DC-SPAN: A Dual Contrastive Attention Network for Multi-View Clustering

Jingyi Chen[1], **Zhibin Dong**[1*], **Tiejun Li**[1*], **Yibo Han**[1]

[1]College of Computer Science and Technology, National University of Defense Technology, Changsha, 410073, China
{chenjingyi20, dzb20, tjli, hyb123}@nudt.edu.cn

Abstract

Multi-view clustering aims to group data by integrating complementary information from multiple views. However, the inherent heterogeneity among views often leads to feature entanglement, severely limiting clustering performance. To address this challenge, we propose DC-SPANa Dual Contrastive Attention Networkgrounded in a disentangle-then-fuse paradigm. DC-SPAN employs a dual-path variational architecture to explicitly decompose each view into shared and private latent subspaces. These representations are then robustly integrated via a Product-of-Experts (PoE) mechanism. At the heart of our model is a novel dual contrastive learning objective that simultaneously encourages alignment of shared components across views and enforces separation of private ones, enabling structured and disentangled representations. A gated attention fusion module further adaptively aggregates these latent factors to yield a unified, discriminative embedding. The overall model is trained end-to-end using a composite loss function that incorporates reconstruction, orthogonality, and contrastive terms, along with a two-stage training scheme for improved stability. Extensive experiments on benchmark datasets demonstrate that DC-SPAN consistently outperforms existing state-of-the-art methods, highlighting its effectiveness and robustness in handling multi-view heterogeneity.

Introduction

In many real-world applications, data are naturally characterized by multiple heterogeneous views, where each view provides a unique and complementary perspective on the same underlying entity. For instance, a news event can be described by video footage, audio recordings, and textual articles. Different from single-view clustering (Sun et al. 2023, 2024a; Sun, Peng, and Ren 2024), multi-view clustering (MVC) (Dong et al. 2023b; Chen et al. 2023; Zhang et al. 2022; Hu et al. 2023; Wen et al. 2020; Wang et al. 2022b) seeks to leverage this rich, multi-faceted information to achieve more robust and comprehensive data partitioning than is possible from any single view alone. Traditional MVC methods often rely on techniques such as subspace learning (Cui et al. 2023; Shi et al. 2024; Zhu et al. 2024; Gu and Feng 2024; Huang et al. 2024a; Gu, Li, and Feng 2024; Huang et al. 2025; Wang et al. 2022a) or graph fusion (Dong et al. 2023a; Zhang et al. 2020; Huang et al. 2024b; Wen et al. 2024; Wang et al. 2022c) to find a consensus representation. While effective for certain data types, these approaches often depend on linear assumptions and can struggle to capture the complex, non-linear relationships inherent in high-dimensional data.

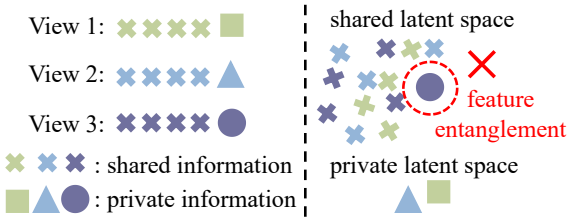


Figure 1: A simple illustration of feature entanglement: without explicit separation of private and shared information, multi-view representations may mix modality-specific and common signals, leading to confounded embeddings.

Recent advancements in deep multi-view clustering have demonstrated significant progress, particularly methods based on contrastive learning (Lin et al. 2021; Trosten et al. 2021). A common strategy in these approaches is to maximize cross-view consistency by treating different views of the same instance as positive pairs and aligning their representations in a shared latent space. However, this paradigm implicitly assumes that discrepancies between views are primarily noise, as shown in Figure 1, thereby neglecting the valuable, discriminative information that is unique to each view. Consequently, view-specific features are frequently erroneously incorporated into shared representations, a phenomenon referred to as feature entanglement.

This entanglement leads to several critical issues. First, it contaminates the shared representation with view-specific noise and irrelevant details, degrading its quality. Second, it forces the model to discard potentially crucial private semantics, thereby diminishing the overall discriminative power of the learned embedding. Finally, by failing to properly model the complementary nature of shared and private information, the resulting representations are often confounded and suboptimal for clustering.

*Corresponding author.

While some recent works have attempted to mitigate this problem through architectural decoupling or by designing separate learning objectives (Xu et al. 2021b; Lu et al. 2024), they often fall short in several aspects. Many lack explicit mechanisms to prevent the mixing of shared and private information, provide insufficient modeling of the private feature space, or employ simplistic fusion strategies that fail to adaptively leverage the complementary strengths of the disentangled factors.

To overcome these limitations, we propose DC-SPAN, a novel dual contrastive attention network. Our framework is designed to explicitly separate, structure, and fuse shared and private information for robust multi-view clustering. The main contributions of our work are as follows:

- We propose a dual-path variational architecture that explicitly disentangles the latent space into shared and private subspaces, effectively mitigating feature entanglement and enhancing model interpretability.
- We employ a PoE inference model to robustly aggregate information for both shared and private factors, which naturally down-weights the influence of noisy or unreliable views. And we introduce a gated attention mechanism that dynamically fuses the shared and private representations, allowing the model to adaptively balance consistency and specificity for each data instance.
- Comprehensive experiments on multiple benchmark datasets demonstrate that DC-SPAN achieves state-of-the-art performance, validating the effectiveness of our approach in handling multi-view heterogeneity.

Related Work

This section provides a concise overview of recent advances in multi-view clustering and the application of contrastive learning.

Multi-view Clustering

MVC aims to derive a unified and coherent partitioning of data by leveraging complementary information from multiple views or modalities (Hu et al. 2023; Sun et al. 2024b; Wang et al. 2022b). Traditional MVC methods can be broadly categorized into subspace-based, graph-based, and matrix factorization approaches. Subspace methods seek a shared latent subspace, with recent works like FSMSC (Chen et al. 2023) using shared anchors and self-supervision to enhance consistency. Graph-based methods, such as DFMVC (Ren et al. 2024) and UMCGL (Du et al. 2024), focus on preserving local geometric structures by integrating view-specific graphs and dynamically fusing them. Matrix factorization techniques, which are often mathematically related to relaxed forms of K-means (McQueen 1967), learn a compact shared representation while preserving data geometry (Wen et al. 2018).

With the advent of deep learning, deep MVC has become the dominant paradigm. Early methods often employed autoencoders to learn latent representations. For instance, AE^2-Nets (Zhang, Liu, and Fu 2019) used a nested autoencoder to implicitly separate shared and private information. Others, like DAMC (Li et al. 2019), utilized adversarial learning to align cross-view distributions. More recent approaches have explored advanced architectures, TTAE (Wang et al. 2025) uses a tensor transformer to model high-order cross-view interactions. To mitigate conflicts during fusion, EPFMVC (Dong et al. 2025b) proposes an enhance then fuse paradigm, which first strengthens view-specific features via channel attention and then progressively fuses views based on a view-graph, starting with the most similar pairs.

Despite their success, a fundamental limitation of many deep MVC methods is their implicit handling of shared and private information. By prioritizing cross-view consistency, these models often inadvertently suppress valuable view-specific characteristics, treating them as noise. This leads to feature entanglement, where the shared representation becomes contaminated with private signals, diminishing its purity and discriminability. While some works have moved towards explicit disentanglement (Xu et al. 2021b; Tang and Liu 2022), they often rely on simplified objectives or structural constraints that may not achieve a comprehensive separation, thus motivating the need for a more principled framework.

Contrastive Learning for Multi-view Clustering

Contrastive learning has emerged as a powerful self-supervised paradigm for learning discriminative representations and has been widely adopted in MVC (Lin et al. 2025; Wang et al. 2023, 2024). The fundamental objective is to pull different views of the same instance (positive pairs) closer in the latent space while pushing apart views from different instances (negative pairs). Early work like CMC (Tian, Krishnan, and Isola 2020) demonstrated the efficacy of maximizing cross-view mutual information. However, a key challenge quickly emerged: aggressive alignment can lead to representation collapse or over-alignment, where view-specific information is lost in the pursuit of consistency, thereby undermining the very benefit of multi-view data.

To address this, more sophisticated contrastive strategies have been developed. These include hierarchical methods like MFLVC (Xu et al. 2022b), which applies contrastive objectives at different feature levels. Predictive mechanisms like COMPLETER (Lin et al. 2021) establishes a theoretical link between view reconstruction and consistency. More recently, decoupled contrastive learning, as seen in DIVIDE (Lu et al. 2024), performs intra-view and inter-view contrastive learning separately to better preserve both view-specific and shared information. To handle incomplete views, SCVT (Dong et al. 2025a) employs a selective alignment strategy, using a view-topology graph to guide contrastive learning exclusively between adjacent views, thereby avoiding forced homogenization and enabling efficient information transfer.

Despite their success, a fundamental limitation of many deep MVC methods is their implicit handling of shared and private information. By prioritizing cross-view consistency, these models often inadvertently suppress valuable view-

specific characteristics, treating them as noise. This leads to feature entanglement, where the shared representation becomes contaminated with private signals, diminishing its purity and discriminability. While recent works like Multi-VAE (Xu et al. 2021b) and DSIMVC (Tang and Liu 2022) attempt explicit disentanglement, their reliance on simplified objectives or structural constraints may not achieve a comprehensive separation, thus motivating our more principled framework.

Method

This section details the proposed DC-SPAN framework, which addresses feature entanglement via a disentangle-then-fuse paradigm. As illustrated in Figure 2, our architecture first decomposes each view's representation into shared and private components. These factors are then integrated to produce a final, discriminative embedding, a process guided by a composite learning objective detailed in the following subsections.

Problem Formulation

Given a multi-view dataset comprising N samples across V views, denoted as $\{\mathbf{x}_i^{(v)} \in \mathbb{R}^{d_v}\}_{i=1,v=1}^{N,V}$, where d_v is the dimensionality of the v-th view, our objective is to learn a compact and informative latent representation $\mathbf{z}_i \in \mathbb{R}^d$ for each sample i. Here, d denotes the dimension of the final fused latent space, which is shared by both the private and shared representations and is used for downstream clustering.

Disentanglement Module

To explicitly counteract feature entanglement, we design a dual-path architecture that decomposes each view's representation into distinct shared and private subspaces. By processing view-invariant and view-specific information through separate pathways, this architectural separation ensures that a pure consensus representation is learned without being contaminated by private characteristics. This approach preserves the integrity of unique features within each view, which is critical for robust downstream clustering.

Private Representation Learning For each view v, a view-specific private encoder $\mathrm{Enc}_{\text{priv}}^{(v)}$ maps the input $\mathbf{x}^{(v)}$ to the parameters of a diagonal Gaussian distribution:

$$\mu_{\text{priv}}^{(v)},\ \sigma_{\text{priv}}^{(v)2} = \mathrm{Enc}_{\text{priv}}^{(v)}(\mathbf{x}^{(v)}), \tag{1}$$

where $\mu_{\text{priv}}^{(v)},\ \sigma_{\text{priv}}^{(v)2} \in \mathbb{R}^d$. The private latent variable is sampled as $\mathbf{z}_{\text{priv}}^{(v)} \sim \mathcal{N}(\mu_{\text{priv}}^{(v)},\ \mathrm{diag}((\sigma_{\text{priv}}^{(v)})^2))$ using the reparameterization trick. These private embeddings capture modality-specific variations and are encouraged to be uncorrelated across views.

To further enhance the robustness and consensus of private information, we introduce a PoE fusion mechanism. Given all views' private posteriors $\{q_{\text{priv}}^{(v)}(\mathbf{z})\}_{v=1}^{V}$, the fused private posterior is computed as:

$$q_{\text{PoE,priv}}(\mathbf{z}) \propto \left(\prod_{v=1}^{V} q_{\text{priv}}^{(v)}(\mathbf{z})\right) \cdot q_{\text{disc,priv}}(\mathbf{z}), \tag{2}$$

where $q_{\text{disc,priv}}(\mathbf{z})$ denotes an optional discriminative expert, such as a clustering objective or regularization term, incorporated in the latent space. This fusion yields a private latent $\mathbf{z}_{\text{priv}}$ that aggregates view-specific information in a probabilistically principled manner, while allowing the injection of task-specific constraints.

Shared Representation Learning Similarly, to extract view-invariant semantics, each input $\mathbf{x}^{(v)}$ is first projected to a common space and then encoded by a shared encoder $\mathrm{Enc}_{\text{shared}}$:

$$\mu_{\text{shared}}^{(v)},\ \sigma_{\text{shared}}^{(v)2} = \mathrm{Enc}_{\text{shared}}(\mathrm{Proj}^{(v)}(\mathbf{x}^{(v)})), \tag{3}$$

where $\mu_{\text{shared}}^{(v)},\ \sigma_{\text{shared}}^{(v)2} \in \mathbb{R}^d$. Each view thus provides a probabilistic estimate of the shared latent variable.

We again employ the PoE principle to fuse these shared posteriors:

$$q_{\text{PoE,shared}}(\mathbf{z}) \propto \left(\prod_{v=1}^{V} q_{\text{shared}}^{(v)}(\mathbf{z})\right) \cdot q_{\text{disc,shared}}(\mathbf{z}), \tag{4}$$

where $q_{\text{shared}}^{(v)}(\mathbf{z}) = \mathcal{N}(\mathbf{z};\ \mu_{\text{shared}}^{(v)},\ \mathrm{diag}((\sigma_{\text{shared}}^{(v)})^2))$ and $q_{\text{disc,shared}}(\mathbf{z})$ is a discriminative expert that can encode additional semantic or clustering constraints. The resulting fused shared posterior $q_{\text{PoE,shared}}(\mathbf{z})$ is sharply peaked in regions where all experts agree, thus providing a consensus-driven shared representation.

This PoE-based fusion for both private and shared latent variables allows the model to effectively integrate multi-view information in a probabilistic manner, resulting in robust latent representations suitable for clustering.

Gated Attention Fusion

To effectively integrate the complementary information from the shared and private representations, we introduce a gated attention fusion module to dynamically weight shared and private features and form the final clustering embedding.

Given the shared representation $\mathbf{z}_{\text{shared}} \in \mathbb{R}^d$ and the private representation $\mathbf{z}_{\text{priv}} \in \mathbb{R}^d$, we first concatenate them to form a joint context vector:

$$\mathbf{z}_{\text{concat}} = [\mathbf{z}_{\text{shared}}; \mathbf{z}_{\text{priv}}] \tag{5}$$

where $[\cdot;\cdot]$ denotes the concatenation operation, $\mathbf{z}_{\text{concat}} \in \mathbb{R}^{2d}$.

This context vector is then passed through a two-layer MLP, which serves as a gating network, to generate a learnable weight vector $\mathbf{w}$:

$$\mathbf{w} = \sigma(\mathbf{W}_2(\mathrm{ReLU}(\mathbf{W}_1\mathbf{z}_{\text{concat}} + \mathbf{b}_1)) + \mathbf{b}_2) \tag{6}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times 2d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d\times d}$ are learnable weight matrices, $\mathbf{b}_1 \in \mathbb{R}^d$ and $\mathbf{b}_2 \in \mathbb{R}^d$ are the corresponding learnable bias vectors, and $\sigma(\cdot)$ is the sigmoid function. The sigmoid activation ensures that the elements of the gating vector $\mathbf{w}$ are constrained to the range $[0, 1]$.

The final fused representation $\mathbf{z}$ is computed via a channel-wise weighted sum:

$$\mathbf{z} = \mathbf{w} \odot \mathbf{z}_{\text{shared}} + (1 - \mathbf{w}) \odot \mathbf{z}_{\text{priv}} \tag{7}$$

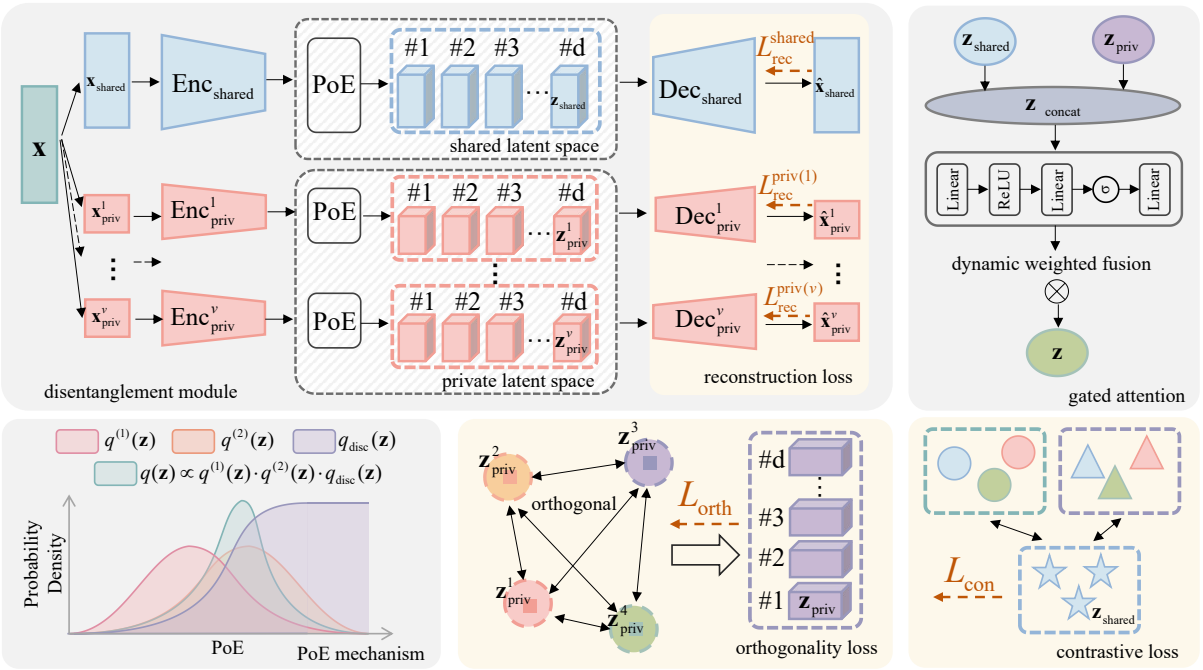


Figure 2: The DC-SPAN framework. To combat feature entanglement, a dual-path architecture disentangles each view into shared and private representations, which are aggregated via a PoE model. A gated attention mechanism then fuses these factors into a final embedding $\mathbf{z}$. The model is trained with a composite objective of reconstruction, contrastive and orthogonality losses.

where $\odot$ denotes the element-wise product. This formulation implements a channel-wise soft selection by forming a convex combination of the shared and private representations. The gating vector $\mathbf{w}$ thus determines the balance, where values approaching 1 favor the shared features and values approaching 0 favor the private features.

This adaptive fusion strategy is crucial for modulating the information flow from the shared and private pathways, ultimately producing a more robust and discriminative representation for clustering.

Learning Objectives

To guide the learning process, we formulate a composite objective function comprising three key terms: a reconstruction loss to preserve information, an orthogonality loss to enforce disentanglement, and a dual contrastive loss to ensure alignment and regularization.

Reconstruction Loss. To ensure the latent representations are informative, we define a reconstruction loss based on reconstructing each input view $\mathbf{x}^{(v)}$ from its shared and private components:

$$\mathcal{L}_{\text{rec}} = \frac{1}{2V} \sum_{v=1}^{V} \left(\|\mathbf{x}^{(v)} - \hat{\mathbf{x}}^{(v)}_{\text{shared}}\|_2^2 + \|\mathbf{x}^{(v)} - \hat{\mathbf{x}}^{(v)}_{\text{priv}}\|_2^2 \right), \tag{8}$$

where $\hat{\mathbf{x}}^{(v)}_{\text{shared}}$ and $\hat{\mathbf{x}}^{(v)}_{\text{priv}}$ are the reconstructions generated by view-specific decoders from $\mathbf{z}_{\text{shared}}$ and $\mathbf{z}^{(v)}_{\text{priv}}$, respectively.

Orthogonality Loss. To explicitly enforce the separation of private information between different views, we impose an orthogonality constraint. This loss minimizes the correlation between the private latent representations of any two distinct views:

$$\mathcal{L}_{\text{orth}} = \sum_{i \neq j} \|(\mathbf{z}^{(i)}_{\text{priv}})^\top \mathbf{z}^{(j)}_{\text{priv}}\|_F^2. \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Dual Contrastive Loss. We introduce a dual contrastive loss that operates on both the shared and private representations to enforce semantic structure. Specifically, for a batch of N samples, we treat the shared representations $\{\mathbf{z}_{\text{shared},i}\}_{i=1}^N$ of all samples as a batch and apply an InfoNCE-style contrastive loss, where positive pairs are different views of the same sample and negatives are other samples in the batch. The same applies to the private representations $\{\mathbf{z}_{\text{priv},i}\}_{i=1}^N$. Thus, the contrastive loss is defined as:

$$\mathcal{L}_{\text{con}} = \sum_{i=1}^{N} -\log \frac{\exp(\text{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(i^+)})/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(k)})/\tau)}, \tag{10}$$

where $\mathbf{z}^{(i^+)}$ denotes the positive pair of $\mathbf{z}^{(i)}$ (i.e., another view of the same sample), $\text{sim}(\cdot)$ is the cosine similarity, and τ is a temperature hyperparameter.

Overall Objective. The final learning objective is a weighted sum of the individual loss components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{orth}} + \lambda_2 \mathcal{L}_{\text{con}}, \tag{11}$$

where λ_1, λ_2 are hyperparameters that balance the contributions of the regularizers.

Algorithm 1: The proposed DC-SPAN

Input: Multi-view data $\{\mathbf{x}_i^{(v)}\}_{i=1,v=1}^{N,V}$; model parameters Θ; hyperparameters λ_1, λ_2; training epochs $T_{\text{pre}}, T_{\text{align}}$
Output: Cluster assignments $\{\mathbf{y}_i\}_{i=1}^{N}$

Stage I: Pre-train with reconstruction loss
for $t = 1$ **to** T_{pre} **do**
    Encode each view to obtain private and shared latents
    Fuse latents and reconstruct inputs
    Update Θ by minimizing reconstruction loss $\mathcal{L}_{\text{rec}}$
end for
Stage II: Fine-tune with full objective
for $t = 1$ **to** T_{align} **do**
    Encode, fuse, and reconstruct as above
    Compute $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{orth}}$, $\mathcal{L}_{\text{con}}$
    Update Θ by minimizing $\mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{orth}} + \lambda_2 \mathcal{L}_{\text{con}}$
end for
Obtain cluster assignments $\{\mathbf{y}_i\}_{i=1}^{N}$ by applying a clustering algorithm (e.g., K-means) on the fused representations
return $\{\mathbf{y}_i\}_{i=1}^{N}$

Optimization Strategy

To improve convergence and disentanglement, we employ a two-stage training strategy. In the first stage, the model is pre-trained using only the reconstruction loss $\mathcal{L}_{\text{rec}}$ to initialize the latent space. In the second stage, we fine-tune the model with the full objective $\mathcal{L}_{\text{total}}$, incorporating orthogonality and alignment constraints for more discriminative representations. This approach stabilizes training and consistently yields better results. The full procedure is detailed in Algorithm 1.

Experiments

In this section, we empirically validate our proposed framework through a series of comprehensive experiments. We first detail the experimental setup, then present a quantitative comparison against state-of-the-art methods, and conclude with in-depth ablation studies and parameter analyses to verify the contribution of each key component.

Experimental Setup

Datasets. We evaluate our model on five widely used multi-view benchmark datasets, summarized in Table 1.

Dataset	Samples	Views	Classes	Dimensionality
ForestTypes	523	3	4	9 / 9 / 9
HW_6Views	2,000	6	10	216 / 76 / 64 / 6 / 240 / 47
Synthetic3d	600	3	3	3 / 3 / 3
Caltech101-7	1474	6	7	48/ 40/ 254/ 1984/ 512/ 928
Wiki	2866	2	10	128 / 10

Table 1: Multi-view datasets in experiments.

Evaluation Metrics. To quantitatively assess the clustering performance, we employ six standard evaluation metrics: Accuracy(ACC) (Xu, Liu, and Gong 2003), Normalized Mutual Information (NMI) (Strehl and Ghosh 2002), Purity (PUR) (Manning 2009), F-score (Powers 2020) , Precision and Recall (Fisher 1936). For all these metrics, a higher value indicates better clustering performance.

Implementation Details. Our model is implemented in PyTorch and all experiments are conducted on a single NVIDIA A100 GPU. We adopt a two-stage optimization strategy, beginning with a pre-training phase focused on reconstruction, followed by a fine-tuning phase using the full objective function. The Adam optimizer is used for both stages. For reproducibility, all experiments are conducted with a fixed random seed.

Comparison Methods. We compare our proposed method against a range of state-of-the-art multi-view clustering algorithms, including both traditional and deep learning-based approaches. Specifically, our baselines include the traditional method LMVSC (Kang et al. 2020) and a suite of deep learning-based approaches, including AE^2-Nets (Zhang, Liu, and Fu 2019), DEMVC (Xu et al. 2021a), SDSNE (Liu et al. 2022), SDMVC (Xu et al. 2022a), DDMvC (Chen et al. 2025b), DFL-NET (Chen et al. 2025a), 3MC (Chen et al. 2025c), and SSLNMVC (Yan, Yang, and Tang 2025).

Result Analysis

Comprehensive evaluations across five benchmark datasets demonstrate the superior performance and robustness of our proposed DC-SPAN framework. As summarized in Table 2, DC-SPAN consistently achieves state-of-the-art or highly competitive results.

On the ForestTypes, HW_6Views, and Wiki datasets, DC-SPAN establishes a clear performance advantage, outperforming the strongest baselines by a significant margin, often in the range of 1% to 10.8%. This underscores our model's enhanced capability to handle cross-view inconsistency. Even on the more structured Synthetic3d dataset, where most deep methods perform well, DC-SPAN maintains a stable lead, further demonstrating its robustness and competitiveness in low-noise environments. On the Caltech101-7 dataset, while the overall performance of all methods is challenged by its complexity, DC-SPAN still secures the best results on most metrics, maintaining a consistent edge over its competitors. This indicates that even in complex image-based scenarios, our learned representations are more semantically discriminative.

These empirical results strongly validate our disentangle-then-fuse paradigm. By explicitly decomposing latent factors and then adaptively integrating them using our tripart objective and gated attention mechanism, DC-SPAN effectively mitigates feature entanglement. Consequently, the model learns a final representation that better captures both cross-view consistency and view-specific diversity, leading to not only higher clustering accuracy but also greater stability and interpretability.

Dataset	Categories	Methods	ACC	NMI	PUR	F-score	Precision	Recall
ForestTypes	Traditional	LMVSC	0.7132	0.5562	0.8050	0.6727	0.6653	0.6803
	Deep	AE^2-Nets	0.4516	0.2205	0.4985	0.4391	0.3975	0.4905
		SDSNE	0.8069	0.5844	0.8069	0.8090	0.8156	0.8136
		DEMVC	0.3728	0.0000	0.3728	0.2025	0.1390	0.2500
		SDMVC	0.5774	0.3303	0.3303	0.5452	0.6666	0.6157
		DDMvC	0.6902	0.4577	0.6902	0.6901	0.7525	0.7360
		DFL-NET	0.7017	0.4717	0.7017	0.7023	0.7592	0.7566
		3MC	0.5430	0.4468	0.6214	0.5162	0.5744	0.5394
		SSLNMVC	0.5086	0.3185	0.5086	0.0983	0.0701	0.1352
		DC-SPAN	**0.8298**	**0.6012**	**0.8298**	**0.8299**	**0.8368**	**0.8314**
HW_6Views	Traditional	LMVSC	0.8510	0.8056	0.8510	0.7723	0.7694	0.7752
	Deep	AE^2-Nets	0.8782	0.8018	0.8782	0.7935	0.7920	0.7949
		SDSNE	0.8760	**0.8775**	0.8760	0.8783	0.8812	0.8760
		DEMVC	0.2745	0.2939	0.2745	0.1766	0.2422	0.2745
		SDMVC	0.4945	0.5190	0.4985	0.4210	0.4596	0.4945
		DDMvC	0.7945	0.7256	0.7945	0.7761	0.8034	0.7945
		DFL-NET	0.7030	0.6683	0.7030	0.6952	0.7059	0.7030
		3MC	0.7845	0.7208	0.7845	0.7864	0.7919	0.7845
		SSLNMVC	0.8005	0.7815	0.8005	0.8007	0.8011	0.8005
		DC-SPAN	**0.9280**	0.8600	**0.9280**	**0.9281**	**0.9297**	**0.9280**
Synthetic3d	Traditional	LMVSC	0.9567	0.8309	0.9567	0.9161	0.9187	0.9197
	Deep	AE^2-Nets	0.6518	0.5045	0.6572	0.6802	0.5860	0.8161
		SDSNE	0.7867	0.4367	0.7867	0.7891	0.8075	0.7867
		DEMVC	0.7150	0.5481	0.7150	0.6953	0.7442	0.7150
		SDMVC	0.9733	0.8819	0.9733	0.9733	0.9733	0.9733
		DDMvC	0.9183	0.7347	0.9183	0.9182	0.9188	0.9183
		DFL-NET	0.9517	0.8156	0.9517	0.9517	0.9521	0.9517
		3MC	0.7850	0.5398	0.7850	0.7782	0.8230	0.7850
		SSLNMVC	0.9617	0.8466	0.9617	0.9615	0.9618	0.9617
		DC-SPAN	**0.9750**	**0.8941**	**0.9750**	**0.9749**	**0.9753**	**0.9750**
Caltech101-7	Traditional	LMVSC	0.5684	0.4574	0.8220	0.5531	0.6965	0.4586
	Deep	AE^2-Nets	0.4468	0.3149	0.7573	0.4622	0.4622	0.3594
		SDSNE	0.6716	**0.6325**	0.8630	0.7322	0.8728	0.4218
		DEMVC	0.5441	0.1670	0.6269	0.5060	0.5715	0.2122
		SDMVC	0.5319	0.6094	0.8684	0.6092	0.8627	0.4073
		DDMvC	0.3758	0.5376	0.8575	0.4795	0.8492	0.3305
		DFL-NET	0.4776	0.5412	0.8507	0.5703	0.8119	0.3303
		3MC	0.4091	0.3948	0.7883	0.5144	0.8471	0.3294
		SSLNMVC	0.4084	0.5325	0.8453	0.5117	0.8477	0.3428
		DC-SPAN	**0.6723**	0.6133	**0.8725**	**0.7360**	**0.8777**	**0.4914**
Wiki	Traditional	LMVSC	0.5780	0.5420	0.6304	0.5011	0.4972	0.5051
	Deep	AE^2-Nets	0.4533	0.4304	0.5241	0.3990	0.4234	0.3772
		SDSNE	0.5607	0.5268	0.6183	0.5553	0.5980	0.5359
		DEMVC	0.2544	0.2409	0.3126	0.2317	0.2738	0.2324
		SDMVC	0.2408	0.1178	0.2711	0.2409	0.2523	0.2409
		DDMvC	0.5803	0.4865	0.6190	0.5864	0.6115	0.5700
		DFL-NET	0.2987	0.2027	0.3029	0.2706	0.2995	0.3049
		3MC	0.5024	0.4762	0.5639	0.5216	0.5709	0.4768
		SSLNMVC	0.3897	0.2937	0.3974	0.3596	0.3637	0.3576
		DC-SPAN	**0.6403**	**0.5477**	**0.6469**	**0.6530**	**0.6776**	**0.6255**

Table 2: Clustering performance comparison on five benchmark datasets. Methods are categorized into traditional and deep learning-based approaches. The best results are highlighted in **bold**, and the second-best are underlined.

Ablation Study on Modules

To validate the contribution of each key component, we conducted a detailed ablation study on the ForestTypes and Synthetic3d datasets, with results summarized in Tables 3 and 4. The findings confirm that each module is integral to the model's performance. Removing the contrastive loss $\mathcal{L}_{\text{con}}$ induced the most significant degradation, with the NMI dropping by over 30.9% on Synthetic3d and 18.6% on Forest-

Types, underscoring its critical role in learning aligned representations. Deactivating the orthogonality loss $\mathcal{L}_{\text{orth}}$ also led to a marked decline, validating its importance for effective disentanglement. Furthermore, replacing the PoE fusion with a simple averaging baseline caused a sharp accuracy drop of nearly 24.6% on Synthetic3d, highlighting the necessity of robust probabilistic aggregation. The consistent superiority of the full model over all ablated variants confirms that our integrated design, where each component synergistically contributes, is key to its success.

$\mathcal{L}_{rec}$	$\mathcal{L}_{orth}$	$\mathcal{L}_{con}$	PoE	ACC	NMI	F-score
✓			✓	0.7539	0.4310	0.7633
✓	✓		✓	0.7386	0.4153	0.7357
✓		✓	✓	0.7826	0.4741	0.6858
✓	✓	✓		0.6258	0.4797	0.7932
✓	✓	✓	✓	**0.8298**	**0.6012**	**0.8299**

Table 3: Ablation study results on ForestTypes. The best results are highlighted in **bold**.

$\mathcal{L}_{rec}$	$\mathcal{L}_{orth}$	$\mathcal{L}_{con}$	PoE	ACC	NMI	F-score
✓			✓	0.9283	0.7521	0.9282
✓	✓		✓	0.9183	0.7282	0.9177
✓		✓	✓	0.9517	0.8276	0.9515
✓	✓	✓		0.9483	0.8138	0.9484
✓	✓	✓	✓	**0.9750**	**0.8941**	**0.9749**

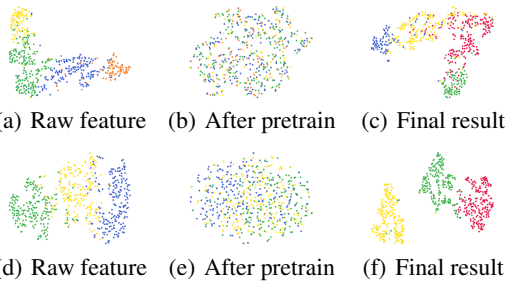
Table 4: Ablation study results on Synthetic3d. The best results are highlighted in **bold**.

Visualization

To qualitatively evaluate the learned representations, we visualize the fused embeddings using t-SNE on the ForestTypes and Synthetic3d datasets at different training stages. As shown in Figure 3, the raw features lack clear cluster structure, and pre-training alone does not yield discriminative embeddings. In contrast, after full training with our composite objective, the final embeddings produced by our full model form distinct, well-separated clusters that align closely with the ground-truth classes. These results intuitively demonstrate the effectiveness of our method in learning representations suitable for clustering.

Parameter Sensitivity Analysis

The final objective function of our model, shown in Equation (12), is balanced by two key hyperparameters, λ_1 and λ_2. To investigate their impact, we conduct a sensitivity analysis on the ForestTypes and Synthetic3d datasets. As illustrated in Figure 4, while the clustering accuracy exhibits some fluctuation as the parameters change, the overall performance trend remains stable across a wide range of values, demonstrating the model's robustness. These results indicate that the model is not overly sensitive to the precise weighting

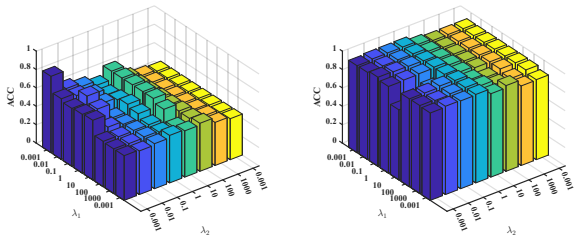


(a) Raw feature (b) After pretrain (c) Final result

(d) Raw feature (e) After pretrain (f) Final result

Figure 3: t-SNE visualization of the learned embeddings at different stages on ForestTypes (top row) and Synthetic3d (bottom row) datasets.

of its core objectives, provided they are within a reasonable range. This robustness is practically advantageous, as it alleviates the need for exhaustive hyperparameter tuning, further highlighting the well-posed nature of our proposed framework.



(a) ForestTypes (b) Synthetic3d

Figure 4: Parameter sensitivity analysis of hyperparameters λ_1 and λ_2 on the ForestTypes and Synthetic3d datasets.

Conclusion

In this paper, we introduced DC-SPAN, a novel deep multi-view clustering framework that mitigates feature entanglement via a disentangle-then-fuse paradigm. Our model employs a dual-path variational architecture with a PoE mechanism to robustly separate and aggregate shared and private representations. This process is guided by a composite objective combining reconstruction, orthogonality and contrastive losses to ensure structured disentanglement and semantic alignment. A gated attention mechanism then adaptively fuses these factors into a discriminative embedding. Extensive experiments validate that DC-SPAN significantly outperforms state-of-the-art methods, confirming that our approach offers a more robust and interpretable solution for multi-view representation learning.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant No. 2022YFB2803405.

References

Chen, Z.; Wu, X.-J.; Xu, T.; and Kittler, J. 2023. Fast self-guided multi-view subspace clustering. *IEEE transactions on image processing*, 32: 6514–6525.

Chen, Z.; Wu, X.-J.; Xu, T.; and Kittler, J. 2025a. DFL-Net: Disentangled Feature Learning Network for Multi-view Clustering. *IEEE Transactions on Knowledge and Data Engineering*.

Chen, Z.; Wu, X.-J.; Xu, T.; Li, H.; and Kittler, J. 2025b. Deep Discriminative Multi-view Clustering. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chen, Z.; Wu, X.-J.; Xu, T.; Li, H.; and Kittler, J. 2025c. Multi-layer multi-level comprehensive learning for deep multi-view clustering. *Information Fusion*, 116: 102785.

Cui, C.; Ren, Y.; Pu, J.; Pu, X.; and He, L. 2023. Deep multi-view subspace clustering with anchor graph. *arXiv preprint arXiv:2305.06939*.

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

Dong, Z.; Jin, J.; Xiao, Y.; Wang, S.; Zhu, X.; Liu, X.; and Zhu, E. 2023a. Iterative deep structural graph contrast clustering for multiview raw data. *IEEE Transactions on Neural Networks and Learning Systems*.

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

Dong, Z.; Wang, S.; Jin, J.; Liu, X.; and Zhu, E. 2023b. Cross-view topology based consistent and complementary information for deep multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19440–19451.

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*, 33: 3399–3412.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2): 179–188.

Gu, Z.; and Feng, S. 2024. From dictionary to tensor: A scalable multi-view subspace clustering framework with triple information enhancement. *Advances in Neural Information Processing Systems*, 37: 103545–103573.

Gu, Z.; Li, Z.; and Feng, S. 2024. Topology-driven multi-view clustering via tensorial refined sigmoid rank minimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 920–931.

Hu, S.; Zou, G.; Zhang, C.; Lou, Z.; Geng, R.; and Ye, Y. 2023. Joint contrastive triple-learning for deep multi-view clustering. *Information Processing & Management*, 60(3): 103284.

Huang, S.; Du, S.; Fu, L.; Wu, Z.; and Wang, S. 2024a. Tensor-derived large-scale multi-view subspace clustering with faithful semantics. *IEEE Transactions on Signal and Information Processing over Networks*, 10: 584–598.

Huang, S.; Fu, L.; Du, S.; Wu, Z.; Vasilakos, A. V.; and Wang, S. 2025. Low-rank tensor learning with projection distance metric for multi-view clustering. *International Journal of Machine Learning and Cybernetics*, 16(1): 25–41.

Huang, Y.-D.; Zhang, G.-Y.; Huang, D.; Wang, C.-D.; Liu, Y.; and Huang, E. 2024b. Confidence-oriented Contrastive Graph Clustering. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

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, Z.; Wang, Q.; Tao, Z.; Gao, Q.; Yang, Z.; et al. 2019. Deep adversarial multi-view clustering network. In *IJCAI*, volume 2, 4.

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.

Lin, Y.; Wang, Y.; Lyu, G.; Deng, Y.; Cai, H.; Lin, H.; Wang, H.; and Yang, Z. 2025. Enhance Multi-View Classification Through Multi-Scale Alignment and Expanded Boundary. In *The Thirteenth International Conference on Learning Representations*.

Liu, C.; Liao, Z.; Ma, Y.; and Zhan, K. 2022. Stationary diffusion state neural estimation for multiview clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 7542–7549.

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

Manning, C. D. 2009. *An introduction to information retrieval*.

McQueen, J. B. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, 281–297.

Powers, D. M. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

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 Thirty-Third International Joint Conference on Artificial Intelligence*, 4842–4850.

Shi, L.; Cao, L.; Wang, J.; and Chen, B. 2024. Enhanced latent multi-view subspace clustering. *IEEE Transactions on Circuits and Systems for Video Technology*.

Strehl, A.; and Ghosh, J. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec): 583–617.

Sun, Y.; Dai, J.; Ren, Z.; Li, Q.; and Peng, D. 2024a. Relaxed energy preserving hashing for image retrieval. *IEEE Transactions on Intelligent Transportation Systems*, 25(7): 7388–7400.

Sun, Y.; Peng, D.; and Ren, Z. 2024. Discrete aggregation hashing for image set classification. *Expert Systems with Applications*, 237: 121615.

Sun, Y.; Qin, Y.; Li, Y.; Peng, D.; Peng, X.; and Hu, P. 2024b. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*.

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26: 824–836.

Tang, H.; and Liu, Y. 2022. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International conference on machine learning*, 21090–21110. PMLR.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *European conference on computer vision*, 776–794. Springer.

Trosten, D. J.; Lokse, S.; Jenssen, R.; and Kampffmeyer, M. 2021. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1255–1265.

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

Wang, J.; Qu, J.; Wang, K.; Li, Z.; Hua, W.; Li, X.; and Liu, A. 2024. Improving the robustness of knowledge-grounded dialogue via contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19135–19143.

Wang, Q.; Tao, Z.; Gao, Q.; and Jiao, L. 2022a. Multi-view subspace clustering via structured multi-pathway network. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5): 7244–7250.

Wang, Q.; Tao, Z.; Xia, W.; Gao, Q.; Cao, X.; and Jiao, L. 2022b. Adversarial multiview clustering networks with adaptive fusion. *IEEE transactions on neural networks and learning systems*, 34(10): 7635–7647.

Wang, Q.; Zhang, Z.; Feng, W.; Tao, Z.; and Gao, Q. 2025. Contrastive Multi-view Subspace Clustering via Tensor Transformers Autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21207–21215.

Wang, S.; Lin, X.; Fang, Z.; Du, S.; and Xiao, G. 2022c. Contrastive consensus graph learning for multi-view clustering. *IEEE/CAA Journal of Automatica Sinica*, 9(11): 2027–2030.

Wen, J.; Zhang, Z.; Xu, Y.; and Zhong, Z. 2018. Incomplete multi-view clustering via graph regularized matrix factorization. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Wen, J.; Zhang, Z.; Zhang, Z.; Wu, Z.; Fei, L.; Xu, Y.; and Zhang, B. 2020. Dimc-net: Deep incomplete multi-view clustering network. In *Proceedings of the 28th ACM international conference on multimedia*, 3753–3761.

Wen, Z.; Wu, T.; Ren, Y.; Ling, Y.; Cui, C.; Pu, X.; and He, L. 2024. Dual-Optimized Adaptive Graph Reconstruction for Multi-View Graph Clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1819–1828.

Xu, J.; Ren, Y.; Li, G.; Pan, L.; Zhu, C.; and Xu, Z. 2021a. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573: 279–290.

Xu, J.; Ren, Y.; Tang, H.; Pu, X.; Zhu, X.; Zeng, M.; and He, L. 2021b. Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9234–9243.

Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; Pu, X.; Yu, P. S.; and He, L. 2022a. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7): 7470–7482.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022b. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16051–16060.

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 267–273.

Yan, W.; Yang, T.; and Tang, C. 2025. Self-supervised Semantic Soft Label Learning Network for Deep Multi-view Clustering. *IEEE Transactions on Multimedia*.

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.

Zhang, J.; Li, L.; Wang, S.; Liu, J.; Liu, Y.; Liu, X.; and Zhu, E. 2022. Multiple kernel clustering with dual noise minimization. In *Proceedings of the 30th ACM international conference on multimedia*, 3440–3450.

Zhang, P.; Wang, S.; Hu, J.; Cheng, Z.; Guo, X.; Zhu, E.; and Cai, Z. 2020. Adaptive weighted graph fusion incomplete multi-view subspace clustering. *Sensors*, 20(20): 5755.

Zhu, P.; Yao, X.; Wang, Y.; Hui, B.; Du, D.; and Hu, Q. 2024. Multiview deep subspace clustering networks. *IEEE Transactions on Cybernetics*, 54(7): 4280–4293.