-view clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1598–1606, 2025.

Junfeng Cao, Wenhua Dong, and Jing Chen. View-unaligned clustering with graph regularization. *Pattern Recognition*, 155:110706, 2024.

Guoqing Chao, Kaixin Xu, Xijiong Xie, and Yongyong Chen. Global graph propagation with hierarchical information transfer for incomplete contrastive multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 15713–15721, 2025.

Qiuyu Chen, Xihong Yang, Zhibing Dong, Qian Qu, En Zhu, and Xinwang Liu. Dynamic weight-enhanced contrastive learning for partially view-aligned clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(8):12426–12438, 2026. doi: 10.1109/TCSVT.2026.3679096.

Shanghui Deng, Xiao Zheng, Chang Tang, Kun Sun, Yuanyuan Liu, and Xinwang Liu. Find true collaborators: Banzhaf index-based cross view alignment for partially view-aligned clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2447–2456, 2025.

Shanghui Deng, Xiao Zheng, Chang Tang, Lianbo Guo, and Xinwang Liu. Game theory inspired cross-view interaction alignment for partially view-aligned clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2026.

Wenhua Dong, Xiao-Jun Wu, Zhenhua Feng, Sara Atito, Muhammad Awais, and Josef Kittler. Probabilistically aligned view-unaligned clustering with adaptive template selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1904–1916, 2026. doi: 10.1109/TPAMI.2025.3618984.

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. doi: 10.1109/TKDE.2023.3270311.

Xiang Fang, Wanlong Fang, Wei Ji, and Tat-Seng Chua. Turing patterns for multimedia: Reaction-diffusion multi-modal fusion for language-guided video moment retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 12509–12518, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758179. URL https://doi.org/10.1145/3746027.3758179.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pp. 524–531 vol. 2, 2005. doi: 10.1109/CVPR.2005.16.

Hang Gao, Zuosong Cai, Yuze Li, Cheng Liu, Gaoyang Li, Ying Li, Wei Du, and You Zhou. Learning whom to align with: Progressive anomaly combination detection for partially view-aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21171–21179, 2026.

Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 121401–121421. Curran Associates, Inc., 2024. doi: 10.52202/079017-3857. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/dbe81b08f7dc4dd8b43bc62dedfd9662-Paper-Conference.pdf.

Wei Guo, Yinghui Sun, Xingfeng Li, Zhenwen Ren, and Quansen Sun. Collaborative anchor graph learning with similarity-guided alignment for multi-view clustering. *IEEE Transactions on Multimedia*, pp. 1–14, 2026. doi: 10.1109/TMM.2026.3704856.

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, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681331. URL https://doi.org/10.1145/3664647.3681331.

Shudong Hou, Le Zhao, Yongchi Fan, Xiaopeng Cheng, and Zhehui Yu. Deep multi-view clustering via dual contrastive consistency fusion. *Pattern Analysis and Applications*, 29(1):48, 2026.

Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. Multi-view spectral clustering network. In *Ijcai*, volume 2, pp. 4, 2019.

Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. Partially view-aligned clustering. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2892–2902. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1e591403ff232de0f0f139ac51d99295-Paper.pdf.

Miao Jia, Zijian Chen, Qing Cheng, and Xingchen Hu. Dynamic prototype adaptation for deep multi-view clustering. In *2025 11th International Conference on Big Data and Information Analytics (BigDIA)*, pp. 8–15, 2025. doi: 10.1109/BigDIA68682.2025.11382808.

Hongyu Jiang, Hong Tao, Zhangqi Jiang, and Chenping Hou. Unaligned multi-view clustering via diversified anchor graph fusion. *Pattern Recognition*, 170:111977, 2026.

Xingfeng Li, Yuangang Pan, Yinghui Sun, Quansen Sun, Ivor W Tsang, and Zhenwen Ren. Fast unpaired multi-view clustering. In *IJCAI*, pp. 4488–4496, 2024.

Xincan Lin, Jie Lian, Zhihao Wu, Jielong Lu, and Shiping Wang. Deep dual contrastive learning for multi-view subspace clustering. *Information Sciences*, 723:122678, 2026. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2025.122678. URL https://www.sciencedirect.com/science/article/pii/S0020025525008114.

Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. COMIC: Multi-view clustering without parameter selection. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5092–5101. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/peng19a.html.

Shengsheng Qian, Dizhan Xue, Jun Hu, Huaiwen Zhang, and Changsheng Xu. Nonparametric clustering-guided cross-view contrastive learning for partially view-aligned representation learning. *IEEE Transactions on Image Processing*, 33:6158–6172, 2024. doi: 10.1109/TIP.2024.3480701.

Yalan Qin, Nan Pu, Hanzhou Wu, and Nicu Sebe. Margin-aware noise-robust contrastive learning for partially view-aligned problem. *ACM Transactions on Knowledge Discovery from Data*, 19 (1):1–20, 2025a.

Yalan Qin, Xinpeng Zhang, Shui Yu, and Guorui Feng. A survey on representation learning for multi-view data. *Neural Networks*, 181:106842, 2025b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106842. URL https://www.sciencedirect.com/science/article/pii/S0893608024007664.

Usha Ruby, Vamsidhar Yendapalli, et al. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng*, 9(10):5393–5397, 2020.

Bohang Sun, Yuena Lin, Tao Yang, Zhen Zhu, Zhen Yang, and Gengyu Lyu. Af-umc: An alignment-free fusion framework for unaligned multi-view clustering. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 104157–104179. Curran Associates, Inc., 2025a. doi: 10.52202/085713-3478. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/965484d5b2b2624ba17295612a3ba7e8-Paper-Conference.pdf.

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. doi: 10.1109/TKDE.2024.3423307.

Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, and Peng Hu. Roll: Robust noisy pseudo-label learning for multi-view clustering with noisy correspondence. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30732–30741, 2025b. doi: 10.1109/CVPR52734.2025.02862.

Jun Wang, Zhenglai Li, Chang Tang, Suyuan Liu, Hao Yu, Chuan Tang, Miaomiao Li, and Xinwang Liu. Scalable cross-view sample alignment for multi-view clustering with view structure similarity. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 69832–69857. Curran Associates, Inc., 2025. doi: 10.52202/085713-2349. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/65338cfb603d4871a2c38e53a3e039c9-Paper-Conference.pdf.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1083–1092, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/wangb15.html.

Xibiao Wang, Hang Gao, Xindian Wei, Liang Peng, Rui Li, Cheng Liu, Si Wu, and Hau-San Wong. Contrastive graph distribution alignment for partially view-aligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 5240–5249, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681048. URL https://doi.org/10.1145/3664647.3681048.

Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149, 2023. doi: 10.1109/TSMC.2022.3192635.

Yi Wen, Siwei Wang, Qing Liao, Weixuan Liang, Ke Liang, Xinhang Wan, and Xinwang Liu. Unpaired multi-view graph clustering with cross-view structure matching. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16049–16063, 2024. doi: 10.1109/TNNLS.2023.3291696.

Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xihong Yang, Siwei Wang, Jiaqi Jin, Fangdi Wang, Tianrui Liu, Yueming Jin, Xinwang Liu, En Zhu, and Kunlun He. Generalized deep multi-view clustering via causal learning with partially aligned cross-view correspondence. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1990–1999, 2025. doi: 10.1109/ICCV51701.2025.00193.

Zhiwen Yu, Ziyang Dong, Chenchen Yu, Kaixiang Yang, Ziwei Fan, and CL Philip Chen. A review on multi-view learning. *Frontiers of Computer Science*, 19(7):197334, 2025.

Changqing Zhang, Yeqing Liu, and Huazhu Fu. Ae2-nets: Autoencoder in autoencoder networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2572–2580, 2019. doi: 10.1109/CVPR.2019.00268.

Liang Zhao, Qiongjie Xie, Zhengtao Li, Songtao Wu, and Yi Yang. Dynamic graph guided progressive partial view-aligned clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):9370–9382, 2024a.

Liang Zhao, Yukun Yuan, Qiongjie Xie, and Ziyue Wang. Anchor based multi-view clustering for partially view-aligned data. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–5, 2024b. doi: 10.1109/ICME57554.2024.10688164.

Liang Zhao, Ziyue Wang, Xiao Wang, Zhikui Chen, and Bo Xu. Incomplete and unpaired multi-view graph clustering with cross-view feature fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22786–22794, 2025.

Liang Zhao, Tianqi Yue, Shubin Ma, Ziyue Wang, ZhiYuan Liu, and Bo Xu. Knnda: A new perspective of alignment recovery for partially view-aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28706–28714, 2026.

A Appendix

A.1 Algorithm Procedure

The overall optimization and clustering procedure of the DeCA framework is summarized in Algorithm 1. First, the k-nearest neighbor graph is constructed, and the network parameters are initialized. Second, the pre-training phase is optimized via the reconstruction loss. Upon the completion of the pre-training phase, the model enters the joint training phase, where network parameters are updated based on the overall joint loss. Once the model converges, the algorithm executes the decoupled feature fusion strategy to generate a unified representation, and applies K-means to output the final clustering results for all samples.

Algorithm 1 Optimization of DeCA

Input: Multi-view data $\{\mathbf{X}^v\}_{v=1}^V$, paired subset $\mathcal{A}_{align}$, trade-offs λ_a, λ_l, pre-training epochs E_{pre}, joint training epochs E_{joint}.
Output: Clustering results.
Init: Construct KNN graphs $\{\mathbf{G}^v\}_{v=1}^V$; initialize network parameters Θ.
Stage 1: Pre-training
for $epoch = 1 \dots E_{pre}$ **do**
  Compute latent representations $\mathbf{H}^v$ via Eq. (3), and decodings $\mathbf{X}^v$, $\mathbf{G}^v_{logits}$ via Eq. (4).
  Compute $\mathcal{L}_{rec}$ via Eq. (5), and update graph autoencoder parameters via backpropagation.
end for
Stage 2: Joint training
for $epoch = 1 \dots E_{joint}$ **do**
  Forward: Project $\mathbf{H}^v$ into the unified semantic space $\mathbf{Z}^v$.
  Losses: Compute $\mathcal{L}_{cscl}$ via Eq. (6), and $\mathcal{L}_{lpca}$ via Eq. (11).
  Backward: Compute $\mathcal{L}_{total}$ via Eq. (12), and update network parameters via backpropagation.
end for
Stage 3: Feature fusion and clustering
Clustering: After convergence, construct the feature matrix $\mathbf{Z}^*$ via Eqs. (13) and Eqs. (14), and then apply K-means.

A.2 Complexity Analysis

Let N, V, d_i, d_e, and E denote the total number of samples, the number of views, the maximum original input dimension, the semantic embedding dimension, and the total number of training epochs, respectively.

Time Complexity. The computational cost of the DeCA framework is dominated by three components: the KNN graph construction phase takes $\mathcal{O}(VN^2d_i)$, the high-frequency matrix multiplications in the GAE optimization and alignment modules consume $\mathcal{O}(VN^2d_e)$ per epoch, and the final K-means clustering requires $\mathcal{O}(I_{iter}NKd_e)$. Overall, the total time cost grows quadratically with the number of samples, and the overall time complexity is upper-bounded by $\mathcal{O}(VN^2d_i + EVN^2d_e)$.

Space Complexity. The spatial bottleneck of the model primarily lies in caching various dense relational matrices, which incurs a dominant memory footprint of $\mathcal{O}(VN^2)$. Meanwhile, storing the original features and latent semantic representations requires $\mathcal{O}(VNd_i + VNd_e)$ of space. Therefore, the overall space complexity of the framework is strictly bounded by $\mathcal{O}(VN^2 + VNd_i)$.

A.3 Sensitivity and Convergence Analysis

Under an unaligned rate of 0.8 on the NoisyMNIST-tune dataset, we perform a grid search for the two hyperparameters λ_a and λ_l within the set $\{0.1, 0.5, 1, 5, 10\}$ and plot the hyperparameter sensitivity as shown on the left side of Figure 3. As observed from the smooth surface in the figure, the clustering performance of DeCA does not exhibit significant fluctuations in response to parameter variations, verifying the high robustness of our method.

The right side of Figure 3 illustrates the overall loss curve during the joint training phase on the MNIST-USPS dataset with an unaligned rate of 0.5. As can be seen from the figure, the overall objective loss drops rapidly during the first 50 epochs of joint training. In the subsequent 250 epochs, the loss continues to decrease and gradually converges.

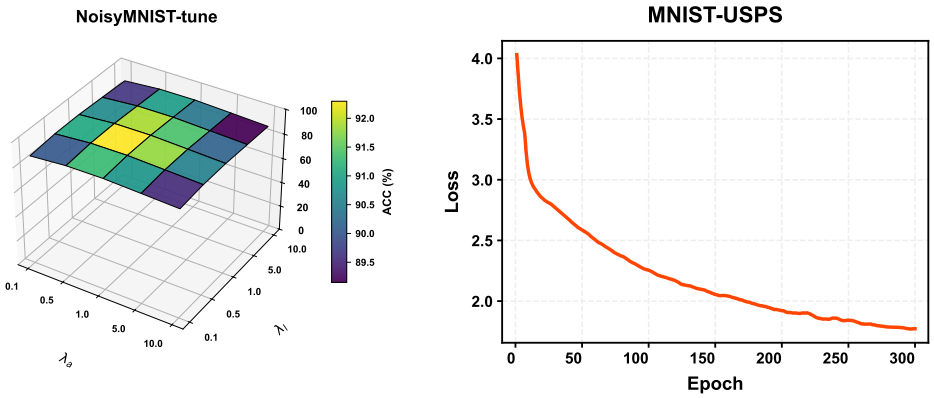

Figure 3: (Left) Robustness evaluation with respect to hyperparameters λ_a and λ_l. (Right) Convergence behavior of the total loss during the training phase.

A.4 Execution Time

Figure 4 illustrates the execution time of our method across ten datasets, which remains nearly identical under the unaligned rates of 0.5 and 0.8, demonstrating exceptional robustness to the degree of unalignment. Figure 5 compares the time costs of our method against eight other methods. As the dataset scale expands, our execution time gradually surpasses multiple methods such as DAGF and CANDY, confirming the highly competitive efficiency of our approach.

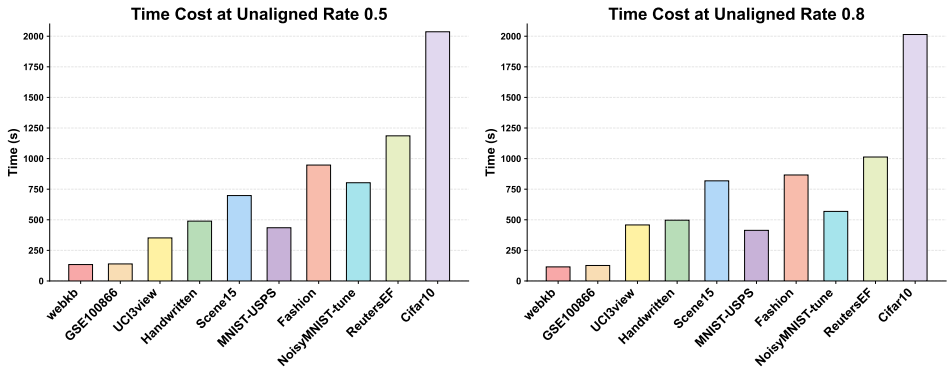

Figure 4: Execution time comparison of DeCA across different datasets at unaligned rates of 0.5 and 0.8.

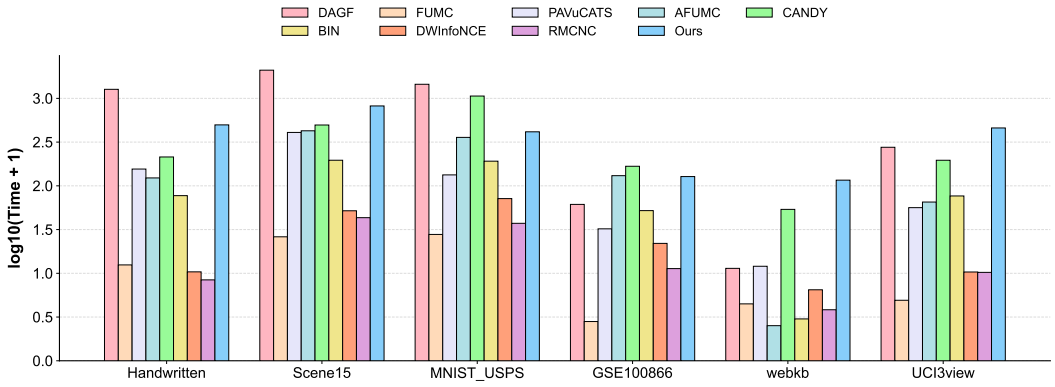

Figure 5: Time cost comparison across different methods on six datasets under an unaligned rate of 0.8.