Learning Whom to Align With: Progressive Anomaly Combination Detection for Partially View-Aligned Clustering

Hang Gao[1], Zuosong Cai[1], Yuze Li[1], Cheng Liu[2,4], Gaoyang Li[3], Ying Li[1*], Wei Du[1*], You Zhou[1]

[1]Key Laboratory of Symbolic Computation and Knowledge Engineering of the Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China
[2]College of Computer Science and Technology, Huaqiao University, Xiamen, China
[3]School of Life Sciences, Nanjing Medical University, Nanjing, China
[4] Department of Computer Science, Shantou University, Shantou China
{gaohang23, caizs23, yuzel23}@mails.jlu.edu.cn, chengliu10@gmail.com, lgyzngc@njmu.edu.cn, {liying, weidu, zyou}@jlu.edu.cn

Abstract

Partially View-aligned Clustering (PVC) addresses the challenge of partial view alignment in multi-view learning by leveraging complementary and consistent information. While existing PVC methods show promise, most rely on distance-based strategies that are sensitive to view-specific details and noise, limiting their robustness. In this work, we propose a novel view alignment strategy that reformulates the alignment task as an anomaly detection problem. Rather than learning a view-alignment matrix that enforces strict one-to-one correspondences across views, we adopt a progressive approach to identify well-aligned samples. Specifically, we sample subsets of data by generating random view combinations from unaligned samples and propose an anomaly combination detection module to evaluate the alignment consistency of these combinations. In addition, our progressive training framework alternates between updating model parameters and selecting high-confidence view combinations for subsequent optimization. By reformulating view alignment as an anomaly detection task, our approach provides a more robust and effective solution to partial view alignment. Experiments on benchmark datasets demonstrate that our method outperforms state-of-the-art approaches in the PVC problem.

Introduction

In real-world applications, data is often represented in multiple modalities, such as images, text, and videos, collectively referred to as multi-view data. Multi-view Clustering (MVC) aims to improve clustering performance by leveraging consistent and complementary information across different views (Han et al. 2022; Hassani and Khasahmadi 2020; Zhang et al. 2022). By effectively capturing both the shared and distinct features within multi-view data, MVC improves clustering accuracy and robustness, making it a powerful approach for solving complex clustering problems across various domains. However, existing multi-view learning methods often assume that all views are perfectly aligned (Jiang et al. 2022; Wan et al. 2022; Pan and Kang 2021; Liu et al. 2022). In practice, this assumption is often violated due to factors such as sensor discrepancies or communication disruptions, leading to partially view-aligned data. In such scenarios, improper alignment can hinder multi-view representation learning, affecting downstream tasks and overall performance. As a result, there is a pressing need for models that can handle the partial view alignment problem.

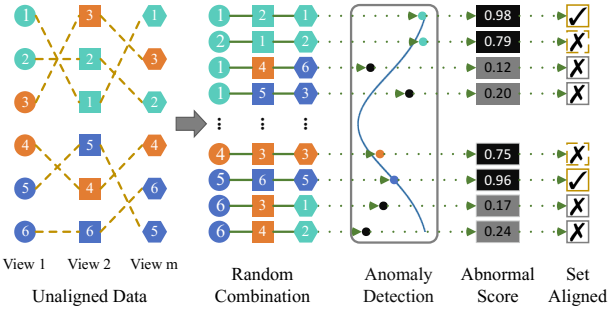

Figure 1: This example demonstrates the core idea of the proposed view alignment method. Distinct shapes correspond to different views, while various colors represent different clusters, and individual instances are denoted by unique numbers. We randomly combine misaligned instances and feed them into the anomaly detection process. Based on the resulting anomaly scores, the most reliable combinations are selected and assigned to the aligned subset.

Recently, several approaches have been proposed to address the challenges of partially view-aligned clustering (PVC), demonstrating promising performance (Li et al. 2024; Qian et al. 2024; Zhao et al. 2023; Gao et al. 2025a). For instance, Huang et al. introduce a differentiable surrogate of the Hungarian algorithm to establish correspondences between partially aligned views, enabling the learning of a shared latent space for better alignment (Huang et al. 2020). He et al. propose a robust variational contrastive learning method that models each sample as a Gaussian distribution, where the mean captures shared semantics and

*Corresponding author.

the variance reflects view-specific details, enhancing alignment and complementarity (He et al. 2024). Wang et al. align semantic similarity graphs using Wasserstein distance to learn view correspondences and apply cross-view contrastive learning to extract shared semantic features (Wang et al. 2024a). While these PVC methods have achieved promising results, they rely heavily on distance-based strategies, such as cross-view instance similarity or graph distances, to establish correspondences. However, these methods are susceptible to perturbations, which may arise from view-specific noise, semantic inconsistencies, or distribution shifts across views. Such perturbations can distort similarity measurements and lead to unreliable alignments. The problem becomes more pronounced in multi-view scenarios with more than two views, where pairwise alignment is less effective. Moreover, many existing methods fail to fully leverage the consistency and complementarity within realigned data to enhance clustering performance.

To address the view alignment problem in partially aligned multi-view data, we propose a novel strategy that reformulates the alignment task as an anomaly detection problem. Instead of directly learning a view-alignment matrix to enforce strict one-to-one correspondences, we adopt a progressive approach that identifies semantically aligned samples across views. Specifically, we generate random cross-view combinations from unaligned samples and assess their alignment consistency using an anomaly combination detection module. Combinations with high scores are regarded as reliable and added to an aligned subset. This process is illustrated in Fig. 1, where random combinations are evaluated, and high-confidence samples are progressively selected for alignment. Our progressive framework alternates between model optimization and high-confidence alignment selection, gradually improving view consistency. Furthermore, contrastive learning is employed to enforce semantic consistency across selected combinations, enhancing both representation learning and alignment robustness. The main contributions of this work could be summarized as follows:

- We introduce a novel perspective by reformulating the view alignment problem as an anomaly detection task. By treating misaligned view combinations as anomalies, our approach effectively identifies inconsistent data distributions, thereby overcoming the limitations of distance-based alignment methods that are sensitive to noise and view-specific discrepancies.
- We propose a progressive view alignment strategy that iteratively updates the aligned set and model parameters. This approach ensures the effective utilization of reliable view combinations, enhancing view correspondences and clustering performance, particularly under noise and view-specific information perturbations.
- Extensive experiments conducted on multiple benchmark datasets, including both partially and fully aligned scenarios, validate the effectiveness and superiority of our framework in addressing view alignment challenges in partially aligned multi-view data.

Related Work

Multi-view clustering (MVC) aims to enhance clustering performance by leveraging the complementary and consistent information across multiple views, with various methods developed based on different underlying assumptions (Liu et al. 2024a; Yan et al. 2023; Ren et al. 2024; Gao et al. 2025b). For instance, Qin et al. address high-dimensional data by enforcing structural and sample assignment consistency across views, ensuring identical numbers of connected components in similarity matrices and learning shared subspace representations through alternating optimization (Qin et al. 2022). Liu et al. propose an anchor-based method that improves anchor quality by incorporating inter-view correlations. By constructing a view graph from aligned anchor graphs and utilizing relationships across views, they enhance anchors and boost clustering performance (Liu et al. 2024b). Similarly, UMCGL (Du et al. 2024) balances consistency and diversity in multi-view graph learning through four modules: multi-channel graph, generative, contrastive, and consensus, addressing challenges such as noise and varying data distributions while improving consistency within views and across groups.

While traditional MVC methods have made significant advances, they typically assume that views are complete and perfectly aligned. In real-world applications, however, these assumptions are often violated, which introduces the challenges of Incomplete Multi-View Clustering (IMVC) and PVC. IMVC explicitly addresses the scenario where some views may be missing, requiring methods that can handle incomplete multi-view data. Recent studies (Lin et al. 2021; Liu et al. 2023a,b) have been proposed to tackle these challenges. For example, SMILE (Zeng et al. 2023) leverages invariant semantic distributions across views to mitigate cross-view discrepancies, allowing effective clustering even in the presence of incomplete data. Similarly, Pu et al. propose adaptive feature imputation with latent graphs, improving clustering by integrating view-specific encoders and adaptive strategies based on global pseudo-labels and local assignments (Pu et al. 2024). Xu et al. employ variational autoencoders with view-specific encoders and a Product-of-Experts framework, incorporating a coherence objective and a Mixture-of-Gaussians prior to enhance shared representations for better clustering performance (Xu et al. 2024).

Compared with MVC and IMVC, PVC specifically addresses the challenge of incomplete cross-view correspondences, with several methods developed to leverage partial correspondences effectively (Zhao and Xie 2024; Yang et al. 2023; Cao, Dong, and Chen 2024; Yu et al. 2021). CGCN (Wang et al. 2024b) introduces an alignment-free cross-view graph contrastive learning framework that utilizes nearest neighbor relationships to learn consistent information across both aligned and unaligned instances, effectively handling unaligned data without requiring explicit alignment. Yang et al. proposed a noise-robust contrastive learning approach to capture cross-view consistency representations, leveraging cross-view Euclidean distance to reconstruct the correspondences between views (Yang et al. 2022). In contrast to the previous methods, UPMGC-SM (Wen et al. 2023) refines view alignments by utilizing cross-view similarity graphs,

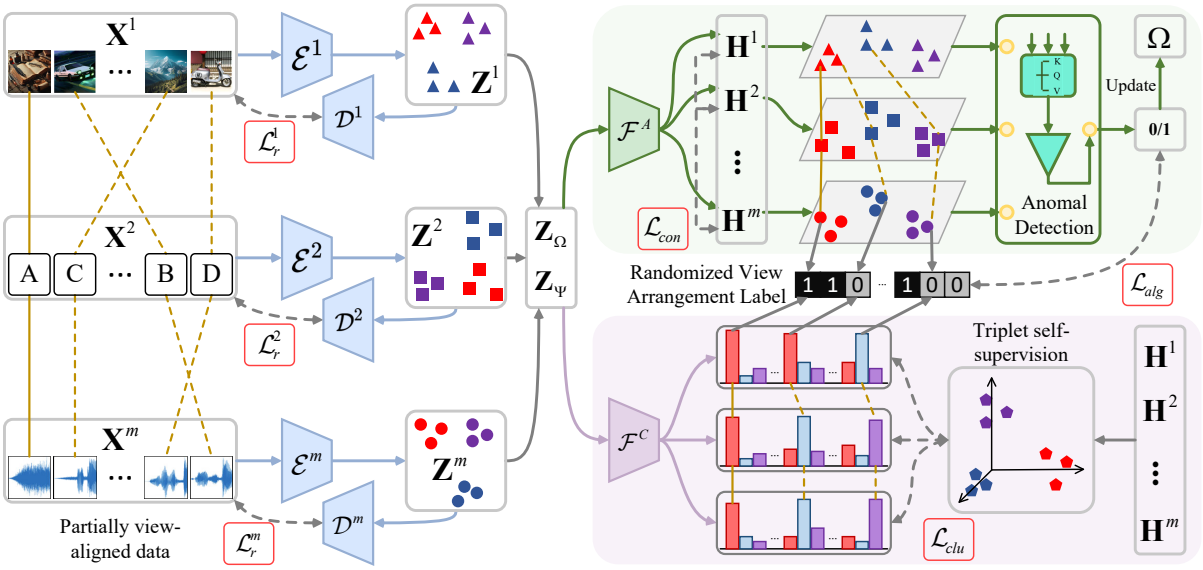

Figure 2: The architecture illustrates how our model processes partial view-aligned data $\{\mathbf{X}^v\}_{v=1}^m$. Each view undergoes view-specific autoencoders to acquire view-specific features $\{\mathbf{Z}^v\}_{v=1}^m$. The green and purple shaded areas represent the progressive view-alignment and consistency clustering modules, respectively. The solid and dashed golden lines represent correct and incorrect correspondences, respectively.

providing a parameter-free solution for multi-view clustering in partially unpaired scenarios.

Unlike previous methods that rely on distance-based alignment, our approach frames view alignment as an anomaly detection task. By identifying misaligned representations as anomalies relative to the expected distribution, we eliminate the need for explicit alignment and use pseudo-labels from the self-supervised consistency clustering module to characterize these outliers. Another key feature is our progressive alignment strategy, which iteratively prioritizes the most reliable re-aligned data, improving robustness to noise and view-specific information perturbations.

Methodology

Preliminaries

Given a multi-view dataset $\{\mathbf{X}^v\}_{v=1}^m$ with c clusters, where m represents the number of views, N is the number of samples, $\mathbf{X}^v = \{\mathbf{X}_1^v, \mathbf{X}_2^v, ..., \mathbf{X}_N^v\}$ denotes the instance set of the v-th view, and $\mathbf{X}_i^v$ is the i-th sample in the v-th view. In the context of partial view alignment, view correspondences may be incomplete. The dataset can thus be partitioned into two subsets: aligned data $\mathbf{X}_\Omega$ and misaligned data $\mathbf{X}_\Psi$, where Ω and Ψ represent the index sets of aligned and misaligned samples, respectively. The primary objective of PVC is to re-align the misaligned instances $\mathbf{X}_\Psi$ and utilize the consistency and complementarity of the re-aligned data to enhance clustering performance.

To tackle the challenges posed by multi-view data, which often contains redundant or noisy information, we first employ view-specific autoencoders to extract discriminative features for each view, as shown in Fig. 2. These autoencoders map the raw data into a compact latent space that retains essential view-specific characteristics while facilitating cross-view alignment. A reconstruction loss is introduced to ensure that the learned features accurately reconstruct the original data, thereby maintaining view-specific information. The reconstruction loss is defined as:

$$\mathcal{L}_r = \sum_{v=1}^{m} \sum_{i=1}^{N} \|\mathbf{X}_i^v - \mathcal{D}^v(\mathcal{E}^v(\mathbf{X}_i^v))\|_F^2, \tag{1}$$

where $\mathcal{E}^v()$ and $\mathcal{D}^v()$ are the encoder and decoder for the v-th view, $\mathbf{Z}_i^v = \mathcal{E}^v(\mathbf{X}_i^v)$ denotes the learned latent feature. This formulation enables efficient feature extraction while ensuring that the latent representations preserve view-specific information, thereby facilitating better alignment for subsequent clustering tasks.

Anomaly Combination Detection Driven Progress View Alignment

Motivation. MVC excels at integrating information across views; however, PVC faces the challenge of missing cross-view correspondences, which impedes the direct exploitation of their consistency and complementary information. Thus, the key task in PVC is to establish these correspondences accurately. We argue that distance-based view alignment methods are susceptible to view-specific information and noise, adversely affecting clustering performance. Consequently, instead of relying on distance-based alignment approaches, we redefine the view alignment problem as an anomaly detection task involving a combination of anomalies. As illustrated in Fig. 2, our view alignment process treats the fused correctly aligned data as normal points and leverages data augmentation to construct outlier points based on clustering-consistent semantics. Without loss of

generality, for two views, normal and anomalous combinations can be formally defined as follows:

$$\mathbf{Y}_{i,j} = \begin{cases} 1, & argmax(\mathbf{U}_i^1) = argmax(\mathbf{U}_j^2) \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathbf{U}$ represents the clustering-consistency semantics extracted from the consistency clustering module, and the $argmax(*)$ function returns the index corresponding to the maximum cluster assignment probability. A combination is considered normal only when instances from each view belong to the same cluster; otherwise, it is classified as anomalous. These normal and anomalous combinations are then used to train the anomaly detection function $g()$. Therefore, our primary task is transformed into learning a consistent clustering pseudo-label to define anomalies, and training an anomaly detection module to identify these anomalies. Notably, these two processes mutually supervise each other within the proposed model.

Since the latent space $\mathbf{Z}$ must capture both view-specific information and noise, data distributions across views may differ. Our view alignment module employs a cross-view contrastive learning strategy to ensure distributional consistency. This involves minimizing the cosine similarity between features from different views, learned through a shared alignment MLP $\mathcal{F}^A$. The process is formulated as:

$$\mathbf{H}^v = \mathcal{F}^A(\mathbf{Z}^v), \tag{3}$$

where $\{\mathbf{H}^v\}_{v=1}^m$ denotes the distribution-consistent semantics, which is constrained by the cross-view contrastive loss applied to the aligned subset, as formulated below:

$$\ell_d(\mathbf{H}^a, \mathbf{H}^b) = \\ -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{d(\mathbf{H}_i^a, \mathbf{H}_i^b)/\tau_D}}{\sum_{j=1}^{N}\sum_{v=a,b} e^{d(\mathbf{H}_i^a, \mathbf{H}_j^v)/\tau_D} - e^{1/\tau_D}}, \tag{4}$$

where τ_D is the temperature parameter, and $d(,)$ represents the cosine similarity between two vectors. Considering the partial view alignment problem, the overall cross-view contrastive loss is given by:

$$\mathcal{L}_{con} = \frac{1}{2}\sum_{a=1}^{m}\sum_{b\neq a} \ell_d(\mathbf{H}_\Omega^a, \mathbf{H}_\Omega^b). \tag{5}$$

To train the anomaly detection $g()$ to identify incorrect view alignments, we utilize known aligned data as normal combinations and construct anomalous combinations through data augmentation on unaligned data. Specifically, random sampling is performed across views within the unaligned data, and clustering-consistency semantics $\mathbf{U}$ are used to determine whether the sampled combinations belong to the same cluster, thereby identifying anomalous view combinations. This process can be formulated as follows:

$$\hat{\mathbf{H}} = \{Sam(\mathbf{H}_\Psi^1), Sam(\mathbf{H}_\Psi^2), ..., Sam(\mathbf{H}_\Psi^m)\}, \\ s.t.\ \mathbf{Y} = 0 \tag{6}$$

where Sam denotes the random sampling function, and then, the objective can be formulated as:

$$\mathcal{L}_{alg} = \|1 - g(\mathbf{H}_\Omega)\|_F^2 + \left\|g(\hat{\mathbf{H}})\right\|_F^2. \tag{7}$$

where $g()$ consists of a multi-head self-attention layer for feature fusion, followed by an MLP and a Sigmoid activation function.

In addition, our progressive view alignment strategy aims to mitigate further the influence of view-specific information and noise on the alignment process. This is accomplished through an iterative procedure that concurrently updates the aligned subset and model parameters. Specifically, after a fixed number of epochs during model training, we update the aligned Ω and unaligned Ψ subsets as follows:

$$\Omega \leftarrow argmax(Top10(g(\hat{\mathbf{H}}))), \tag{8}$$

This formula selects the top $N/10$ best random combinations from the unaligned subset Ψ as aligned data. Subsequently, the aligned subset Ω is used to update the model parameters further.

Triplet Self-supervised Consistency Clustering

We introduce a triplet self-supervised consistency clustering module with two primary objectives: (1) generating consistency clustering pseudo-labels for each view and (2) mitigating the influence of incorrect view alignments on the final clustering results. As previously mentioned, the anomaly detection mechanism in our progressive view alignment module relies on clustering pseudo-labels to detect misalignments. Therefore, the quality of these pseudo-labels directly impacts alignment performance.

To address these objectives, we leverage a triplet self-supervised consistency learning framework that integrates cross-semantic, cross-view, and cross-feature self-supervision to learn clustering-consistent semantics. In particular, we introduce a shared clustering MLP, denoted as $\mathcal{F}^C$, to learn clustering-consistency semantics $\{\mathbf{U}^v \in \mathbb{R}^{N\times c}\}_{v=1}^m$ from the view-specific latent features $\{\mathbf{Z}^v\}_{v=1}^m$. This process can be mathematically represented as follows:

$$\mathbf{U}^v = Softmax(\mathcal{F}^C(\mathbf{Z}^v)). \tag{9}$$

The matrix $\mathbf{U}^v$ represents the probability distribution of each instance assigned to a specific cluster. Specifically, each entry $u_{i,j}^v$ in $\mathbf{U}^v$ corresponds to the probability that the i-th instance in the v-th view is assigned to the j-th cluster. Inspired by Contrastive clustering (Li et al. 2021), we aim to achieve clustering consistency by ensuring that instances of each sample across different views share the same clustering assignment vector. The cross-view self-supervision is then defined using the following contrastive loss:

$$\ell_v(\mathbf{U}^a, \mathbf{U}^b) = -\frac{1}{c} \\ \sum_{i=1}^{c} \log \frac{e^{d(\mathbf{U}_{:,i}^a, \mathbf{U}_{:,i}^b)/\tau_C}}{\sum_{j=1}^{c}\sum_{v=a,b} e^{d(\mathbf{U}_{:,i}^a, \mathbf{U}_{:,j}^v)/\tau_C} - e^{1/\tau_C}}, \tag{10}$$

where τ_C is the temperature parameter and $\mathbf{U}_{:,i}^a$ denotes the i-th column of the matrix $\mathbf{U}^a$ in the a-th view. To avoid

Algorithm 1: Optimization Algorithm

Input: Partially view-aligned dataset $\{\mathbf{X}^v\}_{v=1}^m$, the index sets of aligned Ω and unaligned Ψ data, and the number of clusters c

Output: Clustering result and trained model

1. Pre-training:

Pre-train autoencoders via reconstruction loss $\mathcal{L}_r$.

2. Training using all data:

while $t < T$ **do**

- Update view-specific autoencoders using $\mathcal{L}_r$ to obtain latent features $\mathbf{Z}$.
- Refine the distribution-consistent semantics $\mathbf{H}$ and clustering-consistent semantics $\mathbf{U}$ from $\mathbf{Z}$ and Ω using $\mathcal{L}_{con}$ and $\mathcal{L}_{clu}$.
- Construct anomalous view combinations $\hat{\mathbf{H}}$ using Eq. 6 based on the unligned subset Ψ.
- Update the anomaly detection g using $\mathcal{L}_{alg}$, and refresh the aligned index set Ω by Eq. 8.
- $t = t + 1$.

end

3. Clustering: The clustering results are obtained from cluster consistency features via argmax.

the occurrence of multiple peaks in the clustering assignment vector, we introduce a cross-feature self-supervision constraint, defined as:

$$\ell_f(\mathbf{U}^v) = \sum_{j=1}^{c}(\frac{1}{N}\sum_{i=1}^{N} u_{ij}^v)\log(\frac{1}{N}\sum_{i=1}^{N} u_{ij}^v) \tag{11}$$

Additionally, to prevent invalid solutions where all instances are assigned to a single cluster while enhancing clustering performance, we combine the clustering-consistency semantics with the distribution-consistent semantics from the progressive view alignment. We then implement cross-semantic self-supervision using the following cross-entropy loss:

$$\ell_s(\mathbf{Q}, \mathbf{U}^v) = -\mathbf{Q}\log\mathbf{U}^v, \tag{12}$$

where $\mathbf{Q}$ is the one-hot encoding obtained by applying K-means clustering on the combination of the distribution-consistent and clustering-consistent semantics, finally, the overall clustering loss is expressed as:

$$\begin{aligned}\mathcal{L}_{clu} = &\frac{1}{2}\sum_{a=1}^{m}\sum_{b\neq a}\ell_v(\mathbf{U}_\Omega^a, \mathbf{U}_\Omega^b) \\ &+ \sum_{v=1}^{m}\ell_f(\mathbf{U}_\Omega^v) + \sum_{v=1}^{m}\ell_s(\mathbf{Q}_\Omega, \mathbf{U}_\Omega^v)\end{aligned} \tag{13}$$

Optimization

We adopt a two-phase strategy: pre-training to optimize view-specific autoencoders, and fine-tuning to update model parameters and cross-view correspondences in an alternating manner. The training process is summarized as follows:

Dataset	Samples	Classes	Features
WebKB	1051	2	{2949,334}
Caltech5V	1400	7	{40,254,928,512,1984}
HandWritten	2000	10	{240,216,76,47,64,6}
BDGP	2500	4	{1750,79}
Fashion	10000	10	{784,784,784}
Reuters	18758	6	{10,10,10,10,10}

Table 1: Statistics of the datasets.

- We adopt a pre-training strategy to optimize the parameters of view-specific autoencoders. In this phase, both the encoder and decoder are trained with a reconstruction loss $\mathcal{L}_r$, ensuring that the input of each view is mapped into a unified latent space of consistent dimensionality.
- The fine-tuning alternates between two key steps: updating model parameters to obtain clustering-consistent semantics, distribution-consistent semantics, and anomaly detection results, and updating the aligned and unaligned index sets, Ω and Ψ. First, the progressive view alignment module is optimized using distribution alignment loss $\mathcal{L}_{con}$ and anomaly detection loss $\mathcal{L}_{alg}$. In contrast, the consistency clustering module uses the triplet self-supervised consistency clustering loss $\mathcal{L}_{clu}$. Then, the most reliable view combinations, identified by anomaly detection, are assigned to the aligned subset Ω for further optimization. These steps are iteratively refined to improve performance.
- Finally, a unified representation is derived by averaging the re-aligned clustering consistency semantics across views. The clustering results are subsequently obtained by identifying the index of the maximum value in the feature dimensions.

The overall optimization process of the proposed method is summarized in Algorithm 1.

Experiments

Experimental Settings

Datasets. Six popular multi-view datasets are used to evaluate the proposed method, including: **WebKB**, **Caltech5V** (Xu et al. 2024), **HandWritten** (Liu et al. 2023b), **BDGP** (Wang et al. 2024a), **Fashion** (Xu et al. 2022), **Reuters** (Yang et al. 2021). A brief description of these datasets is summarized in Table 1.

Compared methods. We compare our method with seven MVC methods, including: LMVSC (Kang et al. 2020), GMC (Wang, Yang, and Liu 2019), SMVSC (Sun et al. 2021), OPMC (Liu et al. 2021), AE^2-NETs (Zhang, Liu, and Fu 2019), FastMICE (Huang, Wang, and Lai 2023), MFLVC (Xu et al. 2022), and seven PVC methods, including: PVC (Huang et al. 2020), MVC-UM (Yu et al. 2021), MvCLN (Yang et al. 2021), SURE (Yang et al. 2022), VITAL (He et al. 2024), CGCN (Wang et al. 2024b), CGDA (Wang et al. 2024a). For a fair comparison, we present the experimental results of all competing methods by either referencing the reported outcomes from relevant studies (Yang et al. 2021; Wang et al. 2024b,a) or by utilizing the released source code with their recommended optimal parameter settings.

Methods	WebKB			Caltech2V			HandWritten-2V			BDGP			Fashion-2V			Reuters-2V		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
GMC	77.93	0.50	1.76	19.71	10.25	0.49	43.50	51.29	30.50	31.40	16.06	2.56	47.28	55.96	23.42	26.98	4.50	0.46
AE^2-NETs	60.61	0.17	1.35	17.43	1.44	0.40	69.10	66.45	56.08	39.16	17.77	6.15	65.43	61.09	47.86	35.49	10.61	8.07
LMVCS	61.27	0.81	1.00	43.14	26.84	18.98	75.55	67.35	59.73	48.20	25.07	17.08	65.99	60.19	49.60	45.13	17.86	17.03
SMVSC	72.31	4.86	13.45	36.78	19.74	13.45	64.20	53.14	42.26	50.88	29.63	24.59	59.95	54.28	43.83	45.72	14.98	17.09
OPMC	55.19	5.59	0.17	37.36	21.09	14.98	56.15	51.34	39.66	46.36	27.52	22.53	57.10	61.27	47.24	40.83	11.54	10.73
MFLVC	66.03	13.43	10.22	41.79	29.33	23.33	54.55	42.12	30.32	59.88	27.82	26.49	80.68	70.16	65.04	37.43	16.50	12.78
FastMICE	78.74	3.23	10.53	47.91	30.75	25.08	72.67	63.94	56.01	57.68	30.60	27.42	70.58	63.22	54.77	34.91	15.96	10.87
PVC	17.32	6.44	15.33	26.64	20.34	12.06	76.45	74.47	66.22	84.76	64.83	66.17	68.34	68.66	54.82	42.07	20.43	16.95
MVC-UM	72.12	0.05	0.82	46.28	31.91	25.33	71.45	69.16	60.47	46.68	21.88	8.81	52.07	47.15	34.21	34.02	11.10	12.25
MvCLN	68.98	4.17	10.53	54.21	42.55	34.26	73.10	70.00	57.93	73.04	46.15	44.28	84.37	75.48	71.46	50.16	30.65	24.90
SURE	63.65	1.62	4.90	32.71	16.34	10.04	77.31	72.42	63.01	79.29	57.95	55.87	84.88	77.19	72.76	50.09	29.61	22.08
VITAL	60.49	12.20	4.29	58.63	48.10	40.49	87.18	76.61	73.79	66.14	50.15	42.63	90.06	81.62	80.20	50.34	35.52	21.92
CGCN	57.09	12.35	0.09	60.29	49.19	40.38	71.15	71.79	61.05	86.64	66.41	69.91	81.35	74.22	68.39	51.81	31.98	26.61
CGDA	78.12	4.47	4.26	53.07	38.96	30.96	83.16	79.24	73.01	90.74	77.40	78.84	OM	OM	OM	OM	OM	OM
Ours	**83.63**	**24.22**	**40.25**	**63.14**	**50.70**	**42.29**	**94.45**	**88.56**	**88.19**	**92.76**	**80.57**	**82.90**	**91.45**	**84.66**	**82.81**	**54.79**	**35.38**	**26.78**

Table 2: Comparison of clustering performance under the partially aligned two-view setting on six benchmark datasets. Best results are in bold, second-best in underline. "OM" denotes "Out of Memory."

Methods	Caltech3V			Caltech4V			Caltech5V			Fashion-3V			Reuters-5V			HandWritten-6V		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
GMC	17.79	6.23	0.18	18.21	8.44	0.20	29.00	15.98	4.23	60.31	63.49	39.24	31.91	9.41	0.35	42.25	51.98	18.00
LMVCS	43.71	30.24	21.04	41.00	22.27	15.64	44.85	27.33	20.01	64.60	57.16	45.45	42.09	15.46	15.72	67.65	65.07	54.19
SMVSC	41.57	18.97	13.77	43.50	19.93	12.52	45.14	22.08	14.57	57.19	53.75	46.03	44.39	13.98	15.15	63.65	59.24	48.72
OPMC	37.00	16.63	11.73	42.50	22.92	16.02	47.43	22.37	17.20	63.06	59.74	47.03	40.88	20.32	12.61	82.15	75.44	68.97
MFLVC	33.57	22.37	15.28	35.64	17.92	11.39	37.29	19.31	13.60	80.46	68.80	64.04	33.54	13.59	9.16	69.35	58.21	48.78
FastMICE	41.74	23.62	16.28	42.21	22.71	15.10	43.21	24.12	18.27	69.85	61.70	57.90	32.25	13.69	8.66	78.99	69.62	63.48
PVC$^+$	44.57	32.43	22.30	42.79	33.50	22.27	44.86	34.78	25.32	66.69	70.27	52.96	40.33	22.53	15.03	29.35	36.83	19.13
MVC-UM	48.42	32.68	25.91	48.71	32.76	25.93	61.14	43.20	36.07	57.30	55.67	42.04	36.74	14.15	15.31	70.90	69.08	58.52
VITAL$^+$	64.00	51.72	43.89	71.17	57.76	51.48	76.14	59.52	52.69	90.52	80.03	80.48	53.18	33.71	20.06	78.70	74.17	67.15
CGCN	63.43	**55.75**	**46.31**	66.50	55.70	48.77	73.93	**66.57**	**61.84**	90.02	82.81	80.34	46.27	28.09	23.97	56.95	51.73	39.66
Ours	**65.29**	53.95	45.87	**71.79**	**59.63**	**53.73**	**78.00**	62.65	59.22	**93.70**	**87.73**	**86.92**	**56.65**	**37.81**	**30.42**	**86.20**	**84.87**	**79.29**

Table 3: Comparison of clustering performance under the partially aligned more than two-view setting on six benchmark datasets. Best results are in bold, second-best in underline.

$\mathcal{L}_{clu}$	$\mathcal{L}_{con}$	$\mathcal{L}_{alg}$	Caltech3V			HandWritten-2V			Reuters-5V		
			ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
			24.50	9.17	2.12	10.00	0.00	0.00	29.72	2.56	0.32
	✓		27.93	19.29	5.91	18.70	22.73	4.77	39.96	14.92	7.74
		✓	24.57	9.72	3.94	22.35	28.36	7.62	35.45	5.05	1.80
	✓	✓	31.00	22.44	7.93	25.90	30.32	6.57	31.06	14.07	0.86
✓			60.07	48.01	40.16	85.15	85.79	79.83	53.14	33.15	28.12
✓	✓		61.21	51.74	41.15	93.00	86.04	85.20	54.66	36.33	28.45
✓		✓	63.07	51.25	44.09	85.25	85.07	79.16	50.04	36.84	30.14
w/o Progressive			64.07	52.65	45.19	93.60	86.92	86.42	54.85	37.36	30.35
Complete			**65.29**	**53.95**	**45.87**	**94.45**	**88.56**	**88.19**	**56.65**	**37.81**	**30.42**

Table 4: Ablation studies on different loss functions and progressive training strategy.

Evaluation metrics. To assess the clustering performance, we utilize three commonly used evaluation metrics: Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI).

Clustering Results Comparision

Comparison in classic settings. To evaluate the effectiveness of the proposed method, we comprehensively compared its clustering performance with that of the state-of-the-art (SOTA) MVC and PVC methods. Given the limitations of traditional MVC methods in addressing partial alignment issues, we first applied PCA for dimensionality reduction and used the Hungarian algorithm for realignment before clustering. Since most PVC methods are evaluated on two-view datasets, we present clustering results with a 50% alignment rate in a two-view setting for fair comparison, as shown in Table 2. The key observations are as follows: 1) PVC methods, designed specifically for partial alignment scenarios, consistently outperform traditional MVC methods, particularly on datasets such as BDGP, HandWritten-2V, and Fashion-2V. This improvement can be attributed to the ability of PVC methods to realign partially aligned data, thereby enhancing clustering outcomes effectively. 2) Compared to other advanced PVC methods, our approach demonstrates consistent superiority across all datasets, highlighting the effectiveness of the proposed progressive view alignment strategy and triplet self-supervised clustering mechanism in handling partial alignment challenges.

Comparision of varying view number. To further evaluate the robustness of our model in multi-view scenarios, we compared clustering performance under 50% alignment rate across varying number of views, as shown in Table 3. For fairness, we extended PVC and VITAL into multi-view settings, denoted as PVC$^+$ and VITAL$^+$, respectively. In contrast, MvCLN, SURE, and CGDA, which rely on mechanisms tailored specifically for two-view setting, cannot be easily adapted to multi-view scenarios. The results demonstrate that VITAL$^+$, CGCN, and our approach consistently exhibit clear advantages in multi-view alignment tasks. Our method achieves the best or second-best clustering performance across all settings, underscoring its robustness in effectively handling view alignment beyond two views.

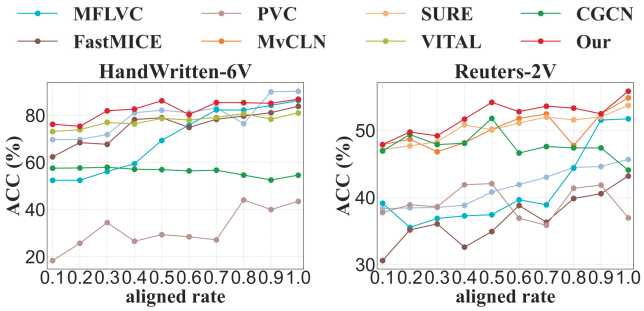

Figure 3: The clustering performance compared with several competing methods under varying alignment ratios.

Comparision of varying alignment ratio. Finally, we assessed the performance of our method under varying alignment ratios, ranging from 10% to 100% in increments of 10%, across two datasets. As shown in Fig. 3, the clustering performance of all PVC methods improves with increasing alignment ratios. However, our method consistently outperforms other SOTA approaches across all alignment ratios. Notably, it maintains competitive performance even at lower alignment ratios, demonstrating its robustness and reliability. These findings suggest that higher alignment ratios enhance clustering performance by providing more cross-view correspondences, and our method remains stable and effective under challenging conditions with limited alignment.

Model Analysis

Ablation Studies. We conducted ablation experiments with different loss combinations to evaluate the effectiveness of the proposed model. By analyzing the clustering results across three datasets in Table 4, it is evident that the best performance is achieved when all three loss functions are utilized simultaneously. The clustering loss $\mathcal{L}_{clu}$ plays a critical role in determining the quality of pseudo-labels, directly influencing both clustering results and view alignment. Consequently, removing the clustering loss results in a significant decline in performance. Additionally, the clustering performance improves markedly when contrastive loss $\mathcal{L}_{con}$ and view alignment loss $\mathcal{L}_{alg}$ are used together, rather than in isolation, underscoring that their combination enhances the effectiveness of view alignment. Finally, "w/o Progressive" denotes excluding the progressive training strategy. The results highlight the effectiveness of prioritizing reliable realigned samples in our progressive training approach.

Effective of View Alignment. To illustrate that reformulating the view alignment task as an anomaly detection problem offers greater robustness than distance-based alignment methods, we visually compare the heatmaps for anomaly scores and cross-view cosine similarities in a two-view setting. As shown in Fig. 4, the anomaly score heatmap reveals a more distinct diagonal block structure, with a sharper contrast between correct and incorrect alignments, thereby emphasizing the efficacy of our view alignment strategy.

To further validate the effectiveness of our model in addressing view alignment tasks involving more than two

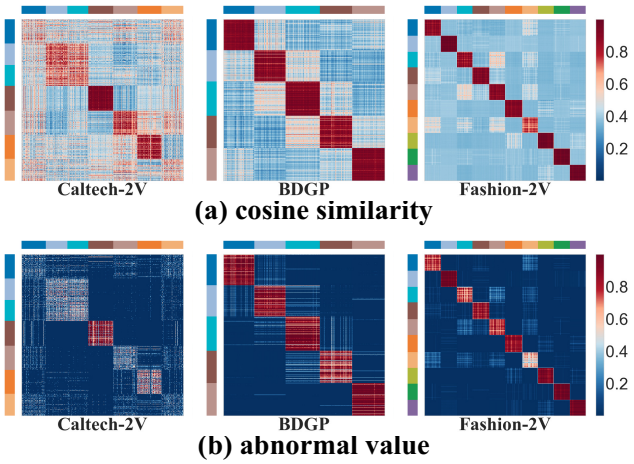

Figure 4: The visualization comparison of cosine similarity (a) and anomaly scores (b) under a 50% alignment rate.

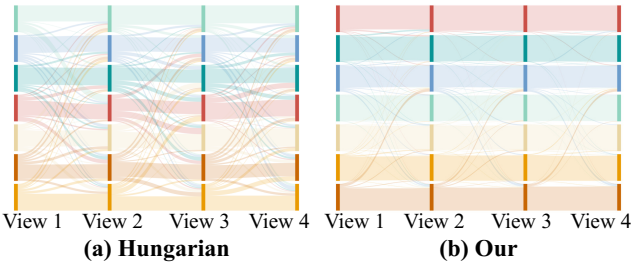

Figure 5: The Sankey diagram compares view alignment on Caltech4V using the Hungarian algorithm and our method. Columns represent views, and colors denote clusters. Ideally, perfect alignment shows a one-to-one correspondence between views within each cluster.

views, we assessed the quality of the constructed cross-view correspondences. Specifically, Sankey diagrams were employed to visually compare the view alignment results produced by the Hungarian algorithm and our proposed method, as illustrated in Fig. 5. The results highlight the superior capability of our method in accurately establishing cross-view correspondences, thereby significantly enhancing the efficiency and effectiveness of model training.

Conclusion

This study introduces a novel framework for PVC, reformulating view alignment as an anomaly combination detection task. Unlike traditional methods sensitive to noise and inconsistencies, our approach combines a progressive view alignment module with a self-supervised consistency clustering module. These modules iteratively realign data and optimize model parameters, effectively addressing noise and incomplete alignments. Experiments on benchmark datasets demonstrate that our method achieve superior view alignment and clustering performance. This work presents a robust solution to PVC challenges and lays the foundation for future advancements in multi-view learning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62372494, 62202334); Natural Science Foundation of Jilin Province (No. 20240302086GX); Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011692, 2023A1515030154).

References

Cao, J.; Dong, W.; and Chen, J. 2024. View-unaligned clustering with graph regularization. *Pattern Recognition*, 110706.

Du, S.; Cai, Z.; Wu, Z.; Pi, Y.; and Wang, S. 2024. UMCGL: Universal Multi-view Consensus Graph Learning with Consistency and Diversity. *IEEE Transactions on Image Processing*.

Gao, H.; Liu, C.; Cai, Z.; Sun, H.; Li, G.; Li, Y.; and Du, W. 2025a. A Novel Approach for Effective Partially View-Aligned Clustering with Triple-Consistency. *IEEE Transactions on Circuits and Systems for Video Technology*.

Gao, H.; Liu, C.; Sun, H.; Li, G.; Li, Y.; Zhou, Y.; and Du, W. 2025b. Incomplete Multi-View Clustering with Cross-View Generation via Pre-trained Transformer. *Pattern Recognition*, 112166.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 2551–2566.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *International conference on machine learning*, 4116–4126. PMLR.

He, C.; Zhu, H.; Hu, P.; and Peng, X. 2024. Robust Variational Contrastive Learning for Partially View-unaligned Clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4167–4176.

Huang, D.; Wang, C.-D.; and Lai, J.-H. 2023. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. *IEEE Transactions on Knowledge and Data Engineering*.

Huang, Z.; Hu, P.; Zhou, J. T.; Lv, J.; and Peng, X. 2020. Partially view-aligned clustering. *Advances in Neural Information Processing Systems*, 33: 2892–2902.

Jiang, G.; Peng, J.; Wang, H.; Mi, Z.; and Fu, X. 2022. Tensorial multi-view clustering via low-rank constrained high-order graph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8): 5307–5318.

Kang, Z.; Zhou, W.; Zhao, Z.; Shao, J.; Han, M.; and Xu, Z. 2020. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 4412–4419.

Li, X.; Pan, Y. P.; Sun, Y.; Sun, Q. S.; Tsang, I. W.; and Ren, Z. 2024. Fast Unpaired Multi-view Clustering. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*.

Li, Y.; Hu, P.; Liu, Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8547–8555.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11174–11183.

Liu, C.; Li, R.; Che, H.; Leung, M.-F.; Wu, S.; Yu, Z.; and Wong, H.-S. 2024a. Latent Structure-Aware View Recovery for Incomplete Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, C.; Li, R.; Wu, S.; Che, H.; Jiang, D.; Yu, Z.; and Wong, H.-S. 2023a. Self-guided partial graph propagation for incomplete multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, C.; Wen, J.; Wu, Z.; Luo, X.; Huang, C.; and Xu, Y. 2023b. Information recovery-driven deep incomplete multi-view clustering network. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, C.; Wu, S.; Jiang, D.; Yu, Z.; and Wong, H.-S. 2022. View-aware collaborative learning for survival prediction and subgroup identification. *IEEE Transactions on Biomedical Engineering*, 70(1): 307–317.

Liu, J.; Liu, X.; Yang, Y.; Liu, L.; Wang, S.; Liang, W.; and Shi, J. 2021. One-pass multi-view clustering for large-scale data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12344–12353.

Liu, S.; Liang, K.; Dong, Z.; Wang, S.; Yang, X.; Zhou, S.; Zhu, E.; and Liu, X. 2024b. 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*, 26151–26161.

Pan, E.; and Kang, Z. 2021. Multi-view contrastive graph clustering. *Advances in neural information processing systems*, 34: 2148–2159.

Pu, J.; Cui, C.; Chen, X.; Ren, Y.; Pu, X.; Hao, Z.; Philip, S. Y.; and He, L. 2024. Adaptive Feature Imputation with Latent Graph for Deep Incomplete Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14633–14641.

Qian, S.; Xue, D.; Hu, J.; Zhang, H.; and Xu, C. 2024. Nonparametric Clustering-Guided Cross-View Contrastive Learning for Partially View-Aligned Representation Learning. *IEEE Transactions on Image Processing*.

Qin, Y.; Feng, G.; Ren, Y.; and Zhang, X. 2022. Consistency-induced multiview subspace clustering. *IEEE Transactions on Cybernetics*, 53(2): 832–844.

Ren, Y.; Pu, J.; Cui, C.; Zheng, Y.; Chen, X.; Pu, X.; and He, L. 2024. Dynamic weighted graph fusion for deep multi-view clustering. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 4842–4850. ijcai. org.

Sun, M.; Zhang, P.; Wang, S.; Zhou, S.; Tu, W.; Liu, X.; Zhu, E.; and Wang, C. 2021. Scalable multi-view subspace clustering with unified anchors. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3528–3536.

Wan, X.; Liu, J.; Liang, W.; Liu, X.; Wen, Y.; and Zhu, E. 2022. Continual multi-view clustering. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3676–3684.

Wang, H.; Yang, Y.; and Liu, B. 2019. GMC: Graph-based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6): 1116–1129.

Wang, X.; Gao, H.; Wei, X.; Peng, L.; Li, R.; Liu, C.; Wu, S.; and Wong, H.-S. 2024a. Contrastive Graph Distribution Alignment for Partially View-Aligned Clustering. In *ACM Multimedia 2024*.

Wang, Y.; Chang, D.; Fu, Z.; Wen, J.; and Zhao, Y. 2024b. Partially View-aligned Representation Learning via Cross-view Graph Contrastive Network. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wen, Y.; Wang, S.; Liao, Q.; Liang, W.; Liang, K.; Wan, X.; and Liu, X. 2023. Unpaired Multi-View Graph Clustering With Cross-View Structure Matching. *IEEE Transactions on Neural Networks and Learning Systems*.

Xu, G.; Wen, J.; Liu, C.; Hu, B.; Liu, Y.; Fei, L.; and Wang, W. 2024. Deep Variational Incomplete Multi-View Clustering: Exploring Shared Clustering Structures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16147–16155.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16051–16060.

Yan, W.; Zhang, Y.; Lv, C.; Tang, C.; Yue, G.; Liao, L.; and Lin, W. 2023. GCFAgg: Global and Cross-view Feature Aggregation for Multi-view Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19863–19872.

Yang, M.; Li, Y.; Hu, P.; Bai, J.; Lv, J.; and Peng, X. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1055–1069.

Yang, M.; Li, Y.; Huang, Z.; Liu, Z.; Hu, P.; and Peng, X. 2021. Partially view-aligned representation learning with noise-robust contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1134–1143.

Yang, W.; Xin, L.; Wang, L.; Yang, M.; Yan, W.; and Gao, Y. 2023. Iterative multiview subspace learning for unpaired multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*.

Yu, H.; Tang, J.; Wang, G.; and Gao, X. 2021. A novel multi-view clustering method for unknown mapping relationships between cross-view samples. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2075–2083.

Zeng, P.; Yang, M.; Lu, Y.; Zhang, C.; Hu, P.; and Peng, X. 2023. Semantic invariant multi-view clustering with fully incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, C.; Li, H.; Chen, C.; Jia, X.; and Chen, C. 2022. Low-rank tensor regularized views recovery for incomplete multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, C.; Liu, Y.; and Fu, H. 2019. Ae2-nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2577–2585.

Zhao, L.; and Xie, Q. 2024. Distribution-level multi-view clustering for unaligned data. *IEEE Signal Processing Letters*.

Zhao, L.; Xie, Q.; Wu, S.; and Ma, S. 2023. An End-to-End Framework for Partial View-Aligned Clustering with Graph Structure. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.